

# **Building Bridges Across Cognitive Sciences Around the World**

## **Proceedings of the 34<sup>th</sup> Annual Meeting of the Cognitive Science Society**

Sapporo, Japan, August 1-4, 2012

*Edited by*

Naomi Miyake  
*University of Tokyo*

David Peebles  
*University of  
Huddersfield*

Richard P. Cooper  
*Birkbeck,  
University of London*



Austin, TX: Cognitive Science Society

ISBN: 978-0-9768318-8-4



*How to cite a paper in these Proceedings:*

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

Author, A. & Author, B. (2012). This is the title of the paper. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. NUMBERS). Austin, TX: Cognitive Science Society.

**APA formatted citation for a published abstract:**

Author, A. & Author, B. (2012). This is the title of the abstract [Abstract]. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (p. NUMBER). Austin, TX: Cognitive Science Society.

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

Author, A. & Author, B. (2012, August). This is the title of the talk or poster. Paper (or Poster) presented at the 34th Annual Conference of the Cognitive Science Society. Sapporo, Japan.

# Table of Contents

<b>Introduction</b>	1
<b>Organizing Committee</b>	3
<b>Program Committee</b>	4
<b>Sponsors</b>	5
<b>Conference Awards</b>	6
<b>Glusko Dissertation Prizes and NSF Fellowships</b>	8
<b>Invited Plenary Presentations</b>	9
<b>Workshops</b>	
<i>Women in Cognitive Science sponsored interactive panel discussion: Professional advancement, leadership and international collaboration</i>	10
Laurie Beth Feldman, Janet van Hell, Judith Kroll, Suparna Rajaram	
<i>Teleoperated Android as a Tool for Cognitive Studies, Communication and Art</i>	12
Shuichi Nishio, Hiroshi Ishiguro	
<i>Modeling the Perception of Intentions</i>	14
Barbara Tversky, Shimon Ullman, Dare Baldwin, Frank E. Pollick, Joshua Tenenbaum, Tao Gao, Peter Pantelis, David Pautler	
<b>Tutorials</b>	
<i>And Now for Something Completely Different: Python in Cognitive Science</i>	16
Mark Andrews, Jesse Diaz	
<i>Full day tutorial on Quantum models of cognition and decision</i>	18
Jerome Busemeyer, Peter Bruza, Taiki Takahashi, Jennifer Trueblood	
<i>Using Bayes to Interpret Non-significant Results</i>	20
Zoltan Dienes	
<i>Nengo and the Neural Engineering Framework: From Spikes to Cognition</i>	22
Chris Eliasmith, Terrence Stewart	
<i>Probability, programs, and the mind: Building structured Bayesian models of cognition</i>	24
Noah Goodman, Joshua Tenenbaum	
<i>Using Machine Learning for Exploratory Data Analysis</i>	26
Joshua Lewis, Virginia de Sa	
<i>Practical Advice on How to Run Human Behavioral Studies</i>	28
Frank Ritter, Jong kim	

## Symposia

<i>Robotics and Emotion</i> .....	30
Rolf Pfeifer, Hiroshi Ishiguro, Yuichiro Anzai, Naomi Miyake	
<i>Computational Models of Intuitive Physics</i> .....	32
Peter Battaglia, Tomer Ullman, Joshua Tenenbaum, Adam Sanborn, Kenneth Forbus, Tobias Gerstenberg, David Lagnado	
<i>The role of comparison in structure learning: Developmental, learning science, and computational perspectives</i> .....	34
Stella Christie, Dedre Gentner, Mutsumi Imai, Etsuko Haryu, Hiroyuki Okada, Ji Y. Son, James Stigler, Leonidas Doumas, Robert G. Morrison, Lindsey E. Richland	
<i>Neural Computations Supporting Cognition: Rumelhart Prize Symposium in Honor of Peter Dayan</i>	36
Kenji Doya, John O'Doherty, Alexandre Pouget, Peter Bossaerts, Nathaniel Daw, Yael Niv	
<i>Thirty years of Marr's Vision: Levels of Analysis in Cognitive Science</i> .....	38
Chris Eliasmith, Tom Griffiths, Valerie Gray Hardcastle, Brad Love, William Bechtel, Richard P. Cooper, David Peebles	
<i>New Frontiers in Computational Models of Grammatical Development</i> .....	40
Micah Goldwater, Scott Friedman, Dedre Gentner, Kenneth Forbus, Cynthia Fisher, Michael Connor, Dan Roth, Franklin Chang, Gary Dell	
<i>Computational, Cognitive, and Neural Models of Decision-making Biases</i> .....	42
Jonathan Malmaud, Joshua Tenenbaum, Peter Dayan, Laurence Maloney, Edward Vul, Nick Chater	
<i>Governing Board Symposium: Cognitive Science and the Learning Sciences</i> .....	44
Naomi Miyake, Roy Pea, Reinhart Pekrun, Kurt Van Lehn, Richard Catrambone, Stella Vosniadou	
<i>How Vertical Spaces Are Perceived and Represented</i> .....	46
Daniele Nardi, Frank Durgin, Kate Jeffery, Steven Weisberg	
<i>What Can Cognitive Science Say or Learn about Economic Crises?</i> .....	48
Magda Osman, Björn Meder, Gerd Gigerenzer, Nick Chater, Daniel Read, Hansjörg Neth	
<i>Dynamic decision making: neuronal, computational, and cognitive underpinnings</i> .....	50
Magda Osman, Maarten Speekenbrink, Peter Dayan, Masataka Watanabe, Nigel Harvey	
<i>Grammatical Approaches to Written and Graphical Communication</i> .....	52
Colin Wilson, Neil Cohn, James Myers, Stephen Goldberg, Ariel Cohen-Goldberg	

## Papers

<i>Constructing a hypothesis space from the Web for large-scale Bayesian word learning</i> .....	54
Joshua Abbott, Joseph Austerweil, Tom Griffiths	
<i>Predicting focal colors with a rational model of representativeness</i> .....	60
Joshua Abbott, Terry Regier, Tom Griffiths	
<i>Gestures in Communication through Line Graphs</i> .....	66
Cengiz Acarturk, Ozge Alacam	

<i>Distributional Learning of Vowel Categories Is Supported by Prosody in Infant-Directed Speech</i> . . . .	72
Frans Adriaans, Daniel Swingley	
<i>Emotion-Based Reinforcement Learning</i> . . . . .	78
Woo-Young Ahn, Olga Rass, Yong-Wook Shin, Jerome R. Busemeyer, Joshua W. Brown, Brian F. O'Donnell	
<i>Cooperation in risky environments: Decisions from experience in a stochastic social dilemma</i> . . . . .	84
Florian Artinger, Nadine Fleischhut, Vittoria Levati, Jeffrey R. Stevens	
<i>Modeling individual differences in socioeconomic game playing</i> . . . . .	90
Derrik Asher, Shunan Zhang, Andrew Zaldivar, Michael Lee, Jeffrey Krichmar	
<i>Lexical Access Across Languages: A Multinomial Model of Auditory Distraction</i> . . . . .	96
Philip Beaman	
<i>Diagramming Phenomena for Mechanistic Explanation</i> . . . . .	102
William Bechtel, Adele Abrahamsen	
<i>Route choice in individuals—semantic network navigation</i> . . . . .	108
Nicole Beckage, Mark Steyvers, Carter Butts	
<i>Competent Deontic Reasoning: The Abstract Deontic Selection Task Revisited</i> . . . . .	114
Sieghard Beller, Andrea Bender	
<i>That's what she (could have) said: How alternative utterances affect language use</i> . . . . .	120
Leon Bergen, Noah Goodman, Roger Levy	
<i>Word predictability and frequency effects in a rational model of reading</i> . . . . .	126
Klinton Bicknell, Roger Levy	
<i>The retriever-connector model: Matching Classroom Data and Agent-Based Computer Models to Simulate Students' Use of Multiple Epistemological Resources</i> . . . . .	132
Paulo Blikstein	
<i>Inducing Mathematical Concepts from Specific Examples: The Role of Schema-Level Variation</i> . . .	138
David Braithwaite, Robert Goldstone	
<i>Learning to recognize unfamiliar voices: the role of language familiarity and music experience</i> . . . .	144
Micah Bregman, Sarah Creel	
<i>Misconceptions Regarding Emergent Phenomena Vary By Domain</i> . . . . .	150
Sarah Brem, Glenda Stump, Gale Sinatra, Raymond Reichenberg, Benjamin Heddy	
<i>Do I know that you know what you know? Modeling testimony in causal inference</i> . . . . .	156
Daphna Buchsbaum, Sophie Bridgers, Andrew Whalen, Elizabeth Seiver, Tom Griffiths, Alison Gopnik	
<i>Revisiting the Relationship between Allocentric-Heading Recall and Self-Reported Sense of Direction</i> . . . . .	162
Heather Burte, Mary Hegarty	
<i>Space-Time Interdependence and Sensory Modalities: Time Affects Space in the Hand But Not in the Eye</i> . . . . .	168
Zhenguang Cai, Louise Connell	

<i>Sentic Panalogy: Swapping Affective Common Sense Reasoning Strategies and Foci</i> .....	174
Erik Cambria, Daniel Olsher, Kenneth Kwok	
<i>Object discovery and inverse physical reasoning</i> .....	180
Christopher Carroll, Charles Kemp	
<i>Category structure modulates interleaving and blocking advantage in inductive category acquisition</i>	186
Paulo Carvalho, Robert Goldstone	
<i>Implicit Learning of L2 word stress rules</i> .....	192
Ricky Chan, Janny Leung	
<i>Generating Realistic Semantic Codes for Use in Neural Network Models</i> .....	198
Ya-Ning Chang, Steve Furber, Stephen Welbourne	
<i>Word Form Encoding in Mandarin Chinese Typewritten Word Production</i> .....	204
Jenn-Yeu Chen, Train-Min Chen	
<i>What Counts in Mandarin Chinese: A Study of Individuation and Quantification</i> .....	210
Pierina Cheung, Peggy Li, David Barner	
<i>The Role of Preview and Incremental Delivery on Visual Search</i> .....	216
Eric Chiu, Michael Spivey	
<i>Determining Relevance: Close Enough is Good Enough</i> .....	222
Sheldon Chow	
<i>Subject Relative Production in SLI Children during Syntactic Priming and Sentence Repetition</i> ..	228
Moreno Coco, Maria Garraffa, Holly Branigan	
<i>Segmenting Visual Narratives: Evidence for Constituent Structure in Comics</i> .....	234
Neil Cohn, Phillip Holcomb, Ray Jackendoff, Gina Kuperberg	
<i>Framing attention in American and Japanese comics</i> .....	240
Neil Cohn, Amaro Taylor-Weiner, Suzanne Grossman	
<i>Early-Talker and Late-Talker Toddlers and Networks Show Different Word Learning Biases</i> .....	246
Eliana Colunga, Clare Sims	
<i>Do you see what I'm singing? Visuospatial movement biases pitch perception</i> .....	252
Louise Connell, Zhenguang Cai, Judith Holler	
<i>Flexible Shortcuts: Linguistic Distributional Information Affects both Shallow and Deep</i>	
<i>Conceptual Processing</i> .....	258
Louise Connell, Dermot Lynott	
<i>Exploring Decision Rules and Sampling Dynamics in Recognition Memory</i> .....	264
Gregory Cox, Richard Shiffrin	
<i>Reverse appraisal: The importance of appraisals for the effect of emotion displays on people's</i>	
<i>decision making in a social dilemma</i> .....	270
Celso de Melo, Jonathan Gratch, Peter Carnevale, Stephen Read	
<i>Neural correlates of social perception: The posterior superior temporal sulcus is modulated by</i>	
<i>action rationality, but not animacy</i> .....	276
Ben Deen, Rebecca Saxe	

<i>The Role of Linguistic Labels in Infants' Categorization: An Eye Tracking Study</i> .....	282
Sophia Deng, Vladimir Sloutsky	
<i>Learning Deterministic Causal Networks from Observational Data</i> .....	288
Ben Devereitt, Charles Kemp	
<i>Enhanced Performance for Recognition of Irrelevant Target-Aligned Auditory Stimuli: Unimodal and Cross-modal Considerations</i> .....	294
Andrew Dewald, Scott Sinnett	
<i>The semantic structure of sensory vocabulary in an African language</i> .....	300
Mark Dingemanse, Asifa Majid	
<i>The Sound of Thickness: Prelinguistic Infants' Associations of Space and Pitch</i> .....	306
Sarah Dolscheid, Sabine Hunnius, Daniel Casasanto, Asifa Majid	
<i>What explains variability in brain regions associate with Theory of Mind in a large sample of neurotypical adults and adults with ASD?</i> .....	312
Nicholas Dufour, Elizabeth Redcay, Liane Young, Penelope Mavros, Joseph Moran, Christina Triantafyllou, John Gabrieli, Rebecca Saxe	
<i>Explanations of Counterfactual Inferences</i> .....	318
Brian Edwards, Lance Rips	
<i>Learning Conceptual Hierarchies by Iterated Relational Consolidation</i> .....	324
James Foster, Fabian Canas, Matt Jones	
<i>"Eyes Closed" and "Eyes Open" Expectations Guide Fixations in Real-World Search</i> .....	330
Tom Foulsham	
<i>Maximum utility unitary coherent perception vs. the Bayesian brain</i> .....	336
Charles Fox, Tom Stafford	
<i>Measuring children's visual access to social information using face detection</i> .....	342
Michael Frank	
<i>The Effects of Feedback During Exploration Depend on Prior Knowledge</i> .....	348
Emily Fyfe, Bethany Rittle-Johnson	
<i>Children's Inferences in Generalizing Novel Nouns and Adjectives</i> .....	354
Annie Gagliardi, Erin Bennett, Jeffrey Lidz, Naomi H. Feldman	
<i>When Suboptimal Behavior is Optimal and Why: Modeling the Acquisition of Noun Classes in Tsez</i> .....	360
Annie Gagliardi, Naomi H. Feldman, Jeffrey Lidz	
<i>Verbal Satiation of Chinese Bisyllabic Words: A Semantic Locus and its Time Course</i> .....	366
Bruno Galmar, Jenn-Yeu Chen	
<i>Online learning of causal structure in a dynamic game situation</i> .....	372
Yue Gao, Eyal Nitzany, Shimon Edelman	
<i>Noisy Newtons: Unifying process and dependency accounts of causal attribution</i> .....	378
Tobias Gerstenberg, Noah Goodman, David Lagnado, Joshua Tenenbaum	
<i>Explaining children's failure in analogy making tasks: A problem of focus of attention?</i> .....	384
Yannick Glady, Jean-Pierre Thibaut, Bob French, Agnes Blaye	

<i>Knowledge and implicature: Modeling language understanding as social cognition</i> .....	390
Noah Goodman, Andreas Stuhlmüller	
<i>Analogical Problem Solving: Insights from Verbal Reports</i> .....	396
Linn Gralla, Thora Tenbrink, Michael Siebers, Ute Schmid	
<i>Comparing the inductive biases of simple neural networks and Bayesian models</i> .....	402
Thomas Griffiths, Joseph Austerweil, Vincent Berthiaume	
<i>Cooperation in Prisoner's Dilemma Game: Influence of Social Relations</i> .....	408
Maurice Grinberg, Evgenia Hristova, Milena Borisova	
<i>Evaluating the Relationship Between Neuropsychological Function and Cognitive Performance</i> ....	414
Glenn Gunzelmann, L. Richard Moore	
<i>The Interplay between Feature-Saliency and Feedback Information in Visual Category Learning Tasks</i> .....	420
Rubi Hammer, Vladimir Sloutsky, Kalanit Grill-Spector	
<i>Self-Terminated vs. Experimenter-Terminated Memory Search</i> .....	426
J. Isaiah Harbison, Erika Hussey, Michael Dougherty, Eddy Davelaar	
<i>The use of ACT-R to develop an attention model for simple driving tasks</i> .....	432
Kerstin Sophie Haring, Marco Ragni, Lars Konieczny, Katsumi Watanabe	
<i>Testing the Split Attention Effect on Learning in a Natural Educational Setting Using an Intelligent Tutoring System for Geometry</i> .....	438
Robert Hausmann, Annalies Vuong	
<i>The effect of "Maverick": A study of Group Dynamics on Breakthrough in Collaborative Problem solving</i> .....	444
Yugo Hayashi	
<i>Tightening up Joke Structure: Not by Length Alone</i> .....	450
Christian F. Hempelmann, Julia M. Taylor, Victor Raskin	
<i>Logical or Pragmatic, as Long as it Suits our Convenience: Scalar Inferences in a Pro-and Contra-attitudinal Context</i> .....	456
Tom Heyman, Walter Schaeken, Katrijn Pipijn	
<i>Do repeated references result in sign reduction?</i> .....	461
Marieke Hoetjes, Emiel Krahmer, Marc Swerts	
<i>When gestures catch the eye: The influence of gaze direction on co-speech gesture comprehension in triadic communication</i> .....	467
Judith Holler, Spencer Kelly, Peter Hagoort, Asli Ozyurek	
<i>Learning from speaker word choice by assuming adjectives are informative</i> .....	473
Alexandra Horowitz, Michael Frank	
<i>Finishing each other's . . . Responding to incomplete contributions in dialogue</i> .....	479
ChrIstine Howes, Patrick G. T. Healey, Matthew Purver, Arash Eshghi	
<i>Identifying representations of categories of discrete items using Markov chain Monte Carlo with People</i> .....	485
Anne Hsu, Jay Martin, Adam Sanborn, Tom Griffiths	



<i>Social Networks are Encoded in Language</i> .....	491
Sterling Hutchinson, Vivek Datla, Max Louwerse	
<i>Memory Indexing of Sequential Symptom Processing in Diagnostic Reasoning</i> .....	497
Georg Jahn, Janina Braatz	
<i>Gestures Alter Thinking About Time</i> .....	503
Azadeh Jamalian, Barbara Tversky	
<i>The Role of Task Characteristics in Children's Scalar Implicature Production</i> .....	509
Leen Janssens, Walter Schaeken	
<i>Learning What is Where from Social Observations</i> .....	515
Julian Jara-Ettinger, Chris Baker, Joshua Tenenbaum	
<i>Generalization of Learning in Games of Strategic Interaction</i> .....	521
Ion Juvina, Christian Lebiere, Cleotilde Gonzalez, Muniba Saleem	
<i>Actively Learning Nouns Across Ambiguous Situations</i> .....	527
George Kachergis, Chen Yu, Richard Shiffrin	
<i>Learning Nouns with Domain-General Associative Learning Mechanisms</i> .....	533
George Kachergis	
<i>Testing a Distinctiveness Explanation of the Primacy Effect in Free Recall Using Event-Related Potentials</i> .....	539
Siri-Maria Kamp, Glen R. Forester, Anthony R. Murphy, Ty Brumback, Emanuel Donchin	
<i>Modeling Learning of Relational Abstractions via Structural Alignment</i> .....	545
Subu Kandaswamy, Kenneth Forbus	
<i>From Hands to Minds: Gestures Promote Action Understanding</i> .....	551
Seokmin Kang, Barbara Tversky, John Black	
<i>An experimental investigation of consistency of explanation and graph representation</i> .....	557
Nana Kanzaki, Kazuhisa Miwa	
<i>The Influence of Virtual Agents' Gender and Rapport on Enhancing Math Performance</i> .....	563
Bilge Karacora, Morteza Dehghani, Nicole Krämer-Mertens, Jonathan Gratch	
<i>A tripartite trans-modal relationship among sounds, shapes and emotions: A case of abrupt modulation</i> .....	569
Shigeto Kawahara, Kazuko Shinohara	
<i>Negating compound sentences</i> .....	575
Sangeet Khemlani, Isabel Orenes, Philip N. Johnson-Laird	
<i>mReactr: A computational theory of deductive reasoning</i> .....	581
Sangeet Khemlani, Greg Trafton	
<i>The Specificity of Online Variation in Speech Production</i> .....	587
Christo Kirov, Colin Wilson	
<i>Running it through the body</i> .....	593
David Kirsh	

<i>A belief-updating model of adaptation and cue combination in syntactic comprehension</i> .....	599
Dave F. Kleinschmidt, Alex B. Fine, T. Florian Jaeger	
<i>A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation</i> .....	605
Dave F. Kleinschmidt, T. Florian Jaeger	
<i>ERP Responses to Violations in Japanese Verb Conjugation Patterns</i> .....	611
Yuki Kobayashi, Yoko Sugioka, Takane Ito	
<i>Event Segmentation of Agent Interactions: Comparing the Whole with Its Parts</i> .....	617
Bryan Koenig, Bryan Koenig, David Pautler, Jonathan Herberg, Kum Seong Wan, Brian Monroe, Brian Monroe, Edwin Wirawan	
<i>Thinking in Patterns: using multi-voxel pattern analyses to find neural correlates of moral judgment in neurotypical and ASD populations</i> .....	623
Jorie Koster-Hale, James Dungan, Rebecca Saxe, Liane Young	
<i>A Unified Model of Categorical Effects in Consonant and Vowel Perception</i> .....	629
Yakov Kronrod, Emily Coppess, Naomi Feldman	
<i>Early and Repeated Exposure to Examples Improves Creative Work</i> .....	635
Chinmay Kulkarni, Steven Dow, Scott Klemmer	
<i>Role of Error Monitoring Mechanisms in Attribution of Sense of Self-Agency</i> .....	641
Neeraj Kumar, Jaison A. Manjaly, Krishna P. Miyapuram	
<i>Pragmatic interpretation of contrastive prosody: It looks like speech adaptation</i> .....	647
Chigusa Kurumada, Meredith Brown, Michael Tanenhaus	
<i>A Graph-Oriented Approach to Measuring Expertise- Detecting Structural Differences between Experts and Intermediates</i> .....	653
Andreas Lachner, Johannes Gurlitt, Matthias Nückles	
<i>Concept learning as motor program induction: A large-scale empirical study</i> .....	659
Brenden Lake, Ruslan Salakhutdinov, Joshua Tenenbaum	
<i>How many kinds of reasoning? Inference, probability, and natural language semantics</i> .....	665
Daniel Lassiter, Noah Goodman	
<i>A Behavioral Investigation of Dimensionality Reduction</i> .....	671
Joshua Lewis, Laurens van der Maaten, Virginia de Sa	
<i>Modeling Melodic Perception as Relational Learning Using a Symbolic-Connectionist Architecture (DORA)</i> .....	677
Ahnate Lim, Leonidas Doumas, Scott Sinnett	
<i>The persisting benefits of using multiple-choice tests as learning events</i> .....	683
Jeri Little, Elizabeth Ligon Bjork	
<i>The perception of simplified and traditional Chinese characters in the eye of simplified and traditional Chinese readers</i> .....	689
Tianyin Liu, Janet Hsiao	

<i>The Chinese Route Argument: Predicting the Longitude and Latitude of Cities in China and the Middle East Using Statistical Linguistic Frequencies</i> .....	695
Max Louwerse, Sterling Hutchinson, Zhiqiang Cai	
<i>Modeling Multiple Strategies for Solving Geometric Analogy Problems</i> .....	701
Andrew Lovett, Kenneth Forbus	
<i>A unified theory of counterfactual reasoning</i> .....	707
Christopher Lucas, Charles Kemp	
<i>Superspace extrapolation reveals inductive biases in function learning</i> .....	713
Christopher Lucas, Douglas Sterling, Charles Kemp	
<i>Does the utility of information influence sampling behavior?</i> .....	719
Doug Markant, Todd Gureckis	
<i>One piece at a time: Learning complex rules through self-directed sampling</i> .....	725
Doug Markant, Todd Gureckis	
<i>Shallow learning as a pathway for successful learning both for tutors and tutees</i> .....	731
Noboru Matsuda, Evelyn Yarzebinski, Victoria Keiser, Rohan Raizada, William W. Cohen, Gabriel Stylianides, Kenneth R. Koedinger	
<i>Going with TRACE beyond Infant Mispronunciation Studies: Lexical Networks and Phoneme Competition</i> .....	737
Julien Mayor, Kim Plunkett	
<i>Causal Status meets Coherence: The Explanatory Role of Causal Models in Categorization</i> .....	743
Ralf Mayrhofer, Anselm Rothe	
<i>Sparse category labels obstruct generalization of category membership</i> .....	749
John V. McDonnell, Carol A. Jew, Todd M. Gureckis	
<i>Automated and Partner-Specific Factors Influencing Lexical Entrainment</i> .....	755
Lisette Mol, Rens Bogers, Tommy Bouwens	
<i>Gesture structure affects syntactic structure in speech</i> .....	761
Lisette Mol, Sotaro Kita	
<i>Modeling Millisecond Time Interval Estimation in Space Fortress Game</i> .....	767
Jungaa Moon, John Anderson	
<i>The Role of Gesture in Second Language Learning: Communication, Acquisition, &amp; Retention</i> ...	773
Laura Morett, Ray Gibbs, Brian MacWhinney	
<i>The Role of Imitation in Generating a Shared Communication System</i> .....	779
Junya Morita, Takeshi Konno, Takashi Hashimoto	
<i>Force Dynamics as a Basis for Moral Intuitions</i> .....	785
Jonas Nagel, Michael Waldmann	
<i>Preservation of the Initial Analysis in Absence of Pragmatic Inference with Japanese Relative Clause Sentences</i> .....	791
Chie Nakamura, Manabu Arai	

<i>The Role of the Amygdala in the Process of Humour Appreciation</i> .....	797
Tagiru Nakamura, Tomoko Matsui, Akira Utsumi, Mika Yamazaki, Kai Makita, Hiroki C. Tanabe, Norihiro Sadato	
<i>The Footbridge Dilemma Reflects More Utilitarian Thinking Than The Trolley Dilemma: Effect Of Number Of Victims In Moral Dilemmas</i> .....	803
Kuninori Nakamura	
<i>Anticipating changes: Adaptation and extrapolation in category learning</i> .....	809
Daniel Navarro, Amy Perfors	
<i>Language-induced Biases on Human Sequential Learning</i> .....	815
Luca Onnis, EriK Thiessen	
<i>Semantic Coherence Facilitates Distributional Learning of Word Meanings</i> .....	821
Long Ouyang, Lera Boroditsky, Michael Frank	
<i>Grounding spatial language in non-linguistic cognition: Evidence for universal and relative spatial semantics in thought</i> .....	827
Michael Pacer, Alexandra Carstensen, Terry Regier	
<i>Elements of a rational framework for continuous-time causal induction</i> .....	833
Michael Pacer, Tom Griffiths	
<i>Musicians are better at learning non-native sound contrasts even in non-tonal languages</i> .....	839
Amy Perfors, Jia Hoong Ong	
<i>Probability matching vs over-regularization in language: Participant behavior depends on their interpretation of the task</i> .....	845
Amy Perfors	
<i>An Operational Model of Joint Attention - Timing of Gaze Patterns in Interactions between Humans and a Virtual Human</i> .....	851
Nadine Pfeiffer-Lessmann, Thies Pfeiffer, Ipke Wachsmuth	
<i>Manipulating Manner: Semantic Representations of Human Locomotion Verbs in English and German</i> .....	857
Katherine Phelps, Steve Duman	
<i>"Less is More" in Bayesian Word Segmentation: When cognitively plausible learners outperform the ideal</i> .....	863
Lawrence Phillips, Lisa Pearl	
<i>Categorial compositionality continued (further): A category theory explanation for the systematicity of recursive cognitive capacities</i> .....	869
Steven Phillips, William Wilson	
<i>Modeling Concept Activation in Working Memory during Online Sentence Processing</i> .....	875
Patrick Plummer, Hsueh-Cheng Wang, Marc Pomplun, Yuhtsuen Tzeng, Keith Rayner	
<i>Exploring the Role of Representation in Models of Grammatical Category Acquisition</i> .....	881
Ting Qian, Patricia Reeder, Richard Aslin, Joshua Tenenbaum, Elissa Newport	
<i>Acoustic analysis supports the existence of a single distributional learning mechanism in structural rule learning from an artificial language</i> .....	887
Okko Räsänen, Heikki Rasilo	

<i>Optimally Designing Games for Cognitive Science Research</i> .....	893
Anna Rafferty, Matei Zaharia, Thomas Griffiths	
<i>Cognitive Workload and the Motor Component of Visual Attention</i> .....	899
Jason Ralph, Wayne Gray, Mike Schoelles	
<i>Order effects in diagnostic reasoning with four candidate hypotheses</i> .....	905
Felix G. Rebitschek, Agnes Scholz, Franziska Bocklisch, Josef F. Krems, Georg Jahn	
<i>Gaze cues in complex, real-world scenes direct the attention of high-functioning adults with autism</i>	911
Elizabeth Redcay, Daniel R O’Young, Lloyd R Slevc, Penelope L Mavros, John D Gabrieli, Pawan Sinha	
<i>Examining the Representation and Understanding of Large Magnitudes Using the Hierarchical Alignment model of Analogical Reasoning</i> .....	917
Ilyse Resnick, Thomas Shipley, Nora Newcombe, Christine Massey, Theodore Wills	
<i>The Development of Joint Belief-Desire Inferences</i> .....	923
Hilary Richardson, Chris Baker, Joshua Tenenbaum, Rebecca Saxe	
<i>Expectations About the Temporal Structure of the World Result in the Attentional Blink and Repetition Blindness</i> .....	929
Cory Rieth, Edward Vul	
<i>Relating Activity Contexts to Early Word Learning in Dense Longitudinal Data</i> .....	935
Brandon Roy, Michael Frank, Deb Roy	
<i>Modeling Cognition: How Fiction Relates to Fact</i> .....	941
Anna-Mari Rusanen, Otto Lappi	
<i>Role of Kolmogorov Complexity on Interest in Moral Dilemma Stories</i> .....	947
Antoine Saillenfest, Jean-Louis Dessalles	
<i>Using Concept Map to Evaluate Learning by Searching</i> .....	953
Hitomi Saito, Yuka Egusa, Masao Takaku, Makiko Miwa, Noriko Kando	
<i>Towards a cognitive science of literary style: Perspective-taking in processing omniscient versus objective voice</i> .....	959
Manami Sato, Hiromu Sakai, Jennifer Wu, Benjamin Bergen	
<i>Beyond one’s own understanding: How text comprehensibility affects laypeople’s decision about scientific claims</i> .....	965
Lisa Scharrer, M. Anne Britt, Marc Stadtler, Rainer Bromme	
<i>Subjective Confidence of Acoustic and Phonemic Representations During Speech Perception</i> .....	971
Jordan Schoenherr, John Logan, Amy Winchester	
<i>Enough is enough: Inductive sufficiency guides learners’ ratings of informant helpfulness</i> .....	977
Patrick Shafto, Hyowon Gweon, Chris Fargen, Laura Schulz	
<i>Investigating the Locus of the Word Frequency Effect in Spoken Word Recognition</i> .....	983
Cynthia Siew, Melvin Yap, Winston Goh	
<i>Zero anaphora and object reference in Japanese child-directed speech</i> .....	989
Cybelle Smith, Michael Frank	

<i>Sources of uncertainty in intuitive physics</i> .....	995
Kevin Smith, Edward Vul	
<i>Change detection under autocorrelation</i> .....	1001
Maarten Speekenbrink, Matthew Twyman, Nigel Harvey	
<i>Using listener gaze to augment speech generation in a virtual 3D environment</i> .....	1007
Maria Staudte, Alexander Koller, Konstantina Garoufi, Matthew Crocker	
<i>The effect of metabolic loading on statistical learning</i> .....	1013
David Stevens, Joanne Arciuli, David Anderson, Mark Williams	
<i>Spaun: A Perception-Cognition-Action Model Using Spiking Neurons</i> .....	1018
Terrence Stewart, Feng-Xuan Choo, Chris Eliasmith	
<i>Perception of Randomness: Subjective Probability of Alternation</i> .....	1024
Yanlong Sun, Hongbin Wang	
<i>Time Course of Inhibitory Control During Analogical Reasoning: An Event-Related Potential Approach</i> .....	1030
Brian Sweis, Krishna Bharani, Robert Morrison	
<i>Effects of explicit knowledge on transfer of visuomotor sequence learning</i> .....	1036
Kanji Tanaka, Katsumi Watanabe	
<i>Spontaneous body movements in spatial cognition</i> .....	1042
Sergiu Tcaci Popescu, Mark Wexler	
<i>Auditory Saliency Using Natural Statistics</i> .....	1048
Tomoki Tsuchida, Garrison Cottrell	
<i>A Multi-Measure Analysis of Context Effects in Multi-Attribute Decision Making: Examining the Similarity, Attraction, and Compromise Effects</i> .....	1054
Takashi Tsuzuki, Jerome Busemeyer	
<i>Mental Arithmetic Efficiency: Interactivity and Individual Differences</i> .....	1060
Frederic Vallee-Tourangeau	
<i>Conceptual alignment in reference with artificial and human dialogue partners</i> .....	1066
Koen van Lierop, Martijn Goudbeek, Emiel Krahmer	
<i>Neural Circuits for Any-Time Phrase Recognition with Applications in Cognitive Models and Human-Robot Interaction</i> .....	1072
Richard Veale, Matthias Scheutz	
<i>An Integrated Model of Associative and Reinforcement Learning</i> .....	1078
Vladislav Veksler, Christopher Myers, Kevin Gluck	
<i>The Impact of Colour Difference and Colour Codability on Reference Production</i> .....	1084
Jette Viethen, Martijn Goudbeek, Emiel Krahmer	
<i>Bayesian Logic and Trial-by-Trial Learning</i> .....	1090
Momme von Sydow, Klaus Fiedler	
<i>Color word learning is a gradual inductive process</i> .....	1096
Katie Wagner, Karen Dobkins, David Barner	

<i>The Role of the Primary Effect in the Assessment of Intentionality and Morality</i> .....	1102
Michael R. Waldmann, Alex Wiegmann	
<i>Children's Causal Learning from Fiction: Assessing the Proximity Between Real and Fictional Worlds</i> .....	1108
Caren M. Walker, Patricia A. Ganea, Alison Gopnik	
<i>Explaining Influences Children's Reliance on Evidence and Prior Knowledge in Causal Induction</i>	1114
Caren M. Walker, Joseph Jay Williams, Tania Lombrozo, Alison Gopnik	
<i>An Abductive Approach to Covert Interventions</i> .....	1120
Hongbin Wang, Yanlong Sun	
<i>Choosing quantity over quality: syntax guides interpretive preferences for novel superlatives</i> .....	1126
Alexis Wellwood, Darko Odic, Justin Halberda, Jeffrey Lidz	
<i>Expertise and the Wisdom of Crowds: Whose Judgments to Trust and When</i> .....	1131
Matthew Welsh	
<i>Complex First? On the Priority of Nouns in Language Acquisition and Evolution</i> .....	1137
Markus Werning	
<i>Order Effects in Moral Judgment. Searching for an Explanation</i> .....	1143
Alex Wiegmann, Yasmina Okan	
<i>Explaining increases belief revision in the face of (many) anomalies</i> .....	1149
Joseph Jay Williams, Caren Walker, Tania Lombrozo	
<i>Olfaction in a hunter-gatherer society: Insights from language and culture</i> .....	1155
Ewelina Wnuk, Asifa Majid	
<i>Learning (to Learn) from Spatial Attention Cues During Infancy</i> .....	1161
Rachel Wu, Natasha Kirkham	
<i>Adaptive Information Search and Decision Making over Single and Repeated Plays</i> .....	1167
Dirk U. Wulff, Thomas T. Hills, Ralph Hertwig	
<i>Stable Self-to-Object Spatial Relations Acquired from Sequential Spatial Learning</i> .....	1173
Chengli Xiao, Fudan Chen	
<i>Mutual Affects in Computer-mediated Collaborative Learning: Positive Feelings Shared by Collaborators Enhance System Evaluations</i> .....	1179
Takashi Yamauchi, Takehiko Ohno, Momoko Nakatani, Yoichi Kato, Art Markman	
<i>Implicit Learning: A Demonstration and a Novel SRT Paradigm</i> .....	1185
Fayme Yeates, Fergal Jones, Andy Wills, Mike Aitken, Ian P. L. McLaren	
<i>The Effect of Semantic Similarity is a Function of Contextual Constraint</i> .....	1191
Hongoak Yun, Gail Mauner, Douglas Roland, Jean-Pierre Koenig	
<i>Mutual Exclusivity and Vocabulary Development</i> .....	1197
Daniel Yurovsky, Ricardo Bion, Linda Smith, Anne Fernald	
<i>Quantitative Linking Hypotheses for Infant Eye Movements</i> .....	1203
Daniel Yurovsky, Shohei Hidaka, Rachel Wu	



<i>Does Statistical Word Learning Scale? It's a Matter of Perspective</i> .....	1209
Daniel Yurovsky, Linda Smith, Chen Yu	
<i>Inferring Covert Events in Logical Metonymies: a Probe Recognition Experiment</i> .....	1215
Alessandra Zarcone, Sebastian Pado, Alessandro Lenci	
<i>Sparse Population Code Models of Word Learning in Concept Drift</i> .....	1221
Byoung-Tak Zhang, Jung-Woo Ha, Myunggu Kang	
<i>The role of recent versus future events in children's comprehension of referentially ambiguous sentences: Evidence from eye tracking</i> .....	1227
Lu Zhang, Lily Kornbluth, Pia Knoeferle	
<i>Updating: Learning versus supposing</i> .....	1233
Jiaying Zhao, Vincenzo Crupi, Katya Tentori, Branden Fitelson, Daniel Osherson	

## Publication-Based Talks

<i>Humor, Emotions and Communication: Human-like Issues of Human-Computer Interactions</i> ...	1238
Pawel Dybala, Kohichi Sayama	
<i>Meta-Representational Competence as an Aspect of Spatial Intelligence</i> .....	1240
Mary Hegarty	

## Posters

<i>Rationality-Guided AGI as Cognitive Systems</i> .....	1242
Ahmed M. H. Abdel-Fattah, Tarek R. Besold, Helmar Gust, Ulf Krumnack, Martin Schmidt, Kai-Uwe Kühnberger, Pei Wang	
<i>Musical Relevance: a Computational Approach</i> .....	1248
Edoardo Acotto, Daniele Radicioni	
<i>Examining the Connection Between Dynamic and Static Spatial Skills and Video Game Performance</i> .....	1254
Deanne Adams, Rich Mayer	
<i>Erroneous Examples Versus Problem Solving: Can We Improve How Middle School Students Learn Decimals?</i> .....	1260
Deanne Adams, Bruce McLaren, Kelley Durkin, Rich Mayer, Bethany Rittle-Johnson, Seiji Isotani, Martin van Velsen	
<i>A Multi-Category Theory of Intention</i> .....	1266
Henny Admoni, Brian Scassellati	
<i>A Narratological Approach for Narrative Discourse: Implementation and Evaluation of the System based on Genette and Jauss</i> .....	1272
Taisuke Akimoto, Takashi Ogata	
<i>A Bayesian Model of the Effect of Object Context on Visual Attention</i> .....	1278
Ben Allison, Frank Keller, Moreno I. Coco	
<i>The Theory of Visual Attention without the race: a new model of visual selection</i> .....	1284
Tobias Andersen, Søren Kyllingsbæk	

<i>The Development of Second-order Social Cognition and its Relation with Complex Language Understanding and Memory</i> .....	1290
Burcu Arslan, Annette Hohenberger, Rineke Verbrugge	
<i>Learning of motor maps from perception: a dimensionality reduction approach</i> .....	1296
Ankit Awasthi, Sadbodh Sharma, Amitabha Mukerjee	
<i>Establishing a Database for Studying Human Face Photograph Memory</i> .....	1302
Wilma Bainbridge, Phillip Isola, Idan Blank, Aude Oliva	
<i>The Representation and Processing of Tense, Aspect &amp; Voice across Verbal Elements in English</i>	1308
Jerry Ball, Christopher Myers	
<i>Cognitive reserve and intelligence: Modulating the effects of damage in ageing dynamical systems</i>	1314
Frank Baughman, Natalie Baughman, Simon Mills	
<i>Verb omission errors: Evidence of rational processing of noisy language inputs</i> .....	1320
Leon Bergen, Roger Levy, Edward Gibson	
<i>Gestural Alignment in Natural Dialogue</i> .....	1326
Kirsten Bergmann, Stefan Kopp	
<i>E Pluribus Multa In Unum: The Rationality Multiverse</i> .....	1332
Tarek R. Besold, Kai-Uwe Kühnberger	
<i>The Role of Semantic Transparency in the Processing of Verb-particle Constructions by French-English Bilinguals</i> .....	1338
Mary-Jane Blais, Laura Gonnerman	
<i>Perception of Ambiguous Drawings and the Construction and Inhibition of its Alternative Interpretation – Reflections on Consciousness</i> .....	1344
Svetoslav Bliznashki, Maria Popova, Boicho Kokinov	
<i>Changes in Foreign Language Skills Over Time</i> .....	1350
Amber Bloomfield, Megan Masters, Steven Ross, Stephen O’Connell, Kassandra Gynther	
<i>Look-Ahead Monte Carlo with People</i> .....	1356
Charles Blundell, Adam Sanborn, Tom Griffiths	
<i>A Study of Loan Color Terms Collocation in Modern Japanese</i> .....	1362
Anna Bordilovskaya	
<i>Intelligibility is Necessary for Scientific Explanation, but Accuracy May Not Be</i> .....	1368
Mike Braverman, John Clevenger, Ian Harmon, Andrew Higgins, Zachary Horne, Joseph Spino, Jonathan Waskan	
<i>Real-time expectations based on context speech rate can cause words to appear or disappear</i> .....	1374
Meredith Brown, Laura C. Dilley, Michael K. Tanenhaus	
<i>Metrical expectations from preceding prosody influence spoken word recognition</i> .....	1380
Meredith Brown, Anne Pier Salverda, Laura C. Dilley, Michael K. Tanenhaus	
<i>Eliciting a Sensemaking Process from Verbal Protocols of Reverse Engineers</i> .....	1386
Adam Bryant, Robert Mills, Gilbert Peterson, Michael Grimaila	

<i>Cooing, Crying, and Babbling: A Link between Music and Prelinguistic Communication</i> .....	1392
Michael Byrd, Casady Bowman, Takashi Yamauchi	
<i>Mothers Do Not Drive Structure in Adult Homesign Systems: Evidence from Comprehension</i> ...	1398
Emily Carrigan, Marie Coppola	
<i>Information Foraging in the Unknown Patches across the Life Span</i> .....	1404
Jessie Chin, Brennan Payne, Andrew Battles, Wai-Tat Fu, Daniel Morrow, Elizabeth Stine-Morrow	
<i>Connectivity Asymmetry Can Explain Visual Hemispheric Asymmetries in Local/Global, Face, and Spatial Frequency Processing</i> .....	1410
Benjamin Cipollini, Janet Hsiao, Garrison Cottrell	
<i>The face inversion effect and evoked brain potentials: Complete loss of configural information affects the N170</i> .....	1416
Ciro Civile, Heike Elchlepp, Rossy McLaren, Aureliu Lavric, Ian P. L. McLaren	
<i>Face recognition and brain potentials: Disruption of configural information reduces the face inversion effect.</i> .....	1422
Ciro Civile, Heike Elchlepp, Rossy McLaren, Aureliu Lavric, Ian P. L. McLaren	
<i>Strength of Perceptual Experience Predicts Word Processing Performance Better than Concreteness or Imageability</i> .....	1428
Louise Connell, Dermot Lynott	
<i>Learning of Relational Categories as a Function of Higher-Order Structure</i> .....	1434
Daniel Corral, Matt Jones	
<i>Gaussian Process Regression for Trajectory Analysis</i> .....	1440
Gregory Cox, George Kachergis, Richard Shiffrin	
<i>Mathematical Modeling of a Biological Odometry</i> .....	1446
Somayeh Danafar	
<i>Solving nonogram puzzles by reinforcement learning</i> .....	1452
Frederic Dandurand, Denis Cousineau, Thomas R. Shultz	
<i>Rational Search of Associative Memory</i> .....	1458
Eddy Davelaar, J. Isaiah Harbison, Erica Yu, Erika Hussey, Michael Dougherty	
<i>Strong structure in weak semantic similarity: A graph based account</i> .....	1464
Simon de Deyne, Daniel Navarro, Amy Perfors, Gert Storms	
<i>Conceptual Event Units of Putting and Taking in Two Unrelated Languages</i> .....	1470
Rebecca Defina, Asifa Majid	
<i>Interpersonal Effects of Emotions in Morally-charged Negotiations</i> .....	1476
Morteza Dehghani, Jonathan Gratch, Peter Carnevale	
<i>Using Accent to Induce Cultural Frame-Switching</i> .....	1482
Morteza Dehghani, Peter Khooshabeh, Lixing Huang, Angela Nazarian, Jonathan Gratch	
<i>How Function Assignment and Word Order are Determined: Evidence from Structural Priming Effects in Japanese Sentence Production</i> .....	1488
Ying Deng, Hajime Ono, Hiromu Sakai	

<i>A Window of Perception When Diverting Attention? Enhancing Recognition For Explicitly Presented, Unattended, and Irrelevant Visual Stimuli by Target Alignment</i> .....	1494
Andrew Dewald, Scott Sinnett	
<i>A Computational Logic Approach to the Suppression Task</i> .....	1500
Emmanuelle-Anna Dietz, Steffen Hölldobler, Marco Ragni	
<i>Confidence in Causal Inferences: The Case of Devaluation</i> .....	1506
Uwe Drewitz, Stefan Brandenburg	
<i>Spatial Co-ordination in Music Tuition</i> .....	1512
Sam Duffy, Patrick G. T. Healey	
<i>Increased Vigilance in Monitoring Others' Mental States During Deception</i> .....	1518
Nicholas Duran, Rick Dale	
<i>Mining Relatedness Graphs for Data Integration</i> .....	1524
Jeremy Engle, Ying Feng, Robert Goldstone	
<i>Stochastic Naming Game and Self-constraining Nature of Synonymy</i> .....	1530
Kerem Eryilmaz, Cem Bozsahin	
<i>Learning unattested languages</i> .....	1536
Sara Finley	
<i>Systemic Expertise: Instructing Non-Artists on Depicting Human Figures in 3-D</i> .....	1542
Nick Flor	
<i>Task switching without knowledge of the tasks.</i> .....	1548
Charlotte Forrest, Heike Elchlepp, Stephen Monsell, Ian P. L. McLaren	
<i>Early effects of word surprisal on pupil size during reading</i> .....	1554
Stefan Frank, Robin Thompson	
<i>The Plausibility of Semantic Properties Generated by a Distributional Model: Evidence from a Visual World Experiment</i> .....	1560
Diego Frassinelli, Frank Keller	
<i>Concepts in context: Evidence from a feature-norming study</i> .....	1566
Diego Frassinelli, Alessandro Lenci	
<i>Learning transfer in small group coordination</i> .....	1572
Seth Frey	
<i>Society Functions Best with an Intermediate Level of Creativity</i> .....	1578
Liane Gabora, Hadi Firouzi	
<i>Does domain size impact speech onset time during reference production?</i> .....	1584
Albert Gatt, Roger P. G. van Gompel, Emiel Krahmer, Kees van Deemter	
<i>Ping Pong in Church: Productive use of concepts in human probabilistic inference</i> .....	1590
Tobias Gerstenberg, Noah Goodman	
<i>Speech Act Recognition in Conversation: Experimental Evidence</i> .....	1596
Rosa Gisladdottir, Dorothee Chwilla, Herbert Schriefers, Stephen Levinson	

<i>Towards historical cognitive science: the case of Ancient Greece</i> .....	1602
Vladimir Glebkin	
<i>Development of Category-Based Reasoning in Preschool-Age Children: Preliminary Results of a Longitudinal Study</i> .....	1608
Karrie Godwin, Bryan Matlen, Anna Fisher	
<i>Is that your final answer? The effects of neutral queries on children's choices</i> .....	1614
Aaron Gonzalez, Patrick Shafto, Elizabeth Baraff Bonawitz, Alison Gopnik	
<i>Abstract language comprehension is incrementally modulated by non-referential spatial information: evidence from eye-tracking</i> .....	1620
Ernesto Guerra, Pia Knoeferle	
<i>Trading in a multiplayer board game: Towards an analysis of non-cooperative dialogue</i> .....	1626
Markus Guhe, Alex Lascarides	
<i>The Characteristics of Usability and Users' Eye movements in Searching for Information in a Hierarchically Organized Information Structure</i> .....	1632
Yoshiko Habuchi, Haruhiko Takeuchi	
<i>A time-invariant connectionist model of spoken word recognition</i> .....	1638
Thomas Hannagan, James Magnusson, Jonathan Grainger	
<i>N-back Performance: Comparing Assessment and Training Performance</i> .....	1644
J. Isaiah Harbison, Sharona Atkins, Michael R. Dougherty	
<i>Pedagogical agents that support learning by explaining: Effects of affective feedback</i> .....	1650
Yugo Hayashi, Mariko Matsumoto, Hitoshi Ogawa	
<i>Knowledge and Political Categorization</i> .....	1656
Evan Heit, Stephen Nicholson	
<i>Going to Extremes: The influence of unsupervised categories on the mental caricaturization of faces and asymmetries in perceptual discrimination</i> .....	1662
Andrew Hendrickson, Paulo Carvalho, Robert Goldstone	
<i>Re-learning labeled categories reveals structured representations</i> .....	1668
Andrew Hendrickson, George Kachergis, Caitlin Fausey, Robert Goldstone	
<i>A Matter of Process Accuracy: Observing or Inferring the Criterion of Few or Many Exemplars</i>	1674
Maria Henriksson	
<i>Identifying Kinematic Cues for Action Style Recognition</i> .....	1679
Shohei Hidaka	
<i>The Atoms of Cognition: A Theory of Ground Epistemics</i> .....	1685
Seng Beng Ho	
<i>Simple heuristic or knowledge-based inference? Model comparison of binary choice inference</i> ....	1691
Hidehito Honda, Toshihiko Matsuka	
<i>Whose turn is it anyway? Same- and cross-person compound contributions in dialogue</i> .....	1697
Christine Howes, Patrick G. T. Healey, Matthew Purver	

<i>Language acquisition in Down Syndrome from embodied perspective: How body constrains language acquisition?</i> .....	1703
Penka Hristova, Hristina Toushek, Georgi Petkov	
<i>The Upbeat of Language: Linguistic Context and Embodiment Predict Processing Valence Words</i>	1709
Sterling Hutchinson, Max Louwerse	
<i>Knowledge-based Modeling in Dynamic Decision Making</i> .....	1715
Angel Iglesias, M. Dolores del Castillo, J. Ignacio Serrano, Jesus Oliva	
<i>Rapid entrainment to spontaneous speech: A comparison of oscillator models</i> .....	1721
Benjamin Inden, Zofia Malisz, Petra Wagner, Ipke Wachsmuth	
<i>An Empirical Study on the Mechanisms of Creativity in Visual Arts</i> .....	1727
Bipin Indurkha, Shinji Ogawa	
<i>Emergence of control in artistic expressions and the process of expertise</i> .....	1733
Chiaki Ishiguro, Takeshi Okada	
<i>Changes in Cognitive Processes upon Learning Mini-Shogi</i> .....	1739
Takeshi Ito, Daisuke Takano, Xiaohong Wan, Keiji Tanaka	
<i>One-shot lotteries in the park</i> .....	1745
Mordechai Juni, Todd Gureckis, Laurence Maloney	
<i>Children's acquisition of fraction knowledge from concrete versus generic instantiations</i> .....	1750
Jennifer Kaminski, Vladimir Sloutsky	
<i>The Role of Imagination in Augmenting Perceptual Representation</i> .....	1756
Seokmin Kang, Gregory Hallman Jr., John Black	
<i>The effects of amnesia on driving performance in elderly drivers.</i> .....	1762
Naoko Kawano, Kunihiro Iwamoto, Kazutoshi Ebe, Katsuyuki Ukai, Yusuke Suzuki, Hiroyuki Umegaki, Tetsuya Iidaka, Norio Ozaki	
<i>From Vectors to Symbols to Cognition: The Symbolic and Sub-Symbolic Aspects of Vector-Symbolic Cognitive Models</i> .....	1768
Matthew Kelly, Robert West	
<i>Sex Differences in the Discrimination of Non-Native Speech Sounds</i> .....	1774
Vera Kempe, John C. Thoresen, Patricia J. Brooks	
<i>Evaluative feedback can improve deductive reasoning</i> .....	1780
Sangeet Khemlani, Adam Moore	
<i>When doing the wrong thing is right</i> .....	1786
David Kirsh, Richard Caballero, Shannon Cuykendall	
<i>Tests and Models of Non-compositional Concepts</i> .....	1792
Kirsty Kitto, Peter Bruza	
<i>Roles of Self Goal Setting in Insight Problem Solving</i> .....	1798
Sachiko Kiyokawa, Katsuyuki Hayashi, Toshihiko Matsuka	
<i>Differences in eye movements between same and other race face recognition.</i> .....	1804
Eve Klama, Fraser Milton	

<i>Different Stable Patterns between Intra- and Inter-personal Systems: Experimental Study on Inter-limb Tapping Coordination</i> .....	1810
Kentaro Kodama, Ryosaku Makino, Nobuhiro Furuyama	
<i>How Can We Live with Overconfident or Unconfident Systems?: A Comparison of Artificial Subtle Expressions with Human-like Expression</i> .....	1816
Takanori Komatsu, Kazuki Kobayashi, Seiji Yamada, Kotaro Funakoshi, Mikio Nakano	
<i>Effect of Social Skills on the Asymmetry in Facial Expressions</i> .....	1822
Masashi Komori, Hiroko Kamide, Satoru Kawamura, Chika Nagaoka	
<i>Reasoning on the Raven's Advanced Progressive Matrices Test with Iconic Visual Representations</i> .....	1828
Maithilee Kunda, Keith McGregor, Ashok Goel	
<i>Corpus-based metrics for assessing communal common ground</i> .....	1834
Roman Kutlak, Kees van Deemter, Chris Mellish	
<i>Getting off at the end of the line: the estimation of large numbers</i> .....	1840
David Landy, Noah Silbert, Aleah Goldin	
<i>Tangent Point Orientation and Anticipated Trajectory Curvature – A Field Study on the Visual Control of High Speed Steering</i> .....	1846
Otto Lappi, Esko Lehtonen	
<i>Arbitrary Category Labels Can Change Similarity Judgments of Human Faces</i> .....	1852
Frankie Lara, Amanda Hahn, Na-Yung Yu, Takashi Yamauchi	
<i>A Critical Look at the Findings of Sergeant (1982)</i> .....	1858
Lyuben Laskin, Meryl Varadinov	
<i>Neural Correlates of Episodic Memory Formation in Audio-Visual Pairing Tasks</i> .....	1864
Chung-Yeon Lee, Beom-Jin Lee, Joon Shik Kim, Byoung-Tak Zhang	
<i>Human Cluster Evaluation and Formal Quality Measures: A Comparative Study</i> .....	1870
Joshua Lewis, Margareta Ackerman, Virginia de Sa	
<i>Learning Cluster Analysis through Experience</i> .....	1876
Joshua Lewis, Virginia de Sa	
<i>The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains</i> .....	1882
Daniel Leyzberg, Samuel Spaulding, Mariya Toneva, Brian Scassellati	
<i>Designing Better Scaffolding in Simulation-Based Learning Environments Teaching Science Systems: A Pilot Study Report</i> .....	1888
Na Li, John Black, Mengzi Gao	
<i>A New Angle on the EMPATH Model: Spatial Frequency Orientation in Recognition of Facial Expressions</i> .....	1894
Rentao Li, Garrison Cottrell	
<i>Learning Image-Derived Eye Movement Patterns to Characterize Perceptual Expertise</i> .....	1900
Rui Li, Jeff Pelz, Pengcheng Shi, Anne Haake	



<i>Multimodal Temporal Perception in Musicians: Evidence for Both Segregated and Supramodal Attentional Systems?</i> .....	1906
Ahnate Lim, Scott Sinnett	
<i>Reexamining Visual Orientation Anisotropies: A Bias Towards Simple Horizontal Stimuli on Temporal Order Judgments</i> .....	1912
Ahnate Lim, Scott Sinnett	
<i>A Bayesian Model of Rule Induction in Raven's Progressive Matrices</i> .....	1918
Daniel R. Little, Stephan Lewandowsky, Thomas L. Griffiths	
<i>Easing and rising of tension from presence of others in player-observer turn-taking in a driving video game: A near-infrared spectroscopy study</i> .....	1924
Tao Liu, Hirofumi Saito, Misato Oi, Matthew Pelowski	
<i>Intentionality of Strong Anticipation in Motor Behaviors</i> .....	1930
Daniel Hsi-wen Liu	
<i>Is embodied cognition infallible or falsifiable? Investigating the thesis as a sound scientific theory</i>	1936
Katherine Livins, Leonidas Doumas	
<i>Natural language – no infinity and probably no recursion</i> .....	1942
Erkki Luuk, Hendrik Luuk	
<i>Modelling the IAT: Implicit Association Test Reflects Shallow Linguistic Environment and not Deep Personal Attitudes</i> .....	1948
Dermot Lynott, Himanshu Kansal, Louise Connell, Kerry O'Brien	
<i>External Working Memory and the Amount of Distributed Cognition</i> .....	1954
Naoki Maeda	
<i>Experimental Investigation of Relationship between Complacency and Tendency to Use Automation System</i> .....	1960
Akihiro Maehigashi, Kazuhisa Miwa, Hitoshi Terai, Kazuaki Kojima, Junya Morita	
<i>Inductive reasoning in the courtroom: Judging guilt based on uncertain evidence</i> .....	1966
Ann Martin, Brett Hayes	
<i>Elaborate Descriptive Information in Indoor Route Instructions</i> .....	1972
Vivien Mast, Cui Jian, Desislava Zhekova	
<i>Connectionist Model Accounting for Retardation of Cognitive-Dissonance Reduction Caused by Attention-Focus Switching</i> .....	1978
Takao Matsumoto	
<i>Investigation of effects of working memory capacity on rule discovery process using eye movement data</i> .....	1984
Miki Matsumuro, Kazuhisa Miwa	
<i>How many Neurons for your 'Grandmother' ? Three Arguments for 'Localised' Representations</i>	1990
Julien Mayor, Kim Plunkett	
<i>Probabilistic generative models for counterfactual reasoning and blame attribution</i> .....	1996
John McCoy, Tomer Ullman, Andreas Stuhlmüller, Tobias Gerstenberg, Joshua Tenenbaum	

<i>Modeling online word segmentation performance in structured artificial languages</i> .....	2002
Stephan Meylan, Chigusa Kurumada, Mike Frank, Benjamin Börschinger, Mark Johnson	
<i>Tradeoff between Problem-solving and Learning Goals: Two Experiments for Demonstrating Assistance Dilemma</i> .....	2008
Kazuhisa Miwa, Hitoshi Terai, Ryuichi Nakaike	
<i>A Quantum Probability-theoretic account of human judgment using Positive-Operator-Valued Measures</i> .....	2014
Takayuki Miyadera, Steven Phillips	
<i>Comparison of neural responses between exogenous and endogenous rule shifting in cued switching task; an ERPs study</i> .....	2019
Maki Miyajima, Atuhito Toyomaki, Ichiro Kusumi, Tsukasa Koyama	
<i>Children's understanding of hidden emotion, theory of mind, and peer relationship</i> .....	2025
Ai Mizokawa, Masuo Koyasu	
<i>The Effect of Visually and Phonologically Misleading Nonwords on Lexical Decisions of Native Japanese Readers</i> .....	2031
Rika Mizuno, Takao Matsui	
<i>The Effects of Mental Imagery and Embodied Action on L2 Word Learning</i> .....	2037
Laura Morett, Ray Gibbs, Brian MacWhinney	
<i>How the Hands Cue the Mind: The Effects of Iconicity and Enactment on Sign Language Acquisition</i> .....	2043
Laura Morett	
<i>Does thinking make you biased? The case of the engineers and lawyers problem.</i> .....	2049
Kinga Morsanyi, Simon Handley	
<i>Inferring aspectuality on French sentences: a minimalist approach</i> .....	2055
Damien Munch, Jean-Louis Dessalles	
<i>Does Analogy Facilitate Transitive Inference in Young Children?</i> .....	2061
Milena Mutafchieva, Kristina Gotseva, Boicho Kokinov	
<i>Cognitive Styles in Two Cognitive Sciences</i> .....	2067
James Myers	
<i>Process of Improvisational Contemporary Dance</i> .....	2073
Yuko Nakano, Takeshi Okada	
<i>Learning Containment Metaphors</i> .....	2079
Sushobhan Nayak, Amitabha Mukerjee	
<i>Interaction of Word Learning and Semantic Category Formation in Late Talking</i> .....	2085
Aida Nematzadeh, Afsaneh Fazly, Suzanne Stevenson	
<i>Modeling dilution effects in perceptual load search tasks</i> .....	2091
Kleanthis Neokleous, Marios Avraamides, Christos Schizas	
<i>A Cultural Decision-Making Model for Negotiation based on Inverse Reinforcement Learning</i> ....	2097
Elnaz Nouri, Kallirroi Georgila, David Traum	

<i>State effects of action video-game playing on visuospatial processing efficiency and attention among experienced action video-game players</i> .....	2103
Takashi Obana, Maria Kozhevnikov	
<i>Variation of Characteristics of Reading and Writing Difficulties in Japanese Children with Learning Disabilities</i> .....	2109
Shino Ogawa, Miwa Fukushima-Murata, Namiko Kubo-Kawai, Tomoko Asai, Hiroko Taniai, Nobuo Masataka	
<i>Dynamic estimation of emphasizing points for user satisfaction evaluations</i> .....	2115
Yoshimasa Ohmoto, Takashi Miyake, Toyoaki Nishida	
<i>The Vowel-Size Relationship Re-Examined Using Speeded Classification.</i> .....	2121
Yuka Ohtake, Etsuko Haryu	
<i>Building Conceptual Dictionary for Providing Common Knowledge in the Integrated Narrative Generation System</i> .....	2126
Kensuke Oishi, Yasunari Kurisawa, Mami Kamada, Itaru Fukuda, Taisuke Akimoto, Takashi Ogata	
<i>Working memory's meager involvement in sentence repetition tests</i> .....	2132
Eve Okura, Deryle Lonsdale	
<i>Changing Discriminatory Norms Using Models of Conceptually-Mediated Cognition and Cultural Worldviews</i> .....	2138
Daniel Olsher	
<i>The Role of Comparison Processes in the Induction of Schemas for Design Styles</i> .....	2144
Takanobu Omata, Keith Holyoak	
<i>A State-Event Transformation Mechanism for Generating Micro Structures of Story in an Integrated Narrative Generation System</i> .....	2150
Kou Onodera, Taisuke Akimoto, Takashi Ogata	
<i>Teaching the Perceptual Structure of Algebraic Expressions: Preliminary Findings from the Pushing Symbols Intervention</i> .....	2156
Erin Ottmar, David Landy, Robert Goldstone	
<i>Improving Representational Competence in Chemistry with Model-Based Feedback</i> .....	2162
Shamin Padalkar, Mary Hegarty	
<i>Cognitive Biases in a Geospatial Intelligence Analysis Task: An ACT-R Model</i> .....	2168
Jaehyon Paik, Peter Pirolli, Christian Lebiere, Matthew Rutledge-Taylor	
<i>Can native-language perceptual bias facilitate learning words in a new language?</i> .....	2174
Bozena Pajak, Sarah Creel, Roger Levy	
<i>Seeing who sees: Contrastive access helps children reason about other minds</i> .....	2180
Kathie Pham, Elizabeth Bonawitz, Alison Gopnik	
<i>Children and Pragmatic Implicatures: A Test of the Pragmatic Tolerance Hypothesis with Different Tasks</i> .....	2186
Katrijn Pipijn, Walter Schaeken	
<i>Why is A Few Sometimes A Lot?</i> .....	2192
Amanda Pogue, Adel Jalabi, Mathieu Le Corre	

<i>Toward machines that behave ethically better than humans do</i> .....	2198
Matthijs Pontier, Johan Hoorn	
<i>Modeling the Influence of Cognitive Fluency and Stereotype Threat on the Processing of Implicit Attitudes</i> .....	2204
Boon-Kiat Quek, Andrew Ortony	
<i>Modeling the Effect of Evaluative Conditioning on Implicit Attitude Acquisition and Performance on the Implicit Association Test</i> .....	2210
Boon-Kiat Quek, Andrew Ortony	
<i>Constraints, Inferences, and the Shortest Path: Which paths do we prefer?</i> .....	2216
Marco Ragni, Jan Wiener	
<i>Think Outside the Box: The Effects of Cognitive Training on Creative Problem Solving</i> .....	2222
Jared Ramsburg, Robert Youmans	
<i>Changing Global Warming Beliefs with Scientific Information: Knowledge, Attitudes, and RTMD (Reinforced Theistic Manifest Destiny Theory)</i> .....	2228
Michael Ranney, Dav Clark, Daniel Reinholz, Sarah Cohen	
<i>Evidence that Threatening Situations Enhance Creativity</i> .....	2234
Sean Riley, Liane Gabora	
<i>Automatic selection of eye tracking variables in visual categorization for adults and infants</i> .....	2240
Samuel Rivera, Catherine Best, Hyungwook Yim, Aleix Martinez, Vladimir Sloutsky, Dirk Walther	
<i>Categorisation in High and Low Schizotypes</i> .....	2246
Aaron Roberts, Nick Braisby	
<i>Inferring Metaphoric Structure from Financial Articles Using Bayesian Sparse Models</i> .....	2252
Martin Sälzle, Mark Keane	
<i>Conceptual Change through Socially Constructive Interaction in the Classroom</i> .....	2258
Moegi Saito, Naomi Miyake	
<i>Strategy Changes in Causal Structure Learning: The Role of Task Complexity</i> .....	2264
Motoyuki Saito, Tsuneo Shimazaki	
<i>The Comprehension of Adjective Metaphors Is Selectively Affected By Negative Meanings Associated With Adjectives As Vehicles</i> .....	2270
Maki Sakamoto, Miho Sumihisa, Takuya Matsumoto, Akira Utsumi	
<i>Problem-Solving Strategy Selection in Relation to Formal Schooling</i> .....	2276
Mennat-Allah Saleh, Christian Sturm	
<i>Shared Book Reading between Mother and Infant Facilitates The Frequency of Joint Attention</i> ..	2282
Ayumi Sato, Ichiro Uchiyama	
<i>A Computational PDP Model for Explaining Automatic Imitation</i> .....	2288
Matthias Scheutz, Bennett Bertenthal	
<i>Taking Development Seriously: Modeling the Interactions in the Emergence of Different Word Learning Biases</i> .....	2294
Savannah Schilling, Clare Sims, Eliana Colunga	

<i>Interactions between abstract actions and apparent distance</i> .....	2300
Kathryn Sears, Jessica Lesky, David Landy	
<i>Vection (self-motion perception) alters cognitive states, cognition of time, mental number line and personality</i> .....	2306
Takeharu Seno, Shuichiro Taya, Yuki Yamada, Keiko Ihaya, Hiroyuki Ito, Shoji Sunaga	
<i>Conscious and unconscious thought preceding complex decisions: The influence of taking notes and intelligence.</i> .....	2310
Aline Sevenants, Dieter Daniëls, Leen Janssens, Walter Schaeken	
<i>When Choice Effects Compete: An Account by Extended EBA Model</i> .....	2315
Kenpei Shiina	
<i>Creative Process of Improvised Street Dance</i> .....	2321
Daichi Shimizu, Takeshi Okada	
<i>Knowing When to Abandon Unproductive Learning</i> .....	2327
Thomas Shultz, Eric Doty, Frederic Dandurand	
<i>Listening to Thematic Music Prior to a Generation Task Causes Thematic Elements to Be Included in a Story Generation Task.</i> .....	2333
Cynthia Sifonis, William Fuss	
<i>Maps in the Head and Maps in the Hand</i> .....	2339
Kenny Skagerlund, David Kirsh, Nils Dahlbäck	
<i>When Students Don't Benefit From Attention Guidance in Animations: The Role of Working Memory in Learning From Animations</i> .....	2345
Irene Therese Skuballa, Rolf Schwonke, Alexander Renkl	
<i>The inductive potential of religion categories in Northern Ireland</i> .....	2351
Kirsty Smyth, Conor Pendergrast, Aidan Feeney, Coley John, Cole Edison, Ulrike Niens	
<i>Does number interference occur during sentence processing?</i> .....	2357
Katja Suckow, Roger P. G. van Gompel	
<i>The effect of text continuity on spatial representation</i> .....	2363
Masashi Sugimoto, Takashi Kusumi	
<i>Stress assignment in the development of reading aloud: Nonword priming effects on Italian children</i> .....	2369
Simone Sulpizio, Simone Sulpizio, Magali Boureux, Cristina Burani, Chizuru Deguchi, Lucia Colombo	
<i>Doppel Teleoperation System: Isolation of physical traits and intelligence for personality study</i> ..	2375
Hidenobu Sumioka, Shuichi Nishio, Erina Okamoto, Hiroshi Ishiguro	
<i>Individuals' process of metaphor interpretations and interestingness cognition</i> .....	2381
Tomohiro Taira, Takashi Kusumi, Akira Utsumi	
<i>Transmission of Rumor and Criticism in Twitter after the Great Japan Earthquake</i> .....	2387
Yuko Tanaka, Yasuaki Sakamoto, Toshihiko Matsuka	

<i>Ontological Properties of Animals in a Children's Dictionary With and Without Common-Sense Knowledge</i> .....	2393
Julia Taylor, Victor Raskin, Christian Hempelmann	
<i>An Experimental Examination of Emergent Features in Metaphor Interpretation Using Semantic Priming Effects</i> .....	2399
Asuka Terai, Robert Goldstone	
<i>An fMRI Investigation of Feature-Emergence-related Activation within Metaphor Comprehension</i>	2405
Asuka Terai, Naoko Kuriyama, Masanori Nakagawa, Kimihiko Yamagishi, Takashi Kusumi, Koji Jimura	
<i>Explanation Reconstruction through Reinterpretation of Key Facts</i> .....	2411
Hitoshi Terai, Kazuhisa Miwa, Shota Matsubayashi	
<i>Automatic Detection of Metonymies using Associative Relations between Words</i> .....	2417
Takehiro Teraoka, Ryuichiro Higashinaka, Jun Okamoto, Shun Ishizaki	
<i>Flexible sequence learning in a SOM model of the mirror system</i> .....	2423
Serge Thill, Josef Behr, Tom Ziemke	
<i>Fuzzy Memory Theory and its Use in Cognitive Science</i> .....	2429
Chris Thornton	
<i>From Head to Toe: Embodiment Through Statistical Linguistic Frequencies</i> .....	2434
Richard Tillman, Vivek Datla, Sterling Hutchinson, Max Louwerse	
<i>Evidence for Modality-Specific Processes in Approximate Numerical Comparison</i> .....	2440
Midori Tokita, Akira Ishiguchi	
<i>Embodied Communication Practices in Instructive Interaction between Musicians</i> .....	2446
Jackson Tolins	
<i>Viewing and performing actions can change what you see</i> .....	2451
Alexia Toskos Dils, Stephen Flusberg, Lera Boroditsky	
<i>Honoring Different Ontological Boundaries: The Role of Language in Category Formation</i> .....	2457
Duc Tran, Hanako Yoshida	
<i>Writing facilitates learning to read in Chinese through reduction of holistic processing: A developmental study</i> .....	2463
Ricky Van Yip Tso, Terry Kit-fong Au, Janet Hui-wen Hsiao	
<i>Temporal Dynamics of Action Perception: The Role of Biological Appearance and Motion Kinematics</i> .....	2469
Burcu Aysen Urgan, Markus Plank, Hiroshi Ishiguro, Howard Poizner, Ayse Pinar Saygin	
<i>Effects of Discourse Goals on the Process of Metaphor Production</i> .....	2475
Akira Utsumi, Kota Nakamura, Maki Sakamoto	
<i>Modeling Efficient Serial Visual Search</i> .....	2481
Bella Veksler, Wayne Gray	
<i>A Conceptual Network-Based Approach to Inferring the Cultural Evolutionary History of the Baltic Psaltery</i> .....	2487
Tomas Veloz, Ilya Tëmkin, Liane Gabora	

<i>Factors influencing children's display of surprise</i> .....	2493
Mandy Visser, Emiel Krahmer, Marc Swerts	
<i>Estimating Semantic Transparency of Constituents of English Compounds and Two-Character Chinese Words using Latent Semantic Analysis</i> .....	2499
Hsueh-Cheng Wang, Li-Chuan Hsu, Yi-Min Tien, Marc Pomplun	
<i>Visual Attention is Attracted by Text Features Even in Scenes without Text</i> .....	2505
Hsueh-Cheng Wang, Shijian Lu, Joo-Hwee Lim, Marc Pomplun	
<i>Implicit theories of the causes of weight gain in adults</i> .....	2511
Nicole Ware, Rachel Dryer	
<i>Relationship between Phonemes and Tactile-emotional Evaluations in Japanese Sound Symbolic Words</i> .....	2517
Junji Watanabe, Yuuka Utsunomiya, Hiroya Tsukurimichi, Maki Sakamoto	
<i>Actor-Observer Asymmetries in Judgments of Intentional Actions</i> .....	2523
Sarah Wellen, David Danks	
<i>Learning Causal Structure through Local Prediction-error Learning</i> .....	2529
Sarah Wellen, David Danks	
<i>Understanding each other: Defining a conceptual space for cognitive modeling</i> .....	2535
Robert West, David Pierre Leibovitz	
<i>Reading direction is sufficient to account for the optimal viewing position in reading: The case of music reading</i> .....	2540
Yetta Kwailing Wong, Janet Hui-wen Hsiao	
<i>Effects of Learning Order and Previous Language Experience in Novel Word Learning</i> .....	2546
Elizabeth Woods, Hanako Yoshida	
<i>Inference and culture : The distinction between low context culture and high context culture as a possible explanation for cultural differences in cognition</i> .....	2552
Hiroshi Yama, Norhayati Zakaria	
<i>"What" versus "How" in Nonvisual Whole-Body Movement</i> .....	2558
Naohide Yamamoto, Dale A. Hirsch	
<i>The Influence of Risk Aversion on Visual Decision Making</i> .....	2564
Ruixin Yang, Garrison Cottrell	
<i>Splitting Visual Focal Attention? It Probably Depends on Who You Are</i> .....	2570
Jit Yong Yap, Stephen Wee Hun Lim	
<i>Roles of Adult's Gestures and Eye Gaze in Whole or Object Part Presenting</i> .....	2576
Tetsuya Yasuda, Harumi Kobayashi	
<i>State-Trace Analysis of Sequence Learning by Simple Recurrent Networks</i> .....	2581
Fayme Yeates, Fergal Jones, Andy Wills, Ian P. L. McLaren	
<i>The Role of Attention in Three-way binding in Episodic Memory</i> .....	2587
Hyungwook Yim, Simon Dennis, Vladimir Sloutsky	



<i>Visual Context Effects on Thematic Role Assignment in Children versus Adults: Evidence from Eye Tracking in German</i> .....	2593
Lu Zhang, Pia Knoeferle	
<i>Argument Homogeneous and Structure Simplicity</i> .....	2599
Niina Ning Zhang	
<i>Modeling a Cognitively Limited Network in an Agent-Based Simulation</i> .....	2603
Changkun Zhao, Ryan Kaulakis, Jonathan Morgan, Jeremiah Hiam, Frank Ritter	
<i>Attention Modeling for Face Recognition via Deep Learning</i> .....	2609
Sheng-hua Zhong, Yan Liu, Yao Zhang, Fu-lai Chung	
<i>Cohesion Grading Decisions in a Summary Evaluation Environment: A Machine Learning Approach</i> .....	2615
Iraide Zipitria, Basilio Sierra, Ana Arruarte, Jon A. Elorriaga	

## Member's Abstracts

<i>The Effects of an Incubation Period on the Metaphor Creation Process</i> .....	2621
Keiga Abe	
<i>Trial Measurement of Implicit Attitude toward Violations in Nursing by the use of Implicit Association Test</i> .....	2622
Yuko Adachi, Shinnosuke Usui, Etsuko Nakagami-Yamaguchi, Akiko Yamada, Keun sik Park, Tatsuya Nakatani	
<i>Frame Augmented Language Model</i> .....	2623
Kisuh Ahn, Eunsuk Lim	
<i>Multi-Voxel Pattern Analysis Applied to the Language Switch in the Bilingual Brain—An fMRI Study</i> .....	2624
Hiroyuki Akama, Miao Mei Lei, Na Li, Brian Murphy	
<i>Visual prosody: The relationships between head movements and the verbs</i> .....	2625
Haruka Amatani	
<i>Determining the effect of ego-involvement on causal reasoning by using contingencies in causal situations</i> .....	2626
Yoshiko Arai	
<i>Which is Stronger? : Discriminative Learning of Sound Symbolism</i> .....	2627
Eiji Aramaki, Sachi Yasuda, Mai Miyabe, Satoshi Miura, Masaki Murata	
<i>The model comparison through orthography, phonology, and semantics</i> .....	2628
Shin-ichi Asakawa	
<i>Age-related Differences in Implicit Memory for Distractor Kanji Characters</i> .....	2629
Akihiro Asano, Etsuko T. Harada, Shoko Saito	
<i>Robots as learning partners in collaborative learning research</i> .....	2630
Jun Ashikaga, Takahiro Nakayama, Sho Inaba, Kenta Iyoki, Naomi Miyake	
<i>What do children hear? Japanese parents' use of numeral classifiers</i> .....	2631
Natsuki Atagi, Catherine Sandhofer	

<i>Comparison and contrast in novel objects categorization: the role of executive functions</i> .....	2632
Luc Augier, Jean-Pierre Thibaut	
<i>Comparison of the brain regions activated during comprehension of action sentences referring to unimanual or bimanual actions: An fMRI study</i> .....	2633
Shunji Awazu, Fumihiko Taya, Sayako Masuda, Shigeru Watanabe	
<i>Effectiveness of Transcranial Direct Current Stimulation on Medial Prefrontal Cortex in Aesthetic Judgement</i> .....	2634
Leila Azari Pishkenari, Hamed Ekhtiari, Mohammad Javad Hatami	
<i>Learning novel words with the help of morphological information</i> .....	2635
Sungbong Bae, Kwangoh Yi	
<i>Distant Border Color Is More Preferred In a Triple Color Combination</i> .....	2636
Ziba Bashardanesh, Ali Yoonessi	
<i>To peek and to peer: "visual" verb meanings are largely unaffected by congenital blindness</i> .....	2637
Marina Bedny, Jorie Koster-Hale, William Johnston, Lindsay Yazzolino, Rebecca Saxe	
<i>Cognitive typology</i> .....	2638
Chuluundorj Begz	
<i>Expert Memory in Blindfold Chess960 - An Interpretation in the light of LIDA</i> .....	2639
Eduardo Bermudez, David Dahmen, Henry Gonzalez	
<i>Vague Linguistic Expressions and the Problem of Equidistance in Verbal Response Scales</i> .....	2640
Franziska Bocklisch, Josef Krems	
<i>Going beyond the headlines: Narratives mitigate intergroup empathy bias</i> .....	2641
Emile Bruneau, Mina Cikara, Rebecca Saxe	
<i>Ordered Information and Word Learning: An Associative Learning Perspective</i> .....	2642
Joseph Burling, Hanako Yoshida	
<i>The Hierarchical Structure of Word Senses</i> .....	2643
Hee-Rahk Chae, Do-Il Hong	
<i>Word Segmentation Difficulty in Fourth Graders with Low Reading Achievement</i> .....	2644
Sau-chin Chen, Jenn-Yeu Chen	
<i>Is the Proximate Unit in Chinese Word Production Motivated by the Visual Prompt of the Task?</i>	2645
Train-Min Chen, Jenn-Yeu Chen	
<i>The Impact of Contextual Cues on Infant Categorization</i> .....	2646
I-Chen Chen, Marin Huang, I-Wen Yu, Pei-Ling Wang, Jon-Fan Hu	
<i>Knowing where to look: Conceptual knowledge guides fixation in an object categorization task</i> ...	2647
Lang Chen, Timothy Rogers	
<i>Mindset interacts with subjective knowledge but not fluency when affecting preferential judgment</i>	2648
Rongjuan Chen, Yasuaki Sakamoto	
<i>A Study of the Syntactic Category of Rhetorical Questions</i> .....	2649
Hongbo Chen	

<i>Eye Movement Patterns Reveal How Oriental People Group Objects</i> .....	2650
Fan-Ning Cheng, Yi-Rong Wu, Jo Pan, Gert Westermann, Hsueh-Chih Chen, Jon-Fan Hu	
<i>Reading and Writing Performance in School-aged Children with Specific Language Impairment or/and with Developmental Coordination Disorder Identified at Preschool Age</i> .....	2651
Rong-Ju Cherng, Hsiang-Chun Cheng, Jenn-Yeu Chen, Chia-Liang Tsai, Miao-Lin Shen	
<i>The development of numerical comparison in 2- to 4-year-old children</i> .....	2652
Pierina Cheung, Mathieu Le Corre	
<i>The examination of physiological factor on two context effects in multi-attribute decision making</i>	2653
Itsuki Chiba, Takashi Tsuzuki, Masashi Soma	
<i>A Classroom Study of Learning to Evaluate Scientific Evidence</i> .....	2654
Clark Chinn	
<i>Impact of Modified Cognitive Behavior Therapy on Children with Autism Spectrum Disorders</i> ...	2655
Suvarna Chinta, Bipin Indurkha	
<i>Can Articulating Aloud Offset Effects of Listening to L1 in a Foreign Accent?</i> .....	2656
Kit Cho, Laurie Beth Feldman	
<i>Blending Narrative Spaces of the Flashback Scenes in the Joint Security Area</i> .....	2657
Hye Rhang Cho, Seung Suk Nam, Sook Whan Cho	
<i>Blending Narrative Spaces of the Reenactment Scenes in the Thin Blue Line</i> .....	2658
Sook Whan Cho, Seung Suk Nam, Hye Rhang Cho	
<i>Ten-month-old Infants Prefer Comforters, not Helpers</i> .....	2659
Hui-Mei Chow, Stephanie Jui-chi Chen, Geroldene H. T. Tsui, Ping-Hui Chiu, Chia-Huei Tseng	
<i>Comprehension of Representational Gestures</i> .....	2660
Kawai Chui	
<i>Task-specific conflict monitoring and cognitive control in the prefrontal cortex</i> .....	2661
Chongwook Chung, Chobok Kim, Jeounghoon Kim	
<i>The Development of Orthographic Awareness for Radical Properties of Chinese Characters in Young Children</i> .....	2662
Yi-Ling Chung, Pei-Yu Luo, Hsueh-Chih Chen, Li-Yun Chang, Jon-Fan Hu	
<i>The Dynamics of Sentence (In)comprehension</i> .....	2663
Gregory Cox, Melody Dye, Seth Frey	
<i>Aquisition of multipliers: learning a new class of number</i> .....	2664
Meghan Dale, Mathieu Le Corre	
<i>Describing faces: conventionalizing ontologies through dialogic interaction</i> .....	2665
Nicolas Davidenko, Gregory Mills	
<i>The effect of speaker's identity on syntactic processing: Evidence from verb-gender agreement in Slovak</i> .....	2666
Doug J. Davidson, Adriana Hanulíková, Manuel Carreiras	

<i>Joint Evaluation and Trend Information Mitigates the Disposition Effect</i> .....	2667
Kyung Soo Do, Hee-Yeon Kim, Rae-yeop Park	
<i>Effects of Scaffolded Feedback and Confidence of Incorrect Answers on Retention</i> .....	2668
Kyung Soo Do, Hanna Kim	
<i>Computing Humorous Metaphors</i> .....	2669
Pawel Dybala, Kohichi Sayama	
<i>Expanding the Transformation Paths in a Mutual Transformation Mechanism of Music and Narrative</i> .....	2670
Jun Endo, Taisuke Akimoto, Takashi Ogata	
<i>Eye tracking differences in respondent behaviour across multiple survey modes</i> .....	2671
Tom Foulsham, Olena Kaminska	
<i>Automatic facilitation of social behavior by implicit inferring of social intention</i> .....	2672
Haruaki Fukuda, Hiroaki Suzuki, Ayumi Yamada	
<i>Quantifying linguistic coordination</i> .....	2673
Riccardo Fusaroli, Kristian Tylén	
<i>How a Quantum Approach to Memory Incorporates Contextuality and Potentiality</i> .....	2674
Liane Gabora, Kirsty Kitto	
<i>The Naturalization of Concepts between Computational Intractability and Cognitive Theories</i> ...	2675
Francesco Gagliardi	
<i>Selection of decision rules in dynamic decision making</i> .....	2676
Jean-François Gagnon, Marie-Ève St-Louis, Sebastien Tremblay	
<i>Annual Cognitive Modeling Competition</i> .....	2677
Kevin Gluck	
<i>Do Young Children Habituate to their Classroom Environment?</i> .....	2678
Karrie Godwin, Anna Fisher	
<i>Phonological neighbourhood clustering effects on verbal short-term memory</i> .....	2679
Winston Goh	
<i>Children's sensitivity to informant's inductive efficiency and learner's epistemic states in pedagogical contexts</i> .....	2680
Hyowon Gweon, Patrick Shafto, Joshua Tenenbaum, Laura Schulz	
<i>The End is Near: Anticipating the end of a sentence</i> .....	2681
Lance Hahn	
<i>Individual differences and phonetic aptitude in the earliest stages of L2 acquisition</i> .....	2682
Adriana Hanulikova, Dan Dediu, Zhou Fang, Jana Basnakova, Falk Huettig	
<i>What do numbers tell you? The effects of data-presentation design on the prolonged use of health-related life-log tools</i> .....	2683
Etsuko T. Harada, Satomi Yoshiyama	
<i>Japanese script types in written names create the images of their referents</i> .....	2684
Aya Hatano, Masahiro Amagase, Jun Kawaguchi	

<i>Is past information useful for evaluating present covariation information? : Effect of irrelevant information on causal judgment</i> .....	2685
Ikuko Hattori, Masasi Hattori	
<i>Visual cognition of "speed lines" in comics: Experimental study on speed perception</i> .....	2686
Hiromasa Hayashi, Goh Matsuda, Yoshiyuki Tamamiya, Kazuo Hiraki	
<i>Ontology Architecture of a Neuro-psychoanalytical, Computational Model</i> .....	2687
Isabella Hinterleitner	
<i>The absence of phonetic symbolism to the novel speech sound -comparison of cross-modal correspondence between Chinese and Japanese speakers using Chinese speech sound-</i> .....	2688
Sachiko Hirata, Shinichi Kita	
<i>Longitudinal observation of action slips: A case study of a young child</i> .....	2689
Naoya Hirose	
<i>The role of social contexts in adults' word learning</i> .....	2690
Masako Hirotani, Koji Shimada, Shuntaro Okazaki, Hiroki C. Tanabe, Norihiro Sadato	
<i>A Cognitive-Educational Approach to the Verb mek-ta in Korean</i> .....	2691
Do-Il Hong, Hee-Rahk Chae	
<i>Connectionist Modeling of Frequency and Regularity in Mandarin Relative Clause Processing</i> ...	2692
Yaling Hsiao, Maryellen MacDonald	
<i>A Corpus Survey of Chinese Individual Classifiers</i> .....	2693
Shuping Huang, Jenn-Yeu Chen	
<i>Property activation during the Interpretation of Noun-Noun Compounds</i> .....	2694
Jie Huang	
<i>Modeling How Naming Experiences Bias Sorting Performance</i> .....	2695
Yu-Sheng Hung, Yun Li, Hsueh-Chih Chen, Jon-Fan Hu	
<i>The Effect of Priming of Individualism/Collectivism on the Müller-Lyer Illusion</i> .....	2696
Yiwon Hyun, Donghoon Lee, Hyunjung Shin, Myeong-ho Sohn	
<i>How novices get skills without supervisors' instructions: through analysis of skills mastery process</i> .....	2697
Jun Ichikawa, Yugo Takeuchi	
<i>The relationship between depressive tendency and relative metacomprehension accuracy.</i> .....	2698
Kenji Ikeda, Yosuke Hattori, Shinji Kitagami	
<i>A Narratological Mechanism for Generating Macro Structures of Story in an Integrated Narrative Generation System</i> .....	2699
Shohei Imabuchi, Takashi Ogata	
<i>A cellular automaton model of ambiguity aversion</i> .....	2700
Kenryo Indo	
<i>The effects of frameworks and examples in learning how to solve word problems</i> .....	2701
Miwa Inuzuka, Hirosuke Tanimoto, Hiroko Kobayashi	

<i>Social Projection as a Universal Strategy in Mental State Inference: Cultural Differences in Utilization of Stereotyping</i> .....	2702
Tatsunori Ishii, Masanori Takezawa	
<i>Functions for mutual interaction with the mind reading</i> .....	2703
Satoru Ishikawa	
<i>The anger superiority effect in children with and without autism</i> .....	2704
Tomoko Isomura, Hiroyasu Ito, Shino Ogawa, Miwa Fukushima, Masahiro Shibasaki, Nobuo Masataka	
<i>Culture, perception, and artistic visualisation: A comparative study of children's drawings in three Siberian cultural groups</i> .....	2705
Kirill Istomin, Jaroslava Bagdasarova, Patrick Heady	
<i>How do children with autism solve logic puzzle?</i> .....	2706
Hiroyasu Ito, Nobuo Masataka	
<i>Effects of Base Rates and Likelihoods on Intuitive Probabilistic Judgments</i> .....	2707
Tomoko Itoh	
<i>Representational Form and Metaphorical Word Use</i> .....	2708
Anja Jamrozik, Micah Goldwater, Eyal Sagi, Dedre Gentner	
<i>Interaction between ability and use of scaffold in EFL vocabulary learning system</i> .....	2709
Felix Jimenez, Masayoshi Kanoh	
<i>Qualitative differences in sequence planning with everyday objects in traumatic brain injured individuals</i> .....	2710
Arianne Johnson, Scott Grafton	
<i>Differential Effects of the Cultural Orientation Dimensions on Global Precedence</i> .....	2711
Mijung Joo, Hyunmin Kang, Hyunjung Shin, Jaesik Lee	
<i>Hierarchical Category Structures Facilitate Acquisition of Probabilistic Relational Categories.</i> ...	2712
Wookyoung Jung, John Hummel	
<i>Dynamic Effects of Perceptual and Categorical Similarity on Recognition Memory</i> .....	2713
George Kachergis, Gregory Cox, Richard Shiffrin	
<i>Asymmetry of McGurk Effect Depending on the Orientation of Faces</i> .....	2714
Masayo Kajimura, Hiroshi Ashida	
<i>An Optimality Theoretical Analysis of Urge Interactions in Toda's (1982) Emotional Fungus-Eater Robots</i> .....	2715
Yasuo Kaneko	
<i>A signal detection analysis of the effects of repeated context on visual search</i> .....	2716
Ryan Kasper, Miguel Eckstein, Barry Giesbrecht	
<i>Cultural variations in authority management in interaction</i> .....	2717
Yasuhiro Katagiri	
<i>Differences in emotional bias effect on working memory in elderly.</i> .....	2718
Maya Katsuhara, Mariko Osaka, Naoyuki Osaka	

<i>Mental rotation of pictured body stimuli: Involvement of visual representation of the stimuli</i> ....	2719
Tsubasa Kawasaki, Takahiro Higuchi	
<i>Extracting the Musical Schema from Traditional Japanese, Chinese and German Folk Songs</i> ....	2720
Akihiro Kawase	
<i>Implicit learning on the order of the visual stimuli under interocular suppression</i> .....	2721
Kaede Kido, Shogo Makioka	
<i>Hierarchical Slow-Feature Models of Gesture Conversation</i> .....	2722
Jiseob Kim, Sooyong Jang, Eun-Sol Kim, Byoung-Tak Zhang	
<i>‘Is this right?’ or ‘Is that wrong?’: Evidence from Dynamic Eye-Hand Movement in Decision Making</i> .....	2723
Eun-Sol Kim, Jiseob Kim, Thies Pfeiffer, Ipke Wachsmuth, Byoung-Tak Zhang	
<i>State Anxiety and the Processing of Covariation Information in Causal Reasoning</i> .....	2724
Young Il Kim, Kyung Il Kim	
<i>Cultural priming and the scene perception</i> .....	2725
Bia Kim, Yoonkyoung Lee, Donghoon Lee, Goeun Lee, HyunJung Shin	
<i>Complex Network Analysis of Social Relationships and Personality from TV Drama Dialogues</i> ..	2726
Joon Shik Kim, Chung-Yeon Lee, Minsu Zhang, Jun-Hee Nam	
<i>A Neuroethological Approach to Robotics</i> .....	2727
DaeEun Kim	
<i>The effectiveness of English teaching integration model based on task-based approach</i> .....	2728
Eun sook Kim	
<i>Nominal Number and Semantics of Common Nouns in Numeral Classifier Languages</i> .....	2729
Jeehoon Kim	
<i>Automaticity in Motor Learning: Evidence from Visuo-motor Tracking Performance and Pupil Dilation</i> .....	2730
Satoshi Kobori, Yosuke Abe, Shogo Nakazono	
<i>Self-Organization of Policy by Symmetric Reasoning and its Application to Reinforcement Learning</i> .....	2731
Yu Kohno, Tatsuji Takahashi	
<i>Study on Facilitation of Problem Posing by Learning Examples through Reproduction</i> .....	2732
Kazuaki Kojima, Kazuhisa Miwa, Tatsunori Matsui	
<i>Egocentric and allocentric frame of reference in virtual maze navigation</i> .....	2733
Takatsugu Kojima	
<i>Consciousness and the language faculty: awareness, qualia and natural language using agents</i> ...	2734
Piotr Konderak	
<i>Three co-creation stages in formation of symbol communication systems</i> .....	2735
Takeshi Konno, Junya Morita, Akihito Kishino, Takashi Hashimoto, Jiro Okuda, Maki Suzuki	
<i>Theory of Mind network encodes how you know what you know in blind and sighted adults</i> .....	2736
Jorie Koster-Hale, Rebecca Saxe, Marina Bedny	

<i>A joint ideomotor effect increases the inter-brain oscillation between two people engaged in a Japanese Ouija board “Kokkuri-san”</i> .....	2737
Kenta Kubo, Kentaro Katahira, Kazuo Okanoya, Masato Okada, Nobuyuki Kawai	
<i>A Framework of Sentence Generation Mechanism for a Narrative Generation System</i> .....	2738
Shinya Kumagai, Sou Funakoshi, Junpei Ono, Taisuke Akimoto, Takashi Ogata	
<i>An on-line model of quantifiers interpretation in Japanese</i> .....	2739
Takeo Kurafuji, Masakatsu Inoue, Michinao Matsui	
<i>Beauty and Cuteness in Peripheral Vision</i> .....	2740
Kana Kuraguchi, Hiroshi Ashida	
<i>Motion-related brain activity enhanced by motion representation during metaphor understanding</i>	2741
Naoko Kuriyama, Asuka Terai, Takashi Kusumi, Masanori Nakagawa, Kimihiko Yamagishi, Koji Jimura	
<i>The influence of redundant and idiosyncratic attributes on coherence within a category and contrast between categories</i> .....	2742
Ikuko Kyoya, Masaomi Oda	
<i>Assessment and refinement of an intelligent tutor for complex decision making</i> .....	2743
Daniel Lafond, Sebastien Tremblay, Michel DuCharme, Marie-Ève St-Louis, Jean-François Gagnon	
<i>Evidence for a phonology-specific learning mechanism</i> .....	2744
Regine Lai, Jeffrey Heinz	
<i>Deceptive strategy choice as decision making under risk</i> .....	2745
Tei Laine, Kayo Sakamoto	
<i>Inner Speech and Task Switching in Adults</i> .....	2746
Lucie Laurent, Jean-Louis Millot, Patrice Andrieu, Valérie Camos, Caroline Floccia, Fabien Mathy	
<i>Enhanced Visual Processing in Perihand Space: Effects of Handedness</i> .....	2747
Nathalie Le Bigot, Marc Grosjean	
<i>The influence of cultural dispositions on the scene perception</i> .....	2748
Yoonkyoung Lee, Bia Kim, Yiwon Hyun, Cheonwoo Shin, Jaesik Lee, HyunJung Shin	
<i>Framing Neuroethics: An Integrated-Unified Scientific Analogy (I-USA) Perspective</i> .....	2749
Sang Bok Lee, Jonathan Jiseop Lee	
<i>How they pick out the answer in multiple choice questionnaire: Independent-self versus interdependent-self</i> .....	2750
Min-Seop Lee, Jeong Ryu, Dayk Jang	
<i>Information Structure, Alternatives, and Scalar Implicatures</i> .....	2751
Chungmin Lee	
<i>Observational category learning increases sensitivity to prototypical and correlational information</i>	2752
Kimery Levering, Kenneth Kurtz	



<i>The thinking behind decisions to spread disaster-related tweets</i> .....	2753
Huaye Li, Rongjuan Chen, Yasuaki Sakamoto	
<i>Size Effect During Emergence of Symbolic Communication System Revealed by Agent-based Modelling</i> .....	2754
Guanhong Li, Takashi Hashimoto	
<i>Understanding human error detection as task interruptions</i> .....	2755
Simon Y. W. Li	
<i>Assessment of children's summarization ability: An alternative measure of reading comprehension</i> .....	2756
Chi-Shun Lien, Hung-Hui Chen	
<i>Cross-linguistic Similarities and Differences in Causative Constructions</i> .....	2757
Eunsuk Lim, Kisuh Ahn, Hee-Rahk Chae	
<i>Unpredictable Grunting in Tennis is More Distracting</i> .....	2758
Ahnate Lim, Alan Kingstone, Scott Sinnett	
<i>Enhancing critical thinking and learning outcomes at the university: A pedagogical perspective</i> ..	2759
Stephen Wee Hun Lim	
<i>Is it Possible to Train the Approximate Number System?</i> .....	2760
Marcus Lindskog, Anders Winman, Peter Juslin	
<i>Cognitive Choreography in Mental Algebra Task</i> .....	2761
John Lindstedt, Wayne Gray	
<i>Undergraduates' online search strategies and visual attention distribution</i> .....	2762
Wan-Yi Liu, Meng-Jung Tsai	
<i>Action and affordances in nominal classifier systems</i> .....	2763
Marit Lobben	
<i>An ERP study on L2 grammatical aspect processing in Japanese</i> .....	2764
Shengyan Long, Yusaku Tsuyuguchi, Manami Sato, Hiromu Sakai	
<i>The Role of Embodiment on Children's Memory Recall through LEGO Robotics Activities</i> .....	2765
Carol M. Lu, John B. Black, Seokmin Kang	
<i>Social elements are not a must for preverbal infants' learning in an interactive event</i> .....	2766
Yuen Ki Ma, Hiu-Mei Chow, Jaclyn Yeung, Anna Wing Yee Ho, Chia-Huei Tseng	
<i>A possible source of phonological deficit in developmental dyslexia: Counter-evidence for universal phonological grammar</i> .....	2767
Norbert Maïonchi-Pino, Yasuyuki Taki, Satoru Yokoyama, Kei Takahashi, Annie Magnan, Hiroshi Hashizume, Jean Écalte, Ryuta Kawashima	
<i>Aberrant sense of agency in patients with schizophrenia: confusion of temporal causality during intentional action</i> .....	2768
Takaki Maeda, Motoichiro Kato, Keisuke Takahata, Tsukasa Okimura, Hajime Asama, Masaru Mimura	
<i>Student perspectives on critical and other thinking skills: Some cultural similarities and differences</i> .....	2769
Emmanuel Manalo, Takashi Kusumi, Masuo Koyasu, Yasushi Michita, Yuko Tanaka	

<i>The role of exploratory decision-making in enhancing episodic memory</i> .....	2770
Doug Markant, Sarah Dubrow, Lila Davachi, Todd Gureckis	
<i>Visual processing deficits due to HIV: Evidence from temporal order judgments</i> .....	2771
Liron Marotz, Scott Sinnett, Cecilia M. Shikuma	
<i>Collaboration for Building Sustainable Knowledge</i> .....	2772
Hiroyuki Masukawa, Ikuro Endo	
<i>Increases in Children's Semantic Organization Predict Category-based Reasoning</i> .....	2773
Bryan Matlen, Karrie Godwin, Anna Fisher	
<i>Efficiency of the Cognitive Bias in Grammar Acquisition</i> .....	2774
Ryuichi Matoba, Makoto Nakamura, Shingo Hagiwara, Satoshi Tojo	
<i>CyclingMusic &amp; CyclingMelody: A System for Enriching Scenery Experience in Cycling by Real-Time Synaesthetic Sonification of Passing Landscape</i> .....	2775
Masaki Matsubara, Satoshi Kuribayashi, Haruka Nukariya, Yasuaki Kakehi	
<i>Does a humanoid robot in front of you activate your mirror neuron system?</i> .....	2776
Goh Matsuda, Kazuo Hiraki, Hiroshi Ishiguro	
<i>The effects of exposure sequence and duration on mere exposure effect</i> .....	2777
Ken Matsuda, Eriko Sugimori, Takashi Kusumi	
<i>The Computational Process of "And-type" Conditionals in Japanese</i> .....	2778
Michinao Matsui	
<i>Connecting input filtering and selection in language evolution</i> .....	2779
Luke Maurits, Tom Griffiths	
<i>When Lexical Development does not Spurt; the Case of Williams Syndrome Children</i> .....	2780
Julien Mayor	
<i>Self-directed information selection aids learning of logical rules</i> .....	2781
John V. McDonnell, Devin Domingo, Todd M. Gureckis	
<i>Making and breaking procedural conventions in dialogue</i> .....	2782
Gregory Mills	
<i>Expanding children's perspectives on art work by robot participation</i> .....	2783
Masaki Miyake, Tomoki Hirano, Naomi Miyake	
<i>The effect of harmonization between word meaning and typography impression on implicit memory</i> .....	2784
Kozue Miyashiro, Etsuko T. Harada	
<i>Short-term memory for tonal and verbal information: Comparison with absolute and non-absolute pitch possessors</i> .....	2785
Shiho Miyazawa, Akihiro Tanaka, Takehiko Nishimoto	
<i>Behavioral priming contributes to the subsequent recognition performance</i> .....	2786
Kiyohumi Miyoshi, Hiroshi Ashida	
<i>How much do you trust me? Economic decision-making and ingroup and outgroup membership</i> .	2787
Rosalba Morese, Daniela Rabellino, Angela Ciaramidaro, Marco R. Elena, Francesca M. Bosco, Rosalba Rosato, Bruno G. Bara	

<i>Using syntactic priming to facilitate the language production of novice Japanese EFL learners ..</i>	2788
Miwa Morishita	
<i>Speed reading training and visual span .....</i>	2789
Aiko Morita	
<i>Analytical method of Japanese folk tale for story generation .....</i>	2790
Hitoshi Morita	
<i>Facilitation and inhibition in the spatiotemporal template for detecting target ring .....</i>	2791
Masayoshi Nagai, Patrick J. Bennett, Allison B. Sekuler	
<i>Analysis method focusing on repeated and not repeated verbalizations during human interface operation .....</i>	2792
Yukari Nagai, Saori Noda, Georgi V. Georgiev, Toshiharu Taura, Deny Willy	
<i>The effects of listener's familiarity with a talker's voice on the speech recognition in noisy condition .....</i>	2793
Chikako Nagaoka, Naoshi Hiraoka, Shintaro Funahashi	
<i>A Comparison of Experienced, Novice Counselor and Non-counselor in Recall of Client-Presented Information in Therapeutic Interview .....</i>	2794
Chika Nagaoka, Sakiko Yoshikawa, Tomoko Kuwabara, Yasuhiro Oyama, Chiriho Hatanaka, Motoki Watabe, Masashi Komori	
<i>The Narrative Structure of Nostalgia Cognition and Film .....</i>	2795
Yuya Naito, Akihito Kanai	
<i>TCieX: An Approach toward Communicating Weight through Pseudo-Haptic Feedback Mechanisms .....</i>	2796
Kumiyo Nakakoji, Yasuhiro Yamamoto	
<i>The effects of paralinguistic cues in a teacher's responses to the students' utterances in a moral class .....</i>	2797
Keiko Nakamoto, Ayumi Nishiyama	
<i>An agent-based model for the emergence of creoles .....</i>	2798
Makoto Nakamura, Shingo Hagiwara, Satoshi Tojo	
<i>Generating new product ideas by assemblage of different product components .....</i>	2799
Jun Nakamura, Yukio Ohsawa	
<i>Postal Addresses as an Assay of Cultural Cognition .....</i>	2800
Hiroko Nakamura, Hiroshi Yama, Gary L. Brase, Nasriah Zakaria, Yoshiko Arai, Norhayati Zakaria, Shafiz Affendi Mohd Yusof, Jun Kawaguchi	
<i>Changes in social motivation, and learning strategies, in PBL. ....</i>	2801
Yoshifumi Nakanishi, Takatoyo Umemoto, Kenshiro Tanaka	
<i>Phoneme exchange in, serial-position effect on, and lexical/semantic contributions to single-word production: An investigation using speech-error induction techniques in Japanese. ....</i>	2802
Masataka Nakayama, Shogo Kajimura, Masashi Sugimoto, Kaori Kuraya, Miyako Inoue, Ryo Ishibashi, Satoru Saito	
<i>An Analysis of Disfluencies in the Actor's Speech for Character Design .....</i>	2803
Seung Suk Nam, Hye Rhang Cho, Sook Whan Cho	

<i>The effect of a childbirth psychoeducation program on postnatal depression</i> .....	2804
Fei Wan Ngai, Sally Wai Chi Chan	
<i>Satisfaction evaluation and time perception for waiting time in ICT usage under dual task situation</i> .....	2805
Sumaru Niida, Satoshi Nakamura, Tomomi Moroga, Etsuko T. Harada, Satoshi Uemura	
<i>Multimodal Interactions Development Process in Collaborative Creation</i> .....	2806
Koshi Nishimoto, Mamiko Sakata	
<i>Task complementarity and response complementarity in the social Simon effect</i> .....	2807
Akio Nishimura, Koh Miyamoto, Kazuhiko Yokosawa	
<i>Reflexive orienting to other's gaze is modulated by the accuracy of recognition of emotional facial expressions.</i> .....	2808
Yuka Nishiyama, Jun Kawaguchi	
<i>A Case Study of Meta-cognitive Exploration of Facial Expressions</i> .....	2809
Takeshige Nishiyama, Hiromi Ochiai, Yuko Toukairin, Masaki Suwa	
<i>A cognitive emotional model for "intrinsic motivation"</i> .....	2810
Kohei Noda	
<i>What is the trial-and-error process of design thinking?</i> .....	2811
Hisataka Noguchi	
<i>Choosing unknown goods: An fMRI study of product choice</i> .....	2812
Ikuya Nomura, Kazuyuki Samejima, Kazuhiro Ueda, Yuichi Washida, Hiroyuki Okada, Takashi Omori	
<i>Clarifying position derived from sophisticated beliefs about the nature of knowledge-to-use</i> .....	2813
Ryota Nomura	
<i>A Method of interviewing to Constructively Generate a Narrative through Interactions between Interviewer and Interviewee - A Case Study to Examine Creative Thoughts of an Architectural Student -</i> .....	2814
Haruka Nukariya, Masaki Suwa	
<i>Do children who experience regret make better decisions? A developmental study of the behavioral consequences of regret</i> .....	2815
Eimear O'Connor, Aidan Feeney, Teresa McCormack	
<i>Analysis of Human Solving Process of Constraint Satisfaction Problem</i> .....	2816
Hidemi Ogasawara, Yoshiyuki Matsuzawa, Masahide Ogawa	
<i>Towards the Development of Integrated Narrative Generation System as the Implementation of Literary Knowledge</i> .....	2817
Takashi Ogata, Taisuke Akimoto	
<i>Effect of Age-related Decline of Task Switching on the Task Sequences that Simulate Real Job</i> ...	2818
Keiji Ogata, Satoru Suto, Kazutaka Ueda, Takatsune Kumada, Tohru Ifukube	
<i>Rhetoric of Biopic and Viewer's Reconsideration : Isn't "story" an Obstacle?</i> .....	2819
Yukiko Ogawa, Akihito Kanai	

<i>Developmental Adjustment of Iconic Language in Care-Takers' Input</i> .....	2820
Masato Ohba, Noburo Saji, Mutsumi Imai, Tomoko Matsui	
<i>Effects of Language on Asynchronized Audiovisual Speech Perception</i> .....	2821
Hitoshi Ohnishi, Kaname Mochizuki	
<i>The effects of regularity in spatial inferences with and without local landmarks on spatial learning</i>	2822
Kayoko Ohtsu, Yoshihiro Ouchi	
<i>Computational model of the meaning acquisition of sentence-final particles</i> .....	2823
Natsuki Oka, Naohiro Nonoguchi, Chie Fukada, Motoyuki Ozeki	
<i>Is the breaking point in mental number line in Japanese children five or ten?</i> .....	2824
Masahiko Okamoto, Sari Nakamura	
<i>Influences of a potential interlocutor on the utterances of the speakers who try to describe objects</i>	2825
Junji Okamoto, Saori Ushiyama	
<i>Delay of word order development in Japanese? Evidence from a preferential looking study with</i> <i>19 and 30-month-old children</i> .....	2826
Akira Omaki, Romy Lassotta, Tessei Kobayashi, Luigi Rizzi, Julie Franck	
<i>Effects of supraliminal and subliminal hint priming on insight problem solving.</i> .....	2827
Ryo Orita, Masasi Hattori	
<i>Transfer of learning from project activities to individual leaning</i> .....	2828
Naoko Osada	
<i>Ranges of storage item sizes in complex working memory span tasks: Latent-variable analysis of</i> <i>the memory load</i> .....	2829
Kazunori Otsuka, Makoto Miyatani	
<i>Does the detection of mind wandering require attentional resources?</i> .....	2830
Sho Otsuka, Takahiro Sekiguchi	
<i>Analysis of the relationship between writing skills and evaluation in an expository writing</i> <i>assignment</i> .....	2831
Hiroko Otsuka, Mio Tsubakimoto, Hiroshi Numata	
<i>Loosely symmetric heuristics as the basis for biases and the empirical Bayes methods</i> .....	2832
Kuratomo Oyo, Tatsuji Takahashi	
<i>How simple explanations change our minds and why we prefer them</i> .....	2833
Michael Pacer, Tania Lombrozo	
<i>The Influence of Culture: Thematic versus Taxonomic Categorization</i> .....	2834
Jo Pan, Yi-Rong Wu, Fan-Ning Cheng, Gert Westermann, Hsueh-Chih Chen, Jon-Fan Hu	
<i>Do Objects Matter for Infants' Formation of a Spatial Category?</i> .....	2835
Youjeong Park, Marianella Casasola	
<i>Cross-level Illusory Conjunction between Implied (Semantic) and Actual (Perceptual) Colors</i> ....	2836
Chan Jeong Park, Anna Wing Yee Ho, Geroldene H. T. Tsui, J. T. Y. Chan, X. Luo, Chia-Huei Tseng	

<i>Judgment under uncertainty is not always certainty-oriented</i> .....	2837
Youngjun Park, Kyungil Kim	
<i>The Effect of Perceptual Complexity on Affective Picture Processing</i> .....	2838
Taejin Park, Soodam Park	
<i>Effect of Saliency-Based Masking in Scene Classification</i> .....	2839
Tae-Suh Park, Byoung-Tak Zhang	
<i>fMRI evidence for sensitivity to coherence and context in the theory of mind network</i> .....	2840
Alexander Paunov, Jorie Koster-Hale, Rebecca Saxe	
<i>The Role of Linguistic Knowledge in Hue Perception</i> .....	2841
Katherine Phelps, Steve Duman, Kevin Gould, Les Sikos	
<i>fMRI of attention and automaticity in judgments from facial appearance</i> .....	2842
Ramsey Raafat, Nikos Konstantinou, Chris Frith, Nilli Lavie, Nick Chater	
<i>Cooperative Behavior in Multicultural Settings: The Contribution of Altruistic Punishment</i> .....	2843
Daniela Rabellino, Rosalba Morese, Angela Ciaramidaro, Bruno G. Bara, Rosalba Rosato, Francesca M. Bosco	
<i>Why so Stressful? The effect of coping strategies and social support to undergraduate students in Korea</i> .....	2844
Young Sun Ryu, Ha Rim Kim, Jeong Ryu	
<i>Automatic Reverse Engineering of Human Behavior Based on Text for Knowledge Acquisition</i> ...	2845
Rafal Rzepka, Kenji Araki	
<i>The Internal Structures of Sound-Symbolic Systems: the Universal and Language-Specific Portions of Sound Symbolism</i> .....	2846
Noburo Saji, Kimi Akita, Mutsumi Imai, Katerina Kantartzis, Sotaro Kita	
<i>Does word order influence non-verbal event description by speakers of OS language?</i> .....	2847
Hiromu Sakai, Takuya Kubo, Hajime Ono, Manami Sato, Masatoshi Koizumi	
<i>Embodied Skill to Activate Communication in TV Shows</i> .....	2848
Rui Sakaida, Masaki Suwa	
<i>Collective decision-making processes in online social networks</i> .....	2849
Yasuaki Sakamoto	
<i>Innovating scuba diving education through enhanced immersion and authenticity within the eDiving® environment</i> .....	2850
Ron Salden	
<i>What is promoted by imitation, what promotes imitation: Relation to understanding of other's mental states</i> .....	2851
Wakako Sanefuji, Tomoka Yamamoto, Ikuko Mohri, Masako Taniike	
<i>Conditions that Modulate Perceptual Interference</i> .....	2852
Ava Santos, Lawrence Barsalou, Christy Wilson	
<i>Contribution of the positive emotional sounds to upward vection</i> .....	2853
Kyoshiro Sasaki, Takeharu Seno, Yuki Yamada, Kayo Miura	

<i>Surmising of location with vague embodied agent's instructions</i> .....	2854
Ryo Sato, Yugo Takeuchi	
<i>Looking at nothing facilitates memory retrieval</i> .....	2855
Agnes Scholz, Katja Mehlhorn, Josef Krems	
<i>Does Talking to a Robot in a High-Pitched Voice Strengthen an Attachment?</i> .....	2856
Ryoko Shibata, Takatsugu Kojima, Chie Fukada, Kaori Sato, Yuki Hachikura, Motoyuki Ozeki, Natsuki Oka	
<i>An explanation for status of listening to the late musical works of Morton Feldman by cognitive point of view</i> .....	2857
Takuro Shibayama, Tatsuji Takahashi	
<i>A tool supporting conflation of video and picture for physical expression</i> .....	2858
Satoshi Shibuya, Ken-ichi Kimura	
<i>Narrative Blending and Tense Aspect by Learners of Korean as a Second Language</i> .....	2859
Eunji Shim	
<i>Reassessing the motivation effect of illustrations in text comprehension</i> .....	2860
Hideaki Shimada	
<i>Brain activity during observation of other's action in live and delayed video-mediated social interaction</i> .....	2861
Sotaro Shimada	
<i>Many faces of diagrams: from general properties to practical advantages and disadvantages</i> .....	2862
Atsushi Shimojima	
<i>Learning Grammar via Statistical Mechanism</i> .....	2863
WonJae Shin, Kathleen Marie Eberhard	
<i>Can you see things from your opponent's point of view? – The relationship between critical thinking dispositions and the ability to articulate views different from one's own</i> .....	2864
Noriko Shingaki, Yukie Tsuzuki	
<i>Magic props induce misdirection differently from the magician's face</i> .....	2865
Marie Shoda, Kazuhiko Yokosawa	
<i>Infants form expectations about others' emotions based on context and perceptual access</i> .....	2866
Amy Skerry, Mina Cikara, Susan Carey, Elizabeth Spelke, Rebecca Saxe	
<i>The conflict response of charitable donations on various decoy options</i> .....	2867
Masashi Soma, Itsuki Chiba, Yuichi Hasimoto	
<i>A Neuro-Robotics Model for the Acquisition of Higher Order Concepts in Action and Language</i> .	2868
Francesca Stramandinoli, Davide Marocco, Angelo Cangelosi	
<i>Representational Translation with Concrete and Virtual Models in Organic Chemistry</i> .....	2869
Andrew Stull, Trevor Barrett, Mary Hegarty	
<i>Language and Number Sense: ANS Representations for 'Most'</i> .....	2870
Yasutada Sudo, Hadas Kotek, Michelle Fullwood, Martin Hackl	

<i>A left cerebral hemisphere's superiority in processing spatial-categorical relation in a non-verbal semantic format. ....</i>	2871
Takashi Suegami, Bruno Laeng	
<i>What role do you play in group activity? Objective evaluation through third parties ....</i>	2872
Noriko Suzuki, Tosihiro Kamiya, Ichiro Umata, Sadanori Ito, Shoichiro Iwasawa, Mamiko Sakata, Katsunori Shimohara	
<i>Do we prefer simple realities or simple descriptions? ....</i>	2873
Colleen Szurkowski, David Landy	
<i>Developing and Testing "Visualizing Connection Note," a Constraint-Free Two-Dimensional Concept-Mapping Tool for Ideation ....</i>	2874
Yuisho Takafuji, Hajime Shirouzu	
<i>Cognitive process of the children with reading difficulties: Analysis of the reading patterns with customizable digital reading software. ....</i>	2875
Maiko Takahashi, Mamoru Iwabuchi, Kenryu Nakamura	
<i>Infant biases in lexical acquisition induced by loosely symmetric reasoning ....</i>	2876
Tatsuji Takahashi, Takumi Kamiya, Takahiro Shimizu, Shuji Shinohara	
<i>Color Affects Face Perception in Schematic Faces ....</i>	2877
Fumiyo Takahashi, Yasuhiro Kawabata	
<i>Neural substrates of grammatical information retrieval during sentence comprehension ....</i>	2878
Kei Takahashi, Satoru Yokoyama, Toshimune Kambara, Ryuta Kawashima	
<i>Soccer as Social Interaction between Observable Bodies ....</i>	2879
Katsuya Takanashi, Kazuki Sekine	
<i>Tactile sensation and onomatopoeia in Japanese ....</i>	2880
Yufuko Takashima	
<i>A remotely operated robot as a research tool to study the effects of different roles for successful collaborative learning ....</i>	2881
Nakayama Takayaro, Ashikaga Jun, Inaba Sho, Iyoki Kenta, Naomi Miyake	
<i>The Influence of Cognitive Functions to Acquire Nursing Skills for Patient Transfer ....</i>	2882
Keiko Takeda, Yoriko Watanabe, Taeko Harada	
<i>The effect of 3D stereoscopic display on spatial cognition: a near-infrared spectroscopy study ....</i>	2883
Yoshiyuki Tamamiya, Kazuo Hiraki	
<i>Understanding displacement of communication by graphical communication tasks ....</i>	2884
Kaori Tamura, Takashi Hashimoto	
<i>Effect of embodied cognition in insight problem solving ....</i>	2885
Masahiko Tamura, Kazuhisa Miwa	
<i>For female eyes only: Comparing the effects of enlarging eyes and irises on facial attractiveness ....</i>	2886
Azumi Tanabe-Ishibashi, Mikina Takahashi, Maya Katsuhara, Kana Kuraguchi, Hiroshi Ashida	
<i>Gender difference of social interaction behavior in child's game ....</i>	2887
Daisuke Tanaka, Shinako Terakawa, Ayumi Seki, Hitoshi Uchiyama, Tatsuya Koeda	



<i>Assignment of accent patterns to nonword items in a rapid reading task by Japanese speakers of the Kansai dialect</i> .....	2888
Yuki Tanida, Yoko Higuchi, Yuri Yano, Satoru Saito	
<i>Experience-based modulation of eye-movement behaviour in dynamic and uncertain visual environments</i> .....	2889
Shuichiro Taya, David Windridge, Magda Osman	
<i>Constructing Social Attitudes through Persuasive Writing Training: A Cognitive Approach in Undergraduate Education of Engineering Ethics</i> .....	2890
Emiko Tayanagi	
<i>The nature of the training effects of mental rotation: the limit for transfer to novel orientation</i> .	2891
Haruna Terada, Hiromi Morita	
<i>Multi-platform Experiment to Discuss Behavioral Consistency across Laboratory and Real Situational Studies</i> .....	2892
Hitoshi Terai, Kazuhisa Miwa, Hiroyuki Okuda, Yuichi Tazaki, Tatsuya Suzuki, Kazuaki Kojima, Junya Morita, Akihiro Maehigashi, Kazuya Takeda	
<i>An Exploration of Crossword Skill</i> .....	2893
Kejkaew Thanasuan, Shane Mueller	
<i>Musical thoughts behind composer's writings</i> .....	2894
Akifumi Tokosumi, Akihiro Kawase	
<i>Not just for consumers: Data and theory show that context effects are fundamental to decision-making</i> .....	2895
Jennifer Trueblood, Scott Brown, Andrew Heathcote, Jerome Busemeyer	
<i>A collinear distractor impairs local element search regardless of its probability occurrence</i> .....	2896
Chia-Huei Tseng, Li Jingling, William Oh	
<i>Can Incubation be efficient in Reviewing?</i> .....	2897
Mio Tsubakimoto	
<i>Constraint discovery hint versus constraint relaxation hint in solving insight problems</i> .....	2898
Syoichi Tsujii, Syoji Hamaguchi, Syota Chimura, Hajime Shirouzu	
<i>Heart rate synchronization in collective creative construction tasks</i> .....	2899
Kristian Tylén, Riccardo Fusaroli	
<i>A Longitudinal Study on the Development of Taiwanese Children's Use of Causal and Anaphoric Cue</i> .....	2900
Yuhtsuen Tzeng, Chiung-hsien Tsai	
<i>Measuring Learners' Awareness through Persona-Conjoint Method</i> .....	2901
Hikaru Uchida, Akiko Orita, Masaaki Kunigami, Takao Terano, Atsushi Yoshikawa	
<i>Inhibitory control in event-based prospective memory task: An examination using the retrieval-practice paradigm</i> .....	2902
Kenta Utsumi, Satoru Saito	
<i>Metacognition in children is specific to domain knowledge</i> .....	2903
Vy Vo, Rosa Li, Nate Kornell, Jessica Cantlon	

<i>A study of cognitive style, visual attention distributions and achievement of Web-based multimedia recipes learning</i> .....	2904
Ching-Yeh Wang, Meng-Jung Tsai	
<i>Does self extend to video game avatars? An ERP study</i> .....	2905
Veronica Weser, Ken Livingston	
<i>Social information aids accuracy but hinders adaptation</i> .....	2906
Thomas Wisdom, Keigo Inukai, Wataru Toyokawa, Kameda Tatsuya	
<i>Searching for something familiar or novel: ERP correlates of top-down attentional selection for specific items and categories</i> .....	2907
Rachel Wu, Gaia Scerif, Richard Aslin, Tim Smith, Martin Eimer	
<i>Practicing “Off ice” Collaborative Learning in a University Ice Hockey Team</i> .....	2908
Masayuki Yamada, Masaki Suwa	
<i>Design of motion using mimetic words</i> .....	2909
Kaori Yamada, Toshiharu Taura, Yukari Nagai	
<i>Stochastic dynamics hidden in Japanese martial arts</i> .....	2910
Yuji Yamamoto, Motoki Okumura, Akifumi Kijima, Keiko Yokoyama, Koji Kadota, Hiroo Suzuki, Kazutoshi Gohara	
<i>A Classification of Manner Adverbs in Korean: A Frame-based Approach</i> .....	2911
Myung-Jin Yang	
<i>Brain activities for different cohesion type on discourse comprehension</i> .....	2912
Ken Yasaka, Satoru Yokoyama, Kei Takahashi, Ryuta Kawashima	
<i>Ad hoc creature: Lost and added in translation from description to depiction</i> .....	2913
Sachi Yasuda, Masashi Okamoto, Eiji Aramaki	
<i>The influence of biological cues on the patterns of categorization in non-mental retarded PDD</i> ..	2914
Hsiang-Chun Yeh, Jon-Fan Hu	
<i>Determining people’s expectations about the form of causal relationships</i> .....	2915
Saiwing Yeung, Christopher Lucas, Tom Griffiths	
<i>The role of linguistic inputs on bilingual language development</i> .....	2916
Michael C. W. Yip	
<i>Nothing-absence difference in causal induction and the pARIs rule</i> .....	2917
Junki Yokokawa, Tatsuji Takahashi	
<i>Proficiency in foreign language reading: the relationship between proficiency test score and reading times</i> .....	2918
Satoru Yokoyama	
<i>Toward a history-sensitive description of syntactic development</i> .....	2919
Masato Yoshikawa	
<i>Pictogram Network to Support English Composition Instructors</i> .....	2920
Sayuri Yoshizawa-Watanabe, Masaaki Kunigami, Satoshi Takahashi, Atsushi Yoshikawa, Takao Terano	

<i>Preschoolers Use Timing of Causal Actions as a Cue for Categorization</i> .....	2921
Yue Yu, Tamar Kushnir	
<i>The Rhetoric of Defamiliarization for Narrative Generation using the Constraints in a</i> <i>Conceptual Dictionary</i> .....	2922
Yike Zhang, Junpei Ono, Takashi Ogata	
<b>Author Index</b> .....	2923
<b>Reviewers List</b> .....	2943

## **Building Bridges across Cognitive Sciences around the World**

Welcome to Sapporo and CogSci 2012, the 34th Annual Conference of the Cognitive Science Society. We hope you enjoy the conference and find the program of workshops, tutorials, talks, symposia, and poster presentations that we have assembled to be both fascinating and inspiring.

CogSci 2012 represents a significant milestone in the history of the conference as it is the first to be located outside of Europe and North America. It is the culmination of several years of planning and the clearest example to date of the desire of the Cognitive Science Society to strengthen its connections with similar organizations in the Asia-Pacific region and elsewhere around the world.

As you browse this program, you will see that CogSci has evolved into a truly international conference with a total of 37 countries being represented. Of the 798 accepted talks and posters, approximately 40% are from East Asia, 37% come from the Americas, and 20% originate from Europe, with the rest coming from Australia, North Africa and other regions of Asia. We hope you will be impressed by the high quality of the research presented and the diverse range of questions being addressed.

In addition to this strong collection of talks and posters, we are delighted to have three world leaders in their fields as plenary speakers – Gerd Gigerenzer, Nancy Nersessian, and Lawrence Barsalou –, not to mention a broad set of symposia exploring the breadth of cognitive science.

As chairs, we organized two symposia to allow distinguished researchers an opportunity to discuss two important contemporary issues in cognitive science. The first takes the thirtieth anniversary of David Marr's landmark posthumous book, *Vision*, to address the question whether his tripartite formulation of levels of analysis is still relevant in the age of reductionist neuroscience and Bayesian analysis. Our second symposium reflects on and explores historical origins and recent developments in robotics research (an area of considerable activity and expertise in Japan) that seek new ways of understanding cognition via the mechanisms and processes of embodiment and emotion.

The Governing Board of the Cognitive Science Society has also organized a symposium to bring together leading thinkers from Asia, Europe and the US to discuss real world implementations of educational innovations based on cognitive and learning science principles and research.

This conference is the result of the hard work of many people. Firstly, we would like to thank the members of the Organizing Committee, the Program Committee, the prize judges, and all the many reviewers for their time and effort. Secondly, we would like to thank the Society's Business Manager, Deborah Gruber, for her work in administering prizes, visas, etc., James Stewart of Precision Conference Solutions for his rapid responses to our questions, and those at Scarritt Group for their organization of the venue and local arrangements. Thirdly, we would in

particular like to thank the Society's Conference Officer, Andy Stull, for his advice and help in coordinating the whole conference, getting the program together, and keeping us on schedule.

We would also like to thank members of The International Association of Cognitive Science (Asia-Pacific) and Japan Cognitive Science Society for their many contributions to this conference and to their progressive drive to strengthen ties between cognitive scientists throughout the Asia-Pacific region and beyond.

CogSci is renowned for the high quality and diversity of the research presented as well as for being a crucial annual opportunity to meet like-minded individuals from across the globe. As you attend the conference we hope you agree with us that this year is no exception. We hope also that the connections you make here bear fruit through new productive collaborations so that the conference can truly achieve its aim of building bridges across cognitive sciences around the world.

Naomi Miyake, David Peebles, Richard P. Cooper

Co-chairs, CogSci 2012

# Organizing Committee CogSci 2012

<b>Program Chairs:</b>	Naomi Miyake, University of Tokyo David Peebles, University of Huddersfield Richard P. Cooper, Birkbeck, University of London
<b>Sponsors Chairs:</b>	Hajime Shirouzu, Chukyo University Christopher T. Kello, University of California, Merced Holger Schultheis, University of Bremen
<b>Awards Chairs:</b>	Michael Pauen, Humboldt University Natalie Sebanz, Central European University Markus Knauff, University of Giessen Ipke Wachsmuth, University of Bielefeld
<b>Member Abstracts Chairs:</b>	Natalie Sebanz, Central European University Sachiko Kiyokawa, Chubu University
<b>Publicity Chair:</b>	Mitchell J. Nathan, University of Wisconsin-Madison
<b>Student Volunteer Chairs:</b>	Christopher Myers, Air Force Research Laboratory Toshihiko Matsuka, Chiba University
<b>Tutorials &amp; Workshops Chair:</b>	Glenn Gunzelmann, Air Force Research Laboratory
<b>Symposia Chair:</b>	Kai-Florian Richter, University of Melbourne
<b>Publication-based Talks Chair:</b>	Kai-Florian Richter, University of Melbourne
<b>Web Chair:</b>	Dongkyu Choi, University of Kansas

## **CogSci2012 Program Committee**

We thank the following people for their generous contribution of time and effort to CogSci 2012.

Altmann, Erik	Fu, Wai-Tat	Myers, Christopher
Aslin, Richard	Fum, Danilo	Navarro, Daniel
Ball, Jerry	Garrod, Simon	Neth, Hansjoerg
Barkowsky, Thomas	Gentner, Dedre	Noelle, David
Barley, Mike	Giudice, Nicholas	Nokes-Malach, Timothy
Beaman, Philip	Goel, Ashok	Oaksford, Mike
Bertel, Sven	Goldstone, Rob	Olivetti, Marta
Best, Brad	Gonnerman, Laura	Pani, John
Billman, Dorrit	Gonzalez, Cleotilde	Papafragou, Anna
Blessing, Stephen	Gunzelmann, Glenn	Peebles, David
Bonnefon, Jean-Francois	Hahn, Ulrike	Pleskac, Timothy
Brighton, Henry	Hampton, James	Ragni, Marco
Brook, Andrew	Hegarty, Mary	Reed, Stephen
Brumby, Duncan	Heit, Evan	Richter, Kai-Florian
Burns, Bruce	Helie, Sebastien	van Rijn, Hedderik
Busemeyer, Jerome	Jacobson, Michael	Scheutz, Matthias
Byrne, Ruth	Jones, Gary	Schmalhofer, Franz
Cacciari, Cristina	Kaschak, Michael	Schmid, Ute
Cangelosi, Angelo	Katz, Irvin	Schoelles, Mike
Casasanto, Daniel	Kemp, Charles	Schultheis, Holger
Cassimatis, Nicholas	Kennedy, William	Sebanz, Natalie
Clancey, Bill	Kintsch, Walter	Shah, Priti
Clement, Catherine	Kiyokawa, Sachiko	Shirouzu, Hajime
Cooper, Rick	Klenk, Matthew	Shultz, Thomas
Cottrell, Gary	Klippel, Alexander	Sloutsky, Vladimir
Cox, Anna	Knauff, Markus	Stracuzzi, David
Dale, Rick	Koedinger, Ken	Sun, Ron
Danks, David	Kokinov, Boicho	Tenenbaum, Josh
DMello, Sidney	Landy, David	Trafton, Greg
Douglass, Scott	Langley, Patrick	Treur, Jan
Estes, Zachary	Love, Brad	Tversky, Barbara
Feeney, Aidan	Luger, George	Upal, M. Afzal
Ferstl, Evelyn	Magnani, Lorenzo	Verguts, Tom
Forbus, Ken	Markman, Art	Waldmann, Michael
Frank, Mike	Matessa, Michael	Wood, Sharon
French, Bob	Miller, Craig	Youmans, Robert
Freudenthal, Daniel	Monaghan, Padraic	Young, Richard

# **CogSci 2012 Sponsors**

*We sincerely thank the sponsors of the 34<sup>th</sup> Annual Meeting of the Cognitive Science Society for their support of the conference awards and the tutorials, and for supporting student participation through reduced registration fees and coverage of travel costs.*

***Air Force Office of  
Scientific Research  
(AFOSR)***

***Asian Office of Aerospace  
Research and Development  
(AOARD)***

***Hakuhodo  
Innovation  
Lab***

***Institute of Education  
Sciences  
(IES)***

***Interdisciplinary  
Transregional  
Collaborative Research  
Center***

***National Science  
Foundation  
(NSF)***

***The Robert J. Glushko  
and Pamela Samuelson  
Foundation***

***Sapporo  
Convention  
Bureau***

***Wiley-Blackwell***

***Japanese Cognitive  
Science Society***



# CogSci 2012 Awards

## Marr Prize

The Marr Prize, named in honor of the late David Marr, is awarded to the best student paper at the conference. All student first authors were eligible for the Marr Prize for the best student paper. The Marr Prize includes an honorarium of \$1,000 and is sponsored by The Cognitive Science Society. The winner of the 2012 Marr Prize for Best Student Paper is:

**George Kachergis, Chen Yu, and Richard M. Shiffrin:** *Actively learning nouns across ambiguous situations* (Friday, 13:00, Track 1)

## Computational Modeling Prizes

Four prizes worth \$1,000 each are awarded for the best full paper submissions to CogSci 2012 that involve computational cognitive modeling. The four prizes represent the best modeling work in the areas of perception/action, language, higher-level cognition, and applied cognition. These prizes are all sponsored by The Cognitive Science Society. The winners of the 2012 Computational Modeling Prizes are:

### Applied Cognition

**Yugo Hayashi:** *The effect of “Maverick”: A study of group dynamics on breakthrough in collaborative problem solving* (Saturday, 14:10, Track 3)

### Perception/Action

**Kevin A. Smith & Edward Vul:** *Sources of uncertainty in intuitive physics* (Friday, 12:40, Track 5)

### Language

**Noah D. Goodman & Andreas Stuhlmüller:** *Knowledge and implicature: Modeling language understanding as social cognition* (Friday, 15:20, Track 1)

### Higher-Level Cognition

**Doug Markant & Todd M. Gureckis:** *Does the utility of information influence sampling behavior?* (Saturday, 12:20, Track 4)

## Cognition and Student Learning (CaSL) Prize

The Cognition and Student Learning (CaSL) Prize is an honorarium of \$1,000 that is awarded to the best paper on research conducted on a topic directly related to cognitive science, educational practice, and subject matter learning. This prize is sponsored by the Institute of Education Sciences (IES). The winner of the 2012 Cognition and Student Learning Prize is:

**Azadeh Jamalian & Barbara Tversky:** *Gestures alter thinking about time* (Thursday, 14:10, Track 4)

## **Student Travel Awards**

Travel awards have been provided to students whose papers were accepted as oral presentations with the highest reviewer rankings, and who indicated a need for travel funding. The Robert J. Glushko and Pamela Samuelson Foundation generously sponsored \$10,000 for student travel awards for these papers. The 2012 Travel Awards went to:

Ricky Chan (Department of English, The University of Hong Kong, Hong Kong)

Gregory Cox (Department of Psychological and Brain Sciences, Indiana University)

Sarah Dolscheid (Max Planck Institute for Psycholinguistics, Nijmegen)

Annie Gagliardi (Department of Linguistics, University of Maryland)

Linn Gralla (Department of Linguistics and Literary Sciences, Universität Bremen)

George Kachergis (Department of Psychological and Brain Sciences, Indiana University)

Jorie Koster-Hale (Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology)

Yakov Kronrod (Department of Linguistics, University of Maryland)

Andrew Lovett (EECS Department, Northwestern University, Illinois)

Doug Markant (Department of Psychology, New York University)

Long Ouyang (Department of Psychology, Stanford University)

Michael Pacer (Department of Psychology, University of California, Berkeley)

Patrick Plummer (Department of Psychology, University of California, San Diego)

Anna Rafferty (Computer Science Division, University of California, Berkeley)

Felix G. Rebitschek (Department of Psychology, University of Greifswald)

Kevin Smith (Department of Psychology, University of California, San Diego)

Sergiu Tcaci Popescu (Laboratoire Psychologie de la Perception, CNRS & Université Paris Descartes)

Tomoki Tsuchida (Department of Computer Science and Engineering, University of California, San Diego)

Joseph Jay Williams (Department of Psychology, University of California, Berkeley)

Ewelina Wnuk (Max Planck Institute for Psycholinguistics, Nijmegen)

Daniel Yurovsky (Department of Psychological and Brain Sciences, Indiana University)

## **Awards Committee**

Markus Knauff (co-chair), Michael Pauen (co-chair), Natalie Sebanz (co-chair), Ipke Wachsmuth (co-chair), Jennifer Wiley (CaSL award coordinator), Felice Bedford, Gary S. Dell, Morton Ann Gernsbacher, Ulrike Hahn, Bernhard Hommel, Boicho Kokinov, Stefan Kopp, Klaus Oberauer, Colleen Seifert, Leon Urbas, David Uttal, Eldad Yechiam.

## **Robert J. Glushko Dissertation Prizes**

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

### **Robert J. Glushko Dissertation Prize Recipients**

The 2012 recipients of the Robert J. Glushko Prizes for Outstanding Doctoral Dissertations / theses in Cognitive Science are:

**Dr. Timothy F. Brady** - 2011 PhD thesis "Structured Representations in Visual Working Memory"

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

**Dr. Jennifer L. Culbertson** - 2010 PhD thesis "Learning Biases, Regularization, and the Emergence of Typological Universals in Syntax"

Department of Cognitive Science, Johns Hopkins University

**Dr. Nazbanou Nozari** - 2011 PhD thesis "Is Comprehension Necessary for Error Detection? A Conflict-based Account of Monitoring in Speech Production"

Department of Psychology, University of Illinois at Urbana-Champaign

**Dr. Steven T. Piantadosi** - 2011 PhD thesis "Learning and the language of thought"

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

**Dr. Rachel Wu** - 2011 PhD thesis "Learning (to Learn) from Spatial Attention Cues During Infancy"

Birkbeck, University of London

## **NSF Funded Joint Conference Grant & Research Fellowships**

In association with the Cognitive Science Society, the US National Science Foundation has funded eight conference grants/research fellowships to US citizens who are enrolled as students at a US institution. The funds support students both to attend the 34<sup>th</sup> Annual Conference of the Cognitive Science Society (CogSci2012) and to participate in a collaborative research project with a sponsoring institution in Japan. The awardees are:

Deanne Adams, David Braithwaite, Heather Burte, Seth Frey, Drew Hendrickson, Laura Morett, Cybelle Smith and Richard Veale.

## Invited Plenary Presentations

### Rumelhart Prize Lecture

Reinforcement Learning in the Mind and Brain: Cinderella at the Cognitive Science Ball

*Peter Dayan*

*Thursday August 2<sup>nd</sup>, 16:30*

### Keynote Talks

Simple Heuristics that Make Us Smart

*Gerd Gigerenzer*

*Thursday August 2<sup>nd</sup>, 09:00*

Building Scientific Cognition: Conceptual Innovation on the Frontiers of Science

*Nancy J. Nersessian*

*Friday August 3<sup>rd</sup>, 09:00*

Situated Conceptualization

*Lawrence W. Barsalou*

*Saturday August 4<sup>th</sup>, 09:00*

## **Women in Cognitive Science sponsored interactive panel discussion: Professional advancement, leadership and international collaboration**

**Laurie Beth Feldman** ([lfeldman@albany.edu](mailto:lfeldman@albany.edu))

Department of Psychology; SS 369  
The University at Albany, SUNY  
Albany, NY 12222, USA

**Judith Kroll** ([jfk7@psu.edu](mailto:jfk7@psu.edu))

Moore Building  
Department of Psychology  
The Pennsylvania State University  
University Park, PA 16802 USA

**Janet van Hell** ([jgv3@psu.edu](mailto:jgv3@psu.edu))

Moore Building  
Department of Psychology  
The Pennsylvania State University  
University Park, PA 16802 USA

**Suparna Rajaram** ([srajaram@notes.cc.sunysb.edu](mailto:srajaram@notes.cc.sunysb.edu))

Department of Psychology  
Stony Brook University  
Stony Brook, NY 11794-2500 USA

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

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

The goal of a small "Connections" conference sponsored by the US National Science Foundation in Japan and the Japan Science and Technology Agency of Japan in 2010 was to establish networking connections between researchers and to strengthen international partnerships for collaboration. American and Japanese participants represented a variety of STEM disciplines. The primary goal of that meeting was to foster the research agendas of the participants. A secondary goal was to "help develop future leaders in science and engineering by encouraging discussion on the institutional environment and culture that are conducive to nurturing women STEM leaders."

The **Interactive Panel Discussion: Professional advancement, leadership and international collaboration** seeks to build on this momentum by bringing together American and Japanese researchers in Cognitive Science at

the 2012 meeting of the Cognitive Science Society in Sapporo, Japan. Speakers include Sanae Ariga (Hokkaido U), Mutsumi Imai (Keio U), Noriko Hoshino (Kobe U Foreign Studies), Naomi Miyake (U Tokyo), Hanako Yoshida (Houston), Laurie Feldman (U Albany & Haskins Labs) and Betty Tuller (NSF). These scientists represent American and Japanese junior and senior researchers, university administrators and program officers/directors of NSF.

Speakers will discuss cross-cultural solutions to foster research productivity and visibility for women scientists. One major theme will be leadership, both how to identify and assume positions that help to develop leadership skills for professional advancement. A related theme is how those experiences do and do not differ across cultures. A second major theme will be how to develop new research collaborations outside of one's primary institution, including international collaborations. All acknowledge that this is not a simple process and often evolves slowly, out of more social networking connections. While such solutions generally occur on an ad hoc basis and vary across individuals, the aim of the WICS workshop is to enable discussion of possible solutions so as to enhance the productivity and visibility for women scientists in cognitive science.

Social Hour and refreshments to follow.

### **Acknowledgments**

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

Partial funding for this event comes from NSF Award BCS-1049764 and is organized in conjunction with the Tokyo office of the NSF Office of International Science and

Engineering. Funds were also contributed by the Office of Support for Female Researchers at Hokkaido University.

# **Full-day Workshop proposal: Teleoperated Android as a Tool for Cognitive Studies, Communication and Art**

**Shuichi NISHIO (nishio@ieee.org)**

Hiroshi Ishiguro Laboratory, Advanced Telecommunications Research Institute International (ATR)  
2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan

**Hiroshi ISHIGURO (ishiguro@sys.es.osaka-u.ac.jp)**

Department of Systems Innovation, Graduate School of Engineering Science, Osaka University  
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

## **Theme and Goals**

Following the successful workshops in 2005 and 2006 on *Android Science*, the aim of this full-day workshop is to introduce and discuss on current insights and future usage of teleoperated androids.

Teleoperated androids, robots owning humanlike appearance equipped with semi-autonomous teleoperation facility, was first introduced to the world in 2007 with the public release of Geminoid HI-1. *Geminoid* is a teleoperated android robot that resembles existing human being. While androids were designed for studying human nature in general, *geminoids* was made to study individual aspects as presence or personality traits, tracing their origins and implementation into robots. Both its appearance that resembles the source person and its teleoperation functionality serves in making Geminoid as a research tool. After the release of Geminoid HI-1, several types of teleoperated androids have been produced: Geminoid F, Geminoid DK, Telenoid R1/R2 and Elfoid P1. While the Geminoids are after real existing persons, Telenoid and Elfoid are attempts to represent human beings in their minimalistic forms; a challenge to see to what extent elements that forms us can be omitted but still able to transfer presence of the teleoperating person.

Since their birth, Geminoids and Telenoids have been used in a variety of domains throughout the world, from studies in various fields such as in cognitive psychology / neuroscience, social psychiatry, developmental psychology, robotics, and human-machine interface to philosophy and art. One example is the *android drama* which showed new possibilities on not only on usage for teleoperated android robots but for artistic representations as well as seeking purity in the natures of human beings.

The past workshops that concentrated on autonomous humanlike robots and androids laid a foundation for android science research, a field that integrates the synthetic approach from robotics with the empirical methodologies of the social sciences. Participants, coming from engineering and the social, cognitive, and biological sciences sought fundamental principles underlying cognition and communication between individuals.

In this workshop, we will focus on the further enhanced and broadened usage of teleoperated androids that can provide new means for cognitive science studies, and can bridge

the gap between cognitive neuroscience and the behavioral sciences, as well as philosophy, social science and arts, leading to a new way of understanding human beings.

## **Topics**

- Using teleoperated androids as a controllable mankind for psychological experiments
- The role of affect and motivation in social development or communication
- Empathic relationships among people and/or robots
- How people become adapted to teleoperated androids
- The evolution, development, and nature of agency, intentionality, or social intelligence
- Models of personal, interindividual, group, or cultural norms
- Cross-modal synchronization or stabilized plasticity in speech and/or gesture
- Teleoperated androids in the society
- Androids working alongside people as peers
- Applications in human environments
- Ethical issues concerning teleoperated androids
- Perception of naturalness, attractiveness, or charisma of teleoperated androids
- Minimal elements required to show human likeness
- The relationship between appearance and perceived behavior
- The Total Turing Test
- Teleoperated androids as communication device
- Elderly care with teleoperated androids
- Using teleoperated androids for artistic expression

## **Importance and Relevance for the conference**

This workshop focuses on and discusses about cross-disciplinary studies under the current usage and future possibilities of teleoperated android robot which can provide new means for cognitive science studies, and can bridge the gap between cognitive neuroscience and the behavioral sciences, as well as philosophy, social science and arts, leading to a new way of understanding human beings. Thus, this will provide opportunities for conference participants to see latest

advances in this area as well as to discuss and find research seeds with researchers of different disciplines.

The organizers has been involved in making and conducting studies on teleoperated androids. Prof. Hiroshi Ishiguro is the inventor of both the notion of android science as well as various teleoperated androids, Geminoid, Telenoid and Elfoid. Dr. Shuichi Nishio has been with Prof. Ishiguro in constructing teleoperated android systems and have conducted various laboratory / field studies with them up to now.

### **Audience**

- Robotics engineers and computer scientists with an interest in cognitive psychology, robotics, human-robot interaction, as well as artificial intelligence, machine learning, pattern recognition and control theory.
- Psychologists and sociologists who are concerned and/or interested with embodied communication or social development
- cognitive scientists who are concerned with the relationship between brain processes and social dynamics; social and comparative biologists;
- Philosophers who are interested in human nature issues such as mind/body separation and interaction;
- Artists and dramatists who are interested in new possibilities of art on human nature;

The workshop is of interest to the target participants because teleoperated androids can work as a test tool for social and cognitive theories. Research in this domain depends on interdisciplinary collaboration between engineers and natural and social scientists.

### **Possible presenters**

- Christian Becker-Asano (University of Freiburg, Germany)
- Thierry Chaminade (University College of London, UK)
- Kazuo Hiraki (Tokyo University, Japan)
- Hiroshi Ishiguro (Osaka University, Japan)
- Shoji Itakura (Kyoto University, Japan)
- Karl MacDorman (Indiana University, US)
- Takashi Minato (ATR, Japan)
- Hideyuki Nakanishi (Osaka University, Japan)
- Shuichi Nishio (ATR, Japan)
- Hideaki Ogawa (Ars Electronica, Austria)
- Kohei Ogawa (ATR, Japan)
- Hirata Oriza (Theater company Seinendan)
- Ayse Saygin (University of California, US)
- Henrik Scharfe (Aalborg University, Denmark)
- Hidenobu Sumioka (ATR, Japan)

### **Estimate of the number of participants**

Thirty participants including organizers and presenters.

### **Publication**

The publication of the workshop will be done in four ways.

1. To our research collaborators:  
We have research collaboration with many laboratory in the world, especially in Japan, Europe and US. We will ask these collaborators to submit papers and to participate in the workshop.
2. Via grant agency:  
This workshop theme, teleoperated android, is now studied under several grants in Japan and Europe. We will ask the grant agency to promote the workshop.
3. To the cognitive science and robotics society:  
This will be done via mailing lists and web pages.
4. To the press people: Our laboratory is accepting more than 50 requests for interviews and shooting from various press in the world. We will advertise the workshop to the press who visited us in the past.

After the workshop has been accepted, we will ask several journals for a special issue so that the fine presentations will appear gathered in much details. Also, we are planning to publish a book that collects the findings and activities related with teleoperated android robots.

### **Special requirements**

If the space allows, we would like to bring our robots and run them throughout the workshop so that people who cannot attend the conference may be able to teleoperate the androids from remote and pseudo-join the conference. In this way, people can see the actual teleoperated androids in use and participants can discuss the real effects of using the robots in the workshop.

This will require: power supplies (100V), a separate room not far away for placing air compressor (because this is noisy) and an Internet connection.

### **Contact**

Dr. Shuichi NISHIO

Affiliation: Hiroshi Ishiguro Laboratory,  
Advanced Telecommunications Research Institute International (ATR)  
Address: 2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan  
Telephone: +81-774-95-1560  
Fax: +81-774-95-1508  
e-mail: nishio@ieee.org

Prof. Hiroshi ISHIGURO

Affiliation: Department of Systems Innovation, Graduate School of Engineering Science, Osaka University



# Workshop on Modeling the Perception of Intentions

**David Pautler (pautlerd@ihpc.a-star.edu.sg)**

Programme in Computational Social Cognition, Institute of High Performance Computing  
1 Fusionopolis Way, 16-16 Connexis  
Singapore 138632

**Keywords:** intention recognition; action understanding; theory of mind; event perception; social cognition; computational modeling; probabilistic inference.

exploring a complex cognitive activity from different disciplinary perspectives.

## Abstract

Attributing intentions to others based on observations of their behavior is a core cognitive ability. It is also a necessary precursor to social judgments such as judgments about responsibility and morality. The seminal work of Heider and Simmel (1944) highlighted the spontaneity, richness, and range of intention attributions that can be elicited by a stimulus as impoverished as moving geometric figures. Subsequent research has revealed a wide range of visuospatial cues that suggest specific intentions as well as observer attributes that influence judgments. How are such cues and observer attributes integrated into an inferred specific intention? A handful of processing models have used frameworks such as schema-matching or probabilistic inference to integrate such cues. This workshop will address two questions: 1) How have different fundamental paradigms fared in the quest for a model of human intention perception? and 2) What questions about this topic are most in need of answers?

## Goals

- To foster on-going communication, and perhaps coordinated research, across disciplines among researchers on this topic.
- To provide an overview for the audience of how different disciplines have approached this topic.

## Why a workshop?

Perception of intentions has been studied by cognitive, social, and developmental psychologists, philosophers, anthropologists, artificial intelligence researchers, and computer vision researchers. Apart from its intrinsic interest, there is the prospect of different approaches informing each other. Because there are several lines of research to cover, and because we want to encourage speakers not just to cover their own lab's work but propose ways of linking with others, a symposium or half-day workshop would not afford enough time.

## How it is relevant?

The topic has drawn much interest across disciplines, and the confirmed speakers represent many of those disciplines. The event could provide a useful example the value of

## The Organizer

While trained as a computer scientist myself, I have led a project on this topic for two years, collaborating closely with cognitive and social psychologists. Together we have completed a literature review spanning the disciplines listed above, designed and built a computational simulation furthering the schema-matching line of research, and published that work in a journal article last year:

Pautler, D., Koenig, B.L., Quek, B.K., Ortony, A. (2011). [Using modified incremental chart parsing to ascribe intentions](#). *Behavior Research Methods* 43(3), 643-665, DOI: 10.3758/s13428-011-0128-2.

## Target Audience

The workshop will be most relevant to those working in high-level perception, theory of mind, and plan recognition, but we expect it to be of general interest to many researchers in cognitive, developmental, and social psychology, as well as artificial intelligence, anthropology, and philosophy. It is hard to predict the number of early arrivals at the first CogSci conference to be held in Asia, but we expect approximately 30 people in the audience.

## Confirmed Speakers

[Barbara Tversky](#)

Department of Psychology Building 420  
Stanford University  
Stanford, CA 94305-2130 USA  
[btversky@stanford.edu](mailto:btversky@stanford.edu)  
+1 650 814 7922  
Fax +1 650-725-5699

[Shimon Ullman](#)

Department of Computer Science And Applied Mathematics  
Ziskind Building, Room 208  
Weizmann Institute of Science  
Rehovot 76100 Israel  
[shimon.ullman@weizmann.ac.il](mailto:shimon.ullman@weizmann.ac.il)  
+972-8-934-2894  
Fax: +972-8-934-2945 / 6023

[Dare Baldwin](#)

Department of Psychology  
Straub Hall  
1227 University of Oregon  
Eugene OR 97403 USA  
[baldwin@uoregon.edu](mailto:baldwin@uoregon.edu)  
+1 541 346-4964

[Frank E. Pollick](#)

Room 702  
Dept of Psychology  
58 Hillhead Street  
Glasgow G12 8QB United Kingdom  
[Frank.Pollick@glasgow.ac.uk](mailto:Frank.Pollick@glasgow.ac.uk)  
+44 (0)141 330 3945

[Josh Tenenbaum](#)

Building 46-4015, 77 Massachusetts Avenue  
Cambridge, MA 02139 USA  
[jbt@mit.edu](mailto:jbt@mit.edu)  
+1 617 452-2010  
Fax +1 617 253-8335

[Tao Gao](#)

Building 46-4053  
Cambridge, MA 02139 USA  
[taogao@mit.edu](mailto:taogao@mit.edu)  
+1 617 324-2895

[Peter Pantelis](#)

Psychology Building  
Busch Campus  
152 Frelinghuysen Road  
Piscataway, NJ 08854-8020 USA  
[peter.pantelis@gmail.com](mailto:peter.pantelis@gmail.com)  
+1 848 445-2576  
Fax: +1 732 445-2263

[David Pautler](#)

Programme in Computational Social Cognition  
Institute of High Performance Computing  
1 Fusionopolis Way  
16-16 Connexis 138632 Singapore  
[pautlerd@ihpc.a-star.edu.sg](mailto:pautlerd@ihpc.a-star.edu.sg)  
+65 6419 1304  
Fax +65 6463 0200

## Archiving and Special Requirements

If videotaping will be available at the conference site, we would like to use the workshop budget of US\$1200 to cover that cost. Videos might be hosted by the organizer's institutional website, <http://cogsys.ihpc.a-star.edu.sg/>.

## Publicity

The most relevant mailing lists we know of are: [Plan Rec Psychonomics](#), and [ACT-R](#). All of those listservs are commonly used for announcements of such events. Furthermore, [CASA](#) (conference on Computer Animation and Social Agents) will be held here in Singapore in May, and we will advertise there.

# And Now for Something Completely Different: Python in Cognitive Science

Mark Andrews ([m.andrews@ntu.ac.uk](mailto:m.andrews@ntu.ac.uk))

Jesse Diaz ([jesse.diaz@ntu.ac.uk](mailto:jesse.diaz@ntu.ac.uk))

Division of Psychology,  
Nottingham Trent University  
Nottingham  
NG1 4B,UK

**Keywords:** Python; Programming; Scientific Computing; Numerical Computing; Computational Modelling; Experimental Design; Stimuli Presentation Software; Data Analysis;

## Objectives and Scope

The objective of this tutorial is to introduce and motivate the use of the Python programming language in cognitive science research. Within the last 10 years, the development of scientific and numerical libraries in Python has grown to the point where Python can now be used as a scientific and numerical computing environment comparable to products like Matlab and Mathematica. As of yet, however, it appears that knowledge of the potential applications of Python to research in cognitive science is still rather limited. The aim of this tutorial, therefore, is to describe these areas of application and to advocate the advantages and appeals of using Python as the principal programming language in cognitive science research. Given the generality of the tools being discussed, it is hoped that this tutorial will have widespread appeal and relevance.

## Outline of Tutorial

The tutorial will be divided into three main parts. The first part introduces the Python language generally. The second introduces numerical and scientific programming in Python. The third part introduces how to develop computer-based psychology and psychophysics experiments using Python.

The tutorial will involve both classroom style lectures with slides and workshop style computer-based worked examples and exercises. The audience are encouraged to bring their own laptop, and all necessary software will be provided in advance.

## General Introduction

In order to introduce Python, we will begin by describing the fundamentals of the Python language. We will also demonstrate how to start an interactive Python session using the ipython environment. The audience will be encouraged to follow the examples themselves using their own computers.

As part of this introduction, we will also compare Python to its alternatives, paying particular attention to comparison with Matlab. This comparison is inevitable, given that Matlab has traditionally been the principal scientific computing tool in cognitive science. Notable points of similarity between Python and Matlab are that both offer an interactive array-processing and visualization environment using high-level dynamic programming languages. Both are designed

for rapid prototyping and development. Both allow for seamless extension using external modules written in compiled languages like C/C++ and Fortran. Notable advantages of Python, however, include that it is a general-purpose language whose application goes far beyond numerical array processing. Python is one of the top five programming language currently in use throughout the world. Python is a remarkably well-designed object-oriented language whose standard library is large and comprehensive. Finally, Python is non-commercial open-source software distributed according to an unrestricted software license. Likewise, its large set of third-party extension modules and libraries are, almost without exception, also distributed using unrestricted or public open-source software licenses.

## Numerical and Scientific Python

The basic Python language as introduced in the previous section lacks n-dimensional numerical arrays and the ability to easily plot and visualize data. These capabilities, in addition to a large number of more special-purpose scientific libraries are provided by the Scipy/Numpy suite of modules. These libraries are seamlessly integrated with ipython to create a rich interactive array-processing and visualization environment, comparable in functionality to Matlab and Mathematica.

We will begin this section by describing ipython's capabilities more extensively than done in the previous section. These include: Interactive high-performance parallel computing for clusters and multicore architectures, an online interactive Notebook comparable to that used in Mathematica, sql-based searchable command histories, in-line graphics, and symbolic mathematics with T<sub>E</sub>X-based output.

Having established how to use ipython, the audience will be encouraged to follow the examples as we discuss the the following topics:

Arrays: General n-dimensional arrays and their operations (e.g. element-wise function application, summing, slicing, indexing, searching) are provided by numpy.

2d visualization: Plotting and visualization, especially of 2d data, are provided by matplotlib, amongst others.

3d visualization: Complex 3d graphics are provided by mayavi.

Parallel computing: Interactive high-performance and parallel programming is a built-in functionality of python.

Integration with C/C++ and Fortran: Interfaces to programs written in compiled languages like C/C++ or Fortran are pro-

vided through the use of interface generators like `swig` and `f2py`.

## Computer-based Experiments

Computer-based cognitive psychology and psychophysics experiments are now almost ubiquitous in cognitive science. While these tasks have been traditionally handled by GUI-based programs like *e-prime* and *superlab*, these programs do not allow for the flexibility and control that is often demanded by researchers. While high-level languages like Matlab are being used as an alternatives to GUI-based programs, Matlab's special-purpose nature is not well suited to the non-numerical programming necessary for experimental stimuli presentation and recording. By contrast, due to the generality of its language, its extensive of widget toolkits (e.g. `wxpython`, `pyGTK`, `pyQt`), and video-game libraries (`pyGame`, `pyglet`), Python allows for considerable flexibility and sophistication in the design experiment software.

Currently, there are at least 4 Python-based stimulus-presentation programs: `Psychopy`, `open-sesame`, `vision-egg`, and `pyepl`. This final section will describe each in brief, but concentrate primarily on `psychopy`.

The aim of this section will be to discuss the principles and functionality of `psychopy` and then to work through examples of simple experiments (e.g. the stroop task, the lexical-decision task). `Psychopy`'s basic object-oriented stimuli and events will be described in order to understand its extensibility. We will, however, also make extensive use of its *builder* interface that can allow from rapid development of code templates. Finally, we will discuss how to interface `psychopy` and Python generally with external devices such as serial response boxes that allow for precise timing of responses.

## The Presenters

The main presenter for this tutorial will be Mark Andrews. Mark Andrews is a Lecturer (Assistant Professor in North American Terminology) in the Division of Psychology, Nottingham Trent University, and has a research affiliate position in the Division of Psychology and Language Sciences, University College London. His teaching primarily involves advanced statistics and experimentation methods. In this capacity, for the past two years, he has taught programming using R and Python to undergraduate and postgraduate students, with student evaluations being overwhelming costive. He has been a Python user for over 10 years, and has extensive experience with all the topics that will be covered in this tutorial. He also is very familiar with the Cognitive Science community, having presented at past conferences often, and being awarded the Computational Language Modelling prize in 2009. Jesse Diaz is a research assistant in Nottingham Trent University, with extensive experience with general Python programming and especially with the use of Python in psychology experiments, both using tools like `psychopy` and by using Python web-application frameworks for online experiments.

## Materials

The use of Python in science is backed by a vibrant community of developers and advocates. We have been in direct contact with principal individuals in this community and they have generously offered their support, both by providing their presentation slides and other teaching materials and by providing their general advice on how to promote Python in settings such as the Cognitive Science tutorials. For example, we have been contact with Dr. Fernando Perez who is a research scientist in neuroscience at UC Berkeley. Dr. Perez is the creator and principal developer of `ipython`. He has kindly offered the extensive teaching materials on the `ipython` computing environment that are at his disposal. Likewise, we have been in contact with Dr. Jonathan Peirce who is an Associate Professor in Psychology in the University of Nottingham. Dr. Peirce is the creator and principal developer of `psychopy`, and has had extensive experience both teaching `psychopy` to students and promoting its use in psychology and cognitive neuroscience. As a result, we have a considerable body of relevant teaching materials to draw upon. Examples are available at sites like following, and elsewhere:

<http://scipy-lectures.github.com>

<http://ipython.org/presentation.html>

# Full Day Tutorial on Quantum Models of Cognition and Decision

**Jerome R. Busemeyer (jbusemey@indiana.edu)**

Cognitive Science, Indiana University, 1101 E. 10th Street,  
Bloomington, IN 47405 USA

**Peter Bruza (p.bruza@qut.edu.au)**

Faculty of Science and Engineering, Queensland University of Technology,  
Brisbane, QLD 4001 Australia

**Taiki Takahashi (taikitakahashi@gmail.com)**

Department of Behavioral Science, Hokkaido University,  
Hokkaido 060-0810 Supporo, Japan

**Jennifer S. Trueblood (jstruebl@indiana.edu)**

Cognitive Science, Indiana University, 1101 E. 10th Street,  
Bloomington, IN 47405 USA

**Keywords:** classical information processing; quantum information processing; logic and mathematical foundation; Bayesian probability, quantum probability; Markov and quantum processes; quantum entanglement; quantum game theory; conceptual combinations; decision making, memory.

causal reasoning, decision making, conceptual combinations, memory recognition, and associative memory. This tutorial is needed to introduce and train cognitive scientists on this promising new theoretical approach to cognitive science.

## General Purpose

This *full day* tutorial is an exposition of a rapidly growing new alternative approach to building computational models of cognition and decision based on quantum theory. The cognitive revolution that occurred in the 1960's was based on classical computational logic, and the connectionist/neural network movements of the 1970's were based on classical dynamical systems. These classical assumptions remain at the heart of both cognitive architecture and neural network theories, and they are so commonly and widely applied that we take them for granted and presume them to be obviously true. What are these critical but hidden assumptions upon which all traditional theories rely? Quantum theory provides a fundamentally different approach to logic, reasoning, probabilistic inference, and dynamical systems. For example, quantum logic does not follow the distributive axiom of Boolean logic; quantum probabilities do not obey the disjunctive axiom of Kolmogorov probability; quantum reasoning does not obey the principle of monotonic reasoning. It turns out that humans do not obey these restrictions either, which is why we consider a quantum approach. This tutorial will provide an exposition of the basic assumptions of classic versus quantum information processing theories. These basic assumptions will be examined, side by side, in a parallel and elementary manner. The logic and mathematical foundation of classic and quantum theory will be laid out in a simple and elementary manner that uncovers the mysteries of both theories. Our main point will be to show that quantum theory provides a unified and powerful explanation for a wide variety of paradoxes found in human cognition and decision ranging across findings from attitudes, inference,

## Presenters

Jerome Busemeyer is a professor of Cognitive Science at Indiana University. He was editor of the *Journal of Mathematical Psychology* and he is now *Associate Editor of Psychological Review*. His research interests include decision making and dynamic modeling. Peter Bruza is a professor of information science and he is a pioneer in the field of quantum interaction (QI). He also serves on the editorial boards of *Information Retrieval*, *Journal of Applied Logic*, *The Logic Journal of the IGPL*. Jerome and Peter are co-authors of a new book "*Quantum models of cognition and decision*" Cambridge University Press, 2012. Taiki Takahashi is a professor at Hokkaido University working in the field of neuroeconomics, but also with expertise in quantum decision theory, and he has published articles on this topic in *Physical Letters A*. Jennifer Trueblood is a PhD student at Indiana University with several publications on the topic of quantum cognition including one in *Cognitive Science*.

## Previous Tutorials and Symposia

This tutorial was presented for a full day at the Cognitive Science meetings in Nashville, 2007, Washington DC 2008, and Amsterdam, 2009, which included around 30 people each time. The ratings obtained from participants after the tutorial were all very good. Also this tutorial follows a symposium on quantum cognition presented at the Cognitive Science meeting 2011, and these papers will appear as a special issue in *Topics in Cognitive Science*. A similar tutorial was presented at the 3<sup>rd</sup> and 4<sup>th</sup> Annual Meetings on Quantum Interaction held

at Saarbruecken, Germany, 2009, and Aberdeen Scotland, 2010 with about 40 participants.

### Participant Background

This tutorial will introduce participants to an entirely new area and no previous experience or background with quantum theory will be assumed. **No background in physics is required.** In fact, except for a few simple examples to motivate the idea, little or no reference to physics will be made during main part of the tutorial. What is required is an elementary background in classic logic and probability. During the tutorial, we will review basic concepts of linear algebra needed for quantum theory. (e.g., vectors, projectors, unitary transformations).

### Material to be Covered

1. The first topic will examine the major differences between classic versus quantum theories of probability. The concept of superposition is introduced and distinguished from classic probability mixtures. The important issue of measurement in classical and quantum systems will be compared and examined. The key to this section will be several dramatic empirical examples illustrating empirical violations of the classic laws of probability (e.g., conjunction, disjunction, total probability) and the parsimonious explanation of all these violations by quantum theory. (1 hr).
2. Next we examine the differences between classical and quantum dynamical systems. The basic idea of a Markov processes will be introduced and compared with quantum processes. (Cognitive architectures and many neural networks can be represented as Markov processes). A parallel development of Markov and quantum processes will be shown. The concept of a state will be distinguished for Markov and quantum systems. The effects of measurement on the state of the system are compared for Markov and quantum systems. A key feature of this section is to show when and how quantum processes depart from Markov processes. (1 hr)
3. The third part will present the details of Matlab and R programs used to compute the choice probability and response time predictions of a dynamic quantum model that has been developed to explain three ongoing research programs in cognitive and decision making: violations of the “sure thing principle” of rational decision theory, violations of dynamic consistency in decisions, and interference of categorization on decisions. (45 min)
4. The fourth part will introduce quantum computing and information processing ideas. The concepts of a bit and a qubit will be contrasted. The concept of a conjunction of properties used in classic information processing theory will be related to the concept of a tensor product space used in quantum theory. The controlled U-gate will be introduced and compared with a production rule. The linear transformation of states used by quantum theories will be related to the distributed representation and content addressable properties of connectionist/neural networks.

The concept of fuzzy representation and probabilistic representation will be discussed and compared for fuzzy set, Bayesian, and quantum theories. The idea of an entangled state will be described. Bell’s inequality will be introduced, and violations found in conceptual combinations are reviewed. The dramatic implications of violations of this inequality for classical theories will be discussed. (45 min)

5. This part will present the details of Matlab programs used to perform quantum computing for some complex information processing tasks. This includes pattern recognition and planning event dependent action sequences under uncertainty. Basic tools of quantum computing will be used including Kronecker products to perform U-gate operations, and partial traces for measurement of components of a complex system. (45 min).
6. This part will detail how quantum theory is being used to model the human mental lexicon. In particular, quantum entanglement will be described as a means of modeling cognitive phenomena in non-reductionist way, e.g., conceptual combinations. A key feature of this section is to introduce formal tools and experimental methods which can determine whether cognitive phenomena can be validly modeled in a decompositional way. (1.0 hr)
7. Review and future directions (30 min).

### References

<http://mypage.iu.edu/~jbusemey/quantum/QuantumCognitionNotes.htm>

Busemeyer, J. R., & Bruza, P. D. (in press). *Quantum models of cognition and decision*. NY: Cambridge University Press.

Bruza, P., Kitto, K., Nelson, D., & McEvoy, C. (2009). Is there something quantum-like in the human mental lexicon? *Journal of Mathematical Psychology*, 53 (5), 362-377.

Busemeyer, J. R., Pothos, E. & Franco, R., Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment ‘errors.’ *Psychological Review*, 118, 193-218.

Cheon, T. & Takahashi, T. (2010) Interference and inequality in quantum decision theory. *Physical Review Letters*, A 375 100.

Pothos, E. M., & Busemeyer, J. R. (2009). A quantum probability explanation for violations of “rational” decision theory. *Proceedings of the Royal Society B*. 276 (1665), 2171-2178

Trueblood, J. S. & Busemeyer, J. R. (2011). A quantum probability explanation for order effects on inference. *Cognitive Science*, 35, 1518-1552.

### Acknowledgments

This tutorial and related research is supported by U.S. National Science Foundation through grants SES-0817965 and SES-0818277 and the Australian Research Council grants DP0773341 and DP1094974.

# Proposal for a tutorial on Using Bayes to Interpret Non-significant Results

**Zoltan Dienes (dienes@sussex.ac.uk)**

School of Psychology, University of Sussex, Brighton, BN1 9QH

**Keywords:** Bayesian inference; Bayes Factors; confidence intervals; likelihood intervals; non-significant results; evidence.

## 1. Outline the objectives and scope of the tutorial.

The purpose of the tutorial is to present simple tools for dealing with non-significant results, an area which cognitive scientists have consistently found problematic. In particular, people will be taught how to apply Bayes Factors and likelihood intervals to draw meaningful inferences from non-significant data, using free easy-to-use on-line software: Software which allows one to determine whether there is strong evidence for the null and against one's theory, or if the data are just insensitive, a distinction  $p$ -values cannot make. These tools have greater flexibility than power calculations and allow null results to be interpreted over a wider range of situations. Such tools should allow the publication of null results to become easier.

The online software for Bayes Factors (with instructions) is here:

[http://www.lifesci.sussex.ac.uk/home/Zoltan\\_Dienes/inference/Bayes.htm](http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm)

And the online software for likelihood intervals here:

[http://www.lifesci.sussex.ac.uk/home/Zoltan\\_Dienes/inference/Likelihood.htm](http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Likelihood.htm)

## 2. Explain how the tutorial will be delivered giving a detailed description of the material that will be covered.

The tutorial will consist of lectures by me; after the second hour people will be able to use their internetted laptops to work through examples on the software, and thereby interactively follow the points I make, and also explore the tools for themselves.

Schedule:

9:30 – 10:30 Basics: The different aims of significance testing and Bayesian inference (including the three moral and inferential paradoxes of significance testing and their solution)

10:45 – 11:45 Bayes Factors

12- 1 Examples with Bayes Factors, to illustrate appropriate and inappropriate use, and robustness checks (Bayesian analysis can of course be misused in ways we will clarify)

2-3 Confidence intervals, likelihood intervals, credibility intervals with examples, including the (little discussed) four principles for using intervals inferentially in theory testing

3:30 – 4:30 Examples showing the complementary strengths and weaknesses of Bayes factors and interval methods for interpreting null results

4:45 – 5:30 Discussion of e.g. any particular data people wish to bring, and free questions

The tutorial will emphasize how statistics, both Bayes factors and interval estimates, can be brought into more intimate contact with theory than has typically been the case, and appropriate ways of doing this. (Interpreting null results requires making contact with theory.)

My emphasis will be practical rather than ideological, though conceptual arguments will be important.

## 3. Justify why it is important to have a tutorial in the proposed area at the conference.

Users of statistics have been criticised for decades for their interpretation of non-significant results. Users have either used null results to count against a theory that predicted a difference (without establishing that the results actually counted against the theory) or ignored the results as uninformative (without establishing that they were). One only need pick up any recent issue of almost any journal to see this. (I don't exclude many of my own papers from this criticism!) In that sense the topic has been important to clarify for a long time. Recently, however there was been a resurgence of interest in Bayesian and likelihood methods, and the recent developments are particularly useful for users of statistics. Little can be more important than that we as a community draw appropriate inferences from data, and get the most from our data. The issues are applicable to the whole community of cognitive scientists, and hence appropriate for a meeting of the Cognitive Science Society. Several strategies for dealing with null results will be taught, as well as reasons why the most common strategies, orthodox as well as Bayesian, can be problematic.

## 4. Specify how relevant the topic is for the conference (i.e., does it focus on an emerging or cross-disciplinary research topic?)

Bayes has been making a resurgence for getting on 10 years now in cognitive science. Part of the interest has been in Bayes as a model for how the mind works. While the workshop has nothing to say on what the best theory of the mind is, theories of how we should analyse data are clearly relevant to theories of how the mind works. More importantly, part of the recent interest in Bayes is precisely on the topic of the workshop – principles and methods for drawing statistical inferences. Indeed, at the meeting of the Society last year Kruschke held a very successful workshop on Bayesian inference based on his book. I will be teaching a slightly different philosophy and different methods (but which complement Kruschke's approach). Kruschke



covered Markov Chain Monte Carlo methods and hierarchical modelling; the workshop will not cover these topics. This workshop offers different tools to the researcher, simple tools for dealing with a t-test (i.e. 1-df contrasts – which is all we are normally really interested in), tools which a researcher could directly use straight after walking out of the workshop (without learning R, BUGS, or anything else). I will make most use of the notion of strength of evidence rather than posterior probabilities. In terms of data, the Bayesian analyses taught just require the sort of summary statistics SPSS or other packages produce. I believe that a majority of people will leave transformed in how they conceive of non-significant results, however they then choose to deal with them.

### **5. State why you are well suited to organize a workshop in the proposed area.**

Dienes (2008) is an introduction to orthodox, Bayesian and likelihood inference which has an associated website with free online software. Dienes (2011) discusses the arguments for Bayes, and also provides practical advice for using Bayes. I have been teaching students to use Bayes at the University of Sussex since 2005 on the undergraduate course Philosophy of Psychology, and the masters course Philosophy of Science, thereby coaching hundreds of students on applying Bayes to over a hundred different papers of their choice. This experience has helped me both pedagogically and in seeing how to apply Bayes in a practical way. I have now submitted (and had reviewed) more than half a dozen standard research papers with Bayesian analyses in them (using the same software that I will be teaching). The Bayesian analyses have not been queried, so my arguments for their use and interpretation seem unproblematic to the community so far! Four of the papers are now published (see

[http://www.lifesci.sussex.ac.uk/home/Zoltan\\_Dienes/inference/Bayes.htm](http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm)

and scroll to bottom for examples of published papers using Bayes as a tool). I have also lectured on using Bayes in China, Norway, Greece and around the UK.

In January I ran the proposed workshop as a one-day national workshop for the Economic and Social Research Council (ESRC) National Centre for Research Methods. Some feedback:

“Thank you again for an extremely informative day that was very well delivered. Sussex students must be very pleased about having such a clear and articulate statistics tutor,” from a UK Professor of Psychology and Research Director of Department. “I’d like to thank you for a very enjoyable and stimulating workshop last Tuesday. Your web page is also extremely helpful,” from a lecturer in Genetics. “I came to the workshop not sure about how useful it would be or how easy to understand – Zoltan made it really interesting and clear with examples. I will definitely use this in my research,” from a psychology postgraduate.

### **6. Identify the likely audience for the tutorial.**

**Specifically, state whether the tutorial will introduce participants to an area, or whether it will cover an advanced topic for participants who already have knowledge in a particular area.**

The audience is anyone who uses statistical inference – i.e. just about everybody attending the Meeting could be interested. I will assume the audience is familiar with a t-test; I will not assume more detailed knowledge. But those with more extensive knowledge will also appreciate the material (I have lectured on the material to undergraduates as well as to statisticians; it has been well received in all contexts).

### **8. Specify any special requirements for the tutorial - particularly, any specialist equipment or software required by participants..**

A laptop per participant, or one laptop between two. Ideally the laptops should be connected to the net.

### **9. Provide full contact details: name of contact person, affiliation, address (including post code/zip and country), telephone, fax, e-mail, names and affiliation of additional author(s).**

Zoltan Dienes, School of Psychology, University of Sussex, Brighton, BN1 9QH, UK, (tel) 44 1273 877335, (fax) 1273 678058, [dienes@sussex.ac.uk](mailto:dienes@sussex.ac.uk)

### **References**

- Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Palgrave Macmillan  
 Website: [http://www.lifesci.sussex.ac.uk/home/Zoltan\\_Dienes/inference/index.htm](http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/index.htm)  
 Dienes, Z. (2011). Bayesian versus Orthodox statistics: Which side are you on? *Perspectives on Psychological Sciences*, 6(3), 274-290.



# Nengo and the Neural Engineering Framework: From Spikes to Cognition

Chris Eliasmith ([celiasmith@uwaterloo.ca](mailto:celiasmith@uwaterloo.ca))

Terrence C. Stewart ([tcstewar@uwaterloo.ca](mailto:tcstewar@uwaterloo.ca))

Center for Theoretical Neuroscience, University of Waterloo  
200 University Ave West, Waterloo, ON, N2L 3G1, Canada

**Keywords:** cognitive modeling; neural engineering; representation; decision making; working memory; cognitive architecture; cognitive control

## Tutorial Objectives

As we learn more about the neural activity underlying cognitive function, there is an increasing demand to explicitly and quantitatively connect cognitive theories to neurological details. Bridging these levels provides benefits in both directions; aspects of the cognitive theory can predict and be constrained by neurological details, and the neurological details can in turn identify important modifications to the overall cognitive theory.

This tutorial introduces the Neural Engineering Framework (NEF; Eliasmith and Anderson, 2003) and the associated open-source toolkit Nengo (<http://nengo.ca>), which offer a general method for implementing high-level cognitive theories using biologically realistic spiking neurons. This approach takes a high-level description of a cognitive theory (in terms of information being represented and transformed) and combines it with relevant anatomical and neurophysiological constraints, producing a detailed mechanistic model of how interacting neurons can efficiently produce the desired behaviour. The resulting models can be run to produce predictions of spike patterns, firing rates, fMRI time-courses, accuracy, reaction times, and overall behaviour. Complete details can be found in the book *How to Build a Brain* (Eliasmith, 2012; to be released by OUP at CogSci 2012).

These methods have been made more accessible by the construction of the software package Nengo, which provides a graphical interface suitable for network construction. This tutorial introduces the NEF theory explaining how high-level function can be systematically related to single cell activity, and provides extensive hands-on experience building these neural models using Nengo. Our central objective is to allow participants to leave the tutorial with a method for constructing cognitive models with spiking neurons, and experience using that method in an intuitive software environment.

## Tutorial Structure

This full-day tutorial combines the theoretical bases of the Neural Engineering Framework with hands-on examples of practically applying these concepts using Nengo. For example, the presentation of the theory for how a scalar value can be represented by the spiking pattern in a group of neurons is paired with a tutorial on using Nengo to generate such a neural group and simulate its behavior over time.

Participants are expected to bring a laptop to follow along with these tutorials (Windows, OS X, and Linux are all supported, and software is provided).

In particular, the tutorial covers using the NEF to represent scalars and vectors, perform linear and nonlinear transformations on these values, and store information over time. These are the basic mechanisms required for a wide range of algorithms, and form the basis for our models of sensorimotor systems, working memory, and cognitive control. This provides participants with basic building blocks for constructing novel neural implementations of a wide variety of cognitive models.

To supplement this, we more closely examine how cognitive theories can be expressed in terms of vectors and transformations. The basic approach of employing *semantic pointers* (vectors that combine the benefits of semantic similarity measures with the compositionality of symbol structures) is described. We show how this method provides a unified approach to many types of cognitive models, including perceptual, symbolic reasoning, and motor control models. For example, we show how to construct a non-classical symbol system, capable of performing the operations required for symbolic cognition. The result is a scalable and efficient neural cognitive architecture, constructed from the basic approaches described in the first half of the tutorial.

Finally, we explore recent results in building whole-brain models using the NEF. This involves a fully integrative model spanning vision, object recognition, working memory, cognitive control, and motor control to produce a neural cognitive architecture. This ~3 million neuron model is built in Nengo, uses images for input, draws digits using a 2-joint arm as its output, and is performs a variety of tasks, including list memory, mental addition, inductive reasoning over symbols, and reinforcement learning. The tutorial covers this model and its behavioural and neurobiological constraints, including the dopaminergic learning system.

Variants of this tutorial were presented at ICCM 2009, CogSci 2010, Telluride 2011, and CogSci 2011. An on-line tutorial is available at <http://nengo.ca>, and significant changes have been made in terms of scaling Nengo models up to larger neuron counts and more complex behaviour.

## Tutorial Justification

The Neural Engineering Framework provides a method to bridge the gap between cognitive and neural theories. Its earlier applications have been to sensory and motor systems, including the barn owl auditory system, rodent navigation, swimming control in zebrafish, and the vestibular ocular

reflex in monkeys. However, these same principles are now being applied to cognitive models. This includes models of serial-order recall (Choo & Eliasmith, 2010), action selection in the basal ganglia (Stewart, Choo, & Eliasmith, 2010), visual working memory (Singh & Eliasmith, 2006), deep belief networks for visual recognition (Tang & Eliasmith, 2010), the Wason card task (Eliasmith, 2005), the Tower of Hanoi task (Stewart & Eliasmith, 2011), and a model of inductive rule generation that received the computational modelling prize in higher-level cognition at CogSci 2010 (Rasmussen & Eliasmith, 2010).

While we find that the Neural Engineering Framework produces extensive new insights into the neural grounding of cognitive function, we also find that the underlying mathematics and a lack of familiarity with biologically realistic neural modeling have been a significant barrier to entry for new researchers. As a result, we feel that a full-day tutorial is most appropriate for introducing the necessary concepts from control theory, signal theory, and theoretical neuroscience. Feedback from previous tutorials has been extremely positive, with participants now using Nengo for their own research and in the classroom at the University of Manchester, Rensselaer, Yale, Franklin & Marshall College, and Stanford.

The NEF provides an exciting new tool for cognitive science, as it provides a technique for producing direct neural predictions from cognitive theory. Furthermore, it leads to important theoretical results as to the relationships between neural properties and the high-level algorithms they are capable of implementing (e.g. the relationship between neurotransmitter re-uptake rate and the 50ms cognitive cycle time; Stewart, Choo, & Eliasmith, 2010).

These consequences are also very general, as the NEF provides techniques that can be applied to a wide variety of cognitive theories. It provides a structure for organizing a high-level description such that it can be implemented by realistic spiking neurons, providing meaningful data in terms of the expected spike patterns, time course, and behavioural accuracy. We have made use of it in a variety of contexts, and have developed tools that support the creation and analysis of these models. Tutorial participants will gain hands-on experience with a tool that helps generate new models and can be applied to existing models. In both cases, these tools will help participants incorporate ever-more-abundant neural data into their research.

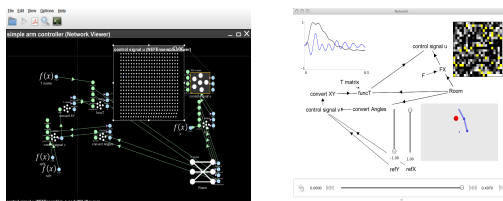


Figure 1: The Nengo interface. Network construction (left) is done either by point-and-click or by Python scripting. Visualization (right) provides on-the-fly control of inputs with plots of spiking activity, decoded representations, etc.

## Audience

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

## Presenters

Chris Eliasmith holds a Canada Research Chair in Theoretical Neuroscience, and is director of the Centre for Theoretical Neuroscience at the University of Waterloo. He has over 50 publications spanning neuroscience, psychology, philosophy, computer science, and engineering, on topics including working memory, mental representation, population coding, neural dynamics, computation, automatic text classification, and cognitive architectures. His recent book, *How to Build a Brain*, and his earlier book, *Neural Engineering*, form the basis for this tutorial.

Terry Stewart is a postdoc in the Centre for Theoretical Neuroscience, and has developed large-scale models with the NEF, including the Tower of Hanoi task, focussing on problems of cognitive control.

## References

- Choo, F., & Eliasmith, C. (2010). A Spiking Neuron Model of Serial-Order Recall. *32<sup>nd</sup> Annual Conference of the Cognitive Science Society*.
- Eliasmith, C. (2005). Cognition with neurons: A large-scale, biologically realistic model of the Wason task. *27<sup>th</sup> Annual Meeting of the Cognitive Science Society*.
- Eliasmith, C. (2012). *How to build a brain: A neural architecture for biological cognition*. New York, NY: Oxford University Press.
- Eliasmith, C., & Anderson, C. (2003). *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. Cambridge: MIT Press.
- Rasmussen, D., & Eliasmith, C. (2010). A neural model of rule generation in inductive reasoning. *32<sup>nd</sup> Annual Conference of the Cognitive Science Society*.
- Singh, R., & Eliasmith, C. (2006). Higher-dimensional neurons explain the tuning and dynamics of working memory cells. *Journal of Neuroscience*, 26, 3667-3678.
- Stewart, T.C., & Eliasmith, C. (2010). Neural symbolic decision making: A scalable and realistic foundation for cognitive architectures. *1<sup>st</sup> Annual Meeting of the Biologically Inspired Cognitive Architectures Society*.
- Stewart, T.C., & Eliasmith, C. (2011). Neural Cognitive Modelling: A Biologically Constrained Spiking Neuron Model of the Tower of Hanoi Task. *33<sup>rd</sup> Annual Conference of the Cognitive Science Society*.
- Stewart, T.C., Choo, X., & Eliasmith, C. (2010). Dynamic Behaviour of a Spiking Model of Action Selection in the Basal Ganglia. *10<sup>th</sup> Int. Conf. on Cognitive Modeling*.
- Tang, Y., Eliasmith, C. (2010). Deep networks for robust visual recognition. *International Conference on Machine Learning*.

# Probability, programs, and the mind: Building structured Bayesian models of cognition

Noah D. Goodman (ngoodman@stanford.edu)

Department of Psychology,  
Stanford University, Stanford CA 94305 USA

Joshua B. Tenenbaum (jbt@mit.edu)

Department of Brain and Cognitive Sciences,  
Massachusetts Institute of Technology, Cambridge MA 02139 USA

## Objectives and scope

Human thought is remarkably flexible: we can think about infinitely many different situations despite uncertainty and novelty. Probabilistic models of cognition (Chater, Tenenbaum, & Yuille, 2006) have been successful at explaining a wide variety of learning and reasoning under uncertainty. They have borrowed tools from statistics and machine learning to explain phenomena from perception (Yuille & Kersten, 2006) to language (Chater & Manning, 2006). Traditional symbolic models (e.g. Newell, Shaw, & Simon, 1958; Anderson & Lebiere, 1998), by contrast, excel at explaining the productivity of thought, which follows from compositionality of symbolic representations. Indeed, there has been a gradual move toward more structured probabilistic models (Tenenbaum, Kemp, Griffiths, & Goodman, 2011) that incorporate aspects of symbolic methods into probabilistic modeling. Unfortunately this movement has resulted in a complex “zoo” of Bayesian models. We have recently introduced the idea that using programs, and particularly *probabilistic programs*, as the representational substrate for probabilistic modeling tames this unruly zoo, fully unifies probabilistic with symbolic approaches, and opens new possibilities in cognitive modeling. The goal of this tutorial is to introduce probabilistic models of cognition from the point of view of probabilistic programming, both as a unifying idea for cognitive modeling and as a practical tool.

The probabilistic programming language Church (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008), mathematically grounded on the stochastic  $\lambda$ -calculus, provides a universal language for representing probabilistic models. We will use Church to introduce key ideas and examples of probabilistic modeling. A Church program represents a probabilistic model, and hence inferences that can be drawn from this model, without committing to a process level implementation of inference. This will allow us to focus the tutorial on structured representations and probabilistic inference phenomena without worrying about the details of inference algorithms (such as Markov chain Monte Carlo) that tutorials on Bayesian modeling often become bogged down in. On the other hand, because there are existing inference tools for Church (e.g. Wingate, Stuhlmüller, & Goodman, 2011), students will get hands-on experience with performing inference over different probabilistic models.

The tutorial will include several in-depth case studies where the probabilistic programming viewpoint is particularly useful. These include intuitive theories, such as naive physics and theory of mind, and models of inductive learning that exhibit learning-to-learn and structured abstraction.

## Tutorial format

This full-day tutorial aims to introduce students to key ideas of, and new tools for constructing, structured probabilistic models. We will assume only basic familiarity with probability and with programming (i.e. minimal mathematical or statistical background). The tutorial will thus be appropriate for a general Cognitive Science audience, as well for practitioners of bayesian modeling who want to learn about probabilistic programming.

We will teach this tutorial drawing on a combination of infrastructure and materials that we’ve have used to teach graduate-level classes at Stanford and MIT (and which has been used by others at UCSD and University College Dublin). In particular, students will use the on-line ChurchServ interface to Church, in order to explore these tools without the need to install special software. This interface has been integrated into a Wiki document on Probabilistic Models of Cognition ([http://projects.csail.mit.edu/church/wiki/Probabilistic\\_Models\\_of\\_Cognition](http://projects.csail.mit.edu/church/wiki/Probabilistic_Models_of_Cognition)) that we will use for portions of the tutorial.

In addition, we will create new examples focussed on aspects of the approach that we expect to be both new and interesting to a Cognitive Science audience. These include models of physics and vision, based on forward-simulation with standard graphics and vision simulators, and models of language understanding that predict detailed, quantitative human data.

We will use the morning session to introduce the ideas of probabilistic modeling and the Church language, to illustrate basic ideas (such as explaining away, and hierarchical models), and to provide hands-on exercises using Church to create models. The afternoon session will be devoted to case studies of more sophisticated applications of these ideas to cognition, including studies from vision, language, and reasoning.

We, the instructors, have extensive experience in probabilistic modeling of cognition and extensive experience teaching courses and tutorials on these techniques. In addition we are active at the forefront of developing probabilistic programming languages, both conceptually and as practical

tools. Both of the instructors have extensive experience teaching tutorials on probabilistic models of cognition specifically from the viewpoint of Church, including courses to graduate students, high-school students, linguists, and psychologists.

Tutorials on Bayesian Models of Inductive Learning have been taught at the Annual Conference of the Cognitive Science Society in 2006, 2008, and 2010 (all co-taught by JBT). This tutorial will be similar in covering the ideas of recent work in Bayesian modeling, but will do so from a different viewpoint and will introduce a different skill set (Church and probabilistic programming). We have presented similar tutorials at the European Summer School For Logic Language and Information 2010 (NDG), the Institute for Pure and Applied Mathematics (NDG and JBT), and several other venues. We will adjust the tutorial based on feedback from those experience as well as the particular audience we expect at Cognitive Science.

## References

- Anderson, J., & Lebiere, C. (1998). *The atomic components of thought*. Lawrence Erlbaum.
- Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *TRENDS in Cognitive Sciences*, 10, 335–344.
- Chater, N., Tenenbaum, J. B., & Yuille, A. (2006, July). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7), 287–291.
- Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: a language for generative models. *Uncertainty in Artificial Intelligence*.
- Newell, A., Shaw, J., & Simon, H. (1958). Elements of a theory of human problem solving. *Psychological Review*, 65(3), 151.
- Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022), 1279.
- Wingate, D., Stuhlmüller, A., & Goodman, N. (2011). Lightweight implementations of probabilistic programming languages via transformational compilation. In *Proceedings of the 14th international conference on artificial intelligence and statistics* (p. 131).
- Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10, 301–308.

# Using Machine Learning for Exploratory Data Analysis

**Joshua M. Lewis**

josh@cogsci.ucsd.edu

Department of Cognitive Science  
University of California, San Diego

**Virginia R. de Sa**

desa@cogsci.ucsd.edu

Department of Cognitive Science  
University of California, San Diego

## Abstract

This tutorial will introduce attendees to fundamental concepts in the clustering and dimensionality reduction fields of unsupervised machine learning. Attendees will learn about the assumptions algorithms make and how those assumptions can cause the algorithms to be more or less suited to particular datasets. Hands-on interaction with machine learning algorithms on real and synthetic data are a central component of this tutorial. Students will use the software platform Divvy (freely available from the Mac App Store or [divvy.ucsd.edu](http://divvy.ucsd.edu)) to visualize and analyze data in real time while testing the concepts learned during formal instruction. We encourage attendees to bring their Mac laptops and their own datasets for the hands-on portion of the tutorial, and if possible to email their datasets ahead of time to [josh@cogsci.ucsd.edu](mailto:josh@cogsci.ucsd.edu).

Attendees will leave the tutorial with a much better understanding of basic concepts in unsupervised machine learning. Pragmatically they will understand when to apply, e.g., k-means to a dataset versus single linkage clustering. Attendees will also learn how to integrate Divvy into their existing research workflow so that they can quickly test and compare machine learning algorithms on their data.

## Objectives and Scope

This tutorial will introduce attendees to fundamental concepts in the clustering and dimensionality reduction fields of unsupervised machine learning. Attendees will learn about the assumptions algorithms make and how those assumptions can cause the algorithms to be more or less suited to particular datasets. Hands-on interaction with machine learning algorithms on real and synthetic data are a central component of this tutorial. Students will use the software platform Divvy to visualize and analyze data in real time while testing the concepts learned during formal instruction. We will encourage attendees to bring their own datasets for analysis in the hands-on portion of the tutorial.

Attendees will leave the tutorial with a much better understanding of basic concepts in unsupervised machine learning. Pragmatically they will understand when to apply, e.g., k-means to a dataset versus single linkage clustering. Attendees will also learn how to integrate Divvy into their existing research workflow so that they can quickly test and compare machine learning algorithms on their data.

## Topics

We will split the tutorial into two sections, a morning section focused on clustering and an afternoon section focused on dimensionality reduction. Both sections will start with a brief (approximately 1.5 hours) formal introduction to mathematical and conceptual underpinnings of the topic, followed by a hands-on lab session applying the concepts learned directly

before. The lab sessions will start with synthetic datasets designed to reinforce conceptual lessons, and then move to real datasets provided by ourselves and the attendees.

The clustering section will cover centroid-based methods (such as k-means), hierarchical methods (such as single linkage), spectral clustering, and probabilistic modeling (such as Gaussian mixture models). The dimensionality reduction section will cover linear methods (such as PCA and projection pursuit) and that nonlinear methods (such as Isomap, tSNE, and Kernel PCA).

Though we will provide formal mathematical characterizations, our focus will be on conceptual differences between techniques, specifically related to choosing the correct technique based on known structure in a dataset. Additionally, we will emphasize that there is no single best clustering or embedding for any given dataset (in other words, there is no universally agreed upon objective function for clustering and dimensionality reduction). One's own analysis goals can play a significant role in, e.g., determining the number of clusters to search for. Finally, on the topic of evaluation we will cover the visualization and interpretation of algorithmic output as well as formal quality measures such as Silhouette for clusterings and Trustworthiness for embeddings.

We will transact our lab sections in Divvy, a free and open-source software platform for performing unsupervised machine learning (see <http://divvy.ucsd.edu> where there is a video of Divvy in action). Divvy will allow attendees to rapidly cluster, reduce and visualize a wide variety of datasets without having to write any code. Divvy can concurrently visualize several perspectives on a dataset and can switch between datasets with one click, even when algorithms are computing in the background. Divvy integrates well with existing research workflows—it can import data from Matlab and R and it exports data and visualizations in standard formats for further analysis.

## Qualifications

Joshua Lewis recently completed his PhD thesis, *Anthropocentric Data Analysis*, on the topic of reintegrating humans into the data analysis process. He is the lead software architect behind Divvy, and has done several studies on the relationship between human reasoning and machine learning. He is a postdoc in UCSD's Natural Computation Lab under the supervision of Virginia de Sa. He has attended CogSci and presented papers every year starting in 2009. Joshua will lead the tutorial.

Virginia de Sa is an associate professor at UCSD in the Cognitive Science department. She has done extensive re-

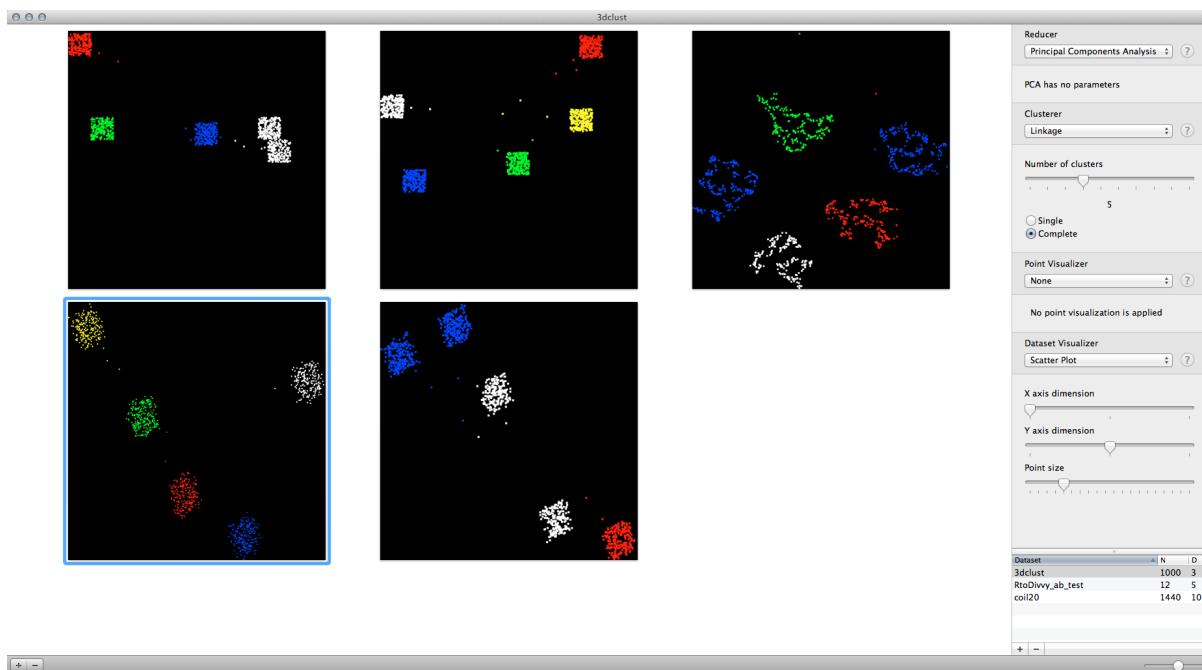


Figure 1: The full Divvy UI. Each visualization represents a different view of the same dataset (generated by combining a dimensionality reduction technique, a clustering technique and a dataset visualizer) and users can set the properties of each view using the tools to the right. A list of datasets resides in the bottom right, allowing the user to switch between them at any time, even while results are computing in the background.

search in the fields of machine learning and human perception and has ten years of experience in teaching undergraduate and graduate courses in data analysis and machine learning to people with both weak and strong mathematical backgrounds. She is the PI on NSF Grant #SES-0963071, which funds Divvy's development. Virginia will assist in developing the tutorial curriculum.

For detailed CVs, please see our websites (listed in the Contact Us section below).

### Relevance to CogSci

Data analysis is a fundamental part of most scientific endeavors, and the judicious application of machine learning techniques to the analysis process is often quite profitable. Further, in the field of Cognitive Science in particular, a basic understanding of machine learning techniques is valuable for interpreting the work done in computer science-focused subfields such as artificial intelligence and computational neuroscience. Clustering and dimensionality reduction are established methodologies for performing data analysis, and this tutorial presents them from the unique perspective of enabling the human researcher to use them wisely.

### Audience

This tutorial will introduce attendees to the area of unsupervised machine learning for data analysis. It does not presuppose any machine learning background and thus will be

appropriate for any graduate student or faculty member interested in integrating machine learning techniques into their research process. On the other hand, it will be less valuable to those researchers who already have extensive experience applying machine learning techniques.

### Attendee Requirements

Divvy a Mac OS X 10.6/10.7 application, and the hands-on portions of the lab will require a Mac laptop with a 64-bit Intel processor (basically any Mac made in the last three years). Attendees will be able to easily download and install the software at the conference or ahead of time. Those who do not have or do not wish to bring a laptop will be grouped with those who do. Hopefully enough attendees can bring laptops to have groups of about 2 to 4 people per analysis team. Last year's conference was suffused with glowing white Apple logos, so we don't think this will be an onerous requirement. Additionally, attendees can submit datasets to us ahead of time that we can integrate into the instruction as examples of real-life data analysis problems.

### Contact Us

Joshua M. Lewis - [josh@cogsci.ucsd.edu](mailto:josh@cogsci.ucsd.edu) - <http://cogsci.ucsd.edu/~josh> - UCSD Cognitive Science - 115 Dufour St, Santa Cruz, CA 95060 - (831)-246-1578

Virginia de Sa - [desa@cogsci.ucsd.edu](mailto:desa@cogsci.ucsd.edu) - <http://cogsci.ucsd.edu/~desa> - UCSD Cognitive Science - 9500 Gilman Dr, La Jolla, CA 92093 - (858)-822-5095

# Practical Advice on How to Run Human Behavioral Studies

**Frank E. Ritter (frank.ritter@psu.edu)**

++1 (814) 865-4453

College of IST, Penn State

University Park, PA 16802 USA

**Keywords:** Psychology experimental method; HCI usability studies.

## (I) Objectives and scope of the tutorial

The lack of materials on the details of running human experiments can lead to a gap between theory and practice, which is particularly acute in cognitive science done outside of psychology departments. Consequently, labs frequently must not only impart these practical skills to students informally but also must address misunderstandings arising from this divorce of theory and practice in their formal education. Researchers in psychology often end up appalled by the lack of this common but undocumented sense when behavioral research is reported by researchers outside of psychology. This tutorial provides practical advice on how to run studies for beginning students and researchers coming starting to run studies.

The details about how to run the studies themselves, how to interact with subjects and so on, are often learned solely through apprenticeship in a psychology or HCI lab. However, many researchers who are running or want to run studies do not have access to learning this tacit knowledge.

This half-day or full-day tutorial will provide participants with an overview of how to run studies with human participants, that is, not how to design or analyze studies but the practicalities of how to setup, debug, and run studies. It will help people running experiments to run them more effectively safely, and comfortably. Our purpose is to provide hands-on knowledge about experimental procedure.

The tutorial will cover the major topics noted in Figure 1. In particular, the tutorial will cover the role of identifying the research problem and reading in the general area; preparation for running a study, including piloting and IRB proposals; preparing to run a formal study, including advertising and recruiting subjects; running study sessions; and wrapping up a study.

## (II) How the tutorial will be delivered

The tutorial will cover the topics in Figure 1 using a lecture/discussion format. The topics will be introduced using a presentation and discussion will follow each section using scenarios and questions included in the book and developed for the Cognitive Science Conference. An early draft (approximately half the current length) of the material is available at [acs.ist.psu.edu/reports/ritterKM09.pdf](http://acs.ist.psu.edu/reports/ritterKM09.pdf), and published copies will be available in the future from Sage.

**Jong W. Kim (Jong.Kim@ucf.edu)**

++1 (814) 865-4453

Psychology, U. of Central Florida

Orlando, FL 32816 USA

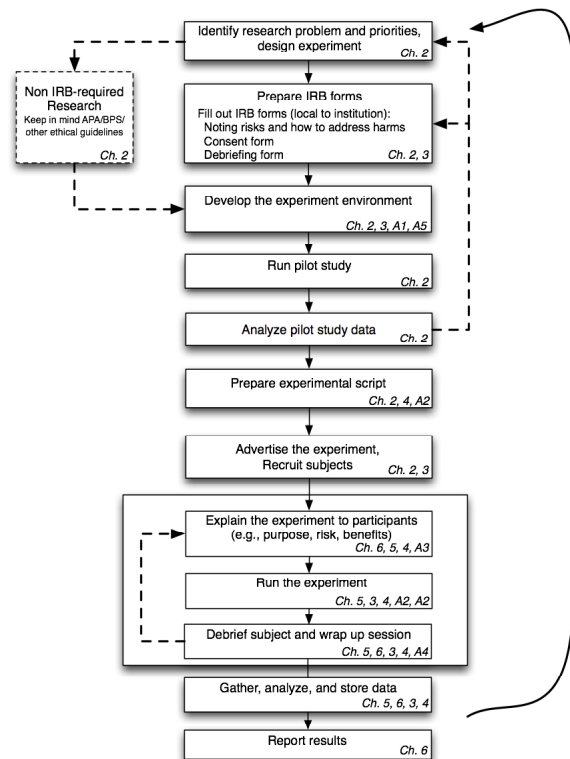


Figure 1. A pictorial summary of the research process with respect to running a human behavioral study. This is similar to, but developed separately from Bethel and Murphy's (2010) figure for human-robotic studies

A copy of the 121 page book as a printout will be provided (assuming that participant numbers can be specified well enough in advance or copied by the conference locally).

## (III) Why the presenter and authors are well suited to give a tutorial in the proposed area

The presenter is well qualified to prepare and present a tutorial in this area. Along with colleagues, Ritter has recently written a book for Sage on this topic (Ritter, Kim, Morgan, & Carlson, in press).

Ritter has also run and directed studies with human participants (e.g., Klein, Bennett, Whetzel, Granger, & Ritter, 2010; Reder & Ritter, 1992; Ritter, Freed, & Haskett, 2005; St. Amant, Horton, & Ritter, 2004; Yeh, Gregory, & Ritter, 2010). His collaborators on this tutorial and book include an industrial engineer (Kim), a research assistant who helps run studies (Morgan), and a professor of psychology who has been a member of an IRB board and



director of a psychology department subject pool (Carlson). While these co-authors will not be presenting, they will help prepare the slides and are co-authors of the book that will be given to attendees.

Ritter is also familiar with tutorials in general because he served as the first co-chair of tutorials at the Cognitive Science Conference in 1999. Since then he has served as tutorial chair or co-chair at the Cognitive Science Conference (2001, 2002, 2004, 2005), and at the International Conference on Cognitive Modeling (2004, 2006, 2007, 2009, 2010, 2012), and was the co-chair of the 2011 HCI Consortium Workshop, which was made up exclusively of tutorials on ways of knowing in HCI. In addition, he gave a tutorial on Soar at HCI International when it was in Japan and two invited lectures in Japan, has hosted a Japanese visitor, and published a paper in Japanese (Ritter, 2009).

This tutorial has been given at the *Behavior Representation in Modeling and Simulation (BRIMS 2012)* conference. The tutorial will be slightly modified for attendees at the Cognitive Science Conference by making it less practitioner/industry oriented, and making it more oriented for Asian and European researchers and for computer scientists. This will mean changing a few slides to represent problems more frequently found in academia than in industry, and assuming slightly different research questions are being asked, for example, a greater emphasis on cognitive science studies and less on controlled observation for product design.

#### **(IV) Why it is appropriate to have a tutorial in the proposed area?**

Practical skills on how to run studies are well known and well taught skills in psychology departments, but often not well known outside of psychology departments. Yet, in cognitive science, if the field believes in building computational models and gathering data to test those models (or starting the other way 'round, or having non-psychologists gather data), for example, work by Morita and colleagues (Morita, Miwa, Kojima, & Ritter, 2011), then how to gather that data is an important skill for every cognitive scientist, no matter their home discipline or outlook.

There are few teaching materials on the practical details on how to run studies, which this tutorial starts to address. So, this tutorial covers an established but not well documented or often formally taught common technique. The tutorial and related book will show that there are important aspects of this technique, and we would argue that without training these aspects are not well known to researchers outside of psychology, and put the resulting researchers and research done by those not trained at risk for failure, interpretable results, or incorrect results.

#### **(V) The likely audience for the tutorial.**

Earlier versions of the material have been used in teaching graduate courses at Carleton University (cognitive science, Canada), U. of Connecticut (human factors, US), Florida Institute of Technology (HCI), U. of Texas at Houston

(medical informatics), Middlesex U. (HCI, UK), Georgia Tech (industrial engineering), and at Penn State (information sciences and HCI). So, we believe that is accessible and useful to undergraduate and graduate students who are working with human participant studies, but are outside of psychology departments.

So, the likely audience for the tutorial are students and researchers outside of psychology departments who are running studies with humans in cognitive science, HCI, and related disciplines. It will also be useful to researchers in industry who are interested in running safer, more efficient, more controlled experiments.

#### **Acknowledgments**

This work was sponsored by ONR (W911QY-07-01-0004 and #N00014-10-1-0401).

#### **References**

- Bethel, C. L., & Murphy, R. M. (2010). Review of human studies methods in HRI and recommendations. *International Journal of Social Robotics*, 2, 347–359.
- Klein, L. C., Bennett, J. M., Whetzel, C. A., Granger, D. A., & Ritter, F. E. (2010). Caffeine and stress alter salivary  $\alpha$ -Amylase levels in young men. *Human Psychopharmacology: Clinical and Experimental*, 25, 359–367.
- Morita, J., Miwa, K., Kojima, K., & Ritter, F. E. (2011). Modeling decision making on the use of automation. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 1971–1976. Cognitive Science Society: Austin, TX.
- Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not the answer. *Journal of Experimental Psychology : Learning, Memory & Cognition*, 18(3), 435–451.
- Ritter, F. E. (2009). 認知モデリングの二つのフロンティア。感情とユーザビリティ (Two cognitive modeling frontiers: Emotions and usability). 認知科学におけるモデルベースアプローチ」 (*Transactions of the Japanese Society for Artificial Intelligence*), 24(2), 245–252. Translated into Japanese by Junya Morita.
- Ritter, F. E., Freed, A. R., & Haskett, O. L. (2005). User information needs: The case of university department web sites. *ACM interactions*, 12(5), 19–27. [acs.ist.psu.edu/acs-lab/reports/ritterFH02.pdf](http://acs.ist.psu.edu/acs-lab/reports/ritterFH02.pdf).
- Ritter, F. E., Kim, J. W., Morgan, J. H., & Carlson, R. A. (in press). *How to run experiments: A practical guide to research with human participants*. Currently 211 pages. Thousand Oaks, CA: Sage.
- St. Amant, R., Horton, T. E., & Ritter, F. E. (2004). Model-based evaluation of cell phone menu interaction. In *Proceedings of the CHI'04 Conference on Human Factors in Computer Systems*, 343–350. ACM: New York, NY.
- Yeh, K.-C., Gregory, J. P., & Ritter, F. E. (2010). One Laptop per Child: Polishing up the XO Laptop user experience. *Ergonomics in Design*, 18(3), 8–13.



# Robotics and Emotion

**Rolf Pfeifer (pfeifer@ifi.uzh.ch)**

Artificial Intelligence Laboratory, Department of Informatics, University of Zurich  
Andreasstrasse 15 (Office 2.32), 8050 Zurich, Switzerland

**Hiroshi Ishiguro (ishiguro@sys.osaka-u.ac.jp)**

Intelligent Robotics Laboratory, Dept. of Systems Innovation,  
Graduate School of Engineering Science, Osaka University,  
1-3, Machikaneyama-Cho, Toyonaka-Shi, Osaka-Fu, Japan

**Yuichiro Anzai (anzai@jsps.go.jp)**

Japan Society for the Promotion of Science  
8, Ichiban-Cho, Chiyoda-Ku, Tokyo, Japan

**Naomi Miyake (nmiyake@p.u-tokyo.ac.jp)**

Graduate School of Education, The University of Tokyo,  
7-3-1, Hongo, Bunkyo-Ku, Tokyo, Japan

**Keywords:** Fungus-eater robots, Emotional fungus-eaters,  
Urge theory.

Robotics studies have developed at many places across the globe and explored many different approaches. Among others, Japan has been one of the leading countries promoting robotics studies, breaking new ground with establishing Human Robot Interaction as an international society and leading the world with very human-like Geminoids. In the 1940's, Masanao Toda, a founder of cognitive science in Japan, proposed a visionary robotic system to explore and understand the function of emotion as a trigger of cognitive mechanisms for survival. For example, when a human was exploring in the ancient wilderness, fear must have worked as a switch between the exploratory mode and the find-an-escape mode to promote survival. The thought-experimental robot, the Fungus Eater, seeded some of the early AI research when it was introduced to psychologists in the U.S. and the Netherlands (Toda, 2000, 1982).

In this symposium, we will look to a few of the starting points of robotics research, like that of Toda's, and explore how this has expanded to include AI and robotics researchers in Europe, particularly with emphasis on embodiment and emotion, and how this has influenced new developments of robotics research in Japan. Our aim is to explore how to extend our understandings of human cognition through the eye of robotics, allowing us to reflect upon ourselves more directly than carefully scripted experiments.

The first speaker, Rolf Pfeifer, who has once connected his work on robot's emotion to Toda's fungus eater robots, will reflect on his own and Toda's work, to foresee the future of cognitive sciences, through the development of robotics studies. The second speaker, Yuichiro Anzai, has witnessed the beginning of the raise of active cognitive sciences in the U.S. in Herb Simon's lab during the late

1970's, and actually worked with Toda at Hokkaido University, to implement Toda's ideas into robotics reality. Reflecting on his own trajectory as a cognitive scientist, reflects on Toda's work to push open new research topics under the theme of robotics and emotion. The third speaker, Hiroshi Ishiguro, as a leading robotics engineer studying human cognition through the eyes of most human-like robots like his own Geminoid, will push him further to integrate neuro-scientific studies, to gain further insights. Putting all three together, we hope to see how the topics of robotics and emotion would re-open a new research field for cognitive sciences.

## Do robots need emotions? An embodied perspective

Rolf Pfeifer (University of Zurich)

Traditionally, in robotics, artificial intelligence, and neuroscience, there has been a focus on the study of the control or the neural system itself. Recently there has been an increasing interest into the notion of embodiment in all disciplines dealing with intelligent behavior, including cognitive science, psychology, philosophy, and linguistics. In an embodied perspective, cognition and emotion are conceived as emergent from the interaction of brain, body, and environment, or more generally from the relation between physical (including physiological) and information (neural, control) processes. It can be shown that through the embodied interaction with the environment, in particular through sensory-motor coordination, information structure is induced in the sensory data, thus facilitating categorization, perception and learning. The patterns thus induced depend jointly on the morphology, the material characteristics, the action and the environment. Because biological systems are mostly "soft", a new engineering discipline, "soft robotics", has taken shape over the last few

years. This is of particular interest to the fields of developmental, social and service robotics, where robots will share their living space with humans and safe and pleasant interaction is at center stage. I will analyze the different roles that emotion can play in this process. Moreover, I will discuss the far-reaching implications of embodiment, in particular of having a soft body, on our view of the mind and human behavior in general: Cognition and emotion are no longer centralized in the brain, but distributed throughout the organism. Because in "soft" systems part of the functionality is in the morphology, the physiology, and the materials, there is no longer a clear separation between control and the to-be-controlled, which implies that we need to fundamentally re-think the notion of control. These ideas will all be illustrated with case studies from biology -- humans and animals -- and robotics. Finally, I will try to establish the relation between these thoughts, Toda's "Fungus Eaters", and his theory of emotion.

### **From fungus-eater robots to emotional, socially interact-able robots**

Yuichiro Anzai (JSPS)

Toda had been trained as a theoretical physicist before he changed his research fields into psychology. This shift contributed to free him from the burden of stereotypic academic psychological thinking, venturing into under-researched topics of emotion, not in isolated fashion but in a more integrated, general-systemic perspectives of his own. Toda's combination of the two fields has given him some unique insights on the nature of man and time. Toda believed that a system of emotions is an evolutionally organized survival mechanism with nearly optimal operating characteristics, which would explain the complexities of our social interaction, altruism, social coalition, as well as the structures of the society. As one of the closest younger colleagues with him, I will develop my own interpretation of his work, to foresee the research only possible with an engineering orientation like robotics.

### **Cognitive Neuroscience Robotics**

Hiroshi Ishiguro (Osaka University/ATR)

Robotics is not just engineering but also human science. The current robotics focuses on *interaction* in addition to *navigation* and *manipulation* that were major topics in previous robotics. Researchers are developing interactive humanoids and androids with humans by using technologies developed in robotics and knowledge found in cognitive science and neuroscience. On the other hand, the developed robots, humanoids and androids, are important research platforms for cognitive science and neuroscience.

Thus, robotics is tightly coupled with the human sciences. For example, we can studies effects of human-like

appearance in inter-personal and social situations by using androids mimicking human appearances. We can study social relationships and the dynamics among humans and robots by using multiple interactive humanoids.

We call this new robotics *cognitive robotics* or *cognitive neuroscience robotics*. This talk will introduce a series of robots developed in Osaka University and ATR and discuss cognitive experiments using them. Especially, the androids give us various insights on human-human and human-robot interactions in inter-personal and social situations.

Researchers joining to the Center Of Excellence program titled *cognitive neuroscience robotics* in Osaka University are sharing this new interdisciplinary research area and working together. This talk also introduces several topics from the program, some of which are related to the topics on studies of emotion.

### **References**

- Toda, M., (2000). The urge theory of emotion and social interaction : Emotions and urges (Revised). SIST Chukyo University Technical Report, No. 1999-1-01.
- Toda, M. (1997). The urge theory of emotion and social interaction: The Humankind as active copers: Human societies and human values
- Toda, M., (1996a). The urge theory of emotion and cognition: The problems of altruism and evolution of systems. SIST Chukyo University Technical Report, No. 95-1.04.
- Toda, M., (1996b). The urge theory of emotion and social interaction: Interpersonal compatibility and close coalitions, SIST Chukyo University Technical Report, No. 96-1-01.
- Toda, M., (1995). The urge theory of emotion and cognition: A decision theoretical model of urge operations, SIST Chukyo University Technical Report, No. 95-1-01.
- Toda, M., (1994a). The Urge Theory of emotion and cognition : Basic structure of the urge operations. SIST Chukyo University Technical Report, No. 94-1-01.
- Toda, M., (1994b). Emotion, society, and versatile architecture, SIST Chukyo University Technical Report, No. 94-1-02.
- Toda, M., (1993). The urge theory of emotion and cognition: Emotion and urge. SIST Chukyo University Technical Report, No. 93-1-01.
- Toda, M., & Higuchi, K., (1990). Common sense, emotion, and chatting, and their roles in interpersonal interactions, SIST Chukyo University Technical Report, No. 90-1-01
- Toda, M., (1982). Man, robot, and society: Models and speculations, The Hague: Martinus Nijhoff Publishing.

# Computational Models of Intuitive Physics

**Peter Battaglia** (pbatt@mit.edu)

**Tomer Ullman** (tomru@mit.edu)

**Joshua Tenenbaum** (jbt@mit.edu; moderator)

MIT, BCS Dept., 77 Massachusetts Ave.

Cambridge, MA 02139 USA

**Adam Sanborn** (a.n.sanborn@warwick.ac.uk)

Dept. of Psychology, University of Warwick

Coventry, CV4 7AL UK

**Kenneth Forbus** (forbus@northwestern.edu)

Northwestern University, EECS Dept., 2133 Sheridan Rd

Evanston, IL 60208 USA

**Tobias Gerstenberg** (t.gerstenberg@ucl.ac.uk)

**David Lagnado** (d.lagado@ucl.ac.uk)

Cognitive, Perceptual, and Brain Sciences Dept., UCL,

26, Bedford Way, London WC1H 0AP, UK

**Keywords:** Intuitive physics; qualitative reasoning; probabilistic programming; Bayesian models; psychophysics.

People have a powerful “physical intelligence” – an ability to infer physical properties of objects and predict future states in complex, dynamic scenes – which they use to interpret their surroundings, plan safe and effective actions, build and understand devices and machines, and communicate efficiently. For instance, you can choose where to place your coffee to prevent it from spilling, arrange books in a stable stack, judge the relative weights of objects after watching them collide, and construct systems of levers and pulleys to manipulate heavy objects. These behaviors suggest that the mind relies on a sophisticated physical reasoning system, and for decades cognitive scientists have been interested in the content of this knowledge, how it is used and how it is acquired. In the last few years, there has been exciting progress in answering these questions in formal computational terms, with the maturation of several different traditions of cognitive modeling that have independently come to take intuitive physics as a central object of study. The goals of this symposium are to: 1) highlight these recent computational developments, focusing chiefly on qualitative reasoning (QR) models and Bayesian perceptual and cognitive models; 2) begin a dialog between leading proponents of these different approaches, discussing a number of dimensions along which the approaches appear to differ and working towards bridging those differences; 3) enrich these models with perspectives from empirical work in cognitive science.

**Background.** The research to be discussed builds on several decades of prior work from multiple traditions in cognitive science. Cognitive psychologists since the 1970s have studied the role that human intuitive physics plays in development, perception, education, and reasoning. Behavioral research with adults focused on identifying errors and biases in people's general understanding and theories about physical rules (McCloskey, 1983), as well as psychophysical studies of how sensory cues drive specific judgments in dynamic displays (Todd & Warren, 1982). Early and ongoing developmental work has identified milestones in cognitive sensitivity and expectations about core physical principles (Baillargeon, 2007). Though these efforts have made significant progress, they did not frame

their results as computational models with sufficient clarity and power to explain people's physical reasoning in complex and varied scenes.

Crucial computational progress has come from the fields of human and computer vision, artificial intelligence (AI), and machine learning. Human and machine vision researchers have recently developed computational models of natural scene understanding (Oliva & Torralba, 2007), but their focus has been on knowledge about the geometry and semantics of scene layouts, not the role of physical constraints and how physical properties are represented and exploited for prediction, reasoning and planning. AI researchers have been developing frameworks for qualitative reasoning (QR) and applying them to physical domains for over 30 years, and these approaches have now matured to the point that they can both solve challenging real-world inference problems and engage directly with behavioral experiments, giving state-of-the-art accounts of people's intuitive reasoning in a wide range of science and engineering domains (Forbus, 2011). The framework of Bayesian reasoning in probabilistic generative models has revolutionized AI and machine learning, and in the last decade has also come to provide a lingua franca for sophisticated reverse-engineering models of human perception, action and cognition (Chater et al, 2006; Tenenbaum et al, 2011). But only in the last few years have Bayesian models been applied to challenging physical reasoning problems, and been shown to give strong quantitative accounts of human physical judgments (Sanborn et al, 2009; Hamrick et al, 2011).

This symposium brings together leading researchers modeling intuitive physics from the QR, Bayesian cognition and perceptual modeling traditions, to discuss highlights of recent models and points of contact and contrast between different modeling approaches. The talks and discussion will explore several axes in the space of possible models, including the following: rational reverse-engineering vs. descriptive or heuristic accounts; qualitative vs. quantitative reasoning; probabilistic vs. deterministic inference; lower-level perceptual vs. higher-level cognitive inferences; implicit vs. explicit reasoning; analog simulation vs. symbolic rule-based representations; the role of memory-, experience- and learning-dependent reasoning; the role of

causal, counterfactual and explanatory reasoning; reasoning about simple rigid bodies vs. complex physical entities and concepts, like non-rigid objects, non-solid substances, fluids, gasses, heat; simple scenarios with few objects moving in simple ways vs. compound scenes of many objects interacting and moving according to complex dynamics.

The speakers come from various avenues of artificial intelligence and cognitive science: Sanborn studies computational models of memory and cognition; Battaglia, computational perception and motor control; Forbus, AI and qualitative reasoning; Tenenbaum, learning and inference in humans and machines.

#### **Sanborn: Reconciling intuitive and Newtonian mechanics for colliding objects**

People have strong intuitions about the masses of objects and the causal forces that they exert upon one another when they collide. These intuitions appear to deviate from Newtonian physics, leading researchers to conclude that people use a set of heuristics to make judgments about collisions. We show that people's judgments about mass are indeed consistent with Newtonian physics, provided uncertainty about the velocities of the objects is taken into account. The resulting rational model of intuitive dynamics easily extends to accommodate other aspects of people's inferences about physical causation, such as judgments of whether one object caused another to move. We argue that intuition and physics need not be divorced, and that a simple psychological process - stochastically approximating Bayesian inference by recalling previous collisions - can bring them together.

#### **Battaglia: Intuitive mechanics in physical reasoning**

I will explore the idea that the brain has an "intuitive mechanics", a realistic model of physics that can estimate physical properties and predict probable futures. This intuitive mechanics is surprisingly faithful to the laws of classical mechanics, it captures statics, dynamics, forces, collisions, and friction. It is fundamentally probabilistic, it supports Bayesian inferences that robustly handle uncertainty, and, like people, its predictions can deviate from objective reality. And, it is resource-bounded, supporting only judgments that can be made based on a few low-precision, short-lived simulations. We conducted a series of psychophysical experiments in which participants made physical judgments about various complex, 3D scenes, and found that this formal model of intuitive mechanics well-predicts people's responses by accounting for their accuracy and several systematic biases. These results suggest that an approximate, probabilistic model of physics forms the basis of human physical reasoning. More generally, this principled computational approach provides a unifying framework for analyzing and understanding this crucial part of human cognition.

#### **Forbus: Qualitative modeling: Capturing human reasoning about the physical world**

There is ample evidence that qualitative representations of space, quantity, and causality capture important regularities of human reasoning about physical situations and systems

(Forbus, 2011). Qualitative reasoning has been used to model intuitive phenomena, such as motion, liquids, and heat. It has also been used to model aspects of the reasoning of scientists and engineers, such as guiding the solution of quantitative problems and extracting insights about complex dynamical systems from visual data. Qualitative representations of space provide a bridge between perception and conceptual knowledge, and can be used to model visual problem solving. When combined with analogical reasoning, qualitative models can provide explanations for aspects of conceptual change (eg. Friedman & Forbus, 2010). This talk will summarize recent work on modeling conceptual change concerning intuitive notions of force, the human circulatory system, and how the seasons change. There is great potential for synthesis between qualitative and Bayesian modeling: Qualitative modeling provides formal languages for hypotheses, while statistical information (in our case, computed automatically via analogical generalization over examples) provides criteria for accepting hypotheses.

#### **Tenenbaum: Integrative perspectives**

I will discuss the prospects for building computational models of intuitive physical reasoning that integrate features of qualitative and probabilistic approaches introduced earlier in the symposium, and present preliminary results on several lines of work exploring this integration. Specific points will include (1) using qualitative reasoning to generate efficient proposals for Monte Carlo-based approximate inference in probabilistic models; (2) using dynamic probabilistic models as the basis for linguistic ascriptions of causal responsibility and explanatory reasoning (joint work with Gerstenberg and Langado); (3) modeling conceptual change in intuitive physics via hierarchical Bayesian inference over symbolic expressions for physical laws (joint work with Ullman).

#### **References**

- Baillargeon, R. (2007). The acquisition of physical knowledge in infancy: A summary in eight lessons. *Blackwell handbook of childhood cognitive development*. Blackwell.
- Chater, N., Tenenbaum, J.B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10, 292-293.
- Forbus, K. (2011). Qualitative Modeling. *WIREs: Cognitive Science*. 2(4), pp 374-391, July/August.
- Friedman, S. & Forbus, K. (2010). An integrated systems approach to explanation-based conceptual change. *Proc. AAAI10*.
- Hamrick, J., Battaglia, P.W., & Tenenbaum, J.B. (2011). Internal physics models guide probabilistic judgments about object dynamics, *Proc. 33rd Ann. Conf. Cognitive Science Society*.
- McCloskey, M. (1983). Intuitive physics. *Scientific American*, 248(4), 122-130.
- Oliva, A. & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12), pp. 520-527.
- Sanborn, A., Mansinghka, V., & Griffiths, T. (2009). A Bayesian framework for modeling intuitive dynamics. *Proc. 31st Ann. Conf. Cognitive Science Society*.
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331 (6022), 1279-1285.
- Todd, J., & Warren, W. (1982). Visual perception of relative mass in dynamic events. *Perception*, 11(3), 325-335.

# The role of comparison in structure learning: Developmental, learning science, and computational perspectives

**Stella Christie (christie@gmail.com) & Dedre Gentner (gentner@northwestern.edu)**

University of British Columbia, Dept. Psychology; Northwestern University, Dept. of Psychology

**Mutsumi Imai (imai@sfc.keio.ac.jp), Etsuko Haryu (haryu@p.u-tokyo.ac.jp) & Hiroyuki Okada (h.okada@eng.tamagawa.ac.jp)**

Keio University, Dept. Psychology; University of Tokyo, Dept. of Education; Tamagawa University, Brain Science Institute

**Ji Y. Son (json2@calstatela.edu) & James W. Stigler (stigler@ucla.edu)**

California State University, Los Angeles, Dept. Psychology; UCLA, Dept. of Psychology

**Leonidas A.A. Doumas (leonidas@hawaii.edu), Robert G. Morrison (Morrison@luc.edu), & Lindsey E. Richland (lrichland@uchicago.edu)**

University of Hawaii, Dept. Psychology; Loyola University, Dept. of Psychology; University of Chicago, Dept. of Human Development

**Keywords:** relational thinking, analogy, comparison, cognitive development, computational modeling

The ability to perceive, comprehend and reason about relations (i.e., *relational thinking*) is central in human cognition. Relational thinking is powerful because it is structured. Specifically, relational thought allows inferences and generalizations that are constrained by the *roles* that elements play, rather than strictly the properties of the elements themselves.

The role of relational comparisons in learning is emerging as an important area of developmental and learning science research. Relational comparisons allow learners to derive symbolic, abstract, and conceptual knowledge representations that are generative, in that children and adults can then use them broadly in new contexts to reason about new elements. Indeed, comparison seems to underlie the very development of the structured relational representations that underlie relational cognition.

This symposium aims to bring to together research on the role of comparison in developmental and adult learning. Specifically, we present research on the role of comparison in the development of spatial reasoning, language learning, adult mathematics learning, and computational approaches to learning structured (i.e., symbolic) representations.

## Christie & Gentner: Domain Specific vs Abstract Language in Spatial Learning

Many studies have suggested that language provides important tools for learning and thinking in cognitive development. In this work we test one specific claim concerning the cognitive effects of language learning: namely, that systematic semantic structure in language can invite correspondingly systematic conceptual structure (Gentner, 2010; Gentner & Christie, 2011). Evidence for this claim comes from prior studies by Loewenstein and Gentner (2005) in which children performed better on a difficult spatial mapping task involving three-tiered structures when they were given the monotonic set of spatial terms *top, middle, bottom*

than when they were given the less systematic set of terms *on, in, under*. To discover the generality of these effects, in this series of studies we asked whether children given *nonspatial* (but systematic) language would still show an advantage in the spatial mapping task. We presented children with a spatial mapping task as in Loewenstein & Gentner (2005). There were three groups: one heard a systematic set of spatial terms (*top/middle/bottom*); one heard a systematic set of nonspatial terms (*one/two/three*); and a third heard a nonsystematic set of nonspatial terms (*dog/pig/cat*). In addition to the standard three-tiered mapping task, we also conducted a vertical-to-horizontal mapping task. The results suggest that (1) children benefit from systematic language; (2) domain-specificity benefits early learning; and (3) at older ages, abstract language can have a larger advantage in a difficult transfer task.

## Imai, Haryu, & Okada: Progressive alignment in verb learning

Verbs should be extended by the sameness of action, whereas nouns should be extended attending to similarity of objects. Children under four years of age easily generalize a novel noun to other objects of like kinds, whereas even 4-year-olds tend to fail extending a novel verb to the same action performed by a different agent or with a different object (Imai et al., 2005, 2008). Children fail to segregate the action from the objects constituting it. In other words, children fail to structurally align action events. Previous research suggests that object similarity between objects in corresponding relational roles can promote structural alignment and help children notice higher-order relational similarity (e.g., Gentner & Toupin, 1986). Borrowing this idea, two experiments examined whether young children's verb generalization would be fostered by similarities between corresponding objects in the two events.

In the first experiment 4 year-old children were shown a video in which a woman was doing a novel action with a novel object, and heard a novel verb. Children were then asked to extend the verb to either a

situation where the action from the video was performed on a novel object (AS), or a novel action was performed on the object from the video (OS). In the AS video, the object was either similar in shape to the object in the original action event (*same object condition*), or dissimilar to the original object (*dissimilar object condition*). Children performed better in the similar object condition, suggesting that object similarity enhanced overall similarity across events and helped children map a novel verb to the same action.

The second study tested whether verb generalization with the help of object similarity can bootstrap 4-year-olds into verb generalization even with perceptually dissimilar objects. Indeed, four-year-olds succeeded in verb generalization across dissimilar objects after having experienced a verb generalization task with similar objects; but they failed when they had experienced verb generalization with dissimilar objects from the beginning.

### **Son & Stigler: Fragmented analogies from procedural understanding of mathematics**

Cross-national comparisons of math pedagogy (e.g., Stevenson & Stigler, 1994) indicate that US classrooms are highly focused on procedures without explanation of their conceptual foundations. The long-term consequences of such pedagogy are dire. Even though the domain of mathematics fundamentally requires an understanding of quantitative relations, students may merely amass a collection of seemingly arbitrary rules along with fragments of relational knowledge. Although analogical processes are typically powerful for reasoning across domains, when rules and procedures are not grounded in relational concepts, students may exhibit fragile or incorrect mappings across contexts thus resulting in inconsistent quantitative reasoning. We examined this hypothesis in a sample of college students (mostly Psychology majors) enrolled in a statistics course. In two studies, students were asked to reason about the results of dividing a positive value,  $a$ , with integers (e.g.,  $a/5$  vs.  $a/9$ ), decimals (e.g.,  $a/.1$  vs.  $a/.05$ ), and variables (e.g.,  $a/n$  vs.  $a/(n-1)$ , given that  $n > 1$ ). Students were asked to indicate which of two given values was larger and why. The integer problem was presented first because it could serve as a potential source for analogical transfer. The first study was conducted with individual interviews where students often chose not to use a pen and paper that was available to them. In study 2, students were asked to write down their choices and rationale. Judgments of quantity in the context of decimals and variables were reliably worse than with integers. Examinations of the rationale given for their choices showed that different numerical contexts yielded distinctly different reasoning strategies. Strategies used for reasoning about integers were either

abandoned or misapplied when reasoning about decimals or variables. Research on analogical reasoning may help educators remedy such fragmented understanding.

### **Doumas: Developing structure**

DORA (Discovery Of Relations by Analogy; Doumas, Hummel, & Sanhofer, 2008) is a symbolic connectionist network that uses time as a signal to dynamically bind distributed (i.e., connectionist) representations of relational roles and objects into explicitly relational (i.e., symbolic) structures. DORA relies on the processes of analogical mapping and intersection discovery to highlight shared abstract properties between separate systems and subsequently predicates these similarities as explicit (i.e., symbolic) representations that can be bound to arguments. Subsequently, DORA can exploit the pattern of activation that emerges between mapped role-filler pairs as a cue to combine these sets of role-filler pairs into a single multi-place relational structure. These processes permit the discovery and predication of shared properties and relations across otherwise different systems and thus allow DORA to learn structured representations from unstructured examples. The DORA model has been used to simulate more than 20 phenomena from child and adult relation learning (e.g., Doumas & Hummel, 2010; Doumas et al., 2008). We propose that DORA's learning mechanism provides an account of how humans learn relational representations and the development of analogical reasoning.

### **References**

- Doumas, L. A. A. & Hummel, J. E. (2010). A computational account of the development of the representations underlying object recognition. *Cognitive Science*, 34, 698-712.
- Doumas, L. A. A., Hummel, J. E., & Sanhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, 115, 1-43.
- Gentner, D. (2010). Bootstrapping children's learning: Analogical processes and symbol systems. *Cognitive Science*, 34 (5). 752-775.
- Gentner, D. & Christie, S. (2010). Mutual bootstrapping between language and analogical processing. *Language and Cognition*, 2(2). 261-283.
- Gentner, D. and Toupin, C. (1986). Systematicity and Surface Similarity in the Development of Analogy. *Cognitive Science*, 10: 277-300.
- Imai, M., Haryu, E., & Okada, H. (2005). Mapping novel nouns and verbs onto dynamic action events: Are verb meanings easier to learn than noun meanings for Japanese children? *Child Development*, 76, 340-355.
- Imai, M., Li, L., Haryu, E., Okada, H., Hirsh-Pasek, K., Golinkoff, R. & Shigematsu, J. (2008). Novel noun and verb learning in Chinese-, English-, and Japanese-speaking children. *Child Development*, 79, 979-1000.
- Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, 50, 315-353.
- Stevenson, H., & Stigler, J. W. (1994). *Learning Gap: Why Our Schools Are Failing and What We Can Learn from Japanese and Chinese Education*. New York, NY: Simon & Schuster.

# Neural Computations Supporting Cognition: Rumelhart Prize Symposium in Honor of Peter Dayan

## Participants

**Kenji Doya (doya@oist.jap)**

Neural Computation Unit, Okinawa Institute of Science and  
Technology, 1919-1 Tancha, Onna  
Okinawa 904-0495 Japan

**John O'Doherty (jodoherty@caltech.edu)**

Division of the Humanities and Social Sciences, California  
Institute of Technology, MC 228-77  
Pasadena, CA 91125 USA

**Alexandre Pouget (alex@cvs.rochester.edu)**

Département de neuroscience fondamentale, Université de  
Genève, 1 rue Michel-Servet  
CH-1211 Geneva 4, Switzerland

**Peter Bossaerts (pbs@hss.caltech.edu)**

Division of the Humanities and Social Sciences, California  
Institute of Technology, MC 228-77  
Pasadena, CA 91125 USA

## Organizers

**Nathaniel Daw (daw@cns.nyu.edu)**

Center for Neural Science  
New York University  
New York, NY, 10003

**Yael Niv (yael@princeton.edu)**

Princeton Neuroscience Institute and Psychology Department  
Princeton University,  
Princeton, NJ, 08544

**Keywords:** neural computation; reinforcement learning;  
inference; uncertainty

## Motivation

Principles of sound statistical inference underpin prominent accounts for a variety of cognitive phenomena, including perception, learning, and decision-making. Linking these building blocks of cognition to the biological substrate that supports them, recent work has investigated how the brain implements probabilistic inference and learning under uncertainty. The interplay between the psychological and biological levels of analysis has shed light on the structure of cognition and computation at both levels.

This symposium builds on Peter Dayan's seminal contributions to linking psychological, neural and computational phenomena. In particular, speakers will discuss recent work growing out of two areas where Dayan made early and fundamental contributions: the brain's mechanisms for reinforcement learning, and neural representations supporting probabilistic inference under uncertainty.

## Reinforcement learning and the basal ganglia

**Authors:** Kenji Doya and Makoto Ito

**Abstract:** The discovery of the parallel between the firing of dopamine neurons and the temporal difference error signal of the reinforcement theory in the 1990s brought a breakthrough in understanding the function of the basal ganglia. Previously the most enigmatic part of the brain is now considered as the center for linking perception, action,

and reward. After more than a decade from the discovery, however, there still remain questions to be answered, such as what striatal neuron firing represents, how and where an action is selected, and how negative reinforcement is realized. Here we review Peter Dayan's seminal contributions and recent developments.

## Fractionating model-based reinforcement-learning its component neural processes

**Author:** John P. O'Doherty

**Abstract:** It has recently been proposed that action-selection in the mammalian brain depends on at least two distinct mechanisms: a model-free reinforcement learning (RL) mechanism in which actions are selected on the basis of cached values acquired through trial and error, and a model-based RL system in which actions are chosen using values computed on-line by means of a rich cognitive model of the decision problem and knowledge of the current incentive value of goals. While much is now known about the putative neural substrates of the model-free RL system and its concomitant temporal difference prediction error, much less is known about how model-based RL is implemented at the neural level. In this talk I will review recent evidence from a series of functional neuroimaging studies in humans supporting the presence of neural signals within a wide expanse of cortex that are relevant to model-based RL. These include, a state-action based prediction error signal within a fronto-parietal network that could mediate learning of the cognitive model, a goal-value signal encoding the value of putative goal-outcomes within the

ventromedial prefrontal cortex, computations corresponding to action-contingency within inferior parietal cortex, and the representation of the effort costs of an action within the dorsomedial frontal cortex. These different computations then need to be integrated in order to construct an overall model-based action-value. Taken together, this evidence suggests that model-based reinforcement-learning theory provides a scaffold upon which a deeper understanding of the functions of a large extent of cortical territory within the mammalian brain can be built.

### **Probabilistic inferences in neural circuits using probabilistic population codes**

**Author:** Alexandre Pouget

**Abstract:** A wide range of seemingly unrelated behaviors can be formalized as instances of probabilistic inferences. This includes odor recognition, sensorimotor transformations, decision making, simple arithmetics and visual search, to name just a few. We will present a neural theory of probabilistic inference in which neurons encode probability distributions using a basis function decomposition of the log probability or log likelihood. This approach makes very specific predictions about the form of the variability in neural responses, as well as about the neural implementation of various probabilistic inferences such as product of distributions, marginalization and sampling. We will discuss several experimental tests of these predictions in the context of arithmetics, visual search and bistable perception.

### **The neural process of subjective belief formation in humans**

**Author:** Peter Bossaerts

**Abstract:** We present a general experimental paradigm with which to study human belief formation from experience. We first establish that human learning proceeds along Bayesian principles, but from subjective albeit robust priors rather than the true prior. Second, properly dissociating neural encoding of values and beliefs, we identify the default mode network as the locus of beliefs learned from frequencies. Third, we study the neural basis for combining objective frequentist information with prior beliefs and discover that Bayesian posterior beliefs are encoded bilaterally in the lateral prefrontal cortex.



# Thirty years of Marr's Vision: Levels of Analysis in Cognitive Science

## Participants

**Chris Eliasmith (celiasmith@uwaterloo.ca)**

Centre for Theoretical Neuroscience, University of Waterloo,  
200 University Avenue West, Waterloo, Ontario, Canada,  
N2L 3G1

**Valerie Gray Hardcastle  
(valerie.hardcastle@uc.edu)**

Departments of Philosophy and Psychology, University of  
Cincinnati,  
Cincinnati, OH 45221, USA

**Tom Griffiths (tom\_griffiths@berkeley.edu)**

Institute of Cognitive and Brain Sciences, University of  
California, Berkeley,  
132 Barker Hall, MC 3190, Berkeley, CA 94720, USA

**Bradley C. Love (b.love@ucl.ac.uk)**

Department of Cognitive, Perceptual and Brain Sciences,  
University College London,  
26 Bedford Way, London, UK, WC1H 0AP

## Discussant

**William Bechtel (bill@mechanism.ucsd.edu)**

Department of Philosophy, University of California, San Diego,  
9500 Gilman Drive, La Jolla, CA 92093, USA

## Organisers

**Richard P. Cooper (R.Cooper@bbk.ac.uk)**

Department of Psychological Sciences,  
Birkbeck, University of London

**David Peebles (d.peebles@hud.ac.uk)**

Department of Behavioural and Social Sciences  
University of Huddersfield

**Keywords:** Marr; levels of description and explanation;  
computational level; algorithmic and representational level;  
implementational level.

## Introduction

Thirty years after Marr's landmark posthumous book, *Vision* (Marr, 1982), the argument for which he is most cited remains the distinction between computational, algorithmic and representational, and the implementation levels. In the interim, many reformulations of this basic distinction have been proposed, but is it still relevant? This symposium will discuss whether there is still a place for the algorithmic and representational level, with its cognitive-level concepts, given the rise in reductionist neuroscience from below and Bayesian analysis from above.

### Marr's Attacks: A Gentle Reminder Chris Eliasmith

Marr's (1982) three levels can be seen as the result of two deep concerns he had about how brain theories were being constructed in his day. We can see these concerns giving rise to 1) an attack on reductionism; and 2) an attack on vagueness.

With respect to reductionism, Marr was interested in ensuring the centrality of not only mechanisms, but also of their function to our generation of brain theories (Marr and Poggio, 1977). With respect to vagueness, Marr wanted to ensure that our high-level descriptions of neural phenomena could be tested against empirical data (Marr, 1975).

Unfortunately, many researchers after Marr seem to have taken his purpose to be a divisive one. Some, such as Pylyshyn (1984), refer to the "three autonomous levels of

description" (p. 259). In contrast, Marr (1982) seems to be suggesting that the intermediate, representational level, is a *bridge* between our more abstract characterizations, and more detailed characterizations (pp. 23-24).

I argue that Marr's levels should be understood in an integrative sense. I show that adopting this perspective provides critical constraints for building Marr-type brain models. I provide the details of one such model: a large-scale simulation of spiking neurons that reproduces detailed neural and behavioral results across a wide array of cognitive and non-cognitive tasks.

In short, adopting Marr's perspective on levels helps pave the way for the kind of unified models of brain function for which he, himself, was striving.

### Bridging Levels of Analysis for Probabilistic Models of Cognition Tom Griffiths

Most probabilistic models of cognition are intended to explain human behavior at the computational level, linking how people act to the solution to an abstract computational problem. This focus is quite different from that of other approaches to cognitive modeling, which tend to emphasize the algorithmic and implementational levels. This raises a number of important questions: When are theories at these different levels incompatible with one another? What are the implications of a computational-level analysis for theories at the other levels? How can we begin to draw connections across levels of analysis, for an integrated account of cognition? I will argue that we can only answer these questions by explicitly taking on the challenge of building a bridge between levels of analysis, considering how

computational-level models can be translated to the algorithmic and implementational levels and how algorithmic- and implementational-level accounts might be cast at the computational level. I will illustrate this argument with examples drawn from recent work looking at Monte Carlo methods as a source of “rational process models” and analyses of the computational-level commitments of artificial neural networks.

## **A New Appreciation for Marr’s Levels:**

### **Understanding How Brains Break**

**Valerie Gray Hardcastle and Kiah Hardcastle**

Much work in the cognitive sciences, including computational neuroscience, now focuses on brains performing less than optimally. That is, while the original programs in artificial intelligence and the like aimed to articulate what thought was in ideal terms, much research now looks at how and why brains or other cognitive engines fail to function as they should. This focus on impairment affects how one can understand Marr’s three levels. In this presentation, we use a method of exploring impulsivity and behavioral inhibition based on a neural network/ population activity model of the cortico-striatal circuitry as a case study to refine Marr’s distinctions. In particular, we will show that the computational level should be redefined, for simply knowing the goal of a computation may not tell us much about why something has gone wrong and why the information-processing device is exhibiting abnormal behavior. We will also argue, as have many others, that the distinction between algorithm and hardware largely collapses when considering the brain.

## **The Primacy of Mechanism in Cognitive Science**

**Brad Love**

Cognitive science is primarily concerned with the “how” questions of brain and behavior. These questions address mechanism, and therefore make contact with Marr’s algorithmic level. From below, mechanistic accounts can be informed, constrained, and inspired by neuroscience. Rather than being reduced by neuroscience, cognitive models are proving valuable in interpreting fMRI data because these mechanisms help neuroscientists understand the function of brain regions. From above, despite many cognitive scientists professing a devotion to the computational level, very few are trained or focus their research on characterizing evolutionary environments, niches, and histories. I will argue that explanations formulated purely at the computational level are not sufficiently constrained, because rational Bayesian models are uninformed by a wide range of process-level data and their assumptions about the environment are generally not grounded in empirical measurement.

Given the recent surge of interest in computational-level theories of cognition, one question is whether integration across algorithmic and computational levels would be

beneficial. One promising avenue for integration is to evaluate the representations on which Bayesian inference operates and the algorithms and heuristics that carry it out as psychological mechanisms. In other words, one means of integration is to evaluate Bayesian models at the algorithmic level. A number of researchers have adopted this strategy and have concluded that humans engage in forms of approximate Bayesian inference that are intended to reflect human capacity limitations. Although an improvement over purely rational approaches, approximate Bayesian models face significant challenges. One challenge is that people are suboptimal for reasons other than capacity limitations. In domains where people’s behavior falls far short of that predicted by rational accounts, the layering of capacity limitations and suboptimality onto the rational account may only serve as a lengthy detour to the algorithmic level.

## **Differentiating While Integrating Levels**

**William Bechtel**

Are all three of Marr’s levels needed? Should they be kept distinct? Symposiasts emphasize how cognitive science is or should integrate Marr’s levels. This is important, but it is also important to emphasize the distinct contributions and methodologies of each level of inquiry. They represent three different perspectives required to understand mechanisms generally, but especially information processing mechanisms. Marr viewed neuroscience of his day as emphasizing the material implementation at the expense of the algorithmic-representational and computational levels, and that has been true of mechanistic science generally. But mechanisms only work insofar as they are organized, and this is especially true of information processing mechanisms that must insure that information is encoded appropriately within the mechanism and made available to the operations that require it. Moreover, it is crucial to understand how a mechanism functions in broader environments that determine the computations it needs to perform (and may fail to perform). Different modes of inquiry are required to examine each of these. This is especially true of the computational perspective, which requires looking outside the mechanism to the environment in which it operates and engaging in appropriate experimental and theoretical studies to understand what those demands really are.

## **References**

- Marr, D. (1975). Approaches to biological information processing. *Science*, 190:875-876.
- Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: W. H. Freeman.
- Marr, D. and Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosciences Research Program Bulletin*, 15, 470-488.
- Pylyshyn, Z. (1984). *Computation and Cognition*, Cambridge, MA: MIT Press.

# New Frontiers in Computational Models of Grammatical Development

**Micah B. Goldwater<sup>1</sup>** ([micahbg@gmail.com](mailto:micahbg@gmail.com)), **Scott Friedman<sup>2</sup>**, **Dedre Gentner<sup>1</sup>**, **Ken Forbus<sup>2</sup>**  
Department of Psychology<sup>1</sup>; Department of Electrical Engineering & Computer Science<sup>2</sup>, Northwestern University  
Evanston, IL 60208 USA

**Cynthia L. Fisher<sup>1</sup>** ([clfishe@cyrus.psych.illinois.edu](mailto:clfishe@cyrus.psych.illinois.edu)), **Michael Connor<sup>2</sup>**, **Dan Roth<sup>2</sup>**  
Department of Psychology; Department of Computer Science, University of Illinois at Urbana-Champaign  
Champaign, IL 61820 USA

**Franklin Chang** ([Franklin.Chang@liverpool.ac.uk](mailto:Franklin.Chang@liverpool.ac.uk))  
School of Psychology, University of Liverpool  
Liverpool, L69 7ZA UK

**Gary S. Dell** ([gdell@cyrus.psych.illinois.edu](mailto:gdell@cyrus.psych.illinois.edu))  
Department of Psychology, University of Illinois at Urbana-Champaign  
Champaign, IL 61820 USA

**Keywords:** computational models; language development; syntax; thematic roles

## Introduction

How children acquire the grammar of their native language has been a central topic in cognitive science since its outset, and has been the focus of much debate. One view assumed an innate Universal Grammar which genetically endowed the child with highly structured knowledge of language (Chomsky, 1965). An opposing position argued against both the assumptions of innate knowledge and structured representations, instead using connectionist architectures with distributed representations to learn grammatical patterns (e.g., Rumelhart & McClelland, 1986).

The field has progressed. There have been many years of rigorous empirical work, detailing the developmental pattern in children. In parallel, AI and cognitive science have made many advances in sophisticated learning algorithms. This symposium brings together models on the forefront of such empirical and computational research. Each model has roots in both sides of the early debate, positing (at least some) structured representations and specifying learning mechanisms. However, the models differ in many crucial ways. They use different computational architectures, learning algorithms, and differ in the knowledge built into the system. These differences in the models reflect and build on different current theories of grammatical development. The models focus on simulating empirical phenomena critical in distinguishing such theories. This symposium presents a unique opportunity to compare these new approaches and invite an open discussion.

## Symposium Structure

This symposium will present three computational models of grammatical development. The first talk, by Cindy Fisher, presents a model rooted in *early abstraction* theories of language development (e.g., Fisher, 2002). The model is implemented in a machine learning architecture that uses its

innate biases linking syntax and semantics and learned grammatical categories to semantically parse sentences and guide word learning. The second talk, by Franklin Chang will describe a connectionist model that does not require explicit thematic roles or innate linking rules to learn syntax, but instead can acquire language from visual-spatial input. The third talk, by Micah Goldwater, presents a third approach—a usage-based model that uses structured symbolic representations and learns abstract thematic roles via analogical abstraction

The symposium begins with a brief introduction, followed by three presentations of computational models, and concludes with a discussion exploring the issues.

**Dedre Gentner** will introduce the symposium. She is the Alice Gabrielle Twight Professor of Psychology and Education at Northwestern University.

**Gary S. Dell** will serve as the discussant. He is Professor of Psychology and Linguistics at University of Illinois at Urbana-Champaign.

We now summarize each talk in turn.

## The Origin of Syntactic Bootstrapping: A Computational Model

Syntactic bootstrapping proposes that children use knowledge of sentence structure in sentence interpretation and verb learning. We present a computational model of the origins of syntactic bootstrapping, based on systems for automatic semantic-role labeling (SRL). SRL models learn to identify sentence constituents that fill semantic roles, and to determine their roles, such as agent, patient, or goal. The present 'BabySRL' instantiates the structure-mapping account of syntactic bootstrapping (Fisher et al., 2010). We assume a structure-mapping process between the nouns in a sentence and the core semantic arguments of the verbs, in which children are biased to create one-to-one mappings. Given this one-to-one mapping bias, the number of nouns in the sentence becomes intrinsically meaningful to toddlers. Second, this account proposes that children's representations of sentences, though partially specified, are couched in

abstract terms, permitting generalization of new syntactic learning to new verbs. We used the BabySRL to investigate the consequences of these assumptions for learning from natural corpora of child-directed speech. The results yield strong evidence that partial sentence representations grounded in a set of nouns are useful as a foundation for further learning. We show (1) that such representations support new learning about English word order (e.g., the first of two nouns is typically an agent), and (2) permit children to learn which words are verbs by tracking their argument-taking behavior in sentences.

**Cynthia L. Fisher** is Professor of Psychology and Linguistics at University of Illinois at Urbana-Champaign.

### The Dual-Path Model

A theory of language should unite acquisition, production, and comprehension. To link acquisition to production, Chang (2002) proposed a connectionist model called the Dual-path model. The model used the same learning mechanism for acquiring abstract English syntactic representations to explain structural priming in adult production (Chang, et al., 2006). The model could learn English and Japanese to similar levels (Chang, 2009). Importantly, the model's incremental planning mechanism could also account for the different direction of heavy NP shift in each language. That is, English speakers tend to postpone long noun phrases (NPs) to the end of sentences, while Japanese speakers tend to place these NPs earlier. Finally, a new version of the model has been developed where thematic roles in the message are replaced with arbitrary spatial pointers. The production or comprehension of words depends on the activation of these pointers, and hence they can simulate eye-tracking in scenes. The model can account for different types of anticipation in English and Japanese during eye-tracking in the visual world (e.g., Kamide et al., 2003). The model highlights the role of non-linguistic processes such as implicit learning and spatial binding in language acquisition and use.

**Franklin Chang** is Lecturer of Psychology at the University of Liverpool.

### An Analogical Learning Model of the Development of Thematic Roles & Structural Priming

This model explores the hypothesis that analogical learning processes can account for the abstraction of thematic roles (Goldwater et al., 2011). We use SME (Falkenhainer et al., 1989) and SAGE (Kuehne et al., 2000) to model analogical mapping and generalization. Learning proceeds via incremental comparison of specific examples to reveal their common structure and create generalizations. This allows the model to evolve abstract roles from verb-specific ones. We assess the abstractness of the model's semantic roles by simulating structural priming in sentence production. To construct an utterance for a new event, it first uses analogical retrieval to find utterances with similar semantic structure in memory. It then creates an analogical mapping between events and transfers the previous sentence structure to construct the new utterance. Hence, structural priming is

shown when a new utterance has analogous structure to a prime utterance.

We assume a usage-based learning trajectory in which children's initial semantic representations are verb-specific (e.g., Tomasello, 2003). Because shared semantic structure is necessary to show priming, early in training priming occurs only across sentences that share verbs. As the model learns, across-verb priming occurs. For example, initially the model can align *giving* events only with other *giving* events, because the roles do not match those of other verbs. By gradual re-representation and generalization, the model develops a hierarchy of more abstract roles.

**Micah B. Goldwater** is a postdoctoral fellow in the Department of Psychology at Northwestern University.

### References

- Chang, F. (2002) Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, 26, 609-651
- Chang, F. (2009) Learning to order words: A connectionist model of heavy NP shift and accessibility effects in Japanese and English. *Journal of Memory and Language*, 61, 374-397
- Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, 113, 2, 234-272
- Chomsky, N. (1965), *Aspects of the Theory of Syntax*, Cambridge, MA: MIT Press
- Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, 41, 1-63
- Fisher, C. (2002). The role of abstract syntactic knowledge in language acquisition: A reply to Tomasello (2000). *Cognition*, 82, 259-278.
- Fisher, C., Gertner, Y., Scott, R., & Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1, 143-149.
- Goldwater, M. B., Friedman, S. E., Gentner, D. G., Forbus, K. F., Taylor, J. L M. (2011). An analogical learning model of the development of thematic roles & structural priming. *Presented at the 36th Annual Boston University Conference on Language Development*.
- Kamide, Y., Altmann, G.T.M., & Haywood, S. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye-movements. *Journal of Memory and Language*. 49, 133-159.
- Kuehne, S. E., Gentner, D. & Forbus, K. D. (2000). Modeling infant learning via symbolic structural alignment. *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*, 286-291
- Rumelhart, D., & McClelland, J., (1986). On Learning the Past Tenses of English Verbs. in McClelland and Rumelhart (eds). *Parallel Distributed Processing*, 216-271
- Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press

# Computational, Cognitive, and Neural Models of Decision-making Biases

**Jonathan Malmaud (malmaud@mit.edu), Joshua B. Tenenbaum (jbt@mit.edu)**

Brain and Cognitive Sciences, 43 Vassar St.  
Cambridge, MA 02139 USA

**Peter Dayan (dayan@gatsby.ucl.ac.uk) Laurence T. Maloney (laurence.maloney@nyu.edu)**

Gatsby Computational Neuroscience Unit  
17 Queen Square  
London, WC1N 3AR England

Dept. of Psychology, 6 Washington Place  
New York, NY 10030 USA

**Edward Vul (evul@ucsd.edu)**

Dept. of Psychology, 9500 Gilman Dr. 0109  
La Jolla, CA 92093 USA

**Nick Chater (n.chater@ucl.ac.uk)**

Division of Psychology and Language Sciences  
26 Bedford Way  
London, WC1H 0AP England

**Keywords:** Computational modeling; Rational analysis; Judgment and decision-making; Biases and heuristics

## Summary

The question for the symposium is how best to understand biases in decision-making, going beyond traditional judgment and decision-making (JDM) accounts such as prospect theory to take a more modern reverse-engineering perspective bridging rational computational, algorithmic, and neural levels of explanation, and viewing decision-making under risk and uncertainty not just as a simple matter of evaluating lotteries but in the context of cognition more broadly, taking seriously learning, perception, motor control, memory, and action planning.

The dominant normative approach to studying decision-making under risk is axiomatic expected utility theory, which argues that any agent obeying seemingly reasonable axioms of choice consistency can be modeled as maximizing the expected utility of its decisions. From decades of research that analyzes people's choices between simple gambles in the lab, it is known that humans routinely violate these axioms. This has forced decision theorists to adopt descriptive models of choice that lack a normative rational in order to account for observed patterns of choice, the most prominent of which is Kahneman's and Tversky's prospect theory for one-shot decisions under risk with immediate outcomes and hyperbolic discounting for decisions involving delayed outcomes.

There are several challenges not addressed by prospect theory and its variants. First, they are silent on the issue of the cognitive mechanisms that are actually responsible for human choice behavior. Second, they do not seem as a practical matter to scale to real-world decision problems, where the space of possible outcomes and actions is not sharply defined, the effects of actions are highly uncertain, and the explicit calculation of expected values is impractical. Third, they do not strongly constrain or give an underlying rationale for the probability weighting function, temporal discounting function, or utility function featured in prospect theory. Thus these models cannot explain why these functions' estimated

forms and parameters seem to be greatly affected by seemingly irrelevant factors of the task framing and setup, such as whether the outcome probabilities are presented numerically in tables or learned through experience and why the evaluation of individual gambles seems to be highly effected by the properties of other gambles in the choice set. More broadly, these theories fail to explain why in day-to-day life human decision-making seems to generally be highly robust and effective while sharply contrasting with normative predictions in the simple, stylized decision tasks commonly used in JDM experiments.

This symposium brings together researchers who represent a variety of perspectives on ways cognitive science can inform our understanding of decision biases to address these challenges, with relevance at all three Marr levels of analysis. Malmaud and Tenenbaum, and Dayan both offer computational-level Bayesian accounts that explain decision-making biases as resulting from reasoning with priors that are adapted for real-world or evolutionary-relevant decision tasks. Malmaud and Tenenbaum explain choices in terms of advanced models from the AI planning literature and animal foraging theory. Dayan offers a neurobiological implementation of inference that spans the Marr levels.

Other approaches relate to algorithms levels of the Marr hierarchy with links to lower and higher levels. Vul offers an algorithmic description of biases as resulting from cognitive limitations associated with reasoning using only a limited number of samples from a posterior over decision parameters. Maloney and Chater link high-level decision-making to known properties of perception and cognition, such as scale-invariance. Maloney gives a unifying account of the probability weighting function as arising from the same principles as perception of continuous quantities in psychophysics. Chater explores the origin of subjective utility and temporal discounting through connections to broader cognitive processes.

One general idea that cuts across all these approaches is that human decision-making can be modeled in a unified way as the result of general cognitive principles that offer prin-

ciplined explanatory accounts of biases in decision-making, rather than via a series of descriptive utility-maximizing models that have undergone ad hoc adjustments to account for a mélange of deviations from a narrow normative standard.

### **Malmaud and Tenenbaum: Prospect theory as rational response**

We will open the symposium by presenting a brief review of the traditions of axiomatic decision theory and descriptive prospect theory, including how sophisticated computational models are beginning to fill in some of prospect theory's known shortcomings as a model of high-level decision-making at the individual level. We will briefly discuss our position that human decision-making is adapted for solving rich, sequential decision problems with structured goals and highly uncertain action-outcome contingencies and as such should not be expected to perform optimally according to narrow normative standards in simple one-shot decision tasks with known contingencies. We will show how modeling human choices as the result of employing state-of-the-art AI methods for planning under uncertainty to a specific class of 'survival' goals commonly studied in the animal foraging literature naturally implies a sequential decision strategy that is compatible with the descriptive predictions of prospect theory. Our approach is also able to make predictions about human behavior for a wide class of tasks for which prospect theory is not applicable. We will present preliminary empirical evidence that these predictions are supported on a specific set of ecologically relevant sequential decision tasks.

### **Dayan: Pavlovian choice illusions**

One useful interpretation of many perceptual illusions is in terms of biases resulting when the mechanisms of inference reflect genetically encoded or learned priors that are inconsistent with a given scene. We will consider how some of the biases of decision-making on which this symposium focuses can be seen as illusions of choice arising from forms of Pavlovian influences that reflect evolutionarily appropriate dispositions. These influences are exerted by various neural systems, notably the neuromodulators dopamine and serotonin and regions of the amygdala, acting on areas such as the striatum that are involved in decision-making. As an example, we will discuss the case of behavioral inhibition, which is one very general response to potential threats, and is closely associated with serotonin. We will show how such inhibition can lead to a particular form of bias in the on-line evaluation of complex options, and show how problems with this bias might even have deleterious psychiatric consequences.

### **Maloney: Ubiquitous log odds**

Similar patterns of distortion are found in visual frequency estimation, frequency estimation based on memory, and in the use of probability in decision-making under risk. Based on joint work with Zhang et al., I will show that probability distortions in all cases (so far) can be approximated by a lin-

ear transformation of the log-odds of probability or relative frequency. The slope and intercept of the linear transformation control probability distortion. Researchers have not been able to predict or explain the values of slope and intercept observed in experiments across tasks or across participants.

In Zhang & Maloney (2012) we focused on one method for presenting probability, the relative frequency of items of one kind in a visual array of  $N$  items. We developed a model of human distortion of relative frequency and demonstrated in two experiments that we can separately control slope and intercept with high accuracy. Our results support the conjecture that probability is systematically adapted to particular tasks much as perceptual information concerning lightness or loudness is transformed. We shown how a simple model based on chunking of information can explain the results we observe with a high degree of precision.

### **Vul: Decision biases and heuristics arising from inference by sampling**

Across many domains, people integrate sophisticated world knowledge with prior expectations nearly optimally, yet when making conscious cognitive judgments, they seem to be grossly irrational. I will explore a potential explanation: that conscious cognitive judgments reflect sample-based approximate inference under constrained cognitive resources. Experiments measuring multiple judgments from individuals with no new information yield evidence for this sampling proposal: any one decision appears to reflect only a small fraction of the information the participant has available, suggesting that each decision is based on only a small number of samples. Here, I will talk about the tradeoffs inherent in using a small number of samples for a decision: why we might want to use few samples, the consequences of using a few samples for judgments, the risks associated with using a few samples when rewards are asymmetric, and how these consequences relate to biases seen in judgment and decision-making.

### **Chater: From cognitive principles to JDM**

This talk will consider how far candidate cognitive principles (such as scale-invariance, relative coding of magnitudes, and incommensurability between distinct dimensions) can provide quantitative and qualitative explanations of results in decision-making. I will illustrate how widespread patterns in JDM (such as constant relative risk-aversion and hyperbolic time-discounting) can be derived; and consider how basic cognitive processes can explain when and in what way, such regularities break down. The aim is to build a theory of JDM built on cognitive principles, rather than rational axiomatic foundations.

# **Governing Board Symposium**

## **Cognitive Science and the Learning Sciences**

### **Participants**

**Kurt VanLehn (Kurt.Vanlehn@asu.edu)**

School of Computing  
Informatics and Decision Science Engineering (CIDSE)  
Arizona State University, Tempe, AZ, USA

**Roy Pea (roypea@stanford.edu)**

School of Education  
Stanford University, Stanford, CA, USA

**Naomi Miyake (nmiyake@p.u-tokyo.ac.jp)**

Graduate School of Education  
The University of Tokyo, Tokyo, Japan

**Reinhard Pekrun (pekrun@lmu.de)**

Faculty of Psychology and Educational Sciences  
Ludwig Maximilians University, Munich, Germany

### **Chairs**

**Richard Catrambone (rc7@prism.gatech.edu)**

School of Psychology, Georgia Institute of Technology  
Atlanta, Georgia, USA

**Stella Vosniadou (svosniad@phs.uoa.gr)**

Department of Philosophy and History of Science  
National and Kapodistrian University of Athens  
Athens, Greece

### **Discussant**

**Stella Vosniadou**

**Keywords:** Learning science, learning tools, constructive interaction, coordinated learning, test anxiety, tutoring system efficacy.

### **Introduction**

The focus of the symposium is on real world implementations of educational innovations based on cognitive and learning science principles and research. These real world implementations can be in physical classrooms, on-line courses, informal educational settings, as well as other learning environments. The innovations can include new ways of conceptualizing and presenting a domain, computer-based multimedia learning tools, and other innovations. The common thread though is that these innovations are beyond lab-testing and are guided by principles and research from the cognitive and learning sciences. The governing board symposium will bring to the conference educational innovation found in different parts of the world (US, Asia, Europe) from distinguished researchers representing a variety of theoretical orientations and focusing on different aspects of the learning process (e.g., cognitive, social, emotional/ motivational).

### **Can intelligent tutoring systems become even more effective than human tutors?**

**Kurt VanLehn**

This talk will start by reviewing reasons why human tutoring should be more effective than computer tutoring. Studies indicate that human tutors do not actually use some of the techniques that they are assumed to use. Moreover, the techniques that they do use are also used by step-based tutoring systems, which are a type of intelligent tutoring system. Thus, it comes as no surprise that step-based tutoring systems and human tutoring are equally effective, as shown in a meta-analysis of content-controlled experiments. This raises the question: what if step-based tutoring systems started using some of the techniques that human tutors

were supposed to use? Would they even become more effective than human tutors?

### **Social foundations of coordinated learning across environments**

**Roy Pea**

A persistent challenge in the learning sciences is accounting for coordinated learning across the socio-cultural environments in which people participate. K-12 aged children have been a special focus of these inquiries, given the preponderance of their awake time for learning outside of school, the recalcitrant problems of transfer of school learning to life, the underuse of funds of knowledge children have from life in school learning, and persistent achievement gaps. Contemporary accounts of K-12 learning over environments, while still attentive to cognitive issues of learning and reasoning in the disciplines, have been making substantive progress on the coordinated learning challenge in their attention to associated learner developments in identity, interests, social networks (and affiliated social learning capital), and examining social learning mechanisms such as imitation, joint visual attention, formative feedback, positioning in discourse, and accountable reasoning and talk in communities of practice. Highlights of recent work on these issues are also imbued with significance for socio-technical design of engaging learning environments that can mediate learning using new social media and mobile technologies. Our NSF-funded LIFE Center (Learning in Informal and Formal Environments) has been pursuing these issues as it seeks to develop and test principles about the social foundations of human learning from infancy to adulthood. Select findings will illustrate these developments towards understanding and designing connected human learning.

## **Bridging cognitive and learning sciences by engineering constructive interaction in Asian classrooms**

**Naomi Miyake**

Real-world learning situations provide us with test fields for our cognitive science theories of how people learn. In this presentation, I report a case where a fundamental framework about how people constructively interact to learn could guide some policy making and practices in classrooms, which could influence the course of change in Japanese school education. The framework is named “constructive interaction,” (Miyake, 1986) which states that two person, when engaged in solving a shared problem, exchange roles of a task-doer who proposes possibilities for solutions and a monitor who reflects upon such proposals, and such role exchange potentially promote each participating individual’s understanding of the problem.

Though group work has been common in Japanese classrooms, such practice has not been guided nor assessed via lenses of cognitive and learning sciences. In the pursuit of acquiring the 21<sup>st</sup> century skills, current classrooms have been trying to shift their practice from teacher-centric, fact-oriented training to learner-centric, knowledge-building learning. In such classes the learners’ activities are often socially interactive, or collaborative. There are many different ways to make a classroom collaborative, sometimes with confusion about which leads to which outcome. In my recent research in promoting collaborative classrooms based on the above framework, I have identified three research questions related to such confusion, created a testable classroom design to answer the questions. The three questions are to confirm that (1) outcomes of constructive interaction are individualistic, not easily shared by other members of the same group (or class), (2) a learner who mostly listens and monitors can still learn as much as more active learners, and (3) for a constructive interaction to lead productive learning, there is no need to socially organize the group, but it is essential for the members to share the desire to solve an apparently shared problem, or understand it. During 2010 and 2011, one hundred and four teachers from elementary to high school devised and delivered such classes in major subject areas, which resulted in higher performance than regular classes, with higher motivation to learn more after the class (<http://coref.u-tokyo.ac.jp/en>). The findings so far show that the answers to the above three questions are positive, as predicted by the basic framework of constructive interaction, making it possible to create design principles for designing more productive collaborative classes around cognitive science frameworks. It has also been shown that this type of cognitive-science-based design principles could guide real learning in real classrooms, and when some basic cognitive science is shared by the practitioners, the outcomes of such classrooms can lead them to develop better practices on their own.

## **Emotions are important for students’ learning and achievement** **Reinhard Pekrun**

Emotions are ubiquitous in academic settings. Students frequently experience emotions such as enjoyment, hope,

pride, anger, anxiety, shame, hopelessness, and boredom in these settings. Moreover, these emotions are likely to influence students’ learning, achievement, and health. Traditionally, they have not received much attention by empirical research, test anxiety studies and attributional research being notable exceptions. During the past ten years, however, there has been growing recognition that emotions are central to students’ learning. In this presentation, I will address the functional relevance of emotions for student learning. Subsequently, I will discuss the origins of these emotions and related educational intervention aiming to promote adaptive emotions that facilitate academic learning. Pekrun’s (2006) control-value theory of achievement emotions will be used as a conceptual framework.

Test anxiety research has shown that anxiety can exert profound effects on academic performance; is this true for other emotions as well? I will discuss five cognitive and motivational mechanisms that can mediate effects on learning: (1) availability of working memory resources; (2) long-term storage of information in terms of retrieval-induced forgetting and facilitation; (3) intrinsic and extrinsic motivation to learn; (4) use of learning strategies; and (5) self-versus external regulation of learning. As a consequence of effects on these processes, emotions can profoundly influence students’ competence development. I will present experimental evidence and findings from two longitudinal studies on upper elementary and university students’ emotions documenting these effects.

Given that students’ emotions are functionally important, their origins and related educational tools to modify these emotions should be considered. Using the control-value theory, I will argue that appraisals of control over achievement activities and outcomes, and of the value of these activities and outcomes, are fundamentally important for emotion arousal in academic settings. By implication, teachers, tasks, and learning environments influence students’ emotions by shaping their perceived control and values, and ways to influence these emotions can be developed by considering these appraisals. One especially important variable shaping students’ appraisals and emotions likely is the cognitive quality of tasks. I will present exemplary evidence from an intervention study which examined the impact of cognitively activating tasks involving mental modeling on students’ emotions in mathematics. The findings suggest that it is possible to promote students’ appraisals and adaptive emotions by shaping tasks and learning environments in cognitively and emotionally activating ways.

## **References**

- Miyake, N. (1986). Constructive interaction and the iterative processes of understanding, *Cognitive Science*, 10(2), 151-177.
- Pekrun, R. (2006). The Control-Value Theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18(4), 315-341.



# How Vertical Spaces Are Perceived and Represented

**Daniele Nardi (d.nardi@uniroma1.it)**

Department of Psychology, Sapienza University of Rome, Via dei Marsi 78  
Rome, 00185 Italy

**Frank H. Durgin (fdurgin1@swarthmore.edu)**

Department of Psychology, Swarthmore College, 500 College Avenue  
Swarthmore, PA, 19081 USA

**Kate J. Jeffery (k.jeffery@ucl.ac.uk)**

Cognitive, Perceptual and Brain Sciences, University College London, 26 Bedford Way  
London, WC1H 0AP, UK

**Steven M. Weisberg (smweis@temple.edu)**

Department of Psychology, Temple University, 1701 N. 13<sup>th</sup> street  
Philadelphia, PA, 19122 USA

**Keywords:** Vertical spaces; volumes; slope or slant; angular scale expansion; spatial abilities; individual and sex differences; place and grid cells.

## Introduction

The purpose of this symposium is to highlight how the vertical dimension is perceived and represented differently from the horizontal dimensions, and the role of this dimension in spatial learning. This is a new and important issue because literature on spatial cognition has hitherto neglected the study of the vertical dimension, under the assumption that space can identically be investigated in the horizontal plane. This notion, while untested, ignores a crucial, unique property of the vertical dimension – that of being parallel to the force of gravity, which poses constraints on affordances and energetic potential. Additionally, the ability to move freely in three dimensions imposes computational complexities not present in two. New research impetus is trying to clarify the role of the vertical dimension in space. The present symposium will try to tie together different perspectives (psychophysics, cognition, neurophysiology), using different animal models (human and non-human), and different experimental methods (real and virtual environments), in order to provide a synthetic view on this issue, and to establish future goals of common interest. We focus here on two aspects of three-dimensional space: surface properties (e.g., hills and valleys) and volumetric properties.

## Surface Properties

A major constraint to terrestrial movement is represented by the inclination of the terrain. While a moderate hill can be a challenge to walk, a steep one can be dangerous and energetically depleting. Therefore, it is ecologically adaptive for our perceptual system to be extremely sensitive in estimating geographical slants. Recent research by Durgin's lab has indeed shown that the coding of slant, and

other angular variables, is exaggerated by the introduction of perceptual biases, such that the range of inclinations relevant to locomotion is more densely coded (Durgin & Li, 2011). This angular scale expansion enables us to more precisely represent small differences in inclinations, and thus may improve the precision of action control. Furthermore, the theory quantitatively predicts the well-reported overestimation of hill slant, suggesting that this phenomenon may not be directly due to a role of effort or physical potential, but to a more general angular coding scheme that is useful for action control as well as route planning and spatial orientation.

Beyond perceptual encoding of slope for action, recent research has demonstrated large individual differences in the ability to rely on slope cues for cognitive spatial tasks, including navigation (Weisberg & Newcombe, 2011). Given that slope is a stable and salient part of the lay of the land, a fundamental question explored in a line of research by Weisberg is whether people are able to take slope into account when building a mental map of the environment – and what makes slope difficult to use for some people. Using a navigation task in a virtual environment, it was found that individual's navigation ability is a crucial factor in determining whether terrain slope facilitated a more accurate representation: in a complex environment, only good navigators were able to take advantage of the information. This result is very important because it can explain previously reported sex-specific difficulties with slope in light of individual differences in broad spatial abilities.

Given the theoretical distinctiveness of the vertical component of space, due to its link with the gravity axis, an important question to investigate is the salience of vertical information in spatial learning tasks. Studies on non-human animal models have indicated that vertical information dominates over horizontal information. Recently, a line of research by Nardi has investigated if the same occurs in humans. Reorientation was tested in a real-world

environment with two different strategies available: one based on directional cues (the slope gradient of the tilted floor) and one based on positional cues (landmarks). Interestingly, slope information did not dominate the reorientation process, as people were equally likely to rely upon either cue. Furthermore, men and women did not significantly differ in their reliance on slope or landmarks, suggesting that in a real environment there is not a sexually dimorphic preference for spatial strategies. However, men exhibited greater overall confidence in solving the task. These findings suggest that the female disadvantage with slope cues, shown in Nardi, Newcombe, and Shipley (2011), could be due – at least partially – to a general difficulty in reorienting, namely lower spatial confidence.

### **Volumetric Properties**

Our evolutionary ancestors, being aquatic, moved freely in all three dimensions, and many animals still do. It is therefore likely that the brain has evolved a method for representing volumetric space in a cognitive map, but the properties of this map are not yet understood. Studies of the representation of space at the single neuron level are providing clues, however. In ordinary, horizontal environments, place cells in the hippocampus encode location while grid cells, in the neighbouring entorhinal cortex, encode an integrated signal of distance and direction. Jeffery's lab has studied for the first time how place and grid cells respond to travel in the vertical dimension. Using rodents as an animal model, it was found that while grid cells are highly insensitive to vertical distances, place cells do show some responsiveness, though at a coarser scale than for horizontal distances (Hayman, Verriotis, Jovalekic, Fenton & Jeffery, 2011). The findings suggest that the representation of vertical space, or perhaps space in the dimension normal to the body plane of the animal, is represented differently, and maybe non-metrically. Preliminary behavioral studies support this notion, finding that rats and mice can encode goal locations in three dimensions but prefer to organize their search behavior in horizontal bands.

From the above findings, it seems that the so-called “cognitive map” of space may perhaps not be uniform in all dimensions, despite our subjective experience to the contrary, a finding that has implications for those navigating in volumetric spaces (astronauts, pilots, deep sea divers, virtual reality explorers etc).

### **References**

Durgin, F. H., & Li, Z. (2011). Perceptual scale expansion: An efficient angular coding strategy for locomotor space. *Attention, Perception & Psychophysics*, 73, 1856-1870. doi: 10.3758/s13414-011-0143-5

Hayman R, Verriotis M, Jovalekic A, Fenton A and Jeffery KJ (2011) Anisotropic encoding of three-dimensional space by place cells and grid cells. *Nature Neuroscience*, 14(9):1182-8.

Nardi, D., Newcombe, N. S., & Shipley, T. F. (2011). The world is not flat: Can people reorient using slope? *Journal of Experimental Psychology: Learning, Memory and Cognition*, 37, 354-367.

Weisberg, S.M. & Newcombe, N. S. The Role of Slope as a Navigational Cue. *Poster presented at the 4<sup>th</sup> annual inter-Science of Learning Center Conference (iSLC)*, Washington, D.C., 3/2011

# What Can Cognitive Science Say or Learn about Economic Crises?

**\*Magda Osman (m.osman@qmul.ac.uk)**

Experimental Biology and Psychology Centre, Queen Mary University London, London, E14NS, UK

**\*Björn Meder (meder@mpib-berlin.mpg.de)**

**Gerd Gigerenzer (gigerenzer@mpib-berlin.mpg.de)**

Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, 14195 Berlin, Germany

**Nick Chater (nick.chater@wbs.ac.uk)**

**Daniel Read (daniel.read@wbs.ac.uk)**

Warwick Business School, University of Warwick, Coventry, CV4 7AL, UK

**Hansjörg Neth (hneth@uni-goettingen.de)**

Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany

**Keywords:** Cognitive Science, Economics, Decision making, Uncertainty, Behavioral Economics, Rationality, Heuristics

## The issue

Economic crises bring to the fore deep issues for the economic profession: why are such crises often not foreseen, and what does this entail for economic theory? In this symposium we also adopt a self-critical analysis, by asking the following: what can the cognitive science community say or learn about cognition and behavior in the context of economic crises? After all, cognitive science shares one of its principle objectives with economics: to investigate and model the principles that underlie and govern human behavior.

## Challenges

The current financial crisis presents us with a real-world example of decision making under uncertainty. Cognitive science offers a variety of theories and models, from probabilistic models of cognition (Chater & Oaksford, 2008) to heuristic approaches (Gigerenzer & Gaissmaier, 2011), each designed to describe decision making under uncertainty. Empirically, the extant methods used to examine this question in both economics and psychology involve simple choice tasks (e.g., lotteries and games with well-defined probabilities and outcomes). *But, are the models sufficient to accurately represent uncertainty, and are the tools adequate for the job of capturing decision making under uncertainty?*

Uncertainty can permeate all aspects of a decision problem, from constructing the action space, to inferring probabilities of outcomes and the behavior of other agents in the situation. For instance, politicians need to decide whether to bail out fragile banks and countries under time pressure, with incomplete information about the problem space, and the necessity to manage conflicting goals (e.g., also considering the needs of their won electorate). Turning situations of this kind into lottery type tasks may in fact be a way of translating the unmanageable (uncertainty) into something manageable (risk), but at the same time the evidence may be giving answers to the wrong kind of questions.

Additionally, there is an issue of scalability. Neoclassical economics assumes that macro-level behavior can be deduced from modeling agents as rational, utility-maximizing individuals. While this oversimplification is often recognized by economists, scaling up to the aggregate level is a necessity when having to inform policy decision. The crucial challenge in revising the microfoundations of economic behavior is how we can build more realistic models, which nevertheless can be scaled up to the aggregate level.

## Goals of the Symposium

The symposium is themed around the target questions: *What can our community say or learn about cognition and behavior in economic crises?*

For instance, could rational or heuristic models help predicting or preventing crises? Or could the psychology of crowds help to explain economic crises? By bringing together researchers with different research perspectives and methodologies, the key objective is to discuss the challenges that real-world problems such as economic crises present us with, and ways in which cognitive science could possibly inform economic theory and policy making. The symposium will consist of a general introduction (Osman, Meder), four talks (Chater, Gigerenzer, Neth, Read) and a discussion (Meder, Osman) involving all participants.

### **Nick Chater**

Chater's work has explored the fundamental principles of cognition, in particular in contexts in which the cognitive system is faced with uncertain inferences (e.g., learning, decision making, reasoning, perception). Recently, his work also concerns applications to policy making.

Vlaev, I., Kusev, P., Chater, N., Stewart, N., & Aldrovandi, S. (2010). Domain effects and financial risk attitudes. *Risk Analysis*, 30, 1374–1386.

Chater, N., & Oaksford, M. (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford: OUP.

### **Gerd Gigerenzer**

Gigerenzer's core research approach has been to understand decision making from the perspective of bounded rationality. This includes heuristic decision making and the development of effective tools for risk communication, with the goal of helping people to make better decisions in an uncertain world.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62, 451–482.

Gigerenzer, G., & Selten, R. (Eds.). (2001). *Bounded rationality: The adaptive toolbox*. Cambridge, MA: MIT Press.

### **Björn Meder**

Much of Meder's work has been concerned with the connections between causal induction and decision making. His recent research focuses on information search, economic psychology, and alternative frameworks of rationality.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review*, 15, 75–80.

Meder, B., Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, 7, 119–148.

### **Hansjörg Neth**

Do people allocate their resources (time, information processing effort) adaptively when facing tasks that vary in their demands and complexity? Neth's empirical work has examined task switching behavior and simple satisficing strategies in cognitive foraging tasks, consumer choice, and financial decision making.

Neth, H., Khemlani, S. S., & Gray, W. D. (2008). Feedback design for the control of a dynamic multitasking system: Dissociating outcome feedback from control feedback. *Human Factors*, 50, 643–651.

Payne, S. J., Duggan, G. B., & Neth, H. (2007). Discretionary task interleaving: Heuristics for time allocation in cognitive foraging. *Journal of Experimental Psychology: General*, 136, 370–388.

### **Magda Osman**

Osman's work explores dynamic decision making and shows that people are sensitive to underlying differences in the stability of the environment when tasked with controlling uncertainty in micro-world dynamic environments.

Osman, M. (2010). Controlling uncertainty: A review of human behavior in complex dynamic environments. *Psychological Bulletin*, 136, 65–86.

Osman, M., & Speekenbrink, M. (2011). Information sampling and strategy development in complex dynamic control environments. *Cognitive Systems Research*, 12, 355–364.

### **Daniel Read**

Within the domain of judgment and decision making, Read has studied a variety of behaviors including seeking (how consumers choose to diversify consumption), intertemporal choice (how people trade off current and future consumption), and decision making under risk.

Read, D. (2007). Time and the marketplace. *Marketing Theory*, 7, 59–74.

Read, D. (2007). Utility theory from Jeremy Bentham to Daniel Kahneman. *Thinking and Reasoning*, 13, 45–61.

# Dynamic decision making: neuronal, computational, and cognitive underpinnings

**Peter Dayan** (dayan@gatsby.ucl.ac.uk)

Gatsby computational neuroscience unit, University College London

**Nigel Harvey** (n.harvey@ucl.ac.uk), **Maarten Speekenbrink\*** (m.speekenbrink@ucl.ac.uk)

Cognitive, Perceptual and Brain Sciences, University College London

**Magda Osman\*** (m.osman@qmul.ac.uk)

Experimental Biology and Psychology Centre, Queen Mary University

**Masataka Watanabe** (watanabe-ms@igakuken.or.jp)

Department of Physiological Psychology, Tokyo Metropolitan Institute of Medical Science

**Keywords:** Dynamic Decision Making, Computational models of reinforcement learning, Animal learning, Applied decision making

## Challenging Issues

As complexity in our everyday environment increases (e.g., mobile applications for monitoring energy consumption), how do we adapt and react to the changing demands placed on us? In dynamic decision making (DDM) problems, the environment changes over time due to previous decisions made and/or factors outside the control of the decision-maker. To maximize his/her reward, an agent effectively needs to control a complex dynamic system. This often involves planning in the face of uncertainty about how decisions change the state of the system and the rewards that can be obtained. Thus, DDM refers to a process by which an agent selects a course of action in a manner that achieves or maintains a desired state in a dynamic environment. This includes balancing exploration and exploitation, distinguishing between different sources of variability within the environment, and tracking the current state of the environment (i.e., filtering).

Thus far there has been little attempt at a synthesis of the amassing research from different areas of cognitive science directed towards understanding DDM. The objective of this symposium is to bring together the latest theoretical approaches and empirical research investigating DDM. The speakers range in expertise from comparative (Prof Watanabe), applied (Prof Harvey) and cognitive psychology

(Dr Osman), computational neuroscience (Prof Dayan), and computational learning theory (Dr Speekenbrink). By bringing these diverse approaches together, the aim is to generate discussion around the following critical question: *What are the processes/mechanisms that enable us to adapt to changes in uncertain environments in terms of the information we process, the decisions we make, and the intrinsic and extrinsic goals that we pursue?* The symposium will consist of a general introduction (Osman), three talks (Dayan, Harvey, Watanabe) and an extended discussion (moderated by Speekenbrink) involving all participants.

## Peter Dayan

Peter Dayan's work in computational and experimental neuroscience has contributed significantly to our understanding of the neural mechanisms underlying DDM and the learning of reward structures therein. Dayan is an expert on reinforcement learning and in recent work, has elucidated the distinction between "model-based" and "model-free" learning and the neural circuits supporting these.

Model-based learning, usually associated with declarative task-knowledge, can support complex planning. Model-free learning, due to its more procedural nature, supports quick and habitual decisions, but will not cope well in an environment that undergoes rapid, abrupt changes. Dayan's recent work has shown how both processes work concurrently to support effective DDM.

Dayan, P. (2009). Goal-directed control and its antipodes. *Neural Networks*, 22, 213-219.

Dayan P., & Daw, N.D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective & Behavioral Neuroscience*, 8, 429-453.

### **Nigel Harvey**

How effective are judgments, and what role do they play in using evidence to plan actions in complex decision making environments? Nigel Harvey's work in cognitive and applied decision making has developed cognitive models of judgments and decisions, and the confidence placed therein, in a variety of domains including economic (e.g., consumer choice behavior), financial and medical (e.g., comparisons of clinical and actuarial judgment). More recently, Harvey has shown that in a variety of decision making situations people decide whether to focus their efforts on acquiring new information from feedback, or whether to implement their extant knowledge.

Harvey, N. (2011). Learning judgment and decision making from feedback: An exploration-exploitation trade-off? In M.K. Dhami, A. Schlottmann and M. Waldmann (Eds.) *Judgment and decision making as a skill: Learning, development, and evolution*. Cambridge: Cambridge University Press.

Reimers, S., Harvey, N. (2011). Sensitivity to autocorrelation in judgmental time series forecasting. *International Journal of Forecasting*, 27, 1196-1214.

### **Magda Osman**

Magda Osman's recent work has advanced the proposal that successful learning in DDM environments can be achieved indirectly via prediction or directly via control processes. In two reviews of DDM, Osman brings together theoretical and empirical research from disparate disciplines spanning engineering, machine learning, management, social and cognitive psychology, and neuroscience, and shows that each has contributed to answering the question: *How do we isolate the effects of our actions from those generated independently in order to achieve desirable outcomes?*

Osman, M. (2010). Controlling uncertainty: A review of human behavior in complex dynamic environments. *Psychological Bulletin*, 136, 65-86.

Osman, M. (2011). The role of feedback in decision making. In *Parkinson's Disease* (Chapter 3), InTech Publishers.

### **Maarten Speekenbrink**

In DDM environments, the potential consequences of actions can change over time, either due to previous actions or external factors. How do people adapt their representations of a task to such abrupt or gradual changes? In contrast with popular beliefs, the findings from Speekenbrink's research suggest that people are generally able to rapidly adapt their predictions to different types of changes in multiple cue environments.

Speekenbrink, M., & Shanks, D.R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General*, 139, 266-298.

Speekenbrink, M., & Shanks, D.R. (2008). Through the looking glass: A dynamic lens model approach to MCPL. In: Chater, N., & Oaksford, M. (Eds.). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford: Oxford University Press. (pp. 409-429).

### **Masataka Watanabe**

Adaptive goal-directed behaviours can be acquired by neuronal mechanisms that can learn and anticipate the possible outcomes of actions, and determine the actions that might be successful for achieving desirable outcomes. Having close anatomical connections with high-order cortical and subcortical limbic areas, the prefrontal cortex (PFC) play the most important role in this process. In several primate studies, Masataka Watanabe has shown that neurons in the lateral PFC integrate cognitive (outcome probabilities) and motivational (rewards) representations that enable adaptive decision making in complex circumstances.

Watanabe, M. (2007). Role of anticipated reward in cognitive behavioral control. *Current Opinion in Neurobiology*, 17, 213-279.

Watanabe, M. (2009) Role of the primate lateral prefrontal cortex in integrating decision-making and motivational information. In: J. Dreher, & L. Tremblay (Eds.). *Handbook of Reward and Decision Making*. Oxford: Academic Press, pp. 79-96.

# Grammatical Approaches to Written and Graphical Communication

**Colin Wilson (wilson@cogsci.jhu.edu)**

Department of Cognitive Science  
Johns Hopkins University

**Neil Cohn (neil.cohn@tufts.edu)**

Department of Psychology  
Tufts University

**James Myers (Lmgmyers@ccu.edu.tw)**

Graduate Institute of Linguistics  
National Chung Cheng University

**Stephen J. Goldberg (sgoldber@hamilton.edu)<sup>1</sup>**

**Ariel M. Cohen-Goldberg (ariel.goldberg@tufts.edu)<sup>2</sup>**

<sup>1</sup>Department of Art History, Hamilton College

<sup>2</sup>Department of Psychology, Tufts University

**Keywords:** grammar; written language; graphical language; orthotactics; comics; characters; calligraphy.

## Introduction

The notion of mental grammar has been at the heart of linguistic theorizing for much of the past century. The arguments for the existence of grammar, a set of mental rules/constraints governing the well-formedness of linguistic structures, are vast and varied, however the central argument that has been made in its favor is that speakers are capable of producing (and understanding) an infinite set of systematically structured words (phonology, morphology) and utterances (syntax).

Within the cognitive sciences, the notion of mental grammar has been reserved to describe competence in spoken language, signed language, and on occasion, narrative structure. This symposium explores the possibility that the notion of grammar should be extended to other cognitive domains, specifically the domain of written and graphical communication. Four different domains representing a broad swath of written communication are considered: letter combinations in spelling (orthotactics), sequential images in comics, the internal structure of individual Chinese characters, and the formal structure of calligraphic scripts. Although these domains all integrate elements of spoken language (e.g., phoneme-grapheme mappings, thought bubbles in comics, etc.), this symposium focuses exclusively on those aspects that are distinct from spoken language: abstract orthography-specific spelling knowledge, the system by which narratives are constructed with sequential graphical panels, the internal formal structure of Chinese characters, and the constraints governing the articulation of brushstrokes in a calligraphic manuscript.

The papers presented here provide theoretical and experimental evidence that the move to extend the notion of

grammar to these domains is substantive and is not simply a metaphor or an analogical borrowing of terminology. These domains are shown to have complex internal structure that is subject to specific constraints on well-formedness. In some cases acceptability judgments and electrophysiological data indicate that speakers have online, synchronic knowledge of these structural constraints. The similarities and differences between the grammars of natural languages and the written/graphical domains presented here will be discussed. Regardless of whether term ‘grammar’ is ultimately applied in these cases, the complexity and systematicity of the cognitive processes underlying these domains must be recognized.

## Modeling the statistical structure of orthographic representations

The study of positional and sequential restrictions on speech sounds (phonotactics) is a traditional subfield of linguistics that has been the subject of much recent experimental and computational research (e.g., Vitevitch et al., 2004; Hayes & Wilson 2008). The possibility that there are analogous, independent restrictions on graphemic representations -- orthotactic constraints on letter or grapheme sequences that are formally similar to those found in spoken language but not reducible to phonotactics or other phonological regularities -- has not been extensively investigated (but see Jespersen 1909-1949; Venezky 1970). In this talk, Wilson evaluates the evidence for an independent orthotactic component by combining a number of methodologies: computational modeling of the mapping from phonology to orthography in spelling, which is plausibly constrained to construct letter strings that are orthotactically acceptable; experimentally elicited judgments of stimuli that differ in spelling but not pronunciation; and prediction of spelling errors made by normal and impaired individuals, which may

similarly respect orthotactic restrictions despite deviating from the intended outputs. The talk also discusses the reciprocal issue, namely how grammatical knowledge of phonotactics can constrain and simplify the mapping from orthography to phonology in reading aloud. The resulting model is one in which individual grammars of sound, spelling, and the mapping between them combine to explain the joint statistical structure of the spoken and written forms of words. (This talk is based on joint work with Mike McCloskey, Simon Fischer-Baum, and Don Mathis).

### **The grammar of visual narratives: Structure and processing of sequential images in comics**

Comics are a ubiquitous form of visual narrative in contemporary society, and nowhere is this more prevalent than Japan, where comics occupy over one-third of all printed material (Gravett, 2004). In this talk, Cohn argues that, just as syntax allows us to differentiate coherent sentences from scrambled strings of words, the comprehension of sequential images in comics also uses a grammatical system to distinguish coherent narrative sequences from random strings of images. First, Cohn will present a theoretical model of the narrative grammar underlying comics—a hierarchic system of constituency structure that constrains the sequences of images. He then will provide an overview of recent research that supports the psychological validity of this grammar, using methods from psycholinguistics and cognitive neuroscience. In particular, Cohn will emphasize that the same neurophysiological responses that appear to violations of syntax and semantics in sentences appear to violations of narrative and semantics in the sequential images of comics. Finally, Cohn considers what ramifications a narrative grammar of sequential images has on theories of verbal narrative and language in general.

### **Levels of analysis in the generalization of Chinese character regularities**

Regardless of calligraphic style, Chinese characters obey strict shape regularities, including restrictions on reduplicated elements and on the location and shape of semantic radicals. In this talk, Myers argues that these regularities should be ascribed to a true grammar. Experimentally collected well-formedness judgments of nonce characters show that the regularities are psychologically active and readily generalize to non-radicals and lexically non-reduplicating character elements (Myers, 2011). New reanalyses show just how far beyond lexical analogy these generalizations can go. Intriguingly, the superficially distinct regularities are derivable from a single abstract structural template that, like metrical feet, shows asymmetric binary branching (Myers, 1996). New cross-regularity priming experiments test whether this template is itself active in character judgments. Together the findings suggest that high-level character grammar is not only real, but akin to prosody in spoken and signed languages.

## **Constraint interaction in the analysis of Chinese calligraphic scripts**

The field of Art History has traditionally treated its object of study—art, in its various physical manifestations—as a phenomenon “out in the world”. Yet, art, like language, is fundamentally a product of the human mind. Goldberg and Cohen-Goldberg argue that the field of Art History can benefit from a mentalist perspective where art is considered the product of artistic/esthetic cognition. Goldberg and Cohen-Goldberg provide a theoretical account of Chinese calligraphy that views calligraphic scripts (seal, clerical, and standard scripts) as the product of a grammar that must simultaneously balance the needs of scriptural well-formedness and legibility (Goldberg, 2004). Borrowing notions from work in theoretical phonology, they argue that calligraphic grammars consist of “markedness” constraints that assure that the calligraphic inscription possess script-typical qualities while “faithfulness” constraints ensure the recoverability of the underlying character. Utilizing this framework, they report novel results concerning 1) the various types of scope that are active within a calligraphic script and 2) the formal relationships that exist between scripts.

### **References**

- Goldberg, S. (2004). The primacy of gesture: Phenomenology and the art of Chinese calligraphy, in A. T. Tymieniecka (Ed.), *Analecta Husserliana LXXXVI* (p. 175-186). Dordrecht: Kluwer Academic Publishers.
- Gravett, P. (2004). *Manga: Sixty years of Japanese comics*. New York, NY: HarperCollins.
- Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3), 379-440.
- Jespersen, O. H. (1909–1949). *A modern English grammar on historical principles*. Heidelberg: Winter.
- Myers, J. (1996). Prosodic structure in Chinese characters. Poster presented at the Fifth International Conference on Chinese Linguistics, National Tsing Hua University, Taiwan.
- Myers, J. (2011). The psychological reality of formal regularities in Chinese characters. Talk presented at the 7th Conference of the European Association of Chinese Linguistics, Venice, Italy.
- Venezky, R. L. (1970). *The structure of English orthography*. The Hague: Mouton.
- Vitevitch, M. S., Armbruster, J., & Chu, S. (2004). Sublexical and lexical representations in speech production: Effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 514-529.



# Constructing a hypothesis space from the Web for large-scale Bayesian word learning

Joshua T. Abbott (joshua.abbott@berkeley.edu)

Joseph L. Austerweil (joseph.austerweil@gmail.com)

Thomas L. Griffiths (tom\_griffiths@berkeley.edu)

Department of Psychology, University of California, Berkeley, CA 94720 USA

## Abstract

The Bayesian generalization framework has been successful in explaining how people generalize a property from a few observed stimuli to novel stimuli, across several different domains. To create a successful Bayesian generalization model, modelers typically specify a hypothesis space and prior probability distribution for each specific domain. However, this raises two problems: the models do not scale beyond the (typically small-scale) domain that they were designed for, and the explanatory power of the models is reduced by their reliance on a hand-coded hypothesis space and prior. To solve these two problems, we propose a method for deriving hypothesis spaces and priors from large online databases. We evaluate our method by constructing a hypothesis space and prior for a Bayesian word learning model from WordNet, a large online database that encodes the semantic relationships between words as a network. After validating our approach by replicating a previous word learning study, we apply the same model to a new experiment featuring three additional taxonomic domains (clothing, containers, and seats). In both experiments, we found that the same automatically constructed hypothesis space explains the complex pattern of generalization behavior, producing accurate predictions across a total of six different domains.

**Keywords:** generalization; concept learning; word learning; Bayesian modeling; online databases

## Introduction

Many problems solved by the mind conform to the same abstract computational formulation: How should a property be generalized to novel stimuli from a set of stimuli observed to have the property? As there are many ways to extend the property that are consistent with some observed evidence, these are problems of *induction*, where the evidence constrains, but does not determine, the solution to a problem. The Bayesian generalization framework (Shepard, 1987; Tenenbaum & Griffiths, 2001) has been remarkably successful at explaining human generalization behavior in a wide range of domains. However, its success is largely dependent on the choice of a hypothesis space and a prior probability distribution on hypotheses, which are usually hand constructed by the researcher for each specific problem. This is unsatisfying practically, because the models do not scale beyond the originally modeled problem, and theoretically, as it is unclear whether their success is due to the cleverness of the modeler and not because of a deep mathematical property of the computational problem that people solve.

One possible solution is to use existing sources of information about the organization of a domain as the basis for specifying a hypothesis space and prior. This helps address both the practical and the theoretical concerns raised by the

Bayesian generalization model. In this paper, we use this approach to show how a hypothesis space and prior can be constructed automatically from a large online database, making it possible to apply the Bayesian generalization framework to a wide range of naturalistic stimuli. We focus on one specific generalization problem, word learning, where people learn new words from observing a few objects that can be labeled with that word. Given that the number of possible extensions of a word is essentially infinite, learning the objects referred to by a word is a very difficult inductive problem (Quine, 1975). Xu and Tenenbaum (2007) showed how the Bayesian generalization framework could be used to explain how people learn new words. However, to construct the hypothesis space of their Bayesian model, Xu and Tenenbaum (2007) elicited approximately 400 similarity judgments from their participants. Clearly this is not practical to extend into every domain where people learn words. Thus, word learning is an appropriate setting for exploring novel methods of constructing hypothesis spaces and prior distributions.

We propose a method for automatically constructing the hypothesis space and prior distribution of a Bayesian word learning model using freely available online resources. In particular, we use WordNet (Fellbaum, 2010; Miller, 1995) as an initial source for automatically creating the hypothesis space, and ImageNet (Deng et al., 2009) as a source of naturalistic images that can be used as stimuli to test the resulting model in behavioral experiments. WordNet is a popular lexical database of English comprised of over 100,000 relational sets of synonyms. ImageNet is a large ontology of images conforming to the hierarchical structure of WordNet, with the aim of providing over 500 high-quality images per noun in WordNet. These resources allow us to construct hypothesis spaces and prior distributions for word learning without eliciting a single judgment from participants and test the resulting model on a much larger scale than was previously possible. We demonstrate that the Bayesian model formulated from WordNet captures participant judgments in two behavioral experiments, addressing the practical and theoretical issues with Bayesian models discussed earlier.

The plan of the rest of the paper is as follows. In the next sections we review the Bayesian generalization model and then examine how Xu and Tenenbaum (2007) constructed the hypothesis space for their Bayesian word learning model. We then show how to build a hypothesis space from WordNet that can be used to evaluate word learning models on a large scale. Afterwards, we present two experiments utilizing this hypoth-

esis space: one that replicates a previous study of adult word learning, and one that investigates word learning for a set of complex concepts in novel domains. Finally, we discuss the implications of our work and future directions for research.

## The Bayesian Generalization Framework

The Bayesian word learning model is a special case of the Bayesian generalization framework. This framework has been used to model generalization in a number of domains including dimensional concepts (Austerweil & Griffiths, 2010; Shepard, 1987; Tenenbaum, 1999), word learning (Xu & Tenenbaum, 2007), numerical concepts (Tenenbaum, 2000), sequential rules (Austerweil & Griffiths, 2011) and rule-based categorical concepts (Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Typically, problems are formulated in this framework as follows: Assume we observe  $n$  positive examples  $\mathbf{x} = \{x_1, \dots, x_n\}$  of concept  $C$  and want to compute  $P(y \in C|\mathbf{x})$ , the probability that some new object  $y$  belongs to  $C$  given the observations  $\mathbf{x}$ . We compute this probability by using a hypothesis space  $\mathcal{H}$ , which is a set of hypothetical concepts, where each hypothesis is defined by the objects that would be members of the concept if the hypothesis were true,  $P(\mathbf{x}|h)$ .

Defining a Bayesian generalization model amounts to defining a hypothesis space  $\mathcal{H}$ , a prior probability distribution over hypotheses,  $P(h)$ , and for each hypothesis, a likelihood function,  $P(\mathbf{x}|h)$ , indicating the probability of observing a set of objects  $\mathbf{x}$  given that the hypothesis is true. A typical definition of the likelihood follows from assuming strong sampling, where objects are generated uniformly at random from the true hypothesis (Tenenbaum & Griffiths, 2001)

$$P(\mathbf{x}|h) = \begin{cases} 1/|h|^n & \text{if } \mathbf{x} \subset h \\ 0 & \text{otherwise} \end{cases}. \quad (1)$$

This likelihood function instantiates the *size principle* for scoring hypotheses: hypotheses containing a smaller number of objects assign greater likelihood than hypotheses with more objects to the same set of objects (Tenenbaum, 1999; Tenenbaum & Griffiths, 2001). The prior distribution over hypotheses,  $P(h)$  depends on the domain and in previous literature has ranged from a simple uniform distribution over the hypothesis space (Shepard, 1987) to a stochastic process over tree structures (Kemp & Tenenbaum, 2009). Given the prior and likelihood, the posterior probability that a hypothesis is true given a set of objects belonging to a novel concept,  $P(h|\mathbf{x})$ , follows from Bayes' rule:  $P(h|\mathbf{x}) \propto P(\mathbf{x}|h)P(h)$ . From this, we can compute the probability that a new object  $y$  is also a member of the concept  $C$  by averaging the predictions of all hypotheses weighted by their posterior probabilities:

$$P(y \in C|\mathbf{x}) = \sum_{h \in \mathcal{H}} P(y \in C|h)P(h|\mathbf{x}), \quad (2)$$

where  $P(y \in C|h) = 1$  if the new object  $y$  is in hypothesis  $h$ , and 0 otherwise.

## Word Learning as Bayesian Inference

Xu and Tenenbaum (2007) derived the hypothesis space for their Bayesian word learning model by applying hierarchical clustering (see Duda & Hart, 1973) to the perceived similarity of every pair of objects. The hypothesis space, prior and likelihood are defined by the tree resulting from hierarchical clustering. Using a tree is well justified from a psychological perspective as children assume the possible referents of novel nouns are tree-structured (Markman, 1991). Nodes in the tree represent potential words (hypotheses) which extend to all the leaves they cover, where the leaves of the tree correspond to the domain of possible objects. The height of a node  $h$  (minimal distance from the node to a leaf) is a measure of the average pairwise dissimilarity of objects covered by node  $h$  and approximates the heterogeneity of the objects that can be called that word. The intuition that more distinctive clusters are more likely to have distinguishing names, was incorporated by defining the prior  $P(h)$  to be proportional to the branch length separating node  $h$  from its parent:

$$P(h) \propto \text{height}(\text{parent}(h)) - \text{height}(h), \quad (3)$$

where  $\text{parent}(h)$  returns the parent of node  $h$ . To incorporate a *basic-level* bias (Markman, 1991; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) in which new words tend to refer more often to a word at an intermediate level in a taxonomy, the prior probability of hypotheses at the basic level were 10 times the value given by Equation 3 (see below for examples). As the height of node  $h$  also approximates the number of objects in the extension of the possible word  $h$ , the likelihood of observing  $n$  objects called word  $h$  is defined as

$$P(\mathbf{x}|h) \propto \left[ \frac{1}{\text{height}(h) + \epsilon} \right]^n, \quad (4)$$

where  $\epsilon$  is a small constant so that the leaf hypotheses (those that refer to only a single object) do not have infinite likelihood (as their height is zero).

Using this framework, Xu and Tenenbaum (2007) accurately predicted how people extend words to new objects depending on the diversity and number of objects labeled with that word. In a set of experiments on both adults and children, they showed participants one or more positive examples of a novel word while manipulating the taxonomic relationship of the objects the word referred to. For example, participants might observe one Dalmatian, three Dalmatians (exemplars at the subordinate-level), a Dalmatian, terrier, and mutt (exemplars at the basic-level), or a Dalmatian, pig, and toucan (exemplars at the superordinate-level) being labeled with a novel word (e.g. “fep”). After observing a word refer to one or three example objects at the subordinate, basic, or superordinate-level, they were asked whether the word referred to novel subordinate, basic, superordinate, and out-of-domain objects.

When participants were given one example of an object that refers to a word (e.g. one Dalmatian), they tended

to select the subordinate-level matches (e.g. the two other Dalmatians) and the basic-level matches (e.g. the two non-Dalmatian dogs). However, when they were shown three subordinate-level examples of a concept (e.g. three Dalmatians), the participants tended to choose only the subordinate-level matches (e.g. they only believed the word referred to the two other Dalmatians). The Bayesian word learning model captured this phenomenon because the prior favors words at the basic-level, but the likelihood favors words at the subordinate-level, and the likelihood's weight increases exponentially in the number of objects.

Unfortunately, the manner in which the hypothesis space was constructed (through hierarchical clustering on pairs of similarity judgments) poses a serious constraint to assessing the model's validity. To construct the hypothesis space in the three domains tested by Xu and Tenenbaum (2007), where there are 15 images per concept, each participant had to provide roughly 400 similarity judgments. To test how well this framework extends to new concepts and domains using their method for constructing the hypothesis space, an impractically large quantity of human judgments would need to be elicited. In the following section, we introduce an alternative method of constructing a hypothesis space for the Bayesian word learning model, which allows for testing the framework without eliciting any judgments from participants.

## Large-Scale Word Learning

Using an online word ontology, we can automatically construct the hypothesis space of a Bayesian word learning model. WordNet is a large lexical database of English represented as a network of words linked by directed edges denoting semantic relatedness (Fellbaum, 2010; Miller, 1995). Its structure was manually designed to group lexical concepts in an "is-a" hierarchy based on the many-to-one mapping of synonyms. For example, a Poodle "is-a" type of dog, thus WordNet has a directed edge from the node for *dog* to the node for *Poodle*. As WordNet is hierarchically structured like the hypothesis space used by Xu and Tenenbaum (2007), it is an ideal candidate for constructing our hypothesis space.

Using a hypothesis space derived from WordNet, we can better test the predictions of different generalization theories for word learning by examining their predictions for a large range of concepts. In the rest of this section, we present the method used to construct a hypothesis space from WordNet and outline the implementations of three generalization models using this hypothesis space for large-scale word learning.

### Constructing a Hypothesis Space

In the context of the Bayesian generalization framework, the hypotheses correspond to subsets of the universe of objects that are psychologically plausible candidates as extensions of concepts (Tenenbaum & Griffiths, 2001). Using WordNet as the basis of our hypothesis space, the set of objects is the set of leaf nodes from the noun-space of the directed graph and the hypotheses correspond to both the inner nodes of the directed graph and the leaf nodes, which distinguish between

objects at the subordinate-level. To construct a hypothesis space from WordNet, we first extracted a tree from the 82,115 noun nodes of WordNet.<sup>1</sup> The nodes are hypotheses, which represent possible words, and form the hypothesis space for the model. From this graph we create a hypothesis space that is a binary matrix,  $\mathcal{H}$ , whose rows are the objects (64,958 leaf nodes from the graph) and columns are the hypotheses (82,115 nodes, 17,157 of which are inner nodes and 64,958 are leaf nodes). Each entry  $(i, j)$  of the matrix  $\mathcal{H}$  denotes whether or not hypothesis node  $j$  is an ancestor of leaf node  $i$  in the WordNet graph (with a 1 indicating it is). The leaf nodes are included as hypotheses so that the model distinguishes between subordinate objects.

### Generalization Models

With a hypothesis space derived from WordNet, we now have the ability to test the Bayesian model of word learning on a much larger scale. In addition, we can use the hypothesis matrix as a feature space for testing alternative models. We compare the Bayesian model against two similarity models: a prototype model and an exemplar model. Given a set of examples  $\mathbf{x} = \{x_1, \dots, x_n\}$  representing some concept  $C$  (where the elements of  $\mathbf{x}$  correspond to rows in the hypothesis matrix  $\mathcal{H}$ ), we can compute a score for each row  $y \in \mathcal{H}$  denoting the probability that  $y$  is also a member of  $C$ . We present the different ways to compute this score below.

**Bayesian model.** This is the Bayesian generalization framework that we discussed earlier. We used strong sampling for the likelihood,  $P(\mathbf{x}|h)$ , which is computed via Equation 1, where the size of  $h$  is the number of nodes that can be reached by a directed path from  $h$ . This simply corresponds to the sum of the elements in the column corresponding to  $h$ .

The prior  $P(h)$  was defined to be Erlang distributed in the size of the hypothesis (a standard prior over sizes in Bayesian models; Shepard, 1987; Tenenbaum, 2000)

$$P(h) \propto (|h|/\sigma^2) \exp\{-|h|/\sigma\}, \quad (5)$$

where the  $\sigma$  parameter was set to 200 by hand fitting the model predictions to all human responses (the same value was used in both experiments). This value favors medium sized hypotheses, which is roughly equivalent to a basic-level bias. The probability that word  $C$  extends to object  $y$  after observing a set of objects called  $C$  is

$$\text{Bscore}(y) = P(y \in C|\mathbf{x}) = \sum_{h \in \mathcal{H}} P(y \in C|h)P(h|\mathbf{x}), \quad (6)$$

where  $P(y \in C) = 1$  if  $y \in h$  and 0 otherwise, and  $P(h|\mathbf{x})$  is the posterior distribution over hypotheses.

**Prototype model.** In this model, we define the prototype of a set of objects,  $x_{\text{proto}}$ , to have those features owned by a majority of the objects in the set. The generalization measure for

<sup>1</sup>Technically WordNet is a directed acyclic graph because some nodes have multiple parents (the method still works in these cases).

an object  $y$  is

$$\text{Pscore}(y) = \exp\{-\lambda_p \text{dist}(y, x_{\text{proto}})\}, \quad (7)$$

where  $\text{dist}(\cdot, \cdot)$  is the Hamming distance between the two vectors and  $\lambda_p$  is a free parameter (for all of the results presented here,  $\lambda_p = 0.15$ , optimized by hand using half-interval search). Pscore was then normalized over all objects  $y$  in the hypothesis space (all leaf nodes).

**Exemplar model.** We define the exemplar model using a similar scoring metric as the prototype model, except rather than computing the distance of object  $y$  to a single prototype vector, we compute a distance for each item  $x_j$  in the set of observations  $\mathbf{x}$ . The exemplar generalization measure is thus computed as

$$\text{Escore}(y) = \sum_{x_j \in \mathbf{x}} \exp\{-\lambda_e \text{dist}(y, x_j)\}, \quad (8)$$

where  $\text{dist}(\cdot, \cdot)$  is the Hamming distance between two vectors and  $\lambda_e$  is a free parameter (for all of the results presented here,  $\lambda_e = 0.20$ , optimized by hand using half-interval search). Escore was then normalized over all objects  $y$  in the hypothesis space (all leaf nodes).

## Behavioral Experiments

To evaluate the performance of our models using the WordNet-based hypothesis space, we conducted two experiments using the paradigm of Xu and Tenenbaum (2007). The first experiment replicates Xu and Tenenbaum (2007) on their three object taxonomies (animals, vehicles, and vegetables), which validates our approach for constructing a hypothesis space from WordNet and using images from ImageNet as stimuli. The second experiment extends the paradigm into three previously unexplored domains (clothing, containers, and seats), which have hierarchical structure, but it is not as clear how well they conform to a natural basic-level taxonomy (Rosch et al., 1976).

### Experiment 1: Validating Our Approach

**Participants.** Thirty four participants were recruited via Amazon Mechanical Turk and compensated \$0.05 for each trial (training set) completed out of twelve possible. Each participant completed as many trials as he or she wished, and twenty unique participants completed each trial. All participant responses were used.

**Stimuli and Procedure.** Within each taxonomy, the stimuli consisted of the images of objects distributed across the superordinate, basic and subordinate-levels, and subsequently split into training and test sets. The training sets were the labeled objects given to participants of which there were four conditions: a single subordinate-level example (e.g. a Dalmatian); three examples of the same subordinate-level object (e.g. three Dalmatians); the subordinate-level object and

two basic-level objects (e.g. a Dalmatian, a Shih Tzu, and a Beagle); and the subordinate object and two superordinate-level objects (e.g. a Dalmatian, a hippopotamus, and a toucan). This corresponds to twelve trials total (four conditions for each of the three object taxonomies).

The test sets were the same regardless of the training set and consisted of eight objects matching the currently tested taxonomy: two subordinate examples (e.g. two other Dalmatians); two basic-level examples (e.g. a Cocker Spaniel and a Corgi); and four superordinate examples (e.g. a cat, a bear, a sea lion, and a horse). There were also sixteen non-matching objects in the test set corresponding to the objects that match the two other taxonomies.

For each trial, participants were instructed that they needed to help a cartoon frog who speaks a different language from us, pick out objects that he wants. The frog shows one or more examples of a novel word (e.g. “dak”) and the participant is instructed to select other items that are a “dak” from the objects comprising the test set. A unique novel word was associated with each of the twelve trials.

**Results.** Figure 1 shows the results of this experiment, along with the predictions of the different generalization models. For each training set condition, the data for each test item has been averaged over participants and domains. The generalization judgments of participants (left-most panel of Figure 1) follows the same qualitative trend as those reported in Xu and Tenenbaum (2007). There is a sharp drop in generalization to basic-level objects when seeing only a single subordinate example compared to the condition when seeing three subordinate examples.

The Bayesian model predictions (second panel from the left) exhibits this same generalization pattern ( $r^2 = 0.98$ ), while the prototype and exemplar models do not ( $r^2 = 0.66$  and  $r^2 = 0.84$ , respectively). This validates our method of automatically creating hypothesis spaces with WordNet.

### Experiment 2: Novel Domains

**Participants.** Thirty six participants were recruited via Amazon Mechanical Turk and compensated \$0.05 for each trial completed out of twelve possible. As in Experiment 1, each participant completed as many trials as he or she wished, and twenty unique participants completed each trial. All participant responses were used.

**Stimuli and Procedure.** Table 1 contains the objects we used for training in the three hierarchical domains (clothing, containers, and seats).<sup>2</sup> As in Experiment 1, the same test objects were used for every training set, and the “non-match” test objects were the objects in the test set which match the two other taxonomies that are not contained in the training set. As before, this corresponds to twelve trials total. The procedure was identical to Experiment 1.

<sup>2</sup>The additional subordinate-level training image and the test images were omitted from Table 1 for brevity.

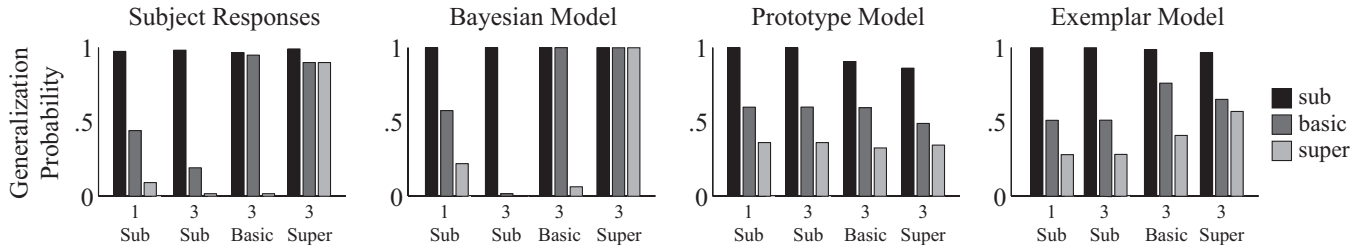


Figure 1: Participant generalization judgments and predictions of the Bayesian, prototype, and exemplar models averaged across the three domains in Experiment 1. The generalizations for non-matching items are omitted for brevity (neither the participants chose nor the Bayesian model predicted non-matching objects, while the prototype and exemplar models predicted non-matches less than 4% of the time for each condition).

**Results.** Figure 2 presents the averaged results of how participants<sup>3</sup> and the Bayesian model generalized the learned words to the test objects based on the observed training set across the different domains in Experiment 2.<sup>4</sup> Across the three domains, the generalization probabilities of the participants and Bayesian model with the same parameters are extremely similar. This is exemplified in the very good quantitative model fit on the averaged data ( $r^2 = 0.95$ ). Furthermore, the hypothesis space constructed automatically from WordNet explains the idiosyncrasies of participant generalization behavior in each domain ( $r^2 = 0.97, 0.88$ , and  $0.91$ , for clothing, containers, and seats respectively). For example, the model accurately predicts that participants would generalize most broadly in the seats domain for the single exemplar and three basic-level exemplar training sets. Additionally, the model captures that people generalized the least in the containers domain for the three subordinate-level exemplar training sets. This would not have been possible if the hypothesis space for each domain had the same structure.

Note that there is a larger amount of variance between model predictions and human performance in Experiment 2 than Experiment 1. We believe that this is due to the domains not conforming to a natural taxonomy. For example, it is unclear if box should be the basic-level category for a mail box and a cigar box; however, this is the basic level of these objects provided by WordNet. Regardless, the good quantitative fit of the Bayesian model’s predictions provides evidence that using WordNet as a hypothesis space for word learning can capture people’s generalizations even for hierarchies without clearly defined basic-level concepts.

## Discussion

Although the Bayesian generalization framework has been extremely successful in explaining human generalization behavior, the hypothesis spaces are typically hand-constructed, which is unsatisfying. In this paper, we explored automatically constructing the hypothesis space using an online re-

source as a potential solution to the methodological challenges posed by this problem. In the first behavioral experiment, we validated that the Bayesian model using this hypothesis space can capture previously found word learning phenomena. In the second behavioral experiment, we showed that the same Bayesian model explains how participants learned words in three novel domains. Using the automatically constructed hypothesis space, the model predicted the subtle changes in participants’ word learning behavior across three domains, thus demonstrating the practical and theoretical benefits of our approach.

In the future, we hope to perform a large scale empirical test of the Bayesian word learning model using more heterogeneous training sets (e.g. one subordinate-level and one basic-level object) and more domains with varied conceptual structure. The larger set of empirical results would enable us to perform a more detailed investigation of the prior knowledge over the types of conceptual structures that people use when they learn words (e.g. do people prefer shallow or deep taxonomies?). Additionally, we hope to incorporate how participant behavior is affected by the visual similarity of the images in the training and tests sets (and its interaction with conceptual structure), which at the moment would not be possible to explore with the Bayesian word learning model.

As word learning is a special case of the more general problem of generalization, our approach potentially could be applied to automatically construct hypothesis spaces for generalization problems in other domains. For example, a Bayesian model of commonsense reasoning could be formulated by automatically deriving hypothesis spaces from ConceptNet (Liu & Singh, 2004) or OpenCyc (Matuszek, Cabral, Witbrock, & DeOliveira, 2006). This follows a development in modern machine learning, which has leveraged online resources to make more successful learning algorithms (Medelyan, Legg, Milne, & Witten, 2009; Ponzetto & Strube, 2006). We hope that this draws a closer connection between computer science and cognitive science, which can lead to more psychologically valid, yet still scalable, artificial intelligence systems.

**Acknowledgments.** This work was supported by the DARPA BOLT contract HR0011-11-2-0009 and grant number IIS-0845410 from the National Science Foundation.

<sup>3</sup>Non-matching objects were only chosen twice (both in the containers domain) and so, they were omitted from Figure 2.

<sup>4</sup>The prototype and exemplar models were omitted from Figure 2 for brevity ( $r^2 = 0.80$  and  $r^2 = 0.90$  averaged over domains, respectively).



















Object level	Clothing		Containers		Seats	
	1	2	1	2	1	2
Subordinate						
Basic						
Superordinate						

Table 1: Training images for Experiment 2.

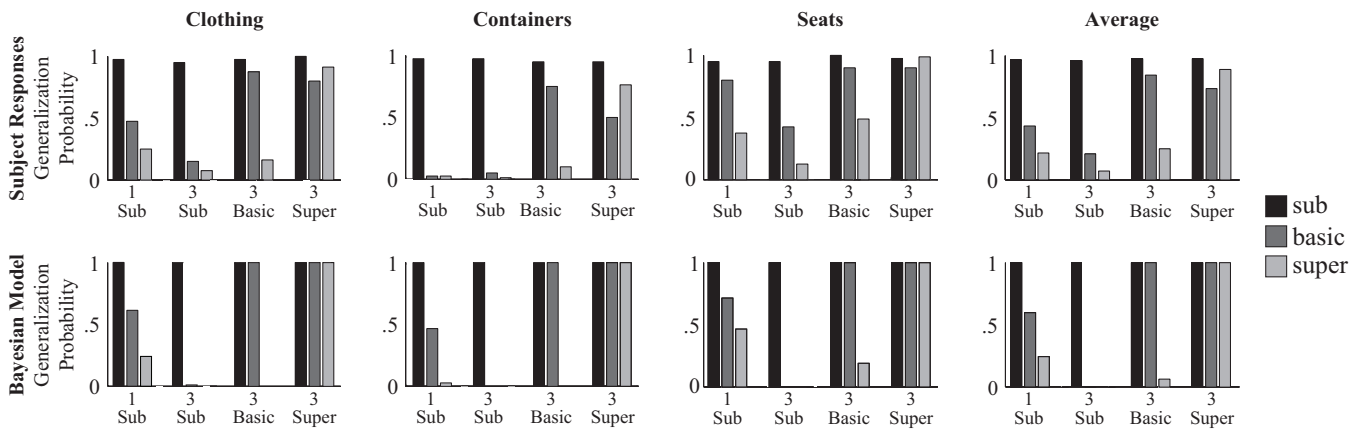


Figure 2: Participant generalization judgments and the predictions of the Bayesian model for Experiment 2. From left to right, the columns present the results for the three taxonomies (clothing, containers, and seats) and average results.

## References

- Austerweil, J. L., & Griffiths, T. L. (2010). Learning hypothesis spaces and dimensions through concept learning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 73–78). Austin, TX: Cognitive Science Society.
- Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, 35, 499–526.
- Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255).
- Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: Wiley.
- Fellbaum, C. (2010). WordNet. *Theory and Applications of Ontology: Computer Applications*, 231–243.
- Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1), 108–154.
- Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, 116(1), 20–58.
- Liu, H., & Singh, P. (2004). ConceptNet - A practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4), 211–226.
- Markman, E. M. (1991). *Categorization and naming in children: Problems of induction*. MIT Press.
- Matuszek, C., Cabral, J., Witbrock, M., & DeOliveira, J. (2006). An introduction to the syntax and content of cyc. In C. Baral (Ed.), *Proceedings of the AAAI 2006 Spring Symposium* (p. 44–49). Menlo Park, CA: AAAI Press.
- Medelyan, O., Legg, C., Milne, D., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9), 1–76.
- Miller, G. (1995). WordNet: a lexical database for english. *Communications of the ACM*, 38(11), 39–41.
- Ponzetto, S. P., & Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the HLT Conference of the NAACL* (p. 192–199).
- Quine, W. V. O. (1975). *Word and object*. MIT Press.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3), 382–439.
- Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, 237, 1317–1323.
- Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems 11* (Vol. 11, p. 59–65). Cambridge, MA: MIT Press.
- Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), *Advances in Neural Information Processing Systems 12* (Vol. 12, pp. 59–65).
- Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4), 629–640.
- Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2), 245–272.

# Predicting focal colors with a rational model of representativeness

**Joshua T. Abbott** ([joshua.abbott@berkeley.edu](mailto:joshua.abbott@berkeley.edu))

Department of Psychology, University of California, Berkeley, CA 94720 USA

**Terry Regier** ([terry.regier@berkeley.edu](mailto:terry.regier@berkeley.edu))

Department of Linguistics, Cognitive Science Program, University of California, Berkeley, CA 94720 USA

**Thomas L. Griffiths** ([tom\\_griffiths@berkeley.edu](mailto:tom_griffiths@berkeley.edu))

Department of Psychology, University of California, Berkeley, CA 94720 USA

## Abstract

Best examples of categories lie at the heart of two major debates in cognitive science, one concerning universal focal colors across languages, and the other concerning the role of representativeness in inference. Here we link these two debates. We show that best examples of named color categories across 110 languages are well-predicted by a rational model of representativeness, and that this model outperforms several natural competitors. We conclude that categorization in the contested semantic domain of color may be governed by general principles that apply more broadly in cognition, and that these principles clarify the interplay of universal and language-specific forces in color naming.

**Keywords:** Language and perception; semantic universals; color naming; representativeness; Bayesian inference.

## Introduction

Do the world's languages reflect a universal repertoire of cognitive and perceptual categories? Or do different languages partition the experienced world in fundamentally different ways? These questions have been pursued in depth in the domain of color naming and cognition (e.g. Berlin & Kay, 1969; Kay & McDaniel, 1978; Lindsey & Brown, 2006; Roberson, Davidoff, Davies, & Shapiro, 2005; Roberson, Davies, & Davidoff, 2000), and current findings suggest an interestingly mixed picture. There are clear universal tendencies of color naming across languages, but there is also substantial cross-language variation (e.g. Regier, Kay, & Khetarpal, 2007), more than is suggested by traditional universalist accounts. At the center of this debate is the disputed role of *focal colors*, or the best examples of named color categories.

It has long been claimed that color naming across languages is constrained by six universal privileged points, or foci, in color space, corresponding to the best examples of what would be described in English as *white*, *black*, *red*, *yellow*, *green*, and *blue*. This view has received empirical support: the best examples of color terms across languages tend to cluster near these six points (Berlin & Kay, 1969; Regier, Kay, & Cook, 2005), and these colors have also been found to be cognitively privileged (Heider, 1972; but see Roberson et al., 2000). A natural and influential proposal (Kay & McDaniel, 1978) is that these privileged colors constitute a universal foundation for color naming, such that languages differ in their color naming systems primarily by grouping these universal foci together into categories in different ways.

Roberson et al. (2000) advanced a diametrically opposed view of color naming, and of the role of best examples in it.

They argued that color categories are not defined around universal foci, but are instead defined at their boundaries by local linguistic convention, which varies across languages. They proposed: "Once a category has been delineated at the boundaries, exposure to exemplars may lead to the abstraction of a central tendency so that observers behave as if their categories have prototypes" (p. 395). On this view, best examples do not reflect a universal cognitive or perceptual substrate, but are merely an after-effect of category construction by language: best examples are derived from language-specific boundaries, rather than boundaries from universal best examples.

A proposal by Jameson and D'Andrade (1997) has the potential to reconcile these two opposed stances. This proposal holds that there are genuine universals of color naming, but they do not stem from a small set of focal colors. Instead, universals of color naming may stem from irregularities in the overall shape of perceptual color space, which is partitioned into categories by language in a near-optimally informative way. This proposal has been shown to explain universal tendencies in the *boundaries* of color categories (Regier et al., 2007). However it has not yet provided an account of *best examples* of these categories, which lie at the heart of the debate.

Here, we address this open issue, completing the reconciliation of the two standardly opposed views. We suggest that best examples are largely universal (in line with the universal-foci view), but nonetheless derived from category boundaries (in line with the relativist view). Specifically, given the independent explanation of category boundaries in terms of the shape of color space, we propose that universal tendencies of best examples are derived from those of boundaries, rather than the other way around as has been traditionally assumed. Moreover, we propose that best examples are derived from category boundaries in an optimal manner, echoing the optimal or near-optimal partition of color space into categories. To pursue this idea, we draw on previous work on a rational model of representativeness, and ask whether the best examples of color categories can be well-predicted as those colors that are most representative of a given category.

The remainder of the paper proceeds as follows. In the next section, we discuss the previous work on representativeness on which we draw, and contrast it with other approaches to that problem. We then describe the color naming data we consider, and a set of competing models that predict the foci



of color categories from the extensions of those categories. We first test these models broadly against data from 110 languages, and then test them in a targeted fashion against the data of a language with an unusual color naming system. In both cases, we find that the rational model of representativeness provides a good fit to the empirical data, and outperforms competing models. We close by discussing the implications of our findings.

## Representativeness

Why do people believe that the sequence of coin flips HHTHT (where H=heads, T=tails) is more likely than the sequence HHHHH to be produced by a fair coin? Using simple probability theory, it is easy to show that the two sequences are in fact equally likely. Cognitive psychologists have proposed that people use a heuristic of “representativeness” instead of performing probabilistic computations in such scenarios (Kahneman & Tversky, 1972). We might then explain why people believe HHTHT is more likely than HHHHH to be produced by a fair coin by arguing that the former is more representative of the output produced by a fair coin than the latter. If this heuristic is a correct account of such inferences, how do we define it? Numerous proposals have been made, connecting representativeness to existing quantities such as similarity (Kahneman & Tversky, 1972), and likelihood (Gigerenzer, 1996). Tenenbaum and Griffiths (2001) took a different approach to this question, providing a *rational analysis* (Anderson, 1990) of representativeness by trying to identify the problem that such a quantity solves. They noted that one sense of representativeness is being a good example of a concept, and showed how this could be quantified in the context of Bayesian inference.

Formally, given some observed data  $d$  and a set of of hypothetical sources,  $\mathcal{H}$ , we assume that a learner uses Bayesian inference to infer which  $h \in \mathcal{H}$  generated  $d$ . Tenenbaum and Griffiths (2001) defined the representativeness of  $d$  for hypothesis  $h$  to be the evidence that  $d$  provides in favor of a specific  $h$  relative to its alternatives:

$$R(d, h) = \log \frac{p(d|h)}{\sum_{h' \neq h} p(d|h')p(h')} \quad (1)$$

where  $p(h')$  in the denominator is the prior distribution on hypotheses, re-normalized over  $h' \neq h$ . This measure was shown to outperform similarity and likelihood in predicting human representativeness judgments for a number of simple stimuli. We propose this measure can also be used to determine focal colors from the set of colors named with a particular color term - that is, the extension of that named color category.

## Representativeness and color foci

Evaluating formal models of representativeness as an account of color foci requires a good source of color naming data. The data we considered were those of the World Color Survey (WCS), which collected color naming data from native speakers of 110 unwritten languages worldwide (Cook,

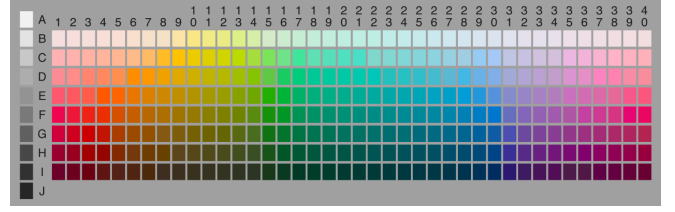


Figure 1: The WCS stimulus array. The rows correspond to 10 levels of Munsell value (lightness), and the columns correspond to 40 equally spaced Munsell hues. The color in each cell corresponds approximately to the maximum available Munsell chroma for that hue-value combination.

Kay, & Regier, 2005). Participants in the WCS were shown each of the 330 color chips from the stimulus array in Figure 1, and were asked to name each chip with a color term from their native language; we refer to the resulting data as “naming data”. Afterwards, participants were asked to pick out those cells in the stimulus array that were the best examples (foci) of each color term they used; we refer to these as “focus data”. The WCS dataset is available at <http://www.icsi.berkeley.edu/wcs/data.html>.

We applied Tenenbaum and Griffiths’ (2001) representativeness model, and a set of natural competitor models, to the problem of predicting best examples of color categories from the extension of those categories. Thus, the models we consider are different formalizations of our central proposal that best examples may be derived from category boundaries. Following Kay and Regier (2003), we represented each color in the stimulus array as a point in 3-dimensional CIELAB color space. For short distances at least, Euclidean distance between two colors in CIELAB is roughly proportional to the perceptual dissimilarity of those colors. For each named color category used by each speaker in each language of the WCS, we modeled that category as a 3-dimensional Gaussian distribution in CIELAB space, and estimated the parameters of that distribution using a normal-inverse-Wishart prior, a standard estimation method for multivariate Gaussian distributions of unknown mean and unknown variance (Gelman, Carlin, Stern, & Rubin, 2004). Specifically, given a set of  $M$  chips  $\mathbf{x}_i$  in color category  $t$  we obtain the estimates:

$$\mu_t = \frac{1}{M} \sum_i^M \mathbf{x}_i$$

$$\Sigma_t = \frac{SS_t + \lambda_0}{n_t + \nu_0}$$

where  $SS_t$  is the sum of squares for category  $t$ :  $\sum_i^M (\mathbf{x}_i - \mu_t)(\mathbf{x}_i - \mu_t)^\top$ ,  $n_t$  is the number of chips in category  $t$  for the current speaker, and  $\lambda_0$  and  $\nu_0$  are the parameters of the prior.  $\lambda_0$  was set by taking an empirical estimate of the variance in CIELAB coordinates over all chips in the stimulus array, and  $\nu_0$  was set to 1. We chose this Bayesian formulation of parameter estimation over standard Maximum Likelihood Es-



timization (MLE) since MLE will result in singular covariance matrices for color categories containing few color chips.

With an estimate of the distribution characterizing the category named by color term  $t$ , we can now adopt the representativeness measure given in Equation 1 to determine how good an example each color chip  $x$  is of a color term  $t$ . Substituting  $x$  in for the observed data  $d$  and  $t$  for hypothesis  $h$  we obtain the expression:

$$R(x, t) = \frac{p(x|t)}{\sum_{t' \neq t} p(x|t')p(t')} \quad (2)$$

where  $p(x|t)$  is computed from the density function of the estimated Gaussian described above and the priors  $p(t')$  are proportional to  $n_{t'}$ , the number of chips in named color category  $t'$ . We test this Bayesian measure against the alternative proposals of representativeness mentioned above (Gigerenzer, 1996; Kahneman & Tversky, 1972): a likelihood model and two similarity models (a prototype model and an exemplar model). In addition, we explore a model that selects as the focus for category  $t$  that chip in the extension of  $t$  that has the highest chroma. Chroma, or saturation, corresponds loosely to how colorful or “un-gray” a given color is, and in exploring this model we follow the suggestion (Jameson & D’Andrade, 1997; Regier et al., 2007) that focal colors tend to be those with high chroma. We note that each of these models captures some variant of the category *central tendency* idea promoted by Roberson et al. (2000), as described above. We present the details of the competing models below. As with the representativeness model, for a given color  $x$  and color term  $t$ , each model assigns a score indicating how good  $x$  is as an example of  $t$ .

**Likelihood model.** In this model, the goodness score of color  $x$  as an example of color category  $t$  is given by the density function of the Gaussian distribution that was fit to the naming data for  $t$ . Thus,

$$L(x, t) = p(x|t) \quad (3)$$

Note that this model is similar to the representativeness model, but without the denominator which captures competition among categories in that model.

**Prototype model.** In this model we define the focus, or prototype, of color category  $t$  to be the mean  $\mu_t$  of the distribution characterizing  $t$ . The score for this measure then becomes the similarity of  $x$  to that prototype:

$$P(x, t) = \exp\{-\text{dist}(x, \mu_t)\} \quad (4)$$

where  $\text{dist}(\cdot, \cdot)$  is the Euclidean distance between two colors in CIELAB color space.

**Exemplar model.** We define the exemplar model using a scoring metric similar to that in the prototype model, except rather than computing the similarity of color  $x$  to a single prototype, we compute its similarity to each color chip that falls

in the extension of category  $t$ , and sum the results. This similarity measure is thus computed as

$$E(x, t) = \sum_{x_j \in \mathbb{X}_t} \exp\{-\lambda \text{dist}(x, x_j)\} \quad (5)$$

where  $\mathbb{X}_t$  is the set of color chips that fall in the extension of category  $t$ ,  $\text{dist}(\cdot, \cdot)$  is the Euclidean distance between two colors in CIELAB space, and  $\lambda$  is a free parameter. For the results presented below,  $\lambda$  was set to the value that yielded the best performance overall, which was 0.25.

**Chroma model.** The score for this model is computed similarly to that for the prototype model, but rather than computing the similarity of color  $x$  to the mean of a distribution characterizing category  $t$ , we compute its similarity to that color chip  $c_t$  which has the highest chroma (saturation) value within the extension of category  $t$ . The chroma values for each chip in the stimulus array are provided with the WCS data. Thus we compute

$$C(x, t) = \exp\{-\text{dist}(x, c_t)\} \quad (6)$$

where  $\text{dist}(\cdot, \cdot)$  is the Euclidean distance between two colors in CIELAB space, and  $c_t$  is the chip within the extension of  $t$  that has the highest chroma value. In the case of ties for  $c_t$  - that is, several chips with the same maximum value for chroma - we randomly select a chip from the set of ties.

## Predicting foci from category extensions

We assessed these models as follows. For each speaker of each language in the WCS, we first considered that speaker’s naming data, and modeled the categories in those data as a set of Gaussians in the manner described above. Then, for each such category, we determined how representative of that category each of the 330 chips in the stimulus array is, according to each model. This yielded, for each model, a ranking of chips in the array by predicted representativeness, and we then compared this model prediction with empirical focus data from the WCS. In the following sections we present both qualitative and quantitative evaluations of the models.

## Distribution of foci

A simple means of assessing the models is to generate predicted focal choices from each model’s ranking of chips, and to then compare those predicted focal choices with the actual focus data of the WCS. Some speakers in the WCS provided more than one focus (best example) for some categories; if a speaker provided  $n$  foci for a given category, we selected the  $n$  top-ranked chips as a given model’s predicted focal choices for that category and speaker. In this manner we obtained, for each model, one predicted focal choice for each empirical focal choice in the data. We then counted the number of times each of the 330 color chips in the stimulus array was selected as a focal choice, yielding a distribution of focal choices over the stimulus array. We then compared the

empirical distribution of foci across the array with the distribution predicted by each of the models. Following Regier et al.'s (2005) empirical analysis of WCS focus data, we plotted these distributions over the chromatic portion of the array, where the 2-dimensional layout makes contours easily interpretable. Accordingly, we did not plot the focal choices for the terms a speaker used to name A0 and J0, corresponding to English focal *white* and *black*. The resulting contour plots, of the empirical WCS focus distribution and the five models' predicted focus distributions, are shown in Figure 2.

The empirical distribution is shown in panel (a), and replicates the findings of Regier et al. (2005). The distribution predicted by the Bayesian representativeness model (panel b) matches this empirical distribution qualitatively fairly well. Moreover, at least on informal inspection, the Bayesian model appears to approximate the empirical distribution more closely than do the competing models. The chroma model (panel f) at first appears to also approximate the empirical distribution fairly well, but closer inspection reveals that several of the peaks of the model distribution do not align correctly with those of the empirical distribution.

This qualitative assessment is reinforced by a quantitative one. The Jensen-Shannon divergence (JSD) is a measure of the dissimilarity between two probability distributions,  $P$  and  $Q$ , defined as

$$\text{JSD}(P||Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M) \quad (7)$$

where  $M = \frac{1}{2}(P + Q)$ , and  $KL(\cdot)$  is the more commonly-known Kullback-Leibler divergence. JSD is closely related to Kullback-Leibler divergence, with the important difference that JSD is always a finite value, ranging from a value of 0 when the two distributions are identical, to a value of 1 when they are maximally different.

We computed the JSD between the WCS empirical focus distribution (normalized so that it may be considered a probability distribution, taken to be  $P$  in Equation 7), and each of the model distributions (similarly normalized, taken to be  $Q$  in Equation 7). The results are shown below in Table 1. The Bayesian model outperforms the other models, diverging less from the empirical distribution than its competitors.

### Rank position of foci

Each model produces as output a ranking of the stimulus chips, where rank is assigned in descending order. Thus, another natural way to assess the models is to note the position of the true empirical focal choice in this ranked list. For example, if a model correctly ranked the true focal chip as the single most representative example of a given color category, it would receive a score of 1/330. As noted previously, sometimes a speaker provided multiple foci for a given color term. To accommodate this we averaged the positional ranking of each focus empirically provided and took the resulting quantity as the model performance for a given color term. In turn, we averaged this performance over the number of color terms a speaker used, then averaged over the number of speakers in

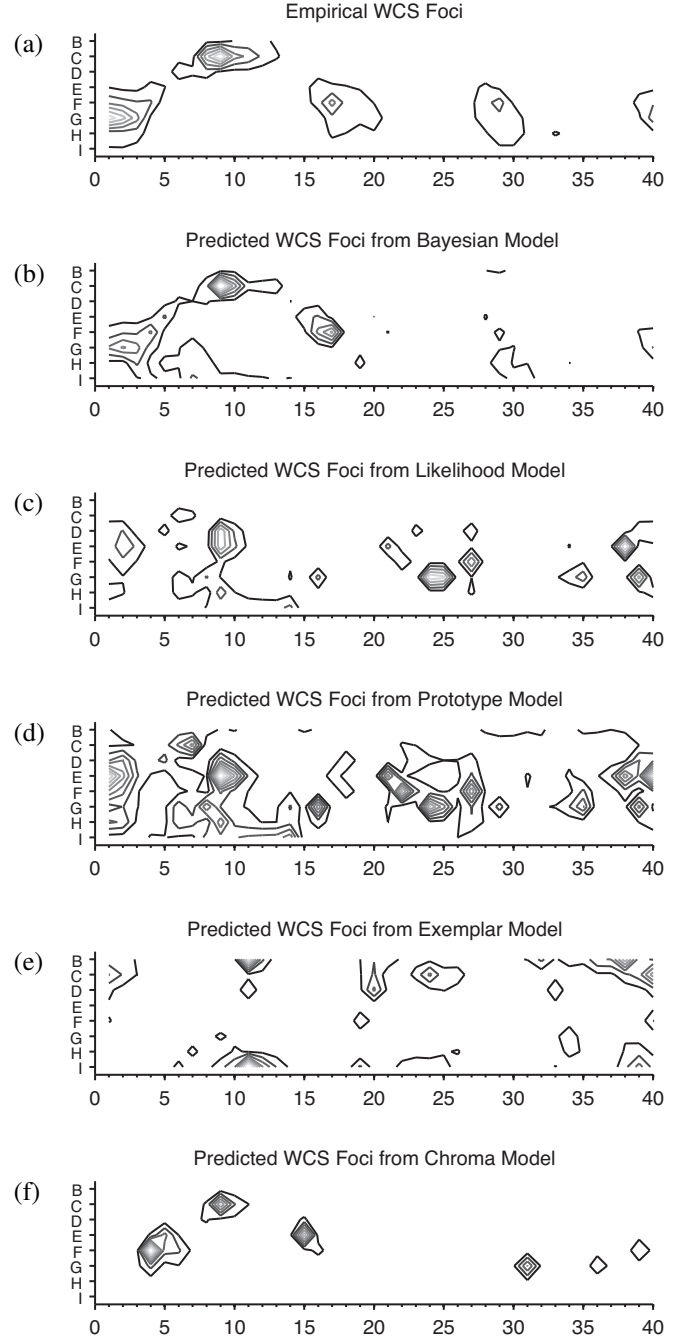


Figure 2: Contour plots of the focus distributions in (a) the WCS, and as predicted by (b) the representativeness model, (c) the likelihood model, (d) the prototype model, (e) the exemplar model, and (f) the chroma model. Each contour line corresponds to 100 focal choices.

a language, and finally computed an average overall model performance for all 110 WCS languages. The average rank position of empirical WCS foci for each model is presented in Table 2.

As before, we find that the Bayesian measure of representativeness outperforms the other models, ranking the true

Table 1: Divergence between empirical WCS focus distribution and model prediction

Model	Jensen-Shannon Divergence
Bayesian	0.0368
Likelihood	0.1977
Prototype	0.1750
Exemplar	0.1760
Chroma	0.1698

Table 2: Average rank position of empirical WCS foci for each model

Model	Average Rank Position
Bayesian	0.1026
Likelihood	0.1381
Prototype	0.1559
Exemplar	0.1457
Chroma	0.2306

foci within the top 11% of chips on average. In comparison, the likelihood model, which has the second highest average, ranks the true foci in the top 14% of chips on average. It is noteworthy that the chroma model, which captures the natural idea that best examples correspond to chroma maxima, performs most poorly, ranking the true foci only within the top 24% of chips.

### A final test: Karajá

So far, we have suggested that color foci may be derived from category boundaries as representative members of a category - and we have shown that this idea accounts well for universal tendencies in focal colors. Thus, foci may inherit their universal tendencies from category boundaries, rather than projecting their universal tendencies to those boundaries. Note, however, that the demonstrations we have seen so far do not discriminate between these two hypotheses. For languages with common color-naming systems, the two hypotheses make the same prediction: foci should tend to fall in the canonical positions shown in Figure 2(a). This is predicted on the traditional universal-foci account, because these are the proposed locations of the universal foci. Roughly the same outcome is predicted by our account, as seen in Figure 2(b).

In a final investigation, then, we attempt to discriminate between these two hypotheses. The hypotheses diverge in their predictions for languages with color categories that have unusual extensions. If foci are a universal groundwork for color naming, then in such unusual cases, foci will fall in the universal (canonical) positions, despite the non-canonicity of the category boundaries. In contrast, our account predicts that in such cases, foci should follow the category boundaries, and fall in non-canonical positions. We test these predictions against a language that is known (Regier, Kay, & Khetarpal, 2009) to have color categories with unusual extensions: Karajá, a language of Brazil.

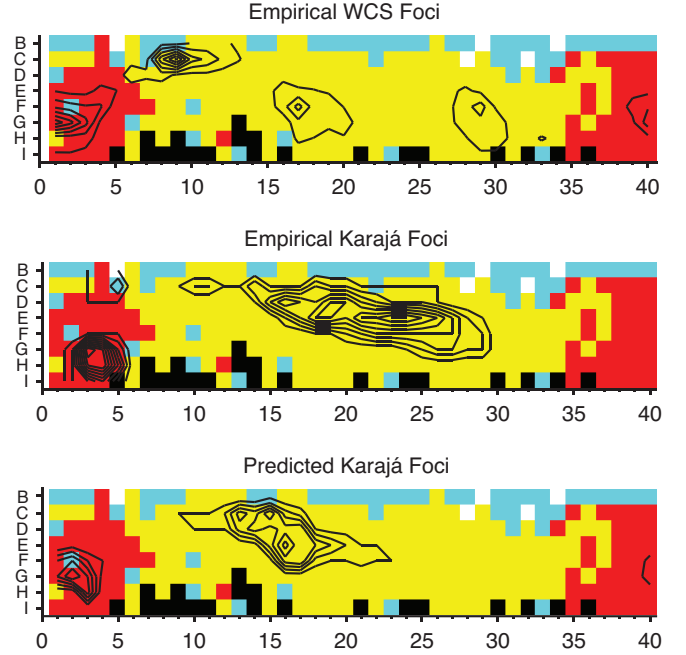


Figure 3: Naming data for the Karajá language, overlaid with contour plots of three different focus distributions: the empirical focus distribution for all languages in the WCS (upper panel), the empirical focus distribution for Karajá itself (middle panel), and the focus distribution predicted by the Bayesian model of representativeness (lower panel).

Figure 3 presents WCS color naming data for Karajá. Here, chips displayed in the same color were named with the same color term by a plurality of participants. These modal naming maps are overlaid with three different focus distributions: the full empirical focus distribution of the WCS (upper panel), the empirical focus distribution from Karajá only (middle panel), and the focus distribution for Karajá predicted by the Bayesian representativeness model (lower panel). The difference between the focus distributions in the top two panels is clearly seen, demonstrating that the foci of Karajá follow the language’s color boundaries and are not in line with the universal foci found across the WCS. Additionally, the focus predictions from the Bayesian model of representativeness follow the empirical Karajá focus distribution relatively closely. As before, these qualitative results are confirmed by a quantitative analysis that measures the Jensen-Shannon divergence between the empirical Karajá focus distribution and the distribution predicted by each of the models. As can be seen in Table 3 below, the Bayesian model outperforms the other models on Karajá considered by itself, not just on the entire WCS dataset. We also examined the rank position of the empirical Karajá foci in the ranking produced by each model, and by this measure as well, the Bayesian model fits the data more closely than the competitors, as shown in Table 4 below.

In sum, when boundaries fall in non-canonical positions,

Table 3: Divergence between empirical Karajá focus distribution and model prediction

Model	Jensen-Shannon Divergence
Bayesian	0.3272
Likelihood	0.4430
Prototype	0.5524
Exemplar	0.5137
Chroma	0.5848

Table 4: Average rank position of empirical Karajá foci for each model

Model	Average Rank Position
Bayesian	0.2064
Likelihood	0.2298
Prototype	0.2877
Exemplar	0.3023
Chroma	0.3199

foci do as well - suggesting that foci may in fact be derived from boundaries. This conclusion is reinforced by the observation that the Bayesian representativeness model predicts foci from boundaries fairly well in this non-canonical case, as well as more generally across the WCS.

## Conclusion

Focal colors, or best examples of color terms, lie at the center of the debate over color naming. These foci have traditionally been viewed either as the underlying source of color naming universals, or as derived from category boundaries that vary with local linguistic convention. In contrast, we have argued for a novel account of this disputed construct, in which focal colors show strong universal tendencies, but are nonetheless derived from category boundaries, as the most representative members of categories. In support of this proposal, we have shown that an existing Bayesian model of representativeness can predict the distribution of focal colors in the world's languages, from category extensions. This account synthesizes traditionally opposed views of color naming (Kay & McDaniel, 1978; Roberson et al., 2000), and accounts for data that challenge the traditional views.

Our proposal also coheres naturally with a recent theoretical account that explains universal tendencies in color naming in terms of optimally informative partitions of an irregularly shaped perceptual color space (Jameson & D'Andrade, 1997; Regier et al., 2007). Significantly, that view explains universal tendencies in color category boundaries without reference to a small set of focal colors, and it leaves the nature of focal colors unexplained. Our proposal fills that gap. Taken together, the two proposals suggest a single overall account of color naming: foci are optimally representative members of categories that are defined at their boundaries - and the boundaries themselves result from near-optimally informative partitions of color space.

**Acknowledgments.** We thank Paul Kay for his helpful comments. This work was supported by grants IIS-0845410 and IIS-1018733 from the National Science Foundation.

## References

- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.
- Cook, R., Kay, P., & Regier, T. (2005). The world color survey database. *Handbook of categorization in cognitive science*.
- Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis*. Chapman & Hall/CRC press.
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review*, 103, 592–596.
- Heider, E. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, 93(1), 10–20.
- Jameson, K., & D'Andrade, R. (1997). *Color categories in thought and language*. Cambridge University Press, Cambridge, UK.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.
- Kay, P., & McDaniel, C. (1978). The linguistic significance of the meanings of basic color terms. *Language*, 610–646.
- Kay, P., & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, 100(15), 9085–9089.
- Lindsey, D., & Brown, A. (2006). Universality of color names. *Proceedings of the National Academy of Sciences*, 103(44), 16608–16613.
- Regier, T., Kay, P., & Cook, R. (2005). Focal colors are universal after all. *Proceedings of the National Academy of Sciences of the United States of America*, 102(23), 8386–8391.
- Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences of the United States of America*, 104(4), 1436–1441.
- Regier, T., Kay, P., & Khetarpal, N. (2009). Color naming and the shape of color space. *Language*, 85(4), 884–892.
- Roberson, D., Davidoff, J., Davies, I., & Shapiro, L. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*, 50, 378–411.
- Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129(3), 369–398.
- Tenenbaum, J. B., & Griffiths, T. L. (2001). The rational basis of representativeness. In J. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 1036–1041).

# Gestures in Communication through Line Graphs

Cengiz Acartürk (ACARTURK@Metu.Edu.Tr)

Özge Alaçam (OZGE@Metu.Edu.Tr)

Cognitive Science, Informatics Institute  
Middle East Technical University, 06800, Ankara, Turkey

## Abstract

Line graphs are widely used in communication settings, for conveying information about states and processes that unfold in time. The communication is achieved by the contribution of other modalities than graphs, such as language and gestures. In a set of experimental investigations, we analyzed the production and comprehension of gestures during communication through line graphs. The findings reveal a systematic use of gestures as well as the limitations of cognitive resources due to the split of attention between the modalities.

**Keywords:** Gesture production; gesture comprehension; graph comprehension; line graphs.

## Line Graphs in Time Domain

Line graphs represent statistical data, most often the relationship between two domain variables. In line graphs, line segments are used for representing the mapping between the values. When used in time domain, line graphs represent the mapping between the values of the domain variable and time. From the perspective of human comprehension, line graphs in time domain have a peculiar characteristic: they represent not only statistical data but also *states* and *processes* that unfold in time, by providing perceptual cues for continuation (Figure 1).

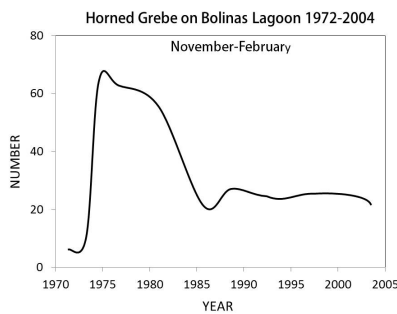


Figure 1: Sample population graph from PRBO (2012); redrawn based on the original.

Accordingly, the population graph in Figure 1 does not only represent the mapping between years and the population of the bird species but also leads to a conceptualization of how population *increases*, *decreases* or *remains stable* in certain periods of time.<sup>1</sup>

<sup>1</sup> Line graphs are generated based on a set of assumptions that specify the way the data points are represented by lines. For instance, according to the original source (PRBO, 2012), the line

Human conceptualization through statistical graphs has been a topic of interdisciplinary research since the past 30 years. The research on graph comprehension has covered a broad range of analyses including the investigations on perceptual processes of graph comprehension (e.g., Cleveland & McGill, 1985), analysis from the perspective of psychology and usability (e.g., Kosslyn, 1989), cognitive models (Lohse, 1993; Peebles & Cheng, 2002), educational psychology and instructional design (Winn, 1987; Mautone & Mayer, 2007). On the other hand, the research on modalities that accompany graphs, such as language and gesture in communication through graphs, has been scarce except for a few studies (e.g., Gerofsky, 2011, on gestures in graphs of polynomial functions). Concerning the relationship between language and gestures, gestures have been considered as having a key role in organizing, conveying spatial information, and preventing decay in visuospatial working memory (Hostetter & Alibali, 2010), thus having the potential to promote learning in educational contexts (Goldin-Meadow, 2010). Analyzing the relationship between graphical cues, language and gestures, the present study investigates communication through line graphs from the perspective of multimodal interaction.

## Communication through Line Graphs

Graphs are abundant both in spoken communication settings (e.g., classroom settings) and in written communication settings (e.g., newspaper articles). Communication through graphs is achieved by means of the contribution of several modalities: language (both in written form and in spoken form), graphical cues in written communication settings, and gestures in spoken communication settings. The previous research on multimodal comprehension reveals a frequent use of spatial terms that convey spatial information in communication through line graphs (Habel & Acartürk, 2007). Moreover, in spoken communication, people tend to produce more gestures when they perform tasks that involve spatial information, compared to tasks with no spatial information (Alibali et al., 2001; Trafton et al., 2006; Hostetter & Sullivan, 2011). Consequently, in communication through line graphs, humans frequently produce gestures that accompany spoken language.

graph in Figure 1 was generated by applying a local regression method called Loess smoothing on data points. The resulting spatial aspects of line graphs, such as smoothness, influence humans' interpretation of the states and processes (Acartürk et al., 2008), a topic beyond the scope of the present study.

Gestures in communication are of different types: the most commonly used ones are deictic (or pointing) gestures and iconic (or representational) gestures. Deictic gestures show objects, people and places, whereas iconic gestures are representations of shape of an object or an action (Özçalışkan & Goldin-Meadow, 2005). In communication settings, deictic gestures facilitate achieving joint attention on objects, whereas iconic gestures overlap with spatial tasks (Alibali et al., 2001; Trafton, et al., 2006). In communication through line graphs, humans may produce both deictic gestures and iconic gestures. It is also not unusual that humans emphasize certain aspects of processes and states represented by line graphs, such as a specific increase, a peak or a stable period of the domain value, in addition to emphasizing an overall pattern. Graphical annotations (also called graphical cues) on graph lines are generally used for this purpose.

The major focus of the present study is to investigate gestures in communication through line graphs, both from a production perspective and a comprehension perspective. For a systematic analysis, we limited the domain of investigation to the relationship between gestures and graphical cues in line graphs (rather than the overall pattern of the graph line). In the first step of the analysis, one group of participants produced gestures during a verbal description task (Experiment 1). We considered the produced gestures as human interpretations of the structural aspects of the states and processes represented by the graphs. The gestures produced by the participants of Experiment 1 were used for designing the stimuli for a comprehension experiment (Experiment 2). This approach resembles what has been termed the “3Ps (Preference-Production-Performance) program” as an empirical method for selecting appropriate representations for abstractions (Kessell & Tversky, 2011). The two approaches are similar; in that, both aim to perform an empirical investigation of the representations rather than leaving the decision for selecting the appropriate representation to intuitions of the graphic designer. Instead of graphic representations, however, we investigated gestures in communication through graphs in a set of consecutive analyses (i.e., the outcome of Experiment 1 was used for preparing the stimuli set in Experiment 2).

## Experiment 1

In Experiment 1, the participants presented verbal descriptions of annotated graphs. Spontaneous gestures of the participants were analyzed in terms of the relationship between the type of the graphical cue and the gesture type.

### Participants, Materials and Design

A total of seven participants (*Mean age* = 25.4, *SD* = 3.78) who were graduate students or teaching assistants from the Faculty of Education, Middle East Technical University (METU) participated in the experiment, five of which reported having teaching experience. The experiment language was Turkish, which was the native language of the

participants. The participants were asked to imagine themselves in an online meeting, in which their task was to present single-sentence summaries of annotated graphs to the audience. According to the scenario, the audience was able to see the participant (i.e., the presenter) but not the graphs. Therefore, the presenter first investigated the graph displayed on a computer screen, then s/he turned towards the audience (an audience picture displayed on another computer screen), and then presented a single-sentence summary of the graph. The participants were not informed that their gestures were in the focus of the experiment. Each participant presented the single-sentence summaries for 14 annotated graphs, thus generating 14 video recordings per participant. The graphs represented populations of bird species in a lagoon. Each graph involved a graphical annotation that emphasized a certain aspect of the information represented, such as a specific increase or a peak. In particular, three types of annotations were used.

- Process annotation: A diagonal arrow that emphasized a specific increase or a decrease.
- Durative state annotation: A horizontal arrow that emphasized a specific period of constant value.
- Punctual state annotation: A point-like circle that emphasized a specific value such as a peak value.

The 14 stimuli involved 2 graphs for familiarization of the participant to the task. The remaining 12 stimuli involved 6 punctual state annotations (2 for the start point of the lines, 2 for middle and 2 for the endpoint of the lines), 4 (diagonal) process annotations and 2 (horizontal) durative state annotations (Figure 2).

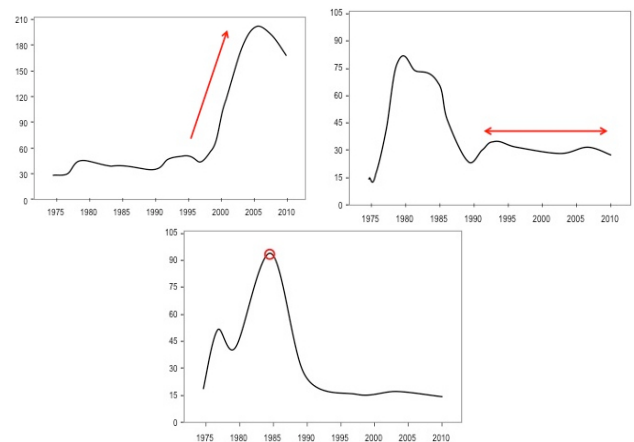


Figure 2: Sample annotated graphs with a process annotation (upper left), a durative state annotation (upper right), and a punctual state annotation (bottom).

Following Gerofsky (2011), we employed the coding scheme proposed by Creswell (2007) for the analysis of (14 graphs x 7 participants) 98 experiment protocols. The Noldus Observer XT event logging software was used for coding. Two coders analyzed the protocols according to the following criteria: For each gesture in the video recording, the coder first classified the gesture in terms of its directionality: having no gesture, no direction, being



vertical, horizontal, diagonal or other.<sup>2</sup> Then the coder identified the following features of each gesture: size (small or big), palm direction (up, down or front), speed (slow or fast) and start position (low, middle or high). In the present study, we focus on the directionality of gestures by leaving the analysis of other features to an extended study. One coder initially coded the entire data, and a second coder, who was blind to the hypothesis, carried out 57% of the dataset. Interrater reliability between coders was calculated by Cohen's kappa. The results revealed an agreement value of .78. According to Landis and Koch (1977), a value above .61 indicates substantial interrater agreement.

## Results

The participants gestured in 86% of the protocols. This number is close to what Hegarty et al. (2005) reported: the participants gestured when they described solutions to mental animation problems in 90% of the cases.<sup>3</sup> Pearson's chi square test and follow-up McNemar tests were conducted to investigate the relationship between the annotation type and the gestures produced by the participants. The test showed a significant effect of annotation type on gesture,  $\chi^2 = 48.1$ ,  $p < .05$ . In particular, for the graphs with process annotations, the participants produced more vertical and diagonal gestures compared to both horizontal gestures,  $\chi^2 = 15.7$ ,  $p < .05$ , and gestures with no direction,  $\chi^2 = 5.88$ ,  $p < .05$ . On the other hand, they produced more horizontal gestures compared to other types of gestures for durative states,  $\chi^2 = 4.08$ ,  $p < .05$ . Finally, for punctual annotations, more non-directed pointing gestures were produced compared to vertical gestures,  $\chi^2 = 16.5$ ,  $p < .05$ , to horizontal gestures,  $\chi^2 = 26.0$ ,  $p < .05$ , and to diagonal gestures,  $\chi^2 = 20.8$ ,  $p < .05$ .

These findings show that, in terms of the categorization of the gestures (cf. McNeill, 2005; Özçalışkan & Goldin-Meadow, 2005) the participants produced iconic gestures for process annotations and durative state annotations. On the other hand, for punctual state annotations, they produced pointing gestures that were ambiguous between iconic (because the pointing gesture was representational) and deictic (by definition).

## Experiment 2

The findings obtained in Experiment 1 suggest that humans produce a specific type of gesture depending on the emphasized aspect of the information represented in the graph. Based on the results obtained in Experiment 1, we investigated comprehension of gestures by humans in

<sup>2</sup> The 'other' category involved beat gestures (simple up-and-down movements without semantic information) or more complex gestures like the combination of vertical, horizontal or diagonal movements.

<sup>3</sup> A further investigation revealed that the five participants who reported teaching experience gestured in 93% of the protocols whereas the two participants who reported no experience in teaching gestured in 68% of the protocols. The finding suggests a potential correlation between teaching experience and gesturing.

Experiment 2. For this, we prepared 14 video recordings in which a narrator presented a single-sentence summary of annotated graphs by producing a relevant gesture concurrently with the spoken description. The verbal description was a single-sentence summary for a graph with process annotation, a graph with durative state annotation or a graph with punctual state annotation. For example, for a graph with a process annotation, the narrator uttered the sentence "[t]he population of coot in the lagoon increased between 1980 and 1985" while producing an upward diagonal gesture that showed an increase. She uttered the sentence "[t]he sanderling population in the lagoon remained stable between 1975 and 1985" accompanied by a horizontal gesture for a graph with a durative state annotation. Finally, for a graph with a punctual state annotation, the narrator uttered the sentence "[t]here exists about 120 terns in the lagoon in the year 2010" accompanied by a pointing gesture (Figure 3). The duration of the video recordings was between 5.3 seconds and 8.6 seconds ( $M = 6.24$ ,  $SD = 0.95$ ).

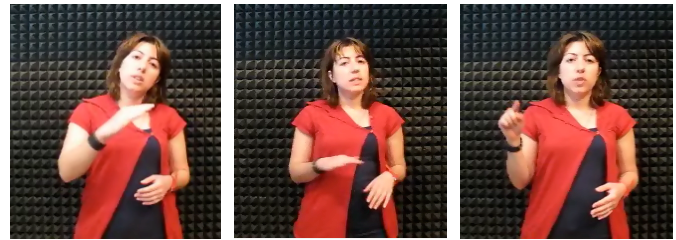


Figure 3: Snapshots from the video recordings with a diagonal gesture for a process annotation (left), a horizontal gesture for a durative state annotation (middle), a pointing gesture for a punctual state annotation (right).

Experiment 2 was conducted in three different conditions. In the first condition, the participants played the videos on the screen one by one and they listened to a single-sentence summary for each graph concurrently. In the second condition, the participants played the same video recordings but the sound was muted, therefore they interpreted what was presented on the screen only. In both the first condition and the second condition, we noted that participants' gaze shifted between the gesture and the face of the narrator. We interpreted this finding as a potential source of attention split. Therefore, in the third condition, we provided the participants with only gestures not the face of the narrator. In all conditions, the participants were asked to predict the described graph among a set of three alternative graphs.

### Condition 1: Concurrent Interpretation of Gestures and Language

**Participants, Material and Design.** Eleven participants ( $Mean\ age = 31.8$ ,  $SD = 5.1$ ), who were either graduate or undergraduate students of METU, participated in the experiment. Each participant was presented 14 video recordings (2 trials and 12 tests). After playing each recording, the participant was asked to choose the described

graph among three alternatives (the alternate graphs were the same except for the graphical annotation). After submitting each choice, the participant reported a subjective evaluation for confidence (“*How confident are you about your judgment?*”) by using a 1 to 3 scale (1 showing a low confidence, 3 showing a high confidence; Beattie and Shovelton, 1999). The stimuli were displayed on a Tobii non-intrusive 120 Hz eye tracker, integrated into a 17” TFT monitor with a resolution of 1024x768 pixels. The spatial resolution and the accuracy of the eye tracker were 0.25° and 0.50° respectively. No time limit was set for the answers. The order of presentation of the stimuli was randomized.

**Results.** The participants exhibited high success rates in predicting the annotated graphs, for all three types of gestures, i.e. the process gesture ( $M = 1.0$ , i.e. 100%), the durative state gesture ( $M = 1.0$ ) and the punctual state gesture ( $M = .93$ ,  $SD = 0.01$ ). The results of an ANOVA test revealed a significant difference between the gesture types,  $F(2, 20) = 5.17$ ,  $\eta^2 = .36$ ,  $p < .05$ : the success rate in punctual states was lower than the other two gesture types. A comparison of the confidence scores reported by the participants, however, revealed no significant difference between the gesture types  $F(2, 20) = 1.86$ ,  $\eta^2 = .16$ ,  $p > .05$ . Finally, the participants spent the longest time to answer punctual state questions ( $M = 7.03$  seconds,  $SD = 2.97$ ), which was longer than both processes ( $M = 5.27$  seconds,  $SD = 1.69$ ) and durative states ( $M = 6.65$  seconds,  $SD = 2.97$ ),  $F(2, 20) = 3.61$ ,  $\eta^2 = .27$ ,  $p < .05$ , without a significant difference between the last two.

### Condition 2: Interpretation of Gestures

The first condition of the experimental investigation employed the most naturalistic setting for an online communication environment: the participants listened to the narrator when she produced the gestures concurrently. In other words, both modalities (i.e., language and gesture) were available to the participants. Therefore, it is not possible to analyze the role of language and gestures separately in comprehension of the presented stimuli. The participants might have used the linguistic information to predict the graph without taking the gestures into account. In the second condition of the study, we asked the participants to predict the described graphs by displaying the video recordings with the sound muted.

**Participants, Material and Design.** Eighteen participants, from METU participated in the experiment ( $Mean\ age = 21.1$ ,  $SD = 1.37$ ). They were presented the same video recordings but they did not hear the narrator. The same experimental procedure was applied as in the previous condition.

**Results.** The participants in Condition 2 exhibited high success rates for processes ( $M = .93$ ,  $SD = .11$ ) and durative states ( $M = .91$ ,  $SD = .19$ ) but a significantly lower success

rate for punctual states ( $M = .55$ ,  $SD = .22$ ),  $F(2, 34) = 25.4$ ,  $\eta^2 = .60$ ,  $p < .05$ . The difference between processes and durative states was not significant. The lack of the language modality resulted in significant differences between the three gesture types in confidence scores,  $F(2, 34) = 18.1$ ,  $\eta^2 = .51$ ,  $p < .05$ . The participants reported lower confidence scores for punctual states ( $M = 2.01$ ,  $SD = 0.42$ ) compared to both processes ( $M = 2.61$ ,  $SD = 0.33$ ) and durative states ( $M = 2.61$ ,  $SD = 0.47$ ). As in Condition 1, the mean response time of the participants in punctual states ( $M = 4.34$  seconds,  $SD = 1.74$ ) was longer than both processes ( $M = 2.66$  seconds,  $SD = 0.80$ ) and durative states ( $M = 2.88$  seconds,  $SD = 1.44$ ),  $F(2, 34) = 10.5$ ,  $\eta^2 = .38$ ,  $p < .05$ , without a significant difference between the last two.

### Condition 3: Attention Split between Gestures and Face

The findings obtained in Condition 1 and Condition 2 show that the lack of linguistic information results in lower success rates in predicting the answers; in particular, in punctual states. The analysis of the eye movements of participants revealed another finding about inspection patterns on the video recordings: the participants shifted their gaze between narrator’s gestures and face both in Condition 1 ( $M = 2.55$ ,  $SD = 0.28$ ) and in Condition 2 ( $M = 2.68$ ,  $SD = 0.35$ ), without a significant difference between the two groups of participants,  $F(1, 26) = 0.83$ ,  $p > .05$ , thus suggesting a potential source of attention split during comprehension. Therefore, a third group of participants were presented narrator’s gestures only, without face and sound.

**Participants, Material and Design.** Twenty-one participants ( $Mean\ age = 21.2$ ,  $SD = 2.37$ ) from METU participated in the experiment. The participants were presented the same stimuli except that the video recordings were cut from the top, so that only the gestures (but not the face) of the narrator were displayed. The same experimental procedure was applied as in the previous conditions.

**Results.** The participants showed high success rates for processes ( $M = .96$ ,  $SD = .10$ ) and durative states ( $M = 1.0$ ) but a relatively lower success rate for punctual states ( $M = .70$ ,  $SD = .19$ ),  $F(2, 40) = 32.0$ ,  $\eta^2 = .61$ ,  $p < .05$ , without a significant difference between processes and durative states. Confidence scores for punctual states ( $M = 2.31$ ,  $SD = 0.40$ ) were also significantly lower than both processes ( $M = 2.69$ ,  $SD = 0.30$ ) and durative states ( $M = 2.76$ ,  $SD = 0.37$ ),  $F(2, 40) = 14.7$ ,  $\eta^2 = .42$ ,  $p < .05$ . Finally, they spent the longest time to answer punctual state questions ( $M = 3.52$  seconds,  $SD = 1.18$ ), significantly different than both processes ( $M = 2.60$  seconds,  $SD = 1.04$ ) and durative states ( $M = 2.41$  seconds,  $SD = 1.02$ ),  $F(2, 40) = 8.84$ ,  $\eta^2 = .31$ ,  $p < .05$ , without a significant difference between the last two.

A comparison between the three groups of participants in the three conditions of Experiment 2 showed that the highest success rate (in predicting the correct annotated



graph that was described in the video recording) was obtained when the participants listened to the single-sentence description of the graphs while playing the video recording. The lack of the language modality, however, resulted in a decrease in success rates. On the other hand, helping the participants to focus on gestures only (by removing narrator's face from the view) resulted in an increase in the success rates (Figure 4).

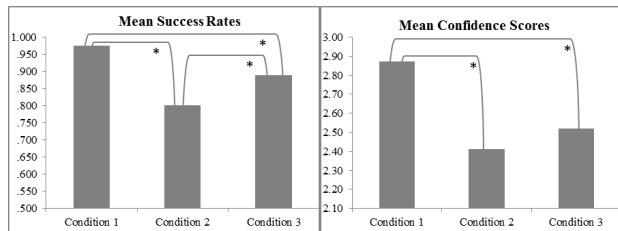


Figure 4: Mean success rates (left) and mean confidence scores (right) in Experiment 2.

For the comparison of the results obtained in the three conditions of Experiment 2, a Games-Howell test was applied since the number of samples for the three groups was not equal and the population variances were significantly different. The test returned a significant difference between the three groups of participants in their overall success rates,  $F(2, 47) = 17.2$ ,  $\eta^2 = .42$ ,  $p < .05$ . Finally, a comparison of the confidence scores between the participant groups showed that the lack of the language modality resulted in lower self-confidence of the participants about their predictions,  $F(2, 47) = 10.3$ ,  $\eta^2 = .30$ ,  $p < .05$  (Figure 4).

## Discussion

In two experiments, we investigated how humans produce gestures (Experiment 1) and comprehend gestures (Experiment 2) when they communicate through graphically annotated line graphs. In Experiment 1, the participants produced more frequent vertical and diagonal gestures to emphasize processes (e.g., *increase*, *decrease*) whereas they produced more horizontal gestures to emphasize durative states (e.g., *remain stable*). Those two types of gesture are known as *iconic gestures* and they overlap with representation of spatial information (Alibali et al., 2001; Trafton et al., 2006). For emphasizing punctual states (e.g., a *peak*), the participants produced *pointing gestures*. In Experiment 2, three groups of participants were presented video recordings and they were asked to predict the described graphs: the video recordings were designed based on the correspondence between diagonal gestures and processes, between horizontal gestures and durative states, and between pointing gestures and punctual states. When gestures were displayed concurrently with linguistic information (Condition 1), the participants showed a high success rate in all gesture types. When language modality was absent, however, they showed a lower success rate and lower self-confidence, in particular in punctual states. These

findings suggest a low efficiency of the pointing gesture (in the form of a deictic pointing gesture) in conveying information about punctual states. On the other hand, vertical and diagonal gestures were efficient in conveying information about processes. Horizontal gestures were efficient in conveying information about durative states. An explanation to these findings may be related to the major roles of iconic gestures and deictic gestures in communication. In contrast to iconic gestures that convey spatial information, the major role of pointing gestures is to attract the attention of the communication partner (McNeill, 2005; Özçalışkan & Goldin-Meadow, 2005). Consequently, further research is needed to identify more appropriate candidates for emphasizing punctual states in graphs. For instance, a circular movement of the index finger might be more appropriate for representing punctual states.

Another finding obtained in Experiment 2 was that participants' back and forth movement of their gazes between the gestures and the face of the narrator is a potential source of attention split during the course of comprehension. Although speech sound was absent (Condition 2) and therefore no linguistically useful information was provided by the narrator (except for the possibility of lip reading), the participants shifted their gaze several times between the narrator's face and the gestures. When the narrator's face was removed from the video recordings (Condition 3), an increase in success rates was observed compared to Condition 2, though the success rates were still lower than the ones obtained when the linguistic information was available (Condition 1). Although this is far from being a naturalistic setting for communication through graphs, the analysis of such boundary cases is necessary for understanding the contribution of separate factors in comprehension. In fact, the findings support the likelihood of the split of attention. A possible explanation may be sought in the domain of the intersection between cognitive science and instructional science, in which the previous studies show that the split of attention between information sources leads to degraded learning outcomes due to limited cognitive resources that are available for understanding the instructional material (Sweller et al., 1998; Mayer & Moreno, 1998). Consequently, the findings suggest that tasks demands may be high in communication through graphs; therefore, attention split should be avoided by, for instance, using small window sizes so that the communication partner is able to attend to both gestures and face in a single fixation.

## Conclusion and Future Work

In communication settings, humans produce gestures when they convey spatial information. As a consequence, in communication through line graphs, gestures are an indispensable part of communication. In this study we investigated how humans produce and comprehend gestures in communication through line graphs. We found that vertical and diagonal gestures efficiently convey information about processes such as increase and decrease,

and horizontal gestures efficiently convey information about durative states. However, pointing gestures are not efficient in conveying information about punctual states, possibly due to their concurrent role as deictic gestures in communication. Our future research will address finding more appropriate gesture candidates for punctual states. The future research will also address the investigation of the interaction between gestures and gradable (scalar) adjectives, gradable adverbs and spatial prepositional phrases and adverbials, e.g. *from*, *to*, and *between*, which are part of the vocabulary in communication through line graphs, in addition to state verbs and verbs of change.

**Acknowledgments.** We thank METU HCI Research and Application Laboratory for their technical support. We also thank four anonymous reviewers for their helpful comments and suggestions. We thank Christopher Habel for stimulating discussions on graphical annotations.

## References

- Acartürk, C., Habel, C., & Çağıltay, K. (2008). Multimodal comprehension of graphics with textual annotations: The role of graphical means relating annotations and graph lines. In J. Howse, J. Lee & G. Stapleton (Eds.), *Diagrammatic Representation and Inference: LNCS* (Vol. 5223, pp. 335-343). Berlin/Heidelberg: Springer.
- Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, 44, 169-188.
- Beattie, G., & Shovelton, H. (1999). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology*, 18, 438-462.
- Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 229, 828-833.
- Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2<sup>nd</sup> ed.). Sage Publications.
- Gerofsky, S. (2011). Mathematical learning and gesture: Character viewpoint and observer viewpoint in students' gestured graphs of functions. *Gesture*, 10(2-3), 321-343.
- Goldin-Meadow, S. (2010). When gesture does and does not promote learning. *Language and Cognition*, 2(1), 1-19.
- Habel, C., & Acartürk, C. (2007). On reciprocal improvement in multimodal generation: Co-reference by text and information graphics. In I. van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds.), *Proceedings of the workshop on multimodal output generation* (pp. 69-80). University of Aberdeen, UK.
- Hegarty, M., Mayer, S., Kriz, S., Keehner, M. (2005). The role of gestures in mental animation. *Spatial Cognition and Computation*, 5, 333-356.
- Hostetter A. B., Alibali M. W. (2010). Language, gesture, action! A test of the gesture as simulated action framework. *Journal of Memory and Language*, 63, 245-257.
- Hostetter, A. B., & Sullivan, E. L. (2011). Gesture production during spatial tasks: Its not all about difficulty. In L. Carlson, C. Hoelscher & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*. Austin, TX.
- Kessell, A. M. & Tversky, B. (2011). Visualizing space, time, and agents: Production, performance, and preference. *Cognitive Processing*, 12, 43-52.
- Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, 3(3), 185-226.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.
- Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human-Computer Interaction*, 8(4), 353-388.
- Mautone, P. D., & Mayer, R. E. (2007). Cognitive aids for guiding graph comprehension. *Journal of Educational Psychology*, 99(3), 640-652.
- Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology*, 90(2), 312-320.
- McNeill, D. (2005). *Gesture and thought*. University of Chicago Press.
- Özçalışkan, S., & Goldin-Meadow, S. (2005). Gesture is at the cutting edge of early language development. *Cognition*, 96, B101-B113.
- Peebles, D. J., & Cheng, P. C.-H. (2002). Extending task analytic models of graph-based reasoning: A cognitive model of problem solving with Cartesian graphs in ACT-R/PM. *Cognitive Systems Research*, 3, 77-86.
- PRBO (2012). *Waterbird Census at Bolinas Lagoon, Marin County, CA*. Public report by Wetlands Ecology Division, Point Reyes Bird Observatory (PRBO) Conservation Science. <http://www.prbo.org/cms/366>, retrieved on January 29, 2012.
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251-296.
- Trafton, J. G., Trickett, S. B., Stitzlein, C. A., Saner, L., Schunn, C. D., & Kirschenbaum, S. S. (2006). The relationship between spatial transformations and iconic gestures. *Spatial Cognition and Computation*, 6(1), 1-29.
- Winn, B. (1987). Charts, graphs, and diagrams in educational materials. In D. M. Willows & H. A. Houghton (Eds.), *The psychology of illustration* (Vol. 1, pp. 152-198). New York: Springer-Verlag.

# Distributional Learning of Vowel Categories is Supported by Prosody in Infant-Directed Speech

Frans Adriaans (adriaans@psych.upenn.edu) and Daniel Swingley (swingley@psych.upenn.edu)

Department of Psychology and Institute for Research in Cognitive Science, University of Pennsylvania  
3401 Walnut Street, Suite 400A, Philadelphia, PA 19104, USA

## Abstract

Infants' acquisition of phonetic categories involves a distributional learning mechanism that operates on acoustic dimensions of the input. However, natural infant-directed speech shows large degrees of phonetic variability, and the resulting overlap between categories suggests that category learning based on distributional clustering may not be feasible without constraints on the learning process, or exploitation of other sources of information. The present study examines whether mothers' prosodic modifications within infant-directed speech help the distributional learning of vowel categories. Specifically, we hypothesize that 'motherese' provides the infant with a subset of high-quality learning tokens that improve category learning. In an analysis of vowel tokens taken from natural mother-infant interactions, we found that prosody can be used to distinguish high-quality tokens (with expanded formant frequencies) from low-quality tokens in the input. Moreover, in simulations of distributional learning we found that models trained on this small set of high-quality tokens provide better classification than models trained on the complete set of tokens. Taken together, these findings show that distributional learning of vowel categories can be improved by attributing importance to tokens that are prosodically prominent in the input. The prosodic properties of motherese might thus be a helpful cue for infants in supporting phonetic category learning.

**Keywords:** Infant-directed speech; phonetic category learning; prosody; computational modeling.

## Introduction

Infants in the first year of life develop knowledge of the phonetic categories that make up the consonants and vowels of their native language (e.g., Werker & Tees, 1984). The early age at which this takes place rules out learning accounts in which semantic contrast in phonologically similar words drives most category learning. As a result, it is assumed that infants learn phonetic categories using an implicit statistical clustering process that relies on separation of the categories in perceptual space. Indeed, 6- and 8-month-old infants have been found to form representations of two distinct categories (e.g., /d/ and /t/) when exposed to an artificially generated bimodal distribution on a distinguishing acoustic dimension, but not when exposed to a unimodal distribution (Maye, Werker, & Gerken, 2002; Maye, Weiss, & Aslin, 2008; see also Cristià, McGuire, Seidl, & Francis, 2011). Further evidence for the plausibility of distributional learning of phonetic category structure comes from analyses of infant-directed speech. Mothers appear to provide their infants with acoustic cues that support distributional learning of phonetic categories (Werker et al., 2007). In particular, infant-directed speech is characterized by expansion of the F1-F2 vowel formant space, which could enhance the separability of vowel categories (Kuhl et al., 1997). Several studies

have used approximations of infant-directed speech tokens as input to computational procedures (such as multivariate Gaussian mixture models) that succeed in learning vowel categories, suggesting that distributional learning could be feasible for infants (de Boer & Kuhl, 2003; McMurray, Aslin, & Toscano, 2009; Vallabha, McClelland, Pons, Werker, & Amano, 2007).

Some caution is appropriate in interpreting these findings. Studies that show the usefulness of distributional cues for category learning have, in large part, been based on analyses (and simulations) of vowel tokens that were elicited in a laboratory setting, and that occurred in a small number of words or nonwords. It is possible that maternal speech under these conditions is different from maternal speech in quotidian home contexts. Analyses of natural, unscripted infant-directed speech recordings show that vowel distributions are highly variable, and that overlap between categories poses a substantial problem for distributional category learning (Swingley, 2009). One possibility suggested by this result is that infants' learning of phonetic categories is guided by additional sources of information, such as the emerging lexicon (Feldman, Griffiths, & Morgan, 2009; Swingley, 2009).

Another possibility, explored here, is that infants are able to succeed in category learning because they have a bias to attend to some tokens more than to others, and that these salient tokens are clearer instances of their categories. If so, the difficulty of distributional category learning is overestimated by considering the whole mass of experienced speech sounds. This notion is indirectly supported by studies showing that infants prefer "motherese" speech over adult-directed speech. Across different languages motherese is characterized by acoustic exaggeration, including higher overall pitch, greater intonation contours, and longer durations (Fernald et al., 1989; Grieser & Kuhl, 1988; Kuhl et al., 1997). These properties have been found to modulate infants' attention, and possibly facilitate language learning by enhancing infants' speech discrimination skills (Fernald & Kuhl, 1987; Karzon, 1985; Liu, Kuhl, & Tsao, 2003; Trainor & Desjardins, 2002).

It remains to be demonstrated that motherese effectively guides the infant's attention to those vowel tokens that are most useful for category learning. Computational models that aim to explain category learning are typically fit to isolated, equally weighted vowel tokens (de Boer & Kuhl, 2003; Vallabha et al., 2007). Such models overlook prosodic context which might make certain vowel tokens more attractive than others, and which thus potentially affects the learnability of vowel categories.

The current study examines the relation between prosodic exaggeration and vowel learning from infant-directed speech. Specifically, we hypothesize that motherese provides the infant with a subset of high-quality learning tokens that improves distributional category learning. First, we analyze prosodic determinants of vowel expansion within infant-directed speech, thereby attempting to predict which vowel tokens in the infant's speech input could be particularly beneficial for phonetic category learning. Second, we simulate the distributional learning of phonetic categories in order to examine whether prosodic focus helps in discovering category structure in cases of large overlap between categories. Importantly, analyses and simulations are done on realistic data, using vowel tokens taken from recordings of natural mother-infant interactions. We thus provide a test of distributional learning in a setting that acknowledges the variability and complexities that are found in real everyday speech.

### Vowel Expansion in Infant-Directed Speech

Earlier studies on vowel expansion compared speech directed to adult listeners and speech directed to infant listeners (Kuhl et al., 1997). While infant-directed speech is often hyperarticulated compared to adult-directed speech, the mechanisms underlying vowel expansion in infant-directed speech are not yet fully understood. It seems likely that the prosodic exaggeration notable in infant-directed speech has an effect on vowel expansion. Here we explore this possibility by asking whether prosodically prominent vowels in infant-directed speech are hyperarticulated relative to parts that are not prosodically highlighted. In analyses of recordings of natural mother-infant interactions, we examine whether prosodic focus predicts vowel expansion (see also Mo, Cole, & Hasegawa-Johnson, 2009). We examine expansion in tokens that were labeled to have focus by human assessors (what we define as "annotated focus"), and also in tokens that were defined as exaggerated on acoustic grounds (higher pitch, greater pitch change, and longer duration; what we define as "acoustic focus"), to determine whether such vowels are more differentiable. Evidence of vowel expansion at prosodically predictable locations in infant-directed speech would indicate that attention to prosody could aid in vowel category learning.

### Methods

Vowel expansion was examined by analyzing vowel productions by one mother ('f1') in the Brent corpus (Brent & Siskind, 2001), available through CHILDES (MacWhinney, 2000). These recordings consist of natural, unscripted infant-directed speech and therefore have no restrictions on the words or vowel types that may occur. Formant (F1, F2) measurements were obtained and hand-checked for 1,166 vowel tokens. Tokens covered the monophthongs of American English (/i/, /ɪ/, /e/, /æ/, /ɑ/, /ʌ/, /ɔ/, /ʊ/, /u/). Measurements taken at 33% and 50% of the vowel's duration were averaged and transformed into *z* scores to neutralize scale differences. Vowel expansion was measured by calculating the

Euclidean distance of each token to the center of the mother's vowel space (Bradlow, Torretta, & Pisoni, 1996). In order to measure prosodic prominence in infant-directed speech each vowel token was judged by a human assessor who indicated whether the vowel occurred in a syllable that the mother was trying to emphasize (*focus* vs. *no focus*). Potential acoustic correlates of focus that were considered were: duration (logarithm of the absolute duration in ms.), pitch (F0 averaged over 33% and 50% measurements), and pitch change (the absolute value of the difference in F0 between measurements at 33% and 50%). The label of "acoustic focus" was assigned to vowels that exceeded the *z*-score of 0.5 for at least one of the three dimensions.

### Results

Table 1 shows the number of focused and unfocused tokens for each vowel. The annotated-focus set contained 336 vowel tokens (28.8% of the total set). The acoustic-focus set had 543 tokens (46.6% of the total set). Figure 1 shows the mean formant frequencies of vowels in focused and unfocused position. Vowels in focused position were further away from the center of the vowel space than vowels in unfocused position.<sup>1</sup>

Stepwise linear regression analyses revealed that annotated focus is a significant predictor of the vowel's distance from the center of the vowel space, independent of vowel type (adjusted  $R^2 = 0.4221$ ; vowel\*\*\*, focus\*\*, vowel:focus *ns*). Vowels in syllables with annotated focus were thus hyperarticulated relative to vowels in unfocused syllables. This confirms the intuition that in natural infant-directed speech mothers exaggerate certain vowels by marking them with sentence focus. Interestingly, vowel expansion did not manifest itself through stretching of the triangle defined by the "point vowels" (/i/, /a/, /u/), but rather followed a consistent pattern of expansion throughout the entire set of monophthongs.

The tokens that had acoustic focus showed very similar results. Stepwise regression revealed that acoustic focus is a significant predictor of vowel expansion (adjusted  $R^2 = 0.4300$ ; vowel\*\*\*, focus\*\*\*, vowel:focus\*). These results indicate that whether infants are able to judge focus (as our annotators did), or whether they simply pay attention to tokens that have extreme values on prosodic dimensions (i.e., "acoustic focus"), the tokens that have focus show expansion, and are thus possibly particularly helpful for the learning of phonetic categories.

In sum, vowels that are prosodically exaggerated might be particularly useful for phonetic learning because they have distributional properties that enhance the separability of vowel categories. The overlap between categories, however, is still substantial. It thus remains to be demonstrated that prosodic highlighting makes a meaningfully large difference in the learnability of vowel categories.

<sup>1</sup>The exception was /ɔ/ and /ʊ/ in the acoustic-focus set. The means of these vowels are unreliable due to their low frequency of occurrence in the data set. (See Table 1.)

Table 1: Frequency of occurrence of vowels in focused and unfocused position.

	/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ʌ/	/ɔ/	/ʊ/	/u/	Total:
	(182)	(320)	(163)	(139)	(112)	(130)	(21)	(22)	(77)	(1,166)
Focus (annotated)	41	72	51	67	32	36	12	5	20	336
No focus (annotated)	141	248	112	72	80	94	9	17	57	830
Focus (acoustic)	105	112	65	95	55	45	16	10	40	543
No focus (acoustic)	77	208	98	44	57	85	5	12	37	623

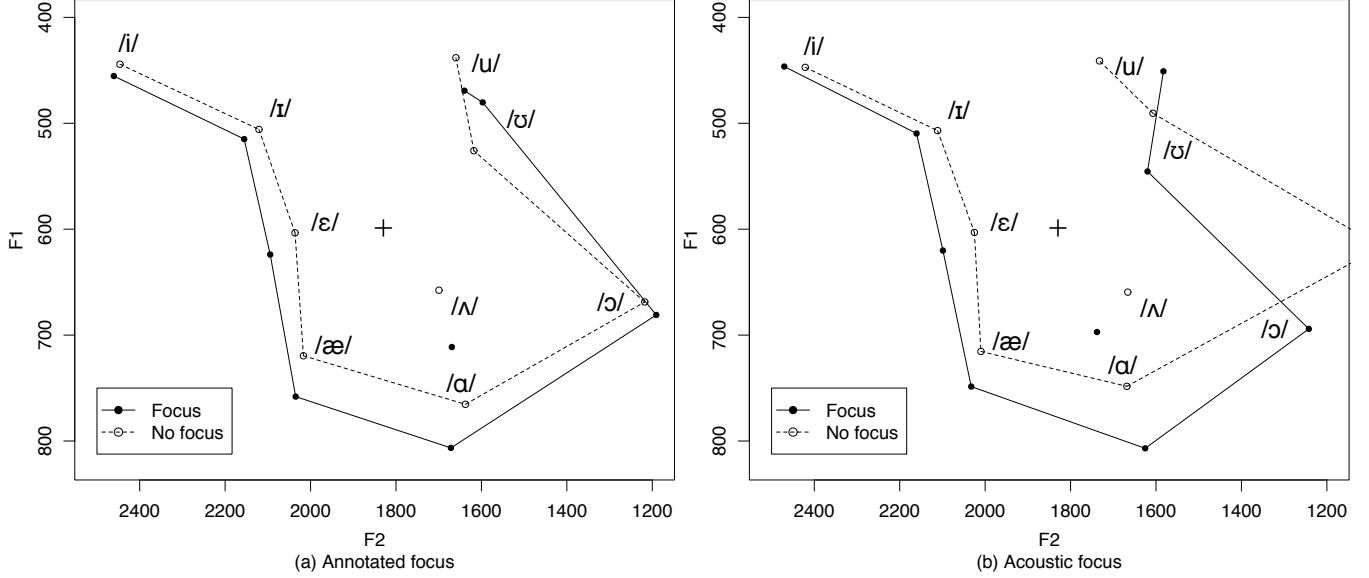


Figure 1: Vowel expansion within infant-directed speech. ‘+’ indicates the center of the mother’s vowel space.

## The Learnability of Vowel Categories

In order to see if prosodically highlighted vowels would be beneficial to infant language learners, we simulate the distributional learning of vowel categories from infant-directed speech. In particular, we examine whether prosodic focus helps in discovering category structure in cases of large overlap between categories. If distributional models of vowel learning show improved performance when trained on prosodically defined subsets of vowel data, then this would constitute evidence that the prosodic properties of motherese support phonetic category learning.

## Methods

The learnability of vowel categories is simulated for two different sets of vowels: /i/, /ɪ/, /ɛ/ and /ɛ/, /æ/, /ɑ/. These sets were chosen because they each contain three vowels that are close in the F1-F2 formant space. As a consequence, the overlap between categories is large, and the learning of these categories poses a substantial problem for distributional learning models. In line with earlier work on computational modeling of phonetic category learning (e.g., de Boer & Kuhl, 2003; McMurray et al., 2009; Vallabha et al.,

2007), we treat categories as multivariate Gaussian distributions. The learning problem is characterized as estimating the parameters (means, covariances and mixing proportions) for these distributions. In our case, categories are defined as 2-dimensional distributions (the  $z$  scores of the first and second formants). Data points are assigned to the category that has the maximum likelihood for that point. Parameters of the Gaussian distributions are estimated using the EM algorithm (Dempster, Laird, & Rubin, 1977) as implemented in the *MCLUST for R* software package (Fraley & Raftery, 2006). All models reported below were trained to discover three ellipse categories. Since vowel ellipses are known to vary in volume, shape, and orientation (e.g., Hillenbrand, Getty, Clark, & Wheeler, 1995), the models were given no information or constraints with respect to volume, shape, or orientation.

In order to assess whether focused tokens were helpful for category learning, models were trained on a subset of the data (either the annotated-focus set or the acoustic-focus set). The Gaussian distributions that were estimated from these subsets were subsequently used to classify all vowel tokens in the data set. We predicted that Gaussian mixture models trained

on a relatively small set of prosodically prominent vowel tokens would provide a better classification of the data than Gaussian mixture models that were trained on the complete set of vowel tokens. Performance of the unsupervised clustering models was assessed by comparing their classification accuracy to a supervised learner that learned three Gaussian categories based on actual vowel category labels. The supervised learner represented an upper bound on the classification accuracy that can be obtained given the maximum likelihood classification criterion that is imposed on the overlapping Gaussian distributions.

## Results

Table 2 shows the classification accuracy for models trained using all tokens, annotated-focus tokens, acoustic-focus tokens, and all tokens’ category labels (this last being the supervised “ideal”). The first thing to note is that the classification accuracy of the supervised learners was below 80%, confirming that overlap between categories was substantial. Considering the unsupervised “All tokens” model, the 12- to 15-percentage-point decline relative to the supervised model shows that the categories are not trivially detectable in the distributions.<sup>2</sup> Using vowel tokens annotated as focused aided accuracy to a small degree in the i-ɪ-ε data set, a result that nevertheless reveals some utility to focus marking given that this model was trained on only 164 data points rather than the entire dataset (which consisted of a total of 665 i-ɪ-ε tokens). However, for the ε-æ-ɑ data set the clustering algorithm was unable to fit a model to the annotated-focus tokens. We believe that this is due to the small size of the annotated-focus data set for ε-æ-ɑ ( $n = 150$ , with only 32 tokens for /ɑ/, see Table 1). Thus, focused vowels are, in at least some cases, variable enough that category solutions are difficult to determine when the quantity of data is very small.

Training on the bigger set of acoustic-focus tokens helped learning substantially, bringing the model within 3 percentage points of the supervised model in the i-ɪ-ε data set, and within 6 percentage points in the ε-æ-ɑ data set. Models that were trained on tokens that were acoustically prominent (long duration, high pitch, greater pitch movement) thus showed substantial classification improvement as compared to models that were prosodically uninformed. To illustrate the performance of different learning models, we display the i-ɪ-ε data along with the classifications that are predicted by different models in Figure 2. Figure 2 shows that only the acoustic-focus training set is able to predict three clearly distinct categories.

As it turns out, tokens that have focus or show acoustic exaggeration have a positive effect on the unsupervised learning of vowel categories. Importantly, these high-quality tokens are easily identifiable based on their prosodic properties. It is thus likely that these tokens are identifiable for infant language learners, and contribute to language learning.

<sup>2</sup>Such a decline is not found in models of the point vowels (i-ɑ-u) alone, for which we found accuracy > 90% for both the supervised

Table 2: Classification accuracy on two different sets of overlapping vowel categories.

Model	Accuracy	
	i-ɪ-ε	ε-æ-ɑ
All tokens	0.6060	0.6449
Annotated focus	0.6331	-
Acoustic focus	0.7008	0.7343
Supervised	0.7278	0.7947

## Discussion

In learning the phonetic categories of their native language, infants face large amounts of variability in the acoustic realizations of different vowel tokens. This poses a substantial problem for the purely bottom-up distributional learning of vowels. Here we presented one possible source of information that may guide phonetic category learning. If infants are able to detect high-quality learning tokens in the input, then they could make considerable progress in category learning. Motherese may play an important role in this process, by bringing such “high-quality” tokens to the infant’s attention through prosodic modifications of the speech stream.

In our clustering experiments, focus as annotated by human listeners was not as effective as “focus” estimated using simple, one-dimensional acoustic measures. It is possible that this difference derived from sample-specific gaps in the number or quality of human-annotated focus tokens for some vowel types; this cannot be ruled out without examining other samples. Furthermore, it is likely that annotators’ judgments of focus were, in some cases, based on their interpretation of the speaker’s intentions: an adult listener might judge a word as being the one the speaker wished to emphasize even if the phonetics were not particularly marked. Still, the superior performance of the model that learned from the tokens that were simply more extreme on at least one of the acoustic dimensions shows that the benefits of “motherese” prosodic highlighting do not depend on possession of a mature capacity for interpreting focus. Sensitivity to simple dimensions like duration or pitch goes a long way.

Infant-directed speech prosody, with its exaggerated prosodic variation, certainly captures infants’ attention, and this may be important for learning. Earlier studies have shown that pitch contours enhance infants’ discrimination skills, since contours increase the acoustic salience of formant frequencies (Trainor & Desjardins, 2002). Such perceptual salience is not taken into account by our model. Our results show that prosody has additional benefits. We find that acoustically exaggerated tokens show a different distribution in the F1-F2 space, with greater distances from the center and enhanced separability of categories. The picture that emerges from earlier studies, combined with the current findings, is

and unsupervised learner.

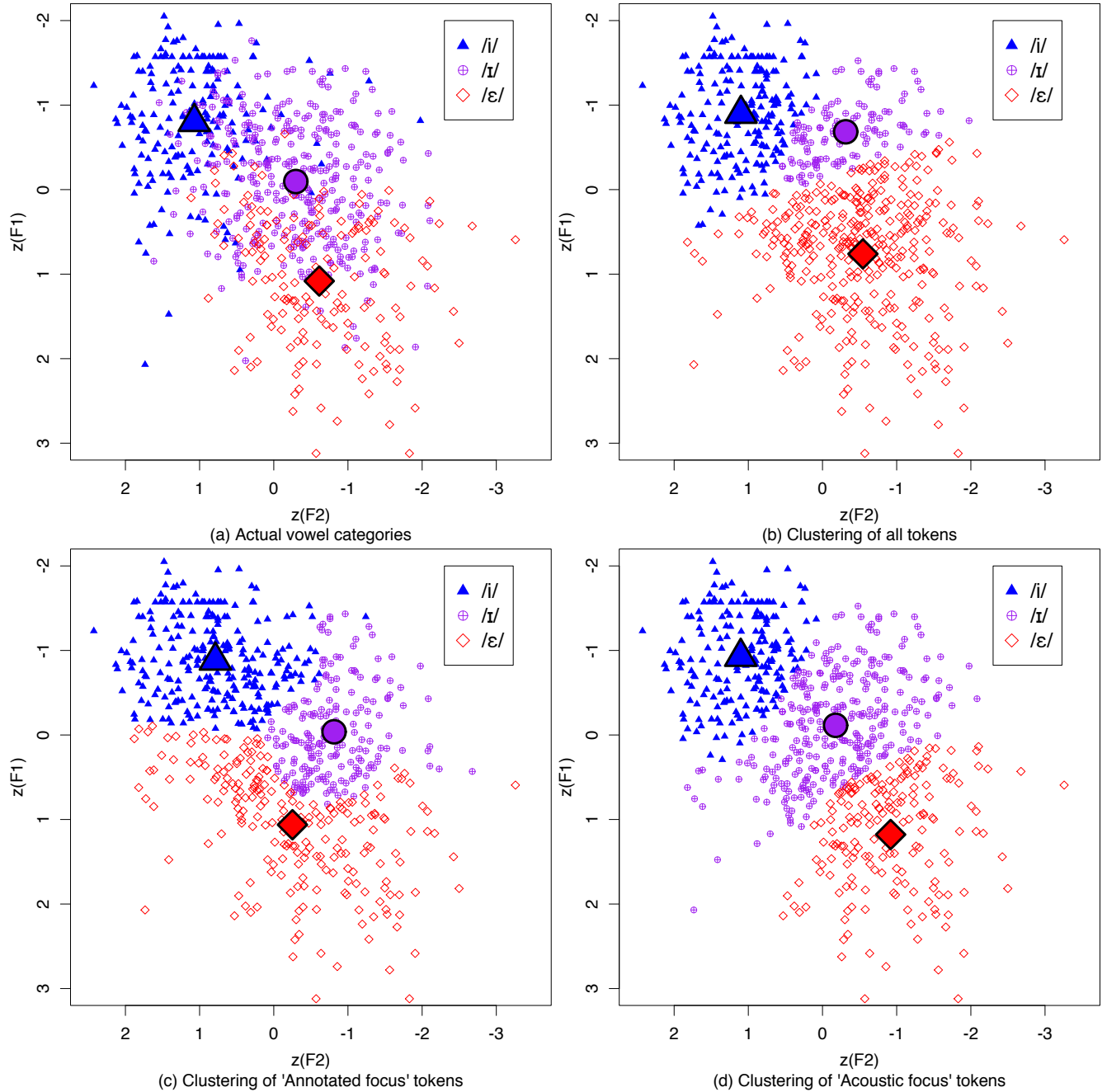


Figure 2: The i-i-ε data set with (a) actual categories, (b) predicted categories based on all tokens, (c) predicted categories based on focused tokens, and (d) predicted categories based on acoustically exaggerated tokens. The means are plotted for each (predicted) category.

that the exaggerated prosody of infant-directed speech may capture infants' attention to speech in a general fashion, and at the same time provide an enhanced speech signal that supports language learning – if infants' category learning favors attention to the most salient instances.

### Acknowledgments

This work was funded by the Netherlands Organisation for Scientific Research (NWO) grant 446.010.027 to F.A. and

NIH grant R01-HD049681 to D.S.

### References

- Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20, 255-272.
- Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cogni-*

- tion, 81, B33-B44.
- Cristià, A., McGuire, G. L., Seidl, A., & Francis, A. L. (2011). Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of Phonetics*, 39, 388-402.
- de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, 4, 129-134.
- Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1-38.
- Feldman, N. H., Griffiths, T. H., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (p. 2208-2213).
- Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, 10, 279-293.
- Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B. de, & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, 16, 477-501.
- Fraley, C., & Raftery, A. E. (2006). *MCLUST Version 3 for R: Normal mixture modeling and model-based clustering* (Tech. Rep.). Seattle, WA: University of Washington.
- Grieser, D. L., & Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology*, 24, 14-20.
- Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97, 3099-3111.
- Karzon, R. G. (1985). Discrimination of polysyllabic sequences by one- to four-month-old infants. *Journal of Experimental Child Psychology*, 39, 326-342.
- Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., et al. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, 277, 684-686.
- Liu, H.-M., Kuhl, P. K., & Tsao, F.-M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science*, 6, F1-F10.
- MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk, volume 2: The database* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, 11, 122-134.
- Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82, B101-B111.
- McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, 12, 369-378.
- Mo, Y., Cole, J., & Hasegawa-Johnson, M. (2009). Prosodic effects on vowel production: evidence from formant structure. In *Proceedings of Interspeech 2009* (p. 2535-2538).
- Swingle, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B*, 364, 3617-3622.
- Trainor, L. J., & Desjardins, R. N. (2002). Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychonomic Bulletin & Review*, 9, 335-340.
- Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America*, 104, 13273-13278.
- Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition*, 103, 147-162.
- Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7, 49-63.



# Emotion-Based Reinforcement Learning

Woo-Young Ahn<sup>1</sup> (ahnw@indiana.edu)

Olga Rasso<sup>1</sup> (rasso@indiana.edu)

Yong-Wook Shin<sup>2</sup> (shaman@amc.seoul.kr)

Jerome R. Busemeyer<sup>1</sup> (jbusemey@indiana.edu)

Joshua W. Brown<sup>1</sup> (jwmbrown@indiana.edu)

Brian F. O'Donnell<sup>1</sup> (bodonnel@indiana.edu)

<sup>1</sup>Department of Psychological and Brain Sciences, Indiana University

<sup>2</sup>Department of Psychiatry, Ulsan University School of Medicine

## Abstract

Studies have shown that counterfactual reasoning can shape human decisions. However, there is a gap in the literature between counterfactual choices in description-based and experience-based paradigms. While studies using description-based paradigms suggest participants maximize expected subjective emotion, studies using experience-based paradigms assume that participants learn the values of options and select what maximizes expected utility. In this study, we used computational modeling to test 1) whether participants make emotion-based decisions in experience-based paradigms, and 2) whether the impact of regret depends on its degree of unexpectedness as suggested by the current regret theory. The results suggest that 1) participants make emotion-based choices even in experience-based paradigms, and 2) the impact of regret is greater when it is expected than when it is unexpected. These results challenge the current theory of regret and suggest that reinforcement learning models may need to use counterfactual value functions when full information is provided.

**Keywords:** Decision making; Bayesian modeling; mathematical modeling; regret; reinforcement learning.

## Introduction

In our daily lives, we constantly face decisions to make and assess the costs and benefits of possible options (e.g., “Should I buy a lottery or just buy a snack with this money?”, “Should I buy Apple or Google stock?”). Usually we know only the outcome of our choices. On rare occasions, we also know what would have happened if we had made different choices (e.g., stock market). Having ‘complete feedback’ (or *full information*) under risk or uncertainty can evoke strong emotions such as regret or disappointment that are triggered by our capacity to reason counterfactually.

The effects of counterfactual reasoning have received much attention, and several theories have been proposed. A growing consensus suggests that disappointment and elation are elicited by comparison between different states (e.g., “my grant was not funded...”) whereas regret and rejoice come from comparison between different choices (e.g., “I should have married another person...”). Also, the unique aspect of regret is a feeling of *responsibility* that comes with negative outcomes from choices.

Among several theories of counterfactual decision-making, *decision affect theory* is regarded as one of the leading models (Mellers, Schwartz, & Ritov, 1999). Decision affect theory assumes that individuals make emotion-based choices and want to maximize subjective expected *pleasure* (or emotion)

rather than to maximize expected return. In decision affect theory, our emotional responses ( $R$ ) are based on obtained outcomes, relevant comparisons, and beliefs about the likelihood of the outcomes:

$$R \propto \text{Chosen Outcome Utility} + \text{Regret / Rejoice} + \text{Disappointment / Elation} \quad (1)$$

All counterfactual terms (regret, rejoice, disappointment, and elation) are weighted by their unexpectedness. Decision affect theory effectively explained various experimental results (Mellers et al., 1999) and Coricelli et al. (2005) used a modified version of the theory to examine the neural correlates of regret using description-based paradigms.<sup>1</sup>

Several studies have examined counterfactual decision-making using experience-based paradigms as well (Lohrenz, McCabe, Camerer, & Montague, 2007; Boorman, Behrens, & Rushworth, 2011; Hayden, Pearson, & Platt, 2009; Yechiam & Rakow, 2011). Although models used in the studies differ slightly from each other, all previous studies used reinforcement learning models, which assume that participants learn about chosen and foregone outcomes from trial-by-trial experience and then choose an option that has the highest expected value.

This study was developed from this gap in the literature: to explain choice behaviors in description-based paradigms with full information, researchers have assumed participants would make emotion-based choices. To explain choice behaviors in experience-based paradigms, researchers have assumed that participants learn the obtained and foregone payoffs and do not make emotion-based choices. We tested whether individuals make emotion-based choices in experience-based paradigms by building computational models for all competing hypotheses. This approach allowed us to quantitatively compare hypotheses in a rigorous way.

Another aim of the study was to test whether regret would be weighted by its unexpectedness (i.e., surprisingness). Mellers et al. (1999) claimed that “...unexpected out-

<sup>1</sup>In description-based paradigms, the outcomes of all options and their probabilities are provided to participants and participants rarely receive feedback. In experience-based paradigms, participants must learn the outcomes or their probabilities from their personal experience (Hertwig, Barren, Weber, & Erev, 2004).

comes have greater emotional impact than expected outcomes.” However, how would you feel given the following scenarios? In scenario 1, an Apple employee told you some inside information about Apple, which would increase its stock price. You believed that this was 80% reliable, but you did not buy the stock whose price sky-rocketed. In scenario 2, an untrustworthy looking stranger told you the same information. You believed he was 20% reliable, but you did not buy the stock, whose price sky-rocketed. According to Mellers et al. (1999), you would experience more regret in scenario 2. However, we hypothesized that scenario 1 would generate more regret because of the unique aspect of regret: a feeling of responsibility. Therefore, we predicted that regret would be weighted by its *expectedness* rather than its unexpectedness. Mellers, Schwartz, Ho, and Ritov (1997) showed that a smaller probabilities of disappointment/elation were associated with greater emotional response. Although Mellers et al. (1999) claimed that the effect of probability would be the same with regret/rejoice, no experiment has directly tested it to our knowledge.

In sum, we designed our experiment to test the following hypotheses. The first hypothesis proposes that participants will learn the chosen and fictive outcomes, compare all available options, and try to maximize their expected return (“Fictive Learning Alone”). The second hypothesis proposes that participants will make emotion-based decisions (i.e., maximize their expected subjective emotion) and their regret will be weighted by its unexpectedness (“Original Regret”). The third hypothesis proposes that participants will make emotion-based decisions and will weight their regret by its expectedness (“Modified Regret”). We designed our experiment to test these hypotheses.

## Method

### Participants

Nineteen healthy individuals (7 men, mean age = 23.0, SD=4.9) participated in the study. Electroencephalography (EEG) was continuously recorded from the scalp, but EEG findings are not reported in this paper. Participants were paid \$10/hr for participation and told that they would earn performance bonuses based on total points earned during the task. In reality, all participants received a fixed amount (\$5) as their bonus money (Lejuez et al., 2003). Study procedures were approved by the Indiana University’s Human Subjects Institutional Review Board.

### Task

All participants completed four separate gambling games, the order of which was randomly mixed for each participant. At the start of each game, participants were told that each game was independent of the previous game(s). In each game (90 trials/game), participants were asked to choose one of two options. One option was a safe option in which participants always won a fixed amount of points (e.g., 11). The other was a risky option in which participants won either larger (e.g., 26)

or smaller points (e.g., 1). The probability of winning larger points was fixed but unknown, and had to be learned from experience. The payoffs of both chosen and unchosen options were revealed on every trial (“full information”). The locations of the options were fixed within games, but randomized across games. Participants were encouraged to choose an option that would maximize their gain. Payoffs were distributed so that the long-term expected values of two options were the same (see Table 1).

Table 1: The payoff distributions of games 1-4. Note that the (long-term) expected values of the safe option (M) and the risky option are the same. M: points of the safe option, L: low (smaller) points, H: high (larger) points, %H: the probability of winning larger points. SD: standard deviation.

Game	M	Risky Option				
		L	H	%H	Mean	SD
1	12	1	56	0.2	12	22.0
2	11	1	26	0.4	11	12.3
3	10	1	16	0.6	10	7.4
4	9	1	11	0.8	9	4.0

The timing and presentation of a trial is illustrated in Figure 1. Each trial started with a message (“WAIT”), which was presented for 1-1.5s. After two options were presented, the participant had 2s to select an option by pressing buttons corresponding in a spatially compatible way to the options. The color of the chosen option remained changed for .6s, and the payoffs of both options appeared for 1s.

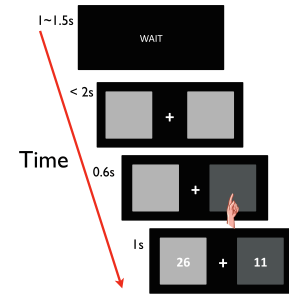


Figure 1: Time course of the gambling task.

### Computational Modeling

Three hypotheses (1. Fictive Learning Alone, 2. Original Regret model, 3. Modified Regret model) were implemented as three distinct reinforcement learning models. They utilized identical learning (probability learning) and choice rules (softmax), but used different value functions. Due to the specific design of the task (only 2 possible payoffs of the risky option in each game), it was assumed that participants would learn the probability of a larger payoff of the risky option (probability learning). In the delta rule (Rescorla & Wagner,

1972), the probability of a larger payoff (H) (the risky option) on the next trial  $t + 1$ ,  $Pr_H(t + 1)$ , is updated as follows:

$$Pr_H(t + 1) = Pr_H(t) + \gamma \cdot [Y(t) - Pr_H(t)] \quad (2)$$

Here  $\gamma$  is the learning rate ( $0 < \gamma < 1$ ) and  $Y(t)$  is the outcome (1 if H, 0 if L) of the current trial  $t$ . We assumed no learning occurred about the safe option because its payoff was always the same (e.g., 11) in a given game. We assumed that the choice of a risky or safe option did not affect the learning rate.<sup>2</sup>

Action selection was implemented via the Luce choice rule (a.k.a. softmax) (Luce, 1959). The inverse temperature parameter ( $\theta$ ) determines the sensitivity of the choice probabilities to the action values. We employed a trial-independent choice rule (Yechiam & Ert, 2007), where  $\theta = 3^c - 1$  ( $0 < c < 5$ ). When  $c$  approaches zero, choices become completely random (exploratory). When  $c$  becomes large, choices become deterministic (exploitive).

$$Pr_R(t + 1) = \frac{e^{\theta \cdot Q_R(t+1)}}{e^{\theta \cdot Q_R(t+1)} + e^{\theta \cdot Q_S(t+1)}} \quad (3)$$

Here  $Q_R(t + 1)$  and  $Q_S(t + 1)$  are action values of choosing the risky (R) and safe (S) options on trial  $t + 1$ , respectively.  $Pr_R(t + 1)$  is the probability of choosing the risky option on trial  $t + 1$ . Next, we describe differences between three competing models (1. Fictive Learning Alone (FLA), 2. Original Regret model, 3. Modified Regret model).

**Fictive Learning Alone (FLA)** The FLA model assumes that participants compute action values of each option separately, then select an option that would maximize their expected return. The action value for the safe option is always the same on each game,  $Q_S(t + 1) = M^\alpha$  ( $0 < \alpha < 1.5$ ). In other words, the chosen outcome utility of  $X$  points ( $u_X$ ) was set to  $X^\alpha$ .  $\alpha$  is a parameter that governs the shape of the utility function. As  $\alpha$  goes to zero, the reward sensitivity diminishes. The action value of the risky option is the sum of two possible utilities, weighted by their probabilities. In other words,  $Q_R(t + 1) = u_H \cdot Pr_H(t + 1) + u_L \cdot Pr_L(t + 1)$ .<sup>3</sup> These action values are entered into Equation 3 to compute the probability of choosing each action on the next trial.

**Original Regret Model** In Regret models (both Original and Modified versions), it is assumed that participants choose an option that maximizes their *subjective expected pleasure* or *emotion* (Mellers et al., 1999). Thus, action values are the weighted sum of expected *emotional responses* ( $R$  in Equation 1), rather than expected *utilities*.

Here we used the notation that  $R_{A(B)}(t + 1)$  is the expected emotional response on trial  $(t + 1)$  when chosen and unchosen

payoffs are  $A$  and  $B$ , respectively. We used Equation 1 to calculate  $R_{M(L)}(t + 1)$ ,  $R_{M(H)}(t + 1)$ ,  $R_{L(M)}(t + 1)$ , and  $R_{H(M)}(t + 1)$ .<sup>4</sup> Following Mellers et al. (1999), we set regret/rejoice and disappointment/elation terms to  $sgn(A - B) \cdot |A - B|^\alpha$  when chosen and unchosen payoffs were  $A$  and  $B$ .<sup>5</sup> We assumed that  $\alpha$  is identical for both counterfactual functions and the chosen outcome utility. Importantly, regret/rejoice or disappointment/elation will be weighted by its surprisingness. We used 1 minus its probability as an index of surprisingness (e.g.,  $1 - Pr_H(t + 1)$ ) (Mellers et al., 1999). For example, suppose a participant chooses the safe option (chosen payoff =  $M$  and the foregone payoff =  $H$ ). Then, the expected emotional response can be expressed as  $R_{M(H)}(t + 1)$  from Equation 1, which is equal to  $M^\alpha + (-1) \cdot |M - H|^\alpha \cdot (1 - Pr_H(t + 1))$ .<sup>6</sup> If the participant chooses the risky option and the chosen payoff is  $L$ , the expected emotional response is  $R_{L(M)}(t + 1)$ .  $R_{L(M)}(t + 1)$  is equal to  $L^\alpha + (-1) \cdot |L - M|^\alpha \cdot (1 - Pr_L(t + 1)) + (-1) \cdot |L - H|^\alpha \cdot (1 - Pr_L(t + 1))$ . Note that the disappointment term was included in this case.  $R_{M(L)}(t + 1)$  and  $R_{H(M)}(t + 1)$  can be calculated in the same way and these terms can be used to calculate action values in Equation 4:

$$\begin{aligned} Q_S(t + 1) &= R_{M(H)}(t + 1) \cdot Pr_H(t + 1) + R_{M(L)}(t + 1) \cdot Pr_L(t + 1) \\ Q_R(t + 1) &= R_{H(M)}(t + 1) \cdot Pr_H(t + 1) + R_{L(M)}(t + 1) \cdot Pr_L(t + 1) \end{aligned} \quad (4)$$

The computed action values are entered into the softmax choice rule in Equation 3 to calculate trial-by-trial probability of choosing a risky (or safe) option.

**Modified Regret Model** This model is identical to the Original Regret model except that regret (but not any other counterfactual comparisons) is weighted by Regret's expectedness. We used regret's probability as its expectedness (e.g.,  $Pr_H(t + 1)$ ). Thus, only  $R_{M(H)}(t + 1)$  and  $R_{L(M)}(t + 1)$  are different between two Regret models because participants experience rejoice, but no regret for  $R_{M(L)}(t + 1)$  and  $R_{H(M)}(t + 1)$ .

**Summary of Three Competing Models** In sum, we compared three different models (specifically, value functions). The FLA model assumes that participants evaluate two options separately and choose the option that maximizes their expected return. The two Regret models assume that participants evaluate anticipated emotional responses and maximize their subjective pleasure. The Regret models, however, make different assumptions about the role of surprisingness when processing regretful outcomes. All three models have three free parameters: learning rate ( $\gamma$ ), utility shape ( $\alpha$ ), and choice consistency ( $c$ ). We used hierarchical Bayesian approach to estimate them, which is useful for reliably estimating group and individual parameters (for a review see Lee, 2011).

<sup>2</sup>We tried several other versions of learning rules (e.g., separate learning rates for chosen and unchosen options) and choice rules (e.g., trial-dependent inverse temperature parameter) that are not reported here, but they did not improve model-fits.

<sup>3</sup> $Pr_L(t + 1) = 1 - Pr_H(t + 1)$ ,  $u_H = H^\alpha$ , and  $u_L = L^\alpha$ .

<sup>4</sup>In all settings,  $L < M < H$  (e.g.,  $L=1$ ,  $M=11$ ,  $H=26$ ).

<sup>5</sup> $sgn(x) = 1$  if  $x > 0$ ,  $-1$  if  $x < 0$ .

<sup>6</sup>The disappointment/elation term is present only for risky choices. The disappointment/elation term is missing in  $R_{M(H)}(t + 1)$  because the safe option was chosen, in which there is only one state.

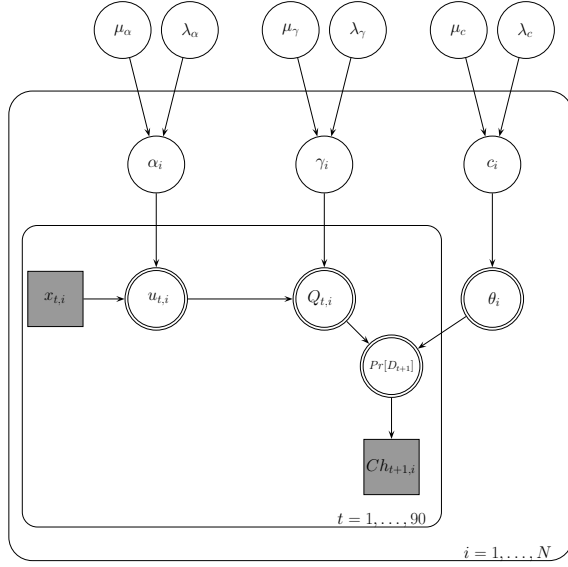


Figure 2: Graphical depiction of the hierarchical Bayesian analysis for three reinforcement learning model.  $R_{A(B)t,i}$  replaces  $u_{t,i}$  for Regret models.

### Graphical Model Implementation - Hierarchical Bayesian Parameter Estimation

Figure 2 shows the graphical representation of all three models. We modeled the variation in  $\gamma_i$ ,  $\alpha_i$ , and  $c_i$  parameters by assuming they have censored Gaussian distributions across participants. (e.g.,  $\gamma_i \sim \text{Normal}(\mu_\gamma, \lambda_\gamma)I(0, 1)$ , where  $\mu_\gamma$  and  $\lambda_\gamma$  are the mean and precision variables of the Gaussian distribution). Mean variables had uniform priors and precision variables had Gamma priors (e.g.,  $\lambda_\gamma \sim \text{Gamma}(.001, .001)$ ). In Figure 2, clear and shaded shapes indicate latent variables and observed variables, respectively. Single and double outlines indicate probabilistic and deterministic functions of input, respectively. Circles and squares indicate continuous and discrete variables, respectively (Lee, 2008). Vectors  $x_{t,i}$  (payoffs) and  $Ch_{t+1,i}$  (choices) were observed and individual ( $\gamma_i$ ,  $\alpha_i$ ,  $c_i$ ) and group parameters ( $\mu_\gamma$ ,  $\mu_\alpha$ ,  $\mu_c$ ,  $\lambda_\gamma$ ,  $\lambda_\alpha$ ,  $\lambda_c$ ) were estimated. We used OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009) to perform Bayesian inference. We used 50,000 posterior samples collected following a total of 30,000 burn-in samples. Multiple chains were used to check convergence and  $\hat{R}$  values indicated that Markov chain Monte Carlo (MCMC) chains converged well with the target posterior distributions. Given that participants' choice behavior varied across games (see Figure 3), we estimated parameters separately for each game (but across all participants within each game). Ideally, model parameters should remain stable across games. Otherwise the model might simply mimic data without providing a coherent theoretical explanation of choice behavior.

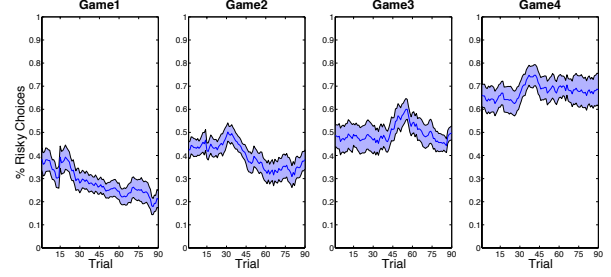


Figure 3: The mean proportions of risky choices over trials on Games 1-4. The blue solid line indicates the group mean on each trial and shaded region indicates  $\pm$ s.e.m. (a moving-average filter was used).

## Results

### Behavioral Results

The proportions of risky choices in each game are plotted in Figure 3. As seen, participants' choice behavior varied across games although the expected values of two options were equated on all games. The mean proportions of risky choices on games 1-4 were .28, .40, .50, and .68 and the differences between games were all significant (games 1 vs 2:  $p < .003$ ; games 2 vs 3:  $p < .004$ ; games 3 vs 4:  $p < .001$ ).

Next, we examined the effect of chosen feedback, foregone feedback, and the magnitude of their difference (Coricelli et al., 2005). For this goal, we performed panel logic regression using the individual random-effects model. The dependent variable was 'switch' (1 if switched from the previous trial, 0 otherwise), and independent variables were the chosen pay-offs (or feedback) ( $fb$ ), the foregone payoffs ( $fgFb$ ), and the magnitude of their difference ( $|fb - fgFb|$ ) on the previous trial (T-1). Table 2 shows that participants were more likely to switch if the chosen feedback was lower ( $p < 3E-16$ ), the foregone feedback was higher ( $p < 2E-13$ ), and the magnitude of the difference was higher ( $p < .011$ ). These results suggest that participants take all three variables into account when making decisions.

To examine the effect of feedback on previous trials, another panel logistic regression analysis was performed, examining how many previous trials ( $fb - fgFb$ ) biased the switch behavior. Figure 4 shows that chosen-foregone pay-offs of up to two previous trials significantly influenced the switch behavior.

Table 2: Regression analysis (panel logit procedure with individual random effect). fb: the chosen payoff (feedback), fgFb: the foregone payoff (feedback).

Variable	Coefficient	Std. Error	$t$	$p$
Constant	.3902	.0186	21.00	<b>&lt;3E-16</b>
fb	-.0064	.0009	-7.44	<b>&lt;2E-13</b>
fgFb	.0027	.0007	3.71	<b>&lt;.001</b>
fb -fgFb	.0025	.0010	2.53	<b>.011</b>

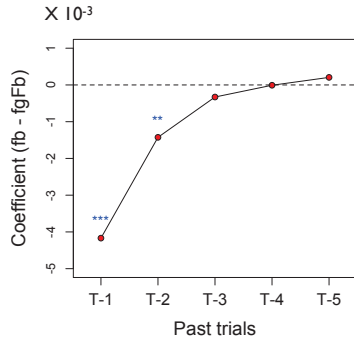


Figure 4: Effects of past outcomes on current choice behavior. fb: the chosen payoff, fgFb: the foregone payoff. \*\*\*  $p < .0001$  \*\*  $p < .001$ .

## Modeling Results

To determine which model best fits our data, we used maximum likelihood estimation (MLE) methods to fit the model to each person and game separately, and then used the Bayesian information criterion (BIC) (Schwartz, 1978) to compare the Bernoulli baseline model, in which the probabilities of two options were equal to the individual's overall proportion of each option (the number of free parameters=1) against three models of interest.<sup>7</sup> The BIC score is a statistic that combines badness of fit with a penalty for the number of parameters. To evaluate the models, we used a BIC change score that measures the improvement of the computational model over the Bernoulli baseline model (BIC change equals the BIC from the baseline model minus the BIC from the cognitive model). Therefore positive BIC changes represent improvement over baseline, and the model with the highest BIC change is considered the best.

Figure 5 shows that the Modified Regret model has the best model fit. When tested across participants, the difference was significant (the Modified vs. Original Regret models:  $p < .005$ , the Modified Regret vs. FLA models:  $p < .05$ ). When the descriptive accuracy was assessed by posterior predictive analysis, the best-fitting model (the Modified Regret model) provided good individual-level model predictions. For example, Figure 6 illustrates a good match between the observed data (Figure 6A) and the model's predictions for a participant's choices (Figure 6B).

Next, we examined whether the parameter values of three models would remain stable across games. Again, ideally model parameters should be similar across different games or tasks. In Figure 7, all parameters of the models were plotted across games 1-4. Clearly, the parameters of the Modified Regret model, which had the best model fit, were the most stable across games. Note that the utility shape ( $\alpha$ ) and consistency ( $c$ ) parameters of FLA and Original Regret models

<sup>7</sup>We are currently working on comparing models by estimating their Bayes factors (Kruschke, 2011)

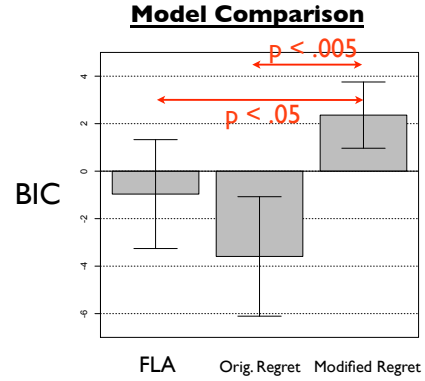


Figure 5: BIC (Bayesian information criterion) scores of three competing models compared to the baseline model. Note that higher BIC indicates a better model fit. Error bars indicate  $\pm$ s.e.m. FLA: Fictive Learning Alone.

varied greatly across games. In sum, the results of both model fit and parameter consistency indicate that the Modified Regret model explains participants' choice behavior best.

## Discussion

The goals of this study were to examine: (1) whether participants make emotion-based choices in experience-based paradigms; (2) whether regret would be weighted by its unexpectedness or expectedness. The modeling results provided strong support for the Modified Regret model: the model had the best model fit and its parameters were the most stable across games, suggesting it might provide a coherent theoretical account for choice behavior across games. The results provide strong support that participants make emotion-based choices and experience greater regret when it was expected rather than when it was unexpected.

We believe this study is the one of the first attempts to incorporate emotion-based decisions into reinforcement learning. Our findings are consistent with previous studies using description-based paradigms that found participants made emotion-based decisions. Our results suggest that reinforcement learning models may need to use value functions that can incorporate emotional components. The results are also consistent with the notion that emotions provide a common currency on how we make decisions under risk or uncertainty (Loewenstein, Weber, Hsee, & Welch, 2001; Weber & Johnson, 2009).

We also believe these results need to be tested in other experience-based paradigms and to determine their generalizability. Some studies found that Bayesian learning models outperformed the delta learning rule (Boorman et al., 2011). Although it is possible that using such a learning model can improve the model fit for all three models, we do not think it will change the main findings of the current study. In sum, we found strong support for the Modified Regret model, which challenges the current theory of regret.



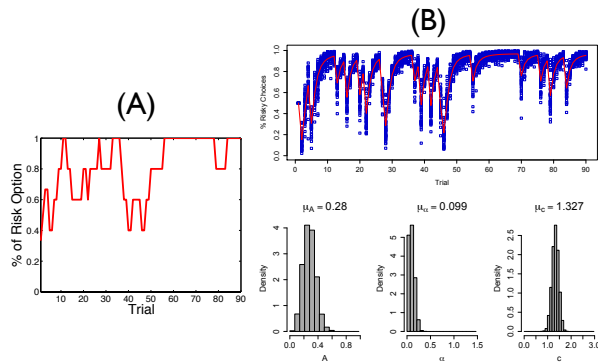


Figure 6: Posterior predictive assessment of the Modified Regret model for one participant. (A) The participant's proportion of risky choices over trials (smoothed with a moving-average filter) (B) posterior predictive distributions for  $Pr_R(t)$ . Small blue squares indicate 50 random samples from the posterior predictive distributions. The red solid line indicates the mean values of the distributions. The participant's model parameter values are in the bottom figure.

## Acknowledgments

This work was supported by the NIMH (R01 MH62150 to BFO), NSF (0817965 to JRB), and Indiana University College of Arts and Sciences Dissertation Fellowship (to WYA).

## References

Boorman, E. D., Behrens, T. E., & Rushworth, M. F. (2011). Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. *PLoS Biology*, 9(6), e1001093.

Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, R. J. (2005). Regret and its avoidance: a neuroimaging study of choice behavior. *Nature Neuroscience*, 8, 1255–1262.

Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2009). Fictive reward signals in the anterior cingulate cortex. *Science*, 324, 948–950.

Hertwig, R., Barren, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15, 534–539.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Academic Press / Elsevier.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin and Review*, 15(1), 1.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical bayesian models. *Journal of Mathematical Psychology*, 55(1), 1–7.

Lejuez, C., Aklin, W., Jones, H., Richards, J., Strong, D., Kahler, C., et al. (2003). The balloon analogue risk task (bart) differentiates smokers and nonsmokers. *Experimental and Clinical Psychopharmacology*, 11(1), 26.

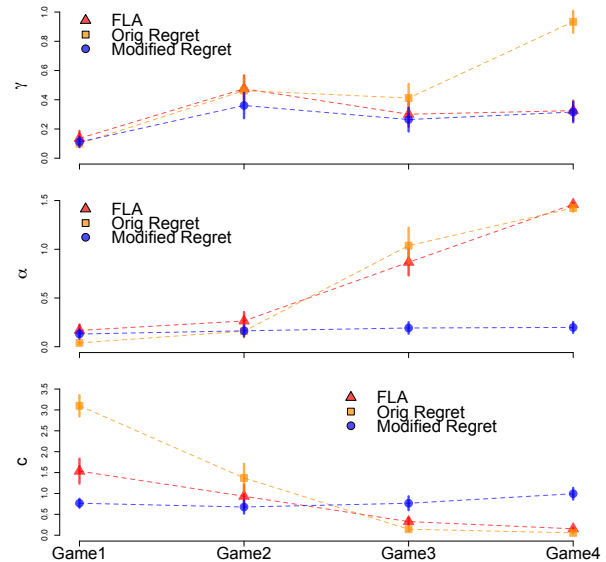


Figure 7: Parameter values of three competing models across games 1-4. Symbols and error bars indicate the means and standard deviations of the posterior distributions, respectively. FLA: the Fictive Learning Alone model.

Loewenstein, G., Weber, E., Hsee, C., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, 127(2), 267.

Lohrenz, T., McCabe, K., Camerer, C. F., & Montague, P. R. (2007). Neural signature of fictive learning signals in a sequential investment task. *Proc. Natl. Acad. Sci. U.S.A.*, 104, 9493–9498.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The bugs project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25), 3049–3067.

Mellers, B., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory. *Psychological Science*, 8(6), 423–429.

Mellers, B., Schwartz, A., & Ritov, I. (1999). Emotion-based choice. *Journal of Experimental Psychology: General*, 128(3), 332.

Rescorla, R. A., & Wagner, A. R. (1972). *A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement* (A. H. Black & W. F. Prokasy, Eds.). Appleton-Century-Crofts.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 5, 461–464.

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, 60, 53–85.

Yechiam, E., & Ert, E. (2007). Evaluating the reliance on past choices in adaptive learning models. *Journal of Mathematical Psychology*, 51, 75–84.

Yechiam, E., & Rakow, T. (2011). The effect of foregone outcomes on choices from experience. *Experimental Psychology*, 1–13.

# Cooperation in Risky Environments: Decisions from Experience in a Stochastic Social Dilemma

**Florian Artinger (artinger@mpib-berlin.mpg.de)**

Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

**Nadine Fleischhut (nadinefl@mpib-berlin.mpg.de)**

Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

**Vittoria Levati (vittoria.levati@univr.it)**

Faculty of Economics, Lungadige Porta Vittoria 41 - 37129 Verona, Italy

**Jeffrey R. Stevens (jeffrey.r.stevens@gmail.com)**

Department of Psychology, 238 Burnett Hall Lincoln, Nebraska 68588-0308, USA

## Abstract

Often in cooperative situations, many aspects of the decision-making environment are uncertain. We investigate how cooperation is shaped by the way information about risk is presented (from description or from experience) and by differences in risky environments. Drawing on research from risky choice, we compare choices in stochastic social dilemmas to those in lotteries with equivalent levels of risk. Cooperation rates in games vary with different levels of risk across decision situations with the same expected outcomes, thereby mimicking behavior in lotteries. Risk presentation, however, only affected choices in lotteries, not in stochastic games. Process data suggests that people respond less to probabilities in the stochastic social dilemmas than in the lotteries. The findings highlight how an uncertain environment shapes cooperation and call for models of the underlying decision processes.

**Keywords:** Decisions from Experience; Cooperation; Risky Choice; Public Good.

## Cooperation in Risky Environments

When people face an opportunity to cooperate, such as when opening a business together or pursuing a joint research project, the outcomes of these enterprises are frequently uncertain. On the one hand, joint enterprises often constitute a social dilemma, where it is in the collective interest of the group to cooperate, yet individually rational to free ride. Despite these incentives, there is overwhelming evidence that many people still engage in cooperation (e.g., Ostrom, 1990). On the other hand, even if people cooperate outcomes often are uncertain due to a risky environment. For instance, even if all business partners cooperate, a new start-up may fail due to external events, such as natural disasters disrupting supplier shipments. Laboratory experiments show that when social dilemmas are embedded in a stochastic environment, cooperation declines sharply (for a review see E. Van Dijk et al., 2004). What has not been addressed is how different levels of environmental risk and the format in which it is presented affect cooperation.

Studies on risky choice find a pronounced difference in behavior depending on how information in lotteries is

presented: whether people sample the distribution of outcomes (*decisions from experience*) or decide based on a summary description of outcomes and probabilities (*decision from description*) (for a review see Rakow & Newell, 2010). In conventional lotteries with described probabilities, people choose as-if they overweight small probabilities as reflected in Prospect Theory (Kahneman & Tversky, 1992). In contrast, people decide as-if they underweight small probabilities if they acquire risk information sequentially by sampling (Hertwig et al., 2004). The difference in choice patterns between decisions from description and experience has been labeled the *description-experience gap* (DE gap).

In lotteries, outcomes depend on environmental risk alone, whereas outcomes in social dilemmas also depend on the choices of other individuals. Stochastic social dilemmas thus combine social uncertainty and environmental risk. Yet our understanding of cooperation in stochastic environments is currently limited to situations in which environmental risk is described by outcomes and probabilities (e.g., Bereby-Meyer & Roth, 2006; Gong et al., 2009; Levati et al., 2009). We argue that real-world risky choices often involve experiencing the outcomes and probabilities of choices rather than receiving their summary statistics. Therefore, examining how risk presentation influences people's decisions is critical to understand how and when people cooperate in risky environments.

There is one important presupposition: risk presentation can influence cooperation only if people are responsive to differences in environmental risk. In lotteries, people's decisions have been found to vary with *different levels of risk*, i.e. for different combinations of outcomes and probabilities while keeping the expected value constant. Analogously, one can describe a stochastic social dilemma by the expected payoffs of cooperation. In a one-shot prisoner's dilemma, people not only cooperate but also respond to different outcomes (Guyer & Rapoport, 1972). Extending this finding to a stochastic setting, the second goal of this study is to establish whether and how different levels of risk affect behavior in one-shot social dilemmas with the same expected payoffs.

Like other types of choices, cooperation is a function of the match between decision processes and the decision-making environment, or what has been labeled ecological rationality. Besides social uncertainty, which has been studied extensively, the levels of environmental risk and uncertainty are critical components of real-world environments that researchers are only recently beginning to appreciate. For instance, cooperation unravels slower in a stochastic social dilemma than in a deterministic one (Bereby-Meyer & Roth, 2006), and groups cooperate more than individuals (Gong et al., 2009). None of the studies, however, addresses how differences in risky environments and the way risk is presented affects cooperation.

## Experiment

The goal of the study is to investigate how risk presentation and different levels of environmental risk affect cooperation in a social dilemma. Even if the outcomes of cooperation also depend on the action of others, the environmental risk affects all who cooperate equally. We thus expect both aspects to influence cooperation in risky environments in the same way as lottery choices with environmental risk alone. To facilitate understanding, we present the detailed hypotheses (see below) after the implementation.

We used a 2 x 2 between-subjects design in which we manipulated risk presentation (description vs. experience) and choice situation (social dilemma vs. lottery). In the *description* condition, subjects received information about how environmental risk influenced outcomes in a *social dilemma* as a probability statement, whereas in the *experience* condition participants sampled to infer the probabilities. To control whether the values and probabilities chosen to implement environmental risk replicated the DE gap, two further groups made decisions in *lotteries*, again either from description or experience. The environmental risk was identical between lotteries and games. To investigate how different levels of risk affect behavior in one-shot social dilemmas, we varied probabilities and outcomes within-subjects while keeping the expected outcomes constant.

## Methods

### Environmental Risk in Social Dilemmas and Lotteries

For the *social dilemma* conditions, we used a stochastic 2-person public goods game (PG) with binary choices. For each choice, participants receive an endowment  $e$  (10€) which they could contribute to a joint project with a randomly matched partner or keep for themselves. Contributions were multiplied by a value ( $msr$ ) and shared equally between both pair members. Denoting  $i$ 's contribution by  $c_i$ , where  $c_i \in \{0, e\}$  and  $i = 1, 2$ ,  $i$ 's payoff is given by

$$\pi_i = e - c_i + \frac{msr}{2}(c_1 + c_2). \quad (1)$$

We impose  $msr \in \{1, 2\}$ . An  $msr > 1$  made it socially optimal to contribute, whereas an  $msr < 2$  rendered free-

riding the dominant strategy for a selfish person, thus creating a social dilemma.

We manipulated environmental risk by assigning the  $msr$  to one of two possible values, representing either a good or a bad event, with a certain probability. In case the bad event occurred (with probability  $p$ ), contributions were multiplied by an  $msr < 1$ , decreasing the value of the public good. When the good event occurred, contributions were multiplied by an  $msr > 1$ , increasing the value of the contributions. The environmental risk only affected what was invested. Cooperation thus represents the risky and non-cooperation the sure option. We chose the two potential  $msr$ -values and corresponding probabilities such that the expected  $msr$ ,  $E[msr]$ , across good and bad event always yielded a social dilemma with  $1 < E[msr] < 2$ .

Table 1 illustrates the eight decision situations employed. Situations 1 to 4 contained *rare* ( $p < 0.25$ ) bad events, analogous to the DE gap studies with lotteries (e.g., Hertwig et al., 2004). Situations 5 and 6 contained more *common* ( $p > 0.25$ ) bad events to test whether the DE gap extends beyond rare events as found by Ludvig and Spetch (2011). We use two different expected  $msr$ , 1.2 and 1.4, to check the robustness of the results. Situations 1 – 6 were designed to extend the findings from the DE gap studies in risky choice to social dilemmas. At the same time, keeping the expected  $msr$  constant across different combinations of probabilities and potential returns allows us to test whether different levels of environmental risk affect choices in the PG in the same way in which they affect choices in lotteries.

Decision situation 7 and 8 explored boundary conditions of a social dilemma and provided a further control of participants' understanding of the incentives. In situation 7, the  $E[msr]$  equaled 1.1, which made it less attractive to cooperate compared to situations 1 – 6. In contrast to the other situations, here the rare event was the good state of the world. Different from situations 1 to 7, the expected  $msr$  of 2.1 in situation 8 did not generate a social dilemma and made it individually and socially optimal to cooperate.

In most studies on the DE gap, the risky option has an expected value that is only marginally higher than the sure option. To avoid floor effects in the social dilemma, we used relatively large expected  $msr$ . This should provide strong incentives to cooperate in the PG but results in a larger difference between the expected  $msr$ -value of the sure option and risky option. To control whether the parameters we chose for implementing environmental risk replicated the DE gap in more standard settings, we ran the same choices as lotteries with identical environmental risks. In the lottery conditions, participants also received an endowment  $e$  and had to decide whether to invest into a risky option. The risky option in each lottery used the same two possible  $msr$  with the same probabilities as the corresponding PG. Yet, while the payoffs in the games also depended on the action of another person, the payoffs in the lotteries only depended on the realized state of the world. The lotteries strip the strategic component away but retain the stochastic component that stems from the environment. We



randomized the order of decision situations in games as well as lotteries, and participants received no feedback about the realized *msr* (or decision of the other group member) after each decision.

### Decisions from Description vs. Decision from Experience

In the *description* conditions, participants received information about environmental risk as a summary statement about probabilities and associated *mrs*-values before they made their decision. In the *experience* conditions, participants sampled the distribution of *mrs*-values by drawing 25 cards from a deck. We used a matched-sampling design based on Ungemach et al. (2010), where people were forced to view a representative sample of the underlying distributions of outcomes. Each card contained a number corresponding to one of the two possible *msr*. For example, in situation 1 the deck had 2 cards with the *msr* 0 and 23 cards with the *msr* 1.30. The sequence of cards was randomized for each participant, yet the two possible *msr* and their frequencies matched exactly the objective probabilities given in the *description* condition. Thus, sampling error could not cause any differences observed between the two conditions.

In the *experience* conditions, we additionally collected time stamps that allowed us to evaluate how long participants viewed a certain card and whether this influenced their decision. To check the accuracy of risk estimates, we also asked participants after the last round how often they saw the two sampled *msr*-values. In the *description* conditions, participants translated the probability statement of the last round into a frequency statement to control whether participants accurately understood the risk.

**Further Tasks** In the *social dilemma* conditions, participants also faced two deterministic PGs with an *msr* of 1.2 and 1.4 (randomized order) after the stochastic situations. This allowed us to investigate how cooperation varies if the stochastic component is removed, since the deterministic games matched the expected *msr* of the stochastic PGs in situations 1, 2, and 5 ( $E[mrs] = 1.2$ ) as well as 3, 4, and 6 ( $E[mrs] = 1.4$ ).

At the end of the experimental session, participants indicated in a questionnaire which of six reasons best explains their decision to invest/not invest into the stochastic PGs: the probability of the *mrs* were (not) sufficiently high, the values of the *mrs* were (not) sufficiently high, conditional cooperation, social uncertainty, greed/opportunism, moral values, or none of these. A section on demographics concluded the experiment.

**Participants and Procedure** We randomly assigned 128 students in Jena, Germany, to one of four sessions. In the social dilemma conditions, participants had to pass control questions to ensure that they understood the impact of

environmental risk and of the other person's choice on their payoffs. All tasks were completed anonymously employing a perfect stranger design. At the end, one decision situation was randomly chosen to determine the payoff. Participants earned on average 14€.

### Hypotheses

**Risk sensitivity in social dilemmas and lotteries** Do different levels of environmental risk affect stochastic PGs in a similar way as they affect lotteries? To test this presupposition, we focus on decisions from description and employ the predictions of Prospect Theory (Kahneman & Tversky, 1992). Using a separate value and weighting function, Prospect Theory transforms the expected outcomes of a lottery into Prospect Theory Values (PTVs), analogous to expected values. When comparing the PTV of a lottery's risky option with a sure option (always 1 in our case), the conventional prediction is that the risky (sure) option is picked if the PTV is larger (smaller). Investment rates into the PG are expected to be lower than in lotteries due to a second source of uncertainty that stems from the other person. Thus, the PTVs based on environmental risk alone are unlikely to be useful. However, the PTVs also produce a ranking of the 8 decision situations in terms of proportion of risky choices. Such a ranking can be applied to both lotteries and stochastic PGs in the description condition. Table 1 lists the PTVs for the eight decision situations of this experiment based on the parameters used by Tversky and Kahneman (1992). From the PTVs, two predictions follow for PGs and lotteries with the same expected *msr*:

(1a) Situations 1 and 3 (bad event occurs with 8%) will lead to a higher number of risky choices than situations 2 and 4 (where the bad event occurs with 20%).

(1b) Situation 5 (6), where the bad event is more common, will lead to more risky choices than situations 1 and 2 (3 and 4).

**Decisions from Description and from Experience** Using lotteries, studies found that experienced small probabilities appear to be underweighted in choices compared to described ones (Hertwig et al., 2004). Extending this choice pattern to social dilemmas leads to the following hypothesis for stochastic PGs and lotteries:

(2) The risky option will be chosen more frequently in the experience condition than in the description condition if the bad event is less likely (situations 1 – 6 and 8), whereas this pattern should reverse for situation 7, in which the good event is less likely.

## Results

### Risk Sensitivity in Social Dilemmas and Lotteries

We would not expect risk presentation to matter unless people are sensitive to different levels of risk in games as they are in lotteries. For the results of hypothesis 1a and 1b,

Table 1: Percentage of subjects investing in PGs / lotteries and differences between description and experience conditions

Decision Situations				Stochastic PG			Lotteries		
#	Risky Option	$E[msr]$	PTV	Desc	Exp	Difference between description and experience conditions	Desc	Exp	Difference between description and experience conditions
<b>One rare event</b>									
1	1.30, 0.92 0, 0.08	1.2	0.93	47	44	-3 ( $\chi^2(1) = 0.06$ , $p = 0.80$ )	78	81	+3 ( $\chi^2(1) = 0.10$ , $p = 0.76$ )
2	1.45, 0.8 0, 0.2	1.2	0.84	28	28	0 ( $\chi^2(1) = 0.00$ , $p = 1.00$ )	44	69	+25 ( $\chi^2(1) = 4.06$ , $p = 0.04$ )
3	1.55, 0.92 0, 0.08	1.4	1.09	66	56	-9 ( $\chi^2(1) = 0.59$ , $p = 0.44$ )	81	88	+6 ( $\chi^2(1) = 0.47$ , $p = 0.49$ )
4	1.80, 0.8 0, 0.2	1.4	1.02	38	38	0 ( $\chi^2(1) = 0.00$ , $p = 1.00$ )	63	78	+16 ( $\chi^2(1) = 1.87$ , $p = 0.17$ )
<i>Mean 1 – 4</i>				45	41	-3 ( $\chi^2(1) = 0.26$ , $p = 0.61$ )	66	79	+13 ( $\chi^2(1) = 5.03$ , $p = 0.03$ )
<b>Two common events</b>									
5	1.80, 0.64 0.20, 0.36	1.2	0.96	25	28	3 ( $\chi^2(1) = 0.08$ , $p = 0.77$ )	34	44	+9 ( $\chi^2(1) = 0.59$ , $p = 0.44$ )
6	1.95, 0.56 0.70, 0.44	1.4	1.21	41	28	-13 ( $\chi^2(1) = 1.11$ , $p = 0.29$ )	44	59	+16 ( $\chi^2(1) = 1.56$ , $p = 0.21$ )
<i>Mean 5 &amp; 6</i>				33	28	-5 ( $\chi^2(1) = 0.33$ , $p = 0.57$ )	39	52	+13 ( $\chi^2(1) = 2.02$ , $p = 0.16$ )
<b>Extreme <math>msr</math></b>									
7	0.75, 0.88 3.50, 0.12	1.1	1.23	19	16	-3 ( $\chi^2(1) = 0.11$ , $p = 0.74$ )	38	16	-22 ( $\chi^2(1) = 3.92$ , $p = 0.05$ )
8	2.20, 0.96 0.30, 0.04	2.1	1.70	91	88	-3 ( $\chi^2(1) = 0.16$ , $p = 0.69$ )	100	97	-3 ( $p = 0.50$ , Fisher's exact test)

we focus on data from the *description* conditions for decision situations 1 to 6.

When comparing decision situations with an  $E[msr] = 1.2$  and  $E[msr] = 1.4$ , cooperation increases with the expected  $msr$ . The deterministic PGs yield a similar pattern: the rate of cooperation is 53% when  $msr = 1.2$  and, 81% when  $msr = 1.4$  ( $\chi^2(1) = 5.74$ ,  $p = 0.02$ ). In the stochastic PGs, the average rate of cooperation is 33% when  $E[msr] = 1.2$  and 48% when  $E[msr] = 1.4$  ( $\chi^2(1) = 4.23$ ,  $p = 0.04$ ). Thus, differences in expected  $msr$  affect behavior even though the social dilemma is maintained and the dominant strategy for a person is not to cooperate. This replicates Guyer & Rapoport (1972) findings and extends it to a stochastic setting. But, besides being sensitive to different expected outcomes, do people react to different levels of risk for constant expected outcomes?

To address this question, we pool our data across situations with expected  $msr$ -values of 1.2 and 1.4 to obtain more reliable results. The mean cooperation rate is 1.7 times higher in situations where the bad event occurs with 8% than in situations where the bad event is common ( $\chi^2(1) = 7.12$ ,  $p = 0.01$ ). Thus, changes in the stochastic environment have a large impact on cooperation. The difference in cooperation between deterministic and stochastic PG with

an 8% chance of a bad event is only 10.5% and not significant ( $\chi^2(1) = 1.62$ ,  $p = 0.20$ ).

To investigate hypotheses 1a and 1b – that situations with 8% receive more investment than situations with 20% –, one can also rely on the pooled data across the  $E[msr]$  of 1.2 and 1.4 because the rankings of PTVs are identical for both. The rate of investment in situations with a probability of 8% compared to 20% sharply drops both for stochastic PGs (from 56% to 33%,  $\chi^2(1) = 7.17$ ,  $p = 0.01$ ) and lotteries (from 80% to 53%,  $\chi^2(1) = 10.12$ ,  $p < 0.001$ ). Paralleling each other, stochastic PGs and lotteries thus are in line with prediction 1a based on Prospect Theory.

For prediction 1b, the data also suggests a decline in cooperation between situations with a probability of 20% and those with two common events. Statistically, however, there is no difference between these two situations, neither for the stochastic PGs (the investment rate is constant at 33%,  $\chi^2(1) = 0.00$ ,  $p = 1.00$ ), nor for lotteries (the investment rate declines from 53% to 39%,  $\chi^2(1) = 2.55$ ,  $p = 0.11$ ). Hypothesis 1b based on Prospect Theory – that the rate of investment is highest with a common event – is neither met in stochastic PGs nor in lotteries.

In summary, we find that different levels of environmental risk both influence choice in the PGs for

decisions from description and result in similar behavior in stochastic PGs and lotteries. Though the data confirm the predictions of Prospect Theory for hypothesis 1a, we did not obtain support for hypothesis 1b for either PGs or lotteries.

### Decisions from Description and from Experience

**Is there a DE gap in lotteries and games?** We initially focus on pooled data from the eight decision situations to start with more reliable results. Hypothesis 2 is directional and states that, except for situation 7, participants should choose the risky option more often in the experience condition. To test this hypothesis, we subtracted the percentage of people contributing in the experience condition from those in the description condition, except for situation 7 where we do the opposite. The results show a positive gap for lotteries ( $\chi^2(1) = 8.24, p = 0.003$ ), with a mean difference between experience and description of 12% (SD = 10%).

Table 1 lists percentage of people investing in experience and description separately for all eight decisions situations in lotteries and stochastic PGs. For lotteries, the predicted difference between the experience and description condition is observed in all situations (including the reversal for situation 7) – except for lottery 8. This lottery shows a ceiling effect because the expected outcome is twice as high as the sure option, so that in both conditions all participants but one invested.

Averaging across lotteries 1-4, which contain a rare event, shows a DE gap of 13% (Table 1). The same DE gap (13%) occurs with lotteries containing a more common bad event (5 and 6, Table 1). The results replicate Ludvig and Spetch (2011), who find the DE gap also for situations with common events. Overall, responses to decisions from description and experience differed in lotteries as predicted based on previous findings. Thus, the parameters we chose for environmental risk replicate the DE gap found in the risky choice literature.

Given that the parameters replicate the DE gap in lotteries and the previous result that people's decisions in games were similarly sensitive to differences in risk as in lotteries, we expected the risk presentation format to influence cooperation as well. The behavior in the stochastic PGs, however, does in this respect not match the behavior in lotteries: the DE gap completely disappears in games ( $\chi^2(1) = 0.38, p = 0.30$ ). The mean difference between experience and description in the stochastic PG is -3% (SD = 6%).

The stochastic PGs stand in stark contrast to the results in the lotteries. In games, 6 out of 8 decision situations show no or only minimal gaps. Experience and description conditions do not differ for any of the decision situations. In fact, situation 7, which is closest in spirit to the situations used in by Hertwig et al., (2004) and Ungemach et al., (2009), shows a strong DE gap in lotteries, but the gap disappears completely in the games.

**Why is there a DE gap in lotteries but not in games?** In the following, we explore reasoning processes in PGs and

lotteries that provide hints to why risk presentation affects lotteries but not stochastic PGs.

One possible explanation underlying this pattern is that participants spend different amounts of time sampling in lotteries and games, which may indicate different search processes. In lotteries, participants spent more time viewing the rare event ( $M = 0.91$  seconds,  $SD = 0.99$ ) compared to the frequent event ( $M = 0.67$  seconds,  $SD = 0.65, t(6400) = 10.01, p < 0.001$ ). Similarly, for the games, participants viewed the rare event ( $M = 0.51$  seconds,  $SD = 0.51$ ) longer than the frequent event ( $M = 0.43$  seconds,  $SD = 0.33, t(6400) = 6.38, p < 0.001$ ). In lotteries, however, participants spent more time sampling than in games for both rare events ( $t(2432) = 12.45, p < 0.001$ ) and frequent events ( $t(10368) = 24.02, p < 0.001$ ). These differences in sampling times thus provide evidence for potentially different search processes in games which appear to pay less attention to the actually observed probabilities compared to lotteries.

To control for the accuracy of risk perception, participants in the experience conditions stated the frequency of the two outcomes in the last situation after they had decided. The actual distribution of outcomes participants saw correlates with the stated frequencies for lotteries ( $r_s = 0.72, p < 0.001$ ) yet to a lesser extent for stochastic PGs ( $r_s = 0.43, p < 0.01$ ). In both conditions participants were calibrated to the actual probabilities and did not underestimate but rather, if anything, overestimated the probability of rare events.

Some researchers suggest that the larger influence of recent events in decisions from experience may drive the DE gap. Hertwig et al. (2004) and Rakow, Demes, & Newell (2008) found a recency effect in decisions from experience but Ungemach et al., (2010) and Hau, Pleskac, Kiefer, & Hertwig (2008) did not. To test for a recency effect, we divided the 25 samples participants draw before each decision into two sets: from 1 to 12 (initial) and from 13 to 25 (latter). Then we computed the expected  $msr$  from the initial samples,  $E[msr]_{1-12}$ , and from the latter samples,  $E[msr]_{13-25}$ . Finally, we compare the number of risky choices made when  $E[msr]_{13-25} > E[msr]_{1-12}$  to the number of risky choices made when  $E[msr]_{13-25} < E[msr]_{1-12}$ . When the  $E[msr]$  of the latter, more recent sample was larger, we find a higher number of risky choices in lotteries ( $\chi^2(1) = 3.77, p = 0.04$ ) but not in games ( $\chi^2(1) = 0.30, p = 0.34$ ). This also suggests that the actual observed probabilities may play a less important role in games than in lotteries.

Finally, for the stochastic PG in description and experience, participants indicated their most important reasons for cooperating as well as not cooperating. This resulted in two statements per participants. Aggregating across both statements, probabilities influenced cooperation decisions in the description condition for 59% of the participants, compared to 39% in the experience condition. In this condition, participants rather emphasized both the value of the  $msr$  they could obtain (20% in experience, and 3% in description) and their expectation whether the other will (not) cooperate, i.e. conditional cooperation (20% in

experience and 11% in description). This indicates that the importance of the probabilities for decisions is further reduced in the stochastic PG in experience.

In summary, participants sampled more quickly in the stochastic PG in the experience condition than in lotteries, as if they were paying less attention to the observed probabilities. In line with this, subjects' risk perception was less accurate in games than in lotteries, and recency – a potential cause of the DE gap – did not play a role in games, whereas we did find a recency effect in lotteries. The questionnaire also highlighted that probabilities were less important in the PG in experience than the size of the values and beliefs about others' behavior. This provides converging evidence that as the probabilities of the risky option lose importance in the games, the DE gap washes out.

### General Discussion

People often cooperate in social dilemmas. We examined how critical aspects of the stochastic environment shape cooperation. First, different levels of environmental risk influence cooperation. Investments in the stochastic PGs match those observed in lotteries, with people preferring an 8% chance of a bad event to a 20% chance for constant expected payoffs. Second, the *msr*-values and probabilities chosen to implement environmental risk replicate the DE gap within individual risky choices in lotteries. That is, people choose the risky option more often when experiencing the risky outcomes compared to when receiving summary descriptions. Our key finding is that, nevertheless, risk presentation matters in lotteries but not in games: no DE gap existed for the social dilemmas. Process data and subjects self-reported reasons for cooperation suggest that the disappearance of the DE gap in games may result from a decision process that emphasizes the size of the outcomes and expectations about others' behavior over outcome probabilities.

In our view, to include environmental risk and decisions from experience into the study of cooperation invites more realism into the laboratory. This study is only a small step to build on insights from research on risky choice for decision situations which combine environmental risk and social uncertainty. In particular, models that focus more on actual decision processes instead of choices alone may provide promising alternative starting points to Prospect Theory, which in our study could not account for the data in the description condition for either lotteries or games. In complex interactive environments, it seems rather likely that non-compensatory decision making emerges. For instance, a lexicographic strategy like the Priority Heuristic (Brandstatter et al. 2006), outlines a sequential decision process which considers outcomes in the first and probabilities only as a second step if no decisions has been made. In a similar fashion, other strategies that do not trade-off reasons may be valuable to model search and decisions processes in situations that combine environmental risk and social uncertainty – and thus also include expectations about

others and further social reasons besides mere outcomes and probabilities.

### Acknowledgments

This research was funded by a research grant of the Arete Initiative at the University of Chicago, with additional support from the Max Planck Society.

### References

- Bereby-Meyer, Y., & Roth, A. E. (2006). The Speed of Learning in Noisy Games: Partial Reinforcement and the Sustainability of Cooperation. *American Economic Review*, 96, 1029-1042.
- Brandstatter, E., Gigerenzer, G., & Hertwig, R. (2006). The Priority Heuristic: Making Choices without Trade-Offs. *Psychological Review*, 113, 409-432.
- Gong, M., Baron, J., & Kunreuther, H. (2009). Group cooperation under uncertainty. *Journal of Risk and Uncertainty*, 39, 251-270.
- Guyer, M. J., & Rapoport, A. (1972). 2× 2 Games Played Once. *The Journal of Conflict Resolution*, 16, 409-431.
- Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, 15, 534-539.
- Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description-experience gap in risky choice: the role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, 21, 493-518.
- Levati, V. M., Morone, A., & Fiore, A. (2009). Voluntary contributions with imperfect information: An experimental study. *Public Choice*, 138, 199-216.
- Ludvig, E. A., & Spetch, M. L. (2011). Of Black Swans and Tossed Coins: Is the Description-Experience Gap in Risky Choice Limited to Rare Events? *PLoS ONE*, 6, e20262.
- Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.
- Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making*, 23, 1-14.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5, 297-323.
- Ungemach, C., Chater, N., & Stewart, N. (2010). Are Probabilities Overweighted or Underweighted When Rare Outcomes Are Experienced (Rarely)? *Psychological Science*, 20, 473-479.
- van Dijk, E., Wit, A., Wilke, H. A. M., & Budescu, D. V. (2004). What we know (and do not know) about the effects of uncertainty on behavior in social dilemmas. In R. Suleiman, D. V. Budescu, I. Fischer, & D. M. Messick (Eds.), *Contemporary psychological research on social dilemmas*. Cambridge University Press.

# Modeling individual differences in socioeconomic game playing

Derrik E. Asher<sup>1</sup>, Shunan Zhang<sup>1</sup>, Andrew Zaldivar<sup>1</sup>, Michael D. Lee<sup>1</sup> and Jeffrey L. Krichmar<sup>1,2</sup>

<sup>1</sup> Department of Cognitive Sciences, University of California, Irvine

<sup>2</sup> Department of Computer Science, University of California, Irvine

## Abstract

Game theory has been useful for understanding risk-taking and cooperative behavior. In the present study, subjects played the Hawk-Dove game with simulated and embodied (robotic) neural agents which used a neurobiologically plausible model of action selection and adaptive behaviors. Subjects had their serotonin levels temporarily altered through acute tryptophan depletion (ATD). The traditional assumption for subject data from Game-theory-ATD or human robot interaction (HRI) studies is that all participants come from the same underlying distribution or same group. We used probabilistic graphical models in order to determine potential sub-group affiliations based on the subjects' responses while playing the Hawk-Dove game. The results from the models indicate sub-groups within a subject population exist. We find that two-group, one that tends toward cooperation and the other that tends toward aggression, best describes the effect of subject behavior in response to ATD and embodiment.

**Keywords:** Adaptive systems; Human robot interaction; Neurotransmitters; Cognitive Robotics; Bayesian inference; Graphical models; Individual Differences.

## Introduction

Economic game theory has had a long, productive history of predicting and describing human behavior in cooperative and competitive situations (Maynard Smith, 1982 ; Nowak, Page, & Sigmund, 2000 ; Skyrms, 2001). The theory of games has also been used to illuminate the neural basis of economic and social decision-making (Lee, 2008 ; Rilling & Sanfey, 2011). However, these studies typically have people play against opponents with set strategies and predictable behavior. Moreover, in most of these studies, subjects are making decisions while sitting in front of an antiseptic computer screen. The present study addresses these issues by having subjects play a socioeconomic game, known as Hawk-Dove, against an autonomous robot with the ability to adapt its behavior to the game situation.

Neuromodulatory systems, such as dopamine and serotonin, appear to be applicable to decision-making in social situations. The serotonergic (5-HT) and dopaminergic (DA) systems oppose each other with respect to predicting punishment (5-HT) versus predicting reward (DA) (Boureau & Dayan, 2011).

We developed a computational model of neuromodulation and action selection based on the assumptions, that dopamine levels are related to the expected reward of an action, and serotonin levels are related to the expected cost or risk of an action (Asher, Zaldivar, & Krichmar, 2010 ; Zaldivar, Asher, & Krichmar, 2010). The model of neuromodulation and action selection demonstrated the ability to adapt to the game situation and its opponent's strategy. The model was embedded in both simulated and embodied neural agents to investigate reciprocal social interactions in games of cooperation

and conflict with people (Asher, Zaldivar, Barton, Brewer, & Krichmar, submitted).

Subjects played a series of Hawk-Dove games against robotic and simulated agents. The effects of serotonergic levels on adaptive behavior in these games were tested by simulating serotonergic lesions in the neural agent, which results in a more aggressive agent, or lowering the CNS serotonin levels of people through a dietary manipulation called acute tryptophan depletion (ATD), which has been shown to decrease cooperation and lower harm-aversion (Crockett, Clark, Tabibnia, Lieberman, & Robbins, 2008 ; Wood, Rilling, Sanfey, Bhagwagar, & Rogers, 2006).

A major finding of the study was that people changed their overall strategies in response to changes in the neural agents state. Subjects tended to deploy either Tit-For-Tat (T4T) or Win-Stay, Lose-Shift (WSLS) strategies during game play. In a T4T strategy, a subject copies the most recent move of the opposing player. In a WSLS strategy, a subject selects the same action that led to a positive payoff in the previous game (Win-Stay), or a different action from the previous game if that action led to zero or negative payoff (Lose-Shift). When playing against a more aggressive neural agent, which had a lesion to its serotonergic system, subjects switched from a Win-Stay, Lose-Shift (WSLS) strategy to a Tit-For-Tat (T4T) strategy. This change in strategy was independent of whether the neural agent was a robot or a computer simulation, and independent of subject tryptophan levels.

In the present study, we test whether embodiment and lowering serotonin has an effect on individual subject behavior during Hawk-Dove game playing by using hierarchical latent mixture models with Bayesian inference. This framework for developing and evaluating structured cognition offers a principled and comprehensive approach for modeling individual differences and their use of cognitive strategies (Lee, 2008 ; Lee, Zhang, Munro, & Steyvers, 2011). The hierarchical nature of the models allows variation in the parameters controlling cognitive processes across individuals to be accommodated. We find that two categories of subjects, one that tends to be more aggressive and one that tends to be more cooperative, best describes subject behavior in response to ATD and embodiment.

## Experiment

### Subjects

Eight subjects (three female; mean age: 26.6 years; standard deviation of age: 3.8 years) participated in this study.

## Hawk-Dove Game

The Hawk-Dove game consisted of a human and a neural agent choosing a single action in response to a territory of interest (TOI). The Hawk-Dove game, which is similar to Prisoner's Dilemma, was chosen because it is amenable to a physical instantiation with a robot. Moreover, it has an additional strategic element since choices are different depending on who arrives at the TOI first. At the start of the game, the TOI and the human subjects' location were randomly placed on a playing grid. The current location of the robot was used as a starting position. The player who arrived at the neutral TOI first had the opportunity to take one of two possible actions: Escalate (i.e. an aggressive, confrontational tactic) or Display (i.e. a nonviolent, cooperative tactic). The player who arrived second responded with one of the two aforementioned actions. After each game, a payoff was calculated. If both players chose Escalate, they received a penalty that is set before the game. If both players chose Display, they split the value of the TOI resource. If one player chose Escalate and the other chose Display, the player who chose Escalate received the entire value of the resource.

We also developed a simulated variant of the Hawk-Dove game, where subjects played against a robot icon on an interactive screen. The same neural model used for the real robot dictated the control of the robots icon and its decision-making. This simulated setup allowed us to judge whether physical embodiment had an effect on human behavior.

## Acute Tryptophan Depletion (ATD)

ATD was used to temporarily alter the levels of serotonin in the brain via a decrease in blood plasma tryptophan, the amino acid precursor to serotonin. Because free blood plasma tryptophan levels, and the corresponding serotonin levels in the brain, vary with the amount of dietary tryptophan and the rate of protein synthesis, these levels can be altered by a low protein diet in combination with a specially prepared 'protein shake'. This protein shake contains an amino acid load (lacking tryptophan), which has two effects. First, it stimulates protein synthesis in the liver, which uses up blood plasma tryptophan. Second, the amino acids that are given in the protein shake compete with tryptophan for transport across the blood-brain barrier, which restricts entry of tryptophan into the brain and leads to lower levels of serotonin in the brain (e.g. Bell, Hood, & Nutt, 2005 ; Hood, Bell, & Nutt, 2005).

## General Procedures

In a double-blind study, human subjects were randomly assigned on the first experimental day to receive either the control mixture (Tryp+) with tryptophan or the mixture without tryptophan (Tryp-). The mixtures were administered as a specially prepared protein shake. The Tryp+ and Tryp- shakes contained 16 and 15 amino acids respectively.

Each subject then returned to participate in the other condition at least seven days later to ensure the return to baseline blood plasma tryptophan levels between experimental days. On the morning of each experimental day, a blood sample

was drawn to determine baseline blood plasma tryptophan levels. Following the blood draw, subjects ingested one of the mixture drinks (either Tryp+ or Tryp-). A second blood sample was drawn approximately 5 hours after ingestion of the mixture to confirm reduction (Tryp- condition) or maintenance (Tryp+ condition) of blood plasma tryptophan levels. Roughly 5.5 hours after consumption of the mixture, human subjects then participated in a series of Hawk-Dove games against a neural agent.

## Experimental Conditions and Data

We were interested in two experimental conditions per subject. In the Simulation vs. Robot condition, subjects would play games against a computer agent or against the robot; In the Tryp+ vs. Tryp- condition, subjects would play games against a neural agent with an intact simulated neuromodulatory system or a simulated lesion of its serotonergic system.

The Escalate or Display decisions for each game were collected for both the subject and neural agent for all games in each condition. The dependent variables of interest are the percentage of games per condition a subject chose the Escalate tactic and the use of either the Tit-For-Tat (T4T) or Win-Stay, Lose-Shift (WSLS) strategies. Human subjects played 20 games of Hawk-Dove per condition. For detailed experimental conditions, see Asher et al. (submitted).

## Data Analysis

### Bayesian Hierarchical Model Approach

To investigate the influences of lowering serotonin levels and of agent embodiment on individual decision-making, we used hierarchical latent mixture models with Bayesian inference. Hierarchical Bayesian inference has been demonstrated as a flexible and interpretable way of extending simple models of cognitive processes (e.g. Lee, 2008 ; Rouder, Lu, Speckman, Sun, & Jiang, 2005 ; Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2010). The hierarchical nature of the models allows variation in the parameters controlling cognitive processes across individuals to be accommodated. The latent mixture nature of the models allows the use of entirely different cognitive strategies across individuals to be modeled.

Formally, we recast the cognitive models as probabilistic graphical models and used Markov Chain Monte Carlo (MCMC) methods for computational Bayesian inference. This approach to Bayesian inference over richly-structured cognitive models has been applied to data covering a diverse set of cognitive skills. For example, Bayesian graphical models have been used to make inferences about the use of strategies, such as WSLS or T4T, from sequences of choice data in bandit problems and other sequential decision-making tasks (e.g. Lee et al., 2011 ; Newell & Lee, 2011).

Using hierarchical latent mixture models, we addressed the question of how ATD and embodiment can affect subjects' decisions to compete (i.e., escalate) or cooperate (i.e., display). We modeled the probability of escalating through a logistic model. Specifically, the logit of the probability of es-

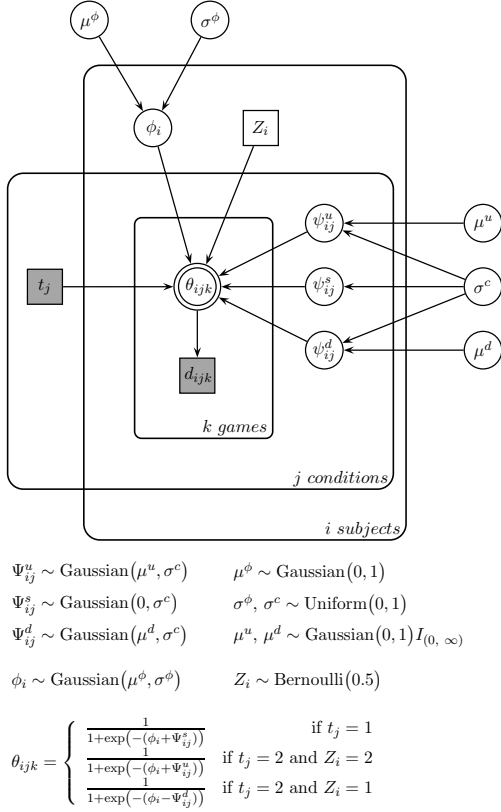


Figure 1: Generic process of escalation or strategy usage under the effect of ATD or embodiment.  $\phi$ : baseline tendency of escalation or strategy usage;  $\Psi^u$ ,  $\Psi^d$  and  $\Psi^s$ : additive effects of escalation or strategy usage associated with the experimental conditions;  $t$ : indicator of experimental condition;  $\theta$ : probability of escalation or strategy usage;  $d$ : observed escalation decision or strategy usage;  $Z$ : latent group indicator;  $\mu^\phi, \sigma^\phi, \mu^u, \mu^d, \sigma^c$ : hyper-parameters of prior distributions. Experimental conditions are either Tryp- vs. Tryp+, or Robot vs. Sim.

calating for each subject in each condition is assumed to follow a Gaussian distribution defined by its mean and variance (hyperparameters in the hierarchical model), with the mean modeled as the sum of the baseline level of escalating for the subject, and an additive effect associated with ATD. To give a full account of the data, the hierarchical model is designed to address individual differences at two levels, the baseline level, which depends on the subjects, and the additive level, which depends on the interaction between subjects and experimental conditions. Our justification is that it is possible that the effect of ATD or embodiment can vary across different individuals, resulting in either an increase or decrease in the likelihood of escalating a fight.

### The Graphical Model

We use graphical models to describe the relationship between subject decision-making (escalation, T4T and WSLs) and

predictors of interest (ATD and embodiment). In total we built six graphical models for every decision-making and predictor pair. In this section, we show a concrete example of how we model the relationship between escalation and ATD; all other models were built along a similar line.

As shown in Figure 1, nodes represent variables of interest, and the graph structure is used to indicate dependencies between the variables. Arrows run into nodes from their direct influences (parents). Formally, the model represents the assumption that, given its parent nodes, each node is independent of all other nodes in the graph except its descendants.

Each subject is assumed to produce data through the same generative model with different parameters. The plate with  $i$  subjects corresponds to independent replications for all subjects. Each subject is assumed to have their own baseline of escalation,  $\phi_i$ , independently drawn from a Gaussian distribution with mean  $\mu^\phi$  and variance  $\sigma^\phi$  (termed as hyper-parameters). Each subject  $i$  is also associated with a latent group identity,  $Z_i$ , with  $Z_i = 1$  indicating that the subject comes from a ‘down group’ that shows less escalation decisions with ATD, and  $Z_i = 2$  indicating that the subject comes from an ‘up group’ that shows more escalation decisions with ATD.

The plate with  $j$  conditions corresponds to independent replications for all experimental conditions.  $t_j$  is an observed variable, with  $t_j = 1$  indicating the control condition (Tryp+), and  $t_j = 2$  indicating the treatment condition where the subjects received ATD (Tryp-).

In the control condition, random fluctuations around the baseline are drawn from a Gaussian distribution with mean of 0 and a variance  $\sigma^c$ . In the ATD condition, the increment in escalation depends on which group the subject is from. For example, if subject  $i$  is from the ‘up group’, the increment of escalation, denoted by  $\Psi_{ij}^u$  in the graphical model, will be drawn from a Gaussian with a positive mean  $\mu^u$  and variance  $\sigma^c$  and added to the baseline  $\phi_i$ . If subject  $i$  is from the ‘down group’, however, the increment  $\Psi_{ij}^d$  will be drawn from a Gaussian with a different positive mean  $\mu^d$  and variance  $\sigma^c$  and deducted from the baseline<sup>1</sup>.

The probability of escalation for subject  $i$  on the  $k$ th game of day  $j$ , denoted by  $\theta_{ijk}$ , is determined through the logistic function by the overall level of escalation. The observed escalation decision of each subject on each game,  $d_{ijk}$ , is a binary variable which is assumed to be independent and identically generated by  $\theta_{ijk}$  through a Bernoulli distribution. Overall, one would determine the likelihood of the observed data for all subjects for each combination of the hyperparameters, and each choice of individual parameters.

We use the conventions of representing continuous variables with circular nodes and discrete variables with square nodes, and unobserved variables without shading and observed variables with shading, and stochastic variables with

<sup>1</sup>Even though in theory,  $\Psi_{ij}^u$  and  $\Psi_{ij}^d$  could flip in sign following the way this model is specified, this almost never happened in the sampling process when the MC chains have converged.

single borders and deterministic variables with double borders. In addition, the plate encloses subsets of the graph that have independent replications in the model. For example, the probability of escalation for each subject in each game is a continuous variable that is not directly observable, but is determined by the overall level of escalation, therefore it is represented by a circular, unshaded, double-bordered node. On the other hand, the binary escalation decision is generated with probability and directly observable, therefore it is represented by a square, shaded node with a single border.

We built another model that captures the relationship between escalation and the effect of embodiment (Robot vs. Simulation). There are only two differences from the graphical model shown in Figure 1. First,  $t_j$  represents embodiment with  $t_1$  indicating robot and  $t_2$  indicating simulation; secondly,  $Z_i = 1$  indicates the latent group that shows more escalation when playing against simulation, and  $Z_i = 2$  indicate the group that shows more escalation when playing against robot.

In addition to escalation decisions, we are also interested in whether general strategies that subjects employed for the Hawk-Dove game were related to ATD and embodiment. Similar to the generative model shown in Figure 1, the observed T4T-type (WSLS-type) decisions are Bernoulli variables generated by probabilities of T4T (WSLS) usage, which are determined by an overall level of T4T (WSLS) usage that is the sum of a baseline level of usage and additive effect from ATD or embodiment.

## Results

### Inferring Strategies

All the graphical models were implemented using WinBUGS which uses MCMC methods. We evaluated all six graphical models by drawing 1000 posterior samples after a ‘burn-in’ period (early steps of MCMC where samples are not recorded so that the Markov chain is allowed time to converge) of 100 samples.

The results showed that both serotonin levels and the embodiment of a robot were influential factors in individual subject decision-making (See Figure 2). For example, panel (a) provides evidence that there are at least two sub-groups within a subject population with respect to how escalation may be altered by ATD. Subject 2, 6 and 8 fall in the ‘down’ group, and subject 1, 3, 4, 5 and 7 fall in the ‘up’ group. Panel (b) shows two sub-groups within a subject population with respect to how escalation may be altered by embodiment. Subject 3, 5 and 6 fall in the ‘down’ group (less escalation when playing with a robot), while all other subjects fall in the ‘up’ group (more escalation when playing with a robot). Individual differences in strategy usage were also affected by ATD and embodiment, as shown in Figure 2c through f. For example, Figure 2c shows individual differences in the effect of ATD on the proportion of T4T-type decisions. The red (dark) dots are subjects who had more T4T-type decisions when tryptophan depleted and the green (light) dots are subjects

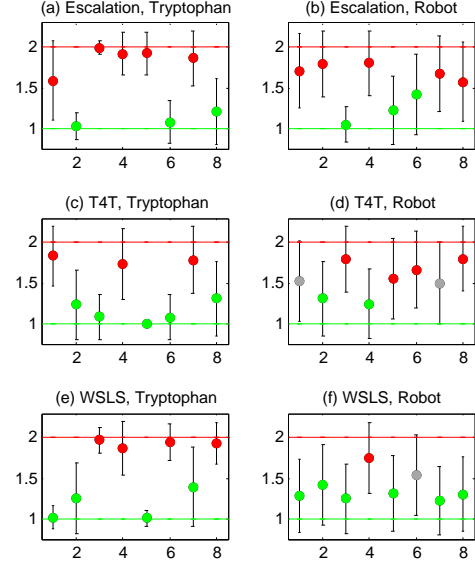


Figure 2: (a): Estimated group identities based on ATD’s effect on escalation. Horizontal axis shows subject indices. Vertical axis shows the posterior mean of the latent variable  $Z$ , with  $Z = 1$  indicating ‘down’ group and  $Z = 2$  indicating ‘up’ group based on the effect of tryptophan depletion. Red (dark) dots are subjects who were more likely to escalate when tryptophan depleted and green (light) dots are subjects who were less likely to escalate when tryptophan depleted. Error bars show the 95% Bayesian CI of the posterior mean. Gray dots imply ambiguous group identities. (b): Estimated group identities based on embodiment’s effect on escalation. Red dots are subjects who were more likely to escalate when playing against a robot. (c): Estimated group identities based on ATD’s effect on T4T usage. Red dots indicate more T4T-type decisions when tryptophan depleted. (d): Estimated group identities based on embodiment’s effect on T4T usage. Red dots indicate more T4T-type decisions when playing against a robot. (e)-(f): Estimated group identities on of the effect of ATD and embodiment on WSLS usage.

who had less T4T-type decisions when tryptophan depleted. In Figure 2d, red (dark) dots represent subjects who had more T4T-type decisions when playing against robot. Similarly, Figure 2e and 2f represent individual differences in how ATD and embodiment affect their usage of WSLS. All conditions show a tendency that subjects can be categorized into sub-groups with regard to how ATD or embodiment may affect strategy usage.

In general, subgroups are more clear when the effect of tryptophan depletion rather than the embodiment type is considered. Potential connections of subgroup identities across conditions for each subject is an interesting point raised here and will be addressed in future research.



## Predictions of Behavioral Patterns

To check the models' ability to describe the data accurately, we examined the posterior predictions of escalation decision and strategy usage for all subjects in all conditions. The posterior predictive is the prediction about observed data for each possible combination of parameter values under the model, where each combination is weighted according to its posterior probability. Our goal is not game-by-game prediction, rather, it is the prediction of overall rate of escalation decision and strategy usage in a specified condition, as captured in Figure 3. The x-axis shows experimental conditions, and the y-axis shows the proportion of escalation decisions or strategy usage. Each gray line is a subject. The colored (dark) lines are posterior predictives (summarized in the same way as the data) for each model, with filled vs. open circles representing whether subjects were inferred in the 'up' or 'down' groups corresponding to the condition (ATD or embodiment). It is clear that our models were able to capture individual differences and fit the data well, especially for the effect of ATD and embodiment on escalation and T4T usage. Predictions of proportions of WSLs-type decisions had larger fluctuations, but the general pattern of change in WSLs usage associated with ATD and embodiment were captured.

## Conclusions

In contrast to the results observed in our population analysis of subject behavior during Hawk-Dove game play (Asher et al., submitted), we found strong influences of lowering serotonin levels and of agent embodiment on individual decision-making by using hierarchical latent mixture models with Bayesian inference. The results from Figure 2 and Figure 3 show that individual differences likely exist within a subject population. Specifically, it appears that there are two groups of subjects within a given population. Our hypothesis is that within Game-theory-ATD and HRI studies, there exist two opposing subgroups within any given population of typical human subjects. These subgroups may exist due to inherent genetic variation. It is interesting to note that the results from Game-theory-ATD and HRI studies are mild when considering a single subject group. However, if one considers that there exist two groups with opposing affects, the results for these kinds of studies may be significantly more robust.

ATD had a strong effect on subject behavior and this behavior could be categorized in two groups: subjects who were more aggressive when tryptophan depleted and subjects who were less aggressive when tryptophan depleted. Aggressive, uncooperative behavior has been reported in behavioral studies in which serotonin levels were lowered through ATD (Schweighofer et al., 2008 ; Tanaka et al., 2007). However, individual variation both due to experience and genetic background can affect behavior. For example, there is widespread variation in the serotonin transporter gene 5-HTTLPR (Homberg & Lesch, 2011). Subjects carrying the short allele variant of the 5-HTTLPR outperform subjects with the long allele in an array of cognitive tasks and show

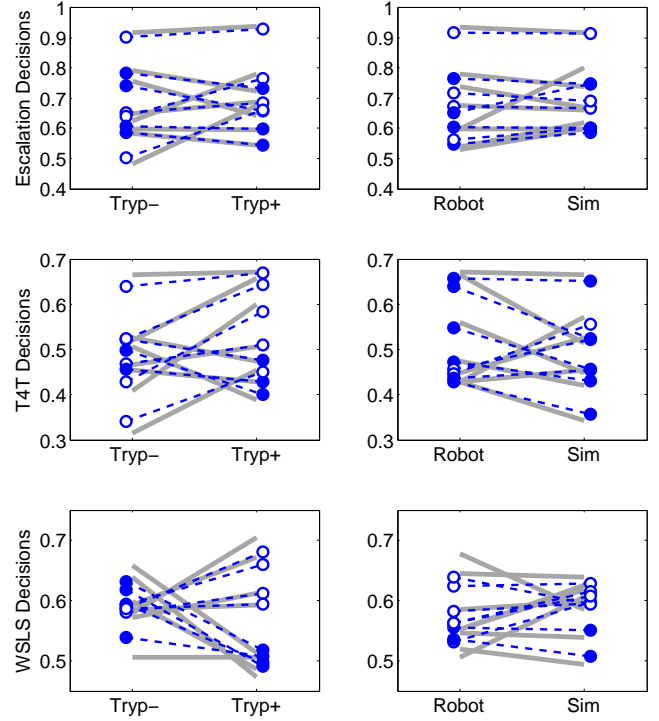


Figure 3: Predicted escalation rate by the model for all subjects with a comparison to the real data. Tryptophan depleted and non-depleted conditions labeled as 'Tryp-' and 'Tryp+'. The broken lines are predictions from the models, with filled vs. open circles representing subjects who show more escalation when tryptophan depleted vs. subjects who show less escalation when tryptophan depleted. The solid lines are the data. Labels of 'Robot' and 'Sim' indicate whether the game was played against the robot or the simulation.

increased social conformity under normal conditions. However, subjects carrying the long allele variant perform better under stressful conditions. The prevalence of these and other genetic polymorphisms in the human population suggests that there is an evolutionary advantage for this variability, such as optimizing competition or cooperation in different situations.

Embodiment had a strong effect on subject behavior and, similar to the ATD effect, this behavior could be categorized in two groups: subjects who were more aggressive when playing against a robot and subjects who were less aggressive when playing against a robot. Playing an opponent who is interactive and personified has previously been observed to evoke strong responses in subjects. For example, in the Ultimatum Game, subjects rejected more offers made by a human partner than those offers made by a computer, suggesting that participants had a stronger emotional reaction to unfair offers from humans than from a computer (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003). Indeed, such embodied models have been shown to elicit strong reactions in humans (Breazeal & Scassellati, 2002 ; Kidd & Breazeal, 2004)

and to exhibit more natural and complex behavior than pure simulations (Krichmar & Edelman, 2002, 2005). However, it appears that individuals responded differently and idiosyncratically to the presence of a robot in the present study.

Our results highlight the following ideas: first, the hierarchical latent mixture model's ability to capture individual differences; secondly, serotonin levels have differing effects on subject decision-making, and lastly embodiment plays a role in how likely subjects are willing to cooperate with an agent. Our results suggest that there are at least two opposing subgroups with respect to Game-theory-ATD and HRI studies. We suggest the possibility that these subgroups may have emerged as a result of genetic variation. A next step towards investigating this hypothesis involves genetic testing of polymorphisms responsible for subject variance and their decision-making in competitive and cooperative games of Game theory along with human robot interactions.

### Acknowledgment

This work was supported by the National Science Foundation (EMT/BSSE Award No.: 0829752) and the Office of Naval Research (Award No.: N000140910036). MDL and SZ acknowledge support from the Air Force Office of Scientific Research Award FA9550-11, and Australian Research Council Grant DP110100797.

### Références

- Asher, D., Zaldivar, A., Barton, B., Brewer, A., & Krichmar, J. (submitted). Reciprocity and retaliation in social games with adaptive agents. *IEEE Transactions on Autonomous Mental Development*.
- Asher, D., Zaldivar, A., & Krichmar, J. (2010). Effect of neuromodulation on performance in game playing: A modeling study. In *IEEE 9th international conference on development and learning* (p. 155-160).
- Bell, C., Hood, S., & Nutt, D. (2005). Acute tryptophan depletion. part II: clinical effects and implications. *Aust N Z J Psychiatry*, 39, 565-574.
- Boureau, Y., & Dayan, P. (2011). Opponency revisited: competition and cooperation between dopamine and serotonin. *Neuropsychopharmacology*, 36, 74-97.
- Breazeal, C., & Scassellati, B. (2002). Robots that imitate humans. *Trends in Cognitive Sciences*, 6, 481-487.
- Crockett, M., Clark, L., Tabibnia, G., Lieberman, M., & Robbins, T. (2008). Serotonin modulates behavioral reactions to unfairness. *Science*, 320, 1739.
- Homberg, J., & Lesch, K. (2011). Looking on the bright side of serotonin transporter gene variation. *Biol Psychiatry*, 69, 513-519.
- Hood, S., Bell, C., & Nutt, D. (2005). Acute tryptophan depletion. part I: clinical effects and implications. *Aust N Z J Psychiatry*, 39, 558-564.
- Kidd, C., & Breazeal, C. (2004). Effect of a robot on user perceptions. In (p. 3559-3564).
- Krichmar, J., & Edelman, G. (2002). Machine psychology: autonomous behavior, perceptual categorization and conditioning in a brain-based device. *Cerebral Cortex*, 12, 818-830.
- Krichmar, J., & Edelman, G. (2005). Brain-based devices for the study of nervous systems and the development of intelligent machines. *Artificial Life*, 11, 63-77.
- Lee, D. (2008). Game theory and neural basis of social decision making. *Nature Neuroscience*, 11, 404-409.
- Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, 15, 1-15.
- Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance on bandit problems. *Cognitive Systems Research*, 12, 164-174.
- Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge University Press.
- Newell, B., & Lee, M. (2011). The right tool for the job? comparing an evidence accumulation and a naive strategy selection model of decision making. *Journal of Behavioral Decision Making*, 24, 456-481.
- Nowak, M., Page, K., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science*, 289, 1773-1775.
- Rilling, J., & Sanfey, A. (2011). The neuroscience of social decision making. *Annual Review Psychology*, 62, 23-48.
- Rouder, J., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). Unknown title. *Psychonomic Bulletin & Review*, 12, 195-223.
- Sanfey, A., Rilling, J., Aronson, J., Nystrom, L., & Cohen, J. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, 300, 1755-1758.
- Schweighofer, N., Bertin, M., Shishida, K., Okamoto, Y., Tanaka, S., Yamawaki, S., et al. (2008). Low-serotonin levels increase delayed reward discounting in humans. *J Neurosci*, 17, 4528-4532.
- Skyrms, B. (2001). The stag hunt. In *Presidential address pacific division of the american philosophical association*.
- Tanaka, S., Schweighofer, N., Asahi, S., Shishida, K., Okamoto, Y., Yamawaki, S., et al. (2007). Serotonin differentially regulates short- and long-term prediction of rewards in the ventral and dorsal striatum. *Public Library of Science ONE*, 12, 2.
- Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E. (2010). Bayesian parameter estimation in the expectance valence model of the Iowa gambling task. *Journal of Mathematical Psychology*, 54, 14-27.
- Wood, R., Rilling, J., Sanfey, A., Bhagwagar, Z., & Rogers, R. (2006). Effects of tryptophan depletion on the performance of an iterated prisoner's dilemma game in healthy adults. *Neuropsychopharmacology*, 31, 1075-1084.
- Zaldivar, A., Asher, D., & Krichmar, J. (2010). *Simulation of how neuromodulation influences cooperative behavior*. Springer-Verlag Lecture Notes on Artificial Intelligence (LNAI 6226).

# Lexical Access Across Languages: A Multinomial Model of Auditory Distraction.

C. Philip Beaman (c.p.beaman@reading.ac.uk)

School of Psychology & Clinical Language Sciences, University of Reading  
Earley Gate, Whiteknights, Reading RG6 6AL, UK

## Abstract

Recall in many types of verbal memory task is reliably disrupted by the presence of auditory distracters, with verbal distracters frequently proving the most disruptive (Beaman, 2005). A multinomial processing tree model (Schweickert, 1993) is applied to the effects on free recall of background speech from a known or an unknown language. The model reproduces the free recall curve and the impact on memory of verbal distracters for which a lexical entry exists (i.e., verbal items from a known language). The effects of semantic relatedness of distracters within a language is found to depend upon a redintegrative factor thought to reflect the contribution of the speech-production system. The differential impacts of known and unknown languages cannot be accounted for in this way, but the same effects of distraction are observed amongst bilinguals, regardless of distracter-language.

**Keywords:** Auditory distraction; bilingualism; memory; MPT models.

## Introduction

Auditory distraction is a simple and inevitable fact of everyday experience, stemming from the role of audition as the “sentinel of the sense” (Handel, 1989; Jones, Hughes & Macken, 2010). A considerable body of experimental data has been amassed, particularly with regard to immediate serial memory (e.g., Jones et al, 2010), indicating that – as a predictor of disruption experienced to the primary task – the lexical content of verbal auditory distracters is less important than the acoustic properties of the signal. For example, to reliably disturb immediate serial recall it is necessary for an auditory stream to consist of multiple, varying items – a single repeated item is much less disruptive (Jones & Macken, 1993). Nevertheless, given the verbal nature of most primary tasks shown to be vulnerable to interference from auditory distracters, it would be surprising if no effect of the lexical properties of the distracters was ever observed.

One task which reliably shows more disruption from meaningful verbal distracters that are semantically related to the material being studied than from semantically unrelated material is categorical free recall. In this task, participants are asked to recall, in any order that occurs to them, a series of items all drawn from the same semantic category (e.g., a fruit, a vegetable, or a four-footed animal) which are presented to them visually, one item at a time. Recall in this task is disrupted by the presence of auditory-verbal distracters but is disrupted more when these distracters are drawn from the same category as the to-be-recalled material. Participants are always asked to ignore anything they may hear, and are never tested on the content of the auditory stream. Results obtained within this task show the extent,

and nature, of the processing to which the auditory distracters are subjected. Similarities and differences between results obtained with category free-recall and with identical distracters applied during immediate serial recall also indicate the generality, and specificity, respectively, of both the auditory distraction effect and memory models which aim to account for this effect.

## The Schweickert (1993) model.

The model tested in this study is Schweickert’s (1993) multinomial model of immediate recall. This model has previously been applied to short-term memory for serial order, in which items must be recalled in the order in which they appeared and are scored as incorrect if an item appears in the wrong position in the serial recall protocol. This model was able to successfully account for the interaction in serial recall data between the frequency of words within the English language (the word frequency effect) and the point at which they were presented in a to-be-recalled list (Hulme, Roodenrys, Schweickert, Brown, Martin & Stuart, 1997). The same model also accounted for a distracter-word frequency effect, that is an effect on immediate serial recall of whether an auditory distracter – presented concurrently with the visual presentation of the to-be-recalled list – was of high or low frequency, with low frequency words causing the most distraction (Buchner & Erdfelder, 2005). As such, the model is a useful one for examining the effects of lexical properties of the auditory distracters, and how these might interact with lexical processing of the to-be-recalled items.

The multinomial model is conceptually straightforward, the structure of the model is given in Figure 1. An item is either directly recalled in an intact form, with probability  $i$ , or else the representation of the item exists only in a degraded form and it must be redintegrated, or reconstructed, which is only possible with probability  $r$ .

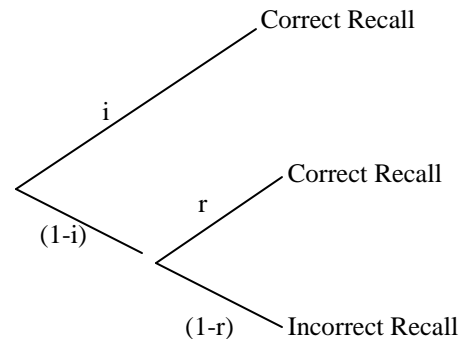


Figure 1. Diagrammatic representation of Schweickert’s (1993) multinomial processing tree model.

The form of the model thus allows for two means by which items can be correctly recalled – they may already exist in an “intact” form and be readily available, or they may require reconstruction. If both of these processes fail, the item cannot be recalled. The model has thus far been applied only to immediate serial recall – that is recall commonly considered to be from “short-term memory” but the existence of two distinct processes, each underlying recall in a different way, calls to mind earlier models previously applied to free recall (e.g., Atkinson & Shiffrin, 1968) which also assumed dual components to recall, so it is of interest to examine whether Schweickert’s model for serial recall can also be applied to free recall, and in particular the free recall of items from within a single category, which will require the model to generate the well-known serial position function typical of free recall, with primacy and extensive recency (Murdoch, 1962), rather than the serial recall curves, with extensive primacy and limited recency, generated by Buchner and Erdfelder (2005) and Hulme et al. (1997).

### Lexicality and recall.

As with all multinomial models, the goodness-of-fit between the model and the data is assessed by finding the values for the free parameters ( $i$  and  $r$ , in this instance) which produce expected data closest to those observed in behavioral testing. A goodness-of-fit test then determines whether the expected values differ significantly from the observed data (Bachelder & Reifer, 1999). In Hulme et al.’s (1997) study,  $i$  was held constant across simulations of different experimental conditions but allowed to vary across serial position to produce the serial position curve indicative of serial recall. That is, for a 7-item to-be-recalled list, different parameter values would exist for  $i_1, i_2, \dots, i_7$ , but these would be identical regardless of experimental condition.  $r$  was held constant within an experimental condition but allowed to vary across conditions. Hulme et al. (1997) argued that variation of  $r$  across experimental condition reflected the effect of word frequency upon the redintegration process, with representations of higher-frequency words supporting the redintegration more effectively than representations of low-frequency words (so  $r_{\text{high-frequency}} > r_{\text{low-frequency}}$ ). It was assumed that verbal short-term memory is essentially a by-product of processes involved in speech perception and speech production (Hulme, Maughan & Brown, 1991), with redintegration an integral part of speech production, representing the “clean-up” of noisy representations (e.g., within an underlying connectionist network). Similarly, Buchner and Erdfelder (2005) concluded that the word-frequency of the *distracters* must impact upon the probability of retrieving an intact representation ( $i$ ) because a model varying  $r$ , but with equivalent values of  $i$  across experimental conditions differed significantly from the data, whereas the expected data from a model with equivalent  $r$  but varying  $i$  across the experimental conditions, such that  $i_{\text{high-frequency distractor}} > i_{\text{low-frequency distractor}}$  were statistically indistinguishable from the

observed data. Buchner and Erdfelder (2005) conclude in favor of an account in which, “low-frequency distracter words require more processing resources that could otherwise have been used for keeping the memory representations of the target words active and intact” (p. 89).

The study by Buchner and Erdfelder (2005) is curious in that there is no necessary *a priori* reason why low-frequency distracters should attract more attention, or require more processing resources, than high frequency distracters – as these authors are careful to note. Previous studies, however, all used immediate serial recall rather than – as studied here – categorical free recall which draws upon semantic memory and appears to be more sensitive to the lexical properties of the auditory distracters than serial recall (Marsh, Hughes & Jones, 2008). In particular, auditory distraction may also occur within a semantic-memory fluency task in which speech production processes presumably play a large part (Jones, Marsh & Hughes, 2012). On this basis, and using the logic employed by Hulme et al. (1997), if it is possible to apply the Schweickert (1993) model to categorical free recall then the lexical effects of the auditory distracters should be most evident on the  $r$  parameter, reflecting interference with speech production systems, rather than the  $i$  parameter which might be interpreted – as, for example, by Buchner and Erdfelder (2005) – as a more general effect, possibly the result of an attentional mechanism drawing off processing resources.

### Modeling Recall and Disruption Within and Across Languages

To test these possibilities and simultaneously test the generality of the Schweickert (1993) model, the model was applied to a set of data obtained from English monolinguals and Welsh-English bilinguals. Bilinguals were used to test the possibility that distraction effects associated with the meaning of speech cannot be inhibited, and by extension the idea that the meaning of speech cannot be ignored. The free recall task was presented in one language (English) with speech distracters in either English or Welsh. The distracting speech (in either English or Welsh) consisted of words related to the same subject, or to a different subject. The typical finding is that both unrelated and semantically related speech (distracter words from the same category as the to-be-recalled items) give a distraction effect, but that there is a greater distraction effect for related speech (Neely & LeCompte, 1999). The effect, even for unrelated speech, is lexical rather than acoustic, because non-words and sinewave speech tokens do not disrupt recall (Marsh et al., 2008).

Where does the disruption originate? If the effect of related speech is conceptual in nature, originating from the organization of the speech planning and production system, then one might expect bilinguals to show equivalent disruptive effects of the meaning of the words regardless of their language of origin (English or Welsh). Conceptual effects of the irrelevant speech arising from

the disruption of speech organization in this way should be reflected in reductions of the  $r$  parameter of the model. Alternatively, if the effect is a non-specific lexical/attentional effect akin to that reported by Buchner and Erdfelder (2005) then the bilinguals might be expected to perform more like monolinguals when the irrelevant speech accesses a lexicon (Welsh) other than the one they are employing for the focal task (English). Any residual difference between the two groups, or between the disruption caused by related and unrelated speech should be accountable in terms of the  $i$  parameter, with lexical/attentional effects reducing the values of this parameter for those conditions that show the most disruption.

For the experiment, twenty-eight English monolinguals and twenty-eight Welsh-English bilinguals each viewed 28 trials of 12 target words, in English, visually-presented for free recall. Stimuli were chosen from semantic categories of the Van Overschelde, Rawson, and Dunlosky (2004) category norms. Items from positions 13-24 in the category-norm lists were used to form target lists and items from positions 1-12 were used as distracters. On half the trials, the auditory distracters were taken from the same category as the targets (e.g., both sets of stimuli were types of animals, and no “shape” exemplars were presented). On the remaining trials, the distracter items were taken from one category of the pair (e.g., fruit) and targets from the other category (e.g., carpenter’s tools). Additionally, half of the distracters were presented in English, and half in Welsh, yielding four separate conditions each experienced by both English monolinguals and Welsh-English bilinguals: English unrelated distracters (EU), English related distracters (ER), Welsh unrelated distracters (WU) and Welsh related distracters (WR). Space precludes a full analysis of the behavioral results, but a bar chart of the overall impact of distracters on both groups is given in Figure 2.

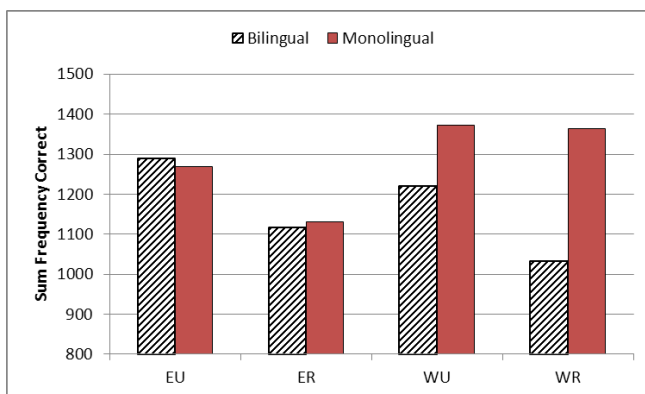


Figure 2. Total frequency of correct recalls across all conditions, summed across serial positions.

### Unrelated distracters across languages.

As with Hulme et al. (1997) and Buchner and Erdfelder (2005)  $i$  was allowed to vary across serial position, thereby

implementing the serial position function, but there was a single value for  $r$  regardless of serial position. multiTree software (Moshagen, 2010) was used to implement the models. In what follows, only models which fit the data are presented graphically.

Examining first the unrelated speech condition for bilingual participants, that is distracters – presented in either English or Welsh – semantically unrelated to the English language targets, the results could be modeled by assuming that neither  $i$  nor  $r$  varied across conditions with no significant difference between observed and expected results,  $G^2 = 15.67$ ,  $df = 11$ ,  $p = .15$ . This confirms the viability of the Schweickert model for categorical free recall and shows that – for Welsh-English bilingual participants – semantically unrelated distracters have an equivalent effect upon free recall of English words regardless of the language of the distracter.

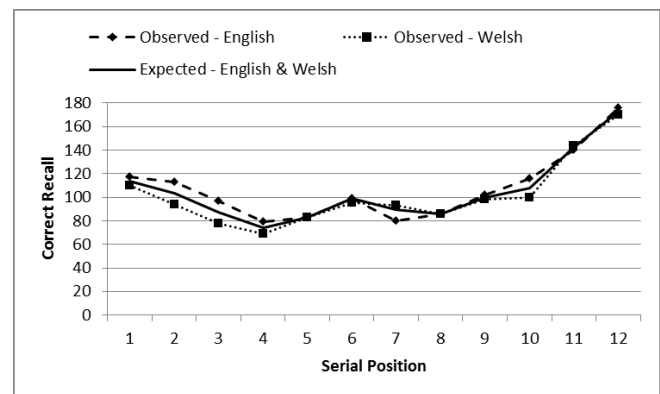


Figure 3. Frequency of correct recalls across serial position by Welsh-English bilinguals for unrelated distracters in English and Welsh. Expected values according to the MPT model are given by the solid line.

For the English monolinguals, a similarly constrained model differs significantly from the data,  $G^2 = 33.95$ ,  $df = 11$ ,  $p < .001$ . Thus, for English monolinguals, there is a difference between unrelated English and unrelated Welsh words as distracters. Relaxing the constraints upon the model by allowing  $r$  to vary across conditions does not improve the fit of the model,  $G^2 = 30.6$ ,  $df = 10$ ,  $p < .001$ . Thus, whatever effect the presence of unrelated verbal distracters in a known language (which have a lexical status) has over the effect of distracters in an unknown language (for which no lexical entry exists), cannot be accounted for within the Schweickert model by a reintegration process. Unfortunately, it is not possible to test the effects of similarly freeing the constraints upon the  $i$  parameter, as investigated by Buchner and Erdfelder (2005), because varying  $i$  across conditions as well as serial positions imposes too few constraints on the model (Bachelder & Reifer, 1999).

### Related distracters across languages.

Applying the model to bilingual English and Welsh speakers exposed to irrelevant distracter speech in either English or Welsh that was semantically related to the English language to-be-remembered stimuli, a model in which  $i$  varied across serial position but  $i$  and  $r$  were identical regardless of the language of the distracter provided a good fit to the data,  $G^2 = 8.67$ ,  $df = 11$ ,  $p = .65$ . Thus, the distraction effects for bilinguals can be modeled using the same parameter values regardless of the language in which the distracters were presented.

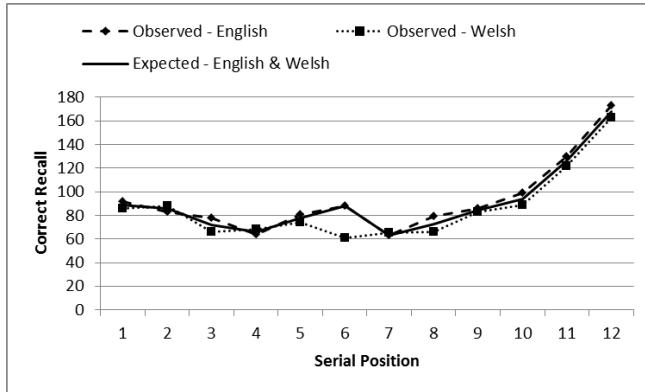


Figure 4. Plot of frequency of correct recalls by Welsh-English bilinguals for semantically related distracters in English and Welsh. Expected values according to the MPT model are given by the solid line.

Unsurprisingly, a similar attempt to model the impact of semantically-related auditory distracters in both English and Welsh on monolingual English speakers was unsuccessful, with the best-fitting model differing substantially from the data,  $G^2 = 60.27$ ,  $df = 11$ ,  $p < .001$ . Allowing  $r$  to vary between Welsh and English distracter conditions was also insufficient to substantially improve the fit of the model,  $G^2 = 26.62$ ,  $df = 10$ ,  $p = .003$ . Thus, in terms of the Schweickert (1993) model, the extra impact, upon a visual-verbal free recall task, of an auditory distracter being in a known language must be upon factors other than redintegration. This is true regardless of whether the auditory distracter is unrelated, or semantically related, to the to-be-remembered targets.

### Comparing unrelated and related distracters within languages.

In addition to looking at the effects of bilingualism upon auditory distraction when the distracters are presented in different languages, it is also of interest to compare the effects of distracters *within* a single language. Using the model to investigate the effects of shifting the language of distracters has revealed that the language of the distracter is irrelevant provided it is a known language (Figures 2-4) and that the difference between known and unknown language distracters cannot be captured by a single redintegrative

factor. This is consistent with reports by Buchner and Erdfelder (2005) that the frequency of occurrence of words presented as distracters impacted upon the  $i$  parameter and not the  $r$  parameter, which they interpret as an attentional effect. However, there are *a priori* reasons to suppose that the difference between semantically-related and unrelated distracters *could* be captured by just such a single, redintegrative factor.

Hulme et al. (1991, 1997) argued that – in immediate serial recall – the effects of word frequency, captured by the  $r$  parameter in the Schweickert model, reflect the operation of a speech production system yoked into supporting recall. In an investigation of the effects of distraction upon a verbal fluency task of the kind frequently used to explore the speech production system, Jones et al. (2012) found an effect of semantically-related speech. Thus, it seems reasonable to suggest,

- 1) In free recall as in serial recall, speech production systems may play a role – perhaps by supporting covert articulatory rehearsal. This may particularly apply to categorical free recall, free recall of items from specific, reasonably circumscribed semantic categories.
- 2) If so, the effects of specifically semantically-related distracters might be traceable to this system via their impact upon the  $r$  parameter in the model.

Applying the model to the data, this time for the effects of related and unrelated English speech upon free recall by monolinguals, a model that does not differ significantly from the data is obtained by varying only the parameter  $r$  between the unrelated and related speech conditions,  $G^2 = 15.51$ ,  $df = 10$ ,  $p = .11$ . The fit of the model is given in Figure 5. To ensure that this fit was not possible simply because there was no difference in the data between the effects of related and unrelated speech, parameter  $r$  was constrained to be equivalent in both conditions. The resulting model differed significantly from both the previous model,  $\Delta G^2 = 11.84$ ,  $df = 1$ ,  $p < .001$  and the data,  $G^2 = 27.35$ ,  $df = 11$ ,  $p = .004$ .

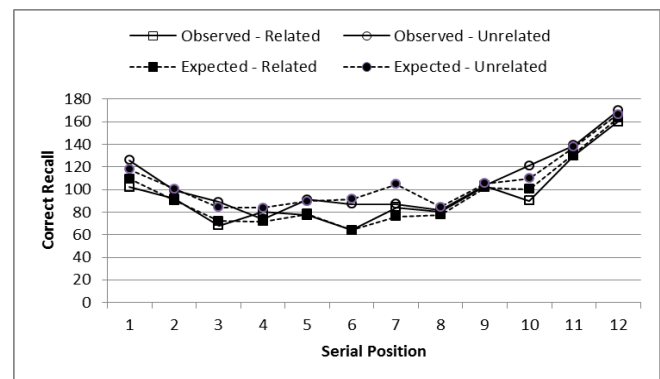


Figure 5. Correct recalls by English monolinguals for semantically related and unrelated distracters in English.



Expected values according to the MPT model are given by the dashed lines and observed values by the solid lines.

Finally, the model was applied to the performance of Welsh-English bilinguals in the presence of English and Welsh distracters that could be either semantically-related, or unrelated to the target lists. Constraining the values of  $i$  to be equivalent regardless of whether the distracters were semantically related or not, but allowing the values of  $r$  to vary, resulted – as in the case of the English monolinguals – in a model that did not differ significantly from the data observed,  $G^2 = 20.67$ ,  $df = 20$ ,  $p = .42$ . This was defined as the baseline model, and the output of this model is shown, for Welsh and English, in Figures 6a and 6b.

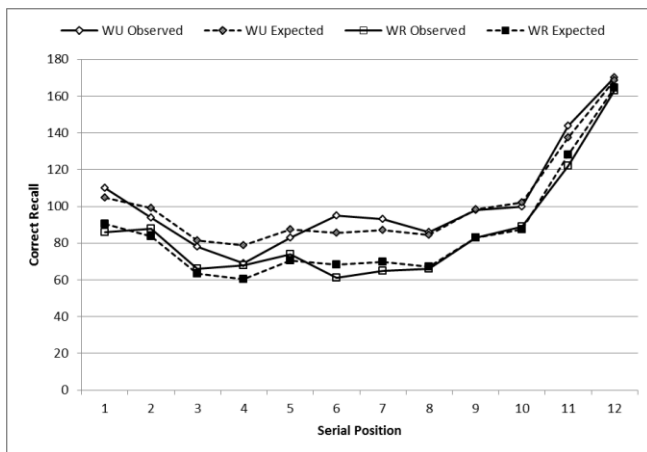


Figure 6a. Frequency of correct recalls across serial position by Welsh-English bilinguals for semantically related (R) and unrelated (U) distracters in Welsh (W). Expected values according to the MPT model are given by the dashed lines, and observed values by the solid lines.

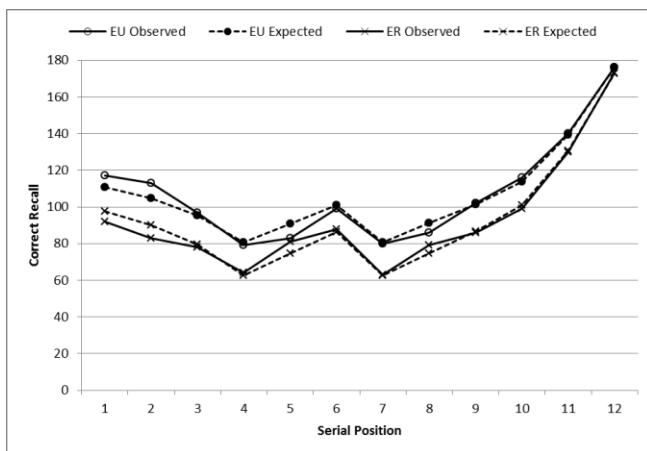


Figure 6b. Correct recalls by Welsh-English bilinguals for semantically related (R) and unrelated (U) distracters in English (E). Expected values according to the MPT model are given by the dashed lines

Additionally constraining the  $r$  values for semantically-related speech distracters to be equivalent across languages, and likewise constraining  $r$  for unrelated distracters to be equivalent across languages, produced a model that did not differ significantly from the baseline model,  $\Delta G^2 = .007$ ,  $df = 2$ ,  $p = .996$ , nor from the data,  $G^2 = 20.67$ ,  $df = 22$ ,  $p = .54$ . However, it was not possible to fit a model in which the  $r$  values were equated within languages, regardless of the semantic relationship between distracters (i.e.,  $r_{EU} = r_{ER}$  and  $r_{WU} = r_{WR}$ ), and were allowed to vary across languages (i.e.,  $r_E \neq r_W$ ) – such a model differed significantly from both the baseline model,  $\Delta G^2 = 57.16$ ,  $df = 2$ ,  $p < .001$ , and from the data,  $G^2 = 77.83$ ,  $df = 22$ ,  $p < .001$ .

## Discussion

The modeling results reported here show that it is possible to extend the Schweickert (1993) model of immediate serial recall to also apply to free recall, consistent with research (such as that of Tan and Ward, 2000), emphasizing similarities between immediate serial and free recall. More importantly, the model also shows that – for bilinguals – the effects of distracters presented in either of their languages are equivalent, even if the primary task on which the impact is observed (free recall in this case) is conducted wholly within one of those languages. Important issues are still to be worked-out with regard to bilingualism, e.g., second-language (L2) proficiency and the age at which L2 was learned, but it is notable that no simple means was found to model, in a similar manner, the effects of English and Welsh distracters on the performance of English-speaking monolinguals. Clearly therefore, for monolinguals, the effects of English and Welsh speech upon categorical free recall performance differed, even when the English distracter speech was semantically-unrelated to the target words. In this, the categorical free-recall task differs from other, notably serial recall, tasks in which foreign speech (including Welsh) has been played to participants, with equivalent effects to speech in their native language (e.g., Jones, Miles & Page, 1990). Whatever the basis for this difference, it cannot be located within a redintegrative stage affected by distracters from known versus unknown languages as manipulation of this parameter did not improve the fit of the model. This is broadly consistent with Buchner and Erdfelder's (2005) finding that varying the word-frequency of distracters within a known language was also more accurately modeled by varying the  $i$  rather than the  $r$  parameter within the Schweickert (1993) model. Unfortunately, known limitations of the modeling methodology employed (principally, the requirement for the model to be identifiable; Bachelder & Reifer, 1999), prevent further exploration of this issue given the experimental design available (for more discussion of this issue, see Buchner & Erdfelder, 2005).

A more interesting, and more positive, finding arising from the current study is that the semantic effects of distraction within a known language *can* be accounted for in terms of a redintegration stage. Comparison of semantically-

related with unrelated English distracter words amongst English monolinguals and Welsh-English bilinguals, and a similar comparison of semantically-related with unrelated Welsh distracter words in the bilingual group, all produce this same result (see Figures 5, 6a and 6b).

This finding is consistent with the data from Jones et al. (2012) showing a semantic distraction effect upon verbal fluency tasks generally considered to tap recall from semantic memory prior to speech-formulation and production, and is consistent with the hypothesis that redintegration reflects a “clean-up” stage in recalling items within the speech production system (Hulme et al., 1991, 1997). “clean-up” may seem to imply simply the filling in of blanks, or correcting of misinformation, within a single already-recalled item by reference to longer-term memory or lexical storage (Levelt, 1999). However, it may fulfill a more important function, namely one of identifying – by means of “cleaning-up” an incomplete or noisy representation – which one of several possible items is the correct one to recall in a particular instance (Nairne, 2003, personal communication). This is likely to be particularly important when, as in the current situation, recall is always from a list in which all of the target items are drawn from the same semantic category and therefore share many semantic and conceptual features. Under such conditions, identifying the correct item from several possible candidates is likely to be particularly important.

In this instance, it seems reasonable to suggest that the presence of semantically-related distracters compromises the redintegrative process at the level of retrieval of the word-concept, resulting in greater distraction than is seen with semantically-unrelated distracters. This suggestion is further supported by the fact that bilinguals show a semantic distraction effect from a language other than the one in which they are nominally working (that is, the English-language memory task). This implies that although the effects of speech on categorical free recall may be lexical (Marsh et al., 2008) the specific effects of semantic distraction across languages are conceptual, not lexical, in nature.

### Acknowledgments

The multinomial modeling research reported here was supported by ESRC grant RES-062-23-1752 to Dylan M. Jones and C. Philip Beaman. Thanks to John Marsh for providing the experimental data and to Robert Hughes for the Welsh language stimuli.

### References

- Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Spence, K. W., & Spence, J. T. *The psychology of learning and motivation* (Volume 2). New York: Academic Press. pp. 89–195
- Bachelder, W. H., & Reifer, D. M. (1999). Theoretical and empirical review of multinomial processing tree modeling. *Psychonomic Bulletin & Review*, 6, 57-86.
- Beaman, C. P. (2005). Auditory distraction from low-intensity noise: A review of the consequences for learning and workplace environments. *Applied Cognitive Psychology*, 19, 1041-1064.
- Buchner, A. & Erdfelder, E. (2005). Word frequency of irrelevant speech distractors affects serial recall. *Memory & Cognition*, 33, 86-97.
- Handel, S. (1989). *Listening: An introduction to the perception of auditory events*. Cambridge, Ma.: MIT Press.
- Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory & Language*, 30, 685-701.
- Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 23, 1217-1232.
- Jones, D. M., Hughes, R. W., & Macken, W. J. (2010). Auditory distraction and serial memory: The avoidable and the ineluctable. *Noise & Health*, 12, 201-209.
- Jones, D. M., & Macken, W. J. (1993). Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 19, 369-381.
- Jones, D. M., Marsh, J. E., & Hughes, R. W. (2012). Retrieval from memory: Vulnerable or inviolable? *Journal of Experimental Psychology: Learning, Memory & Cognition*. In press.
- Jones, D. M., Miles, C., & Page, C. (1990). Disruption of proof-reading by irrelevant speech: Effects of attention, arousal or memory? *Applied Cognitive Psychology*, 4, 89-108.
- Levelt, W. J. M. (1999). Models of word production. *Trends in Cognitive Sciences*, 3, 223-232.
- Marsh, J. E., Hughes, R. W., & Jones, D. M. (2008). Auditory distraction in semantic memory: A process-based approach. *Journal of Memory & Language*, 58, 682-700.
- Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, 42, 42-54.
- Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64, 482-488.
- Neely, C. B., & LeCompte, D. C. (1999). The importance of semantic similarity to the irrelevant speech effect. *Memory & Cognition*, 27, 37-44
- Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition*, 21, 168-175.
- Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory & Language*, 50, 289-335.



# Diagramming Phenomena for Mechanistic Explanation

**William Bechtel (bechtel@ucsd.edu)**

Department of Philosophy, University of California, San Diego  
La Jolla, CA 92014 USA

**Adele Abrahamsen (aabrahamsen@ucsd.edu)**

Center for Research in Language, University of California, San Diego  
La Jolla, CA 92014 USA

## Abstract

As part of an inquiry into how diagrams figure in scientific practice, we examine diagrams that represent phenomena involving circadian rhythms. Different diagrammatic formats are developed and revised over time to best represent different phenomena for which explanations will be sought. Some diagrams are less transparent than others, so learning is often required in order to see the information conveyed.

**Keywords:** Diagrams; Graphs, Mechanistic explanation, Visual representation; Circadian rhythms.

## Introduction

The notion of representation covers a lot of territory in cognitive science, encompassing both internal and external encodings of information and a variety of formats. Cognitive scientists have long focused on language-like internal representations, with some dispute over possible ways they might be supplemented by analog formats. Recent years have brought increased attention to external representations and especially to those incorporating analog formats—that is, diagrams. For example, Hegarty (2004) has shown how individuals perform simulations with diagrams in solving problems, and Cheng (2011) has explored how alternative diagramming techniques can foster learning. However, except for the pioneering analysis by Nessessian (2008) of the role of diagrams in Maxwell’s discoveries, there has been little investigation of the use of diagrams in science. Almost all scientific papers include diagrams, and readers often focus on these as they navigate a paper. They are well suited not only for displaying instruments, techniques, multistep procedures, and results but also scientific reasoning. Most generally this involves the construction, evaluation and revision of hypotheses but our particular interest is in sciences pursuing mechanistic explanations, notably the life sciences.

The project of explaining a phenomenon by identifying and understanding the mechanism responsible for it has roots in the scientific revolution beginning in the 16th century. Descartes posited that phenomena such as magnetic attraction are generated by the coordinated activities of constituent parts (in his case, hypothesized corpusecles). He applied the idea of contact action between particles to explain not just physical phenomena, but nearly all phenomena exhibited by living organisms. The only exceptions were reasoning and language use, which he attributed to an immaterial mind because he could not conceive of a mechanism capable of constructing novel, semantically appropriate

thoughts or sentences. The idea of mechanistic explanation quickly took root in biology. Although resisted by vitalists, who contended that something beyond physical processes was required for the functions of life, other early investigators of physiological phenomena embraced mechanistic explanations. As their inquiries progressed, researchers expanded the range of operations involved in biological mechanisms beyond Cartesian physical contact. Newtonian forces, chemical bonding, and electrical conductance were among the operations used in explaining such phenomena as metabolism, nerve transmission, and heredity.

Mechanistic explanation was largely overlooked by 20<sup>th</sup> century philosophers of science, who drew from physics the idea that scientists explain phenomena by deriving them from laws (Hempel, 1965). More recently, philosophers focusing on the life sciences have moved the spotlight once again to mechanistic explanation (Bechtel & Richardson, 1993/2010; Bechtel & Abrahamsen, 2005; Machamer, Darden, & Craver, 2000). Typically, life scientists treat the system that generates a phenomenon as a mechanism. They decompose it into parts and operations and then recompose it (conceptually, physically, or mathematically) to arrive at an account of how the coordinated performance of these operations could indeed generate the phenomenon.

Although one may try to describe linguistically the parts and operations of a mechanism and how they interact, often telling a narrative about how each part in succession performs its operation, diagrams generally provide a more useful representational format for conceptualizing and reasoning about a mechanism. Parts may be represented by labels, symbols, or abstract shapes, and the operations by which they interact represented by arrows. Diagrams can illustrate the structural and functional relations between many components and allow viewers to direct their attention successively to different activities that may be occurring concurrently in the mechanism.

The initial step in mechanistic research, though, is delineation of the phenomenon to be explained, and that is where we begin our inquiry into diagrams. Linguistic descriptions of phenomena have been the focus in many philosophical accounts of mechanistic explanation (e.g., “proteins are synthesized by constructing strings of amino acids in the order specified in a sequence of DNA”). However, scientists typically work with much more specific accounts of phenomena, often incorporating numerical values determined in their research. Frequently the numerical data relied upon in char-

acterizing the phenomenon is presented in tables. As Bogen and Woodward (1988) made clear, however, explanations are directed not at the data but rather at the pattern extracted from the data—the phenomenon. Some data patterns can be captured in one or a few equations, such as  $\Delta I / I = k$  (Weber's law). Even when equations suffice, but especially when they do not, scientists turn to diagrams to present the phenomenon. Diagrams turn out to be extremely useful for phenomena that exhibit interesting dynamics—patterns of change over time. Well-known examples include tide tables and EEG recordings.

To gain an understanding of these uses of diagrams in the actual practice of science, we focus on a domain of biology in which dynamics are fundamental: circadian rhythm research. Circadian rhythms are oscillations in activity with an approximately 24-hour cycle. They are endogenously generated but entrained to the timing of the day-night cycle in specific locales at different times of the year. They have been identified wherever sought, not only in animals but also in plants, fungi, and even cyanobacteria. They regulate a vast array of physiological processes (e.g., basic metabolism and body temperature) and behaviors (e.g., locomotion and reaction times in cognitive tasks).

Diagrams, of course, are processed visually. Although visual processing is highly complex (involving nearly half of the cortex in primates; see van Essen & Gallant, 1994), seeing often seems transparent: as we look out into the world, we have the impression that we directly perceive the identity and arrangement of objects in the visual field. The apparent transparency of diagrammatic representations of phenomena is part of their appeal—but this is deceptive. Transparent seeing is often the result of a great deal of learning. There are numerous experiments showing this, using complex scenes or illusory stimuli, but diagrams offer potent demonstrations as well. Some techniques of diagramming are so familiar and straightforward that we readily *see* what the diagram is meant to convey. But other techniques are new to us, and we must go through a process of learning before we see what is presented in the diagram. This is clearly true of scientific diagrams, as we will illustrate in examining what are likely for most readers to be unfamiliar diagrammatic formats developed by circadian researchers.

A related characteristic of diagrams on which we will focus is that scientists develop techniques for diagrammatic representation over time. Sometimes in seeking to represent new phenomena, they can borrow a format that had been developed elsewhere and is already well understood. However, existing formats may not offer the best vehicle for revealing what is significant in the phenomenon or for engaging in further reasoning about it. This drives scientists to develop new representational formats. Like other innovators, scientists typically must apply multiple rounds of revision to their first attempts at novel diagrams, finally achieving a format that meets their cognitive needs. Accordingly, representational formats employed in a science are not static but are developed and changed, and such changes can in

turn alter the cognitive processes of the scientists who construct them.

In this paper we examine several diagrammatic formats that researchers have developed to represent circadian phenomena. We especially focus on how these formats were introduced and revised and show what users of the diagrams must learn in order to interpret them.

## Diagrammatic Representations of Circadian Rhythms

A very familiar way of representing rhythmic oscillation is to employ a Cartesian coordinate system in which time is presented on the abscissa and values of a variable of interest on the ordinate. Many such examples can be found in diagrams of circadian phenomena. Figure 1 shows Aschoff & Wever's (1981) plot of potassium levels in urine samples taken every four hours from six individuals. To make the circadian pattern immediately apparent, they shaded the hours of darkness and connected each individual's data points to yield six superimposed line graphs.

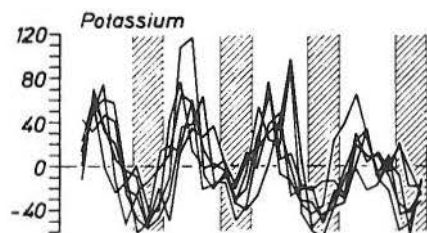


Figure 1. Circadian oscillations in potassium levels for six individuals across four days, measured every four hours (From Aschoff & Wever, 1981).

Such diagrammatic representations make manifest whether the oscillations are regular and sustained (versus damped), the duration of each cycle, and the average amplitude and extent to which it varies across cycles. These can also be compared across diagrams for different individuals (as in Figure 1), variables, or conditions. But there are limitations to the transparency of such diagrams; some types of information are much less perspicuously represented. Accordingly, circadian rhythm researchers have developed other diagrammatic formats, each making particular phenomena manifest.

## Actograms

Even when Cartesian plots incorporate some representation of a reference oscillation (e.g. the light-dark cycles in Figure 1), they are not the ideal display for tracking the variable of interest with respect to those cycles across multiple days. A far more suitable format for this purpose is an actogram, in which stacked horizontal lines represent successive days, and short vertical lines mark each time the variable exceeds some threshold. (Where the vertical lines are dense, some investigators simplify the plot by substituting a solid horizontal bar.) Visual inspection then quickly makes clear any systematic changes across days in the circadian cycles.

The technique of representing activity in actograms appears to have been developed by Johnson (1926), who was investigating the nocturnal versus diurnal behavior of various mammals. Johnson devised the use of a disk rotated by a clock on which movements of a mouse in a cage were recorded as deflections in an otherwise smooth tracing (Figure 2, left side). While one could compare multiple circular tracings to assess changes or stability over successive days, Johnson introduced the actogram as a better format for this purpose. In essence, he unrolled each circular tracing (one day's data) into a straight line and placed each line below the previous one such that the hours of all days were in alignment. In this first actogram (Figure 2, right side), three sets of lines were presented so as to compare mice from three different environments (greenhouse, from lab to woods, and woods). Within a set, the top line is Day 1, second line is Day 2, and so forth. Comparison was made more precise by running a vertical line at two-hour intervals, beginning and ending at 6 am. A shorter vertical line marked the onset of sunset, a convention not maintained by subsequent researchers. A major virtue of the actogram is that viewers can employ their visual ability to detect differences in the pattern of marks so as to compare circadian rhythms across days or conditions. It can be seen that these nocturnal animals are active primarily at night, with variations between conditions in onset time and in the extent of daytime activity but considerable stability across days within each environmental condition.

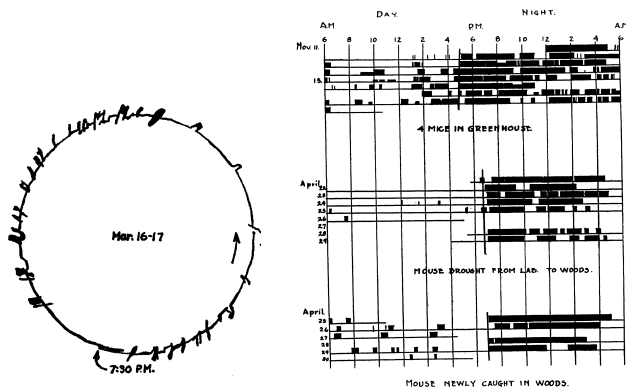


Figure 2. On the left, Johnson's (1926) tracing from a rotating disk illustrating periods of activity by a mouse. On the right, his first actogram in which activity for each day under different conditions is shown on successive lines and the time of sunset is indicated by a vertical line.

In a subsequent investigation (Figure 3) Johnson showed the effects of exposing a deer mouse to different light-dark conditions (as indicated by labels along the right edge rather than by spatially separating conditions): (1) normal light-dark cycle (light during daytime hours, dark at night); (2) constant darkness for several 24-hour periods; (3) reversed light-dark cycle (dark from 8 am to 8 pm, then light from 8 pm to 8 am); (4) again, constant darkness. One can immediately see that the periods of activity showed little change when the mouse was transferred from normal light-dark

cycles to constant darkness, but shifted dramatically when light was reintroduced in reverse: with the dark hours now in daytime, the mouse became active during day rather than night hours. This altered pattern was maintained when constant darkness was reintroduced. Thus, once entrained to a particular light-dark pattern (normal or reversed), mice kept to the same activity cycle when the entrainment stimulus (light) was removed.

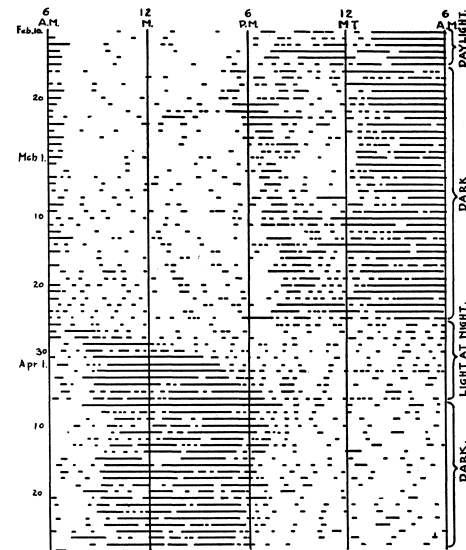


Figure 3. Johnson's (1926) second actogram, which shows a mouse's activity as light-dark conditions were changed as indicated on the right.

Since introduced by Johnson, the actogram has become a standard way of representing circadian activity, especially in animals. While the basic format has been preserved (one line per 24 hours, hours aligned vertically, active times marked along the line), many variations have been introduced to make specific features of circadian behavior more explicit. As just illustrated, chronobiologists are particularly interested in activity under constant darkness (known as *freerunning* behavior). Since this condition eliminates any effects of daily entrainment to sunlight, it can reveal the animal's normal endogenously generated rhythm or (as in Figure 3) the enduring effects of resetting that rhythm via an abnormal light-dark condition. Various conventions have been adopted for conveying lighting conditions visually, rather than by labels along one side as in Figure 3. One is background shading of the actogram across those hours the organism is in darkness (similar to the shading superimposed on the Cartesian plot in Figure 1). Another common convention is horizontal bars at the top or bottom of the actogram, in which white indicates hours of light and black indicates hours of darkness. Thus, in Figure 4 the top bar indicates the normal light-dark condition used as a baseline on days 1-7 and the bar below it indicates that constant darkness was imposed thereafter. It can be seen that once in constant darkness, the mouse's activity begins about a half hour earlier each day. From this it can be concluded that the endogenous period is about 23.5 hours.

A variety of conventions have been developed to indicate temporary changes in conditions. In Figure 4, the gray arrow indicates a day on which a light pulse was presented four hours after activity onset. This not only caused activity to mostly cease for that evening, but also inserted a phase delay the next day into what was otherwise a continuing pattern of phase advance due to constant darkness.

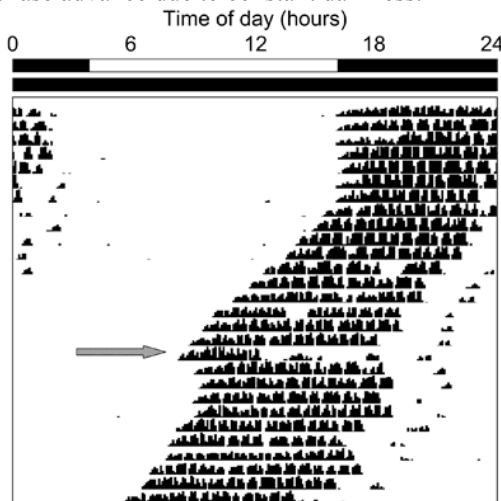


Figure 4. Contemporary actogram in which the top bars indicate a normal light-dark condition for the first seven days and constant darkness for subsequent days. The grey arrow identifies the day a light pulse was administered. (From <http://www.photosensorybiology.org/id16.html>.)

A major innovation in actograms was the introduction of double-plotting—a procedure in which data from the next 24-hour period is plotted not only on the next line but also to the right of the current data. Thus, each line shows data from 48 hours, but the left half of the actogram stacks all 24-hour periods as usual (as does the right half, redundantly). One of the first uses of this technique, by Pittendrigh (1960), well illustrates its particular advantage when activity periods extend across the 24-hour boundary. Pittendrigh was seeking to represent what he called an “after effect”: altered circadian rhythms in the activity of a nocturnal animal after exposure to continuous light. He began with normal light-dark cycles (LD). As shown in the top third of Figure 5, the mouse is inactive when the light is “ON” and becomes active when the light switches to “OFF” (with some subtleties in the data we need not discuss). When he then imposed continuous light (LL), the resultant progressive delay in activity onset indicated the mouse’s day had been stretched longer than 24 hours. After a few weeks, however, it spontaneously shifted back to a nearly 24-hour period. Since double-plotting was new, Pittendrigh marked the divide between the two 24-hour periods with a double vertical line, a convention later dropped as researchers became accustomed to double-plotting. (The convention of indicating lighting conditions with white/black horizontal bars had not yet been adopted, hence Pittendrigh’s ON/OFF markers). The virtue of double-plotting is that one can easily see the full active phase even when it crosses the 24-hour boundary.

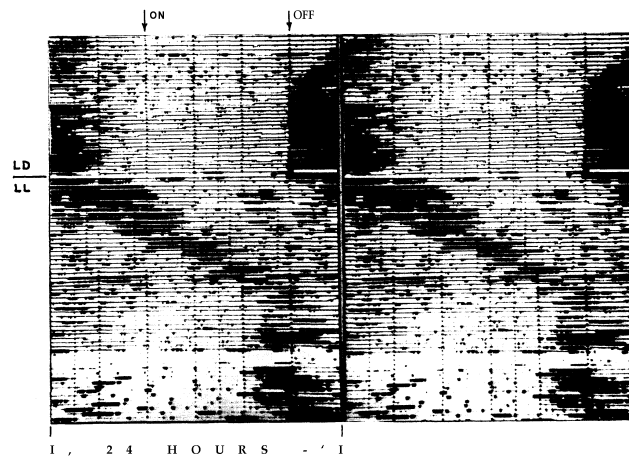


Figure 5. Double-plotted actogram from Pittendrigh (1960), in which data across the 24-hour boundary can be viewed in the middle of its redundant 48-hour display.

Because they provide an effective visual display by which researchers can immediately see variations in patterns of circadian activity, especially in relation to different lighting conditions, actograms have remained an important part of the toolkit for those circadian researchers who use animal behavior measures.

### Phase Response Curves

The circadian cycle can be thought of as a progression through phases, beginning at the time designated hour 0 (typically dawn) and ending at hour 24. The term *phase* may be used with respect to particular points on the curve (the peak and the trough being of particular interest) or for intervals (e.g., subjective day and subjective night under constant darkness). When two different cycles are closely aligned, they are said to be *in phase*. As already noted in the discussion of Figure 4 above, an actogram can show whether and by how much a rhythm is reset following a light pulse (its phase shift). As they explored this phenomenon in the 1950s, circadian researchers soon recognized that the direction and extent of the phase shift depended on the timing of light-dark cycles relative to the animal’s current circadian cycles. Individual instances of resetting could be shown in actograms, but to identify and represent the systematic pattern of resetting, researchers developed what are known as *phase response curves*.

Examining one of the first attempts to represent the effect of light on circadian phase makes evident the challenges in developing an easily interpreted diagrammatic format. Hastings and Sweeney (1958) grew *Gonyaulax polyedra*, a photosynthetic marine dinoflagellate that produces luminescence when disturbed, first under a normal light-dark cycle and then in constant darkness. Next they exposed these organisms to a three-hour pulse of light, varying the time of the pulse so as to determine how much that shifted the time of maximal luminescence. Their results are shown in Figure 6, where hour 0 is the onset of constant darkness. The time of maximal luminescence in control organisms (who were

not exposed to the pulse) is shown by the vertical lines at 7, 31, and 55 hours. The horizontal lines represent organisms exposed to the three-hour pulse at different hours of delay after the onset of darkness (3, 7, 11, . . .). Curves have been fit to data points marked by small triangles, which indicate the time of their subsequent maximum luminescence. The distance of each data point from the nearest vertical line represents the degree of advance or delay. It can be seen by following the horizontal line labeled “23” that organisms exposed to a 3-hour light pulse beginning 23 hours after onset of darkness show maximum luminescence at hour 32 rather than 30, a phase delay of 2 hours. In contrast, pulses beginning 7 hours after darkness produce a large phase advance. While this diagram does encode the crucial information, interpreting it takes considerable effort.

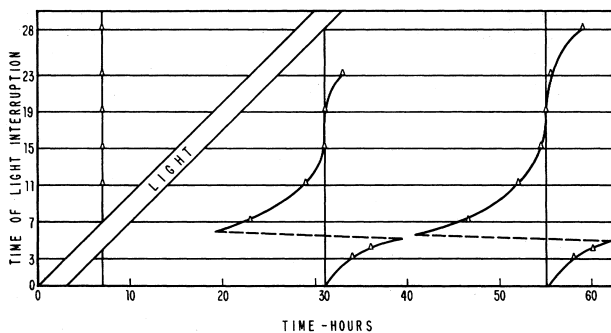


Figure 6. Hastings and Sweney's (1958) diagram showing the changes in peak luminescence of *Gonyaulax polyedra* after 3-hour light pulses. These changes are represented by the distance left or right from the vertical lines.

Shortly thereafter DeCoursey (1960) introduced a different format for representing the same information (Figure 7). In her study of flying squirrels kept in constant darkness she indicated on the abscissa the time of a ten-minute light pulse relative to the usual onset time of an animal's running-wheel activity. The data points indicate the consequent advance or delay in onset of running for two squirrels. With this representation, it is easy to see that light around the beginning of this nocturnal animal's usual activity period (corresponding to the beginning of its subjective night) delays its activity, whereas light 8 to 12 hours later (corresponding to the end of its subjective night) advances its activity. Light during its subjective day (from approximately 12 hours to 0 hours) has no effect. Having represented the phenomenon this way, one can also readily make sense of it—light exposure during subjective day does not indicate a need to reset the phase of one's activity, whereas light at the beginning of subjective night indicates either that the endogenous rhythm is out of synchrony with the external environment or that the period of daylight has expanded. The appropriate adjustment is to delay activity. Likewise, light experienced at the end of the subjective night indicates a need to stop its activity sooner.

DeCoursey's phase response curves quickly became the established means of representing the effect of a stimulus on the phase of an organism's circadian oscillation, although later researchers simplified the abscissa to circadian time (0-

24 hours) and, often, flipped the ordinate so that advances are shown as up and delays as down. Represented in this fashion, as in Figure 8, researchers were able to contrast two patterns of resetting—one producing gradual advances or delays (*Type 1*) versus an alternative (*Type 0*) in which, rather than small advances or delays, at a critical point the organism exhibits a large delay. If this delay is more than 12 hours, it can be seen as a large advance.

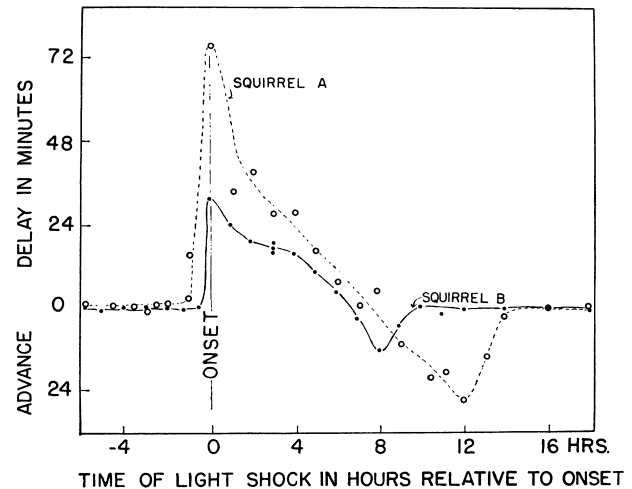


Figure 7. DeCoursey's (1960) phase response curve, which shows the advance or delay of a rat's activity onset for light pulses at different times relative to normal activity onset.

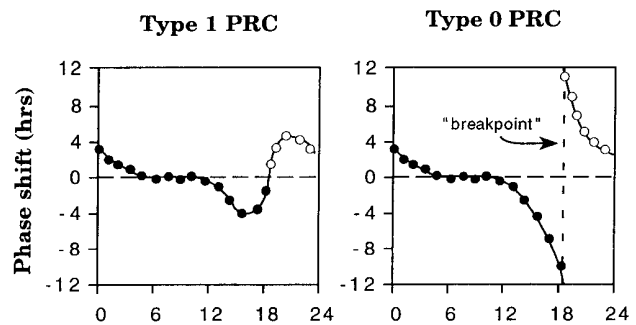


Figure 8. Type 1 and Type 0 phase response curves, with a simplified abscissa (0=dawn). From Johnson (1999).

To facilitate characterization of these two types of resetting, circadian researchers developed yet another diagram format, the phase transition curve, in which the new phase, not the amount of advance, is plotted on the ordinate. Diagonal lines indicate the situation in which the phase does not change. To see what happens in Type 0 resetting, one must plot 48 hours on the ordinate. As seen in figure 9, Type 1 resetting is characterized by a curve that stays very close to the diagonal and so approximates a slope of one (from which the name Type 1 is derived). In Type 0 resetting there is an abrupt jump from one diagonal to another, approximating a slope of 0. A virtue of the phase transition curve is that it makes clear the relation of the new phase to the old, which is not directly displayed in phase response curves.

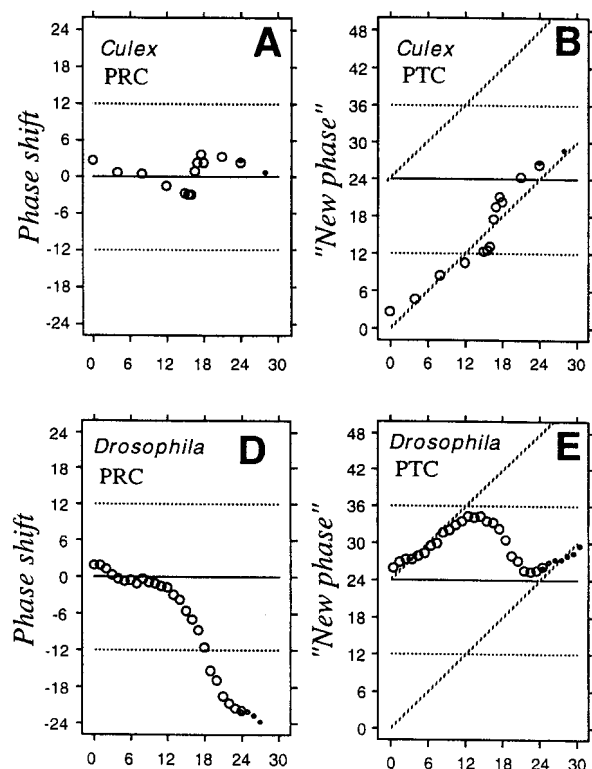


Figure 9. Resetting represented in both phase response curves (left) and phase transition curves (right). From Johnson (1999).

## Discussion

Of the representational formats used by scientists in performing and communicating their research, diagrams are particularly important but not yet extensively studied. We began our exploration of this topic by examining the use of diagrams to delineate phenomena. More specifically, we examined how actograms, phase response curves, and phase transition curves each highlight, and enable scientists to see, different circadian phenomena. Once delineated, such phenomena are explained by working out the responsible mechanism, and in later work we will examine other kinds of diagrams and their roles in mechanistic explanation.

We have drawn attention to the fact that scientists are often developing new representational formats to make manifest the specific phenomena in which they are interested. It is important to note that the researchers who devised the formats in Figures 1-9 were not seeking the best way of representing all circadian phenomena in a single diagram; rather, each sought to elucidate a specific phenomenon, such as changes in phase or the susceptibility to entrainment by light. In highlighting one phenomenon, each format either obscures or leaves out others. Actograms make clear how the phase of rhythms changes when an animal is switched to a free-running condition or exposed to a specific perturbation, but they do not show whether the rhythm is dampening or how the phase would change across the full range of possible times of perturbation. Phase response curves do this

latter job, but do not display phase changes after a switch to free-running.

New representational formats make new cognitive demands on audiences. If these formats were new to you, you also experienced the learning that is required to see what each diagrammatic format is representing. Only after learning to see the phenomenon in the diagram can the scientist use that format to efficiently reason about the phenomenon and begin the work of explaining it.

## References

- Aschoff, J., & Wever, R. A. (1981). The circadian system of man. In J. Aschoff (Ed.), *Biological rhythms. Volume 4 of Handbook of Behavioral Neurobiology* (pp. 311-331). New York: Plenum.
- Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36, 421-441.
- Bechtel, W., & Richardson, R. C. (1993/2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge, MA: MIT Press. 1993 edition published by Princeton University Press.
- Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophical Review*, 97, 303-352.
- Cheng, P. C. H. (2011). Probably good diagrams for learning: Representational epistemic recodification of probability theory. *Topics in Cognitive Science*, 3, 475-498.
- DeCoursey, P. J. (1960). Daily light sensitivity rhythm in a rodent. *Science*, 131, 33-35.
- Hastings, J. W., & Sweeney, B. M. (1958). A persistent diurnal rhythm of luminescence in *Gonyaulax polyedra*. *Biological Bulletin*, 115, 440-458.
- Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Science*, 8, 280-285.
- Hempel, C. G. (1965). Aspects of scientific explanation. In C. G. Hempel (Ed.), *Aspects of scientific explanation and other essays in the philosophy of science* (pp. 331-496). New York: Macmillan.
- Johnson, C. H. (1999). Forty years of PRCs-What have we learned? *Chronobiology International*, 16, 711-743.
- Johnson, M. S. (1926). Activity and distribution of certain wild mice in relation to biotic communities. *Journal of Mammalogy*, 7, 254-277.
- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67, 1-25.
- Nersessian, N. (2008). *Creating scientific concepts*. Cambridge, MA: MIT Press.
- Pittendrigh, C. S. (1960). Circadian rhythms and the circadian organization of living systems. *Cold Spring Harbor Symposia on Quantitative Biology*, 25, 159-184.
- van Essen, D. C., & Gallant, J. L. (1994). Neural mechanisms of form and motion processing in the primate visual system. *Neuron*, 13, 1-10.

# Route choice in individuals—semantic network navigation

Nicole M. Beckage (nbeckage@uci.edu), Mark Steyvers (mark.steyvers@uci.edu)

Department of Cognitive Sciences, Social and Behavioral Sciences Gateway Building  
Irvine, CA 92697 USA

Carter T. Butts (buttsc@uci.edu)

Institute of Mathematical and Behavioral Sciences & Department of Sociology  
Donald Bren Hall Irvine, CA 92697 USA

## Abstract

In a novel experimental task, individuals are asked to navigate from a start word to a goal word through a semantic network. In this forced-choice task, individuals perform with a high success rate (73%) and frequently navigate to the target in the minimal number of required steps (22%). We utilize these experimental results to explore different search and decision strategies. Our descriptive modeling results suggest individuals are not guessing at random (or utilizing only local information) and that knowledge of the global structure is necessary for individuals to succeed. We further show that a latent semantic space model, such as word association space, can capture much of the global semantic knowledge necessary to explain participant decisions. We suggest that performance in this task might capture some of the underlying structure of semantic memory and, importantly, search within memory.

**Keywords:** Semantic network, navigation, semantic memory, network navigation, search in memory

## Introduction

Much work within computer science and informatics has looked at humans as information foragers (Fu & Pirolli, 2007). In many analyses, foragers rely on the structure of the environment for information foraging cues. While we know humans are able to search and gather information from a variety of environments (Fu & Pirolli 2007; MacGregor et al, 1986), we ask what happens when individuals are themselves responsible for both the structure in the environment as well as for searching on and within the structured environment. We use semantic navigation as a task in order to examine this aspect of search.

Semantic navigation includes any and all orientation and search within semantic knowledge. This could be due to communication between individuals, comprehension of auditory and visual language or encoding and retrieval of vocabulary within memory. Semantic space is unique in that it has been shaped not only by individual experience but also through cultural and historical contexts. Unlike searching on the web (Fu & Pirolli, 2007) database menus (MacGregor et al, 1986) or Wikipedia (West et al, 2009), semantic search requires searching on a naturally evolved but explicitly learned representation.

This added level of implicit knowledge of the structure of the environment might allow for foragers within semantic space to utilize the environment more effectively and quickly. We know from past work, that humans are already very good at navigation even in foreign environments such as the web (Fu & Pirolli, 2007). We set out here to see if

individuals can explicitly navigate semantic space and how much individuals rely on local information, available in many types of foraging tasks, as well as global information, which is available through previous linguistic experience. We give individuals a start location in the semantic network and ask them to build chains of associates to get to a target location. This data provides us with the decisions of individuals which can in turn tell us about the underlying navigation process and semantic environment.

Results from computer science have also suggested what types of networks are most easily navigable and Kleinberg (2000) has done simulation work considering what properties networks must have for humans to successfully navigate through them. This work operates within a *message passing* paradigm, in which navigation occurs via a series of independent, uncoordinated routing decisions in which each node selects a neighbor to serve as the next decision maker in the passing chain. An important complement to this family of problems is a *route choice* paradigm, in which a single decision maker identifies the entire path to be followed.

More recently, work exploring human navigation of semantic network paths (from start to goal) has been conducted within the route choice paradigm. Specifically, in relation to this work, this has been studied in Wikipedia where individuals are given a start Wikipedia page and asked to navigate to a goal page (West et al, 2009). Whereas this work does rely on human cognition, the results of this work focus mostly on computer science implications. We hope to expand this work by exploring the cognitive implications of an individual's decisions.

To achieve this goal, we consider how humans navigate through realistic semantic network representations and, therefore, consider a novel task in which individuals are asked to navigate in a predefined semantic space. Because so little work has been done on semantic network navigation by a single individual, we set out in this paper to answer some fundamental questions. The most obvious being whether individuals *can* navigate a semantic network without being given explicit global information regarding network structure. To foreshadow our results, they can and will do so quite well in specific situations. This leads to other questions, such as what type of information might individuals be using in making routing decisions; how much local information is utilized in our specific network representation versus how much knowledge comes from global language knowledge. More generally, what can this tell us about human cognition, navigation and search?



We consider semantic space because individuals receive a great deal of linguistic input through a variety of different media. While we do not believe that every individual has the same semantic representations or knowledge, an important goal of language is communication, which facilitates the need for convergence to a similar, if not identical, representation. The fact, however, that imposing a pre-determined structure does not disallow success within this task suggests that even an impoverished representation of semantic knowledge still contains enough information for participant success.

We begin by describing the semantic network and the experimental task. Then we discuss the performance results and examine them in light of descriptive and cognitive interpretations. We consider descriptive statistics and qualitative models to help build the foundation for future modeling work. Our results importantly suggest that individuals have a specific route choice strategy, and that this strategy is greatly impacted by the similarity of an option to the end word. That is to say, individuals have some idea of distance from their current location to the goal and are often able to use this global information to navigate to the goal. With these main results, we then discuss the future for navigation models and their impact on our understanding of human cognition, navigation and search.

## Methods

### Semantic Network

Our task is rooted in the idea that individuals use both global and local information from the network. However, it is difficult to measure a semantic network for each individual, and moreover, sampling an individual's semantic network may bias the network and participant responses. To get away from these issues, we assume that individuals have similar semantic representations and that these representations can be approximated by a network based on the Florida Association Norms (Nelson et al., 2004). While it seems unlikely that each individual has precisely the same network, an important goal of language is communication with others, suggesting that convergence on the same underlying network would be highly beneficial. Such a network could be recovered through an aggregation process, such as the Florida Association Norms (2004).

The Florida Association Norms (2004) were generated by asking participants to indicate the first semantically related word that came to mind when given a cue word. Because this was asked of many participants, we have many different associations as well as a population level proportion of responses to each cue. For example the word DOG might often elicit CAT but a measurable proportion of participants may respond with BONE. We consider a directed link between words to exist from cue to response if the cue word reliably generated the response word. Each association also has a weight equal to the probability of its elicitation. For example we consider both CAT and BONE to have a link from DOG but the link to CAT receives a higher weight

since more individuals responded with CAT. This network is not symmetric. For example there exists a link from CAT to DOG but not from BONE to DOG.

Altogether 5008 words are included in the association norms with most responses being asked about as cues. However, in our experiment, we use a subset of the network. We trim the 5008-word network by including only words that had more than three words leading in as well as three leading out; we further removed the weakest connections when there were more than 12 associates. In cases where there were multiple associates with minimum strength, all were removed even when the resulting set was less than 12. We trimmed the network so that individuals would have fewer choices to sift through, were less likely to end up selecting an option that led to limited choices and to prevent trials with only a few successful paths. Limiting in-associations resulted in removing words like MOO since it is only generated in response to COW and thus all successful paths require going through COW. Further, LEFT elicits only the response RIGHT so we exclude LEFT and other similar words since it results in loops, or in the more general case, very limited options. This trimming resulted in a smaller network consisting of 2392 words. This network maintains the small world structure of the full network with a short average path length (4.19), a small overall diameter (8) and is a fully connected graph.

### Word Navigation Task

In this task, individuals were given a start word and asked to navigate to a goal word. They were presented with between 3 and 12 associations of the start word and asked to pick the option that they believed would get them closer to the goal word. The selection was then centered in the screen and the next available options were generated from this word. See Figure 1 for a screen shot from the actual experiment. Each subject repeated this process until he or she reached the goal word or made a total of 25 choices (steps). Individuals could also select an undo button, which took them back to their previous decision. This incremented the 25-step count and could be repeated until the start word was reached.

Because the options individuals received were based on the association norms, we had participants complete a quick version of the association norms task. We selected 50 words that were included in the original norming study but were excluded from our experiment based on our above network trimming. After participant completion of the association task, they received verbal instructions for the word navigation task—they were told that the choices they would be offered were generated in the same manner as the task that they just completed. This was done to aid in task understanding and minimize frustration during the task. Participants then began the computer task in which they again received written instructions, an example trial and 3 practice trials before a final set of written instructions. The example trial explained the layout of the experiment whereas the practice trials allowed them to try simplified variants of the word navigation task. On the practice trials,



participants received feedback as to how many steps it took them as well as what an optimal (fewest number of intermediate words) solution would have been. After the three practice trials, they were given a few lines reminding them of the goal of the task and the opportunity to ask the experimenter should they have any questions.

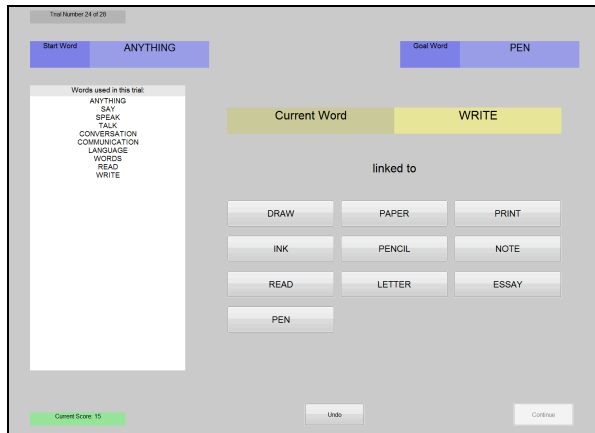


Figure 1: Screen shot of the experiment.

The test trials consisted of 28 trials divided into 4 blocks. Each block contained 7 trials, 3 requiring minimally 3 decisions, and 2 requiring each of 4 or 5 decisions. Trials were prescreened in a pilot study. Problems that were successfully completed in 15 steps or less by at least 1 out of 3 participants were selected. The block order was randomized and trials within each block were also randomized. When the 25-step limit was reached a screen popped up that said “Thank you for trying. You were # steps away,” where # was the number of words between the last word clicked and the goal. They could always see previously selected words as well as the start word, goal word and current word. At the end of the first block, participants received feedback on the overall number of steps taken in that block. In each subsequent block, they could see their current block score at the bottom left of the screen. After completion of each block, a screen reported their overall performance on the completed block as well as the minimum score on any block thus far.

Overall, 53 undergraduates at University of California, Irvine were run in an experiment that lasted maximally 1 hour. Two participants did not complete the task in the allotted time and their data were excluded before analysis. All participants received course credit for completion of the experiment. To prevent meaningless clicking, an experimenter was within earshot for the length of the experiment and participants were warned that they would not receive credit unless they completed at least one trial. Every participant satisfied this requirement.

## Results

### Experimental Results

The first important result of this work is that individuals can reliably navigate semantic networks, moving from start to goal words in a relatively small number of steps. Every trial was solved by at least 15.1% of individuals with the average trial being solved 73.3% of the time and maximal success rate at 92.9%. The information in the semantic network is sufficient for individuals to navigate effectively. Individual performance over all trials varied from 28.6% correct to 92.9%. Further 22.2% of trials were solved in the shortest number of intermediate steps.

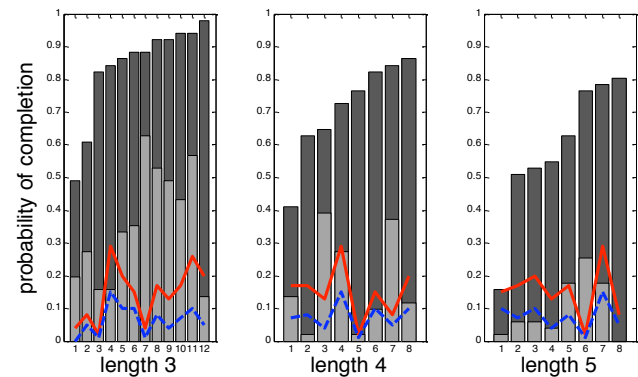


Figure 2: Subject performance on trials of varying min. length. Dark gray bars: proportion of trials correctly solved, light gray: proportion of trials solved in minimal number of steps. The solid line indicates unweighted random walk performance and the dashed, weighted random walk.

We expected to see a difference in success rate based on the number of minimum decisions required and figure 2 shows a general population level trend that trials requiring minimally 3 decisions have the highest success rates; however, the variation across trials suggests that more is going on. Figure 2 shows the results for each problem organized by minimum number of required steps. The first frame contains trials with minimally 3 decisions, the second 4 decisions, etc. The dark gray bar indicates the proportion of individuals who correctly solved that trial in 25 steps or less. The light gray bar indicates the proportion who solved that specific trial in the minimum number of decisions. The trials are rank ordered from least to most solved within each set. We see a general trend here that problems requiring fewer minimum steps are solved more often. This is an interesting finding since participants are not told the minimum length of a trial. We are also not considering the strength of the connections, the number of options or how quickly these problems were solved—instead the fact that the minimum number of steps can be used to help explain performance suggests that the information individuals are using during this task is sensitive to distance in the network. Salient information about distance seems to be present locally given the correlation between distance and

performance. Further, we find a similar relationship between optimal performance and minimum distance with shorter trials finished more optimally.

The trend suggesting that problems with fewer minimum steps are easier does not capture the full complexity of the task or responses chosen. With a closer examination of the results in figure 2, we see that there are many trials that violate this trend. For example, there are a few trials of medium length that are more often solved than shorter trials (and some that are more often optimally solved). It is also interesting to note that the trial that is solved most often has one of the lowest rates of “optimal” performance. This may suggest that our definition of optimal is not the correct baseline for human performance.

We are also interested in capturing the descriptive trends within individual trials. To understand what these trends might look like, we explore one problem in depth. Figure 3 shows the breakdown of a single trial. Only correct responses are included and the weight of the arrow indicates the proportion of individuals who chose that path. This trial has a high success rate and a high percentage of minimum distance paths. The minimum distance path runs along the left. This figure helps illuminate the cognitive process that may be underlying the strategy.

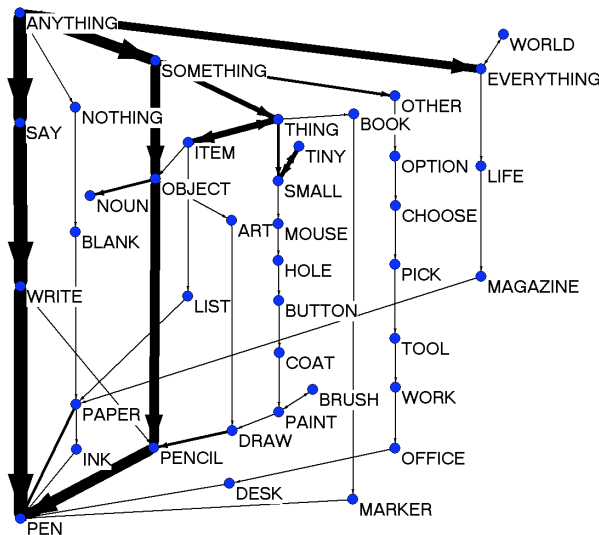


Figure 3: Network based on successful trials; participant responses from start word ANYTHING to goal word PEN.

For instance we can see that each option of ANYTHING led to at least one correct response—that is to say, individuals do not remove their chance of success by picking the wrong option at the first decision. Another important feature is that there are multiple successful paths. Looking closely we can see that individuals are utilizing the undo button to further explore the semantic space. For example an individual at the word PAINT selected BRUSH, but then decided to go back to PAINT. While it is difficult to say exactly why s/he made that choice, it would seem plausible that s/he went back to PAINT because s/he felt

that it was closer to the goal word of PEN than BRUSH was. Another interesting result that we can see from a detailed examination is that the word definition might change based on the goal word. This is most clear in the path that goes from ANYTHING→EVERYTHING→WORLD→EVERYTHING(undo)→LIFE→MAGAZINE→PAPER→PEN. Here we see that the undo button was used to back up after making a decision to go to WORLD. Further, this individual selected the word LIFE from a list of words associated to EVERYTHING. This suggests that the definition coming to mind was one of living, however, with the goal of PEN, s/he utilized LIFE to get to MAGAZINE, suggesting an interpretation of life magazine. We can also see that s/he does a similar thing in going from MAGAZINE to PAPER (likely newspaper) but then from PAPER (something to write on) to PEN.

### Random Walk model

Though the data suggest that individuals are able to solve these semantic navigation problems (to varying degrees) it is possible that participants are guessing and that the structure of the network allows for high rates of success. To test this assumption we considered two types of random walk models. The first is a random walk model that simply checks if the goal word is present and if it isn't, randomly selects from the available options. The second random walk model picks the goal word if present and otherwise randomly selects from the available options with a probability distribution equal to the association norms data (e.g. if CAT was the response to DOG 80% of the time, this random walk would pick CAT 80% of the time as well). In figure 2, the two random walk models are indicated by a solid red (unweighted) and dashed blue (weighted) line. Both random walk models perform worse than our participants. The general trend does not follow that of the participants—problems frequently solved by random guessing are not those that participants most often solved. This confirms the hypothesis that individuals are utilizing global information present in semantic space in the task.

### Descriptive Geodesic model

Now that we know individuals are not guessing at random, we combine the results suggested by the data to build a descriptive model of the decisions individuals make. To do this we consider the geodesic distance (number of steps between two nodes) of the current word to goal word to see if individuals are more likely to select words with lower geodesic. We know that individuals do not always pick options that decrease the geodesic because that would result in optimal performance. However, we can plot the distribution of subject choices and the distribution of all options to see if some of the success can be explained by sensitivity to geodesic distance. Figure 4, top graph, shows the distribution of geodesic from current word (WRITE in figure 1) to goal word (PEN in figure 1) along the x-axis. Options (grey buttons in figure 1) to goal word fall along the y-axis. Here we can see that proportion of choices made

by individuals, as indicated by the size of the box, look different than the full set of options. This is particularly pronounced at low geodesics (heavy weight along the near-diagonal indicates more optimal decisions). The difference between subject choices and options becomes almost unrecognizable as the distance between current and end word increases to a geodesic near 4 or greater. This suggests that individuals have knowledge of the general location of the goal word and that this becomes more accurate as they get closer to in minimal number of steps to the goal. It also suggests that individuals may be picking up on a gradient but that they might be guessing until they get close enough to the goal word to find the gradient.

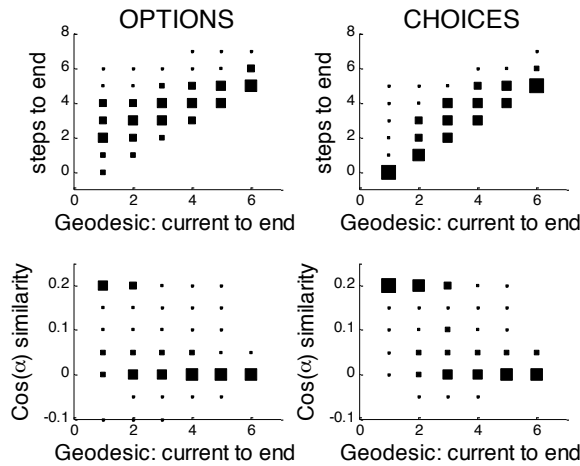


Figure 4: The top panels, based on geodesic distance, represent the distribution of options (left) and participant choices (right) with the size of the square indicating frequency of observation. The bottom panels capture local language-level information based on cosine similarity in latent space.

### Latent Space model

While we consistently have been talking about navigation within a network, it is not necessary to assume individuals represent a complete network in memory. Instead it is likely that the representation is a reduction of this network—a summary of global information that allows individuals to locate their current position in the network as well as access local knowledge of nearby words. One reduction that has been widely studied in semantic memory is that of a reduction of the dimensionality of the semantic space by vector decomposition (latent space analysis, e.g. Landauer & Dumais, 1997; Burgess & Lund, 2000). While we considered a variety of latent spaces, we present data only on the symmetric WAS space from Steyvers et al. (2004). Our goal is to capture the same population level trends as we did previously with the descriptive geodesic model but with a semantic latent space that would be more cognitively plausible and require less information than the full network. To do this, we again consider figure 4 (lower panels) and

the geodesic between current and goal (x axis) but compare that to the latent space cosine similarity of options (panel 1, lower row) and subject choices (panel 2). Whereas, in the geodesic case, weight along the diagonal indicates more optimal choices, high cosine similarity suggests nearness to the target. In this graph, we expect higher weight on similarity judgments to capture more optimal choices. We again see a noticeable difference between selected choices in comparison to all options, moreover, the general trends of the latent space model follow a similar pattern to that of the geodesic in that individuals’ choices are indistinguishable from guessing at higher geodesic. The latent space model offers an explanation—the cosine similarity is near zero for all options such that participants may not be able to use similarity and instead resort to guessing

### Discussion and Implications

Our results suggest that individuals are succeeding by utilizing information present in the network in order to get closer to the goal word. We know that their decisions can be explained at least in part by the local information in the options, especially relative to the goal word. Their success is not, however, based on random guessing or strong associates. This is an interesting finding since it suggests that the information individuals are using is not captured by the environment of the free association task alone. Further, the paths individuals do end up utilizing appear to suggest that the semantic space may be changed and altered by the goal word implying that individuals have a direct influence on their environment. That is to say, the entire structure of the network may be influenced and changed based on the goal. Though we only gave one example in the text, it is not unique. Individuals often interpret words in light of the goal word as opposed to the current word. This adds a dynamic component to network structure that we know exists in memory and knowledge more generally. This task, further gives us a way to study the dynamic nature of semantic knowledge and the role of context in speech.

We also see that there is a large variance in potentially successful paths and that the shortest path is not always the most salient to individuals. While we have not specifically analyzed the difference between shortest paths and participant paths, West et al. (2009, 2012) have looked more closely into this question in Wikipedia navigation and suggest that shortest paths often require out-of-the-box thinking whereas paths that are a bit longer allow for a more obvious chain of associates. We hope to test this directly utilizing our data in the future.

Another important finding is that individuals seem to be making more optimal decisions (ones that get them closer to the end word) when they are already close to the end word. This suggests that individuals can intuit how far away the goal word is without having exact knowledge of the space. This is a result that has been found in most navigation studies (e.g. West et al. 2009) but most studies suggest that the way individuals get closer to the goal is by navigating to hubs with many out-links and utilizing these hubs to get to

an area of the network. However, in our study, we thresholded much of the hub structure away by allowing maximally 12 options for any word. Participants could not simply navigate to a central hub and then jump towards the goal word. Instead, we believe that this ability to perform more optimally when closer comes from the fact that individuals have a semantic representation that allows them to compute distance between two words but that this semantic representation is limited to identification of a relative location. Further, the ability to identify a word as near requires a level of information about current location and goal location that is not always available, specifically when individuals are further away from the goal. Going back to Figure 4, we see that individual choices look very similar to options both in geodesic and WAS space when many words are needed to complete the trial. This suggests that, if individuals are far enough away, guessing might be their main strategy. However, guessing may be the best thing for individuals to do since there is little information available to them (as captured by WAS) and, based on network structure alone, often places them in a better or equal position (in terms of geodesic) than before.

WAS space captures most of these global trends. Particularly, WAS space is often near zero unless there is a strong similarity between words—implying that they are close enough in the network that individuals can sense it. This space also captures the noisiness of relative distance. Since individuals only have a very general idea of goal location, any estimates of which choice is closer to the goal is less exact as the distance between choice and goal increases—which is captured by cosine similarity.

In the future we hope to extend our understanding of network navigation through a relational event model (Butts, 2008). That is to say, we can assume every decision is independent once we condition on the goal word and current word. With independence, we can apply a multinomial logistic regression on linguistic and network-based covariates. We hope to use this model to show how the current and goal word influence decisions as well as more specifically exploring the specifics of language in this task.

We also hope to experimentally test subjects on other types of networks. Since all subjects in this study are natural "experts" in language, there is still the question of whether a more limited level of prior knowledge of the underlying network is still adequate to allow successful navigation and search. Work on folk knowledge of networks suggests that individuals are not very good at reconstructing social networks (Freeman et al., 1987) but our results suggest that success on this task may not require an accurate or even complete network representation, since most individuals succeeded on a variety of problems even though our underlying network of the task is impoverished.

Not only do the applications extend beyond cognitive understanding, but the fact that individuals can navigate suggests that network representations are useful. While we consider semantic space here, many other types of knowledge can be represented as a network, such as social

relationships or a schedule. We believe that the results in this paper speak much more broadly about navigation than they do about language navigation specifically. A model of network navigation may be useful in explaining search, decision-making and even memory. Network structure captures many naturalistic relationships. However, unless we understand ways in which individuals are able to navigate this type of structure, we cannot utilize this representation in cognitive architectures. With this paper, we've begun to address the first concerns of understanding how individuals navigate a network structure, providing us with a new direction for navigation within memory.

## Acknowledgments

This work was funded in part by an NSF GRFP to the first author and by ONR/MURI under grant number N0014-08-1-1015 to Carter Butts. Thanks to Michael Yi for help in concept development and Sarah Hunt for running subjects.

## References

- Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory, *Cognitive dynamics: conceptual and representational change in humans and machines*.
- Butts, C. (2008) A relational event framework for social action. *Social Methodology* 38,155-200.
- Freeman, L.C., Romney, A.K., Freeman, S.C. (1987) Cognitive structure and informant accuracy. *American Anthropologist*.
- Fu, W.T. & Pirolli, P. (2007). SNIF-ACT: A cognitive model of user navigation on the world wide web. *HCI*
- Kleinberg, J. M. (2000) Navigation in a small world. *Nature*, 406(6798).
- Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory. *Psychological Review*, 104, 211-240.
- MacGregor J, Lee, E. & Lam, N (1986). Optimizing the Structure of Database Menu Indexes. *Human Factors*
- Newman, M.E.J. (2003). The structure and function of complex networks. *SIAM Review* 45, 167–256
- Nelson, D.L., McEvoy, C.L., & Schreiber, T.A., (2004). The University of South Florida word association, *Behavioral research methods*.
- Steyvers, M., Shiffrin, R.M., & Nelson, D.L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. *Experimental cognitive psychology and its applications*
- Watts, D.J. & Strogatz, S.H. (1998) Collective dynamics of "small-world" networks, *Nature*, 393, 440–442.
- West R., Pineau, J. & Precup, D. (2009) Wikispeedia: An online game for inferring semantic distances between concepts. *IJCAI '09*

# Competent Deontic Reasoning: The Abstract Deontic Selection Task Revisited

Sieghard Beller (beller@psychologie.uni-freiburg.de)

Andrea Bender (bender@psychologie.uni-freiburg.de)

Center for Interdisciplinary Research (ZiF), Bielefeld University

D-33615 Bielefeld, Germany

Department of Psychology, University of Freiburg

D-79085 Freiburg, Germany

## Abstract

The abstract deontic selection task was developed with the aim of demonstrating abstract rule use in a specific domain (i.e., deontic rules). Yet, the solution rate, while being substantially higher than with abstract *non-deontic* tasks, did not reach the level obtained with *concrete* deontic tasks. What are the reasons for this—difficulties with abstract rule use? A task analysis based on deontic principles uncovers several problems with the formulation of the task. Three experiments replicate the difficulties with the original task and show that performance increases, when the problems are resolved. The results provide novel insights into the interpretation of deontic rules and into the role that such content-specific, but abstract tasks can play for the study of reasoning processes.

**Keywords:** Deontic reasoning; social rules; deontic selection task; dual process theory; pragmatic reasoning schemas.

## Introduction

Abstract reasoning is typically considered as more difficult and error prone than contextualized reasoning. A widely accepted and often cited case in point is *Wason's Selection Task*: The classical abstract task demands testing whether an arbitrary conditional statement is true or false (Wason, 1966) and is solved correctly only by few participants (10-20%). In contrast, structurally similar but contextualized versions that demand testing whether or not a concrete deontic conditional is being followed (e.g., Cox & Griggs, 1982) are solved correctly by the majority of participants (about 75%; Evans, 2003, p. 456; cf. Beller, 2010).

According to the *dual-process account* (Evans, 2008), abstract reasoning is characterized as domain-general, rule-based, sequential, controlled, and slow, and is limited by working memory capacity ("System 2 reasoning"); contextualized reasoning, on the other hand, is characterized as domain-specific, associative, parallel, automatic, and rapid, and is independent of working memory ("System 1 reasoning"). However, this distinction is often not clear cut: On the one hand, there are many examples for people making use of heuristic, associative System 1 processes when thinking about abstract, de-contextualized tasks; in the classical abstract selection task, for instance, they tend to apply the matching heuristic (Evans, 2003). On the other hand, people can also engage in analytic, rule-based System 2 processes when thinking about concrete, contextualized tasks as argued in Beller and Spada (2003).

The abstract deontic selection task is an interesting hybrid: It is domain-specific and contextualized, as it refers to a domain and a contextual framing we all are familiar with (to search for violators of a deontic regulation). But, instead of presenting a concrete, familiar rule like the *Drinking Age* rule "If a person is drinking beer, then that person must be over 16 years of age" (Cox & Griggs, 1982), which may activate an instance-based mode of reasoning, the task is formulated in an abstract way. Finding people's reasoning performance to be as accurate as with concrete, familiar material would thus provide evidence for abstract rule use in a specific domain (Smith, Langston & Nisbett, 1992).

The original version, introduced as "permission problem", reads as follows (Cheng & Holyoak, 1985, p. 403):

*Suppose you are an authority checking whether or not people are obeying certain regulations. The regulations all have the general form, "If one is to take action A, then one must first satisfy precondition P." In other words, in order to be permitted to do "A", one must first have fulfilled prerequisite "P". The cards below contain information on four people: One side of the card indicates whether or not a person has taken action "A", the other indicates whether or not the same individual has fulfilled precondition "P". In order to check that a certain regulation is being followed, which of the cards below would you turn over? Turn over only those that you need to check to be sure.*

Card (1) "Has taken action A"

Card (2) "Has not taken action A"

Card (3) "Has fulfilled precondition P"

Card (4) "Has not fulfilled precondition P"

Which cards does one need to check? This question is not too difficult to answer: The regulation is being followed if a person, who takes action "A", has fulfilled precondition "P". Therefore two cards need to be checked: Card (1) "Has taken action A" in order to find out whether this person has fulfilled precondition "P", and Card (4) "Has not fulfilled precondition P" in order to exclude that this person has taken action "A", as otherwise the regulation would be violated.

The task is about an abstract regulation. This should pose no problems as, according to *Pragmatic Reasoning Schema Theory* (Cheng & Holyoak, 1985), people possess abstract schemas for handling such rules. People should thus be able to solve the abstract deontic task as easily as they solve concrete ones. However, this is not the case as indicated by the meta-analysis reported in Beller (2010, p. 127): Instead, the

solution rate drops from 73.8% in concrete deontic selection tasks ( $N = 1,010$ ; 26 Experiments) to 58.4% in abstract deontic tasks ( $N = 320$ ; 10 Experiments), indicating—at first glance—some difficulties with abstract rule use in the sense of Smith, Langston and Nisbett (1992).

According to *Social Contract Theory* (Cosmides, 1989), this result comes not as a surprise because the abstract deontic task lacks a clear cost-benefit structure necessary to identify rule violators: persons who take the benefit (e.g., the beer) without “paying the costs” (i.e., fulfilling the age requirement). But if cost-benefit information were indeed necessary for the solution, should we then not expect a much stronger decrease in the solution rate than the observed 15%? And if, on the other hand, missing cost-benefit information is not the reason, what else then could be responsible for the reduced solution rate? In order to answer these questions, the “permission problem” needs to be analyzed in more detail.

## A Task Analysis

### How to Solve the Task

As suggested by Beller (2008), the first step in solving a deontic task is to identify which action constraint is imposed by a deontic rule (*constraint principle*). In the “permission problem”, the restriction concerns action “A”, and the condition is “P”. According to the description, condition “P” is necessary for the permission to do “A”. Consequently, if “P” is not fulfilled then action “A” is forbidden. Assuming that “P” is the only relevant condition (*exhaustivity*), the deontic constraint can then be represented according to the *equivalence principle* by the ban “If condition P is not fulfilled then action A is forbidden; otherwise it is allowed”:

Ban (B):  $\neg \text{condition\_P} \leftrightarrow \text{forbidden}(\text{action\_A})$

After the deontic action constraint has been identified, the second step consists of drawing the appropriate deontic inferences. From a deontic norm like ban (B), three types of inferences are possible (Beller, 2008): *Forward inferences* from the condition side of the norm to the deontic status of the regulated action (i.e., whether it is forbidden or allowed, obligatory or not), *backward deontic-to-factual inferences* from the deontic status of the regulated action to the condition (i.e., whether or not the condition is fulfilled), and *backward factual-to-deontic inferences* from information on whether or not the regulated action is taken to the deontic status of the condition (e.g., whether or not it is deontically necessary).

In order to come up with the correct solution to the selection task (“A & not-P”), one has to check *each* of the persons and to infer whether he or she might have *violated* the rule (i.e., has performed the banned action “A” without fulfilling condition “P”). The inference process might proceed as follows: For card (1) “Has taken action A”, one can conclude by a backward factual-to-deontic inference that this person must fulfil condition “P”; otherwise the rule is violated. For card (2) “Has not taken action A”, one can conclude by corresponding backward inferences that this person need not, but may fulfil condition “P”. Therefore, the rule cannot be

violated. For card (3) “Has fulfilled condition P”, one can conclude by forward inferences that this person may, but need not take action “A”; again, the rule cannot be violated. Finally, for card (4) “Has not fulfilled condition P”, one can infer by a corresponding forward inference that this person may not take action “A”; otherwise the rule is violated.

### Problems in the Original Task

Unfortunately, Cheng and Holyoak's (1985) original task is formulated in a way that renders it difficult to interpret the deontic regulation and to understand the task requirement correctly, with direct effects on how the task is solved.

With regard to the interpretation of the deontic regulation, several formulations are problematic: First, by stating that “The regulations *all* have the general form ...” (italics added), the “permission problem” suggests that not a single rule has to be checked, but a set of *possibly different* rules. Second, the formulation “If one is to take action A ...” can be understood as referring to an intention (“if someone *wants* to take action A ...”). However, it is not the *intention* that is deontically constrained, but *doing* the action. Third, the condition is described as a composition of “P” and an additional temporal constraint: “P” has to be fulfilled *prior* to action “A”. This might initiate some temporal reasoning, which cannot be resolved unequivocally, as the necessary temporal information is missing on the cards. Fourth, the precondition has to be fulfilled prior to the action, but the regulation mentions the two elements in the reverse order. And finally, the regulation qualifies condition “P” as *necessary* with the modal *must*, but does not specify whether “P” is *sufficient* for the permission to take action “A”.

With regard to the task requirement, two formulations are problematic: The first is the instruction “to check that a *certain* regulation is being *followed*” (italics added). It emphasizes that one particular regulation has to be checked. At the very beginning, the task speaks of regulations in the plural form, which may cause uncertainty about which specific regulation out of this set is meant. More severely, this instruction emphasizes rule *following*, whereas concrete selection tasks like the *Drinking Age Problem* focus on rule *violation*. This may appear to be a subtle distinction, but in fact it is one that makes an important difference: Whether a person does follow the regulation can be checked by turning over the “A”-card (1), but not by turning over the “not-P”-card (4), as rule following is relevant only for persons to which the rule *applies*: Person (1), who has taken the critical action “A”, is subject to the rule. If it turns out that this person has fulfilled prerequisite “P”, it is clear that he or she has *followed* the rule; otherwise he or she has *violated* it. For person (4), who has not fulfilled prerequisite “P”, the case is different: If this person has taken action “A”, it is clear that this person is subject to the rule and has *violated* it, but if this person has *not* taken action “A”, the rule is simply *not applicable*, and hence it is not possible to detect rule following. People, therefore, may neglect the “not-P”-card and select only the “A”-card. The second problematic formulation is the instruction “Turn over only those [cards] that you need to

check to be sure”. This might cause people to be *cautious* not to select too many cards. Some might think that finding one rule follower (or rule violator) would be sufficient.

In summary, the formulation of the task makes it difficult to extract the relevant deontic information, and the instruction induces a general preference for choosing only few cards (caution) as well as a specific preference for the “A”-card alone (to check rule following). In combination, this leads to a reduction of selection task performance and, accordingly, to an underestimation of people’s deontic competencies. By eliminating the problematic formulations, performance should increase. In the following, three experiments test this hypothesis. The results will be discussed with regard to the role this kind of content-specific, but abstract tasks can play in gaining new insights into content-specific reasoning.

### Three Experiments

In each of the experiments (taken from Beller, in press, and described here together for reasons of space), the original task from Cheng and Holyoak is compared to four new ones. By varying the formulation of the deontic rule (strong vs. weak; backward vs. forward), the question of *rule interpretation* is addressed; the problems of the *instruction* are addressed by focusing on rule violation instead of rule following.

The new tasks were constructed according to the schema shown below. Clues about intentions and temporal conditions are avoided; the instruction emphasizes that each person should be checked for rule violation; and explicit negatives are used to clarify when action “A” is *not* taken and when condition “P” is *not* fulfilled:

*Imagine you are a member of an authority that checks whether people conform to or violate a particular rule. The rule is: {one of the new rules from below}. The “cards” presented below represent four persons. On one side of each card is written whether or not the respective person takes action “A”, on the other side is written whether or not he or she fulfills condition “P”. Your task: Indicate all cards that you have to turn over—all of which you need to know the information on the back—in order to find out whether the respective person violates the rule.*

*Person (1) “Takes action A”*

*Person (2) “Does not take action A”*

*Person (3) “Fulfills condition P”*

*Person (4) “Does not fulfill condition P”*

Each task refers to a single deontic rule, which is given without further clarifications. In Experiment 1, the following four rules were used:

*Obligation O1: “If a person takes action A, then he or she must fulfill condition P.”*

*Release R1: “If a person does not take action A, then he or she need not fulfill condition P.”*

*Ban B1: “If a person does not fulfill condition P, then he or she must not take action A.”*

*Permission P1: “If a person fulfills condition P, then he or she may take action A.”*

All four rules can logically be derived from ban (B), with obligation O1 being analogous to the original rule. Two rules use a strong deontic modal and describe a deontic constraint explicitly (O1 and B1), the two other rules use a weak deontic modal (R1 and P1). Taken literally, these latter rules cannot be violated in the deontic sense, as they do not express a deontic constraint explicitly.

For Experiment 2, these rules were extended by one sentence that make their complementary side explicit in order to facilitate to derive ban (B) from the weak rules:

*Obligation O2: “If a person takes action A, then he or she must fulfill condition P; otherwise he or she need not fulfill it.”*

*Release R2: “If a person does not take action A, then he or she need not fulfill condition P; otherwise he or she must fulfill it.”*

*Ban B2: “If a person does not fulfill condition P, then he or she must not take action A; otherwise he or she may take it.”*

*Permission P2: “If a person fulfills condition P, then he or she may take action A; otherwise he or she must not take it.”*

In Experiment 3, it was manipulated, how the deontic modalities are expressed in the rules: not by deontic modals (e.g., *must not* or *may*) as in Experiment 1 and 2, but by semantically equivalent deontic verbs (e.g. *to forbid* or *permit*). With the verbs, explicit negations can be avoided, thereby eliminating another potential difficulty:

*Obligation O3: “If a person takes action A, then he or she is obliged to fulfill condition P; otherwise he or she is released from fulfilling condition P.”*

*Release R3: “If a person does not take action A, then he or she is released from fulfilling condition P; otherwise he or she is obliged to fulfill condition P.”*

*Ban B3: “If a person does not fulfill condition P, then action A is forbidden; otherwise action A is permitted.”*

*Permission P3: “If a person fulfills condition P, then action A is permitted; otherwise action A is forbidden.”*

For the original task, it is expected that people tend to avoid selecting too many cards (*caution hypothesis*) and tend to neglect the “not-P” card and to select the “A”-card alone (*rule-following hypothesis*) due to the various problems in the formulation of this task. For the new tasks, it is expected that people infer ban (B) more easily due to the clearer formulation. With ban (B) in mind, they should select the cards “A” and “not-P” when asked to check individuals for rule violation, independent of how the rule was formulated (*rule-violation hypothesis*). If asked for the literal meaning of the rules, however, people should differentiate between strong rules that can be violated by a person and weak rules that cannot be violated (*rule-evaluation hypothesis*).



## Method

**Materials.** In each experiment, five deontic selection tasks were used: the original task and four new tasks. The new tasks were constructed according to the general schema shown above and differed only in the way the rule was formulated. In addition to the selection tasks, each experiment was supplemented with a second task. In Experiment 1, a rule *evaluation task* was used in order to check which of the rules O1, R1, B1, and P1 participants consider as violable in a strict deontic sense. The instruction asked participants to *consider for each statement whether it expresses a rule that can in fact be violated by a person*. In Experiment 2, this instruction was used to compare the weak rules from Experiment 1 with the corresponding explicit rules from Experiment 2 (i.e. R1, P1, R2, and P2). In Experiment 3, participants were asked to rank the rules O3, R3, B3, and P3 with respect to comprehensibility (from 1 = best to 4 = worst).

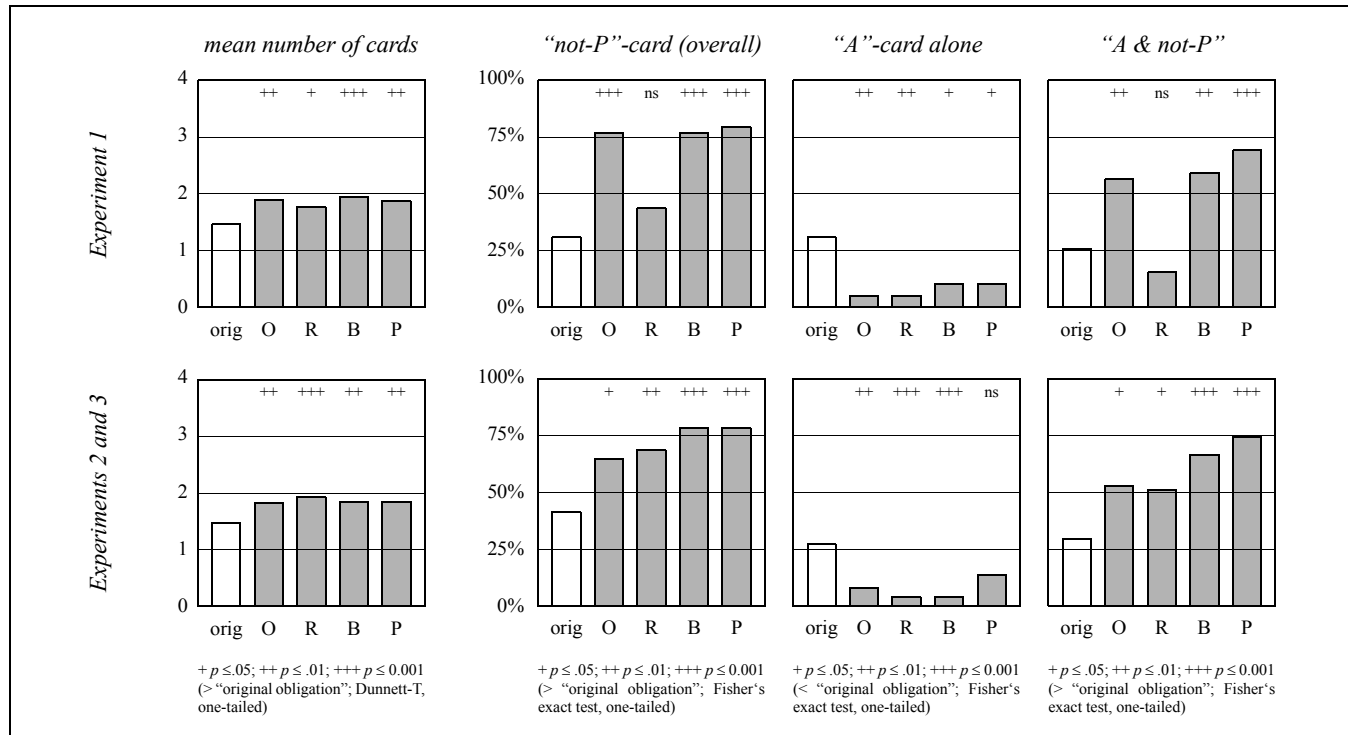
**Participants.** All participants were from the University of Freiburg and had no prior experience with the selection tasks. In Experiment 1, a total of 195 students (39 per condition) volunteered to participate for research credit (72 male, 123 female; mean age 23.2 years,  $SD = 3.57$ ). In Experiment 2, a total of 125 students (25 per condition) volunteered to participate for research credit (36 male, 89 female; mean age 23.1 years,  $SD = 5.27$ ). In Experiment 3, a total of 130 students (26 per condition) volunteered to participate as part of a study on a different subject for which they were paid (37 male, 91 female; mean age 23.5 years,  $SD = 3.94$ ).

**Design and Procedure.** These were the same for all three experiments: Each participant received a booklet with general instructions, one selection task, and the additional task (on separate pages). In both tasks, two orders of answer options were permuted equally frequently across conditions. The different versions of the booklet (with the different selection task versions varying between subjects) were randomly assigned to participants. They were investigated in small groups, were instructed to work on the tasks in the given order, and were granted as much time as they needed.

## Results

In the original selection task, people were expected to be cautious not to select too many cards, which should decrease the mean number of selected cards. With the focus on rule following, people should tend to neglect the “not-P” card and to select the “A”-card alone instead of the cards “A & not-P”. In the new tasks, performance should improve with respect to all these indicators; and exactly this was found.

**Overall analyses of the selection task data.** In Experiment 1, the five experimental groups differed significantly in the mean number of chosen cards ( $F(4, 190) = 3.99$ ;  $p = .004$ ) as indicated by an analysis of variance, and also in the overall frequency of the “not-P”-card ( $L^2 = 34.4$ ;  $N = 195$ ;  $df = 4$ ;  $p < .001$ ), in the frequency of the “A”-card alone ( $L^2 = 14.2$ ;  $N = 195$ ;  $df = 4$ ;  $p = .007$ ), and in the frequency of the combination “A & not-P” ( $L^2 = 36.2$ ;  $N = 195$ ;  $df = 4$ ;  $p < .001$ ) as indicated by the likelihood ratios. Almost all effects were in the predicted directions (cf. Figure 1).



**Figure 1:** Four deontic indices for the original selection task (white bars) and the four new tasks Obligation, Release, Ban, and Permission (grey bars) in Experiment 1 (top row;  $N = 195$ ) and aggregated across Experiments 2 and 3 (bottom row;  $N = 255$ ).



As explained above, the selection tasks used in Experiment 2 and 3 differed only in one aspect: in how the deontic modality was formulated. They were thus analyzed jointly by means of an analysis of variance and log-linear analyses (Kennedy, 1992). As none of these analyses indicated a main effect "Experiment" nor an interaction "Experiment"  $\times$  "Condition" (for all  $F$ s:  $p > .544$ ; for all  $L^2$ :  $p > .553$ ), aggregating the data over the two experiments seemed justified. This also means that it made no difference for the solution of the tasks whether the deontic modality in the rule was expressed by modals with explicit negation (Experiment 2) or by verbs without explicit negation (Experiment 3). As in Experiment 1, the five experimental groups differed in the mean number of chosen cards ( $F(4, 245) = 4.75$ ;  $p = .001$ ), in the overall frequency of the "not-P"-card ( $L^2 = 20.9$ ;  $N = 255$ ;  $df = 4$ ;  $p < .001$ ), in the frequency of the "A"-card alone ( $L^2 = 18.1$ ;  $N = 255$ ;  $df = 4$ ;  $p = .001$ ), and in the frequency of the combination "A & not-P" ( $L^2 = 25.2$ ;  $N = 255$ ;  $df = 4$ ;  $p < .001$ ). Again, almost all effects were in the predicted directions (cf. Figure 1).

**Original Selection Task.** As predicted by the *caution hypothesis*, participants selected too few cards, that is, less than 2.0, the value expected according to the correct "A & not-P" solution (Experiment 1: 1.462;  $t(38) = -6.06$ ;  $p < .001$ ; Experiments 2&3: 1.471;  $t(50) = -5.17$ ;  $p < .001$ ). As predicted by the *rule-following hypothesis*, the main problem was to recognize the relevance of the "not-P"-card (Experiment 1: 30.8% "not-P"; Experiments 2&3: 41.2%). The proportion of the "A"-card alone was rather high (Experiment 1: 30.8%; Experiments 2&3: 27.5%), and the frequency of the correct deontic solution "A & not-P" was low (Experiment 1: 25.6%; Experiments 2&3: 29.4%).

**New Selection Tasks.** The performance in the new tasks differed significantly from the original task in all aspects: The mean number of chosen cards was higher in all four conditions (Experiment 1: 1.872 on average; Experiments 2&3: 1.858). The relevance of the "not-P"-card was clearly recognized (Experiment 1: in three out of four conditions; Experiments 2&3: in all four conditions), the proportion of the "A"-card alone was reduced (Experiment 1: in all four conditions; Experiments 2&3: in three out of four conditions) and, complementarily, the correct combination "A & not-P" increased (Experiment 1: in three conditions; Experiments 2&3: in all four conditions). This pattern indicates that, in most conditions, the majority of participants inferred ban (B) and identified rule violators more often than in the original task, as predicted by the *rule-violation hypothesis*.

**Additional Tasks.** The additional task in Experiment 1 compared the weak rules permission P1 and release R1 with the two strong rules ban B1 and obligation O1. According to the *rule-evaluation hypothesis*, participants should indicate that weak rules cannot be violated in the deontic sense, only strong rules can. To test this hypothesis, a strength-index was calculated by adding 1 for each strong rule that a person marked as violable, subtracting 1 for each weak rule, and dividing the result by 2. This renders a maximum score of 1 if the two strong rules were marked only, a minimum score of -1 if the two weak rules were marked only, and a score of

0 if strong and weak rules were marked in a balanced way. An analysis of variance indicated that the strength-indices of the five selection task conditions did not differ from one another ( $F(4, 190) = .491$ ;  $p = .742$ ). The overall index was positive (.667) and different from zero ( $t(194) = 22.3$ ;  $p < .001$ ). The data clearly support the *rule-evaluation hypothesis*: The strong rules were regarded as violable by most participants (obligation: 80.0%; ban: 84.6%), the weak rules were not (release: 5.6%; permission: 25.6%;  $N = 195$ ).

The additional task in Experiment 2 compared the two weak rules from Experiment 1 (P1 and R1) that left the underlying deontic constraint implicit with the two explicit rules from Experiment 2 (P2 and R2) that expressed this constraint. A strength-index was calculated analogously to Experiment 1. If participants considered the explicit rules as the only violable ones then a high positive value should result. An analysis of variance indicated that the strength-indices of the five selection task conditions did not differ from one another ( $F(4, 120) = .612$ ;  $p = .655$ ). The overall index was positive (.652) and different from zero ( $t(124) = 14.9$ ;  $p < .001$ ). The data again support the *rule-evaluation hypothesis*: The explicit rules were regarded as violable by most participants (R2: 79.2%; P2: 81.6%), the implicit rules were not (R1: 11.2%; P1: 19.2%).

The additional task in Experiment 3 asked participants to rank the four rules from "easiest to understand" (= 1) to "most difficult to understand" (= 4). Two rules expressed an obligation and were formulated as backward rules: Obligation O3 focused on the obligation in the if-then clause, release R3 in the otherwise clause. The two other rules expressed a ban and were formulated as forward rules: Ban B3 focused on the ban in the if-then clause, permission P3 in the otherwise clause. All rules thus expressed a deontic constraint. But were they therefore equally easy to understand, too? In order to answer this question, an analysis of variance on the rank values was conducted with one within-subject factor "rule" (rank value assigned to each of the rules) and one between-subjects factor "condition" (the five selection task conditions). As comprehension of the rules should not depend on the type of selection task that individuals had solved before, an effect "condition" was not expected—and was not found either ( $F(4, 124) = .990$ ;  $p = .416$ ). But, a strong rule effect was found ( $F(3, 372) = 210.6$ ;  $p < .001$ ) with a clear rank order: Permission was regarded as most comprehensible ( $m = 1.26$ ), followed by the ban ( $m = 2.19$ ), the obligation ( $m = 2.92$ ), and the release from obligation ( $m = 3.60$ ). An interaction ( $F(12, 372) = 4.13$ ;  $p < .001$ ) indicated a slight difference between the experimental groups: The one exception from the general order *permission < ban < obligation < release* occurred in the group that had solved the selection tasks with obligation O3; here, the obligation switched to the second position: *permission < obligation < ban < release*.

## Conclusions

Any sound test of a psychological theory requires that we understand the applied behavioral paradigms in all relevant aspects. The task analysis presented in this paper (cf. Beller,

in press) indicated that a great deal of the difficulties with the original abstract deontic selection task arose from an inadequate formulation of the task. The experimental data confirm this analysis: They provide evidence for the specific difficulties with the original task and show that performance can be substantially improved, if the selection task is formulated more clearly with regard to its deontic nature.

The abstract deontic selection task was introduced in order to demonstrate that people possess abstract reasoning schemas for reasoning from deontic rules (Cheng & Holyoak, 1985). The reduced performance with the original version (as compared to concrete versions) was attributed by some scholars to lacking cost-benefit information, which is characteristic for social contract rules (e.g., Cosmides, 1989). This explanation can now be rejected: As all the tasks used in Experiments 1 to 3 have the same cost-benefit structure, the differences between the original task and the new tasks unequivocally show that the reduced performance in the original task must have other reasons.

What do the experimental data tell us about people's deontic competencies? Performance was best with tasks that combined a genuine violation instruction with a genuine permission rule: With 72.2% of participants choosing "A & not-P" on average across the three experiments, these tasks reliably approached the average result from the concrete deontic tasks (73.8%, according to Beller, 2010, p. 127). This finding may count as an indication for rule guided behavior in a specific content domain (Smith, Langston & Nisbett, 1992), that is, for System 2 reasoning in a domain-specific, contextualized task. Permission rules were also rated as most comprehensible even though rules like the one used in Experiment 1 do not express a deontic constraint explicitly. Consequently, before people can come up with the "A & not-P" solution in the selection task, they must infer from the rule which action is banned under which condition, and the data are consistent with the assumption that they do this according to the principles suggested in Beller (2008). As a consequence, the exact wording of the rule—whether it is formulated forwards or backwards, as permission or as obligation—is not relevant for the identification of rule violations, as long as people are able to infer the adequate deontic regulation. In this sense, the results constitute an instance of the rule-change phenomenon in showing that people infer cases of rule violation independently of their relation to the "logical form" of the conditional statement (Cosmides, 1989; cf. Beller, 2001; 2010; Beller & Spada, 2003).

Beyond that, tasks like the abstract deontic selection task are important tools for studying domain-specific reasoning processes as they combine two features: They are *content-specific* as they refer to a particular content domain (like deontic norms), and *abstract* as they do not refer to a specific instance, thereby avoiding content effects due to the experience that people have with particular instances (for a causal example see Beller & Kuhn münchen, 2007). As with content effects in general (cf. Beller & Spada, 2003), specific instances of deontic rules may facilitate performance, if experiences with a rule are available directly from memory (for an example see Beller, 2001). Likewise, experience with a

specific instance might also suppress performance, if inhibiting aspects like additional conditions or additional social norms are evoked by its content (for an example see Beller, Bender & Song, 2009). Our theories of reasoning—even the content-specific ones—are not formulated for single instances (like a specific drinking-age rule), but refer abstractly to classes of comparable situations (like deontic norms in general). Tasks that are both, content-specific and abstract, are thus a mean for testing domain-specific theories in a purer way, and can provide a baseline that help us to detect and to assess the extent of such instance-specific content effects.

**Acknowledgements.** We are grateful to Lukas Bischof, Miriam Hansen, Gregory Kuhn münchen, and Nikol Rummel for helping with data collection, and to Lisa Hüther for valuable comments on earlier versions of this article.

## References

- Beller, S. (2001). A model theory of deontic reasoning about social norms. In J. D. Moore, & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.
- Beller, S. (2008). Deontic norms, deontic reasoning, and deontic conditionals. *Thinking and Reasoning*, 14, 305-341.
- Beller, S. (2010). Deontic reasoning reviewed: psychological questions, empirical findings, and current theories. *Cognitive Processing*, 11, 123-132.
- Beller, S. (in press). Concrete problems in the abstract deontic selection task—and how to solve them. *The Quarterly Journal of Experimental Psychology*.
- Beller, S., Bender, A., & Song, J. (2009). Conditional promises and threats in Germany, China, and Tonga: Cognition and Emotion. *Journal of Cognition and Culture*, 9, 115-139.
- Beller, S., & Kuhn münchen, G. (2007). What causal conditional reasoning tells us about people's understanding of causality. *Thinking and Reasoning*, 13, 426-460.
- Beller, S., & Spada, H. (2003). The logic of content effects in propositional reasoning: The case of conditional reasoning with a point of view. *Thinking and Reasoning*, 9, 335-378.
- Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17, 391-416.
- Cosmides L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, 31, 187-276.
- Cox, R. J., & Griggs, R. A. (1982). The effect of experience on performance in Wason's selection task. *Memory & Cognition*, 10, 496-502.
- Evans, J. St. B. T. (2003). In two minds: dual process accounts of reasoning. *Trends in Cognitive Sciences*, 7, 454-459.
- Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, 59, 255-278.
- Kennedy, J. J. (1992). *Analyzing qualitative data*. New York: Praeger.
- Smith, E. E., Langston, C., & Nisbett, R. E. (1992). The case for rules in reasoning. *Cognitive Science*, 16, 1-40.
- Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology*. Harmondsworth, UK: Penguin.

# That’s what she (could have) said: How alternative utterances affect language use

Leon Bergen (bergen@mit.edu)<sup>1</sup>, Noah D. Goodman (ngoodman@stanford.edu)<sup>2</sup>, Roger Levy (rlevy@ucsd.edu)<sup>3</sup>

<sup>1</sup>Department of Brain and Cognitive Sciences, MIT, Cambridge MA 02139,

<sup>2</sup>Department of Psychology, Stanford University, Stanford CA 94305,

<sup>3</sup>Department of Linguistics, UC San Diego, La Jolla CA 92093

## Abstract

We investigate the effects of alternative utterances on pragmatic interpretation of language. We focus on two specific cases: specificity implicatures (less specific utterances imply the negation of more specific utterances) and Horn implicatures (more complex utterances are assigned to less likely meanings). We present models of these phenomena in terms of recursive social reasoning. Our most sophisticated model is not only able to handle specificity implicature but is also the first formal account of Horn implicatures that correctly predicts human behavior in signaling games with no prior conventions, without appeal to specialized equilibrium selection criteria. Two experiments provide evidence that these implicatures are generated in the absence of prior linguistic conventions or language evolution. Taken together, our modeling and experimental results suggest that the pragmatic effects of alternative utterances can be driven by cooperative social reasoning.

**Keywords:** Pragmatics; Communication; Bayesian modeling

## Introduction

A central observation in the field of pragmatics is that alternative utterances affect our interpretation of language. If the teacher says, “Some of the students passed the test”, then this means that not all of them passed, because the teacher would have said so if they did. If someone is asked what they ate at the restaurant, and they say “salad”, then this means that they did not also get the lobster; otherwise they would have said so. If someone says, “I got the car to turn on,” then this means that turning on the car involved something more unusual than just turning the key. If it hadn’t, they could have just said, “I turned on the car.” Many other cases like this were described in Grice’s (1975) classic.

Horn (1984) proposed a unified account of these disparate cases in terms of his Q and R-Principles. These principles describe how conversational partners are expected to communicate with each other. The Q-Principle states: Say as much as you can. The R-Principle states: Say no more than you must. These principles explain how counterfactual utterances like the ones above have their effect on meaning. If the speaker behaves according to the Q-Principle, then when she says that some of the students passed the test, this must mean that she said all that she could. In particular, she must not have been in a position to say that all of the students passed. Similarly, if the speaker is following the R-Principle, then when she reports that she got the car to turn on, this means that a simpler utterance would not have sufficed to convey her meaning. In particular, simply saying that she turned on the car would not have conveyed her meaning.

A basic question about these principles (or Grice’s related maxims of conversation) is the extent to which they capture people’s online reasoning when they pragmatically interpret

language. The alternative is that such explanations merely provide a succinct way of summarizing pragmatic phenomena, and that pragmatic meanings are learned as part of the grammar, i.e. conventionalized as part of the language. Researchers have argued that some types of pragmatic meanings are computed from the grammar, and not from cooperative social reasoning (Chierchia, 2004). Intuitively, however, some pragmatic inferences generalize to settings in which they could not have been previously learned. Consider the salad/lobster inference described above. This inference is highly context-dependent, and must require a reasoning process that extends beyond what has been learned in the grammar. But then where is the boundary between conventionalized and socially-derived implicatures?

Here we investigate these questions using experiments and computational modeling. Despite the apparent simplicity of explanations in terms of the Q and R-principles, it has been notoriously difficult to develop a formal framework that captures these principles (or the maxims of conversation). In addition, there has been little empirical work investigating whether pragmatic inferences rely on conventionalized meanings. We will be looking at the minimal ways for contrasts between alternative utterances to drive pragmatic interpretation. Specifically, we will be looking at cases in which there are no linguistic conventions whatsoever. If people’s pragmatic interpretations show the same sensitivity to contrast in these settings, it will provide evidence that social reasoning explains more, rather than less, of their pragmatic abilities.

We focus on two traditional examples of counterfactual reasoning in pragmatics. The first, *scalar implicatures*, arise because of the contrast between words that fall on an increasing scale of informativeness. The less informative meaning is typically strengthened to the complement of the more informative meaning, as in the case of “some” vs. “all” above. Because the term “scalar implicature” is sometimes reserved to refer to cases where lexical items fall on a canonical scale of informativeness, we will use the term *specificity implicatures* to describe the strengthening of less informative meanings even in the absence of such a canonical scale. The second, which we will call *Horn implicatures*, guides the interpretation of utterances that differ in their complexity (Horn, 1984). Typically, more complex constructions receive marked (or less probable) interpretations. The car case above is an example of a Horn implicature, assuming (as is plausible) that the two expressions have the same literal content.

These two kinds of implicatures will allow us to explore pragmatic contrast effects along two distinct dimensions: informativeness and cost. While we will model both of these ef-

fects as recursive social reasoning, it will emerge that the simplest model of such reasoning that can account for specificity implicatures is insufficient to explain Horn implicatures. We begin with the simpler version of the model and later enhance the model to account for both kinds of effects.

### Specificity Implicatures

The Gricean tradition views pragmatics as a special domain of social cognition. Pragmatics, on this approach, is the study of social agents who want to cooperate with each other to exchange information. Pragmatic phenomena arise as a result of these goals and the agents' reasoning about each other.

Here we develop a model of ideal discourse between two rational agents, a speaker and listener, each with distinct social goals. The speaker wants to communicate a specific meaning to the listener, requiring her to reason about how the listener will interpret her possible utterances. The listener, in turn, wants to determine what meaning the speaker intended to convey, requiring him to reason about which meaning would have led the speaker to send her utterance. The speaker and listener are modeling each other; crucially, the listener takes into account the speaker modeling him, and the speaker takes *this* into account, and so on. In other words, the speaker and listener have *common knowledge* of each others' communicative goals (Lewis, 1969; Clark, 1996).

This recursive social reasoning bottoms out when the listener stops reasoning about the speaker's intentions. In this *base case*, the listener uses his knowledge of the language's semantics or contextual iconicity to interpret the utterance.

We now turn to the formal specification of the model. The literal content of the utterances is specified by a lexicon  $\mathcal{L}$ , which maps each utterance to a truth function on meanings. If an utterance has no conventional or iconic meaning, then it is given the all-*true* function (i.e. it is a tautology). The listener has a prior distribution  $P$  over meanings; in the base case, the listener uses Bayesian inference to update her belief about the intended meaning given the utterance's literal meaning. More precisely, the listener conditions the prior distribution on the utterance being true, essentially filtering  $P$  through the literal meaning, leading to a new distribution  $L_0$  with support only on meanings that are consistent with the utterance. That is:

$$L_0(m|u, \mathcal{L}) \propto \mathcal{L}_u(m)P(m), \quad (1)$$

where  $m$  is a meaning,  $u$  is the utterance sent by the speaker,  $\mathcal{L}_u$  is a function from meanings to  $\{0,1\}$ , with  $\mathcal{L}_u(m) = 1$  if  $m$  is in the denotation of  $u$ , and  $P$  is the listener's prior distribution over meanings.

Social reasoning enters the model through a pair of recursive formulas that describe how the speaker and listener reason about each other. The formulas describe Bayesian agents  $S_n$  and  $L_n$  of increasing sophistication. The least sophisticated speaker  $S_1$  reasons about the base, "literal" listener  $L_0$ ; a slightly more sophisticated listener  $L_2$  reasons about this speaker; and so on. The speaker  $S_n$  has a utility function  $U_n$  that simultaneously accounts for how informative an utterance is for listener  $L_{n-1}$  as well as for its complexity or cost.

This is intended to capture both the Q and R-Principles. The speaker's choice of utterance is determined by a softmax decision rule that describes an approximately optimal Bayesian decision-maker (Sutton & Barto, 1998). The listener  $L_n$  interprets an utterance by using Bayesian inference to integrate his prior expectations over meanings, given by  $P$ , with his model of  $S_{n-1}$ , which determines how likely the speaker would have been to use that utterance given each possible meaning.

This recursive model is defined as follows. The speaker's conditional distribution over utterances given interpretations is defined as

$$S_n(u|m) \propto e^{\lambda U_n(u|m)}, \quad (2)$$

where  $\lambda > 0$  is the gain on the speaker's softmax decision rule.  $U_n(u|m)$  is the speaker's expected utility from uttering  $u$  to convey  $m$ , defined as

$$U_n(u|m) = \log(L_{n-1}(m|u)) - c(u). \quad (3)$$

Here  $c(u)$  is the cost of uttering  $u$  (in, e.g., time and effort); the other term measures the communicative benefit of  $u$  as the number of bits of information remaining between the listener's posterior distribution  $L_{n-1}(m|u)$  and the true meaning  $m$  (Frank, Goodman, Lai, & Tenenbaum, 2009). Substituting equation 3 into equation 2, we see that

$$S_n(u|m) \propto (L_{n-1}(m|u)e^{c(u)})^\lambda. \quad (4)$$

Hence the speaker prefers low-cost utterances and also prefers to choose an utterance more as the listener is more likely to pick the correct meaning given the utterance. The listener's higher-order interpretations are simply defined as

$$L_n(m|u) \propto P(m)S_{n-1}(u|m). \quad (5)$$

The model defined here is very similar to the iterated best response model of (Jäger & Ebert, 2009).

In situations where the speaker has a choice between utterances one of whose literal meanings is a subset of the other, this model induces an inference that the utterance with the broader meaning should be interpreted as indicating that the narrower meaning does not hold—which we term a specificity implicature. To see why, suppose that there are two possible meanings, *pyramid* and *cube*, and two utterances, "pyramid" and "shape", both of which have equal cost. The literal listener interprets "pyramid" as meaning *pyramid* with probability 1, due to the truth-conditional component of literal interpretation, and interprets "shape" as meaning either *pyramid* or *cube* with probabilities based on the prior,  $P(\text{pyramid})$  and  $P(\text{cube})$  respectively. For the speaker  $S_1$  reasoning about the literal listener, conveying *pyramid* with "pyramid" has higher utility than conveying it with "shape", since the former term ensures the proper interpretation.  $S_1$  is thus more likely to say "pyramid" than "shape" when she means to convey *pyramid*; and she will obligatorily say "shape" when she means to convey *cube*. The more sophisticated listener  $L_2$  uses  $S_1$ 's distributions rather than literal meaning, and thus prefers to interpret "shape" as *cube* rather than *pyramid*, since

the likelihood of “shape” is greater when *cube* is to be conveyed than when *pyramid* is. As the speaker and listener reach higher levels of recursive reasoning, these tendencies to say “pyramid” for *pyramid* and to interpret “shape” as *cube* continue to strengthen, both ultimately asymptoting at probability 1 (Figure 2, pink bars; the asymptotes do not depend on  $\lambda$  or  $P$ , so long as  $\lambda > 1$  and  $P(m) > 0$  for all  $m$ ).

## Experiment 1

Experiment 1 investigated whether people will draw specificity implicatures in a novel communicative setting. We investigated this by looking at the simplest setting in which specificity implicatures are possible, a language with only two messages and two meanings. By varying the (non-conventional) semantic content of the messages, we can determine whether competing messages influence interpretations here as they do in richer, conventionalized settings.

We presented people with a simple communication game, which they played with a partner. In the game, one player was the “speaker”, who had a specific meaning to communicate, and one player was the “listener”, who had to infer this meaning based on the message sent by the speaker. The meaning for the speaker to communicate was randomly chosen to be either a pyramid or a cube. The speaker had the choice between two messages to send the listener: a shape with an iconic relationship with one of the meanings (a triangle for the pyramid), or an “alien” symbol with no obvious connection to either meaning (see Figure 1). If people’s reasoning about semantic competition extends to novel settings, then the alien symbol will get a strengthened interpretation: it should be interpreted as the cube, which is the meaning that the speaker could not directly pick out. Likewise, the speaker will recognize that choosing the alien symbol is more likely to communicate the meaning for which there is no iconic symbol available. We thus predict that the speaker will use the alien symbol to communicate this meaning.

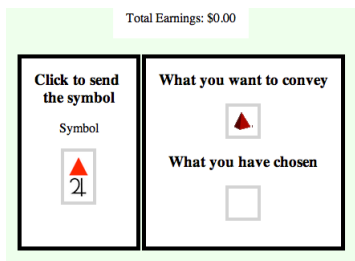


Figure 1: Experiment 1 game interface for speaker.

## Methods:

We recruited 40 participants from Amazon Mechanical Turk. They were paid a small amount of money for participation, in addition to performance-based bonuses, as described below.

Participants were told that they had arrived on an alien planet that contained two objects, a pyramid and a cube. Their

goal was to successfully play a communication game with a partner, given these objects and two messages that they could send, a triangle and an alien symbol. Participants received 10 rounds of practice as the speaker in order to familiarize with the interface; however, they did not play these rounds with a partner or receive any feedback.

The communication game consisted of five rounds. In each round, participants were randomly assigned a partner and a role as either the speaker or listener. Between rounds, participants were told that they were being randomly matched with a partner. Participants were never identified to each other. The speaker was shown a randomly chosen object that needed to be communicated, and given the choice of the two messages to send. Once the speaker clicked on a message, it was sent to the listener, who was asked to determine what meaning was intended. The listener was given the choice of the two objects; once the listener clicked on one of these objects, the speaker and listener were informed whether their communication was successful. If they were successful, they each received a small bonus payment of \$0.06 for that round.

## Results and Discussion

There were two questions of interest in this experiment. The first was whether listeners would interpret the alien symbol as the unnamed object (i.e. the object without an iconic message), i.e. the cube. On every trial, the listener interpreted the alien symbol as the unnamed object and the iconic symbol as the corresponding object. The second question was whether the speaker would choose the alien symbol to convey the unnamed object. Participants selected the alien symbol on every trial on which they needed to communicate the unnamed object; and they selected the name on all but two trials on which they needed to communicate the named object. These results are shown in Figure 2. (The displayed model predictions were not sensitive to the value of the model parameters.)

These results provide evidence that participants were sensitive to the semantic contrast between available utterances. Listeners inferred that the speaker would have only used the alien symbol if she needed to communicate the unnamed object. Speakers similarly inferred that the listener would interpret the alien symbol as the unnamed object, and only chose the alien symbol in order to communicate this object.

## Horn’s Principle

HORN’S PRINCIPLE describes the effects of lexical competition when utterances differ in cost instead of semantic content. The principle states that phrases that are “costlier”—e.g., longer, or involving less-frequent subexpressions—are associated with less probable meanings. For example, *I turned on the car* and *I got the car to turn on* have approximately identical literal meaning; but most speakers would use the shorter sentence to refer to the typical turning of a car key and the longer sentence to some less typical manner of turning on the car. This is the efficient mapping between form and meaning; Horn (1984) and others have documented many instances of such efficient mappings in language.

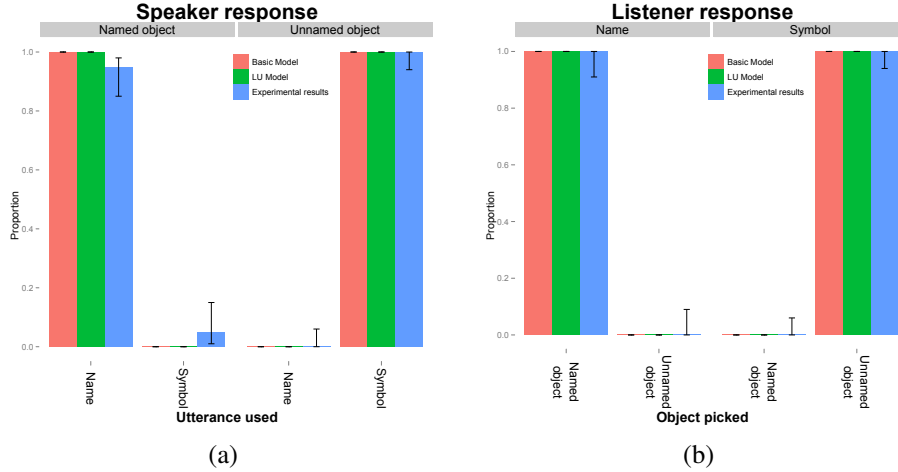


Figure 2: Experiment 1 results and model predictions. (a) Speaker responses given the goal of communicating the named object (left) or the unnamed object (right). The y-axis is the proportion of trials on which the speaker chose each utterance. Error bars are 95% confidence intervals. (b) Listener responses when the speaker chose either the name (left) or the alien symbol (right). The y-axis is the proportion of trials on which the listener chose each object.

Perhaps surprisingly, however, the model of intuitive cooperative communication we have introduced fails to predict Horn’s principle. Consider the problem of a one-shot speaker-listener signaling game with two utterances, “expensive” and “cheap” (the costs of these utterances reflect their names), and two meanings, *likely* and *unlikely*; nothing distinguishes the utterances other than their cost, and each has the all-true meaning. The literal listener  $L_0$  interprets both utterances identically, matching the prior probabilities of the meanings.  $L_0$ ’s interpretation thus provides no information with which the speaker  $S_1$  can distinguish among the utterances; the only thing distinguishing the utterances’ utility is their cost. This leads to an across-the-board dispreference on the part of  $S_1$  for “expensive”, but gives no starting point for more sophisticated listeners or speakers to break the symmetry.

In fact, Horn’s principle has been extraordinarily difficult to derive within other formal frameworks as well. The problem we posed is equivalent to the multiple equilibrium problem for signaling games in economics, which has been investigated for the last 30 years (Cho & Kreps, 1987; Chen, Kartik, & Sobel, 2008). The fundamental difficulty involves ruling out inefficient equilibria, e.g. those in which “expensive” is associated with *likely* and “cheap” with *unlikely*, in the absence of prior conventions or ad-hoc rules for choosing among equilibria. Some recent work by linguists (van Rooij, 2004, 2008; Franke, 2009) has attempted to derive Horn’s principle by using evolutionary game theory or hybrid game-theoretic models, but none has found a derivation for one-shot signaling games, without use of equilibrium refinement criteria specifically designed to pick out the desired equilibria.

Here we propose exactly such a derivation, by revisiting the assumption in our base model regarding the nature of literal meaning in the absence of prior conventions. Earlier in this section we assumed that the absence of prior convention should be represented as a single lexicon  $\mathcal{L}$  in which all utterances have the all-true (tautological) meaning. We revise that assumption in two respects. First, for any given utterance we allow  $\mathcal{L}_u$  to assume either truth value  $\{0, 1\}$  for each meaning,

allowing the lexicon to assign non-trivial semantic content to utterances. Second, we allow for *lexical uncertainty*, where the speaker and listener can reason about distributions over multiple lexica. In the signaling game described above, for example, lexicon  $\mathcal{L}^1$  might assign the meaning *likely* to “expensive” and the all-true meaning to “cheap”, whereas lexicon  $\mathcal{L}^2$  might assign the meaning *unlikely* to “expensive” and the same all-true meaning to “cheap”. The absence of prior conventions then means that the *marginal* interpretation, across lexica, is the same for all utterances.

Including lexical uncertainty generalizes the previous model; the base listener  $L_0$  remains unchanged from equation 1, but the more sophisticated speaker and listener are defined by:

$$S_n(u|m, \mathcal{L}) \propto e^{\lambda U_n(u|m, \mathcal{L})} \quad (6)$$

$$L_n(m|u) \propto \sum_{\mathcal{L}} P(m) P(\mathcal{L}) S_{n-1}(u|m, \mathcal{L}) \quad (7)$$

where

$$U_n(u|m, \mathcal{L}) = \begin{cases} \log(L_0(m|u, \mathcal{L})) - c(u) & \text{if } n = 1 \\ \log(L_{n-1}(m|u)) - c(u) & \text{if } n > 1. \end{cases} \quad (8)$$

We take  $P(\mathcal{L})$  to be the uniform distribution over all seven logically possible lexica in which every utterance assigns “true” to at least one meaning and every meaning is assigned “true” by at least one utterance.

Because this new *lexical-uncertainty* model reduces to the base model when conventions for literal meanings are already established and there is only a single lexicon  $\mathcal{L}$ , the new model continues to properly handle specificity implicature (Figure 2, green bars). Furthermore, the new model derives Horn’s principle. Consider the case we proposed above with two lexica,  $\mathcal{L}^1$  interpreting “expensive” as *likely* and  $\mathcal{L}^2$  interpreting “expensive” as *unlikely* (and both giving a trivial interpretation to “cheap”). Due to the role of the prior  $P(m)$ , the base listener  $L_0$  associates “cheap” with *likely* for both lexica. Now consider two speakers who can use the expensive utterance to precisely communicate their meaning: speaker  $S_1(\cdot|likely, \mathcal{L}^1)$  who wants to communicate *likely* and is using



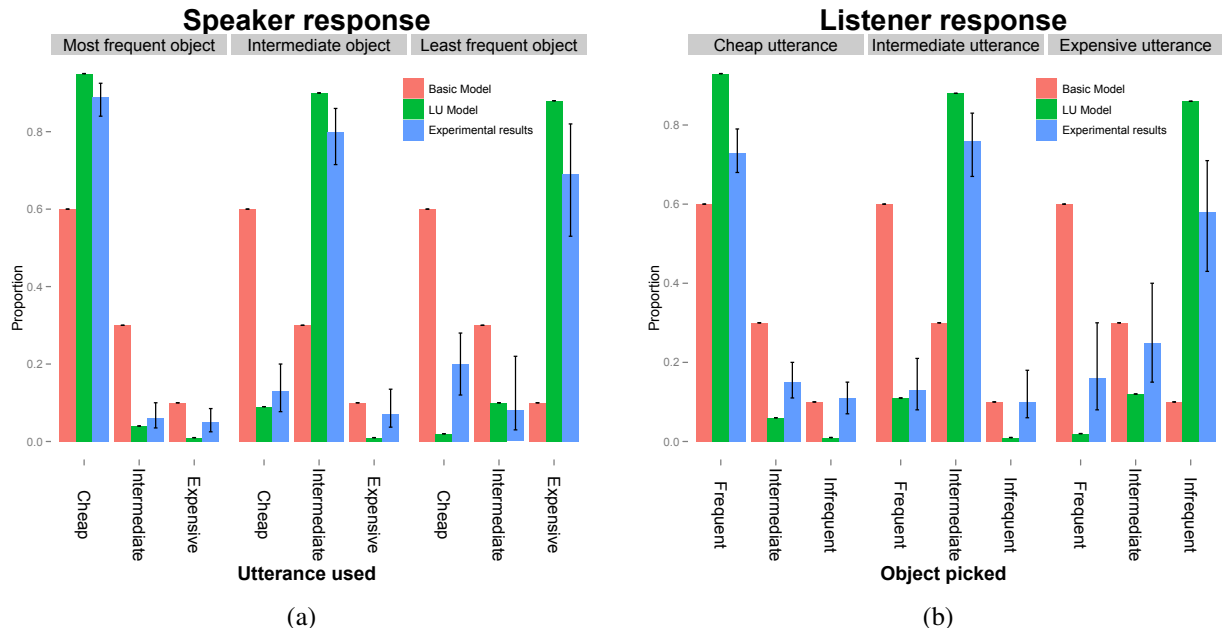


Figure 3: Experiment 2 results and model predictions. (a) Speaker’s message choice given that she needed to convey the most frequent (left), intermediate (center), or least frequent object (right). The y-axis is the proportion of trials on which each message was chosen. Error bars are 95% confidence intervals. (b) Listener’s object choice after receiving the cheapest (left), intermediate (center), or most expensive message (right). The y-axis is the proportion of trials on which each object was chosen.

lexicon  $\mathcal{L}^1$  and speaker  $S_1(\cdot|unlikely, \mathcal{L}^2)$  who wants to communicate *unlikely* and is using lexicon  $\mathcal{L}^2$ . For the speaker  $S_1(\cdot|likely, \mathcal{L}^1)$ , the extra precision of “expensive” is not valuable, because the base listener  $L_0$  will also interpret “cheap” as *likely*. However, for the speaker  $S_1(\cdot|unlikely, \mathcal{L}^2)$ , the extra precision of “expensive” is valuable, because it overrides the base listener’s prior bias against *unlikely*. This breaks the symmetry and leads  $L_2$  to start to prefer the efficient mapping, a preference that gets magnified by  $S_3$ ’s reasoning and continues to get magnified at higher levels of recursive inference.

The lexical uncertainty model can predict correct Horn equilibria beyond the case with two meanings and utterances; predictions (approximated by averaging over a finely gridded approximation to the parameter space) for the case with three meanings and utterances are shown in Figure 3.

## Experiment 2

In Experiment 2 we investigated whether people can coordinate on Horn’s principle in the absence of prior linguistic conventions. This experiment was designed to be a minimal test of this question: people were placed in the simplest setting in which Horn’s principle is possible. People played a communication game with a partner as in Experiment 1. There were three possible meanings for the speaker to communicate, which differed in how frequently they appeared in the game. The speaker was able to communicate the intended meaning by sending one of three messages, which differed only in their cost to the speaker—none was iconic. If the

same reasoning that gives rise to Horn’s principle extends to this novel setting, then we expect the speaker and listener to coordinate on the efficient mapping of meanings to messages.

## Methods:

We recruited 140 participants from Amazon Mechanical Turk, who were paid for participation in addition to bonus payments described below.

The interface and instructions for the experiment were very similar to Experiment 1. Participants were told that they landed on an alien planet containing three kinds of objects and three messages that could be used to communicate these objects. They were told that these objects occurred with different frequencies on the planet; one occurred 60% of the time, one occurred 30%, and the last occurred 10%. Of the three messages, one was free, one cost \$0.01, and the last cost \$0.02. The object frequencies and message costs were randomized between subjects. Participants received 10 practice rounds without a partner to familiarize them with the interface, object frequencies, and message costs, but received no feedback during these rounds.

Each participant played 5 rounds of the game with a partner randomly assigned each round. The game was the same as in Experiment 1, with two changes. The object that the speaker needed to communicate was randomly sampled according to the frequency of the objects on the alien planet (so, e.g., the most frequent object was sampled 60% of the time). Second, the speaker was charged the cost of the message sent.

Table 1: Experiment 2 analyses

Role	Response	Comparison	t-value	p-value
Listener	Frequent object	Cheap utterance > intermediate, expensive	6.07	0.001
	Intermediate object	Intermediate utterance > cheap, expensive	5.31	0.001
	Unlikely object	Expensive utterance > cheap, intermediate	5.55	0.001
Speaker	Cheap utterance	Frequent object > intermediate, unlikely	8.27	0.001
	Intermediate utterance	Intermediate object > frequent, unlikely	6.21	0.001
	Expensive utterance	Unlikely object > frequent, intermediate	5.55	0.001

## Results and Discussion

Human speaker and listener choices are shown in figure 3, alongside the predictions of our base model (pink bars), which does not predict the Horn equilibrium, and our lexical-uncertainty model (green bars), which does. We first analyzed whether listeners interpreted messages according to the efficient mapping, by carrying out three mixed logit regressions with random intercepts for participants. We analyzed whether, e.g., listeners responded with the frequent object more often when they received the cheap utterance than the other utterances. These comparisons are shown in Table 1. They show that the listener's responses were consistent with the efficient mapping. We next analyzed whether speakers chose messages efficiently. We addressed this question in a similar manner to the previous one, carrying out three mixed logit regressions with random intercepts for participants. The comparisons in table 1 show that the speaker was more likely to use the cheap utterance given the frequent object than given the other objects, and similarly for the other utterances.

By design, each participant only played five rounds of the game. This was done to ensure that observed efficiency effects were due to cooperative reasoning, and not due to language evolution. To validate this design, we analyzed whether participants played differently on the first round than on future rounds. To do this, we performed by-subject ANOVAs on the speaker and listener responses to determine whether there were main effects or interactions from the first round. For five of the six speaker and listener response types, there was no main effect of the first round or interaction with the object to communicate or message received ( $p > 0.05$ ). For the intermediate-cost message, there was a small but significant interaction between the first round and the object to communicate ( $p < 0.05$ ). These analyses provide evidence that learning or language evolution are not driving our results.

These results provide evidence that people can construct efficient strategies for communication on-line, in the absence of prior linguistic conventions. This suggests that Horn's principle, as it applies to ordinary language use, may arise from cooperative reasoning between people trying to communicate with each other, rather than from language evolution.

## Discussion

We have investigated two kinds of pragmatic contrast effects. First, we looked at the effect of varying the specificity, or informativeness, of the utterances available to the speaker. A simple model of social cognition was able to account for the strengthening of a non-specific utterance's interpretation. We found in Experiment 1 that listeners inferred this strengthened interpretation, and that speakers anticipated this, in the absence of any prior linguistic conventions.

We next turned to Horn's principle and the effect of varying the relative cost or complexity of utterances. We first found that the simple model of social cognition cannot explain Horn's principle, because it does not have the resources to exploit the asymmetry between more and less costly utterances. However, once the model included uncertainty about the underlying lexicon, it was able to predict Horn's principle, and explain this asymmetry: under the model, only someone who wanted to communicate an unlikely meaning would have an incentive to use an expensive utterance. Notably, this is the first proposed model of Horn's principle which does not rely on specialized equilibrium selection criteria. In Experiment 2, we found evidence that people expect costlier utterances to correspond to less frequent meanings, as predicted by Horn's principle.

Our results suggest that in two important cases, people's pragmatic knowledge extends beyond learned grammatical knowledge to novel, non-linguistic communicative scenarios. While we have not settled the question of how people arrive at ordinary pragmatic inferences, our experimental and modeling results do provide evidence that linguistic conventions are not necessary for them—social cognition will suffice.

## References

- Chen, Y., Kartik, N., & Sobel, J. (2008). Selecting cheap-talk equilibria. *Econometrica*, 76(1), 117–136.
- Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. *Structures and beyond*, 3, 39–103.
- Cho, I., & Kreps, D. (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2).
- Clark, H. (1996). *Using language*. Cambridge University Press (Cambridge England and New York).
- Frank, M. C., Goodman, N. D., Lai, P., & Tenenbaum, J. B. (2009). Informative communication in word production and word learning. In *Proc. cog. sci. soc.*
- Franke, M. (2009). Interpretation of optimal signals. *New perspectives on games and interaction*, 297–310.
- Grice, H. (1975). Logic and conversation. 1975, 41–58.
- Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications* (pp. 11–42).
- Jäger, G., & Ebert, C. (2009). Pragmatic rationalizability. In *Proceedings of sinn und bedeutung* (Vol. 13, pp. 1–15).
- Lewis, D. (1969). *Convention: a philosophical study*. Harvard University Press.
- Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction* (Vol. 28). Cambridge Univ Press.
- van Rooij, R. (2004). Signalling games select horn strategies. *Linguistics and Philosophy*, 27(4), 493–527.
- van Rooij, R. (2008). Games and quantity implicatures. *Journal of Economic Methodology*, 15(3), 261–274.



# Word predictability and frequency effects in a rational model of reading

Klinton Bicknell<sup>1</sup> (kbicknell@ucsd.edu) & Roger Levy<sup>2</sup> (rlevy@ucsd.edu)

<sup>1</sup>Department of Psychology, UC San Diego, La Jolla, CA, USA

<sup>2</sup>Department of Linguistics, UC San Diego, La Jolla, CA, USA

## Abstract

This paper presents results from the first rational model of eye movement control in reading to make predictions for the full range of the eye movement record. The model identifies the text through Bayesian inference and makes eye movement decisions to maximize the efficiency of text identification, going beyond leading approaches which select model parameters to maximize the fit to human data. Two simulations with the model demonstrate that it can produce effects of word predictability and frequency on eye movements in reading similar to those produced by humans, providing evidence that many properties of human reading behavior may be understood as following from the nature of efficient text identification.

**Keywords:** eye movements; reading; rational analysis; computational modeling

## Introduction

During reading, comprehenders must decide when and where to move their eyes 3–4 times every second. Over the past decades, it has been demonstrated that comprehenders make these rapid, fine-grained decisions by combining information from a range of sources including visual input, the motor system, and linguistic knowledge (for reviews see Rayner, 1998, 2009), making reading one of the most complex learned tasks that humans face every day. Gaining a better understanding of this process promises to yield insights about how readers deploy linguistic knowledge for real-time comprehension as well as about how humans learn to perform complex tasks more generally. In this paper, we present the first rational model of eye movement control in reading that makes predictions for the full range of the eye movement record. We model readers as performing Bayesian inference on the identity of the text, combining their probabilistic language knowledge (the prior) with noisy perceptual input about the text (the likelihood) to form and repeatedly update a posterior distribution over the possible text identities. The model uses a parameterized behavior policy for determining when and where to move the eyes, which is sensitive to the posterior distribution over the text, and with parameters selected to optimize identification efficiency. We evaluate the model by examining the effects it produces for two linguistic variables: word frequency and predictability. We present the results of two simulations showing that the model produces effects of these variables similar to those of humans, across four different eye movement measures reflecting both the locations and durations of fixations. The success of the model in deriving these effects from principles of probabilistic inference and rational action suggests that many aspects of human reading behavior may be profitably understood as properties of the set of efficient solutions to the problem of reading.

This model goes beyond leading models of eye movement control in reading such as E-Z Reader (Reichle, Rayner, &

Pollatsek, 2003) and SWIFT (Engbert, Nuthmann, Richter, & Kliegl, 2005) in two ways. First, while those models select parameters to maximize the fit to human data, the current work selects parameters to optimize the efficiency of reading, here characterized as rapid and accurate identification of the contents of the text. To the extent that the model behavior reproduces effects seen in human data, then, it enables understanding those effects as resulting from the properties of efficient solutions to the task. Second, the current work includes a model of the process of identification from visual input, and in so doing derives effects of linguistic variables (such as word frequency and predictability) as resulting from efficient identification, while models such as E-Z Reader and SWIFT directly specify the effects of linguistic variables on eye movement behavior through functions whose form is stipulated exogenously to the model. Modeling identification from visual input should allow for the model to be used to understand a range of effects that are known to influence eye movements but which leading approaches cannot capture, such as information density within words (Hyönä, Niemi, & Underwood, 1989), word misidentification (Slattery, 2009; Levy, Bicknell, Slattery, & Rayner, 2009), and visual neighborhoods (Pollatsek, Perea, & Binder, 1999). The model also goes beyond the only previous rational model of eye movement control in reading, Mr. Chips (Legge, Hooven, Klitz, Mansfield, & Tjan, 2002), in making predictions about not only the location of fixations but also their duration, which is important to gaining a full understanding of a range of effects on eye movements in reading, especially the effects of linguistic variables.

In the following section, we describe our rational framework for reading and the details of our model of eye movement control in reading. We then focus the remainder of the paper on using the model to understand the effects on eye movements in reading of word frequency and predictability, two of the most reliable linguistic effects in the eye movement record. We first use the model qualitatively to provide explanations for why the effects of these variables seen empirically should result from efficient reading behavior, and then present the quantitative results of two simulations demonstrating that these effects are evident in the model's behavior.

## Reading as Bayesian inference

In the proposed framework, we model the goal of reading as efficient text identification. While it is clear that this is not all that readers do – inferring the underlying structural relationships among words in a sentence and discourse relationships between sentences that determine text meaning is a fundamental part of most reading – all reader goals necessarily in-

volve identification of at least part of the text, so we take text identification to be a reasonable first approximation. There are two sources of information relevant to this goal: visual input and language knowledge, which the model combines via Bayesian inference. Specifically, it begins with a prior distribution over possible identities of the text given by its language model, and combines this with noisy visual input about the text at the eyes' position (giving the likelihood term) to form a posterior distribution over the identity of the text taking into account both the language model and the visual input obtained thus far. On the basis of the posterior distribution, the model then decides whether or not to move its eyes (and if so where to move them to) and the cycle repeats.

An implemented model in this framework must formalize a number of pieces of the reading problem, including the possible actions available to the reader and their consequences, the nature of visual input, the nature of language knowledge, a means of combining visual input with prior expectations about the form and structure of the text, and a behavior policy determining how the model will choose actions on the basis of its posterior distribution over the identity of the text. In the remainder of this section, we present the details of our formalizations of these pieces.<sup>1</sup>

### Formal problem of reading: Actions

We assume that on each of a series of discrete timesteps, the model obtains visual input around the current location of the eyes, and then chooses between three actions: (a) continuing to fixate the currently fixated position, (b) initiating a saccade to a new position, or (c) stopping reading. If the model chooses option (a), time simply advances, and if it chooses option (c), then reading immediately ends. If a saccade is initiated (b), there is a lag of two timesteps, representing time required to plan a saccade, during which the model again obtains visual input around the current position, and then the eyes move toward the intended target. Because of motor error, the actual landing position of the eyes is normally distributed around the intended target with standard deviation given by a linear function of the intended distance, with parameters taken from Engbert et al. (2005).<sup>2</sup>

### Noisy visual input

The visual input obtained by a reader on a given timestep is generated from the following process, independently for each character position. Each letter is represented as a 26-dimensional vector, where a single element is 1 and the others are zeros, and visual input about a letter is a sample from a 26-dimensional Gaussian with a mean equal to the letter's true identity and a diagonal covariance matrix  $\Sigma = \lambda^{-1}I$ , where  $\lambda$  is the reader's visual acuity at that position. Higher visual

acuity, then, means a lower sample variance, yielding higher quality visual input. We use the visual acuity function from Engbert et al. (2005), in which  $\lambda$  decreases exponentially with retinal eccentricity and decreases asymmetrically, falling off more slowly to the right than the left.<sup>3</sup> In order to scale the quality of visual information, we multiply each acuity  $\lambda$  by the overall visual input quality  $\Lambda$  (values given in the simulations below.) Visual input about non-alphabetic characters is veridical knowledge of their identity. Visual input is limited to the 19 character positions with the highest acuity (eccentricities between -7 and 12), roughly corresponding to estimates that readers of English obtain useful information from about 19 characters, and more from the right of fixation than the left (Rayner, 1998). Note that in the model each letter is equally confusable with all others, following Norris (2006, 2009), but ignoring work on letter confusability (which could be added to future model revisions; Engel, Dougherty, & Brian Jones, 1973; Geyer, 1977).

### Language knowledge

In general, any generative model of linguistic knowledge that assigns probabilities to text can be used as the prior distribution on the identity of the text. For the simulations in this paper, we use very simple probabilistic models of language knowledge: word  $n$ -gram models (Jurafsky & Martin, 2009). These models encode the probability of each word conditional on the  $n - 1$  previous words. While this is obviously a crude representation of the rich knowledge of language that human readers have, it serves here to illustrate the qualitative effects of using linguistic context in reading.

### Inference about text identity

Given both visual input and language knowledge, the model makes inferences about the identity of the text  $w$  via standard Bayesian inference, where the prior is given by the probability of generating text identity  $w$  from the language model and the likelihood is the probability of generating the visual input  $\mathcal{I}$  from text with identity  $w$  under the visual input model:

$$p(w|\mathcal{I}) \propto p(w)p(\mathcal{I}|w).$$

### Behavior policy

The model uses a simple policy with two parameters,  $\alpha$  and  $\beta$ , to decide between actions based on the marginal probability  $m$  of the most likely character  $c$  in each position  $j$ ,

$$m(j) = \max_c p(w_j = c)$$

where  $w_j$  indicates the character in the  $j$ th position. A high value of  $m$  indicates relative confidence about the character's identity, and a low value relative uncertainty.

Given the values of this statistic  $m$ , the model decides between four possible actions, as illustrated in Figure 1. If the

<sup>1</sup>See Bicknell and Levy (2010b) for further computational details.

<sup>2</sup>In the terminology of the literature, the model has only random motor error (variance), not systematic error (bias). Following Engbert and Krügel (2010), systematic error may arise from Bayesian estimation of the best saccade distance.

<sup>3</sup>While we call refer to it here as visual acuity, it is clear from the asymmetric nature of this function that it also has an attentional component. For now, however, we make the simplifying assumption that it is unchanging over time.

- (a)  $m = [.6, .7, \mathbf{.6}, .4, .3, .6]$ : Keep fixating (3)
- (b)  $m = [.6, .4, \mathbf{.9}, .4, .3, .6]$ : Move back (to 2)
- (c)  $m = [.6, .7, \mathbf{.9}, .4, .3, .6]$ : Move forward (to 6)
- (d)  $m = [.6, .7, \mathbf{.9}, .8, .7, .7]$ : Stop reading

Figure 1: Values of  $m$  for a 6 character text under which a model fixating position 3 would take each of its four actions, if  $\alpha = .7$  and  $\beta = .5$ .

value of this statistic for the current position of the eyes is less than the parameter  $\alpha$ , the model chooses to continue fixating the current position (1a). Otherwise, if the value of  $m(j)$  is less than the parameter  $\beta$  for some leftward position, the model initiates a saccade to the closest such position (1b). If no such positions exist to the left, then the model initiates a saccade to  $n$  characters past the closest position to the right for which  $m(j) < \alpha$  (1c).<sup>4</sup> Finally, if no such positions exist to the right, the model stops reading (1d). Intuitively, then, the model reads by making a rightward sweep to bring its confidence in each character up to  $\alpha$ , but pauses to move left to reread any character whose confidence falls below  $\beta$ .

## Predictability and frequency in rational reading

The general findings about the effects of word predictability and frequency on eye movements in reading can be summarized relatively simply: words that are less predictable and lower frequency tend to receive more and longer fixations (Rayner, 1998, 2009). Here we describe intuitions for why our model should qualitatively reproduce these effects.

### Predictability

The basic intuition for why the model should produce effects of word predictability is very closely related to the reason for frequency effects in isolated word recognition reaction times given by Norris (2006, 2009). In short, the lower the prior probability of a word, the more visual input about it is needed to become confident in its identity. A bit more formally, this intuition is clearest if we make the simplifying assumption that prior to obtaining any visual information about a word, the model has near-veridical knowledge of the preceding context. In that case, the probability of the true identity of the word is given by the word’s predictability in context  $\pi$ . Visual input about the word will then (on average) increase the probability of the word’s true identity under the model’s beliefs. Recall that under our behavior policy, the eyes will remain in this position until the model’s confidence in the identity of the character at that position exceeds the threshold  $\alpha$ . Because information is being obtained about the entire word simultaneously, the probability of the identity of the fixated character is closely tied to the identity of the entire word. Specifically, the model’s confidence in the identity of the word gives a lower bound on the model’s confidence in the identity of a charac-

ter within that word. Thus, the initial probability of the true identity of the fixated character will start at or above the initial probability  $\pi$  of the true word, and – when the word is identified correctly – the model’s confidence about the identity of the fixated character is likely to reach the threshold  $\alpha$  near the same time that confidence about the identity of the fixated word reaches the threshold. As a consequence, the amount of visual input that is needed to reach the threshold which initiates a saccade is largely a function of the distance between  $\pi$  and  $\alpha$ . For more predictable words,  $\pi$  is closer to  $\alpha$ , so less visual input will be needed on average to reach  $\alpha$ , translating into shorter and fewer fixations on the word.

### Frequency

The most obvious intuition for the effect of frequency in the model is parasitic on the effect of predictability: words that are lower frequency are less predictable on average. Thus, as with words of higher predictability, there should be on average shorter and fewer fixations on words of high frequency.

## Simulation 1: full model

We now assess the effects of word predictability and frequency that the model does in fact produce. We use the model to simulate reading of a modified version of the Schilling corpus (Schilling, Rayner, & Chumblay, 1998) of typical sentences used in reading experiments. The arguments just described predict qualitatively that the model will make more and longer fixations on words of lower predictability and frequency. In addition, we quantitatively compare the model’s frequency effects to those of human readers of the Schilling corpus, which have been reported by Pollatsek et al. (2006).

### Methods

**Model implementation** We implemented our model with weighted finite-state automata (wFSAs) using the OpenFST library (Allauzen, Riley, Schalkwyk, Skut, & Mohri, 2007). While inference in the wFSA is exact, for efficiency we used Monte Carlo sampling from the wFSA to estimate the model’s confidence  $m$  in each character position.

**Model parameters and language model** We set the overall visual input quality  $\Lambda$  to 4. The model’s language knowledge was an unsmoothed bigram model created using a vocabulary set consisting of the 500 most frequent words in the British National Corpus (BNC) as well as all the words in our test corpus. From this vocabulary, we counted every bigram in the BNC for which both words were in vocabulary. Due to the intense computation required for exact inference, we then trimmed this set by removing rare bigrams that occur less than 200 times (except that we do not trim any bigrams that occur in our test corpus). This left a set of about 19,000 bigrams, from which we constructed the bigram model.

**Optimization of policy parameters** We define reading efficiency  $E$  to be an interpolation of speed and accuracy

$$E = (1 - \gamma)L - \gamma T$$

<sup>4</sup>The role of  $n$  is to ensure that the model does not center its visual field on the first uncertain character. For the present simulations, we did not attempt to optimize this parameter, but fixed  $n$  at 3.

where  $L$  is the log probability of the true identity of the text under the model's beliefs at the end of reading,  $T$  is the number of timesteps before the model stopped reading, and  $\gamma$  gives the relative value of speed. For the present simulations, we use  $\gamma = .05$ , which produces reasonably accurate reading. To find optimal values of the policy parameters  $\alpha$  and  $\beta$  for this definition of efficiency, we use the PEGASUS method (Ng & Jordan, 2000) to transform this stochastic optimization problem into a deterministic one on which we can use standard optimization algorithms. We then use coordinate ascent (in logit space) to find the optimal values of  $\alpha$  and  $\beta$ . This procedure resulted in optimal values  $\alpha = .88$  and  $\beta = .98$ .<sup>5</sup>

**Test corpus** To ensure that results did not depend on smoothing, we tested the model only on sentences from the Schilling corpus in which every bigram occurred in the BNC. Unfortunately, only 8 of the corpus sentences initially met this criterion, so we made single-word changes to 25 more (mostly proper names and rare nouns), producing a total of 33 sentences for which every bigram occurred in the BNC.

**Analysis** We used the model to perform 50 stochastic simulations of the reading of our modified version of the Schilling corpus. For each run, we calculated four standard eye movement measures for each word in the corpus: first fixation duration, gaze duration (defined to be the sum of all first pass fixations), skipping probability (whether or not word was directly fixated), and refixation probability (the probability of more than one first pass fixation). We then averaged each of these four measures for each word token in the corpus, yielding a single mean value for each measure for each word.

In order to compare the fixation duration measures to humans, we converted the model's timesteps into milliseconds. We performed this scaling by multiplying the duration of each fixation by a conversion factor set to be equal to the mean human gaze duration divided by the mean model gaze duration for the highest frequency bin. That is, we scaled the model predictions to exactly match the human mean for gaze durations in the highest frequency bin.

## Results

For each word in our modified version of the Schilling corpus, we defined its predictability to be its probability under the bigram language model, and we defined its frequency to be its overall probability in the data from which the bigram language model was constructed.

**Predictability** Figure 2 (red lines) shows the effect of predictability on the four aggregate measures. As predicted by both the intuition given above, and in agreement empirical human data, there are shorter fixations, more skipping, and

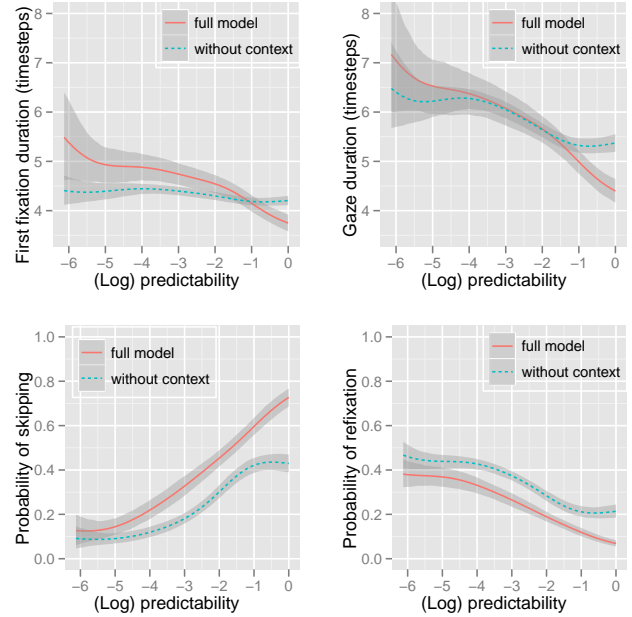


Figure 2: Effects of word predictability in both models on first fixation durations, gaze durations, the rate of skipping, and the rate of making a refixation, as estimated using Gaussian kernel regression with standard deviation equal to 1/8th of the range of log-predictability values. The 95% confidence intervals are bootstrapped from 1000 dataset replicates.

fewer refixations for more predictable words.

**Frequency** Figure 3 (red lines) shows the effects of frequency (binned by rounding down, to facilitate comparison to Pollatsek et al., 2006) on the four aggregate measures. The results across all four measures show a reasonable quantitative fit to the human data (blue lines). Further, comparing the overall size of the effect (i.e., the difference of the highest and lowest frequency bins) of the model to the human data shows a striking fit in effect direction and magnitude for all four measures. One unpredicted result here, however, is that the effect of frequency on the duration measures does not appear completely monotonic.

## Discussion

In summary, these results demonstrate that effects of predictability and frequency in the model's behavior resemble that of human readers in many respects. Predictability effects on all four aggregate measures are monotonic and in the same direction as predicted. Frequency effects on all four measures are in the same direction as predicted, and the total magnitude of the effect is quite similar to that displayed by human readers, despite the fact that we have not made any attempt to fit the human data, excepting only the scaling parameter that converts model timesteps to milliseconds. Overall quantitative fits on all four measures showed reasonable agreement to human data, but the fixation duration measures displayed

<sup>5</sup>It may at first seem puzzling that  $\alpha < \beta$ . However, this is a general property of optimal behavior for the model. While saccades to leave a character are initiated as soon as confidence  $m > \alpha$ , because of the saccade execution delay,  $m$  is usually substantially higher than  $\alpha$  when the eyes leave the character. Hence, it is a reasonable strategy for the threshold for regressions  $\beta$  to be accordingly higher. See also Bicknell and Levy (2010b) for further discussion.

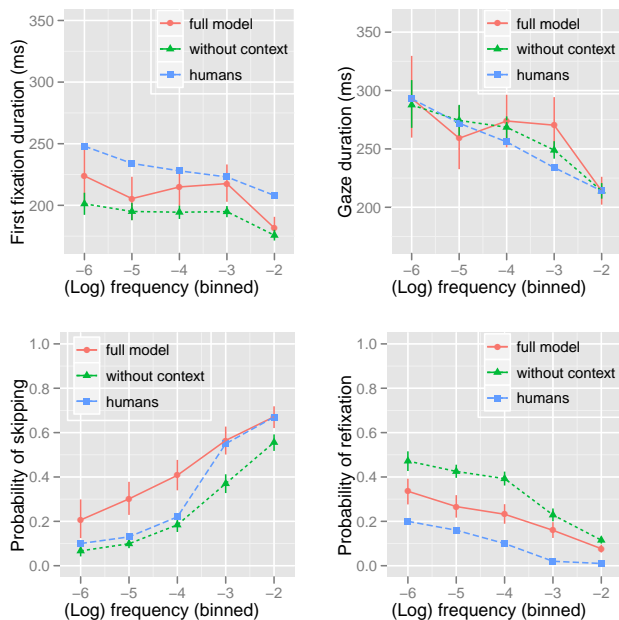


Figure 3: Effects of word frequency in both models on first fixation durations, gaze durations, the rate of skipping, and the rate of making a refixation. The 95% confidence intervals are bootstrapped from 10000 dataset replicates. Mean values from human readers of the Schilling corpus reported by Pollatsek et al. (2006) are shown for comparison.

some non-monotonicity.

It is perhaps unsurprising that predictability seems to have the most consistent effect, given the large role that predictability plays in the model, and the relatively straightforward predictions made previously. More surprising are the apparent non-monotonicities in the predictions for how the fixation duration measures should vary with respect to word frequency. One possibility is that these arise from our artificial removal of many low-frequency words from the language model, which may have meant that some of the low-frequency words in the Schilling corpus had artificially few visual neighbors, yielding an anti-frequency effect. The next simulation investigates this hypothesis.

### Simulation 2: Model without context

The main goal of Simulation 2 is to explore the possibility that removing low frequency words from the model’s vocabulary (which was necessary for computational efficiency) contributed to the non-monotonicities we observed in the effects of word frequency on fixation durations. Our strategy is to simplify the language model, which makes the computations faster to carry out, allowing for the use of a larger vocabulary. Specifically, we replace the previous bigram language model, which made use of linguistic context, with a unigram language model that includes only word frequency information and cannot make use of linguistic context. This simplified language knowledge also allows us to test how the model’s

predictions change when it can no longer make use of linguistic context to help recognize words.

## Methods

Except the following, the methods were identical to those of Simulation 1. We replaced the bigram language model with a unigram language model. Training was performed in the same manner, except that instead of including only the most common 500 words in the BNC, we included all words that occur at least 200 times (corresponding to a frequency of 2 per million; about 19,000 words). Finally, we increased the overall visual input quality  $\Lambda$  from 4 to 10. Because the new language model gives poorer information about the text, more visual input is needed to reach similar levels of confidence in word identities. Increasing the overall input quality to 10 results in the new model taking a similar number of timesteps to read a sentence as the previous model.

## Results and discussion

**Predictability** Figure 2 (green lines) shows the effect of predictability on the four aggregate measures for the model without context. Because the model does not make use of linguistic context in identifying words, any apparent effects of predictability must reflect effects of other variables correlated with predictability (e.g., frequency and length). We can then use these results as a baseline to determine the amount of the full model’s apparent predictability effect that was in fact driven by predictability. The results across all four measures show that predictability effects are smaller for this model without context, indicating that the full model’s use of context was important in producing its predictability effects.

**Frequency** Figure 3 (green lines) shows the effect of frequency on the four aggregate measures. Across all four measures, the size of the frequency effect in this model also shows a reasonable quantitative fit to human data, although the refixation rates and first fixation durations are about twice as far from human data as the full model. As with the full model, however, the direction and magnitude of all frequency effects is a close match to human data. The higher refixation rate and lower word skipping rate of this model relative to the full model likely reflect the model’s poorer language knowledge (cf. Bicknell & Levy, 2010a). Finally, and most importantly, we see that the problem of non-monotonicity is substantially reduced for first fixation durations and completely eliminated for gaze durations, supporting our argument that trimming the vocabulary may have been responsible for some of the non-monotonicity in the previous simulation results.

## General discussion

In this paper, we presented the first rational model of eye movement control in reading to make predictions for the entirety of the reading record. We gave intuitions for why it should produce effects of word predictability and frequency qualitatively similar to those produced by human readers, and presented two simulations empirically testing the effects of

these variables on model behavior. Simulation 1, using a full version of the model with parameters selected to maximize the agent's reading efficiency, demonstrated that the model yields effects of frequency and predictability that are qualitatively – and in frequency's case, quantitatively – similar to those of human readers, though the predictions for fixation durations on words of intermediate frequency did not appear completely monotonic. We hypothesized that these non-monotonicities may have been a result of the full model's small vocabulary, which had to be artificially limited for technical reasons. Simulation 2 tested this hypothesis using a model with simpler language knowledge but a larger vocabulary, and provided some evidence that alleviating this limitation helps to make the frequency effects more monotonic. In addition, by demonstrating that a model that cannot make use of predictability information shows smaller apparent predictability effects, Simulation 2 demonstrated that the predictability effects obtained for the full model were not likely to have been merely an artifact of the correlation between word predictability and other variables such as word length.

Taken together, these results demonstrate that the rational reading framework can produce reasonable effects of word predictability and frequency on four aggregate measures of eye movement behavior: first fixation durations, gaze durations, skip rates, and refixation rates. While the quantitative fit to human data is not perfect, the fact that it is such a good match is striking given that we fit no free parameters to human data, except the conversion of timesteps to milliseconds – a parameter that all timestep-based models must include. (In future work, determining the model's best possible fit to human data will require tuning the only two other truly free parameters of our model – the agent's value of speed relative to accuracy  $\gamma$  and the overall visual input quality  $\Lambda$ .) Instead of being selected to maximize the model's fit to human data, the policy parameters  $\alpha$  and  $\beta$  of our model were set to values that optimized the efficiency with which the model identified the text, given the agent's particular goal function. Future work must be done to explore the predictions of our model for a wider range of eye movement phenomena observed in reading, extending our analyses of the model's behavior both with more dependent measures, such as character landing positions within words and regressive saccades, and with more independent variables, such as word length.

## Acknowledgments

This research was supported by NIH Training Grant T32-DC000041 from the Center for Research in Language at UC San Diego to K. B. and by NSF grant 0953870 and NIH grant R01-HD065829, both to R. L.

## References

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., & Mohri, M. (2007). OpenFst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Ninth International Conference on Implementation and Application of Automata*, (CIAA 2007) (Vol. 4783, p. 11-23). Springer.

Bicknell, K., & Levy, R. (2010a). Rational eye movements in reading combining uncertainty about previous words with contextual probability. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1142–1147). Austin, TX: Cognitive Science Society.

Bicknell, K., & Levy, R. (2010b). A rational model of eye movement control in reading. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ACL* (pp. 1168–1178). Uppsala, Sweden: Association for Computational Linguistics.

Engbert, R., & Krügel, A. (2010). Readers use Bayesian estimation for eye movement control. *Psychological Science*, 21, 366–371.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, 112, 777–813.

Engel, G. R., Dougherty, W. G., & Brian Jones, G. (1973). Correlation and letter recognition. *Canadian Journal of Psychology*, 27, 317–326.

Geyer, L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, 22, 487–490.

Hyönä, J., Niemi, P., & Underwood, G. (1989). Reading long words embedded in sentences: Informativeness of word halves affects eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 142–152.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Legge, G. E., Hooen, T. A., Klitz, T. S., Mansfield, J. S., & Tjan, B. S. (2002). Mr. Chips 2002: new insights from an ideal-observer model of reading. *Vision Research*, 42, 2219–2234.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 21086–21090. (Correction in: *Proceedings of the National Academy of Sciences of the United States of America*, 107, 5260)

Ng, A. Y., & Jordan, M. (2000). PEGASUS: A policy search method for large MDPs and POMDPs. In *Uncertainty in Artificial Intelligence, Proceedings of the Sixteenth Conference* (pp. 406–415).

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, 113, 327–357.

Norris, D. (2009). Putting it all together: A unified account of word recognition and reaction-time distributions. *Psychological Review*, 116, 207–219.

Pollatsek, A., Perea, M., & Binder, K. S. (1999). The effects of “neighborhood size” in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 1142–1158.

Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the E-Z Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology*, 52, 1–56.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372–422.

Rayner, K. (2009). The 35th Sir Frederick Bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, 62, 1457–1506.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, 26, 445–526.

Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition*, 26, 1270–1281.

Slattery, T. J. (2009). Word misperception, the neighbor frequency effect, and the role of sentence context: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 1969–1975.



# The retriever-connector model: Matching Classroom Data and Agent-Based Computer Models to Simulate Students' Use of Multiple Epistemological Resources

Paulo Blikstein (paulob@stanford.edu)

TLT Lab, School of Education and (by courtesy) Computer Science Department, Stanford University  
520 Galvez Mall, CERAS 232, Stanford, CA 94305 USA

## Abstract

Utilizing data from a classroom intervention with 8<sup>th</sup> graders, I employ agent-based computer modeling to simulate the cognitive processes at play during the intervention, in which students transition between using multiple epistemological resources. The model substantiates the hypothesis of manifold epistemological resources, which can be activated with simple prompts and have a non-linear impact on learning.

**Keywords:** Cognitive modeling; agent-based modeling; classroom research; epistemological resources.

## Introduction

Agent-based modeling (ABM) has been increasingly used by scientists to study a wide range of phenomena such as the interactions of species in an ecosystem, the collisions of molecules in a chemical reaction, and the food-gathering behavior of ants (Bonabeau, Dorigo, & Theraulaz, 1999; Wilensky & Reisman, 2006). Such phenomena, in which the *agents* in a system (molecules, or ants) follow simple rules and interaction patterns, but exhibit complex emergent macroscopic behaviors, are studied in a young interdisciplinary field called complex systems or complexity studies (Holland, 1995). Although complex-systems perspectives initially arose from the natural sciences, complexity, emergence, and multi-level descriptions of phenomena are all highly relevant to social science research. In fact, recent decades have observed a surge in social-science studies employing ABM (Epstein & Axtell, 1996; Axelrod, 1997). Recently, ABM has also been used to illustrate aspects of cognitive development (Abrahamson & Wilensky, 2005; Blikstein, Abrahamson & Wilensky, 2006; Smith & Conrey, 2006), and collaboration and group work in classrooms (Abrahamson, Blikstein, & Wilensky, 2007).

ABM has the potential to advance theory in multiple ways, which I illustrate in this paper: (a) explicitizing—ABM demands a high level of specificity in expressing a theoretical model, and it provides the tools and standard practices to express those models; (b) dynamics—ABM enables researchers to mobilize an otherwise static list of conjectured behaviors and observe the macroscopic patterns that may unfold; (c) emergence—ABM can examine cognition and social behaviors as a collection of decentralized, simple rules; and (d) interdisciplinary collaboration—the *lingua franca* of ABM enables researchers from different fields to understand, critique and challenge each other's theories by modifying and extending the computational algorithms that underlie their theoretical models.

## Relevance to cognitive research

Agent-based modeling in cognitive research could address the limitations of current methodologies. First, because experiments with human subjects cannot be indefinitely conducted, replicating findings or exploring a wide parameter space is costly and oftentimes impossible. In the case of research in schools, once the classroom data are collected, the researchers can revisit the videotapes and transcriptions; however, they can never relive the situations. Second, as the field moves toward theories that conceptualize learning as a dynamic and adaptive phenomenon, the traditional medium of scientific discourse—static linear text—becomes limited in its capacity to express these theories. Both of these flaws could be addressed with a set of dynamic, adaptive computer models of learning. Third, tools such as brain imaging cannot yet offer the speed and resolution required to evaluate complex learning processes at a neuronal level, so such models are still far from being applicable to real classrooms. Lastly, ethnographic or micro-genetic methods still cannot offer a 'runnable,' systemic, task-independent account of human learning.

The ultimate goal of using agent-based simulation to explore human learning is to enable researchers to generalize and play "what-if" scenarios using in-depth interviews and ethnographic data and to help them investigate internal cognitive structures by observing external behaviors.

This work builds on previous seminal contributions to the field, in which theoretical models of cognition were implemented by using computer programs to attempt to predict human reasoning (Newell & Simon, 1972) in tasks such as shape classification (Hummel & Biederman, 1992), language acquisition (Goldman & Varna, 1995), memory (Anderson, Bothell, Lebiere, & Matessa, 1998); and other more general-purpose models (Anderson, 1983; Anderson & Bellezza, 1993; Anderson & Lebiere, 1998). My design, however, differs from extant approaches in two ways: (1) *Grain Size: Selecting a unit of analysis toward bridging the micro and macro perspective on learning.* Theories which slice human learning into diminutive pieces, when reintegrated into the larger context of classroom learning, could not account for any meaningful macro-cognitive phenomena, and (2) *Accessibility: Democratizing modeling-based research.* Most computational theories of the mind are so mathematically complex that only specialized researchers can examine and critique them; the intricacies

and jargon of these theoretical models render them incomprehensible for teachers, educators, and policymakers. Conversely, the computer language that I have used for modeling, NetLogo (Wilensky, 1999), has been developed for non-programmers so that users could not only run models but also modify their rules and compare scenarios.

My theoretical inspiration comes from the work of Minsky (1986), and Collins (1978). My computer-based models of human learning postulate non-intelligent cognitive entities with simple rules from which intelligent behavior emerges, or simple individual classroom behaviors that result in complex group-level patterns. To generate and validate such models, ABM tools enable researchers to initially feed a computer model with data from real-world experiments, such as classroom observations or clinical interviews and to subsequently simulate hypothesized scenarios in a safe virtual environment. Researchers from diverse disciplines (and with little, if any, programming background) can embody and articulate their theoretical models in a shared medium with shared nomenclature and shareable/replicable data, thus facilitating interdisciplinary discourse and critique.

However, the work described in this paper is not attempting to *reproduce reality*, which is oftentimes understood to be the goal of a computer model. My objective is to instantiate possible theories of learning in the agent-based form and use the data to qualitatively validate the models, with the goal of advancing theory. However, unlike classical cognitive models, this category of ABM models needs to be much more stylized and simple, as this paper will describe.

### Personal epistemologies & resources

Traditional research on personal epistemologies (Hofer & Pintrich, 2002) has considered them as stable beliefs. However, evidence of variability in student epistemologies suggests the need for more complex models (Hammer & Elby, 2002). The activation of the students' different epistemological resources might depend on context, as shown by Rosenberg, Hammer, & Phelan (2006). In other words, students might instantiate different epistemologies as they perceive contextual cues about the most efficient approach in a given situation. In the Rosenberg et al. case study, a brief epistemological intervention by an 8<sup>th</sup>-grade science teacher led to the students abruptly shifting from one epistemological mode to another. The narrative tells the story of a group of students who were given the task of explaining the rock cycle. For the first few minutes, before the teacher's intervention, they fail to engage in any productive work or to construct a coherent explanation of the rock cycle. Students employ a 'brute force' approach by quickly trying out several short explanations without evaluating if the elements of their explanations make sense together. They generate fragmented descriptions, which do not survive simple logical inference. Rosenberg et al. state that the reason is epistemological and that "They are treating knowledge as comprised of isolated, simple pieces of

information expressed with specific vocabulary and provided by authority." (Rosenberg, et al., 2006, pp. 270)

The authors provide three pieces of evidence for this hypothesis: (1) the students organize their efforts around retrieving information from worksheets, (2) they focus on terminology, and (3) they combine information and construct sentences to present a formal ordering rather than a causal sequence. The narrative goes on to describe how the teacher, realizing the ongoing failure, stops the activity and tells the students: "So, I want to start with what you know, not with what the paper says."

Abruptly, the students change their approach toward engaging in the activity. They immediately start to focus on the elements of the rock cycle that they understand, and they rebuild the story from there. Within minutes, one of the students comes up with a rather complete explanation:

"OK, the volcano erupts, and lava comes out. Lava cools and makes igneous rock. Rain and wind cause small pieces of rock to break off. Sediments form, and rain and wind carry it away, and rain and wind slow down and deposit sediments and this happens over and over again to form layers." (Rosenberg, et al., 2006, pp. 274)

It is impressive how the students, focusing on a single element of the story ("Lava comes out"), correctly connect all of the other pieces of the explanation. Although the "lava comes out" piece was the first to be mentioned, they realized that for lava to come out, the volcano has to erupt; similarly, if the lava comes out and is hot, it has to cool down. For the students to generate a coherent explanation, it was crucial for them to concatenate information while making sense of the connection rules, and they resorted to worksheets fewer times than in the previous activity.

In this paper, my goal is to employ ABM to help model what occurred during those 15 minutes and to answer two research questions concerning the abrupt epistemological shift observed: (1) what caused the two modes to generate very diverse student performance? and (2) how could a brief intervention cause such dramatic change? I built a model that simulates the construction of declarative knowledge in terms of two basic cognitive operations: retrieving information from external/internal sources and applying concatenation rules to join *content pieces*. I expect to answer the research questions by exploring the parameter space of the model for number, type, and efficiency of retrievers and connectors; this might result in emergent behaviors similar to those observed by Rosenberg et al. I warn the reader that the goal is to match an overall reference pattern based on simple theoretical assumptions about learning. The nature of ABM is such that this simplicity is required to generate a manageable parameter space.

### The Agent-Based Model

In the model, the world outside of the mind is composed of various disconnected *content pieces*, represented as green agents. A piece could be a simple statement, such as "Lava comes out of volcanoes," "Lava shoots up," or "Water erodes rocks." These pieces are retrieved by special agents,



called *retrievers* – represented as red agents – and accommodated into the simulated mind. Here, they interact with pre-existing structures until they connect to one of them, making use of a third type of agent, the *connectors*. These pre-existing structures could form an emergent, dynamic network with “hub ideas” (highly connected ideas) and peripheral ideas.

Therefore, the model consists of three independent elements: connectors, retrievers, and content pieces (line-shaped white agents, red agents, and green agents, respectively). Content pieces could be gathered by retrievers to simulate external information being exposed to the mind. Furthermore, the retrievers’ movement simulates the exposure and the searching for different content pieces over time. The recently retrieved information could be connected to preexisting mental structures to form a new knowledge “assembly.” This knowledge-construction process is represented by content pieces being connected to each other via connector agents. Accordingly, the students’ explanations are the ad hoc result of *pieces* of content collected by *retrievers* outside of the mind and assembled by *connectors* inside the mind. For the sake of the model’s simplicity, the internal and external processes occur in the same environment and have not been distinguished visually.

In the simulated world, the content pieces can have a different *stickiness* to the retrievers, thus the model can differently evaluate content from various sources. Content given from authority can have a different effect than previous knowledge in the virtual mind. In addition, retriever agents have been defined hypothesizing that content cannot simply enter the mind as raw information. Information needs to be retrieved and subsequently connected by internal knowledge structures (Piaget, 1952). Furthermore, retrieved content cannot be accessed until it is evaluated and internalized by connectors (Figure 1).

In the real classroom situation, students would assemble a textual explanation, such as: [*the volcano erupts + lava comes out + lava cools + lava makes igneous rock*]. In this model, textual explanation pieces are replaced with numbers, and a correct explanation is simply an ascending sequence of numbers (1, 2, 3, etc.), which reflects the nature of the task that students had at hand (sequencing information). I am aware of the limitations of the chosen approach, but this design principle was appropriate given the research questions and the target results.

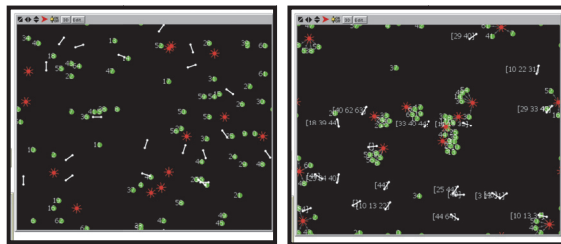


Figure 1. The model’s initial state (left) and after some steps (right), with ‘clusters’ of content form around retrievers.

## Experiments and Data

Contrary to most cognitive modeling software, this model does not attempt to simulate human thinking in its immense range of complexities and detail. I selected, conversely, the particular features of learning processes that will possibly enable me to pair the model’s data and the observations of Rosenberg et al. Because I am only modeling the agent’s ability to construct correct connections between pieces, I am ultimately investigating the computational cost and accuracy in building probabilistic cognitive structures. “Success” in the model is defined by the correct assemblage of a *sentence* without errors (i.e., with all numbers correctly places in an ascending order). I measure the *time to completion* of the sentences as well as the *error rate* in building them. The final performance measure is the ratio between the *average time to completion* and the *average error rate*, which I call the *cost of accuracy*.

### First experiment: Effect of Retrieval Skills

The first round of model runs compares retrievers with different performances or *stickinesses*. When retrievers encounter content pieces, they stick to them and carry them to the connectors. Low-performing retrievers, however, might collide with a piece but fail at sticking to it. The net effect of a low-performing retriever is to bring fewer pieces to the connectors per unit of time. The performance property of retrievers is loosely analogous to the students’ short-term memory skills or the amount of sheer information they can gather within the environment. One conclusion from this simple experiment is that retrievers have a small impact on overall task performance. Dropping retriever success rates from 90% to 20% results in a modest 16% increase in time for the completion of the task; therefore, within the initial parameters and assumptions of the model, retrievers appear not to be the *controlling phase* of the process. This is a key *qualitative* result of the model: good information retrieval skill does not cause abrupt gains in learning. It is important here to restate that my goal is not to present a calibrated computer model that would emulate precise response times of the human brain. Rather, I am advancing the understanding of a cognitive task by using computer modeling to explicitize certain assumptions and refine our understanding of the problem. In this case, for example, the simple experiment drew my attention to the fact that retrieving information should be relatively faster than connecting information to existing knowledge structures, and therefore, an improvement in the quantity of available information or the speed of retrieval would have a relatively low impact on performance. The data from Rosenberg et al. qualitatively corroborate this hypothesis: during the first narrative, with books and worksheets readily accessible but with weak *connecting skills*, students were unable to weave a coherent explanation. From the narrative, it is clear that if students were given more time or more informational resources to complete the task, the impact in task performance would *not* have been significant. In other words, my model replicates one of the classroom

observations of Rosenberg et al.: the controlling phase of the students' cognitive work was not *information retrieval*, and the cause of students' failure in explaining the rock cycle was not due to a lack of information, a lack of time to retrieve the correct information, or weak memorizing skills.

## Second experiment: Effect of Connecting Skills

The goal of the second experiment was to investigate the influence of the connectors' performance on overall task completion time and accuracy. Connectors, in the model, represent more elaborate cognitive agents, which can evaluate different pieces of information and link them based on a simple rule (build ascending sequences of numbers.) Connectors can make "mistakes," for example, wrongly appending the number 41 to the otherwise correct ascending sequence [3, 45, 67]. The probability of such mistakes is controlled by an internal variable within each connector agent (*connector-strength*). The following plots show the impact on time to completion, and accuracy, for different values of *connector strength* (from 10% to 95% of probability of a wrong connection).

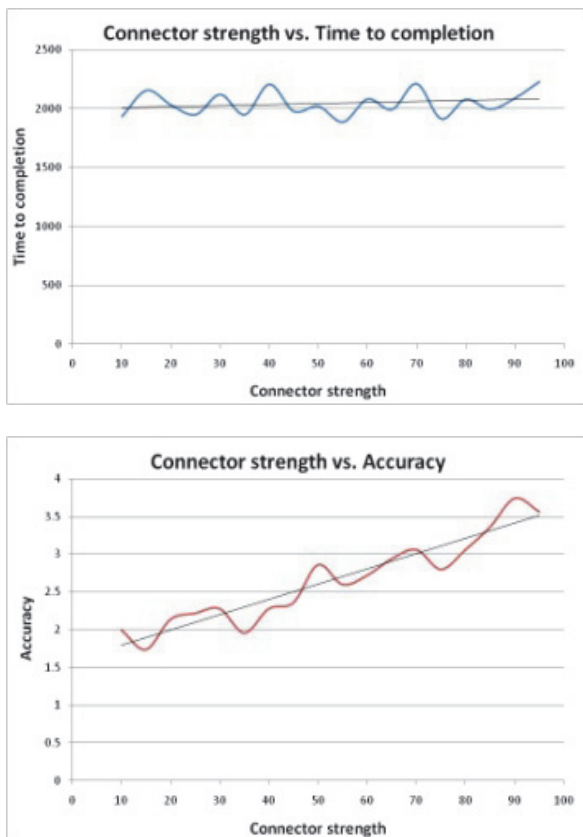


Figure 2. Connector strength, completion, and accuracy<sup>1</sup>.

<sup>1</sup> Note that each data point is an average of 50 model runs. Given the qualitative interpretation of the results and the limited space, I considered that error bars and more detailed statistics would not be informative to the research questions and would add unnecessary information.

At first sight, the *Connector strength vs. Time to Completion* plot (Figure 2, top) suggests that "Connector strength" has no impact on overall performance. However, even though the time to complete the task remains roughly the same, accuracy increases significantly (Figure 2, bottom). Combining the two plots (not shown) suggests a reasonable linear fit between computational cost of accuracy and connector strength, which suggests that increasing the *skill* of the connectors has a much greater impact on overall task performance than increasing the retrievers' skill (see previous experiment). This result confirms a second expected behavior, which is also qualitatively in agreement with the data from Rosenberg et al. When the students were told to "start from what they already knew" and evaluate the connections among the different phases of the rock cycle using previous knowledge (i.e., 'if lava is hot, it must cool down'), their performance increased significantly.

This second experiment hints that *connecting skills* are more significant for task performance than *retrieving skills*. However, the cost of training skilled connectors is still unknown; hence, comparing "unskilled but fast" and "skilled but slow" is crucial, which I attempt to illuminate in the next section.

## Third experiment: Explanation Complexity

The third experiment was aimed at discovering the impact of explanation complexity on performance. In this model, the complexity of the explanations is represented by the 'sentence-size,' which is the target number of *knowledge pieces* that the connectors need to put together (e.g., sentence size 3 would be "volcano erupts" + "lava comes out" + "lava cools"). The following plot shows a comparison between sentence sizes 2 and 3, for different values of connector strength.

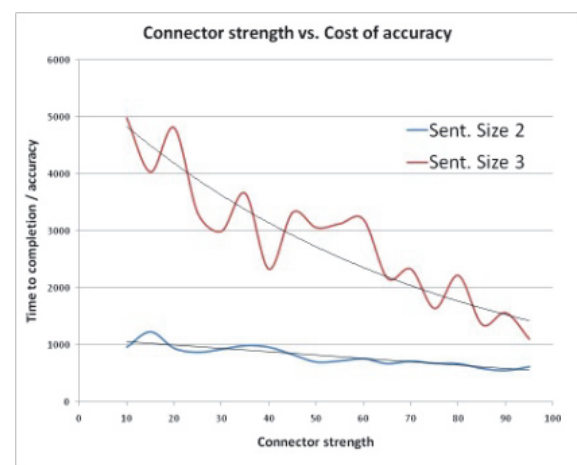


Figure 3. Time to completion divided by the correctness (y axis) and the connector accuracy (x axis.)

One result of this experiment is that while the impact of increasing values of connector strength is linear for sentence size 2, it is roughly exponential for sentence size 3. This

finding suggests that for assembling ‘simple’ content the gain that students obtain from improved connecting skills is much lower than when they are struggling with complex knowledge.

Again, this finding seems to fit with Rosenberg et al.’s narrative. Even in the first moment of the narrative, when students are trying assemble explanations based on worksheets using a brute-force approach (quickly trying many different pairs), they were able to assemble a number of “sentence-size 2” explanations, such as [igneous rock forms] + [weathering occurs]. However, in that first part of the narrative, the students were never able to form “sentence size 3” explanations, which would require extra steps: connecting that initial pair of pieces to a third piece and evaluating all possible pieces for their fit. In the second part of the narrative, after only a few minutes, by trying to expand their explanation *making sense* of the connections between pieces (and not using the brute force approach), students formed a sentence size 4 explanation, and a few minutes later they formed a sentence size 10 explanation:

“Bethany: Listen up! OK, the volcano erupts [1] and lava comes out [2]. Lava cools [3] and makes igneous rock [4]. Rain and wind cause small pieces of rock to break off [5]. Sediments form [6], and rain and wind carry it away [7], and rain and wind slow down and deposit sediments [8] and this happens over and over again to form layers [9]. OK, so water is added to this [10]...” Rosenberg, Hammer, & Phelan (2006), pp. 274

To further investigate the role of the increase in sentence sizes to the overall cost of accuracy, I ran the model for sentence size 4 as well. The results, comparing sizes 2, 3 and 4, are in the following two plots:

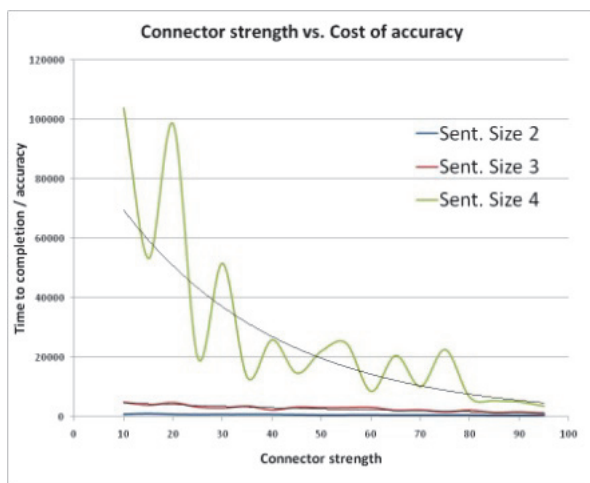


Figure 4. Time to completion divided by the correctness (y axis) and the connector accuracy (x axis.)

To understand Figure 4, it is important to comprehend the intuition behind the results. Essentially, I am comparing a “brute force” versus a “smart” approach for assembling sequences of different sizes. For sentence size 4 (SS4), with low values of connector strength (CS), it is virtually

impossible to assemble a correct explanation (see the very high values of the top curve). For CS 10%, increasing SS from 2 to 4, the accuracy drops by a factor of 100. Increasing SS from 2 to 3, the accuracy drops 5 times. Figure 4, therefore, shows that increasing sentence sizes has a dramatic impact on performance. The important finding here is that this differently impacts “long” and “short” explanations. For SS 2, brute force assemblage is not so costly and works relatively well, so there would be no benefit for developing connecting skills. However, for SS 3 and 4, this ‘brute force’ (low CS) assemblage breaks down.

The events in Rosenberg et al. narrative tell a similar story. In the first half of the class, when students were using brute force methods instead of their own connecting skills, they could not go much further than assembling simple, “SS 2,” explanations. When they activated their ‘connectors,’ prompted by the teacher’s intervention, they switched from a brute force to a “sense-making” mode, in which most of their energy was spent connecting pieces instead of retrieving and randomly connecting them. That shift enabled them to assemble seamlessly explanations of SS = 10.

## Conclusion, limitations, and implications

Throughout this paper, I tried to pair computer model data with real classroom data. In the three experiments, I searched for instances that would resemble what Rosenberg et al. described in their classroom observations. The model seems to validate key elements of those observations:

1) The students’ failure in the first half of the narrative was epistemological (i.e., resulting from a particular approach toward learning) and not due to a lack of memorization or information retrieving skills (the first experiment).

2) The fundamental mathematical basis of the model, from which all other behaviors emerge, is that *brute-force methods are efficient for short sequences, but for long sequences, as the combinatorial space greatly increases, their performance drops accordingly*. In the high connector strength mode, the size of the sentence has a much lesser impact because of the evaluative rule of the connector: any connection will take the exact same computational time for any sentence size. This seems to be the case in the classroom, where the students could assemble long explanations quickly once they were in a ‘high connector strength’ mode.

3) In this simulated environment, I was able to verify that for learning intricate content (here, I equate that to assembling long explanations), there is a significant non-linear payoff from investing in sense-making skills, (connector strength) as opposed to memorizing skills (retrieving speed). For simple content (involving the connection of two content pieces), however, *sheer memorization can even outperform sense-making skills*. The data show that the payoff of improved connector strength only manifests itself after CS 80% (Figure 2, 3, 4).

4) Abrupt, non-linear shifts in student understanding are indeed possible, even within very short periods of time, by

activating different cognitive resources and different epistemological modes.

### Limitations and implication for design

The classroom data used in this paper was chosen because it described a relatively uniform macroscopic behavior that clearly derived from a change in simple, local rules. I acknowledge that many other typical classroom interventions might not exhibit such a uniform behavior. The goal of this model and paper, however, was not to match a computer model to a precise mechanism in the brain. Rather, my goal was to produce the “simplest possible” model that would exhibit the observed behaviors and generate further insight into the research questions. In that sense, this was a theoretical exercise made possible by formalizing the problem as agent rules. Therefore, given the assumptions of the model, I suggest that some possibly overlooked elements in classroom implementation might be more important than one would suspect: (1) the radically different payoffs for improving the *speed of retrieval* versus *sensemaking*, and the determination of which is the controlling phase in the learning process in different scenarios, (2) the non-linear impact of *sentence-sizes* on performance and accuracy, (3) the unexpected success of “brute force” methods for small sentences.

Given the limited space, it is impossible to go into detail about all possible implications, but one implication is very significant. *In earlier grades, exposed to simpler content, students might learn that brute-force methods ‘work.’ In later grades, they might insist on using this method, which would break down because the content is more complex.*

The computational task is of course an approximation of a real classroom task, and as with any model, it can only capture a portion of the real-world complexity. However, my goal here was to demonstrate the potential of agent-based models as a powerful and useful formalism for cognitive theory. This work could potentially have implications for the practice of curricular designers, teachers, and policy makers by offering researchers accessible, transparent tools to simulate, model and test hypotheses about human cognition in social contexts and to pair model data with real classroom data.

### Acknowledgements

Thanks to Uri Wilensky, Dor Abrahamson and Shima Salehi for their input on previous versions of this work, and David Hammer and colleagues for the classroom data.

### References

- Abrahamson, D., & Wilensky, U. (2005). *Piaget? Vygotsky? I'm Game: Agent-Based Modeling for Psychology Research*. Paper presented at the Annual Meeting of the Jean Piaget Society, Vancouver, Canada.
- Abrahamson, D., Blikstein, P., & Wilensky, U. (2007). Classroom model, model classroom: Computer-supported methodology for investigating collaborative-learning pedagogy. *Proceedings of the Computer Supported Collaborative Learning (CSCCL) Conference*, pp. 46 - 55.
- Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.
- Anderson, J. R., & Bellezza, F. S. (1993). *Rules of the mind*. Hillsdale, N.J.: L. Erlbaum Associates.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, N.J.: Lawrence Erlbaum Associates.
- Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*(38), 341-380.
- Axelrod, R. M. (1997). *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton, NJ: Princeton University Press.
- Blikstein, P., Abrahamson, D., & Wilensky, U. (2006). *Minsky, Mind, and Models: Juxtaposing Agent-Based Computer Simulations and Clinical-Interview Data as a Methodology for Investigating Cognitive-Developmental Theory*. Paper presented at the Jean Piaget Society Annual Meeting, Baltimore, USA.
- Blikstein, P., Abrahamson, D., & Wilensky, U. (2009). *Toward a Framework for Cognitive Research Using Agent-Based Modeling and Complexity Sciences*. Paper presented at the American Educational Research Association Annual Meeting, San Diego, CA, USA.
- Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm Intelligence: from natural to artificial systems*. London: Oxford University Press.
- Collins, A. M. (1978). *Fragments of a theory of human plausible reasoning*. Urbana-Champaign, Illinois: ACL.
- Epstein, J., & Axtell, R. (1996). *Growing artificial societies: Social science from the bottom up*. Washington: BIP.
- Goldman, S. R., & Varma, S. (1995). CAPing the construction-integration model of discourse comprehension. In *Discourse Comprehension: Essays in Honor of Walter Kintsch*. Erlbaum.
- Hammer, D., & Elby, A. (2002). On the form of a personal epistemology. In B.K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 169–190). Mahwah, NJ: Lawrence Erlbaum.
- Hofer, B. K., & Pintrich, P. R. (Eds.). (2002). *Personal epistemology: the psychology of beliefs about knowledge and knowing*. Mahwah, NJ: Lawrence Erlbaum.
- Holland, J. (1995). *Hidden order: How adaptation builds complexity*. Reading, MA: Helix Books/ Addison-Wesley.
- Hummel, J. E., & Biederman, I. (1992, Jul 1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, 3(99), 480-517.
- Minsky, M. L. (1986). *The society of mind*. New York, N.Y.
- Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Piaget, J. (1952). *The Origins of Intelligence in Children*. New York, NY: International University Press.
- Rosenberg, S., Hammer, D., & Phelan, J. (2006) Multiple Epistemological Coherences in an Eighth-Grade Discussion of the Rock Cycle. *J. of the Learning Sciences*, 15(2), pp. 261-292.
- Smith, E. R., & Conrey, F. R. (2007). Agent-Based Modeling: A New Approach for Theory Building in Social Psychology. *Personality and Social Psychology Review*, 11(1), 87-104.
- Wilensky, U. (1999). NetLogo. Evanston, IL: Center for Connected Learning and Computer-Based Modeling. <http://ccl.northwestern.edu/netlogo>.
- Wilensky, U., & Reisman, K. (2006). Thinking like a Wolf, a Sheep or a Firefly: Learning Biology through Constructing and Testing Computational Theories. *Cognition & Instruction*, 24(2), pp. 171-209.

# Inducing Mathematical Concepts from Specific Examples: The Role of Schema-Level Variation

David W. Braithwaite (dwbraith@indiana.edu)

Robert L. Goldstone (rgoldsto@indiana.edu)

Indiana University, 1101 E. 10<sup>th</sup> Street  
Bloomington, IN 47405 USA

## Abstract

Previous research suggests that comparing multiple specific examples of a general concept can promote knowledge transfer. The present study investigated whether this approach could be made more effective by systematic variation in the semantic content of the specific examples. Participants received instruction in a mathematical concept in the context of several examples, which instantiated either a single semantic schema (non-varied condition) or two different schemas (varied condition). Schema-level variation during instruction led to better knowledge transfer, as predicted. However, this advantage was limited to participants with relatively high performance before instruction. Variation also improved participants' ability to describe the target concept in abstract terms. Surprisingly, however, this ability was not associated with successful knowledge transfer.

**Keywords:** mathematics; analogy; comparison; schemas; instruction; transfer

## Introduction

Part of the power of mathematics lies in its generality. The same mathematical formulae may be used to understand the growth of slime molds or the accumulation of interest from investments, the probabilities of hands in poker or outcomes of scientific experiments, and the oscillations of mechanical or electromagnetic systems. In order to fully realize this power, however, learners must be able to recognize and apply mathematical concepts in contexts different from those in which they were learned – that is, to transfer their mathematical knowledge from learned to novel contexts.

Learners' difficulties in achieving such transfer are well-documented (Novick & Holyoak, 1991; Ross, 1987). One reason may be that, when a general idea is learned in the context of specific examples, learners' concepts become tied to the details of the examples, inhibiting their ability to recall the concept or apply it correctly when faced with cases that do not share similar details (Ross, 1987). This difficulty may be especially strong when the examples are presented in a perceptually detailed format (Kaminski, Sloutsky, & Heckler, 2008), and is likely to be more serious for domain novices than experts (Novick & Holyoak, 1991).

One way to address this difficulty is to present mathematical ideas in abstract form, without specific examples. Such an approach has indeed been shown to promote transfer in some cases (Kaminski et al., 2008). However, in other cases, learners have experienced serious difficulties with abstractly-presented mathematics, despite being competent with the same mathematics encountered in

familiar contexts (Nuñez, Schliemann, & Carraher, 1993). In such contexts, learners can apply intuitions from everyday life to help in understanding the mathematical ideas involved. Abstract presentation of mathematical ideas therefore risks sacrificing learning for the sake of transfer.

It may, then, be desirable for learners to encounter mathematical ideas in a way that leverages their intuitive understanding of specific examples, while also drawing attention to the abstract structure present in those examples. Research on analogy suggests that this goal might be achieved through presentation of multiple specific examples followed by comparison (Gentner, Loewenstein, & Thompson, 2003; Gick & Holyoak, 1983). Comparing examples encourages learners to align their corresponding elements, and thereby to notice their common relational structure. Awareness of this structure, in turn, can facilitate understanding of new cases with the same structure. Thus, learning mathematical ideas by studying and then comparing multiple examples may enable learners to gain intuitive accessibility without losing generality.

The question then arises as to how the examples which will instantiate a mathematical concept during learning are to be chosen. Central to this question is the issue of how much, and in what ways, the examples should differ from each other. If, as the above research suggests, learners induce concepts that incorporate commonalities among the examples, it seems desirable that the examples should share the mathematical structure in question, but should not share other extraneous details. Extraneous commonalities might be misunderstood as part of the concept to be learned, limiting learners' ability to generalize (Medin & Ross, 1989), and so defeating the purpose of using multiple examples in the first place. These observations suggest that extraneous aspects should be systematically varied across examples, while holding mathematical structure constant.

The present study investigates the effects on mathematical concept learning of a particular type of variation among examples: variation at the level of "semantic schemas." This term here refers to structures more general than specific examples but less general than mathematical structure. Consider the three combinatorics problems shown in Figure 1. Problems (a) and (b) share a schema, termed "Objects Selected in Sequence" (OSS), in which a sequence of selections is made from a fixed set of options. Problem (c), by contrast, belongs to a different schema, termed "People Choosing Options" (PCO), in which several people each choose once from a fixed set of options.



(a) A piano student, when bored, plays random sequences of notes on the piano, using sequences of a fixed length, and choosing from a fixed set of notes. How many different sequences are possible, if there are 5 possible notes and the sequences are 6 notes long?	(b) A website generates user passwords by selecting a certain number of characters randomly from a fixed set of characters. How many different passwords are possible, if the passwords are 6 characters long and there are 5 permissible characters?	(c) A marketing research company conducts a taste test survey. Several consumers are each asked to choose their favorite from among several pizza flavors. How many different results of the survey are possible, if there are 6 consumers and 5 pizza flavors?
---	--	--

Figure 1. Three combinatorics problems.

Of course, all three problems share the same mathematical structure (discussed further in the Methods section), and the differences between them would likely not seem important to a mathematics expert. For mathematics novices, however, semantic schemas are known to exert a strong influence on the mathematical interpretation of contextualized problems. For example, Bassok, Wu, and Olseth (1995) found that learners were more likely to solve correctly problems in which schematic and mathematical roles were matched consistently with their default expectations than problems in which such matches were inconsistent. In light of the preceding discussion, learning about a mathematical structure via several examples based on the same schema might lead learners to induce concepts tied to that particular schema, and thus to perform poorly on problems involving other schemas. Conversely, systematic variation of the schemas encountered during learning should lead to induction of more general concepts and thus to more successful transfer to novel problems.

This hypothesis was investigated in the present study. Combinatorics problems were used as the domain for study and transfer for several reasons. First, the discovery of better methods for learning and teaching combinatorics would have considerable practical value due to the foundational role of combinatorics in applied mathematics – in particular, probability and statistics. Second, mathematics learners are known to have considerable difficulty correctly applying combinatorics methods to novel problems (Bassok et al., 1995; Ross, 1987). Finally, semantic schemas are known to play a role in the mathematical interpretation of combinatorics problems (Bassok et al., 1995).

## Methods

### Participants

Participants were 109 Indiana University undergraduate students, who participated in partial fulfillment of a course requirement.

## Materials

Sixteen story problems were constructed as stimuli. All of the problems had the same mathematical structure: Sampling with Replacement (SWR), in which multiple selections are made from a fixed set. The number of possible joint outcomes in such a case is given by the expression  $m^n$ , where  $m$  is the number of elements of the set and  $n$  is the number of selections, or sampling events.

The sixteen problems belonged to four different schema categories. The first two categories were those already illustrated above: PCO and OSS (OSS: Figure 1a-b, PCO: Figure 1c). Problems in these categories were used as learning examples. The other two categories were Options Assigned to Places (OAPlc) and Objects Assigned to People (OAPpl), illustrated below (Figures 2a and 2b respectively). OAPlc and OAPpl problems served as pretest and transfer problems. Note that in the learning examples (OSS and PCO) and OAPlc problems, people are either doing the choosing or are not mentioned at all. In OAPpl, by contrast, people are being chosen instead of choosing. Due to this role reversal relative to the learning examples, transfer to OAPpl problems was expected to be particularly difficult, as found in previous research (Ross, 1987).

(a) A homeowner is going to repaint several rooms in her house. She chooses one color of paint for each of the rooms. In how many different ways can she paint the rooms, if there are 3 colors and 5 rooms?	(b) A prize drawing is held at a small office party, and each of several prizes is awarded to one of the employees. In how many different ways can the prizes be awarded, if there are 6 prizes and 4 employees?
---	---

Figure 2. Combinatorics problems from the (a) OAPlc and (b) OAPpl categories.

Each problem category contained two pairs of problems, for a total of four problems. The problems within a pair involved the same back story but different numbers, while the two pairs within each category involved different back stories (and different numbers from each other). The order in which the two critical numbers, i.e. the size of the sampled set and the number of sampling events, were presented was varied among questions so that it could not serve as a cue to match the numbers to their respective roles.

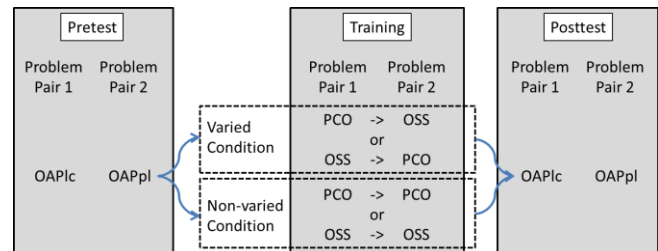


Figure 3. Summary of experimental design.

The experiment employed a pretest-training-posttest design, summarized in Figure 3. The pretest consisted of one OAPlc problem pair and one OAPpl problem pair, for four problems altogether. The posttest consisted of the other OAPlc problem pair followed by the other OAPpl problem pair. Thus, all eight OAPlc and OAPpl problems appeared in either the pretest or the posttest.

The training consisted of worked solutions to four problems drawn from the PCO and OSS categories. Participants were assigned randomly to one of two training conditions. In the varied condition, participants were shown one pair of problems from each category, either PCO followed by OSS or vice versa (these two possible orders were balanced across participants). In the non-varied condition, participants were shown two pairs of problems from the same category, either both PCO or both OSS (again, the two possibilities were balanced across participants). If a certain problem category was shown in a given position (either first pair or second pair), it was always the same problem pair regardless of condition. For example, if PCO problems were shown first in the varied condition, they were the same problems that were shown first in the non-varied condition. An important consequence of this design is that each training problem was shown equally often across the two conditions.

## Procedure

Participants were randomly assigned to receive one set of OAPlc / OAPpl problems as pretest. The pretest problems were displayed to participants on a computer monitor together with a virtual calculator, which participants were encouraged to use as needed. Only one problem appeared on the screen at a time. Two spaces were provided below each problem: one in which to show work, and another in which to write the final answer. Participants were required to show their work and enter some number as their final answer before they could proceed to the next question.

After the pretest, answers were scored for correctness, and participants were classified as high pretest performers if they answered at least 50% of the pretest problems correctly and low pretest performers otherwise. They were then assigned randomly to one of the two training conditions with the constraint that, at each level of pretest performance, the number of participants in each condition was balanced. This manipulation was intended to reduce differences in pretest scores between training conditions.

The training problems corresponding to participant's training conditions were then presented in the same way as the pretest problems. However, after completing each problem, participants were shown the correct answer together with a brief explanation of how the answer was calculated and why this calculation was appropriate. These explanations utilized exponential notation but did not show the general expression  $m^n$ . Instead, they only showed specific versions of this expression instantiated with the numbers used in the problem. The explanation for a given problem did not differ between training conditions.

After completing each pair of training problems, participants were asked to choose from a list of options the correct method of solving problems like those just seen, independent of the specific numbers involved. For example, the correct answer to this question after the problems involving pizza flavors (Figure 1c above) was "Multiply the number of pizza flavors by itself as many times as there are consumers." Participants who chose incorrectly were not allowed to proceed until they chose the correct answer.

After answering the above question for the *second* pair of training problems (only), participants were asked to choose from a list of options the correct mapping between elements of the preceding two problem pairs. For example, the correct answer to this question if the preceding problem pairs involved a website generating passwords and consumers tasting pizza flavors (Figure 1b and 1c) was "The length of the note sequences corresponds to the number of consumers, and the number of possible notes corresponds to the number of pizza flavors." The purpose of this question was to encourage participants to think about the shared structure of the training problem pairs. After answering this question, participants were asked to describe, in free-response format, a general method for solving problems like those just seen. No feedback was given for either of these questions.

Finally, participants were administered the posttest. The posttest utilized whichever set of OAPlc / OAPpl problems had not been presented during the pretest, and the procedure was in all ways the same as for the pretest.

## Coding

For each problem, participants were assigned a score of 1 if their answer was correct and 0 otherwise.

Responses to the free-response question regarding a general solution method posed at the end of the training were coded on a 0-2 scale in each of two respects. For the first respect, Correctness, responses were assigned a score of 2 if they indicated that the number of elements in the sampled set should be raised to the power of the number of sampling events (or multiplied by itself as many times as the latter). Responses which implicated exponentiation but did not correctly identify the base and exponent were assigned a score of 1, and all other responses received a score of 0. The second respect, Abstractness, was intended to measure how well participants had generalized beyond the specific details of the learning examples. Responses were assigned a score of 2 if they referred to the two numbers using general words, such as "the options" (for the size of the sampled set) or "the number of times they are able to be chosen" (for the number of sampling events). Responses which used general words for one but not the other number were assigned a score of 1, and all other responses received a score of 0. All responses were coded by two independent coders, and all disagreements were resolved through discussion. In the analyses detailed below, scores of 0 and 1 were combined for both correctness and abstractness, so that responses were classified as either correct (2) or not correct (0 or 1) and abstract (2) or not abstract (0 or 1).

## Results

Average pretest and posttest scores are shown in Figure 4. Participants demonstrated considerable improvement on posttest, but the amount of improvement varied by problem category. The data were entered into a 2 (test section: pretest or posttest)  $\times$  2 (problem category: OAPlc or OAPpl) within-subjects ANOVA. The main effects of both factors and the interaction between them were all significant (test section:  $F(1,108)=69.8$ ,  $p<.001$ ; problem category:  $F(1,108)=14.6$ ,  $p<.001$ ; interaction:  $F(1,108)=16.4$ ,  $p<.001$ ). Participants improved from pretest (0.216) to posttest (0.489), but this improvement was greater for OAPlc (0.225 to 0.638) than for OAPpl (0.206 to 0.339).

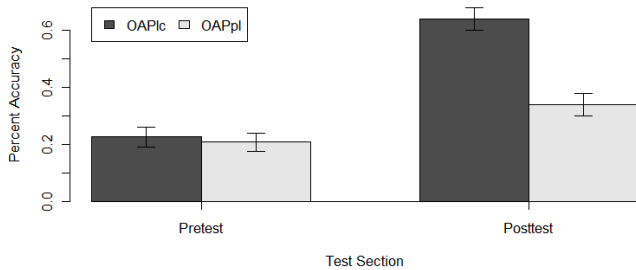


Figure 4. Pre and posttest accuracy by problem category<sup>1</sup>.

Figure 5 shows average transfer scores, defined as the difference between posttest and pretest scores, for each training condition, among low and high pretest performers. Transfer scores were submitted to a 2 $\times$ 2 $\times$ 2 mixed ANOVA with training condition (varied vs. non-varied) and pretest performance (low or high) as between-subjects factors and problem category (OAPlc or OAPpl) as a within-subjects factor. The main effect of pretest performance was significant,  $F(1,105)=66.6$ ,  $p<.001$ , indicating more improvement from pretest to posttest among low pretest performers (0.404) than high pretest performers (-0.056). Also, the effect of problem category was significant,  $F(1,105)=12.3$ ,  $p=.001$ , indicating greater improvement on OAPlc (0.413) than on OAPpl (0.133). Problem category did not interact significantly with any of the other factors.

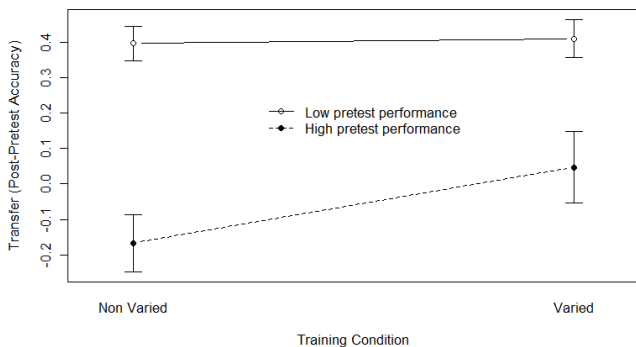


Figure 5. Transfer by condition and pretest performance.

More importantly, the main effect of training condition was significant,  $F(1,105)=4.0$ ,  $p=.049$ , indicating greater improvement in the varied (0.305) than in the non-varied (0.201) condition. However, this effect was qualified by a marginally significant condition by pretest performance interaction,  $F(1,105)=3.1$ ,  $p=.08$ . Consequently, the same model (excluding the pretest performance factor) was applied separately to the data from low and high pretest performers. This analysis found a significant effect of training condition among high performers,  $F(1,29)=.706$ ,  $p=.022$ , indicating higher transfer in the varied condition (0.047) than in the non-varied condition (-0.167), but no effect of training condition among low performers,  $F(1,76)=.042$ ,  $p=.838$  (varied: 0.410, non-varied: 0.397).

In addition to the effect of training condition on transfer, we were also interested in whether training condition affected participants' ability to induce a general method for solving SWR problems. The proportion of participants providing correct and abstract solution descriptions (i.e. receiving scores of 2 on the correctness and abstractness scales) within each training condition are shown in Figure 6. In the varied condition, 40% of participants' solutions were scored as correct, 62% as abstract, and 29% as both correct and abstract. In the non-varied condition, 56% of participants' solutions were scored as correct, 39% as abstract, and 20% as both correct and abstract.

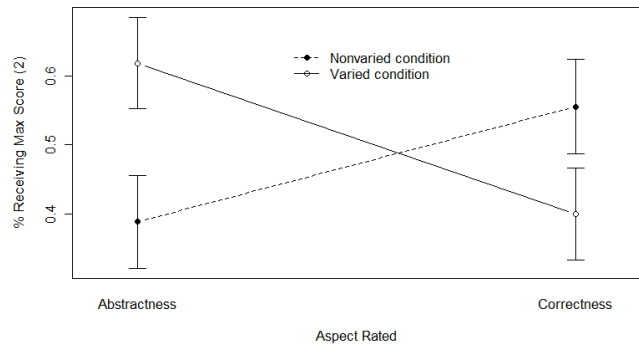


Figure 6. Percent generating correct or abstract general solutions by training condition.

The Breslow-Day test, a non-parametric test for stratified analysis of 2 $\times$ 2 tables, was applied to the frequencies of best (2) and other (0-1) scores within each training condition (varied or non-varied) for each aspect rated (correctness or abstractness). The relative frequencies of best vs. other scores between training conditions differed significantly according to aspect rated,  $p=.004$ . In other words, the effectiveness of varied relative to non-varied training was greater with respect to abstractness than with respect to correctness. To further clarify this effect, Pearson's Chi-square tests were applied to the contingency tables of best vs. other scores by training condition separately for each measurement respect. These analyses found that abstract solutions were more common in the varied than in the non-

<sup>1</sup> Here and elsewhere, error bars indicate standard errors.



varied condition,  $p=.028$ , but the proportion of correct solutions did not differ by training condition,  $p=.152$ .

Were participants who provided solutions that were abstract, correct, or both more likely to perform well on posttest? Average posttest scores among participants displaying each combination of solution abstractness and correctness are shown in Figure 7. (Participants were approximately equally distributed over these combinations.) Scores were virtually identical for each of these combinations: 0.50 for both correct and abstract, 0.49 for neither abstract nor correct, 0.48 for abstract but not correct, and 0.48 for correct but not abstract. A mixed ANOVA applied to posttest scores with solution correctness (correct or not), solution abstractness (abstract or not), pretest performance, and training condition as between-subjects factors and problem category as a within-subjects factor found no significant main effects of solution correctness or abstractness, no significant interaction between them, and no significant interaction of either or both with any other factor. (None of these effects were significant when transfer rather than posttest scores were entered into the model.)

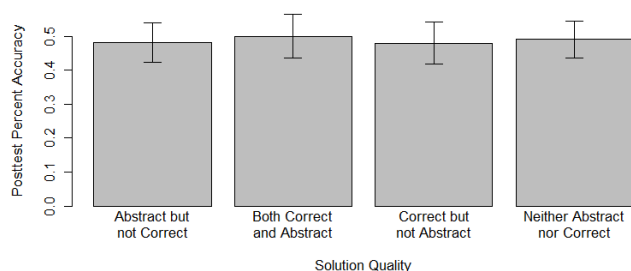


Figure 7. Average transfer scores by correctness and abstractness of generated solution and test problem pair.

## Discussion

This experiment investigated whether exposure to multiple examples of an abstract mathematical concept followed by comparison among them would lead to better induction of the general concept when the semantic schemas of the examples were systematically varied during learning than when all examples were based on the same schema. As predicted, participants in the varied condition both induced more abstract solution methods for SWR problems, and showed greater improvement on a transfer test requiring them to apply such methods. These results suggest that schema-level variation of examples can be an effective way to promote transfer.

Caution is necessary in interpreting these results because the advantage of the varied over the non-varied condition in promoting transfer was almost entirely driven by high pretest performers. Low pretest performers did not benefit from the varied condition, although they were not hurt by it either. A possible reason is that the dissimilarity between examples in the varied condition made it difficult to notice their shared structure. This difficulty might be overcome by presenting several examples from the same schema, thus facilitating comparison and alignment of the examples,

before introducing schema-level variation. Consistent with this view, Kotovsky and Gentner (1996) found that children initially presented with several examples sharing both abstract structure and superficial details were later able to notice shared structure even in the absence of superficial similarity. Similarly, Elio and Anderson (1984) found that category learning was better after a learning schedule beginning with low variation among exemplars and later progressing to more variation, as opposed to one beginning with and maintaining a high level of variability.

Interestingly, Elio and Anderson (1984) also found that when learners were specifically instructed to take an analytical approach to category learning, the effectiveness of training with initially high variability improved. Similarly, high pretest performers in the present study, who may have been better equipped to take an analytical approach to learning the SWR concept, derived greater benefits from varied relative to non-varied training. One account for this result is that good learners are more attentive to the features and relations that are relevant to domain principles. Consequently, good learners would be less likely to be distracted by – and more likely to benefit from – variation in extraneous features and relations. Considering this conclusion together with the previous one regarding weaker learners, the best instructional approach might be an adaptive one, beginning with examples drawn from a single schema and transitioning to schema-level variation once learners demonstrate understanding of the target concept in the context of the initial schema. This interesting possibility deserves further investigation.

However, the observed advantage of the varied training for high pretest performers must also be interpreted with caution. Transfer scores among high pretest performers were rather low, averaging around zero in the varied condition and below zero in the non-varied condition. One interpretation of these data is that varied training merely helped to avoid negative transfer, and did not actually benefit learners. On the other hand, high pretest performers might be expected to show regression to the mean on posttest, resulting in negative scores on our measure of transfer. In this case, the actual (slightly above zero) transfer scores in the varied condition would represent a positive effect of training. It is difficult to disambiguate between these possibilities due to the lack of a control condition in the present study. Also, the inclusion of particularly difficult transfer problems, i.e. those in the OAPpl category, may have obscured the presence of positive transfer by bringing down the overall average. The beneficial effects of schema-level variation might be better explored in future studies by using a wider range of relatively easy transfer problems.

In addition to their differing effects on transfer, the varied and non-varied training conditions also led to differing levels of success in describing general solutions for SWR problems. In particular, while participants in both conditions were equally able to describe correct solutions, those in the varied condition were better able to characterize the elements of those solutions in abstract, general terms.

Previous research has demonstrated that comparison between multiple analogous examples can lead participants to induce their shared abstract structure (Gentner et al., 2003; Gick & Holyoak, 1983). The present findings build on that principle by suggesting that if the examples in question share semantic content not intrinsic to the desired structure, learners may induce a more limited, less general concept than if such extraneous semantic content is systematically varied across learning examples. Moreover, not only superficial elements but also more abstract semantic structures, such as the schemas of the present study, can count as extraneous content in this context. This conclusion implies that instructional design in mathematics could benefit from attention to variation of semantic schemas across examples of a given concept.

Although the varied condition led both to more abstract described solutions and to better transfer performance, the former effect did not mediate the latter as expected. In fact, participants who succeeded in describing general solutions were not more likely than other participants actually to demonstrate successful transfer. This result is surprising in light of previous research, in which the quality of participants' generalizations following exposure to multiple examples of a concept *did* predict their ability to apply the concept to novel cases (Gick & Holyoak, 1983; Novick & Holyoak, 1991). Several explanations are possible for this dissociation of described solution methods and problem-solving performance.

First, participants may not have attempted to apply their described solutions during the transfer test, possibly due to failure to recall the solutions or failure to recognize their relevance. However, these possibilities seem unlikely given that the transfer test was administered immediately after participants described their general solutions, and that the problems in the transfer test were presented in the same format and with very similar wording to those in the training. Second, participants may have attempted to apply their solutions, but failed to do so successfully on either or both pairs of transfer problems. Such failure might have been due either to inability to map the elements of the transfer problems to the roles mentioned in their solutions, or to inability to apply the solution procedure despite having correctly mapped the corresponding elements. Both of these issues have been implicated in failures of analogical transfer in mathematics learning (Novick & Holyoak, 1991). Future research might disambiguate between these possibilities by, on the one hand, directly testing whether participants could map elements in the transfer problems to those in training problems, and on the other hand, testing the effects of providing such a mapping to participants.

Regardless of why posttest performance was not predicted by participants' ability to describe correct and general solution methods, it is clear that such ability was not the cause of the superior transfer observed in the varied over the non-varied condition. The question then arises: what *was* the cause for that advantage in transfer? Because this advantage was dissociated from explicit, articulable

knowledge of how to solve the problems, it seems likely to relate to some form of implicit knowledge, e.g. improved perception / encoding of problems or improved procedural skill. Because the procedures required were essentially the same across problems and conditions, the perceptual explanation seems more likely. The varied condition may have encouraged learners to encode the elements of the problems in terms of their general roles in the mathematical structure of SWR, rather than in terms of their more specific roles in one or another semantic schema. Such improved encoding could, in turn, have facilitated application of the solution procedures learned during training to the transfer problems. This explanation is admittedly speculative, but offers a promising direction for future research.

## Acknowledgments

This research was supported by National Science Foundation REESE grant 0910218.

## References

- Bassok, M., Wu, L.-ling, & Olseth, K. L. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory & Cognition*, 23(3), 354-367.
- Elio, R., & Anderson, J. R. (1984). The effects of information order and learning mode on schema abstraction. *Memory & cognition*, 12(1), 20-30.
- Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, 95(2), 393-405.
- Gick, M. L., & Holyoak, K. J. (1983). Schema Induction and Analogical Transfer. *Cognitive Psychology*, 15, 1-38.
- Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The Advantage of Abstract Examples in Learning Math. *Science*, 320(April), 454-455.
- Kotovskiy, L., & Gentner, D. (1996). Comparison and Categorization in the Development of Relational Similarity. *Child Development*, 67(6), 2797-2822.
- Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem solving, and induction. In R. J. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence*, Vol. 5 (pp. 189-223). Hillsdale, N.J.: Lawrence Erlbaum.
- Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of experimental psychology: Learning, memory, and cognition*, 17(3), 398-415.
- Núñez, T., Schliemann, A. D., & Carraher, D. W. (1993). *Street Mathematics and School Mathematics*. Cambridge University Press.
- Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(4), 629-639.

# Learning to recognize unfamiliar voices: the role of language familiarity and music experience

**Micah R. Bregman (mbregman@cogsci.ucsd.edu)**

Department of Cognitive Science, UC San Diego, 9500 Gilman Dr. M/S 0515  
La Jolla, CA 92093 USA

**Sarah C. Creel (creel@cogsci.ucsd.edu)**

Department of Cognitive Science, UC San Diego, 9500 Gilman Dr. M/S 0515  
La Jolla, CA 92093 USA

## Abstract

Speech not only transmits semantic information through words and syntax, but also provides cues to a talker's identity. Differences in a listener's ability to recognize voices can be attributed to their language background, and in rare cases voice recognition can be selectively damaged in neurological patients. In this study we investigated a group of Korean-English bilinguals and non-Korean speakers' ability to learn to recognize unfamiliar Korean and English talkers by voice, and to generalize to utterances not heard during training. We observed an interaction between language background and stimulus language for speed of learning, however generalization performance indicated no such interaction when compared to baseline performance. Bilinguals' performance recognizing English (but not Korean) voices, was predicted by the age they learned English. We also observed that individuals who actively participated in music production exhibited significantly faster task learning than those who did not produce music. This study indicates that language background has a gradient effect on voice learning among bilinguals, and that non-linguistic auditory processing differences, such as music perception, impact voice identification.

**Keywords:** speech perception; music perception; voice; voice identification; individual differences; bilingualism

## Introduction

Speech is generally studied primarily for its ability to communicate semantic meaning from one individual to another. Many complex animal communication systems such as birdsong, however, evolved primarily to communicate more basic information, providing cues that other conspecific listeners use to evaluate fitness and individual identity. Any comprehensive understanding of the evolutionary origins of speech and language will draw both upon the role communication signals play in transmitting semantic meaning, as well as their role in providing cues to identity.

Human speech contains many acoustic cues that listeners use to recognize, for example, a talker's age, gender, emotional state, or even their identity. Collectively, these elements of the speech signal are known as "indexical cues". Voice recognition, or talker identification, is an important aspect of speech perception, and one that has been relatively little studied. Although often considered separate from the core speech perception system (some

neuroimaging results support this perspective, e.g. Belin, Fecteau, & Bédard, 2004), several studies suggest that talker-specific acoustic cues are intertwined with speech recognition. For example, listeners are better able to understand speech from familiar talkers than unfamiliar ones (Nygaard & Pisoni, 1998).

While several studies have characterized severe disability in voice identification, few have attempted to investigate differences among individuals' abilities to recognize voices, although the existence of dramatic individual differences has been noted for many years (Pollack, Pickett, & Sumby, 1954). In clinical cases, voice recognition can be lost completely in individuals with a neuropsychological disorder known as phonagnosia (Van Lancker, Kreiman, & Cummings, 1989). In a pioneering study, Goggin, Thompson, Strube, & Simental (1991) demonstrated that monolingual English speakers were better able to identify the voices of English-German bilinguals when listening to those individuals speak English than when they spoke German. This suggested that, despite many shared acoustic features (both English and German stimuli shared the acoustic features imparted by a particular talker's vocal tract), the listener's language background had a strong impact on their ability to recognize the voices. This study suggested that differences in phonological processing that arise from linguistic knowledge are important in voice recognition.

Goggin et al. (1991) observed no difference in performance on a voice recognition task for English-Spanish bilinguals when tested on English vs. Spanish speaking voices. They suggested that bilinguals might have equal ability recognizing voices from either language since they have extensive phonological knowledge of both. Bilinguals, however, are heterogeneous in their language background, and it may be the case that late learners, or those dominant in one of their languages do exhibit the voice identification deficits identified in monolinguals.

A recent study demonstrated that differences in phonological processing *within* a language can also affect voice identification. Individuals with dyslexia are significantly impaired in their ability to recognize voices relative to controls, but only in their native language (Perrachione, Del Tufo, & Gabrieli, 2011). This result implies that individual differences in phonological

processing, even among those who share a language background, can dramatically impact listeners' abilities to recognize voices.

Outside clinical populations, what other differences might affect voice recognition accuracy? One possibility is music experience. Extensive musical training may benefit the neural encoding of speech by driving brain networks involved in both speech and music perception to function with higher precision than normally necessary for speech perception alone (Patel, 2011). In fact, musicians have been demonstrated to outperform non-musicians on speech perception tasks, including enhanced perception of speech in noise (Parbery-Clark, Skoe, Lam, & Kraus, 2009) as well as enhanced second language phonological ability in bilinguals (Slevc & Miyake, 2006). Do differences in music background or music perception affect voice recognition ability?

In this study, we investigated these questions in a group of Korean-English bilinguals and a second group of non-Korean speakers. We examined whether differences in language and music background, as well as individual differences in music perception and phonological working memory, affected participants' abilities to learn to recognize a set of unfamiliar voices. We also tested recognition of novel sentences spoken by these voices.

## Methods

### Participants

We tested 48 participants, 22 of whom were bilingual, and spoke Korean and English fluently. The remaining 26 participants had no background or experience with Korean. All Korean-English bilingual participants learned Korean as their first language or in parallel with English, and learned English between 1-17 years of age (mean=7.1 years). All subjects studied at UC San Diego and received course credit for participation. All procedures were part of a protocol approved by the UC San Diego Human Research Protections Program.

### Stimuli

We recorded 15 Korean sentences spoken by each of four female native Korean speakers and 15 English sentences spoken by four female native American English speakers. English sentences were selected from the SPIN sentence set. All chosen sentences were high predictability, e.g. "He caught the fish in his net" (Kalikow, Stevens, & Elliott, 1977). Korean sentences were simple, high predictability, and of similar syllabic length to the English sentences, written by a native Korean speaker, e.g. "공책을 집에 놓고 왔다" ("Gongchek eul jibeh nohgo watda," "I left the notebook at home"). Recordings were made in a sound isolated recording booth, and each monaural recording was trimmed to begin at sentence onset and normalized to a mean of 70dB.

## Procedure

**Voice Learning Task** Participants learned to associate 20 training stimuli (5 sentences x 4 voices) with one of four cartoon objects, which differed in both shape and color. Each cartoon object represented a single talker. We chose cartoon objects rather than faces to control for differences in face discriminability across participants. To initiate a trial, participants clicked a cross in the center of the screen. On each trial, audio playback began simultaneously with the display of the two cartoon objects, one on the left and one on the right, equidistant from the center cross. During each training trial, participants clicked one of the two objects with the computer mouse and after clicking, the correct object remained on the screen to provide feedback until they made a second confirmation click.

Training blocks of 60 trials each were presented (with stimuli randomized within each block) until participants reached 85% correct—that is, they chose the target object on at least 51 of 60 trials in a single block (chance=50%). After reaching criterion, participants completed two test blocks, each with 120 trials. During test blocks, no feedback was provided and the screen was blank after making a response. Test blocks contained 60 trials encompassing the 20 training stimuli, as well as 60 trials containing 5 novel sentences produced by the 4 learned voices. The second test block contained 60 trials of the 20 stimuli learned during training and an additional 5 novel sentences. After completing the training and testing process for one language, participants completed the process in the other language (English or Korean). The language of the first block (Korean or English), the cartoon objects associated with each voice, and the positions of the two images on the screen on each trial were counterbalanced across subjects.

**Behavioral assessments** In addition to completing the voice learning task, participants completed assessments to identify individual differences in language and music background and perception. They completed a questionnaire describing their music training, including formal training and current performance activity. To assess their dominant language, bilingual subjects completed a bilingual dominance survey (BDS; Dunn & Fox Tree, 2009) and a picture naming task assessing lexical inventory in English and Korean (modified from Gollan, Weissberger, Runnqvist, Montoya, & Cera, 2011). All participants completed the pitch contour subtest from the Montreal Battery for the Evaluation of Amusia (MBEA) to measure differences in music perception ability (Peretz, Champod, & Hyde, 2003). During the MBEA test, participants heard 2 example melody pairs followed by 31 test melody pairs. For each pair, they provided a same/different judgment. All melody pairs had the same melodic contour and there were no out-of-key notes, making it a fairly subtle change. Each participant's score was recorded as the number of correct responses (observed range = 12-30, mean = 23.5).

For the Korean-English bilingual participants, language dominance measured using the BDS ranged from -15 (English dominant) to 20 (Korean dominant) and averaged -

0.22. Performance on the lexical naming task ranged from -27 to 18, with a mean of -9.48. These bilingual dominance measures were highly correlated ( $r=0.78$ ), and both BDS and MiNT scores were highly correlated with the age English was learned ( $r=0.92$  and  $r=0.75$ , respectively).

Phonological working memory was estimated by measuring each participant's digit span. Digit span has been used as an index of phonological working memory in many experiments (Baddeley & Hitch, 1977). Participants heard a series of 16 audio recordings with a female voice reading random sequences of English digits at a rate of 1 digit per second. Two sequences for each length were presented, in order, from 2-9 digits. After each recording, participants repeated the numbers they had heard. Scores were recorded as the number of sequences correctly repeated, with a maximum score of 16 (observed range = 7-15, mean = 10.7). Digit spans did not differ between language groups (Welch's  $t(45.95)=0.83$ ,  $p=0.41$ ).

## Results

### Language familiarity predicts learning speed

Previous research suggests that familiarity with a language is predictive of performance on voice identification tasks. However, its role predicting learning rate for unfamiliar voices has not been explicitly tested. We contrasted 22 Korean-English bilinguals with 26 listeners who did not speak Korean. We measured the number of blocks required to reach a criterion of 85% correct within a single block. A 3-way mixed model ANOVA (Figure 1) with Participant Language (English-only, Korean-English; between-participants), Talker Language (English, Korean; within-participants) and block order (English first vs. Korean first; between-participants) revealed no significant main effects of participant language background ( $F(1, 44)=3.19$ ,  $p=0.08$ ), stimulus language ( $F(1, 44)=0.44$ ,  $p=0.51$ ), or block order ( $F(1, 44)=1.09$ ,  $p=0.30$ ). However, there was a strong interaction between stimulus language and language background ( $F(1, 44)=24.02$ ,  $p<0.0001$ ).

Individually, Korean-English bilingual participants were faster to learn Korean talkers ( $M=1.9$  training blocks) than English talkers ( $M=3.5$  blocks; paired  $t$ -test  $t(21)=-3.03$ ,  $p=0.006$ ). Similarly, English-speaking participants learned English voices ( $M=2.5$  blocks) faster than Korean voices ( $M=4.5$  blocks; paired  $t$ -test  $t(25)=4.14$ ,  $p=0.0003$ ). No other interactions were statistically significant (all  $F$ s  $< 0.08$ ,  $p$ s  $> 0.78$ ). Together, these data show that differences in learning rates are present as a function of language background.

We then looked at participants' maximum accuracy on training trials. Although trained to reach a criterion of 85% correct in a block, some participants achieved higher accuracy than others. Again we observed an interaction

between language background and stimulus language (Figure 2a) in the maximum accuracy reached. A 2-way mixed ANOVA indicates no main effects of language background ( $F(1, 46)=2.08$ ,  $p=0.16$ ) or stimulus language ( $F(1, 46)=1.11$ ,  $p=0.30$ ), but a strong interaction ( $F(1, 46)=15.51$ ,  $p=0.0003$ ).

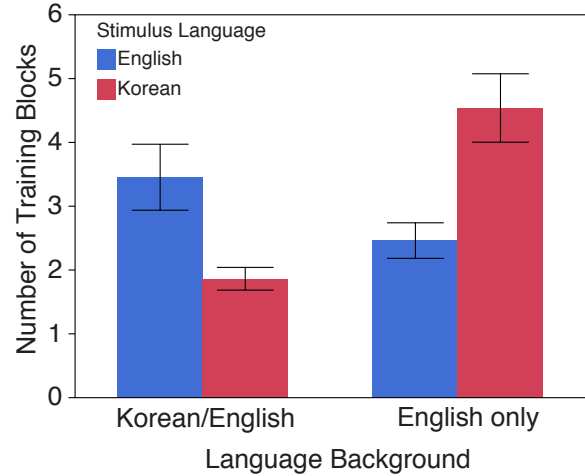


Figure 1: Korean-English bilinguals required fewer training blocks to reach 85% correct recognizing Korean speaking voices (red bars) than English speaking voices (blue bars). Non-Korean speakers show the opposite effect. Bars indicate mean number of training blocks  $\pm$  s.e.

However, we observed no difference in performance between training and generalization test trials in the 40 participants who reached 85% correct after a maximum of 9 training blocks. For each of these participants, we calculated a “generalization penalty” by subtracting the proportion of correct responses to novel tokens of learned talkers with the proportion of correct responses to trained talkers. All stimuli were interleaved and collected in the same test block. We computed a 2-way mixed model ANOVA predicting participant's generalization penalty using language background (between participants) and stimulus language (within participants) as factors (Figure 2b). We observed no main effect of language background (Korean-English vs. English-only; between participants,  $F(1, 39)=2.45$ ,  $p=0.13$ ), no main effect of stimulus language (within participants,  $F(1, 39)=1.72$ ,  $p=0.20$ ) and no interaction between language background and stimulus language ( $F(1, 39)=0.17$ ,  $p=0.68$ ). While language background appears to be important for *learning* to distinguish unfamiliar voices, it does not appear to constrain generalizing to new utterances after the voices have been learned, at least within the short retention period required in this experiment.

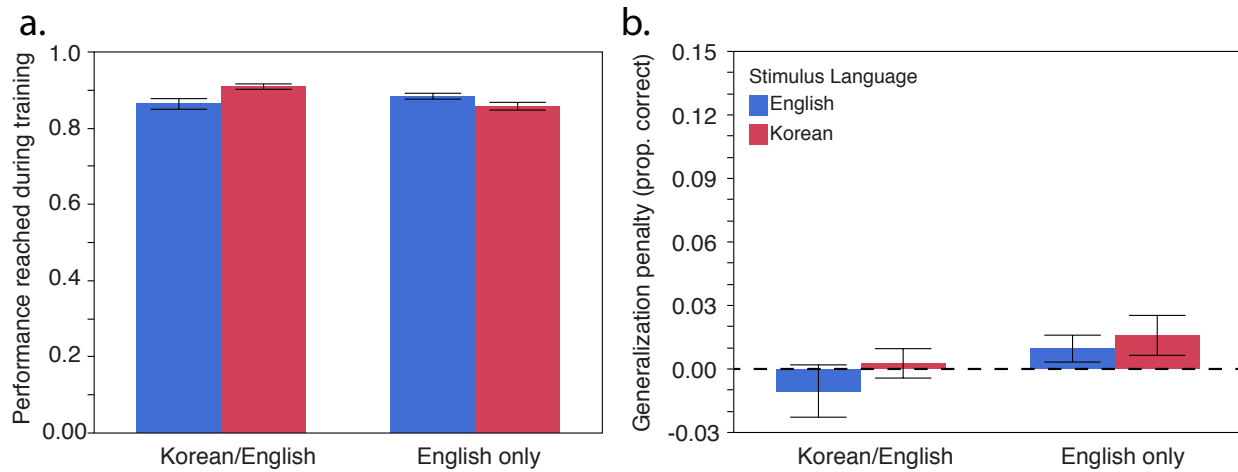


Figure 2: (a) Korean-English bilinguals were slightly more accurate at identifying the correct voice on novel sentences for Korean stimuli (red bars) than for English stimuli (blue bars). Non-Korean speakers show the opposite effect. (b) There were no generalization differences between groups

### Bilinguals' age of L2 acquisition predicts learning speed in L2, but not L1.

We further explored whether individual differences in age of learning English or relative dominance of English or Korean were predictive of task performance among the bilingual subjects. To do so, we computed the correlation between age of English onset (which was the second language for all bilingual participants) with their voice learning rate. Among Korean-English bilinguals, blocks to criterion on English talkers was positively correlated with the age they began learning English (Figure 3a,  $r(20)=0.62$ ,  $p=0.002$ ), while it is uncorrelated for Korean-language stimuli ( $r(20)=0.24$ ,  $p=0.28$ ).

We then separated Korean-English bilingual participants into two groups based on a median split of acquisition age:

those who learned English at or before 5 years old (early learners,  $n=12$ , mean age=3.3 years, mean BDS=-7.8, mean MiNT=-15.6) and those who learned after 5 years old (late learners,  $n=10$ , mean age=10.7 years, mean BDS=6.9, mean MiNT=-4.3). We then conducted a 2-way mixed model ANOVA with factors of Participant Language (between participants; English-only, early-English Bilingual, late-English Bilingual) and Talker Language (within participants). There was a main effect of language background ( $F(2, 45)=4.73$ ,  $p=0.014$ ), no main effect of stimulus language ( $F(1, 45)=1.31$ ,  $p=0.26$ ) and an interaction between language background and stimulus language ( $F(2, 45)=15.91$ ,  $p<0.0001$ ). This interaction resulted from three different patterns of talker learning. Early-learning bilinguals did not differ in their acquisition rate for Korean and English stimuli (paired  $t(11)=-1.74$ ,  $p=0.11$ ). However, late-English-learning bilinguals learned

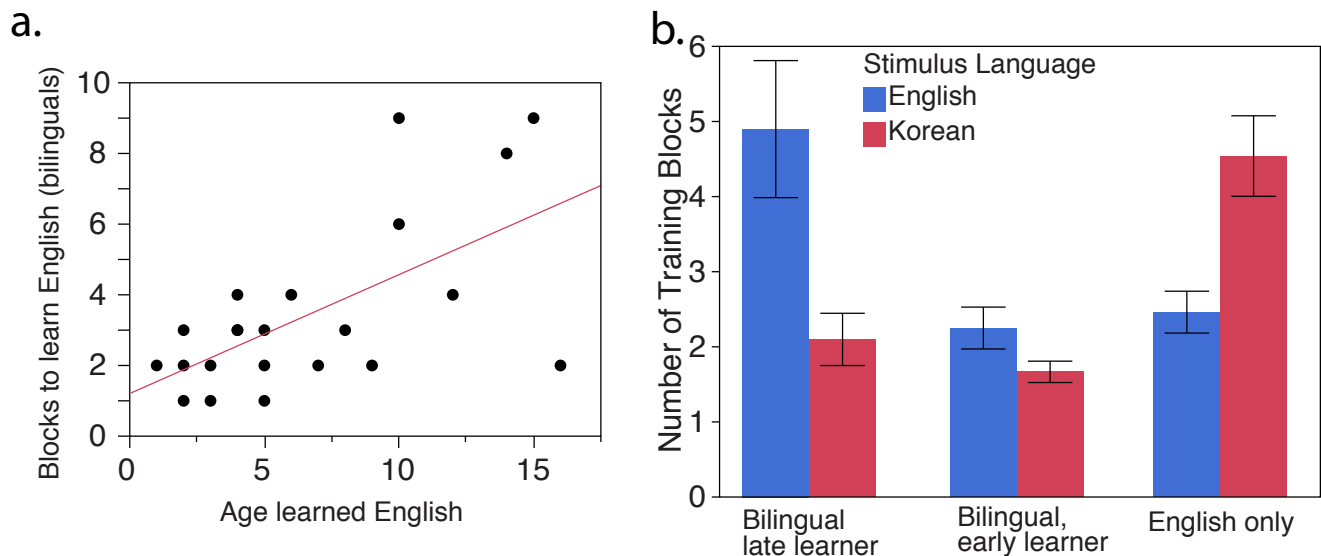


Figure 3: (a) The number of blocks to learn English voices was correlated ( $r=0.62$ ,  $p=0.002$ ) with the age Korean-English bilinguals learned English. (b) Number of training blocks to reach criterion of 85% on each stimulus language for Korean-English bilinguals who learned English late ( $n=10$ ), early ( $n=12$ ) and English-only speakers ( $n=26$ ). Each bar represents the mean number of training blocks (60 trials/block)  $\pm$  s.e.



Korean stimuli faster than English stimuli (paired  $t(9)=-2.87$ ,  $p=0.018$ ), and, as reported above, non-Korean speakers learned English stimuli faster than Korean stimuli.

Taken together, these results are consistent with prior work suggesting that phonological processing is an important element of voice recognition. Our result extends previous work by demonstrating a *gradient* effect of bilingualism. Rather than showing similar patterns of behavior in both languages, age of acquisition is an important predictor of performance recognizing voices in L2, but not L1.

### Music experience predicts learning rate

We collected several behavioral measures of individual differences in auditory perception from our participants (see methods). Our hypothesis was that, since differences in individuals' language profiles (e.g. language familiarity, dyslexia) contribute to differences in voice learning, we might also observe differences among participants due to individual differences in auditory processing that are not strictly linguistic: pitch perception, music background, and music perception ability. We report the correlations between each of these measures and three performance measures: learning rate, generalization performance, and pitch shifted generalization performance (Table 1).

Several previous studies have identified perceptual advantages for individuals with extensive musical training. In particular, musicians have shown better brainstem encoding of pitch (Wong, Skoe, Russo, Dees, & Kraus, 2007), and high musical ability is associated with better second language phonology (Slevc & Miyake, 2006). Is musical experience important for learning to recognize voices?

**Table 1.** Correlations between music measures and voice recognition

	Years Training	MBEA Score	Tone Thres.	<i>Generalization</i>		
				<i>Learning rate (blocks)</i>	<i>Unshifted</i>	<i>Shifted</i>
Years Training	1.000	0.10	-0.02	<b>-0.42</b>	0.029	-0.095
MBEA Score		1.000	-0.26	-0.19	-0.048	-0.199
Tone Thres.			1.000	0.13	0.101	0.208
Learning rate				1.000	-0.202	0.246

We measured musical perceptual ability with the melody contour subtest of the MBEA (Peretz, Champod, & Hyde, 2003), and a pitch discrimination threshold task. Pitch difference threshold and MBEA did not correlate significantly with voice learning or generalization ability. However, measures of musical activity did show a relationship to voice learning rate. Participants who were currently active in producing music at least 1 hour per week

when the experiment was conducted ( $n=11$ ; musical training averaged 12.0 years, range 6-22 years) learned to recognize voices on average in fewer training blocks than those who were not active musicians ( $n=37$ ; who had less musical training, averaging 5.2 years, range 0-27; Welch's  $t(34.23)=-2.52$ ,  $p=0.017$ ). This difference seems to have been driven by musicians' more rapid learning for voices speaking the subject's non-dominant language. When tested on the non-dominant language (Korean for non-Korean speakers, English for Korean-English bilinguals), musicians learned faster than non-musicians (mean=2.71 blocks vs. 4.62 blocks, Welch's  $t(44.28)=-3.07$ ,  $p=0.004$ ). However, when learning to recognize voices in their dominant language, we observed no effect of music background (mean=2.00 blocks for musicians vs. 2.40 blocks for non-musicians, Welch's  $t(17.02)=-0.59$ ,  $p=0.56$ ).

As there are multiple ways of assessing music experience, we also considered the effect of years of musical training (this did not overlap completely with current musical practice). Years of training correlated negatively with average number of training blocks to reach criterion ( $r(46)=-0.42$ ,  $p=0.0036$ ). Again, the relationship to music training is driven by the non-dominant language ( $r(46)=-0.40$ ,  $p=0.006$ ); musical training was not significantly correlated with learning rate for voices in the dominant language ( $r(46)=-0.22$ ,  $p=0.13$ ).

### Discussion

Previous studies demonstrated that individual differences in phonological processing due to language background and dyslexia are important predictors of voice identification ability. The results of the current study extend these findings in a few important respects. In both adults and infants, knowledge of a language improves ability to recognize voices in that language (Goggin et al., 1991; Johnson, Westrek, Nazzi & Cutler, 2011; Perrachione et al., 2011). We extended this work by investigating both monolinguals and bilinguals, and looking at the bilingual participant's language dominance. Not only did we find a crossover interaction between listeners' native-language backgrounds and talkers' language, but we also found that early second-language acquisition facilitated talker learning without loss in performance on the first language. This acquisition effect—if viewed as such—is particularly interesting because it mimics acquisition of phonology: as age of acquisition increases, receptive and productive phonology are less native-like (Flege et al., 2006; Oh et al., 2011).

We also observed significantly faster voice learning for participants with more extensive musical training, particularly those actively involved in music production. This could be associated with changes in auditory encoding that have been observed among musicians that give rise to differences in pitch, music and speech perception. Our result extends this area of research, suggesting that not only is speech comprehension enhanced, but perception of indexical features in the speech signal may be enhanced as well. The effect of music experience appeared only to apply

to participants' learning to recognize voices in a less familiar language. We point out, however, that this study does not actually manipulate music training, so we cannot assert that it *causes* improvement in learning to recognize voices. Perhaps some third variable—inherent or learned individual differences in auditory perception—confers benefits to both voice recognition and music production.

Further work is also needed to identify whether the kinds of individual differences that give rise to enhanced voice recognition also extend to other indexical cues. Are individuals who performed better on individual recognition tasks also more sensitive to acoustic cues such as a talker's emotional state, age, or gender?

We explored how language experience and non-linguistic factors contributed to talker identification in two different languages. Native-language talkers were learned faster than second-language or unfamiliar-language talkers, and among bilinguals, earlier L2 acquisition predicted faster learning. Further, some measures of music experience predicted faster learning in the less-familiar language. Our work suggests a role for early language learning, or at least extent of exposure, in talker identification. This is consistent with a tight linkage between language processing and talker identification, which presents an interesting puzzle given the evidence of specialized neural mechanisms for speech recognition and talker identification.

### Acknowledgements

MRB was supported by the Kavli Institute for Brain and Mind at UC San Diego, and SCC was supported by an NSF CAREER Award (BCS-1057080). We would like to acknowledge the help of undergraduate research assistants Shawn Cho and Hye Young Lee who were instrumental in developing Korean language stimuli and collecting data on participant's language background.

### References

Baddeley, A. D., & Hitch, G. J. (1977). Working Memory. The psychology of learning and motivation: advances in research and theory.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences*, 8(3), 129-35.

Dunn, A. L., & Fox Tree, J. E. (2009). A quick, gradient Bilingual Dominance Scale. *Bilingualism: Language and Cognition*, 12(03), 27-

Flege, J. E., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics*, 34(2), 153-175.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & cognition*, 19(5), 448-58.

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2011). Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary norms for young and aging

Spanish-English bilinguals. *Bilingualism: Language and Cognition*, 1-22.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, 14, 1002-1011.

Kalikow, D., Stevens, K. N., & Elliott, L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, 61(5), 1337.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & psychophysics*, 60(3), 355-76.

Oh, G. E., Guion-Anderson, S., Aoyama, K., Flege, J. E., Akahane-Yamada, R., & Yamada, T. (2011). A one-year longitudinal study of English and Japanese vowel production by Japanese adults and children in an English-speaking setting. *Journal of Phonetics*, 39(2), 156-157. d

Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009). Musician enhancement for speech-in-noise. *Ear and hearing*, 30(6), 653-61.

Patel, A. D. (2011). Why would Musical Training Benefit the Neural Encoding of Speech? The OPERA Hypothesis. *Frontiers in psychology*, 2, 142.

Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of Musical Disorders The Montreal Battery of Evaluation of Amusia. *Annals of the New York Academy of Science*, 999, 58-75.

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. E. (2011). Human voice recognition depends on language ability. *Science* (New York, N.Y.), 333(6042), 595.

Pollack, I., Pickett, J., & Sumby, W. (1954). On the identification of speakers by voice. *Journal of the Acoustical Society of America*, 26(3), 403-406.

Slevc, L. R., & Miyake, A. (2006). Individual differences in second language proficiency: Does musical ability matter? *Psychological Science*, 17(8), 675-681.

Van Lancker, D. R., Kreiman, J., & Cummings, J. (1989). Voice perception deficits: neuroanatomical correlates of phonagnosia. *Journal of clinical and experimental neuropsychology*, 11(5), 665-74.

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience*, 10(4), 420-2.



# Misconceptions Regarding Emergent Phenomenon Vary By Domain

**Sarah K. Brem** ([sarah.brem@asu.edu](mailto:sarah.brem@asu.edu))

School of Social and Family Dynamics, Mail Code 3701  
Arizona State University  
Tempe, AZ 85287-3701

**Glenda S. Stump** ([glenda.stump@asu.edu](mailto:glenda.stump@asu.edu))

School of Social and Family Dynamics, Mail Code 3701  
Arizona State University  
Tempe, AZ 85287-3701

**Gale M. Sinatra** ([gsinatra@usc.edu](mailto:gsinatra@usc.edu))

Rossier School of Education, 3470 Trousdale Parkway  
University of Southern California  
Los Angeles, CA 90089

**Raymond Reichenberg** ([raymond.reichenberg@asu.edu](mailto:raymond.reichenberg@asu.edu))

Mary Lou Fulton Teacher's College, Mail Code 3151  
Arizona State University  
Tempe, AZ 85287-3151

**Benjamin Heddy** ([heddy@usc.edu](mailto:heddy@usc.edu))

Rossier School of Education, 3470 Trousdale Parkway  
University of Southern California  
Los Angeles, CA 90089

## Abstract

Although the study of how learners approach emergent phenomena is relatively new, a consistent set of misconceptions associated with emergence have been documented. However, little consideration has been given as to whether some misconceptions manifest more frequently in one domain than another, or take on a different character depending on the agents or phenomenon involved. We examined participants' explanations of emergent phenomena from three domains. We found significant differences between domains, showing greater or lesser evidence of misconceptions. We propose that novices bring prior knowledge, folk psychology, and folk biology to bear in determining the capabilities of the agents involved in a phenomenon, and that these beliefs guide their explanations. We believe that the study of how people perceive emergence would benefit from drawing upon research on folk theories, anthropomorphism, developmental constraints, and other areas that will help us understand how learners characterize agents, environments, and their interactions.

**Keywords:** emergence; complexity theory; misconceptions; science education; folk biology; folk psychology

## Introduction

We live in a complex world. Not just in the everyday sense, but in a mathematical, or scientific, sense. Many of the phenomena we encounter in everyday life are of the emergent, or complex, sort. Emergent phenomena play a

central role in every scientific discipline. Color and convection are emergent phenomena, as are weather patterns, earthquakes, and the evolution of galaxies. The activities carried out by ant colonies, bee hives, and American voters exhibit emergence, as does the co-evolution of flowers and bees. There is, therefore, potential for great benefit in developing an understanding of how emergent systems arise and behave, but this is a difficult task for learners.

In non-complex, or "direct," systems, the overall behavior of the system and its outcomes tends to be deterministic, linear, and predictable, often organized by a centralized process or individual leader. The circulatory system (Chi, 2005, in press) follows a clear path, each step having a clear purpose in a system that is regulated by nerves keeping the heart beating at a regular pace. A problem in Newtonian mechanics, say, the building of a bridge, involves identifying forces that sum to indicate the stresses on a particular element of the bridge, stressors which can be offset directly by adding new elements in a cumulative way, or inventing new structures that alter how the forces sum.

In contrast, complex systems are unpredictable, nonlinear, and give rise to novel, or emergent, behavior (Holland, 1999). Even when the exact rules governing each agent are fully specified, the resulting phenomenon will be unpredictable and irreducible, and adding new elements multiplies the possible interactions between agents, leading to unexpected, non-replicable outcomes. There are no set

leaders or controllers; an agent may appear to have such powers, such as the lead goose in a flock, but they are in the lead simply because the other geese fell in line with them.

It is not surprising that the naïve learner struggles when faced with explaining and understanding emergent phenomena. Previous studies have well-documented the challenges that emergent systems present for learners (Jacobson, 2001; Hmelo-Silver & Azevedo, 2006). Learners new to emergent systems are likely to expect clear patterns of cause-and-effect, purposeful encounters between agents, and a central control system that oversees their movements and actions (Chi, 2005; Resnick, 1996). Complex systems introduce the notion that order can emerge out of random interactions; stochastic movements and actions are central to how emergent phenomena arise and evolve. This challenges the commonsense belief that ascribes purposefulness to events in the universe (Jacobson, 2001). In a deterministic, linear system, small changes to the starting conditions generally lead to small changes in outcomes. In a complex system, the smallest change can have a drastic effect on the outcome, as interactions amplify and reshape the actions of individual agents and their encounters with their environment. The learner tries to make sense of these patterns by invoking centralized control, in the form of a “queen” bee or other sort of leader, who has knowledge of the entire goings-on and can shift the pattern according to their needs or the needs of the group. But there is no leader to be found.

The study of how people perceive, explain, and understand emergent systems is a relatively new area, and both logistical considerations and underlying assumptions have led us, as researchers, to pursue our investigations through a close examination of single phenomenon (Chi, in press), a single self-contained ecosystem (Hmelo-Silver & Pfeffer, 2004), or a small set of phenomena within a single domain, such as chemistry (e.g., Talanquer, 2008; Rappoport & Ashkenazi, 2008) or evolution (e.g., Evans, 2001; Poling & Evans, 2004).

Logistically, this approach has been fostered by the fact there are only a few tools for modeling emergent systems (e.g., Colella, Klopfer & Resnick, 2001; Tisue & Wilensky, 2004), and creating such models is highly time-consuming. Data collection is similarly time-consuming, and research has focused heavily on case studies and close observation of small groups in complex settings (e.g., Charles & d'Apollonia, 2004; Wilensky & Resnick, 1999), or larger groups interacting with a single aspect of a phenomena. We have spent little time looking at whether and how learners' actions and ideas change from one domain to the next.

Theoretically, too, there is a basic reason why we might believe that, with time, these individual studies would come together to create a fuller account of how learners perceive, represent, and think about emergent systems. To be considered an emergent system, a phenomenon must meet specific mathematically-defined criteria. At the mathematical level, diffusion, chemical bonds, and traffic jams have a great deal in common; emergent phenomenon

belong to a class that is unaffected by features specific to a particular domain.

It remains to be seen, then, to what degree the abstract features and behaviors of an emergent system actually do cross domain boundaries for learners. If I become proficient in understanding diffusion, for example, will this make learning about traffic jams or chemical bonds any easier for me? We not only do not have such comparisons, we lack assessments that allow us to directly compare a learner's understanding of the principles of emergence in one domain to their understanding in another.

Our goal for the project we describe here was to create an assessment of key aspects of emergent systems that could be applied across topics. Learners' understanding of each phenomenon was assessed using an instrument that had identical stems representing each component of emergence, into which the particular phenomenon could be inserted. If learners respond to the basic patterns that underlie the phenomenon, they should show similar levels of understanding (or misunderstanding) across domains. If a person showed a low understanding of diffusion, they should show a similarly low understanding of geese flocking, as the same basic principles are needed to make sense of these two phenomena. Thus, we can determine whether their knowledge is tied to a particular system, or is available at a more abstract level.

If domain knowledge plays a role, however, then responses may be influenced by folk theories that ascribe greater capacity for volition, control, and decision making to certain sorts of entities, and less to others. A bacterium may be seen as relatively incapable of engaging in goal-oriented behavior, of communicating with other bacteria, or of choosing a course of action. A goose, given that it has a brain and nervous system, may be seen as more capable. The social behavior of ants, likewise, may cause us to assume they are more intentional and communicative, and less driven by instinct and environment. Given people's bias to find centralized, intentional causes for emergent patterns, it may be harder to believe that animals with greater neurological development are subject to random processes, do not engage in communication and group planning, and do not make choices when they act.

We therefore decided to test our domain-general stems in the context of three different domains: unicellular slime molds aggregating before sporing, ants, a social animal, foraging for food, and geese, the most “advanced” neurologically, flocking. These three entities not only come from different locations on the phylogenetic tree, they differ in the ways that we describe above.

We designed a study in which participants watched simulations of three phenomena. Participants responded to an open-ended written protocol that used the same questions for each domain, altered only to refer to the domain at hand. We then coded participants' responses and compared their conceptualizations of each phenomena.

If misconceptions arise due to the features that all emergent phenomenon have in common, we would expect

responses to our probes to be roughly the same across the simulations. If, however, domain-specific considerations come into play, we expect to see misconceptions arising more often in some domains than in others. In particular, we predicted that misconceptions would be invoked less often with regard to slime molds than ants or geese. In addition, we predicted ants would give rise to the greatest number of misconceptions due to the familiarity of their social nature.

## Methods

### Participants

Forty participants, undergraduates from a large Southwestern University, completed the written protocol, receiving \$20 in compensation. None had formal training in emergence or complexity theory. The protocol took approximately 60 minutes to complete.

### Materials and Procedure

The participants completed a written protocol posing questions about three emergent phenomena: geese flocking, ants foraging for food, and slime molds aggregating to spore. The order in which the phenomena were presented was counterbalanced across participants.

Each phenomenon was illustrated with a NetLogo simulation that had been video-captured, so that the same run could be shown to all participants. Agents were represented by icons that captured the basic appearance of the real entity (i.e., ant, slime mold, goose), and participants were told before the simulation began what symbols would be used, and what they would mean.

Each simulation lasted approximately 90 seconds, and were roughly divided into three phases. First, there was a brief section that allowed participants to orient themselves to the simulation, the symbols used for the agents and environmental objects, and the agents' behavior. Next, the behavior began to develop emergent properties, i.e., patterns began to form at the group level. Finally, the simulation reached equilibrium (in the case of the flocking and slime mold simulations) or the agents achieved their goal (of finding food, in the ant simulation). Pilot runs of the simulations and the responses of the participants suggested that few had any difficulties mapping the real phenomenon onto the simulation.

There were a total of 7 questions. We began each of the three protocol sections with broad questions (e.g., #1 and #2 below), moving to more specific questions that capture key aspects of emergent phenomena (e.g., #3 and #4 below). The questions were designed to be as similar as possible across entities, substituting in the appropriate phrases:

1. Describe the patterns that the *ants/geese/slime mold organisms* make in as much detail as possible.
2. What do you think causes the *ants/geese/slime molds* to make the patterns you see?

3. Do you think that there are special leader *ants/geese/slime molds* that signal the others to *follow them/ follow them/ come to them and form clusters*?
4. If we make a new video with the same *ants and food/geese/slime molds* in the same starting positions, how similar do you think the patterns they form will be?

We instructed participants to write as much as they could, giving as much detail as they could provide. They were told to answer the questions in the order they were given, and not to go back and change any of their answers.

## Results

To create the coding system used to categorize participants' answers, two of the authors (SKB and GSS), conducted an open coding of the data, allowing codes to emerge, rather than searching for codes based on existing expectations and hypotheses. We iteratively coded sections of the data and discussed our codes, coming to agreement on 13 themes that were present across all domains. No theme arose that was not present at least once in each domain. Two of the other authors (RR and BH) who did not have contact with GSS and SKB during the development of the codes, applied the codes to the data. Both were blind to the hypotheses regarding the relationship between a given domain and possible patterns for that domain. They applied the codes with 94.5% agreement, resolving disagreements through discussion. Less than 5% of the answers were deemed uncodable. Multiple codes could be applied to a single answer, if all of the criteria for each code was met.

Although we arrived at 13 themes, not all of these codes spoke to the issue of misconceptions about emergence. Some were not directly relevant (e.g. "descriptive," used when for a play-by-play description of the simulation, without interpretation, or "external factors," when characteristics of the simulation itself were the focus, instead of the content.) Other codes proved complex and potentially misleading because they captured different concepts for different individuals (this is further discussed below).

We chose to focus on the codes that, based on prior research, are most central to identifying whether participants hold a misconception or correct representation. We chose four that directly invoke misconceptions well-documented in the literature, and two providing evidence that the participant held a correct representation of the phenomenon (see Table 1 for a description of the codes, and a sample response). Regarding the misconceptions, people tend to assume that there is a controlling force directing the agents, that the agents are communicating and cooperating, that certain agents have special powers that allow them to direct the pattern, and that all of the agents are acting out of a sense of purpose. In contrast, the last two suggest understanding of two basic principles of emergence: the

lack of centralized control, and the role of stochastic processes.

Table 1: Codes Used in The Analysis

Code	Description	Example
Centralized Control	Reference to a group member determining, directing, or guiding the actions of the rest of the members.	<i>The slime mold that produces more pheromones will signal the others to come and cluster around them.</i>
Cooperation	Agents cooperatively determine their behavior, or work together as a group.	<i>Each geese are telling each other where to move</i>
Differentiation	Members have different abilities, roles, or attributes from one another.	<i>Maybe the ones clustered have a similar pheromone (sic). Also the levels of what they give off could be different.</i>
Goal-oriented	Specifies a behavior being performed to fulfill a specific, stated purpose.	<i>He takes the path he does because he's searching for food.</i>
Lack of Central Control	Refers to lack of a group member determining, directing, or guiding the actions of other members.	<i>I think the ants are generally just concerned with getting some food, so following a leader was not the goal. They all followed very similar paths because they shared a common goal.</i>
Random Processes	Describes movement or other activity explicitly as "random." Does not include descriptors such as "haphazard" or "scattered," but only those which or seem to be referring to a reasonably accurate version of statistical randomness.	<i>I think this is a random pattern while they search for food. The first makes random search while the followers more directly follow.</i>

Our first step in analyzing the data was to calculate the means and standard errors for each domain on each of these 6 codes (see Table 2 for descriptive and inferential statistics). We took a "token-counting" approach, adding up the number of times a particular code was used a participant in a domain. This gives a sense of how strongly the participants relied on a particular conception in providing their explanations for the phenomena. Since a participant could have invoked a particular concept or principle in answering each question, the maximum token score for each code is seven.

Table 2. Descriptives For Each Domain, By Code (alpha corrected to account for experiment-wide error)

Code	Slime	Ants	Geese	F(2, 78)
Centralized Control	0.23 (.07)	3.35 (.27)	1.18 (.21)	64.22***
Cooperating Agents	1.08 (.17)	0.55 (.13)	1.50 (.27)	8.10**
Differentiated	1.05 (.22)	3.95 (.26)	1.85 (.28)	35.86**
Goal-Oriented	2.70 (.29)	3.77 (.25)	3.03 (.29)	5.34*
Lack Central Control	0.78 (.09)	0.32 (.17)	0.38 (.13)	3.40 <sup>m</sup>
Random Processes	0.70 (.22)	0.48 (.14)	0.38 (.13)	1.51 <sup>ns</sup>

We also calculated participants' use of these codes using a binary process; a score of "1" meant that the participant used the code at least once within that domain; "0" indicated they did not use the code at all in that domain. The results were quite similar; the same pattern of significant findings was found for all but two of the codes. In the case of "goal-oriented," the differences decreased, and the F-value fell to 2.69 (non-significant). In the case of "lack of centralized control," differences increased, and the F-value rose to 16.98 ( $p < 0.01$ ). We believe these differences are due to the following: in the case of goal-orientation, almost every participant invoked it at least once in every domain, restricting the range when using binary scoring. The opposite is true in the case of a lack of centralized control; so few participants invoked this, a binary analysis was able to detect differences that were swamped by non-responses in the token analysis.

### Other Points of Interest

As noted above, there were themes in the data that were not as central, but might shed some light on how participants conceptualized emergence. For example, in reviewing the protocols, we found that participants were expressing quite different ideas, even when using similar language. In one

case, we inquired as to whether the participants believed that the agents followed rules in carrying out their actions. The consensus seemed to favor that they did not follow rules (approximately 2/3 of the replies).

However, we also discovered that participants had different ideas about what constituted a “rule.” For some, following a rule meant making a conscious decision based on a learned rule—we learn to stop at red lights, for example, and we (usually) follow that rule. For others, rules could also be innate or instinctual. Some even went so far as to explicitly state “if you mean conscious rules, no, but if you mean instincts, then yes, they follow instincts.”

Because we could not go back and further probe to see what meaning of “rule” each participant used, we excluded it from further analysis, but we think this disagreement about what constitutes a rule is interesting, for reasons we will address in the discussion.

## Discussion

As predicted, slime molds elicited fewer misconceptions overall than the other two entities, and produced the lowest level of misconception use in three of the four categories associated with misconceptions. Ants also elicited the predicted performance, with the highest use of misconceptions in three of the four categories. Trends in the two categories related to correct conceptions were unclear; reference to an explicit lack of centralized control was only significant in the binary coding of the data; this trend did favor our prediction, in that slime molds had the highest invocation of this correct concept. However, random processes produced no significant differences in either analysis.

The mathematical and scientific power of complexity theory and emergence comes from the ability of these theories to draw bridges between seemingly disparate disciplines. These isomorphisms, along with the tremendous overhead involved in modeling any phenomenon, have led educational researchers and cognitive scientists to focus on a small sub-set of emergent phenomena, reasonably inferring that the errors that crept up in one domain would appear in another, given the underlying similarities. Even if novices did not know the phenomena were isomorphs, the phenomena were governed by the same principles, manifest in similar ways, and created similar puzzles.

There was indeed some similarity in how the participants responded to phenomena; of the 13 themes we identified, every one was present at least once in each domain. That suggests a relatively high degree of consistency across domains in terms of how participants describe and explain phenomena.

However, we also believe that researchers have not been paying enough attention to the fact that each phenomenon is carried out by a different cast of characters—be it molecules, ants, geese, or air streams—and novices might rely on prior knowledge and beliefs in perceiving a phenomenon and devising an explanation for it. As a result, the likelihood of invoking a particular concept or

misconception is not just due to the phenomenon’s abstract characteristics, but also what the novice brings to the phenomenon, perhaps in the form of specific knowledge of the entities, or perhaps in the form of folk theories.

The problem of interpreting participants’ use of the word “rules” illustrates our account well. Differences in one’s beliefs about what constitutes a rule, and whether an entity is capable of acting on rules created differences in the way participants responded. Most participants seem to think of rules as learned guides to appropriate behavior to which one consciously refers. They were reluctant to ascribe that ability to the entities we used in this study. Those who believed that instincts could be thought of as rules were much more willing to think of agents as following rules.

Similarly, we believe that the differences between the three domains on the codes we examined are driven less by specific knowledge of emergence, and more by one’s beliefs about the entities’ capacity for thinking, consciousness, and deliberate decision making. They invoked misconceptions less for slime molds, unicellular microorganisms, than for geese and ants, and there is some evidence that they were more likely to invoke a correct explanation for the slime molds, based on a lack of centralized control and random processes.

Their accounts of ants consistently showed the greatest number of misconceptions; only for cooperation did geese outperform ants. Anecdotally, a fair number of participants spontaneously stated that they knew how ants worked because (a) they have had extensive experience with ants, usually trying to get them out of their houses, and (b) because they had seen the movie ‘Antz.’ They were also inclined to mention that ants were social creatures, and that ability suggested organization, and the mental capacities needed to create organization.

At very least, this should serve as a warning to those of us engaging in research about complexity and emergence. By relying on one or a few phenomena to characterize how people perceive, explain, and understand emergence more generally, we may be greatly over or underestimating the abilities depending on the domains and agents we choose. Testing a model on a variety of phenomena, with agents at different levels of familiarity and perceived cognitive capacity would be a good step to take before deciding that a particular pattern of results arises because of the character of emergent systems generally, and because of the specific properties of the agents and the activities they undertake.

We believe, however, that there is something more interesting going on here, and that a better understanding of how people experience emergence will require us to draw upon research into folk theories (Arico, Fiala, Goldberg & Nichols, 2011) developmental constraints (e.g., intentionality, teleology; Sinatra, Brem & Evans, 2008), anthropomorphism (Tamir & Zohar, 2006), and sociocultural accounts of the differences in ways that different groups of people characterize animals, people, and objects. We need a better sense of how people characterize agents and their abilities to exhibit volition, teleological

thinking, to communicate, and to control themselves, others, and their environment.

We hypothesize that the greater the perceived capabilities of an agent or group of agents, the more likely it is that people will reject explanations that invoke emergence in favor of accounts that give agents greater control over the events that occur. Alternatively, the differences between simulations were due to differences in the phenomena we chose. It may be, for example, that flocking seems to require greater mental skill than following a pheromone trail.

As a first step in addressing these hypotheses, we are currently running a study in which we present isomorphic phenomena across different levels of agents (physical, "lower animal," "higher animal," and human), and are also gathering data about the perceived capacities of each of these agent types. We believe that simulations depicting agents deemed more mentally capable will correlate with greater misconceptions, even when the underlying mechanisms are actually identical. However, if it is the type of phenomenon that drives the differences in our first study, this should surface in this study; responses should be more similar by phenomenon than by level of agent.

In either case, having a better understanding of how perceptions of phenomena and agents vary, this should improve our ability to understand how people look at emergent phenomena, and suggest ways to dispel misconceptions.

### Acknowledgments

This work was made possible by a grant from the National Science Foundation to the first and third authors (0910115). We also thank Katherine G. Nelson and Susan Shapcott for their assistance in preparing this manuscript.

### References

- Arico, A., Fiala, B., Goldberg, R.F., Nichols, S. (2011). The folk psychology of consciousness. *Mind and Language*, 26, 327-352.
- Charles, E.S. & d'Apollonia, S.T. (2004). Developing a conceptual framework to explain emergent causality: Overcoming ontological beliefs to achieve conceptual change. *Proceedings of the 26<sup>th</sup> Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Erlbaum Associates.
- Chi, M.T.H. (2005). Common sense conceptions of emergent processes: Why some misconceptions are robust. *Journal of the Learning Sciences*, 14, 161-199.
- Chi, M.T.H., Roscoe, R., Slotta, J., Roy, M., & Chase, M. (in press). Misconceived causal explanations for "emergent" processes. *Cognitive Science*.
- Colella, V.S., Klopfer, D., Resnick, M. (2001). *Adventures in Modeling: Exploring Complex, Dynamic Systems with StarLog*. Willston, VT: Teachers College Press.
- Evans, E.M. (2001). Cognitive and Contextual Factors in the Emergence of Diverse Belief Systems: Creation versus Evolution. *Cognitive Psychology*, 42, 217-266.
- Hmelo-Silver, C.E. & Azevedo, R. (2006). Understanding Complex Systems: Some Core Challenges. *Journal of the Learning Sciences*, 15, 53-61.
- Hmelo-Silver, C.E. & Pfeffer, M.G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science*, 28, 127-138.
- Holland, J. H. (1999). *Emergence: From Chaos to Order*. New York: Basic Books.
- Jacobson, M.J. (2001). Problem solving, cognition, and complex systems: Differences between experts and novices. *Complexity*, 6, 41-49.
- Poling, D. A., & Evans, E. M. (2002). Why do birds of a feather flock together? Developmental change in the use of multiple explanations: Intention, teleology, essentialism. *British Journal of Developmental Psychology*, 20, 89-112.
- Rappoport, L.T. & Ashkenazi, G. (2008). Connecting Levels of Representation: Emergent versus subemergent perspective. *International Journal of Science Education*, 30, 1585-1603.
- Resnick, M. (1996). Beyond the centralized mindset. *Journal of the Learning Sciences*, 5, 1-22.
- Sinatra, G.M., Brem, S.K., & Evans, E.M. (2008). Changing minds? Implications of Conceptual Change for Teaching and Learning about Biological Evolution. *Evolution Education and Outreach*, 2, 189-195.
- Talanquer, V. (2008) Students' predictions about the sensory properties of chemical compounds: Additive versus emergent frameworks. *Science Education*, 92, 96-114.
- Tamir, P. & Zohar, A. (2006). Anthropomorphism and teleology in reasoning about biological phenomena. *Science Education*, 75, 57-67.
- Tisue, S. & Wilensky, U. (2004). NetLogo: A simple environment for modeling complexity. *International Conference on Complex Systems*. Boston.
- Wilensky, U. & Resnick, M. (1999). Thinking in Levels: A Dynamic Systems Approach to Making Sense of the World. *Journal of Science Education and Technology*, 8, 3-19.

# Do I know that you know what you know? Modeling testimony in causal inference

Daphna Buchsbaum<sup>1</sup>, Sophie Bridgers<sup>1</sup>, Andrew Whalen  
Elizabeth Seiver, Thomas L. Griffiths, Alison Gopnik

{daphnab, sbridgers, awhalen, seiver, tom\_griffiths, gopnik}@berkeley.edu  
Department of Psychology, University of California, Berkeley, Berkeley, CA 94720 USA

## Abstract

We rely on both our own observations and on others' testimony when making causal inferences. To integrate these sources of information we must consider an informant's statements about the world, her expressed level of certainty, her previous accuracy, and perhaps her apparent self-knowledge – how accurately she conveys her own certainty. It can be difficult to tease apart the contributions of all these variables simply by observing people's causal judgments. We present a computational account of how these different cues contribute to a rational causal inference, and two experiments looking at adults' inferences from causal demonstrations and informant testimony, focusing on cases where these sources conflict. We find that adults are able to combine social information with their own observations, and are sensitive to the reliability of each. Adults are also sensitive to the accuracy, certainty, and self-knowledge of the informant, a result confirmed by comparing predictions from models with and without these variables.

## Introduction

People face challenging causal learning problems on a daily basis. They have a variety of information they can use to help solve these problems, including directly observed patterns of cause and effect, and social data such as others' statements about existing causal relationships. Having multiple sources available should enhance our causal reasoning, but integration can be difficult, especially when sources disagree. If an informant's causal statements contradict our causal observations, which source should we trust? Informants can be ignorant, mistaken, even deceptive, so one might think we should always trust what we see over what we hear. Yet the world is unpredictable: Observing a phenomenon once does not mean we will reliably observe it again. How do we evaluate these sources and determine which to rely on? We attempt to better understand how people combine different sources of information when making causal inferences, how they integrate their own observations with conflicting testimony, and how this affects their future evaluation of the social informant.

Informant testimony is a key type of social information that guides learning across domains, but the role of testimony in causal learning has not been extensively explored. Research on how we incorporate information from the social context into our causal judgments has shown that both children and adults are skilled causal learners, and can use information from social demonstration to inform their causal inferences (e.g., Kushnir, Wellman, & Gelman, 2008; Sobel & Sommerville, 2009; McGuigan, Makinson, & Whiten, 2011). This work has investigated how we learn by observing other people (e.g., Goodman, Baker, & Tenenbaum, 2009; Buchsbaum, Gopnik, Griffiths, & Shafto, 2011), and by observing different types of people (e.g., Kushnir et al., 2008),

but has not examined in detail how we make inferences about their credibility based on their causal statements. Here, we explore how people combine information from causal observations and testimony, both to make causal judgments and to evaluate the informants themselves.

Recently, there has been a growing literature on how people, especially children, evaluate informants (e.g., Borckardt, Sprohge, & Nash, 2003; Corriveau, Meints, & Harris, 2009; Koenig & Harris, 2005), including how children integrate their prior knowledge with informant testimony (e.g., Jaswal, 2010; Jaswal & Markman, 2007). Integrating testimony with our observations is particularly challenging because multiple aspects of informants and their testimony contribute to the value of the information they provide, and to how much they should be trusted in the future. These aspects include the level of certainty informants express, their past accuracy, and their self-knowledge – how well their certainty reflects their true knowledge (for an exploration of a similar idea in the context of eye-witness testimony see Tenney, Small, Kondrad, Jaswal, & Spellman, 2011). The difficulty of combining information from what someone says and what we see is especially apparent when an informant's statements appear to be incorrect. It could be that the informant is actually right, and our own observations were inaccurate. Alternatively, the informant may be a knowledgeable person who has simply misspoken, or she could truly be clueless. Finally, an informant could express certainty or uncertainty about her knowledge, shedding a different light on her inaccuracy. Deciding whether an informant has erred, and if so, why, and whether to trust her in the future is therefore a complex problem.

Bayesian modeling provides a mechanism for explicitly representing the contributions of different sources of information to judgments about causal structure and informant credibility. Previous work has used such models to explore the role of social observations in causal learning (e.g., Goodman et al., 2009), and to evaluate the role of informant knowledgeability and helpfulness (Shafto, Eaves, Navarro, & Perfors, in press). Here, we present a model that helps us evaluate the roles of observed cause and effect patterns, as well as an informant's expressed certainty, current and past accuracy, and awareness of her own knowledge level, when making a causal inference.

In this paper, we first review a study exploring how preschoolers combine information from informant testimony with conflicting information from observed causal data. Next, we introduce a computational model of causal inference from testimony that explicitly represents the roles of informant certainty, accuracy, and self-knowledge, as well as direct causal observations, allowing us to assess the

<sup>1</sup>These authors contributed equally to this work.

contributions of each to a rational causal inference. We then present a series of adult experiments motivated by both the model and the child experiments. Finally, we conclude by discussing how predictions by models including some or all of these variables provide us with further insight into our ability to learn from multiple sources, and the information we use to determine when to trust what other people say.

### Children’s Causal Inferences from Testimony

Bridgers, Buchsbaum, Seiver, Gopnik, and Griffiths (2011) presented preschoolers with either an informant who claimed to know which of two blocks was better at activating a machine or an informant who claimed to be guessing, and with observed statistical data that contradicted the informant’s claim. The study investigated which source of information (the person or the data) children would rely on, as well as how likely children would be to trust the informant in a new situation. Though both informants made incorrect predictions, the naïve informant demonstrated more self-knowledge because she knew she did not know, while the knowledgeable informant was unaware she was mistaken.

Results from this study imply that preschoolers are sensitive to the certainty and accuracy of an informant – they were more likely to trust the informant’s endorsement over the data when the informant was knowledgeable than when she was naïve, and were more likely to trust the knowledgeable informant before her inaccurate statements than afterwards. However, these results also suggest that children may not be as sensitive to an informant’s level of self-knowledge since children were as likely to trust the knowledgeable informant (who was mistaken in her certainty) as they were to trust the naïve informant (who was correctly uncertain) in a new situation.

Intuitively, an informant’s certainty, past accuracy, and self-knowledge should all be useful indicators of her credibility. However, it is challenging to infer the influence of each of these variables and of the causal data simply by examining people’s resulting inferences. For example, does children’s failure to differentiate between the informants in Bridgers et al. (2011) mean they lack a concept of self-knowledge altogether or that they weigh other cues to reliability more heavily? Given children’s performance, would adults trust a previously inaccurate but uncertain informant over one who was certain and inaccurate, or would they also use a simpler strategy, for instance mistrusting anyone who was previously incorrect? A computational model of how people combine information from both observed data and an informant to determine the likelihood of a causal relationship could help clarify the factors impacting people’s resolution of the conflict, and their decision of whether or not to trust the informant in the future.

### Modeling Causal Inference From Testimony

People may take into account a variety of social information when making causal inferences from testimony. As noted earlier, there is evidence that both children and adults are sensitive to an informant’s expressed certainty and previous

accuracy. People may also have pre-existing assumptions about how knowledgeable others tend to be, and how often others make mistakes in their assertions. Finally, there is some evidence that at least adults are sensitive to others’ self-knowledge (Tenney et al., 2011).

This social information also interacts with the individual’s own causal observations. It can therefore be difficult to determine which of these variables contribute to people’s resulting causal inferences. We present an explicit model of how these variables could interact. We then evaluate the roles of these different variables by comparing people’s causal judgments to those that would be normative under our model, as well as under simpler models that do not explicitly represent the informant’s knowledge and self-knowledge.

Our model is defined in terms of observed variables representing causal outcomes, statements by an informant about the causal strengths of potential causes, and about her level of certainty about her causal knowledge. The model also has hidden variables representing the actual causal strengths of the potential causes, the informant’s general level of knowledgeability, her specific knowledge of the individual causes, and her level of self-knowledge – how well she knows what she knows. We capture the complex relationships among these variables in a graphical model (see Figure 1).

In this model, we assume that all the variables are binary valued, as they were presented to children in the Bridgers et al. (2011) experiments. Each cause  $c$  has a causal strength  $w_c$ , such that  $p(e_c = 1 \mid c, w_c) = w_c$  for effects  $e_{c,i}$  where  $p(w_c = \rho) = \gamma$  and  $p(w_c = 1 - \rho) = 1 - \gamma$ . Here,  $\rho$  is some relatively high probability of effect, corresponding to the causal strength “almost always makes it go” in Bridgers et al. (2011), with  $1 - \rho$  corresponding to “almost never makes it go.”

The informant’s prediction  $r_c$  about the causal strengths of each cause depends on the true causal strength  $w_c$ , and on her knowledge of the cause  $k_c$ . Here, we assume that  $k_c \in \{0, 1\}$ , corresponding to two possible states of knowledge of a cause: guessing and knowing. If  $k_c = 1$  (the informant knows about the causal strength of  $c$ ) then  $p(r_c = w_c \mid k_c = 1, w_c) = 1 - \epsilon$ , meaning an informant with knowledge of cause  $c$  will predict the true value of  $w_c$  with probability  $1 - \epsilon$ , but with small error probability  $\epsilon$  will report the incorrect value. On the other hand, if  $k_c = 0$  and the informant is guessing about cause  $c$ , then  $p(r_c = w_c \mid k_c = 0, w_c) = p(r_c = w_c \mid k_c = 0) = 0.5$ , that is, the informant will choose uniformly at random between the two possible causal strengths.

We assume that the probability of the informant knowing about a particular cause depends on the informant’s global knowledgeability  $g \in \{0, 1\}$ , with the informant having probability  $\kappa$  of being globally knowledgeable. If  $g = 1$  then the informant is globally knowledgeable and  $p(k_c = 0 \mid g = 1) = 1 - \tau$  and  $p(k_c = 1 \mid g = 1) = \tau$ , that is, the informant is knowledgeable about cause  $c$  with some relatively high probability  $\tau$ . Conversely, if  $g = 0$  and the informant is globally ignorant then  $p(k_c = 0 \mid g = 0) = \tau$  and  $p(k_c = 1 \mid g = 0) = 1 - \tau$ .

Finally, we need to represent the informant’s statement



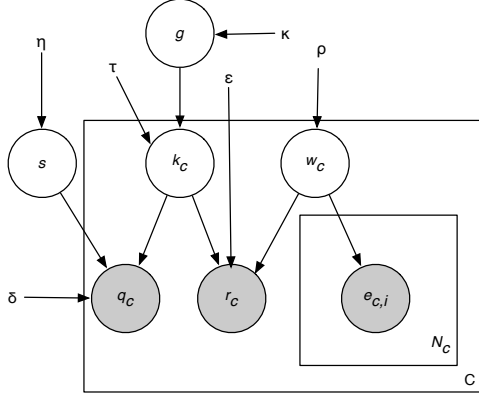


Figure 1: Causal testimony graphical model

about her knowledge of cause  $c$ . The informant’s statement  $q_c$  depends on her knowledge  $k_c$ , and her level of self-knowledge  $s$ . We assume that  $s \in \{0, 1\}$ , corresponding to two possible states of self-knowledge: accurate and inaccurate. If  $s = 1$  (the informant has accurate self-knowledge) then  $p(q_c = k_c | s = 1, k_c) = 1 - \delta$ , meaning the informant will accurately report her level of knowledge  $k_c$  with probability  $1 - \delta$ , but with small error probability  $\delta$  will report her level of knowledge inaccurately.

If  $s = 0$  and the informant has inaccurate self-knowledge then  $p(q_c = k_c | s = 0, k_c) = p(q_c = k_c | s = 0) = 0.5$ , that is, the informant will choose uniformly at random when stating her knowledge of the causal system. We assume that any given informant has probability  $\eta$  of having accurate self-knowledge, and  $1 - \eta$  of having inaccurate self-knowledge.

We assume  $p(\text{choose } c) \propto p(\text{effect} | c, \text{obs})$ , meaning people choose causes in proportion to how likely they think they are to produce the effect, given their observations (including the informant’s statements). This is computable from the model and the dependencies defined in our graphical model (see Figure 1). To evaluate our model, we conducted a series of experiments with adults, exploring whether they can successfully use an informant’s certainty, accuracy, and self-knowledge when making causal inferences.

## Experiment 1: Adult Inferences from Testimony

We investigate how adults resolve a conflict between an informant’s explanation of how a causal system works and actual demonstrations of that system, closely following the procedure of Bridgers et al. (2011). We hypothesized that like preschoolers, adults would be sensitive to the certainty and accuracy of the informant, and be more likely to trust an informant who claimed to be knowledgeable over one who claimed to be naïve, and less likely to trust a previously inaccurate informant. However, unlike children, we predicted that adults would be sensitive to an informant’s level of self-knowledge, and would be more likely to extend their trust to a previously incorrect informant who had claimed ignorance than a previously incorrect informant who had claimed knowledgeability.

## Methods

**Participants** A total of 204 participants were recruited: 100 were UC Berkeley undergraduates who received course credit and 104 were Mechanical Turk workers who were compensated \$0.50. Participants were randomly assigned to one of two experimental groups: the Knowledgeable condition ( $n = 103$ ) or the Naïve condition ( $n = 101$ ).

**Stimuli** The experiment was a web-administered survey involving text and pictures. An image of a brown-haired woman was the informant, and an image of a blonde woman was her assistant. The machine was an image of a green box with a black top. The activated machine had a yellow top and musical notes were placed around it. The blocks were a green rectangle, a pink disk, an orange cube, and a blue cylinder.

**Procedure** First, a woman named Ann (the *informant*) introduced a machine that could light up and play music when certain blocks were placed on top. She then introduced two different blocks and explained that one block almost always activated the machine (the *endorsed* block), while the other block almost never did (the *unendorsed* block). In the Knowledgeable condition, the informant claimed that she really knew which block was better at activating the machine, while in the Naïve condition, the informant claimed that she was just guessing. Besides this difference, the procedure was identical across conditions. Ann then said she needed to leave, and her assistant Jane continued the experiment.

Jane first asked participants to rate how likely each block would be to activate the machine on a scale from 0 (definitely will not) to 10 (definitely will) (the *prior* rating). Jane demonstrated each block on the machine, providing probabilistic evidence that contradicted Ann’s claim: the endorsed block only activated the machine 2/6 times, while the unendorsed block activated it 2/3 times.<sup>1</sup> Participants were then again asked to rate how likely each block would be to activate the machine (the *causal* rating).

Finally, Ann returned with two new blocks, and in both conditions, claimed she *knew* that one block almost always activated the machine and that the other almost never did. Ann then left once more, and Jane asked the participants to rate how likely they thought these new blocks were to activate the machine (the *generalization* rating).

## Results and Discussion

For a summary of the results see Table 1. We analyzed causal efficacy ratings with a  $2 \times 3 \times 2$  repeated measures ANOVA, with endorsement (endorsed or unendorsed), and rating phase (prior, causal, generalization) as the within subject variables, and knowledge condition (Knowledgeable or Naïve) as the between subjects variable. There was a main effect of endorsement – adults rated the endorsed block more highly across phases and conditions ( $F(1, 1206) = 77.69$ ,  $\text{MSE} = 372.1$ ,  $p < 0.001$ ). There was also an effect of endorsement

<sup>1</sup>This pattern of probabilistic data is the same as was used in Experiment 3 of Kushnir and Gopnik (2007).

$\times$  condition ( $F(1, 1206) = 73.19$ ,  $MSE = 350.5$ ,  $p < 0.001$ ), with the endorsed block rated higher in the Knowledgeable condition across phases, and of endorsement  $\times$  phase ( $F(2, 1206) = 62.18$ ,  $MSE = 297.8$ ,  $p < 0.001$ ), with the rating of the endorsed block decreasing and of the unendorsed block increasing in the causal phase. Finally, there was a significant three-way interaction of endorsement  $\times$  phase  $\times$  condition ( $F(2, 1206) = 13.89$ ,  $MSE = 66.5$ ,  $p < 0.001$ ), indicating that the degree to which the ratings change between phases varied by the claimed knowledge level of the informant and whether the block was endorsed.

We explored the particulars of these findings via planned  $t$ -tests. In the prior phase of both conditions, adults were more likely to give the endorsed block a higher rating (paired  $t$ -tests, Knowledgeable:  $t(102) = 17.76$ ,  $p < 0.001$ . Naïve:  $t(100) = 5.02$ ,  $p < 0.001$ ) though participants in the Knowledgeable condition gave the endorsed block a higher rating than those in the Naïve condition (two sample  $t$ -test,  $t(202) = 5.72$ ,  $p < 0.001$ ). These results suggest that before seeing any data, adults in both conditions were likely to trust the informant’s testimony, but were also sensitive to the certainty expressed by the informant.

In the causal phase, there was no difference in adults’ ratings of the endorsed and unendorsed blocks in the Knowledgeable condition (paired  $t$ -test,  $t(102) = 1.39$ ,  $p = 0.17$ ), while adults in the Naïve condition gave the unendorsed block a higher rating (paired  $t$ -test,  $t(102) = 7.18$ ,  $p < 0.001$ ). Adults in the Knowledgeable condition gave the endorsed block a higher rating than adults in the Naïve condition (two sample  $t$ -test,  $t(202) = 2.00$ ,  $p < 0.05$ ) and vice versa for the unendorsed block (two sample  $t$ -test,  $t(202) = 3.62$ ,  $p < 0.001$ ). The fact that in the causal phase adults thought the two blocks had approximately equal causal efficacy in the Knowledgeable condition but rated the unendorsed block more highly in the Naïve condition suggests that participants were responding to both the observed statistical data and the claimed knowledge level of the informant.

Finally, in the generalization phase, adults in both conditions gave the endorsed block higher ratings (paired  $t$ -tests, Knowledgeable:  $t(102) = 13.21$ ,  $p < 0.001$ . Naïve:  $t(102) = 8.23$ ,  $p < 0.001$ ). Unlike the previous phases, there was no difference between conditions in ratings of the endorsed block (two sample  $t$ -test,  $t(202) = 0.82$ ,  $p = 0.41$ ).

We can also compare ratings in the two no-data phases – prior and generalization – to capture how participants’ evaluation of the informant might have changed after receiving evidence about her accuracy in the intervening causal phase. Adults’ ratings decrease between prior and generalization in the Knowledgeable condition (paired  $t$ -test  $t(102) = 3.30$ ,  $p < 0.01$ ), while they increase in the Naïve condition (paired  $t$ -test,  $t(100) = 2.12$ ,  $p < 0.05$ ). This difference suggests that adults may actually be sensitive to an informant’s self-knowledge, increasing their trust in an informant who was incorrect but uncertain in the past over an informant who was incorrect but certain. Our results in the generalization phase

Table 1: Mean ratings and standard errors for Experiment 1

Mean Rating (std err)	Endorsed	Unendorsed
Prior Naïve	6.46 (.24)	4.16 (.25)
Prior Knowledgeable	8.22 (.19)	1.94 (.19)
Causal Naïve	4.09 (.18)	6.29 (.20)
Causal Knowledgeable	4.66 (.22)	5.18 (.23)
Gen Naïve	7.20 (.25)	3.27 (.25)
Gen Knowledgeable	7.45 (.18)	2.71 (.20)

Table 2: Mean ratings and standard errors for Experiment 2

Mean Rating (std err)	Endorsed	Unendorsed
Prior Naïve	6.58 (.31)	3.68 (.32)
Prior Knowledgeable	8.61 (.18)	1.47 (.18)
Causal Naïve	1.15 (.35)	9.45 (.11)
Causal Knowledgeable	2.82 (.54)	8.0 (.45)
Gen Naïve	7.0 (.44)	3.47 (.47)
Gen Knowledgeable	5.88 (.52)	4.15 (.51)

imply that adults are willing to trust both informants more or less equally regardless of their level of self-knowledge. However, due to the stochastic nature of the data, participants may have made excuses for the discrepancy between the data and the knowledgeable informant’s endorsement, possibly appealing to hidden causes such as a faulty machine part that would explain away the conflict. In Experiment 2, we contrasted an informant’s testimony with deterministic data to see if increasing the apparent inaccuracy of the informant would reveal a use of informant self-knowledge.

## Experiment 2: Deterministic Data

We replicated Experiment 1 but with deterministic data, to explore how changing the strength of the data might impact adults’ inferences. We predicted that adults would weight conflicting deterministic data more heavily than conflicting probabilistic data, and would therefore prefer the unendorsed block more often in the causal phase. We also predicted that the stronger data would exaggerate the knowledgeable informant’s lack of self-knowledge, leading adults to consider the naïve informant’s testimony as more reliable than the knowledgeable informant’s in the generalization phase.

## Method

**Participants** A total of 74 participants recruited from Mechanical Turk were compensated \$0.50 and randomly assigned to the Knowledgeable condition ( $n = 34$ ) or the Naïve condition ( $n = 40$ ).

**Stimuli** Stimuli were identical to those in Experiment 1.

**Procedure** The procedure was the same as in Experiment 1 except that the endorsed block activated the machine 0/6 times, while the unendorsed block activated it 6/6 times.

## Results and Discussion

We performed the same analyses for Experiment 2 as we did for Experiment 1. Due to limited space, we discuss only the most relevant results here. Results are summarized in Table 2.

Adults in both conditions of Experiment 2 gave lower ratings to the endorsed block in the causal phase as compared to Experiment 1 (two sample t-tests. Knowledgeable:  $t(135) = 3.77$ ,  $p < 0.001$ . Naïve:  $t(139) = 7.90$ ,  $p < 0.001$ ). Adults thus were sensitive to the increased strength of the data in the second experiment, and were less likely to trust the knowledgeable informant’s claim over the data than participants who observed probabilistic data.

In the generalization phase of Experiment 2, adults rated the endorsed block in the Naïve condition as more causally efficacious than in the Knowledgeable condition, though this effect was only marginal ( $t(72) = 1.69$ ,  $p < 0.10$ ). Thus, adults in the Knowledgeable condition were less inclined to trust the informant’s endorsement than adults in the Naïve condition. This finding suggests that differences in self-knowledge *do* impact adults’ evaluations of informants since adults appeared to place more confidence in the statement of the person whose prior certainty reflected their accuracy.

Comparing across experiments, we found that adults in the Knowledgeable condition of Experiment 1 gave the endorsed block a higher generalization rating than those in Experiment 2 (two sample t-test.  $t(135) = 3.77$ ,  $p < 0.001$ ). On the other hand, there was no difference in adults’ generalization ratings of the endorsed block in the Naïve condition (two sample t-test.  $t(135) = 0.41$ ,  $p = 0.68$ ) across experiments. As predicted, increasing the strength of the conflicting data magnified the knowledgeable informant’s inaccuracy. However, since the naïve informant claimed ignorance, this change did not affect how adults evaluated future information from this informant. In general, trust in the knowledgeable informant decreased with increasingly conflicting data (i.e. from situations of no conflict (prior phases) to situations of prior conflicting data (generalization phases) to situations of directly conflicting data (causal phases)).

Finally, comparing between phases of Experiment 2, in the Knowledgeable condition, adults’ ratings of the endorsed block decrease between prior and generalization phases (paired t-test,  $t(33) = 4.99$ ,  $p < 0.001$ ). Thus, adults are likely to initially trust the endorsement of an informant, whereas they are less likely to extend that trust to a new situation (the generalization phase) after observing evidence that contradicts the informant’s prior claim. Conversely, in the Naïve condition there was no difference between adults’ prior and generalization phase ratings (paired t-test,  $t(39) = 0.79$ ,  $p < 0.43$ ). Thus, adults’ evaluation of the credibility of the naïve informant does not appear to have changed after observing the conflicting data. This further suggests adults are sensitive to self-knowledge when determining the usefulness of an informant’s statement: They found a previously uncertain, inaccurate informant more trustworthy than a previously certain, inaccurate one.

One alternative account is that adults consider the knowledgeable informant to be deceptive rather than having poor self-knowledge. However, if adults suspect the knowledgeable informant is deceiving them by saying the opposite

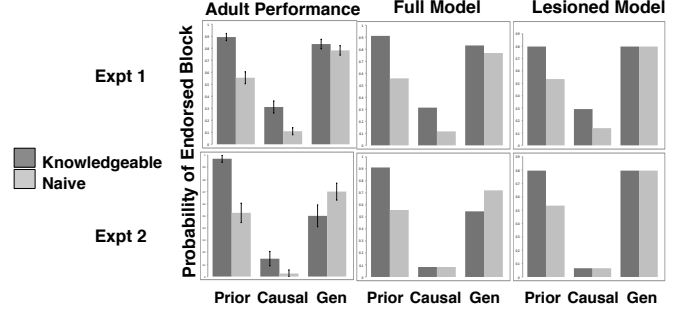


Figure 2: Adult performance vs model predictions

of what she knows, then we would expect them to go against her endorsement in the generalization phase, rating the unendorsed block as more causally efficacious than the endorsed block. However, while adults’ ratings of the endorsed block decrease in the generalization phase, they are still greater than their ratings of the unendorsed block, implying that adults view this informant as less reliable rather than intentionally dishonest.

## Modeling People’s Inferences

We can use our model to further test whether an informant’s expressed certainty, accuracy, and apparent self-knowledge inform adults’ causal judgments. We first evaluate the model by fitting it to data from Experiment 1. In order to be consistent with Bridgers et al. (2011), where children were asked to choose the better block, we assumed that for each set of ratings adults would choose the block they had rated as most likely to be effective. We then optimized the log likelihood of these choices under the model. The best fitting model corresponded to model parameters of  $\rho = 0.91$ ,  $\gamma = 0.40$ ,  $\epsilon = 0.01$ ,  $\delta = 0.1$ ,  $\tau = 0.91$ ,  $\kappa = 0.74$  and  $\eta = 0.93$ , (Pearson’s  $r=0.999$ ,  $p < 0.001$ ). However, a reasonable range of values around these settings also fit the data well. Of particular note are the values of  $\kappa$  and  $\eta$ , corresponding to a belief that most people have good general knowledge, but a substantial minority are relatively clueless, and that almost everyone is aware of their own knowledge level, but a small number of people tend to inaccurately assess what they know.

We tested the generalization of these model parameters by looking at how well they predict adult performance in Experiment 2. The parameters fit to the Experiment 1 data also provide a good fit to the results of Experiment 2 (Pearson’s  $r=0.9664$ ,  $p < 0.001$ ). This suggests that our model is accurately capturing human performance, so we can use it to tease apart the contributing variables to adult causal inferences.

We then conducted a nested model comparison, examining whether including the informant’s global knowledgeability and their self-knowledge adds explanatory value to the model, by creating a series of “lesioned” models, lacking global knowledge  $g$  and self knowledge  $s$ . This approach controls for the additional free parameters in the more complex model. Removing global knowledge corresponds to a model that assumes that all informants have the same probability of knowing about all causes, and that an in-

formant knowing about one cause does not make her any more likely to know about others. Removing self-knowledge corresponds to a model where informants' statements of certainty always reflect their true knowledge – if they say they know something, then they must really know it.

Compared to a model lacking both global knowledge and self-knowledge variables, adding global knowledge to the model resulted in a marginally significant ( $\chi^2(1) = 3.29, p = .07$ ) improvement in model fit. Adding self-knowledge on its own did not improve model performance ( $\chi^2(1) = 0.392, p = 0.53$ ), however adding both self-knowledge and global knowledge variables significantly improved model fit over having only global knowledge ( $\chi^2(1) = 22.04, p < 0.001$ ), or having neither ( $\chi^2(1) = 25.30, p < 0.001$ ) (see Figure 2).

Qualitatively, the addition of global knowledge and self-knowledge only modestly improves the model fit to Experiment 1, their biggest effect is on the fit to Experiment 2. Of particular interest in Experiment 1, the full model and the “lesioned” model (without both variables) appear to make similar predictions about adults' performance in the generalization phase, suggesting that contrary to our initial intuitions, even with a concept of self-knowledge it may still be rational to extend trust equally to both informants. However, these two models make different predictions for generalization performance in Experiment 2, with the full model more accurately capturing adults' inferences. This supports the interpretation that in the first experiment, participants continued to extend trust to the knowledgeable informant not because they lacked a concept of self-knowledge but by explaining away the informant's apparent incorrectness, inferring that the ambiguous data could have been “unlucky.” However, in Experiment 2, where the data more strongly supports the inference that the informant was incorrect, these model results suggest that it requires both a concept of general knowledge (“if this person was wrong before, they're more likely to be wrong again”), and self-knowledge (“they said they ‘knew’ before and they didn't, why should I think they know now?” vs. “they said they didn't know, so it's okay that they were wrong”), in order to infer that the naïve informant is more deserving of trust in the generalization phase.

Overall, our nested model comparison demonstrates that adults take into account the informant's past performance when deciding how much to weight their current testimony, and in particular that adults are sensitive to both the apparent knowledgeability of the informant and their level of self-knowledge.

## Conclusion

We examined how people combine an informant's statements about a causal system with direct observations of that system, and how this influences their evaluation of the informant's knowledgeability and credibility. Together, Experiments 1 and 2 suggest that adults are weighting and integrating evidence from both observed data and the informant in their causal inferences, and that their trust in the informant is moderated by the degree to which the informant's testimony

conflicts with the data. Adults were sensitive to both the informant's certainty and accuracy, and to how well the informant's certainty reflected her accuracy. These findings support our intuition that self-knowledge is a valuable cue adults can use to determine the trustworthiness of an informant's testimony. The close fit of the model to adult performance, and its ability to generalize from Experiment 1 to Experiment 2, confirms that adults are rationally integrating their direct observations with testimony from a social informant when making causal inferences. Our model also strongly suggests that representing self-knowledge is necessary to making these inferences, and that adults could not be using a simpler strategy such as only tracking previous inaccuracy. Overall, these results provide us with further insight into how we learn from and evaluate the sources of information available to us and in particular, revealing that knowing that you do not know can be just as important as knowing that you know.

**Acknowledgments.** This work was supported by grant number FA-9550-10-1-0232 from the Air Force Office of Scientific Research, the NSF Graduate Research Fellowship, NSF Grant BCS-1023875, and the McDonnell Foundation Causal Learning Initiative.

## References

- Borckardt, J., Sprohge, E., & Nash, M. (2003). Effects of the inclusion and refutation of peripheral details on eyewitness credibility. *Journal of Applied Social Psychology, 33*(10), 2187–2197.
- Bridgers, S., Buchsbaum, D., Seiver, E., Gopnik, A., & Griffiths, T. L. (2011). Which block is better at making the machine go?: How children balance their trust in an informant vs. the data. *Poster presented at Biennial Meeting of the Cognitive Development Society.*
- Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition, 120*(3), 331–340.
- Corriveau, K., Meints, K., & Harris, P. (2009). Early tracking of informant accuracy and inaccuracy. *British Journal of Developmental Psychology, 27*(2), 331–342.
- Goodman, N. D., Baker, C. L., & Tenenbaum, J. B. (2009). Cause and intent: Social reasoning in causal learning. *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*
- Jaswal, V. K. (2010). Believing what you're told: Young children's trust in unexpected testimony about the physical world. *Cognitive Psychology, 61*, 248–272.
- Jaswal, V. K., & Markman, E. M. (2007). Looks aren't everything: 24-month-olds' willingness to accept unexpected labels. *Journal of Cognition and Development, 8*(1), 93–111.
- Koenig, M., & Harris, P. (2005). The role of social cognition in early trust. *Trends in Cognitive Sciences, 9*(10), 457–459.
- Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology, 43*(1), 186–196.
- Kushnir, T., Wellman, H., & Gelman, S. (2008). The role of preschoolers' social understanding in evaluating the informativeness of causal interventions. *Cognition, 107*(3), 1084–1092.
- McGuigan, N., Makinson, J., & Whiten, A. (2011). From over-imitation to super-copying: Adults imitate causally irrelevant aspects of tool use with higher fidelity than young children. *British Journal of Psychology, 102*, 1–18.
- Shafto, P., Eaves, D., Navarro, D. J., & Perfors, A. (in press). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science.*
- Sobel, D. M., & Sommerville, J. (2009). Rationales in children's causal learning from others' actions. *Cognitive Development, 24*(1), 70–79.
- Tenney, E. R., Small, J. E., Kondrad, R. L., Jaswal, V. K., & Spellman, B. A. (2011). Accuracy, confidence, and calibration: How young children and adults assess credibility. *Developmental Psychology, 47*(4), 1065–1077.

# Revisiting the Relationship between Allocentric-Heading Recall and Self-Reported Sense of Direction

Heather Burte (Burte@psych.ucsb.edu)

Department of Psychology, UCSB  
Santa Barbara, CA 93106 USA

Mary Hegarty (Hegarty@psych.ucsb.edu)

Department of Psychology, UCSB  
Santa Barbara, CA 93106 USA

## Abstract

Sense-of-direction (SOD) has been described as a system that tracks the body's facing direction relative to an environmental reference frame (allocentric heading). To study this system, Sholl, Kenny, and DellaPorta (2006) developed a heading-recall task and found that task accuracy correlated highly with self-reported SOD measures. This study attempts to replicate and extend their findings, by increasing task accuracy, and testing alternative hypotheses about factors that could affect task performance. In a heading-recall task, participants estimated allocentric heading from pictures of familiar locations on a college campus. Previous results were replicated, but a weaker relationship between SOD and performance, and a novel relationship between location familiarity and performance were found. These results provide support for a human allocentric heading system but suggest that self-reported SOD potentially measures a range of abilities and not solely the operation of this system.

**Keywords:** allocentric heading; sense of direction; spatial orientation; spatial memory; head-direction cells; heading-recall.

## Introduction

In everyday situations, people describe their ability to accurately navigate through cities or neighborhoods using phrases like 'I have a great sense-of-direction' or 'I lack a sense-of-direction'. Kozlowski and Bryant (1977) transformed these colloquial assessments into a 7-point scale which assessed sense-of-direction (SOD). They found that these assessments were related pointing ability to familiar landmarks and updating one's location while traveling in an underground maze. Kozlowski and Bryant used a single item scale: "How good is your sense-of-direction?" Other researchers have measured SOD in a multi-faceted way. For example, the Santa Barbara Sense of Direction scale (SBSOD) is a 15-item scale that asks people about a variety of environmental tasks, such as giving directions and estimating distances, as well as their "sense-of-direction" (Hegarty, Richardson, Montello, Lovelace, & Subbiah, 2002). Hegarty et al. found that this measure of self-assessed SOD is related to several different environmental-scale tasks, including learning the layout of a new place, blindfolded updating, and pointing to familiar landmarks. These environmental-scale tasks require locomotion and integration from multiple viewpoints to acquire and access spatial knowledge. As such, the SBSOD

scale was created around the idea that one's SOD is multi-faceted.

Recently, Sholl, Kenny, and DellaPorta (2006) proposed that SOD is single-faceted, and relates to the performance of a head-direction system in humans, similar to that found in animals. The head-direction system in rats was first discovered by Ranck (1984), who identified brain cells that fire when an animal's head is facing a specific direction. The directions that these cells respond to are not directions based on the axis of the body (also called egocentric headings). They respond to the angles between the forward axis of the body and a reference direction that is grounded in the environment (i.e. the animal's allocentric heading). An example of one allocentric reference system is the cardinal directions, but head-direction cells use the environment's intrinsic structure, not cardinal directions.

Sholl, et al.'s (2006) goal was to discover if humans have an allocentric-heading system that is functionally similar to the head-direction system of animals and to elucidate the functional architecture of this system, including its inputs, outputs, organization, representations, and computations. To accomplish that goal, they developed an allocentric-heading recall task in which students were shown a picture of a familiar landmark on their campus, and had to indicate the direction (with respect to the global environment) from which the photo was taken. They found that a person's current facing direction influences their accuracy and decision latency in recalling allocentric headings: when a person's facing direction matches the allocentric direction to be recalled, there is a facilitation effect; and, when the facing direction is 180° from the allocentric direction to be recalled, there is a detrimental effect. According to the author's, one's current body-direction signals interfere with retrieval of allocentric-headings being remembered from other locations, at which one's body-direction signals were different. These results would be predicted if the human allocentric-heading system works similarly to the animal head-direction system. Sholl et al. also found strong correlations between performance on the heading recall task and both Kozlowski & Bryant's (1977) single-item question (K&B) and the SBSOD. They proposed that SOD measures a single-faceted ability, which reflects the operation of a human head-direction system.

To expand upon Sholl et al.'s (2006) findings, the goal of

this paper is to replicate Sholl et al.'s findings in a different location, test new hypotheses, and to provide further evidence on the single- or multi-faceted nature of what is measured by self-reported SOD measures.

### **Allocentric-Heading Recall and SOD**

The heading-recall task used by Sholl et al. (2006) was a four-alternative, forced-choice task, using campus pictures as stimuli. The pictures were taken from magnetic north, east, south or west. Magnetic compass directions were used because the intrinsic structure of the Boston College campus is aligned as such (and will also be used in the following experiment as the UCSB campus is similarly aligned). However, while cardinal directions will be used for simplicity in writing this article, it should be noted that cardinal directions were never used in written or verbal instructions, as the task can be completed without using cardinal directions.

First, we will define key terminology used: picture heading is the photographer's orientation when taking the picture; default heading is the orientation of participant before each trial; response heading is the participant's response orientation that s/he moved to, decision latency is the participant's time to decide on a response heading and rotation time is the time taken to rotate from the default to the response heading.

In the heading-recall task, participants were asked to indicate picture heading by rotating in a chair. According to Sholl et al. (2006), when viewing a building, the allocentric-heading of that view is stored in memory and is linked to signals of body-direction. Upon seeing a picture of that building, a person recognizes the building, and then recalls the allocentric-heading from spatial memory. Therefore, participants can rotate in a chair to replicate the picture heading because they can compare their current body-direction to the body-direction signals from their memory and move to face the picture heading. The two main measures of this task were accuracy and decision latency. Participants responded more accurately and faster when the picture heading was consistent with their default heading; therefore, accuracy increased and decision latency decreased with increasing angular deviation between the default and picture heading.

Sholl et al. (2006) found that self-reported SOD was related to accuracy in the heading-recall task, especially at the extremes of the SOD scale, and concluded that SOD measures reflect people's awareness of their own allocentric-heading abilities. In their first experiment, accuracy in heading-recall was correlated .74 with the SBSOD and .68 with the K&B scale. However, the SOD scales were administered at the end of the study and so participants' self-assessed SOD ratings could have reflected an assessment of their performance on this task, rather than a more general assessment of their abilities (cf. Heth, Cornell, & Flood, 2002). Thus, the correlations might be inflated. In our study, participants completed the SOD scales before the heading-recall task.

Another concern is that some of Sholl et al.'s, (2006) participants performed very poorly on the heading-recall task, with only 18/40 participants in Experiment 1 and 10/19 participants in Experiment 2 surpassing a 50% accuracy rate. Low accuracy could reflect failure to understand the task, because the heading-recall task is abstract, unlike everyday directional tasks. In fact, Sholl et al. reported instructional difficulties. Thus, the high correlations with SOD measures may reflect the fact that those with poor SOD were unable to understand the task.

The goals of this study are (1) to replicate the results of the heading task in a new context, (2) to maximize accuracy, (3) to reassess the relationship between self-assessed SOD and allocentric-heading recall, and (4) to test two alternative hypotheses. First, this study serves to replicate the methods used by Sholl et al. (2006) and confirm that their experimental effects are robust with differing campus locations, target pictures and participants. Second, we attempt to maximize accuracy in the heading-recall task by offering more practice trials and feedback to participants in the instruction phase, to ensure that participants understood the task. Third, this study reassesses the relationship between SOD and heading-recall when measures of SOD are taken before the heading-recall task rather than after.

Fourth, two alternative hypotheses were tested. The first alternative hypothesis was that performance on the heading task would be correlated with familiarity. Sholl et al. (2006) found no correlations of performance on the heading recall task with familiarity of the landmarks or distance to target. They used landmarks that had been rated as highly familiar by other students, but did not assess familiarity in the context of their experiment. With regards to familiarity, if one must recognize the target before the allocentric-heading can be retrieved from memory, then familiarity might be related to performance on the heading-recall task. Other studies have found that familiarity predicted directional accuracy on a mental wayfinding task (Prestopnik & Roskos-Ewoldsen, 2000). Therefore, our experimental participants completed a familiarity rating task, to test the hypothesis that familiarity is related to accuracy and decision latency.

The second alternative hypothesis concerns unfamiliar targets or targets for which participants cannot retrieve an allocentric-heading straight from memory. In these cases, people might perform the heading task by imagining walking a route from the experiment location to the target location. In this case, target distance should be correlated with decision latency. This prediction is based on the assumption that mental route taking is an analog process similar to mental rotation (Shepard & Metzler, 1971). Just as participants take longer to mentally rotate with larger angles, so might participants take longer to calculate allocentric-heading with larger distances, if they use a mental walk strategy. Just and Carpenter (1985) found that participants with poor spatial abilities rotated at a slower rate than those with good spatial ability, so the relationship between distance and decision latency might be particularly

strong for poor SOD participants. Sholl et al. (2006) failed to find correlations between objective distance and decision latency, but increasing the task understanding of poor SOD participants might reveal these participants' use of the "mental walk" strategy. Therefore, we tested the hypothesis that participants' estimated distances for each landmark would be related to decision latency on the heading-recall task.

## Method

**Pretesting of Stimuli** Twenty students (8 males and 12 females) rated 124 photographs of the University of California, Santa Barbara (UCSB) campus for familiarity and confidence in knowing the location from which the photograph was taken. The photographs were taken from four different headings (facing north, south, east and west). On the basis of this pretesting, 40 photographs (10 from each heading) were selected for the main study. The selected photographs did not differ in familiarity or rated confidence of location across headings. The ratings of familiarity were similar to those reported by Sholl et al., (2006) with the grand mean for the forty photographs being 1.6 on a scale from 1 to 7 with 1 being "Very familiar" and 7 being "Very unfamiliar".

**Participants** Sixty-one students (29 males and 32 females) participated in the main experiment to fulfill a research participation requirement. Participants were required to have spent at least two full quarters on campus. Each participant was assigned to one of the default headings (N, S, E, or W).

**Materials** The experiment took place in a room on campus that was aligned with the main axes of the campus (and the cardinal directions). The experimental room had one west-facing window that was open during the experiment. The view directly out that window was of a courtyard and another large (three storey) adjacent building. However, if one stood next to the window, one could see the mountains and ocean (major orientation markers for the campus), and a few major buildings. Therefore, the window afforded excellent views for initial orientation to the campus (when standing by the window), but only basic information while participants were seated at a desk when completing the experimental tasks.

Markers on the floor denoted four cardinal directions (which were also the default and response headings). Experimenters arranged a swivel chair and desk towards the assigned default heading before the participant arrived. Assigned default headings are used to determine if a participant's actual heading differentially affected the retrieval of picture headings.

A trial started with viewing a photograph of campus on a computer. Participants determined the direction (with respect to the campus environment) in which the photographer stood when taking the photograph (i.e., picture heading) and turned in the chair to reproduce that orientation. For example, if the photograph was taken facing

south, and the participant's default heading was facing north, the participant should turn 180° to face south. In addition to accuracy in completing this task, latency (time to complete the task) was recorded. Latency was recorded by computer and by the experimenter using a stop-watch, so that decision latency and rotation latency could be separately calculated. Both the computer and experimenter started timing when the picture was shown to the participant. The computer stopped timing when the participant indicated via a button press that s/he was about to turn (decision latency). Then the participant turned and indicated to the experimenter when s/he had finished rotating (total latency). The rotation latency was acquired by subtracting decision latency from the total latency. Participants were asked to rotate using the shortest angle.

**Design.** The methodology of the study was both experimental and correlational. The experimental factors were picture heading (within subjects) and default heading (between subjects). Participants were randomly assigned to one of the four default headings and completed forty trials, ten for each picture heading. Accuracy and latency were correlated with self-assessed SOD, average familiarity and accuracy of distance estimates.

**Procedure** Participants were introduced to the experiment, completed a demographics questionnaire, and completed the K&B and SBSOD rating scales. Next, participants were asked to orient to the layout of campus while looking out the window. The experimenter pointed to major points-of-reference (ocean and mountains) and then asked the participant to point towards four major campus buildings, to ensure that s/he was oriented to the global layout of the campus. The experimenter provided feedback, if needed, but most participants oriented and pointed correctly.

Participants were then introduced to the heading-recall task and presented with twelve practice trials in a fixed order. Participants were given feedback, and told the correct answer for any incorrectly answered trials. Next, the forty experimental trials were completed without feedback.

Afterwards, participants completed a distance estimation task, in which they estimated straight-line distances from their current location to the forty photograph locations, using a visually-presented standard unit (20 meters in length). Participants were given two practice distance estimation trials with correct answers provided as feedback. Then the task was completed for all forty photographs without feedback. Finally, participants rated their familiarity with each photograph location on a 7-point Likert scale.

The major procedural differences from Sholl et al. (2006) were that (1) more detailed instructions were provided, (2) more practice heading-recall trials were given, (3) the SOD scales were answered before the heading-recall task rather than after, and (4) distance estimation and familiarity tasks were used to test alternative hypotheses.



## Results

**Pretest and experimental photograph familiarity** The familiarity ratings for pictures from the pretest and from the main experiment correlated significantly,  $r(38) = .63$ ,  $p < .001$ . The mean familiarity in the main experiment was 2.3, which ranged from 1.1 to 3.9 across participants and from 1.1 to 5.3 across pictures. Even though there was a strong correlation between the two familiarity measures, participants in the main experiment rated some pictures as unfamiliar. Four pictures (three east facing and one north facing) had familiarity ratings that exceeded 2.5  $SD$  above the mean familiarity and were removed from analyses.

**Accuracy** To aggregate across default-heading conditions, a new variable called heading disparity was created to denote the angle disparity between the default heading and the picture heading for the four different default headings. For example, if the picture heading is aligned with the default heading, then these responses would be labeled as  $0^\circ$  heading disparity. A 2 (Gender) X 4 repeated measures (Heading disparity:  $0^\circ$ ,  $90^\circ$ ,  $180^\circ$ ,  $270^\circ$ ) ANOVA comparing mean accuracy indicated a main effect of heading disparity,  $F(3,177) = 7.73$ ,  $MSE = .22$ ,  $p < .001$ . The mean correct proportions by heading disparity are shown in Figure 1. Post hoc tests revealed that the  $180^\circ$  condition was the least accurate, the  $90^\circ$  condition was less accurate than the  $0^\circ$  condition and the  $270^\circ$  condition was midway between  $0^\circ$  and  $90^\circ$ . This can be interpreted as an inhibitory effect of having one's body positioned  $180^\circ$  away from the response of one's head-direction system, and is predicted if the human head-direction system operates similarly to that of animals. There were no other main effects or interactions.

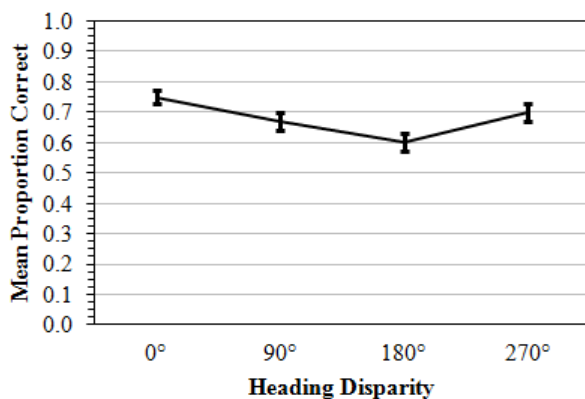


Figure 1: Mean accuracy rate as a function of heading disparity. Error bars are the standard errors of the mean.

To further examine the conditions that lead to the previous analysis, a 2 (Gender: male, female) X 4 (Default heading: N, E, S, W) X 4 repeated measures (Picture heading: N, E, S, W) ANOVA compared mean accuracy. A main effect of picture heading was found,  $F(3,159) = 20.62$ ,  $MSE = .45$ ,  $p < .001$ , with north-facing pictures ( $N = 78\%$ ) and west-facing pictures ( $N = 74\%$ ) having the highest accuracy, south ( $N = 66\%$ ) with moderate accuracy and east

( $N = 58\%$ ) with the lowest accuracy. The mean proportion correct by picture and default heading is shown in Figure 2. This main effect was qualified by an interaction of picture heading with default heading,  $F(9,159) = 3.44$ ,  $MSE = .08$ ,  $p = .001$ . Accuracy was highest when picture and default headings were aligned and lowest when the default and picture headings were misaligned by  $180^\circ$ . Supporting the previous analysis, these findings confirm our main finding that your current heading affects the accuracy with which you can recall allocentric-heading. Aligned headings are easier to recall and  $180^\circ$  unaligned heading are harder to recall. There were no other main effects or interactions.

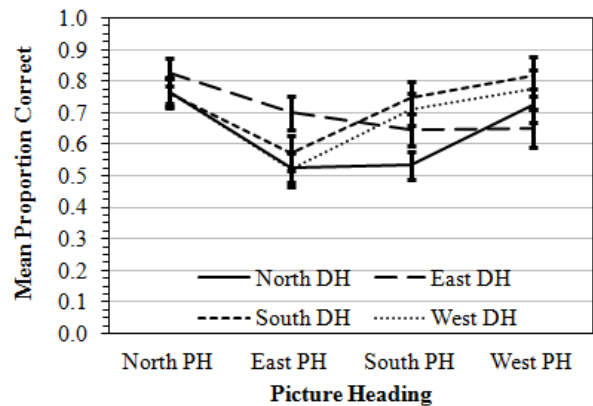


Figure 2: Mean accuracy rate as a function of picture heading (PH) and default heading (DH). Error bars are the standard errors of the mean.

However, there were some qualifications to this result, (1) with north picture headings there was no difference between default headings, and (2) west picture headings were highly accurate for north and south default headings. These findings might have been particular to the campus used as there are large global orientation cues, such as the local mountains when facing North and Isla Vista (an undergraduate housing area) when facing West. In debriefing, some participants reported using heuristics such as determining if the picture heading faced the mountains or Isla Vista.

**Rotation time** A one-way, repeated measures ANOVA investigated the effects of turn magnitude on rotation times. There was a significant linear trend,  $F(1,59) = 145.23$ ,  $MSE = 19.42$ ,  $p < .001$ , with rotation times of 1.56, 2.10, and 2.36 seconds for rotations  $0^\circ$ ,  $90^\circ$  and  $180^\circ$ , respectively. The magnitude of the turn accounted for 71.1% of the variability in turn time, indicating that decision latency was successfully separated from the time to physically turn.

**Decision latency** Outliers greater than 2.5  $SD$  above each participant's mean correct decision latency (3.3% of trials) were recoded to the mean and participants with less than 50% accuracy on the direction task were removed from all decision latency analyses. This was done, as there would be too few decision times to provide a meaningful measure for these participants. Fifty of the 61 participants (82%, 26



male, 24 female) had more than 50% accuracy in the present experiment, in contrast with only 18 of 32 (56%) participants in Sholl et al.'s first experiment. Thus, performance was generally more accurate in the current study.

A 2 (Gender) X 4 repeated measures (Heading disparity: 0°, 90°, 180°, 270°) ANOVA indicated a marginal main effect of heading disparity,  $F(3,144) = 2.45$ ,  $MSE = 7.71$ ,  $p = .07$ . Post hoc tests revealed the 180° and 90° heading-disparity conditions had longer decision latencies than the 0° condition. While suggestive, this pattern is not exactly what is predicted by the animal model and it is only marginally significant, but interestingly, a similar pattern was observed by Sholl et al. It might be due to the specific environments used in both experiments. The mean decision latency as a function of heading disparity is shown in Figure 3.

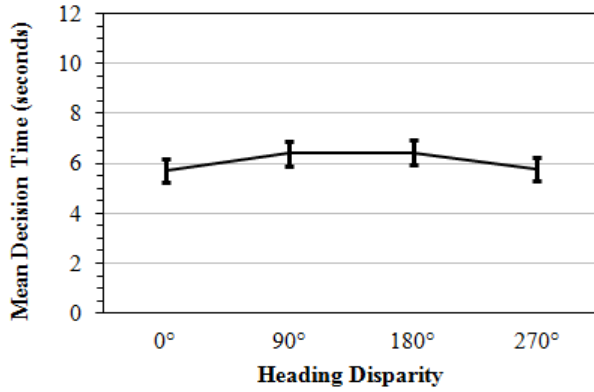


Figure 3: Mean decision latency as a function of heading disparity. Error bars are the standard errors of the mean.

**Correlations with Self-Reported Sense of Direction** The correlation between the two self-reported sense of direction measures was  $r(59) = .60$ ,  $p < .001$ . Mean accuracy and mean correct decision latency were correlated with participants' SOD measures, as shown in Table 1. Both of the SOD measures were positively correlated with mean accuracy and negatively correlated with mean correct decision latency, as expected. However, the correlations were substantially lower than those found by Sholl et al. (2006), and were significant only for accuracy.

**Photograph familiarity** Significant correlations between participants' mean familiarity rating (averaged across the 36 pictures) and their mean accuracy on heading-recall were found,  $r(59) = -.40$ ,  $p < .001$ , supporting our familiarity hypothesis. Therefore, as familiarity approaches 1 for 'very familiar', accuracy increases, however, there was no significant correlation between participants' familiarity and decision latency. Correlating mean familiarity per picture (averaged across individuals) with mean accuracy and decision latency per picture resulted in significant correlations (Table 2).

As seen in Table 1, SOD measures were negatively correlated with participants' mean ratings of familiarity of the landmarks,  $r_{K\&B}(59) = -.34$ ,  $p < .01$ ,  $r_{SBSOD}(59) = -.27$ ,

$p < .05$ , indicating that participants with good SOD rate their familiarity closer to 1 for 'Very familiar'. Furthermore, if one controls for mean landmark familiarity, correlations between accuracy and the SOD measures drop, to  $r_{K\&B}(59) = .19$ ,  $p = .14$  and  $r_{SBSOD}(59) = .29$ ,  $p < .05$ . In sum, the only significant correlation after controlling for familiarity is between accuracy and SBSOD. Therefore, the strong correlation between accuracy and SOD is partially due to high SOD participants being more familiar with the buildings shown in the pictures. In contrast with Sholl et al.'s conclusion that SOD reflects only ability to recall the allocentric heading of the picture, we have found that SOD is related to familiarity. Thus failure to recognize the locations from the pictures may be a source of error in this task.

Table 1: Mean Accuracy and Correct Decision Latency Correlations with Sense of Direction Ratings.

	Accuracy <sup>a</sup>	Decision Latency <sup>b</sup>	Familiarity <sup>a</sup>
K&B	.30*	-.13	-.34**
SBSOD	.37**	-.22	-.27*
Accuracy	--	-.03	-.40**
Familiarity	--	.02	--

\* p-value < 0.05; \*\* p-value < 0.01;

<sup>a</sup>  $N = 61$ ; <sup>b</sup>  $N = 50$

Table 2: Correlations of Familiarity with Accuracy and Correct Decision Latency, with Participants and Pictures as the Unit of Analysis

	Accuracy <sup>a</sup>	Decision Latency <sup>b</sup>
Across participants	-.40**	.01
Across pictures	-.54**	.33*

\* p-value < 0.05; \*\* p-value < 0.01; <sup>a</sup>  $N = 61$ ; <sup>b</sup>  $N = 50$  across participants;  $N = 34$  across pictures

**Distance Estimation** To test our mental walk hypothesis, we correlated mean correct decision latency per picture (averaged across individuals) with participants' estimates of the distance to each picture location. The correlation was not significant,  $r(34) = .23$ ,  $p = .17$ , providing no evidence for the mental walk hypothesis.

## Discussion

Using a heading-recall task, we replicated Sholl et al.'s (2006) finding that individuals can recall allocentric-directional information from pictures and that their performance is related to SOD. We replicated their finding that the least accurate directional estimates come from heading disparities of 180° and that the longest decisions latencies come from heading disparities of 180° and 90°. These results provide support for the theory that humans have an allocentric-heading system similar to those found in animals. We also replicated a significant correlation

between SOD measures and heading-recall measures. However, while our correlations reached significance, they were noticeably lower than those observed by Sholl et al. In addition, we found significant correlations between performance measures and familiarity that Sholl et al. did not find. But we failed to find support for our hypothesis that decision latency would be correlated with estimated distance.

We successfully replicated Sholl et al.'s experimental findings and were also successful in increasing the general accuracy level on the heading-recall task. Thus, people can be quite accurate in providing allocentric-heading for pictures, when adequate training and feedback are provided. Our results demonstrate that the effects replicate across campuses. However, our results also indicate that specific aspects of the local environment, such as the nearby mountain range, may also have affected the accessibility of the views. This study and Sholl et al.'s study were conducted on campuses with structures intrinsically aligned with magnetic compass directions. Using campuses with a less regular structure, or different allocentric reference systems for the pictures, would allow for further testing of the generality of these results.

Our study also attempted to replicate Sholl et al.'s high correlations between SOD measures and heading-recall performance. In contrast we found moderate significant correlations. There are two potential causes for the reduced correlations: administration of SOD measures before the heading-recall task and the increase in accuracy resulting from better instructions and additional practice trials.

With regards to our alternative hypotheses, the hypothesis that heading-recall performance measures would be correlated with familiarity found support. Since recognition is the likely first step in recalling an allocentric-heading, recognizing the view of the location is likely a first step in completing the task. It is possible that poor SOD individuals require more experience with locations than good SOD individuals to attain similar levels of recognition performance. Although Sholl et al. (2006) did not find effects of familiarity, they did not measure familiarity of their experimental participants and we found that the familiarity ratings from our pretest were not perfectly correlated with the familiarity ratings from our experimental participants. In summary, we found that significant correlations between accuracy and SOD are partially due to familiarity. In contrast to Sholl et al.'s conclusion that SOD reflects solely the ability to recall the allocentric heading of the picture, we found that SOD, familiarity and allocentric-heading accuracy are all related.

Our hypothesis that decision latency would be related to distance estimates was not supported. This suggests that people do not accomplish this task by imagining a mental walk to the locations in the pictures, or at least that this mental walk process is not an analog process. On the other hand, in debriefing interviews, many participants mentioned imagining how they would travel past the target location or extrapolating allocentric-heading from the direction of the

target location. This, and the fact that global orienting cues (like the mountains) seem to have affected performance, suggests that there may be several strategies employed in this task.

In conclusion, this study has replicated the result that ability to judge the heading from which a picture was taken is related to one's current heading, and provides motivation for further studying the possibility of a human allocentric orientation (or head-direction) system. On the other hand, our results do not support the view that self-reported SOD measures simply reflect the operation of a human head-direction system. Previous studies have found that SOD measures are related to multiple spatial skills, including learning spatial layout and updating, and the correlations we observed between SOD and the heading recall task are similar in size (in the moderate range) to the correlations typically found with these other tasks. Thus it is likely that self-report SOD measures reflect a range of navigation abilities, not just the operation of a head-direction system, and future studies of this system should rely on objective measures of performance, such as the heading-recall task, rather than relying on self-reports as a measure of a human allocentric heading (or head-direction) system.

## References

- Bryant, J. K. (1982). Personality correlates of sense of direction and geographic orientation. *Journal of Personality and Social Psychology*, 43, 1318-1324.
- Hegarty, M., Richardson, A. E., Montello, D. R., Lovelace, K., & Subbiah, I. (2002). Development of a self-report measure of environmental spatial ability. *Intelligence*, 30, 425-447.
- Heth, C. D., Cornell, E. H., & Flood, T. L. (2002). Self-ratings of sense of direction and route reversal performance. *Applied Cognitive Psychology*, 16, 309-324.
- Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, 92(2), 137-172.
- Kozlowski, L. T. & Bryant, K. J. (1977). Sense of direction, spatial orientation, and cognitive maps. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 590-598.
- Prestopnik, J. L., & Roskos-Ewoldsen, B. (2000). The relations among wayfinding strategy use, sense of direction, sex, familiarity, and wayfinding ability. *Journal of Environmental Psychology*, 20(2), 177-191.
- Ranck, J. B., Jr. (1984). Head-direction cells in the deep cell layers of dorsal presubiculum in freely moving rats. *Society of Neuroscience Abstracts*, 10, 599.
- Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171, 701-703.
- Sholl, M. J., Kenny, R. J., & DellaPorta, K. A. (2006). Allocentric-heading recall and its relation to self-reported sense-of-direction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(3), 516-533.

# Space-Time Interdependence and Sensory Modalities: Time Affects Space in the Hand But Not in the Eye

Zhenguang G. Cai (zhenguangcai@gmail.com)

Louise Connell (louise.connell@manchester.ac.uk)

School of Psychological Sciences, University of Manchester, Coupland 1, Oxford Road  
Manchester, M13 9PL, UK

## Abstract

Time and space are intimately related, but what is the real nature of this relationship? Is time mapped metaphorically onto space, or do the two domains share a common representational format? In the present paper, participants touched (but could not see) physical sticks while listening to an auditory note. Judgements of stick length were affected by concurrent note duration, but not vice versa. When participants were allowed to see as well as touch the sticks, however, the effects reversed. These findings run counter to the spatial metaphor account of time, which claims that effects of space on time should always be stronger than those of time on space. Rather, our findings support the spatial representation account, in which time and space share a common neural substrate that may be affected by concurrent temporal or spatial information, depending on the perceptual acuity of the modality used to perceive space.

**Keywords:** Time; space; representation; haptic perception; visual perception; sensory dominance; metaphor

Though our immediate perception of the world is limited to our senses such as vision and hearing, we use these senses to perceive and represent other dimensions of the world besides colours and sounds. For instance, we can perceive the spatial information of an object (e.g., its length, height and size) by looking at it or touching it. How we perceive and represent more abstract domains such as time, however, has been a perennial philosophical question. Many researchers have suggested that abstract domains are grounded to some extent in more familiar concrete domains that we develop through sensorimotor experience (e.g., Barsalou & Wiemer-Hastings, 2005; Gibbs, 2006; Lakoff & Johnson, 1980, 1999). Time, for example, can be understood through the domain of space, as reflected in our use of language. Speakers of English often talk about time in spatial terms (e.g., *a long/short time*) and sometimes space in temporal terms (e.g., *I am 5 minutes from the airport*). A range of studies has provided evidence that these linguistic expressions reflect a deeper conceptual bridge between time and space. For example, people perceive the passage of time either as if they are moving in space towards the future, or as if the future is moving towards them (e.g., Boroditsky & Ramscar, 2002; McGlone & Harding, 1998). Other studies have shown that space affects the perception of temporal durations such that people experience longer subjective time when they imagine themselves inside a larger scale model of a room than inside a smaller one (DeLong, 1981), with a larger square than a smaller one (Xuan, Zhang, He, & Chen,

2007), and with a longer line than a shorter one (Casasanto & Boroditsky, 2008).

There are two alternative accounts of the relationship between time and space representations. According to the *spatial metaphor* account, people employ spatial metaphors in thinking or talking about time such that they use their concrete spatial experience to support their understanding of abstract time processing (Boroditsky, 2000; Gibbs, 2006; Lakoff & Johnson, 1980, 1999). The temporal relation of two events can be expressed metaphorically as a relation between two locations in space (e.g., tomorrow is ahead of yesterday). Similarly, a temporal duration can be metaphorically envisioned as the distance from a spatial location representing the onset of the duration and a spatial location representing the offset of the duration. Critically, the spatial metaphor account assumes that time and space remain two separate representational systems with an asymmetric mapping between them: concurrent spatial information should always affect its dependent domain of time to a greater extent than concurrent temporal information can affect space (Casasanto & Boroditsky, 2008; Casasanto, Fotakopoulou, & Boroditsky, 2010; Merritt, Casasanto, & Brannon, 2010).

Alternatively, according to the *spatial representation* account of time, temporal and spatial information are processed in a common neural substrate and share representational and attentional resources. Time is closely related to space in action and perception (e.g., Walsh, 2003): space and time are often coordinated in action and correspond to each other in movement (e.g., things travel a certain distance in a certain time). Thus, temporal duration and spatial distance may share a representational format (e.g., Locke, 1689/1995), such that two events are separated by a particular duration in the same way that two locations are separated by a particular distance. Some stronger versions of spatial representation theories have argued that time, space and number all share a common magnitude representation (Burr, Ross, Binda & Morrone, 2010; Walsh, 2003), but a weaker version of the spatial representation theory of time does not necessarily require the magnitude assumption. Critically, rather than comprising separate representational domains, time and space occupy an overlapping temporo-spatial representation that may be affected by concurrent temporal or spatial information. Since the same representation can subserve both temporal and spatial processing, the spatial representation account thus differs from the spatial metaphor account in allowing

the effects of time on space to be as strong as or stronger than the effects of space on time, depending on factors we describe below.

Empirical evidence has thus far favoured the spatial metaphor account, with the strongest evidence coming from studies showing apparently robust asymmetric effects of space on time in nonlinguistic paradigms. For example, Casasanto and Boroditsky (2008; see also Casasanto et al., 2010) showed participants a horizontal line onscreen and then asked them to reproduce either the length of the line or its duration of presentation. They found that people's estimates of the line's duration increased as a function of its length, but that estimates of length remained unaffected by the duration of the line onscreen. Furthermore, the same pattern emerged whether the line was static or grew to its full length, when the line was replaced with a moving dot, or when a concurrent auditory note provided an additional source of temporal information. A later variant of this nonlinguistic task, where participants categorised the length or duration of a line as long or short according to learned standards, did find an effect of time on space (Merritt et al., 2010), but since this effect was smaller than that of space on time, the asymmetric hypothesis of the spatial metaphor account was supported.

The above studies all use the visual modality to present spatial information. However, spatial representations are not themselves visual, and are rather handled by a multimodal or supramodal system that draws perceptual input from visual, haptic, or auditory modalities (or even from linguistic descriptions) in order to create a common spatial representation (Bryant, 1992; Giudice, Betty, & Loomis, 2011; Lacey, Campbell & Sathian, 2007). Visual perception has the best spatial acuity (i.e., the sharpest or most detailed resolution) of all human perceptual modalities, and so spatial representations resulting from vision have a level of specificity that is not found in spatial representations resulting from other perception. Therefore, the asymmetric effects of space on time found by Casasanto and colleagues may be due to the high spatial acuity from vision being relatively impervious to distortion rather than to an asymmetric mapping between domains.

In the present paper, we examined the interaction of time and space using touch rather than vision. Participants perceived spatial information regarding the length of a stick via haptic (i.e., tactile and proprioceptive) perception while concurrently perceiving a note for a particular duration. As in Casasanto and Boroditsky (2008), participants attended to both the spatial length and temporal duration in each trial and then reproduced either length or duration. If the spatial metaphor account is correct, any effects of time on spatial judgements should be substantially weaker than the usual effects of space on temporal judgements. In contrast, if the spatial representation account is correct, then whether time affects space depends on the relative acuity of spatial representations. Though space can be perceived either visually or haptically, research has suggested that haptic-spatial representations are more prone to distortion than

those of vision (e.g., Lederman, Klatsky & Barber, 1985); hence, we predicted that haptic space would be susceptible to interference from concurrent temporal information. Furthermore, since haptic-spatial representations are less acute than visuo-spatial representations (e.g., Schultz & Petersik, 1994), they may not be able to distort time as visuo-spatial representations do. Thus, when spatial information relies on touch, we expected the effect of time on space to be substantially stronger than the effects of space on time.

## Experiment 1

In this study, people were presented with a stick that they could touch but not see, so information regarding spatial length was haptically (but not visually) perceived while hearing a concurrent note for a particular duration. We then asked participants to reproduce either the spatial length of the stick by holding their hands apart (still with no visual feedback) or the temporal duration of the note by holding down a button. Following the spatial representation account, we expected concurrent temporal duration to affect the reproduction of spatial length, but for spatial concurrent information to have limited or no effects at all on the reproduction of duration.

## Method

**Participants** Thirty-two right-handed native speakers of English were recruited from the University of Manchester community (30 women, mean age = 19.2; two were later excluded from data analysis; see below). They all had normal or corrected-to-normal vision and had no hearing impairments. Participants received £5 or course credits for their participation.

**Materials** Eight rigid, hollow plastic sticks (ca. 16 mm in diameter) were divided into varying lengths (100 – 450 mm in steps of 50 mm). Eight sine waveform notes of 440 Hz were created in varying durations (1000 – 4500 ms in steps of 500 ms) with Audacity (Version 1.2.6). Crossing stick lengths with note durations, we created 64 stick-note stimulus sets. Each stimulus set was then combined with a length or duration reproduction task and divided into two stimulus lists, such that if a stimulus set occurred in List 1 with a length task, it occurred in List 2 with a duration task (i.e., task was counterbalanced across stick length and note duration). Each list thus had 32 stick-note pairs, half with a length task and the other half with a duration task.

**Procedure** Each participant was individually tested in a cubicle. The participant sat at a table with a response button box on his or her lap, and placed the hands and forearms through the gap at the bottom of a barrier, with a cape fastened around the neck to block all visual access to the hands and arms (see Figure 1). During the testing procedure, the experimenter (first author) sat at right angles to the participant and had a box to one side containing the eight sticks. The experiment was run with Superlab 4.0,

with the order of trials individually randomized per participant. In each trial, the experimenter placed the relevant stick (as designated by the experimental programme) on the table and the participant pressed against the ends of the stick with index fingers; at point of contact, the experimenter pressed a key to begin playing the note. When the note stopped, the participant let go of the stick and withdrew the hands to the base of the barrier (i.e., to disrupt hand positioning so stick length was not passively preserved between the index fingers). The experimenter then returned the stick to the box and verbally instructed which judgement the participant was to make (as designated by the experimental programme). When the experimenter said “Time”, the participant held down a button on the response box (located on the lap) for the same duration as the note. When the experimenter said “Length”, the participant reached forward (until they touched a board held up by the experimenter) and indicated the length of the stick between the index fingers; the experimenter then removed the board and took a photograph of the hands' position using a fixed camera. Use of the board (at location 'X' in Figure 1) ensured that the participants' hands were at a fixed distance from the camera. The photographs were taken at a resolution that allowed distance discrimination finer than 1 mm. Each participant performed a practice session of 4 trials before the real experiment, and the whole procedure lasted about 30 minutes.

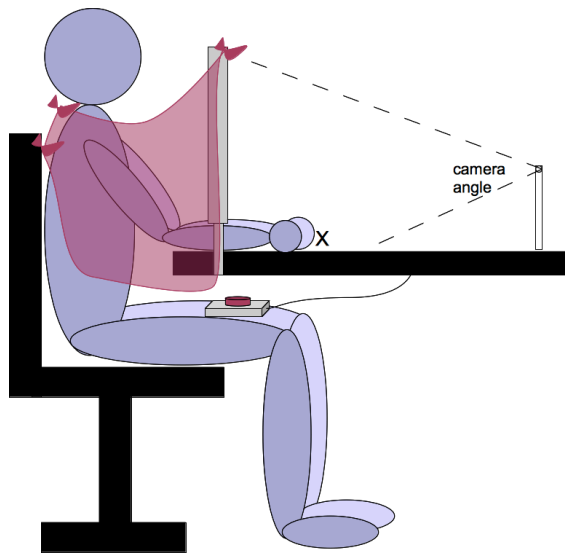


Figure 1: Schematic of the experimental setup: 'X' marks the location of both haptic perception and reproduction of length. The cape and barrier (both opaque) were used in Experiment 1 to block visual access to spatial information, and were absent in Experiment 2 to allow access.

**Measures** Duration reproductions in milliseconds were measured from onset to release of the response button). Length reproductions were measured by the first author from digital photographs by presenting each picture (condition-blind) and clicking on the centre of the left and

right index fingertips; distance was calculated as the difference between x-coordinates. For reliability analysis, the second author blind-coded a random 12% sample of pictures: agreement between coders was very high ( $r = .999$ ) and accurate to within 1 mm distance. All references to length are in mm.

**Design & Analysis** We excluded failed trials in which the participant did not proceed as instructed (e.g., wrong key presses; missed trials), and then removed outliers more than 2.5 SDs away from the mean for each length or duration condition. The data trimming resulted in the exclusion of less than 2% of either the length or duration trials. Following the criterion in Casasanto and Boroditsky (2008, p. 581), two participants who did poorly in either the length or duration judgements (i.e., when the regression coefficient fell below 0.5 in either the regression of reproduced durations with note duration or reproduced lengths with stick length) were excluded from the analysis.<sup>1</sup> We then used linear mixed effects (LME) modelling to analyse condition means for each participant (e.g., average reproduced duration per participant was regressed on each different stick length). The final model always included the fixed effect; the random effects always included the participant intercept.<sup>2</sup> Regression coefficients are reported as unstandardised  $\beta$  values with standard errors.

## Results and discussion

Reproduced length was significantly affected by experienced duration,  $\beta = 0.0033$ ,  $SE = 0.0015$ ,  $t(209) = 2.27$ ,  $p = .024$ , but reproduced duration was unaffected by stick length,  $\beta = 0.113$ ,  $SE = 0.114$ ,  $t(209) = 0.99$ ,  $p = .324$ . Sticks that were accompanied by a longer duration note were judged to be longer in length, and sticks accompanied by a shorter duration note were judged to be shorter in length (see Figure 2). People's judgements of spatial distance perceived through touch were influenced by their temporal experience, but not vice versa. Both spatial and temporal estimates were highly accurate: reproduced durations were well predicted by actual note duration,  $\beta = 0.771$ ,  $SE = 0.014$ ,  $t(209) = 53.56$ ,  $p < .0001$ , and reproduced lengths were well predicted by actual stick length,  $\beta = 0.818$ ,  $SE = 0.011$ ,  $t(209) = 76.20$ ,  $p < .0001$ .

The results of the experiment support the spatial representation rather than spatial metaphor account of time. When space is haptically perceived, it does not affect time perception; instead, time interferes with the perception of haptic space. Our findings stand in direct contrast to those of previous studies that found visual space influenced time but not the other around (Casasanto & Boroditsky, 2008; Casasanto et al., 2010; Merritt et al., 2010). These

<sup>1</sup> The inclusion of these two participants did not change the statistical pattern of the results in this experiment.

<sup>2</sup> The random subject slope did not significantly contribute to the model fit in any of the LME analyses; thus, we did not include it as a random effect.

discrepancies can be attributed to the different acuities of spatial representations in different modalities, as haptic-spatial representations (as in our Experiment 1) are of lower acuity than visuo-spatial representations (as in previous studies), and hence are prone to distortion by temporo-spatial information to a greater extent. Such an account then predicts that if space is visually perceived, the effects in Experiment 1 will be reversed. That is, highly acute visual perception of the stick will affect participants' time judgement, but spatiotemporal information will not be powerful enough to affect the vivid visuo-spatial memory in the length task. We test this hypothesis in Experiment 2.

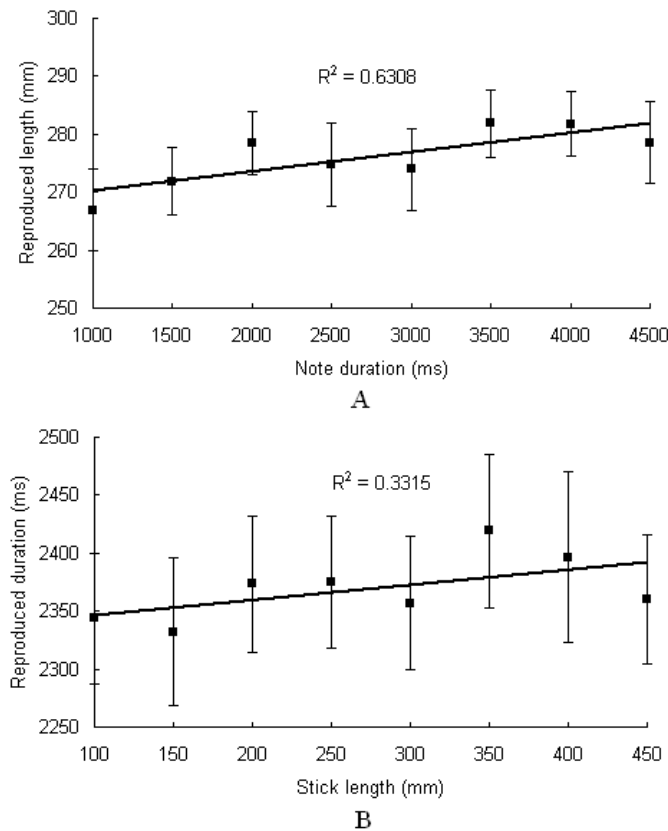


Figure 2: Effects of time on space for haptic perception in Experiment 1 (A), with no corresponding effects of space on time (B). Error bars show one SE.  $R^2$  fit is for graphed means.

## Experiment 2

This study used the same paradigm as Experiment 1 with one exception: people were allowed to see as well as touch the stick, so information regarding spatial length was both haptically and visually perceived. Since the visual modality tends to be dominant in perception (e.g., participants tend to report only visual perception when a visual stimulus is simultaneously presented with an auditory or haptic stimulus: Colavita, 1974; Hartcher-O'Brien et al., 2008), we expected the high spatial acuity of vision in Experiment 2 to affect

temporal judgements but not vice versa (i.e., a restoration of the usual asymmetric effect of space on time).

## Method

**Participants** Twenty-six participants were recruited as in Experiment 1 (22 women, mean age = 19.3; six were later excluded from data analysis; see below).

**Materials** As per Experiment 1.

**Procedure** The procedure was the same as in Experiment 1, except 1) the cape and barrier were removed (see Figure 1) so that participants could see the stick as well as touch it, and see their hands when reproducing length; and 2) the stick was presented at jittered transverse positions in order to discourage participants from using the visual cues of the desk (e.g., distance from side edge) when reproducing the length of the stick.

**Measures** As per Experiment 1. Double-coding of 15% of the lengths shows very high agreement between the two coders ( $r > .999$ ) and accurate to within 1 mm distance.

**Design & Analysis** The same data trimming method as in Experiment 1 resulted in the removal of less than 2% of either the length or duration trials. Six participants were excluded according to the exclusion criterion adopted in Experiment 1.<sup>3</sup>

## Results and discussion

Reproduced length was unaffected by experienced duration,  $\beta = 0.0016$ ,  $SE = 0.0016$ ,  $t(139) = 0.98$ ,  $p = .329$ , but reproduced duration was significantly affected by stick length,  $\beta = 0.325$ ,  $SE = 0.133$ ,  $t(139) = 2.44$ ,  $p = .016$ . Actual durations that were accompanied by shorter sticks were judged to take less time than durations that were accompanied by longer sticks (see Figure 3). People's judgements of time were influenced by their visual-haptic perception of spatial distance, but not vice versa. Both spatial and temporal estimates were again highly accurate: reproduced durations were well predicted by actual duration,  $\beta = 0.773$ ,  $SE = 0.017$ ,  $t(139) = 43.72$ ,  $p < .0001$ , and reproduced lengths were well predicted by actual length,  $\beta = 0.739$ ,  $SE = 0.010$ ,  $t(139) = 70.71$ ,  $p < .0001$ .

Results in Experiment 2 thus demonstrated that when space was perceived in vision, the effects in Experiment 1 were reversed; that is, visual space influenced time but not the other way round, just as found in previous studies by Casasanto and colleagues. As predicted by the spatial representation account of time, the ability of time to affect space depends on the relative acuity of spatial representations.

<sup>3</sup> Again, the inclusion of these 6 participants did not change the statistical pattern in the experiment.



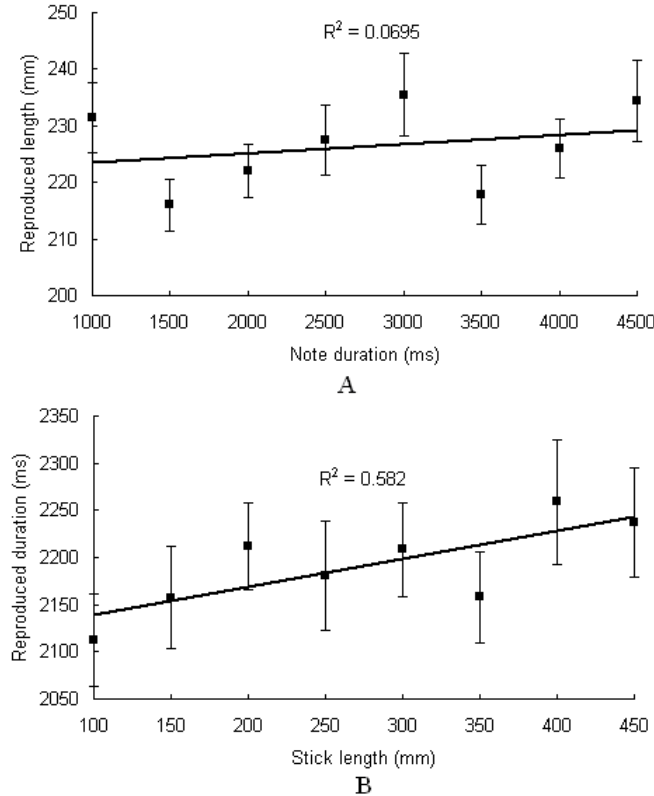


Figure 3: No effects of time on space for visuo-haptic perception in Experiment 2 (A), with instead effects of space on time (B). Error bars show one SE.  $R^2$  fit is for graphed means.

## General discussion

Two experiments revealed a double-disassociation of time and space effects according to sensory modality: time influenced haptic space but not the other way around, and visual space influenced time but not the other way round. The latter findings are in line with previous observations that time perception is subject to spatial interference (Casasanto & Boroditsky, 2008; DeLong, 1981; Xuan et al., 2007). However, when space is perceived haptically, concurrent spatial information fails to affect time perception; on the contrary, the perception of haptic space is influenced by concurrent temporal information. Such findings are, to our best knowledge, the first clear demonstration of a “reverse” asymmetry between space and time, i.e., time affects space to a greater extent than space affects time. This reverse asymmetry is therefore inconsistent with the spatial metaphoric mapping account of time representation (Casasanto & Boroditsky, 2008; Casasanto et al., 2010; Merritt et al., 2010), according to which space should always have a greater effect on time than time on space, as temporal perception metaphorically employs spatial representations. Instead, our findings are more consistent with the spatial representation account, according to which space and time share a common representation that is subject to interference from either

direction.

The spatial representation account thus allows for a two-way interdependence between time and space, which is mediated by the acuity of the sensory modality in which space is perceived. Highly sharp and stable visuo-spatial representations exert a strong influence on time judgements and are relatively impervious to temporal interference, while more distortable haptic-spatial representations are not acute enough to influence time and instead are prone to interference from temporal information. This spatial representation account is also consistent with the findings of Merritt et al. (2010), who found symmetric effects between space and time in rhesus monkeys but not in humans. Merritt et al. argued that one explanation for the discrepancy between humans and monkeys is that human language facilitates the use of metaphoric mappings in spatial representations of time thinking; monkeys, lacking space-time metaphors, also lack asymmetric mappings between the domains. However, it is possible for human language to facilitate greater precision in visuo-spatial tasks without recourse to time-space metaphoric mappings. In their paradigm, Merritt and colleagues required participants to memorise two standard reference lines: one short (6 cm) and one long (24 cm). When later presented with another line, monkeys had only their visuo-spatial memory of the reference lines to help them decide if this new line was long or short, whereas humans also had a verbal numeric label available for what constituted long or short. Previous work has shown that availability of verbal numerical labels enhance accuracy in dot estimation tasks (Izard & Dehaene, 2008; Pica, Lemer, Izard, & Dehaene, 2004), and that verbal shadowing disrupts spatial memory in adults so that they show behaviour patterns similar to young children and rats (Hermer-Vasquez, Spelke, & Katnelson, 1999). It is therefore possible that availability of number words helped to preserve spatial acuity of the reference lines in humans (thus rendering spatial memory less susceptible to temporal interference), whereas lack of number words in monkeys allowed their spatial memory of the reference lines to be distorted by temporo-spatial information.

It should be noted that space-time interdependence may arise from other shared dimensions such as quantity or magnitude, on which space and time are closely interconnected (e.g., *more* space travelled in *more* time). In other words, the underlying representation of both space and time (and number) may be magnitude-based (Burr et al., 2010; Walsh, 2003), which therefore gives rise to the interdependence between space and time. Though such an account is compatible with our data, it would require that magnitude information from haptic space be less acute than magnitude information from visual space, an assumption that has yet to be tested. The spatial representation account of time that we put forward here can explain the current effects in terms of differential perceptual acuity without positing a magnitude system.

Finally, our study has implications beyond space-time interdependence. It suggests that previous findings of

space-time asymmetry have more to do with differential acuity in perceiving space than the use of linguistic metaphor extending into nonlinguistic thought, thus casting considerable doubt on space-time asymmetric as evidence for the effects of language on thought (e.g., Boroditsky, 2000; Whorf, 1956). Furthermore, previous research has shown that visuo-spatial and haptic-spatial information are functionally equivalent (e.g., Giudice, et al, 2011), therefore suggesting a common storage (e.g., Lacey et al., 2007). Our findings lend further support to such a conclusion. That is, in order for time to interact with both haptic space and visual space, spatial information in these different modalities should be encoded in the same format.

In conclusion, the present experiments show that time is not asymmetrically dependent on space, and hence offer evidence against the spatial metaphor account of time representation. Rather, time and space share a common spatial representation: time *affects* spatial information that emerges from relatively low-acuity perceptual modalities like touch, and time *is affected by* spatial information from relatively high-acuity perceptual modalities like vision.

### Acknowledgments

This work was supported by a research project grant from the Leverhulme Trust (F/00 120/CA).

### References

- Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thought*. New York: Cambridge University Press.
- Boroditsky, L. (2000). Metaphoric Structuring: Understanding time through spatial metaphors. *Cognition*, 75, 1-28.
- Boroditsky, L., & Ramscar, M. (2002). The roles of body and mind in abstract thought. *Psychological Science*, 13, 185-189.
- Bryant, D. J. (1992). A spatial representation system in humans. *Psychology*, 3(16), space 1.
- Burr, D. C., Ross, J., Binda, P., & Morrone, M. C. (2010). Saccades compress space, time and number. *Trends in Cognitive Sciences*, 14, 528-533.
- Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, 106, 579-593.
- Casasanto, D., Fotakopoulou, O., & Boroditsky, L. (2010). Space and Time in the Child's Mind: Evidence for a Cross-Dimensional Asymmetry. *Cognitive Science*, 34, 387-405.
- Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, 16, 409-412.
- DeLong, A. J. (1981). Phenomenological space-time: toward an experiential relativity. *Science*, 213, 681-683.
- Gibbs, R. (2006). *Embodiment and cognitive science*. New York: Cambridge University Press.
- Giudice, N. A., Betty, M. R., & Loomis, J. M. (2011). Functional equivalence of spatial images from touch and vision: Evidence from spatial updating in blind and sighted individuals. *Journal of Experimental Psychology: Learning Memory and Cognition*, 37, 621-634.
- Hartcher-O'Brien, J., Gallace, A., Krings, B., Koppen, C., & Spence, C. (2008). When vision 'extinguishes' touch in neurologically-normal people: Extending the Colavita visual dominance effect. *Experimental Brain Research*, 186, 643-658.
- Hermer-Vazquez, L., Spelke, E. S., & Katsnelson, A. S. (1999). Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology*, 39, 3-36.
- Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, 106, 1221-1247.
- Lacey, S., Campbell, C., & Sathian, K. (2007). Vision and touch: multiple or multisensory representations of objects? *Perception*, 36, 1513-1521.
- Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago and London: The University of Chicago Press.
- Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. Chicago: University of Chicago Press.
- Lederman, S. J., Klatzky, R. L. & Barber, P. (1985). Spatial and movement-based heuristics for encoding pattern information through touch. *Journal of Experimental Psychology: General*, 114, 33-49.
- Locke, J. (1689/1995). *An essay concerning human understanding*. Amherst: Prometheus Books.
- McGlone, M. S., & Harding, J. L. (1998). Back (or forward?) to the future: The role of perspective in temporal language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1211-1223.
- Merritt, D. J., Casasanto, D., & Brannon, E. M. (2010). Do monkeys think in metaphors? Representations of space and time in monkeys and humans. *Cognition*, 117, 191-202.
- Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, 306, 499-503.
- Schultz, L. M. & Petersik, J. T. (1994). Visual-haptic relations in a two-dimensional size matching task. *Perceptual and Motor Skills*, 78, 395-402.
- Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, 7, 483-488.
- Whorf, B. (1956). *Language, Thought, and Reality: selected writings of Benjamin Lee Whorf*, ed. J.B. Carroll. Cambridge, MA: MIT Press.
- Xuan, B., Zhang, D., He, S., & Chen, X. (2007). Larger stimuli are judged to last longer. *Journal of Vision*, 7, 1-5.



# Sentic Panalogy: Swapping Affective Common Sense Reasoning Strategies and Foci

Erik Cambria, Daniel Olsher, Kenneth Kwok

{cambria, olsher, kenkwok}@nus.edu.sg

Cognitive Science Programme, Temasek Laboratories

National University of Singapore, 5A Engineering Drive 1, Singapore 117411

<http://sentic.net>

## Abstract

An important difference between traditional AI systems and human intelligence is our ability to harness common sense knowledge gleaned from a lifetime of learning and experiences to inform our decision-making and behavior. This allows humans to adapt easily to novel situations where AI fails catastrophically for lack of situation-specific rules and generalization capabilities. In order for machines to exploit common sense knowledge in reasoning as humans do, moreover, we need to endow them with human-like reasoning strategies. In problem-solving situations, in particular, several analogous representations of the same problem should be maintained in parallel while trying to solve it so that, when problem-solving begins to fail while using one representation, the system can switch to one of the others. Sentic panalogy is a technique that aims to emulate such process by exploiting graph-mining and dimensionality-reduction techniques to dynamically interchange both different reasoning strategies and the foci around which such strategies are developed.

**Keywords:** AI; NLP; Cognitive systems; Sentic computing.

## Introduction

Emotions are different Ways to Think (Minsky, 2006) that our mind triggers to deal with different situations we face in our lives. Our decision-making and problem-solving skills, in fact, are strictly dependent on both our common sense knowledge about the world and the appraisal associated with this (Scherer, Shorr, & Johnstone, 2001). The capability to accordingly compress and exploit such information, which we term affective common sense reasoning (Cambria, Olsher, & Kwok, 2012), is a fundamental component in human experience, cognition, perception, learning, and communication.

For this reason, we cannot prescind from emotions in the development of intelligent user interfaces: if we want computers to be really intelligent, not just have the veneer of intelligence, we need to give them the ability to recognize, understand, and express emotions. Furthermore, in order not to get stuck and to be able to tackle different problems from different perspectives, an intelligent machine should not have a unique way to deal with a task, but rather be endowed with different reasoning strategies and with the capability to accordingly switch among these.

This work further develops a recently proposed approach (Cambria, Mazzocco, Hussain, & Durrani, 2011) for the emulation of the human capability to switch between different perspectives and find novel ways to look at things. Such approach is inspired by Minsky's notion of 'panalogy' (parallel analogy), which states that several analogous representations of the same problem should be maintained in parallel while trying to solve it (Minsky, 2006).

To show the effectiveness of the proposed approach, termed sentic panalogy, we employ it for the natural language processing (NLP) task of sentiment analysis, for which a faceted and nuanced analysis is mostly needed.

The structure of the paper is as follows: the first section provides some background information on sentiment analysis; the second section introduces the concept of affective common sense reasoning and explains why and how this can aid sentiment analysis; the third and fourth sections describe the implementation of the switch among different strategies and among the foci around which such strategies are developed, respectively; the fifth section provides an evaluation of the proposed approach; the last section, finally, comprises concluding remarks and future directions.

## Sentiment Analysis

Sentiment analysis is a branch of the broad field of text data mining and refers generally to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents. It can be viewed as an extension of data mining or knowledge discovery from (structured) databases (Fayyad, Piatetsky, & Smyth, 1996; Simoudis, 1996). As the most natural form of storing information is text, sentiment analysis is believed to have a commercial potential higher than that of data mining. Sentiment analysis, however, is also a much more complex task as it involves dealing with text data that are inherently unstructured and fuzzy. It is a multi-disciplinary research area that involves the adoption of techniques in fields such as text analysis, information retrieval and extraction, auto-categorization, machine learning, clustering, and visualization.

Most of the existing approaches to opinion mining and sentiment analysis rely on the extraction of a vector representing the most salient and important text features, which is later used for classification purposes. Some of the most commonly used features are term frequency (Wu, Luk, Wong, & Kwok, 2008) and presence (Pang, Lee, & Vaithyanathan, 2002). The latter, in particular, is a binary-valued feature vectors in which the entries merely indicate whether a term occurs or not, and formed a more effective basis for polarity classification. This is indicative of an interesting difference between typical topic-based text categorization. While a topic is more likely to be emphasized by frequent occurrences of certain keywords, overall sentiment may not usually be highlighted through repeated use of the same terms.

Differently from topics, in fact, sentiments can often be expressed in a more subtle manner, making it difficult to be identified by specific keywords, especially when considering multiple domains. Humans readers do not face such difficulty as they can infer the cognitive and affective information associated with natural language text through their affective common sense knowledge, that is, obvious or widely accepted things that people normally know about the world, but which are usually left unstated in discourse, e.g., that people smile when they are happy and things fall downwards (and not upwards). An important feature of affective common sense reasoning, in fact, is the sensitivity to nuanced readings of natural language.

A sentence can be read differently depending on nuances in opinionated text and such nuanced reading can lead to markedly different reasoning trajectories. The first step in human cognitive and affective information processing, in fact, is in an appraisal of the current situation (Scherer et al., 2001). In order to accordingly infer semantics and sentics (Cambria, Benson, Eckl, & Hussain, 2012), i.e., the cognitive and affective information associated with natural language text, next-generation sentiment analysis methods need to go beyond a mere word-level analysis and use affective common sense reasoning to better grasp the conceptual rules that govern sentiment and the clues that can convey these concepts from realization to verbalization in the human mind.

### Affective Common Sense Reasoning

Current thinking in cognitive psychology suggests that humans process information at a minimum of two distinct levels. There is extensive evidence for the existence of (at least) two processing systems within the human brain, one that involves fast, parallel, unconscious processing, and one that involves slow, serial, more conscious processing (Kirkpatrick & Epstein, 1992; Chaiken & Trope, 1999; Smith & DeCoster, 2000; Epstein, 2003; Kahneman, 2011). Dual-process models of automatic and controlled social cognition have been proposed in nearly every domain of social psychology.

Evidence from neurosciences supports this separation, with identifiably different brain regions involved in each of the two systems (Lieberman, 2007). Such systems, which we term U-level (unconscious) and C-level (conscious), can operate simultaneously or sequentially, and are most effective in different contexts. The former, in particular, works intuitively, effortlessly, globally, and emotionally. The latter, in turn, works logically, systematically, effortfully, and rationally. According to different contexts and purposes, moreover, the systems should be capable to dynamically swap both different reasoning strategies and the foci around which such strategies are developed.

In this work, we emulate such dual-process model through an ensemble application of dimensionality-reduction and graph-mining techniques on AffectNet (Cambria & Hussain, 2012), an affective common sense knowledge base built upon WordNet-Affect (WNA) (Strapparava & Valitutti, 2004), a

linguistic resource for the lexical representation of affect, and ConceptNet (Havasi, Speer, & Alonso, 2007), a semantic network of common sense knowledge. In particular, multi-dimensionality reduction techniques are employed on AffectNet to dynamically configure it and, hence, to model the switch between different reasoning strategies, while graph mining and clustering methods are applied to model the switch between the foci around which those strategies are developed, in order to accordingly exploit the different facets of the affective common sense knowledge base.

### Swapping Reasoning Strategies

To some extent, our reasoning capability can be re-conducted to the identification of useful patterns in our acquired knowledge about the world. Our experience and common sense knowledge is likely to be organized in our mind as interconnected concepts and events and most of these links are weighted by affective information, as we tend to forget or hardly recall memories that are not associated with any kind of positive or negative emotion. Therefore, the human capacity to envision possible outcomes of a decision might lie both in the capability of crawling the semantic network of concepts we have acquired through experience (C-level), and in the capability of summarizing the huge amount of inputs and outputs of previous situations to find useful patterns that might work at the present time (U-level).

The latter capability, in cognitive science, is termed ‘compression’ and refers to transforming diffuse and distended conceptual structures that are less congenial to human understanding so that they become better suited to our human-scale ways of thinking. Compression is hereby implemented by representing affective common sense knowledge in a way that it is neither too concrete nor too abstract with respect to the detail granularity needed for performing a particular task.

To this end, we first generate a matrix representation of AffectNet by applying blending (Havasi, Speer, Pustejovsky, & Lieberman, 2009), a technique that performs inference over multiple sources of data simultaneously, taking advantage of the overlap between them. In particular, the alignment of ConceptNet and WNA yields  $A$ , a matrix in which common sense and affective knowledge coexist, i.e., a matrix  $14,301 \times 117,365$  whose rows are concepts (e.g., ‘dog’ or ‘bake cake’), whose columns are either common sense and affective features (e.g., ‘isA-pet’ or ‘hasEmotion-joy’), and whose values indicate truth values of assertions. Therefore, in  $A$ , each concept is represented by a vector in the space of possible features whose values are positive for features that produce an assertion of positive valence (e.g., ‘a penguin is a bird’), negative for features that produce an assertion of negative valence (e.g., ‘a penguin cannot fly’), and zero when nothing is known about the assertion. The degree of similarity between two concepts, then, is the dot product between their rows in  $A$ . The value of such a dot product increases whenever two concepts are described with the same feature and decreases when they are described by features that are negations of each other.

In particular, we use truncated singular value decomposition (TSVD) (Wall, Rechtsteiner, & Rocha, 2003) in order to obtain a new matrix containing both hierarchical affective knowledge and common sense. The resulting matrix has the form  $\tilde{A} = U_k \Sigma_k V_k^T$  and is a low-rank approximation of  $A$ , the original data. This approximation is based on minimizing the Frobenius norm of the difference between  $A$  and  $\tilde{A}$  under the constraint  $\text{rank}(\tilde{A}) = k$ . For the Eckart-Young theorem (Eckart & Young, 1936), it represents the best approximation of  $A$  in the least-square sense, in fact:

$$\begin{aligned} \min_{\tilde{A} | \text{rank}(\tilde{A})=k} \|A - \tilde{A}\| &= \min_{\tilde{A} | \text{rank}(\tilde{A})=k} \|\Sigma - U^* \tilde{A} V\| \\ &= \min_{\tilde{A} | \text{rank}(\tilde{A})=k} \|\Sigma - S\| \end{aligned}$$

assuming that  $\tilde{A}$  has the form  $\tilde{A} = USV^*$ , where  $S$  is diagonal. From the rank constraint, i.e.,  $S$  has  $k$  non-zero diagonal entries, the minimum of the above statement is obtained as follows:

$$\begin{aligned} \min_{\tilde{A} | \text{rank}(\tilde{A})=k} \|\Sigma - S\| &= \min_{s_i} \sqrt{\sum_{i=1}^n (\sigma_i - s_i)^2} = \\ &= \min_{s_i} \sqrt{\sum_{i=1}^k (\sigma_i - s_i)^2 + \sum_{i=k+1}^n \sigma_i^2} = \sqrt{\sum_{i=k+1}^n \sigma_i^2} \end{aligned}$$

Therefore,  $\tilde{A}$  of rank  $k$  is the best approximation of  $A$  in the Frobenius norm sense when  $\sigma_i = s_i$  ( $i = 1, \dots, k$ ) and the corresponding singular vectors are the same as those of  $A$ . If we choose to discard all but the first  $k$  principal components, common sense concepts and emotions are represented by vectors of  $k$  coordinates: these coordinates can be seen as describing concepts in terms of ‘eigenmoods’ that form the axes of AffectiveSpace, i.e., the basis  $e_0, \dots, e_{k-1}$  of the vector space. For example, the most significant eigenmood,  $e_0$ , represents concepts with positive affective valence. That is, the larger a concept’s component in the  $e_0$  direction is, the more affectively positive it is likely to be. Concepts with negative  $e_0$  components, then, are likely to have negative affective valence.

Thus, by exploiting the information sharing property of TSVD, concepts with the same affective valence are likely to have similar features - that is, concepts conveying the same emotion tend to fall near each other in AffectiveSpace. Concept similarity does not depend on their absolute positions in the vector space, but rather on the angle they make with the origin. For example we can find concepts such as ‘beautiful day’, ‘birthday party’, ‘laugh’, and ‘make person happy’ very close in direction in the vector space, while concepts like ‘sick’, ‘feel guilty’, ‘be laid off’, and ‘shed tear’ are found in a completely different direction (nearly opposite with respect to the centre of the space). By reducing the dimensionality of the matrix representation of  $A$ , AffectiveSpace compresses the feature space of affective common sense knowledge into one that allows to better gain global insight and human-scale understanding.

The number  $k$  of singular values selected for building AffectiveSpace, in fact, is a measure of the trade-off between precision and efficiency in the representation of the affective common sense knowledge base. Switching between different vector space dimensionalities can be seen as looking at the data from many different points of view. Balancing the number of singular values discarded when synthesizing AffectiveSpace, hence, corresponds to calibrate the affective common sense knowledge representation in a way that it is neither too concrete nor too abstract with respect to the detail granularity needed for performing a particular task. Different  $k$  values, for example, work differently according to the affective dimension we consider, e.g., for Pleasantness the best  $k$  appears to be closer to 100, while for Sensitivity a space of about 70 dimensions appears to be enough for precisely and efficiently represent affective common sense knowledge.

The capability to look at things from a different perspective, moreover, can be emulated by applying different space transformations to AffectiveSpace. The distribution of the values of each AffectiveSpace dimension is bell-shaped, with different centers and different degree of dispersion around them. In order to more uniformly distribute affective common sense knowledge in the vector space, an alternative strategy to represent AffectiveSpace consists in centering the values of the distribution of each dimension on the origin and in mapping dimensions according to a transformation  $x \in \mathbb{R} \mapsto x^* \in [-1, 1]$ . This transformation is often pivotal for better clustering AffectiveSpace as the vector space tends to have different grades of dispersion of data points across different dimensions, with some space regions more densely populated than others.

The switch to a different space configuration helps to distribute data more uniformly, possibly leading to an improved (or, at least, different) reasoning process. Switching between different space configurations, in fact, changes how each dimension is influent in the vector space representation of AffectNet and, hence, changes how we are looking at the affective common sense knowledge because similarity in AffectiveSpace does not depend on concepts’ absolute position, but rather on the angle they make with the origin of the vector space. In particular, the transformation  $x_{ij} \mapsto x_{ij} - \mu_i$  is first applied, being  $\mu_i$  the average of all values of the  $i$ -th dimension. Then a normalization is applied, combining the previous transformation with a new one  $x_{ij} \mapsto \frac{x_{ij}}{a \cdot \sigma_i}$ , where  $\sigma_i$  is the standard deviation calculated on the  $i$ -th dimension and  $a$  is a coefficient that can modify the same proportion of data that is represented within a specified interval.

Finally, in order to ensure that all components of the vectors in the defined space are within  $[-1, 1]$  (i.e., that the Chebyshev distance between the origin and each vector is smaller or equal to 1), a final transformation  $x_{ij} \mapsto s(x_{ij})$  is needed, where  $s(x)$  is a sigmoid function. Different choices for the sigmoid function may be made, influencing how ‘fast’ the function approaches 1 while the independent variable approaches infinity.

Combining the proposed transformations, two possible mapping functions are expressed in the following formulae:

$$x_{ij}^* = \tanh\left(\frac{x_{ij} - \mu_i}{a \cdot \sigma_i}\right)$$

$$x_{ij}^* = \frac{x_{ij} - \mu_i}{a \cdot \sigma_i + |x_{ij} - \mu_i|}$$

This space transformation leads to two main advantages, which could be of notable importance depending on the problem being tackled. First, this different space configuration ensures that each dimension is equally important by avoiding that the information provided by dimensions with higher (i.e., more distant from the origin) averages predominates. Second, normalizing according to the standard deviations of each dimension allows a more uniform distribution of data around the origin, leading to a full use of information potential.

### Swapping Reasoning Foci

The capability of switching among different Ways to Think can be thought as changing the foci around which we develop our different reasoning strategies. Such approach can be implemented in AffectiveSpace by changing the centroids around which the vector space is clustered. Such a clustering process is implemented by adopting a  $k$ -medoids approach (Kaufman & Rousseeuw, 1990) to partition the given observations into  $k$  clusters around as many centroids, trying to minimize a given cost function. Differently from the  $k$ -means algorithm, which does not pose constraints on centroids,  $k$ -medoids do assume that centroids must coincide with  $k$  observed points.

The most commonly used algorithm for finding the  $k$  medoids is the partitioning around medoids (PAM) algorithm, which determines a medoid for each cluster selecting the most centrally located centroid within the cluster. After selection of medoids, clusters are rearranged so that each point is grouped with the closest medoid. Since  $k$ -medoids clustering is a NP-hard problem, different approaches based on alternative optimization algorithms have been developed, though taking risk of being trapped around local minima. We use a modified version of the algorithm recently proposed by Park and Jun (Park & Jun, 2009), which runs similarly to the  $k$ -means clustering algorithm.

This has shown to have similar performance when compared to PAM algorithm while taking a significantly reduced computational time. In particular, we have  $N$  concepts ( $N = 14,301$ ) encoded as points  $x \in \mathbb{R}^p$  ( $p = 100$ ). We want to group them into  $k$  clusters and, in our case, we can fix  $k = 24$  as we are looking for one cluster for each sentic level of the Hourglass of Emotions (Cambria, Livingstone, & Hussain, 2012), a novel biologically-inspired and psychologically-motivated emotion categorization model, based on Plutchik's studies on human emotions (Plutchik, 2001), that can potentially describe any human emotion in terms of four independent but concomitant dimensions, whose different levels of activation make up the total emotional state of the mind.

Specifically, we need to cluster AffectiveSpace four times, once for each dimension. According to the Hourglass categorization model, in fact, each concept can convey, at the same time, more than one emotion (which is why we get compound emotions) and this information can be expressed via a sentic vector specifying the concept's affective valence in terms of Pleasantness, Attention, Sensitivity, and Aptitude. Therefore, given that the distance between two points in AffectiveSpace is defined as  $D(a, b) = \sqrt{\sum_{i=1}^p (a_i - b_i)^2}$ , the used algorithm, applied for each of the four affective dimensions, can be summarized as follows:

1. Each centroid  $C_n \in \mathbb{R}^{100}$  ( $n = 1, 2, \dots, k$ ) is set as one of the six concepts corresponding to each  $s$  in the current affective dimension
2. Assign each record  $x$  to a cluster  $\Xi$  so that  $x_i \in \Xi_n$  if  $D(x_i, C_n) \leq D(x_i, C_m)$   $m = 1, 2, \dots, k$
3. Find a new centroid  $C$  for each cluster  $\Xi$  so that  $C_j = x_i$  if  $\sum_{x_m \in \Xi_j} D(x_i, x_m) \leq \sum_{x_m \in \Xi_j} D(x_h, x_m) \quad \forall x_h \in \Xi_j$
4. Repeat step 2 and 3 until no changes on centroids are observed

After such a clustering process (performed at U-level), concepts that are semantically and affectively related to the input data can be intuitively retrieved by analogy and unconsciously crop out to the C-level. According to the initial centroid we choose, the final clusterization of AffectiveSpace can be very different. Hence, the way such initial medoids are chosen can be re-conducted to the human capability to switch between different perspectives to grasp the different facets of a problem.

At C-level, moreover, reasoning is performed by exploiting AffectNet's connectivity to find semantically and affectively related concepts by means of spectral association (Havasi, Speer, & Holmgren, 2010). Spectral association is a technique that involves assigning activation to seed concepts and applying an operation that spreads their values across the graph. This operation, an approximation of many steps of spreading activation, transfers the most activation to concepts that are connected to the seed concepts by short paths or many different paths in affective common sense knowledge.

Seed concepts can also be associated with negative activation values in order to reduce the spreading operation in the parts of the graph we are specifically not interested in. If we want to find concepts semantically related to 'bank' as a financial institution without getting concepts related to 'river bank', for example, we can set as positive seeds concepts like 'money', 'savings', or 'investment', and, as negative seeds, concepts like 'river', 'water', or 'shore'. The outcomes of spectral association can be very different according to which seeds we select as starting points for the spreading activation steps. Since spectral association involves TSVD, results also depend on the number  $k$  of singular values selected.

While choosing different  $k$  values can be seen as developing different reasoning strategies, choosing different seeds can be associated to changing the foci around which those strategies are developed. Through spectral association, positive and negative values of these concepts are spread across the graph representation of AffectNet, resulting in a set of contextually semantic related instances. Letting a machine switch between such seeds according to its own intuition (e.g., concepts obtained through AffectiveSpace at U-level) can be re-conducted to the human capability to change the foci around which different reasoning strategies are developed and, hence, to iterate on the ways to look at a problem until one that works is found.

## Evaluation

In order to efficiently and timely swap different reasoning strategies and foci, we perform all the computations (relative to the most significant configurations) a priori and save the results in a semantic-aware format, using an approach previously adopted for building SenticNet (Cambria, Havasi, & Hussain, 2012). The result is a system for affect recognition that has multiple ways to deal with natural language semantics and sentics. We tested sentic panalogy on a benchmark for affective common sense knowledge (BACK) built by applying CF-IOF (concept frequency - inverse opinion frequency), a technique similar to TF-IDF, on a 5,000 blogpost database extracted from LiveJournal<sup>1</sup>, a virtual community of users who keep a blog, journal, or diary. An interesting feature of this website is that bloggers are allowed to label their posts with both a category and a mood tag, by choosing from predefined categories and mood themes.

CF-IOF identifies common domain-dependent semantics in order to evaluate how important a concept is to a set of opinions concerning the same topic. Firstly, the frequency of a concept  $c$  for a given domain  $d$  is calculated by counting the occurrences of the concept  $c$  in the set of available  $d$ -tagged opinions and dividing the result by the sum of number of occurrences of all concepts in the set of opinions concerning  $d$ . This frequency is then multiplied by the logarithm of the inverse frequency of the concept in the whole collection of opinions, that is:

$$CF-IOF_{c,d} = \frac{n_{c,d}}{\sum_k n_{k,d}} \log \sum_k \frac{n_k}{n_c}$$

where  $n_{c,d}$  is the number of occurrences of concept  $c$  in the set of opinions tagged as  $d$ ,  $n_k$  is the total number of concept occurrences, and  $n_c$  is the number of occurrences of  $c$  in the whole set of opinions. A high weight in CF-IOF is reached by a high concept frequency in a given domain and a low frequency of the concept in the whole collection of opinions. Specifically, we exploited CF-IOF weighting to filter out common concepts in the LiveJournal corpus and detect relevant mood-dependent semantics for each of the Hourglass sentic levels. The result was a benchmark of 2000 af-

fective concepts that were screened by 21 English-speaking students who were asked to map each concept to the 24 different emotional categories that form the Hourglass of Emotions. BACK's concepts were compared with the classification results obtained by applying the AffectiveSpace process, spectral association, and sentic panalogy. Results showed that sentic panalogy achieves +9.7% and +6.2% accuracy than the standard (i.e., 100-dimensional) AffectiveSpace process and the default (i.e., fixed on the Hourglass sentic levels) spectral association, respectively.

## Brain-Inspired Sentiment Analysis

In order to test sentic panalogy also within a real-world scenario, we developed a brain-inspired software engine for sentiment analysis. This software engine consists of four main components: a pre-processing module, a semantic parser, a target spotting module, and an affect interpreter.

The pre-processing module firstly interprets all the affective valence indicators usually contained in opinionated text such as cross-linguistic onomatopoeias, exclamation words, degree adverbs, and emoticons. Secondly, it lemmatizes text and splits the opinion into single clauses according to grammatical conjunctions and punctuation. Then, the semantic parser deconstructs text into concepts using a lexicon based on sequences of lexemes that represent multiple-word concepts extracted from AffectNet.

The target spotting module aims to individuate one or more sentiment targets, e.g., people, places, events, and ideas, from the input concepts. This is done by projecting the retrieved concepts into both the graph and the vector space representation of AffectNet, in order to assign these to a specific conceptual class. The categorization does not consist in simply labeling each concept, but also in assigning a confidence score to each category label, which is directly proportional to the value of belonging to a specific conceptual cluster (number of steps in the graph and dot product in the vector space). The affect interpreter, similarly, projects the retrieved concepts into the vector space representation of AffectNet, in order to assign these to a specific affective class and, hence, calculates polarity in terms of the Hourglass dimensions.

## Approach Comparison

In order to evaluate the different facets of the engine from different perspectives, we used a PatientOpinion<sup>2</sup> dataset and compared results obtained using standard AffectiveSpace, default spectral association, and sentic panalogy. The resource is a dataset obtained from PatientOpinion, a social enterprise pioneering an on-line feedback service for users of the UK national health service to enable people to share their recent experience of local health services on-line. It is a manually tagged dataset of 2,000 patient opinions that associates to each post a category (namely, clinical service, communication, food, parking, staff, and timeliness) and a positive or negative polarity.

<sup>1</sup><http://livejournal.com>

<sup>2</sup><http://patientopinion.org.uk>

Category Label	Affective-Space	Spectral Association	Sentic Panalogy
clinical service	62.4%	71.5%	80.0%
communication	62.3%	59.8%	71.2%
food	70.7%	69.6%	78.5%
parking	56.3%	53.7%	69.4%
staff	58.5%	49.2%	63.9%
timeliness	69.2%	61.9%	75.1%

Table 1: F-measures relative to PatientOpinion evaluation.

We used it to test the detection of opinion targets and the polarity associated with these (F-measure values are reported in Table 1).

## Conclusion

Sentic panalogy is novel approach to affective common sense reasoning inspired by Minsky's notion of parallel analogy. It employs different KR strategies and reasoning techniques to maintain several analogous representations of the same problem so that, when a particular strategy begins to fail, the system can switch to one of the others. In the future, we plan to develop heuristics to swap reasoning strategies and foci in real-time, rather than performing all the computations a priori, in order to pave the way for more brain-inspired approaches to affective common sense reasoning.

## References

- Cambria, E., Benson, T., Eckl, C., & Hussain, A. (2012). Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality. *Expert Systems with Applications*, <http://dx.doi.org/10.1016/j.eswa.2012.02.120>.
- Cambria, E., Havasi, C., & Hussain, A. (2012). SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *FLAIRS*. Marco Island.
- Cambria, E., & Hussain, A. (2012). *Sentic computing: Techniques, tools, and applications*. Berlin Heidelberg: Springer.
- Cambria, E., Livingstone, A., & Hussain, A. (2012). The hourglass of emotions. In *Cognitive behavioral systems*. Berlin Heidelberg: Springer.
- Cambria, E., Mazzocco, T., Hussain, A., & Durrani, T. (2011). Switching between different ways to think: Multiple approaches to affective common sense reasoning. In *LNCIS* (Vol. 6800, p. 56-69). Berlin: Springer-Verlag.
- Cambria, E., Olsher, D., & Kwok, K. (2012). Sentic activation: A two-level affective common sense reasoning framework. In *AAAI*. Toronto.
- Chaiken, S., & Trope, Y. (1999). *Dual-process theories in social psychology*. New York: Guilford.
- Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*(3), 211-218.
- Epstein, S. (2003). Cognitive-experiential self-theory of personality. In *Comprehensive handbook of psychology* (Vol. 5, p. 159-184). Hoboken, NJ: Wiley & Sons.
- Fayyad, U., Piatetsky, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. In *Advances in knowledge discovery and data mining* (p. 1-36). Cambridge: MIT Press.
- Havasi, C., Speer, R., & Alonso, J. (2007). ConceptNet 3: A flexible, multilingual semantic network for common sense knowledge. In *RANLP*. Borovets.
- Havasi, C., Speer, R., & Holmgren, J. (2010). Automated color selection using semantic knowledge. In *AAAI CSK*. Arlington.
- Havasi, C., Speer, R., Pustejovsky, J., & Lieberman, H. (2009). Digital intuition: Applying common sense using dimensionality reduction. *IEEE Intelligent Systems*, *24*(4), 24-35.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
- Kaufman, L., & Rousseeuw, P. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley-Interscience.
- Kirkpatrick, L., & Epstein, S. (1992). Cognitive experiential self-theory and subjective probability: Further evidence for two conceptual systems. *Journal of Personality and Social Psychology*, *63*, 534-544.
- Lieberman, M. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, *58*, 259-89.
- Minsky, M. (2006). *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. New York: Simon & Schuster.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP* (p. 79-86). Philadelphia.
- Park, H., & Jun, C. (2009). A simple and fast algorithm for k-medoids clustering. *Expert Systems with Applications*, *36*(2), 3336-3341.
- Plutchik, R. (2001). The nature of emotions. *American Scientist*, *89*(4), 344-350.
- Scherer, K., Shorr, A., & Johnstone, T. (2001). *Appraisal processes in emotion: Theory, methods, research*. Canary: Oxford University Press.
- Simoudis, E. (1996). Reality check for data mining. *IEEE Expert*, *11*(5).
- Smith, E., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychological Review*, *4*(2), 108-131.
- Strapparava, C., & Valitutti, A. (2004). WordNet-Affect: An affective extension of WordNet. In *LREC*. Lisbon.
- Wall, M., Rechtsteiner, A., & Rocha, L. (2003). Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis* (p. 91-109). Springer.
- Wu, H., Luk, R., Wong, K., & Kwok, K. (2008). Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems*, *26*(3).

# Object Discovery and Inverse Physical Reasoning

**Christopher D. Carroll (cdcarroll@gmail.com)**

Department of Psychology, Carnegie Mellon University  
5000 Forbes Ave, Pittsburgh, PA 15213

**Charles Kemp (ckemp@cmu.edu)**

Department of Psychology, Carnegie Mellon University  
5000 Forbes Ave, Pittsburgh, PA 15213

## Abstract

Unobserved objects are typically discovered by making backward inferences from effects to causes. The inverse reasoning account proposes that inferences of this kind are carried out by postulating unobserved causes that best support the forward inference from causes to effects. We evaluated the inverse reasoning account by asking people to reason about hidden attractors and repellers that caused an observed particle to move about an arena. We found that people often evaluated specific hypotheses in a manner consistent with the inverse reasoning account but that hypothesis discovery involved processes that were inconsistent with inverse reasoning.

**Keywords:** object discovery; inverse reasoning; inverse problem; Bayesian inference; physical reasoning

## Introduction

Inferences about unobserved objects are common in both scientific and everyday reasoning. Scientists originally postulated the existence of the planet Neptune to explain perturbations in the orbit of Uranus. Similarly, a jilted lover may postulate the existence of a romantic competitor in order to explain the behavior of his or her partner. This paper describes an experimental study of object discovery that is loosely inspired by the discovery of Neptune. Participants observed particles that moved along paths such as the one in Figure 1 and attempted to infer the unobserved attractors and repellers responsible for the particle's motion.

Object discovery typically involves reasoning from effects (e.g., an observed motion) to causes (e.g., an unobserved attractor). Here we refer to inferences from causes to effects as *forward inferences* and inferences from effects to causes as *backward inferences*. We explore the hypothesis that forward and backward reasoning are tightly coupled, and that backward inferences are made by postulating unobserved causes that best support the forward

inference from causes to effects. We refer to this approach as *inverse reasoning* because it achieves backward reasoning by inverting the process of forward reasoning.

One natural way to formalize the inverse reasoning approach makes use of Bayesian inference, which specifies the normative relationship between backward and forward reasoning. Specifically, given some observations  $D$  and a hypothesis  $H$  about the existence and properties of the unobserved causes, Bayes' theorem requires that

$$P(H|D) \propto P(D|H)P(H). \quad (1)$$

Backward and forward reasoning are captured by the posterior  $P(H|D)$  and likelihood  $P(D|H)$ , respectively. Bayes' theorem therefore suggests that backward reasoning should be carried out by combining the forward inferences specified by the likelihood with judgments of plausibility specified by the prior  $P(H)$ . In our setting, this approach suggests that a configuration of unobserved attractors and repellers is a good explanation for a particle's motion to the extent that (1) the configuration predicts the particle's motion and (2) the configuration is relatively parsimonious. Inverse reasoning implies that backward inferences will be consistent with forward inferences, but does not imply that backward inferences will always be accurate. Studies of physical reasoning have documented situations where people's forward inferences deviate from the predictions of classical mechanics (e.g., Clement, 1982; McCloskey, 1983), and faulty forward inferences could produce faulty backward inferences through inverse reasoning.

The inverse reasoning approach has a mixed record as an account of human reasoning. On one hand, the approach has been successfully used to develop models of causal reasoning (e.g., Griffiths & Tenenbaum, 2005), perception (e.g., Yuille & Kersten, 2006), sensorimotor control (e.g., Kording & Wolpert, 2006) and social reasoning (e.g., Baker,

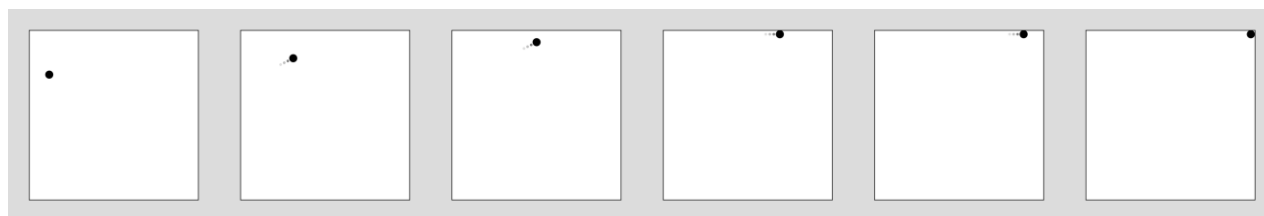


Figure 1: This sequence of bird's-eye-view snapshots shows a particle's motion over time. The particle in this "wall-motion" scene moved along a diagonal path until it reached the top wall. It then continued along the wall.

Saxe, & Tenenbaum, 2009). On the other hand, psychologists have documented several respects in which backward inferences seem inconsistent with inverse reasoning (e.g., Kahneman, Slovic, & Tversky, 1982; Fernbach, Darlow, & Sloman, 2011). People often, for example, erroneously ignore or underutilize the prior when estimating the posterior (Bar-Hillel, 1980).

Based on these findings it is not clear whether the object discovery task considered in this paper should produce results that are consistent with inverse reasoning. Because physical reasoning is a core aspect of cognition that is present early in development (Spelke, Breinlinger, Macomber, & Jacobson, 1992), one might expect that backward physical reasoning will tend to be consistent with normative inverse reasoning. Previous studies of physical reasoning provide some evidence for this claim. For example, Sanborn, Mansinghka, and Griffiths (2009) found that backward inferences about the relative masses of colliding objects were consistent with a Bayesian account of inverse reasoning. Object discovery, however, appears to be more challenging than the tasks considered by previous studies of backward physical reasoning. Inferring hidden properties of observed objects (e.g., the mass of a colliding object) is a relatively well-constrained problem, but object discovery is a more open-ended problem that involves inferring the existence and number of the hidden objects, the locations of those objects, and the properties of those objects. To preview our results, we found that when participants evaluated specific hypotheses about the locations and properties of the hidden objects, their inferences were broadly consistent with inverse reasoning. When asked to generate their own explanations, however, many participants gave responses that were incompatible with the inverse reasoning account.

## Experimental overview

To explore the problem of object discovery we conducted an experiment where participants reasoned about “attractors” and “repellers” that controlled the movements of some observed “particles.” The attractors and repellers were unobserved, and participants attempted to infer their locations given the observed particle motions.

There were three experimental phases: the discovery, prediction, and evaluation phases. In the discovery phase, participants observed the motion of a particle and were asked to infer the locations of hidden attractors and repellers. In the prediction phase, participants were given the locations of one or more attractors or repellers and were asked to predict the trajectory that a particle would follow. In the evaluation phase, participants were given two possible explanations of a particle motion and were asked to decide which explanation was better. Note that the discovery and evaluation phase both assessed backward reasoning and that the prediction phase assessed forward reasoning.

The simplest possible observed trajectory is a straight line, and the obvious explanation for this trajectory is that the particle is either moving towards an attractor or moving away from a repeller. The particle motions presented in the discovery phase (Figure 2.i) include some of the next simplest cases. Each motion can be explained in at least two ways. First, there is a parsimonious explanation that invokes a relatively small number of stationary attractors and repellers. For example, the “wall-motion” scene (Figure 1 and Figure 2.a.i) can be explained by assuming a single repeller (see the first row of Figure 2.ii). Second, each explanation had a less parsimonious explanation where the particle always moved directly towards an attractor or directly away from a repeller, but where the attractors and repellers spontaneously appeared, disappeared, or moved. The second row of Figure 2.ii shows a less parsimonious explanation of the wall-motion scene.

Our primary goal is to explore whether participants generate the parsimonious explanations during the discovery and evaluation phases. If participants agree that the parsimonious explanations are in fact parsimonious and valid, then the inverse reasoning account predicts that these explanations should be generated during the discovery phase and rated favorably during the evaluation phase. If participants fail to generate these explanations in the discovery phase but tend to prefer them in the evaluation phase, this result would be inconsistent with the inverse reasoning view.

The prediction phase asked participants to generate particle trajectories for several kinds of configurations. Each configuration can be viewed as an explanation (plausible or implausible) of a motion observed during the discovery phase. Some of the prediction trials presented participants with their own explanations from the discovery phase. For our purposes, however, the most important prediction trials are those that presented participants with the parsimonious explanations for the three motions in Figure 2.i. Including these trials allowed us to assess whether participants agreed that the parsimonious explanations could in fact explain the observed motions – if not, it would be unsurprising if these explanations were rarely chosen during the discovery phase.

## Method

### Participants

Thirty undergraduates at Carnegie Mellon University participated for course credit.

### Materials and Procedure

Participants were asked to imagine themselves working for a scientist who studies “attractors” and “repellers.” The instructions explained that the participants would view scenes where “particles” moved within a rectangular arena. Participants learned that the particle motions were caused by attractors and repellers located outside the arena.



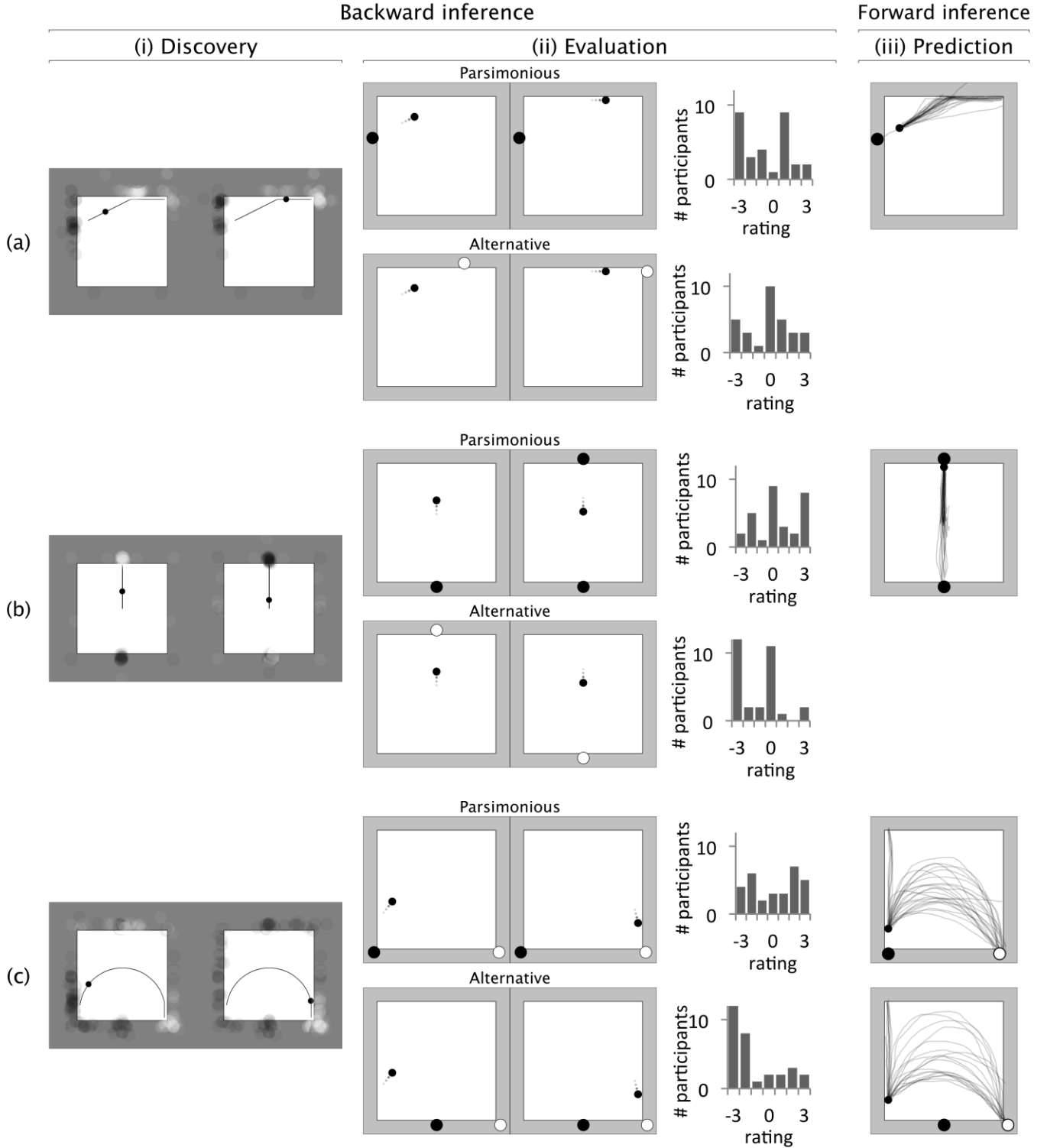


Figure 2: Experimental method and results for the (a) wall-motion, (b) center-return, and (c) curved-motion scenes. (i) The discovery phase. The paths illustrate the motion of the particle, and the circles illustrate the location of the particle in the first and second response pictures. (The particle in the center-return scene moved from its initial position to the top wall, paused, and then returned to the center.) The surrounding area is a heatmap. Areas where attractors were often placed are shown as brighter areas and areas where repellers were often placed are shown as darker areas. (ii) The evaluation phase. The pictures at left show the experimenter-provided explanations. Repellers and attractors are shown as large black and large white circles, respectively. The histograms show the preference ratings (-3 = strongly preferred own explanation; 3 = strongly preferred experimenter-provided explanation). (iii) The prediction phase. The paths in the figure represent the paths drawn by the participants. Trivial prediction trials (i.e., those involving a single attractor or repeller) are not shown.

Participants then viewed three scenes that demonstrated the properties of the attractors and repellers. Each scene was displayed as a sequence of bird's-eye-view snapshots showing the motion of the particle over time. The first two scenes showed that particles move towards attractors and away from repellers, depicted as green and red circular objects, respectively. The third scene showed that a particle placed between two attractors moved towards the closer one, and the instructions explained that distant attractors and repellers exert less force than close ones.

### Discovery phase

Participants were asked to explain a number of scenes where the attractors and repellers were not visible. After completing a practice trial, participants generated explanations for the three scenes in Figure 2.i. Participants also generated explanations for 12 variants of the three primary scenes, but we do not discuss these results here because the variant scenes did not have analogues in the prediction and evaluation phases. In the wall-motion scene (Figure 2.a.i), the particle traveled along a diagonal until it reached the top wall of the arena. It then continued along the top wall of the arena. In the center-return scene (Figure 2.b.i), the particle moved from the center of the arena to the top wall, paused, and then returned to the center. In the curved-motion scene (Figure 2.c.i), the particle moved along a curved path from the lower-left corner of the arena to the lower-right corner of the arena.

Participants explained each particle motion by specifying where the attractors and repellers would have been at two different points in the particle's motion (see Figure 2.i). The instructions explained that the participants were being asked to report the locations of the attractors and repellers in two distinct response pictures because "there may be some situations where you think that something has changed." Responses were made using a computer interface that showed the two response pictures and a summary of the to-be-explained particle motion. Participants could place attractors and repellers by clicking on any location outside the arena. Participants could move or erase placed attractors and repellers. A "reuse" button located between the two response pictures copied the attractors and repellers in the first picture to the second picture.

Participants were allowed to provide up to three explanations for each scene. Each explanation was entered on a separate screen. Participants were allowed to provide written explanations to supplement the picture-based explanations, but few participants did so.

After providing the explanations, the participants rated each provided explanation on a scale ranging from 1 (very unlikely to be the true explanation) to 7 (very likely to be the true explanation). Participants were also asked to rate the likelihood that the true explanation was "fundamentally different" from the provided explanation(s).

### Prediction phase

Participants were asked to predict the particle paths given the locations of the attractors and repellers. Figure 2.iii

presents some of the prediction trials. The prediction pictures in Figure 2.a.iii, b.iii, and c.iii top correspond to the parsimonious explanations. There were other prediction trials that corresponded to less parsimonious explanations. Three other prediction trials presented each participant with the configurations that corresponded to his or her own explanations in the discovery phase.

### Evaluation phase

In the evaluation phase, participants once again viewed the wall-motion, center-return, and curved-motion scenes. In explaining each scene, participants chose between their own explanations and the parsimonious explanation. The parsimonious explanations are shown as the first, third, and fifth rows in Figure 2.ii.

For each forced choice, the participant rated his or her preferred explanation as "much more," "more", or "slightly more" likely to be the true explanation than the competing explanation. Because participants occasionally generated the parsimonious explanations themselves, participants were sometimes presented with a choice between two identical explanations. For these situations, participants were provided with a "these explanations are identical" button. We coded responses on a scale ranging from -3 (own explanation "much more likely" to be the true explanation) to 3 (parsimonious or alternative explanation "much more likely" to be the true explanation). When a participant claimed that the explanations were identical, his or her preference was coded as 0.

Three other trials required the participants to choose between their own explanations and some less parsimonious explanations. These alternative explanations, shown in the second, fourth, and sixth rows of Figure 2.ii, required additional assumptions to explain the particle motion. These trials served to control for the task demand of asking the participants to choose between their own explanation and an experimenter-provided explanation. To further limit any task demands, all competing explanations were described as responses provided by other participants.

## Results

The inverse reasoning account predicts that participants ought to generate the parsimonious explanations during the discovery phase and endorse them during the evaluation phase. In contrast, we found that participants rarely generated the parsimonious explanations during the discovery phase but often preferred them during the evaluation phase. We begin by documenting this general result and then provide more detailed descriptions of the results for the discovery and prediction phases.

### Parsimonious explanations

A wall-motion explanation was coded as parsimonious when it invoked a single stationary attractor or repeller. A center-return explanation was coded as parsimonious when it invoked two balancing repellers above and below the arena or two balancing attractors to the left and right of the

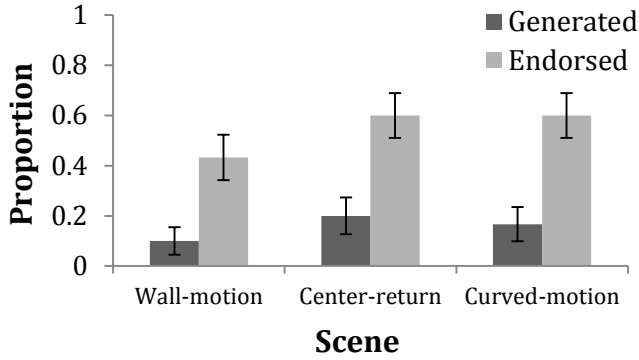


Figure 3: Proportions of the participants generating a parsimonious explanation in the discovery phase and endorsing the parsimonious explanation in the evaluation phase.

arena. A curved-motion explanation was coded as parsimonious when it invoked exactly two stationary attractors or repellers and did not invoke any moving, appearing, or disappearing attractors and repellers. Our coding criteria were intended to be conservative: note, for example, that any curved-motion explanation with two stationary objects was coded as parsimonious regardless of the locations of these objects.

Figure 3 shows that participants rarely generated parsimonious explanations in the discovery phase but often preferred them in the evaluation phase. The differences between the rates of generation and endorsement were significant for each scene (Fisher's exact test yields  $p < .01$  in all cases). This finding cannot be attributed to task demands alone: as shown by the distribution of the preference ratings in Figure 2.ii, participants did not prefer non-parsimonious explanations (rows two, four, and six) to the same extent that they preferred the parsimonious explanations (rows one, three, and five).

## Discovery

The difference between the results for the discovery and evaluation phases suggests that object discovery in our paradigm is not accurately characterized as inverse physical reasoning. Figure 2.i gives some sense of how participants were approaching the discovery task. Each plot in this column is a "heatmap:" locations where participants often placed attractors are shown as brighter areas and locations where repellers were often placed are shown as darker areas.

For the wall-motion trials, 14 participants posited one hidden object along the particle's diagonal trajectory and one hidden object along its horizontal trajectory, 5 participants posited a single attractor or repeller that moved, and 8 participants generated combinations or variations of those explanations. For the center-return trials, 12 participants posited appearing and disappearing attractors and repellers along each path of motion, 9 generated an explanation that involved balancing attractors or repellers but also invoked other attractors and repellers (e.g., had balancing repellers in the second response picture but

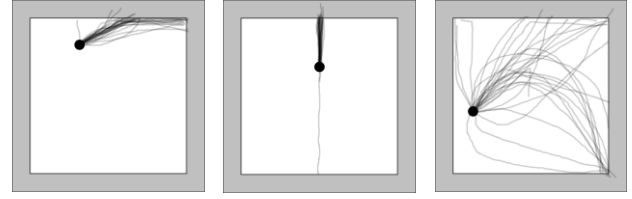


Figure 4: Particle motions predicted by participants given their own explanations for the wall-motion, center-return, and curved-motion scenes.

posited an attractor in the first response picture), and 5 participants generated other explanations. Responses to the curved-motion scene were more variable, and there was less agreement on the locations of the attractors and repellers (see Figure 2.i.c). Most of the participants posited multiple attractors and repellers that were simultaneously present. For example, 7 participants posited three or more stationary and constantly present attractors and repellers, and 12 participants posited two or more attractors or repellers that were simultaneously present at some point during the motion but were either non-stationary or not constantly present. The remaining non-parsimonious explanations most commonly posited a single attractor or repeller along the particle's path of motion in each response picture. Overall, then, responses to the discovery phase reveal a variety of strategies, but one consistent element is that many participants placed objects in line with a particle's instantaneous direction of motion.

## Prediction

Figure 2.iii summarizes the responses on selected prediction trials. The critical question for present purposes is whether participants agreed that the parsimonious explanations would indeed account for the observed motions during the discovery trials. When provided with the parsimonious explanations, 15 of the 30, 24 of the 30, and 23 of the 30 participants predicted that the particle would approximately reproduce the particle motions from the discovery trials for the wall-motion, center-return, and curved-motion scenes respectively. Note that these counts are substantially higher than the number of participants who generated the parsimonious explanations during the discovery phase. The prediction data therefore provide further evidence that some participants failed to generate the parsimonious explanations during the discovery phase even though they considered these explanations to be valid.

Although participants did not always predict that the parsimonious explanations would produce the observed motion, their predictions were usually sensible given some additional assumptions. For example, various participants seemed to assume that friction would stop the particle when it hit the wall in the parsimonious wall-motion prediction trial, that momentum would carry the particle past the center in the parsimonious center-return prediction trial, and that the motion of the particle would be influenced *only* by

nearby repeller in the parsimonious curved-motion prediction trial.

Figure 4 shows the participants' predictions given their own explanations for the wall-motion, center-return, and curved-motion scenes. Some predicted motions diverged dramatically from the particle motion in the to-be-explained scene, and the discrepancies for the curved-motion scene were especially dramatic. These discrepancies should be interpreted cautiously, however, because the participants may have made different assumptions during the discovery and prediction phases. For example, it was natural to assume that the particle had an initial velocity in the first response picture of a trial in the discovery phase (the particle had already moved), but there was no reason to assume a particle velocity in the prediction phase. As a result, future studies are needed before concluding that participants sometimes generate explanations that are truly incompatible with the trajectories that they have observed.

## Discussion

Our data support the conclusion that hypothesis evaluation is consistent with the inverse reasoning account but that hypothesis discovery is not. In some respects, the failure of the participants to discover the parsimonious explanations is quite surprising. The parsimonious explanations were straightforward, requiring the participant to posit at most two stationary attractors or repellers. It should have been possible for participants to discover the parsimonious explanations, and some of them indeed did so. In other respects, the failure of the participants to discover the parsimonious explanation makes sense. Even in the simple object discovery task presented in this paper, there are infinitely many explanations that might be considered. The inverse reasoning account is often unhelpful in these situations. Bayes' theorem admonishes the reasoner to consider all the possible explanations for the observations, but does not provide guidance when doing so is impossible.

Although generating the best explanation from an infinite class may be computationally challenging, evaluating the merits of a handful of selected hypotheses seems substantially easier. It is therefore not surprising that hypothesis evaluation was broadly – although perhaps not absolutely – consistent with the normative inverse reasoning account. The dissociation between discovery and evaluation is consistent with the view that people rely on non-Bayesian strategies to generate candidate explanations for evaluation, but are able to approximate Bayesian reasoning when deciding which of these candidates is best (Bonawitz and Griffiths, 2010).

Participants may have used several different kinds of strategies to generate candidate explanations during the discovery phase of our experiment. For example, an initial explanation might have been generated using the idea that objects often move directly towards attractors. If needed, this initial explanation might have been improved using search heuristics such as hill-climbing. The process of discovery might also rely on analogical reasoning—for

example, many participants explained the curved-motion scene by placing a repeller at the focal point of the curve, and it is tempting to view this inference as a loose analogy to orbital motion. Like any other kind of creative behavior, object discovery is likely to be difficult to characterize in full detail. Future studies, however, can aim to characterize some of the psychological processes involved.

## Acknowledgments

This work was supported in part by NSF grant CDI-0835797 and by the Pittsburgh Life Sciences Greenhouse Opportunity Fund.

## References

- Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*, 329-349.
- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica* (44), 211-233.
- Bonawitz, E. B., & Griffiths, T. L. (2010). Deconfounding hypothesis generation and evaluation in Bayesian models. In S. Ohlsson, & R. Catrambone (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2260-2265). Austin, TX: Cognitive Science Society.
- Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics*, *50*, 66-71.
- Fernbach, P. M., Darlow, D., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, *21* (3), 329-336.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334-384.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.
- Kording, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, *10* (7), 319-326.
- McCloskey, M. (1983). Intuitive physics. *Scientific American*, *24*, 122-130.
- Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117* (4), 1144-1167.
- Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99* (4), 605-632.
- Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10* (7), 301-308.

# Category structure modulates interleaving and blocking advantage in inductive category acquisition

**Paulo F. Carvalho (pcarvalh@indiana.edu)**

Department of Psychological and Brain Sciences, 1101 E 10th St  
Bloomington, IN 47405 USA

**Robert L. Goldstone (rgoldsto@indiana.edu)**

Department of Psychological and Brain Sciences, 1101 E 10th St  
Bloomington, IN 47405 USA

## Abstract

Research in inductive category learning has demonstrated that interleaving exemplars of categories results in better performance than presenting each category in a separate block. Two experiments indicate that the advantage of interleaved over blocked presentation is modulated by the structure of the categories being presented. More specifically, interleaved presentation results in better performance for categories with high within- and between-category similarity while blocked presentation results in better performance for categories with low within- and between-category similarity. This interaction is predicted by accounts in which blocking promotes discovery of features shared by the members of a category whereas interleaving promotes discovery of features that discriminate between categories.

**Keywords:** category learning; order effects; interleaving;

## Introduction

How to present information so that learning and memory are optimized is an important issue in teaching and training contexts (Rohrer & Pashler, 2010). It has long been demonstrated that spacing repeated presentations of the same information results in better memory than repeating the same information at a single occasion, even when time and number of presentations are equated (Ebbinghaus, 1885). This memory phenomenon, known as the “Spacing Effect,” is a highly robust finding (Delaney, Verkoeijen, & Spiguel, 2010; Proctor, 1980) that has been shown both in experimental situations with words and pictures and more applied situations such as flashcard studying (Kornell, 2009). Although demonstrating the critical importance of carefully considering how to present information, the importance of maximizing memory for specific concepts or problems might not be as relevant as learning general concepts. Indeed, in educational contexts, often times inferring what characterizes or defines a concept or problem is more relevant than memorizing a single instance of that concept or fact. In this sense, a more interesting question might be to know whether the way instances are presented influences inductive learning and subsequent generalization of the acquired knowledge.

The question of how to present information in order to optimize category learning and generalization has been raised before and several proposals have been put forward. Some of these proposals are related to the categories being

taught. For example, Elio and Anderson (1984; see also, Sandhofer & Dumas, 2008) have proposed that learning should start with low variability items and later introduce items with greater variability. Another proposal is that items that present the same generalization should be presented close together in temporal sequence (e.g., Elio & Anderson, 1981; Mathy & Feldman, 2009).

More recently, researchers have proposed interleaving items from the categories being taught (i.e., presented in alternating fashion), rather than grouping items together from the same category. The advantage of alternating categories has been observed in different kinds of concepts such as artists’ styles (Kornell & Bjork, 2008), bird species (Kornell, Castel, Eich, & Bjork, 2010; Wahlheim, Dunlosky, & Jacoby, 2011), novel category learning in children (Vlach, Sandhofer, & Kornell, 2008) and also mathematical operations in primary school students (Taylor & Rohrer, 2010).

Initial accounts of this advantage for interleaving related the interleaved presentation with spacing of exemplars. However, Kang and Pashler, (2012) used a procedure similar to the one used by Kornell and Bjork (2008), but with added presentation conditions. In one experiment, the authors compared categorization performance in a generalization test preceded by one of four conditions: (1) blocked, (2) interleaved, (3) blocking in which every presentation of a painting was followed by an unrelated filler task (Temporal Spaced Condition), and (4) when all exemplars from the same painter were presented simultaneously (Blocked Simultaneous Condition). The results showed that only the interleaved condition resulted in better performance than the blocked condition, thus providing evidence that greater temporal spacing of presentations is not the critical factor in the interleaved advantage. The authors argue that the real advantage of interleaving might be a result of the greater opportunity to contrast and compare examples, making the differences between the artists’ styles more salient (see also, Goldstone, 2003; Goldstone & Steyvers, 2001).

Further evidence for this proposal was provided in a second experiment in which the interleaved and blocked conditions were compared to a simultaneous presentation of two paintings by different artists. This latter condition

resulted in similar performance to interleaved presentation and better performance than the blocked condition.

Interleaving examples from different categories, thus, may improve inductive learning because it promotes discrimination between the categories. Rapid alternation between examples of different categories, as well as simultaneous presentation of multiple categories, might lead to enhanced attention to the features that discriminate between the categories and differentiation of the dimensions on which the categories vary (Goldstone & Steyvers, 2001; Kang & Pashler, 2012; Nosofsky, 1986). However, there are also situations in which these conditions are not ideal and previous studies have also demonstrated an advantage of blocked presentation (Goldstone, 1996; Kurtz & Hovland, 1956).

Goldstone (1996) proposed that category learning might be difficult for two different reasons: high between-category similarity or low within-category similarity. High between-category similarity refers to category structures in which both categories share most of their features, making discriminating the categories a matter of finding subtle differences between exemplars (as is the case of distinguishing between alligators versus crocodiles, for example). Low within-category similarity, on the other hand, refers to category structures in which the exemplars of the same category share very few features (as is the case for the category “animal,” for example).

Each one of these two kinds of categories requires different mechanisms for efficient category learning. In the case of high between-category similarity, one has to identify subtle differences between categories, which might be facilitated by frequent alternation. However, rapid alternation between two categories with low within-category similarity will not allow for the identification of the relevant properties characteristic of a category. In this case it might be more beneficial to block exemplars of each category separately so that the learner can identify the shared features among members of a category hidden within their diversity.

By this analysis, the best presentation schedule would depend upon the learning situation at hand. In this paper we approach this question directly by manipulating only the characteristics of the stimuli presented during learning. We aim to demonstrate that interleaving or blocking can both be beneficial for inductive category learning. In Experiment 1 participants were taught three categories with high within- and between-category similarity, in both an interleaved and blocked schedule. In Experiment 2 we employ the same learning procedure but using three categories that possess low within- and between- category similarity. Performance during category learning and in a subsequent generalization task are measured.

### Experiment 1: High Similarity

In this experiment we aim to replicate previous results showing an advantage for interleaved presentation of categories for category learning. Participants were presented with 3 categories (either blocked or interleaved). These 3

categories had high within- and between- category similarity. In this sense, every stimulus from the same category shared several features, some category relevant and others not, while they also shared almost every feature with every stimulus from the other categories. We believe this is a situation close to previous work on sequencing effects in category learning (e.g., Kornell & Bjork, 2008) and literature on perceptual discrimination learning showing an interleaving advantage (Mitchell, Kadib, Nash, Lavis, & Hall, 2008; Mundy, Honey, & Dwyer, 2007). Under this situation, rapid alternation between categories will lead to the direction of attention to the diagnostic features because identifying what is changing from category to category is relatively easy given the high overall similarity.

### Method

**Participants** Forty-four Indiana University undergraduate students participated in this experiment in return for partial course credit. All participants completed every condition. Fifteen participants did not reach the criterion of 34% correct responses during categorization learning and were excluded from further analyses.

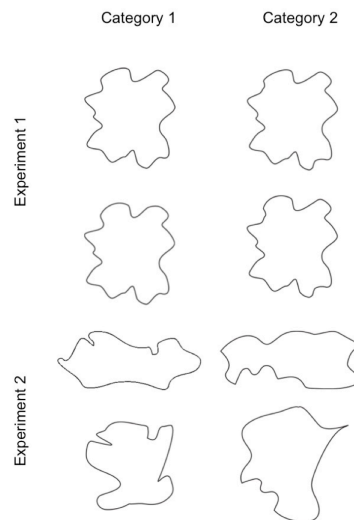


Fig. 1: Examples of stimuli used in Experiments 1 and 2.

**Apparatus and stimuli** In this and the subsequent experiment, blob figures were used as stimuli (see Fig. 1). All blobs were created by randomly generating curvilinear segments. The blobs used were very similar, sharing almost every feature. Variation within each category was exactly the same for all categories, so that a difference that could exist between two exemplars in Category 1 would also exist between 2 exemplars of Categories 2 and 3. As a cover story, participants were told that a recent expedition to Mars recovered several cells of alien organisms. Each cell could be categorized into one of 3 species based solely on its perceptual features. Stimuli were presented on a computer screen and participants responded using keys on the keyboard with a consistent mapping to the category assignment.

**Design and procedure** This experiment had two conditions (blocked category learning and interleaved category learning), manipulated within-subject. Each condition was composed by 2 phases. The first phase was a category learning task. During this task participants were presented with a stimulus in the center of the screen for 500 ms. After the blob was removed, the participant was asked to classify the blob into one of 3 species (Q, Y or P) by pressing the corresponding key on the keyboard. After a participant's response, the blob was presented again for 2000 ms together with the presentation of feedback on the accuracy of their response and the correct species of the blob (e.g., "CORRECT! This cell belongs to species Q" or "Sorry, that is INCORRECT! This cell belongs to species Y").

A 1000 ms inter-trial interval followed and then a new trial began. In the blocked condition, the categories presented alternated 25% of the time while in the interleaved condition they alternated 75% of the time. Thus, in the interleaved condition, the probability of a blob being followed by a blob of the same category was low, whereas for the blocked condition this probability was high. We used this probabilistic approach rather than creating purely interleaved or blocked conditions in order to diminish the possibility that participants noticed the pattern of alternation in responses, which would affect categorization accuracy. Furthermore, if a purely blocked condition had been used there would be no way to guarantee participants' attention to the task, as there would be no uncertainty as to the correct categorization. This approach has been used before in similar tasks with successful results (Carvalho & Goldstone, 2011; Goldstone, 1996).

The learning phase was composed of 4 blocks. Each block had 48 trials (8 exemplars of each category repeated 2 times each). After the 4<sup>th</sup> block of categorization learning a new set of instructions was presented on the screen and the second phase began. This second phase was a generalization task during which 48 stimuli were shown in random order – the 24 blobs participants studied during the learning task and 24 new stimuli. The new stimuli were generated in the same manner as training stimuli, with new instantiations of the unique features. Each stimulus was presented in the center of the screen for 500 ms, after which participants were asked to classify it into one of the species just learned. After a 1000 ms inter-trial interval, a new trial would begin. No feedback was provided during this phase.

The two learning conditions (blocked vs. interleaved) differed only in the frequency of category change and the species labels. In one of the conditions Q, Y and P labels/keys were used, while in the other, A, G and L were used, by random assignment. Which condition was presented first was counterbalanced across participants and the allocation of the stimuli to each category and condition was randomized across participants.

## Results and Discussion

The top panel of Fig. 2 depicts the main results of the categorization task. A 2 x 4 repeated measures ANOVA

with learning condition (blocked vs. interleaved) and learning block (1 vs. 2 vs. 3 vs. 4) as factors revealed main effects of learning block,  $F(3,81) = 59.42$ ,  $p < .00001$  and condition,  $F(1,27) = 18.58$ ,  $p = .0002$ , but no interaction between the two variables,  $F(3,81) = 1.43$ ,  $p = .24$ . Thus, there is an improvement in categorization accuracy during the task, regardless of presentation schedule, and blocked presentation results in the overall best accuracy rates.

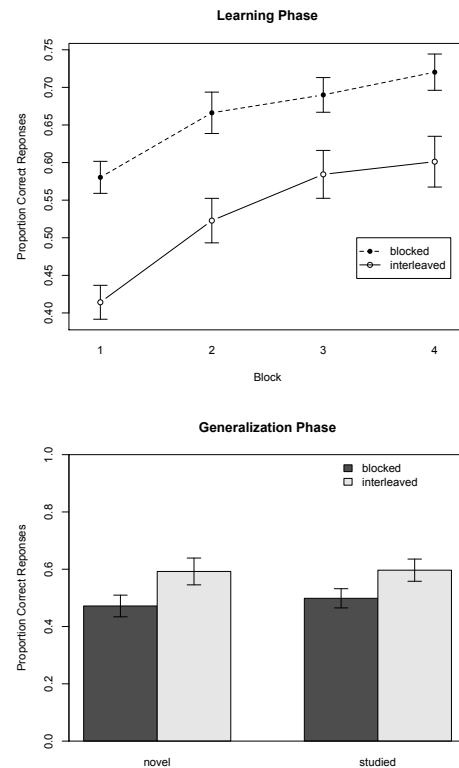


Fig. 2: Results from Experiment 1. The top graph represents accuracy during the learning phase for both conditions across the four blocks of trials. The bottom graph represents accuracy during the generalization phase for novel and studied stimuli for each condition. Error bars represent standard errors.

The results from the generalization task are depicted in the bottom panel of Fig. 2. A 2 x 2 repeated measures ANOVA with learning condition (blocked vs. interleaved) and stimulus presented (novel vs. studied) as factors, revealed a main effect of condition,  $F(1,27) = 4.50$ ,  $p = .04$ , with higher accuracy for the interleaved condition, but no main effect of stimulus presented,  $F(1,27) < 1$ , or interaction between the two variables,  $F(1,27) < 1$ .

Overall these results replicate previous demonstrations that interleaving very similar categories during learning results in better memory for the presented items (as seen by the higher accuracy rates for the studied stimuli) but also better generalization to novel items. The finding that blocked presentation results in higher accuracy rates during learning has also been demonstrated before and has been explained by a relative ease in performing the training task (Taylor & Rohrer, 2010), which would make the interleaved

condition a desirable difficulty in this kind of task (Kornell & Bjork, 2008).

The results are consistent with the hypothesis that interleaved presentation promotes the discovery of features that discriminate between alternative categories, which is important when categories are highly similar.

## Experiment 2: Low Similarity

In this Experiment we analyze the effect of the schedule of presentation in a categorization situation different from that of Experiment 1. We tried to create a category learning task in which blocked presentation of the category exemplars would allow for the discovery of hard-to-find common features among dissimilar category members.

We employ the same procedure as in Experiment 1 but the 3 categories now have low within- and between-category similarity. In this sense, exemplars from the same category share only one feature (the category relevant feature) while differing in all other features. Moreover, exemplars from different categories are also highly dissimilar, sharing none of their features. Thus, in this condition, it is not expected that interleaving will have a beneficial effect on learning. Rapid alternation of categories would not allow for the identification of important features because all the features change from trial to trial. On the contrary, we expect blocked presentation to result in superior category learning and performance on a later generalization task, by promoting comparison of objects from the same category and thereby promoting discovery of their subtle commonality.

## Method

**Participants** Thirty-two Indiana University undergraduate students participated in this experiment in return for partial course credit. All participants completed every condition and all reached the criterion of 34% correct responses or more during categorization learning.

**Apparatus and stimuli** The same kind of blob figures as in Experiment 1 were used. However, the blobs used in this experiment differed from every other blob in all the segments except the one that defined each one of the six categories (each one of the discriminating features). In this experiment, Blobs also differed in their overall shape, some being more circular and others more elongated. All other details were kept the same as in Experiment 1.

**Design and procedure** This experiment followed the same procedure as Experiment 1 with the exception of the stimuli used.

## Results and Discussion

The graph in the top panel of Figure 3 depicts response accuracy during category learning as a function of both learning block and schedule of presentation.

A 2 x 4 repeated measures ANOVA with both learning condition (blocked vs. interleaved) and learning block (1 vs. 2 vs. 3 vs. 4) as repeated measures factors, revealed a main effect of learning block,  $F(3, 93) = 169.74$ ,  $p < .0001$ ,

showing an overall improvement for both presentation conditions. Moreover, categorization is overall better for the blocked condition when compared to the interleaved condition,  $F(1, 31) = 29.03$ ,  $p < .0001$ . Finally, a significant interaction between presentation condition and learning block was also found,  $F(3, 93) = 7.06$ ,  $p = .0002$ , indicating that the difference in accuracy between the two conditions decreases as the categorization progresses.

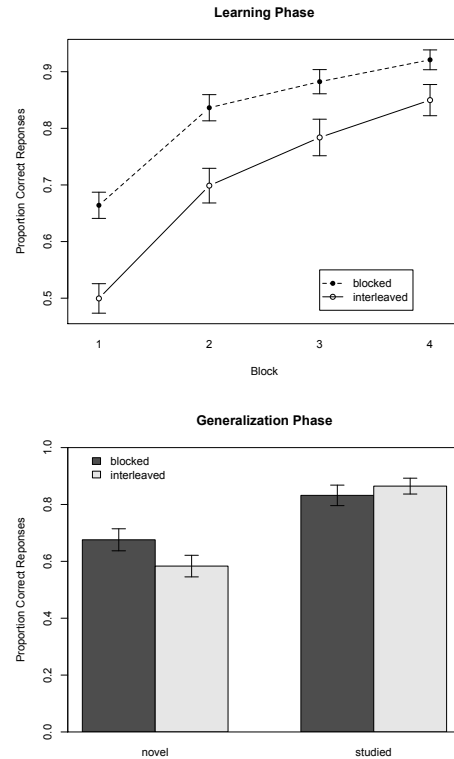


Fig. 3: Results from Experiment 2. The top graph shows accuracy during the learning phase for both conditions across the four blocks of trials. The bottom graph shows accuracy during the generalization phase for novel and studied stimuli for each condition.

The bottom panel of Fig. 3 shows accuracy performance in the generalization task as a function of stimulus presented (novel vs. studied) and presentation condition during learning (blocked vs. interleaved). A 2 x 2 repeated measures ANOVA with both presentation condition (blocked vs. interleaved) and stimulus presented (novel vs. studied) as factors revealed a main effect of stimulus,  $F(1, 31) = 74.20$ ,  $p < .00001$ , with higher accuracy for the studied than novel stimuli. There was no significant main effect of presentation condition during learning,  $F(1, 31) = 1.14$ ,  $p = .30$ . However, the interaction between the two variables was significant,  $F(1, 31) = 14.61$ ,  $p = .0006$ . To further investigate this interaction, pairwise *t*-test comparisons were performed correcting for the error associated with multiple comparisons using Bonferroni correction (critical  $\alpha = .025$ ). This analysis revealed that blocked presentation results in higher accuracy in categorizing novel stimuli,  $t(31) = 2.47$ ,  $p = .02$ , but there is



no difference between the two conditions for studied stimuli,  $t(31) = 1.21, p = .24$ .

Thus, the results of this experiment demonstrate an advantage of blocked presentation of category exemplars in accuracy during learning, although participants in the interleaved condition approach that performance by the last trial. Perhaps more interestingly, blocked presentation resulted in better performance when categorizing new stimuli when compared to interleaved presentation. However, for studied stimuli no difference was found.

These results directly challenge the proposal that interleaving categories during learning improves performance due to the increased spacing of exemplars. They are, however, consistent with our proposal that the interleaving advantage was due to greater opportunity for contrasting categories and discovery features that distinguish between categories. When finding feature differences between categories is easy, as in Experiment 2, then the more important task is to find features shared by members of the same category, a task promoted by blocking. This clearly demonstrates that the category structure has a modulating effect on the advantage of one sequencing over another.

## General Discussion

Learning the characteristics of concepts and categories inductively takes place frequently. In learning contexts, it is also important to be able to generalize that knowledge to new exemplars. In this work we studied the interaction between object presentation during learning and the structure of the category being taught for inductive learning and generalization optimization.

Experiment 1 replicated previous results demonstrating that interleaving categories being taught results in better memory and generalization for those categories. Additionally, in Experiment 2 we presented results pointing to an advantage of blocking by category. In Experiment 2 generalization was better for categories that had their exemplars presented in adjacent temporal sequence during learning.

Critically, the only difference between these two experiments was the properties of the categories being taught. While in Experiment 1, the three categories were highly similar among each other and were constituted by exemplars that were also very similar, in Experiment 2 the similarity within and between the categories was low, resulting in categories that did not share any features and were composed of stimuli that shared only one feature with other objects in their same category.

A possible explanation to at least some of the results presented here could be that interleaving constitutes a “desirable difficulty” (Bjork, 1994). In both experiments, blocked presentation during the learning phase resulted in a higher proportion of correct responses. This could indicate that interleaved presentation involved a greater effort on the participants’ part, maintaining their attention to the stimuli and leading to better performance in the subsequent task.

Although this can, in fact, account for the results found in Experiment 1, that is not the case for Experiment 2. In that experiment, the condition resulting in greater accuracy during learning also resulted in better generalization.

In line with Goldstone (1996) we propose that interleaving categories allows for the identification of the features that are different between each category while blocked presentation promotes the identification of the features that are common among stimuli from the same category. This dichotomy is the result of the same principle: the opportunity to compare and contrast the properties of the categories, what will emphasize different features in different situations,

We further build on this proposal by suggesting a mechanistic account of how it takes place in a trial by trial basis. It has previously been demonstrated that during category learning participants take into account information from the previous few trials to decide whether a stimulus belongs to one category or another (Stewart & Brown, 2004; Stewart, Brown, & Chater, 2002). We further propose, that when two successive stimuli are similar and in different categories, participants’ attention will be directed to the differences between exemplars by a comparison process. On the other hand, if stimuli are very dissimilar but in the same category, participants’ attention will be directed towards the similarities between successive stimuli. This proposal can account for both the advantage of interleaving over blocking with high-similarity categories and the reversal with low-similarity categories.

Rapid alternation of categories allows participants to identify differences between categories, which will be particularly beneficial if those differences are hard to detect, as in the case of the stimuli used in Experiment 1 and the artists’ styles or bird species used in previous studies. Infrequent alternation of categories, on the other hand, will allow participants to identify the commonalities within each category, which is particularly beneficial if categories are composed by members with high variability, like the ones used in Experiment 2.

In appreciating the benefits of blocking it is important to keep in mind that not all concept learning takes place by identifying discriminating features among categories. For example, sometimes it is possible to create an absolute characterization of a category in terms of its prevalent features, regardless of their discriminative values (Markman & Ross, 2003). Furthermore, in other situations, memorizing instances might be a highly useful strategy.

In sum, the results presented here show that there may not be a single answer to the question of how should an instructor sequence information so that the learner acquires the knowledge and is able to generalize it more efficiently. Best sequencing practices will depend on the nature of the categories being sequenced.

## Acknowledgments

This research was supported in part by National Science Foundation REESE grant 0910218 and Department of

Education IES grant R305A1100060. PFC was also supported by a Fulbright Research Fellowship and Graduate Fellowship SFRH/BD/68554/2010 from the Portuguese Foundation for Science and Technology (FCT). The authors would like to thank Rachel Selonick for her help with data collection.

## References

- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press.
- Carvalho, P. F., & Goldstone, R. L. (2011). Sequential similarity and comparison effects in category learning. In C. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Delaney, P. F., Verkoeijen, P. P. J. L., & Spiguel, A. (2010). Spacing and Testing Effects : A Deeply Critical, Lengthy, and At Times Discursive Review of the Literature. *Psychology of Learning and Motivation*, 53(10), 63–147.
- Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology*. City. New York, NY: Teachers College, Columbia University.
- Elio, R., & Anderson, J. R. (1981). The effects of category generalizations and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory*, 7(6), 397-417.
- Elio, R., & Anderson, J. R. (1984). The effects of information order and learning mode on schema abstraction. *Memory & cognition*, 12(1), 20-30.
- Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & cognition*, 24(5), 608-28.
- Goldstone, R. L. (2003). Learning to perceive while perceiving to learn. In R. Kimchi, M. Behrmann, & C. Olson (Eds.), *Perceptual organization in vision: Behavioral and neural perspectives*. New Jersey: Lawrence Erlbaum.
- Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, 130(1), 116-139.
- Kang, S. H. K., & Pashler, H. (2012). Learning Painting Styles: Spacing is Advantageous when it Promotes Discriminative Contrast. *Applied Cognitive Psychology*, 26, 97-103.
- Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, 23(9), 1297-1317.
- Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: is spacing the “enemy of induction”? *Psychological science*, 19(6), 585-92.
- Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and aging*, 25(2), 498-503.
- Kurtz, K. H., & Hovland, C. I. (1956). Concept learning with differing sequences of instances. *Journal of experimental psychology*, 51(4), 239.
- Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, 129(4), 592-613.
- Mathy, F. & Feldman, J. (2009). A rule-based presentation order facilitates category learning. *Psychonomic Bulletin & Review*, 16, 1050-1057.
- Mitchell, C. J., Kadib, R., Nash, S., Lavis, Y., & Hall, G. (2008). Analysis of the role of associative inhibition in perceptual learning by means of the same-different task. *Journal of experimental psychology. Animal behavior processes*, 34(4), 475-85.
- Mundy, M. E., Honey, R. C., & Dwyer, D. M. (2007). Simultaneous presentation of similar stimuli produces perceptual learning in human picture processing. *Journal of experimental psychology. Animal behavior processes*, 33(2), 124-38.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of experimental psychology. General*, 115(1), 39-61.
- Proctor, R. W. (1980). The Influence of Intervening Tasks on the Spacing Effect for Frequency Judgments. *Journal of Experimental Psychology: Human Learning and Memory*, 6(3), 254-266.
- Rohrer, D., & Pashler, H. (2010). Recent Research on Human Learning Challenges Conventional Instructional Strategies. *Educational Researcher*, 39(5), 406-412.
- Stewart, N., & Brown, G. D. A. (2004). Sequence effects in the categorization of tones varying in frequency. *Journal of experimental psychology. Learning, memory, and cognition*, 30(2), 416-30.
- Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(1), 3-11.
- Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, 24(6), 837-848.
- Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children’s memory and category induction. *Cognition*, 109(1), 163-7.
- Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: an investigation of mechanisms, metacognition, and aging. *Memory & cognition*.

# Implicit Learning of L2 Word Stress Rules

Ricky Chan ([rickychan0809@gmail.com](mailto:rickychan0809@gmail.com))

School of English, University of Hong Kong, Pokfulam Road, Hong Kong

Janny Leung ([hiuchi@hku.hk](mailto:hiuchi@hku.hk))

School of English, University of Hong Kong, Pokfulam Road, Hong Kong

## Abstract

This paper reports two experiments on the implicit learning of second language word stress rules and presents a methodological innovation. In both experiments L1 Cantonese L2 English participants practised pronouncing two-syllable Spanish words. Learning of a hidden stress regularity was measured by a judgment task. We assessed participants' awareness of the stress rule by verbal reports, confidence rating and a novel methodology: inclusion-exclusion production tasks adapted from Destrebecqz and Cleeremans (2001). Experiment 1 demonstrated the implicit learning of association between the ending phoneme and word stress and experiment 2 the implicit learning of a more abstract rule of stress placement. We conclude that L2 word stress rules may be learnt implicitly.

**Keywords:** implicit learning, word stress, Spanish, binary confidence ratings, process dissociation procedure

## Introduction

Previous studies of implicit learning have demonstrated that small-scale artificial grammar (Reber, 1967) and sequence of visual stimuli presentation in Serial Reaction Time (SRT) task may be learnt without awareness. In the domain of language, while early first language acquisition is essentially implicit, it remains cloudy whether implicit learning is relevant to second language acquisition (SLA) at all. The theoretical and pedagogical significance of implicit learning in SLA has stimulated recent research interests (e.g. Leung & Williams, 2011; Robinson, 1996, 2005; Saffran, 2001; Saffran et al., 2008; Rebuschat & Williams, 2011).

In the realm of L2 phonology, it has been shown that phonotactics (Dell, Reed, Adams, & Meyer, 2000; Onishi, Chambers, & Fisher, 2002; Warker & Dell, 2006), segmental features (Goldrick, 2004) and metrical stress rules (Gerken, 2004; Gerken & Boltt, 2008) may be learnt implicitly. Yet, to our knowledge, it is not known whether word stress regularities can be learnt without awareness, despite their importance in the parsing of speech stream. In the context of SLA, previous studies on the learning of lexical stress revealed that learners from a non-stress language background may have a different stress system than native speakers do, leading to a non-native "accent" of stress. For example, Mandarin speakers associated high flat tone with stress and Cantonese learners pronounce English stress and unstressed syllables as though they were high and low tones in their mother tongue (Chao, 1980). In light of the difficulty of learning stress, L2 stress pedagogy begs for further research. A crucial theoretical issue in implicit learning research is whether the knowledge acquired is abstract rules or

memorised chunks. Knowlton and Squire (1994, 1996) have proposed that both abstract information about grammar patterns and concrete information about the learning materials may be involved in implicit learning and they both contribute to performance in grammaticality judgment tasks. Whether abstract rules or chunks are learnt is of great relevance to the learning of prosody in that it would shed lights on how knowledge of prosodic signals is represented.

Relevant studies have been met with methodological challenges. Critics have challenged the validity of assessing awareness by verbal reports. The inability to verbalise knowledge might reflect low confidence on the part of the participants, lack of appropriate vocabulary to describe their knowledge or the intrinsic complexities of the regularity, and the knowledge assessed by verbal reports might not be responsible for performance on the measure of learning (Dienes & Berry, 1997; Shanks & St. John, 1994). Dienes et al. (1995) have sought to improve awareness measurement by asking participants to rate their confidence in judgment; a lack of correlation between judgment performance and confidence rating suggests that knowledge is unconscious (zero-correlation criterion). Tunney and Shanks (2003) found that binary confidence ratings are more sensitive to low levels of awareness than continuous ratings and were therefore adopted in our experiment.

More recently, Destrebecqz & Cleeremans (2001) have developed an objective awareness measure called the "method of opposition", modelled on the process dissociation procedure, a methodology framework first proposed by Jacoby (1991) to differentiate between the influences of implicit and explicit knowledge on performance. The method of opposition was applied to the SRT task (e.g. Curran, 2001; Haider, Eichler, & Lange, 2010). After completing the SRT task, participants were told there was a hidden sequence governing the presentation of visual stimuli and they were asked to complete free-generation tasks under both inclusion and exclusion instructions. Participants were asked to press response keys in an order that followed the sequence (inclusion condition) and that did not resemble the sequence (exclusion condition). According to the Global Workspace theory (Baars, 2003), when knowledge becomes conscious, the possibility for voluntary control of performance is opened up. A difference between inclusion and the exclusion performance indicates top-down processing and thus explicit knowledge. On the contrary, people with no explicit knowledge would tend to perform equally well in both inclusion and exclusion tasks (inclusion = exclusion) as they do not have control over how the implicitly learnt knowledge

influences behaviour (Curran, 2001). The experiment below demonstrates how the method of opposition could be employed as an awareness measure in other implicit learning tasks.

## Experiment 1

### Method

**Materials** The learning target was a simplified Spanish stress rule: words that end in “o” have their stress on the penultimate syllable (e.g. *busco* and *burro*) and words that end in “ar” on the last syllable (e.g. *gustar* and *tumbar*). Two-syllable real Spanish words were employed as stimuli and they were generated, using a Spanish male diphone database<sup>1</sup>, by the MBROLI speech synthesizer (Dutoit, Pagel, Pierret, Bataille, & Vrecken, 1996). The specific values were based on those used by Enríquez, Casado and Santos (1989): 100Hz and 60 ms for unstressed vowels and 116 Hz and 120ms for stressed vowels. Albeit the artificiality of the synthesizer, it was still preferred to voice recordings because a Spanish native speaker is likely to be biased against incorrectly stressed words, which raises a possibility that participants rely on the speaker’s fluency rather than knowledge of stress placement in the judgement task.

**Procedure** 37 university students (14 males and 16 females, *Age* = 21.8 years old) were recruited as the experimental group and 15 university students aged 20 to 26 (7 males and 8 females, *Age* 21.4) as the control group. All of them were native Cantonese speakers with English as an L2. None of them reported any knowledge of Spanish or Portuguese. They were told that the experiment aimed to study how people learn words. The experiment took around 25 minutes to complete.

**Training phase:** The training phase consists of 64 randomised trials, each containing a Spanish word and its English translation (See Fig 1, left). A set of 16 Spanish words, half of which end with *-ar* and the other half *-o*, was repeated four times. Participants repeated aloud after the recording. This provided participants with exposure to the target stress rules without explicitly directing their attention to them.

**Testing Phase (pronunciation judgement):** The testing phase consists of 40 trials, each of which includes an English verb and two sound icons (See Fig 1, right). Participants pressed the corresponding keys to listen to two audio presentations and choose the one that “sounds better” to them. As Scott and Dienes (2008) have shown that familiarity is the essential source of knowledge in artificial grammar learning, preference judgment is used rather than accuracy judgment (e.g., “choose the correct one”) because the former may encourage the use of intuition and discourage rule search during the testing phase (Rebuschat & Williams, 2011).

speak

hablar

like

Word 1

Word 2

**Figure 1: A sample trial used in learning (left) and testing phase (right) in experiment 1**

All the 16 critical items (8 *-ar* ending and 8 *-o* ending) were novel to the participants. Those sound pairs differed only in stress placement. Other previously seen items were randomly shuffled with the critical items so as to reduce the likelihood of participants consciously deducing the difference between the sound pairs.

To test whether, for *-ar* ending words, participants’ judgements were based on the *-ar* ending, *-r* ending or the vowel “a” at the end, 8 *-a* ending words were added in the testing phase. Like the *-o* ending words, words that end in “a” have their stress on the penultimate syllable and those sound pairs differed only in stress placement.

**Inclusion-exclusion Tasks.** Adapted from Destrebecqz and Cleeremans (2001), the tasks required participants to read aloud 8 two-syllable words in each of the two conditions: 1) “as similarly to Spanish pronunciation as possible” (inclusion) and 2) “as differently from Spanish pronunciation as possible” (exclusion). A small dot was given in each word (see Fig 2) to indicate syllabification which serves to remind the participants that all words consist of two syllables. Their voices were recorded.

Tum.bar

**Figure 2: A sample trial in the inclusion-exclusion tasks**

**Verbal reports.** Participants were probed for whether they had any knowledge about the pronunciation patterns and stress rules of Spanish words. They were also told there were underlying stress patterns and if they could not report knowledge of the regularities, they would be asked to provide as many guesses as possible.

All the items used are presented in table 1 below.

<sup>1</sup> <http://tcts.fpms.ac.be/synthesis/mbrola.html>

A: Items used in training (16 items)			
Words ended in "ar"		Words ended in "o"	
Hablar	Barrar	Soplo	Borro
Rascar	Roncar	Marco	Busco
Tumbar	Lanzar	Zumbo	Forzo
Contar	Gustar	Tosto	Gasto
B: Critical items used in the testing phase (16 items)			
Words ended in vowels		Words ended in consonants	
Probar	Tratar	Broto	Grabo
Juzgar	Firmar	Colgo	Formo
Saltar	Cantar	Bastar	Pinto
Montar	Costar	Falto	Junto
C: -a ending words used in the testing phase (8 items)			
Roba	Bota	Cita	Dota
Mata	Vota	Paga	Toma
D: Items used in inclusion-exclusion tasks (16 items)			
Inclusion		Exclusion	
Ha.llo	Lle.nar	Lla.mar	Ce.rró
Lle.vo	Cho.car	Ten.tar	Llo.ro
Tum.bar	Bre.go	Chu.par	Pen.so
Man.do	Tar.dar	Car.go	Cor.tar

**Table 1: Items used in training, testing, and inclusion-exclusion tasks in experiment 1**

## Results and Discussion

**Overall performance** Participants' performance on the pronunciation judgment task served as the measure of learning. The experimental group attained an average of 59.6 % accuracy ( $SD = 1.41$ ;  $SEM = .23$ ) on the 16 critical items. Further analysis using *t*-test showed that their performance was significantly above chance,  $t(36) = 6.57$ ,  $p < .001$ ,  $d = 1.55$ . They achieved above-chance-level performances on both -ar ending and -o ending words,  $M = 64.2\%$ ,  $SD = 1.42$ ,  $SEM = .24$ ,  $t(36) = 4.81$ ,  $p < .001$ ,  $d = 1.13$  and  $M = 55.1\%$ ,  $SD = 1.13$ ,  $SEM = .19$ ,  $t(36) = 2.16$ ,  $p = .019$ ,  $d = .51$  respectively. Their performance on -ar ending words was significantly better than that of -o ending words,  $t(69) = 2.41$ ,  $p < 0.01$ ,  $d = .57$ , suggesting more learning of -ar ending words than -o ending words in relation to stress placement.

**Awareness measures** Verbal reports and inclusion-exclusion tasks were used to determine whether the acquired knowledge was conscious or not.

**Verbal reports:** 32 out of 37 participants remained unaware of the underlying stress rules based on verbal reports. One participant was able to verbalize the whole target stress rules.

Four participants reported some knowledge that overlaps with our target rules (i.e. "r" is stressed; intonation falls when a words ends in -o; stress is related to the "ar" ending; stress is related to word length). They were classified as aware. In fact many participants did not notice any stress patterns and were surprised when they were asked to guess.

**Inclusion-exclusion tasks:** Participants' audio recordings for the tasks were analysed using *Praat* to locate their stress placement based on the fundamental frequencies (F0) of the two syllables, as it was found that Chinese speakers rely on F0 as the most important cue in stress judgement (Wang, 2008). The difference between inclusion and exclusion scores was calculated. The five aware participants scored higher under the inclusion instruction than the exclusion instruction (from +2 to +3), showing congruence with their awareness level revealed in verbal reports. 15 out of the 32 remaining participants scored equally for both tasks and they were classified as our truly unaware participants. The other 17 participants who showed some difference in their scores for both tasks (from -3 to +3) were re-classified as aware and were not included in our analysis of unaware data.

**Pronunciation Judgment Task:** The 15 unaware participants achieved an average of 58.8% accuracy ( $M = 9.4$ ;  $SD = 1.20$ ;  $SEM = .32$ ) on the 16 critical items and their performance was significantly above chance,  $t(14) = 4.37$ ,  $p < .01$ ,  $d = 1.65$ . Their performances on both -ar ending and -o ending words are both significantly above chance,  $M = 59.1\%$ ,  $SD = 1.12$ ,  $SEM = .30$ ,  $t(14) = 2.44$ ,  $p = .014$ ,  $d = .92$  and  $M = 58.3\%$ ,  $SD = 1.07$ ,  $SEM = .29$ ,  $t(14) = 2.32$ ,  $p = .018$ ,  $d = .88$  respectively. There is evidence of implicit learning of L2 ending-phoneme-to-stress rules by young Cantonese-speaking adults with only short and limited exposure. However, their accuracy on -ar ending words was not significantly higher than that of -o ending words,  $t(28) = 0.16$ ,  $p = 0.44$ .

**Control:** The control group completed only the pronunciation judgement task. Their accuracy on critical items was not significantly different from chance,  $M = 47.1\%$ ,  $SD = 1.09$ ,  $SEM = .29$ ,  $t(14) = 1.61$ ,  $p = .065$ .

While previous studies suggest that novel L2 stress perception may pose challenges to learners from a non-stress language background, participants' proficiency in English and their prior general knowledge of word stress might have been helpful in this experiment. Moreover, it was found that participants in general performed significantly better for -ar ending words than -o ending words. This may be explained by their prior linguistic knowledge. A previous study (Bailey, Plunkett, & Scarpe, 1999) showed that native English speakers, when learning a novel stress pattern, had a significant bias for non-word-final stress. In our study, when participants were asked if they had any feelings about pronunciation patterns of Spanish words, 12 participants mentioned "the intonation of the words tended to go up", "the last syllable seemed to be louder and higher in pitch" and "stress tends to lie on the final syllable", despite the fact that words with a word-final stress and a non-word-final stress appeared equally frequently in the experiment. These



statements are an indication that stress patterns that are distinct from English (word-final stress in this case) might have appeared more salient to our participants given their prior linguistic experience. In addition, heavy syllables tend to be stressed in English. This may explain why there appeared to be more learning of *-ar* ending words (which has a heavy syllable at the end) than that of *-o* ending words.

It is worth mentioning that there is a medium negative correlation between participants' accuracy on *-ar* ending words and that of *-a* ending words,  $r(35) = -0.51$ ,  $p = .001$  for all participants and  $r(13) = -0.72$ ,  $p = .002$  for unaware participants. On the other hand, a medium positive correlation was found between their accuracy on *-o* ending words and that of *-a* ending words for both the whole experimental group,  $r(35) = 0.49$ ,  $p = .002$ , and the unaware participants,  $r(13) = 0.64$ ,  $p = .001$ . These suggest that participants might have made their judgements based on the last phoneme (i.e. the *-r* ending instead of the vowel *a*) in the last syllable. While it is still unclear whether participants learnt an abstract rule or a set of probabilities, their higher sensitivity to the difference between open and closed syllables provided initial evidence of rule learning. This issue was addressed in a follow-up experiment, whose design was similar to experiment 1.

## Experiment 2

### Method

**Materials** The learning target was more complicated stress regularities: words that end in a consonant (closed syllable) have their stress on the final syllable (e.g. *felol* and *cerroz*) and words that end in a vowel on the last syllable (e.g. *pato* and *bona*). Two-syllable nonwords with four or five phonemes (vowel-ending words and consonant-ending words respectively) were generated by the same speech synthesizer based on Spanish pronunciation. The words were combination of phonemes illustrated below:

1. First phoneme: /b/, /b/, /d/, /d/, /f/, /g/, /h/, /k/, /k/, / / (l), /m/, /n/, /p/, /s/, /s/, /t/, /t/ or /v/.
2. Second phoneme: /a/, /e/ or /o/.
3. Third phoneme: /b/, /b/, /d/, /d/, /g/, /k/, /k/, /l/, /m/, /n/, /n/, /p/, / /, /r/ (trilled), /s/, /t/, /t/ or /v/.
4. Fourth phoneme: /a/, /e/ or /o/
5. Fifth phoneme (consonant-ending words only): / /, /l/, /θ/ (z).

**Procedure** 44 participants were recruited. The experimental group consisted of 22 young adults (9 male and 13 female, *Age* = 21.0 years old). The other 22 young adults (12 male and 10 female, *Age* = 22.3 years old) were the control group. They were all native Cantonese speakers with English as an L2 with no knowledge of Spanish or Portuguese. The experiment took around 30 minutes to complete.

**Training phase:** Participants were presented with a Spanish word (See Fig 2, left). They repeated aloud after the recording. 36 Spanish-based words, half of which end in a

vowel and the other half consonant, were repeated three times to form 108 trials.

**Testing Phase (pronunciation judgement):** Participants clicked to listen to two possible Spanish pronunciations (See Figure 2, right) and chose the one that “sounds better” to them.

domal | Word 1 Word 2

**Figure 2: A sample trial used in learning (left) and testing phase (right) in experiment 2**

**Confidence rating:** A binary confidence rating was included to further improve awareness assessment. After making a judgement, participants were asked to indicate whether they made a guess or were certain about their choice.

Critical data came from 18 novel words, half of which ends in /a/, /e/ or /o/ and the other half /ɔ/, /l/ and /z/. To determine whether participants had learnt an abstract rule or memorised chunks, 12 “extension items”, half of which ends in /i/ or /u/ (vowel ending) and the other half /d/, /x/ (consonant ending) were included. Those sound pairs differed only in placement of stress.

**Inclusion-exclusion tasks:** Same as experiment 1, except that there were 18 items for each condition.

**Verbal reports:** Same as experiment 1.

A: Items used in training (36 items)					
-a ending	-o ending	-e ending	-r ending	-l ending	-z ending
Beba	Gobo	Coge	Botor	Bogel	Cerroz
Bona	Navo	Dade	Coder	Domal	Gapez
Cepa	Pato	Fane	Llaner	Debal	Hocaz
Doca	Seco	Mete	Penar	Felol	Natoz
Hara	Sorro	Tome	Socor	Madel	Tobaz
Llada	Telo	Vese	Tevar	Sasol	Verez
B: Critical items used in the testing phase (18 items)					
-a ending	-o ending	-e ending	-r ending	-l ending	-z ending
Dada	Goto	Cebe	Decar	Mebel	Cepez
Moda	Llemo	Sage	Llager	Savol	Gadoz
Teca	Savo	Tope	Tomor	Sotal	Today
C: Extension items used in the testing phase (12 items)					
-i ending		-u ending	-d ending		-j ending
Llepi		Dotu	Saded		Gotej
Gomi		Sacu	Cemod		Llecaj
Cabi		Tedu	Tobad		Dapoj
D: Items used in inclusion-exclusion tasks (18 items)					
-a ending	-o ending	-e ending	-r ending	-l ending	-z ending
Ho.na	So.to	Ta.re	To.bar	Vo.sal	De.rraz
Ce.ba	Galo	Lle.de	Ca.mer	Ba.pel	Fo.gez
No.ca	Ba.vo	Me.te	Se.nor	Pe.col	Da.doz
Ga.ba	Me.no	Va.de	Bo.ver	Ne.bol	Da.coz
Tela	Co.po	Se.ge	Ce.ror	Fo.tal	Te.maz
Do.sa	Be.to	So.ne	Lle.car	Par.rel	Ho.dez

**Table 2: Items used in training, testing, and inclusion-exclusion tasks in experiment 2**

## Results and Discussion

**Verbal reports:** Among the 22 participants, one participant was able to report our target rules. Two other participants mentioned “rising intonation for -z ending words” and “stress is affected by length of the word”, which overlap with our target rules. These three participants were classified as aware. Still, many participants reported they paid no attention to any stress patterns and were surprised when they were asked to guess.

**Confidence rating:** The Chan difference score (Chan, 1992; Dienes et al, 1995) was computed for each participant to determine whether learning was implicit by the zero correlation criterion. It is the difference between the proportion of ‘certain’ responses which were correct (hit) and those which were incorrect (false alarm). Participants with a positive score were classified as aware and those score 0 or below were classified as unaware.

**Inclusion-exclusion tasks:** Participants who scored the same in inclusion and exclusion tasks were classified as unaware; otherwise they were classified as aware.

**Pronunciation Judgment Task:** Only 8 out of 22 participants were classified as unaware based on the above criteria. Their average accuracy in the judgement task was 68.1% ( $M = 12.25$ ;  $SD = 1.67$ ;  $SEM = .59$ ) on the 18 critical items, which was significantly above chance,  $t(7) = 5.51$ ,  $p < .001$ ,  $d = 1.95$ . They achieved better-than-chance accuracy for both vowel and consonant ending words,  $t(7) = 2.12$ ,  $p = .03$ ,  $d = 0.75$  and  $t(7) = 3.86$ ,  $p = .003$ ,  $d = 1.37$  respectively, but accuracies on vowel and consonant ending words are not significantly different,  $t(14) = .30$ ,  $p = .39$ . These serve as evidence of learning the target regularities.

An analysis of their accuracy on extension items showed that their accuracy was significantly higher than chance,  $t(7) = 4.73$ ,  $p = 0.001$ ,  $d = 1.67$ . They achieved better-than-chance accuracy for both vowel and consonant ending words,  $t(7) = 5.00$ ,  $p < .0001$ ,  $d = 1.77$  and  $t(7) = 2.39$ ,  $p = .023$ ,  $d = 0.85$  respectively, with no significant difference on accuracies for vowel and consonant ending words,  $t(14) = 1.25$ ,  $p = .15$ . Such findings revealed that the resultant knowledge was abstract rules rather than memorized chunks. Participants were sensitive to the difference between the ending phoneme (vowel or consonant) and syllable type (open and closed syllables) in relation to stress placement and were able to apply their abstract knowledge to novel items.

**Control:** The control group completed only the testing phase. Their accuracy on critical items was not significantly higher than chance,  $M = 52.0\%$ ,  $SD = 1.24$ ,  $SEM = .27$   $t(21) = .74$ ,  $p = 2.33$ .

## Conclusion

The two experiments above suggest that, under incidental learning conditions, young adults were able to learn L2 word stress rules, which are supra-segmental phonological patterns, with only short and limited exposure. The knowledge obtained was implicit, abstract, and may be applied to novel items. Their lack of awareness was verified by verbal reports, binary confidence ratings and process dissociation procedure

in the experiments. We conclude that young adults are able to acquire L2 word stress implicitly.

The present study bears theoretical, methodological and pedagogical significance. On the theoretical level, it extends previous findings on the implicit learning of language: not only may syllable regularity (at the segmental level) and metrical stress rules be learnt implicitly (as shown in previous studies), but sensitivity to lexical stress rules (at the supra-segmental level) may also develop without awareness. This raises the possibility of implicit learning of other kinds of prosodic rules such as tonal rules. Importantly, the rules and stimuli in the experiment were all based on a natural language and so there is little question of transferability of findings to the context of language acquisition.

Our study is also methodologically interesting as it is the first implicit learning study which integrates verbal reports, subjective measures of confidence and process dissociation procedure to establish strict criteria for awareness measurement. It is also, to our knowledge, one of the first successful attempts to apply the process dissociation procedure outside of SRT tasks. Based on our data, the verbal reports were useful in identifying three aware participants who were able to verbalize partial knowledge of the target rules. However, other participants whose awareness was not reflected in the verbal reports nevertheless performed differently in the inclusion-exclusion tasks, which appeared to be more sensitive in identifying participants with low confidence or low awareness about the rules. The findings demonstrated the usefulness of inclusion-exclusion tasks as an objective measure of participants’ awareness, and their potentially higher sensitivity than verbal reports.

From the perspective of L2 pedagogy, the present study provides an insight into how L2 stress patterns may be taught and learnt. The possibility of learning L2 stress patterns implicitly offers an alternative to the widely-adopted explicit approach to teaching word stress. It would be theoretically and practically interesting to determine the relative effectiveness of explicit and implicit teaching and learning of word stress and explore their potential synergetic effects in different settings.

Other further research directions include exploring 1) the implicit learning of other supra-segmental phonological rules such as tonal rules; 2) whether the ability to implicitly learn L2 phonological rules is sensitive to individual differences; and 3) whether age has a significant impact on the implicit learning of L2 phonological patterns.

## References

- Baars, B. J. (2003). How brain reveals mind: Neuroimaging supports the central role of conscious experience. *Journal of Consciousness Studies*, 10(9-10), 100-114.
- Bailey, T., Plunkett, K., & Scarpe, E. (1999). A cross-linguistic study in learning prosodic rhythms: Rules, constraints, and similarity. *Language and Speech*, 42(1), 1-38.

- Chan (1992). *Implicit cognitive process: Theoretical issues and applications in computer systems design*. Unpublished DPhil thesis, University of Oxford.
- Chao, Y.R. (1980). Chinese tone and English stress. In Waugh, L. R. & VanSchooneveld, C. H. (eds.) *The Melody of Language*, Baltimore, MD: University Park Press, 41-44
- Curran, T. (2001). Implicit learning revealed by the method of opposition. *Trends in Cognitive Sciences*, 5, 503-504.
- Dell, G. S., Reed, K. D., Adams, D. R., & Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning Memory, and Cognition*, 26(6), 1355-1367.
- Destrebecqz, A., & Cleeremans, A. (2001). Can sequence learning be implicit? New evidence with the process dissociation procedure. *Psychonomic Bulletin & Review*, 8(2), 343-350.
- Dienes, Z., Altmann, G.T.M., Kwan, L. & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21, 1322-1338.
- Dienes, Z., & Berry, D. C. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin and Review*, 4, 3-23.
- Dutoit, T., Pagel, N., Pierret, N., Bataille, O., & Vrecken, V. d. (1996). The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. *Proceedings of ICSLP'96, Philadelphia*, 3(1393-1396).
- Enríquez, E. V., Casado, C., & Santos, A. (1989). La percepción del acento en español. *Lingüística Española Actual*, 11, 241-269.
- Gerken, L. A. (2004). Nine-month-olds extract structural principles required for natural language. *Cognition*, 93, B89-B96.
- Gerken, L. A., & Boltt, A. (2008). Three examples allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Language Learning and Development*, 4, 228-248.
- Goldrick, M. (2004). Phonological features and phonotactic constraints in speech production. *Journal of Memory & Language*, 51, 586-603.
- Haider, H., Eichler, A., & Lange, T. (2010). An old problem: How can we distinguish between conscious and unconscious knowledge acquired in an implicit learning task? *Consciousness and Cognition*.
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language*, 30, 513-541.
- Juffs, A. (1990). Tone, syllable structure and interlanguage phonology: Chinese learners' stress errors. *International Review of Applied Linguistics*, 28(2), 99-118.
- Knowlton, B. J., Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 20, 79-91.
- Knowlton, B. J., Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22, 169-181.
- Leung, J., & Williams, J. N. (2011). The Implicit Learning of Mappings between Forms and Contextually-Derived Meanings. *Studies in Second Language Acquisition*, 33.
- Onishi, K., Chambers, K., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition*, 83, B13-B23.
- N.Warner (eds.). Berlin: Mouton de Gruyter, 203-240.
- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behaviour*, 6, 855-863.
- Rebuschat, P., & Williams, J. N. (2011). Implicit and explicit knowledge in second language acquisition. *Applied Psycholinguistics*, Available on CJO 2011 doi:10.1017/S0142716411000580
- Robinson, P. (1996). Learning simple and complex second language rules under implicit, incidental, rule-search, and instructed conditions. *Studies in Second Language Acquisition*, 18, 27-67.
- Robinson, P. (2005). Cognitive abilities, chunk-strength, and frequency effects in implicit artificial grammar and incidental L2 learning: Replications of Reber, Walkenfeld, & Hernstadt (1991) and Knowlton & Squire (1996) and their relevance for SLA. *Studies in Second Language Acquisition*, 27, 235-268.
- Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, 44(493-515).
- Saffran, J. R., Hauser, M., Seibel, R., Kapfhamer, J., Tsao, F., & Cushman, F. (2008). Grammatical pattern learning by human infants and cotton-top tamarin monkeys. *Cognition*, 107, 479-500.
- Scott, R., & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology: Learning Memory, and Cognition*, 34, 1264-1288.
- Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning system. *Behavioral and Brain Sciences*, 17, 367-447.
- Tunney, R. J. & Shanks, D. R. (2003). Does opposition logic provide evidence for conscious and unconscious processes in artificial grammar learning? *Consciousness and Cognition*, 12, 201-218.
- Wang, Q. (2008). L2 stress perception: The reliance on different acoustic cues. *Speech Prosody 2008*. Campinas, Brazil, 135-138
- Warker, J. A., & Dell, G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 387-398.



# Generating Realistic Semantic Codes for Use in Neural Network Models

**Ya-Ning Chang (ya-ning.chang@postgrad.manchester.ac.uk)**

Neuroscience and Aphasia Research Unit (NARU), University of Manchester,  
Brunswick Street, Manchester, M13 9PL, UK

**Steve Furber (steve.furber@manchester.ac.uk)**

Advanced Processor Technologies Group, University of Manchester, Oxford Road, Manchester, M13 9PL, UK

**Stephen Welbourne (stephen.welbourne@manchester.ac.uk)**

Neuroscience and Aphasia Research Unit (NARU), University of Manchester,  
Brunswick Street, Manchester, M13 9PL, UK

## Abstract

Many psychologically interesting tasks (e.g., reading, lexical decision, semantic categorisation and synonym judgement) require the manipulation of semantic representations. To produce a good computational model of these tasks, it is important to represent semantic information in a realistic manner. This paper aimed to find a method for generating artificial semantic codes, which would be suitable for modelling semantic knowledge. The desired computational criteria for semantic representations included: (1) binary coding; (2) sparse coding; (3) fixed number of active units in a semantic vector; (4) scalable semantic vectors and (5) preservation of realistic internal semantic structure. Several existing methods for generating semantic representations were evaluated against the criteria. The correlated occurrence analogue to the lexical semantics (COALS) system (Rohde, Gonnerman & Plaut, 2006) was selected as the most suitable candidate because it satisfied most of the desired criteria. Semantic vectors generated from the COALS system were converted into binary representations and assessed on their ability to reproduce human semantic category judgements using stimuli from a previous study (Garrard, Lambon Ralph, Hodges & Patterson, 2001). Intriguingly the best performing sets of semantic vectors included 5 positive features and 15 negative features. Positive features are elements that encode the likely presence of a particular attribute whereas negative features encode its absence. These results suggest that including both positive and negative attributes generates a better category structure than the more traditional method of selecting only positive attributes.

**Keywords:** semantics; semantic representations; neural networks; computational modelling; connectionist models.

## Introduction

Computational models are frequently used to simulate human behavioural data and help understand the underlying cognitive processes. Any type of computational model requires decisions to be made about what representation scheme to use. Semantic representations are particularly important for models of many linguistic processes including spoken and written language. This paper aims to find a method of generating semantic representations, which can fulfil a set of requirements derived from the constraints imposed by incorporating semantic knowledge within a large-scale connectionist model. A list of criteria that we

considered essential for sophisticated and efficient simulation using a connectionist model includes: (1) Binary coding: a binary coding scheme is essential for use in connectionist models because the models consist of many neuron-like units whose activation values vary between 0 and 1. The models are trained to match their activation values to predefined targets, which need to be at the extreme ends of the possible activations; (2) Sparse coding: a sparse coding scheme is one in which an item is represented by using a small number of active units in each vector. A sparse representation is attractive from a computational viewpoint because it allows efficient computation. By controlling sparseness, the redundancy of a code can be minimized and learning is generally fast and relatively easy. Importantly, it is likely to reflect the natural structure of the representation system in the brain; (3) Fixed number of active units in each vector: the idea of using a fixed number of active units in each semantic vector is not common in most existing coding schemes. However, it has an advantage that this coding is uniform and it makes sense to think about how similar items are by measuring the Euclidian distance between them – if items have different numbers of features then measuring Euclidian distance does not give a good indication of similarity (Furber, Bainbridge, Cumpstey & Temple, 2004). For connectionist models, there is a particular reason to want to adopt a fixed number of active units, which is that only the active units can contribute to activation in later layers. Units with a zero level of activation do not propagate information in the network and therefore do not generate any weight updates in response to the error signal; (4) Scalable semantic vectors: to keep the simulations computationally tractable, it is important to keep the size of semantic vectors manageable. Vector size is an important design consideration because it determines how many units in the model are needed for modelling semantic knowledge. Given a code length  $n$ , it has a maximum theoretical number of items that it can code for, which is  $2^n$ . The capacity increases dramatically as the code length grows. Thus, the selection of the vector length also needs to consider the number of items to be represented; (5) Preservation of inherent semantic structure: the most important criterion is that the semantic vectors can support human-like semantic classifications. They need to preserve

the inherent semantic structure of the lexicon. Words which are semantically similar should be represented by vectors that are relatively close in the semantic space; by contrast, semantically unrelated words should tend to be far from each other. Preserving these semantic relationships will allow for the possibility of modelling tasks like categorization and synonym judgment, which are commonly used to probe semantic effects.

### Review of existing semantic representation schemes

Several semantic representation schemes have been proposed either for behavioural studies or for use in computational modelling (Dilkina, McClelland & Plaut, 2008; Harm & Seidenberg, 2004; Plaut, 1997; Rogers, Lambon Ralph, Garrard, Bozeat, McClelland, Hodges & Patterson, 2004). These schemes are based on different techniques and there does not seem to be a consensus view as to how to produce a set of representations. It is therefore important to review these competing coding schemes from a modelling perspective using the criteria described above.

**Feature Norms** One traditional method is to obtain the feature norms through experiments (e.g., McRae, Cree, Seidenberg & McNorgan, 2005). In these experiments, subjects are given a list of words and asked to write down attributes about each word. To categorise the attributes and make proper constraints on subjects' responses, some lexical relations such as "is" and "has" are used to prompt subjects to list the features of the stimulus word. The most commonly listed features for a particular word are then considered as the core semantic attributes for that word. The collected attributes for an item can be easily converted into binary codes with the presence of an attribute coded as "1" and the absence as "0". Controlling for sparseness is not so easy, but it may be possible to rank the features by the number of subjects that identified them, and use this as a method for deciding which features to drop. Moreover, this method is not very flexible and practically can only be used for a small set of words.

**Arbitrary Features** Another way to generate semantic representations is to use random features. Features for a word are assigned randomly but the assignments may still respect broader aspects of semantic knowledge such as category knowledge. For example, the words within the same category can be designed to share more features than words belonging to different categories. This method has been applied to various computational studies designed to capture abstract semantic properties including simulations of lexical decision (e.g., Plaut, 1997) and semantic impairment (e.g., Rogers et al., 2004). The features for an item are assigned manually and are binary codes. The control of sparseness can be achieved by adjusting the fraction of the number of active units in a vector over the code length. The fixed number of active units in a vector is also controllable. In addition, the size of vector length is scalable and determined by the modellers. Although this

coding scheme is good for producing coarsely structured semantic representations, it is not easily scalable and it would be very difficult to generate an artificial semantic structure that can capture the complexity found in human semantics.

**Co-occurrence Statistics** Semantic representations can also be derived from very large text corpora by evaluating which words appear in similar types of documents or co-occur within a fixed window. Several semantic representation schemes have been developed on the basis of this statistical co-occurrence including Latent Semantic Analysis (LSA) (e.g., Landauer, Foltz & Laham, 1998), Hyperspace Analogue to Language (HAL) (e.g., Lund & Burgess, 1996) and Correlated Occurrence Analogue to Lexical Semantics (COALS) (Rohde et al., 2006). These methods are all based on similar ideas but they are slightly different in the ways they collect data and deal with the high-dimensional co-occurrence matrices. LSA derives vectors based on a collection of segmented documents in which the number of occurrences of a word in various types of documents is computed as an element in the high-dimensional co-occurrence matrix. The dimensionality of the matrix is then reduced by using Singular Value Decomposition (SVD) while preserving the semantic relations between words as much as possible. Unlike LSA, the derivations in both HAL and COALS are based on words co-occurring within a fixed window in an un-segmented document. The key differences between these three systems are in their ways of expressing the tendency of two words to co-occur: LSA computes the cosines between the vectors of two words, HAL uses distance measure and COALS uses the correlation measure. In addition, HAL reduces the dimensionality of the matrices by eliminating all but the few thousand columns with the largest variant values, which is different from the SVD technique adopted by both LSA and COALS (see Rohde et al., 2006 for more detailed comparisons).

The semantic vectors generated by reducing a high-dimensional matrix are typically real-value vectors but COALS also provides binary-valued vectors. For the other two systems, however, it is relatively easy to convert the real values to binary values by thresholding. The vectors with values greater than a certain level are replaced with the value "1" and all others are replaced with the value "0". Sparse coding can be enforced by adjusting the threshold level used when converting real-value vectors into binary. Similarly, the fixed number of active units in a vector can be designed by modellers during the binarization process by restricting the number of 1's to the top *n* elements of the vector. By using the co-occurrence statistics, the sizes of semantic vectors for lexical items are scalable, which is particularly suitable for the computational modellers seeking a set of representations with low computational cost. The key advantage of this scheme is that it should be able to generate realistic semantic codes for any word lists of any length provided that the latent semantic information contained in the structure of large corpora is sufficiently

detailed and can be extracted efficiently.

**WordNet** WordNet is an online semantic database, which was developed by Miller in 1990. Information in WordNet is organised by many synonymous sets. These sets are linked by their lexical relations such as “is a” or “is part of” relations. A unique feature of WordNet is that it provides multiple word senses, which can be obtained from the database separately while other semantic systems do not distinguish between word senses. Similar to the feature norms, the attributes generated from WordNet have direct semantic interpretations. The semantic vectors generated by WordNet are binary to represent the presence and the absence of attributes, and generally rather sparse. But the number of semantic features for each word is not fixed and the range could be very wide. The size of the semantic vectors is less flexible because the size depends on how words relate to each other within the word list of interest. As a general rule the longer the list of lexical items to be coded the longer resultant semantic vector. Since the vectors are, directly derived from many synonymous sets in Word Net based on the researchers’ semantic knowledge, the semantic structure is likely to be well preserved.

Table 1 summarises the results of these evaluations. Among these, COALS appears to be the best choice because it satisfied most of the criteria than other systems.

Table 1: Summary of the evaluations of different semantic representation schemes

	Feature Norms	Arbitrary Features	Co-occurrence Statistics			Word Net
			LSA	HAL	COALS	
Criterion 1 (Binary coding)	✓	✓	Δ	Δ	✓	✓
Criterion 2 (Sparse Coding)	Δ	Δ	Δ	Δ	Δ	✓
Criterion 3 (Fixed Number of Active Units)	X	Δ	Δ	Δ	Δ	X
Criterion 4 (Scalability)	X	✓	✓	✓	✓	X
Criterion 5 (Semantic Structure)	✓	X	✓	✓	✓	✓

Note: ✓: good fit; Δ: can be adapted to fit; X: poor fit or difficult to support

## Method

The correlated occurrence analogue to the lexical semantics (COALS) system (Rohde et al., 2006) is designed to be very flexible. Although two of the criteria (i.e., sparse coding and fixed number of active units) are outside the scope of the original COALS system, they could be easily achieved by manipulating the semantic vectors generated from the system. However, it is crucial to examine whether the semantic codes generated from COALS preserve enough semantic structure that they can be used to predict the human semantic data. In addition it is important to investigate the best method of transforming the COALS vectors into binary codes. To generate binary vectors Rohde and colleagues simply set negative components to 0 and positive components to 1 based on the original real-valued

vector from the SVD. This means that information contained in negative parts of the vector is lost. Thus the questions asked here are whether negative components also contribute to a good semantic similarity structure and, if so, what is the optimum number of positive and negative features required to produce a best fit to human data. The following sections describe how to generate semantic vectors based on COALS in a way that satisfies all the criteria discussed previously. We then go on to compare the performance of the vectors on a semantic categorisation task using human data taken from Garrard et al.’s (2001) categorisation study.

## Generating Semantic Vectors based on COALS

To explore whether negative components were as important as positive components, a binarization process of coding both positive and negative components were used. A 100-dimensional semantic vector was generated for all items in the Garrard et al.’s (2001) word list except one item “watering can”, which was discarded because the system did not support the compound words. The 100-dimensional vector was duplicated to create two 100-dimensional vectors with the first 100 dimensions coding the positive elements and the second 100 dimensions coding the negative elements. The two 100-dimensional vectors were combined into a 200-dimensional semantic vector. The first half of the 200-dimensional vector contained only the positive components and the second half of the 200-dimensional vector contained only the absolute values of negative components. Assuming the best number of positive and negative components is  $n$  and  $m$  respectively then the top  $n$  values of the first half of the combined vector and the top  $m$  values of the second half of the vector were selected as the key features. All the selected features were set to 1 and others are 0. This resulted in a 200-dimensional binary vector with the property that matching elements from the two halves of this vector (e.g., the 1<sup>st</sup> and 101<sup>st</sup> elements) code for the same feature and have a special relationship whereby if one is on then the other must be off. (Note: both paired vectors can be off indicating that neither the presence nor the absence of this feature is particularly important to the meaning of the item.)

## Testing Procedure

To determine the usefulness of the binary semantic codes we examined the relationship between the category structures derived from the artificially generated sets of semantic vectors and human data taken from Garrard et al.’s (2001) study. In their study, Garrard and colleagues asked subjects to categorise items into a living thing group and nonliving thing group. On a finer scale, the living thing group can be divided into animals, birds and fruit and the nonliving thing group also can be subdivided into household objects, tools and vehicles. There were in total 6 subgroups. Semantic vectors for 61 items in the Garrard et al.’s (2001) list were generated using the method described above. Each vector had a length of 200. The  $n$  features of the first half of

the semantic vector represent the important positive features and the  $m$  features of the second half of the semantic vector indicate the important negative features. The numbers of positive features ( $n$ ) and negative features ( $m$ ) were varied to determine the optimum values of  $n$  and  $m$ .

### Evaluation of semantic vectors

Two parameters based on semantic distances between words can be used to evaluate the match with the semantic structure in the human data: distance validity index ( $DVI$ ) and distance ratio ( $DR$ ).  $DVI$  counts the number of groups where the within group distance (i.e., the averaged Euclidean distance between items in the same group) is smaller than all the between group distances (i.e., the averaged Euclidean distance between items in the different groups). The larger the value of  $DVI$  the better the semantic categories have been partitioned. This is rather coarse measure of semantic structure and for this data the value of  $DVI$  ranges from 1 to 6 (i.e., the number of subgroups). The expected best value is  $DVI=6$  indicating that all the within group distances are smaller than between group distances.  $DR$  computes the average of all the distance ratios. The distance ratio is the sum of between group distances to the sum of within group distances. Ideally there will be a larger between group distance and a smaller within group distance so that  $DR$  should be as large as possible. It should be noted that the value of  $DR$  is also positively correlated with the total number of features within a vector because it is computed on the basis of the Euclidean distance. The distance for the vectors having more features is generally larger than that for the vectors having less features. This indicates that  $DR$  is only a useful comparator for code sets with the same number of features.

Thus far we have tacitly assumed that the subgroups will be exactly the same as those from human data. However, even if a set of semantic codes can be shown to have a  $DVI$  of 6 and a high  $DR$ , it cannot be guaranteed that all its items would actually be categorized into the correct groups based solely on their intrinsic correlations. To evaluate this we needed to test whether the clustering results based on the intrinsic correlations among semantic vectors were similar to semantic categories from human data. We tested this by using the adjusted rand index ( $ARI$ ) (Hubert & Arabie, 1985).  $ARI$  is commonly used to measure the similarity between two different ways of partitioning a set of items. To compare the partitions of human data and the artificial semantic codes,  $ARI$  counts the number of agreements and disagreements between them. It ranges from 0 to 1, with 0 indicating the two partitions are completely different and with 1 indicating the two partitions are exactly the same.

All three indices ( $DVI$ ,  $DR$  and  $ARI$ ) were used to evaluate the semantic vectors. The maximum number of active features including positive and negative in a semantic vector was set to 40 and the minimum was 10. Thus, the population sparseness ranged from 0.05 to 0.2. The numbers of positive ( $n$ ) and negative ( $m$ ) features varied in a complementary manner which was dependent on the total

active features ( $t$ ). To find out what were the optimum numbers of  $n$  and  $m$ , 24 different combinations of positive and negative features were assessed by using the three indices:  $DVI$ ,  $DR$  and  $ARI$ . The evaluations were performed in two steps. The first step was to compare different sets of semantic vectors based on the predefined categories by choosing the combinations with a  $DVI$  of 6 and using the  $DR$  score to select the top candidates within groups with the same number of features. The second step was to use the  $ARI$  score as an independent additional test to confirm that the candidate with the highest  $ARI$  was also one of the possible candidates from the first step.

## Results

### Searching the best semantic vectors

Table 2 shows the results of the six candidates from 24 sets of semantic codes with different combinations of positive and negative features, in which they had the maximum possible number of  $DVI$  (6). The set with 5 positive and 15 negative features (ID 2) had the largest value of  $ARI$ . This set was one of those with the maximum value of  $DVI$  and the value of  $DR$  for this set was also larger than that for other candidate sets with the same total number of features. These results suggest that for this application a set of semantic vectors with 5 positive features and 15 negative features best captures the semantic categories generated from human data. It is also interesting to note that among the top candidates ID 1 was the only one with only positive features and its  $ARI$  was much lower than that for all the other possible candidates. The differences between  $ARI$  for the top candidate and for the two other candidates (ID 3 and ID 5) which included both positive and negative features were relatively small, suggesting that the exact number of positive and negative features may not be critical. However the majority of candidate codes (4/6) did have more negative than positive features.

Table 2: Results of searching the best semantic vectors

ID	Total	Positive	Negative	$DVI$	$DR$	$ARI$
1	10	10	0	6	5.62	0.38
2	20	5	15	6	5.74	<b>0.57</b>
3	20	10	10	6	5.70	0.54
4	30	5	25	6	5.83	0.45
5	30	10	20	6	5.77	0.52
6	40	5	35	6	5.84	0.49

Note: ID: identification;  $DVI$ : distance validity index;  $DR$ : distance ratio;  $ARI$ : adjusted rand index

To further test the significance of including negative codes, 20 sets of semantic codes for each of three groups (positive, neutral and negative-biased) were generated. Each set had the same number of features, ranging from 10 to 48. In the positive group, the vector included only positive

features whereas the vector in the neutral group had an equal number of positive and negative features. In the negative-biased group, the vector had more negative than positive features with a ratio of 3:1. One-tailed paired t-tests were conducted to compare both the neutral and negative-biased groups to the positive group, where the three indices were used as dependent variables. As predicted, the *DVIs* and *ARIs* of the neutral and negative-biased groups were significantly higher than that for the positive group (Table 3). For the *DRs*, the difference between the negative-biased and the positive groups was not significant while there was a significantly lower mean *DR* for the neutral group than for the positive group. The comparison between the negative-bias and neutral groups showed that both *DVI* and *DR* were higher for the negative-biased group than for the neutral group and there was no difference in their *ARIs*. Overall the results demonstrated the negative-biased group was superior to both the positive and neutral groups, confirming that the inclusion of negative codes is important to capturing the way semantic knowledge is represented in humans.

Table 3: Results of Significance Tests

Group	Index	MD	MSE	t Value	df	P Value
Neutral -Positive	<i>DVI</i>	0.05	0.01	3.63	19	.001*
	<i>DR</i>	-0.10	0.03	-3.08	19	.003*
	<i>ARI</i>	0.45	0.14	3.33	19	.002*
Negative-Biased -Positive	<i>DVI</i>	0.04	0.02	1.78	19	.046*
	<i>DR</i>	-0.04	0.03	-1.25	19	.114
	<i>ARI</i>	0.75	0.12	6.10	19	<.001*
Negative-Biased -Neutral	<i>DVI</i>	0.30	0.11	2.85	19	.005*
	<i>DR</i>	0.06	0.01	7.58	19	<.001*
	<i>ARI</i>	-0.02	0.02	-0.87	19	.197

Note: *DVI*: distance validity index; *DR*: distance ratio; *ARI*: adjusted rand index  
MD: mean difference; MSE: mean standard error; \*: *P* value is significant at the .05 level.

## Hierarchical Clustering Analysis

Figure 1 shows the hierarchical clustering based on the optimum vectors indicated in Table 2.

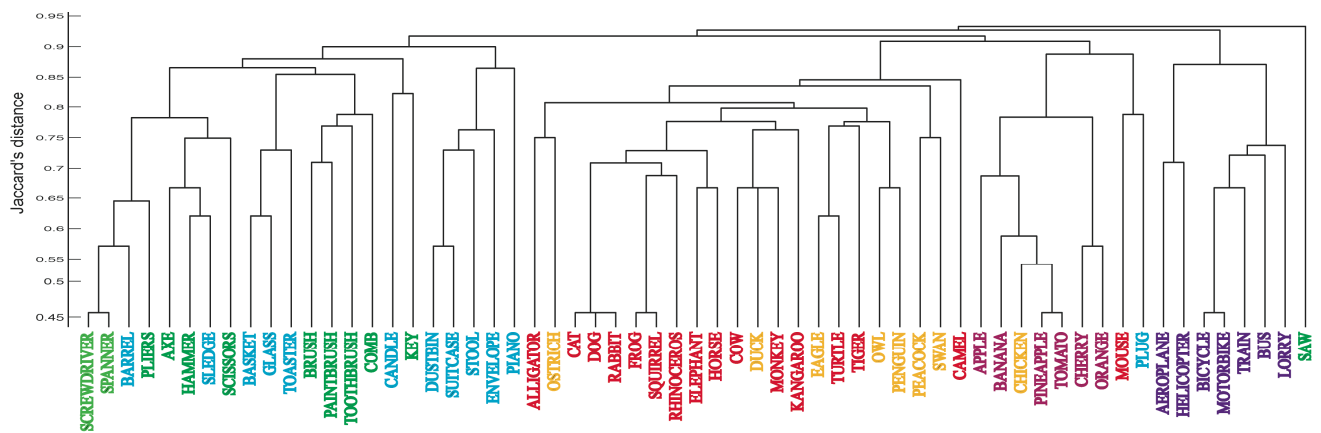


Figure 1: Hierarchical cluster analysis of 61 words based on the 200-dimensional semantic vectors with 5 positive and 15 negative features. (Items coloured base on human categories: Animal (Red), Birds (Orange), Fruit (Purple), Tool (Green), H'hold (Dark Blue), Vehicles (Purple)).

The y axis shows the Jaccard's distance, a measure of similarity between words. The lower the value the more similar the clusters are. The semantic vectors can accurately represent the semantic categories at a coarse scale, which means that the living things and nonliving things are well separated. To compare the clustering results with human data collected by Garrard et al. (2001), the items in Figure 1 were coloured according to its category in the human semantic data. This clearly shows whether the clustering results were consistent with human categories. Ideally, items with the same colour would be clustered together, indicating the items are clustered into the same group as in the human category. Most of the items are correctly clustered. However there are a few interesting exceptions, for example, the word "chicken" was clustered into the fruit category (items coloured in purple) based on the artificial semantic codes, while it should have been clustered into the bird category (Orange). Presumably this is because in many texts the word "chicken" might more frequently co-occur with other food (including fruit) in the kitchen context so this category might be more accurately described as food. Within the nonliving things, it appears that the broader category of tool is well distinguished from the vehicle group but the boundaries between tools and household items is less clear. It is likely that most of the tools and household objects tend to occur in a similar context in the text so it would be difficult to differentiate them in a fine scale by using the co-occurrence statistic approach.

## General Discussion

Several schemes for generating semantic codes have been reviewed in this paper with a focus on the requirements of computational modelling. The primary aim was to determine an appropriate system for representing semantic knowledge, which could be used for a large-scale computational modelling of semantic related tasks.

The desired computational criteria were as follows: (1) binary coding; (2) sparse coding; (3) fixed number of active units in a vector; (4) scalable vectors; (5) preservation of inherent semantic structure. The COALS system (Rohde et al., 2006) provided the best fit to the criteria. The original COALS system discretizes the real-valued vectors based only on the positive components. However, we evaluated codes with varying numbers of positive and negative features by comparing the semantic categories generated from the artificial semantic codes with human category data from Garrard et al.'s (2001) study. The results showed that a set of semantic vectors having 5 positive and 15 negative features could best account for the human semantic categories. It was perhaps surprising that the best binary vectors found here had more negative features than positive features (i.e., 15 negative features v.s. 5 positive features). This is at variance with the prevalent assumption that only positive components should be used to construct semantic codes. Positive features reveal what attributes are likely to be present; in contrast, the negative features provide information about what attributes are likely to be absent. So this result implies that knowledge of the absence of features (e.g., knowing that washing machines cannot walk) is as important as knowledge of positive features. Given both of these types of information, it is possible to separate categories on the basis of their distance in the semantic space, as the optimisation results shown in Table 2. Further significance tests reveal that there was a clear trend for semantic vectors containing both positive and negative features to show a more human-like semantic structure, suggesting that this may be a generally applicable principle. Overall the best performing set of semantic vectors matched the human categorisation data quite well. Further work may be conducted to investigate the performance of the present semantic codes on other types of semantic tasks (e.g., synonym judgement) for a more complete evaluation.

In addition, there are some inherent limitations of using these semantic vectors. The first is that the semantic features are not interpretable because they only encode the semantic regularities among word meanings. What exactly the feature represents is difficult to interpret. But this is only a problem in applications where a direct interpretation of features is required. Another limitation is that it can generate good semantic vectors for most of the uninflected words but it could be difficult to properly account for the deeper meaning of words like morphological regularities (e.g., bake/baker) (Harm, 2002), which would require some additional coding.

To summarise, a novel semantic representation scheme was produced based on modifications to the COALS system. This was evaluated against human categorisation data, and the resultant coding scheme was able to reproduce the human data quite closely. The key finding was that the negative features, which indicate what attributes definitely do not belong to a lexical item, were at least as important as the positive features. The semantic system developed here can be applied to generate semantic codes for a larger word

list used to train more sophisticated computational models.

## Acknowledgments

This research was supported by grants under the Cognitive Foresight Initiative (jointly funded by EPSRC, MRC and BBSRC - EP/F03430X/1) and the Neuroscience Research Institute at the University of Manchester.

## References

- Dilkina, K., McClelland, J. L. & Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology*, 25(2), 136-164.
- Furber, S. B., Bainbridge, W. J., Cumpste, J. M. & Temple, S. (2004). A sparse distributed memory based upon n-of-m codes. *Neural Networks*, 17.
- Garrard, P., Lambon Ralph, M. A., Hodges, J. R. & Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, 18(2), 125-174.
- Harm, M. W. (2002). Building large scale distributed semantic feature sets with wordnet. Pittsburgh, Carnegie Mellon University.
- Harm, M. W. & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, 111(3), 662-720.
- Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193-218.
- Landauer, T. K., Foltz, P. W. & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3), 259-284.
- Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods Instruments & Computers*, 28(2), 203-208.
- McRae, K., Cree, G. S., Seidenberg, M. S. & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4), 547-559.
- Miller, G. A. (1990). Introduction to wordnet: An on-line lexical database\*. *International Journal of Lexicography*, 3(4), 235.
- Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes*, 12(5-6), 765-805.
- Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R. & Patterson, K. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, 111(1), 205-235.
- Rohde, D. L. T., Gonnerman, L. & Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *COMMUNICATIONS OF THE ACM*, 8, 627-633.

# Word Form Encoding in Mandarin Chinese Typewritten Word Production

Jenn-Yeu Chen<sup>1,2</sup> (psyjyc@mail.ncku.edu.tw) and Train-Min Chen<sup>1</sup> (trainmin@gmail.com)

<sup>1</sup> Department of Psychology, National Cheng Kung University, Tainan, Taiwan ROC

<sup>2</sup> Institute of Teaching Chinese as a Second Language, National Taiwan Normal University, Taipei, Taiwan ROC

## Abstract

Employing the implicit priming task, we examined whether Chinese words that shared the initial onset consonant could be typed, using the phonetic-based method (called zhuyin), with faster response times than words that did not share the initial onset consonant. We also examined the effect of sharing the initial tonal syllable. A significant onset preparation effect and a significant syllable preparation effect were both observed. The latter was found to vary linearly with the number of segments in the syllable. The slope of 63 ms was similar to the 70-ms onset effect, suggesting that the syllable effect was segment-based. The results contrasted with the lack of an onset effect previously reported for speaking, and were interpreted as supporting the Output Constraint Hypothesis which states that the kind of outputs a production system is designed to produce (speaking vs. typing) can flexibly and adaptively alter the way the system is organized and operates.

**Keywords:** Chinese; Typing; Speaking; Phonological Encoding; Word-Form Encoding

## Introduction

With the popularization of computers and internet, typing has become a new form of communication that may someday dominate our social life. It is, therefore, of interest to study the cognitive processes involved in typing, in particular, how typing as a language production activity may differ from speaking. Past research has studied typing more as a skilled motor activity during transcription (Shaffer, 1975; Sternberg et al., 1978; Rumelhart & Norman, 1982; Norman & Rumelhart, 1983; Salthouse, 1986; Crump & Logan, 2010a, 2010b) than as a language production activity.

In a spoken language production model (e.g., Dell, 1986; MacKay, 1987; Levelt et al., 1999), word form encoding refers to the hierarchically organized processes that translate the semantic/syntactic representation of a word into its phonological and phonetic forms. The processes involve retrieving the structural frame and the phonemic segments of a word, followed by assigning, in a sequential order, the segments to their categorized slots in the frame. An issue under much debate concerns the units that make up the stored phonological representation of a word and drive the phonological encoding process. In some models, they contain the syllables (e.g., Dell, 1986; MacKay, 1987), while in others they do not (e.g., Levelt et al., 1999). Prevailing evidence from Indo-European languages suggests that the units are the segments (e.g., Meyer, 1990, 1991; Roelofs & Meyer, 1998; Roelofs, 1999; Schiller, 1998, 2004; Schiller & Costa, 2006). But in Mandarin Chinese, they appear to be the syllables (J.-Y. Chen et al., 2002; J.-Y. Chen et al., 2003; T.-M. Chen et al., 2007; O'Seaghdha et al., 2010). The

varying units of a word's stored phonological representation in different languages may have something to do with the rhythmic structures of the languages (Cutler et al., 1986; Otake et al., 1996).

The units may also vary with different kinds of outputs targeted by different production tasks within the same language, e.g., typing as opposed to speaking. Mandarin Chinese provides an excellent testing bed for evaluating this hypothesis. A popular method of typing in Chinese uses the phonetic alphabet called zhuyin in Taiwan and pinyin in China. For example, to type the character 潔 ('clean', jie2) in zhuyin, the onset consonant j, the medial glide i, the rhyme e, and the tone 2 are typed on a keyboard sequentially, followed by the selection among a list of homophones. To this extent, zhuyin typing requires accessing the phonological codes of the character, much like speaking. Employing a traditional (unmasked) priming task and comparing naming with zhuyin typing, our previous study (Chen & Li, 2011) investigated whether syllable onset priming was absent in naming, which would be consistent with our findings for speaking (J.-Y. Chen et al., 2002; J.-Y. Chen et al., 2009; O'Seaghdha et al., 2010), but might be present in zhuyin typing.

A possible reason for predicting onset priming in zhuyin typing is that the output of zhuyin typing consists of discrete manual keystrokes that correspond to the onset, medial, rhyme and tone of a syllable. This is different from the output of naming (and speaking), which consists of syllable-sized articulatory gestures. That is, the different output characteristics constrain the way a word is planned during speaking and typing. Indeed, Berg (2002) observed that slips of the key resembled slips of the pen, but not slips of the tongue. He suggested that 'speaking is characterized by a hierarchical strategy of activation while typewriting is subject to the so-called staircase strategy of serialization in which activation is a function of linear distance' (p.185). Although such a prediction seems obvious and only expected, there are reasons to predict otherwise too. Studying handwriting, Kandel and colleagues (Kandel et al., 2006; Lambert et al., 2007; Alvarez et al., 2009) found that interletter intervals were longer between syllables than within syllables and that the number of syllables of a word modulated the time course of handwriting production, indicating that word syllable structure constrains motor production both in French and Spanish. Given that very similar processes are believed to underlie typing and writing (Berg, 2002), it is reasonable to assume that syllables are also essential units of processing in typing. Direct evidence for this assumption has also been reported (Nottbusch et al., 2005). The results from primed naming and primed zhuyin

typing showed significant onset priming for zhuyin typing but not for naming, supporting the hypothesis that the units of the stored phonological representation of a word vary with different kinds of outputs targeted by different production tasks within the same language. We will refer to this hypothesis as the Output Constraint Hypothesis (OCH).

In the present study, the OCH was tested further with the implicit priming task. The implicit priming task (also known as the form preparation task) has been used extensively in investigating the word form encoding processes in spoken production (Meyer, 1990). The task requires the participants to learn a set of prompt-target word pairs during the learning phase. During the testing phase, the prompt words are shown one at a time and the participants have to say the corresponding target words as responses. The target words are arranged in a homogeneous context such that they share the initial portion of their phonological forms (e.g., the onset consonant). In a heterogeneous context, the same target words are re-arranged such that they no longer share the initial portion of the phonological forms. Response latencies tend to be shorter when the target words are produced, upon the presentation of the prompt words, in the homogeneous context than in the heterogeneous context. The response benefit is attributed to the suspension-resumption mechanism in the production system, according to which the system prepares a word from left to right in an incremental fashion, and it can prepare the word as far to the right as the left portion is known. The system suspends the operation when everything that is known has been prepared, and resumes operation as soon as information about the rest of the word is received. It is assumed that the portion that can be prepared by the system represents the unit of word form encoding during spoken production (Roelofs, 1997a, 1997b). Sufficient evidence has indicated that this unit must be the size of a syllable in Mandarin Chinese, but can be a phonemic segment in English and Dutch.

Because previous studies in Mandarin Chinese have already consistently documented no onset preparation effect with an implicit priming task in speaking (Chen, Chen, & Dell, 2002; O'Seaghdha, Chen, & Chen, 2010), the present study examined typing only, but contrasted the findings with those reported for speaking base on the same materials. The OCH predicts that the onset segment of a Chinese word can be prepared during zhuyin typing, i.e., an onset preparation effect is predicted in an implicit priming task of zhuyin typing.

In addition to the syllable onset, the full syllable was also examined. If phonological encoding in zhuyin typing is segment-driven as predicted by the OCH, a syllable preparation effect that is a function of the number of segments in a syllable should be observed.

## Method

### Participants

Sixteen native Mandarin Chinese speakers from the student body of National Cheng Kung University were recruited for

the onset experiment, and another sixteen for the syllable experiment. They were all native and habitual zhuyin typists, i.e., they learned the zhuyin typing when they first learned typing and have been typing in zhuyin ever since. All the participants had normal or corrected-to-normal vision and they were paid for participation.

### Apparatus and Materials

The experiment was programmed in DMDX (Forster & Forster, 2003) and run on a personal computer (Intel® Core™2 Quad CPU, Q6600@2.40GHz) with a 20-inch LED screen (32bits, 1400x1050 pixels, 8-ms refresh rate) and a standard keyboard that included marks of the zhuyin letters.

The stimulus materials for the onset experiment were disyllabic words taken out of Experiment 5 of J.-Y. Chen et al. (2002). They consisted of four sets of prompt-target word pairs, with four pairs in each set. The prompt and the target in a pair bore clear semantic or associative relationship such that they could be learned easily. The target words were chosen such that they shared the same onset consonant of the first syllable in a set. Across the four sets, four different onset consonants were used (m, d, sh, l). These formed the homogeneous sets (see Set 1-4 in Table 1). The same target words were reshuffled to form the four heterogeneous sets such that within a set the target words no longer shared the onset consonant (see Set 5-8 in Table 1). The arrangement of the stimulus materials was identical to that of J.-Y. Chen et al. (2002) Experiment 5.

The stimulus materials for the syllable experiment were disyllabic words taken out of Experiment 3 of T.-M. Chen & J.-Y. Chen (2006) (see Table 2), and arranged in the same way. The target words in a homogeneous set shared the first tonal syllables.

Table 1: Target words arranged as homogeneous sets (1-4) and heterogeneous sets (5-8) for the onset experiment.

Homogeneous				
Set	1	2	3	4
Heterogeneous	mo1-cai3 摸彩 draw lots	da1-ing4 答應 promise	shu1-fa3 書法 calligraphy	luo1-suo1 囉唆 nagging
	ma2-que4 麻雀 sparrow	de2-kuo2 德國 Germany	shi2-yan4 實驗 experiment	li2-ge1 驪歌 farewell song
	mu3-dan1 牡丹 peony	du3-buo2 賭博 gambling	she3-qi4 捨棄 abandon	la3-ba1 喇叭 horn
	mi4-yue4 蜜月 honeymoon	di4-yu4 地獄 hell	shou4-ruo4 瘦弱 weak	lu4-shi1 律師 lawyer



Table 2: Target words arranged as homogeneous sets (1-4) and heterogeneous sets (5-8) for the syllable experiment.

Homogeneous				
Set	1	2	3	4
Heterogeneous	5	xi1-gua1 西瓜 watermelon	hong2-shui3 洪水 flood	jia1-fa3 加法 addition
	6	xi1-fan4 稀飯 porridge	hong2-guan1 宏觀 macroscopic	yi4-wen2 軼聞 anecdote
	7	xi1-guan3 吸管 straw	jia1-bin1 嘉賓 honored guests	yi4-wei4 異味 peculiar smell
	8	xi1-shui3 溪水 stream	hong2-mo2 虹膜 iris	jia1-shi4 家事 household duty
			jia1-yao2 佳餚 delicacy	yi4-ren2 藝人 entertainer
				yi4-chu4 益處 benefit

## Design and Procedure

The design and the procedure were identical to the experiments where we took the materials from. Each pair in a set was repeated four times (the Repetition factor) so that there were 16 pairs and they appeared in a random order within a block. Half of the participants received the homogeneous sets first and the other half received the heterogeneous sets first (the Sequence factor). The participants went through the round of homogeneous and heterogeneous sets (the Context factor) three times (the Round factor) and in the same sequence. The type of onset consonants or syllables constituted another factor (Onset). The Sequence factor was a between-subjects factor while the rest were within-subjects factors.

During the learning phase, the participants were shown the four pairs of words in a set. They learned the association between the two words in each pair until they had memorized the pairs well. Then the target words were cued one at a time by their associated prompt words. When the participants succeeded in reporting the target words correctly without hesitation, they proceeded to the testing phase. Otherwise, they repeated the learning phase.

During the testing phase, each trial began with a 1000-Hz warning tone and two short dashed lines flanking a blank space at the center of the screen. The tone and the dashed lines appeared for 200 ms. The prompt word appeared in the previously flanked space 600 ms later. The prompt word stayed on the screen for 150 ms. Another 1850 ms elapsed before the trial ended. The participants were told to type in zhuyin the target word upon seeing the prompt word, as quickly and accurately as possible. The participants entered the zhuyin letters in the English input mode. Accordingly, no homophonous characters were shown for selection after the zhuyin letters of a character have been entered. Response latencies were measured, to the accuracy of milliseconds, from the presentation of the prompt word to

the striking of the first zhuyin letter. If no response was initiated within 2000 ms of the presentation of the prompt word, a feedback tone of 500 Hz was sounded for 200 ms. The next trial began after another 200 ms. A practice session containing four trials was given before the experiment began. The participants were seated 60 cm from the screen. Each character measured 1.6 cm in height and 1.1 cm in width at that viewing distance.

## Results

### Onset Experiment

Error rates were 2% in the homogeneous condition and 4% in the heterogeneous condition. Response times ranged from 374 to 1946 ms for the homogeneous trials (mean: 825, SD: 190), and ranged from 354 to 1988 ms for the heterogeneous trials (mean: 894, SD: 195). No apparent outliers were noted. All response times (RTs) for the correct trials were then analyzed using a linear mixed model (Statistical Analytic System, the PROC MIXED procedure) with subjects and items as random-effect variables and context, sequence, round, repetition as fixed-effect variables. Most notable in the analysis was the significant main effect of context:  $F(1, 14) = 33.67, p < .0001$ . The mean RT was 824 ms in the homogeneous context and 894 ms in the heterogeneous context. The difference represents an onset preparation effect of 70 ms. The context effect varied with sequence:  $F(1, 1465) = 4.92, p < .03$ , being greater when the heterogeneous sets appeared first than when the homogeneous sets appeared first. The context by sequence interaction also varied with round:  $F(2, 1465) = 8.46, p < .0005$ . The three-way interaction is manifested as the context effect displaying an increasing trend when the homogeneous trials were done first and a decreasing trend when the heterogeneous trials were done first (see Table 2). The remaining effects are not enumerated here because they were either non-significant ( $p$ 's  $> .06$ ) if they involved the context factor, or significant but did not involve the context factor. Table 3 summarizes the results by presenting the mean RTs as a function of context, round, and sequence.

Table 3: Mean RTs (SEs in the parentheses) as a function of context, round, and sequence for the onset experiment.

Sequence	Round	Homo- geneous Context	Hetero- geneous Context	Prepara- tion Effect
Homo- geneous First	1	909 (18)	932 (24)	23
	2	824 (20)	875 (18)	51
	3	789 (18)	848 (19)	59
	<b>Overall</b>	<b>841 (12)</b>	<b>885 (12)</b>	<b>44</b>
Hetero- geneous First	1	835 (20)	964 (17)	129
	2	787 (19)	876 (18)	89
	3	796 (22)	866 (19)	70
	<b>Overall</b>	<b>806 (12)</b>	<b>902 (11)</b>	<b>96</b>

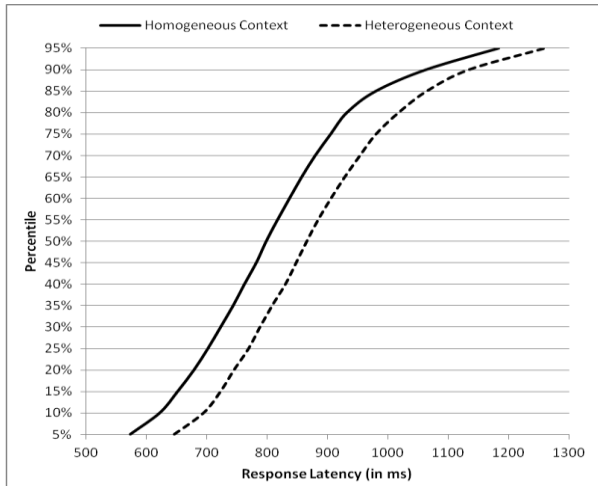


Figure 1: Cumulative response latency distributions for the homogeneous context and the heterogeneous context for the onset experiment.

Although the participants were told to complete all key presses without hesitation or pauses once a response was initiated, it is possible that the homogenous context might have encouraged a strategic behavior in them to start the first key press without having planned for the subsequent keys. If this was the case, the duration of a response should be longer in the homogeneous context than in the heterogeneous context. Unfortunately, response durations were not available to rule out this possibility. However, we plotted cumulative response latency distributions for the two conditions, following Damian and Stadthagen-Gonzalez (2009). As Figure 1 shows, the differences between the two distributions are relatively uniform across the entire spectrum of response latencies, suggesting that the strategy of immaturely starting responses on homogeneous trials was not used by our participants. The similar distributions of the two conditions also rule out the possibility that participants were able to locate the first key and initiate a response on homogeneous trials sooner than on heterogeneous trials, where the first keys were different and took time to locate.

### Syllable Experiment

Error rates were 2% in the homogeneous condition and 5% in the heterogeneous condition. Response times less than 250 ms were excluded, making up 0.8% of all trials, before they were subject to the same kind of analysis as in the onset experiment. The context effect was significant, with the homogeneous RTs being 255 ms faster, on the average, than the heterogeneous RTs (620 ms vs. 875 ms):  $F(1, 14) = 69.98$ ,  $p < .0001$ . The context effect did not vary with sequence ( $p > .8$ ), but it increased significantly across rounds ( $p < .01$ ). The three-way interaction involving context was not significant,  $p > .19$ . Table 4 summarizes the results of the syllable experiment.

The cumulative distribution plot of Figure 2 shows no clear evidence of strategic responding for the homogeneous

trials. To examine if the syllable preparation effect increased with the number of segments in the syllable (tone being counted as a segment), a by-item linear regression analysis was performed, which revealed a slope of 63ms. This is fairly close to the 70 ms onset preparation effect.

Table 4: Mean RTs (SEs in the parentheses) as a function of context, round, and sequence for the syllable experiment.

Sequence	Round	Homo- geneous Context	Hetero- geneous Context	Prepara- tion Effect
Homo- geneous First	1	669 (29)	906 (17)	237
	2	591 (29)	870 (16)	279
	3	584 (32)	858 (14)	274
	<b>Overall</b>	<b>615 (18)</b>	<b>878 (9)</b>	<b>263</b>
Hetero- geneous First	1	678 (13)	916 (12)	238
	2	619 (18)	857 (10)	238
	3	577 (19)	842 (12)	265
	<b>Overall</b>	<b>625 (11)</b>	<b>872 (7)</b>	<b>247</b>

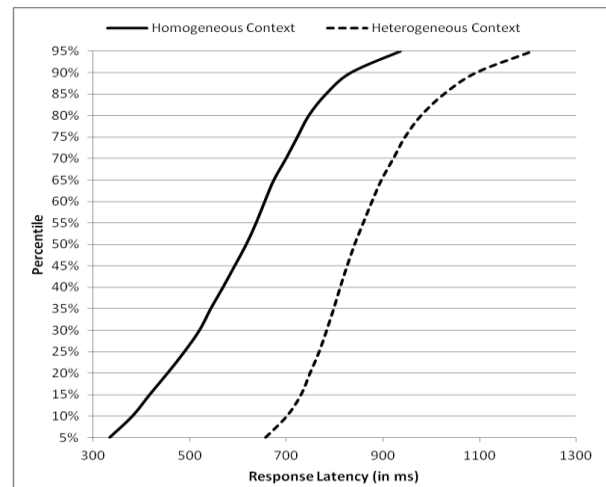


Figure 2: Cumulative response latency distributions for the homogeneous context and the heterogeneous context for the syllable experiment.

### Discussion

Employing the implicit priming task, a widely used tool for studying word form encoding in spoken production, we examined whether words that shared the initial onset consonant could be typed with faster response times than words that did not share the initial onset consonant. The result of the experiment was clear. There was a significant and sizeable onset preparation effect when words to be typed shared the initial onset consonant (70 ms). This contrasted interestingly with the small and non-significant onset effect observed in our previous work when the task was speaking (J.-Y. Chen et al., 2002, Experiment 5 with the same material: -1 ms; O'Seaghdha et al., 2010, Experiments 1-4 and 7 with different materials: 3, -6, 3, 4, 2 ms). We also observed a large tonal syllable preparation

effect which varied with the number of segments in the syllable. The slope of 63 ms was similar to the 70 ms onset preparation effect, suggesting that the syllable preparation effect was likely segment-based. This also contrasted interestingly with the syllable preparation effect previously observed for speaking, which could be more unambiguously attributed to the syllable. Together, these results extended the previous findings of significant onset priming for typing (30 ms) but no significant onset priming for naming (-5.6 ms) when an unmasked priming task was employed (Chen & Li, 2011).

Speaking a Chinese word and typing it in zhuyin take the same input for processing, i.e., the concept of the word. They also require retrieving the same phonological representation of the word. However, the two tasks have different goals, aiming at different outputs. The findings of the present study as well as the earlier one (Chen & Li, 2011) suggest that the form of the output can constrain the internal organization/mechanism of the production system. Speaking, aiming at syllable-sized gestures, requires a syllable-driven word form encoding process. Typing, aiming at segment-sized gestures, involves segment-sized word form encoding process. Thus, all production systems are not organized in the same way. The kind of outputs a production system is designed to produce can flexibly and adaptively alter the way the system is organized and operates.

It has been suggested that the traditional priming paradigm and the implicit priming paradigm tap different levels of word form encoding process (Levelt et al., 1999; Cholin, Schiller, & Levelt, 2004). Traditional priming works to pre-activate the segments of a word, facilitating its phonological encoding process. Its effect takes place at the early stage of phonological encoding. Implicit priming is said to work at this early stage of phonological encoding as well as at the later stages of phonological encoding and phonetic encoding (i.e., online syllabification and possibly accessing the mental syllabary). Because onset priming in zhuyin typing was observed with both the explicit priming paradigm and the implicit priming paradigm, it may be concluded that the production system respects the form of its output and gets ready for that form at the stage as early as the beginning of the word form encoding process.

One caveat against the above conclusion is that typing is typically much slower than speaking, indicating low automaticity, and this is perhaps the reason that typing is less hierarchically organized than speaking (Berg, 2002). Future work will investigate this with professional typists.

The output constraint is not unique to the production system only, but finds an analog in the perception system too, where it is the input that constrains how the perception system is organized and operates. For example, research has shown that the structural and functional basis of word recognition and reading varies between an alphabetic writing system like English and a logographic writing system like Chinese (Perfetti, Liu, & Tan, 2005; Tan, Spinks, Eden, Perfetti, & Siok, 2005; Tzeng & Hung, 1981; Kuo, Yeh, Duann, Wu et al., 2001; Kuo, Yeh, Lee, Wu et al.,

2003; Siok, Niu, Jin, Perfetti, & Tan, 2008; Siok, Perfetti, Jin, & Tan, 2004).

When building a model of language processing, a universal one is always preferred. But, even a universal model needs to incorporate flexible parameters and constraints to accommodate the variations across languages and tasks. One source of such constraints might be sought from the input and output a particular language system is designed to process. This view carries the Gibsonian tradition of emphasizing the role of environment in perception (Gibson, 1986), but extends it to production.

## Acknowledgments

This work was supported by the NSC 100-2410-H-006-023-MY3 grant.

## References

- Alvarez, C. J., Cottrell, D., & Afonso, O. (2009). Writing dictated words and picture names: Syllabic boundaries affect execution in Spanish. *Applied Psycholinguistics*, 30, 205-223.
- Berg, T. (2002). Slips of the typewriter key. *Applied Psycholinguistics*, 23, 185-207.
- Chen, J.-Y., Chen, T.-M., & Dell, G. S. (2002). Word-form encoding in Mandarin Chinese as assessed by the implicit priming task. *Journal of Memory and Language*, 46, 751-781.
- Chen, J.-Y., & Li, C.-Y. (2011). Word form encoding in Chinese word naming and word typing. *Cognition*, 121, 140-146.
- Chen, J.-Y., Lin, W.-C., & Ferrand, L. (2003). Masked priming of the syllable in Mandarin Chinese speech production. *Chinese Journal of Psychology*, 45, 107-120.
- Chen, J.-Y., Chen, T.-M., & O'Seaghdha, P. G. (2009). Word form encoding in Chinese begins with the syllable: Further evidence from masked primed picture naming. Poster presented at the 50th Annual Meeting of the Psychonomic Society, November 19-22, Boston.
- Chen, T.-M., Dell, G. S., & Chen, J.-Y. (2007). A cross-linguistic study of phonological units: syllables emerge from the statistics of Mandarin Chinese, but not from the statistics of English. *Chinese Journal of Psychology*, 49, 137-144.
- Cholin, J., Schiller, N. O., & Levelt, W. J. M. (2004). The preparation of syllables in speech production. *Journal of Memory and Language*, 50, 47-61.
- Crump, M. J. C., & Logan, G. D. (2010a). Episodic contributions to sequential control: Learning from a typist's touch. *Journal of Experimental Psychology: Human Perception and Performance*, 36, 662-672.
- Crump, M. J. C., & Logan, G. D. (2010b). Hierarchical control and skilled typing: Evidence for word-level control over the execution of individual keystrokes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 1369-1380.
- Cutler, A., Mehler, J., Norris, D., Segui, J. (1986). The syllable's differing role in the segmentation of French and

- English. *Journal of Memory and Language*, 25, 385-400.
- Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35, 116-124.
- Gibson, J. J. (1986). *The ecological approach to visual perception*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Kandel, S., Alvarez, C. J., & Vallee, N. (2006). Syllables as processing units in handwriting production. *Journal of Experimental Psychology: Human Perception and Performance*, 32, 18-31.
- Kubozono, H. (1996). Speech segmentation and phonological structure. In T. Otake and A. Cutler (Eds.), *Phonological structure and language processing* (pp. 77-94). Berlin: Mouton de Gruyter.
- Kuo, W.-J., Yeh, T.-C., Duann, J.-R., Wu, Y.-T., Ho, L.-T., Hung, D., Tzeng, O. J.-L., Hsieh, J.-C. (2001). A left-lateralized network for reading Chinese words: a 3 T fMRI study. *Neuroreport*, 12, 3997-4001.
- Kuo, W.-J., Yeh, T.-C., Lee, C.-Y., Wu, Y.-T., Chou, C.-C., Ho, L.-T., Hung, D. L., Tzeng, O. J.-L., & Hsieh, J.-C. (2003). Frequency effects of Chinese character processing in the brain: an event-related fMRI study. *NeuroImage*, 18(3), 720-730.
- Kureta, Y., Fushimi, T., & Tatsumi, I. F. (2006). The functional unit of phonological encoding: Evidence for moraic representation in native Japanese speakers. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 32, 1102-1119.
- Lamert, E., Kandel, S., Fayol, M. & Espéret, E. (2008). The effect of the number of syllables on handwriting production. *Reading & Writing*, 21, 859-883.
- Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral & Brain Sciences*, 22, 1-75.
- MacKay, D. G. (1987). *The organization of perception and action*. New York: Springer-Verlag.
- Meyer, A. S. (1990). The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language*, 29, 524-545.
- Meyer, A. S. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory and Language*, 30, 69-89.
- Norman, D. A., & Rumelhart, D. E. (1983). Studies of typing from the LNR Research Group. In W. E. Cooper (Ed.), *Cognitive aspects of skilled typewriting* (pp. 45-65). New York: Springer-Verlag.
- Nottbusch, G., Grimm, A., Weingarten, R., Will, U. (2005). Syllabic structures in typing: Evidence from deaf writers. *Reading and Writing*, 18, 497-526.
- O'Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition*, 115, 282-302.
- Otake, T., Hatano, G., Cutler, A. & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, 32, 258-278.
- Perfetti, C. A., Liu, Y., & Tan, L. H. (2005). The lexical constituency model: Some implications of research on Chinese for general theories of reading. *Psychological Review*, 112, 43-59.
- Roelofs, A. (1997a). Syllabification in speech production: Evaluation of WEAVER. *Language and Cognitive Processes*, 12, 657-693.
- Roelofs, A. (1997b). The WEAVER model of word-form encoding in speech production. *Cognition*, 64, 249-284.
- Roelofs, A., & Meyer, A. S. (1998). Metrical structure in planning the production of spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24, 922-939.
- Roelofs, A. (1999). Phonological segments and features as planning units in speech production. *Language and Cognitive Processes*, 14, 173-200.
- Rumelhart, D. E., & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive Science*, 6, 1-36.
- Salthouse, T. A. (1986). Perceptual cognitive, and motoric aspects of transcription typing. *Psychological Bulletin*, 99, 303-319.
- Schiller, N. O. (1998). The Effect of visually masked syllable primes on the naming latencies of words and pictures. *Journal of Memory and Language*, 39, 484-507.
- Schiller, N. O. (2004). The onset effect in word naming. *Journal of Memory and Language*, 50, 477-490.
- Schiller, N. O., & Costa, A. (2006). Activation of segments, not syllables, during phonological encoding in speech production. *The Mental Lexicon*, 1, 231-250.
- Shaffer, L. H. (1975). Control processes in typing. *Quarterly Journal of Experimental Psychology*, 27, 419-432.
- Siok, W. T., Niu, Z., Jin, Z., Perfetti, C. A., & Tan, L. H. (2008). A structural-functional basis for dyslexia in the cortex of Chinese readers. *PNAS*, 105, 5561-5566.
- Siok, W. T., Perfetti, C. A., Jin, Z., & Tan, L. H. (2004). Biological abnormality of impaired reading is constrained by culture. *Nature*, 431, 71-76.
- Sternberg, S., Monseil, S., Knoll, R. L., & Wright, C. E. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In G. E. Stelmach (Ed.), *Information processing in motor control and learning* (pp. 117-152). New York: Academic Press.
- Tan, L. H., Spinks, J. A., Eden, G. F., Perfetti, C. A., & Siok, W. T. (2005). Reading depends on writing, in Chinese. *PNAS*, 102, 8781-8785.
- Tzeng, O. J. L. & Hung, D. L. (1981). Linguistic determinism: A written language perspective. In: O. J. L. Tzeng & H. Singer (eds.), *Perception of print: Reading research in experimental psychology* (pp. 237-255). Hillsdale, NJ: Erlbaum.

# What Counts in Mandarin Chinese: A Study of Individuation and Quantification

Pierina Cheung

mpcheung@uwaterloo.ca  
Department of Psychology  
University of Waterloo

Peggy Li

pegs@wjh.harvard.edu  
Laboratory for Developmental Studies  
Harvard University

David Barner

barner@ucsd.edu  
Department of Psychology  
University of California, San Diego

## Abstract

By some accounts, speakers of classifier languages such as Mandarin or Japanese, which lack count-mass syntax, require classifiers to specify individuated meanings of nouns. This paper examines this view by testing how Mandarin speakers interpret bare nouns and use classifier knowledge to guide quantification in four studies. Using a quantity judgment task, Study 1 found that Mandarin speakers interpret nouns like English speakers, regardless of their syntactic status as mass or count in English. Study 2 showed that Mandarin speakers quantified broken objects like English adults, again suggesting that Mandarin nouns specify criteria of individuation. Studies 3 and 4 together showed that classifiers are not typically required for individuation, except when the reference of nouns is semantically ambiguous (e.g., *rock*, *string*) and can denote either objects or substances. In sum, we argue that individuation can be specified lexically in classifier languages like Mandarin, and does not depend on classifier syntax.

**Keywords:** individuation; quantification; nouns; classifiers; word learning; Mandarin; mass-count syntax.

## Introduction

Languages differ in how they express reference to kinds of things. In English and other Indo-European languages, countable things like dogs and cups are typically referred to using count syntax (e.g., *those are dogs*), whereas uncountable entities like milk and sand are expressed as mass nouns (e.g., *that is some milk*). However, other languages, like Mandarin Chinese, make no such syntactic distinction. Instead, nouns in Mandarin, and related classifier languages like Japanese and Tsotsil Mayan, act much like mass nouns in English (Allan, 1980; Chierchia, 1998). Nouns cannot co-occur directly with numerals (*\*san bi* ‘three pen’), but instead require classifiers (CLs) for counting (*san CL-zhi bi* ‘three pens’, is literally translated to ‘three CL-stick pen’). Based on this syntactic distinction, some researchers have argued that nouns in classifier languages may not specify individuation lexically. Instead, languages like Mandarin may rely on classifiers – i.e., words like “bit” and “piece” – to syntactically specify units of individuation, resulting in a fundamental difference in how nouns encode meaning cross-linguistically (e.g., Borer, 2005; Huang & Lee, 2009; Lucy, 1992).

For example, according to Lucy (1992), in classifier languages such as Yucatec Mayan, all lexical nouns “are unspecified as to unit since they all require supplementary marking (i.e., numeral classifiers) in the context of numeral modification” (p. 73). Similarly, in her discussion of Mandarin Chinese, Borer (2005) argues that, “the need for a classifier projection to license counting vs. the absence of

classifiers in the context of mass interpretation confirms the claim that in the absence of classifiers, [noun] predicates in Chinese are interpreted as mass” (p. 108). Under this account, classifiers do not merely reflect the meaning provided by the noun, but actually supply units of individuation and quantification, just as English mass nouns require unitizers like “piece” to specify the unit.

Several studies have provided evidence for the view that only count nouns in mass-count languages lexically specify units of individuation. In one study, using a word extension task, Lucy found that when presented with an entity (e.g., a cardboard box), and asked to judge which of two alternatives was more similar, English speakers preferred a shape-matched choice (e.g., a plastic box) whereas Yucatec Mayans divided their choices between the shape-matched choice and a substance-matched alternative (e.g., a piece of cardboard; see also Lucy & Gaskins, 2001). In a subsequent study, Imai and Gentner (1997) found a similar result in Japanese speakers who were more likely to extend novel words on the basis of substance than on the basis of object kind relative to English speakers.

In more recent work, Huang and colleagues (Huang & Lee, 2009; Huang, 2009) used familiar words to examine noun semantics in Mandarin-speaking adults and children. Using a picture verification task, they found that Mandarin-speaking adults judged sentences containing a bare noun (*yizi* ‘chair’) as acceptable when these nouns were used to refer to either a whole object or just a piece of an object (e.g., *yizi*, or ‘chair’, was acceptable for a whole chair or half of a chair). However, when a sortal classifier was added to the noun (*zhang yizi* ‘a chair’), adults rejected pictures depicting object parts, while children continued to accept them. Based on this finding, they concluded that, first, learning sortal classifiers “initiates children into learning how individuals and non-individuals are encoded in the language” (Huang, 2009: 150), and second, nouns do not have individuated meanings independent of classifiers (see also Borer 2005). Thus, on their view, the combination of a classifier and noun specifies criteria for individuation.

Huang and Lee’s interpretation of these findings is tempered, however, by the fact that many of the nouns they considered to be count in English were in fact syntactically flexible, and could be used as either count or mass in English. For example, the word ‘apple’ in English can refer to either individuals or nonindividuated stuff, depending on syntax (e.g., *some apple* vs. *some apples*). If we assume that noun meanings are the same cross-linguistically, Mandarin speakers might also be willing to accept whole and parts for the flexible nouns in a bare noun phrase because of the different meanings these nouns allow, just as English

speakers might be willing to accept either whole or parts depending on the syntax affixed to the flexible noun.<sup>1</sup>

In support of the view that noun meanings do not differ between mass-count and classifier languages, several recent studies show that count syntax is not necessary for individuation, and that both English mass nouns and bare nouns in classifier languages can specify individuation. First, several researchers have argued that English mass nouns are not limited to denoting non-individuals (e.g., Barner & Snedeker, 2005; Chierchia, 1998). Take, for example, the English mass noun ‘furniture,’ ‘a piece of furniture’ cannot refer to just a leg of a chair, but must denote a whole individual (e.g., a chair). Only ‘a piece of a piece of furniture’ can refer to the leg of a chair. This suggests that mass nouns like ‘furniture’ do provide natural atomic units for counting, namely anything that counts as a “piece” (Doetjes, 1997). This intuition has been supported by experimental studies that probe how mass-count syntax affects quantity judgments. When asked to decide which of two sets contains “more furniture,” participants base quantity judgments on number (e.g., judging that six tiny pieces of furniture are more furniture than two large pieces), despite basing judgments on volume for other mass nouns that denote substances (Barner & Snedeker, 2005). These findings show that English mass nouns can specify individuation, despite lacking overt count syntax.

Moreover, recent studies have found evidence that many nouns in classifier languages also supply criteria for individuation (Barner, Inagaki & Li, 2009; Li, Chen, Barner, & Carey, under review). In the absence of classifiers, speakers of both Japanese and Tsotsil Mayan based quantity judgments on number to the same extent as English speakers for words like ‘cup’, ‘furniture,’ and ‘ketchup’. For mass-count flexible nouns in English such as ‘string’ and ‘apple,’ English speakers quantified by number when the nouns were presented in count syntax (more apples) and by volume when in mass syntax (more apple). Japanese speakers, who received no syntactic cues, made quantity judgments in-between the count and mass groups of English speakers’ judgments, sometimes judging by number and sometimes by volume. This is consistent with the hypothesis that both count and mass readings are available to Japanese and English speakers for flexible nouns, and that syntax selects from universally available lexical meanings.

Subsequent studies have also found that cross-linguistic differences may be much smaller than first reported, are present only when entities are physically ambiguous, and can be made to disappear depending on testing context (e.g., Li, Dunham, & Carey, 2009). Importantly, several studies

have argued that cross-linguistic differences are more likely attributable to lexical statistics rather than to noun semantics. For example, English subjects may be more likely to infer that novel nouns denote object kinds because count nouns are more frequent than mass nouns in English. Speakers of classifier languages, however, need not make such syntactic inferences, and thus may rely more on the physical properties of novel referents to make their judgments, resulting in more variable responding for ambiguous entities (e.g., Imai & Mazuka, 2003; Colunga & Smith, 2006; Li & Gleitman, 2001; Barner et al., 2009).

In summary, a review of recent work on cross-linguistic individuation provides mixed evidence for the claim that, in absence of classifiers, nouns do not specify individuation in languages like Mandarin, Japanese, and Tsotsil Mayan. We believe, however, that the current body of evidence more strongly supports the position that noun semantics are not different cross-linguistically, and that some nouns in classifier languages do provide criteria for individuation just like nouns in mass-count languages. The current study provides even stronger evidence for this position. We assessed how speakers of Mandarin Chinese interpret familiar nouns (Study 1), whether they accept parts of broken objects as units for quantification (Study 2), and whether classifiers change how nouns are interpreted, or are instead semantically inert (Studies 3 and 4).

## Study 1

Using a quantity judgment task (Gathercole, 1985; Barner & Snedeker, 2005), we tested the hypothesis that bare nouns in Mandarin do not individuate unless classifiers are present. We reasoned that, if bare nouns do not individuate in absence of classifier syntax, Mandarin speakers should quantify by volume rather than by number, or quantify randomly across different types of nouns. On the other hand, if nouns can lexically specify individuation, even in absence of classifiers, Mandarin speakers should quantify by number for nouns denoting object kinds (e.g., chair), and by volume for nouns denoting substance kinds (e.g., mustard). For nouns that are used flexibly as either mass or count in English (e.g., string, apple), Mandarin judgments should fall in-between the mass and count judgments, and should vary from one item to the next, depending on the degree to which each word favors an individuated meaning cross-linguistically. To explore this, we tested subjects with two kinds of flexible words – those that continue to apply to a referent in both mass and count forms after the thing has been cut into pieces (e.g., string) and those that can only be used in mass syntax to name the cut referent (e.g., apple).

## Method

**Participants** Fifty-six participants were recruited from universities in Taiwan, with 14 participants assigned to one of the following four noun types (categorized according to their English syntax): count nouns, mass nouns, ‘apple’ type

<sup>1</sup> Other issues such as object functionality arise with Huang and Lee’s study. For example, subjects sometimes noted that the part of a depicted object could still potentially function as a whole individual of that kind (e.g., a torn pair of pants as *kuzi* ‘pants’ could still function as a pair of pants). Thus, it seems likely that results would have differed if they had tested subjects with only translations of English count nouns, using only pictures of clearly non-functional parts.

flexible nouns, and ‘string’ type flexible nouns.<sup>2</sup>

**Procedure** All participants completed a quantity judgment task. They were shown photographs of two characters: one had two large objects or two large portions of substances and the other had four small objects or four small portions of substances. The combined volume of the four small objects or portions was always less than that of the two large objects or portions. Participants were asked to choose which of the two had “more”. Instructions were written in Chinese above the photographs, and all questions were presented without classifiers (*Shui you bijiao duo* [noun]?; Who has more [noun]?).

There were eight nouns for each of the four noun types. For example, the ‘count’ condition included nouns such as ‘bag’ and ‘balloon,’ and the ‘mass’ condition included nouns such as ‘black pepper’ and ‘mustard.’ The two flexible noun lists differed with respect to the salience of their individuated meanings, and in particular, whether their count forms could be used to name pieces of their referents (e.g., half a rock, or half an apple). Eight words satisfied this criterion (e.g., apple, donut), and the remaining eight did not (e.g., rock, string). We will henceforth refer to the first flexible list as ‘Flexible A’ and the second as ‘Flexible B.’<sup>3</sup>

## Results and Discussion

An ANOVA comparing noun types (Count, Flexible A, Flexible B, vs. Mass), with percentage of judgments by number as a dependent variable, found a significant difference across noun types ( $F(3, 52) = 24.88, p < .001, \eta_p^2 = 0.59$ ;  $F(3, 28) = 1444.76, p < .001, \eta_p^2 = 0.99$ ). Pair-wise t-tests by subjects-analysis revealed that judgments based on number were most frequent for count nouns (100%), and least often for substance-mass nouns (0%): Count > Flexible A and Flexible B > Mass (Count vs. Flexible A:  $t(26) = 2.15, p < .05$ ; Flexible A vs. Flexible B:  $t(26) = .74, n.s.$ ; Flexible B vs. Mass:  $t(26) = 4.92, p < .001$ ). Replicating Barner et al. (2009)’s results with Japanese speakers, quantity judgments by number for mass-count flexible nouns were in-between count nouns and mass nouns (Flexible A: 75.0%; Flexible B: 62.5%). These results indicate that Mandarin speakers share the same conceptual distinction on the two kinds of flexible nouns as English speakers. Across languages, the referents of these flexible nouns can be represented either as objects or as the stuff that

forms them. English speakers rely on mass-count syntactic cues when making judgments; however, since Mandarin lacks such cues, speakers relied instead on each referent’s physical properties.

Next, we conducted an items-analysis to examine whether adults responded differently to the two types of flexible nouns. Since participants were more likely to stick to one way of responding throughout the study for flexible nouns, item-analysis was more sensitive to detecting differences across noun types. We found a similar pattern of results (Count > Flexible A > Flexible B > Mass), but the items-analysis also revealed that participants quantified more by number for Flexible A than Flexible B nouns (Count vs. Flexible A:  $t(22) = 13.13, p < .001$ ; Flexible A vs. Flexible B:  $t(14) = 5.58, p < .001$ ; Flexible B vs. Mass:  $t(14) = 53.46, p < .001$ ). Mandarin speakers were slightly more likely to quantify by number for flexible nouns if their English count-noun equivalent only applied to whole referents (e.g., apple, donut), relative to flexible nouns whose English count-noun equivalents applied equally well to a whole object or its parts (e.g., string, rock).

Overall, this set of data suggests that Mandarin noun meanings do not differ fundamentally from nouns in English, and that semantic criteria which predict mass-count usage in English predict the judgments of subjects tested in Mandarin. These data suggest that semantic differences in nouns drive syntactic usage in English, rather than syntax driving the creation of new meanings.

## Study 2

Study 1 provides one form of evidence against the claim that count syntax is necessary for individuation. In Study 2, we sought converging evidence for this claim using a different method. As noted by Huang and Lee (2009), if Mandarin nouns do not specify individuation, then Mandarin speakers should differ from English speakers with respect to how they refer to the parts of broken objects. Previous studies have shown that English-speaking children, unlike adults, often treat parts of objects as units for quantification (e.g., three pieces of a broken fork as being “more forks” than two whole forks; Brooks, Pogue, & Barner, 2011; Shipley & Shepperson, 1990). By some accounts, these failures suggest an inability to use the semantic criteria of nouns to guide quantification. Thus, if Mandarin nouns lack criteria of individuation, then adult speakers should resemble English-speaking children, and should treat pieces of broken objects as units of quantification.

In their study, Huang and Lee (2009) found that Mandarin adults often accepted bare nouns as labels for parts of broken objects. However, as mentioned above, their study included many flexible nouns, whose referents may also be construed as unindividuated by speakers of English when count syntax is not provided (e.g., some apple). In Study 2, we addressed this concern by using nouns that were unambiguously count in English. Also, we varied the syntactic framing of nouns by testing some subjects with

<sup>2</sup> English language is part of the school curriculum in Taiwan, and thus participants in our study would have received training in English. Although proficiency with English can potentially influence participants’ responses, our participants likely do not speak English fluently on a daily basis (see Yeh & Gentner, 2005).

<sup>3</sup> Most nouns were selected using the MacArthur Communicative Development Inventory (Fenson 1994). A group of 13 English speakers provided ratings that corroborated our categorization of whether the noun was a count noun, mass noun, or mass-count flexible noun. Another group of 12 English speakers verified the distinction between flexible A and flexible B nouns. They were asked to judge for each flexible noun whether cutting the thing in question would result in two (“Imagine one [noun]. Imagine that it is cut in half. Are there now two [noun]s?”).

classifiers and some without. If Mandarin nouns do provide criteria of individuation, then Mandarin speakers should behave like their English counterparts and quantify by whole objects regardless of whether a classifier is present. However, if nouns do not provide criteria of individuation, Mandarin speakers should only reliably choose the side with whole objects when a classifier is present.

## Method

**Participants** Twenty-one native Mandarin-speaking adults who had not participated in Study 1 were recruited from universities in Taiwan, and were assigned randomly to one of two conditions.

**Procedure** There were two tasks. In the quantity judgment task, one of the two characters always had two whole objects while the other character had one object cut into three pieces. The objects tested were named by count nouns in English (e.g., cup, ball, shoe), and were a subset of nouns from Study 1. In the counting task, participants saw either three or four objects, one of which was cut into three pieces. They were asked to count the set using a noun (e.g., How many [shoes] are there?) and to give a numerical response. The quantity judgment task was always presented before the counting task.

Participants were tested in Mandarin, and heard instructions containing either a bare noun phrase ( $n=10$ ) or a sortal classifier-noun phrase ( $n=11$ ). In the quantity judgment task, participants were asked, *Shui you bijiao duo* (CL) [noun]? (Who has more (CL) [noun]?). In the counting task, participants were asked, *Zheli you duoshuo* (CL) [noun]? (Here have how-many (CL) [noun]?).

## Results and Discussion

Participants overwhelmingly gave whole object responses in both tasks regardless of whether the sortal classifier was present. For both conditions, responses were near 100% on average for the quantity judgment task and at 100% for the counting task. For the quantity judgment task, there was no significant difference in how often participants gave whole object responses between the classifier (90.9%) and bare noun conditions (100%;  $t(19) = .95$ ,  $p = .35$ ).<sup>4</sup> The finding that adults counted and quantified whole objects in the bare noun condition suggests that judgments were guided by lexical criteria of individuation rather than by classifier syntax. This provides further evidence that nouns in Mandarin do provide criteria of individuation

## Study 3

Studies 1 and 2 show that sortal classifiers are not required to specify individuation for nouns in Mandarin. However, just as English count syntax can disambiguate meanings for flexible nouns, one might expect that sortal classifiers can do the same in Mandarin. To explore this, Study 3 tested

<sup>4</sup> Non-parametric tests using Mann-Whitney U revealed the same pattern of results and found no difference across the two conditions.

subjects using the flexible nouns from Study 1, and manipulated whether the words were presented with classifiers using a quantity judgment task. We predicted that, with the addition of a sortal classifier, Mandarin speakers should unambiguously quantify by number, just as English speakers do when presented with flexible nouns in count syntax.

## Method

**Participants** Sixty-four native Mandarin-speaking participants were recruited in Taiwan as in Study 1.

**Procedure** All participants completed a quantity judgment task. Half of the participants were tested on the Flexible A list, and half on the Flexible B list from Study 1; half of each group was assigned to the bare noun condition and half to the classifier condition, resulting in 16 subjects per group. In the classifier condition nouns were presented with a sortal classifier, whereas nouns in the bare noun condition were not. All else was identical to Study 1.

## Results and Discussion

An ANOVA with Noun Type (Flexible A vs. Flexible B) and Syntactic Frame (Bare vs. Classifier) as between subjects factors found a significant effect of Syntactic Frame ( $F(1,60) = 8.19$ ,  $p < 0.01$ ,  $\eta_p^2 = .120$ ;  $F(1, 14) = 47.15$ ,  $p < 0.001$ ,  $\eta_p^2 = .771$ ). Participants quantified significantly more by number in the classifier condition (85.2%) than in the bare noun condition (62.9%; see Figure 1). For the items-analysis, but not the subjects-analysis, there was a main effect of Noun Type,  $F(1,14) = 12.62$ ,  $p < 0.01$ ,  $\eta_p^2 = 0.47$ . Subjects quantified by number slightly more for the Flexible A list relative to the Flexible B list (80.8% vs. 67.3%). There was no interaction between Syntactic Frame and Noun Type.

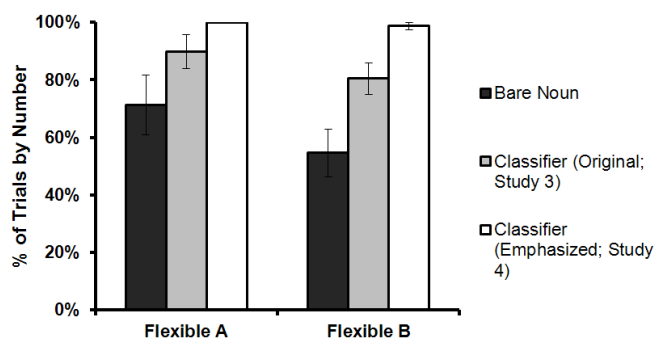


Figure 1. Percentage of judgments by number for flexible A and B nouns across the three conditions - the bare noun and original classifier conditions (from Study 3), and the classifier emphasized condition (from Study 4).

To conclude, we found that classifiers did affect quantity judgment for flexible nouns, leading to more judgments based on number relative to the bare noun condition. However, somewhat surprisingly, the presence of a sortal classifier did not lead participants to choose by number



100% of the time, as one would expect if the classifier were the primary cue for specifying individuation in Mandarin. This finding also suggests that the possibility that adults in our study were implicitly adding the classifier they have come to associate with the noun when making quantity judgments in the bare noun condition seems unlikely.

## Study 4

Although it is possible that participants in Study 3 were relatively insensitive to classifiers when interpreting ambiguous nouns, it is also possible that subjects simply failed to notice their presence when reading the study stimuli. In Study 4 we addressed this concern by underlining the classifiers to emphasize their presence. We expected that when the classifiers are salient to subjects, they should use them to disambiguate the interpretation of flexible nouns.

## Method

**Participants** Twenty additional participants were recruited as in the previous studies.

**Procedure** Procedures were identical to Study 3, with the exception that classifiers were underlined in the written instructions.

## Results and Discussion

With the classifier emphasized, participants now quantified by number 100% of the time for both Noun Types (Flexible A: 100%; Flexible B: 99%; see Figure 1). An ANOVA with Noun Type (Flexible A vs. Flexible B) and Classifier Presentation (Original, Emphasized) as between subjects factors found a significant effect of Classifier Presentation ( $F(1,48) = 7.70, p < 0.01, \eta_p^2 = .138$ ;  $F(1,14) = 65.92, p < 0.001, \eta_p^2 = .825$ ). Participants quantified by number significantly more often when the classifier was underscored (99.4%) than when it was not (85.2%; see Figure 1). No other effects were found by the subjects-analysis. The items-analysis again revealed an effect of Noun Type ( $F(2,14) = 12.62, p < .01, \eta_p^2 = .474$ ). Also, there was a significant interaction between Noun Type and Classifier Presentation in the items-analysis ( $F(2, 14) = 5.23, p < .05, \eta_p^2 = .272$ ). This was driven by the fact that Noun Type only mattered for the original presentation (Study 3) but did not matter for the new presentation (Study 4), since participants were at ceiling in quantifying by number.

Underlining the classifiers led our subjects to quantify by number 100% of the time, suggesting that classifiers can disambiguate between ambiguous noun meanings in Mandarin, much like mass-count syntax in English. However, unlike English subjects who rarely ignore mass-count information, subjects in Study 3 sometimes ignored classifiers when reading instructions. This suggests that classifiers may add little information that is not already provided by the head noun, and typically act mainly as syntactic agreement. In the case of flexible nouns, subjects may draw on their knowledge of the relative frequency of individuated and unindividuated usages – e.g., quickly assuming that ‘apple’ should get an individuated

interpretation because it does most of the time in ordinary speech. Consistent with this idea, recent work on online sentence comprehension in Mandarin suggests that mensural classifiers (i.e., measure words) has a stronger influence on referential selection than sortal classifiers (Klein, Carlson, Li, & Tenenhaus, in press).

## General Discussion

Four studies investigated the claim that bare nouns in Mandarin Chinese do not specify criteria of individuation, and that individuation is introduced by sortal classifiers. Study 1 found that Mandarin speakers do not differ from speakers of English when making quantity judgments for familiar nouns. For example, Mandarin speakers based judgments almost exclusively on number for English count nouns. For mass-count flexible nouns such as ‘apple’ or ‘rock’, Mandarin speakers relied on lexical semantics to determine the units for counting, and made judgments that were roughly between those of English mass and count judgments, suggesting that across languages speakers can access both the individuated and unindividuated interpretations of flexible words. Overall, the current results were similar to those of English and Japanese speakers reported in Barner et al. (2009), and suggest that nouns in Mandarin individuate, despite lacking count syntax, and do so even when classifiers are not explicitly used.

Consistent with this, Study 2 found that Mandarin-speaking adults did not quantify parts of broken objects like English-speaking preschoolers (see Brooks et al., 2011; Shipley & Shepperson, 1990), which provides evidence against the claim that Mandarin nouns lack criteria of individuation. Together, the findings in Studies 1 and 2 suggest that Mandarin noun meanings are no different than noun meanings in English - Mandarin nouns like *yizi* ‘chair’ or *pingguo* ‘apple’ denote kinds of countable individuals.

If individuation is not specified by classifier syntax, what is the role of sortal classifiers in noun phrases? Are classifiers completely inert semantically, or can they sometimes contribute to the compositional semantics of a noun phrase? Findings from Studies 3 and 4 shed light on these questions by testing mass-count flexible nouns. Here, we found an effect of classifier syntax on quantity judgments; participants were more likely to base judgments on number when classifiers were added to flexible nouns. However, adding a classifier did not always have this effect; instead, as shown by Study 4, subjects attended systematically to classifiers only when their presence was made highly salient to subjects. Our suggestion is that subjects typically ignore classifiers because nouns normally provide the relevant content themselves. For adults, lexical meanings supply criteria of individuation, and may be supplemented by knowledge of the relative frequency of different meanings in everyday speech (e.g., the fact that ‘apple’ is frequently used to refer to whole apples rather than to apple-stuff).

The view that lexical semantics can provide the criteria of individuation for nouns in classifier languages such as

Mandarin and Japanese corroborates previous studies in English where English-speaking adults quantify over individuals when nouns such as ‘furniture’ or ‘jewelry’ are used in mass syntax (Barner et al., 2009; Barner & Snedeker, 2005). This suggests that syntax is not the only means that supplies criteria of individuation. In English, the individuation can be expressed syntactically, if the word is used in count syntax (e.g., apples) or through the lexical concept itself (e.g., furniture).<sup>5</sup> In languages that lack count syntax such as Mandarin, however, the lexical concept alone can determine individuation.

To summarize, the current studies provide strong evidence that nouns have similar semantic content cross-linguistically, regardless of variation in their syntactic expression. Nouns in classifier languages such as Mandarin and Japanese encode individuation like nouns in English, and can express this content without requiring the overt use of classifiers. Not only are classifiers unnecessary for individuation in Mandarin, but they also appear to be relatively weak cues to meaning. Unless their presence in a sentence is explicitly highlighted, Mandarin speakers often overlook them, and rely instead on the nouns themselves to determine interpretation. This suggests that in Mandarin, when the meaning of a noun phrase is ambiguous, speakers may rely on other contextual and pragmatic information rather than syntactic cues to disambiguate reference. In contrast, in mass-count languages like English, mass-count syntax often performs this disambiguating function. Outside these ambiguous cases, like string, apple, rock, etc., speakers of classifier languages converge on similar interpretations as speakers of mass-count languages, suggesting that nouns encode individuation in the same way across syntactically diverse languages.

## Acknowledgments

We thank Su-chin (Susan) Shih, Lichun Chang, Alice Fang, Ally Chuang, Paul Chien, Te-Hsin Liu, and Becky Huang, Yaling Hsiao, and Lai Yin Yung for assistance with data collection. Finally, we thank Susan Carey for discussions on experimental manipulations.

## References

- Allan, K. (1980). Nouns and Countability. *Language*, 56, 541–567.
- Barner, D., Inagaki, S., & Li, P. (2009). Language, thought, and real nouns. *Cognition*, 111, 329–344.
- Barner, D., & Snedeker, J. (2005). Quantity judgments and individuation: Evidence that mass nouns count. *Cognition*, 97, 41–46.
- Borer, H. (2005). *In name only*. Oxford: Oxford University Press.
- Brooks, N., Pogue, A., & Barner, D. (2011). Piecing together numerical language: children’s use of default units of quantification. *Developmental Science*, 14, 44–57.
- Chierchia, G. (1998). Plurality of mass nouns and the notion of ‘semantic parameter’. *Events and Grammar*, 70, 53–103.
- Colunga, E., & Smith, L.B. (2003). The emergence of abstract ideas: Evidence from networks and babies. In L. Saitta (Ed.), *Philosophical Transactions by the Royal Society B. Theme Issue: ‘The abstraction paths: from experience to concept’*, 1205–14.
- Doetjes, J. (1997). *Quantifiers and selection: On the distribution of quantifying expressions in French, Dutch and English*. Ph. D. thesis, Leiden University, HAG, The Hague.
- Fenson, L., Dale, P.S., Reznick, J.S., Bates, E., Thal, D., & Pethick, S. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59(5), 1–173.
- Gathercole, V.C. (1985). ‘He has too much hard questions’: The acquisition of the linguistic mass-count distinction in much and many. *Journal of Child Language*, 12, 395–415.
- Huang, A. (2009). *Count-mass distinction and the acquisition of classifiers in Mandarin-speaking children*. Master’s thesis, Chinese University of Hong Kong.
- Huang, A., & Lee, T. H-T. (2009). Quantification and individuation in the acquisition of Chinese classifiers. In Otsu, Y. (Ed.), *Proceedings of the 10th Tokyo Conference on Psycholinguistics* (pp. 117–141). Japan: Keio University.
- Imai, M., & Gentner, D. (1997). A cross-linguistic study on early word meaning. Universal ontology and linguistic influence. *Cognition*, 62, 169–200.
- Imai, M., & Mazuka, R. (2007). Revisiting language universals and linguistic relativity: language-relative construal of individuation constrained by universal ontology. *Cognitive Science*, 31, 385–414.
- Klein, N., Carlson, G.N., Li, R., Jaeger, T.F., Tanenhaus, M.K. (In press). Classifying and massifying incrementally in Chinese language processing. In Diane Massam (Ed.), *A Cross Linguistic Exploration of the Count Mass Distinction*. Oxford University Press. Oxford, England.
- Li, P., Dunham, Y., & Carey, S. (2009). Of substance: The nature of language effects on entity construal. *Cognitive Psychology*, 58, 487–524.
- Li, P., Chen, F., Barner, D., & Carey, S. (under review). Concepts of object and substance kinds: A comparison of speakers of English and of Tsotsil Mayan.
- Lucy, J. (1992). *Grammatical Categories and Cognition*. Glasgow, Scotland: Cambridge University Press.
- Lucy, J. & Gaskins, S. (2001). Grammatical categories and the development of classification preferences: a comparative approach. In S.C. Levinson & M. Bowerman (Eds.), *Language Acquisition and Conceptual Development* (pp. 257–83). Cambridge: Cambridge University Press.
- Shipley, E., & Shepperson, B. (1990). Countable entities: Developmental changes. *Cognition* 34, 109–136.

<sup>5</sup> Interestingly, object-mass nouns such as furniture are often used in count syntax in other languages (e.g., French). This suggests that in languages where nouns are obligatorily marked as mass or count, individuation must be syntactically licensed.

# The Role of Preview and Incremental Delivery on Visual Search

Eric M. Chiu (echiu@ucmerced.edu)

Michael J. Spivey (spivey@ucmerced.edu)

Cognitive and Information Sciences, 5200 North Lake Road,  
Merced, CA 95343 USA

## Abstract

Recent studies show that visual search often involves a combination of both parallel and serial search strategies. Consequently, computational models and theoretical accounts of visual search processing have evolved from traditional parallel or serial descriptions to a continuum from “efficient” to “inefficient.” In our first experiments (1a & 1b), we demonstrate with various conditions that search efficiency does not increase with *simultaneous* delivery of target features in a conjunction-search task. In the second experiment, we explore effects of *incremental* non-linguistic information delivery and discover improvement of search efficiency. We find a facilitatory effect when non-linguistic visual delivery of target features is concurrent with the visual display onset, but not when the target features are delivered prior to display onset. The results support an interactive account of visual perception that explains linguistic and non-linguistic mediation of visual search as chiefly due to the incrementality of target feature delivery once search has begun.

**Keywords:** visual search, incremental, conjunction, efficient

## Introduction

The present study is part of a research program that explores the degree to which the incremental processing of spoken words in a full sentence can interact with concurrent visual search processes.

Traditionally, two contrasting perspectives have plagued the field of attention in visual search. The *serial-processing* perspective claims that observers allocate complete attentional resources discretely and wholly to individual objects, one at a time (Treisman & Gelade, 1980; Treisman, 1988). Conversely, biased competition has been found to be mediated by neural mechanisms in the extrastriate visual cortex, which forms a persuasive line of reasoning not in favor of the serial-processing perspective but for the *parallel-processing* perspective, which claims attention is better characterized as a function of partially active representations of objects simultaneously contending for probabilistic mappings onto motor output (Desimone & Duncan, 1995; Reynolds & Desimone, 2001; Desimone, 1998). Single-feature visual search has been demonstrated to be relatively unaffected by the number of distractors, often inducing a perceptual “pop-out” effect. In contrast, two-feature conjunction-searches typically produce a linear increase in reaction time (RT) as the number of distractors increase. However, as we will demonstrate these apparent dichotomous perspectives may not be from two contrasting

fundamentals but rather a product of a single process better described as a graded enhancement of feature salience and supported by observations of improvement in visual search tasks (Olds, Cowan, & Jolicoeur, 2000a, 2000b, 2000c).

Olds and colleagues (2000a, 2000b, 2000c) observed, in a series of experiments, some facilitatory effects as a result of the very brief duration when search displays had only single-feature distractors. Although observers’ responses were not as fast as with pure “pop-out” displays, Olds and colleagues (2000a, 2000b, 2000c) illustrated a graded improvement of search efficiency, by presenting single-feature visual search pop-out displays for very brief durations (in some conditions less than 100 milliseconds) before transitioning them to conjunction-search displays. Findings like this “search assistance,” along with signal detection theory analyses of visual search data (Eckstein, 1998), and a lack of a bimodal search efficiency distribution (Wolfe, 1998), has replaced the serial-parallel dichotomy account with a continuum of search efficiency (e.g., Nakayama & Joseph, 1998).

Work by Spivey, Tyler, Eberhard, and Tanenhaus (2001) illustrates a different kind of “search assistance” phenomenon. Observers in an *Audio/Visual Concurrent* (A/V-concurrent) condition, where the conjunction-search display is presented concurrently with target identity via auditory linguistic queries (e.g. “Is there a red vertical?”), showed dramatically improved search efficiency compared to an *Auditory-First* control condition, where the same spoken query of target identity was provided prior to visual display onset. The findings suggest that upon hearing the first-mentioned adjective in the spoken query, visual attention is able to begin the search with only that feature, thus initiating the process more efficiently, resembling a single-feature search. Upon hearing the second adjective, several hundred milliseconds later, observers can then quickly find the target among the now attended subset of objects. Additionally, Reali, Spivey, Tyler, and Terranova (2006) implemented quantitative localist attractor simulations to extend the generalizability of the improvement in visual search efficiency when the identity of the conjunction target is delivered incrementally via a spoken target query while the stimulus display is visible, rather than prior to stimulus onset.

Subsequently, Gibson, Eberhard, and Bryant (2005) found with faster speech (4.8 syllables/second vs. 3.0 syllables/second) the A/V-concurrent condition no longer provides an enhanced efficiency effect on conjunction-

search tasks, indicating that improvement in visual search efficiency is affected by speech rate.

Though more recently, experiments by Jones, Kaschak, and Boot (2011) used eye-tracking to examine an alternative view to one that proposes search efficiency is increased due to language enhancing perceptual processing. Jones and colleagues (2011) observed patterns of eye movements suggesting increased efficiency with concurrent speech was not likely due to linguistic enhancement of perceptual processes but instead delaying the onset of target-seeking eye movements. They contend the findings by Gibson et al. (2005) are better explained by this “preview” of search display (when observers are presented with the search display prior to being notified of the target object’s identity) because slower speech provides observers with more search display viewing time, which provides additional information about potential target locations independently of the information conveyed by auditory linguistic speech stream.

The purpose of the present study was to, first, examine the role of preview of search display on visual processing and to, second, further understand exactly how language comprehension and visual search interact in real-time.

### Experiment 1a

In this experiment, we utilized visual cues to deliver simultaneously a two-feature target identity in a conjunction-search task.

#### Method

In this experiment we utilized three SOA, stimulus onset asynchrony, conditions (0-ms, 350-ms, and 750-ms) when identifying the target object. Participants were either presented with the target identifying visual cue simultaneously with the search display (0-ms SOA) or with either a 350-ms or 750-ms delay after onset of search display. All three SOAs appeared equally and randomly.

**Participants** One hundred and fifty-seven University of California, Merced undergraduate students received course credit for participating in this experiment. Participants who were unable to perform the task with an accuracy of 80% or better were removed from the analysis. Twenty-four participants did not meet this requirement thus were removed from the analysis. Additionally, all incorrect responses and trials with RTs greater than 2.5 interquartile ranges (IQR) from the median were also omitted (IQR was used for data culling because of its superior resistance to the influence of outliers).

**Stimuli and Procedure** Each stimulus bar subtended  $2.8^\circ \times 0.4^\circ$  of visual angle and neighboring bars were separated from one another by an average of  $2.0^\circ$  of visual angle. Target identifying visual cues were either red or green horizontal bars that appeared at the top and bottom of the search display or were red or green vertical bars that

appeared on the left and right of the search display. Dimensions of the visual cues were designed to resemble the dimensions of the stimulus objects but four times larger.

The first block was referred to as the “practice” block, consisting of 32 trials, and was followed by an experimental block with 96 trials. Participants were instructed to respond to each display as quickly and accurately as possible by pressing the labeled “YES” button on the keyboard if the target was present in the display and the labeled “NO” button if it was absent.

The target object was present or absent in half of the trials. Moreover, we utilized four set sizes of objects (5, 10, 15, and 20), which appeared equally and randomly. Given two target features (color: red or green, and orientation: vertical or horizontal) four unique targets appeared equally and randomly throughout the trials. The duration of the entire experiment was approximately fifteen minutes. Two 20” Apple iMacs were used to run the experiment. The experiment was programmed and executed using Mathwork’s MATLAB software.

### Results and Discussion

In this experiment we demonstrate with various conditions that search efficiency does not increase in a conjunction-search task when target features are delivered simultaneously, despite having time to preview the search display. The RT-by-set-size functions for target-present trials (filled symbols) are shown in Figure 1 and Figure 2 for target-absent trials (open symbols) in the three SOA conditions, 0-ms (circles), 350-ms (diamonds), and 750-ms (triangles). We should note at this time that RT’s were recorded from display onset, irrespective of condition, until a response was made. Next to each graph line is the best-fit linear equation and the proportion of variance accounted for ( $r^2$ ). Error bars indicate standard error of the mean. In the 0-ms SOA control condition, the RT-by-set-size function was highly linear in both target-present,  $r^2 = .994$ , and target-absent trials,  $r^2 = .984$ , as typically observed in standard conjunction-search tasks. Similarly, the RT-by-set-size functions for the 350-ms and 750-ms SOA conditions were highly linear in target-present trials,  $r^2 = .925$  and  $r^2 = .992$ , and target-absent trials,  $r^2 = .977$  and  $r^2 = .961$ , respectively.

Since our primary interest is to assess the effects of preview on visual search efficiency, analysis in this experiment compared the 350-ms and 750-ms SOA conditions to the 0-ms SOA control condition. Overall mean RTs, as well as y-intercepts, were significantly slower in the 350-ms and 750-ms SOA conditions because delivery of target identity was delayed by 350-ms and 750-ms, respectively, relative to the 0-ms SOA control condition for both target-present,  $t(132) = 2.38$ ,  $p = .017$ , and  $t(132) = 8.21$ ,  $p < .001$ , and for target-absent,  $t(132) = 4.05$ ,  $p < .001$ ,  $t(132) = 9.31$ ,  $p < .001$ , trials. Similar to previous observations, mean accuracy was 94.7% for all three conditions (Spivey et al., 2001; Reali et al., 2006).

For all experiments in this report the most important analysis is in comparison of the slopes of functions relating RT to set size. This slope value is an indicator of how efficient the search process is, that is, how much it resembles a serial process where each new distractor object increases RT by a sizeable fixed duration, or how much it resembles a parallel process where each new distractor object increases RT by little or no amount. The slopes of the RT-by-set-size functions reveal that 350-ms and 750-ms SOA conditions did not produce more efficient visual search compared with the 0-ms SOA control conditions (see fig. 1 & 2). Contrary to findings by Jones et al. (2011) an analysis revealed slopes for the 350-ms and 750-ms SOA conditions compared to the 0-ms SOA control condition were not significantly different for target-present trials (22.4 ms/item & 20.5 ms/item vs. 19.6 ms/item),  $t(132) = 0.61$ ,  $p = .543$ , and  $t(132) = 0.21$ ,  $p = .835$ , and target-absent trials (37.0 ms/item & 35.7 ms/item vs. 41.9 ms/item),  $t(132) = -0.99$ ,  $p = .323$ , and  $t(132) = -1.26$ ,  $p = .207$ .

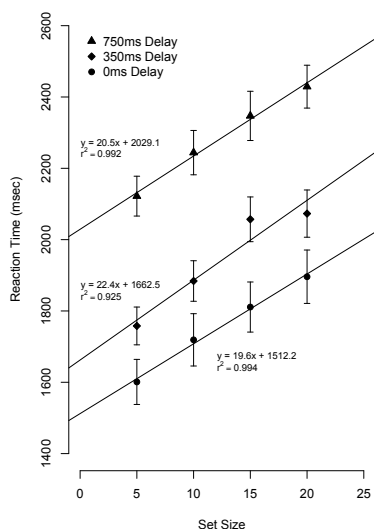


Figure 1. Results from Experiment 1a shown separately for target-present trials.

Similar to Spivey et al. (2001) and Realí et al. (2006), we found a near 2:1 ratio between target-absent and -present trials in all three conditions (37.0-ms/item vs. 22.4 ms/item for 0-ms SOA, 35.7 ms/item vs. 20.5 ms/item for 350-ms SOA, and 41.9 ms/item vs. 19.6 ms/item for 750-ms SOA). This 2:1 ratio between target-absent and -present trials has been regarded as consistent with a standard serial search account.

The results of this experiment indicate that simply delivering target identity simultaneously in a conjunction-search task with a variety of SOAs so that observers are allowed preview time does not substantially affected search efficiency. The results observed in all three conditions are of the type that are traditionally interpreted as consistent with

the construction of a conjunction template of the target object followed by a serial process whereby discretely comparing each display object with the target template (Treisman & Gelade, 1980).

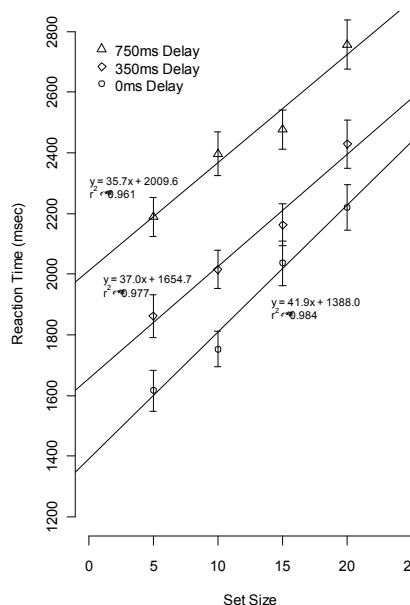


Figure 2. Results from Experiment 1a shown separately for target-absent trials.

## Experiment 1b

In this experiment we extended the methods in Experiment 1a to first, mimic the duration (1500-ms) of the auditory linguistic query, which identified the target object in previous work by Spivey et al. (2001) and to, secondly, explore the effects of a relatively long preview duration of search display on visual search processing.

## Methods

The method in the experiment follows that of Experiment 1a with the exception that only two SOAs (0-ms and 1500-ms) were used for the target identifying visual cue.

**Participants** Fifty-nine University of California, Merced undergraduate students received course credit for participating in this experiment. Five participants were unable to perform the task with an accuracy of 80% or better and were subsequently removed from the analysis. As with Experiment 1a, all incorrect responses and trials with RTs greater than 2.5 IQRs from the median were also omitted.

**Stimuli and Procedure** The same stimuli and target identifying visual cues from Experiment 1a were used in this experiment. Participants were presented with both SOAs equally and randomly in a within-subjects

experimental design. The same testing apparatuses and software were used in this experiment as the last.

## Results and Discussion

As with Experiment 1a, we continue to demonstrate with a slightly different condition that search efficiency does not increase with simultaneous delivery of target feature in a conjunction-search task, despite having time to preview the search display. Figure 3 shows the RT-by-set-size functions for target-present trials (filled symbols) and target-absent trials (open symbols) in the 0-ms SOA (triangles) and 1500-ms SOA (circles). In the 0-ms SOA control condition, the RT-by-set-size function was highly linear in both target-present,  $r^2 = .995$ , and target-absent trials,  $r^2 = .979$ , as typically observed in standard conjunction-search tasks. Similarly, the RT-by-set-size functions for the 1500-ms SOA condition was highly linear in target-present trials,  $r^2 = .975$ , and target-absent trials,  $r^2 = .958$ .

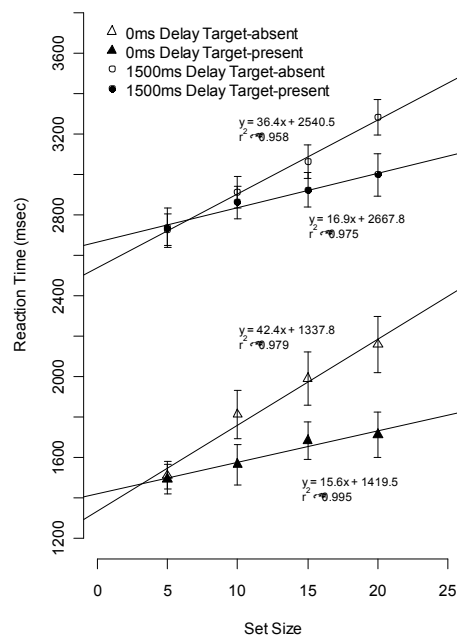


Figure 3. Results for Experiment 1b shown separately for target-present and -absent trials.

Overall mean RTs, as well as y-intercepts, were significantly slower in the 1500-ms SOA condition because delivery of target identity was delayed by 1500-ms relative to the 0-ms SOA control condition for both target-present,  $t(53) = -3.05$ ,  $p = .002$ , and target-absent,  $t(53) = -3.06$ ,  $p < .002$ , trials. Similar to previous observations, mean accuracy was 94.5% for both conditions.

The slopes of the RT-by-set-size functions reveal that the 1500-ms SOA condition did not produce more efficient visual search compared with the 0-ms SOA control

conditions (see fig. 3). An analysis revealed slopes for the 1500-ms SOA condition compared to the 0-ms SOA control condition were not significantly different for target-present trials (16.9 ms/item vs. 15.6 ms/item),  $t(53) = 0.22$ ,  $p = .825$ , and target-absent trials (36.4 ms/item vs. 42.4 ms/item),  $t(53) = -0.85$ ,  $p = .398$ .

Consistent with Experiment 1a, we found a near 2:1 ratio between target-absent and -present trials in both conditions (36.4 ms/item vs. 16.9 ms/item and 42.4 ms/item vs. 15.6 ms/item).

The results of this experiment continue to indicate that simply delivering target identity simultaneously in a conjunction-search task with a relatively long SOA (1500-ms). From Spivey et al. (2001) participants were unable to effectively utilize a noteworthy preview to improve search efficiency. Although a 1500-ms SOA mimics the overall duration of linguistic query it fails to simulate the incremental delivery of information characteristic of speech.

## Experiment 2

In this experiment, we explore effects of incremental non-linguistic information delivery on visual search processing by visually replicating the temporal characteristics of the auditory linguistic query that was used to identify the target object in previous work by Spivey et al. (2001).

## Methods

In this experiment, to simulate the auditory-first and A/V-concurrent condition in Spivey et al. (2001), we utilized two slightly different conditions. A *cue-first* condition similar to the auditory-first condition, delivered target identity incrementally via a visual cue prior to display onset, and a *cue-concurrent* condition similar to the A/V-concurrent condition, delivered target identity incrementally via the identical visual cue but concurrently with display onset. In Spivey et al. (2001) all participants failed to report experiencing any difference in display onset timing, thus auditory-first and A/V-concurrent conditions appeared randomly in a mixed trial design. Since in our experiment the difference between timing of display onset for cue-first and cue-concurrent trials was much more apparent, due to the unimodal nature of the task, we opted for a blocked trial design.

**Participants** Forty-six University of California, Merced undergraduate students received course credit for participating in this experiment. Eight participants were unable to perform the task with a minimum accuracy of 80% and were subsequently removed from the analysis. As with the previous experiments, all incorrect responses and trials with RTs greater than 2.5 IQRs from the median were also omitted from the analysis.

**Stimuli and Procedure** Stimulus objects were identical to experiments 1a and 1b. In order to visually simulate the incremental information delivery of the spoken query (e.g.,

“Is there a red vertical?” 500-ms to utter the first feature color, “red” or “green,” and 1000-ms to utter the second feature orientation, “vertical” or “horizontal”) in Spivey et al. (2001), target identifying visual cues began as all red or all green horizontal and vertical bars that appeared on all sides (top, bottom, left, and right) of the search display for 500-ms to identify the color of the target. To identify the orientation of the target, the visual cue then transitioned to grey horizontal or vertical bars that appeared either at the top and bottom or the left and right of the search display, respectively, for 1000-ms before disappearing. Dimensions of the visual cues were identical to the previous experiments.

Prior to participating in the experimental blocks observers participated in two practice blocks (one of each cue-first and cue-concurrent) consisting of 32 total trials and was followed by two experimental blocks with 64 trials each for a total of 128 trials. One experimental block contained cue-first trials only and the other contained cue-concurrent trials only. The order of the experimental blocks (cue-first first or cue-concurrent first) was randomly assigned to participants, each order was used equally. Participants were instructed to respond to each display as quickly and accurately as possible by pressing the labeled “YES” button on the keyboard if the target was present in the display and the labeled “NO” button if it was absent.

The target object was present or absent in half of the trials. Furthermore, we utilized four set sizes of objects (5, 10, 15, and 20), which appeared equally and randomly. Given two target features (color: red or green, and orientation: vertical or horizontal) four unique targets appeared equally and randomly throughout the trials. The duration of the entire experiment was approximately 20 minutes. Two 20” Apple iMacs were used to run the experiment. The experiment was programmed and executed using Mathwork’s MATLAB software.

## Results and Discussion

In this experiment we demonstrated a facilitatory effect when visual non-linguistic delivery of target features is presented concurrently with the visual display onset, but not when the target features are delivered prior to display onset. Figure 4 shows the RT-by-set-size functions for target-present trials (filled symbols) and target-absent trials (open symbols) in the cue-first (triangles) and cue-concurrent (circles) conditions. In both target-present and -absent trials the RT-by-set-size function was linear for both the cue-concurrent condition,  $r^2 = .768$ ,  $r^2 = .962$ , respectively, and the cue-first condition,  $r^2 = .314$ ,  $r^2 = .698$ , respectively, which is typically observed in standard conjunction-search tasks. Overall mean RT, as well as y-intercepts, were significantly slower in the cue-concurrent condition because delivery of target identity was delayed by 1500-ms relative to the cue-first control condition for both target-present,  $t(37) = 4.49$ ,  $p < .001$ , and target-absent,  $t(37) = -4.32$ ,  $p < .001$ , trials. Mean accuracy was 94.0% for both conditions.

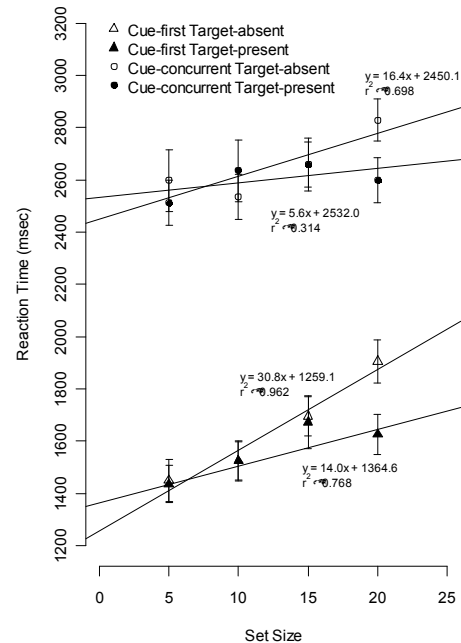


Figure 4. Results from Experiment 2 shown separately for target-present and -absent trials.

The slopes of the RT-by-set-size functions reveal that the cue-concurrent conditions produced more efficient visual search compared with the cue-first control conditions (see fig. 4). An analysis revealed slopes for the cue-concurrent condition compared to the cue-first control condition were significantly different for target-present trials (5.6 ms/item vs. 14.0-ms/item),  $t(37) = -2.77$ ,  $p = .010$ , and target-absent trials (16.4 ms/item vs. 30.8 ms/item),  $t(37) = -2.75$ ,  $p = .006$ . Furthermore, we found a near 2:1 ratio between target-absent and -present trials in both cue-concurrent conditions (16.4 ms/item vs. 5.6 ms/item) and cue-first conditions (30.8 ms/item vs. 14.0-ms/item), regarded as consistent with a standard serial search account.

The results of this experiment indicate that visual non-linguistic delivery of target features presented incrementally and concurrently with the visual display onset has a facilitatory effect on visual search efficiency, but not when the target features are delivered prior to display onset. The results observed in the cue-first condition are of the type that are traditionally interpreted as consistent with the construction of a conjunction template of the target object followed by a serial process of sequentially comparing each display object with the target template (Treisman & Gelade, 1980). Conversely, the results in the cue-concurrent condition, which simply involved shifting the relative timing of display onset relative to target identity cue, are more consistent with a parallel or “partial parallel” (Maioli, Benaglio, Siri, Sosta, & Cappa, 2001) search process, which

is observed in the similarly shallower slopes in the RT-by-set-size functions.

## General Discussion

In a series of experiments we made strides toward understanding exactly how language comprehension and visual search interact in real-time. We demonstrated with various conditions that search efficiency does not increase with simultaneous delivery of target features in a conjunction-search task despite relatively lengthy previews of search display, 1500-ms in some conditions (Experiment 1a & 1b). We then explored the effects of incremental non-linguistic information delivery by visually simulating auditory linguistic queries and discovered an improvement of search efficiency where facilitatory effect only occurred when visual non-linguistic delivery of target features was concurrent with the visual display onset, and not when the target features were delivered prior to display onset.

In conclusion, our findings suggest that it is the incremental nature of target delivery (whether via speech perception or visual perception) that allows the visual search process to begin when only a single feature of the target identity has been heard. When the initial feature is identified the search proceeds in an efficient nearly-parallel fashion so when the second adjective is presented, a substantial amount of the target identification process has already been completed, and as a result the presence of multiple distractors is less disruptive. These results support an interactive account of visual perception that explains linguistic and non-linguistic mediation of visual search as chiefly due to the incrementality of target feature delivery once search has begun. Future research on understanding exactly how language comprehension and visual search interact in real-time will benefit greatly from the development of further experimental tests such as this.

## Acknowledgments

We are grateful to Andreas Kolling for help with the MATLAB code for the experiments. Additional thanks to Monica Yanez, Markie Johnson, Norma Cardona, Mauricio Cifuentes and Courtney Griffin-Oliver for help with collecting data.

## References

- Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 353(1373), 1245.
- Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1), 193–222.
- Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attention processing. *Psychological Science*, 9, 111–118.
- Gibson, B. S., Eberhard, K. M., & Bryant, T. A. (2005). Linguistically mediated visual search: The critical role of speech rate. *Psychonomic Bulletin and Review*, 12(2), 276.
- Jones, J. J., Kaschak, M. P., & Boot, W. R. (2011). Language mediated visual search: The role of display preview. *Cognitive Science Proceedings*, 2739–2744.
- Maioli, C., Benaglio, I., Siri, S., Sosta, K., & Cappa, S. (2001). The integration of parallel and serial processing mechanisms in visual search: Evidence from eye movement recording. *European Journal of Neuroscience*, 13(2), 364–372.
- Nakayama, K., & Joseph, J. S. (1998). Attention, pattern recognition, and pop-out in visual search. *The attentive brain*, 279–298.
- Olds, E. S., Cowan, W. B., & Jolicoeur, P. (2000a). Partial orientation pop-out helps difficult search for orientation. *Perception & psychophysics*, 62(7), 1341–1347.
- Olds, E. S., Cowan, W. B., & Jolicoeur, P. (2000b). The time-course of pop-out search. *Vision Research*, 40(8), 891–912.
- Olds, E. S., Cowan, W. B., & Jolicoeur, P. (2000c). Tracking visual search over space and time. *Psychonomic Bulletin and Review*, 7(2), 292–300.
- Real, F., Spivey, M. J., Tyler, M. J., & Terranova, J. (2006). Inefficient conjunction-search made efficient by concurrent spoken delivery of target identity. *Perception and Psychophysics*, 68(6), 959.
- Reynolds, J., & Desimone, R. (2001). Neural mechanisms of attentional selection. *Visual attention and cortical circuits (Braun J, Koch C, Davis JL, eds)*, 121–136.
- Spivey, M. J., Tyler, M. J., Eberhard, K. M., & Tanenhaus, M. K. (2001). Linguistically mediated visual search. *Psychological Science*, 12(4), 282–286.
- Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95(1), 15–48.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, 12(1), 97–136.
- Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science*, 9, 33–39.



# Determining Relevance: Close Enough is Good Enough

Sheldon J. Chow (schow55@uwo.ca)

Department of Philosophy, Stevenson Hall, The University of Western Ontario  
London, Ontario, N6A 5B8 Canada

## Abstract

Humans exhibit characteristic success in considering what is relevant in their cognitive tasks. Yet understanding how relevance is determined in cognition remains a problem. This paper seeks to make headway on this problem. The relevance problem is first introduced. Sperber and Wilson's influential theory of relevance is then discussed, but dismissed as inadequate. Some conditions are identified that an adequate definition of relevance might reasonably be expected to satisfy. A novel way to conceive of relevance is suggested which proves to be useful in understanding human cognitive performance.

**Keywords:** Relevance; concepts; file model; cognitive architecture.

## The Relevance Problem

A longstanding problem in philosophy and cognitive science is understanding how we determine what is relevant to our cognitive tasks. The cognitive systems paradigmatically responsible for general reasoning and decision-making—so-called central systems—admittedly allow for free exchange of information. A dream of a snake biting its own tail, for example, bore in interesting and important ways on Kekulé's theorizing of the benzene molecule. A consequence of such a free exchange of information, however, is that, provided an appropriate set of background beliefs, any representation held by an agent can *in principle* be relevant to a given cognitive task in central cognition. Who won tonight's football game is *prima facie* irrelevant to whether there is beer in your friend's fridge. But if you believe that your friend's favourite football team played tonight, and that your friend usually overindulges in beer consumption whenever his favourite team wins, then your belief about who won tonight's game actually *is* relevant to your belief about beer in your friend's fridge. Since relevance can be determined only with respect to one's background beliefs, there is no way to circumscribe *a priori* the subset of representations that are relevant in a given occasion of reasoning or inference. Let us express this problem thus:

**The relevance problem** How a cognitive system considers only what is (mostly) relevant, or equivalently, how a cognitive system knows what is relevant.

Despite the fact that the relevance problem introduces a computational problem of sifting through heaps of information to decide what bears on the task at hand, humans seem to determine what is relevant in their cognitive tasks quickly and rather easily. This is not to say that we always know and consider what is relevant. For we often fail to do so, especially

when cognitive demands are high and/or when cognitive resources are low. Nevertheless, humans characteristically exhibit reasonable levels of success at identifying representations which are relevant to the task at hand. Such reasonable levels of success cannot be ascribed to chance or luck. Therefore, we are left to explain how humans (seem to) solve the relevance problem, short of considering the totality of one's beliefs (Fodor, 1987, 2000).

However, the inquiry into how relevance is determined is constrained by how relevance is defined. Indeed, whether and the extent to which relevant representations are picked out and brought to bear on a cognitive task will depend on what property we are concerned with. Yet, defining relevance is not an easy task. In this paper I provide a cursory overview of what is arguably the most influential account of relevance to date, namely Sperber and Wilson's *Relevance Theory*. I will then be able to use their Relevance Theory as a basis from which to propose an understanding of relevance that is supported by a view of concepts and cognition which draws on current theories in cognitive science, as well as Fred Dretske's information-theoretic epistemology.<sup>1</sup> This will allow me to show how relevance, or something like it, can be understood as naturally arising from human cognitive architecture, thus enabling the characteristic performance we observe in human reasoning.

## Sperber and Wilson's Relevance Theory

Sperber and Wilson (1986/1995) developed their Relevance Theory in the context of communication and pragmatics. Humans tend to have an easy time communicating with each other, despite the fact that the meanings of utterances are enormously underdetermined. A simple example: Alice says to Bob, "Isn't that cute?" while nodding toward a chipmunk scurrying up a tree; Bob knows that by "that" Alice is referring to the chipmunk, and not to the birds in the other branches, the tree itself, or whatever else was within his perceptual field at the time of her utterance. According to Sperber and Wilson (SW henceforth), Bob understands that Alice was referring to the chipmunk because the stimulus of the

<sup>1</sup> It should be kept in mind throughout the discussion that I am not intending to develop a full theory of relevance. I therefore do not offer any claim of completeness. Such a task deserves a more extensive treatment than what I can provide in this short paper. My contribution might be understood as a useful preliminary discussion of how the problem of relevance might be stated, and an identification of some conditions that an adequate definition of relevance might reasonably be expected to satisfy. Nevertheless, I shall suggest a potential way to conceive of relevance which proves to be useful in understanding human cognitive performance.

chipmunk running up the tree was *relevant* (or at least more so than any other present stimulus).

Although SW's Relevance Theory is mainly concerned with verbal/ostensive communication and comprehension in particular, they claim that their theory can be extended to the inference processes of ordinary thinking (1986/1995, p. 67; p. 75). This is because SW ground their account of relevance in a fundamental and general view of human cognition. Specifically, SW claims that cognition always tends toward efficiency (i.e., maximizing gains and minimizing costs), and furthermore that human cognition succeeds in increasing its efficiency by having a tendency toward *maximizing relevance*. According to their Relevance Theory, the relevance of an input to a process is a function of *cognitive effect* on the one hand, and *processing effort* on the other.<sup>2</sup>

It should be obvious that relevance, assessed in terms of cognitive effect and processing effort, comes in degrees. In addition, some input may yield greater cognitive effects on some occasions and less effects on others, or, depending on circumstances related to fatigue or stress, the same input may be more or less easy to process at different times. Thus, relevance is a relative property—relative to an individual and to a time. SW therefore provide the following two definitive conditions for relevance:

1. *Ceteris paribus*, the greater the cognitive effects achieved by an individual by processing some input, the more relevant that input is to that individual at that time.
2. *Ceteris paribus*, the greater the effort expended by an individual by processing some input, the less relevant that input is to that individual at that time.

When SW claim that human cognition has a tendency toward maximizing relevance—which they assert as The Cognitive Principle of Relevance<sup>3</sup>—they mean that human cognition is geared toward allocating cognitive resources to processing available inputs (from the environment or from memory) so as to maximize expected cognitive effects for the least expected effort.

Sperber and Wilson (1995) make the point that an individual is not interested in cognitive effects *per se*, but only insofar as such cognitive effects contribute to achieving certain cognitive goals, or otherwise fulfilling its functions. For

---

<sup>2</sup>SW develop their account in terms of constructing an appropriate context within which to process inputs. A context is simply just a set of assumptions within which input can be processed. For simplicity, I will pass over this aspect of their account; no harm is done to the central points of the present paper.

<sup>3</sup>SW's Relevance Theory makes claims about human cognition in general, but an important consequence of their Cognitive Principle, and one that is the basis for their work in pragmatics, is the Communicative Principle of Relevance: *Every ostensive stimulus conveys a presumption of its own optimal relevance*. This Communicative Principle of Relevance is really the centerpiece of SW's Relevance Theory. However, since communication and pragmatics are not the focus of this paper, the Communicative Principle will not be discussed any further.

indeed, there may be cognitive effects that are not worth having, or that contribute negatively to the individual's cognitive functions or goals. Thus, SW define a *positive cognitive effect* as a cognitive effect that contributes to the cognitive goals or functions of the individual in a positive way. This is typically achieved by alterations in the individual's beliefs. In SW's words, positive cognitive effects produce an "epistemic improvement" (Sperber & Wilson, 1995, p. 266); they make a "worthwhile difference" to the cognitive functioning of the individual (Wilson & Sperber, 2006, p. 608).

Thus far I have given a brief outline of SW's Relevance Theory. Nevertheless, there are a number of ways in which Relevance Theory is inadequate. For brevity, I mention here only two inadequacies, but these are most damaging. First, it is grossly unclear how to quantify cognitive effect and effort so as to make sense of the ratio between the two in SW's definition of relevance. Second, SW's conception of cognition is strictly in terms of deduction—processes performed by a "deductive device" that is supposed to "model the system used by human beings in spontaneous inference" (Sperber & Wilson, 1986/1995, p. 94). Of course, there is wide consensus now that spontaneous inference is not strictly deductive, but is to a large extent non-demonstrative. SW's account of cognition completely excludes such non-demonstrative cognition, as well as other aspects of cognition such as perception.<sup>4</sup>

In light of these shortcomings, the remainder of this paper is devoted to offering the beginnings of a more adequate account of relevance in cognition. I will begin by outlining a psychologically plausible theory of cognition. We shall see that this will allow a more naturalistic understanding of SW's notion of "positive cognitive effects", but more importantly it has the potential to deliver a desired account of relevance.

## Concepts and Cognition

A number of cognitive scientists and philosophers have converged on an idea about the nature and function of mental representations (e.g., Evans, 1982; Fodor, 2008; Lawlor, 2001). The idea is a *file model* of cognition (Kahneman, Treisman, & Gibbs, 1992; Pylyshyn, 2003; Treisman, 1982; Treisman & Schmidt, 1982). There is no standard doctrine accepted by everyone who endorses the file model, but there is some common ground that can be identified.

According to the file model of cognition, one has a mental "file" for each object that one has beliefs about. Each mental file has a "label" that both picks out the file, and refers to the

---

<sup>4</sup>By the second edition of *Relevance* (1995), Sperber and Wilson had disavowed such a central systems architecture in favour of the massive modularity hypothesis. Though I note this only in passing, I believe that certain complications arise for their Relevance Theory within a massively modular architecture. This is a matter to be discussed on some other occasion, however. Sperber (2005) has recently suggested that cognitive efficiency, in terms of maximizing relevance, is achieved biologically—specifically, by optimally allocating energy in the brain. His proposal is to think of maximizing cognitive effects and minimizing effort in terms of noncognitive physiological processes. However, this entails an account of relevance entirely different from SW's Relevance Theory, and Sperber has not developed such an account in any detail.

object that is associated with the file. Further, each file holds a number of “notes” or “memos” that contain various kinds of information or beliefs, or representations more generally, about the given object.

It is important for the present purposes to understand that the file model can be conceived more particularly to be a theory of the structure of *concepts*. On this reading, a concept names a file that contains representations about the things in the concept’s extension. When thinking about cats, for instance, one calls upon one’s CAT file, which may contain such representations as *is a living thing, is a furry creature, is a quadruped, is grandma’s favourite animal*, etc.<sup>5</sup> Notice now that what is contained in the file—what notes there are in the file—name other concepts: LIVING THING, FURRY CREATURE, QUADRUPED, GRANDMA’S FAVOURITE ANIMAL, etc. This is not classical decomposition. Rather, concept files contain notes that convey information about the concept in question. Hence, unpacking the notes of one’s CAT concept file is unpacking one’s beliefs and other stored knowledge about cats, the representations of which invoke further concepts. Moreover, whatever is inferred from the unpacked representations may excite further concepts, thereby bringing to mind files and representational content of their own. Thus, such a conceptual system forms a vast network among the beliefs and other representations which are contained in concept-files.

A very important feature of this account is that concepts do not exist independently of one another. Rather, this is a view of concepts wherein vast connections exist between them in virtue of their content. Every representation is a part of an interconnected network, where activation can spread through the system depending on the strengths and the organization of the connections between representations in the network. In principle, the structure allows for any content to activate any other content, which is a hallmark of central cognition (cf. Viger, 2006b).

We might therefore see that the file model of cognition allows for the following sort of picture to underwrite thought and inference. When one is presented with a given cognitive task, the nature and content of the task initially activates and primes<sup>6</sup> a focused set of concepts and contents. Suppose, for instance, one is presented with the task of estimating whether it is likely or unlikely that there will be a plane crash within the month. This task would activate conceptual content concerning *planes, crashes, likelihoods, estimation, months, timeframes*, and more. Certainly, various parameters will constrain what conceptual content gets activated. Some parameters include those that are given by the language used (e.g., using “plane” rather than “jet”), as well as those that are suggested by the nature of the task (e.g., the task elicits a course-grained subjective likelihood assignment to a future event as opposed to, say, a fine-grained numerical subjective

probability). Other parameters will have to do with factors affecting long-term memory recall, such as which concepts an individual possesses, the relative strengths at which conceptual content is stored in long-term memory, the ease with which conceptual content is activated (perhaps based on past activations), and the existence and relative strength of connections between concepts and conceptual content established by past inferences. In addition to these constraints, limits on time and cognitive resources will restrict what and how many conceptualizations occur. Yet, even with these constraints, a considerable amount of conceptual content may still get activated.

This has been a brief overview of the file model of cognition. Before I continue, I should note that the file model is best viewed as a useful analogy for thinking about concepts and the relations among them and their contents. The model is advanced here merely as a gesture toward a possible cognitive conceptual architecture. If the file model is to become a viable basis for a definition of relevance, it will need much further development and theoretical refinement—a task that is beyond the scope of this paper. Nevertheless, the main proposal here does not depend on the truth of the file model, specifically. It requires only that there exist highly organized systems of knowledge or representations in cognition which bear certain connections to one another. The file model gives us a tentative idea of what these systems might be. Yet, that a number of cognitive scientists have converged on it to model various aspects of cognition is suggestive of its plausibility.<sup>7</sup>

## Reinterpreting Relevance

Let us now turn back to relevance. There is something intuitively right about the idea that relevance has something to do with what SW call “positive cognitive effects”. For indeed, it seems as if processing irrelevant information would not generally yield anything positive for the cognitive system. Yet, we would need a more precise understanding of positive cognitive effect which avoids the pitfalls of SW’s account, but which is somehow tied to relevance. The account of concepts and cognition in the previous section delivers this.

To see this, let us begin by drawing some lessons from Fred Dretske (1981). In developing his information-theoretic epistemology, Dretske commented that there is “a *relativization* of the information contained in a signal because *how much*

<sup>5</sup>The convention I adopt here is to represent concepts in small caps, contents in italics, and uses in quotes.

<sup>6</sup>For simplicity, I will often use “active”/“activated” to refer to both active/activated and primed content.

<sup>7</sup>Moreover, the file model lends itself to being interpreted within certain current neurological theories of conceptual cognition. For example, Barsalou’s (1999) *perceptual symbols systems* view of cognition, Patterson, Nestor, and Rogers’ (2007) theory of a *semantic hub*, and Damasio and Damasio’s theory of *convergence zones* (A. R. Damasio, 1989; H. Damasio, Tranel, Grabowski, Adolphs, & Damasio, 2004; Tranel, Damasio, & Damasio, 1997), each posits the existence of neurological sites that store and enact a “code” which specifies specific neural patterns to be (re)activated during mental representation. My preferred (albeit tentative) interpretation is to understand concept-files to be roughly analogous to the posited sites where the codes are stored; opening files can then be understood as instantiating the appropriate codes to (re)activate neural patterns for representation; and file labels can be understood as lexical terms that name concepts in natural language. (This last point was suggested to me by Chris Viger.)

information a signal contains, and hence *what* information it carries, depends on what the potential receiver already knows about the various possibilities that exist at the source” (p. 79). Although we might not know (or even cannot know) absolute measures associated with the amount of information generated by an event or carried by a signal, Dretske believes that *comparisons* can be made, “in particular comparisons between the amount of information generated by the occurrence of an event and the amount of information a signal carries about that event” (p. 54).

According to Dretske, our concepts are cognitive structures that enable us to extract and exploit certain information in the environment. Learning or enriching a concept provides us with the ability to encode (or digitalize) certain aspects of our (analog) sensory experience in such a way that we are able to cognitively respond in certain ways which we would not have been able to otherwise. Having the concept DAFFODIL, for example, enables one to see a daffodil *as* a daffodil (as opposed to a flower, or some green and white object), and thus allows one to have daffodil-thoughts and to cognize daffodil-stimuli in certain ways. Importantly, then, what information one can encode from stimuli crucially depends on what one already knows about the objects of the stimuli, or in the terms of the present account, on what and the way in which conceptual contents are coded in one’s conceptual architecture.<sup>8</sup>

A natural account of relevance follows from this picture in terms of the amount of information received from a source (such as a stimulus). More specifically, the greater the amount of information received, the greater the relevance of that information. For example, suppose that Alice and Bob are on a nature walk. Alice is a botanist, whereas Bob never cared for plant science. As Alice and Bob gaze upon the flora of the forest floor, they both cognitively extract a vast amount of information from their respective perceptual scenes. However, Alice’s conceptual knowledge is so rich that she is able to extract more specialized information than Bob does or even can, having to do with the various kinds of plants that they come across. In this way, the perceptual scene carries more information for Alice than for Bob. Of course, they both process the *same* information, but Alice can *cognitively extract* more information. Whereas Bob simply sees a plant, Alice sees Blindwood ivy; whereas Bob simply sees flowers, Alice sees daffodils. Importantly, certain information in Alice’s and Bob’s perceptual scene is very *relevant* to Alice but not so relevant to Bob. And the information from the perceptual

scene that Alice finds relevant is just that information she is able to extract via her conceptual knowledge. On the other hand, such information is not as relevant to Bob because he cannot represent the information in the same ways, since he lacks the conceptual wherewithal to do so.<sup>9</sup>

Therefore, I suggest that the relevance of a stimulus to a given cognitive system (or agent) depends on the amount of information received from that stimulus. Since the amount of information received depends on one’s conceptual wherewithal to attend to and code specific information in certain ways, whether and the extent to which something is relevant is dependent on the informational content of one’s concepts. But this is not the entire story, since relevance will also depend on the context and cognitive task. Suppose, for instance, that both Alice and Bob are botanists, but Alice is interested in finding a rare flower while Bob is interested in seeing a specific species of ivy. Both Alice and Bob could code the same information in the same ways, but because of their different goals and cognitive tasks, Alice will find flower information more relevant than Bob, and Bob will find ivy information more relevant than Alice.

This can be easily accounted for, however, once it is understood that, in setting up one’s goals and preparing for one’s cognitive task, one requisitely activates a number of concepts with specific conceptualizations and representations, as described above. This will in a sense serve as a filtering mechanism for focusing attention. Alice will thus be (cognitively/conceptually) geared to attend to specific information related to a specific flower whereas Bob will be (cognitively/conceptually) geared to attend to specific information related to ivies, as each will have prepared a set of concepts and representations upon embarking on their respective cognitive tasks. The set of concepts and representations that gets activated when one prepares for a given cognitive task will tend to be relevant, although this may not always be the case. What gets activated will depend on previous experience, past activations, and the extant relations and connections among the activations. These relations will constrain and guide inference.

Thus, on this view, the degree to which information is relevant is a matter of the *informativeness* of information. In other words, relevance is conceived to be a measure of how much information gets encoded given one’s cognitive wherewithal and preparation (i.e., the activations and connective relations among one’s conceptual representations). The same information can be more or less informative, and hence rele-

<sup>8</sup>According to information theory, and thus according to Dretske, a signal carries 0 information if one already knows the message the signal carries. I disagree, since for a signal to carry 0 information, one would have to be unable to extract and conceptualize any information from the source in question (cf. Gabbay & Woods, 2003). And such a thing does not seem to be likely, or even possible, in human cognitive practice. Processing information one already knows may be *redundant*, but it may also be useful to strengthen or reaffirm one’s beliefs. A corollary of this view is that a signal will always carry information for human cognitive systems; the *amount* of information in a signal will depend on what the receiver knows, or more specifically, on activated conceptual representations. More on this presently.

<sup>9</sup>We might consider here a situation in which Bob is trying to learn some botany, in which case the information received from the stimuli given by the forest floor would seem to be more relevant to Bob than Alice, since Alice already knows what Blindwood ivy is and looks like, etc. However, this situation is a shift in context from the scenario described in the main text. Specifically, Bob’s cognitive task has changed, and therefore so has his cognitive preparation. The difference in cognitive preparation, along with, say, Alice’s instruction, will entail a difference in the kinds and amounts of information extracted from the environment, and hence a difference in relevance. See below.

vant, depending on the agent and cognitive task. Understood this way, relevance is not just a property of information *per se*—not just a matter of what information gets processed—but a matter of *how information gets processed*.

If this is right, then positive cognitive effect can be understood in terms of the informativeness of information that is processed—processing information yields positive cognitive effects insofar as the results are informative. But this is just to refer to the degree of relevance of the information in question. This means that positive cognitive effect is yielded by processing relevant information. We therefore have our intuitive connection between positive cognitive effect and relevance borne out by the present account: Processed information has positive cognitive effect *because* it is informative; and the degree to which it is informative is the degree to which the information is relevant. Cognitive effects and effort therefore do not define relevance. Rather, concepts and the extant relations between their contents will facilitate cognitive effects and effort in processing information. Hence, it is structured concepts that deliver relevance, which in turn produces cognitive effects with little effort.

This account of relevance is definitely an improvement over SW's account. On the one hand, there is no problem of how to quantify cognitive effect and effort to make sense of the ratio between the two; instead, relevance is measured directly in terms of degree of informativeness. On the other hand, there is no conceiving cognition strictly in terms of deduction; the adopted view of cognition is amenable to deductive and non-demonstrative inference alike.

## How Do We Determine What is Relevant?

We are now in a position to see how the foregoing account of concepts and cognition can help to explain our characteristic levels of success in our reasoning. Given the present account, we have more constraints on our reasoning than we may know. Specifically, a certain kind of relevance is determined by the extant relations among the contents of our concepts. The kind of relevance I have in mind is *de facto* relevance (cf. Gabbay & Woods, 2003), in which information appears to be relevant due to the architectural characteristics of cognition. More specifically, and to continue the line of reasoning in the previous section, something is (more or less) *de facto* relevant if it appears to be (more or less) informative when processed against a given set of activated concepts. In this way, *de facto* relevance cannot be determined *a priori*, as should be expected. Instead, it simply arises out of the nature and structures of our concepts.

The *de facto* relevance established by the extant relations within and between activated conceptual contents appears to be enough for humans to get by on. The kind of relevance that matters to the relevance problem spelled out in the first section of this paper, on the other hand, is a kind of *objective* relevance which exists independently of cognizers. Many examples of objective relevance come from science. For example, the motions of terrestrial objects are relevant to the

motions of the planets, and this is an objective fact, but we did not know this until Newton came along.<sup>10</sup>

There will certainly be times when we fail to process objectively relevant information, or when we process information that is not very objectively relevant at all. In some cases we may end up processing some objectively *irrelevant* information. Moreover, inevitably there will be cases in which we fail to process *de facto* relevant information, due to cognitive limitations, fatigue, stress, or some other extraneous factor. Under satisfactory conditions, however, our activated concepts, with their contents and extant relations, provide a network that informs cognition of what is *de facto* relevant, and constrains and guides its processing accordingly. The situation may not be ideal, but it is good enough for us to get by on—indeed, such is to be expected from satisficing organisms. On the other hand, when we enter into certain high-stake arenas, such as science or philosophy, we alter our standards, and *de facto* relevance is no longer good enough. In such circumstances, objective relevance is sought, and this is likely why progress and getting things right are much more difficult to achieve in these endeavours.

At the same time, however, we might understand the foregoing account of concepts and cognition as contributing to how humans manage to solve, not the relevance problem stated above, but a more fundamental problem. To see what I mean, let us consider Daniel Dennett's (1984) example of fixing a midnight snack. He noticed that such a mundane task requires copious amounts of knowledge: "We know trillions of things; we know that mayonnaise doesn't dissolve knives on contact, that a slice of bread is smaller than Mount Everest, that opening the refrigerator doesn't cause a nuclear holocaust in the kitchen" (p. 136). Dennett also noticed that we must possess a system of representing this knowledge in such a way that it is accessible on demand. This would require a system that is organized in such a way that achieves the efficient representation and access we observe in humans. In his words: "A walking encyclopedia will walk over a cliff, for all its knowledge of cliffs and the effects of gravity, unless it is designed in such a fashion that it can find the right bits of knowledge at the right times, so it can plan its engagements with the real world" (pp. 140-1). From these considerations, we might say that humans solve a *representational relevance problem*:

**The representational relevance problem** How a cognitive system embodies the informational organization, and enables access to the relevant information, that seems to be required for human-like cognitive performance.

I believe that the present account of concepts and cogni-

<sup>10</sup>Notice that objective relevance also cannot be determined *a priori*; but rather than against a set of beliefs, relevance is determined against a background of facts and phenomena. The motions of terrestrial objects, for instance, are not objectively relevant *simpliciter*; it is objectively relevant with respect certain phenomena, such as planetary motion.

tion is precisely what enables humans to solve the representational relevance problem. For as I have illustrated, concepts are organized in such a way that the extant relations between their contents facilitates access to *de facto* relevant information. Such information may not be objectively relevant, but it will almost certainly be the kind of information that is needed to guide successful action, and this is all that is needed for human-like performance.

It is interesting to note, however, that it seems that much of the information that is *de facto* relevant turns out to be objectively relevant a lot of the time. This is evident from how humans get on in the world, and the success rate of many human inferential endeavours. I can only speculate why this is so: it is likely an outcome of some evolutionary process. This would explain our reasonable levels of success in bringing to bear objectively relevant information on our cognitive tasks. I admit that this is not a deep explanation. However, if we conceive our conceptual system to have evolved to track *things in the world*, then it should not be much of a mystery why our conceptual wherewithal reflects the organized structure of information in the world, including objective relevance relations. In this way, then, we can conceive *de facto* relevance to be built up by systems that track objective relevance. And, just like any cognitive system that tracks stuff in the world, sometimes things work out and sometimes things go awry; and sometimes cognitive systems track truths but not all the time (such as the perceptual systems; cf. Viger, 2006a). It seems, however, that *in the main* cognition tracks truths in the world, and is quite good at it. Thus, the *de facto* relevance embodied by the relations within and between concepts by and large reflects objective relevance in the world. This is what makes *de facto* relevance close enough to objective relevance; and close enough is good enough.

### Acknowledgments

Thanks to Chris Viger for all his comments and suggestions. Thanks also to the anonymous referees whose comments have improved this paper.

### References

- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-660.
- Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33, 25-62.
- Damasio, H., Tranel, D., Grabowski, T. J., Adolphs, R., & Damasio, A. R. (2004). Neural systems behind word and concept retrieval. *Cognition*, 92, 179-229.
- Dennett, D. C. (1984). Cognitive wheels: The frame problem of AI. In C. Hookway (Ed.), *Minds, machines, and evolution* (p. 129-152). Cambridge: Cambridge University Press.
- Dretske, F. I. (1981). *Knowledge and the flow of information*. Cambridge, MA.: The MIT Press.
- Evans, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.
- Fodor, J. A. (1987). Modules, frames, fridgeons, sleeping dogs and the music of the spheres. In J. L. Garfield (Ed.), *Modularity in knowledge, representation and natural-language understanding* (p. 25-36). The MIT Press.
- Fodor, J. A. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. Cambridge: The MIT Press.
- Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. Oxford: Clarendon Press.
- Gabbay, D., & Woods, J. (2003). *A practical logic of cognitive systems, Volume 1. Agenda relevance: A study in formal pragmatics*. Amsterdam: Elsevier.
- Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, 24, 175-219.
- Lawlor, K. (2001). *New thoughts about old things: Cognitive policies as the ground of singular concepts*. New York: Garland Publishing.
- Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8, 976-987.
- Pylyshyn, Z. W. (2003). *Seeing and visualizing: It's not what you think*. Cambridge: The MIT Press.
- Sperber, D. (2005). Modularity and relevance: How can a massively modular mind be flexible and context-sensitive? In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Structure and contents* (p. 53-68). Oxford: Oxford University Press.
- Sperber, D., & Wilson, D. (1986/1995). *Relevance: Communication and cognition*. Oxford: Blackwell (2nd edition, 1995).
- Sperber, D., & Wilson, D. (1995). Postface to the second edition of *Relevance: Communication and Cognition*.
- Tranel, D., Damasio, H., & Damasio, A. R. (1997). A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia*, 35, 1319-1327.
- Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance*, 8, 194-214.
- Treisman, A., & Schmidt, H. (1982). Illusory conjunction in the perception of objects. *Cognitive Psychology*, 14, 107-142.
- Viger, C. (2006a). Is the aim of perception to provide accurate representations? a case for the "no" side. In R. J. Stainton (Ed.), *Contemporary debates in cognitive science* (p. 275-288). Malden, MA: Blackwell Publishing.
- Viger, C. (2006b). Symbols: What cognition requires of representationalism. *Protosociology: The International Journal of Interdisciplinary Research*, 22, 40-59.
- Wilson, D., & Sperber, D. (2006). Relevance theory. In L. R. Horn & G. Ward (Eds.), *The handbook of pragmatics* (p. 607-632). Malden, MA: Blackwell Publishing.

# Subject Relative Production in SLI Children during Syntactic Priming and Sentence Repetition

**Moreno I. Coco**  
([mcoco@staffmail.ed.ac.uk](mailto:mcoco@staffmail.ed.ac.uk))  
School of Informatics (ILCC)  
University of Edinburgh  
10 Crichton Street  
Edinburgh, EH8 9AB

**Maria Garraffa**  
([mgarraff@staffmail.ed.ac.uk](mailto:mgarraff@staffmail.ed.ac.uk))  
School of Psychology  
University of Edinburgh  
7 George Square  
Edinburgh, EH8 9JZ

**Holly P. Branigan**  
([holly.branigan@ed.ac.uk](mailto:holly.branigan@ed.ac.uk))  
School of Psychology,  
University of Edinburgh  
7 George Square  
Edinburgh, EH8 9JZ

## Abstract

Children with Specific Language Impairment (SLIC) experience difficulties in processing Subject relative clauses (SRC). This has been interpreted as evidence that they lack syntactic representations for SRC. Our study investigates the spontaneous production of SRC in typically developing children (TDC) and SLIC in a structural priming paradigm, and compares their performance in a sentence repetition task. We demonstrate that SLIC are much more likely to produce SRC during priming than in sentence repetition; moreover, when primed, their performance matches TDC's baseline (unprimed) performance. Furthermore, we design two simple unsupervised Bayesian models, and predict the developmental group (SLI, TD) and priming condition (Primed, Non-Primed). Overall, this study shows that SLIC can spontaneously produce SRC when primed, suggesting their impairment is related to working memory, rather than a deficit in syntactic knowledge.

**Keywords:** specific language impairment; language development; syntactic priming; sentence repetition; Bayesian data analysis.

## Introduction

Subject relative clauses (SRC) such as *the cat that's on the table* are generally early acquired, at around 3 years in typically developing children (TDC; e.g., Crain et al. 1990). However, children with Specific Language Impairment (SLIC) display difficulties in producing subject (and object) relative clauses (Novogrodsky & Friedmann, 2006). Preschool SLIC show a delayed onset of relative clause production, and frequent omission of the complementizer in both elicitation and spontaneous production (Contemori & Garraffa, 2010). This difficulty extends to repetition of sentences involving SRCs. This is particularly interesting, because recent research has suggested that in TDC, prior exposure to even difficult structures can facilitate their subsequent production (e.g., Bencini & Valian 2008). Such effects have been identified as manifestations of syntactic priming, whereby an abstract syntactic representation is facilitated (Bock, 1986). In adults, syntactic priming appears to be implicated in sentence repetition (Potter & Lombardi, 1998). It is therefore striking that SLIC do not show facilitated production of SRC in sentence repetition, as we would expect a benefit from a syntactic priming effect, enhanced by lexical repetition.

Previous research has therefore proposed that SLIC do not have a syntactic representation of SRC (Conti-Ramsden et al., 2001). In this paper, we consider an alternative hypothesis, namely that SLIC's poor performance in sentence repetition does not reflect a lack of syntactic knowledge, but rather a task-specific difficulty, which may be related to working

memory demands. We investigate this hypothesis by comparing SLIC's and TDC's production of SRCs in tasks where they are explicitly elicited (in a sentence repetition task) and when they are implicitly elicited (in a picture-description syntactic priming paradigm). We use Bayesian Data Analysis to investigate our hypothesis, and design a series of Bayesian models to tackle it.

## Experiment

Substantial research has used a syntactic priming paradigm to demonstrate that people use abstract syntactic representations to process language (Bock, 1986; Pickering & Branigan, 1998). In such research, speakers show an increased tendency to use a particular structure after previously encountering the same structure (even with different lexical content). These effects have been argued to provide evidence about syntactic representation (Branigan et al., 1995). Recent research on TDC has therefore used syntactic priming to provide evidence for the early acquisition of TDC's syntactic representations, e.g., passive constructions (e.g., Bencini & Valian 2008): If children are more likely to produce a particular structure after previous exposure to it, it implies that they have an abstract representation for that structure, which can be facilitated through residual activation or implicit learning (Chang et al., 2006; Pickering & Branigan, 1998). The former may explain short term priming effects; the latter may explain long term and cumulative priming effects. We argue that a syntactic priming paradigm can similarly be used to examine whether SLIC have access to an abstract representation of SRC, whose availability can be incremented through priming.

To do this, we used a Snap priming paradigm (Branigan et al., 2005), in which SLIC and TDC engaged in a card game that involved three elements: 1) listening to the experimenter describe a picture (using either a simple noun or an SRC), 2) describing their own picture, 3) and deciding whether or not the two pictures matched<sup>1</sup>.

In the analysis, we assess whether the two groups (TDC, SLIC) differ in their production of SRC (Model 1); and test if any difference depends on the priming condition (i.e., whether the experimenter's description involved a simple noun or an SRC; Model 2). We expect TDC to have a higher production of SRC than SLIC. However, we expect SLIC to perform better in the primed than in the non-primed condi-

<sup>1</sup>On three quarters of trials, the pictures did not match, and on one quarter - Snap trials - they did match

tion. If so, this would provide evidence that SLIC have a syntactic representation of SRCs, whose retrieval is facilitated by prior exposure. This in turn would imply that SLIC's observed impairments in SRC production have a different source than lack of syntactic knowledge of SRCs.

We then compare the two groups' performance on Snap trials only (i.e., where the child's picture matched the experimenter's picture, and hence the child could repeat the experimenter's description verbatim), with their performance in a task where they were explicitly asked to repeat the experimenter's sentence <sup>2</sup> (Model 3). Thus we compare children's production of SRC sentences, when the children produce those sentences freely in response to a picture stimulus (priming) and when they produce them as part of an explicit repetition task (repetition). If their impairment is fundamentally syntactic, then SLIC are expected to be equally bad on both tasks, and the TDC equally good on both tasks. If instead the impairment has a non-syntactic source, then children may show differential performance on the two tasks. In particular, if SLIC have a working memory impairment, they may find it difficult to produce SRC in a task where they must retain in working memory the meaning to be communicated (as well as its syntactic form).

We then investigate cumulative priming, i.e., whether the likelihood of producing an SRC increases as a function of priming instances (Kaschak et al., 2011), which is assumed to reflect implicit learning. Given previous evidence that SLIC have impairments in implicit learning (Tomblin et al., 2007), we should observe no such cumulative effect benefit in SLIC.

Finally, we move on to possible applications of our study, and investigate the problem of classification, and the related issue of diagnostic statistics to assess impairment (here, SLI). We use Bayesian inference to built two classifiers: in the first (Model 5), we categorize the development group (SLI, TD) based on the number of SR produced; and in the second (Model 6), we use the same measure to categorize development group and priming (Primed, Non-Primed). The retrieval performance of the two Bayesian models is compared with logistic regression classifiers, binomial (Model 5), and multinomial (Model 6) on F-score, an aggregate measure of precision and recall.

## Method

We compared SLICs (and control TDCs) production in a syntactic priming paradigm using a picture description SNAP task. Thirty-eight (19 SLI, 19 TD) pre-school monolingual Italian children participated (mean = 5.4 years, Non-verbal IQ > 92). SLIC were selected on the basis of their general comprehension and expressive abilities, measured on MLU and on performance on standardized language tests (expressive abilities: Frog story; receptive lexicon: PPVT; receptive grammar: TCGB).

Children described target pictures after hearing the experimenter describe a prime picture with an SRC (*a cat that's on a wall*) or a simple noun (*a cat* in a within-participants manipulation (priming). There were 24 prime/target pairs (where the experimenter's and child's pictures, and hence necessarily

their descriptions, were different), and 8 'snap' pairs (where the experimenter's and child's pictures were identical, and they could therefore use identical descriptions).

The same groups of children took also part in a sentence repetition task, where they had to repeat, verbatim, 10 SRC sentences produced by the experimenter. We compare their performance in this task with their performance on the 'snap' trials (8 per participant), in which they described the same picture as the experimenter had just described (and hence could repeat the experimenter's utterance verbatim).

## Data Analysis

We adopt a Bayesian approach for the analysis of this study, with the aim of uncovering the parameters responsible for generating the observed data, (e.g., the performance of the two populations of children), infer their distribution, and have an estimate of the uncertainty in our hypothesis (Kruschke, 2010).

In the next sections, we describe four Bayesian models used to test our hypotheses. Then, we apply Bayesian analysis on a classification task and infer, in an unsupervised way, developmental group (SLIC, TDC) and priming condition (Primed, Non Primed) using the number of produced SRC as our dependent measure. We compare the performance of these two Bayesian classifiers with logistic regression models, which are trained on the same set of data. We report F-scores to assess the overall performance of the models, and precision and recall for the different classes to evaluate more in depth the classification performance.

## Models

All Bayesian analyses presented here are performed using OpenBugs as implemented by the R packages BRugs and R2WinBugs (Sturtz et al., 2009; Kruschke, 2011).

**Model 1: Group analysis** The first model estimates how different the two groups of children (TDC, SLIC) are in producing SRC. We model the occurrence of an SRC production through a Bernoulli process (e.g., coin toss) with a given probability of success, whereby the total number of occurrences in a series of trials follows a binomial distribution. For each group of children, the number of SRC produced is denoted by the vectors  $x$  and  $y$ , which represent independent samples from two different binomial distributions, in a total of  $N$  trials ( $N = 24$ ), with respective probabilities  $p$  and  $q$  underlying SR production. Each group consists of  $S$  observations, equal to the number of children in that group. The index  $i$  in  $x_i$  and  $y_i$  refers to subject  $i$  in each group. The likelihoods of the data are given by:

$$P(x_1, x_2, \dots, x_S | p) = \prod_{i=1}^S \binom{N}{x_i} p^{x_i} (1-p)^{N-x_i}, \quad (1)$$

in the case of the TDC group, and an analogous equation for the SLI group by replacing  $p$  by  $q$ . As prior knowledge, we assume that both  $p$  and  $q$  are independently drawn from  $\sim \text{Beta}[1, 1]$ . This allows us to independently estimate the underlying production rate for TDC and SLIC.

<sup>2</sup>Conducted on the same groups of children.



**Model 2: Effects of priming** In the second model, we compare the effect of priming on the production of SRC for the two groups of children. Here, our dependent measure is the number of SRC produced for each trial, i.e., we aggregated over participants. Every trial can belong to one of the two Priming states (Primed, Not Primed). So, each observation  $n_t, (t = 1, \dots, N)$  is the number of SRC produced in trial  $t$ , to which a particular state of priming  $c_t$  is associated:  $c_t = 1$  refers to a Non-Primed trial, and  $c_t = 2$  to a Primed one. Since trials are independent, and we know their priming status, given by the vector  $c$ , the likelihood of the performance of a group can be computed as a product over trials:

$$P(n_1, n_2, \dots, n_N | p, c) = \prod_{t=1}^N \binom{S}{n_t} p(c_t)^{n_t} (1 - p(c_t))^{S-n_t} \quad (2)$$

We assume a uniform prior for  $p(1)$ , and the relation  $p(2) = p(1) + \theta$ , where the difference parameter  $\theta$ , is also given a uniform prior,  $\theta \sim \text{Uniform}[0, (1 - p(1))]$ . We apply this model, independently to SLIC and TDC. In the former case,  $n$  is calculated from the performance of the SLIC; in the latter case,  $n$  is obtained from the TDC group. This model allows us to infer 4 parameters.

**Model 3: Sentence Repetition vs Priming** In this third model, adapted from Model 2, we explore more closely the role of working memory on the production of SRC by comparing SRC repetition on SNAP trials in the priming task<sup>3</sup>, where participants could repeat the experimenter’s sentence, with SRC elicited through sentence repetition.

Every trial can belong to one of the two tasks (Repetition, Priming). So, each observation  $n_t, (t = 1, \dots, N)$  is the number of SRC produced in trial  $t$  in Repetition ( $c_t = 1, N = 10$ ) or Priming ( $c_t = 2, N = 8$ ). As for Model 2, we assume a uniform prior for  $p(1)$ , and the relation  $p(2) = p(1) + \theta$ , where the difference parameter  $\theta$ , is also given a uniform prior,  $\theta \sim \text{Uniform}[0, (1 - p(1))]$ . This model is applied independently to SLIC and TDC.

**Model 4: Cumulative priming** Finally, in our fourth model, we quantify the rate at which cumulative priming occurs for the two groups of children. For this model, we represent the data as two matrices  $X_{S \times N}$  (TDC) and  $Y_{S \times N}$  (SLI), where each entry, for example  $x_{it} \in \{0, 1\}$  is a Bernoulli random variable denoting production or not of a relative clause for each subject ( $i = 1, \dots, S$ ), in each particular trial ( $t = 1, \dots, N$ ). We can calculate the likelihood of the data from each group, by using the following expression:

$$P(X | p_0, a, k) = \prod_{i=1}^S \prod_{t=1}^N p(t)^{x_{it}} (1 - p(t))^{1-x_{it}} \quad (3)$$

In this model, the probability of producing an SRC is given by  $p(t) = p_0 + ak(t)$ , and it increases linearly with the number of priming trials  $k(t)$  that have occurred up to trial  $t$ .

<sup>3</sup>Note that we exclude this data from any other model of priming.

Table 1: Observed production performances (in percentage) for the two groups on the repetition and priming task (also divided by Priming condition)

Group	Repetition	Priming SNAP	Primed	Non-Primed
TDC	90%	56%	57%	19%
SLIC	16%	39%	24%	0.9%

In analogy, for the SLI group we have  $q(t) = q_0 + bk(t)$ . We must estimate 4 parameters:  $p_0, a, q_0, b$ . As priors, we use non-informative uniform priors for the intercepts, e.g.  $p_0 \sim \text{Uniform}(0, 1)$ , and normal distributions for the slopes, e.g.,  $a \sim \text{Normal}(0, 1)$ . We expect to find a lower rate of cumulative priming for SLIC than for TDC ( $a > b$ ), due to their impairment of implicit learning.

All models are run for 50000 iterations with a thinning of 10. If the chains have not converged, and are highly auto-correlated, we keep updating until convergence. We discard the first 5000 iterations of the Monte Carlo Markov Chains (MCMC) sample as burn-in, to calculate the posterior distribution of the parameter values. A summary of the results is reported in Table 2.

## Results

Before going into the modeling results, we report the observed performance on the priming task, as well as on the repetition task for the two groups (refer to Table 1).

It is clear that during the repetition task SLIC are much more impaired than TDC in producing SRC. However, their production rates improves when we look at repetition during the priming (SNAP task) trials, where SLIC are twice as good compared to the sentence repetition result. When looking at the performance within the Priming task, we see an effect of priming on both TDC and SLIC. The production rate of SRC is more than doubled in Priming trials compared to the Non-Primed condition. These results are largely confirmed in our inferential analysis.

As expected, in Model 1 we find that  $p > q$ : TDC are more likely to produce SRC than SLIC (refer to Table 2). This result indicates that SLIC experience more difficulties, in fact twice more so ( $p/q \approx 2$ ), in producing SRC than TDC. However, in order to understand whether structural priming is actually increasing the production rate of SRC in SLIC, and quantify any difference with TDC, we included the priming variable in Model 2.

Here, we find that TDC are more likely to produce SRC, especially in Primed trials. Crucially, however, we also find a strong effect of Priming in SLIC. In fact, we observe that SLIC are 20% more likely to produce an SRC when primed than when not primed. Moreover, we see that the probability of producing an SRC by a SLIC who is primed becomes indistinguishable, if not higher, than that of a TDC who is not primed. This result demonstrates that the impairment displayed by SLIC cannot be attributed to a lack of syntactic knowledge, but rather to some other aspect of cognition. In fact, when the relevant structural representations are facilitated through syntactic priming, SLIC can spontaneously pro-

duce relatively complex structures such as SRC. To examine the possible role that working memory might play in SLIC's previously observed impairment in SRC production, in Model 3, we compared syntactic priming with a task that explicitly taps working memory, namely sentence repetition.

The result of Model 3 clearly shows that the SLIC perform very differently in the two tasks, Snap and repetition (see Table 2). This is especially evident when looking at the difference in probability between the two tasks: SLIC are much more likely to produce an SRC during the Snap task than during the repetition task, whilst for TDC, the difference  $p(2) - p(1)$  is almost negligible.

This result is intriguing as it implies that given a task in which the child has to actively retrieve syntactic material to communicate a contextually supported proposition (provided by the visual stimulus), their impairment is much less pronounced. This pattern is consistent with a working memory impairment in which SLIC children have difficulty in building, maintaining, and retrieving a representation of an entire sentence. In order to better explore the role of working memory on the production rate, we turn to cumulative priming. If implicit learning, which is hypothesized to underlie long term and cumulative priming effects, is impaired in SLIC, then we might observe a weaker, or almost null, effect of cumulative priming in that group.

Interestingly, the results of Model 4 show that the intercept parameter  $p_0$ , which indicates the baseline probability for TDC to produce an SRC, is markedly lower than the value observed in the previous model (refer to Table 2 for the list of models). The intercept  $q_0$ , instead, is reasonably similar in value across the two models. When we look at the slope parameters ( $a, b$ ), we observe a different rate for the two groups of children. In particular, TDC experience more cumulative priming in producing SRC, compared to SLI children. This result indicates that although cumulative priming occurs in both groups, this effect is more prominent in TDC. Moreover, it seems that cumulative priming plays a crucial role on the likelihood of producing SRC, as it emerges when comparing the estimates of this model with Model 1. The diminished effect of cumulative priming in SLIC is in keeping with evidence that SLIC have impaired implicit learning. At the intercepts, in fact, the two groups display a similar probability of producing SRC, and the difference becomes more prominent when previous exposure to SRC increases.

In the next section, we shift the focus from quantification to prediction, and apply Bayesian analysis to make inferences about our data. In particular, we implement two Bayesian classifiers to predict development group and priming in an unsupervised way, i.e., we assume there is no prior knowledge about such classification, and we want to infer it from the data. In order to validate the performance of such models, we compare the classification performance of the two Bayesian models with the output of logistic classifiers. We report the F-score, which is a measure of test's accuracy based on the weighted average of precision and recall:  $F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$ . Precision is the fraction of retrieved instances that are relevant, e.g., the number of correctly categorized children in a certain group,  $\frac{tp}{tp+fp}$ , and recall is the

Table 2: Mean and 95% Highest Density Intervals for the posterior distributions of parameters for the different models.

Model 1: Groups		
Parameter	Mean	95%HDI
$p_{TD}$	0.4655	(0.4195, 0.5116)
$q_{SLI}$	0.2490	(0.2104, 0.2900)
Model 2: Priming		
Parameter	Mean	95%HDI
$p_{TDPrimed}$	0.6392	(0.5747, 0.7018)
$p_{TDNotPrimed}$	0.2961	(0.2397, 0.3577)
$q_{SLIPrimed}$	0.3521	(0.2924, 0.4153)
$q_{SLINotPrimed}$	0.1479	(0.1052, 0.1958)
Model 3: Task Comparison		
Parameter	Mean	95%HDI
$p_{TDSnap}$	0.7639	(0.7165, 0.8086)
$p_{TDRepetition}$	0.7575	(0.7086, 0.8021)
$q_{SLISnap}$	0.4020	(0.3216, 0.4844)
$q_{SLIRepetition}$	0.1672	(0.1052, 0.1958)
Model 4: Cumulative Priming		
Parameter	Mean	95%HDI
$p_{0InterceptTD}$	0.2702	(0.1926, 0.3566)
$q_{0InterceptSLI}$	0.2085	(0.1344, 0.2870)
$a_{SlopeTD}$	0.0357	(0.0226, 0.048)
$b_{SlopeSLI}$	0.0035	(-0.0049, 0.0199)
Model 5: Group Inference		
Parameter	Mean	95%HDI
$p_{TD}$	0.4494	(0.4030, 0.5060)
$q_{SLI}$	0.2034	(0.1429, 0.2743)
Model 6: Group and Priming Inference		
Parameter	Mean	95%HDI
$p_{TDPrimed}$	0.2707	(0.2274, 0.3216)
$p_{TDNotPrimed}$	0.2143	(0.1467, 0.2828)
$q_{SLIPrimed}$	0.1549	(0.1019, 0.2316)
$q_{SLINotPrimed}$	0.1130	(0.084, 0.1413)

fraction of relevant instances that have been retrieved,  $\frac{tp}{tp+fn}$ , e.g., the correctly categorized children off the total number of possible children in that group; where  $tp$  (true positive) is the number of instances correctly retrieved,  $fp$  (false positive) are the instances wrongly retrieved and  $fn$  those instances which should have also been retrieved. We report precision and recall for each category to distinguish what the model found easier to classify, from what found it more difficult.

### Inference of developmental categories and priming

In the first Bayesian classifier (Model 5), our goal is to predict the development group of a child (SLI, TD) by considering blindly their SR production scores. Therefore, we don't assume a-priori that there is a division between TD and SLI, and mix the data between the two different groups. Our task is to infer correctly which data belongs to each category. We implement an extension of Model 1 as follows.

To each observation  $z_i$ , ( $i = 1, \dots, 2S$ ), corresponding to the

number of SR clauses produced by all subjects of the experiment, we assign a hidden state  $c_i \in \{1, 2\}$ , which refers to the subject being an SLI (1) or a TD (2). If the hidden state of each subject is known, the total likelihood of the data can be computed as:

$$P(z_1, z_2, \dots | c_1, c_2, \dots) = \prod_{i=1}^{2S} \binom{N}{z_i} p(c_i)^{z_i} (1 - p(c_i))^{1-z_i} \quad (4)$$

where  $p(1)$  is the probability of producing a relative clause in the SLI group, and  $p(2)$  of the TD group. This corresponds to a mixture model, where we infer the two distributions present in the data set (estimate  $p$ ) and the probabilities  $\pi_i(1)$  and  $\pi_i(2)$  that each observation  $z_i$  belongs to distribution 1 or 2 (estimate  $c$ ). We assume a Dirichlet prior for the category assignment and a uniform distribution for  $p$ , similar to previous models.

The final model (Model 6) is an extension of Model 5, where, beside the development group, we want to infer also the priming condition (Primed, Non Primed). So, the only real difference is that we assume four categories  $c_i \in \{1, 2, 3, 4\}$  where 1 is SLI-Non priming, 2 is SLI-Priming, 3 is TD-Non Priming and 4 is TD-Priming. Again, the model is blind to these categories, and our goal is to infer them from the data. We proceed as in Model 5 and we obtain a posterior (discrete) distribution for the categorical classification of each observation, namely what is the probability that an observation comes from category 1, 2, 3 or 4. The most likely category for each observation can be obtained by considering the median of such a posterior distribution, i.e. the category that obtains a probability higher than 1/2 for that observation.

**Model performance** We compare the categorization performance of these two unsupervised models against two fully supervised logistic classifiers, Binomial (BL) for Model 5, and multinomial (ML) for model 6. The logistic classifiers are built using the R libraries `glm` and `mlogit`. We first fit a generalized linear model, binomial with logit link, and a multinomial regression to obtain the regression coefficients. Our dependent measure is the number of SRC produced and the independent predictor is the category: group for the binomial logistic classifier, group and priming for the multinomial classifier.

From the regression coefficients, we derive the logits, which exponentiated give us the unscaled probabilities of observing a certain category, e.g., SLI, associated to a certain SR production score. The unscaled probabilities are normalized to range between 0 and 1<sup>4</sup>.

For the sake of completeness, we compare the estimates of the parameters for these two models, where categories have to be inferred, with those where such information was explicitly modeled (Model 1 and Model 2); see Table 2 for the actual values. We observe that the estimates are close between

<sup>4</sup>More details can be found on the online tutorial provided by UCLA: Academic Technology Services, <http://www.ats.ucla.edu/stat/r/dae/mlogit.htm>

Table 3: Classification Performance of the Bayesian Models and the Logistic classifiers: F-scores, precision (left value) and recall (right value) for the different categories (TD and SLI): where P (Primed) and N-P (Non Primed)

(a) Model 5 vs Binomial Classifier

Model	F-score	Categories	
		TD	SLI
Model 5	0.85	(0.8, 0.84)	(0.83, 0.78)
BL	0.85	(0.8, 0.84)	(0.83, 0.78)

(b) Model 6 vs Multinomial Classifier

Model	F-score	SLI N-P	SLI P	TD N-P	TD P
Model 6	0.55	(0.76, 0.88)	(0.42, 0.15)	(0.35, 0.26)	(0.53, 0.94)
ML	0.56	(0.65, 0.78)	(0.36, 0.47)	(0.4, 0.21)	(0.83, 0.78)

Model 1 and Model 5; whereas there is a more marked difference between Model 2 and Model 6. As the number of categories to be inferred from the data increases, the task becomes more challenging.

When comparing the classification performance, we find that Model 5 and BL are perfectly equivalent, both in terms of F-score, and on the precision and recall of the group category. Obviously, the main outstanding difference is that the Bayesian model is unsupervised, while the logistic classifier is fully supervised. Moreover, it is interesting to notice that SLIC and TDC are recognizable as separate populations, rather than being outliers of the same distribution.

On the multi-class task, in contrast, we find that ML is slightly better than Model 6 on the F-score. When looking at precision and recall for the different classes, we find both models achieving higher classification performances on the two most extreme classes (SLI Non Primed and TD Primed), with Model 6 having a better recall than ML, which instead has a better precision. It is interesting to notice that both models fail to account for the two categories SLI Primed and TD Non Primed. The reason is that SLIC Primed perform as well as TDC Non primed. This result confirms further our main hypothesis that SLIC have syntactic knowledge of SRC, and can achieve performance comparable to TDC when the relevant representations are facilitated through syntactic priming.

## General Discussion

Typically developing children are able to process SRC around the age of 3 years (Crain et al., 1990). Yet children with SLI show difficulties with these structures at the same age (Novogrodsky & Friedmann, 2006). In particular, the finding that SLIC cannot repeat verbatim SRC sentences produced by an experimenter has been taken as evidence that SLIC do not have a syntactic representation of relative clauses (Conti-Ramsden et al., 2001). To challenge this claim, we compared SLIC and TDC's production of SRCs in two different tasks, one of which implicitly elicited repetition of SRC sentences (syntactic priming in a 'Snap' task) and one of which explic-

itly elicited repetition of SRC sentences (sentence repetition task), in a series of Bayesian Models. We then used such models to infer developmental group and priming condition on a trial-by-trial basis by looking at the number of SRCs produced.

We found that although TDC display a higher production of SRC than SLIC (Model 1), SLIC nevertheless are more likely to produce an SRC when primed, i.e., after hearing an SRC (with different lexical content) than after hearing a simple noun (Model 2). This suggests that SLIC have an abstract representation of SRC that they can recruit during production, when it has been facilitated through previous use. Crucially, however, SLIC performed worse in a sentence repetition task than in 'snap' priming trials in which they could repeat verbatim the experimenter's sentence (Model 3). In other words, the same children who performed poorly in a task that required explicit repetition of sentences performed significantly better in a priming task, where implicit repetition could be realized through incremental production of a sentence and did not require the (meaning and structure of the) entire sentence to be maintained in working memory. This suggests that SLIC's impairment on these structures is related to working memory. Furthermore, we found that whereas TDC showed a large cumulative priming effect (i.e., were much more likely to produce an SRC after several exposures), SLIC showed a markedly reduced, if not negligible, cumulative effect (Model 4).

Taken together, our results suggest that SLIC's poor performance on SRCs may reflect a working memory impairment that affects on-line processing, rather than the absence of a syntactic representation for SRCs (i.e., a deficit in syntactic knowledge). In a task that implicates all stages of language production, and in which production can occur incrementally without the necessity of retaining in working memory a representation of the entire sentence, SLIC are able to spontaneously produce SRCs after being exposed to them; furthermore, their spontaneous production of SRCs in this context is less impaired than their elicited repetition of SRCs, relative to TDC. However, the finding that SLIC show reduced cumulative priming suggests that they may also have impaired implicit learning, which may have far-reaching implications for their acquisition of language. The classification performance of our Bayesian models demonstrates moreover that we can infer the developmental group of a child very accurately (Model 5) with the same accuracy as a Binomial Logistic Classifier, with the clear advantage that the Bayesian model is unsupervised. Crucially, our classification performance degrades when trying to infer the group and the Priming condition (Model 6), because SLIC Primed perform as well as TDC Non-Primed.

In conclusion, we suggest that modeling differences in syntactic performance in syntactic priming and other experimental paradigms has great potential for investigating the nature of impairments in syntactic representations versus other aspects of cognition in SLIC and other atypical populations.

## References

- Bencini, G. M. L. & Valian, V. (2008). Abstract sentence representation in 3-year-olds: Evidence from comprehension and production. *Journal of Memory and Language*, 59, 97–113.
- Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18, 355–387.
- Branigan, H., McLean, J., & Jones, M. (2005). A blue cat or a cat that is blue? Evidence for abstract syntax in young children's noun phrases. In Brugos, A., Clark-Cotton, M., & Ha, S. (Eds.), *The proceedings of the 29th annual Boston University conference on language development*, (pp. 109–121).
- Branigan, H., Pickering, M., Liversedge, S., Stewart, A., & Urbach, T. (1995). Syntactic priming: Investigating the mental representation of language. *Journal of Psycholinguistic Research*, 24, 489–506.
- Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, 113, 234–272.
- Contemori, C. & Garraffa, M. (2010). Comparison of modalities in SLI syntax: A study on the comprehension and production of non-canonical sentences. *Lingua*, 120, 1940–1955.
- Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for SLI. *Journal of Child Psychology and Psychiatry*, 42, 741–748.
- Crain, S., McKee, C., & Emiliani, M. (1990). Visiting relatives in Italy. In J. deVilliers & L. Frazier (Eds.), *Language Processing and Language Acquisition* (pp. 335–356). Dordrecht: Reidel.
- Kaschak, M., Kutta, T., & Jones, J. (2011). Structural priming as implicit learning: Cumulative priming effects and individual differences. *Psychonomic Bulletin and Review*, 18, 1133–1139.
- Kruschke, J. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, 14, 293–300.
- Kruschke, J. (2011). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Englewood Cliffs, NJ: Academic Press: Elsevier.
- Novogrodsky, R. & Friedmann, N. (2006). The production of relative clauses in SLI: A window to the nature of the impairment. *Advances in Speech-Language pathology*, 8, 364–375.
- Pickering, M. J. & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, 39, 633–651.
- Potter, M. C. & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language*, 38, 265–282.
- Sturtz, S., Ligges, U., & Gelman, A. (2009). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, 12, 1–16.
- Tomblin, J. B., Mainela-Arnold, E., & Zhang, X. (2007). Procedural learning in adolescents with and without specific language impairment. *Language Learning and Development*, 3, 269–293.

# Segmenting Visual Narratives: Evidence for Constituent Structure in Comics

**Neil Cohn (neilcohn@emaki.net), Phillip Holcomb**

Department of Psychology, 490 Boston Ave  
Medford, MA 02155 USA

**Ray Jackendoff**

Center for Cognitive Studies, Miner Hall  
Medford, MA 02155 USA

**Gina Kuperberg**

Department of Psychology, 490 Boston Ave  
Medford, MA 02155 USA

Department of Psychiatry and Athinoula A. Martinos Center for Biomedical Imaging  
Massachusetts General Hospital, Bldg 149, 13th Street  
Charlestown, MA 02129 USA

## Abstract

We have proposed that visual narratives in comics are organized with a hierarchic narrative “grammar.” Inspired by classic “click” studies of syntax, we inserted blank “disruption” panels Before, At, or After the constituent boundaries of comic strips. In self-paced viewing, Experiment 1 found that blanks After the boundary were viewed slower than Before or At the boundary. Panels immediately following blanks were slower than corresponding panels in sequences without blank panels, but only when placed Before or After the boundary. Three ordinal panel positions following the boundary, panels following blanks At the boundary or with No-Blanks were viewed faster than panels following blanks After or Before the boundary. This supports constituency because disruptions had greater impact within, as opposed to between, constituents. Using ERPs, Experiment 2 found a larger anterior negativity to blanks within constituents (Before/After) than between constituents (At). This indicates disruptions of constituents are recognized before reaching a subsequent panel. A larger P600 appeared to blanks After the boundary than in the first constituent (Before/At). This positivity may reflect a reanalysis reflecting the inability to integrate all prior panels into a single constituent, since they are divided by the constituent boundary.

**Keywords:** Grammar; Constituent Structure; Discourse; Narrative; Comics; ERPs; Left Anterior Negativity.

## Introduction

Drawings have conveyed narratives through sequences of images for thousands of years, but in contemporary society they appear most prevalently in comics. In comparison with research on the structure and comprehension of verbal narrative, however, little is known about mechanisms of processing sequential images. For example, are narrative units integrated linearly across each adjacent relationship or they organized into hierarchic constituents?

## Background

Since the 1950s, research on language has stressed that sentences are organized into hierarchic constituents, rather than linear word-to-word connections (e.g., Chomsky, 1957).

An analogous distinction between local relationships and hierarchic segmentation has underlined research of text and discourse. Traditional approaches emphasize coherence relationships between individual sentences (Halliday & Hasan, 1976; Zwaan & Radvansky, 1998). Participants can intuitively segment texts into consistent groupings (Gee & Grosjean, 1984; Mandler & Johnson, 1977), suggesting that readers’ comprehension extends beyond adjacent relationships between individual sentences.

Some theories have formally described hierarchical relationships in narrative, such as “story grammars” (e.g. Mandler & Johnson, 1977; Rumelhart, 1975; Thorndyke, 1977), which use phrase structure rules to organize narratives around characters’ goal-driven events. Constituent structure in these models was examined using clustering models (Gee & Grosjean, 1984; Mandler, 1987; Mandler & Johnson, 1977), similar to those employed by early psycholinguistic research on syntactic relations (e.g. Levelt, 1970). Participants divided a narrative into logical groupings, and these intuitions were then submitted to the algorithms in hierarchic clustering models. These analyses yielded tree structures that closely correlated with the models predicted by the original theories.

As in verbal discourse, theories of visual narrative have also emphasized linear relationships between individual images (McCloud, 1993). However, beyond such linear relationships, participants also intuitively divide visual narratives into segments (Gernsbacher, 1985). Again, this suggests that comprehension may extend beyond adjacent relationships.

Recently, Cohn (2003, In Press) has proposed a theoretical model of narrative structure that formalizes the

constituent structure of visual narrative. Like story grammars' treatment of sentences, this model organizes panels into hierarchic constituents (though important differences distinguish this approach from story grammars, and this model could potentially provide an alternative approach for describing the structure of verbal discourse).

Figure 1 illustrates the narrative structure for a 6-panel *Peanuts* comic strip. This sequence depicts a baseball game that starts with Lucy hitting a ball, which allows Charlie Brown to run home and score, escaping a tag by Schroeder. The first panel shows Lucy tossing a ball. This panel functions as an Initial, initiating the interactions in the sequence. In the second panel, Lucy hits the ball, a narrative "Peak" as the culmination of the initiated action. Together, these two panels act as an Initial constituent that propels the rest of the sequence. A second constituent begins with Schroeder waiting for the ball—nothing happens here except a set-up of the characters (Establisher). The next Initial begins the new set of events, which climax in the penultimate Peak panel—Charlie interrupts the catch by sliding into the base. The last panel then resolves this interaction (Release). Together, these panels act as a Peak that is set up by the constituent-level Initial of the first panels. Thus, at a higher level of processing, the first constituent (Lucy hitting the ball) acts as an Initial, which facilitates the second constituent (Charlie scoring), a Peak. As a result, the narrative structure operates on both the panel level and the level of whole constituencies.

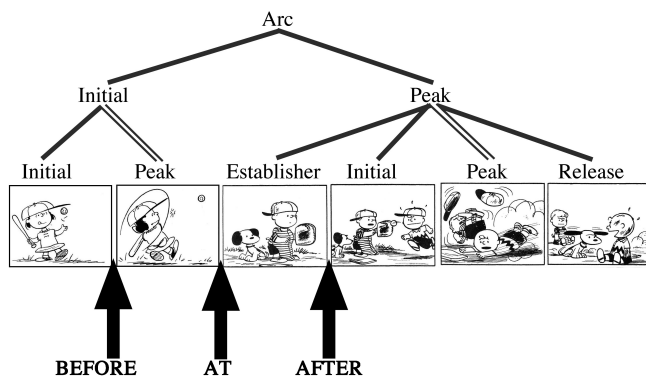


Figure 1: Narrative structure in a comic strip with markings Before, At, and After the narrative constituent boundary.

While clustering analyses have shown evidence for hierarchic structure in verbal discourse, no studies to date have examined constituent in sequential images. Furthermore, evidence that narrative constituent structure is used during *online* processing has yet to be explored in any modality. Here we describe two experiments to determine whether constituents are used when processing sequential images.

## Experiment 1: Self-paced Viewing

In a classic experiment on syntactic structure, Fodor and Bever (1965) pioneered a "click technique" where they played a verbal sentence in one ear of a participant, and then introduced short bursts of white noise ("clicks") in the other ear. They reasoned that, if clauses constitute the perceptual processing units of sentences, clicks that disrupt those units would be harder to discern than clicks occurring between clauses. They found that participants were better able to recall clicks that were placed at the clause boundaries than those before or after it. Also, participants tended to falsely recall clicks within constituencies as occurring within the constituency break. Subsequent studies found similar findings, with the overall interpretation that such disruptions reflect the psychological validity of a syntactic constituent structure (for review see Garrett & Bever, 1974).

In Experiment 1, we used an analogous "disruption" paradigm to determine whether the comprehension of visual sequences also draws upon a constituent structure in narrative sequences. We measured viewing times in graphic sequences where blank white "disruption" panels were inserted Before, At, or After the narrative constituency boundary.

## Methods

### Participants

60 self-defined comic readers (35 male, 27 female, mean age = 24.03) from the Tufts University student population and surrounding neighborhoods were paid for their participation.

### Stimuli

Novel 6-frame long sequences were created (160 sets) using individual panels culled from several volumes of Charles Schulz's *Peanuts*. Using Cohn's (In Press) model of visual narrative, we designed sequences that had two narrative constituents. This was confirmed using a behavioral task in which 20 participants drew lines to divide strips into two parts. Our final 120 strips had a 71% agreement on where constituent boundaries were located. Each sequence had constituent boundaries appearing after panel 2, 3, or 4 (40 of each type). Using these strips, we then inserted blank "disruption" panels Before, At, or After the constituency boundaries (as notated at the bottom of Figure 1), along with No-Blank control sequences. Because of the variation in position of the boundary, blank panels across all sequences could appear between the second and fifth panel positions, and items were counterbalanced across four lists such that each participant only viewed a sequence once. 15 fillers in each list had two successive blank panels and one-third of 75 additional no-blank fillers had violations of coherence.

## Procedure

Participants viewed comic panels one frame at a time on a computer screen, self-pacing their way through the sequence. After each sequence, they rated the coherence of the strip on a 1 to 5 scale.

## Results

Three-way ANOVAs showed that the position of the blank panel significantly impacted how fast it was viewed (see Figure 1),  $F(2,118)=12.93$ ,  $p<.005$ ,  $F(2,238)=7.26$ ,  $p<.005$ . This effect arose because blanks After the boundary were viewed significantly slower than blanks Before and At the boundary, (all  $t_s < -2.7$ , all  $p_s < .01$ ). There were no difference in viewing times between blanks appearing Before and At the boundary,  $t(59)=-.964$ ,  $p=.339$ ,  $t(119)=-1.09$ ,  $p=.280$ .

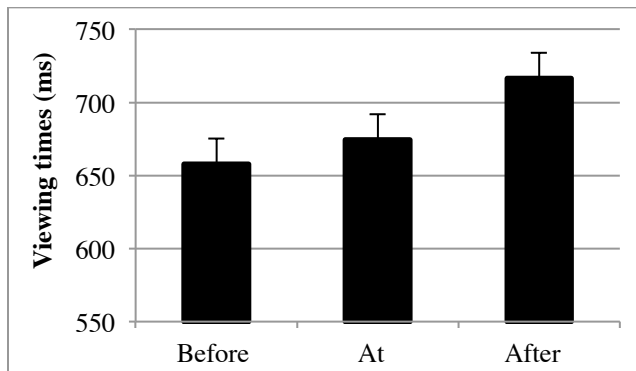


Figure 2: Viewing times to blank panels.

The placement of blank panels in a sequence also impacted the processing of subsequent panels. Viewing times to panels appearing immediately after the blank differed from corresponding panels in sequences that had No-Blank (i.e., Blank +1, see Figure 3). Overall, we found a main effect of Disruption (blank, no blank),  $F(1,59)=6.34$ ,  $p<.05$ ,  $F(2,119)=14.11$ ,  $p<.001$ , a main effect of Position (Before, At, After) in the subjects analysis,  $F(2,118)=3.11$ ,  $p<.05$ ,  $F(2,238)=.472$ ,  $p=.617$ , but no interaction between them,  $F(2,118)=1.3$ ,  $p=.276$ ,  $F(2,238)=1.57$ ,  $p=.210$ . Panels following blanks Before the narrative constituent boundary were viewed significantly slower than their corresponding No-Blank panels,  $t(59)=-3.11$ ,  $p<.005$ ,  $t(119)=-3.48$ ,  $p<.005$ , as were panels following blanks After the boundary, a trending difference in the items analysis,  $t(59)=-1.34$ ,  $p=.184$ ,  $t(119)=-1.8$ ,  $p=.074$ . Panels following blanks At the boundary were not slower than corresponding panels in No-Blank sequences,  $t(59)=-.878$ ,  $p=.383$ ,  $t(119)=-1.56$ ,  $p=.121$ .

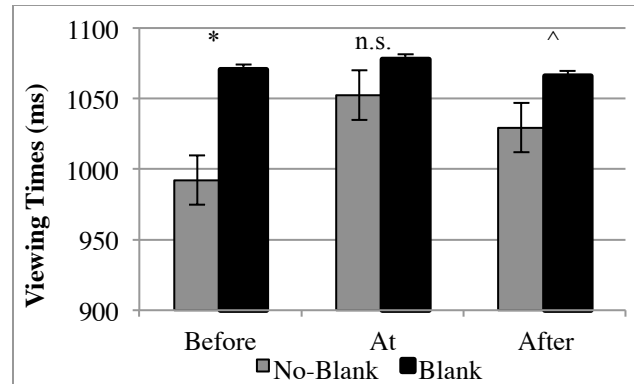


Figure 3: Viewing times to panels immediately after the blank panel (Blank +1) compared with corresponding panels in sequences with No-Blanks.

Delayed effects were also found three panel positions after the narrative constituent boundary (Boundary+3, see Figure 4),  $F(3,177)=2.90$ ,  $p<.05$ ,  $F(3,357)=5.24$ ,  $p=.005$ . Panels following a blank After the boundary were viewed slower than panels following blanks At the boundary and panels in sequences with No-Blank, (all  $t_s > -3.36$ , all  $p_s < .05$ ). Panels following blanks Before the boundary were viewed slower than those following blanks At the boundary,  $t(59)=1.49$ ,  $p=.142$ ,  $t(79)=2.12$ ,  $p<.05$ , and trending to be slower than panels in sequences with No-Blanks,  $t(59)=-1.56$ ,  $p=.124$ ,  $t(79)=-1.77$ ,  $p=.081$ . However, viewing times did not differ between panels following blanks Before and After the boundary, or between panels following a blanks At the boundary and in sequences with No-Blank, (all  $t_s < -1.34$ ,  $p > .184$ ).

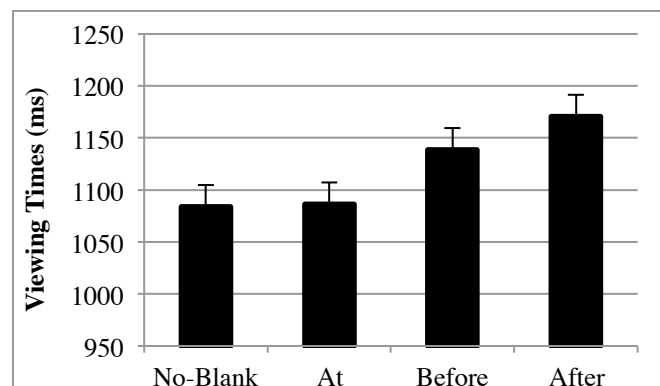


Figure 4: Viewing times to panels three positions after the narrative constituent boundary (Boundary +3) in all sequence types.

## Discussion

Blank panels were viewed slower when they appeared After compared to Before or At the narrative constituent boundary. Panels immediately following blanks were slower than corresponding panels in sequences without a preceding

blank panel, but only when placed Before or After the boundary. A delayed effect of this disruption appeared three panels after the narrative constituent boundary. At this position, panels following blanks At the boundary or with No-Blanks were viewed faster than panels following blanks After or Before the boundary. These results are consistent with the presence of constituent structure to visual narrative: a blank panel disruption had greater impact within (Before/After) as opposed to between (At) constituents.

## Experiment 2: Event-Related Potentials

Different ERP components have been associated with manipulations of semantic and syntactic constraints. For example, the N400 ERP component has been associated with semantic processing of words (Kutas & Hillyard, 1980), individual visual images (Holcomb & McPherson, 1994), and images in visual narratives images (Cohn, Paczynski, Jackendoff, Holcomb, & Kuperberg, 2012; West & Holcomb, 2002). The N400 effect has been suggested as indexing the spreading activation of a word or image's meaning as it integrates with its preceding context to the information stored in semantic memory (Kutas & Federmeier, 2011).

In contrast, two ERP components have been linked to violations of grammatical structure during sentence processing. First, the P600 is a positive deflection that peaks from 600-800ms (Osterhout & Holcomb, 1992). It is seen to syntactic anomalies and ambiguities and has been interpreted as reflecting a process of continued analysis and/or repair as structural and semantic information are integrated to make sense of a sentence (Friederici, 2002; Kuperberg, 2007).

Waveforms resembling the P600 have also been found in domains outside of language. For example, in studies using silent movie clips of everyday events, a P600 was seen to "action violation" endings in which a predicted action was carried out with an incongruous object (e.g. a person preparing to cut a piece of bread followed by an image of the person attempting to cut the bread with an iron). This suggested that the P600 effect may reflect the integration of structural and semantic information around an event beyond linguistic processing (Sitnikova, Holcomb, & Kuperberg, 2008).

Certain syntactic operations have also been tied to a left-anterior negativity (LAN), which falls in the same time window as the N400 (between 300 and 500ms), but is distributed over frontal and left lateralized regions (Neville, Nicol, Barss, Forster, & Garrett, 1991). This component has been associated with violations of syntactic constituent structure in sentences (Friederici, 2002).

Anterior negativities outside of language have also been associated with violations of hierarchic structure in music. Patel and colleagues (1998) found a P600 to structural violations in musical sequences (e.g., a nearby key chord or a distant-key chord appearing after an otherwise in-key musical sequence). Another negative-going effect, distributed over right anterior and temporal sites, also

appeared between 300 and 400 milliseconds. This (*early right anterior negativity*) has led researchers to argue for overlap in the neural resources used to process structure in both music and language (Koelsch, Gunter, Wittfoth, & Sammler, 2005; Patel, et al., 1998).

Only one study has previously hinted at a LAN during the processing of visual sequences. Cohn et al. (2012) manipulated the presence or absence of narrative structure and semantic relatedness between panels in visual narratives. A larger negativity was found to sequences of scrambled panels (which had no semantic relations and no narrative structure) than to sequences with only narrative structure but no semantic theme (i.e., the sequence followed a narrative arc, but used panels from various sequences which had no meaningful relationship with each other). However, this negativity only appeared in a localized left anterior region, and was larger in participants with higher comic reading expertise. This left anterior effect was distinguished from a more widespread N400 effect that was greater in magnitude in these sequences that lacked semantic relations than to panels in sequences with no narrative structure, but with semantic relations to a general theme (like baseball), and to semantically and narratively congruous Normal sequences. We suggested that the left localized negativity—distinguished from the N400—might be analogous to the left anterior negativity effect seen to violations of structure in language. However, because this study did not directly introduce violations to narrative structure, this interpretation was speculative.

In Experiment 2, we predicted that violations of constituency in visual narratives would elicit similar ERP effects as violations of structure during sentence processing: P600 and LAN effects. We used the same set of stimuli from Experiment 1 and measured ERPs directly at blank panels.

## Methods

### Participants

24 self-defined comic readers from Tufts University (8 male, 16 female, mean age = 19.9) participated in the study for compensation.

### Stimuli

The same stimuli were used in Experiment 2 as in Experiment 1.

### Procedure

Participants viewed each sequence one panel at a time on a computer screen while ERPs were measured to all panels. After each sequence, participants rated its coherence on a 1 to 5 scale.

## Results

At the blank panel, between 500 and 700 milliseconds, repeated measures ANOVAs revealed significant interactions between Disruption Position, Region, and



Electrode Site at midline regions, and interactions between Disruption Position, Region, and Hemisphere at peripheral regions of the scalp (all  $F_s > 3.24$ , all  $p_s < .05$ ). Follow-ups showed a larger negativity to blank panels that appeared Before the boundary than to those appearing At the boundary at left anterior regions (F3, FC5, F7). Blanks appearing After the boundary evoked a larger negativity than those At the boundary again in left anterior regions (F3, FC5, F7) as well as at FPz and FP2. ERPs to Blank panels appearing Before or After the boundary did not differ in amplitude.

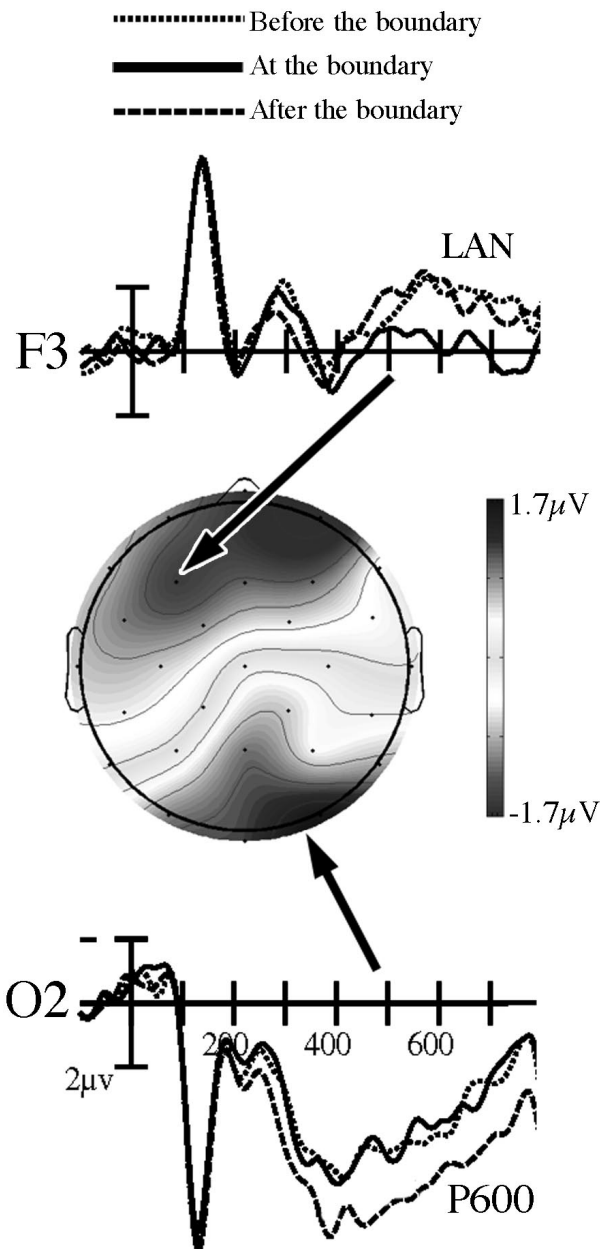


Figure 5: Amplitudes for blank panels at left anterior (F3) and posterior (O2) sites. Scalp map shows the contrast between blank panels After and At the boundary from 500 and 700ms.

At posterior sites, a larger P600 appeared to blank panels placed After the boundary (within the second constituent) than those placed Before or At the constituency boundary. The contrast between the After and At conditions revealed a significant effect at O1, Oz, and O2, and between the After and Before conditions showed an effect at O2.

## Discussion

A larger negativity was found to disruption blank panels that appeared within constituents (Before/After) than to blanks that appeared between constituents (At). This negativity localized to anterior, and somewhat left lateralized, locations. This is consistent with the lateralized anterior distribution of negativity effects shown to violations of syntax in sentences (Neville, et al., 1991) and music (Koelsch, et al., 2005; Patel, et al., 1998). It is important to note that it is only possible to confirm or verify that a blank panel has actually disrupted a narrative constituent once the subsequent panel is reached. However, we still saw a larger anterior negativity where the disruption occurred Before than At the boundary on the blank panel itself, prior to confirmation on the subsequent panel. This suggests that the brain may make online predictions about the building of constituent structure as a narrative progresses.

In addition to the anterior negativity effect, we also saw a P600 effect to blank panels appearing After (versus Before or At) the boundary. Because this positivity appeared to blanks only after a new constituent had been reached, we suggest that it reflected the failure of integrating all of the prior panels into a single constituent. This blank followed a panel after the constituent boundary, meaning that the preceding panel would be unable to be integrated into a single constituent with the other prior panels, thereby evoking a reanalysis of starting a new constituent (Friederici, 2002; Kuperberg, 2007).

## General Discussion

The results of Experiment 1 suggest that the introduction of a blank panel into a sequence has a greater impact on viewing times of subsequent panels if it falls within a narrative constituent (Before/After a narrative boundary) than between narrative constituents (At a narrative boundary). This provides evidence that comprehenders use narrative boundaries during the processing of visual sequences. The results of Experiment 2 support this interpretation. Moreover, because effects were seen on the blank panel itself, this further suggests that comprehenders of visual sequences make active predictions about narrative constituent structures.

Taken together, these results show that disruptions within a narrative constituent have a greater impact on processing than disruptions between narrative constituents. This suggests that the comprehension of sequential images draws upon a narrative structure that is organized into constituents, analogous to grammatical structure in language.

## Acknowledgements

This work was supported by NIMH (R01 MH071635) and NARSAD (with the Sidney Baer Trust), as well as funding from the Tufts Center for Cognitive Studies. Suzanne Grossman, Chelsey Ott, and Patrick Bender their aid in designing stimuli and gathering data. Fantagraphics Books is thanked for their generous donation of *The Complete Peanuts*.

## References

- Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.
- Cohn, N. (2003). *Early Writings on Visual Language*. Carlsbad, CA: Emaki Productions.
- Cohn, N. (In Press). Visual narrative structure. *Cognitive Science*.
- Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012). (Pea)nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive Psychology*, 65(1), 1-38. doi: 10.1016/j.cogpsych.2012.01.003
- Fodor, J., & Bever, T. G. (1965). The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior*, 4(5), 414-420.
- Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, 6(2), 78-84.
- Garrett, M. F., & Bever, T. G. (1974). The Perceptual Segmentation of Sentences. In T. G. Bever & W. Weksel (Eds.), *The Structure and Psychology of Language*. The Hague: Mouton and Co.
- Gee, J. P., & Grosjean, F. (1984). Empirical evidence for narrative structure. *Cognitive Science*, 8, 59-85.
- Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive Psychology*, 17, 324-363.
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Holcomb, P., & McPherson, W. B. (1994). Event-Related Brain Potentials Reflect Semantic Priming in an Object Decision Task. *Brain and Cognition*, 24, 259-276.
- Koelsch, S., Gunter, T. C., Wittfoth, M., & Sammler, D. (2005). Interaction between syntax processing in language and in music: An ERP study. *Journal of Cognitive Neuroscience*, 17(10), 1565-1577.
- Kuperberg, G. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146, 23-49.
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, 62(1), 621-647.
- Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potential reflect semantic incongruity. *Science*, 207, 203-205.
- Levelt, W. (1970). A scaling approach to the study of syntactic relations. In G. B. Flores d'Arcais & W. Levelt (Eds.), *Advances in psycholinguistics* (pp. 109-121). New York: Elsevier.
- Mandler, J. M. (1987). On the psychological reality of story structure. *Discourse Processes*, 10, 1-29.
- Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, 9, 111-151.
- McCloud, S. (1993). *Understanding Comics: The Invisible Art*. New York, NY: Harper Collins.
- Neville, H. J., Nicol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of Cognitive Neuroscience*, 3(2), 151-165.
- Osterhout, L., & Holcomb, P. (1992). Event-related potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31, 758-806.
- Patel, A. D., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. J. (1998). Processing syntactic relations in language and music: An event-related potential study. *Journal of Cognitive Neuroscience*, 10(6), 717-733.
- Rumelhart, D. E. (1975). Notes on a schema for stories. In D. Bobrow & A. Collins (Eds.), *Representation and understanding* (Vol. 211-236). New York, NY: Academic Press.
- Sitnikova, T., Holcomb, P. J., & Kuperberg, G. (2008). Neurocognitive mechanisms of human comprehension. In T. F. Shipley & J. M. Zacks (Eds.), *Understanding Events: How Humans See, Represent, and Act on Events* (pp. 639-683): Oxford University Press.
- Thorndyke, P. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology*, 9, 77-110.
- West, W. C., & Holcomb, P. (2002). Event-related potentials during discourse-level semantic integration of complex pictures. *Cognitive Brain Research*, 13, 363-375.
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2), 162-185.

# Framing attention in American and Japanese comics

Neil Cohn (neilcohn@emaki.net), Amaro Taylor-Weiner, Suzanne Grossman

Department of Psychology, 490 Boston Ave  
Medford, MA 02155 USA

## Abstract

Research has shown that Americans focus more on focal objects of a scene while East Asians attend to the surrounding environment (Nisbett, 2003; Nisbett & Miyamoto, 2005). The panels of comic books—the sequential frames of images—highlight aspects of a scene comparably to how attention focuses on parts of a spatial array. Thus, comparison of American and Japanese comics can inform cross-cultural cognition by looking at the expressive mediums produced by these cultures. We compared the framing of figures and scenes in the panels of two genres of American comics (Independent and Mainstream) with mainstream Japanese “manga.” Both genres of American comics focused on whole scenes as much as individual characters, while Japanese manga individuated characters and parts of scenes. We argue that this framing of space in comics simulates a viewer’s integration of a visual scene, and is consistent with cross-cultural differences in the direction of attention.

**Keywords:** Cultural Psychology; attention; comics; Japan; manga.

## Introduction

Cross-cultural research shows that East Asians and Westerners differ in their direction of attention (Nisbett, 2003; Nisbett & Miyamoto, 2005). Beyond studying attention through perception, cognition can also be compared through cultural production (Morling & Lamoreaux, 2008), as in artistic expression (Masuda, Gonzalez, Kwan, & Nisbett, 2008). Comic books provide an ideal place to analyze the direction of attention, because panels act like windows onto a scene (Cohn, 2007). Thus, analysis of panels in Asian and American comics provides a place to look for cultural differences in cognition through creative expression.

## Attention across Cultures

Over the past decade, various research has shown that Asians and Americans direct their perception to aspects of visual scenes in different ways (Nisbett, 2003; Nisbett & Miyamoto, 2005). On the whole, Americans focus more on focal objects and characters with agency than on aspects of the background, while Asians attend to aspects of the whole environment or to characters’ relationship to the contextual environment.

These findings have been consistent across numerous behavioral paradigms. After viewing video scenes, Americans mostly describe the salient objects, while Asians

describe significantly more aspects of the surrounding context (Masuda & Nisbett, 2001). Americans also tend to notice changes to focal objects in animations that feature slight changes to a single scene, while Asians pick up on changes to the broader environment and relations between objects (Masuda & Nisbett, 2006). When recalling scenes where the background is changed from its original context, Americans are unaffected while Asians’ memory appears impaired (Masuda & Nisbett, 2001), and Americans’ eye movements fixate sooner and longer on focal objects, while Asians make more saccades to elements of the background (Chua, Boland, & Nisbett, 2005). Additionally, when viewing photographs of objects, fMRI studies show that Americans have stronger activation than Asians in brain regions associated with the storing of semantic information about object properties (Gutchess, Welsh, Boduroglu, & Park, 2006). All of this work supports that Americans focus more on focal objects while Asians attend more to aspects of environments and relationships.

Research has also suggested that preferences for attention permeate into artistic representations. Masuda, Gonzalez, Kwan, and Nisbett (2008) looked at a corpus of artwork, and found that Western paintings emphasized the focal objects and figures, while East Asian paintings emphasized the broader context and environment. This trend was reinforced in drawings and photographs of figures and scenes produced by individuals from these cultures. Thus, these cognitive preferences for attention extend into artistic expression, and other contemporary media produced by these cultures might be expected to show further evidence of these trends.

## Comic Panels as Units of Attention

Comic books are an ideal place to examine attention in artistic expression. Because comic panels act as windows on a visual story, they can serve as graphic equivalents of a “spotlight of attention” for the fictitious scene. To this end, Cohn (2007) has described comic panels as “attention units” that highlight parts of a scene in different ways. Within a sequence of images, a scene may have two types of elements: *Active entities* are those that repeat across panels by engaging in the actions and events of the sequence, while *inactive entities* are elements of the background. Panels can be categorized related to these elements (and depicted in Figure 1):

1. *Macro* – depict multiple active entities
2. *Mono* – depict single active entities
3. *Micro* – depict less than one active entity (as in a close up)
4. *Amorphic* – depict no active entities (i.e., only inactive entities)

In one of the first comparisons of American and Japanese comics, McCloud (1993, pp. 77-81) coded types semantic relationships between juxtaposed panels. He found that American and European authors primarily used transitions showing actions with clear temporal change, followed by shifts between characters and locations. Manga similarly showed shifts in actions, characters, and locations.

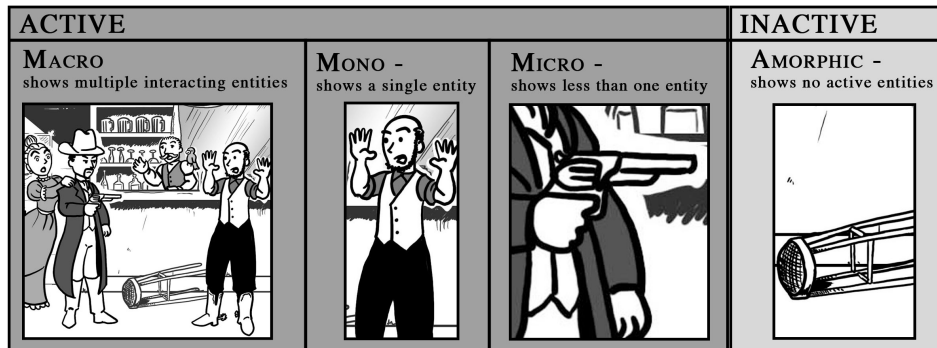


Figure 1: Framing of attention in visual narrative.

These categories are distinguished by the amount of information they contain, which decreases successively: Macros contain more active information than Monos, which show more than Micros, which show more than Amorphic panels. These ways of highlighting attention are similar to, though not the same as, types of film shots. For example, a Macro may involve a long shot to capture the most information possible, but a panel showing only the hands of individuals exchanging a piece of paper would be a Macro (because they involve multiple characters) that uses a close up shot. In this way, close ups are also not always Micros, but they vary based on how much information they window.

### Cross-cultural Comparison of Comics

With the growing influx of Japanese manga (“comics”) into America over the past several decades, many comparisons have been made between the techniques of Japanese and American authors (Cohn, 2010, 2011; McCloud, 1993, 1996). Japanese manga come from a different cultural context than that of American comics. While comics in America have historically appealed to a particular subculture, manga in Japan are treated much the same as movies, television, or textual books. Manga are widely read by all ages, have many genres, and, in fact, are so popular that they constitute nearly one-third of all printed material (Gravett, 2004; Schodt, 1983). Though Japanese manga were influenced by American authors early in their historical development (Gravett, 2004), they developed largely in isolation over the past 60 years. With increased importation of manga into America starting in the 1980s, the differences between narrative techniques that emerged from these separate traditions have become quite salient to readers, authors, and scholars of comics in America.

However, unlike American and European books, manga also transitioned to different aspects of the environment of a scene. McCloud attributed these differences to an “artistic culture” of Japan that focused on “being there over getting there.”

This hypothesis extended McCloud’s (1993, pp. 77-81; 1996) larger proposal that

manga allow a reader to take more of a subjective viewpoint on a story than American and European comics. He based this on the greater focus on environmental aspects in storytelling, more “subjective” types of motion lines (where a reader appears to move at the same pace as a moving object, as opposed to seeing it move in front of them), and subjective viewpoints in panels, which show the viewpoint of a character in the narrative. In order to test this broad claim directly, Cohn (2011) coded a corpus of comics and manga for this last type of subjectivity, where panels depict the viewpoint of a character in the narrative. More subjective panels were used in Japanese manga than American comics. This provided evidence that manga do indeed use more subjective viewpoints, at least across one measurable dimension.

Cohn’s (2011) study also examined the attentional types of panels described above. Nearly 60% of American panels were Macros, with only 35% Monos and 5% Micros (Amorphics were not yet theorized as a category, and were likely mixed in Monos and Micros). However, Japanese manga used almost as many Macros (57%) as Monos (43%), and more Micros (10%) than American comics. Because manga featured less than the whole scene in over half of all panels, it implies that the Japanese are as interested in the component parts of a scene as much as the whole scene. These results also suggest that the narrative structure of manga demands the inferential construction of whole scenes more than American comics (Cohn, 2010). These findings of more Micros in Japanese manga are also consistent with claims by Toku (2001, 2002) that manga influences Japanese children’s drawings. She found that Japanese children draw far more variable viewpoints than American children, particularly “exaggerated” close-ups.

While these studies have indicated that comic panels differ between cultures, variability may exist by looking within cultures. Obvious variability can be found in the diversity of American graphic styles compared to the far more uniform drawing style in manga. Graphic styles are

particularly pronounced between genres, which in America differ greatly between the more “serious” graphic novels and mainstream comics. Styles in genres of Japanese comics also vary, but mostly conform to a stereotypical style of big eyes, pointy chins and noses, and big hair. The diverse styles used in American comics have been likened to types of “dialects”, compared with “accents” in manga genres, which vary on a common schema (Cohn, 2010).

Variation between genres may apply to the level of panels as well, and can thereby inform about the framing of attention. In an early study, Neff (1977) found that panels use types of film shots differently between various genres of American comics. Wide shots (Long and Medium) far outnumbered Close shots (Close and Close ups) in panels for all genres. However, there were far less Close shots in Adventure and Romance comic panels than in Mystery and Alien Beings comics. These findings imply that different genres of American books do highlight different aspects of a visual scene. However, the sample size in this study was somewhat limited in scope—only two pamphlet-sized comics were analyzed per genre—making the results hard to generalize.

Given these precedents, the present study examined comic panels both within and between cultures by manipulating country of origin and genre. We compared the panels of “mainstream” Japanese manga with the two major populations of American comics: Mainstream and Independent (“Indy”) books. Mainstream books from both America and Japan were chosen because they are the most popular and most stereotypical instances of their respective comic cultures. American Indy books were chosen because they feature a different artistic movement in America that contrasts the Mainstream genres (discussed below). Thus, if variation occurs between the structures of comics from America, we may expect it between Mainstream and Indy comics.

If panel types of all three populations differ, it would imply sub-cultural “artistic” contexts vary related to narrative techniques of particular traditions. If Japanese panel types are similar to Mainstream comics yet different from Indy comics, it would imply that the framing of a scene differs based on genre, even cross-culturally. In contrast, if American genres do not differ from each other, yet both differ from Japanese manga, it would imply cultural differences beyond the contexts of genre.

To this end, if both American genres do differ from Japanese panels, we would predict the results to reflect the findings of Masuda et al. (2008) for art and photographs by Asians and Americans. Similar results would expect American comics to focus more on focal objects through Monos and Amorphics. Meanwhile, Japanese panels should show the opposite: here we would predict more Macros to focus on the relationships between characters in whole scenes.

## Methods

### Materials

Thirty graphic books were chosen at random from a corpus of over 200 comics donated from various comic companies. We coded 300 panels in each book for the properties of attentional panel type and shot type. 10 books were chosen from each of three populations: “mainstream” Japanese Manga, Mainstream American comics, and Indy American comics. In order to operationalize how these populations are identified, it is useful to discuss their differences.

Mainstream and Indy books differ greatly in graphic styles, genres, formats, publishers, and often readership. Mainstream comics primarily feature drawing styles common to superhero comics (dynamic linework, muscular figures, brighter colors), and focus on the genres of superheroes, horror, and science fiction. Mainstream books are also often produced by specific publishers and are serialized in pamphlet style formats that are only sometimes afterwards collected into books. Mainstream comics are sold primarily through specialty comic books stores. In contrast, Indy books use more variable graphic styles (particularly more cartoony and “artistic” styles) with more “serious” or dramatic genres (such as memoir, drama, etc.). Different publishers are known for producing Indy books and Mainstream comics, and they appear mostly in book formats (“graphic novels”). Indy books are often sold in comic books stores, but also have a much higher distribution into regular bookstores.

While some overlap in readership does exist between Mainstream and Indy comics, they largely appeal to different groups of people. Readers of Mainstream comics often read serializations that appear each month. They often are very devoted to their favorite comics, and American comics often target the writing with this consistent readership in mind, evident through frequent references to previous storylines. Indy comics have more varied readership because they are not serialized volumes. Often, Indy books are produced in single editions, and thus do not have consistent readership (though readers may follow particular authors’ works). Readers of Japanese manga are often more similar to Mainstream American comics—they have their favorite comics which are released weekly in large anthologies. While readership of manga is larger on the whole in Japan than America, there is no reason to believe that comics in either country are explicitly made with any expectation that readers will be more or less proficient in understanding them.

Additionally, while some crossover exists in readership between American genres, most authors of Mainstream and Indy books remain independent to their genres. Mainstream and Indy books are also created with a slightly different process. Mainstream comics are largely made by an industry-line style committee (Duncan & Smith, 2009) consisting of a writer, penciler, inker, colorist, etc. While an editor coordinates their efforts and oversees the plotline, for the most part these creators are free to follow their own

styles of writing and artwork. In contrast, Indy comics are more often drawn and written by individual authors. Japanese manga typically combine these methods. They are usually attributed to a sole author, who then employs a team of uncredited assistants who complete the more menial aspects of the drawings, like shading or drawing backgrounds (Schodt, 1983). While these creative processes may vary between countries/genres, the finished products largely reflect the intuitions of the authors or creative teams.

In this study, we distinguished American Mainstream and Indy books by criteria of graphic style, genre, and publishers. Mainstream books ranged in publication date from 1992 to 2005 with a mean of 2002, while Indy comics were published between 1991 to 2008, with a mean at 2003. Japanese books featured more consistent visual styles, following the stereotypical “standard graphic dialect” of Japanese comics (Cohn, 2010). However, since genres in Japan do not align neatly with those in America (Shonen “boys comics,” Shojo “girls comics,” and Gekiga “serious comics”), books were chosen that reflected the genre closest to Mainstream American comics—those focusing primarily on action/adventure themes (Shonen “boys comics”). Only English translations of manga were analyzed in the study due to their availability in our donated corpus, though manga were attributed to their original Japanese publication dates, from 1984 to 2005 with a mean of 1999.

Thus, our analysis contrasted either genre or country of origin. American Mainstream books shared a similar overall genre (action/adventure) with Japanese manga, though they came from the same country of origin as American Indy comics. All of the chosen books were widely read and popularly distributed throughout comic readership, and from major publishers—i.e., none of the books were obscure or minimally distributed. Books analyzed are provided online at [http://www.emaki.net/CTG\\_FramingAppendix.html](http://www.emaki.net/CTG_FramingAppendix.html).

## Areas of analysis

All books were coded for their attentional Panel Type—the way in which panels highlight attention in the various types of attentional categories previously discussed (Macro, Mono, Micro, Amorphic). Panels that could not be recognizably coded into these categories were identified as “Ambiguous.” Two researchers independently coded each book’s properties, and were consistent in their codings ( $Kappa=.785$ ,  $p<.01$ ). Final analyses used the mean between coders’ scores for each book.

Populations were fairly similar in the number of pages/book and panels/page analyzed. Indy comics averaged

56.6 pages/book and 5.99 panels/page, while Mainstream comics averaged 62.6 pages with 5.12 panels/page. Manga used 65.2 pages/book with 4.75 panels/page.

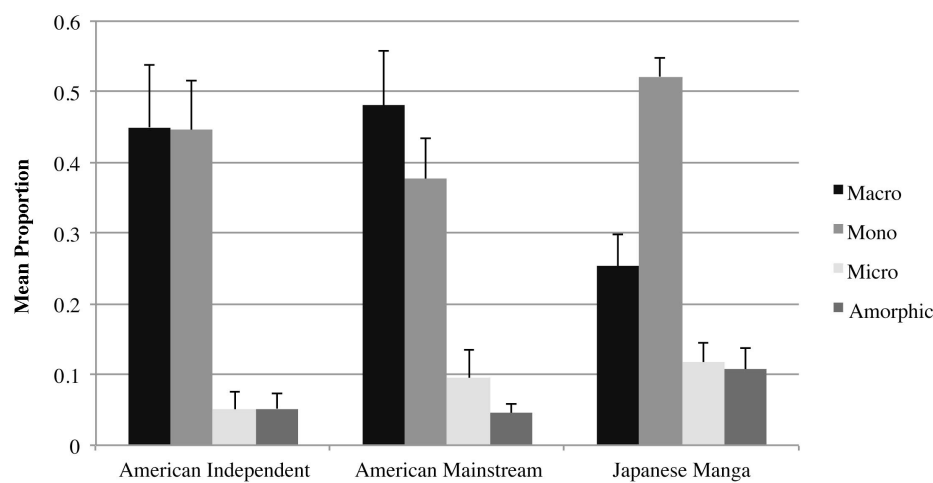


Figure 2: Relative proportion of panel types across American Mainstream comics, American Indy comics, and Japanese Manga.

## Results

### Panel Type

The analysis of attentional framing of panels found main effects for Panel Type,  $F(3,81)=89.71$ ,  $p<.001$ , with a Panel Type by Group interaction,  $F(6,81)=5.68$ ,  $p<.001$ . Main effects between Groups were not significant,  $F(2,27)=1.37$ ,  $p=.269$ .

As depicted in Figure 2, Indy and Mainstream comics used many Macros and Monos, with minimal Micros and Amorphics. Within Indy comics, overall differences were found between panel types,  $F(3, 27)=27.34$ ,  $p<.001$ , as well as between each pair of types (all  $t>5.81$ , all  $p<.001$ ), except the near equal means for Macros with Monos, and Micros with Amorphics. Mainstream panels also differed between all types,  $F(3,27)=30.05$ ,  $p<.001$ . These books featured only slightly more Macros than Monos, which was not statistically significant. Micros and Amorphics numbered far fewer overall, though there were almost twice as many Micros as Amorphics,  $t(9)=2.14$ ,  $p=.06$ . All other panel types featured significant contrasts (all  $t>5.55$ , all  $p<.001$ ).

Finally, Manga also showed main effects between panel types,  $F(3,27)=64.00$ ,  $p<.001$ . Monos far outnumbered other types, with roughly half as many Macros, and far fewer Micros and Amorphics. All types differed from each other, (all  $t>3.16$  or  $<-7.3$ , all  $p<.05$ ), except Micros and Amorphics.

Across the three populations, differences were found between each Panel Type (all  $F_s>2.8$ , all  $p_s<.01$ ). Indy and Mainstream comics showed no differences for any of the panel types (all  $t_s<1.8$ , all  $p_s>.11$ ). Indy panels differed

from Manga for all types (all  $t_s > 2.6$ , all  $p_s < .05$ ) except Monos, while Mainstream panels differed from Manga on all types (all  $t_s > 2.9$ , all  $p_s < .01$ ) except with regard to Micros.

## Discussion

This study analyzed how various cultures' comic panels frame a fictitious scene as a way to gain insight on how these cultures may direct attention. We compared Mainstream and Indy genres of American comics with "mainstream" Japanese manga. Even more than in Cohn's (2011) study, Japanese panels highlighted individual elements of scenes more than American books. Japanese manga were found to have far more Monos than any other type of panel, followed by Macros, and small proportions of Micros and Amorphics. Both Mainstream and Indy American comics had near equal proportions of Macros and Monos, again with small proportions of Micros and Amorphics.

In the analysis of panels types between cultures, manga used significantly more Monos, Amorphics, and Micros than did both types of American comics. American comics did not vary in their panel types between genres, despite surface stylistic differences. Thus, though Japanese manga and Mainstream American comics were similar in terms of "mainstream" appeal and action/adventure themes, this similarity did not influence the framing of scenes. These results suggest that the primary difference between these populations of comics are that of country of origin: The framing of entities in American comic panels differ from Japanese panels, though American comic genres do not differ substantially from each other.

What can these results offer to our understanding of cross-cultural attention and cognition? On the whole, the framing of attention in both genres of American comics focused more on detailing a whole scene as much as, if not more than, individual characters, as indicated by the prevalence of Macros over Monos. In contrast, Japanese manga directed attention toward details in the scene through Monos, Micros, and Amorphics, in lieu of actually showing full scenes in Macros.

These results seem to run counter to the cross-cultural research on attention. As suggested by the analysis of art and photographs in Masuda, Gonzalez, Kwan, and Nisbett (2008), wouldn't we expect American comic panels to focus more on the primary objects of a scene (i.e., Monos) because of a preference for objects over environments? Shouldn't Japanese panels focus more on scenes as a whole (i.e., Macros)? If these results are to be taken directly, they provide counter-evidence for the claims made by Nisbett and Masuda with regard to the manifestation of attention in popular culture.

One interpretation of these results is that cross-cultural panel framing has nothing to do with attention, but rather reflects the expertise of each cultures' readers. Manga are far more prevalent in Japan than in America, and thus Japanese may have a greater expertise in general in reading

sequential images. American comics may be geared towards less experienced readers, and thus they need to be constantly reminded of the elements in a scene with more Macros. On the other hand, more experienced readers in Japan may be able to retain or construct the whole scene without being presented with it.

Thus, under this interpretation, attention is not a factor at all. We find this explanation to be unfeasible. Manga do indeed have wider readership across the country of Japan compared with the readership of comics across America. However, American comics, particularly Mainstream comics, are targeted towards a consistent readership. These readers are often serious and devout fans, and would have as much if not greater fluency in their visual language than casual manga readers in Japan. Thus, attributing these findings to expertise alone seems unlikely.

Though these results on the surface appear to contrast previous findings on attention, we suggest another interpretation of these results that is indeed more consistent with the research by Nisbett and Masuda. Comic panels are not isolated images like the photos and drawings, but are instead meant to be read (and are created) in a sequence. A sequence of images in comics act as a simulation of how an individual might view a fictitious visual scene in front of them (a similar argument for film shots is made by Levin & Simons, 2000). This simulation of attention across sequential images is different from the treatment of attention in individual images, like in the study by Nisbett and Masuda (2003).

Like in attention, readers track only the most important aspects of a sequence to establish the continuity of the narrative. Non-relevant information may then go unattended by the "spotlight of attention" across panels, as happens in change blindness paradigms (Levin & Simons, 2000). There are thus two strategies a comic author can use when creating comic. They can either show a full scene (Macro) and rely on the reader's attentional intuitions to discern the most important parts, or they can use panels to directly highlight only those salient parts directly (Monos, Micros, Amorphics), omitting what is unimportant altogether. This use of panels would heighten the "subjective viewpoint" of panels simulating attention.

These and previous data suggest that American comics more consistently use the first option: letting the reader direct their own attention across panels to find the most relevant aspects of continuity, while letting less important elements simply go unattended. This is suggested by the larger amounts of Macros found in American comics of both genres. In contrast, Japanese manga do more to simulate the perception of a reader's attention, evident in greater use of Monos, Micros, and Amorphic panels. That Japanese manga use a strategy that is more subjective of the way attention may be directed is consistent with McCloud's (1993, 1996) claim that manga allow a reader to take more of a subjective viewpoint on a story. It also is supported by previous corpus analysis showing that "subjective panels"—panels that directly show the viewpoint of a character in the

narrative—are more plentiful in Japanese manga than American comics (Cohn, 2010).

These different strategies of depicting actions by simulating attention also reflect the way in which attention may be different between readers of different cultures. Manga panels highlight individual elements of a scene or environment because that would be how Japanese readers' attention would fall on elements of a visual array, and out of this information would need to integrate these parts into a coherent whole. In contrast, because American readers will naturally pick out the focal characters of the scene, American comics can use more Macros, assuming attention will be directed to the important elements of interest automatically. In this way, panels from comics and manga reflect how a Japanese or American reader might look at a visual scene if the whole array were in front of them, thereby echoing the differences in cultural windowing of attention.

By analyzing comics with a clearly defined categorization system, we have shown that visual narratives are bound by cultural conventions that create patterns in the ways that Japanese and American comic authors window attention onto visual scenes. We propose that these results are consistent with the cross-cultural research showing differences in how Asians and Americans perceive and attend to their visual environment (Nisbett, 2003; Nisbett & Masuda, 2003; Nisbett & Miyamoto, 2005), and lend further support to efforts to study cognitive process through creative cultural expression.

## Acknowledgments

Thanks go to the Tufts Center for Cognitive Studies for funds supporting this research and to these publishers for their generous donation of books: Dark Horse Comics, Drawn & Quarterly, First Second Books, IDW Publishing, Oni Press, Top Cow, and Viz Media. Early drafts benefited from feedback from Ray Jackendoff, Igor Grossman, and Taka Masuda and the Culture and Cognition Lab at the University of Alberta.

## References

- Chua, H. F., Boland, J. E., & Nisbett, R. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences*, 102(35), 12629-12633.
- Cohn, N. (2007). A visual lexicon. *Public Journal of Semiotics*, 1(1), 53-84.
- Cohn, N. (2010). Japanese Visual Language: The structure of manga. In T. Johnson-Woods (Ed.), *Manga: An Anthology of Global and Cultural Perspectives* (pp. 187-203). New York: Continuum Books.
- Cohn, N. (2011). A different kind of cultural frame: An analysis of panels in American comics and Japanese manga. *Image [&] Narrative*, 12(1), 120-134.
- Duncan, R., & Smith, M. J. (2009). *The Power of Comics*. New York: Continuum Books.
- Gravett, P. (2004). *Manga: Sixty Years of Japanese Comics*. New York, NY: HarperCollins.
- Gutchess, A., Welsh, R., Boduroglu, A., & Park, D. (2006). Cultural differences in neural function associated with object processing. *Cognitive, Affective, & Behavioral Neuroscience*, 6(2), 102-109.
- Levin, D. T., & Simons, D. J. (2000). Perceiving Stability in a Changing World: Combining Shots and Integrating Views in Motion Pictures and the Real World. *Media Psychology*, 2(4), 357-380.
- Masuda, T., Gonzalez, R., Kwan, L., & Nisbett, R. E. (2008). Culture and Aesthetic Preference: Comparing the Attention to Context of East Asians and Americans. *Personality and Social Psychology Bulletin*, 34(9), 1260-1275. doi: 10.1177/0146167208320555
- Masuda, T., & Nisbett, R. (2001). Attending Holistically Versus Analytically: Comparing the Context Sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, 81(5), 922-934.
- Masuda, T., & Nisbett, R. (2006). Culture and Change Blindness. *Cognitive Science*, 30, 381-399.
- McCloud, S. (1993). *Understanding Comics: The Invisible Art*. New York, NY: Harper Collins.
- McCloud, S. (1996, April 1996). Understanding Manga. *Wizard Magazine*, 56, 44-48.
- Morling, B., & Lamoreaux, M. (2008). Measuring Culture Outside the Head: A Meta-Analysis of Individualism—Collectivism in Cultural Products. *Personality and Social Psychology Review*, 12(3), 199-221. doi: 10.1177/1088868308318260
- Neff, W. A. (1977). *The Pictorial and Linguistic Features of Comic Book Formulas*. Doctoral Dissertation, University of Denver, Denver, CO.
- Nisbett, R. (2003). *The Geography of Thought: How Asians and Westerners Think Differently... and Why*. New York: Nicholas Brealy Publishing Ltd.
- Nisbett, R., & Masuda, T. (2003). Culture and Point of View. *Proceedings of the National Academy of Sciences*, 100(19), 11163-11170.
- Nisbett, R., & Miyamoto, Y. (2005). The influence of culture: holistic versus analytic perception. *Trends in Cognitive Sciences*, 9(10), 467-473.
- Schodt, F. L. (1983). *Manga! Manga! The World of Japanese Comics*. New York: Kodansha America Inc.
- Toku, M. (2001). Cross-Cultural Analysis of Artistic Development: Drawing by Japanese and U.S. children. *Visual Arts Research*, 27, 46-59.
- Toku, M. (2002). *Children's Artistic and Aesthetic Development: The Influence of Pop-Culture in Children's Drawings*. Paper presented at the 31st INSEA Convention, New York, NY.



# Early-Talker and Late-Talker Toddlers and Networks Show Different Word Learning Biases

**Eliana Colunga (eliana.colunga@colorado.edu)**

Department of Psychology and Neuroscience, 345 UCB  
Boulder, CO 80309-0345 USA

**Clare E. Sims (clare.holtpatrick@colorado.edu)**

Department of Psychology and Neuroscience, 345 UCB  
Boulder, CO 80309-0345 USA

## Abstract

In typical development, word learning goes from slow and laborious to fast and seemingly effortless. Typically developing 2-year-olds are so skilled at learning noun categories that they seem to intuit the whole range of things in the category from hearing a single instance named – they are biased learners. This is not the case for children below the 20<sup>th</sup> percentile on productive vocabulary (*late talkers*). This paper looks at the individual vocabularies and word-learning biases of late- and early-talking toddlers. Experiment 1 shows that neural networks trained on the vocabularies of individual late talkers learn qualitatively different biases than those trained on early talker vocabularies. Experiment 2 confirms the novel predictions made by the simulations about word learning biases in late- vs. early-talking children. The implications for diagnosis and intervention are discussed.

**Keywords:** Late talkers; early talkers; computational models; neural networks, word learning.

## Introduction

There is extraordinary variability in the vocabularies of very young children. A two-year-old in the lower 10<sup>th</sup> percentile may produce around 10 words whereas a two-year-old in the top 10<sup>th</sup> percentile will produce well over 300 (Fenson, Dale, Reznick, Thal, Bates, Hartung, Pethick, & Reilly, 1993). In general, the course of word learning proceeds from slow, effortful learning of nouns and of the range of things that belong in a category, to very rapid learning of object names. Indeed, typically developing 2-year-olds are so skilled at learning new nouns that they seem to intuit the whole range of things in a named category from a single naming experience. This is not necessarily the case for children below the 15<sup>th</sup>-20<sup>th</sup> percentile on productive vocabulary, or *late talkers*. Why do some children learn words quickly and early and others learn words slowly, maybe even showing effects that persist into adolescence? This paper looks at two possible contributing, and interrelated, factors: noun vocabulary composition and word learning biases

The evidence suggests that children are skilled noun learners because they know about the different kinds of properties that are relevant for categorizing different kinds of things. Typically-developing children show word learning biases that are specific to different kinds of things: they generalize names for solid objects by shape and names

for non-solid substances by material (e.g., Soja, Carey, & Spelke, 1991).

The evidence also suggests that children learn how to learn nouns – and learn how different kinds of properties are relevant for different kinds of things – as a consequence of learning names for things. Each noun the child learns appears to teach the child something general about how to learn new nouns that name things of that same kind, and critically, at the same time, this learned general knowledge constrains and facilitates the types of nouns the child will learn next. Through computational models and a study with toddlers, we show that even before they turn 2, late- and early-talker toddlers show different word learning biases.

## Vocabulary composition and word learning biases

The relationship between vocabulary composition and word learning biases has been typically characterized in one of two ways: abstract knowledge guides, facilitates and indeed allows word learning (e.g., Soja et al, 1991; Gelman & Bloom, 2000) or the words that have been learned give rise to, create, and in fact constitute generalized knowledge about word learning (e.g., Colunga & Smith, 2005, Samuelson, 2008). We would like to bypass the debate on whether word-learning biases are the egg to the vocabulary chicken or the other way around and focus instead on the interrelationship between these two factors.

In the domain of names for objects and substances, and in typical development, vocabulary structure and abstract knowledge in the form of kind-specific generalizations appear to be tightly coupled. First, the tendency to attend to shape in the specific context of naming artifacts emerges as children learn nouns, becoming particularly robust around the time children have between 50 to 150 nouns in their productive vocabulary (Gershkoff-Stowe & Smith, 2004). Second, the order of development of these word learning biases reflects the statistical structure of early noun vocabularies, (Samuelson & Smith, 1999; Colunga & Smith, 2005). Third, changing 17-month-olds' vocabulary composition by intensively teaching them names for artifacts yields an early bias to generalize names for artifacts by shape *and accelerates learning of object names outside of the lab*, causing a dramatic increase in vocabulary size for children in the experimental training group but not for those in the control groups (Smith, Jones, Landau, Gershkoff-Stowe & Samuelson, 2002). Fourth, computational models

trained on the structure of the average 30-month-old vocabulary, show word learning biases like those of young children when processing new objects (Colunga & Smith, 2005), and further the structure of the training set affects subsequent training, facilitating the learning of some sorts of categories but hindering others (Colunga, in prep). Altogether, these results suggest a developmental feedback loop between learning object names, developing biases to attend to the relevant properties for artifacts, and the learning of more object names.

## Late Talkers

Children below the 15<sup>th</sup>-20<sup>th</sup> percentile on normative measures of productive vocabulary size, so-called late talkers, are not a homogenous group in terms of their developmental outcomes: some catch up, and some show persistent delays (Rescorla, 2002, Rescorla, Roberts, & Dahlsgaard, 1997). However, Rescorla and colleagues argue against considering late talkers and typically developing children as distinct groups, and argue instead for conceptualizing them in terms of a “language endowment spectrum.” Importantly, although there is continuity in vocabulary measures at the group level, the outcome for individual children cannot be accurately predicted on the basis of vocabulary production or comprehension (e.g., Desmarais, Meyer, Bairati & Rouleau, 2008).

The literature briefly reviewed above suggests that, in typical development, the words a child knows and what the child knows about learning words in general go hand in hand, and that learning names for categories of things organized by shape speeds up learning nouns. However, this may not be the case for all children. Unlike typically developing children, late talkers do not systematically extend the name of a novel solid object to other objects that match it in shape, and in fact, in one study, almost half of the late talkers systematically extended the novel name of a solid object to others matching in texture rather than shape (Jones, 2003). Recent evidence suggests that the vocabularies of children of different language abilities may be structured differently (Colunga & Sims, 2011; Beckage, Smith & Hills, 2011). These findings suggest that late talkers may not just be limited in their production of object names (the measure that defines them as late talkers) but also deficient in the processes that subserve the acquisition of new words and in their knowledge about those categories. The crucial question, then, is whether these differences in vocabulary composition are differences that matter. Do the different nouns late- and early-talkers know yield different word learning biases? In two experiments we test the relationship between vocabulary composition and word learning biases, first in neural networks (Experiment 1) and then with 1-year-old toddlers in the lab (Experiment 2). For the purposes of this paper we will focus on contrasting the vocabularies of children on the two opposite ends of the spectrum, late talkers and early talkers.

If the differences in vocabulary structure can, to some extent, explain the differences in language ability, we would expect late talker vocabularies to yield different word

learning biases than early talker vocabularies. More specifically, we would expect early talker vocabularies to yield word learning biases that would facilitate the learning of a vocabulary structured like the MCDI – highlighting shape similarities for solids and material similarities for non-solids. In contrast, we would expect networks trained on late talkers’ vocabularies to generalize more variable word learning biases, and perhaps even biases that would be unhelpful in learning early vocabularies.

## Experiment 1

### Method

**Materials.** The vocabulary measure used was the Bates-MacArthur Communicative Developmental Inventory toddler version (MCDI) both to select children and to measure vocabulary composition. This is a parent checklist that asks parents to indicate the words that their child *produces* and although it is imperfect as a measurement instrument (Fenson, et al, 1994) it appears to be reliable and to be systematically related to children’s performances in a variety of laboratory measures of word learning, including especially their word-learning biases in the Novel Noun Generalization (NNG) task (e.g., Landau, et al, 1988).

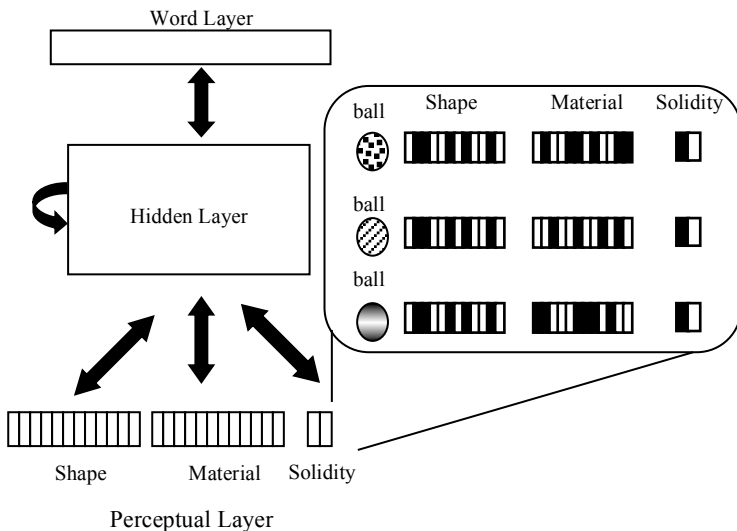
**Participants.** The vocabularies of 15 late talkers and of 15 early talkers were selected out of a pool of 148 parent-filled MCDI forms for children between 18-30 months of age. The criterion for inclusion was that there existed a vocabulary form from a child matching in age to within 5 days in both the late talker and the early talker groups. Late talkers fell under the 25<sup>th</sup> percentile; early talkers were above the 75<sup>th</sup> percentile according to the MCDI norms.

The ages for the two language groups ranged from 18.49 months to 28.26 months ( $M=23.14$  and  $23.15$  for late and early talkers respectively). Vocabulary sizes for the late talker group ranged between 15 and 425 words ( $M=132.53$ ); for the early talker group vocabulary size was between 158 and 664 words ( $M=457$ ).

The noun vocabularies for each individual child were characterized by looking at the proportion of nouns they knew for each of the following categories: 1) solid things alike in shape (e.g., spoon), 2) solid things alike in material (e.g., chalk), 3) solid things alike in both shape and material (e.g., penny), 4) non-solid things alike in shape (e.g., bubble), 5) non-solid things alike in material (e.g., milk), 6) non-solid things alike in both (e.g., jeans). Nouns in children’s vocabularies were classified as falling in each of these categories according to adult judgments made for each of the nouns in the MCDI reported in Samuelson & Smith, 1999. The training sets were then constructed to mimic the vocabulary composition of each child (see below).

**Architecture.** The computational models are a modified version of the ones Colunga & Smith, 2005. The main difference is that these networks were trained using the Leabra algorithm, an algorithm that combines Hebbian and error driven learning (O’Reilly, 1996), instead of Contrastive Hebbian Learning as in the original simulations.

The networks are organized as follows: Words are represented discretely (as single units) and are input on the Word Layer (Figure 1). Referents are represented as distributed patterns over several dimensions on the Perception Layer. For example, the shape and material of an object (say the roundness of a particular ball and its yellow rubbery material) are represented by an activation pattern along the Perception layer. Solidity and Non-solidity are represented discretely; one unit stands for Solid and another for Non-Solid. Finally, there is a hidden layer that is connected to all the other layers and to itself. These networks have been shown to model performance in an analog of the NNG Task when trained on vocabularies structured as those of the average 30-month-old.



**Figure 1: Architecture of the network and example input patterns.**

**Training.** The networks are trained with categories presenting the same correlational structure as each individual child's noun vocabulary. On each training trial, a word is paired with a referent. The patterns associated with each word are determined by adult judgments of the early noun corpus. For example, adults judged balls to be similar in shape but different in material. To simulate this, we randomly selected an input vector to represent ball shape. Then on individual training trials, we paired that pattern with the label ball and a randomly selected material pattern (Figure 1). We do this for each noun in the training set. Each network was trained in this way for its simulated vocabulary until they reached asymptotic (and near perfect) performance. This part of the simulation is intended to put into the networks the lexical knowledge that the individual child would bring to the laboratory NNG task.

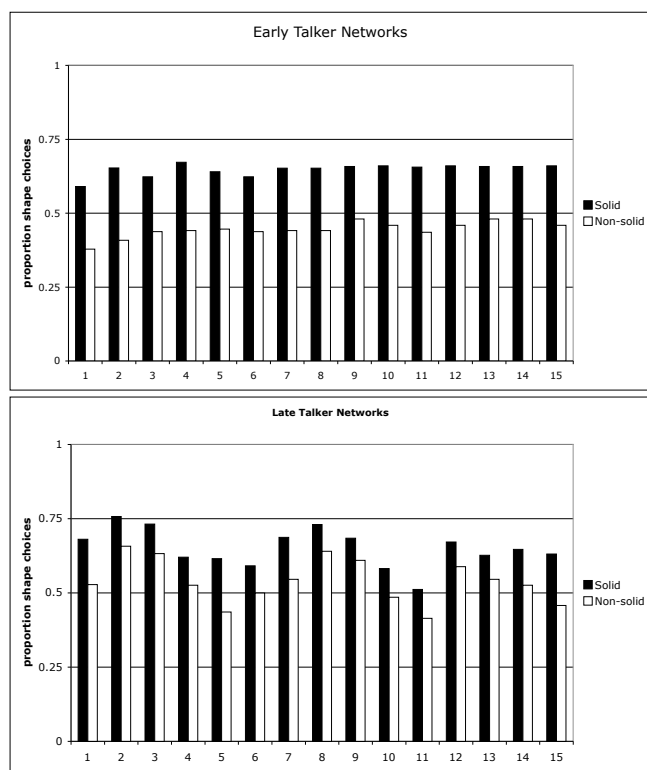
Because we are interested in the consequences of different vocabulary structures regardless of their size, all networks were trained to learn 24 nouns, proportionally structured like their corresponding child's vocabulary. Thus, the only difference between networks were the differences in vocabulary composition for each individual child.

**Testing.** The question is what sort of word learning bias will the networks learn given different vocabulary structures. We address this question in a virtual version of the NNG task. On each test trial of the virtual NNG task, we presented the network with three novel entities (one at a time) on the perception layer – an exemplar, and two choice items, one matching the exemplar in shape only and one matching in material only. For each of these three inputs, we recorded the resulting pattern of activation on the hidden layer. This is a measure of how the network represents these items. If the network emphasizes the shape of the item then the similarities of the internal representations for the exemplar and its shape matching choice should be *greater* than the similarity of the internal representations for the exemplar and the material matching choice. If, however, the internal representations highlight the material of the items, then the similarity of the internal representations for the exemplar and the shape matching choice should be *less* than the corresponding similarity of the exemplar and the material matching choice. We used these similarities along with Luce's choice rule to calculate probability of choice using these similarity measures in order to predict performance in the novel noun generalization task.

In previous work these models have been used to demonstrate the plausibility of the idea that the correlations in the early noun lexicon are sufficient to create second order generalizations – knowledge that any solid thing should be named by shape, and any non-solid thing should be named by material. The present simulations extend this work to variable vocabularies of individual children in the bottom and top ends of the language endowment spectrum.

## Results

The networks' predictions for each of the fifteen vocabularies of early talkers and late talkers are shown in Figure 2. In short, using a cut-off of at least two standard deviations above or below the 50% chance level mark, all networks in the early talker group show a shape bias for solids, and 12/15 early talker networks show a material bias for non-solids as well. In contrast, 12/15 late talker networks show a shape bias for solids and only 3/15 show a material bias for non-solids. Interestingly, 6/15 late-talker networks show a shape bias for *non-solids*, a novel prediction that has not been empirically tested so far. To further analyze the networks' performance, networks were classified according to the observed generalization patterns: *correct* if they showed a shape bias for solids and a material bias for non-solids, *half-right* if they showed the appropriate shape bias for solids but no consistent bias for material, or *wrong*, if they showed an incorrect overgeneralized shape biased to non-solids. A chi-square test showed these types of word learning biases were distributed differently in late talker and early talker networks,  $X^2(2,15)=14.743$ ,  $p=.0006$  (Yates'  $p=0.004$ ).



**Figure 2. Predicted proportion of shape choices for each of the early talker and late talker networks**

## Discussion.

The results of the simulations suggest that the differences in noun vocabulary composition between late- and early-talking children may result in differences in word learning biases. The word learning biases learned by these networks can be interpreted as predictions at the group level. First, the networks make a novel prediction about early talkers. A majority of the early talker networks show material biases for non-solids. Previous findings have shown that children at this age (18- to 30-month-olds) show a material bias for non-solids only when offered extra cues. For example, Soja (1992) showed older 2-year-olds have a material bias when offered supporting syntactic and visual cues, and Colunga & Smith (2005) showed an early material bias for non-solids that were presented in simple shapes for older 1-year-olds. However, children in general do not show a robust material bias for non-solids until around age 3 (Samuelson & Smith, 1999). Thus, this is a novel prediction that warrants testing: the networks predict that early talkers, unlike the general population, will show an early material bias for non-solids even without supporting cues.

The networks also make predictions about the patterns of novel noun generalizations one should expect to see in late talkers between 18 and 30 months of age. As a group, late talkers should show a shape bias for solids, with about half of them overgeneralizing this shape bias to non-solids as well. In Experiment 2 we test these predictions with late- and early-talker toddlers in the lab. Additionally, we run neural network simulations based on the composition of the

individual vocabularies of these children to replicate the pattern found in Experiment 1.

## Experiment 2

### Method

**Participants.** Nine late talkers (5 girls) and 8 early talkers (4 girls) between the ages of 18 and 22 months ( $M=19.4$ ) were selected out of 32 children recruited as part of a larger study. As in Experiment 1, the criterion for inclusion was scoring at or below the 25<sup>th</sup> percentile for late talkers and at or above the 75<sup>th</sup> percentile for early talkers. MCDI scores ranged from 5<sup>th</sup> to 20<sup>th</sup> percentile ( $M=8.9$ ) for the late talkers and between 75<sup>th</sup> and 99<sup>th</sup> percentile for early talkers ( $M=91$ ). Vocabulary sizes for the late-talker group ranged between 9 and 82 words ( $M=33$ ) and between 151 and 526 words ( $M=376.3$ ) for the early-talker group.

**Stimuli.** The stimuli consisted of a warm up set, a solid set and a non-solid set. The warm up set had an exemplar, a red plastic ball, two other balls (a tennis ball and a green and blue rubber ball), a plastic spoon, a toy carrot, and a toy cat.

The solid set consisted of an exemplar, an orange fuzzy round container, and 5 test items: two items matching the exemplar in shape alone (iridescent green bumpy round container and golden glittery round container), two items matching the exemplar in material (fuzzy blue irregular ring and fuzzy orange hook-like shape, and one matching in color (orange mesh polyhedron). The non-solid set was similarly structured and consisted of an exemplar (purple craft sand mixed into Noxzema in a rounded elongated x-like shape), two material matches (green sand + Noxzema in an asymmetric s-like shape and red sand + Noxzema in a lollipop-like shape), two shape matches (elongated x-like shapes made out of sawdust or purple shaving cream), and a color match (purple hair gel in an hourglass shape). All non-solids were presented on flat, square, plastic foam boards.

**Procedure.** In the warm-up phase, the experimenter presented all six toys to the child and allowed him or her to look at them and handle them for 30 seconds. Then the objects were removed and immediately placed back on the table outside of the child's reach. The child was then shown the exemplar ball and told, "look at this ball." Then they were asked to "get a ball" or get "another ball." If the child failed to retrieve a ball, the child was asked one more time, and finally was told "here's another ball," handed the ball, and allowed to get it one more time on request. If the child got one of the non-ball distracter items, they were told, "that's not a ball, that's a \_\_\_\_\_", then the distracter was replaced on the table, and the child was asked again for it.

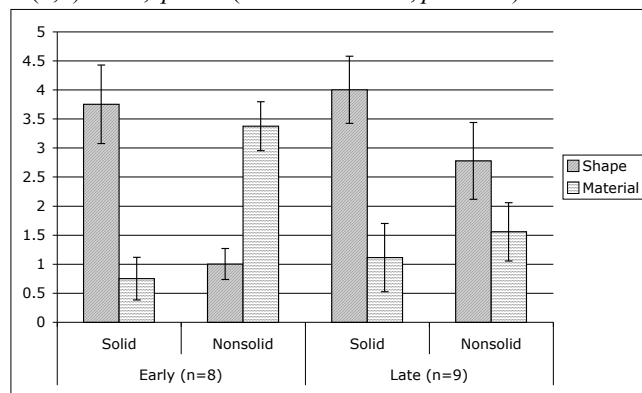
The procedure during the test phase with the solid and non-solid novel sets was the same, except that no feedback was offered. Children were shown the exemplar and told, "Look at this dax" and then asked to "get a dax" or "get another dax" for the solid set or "get more dax" or "get some dax" in the non-solid set. Children were asked to get another (or more) until they indicated that there were no more. Thus, solids were presented with count syntax supporting an object construal and non-solids were

presented with mass syntax supporting a substance construal (Soja, 1992). The solid set was always presented before the non-solid set, and there was a 5-minute break and a change in testing rooms in between the two test sets.

**Coding.** To incorporate order information into children's choices, and because all children made at least three choices for each test set, their choices were coded as follows: 3 points for the item that was 1<sup>st</sup> choice, 2 points for the 2<sup>nd</sup> choice, 1 point for the 3<sup>rd</sup> choice, and 0 points for other.

## Results

**Simulations.** The simulations based on individual children's vocabularies showed patterns comparable to the ones in Experiment 1. For the early talker networks, 6/8 showed shape and material biases, and the other two showed only a shape bias and no robust material bias. None of the early talker networks showed incorrect biases. For the late talker networks, all eight networks showed a shape bias for solids, but only one showed a material bias for non-solids<sup>1</sup>. In addition, 4/8 late talker networks showed an overgeneralized shape bias for non-solids. A chi-square test showed these types of word learning biases were distributed differently in late talker and early talker networks,  $X^2(2,8)=7.77$ ,  $p=.02$  (Yates'  $X^2=4.54$ ,  $p=0.103$ ).



**Figure 3. Scores for shape and material matches for solids and nonsolids for early- and late-talking toddlers.**

**Word learning biases.** The simulations in Experiment 1 predicted that early and late talkers would show different word learning biases, and predict specific patterns of novel noun generalizations for solids and non-solids for these two groups of children. We first look at the data of all children together and then evaluate the predictions for each language group. We submitted both groups of children's scores for the shape and material test items for the solids and non-solid sets to a 2 (language group: early, late) x 2 (solidity: solid, nonsolid) x 2 (dimension: shape, material) mixed ANOVA. Figure 3 shows the average score for the items that matched the exemplar in shape or material for the solid and nonsolid sets for both language groups. There was a main effect of dimension,  $F(1,29) = 4.77$ ,  $p = .045$ ,  $\eta^2 = .24$ ; overall shape matches had higher scores than material matches. There was

<sup>1</sup> One late talker child had no nouns, so no network was ran for that child. Thus, only 8 late talker networks were ran.

also a significant interaction between solidity and dimension,  $F(1,15) = 15.6$ ,  $p=.001$ ,  $\eta^2 = .51$ . Post-hoc tests showed that across both language groups, children were more likely to choose the shape over the material match for the solid set,  $t(16) = 4.03$ ,  $p=.001$ , but not for the nonsolid set,  $t(16) = -.613$ , n.s. The three-way interaction between language group, solidity, and dimension was marginally significant,  $F(1,15) = 4.33$ ,  $p=.055$ ,  $\eta^2 = .22$ .

The language-group-specific predictions made by the models were tested by analyzing the two groups separately. First, the prediction that early talkers would show a robust shape bias for solids and a robust material bias for nonsolids was confirmed by a 2 (solidity) x 2 (dimension) ANOVA revealing a two-way interaction between solidity and dimension,  $F(1,7) = 26.15$ ,  $p = .001$ ,  $\eta^2 = .78$ . Furthermore, planned comparisons (all two-tailed) showed that this interaction came from early talkers' shape bias for solids ( $t(7)=3.06$ ,  $p=.018$ ) and material bias for non-solids ( $t(7)=-4.46$ ,  $p=.003$ ). Second, a similar analysis on late talkers' scores revealed a main effect of dimension,  $F(1,8) = 5.5$ ,  $p=.047$ ,  $\eta^2 = .41$ , and no other main effects or interactions. Planned comparisons showed that late talkers had a shape bias for solids,  $t(8) = 2.57$ ,  $p=.033$ , but did not overgeneralize the shape bias to non-solids as a group,  $t(8) = 1.1$ , n.s. However, 4 out of the 9 late talkers in the study showed a shape bias for non-solids (a difference score of more than 3), and none of the early talkers did.

## Discussion

The results of Experiment 2 confirm the predictions of the simulations in Experiment 1. Early talkers show a shape bias for solids and a material bias for non-solids; late talkers show a shape bias for solids that can be over-generalized to non-solids. It is important to note that these predictions work at the group level and not at the level of individuals. For example, although four late talkers showed an overgeneralized shape bias for nonsolids in both the behavioral tasks and in the network simulations, these were not the same children; only two children showed this bias in both the simulations based on their vocabularies and their performance in the behavioral task. The behavioral task, and probably the vocabulary measure as well, lack the finesse to make predictions at the individual level based on a single data point. We return to this point in the general discussion.

The results of experiment 2 are in line with previous work noting a relationship between the number of nouns in a child's vocabulary and their word learning biases, but they extend it in important ways. The finding that early talkers show robust word learning biases for both solids and non-solids at not even two years of age is new. Although one might have predicted this pattern of results a priori from either the empiricist or the rationalist sides of the word learning debate, or even just from the idea that early talkers might excel across tasks, the prediction came from the models. Harder to predict without the networks, however, is the pattern found for the late talkers. In fact, at first glance it seems to contradict what we know about late talkers; that 2- to 3-year-old late talkers lack a shape bias while their same-

aged peers already have a well-established bias. However, the prediction from the networks, and our findings on the patterns of word learning biases in very young late talkers, before the age of 2, can help us understand the processes underlying word learning in general.

Gershkoff-Stowe and Smith (2004) followed eight children as they learned their first 100 nouns, looking at their word learning biases for solids and their vocabulary growth every three weeks. Their results show that as children's noun vocabulary increases, so does their attention to shape. They set the emergence of the shape bias at around the time children acquire 50 nouns. Our results suggest that this relationship may be different for late talkers. None of the late talkers in Experiment 2 reached the 50-noun mark (though a couple were on the cusp), and yet they overall showed a robust preference for shape for the solid set in our task. Curiously, although attention to shape increased with vocabulary size in Gershkoff-Stowe's study, the lower vocabulary group did show a preference of shape over material. This suggests an intriguing possibility: These models do not make a distinction between naming and non-naming contexts. It is possible that the shape preference for solids here is not a true shape bias, but rather an overgeneralized heightened attention to shape. The fact that about half of the late talkers showed an overgeneralized shape bias for non-solids suggests that this may be the case.

## General Discussion

The work presented here makes two main contributions. First, the findings of these two studies show that late talkers and early talkers know different sorts of nouns that lead to qualitatively different word learning biases. Importantly, these differences are shown within a computational model that has been previously shown to capture various aspects of novel noun learning, suggesting a promising use for process-level computational models. Efforts to tease apart the contributions of different factors to outcomes in late talkers have come up with some characteristics that put children at higher risk, but the underlying mechanisms are not well understood. The work of Ziegler and colleagues in the domain of dyslexia offers a good example of the potential for using computational models – and specifically models that operate at the mechanistic level – in simulating individual differences and further understanding subtypes in atypical development (Ziegler, Castel, Pech-Georgel, George, Aario, & Perry, 2008). Thus, the models presented here are a promising first step in leveraging computational models to aid in the understanding of why some late talkers catch up and others do not.

Second, these models represent an important extension over previous word-learning modeling efforts in that they go beyond modeling the performance of the mythical average child to making predictions about the performance of individual children, and of children who are both at the top and at the bottom of the vocabulary spectrum. In so doing, the simulations presented here make novel and testable predictions. The relationship between vocabulary

composition and word learning biases modeled here – the words you know determine the way you learn new words, which constrains and facilitates the words you will know next, and so on – opens a new way of thinking about computational models, to capture not only averages and not only individuals, but individual *trajectories*. If we can build computational models that can successfully capture this self-constructing developmental loop, the implications for early diagnosis, designing early interventions, and understanding the mechanisms that underlie word learning in typical and atypical development are far-reaching.

## References

- Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. *Psychological Review*, 112(2), 347–382.
- Colunga, E. & Sims, C. (2011). Early Talkers and Late Talkers Know Nouns that License Different Word Learning Biases, in *Proc. of Cog. Sci. Society*, 32. pp. 2250-2255.
- Desmarais, C. S., Meyer, F., Bairati, I., & Rouleau, N. (2008) Systematic review of the literature on characteristics of late-talking toddlers. *Intl. J. of Language and Communication Disorders*. 43(4) pp:361-389.
- Fenson, L., Dale, P., Reznick, J. S., Thal, D., Bates, E., Hartung, J., Pethick, S., & Reilly, J. (1993). *The MacArthur Communicative Development Inventories: User's guide and technical manual*. San Diego, CA:
- Gelman, S. A., & Bloom, P. (2000). Young children are sensitive to how an object was created when deciding what to name it. *Cognition*, 76(2), 91-103.
- Gershkoff-Stowe, L., & Smith, L.B.(2004). Shape and the first hundred nouns. *Child development* 75(4),1098-1114.
- Jones, S. S. (2003). Late talkers show no shape bias in a novel name extension task. *Dev. Science*, 6(5), 477-483.
- Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3, 299-321.
- Rescorla, L. (2002). Language and reading outcomes to age 9 in late-talking toddlers. *Journal of Speech, Language, and Hearing Research*, 45, 360-371.
- Rescorla, L., Roberts, J. & Dahlsgaard, K. (1997). Late talkers at 2: Outcome at age 3. *Journal of Speech & Hearing Research*, 40, 556-566.
- Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition*, 73(1), 1-33.
- Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. . (2002). Creating a shape bias creates rapid word learners. *Psychological Science*, 13, 13-19.
- Soja, N. N., Carey, S., & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meaning. *Cognition*, 38(2), 179-211.
- Ziegler, J. , Castel, C., Pech-Georgel, C., George, F., Alario, F-X., & Perry, C. (2008). Developmental dyslexia and the dual route model of reading: Simulating individual differences and subtypes. *Cognition*, 107, 151–178.

# Do You See What I'm Singing? Visuospatial Movement Biases Pitch Perception

Louise Connell (louise.connell@manchester.ac.uk)

Zhenguang G. Cai (zhenguang.cai@manchester.ac.uk)

School of Psychological Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, UK

Judith Holler (judith.holler@mpi.nl)

Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD, Nijmegen, Netherlands

School of Psychological Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, UK

## Abstract

The nature of the connection between musical and spatial processing is controversial. While pitch may be described in spatial terms such as “high” or “low”, it is unclear whether pitch and space are associated but separate dimensions or whether they share representational and processing resources. In the present study, we asked participants to judge whether a target vocal note was the same as (or different from) a preceding cue note. Importantly, target trials were presented as video clips where a singer sometimes gestured upward or downward while singing that target note, thus providing an alternative, concurrent source of spatial information. Our results show that pitch discrimination was significantly biased by the spatial movement in gesture. These effects were eliminated by spatial memory load but preserved under verbal memory load conditions. Together, our findings suggest that pitch and space have a shared representation such that the mental representation of pitch is audiospatial in nature.

**Keywords:** mental representation; pitch perception; music; gesture; spatial representation; metaphor

## Introduction

Musical and spatial processing are interlinked, but the exact nature and extent of the connection is controversial. People with amusia (i.e., an impaired ability to discriminate pitch) have corresponding spatial deficits in some reports (Douglas & Bilkey, 2007), but others have failed to replicate the association (Tillman et al., 2010; Williamson, Cocchini, & Stewart, 2011). People have been found to map musical pitch to vertical spatial locations (Pratt, 1930; Rusconi, Kwan, Giordano, Umiltà, & Butterworth, 2006), but they are also willing to map it to psychophysical luminosity and loudness (Hubbard, 1996; McDermott, Lehr, & Oxenham, 2008), and to words denoting emotion, size, sweetness, texture and temperature (Eitan & Timmers, 2010; Walker & Smith, 1984). Thus, while pitch may be described in spatial terms such as “high” or “low”, it remains unclear whether pitch and space are merely two amongst many associated dimensions or whether the representation of pitch is fundamentally spatial.

Pitch is a psychoacoustic property that corresponds to waveform frequency; its representation involves the primary auditory cortex but the full neural specification of pitch processing is still not well understood (e.g., Bendor, 2011). Space is a physical property of the three-dimensional body we occupy and the world through which we move, and is represented in a multimodal or supramodal system that takes input from vision, touch, and other perceptual modalities in

order to create a common spatial code (Bryant, 1992; Giudice, Betty, & Loomis, 2011; Lacey, Campbell, & Sathian, 2007). Numerous studies have shown that activating pitch also activates space along the vertical axis. A high-pitch prime leads people to explicitly relate it to a high spatial location (Pratt, 1930), and to implicitly attend to a visual target (Walker et al., 2010) or make a manual response (Rusconi et al., 2006) in a high spatial location. However, the above findings cannot distinguish between an *associative mapping* explanation, where representations of pitch and space are separate but linked, and a *shared representation* explanation, where pitch and space share common representational and processing resources.

According to an associative mapping explanation, the representation of musical pitch is purely auditory in nature. An individual's perception of a note's pitch would essentially comprise a modality-specific auditory representation of its sound frequency, and one would recall its pitch as a simulation (i.e., a partial replay of the neural activation that arose during experience: Barsalou, 1999) of that frequency. Perceiving a high pitch note rapidly activates a high spatial location because the two representational dimensions are directly associated, as are the dimensions of pitch and loudness (McDermott et al., 2008) or pitch and happiness (Eitan & Timmers, 2010). Notwithstanding these associations, pitch perception and discrimination itself remains an exclusively auditory matter. Conversely, a shared representation explanation for pitch/space effects would hold that the representation of musical pitch is audiospatial in nature. Here, an individual's perception of a note's pitch would comprise an audiospatial representation of both its sound frequency and its height on the vertical axis. One would then recall its pitch as an auditory and spatial simulation of that frequency and height. People may therefore be willing to map musical pitch to other dimensions because they all share a common spatial grounding (i.e., are mediated by space): for example, both loudness (Eitan, Schupak, & Marks, 2010) and emotional valence (Meier & Robinson, 2004) show similar effects to pitch in vertical space. Pitch perception and discrimination, therefore, is obligatorily audiospatial.

In the present studies, we aimed to distinguish between these two explanations by using a basic psychophysical task of pitch discrimination, where participants must judge whether a target vocal note is the same as (or different from) a preceding cue note. Importantly, target trials were presented as video clips where a singer sometimes gestured upward or downward while singing that target note, thus



providing an alternative, concurrent source of spatial information. Signal detection analysis then allowed us to isolate the response criterion of pitch discrimination (i.e., underlying bias towards the belief that pitch has or has not changed), for which the two accounts produced differing predictions. An associative mapping explanation of the pitch/space relationship would predict that a concurrent spatial stimulus should have no effect on response criterion. Because pitch representations are purely auditory, people can discriminate pitch on the basis of auditory frequency, regardless of what other processing might be taking place in the spatial system. Only in the shared representation account, where the audiospatial representation of musical pitch cannot be disentangled from the visuospatial representation of vertical gesture, would a criterion shift emerge. Because pitch representation is audiospatial, people cannot discriminate pitch without being biased by concurrent spatial movement.

### Experiment 1: Biasing Pitch

In this and the following experiments, participants watched target trials of an actor gesturing while singing a particular musical note. Gestures frequently and effectively communicate spatial information to recipients that goes beyond what is conveyed in speech (Graham & Argyle, 1975; Holler, Shovelton, & Beattie, 2009). Our nonlinguistic combination of gesture and pitch stimuli therefore allowed us to embed spatial information in a naturalistic context to which people are sensitive, but in a less obtrusive manner than pairing pitch with (for example) geometric shapes.

Our hypotheses were simple. If the shared representation explanation is correct and pitch representations are audiospatial, then spatial information in concurrent gesture should influence pitch discrimination in two specific ways. First, the spatial movement of gesture should bias participants towards the belief they had perceived a pitch movement (i.e., that the target note was different to the cue). Furthermore, participants should be sensitive to the direction of spatial movement, where downward gestures would make pitch appear lower in frequency, and upward gestures would make pitch appear higher in frequency. On the other hand, if the associative mapping explanation is correct, then none of these effects would appear.

### Method

**Participants** Thirty-two native speakers of English from the University of Manchester took part in the experiment. Five were replaced when funnel debriefing indicated they were aware of the potential effect the gestures could have had on their pitch discrimination judgements. All were right-handed, had no hearing impairment, and were non-musicians (i.e., not musically trained). They received course credits or £4 for participation.

**Materials** Target notes consisted of 16 vocal notes, sung by professional actors/singers on a major scale from A2 (110 Hz) to A3 (220 Hz) for the male actor, and from A3 (220

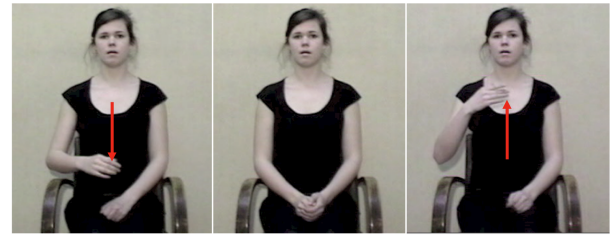


Figure 1: Stills from video stimuli, showing a singer gesturing downward, at rest with no gesture, and gesturing upward. Arrows indicate extent and direction of movement.

Hz) to A4 (440 Hz) for the female actor. The fundamental frequency of these vocal notes was a maximum of 17 cents (17% of a whole tone) away from the intended pitch. Each actor was filmed moving the right hand downward or upward for the duration of the note, (i.e., downward or upward gesture), or resting their hands naturally on the lap (i.e., no gesture) (see Figure 1). In order to ensure stimulus consistency in gestural and vocal behaviour, we overdubbed the best gesture videos with the best target notes, and ensured each final stimulus was a seamless synchronization of mouth movement, gesture movement, and sung vocal. All 48 target videos lasted 1.4 seconds.

Cue notes consisted of synthesized notes at the same fundamental frequencies as the target notes, created with Garageband software with the Classical Ensemble voice (which sounded like a mixed choir of male and female vocalists). We chose to use synthesized human voices in order to avoid the spatial pitch characteristics associated with musical instruments (e.g., horizontal for a piano, vertical for a clarinet), and to give cue notes a similar timbre to target notes while still allowing us to use the same type of cue for male and female actors' notes. We then edited the synthesized cue notes in Audacity to replicate the target sung notes' frequency exactly (same pitch), or to shift it one semitone up (higher pitch) or down (lower pitch).

Cue and target stimuli were paired so that each target note (accompanied by a downward gesture, no gesture, or upward gesture) was preceded by a cue note of the same pitch, higher pitch, or lower pitch. We divided these 144 pairs into two materials lists, where both lists included all 48 same-pitch pairs, and the remaining stimuli were distributed so each list had 24 higher-pitch and 24 lower-pitch pairs (i.e., an equal number of same and different pitch).

**Procedure** Participants were instructed that they should watch videos of professional actors singing musical notes, and, in each case, judge as quickly and accurately as possible whether the actor's sound was the same pitch as an earlier musical note. The experiment was run with Superlab 4.0 on a MacBook laptop, with videos displayed onscreen at approximately 14 x 10.5 cm. Participants were randomly assigned to one of the two material lists and were tested individually in a lab cubicle. In each trial, they first saw a fixation cross for 500 ms, then heard the synthesized note, and immediately afterwards saw the target note video. After



the video, a screen appeared with the prompt “SAME DIFFERENT”, and participants were asked to press the left-hand key on a response box if they thought the actor's sound was the same pitch as the earlier musical note, or the right-hand key if they thought it was a different pitch to the earlier musical note (left/right mapping to same/different responses counterbalanced). If participants pressed the “different” key, another screen appeared with the prompt “HIGHER LOWER”, and participants were asked to press the left-hand key if they thought the actor's sound was a higher pitch than the earlier musical note, or the right-hand key if they thought it was a lower pitch (left/right mapping to higher/lower responses counterbalanced). There was a blank of 500 ms between trials.

Within each lost of materials, stimuli were arranged into six blocks so that each of the 16 target notes appeared once per block (gestures counterbalanced). The order of blocks was fixed but presentation of trials within a block was randomized per participant. Participants performed 4 practice trials before the main experiment, and the whole procedure lasted for about 15 minutes.

**Design & Analysis** We ran two stages of analysis of variance, each with a single within-participants factor of gesture (downward, no-gesture, upward) and effect sizes reported as partial eta-squared ( $\eta_p^2$ ). First, signal detection analysis examined performance on the same/different judgments to determine if gesture affected people's response bias and sensitivity in pitch discrimination. “Different” responses to different-pitch targets constituted hits, and those to same-pitch targets constituted false alarms. For each participant, we then calculated criterion  $c$  (criterion or bias) and  $d'$  (sensitivity) statistics for each gesture condition (e.g., Stanislaw & Todorow, 1999).

Second, we examined the trajectory of error to determine whether downward gestures made notes seem lower than they really were (and upward gestures higher). Each error in the same/different and higher/lower judgments represented an upward or downward response trajectory: for example, a downward trajectory was one where (1) a same-pitch target was judged to be lower in pitch, (2) a higher-pitch target was judged to be the same pitch, or (3) a higher-pitch target was judged to be lower in pitch. For each participant, we calculated the proportion of downward errors out of all errors in each gesture condition. Four participants with empty cells (i.e., perfect accuracy in one or more conditions) were excluded from trajectory analysis.

## Results & Discussion

People found the pitch discrimination task moderately difficult, with overall accuracy of 71.1%. Signal detection analysis supported the shared representation prediction that the spatial movement in gesture would affect pitch discrimination. There was a criterion difference between gesture types,  $F(2, 62) = 4.57$ ,  $p = .014$ ,  $\eta_p^2 = .129$ , as shown in Figure 2 (left panel). Most trials showed a bias towards “same” responses (i.e.,  $c > 0$ ), but planned comparisons showed this bias was weaker for downward ( $p$

$= .006$ ,  $\eta_p^2 = .187$ ) and upward ( $p = .011$ ,  $\eta_p^2 = .156$ ) gestures compared to when notes were unaccompanied by gesture. Upwards and downward gestures had the same response bias ( $p = .999$ ). Participants' increased propensity to make “different” responses in the presence of gesture did not affect their overall sensitivity in pitch discrimination,  $F(2, 62) = 2.04$ ,  $p = .139$ ,  $\eta_p^2 = .062$ , with equivalent performance in no-gesture ( $d' = 1.79$ ), downward ( $d' = 1.99$ ) and upward ( $d' = 1.83$ ) gesture conditions.

Analysis of error trajectory also followed predictions (see Figure 3, left panel). The nature of errors that people made was influenced by gesture,  $F(2, 54) = 9.23$ ,  $p < .001$ ,  $\eta_p^2 = .255$ . Specifically, planned comparisons showed that, relative to the no-gesture condition, downward gestures increased the number of downward trajectory errors ( $p = .007$ ,  $\eta_p^2 = .205$ ) while upward gestures reduced them ( $p = .043$ ,  $\eta_p^2 = .105$ ).<sup>1</sup>

## Experiment 2: Spatial Memory Load

If the shared representation explanation of pitch/space effects is correct, then the criterion shift and error trajectory in Experiment 1 emerge from an overlapping spatial representation of gestural movement and pitch. A spatial memory load should therefore attenuate these effects by occupying resources required for audiospatial pitch discrimination. Holding a spatial load in memory should remove the biasing effect of spatial movement on pitch discrimination, meaning that people will remain quite liberal in their tendency to assume that notes are the same. Consequently, the direction of spatial movement should no longer drive the trajectory of error to the same extent.

## Method

**Participants** Thirty-two new participants took part under the same criteria as Experiment 1. Five participants were replaced for awareness of the gesture effect. All had adequate recall of the spatial memory load (i.e., correctly recalled four or more out of six grids, see Materials).

**Materials** Stimuli were as per Experiment 1. In addition, items in the spatial memory task consisted of six different 3-by-3 grids (plus one for practice) in which five random cells had been filled with an X.

**Procedure** Instructions were identical to Experiment 1 except that participants were asked to hold in memory a visually-presented spatial grid during each block of the task. Before each of the six blocks, participants saw a spatial grid onscreen and could study it until they were satisfied they had memorised it. At the end of the block, participants were

<sup>1</sup>Although our participants were not musically trained, this fact did not preclude some level of knowledge about music; at the end of the experiment, we therefore gave participants a questionnaire to probe their exposure to music instruction (e.g., experience of playing a musical instrument, ability to distinguish pitch differences in staff notation). Musical knowledge was unrelated to either global response criterion  $r(30) = -.055$ ,  $p = .765$ , or downward error trajectory  $r(26) = -.182$ ,  $p = .354$ , though it did correlate positively with overall sensitivity  $r(30) = .597$ ,  $p < .001$ .

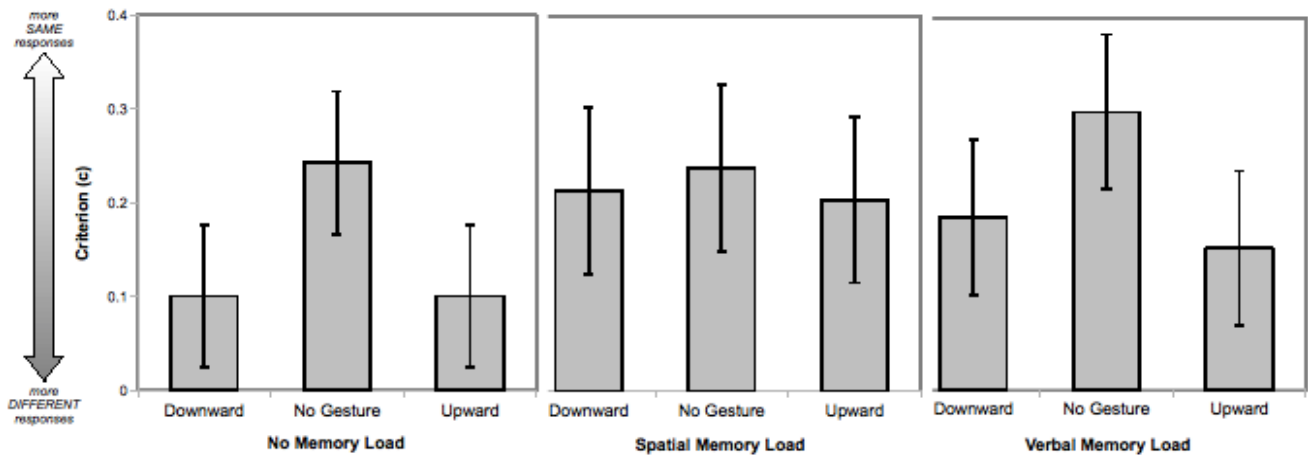


Figure 2: Response criterion in pitch discrimination (i.e., bias towards belief that target note was same as / different to cue) per gesture condition in Experiment 1-3. Error bars show within-participants 95% CI.

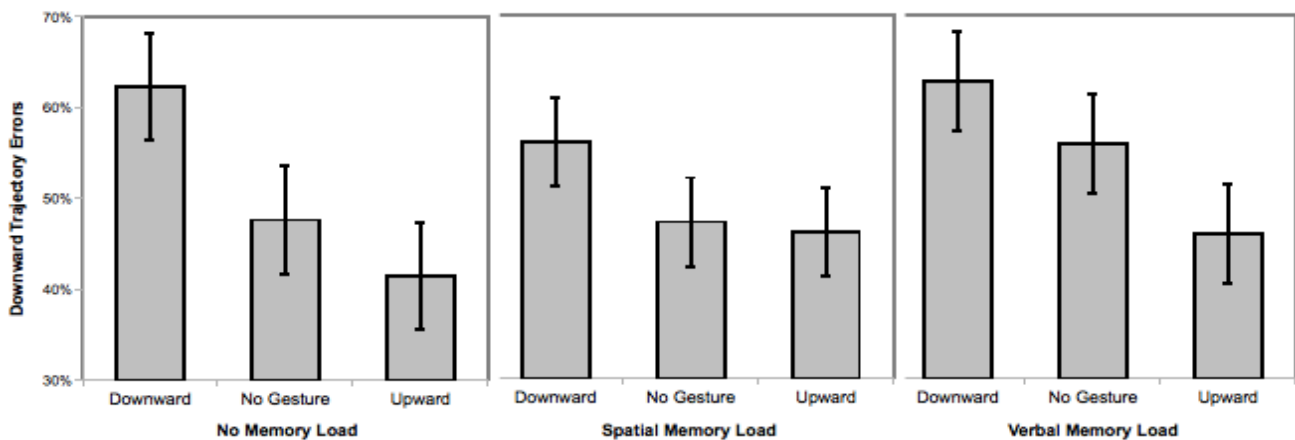


Figure 3: Proportion of pitch discrimination errors that expressed a downward trajectory (i.e., where participants thought the note was lower in pitch than reality) per gesture condition in Experiments 1-3. Error bars show within-participants 95% CI.

asked to recall the grid by drawing the positions of the Xs on a blank grid; these were later coded for accuracy (a grid must be perfectly recalled to qualify as an accurate response). The experiment lasted approximately 20 minutes.

**Design & Analysis** As in Experiment 1. Six participants with perfect accuracy in one or more conditions were excluded from trajectory analysis.

## Results & Discussion

Overall accuracy was similar to Experiment 1 at 73.8%. Signal detection analysis confirmed the predictions of the shared representation account that a spatial memory load would eliminate the biasing effect of spatial movement on pitch discrimination. There was no longer any criterion difference between gesture types,  $F(2, 62) = 0.15$ ,  $p = .856$ ,  $\eta_p^2 = .005$  (see Figure 2, centre panel): a similar bias towards “same” responses appeared for downward, upward and no-gesture conditions (all  $p$ s  $> .3$ ,  $\eta_p^2$ s  $< .009$ ). Sensitivity of pitch discrimination was unaffected by gesture,  $F(2, 62) = 1.11$ ,  $p = .176$ ,  $\eta_p^2 = .054$ : no-gesture  $d'$

$= 2.01$ , downward gesture  $d' = 1.92$ , upward gesture  $d' = 1.88$ .

Analysis of error trajectory showed attenuated effects compared to Experiment 1 (see Figure 3, centre panel). Spatial movement in gesture had an influence on the direction of error,  $F(2, 50) = 3.61$ ,  $p = .034$ ,  $\eta_p^2 = .191$ . Downward gestures led to more downward trajectory errors than no gesture ( $p = .034$ ,  $\eta_p^2 = .127$ ), but upward gestures did not reduce their occurrence relative to no gesture ( $p = .406$ ,  $\eta_p^2 = .002$ ).<sup>2</sup>

## Experiment 3: Verbal Memory Load

While the results of Experiment 1 support the shared representation account of pitch/space effects, it is possible that participants were silently labelling the pitch of the target notes as “higher” or “lower” in preparation for the discrimination task. The spatial movement in gesture could then have interacted with the representation of this verbal

<sup>2</sup> Musical knowledge was again unrelated to response criterion  $r(30) = -.196$ ,  $p = .282$ , and error trajectory  $r(24) = .084$ ,  $p = .984$ , and correlated with overall sensitivity,  $r(30) = .546$ ,  $p = .001$ .

label rather than inducing a bias in pitch discrimination itself. We therefore examined the origin of the criterion shift by replicating the task while participants held a verbal load in memory to block a linguistic labelling strategy. If the shared representation explanation is correct, then the criterion shift of Experiment 1 should emerge unscathed. Furthermore, if the criterion shift re-emerges under a verbal memory load, it will verify that the cancelled effects in Experiment 2 were not due to generic processing difficulties under memory load conditions but rather were specific to spatial content.

## Method

**Participants** Thirty-two new participants took part under the same criteria as Experiment 1. Three participants were replaced for inadequate recall of the verbal memory load (i.e., anyone who recalled fewer than four out of six diphone sequences, see Materials).

**Materials** Stimuli were as per Experiment 1. In addition, items in the verbal memory task consisted of six different sequences of three nonsense diphones (e.g., [tɛ kæ vo]); one further sequence was used for practice. Each sequence was recorded by a male native speaker with clear enunciation.

**Procedure** Instructions were identical to Experiment 1 except that participants were asked to hold in memory an auditorily-presented diphone sequence during each block of the task. Before each of the six blocks, participants listened to a diphone sequence three times and repeated it back to the experimenter (second author); if there were any errors in repetition, the experimenter enunciated the sequence again until participants got it right. At the end of each block, participants recalled aloud the memorised diphone sequence to the experimenter who transcribed it and later coded it for accuracy (a sequence must be perfectly recalled to qualify as an accurate response). Participants were familiarised with diphone recall during the practice session. The experiment took approximately 20 minutes to complete.

**Design & Analysis** As in Experiment 1. Three participants with perfect accuracy in one or more conditions were excluded from trajectory analysis.

## Results & Discussion

Overall accuracy was similar to Experiment 1 at 69.3%. Signal detection analysis replicated the findings of Experiment 1, and confirmed that the biasing effect of spatial movement on pitch discrimination was not due to a verbal labelling strategy. Figure 2 (right panel) shows the criterion difference emerged between gesture types,  $F(2, 62) = 3.39, p = .040, \eta_p^2 = .098$ . As before, the bias towards “same” responses was weaker for downward ( $p = .040, \eta_p^2 = .095$ ) and upward ( $p = .009, \eta_p^2 = .168$ ) gestures compared to when notes were unaccompanied by gesture. Upwards and downward gestures had the same response bias ( $p = .549$ ). Since these “different” responses were distributed across both correct and incorrect trials,

sensitivity of pitch discrimination did not change with gestural movement,  $F(2, 62) = 1.78, p = .176, \eta_p^2 = .054$ , with equivalent performance in no-gesture ( $d' = 1.68$ ), downward ( $d' = 1.89$ ) and upward ( $d' = 1.84$ ) gesture conditions.<sup>3</sup>

Analysis of error trajectory again replicated Experiment 1 (see Figure 3, right panel). The nature of errors in pitch discrimination was influenced by the spatial movement in gesture,  $F(2, 56) = 6.62, p = .003, \eta_p^2 = .191$ . Downward gestures marginally increased the frequency of downward trajectory errors compared to no gesture ( $p = .052, \eta_p^2 = .092$ ) while upward gestures reduced their occurrence ( $p = .034, \eta_p^2 = .115$ ).

## General Discussion

In the present paper, we show that concurrent visuospatial movement biases pitch discrimination. Viewing upward and downward gestures biased people towards believing they had perceived a change in pitch, despite an underlying tendency to assume that all notes were the same. Indeed, when we examined the pattern of errors that people made, we found that the direction of gesture was also driving the direction of error: downward gestures made notes seem lower in pitch than they really were, and upward gestures made notes seem higher in pitch than they really were. These effects were not due to a verbal labelling strategy as they were preserved under verbal memory load. However, their disappearance under spatial memory load conditions indicates that the biasing effect is spatial in origin. Together, these findings support the shared representation explanation for the relationship between pitch and space.

When people hear a musical note, its pitch is not just represented in the auditory modality. Rather, its representation is audiospatial, in that it comprises both an auditory and spatial representation of the note's frequency. However, things become more complicated when people watch someone singing a note. On the one hand, if the singer remains still, then the same story applies: the audiospatial representation still reflects the note's pitch. But, on the other hand, if the singer gestures with an upward or downward movement, then both the visual gesture and auditory note require representational resources in the vertical spatial axis. Hence, the spatial information in the gesture is co-perceived with that in the note, and results in an audiospatial representation of the note's pitch that has been modulated by the direction of spatial movement in the gesture.

While many previous studies have examined the relationship between pitch and vertical space, they could not determine the nature of pitch representation because both associative mappings and shared representations would lead a pitch stimulus to prime its associated spatial location and facilitate motor responses to that location (e.g., Rusconi et al., 2006). However, a mapping from high pitch to high spatial location would be static, and could not explain why

<sup>3</sup> Musical knowledge was again unrelated to response criterion  $r(30) = -.040, p = .828$ , or error trajectory  $r(27) = .044, p = .820$ , and correlated with overall sensitivity,  $r(30) = .394, p = .026$ .

the spatial movement in gesture biased participants towards believing they had perceived a movement in pitch. A dynamic, shared representation of pitch and space, where pitch is represented not only in terms of spatial position but also movement and direction, is consistent with our results.

There are several possibilities as to how and why musical pitch is represented in vertical space, and not in some other spatial dimension. When speaking, producing a pitch higher than normal voice frequency moves the larynx upward in the throat, and producing a lower pitch moves it downward. Furthermore, breathing from the top of the lungs by raising and lowering the shoulders tends to produce higher-pitch vocal notes, while breathing from the bottom of the lungs by tensing and relaxing the thoracic diaphragm tends to produce lower-pitch, resonant notes. Thus, cumulative experience of our own voices provides a possible vertical grounding for vocal pitch, which could then generalize to pitch of other people's voices or musical instruments, and so on. While some have claimed the appearance of pitch/space effects in young infants means the connection between domains is innate (Walker et al., 2010), even 3-4 month old babies have considerable experience of vocalisation. A conservative estimate of one hour per day crying, fussing etc. (e.g., Michelsson, Rinne, & Paajanen, 1990) provides a 4-month old infant with over 100 hours experience of vocal pitch under various body configurations. Since infants of that age can learn statistical regularities in the environment with only a few minutes' exposure (Kirkham, Slemmer & Johnson, 2002), it seems premature to assume they could not have learned to represent pitch spatially.

Future research will need to determine whether pitch/space effects emerge from a learned or innate mechanism, but, whatever its origin, the present paper demonstrates that pitch is fundamentally audiospatial. The nature of the link between musical and spatial processing is one of shared representation.

## Acknowledgments

This work was supported by a research project grant from the Leverhulme Trust (F/00 120/CA).

## References

- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavior and Brain Sciences*, 22, 577–660.
- Bendor, D. (2011). Does a pitch center exist in auditory cortex? *Journal of Neurophysiology*, 107, 743–746.
- Bryant, D. J. (1992). A spatial representation system in humans. *Psychology*, 3(16), space 1.
- Douglas, K. M., & Bilkey, D. K. (2007). Amusia is associated with deficits in spatial processing. *Nature Neuroscience*, 10, 915–921.
- Eitan, Z., Schupak, A., & Marks, L. E. (2008). Louder is higher: Cross-modal interaction of loudness change and vertical motion in speeded classification. *Proceedings of the 10th International Conference on Music Perception and Cognition* (pp. 1-10). Adelaide, Australia: Causal Productions.
- Eitan, Z., & Timmers, R. (2010). Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition*, 114, 405–422.
- Giudice, N. A., Betty, M. R., & Loomis, J. M. (2011). Functional equivalence of spatial images from touch and vision: Evidence from spatial updating in blind and sighted individuals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 621–634.
- Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology*, 10, 57–67.
- Holler, J., Shovelton, H., & Beattie, G. (2009). Do iconic hand gestures really contribute to the communication of semantic information in a face-to-face context? *Journal of Nonverbal Behavior*, 33, 73–88.
- Hubbard, T. L. (1996). Synesthesia-like mappings of lightness, pitch, and melodic interval. *American Journal of Psychology*, 109, 219–238.
- Kirkham, N.Z., Slemmer, J.A., & Johnson, S.P. (2002). Visual statistical learning in infancy: Evidence of a domain general learning mechanism. *Cognition*, 83, B35–B42.
- Lacey, S., Campbell, C., & Sathian, K. (2007). Vision and touch: Multiple or multisensory representations of objects? *Perception*, 36, 1513–1521.
- McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2008). Is relative pitch specific to pitch? *Psychological Science*, 19, 1263–1271.
- Meier, B. P., & Robinson, M. D. (2004). Why the sunny side is up: Associations between affect and vertical position. *Psychological Science*, 15, 243–247.
- Michelsson, K., Rinne, A., & Paajanen, S. (1990). Crying, feeding and sleeping patterns in 1 to 12-month-old infants. *Child: Care, Health and Development*, 16, 99–111.
- Nygaard, L. N., Herold, D. S., & Namya, L. L. (2009). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive Science*, 33, 127–146.
- Pratt, C. C. (1930). The spatial character of high and low tones. *Journal of Experimental Psychology*, 13, 278–285.
- Rusconi, E., Kwan, B., Giordano, B. L., Umiltà, C., & Butterworth, B. (2005). Spatial representation of pitch height: The SMARC effect. *Cognition*, 20, 1–17.
- Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments & Computers*, 31, 137–149.
- Tillmann, B., Jolicœur, P., Ishihara, M., Gosselin, N., Bertrand, O., et al. (2010). The amusic brain: Lost in music, but not in space. *PLoS ONE*, 5(4), e10173.
- Walker, P., Brenner, G., Spring, J., Maststock, K., Slater, A., & Johnson, S. (2010). Preverbal infants' sensitivity to synaesthetic cross modality correspondences. *Psychological Science*, 21, 21–25.
- Walker, P., & Smith, S. (1984). Stroop interference based on the synaesthetic qualities of auditory pitch. *Perception*, 13, 75–81.
- Williamson, V., Cocchini, G. & Stewart, L. (2011). The relationship between pitch and space in congenital amusia. *Brain and Cognition*, 76, 70–76.

# Flexible Shortcuts: Linguistic Distributional Information Affects both Shallow and Deep Conceptual Processing

Louise Connell (louise.connell@manchester.ac.uk)

School of Psychological Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, UK

Dermot Lynott (dermot.lynott@manchester.ac.uk)

Decision and Cognitive Sciences Research Centre, Manchester Business School, University of Manchester  
Booth Street West, Manchester M15 6PB, UK

## Abstract

Previous research has shown that people use both embodied perceptual simulations and linguistic distributional knowledge during conceptual processing, with linguistic information especially useful for shallow tasks and rapid responding. Using two conceptual combination tasks, we show that this linguistic shortcut is evident in both shallow and deep conceptual processing of novel stimuli. Specifically, in both shallow sensibility judgement and deep interpretation generation tasks, people use the linguistic shortcut as a “quick and dirty” guide to whether the concepts are likely to combine in a coherent situated simulation. Linguistic distributional frequency predicts both the likelihood and timecourse of rejecting a novel word compound as nonsensical or uninterpretable. However, it only predicts the timecourse of successful processing in shallow sensibility judgement because deeper interpretation generation requires conceptual processing in the simulation system.

**Keywords:** conceptual combination; linguistic distributional information; embodied cognition; simulation.

## Introduction

The embodied simulation view of conceptual representation holds that the same neural systems that are responsible for representing information during perception, action, and introspection are also responsible for representing (or *simulating*) the same information during conceptual thought (e.g., Barsalou, 1999; Glenberg, 1997). Furthermore, concepts do not exist in a representational vacuum, but rather are situated within a broader situational context that includes perceptual, motor, affective and social information on how that concept has been experienced in the past (e.g., Barsalou & Wiemer-Hastings, 2005; Lynott & Connell, 2010a). A *cactus*, for example, can potentially include visual information (e.g., its green colour and prickly surface), tactile information (e.g., the sharpness of its spines), and affective information (e.g., negative valence for anyone who has spent days picking spines from skin), all situated relative to other concepts (e.g., in a desert location or as a pot plant on a kitchen windowsill).

However, the simulation system does not act alone. People are sensitive to distributional, statistical patterns in language and the wider environment, and this sensitivity provides a powerful generalised learning mechanism from early infancy (Aslin et al., 1998; Kirkham et al., 2002). Even in adults, the linguistic system contains statistical

distributional information in a dynamic web of word-to-word (and phrase-to-phrase) associations that is powerful enough to support superficial strategies in a broad range of linguistic and conceptual tasks (e.g., Barsalou, Santos, Simmons & Wilson, 2008; Louwerse & Jeuniaux, 2008; Lynott & Connell, 2010a). The linguistic and simulation systems are closely interconnected and mutually supportive; linguistic information can activate simulation information, which may in turn activate further linguistic information, and so on. For example, when the word “cactus” is encountered, closely related linguistic tokens such as “prickly” and “sharp” will be activated, which will in turn begin to activate their relevant grounded representations in the simulation system, thus drawing attention to the visual and haptic modalities. Because their structures are both based on experience, the linguistic and simulation systems mirror each other to a certain extent, which suggests that information from language alone can approximate the perceptual, motor, affective, etc. content of concepts. Supporting this view, Louwerse and Connell (2011) have shown that linguistic distributional information is capable of distinguishing words on the basis of their perceptual modality. Words like *rustling*, *glistening*, and *freezing* refer to object properties in particular perceptual modalities (i.e., auditory, visual, and haptic) and occur in language with particular usage patterns. Louwerse and Connell showed that statistical analysis of these distributional patterns (based on 5-gram co-occurrence frequencies from a large corpus) produced three clusters that corresponded to auditory, visuohaptic and olfactogustatory modality groups. In other words, although auditory words were distinct, distributional information could not distinguish vision from touch, nor smell from taste. These three “linguistic modalities” (i.e., modality-specific clusters within the linguistic system) of auditory, visuohaptic and olfactogustatory words are therefore a coarse-grained approximation of the perceptual reality of five modalities. Linguistic distributional information is, at best, a blurred mirror of the simulation system.

The essential difference between the two systems is that the linguistic system is best for “quick and dirty” judgements, while the simulation system is best for deeper conceptual processing. When a word such as “cactus” is heard or read, both systems are kickstarted but the linguistic system peaks in activation (e.g., spreads activation to other

tokens “prickly”, “sharp”, and so on) before the simulation system peaks (e.g., forms a visual, haptic, affective situated simulation of a *cactus*). The linguistic system thus has the potential to act as a shortcut and provide a response before the relatively more expensive simulation system is fully engaged. Support for this idea comes from Louwerse and Connell (2011), who compared the abilities of the linguistic and simulation systems to predict modality switching costs in property verification tasks. Switching costs refer to the finding that people are slower to confirm that a perceptual property is true of an object (e.g., auditory *leaves can be rustling*) when it follows a property from a different modality (e.g., visual *dew can be glistening*), and this processing cost is assumed to arise from the re-allocation of attention between modality-specific areas during perceptual simulation of the object property in question (Pecher, Zeelenberg & Barsalou, 2003). When Louwerse and Connell examined whether switching costs were best predicted by “linguistic modalities” (i.e., auditory, visuohaptic, and olfactogustatory word clusters) or actual perceptual modalities (i.e., auditory, gustatory, haptic, olfactory, and visual categories, based on human ratings), they found that the linguistic shortcut was the best predictor of fast responses, whereas perceptual simulation of five modalities was the best predictor of slow responses. In short, the linguistic system offers a fuzzy approximation that can provide an adequate heuristic in certain tasks, whereas the simulation system provides representational precision for more complex and precise conceptual processing.

## The Current Study

Although Louwerse and Connell's (2011) study offers important evidence for the role of the linguistic shortcut conceptual processing, it is based on the retrieval of familiar information that is always expected to be successful. Most of human cognition is not like that, however. In order to function in a normal environment, we must be able to represent new concepts and process unfamiliar information, and work within the constraint that our conceptual processing is not always successful. Indeed, one of the key issues of a cognitive system with limited resource capacity is that not everything *should* be processed; a cognitive triage mechanism – an automatic means to determine whether it is worth expending precious representational and executive resources on a particular conceptual task, or whether it should be abandoned pending further clarification / information – would offer an invaluable aid to efficient functioning. A strong test of the linguistic shortcut hypothesis would therefore predict that use of the shortcut should be evident in the processing of (1) novel stimuli, (2) for successful responses in relatively shallow conceptual tasks, and (3) for apparent failures where a process is halted as not worth the effort, regardless of the depth of processing ostensibly involved in the task.

In the present experiments, we examined the role of the linguistic shortcut in conceptual combination using both shallow and deep processing tasks. Conceptual combination

is the process of understanding novel word compounds such as *cactus beetle* or *elephant complaint*, and is predicated upon the inherently constructive nature of cognition that allows us to represent new concepts by mentally manipulating old ones. For example, a *cactus beetle* may be represented as a beetle that feeds on cacti, or as a green and prickly beetle; both are equally valid end products of a successful combination process. Recently, Lynott and Connell (2010a) proposed the Embodied Conceptual Combination (ECCo) theory, which argues for a distinct role for the linguistic system during conceptual combination that complements that of the simulation system. Specifically, if the two nouns in a compound have little shared statistical, distributional history from language use, then the linguistic system offers people a reasonable heuristic for rejecting the compound as incomprehensible without expending much cognitive effort in attempting to combine the concepts. In contrast, if the nouns have frequently been encountered in close proximity to one another, then the linguistic system offers people a reasonable heuristic for accepting that the concepts can probably be combined in a shared, situated simulation.

Both sensibility judgement and interpretation generation tasks are commonly used in conceptual combination studies, but they differ in the required depth of processing (Lynott & Connell, 2010a). Sensibility judgement (Experiment 1) is relatively shallow because it simply asks people whether or not a particular compound makes sense. Interpretation generation (Experiment 2) is relatively deep because it asks people whether or not they can think of a meaning for a particular compound, and, if so, to specify the meaning. We therefore expected the linguistic system to play a differential role in conceptual combination according to task requirements: as a shortcut for both accepting and rejecting compounds in sensibility judgements, but only for rejecting compounds in interpretation generation because successful processing requires detailed representation in the simulation system.

## Experiment 1: Sensibility Judgement

In this experiment, we presented people with novel noun-noun compounds in a forced-choice sensibility judgement task, where they pressed “yes” if they thought the compound phrase made sense, and pressed “no” if they thought it was nonsense. Similar methods have been used in a number of previous conceptual combination studies (e.g., Gagné & Shoben, 1997; Estes, 2003; Tagalakis & Keane, 2006). We measured response times to press both “yes” (i.e., accept as sensible because of successful combination) and “no” (reject as nonsense because of failed combination) keys. Following ECCo's proposal regarding the nature of the linguistic shortcut in sensibility judgements, we predicted inverse effects for acceptance and rejection of compounds. Linguistic distributional frequency (i.e., how frequently the two nouns have shared a context) should be *negatively* related to acceptance rates and times because high-frequency compounds will quickly appear sensible: the

linguistic shortcut allows people to assume the concepts in question can combine merely because their two nouns have been frequently juxtaposed. In contrast, linguistic distributional frequency should be *positively* related to rejection times because low-frequency compounds will quickly appear nonsensical: the linguistic shortcut allows them to be dismissed out of hand rather than attempting a costly and potentially pointless combination effort in the simulation system.

## Method

**Materials** Forty one noun-noun compounds were used in this study: 27 novel test items and 14 lexicalised filler items. Test items comprised novel noun-noun compounds (e.g., *octopus apartment*, *elephant complaint*, *whale knife*) with a British National Corpus phrase frequency greater than 20 (BNC, 2001), and featured a range of concept types (i.e., artifacts, natural kinds, abstract concepts). Filler items were lexicalised noun-noun compounds (e.g., *hospital wing*, *guerrilla warfare*) with a BNC frequency greater than 20, and were included to provide a baseline of highly sensible combinations to ensure that participants attended to the task.

In order to approximate the linguistic distributional information available to novel compounds, we carried out a corpus analysis using the Web 1T 5-gram corpus (Brants & Franz, 2006), which contains over a trillion tokens culled from Google indices and thus allows extensive analysis of linguistic distributional patterns<sup>1</sup>. For each compound, we calculated the cumulative 5-gram frequency of occurrence between the modifier and head nouns (e.g., the summed count of *octopus ... apartment* with zero, one, two and three intervening words: for a similar approach, see Louwerse & Connell, 2011). Finally, frequencies were log-transformed as  $\ln(f + c)$ , where  $f$  is the raw frequency and  $c$  is a constant (minimum non-zero frequency) added to all values to enable log calculations of zero counts.

All novel compounds were potentially sensible because they had been successfully interpreted by a majority of participants in previous studies (Lynott & Connell, 2010b). Critical to our present purposes, data from an offline pretest (i.e., an open-response task under no time constraints:  $N = 20$ ) showed no reliable relationship between items' linguistic distributional frequency and success rate of interpretation,

<sup>1</sup>Note that a broader co-occurrence measure like LSA (Landauer & Dumais, 1997) is not the same as the 5-gram frequency count we use here. LSA measures co-occurrence over a broad paragraph-length window before reducing the total matrix to approximately 300 dimensions, so distance between words can be calculated as the cosine of the angle between two points in this high-d space. LSA scores between words therefore reflect a broad linguistic similarity, such that synonyms, which often occur in the same general contexts, should receive a high score. In contrast, n-gram frequencies measure co-occurrence within a narrow window of local context (i.e., with 0-3 intervening words for 5-grams). N-gram frequencies between words therefore reflect whether words are used in close proximity with one another. They do not reflect similarity of meaning because synonyms, which occur within 0-3 words of each other only rarely, should receive a low score.

$r(25) = .170$ , one-tailed  $p = .198$ .

**Participants** Twenty-four native speakers of English completed the experiment for a nominal sum. One participant was excluded for judging a majority of lexicalised filler items as nonsensical.

**Procedure** Participants were told that they would be presented with two-word phrases onscreen; some of these phrases would be familiar to them, while others would not. They were instructed to press the key labelled “Yes” to indicate that the phrase made sense or to press the key labelled “No” to indicate the phrase was nonsense. All responses were made with the participant’s dominant hand.

Each trial began with the word “Ready” appearing on the screen for 2000 ms, followed by the compound which remained onscreen until the participant made a decision. Response times were recorded in seconds from the onset of the compound until the participant’s keypress (“Yes” or “No” button). There was a blank screen interval of 1000 ms until the start of the next trial. Each participant saw all compounds presented in a different random order. The experiment took less than 10 minutes to complete.

**Design & Analysis** Response decision data (i.e., whether a compound was accepted or rejected) were analysed in a mixed-effects logistic regression model (logit link function) with crossed random factors of participants and items. The inclusion of items was empirically validated because it improved model fit over participants alone,  $\chi^2(1) = 38.31$ ,  $p < .0001$ . Linguistic frequency (i.e., log 5-gram frequency per compound) acted as a fixed predictors variable. Response time data were analysed in a mixed-effects linear regression model with participants as a random factor. Items were not included as a crossed random factor because it did not further improve model fit,  $\chi^2(1) = 2.55$ ,  $p = .111$  (Baayen, Davidson & Bates, 2008). Response decision (i.e., yes or no) and linguistic frequency (i.e., log 5-gram frequency per compound) acted as fixed interacting predictors variables. The primary advantages of mixed effects analysis as regards the present experiment is that it can determine the effect of item-level predictors while simultaneously taking participant variability into account, and that it offers greater power than analysing aggregated responses over participants or items (Baayen et al., 2008; Locker, Hoffman & Bovaird, 2007). Regression coefficients are reported as unstandardized  $\beta$  values. Effect size  $r$  for each predictor was calculated from  $t$  (Cohen, 1988).

## Results & Discussion

Data points more than 2.5 standard deviations from each participant’s mean time per response decision were removed as outliers: 1.6% for “yes” responses and 2.4% for “no”.

**Acceptance / Rejection Rates** Overall, 31.6% of novel compounds were judged as sensible and 68.4% as nonsense. As predicted, the likelihood of accepting a noun-noun



compound as sensible increased with linguistic distributional frequency,  $t(606) = 4.63$ ,  $p < .0001$ ,  $\beta = 0.251$ ,  $r = .185$ . Even though all the compounds were novel stimuli with no pre-specified definition, the fact that two nouns had been relatively frequently juxtaposed was enough to allow their combination to seem sensible.

**Acceptance / Rejection Times** Sensibility acceptance times ( $M = 2.625$ ,  $SE = 0.096$ ) were generally slower than rejection times ( $M = 2.364$ ,  $SE = 0.144$ ),  $t(557.1) = 3.03$ ,  $p = .003$ ,  $\beta = 1.151$ ,  $r = .127$ . Linguistic frequency had a marginally positive effect on overall response times,  $t(556.5) = 1.89$ ,  $p = .059$ ,  $\beta = 0.084$ ,  $r = .080$ ; but critically interacted with response decision to produce a negative effect on acceptance times,  $t(556.8) = -2.70$ ,  $p = .007$ ,  $\beta = -0.187$ ,  $r = .114$ . Separate analysis of “yes” and “no” responses showed the predicted inverse effects (see Figure 1). The time taken to accept a novel compound as sensible decreased with greater linguistic frequency,  $t(162.9) = -2.62$ ,  $p = .005$ ,  $\beta = -0.140$ ,  $r = .201$ , whereas the time to judge a compound as nonsense increased with linguistic frequency (i.e., low frequency compounds were rejected quickly, high frequency compounds were not),  $t(376.8) = 1.77$ ,  $p = .039$ ,  $\beta = 0.077$ ,  $r = .091$ .

In other words, the linguistic shortcut acts to facilitate shallow conceptual combination by providing an heuristic of sensibility. Higher linguistic distributional frequency facilitates acceptance of a novel stimulus: words that often share a local context are quickly and frequently judged to be a sensible phrase, which constitutes successful (albeit “quick and dirty”) processing of the combination. Lower

linguistic frequency, however, facilitates rejection: words that rarely share a context are quickly and frequently judged to be a nonsensical phrase, which may appear to constitute a failed conceptual combination process, but is perhaps better regarded as successful avoidance of a potentially costly but fruitless cognitive effort. Of course, participants do not have to rely solely on this linguistic shortcut just because it exists, and are free to base their sensibility judgements on the simulation system. Nevertheless, the results of this experiment demonstrate a statistical tendency to use linguistic distributional information as a sensibility heuristic, even when individual differences between participants and items are partialled out. We return to this issue in the general discussion.

## Experiment 2: Interpretation Generation

While the previous experiment examined a relatively shallow form of conceptual combination (i.e., judging whether a noun-noun compound made sense, but without having to specify why), this experiment focuses on a deeper form of processing by asking people to provide an actual interpretation for each compound. As before, we used a forced-choice task, where participants pressed “yes” if they could think of a meaning for the compound phrase (and then told us the meaning they had generated), and pressed “no” if they could not. Because the interpretation generation task invites deeper processing than sensibility judgement by asking people to think of a meaning, previous research has found it leads to more liberal use of “yes” decisions to novel compounds (Tagalakakis & Keane, 2006). We therefore expected a larger proportion of items to be accepted than in

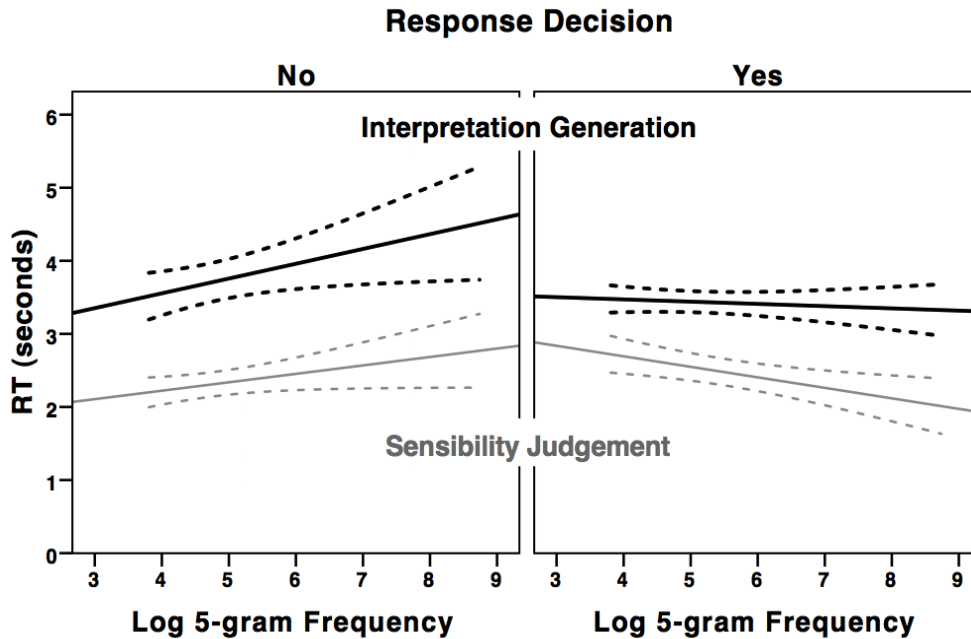


Figure 1: Regression plots of linguistic distributional frequency against model predicted response times for rejection (“no” decision) and acceptance (“yes” decisions) of novel noun-noun compounds in Experiment 1’s sensibility judgement and Experiment 2’s interpretation generation tasks. Dotted lines represent 95% confidence intervals around the mean. All fits except “yes” responses in interpretation generation are significant at  $p < .05$ .



Experiment 1, but, as for sensibility judgement, we expected this acceptance rate to be positively related to linguistic distributional frequency. The linguistic shortcut should quickly make high-frequency compounds appear interpretable, and – because most people can generate meanings for these items when they try – their subsequent combination in the simulation system is likely to succeed. Acceptance times thus reflect the latency of full conceptual combination, and as such should not be predicted by mere linguistic frequency. Rejection times, on the other hand, should show the same positive relationship with linguistic frequency that we saw for sensibility judgement: words that seldom appear in the same contexts will be quickly and frequently rejected as uninterpretable because the linguistic shortcut suggests their concepts may not combine.

## Method

**Materials** As per Experiment 1.

**Participants** Eighteen native speakers of English completed the experiment for a nominal sum.

**Procedure** Instructions were identical to Experiment 1 except that participants were asked to press the key labelled “Yes” to indicate that “Yes, I can think of a meaning” (whereupon a screen appeared for them to type in the interpretation just generated), or to press the key labelled “No” to indicate that “No, I cannot think of a meaning”. The experiment took approximately 20 minutes to complete and had a short, self-paced, break halfway through.

**Design & Analysis** Data were analysed with crossed random factors because model fit improved with the inclusion of items for both logistic regression of response decision data,  $\chi^2(1) = 69.95$ ,  $p < .0001$ , and linear regression of response time data,  $\chi^2(1) = 6.17$ ,  $p = .013$ . All other details were the same as Experiment 1.

## Results & Discussion

2.2% of “yes” responses to novel compounds resulted in blank or invalid interpretation (e.g., “a”, “I don’t know”) and were excluded from analysis as they did not represent successful combination. Data points more than 2.5 standard deviations from each participant’s mean per response decision were removed as outliers: 1.3% for “yes” responses and 2.8% for “no”.

**Acceptance / Rejection Rates** Overall, 68.5% of compounds were accepted and successfully interpreted and 31.5% were rejected as uninterpretable. Each compound had a variety of different, coherent interpretations, such as a *whale knife* as “A knife that has a picture of a whale on it” or “knife used by whalers”, or an *elephant complaint* as “a large complaint” or “a complaint about elephants in the area”. As predicted, the likelihood of successfully interpreting a noun-noun compound increased with linguistic distributional frequency,  $t(439) = 2.10$ ,  $p = .019$ .  $\beta$

$= 0.145$ ,  $r = .100$ .

**Acceptance / Rejection Times** Interpretation times ( $M = 3.348$ ,  $SE = 0.110$ ) were marginally faster than rejection times ( $M = 3.713$ ,  $SE = 0.194$ ),  $t(413.7) = 1.77$ ,  $p = .077$ ,  $\beta = 0.974$ ,  $r = .087$ . Linguistic frequency had an overall positive relationship with response times,  $t(117.0) = 2.14$ ,  $p = .034$ ,  $\beta = 0.209$ ,  $r = .194$ ; but, critically, it negatively interacted with response decision,  $t(413.8) = -2.44$ ,  $p = .015$ ,  $\beta = -0.259$ ,  $r = .119$ . Results for separate analysis of “yes” and “no” responses were as predicted (Figure 1). The time taken to accept and interpret a novel compound was unaffected by linguistic distributional frequency,  $t < 1$ . Like sensibility judgements, however, the time to reject a compound as uninterpretable increased with linguistic frequency,  $t(120.7) = 2.74$ ,  $p = .007$ ,  $\beta = -0.256$ ,  $r = .242$ .

In both shallow sensibility judgement and deep interpretation generation tasks, people use the linguistic shortcut as a “quick and dirty” guide to whether the concepts are likely to combine in a coherent situated simulation. Building a representation that is detailed enough to provide an interpretation is a function of deep conceptual processing in the simulation system, and took some 700 ms longer than accepting a compound as sensible. This extra depth of processing meant that successful interpretation times were no longer predicted by information from the linguistic system. Rejection times were also slower for interpretation generation than for sensibility judgement, and the 1300 ms difference suggests that at least some “no” responses resulted from tried-and-failed conceptual combination in the simulation system. However, the fact that rejection times were still strongly predicted by linguistic distributional frequency shows that the linguistic shortcut offered an important heuristic for avoiding this resource-wasting event.

## General Discussion

There are three novel findings in the present paper. First, we show that linguistic distributional frequency can predict not only the timecourse of successful conceptual processing (i.e., “yes” responses in sensibility judgement), but also the timecourse and likelihood of failure (i.e., “no” responses). Second, use of this linguistic shortcut extends beyond simple retrieval into the processing of novel stimuli in conceptual combination. The more often two words have appeared in close proximity to one another, the faster people are to accept the compound as sensible and the slower they are to reject it as uninterpretable nonsense. Third, we show that the influence of such linguistic shortcuts is not restricted to shallow conceptual tasks, but is also useful in deeper conceptual processing as a form of cognitive triage. The less often two words have appeared in close proximity, the faster people reject their compound as uninterpretable rather than risk costly failure in the simulation system. These findings support theories that argue for complementary roles of the linguistic and simulation systems in conceptual combination (Lynott & Connell,

2010a) and conceptual processing more generally (Barsalou et al., 2008; Louwerse & Jeuniaux, 2008).

But isn't all this just standard *word frequency effects*? In a word, no. We can't observe the above range of effects in conventional psycholinguistic tasks such as lexical decision or word naming. Firstly, responses in lexical decision and naming tasks are either correct or incorrect (e.g., correctly rejecting a non-word), whereas novel compounds do not necessarily have a "correct" interpretation. Rather, an individual's processing of a compound is either successful or unsuccessful, and even an "unsuccessful" outcome may represent the most efficient use of cognitive resources. Secondly, lexical decision and naming tasks rely solely on the recognition of known concepts, while conceptual combination tasks require the processing of new conceptual entities. Thus, the paradigm in this paper allows us to examine the conceptualisation of novel stimuli at two depths of processing, and demonstrate how the linguistic shortcut offers a useful heuristic in both shallow and deep tasks.

Of course, participants do not have to rely solely on a linguistic shortcut just because it exists. An individual may double-check apparently sensible or apparently uninterpretable compounds within the simulation system by actually attempting to combine the concepts. Indeed, it is possible that some particularly cautious individuals may even base every sensibility judgement on whether the concepts can combine into a coherent simulation. However, an easy shortcut is hard to refuse. Because the linguistic shortcut is faster and computationally cheaper than basing a judgment on the simulation system, and because on-the-fly conceptual processing does not have to be perfect (only "good enough": Ferreira, Bailey & Ferraro, 2002), participants can safely exploit it most of the time.

## Acknowledgements

This research was supported in part by the UK Economic and Social Research Council [grant RES-000-22-3248].

## References

- Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9, 321–324.
- Baayen, R. H., Davidson, D.J., & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.
- Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577–609.
- Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, & A. C. Graesser, A. (Eds.). *Symbols, embodiment, and meaning*. Oxford, UK: Oxford University Press.
- Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. A. Zwaan, *Grounding cognition: The role of perception and action in memory, language, and thinking* (pp. 129–163). Cambridge, UK: Cambridge University Press.
- The British National Corpus, Version 2 (BNC World) (2001). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available at <http://www.natcorp.ox.ac.uk>.
- Brants, T., & Franz, A. (2006). *Web 1T 5-gram Version 1*. Philadelphia: Linguistic Data Consortium.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.
- Connell, L., & Lynott, D. (2011). Modality switching costs emerge in concept creation as well as retrieval. *Cognitive Science*, 35, 763–778.
- Estes, Z. (2003). Attributive and relational processes in nominal combination. *Journal of Memory and Language*, 48, 304–319.
- Ferreira, F., Ferraro, V., & Bailey, K. G. D. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11, 11–15.
- Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23, 71–87.
- Glenberg, A. M. (1997). What is memory for? *Behavioral and Brain Sciences*, 20, 1–55.
- Kirkham, N.Z., Slemmer, J.A., & Johnson, S.P. (2002). Visual statistical learning in infancy: Evidence of a domain general learning mechanism. *Cognition*, 83, B35–B42.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis Theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211–240.
- Locker, L., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods*, 39, 723–730.
- Louwerse, M. M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, 35, 381–398.
- Louwerse, M. M., & Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. In M. de Vega, A. Glenberg, & A. C. Graesser (Eds.), *Symbols, embodiment, and meaning*. Oxford University Press.
- Lynott, D., & Connell, L. (2010a). Embodied conceptual combination. *Frontiers in Psychology*, 1(216), 1–14.
- Lynott, D., & Connell, L. (2010b). The effect of prosody on conceptual combination. *Cognitive Science*, 34, 1107–1123.
- Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2003). Verifying different-modality properties for concepts produces switching costs. *Psychological Science*, 14, 119–124.
- Tagalakakis, G., & Keane, M. T. (2006). Familiarity and relational preference in the understanding of noun-noun compounds. *Memory & Cognition*, 34, 1285–1297.

# Exploring Decision Rules and Sampling Dynamics in Recognition Memory

Gregory E. Cox (grcox@indiana.edu)

Richard M. Shiffrin (shiffrin@indiana.edu)

Department of Psychological and Brain Sciences, Indiana University  
1101 E. Tenth St., Bloomington, IN 47405 USA

## Abstract

Cox and Shiffrin (2012) introduced a dynamic model of recognition memory that was capable of making simultaneous predictions for accuracy and mean response time. The present paper extends that work by investigating the assumptions underlying the model's decision process, in particular those pertaining to the process by which features are sampled at test and the processes by which evidence for an "old"/"new" recognition decision accumulates. These assumptions are tested against empirically collected response time distributions. Evidence is found that sampling dynamics can change in response to instructions, and that competition between accumulators for "old" and "new" evidence is not necessary to capture at least the recognition data presented here.

**Keywords:** Recognition memory; response time; memory models.

## Introduction

Cognitive scientists have long understood the utility of timing information for making inferences about the cognitive processes underlying behavior. Yet, the field of recognition memory has only made limited use of this rich source of data. This is partially because of a lack of available data to constrain dynamic models of recognition, and partially because most theories of recognition are themselves static: Recognition decisions are presumed to operate on a fixed value of "familiarity", which is a signal indicating the presence or absence of an item in memory. If the familiarity is above a particular criterion, a participant declares the item "old", otherwise it is considered "new". Most extant models of memory are concerned purely with how this familiarity value is calculated for different items. Some of these models assume that an item is compared to many individual memory traces (e.g., Hintzman, 1988; Shiffrin & Steyvers, 1997; McClelland & Chappell, 1998), to an aggregate of several memory traces (Murdock, 1982), or to the present context (e.g., Dennis & Humphreys, 2001; Howard & Kahana, 2002).

We have recently proposed (Cox & Shiffrin, 2012) that taking a dynamic, as opposed to static, approach to recognition will shed light on many previously vexing issues. Our original work was concerned with the problem of criterion setting in the case of widely-varying stimulus materials, where the absolute value of familiarity may fluctuate wildly between different test items. Although, in principle, criteria might be learned over time (c.f., Turner, Van Zandt, & Brown, 2011), we showed that the problem of learning stimulus-specific criteria is bypassed entirely if one treats recognition decisions as inherently dynamic in nature. In this case, rather than basing a decision on a fixed asymptotic level of familiarity, recognition can make use of the *changes in* familiarity over time as information is gathered from the test stimulus. The shape of

the "familiarity profile" generated by this gradual accrual of information is invariant to many of the factors that would result in different levels of asymptotic familiarity, e.g., amount of prior exposure to the stimulus and degree of encoding.

Although the model proposed by Cox and Shiffrin (2012) was originally motivated to address the criterion setting problem, it has potential application as a general model of recognition. We showed that the predictions of the model for accuracy and mean response times were in accord with published data on recognition memory, but left many questions unanswered. For example, we assumed that the decision took the form of a race between two parallel accumulators, one leading to an "old" response and the other to a "new" response. This decision mechanism was not explicitly compared to other possible mechanisms, e.g., those with non-independent accumulators like a random walk or diffusion process (e.g., Ratcliff, 1978). Further, we assumed that the times between sampling events—i.e., sampling a feature from the test stimulus—were drawn independently and identically from an exponential distribution, but made no attempt to investigate whether this sampling process might differ between study and test and between conditions.

At the level of mean RT, these sets of predictions would be difficult to disentangle from those arising from other assumptions regarding correlated accumulators or different sample time distributions. Thanks to the generous contributions of Chris Donkin and Andrew Heathcote, we are now able to test the assumptions of our original dynamic recognition model regarding the decision mechanism and sampling dynamics. In so doing, we have taken yet more steps toward developing a general theory of recognition that is able to account for both accuracy and response time.

## The Model

The modeling framework used in this paper is a direct outgrowth of the dynamic recognition model proposed by Cox and Shiffrin (2012), which is itself based on the REM model (Shiffrin & Steyvers, 1997). The central tenet of the dynamic recognition model is that information about a test item is sampled over time and added to a memory probe. When the probe is compared to the contents of memory at time  $t$ , it produces a particular familiarity value,  $\phi(t)$ . How this value changes over time constitutes evidence for an old/new recognition decision: decrements in  $\phi(t)$  (relative to  $\phi(t-1)$ ) are evidence that the item is new, because additional item information *reduces* its familiarity. Contrariwise, increases in  $\phi(t)$  are evidence that the item is old, because new item information *increases* its familiarity. We now give a more detailed exposi-

tion of the model, although the reader is directed to Cox and Shiffrin (2012) for additional information.

## Representations and Encoding

An episodic memory trace for an item is represented as a vector of features, each of which is binary with an equal *a priori* probability of being “0” or “1”. These features are divided into two kinds: *content* features which arise from the item itself, which may include information about its sensory characteristics or semantics, and *context* features which arise from the situation in which the item was encountered. Episodic memory is presumed to consist of a (very large) set of such traces from across the life span, although for practical purposes, we restrict ourselves to modeling just a subset of these traces (see below). In addition, we assume that two traces of the same item need not be encoded using the same features. For example, a particular apple might be encoded primarily with features pertaining to its physical appearance on one occasion, but on another the apple might be encoded primarily with features pertaining to its taste. This encoding variability is captured in the model by a parameter  $\alpha$ , the probability that any one feature is shared between two traces of the same item (independent of whether a *value* is stored for that feature). Likewise, the parameter  $\gamma$  denotes the probability that a feature is shared between two traces of *different* items. The exact values of these parameters will depend on the kinds and varieties of items presented during study and test and are at least partially under the decision maker’s control.

A memory trace is formed whenever an item is encountered, e.g., on a study list<sup>1</sup>. Memory traces are formed by copying into long-term memory a short-term memory representation of the item. The short-term memory representation is built up over time. Context features are presumed to be already present in the representation, since they are persistent in the environment. Content features are gradually sampled from the test item and added to the short-term memory representation. Such features are sampled at random, with replacement, and are copied into the STM representation correctly with probability  $c$ , otherwise a random value (either “0” or “1”) is copied. Note that, because the sampling occurs with replacement, a new sampled value can replace one that was previously present in the STM representation, even if the new value was copied incorrectly. Finally, due to capacity limitations, we assume that at most  $N_c$  content features and  $N_x$  context features are available to enter into STM.

At study, the sampling process is terminated at stimulus offset after some specified time,  $T_s$ . We assume that the dynamics of the sampling process are governed by a homogeneous Poisson process, that is, the time between feature samples is exponentially distributed with rate  $\rho_s$ . Thus, the number of samples obtained at study (which may involve “overwriting” a previously sampled feature value) after  $T_s$  time

units is Poisson distributed:

$$\Pr(k \text{ samples}; \rho_s, T_s) = \frac{(\rho_s T_s)^k}{k!} \exp(-\rho_s T_s).$$

At that point, the STM representation of the study item, comprised of the current context features plus whatever content features were sampled from the item in the time available, is copied into a long-term memory trace. Again, any one feature may be incorrectly copied from the STM representation into LTM with probability  $c$ , otherwise a random value is stored. Thus, a memory trace consists of  $N_x$  context features, some of which may have been copied incorrectly, and  $N_c$  content features, some of which have (correctly or incorrectly) sampled values and others of which are “blank”, indicating that no value was sampled for that feature.

At test, the STM representation of the test item serves as a probe that is compared to the contents of memory. This comparison is made after each sampling event  $t$ , resulting in a familiarity value,  $\phi(t)$ . The change in  $\phi(t)$  from one sampling event to the next constitutes the evidence for making a recognition decision. We now turn to the details of how the comparison between the probe and memory is made and how  $\phi(t)$  is calculated.

## Computing Familiarity

To compute a familiarity value, after each sampling event, the STM representation of the test item is used as a probe and is compared, in parallel, to all traces in episodic memory. Each comparison results in a measure of the similarity between the probe and a given trace, denoted  $\lambda_i(t)$  (where  $i$  indexes the trace in memory). The similarity measure takes the form of a likelihood ratio: the likelihood that the probe and the trace encode the *same* item versus the likelihood that they encode *different* items.

**Likelihood Calculation** The evidence that goes into computing these likelihoods comes from the features that have been sampled thus far and added to the probe (by time  $t$ ) and from the features that had been stored in the trace at study. For any one feature, there are five possible outcomes of the comparison:

1. The probe and trace both have a value stored, and the value matches ( $M$ ).
2. The probe and trace both have a value stored, and the values do not match ( $N$ ).
3. The probe has a value stored, but the trace does not ( $P$ ).
4. The trace has a value stored, but the probe does not ( $T$ ).
5. Neither the trace nor the probe have a value stored.

Although the full model described in Cox and Shiffrin (2012) makes use of all these possible outcomes, for present purposes we are only concerned with the first two. This is because outcomes  $P$  and  $T$  are only indicative of a non-match if traces of the same item are more likely to share features than are traces of different items (i.e., if  $\alpha > \gamma$ ). In the simulations reported here, we deal only with items that are of the

<sup>1</sup>A memory trace would also be formed after each test trial. We do not attempt to model this here, but the formation of memory traces at test may be necessary to explain several phenomena in recognition memory (Criss, Malmberg, & Shiffrin, 2011).

same type (namely, nouns) and so we assume that the degree of encoding variability is uniform, i.e.,  $\alpha = \gamma$ .

Although the reader is referred to Cox and Shiffrin (2012) for the details of these derivations, the following are the conditional probabilities needed to compute a match:

$$\begin{aligned}\Pr(M|\text{Same}) &= c + (1 - c)\frac{1}{2} & \Pr(M|\text{Different}) &= \frac{1}{2} \\ \Pr(N|\text{Same}) &= (1 - c)\frac{1}{2} & \Pr(N|\text{Different}) &= \frac{1}{2}.\end{aligned}$$

Because features are sampled independently, these probabilities are multiplied together for as many feature comparisons result in a  $M$  or  $N$  outcome, as appropriate to obtain the likelihood under each hypothesis, *same* or *different*. Letting  $N_M$  and  $N_N$  denote the number of feature-value matches and mismatches, respectively, the likelihood ratio for a trace  $i$  can thus be written:

$$\begin{aligned}\lambda_i(t) &= \left[ \frac{\Pr(M|\text{Same})}{\Pr(M|\text{Different})} \right]^{N_M} \left[ \frac{\Pr(N|\text{Same})}{\Pr(N|\text{Different})} \right]^{N_N} \\ &= (1 + c)^{N_M} (1 - c)^{N_N}.\end{aligned}\quad (1)$$

**Familiarity Calculation** A likelihood ratio  $\lambda_i(t)$  is produced, in parallel, for all traces in memory. However, the vast majority of traces in memory will be an extremely poor match to the probe and produce very low likelihood ratios, either because these traces encode different items (differ in content features) or because they were encoded in vastly different situations (differ in context features). Thus, we assume that only some traces in memory are “active” at any one time. For a trace to be activated in response to a probe, it must produce a likelihood ratio greater than a threshold  $\theta$ . For simplicity, we fix  $\theta = 1^2$ . We denote the set of active traces, i.e., those for which  $\lambda_i(t) > \theta$  by  $A(t)$  (of size  $|A(t)|$ ). Only those traces in  $A(t)$  contribute to the familiarity value  $\phi(t)$ . Because traces that fail to match on either content (i.e., they encode the same item) or context (i.e., they were encoded under similar circumstances, e.g., in a memory study list) are extremely unlikely to pass threshold and entire  $A(t)$ , we only model storage of the  $N$  items from the study list plus  $K$  traces of the test item from different contexts.

Having computed a match value  $\lambda_i(t)$  for each trace in memory and selected the set of activated traces  $A(t)$ , the familiarity value,  $\phi(t)$ , is simply the average likelihood over all active traces, i.e.,

$$\phi(t) = \frac{1}{|A(t)|} \sum_{i \in A(t)} \lambda_i(t). \quad (2)$$

## Decision Mechanism

Our model assumes that recognition decisions are based not on the absolute value of familiarity, but rather on how this value changes over time. Because the likelihood ratio scale is

<sup>2</sup>The full model assumes that  $\theta$  depends on the ratio  $\frac{\alpha}{\gamma}$ , but these are assumed equal in the subsequent simulations and so play no role in the setting of the activation threshold.

highly skewed, we first take the logarithm  $\log \phi(t)$  (this transformation does not alter the underlying logic of the model); the evidence for recognition decisions is then  $\nabla \log \phi(t) = \log \phi(t) - \log \phi(t - 1)$ . Positive values of  $\nabla \log \phi(t)$  are considered evidence that the test item is “old” while negative values are treated as evidence that the item is “new” (i.e., has not been encountered in the current context).

In our original model, positive and negative changes in familiarity feed into two parallel, non-interacting accumulators,  $B^+(t)$  and  $B^-(t)$ . That is,

$$\begin{aligned}B^+(t) &= \sum_{\tau=0}^t \begin{cases} \nabla \log \phi(\tau) & \text{if } \nabla \log \phi(\tau) > 0 \\ 0 & \text{if } \nabla \log \phi(\tau) \leq 0 \end{cases} \\ B^-(t) &= \sum_{\tau=0}^t \begin{cases} \nabla \log \phi(\tau) & \text{if } \nabla \log \phi(\tau) < 0 \\ 0 & \text{if } \nabla \log \phi(\tau) \geq 0 \end{cases}.\end{aligned}$$

The final recognition decision is, then, a race between these two accumulators: the predicted response is given by whichever accumulator reaches its threshold first ( $\beta^+$  or  $\beta^-$  for  $B^+(t)$  and  $B^-(t)$ , respectively).

The predicted response time is related to the number of sampling events  $t$  needed for the first of the accumulators to reach its threshold. In our previous work, we assumed, as is common in many models that posit sequential independent feature samples from a test stimulus (e.g., Brockdorff & Lamberts, 2000; Mewhort & Johns, 2005), that the times between sampling events were exponentially distributed with a fixed, uniform rate  $\rho_t$  (which may be different than the sampling rate at study,  $\rho_s$ ). This has the consequence that the observed response time, given that  $t$  samples were needed for one of the accumulators to reach threshold, is drawn from a Gamma distribution (the convolution of  $t$  independent and identically distributed exponential distributions).

## Simulations

Although our previous work on a dynamic model for recognition memory included predictions for mean response times, a much stronger test of the model is to compare its predictions to empirically collected *distributions* of response times. Unfortunately, empirical RT distributions in recognition memory are still somewhat rare (despite the fact that one of the leading models of response time, the diffusion model of Ratcliff, 1978, was originally developed to account for RT distributions in recognition memory!). Thus, we once again express our gratitude to Chris Donkin and Andrew Heathcote for providing us with empirical RT distributions against which to compare the predictions of our model. This comparison affords special insight into two features of our model that had to be simply assumed in our earlier work: first is the distribution of times between samples at test. Second is the assumption of independence between the two accumulators.

In obtaining a correspondence between the model and data, we fixed most of the parameters involved at the values given in Table 1. Our primary goal in these simulations was not

Table 1: Summary of the fixed parameters of the model, along with their values used in the present simulations.

Parameter	Value	Description
$N_c$	30	Maximum number of content features.
$N_x$	30	Number of context features.
$c$	0.85	Probability of correct feature copy.
$K$	200	Number of history traces of a test item available to be activated.
$\theta$	1	Minimum likelihood needed to enter the set of active traces.
$\rho_s$	60	Feature sampling rate at study.

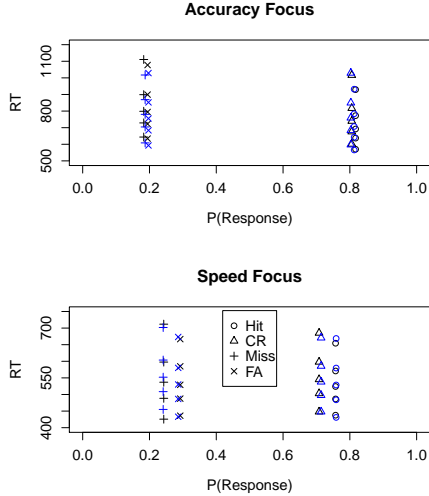


Figure 1: Observed group performance and RT quantiles (10%, 30%, 50%, 70%, 90%) are shown in black, with model predictions using the independent accumulator mechanism in blue.

to obtain the best quantitative fit possible, which could potentially require all model parameters to be freely varied. Rather, we wished to capture the qualitative features of the data whilst varying only a small number of parameters, thereby allowing a more direct interpretation of the model’s predictions.

## Sampling Dynamics

Heathcote and Donkin (2012) recently collected both accuracy and response times in a recognition memory paradigm. Participants studied lists of 25 words each. At the conclusion of each study list, participants were tested in a standard old/new recognition paradigm. The test lists were unbiased (i.e., composed of an equal number of studied and unstudied words). On half of the test lists, participants were instructed to try to be as accurate as possible without taking too long to make their response (“accuracy focus”), while on the other half, participants were instructed to be as fast as possible without sacrificing accuracy (“speed focus”). The resulting group mean accuracy and RT quantiles (along with model fits) are plotted in Figure 1.

To replicate this procedure in our model, we assumed that 25 items were studied for  $T_s = 1$  time unit, fixed the other model parameters at the values given in Table 1, and simu-

Table 2: Best-fitting parameter values for the independent accumulator mechanism, given the Heathcote and Donkin (2012) recognition data.

Parameter	Condition	Value
$\beta^+$	Accuracy	13
	Speed	6
$\beta^-$	Accuracy	-17
	Speed	-9.4
$\rho_t$	Accuracy	158 samples per second
	Speed	105 samples per second
$T_n$	Accuracy	297 ms
	Speed	266 ms

lated 1000 study/test lists. To fit the model to the observed RT distribution data, we wished to optimize four parameters, which we allowed to vary between the speed and accuracy focus conditions: The first two of these are the thresholds for the old and new accumulators,  $\beta^+$  and  $\beta^-$ , respectively. We assume that participants can adjust their decision criteria in response to instructions, with lower thresholds leading to faster but potentially more error-prone responses. Two additional variables were allowed to vary between condition: the sampling rate at test,  $\rho_t$ , and the amount of “non-decision time” per trial,  $T_n$ . Non-decision time is introduced to account for any processes that are not being explicitly modeled, e.g., the time required to initially attend to the test stimulus and the time required to actually execute the motor response. Just as we assume that participants are able to adjust their decision bounds, we assume that instructions can induce participants to devote more resources toward particular components of the recognition process. For example, speed instructions may lead to faster execution of the motor response—leading to reduced non-decision time—but the added time pressure may result in less efficient extraction of information from the test stimulus—and therefore a lower average sampling rate.

Indeed, this is the pattern found in the fitted parameters, shown in Table 2 (which achieved an adjusted  $R^2 = .98$  between predicted and observed RT quantiles). The thresholds are farther apart in the accuracy condition than in the speed condition, as one would expect. In addition, non-decision time is slightly lower in the speed condition, which may be attributed to a slight decrease in the time needed to execute the response resulting from practice in the blocks of speed trials. Most interesting, however, is that although response thresholds are lower—and responses correspondingly faster—in the speed focus condition, the sampling rate is estimated to be substantially lower in the speed condition<sup>3</sup>. In other words, although participants appear willing to make responses on the basis of less evidence when encouraged to produce fast responses, participants appear to be collect this evidence less efficiently. One possible explanation for this is, to paraphrase Starns, Ratcliff, and McKoon (2012), that the additional metacognitive demands in the speed condition (e.g., the need to monitor response time and avoid pure guessing)

<sup>3</sup>In addition, both estimated sampling rates at study are faster than the sampling rate assumed at study, which was fixed at 60 samples per second.

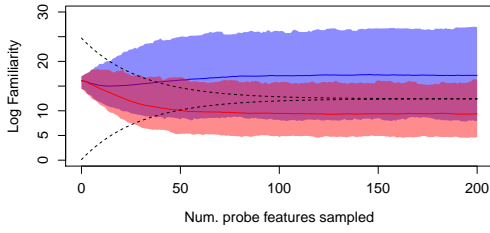


Figure 2: Collapsing thresholds (dashed lines) relative to the mean familiarity value for targets (blue line) and foils (red line). Light regions around the mean familiarity show the area between the 2.5% and 97.5% quantiles of the familiarity distribution at each time.

divert resources from processing the test stimulus itself.

### Correlated Accumulators

Another feature of our model that was previously untested is the assumption of independence between the positive and negative accumulators. This independence places our model in a different class than random walk or diffusion models (e.g., Ratcliff, 1978), which assume that evidence in favor of one option (e.g., “old”) is equivalent evidence *against* the other option (e.g., “new”). A random walk model may thus be seen as a model with two perfectly anti-correlated accumulators.

The primary reason for choosing an independent accumulator structure in our original work was the nature of the evidence on which we presume recognition decisions are based. In particular, because there is a maximum number of features that may be sampled, there is also a maximum (and minimum) value of familiarity that could result from a test item. Consider the case where, by a lucky happenstance, all the features sampled and added to the probe at time  $t$  perfectly match a single trace stored in memory. Further assume that this single trace is the only activated trace. Then, from equations 1 and 2,  $\phi(t) = (1 + c)^{N_c + N_x}$ . Because  $N_c$  and  $N_x$  are fixed, this is the maximum possible familiarity value. Clearly, any subsequent feature samples can only lead to a zero or negative change. Thus, if the threshold on the positive counter has not yet been reached, and the positive and negative counters are anti-correlated, the “old” threshold will never be attained because no further positive changes are possible.

Thus, if the accumulators are to be perfectly anticorrelated, as in a random walk, their thresholds cannot be fixed across time. This is not merely a feature of this particular model, but any model that places a limit on the amount of evidence that may be accrued over time. Thus, we introduce a rule by which the accumulator thresholds may collapse over time (for examples of collapsing thresholds in other domains, see Balakrishnan & Macdonald, 2011; Frazier & Yu, 2008). This rule is but one of many possible rules, but is based on the principle that the thresholds should be reduced in proportion to the amount of information that remains to be

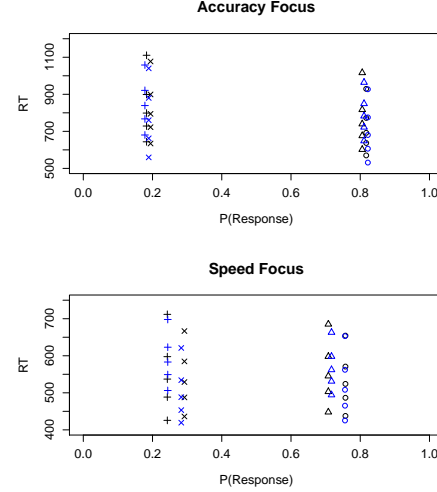


Figure 3: Observed group performance and RT quantiles (10%, 30%, 50%, 70%, 90%) are shown in black, with model predictions—using the correlated accumulators with collapsing thresholds—in blue.

sampled from the stimulus. On any feature sampling event, each of the  $N_c$  content features has an equal chance of being sampled. Thus, the probability that any one feature has *not* been sampled after  $t$  sampling events is  $\left(1 - \frac{1}{N_c}\right)^t$ . Imagine that the decision maker has a number of features  $N_{max}$  that they consider sufficient for making a recognition decision (Diller, Nobel, & Shiffrin, 2001). Then, the probability that  $N_{max}$  features have been sampled after  $t$  sampling events is  $\sigma(t) = \left[1 - \left(1 - \frac{1}{N_c}\right)^t\right]^{N_{max}}$ . Thus, given initial thresholds  $\beta^+(0)$  and  $\beta^-(0)$ , the thresholds collapse toward one another in a symmetric fashion:

$$\beta^+(t) = \beta^+(0) - \sigma(t) \left( \frac{\beta^+(0) - \beta^-(0)}{2} \right) \quad (3)$$

$$\beta^-(t) = \beta^-(0) + \sigma(t) \left( \frac{\beta^+(0) - \beta^-(0)}{2} \right). \quad (4)$$

Different choices of  $N_{max}$  can lead to very different threshold behavior, however both our own investigations (not reported here for space) and the empirical work of Balakrishnan and Macdonald (2011) suggest that thresholds begin rather far apart and thereafter converge relatively quickly. Such behavior is consistent with low values of  $N_{max}$ ; in the absence of explicit evidence otherwise, we choose here  $N_{max} = 1$ . The resulting behavior of the thresholds relative to the mean familiarity value for targets and foils is shown in Figure 2.

Using this new decision rule, we again fit the model to Heathcote’s data, the only difference being that now the *initial* threshold value, rather than a constant threshold, for each accumulator was varied. The resulting best-fitting parameter values are shown in Table 3, with model fits in Figure 3 (adjusted  $R^2 = 0.89$  between predicted and observed RT quantiles). It is apparent that this new mechanism, more akin to



Table 3: Best-fitting parameter values for the correlated accumulator mechanism.

Parameter	Condition	Value
$\beta^+(0)$	Accuracy	8.6
	Speed	3.2
$\beta^-(0)$	Accuracy	-16
	Speed	-9
$\rho_t$	Accuracy	99 samples per second
	Speed	138 samples per second
$T_n$	Accuracy	357 ms
	Speed	367 ms

that of a random walk model, while capable of fitting accuracy just as well as the independent accumulator model does not fit RT quantile data as well, at least when assuming that the time between samples in exponentially distributed. As with the independent accumulator model, initial thresholds are have lower absolute value under speed focus relative to accuracy focus. However, in this case, non-decision time is estimated to be roughly equal between the two conditions, with increased sampling rate in the speed focus condition, as might be expected if speed instructions encouraged greater attention to the stimulus. Although the degree of fit is poorer when using correlated accumulators, the fit is certainly not too bad, not enough to rule out this mechanism as a plausible one for recognition decisions.

## Discussion

Our original work on a dynamic model for recognition memory (Cox & Shiffrin, 2012) represents one of the few attempts to link a full-fledged model of memory (in this case, an extension of the REM model; Shiffrin & Steyvers, 1997) to a decision mechanism capable of predicting both accuracy and response time. By comparing the model's predictions against entire RT distributions, we have been able to show that while the original assumptions of the model are viable, there are other possible routes to explore. These include the effect of task instruction (speed vs. accuracy focus) on sampling dynamics at test as well as a correlated counter mechanism, although a more thorough investigation of these mechanisms and the meaning of their parameters is in order.

Our model is admittedly complex, however, and incorporates many sources of variability. A correlated accumulator mechanism could produce superior predictions given different parameters for the underlying memory process. Indeed, our ongoing work in fitting our own data and data from (Starns et al., 2012) strongly suggests correlation between accumulators, even if the current data of Heathcote and Donkin (2012) do not require them. Just as task demands may influence sampling dynamics, they may also lead to different ways of balancing evidence in favor of "old" and "new" responses. Further, the dynamics of the sampling process may themselves be non-stationary, with the sampling rate changing over time (Hockley & Murdock, 1987) or different features being detected at different rates (Brockdorff & Lamberts, 2000). Finally, although the current paper is primarily

exploratory, we expect that additional models of the recognition process will be developed. As part of this venture, more attention must be paid to the flexibility of such models, and we believe that exercises such as those in this paper can help illuminate the space of possible mechanisms available to modelers of recognition memory. The complexity of the models presented here is balanced by the range of data they may be expected to explain, and the present work represents just one of many forays that will be necessary to develop a general theory of recognition memory.

## References

- Balakrishnan, J. D., & Macdonald, J. A. (2011). Performance measures for dynamic signal detection. *Journal of Mathematical Psychology*, 55, 290–301.
- Brockdorff, N., & Lamberts, K. (2000). A feature-sampling account of the time course of old-new recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(1), 77–102.
- Cox, G. E., & Shiffrin, R. M. (2012). Criterion setting and the dynamics of recognition memory. *Topics in Cognitive Science*, 4(1), 135–150.
- Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, 64, 316–326.
- Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, 108(2), 452–478.
- Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC-REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(2), 414–435.
- Frazier, P. I., & Yu, A. J. (2008). Sequential hypothesis testing under stochastic deadlines. In *Advances in neural information processing systems* (Vol. 20, pp. 465–472). Cambridge, MA: MIT Press.
- Heathcote, A., & Donkin, C. (2012). *Recognition data*. Unpublished.
- Hintzman, D. L. (1988). Judgements of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, 95(4), 528–551.
- Hockley, W. E., & Murdock, B. B. (1987). A decision model for accuracy and response latency in recognition memory. *Psychological Review*, 94(3), 341–358.
- Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46, 269–299.
- McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, 105(4), 724–760.
- Mewhort, D. J. K., & Johns, E. E. (2005). Sharpening the echo: An iterative resonance model for short-term recognition memory. *Memory*, 13(3/4), 300–307.
- Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89(3), 609–626.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2), 59–108.
- Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4(2), 145–166.
- Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology*, 2012(1), 1–34.
- Turner, B. M., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological Review*, 118(4), 583–613.



# Reverse appraisal: The importance of appraisals for the effect of emotion displays on people's decision making in a social dilemma

**Celso M. de Melo** (demelo@ict.usc.edu)

**Jonathan Gratch** (gratch@ict.usc.edu)

Institute for Creative Technologies, University of Southern California,  
12015 Waterfront Drive, Building #4 Playa Vista, CA 90094-2536, USA

**Peter Carnevale** (peter.carnevale@marshall.usc.edu)

USC Marshall School of Business, Los Angeles, CA 90089-0808, USA

**Stephen Read** (read@rcf.usc.edu)

USC Department of Psychology, Los Angeles, CA 90089-1061, USA

## Abstract

Two studies are presented that explore the interpersonal effect of emotion displays in decision making in a social dilemma. Experiment 1 ( $N=405$ ) showed that facial displays of emotion (joy, sadness, anger and guilt) had an effect on perception of how the person was appraising the social dilemma outcomes (perception of appraisals) and on perception of how likely the person was to cooperate in the future (perception of cooperation). Experiment 1 also showed that perception of appraisals (partially and, in some cases, fully) mediated the effect of emotion displays on perception of cooperation. Experiment 2 ( $N=202$ ) showed that manipulating perception of appraisals, by expressing them textually, produced an effect on perception of cooperation thus, providing evidence for a causal model where emotion displays cause perception of appraisals which, in turn, cause perception of cooperation. In line with Hareli and Hess' (2010) findings and a social-functions view of emotion, we advance the reverse appraisal proposal that argues people can infer, from emotion displays, how others are appraising a situation which, in turn, support inferences that are relevant for decision making. We discuss implications of these results and proposal to decision and emotion theory.

**Keywords:** Emotion Displays, Decision Making, Social Dilemma, Appraisal Theories, Reverse Appraisal

## Introduction

Recent decades have seen growing interest on the interpersonal effect of emotion in decision making (e.g., Van Kleef, De Dreu, & Manstead, 2010). Complementing research on the impact of emotion in one's own decision making (for a recent review see Blanchette & Richards, 2010), this research explores how one's emotion displays impact another's decision making and emphasizes that emotional expressions are not simple manifestations of internal experience; rather, expressions are other-directed and communicate intentions, desired courses of actions, expectations and behaviors (Frijda & Mesquita, 1994; Keltner & Kring, 1998). The expression of emotion has also been argued to play a significant role in emergence of cooperation in social dilemmas (Boone & Buck, 2003; Frank, 1988). Social dilemmas, such as the prisoner's dilemma, are situations where people must choose between pursuing their own self-interest and collect a short-term reward or trust another person to reach mutual cooperation

and maximize joint long-term reward (Kollock, 1998). Empirical evidence confirms that facial displays of emotion can impact cooperation (e.g., de Melo, Carnevale, & Gratch, 2012; Schug, Matsumoto, Horita, Yamagishi, & Bonnet, 2010). However, the mechanism by which such social effects of emotion are achieved is less well understood.

In this paper we address this issue and look at appraisal theories of emotion to understand what is the information people retrieve from emotion displays and how is that accomplished. In appraisal theories, emotion displays arise from cognitive appraisal of events with respect to an agent's goals, desires and beliefs (e.g., is this event congruent with my goals? Who is responsible for this event?). According to the pattern of appraisals that occurs, different emotions are experienced and displayed. Since displays reflect the agent's intentions through the appraisal process, it is also plausible to ask whether people can infer from emotion displays the agent's goals by reversing the appraisal mechanism. We refer to this proposal as *reverse appraisal*. The intuition is that if appraisal, abstractly, is a function that maps from <event, mental state> to emotion, reverse appraisal is a function that maps from <event, emotion> to mental state. Empirical evidence is still scarce but in a recent study Hareli and Hess (2010) showed that people could, from expressed emotion, make inferences about the character of the person displaying emotion. For instance, a person who reacted with anger to blame was perceived as being more aggressive, self-confident but also as less warm and gentle than a person who reacted with sadness. Moreover, the results showed that such inferences were mediated by appraisal variables. In our case, reverse appraisal predicts that people infer, from emotion displays, how the counterpart is appraising the social dilemma outcomes; then, from these *perceptions of appraisal*, people infer how likely the counterpart is to cooperate in the future, which we refer to as *perceptions of cooperativeness*. This causal model is shown in Figure 1. The goal of the paper is to establish this causal model.

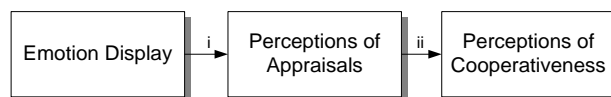


Figure 1: Causal model for the effect of emotion displays in cooperation in a social dilemma.

In a previous study (de Melo, Carnevale, & Gratch, 2011) we began gathering evidence for this causal model. Participants played the iterated prisoner’s dilemma with two computer players, or *agents*, that, even though following the same strategy to choose their actions, showed emotion displays that reflected different social value orientations (e.g., cooperative or competitive). Computer agents that show emotion have been argued in the past to be a useful research tool for basic human-human interaction research (Blascovich et al., 2001). In our case, following the intuition from appraisal theories that what matters is not the emotion display in itself but the appraisals that elicited it, the agents differed only in the context in which joy was expressed. For instance, a cooperative agent showed a smile in mutual cooperation, whereas a competitive agent showed a smile when the agent exploited the participant. As predicted by appraisal theories, the results showed that people interpreted the same smile differently and cooperated more with the cooperative than the competitive agent. In the present paper we go further and study whether the information people retrieve from emotion displays pertains to how the counterpart is appraising the interaction and, if such perceptions of appraisals lead to inferences about the counterpart’s likelihood of cooperation. To accomplish this, in a first experiment we asked participants to imagine playing the prisoner’s dilemma with different agents; participants were always told the same outcome occurred but were shown videos of different emotional reactions from the counterpart and were, then, queried about how they thought the counterpart was appraising the situation and how likely it was to cooperate in the future. We hypothesized that: following previous findings, emotion displays would impact perceptions of the counterpart’s cooperativeness (H1); and, according to reverse appraisal, emotion displays would impact perceptions of the counterpart’s appraisals (H2). A statistical analysis of mediation (Preacher & Hayes, 2008) was also conducted to test whether appraisal variables (conduciveness to goals or blameworthiness) mediated the effect of emotion displays on people’s perception of the counterpart’s cooperativeness. The hypothesis was that appraisal variables would mediate, at least partially, the interpersonal effect of emotion (H3). To further test the mediating role of appraisals, a second experiment explicitly manipulated appraisals and measured the effect on people’s perception of how cooperative the counterpart was. The manipulation consisted of having the agents, instead of showing facial displays of emotion, express how they were appraising the outcome through text (e.g., “I really don’t like this outcome and I blame you for it”). The hypothesis was that, in line with reverse appraisal, expression of appraisals would lead to effects on perception of the counterpart’s cooperativeness that were consistent with findings in the previous experiment (H4). This experiment, thus, establishes the remaining link (ii, Figure 1) in the proposed causal model. Therefore, Experiments 1 and 2, together, provide experimental evidence for the proposed model (Spencer, Zanna, & Fong, 2005).

## Experiment 1

### Method

**Scenarios.** Participants imagined playing the iterated prisoner’s dilemma with agents that displayed emotion. Each scenario pertained to the first round (of a 5-round game) and corresponded to a particular outcome of the game. Participants were then shown a video of the agent reacting emotionally to the outcome. Following the approach by Kiesler, Waters and Sproull (1996), and similarly to our previous study (de Melo et al., 2011) the game was recast as an investment game.

**Design.** The experiment followed a mixed design with two factors: Outcome (between-participants) with 4 levels (one for each possible outcome of the game); and, Emotion (repeated-measures) with 5 levels (Neutral vs. Joy vs. Anger vs. Sadness vs. Guilt). Building on experience from previous studies, we only explored 4 emotions and did not consider, for the time being, a full factorial design but, rather, only pairings of outcome and emotion that produced effects in those studies, as shown in Table 1. Considering only this subset of the possible pairings had, at least, two advantages: (1) each participant experienced at most 3 pairings (as opposed to 5 if all were considered), which constrained total participation time and, thus, reduced fatigue and boredom effects; (2) pairings that did not have a clear intuitive interpretation (e.g., displaying sadness or anger in mutual cooperation) were excluded from analysis.

Table 1: Emotions explored in Experiment 1.

		<i>Agent</i>	
		Cooperation	Defection
<i>Participant</i>	Cooperation	Neutral, Joy	Neutral, Joy, Guilt
	Defection	Neutral, Anger, Sadness	Neutral, Joy, Anger

**Emotion displays.** In this experiment, participants watched videos of agents expressing facial emotion displays. Three agents were used—Ethan, William and David—and the respective facial displays are shown in Figure 2. These facial displays were validated elsewhere (de Melo et al., 2012). The agents were referred to by their names throughout. Each participant saw a different agent in each condition, and they were randomly assigned to conditions.

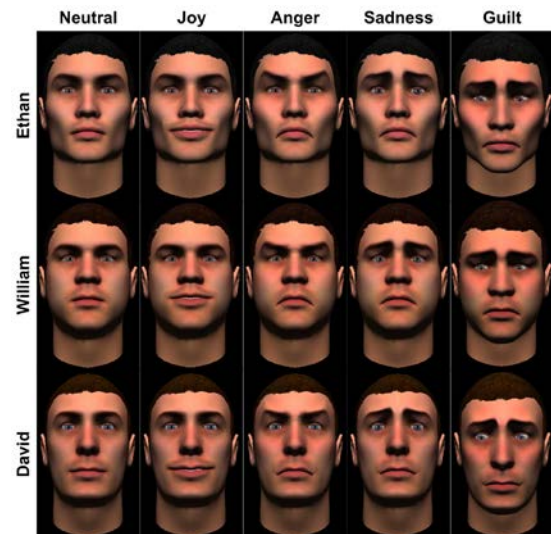


Figure 2: The emotion facial displays used in Experiment 1.

**Measures.** After watching the video of the agent's emotional reaction, we asked participants the following questions (the questions referred to the agents by their respective names): How much did the agent experience each of the following emotions: a) Joy b) Sadness c) Anger d) Guilt? (scale goes from 1, *not at all*, to 7, *very much*).

Even though several appraisal theories have been proposed (Ellsworth & Scherer, 2003), there tends to be agreement on which appraisals predict the emotions we consider in this experiment: joy occurs when the event is conducive to one's goals; sadness occurs when the event is not conducive to one's goals; anger occurs when the event is not conducive to one's goals and is caused by another agent; guilt occurs when the event is not conducive to one's goals and is caused by the self. Thus, two appraisal variables are of relevance here: (a) *conduciveness to goals*, which measures whether the event is consistent or inconsistent with the individual's goals; and, (b) *blameworthiness*, which measures whether the self or another agent is responsible for the event. After watching the video of the agent's emotional reaction, participants were asked the following questions about how was the agent appraising the outcome (Ellsworth & Scherer, 2003):

1. How pleasant for him was it to be in this situation?
2. At the time of experiencing the emotion, do you think he perceived that the consequences of the event did or would bring about positive, desirable consequences for him (e.g., helping him reach a goal, or giving pleasure)?
3. Was the situation obstructive or conducive to his goals?
4. Was what happened something that he regarded as fair?
5. How much did you think he blamed himself for the event?
6. How much did you think he blamed you for the event?

Following the appraisal perception questions, we asked the participant one question about perception of the agent's cooperativeness (scale goes from 1, *not at all*, to 7, *very much*): How likely is he to *cooperate* in the next round? (where "cooperate" was replaced by the respective action in the investment game).

**Participants.** We recruited four-hundred and five ( $N=405$ ) participants online using Amazon Mechanical Turk. This resulted in approximately 100 participants for each outcome. Gender distribution was as follows: *males*, 47.4%; *females*, 52.6%. Age distribution was as follows: *18 to 21 years*, 14.1%; *22 to 34 years*, 59.5%; *35 to 44 years*, 13.6%; *45 to 54 years*, 7.9%; *55 to 64 years*, 4.2%; *65 years and over*, 3.0%. Most participants were from the United States (57.8%) and India (29.6%). The education level distribution was as follows (current or expected degrees): *high school*, 15.8%; *college*, 57.5%; *Masters*, 23.0%; *Ph.D. or above*, 3.7%. Education majors and profession were quite diverse. Participants were paid USD \$1.02 and average participation time was 23 minutes.

## Results

**Effects on perception of cooperativeness.** For each outcome, we conducted a repeated-measures ANOVA to analyze the effect of emotion display on perception of cooperativeness. The means, standard deviations, significance levels and effect sizes are shown in Table 2. In the table, "Participant exploits" corresponds to the outcome where the agent cooperated but the participant defected. "Agent exploits" corresponds to the outcome where the participant cooperated but the agent defected.

Table 2: Effects on perception of cooperativeness.

<i>Mutual cooperation (n=103)</i>		<i>Participant exploits (n=101)</i>	
Neutral	3.18 (1.613)	Neutral	3.67 (1.715)
Joy	4.70 (1.739)	Anger	2.81 (2.077)
-	-	Sadness	2.99 (1.841)
Sig. (r)	.000* (.542)	Sig. (partial $\eta^2$ )	.000* (.078)
<i>Agent exploits (n=98)</i>		<i>Mutual defection (n=103)</i>	
Neutral	3.11 (1.716)	Neutral	3.55 (1.856)
Joy	2.37 (1.755)	Joy	3.47 (1.835)
Guilt	4.56 (2.081)	Anger	3.53 (2.100)
Sig. (partial $\eta^2$ )	.000* (.286)	Sig. (partial $\eta^2$ )	.920 (.001)

\*  $p < .05$ .

**Effects on perception of appraisals.** Questions 1 to 4 were highly correlated ( $\alpha = .850$ ) and, thus, were collapsed (averaged) into a single measure called *conduciveness to goals*. For each outcome, we conducted a repeated-measures ANOVA to analyze the effect of emotion display on conduciveness to goals, self-blameworthiness (question 5) and participant-blameworthiness (question 6). Means, standard deviations, significance levels and effect sizes are reported in Table 3.

Table 3: Effects on perception of appraisal.

	Conduciveness to Goals	Self-Blame	Participant-Blame
<i>Mutual cooperation (n=103)</i>			
Neutral	3.68 (0.896)	2.97 (1.620)	3.23 (1.746)
Joy	5.51 (0.852)	2.64 (1.652)	2.57 (1.551)
Sig. (r)	.000* (.853)	.067 (.181)	.000* (.362)
<i>Agent exploits (n=98)</i>			
Neutral	4.19 (1.089)	2.80 (1.699)	2.88 (1.633)
Joy	5.92 (0.823)	3.05 (2.078)	2.83 (1.759)
Guilt	3.23 (1.179)	4.39 (1.672)	2.84 (1.558)
Sig. (partial $\eta^2$ )	.000* (.671)	.000* (.224)	.950 (.000)
<i>Participant exploits (n=101)</i>			
Neutral	3.56 (1.038)	2.85 (1.676)	2.79 (1.768)
Anger	2.19 (0.868)	3.49 (1.659)	5.20 (1.588)
Sadness	2.40 (0.901)	4.56 (1.590)	3.92 (1.730)
Sig. (partial $\eta^2$ )	.000* (.545)	.000* (.248)	.000* (.466)
<i>Mutual defection (n=103)</i>			
Neutral	3.72 (0.757)	2.92 (1.453)	3.12 (1.635)
Joy	5.32 (0.856)	2.29 (1.493)	2.39 (1.523)
Anger	2.69 (0.856)	3.36 (1.726)	5.02 (1.621)
Sig. (partial $\eta^2$ )	.000* (.733)	.000* (.119)	.000* (.477)

\*  $p < .05$ .

**Mediation analysis.** In this subsection we present a causal steps approach multiple mediation analysis (Preacher & Hayes, 2008) of perceptions of appraisal on the effect of emotion displays on perception of cooperativeness. This method is an extension to multiple mediators of the single-mediation analysis proposed by Baron and Kenny (1986). Figure 3 summarizes the mediation model. The independent variables (IVs) were the classification questions for perception of joy, anger, sadness and guilt. The dependent variable (DV) was perception of cooperativeness. The proposed mediators were the perception of appraisal variables: conduciveness to goals, self-blame and participant-blame. According to this approach, there is

Table 4: Mediation analysis of perceptions of appraisals on the effect of emotions on perception of cooperativeness.

		IV → Mediators (a paths)			Mediators → DV (b paths)			Total Effect (c path)	Direct Effect (c' path)	Indirect Effect (ab paths)			
		Cn	SB	PB	Cn	SB	PB			Tot	Cn	SB	PB
mutual coop	Joy	.457* (.000)	-.015 (.793)	-.125 (.037)	.305* (.026)	-.023 (.820)	-.124 (.221)	.372* (.000)	.217* (.011)	.155* (.011)	.139* (.025)	.000 (.862)	.016 (.285)
	Anger	.501* (.000)	-.140* (.002)	-.360* (.000)	.234 (.067)	.254* (.001)	-.070 (.332)	-.023 (.675)	-.129 (.103)	.107 (.076)	.117 (.066)	-.036* (.020)	.0251 (.332)
mutual defect	Joy	-.411* (.000)	.235* (.000)	.624* (.000)	.153 (.122)	.261* (.001)	-.131 (.099)	.049 (.395)	.131 (.089)	-.083 (.127)	-.063 (.122)	.061* (.004)	-.081 (.098)
	Anger	.496* (.000)	-.149* (.003)	.065 (.132)	-.597* (.000)	.204* (.000)	-.152* (.023)	-.336* (.000)	-.001 (.993)	-.336* (.000)	-.296* (.000)	-.030* (.021)	-.010 (.206)
agent exploits	Joy	-.480* (.000)	.469* (.000)	-.033 (.548)	-.467* (.000)	.127* (.038)	-.127 (.055)	.521* (.000)	.232* (.003)	.288* (.000)	.224* (.000)	.060* (.043)	.004 (.565)
	Guilt	-.136* (.000)	.393* (.000)	.153* (.004)	.576* (.000)	.014 (.842)	.075 (.226)	*.076 (.143)	-.015 (.798)	-.061 (.060)	-.078* (.001)	.005 (.841)	.012 (.260)
human exploits	Sad	-.196* (.000)	.034 (.500)	.543* (.000)	.568* (.000)	-.002 (.977)	.109 (.127)	-.112* (.038)	-.060 (.350)	-.052 (.198)	-.111* (.000)	-.000 (.977)	.059 (.127)
	Anger												

Note. Cn = Conduciveness to goals; SB = Self-Blame; PB = Participant-Blame; CP = Coping Potential.

Values correspond to unstandardized regression coefficients (*p* values in parentheses).

\* *p* < .05.

mediation by a specific mediator  $M_x$  if: (1) the path,  $a_x$ , from the IV to the mediator is significant; (2) the path,  $b_x$ , from the mediator to the DV, when controlling for the IV, is significant; (3) the indirect effect,  $a_x b_x$ , from the IV to the DV, when controlling for the mediator, is significantly different than zero and greater than zero by a non-trivial amount. Moreover, there is mediation of the *set* of mediators when the sum of the indirect effects of all mediators is significantly different than zero. Furthermore, there is full mediation when the direct effect,  $c'$ , of the IV on the DV, when controlling for all the mediators, is non-significant. Table 4 shows the analysis: the shaded cells on the *a*, *b* and *ab* path columns represent that the causal-step requirement on the respective path has been passed.

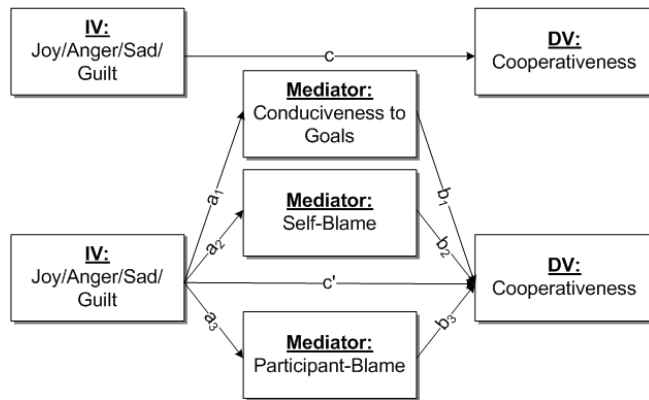


Figure 3: The multiple mediation model.

## Discussion

The results in Table 2 show that emotion displays impacted perception of the counterpart's cooperativeness: in mutual cooperation, participants perceived the happy agent to be more likely to cooperate than a neutral agent; when the participant exploited, participants perceived the angry or sad

agents to be less likely to cooperate than the neutral agent; finally, when the agent exploited, participants perceived the guilty agent to be more likely to cooperate than the neutral agent which, in turn, was more likely to cooperate than the happy agent. Notice these effects are compatible with findings in the literature (e.g., Van Kleef et al., 2010). Hypothesis H1 was, thus, confirmed. The results in Table 3 show that emotion displays impacted perception of appraisals in a way that was consistent with appraisal theories (Ellsworth & Scherer, 2003). For instance, happy agents were perceived to find the outcome conducive to their goals; sad or guilty agents were perceived to blame themselves for the (negative) outcome; and, angry agents were perceived to blame the participant for the (negative) outcome. Therefore, hypothesis H2 was also confirmed. Finally, the results in Table 4, show that appraisal variables, partially and sometimes fully, mediated the effect of emotion displays on perception of cooperativeness. For instance, in mutual cooperation, conduciveness to goals partially mediated the effect of joy; and, when the agent exploited, conduciveness to goals and self-blame fully mediated the effect of joy and partially mediated the effect of guilt. Hypothesis H3 was, thus, confirmed.

## Experiment 2

Spencer et al. (2005) argue that showing mediation statistically, as proposed by Baron and Kenny (1986), is no substitute to showing mediation experimentally. As an alternative to the statistical approach, they propose the experimental-causal-chain design, where each link of the proposed causal model is shown experimentally. Applying the experimental-causal-chain approach to our case means showing, experimentally, each of the causal links in the proposed causal model (Figure 1). The effect of emotion displays on perception of appraisals (causal link i) was already shown, experimentally, in the previous experiment.

In turn, the goal of Experiment 2 was to show experimentally the effect of perception of appraisals on perceptions of cooperativeness (causal link ii).

## Method

Experiment 2 used the same design and scenarios as Experiment 1. However, the manipulation consisted, instead of emotion displays, of textual expression of the appraisals. The mapping of emotion to appraisals followed predictions from appraisal theories (Ellsworth & Scherer, 2003) and is shown in Table 5. Participants were still introduced to the agents they imagined playing with, however, only a static image was shown of the (neutral) face. The textual expression of appraisals was simulated by typing at the bottom of the screen, as if simulating a chat interface.

Table 5: Mapping of emotions to expression of appraisals.

Emotion	Appraisal Expression
Neutral	I neither like, nor dislike this outcome
Joy	I like this outcome
Anger	I do NOT like this outcome and I blame YOU for it
Sadness	I do NOT like this outcome
Guilt	I do NOT like this outcome and I blame MYSELF for it

Regarding measures, after watching the video of the agent's reaction, we asked participants the same question about perception of cooperativeness as in Experiment 1.

We recruited two-hundred and two ( $N=202$ ) participants online using Amazon Mechanical Turk. This resulted in approximately 50 participants for each outcome. Gender distribution was as follows: *males*, 51.0%; *females*, 49.0%. Age distribution was as follows: *18 to 21 years*, 10.4%; *22 to 34 years*, 56.4%; *35 to 44 years*, 12.9%; *45 to 54 years*, 12.4%; *55 to 64 years*, 5.9%; *65 years and over*, 2.0%. Most participants were from the United States (66.3%) and India (22.8%). The education level distribution was as follows (current or expected degrees): *high school*, 15.3%; *college*, 62.9%; *Masters*, 18.3%; *Ph.D. or above*, 3.5%. Education majors and profession were quite diverse. Participants were paid USD \$1.02 and average participation time was 25 minutes.

## Results

For each outcome, we conducted a repeated-measures ANOVA to analyze the effect of emotion display on perception of cooperativeness. The means, standard deviations, significance levels and effect sizes are shown in Table 6. If we collapse the data from the two experiments, it becomes possible to analyze whether there was any interaction between Sample (Experiment 1 vs. Experiment 2) and Emotion (Neutral vs. Joy vs. Anger vs. Sadness vs. Guilt). Because the argument is that appraisals are part of the information retrieved from emotion displays, we expected there to be no interactions. Table 6 also shows these interactions.

## Discussion

The results in Table 6 show that expression of appraisals impacted perceptions of cooperativeness: in mutual cooperation, participants perceived the happy agent to be more likely to cooperate than a neutral agent; when the participant exploited, participants perceived the angry or sad agents to be less likely to cooperate than the neutral agent; finally, when the agent exploited, participants perceived the

guilty agent to be more likely to cooperate than the neutral agent which, in turn, was more likely to cooperate than the happy agent. Notice these are the same patterns as in Experiment 1 and, thus, our hypothesis H4 was confirmed. Finally, notice the Sample x Emotion interaction was not significant for any of the outcomes.

Table 6: Effects on perception of cooperativeness.

<i>Mutual cooperation (n=52)</i>		<i>Participant exploits (n=48)</i>	
Neutral	3.27 (1.693)	Neutral	3.65 (1.839)
Joy	4.85 (1.841)	Anger	2.73 (2.029)
-	-	Sadness	3.00 (1.935)
Sig. (r)	.000* (.654)	Sig. (partial $\eta^2$ )	.001* (.133)
Sample x		Sample x	
Emotion, Sig. (r)	.267 (.005)	Emotion, Sig. (r)	.963 (.000)
<i>Agent exploits (n=52)</i>		<i>Mutual defection (n=50)</i>	
Neutral	4.06 (1.650)	Neutral	3.60 (1.990)
Joy	2.81 (2.077)	Joy	3.20 (1.874)
Guilt	5.31 (1.639)	Anger	3.46 (2.159)
Sig. (partial $\eta^2$ )	.000* (.354)	Sig. (partial $\eta^2$ )	.332 (.022)
Sample x		Sample x	
Emotion, Sig. (r)	.473 (.005)	Emotion, Sig. (r)	.701 (.002)

\*  $p < .05$ .

## General Discussion

This paper presents insight into the mechanism for the social effects of emotion in a social dilemma. Two experiments were described that suggest a causal model (Figure 1) where emotion displays lead the receiver to infer how the sender is appraising the ongoing interaction and, these perceptions of appraisal, in turn, lead to inferences about the sender's propensity for cooperation in the dilemma. The experiments support this model by establishing experimentally both links in the model (Spencer et al., 2005) and by providing statistical evidence that perceptions of appraisal mediate the effect of emotion displays on perceptions of cooperativeness (Preacher & Hayes, 2008). We refer to the mechanism suggested by this causal model as reverse appraisal.

**Implications for emotion theory.** The results presented in the paper provide further evidence that emotions serve important social functions (Frijda & Mesquita, 1994; Keltner & Kring, 1998). For instance, whereas anger signaled a willingness to punish a non-cooperator with non-cooperation, guilt signaled regret for one's non-cooperation and a willingness to cooperate in the future. The paper further proposes that reverse appraisal is a useful framework for the social functions of emotion. Reverse appraisal suggests that an important component of the information retrieved from emotion displays refers to how the counterpart is appraising the ongoing interaction. This information is then used to infer the counterpart's mental states, such as his or her propensity for cooperation in a social dilemma. Finally, reverse appraisal can potentially generalize beyond social dilemmas and be viewed as a general mechanism for interpretation of emotion displays. This is, in fact, a promising line of future inquiry.

**Implications for decision-theory.** Van Kleef et al. (2010) argue emotion displays can influence people's decision

making through affective processes (e.g., emotional contagion) or inferential processes. Regarding the latter, Van Kleef et al. suggest that “each discrete emotion has its own antecedents, appraisal components, relational themes, and action tendencies” (p.48) and, thus, “observing a particular emotion in another person provides relatively differentiated information about how that person regards the situation” (p.53). This paper provides empirical evidence for such “differentiated information” in emotion displays which, as suggested by the results, pertains to perceptions of appraisal. Our results also reinforce findings regarding the importance of context for the interpretation of emotion (de Melo et al., 2011; Hareli & Hess, 2010; Van Kleef et al., 2010). For instance, the results showed that people reacted differently to the same smile when it was shown after mutual cooperation or after the agent exploited them. Finally, the results support Schug et al.’s (2010) contention that people are capable of identifying non-cooperators from emotion displays and punish accordingly.

**Limitations and future work.** Since the results from our experiments were promising, we plan in the near future to repeat both experiments with the complete factorial design. Such design will clarify the effect of emotion in pairings that were left unexplored (e.g., expression of anger in mutual cooperation). Moreover, complementing this paper’s focus of comparing the effect of emotion within the same outcome, the factorial design will allow us to study the effect of emotion *across* outcomes. Another limitation we intend do address is that the current causal model does not predict what will people decide after making inferences about the counterpart’s likelihood of cooperation. Notice the link between perceptions of cooperation and the decision to cooperate is not simple. For instance, whereas a pro-social might cooperate, a pro-self might exploit the cooperator (Steinel & de Dreu, 2004). Finally, to understand how reverse appraisal generalizes beyond the prisoner’s dilemma, it is important to assess whether perceptions of appraisal also mediate the effects of emotion displays in more social decision making tasks such as other social dilemmas (e.g., public goods), trust games and negotiation.

### Acknowledgments

This work was sponsored by the Fundação para a Ciência e a Tecnologia (FCT) grant #SFRH-BD-39590-2007; the Air Force Office of Scientific Research under grant FA9550-09-1-0507; and, the NSF under grant IIS-0916858. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

### References

Baron, R., & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *J. Pers. Soc. Psychol.*, 51, 1173-1182.

Blanchette, I., & Richards, A. (2010). The influence of affect on higher level cognition: A review of research on interpretation, judgment, decision making and reasoning. *Cognition & Emotion*, 15, 1-35.

Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, C., & Bailenson, J. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychol. Inq.*, 13, 103-124.

Boone, R., & Buck, R. (2003). Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. *J. Nonverbal Behav.*, 27, 163-182.

de Melo, C., Carnevale, P., & Gratch, J. (2011). Reverse appraisal: Inferring from emotion displays who is the cooperator and the competitor in a social dilemma. In *Proceedings of CogSci 2011*.

de Melo, C., Carnevale, P., Gratch, J. (2012). The impact of emotion displays in embodied agents on emergence of cooperation with people. *Presence-Teleop. Virt.*, 20, 449-465.

Ellsworth, P., & Scherer, K. (2003). Appraisal processes in emotion. In R. Davidson, K. Scherer, H. Goldsmith (Eds.), *Handbook of Affective Sciences* (pp. 572-595). New York, NY: Oxford University Press.

Frank, R. (1988). *Passions within reason*. New York, NY: Norton.

Frijda, N., & Mesquita, B. (1994). The social roles and functions of emotions. In S. Kitayama & H. Markus (Eds.), *Emotion and culture: Empirical studies of mutual influence* (pp. 51-87). Washington, DC: American Psychological Association.

Hareli, S., & Hess, U. (2010). What emotional reactions can tell us about the nature of others: An appraisal perspective on person perception. *Cognition & Emotion*, 24, 128-140.

Keltner, D., & Kring, A. M. (1998). Emotion, social function, and psychopathology. *Rev. Gen. Psychol.*, 2, 320-342.

Kiesler, S., Waters, K., and Sproull, L. (1996). A prisoner’s dilemma experiment on cooperation with human-like computers. *J. Pers. Soc. Psychol.*, 70, 47-65.

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annu. Rev. of Sociol.*, 24(1), 183-214.

Preacher, K., & Hayes, A. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav. Res. Meth.*, 40, 879-891.

Schug, J., Matsumoto, D., Horita, Y., Yamagishi, T., & Bonnet, K. (2010). Emotional expressivity as a signal of cooperation. *Evol. Hum. Behav.*, 31, 87-94.

Spencer, S., Zanna, M., & Fong, G. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *J. Pers. Soc. Psychol.*, 89, 845-851.

Steinel, W., & de Dreu, C. (2004). Social motives and strategic misrepresentation in social decision making. *J. Pers. Soc. Psychol.*, 86(3), 419-434.

Van Kleef, G., De Dreu, C., & Manstead, A. (2010). An interpersonal approach to emotion in social decision making: The emotions as social information model. *Adv. Exp. Soc. Psychol.*, 42, 45-96.

# Neural correlates of social perception: The posterior superior temporal sulcus is modulated by action rationality, but not animacy

Ben Deen (bdeen@mit.edu) and Rebecca R. Saxe (saxe@mit.edu)

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology  
Cambridge, MA 02139

## Abstract

Recent research has investigated the neural basis of social perception, the ability to make high-level social inferences from perceptual information. The right posterior superior temporal sulcus (pSTS) has been identified as a candidate region for this ability, but the specific processes to which the pSTS contributes remain unclear. In the present study, we investigated the neural correlates of social perception using simple animated geometric shape stimuli, separately manipulating the perceived animacy, goal-directedness, and path rationality in the animations. We did not find an increased pSTS response to animate or goal-directed animations. However, we found that across conditions, the pSTS response tracked path rationality, with stronger responses to irrational paths. This is consistent with prior neuroimaging research on the perception of human actions, and supports the claim that the pSTS is involved in action understanding.

**Keywords:** social perception, fMRI, superior temporal sulcus

## Introduction

Humans have a remarkable ability to infer the dispositions and intentions of other agents from perceptual information, and specifically from motion patterns such as hand and body motion, gaze shifts, and facial motion. This ability, termed social perception, comprises a number of subprocesses: the detection of agents in an environment, perceptual analysis of their motion, inference about social properties from the agent's actions and their context, and prediction of future actions based on these properties.

Recent research has begun to probe the neural basis of these processes, although the relevant brain regions and their specific functional role is still debated. One line of research has pointed to the right posterior superior temporal sulcus (pSTS) as a critical region for social perception (Allison, Puce, & McCarthy, 2000). This region responds more strongly to (human) biological motion than motion of inanimate objects (e.g. Grossman et al., 2000; Pelphrey et al., 2003). These responses might relate to the detection or perceptual analysis of biological motion, to higher-level processing of the intentions underlying the actions, or to some combination thereof.

Another set of studies indicates that the pSTS response to human actions is modulated by inferred intentions. Specifically, actions that violate inferred intentions in a given context, such as twisting empty space next to a gear rather than a gear itself, elicit a stronger pSTS response than the expected actions, across a range of contexts and specific actions (Brass, Schmitt, Spengler, & Gergely, 2007;

Pelphrey, Morris, & McCarthy, 2004; Pelphrey, Singerman, Allison, & McCarthy, 2003; Saxe, Xiao, Kovacs, Perrett, & Kanwisher, 2004; Vander Wyk, Hudac, Carter, Sobel, & Pelphrey, 2009). Such actions have been referred to as incongruent, irrational, or unexpected.

This effect has been interpreted as evidence that the pSTS is sensitive to the goals or intentions underlying human motion. For instance, Pelphrey et al. (2004) argued that the pSTS is involved in predicting actions in a given context based on an "intentional stance," in which actions are determined by a goal state and an assumption that the agent will choose the most efficient means to achieve the goal given situational constraints. They proposed that when this prediction is violated, the pSTS must engage in extra processing to explain the observed action in other terms, which would explain its stronger response to unexpected actions.

Another line of research supporting the role of pSTS in action understanding as employed animations of simple geometric shapes as stimuli (Castelli, Happé, Frith, & Frith, 2003; Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007). These studies have found a stronger pSTS response to animations depicting social interactions between animate shapes, compared with animations of shapes moving as inanimate physical objects. This demonstrates that the role of the pSTS extends to animations that lack the form and motion kinematics of humans, but imply intentional action. However, such comparisons have been largely visually uncontrolled, and could also reflect one of a number of processes: detecting agents, processing of their motion or intentions, or processing of interactions between multiple agents.

The present study aimed to investigate the neural correlates of social perceptual processes, using geometric shape stimuli. In particular, we use dot-chain stimuli perceived as slithering snakes or worms, which provide a strong percept of animacy without the need for multiple, interacting agents (Gao, New, & Scholl, 2011). This ensures that any effects observed do not relate to processing interactions between agents (c.f. Centelles, Assaiante, Nazarian, Anton, & Schmitz, 2011). To investigate each of the subprocesses listed above, we separately manipulated the perceived animacy, goal-directedness, and path rationality (or expectedness) of the animations. We first performed a behavioral study, eliciting judgments about these animations on various dimensions. The animations were then used as stimuli for an fMRI experiment, to investigate the response of the pSTS, as well as motion-sensitive area MT+, as a control region.



## Methods

### Experiment 1: Behavioral study

**Participants** For the behavioral study, responses were gathered using Amazon Mechanical Turk. There were 16 types of animation per condition, and 15 responses were elicited for each animation, yielding a total of 240 responses per condition. Participants were constrained to be from the United States, and to have a minimum 95% approval rating from prior Turk studies. The survey included several foil questions (e.g., what is the color of the dots?), and responses with incorrect answers to these questions were rejected.

**Stimuli** The stimuli consisted of a set of 4s-long animations of dots (i.e. circles) and dot-chains moving within a square-shaped environment, with walls present in some conditions as obstacles. For the head dot of the snake, motion was determined using the chase-subtlety algorithm from Gao, Newman, & Scholl (2009). In this algorithm, the velocity of the dot has a fixed magnitude, with a direction that updates periodically (every 5 frames or .167s, in the present study). The direction is chosen probabilistically: if the angle that directs the dot toward its goal is denoted  $\alpha$  and the subtlety parameter is denoted  $\gamma$ , the new direction is chosen from a uniform distribution over the interval  $[\alpha - \gamma, \alpha + \gamma]$ , where  $\gamma = \pi/12$  in this study. This results in a dot that takes a slightly winding path toward a goal. Tail dots in the chain, if present, followed the path taken by the head dot with a slight lag.

Conditions 1-4 were intended to manipulate animacy and goal-directedness in a 2x2 design (see figure 1 for a schematic depiction of each condition). Animacy was modulated by the presence or absence of six tail dots, leading to the percept of a worm or snake. Goal-directedness was modulated by the presence of a goal-dot at the end of the trajectory.

Conditions 5-8 were intended to manipulate path rationality. Stimuli in conditions 5-7 were considered animate and goal-directed, but involved trajectories with a bend, which was either around a wall or around nothing. These conditions were 5) rational (full wall), 6) semi-rational (half of a wall), and 7) irrational (no wall). As a visual control, walls were added to conditions 1-4 and 7-8, which were not relevant to the paths. In condition 8 (wander), the dot-chain had no goal and increased subtlety ( $\gamma = \pi/4$ ), leading to a percept of a randomly wandering snake, intended as a highly irrational, unexplainable action.

For each condition, 4 specific animations were designed with distinct trajectories; these stimuli were rotated 0°, 90°, 180°, and 270° to create 16 animations per condition. Two visual confounds should be noted: the animate condition had more dots and therefore more motion than the inanimate condition; and the wander condition had more changes in motion than all other conditions. These issues are further discussed below.

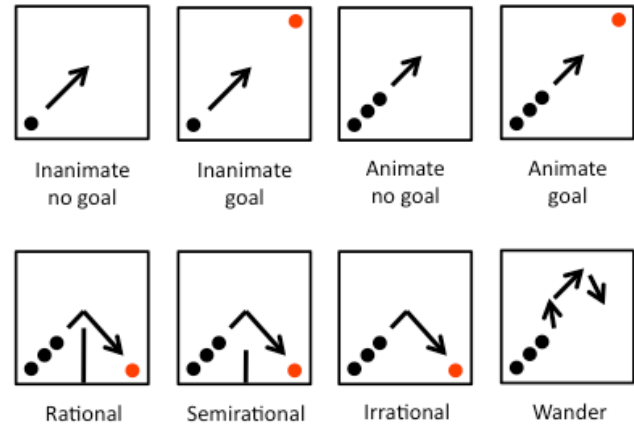


Figure 1: Schematic illustrations of the animation stimuli (not to scale). Note that the actual trajectories were not straight lines, but winding paths (see Methods section).

**Behavioral measures** Participants viewed the animations and were asked to respond to the following questions on a seven-point scale: 1) How much did the moving dot look like a living, animate thing, as opposed to an inanimate physical object? 2) To what extent did the moving dot appear to have a goal or goals? 3) To what extent did the dot's path seem strange or irrational? Additionally, several foil questions were asked to ensure meaningful responses.

**Data analysis** We performed several planned unpaired two-sample *t*-tests to test the specific effects of interest. We first tested the effect of having a tail (conditions 3 and 4 versus 1 and 2) on animacy ratings. We then tested the effects of having a goal dot (2 and 4 versus 1 and 3), of path irrationality (7 versus 5), and of wandering over irrationality (8 versus 7), on both goal-directedness and irrationality ratings. Additionally, we performed a post-hoc test for the effect of having a goal dot for animations with a tail (4 versus 3) on animacy ratings.

### Experiment 2: fMRI study

**Participants** 20 subjects (aged 19-28, mean 23.1; 10 female) were recruited for the fMRI study. All participants had normal or corrected-to-normal vision and no history of neurological or psychiatric disorders, and gave written, informed consent in accordance with the requirements of the MIT institutional review board.

**Stimuli** The animations used in the fMRI experiment were the same as those used in the behavioral study. Stimuli were presented in a jittered, event-related design, with a variable inter-stimulus interval of 0-15 seconds, during which a central fixation cross was presented. The experiment comprised 8 blocks lasting 9 minutes and 44 seconds, each containing 8 stimuli per condition, for a total of 64 stimuli per condition. Participants performed a one-back task on animations during the scan, to maintain attention; repeat trials were not included in the analysis.



Additionally, each subject received a localizer scan intended to define the pSTS and motion-sensitive area MT. This consisted of three conditions in a blocked design: biological motion (point-light displays [PLDs] depicting human motion; cf Grossman et al. 2000), scrambled motion (PLDs with initial dot positions scrambled), and static luminance change (static dots changing in luminance). Each subject received 2 or 3 runs lasting 7 minutes and 24 seconds each, and comprising 6 12s-long blocks per condition separated by a 12s interstimulus interval. Participants performed a one-back task on individual animations within the blocks, to maintain attention.

**fMRI Data Acquisition** Data were acquired on a 3T Siemens Tim Trio scanner, with a 32-channel head coil. Following high-resolution anatomical scans, functional images were acquired with an echo planar imaging pulse sequence sensitive to blood-oxygen-dependent (BOLD) contrast (repetition time [TR] = 1s, echo time [TE] = 30ms, flip angle = 70°, voxel size 3x3x3mm, matrix 64x64, 16 axial slices). Because of our interest in specific brain regions, we used a sequence with limited coverage (of visual cortex and the STS), but a TR of 1s for increased power and temporal resolution. The first four volumes of each acquisition were discarded to allow the system to reach steady state. For localizer scans, a similar pulse sequence was used, but with TR=2 and full brain coverage (32 axial slices).

**fMRI Data Analysis** Preprocessing and analysis of fMRI data was carried out using the FMRIB Software Library (FSL) version 4.1.8, supplemented with Freesurfer 4.5. Preprocessing steps included rigid-body motion correction, correction for interleaved slice timing, brain extraction, spatial smoothing (5mm FWHM Gaussian kernel), and highpass temporal filtering (100s cutoff). Functional images were registered to anatomical images using Freesurfer's bbrgister; anatomical images were in turn normalized to MNI space using FSL's nonlinear registration image registration tool (FNIRT).

For data analysis, whole-brain general linear model-based analyses were initially performed for the main task and localizer, for the purpose of defining regions-of-interest (ROIs) in individual subjects. Regressors were defined as boxcar functions with nonzero values during the duration of the stimuli; these were then convolved with a canonical double-gamma hemodynamic response function. FSL's FILM prewhitening was applied to account for residual autocorrelation. Statistical maps were thresholded with an initial cutoff of  $Z > 2.3$ , followed by Gaussian random field theory-based thresholding with a cluster-wise threshold of  $P < .05$ , to correct for multiple comparisons.

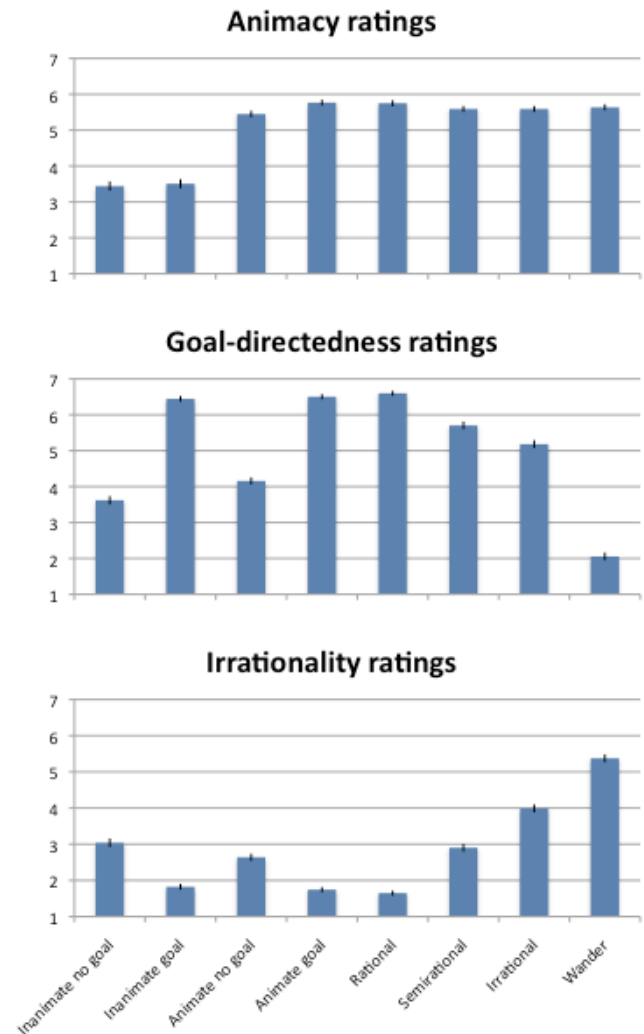


Figure 2: Behavioral responses. Plot of responses to three questions—regarding animacy, goal-directedness, and irrationality—for the eight conditions. Error bars give standard error.

To define the right pSTS, the main task was used rather than the localizer, because the latter did not consistently yield pSTS responses in individual subjects. The contrast of all conditions versus rest in the main task was used to define the pSTS, because this contrast is orthogonal to any balanced between-condition comparison. As a control, we investigated responses in right MT+, a motion-sensitive region thought include retinotopic areas MT, MST, and possibly others (Amano, Wandell, & Dumoulin, 2009). This was defined using the localizer scan, by contrasting scrambled motion with static luminance change. Regions were defined as all active voxels within a 7.5mm-radius sphere around the peak coordinate within an anatomical search space, intersected with a gray matter mask derived using Freesurfer. The search spaces consisted of the STS (for pSTS), and lateral occipito-temporal cortex (for MT+).

Mean betas values across the ROI were extracted for each subject. Planned, paired two-sample *t*-tests were performed

for each ROI, testing for effects of 1) animacy (conditions 3 and 4 vs 1 and 2), 2) goal-directedness (2 and 4 vs 1 and 3), 3) irrationality (7 vs 5), and 4) wandering over irrationality (8 vs 7). Responses were averaged when combining two conditions. Additionally, a post-hoc test assessed the effect of goal-directedness for animations with a tail (condition 4 versus 3) on the pSTS response.

## Results

### Behavioral results

The behavioral results are shown in figure 2. As predicted, the presence of a tail or dot-chain significantly increased the percept of animacy ( $t[958] = 19.43, p < 10^{-70}$ ). In spontaneous post-scan self-reports from subjects who participated in the fMRI experiment (a separate group of subjects), many described the dot-chain stimuli as either a “worm,” “snake,” or “tadpole,” that was “swimming” or “wiggling.” Additionally, we observed that for stimuli with tails, animations that contained a visible goal dot were rated as more animate than those without. A post-hoc test of this difference was significant ( $t[478] = 2.56, p < .02$ ).

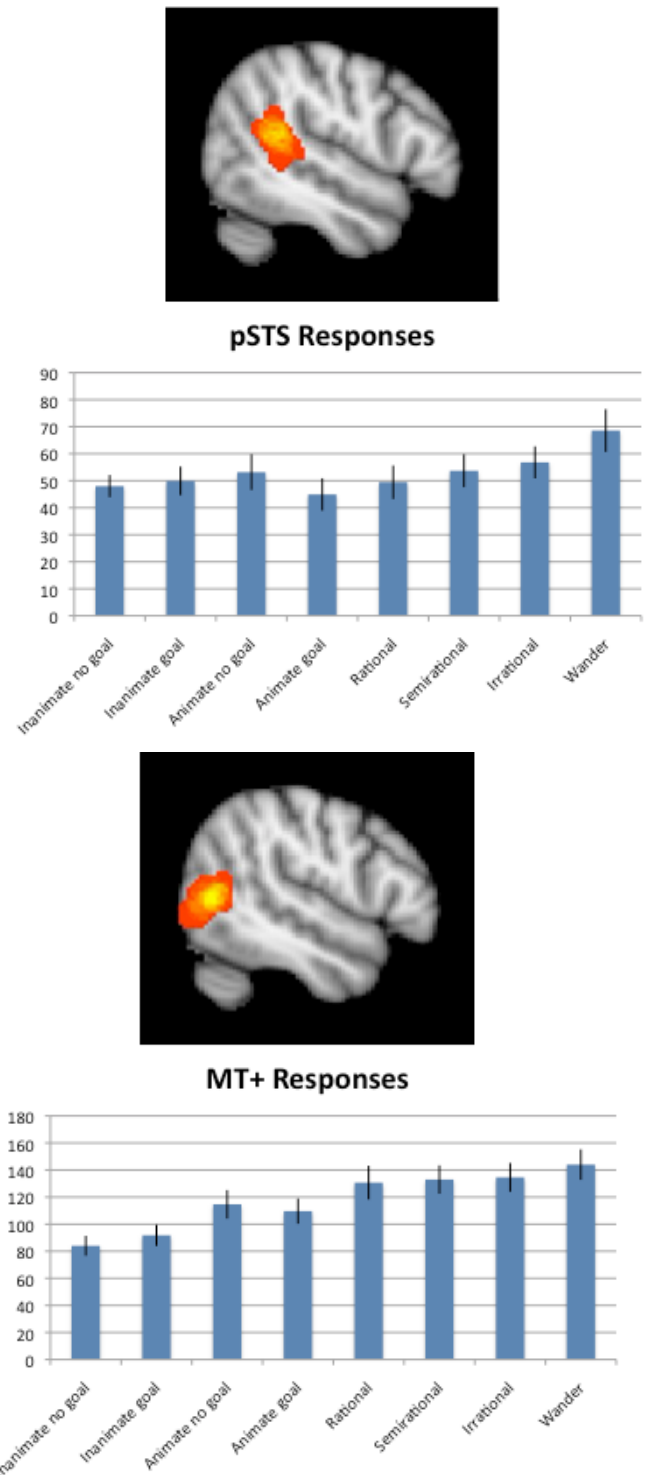
Ratings of goal-directedness were increased by the presence of a goal dot, as expected ( $t[958] = 25.40, p < 10^{-108}$ ). Additionally, goal-directedness ratings were lower for irrationality than rational stimuli ( $t[478] = -13.86, p < 10^{-36}$ ), and for wandering than irrational stimuli ( $t[478] = -24.79, p < 10^{-87}$ ).

Ratings of irrationality were higher for irrational than rational paths, as expected ( $t[478] = 18.00, p < 10^{-54}$ ). Additionally, they were higher for wandering than irrational paths ( $t[478] = 8.96, p < 10^{-17}$ ). The presence of a goal dot also influenced irrationality ratings, with higher ratings for stimuli without a visible goal ( $t[958] = -11.27, p < 10^{-27}$ ). Thus, we found an inverse relationship between ratings of goal-directedness and irrationality: namely, animations depicting an efficient path toward a clear goal were rated as highly goal-directed and rational, while paths that lacked a clear goal or used an inefficient trajectory were rated as less goal-directed and more irrational.

### fMRI results

Results from the right pSTS ROI analysis are shown in figure 3. The ROI was found in 19 out of 20 subjects. We found no effect of the animacy manipulation ( $t[18] = -.03, p = .98$ ), nor the goal dot manipulation ( $t[18] = -1.17, p = .26$ ). However, the pSTS did respond more strongly to irrational than rational stimuli ( $t[18] = -2.25, p < .05$ ), and to wandering than irrational stimuli ( $t[18] = 3.03, p < .01$ ).

Additionally, we observed that among animate stimuli, or stimuli with a tail, the pSTS had a lower response to animations with a visible goal dot. A post-hoc test for this comparison was significant ( $t[18] = -2.30, p < .05$ ). Thus, the pSTS response to animate stimuli tracked behavioral ratings of irrationality, but did not correspond to ratings of animacy or goal-directedness.



Figures 3 (above) and 4: Mean beta values extracted from right posterior superior temporal sulcus (pSTS, figure 3) and right MT+ (figure 4) regions of interest (ROIs). Error bars give standard error. The images above the bar plots show the locations of ROIs across subjects: for each voxel, the value plotted is the fraction of subjects whose ROI contained this voxel.

Results from the right MT+ ROI analysis are shown in figure 4. The ROI was found in 19 out of 20 subjects. For this ROI, there was a main effect of the animacy manipulation ( $t[18] = 5.57, p < 10^{-4}$ ). This is to be expected for a retinotopic region, insofar as the dot-chain stimuli occupied more of the visual field than the individual dot stimuli, and therefore this difference may not reflect the processing of animacy.

There was no effect of goal-directedness ( $t[18] = .52, p = .61$ ) or irrationality ( $t[18] = .97, p = .34$ ) on the MT+ response. These comparisons were tightly controlled for the magnitude and direction of motion, so no differences relating to motion processing were expected.

There was an effect of wandering over irrational stimuli in MT+ ( $t[18] = 2.71, p < .02$ ). This effect may also result from motion processing. Although the magnitude of motion is equated across wander and irrational conditions, the direction and derivatives thereof are not controlled. A larger number of changes in motion direction in the wander condition may have lead to decreased adaptation of direction-specific neural responses in MT+, and therefore an increased BOLD signal.

## Discussion

We have shown behaviorally that dot-chain stimuli governed by a simple motion algorithm can evoke a strong percept of animacy, and in certain conditions, goal-directedness, replicating and extending the findings of Gao et al. (2011). Furthermore, we found that the right pSTS response to these stimuli is not stronger for stimuli rated as animate or goal-directed. However, this response was modulated by path irrationality: for conditions 3-8 (the conditions rated as highly animate), pSTS activity corresponds well with irrationality ratings, as can be seen by comparing figures 2 and 3. By comparison, activity in right MT+ was not generally modulated by irrationality, instead tracking the amount of motion and change in motion in the stimuli, as expected.

Several results of interest came from our behavioral analysis. We found that ratings of goal-directedness and path irrationality had an inverse relationship. Straight paths without goal dots were rated as more irrational than those with visible goals, and inefficient trajectories toward a goal were rated as less goal-directed than efficient trajectories. Thus, these ratings may have both derived from a common implicit quantity, perhaps corresponding to the extent to which an action can be explained in terms of perceptible goals and environmental constraints (e.g. Gergely & Csibra, 2003).

Furthermore, we found that for dot-chain stimuli, which were perceived as highly animate, the presence of a goal dot had a small but significant influence on ratings of animacy. This is consistent with the hypothesis that goal-directedness provides a cue to animacy (e.g. Shultz and McCarthy, 2011). This result was unexpected and assessed with a post-hoc test, and thus should be independently replicated; however, we note that we have another, unpublished dataset

consisting of Mechanical Turk responses to similar stimuli, in which this effect was also observed.

Our imaging results show that with these stimuli, the pSTS response is not modulated by a large difference in perceived animacy between dot-chain and individual dot stimuli. This result appears inconsistent with claims that the pSTS is generally involved in the detection of agents or animate beings (e.g. Gobbini et al., 2011; Shultz and McCarthy, 2011). This finding is not directly inconsistent with any prior empirical result in the literature, to our knowledge, because prior contrasts involving animacy (e.g. faces versus nonfaces, biological motion versus scrambled motion, Heide-Simmer animations versus control animations) have been confounded with other factors (such as specific static or dynamic visual properties, the presence of a human, or the presence of an interaction), and thus cannot be considered pure animacy contrasts.

Another interpretation of these data is that the pSTS is involved in the detection of animacy, but relies on local cues such as the motion of individual dots in our animations, which are similar for the animate and inanimate conditions. This interpretation must invoke other processes to explain the large behavioral difference in animacy judgments for dots and dot-chains. However, this explanation appears inconsistent with the fact that the pSTS response to human motion is modulated by global form, and not just local cues (e.g. Grossman et al. 2000), unless this modulation relates to a process separate from agent detection.

The pSTS response in our data was also not increased by perceived goal-directedness. This is consistent with findings that the right pSTS responds similarly to intentional and externally caused human movements (Morris, Pelphrey, & McCarthy, 2008), and to goal-directed and non-goal-directed actions by robots (Shultz and McCarthy, 2011). This result might be interpreted as evidence against a role of pSTS in processing action goals. However, given the inverse relationship observed between ratings of goal-directedness and irrationality, there is another potential explanation. This region may apply an assumption that actions by animate beings are intentional, and attempt to explain all such actions. In this case, actions with a visible goal may be easier to explain, and thus evoke a weaker pSTS response, as observed for animate stimuli.

While the pSTS response in the present study did not increase with animacy or goal-directedness, it did track the perceived irrationality of the actions depicted. This is consistent with prior findings of irrationality effects during the perception of human actions, as described above, and extends these results to nonhuman agents depicted by simple geometric shape animations. Thus, whatever computations underlie this irrationality effect are likely similarly applied to the actions of human and nonhuman agents.

We note that animations in the rational, irrational, and semirational conditions were perfectly controlled for visual motion; therefore motion cannot be driving the differences observed. The wander condition did have a motion change

confound, as noted above. However, given the similar pSTS response to animate and inanimate conditions, which had a substantial difference in visual motion, we consider it implausible that the high response to the wander condition in this region results from motion properties.

As discussed above, the irrationality effect has been interpreted as supporting a role of the pSTS in action understanding, or inferring goals of actions and predicting future actions based on these goals. There are a number of interpretations of the irrationality effect consistent with this claim. For instance, this response might relate to the inference of a more complex goal structure underlying irrational actions. On this hypothesis, the pSTS tries to rationalize all actions, including ostensibly irrational ones, and simply requires a more complex explanation for the latter, perhaps positing extra goals that weren't immediately inferred from the context. Another possible interpretation is that this response constitutes an error detection signal for actions. On this hypothesis, the pSTS response doesn't reflect a reappraisal of the causal structure behind an irrational action, but simply reflects a signal indicating that the inferred structure was not correct. Future research should attempt to distinguish between these hypotheses.

Another question is of the specificity of this effect to actions. Does the right pSTS respond to any unexpected event, or more specifically, to unexpected visual motion events? While our current data doesn't speak to this question, Saxe et al. (2004) showed that while the pSTS responds more strongly when a walking human pauses behind a bookshelf than when he walks without pause, this isn't the case for gliding objects. This provides some preliminary evidence that this effect is specific to intentional actions, but this question should be followed up in subsequent studies.

In sum, we have shown that the pSTS response to animations of geometric shape motion is not increased by animacy or goal-directedness, but is modulated by action rationality. Future research should explore the computations that underlie this effect, and their precise contribution to action understanding.

## Acknowledgments

We are grateful to Hilary Richardson and Nicholas Dufour for help with data collection, and Tao Gao for helpful discussion. This research was funded by the David and Lucile Packard foundation.

## References

Allison, T., Puce, A., McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Sciences*, 4(7), 267-278.

Amano, K., Wandell, B.A., Dumoulin, S.O. (2009). Visual field maps, population receptive field sizes, and visual field coverage in the human MT+ complex. *Journal of Neurophysiology*, 102(5), 2704-2718.

Brass, M., Schmitt, R.M., Spengler, S., Gergely, G. (2007). Investigating action understanding: Inferential processes versus action simulation. *Current Biology* 17, 2117-2121.

Castelli, F., Happé, F., Frith, U., Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage* 12(3), 314-325.

Centelles, L., Assaiante, C., Nazarian, B., Anton, J.-L., Schmitz, C. (2011). Recruitment of both the mirror and the mentalizing networks when observing social interactions depicted by point-lights: A neuroimaging study. *PLoS One* 6(1), e15749.

Gao, T., Newman, G.E., Scholl, B.J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, 59(2), 154-179.

Gao, T., New, J.J., Scholl, B.J. (2011). Perceived biological agency in a slithering snake animation. Poster presented at the annual meeting of the *Vision Sciences Society*, 5/10/11, Naples, FL.

Gergeley G., Csibra G. (2003). Teleological reasoning in infancy; the naïve theory of rational action. *Trends in Cognitive Science*, 7(7), 287-292.

Gobbini, M.I., et al. (2011). Distinct neural systems involved in agency and animacy detection. *Journal of Cognitive Neuroscience*, 23(8), 1911-1920.

Gobbini, M.I., Koralek, A.C., Bryan, R.E., Montgomery, K.J., Haxby, J.V. (2007). Two takes on the social brain: a comparison of theory of mind tasks. *Journal of Cognitive Neuroscience*, 19(11), 1803-1814.

Grossman, E.D., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., et al. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience*, 12(5), 711-720.

Morris, J.P., Pelphrey, K.A., McCarthy, G. (2008). Perceived causality influences brain activity evoked by biological motion. *Social Neuroscience*, 3(1), 16-25.

Pelphrey, K.A., Mitchell, T.V., McKeown, M.J., Goldstein, J., Allison, T., McCarthy, G. (2003). Brain activity evoked by the perception of human walking: Controlling for meaningful coherent motion. *The Journal of Neuroscience*, 23(17), 6819-6825.

Pelphrey, K.A., Morris, J.P., McCarthy, G. (2004). Grasping the intentions of others: The perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *Journal of Cognitive Neuroscience*, 16(10), 1706-1716.

Pelphrey, K.A., Singerman, J.D., Allison, T., McCarthy, G. (2003). Brain activation evoked by perception of gaze shifts: The influence of context. *Neuropsychologia*, 41(2), 156-170.

Saxe, R., Xiao, D.-K., Kovacs, G., Perrett, D.I., Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia*, 42(11), 1435-1446.

Vander Wyk, B.C., Hudac, C.M., Carter, E.J., Sobel, D.M., Pelphrey, K.A. (2009). Action understanding in the superior temporal sulcus region. *Psychological Science*, 20(6), 771-777.

# The Role of Linguistic Labels in Infants' Categorization: An Eye Tracking Study

**Wei (Sophia) Deng (deng.69@osu.edu)**

Department of Psychology and Center for Cognitive Science  
The Ohio State University  
209C Ohio Stadium East, 1961 Tuttle Park Place  
Columbus, OH 43210 USA

**Vladimir M. Sloutsky (sloutsky.1@osu.edu)**

Department of Psychology and Center for Cognitive Science  
The Ohio State University  
208C Ohio Stadium East, 1961 Tuttle Park Place  
Columbus, OH 43210 USA

## Abstract

How do words affect categorization? According to one theoretical account, even early in development, labels are category markers and are different from other features. According to another theory, early in development, labels are part of the input and are no more than other features. The current study addressed this issue by examining the effects of labels on category learning in 8- to 12- month infants. Infants were familiarized with exemplars from one category in either label-defined or motion-defined condition and then tested with prototypes from learned category and novel category. Eye tracking results indicated that infants exhibited better category learning in the motion-defined than in the label-defined condition. These results provide little evidence for the idea that labels are category markers that facilitate category learning.

**Keywords:** Cognitive Development, Categorization, Attention, Label, Psychology, Human Experimentation.

## Introduction

The ability to form categories is an important component of human cognition (see Murphy, 2002, for a review). It has been well established that this ability appears early in development, with young infants capable of forming categories at an early age (Eimas & Quinn, 1994; Oakes, Madole, & Cohen, 1991), and a substantial body of experimental evidence suggest that linguistic labels affect this process (Balaban & Waxman, 1997; Waxman & Markow, 1995; Fulkerson, Waxman, & Seymour, 2006; Waxman & Booth 2003; Robinson & Sloutsky, 2007). However, the mechanisms underlying the role of labels remain unclear, which has generated considerable debate. Some have argued that words denoting categories have the special status of category markers and, as such, they facilitate category learning at the earliest stages of word learning. At the same time, others have argued that early in development words are akin to other features, but they may become category markers in the course of development.

According to the former theory, early in development, young infants have general assumptions that words but not other kinds of auditory inputs are category markers which

denote categories and “infants embark on the task of word learning equipped with a broad, universally shared expectation, linking words to commonalities among objects” (Waxman, 2003). According to this view, linguistic labels refer to category information and facilitate infants’ category learning. There is much evidence consistent with this view. First, some researchers have demonstrated that labels may facilitate infants’ categorization above and beyond other kinds of auditory input (Balaban & Waxman, 1997; Fulkerson & Haaf, 2003; Fulkerson et al., 2006; Ferry, Hespos, & Waxman, 2010). Second, facilitative effects of linguistic labels were demonstrated for basic-level as well as superordinate or global categories (Balaban & Waxman, 1997; Waxman & Markow, 1995; Waxman & Booth, 2003). Finally, effects of labels have been shown to affect infants’ performance on a variety of cognitive tasks, such as inductive inference (Graham, Kilbreath, Welder, 2004; Welder & Graham, 2001) and object individuation (Xu, 2002; Xu, Cote, & Baker, 2005).

There are challenges, however, to the idea that linguistic labels are category markers that facilitate categorization early in development. There are theoretical proposals arguing that, at least early in development, labels are features of items (similar to color or shape) rather than category markers (Sloutsky & Lo, 1999; Sloutsky & Fisher, 2004; Sloutsky, Lo, & Fisher, 2001) and the contribution of linguistic labels is driven by attentional rather than conceptual factors (Napolitano & Sloutsky, 2004; Sloutsky & Napolitano 2003). There is also evidence that auditory input overshadows (or attenuates processing of) corresponding visual input (Lewkowicz, 1988a, 1988b; Napolitano & Sloutsky, 2004; Robinson et al., 2005; Robinson & Sloutsky, 2004a, 2004b; Sloutsky & Napolitano, 2003). Furthermore, many of the studies examining the effects of labels on infants’ category learning compared the effects of labels with those of unfamiliar sounds, but not with a silent condition. When a salient baseline was introduced (e.g., Robinson & Sloutsky, 2007), labels did not facilitate infants’ category learning above the

silent baseline (see also Robinson & Sloutsky, 2008, for similar findings on individuation tasks).

There is also recent evidence by Deng and Sloutsky (2012) who demonstrated that labels function the same way as other features for young children, but they may become category markers in the course of development. In particular, Deng and Sloutsky used a variant of Yamauchi & Markman's (1998, 2000) paradigm and pitted category label against a highly salient visual feature. They found that unlike many adults who relied on category label, children relied on the salient feature.

If labels are not category markers for preschoolers, how they can be category markers attracting attention to within-category commonalities for infants (e.g., Balaban & Waxman, 1997; Ferry, Hespos, & Waxman, 2010; Waxman & Markow, 1995)? Although there is evidence that labels do little above and beyond a silent condition (Robinson & Sloutsky, 2007), none of the studies claiming that labels are category markers compared effects of labels with those of highly salient visual features. How do labels affect (a) patterns of attention and (b) the outcome of category learning? And how do these effects differ from those of highly salient visual features? The goal of this research is to answer these questions by using a combination of eye tracking and a more traditional novelty preference paradigm.

## Overview of Current Study

The goal of the study reported here was to examine the role of labels in category learning in infancy. The experiment consisted of two between-subjects conditions: label-defined condition and motion-defined condition. In both conditions, infants were familiarized with the exemplars from one category and then tested with the prototype of this category and that of the contrast category. Infants saw the same testing stimuli in both conditions and neither label nor motion was provided during testing. Eye gaze data were collected from infants while being trained and tested in both conditions. If labels are category markers and are able to direct attention to the category-relevant information, then infants should learn better when labels are provided. This should not be the case, however, if labels are part of input rather than category markers.

## Method

### Participants

Thirty-eight infants (16 boys and 22 girls) ranging in age from 8 to 12 months ( $M = 10$  months, 14 days;  $SD = 1$  month, 27 days) participated in this experiment. Data provided by 2 infants were excluded from analyses due to fussiness and 6 infants were excluded for not looking at a single test trial.

### Apparatus

A Tobii T60 eye-tracker with the sampling rate of 60 Hz (i.e., 60 gaze data points per second for each eye) was used

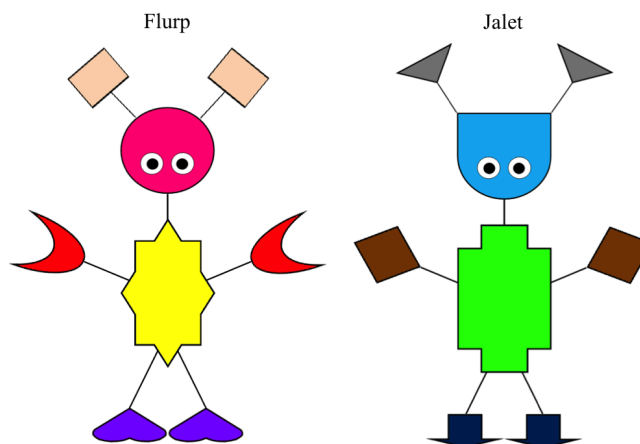


Figure 1. Prototypes of stimuli from Categories A and B.

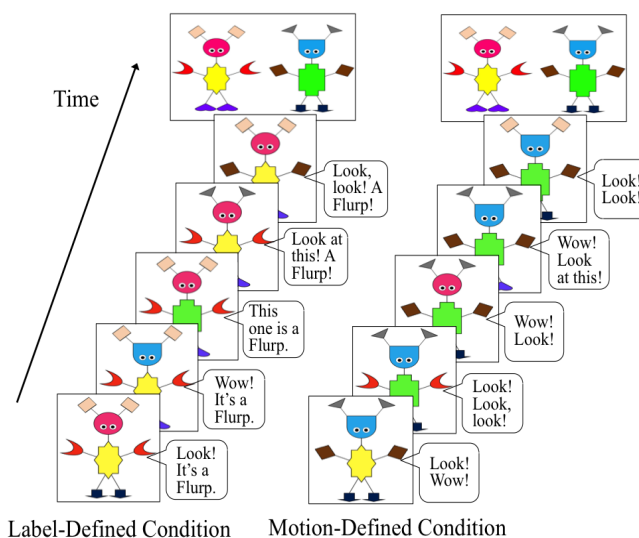


Figure 2. Example Stimuli.

to collect eye gaze data. The eye-tracker is integrated into a 17-inch computer monitor and located on a table inside a booth enclosed by black curtains. A trained experimenter monitored the experiment using Tobii Studio gaze analysis software installed on a 19-inch Dell OptiPlex 755 computer outside the booth. A video stream displaying participants' activities was projected onto a 9-inch black and white Sony SSM-930 CE television for experimenter's online monitoring. Two Dell speakers were located behind a black curtain on each side of the eye-tracker.

### Materials and Design

The materials were colorful drawings of artificial creatures and novel labels "flurp" and "jalet". The items had five features varying in color and shape and two categories were formed by different feature values (see Figure 1). As shown in Table 1, the two categories have a family-resemblance structure, which was derived from two prototypes (A0 and B0) by modifying the values of one of five features – head, antenna, hands, body, or feet. For example, to produce the



Table 1. Category structure used in learning.

Category A							Category B						
Stimuli	Head	Body	Hands	Feet	Antenna	Label/ Motion	Stimuli	Head	Body	Hands	Feet	Antenna	Label/ Motion
A1	1	1	1	1	0	1	B1	0	0	0	0	1	0
A2	1	1	1	0	1	1	B2	0	0	0	1	0	0
A3	1	1	0	1	1	1	B3	0	0	1	0	0	0
A4	1	0	1	1	1	1	B4	0	1	0	0	0	0
A5	0	1	1	1	1	1	B5	1	0	0	0	0	0
A0	1	1	1	1	1	1	B0	0	0	0	0	0	0

Note. The value 1 = any of five dimensions identical to "Flurp" (see Figure 1). The value 0 = any of five dimensions identical to "Jalet" (see Figure 1). A0 and B0 are prototypes of each category.

stimulus A1, the value of the antenna was changed from 1 to 0 so that it had four features consistent with the prototype A0 and one feature consistent with the prototype B0. See Figure 2 for example of stimuli.

There were two between-subjects conditions (label-defined vs. motion-defined). In the label-defined condition, the value of label did not vary across the exemplars; whereas in the motion-defined condition, a pattern of motion did not vary across the exemplars. In particular, one of the features (the feet) was animated to be highly salient using Macromedia Flash MX software and all the members of a given category had feet with the same pattern of motion. For "flurps", the feet were purple, heart-shaped, and stretched up and down; whereas, for "jalets", the feet were dark blue, arrows-shape, and moved sideways.

## Procedure

Infants were seated on parents' laps approximately 60 cm away from the eye-tracker. Parents were instructed not to interact with infants and not to point or label any of the stimuli. Prior to the experiment, infants completed a 5-point calibration sequence. The calibration points consisted of dynamic kitten images appearing in different locations on the screen, with "bounce" sound. After successful calibration, a colorful picture of a baby playing with several different toys was presented on the screen to keep infants' attention.

When infants and parents were ready to begin, an experimenter started the experiment by pressing the space bar. The picture of baby disappeared and infants were presented 20 familiarization trials and 4 test trials. The trials were mixed and pseudo-randomly assigned into 4 blocks, with 5 familiarization trials followed by 1 test trial in each block. On each familiarization trial, infants saw a creature produced from the structure of one category shown in Table 1 with a white background lasting for 8000 msec and heard a phrase starting at the onset of each trial. In label-defined condition, a labeling phrase (e.g., "Look! This is a Flurp") was presented in each trial and lasted for approximately 2800 msec. However, in motion-defined condition, the phrase was not labeled (e.g., "Look at this one!"). The feet of the creature moved after the phrase and the motion lasted for 3000 msec. The onset of motion in motion-defined condition was approximately the same as that of label in

label-defined condition. After familiarization, infants were tested with preference trial (in each block). Each test trial consisted of the prototype of the category that infants were trained with in familiarization and the prototype of the contrast category, and presented without either label or motion for 8000 msec. A dynamic bouncing ball was presented as an attention-engager between trials within each block. A short cartoon video was presented between blocks in order to let infants have a rest. All gaze data were recorded by the computer using Tobii Studio gaze analysis software.

## Results

Analyses presented below primarily focused on the percentage looking to the prototype of novel category at test and on the patterns of attention during familiarization phase and testing phase. Gaze data were exported from the computer using Tobii Studio gaze analysis software. Scenes were created for every trial (both familiarization and test) and eighteen areas of interest (AOIs) for fixations were defined: one rectangle surrounding the creatures on familiarization trials, two rectangles surrounding two prototypes respectively on test trials, and fifteen ellipses surrounding each feature of the creatures on both familiarization and test trials. These data were used to calculate (1) novelty preference score based on the proportion of looking time to the prototype of novel category as compared to the total looking time to both of the prototypes at test; and (2) patterns of attention based on the proportion of looking time to different features on both familiarization and test trials.

## Novelty Preference Scores

To examine how labels or patterns of motion affected infants' categorization, a novelty preference score was calculated for each test trial: accumulated looking time to the prototype from novel category divided by the overall looking time for both stimuli. The main results are presented in Figure 3. The data were submitted to a 4 (block: 1 vs. 2 vs. 3 vs. 4) by 2 (condition: label-defined vs. motion-defined) mixed ANOVA, with block as a within-subjects factor and condition as a between-subjects factor. There was a significant main effect of condition,  $F(1, 28) = 11.51$   $MSE = 0.24$ ,  $p < .01$ ,  $\eta_p^2 = 0.291$ , with infants looking substantially

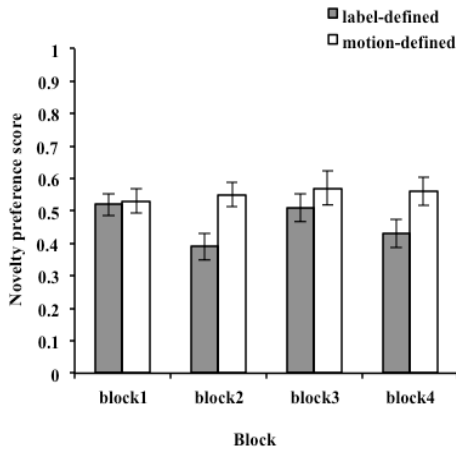


Figure 3. Infants' novelty preference score at test in 4 blocks for label-defined and motion-defined conditions.

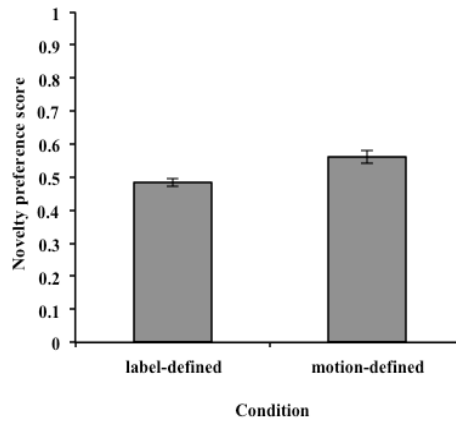


Figure 4. Infants' novelty preference score averaged across four blocks at test.

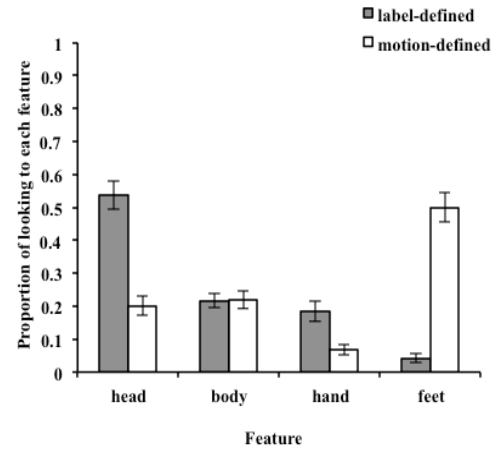


Figure 5. Infants' proportion of accumulated looking time to each feature averaged across 4 blocks at familiarization.

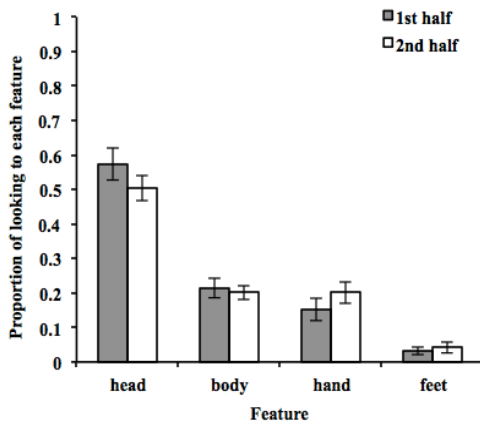


Figure 6. Infants' proportion of looking time to each feature in 1<sup>st</sup> and 2<sup>nd</sup> half familiarization trial in label-defined condition.

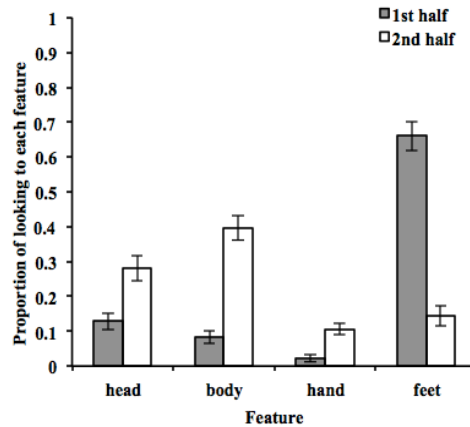


Figure 7. Infants' proportion of looking time to each feature in 1<sup>st</sup> and 2<sup>nd</sup> half familiarization trial in motion-defined condition.

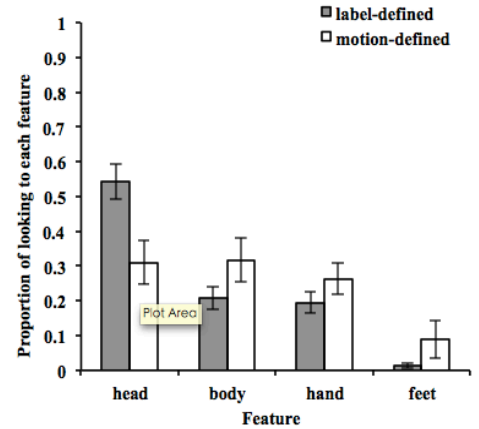


Figure 8. Infants' proportion of accumulated looking time to each feature averaged across 4 blocks at test.

longer to the novel category in motion-defined condition compared to the label-defined condition. However, neither the main effect of block ( $p = .45$ ) nor the interaction between block and condition ( $p = .34$ ) was found. Therefore, data were collapsed across blocks for each condition and the results are presented in Figure 4. As shown in Figure 4, infants looked significantly longer to the novel category in motion-defined condition than in label-defined condition, independent sample  $t(28) = 3.71$ ,  $p < .01$ ,  $d = 1.44$ . In addition, infants' novelty preference score was significantly higher than chance in motion-defined condition

one-sample  $t(11) = 3.27$ ,  $p < .01$ ,  $d = 0.94$ ; whereas in label-defined condition, infants' looking to the novel category was not different from chance,  $p = .19$ .

### Patterns of Attention

To determine if moving feet served as an engager and pushed infants' attention to other features, we compared the accumulated looking time to the stimuli on familiarization trials in two conditions. The gaze data (accumulated looking time) were submitted to a 4 (block: 1 vs. 2 vs. 3 vs. 4) by 2 (condition: label-defined vs. motion-defined) mixed



ANOVA, with block as a within-subjects factor and condition as a between-subjects factor. Results revealed a main effect of block,  $F(3, 84) = 12.87$ ,  $MSE = 0.00$ ,  $p < .01$ ,  $\eta_p^2 = 0.315$ , with infants' accumulated looking time decreasing through blocks. However, infants' accumulated looking time did not differ between label-defined and motion-defined conditions,  $p = .44$ .

Because the comparable accumulated looking during familiarization trials resulted in different outcomes of category learning in two conditions, we deemed it necessary to examine how attention was distributed among different features of the stimuli at both familiarization and test. The proportion of looking time to different features on both familiarization (averaged across five trials within each block) and test trials (averaged across old and new prototypes) were calculated and the data were submitted to two separate 4 (feature: head<sup>1</sup> vs. body vs. feet) by 4 (block: 1 vs. 2 vs. 3 vs. 4) by 2 (condition: label-defined vs. motion-defined) mixed ANOVAs, with feature and block as within-subjects factors and condition as a between-subjects factor. For familiarization trials, because there was no main effect of block ( $p = .16$ ), these data were collapsed across four blocks. As shown in Figure 5, there was an interaction between feature and condition,  $F(3, 84) = 44.74$ ,  $MSE = 0.80$ ,  $p < .01$ ,  $\eta_p^2 = 0.615$ . Infants' accumulated more looking to the head in label-defined condition compared to motion-defined condition, independent sample  $t(27.7) = 6.46$ ,  $p < .01$ ,  $d = 2.25$ ; whereas they looked significantly longer at the feet, independent sample  $t(13.4) = 9.9$ ,  $p < .01$ ,  $d = 4.45$ , and the hands, independent sample  $t(25.2) = 3.40$ ,  $p < .01$ ,  $d = 1.14$ , in motion-defined condition than in label-defined condition. Infants accumulated comparable looking to body in both conditions,  $p = .91$ . Since the label and the moving feet were not presented for the entire 8-second trial in training, we further examined infants' pattern of attention by comparing their accumulated looking to each feature in the first half of the trial to that in the second half of the trial. Data were submitted to two 4 (feature: head vs. body vs. hands vs. feet) by 2 (time course: 1<sup>st</sup> half vs. 2<sup>nd</sup> half) repeated measures ANOVAs for the label-defined and motion-defined conditions respectively. As shown in Figure 6, in the label-defined condition, there was a main effect of feature,  $F(3, 51) = 43.25$ ,  $MSE = 1.63$ ,  $p < .01$ ,  $\eta_p^2 = 0.718$ , and an interaction between feature and time course,  $F(3, 51) = 5.74$ ,  $MSE = 0.02$ ,  $p < .01$ ,  $\eta_p^2 = 0.253$ . Infants looked significantly longer at the head in both of the 1<sup>st</sup> and 2<sup>nd</sup> half of the familiarization trial, Bonferroni  $ps < .01$ , but shorter at the feet,  $ps < .05$ . In the motion-defined condition, as shown in Figure 7, there was a main effect of feature,  $F(3, 33) = 42.63$ ,  $MSE = 0.46$ ,  $p < .01$ ,  $\eta_p^2 = 0.795$ , and an interaction between feature and time course,  $F(3, 33) = 57.67$ ,  $MSE = 0.79$ ,  $p < .01$ ,  $\eta_p^2 = 0.840$ . In contrast to infants who biased to the head in the label-defined condition, infants' looking was more widely distributed in the motion-defined condition: they looked longer at the moving feet during the first 4-second familiarization trial,

Bonferroni  $ps < .01$ , but they looked longer at the body and head during the latter half of familiarization trial (looking time: body > head > feet > hand; only for the last pair, Bonferroni  $p < .01$ ).

Similar analyses were conducted to examine the patterns of attention on test trials (recall that neither labels nor motion was presented during these trials). Since there was no main effect of block ( $p = .42$ ), the data were collapsed across blocks and results are presented in Figure 8. Similar to familiarization, there was an interaction between feature and condition,  $F(3, 84) = 4.94$ ,  $MSE = 0.18$ ,  $p < .01$ ,  $\eta_p^2 = 0.150$ . Infants looked longer at the head in label-defined condition compared to motion-defined condition, independent sample  $t(28) = 2.87$ ,  $p < .01$ ,  $d = 1.11$ . However, there was little evidence that infants in motion-defined condition just focused on feet to categorize; their looking instead was more widely distributed across the features,  $ps > .24$ .

## Discussion

In the study reported here, we investigated the role of labels in infants' categorization by comparing effects of labels with those of highly salient visual features. By using a combination of eye tracking and a more traditional novelty preference paradigm, we examined the patterns of attention and the outcome of category learning.

The current study reveals several important findings. First, infants exhibited better category learning when they saw a salient visual feature (i.e., moving feet) compared to when they heard a label. Second, although infants accumulated comparable looking during learning in motion-defined condition compared label-defined condition, their attention was more distributed among different features when there was a salient visual feature whereas they spent most of their time looking at the head when they heard a label. Third, there was little evidence of labels facilitating category learning in infants.

Many studies have examined the role of labels in infants' categorization and there is little agreement on the role of labels being category markers or features. However, none of the studies examining the effects of labels on categorization in infancy demonstrated that these effects are greater than those of highly salient features. By comparing the outcome of category learning and examining the patterns of attention in label-defined and motion-defined conditions, the current study has provided new evidence as to how labels may affect category learning in infancy. The results indicate that a pattern of motion has a greater facilitative effect on category learning than the label.

The current study raises an interesting question about the role of labels in infants' categorization by comparing the effects of label to those of a salient feature and the results suggest the labels may start out as features rather than category markers. Future research will examine whether labels merely fail to facilitate category learning or they actually hinder infants' category learning in infancy.

<sup>1</sup> Antenna are combined with head as one level of feature.

## Acknowledgments

This research is supported by the NSF grant BCS-0720135 and by NIH grant R01HD056105 to VMS.

## References

- Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology*, 64, 3-26.
- Deng, W., & Sloutsky, V. M. (2012). Carrot-eaters and moving heads: Salient features provide greater support for inductive inference than category labels. *Psychological Science*, 23, 178-186.
- Eimas, P. D., & Quinn, C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. *Child Development*, 65, 903-917.
- Ferry, A., Hespos, S.J., & Waxman, S. (2010). Language facilitates category formation in 3-month-old infants. *Child Development*, 81, 472-479.
- Fulkerson, A. L., & Haaf, R. A. (2003). The influence of labels, non-labeling sounds, and source of auditory input on 9- and 15-month-olds' object categorization. *Infancy*, 4, 349-369.
- Fulkerson, A. L., Waxman, S. R., & Seymour, J. M. (2006). Linking object names and object categories: Words (but not tones) facilitate object categorization in 6- and 12-month-olds. In Bamman, D., Magnitskaia, T., & Zaller, C (Eds.) *Supplement to the Proceedings of the 30th Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.
- Graham, S. A., Kilbreath, C. S., & Welder, A. N. (2004). Thirteen-month-olds rely on shared labels and shape similarity for inductive inferences. *Child Development*, 75, 409-427.
- Lewkowicz, D. J. (1988a). Sensory dominance in infants: 1. Six-month-old infants' response to auditory-visual compounds. *Developmental Psychology*, 24, 155-171.
- Lewkowicz, D. J. (1988b). Sensory dominance in infants: 2. Ten-month-old infants' response to auditory-visual compounds. *Developmental Psychology*, 24, 172-182.
- Murphy, G. L. (2002). *The Big Book of Concepts*, MIT Press.
- Napolitano, A. C., & Sloutsky, V. M. (2004). Is a picture worth a thousand words? The flexible nature of modality dominance in young children. *Child Development*, 75, 1850-1870.
- Oakes, L. M., Madole, K. L., & Cohen, L. B. (1991). Object examining: Habituation and categorization. *Cognitive Development*, 6, 377-392.
- Robinson, C. W., & Sloutsky, V. M. (2004a). Auditory dominance and its change in the course of development. *Child Development*, 75, 1387-1401.
- Robinson, C.W., & Sloutsky, V. M. (2004b). The effect of stimulus familiarity on modality dominance. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the XXVI annual conference of the Cognitive Science Society* (pp. 1167-1172). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Robinson, C. W., Howard, E. M., & Sloutsky, V. M. (2005). Mechanisms underlying the effects of labels on cognitive development. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the XXVII annual conference of the Cognitive Science Society* (pp. 1878-1882). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Robinson, C. W., & Sloutsky, V. M. (2007). Linguistic labels and categorization in infancy: Do labels facilitate or hinder? *Infancy*, 11, 233-253.
- Robinson, C. W., & Sloutsky, V. M. (2008). Effects of auditory input in individuation tasks. *Developmental Science*, 11, 869-881.
- Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, 133, 166-188.
- Sloutsky, V. M., & Lo, Y.-F. (1999). How much does a shared name make things similar? Part 1: Linguistic labels and the development of similarity judgment. *Developmental Psychology*, 35, 1478-1492.
- Sloutsky, V.M., Lo, Y.-F., & Fisher, A.V. (2001). How much does a shared name make things similar? Linguistic Labels and the development of inductive inference. *Child Development*, 72, 1695-1709.
- Sloutsky, V. M., & Napolitano, A. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child development*, 74, 822-833.
- Waxman, S.R. (2003). Links between object categorization and naming: origins and emergence in human infants. In D.H. Rakison & L.M. Oakes (Eds.), *Early category and concept development: Making sense of the blooming, buzzing confusion* (pp. 213-241). London: Oxford University Press.
- Waxman, S. R., & Booth, A. E. (2003). The origins and evolution of links between word learning and conceptual organization: New evidence from 11-month-olds. *Developmental Science*, 6, 130-137.
- Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- 13-month-old infants. *Cognitive Psychology*, 29, 257-302.
- Welder, A. N., & Graham, S. A. (2001). The influences of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. *Child Development*, 72, 1653-1673.
- Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, 85, 223-250.
- Xu, F., Cote, M., & Baker, A. (2005). Labeling guides object individuation in 12-month-old infants. *Psychological Science*, 16, 372-377.
- Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, 39, 124-148.
- Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 776-795.

# Learning Deterministic Causal Networks from Observational Data

**Ben Deverett**

ben.deverett@mail.mcgill.ca  
McGill University

**Charles Kemp**

ckemp@cmu.edu  
Carnegie Mellon University

## Abstract

Previous work suggests that humans find it difficult to learn the structure of causal systems given observational data alone. We show that structure learning is successful when the causal systems in question are consistent with people’s expectations that causal relationships are deterministic and that each pattern of observations has a single underlying cause. Our data are well explained by a Bayesian model that incorporates a preference for symmetric structures and a preference for structures that make the observed data not only possible but likely.

**Keywords:** structure learning, causal learning, Bayesian modeling

Causal networks have been widely used as models of the mental representations that support causal reasoning. For example, an engineer’s knowledge of the local electricity system may take the form of a network where the nodes represent power stations and the links in the network represent connections between stations. Causal networks of this kind may be learned in several ways. For example, an intervention at station A that also affects station B provides evidence for a directed link between A and B. Networks can also be learned via instruction: for example, a senior colleague might tell the engineer that A sends power to B. Here, however, we focus on whether and how causal networks can be learned from observational data. For example, the engineer might infer that A sends power to B after observing that A and B are both inactive during some blackouts, that B alone is inactive during others, but that A is never the only inactive station.

A consensus has emerged that causal structure learning is difficult or impossible given observational data alone. For example, Fernbach and Sloman (2009) cite the results of Steyvers, Tenenbaum, Wagenmakers, and Blum (2003), Lagnado and Sloman (2004), and White (2006) to support their claim that “observation of covariation is insufficient for most participants to recover causal structure” (p 680). Here we join Mayrhofer and Waldmann (2011) in challenging this consensus. We show that people succeed in a structure learning task when the causal systems under consideration are aligned with intuitive expectations about causality. Previous studies suggest that people expect causal relationships to be deterministic (Schulz & Somerville, 2006; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008), and expect that any pattern of observations tends to be a consequence of a single underlying cause (Lombrozo, 2007). We ask people to reason about systems that are consistent with both expectations, and find that structure learning is reliably achieved under these conditions.

A previous study by White (2006) asked participants to learn the structure of deterministic causal systems from observational data alone. The structures involved were five-node

networks where the nodes represented population levels of five different species. White’s task proved to be difficult, and performance was poor even when White gave his participants explicit instructions about how to infer causal structure from observational data. Here, however, we demonstrate that both structures considered by White can be reliably learned in the context of the experimental paradigm that we develop.

Given that humans perform well on the structure learning tasks that we consider, it is natural to ask how this performance is achieved. Mayrhofer and Waldmann (2011) propose that learners rely on a “broken link” heuristic and identify the structure that minimizes the number of cases where a cause is present but an effect is absent. They contrast their heuristic-based approach with Bayesian accounts of structure learning that rely on patterns of conditional independence between variables. We propose a Bayesian account that falls in between these two alternatives. Like Mayrhofer and Waldmann, we believe that models which track patterns of conditional independence are often too powerful to capture the inferences made by resource-bounded human learners. Unlike Mayrhofer and Waldmann, we argue that a Bayesian approach is nevertheless useful for explaining why humans succeed in the tasks that we consider. In particular, we show that human inferences are influenced by two factors that are naturally captured by the prior and the likelihood of a Bayesian model—a preference for symmetric structures, and a preference for structures that explain the observed data without needing to invoke coincidences. We demonstrate that incorporating these factors allows a Bayesian model to account for our data better than an approach that relies on the broken-link heuristic alone.

## Bayesian Structure Learning

The causal systems that we consider are simple activation networks. Each network can be represented as a graph  $G$  which may include cycles. Figure 1a shows one such graph and a data set  $D$  generated over the graph. Each row  $d_i$  in the data set  $D$  represents an observed pattern of activation—for example, the first row represents a case where nodes A, C and D are observed to be active and node B is observed to be inactive. We will assume that each row  $d_i$  is generated by activating a randomly chosen node then allowing activation to propagate through the network. For example, Figure 1b shows that if A is the randomly activated node, the final pattern of activation will match the first row of matrix  $D$  in Figure 1a.

The activation networks that we consider have three important properties. First, all causal links are generative, and these generative links combine according to an OR function. For example, node C in Figure 1a will be active if node A is

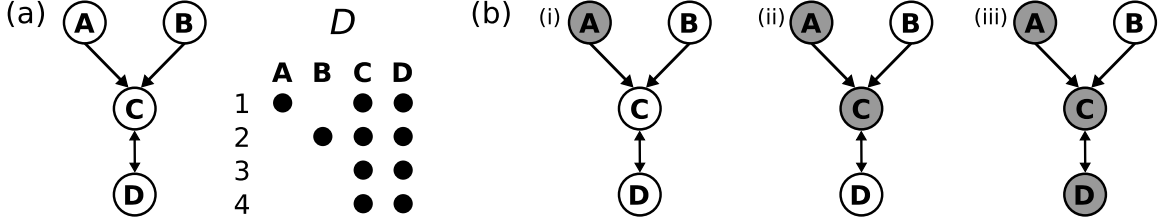


Figure 1: (a) A simple activation network and a data set  $D$  generated over the network. Each row of matrix  $D$  is an observation which indicates that some of the nodes in the network are active. (b) The first observation in (a) is generated when node A spontaneously activates and activation propagates through the network.

active or if node B is active. Second, all causal links are deterministic. Third, spontaneous activations are rare: at most one node in the network can spontaneously activate at any time, which means that each observed pattern of activation can be traced back to a single root cause. For example, the spontaneous activation of node A is the root cause of the activation pattern in the first row of matrix  $D$  in Figure 1a. Our assumptions that causes are rare and have deterministic effects are conceptually related to the work of Lu et al. (2008) on “sparse and strong” priors for causal learning. Note, however, that our notion of rarity is different from their notion of sparsity. Their notion captures the expectation that each node in a causal graph is expected to have at most one strong cause, but ours captures the idea that each pattern of observations  $d_i$  is expected to have a single underlying cause. For example, the activation network in Figure 1a is inconsistent with their notion of sparsity, since A and B are both strong causes of C. This network, however, is consistent with our notion of rarity as long as the base rates of A and B are both very low, which means that at most one of these nodes will spontaneously activate at any time.

Because the networks we consider may include cycles, they are different from standard Bayesian networks. If desired, however, our activation networks can be represented as dynamic Bayesian networks where the cycles are unrolled in time (Rehder & Martin, 2011). For our purposes, however, it will be simplest to work with graphs that may include cycles.

Given a data set  $D$  generated from an unknown network  $G$ , a probability distribution over the possible networks can be computed using Bayes’ rule:

$$P(G|D) \propto P(D|G)P(G) = \left[ \prod_i P(d_i|G) \right] P(G), \quad (1)$$

where we have assumed that the rows  $d_i$  in the matrix  $D$  are independently generated over the graph. We will consider two versions of the prior  $P(G)$  and two versions of the likelihood term  $P(d_i|G)$ .

The first version of the likelihood term assumes that observation  $d_i$  resulted from the spontaneous activation of a single node in the graph. We sum over all nodes  $n$  that may have activated spontaneously:

$$P(d_i|G) = \sum_n P(d_i|G, n)P(n). \quad (2)$$

$P(d_i|G, n)$  is either 1 or 0 depending on whether  $d_i$  is the ob-

servation pattern produced by activating node  $n$  then allowing activation to propagate through the graph. The prior distribution  $P(n)$  is uniform, which captures the assumption that all nodes are equally likely to activate spontaneously. We refer to Equation 2 as the *probabilistic* likelihood.

The second version of the likelihood term depends only on whether observation  $d_i$  is consistent with  $G$ , and will be called the *logical* likelihood:

$$P(d_i|G) = \begin{cases} 1 & \text{if } d_i \text{ is consistent with } G \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Observation  $d_i$  is consistent with  $G$  if  $d_i$  can be produced by activating some node in  $G$  and allowing activation to propagate through the graph.

The first version of the prior  $P(G)$  in Equation 1 corresponds to a uniform distribution over the full space of graphs. The second version captures a preference for graphs that are symmetric. Perceptual research has documented a preference for symmetric stimuli, and this preference can be viewed as an instance of a more general preference for stimuli that display “good form.” We hypothesized that a graph shows “good form” if many of its nodes play similar roles. For example, nodes A and B in Figure 1a play similar roles, and exchanging the labels on these nodes leaves the structure of the graph unchanged. The symmetry score of a graph can be formally defined as the number of graph automorphisms, or the number of node permutations that leave the structure of the graph unchanged. For a given number of nodes, the graph with no edges and the fully connected graph will share the highest possible symmetry score, because all possible node permutations leave the structure of these graphs unchanged. We used these symmetry scores to define a prior  $P(G) \propto s(G)$ , where  $s(G)$  is the symmetry score of graph  $G$ .

Combining the two likelihoods and the two priors produces a total of four different models. The “logical uniform” (LU) model produces a posterior distribution  $P(G|D)$  that assigns equal probability to all graphs  $G$  that are consistent with the data. The LU model is consistent with the broken link heuristic described by Mayrhofer and Waldmann (2011), which assesses how well graph  $G$  accounts for data  $D$  by counting the number of times that a parent node is active and a child node is inactive. In our setting, a graph is deemed consistent with data  $D$  if and only if the graph has a broken link count of zero. When applied to our experimental stimuli, the LU

model therefore makes identical predictions to a model which assumes that people choose a graph that minimizes the broken link count, and that people are indifferent among graphs that satisfy this criterion.

Like the LU model, the “probabilistic uniform” (PU) model assigns nonzero probability only to graphs that are consistent with the data. The PU model, however, allows for cases where a data set  $D$  is consistent with two graphs but better explained by one graph than the other. Consider a three-node problem where  $D$  includes 6 observations and where each observation indicates that nodes A, B and C are all active. The data are consistent with a causal chain where A sends an arrow to B and B sends an arrow to C. The data, however, are not typical of a chain, since the chain hypothesis requires the assumption that A spontaneously activated 6 times in succession, which seems like a big coincidence. The data are also consistent with a fully connected graph, and now no coincidence must be invoked, since all nodes end up active regardless of which node activates first. As this example suggests, comparing the logical models with the probabilistic models will allow us to evaluate whether people’s inferences depend on probabilistic factors like “degree of coincidence” that go beyond consistency with the data.

The “logical symmetry” (LS) and “probabilistic symmetry” (PS) models are directly comparable to the LU and PU models, except that they incorporate a preference for symmetric graphs. Comparing the symmetry models and the uniform models will allow us to evaluate whether people bring *a priori* expectations about the underlying structure to the task of structure learning.

## Structure learning experiment

We designed an experiment to explore whether humans are capable of learning the structure of an activation network given observational data alone, and to evaluate the four models just presented.

**Participants.** 36 members of the CMU community participated in exchange for pay or course credit.

**Design.** The experiment included 34 blocks, each of which included one or more observations generated over an unobserved network. 32 of the blocks involved networks with three nodes, and the final two blocks involved networks with five nodes. The *characteristic data set* for a network is a set of observations that result from spontaneous activations of each node in the network. Given any network with three nodes, there are 64 possible graphs, but the characteristic data sets for these graphs include only 9 qualitatively different types. Representatives of each type are shown in Figures 2a through 2i. Among the blocks of three-node networks, these nine types were each presented twice, making 18 blocks with three observations each. An additional nine blocks with six observations each were created by including two copies of a characteristic data set per block. Five additional blocks each had two or fewer observations, and are shown in Figures 2j through 2n. These 32 blocks were presented in random or-

der, followed by two final blocks for the five-node networks (Figure 5). These five-node networks are identical to causal structures previously studied by White (2006). The observations within all blocks were shown in random order.

**Materials and Procedure.** The nodes in each network appeared as rectangles on screen, and active and inactive nodes had different colors. Participants were told that these rectangles were detectors that “detect a rare type of particle called the mu particle.” Participants were told that the detectors were connected by directed satellite links, and that an “active detector always activates all detectors that it points to.” To reinforce this information, participants were given an example like Figure 1 where they observed a single detector activating and activation subsequently propagating over the network.

Participants then worked through the 34 blocks. Within each block the observations were presented one at a time. After seeing all observations for a given block, participants drew a graph on screen to indicate their best guess about the structure of the underlying network and rated their confidence in their guess on a seven point scale. To minimize memory demands, all observations within a block were retained on screen after being presented, which means that all observations were visible when participants reached the graph-drawing stage. Each previous observation appeared as a panel with detectors, and every edge that participants added during the graph drawing stage was simultaneously added to each of these panels. This design choice was intended to make it as easy as possible for participants to see whether the graph that they had drawn was consistent with all observations for that block.

**Results.** We focus first on results for the three-node networks. The first nine panels in Figure 2 show the most popular graphs for the nine characteristic data sets, and the remaining panels show results for the blocks with two or fewer observations. In each case the most common response is consistent with the data set, indicating that participants understood the task and were successfully able to discover causal structure given observational data alone. In particular, note that all 36 participants discovered the common effect structure in Figure 2d and the common cause structure in Figure 2f. Steyvers et al. (2003) found that these structures are difficult for learners to distinguish in a probabilistic setting, but our data suggest that they are easy to learn in our deterministic setting.

Figure 2 also includes predictions of the PS model, and correlations between the model and the data are shown. Results for all four models across the first 32 blocks of the experiment are shown in Figure 3. The first correlation in each panel shows the performance of a model across the entire set of blocks, and the correlation in parentheses shows the average single-block correlation. The PS model performs best overall, suggesting that the probabilistic likelihood and the symmetry prior are both required in order to capture human judgments. A bootstrap analysis indicates that the overall and average single-block correlations achieved by the PS model are reliably higher than the correlations achieved by the PU

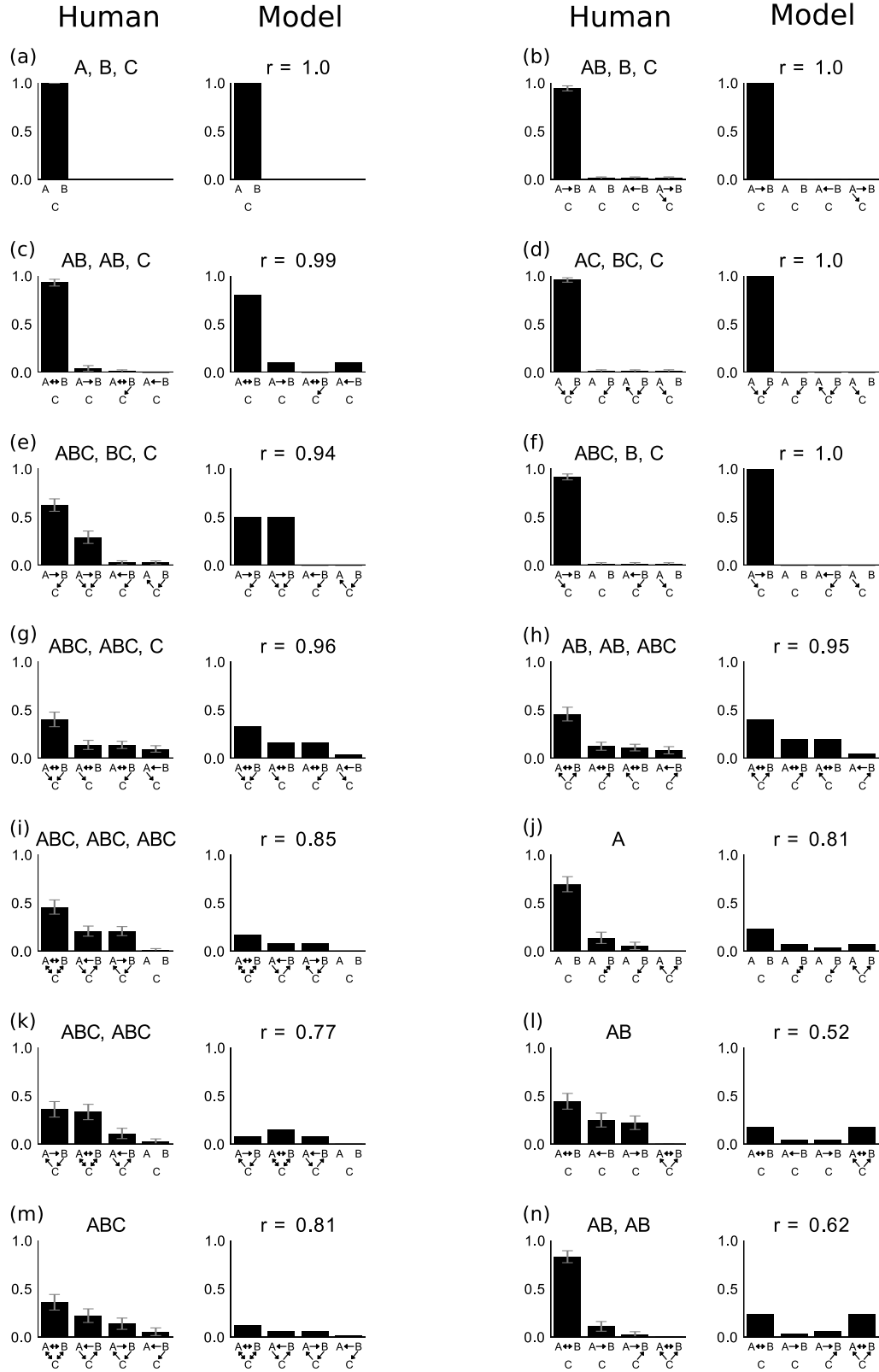


Figure 2: Participant responses and predictions of the PS model for 14 patterns of observations. The observed data are shown above the left plot in each panel, and the correlation between model predictions and human responses is shown above the right plot. The four structures in each plot always include the top two structures chosen by humans and the two most probable structures according to the model.

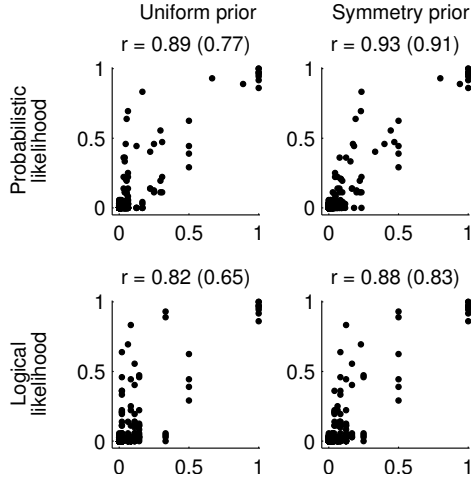


Figure 3: Comparison of the complete set of responses for the first 32 experiment blocks with the predictions of four models. The first correlation in each panel shows correlations based on the complete set of responses, and the correlation in parentheses shows the average correlation across the 32 individual experiment blocks.

and LS models ( $p < 0.01$  in all cases).

The main shortcoming of the logical likelihood is that it leads to predictions that are too diffuse. The structure preferred by participants is typically one of the most probable structures according to the LU model, but the model often assigns the same probability to many other structures. For example, after observing “ABC” three times in succession, the LU model assigns the same probability to all 51 structures that can generate the observation “ABC,” including causal chains over these variables. In contrast, the PU model assigns highest probability to the 18 structures that can *only* generate the observation “ABC.”

Although the PU model performs better than the LU model, its predictions are still more diffuse than the human responses. As just mentioned, the PU model predicts that 18 different structures are equally likely after observing “ABC” three times, but participants overwhelmingly prefer the top three structures shown in Figure 2i. The symmetry prior allows the PS model to capture this preference: note that the fully connected graph is the most symmetric structure that can only generate “ABC,” and the two cycles are the next most symmetric structures that meet this criterion.

To further evaluate the difference between the probabilistic and the logical likelihood, we examined the learning curves that result when the same observation is presented multiple times. The 34 blocks in the experiment include blocks where observation “ABC” is presented once, twice, three times, and six times. Figure 4b shows model predictions for these four blocks, where each bar represents the probability mass assigned to structures that can only generate “ABC.” The learning curves for the LU and LS models are flat—these models

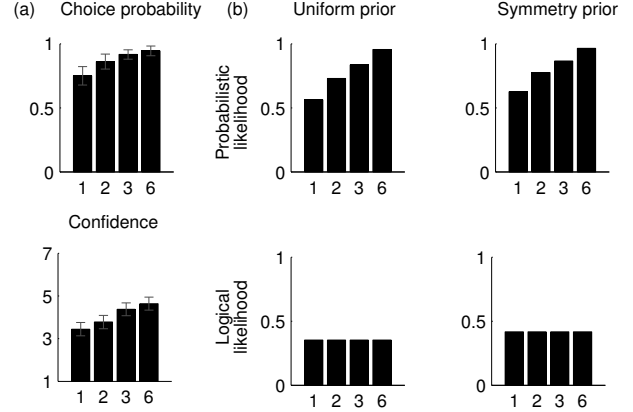


Figure 4: Inferences after observation ABC is presented one, two, three or six times. (a) Proportion of structures chosen by humans that can only generate observation ABC (top); Average confidence ratings (bottom). (b) Probability assigned to structures that can only generate ABC by four models.

are sensitive to whether or not a structure is consistent with an observation, but the number of times that the observation appears is irrelevant. In contrast, the PU and PS models become increasingly confident that the underlying structure can only generate the observation “ABC.” Figure 4a indicates that participants show a similar learning curve, and become increasingly confident in their responses as the number of observations increases. Bootstrap analyses indicate that the differences between the first and the final bars are statistically significant for both plots in Figure 4a ( $p < 0.001$ ).

Figure 5 shows the most popular graphs chosen for the two five-node blocks. Each set of observations is consistent with only one structure, and participants were reliably able to discover these structures. Figure 6 compares our results to those reported by White, who found that relatively few participants were able to discover these five-node structures. There are at least two reasons why these tasks may have produced different results. First, our particle-detector scenario may be more intuitive than White’s task which required inferences about changes in the populations of species over time. Second, we asked participants to reason about the five-node structures following 32 inferences about three-node structures, which means that practice and familiarity with the task may have contributed to their performance. Future studies are needed to isolate the critical differences between these paradigms, but for now we can conclude that there are conditions under which people reliably discover White’s five-node structures from observational data alone.

Taken overall our results support two general conclusions. First, humans succeed at structure learning when causes are strong and when each observation has a single root cause. Because our cover story made these conditions quite clear, our data suggest that people reason accurately about deterministic systems where causes are rare, but not that people sponta-

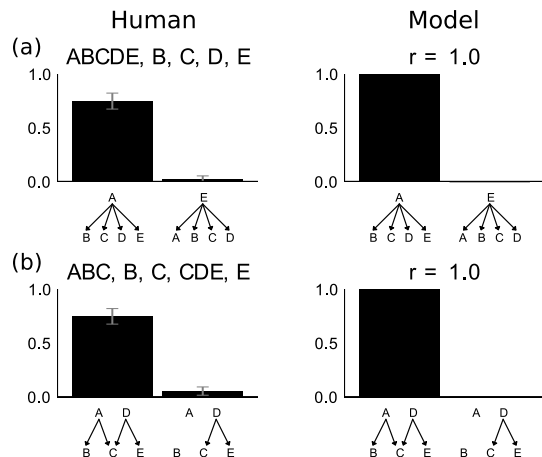


Figure 5: Data sets, human responses and model predictions for the final two experiment blocks. All four models make the same prediction, because in both cases only one structure is consistent with the observations.

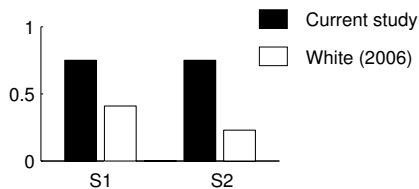


Figure 6: Comparison between our results and results reported by White (2006). The bars show the proportion of participants who successfully learned the five-node structures S1 and S2.

neously bring these assumptions to causal learning problems. Previous studies, however, suggest that both the determinism assumption and the rarity assumption may both apply more generally (Lu et al., 2008; Lombrozo, 2007)

The second general conclusion is that structure learning in our task cannot be adequately characterized as a search for a structure that is consistent with the observed data. At least two additional factors play a role: humans are sensitive to whether or not the observed data are typical of a given structure, and humans have *a priori* preferences for certain kinds of structures including symmetric structures. The PS model illustrates that these factors can be captured by the likelihood and prior of a Bayesian model, and demonstrates the value of the Bayesian approach to structure learning.

## Conclusion

Previous studies have found that structure learning from observational data is difficult. In contrast, our data suggest that humans find structure learning relatively easy in settings where causes act deterministically and where each observation has a single root cause. Future studies can consider relaxations of these conditions and explore whether humans still succeed at structure learning when causes are strong but not fully deterministic, and when most but not all observations

have a single root cause.

Our data are consistent with the recent work of Mayrhofer and Waldmann (2011), who also report positive results for learning from observational data. Mayrhofer and Waldmann (2011) argue that humans succeed at structure learning by relying on simple heuristics, but we found that their “broken link” heuristic accounted less well for our data than a Bayesian model that incorporates a probabilistic likelihood term and a symmetry-based prior. There may be alternative heuristics that can implement the computations required by our Bayesian model, but we believe that any successful account of our data will need to incorporate an *a priori* preference for symmetric structures, and a preference for structures that make the observed data not only possible but likely.

**Acknowledgments.** We thank Alan Jern for assistance with the experiment. This work was made possible by a training program in Neural Computation that was organized by the Center for the Neural Basis of Cognition and supported by NIH R90 DA023426.

## References

- Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 35(3), 678–93.
- Lagnado, D., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 856–876.
- Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, 55, 232–257.
- Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, 115(4), 955–984.
- Mayrhofer, R., & Waldmann, M. R. (2011). Heuristics in covariation-based induction of causal models: sufficiency and necessity priors. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3110–3115). Austin, TX: Cognitive Science Society.
- Rehder, B., & Martin, J. B. (2011). A generative model of causal cycles. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2944–2949). Austin, TX: Cognitive Science Society.
- Schulz, L. E., & Sommerville, J. (2006). God does not play dice: causal determinism and children’s inferences about unobserved causes. *Child Development*, 77(2), 427–442.
- Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, 27, 453–489.
- White, P. A. (2006). How well is causal structure inferred from cooccurrence information? *European Journal of Cognitive Psychology*, 18(3), 454–480.



# Enhanced Performance for Recognition of Irrelevant Target-Aligned Auditory Stimuli: Unimodal and Cross-modal Considerations

**Andrew D. Dewald** ([adewald@hawaii.edu](mailto:adewald@hawaii.edu))

Department of Psychology, University of Hawaii at Manoa  
2530 Dole Street, Honolulu, HI 96822 USA

**Scott Sinnett** ([ssinnett@hawaii.edu](mailto:ssinnett@hawaii.edu))

Department of Psychology, University of Hawaii at Manoa  
2530 Dole Street, Honolulu, HI 96822

## Abstract

Task-irrelevant stimuli are later recognized at enhanced levels providing that they had previously appeared with a task-relevant target (Dewald & Sinnett, submitted; Seitz & Watanabe, 2003, 2005). The present investigation explores this notion in the auditory sensory modality. Participants listened to a stream of auditory sounds and spoken words with the instruction to detect repetitions in only the sound stream (i.e., ignore the words). A surprise test measured recognition for the previously played words. Overall, when comparing target-aligned and non-aligned information in a later recognition task, facilitation was observed for words that had been aligned with target repetitions, despite equal presentation frequency and being irrelevant to the primary repetition task. This enhancement was mediated by the sensory modality of presentation in the surprise recognition task. Congruent auditory presentations between the exposure and recognition tasks yielded improved performance, and under cross-modal presentations the magnitude of the enhancement was greatest.

## Introduction

An emerging body of research has explored how unattended information (explicitly or implicitly presented) that appears simultaneously with a task-target in a separate task is later recognized in a subsequently presented surprise recognition test (Dewald, Sinnett, & Doumas, 2011; Seitz & Watanabe, 2003, 2005; Swallow & Jiang, 2010, 2011; Tsushima, Sasaki, & Watanabe, 2006; Tsushima, Seitz, & Watanabe, 2008). These investigations have generally consisted of a primary task involving visually presented task-relevant and previously irrelevant stimuli, with a later surprise recognition test for the irrelevant stimuli. Collectively, the findings suggest that the task-irrelevant stimuli are indeed processed, as long as they were previously aligned with a task-relevant target. However, two opposing patterns have been observed, with facilitatory (see for example Seitz & Watanabe, 2003; Watanabe, Nañez, & Seitz, 2001) or inhibitory (see for example Tsushima et al., 2006, 2008) effects for the previously aligned, but irrelevant material, seemingly dependent on whether it had originally been presented below or above, respectively, the threshold for conscious awareness.

Seitz and Watanabe (2003) required participants to identify differently colored target letters in a rapid serial visual presentation (RSVP) of the letters. Importantly, an array of randomly moving dots (irrelevant to the task) was

presented in the background during the letter identification task, of which a small subset moved coherently. Despite the coherent motion being implicit (i.e., participants were unable to reliably detect the coherent motion), participants were proficient at discriminating the motion (still presented below threshold) in a later discrimination task as long as it had previously been exposed simultaneously with the presence of a differently colored target letter during the identification task.

It is important to note that the initial motion presentation was implicit. This is a key point, as when presenting the coherent motion above threshold (i.e., 50% of the dots moved coherently), Tsushima et al (2008) showed an inhibition for target-aligned motions. Thus, it appears that facilitatory and inhibitory effects might be dependent on whether the initial presentation was presented below or above threshold, respectively.

As the majority of this research has used a relatively non-complex and simple stimuli (e.g., coherent motion in an array of moving dots), we recently examined (see Dewald et al., 2011) how a more salient stimulus, but irrelevant to the primary task, might affect later processing during an adapted inattention blindness (IB) task (see for example Rees, Ruth, Frith, & Driver, 1999; Sinnett, Costa, & Soto-Faraco, 2006). Participants were required to monitor a stream of pictures that had written words superimposed on top of each image, and respond to immediate repetitions in the picture stream while ignoring the word stream. A subsequently presented surprise recognition test for the previously ignored words demonstrated that words that had been temporally aligned with the presence of a task-target (i.e., an immediately repeating picture) were recognized significantly below chance levels (i.e., inhibited), while words that had been temporally aligned with non-targets (i.e., a non repeating picture) were recognized at chance levels.

It is important to note that the highly salient stimuli (written words) were presented for 350 ms (i.e., more than enough time for explicit perception). However, the aligned words were nonetheless inhibited in a later surprise recognition test. This finding dovetails with the earlier described findings by Tsushima et al (2008), who used target-aligned coherent motion as a stimulus and observed an analogous inhibition of motion discrimination for target-

aligned motions. That is, the coherent motion was also presented above threshold.

Other experimental paradigms have utilized different approaches to further investigate the way that the temporal alignment of task relevant and irrelevant information affects the later recognition of the irrelevant stimuli. Interestingly, despite the explicit presentation of their stimuli, the opposite findings, have been demonstrated, with a facilitation being observed for seemingly analogous conditions that elicited an inhibition. For instance, Swallow and Jiang (2010; see also Lin et al, 2010) completed a series of experiments suggesting an “attentional boost” (i.e., facilitation) for simultaneously presented, suprathreshold information in a dual-task paradigm. Participants monitored a stream of pictures of various scenes, while a series of items (small black superimposed shapes) was simultaneously paired with the presentation of each picture. The task was to remember as many of the presented scenes as possible, in addition to monitor the distractor stream of shapes for the presence of a color change. In a subsequent forced choice recognition test for the picture scenes, an enhanced recognition for pictures that had been previously presented simultaneously with the presence of a target (i.e., differently colored shape) in the distractor stream was observed.

Of particular note to Swallow and Jiang’s (2010) findings is that participants were required to attend to *both* streams of information simultaneously. Recall that in the paradigms utilized by Dewald et al (2011) and Tsushima et al (2006, 2008), participants were instructed to detect a target in one stream, but explicitly instructed not to attend to the other stream of irrelevant information, which constituted the information to be recognized in the subsequent surprise test. The division of attention between both streams of information in Swallow and Jiang’s paradigm could be the reason why a facilitation was observed in their results, rather than an inhibition.

A recent investigation by Dewald and Sinnett (submitted) aimed to more closely align their paradigm with the original procedure used by Tsushima et al (2006; see also, Seitz & Watanabe, 2003; Watanabe et al, 2001). In their experiment the exact same motion was paired with all of the target letters. This is different from the paradigm used by Swallow and Jiang (2010) or Dewald et al (2011), where a number of different pictures or words (irrelevant items in the primary task, but the items of interest in the subsequent recognition test), respectively, were aligned with targets in the primary task. In fact, the initial paradigm used by Dewald et al (2011) had 50 different words serving as the irrelevant stimuli during the picture repetition detection task. Therefore, in our subsequent work (Dewald & Sinnett, submitted) this relatively infrequent exposure rate (each of the 50 aligned words was presented only four times) was increased to an exposure rate more similar to Tsushima et al (2006, 2008). Accordingly, only a single word was aligned with all of the targets from the primary task of detecting picture repetitions. This enabled the number of instances that this task-irrelevant word was presented to be greatly

increased (from two times per word per participant, to 120 times). If the premise is that explicit information is later inhibited during a recognition task if it had been aligned with a task-relevant target, then an inhibition should have been observed for these items, especially given that the paradigm better replicated the original work demonstrating such an inhibition by Tsushima et al (2006). However, a facilitation for aligned words was seen, suggesting that the relationship between inhibition and facilitation is more complex than just whether the previously aligned stimuli were explicitly or implicitly presented. It is likely possible that the salience of the stimuli is also important, as an ignored written word is arguably processed to a higher level (i.e., semantically) than the ignored coherent motion of a moving array of dots.

Assuming that the facilitatory effect for explicitly presented, but ignored, stimuli is driven by whether the irrelevant stimulus is temporally aligned with the presence of a task-relevant target, it is important to extend these results to other sensory modalities. Despite vision being the dominant sense in humans (Chandra, Robinson, & Sinnett, 2011; Colavita et al., 1974; Posner et al., 1980; Sinnett et al., 2007), it is clear that the human perceptual experience is a result of multisensory information. Thus, it is important to explore if these inhibitory and facilitatory effects extend to other sensory modalities, as this will further inform how information is processed both within, and across modalities. For instance, Sinnett et al. (2006) demonstrated that when attentional reservoirs were depleted by a primary task, inattention blindness for spoken word perception was interrupted to the same degree as visual word recognition, however performance improved under multimodal conditions. Furthermore, a recent investigation by Dewald and Sinnett (2011a) extended this finding to the auditory modality by including an additional analysis for items that had previously appeared simultaneously with targets in the separate task. In this case, an inhibition for spoken words (explicitly presented) was observed. However, it should be noted that this investigation also used a large number of auditory words as irrelevant stimuli (i.e., identical to the visual paradigm incorporated in Dewald et al., 2011), therefore it is unknown if the inhibition for target-aligned stimuli will extend to conditions with an increased exposure rate (i.e., similar to the visual condition of Dewald & Sinnett, submitted; see also Tsushima et al., 2006, 2008).

Addressing precisely this, we adapted the same paradigm utilized in Dewald & Sinnett (submitted) to an auditory presentation, with the primary task to detect target repetitions in the sound stream, and the secondary task to subsequently recognize the previously ignored words that had been played simultaneously with the sounds. Critically, only one specific spoken word was presented to participants. This increased exposure rate lead to an enhanced performance rather than inhibited (i.e., akin to the attentional boost observed by Swallow & Jiang, 2010). If visual findings (Dewald & Sinnett, submitted) extend to the auditory modality, then we predict that the higher exposure

rate of the irrelevant spoken word will lead to a later facilitation in recognition for that word, as long as it had appeared simultaneously with an attended target in the previous task<sup>1</sup>.

A separate but equally important aspect of this research explores the modality in which the surprise test is presented. The use of dual task paradigms is pervasive in the cognitive sciences. For instance, seminal studies on dichotic listening (Broadbent, 1958; Treisman, 1964) presented participants with orthogonal messages to each ear, with the instruction to only direct their attention to a single channel of information. After this initial task, an unexpected test was given to assess the ability to recognize or recall the information that had been presented previously at the ignored ear. Of key importance here is the consideration of the sensory modality that the surprise test was presented in, respective of the initial exposure during the primary task. For instance, in these classic studies, the surprise test that probed participants' ability to process the originally presented irrelevant information was always given in the same modality as the original presentation during exposure in the primary task (auditorily).

To the best of our knowledge, this congruency in modality presentation between exposure and recognition tests has never been systematically manipulated. Therefore, given the recent interest in extending inattention blindness (see Sinnett et al., 2006) and target-alignment findings (see Dewald & Sinnett, 2010, 2011a,b) to the auditory modality and across modalities, it is important to explore how presenting the surprise test in a congruent modality would affect results, if at all. Thus, in addition to the aforementioned goals of extending this paradigm to the auditory modality and incorporating a higher exposure rate, we also presented the surprise recognition test in the same or different sensory modality, or across modalities. If primary and secondary task modality congruence is a factor, then we would expect improved results for congruent matchings vs. incongruent matchings, and potentially an additional enhancement for multimodal presentations given that performance is generally enhanced for multisensory presentations (see Driver & Spence, 2004).

## Method

**Participants.** Fifty-one participants were recruited from the University of Hawai'i at Manoa in exchange for course credit. Each participant completed the same auditory repetition detection task, but were divided across three different types of surprise recognition tests: visual only ( $n=18$ ), auditory only ( $n=17$ ), or cross-modal ( $n=16$ ). Participants were naïve to the experiment and had normal or corrected to normal hearing.

**Materials.** A total of 16 one to two syllable, high-frequency English words (average length of 5 letters, range

of 4-6 letters) were selected from the MRC psycholinguistic database (Wilson, 1988). The overall average frequency of the 150 selected words was 120 per million, ranging between 28 and 686. A native English speaker's voice was recorded reading the list three times, after which three blind listeners chose the best exemplar of each spoken word (a fourth listener was recruited in order to break a tie when needed). The selected recordings were edited to have the same length of presentation (350 ms) and average amplitude. The sound stimuli were extracted from a database of 100 familiar sounds and were also edited to 350 ms and similar average amplitude (see Sinnett et al., 2006). A stream of 960 sound-word concatenated items was created. Repeated sounds acted as the task relevant-targets. The presentation stream was broken into eight blocks of 120 trials in which an immediate sound repetition occurred on average one out of every eight trials, equating to 15 task-relevant target repetitions per block.

Eight spoken words (of the original 16) were randomly selected to overlay the 960 trial sound stream. This was done to parallel the quantity of items and exposure to irrelevant stimuli (see Dewald & Sinnett, submitted; Tsushima et al., 2006) as well as the dependent measure employed by Dewald and Sinnett (submitted; i.e., the analogous experiment in the visual domain). The presentation was pseudorandomized so that on average one out of every eight trials was an immediate sound repetition (and therefore the presentation of the same superimposed task-irrelevant target word). Critically, only one superimposed spoken word was aligned with all of the immediately repeated sounds for each participant. This single word was randomized between the eight words between participants, so as to control for any possible differences that may have existed regarding the saliency of any particular word. Lastly, it is important to note that both aligned and non-aligned words were exposed to participants in equal amounts.

A surprise recognition test for the presented words was administered after the completion of the repetition detection task. The test consisted of a total of sixteen words from which half came from the previously heard sound stream, while the other half consisted of foil words that had never been heard before (average frequency of 236 per million with a range of 165-399). The word recognition tasks were randomized and presented by DMDX software (<http://www.u.arizona.edu/jforster/dmdx.htm>) one at a time, in either the visual or auditory modality, or across modalities. For the visual presentation the words were written in bold, capitalized letters in Arial font at a size of 24 points, and remained on the screen until a response was made. For auditory presentations the words were spoken just as they were in the initial repetition detection task, albeit without the accompanying sounds. The sound stream and relevant surprise tasks were presented from two external speakers, equidistant to the computer screen. Cross-modal presentations involved the written word on the screen with the spoken word presented simultaneously.

<sup>1</sup> Note, target aligned and non-aligned words were themselves exposed in equal proportions. The higher frequency relates to comparisons to previous, but analogous research (see Dewald et al., 2011; Dewald & Sinnett, 2011, submitted).

**Procedure.** Participants were required to attend to the sound stream (i.e., they were explicitly instructed to ignore the simultaneously presented, overlaid spoken words) and respond to immediate sound repetitions by pressing the ‘G’ key on the keyboard of the computer. Each item in the sound-word presentation was presented for 350 ms with a 150-ms inter-stimulus interval (ISI; silence) for a stimulus onset asynchrony (SOA) of 500 ms. Before the first experimental block, a training block of eight trials was given and repeated until participants were familiar and comfortable with the task.

Immediately after the repetition detection task, the surprise word recognition test was administered to all participants (modality type of surprise task dependent on condition). Participants were instructed to press the “B” key if they had heard the word during the repetition detection task or, instead, the “V” key if they had not heard the word before.

## Results

To assess whether recognition performance was modulated by target alignment or the modality of presentation of the surprise task, a two-way, repeated measures ANOVA was conducted, with surprise test modality (auditory, visual, or cross-modal) as a between-subjects factor, and target alignment (target-aligned or non-aligned) as a within subjects factor. A main effect for target alignment confirmed that, overall, word recognition performance was significantly better for *target-aligned* words (72.5%) when compared to *non-aligned* words (58.8%) ( $F(2, 48) = 3.54, p = .03$ ). No main effect was found for modality of presentation ( $F(2, 48) = 1.58, p = .211$ ). Additionally, and of key importance to this analysis, an interaction was observed between target-alignment and modality of presentation ( $F(2, 48) = 3.08, p < .05$ ), suggesting that the modality of presentation in the recognition task played a role in recognition performance between *target-aligned* and *non-aligned* words. To further explore this interaction, each surprise recognition condition (auditory, visual, or cross-modal) is individually analyzed below.

**Visual surprise recognition test (VR):** Overall task performance for the surprise test was 65.8%, which was statistically different from chance ( $t(46) = 3.08, p = .001$ ). More importantly, the recognition for the *target-aligned* words (61.0%,  $SE=.11$ ) was not statistically different from *non-aligned* words (61.5%,  $SE=.57; t(17) = .03, p = .97$ ; see Figure 2). Note that the overall performance was higher due to increased performance on correctly rejecting foils. The correct rejection of foil words was compared with overall performance for *target-aligned* and *non-aligned* words. There was a significant difference between recognition for *target-aligned* words and correct rejections (*target-aligned*: 61.0%,  $SE=.11$  vs. *CR*: 79.6%,  $SE=.04, t(17)=2.14, p = .02$ ) as well as a significant difference between correctly recognizing *non-aligned* words and correct rejections (*non-aligned*: 61.5%,  $SE=.57$  vs. *CR*: 79.6%,  $SE=.04, t(17)=2.47, p = .02$ ). Lastly, confirming that participants were able to successfully perform the initial repetition task, the target

repetitions were significantly detected (**Hits**: 68%  $SE=.18$  vs. **Misses**: 28%  $SE=.04, t(17)=9.42, p < .001$ ).

**Auditory surprise recognition test (AR):** Overall task performance was 71.0%, which was statistically different from chance ( $t(45) = 4.47, p = .001$ ). Contrary to the visual condition, recognition for *target-aligned* words (76.4%,  $SE=.10$ ) was significantly better than *non-aligned* words (58.8%,  $SE=.05, t(16) = 2.37, p = .05$ ; see Figure 1). Again, unlike the first experiment, there was no difference in performance between *target-aligned* word recognition and correct rejections of foils (*target-aligned*: 76.04%,  $SE=.10$  vs. *CR*: 77.7%,  $SE=.06, t(16)=0.09, p = .967$ ). However, there was a significant difference between *non-aligned* word recognition and correctly rejecting foil words (*non-aligned*: 58.8%,  $SE=.05$  vs. *CR*: 77.7%,  $SE=.06, t(16)=2.35, p = .03$ ). Lastly, participants accurately detected target repetitions (**Hits**: 73%  $SE=.10$  vs. **Misses**: 23%  $SE=.11, t(16)=8.47, p < .001$ ).

**Cross-modal surprise recognition test (CR):** Overall word recognition was 69.2%, which was statistically better than chance ( $t(43) = 4.28, p = .001$ ). Recognition for *target-aligned* words (81.2%,  $SE=.10$ ) was significantly better than *non-aligned* words (55.7%,  $SE=.05; t(15) = 2.59, p = .05$ ; see Figure 1). There was no difference in performance between *target-aligned* word recognition and correct rejections of foils (*target-aligned*: 81.2%,  $SE=.10$  vs. *CR*: 70.8%,  $SE=.06, t(15)=0.82, p = .422$ ). Again, however, there was a significant difference between *non-aligned* word recognition and correctly rejecting foil words (*non-aligned*: 58.7%,  $SE=.05$  vs. *CR*: 70.8%,  $SE=.06, t(15)=2.07, p = .05$ ). Additionally, participants accurately detected immediate sound target repetitions in the primary task (**Hits**: 75%  $SE=.13$  vs. **Misses**: 24%  $SE=.03, t(15)=9.23, p < .001$ ).

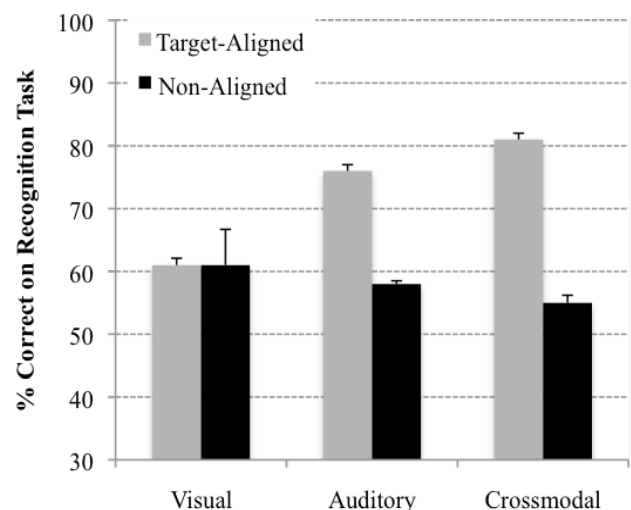


Figure 1. Recognition percentages for Target-Aligned (grey bar) and Non-Aligned (black bar) words dependent on the modality of the surprise test.

## Discussion

The present findings extend investigations exploring how above threshold, but unattended information is processed when it appears simultaneously with an attended target. We demonstrate that in all conditions, the overall recognition of the words in the surprise task was better than chance, despite attention not being directed to the words. This finding is contrary to what can be predicted from inattention blindness investigations in that the overall recognition of the words should have been at chance, or perhaps even inhibited for the irrelevant stimulus (Dewald et al, 2011; Rees et al, 1999; Sinnett et al, 2006). It is likely that the increased exposure to (i.e., fewer items) to the irrelevant stimuli drove this effect. More importantly, the interaction between target-alignment and the modality of the surprise test was significant, driven by performance for *target-aligned* words being statistically better than non-aligned words in the **auditory recognition** (76% vs. 58%) and **cross-modal recognition** (81% vs. 55%) conditions. Note, when comparing the magnitude of the enhancement between these two conditions, there was (no) additional improvement for cross-modal conditions. Accordingly, this suggests that, at least in the present case, temporally aligning explicitly presented, irrelevant, auditory stimuli with relevant auditory target stimuli facilitates subsequent recognition of the irrelevant stimuli, but only if the recognition test is presented in the same modality as the initial task, or across modalities. Thus, an “attentional boost” (see Swallow & Jiang, 2010) for irrelevant stimuli was observed, as long as they were initially presented simultaneously with a target in the picture repetition task, despite not receiving direct attention.

Interestingly, when the surprise recognition task was presented in an incongruent modality from the exposure (i.e., the visual surprise recognition task) there was no difference between *target-aligned* and *non-aligned* words. Previous findings have demonstrated an inhibition for both visually aligned (Dewald et al., 2011) and auditorily aligned words (Dewald & Sinnett, 2011a) in a dual-task paradigm, despite not controlling for task modality congruence (i.e., the visual example presented information in the congruent modality, while the auditory example also presented the surprise test in the visual modality). However, it should be noted that both of these examples used a much lower exposure rate (i.e., 50 words had been aligned in the repetition detection task rather than only one that was repeated 120 times). This could possibly explain why we failed to observe a difference in the visual task condition here.

Further complicating the matter, Dewald and Sinnett (submitted) observed an enhancement for target-aligned words under nearly isomorphic conditions as utilized here, with the only difference being that the initial repetition detection task was presented in the visual modality. In fact, all research conducted thus far, exploring the fate of irrelevant auditory words either target-aligned or not, has presented the words in the visual modality during the

subsequent recognition task (Dewald & Sinnett, 2011a; Sinnett et al, 2006; see Dewald & Sinnett, 2011b for a cross-modal example). It is possible that the incongruence between the modality of presentation and subsequent recognition may in some way affect the recognition of the previously unattended items. Thus, further research should explore whether the observed enhancement for visually aligned words when tested with a visually presented surprise task extends to incongruent task modality presentations (or across modalities).

In the present experiment the cross-modal presentation lead to the greatest magnitude of enhancement for the previously aligned words in the surprise recognition test. This aligns well with previous investigations of attentional allocation across sensory modalities in perceptual and recognition tasks, suggesting that cross-modal presentations generally lead to superior to performance when compared to unimodal presentations (Dewald & Sinnett, 2011b; Duncan, Martens, & Ward, 1997; Sinnett et al, 2006; Toro, Sinnett, & Soto-Faraco, 2005).

The possibility that the presentation modality of the surprise test could modulate performance has yet to be fully explored. The resulting recognition performance, observed when the surprise task was incongruent to the initial exposure, necessitates an investigation into the nature of modality congruence in dual-task paradigms. Clearly, the current findings establish that attention must be paid to future methodologies utilizing a dual-task paradigm and the implications of modality congruence between the primary and secondary task. Furthermore, previous investigations that employ incongruent modalities between exposure and recognition must be revisited (Dewald et al, 2011, Sinnett et al, 2006).

Combined, the present findings and previous research offer insight into how irrelevant information is processed when it is presented simultaneously with an attended target. Under certain circumstances, as demonstrated here (see also Dewald et al., 2011 Dewald & Sinnett, submitted; Seitz & Watanabe, 2003; Tsushima et al., 2008), unattended stimuli can be perceived and affect behavior. Here we extend findings from previous research into the auditory sensory modality, and show a facilitation (i.e., attentional boost) for a highly exposed stimulus that was aligned with a target in the previous task when compared with items that were not aligned. More importantly however, we demonstrate that careful consideration must be given to the modality of presentation of dual-task paradigms in general.

## References

- Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Science U.S.A.*, 90, 5718–5722.
- Borst, A., & Egelhaaf, M. (1989). Principles of visual motion detection. *Trends in Neurosciences*, 12(8), 297-306.



- Broadbent, D.E. (1958). *Perception and communication*. London: Pergamon Press.
- Chandra, M., Robinson, C. W., & Sinnett, S. (2011). Coexistence of multiple modal dominances. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*
- Cherry, E.C. (1953). Some experiments on recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, 25, 957-979.
- Dewald, A.D., & Sinnett, S. (submitted). A window of perception when diverting attention? Enhancing recognition for explicitly presented, unattended, and irrelevant visual stimuli by target alignment. *Submitted to the Proceedings of the 34th Annual Conference of the Cognitive Psychology Society*.
- Dewald, A.D., & Sinnett, S. (2011a). An inhibited recognition performance for explicitly presented target-aligned irrelevant stimuli in the auditory modality. *Proceedings of the 33rd Annual Conference of the Cognitive Psychology Society*.
- Dewald, A.D. & Sinnett, S. (2011b). A multimodal investigation of recognition performance for target-aligned but irrelevant stimuli. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.
- Dewald, A.D., Sinnett, S., & Dumas, L.A.A. (2011). Conditions of directed attention inhibit recognition performance for explicitly presented target-aligned irrelevant stimuli. *Acta Psychologica*, 138, 60-67.
- Duncan J, Martens S, Ward R (1997), Restricted attentional capacity within but not between sensory modalities. *Nature* 387(6635):808-10
- Driver, J., & Spence, C. (2004). Cross-modal spatial attention: Evidence from human performance. In C. Spence & J. Driver (Eds.), *Cross-modal space and cross-modal attention*. Oxford, UK: Oxford University Press.
- Lin, J.Y., Pye, A.D., Murray, & Boynton, G.M. (2010). Enhanced memory for scenes presented at relevant points in time. *PLoS Biol*, 8(3), E1000337.
- Rees, G., Russell, C., Frith, C. D., & Driver, J. (1999). Inattention blindness versus inattentional amnesia for fixated but ignored words. *Science*, 286, 2504-2507.
- Roelfsema, P. R., van Ooyen, A., & Watanabe, T. (2009). Perceptual learning rules based on reinforces and attention. *Trends in Cognitive Sciences*, 14(2), 64-71.
- Seitz, A. R., Kim, R., & Shams, L. (2006). Sound facilitates visual learning. *Current Biology*, 16, 1422- 1427.
- Seitz, A. R. & Watanabe, T. (2003). Psychophysics: Is subliminal learning really passive? *Nature*, 422, 36.
- Seitz, A. R. & Watanabe, T. (2005). A unified model for perceptual learning. *Trends in Cognitive Science*, 9 (7), 329-334.
- Sinnett, S., Costa, A., & Soto-Faraco, S. (2006). Manipulating inattention blindness within and across sensory modalities. *Quarterly Journal of experimental Psychology*, 59(8), 1425-1442
- Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 174-215.
- Swallow K. M., & Jiang, Y. V. (2010). The attentional boost effect: Transient increases in attention to one task enhance performance in a second task. *Cognition*, 115, 118-132.
- Swallow K.M., & Jiang, Y. V. (2011). The role of timing in the attentional boost effect. *Attention, Perception, and Psychophysics*, 73, 389-404.
- Toro, J.M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, 97, 25-34
- Triesman, A.M. (1964). Selective attention in man. *British Medical Bulletin*, 20, 12-16.
- Tsushima, Y., Sasaki, Y., & Watanabe, T. (2006). Greater disruption due to failure of inhibitory control on an ambiguous distractor. *Science*, 314, 1786-1788.
- Tsushima, Y., Seitz, A. R., & Watanabe, T. (2008). Task-irrelevant learning occurs only when the irrelevant feature is weak. *Current Biology*, 18(12), 516-517.
- Watanabe, T., Náñez, Y., & Sasaki, S. (2001). Perceptual learning without perception. *Nature*, 413, 844-848.
- Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, version 2. *Behavioural Research Methods, Instruments and Computers*, 20, 6-11.

# The semantic structure of sensory vocabulary in an African language

Mark Dingemanse<sup>1</sup> (mark.dingemanse@mpi.nl) and Asifa Majid<sup>1,2</sup> (asifa.majid@mpi.nl)

<sup>1</sup>Max Planck Institute for Psycholinguistics, 6500AH Nijmegen, The Netherlands

<sup>2</sup>Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, NL

## Abstract

The widespread occurrence of ideophones, large classes of words specialized in evoking sensory imagery, is little known outside linguistics and anthropology. Ideophones are a common feature in many of the world's languages but are underdeveloped in English and other Indo-European languages. Here we study the meanings of ideophones in Siwu (a Kwa language from Ghana) using a pile-sorting task. The goal was to uncover the underlying structure of the lexical space and to examine the claimed link between ideophones and perception. We found that Siwu ideophones are principally organized around fine-grained aspects of sensory perception, and map onto salient psychophysical dimensions identified in sensory science. The results ratify ideophones as dedicated sensory vocabulary and underline the relevance of ideophones for research on language and perception.

**Keywords:** semantics; sensory vocabulary; ideophones; sensory perception; mental lexicon

## Introduction

Ideophones are marked words that depict sensory imagery, like *sinisinisini* 'closely woven' and *saaa* 'cool sensation' in Siwu, a language of Ghana (Dingemanse, 2011a), or *gorogoro* 'rolling' and *pikapika* 'shiny' in Japanese (Kita, 1997). These highly specific renditions of sensory perceptions — the precise texture of an object felt, the manner of motion of a figure, the visual appearance of a surface — are a common feature of many of the world's languages, with some languages having ideophone inventories numbering into the thousands (Doke & Vilakazi, 1953; Kakehi, Tamori, & Schourup, 1996); but they are underdeveloped in English and other Indo-European languages (Nuckolls, 2004).

The widespread occurrence in natural languages of large classes of words specialized for depicting sensory perception is of significance to research examining the relation between language and perception. Yet, so far, ideophones have been a secret well-kept from cognitive science, studied mostly in lesser known languages by field linguists and anthropologists (Nuckolls, 1996; Voeltz & Kilian-Hatz, 2001). Although the tight link between language and perception in ideophones has long been recognized (Westermann, 1927), the study of their semantic structure has lagged and important questions remain unanswered. What is the link between ideophones and sensory imagery? How should native speakers' knowledge of these sensory words be characterized? What aspects of sensory perception are involved in their representation in the mental lexicon?

Previous semantic classifications of ideophones fail to answer such questions because they usually do not represent native speaker knowledge but reproduce the linguistic categories of the analyst. An example is Alexandre (1966), who classified ideophones in Bulu (a Bantu language from Cameroon) into the five-senses folk model of his own native language: "auditive, visual, tactile, gustative, and olfactive". However, since sensory science shows a larger taxonomy of the senses (Moller, 2002) — including not just the external senses but also interoception (sensitivity to inner physiological conditions) and proprioception (sense of balance and body posture) — we cannot assume the universal relevance of a Western folk model of perception to the classification of ideophones. Other studies of the semantic structure of ideophones have used evidence from lexical collocations, inferring for instance the sensory modes of ideophones from the verbs with which they co-occur (Diffloth, 1976; Awoyale, 1983). Such methods are likely to provide more insight than classifications based on analysts' intuitions, but since the verbs may underspecify the semantic space and since not all ideophones appear in regular collocation with verbs, they may provide an incomplete picture.

Recently, some studies have begun to investigate the semantic structure and sound-symbolic nature of ideophones in Japanese. In a learning study with infants, Imai et al. (2008) built novel words on the template of Japanese ideophones ('mimetics') for manners of motion, and found that these sound-symbolic forms facilitate early verb learning. Osaka (1990) investigated Japanese ideophones of crying, laughing, and talking using a similarity judgment task and found they could be arranged on sensory scales of intensity. In a follow-up study, Osaka and Osaka (2005) focused on laughter ideophones using fMRI, and found they activated striatal reward areas, which they connect to the image-evoking qualities of ideophones. Although interesting in their own right, none of these studies address the domain of ideophones as a whole, nor do they articulate what the principles of organization within this domain might be. Here we study a representative set of ideophones from Siwu, a Kwa language spoken in Ghana, West Africa — a linguistic region known for its extensive ideophone systems (Blench, 2010). The aim was to explore the semantic structure of the domain by capturing native speakers' intuitions.

How might ideophones be organized in the mental lexicon? One possibility is that ideophones are organized in terms of sensory perception (Kita, 1997). Another possibility is that they are organized in terms of semantic dimensions like activity, potency and evaluation, factors

commonly found in multidimensional scaling analyses of affective vocabulary (Osgood, Suci, & Tannenbaum, 1957; Osaka, 1990). A third possibility is that both principles operate, with dimensional and categorical properties working in parallel, perhaps at different levels of granularity. In order to test these hypotheses, we obtained similarity judgments of ideophones using a pile-sorting task.

Another matter of investigation concerns the nature of ideophones as sensory vocabulary. Work in the domain of touch for instance has identified psychophysical dimensions such as rough-smooth, hard-soft, springiness, and firmness (Yoshida, 1968). We considered how the distinctions encoded by ideophones map onto these dimensions, and what this may reveal about ideophones as sensory words and about the construction of sensory vocabularies in general.

## Methods

We used a pile-sorting task (Bernard, 2006; Weller & Romney, 1988) to collect similarity data for a frequency-based selection of 58 ideophones. Sorting tasks have been conducted with a wide variety of items, from English nouns (Miller, 1969) and texture words (Bhushan, Rao, & Lohse, 1997) to Navaho food concepts (Perchonock & Werner, 1969). Hierarchical clustering and multidimensional scaling were used to analyze the similarity data to uncover the underlying dimensions used to categorize the items.

## Participants

Fourteen participants (ten men), with a mean age of 37.1 years, took part in the sorting task. All were native speakers of Siwu. All were literate in Siwu and in Ewe, a regional language of wider communication.

## Materials and design

Ideophones are a class of words distinct from nouns and verbs in the Siwu language. They are formally identifiable in terms of phonotactics, word forms, expressive morphology, syntax, and prosody (Dingemanse, 2011b, pp. 133–160). There are at least 400 ideophones in Siwu. It would clearly be difficult to investigate all ideophones in a single sorting task. A selection of ideophones was therefore made using frequency as a criterion, since this is one measure of the representativeness of the ideophone inventory.

A total of 58 ideophones were selected as follows: a first selection included all ideophones which occurred at least twice in a corpus of five hours of naturally occurring conversations recorded in informal situations (Dingemanse, 2011b). There were 38 such ideophones. A second selection included ideophones which occurred more than three times in responses to elicitation tasks probing different perceptual domains (Majid & Levinson, 2007, 2011). There were 26 such ideophones. Discounting overlaps between the two sets left 20 new ideophones, bringing the total at 58 (Table 1). Each of the 58 ideophones was printed on a 105x35mm card in Siwu orthography.

## Procedure

The cards were presented to participants in a 6x10 array arranged according to a fixed random order. While laying the cards out, the ideophones were read out-loud to make sure they were familiar to participants. Three items, *belele*, *lelele*, and *melemele*, were not always recognized immediately. When this happened, the ideophones were presented in an example sentence from the conversational corpus. Participants were always able to recognize the words in context.

Instructions were given in Siwu: participants were asked to arrange the cards into groups of similar items. The instruction was to “place together words that belong together, words that are akin” (*atɔ̃mɛ wa lõkote, wa lõde manyibi*). Participants were given no explicit criteria for judging similarity and were told they could create as many piles as they liked with as many items per pile as they wished. For three participants who asked for clarification, the sorting procedure was demonstrated using ideophones not included in the set.

Cards were not literally piled but grouped together on a flat surface so all items were kept in view. Participants took between 30 and 70 minutes to group all 58 ideophones. Once participants had completed their groupings, they were debriefed and asked to describe each group.

Table 1: Siwu ideophones used in the sorting task

---

<i>sinisini</i>	closely woven (01),	<i>pepepepe</i>	precisely (02),
<i>nyanyariĩ</i>	dirty (03),	<i>nyagbalaa</i>	pungent/sour (04),
<i>wɔ̃rɔ̃wɔ̃rɔ̃</i>	spotted (05),	<i>m̃ɛr̃m̃ɛr̃ɛ</i>	tasty (06),
<i>yululu</i>	cold (07),	<i>pɔ̃tɔ̃pɔ̃tɔ̃</i>	dirty/muddy (08),
<i>pɔ̃kɔ̃sɔ̃ɔ̃</i>	slow (09),	<i>minimini</i>	round (10),
<i>gbiim</i>	sound of explosion (11),	<i>tagbaraa</i>	long (12),
<i>gbegbe</i>	tough (13),	<i>qɔ̃bɔ̃rɔ̃ɔ̃</i>	soft (14),
<i>belele</i>	broad and extended (15),	<i>shu</i>	sound of ignition (16),
<i>tititi</i>	big and wide (17),	<i>saaa</i>	cool sensation (18),
<i>dzooroo</i>	far (19),	<i>wosoroo</i>	rough (21),
<i>gɔ̃dɔ̃rɔ̃</i>	crooked (21),	<i>gbidii</i>	excessive activity (22),
<i>bebrebee</i>	many (23),	<i>melemele</i>	talkative (24),
<i>pɔ̃lɔ̃pɔ̃lɔ̃</i>	smooth (25),	<i>pelee</i>	completely (26),
<i>lelele</i>	full to the brim (27),	<i>safaraa</i>	rough/coarse-grained (28),
<i>gelegele</i>	shiny (29),	<i>q̃ɛkper̃ɛ</i>	fine-grained (30),
<i>kpɔ̃</i>	sound of impact (31),	<i>kunukunu</i>	completely empty (32),
<i>wĩr̃wĩr̃</i>	small things dispersed (33),	<i>tsuru</i>	sound of sth. rapidly passing by (34),
<i>teteree</i>	loud (35),	<i>kp̃ãũ</i>	big, enormous (36),
<i>waa</i>	sound of water gushing (37),	<i>fiẽfiẽ</i>	silky (38),
<i>buaa</i>	tasteless (39),	<i>mlamla</i>	quickly (40),
<i>kananaa</i>	silent (41),	<i>kr̃ɔ̃kr̃ɔ̃</i>	pleasantly smelling (42),
<i>kpɔ̃lɔ̃kpɔ̃lɔ̃</i>	slippery (43),	<i>gbugburu</i>	tough (44),
<i>gbogboro</i>	tough (45),	<i>kekei</i>	small (46),
<i>kp̃inakp̃ina</i>	black (47),	<i>sɔ̃dzɔ̃lɔ̃ɔ̃</i>	oblong (48),
<i>giligili</i>	round (49),	<i>kpokporo</i>	hard (50),
<i>fututu</i>	pure white (51),	<i>mĩɔ̃mĩɔ̃</i>	pointy (52),
<i>kpoo</i>	silent (53),	<i>wurufuu</i>	fluffy (54),
<i>kpu</i>	sound of impact (55),	<i>yuãyuã</i>	burning sensation (56),
<i>nỹɛ̃k̃nỹɛ̃k̃</i>	very sweet (57),	<i>fũ̃ɛ̃fũ̃</i>	soft-malleable (58)

---



## Results

Dissimilarity matrices were created for each participant and then summed to create the equivalence matrix used for statistical analysis.

### Cluster analysis

Hierarchical cluster analysis of the data was performed using the average-linkage-between-groups method, which does not presuppose a particular type of structure in the data. Figure 1 below shows the dendrogram for the cluster analysis.

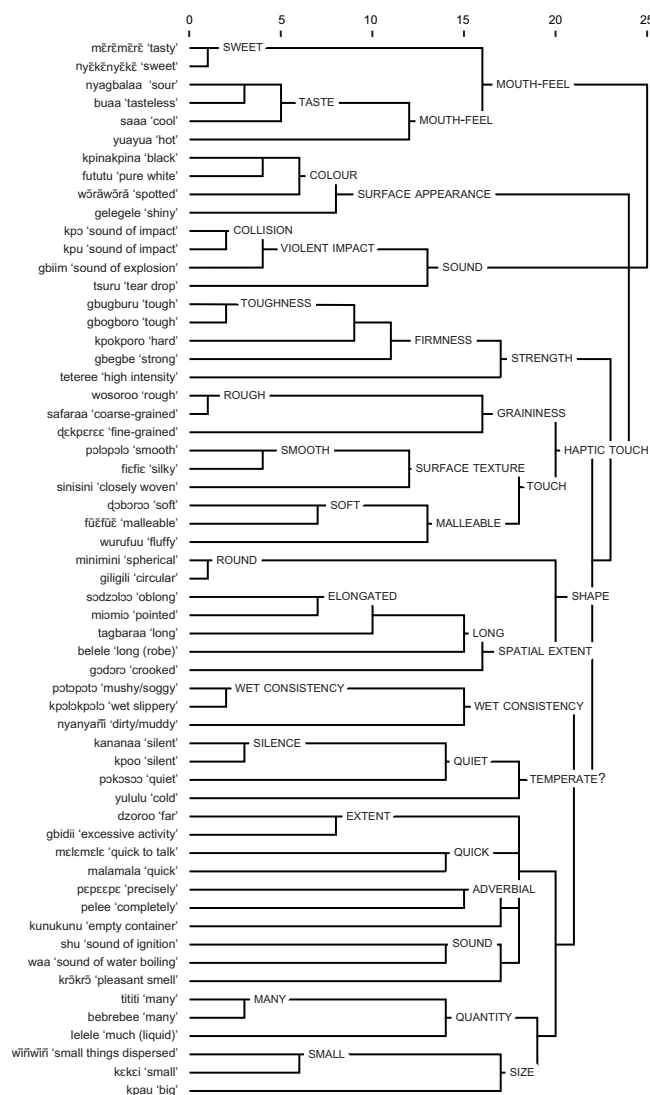


Figure 1: Dendrogram using average linkage

All clusters with a relatively high internal similarity (distances of 10 or lower) can be identified in terms of aspects of sensory perception: SWEET, TASTE, SURFACE APPEARANCE (including COLOUR), VIOLENT IMPACT (including COLLISION), TOUGHNESS, ROUGH, SMOOTH, SOFT, ROUND, ELONGATED, WET CONSISTENCY, SILENCE, EXTENT,

MANY and SMALL. Many of these clusters also join to form coherent higher-order groups. Such higher-order clusters (distances of 20 or lower) include MOUTH-FEEL (merging SWEET and TASTE), SOUND, STRENGTH, HAPTIC TOUCH (merging GRAININESS, SURFACE TEXTURE, and MALLEABLE), SHAPE (merging ROUND and SPATIAL EXTENT), and four clusters that are extended with one member each: WET CONSISTENCY, QUIET, QUANTITY, and SIZE. There is a small residue of words that were rarely grouped with other words, or grouped in very different ways by participants.

The cluster analysis showed two main things. First, it showed that sensory perception is the most salient organizing principle in the domain. This is confirmed by the descriptions given by participants in the debriefing; participants explained their groupings in terms of sensory modalities or acts of sensory perception. Second, it identified a number of coherent sensory categories (and interrelations between categories) at a finer grain than traditional five-senses classifications, providing insight into the different kinds of sensory imagery encoded by ideophones.

### Multidimensional scaling

Multidimensional scaling was used to test whether similarity judgments from ideophones were organized in terms of the three well-known dimensions of affective meaning: activity, evaluation and potency (Osgood et al., 1957; Osaka, 1990). A three-dimensional MDS solution showed an RSQ of .80 and a Kruskal stress value of .21 (for 2 dimensions stress was .27, while for 4 dimensions it was only slightly lower at .20). Within the MDS analysis, the major groups identified by the cluster analysis remain largely intact. This lends additional credibility to the sensory categories identified there and to the interpretation that aspects of sensory perception are an important organizing principle in the domain of ideophones.

Does the MDS analysis additionally support a dimensional interpretation? Dimension 1 features TASTE and SHAPE on opposite poles, reinforcing the strong internal coherence of these clusters and their mutual incompatibility in terms of sorting decisions. It does not easily yield a simple dimensional interpretation. Dimension 2 pulls apart a cluster of SOUND ideophones on one side with SURFACE APPEARANCE on the other. While this might be construed as an opposition of events versus states, suggestive of the dimension of activity, it is more plausible that, just as in Dimension 1, these particular clusters are pulled apart simply as a result of their high internal coherence coupled with their semantic incompatibility. The cluster analysis supports this interpretation.

The connotative dimensions of evaluation (good/bad) and potency (strong/weak) do not seem to play a role either. For instance, if evaluation were relevant, we might expect ideophones like *buàà* 'tasteless' and *nyanyarĩĩ* 'dirty' to cluster together, and to be far apart (on one dimension) from ideophones like *mêrêmêrê* 'delicious' and *nyêkênyêkê* 'intensely sweet', but this is not the case. If potency were

relevant we might expect ideophones that differ mainly in terms of intensity to be distant from one another, but they are not.

The nature of the clusters identified both in the cluster analysis and the MDS analysis provides a likely explanation of why the generic connotative dimensions may be absent. The clusters concern aspects of sensory experience from a diverse range of modalities. The sensory modalities differ in discrete, qualitative ways, and this appears to be reflected in our findings.

## Discussion

### Ideophones as dedicated sensory words

Our results suggest that sensory perception is the main organizing principle in the lexical domain of ideophones, despite other plausible organizing principles. As noted, one might expect generic dimensions commonly found in semantic differential studies to be apparent in this domain. Alternatively, sorting could be done on the basis of other principles, such as word length (grouping *saaa* ‘cool sensation’ with *kpoo* ‘silence’ or *tagbaraa* ‘long’ with *pəkəsəə* ‘slow’) or similarity in sound or spelling (*giligili* ‘circular’ with *gelegele* ‘shiny’). Neither the generic dimensions nor the non-semantic features play an appreciable role in the similarity data. This provides independent confirmation of claims made in the anthropological and linguistic literature that ideophones are dedicated sensory words. It also meshes well with empirical findings on the use of ideophones. Data from a corpus of everyday conversations shows that Siwu speakers use ideophones to demonstrate expertise (communicating very precisely about perceptual qualities) and also to share in sensory spectacles in storytelling (Dingemanse, 2011b, pp. 251–300). Both uses rely on the nature of ideophones as marked words that depict sensory imagery.

Previous explorations of the semantic domain of ideophones have mostly relied on informal judgments by the analyst. There is a danger that they do little more than reproduce folk models of the analysts’ metalanguage. In contrast, the sorting task allows us to capture a semantic-perceptual space based on native speaker judgments. The fine-grained categories that emerged from the cluster analysis (for instance FIRMNESS, SMOOTHNESS, or SURFACE APPEARANCE) are anchored in the data and agree with speakers’ own explanations of their groupings. Therefore they have a different ontological status than coarse-grained labels like Alexandre’s (1966) five-sense classification of ideophones. Moreover, the cluster analysis brings to light structures that do not easily fit into such broad classifications: categories that may combine content from multiple modalities (e.g. WET CONSISTENCY, which may combine vision and touch), or categories at a finer grain (as in the subcategories of HAPTIC TOUCH, which include SURFACE TEXTURE, MALLEABILITY and GRAININESS).

### Sorting task reproduces psychophysical dimensions

The clusters related to haptic touch provide a useful illustration of the nature of ideophones as sensory vocabulary. Ideophones encoding haptic touch sensations are common in Siwu (Dingemanse, 2011a), and are accordingly well-represented in the frequency-based selection of ideophones used in the sorting task. Independently, we know that Siwu speakers describe tactile stimuli predominantly with ideophones (Dingemanse, 2011b).

Psychophysical research in the domain of haptics has identified a number of salient dimensions of touch perception: rough-smooth, hard-soft, springiness, sticky-slipperiness, and firmness (Yoshida, 1968; Guest et al., 2010; Klatzky, Lederman, & Reed, 1987). The clusters identified through the cluster analysis correspond to these dimensions quite closely: we find ROUGHNESS, SMOOTHNESS, SOFTNESS, MALLEABILITY (the equivalent of springiness), and FIRMNESS. So the distinctions encoded by texture ideophones appear to map well onto the psychophysical properties of the domain.

What is it that makes ideophones good at encoding fine semantic distinctions in ways that reproduce psychophysical dimensions? One reason may be the nature of ideophones as a form class in natural languages. Guest et al. (2010) studied English texture words in order to develop a touch lexicon. They considered a mix of adjectives (rough, soft), verbs (burning, vibrating), terms derived from verbs (sticky, prickly), and source-based adverbs that use simile (sandy, woolly, furry). They found that when the set of words was systematically narrowed to improve the coverage of the touch lexicon, many of the source-based terms were removed, suggesting that abstract (i.e. non-source-based) terms and uniformity in linguistic resources may lead to a better coverage of the semantic-perceptual space. These two features, abstractness of meaning and uniformity of linguistic form, are precisely characteristic of ideophones.

Another reason may be that ideophones are special in their mode of signification. Across languages, ideophones tend to be produced with prosodic foregrounding and expressive features like reduplication and lengthening: signs of the fact that they are depictions, as opposed to descriptions, of sensory imagery (Kunene, 1965; Nuckolls, 1996; Dingemanse, 2011a). Being depictions, ideophones often employ several forms of sound-symbolism or iconicity (Westermann, 1927; Diffloth, 1972; Awoyale, 1983). The iconic use of verbal material, more gradient and less arbitrary than ordinary words, may make ideophones especially fit for representing sensory imagery. Three broad types of iconic mappings have been identified in Siwu ideophones (Dingemanse, 2011c), and similar mappings are found in other ideophone inventories (Tufvesson, 2011). Such iconic mappings allow ideophones to move beyond the imitation of singular events towards perceptual analogies and generalizations of event structure.

The close link between the structural properties of ideophones and the deeply sensory nature of their meanings suggests that ideophones constitute a promising area for research in embodied cognition and sensory science (Barsalou, 1999; Kita, 1997; Akita, 2010).

### Organizing principles and level of granularity

Earlier, we mentioned three hypotheses about the organization of ideophones in the mental lexicon: they may be organized in terms of aspects of sensory perception, in terms of generic connotative dimensions, or in terms of a combination of both. Analysis of the similarity judgment data revealed an array of fine-grained sensory categories, suggesting that for this set of ideophones, sensory perception is the most salient organizing principle. Connotative dimensions like activity, potency and evaluation, on the other hand, appear to play no role in the organization of the domain at this level of granularity.

Level of granularity may be a key concept here. The present study took a selection of high frequency ideophones across subdomains. Earlier studies, in contrast, have tended to focus on small subsets of ideophones from circumscribed semantic domains. Osaka (1990) for instance collected similarity data for Japanese ideophones in the semantic domains of talking, crying, and laughing, and found that in each of these domains it was possible to place the ideophones on sensory scales of intensity. This is not surprising: within any sufficiently semantically homogenous cluster of words, the broad connotative dimensions identified by Osgood et al. (1957) will likely play some role (Boster, 2005). This suggests that the third hypothesis, which holds that ideophones are organized by a combination of sensory aspects and generic connotative dimensions, is the most plausible one.

Combining the results from earlier studies and the findings of the present study, we suggest that the word class of ideophones as a whole is organized in terms of aspects of sensory perception, broadly construed; that the semantic distinctions made by ideophones in specific sensory domains may closely follow psychophysical properties; and that connotative dimensions like activity, potency and evaluation may play a role in the internal organization of subsets of sensory words that are sufficiently semantically homogenous.

### Conclusions

We have shown that the lexical space of ideophones is organized around aspects of sensory perception. The sensory categories identified through hierarchical cluster analysis are more fine-grained than previous classifications, demonstrating the utility of sorting tasks to capture native speaker intuitions and to map out semantic structures in the mental lexicon. In addition, several of the clusters that emerge from the analysis appear to map neatly onto psychophysical dimensions identified in other research, suggesting that ideophones constitute a sophisticated sensory vocabulary.

Our findings set the scene for cross-linguistic studies of ideophones in the future. Are the same aspects of sensory perception reproduced across languages, suggesting a shared semantic-perceptual space? Do languages impose their own organization on the domain? Does the finer organization of ideophones in specific sensory modalities match relevant psychophysical dimensions, as we see in Siwu ideophones for haptic touch? What are the cognitive consequences of having such dedicated sensory words in terms of mental representation, production and comprehension? And how do the design principles of ideophones relate to their fitness to evoke sensory imagery? The domain of ideophones is ripe for cross-linguistic investigation of these questions.

Ideophones are an example of the kind of linguistic structure that is in danger of being overlooked if the cognitive sciences keep focusing on the very thin slice of human behavior exhibited by Western populations (Henrich, Heine, & Norenzayan, 2010; Majid & Levinson, 2010). Here we hope to have shown that they provide a fruitful avenue to explore how languages can encode sensory perception, and that they are a powerful reminder of the fact that cross-linguistic diversity is an asset rather than an obstacle in the study of language and mind.

### Acknowledgments

We thank the Mawu people of Akpafu and Lolobi for supporting research on the Siwu language, and we are especially grateful to all participants in Mempeasem, Todzi and Adakɔ for their involvement. *Mi ndo karabra lo!* We also thank two anonymous reviewers for helpful comments. This work was funded by the Max Planck Institute for Psycholinguistics.

### References

- Akita, K. (2010). An embodied semantic analysis of psychological mimetics in Japanese. *Linguistics*, 48, 1195–1220. doi:10.1515/LING.2010.039
- Alexandre, P. (1966). Préliminaire à une présentation des idéophones Bulu. In J. Lukas (Ed.), *Neue Afrikanische Studien, Hamburger Beiträge zur Afrika-Kunde* (pp. 9–28). Hamburg: Deutsches Institut für Afrika-Forschung.
- Awoyale, Y. (1983). On the semantic fields of Yoruba ideophones. *Journal of the Linguistic Association of Nigeria*, 2, 11–22.
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(04), 577–660.
- Bernard, H. R. (2006). *Research methods in anthropology*. Rowman Altamira.
- Bhushan, N., Rao, A. R., & Lohse, G. L. (1997). The texture lexicon: Understanding the categorization of visual texture terms and their relationship to texture images. *Cognitive Science*, 21(2), 219–246. doi:10.1016/S0364-0213(99)80023-8
- Blench, R. (2010). The sensory world: ideophones in Africa and elsewhere. In A. Storch (Ed.), *Perception of*

- the Invisible: Religion, Historical Semantics and the Role of Perceptive Verbs*, Sprache und Geschichte in Afrika (pp. 275–296). Cologne: Köppe.
- Boster, J. S. (2005). Emotion categories across languages. In C. Lefebvre & H. Cohen (Eds.), *Handbook of Categorization in the Cognitive Sciences* (pp. 187–223). Amsterdam: Elsevier Science.
- Diffloth, G. (1972). Notes on expressive meaning. *Chicago Linguistic Society*, 8, 440–447.
- Diffloth, G. (1976). Expressives in Semai. *Oceanic Linguistics Special Publications*, (13), 249–264.
- Dingemanse, M. (2011a). Ideophones and the aesthetics of everyday language in a West-African society. *The Senses and Society*, 6(1), 77–85. doi:10.2752/174589311X12893982233830
- Dingemanse, M. (2011b). *The Meaning and Use of Ideophones in Siwu* (PhD dissertation). Radboud University, Nijmegen. Retrieved from <http://thesis.ideophone.org/>
- Dingemanse, M. (2011c). Ezra Pound among the Mawu: Ideophones and Iconicity in Siwu. In P. Michelucci, O. Fischer, & C. Ljungberg (Eds.), *Semblance and Signification, Iconicity in Language and Literature* (pp. 39–54). Amsterdam: John Benjamins.
- Doke, C. M., & Vilakazi, B. W. (1953). *Zulu-English Dictionary* (2d ed.). Johannesburg: Witwatersrand University Press.
- Guest, S., Dessirier, J. M., Mehrabyan, A., McGlone, F., Essick, G., Gescheider, G., Fontana, A., et al. (2010). The development and validation of sensory and emotional scales of touch perception. *Attention, Perception, & Psychophysics*, 73(2), 531–550. doi:10.3758/s13414-010-0037-y
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The Weirdest People in the World? *Behavioral and Brain Sciences*, 33(2-3), 61–83. doi:10.1017/S0140525X0999152X
- Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, 109(1), 54–65. doi:10.1016/j.cognition.2008.07.015
- Takehi, H., Tamori, I., & Schourup, L. (1996). *Dictionary of Iconic Expressions in Japanese*. Berlin/New York: Mouton.
- Kita, S. (1997). Two-dimensional semantic analysis of Japanese mimetics. *Linguistics*, 35, 379–415.
- Klatzky, R. L., Lederman, S. J., & Reed, C. (1987). There's more to touch than meets the eye: The salience of object attributes for haptics with and without vision. *Journal of Experimental Psychology: General*, 116, 356–369. doi:10.1037/0096-3445.116.4.356
- Kunene, D. P. (1965). The ideophone in Southern Sotho. *Journal of African Languages*, 4, 19–39.
- Majid, A., & Levinson, S. C. (2007). Language of perception: overview of field tasks. In A. Majid (Ed.), *Field Manual Volume 10* (pp. 8–9). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from <http://fieldmanuals.mpi.nl/volumes/2007/language-of-perception-overview/>
- Majid, A., & Levinson, S. C. (2010). WEIRD languages have misled us too. *Behavioral and Brain Sciences*.
- Majid, A., & Levinson, S. C. (2011). The Senses in Language and Culture. *The Senses and Society*, 6(1), 5–18.
- Miller, G. A. (1969). A psychological method to investigate verbal concepts. *Journal of Mathematical Psychology*, 6(2), 169–191. doi:10.1016/0022-2496(69)90001-7
- Moller, A. R. (2002). *Sensory systems: anatomy and physiology*. San Diego, Calif.; London: Academic.
- Nuckolls, J. B. (1996). *Sounds Like Life: Sound-Symbolic Grammar, Performance, and Cognition in Pastaza Quechua*. New York: Oxford University Press.
- Nuckolls, J. B. (2004). To be or to be not ideophonically impoverished. In Wai Fong Chiang, E. Chun, L. Mahalingappa, & S. Mehus (Eds.), *SALSA XI: Proceedings of the Eleventh Annual Symposium about Language and Society*, Texas Linguistic Forum (pp. 131–142). Austin.
- Osaka, N. (1990). Multidimensional analysis of onomatopoeia: A note to make sensory scale from words. *Studia phonologica*, 24, 25–33.
- Osaka, N., & Osaka, M. (2005). Striatal reward areas activated by implicit laughter induced by mimic words in humans: a functional magnetic resonance imaging study. *Neuroreport*, 16(15), 1621–1624.
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The Measurement of Meaning*. Urbana: University of Illinois Press.
- Perchonock, N., & Werner, O. (1969). Navaho systems of classification: some implications for ethnoscience. *Ethnology*, 8(3), 229–242.
- Tufvesson, S. (2011). Analogy-making in the Semai Sensory World. *The Senses and Society*, 6(1), 86–95. doi:10.2752/174589311X12893982233876
- Voeltz, F. K. E., & Kilian-Hatz, C. (Eds.). (2001). *Ideophones*. Typological Studies in Language. Amsterdam: John Benjamins.
- Weller, S. C., & Romney, A. K. (1988). *Systematic data collection*. Newbury Park, CA: Sage Publications.
- Westermann, D. H. (1927). Laut, Ton und Sinn in westafrikanischen Sudansprachen. In F. Boas (Ed.), *Festschrift Meinhof* (pp. 315–328). Hamburg: L. Friederichsen.
- Yoshida, M. (1968). Dimensions of tactual impressions (1). *Japanese Psychological Research*, 10(4), 123–137.

# The Sound of Thickness: Prelinguistic Infants' Associations of Space and Pitch

Sarah Dolscheid<sup>1,2</sup>    Sabine Hunnius<sup>3</sup>    Daniel Casasanto<sup>4</sup>    Asifa Majid<sup>1,3</sup>  
(sarah.dolscheid@mpi.nl) (s.hunnius@donders.ru.nl) (casasand@newschool.edu) (asifa.majid@mpi.nl)

<sup>1</sup>Max Planck Institute for Psycholinguistics, Nijmegen, NL

<sup>2</sup>International Max Planck Research School for Language Sciences, Nijmegen, NL

<sup>3</sup>Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, NL

<sup>4</sup>Department of Psychology, The New School for Social Research, New York, USA

## Abstract

People often talk about musical pitch in terms of spatial metaphors. In English, for instance, pitches can be *high* or *low*, whereas in other languages pitches are described as *thick* or *thin*. According to psychophysical studies, metaphors in language can also shape people's nonlinguistic space-pitch representations. But does language establish mappings between space and pitch in the first place or does it modify preexisting associations? Here we tested 4-month-old Dutch infants' sensitivity to height-pitch and thickness-pitch mappings in two preferential looking tasks. Dutch infants looked significantly longer at cross-modally congruent stimuli in both experiments, indicating that infants are sensitive to space-pitch associations prior to language. This early presence of space-pitch mappings suggests that these associations do not originate from language. Rather, language may build upon pre-existing mappings and change them gradually via some form of competitive associative learning.

**Keywords:** cross-modal; multisensory; metaphors; synaesthesia; infant perception; language acquisition; language of perception; preferential looking

## Introduction

Does a cake taste yellow? Or a tone played by a trumpet sound scarlet? For some people they do. Yet synaesthesia, a condition in which stimulation of one sensory modality induces systematic perceptual experiences in another modality, is relatively rare. Other types of cross-modal associations, however, can be found in non-synaesthetes too. Psychophysical studies have shown that adults and children without synaesthesia associate higher pitches with sharper edges (Marks, 1987; Parise & Spence, 2009), positions higher in space (e.g., Ben-Artzi & Marks, 1995; Melara & O'Brien, 1987), lighter color (Hubbard 1996; Marks 1989; Melara 1989), and increasing brightness (Marks, 1987).

Even infants seem to be sensitive to some of these associations. Cross-modal correspondences between loudness and brightness have been demonstrated in 20- to 30-day-old infants (Lewkowicz & Turkewitz, 1980). In a preferential looking task, 3- to 4-month-old infants preferred congruent trials – in which visuospatial height and pitch height corresponded – over incongruent trials. (Walker et al., 2010); that is, infants looked longer at a ball moving upwards if it was accompanied by a rising pitch than if it was accompanied by a falling pitch. Also, prelinguistic infants under 1-year-old 'matched' visual arrows pointing

up or down with tones sweeping up or down in frequency (Wagner et al., 1981).

These findings have led to the assumption that cross-modal mappings are innately hardwired in the brain (Mondloch & Maurer, 2004) and represent an unlearned aspect of perception (Walker et al., 2010). Accordingly, some of these associations are posited to be universal (Marks, Hammeal, & Bornstein, 1987; see also Spence, 2011).

However, there are other findings that seem to be at odds with these conclusions. A number of cross-modal correspondences are only acquired later in the course of development. For instance, even 9-year-old children are not able to systematically match size and pitch, a task that is consistently solved by adults (Marks et al., 1987; but see Mondloch & Maurer, 2004). Cross-modal correspondences are also affected by developmental changes. Unlike adults and older children, 2-year-olds consistently map light grey to smaller objects and dark grey to bigger objects (Smith & Sera, 1992). In the course of language acquisition, however, children's associations gradually shift. As a result, it has been suggested that language may have an impact on the trajectory of cross-modal relations (Smith & Sera, 1992).

On one hand, language appears to mirror cross-modal experience. The auditory domain, for instance, is often linguistically encoded in terms of other sensory modalities (Williams, 1976). People use metaphors like "soft" voice, "dark" timbre or "high" pitch, suggesting that language echoes cross-modal perceptual impressions. On the other hand, language also seems to affect cross-modal associations. For example, Martino & Marks (1999) suggest that cross-modal effects, like the association between space and pitch, may be mediated by language. Various tasks show correspondences between spatial height and pitch consistent with high-low metaphors in language (Rusconi, et al, 2006, Evans & Treisman, 2010). However, since linguistic labels and height-pitch associations merely coincide, the direction of influence is hard to establish and the contribution of language remains unclear.

Cross-linguistic comparison provides one way to overcome this limitation. Not every language uses the same metaphors for pitch (Ashley, 2004; Levinson & Majid, 2007). For example, while languages like English and Dutch talk about pitch in terms of "height", other languages like, Farsi, Turkish and Zapotec (spoken in Mexico) describe

high-frequency pitch as “thin” and low-frequency pitches as “thick” (Shayan, Ozturk, & Sicoli, 2011). To find out whether these differences in spoken metaphors correspond to different mental representations of pitch, Dolscheid and colleagues, conducted a series of nonlinguistic psychophysical experiments in adult speakers of Dutch (a “height” language) and Farsi (a “thickness” language). Participants were asked to reproduce musical pitches that they heard in the presence of irrelevant spatial information, i.e. lines varying either in height or in thickness (Dolscheid et al., submitted). Dutch speakers’ pitch estimates were significantly modulated by spatial height but not by thickness. Conversely, Farsi speakers’ pitch estimates were modulated by spatial thickness but not by height. Overall, the results indicated that nonlinguistic pitch-space associations follow language-specific vocabulary, suggesting that cross-modal pitch representations are language-specific.

At this point, however, it is unclear whether language establishes cross-modal mappings between space and pitch in the first place, or whether it merely modifies preexisting associations. While some researchers stress the relevance of language in concept formation (e.g. Gopnik & Meltzoff, 1997), others argue that conceptual representations must precede the acquisition of language (e.g. Bloom, 2000; Bloom & Keil, 2001). The former position allows for a stronger role of language in space-pitch associations; while the latter position suggests that children are likely to have some notion of space-pitch correspondences prior to learning language. Consistent with this latter view, infants seem to be sensitive to height-pitch mappings even prelinguistically (Walker et al., 2010). Critically, however, nothing is known about the origins of thickness-pitch relationships. Do children also have this mapping available to them prior to language, or is it only learned on the basis of language-input?

One possibility is that children could start out with both a height-pitch and a thickness-pitch mapping even before they learn language. The strength of these mappings might then subsequently be adjusted, according to the relative frequencies of space-pitch metaphors in the languages children acquire (Casasanto 2008, 2010; Dolscheid et al., submitted; Smith & Sera, 1992). Alternatively, height-pitch and thickness-pitch associations might follow different trajectories. Whereas height-pitch associations are available to prelinguistic infants, the thickness-pitch mapping might only be learned later. Metaphors in language could provide one possible way to learn this association. Using thickness terminology to refer to pitch may invite speakers to align correspondent representations and extract similarities between space and pitch in a process called structural alignment (see e.g. Boroditsky, 2001; Gentner, 2003). In line with this proposal, Shayan et al. (submitted) found that Turkish and Farsi 2- to 5-year-olds were able to successfully map thickness to pitch but same-aged German children (who like English and Dutch speakers do not have a thickness metaphor) were not able to make this association

successfully. This is consistent with the proposal that language input promotes cross-modal associations between thickness and pitch. Note, however, that these results do not rule out the possibility that the thickness-pitch mappings were available to all infants prelinguistically, but are no longer equally available to German children.

In order to determine the prelinguistic availability of space-pitch mappings, we tested 4-month-old Dutch babies using a preferential looking paradigm. To investigate height-pitch correspondences, we followed Walker et al.’s (2010) procedure. Infants watched a ball moving up and down the screen accompanied by the sound of a sliding whistle. The whistle’s fundamental frequency changed at a constant rate. In the congruent condition, the pitch of the sound “rose” and “fell” in accordance with the movement of the ball. In the incongruent condition, the pitch of the sound “rose” and “fell” in opposition to the movement of the ball (see Fig. 1a). Walker et al. (2010) reported that infants looked longer at the congruent compared to the incongruent condition, suggesting an early preference for pitch-height congruencies.

In a second step, we tested prelinguistic infants in a thickness-pitch task analogous to the height-pitch task. Instead of balls moving up and down the screen, a vertical tube varied in thickness, changing continuously from thin to thick (see Fig. 1b).

We reasoned that if both, height-pitch and thickness-pitch mappings are available to infants prelinguistically, infants should prefer both congruent height-pitch and congruent thickness-pitch stimuli over incongruent ones. If however, height-pitch and thickness-pitch relationships follow different developmental trajectories, with thickness mappings only becoming acquired later, then infants should show preferences for congruent height-pitch stimuli but not for congruent thickness-pitch stimuli.

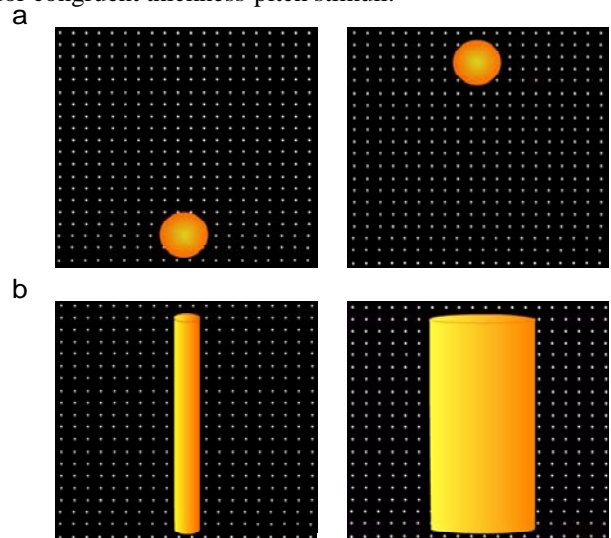


Figure 1: Examples of animations presented as stimuli in Experiment 1 (Panel a) and Experiment 2 (Panel b). In (a) the extremes of the ball’s vertical trajectory are shown. In (b) the extremes of thickness are depicted. The images are reproduced to scale.



## Experiment 1: Auditory pitch and visuospatial height

### Methods

**Participants** Ten male and ten female infants completed the first (pitch-height) experiment (mean age = 129 days, range: 113 to 138 days). Another seven infants were tested, but not included in the analyses: one infant was excluded due to experimenter error; a further six infants were excluded due to excessive fussiness.

**Materials and Procedure** QuickTime animations were presented on a 102 x 76 cm Sony LCD screen using HABIT software. Animations appeared within a 67 x 67 cm screen area ( $25.6^\circ \times 25.6^\circ$ ), and lasted a maximum of 60 s. Before each animation, a flashing light appeared to ensure that infants attended to the screen. Infants sat in a Maxi-Cosi infant seat which was placed on their parent's lap, viewing the animations from a distance of 1.50 m. Infants' visual fixations were monitored and recorded on video. Animations were stopped if the infant looked elsewhere for a single period of 1 s or more. The total duration an infant looked directly at the animation was logged online during the experiment using HABIT software and written to an output-file. Additionally, looking times were determined by a subsequent frame-by-frame coding of the digitized video using SuperCoder. Coding was performed by a coder blind to the experimental condition. 25 percent of the data was double-coded by a second person, also blind to the condition.

Infants watched a 10-cm ( $4^\circ$ ) diameter orange ball moving up and down a 50-cm vertical trajectory in front of a 20 x 20 grid of small, white dots on a black field. The ball moved at a constant speed of 20 cm/s and paused for 42 ms at each endpoint. Animations were accompanied by the sound of a sliding whistle (a sinusoidal tone). The fundamental frequency of the sound changed at a constant rate, between 300 and 1700 Hz over 2.5 s, coinciding with a single phase of the animation (e.g. the ball moving up). The amplitude of the sound increased and then decreased between 47 and 84 dB within each phase of the animation, peaking at 1000 Hz. Amplitude thus changed about twice as fast as pitch to ensure that variation in perceived pitch and loudness were not confounded.

Every infant viewed three congruent animations interleaved with three incongruent animations. Half of the children watched a congruent animation first and the other half watched an incongruent animation first. During the whole experiment, parents were listening to music via headphones. Since parents could not hear the sliding sounds, they were unable to distinguish between experimental conditions (the spatial trajectories of the stimuli did not differ between conditions). We therefore ensured parents could not bias their infant's looking behavior.

### Results

A high level of agreement was confirmed between the two observers in their coding of each infant's individual looking times (mean Pearson's  $r(28) = .99$ ,  $p < .001$ ).

14 of the 20 infants looked longer at the congruent animation than at the incongruent animation. On average, infants looked at the congruent animations for 31.7 s ( $SD = 11.4$ ) and at the incongruent animations for 26.1 s ( $SD = 13.3$ ). A paired-samples t-test confirmed that infants looked significantly longer at the congruent animations,  $t(19) = 1.99$ ,  $p = .03$ ,  $d = 0.45$  (one-tailed).

## Experiment 2: Auditory pitch and visuospatial thickness

### Methods

**Participants** Ten male and ten female infants completed the second (pitch-thickness) experiment (mean age = 127 days, range: 113 to 138 days). An additional eight infants were tested, but their data was not analyzed: one infant was excluded due to technical problems; a further 7 infants were excluded due to fussiness.

**Materials and Procedure** The same procedure as in Experiment 1 was used. This time, infants watched a vertical orange tube that varied in thickness, changing continuously from thin to thick (see Figure 1).

The animation was presented on a 20 x 20 grid of small, white dots on a black field, as in Experiment 1. The tube was 60 cm long ranging from 6 to 26 cm in width. It expanded at a constant speed of 8 cm/s and paused for 42 ms at each endpoint. Animations were accompanied by the sound of the exact same sliding whistle as in Experiment 1. The fundamental frequency of the sound changed at a constant rate, between 300 and 1700 Hz over 2.5 s, coinciding with a single phase of the animation (i.e. during tube expansion). Each infant viewed three congruent animations interleaved with three incongruent animations, with half of the children watching a congruent animation first and the other half watching an incongruent animation first. During the whole experiment, parents were listening to music via headphones.

### Results

A high level of agreement was confirmed between the two judges in their estimates of each infant's individual looking times (mean Pearson's  $r(28) = .99$ ,  $p < .001$ ).

13 of the 20 infants looked longer at the congruent animations than the incongruent animations. On average, infants looked longer at the congruent 24.4 s ( $SD = 11.8$ ) than the incongruent animations 19.4 s ( $SD = 11.5$ ). A paired-samples t-test confirmed that infants looked significantly longer at the congruent animations,  $t(19) = 2.19$ ,  $p = .02$ ,  $d = 0.43$  (one-tailed).

## Between experiment comparison

Dutch infants looked significantly longer at cross-modally congruent stimuli in both experiments. While this suggests a comparable starting point for both thickness-pitch and height-pitch mappings, it is nevertheless possible that infants display differential preference with respect to the two mappings. We therefore compared the results of the two previous experiments directly.

## Results

Submitting looking times to a 2 (Space: height vs. thickness) by 2 (Congruency: congruent vs. incongruent) mixed ANOVA yielded a significant main effect of Space ( $F(2,38) = 4.40, p = .04, \eta_p^2 = 0.10$ ), showing that looking times differed between height and thickness stimuli. Infants looked longer at height stimuli as compared to thickness stimuli, indicating that perhaps height was more salient for them. However, no interaction between Space and Congruency ( $F(2,38) = 0.03, ns, \eta_p^2 = 0.00$ ) was observed. There were thus no indications that looking time reductions induced by incongruency differed between the two experiments. In line with this, percentage reduction in looking time across experiments was of comparable size, i.e., 18% for the height-pitch experiment and 20% for the thickness-pitch experiment.

## General Discussion

Our results demonstrate that prelinguistic infants are sensitive to correspondences between auditory pitch and spatial information of two different types, visuospatial thickness as well as height. Dutch infants looked significantly longer at cross-modally congruent stimuli in both experiments, suggesting a comparable starting point for height-pitch mappings and thickness-pitch mappings.

It is possible that these mappings are only present in very young infants but get lost in the course of development due to neuronal pruning. Whereas 2-to 3-month-old infants were found to be sensitive to arbitrary associations between colors and shapes, 8-month-old infants no longer show this early synaesthetic association (Wagner & Dobkins, 2011). Does the same developmental trajectory hold true for space-pitch mappings? Unlike synaesthetic color-shape associations that seem highly individualized and thus unspecific (e.g., one infant might associate triangles with green, and another with red), space-pitch associations follow a specific pattern, showing the same congruity preferences found in languages. It is therefore possible that space-pitch mappings persist during infancy and childhood. In line with this suggestion, sensitivity to height-pitch associations has been reported in 6-month-olds (Braaten, 1993) as well as in children aged 4 to 5 years (Roffler & Butler, 1967). On the other hand, 2- to 5-year-old German speaking children have been found to be insensitive to the thickness mapping (Shayan et al., submitted). There is also contradictory evidence regarding children's sensitivity to size-pitch associations. Whereas Marks et al. (1987) report that

children are unable to systematically map size (big vs. small) to pitch until they are 13 years old, Mondloch and Maurer (2004) find evidence for size-pitch congruency effects in children as young as 3 years of age.<sup>1</sup> Details about the developmental trajectory of space-pitch mappings thus remain unclear and are subject to future research.

One aspect that seems to facilitate detecting cross-modal associations is motion (see also Jeschonek, Pauen & Babocsai, in preparation). Mondloch and Maurer presented children with moving balls that differed in size; while Marks et al. and Shayan et al. used static stimuli. In the present study, too, the dynamic display of spatial information (up- and downward movement or horizontal expansion) and pitch (presented as glides) may have facilitated the detection of corresponding information. Displaying stimuli dynamically and in synchronicity could direct infants' attention to the relational correspondences across modalities. However, movement by itself cannot explain the pattern of results: infants must still align stimuli attributes that are congruent to each other.

## Language acquisition and cross-modal associations

Our findings demonstrate that both height-pitch and thickness-pitch correspondences are perceived before the infant has mastery of language. While this finding is consistent with the view that representations precede language (e.g. Bloom, 2000), it does not entail that these associations are fixed. Language could still influence the structure and content of preexisting mental representations via simple learning mechanisms. In the course of language acquisition, the relative strengths of different space-pitch mappings could be adjusted according to the language-specific frequencies of metaphors that children acquire (Casasanto 2008, 2010). Over time, speakers of a "height" language like Dutch would strengthen the height-pitch mapping at the expense of the thickness-pitch mapping – and vice versa for speakers of a "thickness" language like Farsi (Dolscheid et al., submitted). Evidence in support of this associative learning account is provided by linguistic training experiments. Dutch speakers, after being trained to use Farsi-like metaphors describing pitch relationships in terms of thickness, demonstrated nonlinguistic thickness-pitch mappings just like Farsi speakers. By contrast, when participants received the same amount of linguistic training with an alternative space-pitch mapping that is not present in any known language, they showed no effect of training (Dolscheid et al., submitted). These training studies demonstrate a causal role for language in strengthening the use of some nonlinguistic mappings more than others.

---

<sup>1</sup> The thickness-stimuli used in Experiment 2 could also be interpreted as a size manipulation. Indeed, even though movement was restricted to the horizontal plane which is characteristic for thickness, there is a concomitant difference in overall size. For the present purposes, nothing rests on being able to make the distinction between thickness and size, per se, since the reported inconsistency in the ability to make the cross-modal mapping to pitch applies equally to both spatial parameters.



While language may enforce particular pitch-space mappings, this proposal has to take into account that metaphors pose additional demands in language acquisition. Pitch metaphors are inherently polysemous; the acquisition of both spatial and sound meanings is likely more complex than when a single meaning has to be acquired (see e.g. Johnson, 1992). Consistent with this proposal French speaking children trained to describe sounds using either the single-meaning terms *aigu* and *grave* (a pair of antonyms used only to label high and low pitches) versus the polysemous words *haut* and *bas* (which are used to refer to pitch and space) were better able to label sound stimuli (Costa-Giomi & Descombes, 1996).

Aside from polysemy, another important attribute of metaphorical language lies in its directionality. Taking spatial metaphors of time as an example, people talk about time in terms of space far more often (“a long vacation”; “a short meeting”) than they talk about space in terms of time (though it occasionally occurs: “I live two minutes from here”) (Casasanto, 2008, 2010; Lakoff & Johnson, 1980). For pitch metaphors the same asymmetry seems to hold, which has also been found to be reflected in adults’ nonlinguistic pitch representations (Casasanto, 2010). Note, however, that our results are agnostic of a space-pitch asymmetry in prelinguistic infants. While we have demonstrated that infants are able to detect space-pitch associations, our tasks do not speak to possible directionality. Future studies are necessary to determine whether language plays a role in introducing this asymmetry (e.g. see Meritt, Casasanto, & Brannon, 2010), or whether it is present independent of language (e.g. see Marks et al., 1987).

### Effects of cross-modal associations on language

Since cross-modal associations are present before children acquire language, it is possible that the associations themselves shape metaphors in language. We find the height-pitch metaphor in languages such as Spanish, German and Polish (Rusconi et al., 2006), as well as non-Indo-European languages, like Japanese and Chinese. In all of these languages, “high” refers to high frequency sounds and “low” to low frequency sounds, but not the reverse. Likewise, psychophysical studies demonstrated that participants associate higher pitches with smaller objects, not with larger objects (e.g. Gallace & Spence, 2006). For the Kpelle and Jabo people in Liberia, this association is also encoded in language: “small” refers to high pitch and “big” refers to low pitch (see e.g., Eitan & Timmers, 2010). Prelinguistic associations, alongside correlations of properties in the real world, may thus serve as guiding principles that constrain the way pitch gets lexicalized, across languages.

Consequently, it might be harder to learn linguistic metaphors that are inconsistent with cross-modal mappings for which there is evidence in the natural world. The results of a training study support this suggestion. Dutch speakers trained to use reversed thickness-pitch mappings

(thick=high, thin=low) were not able to master this association, even though they could learn the comparable congruent mapping (Dolscheid et al., submitted). It thus appears that language cannot easily retrain mappings that are supported by correlations present prelinguistically and/or supported by real world experience. Early sensitivity to certain mappings might therefore constrain the set of cross-modal associations that are likely to be observed in language and the mind.

### Origins of cross-modal mappings?

Are cross-modal mappings innate? Based on the current evidence, we can only conclude that cross-modal associations between space and pitch are present from very early. By the age of 4 months, however, infants may well have encountered enough relevant co-occurrences in their interaction with the world to have learned these mappings. Thickness-pitch mappings seem especially prevalent: thicker strings produce lower notes, bigger bells have lower chimes, and people with bigger (‘thicker’) bodies tend to have lower voices. While infants may have internalized these regularities, the case for innate height-pitch mappings is not conclusive (see also Walker et al., 2010).

### Conclusions

No matter whether cross-modal associations are inborn or learned, the finding that both height-pitch and thickness-pitch mappings can be observed in infants as young as 4 months of age constrains theorizing about the role that language plays in shaping nonlinguistic mental representations of pitch.

Our data show that space-pitch associations are present prior to language, suggesting that language is unlikely to create cross-modal mappings between space and pitch, even if language seems to play this role in other domains (Gentner, 2002).

It appears that both the height-pitch mapping found in languages like Dutch and the thickness-pitch mapping found in languages like Farsi are already present in prelinguistic infants’ minds. This suggests that people who use different spatial metaphors for pitch in their native languages come to think about pitch differently not because language instills in them one cross-modal mapping instead of the other, but rather because language strengthens one pre-existing mapping at the expense the other, via some form of competitive associative learning (Casasanto, 2008, 2010; Dolscheid, et al., submitted). The precise learning mechanisms that give rise to cross-linguistic differences in pitch representation, and the underlying neural mechanisms, remain topics for future research.

### Acknowledgments

We would like to thank Sho Tsuji, Manu Schuetze, Angela Khadar, Margret van Beuningen, Nienke Dijkstra, Laura Arendsen, Dirkje van der Aa & Webb Phillips.

## References

- Ashley, R. (2004). Musical pitch space across modalities: Spatial and other mappings through language and culture. In P. W. S. Lipscomb, R. Ashley, R. Gjerdingen (Eds.), *Proceedings of the 8th International Conference on Music Perception and Cognition*. Adelaide: Causal Productions.
- Ben-Artzi, E., & Marks, L. E. (1995). Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics*, 57, 1151-1162.
- Bloom, P. (2000). How children learn the meanings of words. Cambridge, MA: MIT Press.
- Bloom, P., & Keil, F. (2001). Thinking through language. *Mind and Language*, 16, 351-367.
- Braaten, R. (1993). *Synesthetic correspondence between visual location and auditory pitch in infants*. Paper presented at the 34th Annual Meeting of the Psychonomic Society.
- Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, 43(1), 1-22.
- Casasanto, D. (2008). Who is afraid of the big bad Whorf? Cross-linguistic differences in temporal language and thought. *Language Learning*, 58(1), 63-79.
- Casasanto, D. (2010). Space for Thinking. In V. Evans & P. Chilton (Eds.), *Language, Cognition and Space: The State of the Art and New Directions* (pp. 453-478). London: Equinox Publishing.
- Costa-Giomi, E., & Descombes, V., (1996). Pitch labels with single and multiple meanings: A study with French-speaking children. *Journal of Research in Music Education*, 44(3), 204-214.
- Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (submitted). The Thickness of Musical Pitch: Psychophysical evidence for linguistic relativity.
- Eitan, Z., & Timmers, R. (2010). Beethoven's last piano sonata and those who follow crocodiles: cross-domain mappings of auditory pitch in a musical context. *Cognition*, 114(3), 405-22.
- Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, 10(1), 1-12.
- Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, 68, 1191-1203.
- Gentner, D. (2002). Analogy in scientific discovery: The case of Johannes Kepler. In L. Magnani & N. J. Nersessian (Eds.), *Model-based reasoning: Science, technology, values* (pp.21-39). New York: Kluwer Academic/ Plenum Publisher.
- Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp.195-235). Cambridge, MA: MIT Press.
- Gopnik, A., & Meltzoff, A. N. (1997). Words, Thoughts, and Theories. Cambridge, MA: MIT Press.
- Hubbard, T. L. (1996). Synesthesia-like mappings of lightness, pitch, and melodic interval. *The American Journal of Psychology*, 109, 219-238.
- Jeschonek, S., Pauen, S., & Babocsai, L. (in preparation). Cross-modal mapping of visual and acoustic displays in infants: The effect of static and dynamic components.
- Johnson, C. J. (1992). Cognitive components of naming in children: Effects of referential uncertainty and stimulus realism. *Journal of Experimental Child Psychology*, 53, 45-71.
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago and London: The University of Chicago Press.
- Levinson, S. C., & Majid, A. (2007). The language of sound. In A. Majid (Ed.), *Field Manual Volume 10* (pp. 29-31). Nijmegen: Max Planck Institute for Psycholinguistics.
- Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory-visual intensity matching. *Developmental Psychology*, 16, 597-607.
- Marks, L. E. (1987). On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology and Human Perception Performance*, 13, 384-394.
- Marks, L. E. (1989). On cross-modal similarity: The perceptual structure of pitch, loudness, and brightness. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 586-602.
- Marks, L. E., Hammeal, R. J., & Bornstein, M. H. (1987). Perceiving similarity and comprehending metaphor. *Monographs of the Society for Research in Child Development*, 52(1), 1-102.
- Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception*, 28, 903-923.
- Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General*, 116, 323-336.
- Melara, R. D. (1989). Dimensional interaction between color and pitch. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 69-79.
- Merritt, D.J., Casasanto, D., & Brannon, E. M. (2010). Do Monkeys Think in Metaphors? Representations of Space and Time in Monkeys and Humans. *Cognition*, 117, 191-202.
- Mondloch, C. J., & Maurer, D. (2004). Do small white balls squeak? Pitch-object correspondences in your children. *Cognitive, Affective & Behavioral Neuroscience*, 4, 133-136.
- Parise, C., & Spence, C. (2009). "When birds of a feather flock together": Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE*, 4, e5664.
- Rusconi, E., Kwan, B., Giordano, B. L., Umiltà, C., & Butterworth, B. (2006). Spatial representation of pitch height: the SMARC effect. *Cognition*, 99(2), 113-29.
- Shayan, S., Ozturk O., & Sicoli, M. (2011). Thickness of pitch, cross-modal metaphors in Farsi, Turkish and Zapotec, *The Senses and Society*.
- Shayan, S., Ozturk, O., Bowerman, M., & Majid, A. (submitted). Spatial metaphor in language can promote the development of cross-modal mappings in children.
- Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology*, 24, 99-142.
- Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception & Psychophysics*, 73, 971-995.
- Roffler, S. K., & Butler, R. A. (1968). Factors that influence the localization of sound in the vertical plane. *The Journal of the Acoustical Society of America*, 43, 1255-1259.
- Wagner, K., & Dobkins, K. R. (2011). Synaesthetic Associations Decrease During Infancy. *Psychological Science*, 22, 1067-1072.
- Wagner, S., Winner, E., Cicchetti, D., & Gardner, H. (1981). "Metaphorical" Mapping in Human Infants. *Child Development*, 52(2), 728.
- Walker, P., Bremmner, J.G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, 21(1), 21-25.
- Williams, J. M. (1976). Synaesthetic adjectives. A possible law of semantic change. *Language*, 52, 461-478.

# What explains variability in brain regions associated with Theory of Mind in a large sample of neurotypical adults and adults with ASD?

Nicholas Dufour<sup>1</sup>, Elizabeth Redcay<sup>2</sup>, Liane Young<sup>3</sup>, Penelope L Mavros<sup>1</sup>, M Joseph Moran<sup>4</sup>, Christina Triantafyllou<sup>1,5</sup>, John Gabrieli<sup>1,5</sup>, and Rebecca Saxe<sup>1,5</sup>

1. Department of Brain and Cognitive Sciences, MIT
2. Department of Psychology, University of Maryland
3. Department of Psychology, Boston College
4. Psychology Department, Harvard University
5. McGovern Institute for Brain Research, MIT

## Abstract

Theory of mind ('ToM') tasks elicit highly reliable neural activity across individuals and experimental paradigms. We compared activity in a very large sample of neurotypical ('NT', N=477) individuals, and a group of high functioning individuals with autism spectrum disorders ('ASD', n=27), using both region of interest ('ROI') and whole-brain analyses. Although ToM activity showed significant and reliable individual differences, these differences were not explained by participant gender or age, or most experimental parameters. Furthermore, there were no differences between ASD and NT individuals. These results imply that the social cognitive impairments typical of ASD can occur without gross changes in the size or response magnitude of ToM brain regions.

**Keywords:** Theory of mind; ASD; fMRI; TPJ; PC; precuneus; MPFC; DMPFC; MMPFC

## Introduction

Theory of Mind ('ToM') is the capacity to represent the mental states of others (Premack & Woodruff, 1978). Individuals with autism spectrum disorders (ASD) appear to have particular difficulty with aspects of ToM. In particular, children with ASD are disproportionately delayed on tasks that tap inferences about other people's beliefs (Baron-Cohen, 1989). The neural mechanism of this deficit remains unknown. However, in neurotypical (NT) adults and children, fMRI studies reveal a remarkable reliable group of brain regions recruited during ToM tasks. These regions include the left and right temporo-parietal junction (RTPJ and LTPJ), right anterior superior temporal sulcus (rSTS), the medial precuneus (PC), and the medial prefrontal cortex (MPFC) (U. Frith & Frith, 2003). Thus, a tempting hypothesis is that dysfunction of the brain regions typically implicated in ToM is responsible for the social cognitive impairments observed in ASD.

Previous attempts to characterize the function of ToM brain regions in adults with ASD have yielded conflicting results. Some studies suggest that ToM regions are hypoactive (i.e., produce a smaller or less selective response, (Kennedy & Courchesne, 2008; Lombardo, Chakrabarti, Bullmore), while other studies find no difference between ASD and NT individuals (Gilbert, Bird, Brindley, Frith, & Burgess, 2008), and still others find the

opposite pattern, hyperactivation, in ASD (Dichter, Felder, & Bodfish, 2009).

One explanation of these conflicting results may be that sample sizes are small, and individual variability is large. Small samples of individuals with ASD are problematic because individuals with ASD may be highly heterogeneous in their neural responses (e.g., Hasson et al., 2009). Small samples of NT participants are equally problematic, because they allow for calculation of only the mean response, not the distribution. Understanding the distribution is critical if neural measures are to be useful in a clinical setting. For most clinical applications, it is more important to be able to describe the neural activity pattern of each specific individual, relative to typical and atypical distributions. For example, using fMRI to help diagnose ASD would require comparing each individual to the typical distribution.

In order to measure the distribution of responses in ToM brain regions of NT participants, we aggregated data collected over 5 years from 477 NT participants. This large sample allowed us to investigate variability in ToM region responses, and measure any difference between NT participants and adults with ASD, with unusually high sensitivity. The main goal of the current paper is therefore to compare the response in these regions in a large sample of NT participants and a moderate sample of high-functioning adults with ASD. In order to do so, we also (i) identify and remove variance in the measured response, associated with basic experimental parameters such as the stimulus modality, number of stimuli, or experimental task, and (ii) test whether the response of ToM regions is related to basic demographic factors that may be relevant for ASD, including gender, age, and IQ.

## Methods

**Typical Participants:** Data were analyzed from 477 NT participants (M=25.2 years, range: 18-69 years; 179 male). IQ was measured in 60 of these participants (IQ 84-141, M=117.5, SD=12.4). Participants provided informed consent, in accordance with the guidelines of the MIT Committee on the Use of Human Experimental Subjects (COUHES), and were compensated approximately \$30 per hour for their time.

**ASD Participants:** 27 individuals with a clinical diagnosis of ASD (M=33.9yrs, range 18-66yrs; 20 male) were

included, having volunteered to participate in one of two (Moran et al., 2011; Redcay et al., 2012) previous studies. The Autism Diagnostic Observation Schedule (ADOS) was administered to 23 of the 27 ASD participants (ADOS communication score  $M=3.2$ ,  $SD=1.3$ ; ADOS social score  $M=5.8$ ,  $SD=1.8$ ). For 24 of the ASD participants, IQ measures were obtained by the Kaufman Brief Intelligence Test (IQ 69-141,  $M=116.3$ ,  $SD=16.8$ ). For direct NT vs. ASD comparison, a set of 24 NT participants (collectively termed ‘matched’) were chosen based on pairwise similarity with the ASD participants on IQ, age, and gender (age 20-54,  $M=29.9$  years,  $SD=8.8$  years; IQ 84-141,  $M=116.3$ ,  $SD=14.5$ ; 19 male).

**fMRI Tasks** All participants were presented with verbal narratives in English that described a character and his/her mental states (Mental condition) or described physical objects and events (Control condition). The stimuli were presented either visually as text on a screen, or aurally through headphones. After reading or hearing the narrative, participants performed one of 4 tasks. These tasks correspond to the functional localizers used in (Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2010; Kliemann, Young, Scholz, & Saxe, 2008; L. Young & Saxe, 2008; L. Young & Saxe, 2009; L. Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010; L. Young, Nichols, & Saxe, 2010; L. Young, Scholz, & Saxe, 2011) and unpublished data.

**fMRI Methods:** Participants were scanned on a 3T Siemens scanner at the Martinos Imaging Center at the McGovern Institute for Brain Research at the Massachusetts Institute of Technology ( $n=468$ ) or at the Massachusetts General Hospital ( $n=36$ ). NT participants were scanned between 2006 and 2011. ASD participants were scanned between 2007 and 2010. Matched NTs were scanned between 2007 and 2010. Functional data were acquired using single echo gradient echo echo-planar-imaging with voxel size  $3.125 \times 3.125 \times 4.000$  mm ( $TE=30$  ms, flip angle= $90^\circ$ , TR either 2.5 ( $n=36$ ) or 2 secs ( $n=468$ )). Participants were scanned on either a 12-channel or a 32-channel receive coil, both Siemens products. Data were analyzed using SPM2 or SPM8 (<http://www.fil.ion.ucl.ac.uk>) and in-house software. The data were realigned to account for motion, smoothed with a 5 mm Gaussian kernel and normalized to a standard template in Montreal Neurological Institute space.

**ROI Analyses:** Six functional ROIs (ROIs) from the ToM network were defined in individual participants, using the contrast Mental>Control, consistent with previous literature (e.g. (U. Frith & Frith, 2003; Saxe & Kanwisher, 2003)): RTPJ, LTPJ, PC, dorsal and middle MPFC (DMPFC and MMPFC) and rSTS.

To identify individually-defined functional ROIs, initial “hypothesis spaces” were defined as the 9mm radius sphere centered about local maxima for each region, in the group random effects analysis performed on all 477 NT participants (see figure 1). Each participant’s contrast image

(Mental>Control) was masked with the six hypothesis spaces; all voxels contiguous with the peak voxel and significant at  $p < 0.001$ , within a 9mm radius, were defined as the ROI. From each ROI three parameters were extracted: the peak voxel t-value, the size of the ROI (number of voxels included), and the mean T. The presence or absence of an ROI was used as a fourth parameter. The reliability of ROI parameters was assessed by split-half analysis. Contrast images were derived from even versus odd runs in each participant. ROIs were picked using a minimum cluster size of 10 and a significance level of  $p < 0.05$ . The correlation of the ROI even and odd parameter values was measured across participants.

Every subject for whom we had complete demographic and experimental data was then included in a multivariate general linear model (GLM). The resulting model was a seven-column (age, gender, group, modality, coil, number of stimuli, and the mean term) predictor matrix and included data from 383 participants. For the binary statistic that indicated whether or not the ROI of interest was identified in a given subject, the GLM presumed a binomial distribution and a logit linker function. The GLM used a normal distribution otherwise. Continuous regressors were mean-centered prior to regression. Correction for multiple comparisons was performed with Bonferroni correction, across all predictors and all dependent measures, within each ROI. In total there were six predictors for the four ROI parameters, a total of 24 comparisons per ROI; thus effects were taken to be significant if  $p < 0.0021$ . Any relationship significant at  $p < 0.01$  is discussed as a ‘trend.’

An identical procedure was conducted for the matched group, except that coil and modality did not vary within and thus were omitted. IQ was added to the predictor matrix, resulting in a total of 20 comparisons per ROI, and a significance threshold of  $p < 0.0025$ . Any relationship found to have a significance  $0.01 < p < 0.0025$  is discussed as a trend.

**Whole-brain analyses:** Whole-brain analyses were conducted for the contrast of interest (Mental>Control), in order to identify effects on the ToM brain regions. To correct for comparisons, nonparametric whole-brain analysis was performed using SnPM (<http://www.sph.umich.edu/ni-stat/SnPM/>). Each test used 3mm variance smoothing and 5,000 permutations, with no global normalization, grand mean scaling, or threshold masking. The corrected  $p$ -value for filtering was 0.05, with a threshold of 3, and a voxel-cluster combining theta value of 0.5. Permutations were repeated for each predictor of interest; all demographic and experimental predictor variables were included as nuisance regressors using modified SnPM plugins. Because (to foreshadow our results) we find a *lack* of significant differences between ASD and NT participants, we also examined the results using a substantially more lenient threshold: regions were considered significant if composed of a contiguous cluster of at least ten voxels at a t-value of 3 or greater, as this

corresponds to  $p < 0.001$  (uncorrected). This more lenient threshold is consequently a more stringent test of the hypothesis that there are no differences between the groups.

## Results

### ROI results

Six functional ROIs (ROIs) from the ToM network were defined in individual participants, using the contrast Mental>Control, consistent with previous literature (U. Frith & Frith, 2003; Saxe & Kanwisher, 2003): RTPJ (in 414/504, or 82.1%), LTPJ (77.2%), PC (84.7%), DMPFC (60.1%), MMPFC (64.7%) and rSTS (65.5%).

The goal of this project is to explain individual differences in the size and magnitude of brain regions involved in ToM. Before testing individual differences, however, it was critical to determine that (i) there was variability in these measures, and (ii) the differences between individuals on these measures are reliable (i.e. that inter-individual differences do not simply reflect noise in the measurement). All ROI parameters showed reasonable variability. The standard deviation of the peak T-value ranged between 1 and 2, and the standard deviation of ROI size (in voxels) ranged from 60 to 90 voxels. In order to test whether this variability reflects stable individual differences, we compared the ROI measurements within individuals. ROIs were picked independently from even and odd runs in the 235 participants from whom we had more than three runs of data. RTPJ was identified in 72% of participants, LTPJ in 66%, PC in 75%, DMPFC in 55%, MMPFC in 53%, and rSTS in 56%. Correlations between the even and odd parameter values (mass, x coordinate, etc.) and across subjects had an average Pearson's  $r$ -value of 0.51. These correlations were all significant at  $p < 0.001$ , and all but two at  $p < 0.0001$ . Thus, the ROI parameters are reliable within subject, making it worthwhile to explain inter-subject variability.

Next we used multivariate general linear regression analyses to estimate whether any variance in the size or

response magnitude of ToM brain regions is explained by ASD status. The first set of analyses compared all of the individuals with ASD ( $n=27$ , 23 male) to all of the NT individuals ( $n=439$ , 179 male). In these analyses, no parameter of any ROI was significantly predicted by the group membership (ASD vs. NT) of the individual ( $p > .09$  for all ROIs). The ASD participants were similar to NT participants on the ROI measures considered; no ASD participant fell outside of 3 standard deviations on any measure or any ROI, and only one ASD participant fell outside 2 SDs. In a second set of analyses, we compared individuals with ASD ( $N=24$ , 19 male) to a group of matched NT individuals ( $N=24$ , 19 male). Again, we found no significant difference between groups on any ROI parameter (all  $p > 0.01$ ). The new parameter of IQ was found to predict larger sized PC ROIs ( $p = 0.0064$ ,  $\beta=2.591\pm2.699$ , +1.7 voxels/IQ point) at the level of a trend. Finally, the effect of ADOS score was considered. For this analysis, participants were restricted to those from the ASD group. None of the parameters significantly predicted any measured ROI parameter, even at the level of the trend.

The choice of coil had the largest effect. The 32-channel coil produced significantly greater peak (means:  $p = 0.0004$ ,  $\beta = 0.610\pm0.470$ , 1.21 units higher in 32-channel ROIs) and mean T values (means:  $p = 0.0006$ ,  $\beta=0.309\pm0.212$ , 0.62 units) in all ROIs except DMPFC and PC compared with the 12-channel coil. PC mean ( $p = 0.0030$ ,  $\beta = 0.253\pm0.260$ , 0.541 units) and peak T ( $p = 0.0039$ ,  $\beta = 0.480\pm0.509$ , 1.01 units) was increased in the 32-channel as well, but at the level of a trend. The 32-channel coil additionally significantly increased the size of the RTPJ ( $p = 0.0001$ ,  $\beta = 40.35\pm30.38$ ), and increased the probability of finding the RSTS ( $p = 0.0087$ ,  $\beta = 1.354\pm1.589$ , 152% more likely) and its size ( $p = 0.0030$ ,  $\beta = 24.495\pm25.203$ , 58.7 voxels larger) at the level of a trend. We also found an unexpected effect of number of stimuli: as the number of stimuli used in the experiment increased, the probability of identifying regions in the medial prefrontal cortex (MMPFC and DMPFC) decreased (means:  $p=0.0073$ ,  $\beta=-0.067\pm0.079$ ,  $\sim -2\%$

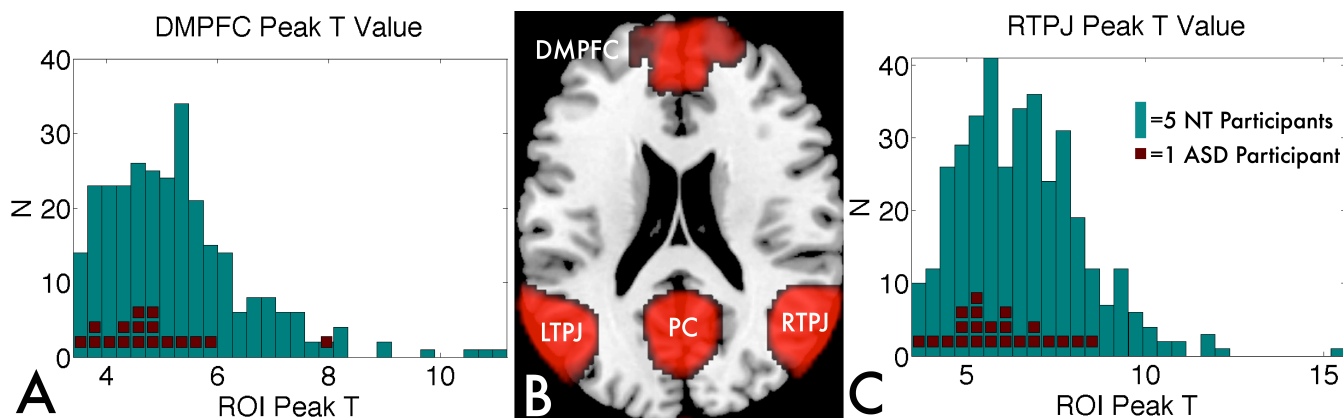


Figure 1: (A) Histogram of the DMPFC ROI peak T value for NT participants (teal) overlaid with the ASD participants (red). (B) Whole brain random effects analysis of the main effect, Mental > Control, in the full sample, corrected for multiple comparisons with permutations; axial slice shown at  $z = 22\text{mm}$ . Visible are RTPJ, LTPJ, DMPFC, and PC. (C) Histogram of the RTPJ ROI peak T value for NT participants (teal) overlaid with the ASD participants (red).

stimulus) at the level of a trend. There were no significant effects of age or gender on any parameter of any region.

In sum, ROI analyses suggest that while individuals differ reliably in the size and response magnitude of brain regions associated with ToM, these neural parameters are not affected by whether an ASD diagnosis. Experimental parameters, such as the MRI coil used, and demographic variables, such as IQ, explain some of the variance across individuals. Within the range of ADOS scores in the current sample, autism severity does not explain variance in ROI parameters, either. However, ROI analyses (and especially the three ROI parameters) provide a limited window on the brain, so to look further for differences between groups in ToM brain regions, we next conducted whole brain analyses.

### Whole brain analysis results

In the whole-brain analyses, the main effect identifies brain regions significantly recruited during Mental compared to Control conditions, controlling for variance explained by any of the nuisance regressors. This analysis identified robust activation in all of the regions previously associated with ToM, including RTPJ, LTPJ, medial PC and posterior cingulate, MPFC, and anterior STS. It also identified activation in other regions, including the left superior frontal gyrus (BA8 and BA6), the left medial frontal gyrus (BA8), regions of right middle frontal gyrus (BA6, BA8, BA9), the right superior temporal gyrus (BA38) and the right inferior frontal gyrus (BA47). Also present was activity in the cingulate (BA24) and anterior cingulate (BA32), as well as the thalamus (BA24) and the right amygdala.

Next, we compared activation in the full sample of individuals with ASD vs. NT. Regions were significant if the difference between activation during mental versus control tasks was greater in one group than in the other. When these analyses were corrected for multiple comparisons using permutations, we observed no regions of significant group differences. A more lenient threshold revealed a small region in the right cingulate gyrus ([2, 14, 22], peak  $T = 3.5$ ,  $128\text{mm}^3$ ) with a greater condition difference in ASD than NT groups. In this region, ASD participants showed greater deactivation in the control condition, but no difference during the Mental condition. There were no regions with greater difference between conditions in the NT participants.

We also compared the ASD group to a smaller matched group. When correcting for multiple comparisons with permutations, we again failed to find any regions significant for the ASD > NT contrast. More lenient traditional thresholds also failed to reveal any significant regions. In the reverse contrast (NT > ASD), a single region was found in the right middle occipital gyrus ([36, -62, -8], peak  $T = 5.09$ ,  $1032\text{mm}^3$ ) when corrected with permutations. This region was again identified using the more lenient threshold ([36, -62, -8], peak  $T = 5.8$ ,  $784\text{mm}^3$ ), along with regions in the left middle temporal gyrus ([-48, 10, -44, peak  $T = 4.44$ ,  $304\text{mm}^3$ ), the right middle posterior cingulate ([26, -68, 12],

peak  $T = 4.32$ ,  $736\text{mm}^3$ ), the left posterior lobe ([-44, -60, 38], peak  $T = 4.31$ ,  $488\text{mm}^3$ ), the left cingulate gyrus ([-16, -56, 26], peak  $T = 4.24$ ,  $424\text{mm}^3$ ), left inferior temporal gyrus ([-58, -28, -20], peak  $T = 4.24$ ,  $208\text{mm}^3$ ), the right posterior insula ([40, -24, 12], peak  $T = 4.12$ ,  $304\text{mm}^3$ ), right precentral gyrus ([32, -26, 68], peak  $T = 3.81$ ,  $168\text{mm}^3$ ), right superior temporal gyrus ([44, -60, 34], peak  $T = 3.77$ ,  $232\text{mm}^3$ ), and the left posterior cingulate ([-14, -54, 14], peak  $T = 3.76$ ,  $104\text{mm}^3$ ).

## Discussion

The main question we sought to address in this paper was whether individuals diagnosed with ASD show differences in the magnitude or extent of activity in ToM brain regions, compared to a large sample of NT participants. To this end, we aggregated data across multiple experiments to produce a large sample of NT individuals ( $N=477$ ) and a moderate sample of high functioning ASD individuals ( $N=27$ ). Before directly comparing them, we tested whether neural responses to Mental stimuli were reliable within participants and variable across participants, in the NT population. They were. Next, we tested whether the magnitude of neural responses to Mental vs Control stimuli differed between groups, either in targeted regions of interest or in whole brain analyses. For the most part, these analyses identified no reliable differences between groups, especially in the previously hypothesized ToM brain regions. These results suggest that differences between these groups of participants in ToM brain regions, if they exist, are small and could not be used to diagnose ASD.

We used two complementary analysis strategies: ROI analyses focused on previous identified ToM brain regions are more sensitive, whereas whole brain analyses find group differences anywhere in the brain, and are less restricted. For both kinds of analyses, we conducted two comparisons by regression with simultaneous nuisance regressors to control for demographic and experimental variance: the ASD group vs. the whole group of NT individuals, and the ASD group vs. NT individuals matched to the ASD group on age, gender, IQ and experimental parameters. For both comparisons, we found no reliable differences between groups in the size, response magnitude, or probability of identifying above-threshold voxels, in any ToM ROI. Indeed, the ROI parameters of individuals with ASD fell squarely within the distribution of typical values, almost never straying more than 2SD from the typical means. Also, ADOS scores of the ASD participants did not predict any ROI parameter, even at the level of a trend.

In the whole brain analyses, the results of group comparisons depended on the thresholds used for correcting for multiple comparisons. Permutation-based correction, which estimates the false positive rate empirically, revealed no significant differences between the two complete groups. When we reduced the sample to just the matched NT group, we found one region, in the right middle occipital gyrus, which showed increased response to Mental than Control stimuli in the NT group, but not the ASD group. However,



since this region did not show a higher response to Mental than Control stimuli in the overall main effect analysis of all participants, and is not typically associated with any kind of social cognition, we are cautious about making strong claims based on this effect.

Because these results suggest a null result - namely, no difference between groups - we also examined the same analyses at a more lenient threshold that could reveal true differences between groups that are just below the threshold for significance. Again we found no regions more active in the full NT sample, compared to the ASD group. A small ( $128 \text{ mm}^3$ ) region in right cingulate gyrus appeared more active in participants with ASD at this threshold; in this region, ASD participants showed greater deactivation to the control condition than NT participants. Reducing the sample to just the matched NT participants, and using the lower threshold, produced a number of small regions showing greater activation in NT than ASD participants. However, none were in any region in the main effect analysis of Mental > Control stimuli. Thus, we could not identify any region that both (a) was reliably recruited for Mental more than Control stimuli in 477 NT individuals, and (b) showed less activity in the same contrast, in individuals with ASD.

Using a similar analysis strategy, we also found that age and gender do not affect activity in ToM brain regions; nor do the modality of the stimuli (visual vs aural) or the experimental task. Thus, although individual differences in ToM brain regions are reliable and robust, they are not explained by simple demographic or experimental variables. The absence of an effect of gender is particularly noteworthy, because the full sample contained a large number of male and female participants. Behavioral measures of ToM often reveal an advantage for female participants (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997); apparently this advantage is not due to grossly different ToM brain regions.

One significant factor was the coil used. The 32-channel coil has documented higher SNR (Triantafyllou, Polimeni, & Wald, 2011); we found that this difference translated into larger ROIs that were more likely to be detected in individual participants. Thus, our results suggest that for individually-defined ROI analyses, the increased SNR of the 32-channel coil provides a clear benefit. On the other hand, increasing the number of stimuli per condition did not have the same benefit: medial prefrontal regions were less likely to be identified, in experiments using more stimuli. This unexpected effect could reflect habituation, after more than 20 stories about characters' false beliefs.

With regard to our key null results, the current study has advantages and disadvantages. On the one hand, the large sample size provides more power and sensitivity to detect effects where they exist. In particular, although our sample of ASD individuals was only moderately large, the very large sample of NT individuals included gives us very high confidence on the true mean of the ROI parameters in NT individuals. Finding that the ASD population mean does

not differ from the NT mean is thus strong evidence that these groups' data cannot be attributed to different population distributions.

However, these results cannot be interpreted as ruling out any differences in the neural mechanisms for ToM in individuals with ASD. One qualification of the current results is that the parameters measured here provide only a limited measure of a region's function. Other measures include the functional connectivity of each region and within-region spatial pattern of responses (Biswal, Zerrin Yetkin, Haughton, & Hyde, 1995; Haxby et al., 2001). Participants with ASD may differ in these other measures of ToM region function (Kleinhans et al., 2008). Indeed work in our lab using multi-voxel pattern analysis (MVPA) demonstrated the existence of reliable differences between ASD and NT individuals (Koster-Hale, Saxe, and Young, submitted).

Another qualification is that the ASD participants in the current sample are very high functioning. Although they meet diagnostic criteria for ASD (and have been shown to have behavioral deficits in ToM tasks in a previous study, Moran et al., 2011), these individuals are highly verbal and pass first-order false belief tasks. Thus, our results do not rule out gross differences in the ToM regions of lower-functioning individuals with ASD. On the other hand, the individuals in our sample are diagnosed with ASD because of disproportionate difficulties with social interaction and communication, and are similar to populations used in previous fMRI studies. Also, we found no evidence that within our participants, increasing ASD severity had any effect on the measured ROI parameters. So the current results imply that social cognitive impairments can occur without gross changes in the size or position of ToM brain regions. Collectively, the current results provide strong evidence that the neural differences between high functioning adults with ASD and NT participants are not due to gross changes in the magnitude of ToM brain region activity.

These results leave open a number of key questions. First, it will be key to identify the neural differences between adults with ASD and NT individuals that account for behavioral differences in ToM. One key possibility is that individuals with ASD are highly heterogeneous, so that different neural sources explain the behavioral delays in different individuals. If so, the group-average analyses used here may have limited sensitivity to detect those differences. Second, the current study focused on adults. It will be important in future research to test whether the developmental trajectory of ToM brain regions differs in children with ASD compared to NT children, even if the mature states of the system are reasonably similar. Finally, it would be useful to extend these analyses to lower-functioning individuals with ASD. Nevertheless, the implication of this study is that social-cognitive impairments can occur without large changes in the activation of ToM brain regions.

## Acknowledgements

This paper is based upon work supported by the Simons Foundation, the National Science Foundation (grant 095518), the Dana Foundation, a National Science Foundation Graduate Research Fellowship (grant 0645960), and a John Merck Scholars Grant. The authors wish to acknowledge Marina Bedny, Emile Bruneau, Hyowon Gweon and Jorie Koster-Hale for collecting fMRI data.

## References

- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the mind in the eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42(2), 241-251.
- Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or asperger syndrome. *Journal of Child Psychology and Psychiatry*, 38(7), 813-822.
- Baron-Cohen, S. (1989). The autistic child's theory of mind: A case of specific developmental delay. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 30(2), 285-297.
- Biswal, B., Zerrin Yetkin, F., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic Resonance in Medicine*, 34(4), 537-541.
- Dichter, G. S., Felder, J. N., & Bodfish, J. W. (2009). Autism is characterized by dorsal anterior cingulate hyperactivation during social target detection. *Social Cognitive and Affective Neuroscience*, 4(3), 215-226.
- Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2010). fMRI item analysis in a theory of mind task. *NeuroImage*.
- Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431), 459-473.
- Gilbert, S. J., Bird, G., Brindley, R., Frith, C. D., & Burgess, P. W. (2008). Atypical recruitment of medial prefrontal cortex in autism spectrum disorders: An fMRI study of two executive function tasks. *Neuropsychologia*, 46(9), 2281-2291.
- Hasson, U., Avidan, G., Gelbard, H., Vallines, I., Harel, M., Minshew, N., et al. (2009). Shared and idiosyncratic cortical activation patterns in autism revealed under continuous real-life viewing conditions. *Autism Research*, 2(4), 220-231.
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425-2430.
- Kennedy, D. P., & Courchesne, E. (2008). Functional abnormalities of the default network during self-and other-reflection in autism. *Social Cognitive and Affective Neuroscience*, 3(2), 177-190.
- Kleinhans, N. M., Richards, T., Sterling, L., Stegbauer, K. C., Mahurin, R., Johnson, L. C., et al. (2008). Abnormal functional connectivity in autism spectrum disorders during face processing. *Brain*, 131(4), 1000-1012.
- Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, 46(12), 2949-2957.
- Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., & Baron-Cohen, S. (2011). Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. *NeuroImage*.
- Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., et al. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences*, 108(7), 2688-2692.
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind. *Behavioral and Brain Sciences*, 1(4), 515-526.
- Redcay, E., Dodell-Feder, D., Mavros, P. L., Kleiner, M., Pearrow, M. J., Triantafyllou, C., et al. (2012). Atypical brain activation patterns during a face-to-face joint attention game in adults with autism spectrum disorder. *Human Brain Mapping*.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people:: The role of the temporo-parietal junction in. *NeuroImage*, 19(4), 1835-1842.
- Triantafyllou, C., Polimeni, J. R., & Wald, L. L. (2011). Physiological noise and signal-to-noise ratio in fMRI with multi-channel array coils. *NeuroImage*, 55(2), 597-606.
- Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, 107(15), 6753.
- Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, 1(3), 333-349.
- Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, 40(4), 1912-1920.
- Young, L., & Saxe, R. (2009). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, 21(7), 1396-1405.
- Young, L., Scholz, J., & Saxe, R. (2011). Neural evidence for "intuitive prosecution": The use of mental state information for negative moral verdicts. *Social Neuroscience*, 6(3), 302-315.



# Explanations of Counterfactual Inferences

Brian J. Edwards (Brian.Edwards@u.northwestern.edu)

Department of Psychology, Northwestern University, 2029 Sheridan Rd., Evanston, IL 60208 USA

Lance J. Rips (Rips@northwestern.edu)

Department of Psychology, Northwestern University, 2029 Sheridan Rd., Evanston, IL 60208 USA

## Abstract

When engaging in counterfactual thought, people must imagine changes to the actual state of the world. In this study, we investigated how people reason about counterfactual scenarios by asking participants to make counterfactual inferences about a series of causal devices (i.e., answer questions such as *If component X had not operated [had failed], would components Y, Z, and W have operated?*) and to explain their reasoning. Participants avoided breaking deterministic causal links (i.e., *W always causes X*), but were willing to break probabilistic causal links (i.e., *W sometimes causes X*) to keep prior causal events in the same states as in the actual world. Participants' explanations supported this pattern of inferences. When the causal links were deterministic, participants reasoned diagnostically to infer that the states of prior causal events would have been different in the counterfactual world. In contrast, when the links were probabilistic, participants cited the links' unreliability as an explanation for why the states of prior causal events would have been the same as in the actual world. Additionally, participants who were told that a component "had failed" (vs. "had not operated") were more likely to attribute the state of that component to it being "internally broken" and infer that causally upstream components would have operated. Our results suggest that people use their explanation of the antecedent event (the "if" clause) to guide their counterfactual inferences. We discuss the implications of these findings for two rival Bayes-net theories of counterfactual reasoning: Pearl's (2000) and Hiddleston's (2005).

**Keywords:** Counterfactuals, causation, explanation

## Introduction

People often engage in counterfactual reasoning (e.g., *If I hadn't partied the night before the exam, then I would have passed the exam*) to second-guess decisions, attribute credit or blame, and diagnose causal relations (see Byrne, 2005, for a review). Reasoning about counterfactual scenarios such as the preceding example requires imagining changes to the actual state of the world—for instance, imagining a counterfactual world in which I hadn't partied the night before the exam. One of the central issues in the study of counterfactual reasoning is how people re-imagine the world to satisfy the antecedent of a counterfactual scenario. (The *antecedent* is the "if" clause, and we will refer to the "then" clause as the *consequent*.) In particular, what types of events do people keep the same in the actual and counterfactual worlds and what types of events do people change?

One way people might reason about counterfactual scenarios, which we will call *pruning theory*, is by using an intervention to change the state of the antecedent event from

the actual state to the counterfactual state and then tracing the consequences of that intervention (Pearl, 2000; see also Woodward, 2003). The intervention severs the causal link between the antecedent and its immediate causes, and as a result of this "graph surgery," the counterfactual states of upstream events would be the same as in the actual world. However, downstream events that are a consequence of the antecedent would change states according to the causal laws governing the system. To illustrate this approach, consider a causal chain  $A \rightarrow B \rightarrow C$  and a counterfactual antecedent *If B had not occurred...* (in the actual world, *A*, *B*, and *C* all occurred). A person using pruning theory would intervene on *B* to change the state of *B* from present to absent. Since upstream events (*A*) are unaffected by this intervention, *A* would still have been present in the counterfactual world. But since *C* is an effect of *B*, *B*'s absence would in turn cause *C* to be absent.

Pruning theory might appeal to reasoners in two ways. First, by keeping all the events that are causally prior to the antecedent in the same states as in the actual world, pruning theory creates a counterfactual world that is maximally similar to the actual world with respect to these prior events. Second, the pruning approach makes counterfactual thinking computationally easy. The strategy of always keeping prior events in their original states allows reasoners to avoid the cognitively challenging process of reasoning backwards to determine the counterfactual states of upstream causes.

However, other researchers have questioned whether the type of change pruning theory proposes is necessarily the most reasonable way to modify the causal system in the counterfactual situation (e.g., Hiddleston, 2005). One criticism of pruning theory is that it is very disruptive to the structure of a causal system and can require reasoners to violate causal laws. Consider a deterministic causal system in which *A*, without exception, always causes *B*. In this setting, one might be reluctant to imagine a counterfactual world in which *A* occurred, but *B* did not occur (e.g., in answering the question *If B had not occurred, would A have occurred?*). Thus, when reasoning about this counterfactual scenario, one might be more likely to infer that the reason *B* did not occur was that *A* did not occur, and the absence of *A* caused *B* to be absent too (Hiddleston, 2005). We will call this alternative *minimal-network theory*. When the causal links are probabilistic (i.e., *A sometimes causes B*), however, minimal-network theory proposes that *A* might or might not have occurred, since either possibility is "legal" in accordance with the system's causal laws.

Table 1 compares the predictions of pruning theory and minimal-network theory for a device in which component A's operating usually causes component B to operate and component B's operating always causes component C to operate (at present, all three components are operating). The device's structure is illustrated as follows:



Table 1: Comparison of Pruning Theory and Minimal-network Theory

	If component B had not operated, would component A have operated?	If component C had not operated, would component B have operated?
Pruning Theory	Yes	Yes
Minimal-network Theory	Maybe	No

Previous empirical work has explored whether people's counterfactual inferences are consistent with either of these two theories of counterfactual reasoning. In one experiment, Sloman and Lagnado (2005) presented people with causal information about a simple rocket-ship device with the causal structure  $A \rightarrow B$  and asked them a variety of counterfactual questions. Sloman and Lagnado found evidence that people engaged in pruning when they were told that a component was *prevented* from operating, but not when told that the component was *observed* not to have operated. However, subtle differences in wording across their experiments led to significantly different patterns of counterfactual inferences, making it difficult to generalize from the data. In another study, Rips (2010) asked people counterfactual questions about three- and four-component mechanical devices. Although participants' counterfactual inferences did not provide strong support for either pruning theory or minimal-network theory, their inferences were more closely aligned with minimal-network theory (see also Dehghani, Iliev, & Kaufmann, 2012).

In the two experiments in this study, we presented participants with counterfactual questions for which pruning theory and minimal-network theory make different predictions. The wording of these questions was manipulated across two between-subjects conditions. One group of participants was told that a component of a mechanical device "had not operated," and another group was told that the component "had failed" (e.g., *If Component B had not operated/had failed...*). The neutral "had not operated" wording does not suggest a particular explanation for the state of the component; however, the "had failed" wording suggests an explanation that is local to the component (e.g., the component is internally broken). Thus, we predict that participants in the *not operated* and *failed* conditions will make different counterfactual

inferences about the operating states of the other components. Specifically, we predict that participants in the *not operated* condition will reason diagnostically about the states of the other components based on the device's causal structure, consistent with minimal-network theory. In contrast, we predict that participants in the *failed* condition will reason that since the antecedent component is broken, its operating state is not diagnostic of the states of the other components. Thus, participants will break the causal links between the antecedent and its causes and infer that causally prior components would have operated in the counterfactual situation, consistent with pruning theory. In addition to examining participants' inferences about which components would and would not have operated in the counterfactual situation, we analyzed participants' explanations of their reasoning. In Experiment 1, we analyzed people's explanations of why they thought the non-antecedent components would or would not have operated. In Experiment 2, we analyzed people's explanations of why the antecedent event would have occurred.

### Experiment 1

Participants in this experiment received a series of problems about a set of eight hypothetical devices, each with four components. For each device, they answered counterfactual questions of the form *If component X had not operated [had failed], would components Y, Z, and W have operated?* and provided explanations justifying their reasoning.

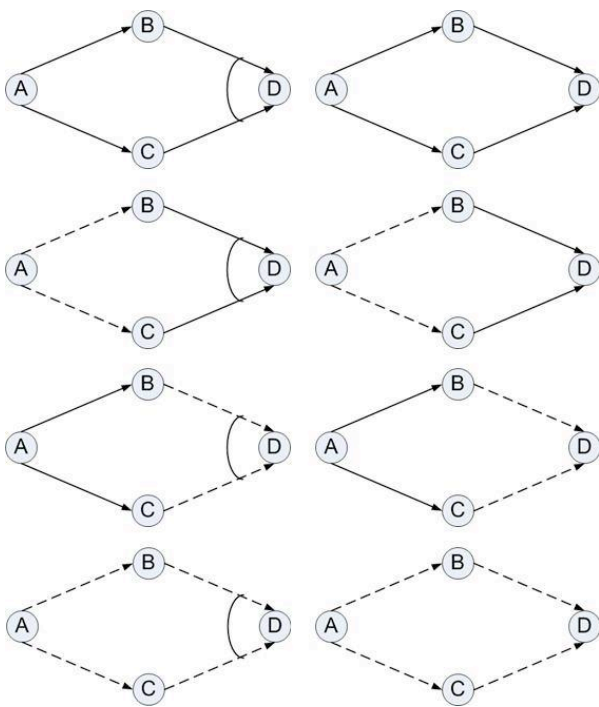
### Method

**Materials.** The questionnaire booklets contained three pages of instructions followed by 24 pages of questions. The instructions explained the experimental task and told participants how to interpret the diagrams of the causal devices on the following pages. Each question page contained a written description of how a device operated (e.g., Component A's operating always causes component B to operate, etc.), which was accompanied by the corresponding diagram in Figure 1.

As shown in Figure 1, there were eight different causal devices, all of which had "diamond" structures. The devices varied in whether the causal links between components were deterministic (solid lines in Figure 1) or probabilistic (dashed lines), and whether components B and C had to operate together to cause D to operate (arc connecting links in Figure 1) or could independently cause component D to operate (no arc). The order of the devices was counterbalanced across participants. We used devices with diamond structures for two reasons. First, previous causal reasoning studies have used diamond structures and have found that people make accurate causal inferences about these systems (Meder, Hagmayer, & Waldmann, 2008, 2009). Second, pruning theory and minimal-network theory make different predictions for many of the counterfactual questions about these devices.

After learning how each device works, participants were told the device's current operating state, which was always that "at present, components A, B, C, and D are all operating." Next, participants were asked a counterfactual question about the device, such as *If component B had not operated, would components A, C, and D have operated?* For each of the eight devices, participants answered three counterfactual questions, one question each with A, B, and D as the antecedent component. Since the devices were symmetric with respect to components B and C, we did not ask a separate question in which C was the antecedent. The order of the antecedent components for these questions (ABD vs. DBA) was balanced across participants.

Figure 1: Causal Devices Used in Experiments 1 and 2



In this figure, solid arrows indicate deterministic links (e.g., A always causes B) and dashed arrows indicate probabilistic links (e.g., A usually causes B). All causal relationships are in the direction shown by the arrows. The arcs indicate that component B and component C operating together cause component D to operate, but component B or component C operating alone never causes component D to operate (jointly caused devices). The absence of an arc indicates that component B or component C operating alone causes component D to operate (separately caused devices).

Participants were randomly assigned to one of two experimental conditions. In the *not operated* condition, participants learned that the antecedent component (component B in the preceding example) "had not operated." In the *failed* condition, participants learned that the antecedent component "had failed."

For each counterfactual question, participants indicated which of the three non-antecedent components would have operated in the counterfactual state. For each component, participants could say that the component (1) would have operated, (2) would not have operated, or (3) might or might not have operated. To gain insight into how participants were reasoning about the counterfactual questions, participants also indicated the order in which they reasoned about the non-antecedent components. After making these inferences, participants justified their answers by responding to the prompt "Please explain why you answered in the way you did."

**Procedure.** Participants received the questionnaire booklet from the experimenter and answered the questions at their own pace. The experiment took approximately 30 minutes to complete.

**Participants.** Participants were 32 undergraduate students at Northwestern University. Participants received course credit for their participation.

## Results and Discussion

We analyzed participants' answers to the counterfactual questions (e.g., *If component B had not operated, would component A have operated?*) to see if their inferences were consistent with minimal-network theory or pruning theory. Responses of "would have operated" were scored as +1, responses of "would not have operated" were scored as -1, and responses of "might or might not have operated" were scored as 0. The mean score for participants was higher in the *failed* condition ( $M = -0.14$ ) than in the *not operated* condition ( $M = -0.43$ ),  $F(1, 32) = 7.07$ ,  $MSe = 7.29$ ,  $p = .01$ .

In two cases, pruning theory and minimal-network theory make the same predictions: (1) when component A was the antecedent, and (2) for the devices in which components B and C must both operate in order for component D to operate (jointly caused devices), when component B was the antecedent and component D was the consequent. In case (1), both theories say that components B, C, and D would all not have operated, and in case (2), both theories say that component D would not have operated. For all the other counterfactual questions, pruning theory predicts that the consequent component definitely would have operated (producing positive scores), whereas minimal-network theory predicts that the consequent component either (a) definitely would not have operated or (b) might or might not have operated (producing negative or 0 scores respectively).

When we restricted our analysis to the cases in which pruning theory and minimal-network theory make different predictions, the mean score for participants in the *failed* condition was 0.17 and the mean score for participants in the *not operated* condition was -0.27. As was the case with the entire data set, the difference between conditions was significant,  $F(1, 32) = 11.96$ ,  $MSe = 6.27$ ,  $p = .002$ . The mean score for the *not operated* condition was significantly less than 0,  $t(17) = -4.50$ ,  $p < .001$ ; however, the mean score

for the *failed* condition was not significantly different from 0,  $t(17) = 1.68$ , n.s.

Next, we examined the serial order (1, 2, or 3) in which participants reasoned about the three non-antecedent components. The most interesting case is the one in which component B was the antecedent since participants could work their way downstream (i.e., reason about component D first) or upstream (i.e., reason about component A first). Most participants (69%) started upstream, reasoning about component A before component D (Binomial test,  $p < .001$ ). The mean serial position for component A was 1.44, whereas the mean position for component D was 2.32. The order in which participants reasoned about the components did not differ across the *failed* and *not operated* conditions.

We also examined participants' explanations of their counterfactual reasoning to see if the explanations were consistent with pruning theory or minimal-network theory. We classified explanations in two ways.

(1) Explanations were coded as *causal backtracking* if participants used the state of the antecedent component to reason diagnostically about the states of upstream components. A sample causal-backtracking explanation was "If B wasn't operating that would mean A wasn't working since A always causes B." Causal-backtracking explanations are consistent with minimal-network theory.

(2) Explanations were coded as *causes are independent of effects* if they suggested that the states of upstream "cause" components are not affected by the states of downstream "effect" components. A sample explanation was "Neither A, B, nor C are dependent on D so they all will have operated." Such an explanation is consistent with pruning theory.

Notice that these three types of explanations are only applicable when there are components that are causally upstream of the antecedent component. Thus, we restricted the following analyses to the counterfactual questions in which B or D was the antecedent. The data were coded by a person who was unfamiliar with the experimental hypotheses, and 25% of the data were coded independently by a second coder. Inter-coder reliability was 90%.

Participants in the *not operated* condition were significantly more likely to provide "causal-backtracking" explanations than participants in the *failed* condition (65% vs. 32% respectively,  $F(1,24) = 12.9$ ,  $MSe = 16.4$ ,  $p = .001$ ). In contrast, participants in the *failed* condition were significantly more likely to provide "causes are independent of effects" explanations than participants in the *not operated* condition (25% vs. 9% respectively,  $F(1, 21) = 5.57$ ,  $MSe = 3.90$ ,  $p = .03$ ). Participants in the *not operated* condition were significantly more likely to provide "causal-backtracking" explanations than "causes are independent of effects" explanations ( $t(14) = 6.33$ ,  $p < .001$ ); however, participants in the *failed* condition did not significantly prefer either type of explanation.

In sum, participants in the *not operated* and *failed* conditions differed in their counterfactual inferences. Participants in the *not operated* condition had a stronger tendency to say that non-antecedent components would not

have operated than participants in the *failed* condition, and they made inferences that were better predicted by minimal-network theory. The analysis of participants' explanations also showed that most participants in the *not operated* condition used causal backtracking to diagnose the counterfactual operating states of upstream components. In contrast, participants in the *failed* condition were more likely than participants in the *not operated* condition to say that the operating states of upstream components were independent of, and could not be diagnosed from the state of the antecedent.

## Experiment 2

The pattern of inferences and reasoning strategies in Experiment 1 suggests that participants in the *not operated* and *failed* conditions may have generated different explanations for why the antecedent component had not operated. We therefore performed a second experiment to investigate the possible relationship between participants' explanations of why the antecedent component had not operated and their counterfactual inferences.

### Method

The experiment contained two parts, an inference task and an explanation task. The same eight causal devices from Experiment 1 were used in Experiment 2 (see Figure 1). As in Experiment 1, participants were randomly assigned to either the *not operated* condition or the *failed* condition.

### Materials.

*Inference task:* The inference task was identical to Experiment 1 except that participants did not provide explanations of their counterfactual inferences during this part of the experiment.

*Explanation task:* In the explanation task, participants described why the *antecedent* component had not operated. Note that this is a different type of explanation than the ones participants provided in Experiment 1; in Experiment 1, participants explained why the *non-antecedent* components would or would not have operated. The explanation-task booklet included three pages of instructions followed by 24 pages of questions. As in the inference task and Experiment 1, participants received information about how the causal devices work and told that "at present, components A, B, C, and D are all operating." Participants in the *not operated* condition were asked questions of the form *If component X had not operated, which of the following would best explain why?* Participants in the *failed* condition were asked a question that was identical except that "not operated" was replaced by "failed." For each device, participants answered this question for each of components A, B, and D as the antecedent. For each participant, the order of the devices, and within each device, the order of the antecedent components, was the same in the inference and explanation tasks.

When component B was the antecedent, participants selected an explanation from the following list:

- (1) Component B was internally broken.
- (2) Factors external to the device prevented component B from operating.
- (3) Component B operates unreliably, and component B just didn't operate this time.
- (4) Component A did not operate, which in turn caused component B not to operate.
- (5) Component A operated, but component B just didn't operate this time because the connection between component A and component B is unreliable.
- (6) Component A operated, but the connection between component A and component B was broken.

The list of explanations was similar when component D was the antecedent, except that "component D" was substituted for "component B" and "component B and/or<sup>1</sup> component C" was substituted for "component A." When component A was the antecedent, only the first three answer choices were included since component A's operation is not caused by other components. The order of the answer choices (above order vs. reverse order) was balanced across participants.

After choosing an explanation, participants rated their confidence on a 0-9 scale with one-point increments, where 0 = "not at all confident" and 9 = "extremely confident."

**Procedure.** Half of the participants completed the inference task followed by the explanation task and the remaining participants completed the explanation task followed by the inference task. Each task took approximately 20 minutes with the entire experiment taking approximately 40 minutes.

**Participants.** Participants were 32 undergraduate students at Northwestern University who had not participated in Experiment 1. Participants received course credit for their cooperation.

## Results and Discussion

**Inference Task.** The inference task replicated the findings of Experiment 1. The mean score for participants in the *failed* condition was significantly higher than for participants in the *not operated* condition. This was true for all counterfactual questions ( $M = -0.16$  vs.  $M = -0.48$  respectively,  $F(1,30) = 14.47$ ,  $MSe = 4.14$ ,  $p < .001$ ) and for the subset of counterfactual questions for which pruning theory and minimal-network theory make different predictions ( $M = 0.27$  vs.  $M = -0.25$  respectively,  $F(1, 30) = 19.27$ ,  $MSe = 4.94$ ,  $p < .001$ ).

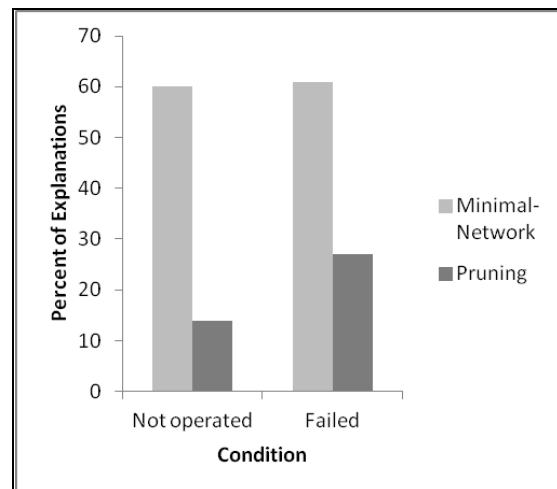
**Explanation Task.** Participants' explanations were coded as consistent with pruning theory, consistent with minimal-network theory, or consistent with neither theory. Explanations 1, 2, and 6 (see Method section) were

<sup>1</sup> If either component B or component C operating alone could cause component D to operate, the "and" wording was used. Otherwise, the "or" wording was used.

classified as pruning explanations. When the links between the antecedent and its causes were deterministic, explanation 4 was classified as a minimal-network explanation. When the links between the antecedent and its causes were probabilistic, explanation 5 was classified as a minimal-network explanation. All other responses were classified as "other." Since neither explanation 4 nor explanation 5 (the two possible minimal-network explanations) is applicable when component A was the antecedent, the following analyses were conducted only for the counterfactual questions in which component B or D was the antecedent.

Participants were significantly more likely to choose minimal-network explanations than pruning explanations (61% vs. 21% respectively,  $t(31) = 5.31$ ,  $p < .001$ ). This pattern was observed in both the *not operated* (60% vs. 14% respectively,  $t(15) = 4.92$ ,  $p < .001$ ) and *failed* conditions (61% vs. 27% respectively,  $t(15) = 2.85$ ,  $p = .01$ ). Notice that participants in the *failed* condition were significantly more likely to choose pruning explanations than participants in the *not operated* condition ( $F(1,29) = 4.56$ ,  $MSe = 3.04$ ,  $p = .04$ ). The results are shown in Figure 2.

Figure 2: Percent of Minimal-Network and Pruning Explanations by Condition



Interestingly, when the causal links between the antecedent and its causes were probabilistic, participants in both conditions were significantly more likely to choose a minimal-network explanation (e.g., Component A operated, but component B just didn't operate this time because the connection between component A and component B is unreliable) than a pruning explanation (e.g., Component B was internally broken; Factors external to the device prevented component B from operating; Component A operated, but the connection between component A and component B was broken), (*Not operated* condition:  $t(15) = 5.81$ ,  $p < .001$ , *Failed* condition:  $t(15) = 5.00$ ,  $p < .001$ ). All these explanations (both the pruning and minimal-network explanations) imply, and in some cases state explicitly, that causally upstream components would have operated. Even



though this counterfactual state is consistent with both pruning theory and minimal-network theory, participants in both conditions preferred minimal-network explanations.

As in Experiment 1, minimal-network theory better explained the inferences of participants in the *not operated* condition compared to pruning theory. While participants in the *failed* condition were more likely than participants in the *not operated* condition to say that non-antecedent components would have operated, participants in both conditions preferred minimal-network explanations over pruning explanations. Thus, Experiment 2 suggests that minimal-network theory might provide a starting point for a good psychological theory of counterfactual reasoning.

## General Discussion

In the two experiments in this paper, we examined (1) participants' counterfactual inferences about the states of variables in a causal system and (2) participants' explanations of their reasoning. Alternative theories of counterfactual reasoning such as pruning theory and minimal-network theory make different predictions about how people should modify (or preserve) the system's causal structure when reasoning about a counterfactual scenario.

A defining characteristic of pruning theory is the proposal that people treat counterfactuals as interventions. Under this account, people should simulate the counterfactual state by intervening on the causal system, and we would expect them to break both probabilistic and deterministic causal links and say that upstream components would have operated. Furthermore, they should endorse an interventionist explanation for the counterfactual state of the antecedent component, such as "factors external to the device prevented the antecedent component from operating."

Our data provide evidence against this hypothesis. Participants in the neutrally worded *not operated* condition made counterfactual inferences that preserved deterministic causal relationships between components' operating states. When the causal links between the antecedent component and its causes were deterministic, participants inferred that the antecedent component's causes would not have operated, which in turn caused the antecedent component not to operate. However, when the causal links were probabilistic, participants inferred that the antecedent component's causes would have operated, but the antecedent component would not have operated because the links were unreliable. These inferences and explanations are consistent with minimal-network theory, which proposes that people should prefer "legal" counterfactual states that preserve the system's (deterministic) causal laws, but they are inconsistent with pruning theory.

We also found that participants in the *not operated* and *failed* conditions reasoned differently about the counterfactual scenarios. The *failed* wording suggested to participants that the antecedent component was internally broken. Accordingly, these participants modified the devices' causal structure by breaking the causal links between the antecedent and its causes, and they inferred that

upstream components would have operated. Other studies that have varied the wording of counterfactual questions have found similar effects (Sloman & Lagnado, 2005).

Each type of wording supports a particular (and different) explanation for the counterfactual antecedent. The differences in participants' explanations across conditions suggest that these explanations may in turn shape participants' counterfactual inferences. Hempel (1965) famously proposed that causal explanations support predictive inferences, and our data suggest such a connection between explanation and inference in counterfactual reasoning (Goodman, 1955). Specifically, we propose that when engaging in counterfactual reasoning, people integrate their explanation of the counterfactual antecedent with their knowledge of the system's causal structure to infer the system's counterfactual state.

## Acknowledgements

We thank Ben Rottman, Steven Sloman, and members of the Northwestern University Higher Level Cognition Laboratory for their valuable feedback on these experiments. We thank Samantha Thompson and Joanna Westerfield for their research assistance. The research was supported by an NSF graduate research fellowship (BJE) and IES Grant R305A080341 (LJR).

## References

- Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality*. Cambridge, MA: MIT Press.
- Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind and Language*.
- Goodman (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.
- Hempel, C. G. (1965). Aspects of scientific explanation. In C. G. Hempel, *Aspects of scientific explanation and other essays in the philosophy of science* (pp. 331-496). New York: Free Press.
- Hiddleston, E. (2005). A causal theory of counterfactuals. *Nous*, 39, 632-657.
- Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review*, 15, 75-80.
- Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition*, 37, 249-264.
- Pearl, J. (2000). *Causality*. Cambridge, UK: Cambridge University Press.
- Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, 34, 175-221.
- Sloman, S. A., & Lagnado, D. A. (2005). Do we "do"? *Cognitive Science*, 29, 5-39.
- Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

# Learning Conceptual Hierarchies by Iterated Relational Consolidation

James M. Foster, Fabián Cañas, & Matt Jones

{james.m.foster, canas, mcj}@colorado.edu

University of Colorado, Department of Psychology & Neuroscience  
Boulder, CO 80309 USA

## Abstract

Learning new concepts is critical to making sense of the world. Research on analogical reasoning suggests structure mapping and schema induction can enable discovery of new relational concepts. However, existing theories of schema induction and refinement are insufficient to explain acquisition of rich, compositional hierarchies of relational concepts. This paper offers a proposal for this sort of representation construction, founded on reinforcement learning to evaluate the predictive usefulness of higher-order relations, together with a mechanism of relational consolidation by which systems of relations (schemas) can be chunked into unitary entities. A computational model of these ideas is outlined and partially tested in simulations and human experiments. Implications and moderating factors for relational consolidation are considered.

**Keywords:** Relational Consolidation; Analogy; Schema Induction; Predication; Refinement; Concept Learning

## Introduction

Consider a second-order same-different task, in which the subject is presented with two pairs of objects and must recognize whether the pairs match in terms of whether their objects are the same or different. The pairs match if both are instances of *sameness* (e.g., A-A, B-B) or if both are instances of *difference* (e.g., A-B, C-D), and they mismatch if one is an instance of sameness and the other an instance of difference (e.g., A-A, B-C). Thompson, Oden, and Boyson (1997) tested naive chimpanzees on this task and found them unable to learn it, unless they were first trained on a first-order same-different task. In the first-order task, a single pair of objects was presented, and subjects learned to associate sameness and difference to two plastic tokens (e.g., a yellow triangle and a red circle, respectively). Thompson et al. argued this training enabled subjects to reduce the second-order same-different task to a first-order task, by mentally replacing each pair of objects with its associated token, and then determining whether those tokens matched (see also Clark, 2006).

Learning higher-order relations, such as in the second-order same-different task, is arguably critical to abstract conceptual development. In this paper, we argue that many concepts reside in relational hierarchies (relations among relations, and so on), and we investigate how such concepts might be learned. Our basic premises are that much structure in the world (or at least its mental representation) is hierarchically compositional, and that discovering (or creating) this structure is a powerful cognitive mechanism for both learning and designing complex systems.

For example, computer architecture, mathematical functions, and natural languages all exemplify multiple

levels of abstraction by chunking systems of relations at one level into building blocks at the next level. In computer architecture, digital logic gates are composed to form adders, which are composed with other digital circuits to form an arithmetic logic unit (ALU), which is a building block in a computer's CPU. Software design manages complexity by continuing this hierarchy, composing primitive functions into more complex functions, and from there to objects and design patterns. The conceptual progression in mathematics proceeds similarly, composing the counting operation to define adding, which is further composed to form multiplying, and then exponentiation. In traditional views of linguistics, phonemes, morphemes, words, and sentences form another example of a relational hierarchy.<sup>1</sup>

These examples motivate our basic research questions. How are relational hierarchies mentally represented? How are these representations learned or constructed through experience? Once a relational concept is learned, how can one discover the higher-order relations in which it can participate?

Here we consider the proposal that much of concept learning is driven by recognizing relational structure through analogy. Research on analogical reasoning has converged on a view that episodes or scenarios are represented as patterns of role binding, in which objects are bound to roles of relations (Gentner, 1983; Hummel & Holyoak, 2003). For example, the fact that the earth revolves around the sun is represented by binding *earth* and *sun* to the first and second roles of a *revolves\_around* relation. An analogy between two scenarios constitutes a determination that they share a common pattern of role binding. For example, in the analogy between the solar system and the atom (Gentner, 1983), each system has the property that the object playing role 1 of *revolves\_around* is the same as the object playing role 2 of *more\_massive\_than*.

Analogy formation can thus be viewed as a search for a pattern of linkage among relations (i.e., in terms of how they are bound to shared objects) that holds in two different scenarios. This linkage pattern is a type of higher-order relation among the linked relations. Theories of schema induction (e.g., Doumas, Hummel, & Sandhofer, 2008) offer one way for such higher-order relations to be learned. When an analogy is formed, an abstract schema is created that captures the common structure discovered by the

---

<sup>1</sup> Relational hierarchies are not taxonomic hierarchies. In a taxonomic hierarchy, each concept or category is a union of lower-level categories. In a relational hierarchy, each *instance* of a concept is a configuration of instances of lower-order concepts.

analogy. The schema can act as a new relational concept, in that it can be analogically aligned with future instances of the higher-order relation it embodies.

Although we agree with theories of schema induction, we argue it is insufficient to explain human relational learning. Schemas are explicit relational structures, and thus they cannot be bound to roles of yet-higher-order relations in the way unitary objects and relations can. The Thompson et al. (1997) study suggests that newly learned relations can only fill roles of other relations if they can be represented as atomic entities. Therefore, to explain acquisition of relational hierarchies, we put forward the hypothesis that useful schemas are eventually replaced (or supplemented) with unitary representations. Thus, a concept that was represented as a system of relations (via the schema) can now be represented as an atomic entity, capable of entering into relations itself. We label this process *relational consolidation*, in a deliberate parallel to theories of episodic memory consolidation (e.g., Squire & Alvarez, 1995).

We further propose that analogy, schema induction, and relational consolidation form a cycle that, when iterated, can produce relational hierarchies of arbitrary depth (height). This form of learning leads to a dualist view of objects and relations, in which (nearly) every concept is both a relational structure among its components and an object capable of participating in relations. The conceptual systems built from this hierarchical relational chunking are potentially quite powerful and flexible.

The remainder of this paper sketches a computational model under development that formalizes these ideas. We report experimental tests and discuss implications of human learning of higher-order relational structures.

## Model

We propose a computational model for learning hierarchies of relational concepts, named APEC for Analogy, Predication, Evaluation, and Consolidation. The first two stages (A, P) of the model draw on prior work on analogy and schema induction (Doumas et al., 2008; Forbus, Gentner, & Law, 1995; Larkey & Love, 2003). The last two stages make new proposals for how schemas are selected (E) and consolidated (C) into new concepts. Altogether, the model progresses through parallel processes of analogy formation, predication of relational structure by schema induction, evaluation and refinement of schemas in a reinforcement-learning setting, and consolidation of useful schemas into new atomic relations. Consolidated relations enter into new analogies, allowing the entire learning process to iterate.

The goal of the model is to identify new higher-order relations that are useful for making predictions and guiding behavior. There are an infinite number of higher-order relations that could be learned from any episode, and thus the challenge is selecting those that carry useful information (analogous to the problem of selecting configural cues in feature-based learning; e.g., Gluck & Bower, 1988). The present model addresses this problem in two ways. First,

analogical mapping identifies higher-order relations that recur across multiple episodes, to determine which schemas to induce (a form of unsupervised learning). Second, schema evaluation determines how useful each higher-order relation is for predicting outcomes or reward, to determine which schemas to consolidate (a form of reinforcement learning).

The model is currently being implemented within Conway's Game of Life (Gardner, 1970), a cellular automaton exhibiting hierarchical emergent structure, to test its ability to discover that structure. The model produces interesting patterns of schema formation and evolution, which will be reported elsewhere. Here we lay out the model's basic architecture and formulation.

## Analogy

APEC represents episodes as systems of role binding among entities, each of which is an instance of a known concept. Every entity is eligible to be bound to a role of one or more other entities, and all entities except primitive objects (used to seed the model) have roles that other entities can bind to. The goal of the analogy component of the model is to find correspondences between episodes that maximally preserve these role-filler relationships (i.e., parallel connectivity).

Formation of an analogy is achieved by a dynamic process of structure mapping. APEC's mapping dynamics are based on a simplified version of the Connectionist Analogy Builder (CAB; Larkey & Love, 2003). For every pair of entities (say,  $a_i$  in episode 1 and  $b_j$  in episode 2), a mapping weight ( $m_{ij}$ ) is defined. Mapping weights evolve according to excitatory and inhibitory dynamics. The raw evidence,  $R_{ij}$ , for mapping weight  $m_{ij}$  is derived by summing the excitation received from all other weights:

$$R_{ij} = \sum_{kl} w_{ijkl} m_{kl}$$

The excitation weight  $w_{ijkl}$  equals 1 if  $m_{ij}$  and  $m_{kl}$  correspond to immediately adjacent and compatible mapping connections (e.g.,  $a_k$  plays role  $r$  in  $a_i$ , and  $b_l$  plays role  $r$  in  $b_j$ ), and it equals 0 otherwise. The raw evidence is filtered through additional inhibitory mechanisms that encourage one-to-one mappings, and the result determines the incremental change to the mapping weights. These dynamics continue until all mapping weights converge to 0 or 1.

Following the MAC/FAC model of analogical retrieval (Forbus et al, 1995), APEC uses a measure of structural match to determine the quality of an analogy. An initial score is assigned to every matched pair of nodes to enforce a size preference. A preference for deep analogies (systematicity) is implemented via a trickle-down method, whereby initial match scores are passed down to increment the scores of matching components. The match scores are summed to get a global measure of structural match quality.

## Predication

If the analogy achieves a minimum match quality, a schema is induced that represents the structural commonalities of



the analogues and encodes the shared pattern of role binding embodied by the analogy. Specifically, an abstract entity is created for every mapping weight in the analogy, and these entities are role-bound to each other if the corresponding entities in the analogues are so bound. Once created, schemas are treated identically to episodes (they are just more abstract). This simple mechanism is drawn from prior work on schema induction (Gick & Holyoak, 1983; Doumas et al., 2008; Kuehne et al., 2000). A schema can be thought of as codifying the higher-order relation embodied by the analogy, hence turning it into an explicit predicate. Aligning the schema with any new episode enables a test of whether that episode instantiates the higher-order relation.

## Evaluation

When a schema is retrieved and compared to a new episode, it is refined, by abstracting the common structure between schema and episode (Doumas et al., 2008). This process is a form of intersection discovery, where the intersection of the set of relations in a schema and episode are encoded as a new schema. In this way, schemas shrink in size because the variability between episodes is abstracted over, leaving only the structure that is consistent across episodes. However, there may also be a need for schema elaboration, where schemas can grow in size (Corral & Jones, in press). We are currently exploring implementing schema elaboration in the model.

In parallel with schema refinement, schemas are evaluated as candidates for consolidation as new relational concepts. New concepts are useful because they can facilitate generalization. Learning about one instance of a concept can be applied to other instances. Jones & Cañas (2010) show how representations can be learned by improving generalization in a reinforcement-learning framework. The basic idea is that reward prediction error (TD error) can be used to determine when generalization from some past stimulus to the current stimulus was or was not helpful. In the present context, if a learner encounters an episode that is alignable with some stored schema, then analogical inference from that schema enables generalization from past instances of that schema. If this inference leads to improved prediction or behavior, then the schema is strengthened, and if not it is weakened. This process tunes generalization to depend more on higher-order relations that are predictive and less on those that are not.

## Consolidation

Relational consolidation is the process of a schema becoming chunked into a unitary concept that can be recognized automatically, retrieved from memory in parallel, and represented as an element of yet-higher-order relations. As summarized in Table 1, consolidation is hypothesized to confer properties to a concept that are not true of (unconsolidated) schemas, because consolidated concepts are recognizable perceptually, without explicit (working-memory dependent) structure mapping.

Table 1. Predicted consequences of consolidation

Not Consolidated	Consolidated
More affected by WM demands	Less affected by WM demands
Quicker at analogical inference, because structure mapping is active	Easier to learn higher-order structure, because instances can be represented by tokens
Serial retrieval	Parallel retrieval

It is important to note that consolidation is not a change in the declarative knowledge embodied by a concept. Rather, it is a proceduralization of the concept that enables future changes in knowledge – similar to the interaction between declarative and procedural knowledge in production systems (Anderson & Lebiere, 1998).

The DORA model of relational predication (Doumas et al., 2008) has an operation similar to relational consolidation, in which it recruits a new proposition node to bind lower-order relations. This new node can be bound to roles of yet-higher-order relations, but the relation is still explicitly structured. In contrast, consolidation might be viewed within the DORA framework as enabling the new proposition node at the top of the relational structure to evolve into a new semantic node at the bottom. We conjecture this difference has important implications for recognition and retrieval of instances of the concept.

Relational consolidation is best explained from the perspective of the MAC/FAC model of analogical retrieval (Forbus et al, 1995). MAC/FAC embodies the assumption that verifying the lower-level elements of an episode (i.e., objects and relations) is fast and automatic, whereas verifying relational structure is slower and requires working-memory resources. In the first stage of analogical retrieval (Many Are Called), the target episode is converted to a flat feature vector that is used to probe all episodic memories in parallel. Importantly, the dimensions in the MAC feature vector are predefined, based on the concepts the learner currently knows. Stored episodes that share content (objects and relations) with the target are retrieved, without regard for how those objects and relations are connected by role binding. In the second stage (Few Are Chosen), the episodes retrieved by the MAC stage are filtered by structural alignment to the target. Only those episodes that are alignable with the target survive this stage.

From the perspective of the MAC/FAC framework, relational consolidation enables a higher-order relation to be chunked and treated as a dimension of the feature vector used for memory probing. Prior to consolidation, retrieval of instances of a higher-order relation require something like the FAC stage, in which subjects explicitly map between those instances and the schema. Following consolidation, retrieval can rely solely on the MAC stage, thus operating much more rapidly and without requiring working memory. We also propose a similar difference for perceptual recognition of instances of the concept in the

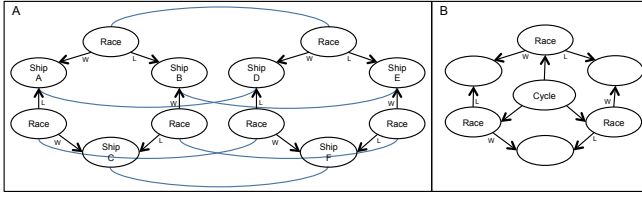


Figure 1: The two ways a predicated relation can be represented or recognized. (A) Before consolidation, episodes must be structurally aligned to a schema. (B) After consolidation, an instance of the concept is explicitly represented and bound to the lower-order relations. The labels on the nodes refer to Experiments 1 & 2.

environment, through the creation of a perceptual detector for each consolidated concept. Figure 1 illustrates the difference between recognizing an instance of a higher-order relation that has versus has not been consolidated.

## Experiments 1 & 2

The goal of the present experiments was to test learning of categories defined purely by higher-order relations. That is, the set of objects and relations present in instances of each category were identical; only the way the relations were linked into a higher-order structure differed. If people can learn this type of category distinction, it would support our basic proposal for how higher-order relations are defined in constructing relational hierarchies.

Each trial showed animations of three spaceships racing each other in pairs. The categories were defined by the two logically possible structures these races could form: a *cycle* (e.g., A beats B, B beats C, C beats A) and an *ordering* (e.g., B beats A, B beats C, C beats A). The races are thus first-order relations between spaceships, and cycle and ordering are the possible higher-order relations (see Figure 1 for an example of the cycle structure).

According to APEC, three types of learning potentially contribute to this task. First, analogical alignment between episodes (trials) leads to induction of schemas capturing their common structure or properties. Some of these schemas will capture the true category structure, whereas others will be based on irrelevant information (e.g., ship color or spatial position). Second, feedback following each trial is used to strengthen or weaken schemas that contributed to each response, so that eventually the correct higher-order relations should come to dominate performance. Third, with sufficient learning, the correct higher-order relations may become consolidated. The results reported below primarily bear on the first of these mechanisms.

## Methods

110 and 62 undergraduates participated in Experiments 1 and 2, respectively. Subjects were told they would observe alien spaceship races, in sessions of three races each. The aliens were said to have two names for possible outcomes of a session, and the subject's task was to learn their meaning.

On each trial, three spaceships (differing only in color) raced in pairs in a sequence of three races. The subject classified the session as "Dekal" or "Koplu" by typing D or K. The correct answer was then displayed. The experiment lasted until the subject met a learning criterion of 8 out of 10 trials correct, or until 25 minutes elapsed (50-70 trials).

Each experiment included two orthogonal manipulations designed to bias attention between objects (spaceships), relations (races), and higher-order relations. In Experiment 1, the trials were described as either "tournaments" or "sessions". In addition, the main task was preceded by a series of footraces among 5 cartoon characters, after which subjects were asked either which character had done best overall or which of two characters had won a specific race. The tournament label and overall-winner question were predicted to make rankings salient, thus shifting attention to higher-order relations.

In Experiment 2, half the subjects were given a first-person perspective by adding a gold star to mark "your ship" on each trial. In addition, the three colors used for the spaceships were either constant or variable from trial to trial. The marked ship and constant colors were predicted to make individual objects more salient, thus shifting attention away from higher-order relations.

## Results

There were no differences across conditions in either experiment in meeting the learning criterion or number of trials to criterion. All subsequent analyses are based on collapsing the groups of both experiments.

The results demonstrate that subjects could learn the difference in categories. 113 subjects (65.7%) met the learning criterion. Figure 2A shows the distribution of trials to criterion for these subjects ( $M = 30.9$ ). Ten subjects learned without making a single error; they were excluded from the remaining analyses, which are based on errors.

Figure 2B shows a backward learning curve, aligned on each subject's last error (for subjects who solved the task but made at least one error). Performance prior to the last error is only slightly above chance, indicating learning was nearly all-or-none. This is consistent with our hypothesis that learning is triggered by inducing the correct schema.

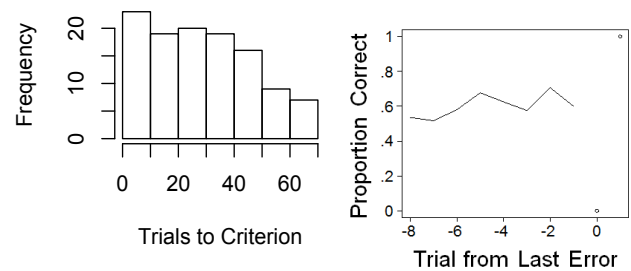


Figure 2. A: Distribution of trials to criterion, for subjects solving the task. B: Backward learning curve.

If learning depends on analogical mapping, then it should be most evident following consecutive stimuli in the same category (assuming subjects most often compare stimuli from consecutive trials). Assuming the schema was induced following the trial of the last error ( $t_{\text{last}}$ ), this predicts the stimuli on trials  $t_{\text{last}} - 1$  and  $t_{\text{last}}$  should tend to be of the same category (see Figure 3). This prediction holds for 59 of the 103 subjects ( $p = .084$ , one-tailed binomial test). If we relax the all-or-none assumption and examine performance of all subjects on all trials  $t$ , we find a highly reliable advantage when trials  $t - 2$  and  $t - 1$  are of the same category (mean difference 4.3%, paired  $t_{169} = 3.42$ ,  $p < .001$ ).<sup>2</sup> This suggests comparison of recent stimuli from the same category had a significant effect on performance.

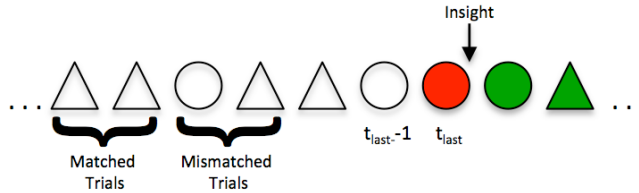


Figure 3. Example sequence of trial types and the most likely moment of schema induction. Triangles indicate ordering trials and circles indicate cycle trials.

## Discussion

Analogy and metaphor are pervasive in cognition (Hofstadter, 2001; Lakoff, 1980) and play a critical role in abstract reasoning. The past three decades of research have led to a strong consensus that analogy hinges on recognition of common relational structure between two or more situations (e.g., Gentner, 1983; Hummel & Holyoak, 2003). This suggests that acquisition of new higher-order relations plays a critical role in human conceptual development.

We propose that many (if not most) abstract concepts exist in relational hierarchies, in which entities are at once relational structures among their components and elements of higher-order relations. The structure-mapping process of analogy (Gentner, 1983) can be viewed as a search for relations among relations, in the sense of how relations are connected to one another by operating on the same objects. Successful analogies—those that lead to useful inferences or predictions—might thus be treated as candidates for new relational concepts.

Our approach builds on models of analogical retrieval (Forbus et al., 1995), structure mapping (Falkenhainer, et al., 1989; Hummel & Holyoak, 1996; Larkey & Love, 2003), predication, and refinement (Doumas et al., 2008). Importantly, our model goes beyond previous models of schema induction (Doumas et al., 2008) by positing relational consolidation as a means for learning new relational concepts. A further contribution of our approach is embedding predication within a reinforcement-learning framework in order to modify analogical similarity, and thus

generalization, by representation change (Jones & Cañas, 2010; see also Tomlinson & Love, 2006).

Taken together, these ideas lead to a model, APEC, which iterates the Analogy, Predication, Evaluation, and Consolidation stages to build relational hierarchies in a long-term conceptual learning system. Our eventual aim for the model is a system that can autonomously discover useful structure in its environment by construction of these relational hierarchies.

Although the present experimental results indicate a role for analogical comparison and schema induction, we do not have strong evidence here for consolidation. Indeed, we believe it more likely that the categories were learned only as schemas. The experiments do provide a test of one fundamental assumption of the model: that people can learn higher-order relations defined solely by the configuration of shared role binding among lower-order relations.

Future experiments could build the tournaments into third-order structures that require consolidation of the tournament type in order to solve the task. Another future experiment is to test transfer of learned relational structure to an alternate domain with different lower-order relations. The lack of effect of our manipulations suggest it is an open question what factors influence the kind of learning tested in these experiments.

Evidence for relational consolidation could also come from process dissociation between the two modes of representation outlined in Table 1. Other evidence may come from neurological studies. We speculate that relational consolidation is implemented neurally by a process of hippocampal-to-cortical feedforward training, in line with models of episodic memory consolidation (Gluck & Myers, 1993; McClelland, McNoughton, & O'Reilly, 1995). The hippocampus is well suited for storing schemas, which are inherently structured, given the conjunctive and localist nature of hippocampal representations.

Relational consolidation is similar to the career-of-metaphor hypothesis, by which a metaphor is originally an analogical mapping between the base and target domain, but it can become conventionalized so that the target is recognized immediately as an instance of the base category (Gentner et al., 2001). This transition from novel to conventional metaphor resembles the transition from non-consolidated to consolidated relations, in that both involve a transition from recognition via structure mapping to more automatic, perceptual recognition. The major difference is that in the career of metaphor, the base concept was already consolidated, and the conceptual change is a form of sense extension of that base concept. The conventionalization process does not create new concepts; it just extends their meaning. Therefore it does not function to build up relational hierarchies. Nevertheless, the two ideas seem intimately linked, in that the career-of-metaphor mechanism might play an important role in extending and refining the meaning of concepts after they have been consolidated.

Language is almost certainly an important factor in relational consolidation. Thompson et al.'s (1997) finding

<sup>2</sup> Two additional subjects were excluded because they experienced no alternation trials before meeting the learning criterion.

of chimps learning higher-order relations depended on initial training with material tokens. In humans, words can act as linguistic tokens (Clark, 2006) and have been shown to aid category learning (Lupyan, Rakison, & McClelland 2007). Son, Dumas, and Goldstone (2010) offer two possible roles of language in relational learning: “(1) words invite learners to compare, highlight, and represent relations (the Generic Tokens [GT] hypothesis), and/or (2) words carry semantic cues to common structure (the Cues to Specific Meaning [CSM] hypothesis)” (p. 55). The lack of significant effect of the CSM manipulation in our Experiment 1 could be explained by the cue word appearing only at the start of the experiment, or by the relatively subtle difference in semantics evoked by the cues “tournament” vs. “session”. A stronger test of CSM would be to cue subjects with category labels whose meanings structurally match or mismatch the category schema.

Finally, we do not claim that relational consolidation is the only mechanism for acquiring new relational concepts. Research on basic-level objects (Rosch et al., 1976) suggests there are truly primitive object concepts that are not originally constructed as relational systems. Clearly a lot of discovery comes from analyzing the substructure of objects, and that process should be included in any complete model. For example, categories can be induced for objects that fill the same roles (Jones & Love 2007). Although we have been working on concept learning through mechanisms of synthesis, a future goal is to explore how combinations of analytic and synthetic mechanisms of relational learning might be more powerful than both alone.

### Acknowledgments

This work was supported by AFOSR Grant FA-9550-10-1-0177 to MJ.

### References

- Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.
- Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, 10(8), 370–374.
- Corral, D., & Jones, M. (in press). Learning of relational categories as a function of higher-order structure. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.
- Dumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, 115(1), 1–43.
- Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, 41(1), 1–63.
- Forbus, K. D., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19(2), 141–205.
- Gardner, M. (1970). Mathematical Games - The fantastic combinations of John Conway's new solitaire game "life". *Scientific American*, 223, 120-123.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2), 155–170.
- Gentner, D., Bowdle, B., Wolff, P., & Boronat, C. (2001). Metaphor is like analogy. In D. Gentner, K.J. Holyoak, & B.K. Kokinov (eds.), *The analogical mind: Perspectives from cognitive science*, 199–253.
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive psychology*, 15(1), 1–38.
- Gluck, M. A. & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, 117, 227-247.
- Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, 3(4), 491–516.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110(2), 220.
- Jones, M., & Canas, F. (2010). Integrating reinforcement learning with models of representation learning. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.
- Jones, M., & Love, B. C. (2007). Beyond common features: The role of roles in determining similarity. *Cognitive Psychology*, 55(3), 196–231.
- Kuehne, S., Forbus, K., Gentner, D., & Quinn, B. (2000). SEQL: Category learning as progressive abstraction using structure mapping. *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society* (pp. 770–775).
- Larkey, L. B., & Love, B. C. (2003). CAB: Connectionist analogy builder. *Cognitive Science*, 27(5), 781–794.
- Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not Just for Talking. *Psychological Science*, 18(12), 1077.
- McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3), 419.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, 8(3), 382–439.
- Son, J. Y., Dumas, L. A. A., & Goldstone, R. L. (2010). When Do Words Promote Analogical Transfer? *The Journal of Problem Solving*, 3(1), 4.
- Squire, L. R. & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current Opinion in Neurobiology*, 5, 169-177.
- Thompson, R. K. R., Oden, D. L., & Boysen, S. T. (1997). Language-naïve chimpanzees (Pan troglodytes) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes*, 23(1), 31-43.
- Tomlinson, M., & Love, B. C. (2006). Learning abstract relations through analogy to concrete exemplars. *Proceedings of the 28th annual conference of the Cognitive Science Society* (pp. 2269–2274).

# “Eyes Closed” and “Eyes Open” Expectations Guide Fixations in Real-World Search

Tom Foulsham (foulsham@essex.ac.uk)

Department of Psychology, University of Essex  
Wivenhoe Park, Colchester, CO4 3SQ, UK

## Abstract

Investigations of search within realistic scenes have identified both bottom-up and top-down influences on performance. Here, I describe two types of top-down expectations that might guide observers looking for objects. Initially, likely locations can be predicted based only on the target identity but without any visual information from the scene (“Eyes closed”). When a visual preview becomes available, a more refined prediction can be made based on scene layout (“Eyes open”). In two experiments participants guessed the location of a target with or without a brief preview of the scene. Responses were consistent between observers and were used to predict the eye movements of new observers in a third experiment. The results confirm that participants use both types of top-down cues during search, and provide a simple method for estimating these expectations in predictive models.

**Keywords:** scene perception; visual search; attention; eye movements

## Introduction

Imagine walking into a friend’s kitchen to look for a coffee mug when you have never been there before. Even before you open the door you would already have a significant amount of knowledge about where this object might be. For example, you would not expect it to be on the floor or near the ceiling, so you would be unlikely to look in these locations. When entering the room, the first glance tells you that, in this particular kitchen, there is a large window on one side of the room and shelving with cupboards on the opposite side. You refine your expectations about where the object will be and subsequently recognize the mug on a shelf.

As this scenario reveals, searching for something in the real world involves not just the matching of visual information to a stored template but also the use of detailed semantic knowledge about scenes and objects. Although visual search has been very well studied in cognitive psychology, this has mostly been in the context of simple search displays and models that predict performance based on target features (e.g., Wolfe, 1998). More recently, there has been significant interest in exploring the mechanisms involved in directing attention during search within natural scenes, and in testing these mechanisms by measuring eye fixations.

In one approach, computational models of bottom-up visual salience have been proposed that select targets based on the degree to which they stand out from their background (Itti & Koch, 2000). By this account, participants will attend

to regions of high salience until they find what they are looking for. However, it is clear from several experiments that, when searchers know what they are looking for, this knowledge, and not simple visual salience, dominates the locations inspected during search (Chen & Zelinsky, 2006; Foulsham & Underwood, 2007).

This has led to more realistic models that combine top-down knowledge of the searcher to prioritize those locations in a scene in which an object is likely to appear. One way to do this is to compare scene locations with a representation of target *appearance*. If one knows that the target is red, locations with this colour should be more likely to be fixated. This principle underlies several models of search guidance (Wolfe, 1994; Navalpakkam & Itti, 2005), and can be successful at predicting fixations in real scenes (Zelinsky, 2008; Kanan et al., 2009).

However, it is clear from the scenario at the beginning of this paper that searchers also have access to detailed expectations about target *location*. There is evidence from several different experiments that these expectations are used to direct attention during search. For example, scrambling an image—so that local visual features remain the same but their configuration is altered—impedes search and alters eye movements (Biederman et al., 1973; Foulsham, Alan & Kingstone, 2011). Objects that are incongruent with their context or out of place may be found more slowly (Henderson, Weeks & Hollingworth, 1999). The contextual guidance model proposed by Torralba et al. (2006) accounts for these effects by combining bottom-up salience with a Bayesian prior for where an object is likely to be, conditioned on the global features of an image. In essence, the model recognizes the gist and layout of a scene (e.g., finding street level in an urban environment), learns the target’s likely location within this representation, and searches accordingly (e.g., by looking for people at street level).

The top-down guidance by context discussed so far emerges early, with the first glance of a scene, but requires visual input in the form of low spatial-frequency features and “gist”. On the other hand, it is likely that the semantic information associated with different objects might include general expectations about position within a scene-centered or person-centered frame-of-reference which could be activated *before* exposure to the to-be-searched scene. The present paper investigates whether these expectations are reliable and whether they effect the distribution of attention in real-world search. If so, they could be incorporated into probabilistic models (e.g., Kanan et al., 2009; Torralba et al., 2006).



I will distinguish between “Eyes closed” expectations, which can be made prior to any perception of the scene, and “Eyes open” expectations, which are affected by a rapid perception of scene gist, as might be available during the first fixation on a scene. I describe two simple experiments to quantify “Eyes closed” and “Eyes open” predictions, and these predictions are then compared to the eye fixations made by independent searchers. If contextual guidance of attention occurs only in response to scene features then “Eyes closed” expectations will not be a good description of where people look during search.

## Experiments 1 and 2

In Experiments 1 and 2, participants guessed where a target object would be located based on very little information.

### Method

**Participants** Eighteen student volunteers (12 females) took part in return for course credit. All participants took part in Experiment 1 first, followed by Experiment 2. The mean age was 19.4 years.

**Stimuli and Apparatus** The stimuli for all experiments were derived from 72 colour photographs of indoor and outdoor scenes collected from the Internet. Scenes were chosen which contained a single example of an easily nameable target object that was not located directly in the centre of the image. The name of this object was the matching target label for the scene.

Each target label was also matched to another scene from the set in which it could plausibly be found. This led to 144 label-scene pairs, half of which were “target-present” trials, where the scene contained the target, and half of which were “target-absent”. The same target labels were used in both experiments.

Target labels were presented in large black font centred above a grey rectangle representing the scene. In Experiment 2, scene images were presented at a resolution of 1024 x 768 pixels. Stimuli were presented on a 19-inch CRT monitor with a refresh rate of 60Hz. Presentation was controlled by PsychoPy (Pierce, 2007), and responses were entered with the mouse.

**Procedure** Figure 1 depicts the procedure. In Experiment 1, participants were instructed to “make their best guess” where a target was located in an image. The experiment began with a practice example of a target and scene. In the experimental trials, a target label was presented alongside a grey rectangle representing the image frame, and participants were instructed to click with a mouse cursor where in the frame they thought the target was located. In order to motivate participants, feedback was given after every 12 trials in the form of a percentage score representing how close their mouse clicks had been to the actual target locations.

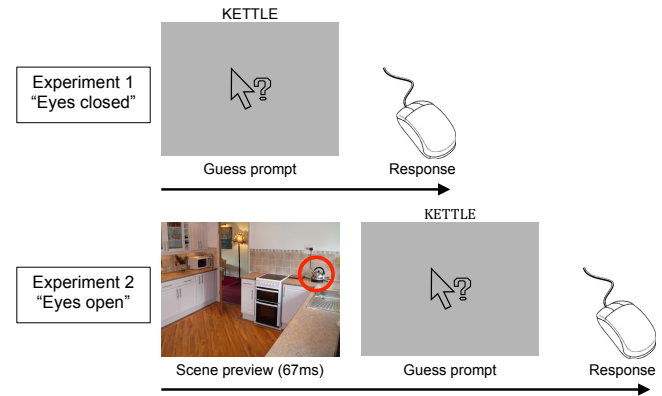


Figure 1: The procedure for one trial in Experiment 1 (top) and Experiment 2 (bottom). The target is highlighted in the scene preview, for display purposes only.

Scores were calculated from the average Euclidian distance between the chosen location and the centre of the target object, in target present trials only, and normalized by the scene diagonal. All 72 target labels were presented in a random order, but the actual scenes were not shown to participants.

In Experiment 2, participants were given a brief preview of the scene in which the target was located before they made their response. In each trial, a text prompt told participants to get ready, and a fixation cross was then presented in the centre of the screen for 1s. The scene was then presented briefly for 67ms, followed by a target label presented alongside a grey rectangle representing the image frame. The brief preview was chosen because it is known that scene gist can be perceived very quickly (Biederman et al., 1973), and also to limit the possibility that targets would be attended during the preview. A pattern mask was not included, and so after-images may have persisted, although the guess prompt had the effect of partly masking the display and drawing attention away from the scene. As in Experiment 1, participants were instructed to guess the location of the target with a mouse click, and feedback was given regarding average performance over the previous 12 trials. In Experiment 2, all 144 label-scene pairs were presented in a random order.

### Analysis and Results

The results were analysed in order to estimate the inter-observer agreement, i.e., the degree to which different participants “guessed” in similar locations for each target label. The approach used closely followed that in previous studies of fixations in real-world search (Torralba et al., 2006; Ehinger et al., 2009). Participant-selected locations were first combined to produce a spatial model of target predictions—an “expectation map”—which was then used to predict search behaviour in Experiment 3.

Expectation maps were formed by representing each participant’s guessed location as a 2-dimensional Gaussian and summing across all participants. The highest points on this map indicate locations that, according to the

participants, are most likely to contain the target. The dispersion of the map will reflect between-participant agreement: maps that cluster into a few small areas signify that participants agreed on where a target was likely to appear. Maps were computed separately for each target label in Experiment 1, and for target-present and target-absent trials in Experiment 2.

The receiver operating characteristic (ROC) curve was used to evaluate expectation maps. The ROC curve is a non-parametric measure of sensitivity originating from signal detection theory. This measure has become common in machine learning, and also in studies of spatial attention and eye movements, as it allows spatial distributions (e.g., salience maps) to be compared to specific locations (e.g., eye fixations). Full coverage of this method can be found elsewhere (e.g., Ehinger et al., 2009). The area under the curve (AUC) was computed as a summary statistic. AUC values indicate the probability that the map will rank a selected location more highly than a non-selected location and range from 0 to 1, with a score of 0.5 indicating chance performance.

**Inter-Observer Agreement** An all-except-one method was used to compute the between-participant agreement for each expectation map. In this analysis, a map was computed based on the responses of all participants except one, and the ROC curve was used to evaluate how well this model predicted the location chosen by the remaining participant. This process was repeated for all participants, and the mean AUC value obtained is an indicator of the inter-observer agreement in guessed locations.

Figure 2 shows a target expectation map for two example target labels from Experiment 1. The first is from a target on which participants showed considerable agreement, while the mouse clicks in the second are more distributed. Table 1 summarizes the between-participant AUC scores across all targets in Experiments 1 and 2. Critically, all the scores are much greater than 0.5, confirming that participants were indeed consistent in the points that they chose. This was true for Experiment 2, where participants saw a brief preview of the search scene, and also for Experiment 1, when participants guessed (“eyes closed”) with only the target identity to go on.

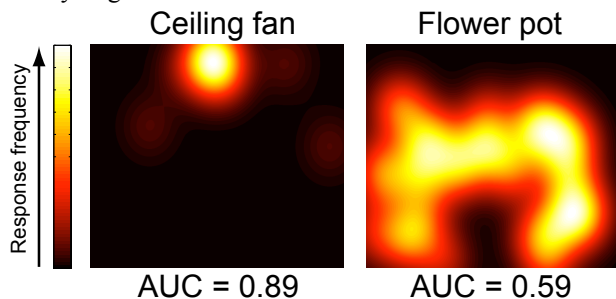


Figure 2: Expectation maps for two targets in Experiment 1. AUC scores represent the inter-observer agreement, which is high for one target (left) and much lower for the other (right).

The targets used were distributed throughout the scene, and could appear anywhere. However, some of the agreement may have originated because, across all trials, objects and mouse clicks were more likely to be in some locations (such as the centre of the image) than others. To control for this, an additional “between-target” analysis was performed using the method described above but with the responses associated with each object used to predict those from a *different* target (e.g., how well do guesses for the location of a flower pot predict those for a ceiling fan or a TV?). This control analysis will therefore quantify convergence that is independent of the particular target. This between-target control was higher than 0.5, probably because some objects were in a similar position. Importantly, inter-observer agreement was significantly higher than the between-target control AUC, in both Experiments (all  $t(71) > 3.8$ ,  $p < .001$ ).

Table 1: Inter-observer agreement in target guesses in Experiments 1 and 2. AUC values give the mean and standard deviation across all targets.

Trial type	AUC	
	Mean	SD
Experiment 1: “Eyes closed”		
All trials	0.79	0.07
Experiment 2: “Eyes open”		
Target-present	0.88	0.08
Target-absent	0.83	0.08
Between-target control	0.71	0.1

Inter-observer agreement was significantly higher in the preview Experiment 2 than in Experiment 1. Moreover, this was the case in both target-present scenes (where participants could have, in theory, perceived the target object during the preview;  $t(71)=5.6$ ,  $p<.0001$ ) and in target-absent scenes (where there was no target to find;  $t(71)=3.5$ ,  $p<.001$ ). In other words, exposure to a brief glimpse of a scene made participants more likely to predict the same location for an object, even when that object was not present. Figure 3 shows the expectation map for the target label “TV”, from responses in Experiment 1 (where participants made a blind guess) and for a target absent trial in Experiment 2 (where participants saw the depicted preview scene which did not contain a TV). Participants responding in Experiment 2 changed their guesses considerably and focused on a spot where a TV might appear.



Figure 3: Expectation maps for the target label “TV” when guessing in Experiment 1 (left) and in Experiment 2 (right, superimposed over the preview scene that was shown).

### Experiment 3

Experiments 1 and 2 confirmed that people were consistent in their expectations about where a named target would be likely to appear in a real world scene. The target maps provide a simple way of representing these expectations. Experiment 3 tested whether the eye movements of a new group of observers could be predicted from the target guesses.

#### Method

**Participants** Eighteen new participants (12 females), who had not taken part in Experiments 1 and 2, were recruited for payment. All participants had normal or corrected-to-normal vision, and their mean age was 22.5 years.

**Stimuli and Apparatus** The same set of target labels and scenes was used as in the previous experiments. To avoid trial-to-trial learning, each scene was presented once only, with half of the scenes containing the target and half without (i.e. matched with a different target label not present in the scene as in Experiment 2). Across participants, each scene appeared in both target-present and target-absent conditions.

Stimuli were presented on a 19-inch monitor positioned 60cm from the observers. Participants rested on a chin-rest, which ensured a constant viewing distance and restricted head movements. Scene images filled the screen, subtending 33 x 26 degrees of visual angle at this viewing distance.

Eye movements were recorded during search using the EyeLink 1000 system (SR Research), which used a desk-mounted camera to record monocular eye position from a video image of the pupil and the corneal reflection. This eye-tracker has a high spatial resolution (error of less than 0.5 degrees on average) and captured eye position at 1000Hz. Samples were parsed into oculomotor events using the EyeLink system’s default algorithm, which identifies saccades and fixations based on velocity and acceleration thresholds. Search responses were entered via a button box.

**Procedure** The experiment began with an eye-tracker calibration (using a nine-dot grid), followed by instructions and 8 practice trials. The experimental trials followed a standard visual search procedure. First, a target label was presented, written in black font in the centre of the screen

for 1s. This was replaced by a fixation point presented in the centre of the screen and participants pressed a button to proceed with the search. At this time the eyetracker checked that fixation was on the centre. The search scene then appeared, and participants were told to press one of two buttons as quickly and accurately as possible to identify whether or not the target was present in the scene. The search response terminated the trial, which ended with a blank screen for 500ms. All 72 trials were presented in the same way, in a random order, and the eye-tracker was recalibrated at the halfway point.

#### Results

**Search Performance** Participants responded accurately on a mean of 89% of all trials. In correct, target-present trials, the mean reaction time was 1350ms (standard error of the mean, SEM = 134) and participants made 5.5 fixations on average, per trial (SEM = 0.4). As with most visual search tasks, target absent trials were responded to more slowly ( $M = 2063\text{ms}$ ,  $\text{SEM} = 237$ ) and with more fixations per trial ( $M = 7.9$ ,  $\text{SEM} = 0.7$ ). Figure 4 gives an example of the locations fixated during a trial.

### TV - Search

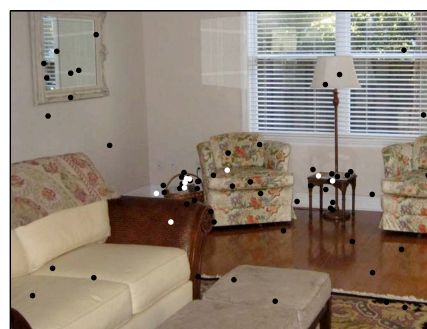


Figure 4: The locations of fixations made by all participants searching for the target “TV” in a target-absent trial. White markers indicate the first fixation in the trial.

#### Predicting Fixation Locations From Expectation Maps

The remaining analyses aimed to assess whether fixation locations in the visual search task could be predicted based on the expectation maps derived from each target in Experiments 1 and 2. As previously, an ROC approach was followed. For each target-scene pair, the analysis asked how well the expectation maps formed from guesses could discriminate between fixated and non-fixated locations. Because it was anticipated from theory and previous experiments that attentional priorities might change over time, separate ROC curves were computed from each participant’s first saccade target (i.e., the location of the first fixation away from the central starting point) and from all fixations in the trial. It is also essential to take into account the general tendencies for fixations (and probably mouse clicks) to be located near to the centre of the image and away from the scene edges.



Table 2: Predicting fixation locations from the guessed locations in Experiments 1 and 2.

Prediction	Trial type	AUC (all saccades)		AUC (first saccade)	
		M	SD	M	SD
Experiment 1: “Eyes closed”	Target-present	0.67	0.12	0.71	0.16
	Target-absent	0.68	0.13	0.73	0.17
	Between-trial control	0.62	0.05	0.67	0.05
Experiment 2: “Eyes open”	Target-present	0.82	0.12	0.83	0.13
	Target-absent	0.76	0.11	0.79	0.10
	Between-trial control	0.64	0.07	0.72	0.11

Therefore, following Ehinger et al. (2009), I computed a between-trial control comparison where the expectation map for one target/scene was used to predict the fixations made while searching for a different object.

Table 2 displays AUC summary statistics for the comparison between expectation maps and fixated locations. There were several noteworthy results. First, all the AUC values are greater than 0.5, showing that fixation locations could be predicted on the basis of the mouse responses made in Experiments 1 and 2. Moreover, in all trial types, the observed AUC values are greater than the between-trial control estimate. This was statistically reliable across the different target/scene pairs (all  $t(71) > 2.6$ ,  $ps < .01$ ) and confirms that the results cannot be attributed to general spatial biases.

In addition, both expectation models were better predictors of the first saccade in a trial than they were of all saccades. This may be because the initial saccade was most likely to move toward the expected location, whereas later saccades might be exploring different areas of the picture. However, the between-trial control also led to higher AUC values when only the first saccade was evaluated, so it seems the first saccade is more predictable *in general*. This might be because of a strong central bias in scene viewing which tends to decrease over time, particularly when viewing starts in the centre (as it did here).

Most importantly, the guesses made by participants who saw a brief preview of the scene (“Eyes open”) were a significantly better predictor of fixation locations than those who guessed blindly without seeing the scene. The best performance came in target-present trials, which indicates that participants in Experiment 2 had seen the target at least some of the time when guessing. Searchers in Experiment 3 were obviously highly likely to look at this correct target location, whereas there was more variability in target-absent images. However, it is important to note that, even when there was no target, the “Eyes closed” guesses of an independent group of participants were a significant predictor of fixation.

**Predicting Between-target Variation** An additional question concerns the relationship between expectations and search performance. If target objects are strongly associated with a particular location then we would expect a

considerable amount of inter-observer agreement in the expectation maps (e.g., compare the two maps in Figure 2). If these expectations are an important factor in real world search, then the inter-observer agreement should correlate with reaction time in Experiment 3.

The mean reaction time was calculated across all participants for each correct, target-present trial, and then correlated with the AUC values representing inter-observer agreement from Experiments 1 and 2. In both cases there was a negative correlation (see Figure 5). When participants were more consistent in their guesses about where an object would appear, this object was found more quickly. The correlation with “Eyes closed” guesses approached significance ( $r = -.21$ ,  $p = .08$ ), while the correlation with “Eyes open” guesses was larger and statistically reliable ( $r = .50$ ,  $p < .001$ ).

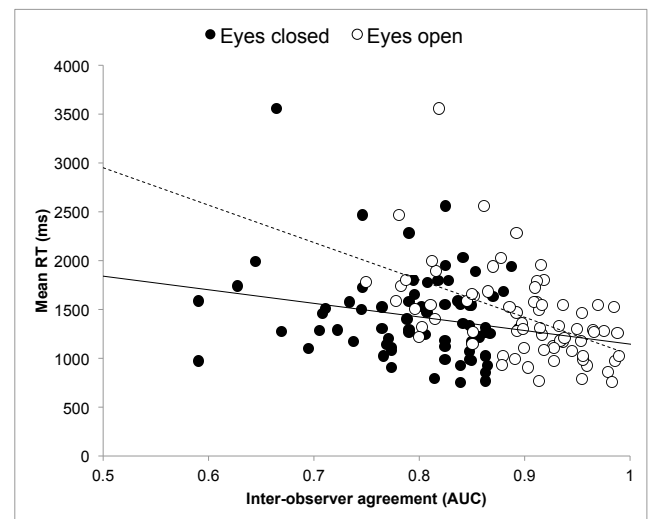


Figure 5: The correlation between inter-observer agreement and search RT. Each data point represents a target/scene pair from Experiments 1 or 2, with least-squares regression lines.

## Discussion

In this paper I have proposed a distinction between the different types of top-down information available in guiding search in real-world scenes. Unlike previous descriptions of contextual effects (Biederman et al., 1973; Torralba et al., 2006), I specifically emphasized the fact that some predictions based on semantic knowledge can be made prior to the onset of the search scene. There were several interesting findings, which point to promising future directions for this approach.

First, participants showed a reliable amount of agreement when asked to blindly guess the target location. Although participants initially found this task unusual they were able to do so quickly and often chose the same locations for an object. There was some variation between different objects, with objects showing the largest amount of agreement those which are strongly constrained to spatial locations (such as light fittings). The method described here could be used in further research to characterise different search objects and their effects on performance. It should be noted that, because the present studies were limited to a fixed image frame on a monitor it mainly measured knowledge about picture composition (e.g., where the horizon is likely to be in a scene). How participants use such information in real life, where frames of reference change with head and body position, remains an open question and could be explored by looking at attention in active, real-world environments (see Foulsham, Walker & Kingstone, 2011).

Second, “Eyes closed” predictions were at least partly separable from those made in response to a preview of the scene. A brief preview of the scene gist, prior to seeing the target label, was enough to increase agreement between observers, even when there was no target to find. In other words, additional information about the scene was used by participants in a consistent way. It would be interesting to determine some of the cues that participants are responding to in this situation, as they could potentially be both appearance-based (selecting something which looked like the target) and location-based (selecting a region where the target might reasonably occur).

Third, the guesses of the participants in Experiments 1 and 2 were reliable predictors of fixation locations in independent searchers in Experiment 3. This was true when participants guessed based on a brief preview of the scene, which confirms that searchers look towards the parts of the scene which are contextually relevant given the gist. This finding, in both target-present and target-absent scenes, is similar to that reported by Ehinger et al. (2009), who used a “context oracle” defined by the responses of independent observers predicting the y-coordinate where pedestrians should occur in street scenes. However, what is surprising in the current experiments is that, even without the scene, participants are able to predict target locations, and these predictions are reflected in fixation behavior. In future work this could be modeled by positing a “blind” statistical prior which could then be refined according to global features

such as those in the contextual guidance model of Torralba et al., (2006).

## References

- Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology*, 97, 22–27.
- Chen, X., & Zelinsky, G. J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research*, 46(24), 4118–4133.
- Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, 17, 945–978.
- Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual saliency in scene perception? *Perception*, 36, 1123–1138.
- Foulsham, T., Alan, R. & Kingstone, A. (2011). Scrambled eyes? Disrupting scene structure impedes focal processing and increases bottom-up guidance. *Attention, Perception and Psychophysics*, 73 (7), 2008–2025.
- Foulsham, T., Walker, E. & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, 51 (17), 1920–1931.
- Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal Of Experimental Psychology: Human Perception & Performance*, 25, 210–228.
- Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10–12), 1489–1506.
- Kanan, C., Tong, M. H., Zhang, L., & Cottrell, G. W. (2009). SUN: Top-down saliency using natural statistics. *Visual Cognition*, 17, 979–1003.
- Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, 45(2), 205–231.
- Peirce, JW (2007) PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1–2), 8–13.
- Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4), 766–786.
- Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, 1, 202–228.
- Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science*, 9(1), 33–39.
- Zelinsky, G. J. (2008). A Theory of Eye Movements During Target Acquisition. *Psychological Review*, 115(4), 787–835.

# Maximum utility unitary coherent perception vs. the Bayesian brain

Charles Fox (charles.fox@sheffield.ac.uk)

Department of Computer Science, University of Sheffield, UK

Tom Stafford (t.stafford@sheffield.ac.uk)

Department of Psychology, University of Sheffield, UK

## Abstract

Our subjective experience of the world is ‘unitary coherent’ (UC). ‘Unitary’ means we only perceive one interpretation at a time rather than a blur of multiple possible worlds. ‘Coherent’ means that we almost always perceive scenes that do not contain contradictory parts. While this form of first-person perceptual experience may seem obvious, it is in opposition to the requirements of optimal decision making, and to some forms of the ‘Bayesian brain’ hypothesis. We hypothesise that there are at least three types of ‘Bayesian’ action selection occurring in cognition, including a ‘maximum utility (MU) percept’ strategy that makes use of UC percepts. We give evidence from a video game experiment that is compatible with MU/UC perception and action selection, and is incompatible with optimal actions. Furthermore, it is compatible with the presence of utility bias in MU/UC perception: by changing the available actions we may be able to manipulate the subject’s percept of a fixed ambiguous stimulus.

**Keywords:** Bayesian; psychophysics; utility; bias; perception

## A paradox about perception

Our subjective experience of the world is ‘unitary coherent’. *Unitary* means we only perceive one interpretation at a time (e.g. either a face *or* a vase in the Rubin Vase illusion) rather than a blur of multiple possible interpretations (never the face *and* vase together). *Coherent* means that we almost always perceive scenes that do not contain contradictory parts. (e.g. we do not see part face and part vase). While the UC nature of perception may seem obvious from subjective experience, it is in opposition to the requirements of optimal decision making, which require consideration of *all* possible interpretations of sensory data (Bernardo & Smith, 2000). In particular, the ‘Bayesian brain’ hypothesis (Doya, Ishii, Pouget, & Rao, 2007) views perception as computing probabilities of many interpretations, and optimal actions would be found by integrating out the utility of all actions under all percepts. If both Bayesian brain research and optimal action theory (Körding & Wolpert, 2004) suggest that perception should operate using a distribution, or ‘Bayesian blur’ of possible percepts, why then is our subjective experience limited to a unitary coherent percept instead? And which unitary coherent percept do we perceive: the most probable one or the most useful one? This study argues that as full Bayesian perception and action selection is computationally hard, an approximation which we call ‘maximum utility (MU)’ perception is a useful surrogate. It then presents evidence in support of the maximum utility perception hypothesis using a video game style experiment.

## Types of perception and action selection

When a unitary coherent percept is required in machine perception, such as the output of a machine vision (Felzenszwalb & Huttenlocher, 2005) or speech recognition system (Young et al., 2006), the maximum a posteriori (MAP) state is often used by system engineers,

$$s_{MAP} = \arg_s \max P(s|d), \quad (1)$$

where  $s$  are world states and  $d$  is the available data. However real-world agents are often required to make actions as well as – or instead of – reporting percepts. In these cases, perceiving the MAP state does not necessarily lead to the best action if the following naive action-selection rule is used as a separate stage following MAP perception,

$$a_{naive} = \arg_a \max U(a, s_{MAP}), \quad (2)$$

where  $U$  is utility,  $a$  are actions. Instead, optimal actions are obtained (Bernardo & Smith, 2000) by maximising *expected* utility (MEU), which requires integrating over the ‘Bayesian blur’ of possible worlds,

$$a_{MEU} = \arg_a \max \int_s U(a, s) P(s|d) ds. \quad (3)$$

MEU action selection has no role for unitary coherent percepts. Instead it must consider *every* interpretation  $s$ . Computational approximations to this integral might ignore some improbable interpretations (Spiegelhalter, Thomas, Best, & Gilks, 2008), but still sum over a *set* of possible world states  $s$  rather than privileging any particular unitary state.

In many cases humans have been claimed to make optimal actions (Griffiths & Tenenbaum, 2006). This may occur for low-level, rapid stimulus-response type actions, and for high-level cognitive decisions such as business and financial decisions. Much recent work in ‘Bayesian Cognitive Science’ proceeds by *assuming* an MEU framework, then reasoning backwards from observed actions to report human priors on various stimuli (Stone, Kerrigan, & Porrill, 2009).

So why then do we bother to perceive UC percepts? Do they have some functional significance as well as being correlates of subjective experience? If they do have a function, this would suggest that Bayesian Cognitive Science’s assumption of optimal action is flawed, and could potentially invalidate some of its reported human priors.

Bayesian inference and hence MEU decision making is generally an NP-hard problem (Cooper, 1990) so is impractical for all but the most constrained percepts and actions.

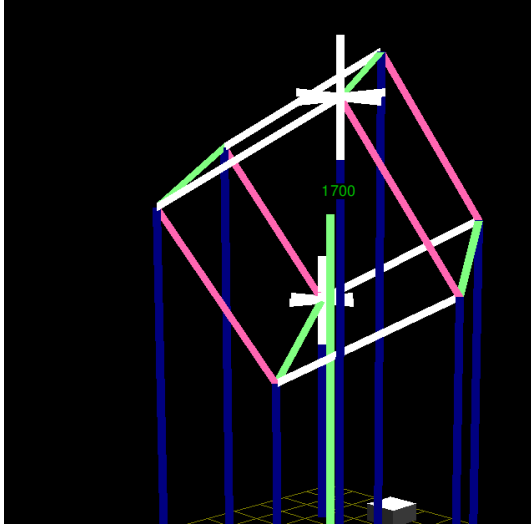


Figure 1: 3D environment used in the training phase of the game. A joystick moves the missile launcher around a 2D ( $x, z$ ) plane on the ground. Pressing and holding the joystick button fires a missile (not shown) vertically upwards ( $y$  axis). Releasing the button detonates the missile. Points are scored for detonation close to the target(s) shown by white crosses. The training phase shown here includes colour, overlap, perspective and support cues to make the cube's configuration unambiguous. These cues are removed in the test phase to leave an ambiguous Necker cube, with ambiguous 3D target positions. Figure is best viewed in colour.

It has been suggested (Gigerenzer & Todd, 1999; Goldstein & Gigerenzer, 2002) that making actions based on a single 'best' percept (such as the 'take the best' heuristic and 'less if more; effect) could be a useful heuristic to speed up the decision making process at the expense of optimality. However the 'percepts' in these cases are high level logical states of the word rather than actual perceptual objects in three dimensional space.

We propose an alternative form of perception and action selection to MAP perception and MEU action selection, which we call *maximum utility perception (MU)*. In MU we choose a UC state and action together,

$$(s_{MU}, a_{MU}) = \arg_{s,a} \max U(a, s) P(s|d) \quad (4)$$

which yields the best possible action assuming that only a single world state can be considered.

### The MU Hypothesis

We hypothesise that humans have at least *three* kinds of decision making behaviour, moving from fast and simple to slow and accurate:

1. *Immediate stimulus-response (S-R)*. A fast association from input data directly to an action. Such mappings do not need to build a UC percept. The could be implemented neurally at the sub-cortical level, such as direct links from supe-

rior colliculus to basal ganglia (Redgrave, Prescott, & Gurney, 1999). For simple mappings, it is possible that simple computational structures such as small neural networks could learn to perform near-optimal action selection, as has been demonstrated by computational experiments (Ramamoorthy & Verguts, 2012). Near-optimal performance in fast, low-level tasks such as reaching and pointing *quickly* at spheres having different locations and utilities (Kording & Wolpert, 2007); and 'simple heuristics' (Kahneman, Slovic, & Tversky, 1982; Gigerenzer & Todd, 1999) would be candidates for this mechanism. For simple tasks this method would give MEU-like results but without explicitly performing the MEU integration.

2. *Fast MU/UC percept-response (P-R)* which achieves suboptimal MU action in reasonable computation time. UC perception could be implemented cortically, with high-level perceptual areas computing a single most *useful* percept of the world, jointly with action selection under utility bias. Evidence for UC perception is found in binocular rivalry experiments (Srinivasan & Nunez, 2006), and in computational models (Riesenhuber & Poggio, 1999) as well as in everyday subjective experience. This paper gives evidence for MU/UC perception.

3. *Full MEU* action selection, via conscious sequential consideration of many possible percepts and responses. This slow type of decision making would occur for example when making a business decision, where several minutes (or even hours or days) are set aside to consciously perceive one possible world at a time, and the effects of many possible actions in them are simulated, and the resulting utilities averaged over. Humans are well-known to be poor at this kind of computation (Kahneman, 2003), and real-life action selection of this type is often performed in the business world by specialised operations researchers making use of computers to calculate the expected utilities (Pourret, Naïm, & Marcot, 2008), rather than relying on their own cognitive faculties.

If multiple decision making systems exist, it seems likely that the basal ganglia system is used to switch between them, for example taking account of time pressures for the type of decision to be made (Lengyel & Dayan, 2008; Redgrave et al., 1999; Daw, Niv, & Dayan, 2005). Strong support for the existence of at least two systems comes from the Ebbinghouse illusion, which produces different perceptual reports in verbal and stimulus-response type actions. It has been shown (Goodale & Milner, 1992) that the motor actions are consistent with optimal MEU-like decisions in the *same* subjects that make incorrect verbal reports.

While research on near-Bayesian optimal decisions of the S-R and Full MEU types abounds, there has been comparatively little work on the role of unitary-coherent perception in decision making. While our subjective experience tells us very clearly that *something* in the brain is computing a UC percept (which is incidentally presented to our conscious experience), and researchers have modelled how MAP percepts could be computed in this way (Riesenhuber & Poggio, 1999)

there has been little study of how this type of perception could be used in action selection as a replacement for S-R and MEU behaviour. Our hypothesis is that MU perception and action is in fact the dominant mode of everyday, aware, perception and action – the type of cognition that occurs consciously but not deliberately.

It is difficult to design experiments to isolate this middle, MU, level of perception, because as soon as subjects know their performance is being monitored they tend to start deliberating as in Full MEU, rather than performing ‘everyday’ perception and action selection. Conversely, if tasks are too low-level and fast-paced, they will use rapid S-R behaviour. Perhaps that is why few experiments have noticed MU effects before. To this end, we have carefully designed a simple 3D perception task, and examine two hypotheses:

*Hypothesis H1* is that there are examples of human behaviour that are consistent with UC perception and inconsistent with both Full MEU (deliberative) and approximate MEU (S-R). A positive result here would stimulate further research into delimiting the circumstances in which the different behaviour types are employed.

*Hypothesis H2* is that the particular kind of UC used in human perception is the MU percept. To find evidence for this stronger hypothesis, we will examine if it is possible to bias the percept from equally a priori probably percepts by altering the available action set, as predicted by MU.

## Methods

A video game – loosely based on “space invaders” – was designed and implemented<sup>1</sup>, having optimal MEU actions that require consideration of multiple scene interpretations, and having MU actions giving suboptimal rewards. If human behaviour in this (or any other) game could be shown to deviate from MEU behaviour and be consistent with MU, then evidence is provided for H1. Further, if the human behaviour is consistent with predictions made by MU selection, then evidence is provided for H2. An overview of the phases of the game is given here, followed by details of each phase.

In phase one of the game, shown in fig. 1, subjects were trained in several rounds to fire missiles from a launcher in a 2D plane, in an unambiguous simulated 3D environment. They received rewards according to how close to aerial targets (shown as white crosses in the figure) they get. After demonstrating that they understand the utility function and controls by passing a second, ‘examination’ phase, they are then tested (phase three) in an ambiguous bi-stable environment. A true MEU strategy would consider both interpretations of this environment, whereas a UC based strategy would use only one and lead to a different action. Phase one consists of several rounds which teach the subject about the game. The choice of tasks here is fairly arbitrary, as the logic of the experiment is that *if* subjects can pass the phase two exam *then* they have demonstrated an understanding of the rules sufficient to play the real test game in phase three. Phases one and two may

be repeated in any order as many times as the subject desires, until the exam is passed successfully.

### Phase 1: exploratory training round

The game environment consists of a visible 2D horizontal plane on which a missile base can move around, a wire-frame cube in the 3D space above the plane, and one or two targets located at vertices of the cube. The environment is drawn using very strong perspective<sup>2</sup>, and the vertices of the cube are connected vertically to the plane by lines to make their 3D locations unambiguous. In addition, edges of the cube are drawn with thick lines of different colours, producing additional disambiguation cues where one lines is seen to cross in front of another.

Subjects control the  $(x, z)$  position of a missile base using an analogue joystick (Logitech Extreme 3D Pro) and fire a missile by pressing and holding the joystick trigger. Once fired, the missile moves upwards (the  $y$  direction) until the trigger is released. The missile then explodes at position  $\mathbf{m} = (m_x, m_y, m_z)$ .  $N \in 1, 2$  targets are present in the environment at positions  $\mathbf{t}^i = (t_x^i, t_y^i, t_z^i)$ ,  $i = 1 : N$ , and a Gaussian reward  $R$  is received and displayed centre-screen after the explosion,

$$R = \sum_{i=1}^N r_i, \quad (5)$$

$$r_i = 100 \times \exp - \frac{(\mathbf{m} - \mathbf{t})(\mathbf{m} - \mathbf{t})^T}{2\sigma^2}. \quad (6)$$

The spread,  $\sigma$  was fixed at a large enough value ( $\sigma = 2\sqrt{3}/\sqrt{2\ln 2}$ ) so that if two targets are present, the score is always highest when firing at the point between them than when firing directly at one of the targets.

In the exploratory round, single targets are presented at different vertices of a fixed cube. Subjects have unlimited time to position the missile base and fire. They are then represented with a visual display of the reward, then the next target is presented. They are given 50 such targets to practice with, no cumulative score display, and are encouraged to experiment to learn about the rewards available at different distances from the target by an introductory message.

### Phase 1: utility training round

In the exploratory round, subjects obtained high scores by firing as close to the target as possible. To help them learn about the shape of the Gaussian utility function, a series of rounds takes place in which fixed cubes are shown and the subject is asked to deliberately score *only* 50, 70 or 90 points. Thus they are encouraged to try aiming at locations at different distances from the target.

### Phase 1: double target training

This round is similar to the first exploratory round, but uses two targets presented together at each trial. By construction

<sup>1</sup>in Python, source code available on request.

<sup>2</sup>OpenGL: gluPerspective(45, 1.0\*width/height, 0.1, 100.0); gluLookAt(7,0.05,7, 0,-1.25,0, 0,1,0). Cube faces are 2\*2 units.

(choice of  $\sigma$ ), the optimal action is now always to aim at the point midway between the targets.

### Phase 1: blackout training rounds

Two further training rounds take place. In the first, the lower half ( $x + z > 0$ ) of the missile-launching area is ‘blacked out’. It is coloured red, and the joystick is unable to move into the red area. In the second round, the situation is reversed at the top half of the grid ( $x + z < 0$ ) is blacked out. Subjects learn that that optimal strategy when faced with a target in the blacked out region is to fire from a position as close to it as possible that is on the centre line.

### Phase 2: examination round

The purpose of the examination round is to demonstrate that the subject has learned the optimal actions for single and double targets, as well as possessing sufficient motor skills to control the game using the joystick. 20 trails are presented in rapid succession (one every 5s) and a cumulative score is maintained. If subjects fail to score 2170 points or more, they are sent back to the exploratory phases then made to repeat the examination (or allowed to leave the experiment). The qualifying score was chosen such that it can only be obtained by using the optimal strategy of aiming closer to the centre of each pair of double targets than to either individual target. Thus by passing the examination phase, subjects demonstrate knowledge of this strategy.

### Phase 3: Ambiguous test round

In trials within this round, a bi-stable ambiguous (Necker) cube is presented to the subject very quickly, at random orientations. In some (80% of) trials there is one target at a random vertex. In others there are two targets, which may be at opposite (10% of trails) or non-opposite (10% of trials) vertices. The percept is made ambiguous by switching the projection from perspective to orthographic, dropping the vertex-to-plane cues, and drawing all edges in white to remove overlap cues. To motivate subjects, they are told that their cumulative score from all rounds of phase three will be their reported result, and that the subject with the highest reported result will receive twenty UK pounds in cash. (Subjects were undergraduate psychology students and were not paid otherwise; but received credit towards their degree for participating.)

### Phase 3: Blackout test rounds

The ambiguous test round is repeated twice more, with blacked out near and far regions as in the learning phase.

### Debriefing

It is crucially important that subjects do not become aware of the ambiguity, because this could allow high-level (Full MEU) reasoning to aim in the centre, and destroy any UC-revealing behavioural effects. For this reason, after phase 3, subjects were told about Necker cubes and asked if they were aware of the Necker ambiguity.

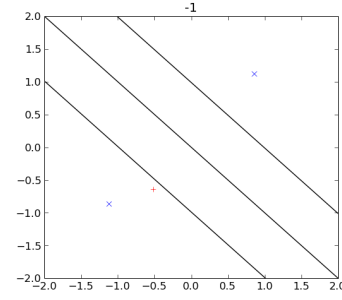


Figure 2: Example of a single-target trail,  $(x, z)$  plane. The viewer’s position is in top right corner. Blue crosses show the two ambiguous target positions resulting from a single target vertex on a Necker cube. The red cross shows the firing position. Black lines show the centre line and the two classification boundaries, dividing the launching area into near (top-right), centre and far (bottom-left) firing regions.

### Processing

25 subjects were tested. Of these, 20 completed the exam and proceeded to generate data in the test phases. In debriefing, no subjects reported awareness of the ambiguity in the Necker cubes. For each trial, the 3D positions of both ambiguous locations of the target or targets were computed. This was achieved by transforming the  $x, z$  joystick co-ordinates into a horizontal and depth pairs,  $(h, d)$ , then flipping the depth coordinate,

$$\begin{bmatrix} h \\ d \end{bmatrix} = H \begin{bmatrix} x \\ z \end{bmatrix}, \quad (7)$$

where  $H$  is the Hadamard matrix,

$$H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \quad (8)$$

The complementary ambiguous location is thus

$$\begin{bmatrix} x' \\ z' \end{bmatrix} = H^{-1} \begin{bmatrix} h \\ -d \end{bmatrix}. \quad (9)$$

Furthermore, the height coordinate was transformed by  $y' = y - cd$ , where  $c$  is a constant ( $c = 0.9$ ) which compensates for the choice of viewing angle in the projection images.

All shots were classified into three regions (fig 2), according to whether their  $(x, z)$  firing locations were closest to the near ambiguous target location, the far ambiguous target location, or the centre line.

## Results

### Non-blackout trials with single targets

In these trials, the MEU action by construction (i.e. choice of  $\sigma$  as described in phase 1) is to fire at the point on the centre line between the two possible ambiguous locations. (fig. 2). In contrast, the MU action is to fire directly at a

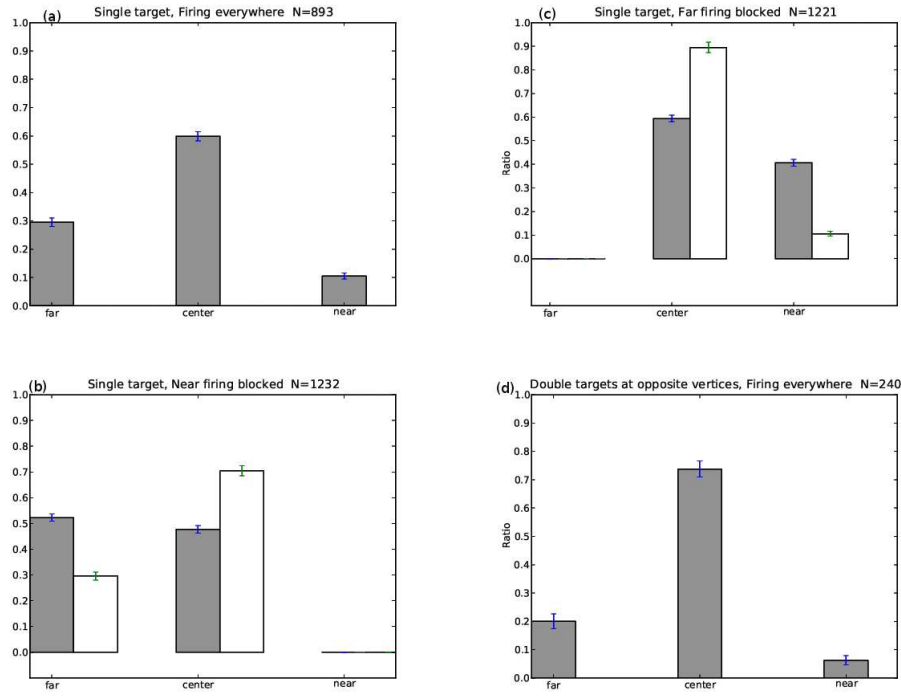


Figure 3: Results. Grey bars: observed frequency ratios, with Beta posterior, one-standard-deviation confidence intervals. White: predictions for blocked firings under the null hypothesis. The null hypothesis is that the unblocked area is unchanged from the 'single target, firing everywhere' case, while the centre is the sum of the centre and blocked firings from that case. All error bars assume IID observations and ignore which observation came from which subject.

randomly-chosen single one of those locations. This is because the MEU action averages over the two possible states of the world, which gives the same calculation as choosing where to fire in an unambiguous double target case; whereas the MU action picks just one interpretation of the target location, then fires directly at it.

Fig. 3a shows the distribution into region classes, over all subjects and trials of this type. Treating each action by each subject as an independent observation (i.e. ignoring subject-specific effects), and beginning with a flat Beta prior over the ratio of shots in each region, we infer posterior ratios, along with uncertainties. The figure shows the mean and one standard deviation error bars inferred about the population ratio of shots fired of each type. Signal detection theory can be used to obtain the  $p$  values, but broadly two ratios are significantly different if pairs of error bars do not overlap.

The results of these trials are surprising but inconclusive. Although target locations are perfectly ambiguous between near and far positions, subjects show a preference for the far target over the near one. That is, they are already interpreting the Necker cubes percept in a biased way, to favour interpretations with the target at the back of the scene.

If the MEU strategy was followed perfectly, we would see all shots fired in the centre and none in the near or far regions. If the MU strategy was followed perfectly, we would see all shots in the near and far regions and none in the centre. Unfortunately, we see shots in all three regions. Whilst this

is incompatible with a pure MEU strategy, it can weakly be explained from a MU perspective: A large number of shots are fired from the centre line, which may be due to subjects losing all depth perception (i.e. not perceiving the cube as 3D at all) and hedging by firing in the centre; they may also be due to limited depth perception resulting in a stable cube percept but an inaccurate joystick placement. (Some subjects commented on the lack of training in the absence of the vertical supports, and consequent loss of skill at pointing to the depth of the targets.) The near and far shots would be correct MU actions, and the centre shots due to a problem with the experiment, requiring a better communication of the depth to the subjects in future versions whilst retaining the ambiguity.

### Blackout trials with single targets

In these trials, the pure MEU action is *still* to fire at the point on the centre line between the two possible ambiguous locations. Points on the centre line are still available during a blackout, so the optimal strategy is unchanged.

Fig. 3b shows the results when the near-side is blacked out. The majority of shots are now fired in the far region. This is consistent with the MU strategy: actually *perceiving* and acting on the Necker cube in the configuration which enables the target to be reached; an optimism bias. If we assume UC perception and action, these new results then show MU-like bias occurring within in. For comparison, we show in white the prediction of a null hypothesis. This is obtained taking



the Single target histogram of fig. 3a and moving its near mass to its centre mass, as would occur if UC percepts were unchanged by the utility bias, and near shots were substituted by firing on the centre line as close to blacked-out near targets as possible. MEU theory is unable to explain why the observed frequencies are so different from this null hypothesis. MU theory explains it easily: there is no utility in *perceiving* near targets; but if they are re-perceived as far targets, then an increased utility can be obtained by firing at them.

Similarly, fig. 3c shows the results when the far-side is blacked out. Again compared to a null hypothesis (white bars) which moves the mass from the far region to the centre region from fig. 3a, we see a significant difference, again suggesting the MU-like change in both percept and action towards the obtainable (non-black-out) target position.

Fig. 3d is shown as a control. It is the distribution from the non-black-out trials having two targets at opposite cube vertices. In these cases, the MEU and MU strategies are the same – fire in the dead centre of the grid, and the results show a significantly increased rate of firing in the centre region over that of fig. 3a. This again supports the presence of MU over MEU, because MEU would give identical results in 3a and 3d but MU would give an increase in the centre region in 3d over 3a, which we do see here.

## Discussion

In informal discussions, we often argue that “perception is obviously UC from subjective experience”. Operationalists challenge this statement and would prefer us to cite an experiment to demonstrate the claim in the third person. While MEU actions are known to occur at both low-level psychophysical tasks and at high-level cognitive reasoning tasks, we have here presented evidence for the existence of a largely unexplored middle ground in which action selection is consistent with MU, and inconsistent with both Full deliberative and S-R approximate MEU, as in Hypothesis H1.

MU, but not MEU, can explain the deviations from the null hypotheses seen in figs. 3b and 3c, and also the difference between figs. 3a and 3d. The match of the data to MU behaviour in these cases gives some support for Hypothesis H2.

It was disappointing for the MU theory that fig. 3a did not present conclusive evidence by itself of MU over MEU, as pure MU would predict all shots to be fired in near and far regions, not in the centre. One way to explain the data away here is that the stimuli used were insufficiently informative to give subjects a sense of space, so they fire in the centre by default in the absence of any meaningful percept. Future work should try to refine the experiment to see if such hypothetical null percepts can be replaced by true percepts, for example by using different input and display systems while retaining the ambiguity in the Necker cube itself. Finally, the assumption that all shots by all subjects are mutually independent is strong, and future work should employ more subjects so that the assumption of independence of shots belonging to each subject can be dropped in the analysis.

## References

- Bernardo, J., & Smith, A. (2000). *Bayesian Theory*. Wiley.
- Cooper, G. F. (1990). The computational complexity of probabilistic inference. *Artif. Intell.*, 42, 393–405.
- Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci*, 8(12), 1704–1711.
- Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (2007). *Bayesian brain*. MIT.
- Felzenszwalb, P. F., & Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1), 55–79.
- Gigerenzer, G., & Todd, P. (1999). *Simple heuristics that make us smart*. OUP.
- Goldstein, D., & Gigerenzer, G. (2002). Models of ecological rationality. *Psych. review*, 109(1), 75.
- Goodale, M., & Milner, A. (1992). Separate visual pathways for perception and action. *TINS*, 15(1), 20–25.
- Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psych. Science*, 17, 767–773.
- Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 93(5), 1449–1475.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge.
- Körding, K., & Wolpert, D. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971), 244–247.
- Körding, K., & Wolpert, D. (2007). *Bayesian statistics and utility functions in sensorimotor control*, in Doya et al.
- Lengyel, M., & Dayan, P. (2008). Hippocampal contributions to control: the third way. In *NIPS*. MIT.
- Pourret, O., Naïm, P., & Marcot, B. (2008). *Bayesian networks: a practical guide to applications*. Wiley.
- Ramamoorthy, A., & Verguts, T. (2012). A computational model of instruction following. *Brain Research, pre-press*, doi:10.1016/j.brainres.2011.12.025.
- Redgrave, P., Prescott, T. J., & Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89(4), 1009–1023.
- Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11), 1019–1025.
- Spiegelhalter, D., Thomas, A., Best, N., & Gilks, W. (2008). *Bugs: Bayesian inference using Gibbs sampling*. Available from <http://www.mrc-bsu.cam.ac.uk/bugs/>
- Srinivasan, R., & Nunez, P. (2006). *Brain networks with distinct spatial and temporal structure*. Abstr. in Assoc. Scientific Study of Consciousness, Oxford.
- Stone, J., Kerrigan, I., & Porrill, J. (2009). Where is the light? Bayesian perceptual priors for lighting direction. *Proc. Royal Soc. B*, 276(1663), 1797–1804.
- Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., et al. (2006). *The HTK book*.



# Measuring children's visual access to social information using face detection

Michael C. Frank

mcf Frank@stanford.edu

Department of Psychology

Stanford University

## Abstract

Other people are the most important source of information in a child's life, and one important channel for social information is faces. Faces can convey affective, linguistic, and referential information through expressions, speech, and eye-gaze. But in order for children to apprehend this information, it must be accessible. How much of the time can children actually see the faces of the people around them? We use data from a head-mounted camera, in combination with face-detection methods from computer vision, to address this question in a scalable, automatic fashion. We develop a detection system using off-the-shelf methods and show that it produces robust results. Data from a single child's visual experience suggest the possibility of systematic changes in the visibility of faces across the first year, possibly due to postural shifts.

Keywords: Social development; face processing; head-camera.

## Introduction

Faces are perhaps the most important source of social information for young children. Infants show a preference for faces and face-like configurations from birth (Johnson, Dziurawiec, Ellis, & Morton, 1991; Farroni et al., 2005), and they will fixate faces to the exclusion of nearly everything else when attending to complex naturalistic stimuli (Frank, Vul, & Johnson, 2009; Frank, Vul, & Saxe, 2011). By their first birthday, they are sensitive to facial information about emotion (Cohn & Tronick, 1983) and social group (Kelly et al., 2005), and they will readily follow gaze to an attended target (Scaife & Bruner, 1975). As they begin to speak and understand language, joint attention becomes a powerful cue for learning the meanings of words (Baldwin, 1991).

To extract all of this important information in the natural environment, infants and children must attend to people's faces. Nearly all of what we know about children's attention to—and understanding of—faces comes from tightly-controlled lab experiments. In such experiments, the stimuli are typically presented in a very accessible format: at eye-level, large enough so that all details can be appreciated. How often do children actually see the faces of the people around them, though? And how often are the faces large enough to discern details from?

Head-mounted cameras provide a new technique for measuring access to faces during development. While the method of placing a miniature camera on the head of an infant or young child is still relatively new, a number of investigators have begun using it to record children's first-person perspective (Yoshida & Smith, 2008; Aslin, 2009; Smith, Yu, & Pereira, in press). Some studies have even used head-mounted eye-trackers to measure what part of the visual scene the child is fixating, a good proxy for what parts of the world

the child is attending to (Franchak, Kretch, Soska, Babcock, & Adolph, 2010).

Of particular interest is the result, reported by Franchak et al. (2010), that 14-month-olds rarely fixated their mother's face, even when she spoke to them directly. They looked instead at her hands or other parts of her body. The authors speculated that this result might have been due to the mother's location, usually high above the child. When mothers were sitting down, their faces were much more visible to their children. In our current investigation we follow up on this suggestion, investigating the possibility that the posture of caregivers and the infant's own posture work together to cause developmental changes in the accessibility of social information.

The introduction of these new methods mean that for the first time, we can see what babies are looking at as they interact with—and learn from—the people around them. This development opens up many new questions for investigation. Yet work of this type is hindered by the tremendously slow and resource-intensive task of manually annotating videos, frame by frame. Up until now, only a few research groups have grappled with the task of how to analyze the massive datasets captured using these methods.

The current study thus serves two purposes. First, it is designed to measure the accessibility of social information—in the form of faces—to infants. To investigate this question across development, we make use of a previously-described dataset (Aslin, 2009), in which a head-mounted camera recorded 2–3 hours of the visual experience of a single child at ages 3, 8, and 12 months (sample frames shown in Figure 1). Second, we investigate the possibility of using automated face detection to measure social information. It might in principle be possible to hand-annotate the presence of faces in each of the million-odd frames in our dataset (such annotation can be done around 4–8 times slower than real-time, yielding around 25–50 hours of total annotation time). For any larger study with more participants, annotation costs would quickly become prohibitive. Our study thus was designed to serve as proof-of-concept for the automated strategy.

Detection of upright faces in static images is widely considered to be a solved problem in computer vision, with the work of Viola and Jones (2004) providing a computationally-efficient solution that is now used in a wide variety of systems and consumer electronics. Nevertheless, the dataset we used presents a distinct set of challenges for such methods. In what follows, we describe our method for handling these challenges using a collection of out-of-the-box techniques from computer vision and machine learning. We end by describing

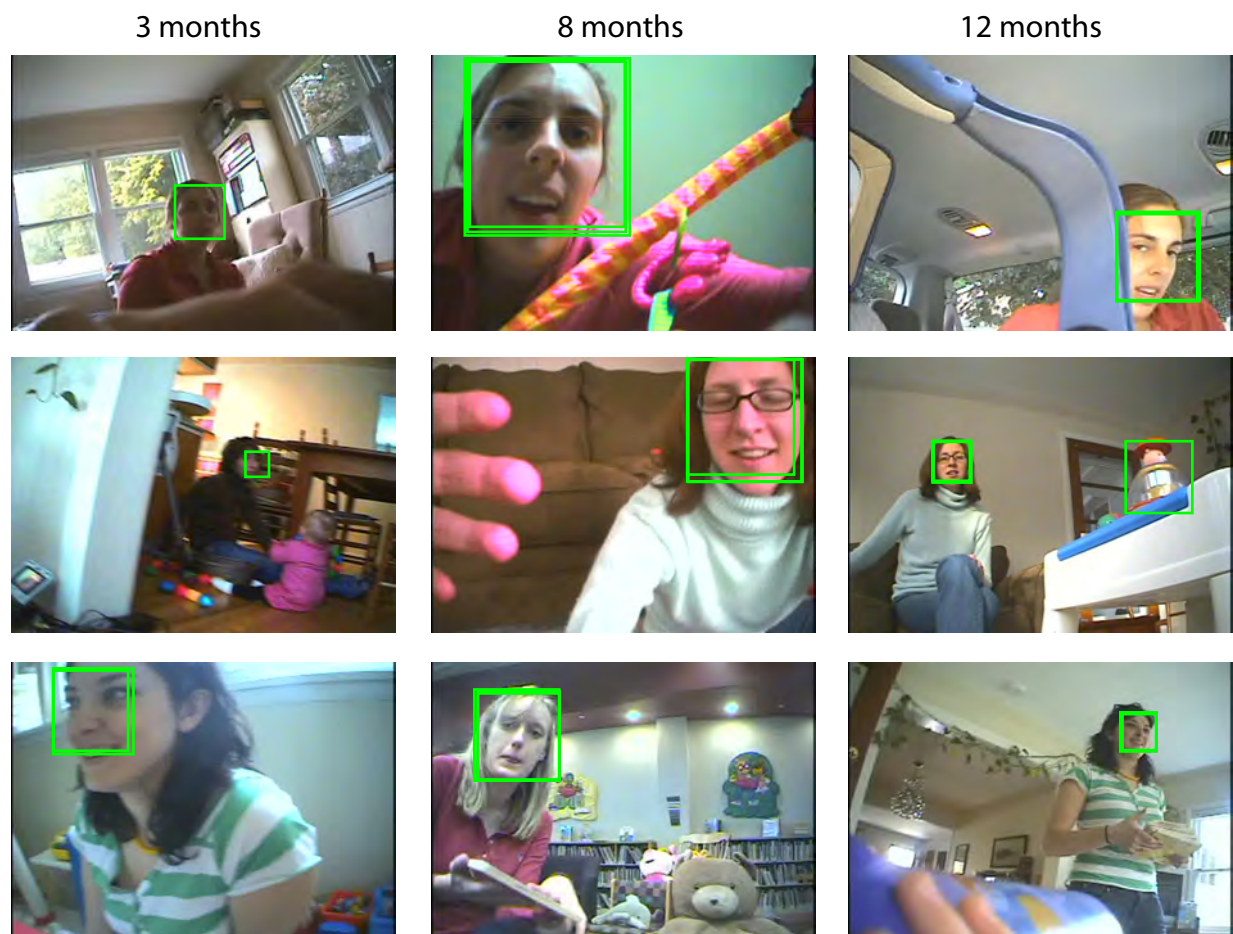


Figure 1: Sample frames showing head-camera data plotted along with face detector data. A separate rectangle is plotted for each active detector. Frames were selected in which annotations and model predictions matched.

developmental changes in the prevalence and size of faces in the field of view of the infant we studied. These changes suggest that there may be a number of important factors influencing the accessibility of social information during early development.

## Methods

Although in principle a single joint detection and tracking system could be constructed to detect faces in head-camera video, in practice such a system would be complex and computationally-intensive. Thus, we pursued a two-step approach to face-detection (Figure 2). We first preprocessed each frame of our data separately using simple but noisy detectors, which find faces in static images. We then tested a number of supervised post-processing models on their performance in picking frames with successful rather than spurious detections.

Because of this two-step scheme, conventional annotations of a gold standard training sample (e.g. face/no face) were not maximally effective. If the detectors did not find a face in a

frame, training the post-processing model that the frame contained a face would be counterproductive. Instead, our strategy was to create two annotated sets. The first was a training set that indicated whether, for each frame, the detectors had correctly identified a face. The second was a generalization dataset that indicated whether a face was in fact present in the frame, allowing us to test what proportion of faces our models identified on a completely independent dataset (different clips from the same corpus). In addition, we annotated the child's posture in each video of the corpus. These annotations (along with the details of the dataset) are described below.

## Data and annotation

**Aslin (2009) head-camera dataset** Data for the study consisted of videos collected on three days during the infancy of a single child, at ages 3, 8, and 12 months. This dataset was originally collected by Aslin (2009); the data are described in detail in Cicchino, Aslin, and Rakison (2010). The method of collection was a small wireless camera mounted

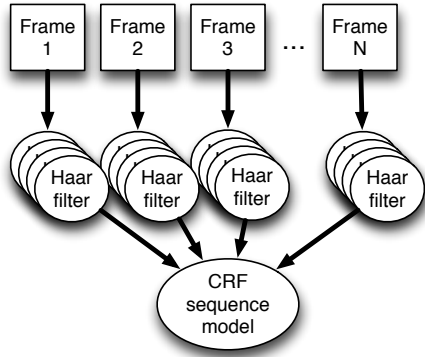


Figure 2: Schematic of our face-processing approach. Step 1: process frames with noisy Haar-style detectors. Step 2: filter detections with conditional random field (CRF) model.

on the infant’s head, allowing recording of a large portion of the infant’s visual field. The camera was a Sony 480TVL CCD “bullet” camera, embedded in a headband and wirelessly transmitted to a digital video recorder. Videos were approximately 126, 190, and 140 minutes long for the segments collected at 3, 8, and 12 months, respectively. Recordings were made while the infant was in a number of different locations, including in the home, on a shopping trip, on a walk, and at a playgroup. Due to the variation in activities across ages, the natural statistics of these three samples were unmatched (likely due to both sampling issues and true differences in the distribution of activities across ages); thus we will not attempt to compare across activity types.

**Annotation of detectors (training set)** We annotated a sample of videos to provide training data for our models. For this annotation effort our goal was to select frames in which the raw face detectors had correctly selected a face (and reject those for which the detections were incorrect). We classed a frame as containing a correct detection if there was at least one detector around the face of a person (thus a frame could still contain some spurious detections, though in practice this was relatively rare). We annotated nine clips of one minute each (16k frames). Three minute-long clips were selected for each age group randomly, with the caveat that they included some correct face detections in each.

**Annotation of face presence (generalization set)** We additionally performed frame-by-frame annotations of whether a face was present in the video frame. We selected 3–4 one-minute clips at each of the three ages for a total of 11 minutes of video at 30 frames per second (20k frames). One-minute clips were selected randomly, again with the caveat that they needed to contain at least some instances of faces. We counted a frame as containing a face when a face was fully visible with no occlusions at three-quarter view or greater (both eyes visible). This stringent annotation criterion was used because occluded or profile-view faces are much less

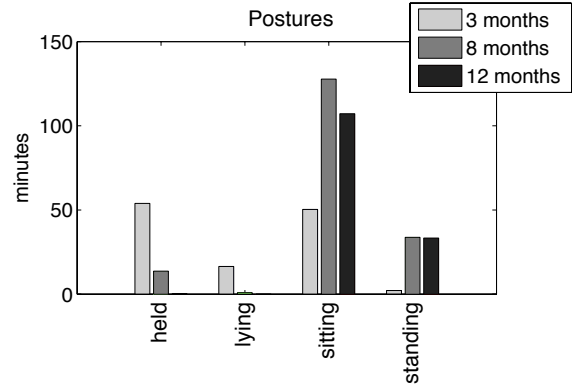


Figure 3: Time spent in each coded posture at each age.

likely to be useful for inferring eye-gaze direction, emotional state, or other social information.

**Posture annotation** We additionally annotated the posture of the child during the videos, in order to use this factor in our analysis of position and size of detected faces. We attempted to estimate the child’s posture wherever possible, categorizing it as lying, sitting, standing, crawling, or being held. Figure 3 shows descriptive data for this measure. Annotation of this measure was somewhat subjective, but inter-rater agreement was relatively high with  $\kappa = .72$  for five categories.

## Models

Although face detection is generally considered to be a solved problem (Viola & Jones, 2004), face detection in developmental, first-person data presents a number of challenges that do not usually occur in static photographs or standard videos. First, faces are often occluded and at odd orientations for children. Second, in our case, the video was transmitted wirelessly and contained some artifacts due to the transmission method. Third, the head-mounted camera was subject to quick movements as the child moved his head, meaning that many methods applicable for scene segmentation or motion tracking in static-camera applications could not be used here. Our modeling goal in this project was to combine computationally-inexpensive techniques to address these challenges.

Our preprocessing step made use of off-the-shelf Haar-style detectors from the OpenCV package (Bradski & Kaehler, 2008). Each frame was processed with four separate detectors: three full-face and one profile detector. These detectors were noisy, capturing many faces but also spuriously identifying many background elements as faces as well (e.g. doorknobs, high contrast windowpanes, see Figure 4). This processing step ran at approx. 10% of real time on a quad-core machine, taking around 4 days to process all detectors.

Next, we trained post-processing models to discriminate valid detections from invalid detections, using our detector-annotated training set. Our primary model of interest was a

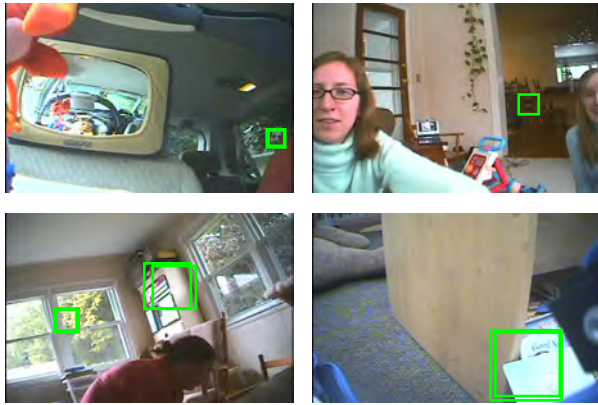


Figure 4: Frames in which CRF model incorrectly predicted that the detectors had correctly identified a face.

conditional random field (CRF) model (Lafferty, McCallum, & Pereira, 2001). CRFs are discriminative sequence models: they take input data of sequences of observations (with some feature set describing each observation) and return a classification of each observation in the sequence. Their key difference from feature-based classifiers (e.g. Naive Bayes or MaxEnt) is their ability to use sequential information; likewise, their key difference from sequence models (e.g. hidden Markov models) is their ability to incorporate rich featural information about each observation. They have been applied successfully to a number of tasks including natural language processing and computer vision. For this application, we used the Matlab CRF toolbox (Schmidt & Swersky, 2008).

We included two other simpler models for comparison: a Naive Bayes (NB) classifier and a hidden Markov model (HMM). The classifier made use of exactly the same feature set but considered each frame in isolation (neglecting sequential dependencies). The HMM considered only the sequence of decisions and the number of detectors that were active. Thus, the difference in performance between the CRF and the classifier provides a rough measure of the contribution of sequential information (provided by the video), while the difference between CRF and HMM provides an estimate of the gain due to adding featural information.

We created a set of binary features to describe the detections in each frame. These included a separate feature for whether each detector was active, a feature for each detector pair to indicate whether the detector centers fell within a certain threshold (5 pixels) of one another, and features for each detector indicating whether it changed in size or disappeared in either the preceding or following frame. We used this feature set to train the models to classify the training data as containing correct or incorrect detections.

Table 1: Model performance on detector-annotated training dataset (“Tr,” 9 minutes, only frames with successful detections) and generalization dataset (“Gen,” 11 minutes, all frames with human-visible faces). P = precision, R = recall, F = F-score (harmonic mean of precision and recall).

	Tr P	Tr R	Tr F	Gen P	Gen R	Gen F
NB	.64	.82	.72	.76	.55	.64
HMM	.72	<b>.85</b>	.78	.81	<b>.57</b>	<b>.67</b>
CRF	<b>.85</b>	.77	<b>.81</b>	<b>.85</b>	.53	.65

Table 2: CRF model performance on generalization training set by age. Prop. faces refers to the proportion of total faces in the gold-standard dataset for that age.

	3 months	8 months	12 months
Precision	.92	.89	.33
Recall	.60	.48	.35
F-score	.73	.62	.34
Prop. faces	.46	.45	.07

## Results

Table 1 shows evaluation results for each of the models on the two datasets we annotated, the detector-annotated training dataset and the gold-standard generalization dataset. (Rather than using a technique like cross-validation to test generalization performance, we report results on both the training data and an independent generalization set that was never used for training). The CRF model performed best on the training data, capturing a slightly better tradeoff between precision and recall. The gain in performance was relatively slight from the HMM to the CRF, indicating that the majority of the value of the CRF was due to the sequential dependencies enforced by the model. Knowing that a previous frame contained a successful detection was helpful in deciding whether the current one did as well.

When we applied the three models to the generalization dataset, F-scores were within a small range of one another, with the HMM outperforming the CRF, perhaps indicating some overfitting of feature weights to the training data. Nevertheless, the CRF produced the highest precision on the generalization dataset. Because our aim was to measure the quantity and spatial distribution of faces at each age, we judged precision more valuable than recall and chose the CRF model for our analysis (though we note that results do not change meaningfully if the other models are chosen).

Performance on the generalization set was highly asymmetric across the three ages, with high precision and recall for the 3-month data, mid-level performance on the 8-month data, and very low performance on the 12-month data (Table 2). A number of experiments attempting age-specific training failed to find major gains in performance by training only on e.g. 12-month data. There were few faces in the 12-month



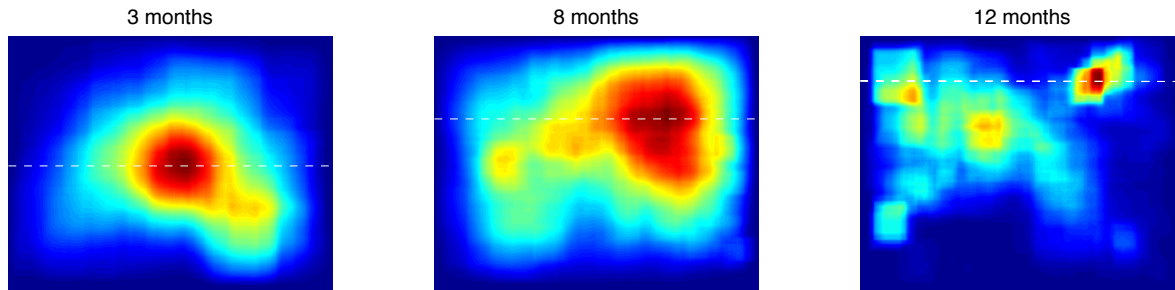


Figure 5: Heat maps showing probability of finding a face in each location of the camera field for 3, 8, and 12-month-old data. Dotted lines show the vertical locations of maxima.

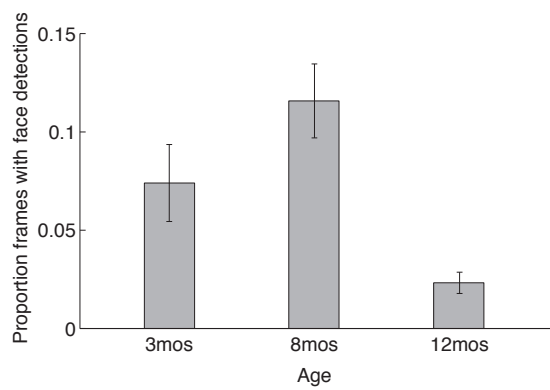


Figure 6: Proportion of faces detected by CRF model at each age. Error bars show standard error across video clips.

data (7% of frames, compared with 46% of frames in the 3-month data), and those that were present were very hard to detect correctly, perhaps because of their small size. Figure 4 shows frames in which the CRF model incorrectly reported a face; these typically showed consistent spurious detections for some superficially face-like configuration of objects.

We evaluated the CRF model on the entire dataset, using the settings established in training. Congruent with the generalization data, we found very few faces in the 12-month data relative to the other two ages. Figure 6 shows the estimated proportion of face-containing frames across clips at each age. Nevertheless, we should be cautious in interpreting these results, due to the relatively small amount of data available in this dataset. It may be the case that these results are skewed due to, e.g., participating in a play-group at 8 months with many children present.

Figure 5 shows a heat map of the probability of finding a face at each location in the camera's field for each of the three samples.<sup>1</sup> Faces were higher in the image plane at 8 and 12

months than at 3 months. This shift could potentially be due to postural differences: the child was more likely to be held or lying down at 3 months and more likely to be sitting or standing at 8 and 12 months. In a sitting or standing position, faces tend to be higher in the visual field than when lying down and looking up over the edge of the crib.

Faces were also different sizes in the older videos. The 3-month videos had a qualitatively different distribution of detected face sizes (Figure 7). We cannot completely rule out the possibility that some of the smaller faces in the 8- and 12-month videos were spurious detections. Nevertheless, the relatively similar distribution for each of these (compared with the drop in precision from 8 to 12 months) suggests that decreasing precision of detections was not the only factor here. Though speculative, a postural explanation for the shift in size might also be proposed: at older ages, the child was less likely to be lying or being held close to the face of a caregiver. Instead, in a seated or standing position, the faces of others would be further away.

Our final analysis directly measured size and vertical position of faces by posture (due to the limited overlap in postures between ages, regression analysis was not possible). The lying posture, seen only at 3 months, had a much larger face size than the other postures (almost 9% of camera field, as opposed to 2.5% for holding, 3% for sitting, and 4% for standing). Both lying and being held also had lower average vertical positions (.50 and .52 respectively, where 1 was the top of the screen) than sitting and standing (.60 and .57, respectively).

## General Discussion

We investigated the possibility of using automated face detection techniques to measure the accessibility of social information to infants. With head-mounted videos from a single infant at 3, 8, and 12 months, we constructed a discriminative model of face-detection that made use of inexpensive but noisy detectors and a secondary filtering step using a conditional random field model. This approach was relatively suc-

<sup>1</sup>For this and the remaining analyses, we averaged across all detections for each frame, potentially including some noise due to spurious detectors in correct frames. Future work should look estimate

the location of correct detections as well as frames in which they occur.

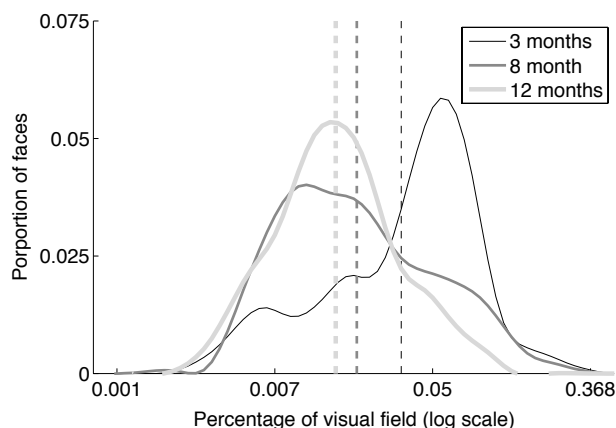


Figure 7: Smoothed histogram of detected face sizes at 3, 8, and 12 months. Height shows proportion of detections at each size; horizontal axis is scaled in log proportion of camera field. Dashed lines give means.

successful in picking out correct detections for the dataset as a whole.

The face-detector data revealed a surprising pattern. Faces were far less frequent in the 12-month data (and harder to detect, providing a potential caveat to our descriptive results). In addition, those faces that were detected in the older part of the dataset were both smaller and higher in the visual field of the infant. These differences seemed related to the distribution of postures across different ages, and indeed size and horizontal position did vary with posture. Nevertheless, further research (and considerably more data) will be necessary to check these conclusions.

The speculative picture that emerges is nevertheless congruent with previous work (Franchak et al., 2010). As children grow and become more adept at locomotion, they create a situation where the faces of others in their environment are further away from them and less visible. While the young infant is constantly having the faces of others pressed into his, the toddler lives in a world populated by knees.

More broadly, the methodological upshot of this work is that head-camera footage may be an extremely valuable tool for studying social attention and access to social information “in the wild.” Nevertheless, this work cannot proceed if hand-annotation is the only solution. Computer vision methods that are appropriate for data of this type must be developed, and this study took a first step in that direction, revealing suggestive developmental differences.

## Acknowledgments

Special thanks to Richard Aslin for generously sharing the primary dataset used in this project. Additional thanks to Ally Kraus for help with data organization and annotation, as well as to Adam Vogel and Evan Rosen for work on an earlier

version of this project. This work was supported by a John Merck Scholars Fellowship.

## References

- Aslin, R. (2009). How infants view natural scenes gathered from a head-mounted camera. *Optometry & Vision Science*, 86, 561.
- Baldwin, D. (1991). Infants’ contribution to the achievement of joint reference. *Child Development*, 875–890.
- Bradski, G., & Kaehler, A. (2008). *Learning opencv: Computer vision with the opencv library*. O’Reilly Media.
- Cicchino, J., Aslin, R., & Rakison, D. (2010). Correspondences between what infants see and know about causal and self-propelled motion. *Cognition*.
- Cohn, J. F., & Tronick, E. Z. (1983). Three-month-old infants’ reaction to simulated maternal depression. *Child Development*, 54, 185–193.
- Farroni, T., Johnson, M., Menon, E., Zulian, L., Faraguna, D., & Csibra, G. (2005). Newborns’ preference for face-relevant stimuli: Effects of contrast polarity. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 17245.
- Franchak, J., Kretch, K., Soska, K., Babcock, J., & Adolph, K. (2010). Head-mounted eye-tracking of infants natural interactions: A new method.
- Frank, M., Vul, E., & Johnson, S. (2009). Development of infants’ attention to faces during the first year. *Cognition*, 110, 160–170.
- Frank, M., Vul, E., & Saxe, R. (2011). Measuring the development of social attention using free-viewing. *Infancy*, 1–21.
- Johnson, M., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns’ preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40, 1–19.
- Kelly, D., Quinn, P., Slater, A., Lee, K., Gibson, A., Smith, M., et al. (2005). Three-month-olds, but not newborns, prefer own-race faces. *Developmental Science*, 8, F31.
- Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. int’l conf. on machine learning (icml)* (pp. 282–289).
- Scaife, M., & Bruner, J. (1975). The capacity for joint visual attention in the infant. *Nature*.
- Schmidt, M., & Swersky, K. (2008). *Conditional Random Field Toolbox for Matlab*. (<http://www.di.ens.fr/~mschmidt/Software/crfChain.html>)
- Smith, L., Yu, C., & Pereira, A. (in press). Not your mother’s view: The dynamics of toddler visual experience. *Developmental Science*.
- Viola, P., & Jones, D. H. (2004). Robust real-time face detection. *International Journal of Computer Vision*.
- Yoshida, H., & Smith, L. (2008). What’s in view for toddlers? using a head camera to study visual experience. *Infancy*, 13, 229–248.

# The Effects of Feedback During Exploration Depend on Prior Knowledge

Emily R. Fyfe (Emily.R.Fyfe@Vanderbilt.Edu)

Bethany Rittle-Johnson (Bethany.Rittle-Johnson@Vanderbilt.Edu)

Department of Psychology and Human Development, Vanderbilt University, Peabody College #552, 230 Appleton Place  
Nashville, TN 37203-5701 USA

## Abstract

Providing exploratory activities prior to instruction has been shown to facilitate learning. However, questions remain regarding the provision of guidance during the exploration phase. In this study, we replicated and extended a previous experiment by examining the effects of feedback during exploratory problem solving for children with varying levels of prior knowledge. Ninety-five children (M age = 8 yrs) solved 12 novel math problems and then received brief conceptual instruction. After solving each problem, they received (a) no-feedback, (b) outcome-feedback, or (c) strategy-feedback. Consistent with the previous experiment, the results resembled an aptitude by treatment interaction. Feedback during exploration prior to instruction improved children's procedural knowledge, but only for those with low prior knowledge. For children with higher prior knowledge, no feedback resulted in better procedural knowledge. Results suggest that providing feedback may not always be optimal.

**Keywords:** Guided Discovery Learning; Feedback; Aptitude by Treatment Interaction; Math Equivalence.

## Guided Discovery Learning

An emerging consensus suggests that people learn best through some form of *guided discovery*, which is defined as exploratory learning with supplemental instructional guidance. Learning tasks are exploratory if learners have not received instruction on how to complete them and instructional guidance encompasses a variety of tools, from in-depth instruction manuals to minimal feedback or hints. For example, Mayer's review (2004) suggests a "mixture of guidance and exploration is needed" (p. 17). Additionally, Alfieri et al.'s (2011) recent meta-analysis revealed the superiority of guided discovery over both pure discovery learning and pure direct instruction.

Providing exploratory activities prior to instruction is one form of guided discovery that has been recommended by researchers in education and cognitive psychology alike (e.g., Hiebert & Grouws, 2007; Schwartz & Bransford, 1998), and it is the form we focus on in this study. For example, several mathematics education researchers suggest, "each person must struggle with a situation or problem first in order to make sense of the information he or she hears later" (Stigler & Hiebert, 1998, p. 3). Similarly, Schwartz and Bransford (1998) suggest that exploratory activities facilitate the development of differentiated knowledge of the target problem space, which prepares learners for subsequent instruction.

There is a growing body of evidence to support the claim that exploration prior to instruction is beneficial (e.g., DeCaro & Rittle-Johnson, 2011; Schwartz & Bransford,

1998; Schwartz & Martin, 2004; Schwartz, Chase, Oppezzo, & Chin, 2011). For example, college students who explored novel examples learned more from a subsequent lecture than students who merely summarized a relevant text prior to the lecture (Schwartz & Bransford, 1998). Further, the timing of exploration and instruction matters. For example, children in elementary school benefited more from solving unfamiliar math problems before receiving instruction rather than vice versa (DeCaro & Rittle-Johnson, 2011).

However, questions remain regarding how and for whom this form of guided discovery is effective. First, should any guidance be provided during the exploratory activity? Mayer's review (2004) indicates that it is the guidance provided *during exploratory problem solving* that is crucial. Second, for whom is this guidance during exploration most advantageous? As noted by Cronbach and Snow (1977), often "the instructional approach that is best on the average is not best for all persons" (p. 1).

## Feedback and Prior Knowledge

Feedback is touted as one form of guidance that may be particularly beneficial during exploration. *Feedback* is any information about performance or understanding that the learner can use to confirm, reject, or modify prior knowledge. For example, Alfieri et al. (2011) specifically recommend "providing timely feedback" as an optimal approach to learning (p. 13). Similarly, Mayer (2004) cites feedback as an effective tool (among others) for keeping learners on track. Further, past research indicates that feedback's primary function is to identify errors and encourage the adoption of correct alternatives (e.g., Kulhavy, 1977), which may be particularly helpful when exploring a novel problem space. Given these positive effects, it seems likely that providing feedback during exploration would be universally beneficial.

However, a growing body of research indicates that the effects of feedback depend on learners' prior domain knowledge (e.g., Fyfe, Rittle-Johnson, & DeCaro, 2011; Kraise, Stark, & Mandl, 2009; Luwel et al., 2011). For example, college students with low prior knowledge learned more about statistics if they received feedback during problem solving than if they did not. However, students with higher prior knowledge did not benefit from such feedback (Kraise et al., 2009). Similarly, Luwel et al. (2011) examined children's performance on a numerosity judgment task that could be solved using one of two correct strategies. Children who knew neither strategy at pretest benefited greatly from feedback in terms of correct strategy selection,

but feedback had a much weaker effect for children who already knew the strategies at pretest. Together, these studies suggest that learners with low prior knowledge should benefit from feedback during exploration, but learners with higher prior knowledge may not.

This idea is consistent with past work on aptitude by treatment interactions (Cronbach & Snow, 1977), which occur when instructional treatments have positive effects for one kind of person, but neutral or even negative effects for another. Importantly, these interactions often occur in the context of differing levels of external guidance. For example, Snow and Swanson (1992) suggest tutors “should provide more scaffolding for less able learners and less scaffolding for more able learners” (p. 610). A large number of aptitude by treatment interactions involve interactions between instructional guidance and learners’ prior knowledge in the target domain (Kalyuga, 2007). For example, learners with low prior knowledge learn more from studying structured worked examples than from solving problems on their own. However, as knowledge increases, independent problem solving becomes the superior learning activity (e.g., Kalyuga & Sweller, 2004). In general, this work supports the notion that providing guidance (i.e., feedback) during exploration prior to instruction may help learners with low prior knowledge, but learners with higher prior knowledge may not need it.

### Previous Experiment

Thus, we compared the effects of feedback (i.e., guidance during exploration) to no feedback (i.e., no guidance during exploration) prior to instruction. We hypothesized that feedback during exploration would result in higher learning than no feedback. However, we expected the effect to be stronger for children with low prior knowledge.

We also explored whether the type of feedback mattered. *Outcome feedback* provides a judgment about the accuracy of the learner’s response, whereas *strategy feedback* provides a judgment about how the learner obtained that response. Outcome feedback has been studied extensively and is generally related to positive outcomes (e.g., Kluger & DeNisi, 1996). In contrast, few empirical studies have examined the effects of strategy feedback (e.g., Luwel et al., 2011). The limited evidence suggests strategy feedback can benefit strategy selection, but more research is needed to examine its effects across tasks and outcome measures.

We examined the effects of feedback in the context of children exploring math equivalence problems (problems with operations on both sides of the equal sign, such as  $3 + 4 + 5 = 3 + \underline{\quad}$ ). These problems are not typically included in elementary mathematics curricula (Rittle-Johnson et al., 2011), and research shows that U.S. children exhibit poor performance on math equivalence problems (e.g., Alibali, 1999; McNeil, 2008). Thus, these problems are novel and difficult for elementary school children, providing an apt domain to investigate exploratory problem solving.

In an initial experiment, children received a tutoring session that included exploratory problem solving followed

by brief conceptual instruction (Fyfe et al., 2011). The session was identical for all children with the exception that the feedback provided after each problem differed by condition. In the strategy-feedback condition, children received feedback on how they solved each problem. In the outcome-feedback condition, children received feedback on their answer to each problem. In the no-feedback condition, children did not receive feedback and were simply told to go on to the next problem. After the tutoring session, children completed a posttest (immediately and after a 2-week delay) that assessed conceptual and procedural knowledge of math equivalence. *Conceptual knowledge* is an understanding of the principles governing a domain and *procedural knowledge* is the ability to execute action sequences to correctly solve problems (e.g., Rittle-Johnson et al. 2011).

In line with our hypothesis, the effects of feedback on procedural knowledge depended upon prior knowledge. For low-knowledge children, feedback during exploration improved their procedural knowledge relative to no feedback. In contrast, for children with higher prior knowledge, no feedback resulted in superior performance than feedback, though this effect was slightly stronger for strategy-feedback than outcome-feedback. There were few effects on children’s conceptual knowledge. Thus, the results resembled an aptitude by treatment interaction. Children with low knowledge benefitted from receiving feedback, but children with higher knowledge benefitted more from exploring independently without feedback.

Although we predicted that prior knowledge would moderate the impact of feedback, we did not have a prior reason to expect a reversal such that feedback would actually harm learning for children with higher prior knowledge. Also, several limitations in the design constrained the strength of the conclusions. First, the manipulation was not as clean or as strong as it could have been. For example, all children were asked to report how they solved each problem, which inevitably guided all children’s attention to their strategies. The strategy-feedback manipulation would be stronger if only children in the strategy-feedback condition were encouraged to attend to their strategy use. Also, the feedback provided in both feedback conditions was relatively vague and not specific to the child’s response. For example, in the strategy-feedback condition, incorrect strategies were referred to as “not a correct way,” which may have been unclear. Further, children in both the strategy-feedback and outcome-feedback conditions were told if their target response (strategy or answer, respectively) was correct, but only children in the outcome-feedback were given additional information (i.e., the correct answer). The contrast between the two feedback conditions could be improved.

Second, we sought to clarify the influences of feedback type during exploration prior to instruction. Given the paucity of research comparing outcome-feedback to strategy-feedback, we wanted to confirm that feedback type is not central to children’s learning during exploration. To address these concerns, we conducted a second experiment



similar to Experiment 1, but with several modifications intended to strengthen the design.

## The Current Experiment

The current experiment was designed to strengthen the condition manipulation in Fyfe et al. (2011) and verify the results with an independent sample. Specifically, we attempted to replicate the finding that low-knowledge children benefit from feedback during exploration prior to instruction, whereas children with higher prior knowledge benefit from no feedback. Additionally, we sought to clarify the influences of outcome-feedback and strategy-feedback to confirm that feedback type did not impact children's learning during exploratory problem solving.

We strengthened the condition manipulation in three ways. First, to differentiate the conditions, we only had children in the strategy-feedback condition report how they solved each problem. Children in the other conditions were asked to report other information to mimic the interaction with the experimenter (i.e., their answer in the outcome-feedback condition and their completion of the problem in the no-feedback condition). Second, we made the feedback more specific by re-voicing the child's response. In the strategy-feedback condition we restated the child's strategy and in the outcome-feedback condition we restated the child's answer. Finally, we did not provide the correct answer in the outcome-feedback condition. In Fyfe et al. (2011), only children in the outcome-feedback condition received additional information (i.e., the correct answer). An alternative solution was to provide children in the strategy-feedback condition with additional information (i.e., a correct strategy). But, telling people how to solve a problem is a form of direct instruction, and we were interested in the guidance provided *prior* to direct instruction. So we eliminated the correct answer in the outcome-feedback condition to enhance parallelism across conditions.

Consistent with Fyfe et al. (2011), we predicted that children who received feedback during exploratory problem solving prior to instruction would exhibit better procedural knowledge of math equivalence than children who did not. However, we expected this effect to be larger for children with lower prior knowledge and to reverse for children with higher prior knowledge. Further, we did not expect any differences in children's conceptual knowledge.

## Method

Elementary school children received a tutoring session that included exploratory problem solving followed by brief conceptual instruction about math equivalence. The presence and type of feedback was manipulated during the exploratory problem solving.

### Participants

Participants were 111 second- and third-grade children. Ten were excluded from participation because they scored above 80% on pretest measures designed to assess children's prior

knowledge of math equivalence. Six additional children were excluded from analysis for not completing all activities. The final sample contained 95 children ( $M$  age = 7 yrs, 11 mo; 60 girls, 35 boys; 97% Black, 3% White).

### Design and Procedure

We used a pretest – intervention – posttest design with a two-week retention test. For the intervention, children were randomly assigned to one of three conditions: strategy-feedback ( $n = 31$ ), outcome-feedback ( $n = 33$ ), or no-feedback ( $n = 31$ ). Children completed the pretest in their classrooms in a 20-minute session. Within 1 week they completed a one-on-one tutoring intervention and posttest in a single session lasting approximately 45 minutes. Approximately two weeks after the intervention session, children completed the retention test in their classrooms.

The intervention began with exploratory problem solving. Children solved 12 novel math equivalence problems (e.g.,  $9 + 7 + 6 = \_\_ + 6$ ). In Fyfe et al. (2011), the problem were presented one at a time on a computer screen. In this study, we presented the problems in paper/pencil format to simulate a more typical classroom activity.

In the *strategy-feedback condition*, children reported how they solved each problem and received feedback on the strategy, which included a re-voicing of their report (“Good job! That is one correct way to solve that problem. [Child's strategy] was a correct way to solve it. / “Good try, but that is not a correct way to solve the problem. [Child's strategy] is not a correct way to solve it.”). The experimenter re-voiced the strategy just as the child reported it to ensure no added information was provided. In the *outcome-feedback condition*, children reported their numerical answer and received feedback on it, which included a re-voicing of their report, but not the correct answer (“Good job! You got the right answer, [child's answer] is the correct answer.” / “Good try, but you did not get the right answer, [child's answer] is not the correct answer.”). In the *no-feedback condition*, children reported when they completed each problem and were then told to move on.

After exploratory problem solving all children received brief conceptual instruction on the relational function of the equal sign. The experimenter provided a definition of the equal sign and explained how the left and right side of a problem were equal, using number sentences as examples (e.g.,  $3 + 4 = 3 + 4$ ). Between the exploratory problem solving and instruction, children completed a brief form of the assessment (midtest) to gauge the immediate effects of exploration prior to instruction.

### Math Equivalence Assessment

The math equivalence assessment, adapted from past work (Rittle-Johnson et al., 2011) was administered at pretest, posttest, and retention test. It included both conceptual (10 items) and procedural (8 items) knowledge subscales. Conceptual items assessed knowledge of the meaning of the equal sign and the structure of equations. Procedural items consisted of math equivalence problems,

and scores were based on children's use of a correct strategy to solve the problem. Example items and scoring are presented in Table 1. A brief version of the assessment (5 more difficult items) was used as the midtest.

Table 1: Example items on the assessment.

Task	Scoring
<i>Procedural Knowledge</i>	
Solve $8 = 6 + \square$ (operation on right side)	Use correct strategy (if unclear, response must be $\pm 1$ of correct answer)
Solve $3 + 4 = \square + 5$ (operations on both sides)	Same as above
Solve $\square + 6 = 8 + 6 + 5$ (blank on left)	Same as above
<i>Conceptual Knowledge</i>	
Define equal sign	Provide relational definition (same amount)
Judge equations such as $3 = 3$ as true or false	Correctly judge equations
Select choice that shows 10¢ is same as 1 dime	Select equal sign

## Analysis and Results

We used a planned contrast analysis of variance model. Because our condition variable had three groups (no-feedback, outcome-feedback, strategy-feedback), we created two coded variables. The first variable (feedback) compared no-feedback to the two feedback conditions combined. This allowed us to address our primary hypothesis regarding the presence or absence of guidance during exploration. The second variable (feedback type) compared outcome-feedback to strategy-feedback and allowed us to explore differences in the type of guidance provided. To evaluate whether condition effects depended on prior knowledge, we included two interaction terms: feedback by prior knowledge and feedback type by prior knowledge. We used procedural knowledge pretest scores as the prior knowledge measure as it is the most relevant domain knowledge for learning during exploratory problem solving. Finally, we included three covariates (children's age as well as procedural and conceptual knowledge pretest scores).

To evaluate children's performance on the assessment we conducted repeated measures ANCOVAs with feedback (feedback vs. none) and feedback type (outcome vs. strategy) as between-subject variables and time (midtest, posttest, retention test) as the within-subject variable. The two interactions and three covariates were also included. We examined procedural and conceptual knowledge separately.

### Pretest

On the pretest, children answered few procedural ( $M = 20\%$ ,  $SD = 18\%$ ) and conceptual ( $M = 19\%$ ,  $SD = 18\%$ ) items correctly. Importantly, there were no differences between conditions on either scale at pretest,  $F$ 's  $< 1$ .

## Procedural Knowledge

Children's procedural knowledge increased from midtest ( $M = 26\%$ ,  $SE = 3\%$ ) to posttest ( $M = 37\%$ ,  $SE = 3\%$ ), and stayed similar two weeks later ( $M = 32\%$ ,  $SE = 3\%$ ).

There were no main effects of feedback or feedback type, nor did feedback type interact with prior knowledge,  $F$ 's  $< 1$ . However, consistent with Fyfe et al. (2011), there was a feedback by prior knowledge interaction,  $F(1, 87) = 4.67$ ,  $p = .03$ ,  $\eta_p^2 = .05$ . As prior knowledge increased, the benefits of feedback decreased ( $B = -1.06$ ,  $SE = 0.49$ ). To help interpret the interaction, we categorized children as having higher prior knowledge (scored above the median on the procedural knowledge pretest measure) or low prior knowledge and examined the main effects of feedback for each group (see Figure 1). For the low-knowledge group, children who received feedback exhibited higher procedural knowledge ( $M = 33\%$ ,  $SE = 4\%$ ) than children who did not receive feedback ( $M = 20\%$ ,  $SE = 5\%$ ),  $F(1, 87) = 4.00$ ,  $p = .05$ ,  $\eta_p^2 = .04$ . For the higher-knowledge group, children who received feedback exhibited lower procedural knowledge ( $M = 28\%$ ,  $SE = 5\%$ ) than children who did not receive feedback ( $M = 50\%$ ,  $SE = 6\%$ ),  $F(1, 87) = 7.54$ ,  $p = .007$ ,  $\eta_p^2 = .08$ . Feedback during exploration was more beneficial than no feedback for children with low prior knowledge, but for children with higher prior knowledge, the reverse was true. Feedback type did not matter, suggesting that both types of feedback were beneficial for low-knowledge children, and both types of feedback were detrimental for higher-knowledge children.

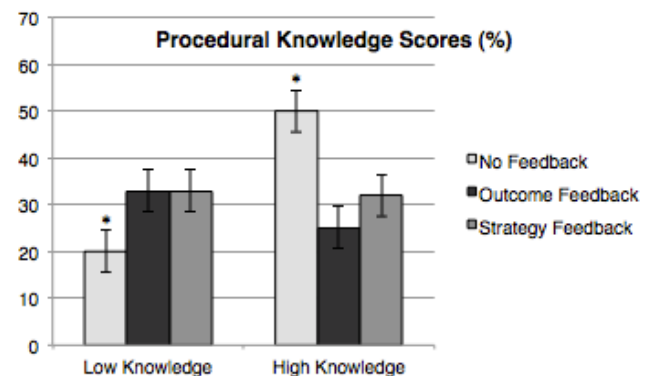


Figure 1: Percent correct on procedural knowledge assessment by condition and prior knowledge. Scores are estimated marginal means based on midtest, posttest, and retention test scores.

## Conceptual Knowledge

Children's conceptual knowledge increased from midtest ( $M = 21\%$ ,  $SE = 2\%$ ) to posttest ( $M = 50\%$ ,  $SE = 2\%$ ) and stayed similar at retention test ( $M = 43\%$ ,  $SE = 2\%$ ).

There were no main effects of feedback or feedback type, nor did feedback type interact with prior knowledge,  $F$ 's  $< 1$ . However, there was a marginal feedback by prior knowledge interaction,  $F(1, 87) = 3.63$ ,  $p = .06$ ,  $\eta_p^2 = .05$ .

As prior knowledge increased, the benefits of feedback tended to decrease ( $B = -0.70$ ,  $SE = 0.37$ ). To help interpret the marginal interaction, we examined the effect of feedback for low- and higher-knowledge children separately (based on a median split of procedural knowledge pretest scores; see Figure 2). For the low-knowledge group, children who received feedback exhibited somewhat higher conceptual knowledge ( $M = 44\%$ ,  $SE = 3\%$ ) than children who did not receive feedback ( $M = 37\%$ ,  $SE = 4\%$ ),  $F(1, 87) = 2.56$ ,  $p = .11$ ,  $\eta_p^2 = .03$ . For the higher-knowledge group, children who received feedback exhibited somewhat lower conceptual knowledge ( $M = 29\%$ ,  $SE = 3\%$ ) than children who did not receive feedback ( $M = 39\%$ ,  $SE = 5\%$ ),  $F(1, 87) = 2.60$ ,  $p = .11$ ,  $\eta_p^2 = .03$ . Although not reliable, particularly when dichotomizing prior knowledge, these results resemble the pattern of findings found for procedural knowledge.

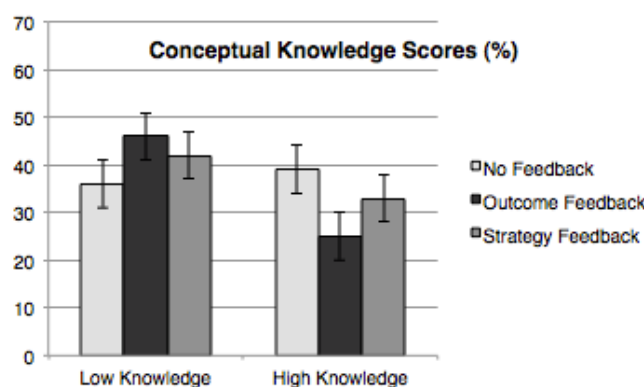


Figure 2: Percentage correct on conceptual knowledge assessment by condition and prior knowledge. Scores are estimated marginal means based on midtest, posttest, and retention test scores.

## Discussion

Guided discovery generally facilitates deeper learning than discovery or instruction alone (e.g., Alfieri et al., 2011; Mayer, 2004). For example, providing exploratory activities with subsequent instruction can be beneficial (e.g., DeCaro & Rittle-Johnson, 2011; Schwartz & Bransford, 1998). However, the amount of guidance provided during the exploratory activities has largely gone unstudied, leaving questions as to how and for whom the guidance can work. In a previous experiment, we attempted to address these questions by examining the effects of feedback during exploratory problem solving prior to instruction. Some children received feedback (on their answer or on their strategy) after solving each problem, while others did not. In this study, we strengthened the condition manipulation and verified the results with an independent sample of children. Our results were consistent with those in Fyfe et al.'s original report. For children with low prior knowledge, feedback led to higher procedural knowledge than no-feedback. But for children with higher prior knowledge, feedback hindered performance relative to no-feedback.

There was a similar, but weaker effect for conceptual knowledge. Feedback type had little effect in general. Overall, we replicated the previous findings and provided evidence for the reliability of the results.

The results are consistent with prior work demonstrating aptitude by treatment interactions, which demonstrate that a single instructional method is often not best for learners with varying levels of prior knowledge (Cronbach & Snow, 1977; Kalyuga, 2007). In particular, a common conclusion is that low-knowledge learners benefit from more guidance, while high-knowledge learners benefit from less guidance (Snow & Swanson, 1992). Aptitude by treatment interactions have been found in a variety of domains including math, science, and problem solving (see Kalyuga et al., 2003). The current study (coupled with Fyfe et al., 2011) extends the aptitude by treatment interaction work to the presentation of feedback during exploratory problem solving. Children who enter the situation with low knowledge of the domain need feedback to improve their knowledge of correct procedures. Children with higher domain knowledge, on the other hand, do not need this feedback and actually perform better without it. This occurred even though higher knowledge children in our study were far from experts and still had a lot to learn.

Despite the growing evidence that prior knowledge moderates the impact of feedback during problem solving (e.g., Fyfe et al., 2011; Kraise et al., 2009; Luwel et al., 2011), the reasons underlying this effect remain unclear. One potential explanation relies on the learner's experience of cognitive load (Paas, Renkl, & Sweller, 2003). For low-knowledge learners, novel tasks can easily overload their working memory; thus, they often need some form of external guidance to reduce cognitive load. In contrast, higher-knowledge learners can use their existing, relevant schemas to help them complete the task without cognitive overload; thus, they often do not need external guidance. This may explain why low-knowledge learners benefited from feedback, but high-knowledge learners did not. It is also possible that differences in motivation would help explain the findings. Children who are more knowledgeable may also be more motivated to learn. In turn, those who are more motivated may thrive in less structured, challenging environments whereas children who are less motivated may not (Schnotz, 2010). Finally, changes in children's strategy knowledge may also play a role. For low-knowledge children, the constraining effects of feedback may have sped up the process of strategy acquisition, which in turn could jumpstart subsequent strategy changes including the strengthening of correct strategies (Siegler, 1996). However, for higher-knowledge children, the constraining effects of feedback may not have been necessary since these children already knew a correct strategy. More work is needed to tease apart these alternative explanations.

Our results also have important implications for research on guided discovery learning. They suggest that prior knowledge (and other learner characteristics) should be considered when evaluating the efficacy of guided discovery

methods (Cronbach & Snow, 1977). They also highlight the need to evaluate and optimize different aspects of guided discovery techniques. We examined the amount of guidance provided during exploration prior to instruction and found that more was not always better. Unfortunately, even when researchers recognize the benefits of combining exploration and instruction, the recommendation is usually to include more guidance (e.g., Alfieri et al. 2011).

Despite the positive contributions of the current study, future research is needed. For example, researchers should continue investigating the effects of feedback type. We did not detect many differences between outcome feedback and strategy feedback, but past research suggests strategy-feedback can be more beneficial, at least in terms of strategy selection (Luwel et al. 2011). Further, research should more carefully address what counts as sufficient prior knowledge. As more research finds that the effectiveness of instruction depends on prior knowledge, instructors will need guidance on how to choose instructional techniques for particular children with particular levels of prior knowledge.

This study extends research on guided discovery learning in which exploration is provided prior to direct instruction. Providing feedback during the initial exploration facilitates learning for low- but *not* higher-knowledge children. Thus, providing feedback may not always be optimal.

### Acknowledgments

The first author is supported by a predoctoral training grant provided by the Institute of Education Sciences, U.S. Department of Education, through Vanderbilt's Experimental Education Research Training grant (ExpERT; David S. Corday, Director; grant number R305B080025). This work was also supported by an NSF CAREER grant (#DRL-0746565) awarded to Bethany Rittle-Johnson.

### References

- Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbarn, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, 103, 1-18.
- Alibali, M. W. (1999). How children change their minds: Strategy change can be gradual or abrupt. *Developmental Psychology*, 35, 127-145.
- Cronbach, L. J. & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.
- DeCaro, M., & Rittle-Johnson, B. (2011, March). Preparing to learn from math instruction by solving problems first. In B. Rittle-Johnson and M. DeCaro (chairs), *When are times for telling? Preparing students to learn from instruction*. Symposium presented at the Society for Research in Child Development Conference, Montreal.
- Fyfe, E. R., Rittle-Johnson, B., & DeCaro, M. S. (2011, Sept.). The effects of feedback during exploratory math practice. Paper presented at the Society for Research on Educational Effectiveness Conference, Washington, DC.
- Hiebert, J., & Grouws, D. (2007). The effects of classroom mathematics teaching on student learning. In F. K. Lester, *Second handbook of research on mathematics teaching and learning*. Charlotte, NC: Information Age Publishing.
- Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, 19, 509-539.
- Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. *Journal of Educational Psychology*, 96, 558-568.
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, 38, 23-31.
- Kluger, A. N., & DeNisi, A. (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119, 254-284.
- Kraise, U.-M., Stark, R., & Mandl, H. (2009). The effects of cooperative learning and feedback on e-learning in statistics. *Learning and Instruction*, 19, 158-170.
- Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, 47, 211-232.
- Luwel, K., Foustana, A., Papadatos, Y., & Verschaffel, L. (2011). The role of intelligence and feedback in children's strategy competence. *Journal of Experimental Child Psychology*, 108, 61-76.
- Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist*, 59, 14-19.
- McNeil, N. M. (2008). Limitations to teaching children  $2 + 2 = 4$ : Typical arithmetic problems can hinder learning of math equivalence. *Child Development*, 79, 1524-1537.
- Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38, 1-4.
- Rittle-Johnson, B., Matthews, P., Taylor, R., & McEldeen, K. (2011). Assessing knowledge of math equivalence: A construct modeling approach. *Journal of Educational Psychology*, 103, 85-104.
- Schnotz, W. (2010). Reanalyzing the expertise reversal effect. *Instructional Science*, 38, 315-323.
- Schwartz, D., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction*, 16, 475-522.
- Schwartz, D., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics. *Cognition and Instruction*, 22, 129-184.
- Schwartz, D., Chase, C., Oppezzo, M., & Chin, D. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, 103, 759-775.
- Siegler, R. (1996). *Emerging minds: The process of change in children's thinking*. NY: Oxford University Press.
- Snow, R. E., & Swanson, J. (1992). Instructional psychology: Aptitude, adaptation, and assessment. *Annual Reviews in Psychology*, 43, 583-626.
- Stigler, J. W., & Hiebert, J. (1998, Winter). Teaching is a cultural activity. *American Educator*, 1-10.

# Children's Inferences in Generalizing Novel Nouns and Adjectives

Annie Gagliardi (acg39@umd.edu)

Erin Bennett (ebennet2@umd.edu)

Jeffrey Lidz (jlidz@umd.edu)

Naomi H. Feldman (nhf@umd.edu)

Department of Linguistics, University of Maryland, College Park, MD 20742 USA

## Abstract

By the time children begin to rapidly acquire new word meanings they are already able to determine the grammatical category of novel words based on syntactic and morphological cues. Here we test whether children can leverage this knowledge when inferring the meaning of a novel word. Through a novel word learning experiment we determine that children can use this information, drawing different conclusions for the most likely meanings of novel words in distinct grammatical categories. We use a Bayesian model to formalize the higher level knowledge that children might have about noun and adjective meanings. Simulations show that children's behavior reflects the type of shift we would predict on the basis of noun and adjective meanings in the English lexicon.

**Keywords:** language acquisition; word learning; Bayesian inference

One of the most striking phenomena in language acquisition is children's ability to rapidly learn the meanings of novel words with only limited exposure. How exactly children do this has been researched extensively, with three lines of inquiry dominating the attempts to formalize this process: *hypothesis elimination* (Berwick, 1963; Pinker, 1989; Siskind, 1996), *associative learning* (Colunga & Smith, 2005; Regier, 2005) and *Bayesian inference* (Xu & Tenenbaum, 2007). Xu and Tenenbaum argue that Bayesian inference is superior to hypothesis elimination and associative learning because it uniquely allows the learner to take advantage of 'suspicious coincidences' when learning words for overlapping concepts. For example, in a word learning experiment they found that when children were shown three Dalmatians labeled with a novel object label, there was a strong bias for children to think that the novel word meant Dalmatian, rather than dog, or animal. This bias was not as strong when children only saw one Dalmatian labeled with the novel label. Neither hypothesis elimination nor associative learning predict the effect of the suspicious coincidence that results from the narrow distribution of exemplars on the kind hierarchy (which is in turn contingent on the number of exemplars). Xu and Tenenbaum's model does predict this effect, via the likelihood term, which takes into account both the number of exemplars and the size of the hypothesis.

One key assumption that Xu and Tenenbaum relied on was that the candidate concepts fell on a hierarchy of kinds. That is, in their model the learner does not have to determine what domain to generalize across, as this domain was given by the kind hierarchy. This assumption has two implications for their model: (1) most of the work in hypothesis selection is being done by the likelihood, as the prior probability of each hypothesis is comparatively much less variable and

(2) it largely limits the model to the discussion of object label learning, as this is the domain that primarily uses the kind hierarchy.

In this paper we probe the predictions of the Bayesian model on different grammatical categories, nouns and adjectives, which tend to draw from different concept hierarchies. This allows us to better test the role of the prior probability of a concept given a grammatical category by letting us examine the link between grammatical category and concept hierarchy. Toward these goals we conducted a word learning experiment that replicates Xu and Tenenbaum's finding with learning novel nouns, and extends the paradigm to novel adjective learning. We find that children use the grammatical category of the novel word to constrain their hypotheses about the meaning of the novel word. This is demonstrated through their sensitivity to the suspicious coincidence in the distribution of exemplars on the kind hierarchy when learning nouns but not adjectives. A Bayesian model that takes into account not only conceptual similarity but also the link between grammatical category and concept matches the qualitative shift between nouns and adjectives seen in the children's data. The model's ability to capture this shift highlights the crucial role that children's prior beliefs contribute to their generalizations in word learning. Through this work we extend the Bayesian model of word learning in ways that make it more realistic with respect to both the structure of natural language and the task faced by a child acquiring novel words.

Our paper is organized as follows. We first present our word learning experiment. We then use a Bayesian model to formalize children's prior distribution over concepts. The next section presents simulations comparing the model to children's behavior. We conclude by discussing the implications that this work has for language acquisition, in particular the importance of considering how a learner's prior knowledge affects the way in which the data from the environment are used in language acquisition.

## Word Learning Experiment

In a novel word learning experiment children were presented with an array of animals and vehicles and taught a novel label (noun or adjective) for a concept. Children were then asked to generalize their inferred concept to novel items. The stimuli allowed generalization along both kind and property dimensions. If children are able to use syntactic information to constrain their inference of words' meanings, then we should expect them to generalize differently when learning nouns versus adjectives.





Figure 1: The stimuli for our experiment included 36 objects in subordinate, basic, and superordinate vehicle and animal categories. Half the items were striped and half spotted.

## Methods

Our experiment tested two groups of children using a between subjects design. The noun group learned two novel nouns, and the adjective group learned two novel adjectives.

**Participants** Participants were 24 children (mean age = 4;0, range = 3;6-5;0) recruited from the greater College Park area as well as an on campus preschool. Children either visited the lab with their parents or were visited by researchers at their preschool. Four children were excluded from the final analysis for the following reasons. One was too shy to interact with the snail and three said they didn't know when they were asked to perform the generalization task outlined below.

**Stimuli** All children were presented with an array of pictures (Figure 1) that included 36 items from two superordinate categories on the kind hierarchy (18 vehicles and 18 animals). Each category had items from several basic levels (animals: 12 dogs, 2 cats, 2 squirrels, 2 owls; vehicles: 12 roofed cars, 2 convertibles, 2 vans, 2 trucks). One basic level from each superordinate category had items from two subordinate level categories (dogs: 6 Dachshunds and 6 Yorkshire terriers, roofed cars: 6 taxis and 6 police cars). There were both striped and spotted items of each item type.

**Procedure** A snail puppet was introduced to the child, and the child was told that the snail spoke a funny snail language that was mostly like English but included some new words. The experimenter explained that they would try to figure out the snail's words by listening to him talk about some of the pictures. Before proceeding further, the experimenter checked that both the snail and the child could see all of the pictures in the array. This ensured that participants were aware of the range of items in the experimental world.

During the **word learning phase** the snail looked at the pictures and pointed out an item from one of the subordinate

Speaker	Utterance	Action
Snail	'This is a <i>blicky one</i> '	points to striped Dachshund 1
Snail	'Look, another <i>blicky one</i> '	points to striped Dachshund 2
Snail	'Here's another <i>blicky one</i> '	points to striped Dachshund 3
Snail	'I'm going to go have a rest'	retreats to shell
Experimenter	'Here are some more pictures. Can you put circles on all the <i>blicky ones</i> to surprise the snail?'	lays out new array of pictures and gives the child a set of rings
Child	—	puts rings on items that match child's hypothesis for the meaning of <i>blicky</i>

Table 1: Sample adjective trial. Noun trials were identical with *blick* substituted for *blicky one*.

level categories (e.g., a striped dachshund). In the noun condition he described it as *a blick*, and in the adjective condition he described it as *a blicky one*. This happened 3 times, with the snail pointing to a different striped dachshund each time. Then the snail would get tired and retire to his shell for a nap.

While the snail slept, the experimenter initiated the **test phase**, during which the child was presented with another array of pictures and asked to place circles on the other *blicks* (noun condition) or *blicky ones* (adjective condition). A single trial is schematized in Table 1. The entire procedure was repeated for a second novel word used to describe another item from a different subordinate level (e.g., a spotted taxi). Order of item (dog before vehicle or vice versa), described subordinate level item (dachshund vs yorkie and taxi vs police car), and described pattern order (striped before spotted and vice versa) were counterbalanced across subjects.

## Results

Each item presented during the word learning phase was consistent with seven candidate concepts (illustrated in Table 2), picking from the kind hierarchy, property hierarchy or combining concepts from both. For data analysis, children's hypotheses were collapsed depending where the generalization fell on the kind hierarchy. Children's choices were coded as follows, with one response recorded per trial. Subordinate responses were recorded if children chose only items from the same subordinate level as the example (e.g., only dachshunds after being presented with dachshunds). Basic responses were recorded if children chose from only the basic level (e.g. either dog type after being presented with

Hypothesis	Dimension	Level
Dachshund	Kind	subordinate
Dog	Kind	basic
Animal	Kind	superordinate
Striped	Property	neutral
Striped $\wedge$ Dachshund	Property $\wedge$ Kind	subordinate
Striped $\wedge$ Dog	Property $\wedge$ Kind	basic
Striped $\wedge$ Animal	Property $\wedge$ Kind	superordinate

Table 2: Candidate concepts, given three exemplars of striped Dachshunds.

dachshunds) or from the basic and subordinate levels. A superordinate response was recorded if children chose only from the superordinate level (e.g., any animal after being presented with dachshunds) or from the superordinate level with any combination of the lower levels. Finally, neutral responses were recorded if children chose from anywhere on the kind hierarchy (e.g., anything from the vehicle hierarchy after being shown a dachshund). All items chosen by children with neutral responses were consistent with the property (either striped or spotted) that they had been taught.

Results are shown in Figure 2(a). In the noun condition, we replicated Xu & Tenenbaum’s finding, uncovering a bias for the subordinate level meaning when all observations fall into the same subordinate level. In the adjective condition, however, we see a different pattern. The placement of the item on the kind hierarchy had no bearing on children’s choices, with the overwhelming majority choosing the neutral interpretation, indicating their belief that the novel adjective’s meaning referred just to the most salient property (striped versus spotted) rather than the kind. Planned comparisons revealed that the proportion of trials that children chose the subordinate and neutral meanings differed significantly by condition (subordinate:  $t(33) = 3.49, p < 0.002$ , neutral:  $t(26) = 3.39, p < 0.003$ ).

## Discussion

These results demonstrate that children use their knowledge of grammatical categories, and the associated kinds of meanings that correlate with these categories, when inferring the meanings of novel words. In particular, they favor concepts from a kind hierarchy for novel nouns, and from a property hierarchy for novel adjectives. In one respect this result is not new, as infants as young as 14 months have been shown to know the mapping between grammatical and conceptual categories (Waxman & Markow, 1998; Booth & Waxman, 2003, 2009). Instead, the novelty is in showing that this mapping constrains children’s inferences. A very low prior probability for a hypothesis on the kind hierarchy blocks it from being determined the most likely for a novel adjective meaning, despite it being the narrowest possible hypothesis.

This finding emphasizes the role of the hypothesis space, since the most likely hypothesis differs depending on the grammatical category of the word being learned. In order

to determine whether children are behaving optimally with respect to a specific hypothesis space (conditioned by grammatical category and the information available to them in the English lexicon), we used a Bayesian model to predict generalization behavior from the nouns and adjectives that are likely to be present in the children’s early lexicons.

## Model

We assume the generative model shown in Figure 3(a). Our model assumes that the snail in our experiment, having chosen a grammatical category for the word he will teach the children, chooses a concept to teach (such as *dog*, *striped*, or *dachshund*), and then independently chooses three objects as examples of that concept.

The children in our experiment inferred what concept a new word referred to based on the grammatical category of the novel word (noun or adjective) and the objects the snail identified as examples of that word. Our model therefore computes the probability of each concept  $C$  for a given grammatical category  $P$  and set of objects  $X$ ,

$$\mathbb{P}(C|X, P) \quad (1)$$

We can use Bayes’ rule to compute the posterior probability over concepts given a set of examples and a word’s grammatical category,

$$\mathbb{P}(C_i|X, P) = \frac{\mathbb{P}(X|C_i) \cdot \mathbb{P}(C_i|P)}{\sum_{C_j \in \{\text{all concepts}\}} \mathbb{P}(X|C_j) \cdot \mathbb{P}(C_j|P)} \quad (2)$$

We assume that the probability of the data  $X$  depends only on the concept  $C$  and is independent of the grammatical category, given the concept. Since the normalizing constant in the denominator will be the same for all candidate concepts, we only need to find the values of  $\mathbb{P}(X|C_i)$  and  $\mathbb{P}(C_i|P)$  for the concepts we are considering.

### Concept Prior: $\mathbb{P}(C|P)$

Following Goodman, Tenenbaum, Feldman and Griffiths (2008) (cf. Austerweil & Griffiths, 2010), we represent concepts according to the concept grammar in Figure 3(b), with nonterminal nodes *Kind* and *Property* representing the dimensions a concept is defined along. Words like *dog* and *striped* are defined along only one of these dimensions (*Kind* and *Property*, respectively). Words like *kitten*, which describes a young cat, are defined along both dimensions (*Kind  $\wedge$  Property*). The derivation of each concept involves first applying a rule determining the dimension of the concept and then applying the dimension-specific rules until all terminal nodes have been identified. For example, in our concept language, the concept *dog* is formed by first applying the rule *Concept*  $\rightarrow$  *Kind* and then applying the rule *Kind*  $\rightarrow$  *dog*.

If we assign probabilities to each of the rules in this concept grammar and assume that the rules are applied independently of one another, then the resulting PCFG will determine the

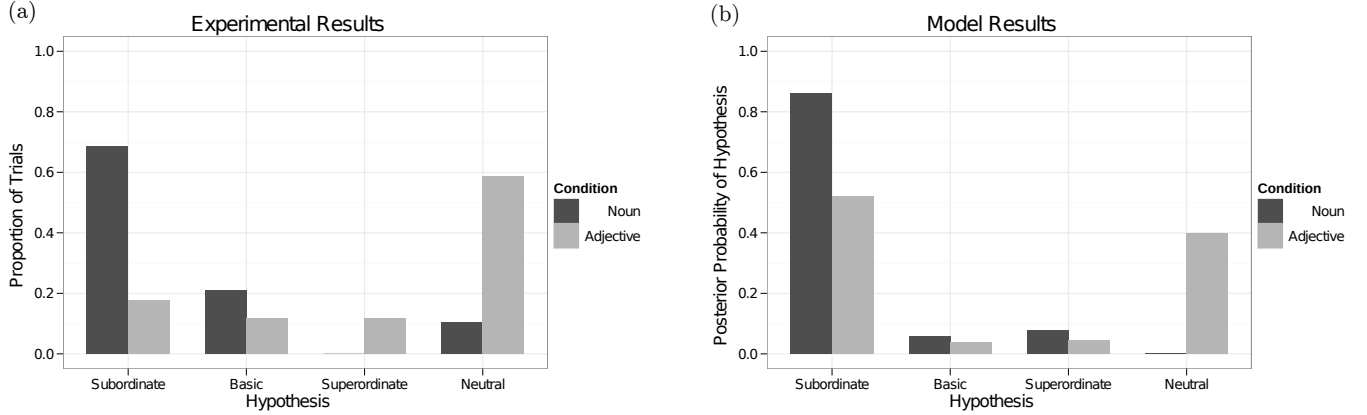


Figure 2: (a) Results of word learning experiment and (b) results of modeling.

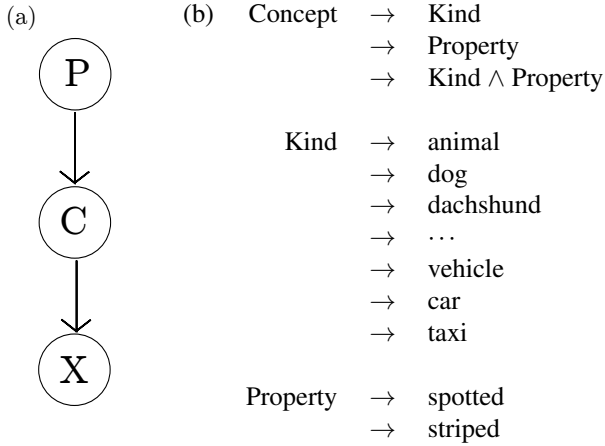


Figure 3: (a) Grammatical categories  $P$  determine the parameters for our prior over concepts  $C$ . Specific objects  $X$  are sampled from the set of items that exemplify a concept. (b) A probabilistic context-free grammar for concepts. Probabilities for each expansion rule are discussed in the Concept Prior section.

probabilities of all the concepts in our experiment. The probability of each concept would be the product of the probabilities of the rules applied to form it,

$$\mathbb{P}(C) = \prod_{R \in \{\text{rules to form } C\}} \mathbb{P}(R) \quad (3)$$

The differences in the types of concepts denoted by nouns and adjectives are represented in our model through differences in the probability distributions over the set of rules that expand *Concept* to particular dimensions. We assume children are computing this prior distribution separately for each part of speech, keeping track of the number of nouns or adjectives whose meanings denote a kind, a property, or both a kind and a property. They can estimate the rule probabilities from these counts using a Dirichlet-multinomial model. Under this model, the prior over dimension expansions based on

	Kind	Property	Both
Noun	335	4	24
Adjective	3	61	2

Table 3: Average counts (rounded) from 22 participants’ ratings of nouns and adjectives as descriptions of kinds, properties, or both.

the counts  $p_{d_i,P}$  of the productions seen by the learner of a particular dimension  $d_i$  for that grammatical category  $P$  is

$$\mathbb{P}(d_i|P) = \frac{p_{d_i,P} + 1}{\sum_{d_j \in \{\text{all dims}\}} p_{d_j,P} + 3} \quad (4)$$

We approximated these production counts from a Mechanical Turk survey where for each word in a vocabulary list of 429 words (363 nouns and 66 adjectives) that 30-month-old children likely know (Dale & Fenson, 1996), we asked adult English speaking participants to judge whether the word was best described as a kind, a property, or both. Different but often overlapping sets of 10 people were asked to respond to each word, and so we had a total of 22 participants in our study. Two participants’ judgments were excluded due to an extraordinarily high proportion of *Both* responses (proportion *Both* > 0.36, over two standard deviations outside the mean proportion of *Both* responses). While the children in our experiment (3-5 year-olds) were much older than 30 months, we believe that this vocabulary list is appropriate for our purposes, since the children in our experiments are almost certainly familiar with these words and differ only in additional words they might know. We assume that the distribution of noun and adjective dimensions in this set of words is representative of that of the larger and more varied set of words that our participants are familiar with. Table 3 shows the average counts of each description for each grammatical category.

For kinds, we assume a structure like Xu and Tenenbaum (2007) where the probability of a concept depends on its distinctiveness. For these measures we use a hierarchical cluster



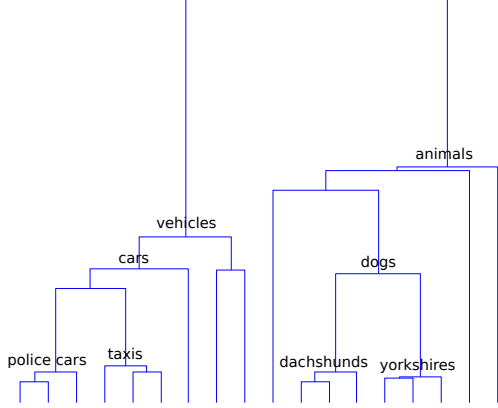


Figure 4: Hierarchical clustering of experimental item similarity.

tree, shown in Figure 4. To make this tree, we conducted a similarity judgment study, similar to Xu and Tenenbaum’s using the items that the snail had labeled in our experiment. Our participants, 26 students from the University of Maryland who received course credit for their participation, rated the similarity of all possible pairs of the 36 pictures on a scale from 1 (not similar at all) to 9 (very similar).

To incorporate cluster distinctiveness, Xu and Tenenbaum measure the branch length (which represents the Euclidean distance) between the concept node and its parent node. By this measure, the further a particular node is from its parent, the more distinct it is considered to be. Where  $\mathcal{K}$  is the set of all *Kind* concepts, the probability of a concept  $C_i$  given that it is defined over the *Kind* dimension is the branch length normed over all *Kind* concepts,

$$\mathbb{P}(C_i|\text{Kind}) = \frac{\text{height}(\text{parent}(C_i)) - \text{height}(C_i)}{\sum_{C_j \in \mathcal{K}} \text{height}(\text{parent}(C_j)) - \text{height}(C_j)} \quad (5)$$

For properties, we assume that in our experiment they are chosen from a multinomial distribution with each property equally likely to be selected. Since there were only two very salient properties in our experiment, we give each property the probability of  $\frac{1}{2}$ ,

$$\mathbb{P}(C|\text{Property}) = \frac{1}{2} \quad (6)$$

**Example Derivation of a Concept Prior** Under this model of the concept prior, the prior probability that the noun *bllick* refers to the concept *Dachshund* will have the following derivation. First, we have production counts for nouns that describe kinds  $p_{\text{Kind}, \text{Noun}}$  that were found in our Mechanical Turk study (we found that on average 335 out of 363 nouns were categorized as kinds). From this production count and the total production counts for nouns, we derive the probability of expanding *Concept* to *Kind*.

$$\begin{aligned} \mathbb{P}(\text{Kind}|\text{Noun}) &= \frac{p_{\text{Kind}, \text{Noun}} + 1}{\sum_{d \in \{\text{Kind}, \text{Property}, \text{Both}\}} p_{d, \text{Noun}} + 3} \\ &= \frac{335 + 1}{363 + 3} = 0.92 \end{aligned} \quad (7)$$

Then we find the probability of the concept being *Dachshund* given that it is defined only along the *Kind* dimension, using the height of the branch *Dachshund* and its immediate parent *dog*. These heights were 0.1259 and 0.3115, respectively.

$$\begin{aligned} \mathbb{P}(\text{dachshund}|\text{Kind}) &= \frac{\text{height}(\text{parent}(\text{dog})) - \text{height}(\text{dog})}{\sum_{C \in \mathcal{K}} \text{height}(\text{parent}(C)) - \text{height}(C)} \\ &= \frac{0.1856}{1.7576} = 0.1056 \end{aligned} \quad (8)$$

Finally, to compute the prior probability of the concept *Dachshund* given that it is a noun, we multiply the probability of expanding *Concept* to *Kind* by the probability of the concept being *Dachshund*.

$$\begin{aligned} \mathbb{P}(\text{Dachshund}|\text{Noun}) &= \mathbb{P}(\text{Kind}|\text{Noun}) \cdot \mathbb{P}(\text{Dachshund}|\text{Kind}) \\ &= 0.92 \cdot 0.1056 = 0.09715 \end{aligned} \quad (9)$$

### Concept Likelihood: $\mathbb{P}(X|C)$

We assume that, given a set of objects that are examples of a concept  $C$ , each object is equally likely to be chosen by the snail.<sup>1</sup> Therefore, the probability of the data given a concept is proportional to the size of the set of things matching that concept. For example, for the concept *dog*, the probability of picking a particular dog, Fido, is inversely proportional to the number of dogs there are in the scene. So if  $n$  objects are chosen by the snail as examples of a concept  $C$ , and these objects are plausible examples of the concept,

$$\mathbb{P}(X|C) = \left( \frac{1}{|C|} \right)^n \quad (10)$$

### Simulations

For each experimental trial we computed the posterior probability over concepts using both the noun and adjective priors. We assumed that on each trial children were sampling a concept from the posterior distribution over concepts given the

<sup>1</sup>Xu and Tenenbaum use a different estimate of category sizes for kinds, which is based on the same heirarchy as their concept prior. We found little difference when we compared the our own likelihood distributions with those computed by Xu and Tenenbaum’s methods on our experimental items. A very similar ordering applied over concepts, and each item was on the same order of magnitude for both measures of the likelihood.

grammatical category of the novel word. Thus the posterior probability over concepts as generated by the model should give us the frequency with which a child should show any given behavior. In order to be able to compare the model to the experimental data, we sorted the concepts into the same categories that we used for analyzing the experimental data: subordinate, basic, superordinate and neutral. The candidate concepts for *striped Dachshund*, along with the levels they mapped on to, are found in Table 2 (in the Results section of the Word Learning Experiment, above).

The results of our model are shown in Figure 2(b). Overall the model captures the qualitative shift seen between noun and adjective generalization in the experimental results, with a much higher posterior probability for the subordinate level given a noun, and a shift of a large part of the probability to the neutral level given an adjective.

## Discussion

In this paper we have shown that while children tend to map novel nouns onto a kind hierarchy, they prefer to map novel adjectives onto a property hierarchy. This behavior is predicted if children use their knowledge of grammatical categories and the distributions of different concept types within these categories to constrain the space of hypothesized meanings when learning novel words. A Bayesian model trained on the distribution of concepts across grammatical categories in the English lexicon predicts a qualitatively similar generalization pattern. Together these results suggest that not only are children able to use what they know about grammatical categories when inferring the meanings of novel words, the way they do this is predicted by the distributions of concept types across grammatical categories in English. Moreover, the constraints imposed on inference by grammatical category are powerful enough to overcome much of the effect of the size principle on the likelihood.

These findings have several implications for language acquisition and models of language acquisition. First, while the ‘size principle’ has received considerable attention as a solution to the word learning problem, this work demonstrates that the beliefs children bring to the word learning task also play a key role in word learning. Second, while a model based on priors derived from the lexicon captured the shift we see between noun adjective generalizations, it does not perfectly predict children’s behavior. Future research will probe whether this can be explained by making a closer approximation of the child’s lexicon (as our Mechanical Turk task may have overestimated the number of *both* concepts), or whether it stems from the learner’s tendency to amplify biases that exist in the input (e.g. Hudson-Kam & Newport, 2009). Third, we can ask how children behave with respect to concept hierarchies in languages that collapse the distinction between nouns and adjectives (e.g., Georgian). Does the size principle play a role only to the extent that nouns are likely to draw from the kind hierarchy? Next, as these beliefs are attributable to the distribution of concept types across grammat-

ical categories in the children’s own lexicons, there are obvious extensions of this work to modeling the infant word learning by weakening (or making nonexistent or unavailable) the link between grammatical category and concept hierarchy. There are several findings that would be interesting to model this way, including (1) that 11-month-olds make the same generalizations for words presented as nouns and adjectives and these generalizations are neutral with respect to kind vs. property meanings (Waxman & Booth, 2003), or (2) that the noun-kind link is established earlier than the adj-property link (Booth & Waxman, 2003, 2009). Finally, we can ask to what degree a group of exemplars’ distribution on a given concept hierarchy is used in acquiring linguistic phenomena that extend beyond word meanings (e.g., word classes).

**Acknowledgments.** This research was supported by NSF IGERT 0801465, a NSF GRF to Gagliardi, and a Baggett Fellowship to Bennett. We would like to thank the UMD Cognitive Neuroscience of Language Lab, the UMD Project on Childrens Language Learning and the UMD Computational Psycholinguistics group for helpful discussion and assistance.

## References

- Austerweil, J. L., & Griffiths, T. L. (2010). Learning hypothesis spaces and dimensions through concept learning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 73–78). Austin, TX: Cognitive Science Society.
- Berwick, R. C. (1963). Learning from positive-only examples: The subset principle and three case studies. In J. G. Carbonell, R. S. Michalski, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (vol. 2). Los Altos, CA: Morgan Kaufmann.
- Booth, A. E., & Waxman, S. R. (2003). Mapping words to the world in infancy: Infants’ expectations for count nouns and adjectives. *Journal of Cognition & Development*, 4, 357–381.
- Booth, A. E., & Waxman, S. R. (2009). A horse of a different color: Specifying with precision infants’ mappings of novel nouns and adjectives. *Child Development*, 80(1), 15–22.
- Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, 112, 347–382.
- Dale, P., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, 28, 125–127.
- Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32, 108–154.
- Hudson-Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, 59, 30–66.
- Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.
- Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, 29, 819–865.
- Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61, 39–91.
- Waxman, S. R., & Booth, A. E. (2003). The origins and evolution of links between word learning and conceptual organization: New evidence from 11-month-olds. *Developmental Science*, 6(2), 130–137.
- Waxman, S. R., & Markow, D. B. (1998). Object properties and object kind: Twenty-one-month-old infants extension of novel adjectives. *Child Development*, 69, 1313–1329.
- Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2), 245–272.

# When Suboptimal Behavior is Optimal and Why: Modeling the Acquisition of Noun Classes in Tsez

Annie Gagliardi (acg39@umd.edu)

Naomi H. Feldman (nhf@umd.edu)

Jeffrey Lidz (jlidz@umd.edu)

Department of Linguistics, 1401 Marie Mount Hall, University of Maryland, College Park, MD 20742 USA

## Abstract

Children acquiring languages with noun classes (grammatical gender) have ample statistical information available that characterizes the distribution of nouns into these classes, but their use of this information to classify novel nouns differs from the predictions made by an optimal Bayesian classifier. We propose three models that introduce uncertainty into the optimal Bayesian classifier and find that all three provide ways to account for the difference between children's behavior and the optimal classifier. These results suggest that children may be classifying optimally with respect to a distribution that doesn't match the surface distribution of these statistical features.

**Keywords:** language acquisition; noun classes; Bayesian classification; statistical learning.

Learners are surrounded by statistical information. Considerable evidence suggests that they can make use of statistics to learn about their environment. For example, when acquiring artificial languages, children track distributional cues that allow them to discover phonetic categories (Maye, Werker & Gerken, 2002), word boundaries (Saffran, Newport & Aslin, 1996), grammatical categories (Mintz, 2003; Reeder, Newport & Aslin 2009, 2010), grammatical dependencies (Gomez & Maye, 2005; Saffran, 2001) and phrase structure (Takahashi, 2009). This leads to a commonly held belief in the language acquisition literature that children are perfect statistical learners (e.g. Elman, Bates, Johnson Karmiloff-Smith, Parisi & Plunkett 1996).

The hypothesis that children are perfect statistical learners predicts that when tested on their ability to generalize aspects of their native language in an experimental setting, children's linguistic knowledge should always reflect the distribution of statistical information in the input. However, this is not always the case. Work by Hudson-Kam and Newport (2009), for example, suggests that children are not perfectly veridical learners, in that they sometimes override statistical patterns in the service of amplifying some other facet of the language they are acquiring. As this work has largely focused on artificial language learning, here we examine another type of non-veridical statistical learning involving the acquisition of noun class (grammatical gender) in a natural language, Tsez. We present evidence showing that children exhibit behavior that is inconsistent with the statistical information available in the input when assigning novel nouns to noun classes. This inconsistent

behavior suggests that there is more to language acquisition than a simple mapping of external statistical information to an internal representation of this distribution. In particular it suggests that properties of the learner shape the statistical information in the input into the subset of information that is used to guide inferences in language acquisition: the intake. We use a Bayesian model of noun classification to probe what underlies the difference in the measureable *input* and the *intake* that children use to acquire noun classes.

As a general framework, we assume that optimal performance in an experimental task involves the following four components:

- (1) Accumulation of knowledge of the statistical distribution of features relating to some phenomenon
  - (2) Observation of features in a novel experimental item
  - (3) Knowledge of which features are relevant for the statistical computation
  - (4) Bayesian computation to determine how to generalize the phenomenon in question to the novel instance
- (1) depends on the learner's ability to observe and encode a statistical distribution of features pertaining to some phenomenon. (2) is similar to (1), but refers to encoding these features given a situation where the learner will be performing a computation to classify or otherwise deal with a novel instance. (3) requires the learner to know which features are relevant for a computation and is by no means trivial, as not every feature related to every phenomenon is relevant to the associated computation. (4) is an assumption that we are making about the kind of computations that learners use distributional information for. While step (4) is often assumed to be the culprit when learners show suboptimal performance in experimental tasks, in principle steps (1) through (3) can also contribute to suboptimal performance.

Our case study on Tsez noun classification examines how each of these pieces could result in a reshaping of the statistical information in the input. We begin with an outline of the distributional information that characterizes Tsez noun classes. We then compare children's use of this information in classification with that of a naïve Bayesian classifier. Finally, we explore three models that introduce uncertainty in levels (1)-(3) from above, in an effort to determine what underlies the difference between children's performance and predictions made by the Bayesian model.

## Tsez Noun Classes

Many languages make use of subclasses of nouns, called noun classes or grammatical gender. The presence and number of noun classes, as well as the distribution of individual nouns into classes varies greatly across languages, but several features remain constant. All noun class systems exhibit some degree of distributional information both internal and external to the noun. Noun internal distributional information consists of commonalities among the nouns in a class, such as semantic or phonological features. Noun external distributional information is made up of class defining information that is separate from the noun, such as agreement morphology that is contingent on noun class. We will look at noun class acquisition in Tsez as a case study.

Tsez, a Nakh-Dagestanian language spoken by about 6000 people in the Northeast Caucasus, has four noun classes. These classes can be characterized based on noun external distributional information (e.g. prefixal agreement on vowel initial verbs and adjectives) (Table 1), and noun internal distributional information (semantic and morphophonological features on the nouns themselves) (Table 2).

Table 1: Noun External Distributional Information.

Class 1	Class 2	Class 3	Class 4
Ø-igu uži	j-igu kid	b-igu k'et'u	r-igu čorpa
I-good boy	II-good girl	III-good cat	IV-good soup
<i>good boy</i>	<i>good girl</i>	<i>good cat</i>	<i>good soup</i>

Table 2: Noun Internal Distributional Information (a selection)

Feature	Value	Class predicted	% class with this feature value	% nouns with this value in predicted class
Semantic	female	2	13	100
Semantic	animate	3	22	100
First Segment	r-	4	9	61

Gagliardi and Lidz (under review) measured noun internal distributional information by taking all nouns from a corpus of Tsez child directed speech, tagging them for potentially relevant semantic and morphophonological cues and using decision tree modeling to determine which features were most predictive of class (cf. Plaster, Harizanov & Polinsky, in press). The features shown in Table 2 are only a selection of the most predictive features of class, with only the most predictive values of these features shown.<sup>1</sup> The full structure of each feature that we assume in our model is given below in Table 3. Each feature has specified values that were

<sup>1</sup> Here we talk about ‘noun classes’ to refer what is often called grammatical gender. One of the cues to noun class is often natural gender, but this is only one of several cues, and many other nouns are in each class that don’t have this (or potentially any) cue predicting their class.

highly predictive of some class and an unspecified value that ranges over all other possible values that were not predictive.

Table 3: Structure of Features

Feature	Specified Values	Unspecified Value
Semantic	male, female, animate	other
First segment	r-, b-	other
Last Segment	i	other

In this paper we will focus on how children use noun internal distributional information. In particular we will look at whether a child can make use of the predictive phonological and semantic information when classifying novel nouns, and how they perform when a noun has two features that make conflicting predictions. Returning to the four components of statistical learning outlined above, we will be looking at

- (1) Whether Tsez children have knowledge of the noun internal distributional information
- (2) Whether they can observe these features on novel nouns
- (3) Whether they assume all features are relevant for classification
- (4) We assume for the purposes of our analysis that the computation they make based on this information is Bayesian.

## Classifying Novel Nouns in Tsez

To assess whether children can use the statistics of noun internal information available in their input, we compare classification of novel nouns by Tsez acquiring children to the classification behavior that is predicted by a Bayesian model trained on the input data from our corpus. We describe the experimental data and the model in turn.

## Classification by Tsez Children

To determine whether or not children classified novel nouns consistently with the predictions made by the probabilities associated with their noun internal features, 10 native Tsez speaking children (mean: 6yrs, range: 4-7yrs) participated in a classification task. Here we give an overview of the experiment; for further details, including adult data, see Gagliardi and Lidz (under review).

**Method** Children were presented with unfamiliar items labeled with novel nouns by a native Tsez speaker. They were instructed to first tell a character to begin eating and then tell the character whether or not to eat the other labeled items. As the both the intransitive (eat) and transitive (eat it) forms for *eat* are vowel initial in Tsez (*-iš* and *-ac'o* respectively), classification of the novel word could be seen on the agreement prefix. Furthermore, intransitive verbs in Tsez agree with the agent (the eater) and transitive verbs agree with the patient (the thing eaten). An example trial is schematized in Table 4.

The test items had either a single noun internal distributional feature from Table 2, or a combination of these features that made conflicting predictions (e.g. semantic = [animate] and initial = [r]). The exact feature combinations used in this experiment, along with the classes each feature predicts, are shown in Table 5. While these only represent a selection of the most predictive features, we focus on them here as they are a representative set of predictive semantic and phonological features.

Table 4: Example Experimental Trial

Speaker	Utterance	Action/Conclusion
Experimenter	kid girl(class2) girl	Points to girl on page
Child	sis, q'ano, ɬono, j-iš one two three CL2-eat One two three, Eat!	Tells <i>kid</i> to start eating using Class 2 prefix <b>j</b> / <b>kid is in Class 2</b>
Experimenter	zamil novel[animate]	Points to unfamiliar animal and labels it with the novel noun <i>zamil</i>
Child	zamil b-ac'xosi aanu zamilCL3-eat-pres.part neg pro isn't eating the <i>zamil</i>	Says whether or not the girl is eating the <i>zamil</i> using Class 3 prefix <b>b</b> / <b>zamil is in Class 3</b>

Table 5: Features Used in Experiment and Simulations

Feature	Value	Class Predicted
Semantic	female	2
Semantic	animate	3
First Segment	r	4
Semantic & First Segment	female & r	2 and 4
Semantic & First Segment	animate & r	3 and 4

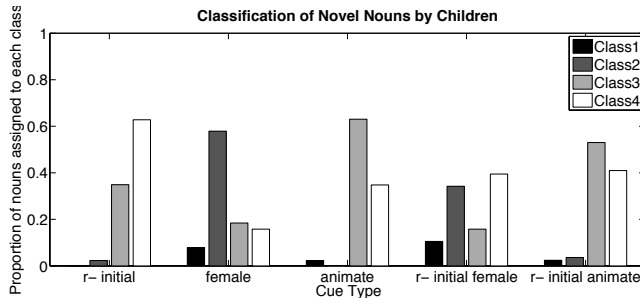


Figure 1: Proportion of novel nouns assigned to each class (by cue type) in the experimental task

**Results** The proportion of nouns that children assigned to each class are shown in Figure 2. When nouns had no conflicting features, children assigned more nouns to the class most strongly predicted by the feature than to any other class. However, when nouns had more than one

feature that made conflicting predictions, children relied more heavily on the phonological feature [r-] than on the semantic feature. This is not likely to be predicted by the distribution of these features in the input, where nouns with the [animate] and [female] values of the semantic feature never occur in Class 4.<sup>2</sup>

### Classification by an Optimal Bayesian Classifier

Given these experimental data, we can evaluate whether children are optimally using the statistics in their input by examining how a Bayesian model would classify each novel noun. That is, what would an ideal learner, exposed to input with these features, do when asked to classify novel words?

Our model is shown in Equation 1. The prior probability of a class  $p(c)$  corresponds to its frequency of occurrence, and the likelihood terms  $p(f|c)$  for each of  $n$  independent features  $f$  can be computed from feature counts in the lexicon.

$$p(c | f_1, f_2 \dots f_n) = \frac{p(f_1 | c)p(f_2 | c) \dots p(f_n | c)p(c)}{\sum_i p(f_1 | c_i)p(f_2 | c_i) \dots p(f_n | c_i)p(c_i)} \quad (1)$$

The results of classification with this model are shown in Figure 2. Just as we did with children, we tested the model on classification with each semantic and phonological feature from Table 2 individually, as well as cases where these features were in conflict with one another. As would be expected based on the relative strength of these features (Table 2), when semantic and phonological features make conflicting predictions the model classifies in line with the predictions made by the semantic feature.

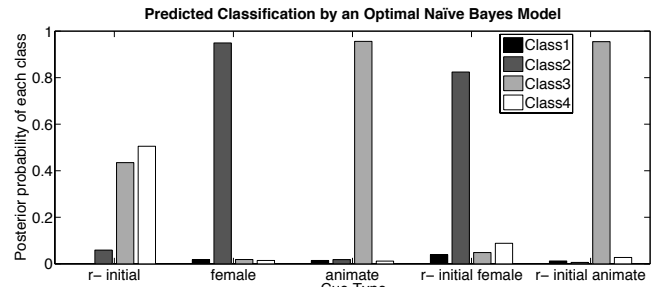


Figure 2: Predicted classification of novel nouns by an optimal naïve Bayesian classifier

The model's classification differs from that of the children in that when features made conflicting predictions the model relied on the statistically strongest cue (the semantic feature), while the children did not rely so heavily on this.

### Predicting Suboptimal Performance

While children roughly align with the model when classifying based on one highly predictive feature, they diverge when features make conflicting predictions. Children appear to use phonological features out of proportion with their statistical reliability. That is, children

<sup>2</sup> For a more detailed description of the results of the experiment, see Gagliardi & Lidz (under review).

appear to prefer the weaker predictions made by the phonological feature to the stronger ones made by the semantic feature. In order to determine the source of this asymmetry it is useful to first consider the fundamental differences between semantic and phonological features that could lead to this kind of behavior, and then to determine where and how these factors could affect our model.

There are several differences between semantic and phonological features that could affect their use in noun classification, but here we will focus on a fundamental difference in how reliably perceived and encoded each feature type may be during early acquisition. Every time a word is uttered (or most of the time, allowing for noisy conditions and fast speech) phonological features are present. However, especially during the early stages of lexical acquisition, the meaning of a word, and thus the associated semantic features, is much less likely to be available or apparent. Below we will consider how this sort of asymmetry could lead to a disparity in the way children end up using them in novel noun classification.

### Three Models of Uncertainty

The difference between semantic and phonological features could affect each of the three components from the schema of noun classification in different ways. In this section we will model each of these to see how building the asymmetry into each level changes the classification by the model.

#### Knowledge of Noun Internal Distributional Information

An asymmetry in the reliability with which semantic and phonological features of nouns are perceived and encoded during word learning could lead to a disparity in the way phonological and semantic features are represented as compared with how they are distributed in the input.

In our first manipulation (the Semantic Incompetence Hypothesis) we examined how classification by the model would be affected if the learner was misrepresenting some proportion of the semantic features that they should have encoded on nouns in their lexicon. We assume that learners represented the remaining proportion of nouns as predicted (accurately observing features during the experiment and assuming that both semantic and phonological features were relevant in classification). In doing this, we assume that learners' beliefs about which features are predictive of which class is built up as they observe different feature values on words belonging to different classes. One way of quantifying this is by modeling the learner's belief about the likelihood terms  $p(f|c)$  from Equation 1 under the assumption that these beliefs are derived from the counts that a learner accumulates of nouns in each class that contain a given feature. We assume learners use a multinomial model with a uniform Dirichlet prior distribution to estimate the proportion of items each class  $c$  that contain a particular value  $k$  for feature  $f$ . Under this assumption, each likelihood term is equal to:

$$p(f = k | c) = \frac{N_{c,f=k} + 1}{N_c + K} \quad (2)$$

where  $N_c$  denotes the number of nouns in the class,  $N_{c,f=k}$  denotes the number of nouns in the class for which the feature has value  $k$ , and  $K$  is the number of possible values for the feature.

We introduce misrepresentation of semantic features into this model by manipulating the number of observations of a noun with a certain feature value in each class. Since the semantic incompetence hypothesis posits that children misrepresent semantic feature values some proportion of the time, we reduce the count of nouns in each class that contain the relevant semantic features, changing them instead to the unspecified feature value [other]. We then compute the posterior probability of noun class membership using these adjusted feature counts. We can use this model to ask how low the counts would have to be in order for children's behavior to be optimal with respect to their beliefs.

We evaluated the model by comparing its behavior to children's behavior from the classification task. The model produced a close fit to the data in each condition (Figure 3). Furthermore, the estimated degree of misrepresentation was highly consistent across all semantic features and conflicting feature combinations. The best fitting level of uncertainty ranged from 0.96-0.91, meaning that children would be only using 4-9% of the semantic cues available to them. A generalized likelihood ratio test in which the level of misrepresentation was held constant across simulations (0.95) demonstrates that our semantic incompetence model significantly outperforms the optimal naïve Bayesian classifier ( $p < 0.0001$ ).

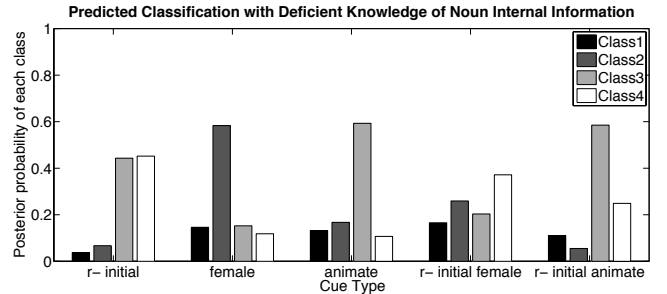


Figure 3: Classification of novel nouns as predicted by a Naïve Bayes Classifier with 95% of predictive semantic features misrepresented as [other].

Although this model produces a close fit to the empirical data, it predicts an extremely high degree of misperception. To understand why this is the case, consider that using likelihood terms for each class that are proportional to the

true empirical counts  $\frac{N_{c,f=k}}{N_c}$  would yield optimal noun

classification performance, regardless of the exact proportion of time children are misrepresenting features. That is, substituting  $\beta * p(f_i|c)$  for each term  $p(f_i|c)$  in Equation 1, where  $\beta$  is a constant denoting the degree of misperception, does not result in any change in the posterior probability distribution. This analysis suggests that changes in model predictions under this account of feature

misrepresentation occur primarily for low empirical feature counts, when the model relies heavily on pseudocounts from the Dirichlet prior distribution.

**Observation of semantic and phonological features on novel nouns** A second possibility is that children have little trouble perceiving, encoding and representing features on the words in their lexicon, but that the semantic features on the experimental items (as they are presented as flat pictures in a book) are unreliably perceived and encoded. We call this the Experimental Reject Hypothesis.

In this manipulation we investigate what would happen if a learner had a lexicon that faithfully represented the predictive features as they were distributed in the input and assumed both semantic and phonological features were relevant to classification, but didn't reliably encode semantic features on experimental items. To do this we use a mixture model, where some proportion of the time ( $1 - \beta$ ) an item that was supposed to have the specified semantic feature value [animate] or [female] (denoted as [spe]) it would be classified as with that value, the rest of the time ( $\beta$ ) it would be classified as if it had the unspecified value [other]. This yields the following model:

$$p(c | f_1, f_2) = (1 - \beta) \frac{p(f_1 = [spe] | c) p(f_2 | c) p(c)}{\sum_i p(f_1 = [spe] | c_i) p(f_2 | c_i) p(c_i)} + \beta \frac{p(f_1 = [other] | c) p(f_2 | c) p(c)}{\sum_i p(f_1 = [other] | c_i) p(f_2 | c_i) p(c_i)} \quad (3)$$

As with the semantic incompetence model, we found the best-fitting value of  $\beta$  and evaluated the model by comparing it to children's behavior. This model again produced a close fit for all feature values (Figure 4). The model showed a consistent degree of misperception across all semantic features and feature combinations. The best fitting level value of  $\beta$  ranged from .49 to .83, where 58% was the best fit overall. This means that children would be misperceiving semantic features on 58% of the experimental items. A generalized likelihood ratio test indicates that the experimental reject model also significantly outperforms the optimal naïve Bayesian classifier ( $p < 0.05$ ).

**Assumption that all features are relevant for classification** The asymmetry between the reliability of perceiving and encoding phonological as compared to semantic features could also engender a bias to prefer phonological information for classification decisions, as phonological information has been reliably available for a longer period of time. Our third model, embodying the Phonological Preference Hypothesis, therefore looked at what would happen if we had a learner that was biased not to use semantic features in classification some proportion of the time, even if these features were represented just as distributed in the input and accurately perceived during the experimental task. We used a second mixture model, this time looking at the mixture of a Bayesian classifier that used both semantic and phonological features, and one that only used phonological features.

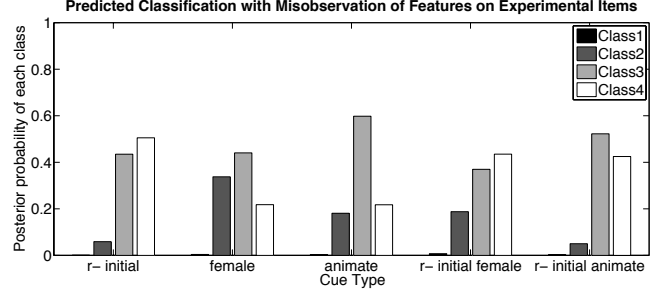


Figure 4: Classification of novel nouns as predicted by a model that misobserves semantic features on experimental items 58% of the time

The crucial difference between this model and the experimental reject model is that in the experimental reject model semantic features are always used, but are encoded as the wrong value (the unspecified [other] value) some proportion of the time, whereas in the phonological preference model, semantic features do not factor into the calculation at all some proportion of the time ( $\beta$ ). The model can be seen in Equation 4.

$$p(c | f_1, f_2) = (1 - \beta) \frac{p(f_1 = [sem] | c) p(f_2 | c) p(c)}{\sum_i p(f_1 = [sem] | c_i) p(f_2 | c_i) p(c_i)} + \beta \frac{p(f_2 | c) p(c)}{\sum_i p(f_2 | c_i) p(c_i)} \quad (4)$$

Again we evaluated the model against the children's classification data and found a close fit (Figure 5). The best fitting value of  $\beta$  ranged from .49 to .83, and was .65 over all, meaning that children would be choosing not to use semantic features on 65% of classification decisions. A generalized log likelihood test showed that this model also significantly outperformed the optimal naïve Bayesian classifier ( $p < 0.0001$ ).

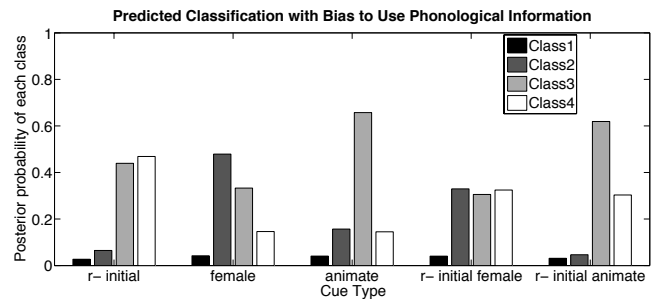


Figure 5: Classification as predicted by a model biased not to use semantic information 65% of the time

## Discussion

Tsez noun classes are characterized by both semantic and phonological features. Children have been shown to be able to use these features when classifying novel nouns. Here we showed that their classification patterns differ from those of an optimal Bayesian classifier when nouns have semantic and phonological features that make conflicting predictions.

We also presented three models that take into account ways in which the difference between semantic and phonological features could lead to children's apparent preference to use the less reliable phonological features. These models examined how classification would look if a learner had (a) misrepresented semantic features in the lexicon, (b) misencoded semantic features during the classification experiment, or (c) developed a bias to use phonological information in noun classification due to its higher reliability in the early stages of lexical acquisition. All three models fit children's data significantly better than the optimal naïve Bayesian classifier did. This suggests that although originally children did not look as though they were behaving optimally with respect to the input, they may well be behaving optimally with respect to their intake, that is, the input as they have represented it.

It is not obvious how one would best to evaluate the alternative models with respect to one another. For example, each model yielded a different best-fit parameter, corresponding to a different degree of misrepresentation or bias. While these best fitting parameters may differ in terms of their 'reasonableness' (i.e. misrepresenting 95% of semantic features in the lexicon at age 6 seems quite high), it isn't immediately clear how to measure reasonableness, or how to compare it across models. Furthermore, it is likely that a combination of all three of these processes (and perhaps more that we haven't considered here) is influencing children's classification decisions. This could potentially be explored through a combined model; however, as all of these models fit the data so closely, it would be difficult to determine which and to what extent each type of misrepresentation or bias is involved.

This work has several important implications for research statistical learning and language acquisition. First, and most broadly, by combining experimental data from children acquiring an understudied language with computational modeling techniques, we found a better understanding of both children's acquisition of Tsez, and the role of statistical cues in language acquisition. Tsez was an ideal language to look at, as feature types differed in their reliability as cues to noun class. However, we expect that these results will be generalizable across languages, as the relative difficulty of acquiring semantic, as compared to phonological, features of words will be consistent cross linguistically.

Second, we identified an area where children's behavior does not appear to reflect the ideal inferences licensed by the statistical patterns in the input. Three models allowed us to investigate the source of this asymmetry. While each model differed in where the asymmetry came from, all employed a weakening of the statistical import of semantic features. This is a distinct pattern from the finding that children learning an artificial language amplify an already strong statistical tendency (Hudson-Kam & Newport, 2009). Further research will determine whether or not these patterns could be in some way related.

Next, we showed that it is possible for a learner to be suboptimal and Bayesian at the same time. That is, we

demonstrated that while children's behavior does not align with the predictions made by the optimal Bayesian classifier, it can be predicted by modifying the terms of this classifier in reasonable ways. Thus we were able to model children's suboptimal behavior using a Bayesian model, rather than adopting some other system of computation.

Finally, our models showed that it is plausible that these children are indeed behaving optimally with respect to some statistical distribution, just not one directly measureable from the input. This point is crucial as researchers extend accounts of statistical learning to a greater range of problems, highlighting the fact that the critical question isn't whether or not children are using statistics to acquire language, but what statistics they are using.

**Acknowledgments** This research was supported by NSF IGERT 0801465 and a NSF GRF to Gagliardi. We would like to thank Masha Polinsky, the UMD Cognitive Neuroscience of Language Lab, the UMD Project on Children's Language Learning and the UMD Computational Psycholinguistics group for helpful discussion and assistance.

## References

- Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.
- Gagliardi, A., & Lidz, J. (Under review) Separating input from intake: Acquiring noun classes in Tsez.
- Gómez, R.L., & Maye, J. (2005). The Developmental Trajectory of Nonadjacent Dependency Learning. *Infancy*, 7, 183–206.
- Hudson Kam, C.L., & Newport, E.L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, 59, 30–66.
- Mintz, T.H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90, 91–117.
- Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82, B101–B111.
- Plaster, K., Polinsky, M., & Harizanov, B. (In Press). Noun Classes Grow on Trees: Noun Classification in the North-East Caucasus. *Language and representations* (tentative). John Benjamins
- Reeder, P.A., Newport, E.L., & Aslin, R.N. (2009). The role of distributional information in linguistic category formation. In N. Taatgen and H. van Rijn (eds), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Reeder, P.A., Newport, E.L., & Aslin, R.N. (2010). Novel words in novel contexts: The role of distributional information in form-class category learning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35, 606–621.
- Saffran, J.R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, 44, 493–515.
- Takahashi, E. (2009). *Beyond statistical learning in the acquisition of phrase structure*. College Park, MD: University of Maryland dissertation.



# Verbal Satiation of Chinese Bisyllabic Words: A Semantic Locus and its Time Course

Bruno Galmar (hsuyueshan@gmail.com)

Institute of Education, National Cheng Kung University, Tainan, TAIWAN

Jenn-Yeu Chen (psyjyc@mail.ncku.edu.tw)

Institute of Cognitive Science, National Cheng Kung University, Tainan, TAIWAN

## Abstract

Verbal satiation of Chinese bisyllabic words was studied in three experiments to ascertain the phenomenon, to track its time course, and to identify its locus. Experiment 1 asked the participants to judge if an exemplar matched a category in 22 blocks of 40 trials each. Within a block, one category appeared 20 times (*repeated* trials) while each of the remaining 10 categories appeared only twice (*baseline* trials). For the first 11 trials, response times (RTs) for the repeated ones were similar to RTs for the baseline ones. For the subsequent trials, *repeated* RTs were slower (by 9 ms) than *baseline* RTs, indicating a satiation effect. Its loci could be orthographic, semantic, or both, or on the associative links between form and meaning. In Experiment 2, category names were not shown. Participants judged if two exemplars belonged to the same category. *Repeated* RTs were faster (by 6 ms) than *baseline* RTs for the first 12 trials. Then, verbal satiation emerged but was short-lived (between the 13<sup>th</sup> and the 17<sup>th</sup> trial) and was of greater magnitude (20 ms) than that observed in Experiment 1. The satiation effect must be semantic, as only meanings were repeated. Experiment 3 asked participants to judge if two category names were identical, mostly an orthographic task. *Repeated* RTs were similar to *baseline* ones across all trials, suggesting no orthographic satiation. The results indicate that semantic satiation of Chinese words can be directly semantic (categorical). Its time course conforms to the habituation model described in Rankin, et al. (2009), i.e., sensitization (semantic priming) before habituation (semantic satiation) and habituation followed by dishabituation (recovery).

**Keywords:** Verbal Satiation, Semantic Satiation, Repetitive Semantic Processing.

## Introduction

For more than a century, the self-reports collected by psychologists (Moulin & Connor, 2006; Severance & Washburn, 1907) mention a loss of the meaning of an alphabetic word following its prolonged viewing (e.g., 3 min.) or its active repetition (oral or written, e.g., for 30 times). This experienced loss of meaning has been coined *semantic satiation*. The term *semantic satiation* emphasizes that the locus of satiation is thought to be semantic. In the present work, we will also use the term *verbal satiation* which is neutral regarding the locus of satiation. For the non-alphabetic Chinese script, according to the self-reports collected by Cheng and Wu (1994), prolonged viewing of a multi-component Chinese character (e.g. 臉, which means a face, can be decomposed at a first level into 月 + 僉) elicits an *orthographic satiation*. The original binding of the

different components is disrupted and the character looks weird. Hence, the primary subjective experience is an orthographic decomposition of the character, not the loss of the character's meaning. Cheng and Lan (2009) advanced that for the Chinese script there is no genuine semantic satiation but mainly orthographic satiation. To our knowledge, no one has tried to demonstrate whether for the Chinese script a verbal satiation with a unique semantic locus exists. If it does, semantic satiation would be genuine and would not be a mere by-product of orthographic satiation or of any pre-semantic satiation. Neither were we aware of any studies that examined Chinese verbal satiation at the word level rather than the character level.

In a most recent study on verbal satiation in English, Tian and Huber (2010) designed a continuous speeded category-matching task to track the time course of verbal satiation, and to identify its locus. In their Experiment 1, subjects performed for each trial a membership task for a category name (the cue) and an exemplar (the target). Half of the 20 trials of an experimental block contained a repeated category name (e.g. VEGETABLE). In half of these trials, the repeated category name was paired with exemplars of the category, producing matching trials (VEGETABLE-CARROT). In the other half of these trials, it was paired with exemplars of non-repeated categories, producing mismatching trials (VEGETABLE-GOLF). These 10 trials constituted the *satiation trials*. The other 10 trials using 10 distinct category names constituted the non-repeated trials, or the *baseline trials*. The satiation trials and the baseline trials were mixed and ordered randomly within a block. In their Experiment 1, Tian and Huber aimed at monitoring continuously the effects of repetition of a category name. They predicted that only response times for satiation trials should slow down as the task progressed if satiation occurred and if the task did not cause fatigue. The locus of any detected satiation in this experiment could be: (a) orthographic if the word's orthographic form representing the category name is satiated, (b) semantic if the meaning of the word is satiated, (c) both orthographic and semantic, or (d) on the associative link between the orthographic and semantic units. To distinguish between these four possible loci, Tian and Huber designed two more experiments based on the same speeded category-matching task. In their Experiment 2, Tian and Huber aimed at isolating a sole semantic locus of verbal satiation. They tested whether verbal satiation of the meaning of a category name (e.g.

VEGETABLE) can be induced without repeating the category name but instead by repeating many of its exemplars (e.g., CARROT, LETTUCE, ...). In this experiment, Tian and Huber used exemplars instead of category names as the repeated words. The task for their subjects was to judge whether two exemplars are members of the same category (e.g. LION-TIGER, CAR-ROSE). Within a block, exemplars of the category to be satiated made half of the stimulus words. Trials with exemplars from 10 other categories served in the baseline trials. The prevalence of the exemplars from the non-presented category name to be satiated should elicit satiation of its meaning. In Experiment 3, Tian and Huber aimed at isolating an orthographic locus of verbal satiation. Subjects had to perform lexical decisions about whether two words were the same word (e.g., VEGETABLE-VEGETABLE, ANIMAL-VEGETABLE). Within a block, the word to be satiated was used repeatedly in half of the trials. Repetition of the same word in this task was assumed to induce a verbal satiation with a lexical locus. In their experiments 2 and 3, Tian and Huber did not observe a satiation effect, but in Experiment 1 they did. Hence, Tian and Huber concluded that verbal satiation occurs in a task only when two conditions are met: (a) a word form is repeated, (b) the meaning of the repeated word is repetitively accessed. Tian and Huber posited that repeatedly viewing a word while thinking of its meaning elicits *associative satiation*: the information-flow channel from the lexical units to the semantic units is satiated which caused the subjective experience of the loss of meaning of a satiated word.

The lack of a satiation effect in Experiments 2 and 3 of Tian and Huber's (2010) study is critical to their interpretation of the satiation effect observed in Experiment 1. Therefore, the findings need to be evaluated with rigor. On the one hand, the satiation effect of Experiment 1 could be due to the concomitant loci of satiation at the levels of form, meaning, and form-to-meaning link. On the other hand, the failure to observe a satiation effect in Experiment 2 could be the result of insufficient number of repetitions (10). In previous studies, the smallest number of repetitions for which semantic satiation effects were observed was 15 (Kounios, Kotz, & Holcomb, 2000).

Our main goal was to observe whether a verbal satiation at the meaning level could occur in Chinese, and if it did, to track its time course. To pursue this goal, we adapted Tian and Huber's (2010) three experiments by doubling the number of repeated trials within a block and by using traditional Chinese multisyllabic words as visual stimuli.

## General Method

### Participants

Participants were Taiwanese students from National Cheng Kung University. The participants were different for each experiment.

### Apparatus

The experiments were programmed with DMDX (Forster & Forster, 2003).

## Materials and Design

The materials consisted of eleven Chinese category names: 蔬菜 (VEGETABLE), 動物 (ANIMAL), 水果 (FRUIT), 疾病 (DISEASE), 親戚 (RELATIVE), 運動 (SPORT), 職業 (OCCUPATION), 國家 (COUNTRY), 城市 (CITY), 公司 (COMPANY) and 樂器 (MUSICAL INSTRUMENT). Twenty exemplars were selected for each category from a Chinese corpus and word lists. Care was taken to ensure the exemplars shared no characters with the category names.

One experimental block contained 40 trials of pair of words as shown in Table 1. We detail the pairing in Experiment 1 to exemplify its general principles. Adaptations for Experiments 2 and 3 are given in the respective Experiment section. Within a block, all the 11 categories were represented but only one category (VEGETABLES in Table 1) was to be satiated through long-term repetition. This latter category is termed the *dominant* category of the block and the 10 others the *non-dominant* categories. The 11 Chinese category names listed before and their respective 20 exemplars were used respectively as cues and targets. Within a block, the 20 *repeated trials* contained as a cue the dominant category name. Ten exemplars from the dominant category served as targets to produce the *repeated match trials*. Ten exemplars from the non-dominant categories served as targets to produce the *repeated mismatch trials*. The 10 non-dominant category names served twice as a cue to produce the 20 *baseline trials*. They were randomly assigned to one or two of the (*baseline match*, *baseline mismatch*) groups to prevent any informed guessing about the pairing of a second occurrence of a non-dominant category name from viewing its first occurrence in a previous trial. Ten exemplars, distinct from the ones in the *repeated mismatch trials*, from the non-dominant categories served as targets and are paired to their respective category label to produce the *baseline match trials*. The 10 remaining exemplars from the dominant category not used in the *repeated match trials* served as targets to produce the *baseline mismatch trials*. Hence in a block, all the 20 dominant exemplars occurred once. Each of the 220 exemplars was used only twice in a set. Each of the 11 category names served as the word to be satiated in one block. The eleven blocks made a set. For the experiment, participants ran 2 sets totaling 22 blocks of 40 trials.

### Procedure

We describe the procedure of Experiment 1 to exemplify the general common procedure to all experiments. Adaptations for Experiments 2 and 3 are given in the respective Experiment section. Subjects faced a screen with a black background. For the first trial of a block, a fixation-cross appeared in white at the center of the screen for 200 ms. Then, a category name was presented in white at the center of the screen. After 1000 ms, an exemplar in white was presented below the cue. Within 2000 ms, subjects had to decide whether the exemplar matched the above category name. The two words remained on the screen until subjects pressed a Right or Wrong button on a gamepad. After pressing one of the buttons, the screen was cleared, and after

300 ms the next trial began with the presentation of the next cue word.

### Data Analysis

Both reaction time and accuracy for each trial were recorded. The first trial of each block was suppressed from the analysis because it tended to deviate from the trend of the rest of the trials which are the primary focus of the study. We analyzed both the RTs and accuracy data as a function of the trial number ranging from 2 to 40. This approach allows the most straightforward visualization of the dynamics of verbal satiation within a block as the task progresses and also equates the baseline/repetition conditions in term of general fatigue along a block. In a first statistical analysis, the range of the 39 trial numbers was a priori and arbitrarily segmented and mapped onto four equally spaced intervals. Mean RTs were calculated separately for Yes and No responses for each subject and separately for the repeated and baseline conditions for each of the four intervals over the 22 blocks. The mean RTs were subject to the analysis of variance. We run another similar statistical analysis with both the number of intervals and their unequal length informed by the data pattern over the 39 trial numbers. As the conclusions derived from the two statistical analyses were congruent, we reported only the results of the latter which offered a finer tracking of the dynamics of verbal satiation. Because of space limitation, only data for Yes responses are presented. Data for No responses were congruent with the ones from Yes responses.

### Experiment 1

In Experiment 1, we aimed at eliciting verbal satiation through 20 repetitive access to a Chinese category name's word form and meaning.

According to three accounts of satiation (associative satiation, semantic satiation, and orthographic satiation), Experiment 1 should elicit verbal satiation. Repetitive access from the same word form to the same meaning unit should satiate the associative link between orthographic and semantic units. Accessing repetitively the meaning of the category name should satiate it. Repetitive access to the same word form should elicit orthographic satiation.

### Method

Forty students (female 33) participated in the experiment. The subjects ran a category-matching task as described in the General Method section.

### Results and Discussion

One participant out of 40 was excluded based on a 90% accuracy criterion. Figure 1 plots the time course of the RTs for the correct Yes responses as a function of the trial number. Both baseline and repeated trials showed a slowing trend which revealed a general fatigue effect.

In Fig. 1, we drew a vertical line between trial number 11 and 12, a cutoff from which repeated RTs appeared to become globally slower than baseline RTs until the end of the block. From this cutoff, we defined two intervals:

position 1 for the trials 2-11 and position 2 for trials 12-40. A repeated measures two-way ANOVA was applied with repetition (2: baseline and repeated) and position (2: position 1 and position 2) as the two within-subject factors. We found both a statistically significant main effect of repetition,  $F(1, 38) = 4.13, p = .049$ ,  $GES = .0016$  ( $GES$  being the generalized eta squared, Bakeman, 2005), a statistically significant main effect of position,  $F(1, 38) = 33.65, p < .001$ ,  $GES = .031$ , and a statistically significant repetition  $\times$  position interaction,  $F(1, 38) = 9.67, p = .0035$ ,  $GES = .003$ . The interaction is characterized by the RTs for the repeated trials being comparable with the RTs for the baseline trials at position 1 but becoming slower at position 2. A paired t-test comparing RTs for repeated and baseline trials at Position 1 revealed no significant difference,  $t(38) = -0.57, p = .58$ . Another paired t-test at Position 2 showed that RTs for repeated trials were significantly slower (by about 9 ms) than RTs for baseline trials,  $t(38) = 4.7, p < .001$ .

Table 1. Block structure in Experiments 1, 2 and, 3.

Trial Number	Repetition Status	Match Status	Experiment 1	Experiment 2	Experiment 3
1	Repeated	Match	蔬菜 / vegetables 玉米 / corn	馬鈴薯 / potato 蘆筍 / asparagus	蔬菜 / vegetables 蔬菜 / vegetables
2	Baseline	Mismatch	疾病 / disease 南瓜 / pumpkin	台北 / Taipei 香菇 / mushroom	運動 / sport 蔬菜 / vegetables
3	Repeated	Match	蔬菜 / vegetables 茄子 / aubergine	茄子 / aubergine 洋葱 / onion	蔬菜 / vegetables 蔬菜 / vegetables
4	Repeated	Match	蔬菜 / vegetables 牛蒡 / burdock	南瓜 / pumpkin 紅蘿蔔 / carrot	蔬菜 / vegetables 蔬菜 / vegetables
5	Baseline	Match	職業 / profession 秘書 / secretary	法國 / France 巴西 / Brazil	親戚 / relative 親戚 / relative
...	...	...	...	...	...
38	Baseline	Mismatch	城市 / city 香菇 / mushroom	足球 / football 芋頭 / taro	城市 / city 蔬菜 / vegetables
39	Baseline	Match	運動 / sport 網球 / tennis	姑姑 / aunt 姪女 / niece	動物 / animal 動物 / animal
40	Repeated	Mismatch	蔬菜 / vegetables 吉他 / guitar	黃瓜 / cucumber 網球 / tennis	蔬菜 / vegetables 國家 / country

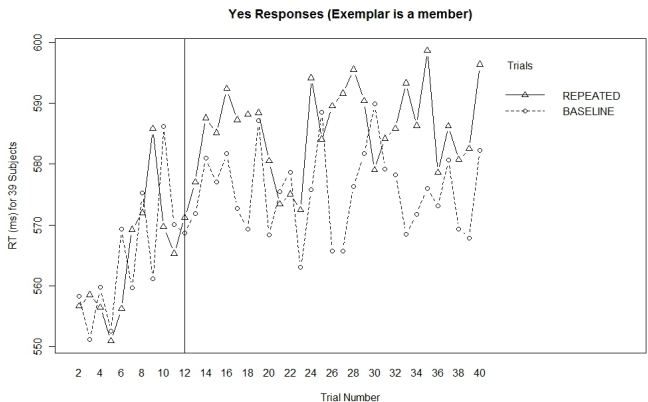


Figure 1. Experiment 1. Reaction times as a function of trial number for Yes responses.

The accuracy data were analyzed in the same way. We found a statistically significant main effect of repetition,  $F(1, 38) = 12.55$ ,  $p = .001$ ,  $GES = .049$ , revealing better accuracy for repeated trials. The main effect of position was not statistically significant,  $F(1, 38) = 0.44$ ,  $p = .51$ ,  $GES = .003$ . There was no repetition  $\times$  position interaction,  $F(1, 38) = 0.034$ ,  $p = .85$ ,  $GES < .001$ , suggesting that the interaction effect in the RT data was not the result of a speed-accuracy tradeoff.

Experiment 1 asked the subjects to determine if an exemplar word matched a category word. In terms of YES responses, response times for the first 11 repeated trials were similar to RTs for the non-repeated baseline trials. After that, RTs for the repeated trials became slower than RTs for the baseline trials. The increased RT difference between the repeated and the baseline trials from position 1 to position 2 was about 9 ms. Not due to a speed-accuracy tradeoff, this difference represented a verbal satiation effect brought about by the repeated processing of the same category word.

## Experiment 2

Experiment 2 investigated whether pure semantic satiation was possible by asking subjects to repeatedly process a category but not its name. According to the three accounts of satiation, Experiment 2 should elicit neither associative satiation nor orthographic satiation because there is no repetition of the category name. However, repetitive access to the meaning of an unshown category name should elicit satiation with a semantic locus.

### Method

Forty-one students (female 26) participated in the experiment. The materials for Experiment 2 were the 220 exemplars from the 11 categories used in Experiment 1. Category labels were not shown. Within a block, the 20 exemplars of the dominant category occurred twice, once as a cue in the *repeated trials* and once as a target. Exemplars from the 10 other categories served as cues in the baseline trials. Each of the 11 category names served as the unshown word to be satiated in one block. For the procedure, the cue word was no more a category name but a changing exemplar. Hence, subjects had to decide whether two exemplars belonged to the same category.

### Results and Discussion

One participant out of 41 was excluded based on a 90% accuracy criterion. Figure 2 plots the time course of the RTs for the correct Yes responses. As for Experiment 1, both baseline and repeated trials showed a slowing trend which revealed a general fatigue effect. In Fig. 2, we drew two vertical lines at trial numbers 13 and 17, two cutoff trial numbers separating the plot in three intervals: position 1 for the trials 2-12, position 2 for trials 13-17, and position 3 for trials 18-40. During position 1, repeated RTs appeared globally faster than baseline RTs which could signal a facilitatory semantic priming effect. Position 2 represents a potential verbal satiation interval during which repeated RTs became slower than baseline RTs. During position 3, the

repeated RTs recovered to be as fast as baseline RTs. This qualitative interpretation of Fig. 2 was confirmed by the thereafter statistical analysis. A repeated measures two-way ANOVA on mean correct RTs was applied with repetition (2) and position (3) as the two within-subject factors. We reported the Huynh-Feldt corrections for all statistical effects involving more than one degree of freedom in the numerator. The main effect of repetition was not statistically significant,  $F(1, 39) = 3.3$ ,  $p = .077$ ,  $GES = .0016$ .

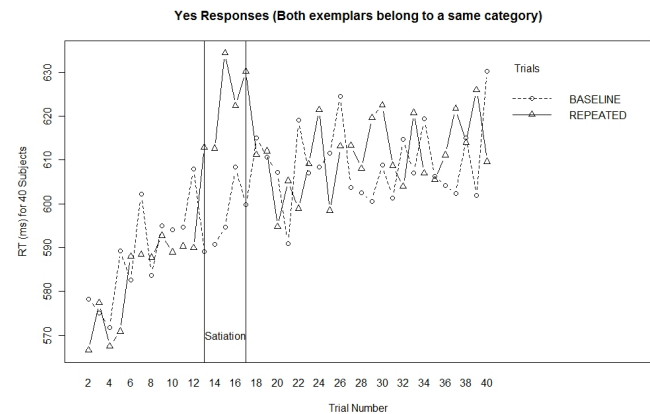


Figure 2. Experiment 2. Reaction times as a function of trial number for Yes responses.

However, we found a statistically significant main effect of position,  $F(2, 78) = 31.5$ ,  $p < .001$ ,  $GES = .026$ , and a statistically significant repetition  $\times$  position interaction,  $F(2, 78) = 11.09$ ,  $p < .001$ ,  $GES = .007$ . The interaction is characterized by the RTs for the repeated trials compared to the RTs for the baseline trials being faster at position 1, slower at position 2 and equally fast at position 3. A paired t-test comparing RTs for repeated and baseline trials at position 1 revealed marginally statistically significant faster reaction times (by 6 ms) for repeated trials,  $t(39) = -1.88$ ,  $p = .068$ , 95% CI [-12.8, 0.48]. Repeated trials are made of exemplars of a same dominant category, therefore the observed facilitation at position 1 for repeated trials revealed the occurrence of semantic priming. Another paired t-test at Position 2 showed that RTs for repeated trials became statistically significantly slower (by 20 ms) than baseline trials,  $t(39) = 4.7$ ,  $p = .003$ , 95% CI [7.5, 33.6]. Such a reversal from semantic priming at position 1 to semantic impairment at position 2 for repeated trials supported the occurrence of verbal satiation at position 2. Another paired t-test at position 3 showed that repeated RTs were no longer slower than baseline RTs,  $t(39) = 0.55$ ,  $p = .58$ , 95% CI [-4.1, 7.1]. Hence, at position 3, verbal satiation vanished.

Analysis of the accuracy data showed a statistically significant main effect of repetition:  $F(1, 39) = 6.27$ ,  $p = .017$ ,  $GES = .018$ , a statistically significant main effect of position:  $F(2, 78) = 3.3$ ,  $p = .043$ ,  $GES = .02$ , and a marginally statistically significant repetition  $\times$  position interaction:  $F(2, 78) = 3.11$ ,  $p = .058$ ,  $GES = .018$ . Of main interest to us, accuracy rates for repeated trials were

statistically significantly higher only for position 1,  $t(39) = 4.05$ ,  $p < .001$ . Then, they decreased to the level of baseline trials for position 2 and 3, ( $t(39) = 0.38$ ,  $p = .7$ ;  $t(39) = 0.54$ ,  $p = .59$ ). Hence, at position 2 RTs for repeated trials increased as shown previously whereas their accuracy rates decreased, indicating no speed-accuracy tradeoff. The RT and the accuracy data constituted convergent evidence towards the occurrence of verbal satiation at Position 2.

In Experiment 2, the main result was that repeated processing of exemplars of a same category increased the RTs differences between repeated trials and baseline trials. We posited that these RTs differences signaled the occurrence of verbal satiation. Verbal satiation occurred roughly at the same trial number than in Experiment 1. Unlike the verbal satiation in Experiment 1, verbal satiation in Experiment 2 was short-lived (limited to trials 13-17) and of higher magnitude (20 ms versus 9 ms). The verbal satiation in Experiment 2 must be purely semantic because the word form of the dominant category name was never shown and the cues in repeated trials had all a different word form. Hence, with Experiment 2, we identified a semantic locus to verbal satiation which could have contributed to the verbal satiation found in Experiment 1.

### Experiment 3

Experiment 3 investigated whether pure orthographic satiation was possible by asking subjects to repeatedly recognize the word form of a category name. According to the three accounts of satiation, Experiment 3 should elicit neither associative satiation nor semantic satiation because there is no need to access repetitively the meaning of the category name. However, repetitive processing of a same word form name should elicit its orthographic satiation as demonstrated in (Cheng & Lan, 2009).

#### Method

Forty-one students (female 20) participated in the experiment. The materials for Experiment 3 were the 11 category names used in Experiment 1. Table 1 illustrates the pairing of trials. The exemplars in Experiment 1 were all replaced by category labels. Hence, within a block, the dominant category label was repeated 20 times as a cue and 20 times as a target. Non-dominant category names served twice as a target. The task became lexical rather than semantic. Subjects had to decide whether the cue and the target were a same word.

#### Results and Discussion

Two participants out of 41 were excluded based on a 90% accuracy criterion. Figure 3 plots the time course of the RTs for the correct Yes responses. Unexpectedly, not only no occurrence of verbal satiation could be observed around the 12-13 trials as in the two previous experiments but it also appeared that repeated RTs were globally equally fast to baseline RTs. We partitioned the 40 trials into three intervals: trials 2-30, trials 31-35 and trials 36-40. The second interval could be the seat of a weak and delayed verbal satiation. We proceeded to a statistical analysis to evaluate this possibility. For Yes responses, a repeated

measures two-way ANOVA on mean correct RT was applied with repetition (2) and position (3) as the two within-subject factors. None of the main effect of repetition, the main effect of position, and the repetition  $\times$  position interaction, ( $F(1, 38) = 0.037$ ,  $p = .85$ ,  $GES < .001$ ;  $F(2, 76) = 0.7$ ,  $p = .47$ ,  $GES = .001$ ,  $F(2, 76) = 2.08$ ,  $p = .14$ ,  $GES = .001$ ) were statistically significant.

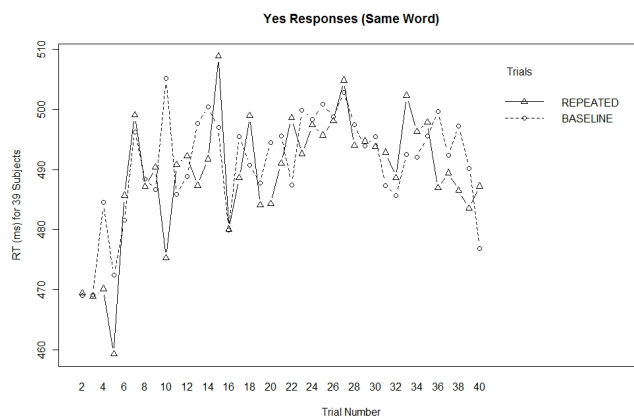


Figure 4. Experiment 3. Reaction times as a function of trial number for Yes responses.

The latter result invalidated our hypothesis of occurrence of a verbal satiation elicited by repetitive processing of a repeated word form. None of the paired t-test comparing RTs for repeated and baseline trials for each of the three positions reached statistical significance. Hence in experiment 3, speed of processing for both conditions was equal.

### General Discussion

The present study explored the existence of verbal satiation in Chinese and questioned whether its nature would be semantic. In both Experiments 1 and 2, we found within 40 trials an interval for which RTs for repeated trials became slower than baseline RTs after being equally fast over the first 11 trials in Experiment 1 and after being slightly faster over the first 12 trials in Experiment 2. In Experiment 3, RTs were similar for repeated trials and baseline trials. For Experiments 1 and 2, we ruled out a speed-accuracy trade-off strategy. Instead, we posited that verbal satiation occurred in both experiments and we advanced that its nature is semantic as it can be induced by mere repetitive processing of exemplars of a category without repetition of the word form of the category label. Before defending our viewpoint, we considered a number of alternative explanations.

#### Did subjects strategically generate exemplars impairing semantic retrieval?

If subjects on seeing a cue category name or an exemplar retrieve strategically from semantic memory a few exemplars for further responding, these exemplars could block competing exemplars impeding semantic processing if the target words differ from the expected exemplars.

Repeated RTs for Yes responses will become increasingly slower than baseline RTs because of the repetitive impediment of blocking on the exemplars of a same category in repeated trials. This specific slowing down of repeated trials should happen from the very beginning within a block. However, our results in Experiments 1 and 2 refuted this latter prediction.

### Does inhibition underlie verbal satiation?

Semantic inhibition of return (IOR) (Weger & Inhoff, 2006) refers to an attentional bias towards *semantic novelty* slowing down processing of repeated or semantically related words compared to unrelated and nonrepeated words. Weger and Inhoff (2006) found that for semantic IOR tended to not occur when repeated/related words with large item variability are mixed with a rather heterogeneous pool of nonrepeated/unrelated words. Hence, both of the designs in Experiment 1 and 2 are not propitious to semantic IOR.

### A semantic habituation model of verbal satiation

In Rankin, et al. (2009), ten characteristics of behavioral habituation are listed. We described thereafter the first five characteristics and showed that data in Experiment 2 in the light of data in Experiment 1 matched realistically a semantic habituation model of verbal satiation. The first three characteristics in Rankin, et al. (2009) shaped the usual time course of habituation: sensitization before habituation and dishabituation following habituation. In Experiment 2, the three intervals of the 40 trials in Figure 2 delineated respectively first the sensitization phase with facilitatory semantic priming for repeated trials, then a habituation phase corresponding to the semantic satiation of the unshown dominant category name, and finally a dishabituation phase for which repeated RTs recovered from habituation to return to the level of baseline RTs. The fourth characteristic of habituation (Rankin, et al., 2009) can be stated as: more frequent stimulations can result in more pronounced response decrement (slowing reaction times in our case). We considered that if we count the number of distinct stimuli to the meaning of the dominant category name, it was greater in Experiment 2 than in Experiment 1, simply because in Experiment 2 many exemplars replaced the dominant category name. We thought that the higher number of external stimuli (exemplars) in Experiment 2 could be translated as a higher frequency of external stimulation to the meaning of the category name. Following the fourth characteristic of habituation, a higher magnitude of verbal satiation could be expected in Experiment 2 than in Experiment 1. The fifth characteristic of habituation can be stated as: the weaker the stimulus, the more pronounced is habituation. We considered that in Experiment 2, the exemplars replacing the dominant category name as a cue constituted weaker semantic stimuli to the meaning of the category name than the category name itself. Hence, using weaker stimuli in Experiment 2 than in Experiment 1, we could expect, according to the fifth characteristic, again a higher magnitude of verbal satiation in Experiment 2. The

dynamics of our enduring verbal satiation in Experiment 1 is akin to the long-term semantic satiation obtained in a semantic generation task by Kuhl and Anderson (2011).

## Conclusion

The semantic satiation account of verbal satiation (Smith & Klein, 1990) which stipulates that semantic units became habituated with repetitive access to the meaning of word was newly validated for Chinese multisyllabic words. The time course of semantic satiation follows the classic habituation model (Rankin, et al., 2009): sensitization (semantic priming) before habituation (semantic satiation) and habituation followed by dishabituation (recovery).

Semantic satiation exemplifies that conceptual processing can be habituated as early suggested by Baars (1987). Hence, the cognitive system would respond to both perceptual information redundancy (see the example of stabilized retinal images (Pritchard, Heron & Hebb, 1960)) and meaning redundancy with the same habituation mechanism.

## Acknowledgments

This work was supported by the NSC97-2410-H-006-095-MY3 grant awarded to Jenn-Yeu Chen as well as a doctoral fellowship grant (NSC100-2420-H-006-007-DR) awarded to Bruno Galmar.

## References

- Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, 37(3), 379-384.
- Baars, B. J. (1988) *A cognitive theory of consciousness*. Cambridge University Press.
- Cheng, C. M., & Lan, Y. H. (2009). An implicit test of Chinese orthographic satiation. *Reading and Writing*, 1-36.
- Cheng, C. M., & Wu, S. J. (1994). Orthographic satiation and disorganization in Chinese. *Advances in the study of Chinese language processing*, 1, 1-30.
- Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior*
- Kounios, J., Kotz, S. A., & Holcomb, P. J. (2000). On the locus of the semantic satiation effect: Evidence from event-related brain potentials. *Memory and Cognition*, 28(8), 1366-1377.
- Kuhl, B., & Anderson, M. (2011). More is not always better: paradoxical effects of repetition on semantic accessibility. *Psychonomic Bulletin & Review*, 1-9. doi: 10.3758/s13423-011-0110-0
- Moulin, C. J., & Connor, A. R. (2006). *Semantic satiation and subjective experience: The strange case of jamais vu*. Paper presented at the International Conference of Memory, Sydney.
- Pritchard R.M., Heron W., & Hebb D.O. (1960). Visual Perception Approached by the Method of Stabilized Images. *Canadian J. Psych.*, 14, 67-77.
- Rankin, C. H., Abrams, T., Barry, R. J., Bhatnagar, S., Clayton, D. F., Colombo, J., et al. (2009). Habituation revisited: an updated and revised description of the behavioral characteristics of habituation. *Neurobiology of learning and memory*, 92(2), 135-138.
- Severance, E., & Washburn, M. F. (1907). The loss of associative power in words after long fixation. *The American Journal of Psychology*, 18(2), 182-186.
- Smith, L., & Klein, R. (1990). Evidence for semantic satiation: Repeating a category slows subsequent semantic processing. *Learning, Memory*, 16(5), 852-861.
- Tian, X., & Huber, D. E. (2010). Testing an associative account of semantic satiation. *Cognitive Psychology*, 60, 267-290.
- Weger, U. W., & Inhoff, A. W. (2006). Semantic inhibition of return is the exception rather than the rule. *Attention, Perception, & Psychophysics*, 68(2), 244-253.



# Online learning of causal structure in a dynamic game situation

Yue Gao (ygao@cs.cornell.edu)

Department of Computer Science, Cornell University  
Ithaca, NY, 14853 USA

Eyal Nitzany (ein3@cornell.edu)

Program in Computational Biology and Medicine, Cornell University  
Ithaca, NY, 14853 USA

Shimon Edelman (edelman@cornell.edu)

Department of Psychology, Cornell University  
Ithaca, NY, 14853 USA

## Abstract

Agents situated in a dynamic environment with an initially unknown causal structure, which, moreover, links certain behavioral choices to rewards, must be able to learn such structure incrementally on the fly. We report an experimental study that characterizes human learning in a controlled dynamic game environment, and describe a computational model that is capable of similar learning. The model learns by building up a representation of the hypothesized causes and effects, including estimates of the strength of each causal interaction. It is driven initially by simple guesses regarding such interactions, inspired by events occurring in close temporal succession. The model maintains its structure dynamically (including omitting or even reversing the current best-guess dependencies, if warranted by new evidence), and estimates the projected probability of possible outcomes by performing inference on the resulting Bayesian network. The model reproduces the human performance in the present dynamical task.

**Keywords:** Temporal learning, causality, structure learning, Dynamic Bayesian graphical model, STDP.

## Introduction

There are many types of cues that an agent can use to learn the causal structure of its interactions with the environment, such as prior knowledge (which constrains the hypothesis space), statistical relations, intervention, and temporal ordering (Lagnado et al., 2007). Among these, temporal ordering is particularly intriguing. First, proximity among cues appears to play a central role in learning structure in time and space (Goldstein et al., 2010). Second, in causal learning, temporal ordering, similarly to intervention, carries with it information regarding the direction of causality, which is crucial for prediction. Finally, putative causal relationships between ordered events that occur in close temporal proximity can be registered by relatively well-understood computational mechanisms akin to those that support synaptic modification in nervous systems.

Both the learning of causal structure (as in model selection) and the modification of its parameters (as in classical schemes such as  $\Delta P$ , *PowerPC*, and Rescorla-Wagner) can be put on a rational basis (Griffiths and Tenenbaum, 2009; Holyoak and Cheng, 2011). Given

the special appeal of temporal proximity as a cue to causal structure and strength, a simple, incremental, heuristic approach to causal learning based on this cue is, however, worth exploring — particularly if such an approach proves effective in dealing with dynamical scenarios where irrelevant variables abound and where the model may need to be modifiable on the fly. In this paper, we describe a dynamical causal Bayesian model that uses temporal proximity among cues to learn its structure and parameters from a continuous stream of observations and action-related rewards in a computer game-related scenario.

## Dynamic causal Bayesian modeling

Consider a dynamic situation described by a set of binary variables  $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$ , whose values occasionally change over time, and where  $X_i = 1$  indicates the presence of some item or feature and  $X_i = 0$  its absence. The causal relationships among the variables in  $\mathcal{X}$ , if any, are initially unknown; each one could be a cause, an effect, or neither. Our approach integrates bottom-up, event-driven learning with top-down revisions as dictated by the model’s self-maintained track record in predicting impending events or the outcomes of actions.

The graph representing the model’s current hypothesis regarding the causal relationships over the members of  $\mathcal{X}$ , which serve as its vertices, has initially no edges. As time progresses, edges are added to the graph according to the temporal order of the observed events and any interventions (the outcomes of model’s own actions). As new edges, corresponding to pairwise hypothesized causal dependencies, are added, the model attempts to integrate the subgraphs they form — “twigs” (Figure 1, left) each consisting of a pair of vertices joined by a directed edge — into a larger (eventually, global) structure.

Note that the twigs are supposed to capture causal “strength,” which we operationalize via a Hebb-like learning mechanism (detailed below), while the final causal model is intended to support probabilistic inference, by being treated as a Bayesian network. The *Union* operation (Figure 2, left), which combines twigs

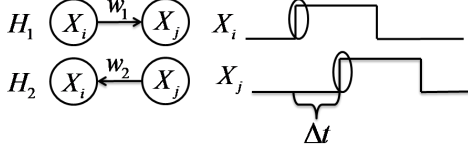


Figure 1: *Left*: the two possible *Twig* structures for two variables  $X_i$  and  $X_j$ . The weights  $w_i$  represent causal strength. *Right*: a new *Twig* is added to the model if a prior event involving some  $X_i$  is found within a short time window  $\Delta t$  of a triggering event involving  $X_j$ .



Figure 2: *Left*: the *Union* operation. In this example, there are four possible *DCBN* outcomes: causal chains  $U_1$  and  $U_2$ , a common effect structure  $U_3$ , and a common cause structure  $U_4$ . *Right*: for each case where a binary effect is driven by a real-valued “strength” link, the *Union* operation adds a hidden *softmax* node, as illustrated here for  $U_3$  by the frame drawn around  $R_{i,j}$  and  $E$  (Lu et al., 2008).

into a dynamic causal Bayesian network (*DCBN*), mediates between these two aspects of the model by inserting as needed “hidden” variables that convert real-valued strength variables into probability distributions (Figure 2, right).

As the network forms, the model becomes ready for generating predictions (inference). Given the state of observations at time  $t$ , it can be used to predict the most likely value for variables of interest at a later time. During this phase, inference is alternated with learning, with the latter being driven by the model’s monitoring of its own predictions and by comparing those to the observed outcomes. The resulting changes may include the model’s representation of the causal structure of the environment: for instance, the direction of some of the twigs (cause-and-effect subgraphs) may be reversed.

**Structure and strength learning.** We now proceed to describe the operation of the model in some detail, starting with twig learning. Every elementary subgraph, or *Twig* =  $\{C, E, w\}$ , consists of a single cause  $C$ , a single effect  $E$ , and the strength or weight  $w$  of their causal connection. Initially, no connections between variables

exist. Similarly to how humans seem to handle causal cues such as temporal ordering and proximity of notable events (Lagnado et al., 2007), the model only seeks to form a *Twig* when some item or feature appears on the scene (i.e., a variable changes state from 0 to 1). The model then scans the recent past, up to a duration of  $\Delta T$ , for a potential cause of the event at hand, in the form of the change in some other variable’s value. Any variable that is on the record as having changed its value (either from 0 to 1 or from 1 to 0) is labeled as a potential causes for the event, forming a *Twig* (Figure 1).

The weight  $w$  is modified via Hebbian learning, specifically, spike timing dependent plasticity (STDP; Caporale and Dan, 2008). This family of temporally asymmetric Hebbian rules affords quick (exponential) learning, as well as unlearning (in “negative” trials, in which the purported effect precedes the cause):

$$\begin{aligned} \Delta w_+ &= A_+ \exp(-\Delta t / \Delta T) \\ \Delta w_- &= A_- \exp(\Delta t / \Delta T) \\ w(t+1) &= \min(w(t) + \Delta w, w_{max}) \end{aligned} \quad (1)$$

where the  $+$  and  $-$  subscripts denote positive and negative trials ( $E$  following or preceding  $C$ , respectively). We set  $A_+ = 1$  and  $A_- = 0.5$ , thus giving more weight to positive evidence. If some *Twig* elements share a common variable, the model attempts to combine them by applying the *Union* operation, as in  $\text{Twig}(X_i, X_j) \cup \text{Twig}(X_i, X_k) \Rightarrow \text{DCBN}(X_i, X_j, X_k)$  (Figure 2), adding, as needed, hidden variables, as described below.

**Learning the *softmax* parameters.** To integrate a representation of causal strength into a probabilistic (Bayesian) model, we follow Lu et al. (2008) by endowing the model with internal states, or hidden variables:  $R_i$  and  $R_j$  in Figure 2, right. The state of each  $R_i$  is related to that of its parent node  $X_i$  through a Gaussian distribution parameterized by the weight  $w_i$ :

$$P(R_i | w_i, X_i) \propto e^{-(R_i - w_i X_i) / 2\sigma_i^2} \quad (2)$$

The binary effect variable  $E = e_i$ ,  $e_i$  being the  $i$ th discrete value of  $E$ , is driven, in turn, by  $\mathbf{R}$  through a *softmax* function:

$$P(E = e_i | \mathbf{R}) = \frac{\exp(\mathbf{w}(:, i)^T \mathbf{R} + \mathbf{b}(i))}{\sum_j \exp(\mathbf{w}(:, j)^T \mathbf{R} + \mathbf{b}(j))} \quad (3)$$

where  $\mathbf{R}$  is the vector that comprises  $R_i$  and  $R_j$ , and  $\mathbf{w}$  and  $\mathbf{b}$  are parameters that are learned as the model is exposed to data, using an iteratively reweighted least squares (IRLS) algorithm (Green, 1984).

**Inference.** We illustrate the inference process, in which the model is used to generate predictions for some



---

**Algorithm 1** Dynamic causal Bayesian model (DCBN)

---

```
1: INITIAL LEARNING
2: Given: variables  $\mathcal{X}$ ; window  $\Delta T$ .
3: Note:  $|\mathcal{X}| = n$  is the number of variables.
4: Note:  $t$  is the current time.
5: for  $i = 1 \rightarrow n$  do
6:   if  $X_i^t == 1$  and  $X_i^{t-1} == 0$  then
7:     for  $j = 1 \rightarrow n$  do
8:       if  $X_j$  preceded  $X_i$  by  $\Delta t < \Delta T$  then
9:         if No  $Twig(X_i, X_j, w_{ij})$  exists then
10:           Compute  $w_{ij}$  (eq. 1);
11:           Create  $Twig(X_i, X_j, w_{ij})$ ;
12:         else
13:           Update  $w_{ij}$  (eq. 1);
14:         end if
15:       end if
16:     end for
17:   end if
18:   Compute  $Union$  over  $Twigs$  to form DCBN;
19:   Train  $softmax$  (eq. 3) for hidden variables;
20: end for
21: INFERENCE AND FURTHER LEARNING
22: while True do
23:   Perform inference on DCBN;
24:   if inference deviates from observation then
25:     Modify  $Twig$  weights  $w$  (eq. 1);
26:     for every  $Twig$  do
27:       if  $w < 0$  then
28:         Reverse the edge;
29:         Re-learn  $w$ ;
30:       end if
31:     end for
32:     If structure changed, recompute  $Union$ ;
33:   end if
34:   Retrain  $softmax$  parameters;
35: end while
```

---

variable values, given others, on an example with an effect  $E$  that depends on two causes,  $X_{1,2}$  (Figure 2, right). Given the values of  $X_{1,2}$ , inference requires integration over the hidden variables:

$$P(E | w_1, w_2, X_1, X_2) = \int \int P(E | R_1, R_2) \prod_{i=1}^2 P(R_i | w_i, X_i) P(X_i) dR_1 dR_2 \quad (4)$$

Because  $R_i$  are unobserved and continuous and their descendants are discrete, exact inference is impossible (Lerner et al., 2001; Murphy, 1999). As an approximation, we sample each  $R_i$ , conditioned on its parent  $X_i$  and weight  $w_i$ . Specifically, if  $X_i = 1$ , we sample from the Gaussian distribution associated with it (eq. 2); if  $X_i = 0$ , we sample from a zero-mean Gaussian distribution, which is the same for all the variables.

We then discretize the integral, with a step size of 0.1

and lower and upper bounds set to  $L = \min_{X_i}(-4\sigma_i)$  and  $U = \max_{X_i}(w_i X_i + 4\sigma_i)$ , respectively:

$$P(E | w_1, w_2, X_1, X_2) = \sum_{R_1=L}^U \sum_{R_2=L}^U P(E | R_1, R_2) \prod_{i=1}^2 P(R_i | w_i, X_i) P(X_i) \quad (5)$$

When the predicted value of  $E$  yielded by the inference step matches the observed value with a high confidence, the model is not modified. Every time the prediction falters, the model learns; both its structure and its parameters can be modified, as described in Algorithm 1.

## The experiments

To evaluate the model, we tested it in an experiment that involved learning in a dynamically unfolding game situation.<sup>1</sup> For the same experiment, we also collected performance data from human subjects.

Most of the published studies of causal learning to date have been conducted in somewhat artificial behavioral settings. In many studies, the task consists of a series of trials, in each of which the subject is presented with a few stimuli — often just two or three items on a blank screen, along with choices that can be made via a key press (e.g., Steyvers, Tenenbaum, Wagenmakers, and Blum, 2003). More elaborate tasks may involve a contraption that displays a few objects whose behaviors may be causally interlinked (e.g., Kushnir, Gopnik, Lucas, and Schulz, 2010). The narrative context that defines the task for the subjects is often couched in causal language (as in “Can you tell what makes the box go?”). In comparison, in the present study the behavioral task involved an arguably more natural situation: playing a computer game, which unfolds in real time, and requires that the subject drive down a track surrounded by various objects, while attempting to accumulate rewards.

The experimental platform we used is an adaptation of a car-racing computer game.<sup>2</sup> The virtual environment through which the subject is driving consists of tracks surrounded by scenes whose composition is controlled. It is flexible enough to support various types of cues to causal structure, including interventions (Lagnado et al., 2007). Moreover, because the game can be played against another subject or against a computer program, it affords the study of social effects in learning (Goldstein et al., 2010).<sup>3</sup>

---

<sup>1</sup>In a separate study, we used the model to replicate successfully some of the standard effects in causal learning, such as forward and backward blocking (Holyoak and Cheng, 2011).

<sup>2</sup><http://supertuxkart.sourceforge.net/> (public domain). We modified the game to support Wiimote and to incorporate our tracks, scenes, and reporting. The modified code is available upon request.

<sup>3</sup>Note that instructions that the subject receives from the experimenter may be considered a kind of social cue.



Figure 3: A typical game scene, as presented to the subjects as part of the instructions for the experiment.

scene type	<i>crate</i>	<i>dog</i>	<i>cat</i>	<i>fox</i>	box contents
1	1	[0]	1	0	$[plunger]_{t+}$
2	1	[1]	1	0	$[cake]_{t+}$
3	1	[0]	1	1	$[plunger]_{t+}$
4	1	[1]	1	1	$[cake]_{t+}$
5	0	0	1	0	$[plunger]_{t+}$
6	0	1	1	0	$[cake]_{t+}$

Table 1: The six scene types in the experiment, with the presence or absence of various objects indicated by 1/0. When *crate* is present, *dog* is hidden inside it (bracketed,  $[\cdot]$ ). The contents of the surprise box (*cake* or *plunger*) become visible only if the subject actively “takes” it (signified by  $[\cdot]_{t+}$ ). Note that *dog* perfectly predicts  $[cake]_{t+}$ , but subjects who miss the significance of *crate* will be unable to distinguish between scenes 1 and 2, or 3 and 4.

## The behavioral experiment

Given the novelty and the potential difficulty of the dynamical learning scenario, in this study we opted for a maximally simple dependency to be learned: a single causal link between two variables. Each scene in the experiment could include any or all of the following objects: a dog, a cat, a fox, and a crate (Figure 3). In addition, in each scene there was a “surprise” box which, if the subject chose to “take” it, revealed the reward: a cake or a plunger, depending on the appearance of other objects in the scene. The subjects were instructed to collect as many cakes as possible, while refraining from taking plungers. Altogether, each subject encountered 252 scenes: 6 different racetracks  $\times$  3 laps  $\times$  14 scenes drawn at random for each track from among the scene types listed in Table 1.

The subjects’ task is best *seen*<sup>4</sup> as learning a directed

<sup>4</sup>The question of what the relationship between *dog* and *cake* in this simple scenario really *is*, causal or associative, is best avoided, given the philosophical issues surrounding causality (Schaffer, 2009). Somewhat paradoxically, the distinction is easier for more complex networks of dependencies

or causal (rather than an undirected or merely associative) relationship, for two reasons: the asymmetry in the temporal structure of each scene encounter and in the functional significance of its components. First, the reward never co-occurred with any of the other variables: rather, it always followed them temporally, and then only if the subject actively intervened by opening the surprise box. Second, it made no sense for the subjects to hypothesize symmetrical functional roles for the reward and for the other variables, given that their goal was formulated exclusively in terms of the reward. In any case, no causal language was used in the instructions given to subjects, which makes the present experimental set up arguably more natural as a platform for exploring simple learning in the wild than those that explicitly require the subjects to seek causal explanations for behavioral outcomes.

Eighteen subjects, recruited online from the Cornell University subject pool, participated in the study for course credit. The dependent variable, Correct, was defined as equal to 1 in trials where the subject opened a box with *cake* or refrained from opening a box with *plunger*. A mixed model analysis using the *lmer* package (Bates, 2005), with a binomial linking function, and with Subject, Track, and Scene as random factors, yielded a significant effect of Lap on Correct ( $z = 7.53$ ,  $p = 5.1 \times 10^{-14}$ ). Averaged over tracks, the subjects’ Correct rate reached 0.61 in the third lap. The far from perfect performance is understandable, given that the inconsequential parts of the game environment (such as a stable with horses, bales of hay, etc.), as well as of the surprise-box scenes themselves, made it difficult for subjects to home in on the truly predictive variable (*dog*). Moreover, in scene types 1 through 4, the dog appears inside a crate and is thus not visible, unless the subject drives through the crate (something that few subjects ventured to do).

The evolution of subjects’ performance over time is illustrated for each scene type in Figure 4, right. Debriefing indicated that subjects generally assumed correctly that the contents of the surprise box could be anticipated by noting which of the other objects were present in the scene. Many of the subjects did not, however, allow for the possibility that a cause may be hidden, which prompted them to invent incorrect explanations for the difference between scene types (1,2) and (3,4). Some subjects also tried to find patterns in the irrelevant variables such as the distances among objects, the curving of the track, and the location of the box with respect to other items.

## Modeling results

We simulated the behavioral experiment by feeding the model incrementally the same sequence of observations among variables, such as those explored by Blaisdell et al. (2006).

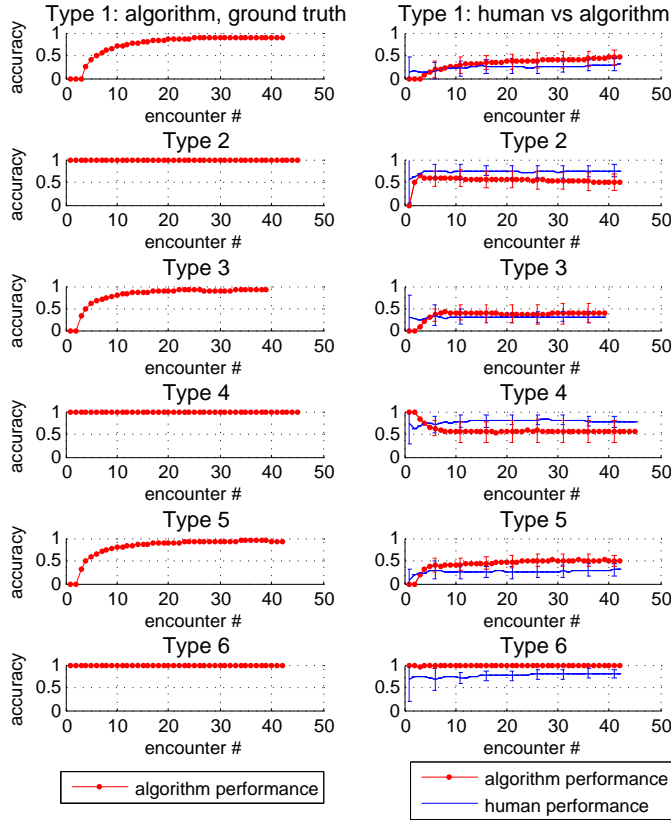


Figure 4: *Left*: the performance of the algorithm on ground-truth data, for each of the six scene types. *Right*: the performance of 18 runs of the algorithm (filled circles) and of the 18 subjects subjects in a real run (means with 95% confidence intervals). The ups and downs in the algorithm’s performance over time are due to its sensitivity to the order of scene appearance (batch algorithms do not exhibit this behavior).

encountered by the human subjects, namely, the values of the four variables listed in Table 1, plus, in the cases where the model decided to open the surprise box, the value of reward. In the first lap (14 scenes; the “initial learning” phase in Algorithm 1), the model was set to open every box. Subsequently, if the model’s decision whether or not to open the box could be made with 95% confidence,<sup>5</sup> it chose the recommended action; otherwise it flipped a coin. If the decision was to open the box, the model used the outcome to adjust its parameters; if not, it simulated an outcome by adding to the predicted value of the reward a random number (distributed uniformly

<sup>5</sup>As decided by a binomial test with a confidence interval of  $\hat{p} \pm z_{1-\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n}$ , where  $\hat{p}$  is the sample proportion of successes in the observed sequence of trials and  $z_{1-\alpha/2}$  is the  $1-\alpha/2$  percentile of a standard normal distribution, with  $\alpha$  being the error percentile and  $n$  the sample size. For a 95% confidence level,  $\alpha = 5\%$  and  $z_{1-\alpha/2} = 1.96$ .

in  $[0, 1]$ ) and used that outcome to learn.<sup>6</sup>

As can be seen in Figure 4, left, when fed ground-truth data, the model learned quickly and reliably.<sup>7</sup> More to the point, when presented with the real sequence of observations, it generally behaved similarly to human subjects (Figure 4, right), reaching a comparable level of performance: 0.66 accuracy in the third lap. As with the human subjects, the effect of Lap was significant ( $z = 2.56$ ,  $p = 0.01$ ). In Figure 4, right, in those cases where the dog was hidden from view (scene types 1 through 4; see Table 1), the human subjects performed poorly, and the algorithm too converged to a chance-level performance.

## Conclusions

Similarly to some other recent studies and models of causal learning (Lu et al., 2008; Lagnado and Speekenbrink, 2010; Bonawitz et al., 2011), the present work focuses on sequential learning and inference. There are also important differences. First, our behavioral setup uses a dynamic video game that subjects readily relate to. Second, the model we develop is rooted in some basic intuitions regarding how animals learn the causal structure of dynamic situation: (1) the importance of close temporal succession of events and outcomes, (2) the utility of neural-like mechanisms that may register it, and (3) a heuristic approach to bootstrapping causal learning from very simple pairwise dependencies gleaned from the data. In those respects, the algorithm we offer is a special-purpose model rather than a general learner.

To ascertain that subjects in our game scenario engage in causal learning and inference, rather than in memorization of contextual cues they believe to be associated with particular outcomes, future experiments will need to include explicit intervention-based tests (cf. Blaisdell, Sawa, Leising, and Waldmann, 2006), including having the subjects manipulate the variables of their choice to test any hypotheses that they may have formed. It would also be interesting to analyze the evolution over time of the subjects’ choices in opening or avoiding reward boxes: early in the experiment, it is rational to open boxes, so as to gather data; as the subjects develop an ability to predict the reward, they should become more choosy. This sequential behavior can then be compared to that of the model (Bonawitz et al., 2011).

The model itself can be improved and extended in several ways. For instance, as it is tested on learning

<sup>6</sup>Without some such mechanism, the model would have no way of recovering from a string of “don’t open” decisions — a problem that is peculiar to models that intersperse learning with inference.

<sup>7</sup>In comparison, a straightforward model selection approach based on maximum likelihood or AIC/BIC optimization, implemented with the Bayes Network Toolbox for Matlab (Murphy, 2001), trained incrementally on the ground truth data, did not converge to the right causal graph for this experiment.

tasks that involve more complex causal structure than that in the present study, it may be necessary to include methods for detecting and “defusing” loops that would otherwise complicate inference. Furthermore, the model can be made to incorporate additional cues to causal structure, in particular, interventions (Steyvers et al., 2003), global contextual cues, and factors such as eligibility traces (Izhikevich, 2007) that would allow it to learn from such cues across multiple time scales. Finally, if equipped with a vision front end and real-valued outputs, a model rooted in the present approach may employ reinforcement learning (Fox et al., 2008; Hosoya, 2009) to master driving around the track and competing directly with a human participant.

**Acknowledgments.** We thank Tamar Kushnir and the members of her lab for valuable comments on the present project. EN was supported by Cornell University’s Tri-Institutional Training Program in Computational Biology and Medicine.

## References

- Bates, D. (2005). Fitting linear mixed models in R. *R News* 5, 27–30.
- Blaisdell, A. P., K. Sawa, K. J. Leising, and M. R. Waldmann (2006). Causal reasoning in rats. *Science* 311, 1020–1022.
- Bonawitz, E., S. Denison, A. Chen, A. Gopnik, and T. L. Griffiths (2011). A simple sequential algorithm for approximating bayesian inference. In L. Carlson, C. Hölscher, and T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pp. 2463–2468.
- Caporale, N. and Y. Dan (2008). Spike timing-dependent plasticity: A Hebbian learning rule. *Annual Review of Neuroscience* 31, 25–46.
- Fox, C. N., N. Girdhar, and K. N. Gurney (2008). A causal Bayesian network view of reinforcement learning. In *Proceedings of the 21th International Florida Artificial Intelligence Research Society Conference, FLAIRS-21*, pp. 109–110.
- Goldstein, M. H., H. R. Waterfall, A. Lotem, J. Halpern, J. Schwade, L. Onnis, and S. Edelman (2010). General cognitive principles for learning structure in time and space. *Trends in Cognitive Sciences* 14, 249–258.
- Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. Royal Stat. Soc. Ser. B* 46, 149–192.
- Griffiths, T. L. and J. B. Tenenbaum (2009). Theory-based causal induction. *Psychological Review* 116, 661–716.
- Holyoak, K. J. and P. W. Cheng (2011). Causal learning and inference as a rational process: the new synthesis. *Annual Review of Psychology* 62, 135–163.
- Hosoya, H. (2009). A motor learning neural model based on Bayesian network and reinforcement learning. In *Proceedings of International Joint Conference on Neural Networks*, Atlanta, GA. IEEE.
- Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cerebral Cortex* 17, 2443–2452.
- Kushnir, T., A. Gopnik, C. Lucas, and L. Schulz (2010). Inferring hidden causal structure. *Cognitive science* 34(1), 148–160.
- Lagnado, D. A. and M. Speekenbrink (2010). The influence of delays in real-time causal learning. *The Open Psychology Journal* 3, 184–195.
- Lagnado, D. A., M. R. Waldmann, Y. Hagmayer, and S. A. Sloman (2007). Beyond covariation: cues to causal structure. In A. Gopnik and L. Schulz (Eds.), *Structure*, pp. 1–48. Oxford University Press.
- Lerner, U., E. Segal, and D. Koller (2001). Exact inference in networks with discrete children of continuous parents. In *Proc. 17th Conf. on Uncertainty in Artificial Intelligence*, pp. 319–238.
- Lu, H., R. R. Rojas, T. Beckers, and A. Yuille (2008). Sequential causal learning in humans and rats. In B. C. Love, K. McRae, and V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pp. 188–195.
- Murphy, K. (1999). A variational approximation for bayesian networks with discrete and continuous latent variables. In *Proc. UAI*, Volume 99, pp. 457–466.
- Murphy, K. P. (2001). The Bayes Net Toolbox for Matlab. *Computing Science and Statistics* 33, 331–351.
- Schaffer, J. (2009). The metaphysics of causation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2009 ed.).
- Steyvers, M., J. B. Tenenbaum, E. J. Wagenmakers, and B. Blum (2003). Inferring causal networks from observations and interventions. *Cognitive Science* 27, 453–489.

# Noisy Newtons: Unifying process and dependency accounts of causal attribution

Tobias Gerstenberg<sup>1</sup> (t.gerstenberg@ucl.ac.uk), Noah Goodman<sup>2</sup> (ngoodman@stanford.edu),  
David A. Lagnado<sup>1</sup> (d.lagnado@ucl.ac.uk) & Joshua B. Tenenbaum<sup>3</sup> (jbt@mit.edu)

<sup>1</sup>Cognitive, Perceptual and Brain Sciences, University College London, London WC1H 0AP

<sup>2</sup>Department of Psychology, Stanford University, Stanford, CA 94305

<sup>3</sup>Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

## Abstract

There is a long tradition in both philosophy and psychology to separate process accounts from dependency accounts of causation. In this paper, we motivate a unifying account that explains people's causal attributions in terms of counterfactuals defined over probabilistic generative models. In our experiments, participants see two billiard balls colliding and indicate to what extent ball A caused/prevented ball B to go through a gate. Our model predicts that people arrive at their causal judgments by comparing what actually happened with what they think would have happened, had the collision between A and B not taken place. Participants' judgments about what would have happened are highly correlated with a noisy model of Newtonian physics. Using those counterfactual judgments, we can predict participants' cause and prevention judgments very accurately ( $r = .99$ ). Our framework also allows us to capture intrinsically counterfactual judgments such as *almost* caused/prevented.

**Keywords:** causality; counterfactuals; attribution; physics.

## Introduction

There has been a longstanding divide in philosophy between two fundamentally different ways of conceptualizing causality. According to *dependency accounts* of causation, what it means for A to be a cause of B is that B is in some way dependent on A. Dependence has been conceptualized in terms of regularity of succession (A is regularly succeeded by B; Hume, 2000 [1748]), probabilities (the presence of A increases the probability of B; Suppes, 1970) or counterfactuals (if A had not been present B would not have occurred; Lewis, 1970). For *process accounts*, in contrast, what it means for A to be a cause of B is that a physical quantity is transmitted along a pathway from A to B (Dowe, 2000).

The psychological literature on causal learning and attribution neatly maps onto the two major accounts in philosophy. On the one hand, people have been shown to use contingency information when drawing inferences about whether and how strongly two events are causally linked (Cheng, 1997). On the other hand, people display a preference to choose causes that influence an effect via a continuous causal mechanism over causes that are connected with the effect through mere dependence (Walsh & Sloman, 2011).

A point that has been raised in favor of process accounts is that they are capable of capturing the semantics of different causal terms. Whereas dependency accounts have mostly focussed on causation and prevention, Wolff (2007) has provided a process account that not only predicts when people use the terms *cause* and *prevent* but also *enable* and *despite*. Following a linguistic analysis of causation by Talmy (1988) in terms of force dynamics, Wolff (2007) argues that the aforementioned causal terms can be reduced to configurations of force vectors. For example, what it means for a patient (P) to

have been caused by an affector (A) to reach an endstate (E) is that P did not have a tendency towards E, A impacted on P in a way that their force vectors were not pointing in the same direction and P reached E. If, in contrast, the force vectors of both P and A point towards E and P reaches E, the model predicts that people will say "A enabled (rather than caused) P". Importantly, according to Wolff's account, the core dimensions which underlie the different causal terms, such as P's tendency towards E, are defined in strictly non-counterfactual terms. Hence, "tendency" is defined as the direction of P's force rather than whether P would reach E in the absence of any other forces.

While the force dynamics model has strong intuitive appeal for interactions between physical entities, it is questionable how it can be extended to capture causal attributions in situations involving more abstract entities. For example, one might legitimately assert that the fall of Lehman Brothers caused the financial crisis or that Tom's belief that he forgot his keys caused him to turn around and go back home. While it is unclear how these causal relationships could be expressed in terms of force vectors, they do not pose a problem for the more flexible dependency accounts. For example, according to a counterfactual account, Tom's belief qualifies as cause of his behaviour if it is true that his behavior would have been different had the content of his belief been different. Hence, there appears to be a trade-off between the semantic richness of process accounts on the one hand and the generality and flexibility of dependency accounts on the other hand.

Rather than fostering the divide between process accounts and dependency accounts, we propose a theory of causal attribution that combines the best of both worlds. In the spirit of Pearl (2000), we model causal attributions in terms of counterfactuals defined over probabilistic generative models. However, we agree with Wolff (2007) that people's causal knowledge is often richer than what can be expressed with a causal Bayes net. We aim to unify process and dependency accounts by showing that people have intuitive theories in the form of detailed generative models, and that causal judgements are made by considering counterfactuals over these intuitive theories. Here we demonstrate the superiority of our approach over existing models of causal attribution in a physical domain. We show that people use their intuitive understanding of physics to simulate possible future outcomes and that their causal attributions are a function of what actually happened and their belief about what would have happened had the cause not been present.



## Overview of Experiments and Model Predictions

Before discussing the predictions of our model and the supporting evidence from four experiments, we describe the domain to which we applied our model. In all experiments, participants saw the same 18 video clips which were generated by implementing the physics engine Box2D into Adobe Flash CS5. Figure 1 depicts a selection of the clips.<sup>1</sup> In each clip, there was a single collision event between a grey ball (A) and a red ball (B) which enter the scene from the right. Collisions were elastic and there was no friction. The black bars are solid walls and the red bar on the left is a gate that balls can go through. In some clips B went through the gate (e.g. clip 18) while in others it did not (e.g. clip 5). In the 18 video clips, we crossed whether ball B went through the gate given that it collided with ball A (rows in Figure 1: actual miss/close/hit) with whether B would go through the gate if A was not present in the scene (columns in Figure 1: counterfactual miss/close/hit). Participants viewed two clips for each combination of actual and counterfactual outcome.

In Experiments 1 and 2, the video clips stopped shortly after the collision event. Participants judged whether ball B will go through the gate (Experiment 1) or whether ball B would go through the gate if ball A was not present in the scene (Experiment 2). In Experiment 3, participants saw each clip played until the end and then judged to what extent ball A caused ball B to go through the gate or prevented B from going through the gate. Finally, in Experiment 4 participants chose from a set of sentences which best describes the clip they have just seen. All experiments were run online and participants were recruited via Amazon Mechanical Turk.

In order to model people’s predictions of actual and counterfactual future outcomes, we developed the Physics Simulation Model (PSM) which assumes that people make use of their intuitive understanding of physics to simulate what will or what might have happened. Hamrick, Battaglia, and Tenenbaum (2011) have shown that people’s stability judgments about towers of blocks is closely in line with a noisy model of Newtonian physics. While in their model, the introduced noise captures people’s uncertainty about the exact location of each block, the noise in our model captures the fact that people cannot perfectly predict the trajectory of a moving ball (cf. Figure 1, clip 1). We introduce noise via drawing different degrees of angular perturbation from a Gaussian distribution with  $M = 0$  and  $SD = \{1, 2, \dots, 10\}$  which is then applied to B’s actual velocity vector (given that it collided with A, clip 1 bottom left) or B’s counterfactual velocity vector (given that A was not present in the scene, clip 1 top left) at the time of collision.

We evaluate the probability that B would go through the gate when A was present,  $P(B|A)$ , or absent  $P(B|\neg A)$  by forward sampling from our noisy versions of Newtonian physics.

<sup>1</sup>All clips can be viewed here: <http://www.ucl.ac.uk/lagnado-lab/experiments/demos/physicsdemo.html>

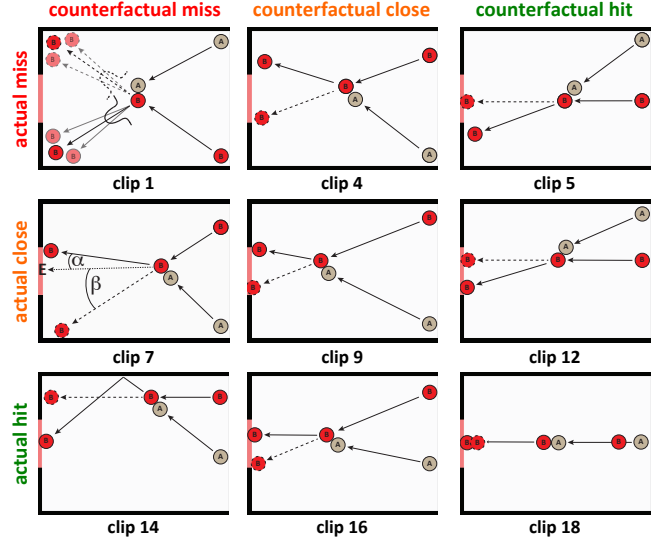


Figure 1: Selection of clips used in the experiment. Solid arrows = actual paths, dashed arrows = counterfactual paths. Clip 1 depicts an illustration of the Physics Simulation Model and clip 7 of the Actual Force Model. *Note:* actual miss = B clearly misses; actual close = B just misses/hits; actual hit = B clearly hits; counterfactual miss = B would have clearly missed; counterfactual close = B would have just missed/hit; counterfactual hit = B would have clearly hit.

For each clip and degree of noise ( $SD$ ), we ran 1000 noisy repetitions of the original clip and counted the worlds in which B goes through the gate given that A was present  $P(B|A)$  or absent  $P(B|\neg A)$ .

## Experiments 1 & 2: Intuitive Physics

The aim of Experiments 1 and 2 was to evaluate how well people make use of their intuitive physical knowledge to predict actual (Experiment 1) or counterfactual (Experiment 2) future states. Participants saw 18 video clips (see Figure 1 for some examples) up to the point shortly after the two balls collided (0.1s). After having seen the clip twice, participants answered the question: “Will the red ball go through the hole?” (Experiment 1,  $N = 21$ ) or “Would the red ball have gone through the goal if the gray ball had not been present?” (Experiment 2,  $N = 20$ ). Participants indicated their answers on a slider that ranged from 0 (“definitely no”) to 100 (“definitely yes”). The midpoint was labeled “uncertain”. After having made their judgment, participants viewed the clip until the end either with both balls being present (Experiment 1) or with ball A being removed from the scene (Experiment 2).

## Results and Discussion

Participants were accurate in judging whether ball B will go through the gate (Experiment 1) or would have gone through the gate (Experiment 2) with a mean absolute difference from the deterministic physics model (which assigns a value of 100 if B goes in and 0 if B does not go in) of 28.6 ( $SD = 29.9$ ) in Experiment 1 and 25.1 ( $SD = 30.5$ ) in Experiment 2. Figure 2 shows the correlation of the PSM with participants’ judgments in Experiment 1 (solid black line) and Experiment 2 (dashed black line) for different degrees of noise. While people’s judgments already correlate quite well with a deterministic Newtonian

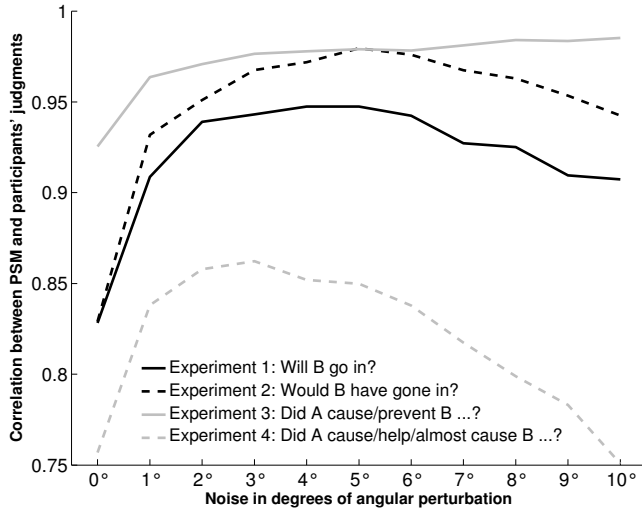


Figure 2: Correlation of the Physics Simulation Model with people's judgments in all four experiments for different degrees of noise.

nian physics model (degree of noise = 0°), introducing small degrees of noise results in much higher correlations with a maximum correlation of  $r = .95$  in Experiment 1 and  $r = .98$  in Experiment 2 for  $SD = 5^\circ$ .

The results of Experiments 1 and 2 show that people are capable of mentally simulating what will happen (Experiment 1) or what would have happened (Experiment 2). Given that each clip stopped very shortly after the collision event, participants' accuracy in judging whether ball B will go in or not is quite impressive.

### Experiment 3: Causation and Prevention

In Experiment 3, we wanted to investigate how people use their intuitive understanding of physics to make judgments about the extent to which one event caused or prevented another event from happening. Unlike in Experiments 1 and 2, participants ( $N = 22$ ) saw each clip played until the end. After having seen each clip twice, participants answered the question "What role did ball A play?" by moving a slider whose endpoints were labeled with "it prevented B from going through the hole" and "it caused B to go through the hole". The midpoint was labeled "neither". The slider ranged from -100 (prevented) to 100 (caused). Participants were instructed that they could use intermediate values on the slider to indicate that ball A somewhat caused or prevented B.

### Model Predictions

**Physics Simulation Model** According to the PSM, people arrive at their cause and prevention judgments by comparing what actually happened with what they think would have happened if the cause event had not taken place. More specifically, our model predicts that people compare  $P(B|A)$ , the probability that ball B will go through the gate given that it collided with ball A, with  $P(B|\neg A)$ , the probability that B would have gone through the gate if A had not been present in the scene. Since participants in Experiment 3 watch the clips until the end, the value of  $P(B|A)$  is certain: it is either 1 when B goes through the gate or 0 when B misses the gate.

In order to determine  $P(B|\neg A)$ , the PSM assumes that people use their confidence in the result of their mental simulation of what would have happened had A not been present.

In general, if  $P(B|A) - P(B|\neg A)$  is negative, participants should say that A prevented B from going through the gate. Intuitively, if it was obvious that B would have gone in had A not been present (i.e.  $P(B|\neg A)$  is high) but B misses the gate as a result of colliding with A (i.e.  $P(B|A) = 0$ ), A should be judged to have prevented B from going through the gate. Similarly, if the difference is positive, participants should say that A caused B to go through the gate. If the chance that B would have gone through the goal without A was low but, as a result of colliding with A, B goes through the gate, A should be judged to have caused B to go through the gate. Clip 1 in Figure 1 shows an example for which our model predicts that participants will say that A neither caused nor prevented B.  $P(B|A)$  is 0 since B does not go through the gate. However,  $P(B|\neg A)$  is also close to 0 since it is clear that B would have missed the gate even if A had not been present in the scene.

**Actual Force Model** The Actual Force Model (AFM) is our best attempt to apply Wolff's (2007) force dynamics model to our task.<sup>2</sup> According to the AFM, participants' cause and prevention judgments are a direct result of the physical forces which are present at the time of collision.

Clip 7 in Figure 1 illustrates how the AFM works. First, a goal vector (dotted arrow) is drawn from ball B's location at the time of collision to an end state (E), which we defined to be in the center of the gate. Second, the angle  $\alpha$  between the velocity vector that ball B has shortly *after* the collision with A (solid arrow) and the goal vector as well as the angle  $\beta$  between the velocity vector that ball B has shortly *before* colliding with A (dashed arrow) are determined. Third, the model predicts people's cause and prevention judgments via comparison of  $\alpha$  and  $\beta$ . In general, if ball B goes in and  $\beta - \alpha$  is greater than 0, the model predicts people will say that A caused B. Conversely, if ball B does not go in and  $\beta - \alpha$  is smaller than 0, the model predicts people will say A prevented B. For situations in which  $\beta - \alpha$  is greater than 0 but B does not go in or  $\beta - \alpha$  is smaller than 0 but B does go in, we fix the model prediction to 0. This constraint prevents the model from predicting, for example, that people will say "A caused B" when B missed the gate.

### Results and Discussion

Figure 3 shows participants' mean cause and prevention judgments for the 18 different clips together with the predictions of the PSM and the AFM. For the particular implementation of the PSM depicted in Figure 3, we directly used participants' judgments from Experiment 2 in which they indicated whether ball B would have gone through the gate if ball A had not been present as the values for  $P(B|\neg A)$ . For example, in clip 5 the ball misses the gate (hence  $P(B|A) = 0$ ) and

<sup>2</sup>While the force dynamics model only makes predictions about which out of several sentences participants will choose to describe a situation, the AFM makes quantitative predictions about the extent to which an event is seen as causal/preventive.

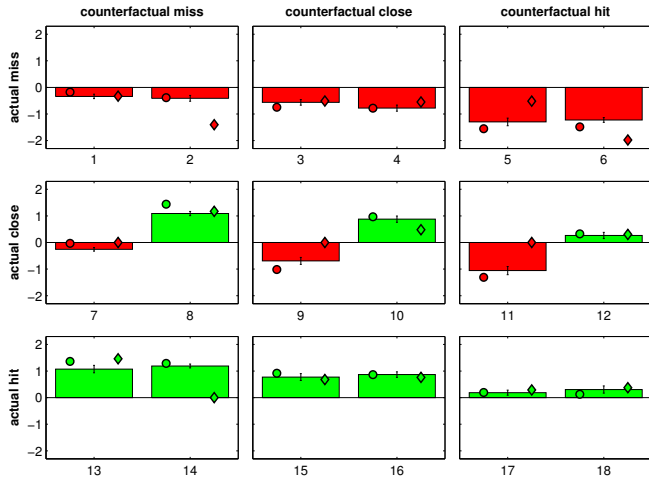


Figure 3: Z-scored mean cause (green) and prevention ratings (red) for the different clips denoted on the x-axes.  $\circ$  = predictions of the Physics Simulation Model ( $r = .99$ ),  $\diamond$  = predictions of the Actual Force Model ( $r = .77$ ). Error bars are  $\pm 1$  SEM.

participants' average confidence rating from Experiment 2 of whether B would have gone through in the absence of A is 97% (hence  $P(B|\neg A) = .97$ ). Thus the PSM predicts that participants will indicate that A strongly prevented B in this clip, because  $P(B|A) - P(B|\neg A)$  is close to the minimum of -1.

Overall, the PSM predicts participants' cause and prevention ratings very well with  $r = .99$  and  $RMSE = 0.02$ . A high median correlation across participants of  $r = .88$  with a minimum of  $r = .61$  and a maximum of  $r = .95$  demonstrates that the good performance of the PSM is not due to a mere aggregation effect. The PSM achieves its high predictive accuracy without the need for any free parameters. We directly used participants' judgments from Experiment 2 to determine the value of  $P(B|\neg A)$  for each clip. Figure 2 shows that participants' judgments also correlate highly with the PSM when we generate  $P(B|\neg A)$  through the noisy simulations of Newtonian physics as described above.

The AFM, in contrast, does not predict participants' judgments equally well with a correlation of  $r = .77$  and  $RMSE = 0.44$ . While the AFM predicts people's judgments for many of the clips, there are a number of clips for which its predictions are inaccurate (most notably: clips 2, 5, 9, 11 and 14).

Interestingly, people's cause and prevention judgments were not affected by the closeness of the actual outcome. That is, participants' cause ratings did not differ between situations in which B just went through the gate (clips 8, 10, 12:  $M = .51$ ,  $SD = .40$ ) compared to situations in which B clearly went through (clips 13 - 18:  $M = .49$ ,  $SD = .42$ ). Similarly, prevention judgments were not different between situations in which B just missed (clips 7, 9, 11:  $M = -.41$ ,  $SD = .43$ ) and situations in which B clearly missed (clips 1-6:  $M = -.47$ ,  $SD = .44$ ).

People's cause and prevention judgments were very well predicted by the PSM. In order to judge whether ball A caused or prevented ball B, participants appear to compare what actually happened with what they think would have happened had A not been present. This very high correlation is achieved

without the need for any free parameters in the model.  $P(B|A)$  is determined by the outcome of the clip and  $P(B|\neg A)$  by participants' judgments in Experiment 2. The PSM also correlates highly with participants' causal attributions when  $P(B|\neg A)$  is treated as a free parameter and estimated via the noisy Newtonian physics model with a maximum correlation of  $r = .99$  (cf. Figure 2).

The AFM which assumes that people arrive at their judgments via comparing instantaneous force vectors, rather than a mental simulation of the full physical dynamics cannot capture people's judgments equally well. Clip 14 (see Figure 1) gives an example in which the AFM gets it wrong. While participants indicate that A caused B to go through the gate (see Figure 3), the AFM model cannot predict this. In this situation, the angle between the velocity vector of B shortly after the collision and the goal vector  $\alpha$  is greater than the angle between the velocity vector of B shortly before the collision and the goal vector  $\beta$ . Hence, the model predicts that A is preventing B but since B does in fact go in, the model's prediction is fixed to 0. In defense of the AFM, it could be argued that clip 14 is better thought of as a causal chain in which A causes B to hit the wall which then causes B to go in. Whether participants would count the static wall as a cause of B going through the gate is an empirical question. In any case, the other problematic clips mentioned above remain. Each of these clips only involves a single interaction.

## Experiment 4: Almost Caused/Prevented

The results of Experiment 3 show that people's cause and prevention judgments are only influenced by their degree of belief about whether the event of interest would have happened without the cause being present and not influenced by how close the outcome actually was. However, often the closeness with which something happened clearly matters to us, such as when we almost missed a flight to Japan or only just made it in time for our talk.

As mentioned above, one of the appeals of process accounts is that they acknowledge the semantic richness of the concept of causation by making predictions about which out of several causal verbs people will choose to describe a particular situation. In this experiment, we will demonstrate that our framework is not only capable of capturing the difference between different causal verbs such as *caused* or *helped* but also predicts when people make use of intrinsically counterfactual concepts such as *almost caused* or *almost prevented*. Current process accounts (e.g. Wolff, 2007) cannot make predictions in these situations as they aim to analyze causality without making reference to counterfactuals.

In Experiment 4, participants ( $N = 41$ ) had to select from a set of seven sentences the one that describes the situation best. The sentences were: A caused / helped / almost caused B to go in the hole; A prevented / helped to prevent / almost prevented B from going in the hole; A had no significant effect on whether B went in the hole or not.



Table 1: Predicted probability of choosing different sentences in Experiment 4.

	outcome	probability
(1) caused	hit	$1 - P(B \neg A)$
(2) helped	hit	$1 - \text{abs}(0.5 - \text{caused})^*$
(3) almost caused	miss	$p(\text{almost } B A) - \text{prevented}$
(4) prevented	miss	$p(B \neg A)$
(5) helped to prevent	miss	$1 - \text{abs}(0.5 - \text{prevented})^*$
(6) almost prevented	hit	$p(\text{almost } \neg B A) - \text{caused}$
(7) no effect	hit/miss	$1 - \max((1), \dots, (6))$

\*rescaled to range from 0 to 1

## Model Predictions

Table 1 gives an overview of the model predictions which are a function of the outcome, that is, whether B went in or not, and the probabilities  $P(B|\neg A)$ ,  $P(\text{almost } B|A)$  and  $P(\text{almost } \neg B|A)$ . For  $P(B|\neg A)$  we can again use participants' judgments from Experiment 2 or the predictions of the PSM.

The model's predictions for *caused* and *prevented* are identical to the predictions in Experiment 3. According to our model, the difference between *caused* and *helped* is an epistemic one. People are predicted to select *helped* when B went in and when they were uncertain about what would have happened had A not been present. However, when it was clear that B would have gone in or would have missed, people are predicted to select *no effect* or *caused*, respectively, and not *helped*. Similarly, when B missed and it was uncertain whether B would have gone in, the model predicts that people select *helped to prevent*.

In order to predict when people select *almost caused* or *almost prevented*, we first have to define the probabilities  $P(\text{almost } B|A)$  and  $P(\text{almost } \neg B|A)$ . These probabilities express the closeness of an alternative counterfactual outcome to the actual outcome. One way to get at the probabilities would be to have participants judge how closely B hit or missed the gate. However, here we used a variation of the PSM to generate these probabilities. For each clip we ran a set of  $100 \times 10$  noisy simulations for different noise levels from  $SD = 1^\circ$  to  $5^\circ$ , whereby the noise was again introduced at the time of collision. If the outcome in the noisy simulations was different from the original outcome in *any* of the ten repetitions in each of the 100 simulated worlds, we counted this as a positive instance. If the outcome in all ten repetitions was the same as the original outcome, we counted this as a negative instance. For example, a value of  $P(\text{almost } \neg B|A) = .87$  in a situation in which B goes through the gate in the original clip, means that in 87 out of the 100 generated worlds, the ball did not go through the gate in at least one of the ten repetitions of each of the worlds. For the remaining 13 worlds, the ball did go in for all ten repetitions. Intuitively, in situations in which the outcome was close, the chances that the outcome in the noisy simulation will be different from the outcome in the original clip in at least one out of ten repetitions are high. However, if ball B clearly missed, for example, it is unlikely

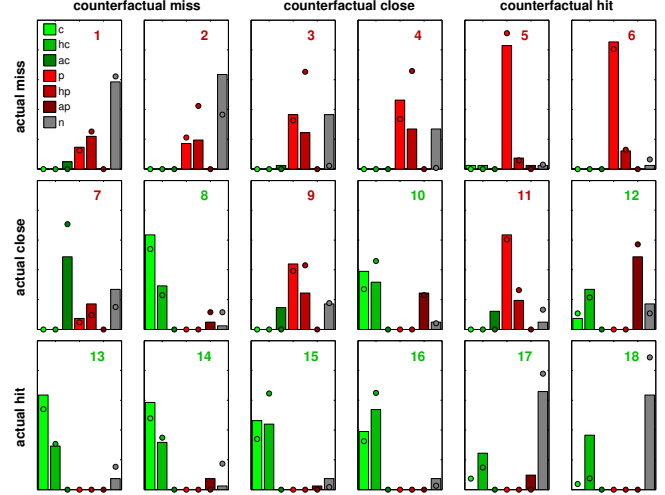


Figure 4: Frequencies with which different sentences were selected in Experiment 4 (bars) and predictions by the Physics Simulation Model (circles),  $r = .86$ . The color of the clip number indicates whether the ball went in (green) or not (red). Note: c = caused, hc = helped (to cause), ac = almost caused, p = prevented, hp = helped to prevent, ap = almost prevented, n = no significant effect.

that there will be a noisy simulation in which the introduced angular perturbation is sufficient to make B go in.

The model predicts that people will select *almost caused* when B just missed (which means that  $P(\text{almost } B|A)$  is high) and the probability that it would have gone in given that A was absent is low. People should select *almost prevented* when B just went in ( $P(\text{almost } \neg B|A)$  is high), and when it was clear that B would have gone in had A been absent ( $P(B|\neg A)$  is high). Finally, if none of these calculations result in a high value, people are predicted to select that A had *no significant effect* on whether B went through the gate.

The model predictions for the 18 different clips can be seen in Figure 4. We used Luce's (1959) choice rule to transform the different probabilities into predictions about the frequencies with which the different sentences will be selected. The model predicts that the modal choice in situations in which B does not go in changes from *prevented* for clips in which it was clear that B would have gone in (clips 5 & 6) to *helped to prevent* in situations in which it was unclear whether B would have gone in (clips 3 & 4). The same switch of the modal response as a function about the certainty of what would have happened is predicted for *caused* (clips 13 & 14) and *helped* (clips 15 & 16). If there was little uncertainty about the counterfactual outcome and it matches the actual outcome, people are predicted to select *had no effect* (clips 1, 2, 17, 18). For clip 7 in which B just misses, people are predicted to select *almost caused* since it was clear that B would have missed but for A. Conversely, for clip 12, the model predicts that people will select *almost prevented*: B just goes in and it was clear that it would have gone in without A being present.

## Results and Discussion

The model predicts the frequencies with which participants select the different sentences very well,  $r = .86$  (cf. Figure 4). Figure 2 shows the correlation when we generate  $P(B|\neg A)$

through noisily perturbing the vector rather than taking participants' ratings from Experiment 2. It predicts the modal choice correctly in 12 out of 18 clips. While participants' modal response does not change between clips 5 & 6 and clips 3 & 4 as predicted by the model, the proportion of *helped to prevent* selections clearly increases. A similar shift is observed between clips 13 & 14 and 15 & 16 for which participants' selection of *helped* increases as a function of the uncertainty over what would have happened.

As predicted by the model, participants' modal response in clip 7 is *almost caused* and in clip 12 *almost prevented*. The variance in responses within a clip is greater for the clips in which the actual outcome was close (middle row) compared to when it clearly missed (top row) or clearly hit (bottom row). For example, in clip 10 in which B just goes in and the counterfactual outcome is close the majority of participants selected *caused* or *helped* while a minority of participants selected *almost prevented*. This pattern closely matches the predictions of our model. Whether a participant is expected to select *caused* or *almost prevented* depends on the participant's subjective belief about the counterfactual outcome. If a participant thought that B would have missed she will say A *caused* or *helped* it. However, if a participant thought that B would have gone in but for A he will select *almost prevented* because B barely goes in.

The close fit between our model prediction and participants' selection of sentences demonstrates that our model is capable of capturing some of the richness of people's causal vocabulary. Our model not only allows to distinguish cases of *causing/preventing* from *helping* but also accurately predicts people's *almost caused/prevented* judgments. Process theories that analyze different causal concepts without the use of counterfactuals (e.g. Wolff, 2007) cannot make predictions about when people will say that something almost happened.

## General Discussion

In this paper, we developed a framework for understanding causal attributions that aims to break the longstanding dichotomy between process accounts and dependency accounts of causation. We showed that people's quantitative cause and prevention judgments (Experiment 3) as well as people's use of different causal verbs (Experiment 4) can be very well predicted by assuming that people compare what actually happened when the cause was present, with what they think would have happened in the absence of the cause. We provided evidence that people use their intuitive understanding of physics to simulate possible outcomes (Experiments 1 & 2). Understanding causal attributions in terms of counterfactuals defined over probabilistic generative models sidesteps the presumed trade-off between flexibility and richness described above. Our model retains the generality of dependency accounts while the use of a generative model based on Newtonian physics allows us to capture some of the richness of people's concept of causation.

According to our account, causal attributions are subjective

and model-dependent. Two observers with a different understanding of the underlying generative model are predicted to reach different causal verdicts for the same clip when their beliefs about what would have happened in the absence of the cause event differ. The noisy Newtonian physics model predicted participants' judgments well in our experiments. However, we are not committed to this particular generative model – indeed, our account predicts that the ways in which people's intuitive understanding of physics is biased will be mirrored in their causal attributions.

While our framework shares some of the key insights of Wolff's (2007) force dynamic account, such as the need for a richer specification of people's causal representations, our proposals are different in critical respects. Most importantly, our accounts differ in the role that counterfactuals play. Wolff (2007) aims to reduce causal attributions to configurations of force vectors and argues that these force representations (which are primary) can then be used for the simulation of counterfactual outcomes (which are secondary). Our account, in contrast, does not try to explain causal attributions in terms of non-causal representations but postulates that causal attributions are intimately linked with the simulation of counterfactuals. Hence, we claim that in order to say whether A caused B, it is necessary to consider what would have happened to B in the absence of A and not sufficient to only consider what forces were present at the time of interaction between A and B.

In future experiments, we will investigate how our account can handle more complex physical interactions and interactions between intentional agents.

## Acknowledgments

We thank Tomer Ullman, Peter Battaglia and Christos Bechlivanis for very helpful discussions. This work was supported by a doctoral grant from the AXA research fund (TG), a John S. McDonnell Foundation Scholar Award (NG), a ONR grant N00014-09-1-0124 (NG, JT), an ESRC grant RES-062-33-0004 (DL) and ONR MURI grants W911NF-08-1-0242 and 1015GNA126 (JT).

## References

- Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367–405.
- Dowe, P. (2000). *Physical causation*. Cambridge University Press.
- Hamrick, J., Battaglia, P., & Tenenbaum, J. (2011). Internal physics models guide probabilistic judgments about object dynamics. In *Proceedings of the 33rd annual conference of the cognitive science society*.
- Hume, D. (2000 [1748]). *An enquiry concerning human understanding: A critical edition*. Oxford University Press.
- Lewis, D. (1970). *Counterfactuals*. Blackwell.
- Luce, R. (1959). *Individual choice behavior: A theoretical analysis*. John Wiley.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.
- Suppes, P. (1970). *A probabilistic theory of causation*. North-Holland.
- Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, 12(1), 49–100.
- Walsh, C. R., & Sloman, S. A. (2011). The meaning of cause and prevent: The role of causal mechanism. *Mind & Language*, 26(1), 21–52.
- Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, 136(1), 82–111.

# Explaining children's failure in analogy making tasks: A problem of focus of attention?

**Yannick Glady, Jean-Pierre Thibaut, Robert French**

{yannick.glady, jean-pierre.thibaut, robert.french}@u-bourgogne.fr  
LEAD-CNRS, UMR 5022, University of Burgundy, Pôle AAFE – Esplanade Erasme  
21065 DIJON. FRANCE

**Agnès Blaye**

agnes.blaye@univ-amu.fr  
Aix-Marseille Université, LPC, 13621 Aix en Provence cedex 1

## Abstract

Analogical reasoning is commonly recognized as essential to human cognition, but young children often perform poorly in the classical A:B::C:? analogical reasoning task. Previous eye-tracking results have shown that children did not visually explore the A:B pair as much as adults in this task. We hypothesized that this lack of exploration could help account for the low scores of children in comparison to adults. The present study shows that children's performance improves significantly if they are required to look at and process the A:B pair before they are shown the full A:B::C:? problem. This confirms our hypothesis that the A:B pair is insufficiently processed by children during the resolution of such problems.

**Keywords:** Analogical reasoning; development; executive functions; cognition.

## Introduction

Analogical reasoning is a central feature of human cognition (Gentner & Holyoak, 1997; Hofstadter, 2001). It is defined as the transfer of a structured set of relations from a source domain to a target domain from which it is more or less distant. A most classical example is the A:B::C:D analogy (e.g., dog:doghouse::bird: ? solution "Nest", in which the "lives in" relation must be abstracted). In other analogy problems, a solution to a source problem can be used to solve a target problem (e.g. Holyoak et al. 1984).

Many experiments have been devoted to the study of ontogenetic changes in the ability of analogical reasoning (Chen, Sanchez, & Campbell, 1997; Gentner, 1988; Goswami & Brown, 1990; Holyoak, Junn, & Billman, 1984; Richland, Morrison, & Holyoak, 2006; Thibaut, French, & Vezneva, 2010a). Data suggest that analogical reasoning can be found present as early as 10 months in very simple experimental settings (e.g., Chen et al, 1997). Children's analogical reasoning capacities improve as their knowledge of the involved relations, or their abilities to resist irrelevant information increase. Several models have been proposed in order to explain these changes. They fall roughly into two subclasses: models that try to explain development of analogical reasoning by the increase of structured knowledge about the world (Goswami, 1992) and models that suggest that the key lies with the maturation of control processes, such as working memory or executive functions (Halford, Wilson, & Phillips, 1998; Richland et al., 2006).

Richland et al. (2006) and Thibaut and colleagues (Thibaut, French, & Vezneva, 2009, 2010a, 2010b; Thibaut, French, Vezneva, Gérard, & Glady, 2011) posited that while knowledge of relations is necessary to do analogy making, executive functions are also involved in solving analogical problems. Thibaut et al. interpreted their results as showing that younger children's difficulties with analogy making arose because of insufficiently developed executive functions, specifically inhibition. In one experiment involving semantic A:B::C: ? analogies with four possible responses Thibaut, French, and Vezneva, (2010b) compared weak and strong analogies (i.e., analogies in which the items of the A:B and C:D pairs were weakly, or strongly, associated). Results revealed poorer results in weak (e.g., shirt:suitcase::toy:box) analogies than in strong ones, especially when the number of distractor items was high (i.e., three vs. one). Importantly, the authors controlled to ensure that the children knew the semantic relations within the pair (i.e., the semantic relations between A and B, and between C and D). Thus, children's failure to map the A:B pair on the potential C:D target pair could not be explained by a lack of knowledge. They showed that a greater number of distractors led to poorer performance in the case of weak analogies. They suggested that for strongly associated A:B and C:D item pairs, children were not interfered with by the semantic distractors. In contrast, when the problem involved weakly associated items, mapping the A:B pair onto the C:D pair requires more than simply accessing the obvious semantic dimensions of the items.

The authors characterized analogy-making as a search through a space of features and potential relations. The number of relations holding between any A:B pair is potentially large because, depending on the context, any number of different relations might be relevant (Chalmers, French, & Hofstadter, 1992; French, 1995; Hofstadter, 1995; Mitchell, 1993; Murphy & Medin, 1985; Thibaut, 1997). As mentioned above, the structure of the search space and the presence or absence of competing non-analogical solutions have an effect on the search, especially for young children, who have greater difficulty handling the cognitive load associated with a more elaborate search of the space of possible solutions.

The notion of "searching in a semantic space" was directly investigated in an eye-tracking study by Thibaut,

French, Missault, Gérard, and Glady (2011; Thibaut & French, submitted). The authors started with the idea that the search space in an analogy task is dynamically created as the result of comparisons between the items that compose the analogy problem and this requires the integration of the various sources of information that are available during the task. They used an eye-tracker because cognitive monitoring is difficult to assess with the sole performance measures (i.e., error measures and reaction times) that are usually used in the literature (e.g., Rattermann & Gentner, 1998; Richland et al., 2006; Thibaut et al., 2010b). Eye-tracking allowed them to study precisely how the space of potential solutions was explored by both children and adults. The idea was to study what parts of the space were explored and exactly when that exploration took place. By manipulating various aspects of analogical problems of the A:B::C:D type, eye-tracking allowed them to probe the factors affecting the search of solution space.

Compared with adults, children obtained poorer results. There were also key differences between adults and children in the temporal organization of their respective search profiles. First, adults focused on the A and B pair at the beginning of the trial, paying less or no attention to C and to stimuli in the solution set. Later they focused on C and the Target, which they compared with the semantically related distractor. At the end of the trial, the Target was their sole focus of attention. By contrast, children organized their search around C on which they actively focused during the entire trial. At the very beginning of the trial they paid more attention to C and B. They began looking at the Target and the semantic distractor earlier than in the adults' case. Thus the main differences between children and adults were that children focused on B and C at the beginning of a trial, compared to A and B for adults, and that the Target and the semantic distractor were focused on earlier by children than by adults. The comparison between error trials and correct trials in the case of children revealed that errors were characterized by longer looking times on C and shorter looking times on A. Overall, the results showed that children organized their search around C and paid less attention to A and B when necessary.

This pattern of results suggests that one reason children might fail in analogy-making tasks is that they do not pay sufficient attention to A and B or do not include them in their search. Recall that the task explicitly requires "finding the item that goes with C". Thus, in order to successfully comply with the task, children have to focus on stimuli other than the ones which are highlighted by the instructions, i.e. the C item and the set of distractors. Specifically, they have to study A and B and integrate information from these items in their search for the "one that goes with C". The executive function framework predicts that children might find it hard to inhibit the search-for-the-one-that-goes-with-C goal in order, first, to study A and B, and, second to compare what they have discovered for this pair and to integrate it in their search for the Target item that goes with C.

This analysis led us to the central prediction of the present paper. We started with the general hypothesis that young children find it hard to follow the instructions, that is, to integrate A and B in their exploration of C and the solution set. In this context, if the way the analogy task is implemented forces them to study and interpret the A:B pair, then they should obtain better results than in the classical situation in which all the stimuli are introduced simultaneously.

Thus, in the present experiment, we compared two conditions, i.e., the Standard condition and an A:B-first condition. In the latter condition, children first saw the A:B pair alone and were asked to describe the relation holding between A and B before they were shown C and the solution set. We hypothesized that the A:B-first condition would force children to focus on this pair which would help them to integrate it in their search for the correct C:Target pair.

## Experiment

The present study more directly tested the influence of A:B in children's analogy making. The reasoning was as follows. If children do not pay enough attention to A:B while making analogies, they should obtain better results with procedures requiring a preliminary treatment and interpretation of the A:B pair. Children were, first, presented the A:B pair alone. Then, they had to study it and explain the semantic relation holding between A and B, *before* they were presented with the other pictures. We predicted that, in this condition, children would have higher scores than children that would see all the stimuli composing a problem simultaneously. Indeed, as suggested by Thibaut, et al., (2011), young children have difficulties not looking at C and the solution set rather than at A and B. In a similar vein, Thibaut and French (submitted) showed that a distinctive feature of errors, compared to correct trials, is an imbalance between A and C in favor of C.

## Methods

### Participants

Subjects were 42 5-year-old preschool children (M = 67.1 months; range, 57 to 77 months). Their participation to the experiment was submitted to informed consent of their parents.

The subjects were equally divided into two groups: Standard Analogies group (N = 21; M = 67.4 months; range, 56-75 months) and A:B-first group (N = 21; M = 66.8 months; range, 59-77 months).

### Materials

The experiment consisted of 14 trials, with 2 training trials and 12 experimental trials (See Table 1 for the list of trials). Analogies were of the A:B::C:? format composed of 7 items (black and white drawings; see Figure 1). The problem consisted of the A:B pair (the source), the C item (the target), and an empty square. The solution set was

composed of four stimuli: the analogical answer, a distractor that was semantically related to the C item, and 2 items that were not semantically related to C. Positions of the different alternatives were counterbalanced.

The trials were presented to the children on a touch screen controlled by an E-Prime® program used to run the experiment.

### Procedure

Children were individually tested in their school, in a quiet room.

First, participants' knowledge of the stimuli used in the experiment was assessed. Each stimulus was introduced alone and participants were asked to name it or, when they did not know its name, to describe its function or a context in which it could be found. Children recognized 98% of the items correctly. The analogy task followed.

The Standard Analogies group was shown all 7 items defining a problem simultaneously. In the first practice trial, the task was explained to children belonging to the Standard Analogies group as follows: "Let me explain how it works. At first, you have to find why these two pictures [showing A and B] go well together. So, why do you think [A] goes with

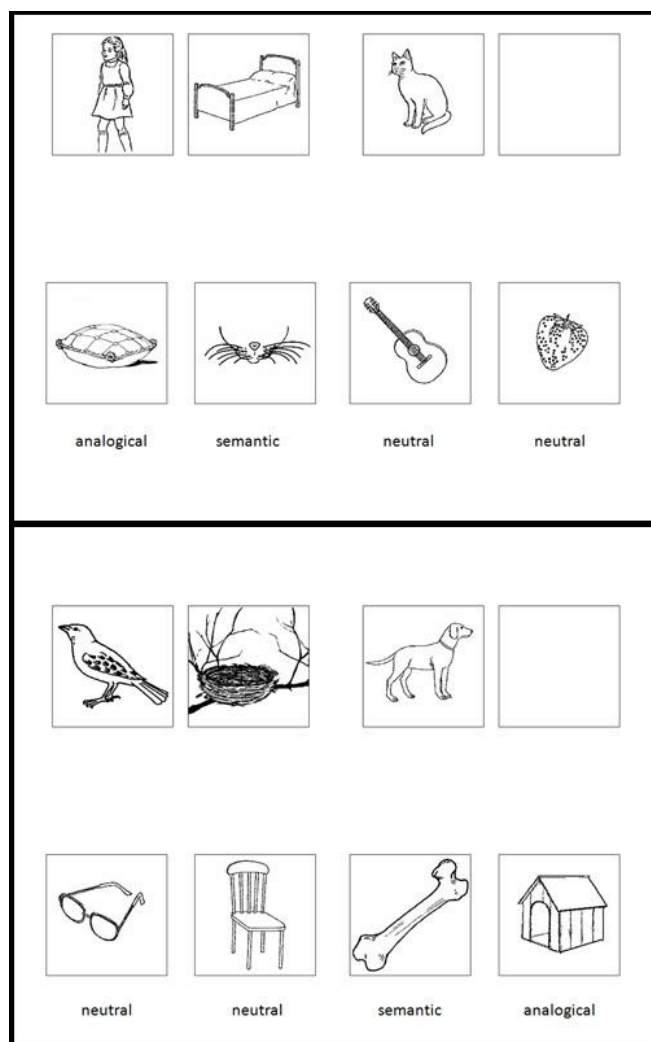
[B]? OK! You see this one [showing C]? It is alone. What you have to do is to find one picture in these four images [showing the four answer options] that goes well with this one [C] in the same way as this one [B] goes with [A] so the two pairs of pictures go together for the same reason. Which picture goes up there [showing the empty slot] with [C] like [B] with [A]? The child gave an answer and justified her choice. Then, the experimenter rephrased the entire trial, explaining and emphasizing why "A and B" and "C and D" go together for the same reason. During the second practice trial, they were asked to do the same. When children did not attend to the A:B pair while explaining their choice, they were asked to do so, and care was taken to ensure that they understood the instructions during the training trials. In the experimental phase, they were asked to do the same thing that was explained to them during the experiment trials and to justify their answer afterward. No feedback was given for the experimental trials.

The A:B-first group was first shown the A:B pair alone and was asked to describe the semantic relation holding between the two drawings: "Why do these two things go together". Once they had given the relation, the

A	B	C	D (Target)	Semantic Distractor	Relation
<b>Practice trials</b>					
Wolf	Meat	Goat	Grass	Horns	<i>Eat</i>
Child	Foot	Elephant	Paw	Giraffe	<i>Part of</i>
<b>Experiment trials</b>					
Shirt	Suitcase	Toy car	Box	Gas pump	<i>Put in</i>
Child	Bed	Cat	Pillow	Whiskers	<i>Sleep on</i>
Pig	Dish	Man	Plate	Watch	<i>Eat in</i>
Man	Nose	Stag	Muzzle	Owl	<i>Breathe with</i>
Glass	Sideboard	Ring	Case	Watch	<i>Put in</i>
Pineapple	Bottle	Orange	Carafe	Strawberry	<i>Put juice in</i>
Train	Rails	Boat	Sea	Crab	<i>Move on</i>
Glove	Hand	Shoe	Foot	Footprints	<i>Put on</i>
Lamp	Socket	Remote control	Battery	Radio	<i>Work with</i>
Bird	Nest	Dog	Doghouse	Bone	<i>Live in</i>
Spider	Cobweb	Bee	Beehive	Flower	<i>Live in</i>
Lock	Key	Bottle	Corkscrew	Glass	<i>Open</i>

Table 1: List of stimuli and relations used to build the analogies of the experiment

experimenter displayed the full set of 7 stimuli defining the problem and asked them to complete the second pair as for the other group. Apart from this preliminary question for the A:B pair, the two practice trials were framed in the same way as in the Standard Analogies group. In other words, after they had mentioned the relationship holding between A and B they were shown the set of stimuli defining a trial and the same instructions as in the Standard Analogies group were given.



**Figure 1:** Two examples of analogies used in the experiment. Analogical: Analogical answer; Semantic: Distractor related to the C item; Neutral: unrelated picture

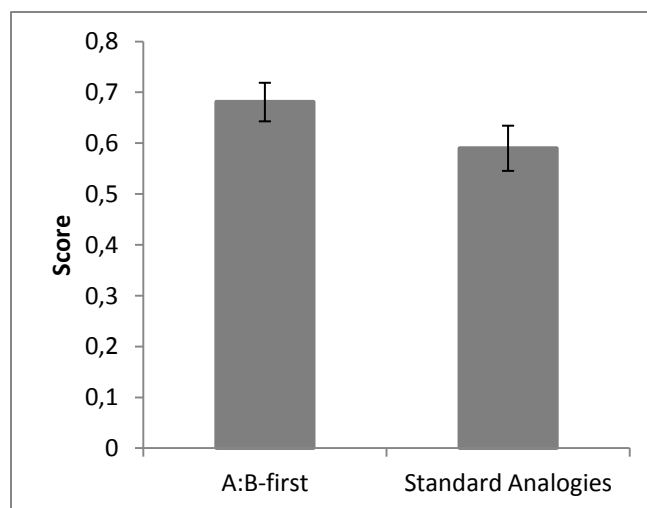
Afterwards, children's understanding of the semantic relation between A and B and between C and D was assessed. They were shown the A:B pairs and were asked *why* the two items of each pair went together. The same was true for the C:D pairs (see Thibaut et al., 2011, for more details).

## Results

We first removed all the trials in which children could not identify one of the semantic relations, either A:B or C:D. As a result, 3% of the trials were removed from subsequent analysis. Note also that in most trials (90% of the cases), children found the target relation that was intended by the experimenter.

We ran a one-way ANOVA on the scores defined as the proportion of correct answers with Condition (Standard Analogies vs. A:B-first) as a between-subject factor.

There was a main effect of condition,  $F(1, 40) = 6.02$ ,  $p < .05$ ,  $\eta^2 = .13$ , with better scores in the A:B-first condition (mean score = .68; see Figure 2) than in the Standard Analogies condition (mean score = .58). These results confirmed our hypothesis that processing the A:B pair first could help children in their search for the analogical answer.



**Figure 2:** Scores of the 5-year-olds in the analogical reasoning test in the two conditions;  $p < .05$ .

When children did not select the analogical match, in 84.5% of the cases, they selected the distractor that was semantically related to C. This result differs significantly from chance (25% of selection, one-sample t test;  $t(41) = 21.31$ ,  $p < .001$ ).

## General discussion

The main purpose of the experiment presented in this paper was to test whether young children's difficulties in analogy-making might result from their difficulties to integrate the A:B pair in the analogy problem. In this study, we conjectured that one source of children's difficulty lies in their search strategy for the task. We suggest that this strategy is, at least in part, induced by the instructions which

require them to find “the item that goes with C.” In the experiment, we directly tested our hypothesis in a condition that required children to first interpret the A:B pair. It was compared with the classical analogy problems. The results confirmed our hypothesis, since children were better in the A:B-first condition.

The experiment is consistent with the idea that children spend less time than adults studying the A:B pair. These data are consistent with Thibaut et al. (2011) eye-tracking data (see also Thibaut & French, submitted) showing that children spend less time on A and B, compared to C, had fewer A:B transitions than adults. The experiment forced them to do what adults do spontaneously and, i.e., inducing the sequential A:B then C:D strategy, which gave rise to higher scores than in the “classical” simultaneous presentation.

Lovett et al. (2009) proposed a two-stage computational model of geometrical A:B::C:? task solving. The program’s performances fitted well with adults performances on Evans’ geometrical problems (Evans, 1968), predicting the different patterns of human answers on each item of the task. This program may also well model children’s pattern of answer observed in this study by modifying some of its processes, like allowing only a shallow first-stage A:B relational description that may result from the lack of treatment of this pair observed in children and/or not allowing the executive to induce another description of the A:B pair.

The increased performance in the A:B-first condition (Thibaut, French, Missault, et al., 2011) are entirely compatible with the executive function view. Given that the instructions prompt them to find a partner for C in the set of solutions, they might find it hard to inhibit the set of stimuli which were explicitly mentioned in the instructions. Another, related interpretation, could involve the representation and maintenance of the sub-goals of the task. This has been suggested for other tasks assessing executive functions (Blaye & Chevalier, 2011; Gruber & Goschke, 2004). Children may represent the main goal of the task, which is to find a picture that is related to the C item, but may have difficulty departing from this goal to achieve a crucial sub-goal – namely, finding which relation has to be used to find the correct answer between the different options related to the C item (analogical answer and distractor). Studying and verbalizing the relation linking the A:B pair may contribute to enhance this sub-goal. In this format, they should not have to generate this sub-goal by themselves. Another interpretation would be that children lack the correct strategy which is to look at the A-B pair first. In this context, our “A:B first” condition provided them with the correct strategy for performing the task. In other words, children would not know how to perform the task or to organize it in order to perform it correctly. This is a plausible hypothesis. However, it is difficult to disentangle what is due to inhibition and/or flexibility mechanisms from what results from an explicit strategy.

The studies in the literature have pointed out two main explanations of children’s failures to do analogies correctly. The first is the role of knowledge (e.g., Gentner, 1988; Goswami & Brown, 1990). The second is related to executive functions. It has been shown that children might have difficulties handling all the information available in the task, such as distractors related to C (see Richland et al., 2006; Thibaut et al., 2010a, b for discussions). The present research demonstrates that the task itself has cognitive constraints which generate a cognitive load that must be coped with by young children. In other words, for adults and most likely children older than 9, the comparison between A:B and C and the potential candidates for a solution is automatically driven by the task instructions (the so-called mapping process). By contrast, for children, temporally leaving aside the instructions “looking for the one that goes with C” in order to compare A with B, generates cognitive load. One might conceive of this as a necessity to temporarily inhibit C and the solution set, or as a necessity to be cognitively flexible, that is to be able to conceive the task under different perspectives (i.e., from an A-B perspective or from a C-solution set perspective and integrate these two perspectives). In sum, the present research has made it clear that the analogy task generates its own demands that cannot be taken for granted, in the case of children.

## Acknowledgement

This research has been supported by a French ANR Grant for the “ANAFONEX” project ANR-10-BLAN-1908-01.

## References

- Blaye, A., & Chevalier, N. (2011). The role of goal representation in preschoolers’ flexibility and inhibition. *Journal of experimental child psychology*, 108(3), 469-83.
- Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, 4(3), 185-211.
- Chen, Z., Sanchez, R. P., & Campbell, T. (1997). From beyond to within their grasp: The rudiments of analogical problem solving in 10- and 13-month-olds. *Developmental Psychology*, 33(5), 790-801.
- Evans, T. (1968). A program for the solution of geometric-analogy intelligence test questions. In M. Minsky (Ed.), *Semantic information processing* (pp. 271-353). Cambridge, MA: MIT Press.
- Forbus, K. D., Usher, J., & Lovett, A. (2008). CogSketch: Open-domain sketch understanding for cognitive science research and for education. *Proceedings of the fifth eurographics workshop on sketch-based interfaces and modeling*.
- French, R. M. (1995). *The Subtlety of Sameness: A Theory and Computer Model of Analogy-Making*. Cambridge, MA: The MIT Press.

- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155–170.
- Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, 59(1), 47-59.
- Gentner, D., & Holyoak, K. J. (1997). Reasoning and learning by analogy. *American Psychologist*, 52(1), 32-4.
- Goswami, U. (1992). *Analogical Reasoning in Children*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Goswami, U., & Brown, A. L. (1990). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition*, 35(1), 69–95.
- Gruber, O., & Goschke, T. (2004). Executive control emerging from dynamic interactions between brain systems mediating language, working memory and attentional processes. *Acta Psychologica*, 115(2-3), 105-21.
- Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. *The Behavioral and Brain Sciences*, 21(6), 803-64.
- Hofstadter, D. R. (1995). *Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. New York, NY: Basic Books.
- Hofstadter, D. R. (2001). Epilogue: Analogy as the core of cognition. In D. Gentner, K. J. Holyoak, & B. Kokinov (Eds.), *The Analogical Mind: Perspectives from Cognitive Science*. Cambridge, MA: The MIT Press.
- Holyoak, K. J., Junn, E. N., & Billman, D. O. (1984). Development of analogical problem-solving skill. *Child Development*, 55(6), 2042–2055.
- Lovett, A., Tomai, E., Forbus, K. D., & Usher, J. (2009). Solving geometric analogy problems through two-stage analogical mapping. *Cognitive science*, 33(7), 1192-231.
- Mitchell, M. (1993). *Analogy-Making as Perception*. Cambridge, MA: The MIT Press.
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92(3), 289-316.
- Richland, L. E., Morrison, R., & Holyoak, K. J. (2006). Children's Development of Analogical Reasoning: Insights from Scene Analogy Problems. *Journal of Experimental Child Psychology*, 94(3), 249-273.
- Thibaut, J.-P. (1997). Similarité et catégorisation. *L'Année Psychologique*, 97, 701-736.
- Thibaut, J.-P., French, R. M., Missault, A., Gérard, Y., & Glady, Y. (2011). In the Eyes of the Beholder: What Eye-Tracking Reveals About Analogy-Making Strategies in Children and Adults. *Proceedings of the Thirty-third Annual Meeting of the Cognitive Science Society* (pp. 453–458).
- Thibaut, J.-P., French, R. M., & Vezneva, M. (2009). Cognitive Load and Analogy-making in Children : Explaining an Unexpected Interaction. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the Thirty-first Annual Cognitive Science Society Conference* (pp. 1048-1053).
- Thibaut, J.-P., French, R. M., & Vezneva, M. (2010a). The development of analogy making in children: cognitive load and executive functions. *Journal of Experimental Child Psychology*, 106(1), 1-19.
- Thibaut, J.-P., French, R. M., & Vezneva, M. (2010b). Cognitive load and semantic analogies: Searching semantic space. *Psychonomic Bulletin & Review*, 17(4), 569-74.
- Thibaut, J.-P., French, R. M., Vezneva, M., Gérard, Y., & Glady, Y. (2011). Semantic analogies by young children: testing the role of inhibition. In B. Kokinov, A. Karmiloff-Smith, & N. J. Nersessian (Eds.), *European Perspectives on Cognitive Science*. New Bulgarian University Press.



# Knowledge and implicature: Modeling language understanding as social cognition

Noah D. Goodman (ngoodman@stanford.edu),

Department of Psychology, Stanford University

Andreas Stuhlmüller (ast@mit.edu),

Department of Brain and Cognitive Sciences, MIT

## Abstract

Is language understanding a special case of social cognition? To help evaluate this view, we can formalize it as the *rational speech-act* theory: listeners assume that speakers choose their utterances approximately optimally, and listeners interpret an utterance by using Bayesian inference to “invert” this model of the speaker. We apply this framework to model scalar implicature (“some” implies “not all”, and “N” implies “not more than N”). This model predicts an interaction between the speaker’s knowledge state and the listener’s interpretation. We test these predictions in two experiments, and find good fit between model predictions and human judgements. **Keywords:** Language; Bayesian model; Scalar implicature

To what extent does language understanding rely on extralinguistic knowledge and processes? One view of language processing suggests that it consists of largely separate, special-purpose faculties; another view, that it depends critically on domain general inference mechanisms, and even on intuitive theories that are not specific to language. Indeed, many thinkers have viewed speech as an action with communicative goals, such as informing a listener (Grice, 1975; Levinson, 2000; Clark, 1996). A listener making this assumption can make stronger inferences than an utterance would allow from its literal meaning—pragmatic effects can strengthen, or change, the interpreted meaning.

Recent work has aimed to formalize the social inference view of pragmatics using tools of Bayesian statistics and information theory (Frank & Goodman, under review); we refer to this formal framework as the *rational speech-act* theory of language understanding. It views pragmatic competence as following naturally from an intuitive theory of speech production, which in turn is a special case of intuitive theory of mind. More precisely, listeners have an internal model which describes speakers as choosing their utterances approximately optimally on the basis of certain social goals, such as conveying information to the listener; listeners then interpret an utterance by using Bayesian inference to “invert” this model of the speaker, drawing conclusions about the world state and speaker’s intention from the utterance and any other relevant world knowledge. These two rationality assumptions, for listener and speaker, have a role similar to those made in ideal observer models of perception (Geisler, 2003): they provide a starting point for a quantitative understanding of the complex interactions involved in language understanding. Indeed, this view has provided good quantitative models for pragmatic inference in a number of simple settings (Frank, Goodman, Lai, & Tenenbaum, 2009; Frank & Goodman, under review).

Because rational decision making predicts that action selection is related to the *expected* utility—a quantity that de-

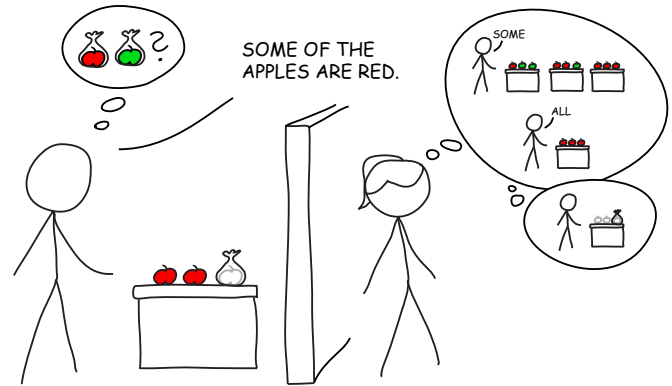


Figure 1: How will the listener interpret the speaker’s utterance? How will this change if she knows that he can see only two of the objects?

pends on the actor’s belief distribution—a listener who views the speaker as rational should be sensitive to the speaker’s belief state. The rational speech-act theory thus predicts an interaction between (shared) knowledge about a speaker’s knowledge state and a listener’s interpretation of his utterance. This is a very general prediction of the framework, which could easily prove to be false—if pragmatic inference is a highly modularized computation, for instance, we would not expect such general knowledge to affect it. Deriving and testing precise predictions about this interaction thus provides an important test of the rational speech-act theory.

If you hear “some of the apples are red,” you will infer that not *all* of the apples are. Pragmatic effects of this sort are called scalar implicatures (Horn, 2004), and they provide a window on the interactions between the language faculty and general cognition. Consider Fig. 1: If the speaker has seen all the apples, his utterance would be interpreted as “some, but not all, of the apples are red.” However, if the speaker had only looked at two of the apples, the listener might draw a different conclusion. Indeed, we show below that the implicature “not all” can be canceled by facts about the speaker’s perceptual access. This interaction between language understanding and general knowledge is not predicted by strongly modular theories of scalar implicature, and even moderately modular theories (Chierchia, Fox, & Spector, 2011) predict that such influences can only turn off scalar implicatures. We show that the interaction of knowledge and implicature is more fine grained: the details of a speaker’s belief distribu-

tion affect the details of an implicature.

This paper provides a formal model of the pragmatic inference that leads to scalar implicatures, building on the rational speech-act framework. We directly model the possibility that the speaker may have incomplete knowledge and the effects this has on the listener’s interpretation. We derive predictions of this model for interpretation of the quantifier “some” and the number words (“one”, “two”, etc.). The model both explains the standard implicatures for these words, and predicts that these implicatures can be canceled, completely or partially, when the speaker has incomplete knowledge. We test these predictions in two experiments and find good fits to human judgements, both qualitatively and quantitatively.

### A rational speech-act model

We view language comprehension as a rational inference based on an intuitive theory of language production. Our setting is illustrated in Fig. 1. The listener infers the world state,  $s$ , given the speaker’s utterance,  $w$ , and shared information about the speaker’s (possibly incomplete) information access,  $a$ . By Bayes’ rule:

$$P_{\text{listener}}(s|w, a) \propto P_{\text{speaker}}(w|s, a)P(s). \quad (1)$$

Where  $P(s)$  captures the listener’s prior beliefs about the world state, and  $P_{\text{speaker}}(w|s, a)$  describes the listener’s intuitive theory of how the speaker chooses words.

The speaker chooses an utterance in accord with Bayesian decision theory (Berger, 1985): she acts to approximately optimize expected utility. Imagine a speaker who makes observation  $o$  about the true state of the world. (For instance, in Fig. 1, the speaker has perceptual access to two of three apples, and observes that those two are red). The speaker selects an utterance  $w$  to convey information about the world state to a listener, and does so by soft-max optimizing expected utility:

$$P_{\text{speaker}}(w|o, a) \propto \exp(\alpha \mathbb{E}_{P(s|o, a)}[U(w; s)]) \quad (2)$$

The speaker’s utility function,  $U(w; s)$ , captures the value of saying  $w$  if the world is actually  $s$ . The expectation is taken over the speaker’s belief state,  $P(s|o, a)$ , because the speaker may still be uncertain about the state of the world. The parameter  $\alpha$  controls the deviation from optimality.

So far, nothing in the model is unique to language—indeed, similar models have been used to model social cognition more generally (Baker, Saxe, & Tenenbaum, 2009; Goodman, Baker, & Tenenbaum, 2009; Ullman et al., 2009). To capture a motivation to be informative, utility must be related to the information conveyed in the utterance.<sup>1</sup> More specifically, utility is related to the amount of information that a *literal* listener would not yet know about state  $s$  after hearing it described by utterance  $w$ —this is the negative *surprisal* (Cover & Thomas, 1991):

$$U(w; s) = \ln(P_{\text{lex}}(s|w)), \quad (3)$$

<sup>1</sup>Other communicative motivations could be added to this utility, such as a complexity term influencing the manner of expression.

where the *literal interpretation* probability  $P_{\text{lex}}(s|w)$  is determined by the lexicon—here we will assume that each utterance has a truth function,  $F_w : s \mapsto \{0, 1\}$ , and the distribution is otherwise uninformative:  $P_{\text{lex}}(s|w) \propto \delta_{F_w(s)}$ .

We assume that the speaker’s access  $a$  is common knowledge of speaker and listener, but the listener still does not know what observation the speaker made, hence:

$$P_{\text{speaker}}(w|s, a) = \sum_o P_{\text{speaker}}(w|o, a)P(o|a, s) \quad (4)$$

In the experiments below (as in Fig. 1), the state of the world is always a set of objects that may have a given property and observations consist of looking at a subset of the objects. Thus the observation probability  $P(o|a, s)$  is a hypergeometric distribution (i.e. the probability of drawing  $N$  balls of a given color, without replacement, from an urn containing a given set of colored balls). In this setting, it is also reasonable to assume that the prior probability,  $P(s)$ , is a binomial distribution (i.e. draws *with* replacement); we will initially assume so, and will later measure the distribution empirically.

Overall, the above equations describe the inferences that a rational listener will make to comprehend a speaker that she believes to be approximately rational and have a goal to be informative. Importantly, these inferences depend on shared knowledge about the aspects of the world to which the speaker has access. Thus the rational speech-act theory predicts that the speaker’s access affects utterance interpretation; we test this prediction below. To derive more specific predictions we must describe the set of alternative utterances.

### The alternatives

We have assumed that the interpretation of an utterance is made with respect to a set of alternative utterances. These alternatives could be all possible utterances, or could be a limited set generated by replacing key words in the actual utterance with related words. The alternatives may, for instance, be generated by a grammatical mechanism as in Fox and Katzir (2011). For our results the details of this process are unimportant; what is important is that there exists a set of alternative expressions and a (truth-functional) literal meaning for each. We make standard assumptions in both respects.

Consider the case of the quantifier words “some” and “all”. Under the standard semantics, “all the balls are red” is true exactly when  $N$  of the  $N$  balls are red, while “some of the balls are red” is true when  $M \geq 1$  of the  $N$  balls are red. In particular, the literal meaning of “some” allows the state where all  $N$  balls are red. We use a lexicon that consists of the standard meanings for “none”, “some”, and “all”. Model predictions are shown in Fig. 2a for the listener’s interpretation of “some of the balls are red” when there are 3 objects, under varying conditions of speaker access.

When the speaker has perceptual access to 3 of the 3 objects (hence complete knowledge) and says “some”, there is a lower probability on 3 than 2—this is the standard “some but not all” implicature. To understand why the model predicts this implicature, we can first simplify the speaker model: in

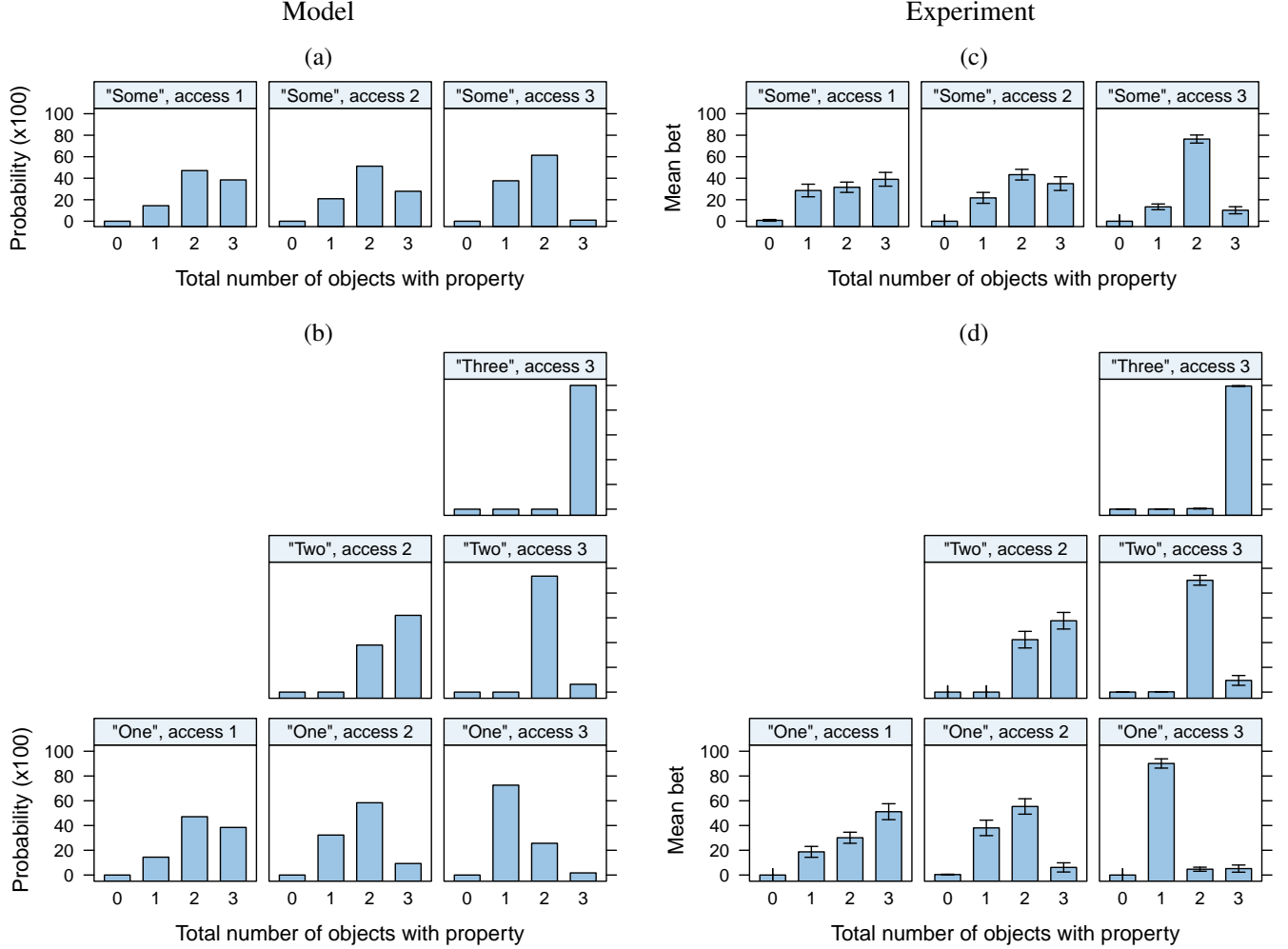


Figure 2: (a,b) Model prediction for probability of each world state (number of objects with property), varying the word the speaker used and the speaker’s perceptual access. The prior is assumed to be binomial with base rate 0.62, and the speaker optimality parameter is set to  $\alpha = 3.4$ . (c,d) Mean participant bet on each world state, varying the word the speaker used and the speaker’s perceptual access. Data have been filtered to include only trials where the participant’s bet that the speaker had complete knowledge was greater than 70 in the expected direction. Error bars are standard error of the mean.

a complete knowledge situation the speaker’s belief distribution is concentrated on the true state, so Eqs. 2, 3, and 4 become:

$$P_{speaker}(w|s,a) \propto \exp(\alpha \ln(P_{lex}(s|w))) \propto (P_{lex}(s|w))^\alpha. \quad (5)$$

Using the standard meanings described above, we see that if the state were 3 the speaker would be more likely to say “all” ( $P_{lex}(s|w)=1$ ) than “some” ( $P_{lex}(s|w)=\frac{1}{3}$ ); conversely, if the state were 2, the speaker would say “some”, since the probability for “all” is now  $P_{lex}(s|w)=0$ . Now consider Eq. 1 and imagine for the moment a uniform prior over states. In this case the listener will infer each state  $s$  in proportion to how likely the speaker was to say “some” given this state. Overall, this leads to the implicature that “some” is unlikely to be interpreted as 3—“some but not all”.

In contrast, when the speaker has only partial access the calculation is more complex, involving the inferred belief distribution of the speaker. Comparing across the three panels of Fig. 2a, we see the probability of 3 is much higher when access is 1 or 2 than when it is 3. When access is 1, no implicature is predicted (the probability of 3 is approximately the same as the probability of 2); when access is 2, only a very slight implicature. Overall, we predict that incomplete speaker knowledge can *cancel* the standard “some but not all” implicature.

The case of numerals (“one”, “two”, ...) is similar but more subtle. It has been argued that number words have a lower-bound meaning (Horn, 1972) (e.g. “two balls are red” means  $M \geq 2$  of the balls are red), and the intuitive, exact, meanings arise as a pragmatic implicature—“one but not two, three, etc.”. In Fig. 2b we show model predictions based on the

lower-bound semantics for number words, varying speaker's access. We see that exact meanings do arise as an inference when the speaker has complete access, but there is an interaction: number words do not receive an exact interpretation when the speaker has incomplete knowledge. Of particular interest is the case where the speaker has seen 2 out of 3 objects and says "one": here a *partial* implicature is predicted, with the probability of 3 low, but 1 and 2 high.

## Experiment 1

Because a rational speaker chooses actions based on *expected* utility, the rational speech-act model predicts an effect of speaker's knowledge on the listener's interpretation of "some" statements. We tested the predictions of the model by putting participants in the role of the listener and asking them to judge the state of the world in scenarios where perceptual access (and hence knowledgeability) of the speaker varied.

### Participants, Materials, and Methods

50 participants were recruited through Amazon's Mechanical Turk crowd-sourcing service and completed the experiment for a small payment.

We constructed six scenarios in which a speaker had three objects that could have (or not) a given property. The speaker then made a statement indicating the number of objects they had looked at and a quantified ("some") statement. We split each scenario into setup and speech act phases. The setup phase named the speaker and described the objects and the relevant property. For example:

Letters to Laura's company almost always have checks inside. Today Laura received 3 letters.

Because our model predicts greater effects when the *a priori* base rate of the property is high (otherwise it is difficult to tell an implicature from an *a priori* belief that it is unlikely for all objects to have the property), we describe all properties as "almost always" holding. To make sure participants attended to the setup, we asked them to report the *a priori* probability that 0, 1, 2, or 3 objects had the property:

How many of the 3 letters do you think have checks inside?

The speech-act phase introduced a speech act in which the speaker both declared how many of the objects they had observed and stated that some objects had the property:

Laura tells you on the phone: "I have looked at 2 of the 3 letters. Some of the letters have checks inside."

We then again elicited judgements about how many objects had the property:

Now how many of the 3 letters do you think have checks inside?

Finally, because the speech act might not be a perfect manipulation of speaker's knowledgeability (for instance, the speaker

may have gained knowledge by another route), we elicited this directly:

Do you think Laura knows exactly how many of the 3 letters have checks inside?

Each response was given by a betting measure: participants were instructed to divide "\$100" among the options, betting to indicate their confidence in each option. For the first two questions there were four options (0–3 of the objects have the property) and for the final question there were two options (the speaker does/doesn't have complete knowledge). Before the experiment began, participants were given a brief warm-up, using unrelated questions, to familiarize them with the betting measure.

Each scenario existed in forms varying speaker access (the number of objects the speakers had looked at) from 1 to 3. Each participant saw each access condition once, in random order, presented using randomly chosen scenarios (with no duplicate scenarios). In terms of our predictions, we have two partial-knowledge conditions (where we expect cancellation of the implicature) and one complete-knowledge "control" (where we expect the standard implicature).<sup>2</sup>

### Qualitative Results

There was no effect of scenario, so we collapse across this factor in all analyses. As expected based related work (Goodman et al., 2009), the speaker's access statement (e.g. "I have looked at 2 of the 3 letters") was not a perfect manipulation of knowledgeability: in the partial access conditions some participants judged that the speaker was likely to know exactly how many objects had the property. (The bet that the speaker had complete knowledge when access=1 was  $M = 27.1$ ,  $SD = 4.9$ ; when access=2,  $M = 34.8$ ,  $SD = 5.7$ ; when access=3,  $M = 93.0$ ,  $SD = 2.7$ .) Since we were interested in the effects of varying knowledgeability, we initially exclude trials in which the knowledge judgement was less than 70 in the expected direction (we come back to the full data set in the quantitative analysis below). Fig. 2c shows the mean of bets on each option, as access varied. As predicted, there was an effect of speaker's access on listener's interpretation (one-way ANOVA with bets on 3 as dependent variable,  $F(2, 102) = 10.18$ ,  $p < 0.001$ ).

We next performed our pre-planned comparisons to check that an implicature was drawn when the speaker had complete knowledge, but not when the speaker had partial knowledge. In the complete access condition bets on 3 were less than bets on 2 (paired, directional t-test,  $t(43) = -10.2$ ,  $p < 0.001$ ). In the partial access conditions the implicature was canceled: bets on 2 did not exceed bets on 3 when speaker had access to 1 object (paired, directional t-test,  $t(31) = 0.77$ ,  $p = 0.78$ ) or when access was 2 (paired, directional t-test,  $t(28) = -0.82$ ,  $p = 0.21$ ). If we look at just bets on 3, we see significantly lower bets in the complete access condition than the

<sup>2</sup>Expt. 1 may be viewed at <http://goo.gl/3S5zz>, Expt. 2 at <http://goo.gl/iSc6o>.

access 1 condition (unpaired, directional t-test,  $t(47) = -4.0$ ,  $p < 0.001$ ) or the access 2 condition (unpaired, directional t-test,  $t(43) = -3.5$ ,  $p < 0.001$ ). While there was no significant implicature in either partial-access condition, there is a slightly greater tendency toward implicature in the access 2 condition than the access 1 condition, as predicted by the model (two-way ANOVA with bet as dependent variable, and access (1 or 2) and state (2 or 3) as independent variables,  $F(2, 294) = 3.77$ ,  $p < 0.05$ ).

Thus, the knowledge of the speaker affected listener's interpretation of "some" in the way predicted by the rational speech-act model. We examine the quantitative fit of the model below.

## Experiment 2

In Expt. 2 we tested the predictions of the rational speech-act model for interpretation of numerals. We again expect to find an effect of speaker's knowledge, but in this case there is a more detailed interaction: the implicature should be canceled when the speaker says "one" after seeing only one object and when the speaker says "two" after seeing two objects, but it should only be partially canceled when the speaker says "one" after seeing two objects—this implies a fine-grained interplay between the speaker's knowledge state and the interpretation of her utterance.

### Participants, Materials, and Methods

50 participants were recruited through Amazon's Mechanical Turk crowd-sourcing service and completed the experiment for a small payment.

We used the same stimuli as in Expt. 1, modifying the scenarios only in the speech act: the speaker now made a statement indicating the number of objects they had looked at and the number that had the property. For instance:

Laura tells you on the phone: "I have looked at 2 of the 3 letters. 1 of the letters has checks inside."

Each scenario existed in forms varying the speaker's access, from 1 to 3, and the number word the speaker used, from 1 to 3; we limited to sensible situations where the word used was no greater than the number of objects seen. Hence we had six conditions, with access/word: 1/1, 2/1, 2/2, 3/1, 3/2, 3/3. In terms of our predictions, we have three partial-knowledge conditions (where we expect partial or complete cancellation of the implicature) and three complete-knowledge "controls" (where we expect the standard implicatures). The order of scenarios and the order of conditions were randomized between participants.

### Qualitative Results

There was again no effect of scenario, so we collapse across this factor. As in Expt. 1, the speaker's access statement was not a perfect manipulation of knowledgeability (bets that speaker had complete knowledge in partial-access conditions:  $M = 42.0$ ,  $SD = 3.4$ , in complete access conditions:  $M = 92.1$ ,  $SD = 1.6$ ). We once again limit to trials with the expected

judgements of knowledgeability (with a threshold of 70). The mean of participants' bets are shown in Fig. 2d. To evaluate the overall effect of access, we performed an ANOVA with access and word as independent measures and bet on 3 as dependent measure. We find a main effect of access ( $F(2, 205) = 6.57$ ,  $p < 0.01$ ), an interaction between word and access ( $F(1, 205) = 34.7$ ,  $p < 0.001$ ), and a main effect of word ( $F(2, 205) = 269.8$ ,  $p < 0.001$ ).

We then explored the results in more detail using planned comparisons to test whether implicatures were drawn (only) when predicted. We found an implicature in the complete access conditions: when the speaker said "two", bets on state 3 were less than on state 2 (paired, directional t-test,  $t(43) = -10.2$ ,  $p < 0.001$ ). When the speaker said "one", bets on state 1 were greater than on state 3 (paired, directional t-test,  $t(42) = -13.1$ ,  $p < 0.001$ ) or state 2 (paired, directional t-test,  $t(42) = -17.1$ ,  $p < 0.001$ ). In contrast, there was no implicature when access was 1 and the speaker says "one"—bets on 1 were not greater than on 2 (paired, directional t-test,  $t(24) = 1.9$ ,  $p = 0.96$ ) or on 3 (paired, directional t-test,  $t(24) = 3.2$ ,  $p = 1.0$ )—and no implicature when access is 2 and the speaker says "two"—bets on 2 were not greater than on 3 (paired, directional t-test,  $t(24) = 1.1$ ,  $p = 0.87$ ). When access is 2 and the speaker says "one" we found the predicted *partial implicature*: bets on state 1 were significantly greater than on state 3 (paired, directional t-test,  $t(25) = -3.9$ ,  $p < 0.001$ ), but not on state 2 (paired, directional t-test,  $t(25) = 1.5$ ,  $p = 0.92$ ).

These results again support the predictions of the rational speech-act model, showing not merely an interaction between the speaker's knowledge and the listener's tendency to draw an implicature, but a fine-grained interaction that is unlikely to result from a modular process of language understanding. In addition, these results support the standard, but controversial (Huang, Snedeker, & Spelke, 2004; Barner & Bachrach, 2010), view that number words have a "lower-bound" semantics which is only strengthened into an exact meaning by pragmatic inference.

### Quantitative model comparison

To evaluate the quantitative model predictions, we first compare model predictions with mean human ratings for the subset of trials in which participants gave knowledgeability ratings in the expected direction (greater than 70, as above). As in the model description, we assume a binomial prior distribution. We fit the prior base rate parameter and the  $\alpha$  parameter by minimizing the root mean squared error (RMSE) of the model predictions and mean data for both experiments. The resulting model fit is  $RMSE = 9.01$  and  $r = 0.96$ . The model predictions with the best-fitting parameters are shown in Fig. 2a,b and show striking correspondence with the human data in Fig. 2c,d.

To consider all responses, including those previously removed due to unexpected knowledge judgements, we extend the model by including an additional knowledge parameter: the probability that the speaker did in fact have complete

knowledge (regardless of perceptual access). Since we measured prior expectations and knowledgeability in each trial, we compute model predictions, trial by trial, using these values. Because the betting interface encouraged participants to round small bets to 0, but probability 0 is very different than a non-zero small probability to the model, we changed 0 and 100 bets to 1 and 99. In addition, we assumed a simple power-law relationship between subjective probability and bets (equivalent to the soft-max used to model the speaker, Eq. 2). The parameter of this power law and the speaker optimality,  $\alpha$ , were then fit by minimizing the RMSE of mean model predictions to mean human judgements in both experiments. The resulting fit was again good: RMSE=8.36,  $r=0.95$ . Looking at individual participants, we find median correlation of  $r = 0.75$  between a participant's judgements and the model predictions based on their prior and knowledge scores. These results suggest that the rational speech-act model is able to capture the quantitative pattern of people's judgements both across the population and within individual participants.

## Conclusion

We have described a rational speech-act model of scalar implicatures and their interaction with speaker knowledge. This model formalizes language understanding as social cognition, with language-specific goals and actions, using the tools of Bayesian statistics. In addition to predicting the standard implicatures ("some but not all", etc) as an inference that depends on the alternative utterances, this model predicted that these implicatures could be canceled, completely or in part, when it was common knowledge that the speaker had incomplete knowledge. Expts. 1 and 2 verified these qualitative predictions, and showed tight quantitative fits with the model.

The predicted interaction between interpretation and speaker's knowledge was not a peculiarity of this set of words or of scalar lexical items, instead it follows from the general fact that rational decision makers must choose actions based on their expected utility. In contrast, a strongly modular of implicatures would predict no such interactions. One could amend these theories (Chierchia et al., 2011) to allow a speaker's ignorance to gate the implicature mechanism, but this still would not predict the fine-grained interactions demonstrated for number words: the details of a speaker's knowledge state influence a listener's interpretation.

Our results support the rational speech-act framework for modeling pragmatics. More generally, they further boost the momentum building for quantitative models of language as a branch of rational social cognition. In the words of Grice (1975): "One of [our] avowed aims is to see talking as a special case or variety of purposive, indeed rational, behavior..."

## Acknowledgments

We are grateful to Mike Frank, Josh Hartshorne, Danny Fox, and Irene Heim for comments on this work. This work was supported by a John S. McDonnell Foundation Scholar Award and ONR grant N00014-09-1-0124.

## References

- Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113, 329-349.
- Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive psychology*, 60(1), 40-62.
- Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer.
- Chierchia, G., Fox, D., & Spector, B. (2011). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In *Semantics: An international handbook of natural language meaning*. Berlin: Mouton de Gruyter.
- Clark, H. H. (1996). *Using language*. Cambridge University Press Cambridge.
- Cover, T., & Thomas, J. (1991). *Elements of information theory*. Wiley Online Library.
- Fox, D., & Katzir, R. (2011). On the characterization of alternatives. *Natural Language Semantics*.
- Frank, M. C., & Goodman, N. D. (under review). *Quantifying pragmatic inference in language games*. (Under review)
- Frank, M. C., Goodman, N. D., Lai, P., & Tenenbaum, J. B. (2009). Informative communication in word production and word learning. In *Proceedings of the 31st annual conference of the cognitive science society*.
- Geisler, W. S. (2003). Ideal observer analysis. In L. Chalupa & J. Werner (Eds.), *The visual neurosciences* (pp. 825-837). MIT press.
- Goodman, N. D., Baker, C. L., & Tenenbaum, J. B. (2009). Cause and intent: Social reasoning in causal learning. In *Proceedings of the 31st annual conference of the cognitive science society*.
- Grice, H. (1975). Logic and conversation. In *Readings in language and mind*. Blackwell.
- Horn, L. (1972). *On the semantic properties of logical operators in english*. Unpublished doctoral dissertation, UCLA.
- Horn, L. (2004). Implicature. *The handbook of pragmatics*, 2-28.
- Huang, Y., Snedeker, J., & Spelke, E. (2004). What exactly do numbers mean. In *26th annual meeting of the cognitive science society, chicago*.
- Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. The MIT Press.
- Ullman, T., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in neural information processing systems* 22.

# Analogical Problem Solving: Insights from Verbal Reports

**Linn Gralla (gralla@uni-bremen.de)**

Department of Linguistics and Literary Sciences, Universität Bremen, Germany

**Thora Tenbrink (tenbrink@uni-bremen.de)**

Department of Linguistics and Literary Sciences, Universität Bremen, Germany

**Michael Siebers (michael.siebers@uni-bamberg.de)**

Cognitive Systems Group, Universität Bamberg, Germany

**Ute Schmid (ute.schmid@uni-bamberg.de)**

Cognitive Systems Group, Universität Bamberg, Germany

## Abstract

Problem solving is a complex cognitive activity that involves the construction of sequences of actions to reach a given goal. One powerful strategy is to identify analogies between the problem at hand and previously encountered ones. Relevant similarities between problems can be detected more easily if there is a high resemblance on the surface or with respect to structure. Earlier theoretical observations and performance data have pointed to two distinct kinds of analogical reasoning, direct solution transfer (transformational analogy) and the creation of a new solution based on adapted past reasoning processes (derivational analogy). In order to gain insights about the cognitive processes involved, we extend this work by an analysis of different kinds of verbal data. Planning protocols were collected prior to problem solving, and retrospective reports, evaluations, and instructions were elicited after the task was completed. Results show that the different kinds of analogical reasoning involved different degrees of analogy awareness, as reflected by the verbalizations. Derivational analogy involved problem solving on a more detailed and structured step-by-step basis than the more superficial transformational strategy, in which a simple matching procedure was employed.

**Keywords:** analogical problem solving, verbal reports, cognitive discourse analysis

## Introduction

People frequently encounter problem situations in their daily lives. Most of these are instantly solved, as humans are well equipped with numerous problem solving strategies. One powerful strategy is the adaptation of previous experiences to solve the newly encountered problem. Analogical reasoning is not only involved in problem solving but also in a number of other human activities, such as use of metaphor, scientific reasoning, humor, and empathy. Gentner, Holyoak & Kokinov (2001) therefore argue that analogical reasoning is at the core of cognition. Hofstadter (2001) supports this view by arguing that all concepts that are used to understand recurrent and new situations are packages of analogies.

Even though analogical reasoning is assumed to be a

ubiquitous and efficient problem solving strategy, participants seldom apply it in experimental settings, unless the analogous nature is directly salient (Gick & Holyoak, 1980; Schmid, Wirth & Polkehn, 1999). If analogies are used in problem solving the transfer can be on different levels of specificity, leading to different cognitive solution steps and strategies. In this paper, we pursue these issues by an analysis of verbal data collected while solving an analogy problem. In the following, we will first discuss previous work on analogical problem solving, with a brief look at the role of language. Next, we introduce a previous study by Schelhorn, Griego, and Schmid (2007) that served as our starting point. Our current study is then presented and discussed in the remainder of this paper.

## Analogical Problem Solving

Analogies are based on shared relations between base and target problem (Gentner, 1983; Clement & Gentner, 1991). By highlighting shared relational structures, analogies connect domains and problems that may appear only marginally similar on the surface. This process involves structural alignment as a crucial component of analogical reasoning. While similarity centers on shared attributes, analogy concerns the alignment of relational structures at a deeper level. According to the *systematicity principle*, the structural relations are connected by one-to-one correspondences (Gentner & Markman, 1997).

As proposed by Carbonell (1986), analogical problem solving can be performed on different levels of abstraction. *Transformational analogy* is based on direct solution transfer, i.e. the solution to a previous problem is slightly altered in a transformation process to solve the new problem. The solution transfer process contains three basic processes. First, the *initial partial matching process* determines if two problems share similar aspects based on state information and operator sequences. Second, the sequence of actions from the retrieved solution is transferred to the new situation in an *analogical mapping process*. Third, the retrieved solution is copied and altered in a

heuristically guided manner to finally satisfy the given constraints (Carbonell, 1986).

*Derivational analogy* follows the same processes of analogical thinking. However, the accessed information is different, since it is based on the preservation and reconstruction of past reasoning processes. In the *initial partial matching process*, significant aspects are considered analogous if they share the same reasoning steps, i.e. the same issues are considered and equivalent decisions are made. Second, in *transfer of earlier derivation*, significant aspects of the reasoning process are recreated. Finally, the retrieved derivation is applied to the current situation “by ‘replaying’ the problem solving episode, checking for each step if the derivation is still applicable in the new problem solving context” (Schmid & Carbonell, 1999: 116). To summarize, in transformational analogy the solution is slightly altered to fit the new problem. In derivational analogy, in contrast, previous reasoning processes are applied and adapted to find a solution. As a result, new solutions are likely to be different from previous ones.

While Carbonell introduced derivational analogy as an artificial intelligence model, humans have also been shown to use this strategy (Schmid & Carbonell, 1999). Experiments showed that a high saliency of analogous elements fosters the use of transformational analogy (Schelhorn et al., 2007). Furthermore, participants were more successful in solving novel problems when they studied examples by using instructions fostering derivational analogy (Kleinbeck et al., 2001).

### Verbal Reports in Studying Problem Solving

Measuring solution times is common in problem solving research (e.g. Funke & Spering, 2006); the analysis is based on assumptions about the time which different processes take. However, solution times do not contribute information about cognitive processes at work during problem solving. A combination with other measures such as verbal or behavioral data can lead to more detailed insights.

The elicitation and analysis of verbal reports is an established method to study the processes involved in human problem solving (e.g. Newell & Simon, 1972; Gick & Holyoak, 1980; Ericsson & Simon, 1984). Ericsson & Simon (1984) outline the different processes that can be accessed by different kinds of verbal reports. They argue that think-aloud protocols and retrospective reports do not modify cognitive processes; the task-oriented processes determine what information is heeded and verbalized. Most information in these reports is still held in short term memory. If information is retrieved from long-term memory in retrospective reports, some information might already be missing or erroneous, for instance with regard to difficulties encountered while solving the problem.

Most studies investigating verbal data along with problem solving focus on the content level, identifying the explicit statements elicited from participants during (or following) a problem solving process. Few studies analyze verbal reports on a deeper linguistic level, identifying more precisely *how*

the different processes are described (Caron & Caron-Pargue, 1987; Wedman, Wedman & Folger, 1996), or which linguistic differences can be found between reports of successful and unsuccessful problem solvers (Roth, 1985).

In this study, the analysis of performance data is supplemented by a content and linguistic analysis of different kinds of verbal reports. By combining those measures we aim to

- identify the processes described in the verbal reports and match those to the different processes proposed for transformational and derivational analogy,
- explore how transformational and derivational analogy can be linguistically distinguished in verbal reports, and
- confirm that a high saliency of analogous elements fosters transformational analogy.

### The Effect of High vs. Low Guidance on the Selected Analogical Transfer

Our experiment is based on a study by Schelhorn et al. (2007) in which the saliency of the correspondences between entities in base and target problem was varied. These authors used the following design to address the influence of saliency on the selected analogical strategy.

First, in order to prime participants for analogical reasoning, they were given two example problems (‘The Fortress’ and ‘Radiation’, cf. Gick & Holyoak, 1980) and their solutions, where the second was solved analogously to the first. While those two problems were purely conceptual, the base and target tasks in the study by Schelhorn et al. concerned a type of path-finding problem called “Eulerian Trail” where a path needs to be found that visits every edge of the graph exactly once. After participants were presented with the *base problem* (Boat; see Schelhorn et al., 2007: Appendix), the solution was explained step-by-step (visually supplemented by a graph). Then participants were given the *target problem* (Birthday) and asked to find a solution. In the *high guidance* condition, the initial letters of the five objects that represented the edges of the graph were identical (cities in the base problem and peoples’ names in the target problem). This was not the case in the *low guidance* condition. After giving the solution, participants were asked to map objects from the base to those of the target. Mapping times were recorded. Participants then completed a *Strategy Assessment Questionnaire* (SAQ), which contained 16 statements that participants should agree or disagree with (see Schelhorn et al., 2007: Appendix). Five statements corresponded to the derivational strategy (e.g. *The “boat” and the “birthday” problem seemed similar but I could not figure out how the solutions were related*), and five corresponded to the transformational strategy (e.g. *It was simple to use the “boat” solution to solve the “birthday” problem by replacing the names of the towns with the names of the people*). The remaining statements were fillers. Finally, biographical information was collected.

Schelhorn et al. (2007) found that participants transferred knowledge from the base to the target problem even in the



absence of surface and structural correspondences, namely by using derivational analogy. Since the derivation process takes longer than direct solution transfer, participants in the *high guidance version* should be faster. However, as participants may solve the problem while reading the task instructions, or take time to re-read parts of the instruction when facing difficulties, solution times as such did not seem to be an accurate measure. Mapping times seemed more informative since participants in the *low guidance* condition needed to map the entities from the base problem to those of the target problem in a separate step to solve the mapping task. Results showed that, as expected, mapping times were significantly longer in the *low guidance* condition. Furthermore, participants in the *low guidance* condition agreed with more SAQ questions that corresponded to the derivational strategy than participants in the *high guidance* condition, indicating that low correspondence hampers direct solution transfer.

In a comparable study, Schmid & Carbonell (1999) report similar performance results. Additionally, they briefly report how the two analogical strategies were expressed in think aloud protocols. In preparation for the analysis of verbal reports in our current study, we revisited the set of 14 think aloud protocols to identify linguistic markers of cognitive processes. This analysis revealed the following general structure in 10 of the protocols:

1. construct the graphs representing cities and locks,
2. connect cities satisfying the given constraints, and
3. check the solution during a final evaluation.

The final stage of evaluation was missing in the remaining four protocols. An analysis of the verbs occurring in the protocols revealed that participants were engaged in a number of mental activities: satisfying constraints (*have to*, *need to*), forming hypotheses (*should*, *could*), gaining insight (*I see*), planning (*want to*, *going to*), and recalling (*seen before*). In those data, the distribution of verbs (categorized following Halliday & Matthiessen, 1999) could to some extent be associated with use of derivational and transformational strategies. For instance, participants using derivational analogy used more verbs of ‘doing’ (*go*, *start*) than participants using transformational analogy.

## The Eulerian Trail: Empirical Study

### Hypotheses

The current study supplements the quantitative analysis of performance data by a qualitative analysis of verbal data. We expected protocols to show transformational and derivational strategies, reflecting the processes proposed by Carbonell (1986). In particular:

- Participants in the high guidance version were expected to state explicitly that they noticed the analogy.
- We assumed that participants using the transformational strategy would explicitly state that they used the base solution for solving the target problem. On a linguistic level, they might use explicit markers of

correspondence, such as “the same as”, “similar”, and “analogous”.

- Reports by participants using derivational analogy should include descriptions of different stages of the problem solving process, such as visualizing the different connections (relationships) between the different points (people). Furthermore, unspecific terms (in general) and structuring devices, i.e. ordinal numbers and temporal connectors were expected to be more frequent.

### Design

In addition to replicating the study by Schelhorn et al. (2007) (*original* condition), we collected different kinds of verbal reports in two further conditions, yielding a 2x3 design (high vs. low guidance, original vs. planning vs. retrospective). In the *planning condition*, participants were asked to write down how they would solve the problem (*planning protocol*) after going through the example problem and viewing the target problem for the first time. Furthermore, they were asked to *evaluate* their plans after completing the mapping task. Participants in the *retrospective condition* were asked to write a *report* on how they solved the target problem, and subsequently to write an *instruction* for a friend to solve this problem. Since these verbal reports were collected at different times relative to the problem solving process, different kinds of information were gathered. We expected that planning protocols and retrospective reports would be most likely to include descriptions of the actual problem solving process. Planning protocols were expected to contain information on spontaneous transfer from base to target problem. Retrospective reports may further include memories of detours and fresh starts, and possibly information on the mapping task and meta-information on the study or the problem solving process. This kind of meta-information may also be reported in evaluation protocols. Instructions, on the other hand, would be highly structured and include generalized steps to solve the given problem.

### Procedure

Participants were recruited by various means, e.g. by a call for participation on LinguistList.org and among students at the University of Bremen. As a consequence, the age range was very wide (22 to 73 years, mean 38,1 years). Participants were randomly assigned to conditions. Performance was analyzed with regard to solution and mapping correctness, mapping times, and the answers given in the SAQ. The elicited verbal reports were analyzed qualitatively with regard to their content, the overall structure, and linguistic markers such as verbs, nouns, structuring devices, and other keywords.

### Results

69 participants (35 female and 34 male) took part in the web-based study. 33 of these were given the *high guidance* condition by the system, and 36 saw the *low guidance*

condition. We collected 20 planning protocols, 21 evaluations, 22 retrospective reports, and 21 instructions (evenly distributed between the two guidance conditions).

**Performance and Strategy Assessment.** Solutions to the target problem did not differ significantly between the *high* (57,6%) and *low guidance version* (66,7%). However, the mapping task was solved significantly better in the *high* (90,9%) than the *low guidance version* (63,9%),  $\chi^2 = 7.01$ ;  $p < .01$ . Also, mapping times in the *high guidance* condition ( $n = 33$ ;  $M = 76.9$  s;  $SD = 64.5$ ) were significantly shorter ( $W = 947$ ,  $p < .001$ ) than in the *low guidance* condition ( $n = 36$ ;  $M = 217.0$  s;  $SD = 206.7$ ).

The answers to the SAQ were analyzed using Mann-Whitney U tests to compare the amount of positively ranked questions belonging to the derivational strategy to those that belong to the transformational analogy. Participants with *high guidance* ( $n = 33$ ) agreed with significantly more statements corresponding to transformational analogy than participants with *low guidance* ( $W = 854.4$ ,  $p < .001$ ). Statements that corresponded to derivational analogy were significantly more often confirmed by participants in the *low guidance* ( $n = 35$ ) than by those in the *high guidance* condition ( $W = 771.5$ ,  $p < .01$ ) (Figure 1). One participant in the *low guidance* condition did not complete the SAQ.

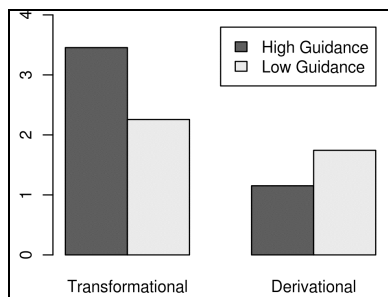


Figure 1: Mean number of agreed SAQ questions.

Running a Kruskal-Wallis test no interaction was found between the verbal elicitation task and mapping times ( $H = 2.8207$ ,  $p = .244$ ). For this reason, and because the results of the performance measures confirmed the findings by Schelhorn et al. (2007), we assume that the elicited verbal data did not affect performance in any substantial way and can therefore contribute to the study of differences in derivational and transformational analogy.

**Verbal Data Analysis.** First we investigated the general structure of the collected 21 planning protocols. In 12 of these, participants stated *noticing the analogy* between base and target problem in the beginning. In 7 protocols this insight was followed by a *description of the graph* and a subsequent description of the *problem solving process*. 17 participants stated the *solution* at the end of the protocol. Similar structures were also found in the retrospective reports, which additionally contained meta-information such as remarks on the study design, the problem solving process, and background knowledge. 7 participants (2 *high guidance*, 5 *low guidance*) furthermore reported how they

solved the mapping task. They reported different strategies, viz. redrawing the graph (4 cases), aligning the entities (once), or matching functions of entities (twice).

88% of all planning protocols in the *high guidance* condition contained statements of *noticing the analogy*, as compared to 40% in the *low guidance* condition. A similar trend emerged in the retrospective reports; 73% of the *high guidance* reports explicitly mentioned noticing the analogies, as compared to 27% in the *low guidance* version. With regard to success, it could be observed that those participants who reported noticing the analogy in their planning protocols succeeded in finding the solution three times more often than those who did not mention the analogy. In retrospective reports no such effect of *noticing the analogy* could be found. A closer analysis showed that some participants reported solving the example problem themselves; if they came up with a wrong solution there, noticing the analogy and transferring the solution to the target would result in a wrong solution.

A closer look at the nature of the descriptions of the analogies revealed a systematic difference. In the *high guidance version* they reflected abstract, general observations of the following kind: “When I read the *birthday problem*, I recognized that it was the *exact same problem* as the *boat problem*”. Here, no alignment of structures or entities is provided. Only protocols collected in the *low guidance version* contained more detailed representations such as: “I recognized right away that the *messages* were analogous to the *locks*.” Here, the entities of the target problem are matched to those of the base problem.

The descriptions of the problem solving processes could be divided into four subcategories. Descriptions of abstract, generalized steps were classified as *general strategy*. If the solution of the base problem was directly transferred to the target problem, this was called *direct solution transfer*. If the strategy (rather than the solution) was described as being transferred, this was categorized as *direct strategy transfer*. If instances of the solution process were specified, the description was classified as *step-by-step*. These categories showed different linguistic markers as illustrated by the following examples:

1. I worked counterclockwise and *connected* as many of the *people* along the edges as possible before working on the *connections* that cut across the middle of my shape. (general strategy)
2. I simply *copied and pasted* the solution from the third problem onto the forth. (direct solution transfer)
3. I *solved* it the *same way*. (direct strategy transfer)
4. Starting from *S*, I *first* connected the group of 3 persons that know each other (*S, B, E*), coming back to *S*. Then from *S* to *R*, and from *R* to the other group of 3 persons that know each other (*R, M, B*), coming back to *R*. Finally from *R* to *E*. (step-by-step)

In detail, the category *general strategy* included

- general statements of the kind ‘find pattern/ mapping’,
- the generalized strategy ‘go through the graph’, and

- the verbs ‘connect’ and ‘draw’ (6 occurrences as compared to 3 in step-by-step descriptions).

The two categories that described direct transfer both contained the markers ‘same’ and ‘again’ (6 times) and the verb ‘solve’ (5 occurrences). But the verb ‘copy’ was solely used in two protocols to describe *direct solution transfer*. And the nouns ‘strategy’ and ‘template’ only occurred in protocols displaying *direct strategy transfer*. Protocols in the category *step-by-step* contained

- first letters or names of entities,
- references to connections drawn by the participants (5 out of 7 cases), and
- a detailed description of the process of drawing the graph, listing the connections and finding a way to satisfy the task constraints (2 cases).

Structuring devices were most frequent in *general strategy* or *step-by-step* descriptions, viz. temporal connectors (13 compared to 3 occurrences in the two *transfer* categories) and the conjunction ‘and’ (27 compared to 2 occurrences).

26 out of 33 *planning protocols* and *retrospective reports* contained strategy descriptions. Of these, 15 were categorized as *general strategy* descriptions and 7 occurrences were found for each of the other three categories respectively. *Step-by-step* descriptions were most frequently used in the *low guidance* version (5 as compared to 2 times). *Direct solution transfer* was more often described in the *high guidance* version (6 as compared to 1 occurrence).

*Instructions* were found to be more structured (structuring devices were used in 50% of the instructions) and more general in describing the steps to be taken (12 out of 20 protocols). Those steps can be summarized as ‘draw the graph & find a pattern by satisfying the given constraints’. Eight instructions included references to task-specific entities (e.g. *lock*, *messenger puzzle*). One participant refused to write an instruction. 16 instructions contained an advice for a specific strategy. The same strategy categories as outlined before could be used for the analysis. A comparison between the strategy described in the instruction and (by the same participant) in the *retrospective report* revealed that people advised a more general strategy (6 cases) or the same strategy (9). No differences with regard to *high* or *low guidance* was found. 15 out of 21 *evaluations* stated if the planned strategy was used; this was mostly the case (10 of 15).

## Discussion

We set out in our study to extend findings previously published by Schelhorn et al. (2007) concerning the use of transformational and derivational strategies in analogical problem solving. Our performance results successfully replicate the earlier findings in that participants given *high guidance* were more likely to use transformational analogy, while participants given *low guidance* could be associated with derivational analogy. The assumption that participants would have to map base and target entities in a separate step

in order to solve the mapping task was confirmed by the descriptions found in 8 protocols.

The analysis of verbal data furthermore provides a range of insights about the cognitive processes involved in analogical problem solving. As a tendency, participants reported recognizing the analogy more often when it was highly salient. Equally unsurprisingly, those recognizing the analogy appeared more likely to succeed in giving the right solution. However, as the *retrospective reports* revealed, participants working their way through the example problem by themselves may give the wrong solution although they noticed the analogy. This observation might explain the low performance in solution correctness of the target problem. The distinctively better performance of participants in the high guidance version on the mapping task supports this view. This interesting possibility could not be detected by performance data alone.

Our qualitative analysis of strategy categories suggested that participants in the *high guidance* version described *direct solution transfer* more often than participants in the *low guidance* version, as opposed to more detailed *step-by-step* descriptions that were associated with *low guidance*. Together these tendencies support the idea that high guidance fosters a transformational strategy involving more superficial and less intricate cognitive processes.

A comparison of the processes identified in our verbal data with the processes hypothesized for transformational and derivational analogy by Carbonell (1986) reveals the following. For transformational analogy, which is associated with *high guidance*, the *initial partial matching process* is expressed in descriptions of *noticing the analogy*. Since the descriptions exhibit a very abstract level of representation, no conclusions can be drawn about the nature of the aspects that are considered analogous. Quite possibly, participants did not need to consider the matter in any more depth (leading to the lack of more detailed alignment descriptions), since a superficial transformation was sufficient. The *analogical matching process* in which knowledge from the base problem is transferred to the target problem is evident in the descriptions of *graph representation*. The following example is representative for the transfer of the sequence of actions: “making nodes for each person and drawing lines between acquaintances”. The process of *alteration of the retrieved solution* is straightforwardly expressed in descriptions of *direct solution transfer*.

In *low guidance* protocols that are associated with the use of derivational analogy, the initial step of matching reasoning processes of base and target problem is also verbally expressed by *noticing the analogy*. These descriptions show a more detailed representation (matching base and target entities), but do not contain information on individual reasoning steps. The transfer of significant aspects of the reasoning process and traces of replaying the problem solving episode can be observed in *step-by-step* descriptions.

Our analysis of verbal reports thus enabled the identification of different cognitive processes involved in analogical problem solving, along with linguistic markers, depending on the degree of guidance which led to the different problem solving strategies previously described as transformational and derivational (Carbonell, 1986). These findings illustrate that the analysis of verbal data contributes to a more detailed understanding of the processes at work during analogical problem solving.

Our elicitation of written data can be regarded as a first broad exploration of the kinds of verbalizations that might be expected along with complex analogical problem solving tasks. Quite typically for free production tasks and low participant numbers that allow for a more or less complete comprehension of the descriptions (rather than performing quantitative computations), the resulting numbers of occurrences of specific phenomena (as reported in this paper) were too small for statistical validation. Nevertheless, the distribution of contents and linguistic markings were both inspiring and suggestive in light of the theoretical background of this study, and thus open up some avenues for further research. In particular, we suggest the following:

- Of the four types of elicited verbalization the planning protocols seemed to be the most informative. Quite unexpectedly, participants seemed to already solve the task while writing up how they would do this, rendering the descriptions rather similar to think aloud data (elicited during, rather than before, problem solving). If this observation can be supported by further studies, it would open up interesting ways of collecting verbal data much more efficiently than possible with think aloud recording.
- The tendency for *low-guidance* participants to produce more detailed procedural descriptions of a derivational problem solving process calls for further exploration. Focusing on this particular aspect, a more controlled elicitation of a larger amount of verbal data should highlight how these matching processes develop over time, as well as the extent to which the two proposed analogical reasoning strategies (transformational and derivational) are systematically distinct. The linguistic markers identified in the present study can serve as a first indication of the ways in which language represents these distinctions.

### Acknowledgments

Funding by the Zentrale Forschungsförderung Universität Bremen is gratefully acknowledged. We thank André Scholz for the technical support.

### References

Carbonell, J. G. (1986). Derivational analogy: A theory of reconstructive problem solving and expertise acquisition. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach – Volume II*. Los Altos: Kaufmann.

- Caron, J., & Caron-Pargue, J. (1987). Towards a psycholinguistic approach of argumentative operators: The ‘Thinking Aloud’ procedure. In F. van Eemeren, R. Grootendorst, J. A. Blair, C. A. Willard (Eds.), *Argumentation: Perspectives and approaches*. Providence USA: Foris Publications.
- Clement, C., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, 15, 89-132.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis. Verbal reports as data*. Cambridge, Mass: MIT Press.
- Funke, J. & Spering, M. (2006). Methoden der Denk- und Problemlöseforschung. In Funke, J. (Ed.), *Enzyklopädie der Psychologie, Band 8*. Göttingen: Hogrefe.
- Gentner, D. (1983). Structure-Mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.
- Gentner, D., & Markman, A. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52 (1), 45-56.
- Gentner, D., Holyoak, K. J. & Kokinov, B. N. (2001). Introduction. In Gentner, D., Holyoak, K.J. & Kokinov, B.N. (Eds.), *The Analogical Mind – Perspectives from Cognitive Science*. Cambridge, MA: MIT.
- Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12, 306-355.
- Halliday, M. A. K & Matthiessen, C. M. (1999). *Construing Experience Through Meaning – A Language Based Approach to Cognition*. London: Continuum.
- Hofstadter, D.R. (2001). Epilogue: Analogy as the Core of Cognition. In Gentner, D., Holyoak, K. J. & Kokinov, B. N. (Eds.), *The Analogical Mind – Perspectives from Cognitive Science*. Cambridge, MA: MIT.
- Kleinbeck, S., Gerjets, P., Scheiter, K., & Schmid, U. (2001). Einfluss derivationaler und transformationaler Beispielformate auf Beispielnutzung und Problemlöseleistung. *Proceedings der 43. Fachtagung experimentell arbeitender Psychologen*.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Engelwood Cliffs, NJ: Prentice Hall.
- Roth, T. (1985). Sprachstatistisch objektivierbare Denkstilunterschiede zwischen „guten“ und „schlechten“ Bearbeitern komplexer Probleme. *Sprache und Kognition*, 4, 178-191.
- Schelhorn, S. E., Griego, J., & Schmid, U. (2007). Transformational and derivational strategies in analogical problem solving. *Cognitive Processing*, 8, 45-55.
- Schmid, U., & Carbonell, J. (1999). Empirical evidence for derivational analogy. *Proceedings der 4. Fachtagung der Gesellschaft für Kognitionswissenschaft* (pp. 116-121). Bielefeld: Bielefeld Universität.
- Schmid, U., Wirth, J., & Polkehn, K. (1999). Analogical Transfer of Non-isomorphic Source Problems. *21st Annual Conference of the Cognitive Science Society*.
- Wedman, J., Wedman, J., & Folger, T. (1996). Analysis of analogical problem-solving processes via think aloud protocols. *Journal of Research and Development in Education*, 30 (1), 51-62.

# Comparing the inductive biases of simple neural networks and Bayesian models

Thomas L. Griffiths (tom.griffiths@berkeley.edu)

Joseph L. Austerweil (joseph.austerweil@gmail.com)

Vincent G. Berthiaume (vberthiaume@berkeley.edu)

Department of Psychology, University of California, Berkeley, CA 94720 USA

## Abstract

Understanding the relationship between connectionist and probabilistic models is important for evaluating the compatibility of these approaches. We use mathematical analyses and computer simulations to show that a linear neural network can approximate the generalization performance of a probabilistic model of property induction, and that training this network by gradient descent with early stopping results in similar performance to Bayesian inference with a particular prior. However, this prior differs from distributions defined using discrete structure, suggesting that neural networks have inductive biases that can be differentiated from probabilistic models with structured representations.

**Keywords:** Bayesian modeling, connectionism, inductive biases, property induction

## Introduction

Cognitive scientists use different mathematical formalisms to model human cognition. Understanding the relationships between these approaches is critical to resolving questions about the nature of the mind. Recently, researchers have debated whether probabilistic or connectionist models of cognition provide better prospects for making progress in cognitive science (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; McClelland et al., 2010). One of the key issues in this debate is that many probabilistic models are defined in terms of structured, discrete representations, while connectionist models use continuous, graded representations that can mimic discrete structure when needed. A possible resolution would be to view probabilistic and connectionist models as lying at different levels of analysis (Marr, 1982), with neural networks a continuous approximation to Bayesian inference over discrete representations. However, this requires establishing whether such an approximation is possible.

To explore this issue, we use the problem of property induction as a case study for investigating the relationship between probabilistic models of cognition and neural networks. Property induction – inferring the properties of a novel object based on the properties of other objects – has played a key role in the debate between probabilistic and connectionist models. An influential probabilistic model explains human property induction in terms of Bayesian inference over discrete representations such as graphs and trees (Kemp & Tenenbaum, 2009), whereas a successful connectionist model explains people’s inferences via continuous representations learned by gradient descent (Rogers & McClelland, 2004).

We use a combination of mathematical analysis and computer simulations to address three questions. First, can a probabilistic model with a discrete representation for a set of objects be approximated by a neural network model with continuous representations? Second, are the solutions that tend

to be found by training neural networks by gradient descent comparable to those produced by Bayesian inference (that is, are the *inductive biases* of these approaches related)? Finally, how compatible are the inductive biases of neural networks with those of structured probabilistic models? We provide positive answers to the first two questions, showing that a simple neural network can always approximate a probabilistic model of property induction, and that training this network using a gradient descent algorithm is similar to Bayesian inference with a particular prior distribution. However, we also show that there remains a significant difference between this prior and distributions based on discrete representations.

## Mathematical analysis

Our mathematical analysis focuses on comparing the model of property induction used by Kemp and Tenenbaum (2009; henceforth KT09) with a linear neural network.

### Setting up the problem

The KT09 model assumes that we want to capture the joint distribution of the elements of continuous  $N$ -dimensional vectors  $\mathbf{x}$  indicating the value of a single property for  $N$  objects.<sup>1</sup> This distribution,  $p(\mathbf{x})$ , results from a diffusion process on a graph. The diffusion process induces a multivariate Gaussian distribution on  $\mathbf{x}$  with mean zero and covariance

$$\Sigma_{\text{discrete}} = \left( \Delta + \frac{1}{\sigma^2} \mathbf{I} \right)^{-1} \quad (1)$$

where  $\Delta$  is the Laplacian of the graph, being  $\mathbf{G} - \mathbf{I}$  for a graph with adjacency matrix  $\mathbf{G}$ , and  $\mathbf{I}$  is the identity matrix.

Now consider a linear neural network model.<sup>2</sup> This model represents an observed  $N \times M$  matrix (the values of  $M$  properties for  $N$  objects) as the product

$$\mathbf{X} = \mathbf{Y}\mathbf{Z} \quad (2)$$

where  $\mathbf{X}$  is the  $N \times M$  matrix of observed objects,  $\mathbf{Y}$  is an  $N \times K$  matrix, and  $\mathbf{Z}$  is a  $K \times M$  matrix. In this model,  $\mathbf{Z}$  is the representation of the set of properties on a hidden layer with  $K$  units (as might be encoded in the weights from an

<sup>1</sup>This formulation is a little counter-intuitive, as the set of objects is fixed but the set of properties is left open (ie. new properties tend to be observed, rather than new objects). This differs from the most intuitive way of thinking about the problem for a neural network, in which the network is trained to predict the properties that objects have, with the set of properties fixed and the set of objects left open.

<sup>2</sup>Neural network models typically use non-linear activation functions at the hidden layer. This complicates the analysis, but we hope to explore the consequences of such non-linearities in future work. We return to this point in the Discussion.

input layer to the hidden layer, with localist coding of properties at the input layer) and  $\mathbf{Y}$  encodes the relation of properties over objects on the hidden layer.<sup>3</sup> A single property vector is generated by multiplying the weight matrix,  $\mathbf{Y}$ , by the vector representing the property,  $\mathbf{z}$ , to obtain  $\mathbf{x} = \mathbf{Y}\mathbf{z}$ . The model is trained by finding weights  $\mathbf{Y}$  and representations  $\mathbf{Z}$  that minimize the error in reconstructing  $\mathbf{X}$ .

### Approximating generalization

It should be clear that the linear neural network can perfectly reproduce any observed matrix  $\mathbf{X}$ , provided  $K$  is greater than or equal to the rank of  $\mathbf{X}$ . This follows simply by thinking about Equation 2 as a set of equations for the entries in  $\mathbf{X}$  where the entries in  $\mathbf{Y}$  and  $\mathbf{Z}$  are free parameters – we can reproduce  $\mathbf{X}$  if we have enough free parameters to construct its linearly independent columns. The more interesting question is thus how the network will generalize. That is, what does it predict for a new property based on what it has learned from the observed properties?

Analyzing generalization requires making assumptions about the nature of the  $\mathbf{z}$  vector for a novel property. If we assume that  $\mathbf{z}$  follows a multivariate Gaussian distribution with mean zero and covariance  $\sigma_z^2 \mathbf{I}$ , we can obtain some results that provide connections between the neural network and Bayesian approaches. This is a reasonable assumption if the weights from the localist node from an unobserved property to the hidden layer are assumed to be independently drawn from a Gaussian distribution. This will be true if the initial weights are drawn from a Gaussian, but as we show below it is also consistent with the implicit prior assumed by gradient descent algorithms.

We can determine the prediction the neural network will make for a new property by asking how  $\mathbf{x}$  is distributed given  $\mathbf{Y}$ . Using standard Gaussian identities,  $\mathbf{x}$  will be multivariate Gaussian with mean zero and covariance

$$\Sigma_{\text{continuous}} = \sigma_z^2 \mathbf{Y}\mathbf{Y}^T \quad (3)$$

since  $\mathbf{x}$  is a linear function of a Gaussian random variable.

Characterizing the distribution on  $\mathbf{x}$  implied by this model makes it straightforward to construct a condition under which the model produces the same joint distribution as a probabilistic model based on any discrete graph structure: This will occur when  $\Sigma_{\text{discrete}} = \Sigma_{\text{continuous}}$ . This can be used to establish a direct connection between the neural network's representations for the objects and the graph Laplacian  $\Delta$ . In particular,  $\mathbf{Y}$  can be obtained from the eigenvectors of  $\Delta$ . If the network is trained from a matrix  $\mathbf{X}$  of property values sampled from  $p(\mathbf{x})$ , then any learning algorithm that produces a representation corresponding to the principal components of  $\mathbf{X}$  will

<sup>3</sup>Since the model is linear, this interpretation can be “transposed” to give another interpretation, where  $\mathbf{Y}$  is the hidden layer representation of the objects and  $\mathbf{Z}$  the weights for the properties. This is a more intuitive way of formulating the model and is also more consistent with connectionist models of these phenomena, as advocated by Rogers and McClelland (2004). However, this interpretation is a little harder to use to get intuitions about the results shown below.

approximate this outcome, with the approximation improving as the number of samples  $M$  increases. Thus, the answer to our first question is that the probabilistic model can be approximated arbitrarily well by a neural network.

Establishing that our simple neural network with continuous representations can potentially approximate the generalization performance of a probabilistic model using a discrete representation raises a different question: Will these models also perform similarly when learning those representations from data? That is, if we train a neural network model on a finite number of samples from  $p(\mathbf{x})$ , will it behave similarly to a probabilistic model that infers a discrete representation from the same data via Bayesian inference? This is a question about the inductive biases of these two different approaches to learning – those factors that lead a learning algorithm to favor one solution over another. In the context of the property induction problem, this question reduces to whether the predictions produced by the neural network after training will be similar to those resulting from Bayesian inference with a particular prior distribution.

### Gradient descent and Bayesian inference

Gradient descent is a standard approach to training a neural network, where the weights are assigned small random values and then modified in the direction indicated by the gradient of the error repeatedly for a fixed number of training iterations. In this section, we summarize results showing that this learning algorithm behaves similarly to Bayesian inference with a Wishart prior on covariance matrices.

For simplicity, we start by considering the problem of updating  $\mathbf{z}$  for a single property, keeping  $\mathbf{Y}$  fixed. In this case the goal is to find the  $\mathbf{z}$  such that  $\mathbf{Y}\mathbf{z}$  minimizes the squared error in reconstructing the corresponding property vector  $\mathbf{x}$ . We can write the objective function as  $(\mathbf{x} - \mathbf{Y}\mathbf{z})^T(\mathbf{x} - \mathbf{Y}\mathbf{z})$ . Differentiating, we obtain the weight update rule

$$\Delta \mathbf{z} = \eta \mathbf{Y}^T (\mathbf{x} - \mathbf{Y}\mathbf{z}) \quad (4)$$

where  $\eta$  is a learning rate (assuming simultaneous updates).

For comparison with performing Bayesian inference, we can derive the estimate for  $\mathbf{z}$  that we would obtain by assuming a Gaussian prior and finding the posterior mean (or the maximum a posteriori value, as they are the same in this case). The Bayesian estimate is

$$\hat{\mathbf{z}} = (\mathbf{Y}^T \mathbf{Y} + \frac{\sigma_x^2}{\sigma_z^2} \mathbf{I})^{-1} \mathbf{Y}^T \mathbf{x} \quad (5)$$

where  $\sigma_x^2$  is the assumed noise variance in  $\mathbf{x}$ .

Inspecting these two equations, we can see that they use two different forms of *regularization* – approaches to controlling the complexity of the solution found by learning. Neural network training typically starts with weights close to zero, so weights grow over successive passes through the data. Stopping early keeps weights smaller. In the Bayesian solution, the ratio of  $\sigma_x^2$  to  $\sigma_z^2$  controls the size of the weights: If  $\sigma_z^2$  is small relative to  $\sigma_x^2$  (i.e., we are more confident in our prior

beliefs than the observed data), the corresponding term can dominate the matrix that is inverted, reducing the weights proportionally. Despite this difference in regularization style, there are cases where they will produce similar results: If  $\mathbf{z}$  is close to zero and  $\mathbf{Y}^T \mathbf{Y}$  is close to  $c\mathbf{I}$ , then  $\hat{\mathbf{z}}$  will equal  $\mathbf{z}$  after one pass of gradient descent with  $\eta = 1/(c + \frac{\sigma_z^2}{\sigma_x^2})$ .

More generally, it is possible to show that the solution produced by a linear neural network trained by gradient descent with early stopping is equivalent to generating a Bayesian estimate with a Gaussian prior (Fleming, 1990; Santos, 1996). When applied to Equation 4, these results indicate that following this learning rule is equivalent to assuming a Gaussian prior on  $\mathbf{z}$  with mean zero and a covariance determined by  $\mathbf{Y}$  and the number of iterations of learning.

While the analysis presented so far has focused on  $\mathbf{Z}$ , the linearity of the network means that learning  $\mathbf{Y}$  can be analyzed in the same way. A Gaussian prior on  $\mathbf{Y}$  implies that the implicit prior on  $\mathbf{Y}\mathbf{Y}^T$  assumed by a neural network trained by gradient descent with early stopping is a Wishart distribution, the distribution obtained by taking the product of two matrices drawn from a multivariate Gaussian (Muirhead, 1982).

## Summary of mathematical results

The key results of the mathematical analyses presented in this section are that the generalization performance of the KT09 model can be approximated by a linear neural network model with continuous representations, and that the inductive bias induced by training the neural network by gradient descent with early stopping should be similar to that of Bayesian inference with a Wishart prior on covariance matrices. These results make two clear predictions: Neural networks should perform best when learning from data whose covariance matrices are Wishart distributed, and we should expect them to perform more similarly to Bayesian models that use a Wishart prior than to models with other priors.

These results also raise a question: How similar is the Wishart distribution to distributions that are based on discrete representations? If the distributions are similar, then the inductive biases of neural networks and probabilistic models with discrete representations will also be similar, meaning that these approaches need not be seen as lying in opposition to one another. If the distributions are different, then there are opportunities to empirically separate these accounts and we cannot view simple neural networks as a scheme for approximating the solutions identified by probabilistic models.

## Simulations

We explored the issues raised by our mathematical analyses through simulations comparing the performance of neural networks and Bayesian models with different prior distributions. The set of priors that we used included the Wishart distribution as well as several distributions based on discrete structures. Following the KT09 model, we included distributions on covariance matrices by defining a distribution on graphs  $\mathbf{G}$  and then deriving a covariance matrix for

each graph. The distributions on graphs we considered were stochastic graph grammars that generate trees, chains, grids, and partitions (Nagl, 1986; Kemp & Tenenbaum, 2008) and Erdős-Rényi random graphs (Erdős & Rényi, 1959).

Our analysis proceeded as follows. For each prior distribution over covariance matrices, we generated  $T$  samples of  $N \times N$  covariance matrices  $\Sigma_1, \dots, \Sigma_T$ . From each covariance matrix, we sampled a  $N \times M$  matrix  $\mathbf{X}$  containing the values of  $M$  features for each of the  $N$  objects ( $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_M]$ ,  $\mathbf{x}_i \sim N(\mathbf{0}, \Sigma)$ ). We then computed the marginal probability of these samples under a Wishart distribution, integrating over its parameters. This let us determine how closely different priors relate to the Wishart distribution.

To compare the different approaches to learning, we applied the neural network and Bayesian models to all of the samples of  $\mathbf{X}$  we had produced. We found  $\mathbf{Y}\mathbf{Y}^T$  at different stopping points and compared this to the true covariance matrix for data generated from each of the different priors. The goal of this first analysis was to evaluate whether the neural network performed best with data whose covariance matrix was Wishart distributed. For the second analysis, we also obtained an estimate of the covariance matrix from each sample using Bayesian inference with each of the different prior distributions and calculated the distance between these covariance matrices and  $\mathbf{Y}\mathbf{Y}^T$ . This allowed us to examine how the distance between the solutions produced by the neural network and Bayesian inference was related to the extent to which the priors were similar to a Wishart distribution.

## Calculating marginal Wishart probabilities

To perform our analysis, we must be able to calculate how close a distribution is to a Wishart. We did this using the marginal probability of a set of covariance matrices under a Wishart, integrating over the parameters of the distribution. The result is a measure of the “Wishartiness” of the covariance matrices, which can be applied to samples from different distributions in order to evaluate their similarity to a Wishart.

Assume we have a Wishart distribution with degrees of freedom  $b$  and covariance center  $\mathbf{S}$ , and that  $\mathbf{S}$  is drawn from an inverse-Wishart distribution with parameters  $a$  and  $\Psi$ . We draw covariance matrices  $\Sigma_1, \dots, \Sigma_T$  from this distribution. The marginal probability of  $\Sigma_1, \dots, \Sigma_T$  given  $a$ ,  $b$ , and  $\Psi$  is

$$p(\Sigma_1, \dots, \Sigma_T) = \int d\mathbf{S} p(\mathbf{S}|a, \Psi) \prod_{t=1}^T p(\Sigma_t|b, \mathbf{S})$$

which yields

$$p(\Sigma_1, \dots, \Sigma_T) = \frac{\Gamma_N(\frac{1}{2}(a + bT)) |\Psi|^{a/2} \prod_{t=1}^T |\Sigma_t|^{(b-N-1)/2}}{\Gamma_N(a/2) (\Gamma_N(b/2))^T |\Psi + \sum_{t=1}^T \Sigma_t|^{(a+bT)/2}}$$

where  $\Gamma_N(\cdot)$  is the multivariate gamma function,

$$\Gamma_N(x) = \pi^{N(N-1)/4} \prod_{j=1}^N \Gamma(x + (1-j)/2)$$



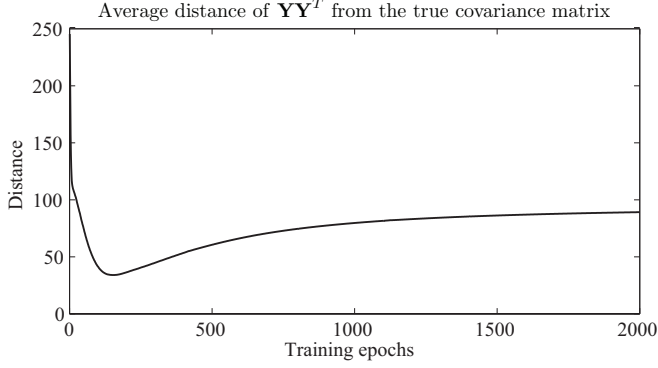


Figure 1: Average distance between the true covariance matrix and the covariance matrix learned by the neural network.

and  $\Gamma(x)$  is the generalized factorial function (Boas, 1983). This is the ratio of the normalization constants for a Wishart and an inverse-Wishart distribution, due to conjugacy.

### Neural network learning

The linear neural network is defined by two matrices: a  $N \times K$  matrix  $\mathbf{Y}$  that maps the properties into the latent space and a  $K \times M$  matrix  $\mathbf{Z}$  that maps the latent space to the objects. We trained the neural network by gradient descent on error, with

$$\Delta \mathbf{y} = \eta (\mathbf{x} - \mathbf{y}\mathbf{Z})\mathbf{Z}^T \quad (6)$$

and Equation 4 as the weight update rules.  $K$  was set to one more than the rank of the object matrix  $\mathbf{X}$ . The weights were initialized to normally distributed random values with mean 0 and variance 0.05. We used a learning rate  $\eta$  of 0.0025 and 2000 training epochs (full passes through the data), which were determined by pilot simulations. At each possible stopping point (epoch), we recorded  $\mathbf{Y}\mathbf{Y}^T$ . Figure 1 shows the average distance between  $\mathbf{Y}\mathbf{Y}^T$  and the true covariance matrix as a function of epoch, which initially decreases and then rises again due to overfitting.

### Priors and Bayesian inference

We considered eight different prior distributions, requiring us to use three different algorithms for Bayesian inference.

**Wishart prior.** The first Bayesian model used a Wishart prior with covariance center  $\mathbf{I}$  and degrees of freedom  $b = 1000$ . Unfortunately the Wishart is not conjugate to the multivariate Gaussian, so we found an estimate of the covariance matrix under this prior using stochastic search with simulated annealing. The state of the search (a covariance matrix) was initialized to a random draw from the posterior distribution using an inverse-Wishart prior (for details, see Gelman, Carlin, Stern, & Rubin, 1995). A new proposed state was then drawn from a Wishart distribution centered at the current state with  $b + N$  degrees of freedom. A Metropolis-Hastings acceptance rule was used to decide whether to replace the current state with the proposed state, based on the product of two ratios of their (unnormalized) posterior probabilities

and the probability of generating the proposed state from the current state and vice versa (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). This probability was annealed by raising the probability to the power  $1/\tau$ , with  $\tau$  decreasing according to a logarithmic schedule.

**Graph grammar priors.** We used four priors based on graph grammars, defining distributions on graphs that correspond to trees, grids, chains, and partitions (Nagl, 1986; Kemp & Tenenbaum, 2008). These random graph grammars are generative processes that start with a single node and then replace a random node in the current graph with two nodes  $J$  times, where  $J \sim \text{Geom}(\theta)$ . Different graph structures result from using different rules for connecting the parents and children of the old node to the new nodes (for the tree grammar, there is also a latent node that cannot contain any objects), and different rules for connecting the new nodes result in different generated graph structures.<sup>4</sup> Afterwards, the objects are assigned to nodes uniformly at random (except not to latent nodes). For example, if the rule for node replacement does not create any edges, then the random graph grammar generates random partitions of the objects.

To convert the graph to a covariance matrix, we follow Kemp and Tenenbaum (2008) by first forming an “entity” graph containing  $N + L$  nodes, where the first  $N$  nodes represent each object and are only directly connected with an edge to their assigned node. Second, we complete the “entity” graph by connecting the last  $L$  nodes to each other according to the result of the previous graph replacement process. Next, we form a  $(N + L) \times (N + L)$  adjacency matrix  $\mathbf{W}$ , where  $1/w_{ij} \sim \text{Exp}(\beta)$  if there is an edge between nodes  $i$  and  $j$  (representing how close nodes  $i$  and  $j$  are). Otherwise,  $w_{ij} = 0$ . This specifies a  $(N + L) \times (N + L)$  covariance matrix for the multivariate Gaussian distribution over the latent and observed variables,  $(\mathbf{E} - \mathbf{W} + \frac{1}{\sigma^2}\mathbf{I})^{-1}$  where  $\mathbf{E}$  is a  $(N + K) \times (N + K)$  diagonal matrix with  $e_{ii} = \sum_j w_{ij}$  and  $\mathbf{I}$  is the  $(N + K) \times (N + K)$  identity matrix. The hidden nodes can be marginalized out analytically, resulting in the  $N$  objects being normally distributed with covariance matrix given by the first  $N \times N$  elements of the original covariance matrix.<sup>5</sup> Bayesian inference was performed with code from <http://charleskemp.com>, which uses stochastic search to find an estimated maximum *a posteriori* covariance matrix for a given set of data.<sup>6</sup>

**Erdős-Rényi priors.** In addition to the four random graph generators from Kemp and Tenenbaum (2008), we used a standard random graph generator: the Erdős-Rényi random graph (Erdős & Rényi, 1959). Each object is represented by a node. Unlike the node replacement grammars, we gener-

<sup>4</sup>For simplicity, we assumed the graph structures are undirected.

<sup>5</sup>It is important to note that this is not equivalent to the first  $N \times N$  elements of the inverse covariance matrix.

<sup>6</sup>The parameters were set to  $\beta = 0.4$  (edge length parameter),  $\sigma^2 = 0.4$  (covariance matrix regularization parameter), and  $\theta = 1 - e^{-3}$  (simplicity bias), which are similar to the values used by Kemp and Tenenbaum (2008). We used the “45” speed setting.



ate random graphs by directly connect pairs of objects with an edge with probability  $p$ . Once the graph is generated, the implied covariance matrix is found by the same procedure as before (except we do not need to perform the additional marginalization step as the initial covariance matrix is already  $N \times N$ ). We considered priors corresponding to  $p \in \{0.1, 0.5, 0.9\}$ . Covariance matrices with these priors were estimated using stochastic search by simulated annealing. The covariance matrix was initialized to a random Erdős-Rényi covariance matrix and proposals were generated from the current state by removing or deleting a random number of edges (such that the number of edges in the proposals were binomially distributed). The search procedure and annealing schedule were otherwise the same as for the Wishart prior.

### The distance between covariance matrices

To analyze the results produced by the neural network and Bayesian models, we needed a measure of the similarity of two matrices. We used a distance metric between positive definite matrices (valid covariance matrices) defined by Förstner and Moonen (1999)

$$d(\Sigma_1, \Sigma_2) = \sqrt{\sum_{i=1}^n \ln^2 \lambda_i(\Sigma_1, \Sigma_2)}, \quad (7)$$

where  $\lambda_i(\Sigma_1, \Sigma_2)$  are the generalized eigenvalues of  $\Sigma_1$  and  $\Sigma_2$ , being the roots of  $|\lambda \Sigma_1 - \Sigma_2| = 0$ . When computing these distances, we used the best stopping point for the neural network (the one resulting in minimal distance). Looking across epochs, we found the value of  $\mathbf{Y}\mathbf{Y}^T$  with the minimal distance to the true covariance matrix and to the eight covariance matrices estimated by Bayesian models with different priors.

### Simulation procedure and results

For each prior, we generated 101 data sets that each consisted of  $T = 100$  covariance matrices. From each matrix, we sampled the values of  $M = 100$  features for  $N = 10$  objects. We then computed the marginal probability of the covariance matrices generated by each prior under the assumption they were drawn from a Wishart distribution, with the median result shown in the top row of Table 1. As expected, the Wishart prior was the most compatible with a Wishart distribution. The discrete priors produced results that were reasonably consistent with the Wishart distribution, while the the Erdős-Rényi generative processes produced results that were poorly characterized as Wishart. We used the data set with the median Wishart value for the subsequent analyses.

Next, we trained neural networks on the object set  $\mathbf{X}$  generated from each covariance matrix sampled from each of the eight priors, and computed the distance between  $\mathbf{Y}\mathbf{Y}^T$  and the true covariance matrix. The results are shown in the second row of Table 1. Performance was statistically significantly better when the true covariance matrices were drawn from the Wishart, consistent with our mathematical analysis.

Finally, we found Bayesian estimates of the covariance matrix for each object set  $\mathbf{X}$  using all eight priors. Stochastic search was run for 20000 iterations in each case. We

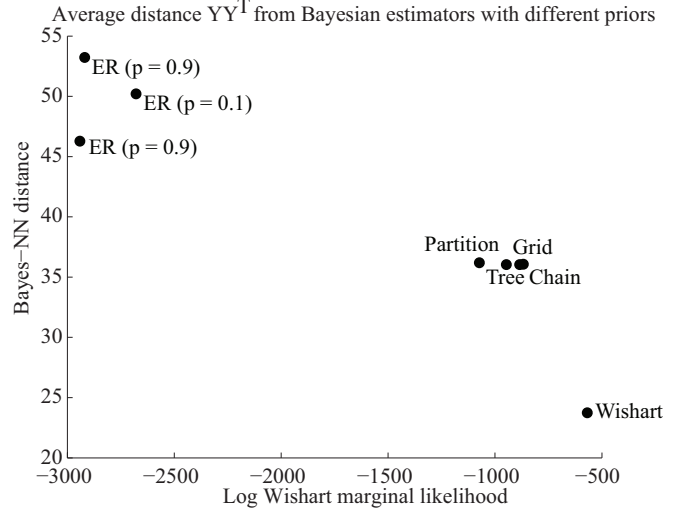


Figure 2: Average (smallest possible) distance of  $\mathbf{Y}\mathbf{Y}^T$  from the Bayesian estimates of the covariance matrix, plotted as a function of the logarithm of the Wishart marginal likelihood for the corresponding prior.

computed the distance between  $\mathbf{Y}\mathbf{Y}^T$  and the Bayesian estimates for each object set, then averaged this quantity across all object sets. The results are shown in the third row of Table 1. As predicted, we found a negative correlation between the distance between estimates and the extent to which the corresponding prior is consistent with a Wishart distribution (as reflected by the marginal probabilities in the first row of Table 1) with  $r = -0.92$  and  $r = -0.83$  for Pearson’s product-moment and Spearman’s rank-order correlation, respectively.<sup>7</sup> A scatterplot showing the relationship between these two quantities is shown in Figure 2.

The variation in how well the neural network approximated the Bayesian estimates with different prior distributions is informative about the inductive biases of neural networks and structured probabilistic models. The neural network was closest in performance to Bayesian inference with a Wishart prior, which is purely continuous. All priors based on discrete structure, in the form of an underlying graph, resulted in statistically significantly worse performance. Within these discrete priors, those based on graph grammars were better approximated than the Erdős-Rényi priors. This pattern of results is interesting from the perspective of the debate between probabilistic and connectionist accounts of property induction, which has focused on discriminating the predictions of probabilistic models using representations based on graph grammars from neural networks. Our results suggest that this may be harder than discriminating probabilistic models that assume arbitrary discrete structure, as in the Erdős-Rényi priors, from neural networks.

<sup>7</sup>We confirmed that this correlation could not be fully explained by the norm of the matrices, but plan on running further simulations to rule out other possible alternative explanations for our results.

Table 1: Properties of different priors and comparison of gradient descent and Bayesian learning

	Wishart	Graph grammar priors				Erdős-Rényi priors		
		Grid	Chain	Tree	Partition	$p = 0.1$	$p = 0.5$	$p = 0.9$
Marginal probability under Wishart	-567.87	-867.68	-884.95	-946.22	-1073.10	-2678.04	-2940.98	-2919.44
Distance of $\mathbf{Y}\mathbf{Y}^T$ from true covariance	14.15 <sup>a</sup>	33.31 <sup>b</sup>	34.18 <sup>b</sup>	32.36 <sup>b</sup>	33.97 <sup>b</sup>	33.93 <sup>b</sup>	31.73 <sup>b</sup>	33.35 <sup>b</sup>
Distance of $\mathbf{Y}\mathbf{Y}^T$ from Bayesian estimate	23.53 <sup>a</sup>	36.04 <sup>b</sup>	36.07 <sup>b</sup>	36.03 <sup>b</sup>	36.19 <sup>b</sup>	50.21 <sup>c</sup>	46.29 <sup>d</sup>	53.23 <sup>e</sup>

Note: In each row, different superscripts indicate statistically significant differences in scores (Bonferroni  $p < .05$ ).

## Discussion

Our analysis of the relationship between probabilistic and connectionist models in the context of property induction has produced several interesting results. First, the generalization performance of a probabilistic model with a discrete representation can be approximated by an appropriately configured linear neural network with continuous representations. Second, training such a network by gradient descent with early stopping is similar to performing Bayesian inference over covariance matrices with a Wishart prior. Finally, prior distributions that assume discrete structure vary in the extent to which they resemble a Wishart prior, and this variation predicts how well Bayesian inference using those prior distributions is approximated by a neural network. However, all prior distributions using discrete structure that we considered resulted in worse approximations than that given with a Wishart prior.

There are limitations in the analyses presented here that we hope to address in future work. As noted earlier, the assumption of linearity in the neural network deviates from standard practice in connectionist modeling. While we do not expect that this will substantially change our results (given that early stopping enforces small weights, effects of the non-linearity should be minimized), further simulations should be conducted to confirm that this is the case. We would also like to explore more sophisticated learning algorithms, such as cascade correlation (Fahlman & Lebiere, 1990), which may result in different inductive biases.

Returning to the questions that motivated our investigation, our results provide a mixed set of answers as to the potential for neural networks to be viewed as a continuous approximation to Bayesian inference over discrete representations. While specific neural networks can always be constructed that emulate the generalization performance of probabilistic models using discrete representations and the inductive biases of neural networks can be expressed in a form that is consistent with Bayesian inference, these inductive biases are quite different from those of Bayesian models using priors defined on discrete objects. Our results suggest that there is room to empirically separate these two approaches, and that identifying neural systems that can approximate arbitrary Bayesian mod-

els may require going beyond simple neural networks that use general-purpose learning algorithms.

**Acknowledgments.** We thank Noah Goodman, Surya Ganguli, and Jay McClelland for discussions and grant number FA-9550-10-1-0232 from the Air Force Office of Scientific Research and a fellowship from the Fonds de Recherche du Québec to VGB for funding.

## References

- Boas, M. L. (1983). *Mathematical methods in the physical sciences* (2nd ed.). New York: Wiley.
- Erdős, P., & Rényi, A. (1959). On random graphs, I. *Publicationes Mathematicae*, 6, 290-297.
- Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In *Advances in Neural Information Processing Systems 2*.
- Fleming, H. E. (1990). Equivalence of regularization and truncated iteration in the solution of ill-posed image reconstruction problems. *Linear Algebra and its Applications*, 130, 133-150.
- Förstner, W., & Moonen, B. (1999). A metric for covariance matrices. In F. Krumm & V. S. Schwarze (Eds.), *Qua vadis geodisa...? festschrift for Erik W. Grafarend on the occasion of his 60th birthday* (p. 113-128). Stuttgart, Germany: Stuttgart University.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.
- Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8), 357-364.
- Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, USA*, 105, 10687-10692.
- Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, 116(1), 20-58.
- Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.
- McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., et al. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8), 348-356.
- Metropolis, A. W., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087-1092.
- Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. New York: John Wiley & Sons.
- Nagl, M. (1986). Set theoretic approaches to graph grammars. In *Proceedings of the 3rd international workshop on graph-grammars and their application to computer science* (p. 41-54). London, UK: Springer.
- Rogers, T., & McClelland, J. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.
- Santos, R. J. (1996). Equivalence of regularization and truncated iteration for general ill-posed problems. *Linear Algebra and its Applications*, 236, 25-33.

# Cooperation in Prisoner's Dilemma Game: Influence of Social Relations

Maurice Grinberg (mgrinberg@nbu.bg)  
Evgenia Hristova (ehristova@cogs.nbu.bg)  
Milena Borisova (borisova\_milena@abv.bg)

Department of Cognitive Science and Psychology,  
New Bulgarian University, 21 Montevideo Street, Sofia 1618, Bulgaria

## Abstract

The paper explores the influence of the type of relations among players on cooperation in the Prisoner's dilemma game. The relations between players are operationalized according to Fiske's relational models theory (Fiske, 1991): communal sharing, authority ranking, equality matching, and market pricing. This is achieved by using various ways of distributing the total payoff gained by a dyad of players in a series of Prisoner's dilemma games: each player receives the total payoff (*unity*), one of the players receives more than the other (*hierarchy*), each player receives half of the total payoff (*equality*), each player receives a portion of the total payoff proportional to his/hers individual payoffs (*proportionality*). For these four conditions, the cooperation rates, the mutual cooperation, the mutual defection, and the payoffs gained are analyzed and compared for a series of forty games. The results show that in the *proportionality* condition there is less cooperation, less mutual cooperation, more mutual defection and less total payoff than in the other three conditions.

**Keywords:** Prisoner's Dilemma, decision-making, cooperation, social interaction, relational models

## Introduction

### Prisoner's Dilemma Game

The Prisoner's dilemma (PD) game is one of the most extensively studied social dilemmas. PD is a two-person game. The interest in studying PD game arises from the idea that many social situations and problems such as overpopulation, pollution, energy savings, participation in a battle, etc. have such a dilemma structure (Dawes, 1980). The payoff table for this game is presented in Figure 1. In the PD game the players simultaneously choose their moves – C (cooperate) or D (defect), without knowing their opponent's choice.

In order to be a Prisoner's dilemma game, the payoffs (see Figure 1) should satisfy the inequalities  $T > R > P > S$  and  $2R > T + S$ . Because of this game structure a dilemma appears – there is no obvious best move. On one hand, the D choice is dominant for both players – each player gets larger payoff by choosing D (defection) than by choosing C (cooperation) no matter what the other player chooses. On the other hand, the payoff for mutual defection (P) is lower than the payoff if both players choose their dominated C strategies (R for each player).

As PD game is used as a model for describing social dilemmas and studying the phenomenon of cooperation, there is a great interest in the conditions that could promote or diminish cooperation.

		Player II	
		C	D
Player I	C	R, R	S, T
	D	T, S	P, P

		Player II	
		C	D
Player I	C	3, 3	1, 4
	D	4, 1	2, 2

Figure 1: Payoff tables for the PD game – with standard notation for the payoffs and an example. In each cell the comma separated payoffs are the Player I's and Player II's payoffs, respectively.

In game theory several assumptions about the game and the players are made. The agents are assumed to be perfectly rational and to have perfect information about the game. Under these conditions, they are supposed to try to maximize their payoffs in a completely selfish manner (Colman, 2003). From this point of view the dominant strategy in the game is defection (in one-shot or in repeated PD games with a fixed and known number). This prediction is in contrast with the behavior of the players observed in laboratory settings or in real life situations.

In human societies, people cooperate all the time and often cooperation is seen as one of the foundations of human civilization (see e.g. Gärdenfors, 2003). Sally (1995) provides a meta-review of the experiments involving PD games published between 1958 and 1995 and shows that in its iterated version (the game is played many times), cooperation choices are made in 20-50 % of the games (mean 47.4 %) and even in one-shot games many players cooperate although much less than in the iterated version.

Several studies have shown how cooperation can emerge from expected utility or anticipatory reinforcement models without any specific relations between the players (Grinberg, Hristova, & Lalev, 2010; and the references there in).

However, it is clear that the deeper understanding of how people make decisions while playing PD games should account for the role of the social relations involved in the interactions. Moreover, as the PD game is central in the

modeling of social interactions it can be used to explore the existence and limits of the relational social types as posited by relational social models (see e.g. Haslam, 2004). Exploring the potential of games like the PD game as modeling relational types is one of the goals of this paper.

### Social Interactions and Cooperation

How decision-making in PD games is influenced by social interactions has been explored in many studies that try to account for the contradiction between the normative predictions and the experimental results in PD games.

Several studies have established the influence of social interaction on cooperation. For instance, Durkin, Frost, Aronov & Breslow (1967) found that cooperative moves double when participants have visual contact with each other compared to the condition where they don't have such contact. Sally (2001) investigated the behavioural changes in participants who know each other or are psychologically or socially close and discussed the importance of such closeness in game strategy building. According to this account, participants play differently depending on how they perceive their opponent – as a friend or a stranger. In both cases, according to Sally (2001), the social interaction is essential and the social dilemmas like PD need to be investigated from the perspective of a general relational theory.

There are several theories that account for the cooperative behavior in PD games in terms of socially established values and stress the importance of social interaction and relationships as tools for achieving cooperation. Among them are theories that explain cooperation by altruism, reciprocity or reputation building.

*Reputation building theory* (Kreps et al., 1982; Andreoni & Miller, 1993) is one of the main theories aimed to explain cooperation in iterated PD game. This theory assumes that players are self-interested (not altruists), but the repetition in iterated PD games creates incentives to cooperate. According to this model, the player is building himself a reputation of a cooperative player and expects that the other player will also cooperate.

*Reciprocity*, according to many researchers, is a widespread norm and is the basis of many relationships and societies (Trivers, 1972). People reciprocate cooperation with cooperation. One of the most studied strategies that are based on reciprocity is the tit-for-tat strategy. A player using this strategy cooperates initially, and then plays the same as his/her opponent did in the previous game. It has been demonstrated in computer tournaments that in the long run the tit-for-tat strategy results in higher payoffs compared to other strategies (Axelrod & Hamilton, 1984; Komorita, Hilty, & Parks, 1991).

Another influential theory about cooperation in PD game is based on the concept of *altruism*. In contrast to reputation building theory, this theory assumes that some players are not strictly self-interested and view more benefit in cooperation than the actual payoffs they receive (Cooper et

al., 1996). From an altruistic perspective, cooperation can yield higher payoffs than defection.

Although these social theories of cooperation have been proposed to explain cooperative behavior unexpected by normative game theory, it is interesting to consider more general social theories that are more closely related to the game theoretic analysis of social relations. In our opinion such a theory is the relational models theory proposed by Alan Fiske (Fiske, 1991).

### Relational Models Theory

Relational models theory (Fiske, 1992; Fiske & Haslam, 1996; McGraw, Tetlock, & Kristel, 2003; Rai & Fiske, 2011) states that there are four basic schemas that are used to build, organize and maintain relationships and interactions among individuals in a society. These models are supposed to be universal and all relations could be described by these models or by combination of them. The four types of relations generate four modes for every aspect of the interactions between people – resource allocation, moral judgments, decision-making, etc.

These four relation models are the following (Fiske, 1992):

- **Communal Sharing** – relations in an undifferentiated group of people with equivalent status. Everyone in a community - which could consist of two members or could be very large – has some rights and some duties. The focus is on commonalities and not on distinctions;
- **Authority Ranking** – implies an ordinal ranking in society and this ranking scheme determines one's relative status. For instance, military hierarchy can be considered a prototype of such relations;
- **Equality Matching** – relations are based on a model of one-to-one correspondence as in turn-taking, tit-for-tat strategies, etc. The social prototype would be friendship networks, in which reciprocity is a norm which rules the distribution of wealth;
- **Market Pricing** – based on a model of proportionality in social relations in which people reduce their interaction to some ratios of utility measures. Examples of relations of this type are the ones governed by prices, rational calculations, expected utilities, etc.

### Payoff Distribution in PD and Fiske's Relational Model Types

Fiske's relational model theory (Fiske, 1992) claims that different relational models influence and are manifested in a lot of domains and activities, e.g. reciprocal exchange, distribution, contribution, work, significance of time, social influence, constitution of groups, motivation, moral judgments, etc.

Here, we focus on the type of distribution of group resources using one and the same game, namely the PD

game. We focus on social interaction related to social exchange as instantiated by contribution and distribution of a common resource. We share the opinion that situations involving exchange are the most appropriate to study the four types of relational models in isolation (Haslam, 2004).

In the classical PD game experiments each player is rewarded according to his/hers personal payoffs. However, Fiske's relational model theory states that in real-life situations the distribution of payoffs and resources depends on the type of the relational model behind the social interaction. There are **four types of distributions**, corresponding to the four relational types described above (Fiske, 1992, Table 1, p. 694):

- **communal sharing** – 'corporate use of resources regarded as common, everything belongs to all together';
- **authority ranking** – 'the higher the person's rank, the more he or she gets';
- **equality matching** – 'to each the same, everyone gets identical shares';
- **market pricing** – 'to each in due proportion'.

## Goals of the Study

The main goal of the present study is to make a first step in the mapping of Fiske's relational models theory to games from game theory. More specifically, we want to study how the four relational types, implemented as distinct payoff distributional models, influence cooperation in PD games. As relational models are complex and encompass various domains, in the present study the focus is on the different **distribution schemas** within the same type of games (the PD game).

We aim to explore what is the influence of the type of relation among players on a set of game outcomes that characterize the playing of a PD game – cooperation, mutual cooperation, and mutual defection. It is also important to check the influence of the distribution model on the overall payoffs that are received – e.g. what type of model is more beneficial in terms of total payoff earned in interactions with the strategic structure of the PD game.

Cooperation is expected to be the highest if the payoff distribution is in accordance with the communal sharing model. Cooperation is expected to be lowest if the distribution follows the rules of market pricing model, e.g. when everyone is rewarded depending on his/her personal contribution – in this scenario we expect more individualistic orientation of the players.

## Method

### Stimuli

A sequence of 40 Prisoner's dilemma games is used in the experiment. All of the games used had the payoff matrix given in Figure 2. At the beginning of the series there were

5 training games (results from these games are not included in the analysis) thus the total sequence comprised 45 games.

		Player II	
		C	D
Player I	C	40, 40	10, 50
	D	50, 10	15, 15

Figure 2: Payoff table for the PD game used in the experiment.

## Experimental Conditions

The **distribution of the total payoff** is varied in accordance with the four relational models described above in a between-subjects design. There are four experimental conditions that differ in the way that the total payoff of a pair is divided between the players in that pair:

- **Unity condition** – each player receives the total payoff earned by the pair (*communal sharing* relational model);
- **Hierarchy condition** – one of the players receives more than the other – 2/3 vs. 1/3 of the total payoff of the pair (*authority ranking* relational model);
- **Equality condition** – each player receives equal portion of the total payoff (*equality matching* relational model);
- **Proportionality condition** – each player receives a share of the total payoff proportional to his/hers individual payoffs (*market pricing* relational model).

## Procedure

Subjects were tested in pairs. After receiving the appropriate instructions for the experimental condition they were in, each dyad played 5 training games, followed by 40 games that were analyzed. The experimenters secured that the participants will not have visual, verbal and any kind of other contact before and during the experiment. Therefore, no player knew who the other player was before the end of the experiment.

Instructions for the experiment explained in details the rules of the game and included several test questions to make sure that the participants understood correctly the rules. There were four instructions that varied only in the explanation for the total monetary payoff distribution. They are quoted below because they define the relational models in the four conditions:

- *Unity condition* – 'Each of you will receive the amount of money you have earned together';
- *Hierarchy condition* – 'You will get 2/3 of the total amount of money of the pair, and the other player will get 1/3 of it (for one of the players). You will get 1/3 of

the total amount of money of the pair, and the other player will get 2/3 of it (for the second player)';

- *Equality condition* – ‘The total amount of money of the pair will be split equally between you and the other player’;
- *Proportionality condition* – ‘The total amount of money of the pair will be split between you and the other player in accordance with the number of points each of you has earned’.

The game was presented in a formal and a neutral formulation. On the interface, the cooperation move was labeled ‘1’ and the defection move was labeled ‘2’. However, further in the paper, for convenience, we will continue to use *cooperation* instead of move ‘1’ and *defection* instead of move ‘2’. Matlab 7.6.0 (R2008a) was used for presenting the game and recording the choices of the players.

After each game the subjects got feedback about their own and the other player’s choice and payoffs in the current game. They could also constantly monitor their own total payoff; the total payoff of the other player; the total payoff for the pair, and the monetary equivalent of the total payoff of the pair (that is to be distributed among them).

Participants were instructed to try to *maximize* the amount of money they will get. Subjects were paid real money accordingly to the final payoff in the game. Players in the *unity* condition received the same amount of money for 1000 points as participants in the other 3 conditions received for 500 points. Thus we tried to equate the absolute magnitude of the monetary payoff that the participants could receive during the experiment.

Each session lasted about 20 minutes.

## Participants

80 participants (47 female, 33 male) took part in the experiment. All of them were university students, mean age 23.3 years.

They were tested in 40 pairs – 10 pairs in each experimental condition. Participants were randomly assigned to the experimental condition. In the *hierarchy condition*, it was randomly determined which player will get 1/3 and which player – 2/3 of the total payoff of the pair.

Subjects who have previously played the Prisoner’s dilemma game were not allowed to participate in the study.

## Results

To explore the influence of payoff distribution model on choices and cooperation in the PD games, the following dependent variables are analyzed: number of **cooperative choices for each player**; number of games with **mutual cooperation in a pair**; number of games with **mutual defection in a pair**. In the figures results are presented in percentages for clarity. However, the analysis is performed using the specified dependent variables.

Average **total payoff for a dyad (in points)** is analyzed to assess which type of payoff distribution led to higher profits.

Each dependent variable is analyzed in ANOVA with distribution model as between-subject factor with 4 levels (*unity* vs. *hierarchy* vs. *equality* vs. *proportionality*).

## Cooperation

The cooperative choices (%) for each distribution type are presented in Figure 3. The analysis shows a significant influence of the distribution type on the number of cooperative moves ( $F(3, 76) = 4.49, p = 0.006$ ).

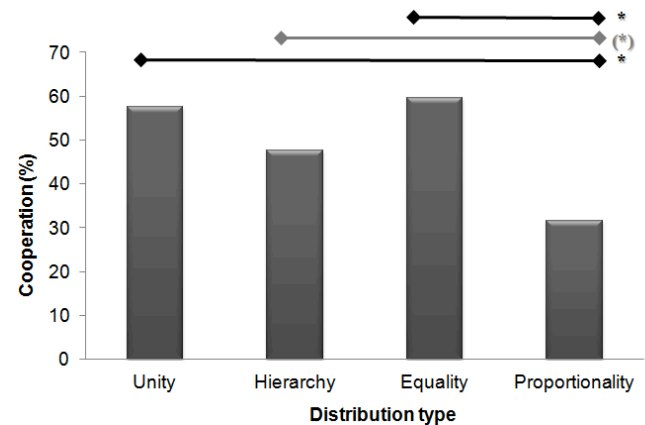


Figure 3: Average percentage of cooperative choices in each distribution condition (\* means  $p < 0.05$ ; \*\* – marginally significant difference).

Post-hoc LSD test shows that the cooperation rate in the *proportionality condition* is significantly lower than the cooperation rate in the *unity condition* ( $p = 0.003$ ) and in the *equality condition* ( $p = 0.002$ ). The difference between cooperation rates in *proportionality condition* and *hierarchy condition* is marginally significant ( $p = 0.065$ ). All other differences are non-significant.

It seems that the type of distributional model influences the cooperation rate. Cooperation is lower when each player gets a portion of the total payoff that is proportional to his/her payoffs during the game. In the terminology of the Fiske’s theory, the market pricing relational model leads to diminished cooperation in comparison to the other three relational models. When the final payoff for the player depends on his/her individual results, the choices are more non-cooperative in comparison to the cases in which the total payoff of the pair is divided between players (no matter in what predefined proportions) and when each player received the total amount earned by both of them.

## Mutual Cooperation

Average percentage of games in which there is mutual cooperation (both players have chosen to cooperate) is presented in Figure 4.



The ANOVA does not identify a statistically significant influence of the distribution type on the number of mutual cooperative game outcomes ( $F(3, 36) = 1.94, p = 0.141$ ). However, a further conducted Post-hoc LSD test shows that difference exists between the *proportionality* and *equality condition* ( $p = 0.038$ ). A marginally significant difference is observed between the *proportionality* and *unity condition* ( $p = 0.079$ ).

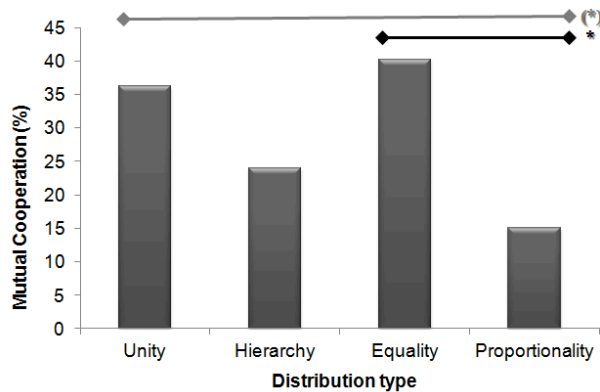


Figure 4: Average percentage of mutual cooperation in a pair in each distribution condition (\* means  $p < 0.05$ ; (\*) – marginally significant difference).

It turns out that mutual cooperation is lower when the payoff of each player depends on his individual contribution compared to the situations in which players divide their joint winnings in equal shares or receive the total amount earned. This is a rather logical result in line with the assumption of the current study that the distribution type representing communal sharing or equality matching will promote collectivistic orientation; while money pricing relationship will most probably trigger an individualistic behavior among subject.

### Mutual Defection

Average percentage of games with mutual defection (both players have chosen to defect) is presented in Figure 5.

For the number of games with mutual defection the ANOVA shows a significant influence of the distribution type ( $F(3, 36) = 3.943, p = 0.016$ ). Post-hoc LSD test identifies differences between the *proportionality condition* and every other condition in the experiment – *unity* ( $p = 0.006$ ), *hierarchy* ( $p = 0.032$ ), and *equality* ( $p = 0.005$ ).

Therefore, it can be concluded that when distribution of payoff is conducted according to individual results, mutual defection is a much more typical choice in comparison to all other distribution cases. In all other conditions this outcome (mutual defection) is relatively low. It should be noted that mutual defection leads to the lowest possible payoff for the pair. As noted in the introduction, the defection is dominant strategy for each individual player; however, mutual defection leads to the worst possible payoff for the society –

thus the dilemma structure of the game arises as the opposition between individual and collective rationality.

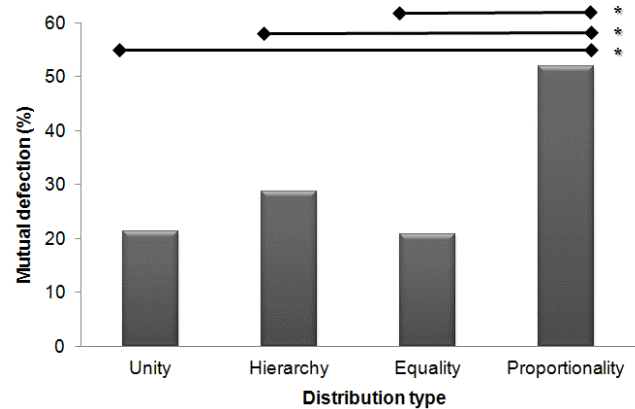


Figure 5: Average percentage of mutual defection in a pair in each distribution condition (\* means  $p < 0.05$ ).

### Average Payoff

The payoff analysis was conducted on the basis of the average payoff per pair (in points) for the sequence of 40 games (Figure 6). The ANOVA shows a significant influence of the distribution type on the payoff ( $F(3, 76) = 3.271, p = 0.026$ ).

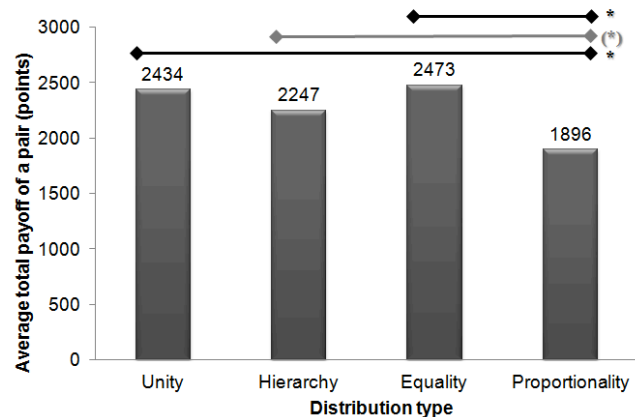


Figure 6: Average payoff per sequence of 40 games for a pair in each distribution condition (\* means  $p < 0.05$ ; (\*) – marginally significant difference).

Significant differences were established through post-hoc LSD test between the *proportionality* and *unity condition* ( $p = 0.011$ ) and between the *proportionality* and *equality condition* ( $p = 0.006$ ). Marginally significant is the difference between the *proportionality* and *hierarchy condition* ( $p = 0.093$ ).

The total payoff of a pair is lower when its distribution among the players depends on the individual contribution and points earned. This is a rather paradoxical result taking

into account that such a distribution is representing the market pricing relational model, which is mostly related to individualistic attitude and profit orientation. However, taking into account that in the *proportionality condition* there is the highest number of games with mutual defection, the result is not surprising and could be explained by the lower payoff that the players get when both are non-cooperative.

## Conclusions and Discussion

The current paper presents an experimental study on the phenomenon of cooperation in Prisoner's dilemma game in the light of Fiske's Relational models theory. The experiment was designed to explore the influence of the different distribution types of the total payoff (thus reflecting the four elementary human relations according to Fiske's theory) on the level of individual and mutual cooperative and non-cooperative behavior.

The results show that the distribution type corresponding to the individualistic Market Pricing relational model is characterized by a lower cooperation, lower mutual cooperation, higher mutual defection and lower total payoff of the participating subjects in comparison to the other distribution situations. When players are rewarded based on their individual results, they cooperate less and receive lower payoffs (for each player and for the dyad of players). This is an interesting result taking into account the fact that in formal game theory, in many experiments, in many real life situations, the players are perceived as individualistic beings. And there are attempts to apply policies aimed at achieving higher collective payoff by profit distribution accordingly to the contribution of each individual. Current results demonstrate that such distribution in fact leads to significantly lower total collective earnings.

An interesting topic for reflections and further research is whether a behavioral model representing competitiveness and profit-orientation might actually be effective in achieving its goals within the Prisoner's dilemma game model, respectfully within real life situations depicting this model.

Another related direction for exploration is *hierarchy condition*. Generally speaking, this condition could lead to higher cooperation despite the difference in received payoffs if the relation between the players is perceived as an in-team relation. This could be achieved by justifying the role attribution in this condition with some game related advantage (e.g. a better strategy in the test games).

A broader implications of these results might be applicable not only to studies of decision making in games, but also to socio-economical policies employed by organizations, governments, etc.

## References

Andreoni, J., & Miller, J. (1993). Rational cooperation in the finitely repeated Prisoner's dilemma: experimental evidence. *The Economic Journal*, 103, 570-585.

- Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211, 1390-1396.
- Colman, A. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences*, 26, 139-198.
- Cooper, R., DeJong, D. V., Forsythe, R., & Ross, T. W. (1996). Cooperation without reputation: Experimental evidence from Prisoner's dilemma games. *Games and Economic Behavior*, 12, 187-218.
- Dawes, R. (1980). Social dilemmas. *Annual Review of Psychology*, 31, 169-193.
- Durkin, J., Frost, M., Aronow, E., & Breslow, M. (1967). Moment of truth encounters in prisoner's dilemma. *American Psychologist*, 22, 595.
- Fiske, A. P. (1991). *Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing*. New York, NY: Free Press.
- Fiske, A. P. (1992). The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological Review*, 99, 689-723.
- Fiske, A. P., & Haslam, N. (1996). Social cognition is thinking about relationships. *Current directions in psychological science*, 5, 143-148.
- Gärdenfors, P. (2003). *How Homo became Sapiens*. Oxford University Press.
- Grinberg, M., Hristova E., & Lalev, E. (2010) Models for cooperative decisions in prisoner's dilemma. In: Nefti, S. & Gray, J. (Eds.), *Advances in Cognitive Systems*. IET: London.
- Haslam, N. (Ed.) (2004). *Relational models theory: a contemporary overview*. Mahwah, NJ: Erlbaum.
- Komorita, S., Hilty, J., & Parks, C. (1991). Reciprocity and cooperation in social dilemmas. *Journal of Conflict Resolution*, 35, 494-518.
- Kreps, D., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated Prisoner's dilemma'. *Journal of Economic Theory*, 27, 245-252.
- McGraw, A. P., Tetlock, P. E., & Kristel, O. V. (2003). The Limits of fungibility: relational schemata and the value of things. *Journal of Consumer Research*, 36, 219-229.
- Rapoport, A., & Chammah, A. (1965). *Prisoner's dilemma: a study in conflict and cooperation*. Univ. of Michigan Press.
- Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, 117, 57-75.
- Sally, D. (1995). Conversation and cooperation in social dilemmas. A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, 7, 58-92.
- Sally, D. (2001). On sympathy and games. *Journal of Economic Behavior & Organization*, 44, 1-30.
- Trivers, R. (1972). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46, 35-37.



# Evaluating the Relationship Between Neuropsychological Function and Cognitive Performance

**Glenn Gunzelmann (glenn.gunzelmann@us.af.mil)**

Cognitive Models and Agents Branch, Air Force Research Laboratory  
2620 Q St, Building 852, Wright Patterson AFB, OH 45433 USA

**L. Richard Moore, Jr. (Larry.Moore.ctr2@wpafb.af.mil)**

L3 Communications  
5950 East Sossaman Road, Suite 102, Mesa, AZ 85212 USA

## Abstract

The last 2 decades have produced a vast literature describing relationships between cognitive performance and neuropsychological data. This literature has provided the foundation for countless theories about the neural correlates of cognitive processing and specific theories regarding the role of different cortical areas in human cognition. In this paper, we examine a particular theory – the error likelihood model (Brown & Braver, 2005) – that attempts to account for the function of a particular brain area (the anterior cingulate cortex). A careful evaluation of behavioral data from humans raises questions about the error likelihood model and the implications of neuropsychological data for understanding cognitive performance.

**Keywords:** Neural Correlates; Anterior Cingulate; Error Likelihood; Cognitive Performance; Change Signal.

## Introduction

The last 2 decades have produced a vast literature describing relationships between cognitive performance and neuropsychological data. This literature has provided the foundation for countless theories about the neurological correlates of cognitive processing and specific theories regarding the role of different cortical areas in human cognition. These theories have had a tremendous impact on cognitive science, as well as the perceptions of the general public about the relationship between neural activity and cognitive processing.

The debate surrounding the role and utility of neuropsychological data in understanding human cognition has been ongoing (e.g., Cacioppo, Berntson, Lorig, Norris, Rickett, & Nusbaum, 2003; Coltheart, 2006; Henson, 2006; Uttal, 2001). Whereas evidence of neural correlates have been found in a variety of contexts (e.g., Cabeza & Nyberg, 1997; 2000), direct mappings of information processing activity to particular brain areas may be too simplistic (c.f., Horgan, 1999; Hubbard, 2003; Sohrabi & Brook, 2005). Instead, we argue that it is necessary to understand in detail both the cognitive behavior and the neuropsychological evidence to accurately understand the relationships between neural activity and cognitive processing.

In this paper, we consider a particular example of this complex relationship. We begin with a description of a task – the *change signal task* – which has been used in research attempting to understand the function of the anterior cingulate cortex (ACC) in humans (Brown & Braver, 2005).

Brown and Braver (2005) used fMRI data from participants performing this task to support a model of ACC function they refer to as the *error likelihood model*.

We conducted an extension of Brown and Braver's study using the same task and present the empirical data from human participants here. A detailed analysis of the change signal task and the human performance data provides alternative explanations for most of the human data captured by the error likelihood model, and raises some cautions for those attempting to interpret the significance of neuropsychological data for understanding the underlying cognitive processes of human cognition.

## The Change Signal Task

The change signal task is a modification of the stop signal task from Logan and Cowan (1984), which Brown and Braver (2005) used in an fMRI study to examine the function of the ACC in responding to potential errors. In the task, participants are presented with an arrow on each trial, which points either to the left or the right. This is the *go signal*. Critically, on 33% of the trials, a second arrow facing in the opposite direction (the *change signal*) is presented at a carefully controlled delay (the *change signal delay*) relative to the onset of the first arrow. In trials where this arrow appears, participants are instructed to withhold their initial response, and make the response associated with the change signal instead.

The change signal delay is manipulated throughout the task to ensure a relatively constant error rate, however, this characteristic of the task is not revealed to the participants. In Brown and Braver (2005), two stimulus colors were used, and the change signal delay was manipulated independently for each of the color conditions to produce different error rates (error likelihood conditions). In one, the change signal delay tended to be longer, leading to a higher error likelihood, while the other condition tended to have shorter change signal delays with a correspondingly lower error likelihood.

The change signal delay was 250ms in both conditions at the start of the study for all participants. Correct responses led to an increase in the change signal delay; 2ms for the low error likelihood condition, and 50 ms for the high error likelihood condition. In both conditions, errors led to a 50 ms decrease in the change signal delay. These parameters were intended to produce error rates of 4% and 50% in the

low and high error likelihood conditions, respectively. Finally, the change signal delay was constrained to be between 20ms and 800ms, and responses taking longer than 1000ms after the go stimulus presentation were identified as lapses and treated as errors. This last manipulation prevented people from waiting for arbitrarily long periods before making their responses.

### Experiment in Brown and Braver (2005)

Brown and Braver (2005) conducted an empirical study to assess the role of the ACC in performing the change signal task. In it, participants completed an average of 535 trials of the task in a single session. While doing the task, fMRI data was collected. Brown and Braver (2005) did not consider in detail the performance data from the study, instead focusing on the fMRI results and their error likelihood model. They did, however, provide supplementary materials that include some additional consideration of the behavioral results.

The change signal task offers interesting challenges for human cognition, and the results presented in Brown and Braver (2005) show that the ACC is sensitive to the differences between go and change trials as well as the error likelihood conditions. Our analysis of the task and data from a replication, however, suggest that many of the findings may reflect artifacts of the task, rather than revealing critical differences in the underlying cognitive processing across conditions by the participants in the study. Before describing this in detail, we provide an overview of Brown and Braver's (2005) error likelihood model, and the relationships between the mechanisms in the model and the fMRI data they presented.

### The Error Likelihood Model

The Brown and Braver (2005) error likelihood model presents a theory of ACC function embodied in a neural network-based computational model. The model posits that the ACC functions to detect the likelihood of an error, given a particular task and stimulus context. As they put it, "[the] ACC learns to signal, via the magnitude of its activity, the predicted likelihood of an error occurring in response to a given task condition" (Brown & Braver, 2005, p 1120). They also describe how this conceptualization of ACC function can account for conflict and error detection phenomena that have been shown in ACC activation patterns (e.g., Botvinick, Braver, Barch, Carter, & Cohen, 2001; Cohen, Botvinick, & Carter, 2000).

Brown and Braver suggest that detecting the likelihood of error plays a key role in cognitive control by serving as an "early warning signal" that can be used to recruit resources for performing the task. Thus, a central claim in their theory and model is that ACC activation is used by higher-level cognitive processes to guide adaptive behavior in the task and improve cognitive performance.

The details of the model are beyond the scope of this paper. However, it makes several important predictions in the context of the change signal task. Most intuitively, it predicts that ACC activation should be higher for the high

error likelihood condition than for the low error likelihood condition. The authors discuss this effect in their model as a consequence of learned associations between the stimulus color and the likelihood of an error.

In addition to the predicted differences in cortical activation for the error likelihood conditions, the model also predicts differences between *change trials* (where a change signal is presented) and *go trials* (where no change signal is presented), with higher activation for change trials due to the signal these trials provide for reinforcement learning processes in the ACC. The fMRI data from humans show the same qualitative trends, providing support for the model.

A critical finding in support of the error likelihood model was that ACC activation was higher for go trials in the high error likelihood condition than it was for go trials for the low error likelihood condition. It is argued that sensitivity to the stimulus color is responsible for this effect, since these trials are equivalent in all other respects. This is also the primary finding that differentiates the error likelihood model from an alternative account, the response conflict model (Botvinick et al., 2001).

### Empirical Study

To better understand human performance in the change signal task, we conducted our own empirical study to obtain detailed data on task performance. In addition to the change signal task, participants performed a two-alternative forced choice (2AFC) task that matched the change signal task, only without any change signals. One motivation for this design was to investigate the role of within task fatigue on changes in response time in the change signal task (see Moore, Gunzelmann, & Brown, 2010).

### Methodology

There were 33 participants in the study (18 female and 15 male; ages between 18 and 50). Each participant performed both tasks in a single session lasting approximately 1 hour (task order was counterbalanced). The design of the change signal task replicated the study described in Brown and Braver (2005), except that our participants performed more trials. Specifically, participants completed 6 blocks of 107 trials for a total of 642 trials in our study.

In Brown and Braver's experiment, the association between stimulus color and error likelihood condition was swapped after participants had completed approximately 80% of the trials. This occurred in our experiment at the midway point. Just as in Brown and Braver (2005), this switch in associations between color and error likelihood condition was not signaled to participants. Only one participant in our study reported noticing this manipulation. In fact, only 9 participants were able to accurately articulate the significance of the stimulus colors in the experiment at all.

### Results

We collected accuracy and response time data from participants performing the task. Unless noted otherwise, the

results presented here only include data from trials where correct responses were made. Furthermore, the data from two participants were excluded from the analyses because one failed to complete the 2AFC task, and the other exhibited an unusually long string of incorrect responses during the change signal task. As in Brown and Braver (2005), trials where no response was made within 1000ms of the onset of the go stimulus were aborted and treated as errors, with change signal delays adjusted accordingly.

Figure 1 shows median participant response times relative to the presentation of the initial go stimulus across each of the 6 blocks of trials. Firstly, response times for the 2AFC task are stable across all blocks, showing no evidence of within-task fatigue during the course of the experiment,  $F(1,19723)=3.144$ ,  $p=.076$ .

Beyond the 2AFC, the results for the change signal task are generally consistent with those obtained in the Brown and Braver (2005) experiment, and there are several features that will be relevant for the rest of the analyses and discussion that follows. First, note that the response times for the change signal task are consistently much longer than for the 2AFC, and that there is a wide disparity between the change high and change low conditions. We will demonstrate how these phenomena are related to the dynamics of the task. The closely coupled go low and go high conditions will also be discussed, and it will be shown that participant behavior was indistinguishable between the two.

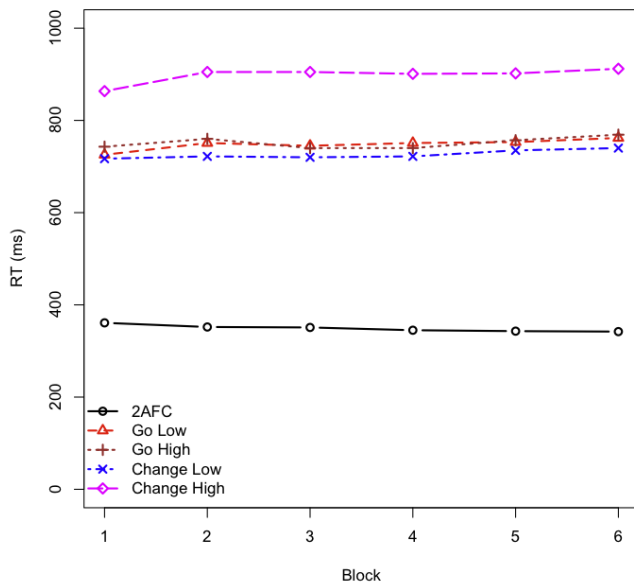


Figure 1: Median reaction times measured from the go signal for the four conditions of the change signal task as well as the 2AFC.

Figure 2 shows a significant ( $F_{Low}(1,3180)=444.4$ ,  $p < .001$ ,  $F_{High}(1,1643)=374$ ,  $p < .001$ ) correlation ( $r = .77$ ) between the change signal delay and the participant reaction time in the change trials. These data illustrate the impact of the change signal delay on overall response times shown in

Figure 1. In fact, if the change signal delay is subtracted from the response times on change trials, the disparity between these conditions nearly disappears (Figure 3).

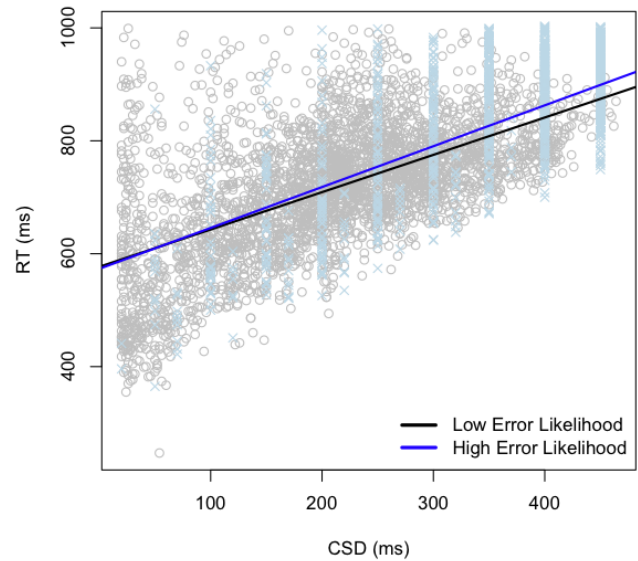


Figure 2: Reaction time as a function of change signal delay in correct change signal trials. Regression lines overlay the lighter scatter plots of each condition.

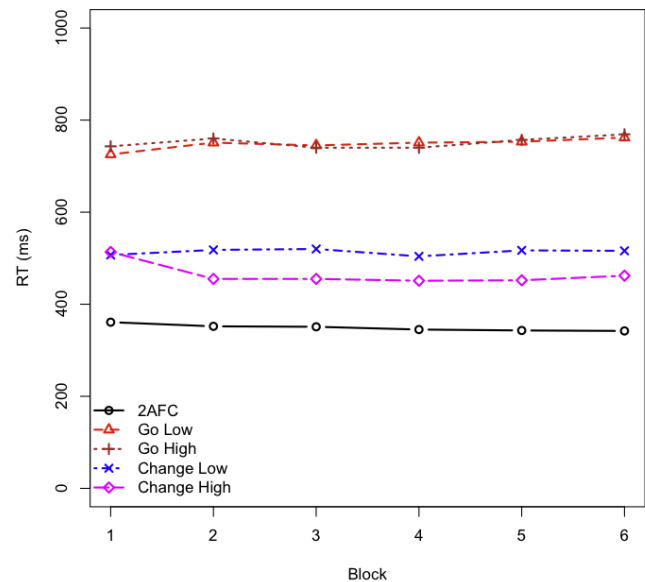


Figure 3: Median reaction times measured from the final stimulus for the four conditions of the change signal task as well as the 2AFC (i.e. the change signal delay has been subtracted from the high error condition reaction times).

One clear consequence of factoring out the change signal delay on change trials is that response times are significantly faster for change trials than for go trials ( $t(10836.23) = -17.2995$ ,  $p < .0001$ ). This effect can be explained in context as a strategic adaptation to the characteristics of the change signal task. Specifically, it is our hypothesis that participants

are intentionally delaying their responses to go signals in hopes of correctly responding to the change signals (Moore, Gunzelmann, & Daigle, 2012).

This perspective accounts for the slower reaction times in the go trials, because it suggests that participants would respond to the go signal only after their strategic delay, or hedge, was complete. It also explains the large difference in response times for the go trials in the change signal task versus response times for the 2AFC task (Figures 1 & 3).

Although we hypothesize that participants also hedge their response in the change conditions, there is no reason for them to delay making a response once a change signal is presented. In change trials, therefore, responses can be initiated as soon as the change signal appears. Moore et al. (2012) present a computational model demonstrating the plausibility of this account.

Another interesting feature in Figure 3 is that median response times for change trials are faster in the high error likelihood condition than in the low error likelihood condition. To understand this average difference, it is necessary to examine the details of human performance in these cases and the characteristics of the task that give rise to the observed results. Figure 4 shows the distribution of response times for the change high and change low conditions (left side), as well as the proportion of lapses in each of the conditions (right side).

Figure 4 illustrates that the difference in response times between the two change conditions in Figure 3 is likely a function of the 1000 ms lapse threshold. In the high error likelihood condition, 15 % of the trials resulted in lapses (see the right half of Figure 4), while only 1% of trials in the low error likelihood condition resulted in lapses. The right side of Figure 4 gives clear evidence that the response time distribution is truncated at the lapse threshold, which has the effect of reducing the median response time for correct responses (lapses are treated as errors).

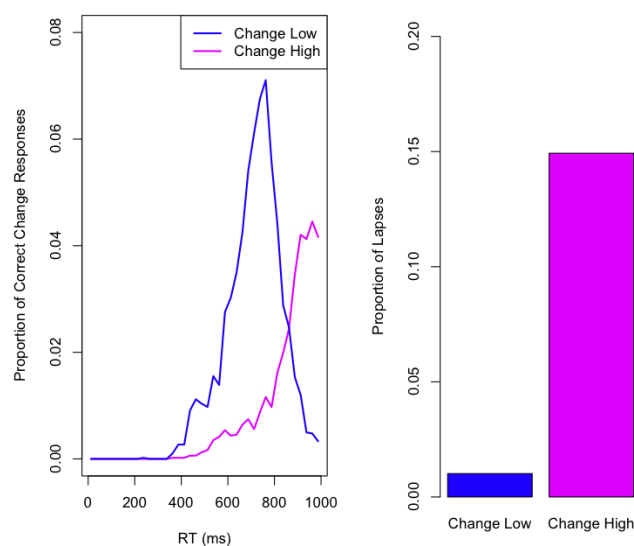


Figure 4: Response distribution for the high and low change conditions, and the proportion of lapses for each.

Lastly Figure 5 shows the distributions for the two go conditions. They align extremely well, and Kolmogorov-Smirnov test shows no statistical difference ( $p=.36$ ). As discussed below, this is an important result, as it generates questions regarding the extent to which people are aware of the significance of the stimulus colors in the task, or the degree to which they are able to use the colors in a meaningful way to adapt to the characteristics and demands of the task. In the next section we compare and contrast the results of our study with the fMRI data described in Brown and Braver (2005).

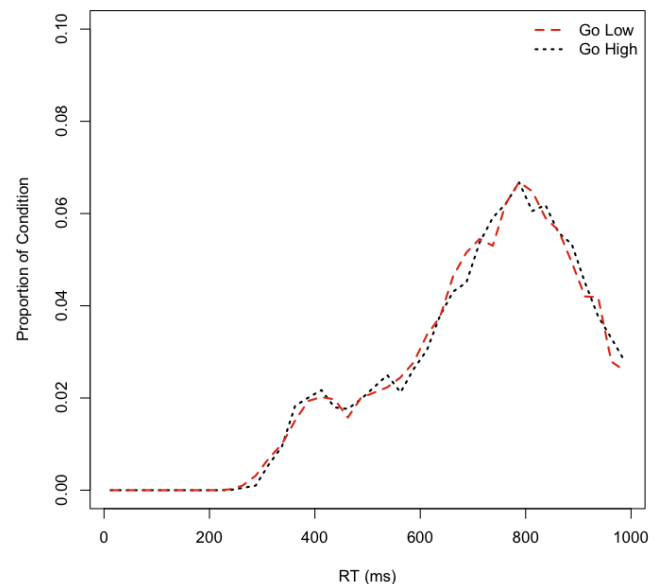


Figure 5: Distribution of response times for go trials in the two error likelihood conditions (accumulated into 25ms bins).

## ACC Activation and Error Likelihood

Evidence for the error likelihood model in Brown and Braver (2005) came from fMRI data. However, to fully evaluate the validity and significance of the model, careful analysis of the task and consideration of human performance data are necessary. Based on the analyses of the response time data presented above, we discuss the fMRI findings in this section, showing that nearly all of the critical phenomena in the fMRI data from the task can be accounted for by simply assuming that longer trials (including the change signal delay) lead to more ACC activation (though this may be too simplistic in general; see Mulert, Gallinat, Dorn, Herrmann, & Winterer, 2003; Winterer, Adams, Jones, & Knutson, 2002).

## Task Artifacts in Performance and ACC Activation

Brown and Braver (2005) cited several findings in their fMRI data to support the error likelihood model of ACC function. Most of these can be related directly to task-driven differences in response time we found in our study. One

example of this is the finding in Brown and Braver (2005) that activation in the ACC was higher for change trials than for go trials. We did find longer response times for change trials, but the evidence suggests that this difference is an artifact of the task-driven change signal delay (Figure 3).

In addition to the main effect of trial type (change versus go), Brown and Braver (2005) also found greater ACC activation for change trials in the high error likelihood condition than in the low error likelihood condition. Once again, this is associated with a significant difference in response time (Figure 1). The response time difference, in turn, is driven by the difference in the change signal delay between the two conditions. When those delays are factored out of the response times, the difference between those conditions disappears (Figure 3).

### The Critical Phenomenon

The only phenomenon that is not captured well by the timing of the presentation of stimuli in the task is the difference in ACC activation between go trials in the two error likelihood conditions that is predicted by the error likelihood model and supported by the fMRI data in Brown and Braver (2005). In this case, the behavioral data diverge from those trends. In fact, our data show essentially equivalent performance, with no significant difference in behavior across conditions (Figure 5).

This result creates an interesting circumstance with regard to assessing the significance of the fMRI data and the implications of the error likelihood model for understanding human cognition. On the one hand, the fMRI data show a significant difference in ACC activation between go trials from the two error likelihood conditions. Importantly, the error likelihood model predicts the fMRI data well, providing a consistent account of neuropsychological data. This is an interesting capacity of the model, and one that adds support to the proposed mechanisms.

On the other hand, while the fMRI data and the model both suggest that the error likelihood conditions are differentiated at a neuropsychological level, there is no evidence that they are differentiated at a behavioral level in our data. Brown and Braver (2005) take the position that the ACC is critical in the recruitment of cognitive control during task performance when the likelihood of making errors is greater. In this case, color provides a cue to differences in difficulty, albeit a cue that is not explicitly described to participants. Importantly, others have failed to replicate the fMRI findings reported by Brown and Braver, even with more explicit cues regarding the error likelihood cues and their significance (Nieuwenhuis, Schweizer, Mars, Botvinick, & Hajcak, 2007).

The critical question is, if the ACC is sensitive to the color of the stimulus as an indication of the likelihood of making an error, why is there no evidence in the behavioral data? The answer to this question is essential to understanding the relationship between cognitive processing and neuropsychological data in the change signal task. We conclude the paper with a discussion of this issue, and more

generally the challenges associated with using neuropsychological findings to inform our understanding of cognitive processes.

### Conclusions and Implications

Brown and Braver (2005) presented provocative neuropsychological data from a novel task, which they used to validate a computational theory of ACC function. As our results and analyses show, however, questions remain about whether it is task artifacts or cognitive phenomena that are responsible for many trends in the fMRI data, and about the implications of the data and the error likelihood model for understanding human cognitive performance and behavior.

Importantly, the model accounts for what appears to be a critical phenomenon in the empirical study – higher ACC activation for go trials in the high error likelihood condition than in the low error likelihood condition. This is, in fact, the only phenomenon predicted by their model that cannot be explained by the timing of the presentation of stimuli in the task, which directly impacts response times as well. Unfortunately, others have failed to replicate that finding (Nieuwenhuis et al., 2007).

Even if the effect is real, questions remain about what these results mean with regard to the underlying cognitive processes. According to Brown and Braver (2005), the ACC is an “early warning system that recruits cognitive control to match its predicted demand” (p.1120). In the context of the change signal task, however, one would expect that recruiting cognitive control would (1) increase explicit awareness about features in the environment related to error likelihood and/or (2) impact human behavior in a manner consistent with the implications of the likelihood of error.

In support of these expectations, Dehaene et al. (2003) found evidence for elevated ACC activation only in circumstances where stimuli creating conflict in a priming task were “consciously detected” (p. 13726). Based upon our results, however, the manipulation of error likelihood was not obvious to participants, and there was no impact on task performance. This creates some challenges that must be addressed to better understand the cognitive processing involved and the significance of ACC activation in the task.

There is evidence in the change signal task that participants adapt to the change signal delay. As they gain experience with the task, average response times increase, reflecting strategic adaptation to the task. However, there is *no* evidence that their adaptation is sensitive to the distinction between error likelihoods signaled by the two stimulus colors. Instead, reaction times for go trials are virtually identical, regardless of the error likelihood condition (Figure 5). This is also true of change trials, when the change signal delay and truncated response distribution in the high error likelihood condition are taken into account (Figures 3 & 4). An interesting follow-up would be to examine human performance if the role of the colors was explicitly explained to participants before the study began (see Nieuwenhuis et al., 2007 for an experiment along these lines).

Of course, this leaves the incongruity between ACC activation and participant performance in these two conditions begging for a theoretical explanation, in addition to questions regarding the replicability of the phenomena (Nieuwenhuis et al., 2007). Our findings expose the discrepancy and reveal the importance of understanding this finding. And, we hasten to add that our empirical findings do not provide evidence to directly contradict the error likelihood model of Brown and Braver (2005). Taken with the failure to replicate the fMRI findings (Nieuwenhuis et al., 2007), however, there is an indication that further research is warranted to understand human performance on the task and the role of the ACC.

Finally, our results suggest in general that fMRI data, like the results presented in support of the error likelihood model, must be interpreted with caution and considered in the context of the performance of participants as well as the context of the task environment that is the focus of study. We have shown that these factors can add important information to inform theories regarding the relationship of neuropsychological data to cognitive processes. It is only by considering multiple sources of evidence that we will be able to arrive at comprehensive theories of human information processing and cognition and how those functions are realized in the brain.

### Acknowledgments

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. This research was sponsored by the Air Force Research Laboratory's Warfighter Readiness Research Division and by grants 09RH06COR and 10RH04COR from the Air Force Office of Scientific Research (AFOSR). We thank Marissa Daigle for data collection.

### References

- Botvinick, M., Braver, T., Barch, D. Carter, C. & Cohen, J. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108(3), 624-652.
- Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, 307(5712), 1118. doi:10.1126/science.1106642
- Cabeza, R., & Nyberg, L. (1997). Imaging cognition: An empirical review of PET studies with normal subjects. *Journal of Cognitive Neuroscience*, 9, 1-26.
- Cabeza, R., & Nyberg, L. (2000). Imaging cognition II: An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience*, 12, 1-47.
- Cacioppo, J. T., Berntson, G. G., Lorig, T. S., Norris, C. J., Rickett, E., & Nusbaum, H. (2003). Just because you're imaging the brain doesn't mean you can stop using your head: A primer and set of first principles. *Journal of Personality and Social Psychology*, 84(4), 650-661.
- Cohen, J. D., Botvinick, M. & Carter, C. S. (2000). Anterior cingulate and prefrontal cortex: Who's in control? *Nature Neuroscience*, 3, 421-423.
- Coltheart, M. (2006). What has functional neuroimaging told us about the mind (so far)? *Cortex*, 42, 323-331.
- Dehaene, S., Artiges, E., Naccache, L., Martelli, C., Viard, A., Schürhoff, F., Recasens, C., Martinot, M. L. P., Leboyer, M., & Martinot, J-L. (2003). Conscious and subliminal conflicts in normal subjects and patients with schizophrenia: The role of the anterior cingulate. *Proceedings of the National Academy of Science*, 100(23), 13722-13727.
- Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? *Quarterly Journal of Experimental Psychology*, 58A(2), 193-233.
- Henson, R. (2006). What has (neuro)psychology told us about the mind (so far)? A reply to Coltheart (2006). *Cortex*, 42, 387-392.
- Horgan, J. (1999). The undiscovered mind: How the human brain defies replication, medication, and explanation. *Psychological Science*, 10(6), 470-474.
- Hubbard, E. M. (2003). A discussion and review of Uttal (2001) The new phrenology. *Cognitive Science Online*, 1, 22-33.
- Logan, G. D. & Cowan, W. B. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review*, 91, 295-327.
- Moore, L. R., Jr., Gunzelmann, G., & Brown, J. W. (2010). Modeling statistical learning and response inhibition with the change signal task. In D.D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th International Conference on Cognitive Modeling*. doi:10.1126/science.1105783
- Moore, L. R., Jr., Gunzelmann, G., & Daigle, M. (2012). One Model, Two Tasks: Decomposing the Change Signal Task. In N. Rußwinkel, U. Drewitz & H. van Rijn (eds.), *Proceedings of the 11th International Conference on Cognitive Modeling* (pp. 224-229), Berlin: Universitaetsverlag der TU Berlin.
- Mulert, C., Gallinat, J., Dorn, H., Herrmann, W. M., & Winterer, G. (2003). The relationship between reaction time, error rate and anterior cingulate cortex activity. *International Journal of Psychophysiology*, 47, 175-183.
- Nieuwenhuis, S., Schweizer, T. S., Mars, R. B., Botvinick, M. M., & Hajcak, G. (2007). Error likelihood prediction in the medial frontal cortex: A critical evaluation. *Cerebral Cortex*, 17(7), 1570-1581.
- Sohrabi, A., & Brook, A. (2005). Functional neuroimaging and its implications for cognitive science: Beyond phrenology and localization. In B. G. Bara, L. Barsalou, and M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Meeting of the Cognitive Science Society* (pp. 2044-2049). Mahwah, NJ: Erlbaum.
- Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*. Cambridge, MA: MIT Press.
- Winterer, G., Adams, C. M., Jones, D. W., & Knutson, B. (2002). Volition to action – An event-related fMRI study. *Neuroimage*, 17, 851-858.



# The Interplay between Feature-Saliency and Feedback Information in Visual Category Learning Tasks

Rubi Hammer<sup>1</sup> (rubih@stanford.edu)  
Vladimir Sloutsky<sup>2</sup> (sloutsky@psy.ohio-state.edu)  
Kalanit Grill-Spector<sup>1</sup> (kalanit@stanford.edu)

1. Department of Psychology, Stanford University

2. Department of Psychology and Center for Cognitive Science, Ohio-State University

## Abstract

What is the role of feedback information in different visual category learning (VCL) scenarios? To address this question we tested participants' performance in VCL tasks in which stimuli varied in three feature dimensions, one of which was relevant for the task and the other two were irrelevant. The relevant feature could be identified based on trial-by-trial feedback. In one condition the task relevant and irrelevant features were highly-salient. In the second condition all features had low-visual-saliency. Feedback information was also manipulated: In the high-information condition the task relevant feature could be identified by the information provided in each trial whereas in the mid-information condition the feedback was ambiguous and information from several learning trials was required in order to confidently identify the relevant feature. Surprisingly, our data shows that mid- and high-information feedback are similarly effective in high-saliency VCL tasks. In contrast, in low-saliency VCL tasks, mid-information feedback impairs learning. We suggest that VCL can be done effectively either when feedback is ambiguous or in low-saliency conditions, but not in scenarios when both challenges occur concurrently.

**Keywords:** Visual category learning; Feedback information; Attentional learning; Perceptual learning; Feature-saliency.

## Introduction

Humans are capable of effectively managing a vast amount of sensory information, rapidly rendering it into a coherent, reliable and meaningful representation of objects and events. This capability depends on two fundamental learning processes: One is perceptual learning which allows identifying subtle, initially hard to detect, differences between stimuli (Goldstone & Barsalou, 1998; Kourtzi, 2010). The other is attentional learning which requires shifting attention to relevant attributes while at the same time filtering out irrelevant, even if salient, visual attributes (Blair, Watson & Meier, 2009; Kalish & Kruschke, 2000; Rehder & Hoffman, 2005). It is known that these two forms of learning allow reducing the probability of future decision errors by improving discriminability among similar objects from different categories, and allowing effective generalization to novel stimuli. To date, it is not clear how these processes interact in different learning scenarios.

Difficulty in perceptual learning tasks is determined by feature-saliency. Difficulty in attentional learning tasks is determined by the numbers of simultaneous perceptual features one has to process when categorizing objects.

Learning trajectories in both attentional and perceptual learning tasks are also determined by the availability of informative feedback. In this study we examine the interaction between perceptual and attentional learning by testing the interaction between feature-saliency and feedback information in visual category learning (VCL) tasks of complex visual stimuli.

We define feature-saliency as the physical dissimilarity between stimuli in a given feature dimension. When objects are perceived as highly dissimilar across a feature dimension, this feature is perceived as more diagnostic than lower-saliency feature dimensions (Chin-Parker & Ross, 2004). In each VCL task in our experiment stimuli differed from one another in three feature dimensions, yet only one was relevant for correctly categorizing the stimuli. In each task we kept the relative feature saliency (low or high) similar across all three feature dimensions making them equal candidates for being perceived as task relevant. Therefore, the diagnostic value of each feature could be determined only by the information provided by feedback.

We define the feedback information level based on its ambiguity. In each learning trial we presented a pair of stimuli and the participant had to determine if these stimuli belong to the same category or different categories. We used two levels of feedback information: 1) In the *high-information* learning condition the feedback in each and every trial provided sufficient information for learning the rule as *same-category pairs* were identical only in the task relevant feature and differed in the two irrelevant ones, whereas *different-categories pairs* were different only in the task relevant feature (and identical in the two irrelevant ones). If  $A$  denotes the relevant feature,  $B$  and  $C$  the irrelevant features, and  $X$  is the outcome of a categorization decision, the only possible trial-by-trial inferred causality (a feasible feature-decision association) in the high-information condition was  $A \rightarrow X \cap A \rightarrow X \cap A \rightarrow X \dots$  2) In the *mid-information* condition each trial was ambiguous as *same-category pairs* were identical in the task relevant feature and one of the two irrelevant features (randomly alternating between the two), whereas *different-categories pair* were different in the task relevant feature and one of the irrelevant features. The causality here was  $(A \cup B \rightarrow X) \cap (A \cup C \rightarrow X) \cap (A \cup C \rightarrow X) \cap (A \cup B \rightarrow X) \dots$  This learning scenario is more likely to require distributing attention between features and integrating information across more trials in order to learn the categorization rule.

We expect VCL efficiency to depend both on feature-saliency and the level of feedback information. Specifically, we hypothesize that when differences across visual feature dimensions are hard to detect, shifting attention from one feature to the other will not be effective due to the poor representation of features. This will become a greater challenge when feedback is ambiguous, making it harder to associate a feature representation with the corresponding decision outcome (Nosofsky & Palmeri, 1996). In contrast, when feature-saliency is high, relevant visual information is readily available, enabling to rapidly shift attention away from irrelevant features. This may allow effective learning even in the face of ambiguous feedback. Therefore, we expect an interactive effect of feature-saliency and feedback information. Nevertheless, based on current knowledge, we cannot predict whether this interaction will manifest as an additive effect or as an augmented interference in which the effect of feedback ambiguity on learning efficiency will be more profound in low-saliency learning conditions.

## Methods

### Participants

Sixty paid adults (36 females), with normal or corrected to normal vision, participated in the experiment. The experiment was approved by the Stanford University IRB.

### Materials and setting

We ran the experiment using Psychtoolbox (MATLAB®) on 1920X1200 pixels computer display. Participants' head was located about 70 cm (~2 feet) from the computer screen such that each one of the two simultaneously presented stimuli occupied approximately 14° of the visual field.

### Stimuli

We used four distinct sets of novel creature-like stimuli. In each set the stimuli varied in three feature dimensions (see examples in Figure 1). Exemplars for each stimulus set were produced from one standard object and three morph targets, each differed from the standard in one feature dimension. For VCL tasks with high-feature-saliency we used high morph values (at least 77%). Low-saliency exemplar pairs differed by small morph values (22-33%). The morphing values we used were determined based on pilot tests such that within each stimulus set differences in all feature dimensions were similarly likely to be detected. We also ensured that in the low-saliency condition differences within each feature dimension will be not detected easily without feedback. For each stimulus set we determined two categories. Members of each category varied in the two irrelevant feature dimensions and were identical in the third. This third feature-dimension was the diagnostic feature-dimension differentiating between the two categories.

### Design

Tasks differed in feature saliency (high-saliency vs. low-saliency), and three levels of feedback information. In high-saliency VCL tasks both within-category and between-

categories differences had high-saliency. In low-saliency VCL tasks both within-category and between-categories differences had low-saliency. In each VCL task participants were trained with one of three levels of feedback information: 1) High-level information feedback that potentially enabled identifying the diagnostic feature within each learning trial; 2) mid-level information feedback, in which each learning trial provided ambiguous information regarding which feature was the relevant one; 3) In addition to these two feedback-based learning conditions, we also tested participants in a control condition in which no feedback was provided to the participants. This provided a useful benchmark for assessing the contribution of feedback information to learning.

Each VCL task was based on a different stimulus set and a unique combination of feature-saliency and feedback information level. In order to prevent cross-conditions differences in categorization performances that derive from differences between stimulus sets, we counterbalanced the tasks across participants such that each one of the four stimulus sets was used in each one of the six experimental conditions the same number of times.

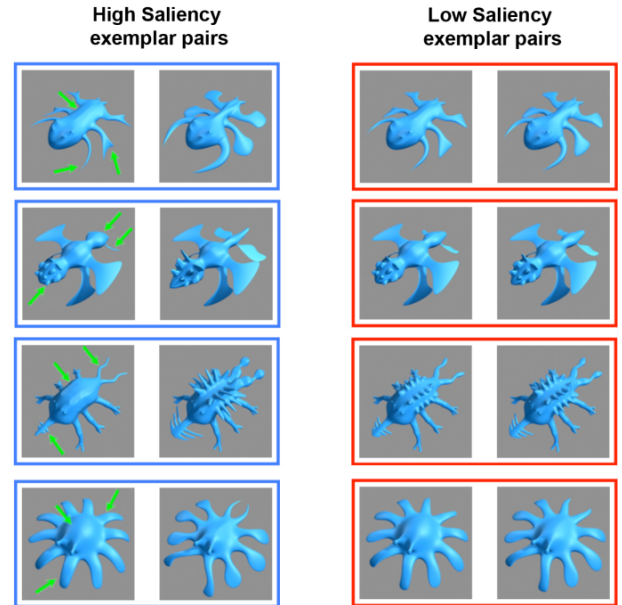


Figure 1: Examples of pairs of creatures from the four stimulus sets. Each row shows examples from a different stimulus set. *Left*: high-saliency pairs; *Right*: low-saliency pairs. Green arrows indicate high-saliency differences in the three feature dimensions in which the paired creatures differ (e.g., in the upper row the two creatures differ in horns, limbs and body-width). In each stimulus set, the same diagnostic feature was used for both the low- and high-saliency tasks. From top to bottom, the task relevant feature was body-width, head-spikes, horns and body curvature.

Each VCL task included seven blocks: four test blocks (denoted as T1-T4) that alternated with three learning blocks (L1-L3). Each block consisted of 24 trials. In each



trial two creatures were presented simultaneously on a computer screen for 2.2 seconds during which the participant had to decide if they belonged to the same or different categories by pressing one of two keyboard keys. Feedback was given during the 0.8 seconds inter-trial interval: In the mid- and high-information feedback conditions, a green square indicated a correct answer and red square an incorrect one. In the control, no feedback, learning blocks and in the test blocks, a yellow square indicated that the response was recorded (Figure 2a).

During test blocks, creature pairs always differed in two feature dimensions. *Same-category pairs* differed in the two irrelevant feature dimensions and were identical in the relevant one (as the right upper pair in Figure 2b). *Different categories pairs* differed in the relevant feature dimension and in one of the irrelevant feature dimensions, but were identical in the other irrelevant dimension (see left lower pair in Figure 2b). This design prevents participants from making the same/different categorization decision based on the overall similarity between stimuli. It also allowed us to keep the statistics of the three features and their pair-wise covariance identical.

For feedback-based learning blocks, pairs of creatures were selected in the following way: 1) In the high-information feedback condition *same-category pairs* were identical in the relevant feature dimension and differed in the two irrelevant feature dimensions. *Different-categories pairs* differed in the relevant feature dimension and were identical in the two irrelevant ones. Thus, each trial indicates either all the within-category variability (*same-category pairs*), or only the diagnostic feature dimension discriminating between categories (*different categories pairs*). 2) For mid-information feedback, *same-category pairs* were identical in the relevant dimension and in one of the two irrelevant dimensions (alternating between the two across different trials). *Different-category pairs* differed in two features: the relevant one and an irrelevant one (again, randomly alternating across trials between the irrelevant two). Although in this case each trial was ambiguous, it was still possible to learn the categorization rule based on the information provided across several trials. Such ambiguity in the feedback keeps the attentional learning aspect of the task more challenging by increasing the probability that participants will divide attention, in each trial, between two feature dimensions that are equally perceived as relevant. 3) The composition of the trials in the learning blocks with no feedback was similar to the one used in the test blocks.

## Procedure

To keep the duration of the experimental session short (~ 75 minutes), each participant performed a combination of three out of six conditions; [2 feature-saliency] X [3 feedback information]. Participants were informed that in each VCL task they have to learn to classify unfamiliar creatures from two distinct subspecies based on one attribute (feature dimension). Participants were also told that any variability in other attributes should be considered as irrelevant and

ignored. Before starting the experimental tasks, participants performed a warm-up VCL task (with a different stimulus set). This enabled the participant to become familiarized with the experimental tasks.

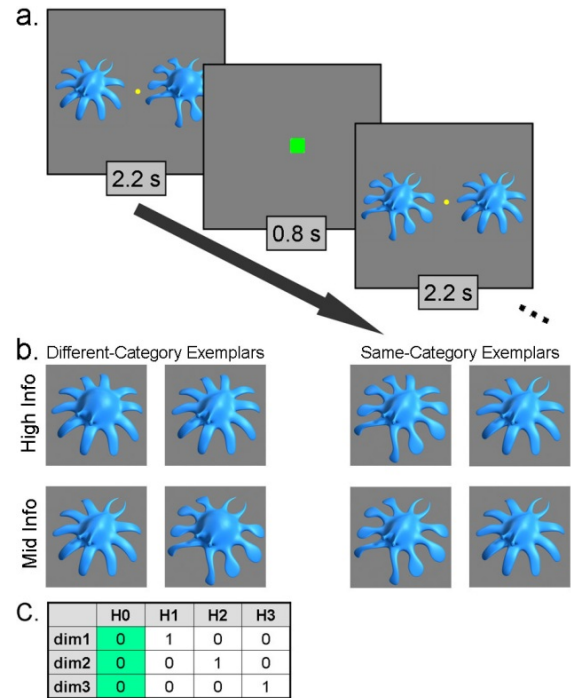


Figure 2: Experimental design. (a) An illustration of an experimental trial. A pair of stimuli presented for 2.2 seconds during which the participant had to judge if the creatures are from same or different categories. This was followed by 0.8 seconds of feedback presentation (e.g., green square indicates a correct answer) after which the next trial started. (b) Examples of different-categories (left) and same-category (right) pairs, high-information (top) and mid-information (bottom) condition. In each of the feedback based learning blocks, the category relation between the paired stimuli could be derived from the feedback. In High-information trials different-categories pairs differ only in the relevant feature, and same-category pairs are identical only in this feature. Such a trial enables effectively pinpointing the relevant feature dimension by eliminating 2 out of three possible hypotheses. Mid-informative trials provide less information since such a trial always leaves two options (out of three). (c) A table describing all possible hypotheses for a given VCL task. When all feature dimensions are salient the participant is only required to decide which feature is relevant (H1 – H3). When features are not salient, VCL requires shifting from H0 (represent a case in which the participant is unaware of either one of the potentially relevant feature dimensions) to the correct hypothesis signaling the relevant feature dimension.

## Performance measurements

We define a “Hit” as correctly deciding that two creatures are members of the same category, and a “False-Alarm” as

incorrectly deciding that creatures of different categories are members of the same category. Based on the Hit and False-Alarm rate we calculated participants' sensitivity using the non-parametric measure A-prime (Grier, 1971; Formula 1).  $A' = 1$  indicates perfect performance and  $A' = 0.5$  indicates chance level.  $0 \leq A' < 0.5$  represent a response confusion.

Formula 1: A-prime calculation.  $H$  denotes the Hit rate, and  $F$  the False-Alarm rate (Hit and False-Alarm rates are calculated based on the 24 trials in each test/learning block).

$$A' = 0.5 + \left[ \text{sign}(H - F) \times \frac{(H - F)^2 + |H - F|}{4 \times \max(H, F) - 4 \times H \times F} \right]$$

### Benchmarks

We evaluated participant performances according to the following benchmarks: Chance performance,  $A' = 0.5$ ; Perfect performance,  $A' = 1$ ; Performance based on systematically referring to an irrelevant feature during a test block or during a no-feedback "learning" block,  $A' = 0.12$ ; Performance based on systematically referring to an irrelevant feature during a learning block with mid-information feedback,  $A' = 0.5$ ; Performance based on systematically referring to an irrelevant feature during a learning block with high-information feedback,  $A' = 0$ .

## Results

Reported data is based on 24 participants in each condition. Excluded from this analysis are cases in which performance level was inconsistent between the test vs. the learning trials within a given VCL task (evident as high performance in the learning blocks, where feedback was available, contrasted with near-chance performance in the following test blocks, where feedback was not available).

### Pre-learning performance

First, we confirmed that the initial performance level, in each of the two feature-saliency conditions, is similar. A two-way analysis of variance (ANOVA) with feature-saliency and feedback information as independent variables, and participants' sensitivity ( $A'$ ) in the first, pre-learning, test block (T-1) as the dependent variable shows no significant interaction  $F(2, 144) = 0.93$ , no main effect of feedback information  $F(2, 144) = 0.91$ , and no main effect of feature-saliency  $F(1, 144) = 2.12, p = 0.15$ .

### Trial-by-trial, feedback-based, learning dynamics

We assessed categorization improvement in the mid- and high-feedback information conditions based on learning trajectories across 72 learning trials (across the three learning blocks, L1 to L3). We calculated performance based on a moving average with a window of six trials. Figure 3 shows rapid learning, with almost identical trajectories, both with mid- and high-information feedback in the high-saliency condition. In contrast, in the low-saliency condition mid-information feedback resulted with significantly lower improvement compared with high-information feedback. Note that high-information feedback

ended with similarly high performance level in both saliency conditions, yet in the low-saliency condition it required more learning trials.

Separate two-way ANOVAs, one for the high-saliency condition and one for the low-saliency condition, with feedback information (mid/high) as the between participants independent variable, learning trial number (1-72) as a within participant independent variable, and participants' percent correct as the dependent variable, show a significant difference in the linear contrast between mid- and high-information feedback learning conditions in the low-saliency condition,  $F(1, 46) = 4.73, p < 0.04$ , but not in the high-saliency condition,  $F(1, 46) = 0.29$ .

Indeed, a three-way ANOVA with feature-saliency, feedback information and learning trial number as independent variables, and participants' percent correct as the dependent variable, confirm that the interaction between feature-saliency and feedback information is significant,  $F(3, 92) = 2.73, p < 0.02$ , partial  $\eta^2 = 0.029$  (Figure 3).

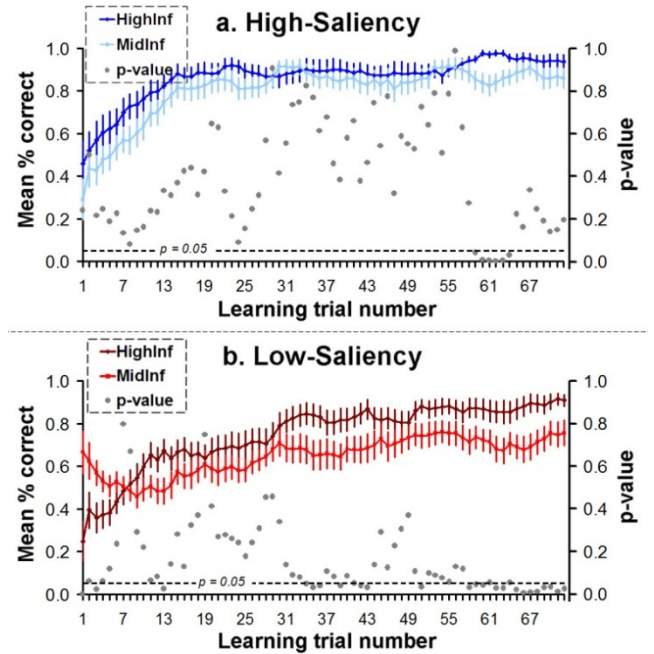


Figure 3: Participants' mean percent correct in the high-saliency (a) and low-saliency (b) conditions in each of the 72 learning trials. The value in each bin is based on moving average with a window of 6 trials (error bars represent standard error of the mean across 24 participants). Gray dots represent the significance level ( $p$ -value) of the difference between mid- and high-information learning in each trial (based on independent sample t-tests; Dashed line marks a significance level of  $p = 0.05$ ). This illustrates a consistent significant difference between the mid- and high-information conditions, particularly in the second half of the learning process, only in the low feature-saliency condition.

### Between test blocks dynamics

To further assess participants learning, we examined their performance in the test blocks where no feedback was given

and all trials had the same composition irrespective of the feedback information condition (paired creatures always differed in 2 features; see Methods). Results are shown in Figure 4. Our analysis shows significant improvement in all learning conditions (all  $p < 0.01$ ). Importantly, we found a significant difference in the learning trajectories between the mid- and high-information feedback conditions in the low-saliency condition but not in the high-saliency condition.

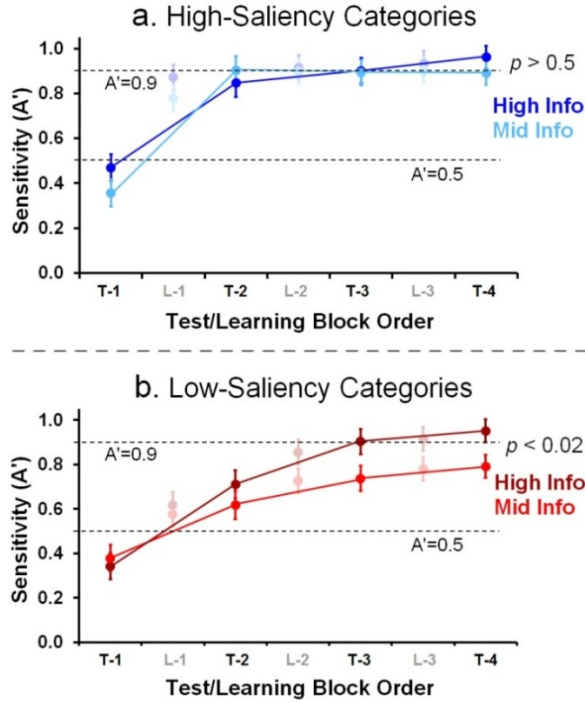


Figure 4: Participants' mean sensitivity (error bars represent standard error of the mean) in the (a) high-saliency and (b) low-saliency conditions. Feedback information levels are marked by different colors. T-1 to T-4: Performance in the test blocks (data points connected with lines); L-1 to L-3: Performance in the learning blocks. When feature-saliency is high, there are no significant differences in VCL efficiency between the mid- and high-information conditions. On the other hand, when the feature-saliency is low and participants were provided with mid-information feedback, performance was lower as compared with the high-information feedback condition.

*High-feature-saliency VCL:* A two-way ANOVA for the high-saliency condition, with feedback information and test block order (T-1 to T-4) as independent variables and participants' sensitivity ( $A'$ ) as a dependent variable shows no significant linear contrast between the mid- and high-information conditions,  $F(1, 46) = 0.04$  (Figure 4a).

*Low-feature-saliency VCL:* In contrast to the above, in the low-saliency condition there was a significant linear contrast between the mid- and high-information feedback conditions,  $F(1, 46) = 6.09$ ,  $p < 0.02$  (Figure 4b).

*Comparing high-feature-saliency and low-feature-saliency VCL:* A three-way ANOVA with feature-saliency

(low/high), feedback information (mid/high) and test block order (T1 to T4) as independent variables, and participants' sensitivity ( $A'$ ) as a dependent variable, confirm that the interaction between feature-saliency and feedback information is close to significant,  $F(3, 92) = 2.84$ ,  $p = 0.06$ , partial  $\eta^2 = 0.030$ .

Next, we examined if differences in sensitivity between the high- and low-saliency conditions are driven mostly by participants' Hit rate or by their False-Alarm rate. We found no significant effect in the False-Alarm rate and a trend in Hit rate: Two-way ANOVAs, one for the high-saliency and one for the low-saliency condition, with feedback information and test block order as independent variables and participants' Hit rate (in the test blocks) as the dependent variable, show a close to significant difference between the mid- and high-information feedback conditions in the low-feature-saliency condition,  $F(1, 46) = 3.65$ ,  $p = 0.06$ , but not in the high-feature-saliency condition,  $F(1, 46) = 0.57$ . A similar analysis with False-Alarm rate as the dependent variable, shows no significant difference between mid- and high-information learning, neither in the low-saliency condition,  $F(1, 46) = 1.49$ ,  $p = 0.23$ , or in the high-saliency condition,  $F(1, 46) = 1.94$ ,  $p = 0.17$ .

These findings are surprising since the main challenge in low-saliency tasks is to learn to identify subtle important differences between similar categories (i.e. avoiding False Alarms) rather than deciding correctly that two apparently similar objects are from the same category (i.e. avoiding Misses). Nevertheless, our findings shows that higher information feedback in low-saliency conditions is mostly helpful in assisting participants performing better by avoiding Misses. We suggest that the lack of significant differences in False-Alarms rate represent, in fact, a response bias exhibited by participants in low saliency conditions – instead of discriminating between categories based on the relevant feature dimension, in the low-saliency tasks participants are more likely to react to any apparent subtle difference among paired creatures as if it is relevant, perhaps due to poor capacity in pinpointing the relevant one.

This interpretation is consistent with the apparent “superior” performance in the first few learning trials in the low-saliency mid-information condition (Figure 3b) where the participants seem to perform better than in the low-saliency high-information condition. In mid-information tasks, a strategy based on deciding “different categories” whenever identifying any difference, is with advantage since the task diagnostic feature is always coupled with an irrelevant one (whereas same-category pairs differ in only one, irrelevant, feature-dimension). That is, in low-saliency mid-information learning conditions people are likely to effectively avoid False-Alarms but for the wrong reason.

#### Performance in the control, no feedback, tasks

Finally, we confirmed that without feedback there is no significant learning in our VCL tasks: A two-way ANOVA conducted for the no feedback VCL tasks with feature-saliency (high/low) and test block order (T1 to T4) as



independent variable, and participants' sensitivity ( $A'$ ) as the dependent variable, shows no significant interaction between feature-saliency and test block  $F(3, 46) = 1.73, p = 0.20$ , and no test block order learning effect  $F(3, 46) = 0.36$ . Note that mean performance in the no-feedback tasks never significantly exceeded values of  $A' = 0.5$ .

## Discussion

We tested the interaction between feature-saliency and feedback information in visual category learning (VCL) tasks as a mean to explore the nature of the interaction between perceptual learning and attentional learning. Simply speaking, perceptual learning is a process that involves improvement in the ability to identify important fine differences between categories, whereas attentional learning improves the ability to filter out irrelevant (even if salient) within category differences (Hammer et al., 2009). Here we show that the interaction between these two processes is more complex than this simplistic view.

We report two important findings: First, perhaps surprisingly, we show that mid-information and high-information feedback are equally effective for learning when stimuli have marked visual differences as in the high-saliency condition. This suggests that when diagnostic visual information is readily accessible, ambiguity in feedback (which is associated with higher attention load) can be resolved with no apparent effort (at least in simple rule learning tasks and when testing typical adults). Second, importantly, there are substantial differences between mid-information and high-information feedback when stimuli are only subtly different. This suggests that low-saliency VCL depends more on informative feedback that serves to orient attention to the relevant feature. Unlike in high-information learning trials, in mid-information learning trials with low-saliency features, participants may not only face difficulties in noticing the relevant feature, but also have difficulties in disassociating it from irrelevant ones. Therefore, participants may have been unaware of the relationship between the feedback and the relevant feature-dimension, which consequently lowered learning effectiveness.

These findings are relevant to the developing debate on the role of attention in perceptual learning: Most findings suggest that perceptual learning requires attention to a target visual feature (Ahissar & Hochstein, 1993; Schoups et al., 2001), or the presence of informative feedback associated with an attended visual feature (Herzog & Fahle, 2002). In contrast, recent findings show that "accidental" perceptual learning can occur (Seitz, Kim & Watanabe, 2009). Nevertheless, this seems to be restricted to learning scenarios with informative feedback where unattended features are strongly correlated with an attended one.

Here we show that when there is only a partial positive correlation between the presentation of a task relevant feature and the presentation of irrelevant features (as it is inherently the case in mid-information feedback conditions), together with lack of explicit information regarding which feature is relevant, there is significant interference with the

learning process. This is evident as significantly less effective learning compared with cases where the relevant feature and irrelevant features are consistently anticorrelated (as in high-feedback-information learning scenarios).

We conclude that the role of attention in visual learning tasks depends on the correlations between relevant and irrelevant features, the nature of information provided by available feedback, and feature-saliency. This suggests that in everyday life scenarios, when making judgments on complex objects in cluttered scenes, the relative contribution of attentional learning and perceptual learning can change quite substantially from one learning scenario to the other. Thus, perceptual learning and attentional learning should not be construed as mutually exclusive processes but rather as complementary processes, and visual learning tasks should be considered as a mixture of these two processes.

## References

- Ahissar, M. & Hochstein, S. (1993) Attentional control of early perceptual learning. *PNAS*, 90, 5718–5722.
- Blair, M. R., Watson, M. R., & Meier, K. M. (2009). Errors, efficiency, and the interplay between attention and category learning. *Cognition*, 112(2), 330-336.
- Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *JEP: LMC*, 30(1), 216-226.
- Goldstone, R. L., & Barsalou, L. (1998). Reuniting perception and conception. *Cognition*, 65, 231-262.
- Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychol Bull*, 75, 424-429.
- Hammer, R., Diesendruck, G., Weinshall, D., Hochstein, S. (2009). The development of category learning strategies: what makes the difference? *Cognition*, 112 (1), 105–119.
- Herzog, M. H., & Fahle, M. (2002). Effect of Grouping in contextual modulation. *Nature*, 415(6870), 433-436.
- Kurtzi, Z. (2010). Visual learning for perceptual and categorical decisions in the human brain. *Vision Res* 50(4), 433-440.
- Nosofsky, R. M., & Palmeri, T. J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bull & Rev*, 3, 222-226.
- Rehder, B., & Hoffman, A. B. (2005). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *JEP: LMC*, 31(5), 811-829.
- Schoups, A., Vogel, R., & Orban, G. (2001). Practicing orientation identification improves orientation coding in V1 neurons. *Nature*, 412(6846), 549-553.
- Seitz, A. R., Kim, D., & Watanabe, T. (2009). Rewards evoke learning of unconsciously processed visual stimuli in adult humans. *Neuron*, 61(5), 700-707.

## Acknowledgments

This research was supported by grants from the NSF (BCS-0720135) and NIH (R01HD056105) to Vladimir Sloutsky, and NIH (R01EY019279-01A1) to Kalanit Grill-Spector.

# Self-Terminated vs. Experimenter-Terminated Memory Search

**J. Isaiah Harbison (isaiah.harbison@gmail.com)**

**Erika K. Hussey (erikahussey@gmail.com)**

**Michael R. Dougherty (mdougher@umd.edu)**

Department of Psychology, University of Maryland at College Park  
College Park, MD 20742 USA

**Eddy J. Davelaar (eddy.davelaar@gmail.com)**

Department of Psychological Sciences, Birkbeck University of London  
Malet Street, WC1E 7HX, London, UK

## Abstract

In most free-recall experiments, participants are given a preset amount of time to search memory. Recently, several studies have examined retrieval in an open-interval design in which the participant, not the experimenter, determines when to terminate memory search. The present study performs the first direct comparison between participant-terminated and experimenter-terminated retrieval. No difference was found in the number of items retrieved from memory; however, inter-retrieval times (IRTs) did differ, such that the participant-terminated paradigm did not show the hyperbolic function typically found when using the experimenter-determined, closed-interval design. We were able to account for this result by equipping a simple relative sampling model with a memory search stopping rule that assumes that giving participants a pre-set retrieval interval causes them to search longer (and tolerate more search failures) than they would in the open-interval design.

**Keywords:** memory; free-recall; stopping rules.

## Terminating Memory Search

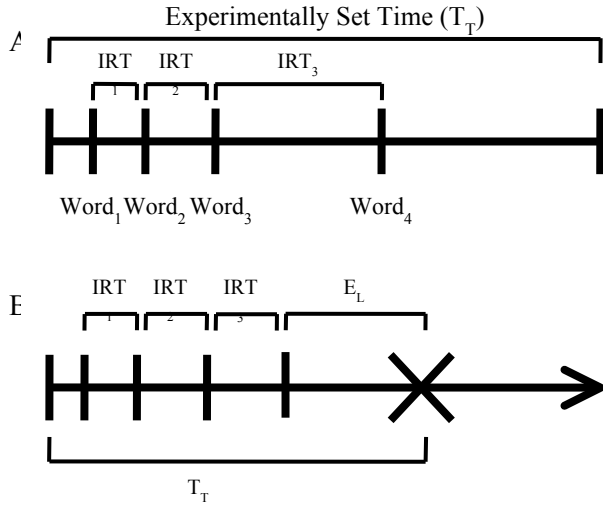
At some point, any search of memory must end. Several recent studies have begun to examine how this decision is made in free recall tasks, where search is often terminated while there remain potentially retrievable items unreported (Dougherty & Harbison, 2007; Harbison, Dougherty, Davelaar, & Fayyad, 2009; Unsworth, Brewer, & Spillers, 2011). These studies used a slightly modified version of the standard list recall paradigm; the only difference is that participants terminate their own memory search. This open-retrieval interval design (henceforth, open-interval design) is in contrast to the standard, closed-retrieval-interval design (closed-interval design) that gives participants a pre-determined retrieval interval. Both the open-interval and closed-interval designs have strengths and weaknesses, but to date, they have not been directly compared. The goal of the present study is to perform this comparison and evaluate how allowing or not allowing participants to terminate their own search influences the variables used to describe memory retrieval.

## Open- and Closed-Interval Designs

The difference between the open-interval and closed-interval recall design is illustrated in Figure 1. The closed-interval design, shown in Panel A, is restricted, in that the retrieval interval participants are given is pre-determined by the experimenter. After the interval has expired, search is terminated for the participant by the experimenter or by the software program used for the experiment. One reason this design has been used is that it allows greater focus on basic processes of memory retrieval, attempting to eliminate individual differences in how long participants spend searching memory. However, it is not necessarily the case that participants search during the entire pre-determined interval. In fact, many process models that have been proposed to account for the retrieval results from the closed-interval design have assumed a stopping decision to be part of the memory search process (e.g., Raaijmakers & Shiffrin, 1981). Moreover, the closed-interval design might induce participants to search memory longer than they normally would, potentially leading to results that do not replicate when participants are free to retrieve and self-terminate memory search.

The open-interval design (panel B of Figure 1) gives participants an unlimited amount of time to retrieve. The principle strength of this design is that it allows for the measurement of memory search termination decisions, including the total time spent in search—determined by the participant—and the exit latency, or the time between the final retrieval and the decision to terminate search. Both measures have proven diagnostic for evaluating memory search stopping rules (Harbison et al., 2009). The design also, arguably, has greater ecological validity: Individuals are unlikely to have a fixed external time limit when searching memory during most everyday tasks outside the lab (for an examination of termination decision in response to external demands, see Davelaar, Yu, Harbison, Hussey, & Dougherty, 2012). However, the open-interval design too has potential weaknesses: Self-termination might prime participants to put less effort into retrieval and therefore provide inadequate data for the purposes of theory testing. If

participants lacked sufficient motivation to adequately search memory, they might recall fewer items in the open-interval paradigm. However, to the best of our knowledge, this account has not been evaluated. Moreover, as far as we are aware, there has been no comparison between open- and closed-interval designs more generally. In what ways do retrieval data obtained from an open-interval design differ from those obtained in the closed-interval design? And, what can the open-interval design tell us about memory retrieval that cannot be discerned from the close-interval design?



**Figure 1.** (A) Closed-interval and (B) open-interval retrieval designs, adapted from Harbison & Dougherty (2007). X indicates the time when a participant decides to terminate memory search; hash marks indicate the time associated with the latency onset of words recalled.  $T_T$  = Total Time Searching;  $E_L$  = Exit Latency; IRT = Inter-Retrieval Time.

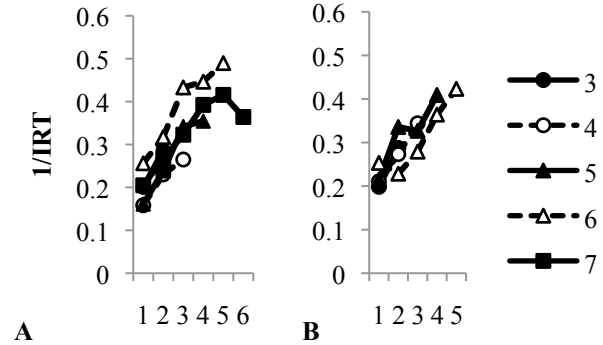
## Previous Results

As no experiment has yet directly compared the open- and closed-interval designs, it is not possible to draw firm conclusions from existing data. However, there is a wealth of data from closed-interval experiments suggesting specific patterns in the temporal dynamics of recall. For our purposes, we focus on the inter-retrieval times (IRTs), or the time between successive retrievals. IRTs have played an important role in constraining theories of memory retrieval (Rohrer, 1996; Wixted & Rohrer, 1994), and they are generally well described by the equation:

$$IRT_i = \frac{\tau}{N - i} \quad (1)$$

for  $i=1,2,\dots,N-1$ , where  $i$  is the inter-response interval starting with the interval between the first and second retrieval,  $\tau$  is the estimated mean retrieval latency, and  $N$  is the total number of items retrieved. Equation 1 captures the key empirical result that IRTs typically follow a hyperbolic function, such that the time between successive retrievals

increases as the number of items retrieved increases (Murdock & Okada, 1970; Polyn, Norman, & Kahana, 2009; Raaijmakers & Shiffrin, 1980; Wixted & Rohrer, 1994).



**Figure 2.** 1/IRT data from (A) Dougherty and Harbison, 2007 and (B) Harbison et al., 2009. The x-axis is the retrieval interval in reverse order, with 1 representing the final interval. The legend indicates the number of items recalled.

A particularly informative way of looking at IRT data is by inverting the IRT (1/IRT) and plotting the results in reverse order along the x-axis, such that the final IRT is in the first position. When plotted this way, Equation 1 predicts that the intercept should be zero. Rohrer (1996) found support for this prediction using the closed-interval design. However, a reanalysis of data from two open-interval experiments (Dougherty & Harbison, 2007; Exp. 1, Harbison et al., 2009) revealed a different pattern: Instead of an intercept of zero, the data from these experiments were best fit by lines with intercepts greater than zero (ranging from .103 to .210), as shown in Figure 2.

What does this result mean for the comparison of open- and closed-interval designs? In particular, could this indicate that something differs in the search process when participants make their own stopping decisions? To answer these questions, we conducted a simulation evaluating the predictions of the relative-strength model of retrieval when stopping decisions were included.

## Simulation

For simplicity and ease of comparison with previous research examining IRT results from the closed-interval design, we followed the same simulation procedure as Rohrer (1996). We used the same relative-strength model, which is nearly identical to the sample and recovery processes of the search of associative memory (SAM) model (Raaijmakers & Shiffrin, 1981). We also used the same activation patterns tested by Rohrer (1996). The model randomly sampled items based on their relative activation and attempted to recover the sampled item based on its absolute activation. An iteration of random sampling and attempted recovery is referred to as a retrieval attempt and each retrieval attempt could either succeed or fail. For the

current simulations, potential retrieval failures are (1) re-sampling an already-outputted item or (2) failing to recover a sampled item due to its absolute activation not meeting the recovery threshold. As we used the same activation patterns as Rohrer (1996), we also used the same recovery threshold, .5.

The one difference between the model tested here and the model used by Rohrer was our use of a stopping rule that terminated the search process. Our model included a stopping rule based on the number of retrieval failures. This stopping rule, native to the SAM model, is the only rule tested so far that has been able to account for both the total time and exit latency data from open-interval experiments (Harbison et al., 2009).

Other than the recovery threshold, the only parameter in the model was the stopping threshold, the number of retrieval failures the model allowed before memory search was terminated. The stopping threshold was varied between 10 and 40 in steps of 10. Each activation pattern was run with each stopping threshold 10,000 times. The dependent variables of interest included the number of items retrieved, the IRTs, and the intercept of the best fitting line for the 1/IRT data.

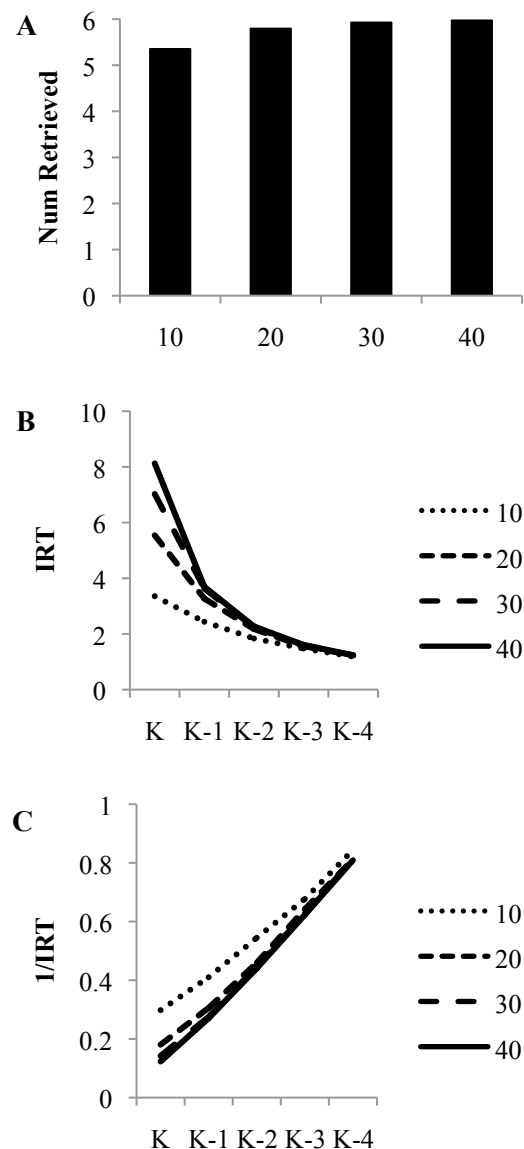
**Table 1.** Mean Simulation Results by Activation Pattern.

Act. Pattern	Variable	Stopping Threshold			
		10	20	30	40
1,1,1,1,1,1	Num Ret	5.60	5.94	5.99	6.00
	Last IRT	3.30	5.17	5.80	6.00
	Intercept	.16	.03	0	0
.5,.6,1,1.2,1.2,1.5	Num Ret	5.35	5.79	5.92	5.97
	Last IRT	3.35	5.54	7.03	8.12
	Intercept	.14	0	-.05	-.06
.4,.5,.6,1,1.5,2	Num Ret	4.35	4.78	4.92	4.97
	Last IRT	3.37	5.73	7.43	8.42
	Intercept	.13	-.02	-.07	-.09

## Results and Discussion

The results are reported in Table 1 and a representative sample of the model's behavior over various activation patterns. The results of the second activation pattern in Table 1 are shown in Figure 3. First, it should be noted that the variation in stopping threshold did not have a large impact on the mean number of items retrieved (see Figure 3a). However, Figure 3c illustrates that the intercept of the inter-retrieval rates did show substantial variation. Consistent with Rohrer (1996), when the stopping threshold was larger (e.g., 40 failures), the intercept was indeed near zero. However, there was a negative correlation between stopping threshold and the intercept, such that at smaller stopping thresholds the intercept was greater than zero. Therefore, the 1/IRT predictions of the relative-strength model appear to be consistent with closed-interval experiments at greater stopping thresholds and open-interval experiments at smaller stopping thresholds.

Investigating the IRT data more closely, Figure 3b shows the last (or K) IRT, the second to last (K-1) IRT, the third to last (K-2) IRT, and so forth for each stopping threshold value tested. The last IRT showed the greatest variation due to changes in the stopping threshold. Weaker relationships between IRT and stopping threshold were observed the further the IRT was from the final IRT. Importantly, there were large variations in the last IRT even when there were only minute changes in the number of items retrieved. For example, while the mean final IRT more than doubled in size when going from a stopping threshold of 10 to 40, the mean number retrieved varied by only ten percent.



**Figure 3.** Simulation results for (A) number of items retrieved, (B) IRT, and (C) 1/IRT functions for 4 stopping thresholds (10, 20, 30, or 40 failures).

What is the source of the relationship between IRTs and the stopping threshold? When using the total failures

stopping rule, the lower the stopping threshold, the fewer the number of possible attempts for retrieving each item, on average. If an item is not retrieved with a minimal number of failures, it will not be retrieved. For example, if the fifth item is not retrieved before 10 retrieval failures have occurred, then search will terminate with only four items retrieved. Since the probability of retrieval failure increases with each item retrieved, the limit on the number of allowable retrieval failures plays a larger role towards the end of retrieval and particularly for the final item retrieved.

Note that these predictions are particular to the total failure stopping rule. Although not presented here, of the four stopping rules tested in Harbison et al. (2009) the only other stopping rule that correctly predicts the general form of the IRTs is the total time rule, though this rule *cannot* account for the systematic variation in total time spent in search. When equipped with the time-since-last-success or the last-IRT stopping rules, the relative-strength model produces IRT predictions that vary substantially from the results of both open- and closed-interval experiments.

The apparent difference in IRT data between the open- and closed-interval designs are accounted for by the relative-strength model as long as the model includes a stopping rule based on total retrieval failures. According to the model, the difference between the open- and closed-interval designs is expected if the designs induce subjects to use different stopping thresholds. What is left to determine is if the pattern is indeed real. Testing this requires a direct comparison between the designs.

## Experiment

Forty-nine participants were randomly assigned into one of two counterbalancing conditions: open-then-closed or closed-then-open. List length was also varied within participant, resulting in a 2 (retrieval block: open vs. closed) x 4 (list length: 5, 7, 9, vs. 11 words) within-subjects design. List length was varied randomly such that all participants were given four study lists of each of four lengths evenly and randomly within each block. List length was systematically varied primarily to prevent participants from learning exactly how many items were on each list and using that information to determine stopping decisions.

**Stimuli** Thirty-six word lists were randomly generated for each participant from a list of 280 high-imagability ( $M = 577/700$ ), high-concreteness ( $578/700$ ), moderate-to-high-frequency (Kucera-Francis frequency = 54), single-syllable nouns drawn from the MRC psycholinguistic database (Wilson, 1988).

**Procedure** For both the open- and closed-interval blocks, participants were given a total of 18 lists consisting of 2 practice trials followed by 16 test trials. During the list presentation of each trial, words were presented sequentially at a rate of 3 seconds per word. Following the list, participants were given a distracter task that consisted of two simple, timed math problems. Each problem contained

three digits and two operands (e.g.,  $3 * 2 + 1$ ). Each component of the problem was presented sequentially at a rate of one second per item. After viewing the final digit of the problem, participants saw an equal sign with a question mark, prompting them to respond with the correct answer.

Participants were then given the opportunity to verbally recall items from the most recent word list. During open-interval trials, participants were told to press the spacebar when they could no longer retrieve additional items from the current memory list; hence, they were given control over when to end the retrieval interval. During closed-interval blocks, participants were given 45 seconds to retrieve the study list items. Based on prior research, we anticipated that a 45-second retrieval interval would provide ample time for most participants to complete the recall task.

All participants were presented with both block types to ensure a proper comparison of IRTs between the open and closed intervals, and the order of block presentation was counterbalanced across participants. All retrievals were made verbally by speaking into a microphone and were digitally recorded for later scoring. The responses for each list were recorded in an audio file and hand coded to extract the time-to-word for each item recalled.

## Results and Discussion

We conducted Jeffreys-Zellner-Siow (JZS) Bayes-factor (BF) tests to verify the results of each significance test. Moreover, some comparisons reported below are expected to support the null hypothesis, and JZS BFs provide a means to assess the degree to which this is indeed the case. Bayes-factor tests reflect the likelihood of support for the null hypothesis over support for the alternative hypothesis, such that coefficients less than 0.3 index strong support for the alternative hypothesis and those greater than 3 index strong support for the null hypothesis (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

**Block Order** We first conducted a manipulation check to determine whether there was an effect associated with block order (open-interval first vs. closed-interval first). A repeated-measures analysis of variance (ANOVA) revealed no effect of Block Order x Block Type for average number of items recalled ( $F(1,47)=0.002$ ,  $p>0.96$ ,  $BF=4.68$ ) or total number of intrusions ( $F(1,47)=0.442$ ,  $p>0.50$ ,  $BF=3.84$ ). Because exit latency cannot be computed for trials in the closed-interval block, we examined the effect of Block Order on exit latency only on open-interval trials, and found no main effect ( $F(1,47)=2.30$ ,  $p>0.13$ ,  $BF=1.70$ ). Finally, there were no reliable effects of Block Order for IRTs at any level of number recalled ( $p's>0.56$ ). Because these early analyses suggest that there are no effects of Block Order, all subsequent analyses will be collapsed across this factor to increase statistical power.

**List Length** We next examined the effect of list length on number recalled. Replicating earlier work, we showed a main effect of number recalled ( $F(3,291)=60.39$ ,  $p<0.0001$ ).



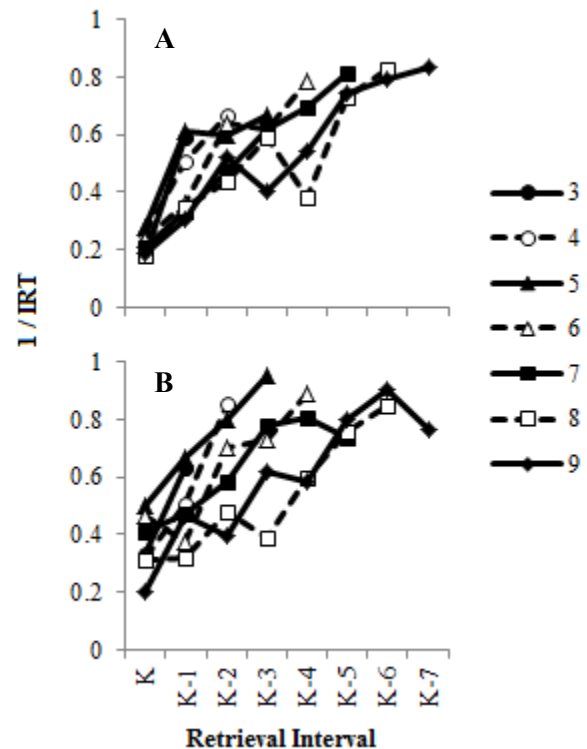
Scheffé post hoc analyses revealed that longer study lists resulted in the retrieval of additional items ( $M_5=4.14$ ,  $M_7=4.75$ ,  $M_9=5.16$ ,  $M_{11}=5.92$ ;  $p's < 0.05$ ). Thus, all subsequent analyses will be collapsed across list length.

**Number Recalled** There was not a significant difference in the number of items retrieved between the closed- ( $M=5.11$ ) and open-interval trials ( $M=4.89$ ;  $t(48)=0.947$ ,  $p>0.33$ ,  $BF=5.78$ ), indicating that people's decisions to terminate retrieval did not impact recall rates. Analyses of intrusion and repetition errors are also consistent with this conclusion: The average number of intrusions did not differ as a function of Block Type ( $M_{closed}=0.351$ ;  $M_{open}=0.441$ ;  $t(48)=-0.972$ ,  $p>0.33$ ,  $BF=5.64$ ). An effect of Block Type did not emerge when intrusions were split into 3 types: a) repetitions ( $t(48)=0.467$ ,  $p>0.64$ ,  $BF=8.04$ ); b) extra-list false alarms, or items recalled that were not presented in any prior study lists ( $t(48)=0.502$ ,  $p>0.61$ ,  $BF=7.91$ ); and, c) intra-list false alarms, or items that were incorrectly output that occurred on previous lists ( $t(48)=1.642$ ,  $p>0.10$ ,  $BF=2.47$ ). Also, intrusion rates did not change as a function of time spent in the experiment ( $p>0.47$ , but see Unsworth et al., 2011). Given these results, we are comfortable concluding that the open-interval design does not differ from the closed-interval design in terms of number and type recalled.

**Exit Latency and Total Time** We found that the current exit latency and total time data were consistent with previous results using an open-interval paradigm. Exit latency decreased as a function of number recalled (Dougherty & Harbison, 2007): Mean within-subject gamma ( $\gamma$ ) correlation coefficients for exit latency and number recalled (mean  $\gamma = -0.139$ ) indicate that participants spend more time deciding to terminate search after the final item is recalled when fewer words are output in a trial (one-sample t-test of  $\gamma$ :  $t(48)=-3.719$ ,  $p<0.001$ ). Also consistent with previous data, the total time spent in search was positively correlated with the number recalled (mean  $\gamma = 0.268$ ;  $t(48)=5.775$ ,  $p<0.0001$ ).

**Inter-Retrieval Times** IRTs were computed by taking the difference between the verbal onset times for each subsequent item recalled in a trial. We conducted these irrespective of the identity of the item recalled (i.e., IRTs were computed to incorporate trials containing intrusions). We first examined IRTs as a function of Block Type (open- vs. closed-interval) for each level of Number Recalled for each participant. Since many participants did not output a full range of Number Recalled levels across both open- and closed-interval blocks, pairwise comparisons were only conducted for subjects that could contribute data to both levels of Block Type for a given Number Recalled level. For example, it was possible that a participant recalled three items in two separate trials of the open-interval block and never recalled three items in any trials of the closed-interval block. Because of this variation in observations, we only

report IRT averages when at least 15 subjects contributed data to both open- and closed-intervals.



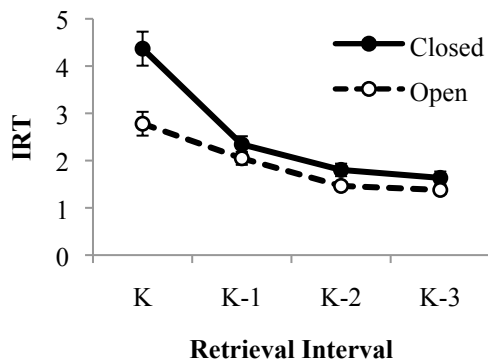
**Figure 4.** Mean IRTs as a function of Number Recalled for (A) Closed- and (B) Open-interval retrieval trials.

Figure 4 illustrates the  $1/IRT$ s for open- and closed blocks as a function of Number Recalled. Closed-interval intercepts were less than the open-interval intercepts for 6 of the 7 different total number of items retrieved. Especially important to this functional relationship is the final, or K, IRT (see Figure 5); there was a main effect of Block Type on the K IRT ( $F(1,47)=28.48$ ,  $p<0.0001$ ,  $BF=1.53 \times 10^{-4}$ ), such that closed-interval trials ( $M=4.462s$ ) led to longer final IRTs than open-interval trials ( $M=2.792s$ ). In fact, the mean K IRT on closed-interval trials was significantly larger than that on open trials for all but two levels of Number Recalled (i.e., 3 and 8,  $p's > 0.38$ ). A sign test of the binomial relationships for the number recalled of all final IRTs reveals that six of the six comparisons favor closed-retrieval intervals to have longer IRTs than open-interval intervals; a one-tailed test assessing the probability of this pattern occurring yielded a p-value of 0.016. Thus, despite the fact that there were no significant differences in overall number recalled, there do appear to be differences in the temporal dynamics between open- and closed-interval results.

## General Discussion

The present experiment directly compared the closed-interval design, in which the experimenter determines the length of the retrieval interval, to the open-interval design, in which participants are allowed to terminate their own

memory search. The IRT functions differed between these designs: Compared to their closed-interval counterparts, open-interval trials resulted in overall shorter final average IRTs.



**Figure 5.** Mean retrieval times for the final (K) IRT, second-to-last (K-1) IRT, third-to-last (K-2) IRT, and fourth-to-last (K-3) IRT.

What do these differences in the IRTs mean? According to the present simulations, these results suggest a difference in the stopping threshold between open- and closed-interval retrieval. As shown in Figure 3B, the relative-strength model, when equipped with the stopping rule supported by the existing data (Harbison et al, 2009), predicts a positive correlation between the final IRT and the stopping threshold. When the stopping threshold is sufficiently large, the final IRT is also large and the IRT predictions are consistent with Equation 1; specifically, the intercept of the inverse of the IRTs is 0 when plotted in reverse order (Rohrer, 1996). However, when the stopping threshold is set to smaller values, the final IRT is also smaller. This decreases the predicted slope (or mean retrieval latency,  $\tau$ ) and increases the intercept, producing results similar to those observed under the open-interval design and inconsistent with Equation 1.

The present research finds systematic differences in the temporal characteristics of memory retrieval between open- and closed-interval designs. These differences are predicted by the relative strength sampling model when equipped with a memory search stopping rule if it can be assumed that the type of retrieval interval influences memory search stopping decisions. In the terms of the model, participants appear to use the same stopping rule but set a higher stopping threshold for closed-interval retrieval than for open-interval retrieval. Participants persist in search longer; however, they do not retrieve more items in the closed-interval design than in the open-interval design. The temporal differences were predicted by the model and observed in the data despite no appreciable differences in the number of items retrieved. These results indicate that participants do not in fact terminate search over-quickly in open- relative to closed-interval designs. Furthermore, as participants were able to retrieve the same amount of items in less time, the results suggest that the open-interval design might provide a

method to measure not only how memory is searched, but also how efficiently memory can be searched.

## Acknowledgments

We gratefully acknowledge NSF (grant BCS-1030831) for support of this research.

## References

- Davelaar, E. J., Yu, E., Harbison, J. I., Hussey, E., & Dougherty, M. R. (2012). A rational analysis of memory search termination. *Proceedings of the 11<sup>th</sup> International Conference on Cognitive Modeling*.
- Dougherty, M. R., & Harbison, J. I. (2007). Motivated to retrieve: how often are you willing to go back to the well when the well is dry? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 1108-1117.
- Harbison, J. I., Dougherty, M. R., Davelaar, E. J., & Fayyad, B. (2009). On the lawfulness of the decision to terminate memory search. *Cognition*, 111, 146-421.
- Murdock, B. B. & Okada, R. (1970). Inter-response times in single-trial free recall. *Journal of Verbal Learning and Verbal Behavior*, 86, 263-267.
- Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, 116, 129-156.
- Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, 88, 93-134.
- Rohrer, D. (1996). On the relative and absolute strength of a memory trace. *Memory & Cognition*, 24, 188-201.
- Rohrer, D. & Wixted, J. T. (1994). An analysis of latency and inter-response time in free recall. *Memory & Cognition*, 22, 511-524.
- Rouder, J. N., Speckman, P., Sun, D., Morey, R., & Iverson, G. J. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16, 225-237.
- Unsworth, N., Brewer, G. A., & Spillers, G. J. (2011). Factors that influence search termination decision in free recall: an examination of response type and confidence. *Acta Psychologica*, 138, 19-29.
- Wilson, M. D. (1988). The MRC Psycholinguistic Database: Machine readable dictionary, Version 2. *Behavioral Research Methods, Instruments and Computers*, 20, 6-11.
- Wixted, J. T. & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review*, 1, 89-106.

# The use of ACT-R to develop an attention model for simple driving tasks

**Kerstin Sophie Haring (ksharing@fennel.rcast.u-tokyo.ac.jp)**

**Katsumi Watanabe (kw@fennel.rcast.u-tokyo.ac.jp)**

Research Center for Advanced Science and Technology, The University of Tokyo  
4-6-1, Komaba, Meguro-ku, Tokyo, 153-8904, Japan

**Marco Ragni (ragni@cognition.uni-freiburg.de)**

**Lars Konieczny (lars@cognition.uni-freiburg.de)**

Center for Cognitive Science, University of Freiburg  
Friedrichstr. 50, 79098 Freiburg, Germany

## Abstract

Driving a car is obviously a complex task and the construction of an ACT-R model of human attention while performing this task is similarly complex along multiple dimensions and presents a challenge to architecture and modeler. This work is a first attempt to develop an integrated driver model of attention in the cognitive architecture ACT-R. The model is able to keep a traffic lane, identifies traffic signs and crossroads in a sparse, simulated environment.

**Keywords:** Driver behavior model; cognitive architecture; ACT-R; Attention

## Introduction

For most of us, driving a car is one of our everyday tasks. But even for experienced drivers, just the task itself it is a cognitive challenging task involving a big range of human senses like sight, hearing, touch and acceleration. And this is not yet considering secondary tasks like talking on the phone or visual distractions like city illuminations. Luckily, most driving task are not as challenging as the Traffic Light Tree in Fig. 1, an artificial scenario by the French sculptor Pierre Vivant.



Fig. 1: The (thankfully not on a crossroad) installed traffic light sculpture by Pierre Vivant.

Current attempts to model human attention while driving a car are realized in a quite more simple environment, yet they are quite an important first step towards the modeling

of these highly complex tasks. Vice versa, it also can provide an indication for the future development of a cognitive architecture by showing what cannot be realized yet.



Fig. 2: Screenshot of the environment interaction with ACT-R. The red circle indicates the current visual focus of attention of the model.

The simulation environment for this model was restricted to the components the cognitive architecture can recognize. Nevertheless, basic driving scenarios simulating human visual attention and driver behavior could be implemented.

The screenshot from the driving environment, which was separately implemented in Lisp for this work, shows from top-down another (blue) vehicle, the focus of attention (red circle) and the navigation point (N) to keep the vehicle in the center of the road. This model focuses on basic reference points like the horizon, a leading car, the border and the center line of the road, crossroads and traffic signs. For example, the model of a driver in the screenshot in Fig. 2 sets the focus of visual attention on the outer border of the road which enables it to reevaluate the center for the N point. In the next step, it will shift the focus of attention to the front and (hopefully) detect the car in front. If so, possible next steps could be the comparison of the distance to a (here fixed) safety distance or an overtaking procedure.

The here presented cognitive model should simulate through ACT-R human attention while driving in a simplified environment and produces the behavior for scenarios with other cars, crossroads and traffic signs.

## The cognitive architecture

The ACT-R (Anderson, 1993; Anderson 2007) cognitive architecture proposes artificial, computational processes that aim to act like a human cognitive system. Most of its basic

assumptions are inspired by the progress of cognitive neuroscience. The tasks that humans can perform should, in theory, consist of a series of discrete operations. ACT-R is primarily used to model experimental psychological data. ACT-R comprises theories about the operation mode of human information processing and describes a comprehensive computer model of human cognition. The architecture is not only a proposed unified theory of cognition, it is also a programming environment, a production system with a development environment where it is for example possible to set parameters or run simulations. ACT-R is a framework in which the researcher can create models (programs) for different tasks. Running this model produces a simulation of human behavior.

As many cognitive architectures, ACT-R contains a number of modules which can be accessed through their limited-size buffers. For each module, a dedicated buffer serves as an interface with that module. The state of ACT-R at a given time is the content of the buffers at that time. Buffers are connected to the modules and are changed by production rules. Every buffer and (nearly) every module can be allocated to a cortex region. This enables an interesting mapping between buffers and neural processes (Anderson 2007).

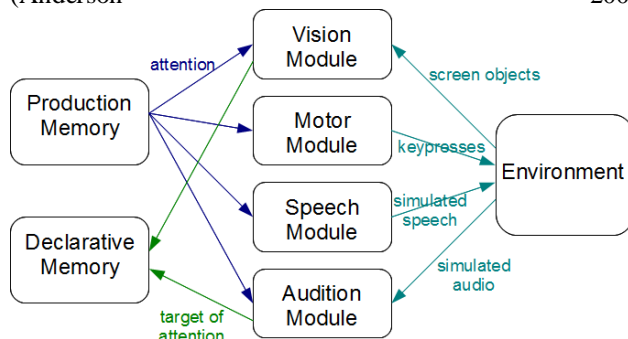


Fig. 3: ACT-R system diagram (Byrne, 2001). The Cognitive Layer and each of the Perceptual-Motor modules run in parallel, but each component is itself serial.

The main assumption of ACT-R is the representation of knowledge as either declarative or procedural knowledge. Declarative knowledge, consisting of facts, is represented in form of *chunks*, or small logical units which encode simple facts (e.g. the fact: “Sapporo is in Japan”). Procedural knowledge, representing knowledge about how we do things, is represented in form of *production rules*, condition-action rules that generate a specific action (e.g. manipulate declarative knowledge) if the conditions of this rule are fulfilled. In other words, ACT-R’s knowledge representation is split in two kind of memory modules, perceptual-motor modules and memory modules.

The diagram in Fig. 3 shows the ACT-R in action. For the visual attention, the environment provides screen objects to the vision module. The target of attention is put into the declarative memory in form of a chunk.

We decided to choose ACT-R for this task because it has a visual search, is a well-accepted cognitive architecture, and was already used in the past to evaluate the attention during a driving task. A crucial advantage of the ACT-R architecture is that the three main components used in this model (control, monitoring and decision-making) can be implemented directly. This takes into account human constraints and results in a cognitive adequate model of human attention.

## Previous work

Most developed approaches can be distinguished into two categories: task specific and generic approaches. *Task specific approaches* such as Cosmodrive (Bellet et al., 2007) and Pelops (Benmimoun, 2004) reproduce the cognitive functions of a car driver. In contrast to task specific approaches, *generic approaches* can model various aspects of human behavior. Therefore, it is necessary for these architectures to include a theory of human information processing. Examples for such architectures in which driver models have been implemented are ACT-R (Anderson, 1993; Salvucci, 2006), SOAR (Aasman, 1995) and QN-MHP (Liu et al., 2006).

Driver models were described by Aasman (1995) in the cognitive architecture SOAR and by Liu (1996) in Queuing Network-Model Human Processor (QN-MHP). Although these models already exist in other cognitive architectures and the central ideas remain the same in any architecture, the ACT-R model of a driver shows a broader spectrum of application (Salvucci 2001; 2006).

Salvucci (2006) developed a first integrated cognitive model of human driving behavior in ACT-R. He showed in his work the generality and the applicability using the cognitive architecture ACT-R for the specific task of driving. His model is designed to keep a standard vehicle on a multi-lane highway with moderate traffic. The model is also able to recognize the distance to a vehicle ahead and to make the decision for overtaking. As driving is a highly complex task and not readily implementable, this model has some limitations. The model solely was meant to interact with a highway environment without recognition of traffic signs, crossings or slip roads. An implementation limitation was the use of the previous version ACT-R 5.0 and its incompatibility to newer versions. It was also not possible to make the ACT-R model interact directly with a driving simulator.

Regardless the challenges, Salvucci proposed to develop a driver model in the context of embodied cognition, task and artifact (ETA) framework (Byrne, 2001), an idea which was adopted in this work.

## Cognitive model

A driver model can be a powerful instrument with several possible fields of application, such as the development of intelligent driver assistant systems. Our model is implemented the newest version ACT-R 6 (Anderson, 2007) and using the standard ACT-R development environment



running on an open source LISP, which not only guarantees support and accountability, but also enables the research community to use the developed model for further research.

## Driver Modeling

We introduce a computational model of human attention in a car driving task implemented in the ACT-R architecture. As described previously, this model aims to account for the embodied cognition, task and artifact (ETA) framework.

The complex task of driving a car can be divided into basic subtasks. These must be integrated and interleaved to achieve the continuously changing parent task. Michon (1985) identified three levels of skills and control for the driving task: operational (control), tactical (maneuvering), and strategical (planning). He claims that a comprehensive model should take into account the various levels.

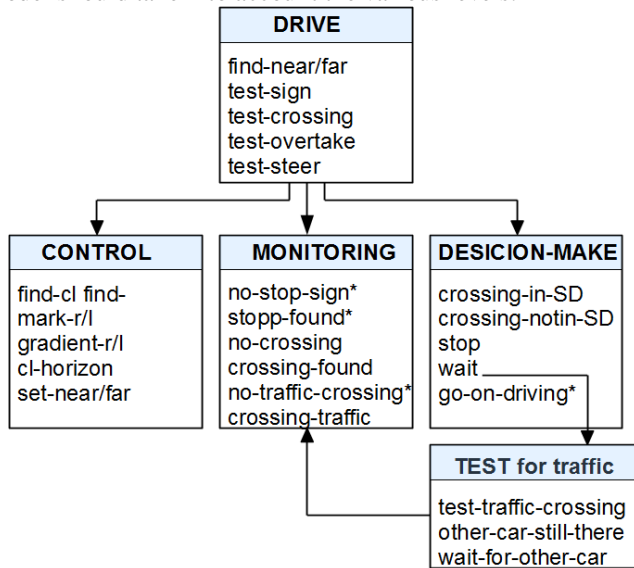


Fig. 4: Schematic representation of the production rules of the driver model in a simple crossroad scenario. The title of a box indicates the current goal and the corresponding production rules. The arrows show the flow of control and the asteriks the return to the parent-goal.

The independent subtasks of the (simplified) parent task *drive* (see Fig. 4) were implemented as *control*, the operational process controlling the input, *monitoring*, the tactical process interacting with the environment, and *decision making*, also analogous to the tactical level of Michon (1985), managing maneuvers like overtaking. These subtasks are processed serially. Every production of the top level goal *drive* has sub-goals, which incorporate the three components.

**Development Environment** The theory of ACT-R is embedded in the ACT-R software in form of Common Lisp functions. This model is implemented in the open source Clozure CommonLisp 1.3 and the current version of ACT-R 6.0 under the operating system Ubuntu 9.04. In order to

make the simulation environment interact with the ACT-R system, it was directly implemented in LISP with simple graphics and the extension with the LTK Lisp Toolkit. As it was not possible to make ACT-R directly interact with a driving simulator, we decided to use a Lisp-implementation of a driving environment.

## Model Specification

As mentioned earlier, human attention during a driving task is quite complex along multiple dimensions. It is not yet possible to model every aspect of human attention within a cognitive architecture for such a complex task. To limit the scope of the project, model is hold quite simple. The model focusses on simple visual perception and attention shifts how they might occur in a sparse, artificial environment.

The first issue to address was to implement the three components control, monitoring and decision making as a loop of cognitive operations in the serial processor of the ACT-R architecture. The UML diagram in Fig. 5 shows the behavior of the cognitive model. This diagram identifies one primary loop, which corresponds largely to the control component in Fig. 4. The primary loop implements the identification of the near and the far point, in other words, the points responsible for the stable navigation in the middle of the road. From the initial state, the model finds the road marks and sets the near point for stable navigation on the road. The model then fires a production rule screening for a traffic sign, changes the state according to the result and sets the far point. In our model, the near and far point are used as control components and explained in detail in the next paragraph. If the model reaches the state *find far* it will continuously repeat the primary control loop.

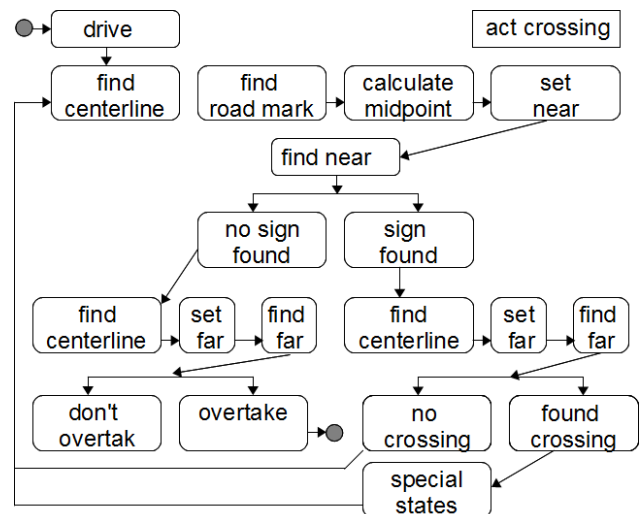


Fig. 5: UML-Diagram of the driver model. To execute the task drive, the model runs through several states.

This primary loop can be extended in case the monitoring component finds a special state like an intersection, that is, the condition part of a production rule investigating the right road marks on the right detects a crossing and the action part of this rule changes the state of the model, testing for other

given constraints. According to the result, the model might change the state or repeat the primary control loop updating the near and the far point for stable navigation.

## Control

The control component of attention while performing a driving task manages the perception of lower level visual cues and the control over the vehicle (e.g., stopping). The model uses the simple concept of two salient visual attributes, based on earlier findings on locomotion (Llewellyn, 1971). Steering is described (Land & Horwood, 1995) as divided in two levels, guidance and stabilization, by using a „far“ and a „near“ region. Models of steering developed under this assumption have been proven to be consistent with empirical evidence.

This task specific information was required to construct the model. An issue to be addressed was what kind of strategies might be used by a human in a driving environment. Salvucci & Gray (2004) base the perception of a cognitive model on a near and a far point for guidance and stabilization. This model extends the idea of two levels to the extend, that with the far point, also other salient attributes are encoded. The visual attention of the model does not only switch in between the near point in the middle of the road and the horizon (or any other straight point ahead), but also encodes crossings, traffic signs or other cars coming from the right hand side on a crossing. For the here created artificial road environment, these two points account to capture the relevant aspects of the environment. The idea behind this wider use of the far point is the possible extension in further implementations. The far point could be used to determine other attributes relevant or irrelevant for the driving task and give an account, how errors while driving (e.g. overlooking of a traffic light before a noisy background during monitoring).

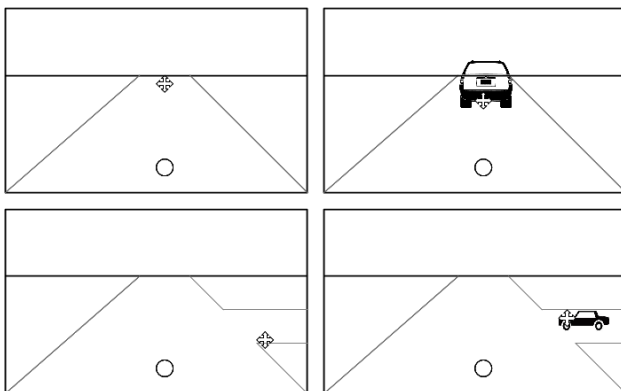


Fig. 6: Near and far points for a straight road with a vanishing point and a road segment with a lead car.

Fig. 6 illustrates the near and far points and how they are set in some possible situations during the driving task. The near point determines the position in the middle of the road. The far point is used to identify the direction of driving and other non-control points like vehicles, crossroads and traffic signs. If there is a lead vehicle, the distance between the two

points is determined, and in case it falls below a certain safety distance, the model can react according to that. At a crossroad the model will set the far point on the border of the crossing or on the vehicle approaching from the right. If the model decides to continue driving, it will not look again for another car at the crossroad, which is surely an issue for future implementations. Stopping is implemented by setting the far point onto the near point. The model will continue a loop until the other vehicle is not on the crossroad anymore and out of the safety distance.

## Monitoring

The monitoring component captures the environment continuously and updates the declarative memory. In the here implemented driving environment, the situation awareness mainly focuses on other vehicles around or traffic signs. The model shifts the focus of visual attention towards a certain object which is then encoded as visual attribute in the declarative memory. This shift is achieved through three different methods of shifting attention: First by specific locations or directions, second by specific characteristics, and third by objects, that have not been in the focus of attention yet. The combination of these methods of attention shift enables the model to create complex search strategies through the production rules. As ACT-R has a build-in memory decay mechanism, it might be possible to predict driver errors because the chunks encoding the current environment decay and can be forgotten if not updated continuously. Another source of possible driver errors could be the potential failure in encoding relevant information (e.g. to overlook a traffic sign or a vehicle).

## Decision Making

The information provided by the control and monitoring component is used to determine if and what decisions must be made on the tactical level concerning the maneuvering (e.g. stopping or overtaking). Our focus on decision making assumes that the attributes in the environment are encoded correctly. The decision how to proceed (in what state of the model) is based on a pattern matching with the knowledge about the environment. If there is no crossing encoded, the corresponding production rule will not fire and the primary loop will continue. The decision whether to stop or to continue driving depends on the encoded traffic sign or on other vehicles. In our environment, the model always recognized these situations correctly, but it would be interesting in future implementations to observe the behavior and decision of the model in case, an error during encoding of attributes occurs.

In order for the model to produce a decision making process similar to humans, encoding a visual attribute and shifting visual attention cannot occur at the same time. For this model, the focus of attention is for example either on the near or far point or encoding a traffic sign.

## Results and Discussion

Obviously, the model presented here does not account for all aspects of human attention during driving, especially not in a naturalistic environment. There are still quite some practical limitations in both, the architecture and the modeling effort itself. This study is an attempt to capture some of the difficult behaviors involved in driving. It also shows some of the limitations of the ACT-R architecture.

This study presents a simple simulation environment and a cognitive model of driver attention during car driving that is able to interact during run-time.

To obtain an integrated driver model of human driving behavior, it is essential to develop models in an architecture which is not task specific and can also model human behavior also in a different context, like ACT-R. This model is a first attempt to recognize, still simplified, traffic signs and crossroads and might make a first step towards the vision of accident-free driving. A majority (over 80%) of the automobile accidents are caused by the driver themselves (Statistisches Bundesamt, 2011). Nearly 16% of the accidents happen while turning and during exit, followed by disregarding the right of way (15%) and not-adapted speed (15%). Theoretically, the cognitive driver model could give a deeper insight for around 30% of the human errors while driving. However, it has to be taken into account that the model is still interacting with a simplified environment and not yet taking into account driver's prior experience, which could be implemented by an increased attention in potentially high accident risk situations. The model and the environment do not present a complete picture of driver behavior yet, but they form a base to extend the ETA framework in any direction.

The ACT-R architecture limits the employment of the three components control, monitoring and decision making by using a serial cognitive processor. The serial processing of the subtasks is typical for the human bottleneck of information processing (Anderson et al, 2004). The resulting model is not an optimal model in a mathematical sense, but approximates human behavior and makes it possible, to mimic human cognitive capacities, simulate the dynamic nature of human driving behavior, and therefore to produce a cognitive adequate model of human driving behavior. If the model is, for example, occupied with an attention shift, it cannot simultaneously update the near point. Also, the model can only fire on production rule at a time and only one visual operation can be executed at a time. These processes take a certain time which are written in an output file. This file contains the time, the active buffer and the according event. This enables the researcher to compare the produced data with human data.

The knowledge representation comprehends declarative knowledge in chunks and procedural knowledge in production rules. For example, the scenario at a crossroad was implemented in 73 explicit production rules, which are highly detailed and is therefore open to future extensions of the model.

This study did not validate the model data so far. Future research could compare the output file data with human data, specially compare the attention shift of the model to human drivers over eye-tracking and the reaction times. But one must remember, that only most critical parts of key scenarios can be validated as no single method is sufficient enough to understand the complex task of human driving behavior yet.

### ACT-R for complex tasks

Modeling such complex tasks in the cognitive architecture ACT-R presents quite some technical challenges to the modeler. For a complex driving task and the validation, the ACT-R model and the participants should interact in the same environment. However, for this validation, it must be possible from the technical side to connect the ACT-R model directly to the simulation environment, which can be technical challenging. Also thinkable is to develop a Lisp version of a driving simulator which can easily interact with ACT-R. If the simulator allows to extract the same information ACT-R does, the output files could be compared, even though the multiple implementation might be a potential source of errors. Also, the current version of ACT-R has some difficulties to directly recognize other components than the already implemented. The attempt to model such complex tasks in generic cognitive architectures show the applicability as well as the still remaining technical limitations.

However, such a complex task might raise the question not only about the limitations of the architecture itself, but also the modeling of human behavior. It might be a good approach to study the key scenarios of human attention during driving in more detail and transfer these results into the model code, breaking down the overall complex task into smaller subtasks in specific situations.

## Conclusion and Outlook

We hope that this research will motivate more members of the computational modeling community to study human attention during driving a car and to overcome the practical limitations. Modeling of such complex tasks holds great promise for meeting the modeling challenges.

The progress to date in the development of cognitive architectures has been impressive, yet scientific gaps, technical challenges and practical issues remain. On one hand, cognitive models help to develop an understanding of driver behavior and aim to provide a theoretical account for human attention while driving. On the other hand, they are powerful and practical tools when implementing human-centered design and real-world applications. First steps towards the examination of the source of human mistakes through distraction from the primary driving task through secondary tasks like dialing a phone haven been taken (Salvucci, 2001) showing the feasibility of the architecture for these task and possible extensions.

The ACT-R architecture enables to elucidate interesting aspects and provides a theory of human attention while

driving. At the same time, human attention during driving is a challenging task for the ACT-R cognitive architecture. It shows the still existing limitations beyond basic laboratory tasks and pushes the research community to expand the architecture towards more complex and finally real-world tasks.

### Acknowledgments

We are grateful for the kind advice and assistance of Prof. Bernhard Nebel, the support of Prof. Wolfram Burgard (University of Freiburg), and for intensive discussions with Dario Salvucci (University of Philadelphia). This work has been partially supported by a grant from the DFG to MR (Project R8-CSPACE in the SFB/TR8 "Spatial Cognition").

### References

- Aasman, J. (1995). Modeling driver behavior in Soar. In: Leidschendam, The Netherlands: KPN Research
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C. & Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review*, 111, 1036
- Anderson, J. R. & Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Lawrence Erlbaum.
- Bellet, T., Bailly, B., Mayenobe, P., & Georgeon, O. (2007). Modelling Driver Behavior in Automotive Environments. Critical Issues in Driver Interactions with Intelligent Transportation Systems. *Cognitive Modelling and Computational Simulation of Driver Mental Activities*.
- Benmimoun, A. (2004). Der Fahrer als Vorbild für Fahrerassistenzsysteme? Ein fahrermodellbasierter Ansatz zur Entwicklung von situationsadaptiven FAS. 13. Aachener Kolloquium Fahrzeug- und Motorentechnik
- Boer, E. R. (1996). Tangent point oriented curve negotiation. *IEEE Proceedings of the Intelligent Vehicles 96 Symposium*
- Byrne, M. D. (2001). ACT-R/PM and menu selection: Applying a cognitive architecture to HCI. *International Journal of Human-Computer Studies*, 55, 41-84
- Land, M., & Horwood, J. (1995). Which part of the road guide steering? *Nature*, 3, 77, 339-340
- Liu, Y. (1996). Queuing network modeling of elementary mental processes. *Psychological Review*, 103, 116-136
- Liu, Y., Feyen, R., & Tsimhoni, O. (2006). Queuing Network-Model Human Processor (QN-MHP): A computational Architecture for Multitasking Performance in Human-Machine Systems. *ACM Transactions on Computer-Human Interaction* 13, 37-70
- Llewellyn, L. (1971). Visual guidance of locomotion. *Journal of Experimental Psychology*, 91, 245-254
- Michon, J. A. (1985). A critical view of driver behavior models: What do we know, what should we do? *Human behavior and traffic safety*, 485-52, Plenum Press
- Pomerlau, D., & Jochem T. (1996). Rapidly adapting machine vision for automated vehicle steering. *IEEE Expert*, 112, 19-27
- Reid, L. D., Solowka, E. N., & Billing, A. M. (1981). A systematic study of driver behavior steering control models. *Ergonomics*, 24, 447-462
- Salvucci, D. D. (2001). Predicting the Effects of In-Car Interface Use on Driver Performance: An Integrated Model Approach. *International Journal of Human-Computer Studies*, 55, 85-107
- Salvucci, D. D. (2006). Modeling Driver Behavior in a Cognitive Architecture. *Human Factors*, 48, 362-380
- Salvucci, D. D., Liu, A., & Boer, E. R. (2001). Control and monitoring during lane changes. *Vision in Vehicles*, 9
- Salvucci, D. D., & Gray, R. (2004). A Two-Point Visual Control Model of Steering. *Perception*, 33, 1233
- Statistisches Bundesamt (online 20.12.2011) [www.destatis.de](http://www.destatis.de)



# Testing the Split Attention Effect on Learning in a Natural Educational Setting Using an Intelligent Tutoring System for Geometry

Robert GM Hausmann ([bhausmann@carnegielearning.com](mailto:bhausmann@carnegielearning.com))

Annalies Vuong ([avuong@carnegielearning.com](mailto:avuong@carnegielearning.com))

Carnegie Learning, Inc.

437 Grant Street, Frick Building, 20th Floor

Pittsburgh, PA 15219 USA

## Abstract

Intelligent tutoring systems (ITS) are a successful application of cognitive science theory to the field of education. Data generated by students using an ITS can also be used to test the external validity of cognitive science principles developed largely in laboratory settings. The present paper collected data from high-school students using two versions of *Cognitive Tutor*, an ITS for Geometry, to assess the impact of eliminating the split-attention effect. The two versions differed in the extent to which the interface required split attention during problem solving. One version used integrated diagrams whereas the other used non-integrated tables and diagrams. Results suggested that students needed fewer problems to master skills in the integrated version, and this was particularly true for mastering difficult skills. This study demonstrates the successful use of cognitive science principles to improve learning through empirically and theoretically derived enhancements to an ITS used in a natural educational setting.

**Keywords:** Intelligent tutoring systems, mathematics instruction, split-attention effect, cognitive load theory.

## Introduction

One of the promises of cognitive science is that it informs the design and implementation of effective instruction and educational tasks (Bruer, 1997). Unfortunately, there is often a disconnect between instructional tasks, as they are originally designed, and the actual implementation of those tasks in the classroom (Stein, Smith, Henningsen, & Silver, 2000; p. 4). One way to partially mitigate this danger is to design instructional tasks in software. Ideally, the design of the software is based on a cognitive theory of learning. Several intelligent tutoring systems (ITS) have been designed based on cognitive theories, including constraint-based reasoning (Mitrovic & Ohlsson, 1999), failure-driven learning (VanLehn, 1988), and the ACT-R theory of human cognition (Anderson, Boyle, Corbett, & Lewis, 1990).

The designers of intelligent tutoring systems face at least two challenges. First, they must demonstrate a learning benefit above and beyond traditional classroom materials and activities. More importantly, an ITS should be able to demonstrate continuous improvements to learning as the theories and empirical findings from cognitive and learning sciences advance.

The purpose of the current paper is to evaluate how enhancements to the ITS *Cognitive Tutor: Geometry* affected student learning in real-world, educational settings. The ITS modifications were based on the predictions of

cognitive load theory, which claims that learning is harmed when a student splits his or her attention across interdependent sources of information. The so-called “split-attention effect” inspired an *in vivo* study in which a single unit from *Cognitive Tutor* was heavily revised to reduce split attention caused by the user interface. The goal of this paper is to extend the generalizability of that *in vivo* experiment by conducting a more in-depth analysis of student learning using data collected from real students using two different versions of the commercially available *Cognitive Tutor*.

## Cognitive Tutor

*Cognitive Tutor* is an intelligent tutoring system inspired by the ACT-R theory of human cognition. *Cognitive Tutor* is based on the pedagogical principle that knowledge is decomposed into knowledge components called *skills*, and learning is maximized when the student is responsible for actively taking each problem-solving step. A cognitive model tracks if the student takes a step off the ideal solution path. Student modeling also allows the tutor to provide immediate feedback, as well as help, in the form of hints, at any step during problem solving. The *Cognitive Tutor* operationally defines “mastery” when the probability that a student knows a skill reaches a threshold of 95%.

*Cognitive Tutor* has been evaluated for its efficacy in both the laboratory and the classroom (Anderson, Corbett, Koedinger, & Pelletier, 1995; Koedinger, Anderson, Hadley, & Mark, 1997), as well in randomized field trials (Ritter, Kulikowich, Lei, McGuire, & Morgan, 2007). A majority of the aforementioned studies used traditional learning materials, such as textbooks and paper-and-pencil homework assignments, as the baseline learning condition. Summarizing over several studies, Corbett (2001) estimates the effect size of *Cognitive Tutor* to be around one standard deviation above traditional instructional materials.

Although effective, there is still room for improvement given that one-on-one human tutoring, when combined with mastery learning, produces about a two standard deviation increase in learning (Bloom, 1984). One way to improve an ITS is to go back and reanalyze the design of individual units and ask if there are any opportunities for enhancing the student’s interaction with the target material. One of the pedagogical commitments of *Cognitive Tutor* is to “minimize working memory load” (Anderson et al., 1995; p.

180). Therefore, the next section discusses cognitive load theory as it applies to geometry instruction.

## Cognitive Load Theory and the Split-Attention Effect

According to cognitive load theory, learning is most likely to take place when the learning environment maximizes *germane* load and minimizes *extraneous* load. Germane cognitive load is defined as “load devoted to the processing, construction and automation of schemas;” whereas extraneous load is defined as “load generated by the manner in which information is presented to learners and is under the control of instructional designers” (Chandler & Sweller, 1991).

Solving problems in geometry is most likely to include both types of load. For example, it is often the case that geometry problems are stated verbally, and they are accompanied by a diagram. The student’s first task is to map the given information, stated in the problem scenario, onto the figure. For example, this would require that the student holds an angle name and its measure in working memory (e.g.,  $m\angle ABC = 15^\circ$ ) while locating the relevant angle in the diagram. This is an example of the *split-attention effect* (Kalyuga, Chandler, & Sweller, 1999).

Holding the angle name and its measure in working memory is not directly relevant to learning how to solve these types of problems; therefore, the working memory load imposed on the student is considered an extraneous load. According to cognitive load theory, instructional designers are recommended that they create a learning environment that minimizes extraneous load caused by the split-attention effect.

Based on the hypothesis that splitting one’s attention across multiple sources of information harms learning, Butcher and Aleven (2008) conducted an *in vivo* experiment where they contrasted classroom learning from two different versions of *Cognitive Tutor: Geometry*. The traditional interface included a verbal statement of the given information, a diagram, and a table. The table was the focus of the student interactions and required students to enter the measure of each angle, as well as a reason justifying the calculation. For the experimental interface, all of the inputs were made directly in the diagram. Students entered their measures and reasons by clicking on angles in the diagram. The learning results from that study suggested that the interactive diagram was easier to learn from, especially in terms of a delayed posttest for numerical test items.

Because the study by Butcher and Aleven (2008) was conducted with a relatively restricted sample of students ( $n = 58$ ) and a single unit of instruction, there is an open question as to whether a change to the interface translates to classroom learning. Will the results replicate when they are implemented in the “wild?”

To address this question, we conducted an analysis of log files generated by students using one of two different versions of *Cognitive Tutor: Geometry* that differed in terms of the split attention required by the user interface.

## Method

### Participants

We compared the usage data from two different versions of *Cognitive Tutor: Geometry* software developed by Carnegie Learning, which are described in the section below. User log files generated by the tutor contain detailed information about every action taken in the interface, including latencies, errors, and access to the various forms of help (i.e., requesting hints, accessing the glossary, reading the lesson page, or studying the interactive example).

Two cohorts of students used two different versions of the software. Approximately 10% of the schools that use Carnegie Learning products were randomly selected to collect log files from their students. For the current study, that translates into approximately  $n = 1,577$  students for the 2009 version and  $n = 2,168$  students for the 2010 version.

### Materials

The interface for several units and sections of the *Cognitive Tutor: Geometry* curriculum were revised to reduce the split-attention effect by using interactive diagrams. Those units include: Pythagorean Theorem, Angle Relationships in a Triangle, and Special Right Triangles.

**Table Interface (v.2009).** Previous versions of *Cognitive Tutor: Geometry* included an interface similar to the one described as the control condition from Butcher and Aleven (2008). The interface included a static diagram, a verbal statement (in paragraph form) of the givens and the sought, and a table of angles in which the student is tasked with calculating the measure and providing a rationale for the calculation (see Fig. 1).

Angle	Measure	Reason
m.∠RTS	30.8	Given
m.∠RST	75.4	Given
m.∠RTU	90	Right Angle
m.∠TRU	59.2	Triangle Sum Theorem
m.∠SRU	90	Right Angle
m.∠SRU	14.6	Triangle Sum Theorem
m.∠TRS	73.8	Angle Addition Postulate

Figure 1: Table Interface (v.2009).

**Interactive Diagram Interface (v.2010).** In an effort to reduce the split-attention effect, several units/sections of geometry were modified to use an “interactive diagram.” The design of the interactive diagrams was similar, but not identical, to the design of the circle tutor used in Butcher and Aleven (2008). All student interactions were handled in the diagram itself. Students had the ability to click on individual angles. Once an angle was selected, a *flyout*

(shown with a blue background in Fig. 2) appeared with several input fields, including the angle measure, a reason field where the student justifies her calculation, and drop-down menus to select the other angles that participate in the target angle's calculation. When the tutor determined that each entry was complete, a summary appeared under the "Diagram Notes," and the diagram itself was labeled with the angle measure.

As previously mentioned, the revised design was intended to reduce working memory load by externalizing some of the information. Because students can *see* an angle in the diagram, they are no longer burdened with holding the angle's name and measure in working memory. Theoretically, this should provide students with more cognitive resources to search the problem space (Larkin & Simon, 1987) and generate domain-relevant inferences.

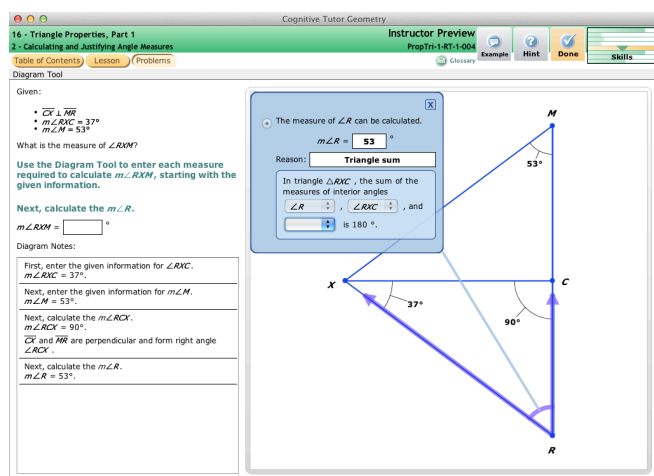


Figure 2: Interactive Diagram Interface (v.2010).

## Results

Due to the large sample sizes, all of the independent, two-sample tests were significant with an alpha level of  $\alpha = .01$ ; thus, instead of reporting  $p$ -values, we relied on Cohen's  $d$  as an effect-size indicator. This allows for a better estimate of the practical significance of the differences.

The results are broken down into two sections. The first section analyzes learning at the unit level. Given that units from v.2009 were reorganized in v.2010, this made one-to-one comparisons at the unit level difficult. We therefore analyzed learning at a finer level of granularity by focusing on the learning of individual skills, regardless of the unit in which the skill appeared.

To control for school-related differences, we replicated all skill-level analyses by restricting our sample to students from the same school. There were small numeric differences in the magnitude of the effect sizes, but they were all in the same direction and interpretation category (e.g., small [ $d = .20$ ], medium [ $d = .50$ ], & large [ $d = .80$ ]); therefore, we collapsed across schools in all subsequent analyses.

## Learning Measured at the Unit Level

The software organizes the learning material hierarchically. Units form the highest level of organization, which are subdivided into sections. Sections are further broken down into problems, which can be further subdivided into individual skills.

We used two different measures to evaluate learning at the unit level. The first was the median number of *Problems* the students solved before graduating the unit. Graduation was defined as mastering all skills. Second, we measured the total amount of *Time* (in minutes) spent in the unit. The results for unit-level measures are summarized in Table 1.

For the problem metric, students using the table interface generally needed to solve more problems in the tutoring system than students using the interactive diagram interface. This was true for Special Right Triangles (5+33 vs. 27 in v.2009 and v.2010, respectively) and Angle Relationships (39 vs. 9+7). The results were reversed, however, for the Pythagorean Theorem unit (5 vs. 9).

The results for the total amount of time spent in the tutor were largely consistent (and correlated with) the number of problems the students solved. The biggest time saving was observed for Angle Relationships (220.01 vs. 70.67+32.22).

Table 1: Unit-level comparisons between the two versions of the software.

Unit	Year	Section Num.	$n$	Problems to Grad.	Time (min.)
Pythagorean Theorem	2009	1	1,577	5	19.34
	2010	1	2,168	9	38.54
Special Right Triangles	2009	1	858	5	9.67
		2	745	33	54.56
	2010	1	1,197	27	67.71
Angle Relationships	2009	1	331	39	220.01
	2010	1	1,213	9	70.67
		2	899	7	32.22

One potential explanation for the results in which the table interface demonstrated better performance than the interactive diagram interface could be due to a difference in the distribution of material across sections. For example, the Special Right Triangles unit originally included two sections: "Finding the Lengths of Sides of a 45-45-90 Triangle" and "Finding the Lengths of Sides of a 30-60-90 Triangle." In the revised version, these sections were combined to form a single section: "Calculating the Lengths of Sides of Special Right Triangles." Combining the sections may have increased the difficulty because students were required to discriminate between the principles necessary to solve two different types of problems.

A similar case could be made for the Pythagorean Theorem unit. The original unit included a section that only required students to solve for the length of the hypotenuse. However, in the revised version, either the hypotenuse or

the leg could be the sought value. Again, students were required to make finer-grained discriminations in the revised version, which could have accounted for the increased time required to graduate from the unit.

### Learning Measured at the Skill Level

As the previous paragraph suggests, the two versions of *Cognitive Tutor* changed in more ways than just the interface. In some cases, the geometry units were rearranged such that the section breakdown was different from one year to the next; therefore, measures of learning could be confounded by section-level changes.

To control for these potentially confounding factors, we measured learning on individual skills, with learning operationally defined as the number of problems the students solved to master each skill. As stated previously, mastery was achieved when the probability of a student knowing a skill reached 95%. To ensure a fair comparison, for this analysis we focused on skills that were consistent between the two versions.

The skills for each unit are presented in separate sections below. The name and id of each skill can be found in the Appendix.

**Special Right Triangles.** Two types of triangles were covered in this section: 45-45-90 and 30-60-90. For the easier triangles (45-45-90), the revised version using interactive diagrams actually led to worse performance in that students needed more problems to master these skills in the revised interface (see Table 2; shaded values). Effect sizes ranged from small ( $d = -0.33$ ) to large ( $d = -2.54$ ).

The reverse, however, was true for the more challenging triangles (30-60-90). The revised interface reduced the number of problems needed to master the associated skills. Effect sizes ranged between ( $d = .33 - .46$ ).

Table 2: Skill comparisons for Special Right Triangles.

Skill ID	2009		2010		$d$
	$n$	$\bar{x}$ ( $SD$ )	$n$	$\bar{x}$ ( $SD$ )	
SR-45_01	871	3.09 (1.9)	991	3.99 (3.41)	-0.33
SR-45_02	785	1.52 (1.06)	1028	6.32 (7.66)	-0.88
SR-45_03	819	3.33 (1.2)	938	16.67 (7.81)	-2.39
SR-45_04	798	2.35 (1.03)	934	16.6 (7.88)	-2.54
SR-30_01	476	21.78 (11.36)	893	17.27 (7.79)	0.46
SR-30_02	558	19.68 (11.8)	991	16.29 (7.65)	0.34
SR-30_03	478	21.56 (12.07)	1000	16.63 (6.96)	0.50
SR-30_04	543	21.08 (12.13)	1009	17.89 (5.93)	0.33
SR-30_05	577	16.29 (10.95)	924	19.5 (5.85)	-0.36
SR-30_06	465	22.31 (12.04)	890	17.57 (7.95)	0.46

Note: Skill IDs refer to Special Right Triangles, followed by the first angle measure (e.g., 45 or 30).

One potential explanation for the inconsistent results is that the previous version was easier than the revised version because it separated the two special right triangles into their own sections (see Table 1). When students solved 45-45-90 problems, they did not have to discriminate between shorter, longer, or equal leg lengths when calculating the non-given side. Restricting our analyses to the 2009 sample, students demonstrated fewer errors while solving 45-45-90 problems ( $M = 5.15$ ,  $SD = 5.93$ ) than 30-60-90 problems ( $M = 39.53$ ,  $SD = 37.35$ ),  $d = 1.29$ . This suggests that, at the section level, the 45-45-90 problems were easier to solve.

**Angle Relationships in a Triangle.** The skills associated with the unit “Angle Relationships in a Triangle” were more consistent. For these skills, the revised interface showed a marked reduction in the average number of problems the students solved before mastering their skills. The effect sizes ranged between medium ( $d = .57$ ) and large ( $d = 2.66$ ), with a majority of the skills falling in the large category (see Table 3).

The lone exception was the first skill, which asks the students to “Enter given value.” It seems that entering the given value was slightly easier in the original table interface that required students to map between the verbal description and entering the given in a table. This might be because the answer of the top row of the table is *always* the given value, whereas a small amount of search is required to enter the given in the interactive diagram.

Table 3: Skill comparisons for Angle Relationships.

Skill ID	2009		2010		$d$
	$n$	$\bar{x}$ ( $SD$ )	$n$	$\bar{x}$ ( $SD$ )	
Ang_Re_01	1140	1.86 (1.27)	901	2.29 (0.98)	-0.38
Ang_Re_02	1244	4.38 (5.83)	891	1.79 (1.56)	0.61
Ang_Re_03 <sup>1</sup>	455	9.86 (5.43)	887	2.19 (2.49)	1.82
Ang_Re_04	455	5.90 (8.8)	887	2.19 (2.49)	0.57
Ang_Re_05	369	24.50 (16.83)	887	2.19 (2.49)	1.85
Ang_Re_06	344	29.80 (16.41)	887	2.19 (2.49)	2.35
Ang_Re_07	344	30.17 (16.35)	887	2.19 (2.49)	2.39
Ang_Re_08	331	31.86 (15.59)	887	2.19 (2.49)	2.66
Ang_Re_09	344	29.79 (16)	887	2.19 (2.49)	2.41
Ang_Re_10	450	10.10 (13.31)	889	3.88 (2.2)	0.65

**Pythagorean Theorem.** The final section in which there were matching skills was the unit on the Pythagorean Theorem. The two skills in this section that fit our criteria both demonstrated an advantage for the interactive diagram. Students using the revised interface needed fewer problems to solve both skills (i.e., calculate the length of the

<sup>1</sup> “Ang\_Re\_03” through “\_09” have the same statistics because the 2010 skill included each of the 2009 variants as sub-skills.

hypotenuse in and out of a contextual scenario). The effect sizes were both considered “large” ( $d > .90$ ; see Table 4).

Table 4: Skill comparisons for Pythagorean Theorem.

Skill ID	2009		2010		$d$
	$n$	$\bar{x}$ ( $SD$ )	$n$	$\bar{x}$ ( $SD$ )	
Pythag_01	1663	15.64 (8.79)	2338	8.84 (4.42)	0.98
Pythag_02	1702	15.38 (8.9)	2338	8.84 (4.42)	0.93

## Discussion

Using data gathered from the use of an intelligent tutoring system (ITS) in natural educational settings, the current study demonstrates how an already effective intelligent tutoring system can be further refined through the application of cognitive and learning theories. The current study draws from research on the “split-attention effect,” which demonstrates that performance on a task is greatly reduced when the student must split her attention across interdependent sources of information. Learning is greatly reduced because working memory is tasked with holding a large number of chunks of information. According to cognitive load theory, when that large burden on working memory is not relevant to abstracting principles from the domain, then this leads to an “extraneous load.” Students are not able to transfer the knowledge that is inferred from problem solving to long-term memory.

The split-attention effect is particularly relevant to solving geometry problems in an ITS, where the student is required to split her attention across a verbal scenario that states given information, a diagram that depicts relationships between segments and angles, and a table that holds information about each angle.

Butcher and Aleven (2008) demonstrated, in an *in vivo* study, that a revised ITS interface can enhance learning both immediately and over the long term. On the basis of their strong results, the *Cognitive Tutor: Geometry* interface was revised to emulate the same type of interaction. With the “interactive diagrams,” students were given the chance to concentrate the focus of their attention on the learning materials.

Although the *in vivo* results were strong, there was a chance that crucial design elements did not get directly translated into the commercial version of the software. A comparison of screenshots between the Butcher and Aleven (2008; Fig. 1) and the current study (Figs. 1 & 2) reveals that there were subtle design differences. For example, the original study modified a unit on circles, whereas the current study mainly concentrated on triangles. There may be subtle content differences that lend themselves more or less well to learning gains through interactive diagrams. Second, the information in the interactive circle diagrams was echoed in a table; whereas, the triangle tutor included a “Diagram Notes” panel with similar, but differently formatted, information.

The current results support the generalization that small design differences can have a measurable impact on learning. At the unit level, there were generally mixed results with some, but not all, units demonstrating a reduced amount of time spend solving problems. Analysis at the more fine-grained level of individual skills yielded more consistent results. Most, but not all, of the skills associated with the interactive diagram showed a positive effect. Some of the skills that showed an increase in the number of problems required to master the skills seemed to fit into one of two categories. Either the skills were very easy (i.e., “enter given”) or they were embedded in a particularly easy unit. In these cases, it might have been better to rely on the old design. For more difficult skills, however, there was a definite advantage to interacting directly with the diagram.

Although the results are encouraging, the current set of analyses could be improved in the following ways. First, this was not an experimental study. Students were not randomly assigned to condition; therefore, the conclusions that we can draw from these analyses are strictly correlational. However, these results are suggestive and point to interesting new research projects. For example, subsequent research should test the hypothesis that interactive diagrams are especially helpful for more difficult topics.

Another improvement on the current analyses would be to assess “robust” learning, which is defined as learning that is retained over a long interval, transfers to new situations, and helps accelerate learning of subsequent material (Koedinger, Corbett, & Perfetti, 2010). Because this was an analysis of the log files generated by student users, we were not privy to the students’ pre- and post-test scores. Future analyses will look at post-requisite materials available in the tutor and evaluate if there is any evidence of transfer or accelerated future learning. Although it may have taken students more problems to master the skills presented in the Special Right Triangles unit, students might be able to transfer their knowledge more accurately when they were required to struggle with deciding which rule applies within the collapsed Special Right Triangle section (e.g., desirable difficulties; Bjork, 1994).

In addition, we would also like to conduct further analyses to determine whether the changes in the design features affected the learning curves of the matched (i.e., comparable) skills.

In conclusion, it is widely acknowledged that learning geometry is challenging. As instructional designers and members of the cognitive science community, it is incumbent upon us to ensure that learning difficult science, technology, engineering, or math (STEM) topics is both efficient and robust. One way to continuously improve our methods of instruction is to keep going back and testing our learning environments against the most recent empirical and theoretical developments. The current study takes an important step in that direction.



## Acknowledgments

The authors would like to thank the schools that allowed us to collect log files from their students. More importantly, we wish to thank all of the anonymous students who used our software to solve their geometry homework problems. Special thanks to Leslie Hausmann for commenting on a previous version of this paper.

## References

- Anderson, J. R., Boyle, C. F., Corbett, A. T., & Lewis, M. W. (1990). Cognitive modeling and intelligent tutoring. *Artificial Intelligence*, 42, 7-49.
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4, 167-207.
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. (J. Metcalfe & A. P. Shimamura, Eds.) *Metacognition Knowing about knowing*. The MIT Press. Retrieved from <http://psycnet.apa.org/psycinfo/1994-97967-009>
- Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6), 4-16.
- Bruer, J. T. (1997). Education and the brain: A bridge too far. *Educational Researcher*, 26(8), 4-16.
- Butcher, K. R., & Aleven, V. A. (2008). Diagram Interaction during Intelligent Tutoring in Geometry: Support for Knowledge Retention and Deep Understanding. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1736-1741). Austin, TX: Cognitive Science Society.
- Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4), 293-332.
- Corbett, A. T. (2001). Cognitive computer tutors: Solving the two-sigma problem. In M. Bauer, P. J. Gmytrasiewicz, & J. Vassileva (Eds.), *User Modeling* (pp. 137-147). Berlin: Springer-Verlag. doi:10.1007/3-540-44566-8
- Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology*, 13(4), 351-371.
- Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8(1), 30-43.
- Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2010). *The Knowledge-Learning-Instruction (KLI) Framework: Toward Bridging the Science-Practice Chasm to Enhance Robust Student Learning*. Cognitive Science. Pittsburgh, PA. Retrieved from <http://www.learnlab.org/documents/KLI-Framework-Tech-Report.pdf>
- Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11, 65-99.
- Mitrovic, A., & Ohlsson, S. (1999). Evaluation of a constraint-based tutor for a database language. *International Journal of Artificial Intelligence in Education*, 10(3-4), 238-256.
- Ritter, S., Kulikowich, J., Lei, P., McGuire, C. L., & Morgan, P. (2007). What Evidence Matters? A randomized field trial of Cognitive Tutor Algebra I. In T. Hirashima, H. U. Hoppe, & S. S.-C. Young (Eds.), *Supporting Learning Flow Through Integrative Technologies* (Vol. 162, pp. 13-20). IOS Press.
- Stein, M. K., Smith, M. S., Henningsen, M. A., & Silver, E. A. (2000). *Implementing Standards-Based Mathematics Instruction: A Casebook for Professional Development*. New York, NY: Teachers College Press.
- VanLehn, K. (1988). Toward a theory of impasse-driven learning. In H. Mandl & A. Lesgold (Eds.), *Learning issues for intelligent tutoring systems* (pp. 19-41). New York: Springer.

## Appendix

Skill ID	Skill Name
SR-45_01	Enter given side length.
SR-45_02	Calculate leg given other leg in a 45-45-90 triangle.
SR-45_03	Calculate hypotenuse in a 45-45-90 triangle.
SR-45_04	Calculate leg given hypotenuse in a 45-45-90 triangle.
SR-30_01	Calculate longer leg given shorter leg in a 30-60-90 triangle.
SR-30_02	Calculate hypotenuse given shorter leg in a 30-60-90 triangle.
SR-30_03	Calculate shorter leg given hypotenuse in a 30-60-90 triangle.
SR-30_04	Calculate hypotenuse given longer leg in a 30-60-90 triangle.
SR-30_05	Calculate shorter leg given longer leg in a 30-60-90 triangle.
SR-30_06	Calculate longer leg given hypotenuse in a 30-60-90 triangle.
Ang_Re_01	Enter given value.
Ang_Re_02	Enter calculated value.
Ang_Re_03	Enter reason of Right Angle.
Ang_Re_04	Enter reason of Triangle Sum.
Ang_Re_05	Enter reason of Angle Addition or Triangle Sum.
Ang_Re_06	Enter reason of Triangle Exterior Angle.
Ang_Re_07	Enter reason of Linear Pair.
Ang_Re_08	Enter reason of Angle Addition.
Ang_Re_09	Enter reason of Isosceles Triangle.
Ang_Re_10	Enter reason of Equilateral Triangle.
Pythag_01	Find hypotenuse in context
Pythag_02	Find hypotenuse out of context

# The effect of "Maverick": A study of Group Dynamics on Breakthrough in Collaborative Problem solving

Yugo Hayashi ([yhayashi@fc.ritsumei.ac.jp](mailto:yhayashi@fc.ritsumei.ac.jp))

College of Information Science and Engineering, Ritsumeikan University,  
1-1-1 Nojihigashi, Kusatsu, Japan

## Abstract

The presented study is concerned with one aspect of the effect of a "whistleblower" or a person arousing or informing a different perspective on a collaborative problem-solving task. Its purpose is to find out, through an experiment, how a whistle-blower (which we called "Maverick") affects the facilitation of a breakthrough in a rule-discovery task. In the experiment two hypotheses were tested: 1) Collaborative problem-solving task is facilitated by contribution of member with different perspective. 2) Problem-solving task is facilitated more as the number of participant with a different perspective increases. In the experiment, several sets of figures were presented in three different settings (without a different perspective, with Maverick, three members with a different perspective), where a group of six members (one human and five conversational agents) collaboratively engaged in a rule-discovery task via a text-based chat system. The experiment revealed an interesting result to the effect that while a different perspective, overall, contributed to the facilitation of problem-solving such contribution was not statistically significant when it was presented by half of the members. The implications of the result were discussed by referring to the related literature in psychology.

**Keywords:** Collaborative Problem Solving; Different Perspectives; Conversational Agent.

## Introduction

In Cognitive Science, many studies have been conducted to investigate what contributes most to collaborative problem solving. Several of them found that different perspectives promote an interaction between a pair and lead to collaborative problem solving (Miyake, 1986; Shirouzu, Miyake, & Masukawa, 2002; Hayashi, Miwa, & Morita, 2006). Others reported that asking reflective questions to conversational partners is a useful interaction strategy for gaining a deeper understanding about the problem (Okada & Simon, 1997; Miwa, 2004). They argued that the use of verbal probes such as providing clarification questions and suggestions prompted the problem solvers' reflective thinking and metacognition. A study by Shirouzu, Miyake, & Masukawa (2002) also suggested that taking different roles is an effective way to reconstruct the external representation of problem solvers and stimulate creative thinking.

Unfortunately, most of the findings on collaborative problem solving in the past have been based on the experimental data using a few participants such as a pair. Only few studies investigated the interactional aspects of collaborative process in a group of people and it remains unknown what kind of group dynamics actually operates in the perception and interpretation of different perspectives

during the task of collaborative problem solving among several people. In the present study, the author will look into the nature of such operation through an experimental set-up where a different perspective is presented to illusory members of a conversational group.

## Integrating different perspectives of others during collaborative problem solving

Cognitive operations such as combing and integrating different perspectives during problem solving are effective strategies for generating new ideas (Finke, Ward, & Smith, 1992). They are also considered to play important roles in several cognitive domains. In the model of Hegelian dialectic thinking, different perspective is regarded as a key concept for creativity. Also important in this model is a process called the 'aufheben' whereby contradicting different opinions are integrated and interpreted into a higher level of concept (Hegel, 1874). In group problem solving, it is assumed that this process of integrating different perspectives can play an important role during collaborative problem solving.

Hayashi & Miwa (2009) is one of the few experimental studies that examined the effect of cognitive operations such as combining and integrating different perspectives upon collaborative problem solving. In that study, they investigated the nature of such operations in a rule-discovery task where each of a pair had a different perspective. Results of the experiment showed that establishing a common ground between the two through such operations is a key to success in problem solving.

It can be expected that such operations would become more difficult when more than two people with different perspectives engage in problem solving. The present study will focus on the nature of group dynamics of collaborative problem solving in a group, which very few studies have investigated and still remains an unsolved issue in cognitive science (see Figure 1 for an experimental set-up of this study).

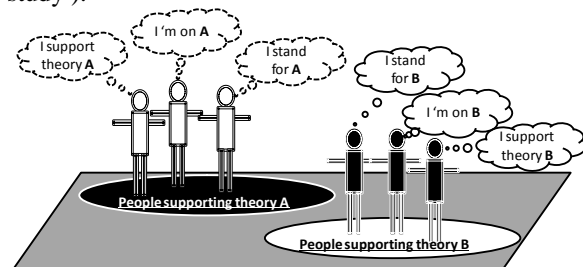


Figure 1. An experimental setting.



## Changing perspectives in group problem solving: Participation of 'whistle blower'

One of the common phenomena observed in problem-solving task is that people are often fixated on or biased in favor of a particular perspective. In such a situation, it is vital that they take a different perspective, but an important question is how we can make that happen.

In the field of organizational study, it has been found that participation by so-called a 'whistle blower' often arouses a different perspective among group members (Elliston, Keenan, Lockhart, & Van Schaick, 1985; Near & Miceli, 1987). In general, a whistle blower is defined as the person who dissents from the laws and systems of the organizations to which she/he belongs. It is said that such person may be perceived as a troublemaker that brings a confusing idea to a group, or, sometimes as a reformer who brings an innovative idea to a group. In either case, whistle blower is known to become a potential provider of an 'anomaly cue' that stimulates reflective thinking. We may predict that participation of whistle blower with a different perspective promote reflective process of collaborative problem solving.

### Aim of this study

Studies discussed above suggest that different perspective provided by whistle blower may provide breakthrough and facilitate their task in collaborative problem solving. The group dynamics of incorporating different perspective may also be affected by several factors such as the number of whistle blower in a group. It may be natural to expect that collaborative problem solving will be facilitated more when a different perspective is suggested by more than one person. The purpose of the present study is to examine these assumptions, by testing the following hypotheses:

- 1) Collaborative problem-solving task is facilitated by contribution of member with different perspective.
- 2) Problem-solving task is facilitated more as the number of participant with a different perspective increases.

### Experiment of design

In this study, we used a modified version of the experimental design in Hayashi et al. (2006), where pairs of participants with different perspectives engaged in a rule-discovery task. The design of the experiment was developed based on the Gestalt theory.

### Controlling the participants' perspective

In the experiment, several sets of random patterns of several figures on a 6 x 6 grid base, each colored black or white, were generated, (see Figure 2). In each set, a pattern consisting of combined square blocks was shown against the background of either black or white background colors. The background color was controlled to derive, through Gestalt effect, the change in problem-solver's perspective (Koffka, 1935).

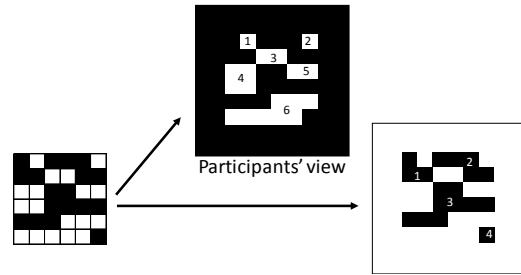


Figure 2. Example of stimuli with ten components (six in white and four in black).

Each set consists of several "objects" (or patterns) in black or white color, each of which consists of a single block or combined multiple blocks. In one example shown in Figure 2, one of the paired objects has a total of ten "components" comprising four black components and six white components. They were shown on a computer display against either black or white background. When a participant is focusing on white components inside a black background, they become the figure and the six components pop out; on the other hand, when a participant is focusing on a black object, it becomes the ground. The default setting was such that the participants easily see figure component in white color (See Figure 2). The alternative perspective is a perspective that suggests figure component in black.

While only two members engaged in the task in Hayashi et al. (2006), in the present study a group of six problem solvers collaboratively worked on the problem solving task through computer terminals connected via a local network. The six members of the problem solvers consisted of one human participant and five chat partners of computer gents. In this study we call member with a different perspective 'maverick'. It is agent that plays the role of group member and focuses on a background color which is different from that the human participant does. Shown in Figure 3 is an illustration of the setting where all six members except one (or Maverick) see four white components against black background.

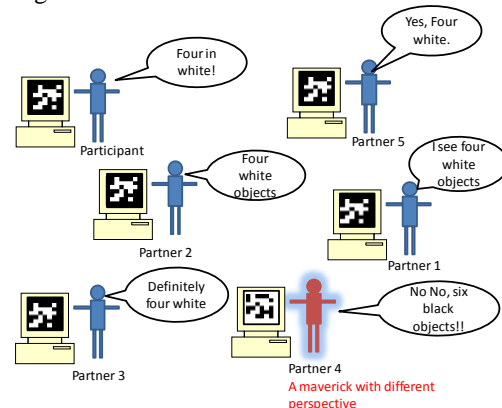


Figure 3. Example of experimental situation.

In the experiment, several types of objects were presented in sequence on a computer display (see Figure 4). For each

object, a square outer-box was shown on the display for one second, which was followed by stimulus picture presented inside the box frame. The number of white components and black components was controlled and the total number of the components presented to the participants was between six and twelve. Sequential pattern of the sums of black components and white components was repeatedly presented (i.e. 6, 8, 10 / 6, 8, 10) (see Table 1).

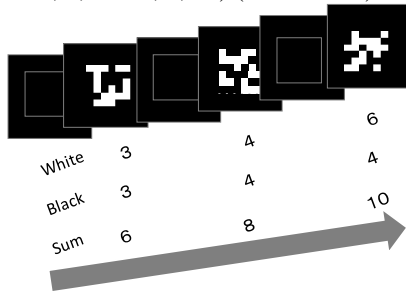


Figure. 4. Series of presented stimuli.

Table 1. Example of sequences of the presented objects.

White	4	6	4	7	2	4	6	5	3
Black (maverick)	2	2	6	5	4	4	4	7	3
Sum of Black and white	6	8	10	12	6	8	10	12	6

In this experiment, the task of a participant is to discover a rule concerning the regularity of the series of sums of the components. If participants are preoccupied to focus on only one of the two background colors, they cannot find a target rule. To make a breakthrough in the task, problem solvers have to take into consideration the number of components that pop out as figure and that of components hidden in the background. In order for a participant to find a rule, he/she needs to focus on the components hidden in the background. In a setting shown in Figure 3, for example, the maverick (partner 4) keeps suggesting an alternative perspective to call their attention to the presence of the black components.

The task required of the participants in the experiment was to type in the chat exchange when engaging in the task. The participants were assigned to always talk first during the chat exchange. Shown in Figure 5 is an illustration of interface record of a text-based exchange among a participant and five other agents that are discussing the problem. During the task, they were able to use buttons at the bottom of the screen to change objects, send messages, and to terminate the experiment. The participants were allowed to write only one sentence of less than 30 characters for each pair of components and were asked to finish the task within 30 minutes.

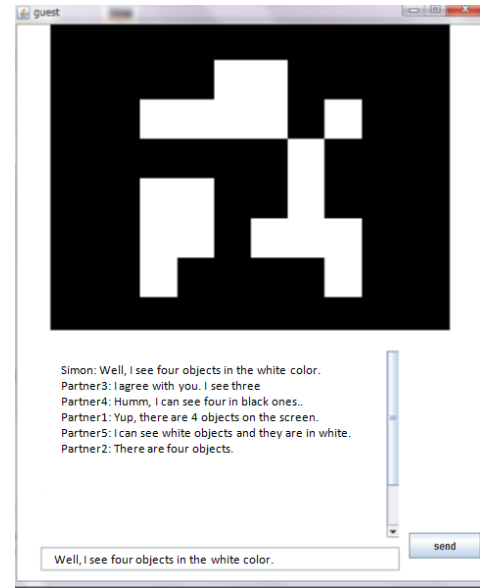


Figure 5. Example of the interface.

## Conversational agents and the experimental system

The system used in the experiment was designed by Java (see Figure 6) and consists of four program modules: (1) Server, (2) Client, (3) Agent, and (4) Problem Generator. Multi-threads, which process all the messages simultaneously, were used for the Server. When messages are sent to the Server, they were re-distributed to all Clients (Agents). The Problem Generator generates the objects that were presented in sequence (see Table 1). This module provided important information about the sequence of the stimuli and the objects presented by GUI, which was used by the agents to generate messages. A simple conversational computer agent used in this study is a typical rule-based system. Based on some pre-defined rules, it can respond meaningfully to sentences that were input by the participants (See Figure 7).

The Semantic Analyzer extracts keywords from input messages and detects keywords relevant to the task. Keywords collected from a previous study were used to build the Dictionary which contained important keywords for the task (Hayashi et al., 2009). Working Memory is created by the Generator, and it consists of two associated database: (a) presented objects (Picture Database), and (b) detected key words (the Semantic Analyzer). Various types of argument statements are stored in Rule Base in the form of 'if-then' format. Definitions from Working Memory are sent to Rule Base to search for matching statements. When there are several overlapping statements, a simple conflict-resolution strategy is utilized. When a matching process ends, selected sentences are sent to the Generator. Then, definitions in Working Memory are updated, and finally, output messages are displayed.

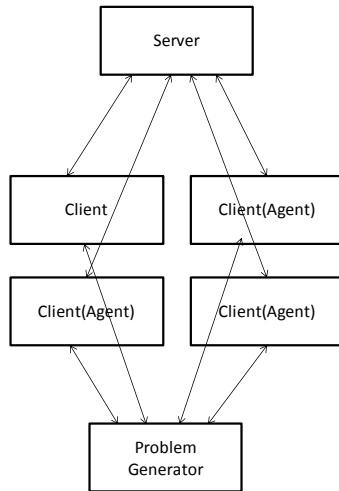


Figure 6. The structure of the system.

The conversational agent described above played the role of partner and produced a virtual experimental environment. The participants of the experiment were instructed that they are interacting with real people, though they were actually interacting with computer agents to solve the task.

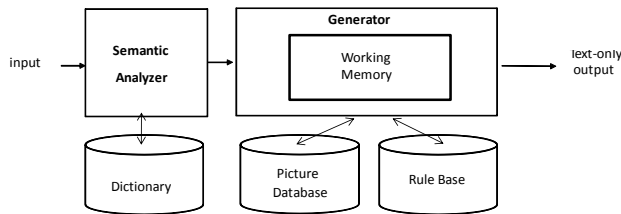


Figure 7. Structure of the agent.

## Experimental Setting

To test the hypotheses, the following three conditions were set up: (1) 6:0 condition, (2) 5:1 condition, and (3) 3:3 condition. The third condition was included to test the second hypothesis. In the first 6:0 condition, all of the members had the same perspective; there were no members with different perspectives in this condition. In the second 5:1 condition, one of the collaborating agents had a different perspective. In the third 3:3 condition, half of the six members had a different perspective. In all conditions, a group of six members consisted of one human participant and five computer agents and a different (or an alternative) perspective based on black background color was provided by computer agent to a human participant (See Figure 8). The participants engaged in the task, without being told that they were interacting with computer agents.

101 undergraduate students participated in the experiment (38 males and 63 females; the average age was 20.3). The experiment took place in a computer room where maximum capacity was 60 people. All participants were randomly assigned to each condition and they were instructed that

they will start the task with someone inside the room. Those participants who did not follow the instructions to answer the final questions or who felt suspicious about their partners were excluded from data. Participants who did not begin the experiment by focusing on the objects by figure color were also excluded. After this screening, the total number of participants that provided to the data was 92 (31 participants each for the 6:0 and the 5:1 condition and 30 participants for the 3:3 condition.)

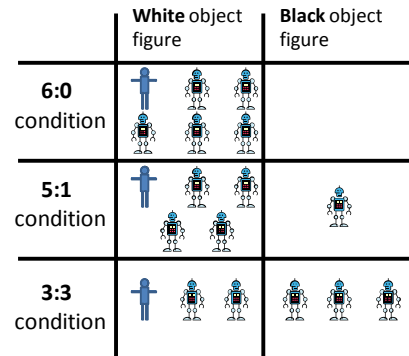


Figure 8. Experimental conditions.

## Dependant variables

After the task, each participant was asked to describe the target rule on an answer sheet. If their answers were in some ways related to the number of black or white components of the objects, they were judged as 'integrated' (e.g. The sums of the black and white components are 6,8,10 and the sequence is repeated in the same order. The difference of the number of the black and white components is between zero to two.) If their answers did not include such information, they were evaluated as 'not integrated'.

All of the answer sheets were also analyzed to evaluate their cognitive process of perspective change. If the conversation data included a description that referred to the background color, it was counted as a token of 'a change of perspective'. (e.g. "I was counting only the white objects, but maybe the black objects have something to do with the target rule...". If their data did not include such information, it was counted as a token of 'no perspective change'. The data was then statistically analyzed for (1) the number of integrated answer, and (2) the process of change in perspective.

## Results

### Problem Solving Performance

The results were analyzed using a 1 x 3 between-subjects factorial design. Figure 9 shows the results of the performance of problem solving. The vertical axis represents the ratio of the problem solving performance, and the horizontal axis represents the experimental condition. The numerals shown on the cylindrical bars indicate the number of participants in each condition.

A Chi square analysis was conducted to verify if the difference of the number of problem solvers who used perspectives of others was statistically significant. An overall analysis suggests that, a group which had a member with a different perspective integrated its perspective into their own perspective more frequently than a group which did not. There were a significant difference among the three conditions ( $\chi^2(2) = 7.189, p < .05$ ). Next, a multiple comparison was conducted on each two conditions using the Fisher's exact test. There was a significant difference between the 6:0 and the 5:1 condition ( $p < .05$ ). On the other hand, the differences between the 6:0 condition and the 3:3 condition were marginal ( $p < .10$ ). There was no significant difference between the conditions of 5:1 and 3:3 ( $p = .41$ ).

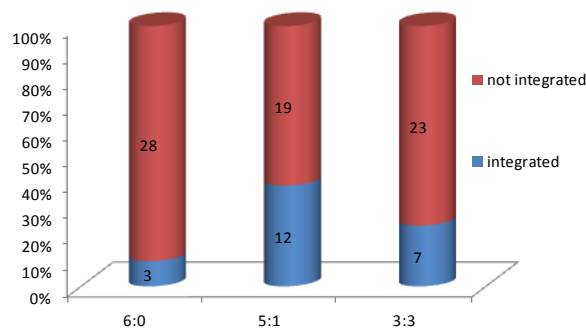


Figure 9. Results of the problem solving performance

### Perspective change process

Figure 10 shows the results of the performance of a change of perspective. The vertical axis represents the ratio of the change of perspective, and the horizontal axis represents the experimental condition. The numerals shown on the cylindrical bars indicate the number of participants in each condition.

A Chi square analysis was conducted to check the difference of the number of problem solvers who integrated their perspectives during the task. Results indicate that there were differences among the conditions ( $\chi^2(2) = 15.230, p < .01$ ). Next, a multiple comparison was conducted on each two conditions using the Fisher's exact test. There were differences between the 6:0 and 5:1 condition ( $p < .01$ ). On the other hand, the differences between 6:0 and 3:3 conditions were marginal ( $p < .10$ ). At last, differences were found between the conditions of 5:1 and 3:3 ( $p < .01$ ).

Like the problem solving performance, an over-all analysis suggests that a group which had members with different perspectives were taking into consideration the perspectives of others more frequently than a group which did not. More importantly, the results showed that the problem solvers were incorporating others' perspectives most when a different perspective was proposed by a single partner; the difference was greater than when different perspectives were proposed by more than one person.

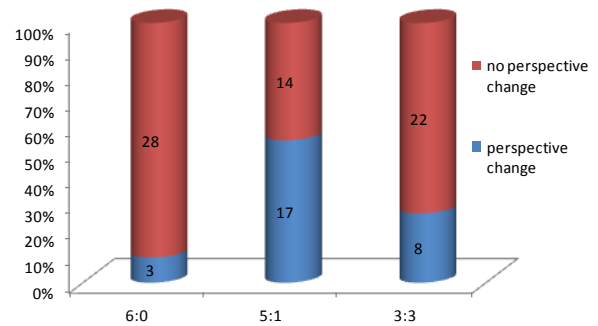


Figure 10. Results of the problem solving process

### The effect of 'Marverick' and experimental condition

The statistical analyses above indicated that in both problem solving performance and perspective change process, a group which had member with a different perspective were incorporating their alternative perspective more frequently than a group which did not have such member. In other words, the over-all analysis revealed the results that support the first hypothesis.

Further analysis showed, however, that the participants were not incorporating an alternative perspective when it was proposed by three members; that is, different to the expectation, the effect of an alternative perspective by a single partner was greater than that by three partners (i.e. half of group members). In other words the results of the experiment did not support the second hypothesis; Marverick outperformed the impact of its alternative perspective than multiple members with an alternative perspectives. In other words, it suggests that people are less inclined to integrate a different perspective when it is suggested from several people. Implications of this finding will be presented below.

### Discussion

It has been pointed out that a different perspective in a group may be favored especially when people face a situation which is comprehensible only vaguely or partially. Though people tend to be influenced by a perspective that is dominant in a group, a breakthrough can take place when a member of a group presents a dissenting view. In psychology this type of social influence is called the "minority influence". In one of the classical studies on 'minority effects', Moscovici & Nemeth (1974) argued that a minority of one is more influential than a minority of more than one. They argued that if one person is consistent with the minority view, over time, it may capture more attention from the majority. In the 5:1 condition of the experiment, suggestions of a different perspective from a single member (i.e., Marverik) was not only consistent but persistent in that it gave suggestions more frequently and over longer time compared to those in the 3:3 condition. That may have aroused the participants to think that the suggestion of

alternative perspective from that person must be very helpful if he/she is so sure of it. One of the possible reasons why the facilitation effect from Marverik was higher in the experiment may have to do with the effect of attention and its consistency. Other studies have shown that the influence of a minority perspective is a desirable condition for increasing the diversity of views, prompting reconsideration, processing information and making a decision (Moscovici, Lage & Naffrechoux, 1969; Nemeth, Brown, & Rogers, 2001). While the situations investigated in these studies are somewhat different from the present study, the result of the experiment showed that minority overall can exert a similar desirable effect to the facilitation of the task of collaborative problem solving where a breakthrough is needed.

The experiment also suggested that a different perspective from a single member may be more easily incorporated to facilitate cooperative problem-solving task than that from multiple members. One of the factors that are involved in the rejection of the second hypothesis may be related to the notion called "groupthink". It is a psychological phenomenon that often occurs when the desire for harmony overrides critical evaluation of own perspective and serious appraisal of alternatives. In a more realistic setting, social influence may force people to adapt a perspective a majority of others take. People may ignore the view of a minority, favoring a certain perspective, whether that perspective is ideal or not. This may take place in the process of decision making in a group where its members want to minimize conflict and reach a consensus decision without critical evaluation. This is similar to a negative effect associated with a whistle blower mentioned above. It may be that this negative effect of groupthink was working for the participants in 3:3 condition.

Also, a different perspective with people with the same opposing perspective may have lead to confusion. This type of confusion is pointed out in the literature of organizational psychology (Pondy, 1967). To confirm why the point investigated by the second hypothesis was rejected, these and other factors may need to be taken into account in a further experiment.

## CONCLUSION

This study investigated the influence of different perspective during collaborative problem solving. The results showed that, while a different perspective is more likely to contribute to problem solving, it was more effective when it was presented by only one person than by several people. The discussion suggested that further study is needed to confirm the latter point.

## Acknowledgments

I want to thank my student advisee Shin Takii (Ritsumeikan University) for helping collect the experimental data. I am deeply gratefully to Hitoshi Ogawa (Ritsumeikan University) for giving me suggestions for the experiment system. I appreciate Hajime Shirouzu (Chukyo University) for mentoring me during the writing action. My deepest

application goes to Kazuhisa Miwa (Nagoya University) who always gives me insightful comments.

This research was partially supported by the Grant-in-Aid for Scientific Research (KAKENHI), The Ministry of Education, Culture, Sports, Science and Technology, Japan (MEXTGrant), Grant No. 23700325

## References

- Elliston, F., Keenan, J., Lockhart, P., & Van Schaick, J., (1985). *Whistle-blowing Research : Methodological and Moral Issues*. New york: Praeger
- Finke, R. A., Ward, T. B., & Smith, S. M., (1992). *Creative Cognition: Theory, Research, and Applications*. Cambridge: The MIT Press
- Hayashi, Y., Miwa, K., & Morita, J., (2006) A laboratory study on distributed problem solving by taking different perspectives. In *Proceedings of the 28th annual conference of the cognitive science society* (pp. 333-338). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hayashi, Y., Miwa, K., (2009) Prior experience and communication media in establishing common ground during collaboration. In *Proceedings of the 31th annual conference of the cognitive science society* (528-531). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hegel, G. W. F., (1874) *The Logic. Encyclopedia of the Philosophical Sciences*. 2nd Edition. London: Oxford University Press
- Koffka, K., (1935) *Principles of gestalt psychology*. Routledge and Kegan Paul
- Miwa, K., (2004) Collaborative discovery in a simple reasoning task. *Cognitive System Research*, 5, 41-62.
- Miyake, N., (1986) Constructive interaction and the interactive process of understanding. *Cognitive Science*, 10, 151-177.
- Moscovici, S., Lage, E., Naffrechoux, M., (1969) Influence of a consistent minority on the response of a majority in a color perception task, *Sociometry*, 32, 365-379.
- Near, J. P., & Miceli, M. P., (1987) Whistle-blowers in organizations: Dissidents or reformers?, *Reserch in Organizational Behavior*, 9, 321-368.
- Nemeth, C., Brown, K., & Rogers, J., (2001) Devil's advocate versus authentic dissent: stimulating quantity and quality, *European Journal of Social Psychology*, 31, 707-720.
- Okada, T., & Simon, H., (1997) Collaborative discovery in a scientific domain. *Cognitive Science*, 21, 109-146.
- Pondy, L. R., (1967), Organizational Conflict: Concepts and models. *Administrative Science Quarterly*, 12, 296-320.
- Shirouzu, H., Miyake, N., & Masukawa, H., (2002) Cognitively active externalization for situated reflection. *Cognitive Science*, 26, 469-501.

# Tightening up Joke Structure: Not by Length Alone

**Christian F. Hempelmann (chempel@purdue.edu)**

CERIAS, Purdue University  
656 Oval Dr., West Lafayette, IN 47907-2086 USA

**Julia M. Taylor (jtaylor1@purdue.edu)**

Department of Computer and Information Technology, Purdue University  
401 N. Grant St., West Lafayette, IN 47907-2021 USA

**Victor Raskin (vraskin@purdue.edu)**

CERIAS, Purdue University  
656 Oval Dr., West Lafayette, IN 47907-2086 USA

## Abstract

The paper seeks to tighten up the notion of joke structure in the context of the Ontological Semantic Theory of Humor for computational use. The method is testing the prior hypothesis that a minimalist version of a joke, consisting only of the setup and punch line, is the most effective one. A small ‘human computation’ pilot study casts serious doubt on this hypothesis.

**Keywords:** humor research; (minimalist) joke structure; setup; punch line; joke versions.

## 1. Introduction

The human ability to communicate is incomplete without humor. If a computational system is ever to approximate this human communicative ability and act as a competent partner in a conversation with a human, humor must be accounted for. Over the last decades, humor research has become an intense multidisciplinary effort with significant contributions from linguistics, psychology, sociology, neuro- and cognitive sciences (Raskin 1985, 2008; Ruch 1998, Oring 1992, Davies 1990, Attardo 1994, Morreall 1983). Along with theories and analyses of human-generated and perceived humor, since the early 1990s, there have been more explorations of computational humor as well, starting with attempts at humor generation through humor detection to semantically based systems (see Section 2.1 below).

Part of the difficulty in relating computational humor-generation and, to a lesser degree, humor-detection systems to human appreciation is the question how much information has to be present in the text of a joke to ensure a successful setup and the most effective punch line. This is precisely what has not been addressed yet on the computational front—how much information is enough and not too much to carry a joke without risking the opposite extremes of crypticality or verbosity.

Two related goals of this line of research are a) to create an NLP system capable of understanding the mechanism of a joke at a level sufficient for providing a punch line to a human-generated setup (even if unintentionally) and b), conversely, for the computer to react competently to a human-generated punch line that follows a setup, generated

by either participant. The first scenario enables the computer to generate humor in reaction to a human cue in human-computer interaction, the second scenario lets the computer identify humor in the same scenario and enables it to react competently to it.

Most existing theories available for humor detection or generation fall short of providing the adequate support for this task. These theories are either too fine-grained to be useful or too coarse to correctly classify any given text as a joke or a non-joke. But our ontological-semantic system provides a sufficiently rich and flexible basis because it operates at the level of human text-meaning processing. In the following, we will summarize the state of the art, introduce our approach, and then discuss a pilot study assessing human appreciation of jokes in variants of different length and types of manipulation.

## 2. Background

### 2.1. State of the Art

The usefulness of and motivations for computational humor, along with its feasibility, have been intensely discussed (see Ritchie 2004, Hempelmann 2008, Taylor 2008, Strapparava et al. 2011 and references in all of these sources). The most useful work on computational humor is based on a humor theory and seeks to gain further insights, to validate, and to improve the theory, while taking advantage of its assets. Work on humor theories has a long history, and, to this day, the true multifaceted nature of humor is still being debated (Raskin 1985, Morreall 1983, Oring 1992, Ruch 1998, Davies 1990, Attardo 1994): there is no universally accepted theory of humor that explains “what is funny, why it is funny, how it is funny, when it is funny, and to whom it is funny.” (Raskin 1985: 5).

The linguistic theories of humor (Raskin 1985, Attardo & Raskin 1991) have reached a level of formal representation that is adaptable for the computation of any humorous text (Raskin et al. 2009a,b). But the best-known and most-used linguistic theory of humor remains the early Script-based Semantic Theory of Humor (SSTH: Raskin 1985).



According to the SSTH, there are two conditions for a text to be humorous:

- A text has to be compatible, fully or in part, with two different scripts.
- The two scripts with which the text is compatible are opposite, and the text must overlap fully or partially with them.

The compatibility of the text with two scripts is the necessary condition for humor; the oppositeness of the scripts is the sufficient condition. The former was to be detected in the course of normal semantic analysis; the latter was not included at that point.

The central concept, that of a script, is defined as “an enriched, structured chunk of semantic information, associated with word meaning and evoked by specific words. The script is also a cognitive structure internalized by the native speaker, and it represents the native speaker’s knowledge of a small part of the world. [...] Formally or technically, every script is a graph with lexical nodes and semantic links between the nodes” (Raskin 1985: 81). Scripts were further developed, formally and computationally, in Ontological Semantics (Nirenburg and Raskin 2004, Raskin et al. 2003), and the current, third stage of the theory, the Ontological Semantic Theory of Humor (OSTH), has the functionality both to perform the computational semantic analysis that establishes the necessary compatibility of scripts and encompasses their sufficient oppositeness.

The scripts can be linguistic, general knowledge, restricted, or individual. Linguistic scripts are known to any “average,” “standard” native speaker (adult, reasonably educated, mainstream culture, etc). General knowledge scripts, such as crossing the street or going to a store, are known to a large number of people and are not affected by their use of language. Restricted knowledge scripts are known to a smaller number of people and are not affected by their use of language either. Individual scripts are “owned” by one person: an example of an individual script would be a child’s memory of her first swim.

The General Theory of Verbal Humor (GTVH: Attardo & Raskin 1991), is an extended, second-stage multidisciplinary theory of humor that is also built upon the notion of script overlap and script oppositeness. The theory, empirically verified in Ruch et al (1993), describes jokes in terms of six knowledge resources: *Script Opposition* (SO), informed largely by linguistics, deals with script overlap and oppositeness presented in Script-based Semantic Theory of Humor (SSTH); *Logical Mechanism* (LM), informed by logic and cognitive psychology, accounts for the way in which the two scripts in the joke are brought together in a faulty, but locally valid way; *Situation* (SI), informed by many disciplines, contains the “props” of the joke, the textual materials evoked by the scripts of the joke that are not necessarily funny; *Target* (TA), informed by sociology, represents any individual or group from whom humorous behavior is expected; *Narrative Strategy* (NS) is the rhetorical structure of the text; *Language* (LA) is the actual

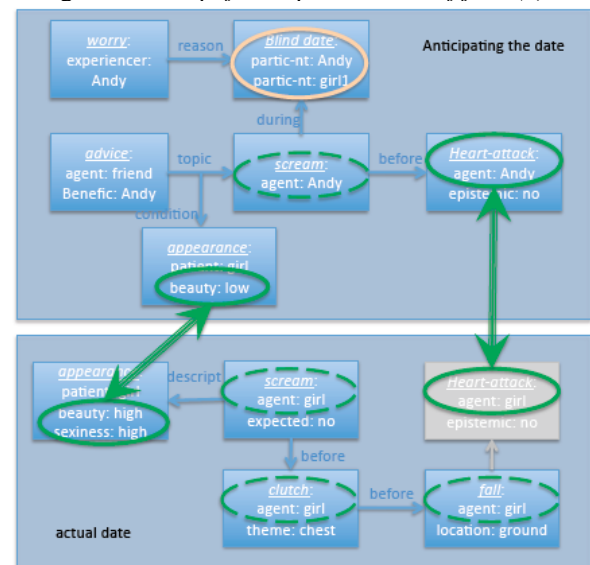
lexical, syntactic, phonological, etc., choices at the linguistic level that instantiate all the other choices. According to the GTVH, each joke can be viewed as a 6-parameter vector (Ruch et al. 1993): Joke = {SO, LM, SI, TA, NS, LA}.

## 2.2. Ontological Semantic Theory of Humor

Since Raskin’s (1985) definition of scripts and the general conditions for a text to be humorous, the definition and interpretation of script overlap and oppositeness have been debated (Attardo et al. 2002, Hempelmann 2003, Taylor 2008). Basing the GTVH on Ontological Semantic Technology (OST) allows a crisper definition of the necessary and sufficient conditions for verbal humor (Raskin et al. 2009a,b).

At the core of OST (Raskin et al. 2010, Hempelmann et al. 2010, Taylor and Raskin 2011, Taylor et al. 2010) are repositories of world and linguistic knowledge, acquired semi-automatically (or, rather, in hybrid automatic and human computation—see Law and von Ann 2011) and used to disambiguate the different meanings of words and sentences and to represent them comprehensively. These repositories consist of the ontology, containing language-independent concepts and relationships between them; one lexicon per supported language, containing word senses anchored in the ontology which is used to represent their meaning; and the onomasticon, which contains names of people, countries, organizations, etc., and their descriptions, also anchoring them in ontological concepts and interlinking them with its other entries.

Figure 1: A simplified representation of joke (4)



The lexicon and ontology are used by the OST Processor, a software that produces Text Meaning Representations (TMRs) from the text that it reads. The format of TMRs conforms to the format and interpretation of the ontology. The processed TMRs are entered into the Information Repository, from which information is used for further



processing and reasoning. Recent implementation of components of the system have produced successful results.

OST progress has enabled enhanced meaning representation of all the components of the joke, shedding light even on such less linguistic ones as the Target and Narrative Strategy (see a much simplified graphic representation of Joke 4 below in Figure 1). OSTH is reconsidering the six GTVH knowledge resources with the additional emphasis on providing the ontological support to tighten and straighten their definitions and conditions of usage. However, one troubling, even if expected result, of the formalization in OSTH is the realization that the SO of GTVH and SSTH was defined inadequately. Because SO constitutes the decisive factor in determining whether a text is a joke, and thus dominates other knowledge resources, the current theories have to be modified and revised to an extent for future research, the rationale and pilot study for which we are presenting here, will help to determine.

### 3. Joke Variants

One of the recent discoveries, part of gaining new insights into the Narrative Strategy within OSTH, is the apparent need of some extra material right before the punch line (see Taylor 2010, 2011). More generally, by observing the coexistence of different versions of the same jokes, we realized that some extra parts of jokes, in the setup and punch line, may have specific functionalities, while others are pure ballast contributing nothing but verbosity. To the best of our knowledge, the contribution of seemingly inessential information in jokes has never been systematically studied.

Understanding the seeming importance of extra material is needed to detect the essential and necessary information for a joke to make sense and to be effective. Kuipers (2006: 204), for example, found that both in the United States and the Netherlands, longer jokes are considered generally funnier than shorter ones. Such information cannot be measured in the number of words but rather by the tightness/non-redundancy of the underlying conceptual structures. An initial approach along these lines, proposing “meaning density” as a factor in joke funniness was presented in Hempelmann (2011). What allows for testing this assumption is the fact that the same joke often exists in several attested versions. One attractively simple hypothesis may thus be that the essential information of all versions of the joke is the conceptual structure of the minimalist version of the joke. The rationale for computing the essential information is to understand the proliferation of multiple versions, of widely varying lengths and genres, of the same joke—in conversational practice, in print and on the Internet—and to test whether their “common core” can carry the joke on its own.

To illustrate this point, let us compare two versions of the following blind date joke:

(1) Danny sets up Andy to go on a blind date with Shirley, a friend of a friend of his. But Andy is a little worried about going out with someone he has never seen

before. “What do I do if she’s ugly?” says Andy, “I’ll be stuck with her all night.” “Don’t worry.” Danny says. “Just go up to her door and meet her first. If you like what you see, then everything goes as planned. If you don’t, just shout Aaaaauuggghhh! clutch your chest and fake a heart attack.” So that night, Andy knocks at Shirley’s door, and when she comes out he is awe-struck at how beautiful and sexy she is. Andy’s about to speak when the girl suddenly shouts, “Aaaaauuggghhh!”, clutches her chest and falls to the ground.

(2) Andy is going on a blind date but is worried that she may turn out to be ugly. A friend advises him to fake a heart attack if it turns out to be the case. When Andy arrives, the door is open by a sexy and beautiful woman, who suddenly clutches her chest and falls to the ground.

Both versions contain the same scripts, roughly corresponding to the anticipation of the blind date and the actual event. The second version is minimalistic in that it contains virtually nothing that can be removed from the text without rendering it incomprehensible and useless as a joke. The first version adds much additional detail. The second version is synthesized, and it is possible that it has lost too much, and some supporting detail would actually improve it. This optimality is of crucial significance in generating a joke by a computer. To put it differently, removing information that is redundant for a plain expository text may result in a significant loss for a joke, as demonstrated in (3).

(3) Andy is going on a blind date but is worried that she may turn out to be ugly. A friend advises him to fake a heart attack then. The date turns out to be sexy and beautiful, but she suddenly clutches her chest and falls to the ground.

It has been established in humor theory that the punch line has to be short, and preferably by far to conclude the joke (see, for instance, Attardo et al. 1994). What has not been adequately researched is the punch line parameters, including its boundaries and most effective delivery mode, especially how minimalistic it can and should be. It has been suggested (see, for instance, Giora 2002) that including a familiar element within an innovative stimulus leads to more pleasure for the subjects than a purely innovative stimulus. Our preliminary research seems to indicate that, while, generally, accompanying information can be removed from the setup, some seemingly disposable elements may have to be left in the punch line.

Thus, if we compare (3) above to (4) below, the former reads more like the serious report of a somewhat funny event than as a joke, while the latter is easier to perceive as a joke.

(4) Andy is going on a blind date but is worried that she may turn out to be ugly. A friend advises him to scream and fake a heart attack then. The date turns out to be sexy and beautiful, but she suddenly screams “aaaauuhhh,” clutches her chest and falls to the ground.

The difference between these two versions is presence of the clause *she suddenly screams ‘aaaauuhhh’* in (4). Its precise contribution to the text is something that we are interested in establishing in this pilot.

#### 4. Pilot Study

To test the hypotheses outlined in the previous section and explore general effects between joke variants to generate more hypotheses for more formal future inquiry, we created a small test corpus for a pilot study.

This corpus of fifty stimuli consists of 10 jokes found online in at least two variants differing in length, complemented by another three synthesized variants for each joke: one is the minimalist version, condensing the joke to a summary only mentioning the necessary and, presumably sufficient information for the joke to be operational; another is this minimalist version together with a dialogue element in the Narrative Structure of the joke, since we realized in creating the non-dialogue minimalist version that the joke seemed to us radically decreased in funniness; the third artificial variant was added to be just that, a control version based on the longer real variant of the joke, to see if artificial manipulation in itself affected perceived funniness. The rationale for the final version is based on the fact that jokes are folkloristic creations optimized by iterations of retelling and not owned by individual authors, a characteristic that does not hold for cut-and-paste online joke collections created to generate traffic. In sum, the five variants<sup>1</sup> for each of the ten jokes are:

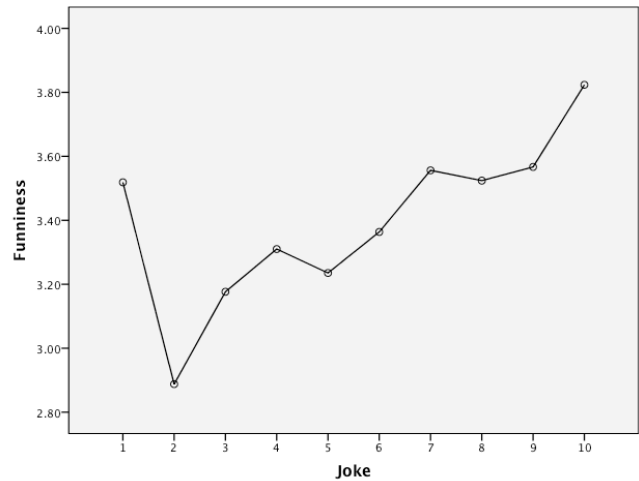
- long non-manipulated version
- shorter manipulated version
- non-dialogue manipulated minimalist version
- dialogue manipulated minimalist version
- longer manipulated minimalist version with paraphrasing

In this pilot study, we recruited raters for the funniness of these variants using Amazon's Mechanical Turk, a method generally deemed valid (Buhrmester et al. 2011) and now proclaimed to be a form of 'human computation' (Law and von Ann 2011), with an incentive of \$0.10 for participation. The Mechanical Turk aims to filter bots and human responders who don't pay attention to the instructions in several ways. This includes the researcher's ability to block certain countries, including those where non-native English issues might affect the research issue, and to select only participants who have had a certain number of approved assignments in the past. In addition to these controls, in a second pilot study we included as the only difference from the first study reported here one additional stimulus that instead of a punch line had the direction to rate it at a given level of funniness. We then excluded all responses who didn't follow this direction under the assumption that the raters didn't read the instructions and clicked through the responses randomly. Interestingly, there was no significant difference in the results to the initial study.

In both versions of this pilot study, the participants, of whom we aimed for 200, were directed to a survey in

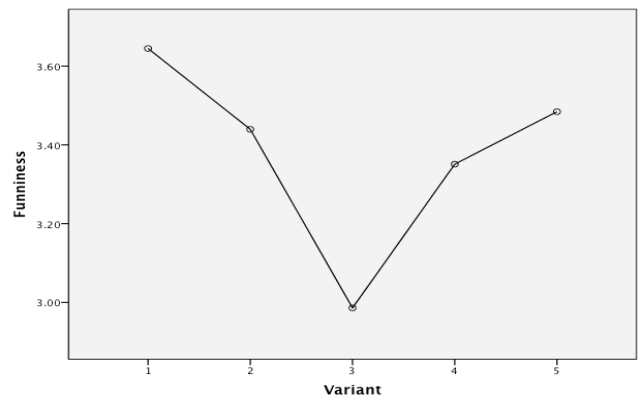
Qualtrics, in which they were presented with a random variant of each of the ten jokes, in random order of joke. Ideally, we would have gathered 40 ratings for each of the 50 variants, but some variation exists because raters who took under 2 minutes for their 10 stimuli or had the same rating for each stimulus were excluded, resulting in a sample of 176 participants for the version without the control stimulus presented here.

Figure 2: Mean funniness for each joke across all variants



An Analyses of Variance (ANOVA) for repeated measurements for the funniness of the ten different jokes across all variants revealed that the jokes were perceived to differ in funniness,  $F(8.38, 1558.03) = 11.38, p < .001$ . Joke 2 ("Matthew Chapter 11") was considered significantly less funny than the other jokes, while joke 10 (cheap parking in New York City) was deemed funnier than all others (see Figure 2). Low numbers did not allow for the exclusion of these two jokes from further analysis homing in on the variants.

Figure 3: Mean funniness of variant type across all jokes



More importantly for the line of inquiry that we are pursuing here, a second ANOVA for repeated measurements for the funniness of the five variants of jokes revealed significant differences,  $F(3.72, 338.48) = 7.43, p <$

<sup>1</sup> For lack of space we can't include all jokes here, but the minimalist versions are in the appendix, while the full text of all variants of all jokes can be found at: [http://web.ics.purdue.edu/~vraskin/joke\\_variants.pdf](http://web.ics.purdue.edu/~vraskin/joke_variants.pdf)

.001. Figure 3 shows that joke variant type 3, the manipulated minimalist version without dialogue, was rated as the least funny and significantly so against all other variant types. This effect is most interesting in contrast to the other minimalist manipulated variant type 4 that does include a dialogue in its Narrative Structure. The importance of this result, privileging dialogue over exposition, warrants further investigation in our continued research.

Table 1 shows the levels of significance for the effects summarized in Figure 3.

Table 1. *Within-subjects contrasts for the pairs of joke variant (v) types.*

	v 2	v 3	v 4	v 5
v 1	2.71	25.91***	4.18*	1.69
v 2		12.13**	0.41	0.20
v 3			7.30**	19.68***
v 4				1.06

Note. Cells contain F-statistics for the contrasts,  $F(1, 91)$ .

\*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ .

Clearly, the minimalist versions fared worst and the longer versions overall were deemed funnier, if not significantly so. But some amount of non-essential information clearly accounts for relevant degrees of funniness of jokes.

## Summary and Outlook

With respect to our initial hypotheses, our results confirm that condensing jokes so they only contain the SO-relevant information is not optimizing their funniness. Something is lost in the process and the difference between the dialogue and non-dialogue manipulated variants seems to point at the importance of that NS factor. A further speculation that should be explored based on these results is that the faultiness of the logical mechanism might no longer be sufficiently hinted at to make it retrievable, rendering the oppositeness of those variants too blunt. In terms of a classic linguistic distinction, these initial findings are pointing at the importance of performance-related factors in jokes, in contrast to purely competence-based relation of information. In other words, joke texts have an aesthetic dimension that has yet to be allocated more clearly to a part of the OSTH model in future follow-up studies with further careful manipulations of joke variants.

## Acknowledgments

The authors are grateful to Ursula Beermann, Department of Psychology, University of California at Berkeley, for her help with the pilot study. They would also like to thank the IRB of Purdue University for the prompt granting of the exemption.

## References

Attardo, S. (1994). *Linguistic Theories of Humor*. Berlin-New York: Mouton de Gruyter.

- Attardo, S., & Raskin, V. (1991). Script theory revis(it)ed: joke similarity and joke representation model. *Humor: International Journal of Humor Research* 4(3-4): 293-347.
- Attardo, S., Hempelmann, C. F., & Di Maio, S. (2002). Script oppositions and logical mechanisms: Modeling incongruities and their resolutions. *Humor: International Journal of Humor Research* 15(1): 3-46.
- Buhrmester, M., Kwang, T., & Gosling, S.D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1), 3-5.
- Davies, C. (1990). *Ethnic Humor Around the World: A Comparative Analysis*. Bloomington, IN: Indiana University Press.
- Hempelmann, C. F. (2008). Computational humor: Going beyond the pun. In V. Raskin (Ed.), *The Primer of Humor Research*, Berlin-New York: Mouton de Gruyter, 333-360.
- Hempelmann, C. F. (2011) Meaning Density: Explicitness is Proverbially and at This Point in This Title Also Obviously the Death of Wit, or is It? *2011 Annual Meeting of the International Society for Humor Studies—ISHS'11*. Boston, MA: Boston University
- Hempelmann, C. F., Taylor, J. M., & Raskin, V. (2010). Application-guided Ontological Engineering. *Proceedings of the International Conference on Artificial Intelligence*, Las Vegas, NE.
- Kuipers, G. (2006). *Good Humor, Bad Taste*. Berlin, New York: Mouton de Gruyter.
- Law, E., and von Ann, L. (2011). *Human Computation*. No place: Morgan & Claypool.
- Morreall, J. (1983). *Taking Humor Seriously*. Albany: SUNY Press.
- Nirenburg, S., & Raskin, V. (2004). *Ontological Semantics*. Cambridge, MA: MIT Press.
- Oring, E. 1992. *Jokes and Their Relations*. Lexington, KY: University Press of Kentucky
- Raskin, V. (1985). *Semantic Mechanisms of Humor*. Dordrecht: Reidel.
- Raskin, V. (1990). Sophistication in humor and beyond. In M. Glazer (ed.), *Abstracts of the Eighth International Conference on Humor*, Sheffield, UK: University of Sheffield Press.
- Raskin, V. (2005). The threshold of triviality in telling tales: Is it inherent in inferences? In S. Attardo and L. Birden (eds.), *Abstracts of ISHS 2005: The 17<sup>th</sup> Annual Conference of the International Society of Humor Studies*, Youngstown, OH: Youngstown State University.
- Raskin, V. (2008). Theory of humor and practice of humor research: Editor's notes and thoughts. In V. Raskin (ed.), *The Primer of Humor Research*. Berlin-New York: Mouton de Gruyter, 1-15.
- Raskin, V., Hempelmann, C. F., & Taylor, J. M. (2009a). How to understand and assess a theory: The evolution of the SSTH into the GTVH and now into the OSTH, *Journal of Literary Theory* 3(2): 285-312.

- Raskin, V., Hempelmann, C. F., & Taylor, J. M. (2009b). Symposium 'From SSTH to GTVH to OSTH—and never ever back!' 2009 Annual Meeting of the International Society for Humor Studies—ISHS'09: International Conference on Humor Research, Long Beach, CA: University of California.
- Raskin, V., Hempelmann, C. F., & Taylor, J. M. (2010). Guessing vs. knowing: The two approaches to semantics in natural language processing, *Annual International Conference Dialogue 2010*, 642-650, Bekasovo (Moscow), Russia.
- Raskin, V., Nirenburg, S., Hempelmann, C. F., Nirenburg, I., & Triezenberg, K. E. (2003). The genesis of a script for bankruptcy in ontological semantics. In G. Hirst & S. Nirenburg (Eds.), *Proceedings of the Workshop on Text Meaning, 2003 NAACL Human Language Technology Conference*, 27-31.
- Ritchie, G. (2004). *The Linguistic Analysis of Jokes*. London-New York: Routledge.
- Ruch, W. (Ed.) (1998). *The Sense of Humor: Explorations of a Personality Characteristic*. Berlin: Mouton de Gruyter.
- Ruch, W., Attardo, S., & Raskin, V. (1993). Toward an empirical verification of the General Theory of Verbal Humor. *Humor: International Journal of Humor Research* 6(2), 123-136.
- Strapparava, C., Stock, O., Mihalcea, R. (2011). Computational humour. In Cowie, R., Petta, P., and Pelachaud, C. (Eds.) (2011), *Emotion-Oriented Systems: The HUMAINE Handbook, Cognitive Technologies*, chapter 6.4. Berlin: Springer.
- Taylor, J. M. (2008). *Towards Informal Computer Human Communication: Detecting Humor in Restricted Domain*, Ph. D thesis, Department of Electrical and Computer Engineering, University of Cincinnati, 2008.
- Taylor, J. M. (2010). Ontology-based view of natural language meaning: The case of humor detection, *Journal of Ambient Intelligence and Humanized Computing* 1(3), 221-234.
- Taylor, J. M. (2011). Does SO2 always result in a joke: How long is long enough? *2011 Annual Meeting of the International Society for Humor Studies—ISHS'11*. Boston, MA: Boston University.
- Taylor, J. M., and Raskin, V. (2011). Understanding the unknown: Unattested input processing in natural language, *Proc. FUZZ-IEEE-11*, Taipei, Taiwan.
- Taylor, J. M., Hempelmann, C. F., & Raskin, V. (2010). On an automatic acquisition toolbox for ontologies and lexicons in Ontological Semantics, *International Conference on Artificial Intelligence*, Las Vegas, NE.
- 2 Matthew has been in business for many years, and suddenly the business is going down the drain. When he looked for advice by opening the Bible on a random page. It read, "Matthew. Chapter 11."
- 3 A store manager overhears one of his salesmen telling to a customer that the store hasn't had something for awhile and it doesn't look like they'll be getting any soon.
- The manager yells after the departing customer to come back next week because surely they'd have it by then.irate, he reprimands his salesman for telling a customer they're out of anything and asks what the customer wanted. It was rain.
- 4 A man gets pulled over by a policeman, who tells the man that his wife fell out of the car about a mile back. The man is relieved because he thought he'd gone deaf.
- 5 The Pope was finally persuaded by his cardinals to sleep with a woman, so that he could better understand the problems of mankind. The Pope agrees, but insists that she has to have certain qualifications: first, she has to be blind, so she cannot see who is doing it to her; second, she has to be mute, so she can't tell anyone what happened; and third, she has to have big tits.
- 6 The door bell rings at the whorehouse. A girl who answers the door, sees a guy with no arms and no legs and asks what he thinks he's going to do in there.
- The guy points out that he was able to ring the bell after all.
- 7 On a bus, a punk kid with red, green and orange hair notices an old guy staring at him. When he asks the old man if he himself never did anything wild in his time, it turns out that the old man once screwed a parrot and was wondering whether the punk was his son.
- 8 One day an angel made a male and a female statues that have faced each other in a park for decades come alive to do anything they wanted for thirty minutes. The two dashed for the bushes, whose branches started shaking while there was giggling and laughter. Fifteen minutes later, they emerged with wide grins on their faces and they still had fifteen more minutes. Then the female statue said to the male that this time he should hold the pigeon down and she'll poop on its head.
- 9 As a woman gets on a bus with her baby, and the driver tells her that hers is the ugliest baby he's ever seen. Angrily, she complains to a man in the rear of the bus that the driver just insulted her. The man suggests she go and tell the bus driver off and offers to hold her monkey for her.
- 10 A man walks into a bank in New York City and asks for a \$4000 loan. The bank teller agrees to accept the man's black Porsche parked in the bank's parking garage as security. A few weeks later the man returns to pay off his loan and the interest of \$11 dollars. The manager wonders why the man needed to borrow \$4000 dollars, since the bank found out that he was a millionaire. The man replies that nowhere else in New York can he park his car for three weeks for \$11 dollars.

## Appendix: Minimalist Joke Variants

1 Andy is going on a blind date but is worried that she may turn out to be ugly. A friend advises him to fake a heart attack then. The date turns out to be beautiful and sexy, but she suddenly clutches her chest and falls to the ground.

# Logical or Pragmatic, as Long as it Suits our Convenience: Scalar Inferences in a Pro-and Contra-attitudinal Context

**Tom Heyman** (tom\_heyman123@hotmail.com)

Department of Experimental Psychology, 102 Tiensestraat  
3000 Leuven, Belgium

**Walter Schaeken** (Walter.Schaeken@ppw.kuleuven.be)

Department of Experimental Psychology, 102 Tiensestraat  
3000 Leuven, Belgium

**Katrijn Pipijn** (Katrijn.Pipijn@ppw.kuleuven.be)

Department of Experimental Psychology, 102 Tiensestraat  
3000 Leuven, Belgium

## Abstract

In the present study we propose a context wherein the endorsement rate of the scalar inference from ‘some’ to ‘not all’ either increases or decreases. It is known that people tend to interpret the quantifier ‘some’ as ‘not all’, though logically some means ‘some and possibly all’. However, we argue that this tendency to derive the scalar inference is variable and depends on the attitude of the reader or listener. When the ‘not all’ interpretation implies a confirmation of one’s attitude, we expect a higher endorsement rate of the inference. On the other hand when the ‘some and possibly all’ interpretation contains pro-attitudinal information, we expect a decrease in endorsement rate. These predictions are derived from Kunda’s theory of motivated reasoning (1990) and are supported by the data. Theoretical implications and suggestions for further research along this line are discussed.

**Keywords:** Scalar inferences; Motivated reasoning; Attitudes; Relevance theory.

## Introduction

Understanding utterances requires knowledge about grammar and the semantics of words, but also about non-linguistic properties such as the speaker’s intentions and the context in which the utterance is expressed. The interpretation of utterances has been studied extensively and belongs to pragmatics (Sperber & Wilson, 1986). Within the field of pragmatics one can distinguish different topics such as deixis, presupposition, speech acts, implicatures (Levinson, 1983). However, in the present paper we will only focus on the latter. An implicature, a term introduced by Grice (1975), is what is suggested by an utterance though not explicitly stated nor logically deducible. Consider the following classic example:

- (1) They got married and had a baby.

Here, it is suggested that they first got married and then had children (i.e., the implicature), though logically the sequence of events could also be reversed. A lot of research has been devoted to a subclass of implicatures called scalar implicatures. When a weaker term (e.g., ‘some’) is used to imply the negation of a more informative, stronger term

(e.g., ‘all’), it is called a scalar implicature. For instance, when confronted with a sentence like (2), people tend to interpret ‘some countries’ as ‘not all countries’. Although logically, ‘some countries’ could mean ‘all countries’, people rarely take this possibility into account.

- (2) Some countries are poor.

The explanation for this tendency can be found in Grice’s cooperative principle (1975, 1989). According to this principle, people should make their contributions such as it is required, when they interact. For that purpose, he proposed four maxims, enabling people to communicate effectively. One of these maxims, the maxim of quantity, states that any contribution to a conversation should be as informative as possible. Returning to the example, if the speaker knows for a fact that all countries are poor, he would have used the stronger quantifier ‘all’. However, the speaker employed the weaker term ‘some’, which is taken to mean that the stronger term is not appropriate (either because the speaker knows that not all countries are poor or because he is unsure that all countries are poor).

Looking through the literature, one can find two contradicting opinions concerning the mechanism underlying the production of scalar inferences (Noveck & Reboul, 2008). The neo-gricean account (e.g., Levinson, 2000) assumes that the ‘not all’ interpretation (also called the pragmatic interpretation) is derived by default. A logical interpretation of ‘some’ may be possible, but only when the pragmatic interpretation is cancelled in a certain context. Chierchia (2004) has also proposed a default theory in which he argues that scalar inferences are derived by default except when the scalar term occurs in a downward entailing context. These contexts include negations, question forms and antecedents of conditionals (Noveck & Reboul, 2008). On the other hand, there are theorists who argue against the default character of scalar inferences. Instead, they defend a contextualist view, which comprehends that the narrowing from ‘some’ to ‘not all’ is entirely determined by the context. The best known of these theories is probably Relevance theory (Sperber & Wilson, 1986, 1995; Wilson &

Sperber 2003). This theory states that people will only derive a scalar inference when it yields sufficient positive cognitive effects. To interpret an utterance as (2), listeners and readers are thought to follow a path of least effort in computing cognitive effects and to stop when their expectation of relevance is satisfied.

Evidence from different experimental studies seems to favour Relevance theory. Studies involving response time measures showed that a logical interpretation of underinformative utterances like ‘some oaks are trees’ takes less time than a pragmatic interpretation (Bott & Noveck, 2004; Noveck & Posada 2003). Also when participants were instructed to respond quickly, they produced less scalar inferences (Bott & Noveck 2004). Moreover, De Neys and Schaeken (2007) found that people made more logical and fewer pragmatic interpretations under high cognitive load. Breheny, Katsos and Williams (2006) took another approach and manipulated the context wherein the implicature was embedded. They found longer reading times in an upper-bound context (i.e., a context that warrants the scalar inference) than in a lower-bound context (i.e., a context that makes the inference inappropriate). Recently, two studies examined the computation of scalar inferences using a visual world paradigm (Grodner, Klein, Carbary & Tanenhaus, 2010; Huang & Snedeker, 2009). Where Grodner et al. (2010) report that the scalar inference is computed immediately, Huang & Snedeker (2009) found a short delay in the computation of the scalar inference. Since the latter is in line with Relevance theory and other experimental studies (Bott & Noveck, 2004; Noveck & Posada, 2003), we can conclude that most of the experimental evidence favours Relevance theory.

As already mentioned, Relevance theory states that the scalar inference from ‘some’ to ‘not all’ is highly dependent on the context. Some theorists have tried to identify contexts in which the availability of the scalar inference varied (Bonneton, Feeney & Villejoubert, 2009; Breheny et al., 2006). Two of them, the lower-bound and upper-bound contexts, were already discussed above. Furthermore, Bonneton et al. (2009) established that the endorsement of the inference declines in face-threatening contexts. In this paper we propose a different context wherein the occurrence of the scalar inference depends on one’s attitudes. Such a context is more subjective (compared with more objective upper/lower-bound and face-threatening contexts) since scalar inferences here depend on feelings and motives of people.

In the literature, one can find many examples of how reasoning is affected by beliefs and attitudes. Performance on the Wason selection task for instance improved significantly when the task rule was contra-attitudinal because participants were motivated to look for disconfirmation, which leads to the correct solution to the task (Dawson, Gilovich & Regan, 2002). Kunda (1990) developed a theory that explained how reasoning might be influenced by attitudes. The theory of motivated reasoning, as it is called, postulates that when one is motivated to

arrive at a particular conclusion, one applies certain strategies that are considered most likely to yield the desired conclusion. Put differently, if confronted with contra-attitudinal information, one is motivated to reject this information (i.e., disconfirmation bias). On the other hand if the information is consistent with one’s attitude, one is inclined to believe in it (i.e., confirmation bias). With this in mind, it should be possible to design a context where the scalar inference is either appropriate or inappropriate. Consider the following example:

- (3) a. in countries where the death penalty is applied, crime has decreased.
- b. in some countries where the death penalty is applied, crime has decreased.
- c. not all countries applying death penalty experience decreasing crime.
- d. some and possibly all countries applying death penalty experience decreasing crime.

Depending on someone’s attitude towards the death penalty, the utterance (3a) is either pro-or contra-attitudinal. For people in favour of the death penalty (3a) is in line with their attitude whereas for people opposed to death penalty it is contrary to their beliefs. If the utterance were prefaced by the scalar “some” as in (3b), it can be interpreted either pragmatically (3c) or logically (3d). The logical interpretation, which is consistent with (3a), is pro-attitudinal for people in favour of death penalty but contra-attitudinal for people opposed to it. The reverse is true for the pragmatic interpretation. Therefore we predict, based on the theory of motivated reasoning, that people are inclined to interpret the scalar pragmatically when they are against the death penalty and logically when they are sympathetic to it. Note that these predictions apply only to the current example. If the word ‘decreased’ were to be replaced by ‘increased’, we would predict the opposite pattern. Furthermore, it is possible that one holds a neutral attitude towards the death penalty. This neutral condition serves as a baseline against which the effect of the pro-and contra-attitudinal context is evaluated. In general, given that the content of the utterance (i.e., 3a) is pro-attitudinal, we expect more logical interpretations while for contra-attitudinal utterances we expect more pragmatic interpretations compared with neutral utterances. In other words, we expect the endorsement rate of the scalar inference in the neutral condition to lie somewhere in between those of the pro-attitudinal and the contra-attitudinal condition.

## Methodology

### Participants

The study has been carried out in two different groups of participants. One group consisted of 73 12<sup>th</sup> grade students at the Groenendaalcollege in Merksem (34 men, 39 women, mean age 17,2), who participated voluntarily. The other participants were 197 first-year psychology students of the

University of Leuven (26 men, 171 women, mean age 18,5), who participated in return for course credit. In the analysis data from both groups are combined because results for each group were very similar.

## Materials and procedure

We used a self-constructed questionnaire to gauge participants' attitudes towards different issues. We asked them to indicate how they felt about e.g. the death penalty, more police on the streets, legalization of soft drugs,... on a scale from 1 to 5, 1 meaning *strong in favour* and 5 meaning *strong in disavour*. They also had to indicate on a similar scale how certain they felt about their answers, with 1 meaning *not sure at all* and 5 *very sure*. The item assessing one's attitude towards the death penalty was the crucial item of the questionnaire. The other items were merely fillers.

To measure whether or not the subjects derived the scalar inference, we took a similar approach as in Bonnefon et al. (2009). Half of the subjects received the following story:

*A research group of sociologists and criminologists did a large-scale study into the relationship between death penalty and crime. From this research has become clear that in some countries where the death penalty is applied, crime has decreased.*

*Do you think then that it is possible that in all countries where the death penalty is applied, crime has decreased?*

Subjects were then asked to circle the correct answer (yes or no). The other half got a slightly adapted story in which the word 'decreased' each time was replaced by 'increased'. Subjects received the questionnaire in small groups and were randomly assigned to one of these two versions. Since the experiment was conducted in Belgium, all materials were in Dutch.

## Results

Based on their responses to the attitude questionnaire, participants were divided into three groups: in favour of death penalty (4 or 5 on the scale), against the death penalty (1 or 2) and neutral towards death penalty (3). These groups were again divided according to attitude strength.

Table 1: Contingency table of content of the utterance against interpretation of the scalar (percentages are in parentheses).

Interpretation of the scalar	Content of the utterance			Total
	Pro-attitudinal	Neutral	Contra-attitudinal	
Logical	30 (34.5%)	7 (17.5%)	8 (8.2%)	45 (20%)
Pragmatic	57 (65.5%)	33 (82.5%)	90 (91.8%)	180 (80%)

Participants scoring 4 or 5 on this second scale were thought to be sure about their attitude whereas 3 or less indicates a rather uncertain opinion concerning death penalty. Participants were divided in these two categories because people with uncertain attitudes might attach less importance to the pro-or contra-attitudinal utterance. As a consequence, their interpretations of the scalar term might be less affected (or even uninfluenced) by the context. Therefore, participants who felt uncertain about their beliefs were excluded from the analysis<sup>1</sup>.

Crossing the type of story (i.e., crime has decreased versus crime has increased) with attitude towards death penalty yields six combinations which can be arranged in three groups or conditions: a pro-attitudinal condition (i.e., crime decreased and in favour of death penalty + crime increased and against death penalty), a contra-attitudinal condition (i.e., crime decreased and against death penalty + crime has increased and in favour of death penalty) and a neutral condition (i.e., crime decreased and neutral + crime increased and neutral). A chi-square test revealed a significant difference in interpretation of the scalar term between these three conditions ( $\chi^2(2, N=225) = 20.14, p < 0.001$ ). As predicted, subjects in the contra-attitudinal condition made more pragmatic interpretations than those who are neutral regarding the death penalty (Table 1). On the other hand, subjects in the pro-attitudinal condition made more logical interpretations compared with the neutral condition. Nevertheless, the tendency to derive the scalar inference seems to be so natural that still two-third of the subjects interpret 'some' as 'not all'. If we look at the data for pro and contra death penalty attitudes separately, we find similar results. For those opposed to the death penalty, the endorsement rate of the scalar inference is 66% in the pro-attitudinal condition and 93% in the contra-attitudinal condition ( $\chi^2(1, N=158) = 17.71, p < 0.001$ ). Subjects in favour of death penalty derived the scalar inference in 62% of the instances when they received the pro-attitudinal story and 86% when they received the contra-attitudinal story ( $\chi^2(1, N=27) = 2.05, p > 0.1$ ). Although this latter result did not reach significance (due to the small sample size), it is clearly in accordance with the other results.

## Discussion

The present research identifies a new context wherein the occurrence of the scalar inference either increases or

<sup>1</sup> Including all participants in the analysis had little effect on the results, probably due to the low number of uncertain attitudes (45 on a total 270). Still, we decided to preserve the distinction between certain and uncertain attitudes.



decreases. It is the first study to demonstrate the crucial role of attitudes in the interpretation of the scalar term 'some'. Depending on one's attitude, the scalar is more likely to be interpreted pragmatically, in contra-attitudinal utterances, or logically, in pro-attitudinal utterances, compared with a neutral condition.

These findings are in line with Kunda's theory of motivated reasoning (1990), in that people are motivated to reach a conclusion consistent with one's attitudes and beliefs. Furthermore, they complement previous studies demonstrating that context plays a crucial role in the interpretation of scalar terms.

The present study does not only differ in the *kind* of contextual manipulation, also the *position* of the scalar term relative to the context deviates from previous research. Formerly, the crucial context occurred either before or right after the scalar term. Here, the contextual manipulation (i.e., the word 'increased' or 'decreased') is situated nine words after the scalar 'some' and this has important theoretical implications<sup>2</sup>. Recent studies have shown that the computation of scalar inferences occurs relatively fast (Grodner et al., 2010; Huang & Snedeker, 2009). Also, Hartshorne and Snedeker (submitted) report that the interpretation of 'some' manifests itself nine words after encountering the scalar term. Given these findings, the contextual manipulation in the present study can initially have no effect on processing and interpretation of the scalar term. Before being confronted with the pro-or contra-attitudinal context, participants have already adopted either a logical or a pragmatic interpretation of the scalar. Thus from the perspective of Relevance theory, people will stop at the optimally relevant interpretation of 'some' without being influenced by the contextual manipulation. However, the ultimate interpretation of the scalar is affected by this context as evidenced by the varying number of logical interpretations in the different conditions. This suggests that readers (and listeners) do not necessarily stop at the first optimally relevant interpretation but rather keep on searching for a more relevant interpretation. In other words, people may temporarily hold a certain interpretation of 'some' but eventually move to a different one due to the context. Further research should be conducted in order to support this claim, especially because it is inconsistent with an assumption of Relevance theory (i.e., that the reader or listener stops at the first interpretation that satisfies his expectations of relevance).

A recent study of Bonnefon, De Neys and Feeney (2011) regarding face-threatening contexts, actually provides evidence for a "second push towards another equilibrium between effect and effort", as they call it. The authors found that logical interpretations took longer and were more difficult to reach in face-threatening contexts. A possible way to reconcile our findings and those of Bonnefon et al. (2011) with Relevance theory is to allow a reconsideration

of the interpretation of the scalar once the meaning of the sentence is fully grasped. This process would only be triggered in certain situations (i.e., pro-or contra-attitudinal context, face-threatening context) and requires processing time and effort. Thus, when people read the word 'increased' or 'decreased' they might revise their initial interpretation of the scalar term but this comes at a cost. Further research should help to determine whether the suggested modification of Relevance theory is valid.

In sum, this study adds to the existing literature in two ways. First, it provides evidence for the existence of a new context wherein the interpretation of the scalar term varies. Second, through the position of the contextual manipulation it challenges the assumption made by Relevance theory that people stop at the first optimally relevant interpretation of the scalar.

## Acknowledgments

We would like to thank Frank Mannaerts for his assistance in collecting the data at the Groenendaalcollege. Correspondence should be addressed to Tom Heyman, Department of Psychology, University of Leuven, Tiensestraat 102, 3000 Leuven, Belgium. E-mail: tom\_heyman123@hotmail.com.

## References

- Bonnefon, J. F., De Neys, W., & Feeney, A. (2011). Processing scalar inferences in face-threatening contexts. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Bonnefon, J.F., Feeney, A. & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition*, 112, 249-258.
- Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, 51, 437-457.
- Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100, 434-463.
- Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In A. Belletti (Ed.), *Structures and beyond: The cartography of syntactic structures* (Vol. 3, pp. 39-103). Oxford: Oxford University Press.
- Dawson, E., Gilovich, T., & Regan, D.T. (2002). Motivated reasoning and performance on the Wason Selection Task. *Personality and Social Psychology Bulletin*, 28, 1379-1387.
- De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, 54, 128-133.

<sup>2</sup> Because death penalty is a single word in Dutch, the contextual manipulation occurs nine words after the scalar (not ten words, as it would be in English).

- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics* (Vol. 3): *Speech acts* (pp. 41-58). New York: Academic Press.
- Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.
- Grodner, D., Klein, N., Carbary, K., & Tanenhaus, M. (2010). "Some," and possible all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, 116, 42-55.
- Hartshorne, J. K., & Snedeker, J. (submitted for publication). The speed of inference: Evidence against rapid use of context in calculation of scalar implicatures.
- Huang, Y. T., & Snedeker, J. (2009). On-line interpretation of scalar quantifiers: Insight into the semantic-pragmatics interface. *Cognitive Psychology*, 58, 376-415.
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108, 480-498.
- Levinson, S.C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
- Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.
- Morris, W. (1969). *The American heritage dictionary of the English language*. New York: American Heritage Publishing Co. Inc.
- Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, 85, 203-210.
- Noveck, I. A., & Reboul, A. (2008). Experimental pragmatics: A Gricean turn in the study of language. *Trends in Cognitive Sciences*, 11, 425-431.
- Sperber, D., & Wilson, D. (1986/1995). *Relevance: Communication and cognition*. Oxford: Blackwell.
- Wilson D, & Sperber D. (2003). Relevance theory. In G. Ward & L. Horn (eds.), *Handbook of Pragmatics* (pp. 607-632). Oxford: Blackwell.

# Do repeated references result in sign reduction?

Marieke Hoetjes (M.W.Hoetjes@uvt.nl)

Emiel Krahmer (E.J.Krahmer@uvt.nl)

Marc Swerts (M.G.J.Swerts@uvt.nl)

Tilburg centre for Cognition and Communication (TiCC), School of Humanities, Tilburg University,  
PO Box 90153, 5000 LE, Tilburg, the Netherlands.

## Abstract

Previous research has shown that repeated references are often reduced compared to initial references. The present study looks at the production of repeated references by signers of Sign Language of the Netherlands (NGT). Participants had to describe figures to an addressee, who had to pick the correct figure from a large group of figures. Several figures had to be described several times. The question was whether there would be reduction in the signed repeated references, as has been found previously for speech and gesture. We found systematic effects of repetition, in that repeated references are shorter, contain fewer signs, and shorter signs than initial references. Moreover, a perception experiment showed that signs produced during repeated references were also considered to be less precise than the signs produced during initial references.

**Keywords** sign language; repeated reference; reduction

## Introduction

Variability is ubiquitous in speech production, with words never pronounced the exact same way more than once. For example, someone might first pronounce the word 'of course' slowly and precisely, followed by an instance where it is pronounced quickly, less precise and more like 'fcourse'. This example of language variability shows that language can be reduced (in this case by shortening and merging words). While various studies have looked at reduction in speech, reduction in signs remains largely unexplored. The present study addresses this point.

## Reduction in spoken repeated references

In conversation, people often produce referring expressions to describe objects in the world around us. The production of repeated references occurs when people refer to the same object more than once in the conversation. Research has found that in speech, these repeated references are often reduced in at least two ways (Aylett & Turk, 2004; Bard, et al., 2000; Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; Fowler, 1988; Fowler & Housum, 1987; Galati & Brennan, 2010; Lam & Watson, 2010). Firstly, repeated references to the same target object usually contain fewer words than initial references (Clark & Wilkes-Gibbs, 1986; Galati & Brennan, 2010). Brennan and Clark (1996) claim that this is due to the fact that people establish so-called

conceptual pacts as more common ground is established over the course of the conversation (debated in e.g. Horton & Gerrig, 2005). Secondly, repeated references are often also reduced acoustically (Aylett & Turk, 2004; Bard, et al., 2000; Fowler, 1988; Fowler & Housum, 1987; Lam & Watson, 2010). Repeated references, when taken out of context and presented to a listener, have been found to be less recognisable for the addressee because their pronunciation is less clear in repeated references than in initial references (Bard, et al., 2000; Galati & Brennan, 2010). Lieberman (1963) found similar acoustic reduction for redundant words, which were shorter and perceived as less intelligible when taken out of context.

There are two dominant views on the reason why referring expressions may be reduced. On the one hand, reduction in referring expressions may be due to speaker oriented causes, such as production and planning processes (Arnold, 2008; Arnold, Kahn, & Pancani, 2012; Bard, et al., 2000; Bard & Aylett, 2005; Ferreira, 2008). On the other hand, reduction in referring expressions may be due to listener oriented causes, such as communicative strategies (e.g. Aylett & Turk, 2004; Fenk-Oczlon, 2001; Lieberman, 1963; Lindblom, 1990; Zipf, 1949). The use of communicative strategies, with speakers as efficient language users, has been shown by a range of studies (for an overview, see Jaeger & Tily, 2011), including Zipf's (1949) Principle of Least Effort, and Shannon's noisy channel model (1948). More recently, Lindblom (1990), in his H&H theory, claims that speakers adapt to the listener's needs, meaning that redundant speech is reduced as long as 'sufficient discriminability' remains. Jaeger (2010) proposed the hypothesis of Uniform Information Density (UID), which states that 'speakers prefer utterances that distribute information uniformly across the signal (information density)' (Jaeger, 2010:25). What this means is that the interaction between speaker and addressee is optimized by the speaker's lengthening or shortening of an utterance, such that the utterance becomes more uniform and optimal for both speaker and addressee.

It can be argued that the reduction in repeated references that previous studies have found is due to the abovementioned processes: when speakers produce repeated references, they fully reproduce those (auditory) aspects of the referring expression that contain important or new

information and are necessary for quick target identification. The less informative aspects of the referring expression may be reduced or omitted, leading to reduced references.

### **Reduction in visual repeated references: gesture and sign language**

Taking into account that communication does not only consist of 'spoken' aspects of speech, but can also contain or consist of visual aspects such as gestures (Kendon, 2004; McNeill, 1992) or signs (Stokoe, 2005), we may wonder whether a reduction process such as described above for spoken repeated references also occurs in the visual domain.

Relevant previous research on gesture has looked at the effect of common ground (Gerwing & Bavelas, 2004; Holler & Wilkin, 2009) and repeated references (de Ruiter, Bangerter, & Dings, 2012; Hoetjes, Koolen, Goudbeek, Krahmer, & Swerts, 2011) on gesture production, albeit with inconclusive results. For example, when we look at repeated references, on the one hand, de Ruiter et al. (2012), when testing their tradeoff hypothesis, found that repetition did not affect gesture rate. On the other hand, Hoetjes et al. (2011) found that both speech and gesture were reduced in repeated references.

There has been a range of research on phonological and phonetic aspects of sign language (Crasborn, 2001; Sandler, 1989; Sandler & Lillo-Martin, 2006; Schembri, et al., 2009; Tyrone & Mauk, 2010), starting with Stokoe (2005) in 1960 proposing that signs in sign languages consist of three main parameters (handshape, location and movement). However, hardly any studies have looked at sign language from the perspective of efficient language use. In this light, it is interesting to see how signs behave with regard to reduction in repeated references. We may wonder what the role of signs is compared to speech and to co-speech gesture. On the one hand, considering that signs, like words, usually convey lexical meaning, it might be the case that reduction in sign is similar to reduction in speech, for example with regard to the semantics that are expressed. On the other hand, signs, unlike words but like co-speech gestures, are a means of communication in the visual domain, and there may be aspects of reduction that are modality specific and thus alike between signs and co-speech gestures. Of course, it could also be the case that signs are not reduced in a way comparable to speech or to co-speech gestures, but that signs, if they are reduced, are reduced in a sign-specific manner.

The only experimental study on sign language we are aware of that can be related to the idea of efficiency of language users in the production of repeated references is the work by Tyrone and Mauk (2010) on phonetic reduction in American Sign Language. In their study, Tyrone and Mauk looked at the production of the sign WONDER in two phonetic contexts and at three signing rates. Their results show that sign lowering occurs with increasing signing rate and can, but not necessarily does, occur in specific phonetic contexts. Another study on variation in sign language, by Schembri and colleagues (2009), looked at naturalistic data

and also found that sign location can vary with signs produced at lower locations than their citation form. However, neither of these studies takes repetition into account as one of the factors influencing sign production.

In the present study we will look at signs of Sign Language of the Netherlands (NGT), to see whether reduction in repeated references, as previously found for speech and gesture, also occurs in sign language. Considering that NGT is a fully fledged sign language and presumably behaves in many respects as a spoken language, we hypothesize that, as in speech, reduction in repeated references will occur. The question is of course how reduction in signs can be measured. In the present study we have decided to measure reduction by combining methods that have been used previously in studies on speech and on gesture. We will look at sign characteristics that we consider comparable with some of the aspects of speech that have been studied previously when looking at reduction, namely number of words, utterance duration and word duration. We will also take precision into account, which has been done in previous studies on gesture. Therefore, in the present study on sign language we will look at the number of signs, utterance and sign duration and at sign precision. We conducted a production task to analyse the first three attributes. Following Hoetjes et al. (2011), we conducted a perception task to analyse the last attribute, sign precision.

### **Production experiment**

To study reduction in repeated references in Sign Language of the Netherlands (NGT), a data set was created consisting of recordings of participants taking part in a director-matcher task. In this task, the director had to describe an object in such a way that the matcher could identify the object from a range of similar looking figures. In the stimuli, there were several figures that had to be described more than once, leading to repeated references to the same item.

### **Participants**

The director-matcher task was done by a total of 14 signers of NGT. The group of participants consisted of 5 male and 9 female speakers, with an average age of 46 years old (range 26-60 years old). The average length of time that the participants had been signing NGT was 23.5 years (range 2-50 years). Participants would take part twice in the experiment; first they were randomly assigned the role of either director or matcher and they would switch roles after doing the experiment once.

### **Stimuli**

Two picture grids, each containing 16 pictures, were used by each director. Each picture grid showed either pictures of people, or pictures of furniture items. The two different domains (people and furniture) were used since previous studies on referring expressions had shown them to be efficient domains for making people produce referring expressions (Koolen, Gatt, Goudbeek, & Krahmer, 2011;

Van Deemter, Gatt, van der Sluis, & Power, in press; Van der Sluis & Krahmer, 2007).

Each picture grid was used for 15 trials, adding up to a total of 30 trials. For the first 15 trials, a people picture grid was used, for the last 15 trials a furniture picture grid was used. Since the participants would do the experiment twice, once in the role of director and once in the role of matcher, two sets of picture grids were used, with different pictures on each picture grid, making sure that the same picture never had to be described across roles. In each trial, there was one target object (marked by a red square around the object), surrounded by 15 distractor objects, which had to be described by the director. The crucial manipulation in the task was that several pictures had to be described repeatedly: in each of the picture grids there were two pictures that had to be described three times. Repeated references to the same object were never one straight after the other. This means that descriptions of other objects were given in between the initial and repeated descriptions of the critical objects. An example of a trial with object description can be seen below in figure 1.



“CHAIR, RED, NOT LEFT, SIDEWAYS TO THE RIGHT, LITTLE BIT BIGGER.”

Figure 1. Picture grid showing a trial, followed by gloss of example initial description of the target object.

## Procedure

The director and the matcher were seated at a table opposite each other. A camera was positioned behind the matcher filming the upper body and hands of the director. The director had a laptop screen to her side and the matcher had a picture card in front of her. The director and matcher could see each other directly, but could not see each other's screen or card. The director was presented with a trial on the computer screen and was asked to provide a description of the target object in such a way that the matcher could distinguish it from the 15 distractor objects. The matcher had a picture card filled with the same 16 objects in front of her, which was not visible to the director. The matcher's card showed the same objects as on the director's screen, but these objects were ordered differently for the director and the matcher. This means that the director could not use

the location of the target object on the grid as part of the description. This was explicitly communicated to the directors. Once the correct object was found, the director went on to the next trial. The entire task took the participants about 20 minutes. After conducting 15 trials from the people domain and 15 trials from the furniture domain, the director and matcher would switch roles to conduct the experiment again, using the other set of picture grids.

## Data analysis

For the purpose of the current analyses, the first and third (hence initial and repeated) descriptions of the four objects that had to be described three times were annotated and analysed. These four objects were never described in the first or last trial. The focus on the initial and repeated descriptions means that the current analyses are based on a data set which consists of eight descriptions (two initial and two repeated descriptions for each of the two picture grids) for each of the 14 participants, leading to a total of 112 object descriptions. We used the multimodal annotation programme ELAN (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006) to annotate the signs. We looked at the duration of the complete descriptions, the number of lexical signs that were produced in the descriptions and the duration of the signs. A separate perception experiment was used to measure sign precision, which will be discussed below under Perception experiment.

The experiment consisted of a 2 x 2 x 2 design with factors Domain (levels: people, furniture), Repetition (levels: initial, repeated), and Picture (levels: one, two). We tested for significance using repeated measures ANOVAs by participants ( $F_1$ ) and by items ( $F_2$ ).

## Results

Firstly, it was found that speakers take significantly less time (in seconds) to describe repeated references ( $M = 14.46$ ,  $SD = 1.46$ ) compared to initial references ( $M = 24.24$ ,  $SD = 2.25$ ),  $F_1(1, 13) = 35.15$ ;  $p < .001$ ,  $F_2(1, 4) = 22.30$ ,  $p < .01$ . For the mean number of signs it was found that speakers produce significantly fewer signs in repeated references ( $M = 5.57$ ,  $SD = .32$ ) compared to initial references ( $M = 8.16$ ,  $SD = .56$ ),  $F_1(1, 13) = 42.5$ ;  $p < .001$ ,  $F_2(1, 4) = 16.59$ ,  $p < .05$ . Moreover, the average duration (in seconds) of signs is shorter in repeated references ( $M = 1.2$ ,  $SD = .054$ ) than in initial references ( $M = 1.47$ ,  $SD = .074$ ),  $F_1(1, 13) = 15.1$ ;  $p < .01$ ,  $F_2(1, 4) = 20.17$ ,  $p < .05$ . In sum: we find systematic effects of repetition, in that repeated references are shorter, contain fewer signs, and shorter signs than initial references. These effects were the same for both domains (furniture and people) and for all pictures; in particular, we found no significant interaction between the factors repetition and domain or repetition and picture. To illustrate, figures 2 and 3 below show a case of reduction in the description of a target object from the furniture domain. In the initial description, the participant takes longer and uses more signs and more precise signs (to

be discussed in the perception experiment below) than in the repeated description.



“SOFA, THREE SEATS, ASKEW, BIG, TO THE RIGHT, TO THE SIDE”

Figure 2. Still and gloss of initial description of a sofa, lasting 48 seconds. Sign depicted in still is SOFA, with a fairly large extension and well defined edges (see arrows).



“SOFA, GREEN, TURNED AROUND, THREE SEATS”

Figure 3. Still and gloss of repeated description of the same sofa as in figure 2, lasting 17 seconds. Sign depicted in still is SOFA, with smaller extension than in figure 2 and without well defined edges (see arrows).

### Conclusion production experiment

The results show that several aspects of NGT were reduced in repeated references. Repeated references produced by signers of NGT were shorter than initial references, and repeated references in NGT contained fewer and shorter signs than initial references. This means that, at least for the aspects taken into account here, repeated references in NGT behaved as previous studies found for repeated references in speech. Repeated references by signers of NGT, containing predictable information, were produced in a more efficient way than initial references.

### Perception experiment

Since it is difficult to define objective measures with which to measure sign precision, a perception experiment was set up in which participants had to judge, in a forced choice

task, which sign they considered to be the most precise, looking at pairs of video clips with signs produced in either initial or repeated references.

### Participants

Twenty-seven first year university students, who had no knowledge of NGT, took part as partial fulfillment of course credits. Non-NGT speaking participants were used on purpose, so that the participants would not know the lexical meaning of the signs but would only judge the signs on their perceived precision.

### Stimuli

The participants were presented with a PowerPoint presentation in which they saw 40 pairs of video clips. Each pair of video clips was presented on one slide. Both video clips showed the same sign, produced by the same signer of NGT, about the same object, as described in the director-matcher task, except in one video clip the sign was produced in an initial reference and in the other video clip the sign was produced in a repeated reference. The order in which the participants were presented with initial versus repeated signs in the video clip pairs was counterbalanced over pairs of video clips (so it was not the case that for each pair the first video clip they saw was always the sign produced in an initial reference).

### Procedure

The participants had to watch the pairs of video clips, one video clip at a time, and were allowed to watch a video clip more than once if they wanted to. The task was to choose for each pair of video clips which sign they considered to be the most precise (the sign in video clip A or B). The task was a self-paced forced choice task and even though the participants were allowed to watch the video clips more than once, they were encouraged to go with their first intuition. The only instruction they were given was to choose which sign they considered to be the “most precise”. No details were given to suggest what the participants should base this judgment on.

### Data analysis

For each pair of video clips, each sign that was considered to be the most precise received a point from each participant. Statistical analyses consisted of repeated measures ANOVAs over proportions, by participants ( $F_1$ ) and by items ( $F_2$ ).

### Results

In line with our earlier results, we find that signs produced during repeated references ( $M = .33$ ,  $SD = .04$ ) were considered to be less precise than the signs produced during initial references ( $M = .67$ ,  $SD = .04$ ),  $F_1(1, 26) = 121.29$ ,  $p < .001$ ,  $F_2(1, 78) = 41.21$ ,  $p < .001$ . The effect was the same for both domains (furniture and people).

## Conclusion perception experiment

The results show that signs produced in repeated references were considered to be less precise than signs produced in initial references. Therefore, it can be concluded that there was also reduction in repeated references when it comes to sign precision.

## Discussion and conclusion

Summarizing the results from the production and perception experiments, we found reduction in repeated references in sign language. We found that repeated references were shorter, contained fewer and shorter signs, and that signs produced in repeated references were considered to be less precise than signs in initial references.

The present results on sign language can be tied in with previous findings, both on speech and on gesture, that language users tend to be efficient by reducing predictable information (e.g. Jaeger, 2010). Relating the results to previous work on speech, we showed that repeated references were shorter and contained fewer signs than initial references, in line with work by Clark and Wilkes-Gibbs (1986) and Galati and Brennan (2010). The result that signs in repeated references were shorter can be related to previous work on speech by Aylett and Turk (2004) and by Lam and Watson (2010) where it was found that predictable speech (through redundancy or repetition) had a shorter duration than unpredictable speech. Our finding that signs in repeated references were considered to be less precise can be viewed to be an extension of the work by Bard et al. (2000), who found that repeated references had a less clear pronunciation than initial references.

When we compare the results from the present study with previous work on co-speech gestures, we can also see clear links. It has been found that gestures with common ground are less precise (Gerwing & Bavelas, 2004) and contain less semantic information (Holler & Wilkin, 2009) than gestures without common ground. This can be related to our findings that signs in repeated references were considered to be less precise and that repeated references in NGT contained fewer signs than initial references. Work on the effect of repeated references on gestures (Hoetjes, et al., 2011) found that repeated references may cause reduction in the number of gestures, as was found in the present study for the number of signs. Moreover, their finding that gestures in repeated references were considered to be less precise than gestures in initial references, can be directly mapped onto the present results for signs. Importantly, the reduction found in the current study can be tied in with work on language efficiency and cannot be explained through a general reduction of descriptions over time (with participants becoming more 'sloppy' in the course of the experiment). In short, the present study is the first study on sign language that shows that signers of NGT behave similarly when describing repeated references as to what previous studies have found for speech and gesture by speakers of spoken languages.

Due to the fact that hardly any previous work has been done on reduction in sign language, the method used in the current study was inspired by relevant previous work on speech and gesture. We looked at fairly rough and modality independent (i.e. applicable to speech, gesture and sign) measures such as duration of the description and number of signs and not at more sign-specific aspects such as exact sign location (as has been done by e.g. Tyrone & Mauk, 2010). Despite the fact that our measures were not based on sign characteristics per se, we were still able to find that reduction in sign language occurred. This shows that it is possible to use such modality independent methods to study reduction in repeated references.

Naturally, the current study leaves room for some discussion. In the perception experiment, we used participants with no knowledge of NGT to judge the precision of signs produced in the production experiment. This was done purposefully, so that the participants were not in any way influenced by the lexical meaning of the signs and could focus only on the precision judgment task. There are reasons to assume that the use of non-NGT signers is indeed a reasonable approach. Research has shown (Brentari, Gonzalez, Seidl, & Wilbur, 2011) that non-signers have a high degree of sensitivity to visual prosodic cues of a sign language. However, future work could include NGT signing participants in the perception experiment. Also, if using NGT signing participants in future work, another possibility would be to set up the task slightly differently by asking participants to judge a sign's intelligibility, as in Bard et al.'s (2000) work on speech, instead of judging its precision.

In sum, the analyses done presently are the first of its kind to show us not only that we can use analyses from related work on speech and gesture and adapt them to analyse signs in repeated references, but also that signers of NGT reduced their repeated references. In fact, the ways in which these repeated references were reduced in NGT are quite similar to what has been found previously for speech and gesture. It is well known that speakers of non-signed languages are communicatively efficient by reducing repeated information, both in speech and in co-speech gestures. This study has shown, for the first time, that signers can design their utterances to be efficient in the same ways.

## Acknowledgments

We would like to thank Axelle Schmit and Manon Yassa for help in collecting and annotating the data, Onno Crasborn and Karin van Nispen for helpful discussions and comments and Martijn Goudbeek for help with the statistics. We received financial support from The Netherlands Organization for Scientific Research, via a Vici grant (NWO grant 27770007), which is gratefully acknowledged.



## References

- Arnold, J. E. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes*, 23(4), 495-527.
- Arnold, J. E., Kahn, J., & Pancani, G. (2012). Audience design affects acoustic reduction via production facilitation. *Psychonomic Bulletin & Review*.
- Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1), 31-56.
- Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, 42, 1-22.
- Bard, E. G., & Aylett, M. (2005). Referential form, duration, and modelling the listener in spoken dialogue. In J. Trueswell & M. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 173-191). Cambridge: MIT Press.
- Brennan, S., & Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology*, 22(6), 1482-1493.
- Brentari, D., Gonzalez, C., Seidl, A., & Wilbur, R. B. (2011). Sensitivity to visual prosodic cues in signers and nonsigners. *Language and Speech*, 54(1), 49-72.
- Clark, H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, 1-39.
- Crasborn, O. (2001). *Phonetic implementation of phonological categories in Sign Language of the Netherlands*. PhD dissertation. Utrecht: LOT.
- de Ruiter, J. P., Bangertner, A., & Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the trade-off hypothesis. *Topics in Cognitive Science*, 4(2), 232-248.
- Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 431-448). Amsterdam/Philadelphia: John Benjamins.
- Ferreira, V. S. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Learning and Motivation*, 49, 209-246.
- Fowler, C. A. (1988). Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech*, 31(4), 307-319.
- Fowler, C. A., & Housum, J. (1987). Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26(5), 489-504.
- Galati, A., & Brennan, S. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, 62, 35-51.
- Gerwing, J., & Bavelas, J. (2004). Linguistic influences on gesture's form. *Gesture*, 4, 157-195.
- Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (2011). GREEBLES Greeble greeb. On reduction in speech and gesture in repeated references. In L. Carlson, C. Hoelscher & T. F. Shipley (Eds.), *33rd Annual Conference of the Cognitive Science Society* (pp. 3250-3255). Boston: Cognitive Science Society.
- Holler, J., & Wilkin, K. (2009). Communicating common ground: how mutually shared knowledge influences speech and gesture in a narrative task. *Language and Cognitive Processes*, 24(2), 267-289.
- Horton, W. S., & Gerrig, R. J. (2005). Conversational common ground and memory processes in language production. *Discourse Processes*, 40, 1-35.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1), 23-62.
- Jaeger, T. F., & Tily, H. (2011). Language Processing Complexity and Communicative Efficiency. *WIREs: Cognitive Science*, 2(3), 323-335.
- Kendon, A. (2004). *Gesture. Visible action as utterance*. Cambridge: Cambridge University Press.
- Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43(13), 3231-3250.
- Lam, T. Q., & Watson, D. G. (2010). Repetition is easy: Why repeated referents have reduced prominence. *Memory and Cognition*, 38(8), 1137-1146.
- Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6(3), 172-187.
- Lindblom, B. (1990). Explaining variation: a sketch of the H and H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403-439). Dordrecht: Kluwer Academic Publishers.
- McNeill, D. (1992). *Hand and mind. What gestures reveal about thought*. Chicago: University of Chicago Press.
- Sandler, W. (1989). *Phonological Representation of the Sign: Linearity and Nonlinearity in American Sign Language*. Dordrecht: Foris.
- Sandler, W., & Lillo-Martin, D. (2006). *Sign Language and Linguistic Universals*. Cambridge: Cambridge University Press.
- Schembri, A., McKee, D., McKee, R., Pivac, S., Johnston, T., & Goswell, D. (2009). Phonological variation and change in Australian and New Zealand Sign languages: The location variable. *Language variation and change*, 21, 193-231.
- Shannon, C. (1948). A mathematical theory of communications. *Bell systems technical journal*, 27(4), 623-656.
- Stokoe, W. C. (2005). Sign language structure: An outline of the visual communication systems of the American Deaf. *Journal of Deaf Studies and Deaf Education* 10(1), 3-37.
- Tyrone, M. E., & Mauk, C. E. (2010). Sign lowering and phonetic reduction in American Sign Language. *Journal of Phonetics*, 38, 317-328.
- Van Deemter, K., Gatt, A., van der Sluis, I., & Power, R. (in press). Generation of referring expressions: Assessing the Incremental Algorithm. *Cognitive Science*.
- Van der Sluis, I., & Krahmer, E. (2007). Generating Multimodal Referring Expressions. *Discourse Processes*, 44(3), 145-174.
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). *ELAN: a Professional Framework for Multimodality Research*. Paper presented at the LREC 2006, Fifth International Conference on Language Resources and Evaluation.
- Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley.

# When gestures catch the eye: The influence of gaze direction on co-speech gesture comprehension in triadic communication

Judith Holler (judith.holler@mpi.nl)<sup>1,2</sup>

Spencer Kelly (skelly@colgate.edu)<sup>3</sup>

Peter Hagoort (peter.hagoort@mpi.nl)<sup>1,4</sup>

Asli Ozyurek (asli.ozyurek@mpi.nl)<sup>1,5</sup>

<sup>1</sup> Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD Nijmegen, The Netherlands

<sup>2</sup> University of Manchester, School of Psychological Sciences, Coupland Building 1, M13 9PL Manchester, UK

<sup>3</sup> Colgate University, Psychology Department, Center for Language and Brain, Oak Drive 13, Hamilton, NY 13346, USA

<sup>4</sup> Donders Institute for Brain, Cognition and Behaviour, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

<sup>5</sup> Centre for Language Studies, Radboud University, Erasmusplein 1, 6525HT Nijmegen, The Netherlands

## Abstract

Co-speech gestures are an integral part of human face-to-face communication, but little is known about how pragmatic factors influence our comprehension of those gestures. The present study investigates how different types of recipients process iconic gestures in a triadic communicative situation. Participants (N = 32) took on the role of one of two recipients in a triad and were presented with 160 video clips of an actor speaking, or speaking and gesturing. Crucially, the actor's eye gaze was manipulated in that she alternated her gaze between the two recipients. Participants thus perceived some messages in the role of addressed recipient and some in the role of unaddressed recipient. In these roles, participants were asked to make judgements concerning the speaker's messages. Their reaction times showed that unaddressed recipients did comprehend speaker's gestures differently to addressees. The findings are discussed with respect to automatic and controlled processes involved in gesture comprehension.

**Keywords:** co-speech iconic gesture; eye gaze; recipient status; communicative intent; multi-party communication.

## Introduction

When we speak, we frequently move our bodies to supplement what we say with co-speech gestures. A large proportion of these gestures are iconic in nature. Importantly, iconic gestures bear a close link with the speech that they accompany on semantic and temporal levels and have therefore been argued to constitute an integral part of human language (McNeill, 1992; Kendon, 2004) and thus of speaker's utterances (i.e., 'composite utterances', Kendon, 2004). While iconic gestures have been shown to fulfill a variety of cognitive functions which appear to benefit the speaker him or herself (e.g., Chawla & Krauss, 1994; Hostetter, Alibali & Kita, 2007), there is a growing body of evidence that their production is also linked to the speaker's communicative intent (Gerwing & Bavelas, 2004; Holler & Stevens, 2007; Kelly, Byrne & Holler, 2011; Özyürek, 2002).

The comprehension of iconic gestures, especially with respect to the attribution of communicative intentions, has been considerably less well researched. What we do know is that iconic gestures successfully communicate semantic information and that recipients integrate this information

with that contained in the accompanying speech (e.g., Holle & Gunter, 2007; Holler, Shovelton & Beattie, 2009; Kelly, Barr, Church & Lynch, 1999; Kelly, Kravitz & Hopkins, 2004; Willems, Özyürek & Hagoort, 2007). However, one limitation of studies on the comprehension of gestures is that many of them have presented stimuli in isolation, that is, video clips showing iconic gestures (and sometimes a torso) but, crucially, no head or facial information, and those studies that have included the face have tended to focus on the lips. In face-to-face communication, however, gestures are not only accompanied by speech and mouth movements, but also by a multitude of additional nonverbal social cues. Instead of focusing our attention solely on speech and gesture when listening to someone speaking we are required to divide our cognitive resources in such a way that allows us to take in and combine all of those cues. How we process and comprehend iconic gestures in more situated contexts that are much closer to real life situations therefore remains a wide-open issue.

Of particular interest in this respect is the influence of eye gaze, one of the most powerful nonverbal social cues (Pelphrey & Perlman, 2009; Senju & Johnson, 2009). Eye gaze is not only an omnipresent contextual cue when observing co-speech gestures, it is also inherently linked to the perception of communicative intent (Kampe, Frith & Frith, 2003; Schilbach et al., 2006) and the regulation of social interaction (Argyle & Cook, 1976; Goodwin 1981; Kendon, 1967). This begs the question of how the co-occurrence of gaze and gesture influences recipients' comprehension. The present study addresses this very question.

To do so, it builds on a couple of recent studies that have begun to focus on the issue of perceived communicative intent in conjunction with gesture comprehension. Kelly et al. (2007, 2010) showed that participants integrated co-occurring information from speech and gesture less strongly when the two modalities were perceived as not intentionally coupled (e.g., male hands gesturing accompanied by a female voice speaking) than when they were perceived as intended to form a composite utterance (e.g., male hands gesturing accompanied by a male voice speaking). This is the first empirical evidence that the perceived intentional

stance of a communicator influences recipients' processing of iconic gesture and speech.

However, as many previous studies in this field, these two studies did not present gestures in their natural context but, instead, in isolation of any facial cues (including eye gaze) with the aim to control for the influence of lip movements. In addition, and in line with the predominant gesture comprehension paradigm at the time, both of the studies used mismatching speech-gesture stimuli, that is, stimuli in which the information provided by speech conflicted with that depicted by the accompanying iconic gestures. Whilst this was an ideal test bed for first enquiries into the semantic integration of speech and gesture, it compromises the generalisability of such findings to more natural speech-gesture utterances, something the present study aims to overcome.

Another recent study that has addressed the topic of communicative intent and gesture comprehension was conducted by Straube et al. (2010). In their study, participants watched video clips of a speaker who looked directly at them or who was oriented away from the camera. In contrast to previous studies on co-speech gesture comprehension, Straube et al.'s paradigm did include the speaker's head and eye gaze, and, furthermore, they avoided the use of mismatching gestures. However, the authors manipulated multiple nonverbal social cues simultaneously (body/torso orientation, gesture orientation, as well as gaze direction), preventing us to draw conclusions about the effect of gaze direction specifically on participants' comprehension. In addition, the information depicted by the gestures used as stimuli was redundant with that in speech (e.g., the speaker referred to a 'round bowl' in speech accompanied by a gesture depicting a round, bowl-like shape). It is therefore not possible to identify whether the differences in participants' comprehension (measured in the form of their neural response and memory performance for the stimuli) between the two conditions (frontal/averted) was due to differences in their perception of speech, gesture, or a combination of the two.

The studies by Kelly et al. (2007, 2010) and Straube et al. (2011) are laudable first attempts tapping the issue of communicative intent and gesture comprehension and useful stepping stones for further investigations on this topic. The present study aims to build on this work by investigating the effect of perceived communicative intent, as signaled through the speaker's eye gaze direction, on the comprehension of iconic gestures. Importantly, this study will be presenting gestures in a more natural context (including the head), manipulating social eye gaze as the only social cue of interest, and it will be based on gestures that match the speech but which are complementary in nature. This, in conjunction with the particular experimental paradigm employed in this study, will allow us to tap into the processing of gesture directly, and to zoom into recipients' processing of the verbal and the gestural components of the speaker's messages separately. We will do so by creating a set-up simulating a triadic

communicative situation involving one speaker and two recipients, combined with a manipulation of the speaker's eye gaze direction which will indicate to participants when they are an addressed and when they are an unaddressed recipient. Thus, it is the first experimental study looking specifically on the effect of social eye gaze on speech and iconic co-speech gestures.

## Method

### Participants

Thirty-two female, right-handed German native speakers participated in the experiment (mean age = 21.6yrs) and were compensated with 8€ payment.

### Design

The study employed an experimental paradigm simulating multi-party communication involving one speaker-gesturer and two recipients, only one of which was a 'real' participant (the other one was fictive). Participants were made to believe that the other participant taking on the role of the second recipient was located in a different room.

A confederate acted as the speaker-gesturer and produced scripted utterances to a video camera (we used pre-recorded instead of live stimuli to ensure that all participants were presented with identical stimuli). These pre-recorded video clips were presented to participants while making them believe that they were engaging in a live communication with the speaker, who they thought was located in yet a different room to themselves but connected to them via a live camera link. (The second recipient was, allegedly, also connected to the person acting as speaker via a live-camera link, in the same way as they were.)

To prevent participants from realising that the speaker was pre-recorded, they were told that the camera link was a one-way connection in that the two recipients were able to hear and see the speaker but that the speaker was not able to hear or see them (and the two recipients were, of course, not able to see or hear each other). They were also told that the speaker had been asked to stand in front of two different cameras, that she knew of the two recipients' presence, and that she had been told that each camera was hooked up to one of the recipients' computer monitors, allowing them to hear and see what she was communicating. In order to create a more plausible situation and convince subjects that the speaker did not memorise the entire set of scripted sentences by heart, the video clips showed the speaker looking down before each sentence was spoken; participants were told that a laptop had been positioned on a table in front of the speaker displaying a black and white drawing accompanied by a couple of words before each trial, and that the speaker had been instructed to communicate the contents of the information displayed on the screen spontaneously and in a way that felt natural to them (no explicit mention of gesture was made). They (the actual participants) were then informed that the speaker would sometimes address them by looking into the camera linked

to their own on monitor, and sometimes the other, second recipient, by looking into the respective other camera. This created two different views for the actual participant, one in which they were directly gazed at, and one in which they observed the speaker's gaze being averted (see Fig 1). Gaze direction, as implemented through this manipulation, constituted our main IV (within-participants). In addition, as a second IV (modality), we manipulated the occurrence of gesture in association with the sentences spoken by the speaker in the video clips (within-participants).

## Stimuli

The experiment set-up described in the preceding section required the creation of four types of video stimuli: a) Direct gaze (speech only), b) Averted gaze (speech only), c) Direct gaze (speech + gesture), d) Averted gaze (speech + gesture) (Fig. 1). In each video vignette, the actor spoke a short sentence (canonical SVO structure), e.g. 'she goes through the list' ('*sie geht durch die Liste*'). Crucially, the verb included in the sentence was always manner *unspecific*, i.e., 'to go through' (*durchgehen*). The iconic gestures accompanying these verbs always specified the manner of action, e.g., to tick items on the list (*abhaken*). This manipulation allowed us to measure participants' comprehension of the gestures independently of speech, without using mismatching gestures (see Introduction). Participants watched the videos on a computer screen in a soundproof experimental test booth; the audio signal was presented via closed-back headphones. Materials were presented with Presentation® software (www.neurobs.com).

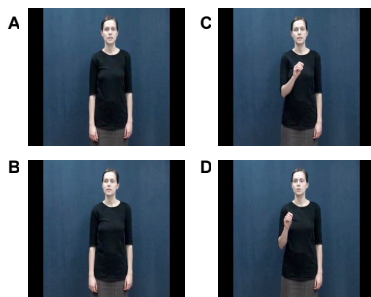


Figure 1: Examples of the four types of video stimuli used, A = Direct gaze (S only), B = Averted gaze (S only), C = Direct gaze (S + G), D = Averted gaze (S + G).

## Procedure

First, participants completed six practice trials (showing a different actor). Before the start of the experiment proper, the experimenter made a fake phone call to check whether 'the other participants' were ready to start.

Participants then watched 160 videos (40 stimuli per condition). Each video clip was followed by a written word (presented in capital letters, centre of screen) that matched either the verb contained in the preceding spoken sentence (*speech-related targets* [20 items per condition]; designed to tap into the processing of the verbal component of the

speaker's message) or the content of the gesture performed by the speaker in the video (*gesture-related targets* [20 items per condition]; designed to tap into the processing of the gestural component of the speaker's message).

## Task

Participants were asked to judge "whether the word displayed on the screen had been mentioned by the speaker in the preceding video", thus requiring 'yes' answers for all speech-related targets, and 'no' answers for all gesture-related targets. Reaction times (RTs) to participants' yes/no answers (delivered via a button box; yes = dominant hand) as well as errors<sup>1</sup> were recorded.

## Results

RTs for gesture and speech-related targets were entered into two separate 2 (gaze: direct vs. averted) x 2 (modality: speech only vs. speech+gesture) repeated measures ANOVA, excluding errors (constituting 2% of the total number of trials) and outliers (2 SD).

### Speech-related targets

Our first comparison concerned participants' responses to the speech-related targets (e.g., 'to go through' (*durchgehen*)) designed to tap primarily into the processing of the verbal component of the speaker's composite utterances. This analysis revealed a significant main effect of modality ( $p = .0001$ ), with slower response times in the speech + gesture conditions than in the speech only conditions. The main effect of gaze was not significant ( $p = .090$ ), and neither was the interaction between gaze and modality ( $p = .870$ ).

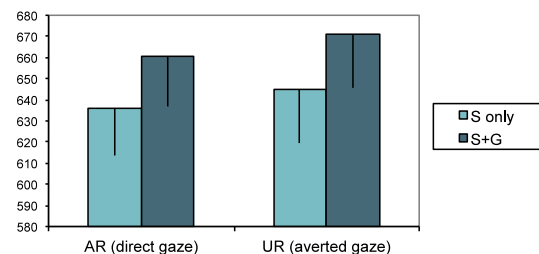


Figure 2: Addressed and unaddressed recipients' (AR/UR) RTs (ms) in the speech-only and speech + gesture conditions for speech-related targets (error bars = SE).

### Gesture-related targets

Our main comparison focused on participants' responses to the gesture-related targets (e.g., 'to tick' (*abhaken*)) intended to tap primarily into the processing of the gestural

<sup>1</sup> Due to restrictions on space, we only report our RT results here. Note, however, that the error rate analysis revealed very few significant differences, and those that did emerge did not relate to the relevant differences in RTs in a meaningful way.

component of the speaker's composite utterances. This analysis revealed no main effect of modality ( $p = .216$ ) and no main effect of gaze ( $p = .087$ ). However, the interaction between gaze and modality was significant ( $p = .045$ ). Independently from the omnibus interaction effect, we carried out two a priori contrasts (UR S+G vs. AR S+G; UR S-only vs. AR S-only). These showed that the interaction is driven by unaddressed recipients taking significantly longer to respond in the S+G condition than addressed recipients in the S+G condition ( $p = .026$ ). The comparison of unaddressed and addressed recipients' RTs in the speech-only conditions was not significant.

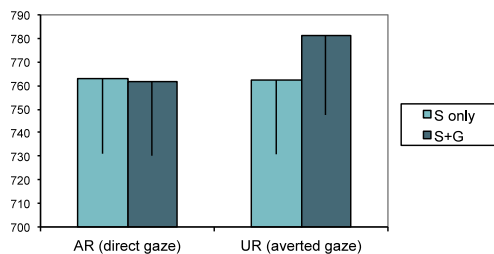


Figure 3: Addressed and unaddressed recipients' (AR/UR) RTs (ms) in the speech-only and speech + gesture conditions for gesture-related targets (error bars = *SE*).

## Discussion

The present study has investigated co-speech gesture comprehension in the presence of eye gaze, a powerful social cue integral to human face-to-face communication. Our findings reveal that recipients' gesture comprehension is indeed influenced by the speaker's eye gaze direction. More specifically, we have shown that, when a speaker's eye gaze is used to signal communicative intent in the sense of address, recipients who are currently unaddressed (but ratified participants in a communication, Goffmann, 1981) do process speech-accompanying iconic gestures differently to addressed recipients. This finding advances our understanding of human communication by pointing to an important way in which pragmatic processes shape and influence the comprehension of co-speech gestures, which, to date, has been addressed by only a very small number of studies (Kelly et al., 1999, 2007, 2010; Straube et al., 2011).

There are at least two competing interpretations of our effects of eye gaze direction on gesture comprehension. One is what we have termed the 'Fuzzy Representation Hypothesis'. According to this hypothesis, due to unaddressed recipients perceiving the speaker's gestures as not intended for them (but for the other, gazed at recipient instead), they interpret the gesture to a lesser degree. As a consequence, they take longer to respond to the gesture-related targets because they have constructed merely a partial, or fuzzy, mental representation of the gestural meaning. They are aware that something relating to the meaning of the word displayed on the screen may have been presented to them, but they have a hard time making a quick

decision on the modality in which this information was presented (since the gestural component of their mental representation is 'incomplete' or 'fuzzy'). Further, we would argue that the underlying mechanism leading to this fuzzy representation would not simply be one of reduced attention. The reason for this is that, first, addressed and unaddressed recipients do not differ in the number of errors they made, and, second, the modality effect we found for the speech-related targets is as strong for unaddressed as for addressed recipients. Thus, our data suggest that unaddressed recipients do process the gestural information but less strongly so, and that the reason for this is a modulation of perceived communicative intent rather than a pure decrease in attention.

An alternative possibility is what we have called the 'Competing Modalities Hypothesis'. The rationale underlying this account is that, while addressed recipients (in both dyadic and multi-party interactions) are expected to engage in mutual gaze with a speaker (Argyle & Dean, 1976; Kendon, 1967), unaddressed recipients are free to disengage from the process of gazing at the speaker. This means that the default situation requires recipients who are directly addressed through a speaker's gaze to split their attention between information coming from multiple modalities including speech, gesture, and gaze (and additional facial cues). Unaddressed recipients, on the other hand, have fewer visual social cues to process since the speaker's gaze is averted from them. They may therefore zoom in on gesture, instead of processing gesture and gaze simultaneously, and may thus have more cognitive resources available to focus on the processing of gesture. As a consequence, in the current paradigm, unaddressed recipients are taking longer than addressed recipients to respond to the gesture-related targets (requiring a 'no' answer) because the gestural component of the speaker's utterance constitutes a more prominent component of their mental representation of the event described (since they focused on gesture more). To declare something as not having been mentioned by the speaker despite the stronger memory trace of the gesture being at the forefront of their mind appears to be a difficult task.

The present study was a fruitful undertaking as it has shown that recipient status can influence gesture processing, but also because it allowed us to formulate two possible accounts of a potential process model explaining those effects. Currently, we are unable to unequivocally declare one of them as the more appropriate one, but further studies are currently underway tackling this issue (this on-going research will also add further insights into participants' visual fixations in the two recipient roles, and it tests recipients' gesture comprehension avoiding the suppression of gestural information/no responses). That said, we believe that the Competing Modalities Hypothesis provides the more intuitive account, and some additional data we have collected speak to this preliminary conclusion, too: as a follow-up analysis, we obtained ratings for all of our stimuli from an independent set of participants which gave us an

insight into the degree of ambiguity/clarity of the individual gesture stimuli in the absence of speech. If the Fuzzy Representation Hypothesis should hold, gestures that are more ambiguous in the absence of speech (i.e., less pantomimic) should cause unaddressed recipients particular problems of interpretation (since they require more interpretation effort and more integration work). Unaddressed recipients should therefore have taken especially long to respond to those iconic gestures. However, when we correlated unaddressed recipients' RTs with the ambiguity ratings of the individual gestures, no relationship of this sort was found. One could argue, of course, that more pantomimic gestures should slow unaddressed recipients down also if we assume that the Competing Modalities Hypothesis holds, since they might 'stick' particularly well in the participant's mind; that is, the gestural components of utterances accompanied by more pantomimic gestures might become particularly prominent parts of unaddressed recipients' mental representations. However, this argument only holds if we assume that unaddressed recipients also integrate the verbal and the gestural information less than addressed recipients. According to the Fuzzy Representation Hypothesis, this would be the case, since processing the gestural information less well will also affect the integration of this information with speech. According to the Competing Modalities Hypothesis, however, unaddressed recipients may integrate speech and gesture to the same extent as addressed recipients, with the addition that they process the gestural information more strongly than them. Therefore, our favoured interpretation is the Competing Modalities Hypothesis, but further research is needed before we can draw firm conclusions.

### **Automatic and controlled processes in gesture comprehension**

With regard to the difference in response patterns for the speech-related and gesture-related targets, our results also relate to the distinction between automatic and controlled processes in co-speech gesture comprehension (Kelly et al., 2010). Bear in mind that these two sets of targets were designed to tap the processing of the speaker's information in different ways. What is striking is that, in the case of speech-related targets - a comparatively easy task - we observed a clear modality effect, as has been demonstrated by previous comprehension studies. The semantic integration of gesture and speech has been argued to be automatic in the sense of a low-level, fast and obligatory process (Kelly et al., 2010; see also Kelly, Özyürek & Maris, 2010). This explanation would account for the intrusion of the gestural information (longer RTs in the speech + gesture conditions) despite participants here having been asked to judge the content of speech only. At the same time, the results show a lack of an effect of our gaze manipulation, indicating that this task (saying 'yes' in response to a visually presented word that was presented auditorily immediately prior to this) might have been so

easily and quickly accomplished that higher-order processes involved in the processing of pragmatic information, and judgements of speaker-intentions in particular, may not have come into play.

Responses to the gesture-related targets tell a different story, however. Here, participants were slower in general, indicating that this task might have been perceived as more difficult (i.e., saying 'no' to indicate that a certain meaning had *not* been mentioned by the speaker, despite the meaning of the word displayed on screen being related to the meaning in the speaker's gesture). This seems plausible since participants here had to consult their mental representation more carefully by actively teasing apart what they heard and what they saw to arrive at a decision. In order to answer accurately, they were essentially required to suppress the gestural information they had received. This contrasts with the speech-target situation where intrusion of the gestural information may have slowed participants down since there was more information to process, but the gestural information did not interfere as such; after all, 'ticking' items on a list is part of the event of 'going through' a list. To answer 'yes', which participants were required to do for speech-related targets, is still correct, even if the gestural information is taken into account. Answering 'no' to the gesture-related targets involved a very different process, as it required the temporary suppression of the gestural information. Consequently, slower and more difficult information processing may have led to the involvement of more controlled, higher-order cognitive operations, which do take into consideration the intentional stance of a speaker.

### **Effects of eye gaze direction on speech processing**

Our results reveal that speech comprehension was not different for addressed and unaddressed recipients<sup>2</sup>. One possibility therefore is that gesture comprehension processes are more sensitive to perceived communicative intent as signalled through a speaker's gaze than the processing of speech is. However, we have to remain cautious with drawing such a conclusion from the present data. This is because the current paradigm required participants to judge speech (i.e., whether a certain word had been *mentioned* in the preceding clip). As a consequence, participants will have devoted much attention to the processing of speech, irrespective of their recipient status, since they were always required to respond. In other words, a paradigm that does not ask participants to explicitly devote attention to the verbal modality might reveal a modulation of eye gaze direction for the processing of speech only utterances also. In fact, based upon the Competing Modalities Hypothesis,

<sup>2</sup> The lack of an effect of recipient status on the processing of speech stands, at first sight, in slight contrast to a study by Schober & Clark (1989) who found overhearers to be slower and less accurate in their understanding of verbal references than addressees. However, in their study, overhearers were not official recipients of the communication, which distinguishes them from unaddressed recipients (Goffman, 1981).

one might expect to see such a modulation, since unaddressed recipients have more cognitive resources available that they are able to devote to the processing of speech. The present study was designed to mainly measure the processing of gesture, and future research is needed to provide more conclusive answers to the question of how speakers' eye gaze direction influences the comprehension of speech.

## Conclusion

In sum, our study suggests that recipients keep an eye on where speakers are looking, and this subtle piece of information has a significant impact on the extent to which co-speech gestures are processed.

## Acknowledgments

We thank four anonymous reviewers for their helpful feedback on an earlier version of this paper. We also thank the European Commission for funding this research (Marie Curie Fellowship, EU Project number: 255569), Manuela Schuetze and Nick Wood for extensive help with creating the experimental materials, Ronald Fischer and Idil Kokal for their help with programming, the participants who took part in our study, as well as the NBL and GSL lab groups, Ivan Toni, Natalie Sebanz and Guenther Knoblich, for valuable feedback in discussions of the design of our study.

## References

- Argyle, M. and Cook, M. (1976). *Gaze and mutual gaze*. Cambridge University Press.
- Chawla, P., & Krauss, R. M. (1994). Gesture and speech in spontaneous and rehearsed narratives. *Journal of Experimental Social Psychology*, 30, 580-601.
- Gerwing, J. & Bavelas, J.B. (2004). Linguistic influences on gesture's form. *Gesture*, 4, 157-195.
- Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press.
- Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, 19, 1175-1192.
- Holler, J., Shovelton, H., & Beattie, G. (2009). Do iconic gestures really contribute to the semantic information communicated in face-to-face interaction? *Journal of Nonverbal Behavior*, 33, 73-88.
- Holler, J., & Stevens, R. (2007). An experimental investigation into the effect of common ground on how speakers use gesture and speech to represent size information in referential communication. *Journal of Language and Social Psychology*, 26, 4-27.
- Hostetter, A. B., Alibali, W. M., & Kita, S. (2007). I see it in my hand's eye: Representational gestures are sensitive to conceptual demands. *Language and Cognitive Processes*, 22, 313-336.
- Kampe, K., Frith, C.D., & Frith, U. (2003). "Hey John": Signals conveying communicative intention towards the self activate brain regions associated with mentalising regardless of modality. *Journal of Neuroscience*, 23, 5258-5263.
- Kelly, S. D., Barr, D., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language*, 40, 577-592.
- Kelly, S. D., Byrne, K., & Holler, J. (2011). Raising the ante of communication: Evidence for enhanced gesture use in high stakes situations. *Information*, 2, 579-593.
- Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, 89, 253-260.
- Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, 21, 260-267.
- Kelly, S. D., Ward, S., Creigh, P., & Bartolotti, J. (2007). An intentional stance modulates the integration of gesture and speech during comprehension. *Brain and Language*, 101, 222-233.
- Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, 26, 22-63.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.
- Özyürek, A. (2002). Do speakers design their cospeech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory and Language*, 46, 688-704.
- Pelphrey, K. A., & Perlman, S. B. (2009). Charting brain mechanisms for the development of social cognition. In J. M. Rumsey & M. Ernst (Eds.), *Neuroimaging in Developmental Clinical Neuroscience*. Cambridge University Press.
- Schilbach, L., Wohlschläger, AM, Newen, A, Krämer, N, Shah, NJ, Fink, GR, Vogeley, K (2006). Being With Others: Neural Correlates of Social Interaction. *Neuropsychologia*, 44, 718-30.
- Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, 21, 211-232.
- Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., & Small, S. L. (2009). Gestures orchestrate brain networks for language understanding. *Current Biology*, 19, 661-667.
- Senju, A., & Johnson, M. H. (2009). The eye contact effect: Mechanisms and development. *Trends in Cognitive Sciences*, 13, 127-134.
- Straube, B., Green, A., Jansen, A., Chatterjee, A., & Kircher, T. (2010). Social cues, mentalizing and the neural processing of speech accompanied by gestures. *Neuropsychologia*, 48, 382-393.
- Willems, R. M., Ozyurek, A., & Hagoort, P. (2007). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, 17, 2322-2333.



# Learning from speaker word choice by assuming adjectives are informative

Alexandra Horowitz

ahorowit@stanford.edu

Department of Psychology  
Stanford University

Michael C. Frank

mcf Frank@stanford.edu

Department of Psychology  
Stanford University

## Abstract

Pragmatic abilities are not only a component of efficient communication; they can also be an important learning mechanism for young children. We discuss four experiments and a corpus analysis to investigate whether children and adults can infer information about a speaker's knowledge based on the choice of an adjective. In Experiments 1 – 3, we found that adults are sensitive to adjective use as an indicator of intended contrast dimension (e.g. that people say “red” if an object could have been blue, but “tall” if it could have been short). In Experiment 4, we found developmental differences between older and younger 4-year-olds: older children were above chance at selecting the referential dimension of interest, while younger children exhibited some contrast inference but a strong color bias. This suggests that by preschool, children are beginning to make inferences from a speaker's word choices, but that there are differences between adjective types. We conducted an exploratory corpus analysis to investigate possible causes for this developmental difference.

Keywords: Pragmatics; adjectives; language development.

## Introduction

A key feature of human language is its ability to convey information efficiently in context. For adults, the ability to make pragmatic inferences—extrapolations about meaning in context—can dramatically facilitate the exchange of information between conversational partners. For example, from “I can't find my left shoe,” we can make the inference that the speaker probably knows where her right shoe is, or else she would have mentioned that both of her shoes were lost. For children, sensitivity to word choice information is instrumental in early word learning; pragmatic inference helps guide both language acquisition and comprehension. In general, recognizing that speakers have chosen to say something in a particular way *because of some communicative goal* is an integral part of understanding language (H. Clark, 1996).

In addition to aiding in the acquisition and comprehension of language, inferences about the pragmatic implications of speakers' wording decisions might also be an important learning mechanism for children. Following our shoe example further: Children who learn to infer implied information embedded in word choices can incorporate implicit knowledge (we are talking about one specific shoe) and move on to acquire additional information (“where was it last seen?”) rather than spending time repeating and confirming implied details (“only one shoe is lost?”). Using speakers' word choices to make broader inferences allows children to learn from both what is stated and what are implied alternatives. The earlier and faster children can recognize implicit contrasts from word choice, the greater their opportunities to make use of this information.

The goal in our studies was thus to investigate whether children can learn from how speakers choose to express themselves. We focused on adjective use as a case study because adjectives are optional and may signify cues to contrast and noteworthy features. As an extension of the principle of contrast—that a contrast in form signifies a contrast in meaning (E. Clark, 1987)—children should interpret that referential expressions modified with adjectives convey different types of information than expressions with bare nouns. This makes adjective interpretation a useful domain for examining how children form implicit inferences from a speaker's word choices. We began by looking at color and size terms because they are some of the earliest-learned and most commonly-used adjectives.

Adults perceive adjective use as marking contrast, but there are differences between their comprehension and production. Grice's maxim of quantity—that speakers should be only as informative as is necessary—predicts that modifiers should be used selectively to disambiguate target referents from contrast sets (Grice, 1975). Indeed, in comprehension, visual search findings reveal that adults process color and size information as it comes online and are faster to locate a modified referent (e.g. “big comb”) when a contrasting competitor item is present (e.g. a small comb) than when a distractor item is present (e.g. a spoon) (Sedivy, Tanenhaus, Chambers, & Carlson, 1999). This indicates that prenominal modified expressions may evoke a contrast set with the referent; adults are sensitive to implied contrast information embedded in adjective use and process prenominal modifiers incrementally to locate and disambiguate a speaker's intended referent. However, adults are not always Gricean in their production; although they rarely produce scalar modifiers without a size contrast set present, they frequently over-modify with color terms (Grodner & Sedivy, 2005; Sedivy, 2003). In all, adults are sensitive to the implications of adjective use, but they are not always maximally informative.

Children seem to be developing similar skills in their preschool-age years. By kindergarten, children can recognize the informativeness of adjective use in both comprehension and production; children are more likely to use an adjective to uniquely identify a big cup from a small cup than when only a single cup is present (Nadig & Sedivy, 2002), suggesting that they are able to consider what level of description is most useful to the perspective of an interlocutor. Preschoolers can also learn to produce unambiguous references for more complex scenes with feedback (Deutsch & Pechmann, 1982). In processing speech, 3-year-olds make use of adjective information as soon as it becomes available; they correctly look to

the bigger of two cars upon hearing “big”, even before they hear the word “car” (Fernald, Thorpe, & Marchman, 2010).

In sum, preschoolers actively process adjective information in production and comprehension as cues for uniquely identifying referents. But can children apply this knowledge to make inferences about what implicit information is conveyed by a speaker’s choice of adjective use? To our knowledge, this is the first investigation of children’s abilities to use referential expressions to make inferences about the broader context.

In this paper, we outline four experiments and a corpus analysis investigating pragmatic inferences from information contained in speaker word choice. In Experiment 1, we used a novel task to examine whether adults use adjective information to infer referential contrast, and found that they performed equally strongly with color and size terms. We slightly modified the language in Experiment 2 to make implied contrast less salient, and found that performance only decreased slightly. In Experiment 3, we examined performance with other context-dependent and context-independent features, and found no differences across modifier types. In Experiment 4, we extended our task to 4-year-olds, and found a developmental change between children younger and older than 4-and-a-half: while older 4s performed above chance, younger 4s showed a strong color bias. We conducted an exploratory corpus analysis to examine possible causes for this developmental difference.

## Experiment 1

Our goal was to examine what information listeners can infer about why a speaker chose to form an utterance in a particular way. Before studying behavior in children, we wanted to confirm our intuitions that adults interpret adjective information contrastively when more than one possible visual contrast is available. In Experiment 1, we used a novel task in which adjective choice was the only informative cue to referential contrast (Figure 1). If adults are able make inferences about a speaker’s intended contrast set from adjective information, then they will be more likely to infer contrast along the referenced dimension rather than another visible but unstated dimension. This is precisely what we found.

## Methods

**Participants** Seventy-two adult participants were recruited from Amazon’s Mechanical Turk (MTurk). They were informed that the task was designed for children. All reported that they were native speakers of English.

**Stimuli** Participants viewed an online storybook with cartoon images of aliens, with one test trial. The test featured a set of three aliens: a training exemplar, and a test set consisting of an alien identical to the training exemplar except for size and an alien identical to the training exemplar except for color (Figure 1). We used the height dimension to reflect size because piloting revealed that scaling total size was construed as reflecting age rather than size.



Figure 1: Participants saw a set of three aliens: a training exemplar (e.g. tall red) followed by a test set containing one alien that differed only by color (e.g. tall blue), and one that differed only by size (e.g. short red). Participants were told that a character uttered an expression modified by an adjective (color or size), and asked to predict which test exemplar represented what that kind of alien usually looked like.

**Procedures** Participants were introduced to a character named Allen the Alien. Allen presented the training exemplar, and said something about it, e.g. “This is a special kind of glorp. This is a red glorp.” For a third of participants the reference was about color, and for another third the reference was about size. Participants then saw the test set and were asked, “What do you think glorps usually look like?”, and prompted to select one of the two images. We measured whether adults used the adjective information to infer a contrast along the referenced dimension. For the remaining third of participants, we measured responses to a bare noun baseline in order to detect any response biases unrelated to adjective information.

## Results and Discussion

The baseline control revealed that when adults had no access to adjective information, they were exactly at chance for the image they selected. Of the 24 baseline participants, they selected the size contrast half of the time ( $n=12$ ) and the color contrast equally as often ( $n=12$ ). Additionally, participants were equally likely to select an image on either the left ( $n=12$ ) or right ( $n=12$ ) side regardless of counterbalancing ( $n=6$  for each combination of contrast type and side presentation). These results indicate that adults showed no selection bias with our stimuli when only visual cues were available.

For the experimental trials, responses were coded as correct if participants selected the alien that differed along the referenced dimension (i.e. chose the alien that differed by size

in size trials, and color in color trials). Participants selected the contrasting dimension more often than chance ( $p < .01$  in an exact binomial test for both conditions) and performance did not differ across the two conditions ( $\chi^2(1) = 0.10$ ,  $p = .75$ , Figure 2).

Adjective information was used informatively to denote contrast along a specified dimension; adults interpreted color or size reference as conveying a noteworthy or contrastive feature that distinguished different aliens. Although adults often have access to prosody and emphatic stress on prenominal adjectives as additional cues to contrastive focus in speech (Ito & Speer, 2008), they were sensitive to the implications of adjective use even in our written task. The overall high level of performance suggests that adults are attuned to the subtle informativeness of word choice information. Though the only difference across all conditions was the use of a size term vs. a color term to describe the training exemplar, adults were able to detect and apply this information when making generalizations about the broader populations. They effectively interpreted the single difference in word choice across trials as signaling either a color or size contrast.

## Experiment 2

In Experiment 1, we tried to make cues to contrast highly salient by drawing attention to the unique referent (i.e., “this is a special kind of alien”). In Experiment 2, we ran the same procedure but used a more neutral introduction (“this is a glorp”) in order to determine whether adults would still use adjective information to infer a contrast dimension. Performance decreased only slightly from Experiment 1.

### Methods

**Participants** Forty-eight MTurk workers were recruited.

**Stimuli and procedure** Stimuli were identical to Experiment 1. The only change to the procedure was a modification of the referential expression. Instead of Allen saying, “This is a special kind of glorp. This is a red glorp,” we changed the wording to, “This is a glorp. This is a red glorp.”

### Results and Discussion

Again, adult participants performed significantly above chance ( $p < .05$  in an exact binomial test for both conditions) with no difference between conditions ( $\chi^2(1) = 0.11$ ,  $p = .74$ , Figure 2). Adults used adjective information to infer a referential contrast dimension even when more overt cues to contrast were removed.

## Experiment 3

In our next experiment, we tested whether inferences about contrast from adjective use would generalize to other properties that are less ubiquitous than color and size. We re-ran the procedure keeping color constant, but used the dimensions of texture (which, like color, remains constant across contexts) and width (which, like size, is relative to context).

## Methods

**Participants** Ninety-six MTurk workers were recruited.

**Stimuli and procedure** Stimuli were similar to Experiment 1, but color was held constant and other contrasts were added. Participants either saw aliens that differed by height and texture (spiky vs. smooth), or by height and width (fat vs. thin). We again used language to highlight contrast salience, e.g. “This is a special kind of glorp. This is a [spiky] glorp.” A height, texture, or width adjective was used.

### Results and Discussion

Responses were coded as correct if participants selected the alien that differed along the referenced dimension. Adult participants performed significantly above chance in exact binomial tests ( $p < .05$ ) for all conditions, and with no differences between conditions ( $\chi^2(2) = 0.55$ ,  $p = .76$ , Figure 2). This result suggests that adults are able to infer relevant contrast information from adjective use across a variety of context-dependent and -independent dimensions. Our findings thus show that adults are sensitive to pragmatic inference from adjective use across a variety of adjective types. Our next experiment tests children’s sensitivity.

## Experiment 4

To assess children’s ability to use adjective choice to infer contrast, we used a similar paradigm. We focused on color and size contrasts with 4 – 5 year-old children because this has been found to be an age of pragmatic development in other studies (Barner, Brooks, & Bale, 2011). If young children infer the potential informativeness of adjective use to convey cues to contrast, then they should be more likely to select the image that differs along the referenced dimension (i.e. color or size). If they are not sensitive to this information, then they should select an image at random or according to a baseline bias for one dimension or the other.

### Methods

**Participants** We recruited 46 four-year-old children from the Bing Nursery School at Stanford University. Pilot testing suggested response differences by older and younger 4-year-olds, so we recruited two age groups: twenty-four children age 4.0 – 4.5 (mean age 4 years 2 months) and 22 children age 4.5 – 5.0 (mean age 4 years 10 months).

**Stimuli** We used a similar task design to the previous experiments, but printed a physical book that the experimenter read with children. Each child received two training and six test trials. Each test trial used a unique set of three aliens: a training exemplar alien and two test exemplar aliens that differed from the training alien each by only color or only size.

**Procedures** The experimenter read the book to each child individually in a quiet room at the Bing Nursery School. Children were introduced to Allen the Alien, and then completed two training trials with familiar items to get them used to the study design (e.g., “This is a special kind of milk. This is

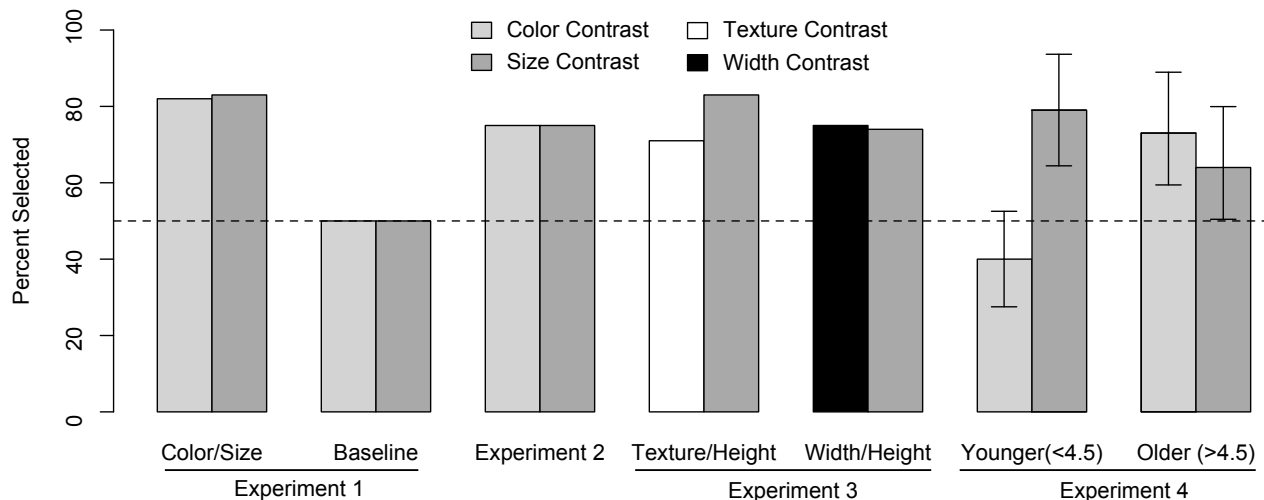


Figure 2: Mean percent correct performance across all conditions of Experiments 1-4. The dashed line represents chance (50%). Error bars represent 95% confidence intervals.

chocolate milk. What does milk usually look like?”). Training trials featured adjectives other than color and size and only one relevant contrast choice (e.g. plain milk vs. orange juice). If children did not select the correct training image, they were prompted until the correct image was chosen.

Training was followed by six test trials. For each test trial, the child was shown a picture of a single exemplar alien and told something about it, e.g. “This is a special kind of glorp. This is a tall glorp.” Children were then shown two pictures, one of an alien that differed from the exemplar by size, and one that differed by color. They were asked, “What do you think glorps usually look like?” Children received six test trials: two trials using size adjectives (e.g. “this is a tall glorp”), two trials using color adjectives (e.g., “this is a red glorp”), and two trials with no adjective to serve as a baseline (e.g., “this is a glorp”). Adjectives were focused with contrastive stress. The experimenter averted her gaze while children pointed to their response. The sessions were video-recorded. Trial types and test pictures were counterbalanced across children, and alien sets were presented in one of two orders. The task took about 10 minutes to complete.

## Results and Discussion

Preschool-aged children did show sensitivity to the implications of adjective use. Overall, preschoolers could pick out adjective information as marking an implied contrast dimension. Nevertheless, we saw an interesting developmental trend in the data (Figure 2).

We analyzed our results using a logit mixed model, predicting correct responses as an interaction between age (older vs. younger) and contrast type with random effects of participant and alien type. There was a significant effect of age, such that older children performed better than younger children ( $\beta = -2.06$ ,  $p < .001$ ). There was also a significant

interaction between age and contrast (color or size) such that older children performed above chance for both color and size trials, and younger children responded above chance for size trials but were only at chance for color trials ( $\beta = -3.25$ ,  $p < .01$ ). Overall, this analysis suggested weak responding by the younger 4s with successes by the older 4s.

To ensure that performance differences were not due to unfamiliarity with the color and size terms, we ran a posttest with a subset of children for each age group ( $n=13$  younger,  $n=12$  older). Younger children produced the correct size term over 80% and color terms 95% of the time. Older children’s production was 94% for size and 99% for color. These data suggest that younger children’s lower performance on color trials was not a result of not knowing their color words.

Baseline responses also indicated a significant developmental change. While younger 4s chose color-matching targets 71% of the time on the baseline trials, older 4s chose color-matching targets 38% of the time. In fact, nearly half ( $n=10$ ) of the younger children selected a color match for all trials in the study, while only one of the older children did. To examine this effect, we replotted our data by proportion of trials on which the color-matching target was chosen (Figure 3). A correct pattern of responding for a color trial would be choosing the size match (hence success on a color trial would be below 50% in Figure 3), while a correct response on a size trial would be choosing the color match (above 50%). Replotted in this manner, we can see that younger 4s are modifying their responses only slightly for color trials and not at all for size trials, while older 4s modify their responding slightly for color trials and considerably for size trials.

We captured this pattern with a second logit mixed model, this time predicting choice of color-matching target as a function of trial type (including baseline), age group, and their interaction. In this analysis, we saw that younger children had

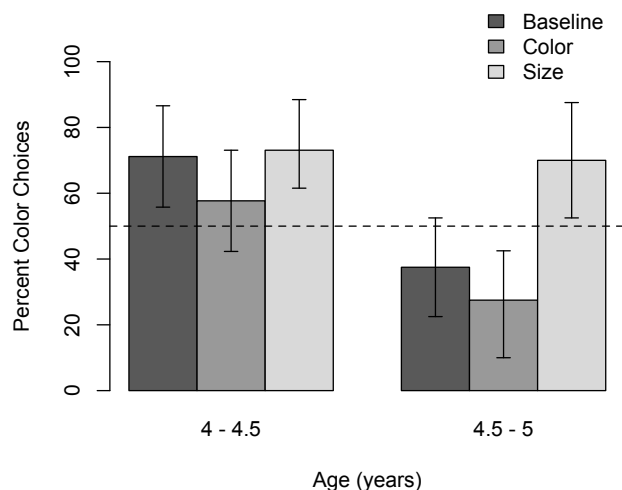


Figure 3: Mean percent color choices for all trials in the three trial types by the children in Experiment 4. Error bars show 95% confidence intervals.

a significant bias for color ( $\beta = 1.32$ ,  $p = .004$ ), and a trend towards differential responding in color trials ( $\beta = -.83$ ,  $p = .09$ ). There was a significant coefficient on older children's bias, indicating more size responding ( $\beta = -2.02$ ,  $p = .002$ ), as well as a significant interaction for size trials, indicating success in overcoming this baseline effect ( $\beta = 1.76$ ,  $p = .04$ ), but only for size trials. Thus, both groups showed some bias in their responding, but older 4s were better able to overcome that bias—at least for size—and make inferences about why a particular adjective was produced.

To summarize: Even within the narrow age range of our sample, there was a developmental difference between performance on color and size adjective trials. Although children at both ages were sometimes able to make contrast inferences, the ability to make these inferences clearly depended on the category of the adjective being used. Our next study looks to children's input to investigate one potential source for this developmental change.

### Corpus Analysis

There are a number of possible reasons for developmental differences in color and size responding. One factor raised by Sandhofer and Smith (2001) was that distinctions in the ways color and size terms are used may promote production for color terms and comprehension for size terms. We conducted a corpus study to investigate how color and size terms are presented in natural speech to children, and whether there are differences in the types of contexts and contrasts in which these terms are typically used.

**Materials** Fourteen full speech files from the Providence corpus of CHILDES (Demuth, Culbertson, & Alter, 2006) were analyzed for adult-to-child speech containing color and

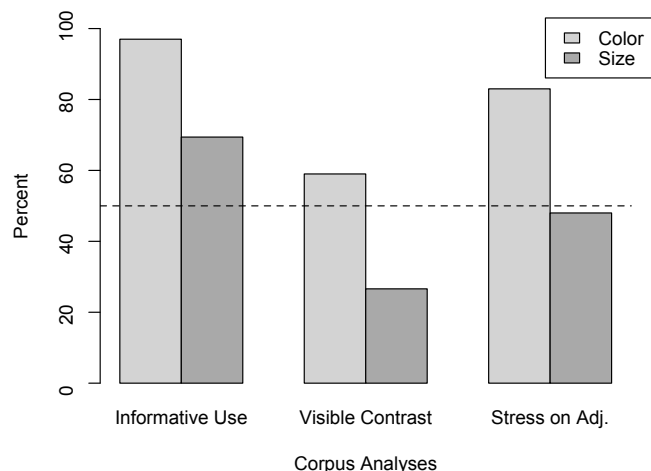


Figure 4: Percent of color and size corpus samples that were categorized by informative use, being in the context of a visible contrast, and exhibiting stress on the adjective.

size use. We used the last two or three samples for five children to most closely match the ages we used in Experiment 4 (age range 3;5.16 – 4;0.02, mean 3;8.06). A second coder blind to the hypotheses coded a random 40% of the samples. Inter-rater agreement was substantial (Cohen's  $\kappa = 0.70$ ).

**Procedures** For each file, all uses of color and size adjectives throughout the written transcript were analyzed along with the video context. For each adjective use, we marked the category (color or size), adjective (e.g. red), position (prenominal, postnominal, or adjective used as noun), stress (whether or not the adjective was marked with emphatic stress), whether a visual contrast was present (e.g. whether a big and small item were both physically available during a reference to size), and whether the term was used in a contrastive or informative context (e.g. did the adjective convey a clear possible learning opportunity for the child).

### Results and Discussion

We analyzed 330 speech instances of adjective use: 228 for size and 102 for color. The majority of the utterances used the adjective in the prenominal position (66% for color and 77% for size), but while 20% of size uses appeared in the postnominal, we only found 2 instances for color. The remaining third of color examples were used for naming or without a noun.

Emphatic stress was placed on color terms more often than size terms (83% for color, 48% for size). Color terms were also used more often with a visual contrast present than were size terms (59% color samples and 27% size samples). Results are plotted in Figure 4. It may be that adult speakers are selecting informative opportunities to use color labels with children at this age. Children may have a firmer grasp on size terms and infer contrast without a visible contrast set present.

For each of the samples, we examined whether or not the adjective was used in an informative context. Contexts were coded as informative if there was verbal reference to contrast, visual contrast, or marked salience to evoke an implied contrast with the prototype. Adjective use was considered uninformative when it did not provide qualifiable information to the child. In our samples, we found that 25% of size term use was uninformative, but this was almost never the case for color, suggesting that color and size terms are produced in different types of contexts. Color terms were nearly always used in instances that provided some type of information to the child (97%), whereas size was used informatively in 70% of instances. In addition, color was used to reference a visible contrast that was present in the immediate scene—as opposed to a reference to an unseen or implicit contrast item—44% of the time but only 16% for size.

It may be the case that children expect color terms to highlight the salience of a particular item in a visible contrast set instead of selecting an implied contrast set. In addition, perhaps the greater information contained in color utterances for children at this age may lead children to be biased to pair items on the basis of color, as in baseline performance in Experiment 4. Although more work is needed to understand the links between corpus distribution and behavior, the differences in use we observed might lead children to interpret color and size information in different ways.

## General Discussion

Can adults and children learn from speakers' choice of a particular adjective? Our results suggest that adults are able to infer the general structure of a category based on the words chosen to describe a specific, anomalous example. Children also showed sensitivity to word choice, though we saw developmental differences between ages four and five. Children older than four-and-a-half sometimes succeeded in making inferences based on word choice, while younger children primarily exhibited a color bias. Our corpus analysis suggests that the language adults use with children around this age may mark color as implying a salient, immediate dimension, whereas size was used for a wider variety of functions.

The ability to make inferences from speakers' word choices may not only reflect more adult-like comprehension, but may also be an important learning mechanism for children. The earlier and faster children can go beyond what is stated at face value, the more opportunities they have to gain further knowledge through pragmatic cues. This kind of inference is consistent with work suggesting that pragmatic mechanisms can be used in a variety of different kinds of inferences: for inferring speaker meaning, learning words, and in this case, inferring facts about the world (Frank, Goodman, Lai, & Tenenbaum, 2009).

Children who are able to recognize that word choices can convey broader information will have greater opportunities for learning about the world because they can recognize both what is explicitly stated and implicitly implied. This abil-

ity allows children both to learn more from each utterance and to increase learning opportunities. Although pragmatic inferences are not always easy for children, our results suggest that these inferences may become an important source of background knowledge about the world.

## Acknowledgments

Thanks to the staff and families at the Bing Nursery School.

## References

- Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, 118(1), 84–93.
- Clark, E. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition*, 1, 33.
- Clark, H. (1996). *Using language* (Vol. 4). Cambridge University Press Cambridge.
- Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and Speech*, 49(2), 137–173.
- Deutsch, W., & Pechmann, T. (1982). Social interaction and the development of definite descriptions. *Cognition*, 11(2), 159–184.
- Fernald, A., Thorpe, K., & Marchman, V. (2010). Blue car, red car: Developing efficiency in online interpretation of adjective-noun phrases. *Cognitive psychology*, 60(3), 190–217.
- Frank, M., Goodman, N., Lai, P., & Tenenbaum, J. (2009). Informative communication in word production and word learning. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 206–211).
- Grice, H. (1975). Logic and conversation. 1975, 41–58.
- Grodner, D., & Sedivy, J. C. (2005). The effect of speaker-specific information on pragmatic inferences. *The processing and acquisition of reference*. MIT Press: Cambridge, MA.
- Ito, K., & Speer, S. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, 58(2), 541–573.
- Nadig, A., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, 13(4), 329.
- Sandhofer, C., & Smith, L. (2001). Why children learn color and size words so differently: Evidence from adults' learning of artificial terms. *Journal of Experimental Psychology: General*, 130(4), 600.
- Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32(1), 3–23.
- Sedivy, J. C., Tanenhaus, M., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71(2), 109–147.

# Finishing each other's ... Responding to incomplete contributions in dialogue

Christine Howes, Patrick G. T. Healey, Matthew Purver, Arash Eshghi

{chrizba, ph, mpurver, arash}@eecs.qmul.ac.uk

Queen Mary University of London

Interaction, Media and Communication Research Group

School of Electronic Engineering and Computer Science, London E1 4NS, UK

## Abstract

A distinguishing feature of dialogue is that contributions can be fragmentary or incomplete. Such incomplete utterances may be later completed by another interlocutor. These cross-person *compound contributions* (CCs) have been hypothesised to be more likely in predictable contexts but the contributions of different sources of predictability has not been systematically investigated. In this paper we present an experiment which artificially truncates genuine contributions in ongoing text-based dialogues, to investigate the effects of lexical, syntactic and pragmatic predictability of the truncation point on the likelihood of one's interlocutor supplying a continuation. We show that what is critical is the actual and presumed accessibility of common ground, and that while people are sensitive to syntactic predictability, this alone is insufficient to prompt a completion.

**Keywords:** Dialogue; compound contributions; common ground.

## Introduction

It is well known that contributions to dialogue are often fragmentary or in some sense unfinished Fernández and Ginzburg (2002). These incomplete utterances may be subsequently completed, either by the original speaker following some response or interruption from an interlocutor, or, by another person (Purver et al., 2009).

These *compound contributions* (CCs) are a paradigmatic feature of dialogue, and cross-person CCs in particular are a key indicator of coordination between interlocutors. Although naturally occurring cross-person CCs and their interpretations have been studied (Lerner, 1996; Purver et al., 2009), there has not previously been a systematic, experimental, attempt to investigate the factors that influence how a completion for an incomplete utterance may be produced. Intuitively, people's willingness to finish another person's incomplete utterance will depend (at least) on how predictable the rest of the utterance is. There are several sources of possible predictability.

*Expansions* are CCs which add material (e.g. an adjunct) to an already complete syntactic element; *completions* are CCs which complete an incomplete element. Conversation analytic (CA) discussions of CCs suggest that they should preferably occur at *transition relevance places* (TRPs), points that are foreseeable by the participants. *Expansions* are CCs with split points at TRPs, and are more common in spoken dialogue (Howes et al., 2011) so ought to be more likely than completions.

**Hypothesis 1** *Cross-person completions are more likely at transition relevance places*

Second, *completions* should tend to occur at syntactically projectable points (e.g. compound turn constructional units Lerner, 1991).

**Hypothesis 2** *Cross-person completions are more likely when they are syntactically predictable.*

A third source of predictability comes from the degree to which the speaker and hearer share, or can be assumed to share, common ground relevant to the CC. If the topic of the utterance is already in the common ground then the content of the completion is more predictable.

**Hypothesis 3** *Cross-person completions are more likely when they address topics that are part of the common ground.*

The effects of these different forms of predictability are directly tested here for the first time using a text chat experiment performed with the DiET experimental platform. The evidence points towards shared knowledge being a key factor with other sources of predictability also contributing.

## Method

In this experiment, to see what factors influence how people respond to unfinished turns and their likelihood of producing a continuation, a number of genuine single contributions in dyadic text-based conversations were artificially split into two parts, using the DiET chat tool.

## The DiET chat tool

The Dialogue Experimental Toolkit (DiET) chat tool is a text-based chat interface into which interventions can be introduced into a dialogue in real time. These interventions can take a number of forms; turns may not be relayed, additional turns may be added, as in Healey et al. (2003), in which spoof clarification requests are added to the dialogue, or turns may be altered prior to transmission. As these manipulations occur as the dialogue progresses, they cause a minimum of disruption to the 'flow' of the conversation.

The DiET chat tool is a custom built Java application, consisting of two main components: the server console and the user interface. The server time-stamps and



stores each key press, and acts as an intermediary between what participants type and what they see. All turns are passed to the server, from where it is relayed to the other participants. Prior to being relayed, real turns can therefore be automatically altered by the server or not relayed, or fake turns can be introduced.

**Character-by-character interface** In the character-by-character version of the DiET chat tool, the user interface consists of a single chat window. Below this, there is a status bar, which indicates if any participants are actively typing (see figure 1).

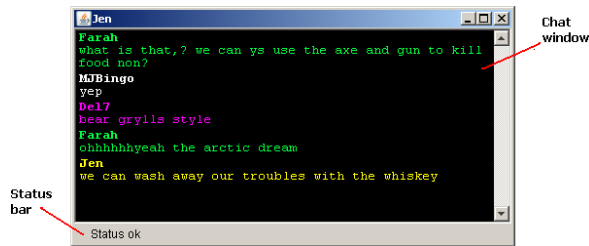


Figure 1: The DiET chat window (as viewed by *Jen*)

Unlike traditional chat interfaces (such as MSN Messenger), users type directly into the same window in which they see their interlocutors' contributions. This means that each character that any of the participants type is displayed in the window at the time it is entered – i.e. users see both their own and their interlocutors' contributions unfold in a character-by-character fashion. Consequently, only one participant may type at a time.

## The intervention

For this experiment, single contributions were artificially split into two parts. The first part was transmitted to the other participant as it was typed, with the turn truncated according to various factors as discussed below. Following a pilot study, which showed that people were more likely to supply a response after a filler “...” or “...?” than if there were no filler (after a filler: 18/26, 69%, no filler: 12/45, 27%;  $\chi^2_{(1)} = 12.24, p < 0.001$ ), the truncated first part of the genuine turn was followed by a text filler. Subsequently, there was a delay of 12 seconds, during which the other person could respond if they wished. Any response was trapped by the server and not relayed to the original sender, before the rest of the original (interrupted) contribution was transmitted.

Split points are manipulated according to measures of a) syntactic and b) lexical predictability calculated as each turn is produced.

## Entropy

Entropy is a measure of uncertainty: the higher the entropy, the higher the uncertainty; and the lower the entropy, the higher the predictability. Here we used two

measures: part-of-speech entropy, to capture the “syntactic” predictability of one part-of-speech (POS) following another; and lexical entropy, to capture the predictability of a particular lexical item following a specific POS. To illustrate the difference: although determiners are predictably followed by nouns, there are lots of different nouns: determiners therefore have a relatively low POS entropy, and a relatively high lexical entropy.

Since predictability depends on dialogue context and topic, entropy values were calculated from a corpus of prior dialogues (53663 word tokens) collected using the same tool and domain (the balloon task – see below).<sup>1</sup> POS tags were generated using the Stanford POS tagger (Toutanova et al., 2003) with a misspellings map for common chat abbreviations and typos. For each POS, entropy was calculated as follows over the observed types of the following POS  $S$  or lexical item  $L$ :

$$H_{pos} = - \sum_S p_S \log(p_S) \quad H_{lex} = - \sum_L p_L \log(p_L)$$

During the experiment, a POS-tagger analysed the strings in real time and triggered an intervention based on these entropy values, and a minimum requirement of 9 words (based on the mean length of all contributions). This manipulation produced a range of interventions with high, medium and low POS entropy, and, independently, high, medium and low lexical entropy.

## Subjects and materials

The experiment was carried out on 16 pairs of students from Queen Mary University of London who were each paid £7.00 or given course credit for providing an hour of their time. The task was the *balloon task* – an ethical dilemma requiring agreement on which of three passengers should be thrown out of a hot air balloon that will crash, killing all the passengers, if one is not sacrificed. The choice is between a scientist, who believes he is on the brink of discovering a cure for cancer, a woman who is 7 months pregnant, and her husband, the pilot. This task was chosen on the basis that it is known to stimulate discussion, leading to dialogues of a sufficient length to enable an adequate number of interventions.

Subjects were seated at desktop computers in separate rooms, asked to input their e-mail address and username and given the task description. They were told that the experiment was investigating the differences in communication when conducted using a text-only interface as opposed to face-to-face, that the experiment would last approximately 45 minutes, and that all turns would be recorded anonymously for later analysis.

## Analysis

Each intervention was annotated according to a number of factors. Firstly, whether or not there was a response

<sup>1</sup>This corpus is small, but extremely domain specific.

to the intervention during the timeout period. If there had been, the type of response was coded according to whether it was a compound contribution (CC), a clarification request (CR) or a yes/no response.<sup>2</sup> These are not mutually exclusive – example (1) is a CR constructed as a CC, and example (2) is a CC and a yes/no answer. The minimum POS entropy was 1.44, maximum 4.16, mean 3.27 (standard deviation 0.87); for lexical entropy those values are 5.59–8.14, mean 7.03 (s.d. 0.60).

- (1) **B:** also surely the guy who knows how to ...  
**N:** fly?  
**B:** fly the baloon should know how to inscrease its height? [DiET CCInd9 1277-80]
- (2) **J:** do you assess their value to society ...  
**Q:** in milliseconds yes =  
**J:** firstim with nick qne wuwi and susie - tom can explain how toise use the hot air balloon before he jumps [DiET CCInd13 2048-51]

The intervened turn was also annotated for whether it was potentially end-complete and could therefore be responded to as if it were a complete contribution. Antecedent end-completeness can be used as a proxy measure for pragmatic completeness, with 40 of the 241 truncated contributions appearing to end in a complete way.

The other major factor predicted to increase production of CCs was whether the subject under discussion was known to be shared. Lexical entropy gives us a measure of the predictability of the local context, with entities and concepts more or less predictable in certain sentential contexts because of the limited domain. However, it does not capture the potential effect on the predictability of local upcoming material of the *shared* context established in the course of any particular conversation between a specific pair of individuals. Each intervened contribution was therefore classified as either contributing to an ongoing topic of discussion, or introducing a new topic, as a loose measure of common ground.

## Results

Of the 241 interventions, 171 elicited a response (71%). A GEE analysis with whether or not there was a response to the intervention as dependent variable<sup>3</sup> with POS and lexical entropy values as covariates, antecedent end-completeness as a fixed factor and participant as subject effect (goodness of fit QIC = 294.562; see table 1) showed a main effect of antecedent end-completeness such that responses were more likely in cases that could be considered complete on their own, showing that people are sensitive to TRPs.

<sup>2</sup>These response types were chosen on the basis of an examination of the response data.

<sup>3</sup>All models in this paper use a binary model with a logit link function and an independent correlation structure unless otherwise stated.

There was also an interaction effect of POS entropy by lexical entropy ( $B = 0.237$ , Wald- $\chi^2 = 5.893$ ,  $p = 0.015$ ). This effect is illustrated in figure 2. Simple slopes analysis (Aiken et al., 1991) showed that responses are more likely in cases where both POS and lexical entropy were high (the highly unpredictable cases) than in cases where one or both levels of entropy were low.

IV	Model effects	
	Wald $\chi^2$	p
Antecedent end-complete (Ant)	4.286	0.038*
Lexical entropy (Lex)	0.148	0.700
POS entropy (POS)	0.593	0.441
Ant $\times$ Lex	3.251	0.071
Ant $\times$ POS	2.546	0.111
Lex $\times$ POS	6.460	0.011*
Lex $\times$ POS $\times$ Ant	0.287	0.592

Table 1: GEE of response or not by lexical entropy, POS entropy and antecedent end-completeness

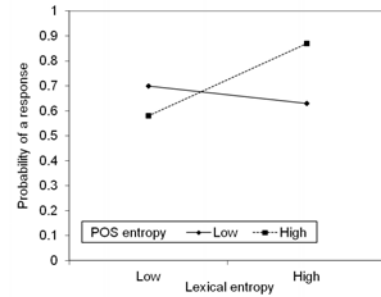


Figure 2: Marginal means of probability of a response by POS entropy  $\times$  lexical entropy

## Type of response

The results outlined above may conflate different effects which are specifically associated with different kinds of response. Analyses were therefore carried out separately on the different types of responses.

Response type		Antecedent end-complete			
		N	%	Y	%
Yes/No	Y	20	15	12	36
	N	118	85	21	64
CR	Y	39	28	2	6
	N	99	72	31	94
CC	Y	62	45	10	30
	N	76	55	23	70
Total		138	69	33	83

Table 2: Response type

The breakdown of the 171 responses is shown in table 2.<sup>4</sup>

<sup>4</sup>Note that there were no differences in types of response according to which filler type was used ('...' or '...?').

Participants were more likely to produce a Yes/No response if the antecedent is end-complete ( $\chi^2_{(1)} = 8.374, p = 0.004$ ), and they are also less likely to respond with a clarification request ( $\chi^2_{(1)} = 7.201, p = 0.007$ ). There is no difference in the proportion of responses constructed as CCs based on whether the antecedent was end-complete or not, which is unexpected given the preference for expansions over completions in corpus studies (Purver et al., 2009).

### CR responses

With the data filtered to responses only, GEE analyses on whether or not the response was formulated as a CR, with the POS and lexical entropy values as covariates and participant as subject effect (goodness of fit = 186.828) showed a main effect of POS entropy (see table 3). Greater syntactic predictability (lower POS entropy) increased the probability of the response being a clarification request.

IV	Model effects	
	Wald $\chi^2$	p
Lexical entropy	2.207	0.137
POS entropy	5.135	0.023*
Lex $\times$ POS entropy	0.176	0.674

Table 3: GEEs CRs by lexical entropy, POS entropy and antecedent end-completeness

CRs are often formulated as CCs, as in (3) which is particularly true where the syntactic category of the next word was highly predictable (independently of lexical entropy). Of the 72 CCs, 21 occurred in syntactically predictable (low POS entropy) conditions with 12 of these also being CRs. Of the other 51 CCs, only 13 were also CRs (57% vs. 25%;  $\chi^2_{(1)} = 6.575, p = 0.010$ ).

- (3) **N:** i think susie because she is t ...  
**B:** a woman?  
**N:** ehe least important out of the three if you think about it ... dr nick is a doctor and could be really useful in the world [DiET CCInd9 1214-7]

### CC responses

IV	Model effects	
	Wald $\chi^2$	p
Antecedent end-complete (Ant)	1.951	0.162
Lexical entropy (Lex)	3.586	0.058
POS entropy (POS)	0.235	0.627
Ant $\times$ Lex	15.835	<0.001**
Ant $\times$ POS	0.018	0.894
Lex $\times$ POS	0.344	0.558
Ant $\times$ Lex $\times$ POS	0.005	0.945

Table 4: GEE of CCs by lexical entropy, POS entropy and antecedent end-completeness

GEE analyses on whether or not the response was formulated as a CC, with the POS and lexical entropy val-

ues as covariates, participant as subject effect and antecedent end-completeness as a fixed effect (goodness of fit = 234.351) showed an interaction between antecedent end-completeness  $\times$  lexical entropy (table 4).

Simple slopes analysis shows that if the next lexical item is unpredictable then you are more likely to formulate your response as a CC if the antecedent is not end-complete. When the antecedent is end-complete (the solid line in figure 3), responses are more likely to be continuations in more highly predictable contexts (as in e.g. (4)), but when it is not end-complete CCs are more likely in the lexically unpredictable cases (as in e.g. (5)).

- (4) **W:** I feel like we should be talking ...?  
**J:** about the prompt?  
**W:** about something important.  
[DiET CCInd16 2846-9]
- (5) **W:** nope we are not god we are ...?  
**M:** [M] and [W] ini lol we are [M] and [W]  
**u** fool lol so s just shut up npw please ad  
**thank u for ur c kindness**  
**W:** not making dis di decision i knw we got bre  
spellintg werrorz man i r we even aloowed to talk  
type in slang?  
[DiET CCInd6 929-32]

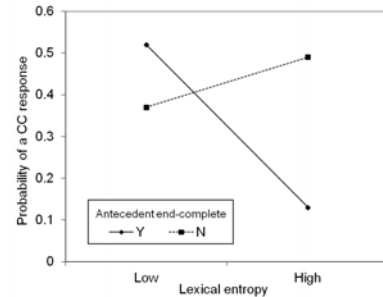


Figure 3: Marginal means of probability of a CC response by lexical entropy  $\times$  antecedent end-completeness

### Context

To test the hypothesis that CCs are more common where participants share information or common ground about the subject under discussion, planned post hoc analyses were carried out using the topic under discussion. Of the 241 intervened contributions, 170 were about an existing topic under discussion, whilst 71 introduced some new topic.

Participants were no more likely to respond if the turn was about the current topic or not; nor were they more likely to respond with a yes/no answer, or a clarification request. However, they were more likely to construct their response as a CC if it was about the current topic than if it was about something else (topic 59/121, 49% vs. Off-topic 13/50, 26%;  $\chi^2_{(1)} = 7.519, p = 0.006$ ).

Adding topic to the GEE model with CC response as dependent variable (QIC = 227.895, table 5)<sup>5</sup> resulted in a three-way interaction effect of lexical entropy  $\times$  POS entropy  $\times$  topic.

IV	Model effects	
	Wald $\chi^2$	p
Antecedent end-complete (Ant)	0.046	0.830
Topic	0.276	0.600
Lexical entropy (Lex)	2.545	0.111
POS entropy (POS)	0.018	0.892
Line number	2.361	0.124
Ant $\times$ Topic	0.381	0.537
Ant $\times$ Lex	3.435	0.064
Ant $\times$ POS	0.183	0.669
Topic $\times$ Lex	2.103	0.147
Topic $\times$ POS	0.281	0.596
Lex $\times$ POS	0.034	0.853
Ant $\times$ Topic $\times$ Lex	0.091	0.763
Ant $\times$ Topic $\times$ POS	0.005	0.946
Ant $\times$ Lex $\times$ POS	0.133	0.716
Topic $\times$ Lex $\times$ POS	8.635	0.003**

Table 5: GEE of type of CC responses by lexical entropy, POS entropy, antecedent end-completeness and topic

Exploring the interaction effect (figure 4) shows that in lexically unpredictable cases, which were syntactically predictable, participants were more likely to construct their response as a CC if they were talking about some topic which they had already been discussing, and which was therefore contextually salient.

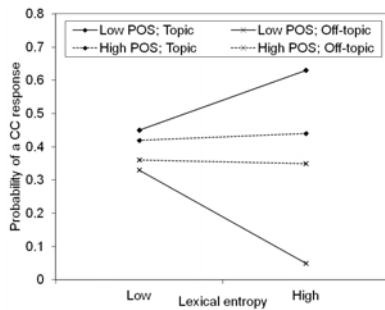


Figure 4: Marginal means of probability of a CC response by lexical entropy  $\times$  POS entropy  $\times$  topic

## Discussion

These results offer some insights regarding the conditions influencing whether and how conversational partners respond to an incomplete utterance, and when they can and do construct those responses as continuations.

There is a response to 71% of the interventions, with this proportion affected by the predictability of the upcoming material. Perhaps counterintuitively, people are

<sup>5</sup>The model also included line number as an additional covariate as it was found that participants were more likely to introduce a new topic later on in the conversation.

more likely to respond to unfinished contributions<sup>6</sup> if both syntactic and lexical items were unpredictable. This is not what we would expect if a simple model of levels of predictability were correct, as intuitively the most predictable cases ought to elicit the most responses. However, it is what we would expect if one of the drivers of human communication is in locally managing and resolving potential sources of misunderstanding (as in the interactive misalignment of Healey, 2008).

The main effect of potential completeness also demonstrates that people are more comfortable responding at all if the other person has reached a potential TRP – backing up findings from corpus studies (Purver et al., 2009) and conversation analysts assertion that people are sensitive to possible endings (Schegloff, 1996).

## Compound contributions

Contrary to Hypotheses 1 and 2, continuations are not more likely at TRPs or syntactically predictable points. What is critical seems to be the actual and presumed accessibility of common ground. If the local content of what comes next is salient from the (presumed shared) context then people will produce completions. They do this by taking advantage of the syntactic structure of the antecedent, but syntactic predictability alone is not sufficient to prompt a completion.

A continuation response is more likely if the antecedent is complete but the next word is predictable (as in e.g. (4)) or if the antecedent is incomplete, suggesting that people complete where they can.

For the cases in which the antecedent is not end-complete, responses were more likely to be constructed as CCs in lexically unpredictable cases. However, if the next lexical item is highly predictable, then it can be interpreted as if it had actually been produced, as in (6). This result is not as surprising as it first appears as in a BNC corpus study (Howes et al., 2011), only 64% of end-incomplete contributions get continued, meaning that 36% never do. These are cases in which the local context is so predictable that it can be taken to be shared without the words themselves being produced.

(6) **T:** its not that fair on the girl doing th ...

**H:** exactly, you need to think of others and not be so selfish :P

**T:** study we should do lots of chatting although i doubt she'll read past the exercise what with it not being standardised etc [DiET CCInd4 685-8]

## Context

The three-way interaction of POS entropy by lexical entropy by topic adds weight to the notion that what is critical to the production of a continuation in response

<sup>6</sup>This could be a genuine difference in text chat because of the availability of other cues in spoken dialogue, but we leave a discussion of this to one side.

to an incomplete utterance is the actual and presumed accessibility of common ground.

If the lexical item is unpredictable then syntactic predictability aids production of CCs in cases where the topic of the truncated contribution is shared, thus acting as a resource which helps frame the offered continuation as such. Syntax does not however help at all in cases where the topic is new so the gist of the contribution cannot be predicted and the predictability of the next word also offers no clues as to a plausible continuation.

This pattern of predictability corresponds to cases in which the high lexical entropy equates to lots of different words of a single type, as in the determiner case, rather than the high lexical entropy being associated with lots of different words of many different types (as with e.g. adverbs). This means that the syntactic category is highly constrained and the additional information associated with contextual salience can significantly narrow down an appropriate continuation.

### Summary

This experiment, to the best of our knowledge the first to ever systematically attempt to induce continuations in an ongoing dialogue, shows that different types of predictability have different effects on what type of response participants produce to incomplete contributions, if any.

It shows that although syntax can be mobilised in constructing a response, it is not the crucial determinant of whether people construct their responses as continuations to the immediately preceding contribution. Participants make use of syntactic predictability only if the context is sufficiently constrained. Though people respect the constraints of the syntax, different points in the sentence do not cause greater difficulty in producing something that syntactically builds off a prior turn. However, that the grammar is a mutually available resource does not mean that it is used in the same way by all interlocutors, as evidenced by the finding that clarification requests are more likely, and more likely to be formulated as continuations, when the syntactic category of the upcoming material is more predictable, as these are cases where the syntax may be exploited to localise the source of a potential misunderstanding.

Another of the main findings is that people are sensitive to potential turn endings. These may be syntactic (in the antecedent end-complete cases) but they are not necessarily so. Some cases which appear to be syntactically incomplete can be responded to as if they are complete, provided that the continuation is highly predictable. If there are indeed cases which are interpreted as complete when they are not – as if the hearer is supplying the missing material internally, but does not necessarily produce it, this has implications for any grammatical or dialogue model. Incomplete syntactic strings must be not only successfully analysed, but also assigned

potentially complete semantic representations.

The evidence from this experiment shows that when people are likely to produce CCs (or produce more CCs) is principally driven by common ground. They are possible (or more likely) when it is shared. How this is cashed out remains to be seen, however, it is apparent that some formal notion of context is crucial for a thorough understanding of CCs, especially if we are to ever hope to model them appropriately in a dialogue system.

### References

- Aiken, L. S., West, S. G., and Reno, R. R. *Multiple Regression: Testing and Interpreting Interactions*. Sage Publications, 1991.
- Fernández, R. and Ginzburg, J. Non-sentential utterances: A corpus-based study. *Traitement Automatique des Langues*, 43(2), 2002.
- Healey, P. G. T. Interactive misalignment: The role of repair in the development of group sub-languages. In Cooper, R. and Kempson, R., editors, *Language in Flux*. College Publications, 2008.
- Healey, P. G. T., Purver, M., King, J., Ginzburg, J., and Mills, G. Experimenting with clarification in dialogue. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*. Boston, Massachusetts, 2003.
- Howes, C., Purver, M., Healey, P. G. T., Mills, G. J., and Gregoromichelaki, E. On incrementality in dialogue: Evidence from compound contributions. *Dialogue and Discourse*, 2(1):279–311, 2011.
- Lerner, G. H. On the syntax of sentences-in-progress. *Language in Society*, pages 441–458, 1991.
- Lerner, G. H. On the “semi-permeable” character of grammatical units in conversation: Conditional entry into the turn space of another speaker. In Ochs, E., Schegloff, E. A., and Thompson, S. A., editors, *Interaction and Grammar*, pages 238–276. Cambridge University Press, 1996.
- Purver, M., Howes, C., Gregoromichelaki, E., and Healey, P. G. T. Split utterances in dialogue: A corpus study. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009 Conference)*, pages 262–271. Association for Computational Linguistics, London, UK, 2009.
- Schegloff, E. A. Turn organization: One intersection of grammar and interaction. In Ochs, E., Schegloff, E. A., and Thompson, S. A., editors, *Interaction and Grammar*, pages 52–133. Cambridge University Press, 1996.
- Toutanova, K., Klein, D., Manning, C., and Singer, Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*, pages 252–259. 2003.

# Identifying representations of categories of discrete items using Markov chain Monte Carlo with People

Anne S. Hsu (anne.hsu@ucl.ac.uk)

Department of Cognitive, Perceptual, and Brain Sciences, University College London, London WC1H 0AP UK

Jay B. Martin (jbmartin@nyu.edu)

Department of Psychology, New York University, New York, NY 10003 USA

Adam N. Sanborn (A.N.Sanborn@warwick.ac.uk)

Department of Psychology, University of Warwick, Coventry CV4 7AL UK

Thomas L. Griffiths (tom\_griffiths@berkeley.edu)

Department of Psychology, University of California, Berkeley, CA 94720 USA

## Abstract

Identifying the structure of mental representations is a basic problem for cognitive science. We present a method for identifying people's representations of categories that are defined over a set of discrete items, such as a collection of images. This method builds on previous work using Markov chain Monte Carlo algorithms as the basis for designing behavioral experiments, and we thus call it discrete Markov chain Monte Carlo with People (d-MCMCP). We illustrate how this approach can be used to identify the structure of visual categories using real images drawn from large databases.

**Keywords:** category representation; Markov chain Monte Carlo; image databases

## Introduction

Humans outperform the most sophisticated computers in their ability to process complex stimuli, such as recognizing faces or comprehending ambiguous linguistic input. These abilities are facilitated by organizing stimuli into categories. People's representations of categories directly affect their behavior: Recognizing scenes, parsing language, and making decisions, for example, are all influenced by people's category representations. Therefore, understanding the structure of these representations is an important goal for cognitive science. Most research on computational models of categorization has tended to use artificial stimuli, because such stimuli lend themselves to controlled experiments and yield results which are easily quantified (e.g., Nosofsky, 1986; Ashby, 1992; Nosofsky, 1987). However, the stimuli constituting real-life categories – such as images or words – are often characterized by complex features that vary along a large number of dimensions that are hard to quantify. In this paper, we present a method for estimating the structure of categories using an arbitrary discrete set of stimuli, making it possible to investigate real-life categories using complex stimuli such as images drawn from large online databases.

Many computational models of categorization can be interpreted as representing a category as a probability distribution over stimuli (Ashby & Alfonso-Reese, 1995). For example, a category  $c$  might be represented by the probability distribution over images  $x$  associated with that category,  $p(x|c)$ .

Using this insight, new experimental methods have been developed for estimating these subjective probability distributions. These methods are based on implementing Markov chain Monte Carlo (MCMC) algorithms, which are widely used in computer science and statistics for sampling from complex probability distributions. The Markov chain Monte Carlo with People (MCMCP) method (Sanborn & Griffiths, 2008; Sanborn, Griffiths, & Shiffrin, 2010) adapts MCMC algorithms so as to sample from subjective probability distributions, such as the distributions over stimuli associated with categories. The MCMCP method has been used to estimate the structure of categories defined on continuous, easily parameterized stimuli, such as stick-figure animals and basic fruit shapes (Sanborn et al., 2010) or computer-generated faces (McDuff, 2010; Martin, Griffiths, & Sanborn, 2012).

While the introduction of MCMCP made it easier to explore complex, high-dimensional representations, the original method could only be used with stimuli that vary along a fixed set of parameterized dimensions. This is a serious limitation for exploring real-life categories. For example, it is difficult to quantify the difference between faces with genuine smiles vs. disingenuous smiles. Here, we present an extension of MCMCP that removes this limitation. Our method, which we call discrete Markov chain Monte Carlo with People (d-MCMCP), allows estimation of probability distributions over arbitrary discrete sets of stimuli. This supports the exploration of categories relating to real-life stimuli such as photographic images, every-day objects, real commercial products, and linguistic materials such as documents and words. Because we no longer need to explicitly parameterize the stimuli being examined, d-MCMCP allows us to exploit the vast array of natural stimuli available from the internet.

The outline of this paper is as follows. The next section introduces the key ideas behind MCMCP. We then present our extension of this method to discrete sets of stimuli. The remainder of the paper focuses on two experiments that demonstrate the utility of this method. The first experiment explores the categories of *happy* and *sad* faces using photographic images, allowing us to compare against previous results obtained using the original MCMCP algorithm applied

to parameterized images of faces (Martin et al., 2012). The second experiment explores people’s concepts of the seasons *Spring*, *Summer*, *Autumn* and *Winter* using a set of 4000 images drawn from online databases.

### Markov chain Monte Carlo with People

Markov chain Monte Carlo algorithms are a class of methods for generating samples from complex probability distributions by constructing Markov chains that converge to those distributions over time (see Gilks, Richardson, & Spiegelhalter, 1996). If we want to draw a sample from the probability distribution  $p(x)$ , we define a Markov chain such that the stationary distribution of that chain is  $p(x)$ , and sample a sequence of states from that chain. If the sequence is long enough, the states of the chain can be treated similarly to samples from  $p(x)$ . The Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) is one of the most popular methods for constructing such a Markov chain. The sequence of states is initialized with an arbitrary value,  $x$ . The next value in the sequence is generated via a two step process. First, a candidate for the next value,  $x'$ , is chosen by sampling from an arbitrary proposal distribution conditioned on  $x$  that is specified by the designer of the algorithm,  $q(x';x)$ . Second, a decision is made as to whether that proposed value will be accepted using a valid acceptance function which is a function of the relative probability of  $x$  and  $x'$  under the target distribution  $p(x)$ . While the original Metropolis algorithm used a different acceptance function, an example of a valid acceptance function is the Barker function (Barker, 1965) which specifies the acceptance probability to be

$$A(x^*;x) = \frac{p(x')}{p(x') + p(x)} \quad (1)$$

and defines a Markov chain that converges to  $p(x)$  provided  $q(x';x)$  is symmetric, with  $q(x';x) = q(x;x')$ .

The Markov chain Monte Carlo with People method uses the idea that categories can be represented as probability distributions over stimuli (Ashby & Alfonso-Reese, 1995). The distribution over stimuli  $x$  for category  $c$ ,  $p(x|c)$  indicates the degree to which a stimulus item  $x$  is perceived to represent a given category  $c$ . In theory, the simplest approach to measuring human categories would be to ask people to rate the degree of category membership for all possible stimuli. However, this has two serious limitations. First, categories span such a large number of possible items that collecting individual ratings of each are not practical. Second, a question such as “How good is this example of a *happy face*?” is difficult to answer, and there is no obvious scale to use for the answer. A solution to the second limitation would be to ask people to make pairwise judgments, i.e. “Which example is a better example of a *happy face*?”. However, this only exacerbates the first limitation because the number of judgments required for all possible pairs of  $n$  items is now on the order of  $n^2$ .

Markov chain Monte Carlo with People addresses the challenge of estimating the distribution associated with a category

by constructing a Markov chain that produces samples from that distribution. The method is based on a correspondence between human choice behavior and the Barker acceptance function. If a task can be constructed in which people are offered a choice between  $x$  and  $x'$  and choose  $x'$  with probability

$$P_{\text{choice}}(x';x|c) = \frac{p(x'|c)}{p(x'|c) + p(x|c)} \quad (2)$$

then this provides a valid acceptance function for a Markov chain that will converge to  $p(x|c)$ . Equation 2 has a long history as a model of human choice probabilities, where it is known as the Luce choice rule or the ratio rule (Luce, 1963; Shepard, 1957). This rule has been shown to provide a good fit to human data when people choose between two stimuli based on a particular property (Bradley, 1954; Clarke, 1957; Hopkins, 1954). The Luce choice rule has also been used to convert psychological response magnitudes into response probabilities in many models of cognition (Nosofsky, 1986; Ashby, 1992; Nosofsky, 1987; McClelland & Elman, 1986).

Based on this correspondence, the MCMCP method implements the Metropolis algorithm, using people’s choices to determine which proposals are accepted (Sanborn et al., 2010). In a standard experiment, people would be asked to make a series of pairwise decisions in which they are asked to choose the best category member from two proposed stimuli. The stimuli that are presented in each decision correspond to the values  $x$  and  $x'$  in the Metropolis algorithm, and the choices that people make determine which proposals are accepted. With enough decisions, MCMCP will converge to samples from the probability distribution associated with that category, and individual stimuli will be encountered with probability given by  $p(x|c)$ . The proposal distribution can be selected by the experimenter, provided it is symmetric in the way required by the Barker acceptance rule.

### Estimating categories for discrete sets of stimuli

The MCMCP method requires defining a proposal distribution  $q(x';x)$  for choosing the next stimulus to present on each trial based on the current stimulus. When stimuli are described by a fixed set of parameters, this is easy – previous work has used Gaussian or uniform distributions to generate proposals for continuous parameters, and a multinomial distribution can be used to propose new values for discrete features. However, real-life categories are not made up of easily parameterized items: Real-life categories apply to stimuli that are difficult to parameterize such as real objects, images, sounds, and words. The lack of parameterization makes it unclear how to propose a reasonable stimulus based on the current stimulus, which is central to the MCMCP algorithm’s efficiency. Hence, in order for MCMCP to measure representations of a wide range of real-life categories, we need to a method for making reasonable proposals when exploring stimuli that are not easily parameterized. In this section, we introduce such a method, which we call discrete Markov Chain Monte Carlo with People (d-MCMCP).



The d-MCMCP procedure adds three steps to MCMCP. The first step is to create a database of stimulus items over which the probability distribution associated with a category is to be estimated. The second step is computing a rough measure of the similarity between all possible item pairs, giving a symmetric similarity matrix,  $S$ . The similarity metric can be chosen as appropriate for the domain, and need only provide a heuristic guide to the perceived similarity of human participants. For example, similarity between color histograms can be used to quantify similarity for color images. The third step is constructing a graph of the stimulus items based on these similarities. A random walk on this graph is then used to define the proposal distribution used in MCMCP.

A key assumption in using the Barker acceptance function is that the proposals must be symmetric. That is the probability of choosing a proposal value given a current value would be the same if the proposal and current values were reversed. In order for this to be true of a random walk on a graph, the edges must be symmetric (ie. the walk can traverse an edge in each direction), and each node in the graph must have the same degree (ie. each node must have the same number of neighbors). Just choosing the  $b$  nearest neighbors (as given by the similarity metric) for each node does not suffice, because while node  $i$  might be one of  $b$  nearest neighbors to node  $j$  the reverse is not does not have to be true. As a result nodes will have different degrees. Taking the union or intersection of edges resulting from considering the nearest neighbors of each item will also result in unequal degrees.

To address this problem, we instead construct the graph that maximizes the similarity along edges while keeping the degree of each node constant. Formally, we want to find

$$\arg \max_G \sum_{ij} G_{ij} S_{ij} \quad \text{s.t.} \quad \sum_j G_{ij} = b; \quad G_{ii} = 0; \quad G_{ij} = G_{ji}$$

where  $G$  is the adjacency matrix of the graph, with  $G_{ij} = 1$  if there is an edge from  $i$  to  $j$  and  $G_{ij} = 0$  otherwise. This is an instance of the *maximum weight b-matching* problem (Papadimitriou & Steiglitz, 1998). Exact algorithms exist for solving this problem, such as the *blossom* algorithm (Edmonds, 1965), but these are impractical for large-scale applications. Consequently, we use an approximate algorithm based on max-product message passing to find a  $b$ -matching (Jebara & Shchogolev, 2006).

Given a graph on stimuli that is a  $b$ -matching, proposals for the d-MCMCP algorithm can be made in a variety of ways. The selected proposal method is held constant throughout the experiment (as is standard in MCMC and MCMCP). The most straightforward proposal method is to choose a proposal uniformly from all  $b$  neighbors, where the value of  $b$  is chosen at the experimenter's discretion. A second method is to make a geometric proposal. Here, the proposal is generated iteratively using a number of steps,  $n_{\text{geom}}$ , that is chosen from a geometric distribution with a fixed parameter. A random walk of length  $n_{\text{geom}}$  is then performed, choosing the next node uniformly from the  $b$  neighbors of the most recent one. The node at the end of the random walk is the proposal.

For all proposal methods it is prudent to allow for some small probability of choosing uniformly from all possible stimulus items to allow the algorithm to move between local maxima.

## Experiment 1: *happy* and *sad* faces

As a first test of the d-MCMCP method, we examined the categories of *happy* and *sad* faces using a database of images of real faces. Previous work had applied the original MCMCP method to estimating these categories using parameterized face stimuli, where a continuous space was derived from eigenfaces computed from the same set of images (Martin et al., 2012). Use of the same image database allows direct comparison of the results of d-MCMCP and MCMCP with a matched stimulus set, and ratings of the emotional content of the resulting faces provide a way to evaluate these results.

### Method

**Participants.** A total of 60 undergraduates participated in exchange for course credit.

**Stimuli.** A database of 1271 images of faces was assembled from the California Facial Expression (CAFE) database, a collection of 1280 normalized  $40 \times 64$  pixel gray-scale portraits containing 64 individuals (Dailey, Cottrell, & Reilly, 2001), expressing approximately eight distinct "FACS-correct" emotions, which are classified according to the taxonomy of the Facial Action Coding System (Ekman & Friesen, 1978).

**Procedure.** Face images were convolved with Gabor filters at 8 scales and 5 orientations. Principal Components Analysis (PCA) was then applied to the whole set of convolved images and the Euclidean distance between the top 50 components was used as the similarity metric for defining the matrix  $S$ . Two graphs  $G$  were produced using the approximate  $b$ -matching algorithm from Jebara and Shchogolev (2006), one with  $b = 6$  and one with  $b = 16$ . This algorithm gives an approximate solution to the  $b$ -matching problem, so there was still some minor variation in the degree of individual nodes. Our empirical evaluation of the performance of the d-MCMCP procedure will thus also help to indicate whether this residual variation affects the results. There is no guarantee that a maximal  $b$ -matching is connected, so we used the largest connected component as the basis for the d-MCMCP procedure. The largest connected component contained 1216 images with  $b = 6$  and all 1271 images with  $b = 16$ .

We compared three different methods for defining the proposal distributions. For all three proposal methods, we allowed for a 10% chance of proposing a jump to a node chosen uniformly at random. The three methods for choosing the remaining proposals were the uniform random walk on the graph with  $b = 6$  (U6), the uniform random walk on the graph with  $b = 16$  (U16), and the geometric proposal with  $n_{\text{geom}} = 0.5$  on the graph with  $b = 6$  (G6).

Participants were randomly assigned to proposal-type conditions. Trials were presented on three different computers, one for each proposal type. Each participant completed trials corresponding to four d-MCMCP chains (two for *happy*

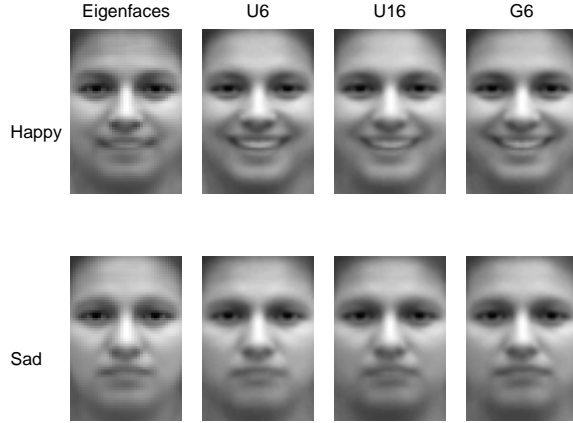


Figure 1: Results of comparing MCMCP using eigenfaces and d-MCMCP with a variety of proposal methods on the same set of face stimuli. Average faces for each type of proposal. Averages are taken across all trials and all four chains corresponding to *happy* and *sad*.

faces, two for *sad* faces). There were 100 trials for each of the four chains. On a given trial, the participant decided which of a pair of faces was either more *happy* or more *sad*. Twelve trials in the beginning were offered as practice, which were not included in the analysis. There were also 40 catch trials with face pairs for which the more *happy* or *sad* face was clearly obvious (in this case, we used the emotion designations in the CAFE database to select faces that should clearly be happy or sad). Thus each participant responded to  $100 \times 4 + 12 + 40 = 452$  trials, which took approximately 25 minutes. The responses were linked in chains of ten participants each: The last trial of each of the four chains was passed along to the next participant as his/her first non-practice trial to form a linked chain of 1000 trials. Participants who did not correctly answer at least 27 catch trials ( $p < .01$  under random guessing) were not included in the results, or added into a chain. We collected two chains of 10 participants for each proposal type, corresponding to four *happy* and *sad* chains with 1000 trials in each chain.

## Results

The images selected on each trial were averaged together to produce the average faces shown in Figure 1. All three proposal methods produced mean faces that appeared reasonably consistent with the target emotions. Also included in Figure 1 are the results reported in (Martin et al., 2012) using MCMCP in a parameterized space based on the eigenfaces derived from the image database we used for d-MCMCP. Qualitatively, the results from d-MCMCP are at least as good and perhaps better than those produced using eigenfaces.

To quantify the performance of the different variants of the algorithm, we conducted a follow-up experiment in which a group of 40 participants recruited via Amazon Mechanical

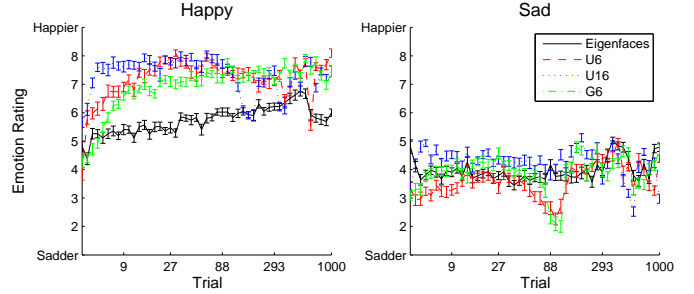


Figure 2: Happiness ratings for average faces for three types of d-MCMCP proposals as well as original MCMCP method as a function of trial number (error bars show one standard error). Averages are taken across the 50 most recent trials (or starting from the first trial for trials less than 50) and across all four chains corresponding to the same emotion (*happy* or *sad*). Also included are face ratings for the results of a previous MCMCP experiment that used eigenfaces derived from the same image database (Martin et al., 2012).

Turk provided ratings of the emotions exhibited by faces derived from our chains. For each proposal type (and for the chains based on eigenfaces used in Martin et al. (2012)), cumulative average faces were computed for each of 40 logarithmically spaced numbers of trials, averaging across all four chains that corresponded to each emotion. For trial numbers greater than 50 images were averaged only over the 50 most recent trials, meaning that no more than 200 faces contributed to any single image. Participants rated the emotion exhibited by each of these mean faces on a scale from 1-9, where 1 indicates “very sad” and 9 indicates “very happy”. All participants rated all faces, and received \$1 in compensation for their time.

The results of our follow-up experiment are shown in Figure 2. The d-MCMCP method results in statistically significantly higher ratings for faces derived from *happy* chains regardless of proposal type, perhaps as a consequence of being able to explore a larger space of faces than the eigenface method. Results for *sad* chains are more comparable. There are no systematic differences between the different proposal types, although the U16 proposal appears to produce happier faces faster than the other two proposals. For both *happy* and *sad* chains there is some variation in the emotion ratings of mean faces over time, consistent with the idea that MCMCP should be exploring the distribution of faces associated with the category (and possibly moving between modes of that distribution) rather than finding the most extreme instance of that category.

## Experiment 2: Seasonal images

Our first experiment indicated that d-MCMCP produced comparable or better performance to MCMCP when applied to a set of stimuli where both methods could be used. In our second experiment, we used d-MCMCP to explore categories defined on a set of stimuli for which there is no simple paramet-

ric representation. Specifically, we explored the categories of images associated with the seasons *Spring*, *Summer*, *Autumn*, and *Winter*, using 4000 images obtained from online image databases. By applying the d-MCMCP procedure to these stimuli, we can identify high probability images and compute informative aggregate statistics for each category, allowing us to answer questions such as what distribution of colors is associated with each season.

## Method

**Participants.** A total of 90 participants were recruited using Amazon Mechanical Turk. Each participant was paid \$1 for completing the 25 minute experiment.

**Stimuli.** A set of 4000 colored season-related images was assembled by searching for public domain web images using the phrases “spring season”, “summer season”, “autumn season”, and “winter season” in Google Image Search and on Flickr.com. The top 500 results for searches on Google and Flickr for each season were downloaded using Bulk Image Downloader. All images were resized so that the maximum dimension was 250 pixels, while preserving the original ratio of image height to width.

**Procedure.** The similarity between all possible image pairs (7998000 pairs for 4000 images) was quantified using both the Basic Color Histogram (BCH) descriptor (Griffin, 2006) and the Scale-Invariant Feature Transform (SIFT; Lowe, 1999). BCH classifies and counts pixels as belonging to one or other of the eleven basic colors (black, white, grey, red, orange, yellow, green, blue, purple, pink, and brown). SIFT applies local filters to transform images into collections of local feature vectors which are invariant to scaling, rotation and translation of the image. Similarity results over all pairs of images for both methods were normalized to have unit variance and then added together, thus yielding a similarity measure which combined results of both BCH and SIFT. The similarity between all pairs was represented as a similarity matrix which was fed into the *b*-matching algorithm. A graph was found using  $b = 5$ , which was the smallest value such that all 4000 images remained fully connected. We used a proposal distribution corresponding to a uniform random walk on this graph.

Each participant made pairwise choices between images by answering questions such as *Which image is more representative of Spring?*. There were 100 trials for each of four chains, one for each season. There were also 12 practice trials, and 40 catch trials for which one image of the pair obviously corresponded to a particular season (as judged by the experimenter). Thus each participant completed 452 trials. Participants who did not at least get 27 catch trials correct were not included in the chains or analysis. We collected data by linking three sets of 10 participants forming three chains of 1000 trials for each of the four seasons.

## Results

The top ten images that were chosen most often over all three chains for each season are shown in Figure 3. Clearly, the im-

ages are very indicative of each season. Figure 4 (a) shows, as a function of the number of trials, the L1 distance between 11-bin color histograms calculated for cumulative images, both between chains for the same season and between chains corresponding to different seasons. Within-chain distance decreases over time, and is typically lower than the similarity between chains, supporting the idea that chains are converging towards different parts of the space of images. Figure 4 (b) shows a simple example of the kind of statistical analyses that can be done on the resulting samples. The color histograms for the different seasons are quite different from one another, and each correspond to a palette that seems consistent with our intuitive representation of each season.

## Conclusion

We have presented a new method for estimating the structure of people’s mental representation of categories, showing that it produces performance that is comparable to existing methods, and can be used with rich sets of complex stimuli such as images derived from online databases. By extending the MCMCP algorithm so that it can be applied to any arbitrary set of stimuli, our d-MCMCP method makes it possible to measure people’s representations of a broader range of natural categories, and in a greater variety of real-world settings. Using our approach, MCMCP algorithms can be applied to large databases which contain discrete items, such as images or text. This has the potential to lead to significant advances for cognitive scientists interested in studying categories in a way that goes beyond simple parameterized stimuli. The results of such an investigation are likely to be valuable to machine learning and computer vision researchers interested in training systems to produce and improve on human performance in categorizing images and other complex stimuli. Conducting experiments using d-MCMCP on a large scale will allow us to build up a catalogue of human category representations, taking a step towards understanding how those categories are formed.

**Acknowledgments.** This work was supported by grant number IIS-0845410 from the National Science Foundation.

## References

- Ashby, F. G. (1992). *Multidimensional models of perception and cognition*. Hillsdale, NJ: Erlbaum.
- Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39, 216-233.
- Barker, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18, 119-133.
- Bradley, R. A. (1954). Incomplete block rank analysis: On the appropriateness of the model of a method of paired comparisons. *Biometrics*, 10, 375-390.
- Clarke, F. R. (1957). Constant-ratio rule for confusion matrices in speech communication. *The Journal of the Acoustical Society of America*, 29, 715-720.
- Dailey, M., Cottrell, G., & Reilly, J. (2001). *California facial expressions (CAFE)*. University of California, San Diego: Unpublished digital images.
- Edmonds, J. (1965). Paths, trees and flowers. *Canadian Journal of Mathematics*, 17, 449-467.

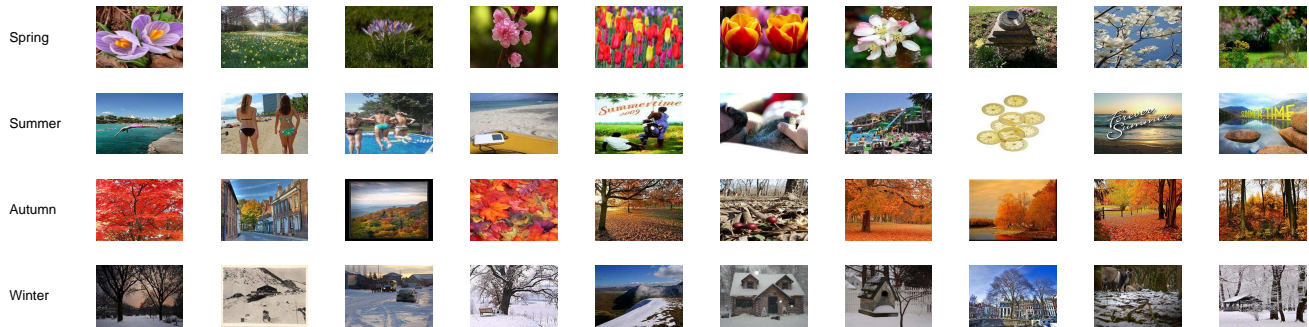


Figure 3: Top ten most popular images over all chains for each season, decreasing in popularity from left to right.

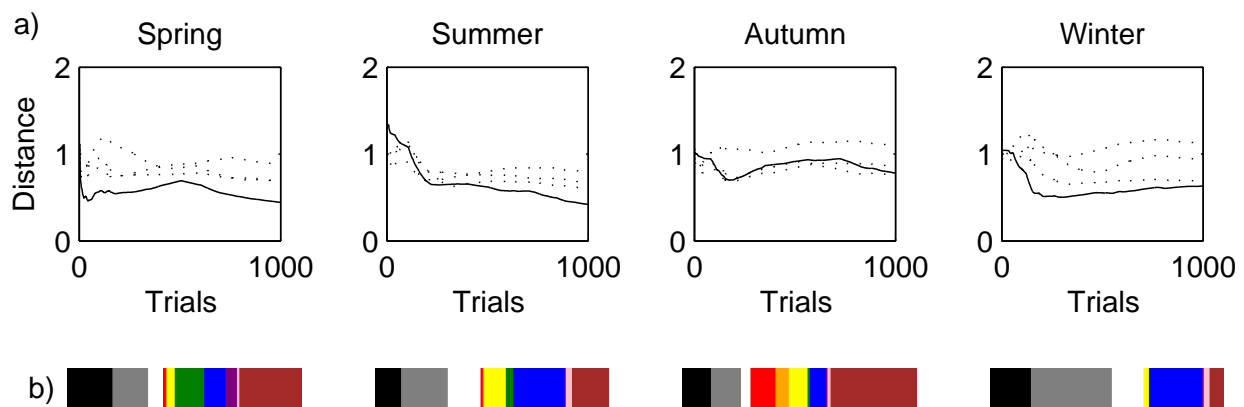


Figure 4: 11-bin color histograms were calculated for all cumulative images in all three chains as a function of the number of trials. a) average L1 distance between the cumulative histogram of a single chain and the other two chains which correspond to the same season (solid line) or the other three chains which correspond to a different season (one dotted line for each other season). b) Color histograms of all images, averaged over all chains for each season (Griffin, 2006).

- Ekman, P., & Friesen, W. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.
- Gilks, W., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Suffolk: Chapman and Hall.
- Griffin, L. D. (2006). The basic colour categories are optimal for classification. *Journal of the Royal Society Interface*, 3, 71-85.
- Hopkins, J. W. (1954). Incomplete block rank analysis: Some taste test results. *Biometrics*, 10, 391-399.
- Jebara, T., & Shchogolev, V. (2006). B-matching for spectral clustering. In *Proceedings of the European Conference on Machine Learning (ECML)*.
- Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision (ICCV)*.
- Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology, volume 1* (p. 103-190). New York and London: John Wiley and Sons, Inc.
- Martin, J. B., Griffiths, T. L., & Sanborn, A. N. (2012). Testing the efficiency of Markov Chain Monte Carlo with people using facial affect categories. *Cognitive Science*, 36, 150-162.
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.
- McDuff, D. (2010). A human-markov chain monte carlo method for investigating facial expression categorization. In *Proceedings of the 10th International Conference on Cognitive Modeling*.
- Metropolis, A. W., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087-1092.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39-57.
- Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 87-108.
- Papadimitriou, C., & Steiglitz, K. (1998). *Combinatorial Optimization: Algorithms and Complexity*. New York: Dover.
- Sanborn, A. N., & Griffiths, T. L. (2008). Markov Chain Monte Carlo with people. In *Advances in Neural Information Processing Systems 20*.
- Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. (2010). Uncovering mental representations with Markov Chain Monte Carlo. *Cognitive Psychology*, 60, 63-106.
- Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22, 325-345.

# Social Networks are Encoded in Language

**Sterling Hutchinson (schthns@memphis.edu)**

Department of Psychology / Institute for Intelligent Systems, University of Memphis  
365 Innovation Drive, Memphis, TN 38152 USA

**Vivek Datla (vvdarla@memphis.edu)**

Department of Computer Science / Institute for Intelligent Systems, University of Memphis  
365 Innovation Drive, Memphis, TN 38152 USA

**Max M. Louwerse (mlouwerse@memphis.edu)**

Department of Psychology / Institute for Intelligent Systems, University of Memphis  
365 Innovation Drive, Memphis, TN 38152 USA

## Abstract

Knowledge regarding social information is thought to be derived from many different sources, such as interviews and formal relationships. Social networks can likewise be generated from such external information. Recent work has demonstrated that statistical linguistic data can explain findings thought to be explained by external factors alone, such as perceptual relations. The current study explored whether language implicitly comprises information that allows for extracting social networks, by testing the hypothesis that individuals who are socially related together are linguistically talked about together, as well as the hypothesis that individuals who are socially related more are talked about more. In the first analysis using first-order co-occurrences of names of characters in the *Harry Potter* novels we found that an MDS solution correlated with the actual social network of characters as rated by humans. In a second study using higher-order co-occurrences, a latent semantic analysis (LSA) space was trained on all seven *Harry Potter* novels. LSA cosine values for all character pairs were obtained, marking their semantic similarity. Again, an MDS analysis comparing the LSA data with the actual social relationships yielded a significant bidimensional regression. These results demonstrate that linguistic information indeed encodes social relationship information and show that implicit information within language can generate social networks.

**Keywords:** social relations; social networks; social cognition; statistical linguistic frequencies

## Introduction

What is the nature of social relations and how can such relations be estimated? Social media, such as Facebook, LinkedIn, and Twitter allow us to answer this question, based on individuals choosing their friends. However, when such deliberate decisions are not readily available, how can social relations be measured and social networks be plotted otherwise?

Social relations can be interpreted in three non-mutually exclusive ways (Fischer, 1982). First, they can be formal in socially recognized roles, such as teacher/student, employer/employee, or father/son. Second, they can be sentimental, as when individuals feel close to others. Finally, a relation can be defined in terms of interactions

and exchanges. The formal, sentimental, and interactive nature of the social relationship can be determined by assessing a number of factors. For example, relationships can be predicted in part by the kinship of the individuals. In families, siblings tend to be close friends. Gender especially impacts the nature of relationships such that if a member of the dyad is a female, the relationship is more likely to be successful (Kim, McHale, & Osgood, 2006; Wright & Scanlon, 1991). Environment tends to weigh heavily in terms of whether or not two individuals are likely to build a relationship together. Proximity has also long been established as a strong predictor for relationships of all varieties, with increased proximity leading to increased likelihood of interpersonal relationships (Ebbesen, Kjos, Konecni, 1976). In addition, ties between locations (e.g., commonly trekked routes) also impact social interaction (Takhetev, Gruz, & Wellman, 2011). Similarly, familiarity fosters attraction between individuals (Reis, Manianci, Caprariello, Eastwick, & Finkel, 2011; Zajonc, 1968; 2001). Further, those who share interests, attitudes, and characteristics are more likely to develop friendships. In fact, any similarity between two individuals promotes the formation of a relationship between them (Bryne, 1971), with important matters (e.g., religious views, political attitudes) given more weight (Touhey, 1972). Emotions also impact relationships. When two individuals first encounter one another, a future friendship becomes more likely if the interaction is positive, whereas a friendship is not apt to blossom if the interaction is negative (Farina, Wheeler, & Mehta, 1991). Even physical features, like smell or appearance influence the relationships we form (Li, Moallem, Paller, & Gottfried, 2007).

After social relations are formed, different factors help these relations to solidify. For instance, Berscheid, Snyder, and Omoto (1989) found that closeness was significantly related to satisfaction of established romantic relationships, as was self-disclosure (Sprecher & Henrick, 2004). Feeney and Noller (1992) argued that individual differences like attachment styles impact the duration of social relationships, as does equity (Hatfield, Traupmann, & Walster, 1978). In addition, when it comes to group relationships, predicting

outcomes becomes more complicated, with members constantly joining and leaving the group. In fact, Kariam, Wang, and Leskovec (2012) found that diffusion growth (i.e., the addition of a new members to a group due to current relationships with group members) limits group size. Such findings indicate that the structure of groups is impacted by various factors.

When these social relationships are formed, how do we plot these relationships as social networks? Social scientists typically rely on interviews. For instance, Fischer (1982) asked respondents who they would share personal information with, who they worked with, and who visited their house, etc. in order to plot personal networks in the San Francisco area. But how could such networks be represented when participants cannot be interviewed as in Fischer's study, or when participants otherwise do not voluntarily release personal information as in social media? One answer to this question might lie in language. Perhaps social networks can also be acquired from and represented implicitly through linguistic sources.

In several studies we have demonstrated that perceptual information that is readily available from the world around us is encoded in language. For instance, Louwerse, Cai, Hu, Ventura, and Jeuniaux (2006) and Louwerse and Zwaan (2009) tested whether language encodes geographical information by correlating statistical linguistic frequencies between cities with the actual physical distances between those cities. Louwerse and Zwaan (2009) further tested the hypothesis by correlating computationally generated semantic relationship cosine values with the longitude and latitude of cities in the US. Recently, Louwerse, Hutchinson, and Cai (in press) found evidence for this hypothesis with Chinese texts predicting locations in China and Arabic texts predicting locations in the Middle East. Louwerse and Benesh (in press) have further extended these findings by demonstrating that the longitude and latitude of cities in the fictional Middle Earth can be predicted using the text of the *Lord of the Rings* trilogy. The semantic associations between cities in a corpus accurately estimated the physical distance between cities, supporting the claim that language encodes geographical information. A similar reasoning can perhaps be applied to social relationships. If the physical distance between individuals is small, their semantic association might be high.

In a number of studies we have shown that perceptual and embodied relations are encoded in language (Louwerse, 2011). Perhaps social relations are also encoded in language, such that computational algorithms can extract such relations from text. In the current paper we tested two hypotheses. First, if individuals are socially related together, they are talked about together. That is, if individuals are in social proximity, they are likely to be found in textual proximity. Second, if individuals are socially related more, they are talked about more.

To test these hypotheses we used the seven fantasy fiction *Harry Potter* novels and extracted the semantic relationships

between the characters and compared them with the actual social relationships between the characters.

## The Social Network of Harry Potter

Social networks are structures that map relationships between individuals. They are complex systems that can be used to examine, predict, and measure various features embedded within a network (see Newman, 2003 for an overview). Nodes represent specific individuals with edges connecting those individuals and representing relational information.

There are several ways social networks are produced. Social networks are often generated manually whereby individuals are linked to others if they are friends, colleagues, family members, etc. Individuals are able to generate their own egocentric social networks representing those other individuals with whom they share a relationship. Of course, the individual generating the network will do so based on the existence and strength of relationships that were generated by and subject to the factors enumerated above (Scott, 1988). This is the technique employed by Muckety LLC (2012). Muckety is a news corporation that manually generates maps of relationship influence between relevant individuals in a network. They manually specify networks of influence where each node is related to numerous other nodes via specific types of relationships (e.g., friend, enemy, relative). These relationships are manually researched using a variety of sources, such as government agencies and organizations, news publications, books, organization web sites, and interviews, and are expectedly costly to produce. Although Muckety generally generates networks representing current political, financial, and educational communities they have also constructed a social network representing each of the relationships between characters and organizations from all *Harry Potter* novels (Rowling, 1998; 1999a; 1999b; 2000; 2003; 2005; 2007).

Although Muckety provided a manually generated relationship network, edge weights between nodes were not provided. We thus computed edge weights as follows. Considering that between any two individuals there exists approximately four friendship links (Backstrom, Boldi, Rosa, Uander, & Vigna, 2012), we calculated a value representing higher-order relationships four degrees away. First order relationships were assigned a value of 1, relationships separated by one friendship link (or degree of separation) were assigned a value of .5, relationships separated by two friendship links were assigned a value of .25, relationships separated by three friendship links were assigned a value of .125, and relationships separated by four friendship links were assigned a value of .0625. To illustrate, *Harry Potter – Ron Weasley* received +1 because they are directly related as friends. *Harry Potter – Percy Weasley* received +.5 because they both share a relationship with *Ron Weasley*. *Harry Potter – Igor Karkaroff* would receive +.25 because *Harry* shares a relationship with someone (e.g., *Dumbledore*) who shares a relationship with



another person (e.g. *Snape*) who directly shares a relationship with *Igor Karkaroff*. This process was repeated until four friendship links were reached.

## Computational Study

The current study investigated whether statistical linguistic information encodes social relationships by testing the hypotheses a) if individuals are socially related together, they talked about together, b) if individuals are socially related more, they are talked about more. Two computational algorithms were used to test these hypotheses. First, we relied on first-order co-occurrence frequencies of character names. Although first-order frequencies are easy to compute, they also come at a price. Due to sparsity problems, they can sometimes give a biased result (Louwerse, 2011). We therefore also used a higher-order co-occurrence algorithm, latent semantic analysis (LSA; Landauer, McNamara, Dennis & Kintsch, 2007).

The seven Harry Potter books were converted to one electronic document used for the research purposes described in this study only. The document consisted of a total of 1,277,991 words. The electronic document was then filtered and all stop words (grammatical items) and punctuation marks were removed, resulting in a final file with 517,501 words and 21,423 paragraphs.

### First-order co-occurrences

In order to determine the first-order co-occurrences of character names, we computed the co-occurrence of all combinations of 56 character names in the Harry Potter novels with name pair in a five-word window. To avoid any biases with single word and two-word names (*Harry* versus *Harry Potter*), we selected the names by which each character was most frequently called while keeping the least ambiguous (e.g., *Ron Weasley* and *Arthur Weasley* are both referred to as *Weasley*, we therefore selected the names *Ron* and *Arthur*).

Although Muckety included 263 nodes (including characters, organizations, and locations), we were only interested in character relationships. We selected 56 characters for the analysis to keep the analysis from becoming too computationally expensive, as each character was paired with every other character. Characters were included on the basis of their prominence in the Harry Potter series, i.e., obscure characters were excluded from the analysis as they shared relationships with the fewest other characters.

These 56 x 56 frequency combinations were entered in an MDS analysis using the SMACOF algorithm. The SMACOF algorithm minimizes the sum of squares of the error by optimizing the fit to the distances (as opposed to the squared distances) and is thus preferred to ALSCAL (Young, 1985). We used default criteria for SMACOF, with the maximum iterations = 100, stress convergence = .0001 and the minimum stress value = .0001. Co-occurrence frequencies converged in 10 iterations with stress = .16.

Similarly, the Muckety scores for all 56 x 56 relations

were entered in an MDS analysis, using the same parameters as for the linguistic data. The MDS converged in 25 iterations, with stress = .13.

To do justice to the 2D structure of the Muckety data, we conducted a bidimensional regression to determine the relationship between the human data and the statistical linguistic frequency data. Tobler (1964) and Friedman & Kohler (2003) introduced bidimensional regressions in order to compute the mapping of any two planes under consideration. Whereas in a unidimensional regression each data point is shifted by intercept and slope, each actual and predicted value of the dependent variable are presented by a point in space, whereby vectors represent intercept and slope.

The bidimensional regression for Muckety and co-occurrence values yielded a moderate correlation,  $r = .43$ ,  $p < .001$ ,  $n = 56$ . The moderate correlation can most likely be attributed to the relatively small size of the corpus, as this impacts first order co-occurrences most (Louwerse, 2011). To account for this sparsity problem, it is often recommended to not so much rely on first-order co-occurrences, but on higher-order co-occurrences.

### Higher-order co-occurrences

To compute the higher-order computational relationship strength values we employed Latent semantic analysis (LSA). Latent Semantic Analysis captures semantic relations by mapping initially meaningless words into a continuous high dimensional semantic space (Landauer, McNamara, Dennis & Kintsch, 2007). More specifically, a first-order process associates stimuli (words) and the contexts they occur in (paragraphs). Stimuli are paired based on their contiguity or co-occurrence. These local associations are next transformed by means of Singular Value Decomposition (SVD) into a small number of dimensions (typically 300) yielding more unified knowledge representations by removing noise.

In the current study the input was the electronic version of the Harry Potter novels, segmented into paragraphs, from which a large term-document was created. For example, if there are  $m$  terms in  $n$  paragraphs, a matrix of  $A = (f_{ij} \times G(j) \times L(i, j))_{m \times n}$  is obtained. The value of  $f_{ij}$  is a function of the integer that represents the number of times term  $i$  appears in document  $j$ ,  $L(i, j)$  is a local weighting of term  $i$  in document  $j$ , and  $G(j)$  is the global weighting for term  $j$ . Such a weighting function is used to differentially treat terms and documents to reflect knowledge that is beyond the collection of the documents. The large matrix of  $A$  has, however, much redundant information, for instance because not every word occurs in every paragraph. Singular Value Decomposition reduces this noise by decomposing the matrix  $A$  into three matrices  $A = U\Sigma V^T$ ; where  $U$  is an  $m$  by  $m$  square matrix and  $V$  is an  $n$  by  $n$  square matrix, with  $\Sigma$  being an  $m$  by  $n$  diagonal matrix with singular values on the diagonal. By removing dimensions corresponding to smaller singular values, the representation of each word is reduced as a smaller vector with each word now becomes a weighted



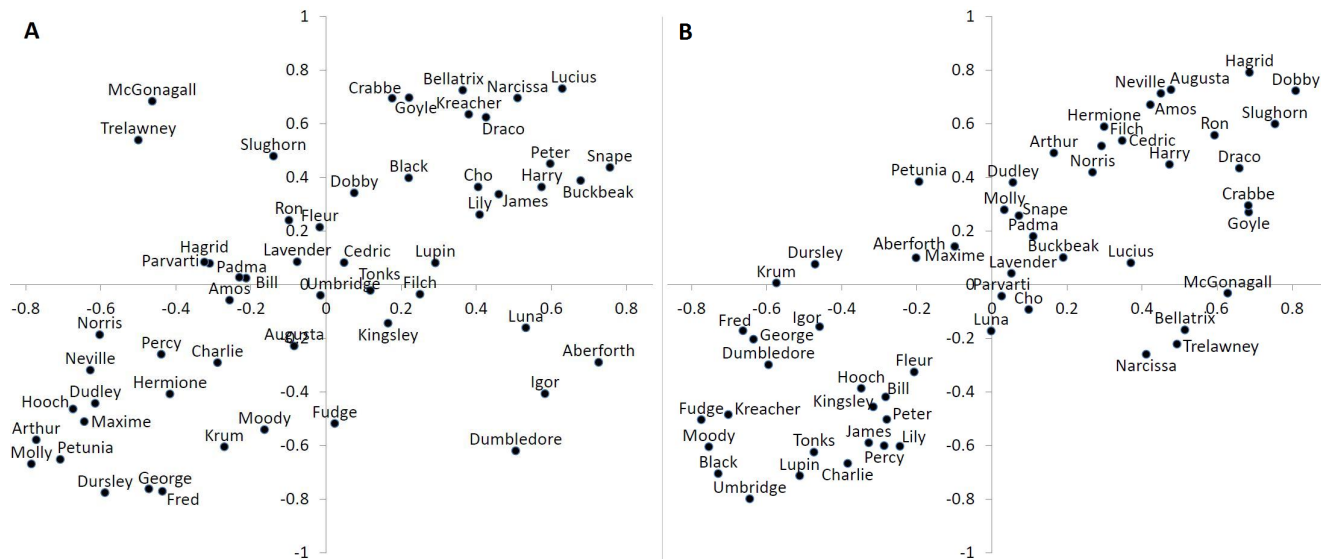


Figure 1: MDS loadings of Muckety scores (x-axis) and LSA higher-order values (y-axis) for the first dimension (Figure 1A) and second dimension (Figure 1B)

vector on 300 dimensions, with only the most important dimensions that correspond to larger singular values being kept (Landauer et al., 2007). The semantic relationship between characters can then be estimated by taking the cosine between two vectors.

As before, the 56 x 56 cosine matrix was submitted to MDS, which converged in 11 iterations with stress = .15. When we compared the two-dimensional loadings of the Muckety scores and the LSA scores in a bidimensional regression, we found, somewhat surprisingly, a weak correlation,  $r = .23$ ,  $p = .08$ ,  $n = 56$ . Yet when LSA values were allowed to populate a three dimensional configuration (stress = .07, convergence in 20 iterations), the bidimensional regression between Muckety scores and the second and third dimension of the LSA MDS yielded a more moderate (and significant) correlation,  $r = .30$ ,  $p = .02$ ,  $n = 56$ .

Upon visual inspection of the MDS plot, the first dimension did not explain social relations, but identified an outlier in the data. The character *Ginny Weasley* had more direct relationships than any of the other characters (except Harry Potter), yet the frequency with which *Ginny* occurred in the text was quite low. To illustrate, the word *Harry* occurred 21,781 times in the text whereas the word *Ginny* only occurred 762 times. After the removal of this outlier we again ran an MDS with two dimensions for both LSA (stress = .13, convergence in 12 iterations) and Muckety values (stress = .09, convergence in 33 iterations) (see Figure 1). The bidimensional regression now yielded a strong correlation between LSA values and Muckety values,  $r = .76$ ,  $p < .001$ ,  $n = 55$ .

The Muckety standard and the LSA estimates are illustrated in Figures 1A and 1B. As both figures show, the correlation between Muckety and LSA loadings are relatively strong. The correlation between Muckety and LSA values for the first dimension is represented in Figure

1A,  $r = .56$ , and the correlation between Muckety and LSA values for the second dimension is represented in Figure 1B,  $r = .76$ .

Both the first-order and the higher-order co-occurrence results demonstrate that it is possible to extract a social network from language using statistical linguistic frequencies of names of individuals.

## Number of Relationships

In this study we also test the hypothesis that characters who were socially related more are talked about more. We therefore computed the frequency of the 56 character used in the previous study. Next, we calculated how many relationships each character had in the Muckety network. For instance, *Harry Potter* shared direct relationships with 37 other characters whereas *Luna Lovegood* only shared direct relationships with 9 other characters (see Table 1).

Name frequency and number of relationships correlated strongly,  $r = .72$ ,  $p < .001$ ,  $n = 56$ , suggesting that individuals who have a large social network appear more in the text.

## Discussion

The current study aimed to determine if language encodes social relationship information. The reported results suggest computationally derived character pair values can explain human relationship strength ratings. If linguistic representations did not encode relationship structures we would have expected our computationally derived relationship strength values not to predict the human scores. They do, however, significantly predict the human Muckety relationship network scores, suggesting that social information is encoded in language, supporting the hypotheses that individuals who are socially related together are talked about together, and those who are socially related more are talked about more.

Table 1: Number of Relationships (Rel) and Frequency (Freq) of Character Names

Character Name	Rel	Freq	Character Name	Rel	Freq	Character Name	Rel	Freq
Aberforth Dumbledore	6	78	Fred Weasley	16	1075	Narcissa Malfoy	8	75
Alastor Moody	6	874	George Weasley	16	898	Neville Longbottom	10	928
Albus Dumbledore	16	3981	Ginny Weasley	19	792	Nymphadora Tonks	9	243
Amos Diggory	4	54	Goyle	3	278	Padma Patil	5	35
Argus Filch	3	335	Harry Potter	37	21781	Parvati Patil	5	168
Arthur Weasley	15	171	Hermione Granger	13	6132	Percy Weasley	7	512
Augusta Longbottom	3	2	Igor Karkaroff	3	321	Peter Pettigrew	8	163
Bellatrix Lestrangle	16	250	James Potter	10	186	Petunia Dursley	5	671
Bill Weasley	13	365	Kingsley Shacklebolt	4	119	Remus Lupin	12	841
Buckbeak	3	131	Kreacher	3	305	Ron Weasley	17	9144
Cedric Diggory	9	813	Lavender Brown	5	286	Rubeus Hagrid	12	2342
Charlie Weasley	12	165	Lily Potter	10	119	Severus Snape	15	2172
Cho Chang	9	261	Lucius Malfoy	10	148	Sirius Black	15	2314
Cornelius Fudge	2	651	Luna Lovegood	9	401	Slughorn	5	425
Dobby	6	613	Madam Hooch	3	52	Trelawney	5	284
Dolores Umbridge	5	663	Madame Maxime	3	201	Vernon Dursley	4	927
Draco Malfoy	16	1719	Mcgonagall	5	818	Viktor Krum	6	561
Dudley Dursley	4	477	Molly Weasley	15	83	Vincent Crabbe	7	268
Fleur Delacour	6	424	Mrs Norris	2	64			

We found evidence supporting both hypotheses. For the first hypothesis we used first-order co-occurrences which yielded an acceptable bidimensional regression coefficient. A higher-order co-occurrence like LSA yielded a high bidimensional regression coefficient, likely because its reduced sensitivity to sparsity problems of the linguistic data.

Even though narrative fictions offers a simulation of the social world around us (Mar & Oatley, 2008), the main conclusion of this study can of course be extended to the non-fictional world. We have already demonstrated this for geographical estimates for cities in the United States using newspapers (Louwerse & Zwaan, 2009), and geographical estimates for cities in the fictional Middle Earth using Lord of the Rings (Louwerse & Benesh, in press). We therefore expect that the method for the fictional Harry Potter novels can be extended to non-fictional texts. For instance, by using newspaper articles social relations among political leaders can be determined. By using blogs and tweets social networks of individuals in these texts can be estimated.

In addition to this application of the conclusion in this study, another important conclusion for the cognitive sciences is that language implicitly encodes information. In other work we have established this for geographical information (Louwerse & Benesh, in press; Louwerse, Hutchinson, & Cai, in press; Louwerse & Zwaan, 2009), bodily information (Tillman, Datla, Hutchinson, & Louwerse, in press) and other perceptual information (Louwerse, 2008; Louwerse & Connell, 2011). The current study shows that this can be extended to social information. Language has evolved such that statistical linguistic frequencies can capture the social relationships in the world

around us, in the fictional world, and even in the wizarding world.

## References

- Backstrom, L., Boldi, P., Rosa,., Ugander., & Vigna, S. (2012). Four degrees of separation. Retrieved January 31, 2012, from the arXiv database.
- Berscheid, E., Snyder, M., & Omoto, A. M. (1989). The relationship closeness inventory: Assessing the closeness of interpersonal relationships. *Journal of Personality and Social Psychology*, 57, 792-807.
- Byrne, D. (1971). *The attraction paradigm*. New York, NY: Academic Press.
- Ebbesen, E. B., Kjos, G. L., & Konecni, V. J. (1976). Spatial ecology: Its effects on the choice of friends and enemies. *Journal of Experimental Social Psychology*, 12, 505-518.
- Farina, A., Wheeler, D. S., & Mehta, S. (1991). The impact of an unpleasant and demeaning social interaction. *Journal of Social and Clinical Psychology*, 10, 351-371.
- Feeney, J. A., & Noller, P. (2011). Attachment style and romantic love: Relationship dissolution. *Australian Journal of Psychology*, 44, 69-74.
- Fischer, C. S. (1982). *To dwell among friends: Personal networks in town and city*. Chicago, IL: University of Chicago Press.
- Friedman, A., & Kohler, B. (2003). Bidimensional regression: A method for assessing the configural similarity of cognitive maps and other two-dimensional data. *Psychological Methods*, 8, 468-491.
- Hatfield, E., Traupmann, J., & Walster, G. W. (1978). Equity and extramarital sexuality. *Archives of Sexual Behavior*, 7, 127-141. Reprinted in M. Cook & G. Wilson (Eds.). (1979). *Love and attraction: An*

- international conference. (pp.309-323). Oxford: Pergamon Press.
- Kariam, S., Wang, D., & Leskovec, J. (2012). The life and death of online groups: Predicting group growth and longevity. *Proceedings of the ACM Conference on Web Search and Data Mining*.
- Kim, J., McHale, S., Osgood, D., & Crouter, A. (2006). Longitudinal course and family correlates of sibling relationships from childhood through adolescence. *Child Development, 77*, 1746-1761.
- Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.
- Li, W., Mollallem, I., Paller, K. A., & Gottfried, J. A. (2007). Subliminal smells can guide social preferences. *Psychological Science, 18*, 1044-1049.
- Louwerse, M. M. (2008). Embodied representations are encoded in language. *Psychonomic Bulletin and Review, 15*, 838-844.
- Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *TopiCS in Cognitive Science, 3*, 273-302.
- Louwerse, M.M. & Benesh, N. (in press). Representing spatial structure through maps and language: Lord of the Rings encodes the spatial structure of Middle Earth. *Cognitive Science*.
- Louwerse, M. M., Cai, Z., Hu, X., Ventura, M., & Jeuniaux, P. (2006). Cognitively inspired natural-language based knowledge representations: Further explorations of latent semantic analysis. *International Journal of Artificial Intelligence Tools, 15*, 1021-1039
- Louwerse, M. M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science, 35*, 381-398.
- Louwerse, M. M., Hutchinson, S., & Cai, Z. (in press). The Chinese route argument: Predicting the longitude and latitude of cities in China and the Middle East using statistical linguistic frequencies. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Louwerse, M. M. & Zwaan, R.A. (2009). Language encodes geographical information. *Cognitive Science, 33*, 51-73.
- Mar, R. A. & Oatley, K. (2008). The function of fiction is the abstraction and simulation of social experience. *Perspectives on Psychological Science, 3*, 173-192.
- Muckety LLC. (2012). *Harry Potter Series* [Graphical Interactive Relationship Influence Map]. Retrieved from <http://www.muckety.com/Harry-Potter-series/5017817.muckety>
- Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM review, 45*, 167-256.
- Reis, H. T., Maniaci, M. R., Caprariello, P. A., Eastwick, P. W., & Finkel, E. J. (2011). Familiarity does indeed promote attraction in live interaction. *Journal of Personality and Social Psychology, 101*, 557-570.
- Rowling, J. K. (1998). *Harry Potter and the sorcerer's stone*. New York, NY: Scholastic Books.
- Rowling, J. K. (1999a). *Harry Potter and the chamber of secrets*. New York, NY: Scholastic Books.
- Rowling, J. K. (1999b). *Harry Potter and the prisoner of Azkaban*. New York, NY: Scholastic Books.
- Rowling, J. K. (2000). *Harry Potter and the goblet of fire*. New York, NY: Scholastic Books.
- Rowling, J. K. (2003). *Harry Potter and the order of the phoenix*. New York, NY: Scholastic Books.
- Rowling, J. K. (2005). *Harry Potter and the half blood Prince*. New York, NY: Scholastic Books.
- Rowling, J. K. (2007). *Harry Potter and the deathly hallows*. New York, NY: Scholastic Books.
- Scott, J. (1988). Social network analysis. *Sociology, 22*, 109-127.
- Sprecher, S., & Hendrick, S.S. (2004). Self-disclosure in intimate relationships: Associations with individual and relationship characteristics over time. *Journal of Social & Clinical Psychology, 23*, 857-877
- Takhteyev, Y., Gruzd, A., & Wellman, B. (2011). Geography of twitter networks. *Social Networks*.
- Tillman, R., Datla, V., Hutchinson, S., & Louwerse, M. M. (in press). From head to toe: Embodiment through statistical linguistic frequencies. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Tobler, W. R. (1964). Bidimensional regression. *Geographical Analysis, 26*, 187-212.
- Touhey, J. C. (1972) Comparison of two dimensions of attitude similarity on heterosexual attraction. *Journal of Personality and Social Psychology, 23*, 8-10.
- Wright, P. H., & Scanlon, M. B. (1991). Gender role orientations and friendship: Some attenuation, but gender differences abound. *Sex Roles, 24*, 551-566.
- Young, F.W. (1985) Multidimensional scaling. In S. Kotz, & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, (Vol. 5, pp. 649-659). New York, NY: Wiley.
- Zajonc, R.B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science, 10*, 224-228.
- Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Applied Social Psychology, 9*, 1-27.

# Memory Indexing of Sequential Symptom Processing in Diagnostic Reasoning

**Georg Jahn (georg.jahn@uni-greifswald.de)**

University of Greifswald, Department of Psychology  
Franz-Mehring-Str. 47, D-17487 Greifswald, Germany

**Janina Braatz (janina.braatz@uni-greifswald.de)**

University of Greifswald, Department of Psychology  
Franz-Mehring-Str. 47, D-17487 Greifswald, Germany

## Abstract

Explaining symptoms by the most likely cause is a process during which hypotheses are activated and updated in memory. By letting participants learn about causes and symptoms in a spatial array, we could apply eye tracking during diagnostic reasoning to trace the activation level of hypotheses across a sequence of symptoms. Fixation proportions on former locations of possible causes reflected the causal strength of initial symptoms, a bias towards focal hypotheses, and the final diagnosis. Looking-at-nothing revealing memory activation consistent with process models of diagnostic reasoning was stable even after one week.

**Keywords:** Diagnostic reasoning, Probabilistic inference, Eye tracking, Order effects, Spatial index

## Introduction

The goal of diagnostic reasoning is to determine the most likely cause of observed symptoms. In routine cases, medical diagnosis may proceed as simple pattern classification. Often, however, symptoms are ambiguous and the clinician has to consider multiple alternative diagnoses. Then, medical diagnosis is a case of hypothesis generation and hypothesis testing (Lange, Thomas, & Davelaar, 2012; Thomas, Dougherty, Sprenger, & Harbison, 2008) as it occurs in science, criminal investigation, or searching for faults in technical systems. Diagnostic reasoning with limited information search, for example, when clinical cases are presented as case histories, requires information integration based on knowledge and multiple probabilistic cues. In the present study, we used the cover story of an accident in a chemical plant, in which workers were affected by one of four chemicals (Mehlhorn, Taatgen, Lebiere, & Krems, 2011). Participants had to decide, which chemical had caused a worker's symptoms. By laying out chemicals and symptoms in a spatial array in a learning phase, we could use eye tracking for process tracing of memory-based diagnostic reasoning to study the updating of diagnostic hypotheses.

In sequential diagnostic reasoning, the first symptom triggers a limited number of candidate hypotheses, which frame the processing of subsequent symptoms. Equally supported alternative hypotheses may be missed or rated less likely than the focal hypothesis. Similar primacy order effects have been documented for judicial decision making and social judgment, for example.

The focal hypothesis or the set of focal hypotheses is held in working memory. If symptoms have to be retained in

working memory as well, capacity limits increase in importance. Cued recall of candidate hypotheses and sequential symptom processing to update the focal hypotheses' degree of support are cognitive processes, which elude observation and are altered if ratings are elicited during reasoning. If external representations of symptoms or knowledge about causes were permanently ready for inspection, patterns of information search could be recorded via behavior records (e.g., Mouselab) or eye tracking. Here, we demonstrate a similar process tracing method (Renkewitz & Jahn, in press) that is suitable for investigating purely memory-based hypothesis generation and symptom processing. It builds on the tendency to direct the gaze to locations where information was presented before when one attempts to retrieve it from memory or reactivates it in working memory (the "looking-at-nothing" phenomenon; Richardson & Spivey, 2000).

Our participants learned about the four chemicals and the symptoms that they could cause in a spatial array. During diagnostic reasoning, symptoms were presented auditorily and the participants' eye movements on the emptied spatial array were tracked. Our goal was to trace the activation, updating, and revision of hypotheses. In particular, we were interested to see whether the activation of initial hypotheses reflected the strength of support by the first symptom, whether focal hypotheses had an advantage over equally supported alternative hypotheses, and whether fixation proportions corresponded to the final diagnosis in trials with ambiguous symptom sequences. Extending previous findings, we demonstrate the stability of looking-at-nothing over an interval of one week.

## Experiment

### Method

**Participants.** Thirty-six students of the University of Greifswald (28 female, 8 male) with a mean age of 22.1 years ( $SD = 2.4$ ) completed the first session. 32 of them returned for the second session 7.3 days later on average ( $SD = 1$ ; range 6 to 10 days).

**Materials.** To prepare for the diagnostic reasoning task, participants learned about four chemicals and possible symptoms. There were six symptom classes each containing two symptoms that are listed in Table 1. The four chemicals and the symptom classes that each could cause were presented in a 2x2 arrangement as shown in Figure 1. The square in the bottom right quadrant measured 9.1° by 9.1° of

visual angle. Symptoms from the symptom class in the top rectangle were “almost always” caused by the respective chemical, those in the middle and bottom rectangles were “occasionally” caused by the respective chemical.

Table 1: Symptom classes and symptoms. The original materials were in German.

Symptom Class	Symptom	Symptom
Eyes	Eyelid swelling	Lacrimation
Respiration	Cough	Difficult breathing
Skin	Acid burn	Rash
Neurological	Paralysis	Speech disorder
Circulatory Pr.	Sweating	Swoon
Pain	Twinge	Sting

As can be seen in Figure 1, each symptom class appeared with two chemicals. For example, “Eyes” symptoms were almost always caused by the top left chemical, but only occasionally by the top right chemical. Such symptoms are denoted “Ab” (frequent for A, occasional for B) or “Ba” (frequent for B, occasional for A) in the following. Furthermore, each chemical shared an occasional symptom class (Circulatory Problems, Pain) with a chemical in the diagonally opposite quadrant. Symptoms from these classes are denoted “ac” in the following.

A single trial in the diagnostic task consisted of a sequence of four symptoms, for example: Eyelid swelling, Cough, Swoon, and Difficult Breathing (Ab\_Ba\_ac\_Ba). Note that in this example, the third symptom (a circulatory problem) disambiguates the symptoms heard up to then and leaves only “A” (the top left chemical in this example) as the final diagnosis.

Ten different item types were constructed that are listed in Table 2. The point in the sequence at which a symptom in combination with foregoing symptoms determined the final diagnosis did vary across item types. In item type 10 the symptom pattern remained ambiguous. The column denoted “Specific symptoms” shows which item types are equivalent regarding the evidence provided by the symptoms irrespective of symptom order.

The symptom orders in Table 2 were used with each of the chemicals in the A-role. This was possible because the chemicals’ symptom patterns were symmetric. Furthermore, all possible assignments of symptoms to item types were constructed with the restriction that no single symptom was repeated in a symptom sequence.

**Procedure.** The experiment consisted of two sessions. In the first session, the participants acquired the knowledge to be used in diagnostic reasoning and then completed two phases of diagnostic reasoning trials. In the first half of diagnostic reasoning trials, the 2x2 arrangement of geometric forms was the same as during learning. In the second half, the arrangement was changed. The bottom pair became the top pair and the top pair became the bottom pair. In the second session, which took place 6 to 10 days later ( $M = 7.1$ ), the participants returned for diagnostic reasoning trials, in which the original arrangement was presented

again. Eye movements were recorded during diagnostic reasoning only.

**Knowledge acquisition.** Participants were instructed that their task would be to determine the cause of a patient’s symptoms. They were told that the patients are workers in a chemical plant, which processes four chemicals. Each patient was affected by exactly one of those chemicals. They should determine, which chemical most likely had caused a patient’s symptoms. Next, they studied Table 1 and worked through test trials until the set of twelve symptoms was once assigned to symptom classes without errors.

Then, participants were told that each chemical could cause three of the six symptom classes and the frequency with which a chemical causes a symptom class would vary. The symptom class shown in line one would be caused “almost always” and the symptom classes shown in lines two and three would be caused “occasionally”. Next, the chemicals with symptom classes were presented as shown in Figure 1. They could be studied until participants felt ready to be tested.

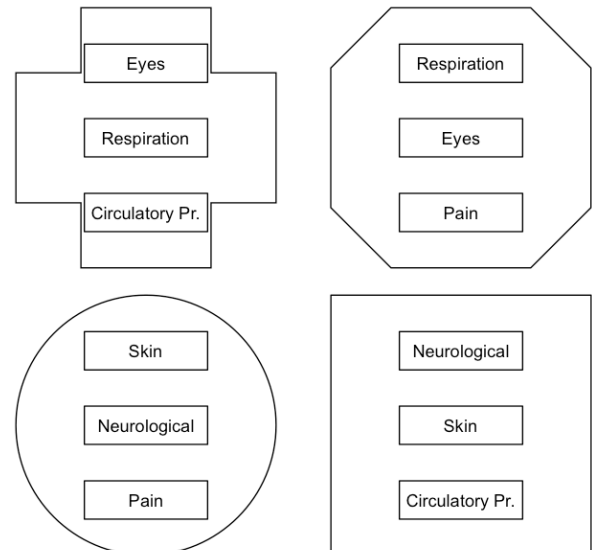


Figure 1: The four chemicals as they were presented in the learning phase. During diagnostic reasoning, the rectangular frames containing the symptom categories were empty.

In each test trial of the learning procedure, the emptied spatial array of geometric forms was shown and a symptom was presented acoustically followed by either “almost always” or “occasionally”. Participants responded by indicating the chemical that causes this symptom with this frequency. The response was given with adjacent keys on a standard keyboard (u, i, j, and k), whose arrangement approximately matched the 2x2 arrangement on the screen. Feedback was provided acoustically with a mellow or an unpleasant tone. After positive feedback, the next trial started automatically. After negative feedback, the filled arrangement was presented until participants hit the space bar to proceed to the next test trial. Testing continued until the set of 20 different testing items presented in random

order was once answered without errors. Learning was completed within 19 min on average ( $SD = 7$ ).

Before diagnostic reasoning in the second session, participants had the opportunity to refresh their knowledge by inspecting the patterns of symptom classes for the four chemicals. These were printed on separate cards within rectangular frames but without the surrounding geometrical forms. There was also one card showing the symptom classes and the single symptoms they contained.

**Diagnostic reasoning.** Each diagnostic reasoning trial started with a fixation cross in the center of the screen for 1000 ms followed by a screen showing the emptied spatial array and the acoustic presentation of the symptom sequence with delays of 3000 ms between symptoms that each lasted 1000 ms. After the fourth symptom, participants indicated their diagnosis with one of the four keys already used during learning. After the response, a confidence rating was collected, which is not reported further in this paper.

In the first session, each participant, worked twice through the 40 possible combinations of chemicals with item types: once viewing the original arrangement and once viewing the flipped arrangement. Participants returning for the second session, worked through the 40 possible

combinations again viewing the original arrangement. In addition, there were four training trials in each session.

The order of the 40 trials in each diagnostic reasoning section was pseudo-random and balanced across participants. For each trial, the actual sequence of symptoms was drawn randomly from the 8 or 4 possible sequences for this combination of item type and chemical in the A-role.

The diagnostic reasoning sections in the first session took approximately 75 min in total. Between sections participants took a rest for about 5 min. The second session took approximately 30 min.

**Apparatus.** The experimental stimuli were presented on a 19" LCD-monitor at a resolution of 800 x 600 pixels, the symptom sequences were presented through headphones. During the diagnostic reasoning phases, eye movements were monitored by a desk-mounted SMI RED eye tracker (Sensomotoric Instruments, Teltow, Germany) with a sampling rate of 60 Hz and an accuracy of approximately 0.5 degrees of visual angle. The eye tracker was calibrated before each diagnostic reasoning phase. Participants sat at a distance of approximately 60 cm to the monitor. Head movements were restrained with a chin rest.

Table 2: The ten item types, the specific symptoms that they contain, the symptom orders, and the chemicals that remain as diagnostic hypotheses after each symptom.

Item type	Specific symptoms	Order	After 1st	After 2nd	After 3rd	After 4th
1	BB	Ba_ac_Ba_ac	B,(A)	A	A	A
2	ABB	Ba_ac_Ba_Ab	B,(A)	A	A	A
3	ABB	ac_Ba_Ba_Ab	A,C	A	A	A
4	AB	ac_Ba_ac_Ab	A,C	A	A	A
5	AA	ac_ac_Ab_Ab	A,C	A,C	A	A
6	AA	Ab_Ab_ac_ac	A,(B)	A,(B)	A	A
7	ABB	Ab_Ba_ac_Ba	A,(B)	A,B	A	A
8	ABB	Ab_Ba_Ba_ac	A,(B)	A,B	A,B	A
9	ABB	Ba_Ba_Ab_ac	B,(A)	B,(A)	A,B	A
10	AABB	Ab_Ab_Ba_Ba	A,(B)	A,(B)	A,B	A,B

## Results

In all non-ambiguous item types the single correct diagnosis is denoted "A" (see Table 2). In the ambiguous item type AABB\_10, both "A" and "B" were correct diagnoses consistent with the pattern of symptoms. In every trial, the following spatial relations held between chemicals in the A-, B-, C-, and D-roles: the B-chemical was located horizontally next to the A-chemical, the C-chemical was diagonally opposite to the A-chemical, and the D-chemical was below or above the A-chemical (see Figure 1). When the layout on the screen was flipped for the second half of the first session, eight participants immediately noticed the flipped layout, eleven participants noticed the change at some point during the 40 trials, and the remaining seventeen apparently did not notice the change at all. Because of this variability, we focus on diagnostic reasoning sections with original layouts.

**Accuracy.** The mean proportion of A-diagnoses for each item type is shown in Figure 2. The item types are ordered by the combination of specific symptoms that they contain and numbered as in Table 2. Overall, accuracy was high. For the ambiguous item type AABB\_10, the mean proportion of A- or B-diagnoses was .99 and .98 in the first and second session, respectively (both  $SEs$  .01).

The five ABB item types did not differ significantly in accuracy. The two AA item types were similar in accuracy as well. Thus, we computed mean accuracy for ABB and AA for a comparison with AB and BB item types in a repeated-measures ANOVA including session and item type (AA, AB, ABB, and BB). Accuracy was higher in the second session,  $F(1, 31) = 5.49$ ,  $MSE = 0.011$ ,  $p = .03$ , the main effect of item type was significant,  $F(3, 93) = 6.43$ ,  $MSE = 0.015$ ,  $p = .002$ , and there was no significant interaction,  $F < 0.9$ . In both sessions, accuracy for AA was similar to AB, higher than ABB (Cohen's  $d = 0.64$  and  $0.77$

in sessions 1 and 2, respectively), but only slightly higher than BB ( $d = 0.34$  and  $0.27$ ).

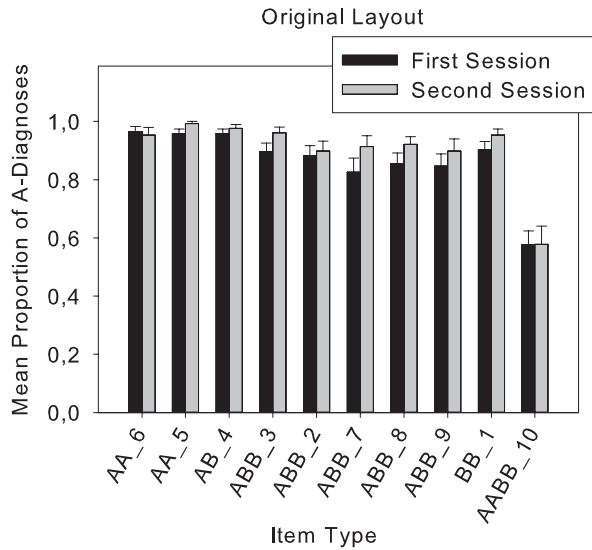


Figure 2. Mean proportions of A-diagnoses; for AABBB items B-diagnoses were correct as well

**Response times.** Response times were measured from the onset of the fourth symptom. Median response times for A-responses were computed after trimming outliers 3 *SD* above the individual session means (2.2 % of all A-responses). As shown in Figure 3, median response time was prolonged for the ambiguous item type AABBB\_10 in both sessions. In the first session, median response time was also prolonged for ABB items, in which the final diagnosis was determined late with the third symptom (ABB\_8 and ABB\_9), and for the BB item. These three item types differed significantly from all other non-ambiguous item types with  $d$ s varying between 0.31 and 1.06. In the second session, these differences between the nine non-ambiguous item types were attenuated,  $F(8,232) = 2.02$ ,  $MSE = 851180$ ,  $p = .045$ .

**Fixation proportions on quadrants.** Gaze data were analyzed for trials with correct responses, in which the original layout had been presented. Trials with more than 40% missing gaze data were discarded (2.4 % in the first session, 4.0 % in the second session). We focus on aggregated gaze data to examine the distribution of gaze allocation between screen quadrants representing diagnostic hypotheses in response to each symptom.

The quadrants of the screen were defined as areas of interest and each trial was divided in four intervals defined by the onsets of the four symptoms and the response after the fourth symptom. For each interval in a trial, the proportions of total fixation time that fell upon the four quadrants were computed and coded as A, B, C, or D according to a quadrant's chemical's role in the respective trial. Means of these fixation proportions were computed separately by session and item type for each participant. For the ambiguous AABBB item type, mean fixation proportions were computed separately for trials with A- and B-responses.

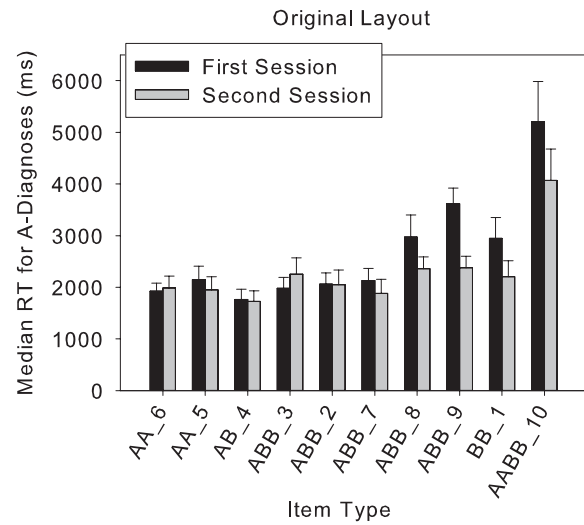


Figure 3. Median response times for A-diagnoses

Plots of mean fixation proportions across the four intervals of a symptom sequence are shown for three exemplary non-ambiguous item types in Figure 4. The three non-ambiguous item types are BB\_1 starting with a Ba-symptom, ABB\_3 starting with an ac-symptom, and ABB\_8 starting with an Ab-symptom. In addition, plots for item type AABBB\_10 separated for trials with A-responses and trials with B-responses are shown.

Apparent from a brief inspection, the item types induced very different fixation patterns, which were replicated for each item type with only small deviations in the second session one week later. In the interval from the onset of the first symptom to the onset of the second symptom, mean fixation proportions reflected how much the first symptom supported the individual hypotheses. With Ba as the first symptom, the largest proportion (nearly 40%) of fixations in the first interval was directed to the B-quadrant, a smaller proportion (about 20%) to the A-quadrant, and only about 10 % to the C- and D-quadrants, respectively. With Ab as the first symptom, the analogous pattern was observed except for the ambiguous AABBB items that were answered with B finally. With ac as the first symptom, both A- and C-quadrants were fixated for a similar proportion (nearly 30%) and longer than the B- and D-quadrants.

As soon as an ac-symptom had occurred with either an Ab- or a Ba-symptom, the diagnosis A was determined. In the following intervals, fixation proportions for all other hypotheses dropped sharply and remained low (third and fourth intervals for BB\_1 and ABB\_3). Thus, later symptoms triggered only fixations to the A-quadrant, but not to quadrants with which they were associated as well and which were fixated in the first interval for the respective symptom. For example, the B-quadrant received a large proportion after Ba as the first symptom but only a small proportion after Ba as the third symptom in BB\_1. Similarly, the C-quadrant received hardly any fixations after ac as a later symptom in BB\_1 and ABB\_8 compared to after ac as the first symptom in ABB\_3.



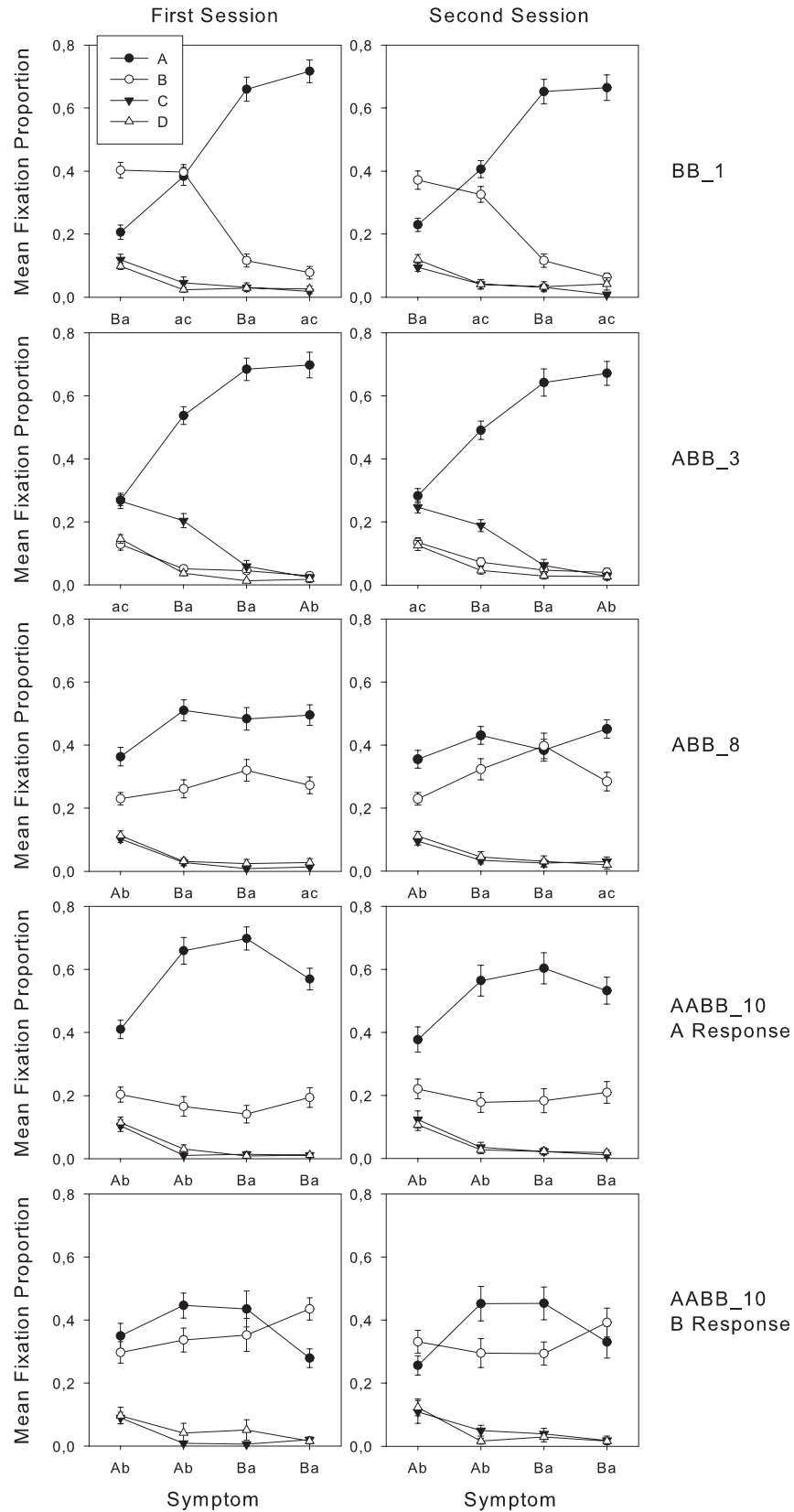


Figure 4: Mean fixation proportions on A-, B-, C, and D-quadrants in each interval of a symptom sequence. Fixation proportions for the ambiguous item type AABB\_10 are shown separately for trials answered with A and those answered with B. Error bars show the standard error of the mean.

When the diagnosis was determined late with an ac-symptom as the fourth symptom (ABB\_8) or not at all (AABB\_10), the alternative hypothesis B received considerable fixation proportions even in the fourth interval. In AABB\_10, fixation proportions were higher for B in the trials that were finally answered with B than in those finally answered with A.

Furthermore, ABB\_8, in which the diagnosis was determined late, showed a strong influence of the first symptom. The A-hypothesis that was supported more strongly by the first symptom and remained consistent with the following symptoms drew a larger fixation proportion than the alternative B-hypothesis in the second interval although the alternative was equally supported. For ABB\_8 in the first session, the A quadrant received a larger proportion even in the third interval although B was more strongly supported at this time. For the ambiguous item type AABB\_10, in which support was equal for A and B, fixation proportions were also influenced by the first symptom but finally rose for B above A when responded with B.

Overall, fixation proportions were similar in both sessions with a tendency towards fixation proportions better reflecting support for alternative hypotheses in the second session.

## Discussion

Process models of diagnostic reasoning postulate selective and changing activation of hypotheses in working memory during sequential symptom processing (Mehlhorn et al., 2011). To observe correlates of these memory dynamics, we have assigned the hypotheses to spatial locations and applied eye tracking for process tracing. Fixation proportions were influenced by memory activation because the possible causes (candidate hypotheses) had been spatially indexed during learning. Thus, process tracing was possible despite purely memory-based diagnostic reasoning. Extending previous successful applications of memory indexing (Renkewitz & Jahn, in press) and previous studies of looking-at-nothing (Richardson & Spivey, 2000), we found surprisingly stable patterns of fixation proportions one week after learning. Associations with locations and geometric figures in long-term memory strongly influenced gaze behavior similar to knowledge triggering fixations in the visual world paradigm.

The first symptom triggered fixations to its possible causes that are set up as focal hypotheses in working memory according to process models. The participants had to remember the symptom class that the first symptom belonged to and which chemicals could cause this symptom class. The result that fixation proportions were higher for more strongly supported hypotheses suggests that relative status as a focal hypothesis and the according activation level in working memory directed fixations. Hence, not just retrieval from long-term memory, but also rehearsal in working memory seems to trigger looking-at-nothing.

Excluded hypotheses did receive hardly any fixations in the subsequent intervals. Thus, fixations were not involuntarily directed towards any location associated with presented symptoms. Instead, symptoms supported the remaining focal hypothesis, and its location was fixated.

The unique information to be gained by memory indexing is clearly shown, for example, in the symptom sequence that starts with a symptom supporting strongly a hypothesis that is not the correct final diagnosis (BB\_1). The time course of fixation proportions reveals the change of the initial focal hypothesis that, of course, left no trace in the final response. And for the ambiguous item, for which the final response varied, memory indexing reveals that the finally chosen hypothesis is reflected in the relative weighting of focal hypotheses right from the beginning of the ambiguous symptom sequence.

Possibly, gaze was not only a correlate of memory activation, but also actively used as a deictic pointer to support or relieve working memory. In this study, gaze as deictic pointer was particularly useful because the spatial array of hypotheses matched the arrangement of response keys. Consequently, fixation proportions may reflect both memory activation and intended memory retention. Nonetheless, memory indexing revealed the current status of hypotheses in diagnostic reasoning, which proves this method as a valuable tool for informing and testing process models of information integration in reasoning and decision making.

## Acknowledgments

This research was supported by German Research Foundation (DFG) Grant JA 1761/7-1. We thank Agnes Scholz and Markus Krüger for helpful comments.

## References

- Lange, N. D., Thomas, R. P., & Davelaar, E. J. (2012). Data acquisition dynamics and hypothesis generation. In N. Rußwinkel, U. Drewitz, & H. van Rijn (Eds.), *Proceedings of the 11th International Conference on Cognitive Modeling* (pp. 31-36). Berlin: Universitätsverlag der TU Berlin.
- Mehlhorn, K., Taatgen, N. A., Lebiere, C., & Krems, J. F. (2011). Memory activation and the availability of explanations in sequential diagnostic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1391-1411.
- Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, 115(1), 155-185.
- Renkewitz, F. & Jahn, G. (in press). Memory Indexing: A novel method for tracing memory processes in complex cognitive tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Richardson, D. C. & Spivey, M. J. (2000). Representation, space and hollywood squares: Looking at things that aren't there anymore. *Cognition*, 76, 269-295.

# Gestures Alter Thinking About Time

**Azadeh Jamalian** (aj2334@columbia.edu)

Teachers College, Columbia University, 525 W. 120<sup>th</sup> Street  
New York, NY 10027 USA

**Barbara Tversky** (btversky@stanford.edu)

Teachers College, Columbia University, 525 W. 120<sup>th</sup> Street  
New York, NY 10027 USA

## Abstract

Can gestures alter thought? Thinking about time is deeply interlinked with actions in space, and gestures are abstracted actions. Four experiments showed that gestures alter thinking about time. Participants heard the same speech accompanied by different gestures. The viewed gestures biased listeners toward circular or linear thinking, toward parallel or sequential thinking, toward moving-ego or moving-time perspectives. Gestures can abstract and show mental models more directly and succinctly than speech.

**Keywords:** Gesture, space, time, metaphor, diagram

## Introduction

As they say, life is just one thing after another. But there is more complexity to thinking about events in time. Historical and autobiographical events are often regarded as on a timeline, but events can also happen simultaneously, not a simple single sequence. Repeating events like seasons, days, and the cell cycle can be regarded as circular. Moreover, reasoning about events in time entails taking a perspective on the timeline. Two common perspectives are *moving-ego*, thinking of yourself as moving along a timeline (we're approaching summer), or *moving-time*, thinking of yourself as stationary on a timeline with events moving past you (summer is approaching) (e. g., Clark, 1973). These perspectives are analogous to a route or intrinsic or egocentric perspective in space; the viewpoint is embedded in space or in time, with ego as the reference (e.g., Levinson, 1996; Tversky, 1996). But just as it is possible to take an external or survey or absolute perspective on space, it is possible to take an external or absolute or calendar view on time, an outside perspective regarding events as ordered by dates. In the case of survey/absolute spatial perspective, the reference points are landmarks and the terms of reference are typically north-south-east-west. For external/absolute/calendar temporal perspective, the reference points are dates or events, and the terms of reference are earlier/later.

Whatever the perspective, how people think about events in time is highly interlinked to actions in space (Talmy, 2000; Tversky, 2011). The strong association between action, space, and time is reflected in the language people use when talking about time, the diagrams they draw when conveying events in time, and the gestures that accompany narratives of events in time. People say time "marches on", we "move through" time, one event occurs "before"

another, "time has passed", and "the future is ahead of us" (e. g., Clark, 1973; Evans, 2003; Lakoff and Johnson, 1999, Moore, 2006; Nunez, 1999). People's diagrams of events in time, such as the meals of a day, are typically ordered in reading order on a horizontal line (Tversky, Kugelmass, & Winter, 1991). When relating events in time, English speakers often move their hands from left to right, event by event (e.g. Cienki, 1998); they point frontwards for the future and backwards for the past (e.g. Cooperrider & Nunez, 2009). Language, diagrams, and gestures are ways of externalizing thought, and are congruent with thinking (Tversky, 2011).

If people use actions in space to express their conceptions of events in time, will seeing different forms of actions in space change their understanding of time? We address this question here, by explaining temporal events with identical language but different gestures.

Speakers everywhere gesture while they speak. Most gestures are redundant with the speech they accompany (McNeill, 1992), but gestures sometimes express information that is not expressed in speech (e. g., McNeil, 1992; Church & Goldin-Meadow, 1986; Goldin-Meadow, Alibali, & Church, 1993; Perry, Church, & Goldin-Meadow, 1988). Although some have questioned the communicative significance of gestures (Krauss, 1998; Rauscher, Krauss, Chen, 1996; Rimè & Shiaratura, 1991), there is good evidence that speakers often intend their gestures to be communicative (e. g., Cohen, 1977; Cohen & Harrison, 1973; Alibali, Heat, & Myers, 2001; Emmorey & Casey, 2001), and that gestures, whether redundant or mismatching, influence addressees' comprehension (Goldin-Meadow & Sandhofer, 1999; Thompson & Massaro, 1994).

Can the unique information in gesture alter listeners' mental models of a highly abstract yet familiar concept? In a series of studies on reasoning about time, we demonstrate that gestures affect addressee's conceptions of time by keeping speech constant but altering gestures.

## 1: Circular vs. Linear Thinking: Diagram

Prior work (Kessell & Tversky, submitted) has shown that people are biased towards linear thinking. Participants were asked to diagram four-step *cyclical* or *sequential* processes. Most participants drew linear diagrams even for cycles. Expecting congruency between conception and visualization, Kessell and Tversky concluded that circular

thinking is harder or less natural than linear thinking. Regarding time as a cycle is difficult because it requires abstraction from a particular instance of an event (i.e. a seed to a flower) to general classes of events. Thinking of time cyclically also requires ignoring the forward progression of time to thinking of time as traveling in a circle with no beginning, middle, or ending.

Even though participants did not produce a preponderance of circular representations for cycles, they did comprehend circular diagrams (Kessell & Tversky, submitted). Could circular hand gestures prime cyclical thinking?

## Method

**Participants.** 63 (40 female, 23 male) volunteers, mostly graduate students from Columbia University, participated.

**Procedure and Design.** All participants consented verbally to participate in the study. An experimenter said to each participant: "I will tell you about some events. I'd like you to think about these events and then construct a simple schematic diagram to convey them." One-third of participants were then told twice about one of the three cycles below:

Cycles
<b>Seed to flower:</b> <ul style="list-style-type: none"> <li>• A seed germinates</li> <li>• A flower grows</li> <li>• The flower is pollinated</li> <li>• A new seed is formed</li> </ul>
<b>Events of a day:</b> <ul style="list-style-type: none"> <li>• Wake up</li> <li>• Go to work</li> <li>• Come home</li> <li>• Go to sleep</li> </ul>
<b>Clothing Cycle:</b> <ul style="list-style-type: none"> <li>• Take clothes out</li> <li>• Wear clothes</li> <li>• Wash clothes</li> <li>• Put clothes away</li> </ul>

Figure 1: Cyclical Stimuli

Each example was identically worded but accompanied by linear, circular, or no gestures. For the *linear* condition, the experimenter made 4 discrete slicing gestures right to left for the 4 stages in the spoken text. For the *circular* condition, the experimenter made 4 pointing gestures at 12, 9, 6, and 3 o'clock for the 4 stages in the spoken text. The right-to-left and counter-clockwise directions were from the experimenter's point of view so compatible with the subject's perspective. For the *no-gesture* group, the experimenter kept her hands in pockets.

## Results

**Coding the diagrams.** Participants' diagrams were coded blindly as either linear, or circular. In circular (or repeating) diagrams the last event was connected back to the first, but not in linear (or ending) diagrams. Two of the diagrams

from the circular-gesture condition, and 2 from the no-gesture condition were coded as "other" (see Figure 2).

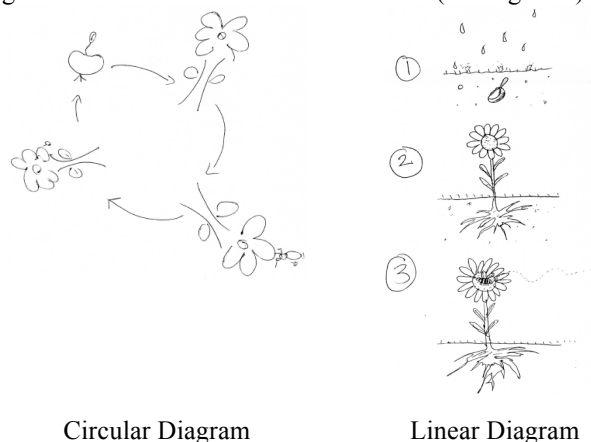


Figure 2: Examples of Diagrams

**Findings.** Of those who saw circular gestures, 66.7% drew circular diagrams. Of those who saw linear gestures, only 14.3% drew circular diagrams whereas 85.7% drew linear ones. As expected, of those who saw no gestures, 66.7% drew linear diagrams. Figure 3 shows the percent of linear, circular, and "other" types of diagrams for the three gesture conditions.

The form of gesture participants saw influenced the diagrams (excluding "other" diagrams) they drew; in a log-linear analysis, the two-way association between gesture condition and diagram type was significant,  $\chi^2(2)=17.668$ ,  $p=.000$ .

Post-hoc analyses showed significant effects of circular vs. linear gesture,  $\chi^2(1)=16.851$ ,  $p=.000$ , and circular vs. no-gesture,  $\chi^2(1)=10.556$ ,  $p=.001$ , on diagrams. No significant differences were found for linear vs. no-gesture conditions,  $\chi^2(1)=0.902$ ,  $p=.342$ . The number of circular diagrams was significantly higher than the number of linear diagrams in the circular-gesture condition,  $\chi^2(1)=4.439$ ,  $p=.035$ . As expected, the number of linear diagrams was significantly greater than the number of circular diagrams in the linear-gesture,  $\chi^2(1)=11.872$ ,  $p=.001$ , and no-gesture conditions,  $\chi^2(1)=4.439$ ,  $p=.035$ .

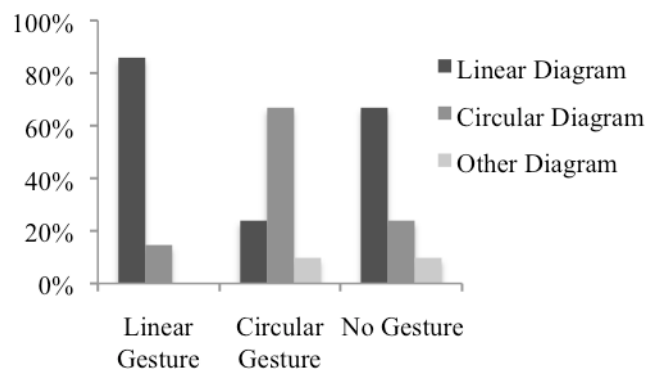


Figure 3: Proportion of linear, circular, and "other" diagrams by gesture conditions

## Discussion

Gestures had powerful effects on people's diagrams of events in time. People were asked to diagram a cyclical sequence of four events. Without gestures, a majority of participants drew linear diagrams. However, with circular gestures, a majority drew circular diagrams. If the way people diagram reflects the way they think, and there is considerable evidence for this (e. g., Tversky, 2011; Tversky, et al., 2002), then we can conclude that gestures affect the way people think about temporal events. However, it could be argued that participants copied the diagram the experimenter drew in the air. The next study obviates that objection by asking participants to make inferences.

## 2: Circular vs. Linear Thinking: Next Step

If seeing circular gestures induces cyclical thinking about time, then when participants are asked what comes after the "last" step they should tend to respond with the "first" step. This tendency should be reduced if linear gestures promote linear thought.

## Method

**Participants.** 60 volunteers, mostly graduate students from Columbia University participated in this study.

**Procedure and Design.** The procedure and design were the same as the previous experiment except that the no-gesture condition was eliminated, only the seed cycle was used, and instead of being asked to produce a diagram, participants were asked: "What comes after the new seed forms?"

## Results

**Coding.** Participants' answers to the question "what comes after?" were coded as linear or circular. Circular answers included repeating the first or any other stage and saying words such as *repeating* and *cycle*. Any other answers, such as "that was the last stage," "nothing," or "a fruit" were coded as linear.

**Findings.** In the circular gesture condition, 90% responded with circular answers, but in the linear gesture condition, only 60% responded circularly (Figure 4). In a log-linear analysis, the two-way association between gesture condition and answer type was significant,  $\chi^2(1) = 7.595$ ,  $p = .006$ . Interestingly, 30% of those who answered circularly in the linear gesture condition seemed unsure about their answers as they answered with a question tone.

## Discussion

The previous experiment had shown effects of gesture form on diagram form. Here, we found effects of gesture on inferences. When asked "what comes after?" once hearing the last of four stages of a cycle, participants who saw circular gestures were far more likely to respond with the first or subsequent step of the cycle than those who saw linear gestures. Will gesture affect other kinds of thinking about time?

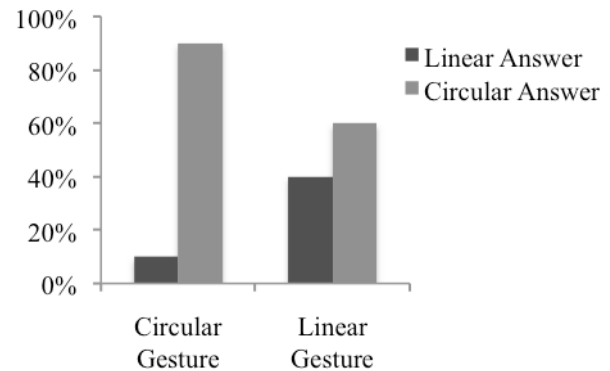


Figure 4: Proportion of linear and circular answers for each gesture condition

## 3: Perspective: Moving Ego/Time

The first two experiments showed that circular gestures promoted cyclical conceptions of time. The next experiment asks whether gestures can bias perspectives on time.

When people are asked "Next Wednesday's meeting has been moved forward two days; when is the meeting now that it has been rescheduled?" half say Friday, and half say Monday (Boroditsky, 2000; McGlone & Harding, 1998). Those answering Friday see themselves as moving through time, taking an ego-moving perspective. Those who answer Monday see themselves as stationary and time as moving past them, taking a time-moving perspective (Boroditsky, 2000; McGlone & Harding, 1998; McTaggart, 1908). In a series of clever experiments, Boroditsky & Ramscar (2002) showed that although people have strong intuitions about which answer is correct, their answers change dramatically depending on how recently they have moved or seen movement in space. For example, people who have just landed at an airport are more likely to take an ego-moving perspective than those waiting to meet passengers. People sitting still but watching things move are more likely to take a time-moving perspective. Will seeing actions in space, notably gesture, have similar effects on temporal perspective taking?

## Method

**Participants.** 40 volunteers (25 female, 15 male), mostly graduate students from Columbia University participated in this study.

**Procedure and Design.** All participants consented verbally to participate in the study. While standing side by side, an experimenter told each participant: "Next Wednesday's meeting has been moved forward two days. What day is the meeting, now that it has been rescheduled?"

Participants were divided into two conditions: (1) forward sagittal gesture, and (2) backward sagittal gesture. In both conditions, the experimenter made a slice in the space in front of her body, with her palm facing her, while saying "next Wednesday's meeting", and then moved her hands away from her body for the *forward-gesture*, and towards

her for *backward-gesture* condition while saying “has been moved forward”. Note that participants and experimenter had identical points of view.

## Results

The majority of participants who saw the forward gesture answered that the meeting was moved to Friday whereas the majority who saw the backward gesture answered that the meeting was moved to Monday (Figure 5). One participant answered “not sure” and another, “Based on your gesture I’d say Friday, but based on your words, Monday”; these were coded as “other” and not included in the statistical analysis. In a log-linear analysis, the two-way association between condition (forward versus backward sagittal gesture), and answer type (Friday versus Monday) was significant,  $\chi^2(1)=21.510$ ,  $p=.000$ .

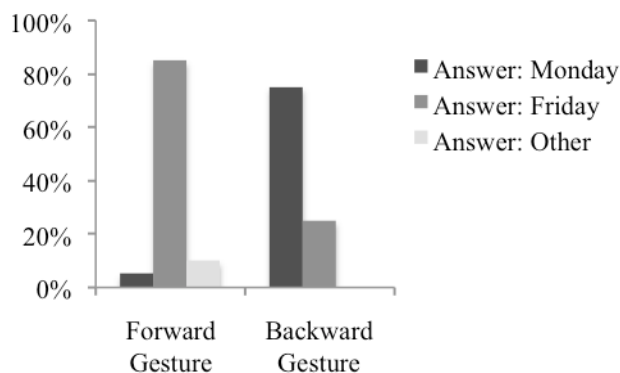


Figure 5: Proportion of participants answering “Friday” and “Monday” in each gesture conditions

## Discussion

When people are told that Wednesday’s meeting was moved forward two days and asked when the meeting is now, half spontaneously take an *ego-moving perspective*, answering Friday, and half take a *time-moving perspective*, answering Monday (e. g., Boroditsky, 2000; McGlone & Harding, 1998; McTaggart, 1908). Actually moving in space biases respondents toward the ego-moving perspective and watching movement from a stationary position biases the time-moving perspective (Boroditsky & Ramscar, 2002). Here, we found that observing representational actions, namely, gestures, also dramatically affected temporal perspective-taking. The experimenter first established a reference point for Wednesday in front of her body. When she gestured in a frontwards direction away from her body, a majority of participants responded that the meeting was moved to Friday, taking an ego-moving perspective, and when she moved her hand in a backwards direction towards her body, a majority of participants responded that the meeting was moved to Monday, taking a time-moving perspective.

Notably, the gestures were along the sagittal front-to-back axis of the body. For English speakers, the ego is the reference point, with future in the front of ego and the past behind (Cooperrider & Nunez, 2009). Here, the

experimenter made a new reference point by placing her hand in front of her body saying “Next Wednesday’s meeting,” so the reference is Wednesday rather than the ego. Then, the experimenter moved her hand along the axis either to the front of the reference point or to the back of the reference point with respect to the body. There is another possible account for the effects of the gestures, and some participants could have adopted one, some the other. Participants could have taken an external or calendar perspective, with the sagittal axis as a timeline and Wednesday as the reference point, with later events away from the body and the earlier events closer to the body. Either way, the gestures disambiguated the language and determined participants’ responses.

## 4: Parallel vs. Sequential Thinking

So far our experiments have shown that gestures alter the way people think of a sequence of events in time. Yet in life, people often have to keep track of events that happen simultaneously, a task that can be difficult (e. g., Bauer & Johnson-Laird, 1993). In one study, students had difficulties comprehending that the two middle steps of a four-step procedure for writing a paper were simultaneous. A diagram showing the simultaneous events side-by-side helped (Glenberg & Langston, 1992). Like diagrams, and in contrast to serial language, gestures can organize things in space and show simultaneity (e. g., Tversky, Heiser, Lee, & Daniel, 2009). Might gestures help people think about parallel events in time?

## Method

**Participants.** 60 volunteers, mostly graduate students from Columbia University participated in this study.

**Procedure and Design.** After receiving verbal consent for participation, an experimenter said to each participant: “I will tell you about a procedure, and then ask you a quick question about it”. Participants were then told the following procedure for writing a paper (based on Glenberg & Langston, 1992): “There are four steps to be taken when writing a paper. The first step is to write a first draft. The next two steps should be taken at the same time: One of the steps is to consider the structure; the other step is to address the audience. The final step is to proofread the paper.”

Participants were divided into two conditions: (1) parallel-gesture, and (2) sequential gesture. For the *parallel-gesture condition*, the experimenter made a slice in the air in front of her face, with her right hand palm facing down, while saying “the first step is to write a first draft”. Next, she made two slices with two hands simultaneously below her first hand gesture, while saying “the next two steps should be taken at the same time”. Next, she moved her right hand back and forth from her wrist, in place, with her left hand still in the air, while saying, “one of the steps is to consider the structure”. Then, she reversed those hand actions while saying, “the other step is to address the audience”. Next, she took away her left hand and made a slice with her right hand facing down, below its previous spot, while saying, “the



final step is proof read the paper.” For the *sequential-gesture* condition, the experimenter made 4 slices with her right hand facing down, from top to bottom on a vertical line in front of her, for the 4 steps in the procedure.

After hearing the description twice, participants were asked: “Here is the question now: According to the procedure I just gave you, what should one do immediately after writing the first draft/ before proof reading the paper?” Half of the participants in each condition were asked about steps *after* writing the first draft, and the other half were asked about steps *before* proof reading the paper.

## Results

**Coding.** Participants’ answers to before/after questions were coded as sequential, parallel, or other. Answers that mentioned only one of the two steps (*considering the structure* or *addressing the audience*) were coded as sequential. Answers that mentioned both steps were coded as parallel. Any other answer was coded as “other”.

**Data analysis.** In the *parallel-gesture* condition, 76.7% mentioned both steps while only 56.7% in the *sequential-gesture* condition gave *parallel* answer. Forty percent of participants in the sequential-gesture condition but only 10% of subjects in the *parallel-gesture* condition mentioned a single step (Figure 6). Four participants in the parallel-gesture condition, and one in the sequential-gesture condition mentioned other steps and were excluded from the data analysis.

In a log-linear analysis, the two-way association between gesture type and answer was significant,  $\chi^2(1) = 6.276$ ,  $p = .012$ . However, the two-way association between question type (before vs. after) with answer (parallel vs. sequential) was not significant,  $\chi^2(1) = 1.988$ ,  $p = .159$ , nor was its three-way association with condition (parallel- vs. sequential-gesture) and answer type,  $\chi^2(1) = 0.114$ ,  $p = .736$ .

In addition, significantly more participants in the parallel-gesture condition gave parallel answers than sequential answers,  $\chi^2(1) = 17.447$ ,  $p = .000$ . There was no significant difference between number of parallel and sequential answers in the sequential-gesture condition,  $\chi^2(1) = 0.866$ ,  $p = .35$ .

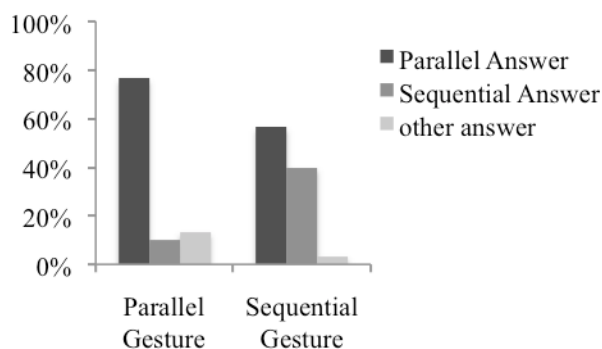


Figure 6: Proportion of parallel, sequential, and “other” answers in parallel- and sequential-gesture conditions

## Discussion

We have shown yet again that gesture influences how people think about time. Previous research (e. g., Bauer & Johnson-Laird, 1993) had shown that people find it difficult to conceptualize parallel events. Here we showed that gestures that indicate the parallel structure of events help people to reason about simultaneity of events.

## General Discussion

One way that people think of events in time is through space, as dots, representing events, on a line, representing time. The line can be regarded as straight, a linear sequence of events, perhaps, as in most narratives, having a beginning, middle, and end. For events that repeat, like the parts of the day or the seasons of the year, the line can be regarded as circular. Simultaneous events can be thought of as parallel lines. The mental time lines typically have a spatial orientation and a perspective. When straight, that line can be regarded as horizontal in reading order, vertical in top-down order (Boroditsky 2001; Tversky, et al., 1996) or sagittal from front to back (Cooperider & Nunez, 2009). Thinking and talking about time use target and reference events and a perspective on time, just like thinking and talking about space (e. g., Talmy, 2000). People can take an external perspective on the line, as in looking at a calendar or a timeline, much like taking an overview of an environment or looking at a map. Alternatively, they can see themselves embedded in time just as they can see themselves embedded in space. The ego can serve as a reference point, located on the timeline. In the ego-moving perspective, ego moves along events on the timeline; in the time-moving perspective, ego is stationary and events move past ego. Either way, changes in time are conceived of as actions in space. If changes in time are conceived of as actions in space, then actions in space might affect conceptions of time. Indeed, Boroditsky and Ramscar (2002) showed exactly that, that moving in space or watching movement in space alters temporal perspective. Here we found that information in gestures but not in speech could also alter people’s conceptions of time. Circular gestures biased thinking about a series of events as a cyclical rather than linear. Frontwards gestures away from the body biased taking an ego-moving perspective on time and gestures toward the body biased taking a time-moving perspective. Finally, gestures that traced parallel paths helped people think about simultaneous events.

Why do gestures have such powerful effects on thought? Many gestures are miniature actions in space that represent actual actions. For representing time, the gestures traced temporal paths in space, and indicated specific events along the paths. In representing paths as lines and events as dots, gestures are like diagrams (Tversky, 2011; Tversky, et al., 2009). The set of gestures both abstracts a model of time and shows it, a more direct way to communicate than purely symbolic speech.



## Acknowledgments

We are grateful to NSF HHC 0905417, IIS-0725223, IIS-0855995, and REC 0440103 for partial support.

## References

- Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, 44, 169–188.
- Bauer, M. I. and Johnson-Laird, P. N. (1993) How diagrams can improve reasoning. *Psychological Science*, 4, 372–378.
- Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition* (75:1), 1–28.
- Boroditsky, L. (2001). Does language shape thought? English and Mandarin speakers' conceptions of time. *Cognitive Psychology*, 43(1), 1–22.
- Boroditsky, L. & Ramscar, M. (2002). The roles of body and mind in abstract thought. *Psychological Science*, 13, 185–189.
- Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition*, 23, 43–71.
- Cienki, A. (1998). Metaphoric gestures and some of their relations to verbal metaphoric expressions. In Jean-Pierre Koenig (Ed.), *Discourse and cognition: Bridging the gap* (pp 189–204). Stanford: CSLI Publications.
- Cohen, A. (1977). The communicative functions of hand illustrators. *Journal of Communication*, 27, 54–63.
- Cohen, A., & Harrison, R. P. (1973). Intentionality in the use of hand illustrators in face-to-face communication situations. *Journal of Personality and Social Psychology*, 6, 341–349.
- Cooperrider, K., & Nunez, R. (2009). Across time, across the body: Transversal Temporal Gestures. *Gesture*, 9(2), 181–206.
- Clark, H. H. (1973). Time, space, semantics, and the child. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language*. Pp. 27–63. New York: Academic Press.
- Emmorey, K., & Casey, S. (2001). Gesture, thought, and spatial language. *Gesture*, 1, 35–50.
- Evans, V. (2003). *The structure of time*. Philadelphia, PA: John Benjamins.
- Glenberg, A. M. & Langston, W. E. (1992). Comprehension of illustrated text: Pictures help to build mental models. *Journal of Memory and Language*, 31, 129–151.
- Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review*, 100, 279–297.
- Goldin-Meadow, S., & Sandhofer, C. M. (1999). Gesture conveys substantive information to ordinary listeners. *Developmental Science*, 2, 67–74.
- Kelly, S. D., & Church, R. B. (1998). A comparison between children's and adults' ability to detect conceptual information conveyed through representational gestures. *Child Development*, 69, 85–93.
- Kessell, A. M. & Tversky, B. (submitted). Linear and circular thinking.
- Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science*, 7, 54–60.
- Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
- Levinson, S. (1996). Frames of reference and Molyneux's question: Cross-linguistic evidence. In P. Bloom, M. A. Peterson, L. Nadel, and M. Garrett, *Space and Language* (pp. 109–169). Cambridge, MA: MIT Press.
- McGlone, M.S., & Harding, J.L. (1998). Back (or forward?) to the future: The role of perspective in temporal language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1211–1223.
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.
- McTaggart, J. (1908). The unreality of time. *Mind*, 17, 457–474.
- Moore, K.E. (2006). Space-to-time mappings and temporal concepts. *Cognitive Linguistics*, 17(2), 199–244.
- Nunez, R.E. (1999). Could the future taste purple? Reclaiming mind, body, and cognition. In R.E. Nunez and W.J. Freeman (Eds), *Reclaiming cognition: the primacy of action, intention and emotion*. Thoverton, UK: Imprint Academic.
- Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, 7, 226–231.
- Rimè, B., & Shiaratura, L. (1991). Gesture and speech. In R. S. Feldman & B. Rimè (Eds.), *Fundamentals of nonverbal behavior* (pp. 239–281). Cambridge, UK: Cambridge University Press.
- Perry, M., Church, R. B., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development*, 3, 359–400.
- Talmy, L. (2000). *Toward a cognitive linguistics*. Cambridge: MIT Press.
- Thompson, L. A., & Massaro, D. W. (1994). Children's integration of speech and pointing gestures in comprehension. *Journal of Experimental Child Psychology*, 57, 327–354.
- Tversky, B. (1996). Spatial perspective in descriptions. In P. Bloom, M. A. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and space*. (pp. 463–491). Cambridge: MIT Press.
- Tversky, B. (2011). Visualizing thought. *Topics in Cognitive Science*, 3, 499–535.
- Tversky, B., Heiser, J., Lee, P. and Daniel, M.P. (2009). Explanations in gesture, diagram, and word. In K. R. Coventry, T. Tenbrink, & J. A. Bateman (Editors), *Spatial language and dialogue*. Oxford: Oxford University Press. Pp. 119–131.
- Tversky, B., Kugelmass, S., & Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology*, 23, 515–557.

# The Role of Task Characteristics in Children's Scalar Implicature Production

**Leen Janssens (Leen.Janssens@ppw.kuleuven.be)**

Laboratory of Experimental Psychology, Tiensestraat 102  
3000 Leuven, BELGIUM

**Walter Schaeken (Walter.Schaeken@ppw.kuleuven.be)**

Laboratory of Experimental Psychology, Tiensestraat 102  
3000 Leuven, BELGIUM

## Abstract

In two experiments, we aimed to show the importance of task characteristics in scalar implicature production. In Experiment 1, we found that five-year-olds were significantly more pragmatic when given an Action-Based Task (ABT), in which they had to respond by performing an action than in a Truth-Value Judgement Task (TVJT), in which they had to evaluate the truth-value of statements. Experiment 2 showed that seven-year-olds were almost exclusively pragmatic on the same ABT and TVJT used in Experiment 1. However, we found a 22% drop in pragmatic responses when the TVJT contained world-knowledge statements (rather than statements about simple objects such as marbles). Together, these two experiments provide evidence that not only the nature of the task, but also its specific content is crucial in determining the extent to which young children produce scalar implicatures.

**Keywords:** scalar implicatures; task characteristics; ABT; TVJT.

## Introduction

People communicate with each other to express what they feel, think, want, etc. Although this seems to happen effortlessly and automatically, the communication process is more than just the simple encoding and decoding of a message by a messenger and a receiver. Not only the literal meaning of a sentence is important, but also the implicit meaning that the speaker wants to communicate. The first systematic attempt to explain how these inferences are derived, belongs to Paul Grice. He offered a comprehensive framework of the mechanics of inferential communication (Grice, 1975). According to Grice, communication is a co-operative enterprise between people, governed by certain relational expectations about how a conversational exchange should be conducted. These relational expectations are called 'maxims' and Grice proposed four of these maxims: the Maxim of Quantity, the Maxim of Quality, the Maxim of Relation and the Maxim of Manner. These maxims respectively imply that interlocutors are always expected to offer contributions which are informative, truthful, relevant to the goals of the conversation and appropriately phrased. Grice introduced the term 'implicature', which refers to the

meaning that is implied by the speaker but not explicitly stated.

Considerable experimental research has been devoted to *scalar implicatures*, i.e. implicatures based on the existence of ordered terms on a scale of informativity (e.g., <all, most, many, some>). The general consensus is that the weaker term (e.g., the quantifier *some*), while logically compatible with a stronger term from the same scale (e.g., *all*), prompts the inference that '*all*' is not the case because the speaker did not use the stronger term. Therefore, the scalar expression '*some*' can be interpreted in two ways: either in an inference-driven, pragmatic reading, which excludes '*all*', or in its literal, semantic meaning, which is compatible with '*all*'.

Recent experimental investigations into children's interpretation of scalar terms have concluded that preschool children are often insensitive to scalar implicatures in tasks involving language comprehension (Chierchia et al., 2001; Noveck, 2001). In these studies, children, although otherwise linguistically competent, were shown to attend only to the logical/semantic meaning of the scalar terms. For example, Noveck (2001) found that 89% of the seven-to-eight-year olds in his study agreed with statements such as 'Some giraffes have long necks.' Such behavior has led Noveck (2001) to conclude that "younger, albeit competent reasoners, initially treat a relatively weak term logically before becoming aware of its pragmatic potential", and that, in this respect, "children are more logical than adults" (Noveck, 2001, p. 165).

The availability of cognitive resources is often used to explain this typically found pragmatic delay in children. As suggested by Noveck (2001), a plausible explanation for this delay is that inferring scalar implicatures requires effort and that children have less cognitive resources available than adults. Two different theories make different predictions regarding this issue. According to the neo-Gricean theorists (e.g., Levinson, 2000), implicature production happens automatically and only its inhibition demands processing costs. Relevance Theory (Sperber & Wilson, 1995) in contrast, suggests that an implicature will only be produced if it is relevant in the context and they state that this production requires additional processing costs. Evidence in favour of Relevance Theory, regarding scalar implicatures, has been presented among others by Noveck and Posada

(2003). Their experiments indicated that pragmatic answers require more time than logical answers. Assuming that longer time is associated with more processing costs, this provides indirect evidence for Relevance Theory.

In contrast to research showing that children initially reason logically, there is also substantial experimental evidence that children are not incapable of drawing scalar inferences and that they are aware of the pragmatic potential of scalar expressions. In these kinds of studies, the prime interest is to discover what conditions facilitate implicature production for children. A key factor seems to be the nature of the task. For instance, Foppolo, Guasti, and Chierchia (2004) conducted experiments concerning the quantitative scale <all, some> using two different tasks: a Truth-Value Judgement Task (TVJT) (Crain & Thornton, 1998), in which participants had to decide whether (under-informative) statements were true or false, and a Felicity Judgement Task (FJT) (Chierchia et al., 2001). In the FJT, participants were presented with a pair of utterances with the same truth-value but different levels of appropriateness and were asked to choose the most felicitous description.

When five-year-olds completed the FJT, the number of pragmatic responses was 95% while the number of pragmatic responses in the TVJT was only 50%.

Pouscoulous et al. (2007) also examined the role of the nature of the task. In their first experiment, they replicated earlier findings showing that nine-year-olds were more likely than adults to consider as true statements such as 'Some turtles are in the boxes' (uttered when all turtles are in the boxes) in a TVJT. In their second experiment, they presented an Action-Based Task (ABT), in which participants did not have to give a metalinguistic evaluation of statements but had to respond by performing an action. Children were presented with five boxes and five tokens. They were asked to adapt the situation to make it compatible with a statement. For example, if they were told "I would like all the boxes to contain a token" and two of the five boxes already contained a token, they were expected to put a token in every empty box. The results showed that, when children were asked to perform an action rather than give a metalinguistic truth evaluation, the number of implicatures made by the children increases.

In our own study we built on these experiments by Pouscoulous et al. (2007).

## Experiment 1

Our primary goal was to directly test the role played by the nature of the task in implicature production by five-year-olds. We therefore made three important changes to the Pouscoulous et al. (2007) study.

First, we presented the same group of children with both a TVJT and an ABT: manipulating the nature of the task within subjects allowed direct comparison between the two tasks.

Second, there was an important difference in content between the ABT and TVJT used by Pouscoulous et al. (2007). Whereas the ABT in Pouscoulous et al. (2007) only used tokens and boxes, in the TVJT, the children were presented with three types of animals that remained in front of them throughout the task. For each statement, they had to focus on one type of animal and ignore the other animals. Since the statements were randomly ordered, they constantly had to switch their attention between the three types, which placed greater demands on information processing than in the ABT. To remedy this problem, we made the two tasks more similar in design.

Third, we measured children's working memory (WM) capacity and compared a group of low WM-span children with a group of high WM-span children. As Pouscoulous et al. (2007) suggested, cognitive resources are important in implicature production and may explain why easier tasks, that require less cognitive resources, lead to more pragmatic answers than more difficult tasks. In adults, it has been shown that burdening WM decreases implicature production by 10% (De Neys & Schaeken, 2007). Consequently, it can be assumed that people with less cognitive resources will be less pragmatic than people with more cognitive resources. But so far, no research has been conducted on children that directly investigated the role of cognitive resources. That is why we will measure WM-capacity in the children in our experiments and investigate whether children with a high WM-capacity produce more scalar implicatures than children with a low WM-capacity.

## Method

**Participants** The sample comprised 48 five-year-olds (28 boys and 20 girls) between the ages of 5.2 and 6.1 with a mean age of 5.6 (SD=1.15), recruited from two different schools in Belgium. All were native Dutch speakers.

**Action-Based Task (ABT)** The ABT consisted of three scenarios, each involving five plastic boxes and five marbles. In the 'All-scenario', all five boxes contained a marble. In the 'None-scenario', all the boxes were empty. In the 'Subset scenario', two boxes contained a marble. In each scenario, a puppet, handled by the experimenter, was used to utter the same four requests: 'I would like all the boxes to contain a marble' ('Ik zou willen dat er in alle dozen een knikker zit'), 'I would like some boxes to contain a marble' ('Ik zou willen dat er in sommige dozen een knikker zit'), 'I would like none of the boxes to contain a marble' ('Ik zou willen dat er in geen van de dozen een knikker zit') and 'I would like some boxes not to contain a marble' ('Ik zou willen dat er in sommige dozen geen knikker zit'). This amounted to a total of 12 requests. The participants were instructed to make changes to the scenario to comply with the puppet's requests. For example, if the puppet said 'I would like all the boxes to contain a marble' in the 'Subset-scenario', the child was expected to put a marble in the three empty boxes.

There were two critical situations and 10 control statements. The first critical statement occurred in the 'All-scenario' when the puppet stated 'I would like some boxes to contain a marble'. If the child interprets 'some' logically, he or she will make no changes to the scenario. However, if the child grasps the implicature, he or she will take at least one of the marbles away. The second critical statement occurred in the 'None-scenario' when the puppet uttered the statement 'I would like some boxes not to contain a marble'. In this case, if the child interprets the statement logically, no action should be taken. A pragmatic interpretation, on the other hand, would require an action (adding at least one marble to the boxes).

For the 10 control statements, there was no distinction possible between pragmatic and logic interpretations. An example is 'I would like all the boxes to contain a marble' in the 'None-scenario'. In this case the child is expected to put a marble in all 5 empty boxes.

**Truth-Value Judgement Task (TVJT)** The children were presented with five boxes and five marbles in the three same scenarios as in the ABT. In each scenario, a puppet made the same four statements (amounting to a total of 12 sentences): 'All the marbles are in the boxes' ('Alle knikkers zitten in de dozen'), 'Some marbles are in the boxes' ('Sommige knikkers zitten in de dozen'), 'None of the marbles are in the boxes' ('Geen van de knikkers zit in een doos') and 'Some marbles are not in the boxes' ('Sommige knikkers zitten niet in de dozen'). After each statement, participants had to decide whether it was true or false. The two critical statements were '*Some marbles are in the boxes*' in the 'All-scenario' and '*Some marbles are not in the boxes*' in the 'None-scenario'. In both cases, 'true' would be the logical answer, whereas 'false' would be the pragmatic answer.

The other 10 statements were control statements (e.g. '*Some marbles are in the boxes*' in the 'Subset-scenario').

**Working Memory Tasks** The children performed three WM-tasks. First, the auditory (phonological loop) component was measured using the Digit Span Forward task in which subjects have to repeat an orally presented list of numbers. The list starts with a sequence of two numbers and keeps increasing until the child makes two errors within one block of the same digit-length. Second, the visual component (visuo-spatial sketchpad) was measured using the Corsi Block Span test. In this test, the children were presented with nine wooden blocks on which the experimenter tapped a pattern and the children were instructed to repeat the sequence. The sequence becomes longer until the child makes two errors within one block of the same difficulty level. The third WM task, which was intended to provide a 'central executive' measure, was the Digit Span Backward task. This task is identical to the Digit Span Forward, except that the subject needs to repeat the numbers in reverse order. The raw scores for each of these tasks (i.e. the total number of correct answers) were

converted into z-scores, which were then added up to compute the WM span.

**Procedure** Each participant was interviewed individually for about 20 minutes. Participants first completed the three WM tasks. The order of the other two tasks was randomized, so that half of the participants started with the TVJT and the other half with the ABT. In both tasks, the experimenter used a puppet called Minnie. In the TVJT, the children were informed that the puppet sometimes says things that are correct and sometimes says things that are wrong. In the ABT, the children were told that the puppet would give instructions regarding the boxes and the marbles and that they would either have to remove marbles, add marbles, or make no changes. Before the start of the experiment, the children were given three practice questions in the ABT. These questions were very similar to the experimental sentences but employed numbers instead of quantifiers. The three training questions were: '*I would like two boxes to contain a marble*', when only one box contained a marble, '*I would like three boxes to contain a marble*', when three boxes contained a marble and '*I would like two boxes to contain a marble*', when three boxes contained a marble. These training questions were constructed so that the participants had to remove marbles, add marbles and change nothing. This way, they got acquainted with all types of actions they would have to perform during the experiment. If the children made errors on these training questions, the experimenter corrected them and explained their mistakes.

## Results

We hypothesized that there would be differences in implicature production and performance between the TVJT and the ABT. Our hypothesis about the difference in performance was confirmed by the finding that the TVJT leads to significantly more errors than the ABT on the control statements (9% versus 5%, respectively. Wilcoxon Matched Pairs test,  $n=23$ ;  $T=57.5$ ;  $p=.011$ ).

With regard to the critical sentences, we hypothesized that the ABT would lead to more pragmatic answers than the TVJT. Again, our hypothesis was confirmed. The children responded pragmatically to the critical sentences in 91% of the instances on the ABT, compared to 70% on the TVJT (Wilcoxon Matched Pairs test,  $n=20$ ;  $T=22.5$ ;  $p=.002$ ). These results are shown in Table 1.

For both tasks, we compared a high WM-span group ( $N=16$ ;  $M=2.13$ ;  $SD=0.82$ ) with a low WM-span group ( $N=16$ ;  $M=-2.37$ ;  $SD=1.42$ ) with regard to the number of correct answers to the control sentences and the number of pragmatic responses. While there were no significant differences in pragmatic processing, the number of correct responses to the unambiguous control sentences differed significantly between the two groups. The high-span group was more accurate than the low-span group on both the ABT (98% vs 91% correct answers; Mann-Whitney U test,

Table 1: Percentage of logical responses in each scenario of the TVJT and ABT (Experiment 1).

Utterance	Task	All-scenario	None-scenario	Subset-scenario
(1) All the marbles are in the boxes.	TVJT	100%	100%	98%
(1) I would like all the boxes to contain a marble.	ABT	98%	100%	100%
(2) Some marbles are in the boxes.	TVJT	<b>23%</b>	100%	96%
(2) I would like some boxes to contain a marble.	ABT	<b>4%</b>	96%	98%
(3) None of the marbles are in the boxes.	TVJT	96%	85%	88%
(3) I would like none of the boxes to contain a marble.	ABT	98%	92%	98%
(4) Some marbles are not in the boxes.	TVJT	94%	<b>37%</b>	58%
(4) I would like some boxes not to contain a marble.	ABT	92%	<b>15%</b>	83%

Note: critical statements are in bold

$n_1=16$ ,  $n_2=16$ ;  $U=96.5$ ;  $z=-1.72$   $p=.04$ ) and the TVJT (94% vs 87% correct answers; Mann-Whitney U test,  $n_1=16$ ,  $n_2=16$ ;  $U=76$ ;  $z=-2.06$ ;  $p=.02$ ).

## Discussion

The ABT led to significantly more pragmatic answers than the TVJT. In addition, the five-year-olds made fewer mistakes on the ABT control statements than on the TVJT control statements.

These results indicate that meta-linguistic tasks are harder than tasks that require no verbal response.

Our results show that even five-year-old children are competent pragmatic reasoners. Their competence is still ‘vulnerable’, but taking into account certain factors such as task complexity, task content, context, training, etc., they are capable of producing scalar implicatures on a high level. This confirms the findings of Pouscoulous et al. (2007). Moreover, the validity of our results was enhanced by manipulating the nature of the tasks within participants and by changing the design of the TVJT to make it more comparable to the ABT. This allows us to attribute the results to the task’s cognitive demands and to conclude that the nature of the task is crucial in implicature processing in five-year-olds.

Our WM-measures revealed no significant differences in implicature processing between a group of low-span children and a group of high-span children. Although the high-span children made significantly fewer errors on the control statements, these WM-results do not allow us to draw firm conclusions about the role of WM in implicature processing.

Remarkably, the five-year-olds in our experiment produced a much higher percentage of pragmatic answers than the children tested in Pouscoulous et al. (2007). They were equally pragmatic on the ABT and more pragmatic on the TVJT than the seven-year-olds and the adults in Pouscoulous et al. (2007), who conclude that “Only 7-year-olds reveal behavior that approaches that of adults among the standard cases and even among them adultlike implicature performance is less likely when it concerns negative sentences” (Pouscoulous et al., 2007, p.371). Since we had only investigated one age-group (five-year-olds) and

since the age of seven is mostly found to be the age at which children really begin to demonstrate pragmatic skills (Guasti et al., 2005), we ran the same experiment with a group of seven-year-olds. We expected them to be even more pragmatic than the five-year-olds. In addition to the ABT and TVJT used in Experiment 1, we included a TVJT that is often used in experimental research on implicatures, i.e. the world-knowledge TVJT from Noveck (2001).

## Experiment 2

### Method

**Participants** Thirty-four seven-year-olds (18 girls, 16 boys) between the ages of 6.9 and 8.5 with a mean age of 7.5 ( $SD=.32$ ) participated in this experiment. All participants were recruited from the same school and were native Dutch speakers.

**TVJT, ABT and WM Tasks** The same TVJT, ABT and three WM tasks were used as in Experiment 1.

**World-knowledge TVJT** In order to investigate whether the specific content of the task plays a role in implicature production, the seven-year-olds conducted a task based on Noveck (2001; Experiment 3). In this task, the children were presented with 30 statements (translated into Dutch) and were instructed to indicate whether or not they agreed with each statement. The sentences were based on three types of information: factually universal (that elephants have trunks is arguably best represented with the quantifier All), factually existential (that birds live in cages is arguably best represented with Some), and absurd (that stores are made of bubbles is arguably false with both kinds of quantifiers). The statements can be categorized in six subgroups:

- (a) Five absurd *All* sentences (e.g. All birds have telephones.)
- (b) Five absurd *Some* sentences (e.g. Some fish are made of leaves.)
- (c) Five true *All* sentences (e.g. All elephants have trunks.)

- (d) Five true (and felicitous) *Some* sentences (e.g. Some flowers are yellow.)
- (e) Five false *All* sentences (e.g. All dogs have spots.)
- (f) Five true (but pragmatically infelicitous) *Some* sentences (e.g. Some giraffes have long necks.)

We were particularly interested in the sentences from category (f). If children agree with such statements they are responding logically, while disagreeing implies a pragmatic response. If we look at the different types of statements, it is clear that switching quantifiers can make (c) interchangeable with (f) as well as (d) with (e). In this way, we created two versions of this task. In each version, both the *All* and the *Some* sentences were randomized, as were the different types of statements.

**Procedure** The procedure was exactly the same as in Experiment 1. However, an additional test was administered after all other tests were performed. All children received a paper with the 30 statements included in the world-knowledge TVJT. These statements were read out to them and they were asked to indicate, for each statement, whether they agreed or disagreed by circling the appropriate answer.

## Results

The TVJT control statements led to 96% correct answers, compared to 99% for the ABT (Wilcoxon Matched Pairs test,  $n=11$ ,  $T=66$ ,  $p=.001$ ). For the control statements of the world-knowledge TVJT, the number of correct answers was 94%, which differed significantly from the ABT (Wilcoxon Matched Pairs test,  $n=3$ ,  $T=42$ ,  $p<.001$ ), though not from the other TVJT (Wilcoxon Matched Pairs test,  $n=9$ ,  $T=133$ ,  $p=.11$ ). Regarding the critical sentences, there were no significant differences between the TVJT and the ABT in the number of pragmatic answers (91% versus 94%, respectively; Wilcoxon Matched Pairs test,  $n=5$ ,  $T=22.5$ ,  $p=.48$ ). In contrast, the world-knowledge TVJT only yielded 69% pragmatic answers, which differed significantly from the other TVJT (Wilcoxon Matched Pairs test,  $n=3$ ,  $T=46.5$ ,  $p=.005$ ) and from the ABT (Wilcoxon Matched Pairs test,  $n=3$ ,  $T=34.5$ ,  $p=.003$ ). The results of the ABT and the TVJT

are shown in Table 2 whereas the results of the world-knowledge TVJT are shown in Table 3.

As in Experiment 1, we compared a group of high WM-span children ( $N=11$ ;  $M=2.32$ ;  $SD=1.07$ ) with a low-span group ( $N=11$ ;  $M=-2.38$ ;  $SD=1.06$ ). No significant differences were found between the two groups on any of the three tasks, neither in pragmatic responses, nor in performance on the unambiguous sentences.

## General Discussion

The two studies reported in this article investigated the role of the task in scalar implicature production in young children. Our goal was to show that the kind of task and even the specific task content has an important impact on scalar implicature production. In Experiment 1, we investigated five-year-old children. We found, as expected, that a more difficult TVJT caused the children to be less accurate and less pragmatic than an ABT in which children did not have to answer verbally. Given our methodological improvements, this difference was not caused by a difference in task design but by a difference in task complexity. Manipulating the nature of the task is sufficient to show that, under the right circumstances, children as young as five years are capable of spontaneously producing implicatures.

In Experiment 2, we investigated a group of seven-year-olds whom we expected to be even more pragmatic than the five-year-olds in Experiment 1. This expectation was confirmed by the results: the pragmatic response rate was so high that it did not lead to a significant difference between the ABT and the TVJT. However, when the children performed a TVJT involving world-knowledge statements, pragmatic responses dropped by 22%. For the world-knowledge TVJT, the children need to rely on the knowledge they have stored in their memory, whereas in the simple TVJT, they just have to rely on the boxes and marbles in front of them, which is less demanding on memory resources.

Table 2: Percentage of logical responses in each scenario of the TVJT and ABT (Experiment 2).

Utterance	Task	All-scenario	None-scenario	Subset-scenario
(1) All the marbles are in the boxes.	TVJT	100%	100%	100%
(1) I would like all the boxes to contain a marble.	ABT	100%	100%	100%
(2) Some marbles are in the boxes.	TVJT	<b>0%</b>	100%	94%
(2) I would like some boxes to contain a marble.	ABT	<b>3%</b>	100%	100%
(3) None of the marbles are in the boxes.	TVJT	97%	94%	97%
(3) I would like none of the boxes to contain a marble.	ABT	100%	100%	100%
(4) Some marbles are not in the boxes.	TVJT	97%	<b>18%</b>	79%
(4) I would like some boxes not to contain a marble.	ABT	100%	<b>9%</b>	91%

Note: critical statements are in bold

Table 3: Percentage of logical responses on the world-knowledge TVJT (Experiment 2).

Sentence type	Correct Response	
Utterances expressed with All		
Absurd (false) (e.g. All birds have telephones)	No	97%
Appropriate (true) (e.g. All elephants have trunks)	Yes	94%
Inappropriate (false) (e.g. All dogs have spots)	No	92%
Utterances expressed with Some		
Absurd (false) (e.g. Some fish are made of leaves)	No	100%
Appropriate (true) (e.g. Some flowers are yellow.)	Yes	82%
<b>Inappropriate (true though pragmatically infelicitous)</b> <b>(e.g. Some giraffes have long necks)</b>	<b>Yes</b>	<b>31%</b>

Note: critical statements are in bold

Another difference between the two TVJT's is that the TVJT with the marbles and the boxes is based on visual input (the marbles and the boxes) whereas the world-knowledge TVJT is not based on visual input. The hypothesis that easier tasks lead to significantly more pragmatic answers than more difficult tasks is based on the assumption that cognitive resources are critical in implicature production (De Neys & Schaeken, 2007). As easier tasks require fewer cognitive resources than complex tasks, more cognitive resources remain available for producing implicatures.

Based on these assumptions, we hypothesized that children with a high WM-capacity would be more pragmatic than children with a low WM-capacity since they have more cognitive resources available. This hypothesis was not confirmed.

Even when we performed the WM-analyses on the combined sample from both experiments (with the highest scoring children in each experiment as the 'high-group' and the lowest scoring children as the 'low-group'), we did not find a significant WM-effect. Although a certain trend can be observed in our WM-data, we are unable to find a single significant WM-effect. However, this is hardly surprising given that the significant WM-effect found in adults is small (De Neys and Schaeken, 2007), which ensures a smooth flow of communication.

In sum, the key finding of the present study is that the nature of the task and the specific task content are very important in scalar implicature production in young children: more cognitive tasks or more cognitive task content cause a decrease in implicature production. This factor has to be taken into account when investigating implicature production in children. Another factor that might need to be taken into account in future research is a

measure of general language ability. Since it was found that metalinguistic tasks are harder than action-tasks, it is plausible that general language ability may account at least partly for these results.

## Acknowledgments

We would like to thank all schools – Sint-Pietersschool Korbeek-Lo', 'Ark Leuven' and 'Gemeentelijke basisschool Pellenberg'- for allowing us to run our experiments. Leen Janssens is a research assistant from the Fund for Scientific Research Flanders.

## References

- Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: evidence for a grammatical view of scalar implicatures. In A. H.-J. Do, L. Dominguez & A. Johansen (Eds.), *Proceedings of the 25<sup>th</sup> Boston University Conference on Language Development* (pp.157-168). Somerville, MA: Cascadilla Press.
- Crain, S., & Thornton, R. (1998). *Investigations in Universal Grammar. A Guide to Experiments on the Acquisition of Syntax and Semantics*. The MIT Press, Cambridge, MA.
- De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, 54(2), 128-133.
- Foppolo, F., Guasti, M. T., & Chierchia, G. (2004). *Pragmatic inferences in children's comprehension of scalar items*. (Talk presented at Second Lisbon Meeting on Language Acquisition, Lisbon, 2004).
- Grice, H. P. (1975). *Logic and Conversation*. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics*, Vol. 3. New York: Academic Press.
- Guasti, M. T., Chierchia G., Crain S., Foppolo F., Gualmini A., & Meroni L. (2005). Why Children and Adults Sometimes (But Not Always) Compute Implicatures. *Language and Cognitive Processes* 20, 667–696.
- Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.
- Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, 78(2), 165-188.
- Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, 85(2), 203-210.
- Pouscoulous, N., Noveck, I., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs and implicature production. *Language Acquisition*, 14(4), 347-375.
- Sperber, D., & Wilson, D. (1995). *Relevance: Communication and Cognition*. Cambridge, MA: Harvard University Press.



# Learning What is Where from Social Observations

Julian Jara-Ettinger (jjara@mit.edu)

Chris L. Baker (clbaker@mit.edu)

Joshua B. Tenenbaum (jbt@mit.edu)

Department of Brain and Cognitive Sciences, MIT  
Cambridge, MA 02139

## Abstract

Observing the actions of other people allows us to learn not only about their mental states, but also about hidden aspects of a shared environmental situation – things we cannot see, but they can, and that influence their behavior in predictable ways. This paper presents a computational model of how people can learn about the world through these social inferences, supported by the same *Theory of Mind* (ToM) that enables representing and reasoning about an agent’s mental states such as beliefs, desires, and intentions. The model is an extension of the Bayesian Theory of Mind (BToM) model of Baker et al. (2011), which treats observed intentional actions as the output of an approximately rational planning process and then reasons backwards to infer the most likely inputs to the agent’s planner – in this case, the locations and states of utility sources (potential goal objects) in the environment. We conducted a large-scale experiment comparing the world-state inferences of the BToM model and those of human subjects, given observations of agents moving along various trajectories in simple spatial environments. The model quantitatively predicts subjects’ graded beliefs about possible world states with high accuracy – and substantially better than a non-mentalistic feature-based model with many more free parameters. These results show the power of social learning for acquiring surprisingly fine-grained knowledge about the world.

**Keywords:** Social Cognition; Theory of Mind; Social Learning; Reinforcement Learning

## Introduction

The most obvious way to learn about the world is by direct observation. You may believe there is a Starbucks across the street from your office because you have passed it many times, and believe it is open at this moment because you just passed by a few minutes ago and saw a number of people going in and out. But many aspects of the world are unobservable and must be inferred indirectly, often based on observing the actions of other people who know or perceive what you do not. Consider the situation of driving or biking and needing to turn left at an intersection onto a busy street, across oncoming traffic. Of course before turning you will check to see whether there are any cars coming down the busy street from the left, but suppose there is a large truck parked on the street, blocking your view so that you cannot see whether there is any oncoming traffic. You may inch out slowly until you can see, but you may also observe what other drivers or pedestrians are doing. If they are in a position to see the oncoming cars that you cannot, and if they are crossing the busy street at the same point you wish to turn, then it is a good bet that your turn would also be safe.

Making such a judgment is literally betting your life on a mental model of another person’s cognitive processes – a Theory of Mind (ToM) (e.g. Dennett, 1987; Wellman, 1990;

Gopnik & Meltzoff, 1997). Implicitly you assume basic aspects of rationality in the person you see crossing the street: that they want to cross safely, that they update their beliefs about the presence of oncoming cars based on what they can see, and that they plan their actions appropriately to achieve their goals given their beliefs. If they are clearly paying attention to the side of the street you cannot see, and they are walking across unhurriedly and unworriedly, it is then a good bet that no traffic is headed imminently toward them; if they are jumping or dashing out of the way, that is another story.

Accounts of distinctively human cognition often emphasize the sophisticated representational power of people’s ToM, as in the capacity to represent arbitrary belief states, false beliefs as well as true ones, and predict how people will act accordingly. But just as or more important is the sophisticated inferential power of ToM: how we can learn about the contents of other agents’ mental states, or even the structure of the world, by reasoning backwards to the best explanations of agents’ observed behaviors. This kind of inverse reasoning underlies not only the traffic example above, but many other situations of practical importance for everyday cognition. For example, if you see people filing out of a new restaurant with contented looks, it is a good bet the food inside is satisfying. If you see someone enter the restaurant with an expression of eager anticipation, then exit a moment later and start looking for a different place to eat, you might guess that the restaurant is unexpectedly closed – or perhaps he mistook it for a different place. If your friend the foodie goes far out of his way while visiting a new city to visit a particular restaurant, you can bet that place is one of the city’s best.

In this paper, we present a computational model of this social-learning capacity – inferring the world’s state from observing other agents’ behavior, guided by ToM. Similar inferential abilities have been studied in infants (Csibra, Biró, Koós, & Gergely, 2003) and adults (Goodman, Baker, & Tenenbaum, 2009), and the latter paper presented a computational model similar to ours in key respects (but focused on causal learning). Our work is the first to test people’s social learning against rational model predictions in a large-scale quantitative experiment, showing that people can form surprisingly accurate fine-grained beliefs about the relative probabilities of different possible worlds from sparse social observations – just a single agent moving along a single goal-directed path of intentional action. We contrast our model with a non-intentional, non-ToM account based on low-level features of the agent’s motion. Even when we introduce many free parameters in the form of variable feature weights,

and optimize their values to best fit people’s world-state inferences, the feature-based alternative performs substantially worse than a ToM-based model with many fewer parameters.

## Computational framework

A rapidly growing body of research suggests that human judgments about intentional agents’ mental states (goals, preferences, beliefs) can be modeled as probabilistic inverse planning, inverse optimal control, or inverse decision-making: Bayesian inferences over predictive models of agents’ rational behavior (Baker, Saxe, & Tenenbaum, 2009; Lucas, Griffiths, Xu, & Fawcett, 2009; Bergen, Evans, & Tenenbaum, 2010; Jern, Lucas, & Kemp, 2012; Baker, Goodman, & Tenenbaum, 2008; Ullman et al., 2010; Tauber & Steyvers, 2011). Here we adopt the Bayesian ToM (BToM) formulation of Baker, Saxe, and Tenenbaum (2011), expressing relations between the world’s state, an agent’s state, and the agent’s observations, beliefs, desires, and actions in terms of a rational-agent model known as a partially observable Markov decision process (POMDP) (Kaelbling, Littman, & Cassandra, 1998). This captures a probabilistic version of the classical rational agent who updates their beliefs to conform with their observations and chooses sequences of actions expected to achieve their desires given their beliefs. The causal schema for BToM is shown in Fig. 1(a).

Baker et al. (2011) used the BToM model to explain human observers’ joint inferences about agents’ beliefs and desires, based on how these mental states guided agents’ actions exploring a small, spatially structured world with different sources of utility (candidate goals) in different locations. Observers had full knowledge of the agent’s situation and world state, but the agent only learned about the world piecemeal (based on line-of-sight perceptual access) as it explored. In contrast, in this paper we consider scenarios where *neither* the agent nor the observer have full access to the state of the world. The agent again has line-of-sight perceptual access, but the observer sees none of the utility sources (candidate goals) in the environment; these must be inferred from observing the agent’s movements. At first blush, this inference problem might seem hopelessly underconstrained; however, we will show that when the observer knows the agent’s preferences, and if those preferences are strong enough, then joint inferences about the agent’s beliefs and the unobservable world state are possible.

To illustrate how this works, consider the scenario shown in Fig. 2. At a certain university food hall, every day at lunchtime three different food carts arrive: an Afghani (A) cart, a Burmese (B) cart, and a Colombian (C) cart. The food hall contains three rooms, West (W), North (N) and East (E), and on any given day, any cart can be in any room. Harold, the student shown in the figure, always prefers to eat at cart A over carts B and C, and prefers to eat at cart B over cart C. Furthermore, carts A and B can be *open* or *closed* when Harold arrives; he only goes to a cart if he sees that it is open. Cart C is always open and is the last resort when all others are

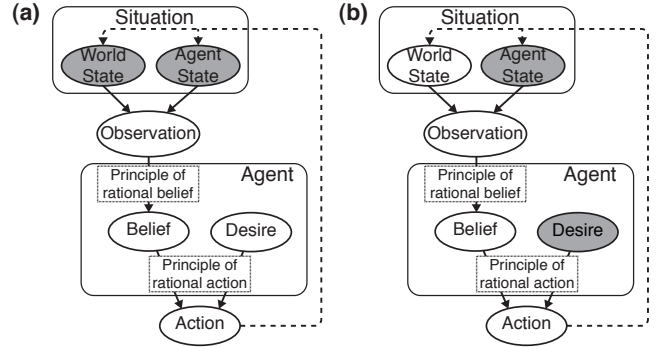


Figure 1: Causal structure of Theory of Mind. Traditional accounts of ToM (e.g., Dennett, 1987; Wellman, 1990; Gopnik & Meltzoff, 1997) have proposed informal versions of these schemata, characterizing the content and causal relations of ToM in commonsense terms, e.g., “seeing is believing” for the principle of rational belief. (a) Schematic of the Bayesian theory of mind (BToM) model proposed by Baker et al. (2011). Grey shaded nodes – World State and Agent State – are assumed to be observed (for the observer; not necessarily for the agent, as described in the main text). (b) Our extension of BToM to allow inference of hidden aspects of the World State by observing an agent’s behavior. Here, the Agent State and Desire are observed, but the World State is only partially observable for both agent and observer.

closed.

Fig. 2(a) shows a hypothetical path that Harold could take, ending in the North room. What, if anything, does this tell us about the cart locations? From where Harold enters the food hall, he can observe the cart in the North room. Next, he checks the East room, indicating that either cart A is not in the North room, or that cart A is in the North room, but is closed. When Harold returns to the North room, only one possibility remains: that he saw cart A in the East room, but it was closed, so he returned to the North room to eat at cart B, which was open (this cart configuration is shown in Fig. 2(d), row 1, column 3). Crucially, this inference also depends on Harold’s *not* checking the West room, which is consistent with several other configurations in Fig. 2(d). In our experiment, 66% of participants rated the correct configuration to be the most likely in this condition (chance = 17%).

## Informal Model Sketch

Fig. 1 sketches the causal schema for BToM. For concreteness, we will describe the content of the model in terms of our food carts scenario, but in principle the BToM framework can be defined over arbitrary state and action spaces. In our food cart examples, there are 24 possible World States: 6 possible cart configurations (shown in Fig. 2(d)) times 4 possible joint combinations of open/closed for carts A and B. There are 12 possible Agent States, one for each grid square in the food hall scenario. Agents’ Observations provide information about the World State, conditioned on the Agent State, and are based on line-of-sight visibility – Fig. 2(b,c) give examples of what can be seen from different vantage points. The observer represents an agent’s Belief as a probability distribution over possible World States. The observer maintains a

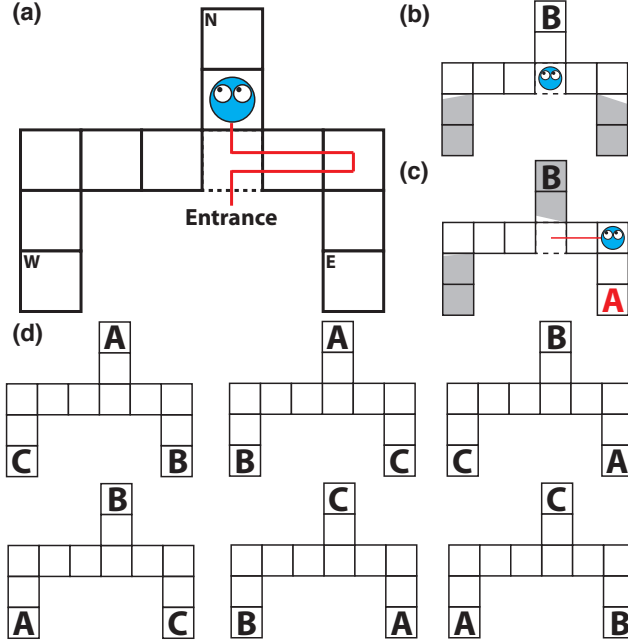


Figure 2: Example experimental stimulus. (a) Example of an observed path. The task is to figure out where each of three food carts is, given the trajectory the agent took. (b) In the agent’s initial position he can observe the North spot with food truck B in it. However, he doesn’t know where his favorite cart A is. (c) Agent’s state when he travels to the entrance of the East hallway. He can now observe cart A being closed and remembers having seen cart B in the North spot. With this information he can deduce that cart C is in the West room and so his best option is to choose the North spot, producing the path shown in (a). (d) Possible configurations the carts can take independent of them being closed or open. In the experiment, subjects ranked these six configurations for each path.

finite set of possible Beliefs the agent could hold, drawn from a prior over initial Beliefs, and simulates the agent’s Belief update for each possible Observation, given the Agent State and World State. An agent’s Desire is captured by utilities for each cart which capture the preference relation  $A \succ B \succ C$ .

Given the representational content of the nodes, BToM expresses the functional form of the causal relations in Fig. 1 in terms of POMDPs, which capture the dual principles of rational Belief and Action in Fig. 1. To generate a POMDP policy for each initial Belief point, we employ an implementation of the SARSOP algorithm (Kurniawati, Hsu, & Lee, 2008), provided by the APPL POMDP solver. These policies represent a predictive, Belief- and Desire-dependent distribution over the agent’s actions.

The schema in Fig. 1 illustrates the conditional dependencies involved in the model of the agent. For clarity, in this informal sketch we suppress the temporal nature of the model; technical details of dynamic inference are provided in Baker et al. (2011). The predictive distribution over the agent’s Action, given its Desire, Beliefs, the World State and the Agent State (abbreviating variable names as A, D, B, W, S respectively) is:

$$p(A|D, W, S) = \sum_B p(A|B, D) \sum_O p(B|O) p(O|W, S). \quad (1)$$

In Fig. 1(b), the World State is unknown, and the problem of “learning what is where” involves inferring the World State, given an agent’s Desire and Action using Bayes’ rule:

$$p(W|A, D, S) \propto p(A|D, W, S) p(W). \quad (2)$$

Intuitively, this involves evaluating the likelihood of every possible World State, given the agent’s Action, Desire and Agent State, and integrating these likelihoods with the prior over possible World States. Evaluation of each likelihood also requires simultaneously inferring and updating the agent’s Beliefs over time.

### An alternative cue-based model

To assess the intrinsic difficulty or logical complexity of our task, we formulated a cue-based alternative to our BToM account of social inference of food cart locations. We name the alternative model F-40; the model considered 7 key features and fit 40 free parameters (one for each feature, plus an additive constant, multiplied by 5 independent response variables) using multinomial logistic regression to minimize the error in prediction of human judgments. The features were chosen to capture key moments in the paths that were strongly indicative of a preferred cart being in a particular location. Specifically, for each room, we assigned a unique vantage point at which the agent could see what was in that room, and could choose to either commit to eating at that room by moving North/South, or commit to moving to another vantage point by moving East/West. The set of vantage points is indicated by the marked cells in Fig. 5. Features Toward and Away were computed for each room by counting the number of times the agent moved to or away from that room, starting from that room’s vantage point. In addition to the 6 Toward and Away features, the 7th feature recorded whether or not the condition was part of the introduction (in which carts could not be closed) or the main experiment. Because of its large number of free parameters, we hypothesized that F-40 would capture those regularities in people’s judgments that could be explained by low-level movement properties.

## Experiment

### Design

Fig. 2 illustrates our experimental design. On each trial, subjects were shown either a complete or an incomplete path that the agent took. They were then asked to rate on a scale from 0 to 10 (with 0 meaning “Definitely Not”; 10 “Definitely”; and 5 “Maybe”) how much they believed each possible configuration of carts was the real one. Fig. 2(d) shows the six possible configurations of carts that subjects rated on each trial. Food cart names as well as stimulus order were randomized across subjects. For simplicity we will refer to the carts as Afghani (A), Burmese (B), and Colombian (C), always with the preference order:  $A \succ B \succ C$ .

In this scenario there are 24 possible worlds (6 possible permutations of the cart's locations multiplied by 4 permutations of carts A and B being open or closed). Stimuli were generated as follows. We assume that the agent always starts at the entrance of the North hallway, being able to choose between entering that hall, going to the West hall, or going to the East hall. An exhaustive list of possible paths was constructed by listing all possible combinations of the short-term goals of the agent (go to entrance of W hall, go to entrance of N hall, and go to entrance of W hall), assuming that the first time a hall is selected it is for the purpose of exploration, and any selection of a hall that had been selected before is for exploitation, meaning the agent has chosen where to eat. From the eleven exhaustively enumerated paths, two paths that only produced permutations of beliefs were removed, leaving a total of 9 complete paths. In addition, 7 incomplete paths (subsequences of the 9 complete paths) which produce different judgments were selected. Lastly, three of these paths were duplicated in initial displays in which all carts are assumed to be open, shown to subjects to familiarize them with the task. This produced a total of 19 different paths (see Fig. 3) for which each subject rated the six possible configurations of carts, for a total of 114 judgments per subject.

## Participants

200 U.S. residents were recruited using the Amazon Mechanical Turk. 176 subjects were included in the analysis, with 24 excluded due to server error.

## Procedure

Subjects first completed a familiarization stage, which began with an explanation of the basic food cart setting, and allowed subjects to provide judgments for three paths where the food carts were assumed to always be open. Next, the possibility that carts could be closed was introduced with a step by step example. The experimental stage immediately followed.

## Results

We begin by analyzing the fit between people's judgments and our two models. Fig. 3 shows the average human rating of the likelihood of each cart configuration, the BToM model, and the F-40 model. In Fig. 3 it is clear that both models perform well in capturing the general contours of the mean subject belief, but with a quantitative difference in their explanatory power.

The BToM model has four parameters that were not fit to the data: three parameters indicating how strong the preference for each food cart is, and a discount parameter indicating the tradeoff between immediate and delayed rewards. Intuitively, these four parameters together determine whether an agent is willing to spend time and energy finding food carts he likes better or whether he should settle for a closer cart. These parameters were set only qualitatively, to ensure that the agent would have a strong preference order that would motivate him to explore the environment until he finds the best option.

In contrast, the F-40 model has forty free parameters fit to the average subject ratings, and so, by construction, the fit is very close to human judgment. Looking deeper into the model, there were no outstanding predictive features of the path that would determine the food cart ordering. That is, F-40 shows a great capacity to mimic human reasoning, but it fails to capture the essence of the task. This is clear in Fig. 4, where we can see that mean human judgments have a  $r = 0.91$  correlation with the BToM model, but a correlation of  $r = 0.64$  with the F-40 model. As we can see in the scatterplot, BToM comes much closer to explaining the variance in the human data, while F-40 is much less accurate overall. Fig. 4(c) shows that for a strong majority of individual subjects, the BToM model provides a superior fit. To further assess the statistical significance of the models we performed a Bootstrap Cross-Validated Correlational Analysis (Cohen, 1995). For 10,000 iterations, we trained F-40 on randomly selected subsets of paths and compared its performance on the remaining untrained paths. This produced average correlations of  $r = -0.0733$ ,  $-0.0832$  and  $0.0830$ , for training sets of size 16, 17 and 18 (and testing sets of size 3, 2, and 1), respectively. A similar analysis with BToM (for which no parameters were fit to data) yielded correlations of  $r = 0.9015$ ,  $0.8922$  and  $0.8714$  for testing sets of size 3, 2 and 1, respectively. These analyses suggest that the feature-based model is not tapping into the cognitive mechanisms underlying human performance, but rather just fitting the data without strong predictive power.

This is clear in Fig. 5, where two paths that contrast the models' performance are shown. For path 1, BToM is capable of realizing that if the carts were set as C, B, A/closed in positions West, North, and East respectively, then the agent would have no reason to visit the West position, since by the time it has observed B and A/closed it already has all the information it needs to make its final choice. It is this type of fine grained reasoning that allows BToM to make subtle inferences when F-40 fails as a result of mimicking the data rather than predicting it.

## Discussion

In this work we have proposed a Bayesian Theory of Mind model to explain how we make sense of the world by observing how others interact with it. Our experiment shows that subjects produce very similar predictions to that of the ideal Bayesian observer. We have compared the BToM model to a feature-based regression model (F-40) that was fit to subjects' mean judgments. Although the F-40 model appears to be a good competitor, we show that at both the individual and average level, the correlation with the BToM model is substantially higher compared to the correlation with F-40. Further analysis showed how BToM is capable of more subtle, fine-grained reasoning, making sensible inferences in several situations where F-40 gives counter-intuitive predictions.

One interesting point is that the BToM model is more sensitive to the precise geometry of the environment than humans

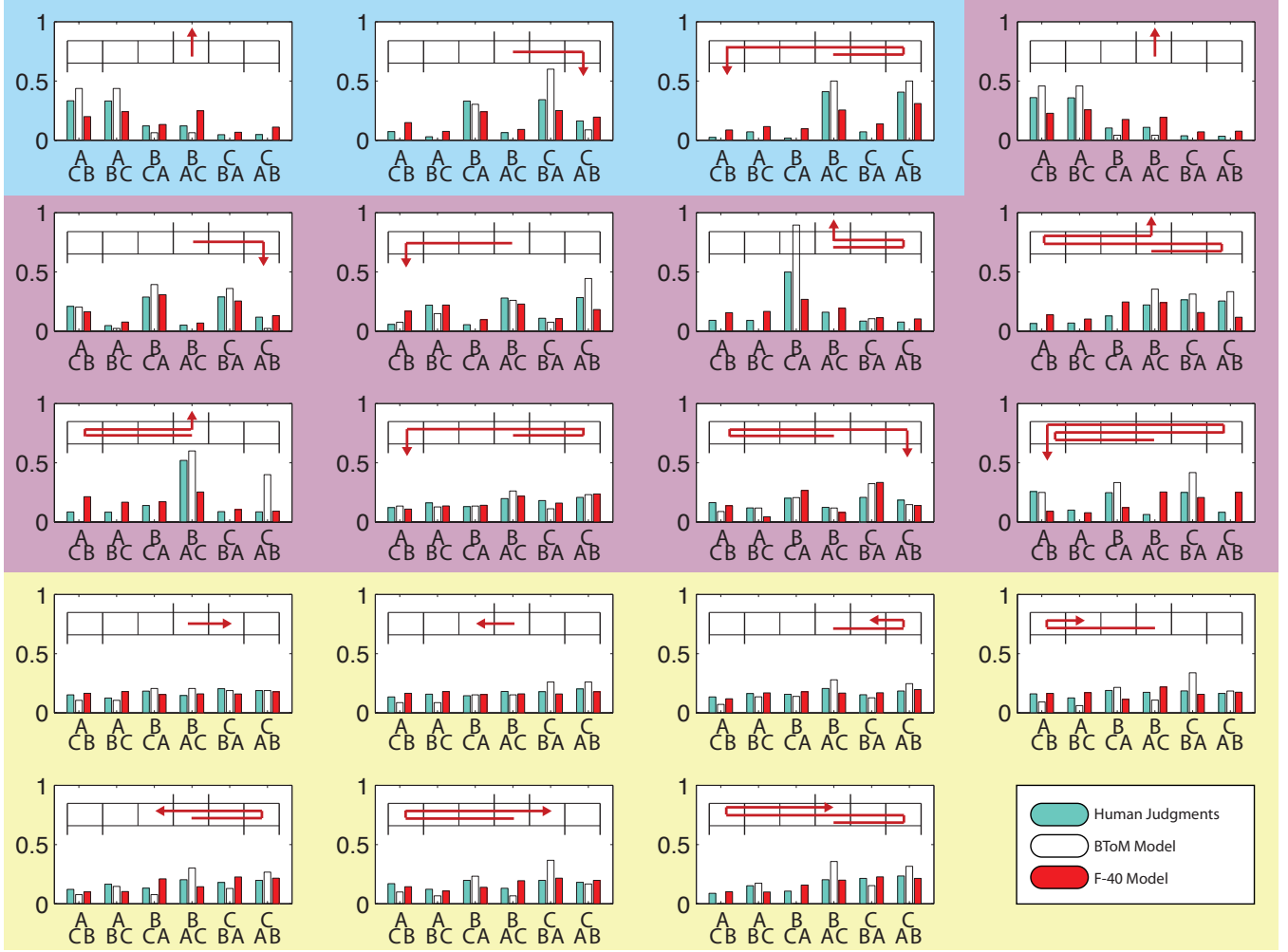


Figure 3: Mean subjects' judgments (normalized degrees of belief in each of six possible configurations of the food carts), along with BToM and F-40 predictions for the 19 displayed paths. The agent's preference order for these carts is always known to be  $A > B > C$ , and carts A and B may be open or closed. The first three conditions were those used in the familiarization phase. The second block used "completed" paths, in which the agent committed to a particular cart in the last frame. The last block of conditions used "incomplete" paths, in which the agent's final destination had not yet resolved.

seem to be. Specifically, because of the asymmetry of the hallway in our experiment, the model assigned a significantly higher cost to checking the West hallway versus checking the East hallway. Thus, when the model observed the agent going West, it reasoned that the agent must have had some prior belief in the presence of a high value cart in the West hallway or a low value cart in the East hallway that made him go through the more lengthy path versus the shorter path to check the East hallway. In contrast, subjects did not appear to be sensitive to the distance mismatch and produced relatively symmetric judgments on paths that had the same structure but traveled in opposite directions. This is particularly evident in the plot (3,1) of Fig. 3. In this path, subjects believed that the agent had already found carts (A) and (B) and therefore had no need to visit the East room. The model however, when observing the agent choose the longer path, reasoned that there was some prior belief the agent had that could have been wrong. This leads the model to consider it possible that

the (B) food truck was in the East room but that the agent had a prior belief that it was closed and therefore did not bother checking it. Analogous disparities were found by Baker et al. (2011), and in ongoing work we are investigating more qualitative representations of the spatial structure of the environment that might support a closer match between BToM model reasoning and human judgments.

In sum, these results show the power of social inference for acquiring surprisingly fine-grained knowledge about the world. ToM is typically thought of as a system of knowledge for reasoning about the mental states and actions of intentional agents, but it is not only that. In the context of a Bayesian framework, actions of other agents become clues to any aspects of the environment that causally influence their behavior – sometimes the only clues available. ToM thus also provides an essential tool for learning about the world.



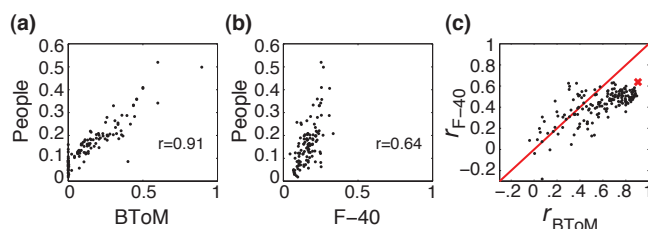


Figure 4: Comparison of models and normalized human judgments. (a) BToM vs. mean normalized human judgments. Each point represents a mean human rating plotted against the corresponding model prediction; there are 114 points in all (19 conditions times ratings for 6 possible cart configurations), with an overall correlation of  $r = 0.91$ . (b) F-40 vs. mean normalized human judgments; analogous to analysis (a), with an overall correlation of  $r = 0.64$ . (c) Scatter plot of individual subjects' correlations with BToM vs. individual subjects' correlations with F-40; 176 points in all (one point for each subject). For 80% of subjects, the correlation between BToM and that subject's ratings is higher than the correlation of F-40 with that subject's ratings. The bold "X" plots the correlation of BToM with the mean human judgments vs. that of F-40.

**Acknowledgements** This work was supported by ARO MURI contract W911NF-08-1-0242 and ONR MURI contract 1015GNA126. We thank the Motion Modeling, Analysis and Planning group at the National University of Singapore for providing the APPL POMDP solver. APPL is available at: <http://bigbird.comp.nus.edu.sg/pmwiki/farm/appl>.

## References

- Baker, C. L., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory-based social goal inference. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society* (pp. 1447–1455).
- Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113, 329–349.
- Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Thirtieth Third Annual Conference of the Cognitive Science Society* (p. 2469–2474).
- Bergen, L., Evans, O. R., & Tenenbaum, J. B. (2010). Learning structured preferences. In *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society* (pp. 853–858).
- Cohen, P. R. (1995). *Empirical methods in artificial intelligence*. Cambridge, MA: MIT Press.
- Csibra, G., Biró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, 27, 111–133.
- Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
- Goodman, N. D., Baker, C. L., & Tenenbaum, J. B. (2009). Cause and intent: Social reasoning in causal learning. In *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society* (pp. 2759–2764).
- Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.
- Jern, A., Lucas, C. G., & Kemp, C. (2012). Evaluating the inverse decision-making approach to preference learning. In *Advances in Neural Information Processing Systems*.
- Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101, 99–134.
- Kurniawati, H., Hsu, D., & Lee, W. (2008). SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *Proc. Robotics: Science and Systems*.
- Lucas, C. G., Griffiths, T. L., Xu, F., & Fawcett, C. (2009). A rational model of preference learning and choice prediction by children. In *Advances in Neural Information Processing Systems 21* (pp. 985–992).
- Tauber, S., & Steyvers, M. (2011). Using inverse planning and theory of mind for social goal inference. In *Proceedings of the Thirtieth Third Annual Conference of the Cognitive Science Society*.
- Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2010). Help or hinder: Bayesian models of social goal inference. In *Advances in Neural Information Processing Systems 22* (pp. 1874–1882).
- Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

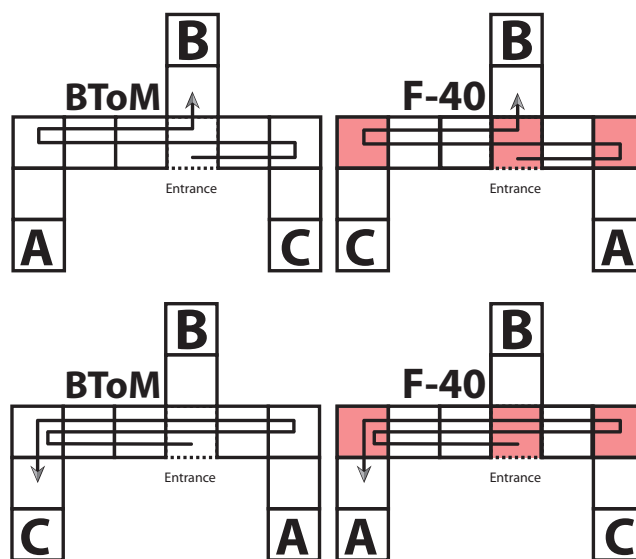


Figure 5: Comparing MAP predictions (the configuration shown for each path) of BToM and F-40 models for two different paths. F-40's errors reveal how a non-mentalistic approach fails to use context specific reasoning to make accurate predictions. For the F-40 model, marked grid squares indicate "vantage points" used to compute the features.

# Generalization of Learning in Games of Strategic Interaction

Ion Juvina ([ijuvina@cmu.edu](mailto:ijuvina@cmu.edu)), Christian Lebiere ([cl@cmu.edu](mailto:cl@cmu.edu))

Department of Psychology, Carnegie Mellon University, 5000 Forbes Ave.  
Pittsburgh, PA 15213 USA

Cleotilde Gonzalez ([coty@cmu.edu](mailto:coty@cmu.edu)), & Muniba Saleem ([msaleem@cmu.edu](mailto:msaleem@cmu.edu))

Department of Social and Decision Sciences, Carnegie Mellon University, 5000 Forbes Ave.  
Pittsburgh, PA 15213 USA

## Abstract

We present a laboratory study investigating the generalization of learning across two games of strategic interaction. The participants' performance was higher when a game was played after, as compared to before, a different game. We found that the generalization of learning from one game to another was driven by both surface and deep similarities between the two games. We developed a computational cognitive model to investigate mechanisms of generalization. Model development highlighted some of the challenges of cognitive modeling in general and modeling strategic interaction in particular. We found that development of reciprocal trust was a key factor that explained the observed generalization effect.

**Keywords:** Cognitive modeling; Game theory; Strategic interaction; Generalization of learning.

## Introduction and Background

Games of strategic interaction have successfully been used to model various real-world phenomena. For example, the game Prisoner's Dilemma has extensively been used as a model for real-world conflict and cooperation (Rapoport, Guyer, & Gordon, 1976). There has been a recent tendency toward studying ensembles of games, as most real-world "games" rarely occur in isolation; more often they take place either concurrently or in sequence (Bednar, Chen, Xiao Liu, & Page, in press). For instance, when games are played in sequence, an effect known as "spillover of precedent" may occur: a precedent of efficient play in a game can be transferred to the next game (e.g., Knez & Camerer, 2000). We refer to this effect as *generalization of learning in games of strategic interaction*. This effect has important practical implications. For example, most organizations employ training exercises to develop cooperation and trust among their employees. The assumption is that what is learned in a very specific, ad-hoc exercise generalizes to organizational life once the training is over.

Research on what factors cause generalization of learning in games of strategic interaction can be summarized as follows: (1) Bednar and colleagues (in press) use the concept of *entropy* or strategic uncertainty to explain when learned behavior is likely to spillover from one game to another. They suggest that prevalent strategies in games with low entropy are more likely to be used in games with high entropy, but not vice versa (Bednar et al., in press). In

other words, individuals develop strategies for easier games and apply them to more complex games. (2) Another explanation says that expecting others to do what they did in the past (and expecting that they will think you will do what you did in the past, etc.) can coordinate expectations about which of many equilibria will happen (Devetag, 2005). (3) Finally, Knez and Camerer (2000) found that generalization of learning across games strongly depended on the presence of superficial, surface similarity (what they call 'descriptive' similarity) between the two games. When the games differed in (what we call) surface characteristics (e.g., actions were numbered differently in the two games) transfer of learning from one game to another did not occur. This result is at odds with what is known from the literature on individual problem solving: generalization of learning is facilitated by our ability to perceive abstract, deep-level similarities, and it can be impeded by the presence of superficial, surface similarities (Holyoak & Thagard, 1995).

In this paper we present an experiment aimed at studying generalization of learning in games of strategic interaction. We want to understand when, why, in which direction, and under what conditions generalization occurs. We also present a computational cognitive model as an aid in our attempt to explain the empirical results and settle any potential inconsistencies in the literature.

The next section introduces the experiment and discusses its results. Then the cognitive model is described and its correspondence with the human data is discussed. The paper ends with a general conclusion.

## Experiment

Due to space limitation, only a brief description of the experiment is given here. A more detailed description was presented elsewhere (Juvina, Saleem, Gonzalez, & Lebiere, submitted). We selected two of the most representative games of strategic interaction: Prisoner's Dilemma (PD) and the Chicken Game (CG). They are both mixed-motive non-zero-sum games that are played repeatedly. Players can choose to maximize short- or long-term payoffs by engaging in cooperation or defection and coordinating their choices with each other. These features give these games the strategic dimension that makes them so relevant to real-world situations (Camerer, 2003). What makes PD and CG particularly suitable for this experiment is the presence of theoretically interesting similarities and differences,



providing an ideal material for studying generalization of learning. Table 1 presents the payoff matrices of PD and CG.

Table 1: Payoff matrices of PD and CG.

PD	A	B	CG	A	B
A	-1,-1	10,-10	A	-10,-10	10,-1
B	-10,10	1,1	B	-1,10	1,1

Both PD and CG have two symmetric (win-win and lose-lose) and two asymmetric (win-lose and lose-win) outcomes. Besides these similarities there are significant differences between the two games. In CG, either of the asymmetric outcomes is more effective in terms of joint payoffs than the [1,1] outcome. This is not the case in PD where an asymmetric outcome [10,-10] is inferior in terms of joint payoffs to the win-win outcome [1,1]. Mutual cooperation in CG can be reached by a strongly optimal strategy (i.e., alternation of [-1,10] and [10,-1]) or a weakly optimal strategy [1,1]. The optimal strategy in PD corresponds to the weakly optimal strategy in CG numerically, while the strongly optimal strategy of alternation in CG shares no surface-level similarities with the optimal strategy in PD. Thus, although mutual cooperation corresponds to different choices in the two games (i.e., surface-level dissimilarity), they share a deep-level similarity in the sense that mutual cooperation is, in the long run, superior to competition in both games. This provides a perfect test for our first hypothesis stating that individuals who have learned how to find an optimal strategy in one game will be more likely to find an optimal strategy in the next game even if those optimal strategies are different across the two games.

In both PD and CG, learning must occur not only at an individual level but also at a dyad level. If learning occurs only in one of the players in a dyad, the outcomes are disastrous for that player, because the best solution also bears the highest risk. For example, if only one player understands that alternating between the two moves is the optimal solution in CG, the outcome for that player can be a sequence of -1 and -10 payoffs. Only if both players understand the value of alternation and are willing to alternate, the result will be a sequence of 10 and -1 payoffs for each player, which in average gives each player a payoff of 4.5 points per round. Thus, the context of interdependence makes unilateral individual learning not only useless but also detrimental. The two players must jointly learn that only a solution that maximizes joint payoff is viable long term. However, this solution carries the most risk and thus it is potentially unstable in the long term. To ensure that the optimal solution is maintained from one round to another, there must exist a mechanism that mitigates the risk associated with this solution. It has been suggested that trust relations are self-sustaining once they have been developed (Hardin, 2002). In situations where there are benefits to individuals that can only be generated through mutual trust, each individual has an incentive to

maintain the relation. A trust relation develops through gradual risk-taking and reciprocation (Cook, Yamagishi, Cheshire, Cooper, Matsuda, & Mashima, 2005). In turn, as trust develops, risk is reduced and the trust relation becomes more stable. Our second hypothesis states that participants develop reciprocal trust throughout the first game, which facilitates learning of the optimal solution in the second game.

## Participants and Design

One hundred and twenty participants were paired with anonymous partners (leading to 60 pairs) and were asked to play the two games in sequence. The 60 pairs were randomly assigned to two conditions defined by the order in which the games were played: PD-CG and CG-PD. Participants played 200 unnumbered rounds of each game. At the end of each game, participants completed a five-item questionnaire assessing: how trustful they were of the opponent; how trustful of them the opponent was; how fair they thought the opponent's actions were; how fair the participants' actions were towards their opponents; and how satisfied they were with the overall outcome of the game.

## Results<sup>1</sup> and Discussion

To study generalization of learning across the two games, we analyzed the outcomes of a game according to when it was played. We also analyzed the round-by-round dynamics of these outcomes. The statistical significance of the observed effects was tested with the aid of Linear Mixed Effects analysis (*lmer* analysis from the LME4 package in R). This analysis was preferred instead of the classical analysis of variance (ANOVA) because the data violated the assumption of normality.

**Similarities and differences** The frequencies of the most relevant outcomes (i.e., the two symmetric ones and an alternation of the two asymmetric ones) are displayed in Figure 1 on a round-by-round basis. The first thing to notice is how different the two games are from each other from a behavioral perspective: the [1,1] outcome increases in PD but decreases in CG; alternation is prominent in CG but almost nonexistent in PD; and the mutually destructive outcome ([-1,-1] in PD and [-10,-10] in CG) is more frequent in PD than in CG. However, in spite of these apparent differences, the two games are similar in the sense that mutual cooperation emerges as the preferred solution and it becomes more and more stable over time. These patterns are in line with previous findings (e.g., Rapoport et al., 1976). Given this deep-level similarity, we expect players to be able to generalize their learning of the optimal strategy across the two games, although surface similarities might impede this process (Holyoak & Thagard, 1995). Since we ran the games in both orders (i.e., PD-CG and CG-

<sup>1</sup> Only a summary of the results is provided here as a context for understanding the cognitive model. A more detailed presentation of the results was given in Juvina et al., submitted.

PD), we can also test whether generalization occurs only in one direction, from low to high entropy, as suggested by Bednar and colleagues (in press).

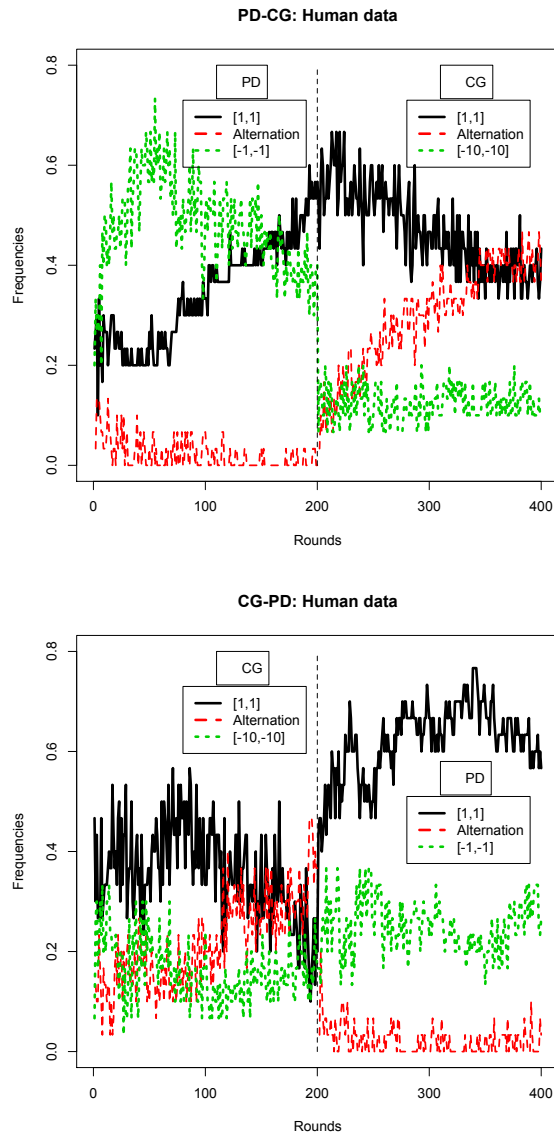


Figure 1: Frequencies of the most relevant outcomes in PD and CG by order (PD-CG on top and CG-PD on bottom) and round averaged across all the human participants.

**Generalization Driven by Surface Similarities** If learning across games is driven by surface similarities, one would expect the strategy that is learned in the first game to be applied in the second game as well, even though it may not be appropriate for the second game. This is indeed the case with regard to the  $[1,1]$  outcome in the PD-CG order: players learn that  $[1,1]$  is long-term optimal in PD and they are more likely to achieve it in the subsequent CG, even though it is only weakly optimal in CG. A LME model with occurrence of  $[1,1]$  as a dependent variable (binomial distribution), order, round, and their interaction as fixed factors, and participant as a random factor was used to test

the observed effects. There was a main effect of order ( $z = 2.21, p = 0.027$ ) and a main effect of round ( $z = -8.171, p < 0.001$ ); the interaction between order and round was also significant ( $z = -7.196, p < 0.001$ ) indicating that the main effect of order is larger at the beginning of the game and it progressively becomes smaller.

In the CG-PD order, if generalization of learning across games were driven by surface similarities, one would expect the strategy of alternating between the two asymmetrical outcomes to be attempted in the second game as well, at least in the beginning of the game. The main effect of order was non-significant ( $z = 1.476, p > 0.10$ ), suggesting that the strongly optimal strategy in CG (alternation) was not transferred as such (based on surface similarities) to PD. There remains the possibility that the  $[1,1]$  outcome was transferred as such from CG to PD. Even though the  $[1,1]$  outcome is only weakly optimal in CG, it was selected with relatively high frequency (see Figure 1) and it might have been considered optimal by some participants. We will revisit this point in the section on combined effects of surface and deep-level similarities.

#### Generalization Driven by Deep-Level Similarities

If learning across games was driven by deep-level similarities, one would expect learning the optimal strategy in the first game to increase the probability of learning the optimal strategy in the second game, even though there is no surface similarity between these strategies. These strategies ( $[1,1]$  in PD and alternation in CG) are similar only on an abstract, deep level: they both aim at maximizing joint payoff in a sustainable way, which in these two games is realistically possible only if the two players make (approximately) equal payoffs on a long run. On a surface level, these two strategies are very different. The  $[1,1]$  strategy in PD requires that players make the same move at each trial and they do not switch to the opposite move. In contrast, the alternation strategy in CG requires that players make opposite moves at each round and they continuously switch between the two moves. A LME model with occurrence of the alternation outcome in CG as a dependent variable, order, round and their interaction as fixed factors, and participant as a random factor revealed a main effect of order ( $z = -2.014, p = 0.044$ ) indicating a higher level of alternation when CG was played after PD, a main effect of round ( $z = 16.205, p < 0.001$ ) indicating that more and more pairs of participants discovered the alternation strategy as the game unfolded, and a significant interaction between order and round ( $z = 8.5, p < 0.001$ ) indicating that the optimal strategy was learned faster when CG was played second. The same analysis was conducted for the occurrence of the  $[1,1]$  outcome in PD and it revealed a main effect of order ( $z = -4.340, p < 0.001$ ) indicating that more pairs of participants discovered the optimal strategy in PD when it was played after CG, a main effect of round ( $z = 10.149, p < 0.001$ ) indicating that more and more pairs of participants found the optimal strategy as the game unfolded, and a significant interaction between order and round ( $z = 12.689,$

$p < 0.001$ ) indicating that the optimal strategy reached a ceiling when PD was played after CG, whereas it increased continuously when PD was played before CG. These results supported our first hypothesis. Specifically, learning the optimal strategy in the first game increased the probability of learning the optimal strategy in the second game, even though the optimal strategies were different in the two games. This generalization effect was significant in both directions (PD-CG and CG-PD) suggesting that entropy (Bednar et al., in press) has little explanatory relevance. If entropy were the causing factor, generalization would have only occurred in one direction.

**Combined Effects of Surface and Deep Similarities** In the case of deep-level generalization, the main effect of order was smaller in magnitude for CG ( $z = -2.014$ ,  $p = 0.044$ ) than for PD ( $z = -4.340$ ,  $p < 0.001$ ). It seems as if CG has a stronger impact on PD than vice versa. A possible explanation for this difference is based on how surface and deep-level similarities combine with each other to drive generalization of learning across games. They may have congruent or incongruent effects. Thus, in the PD-CG order, surface and deep-level similarities act in a divergent, incongruent way: surface similarity makes it more likely that the [1,1] outcome is selected whereas deep-level similarities make it more likely that the alternation outcome is selected. In other words, generalization based on surface similarity interferes with generalization based on deep-level similarity. In contrast, in the CG-PD order, both types of similarities act in a convergent, congruent way: they both increase the probability that the [1,1] outcome is selected. There is no impeding effect of surface similarity on PD because there is no optimal strategy in CG that is similar enough to a non-optimal or sub-optimal strategy in PD. The impeding and/or enabling effects of surface similarities on deep-level generalization are revisited in the modeling section.

**Reciprocal Trust** In addition to game choices, we analyzed the debriefing questionnaires that were administered at the end of each game. Since the answers to these questions were highly correlated with each other for any one individual participant, we averaged them in one composite variable that we call Reciprocal Trust. Since the debriefing questions were administered twice (at the end of each game) we refer to them as T1 and T2. We calculated correlations between these two trust variables and the variables indicating mutual cooperation in the two games. Spearman's  $\rho$  coefficient was used for correlations because the data failed to meet the normality assumption. We found that the more frequent mutual cooperation was in the first game the more likely the players were to rate each other as trustworthy at T1 ( $r = 0.75$ ,  $p < 0.001$  for PD and  $r = 0.42$ ,  $p < 0.001$  for CG). In addition, the more trustworthy players rated each other at T1, the more likely they were to enact mutual cooperation in the second game ( $r = 0.28$ ,  $p = 0.03$  for CG and  $r = 0.47$ ,  $p < 0.001$  for PD). Finally, mutual cooperation in the second

game predicted high levels of trust at T2 ( $r = 0.67$ ,  $p < 0.001$  for CG and  $r = 0.88$ ,  $p < 0.001$  for PD). As expected, the level of reciprocal trust increased from T1 to T2 (mean<sub>T1</sub> = 11.8, mean<sub>T2</sub> = 14.1,  $t = -3.247$ ,  $p = 0.001$ ). These correlations between trust and the frequency of mutual cooperation corroborate our second hypothesis. They suggest that generalization of learning driven by deep-level similarity is facilitated by development and maintenance of reciprocal trust. This finding will be essential for model development.

## A cognitive model of generalization of learning

Modeling generalization of learning across games of strategic interaction provides an opportunity to address some of the ongoing challenges of computational cognitive modeling. Three of these challenges are particularly relevant here and are described below as the model is introduced. The model is developed in ACT-R and it will be made freely available to the public on the ACT-R website<sup>2</sup>.

## Interdependence

In games of strategic interaction, players are aware of each other and their interdependence. In a previous study we showed that game outcomes were influenced by players' awareness of interdependence. In PD, the more information the two players in a dyad had about each other's options and payoffs the more likely they were to establish and maintain mutual cooperation (Martin, Gonzalez, Juvina, & Lebiere, submitted). Consequently, a cognitive model playing against another cognitive model in a simultaneous choice paradigm needs to develop an adequate representation of the opponent. We use instance-based learning (IBL) and sequence learning (SL) (Gonzalez, Lerch, & Lebiere, 2003) to ensure that the opponent is dynamically represented as the game unfolds. Specifically, at each round in the game an instance (snapshot of the current situation) is saved in memory. The instance contains the previous moves of the two players and the opponent's current move. Saved instances are used to develop contextualized expectations about the opponent's moves based on ACT-R's memory storage and retrieval mechanisms (Anderson, 2007). Expectations can explain some of the spillovers across games (Devetag, 2005).

## Generality

Before one attempts to build a model of generalization of learning across two games, one needs to have a model that is able to account for the human data in both games. Although by and large cognitive models are task-specific, there is a growing need to develop more general, task-independent models and there are a few precedents: Lebiere, Wallach, and West (2000) developed a model of PD that was able to account for human behavior in a number of other 2X2 games; and Salvucci (under revision) developed a "supermodel" that accounts for human data in seven

<sup>2</sup> <http://act-r.psy.cmu.edu/>

different tasks. We build upon these precedents of generality by developing a single model to account for round-by-round human data in both PD and CG. We achieve this generality by using instance-based learning for opponent modeling (as described in the previous section) and reinforcement learning for action selection. Both instance-based learning and reinforcement learning are very general learning mechanisms that can produce different results depending on their input. Specifically, at each round in the game, the model predicts the opponent's move based on the opponent's past behavior and selects its own move based on the utilities of its own past moves in the current context. The input that the model receives as it plays determines the model's behavior. The input is represented by opponent's move, own move, and the payoffs associated with these moves.

An important question is what constitutes the reward from which the model learns the utilities of its actions (moves). Players may try to maximize their own payoff, the opponent's payoff, the sum of the two player's payoffs, the difference, etc. Thus, a large number of reward structures can be imagined. A complicating assumption is that the reward structure might change as the game unfolds depending on the dynamics of the interaction between the two players. This indeed seems to be the case here, as we have realized after a large number of model explorations: no single preset reward structure is sufficient to account for the human data. One could try to computationally explore the space of all possible reward structures and their combinations to find the one that best fit the human data, but the value of this approach is questionable, because it may lead to a theoretically opaque solution. Instead, we chose to employ a theoretically guided exploration that drastically reduces the number of possible reward structures and, more importantly, gives us a principled way to describe the dynamics of players' motives as the game unfolds (see its description in the next section).

### Generalization of learning

When the model relies only on the two learning mechanisms described above (i.e., instance-based learning and reinforcement learning) it is able to only account for the generalization driven by surface similarities. Thus, the opponent is expected to make the same move in a given context as in the previous game. Similarly, an action that has been rewarded in the first game tends to be selected more often in the second game. It is impossible in this framework to account for generalization driven by deep-level similarities. For example, if the opponent used to repeat move B when it was reciprocated in PD, there is no reason to switch to alternation between A and B when none of these moves are reciprocated in CG. Moreover, learning within a game may in fact hinder generalization of learning across games if surface similarities are incongruent with the optimal solution in the target game. To find a solution to the deep generalization problem, we need to return to a theoretical and empirical analysis of the two games.

As mentioned in the introduction, in both PD and CG the long-term optimal solution bears the highest risk and, thus, it is unstable in the absence of reciprocal trust. We indeed found that self-reported trust increases after game playing and it positively correlates with the optimal outcome. It may well be that trust explains the deep-level generalization of learning across games. Players learn to trust each other and this affects their reward structure.

Recent literature on trust (e.g., Castelfranchi & Falcone, 2010) suggests that trust is essentially a mechanism that mitigates risk and develops through risk-taking and reciprocity. Inspired by this literature, we added a "trust accumulator" to our model – a variable that increases when the opponent makes a cooperative (risky) move and decreases when the opponent makes a competitive move. In addition, a variable called "willingness to invest in trust" was necessary to overcome situations in which both players strongly distrust each other and persist in a mutually destructive outcome, which further erodes their reciprocal trust, and so on. In such situations, the empirical data shows that players make attempts to develop trust by gradual risk-taking. When these attempts are reciprocated, trust starts to re-develop. In the absence of reciprocation these attempts are discontinued. The willingness to invest in trust increases with each mutually destructive outcome and decreases with each attempt to cooperate that is not reciprocated.

The variables "trust accumulator" and "willingness to invest in trust" are used to determine the dynamics of the reward structure. They both start at zero. When they both are zero or negative, the two players act selfishly by trying to maximize the difference between their own payoff and the opponent's payoff. This quickly leads to the mutually destructive outcome, which decreases trust but increases the willingness to invest in trust. When the latter is positive, a player acts selflessly, trying to maximize the opponent's payoff. This can lead to mutual cooperation and development of trust or players may relapse into mutual destruction. When the trust accumulator is positive, a player tries to maximize joint payoff and avoid exploitation. Thus, the model switches between three reward functions depending on the dynamics of trust between the two players. This mechanism provides a principled solution to the problem of selecting the right reward structure and in the same time solves the generalization problem: due to accumulation of trust in the first game, the model employs a reward structure that is conducive to the optimal solution and thus better performance in the second game.

### Modeling results

A cognitive model incorporating the principles described above was developed and fit to the human data presented in the previous section. Fitting the model to the human data was done manually by varying a number of parameters (of which some are standard in the ACT-R architecture and others were introduced as part of the trust mechanism) and trying to increase correlation ( $r$ ) and decrease root mean square deviation (RMSD) between model and human data.

The results of the current best model ( $r = 0.89$ ,  $\text{RMSD} = 0.09$ ) are presented in Figure 2.

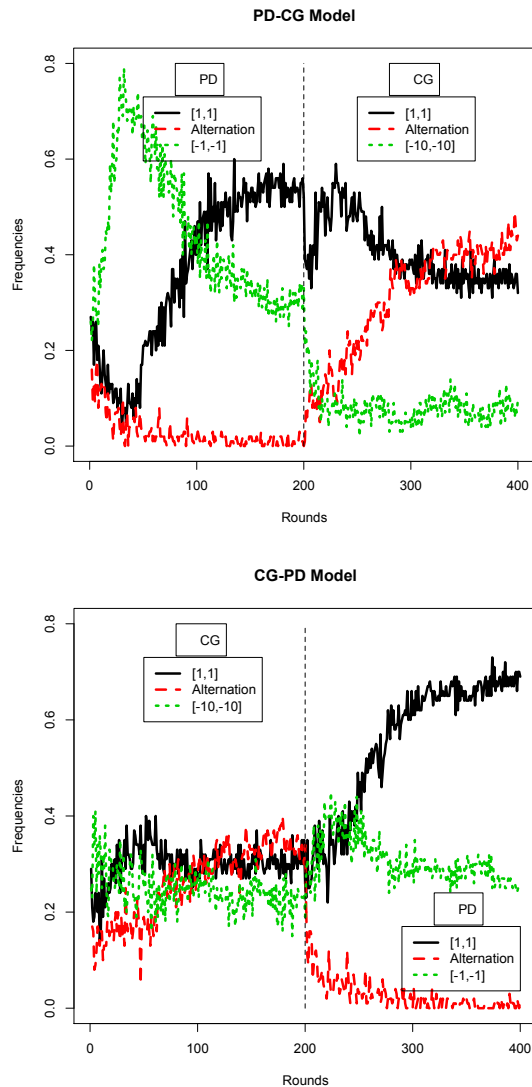


Figure 2: Results of model simulation.

Overall, the model matches the trends in the human data reasonably well (compare to Figure 1). More importantly, the generalization effects are also accounted for.

## Discussion and Conclusion

We found that generalization of learning across two games of strategic interaction is driven by deep-level similarities between the two games. Surface similarities may facilitate or hinder generalization depending on whether they are congruent or incongruent with the optimal solution. We used one cognitive model to account for human data in both games. This model helped to explain the observed generalization effect: reciprocal trust was necessary to mitigate the risk associated with the long-term optimal solution. We can conclude that some of the factors suggested in the literature are not necessary (entropy,

cognitive load) or insufficient (expectations, surface similarities), while others are essential (deep-level similarity and reciprocal trust) for generalization of learning in games of strategic interaction.

## Acknowledgments

This research was supported by the Defense Threat Reduction Agency (DTRA) grant number: HDTRA1-09-1-0053 to Cleotilde Gonzalez and Christian Lebiere.

## References

- Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford University Press.
- Bednar, J., Chen, Y., Xiao Liu, T., & Page, S.E. (in press). Behavioral Spillovers and Cognitive Load in Multiple Games: An Experimental Study. *Games and Economic Behavior*.
- Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, New Jersey: Princeton University Press.
- Castelfranchi, C., & Falcone, R. (2010). *Trust Theory: A Socio-Cognitive and Computational Model*. John Wiley and Sons.
- Cook, K. S., Yamagishi, T., Cheshire, C., Cooper, R., Matsuda, M., & Mashima, R. (2005). Trust Building via Risk Taking: A Cross-Societal Experiment. *Social Psychology Quarterly*, 68(2), 121-142.
- Devetag, G. (2005). Precedent transfer in coordination games: An experiment. *Economics Letters*, 89, 227-232.
- Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, 27, 591-635.
- Hardin, R. (2002). *Trust and Trustworthiness*. New York: Russell Sage Foundation.
- Holyoak, K. J., & Thagard, P. R. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.
- Juvina, I., Saleem, M., Gonzalez, C., & Lebiere, C. (submitted). Generalization of learning in conflict games: Effects of surface and deep level similarities. *Organizational Behavior and Human Decision Processes*.
- Knez, M., & Camerer, C. (2000). Increasing Cooperation in Prisoner's Dilemmas by Establishing a Precedent of Efficiency in Coordination Games. *Organizational Behavior and Human Decision Processes*, 82, 194-216.
- Lebiere, C., Wallach, D., & West, R. L. (2000). *A memory-based account of the prisoner's dilemma and other 2x2 games*. Paper presented at the International Conference on Cognitive Modeling.
- Martin, J. M., Gonzalez, C., Juvina, I., Lebiere, C. (submitted). Interdependence information and its effects on cooperation.
- Rapoport, A., Guyer, M.J., & Gordon, D.G. (1976). *The 2X2 game*. Ann Arbor, MI: The University of Michigan Press.
- Salvucci, D.D. (under revision). Integration and reuse in cognitive skill acquisition. *Cognitive Science*.

# Actively Learning Nouns Across Ambiguous Situations

George Kachergis, Chen Yu, and Richard M. Shiffrin

{gkacherg, chenyu, shiffrin}@indiana.edu

Department of Psychological & Brain Science / Cognitive Science Program  
Bloomington, IN 47405 USA

## Abstract

Previous research shows that people can use the co-occurrence of words and objects in ambiguous situations (i.e., containing multiple words and objects) to learn word meanings during a brief passive training period (Yu & Smith, 2007). However, learners in the world are not completely passive, but can affect how their environment is structured by moving their heads, eyes, and even objects. These actions can indicate attention to a language teacher, who may then be more likely to name the attended objects. Using a novel active learning paradigm in which learners choose which four objects they would like to see named on each successive trial, this study asks whether active learning is superior to passive learning in a cross-situational word learning context. Finding that learners perform better in active learning, we investigate the strategies that were most successful, discuss the implications, and model the results.

**Keywords:** active learning; statistical learning; cross-situational learning; temporal contiguity; language acquisition

## Introduction

Human infants learn words quite quickly despite many challenges facing them, including uncertainty and ambiguity in the language environment. Recent research has studied how learners may acquire word meanings from regularities in the co-occurrence of words and referents (e.g., objects). Such cross-situational statistical word learning relies on two assumptions: 1) that spoken words are often relevant to the visible environment, and 2) that learners can to some extent remember the co-occurrence of multiple words and objects in a scene. Thus, as words and their intended referents are observed in different situations over time, learners can apprehend the correct word-object mappings. Relying only on the regularity of the linguistic environment and basic memory and attention processes, this may be an important method of learning nouns for infants, and even adult travelers.

In adult cross-situational learning studies (e.g., Yu & Smith 2007), participants are asked to learn the meaning of alien words by watching a series of training trials. On each trial learners see an array of unfamiliar objects (e.g., four sculptures) and hear pseudowords (e.g., *stigson*, *bosa*). The meaning of each pseudoword is ambiguous on a given trial, because although each word refers to a single onscreen object, the intended referent is not indicated. In a typical learning scenario, participants attempt to learn 18 word-object pairings from 27 trials, with four words and four objects given per trial. In this design, each word-referent pair is presented six times over the five-minute training period. Learning a correct word-object pairing requires some form of accumulation of word-object co-occurrences.

When tested on each word and given four trained objects to choose from, participants chose the correct object for half of the 18 words, on average (Yu & Smith, 2007).

However, in the real world even infant learners are not passive observers, merely watching the world go by. As learners shift their attention, their eyes, head and hands move, changing the objects in their view. If caregivers notice these attention shifts, they may be more likely to name objects that are currently being attended to. Thus, learners may in essence be able to increase the likelihood an object is named by shifting their attention to include this object. This is a form of active learning, a concept studied extensively in machine learning (cf. Settles, 2009), in which a learner can query an information source for the labels of particular data points.

In this study, we introduce active cross-situational word learning, in which learners choose which four objects they would like to see named on each successive trial. Thus, learners control when to repeat pairs, when to stop experiencing pairs they feel they know, and when to attempt to learn more pairs. This gives us a glimpse of their preferred strategies. For example, participants may choose to repeat a single pair from the previous trial, and leverage working memory to quickly learn that the repeated word refers to the repeated object, while ignoring the other three word-object pairs on the trial. Equivalently, a learner may prefer to repeat three pairs from the previous trial, and quickly learn the novel pairing that was not present. Kachergis, Yu, & Shiffrin (2009a) manipulated this sort of temporal contiguity in a passive cross-situational study and found not only that repeated pairs are learned more easily, but so are unrepeated pairs in conditions with some repeats. This suggests that simple inference supported by working memory is not the only learning mechanism at work.

In fact, investigating active learning can reveal what information and mechanisms a learner has at their disposal, and characterizing the observed strategies—and their performance—will motivate learning models. For example, our recent associative model of cross-situational learning assumes that learners have access to both their familiarity and their uncertainty about the word-object pairings present on a given trial, and that attention competes for uncertain stimuli and for already-strong pairings (Kachergis, Yu, & Shiffrin, 2012). This model matches adult behavior in passive cross-situational experiments investigating mutual exclusivity, a bias to find 1-to-1 word-object mappings that is present even in 2.5-year-olds (Markman & Wachtel, 1988). If active learners have access to their knowledge of pairing strength and stimulus uncertainty, these cues can be combined to produce a few active learning strategies. One

strategy is to choose one object you have never seen before (i.e., one with maximal uncertainty), and fill the remaining three slots on the trial with familiar objects. Alternatively, learners may choose novel combinations of familiar objects in order to disambiguate mappings; we have previously found such contextual diversity to aid learning (Kachergis, Yu, & Shiffrin, 2009b). Detailed analysis of active learning strategies can reveal what knowledge is available to learners and how they attempt to employ it to learn the correct mappings. It may even be that people are worse at actively structuring the learning environment than the randomly-constructed passive trial sequences they normally experience in word-learning experiments.

In the Experiment, participants do two blocks of passive cross-situational learning, as well as of two blocks of active cross-situational learning in which they choose the objects that they see named on each successive trial. Although there are many other possible formulations of active cross-situational learning, we choose this instantiation because it most closely matches the passive task, and it somewhat matches the real world, where learners can attend to objects and likely increase the chance of a teacher labeling those objects.

## Experiment

Participants were asked to learn 18 word-referent pairs from a series of individually ambiguous training trials using the cross-situational word learning paradigm (Yu & Smith, 2007). Each training trial was comprised of a display of four novel objects and four spoken pseudowords. With no indication of which word refers to which object, learners have little chance of guessing the four correct word-referent mappings from the 16 possible pairings. However, since words always appear on trials with their proper referents, the correct pairings may be learned over the series of trials.

The key manipulation of this study is to allow learners in active conditions to choose which four objects they want to see named on the next trial. In both conditions, 18 word-referent pairs were experienced over a series of 27 training trials. Importantly, the same pair was never allowed to appear in neighboring trials in passive conditions. In both conditions, each pair could only appear six times during the training session. Thus, both the number of exposures per pair and the ambiguity on each trial (i.e., number of pairs) were matched in active and passive learning conditions. In order to compare passive and active learning performance, each participant underwent two training and test blocks of each.

## Subjects

Participants were 41 undergraduates at Indiana University who received course credit for participating. None had participated in other cross-situational experiments.

## Stimuli

Each training trial consisted of an array of four uncommon objects (e.g., sculptures) and four spoken pseudowords. The 72 pseudowords generated by computer are phonotactically-

probable in English (e.g., “bosa”), and were spoken by a monotone, synthetic female voice. These 72 objects and 72 words were randomly assigned to four sets of 18 word-object pairings, one set for each training condition.

Training for each condition consisted of 27 trials. Each training trial began with the appearance of four objects, which remained visible for the entire trial. After 2s of initial silence, the four words were heard in a random order (1s per word, with 2s of silence after each) for a total duration of 14s per trial.

## Procedure

Participants were told they would see a series of trials with four objects and four alien words, but that the order of presentation of the words was random. They were also told that their knowledge of which words belong with which objects would be tested at the end. In the active learning conditions, participants were instructed that they would be able to choose four objects they wanted to see named next. In active learning training blocks, after each trial a display of all 18 objects in the to-be-learned set was shown, and participants chose four to be named on the next trial. Objects that had already been chosen six times were not

After each training block, participants’ knowledge of word-object mappings was assessed using 18-alternative forced choice (18AFC) testing: on each test trial a single word was played, and the participant was instructed to choose the appropriate object from a display of all 18 trained objects. Each of the 18 words was tested once in a random order.

Every participant did four blocks of training and testing: half did two active learning blocks followed by two passive learning blocks, and the other half did the reverse.

## Results & Discussion

A repeated measures ANOVA on accuracy<sup>1</sup> by training type (active or passive) and training type repetition (1<sup>st</sup> or 2<sup>nd</sup>), nested by condition order (active-first or passive-first) revealed a significant main effect of training type ( $F(1,39) = 15.17$ ;  $p < .001$ ). Test performance after active learning is far better than after passive learning (active  $M = .59$ ; passive  $M = .35$ ), confirming that adults can use knowledge of their internal state to structure their environment for better learning. Moreover, participants did not improve much on their second block of either training type: there was no significant effect of repetition ( $F(1,39) = 2.08$ ;  $p = .15$ ). There was no significant interaction of condition order and repetition ( $F(2,38) = 1.62$ ;  $p = .20$ ), nor of training type and repetition ( $F < 1$ ), but there was a significant interaction of training type and condition order ( $F(2,38) = 4.53$ ;  $p < .05$ ). As shown in Figure 3, doing active learning first improves performance in the passive conditions (passive  $M = .30$  if passive-first,  $M = .39$  if active-first). This is somewhat surprising, as it is easy to imagine that doing passive first

<sup>1</sup> Data from one subject were excluded after it was found that their average performance in all four blocks was below chance (chance in an 18AFC test is .056).



might give learners an idea for better active learning strategies, whereas it is difficult to see how practice at active learning can improve one's performance in conditions with no command. However, it may be that active learning also allows learners to practice different rehearsal strategies, and helps them choose better ones even when they cannot control the structure of the trials. In any case, individual performance after the different types of training was significantly correlated (Pearson's  $r=.62$ ,  $t(38)=4.81$ ,  $p<.001$ ). Figure 2 shows that almost every participant performed at least as well after active training as passive training.

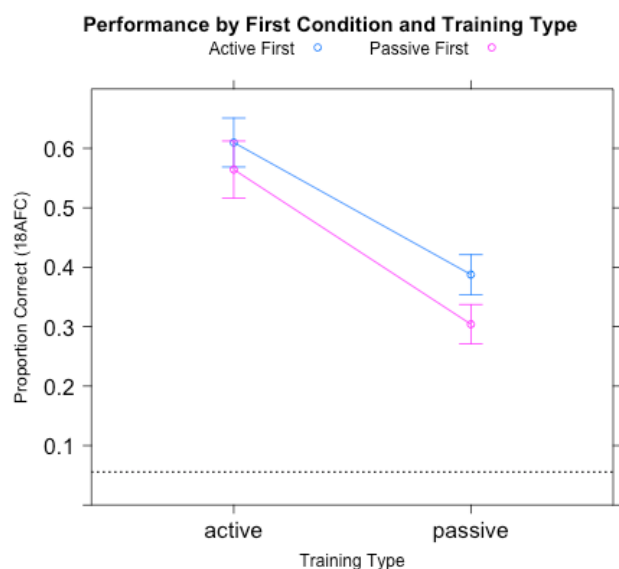


Figure 1: Accuracy by type of first condition and training type in the Experiment. Active learning resulted in far higher test performance than passive learning. Moreover, learners who did active learning first performed better in the passive learning conditions. Error bars show  $\pm$ SE.

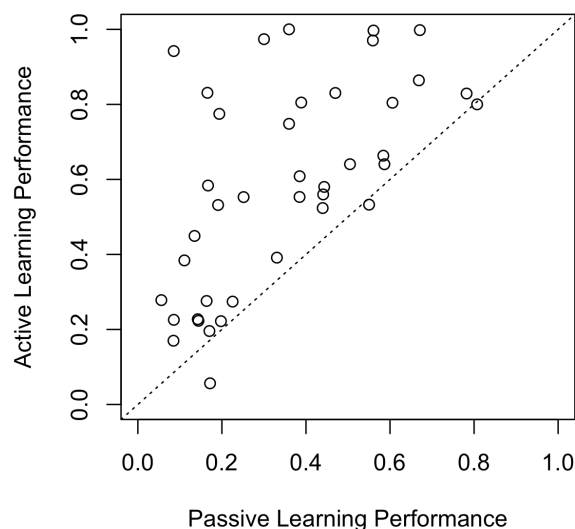


Figure 2: Comparison of performance after passive vs. active learning for each participant. Performance after the

two types of training is correlated ( $r=.62$ ), but learners are almost universally better after active training.

Given that adults can actively structure their environment in order to effectively learn the word meanings, we next investigate the strategies effective learners use to disambiguate mappings. However, we must first consider that there are many strategies, and that not all of them result in swift learning. Performance in cross-situational word-learning is typically highly variable, both within- and between-subjects. This is likely because what is learned on a given trial depends on what has been learned on all previous trials, and both the ambiguity on each trial and the fallibility of human memory means that people often learn different things. Giving learners an opportunity to structure training may yield a more diverse set of learning states, and thus may increase variability in performance. Figure 3 shows a histogram of learning performance after each block of active and passive learning. While accuracy after passive training is unimodal and positively skewed, accuracy after passive learning looks roughly bimodal, with peaks at .25 and at .95, which may reflect strategies of differing utility. In the following analysis, we will examine strategy differences both by doing a median split on the performance of active learners and analyzing the strategies used by each group, and by clustering the active training trials and looking at the mean performance of each cluster.

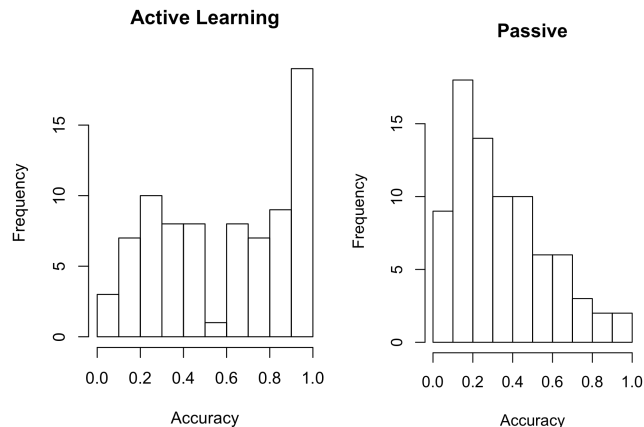


Figure 3: Histograms of performance after active learning (left) and passive learning (right) training blocks (2/subject/condition). Accuracy after active learning is bimodal, indicating that some strategies are quite successful while others are mediocre.

As mentioned earlier, one obvious active learning strategy is to choose to repeat some pairs from this trial on the next trial. If constructed randomly, a given trial would contain only .22 pairs repeated from the previous trial. In our passive training conditions, no pairs were allowed to repeat. The overall mean number of repeated pairs selected by active learners was 1.5—learners are using repetition to disambiguate pairs. To distinguish individual strategies (e.g., repeating one vs. repeating three), we clustered the

trial-by-trial number of repetitions chosen in each active learning training block. Using partitioning around medoids we found two clusters, estimated by the optimum average silhouette width (Kaufman & Rousseeuw, 1990). Cluster 1 contained 33 of the active training structures, and Cluster 2 contained the other 47. Figure 4 shows the trial-by-trial average number of repeated pairs for each cluster. Although people in both clusters initially repeat around one pair per trial, learners in Cluster 1 soon began to repeat two or more pairs on average, while those in Cluster 2 stayed closer to one repeat, until the last few trials<sup>2</sup>. Overall, Cluster 1 repeated 1.9 pairs per trial, significantly more than Cluster 2's mean of 1.1 repetitions (Welch  $t(60.7) = 9.09, p < .001$ ). From Figure 5, which shows how many pairs were repeated trial-by-trial in each cluster, it is clear that learners in Cluster 2 often chose to repeat single pairs until the very end. Cluster 1 shows a much more varied approach, repeating anywhere from one to three pairs. It turns out that these strategy clusters—constructed solely from the active training data—result in different overall levels of performance: Cluster 1's mean of .71 is significantly higher than Cluster 2's mean of .50 (Welch  $t(69.8) = 3.00, p < .01$ ). Repeating more than one pair seems to be a good strategy—indeed, the mean number of pairs repeated per trial in active training is correlated with learning (Pearson's  $r = .30, t(78) = 2.82, p < .01$ ). Corroborating this clustering result, a median ( $Mdn = .61$ ) split on active learning performance identifies a similar grouping: Cluster 1 contained 22 of the 33 better blocks, whereas Cluster 2 contained 30 of the 47 worse blocks ( $\chi^2 = 6.05, p = .01$ ). A graph of the active learning blocks identified by median split looks much like Figure 5, showing that better learners repeat more pairs.

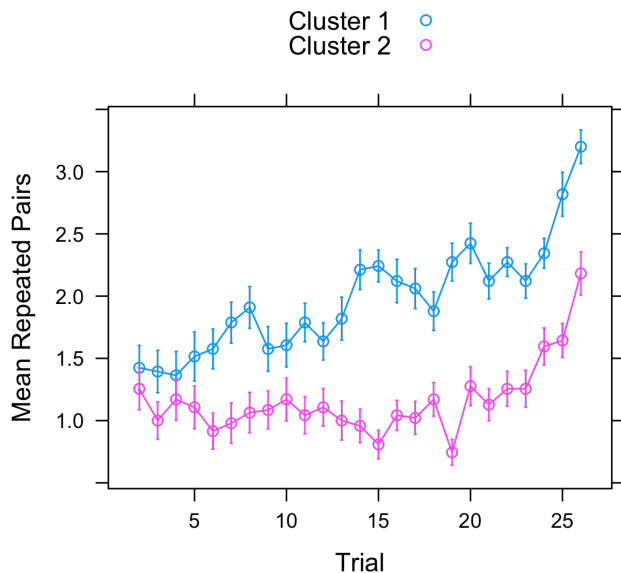


Figure 4: The mean number of word-object pairs repeated on consecutive trials by the two clusters of active learners. Learners in Cluster 1 repeated more pairs per trial than Cluster 2, except at the beginning, when both repeated  $\sim 1$ . Error bars show  $\pm$ SE.

<sup>2</sup> Due to the constraint of each pair appearing only six times—as in passive training—there are only a few objects remain to choose from, with the final trial being completely determined.

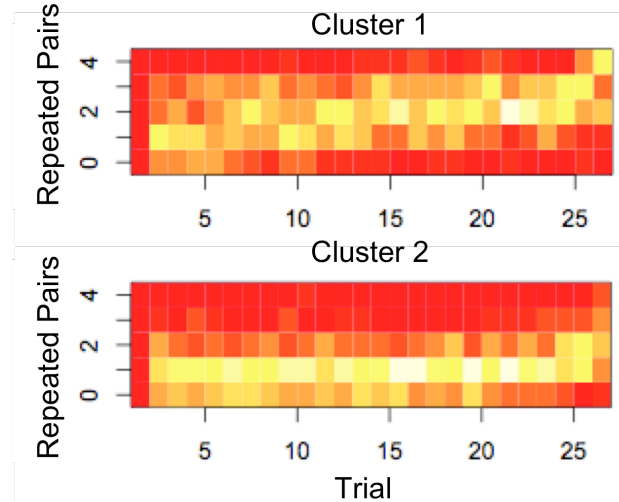


Figure 5: The number of pairs active learners chose to repeat on each consecutive trial, accumulated for each of the two clusters (red=0, white=27). Learners in Cluster 2 most often repeated one—or even zero—pairs, while Cluster 2 chose anywhere from one to three repeats per trial.

How is it that repeating more than one pair on each trial can further improve learning? Working memory can likely be used to segregate repeated and unrepeated pairs on a given trial. Thus, choosing one or three objects for repetition allows the learner to infer that the single repeated or new word goes with the repeated or new object. Repeating two pairs also yields information—only 8 associations are reasonable using repetition information, instead of 16 on a normal trial—but is most useful if a learner already knows one of the repeated pairs: then they may learn the unknown pair, and practice the known pair. To elucidate what learners are doing in active training when repeating multiple pairs, we extend a recent associative model of cross-situational word learning with a working memory mechanism.

## Model

The Experiment showed that adults learn many more words from active cross-situational training than passive training. Our analysis of active learning strategies found that most people repeated one or more pairs in consecutive trials, and that repeating more pairs helped: many excellent learners repeated close to two pairs per trial. To understand how this is helpful, we will introduce and then extend an associative model of cross-situational word learning proposed by Kachergis, Yu, and Shiffrin (2012).

The model assumes that learners do not equally attend to all word-object pairings on a trial (i.e., store all co-occurrences). Rather, selective attention on a trial is drawn to strengthen associations between words and objects that have co-occurred previously. This bias for familiar pairings competes with a bias to attend to stimuli that have no strong associates (e.g., as a novel stimulus). The competing familiarity and uncertainty biases allow the model to exhibit fast mapping, since a novel word-novel object combination will demand more attention, and mutual exclusivity: a novel

word will only become weakly associated with an already-known referent (Kachergis, Yu, & Shiffrin, 2012). For example, suppose word  $w_1$  and object  $o_1$  have appeared together and are thus somewhat associated, while  $w_7$  and  $o_7$  are novel. Given a trial with both pairs:  $\{w_1, o_1, w_7, o_7\}$ ,  $w_1-o_1$  demands more attention than  $w_7-o_1$ ,  $w_1-o_7$ , or  $w_7-o_7$ , since  $w_1-o_1$  is stronger than baseline. However, attention is also pulled individually to  $w_7$  and to  $o_7$ , since both of these novel stimuli have no strong associates. Uncertainty is measured by the entropy of each stimulus' association strengths. Because of the high joint uncertainty of  $w_7$  and  $o_7$ , more attention is given to the association  $w_7-o_7$ . Thus, attention is mostly divided between  $w_1-o_1$  and  $w_7-o_7$ , although the other pairings will be strengthened a bit.

Formally, let  $M$  be an  $n$  word  $\times$   $n$  object association matrix that is incrementally built during training. Cell  $M_{w,o}$  will be the strength of association between word  $w$  and object  $o$ . Strengths are subject to forgetting (i.e., general decay) but are augmented by viewing the particular stimuli. Before the first trial,  $M$  is empty. On each training trial  $t$ , a subset  $S$  of  $m$  word-object pairings appears. If new words and objects are seen, new rows and columns are first added. The initial values for these new rows and columns are  $k$ , a small constant (here, 0.01).

Association strengths are allowed to decay, and on each new trial a fixed amount of associative weight,  $\chi$ , is distributed among the associations between words and objects, and added to the strengths. The rule used to distribute  $\chi$  (i.e., attention) balances a bias for attending to unknown stimuli with a bias for strengthening already-strong associations. When a word and referent are repeated, extra attention (i.e.,  $\chi$ ) is given to this pair—a bias for prior knowledge. Pairs of stimuli with no strong associates also attract attention, whereas pairings between uncertain objects and known words, or vice-versa, draw little attention. To capture stimulus uncertainty, we allocate strength using entropy ( $H$ ), a measure of uncertainty that is 0 when the outcome of a variable is certain (e.g., a word appears with one object, and has never appeared with any other object), and maximal ( $\log_2 n$ ) when all of the  $n$  possible object (or word) associations are equally likely (e.g., when a stimulus has not been observed before, or if a stimulus were to appear with every other stimulus equally). In the model, on each trial the entropy of each word (and object) is calculated from the normalized row (column) vector of associations for that word (object),  $p(M_{w,\cdot})$ , as follows:

$$H(M_{w,\cdot}) = - \sum_{i=1}^n p(M_{w,i}) \cdot \log(p(M_{w,i}))$$

The update rule for allocating attention and adjusting strengths for the stimuli presented on a trial is:

$$M_{w,o} = \alpha M_{w,o} + \frac{\chi \cdot e^{\lambda \cdot (H(w) + H(o))} \cdot M_{w,o}}{\sum_{w \in S} \sum_{o \in S} e^{\lambda \cdot (H(w) + H(o))} \cdot M_{w,o}}$$

In this equation,  $\alpha$  is a parameter governing forgetting,  $\chi$  is the weight being distributed, and  $\lambda$  is a scaling parameter

governing differential weighting of uncertainty and prior knowledge (familiarity). As  $\lambda$  increases, the weight of uncertainty (i.e., the exponentiated entropy term, which includes both the word's and object's association entropies) increases relative to familiarity. The denominator normalizes the numerator so that exactly  $\chi$  associative weight is distributed among the potential associations on the trial. For stimuli not on a trial, only forgetting operates. After training, a learner is tested with each word and chooses an object from  $n$  alternatives in proportion to the association strengths of each alternative to that word.

Using competing biases for familiar pairings and uncertain stimuli, this associative model learns on a trial-by-trial basis by distributing attention in a way that corresponds with both our intuitions about word-learning and a number of empirical findings. However, although this model does exhibit training order effects, it has no working memory component that would confer additional benefit for successively repeated pairs. Thus, we augment the **baseline** model with a mechanism that segregates words and objects repeated from the last trial from unrepeated stimuli, and only strengthens associations within these subsets. This working memory (**WM**) model will learn better than the baseline model whenever there are repetitions, because of the 16 possible associations on the trial, it will not attend to the spurious ones between repeated stimuli and unrepeated stimuli: 6 in the case of one or three repeated pairs, and 8 in the case of two repeated pairs. To estimate whether people are attending more to the repeated or unrepeated stimuli, we added an attention parameter  $\beta$  to the WM model that apportions more weight to associations between repeated stimuli as  $\beta$  approaches 1, and more weight to unrepeated pairs as  $\beta$  approaches 0. When  $\beta = .5$ , the attention given to repeated vs. unrepeated associations is proportional to the size of each subset.

Three parameters ( $\chi$ ,  $\alpha$ , and  $\lambda$ ) were fit to each active training order for the baseline model, and four ( $\chi$ ,  $\alpha$ ,  $\lambda$ , and  $\beta$ ) were fit to the WM model. Fitting only to the overall mean accuracy of each active training order—two conditions per learner—does not capture detail of repetition's effect on accuracy, which may vary in different active training sessions. Instead, we fit to the accuracy for each subgroup of pairs that were repeated different numbers of times (0-5, as each pair was seen 6 times). An ANCOVA shows number of repetitions significantly affected accuracy ( $F(1,209) = 8.50, p < .01$ ), discussed in more detail later.

## Results & Discussion

Overall, both models achieved quite good fits to the data, with  $R^2 = .901$  for the baseline model, and  $R^2 = .925$  for the WM model. The WM model's BIC was 577.7 and the baseline model's BIC was 565.9, so the WM model is preferred, despite the additional parameter. Figure 5 shows mean accuracy for humans and both models on the subsets of pairs that were repeated on pairs of consecutive trials. Accuracy increases from 0 to 3 repetitions, while the few people who repeated pairs 4 or 5 times improved less, though with great variability.

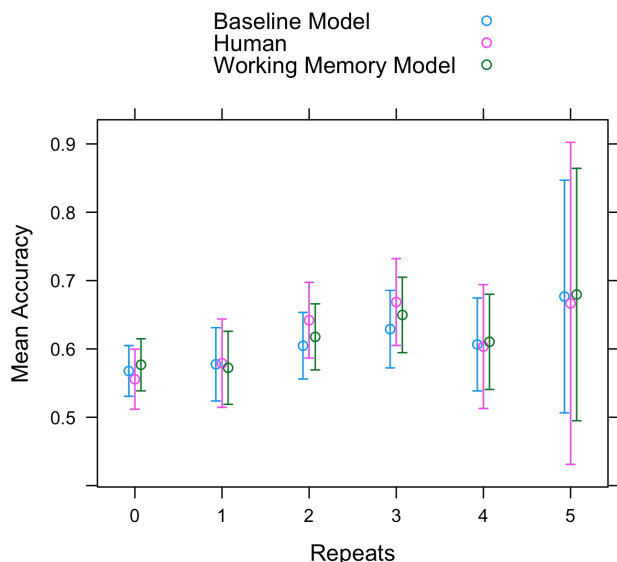


Figure 5. Human and model accuracy on actively-learned subsets of items that were repeated 0-5 pairs of trials (not necessarily consecutive for all repetitions—except in the rare case of 5 repetitions). Error bars are  $\pm$ SE.

Given the large number of repetitions used by active learners, it is surprising that the baseline model can approach the fit of the WM model without explicit awareness of repetitions. This may indicate that individual differences (e.g., in learning rate) contribute much of the variability. However, the WM better accounts for the data, and contains a parameter,  $\beta$ , that should be valuable in our pursuit to understand the range of strategies. Do learners focus more on repeated ( $\beta \approx 1$ ) pairs, unrepeatd pairs ( $\beta \approx 0$ ), or do they split attention ( $\beta \approx .5$ )? Figure 6 shows the trimodal distribution of the estimated  $\beta$  values: many people focused almost exclusively on learning the repeated pairs, but several attended only to unrepeatd pairs, and the majority split attention roughly equally. Once again, we see individual differences spanning the range of possibilities, although the peaks are of interest. However,  $\beta$  values were uncorrelated with accuracy ( $r = .06$ ), and people with modal  $\beta$  values showed no different accuracy, on average. Thus, the WM model found three attention strategies for repeated pairs, but the strategies alone do not predict performance.

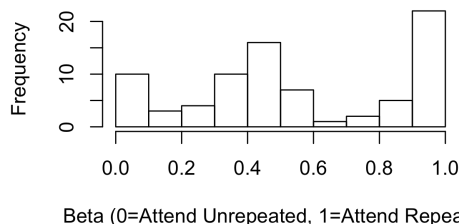


Figure 6: Histogram of best-fitting  $\beta$  values, showing a trimodal distribution peaked (highest to lowest) at 1—attend repeated pair, .5—split attention, and 0—attend unrepeatd.

## General Discussion

Active learning can speed language acquisition if the learner can implement an appropriate strategy based on the

information available to them. In the context of cross-situational word learning, we have shown that many adults can generate strategies that improve their overall learning. Indeed, people who did active learning first were better at passive learning, suggesting that some strategies carried over, which is somewhat puzzling because many of the active learning strategies involved trial-to-trial repetitions of at least one word-object pair—many more, in the most successful active learning blocks. Given that active learners were using many repetitions, but with apparently diverse strategies and outcomes, we extended a word-learning model with a working memory mechanism to attempt to see how people were leveraging repetitions. Overall, the model accounted for active learning accuracy very well, but parameters told of a plurality of strategies: many people ignore unrepeatd pairs while several only attend to these pairs, but the majority fall roughly in the middle, attending to both repeated and unrepeatd pairs. It may be that this focus often shifts during a block, as knowledge develops. Future work should also focus on predicting which pairs people will choose next, perhaps based on their current knowledge state.

In summary, active noun learners use many repetitions, and successfully learn far more than in passive training. Infants may also benefit from such repeated labeling, and fortunately there is much autocorrelation in scenes (as you turn your head or shift your eyes, many objects remain in view) and in language (conversations drift over minutes). Moreover, we suggest that infants likely influence their learning environment in a way that is analogous to the active learning paradigm we present here. By choosing to look longer at some objects, they may increase the likelihood that a caregiver will label one of those objects. Active learning is clearly a powerful learning aid, and with better understanding it can likely be harnessed in education to speed learning in many domains.

## References

- Kachergis, G., Yu, C., & Shiffrin, R. M. (2009a). Temporal contiguity in cross-situational statistical learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of CogSci 31*. Austin, TX: Cognitive Science Society.
- Kachergis, G., Yu, C., & Shiffrin, R. M. (2009). Frequency and contextual diversity effects in cross-situational word learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of CogSci 31* (pp. 755-760).
- Kachergis, G., Yu, C., & Shiffrin, R.M. (2012). An associative model of adaptive inference for learning word-referent mappings. *Psychonomic Bulletin & Review*.
- Kaufman, L. and Rousseeuw, P.J. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley, New York.
- Markman, E.M. & Wachtel, G.F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20, 121-157.
- Settles, B. (2009). Active learning literature survey. *Computer Sciences Technical Report 1648*, Uni. of Wisconsin-Madison.
- Yu, C. & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18, 414-420.

# Learning Nouns with Domain-General Associative Learning Mechanisms

George Kachergis

gkacherg@indiana.edu

Department of Psychological & Brain Science / Cognitive Science Program  
Bloomington, IN 47405 USA

## Abstract

Associative learning has been meticulously studied in many species, and diverse effects have been explained using a handful of basic assumptions and mechanisms. Human language acquisition proceeds remarkably quickly and is of great interest, but is arguably more difficult to capture under the microscope. Nonetheless, empirical investigations have led researchers to theorize a variety of language learning principles and constraints. While there may indeed be language-specific learning mechanisms that are distinct from more universal associative learning mechanisms, we seek to explain some basic principles of language acquisition using domain-general mechanisms. Using an experiment and a model, we show how the principles of mutual exclusivity—an assumption of 1-to-1 word-object mappings, contrast, and other constraints related to fast mapping may stem from attention mechanisms attributed to associative learning effects such as blocking and highlighting, but directed by competing biases for familiar and unfamiliar pairs instead of surprise.

**Keywords:** statistical learning; language acquisition; cross-situational learning; associative learning; attention

## Introduction

All organisms learn, but only humans master human languages. Since many neural structures and basic learning mechanisms are conserved across species, it bears asking how much of human language learning can be explained with domain-general mechanisms, without appealing to innate (i.e., evolved) linguistic knowledge, exemplified by the work of Noam Chomsky, or domain-specific principles and constraints, whether innate or developed early in life (e.g., Markman, 1992).

One essential part of language learning is learning word-object mappings—nouns. Two border collies have been shown to learn hundreds of nouns over years of training (Pilley & Reid, 2011; Kaminski, Call, & Fischer, 2004). Of course, this feat pales in comparison to human language learning: infants begin producing words at 1 year, and by the end of high school have command of 60,000 words, conservatively (Bloom, 2000). However, both dogs and infants have been shown to fast map: given a new word, they will choose a new object over an object with a known label, and retain the mapping weeks later (see Bloom, 2000). Fast mapping is a powerful ability for word learning, but is it based on domain-general or domain-specific learning mechanisms?

One approach to studying language acquisition views word learning as a problem of induction with an enormous hypothesis space, and proposes a number of constraints to restrict the space (Markman, 1992). In this view, infants generate hypotheses that are consistent with this set of constraints and principles. The present paper is concerned

with a subset of these principles that relate to how people map new words to objects.

Mutual exclusivity (ME) is the assumption that every object has only one name (Markman & Wachtel, 1988). A fill-the-lexical gap bias, which causes children to want to find a name for an object with no known name, has also been proposed (Clark, 1983) and argued (Merriman and Bowman, 1989). When given a set of familiar and unfamiliar objects, it has been shown that 28-month-olds assume that a new label maps to an unfamiliar object (e.g., Mervis & Bertrand, 1994). Similarly, the principle of contrast states that an infant given a new word will seek to attach it to an unlabeled object (Clark, 1983). Fill-the-gap, ME, and contrast make many of the same predictions made by the more general novel name-nameless category principle (N3C), which states that novel labels map to novel objects (Golinkoff, Mervis, & Hirsh-Pasek, 1994).

It is not our goal to explore the overlapping and nuanced ways that these various principles interact. Indeed, we hope to avoid this confusing plurality of explanations by showing that many of the behaviors ascribed to these theories can be explained by domain-general learning mechanisms uncovered by studies of associative learning. Nor are we the first to suggest that human language acquisition—as fast and yet difficult as it is—can be explained with domain-general learning mechanisms: Smith (2000) argued as much, and much recent work in statistical learning (described below) is motivated by this premise. Recent work has even found that children show a 1-to-1 bias in domains other than language: voices to faces (Moher, Feigenson, & Halberda, 2010) and actions to objects (Childers & Tomasello, 2003). However, few direct analogies have been drawn between the models and paradigms of word learning and associative learning, but see Ramscar et al. (2010). After introducing some associative learning paradigms and linking them to word learning, we discuss how universal attentional biases may account for many behaviors observed across domains. Finally, we report a new empirical word learning study using an associative learning highlighting design, and explain the results with a word-learning model that has competing attentional biases for familiarity and uncertainty.

## Associative Learning

Associative learning paradigms typically present one or more perceptual cues (e.g., objects, sounds), learners make a response (e.g., a button press), and feedback is given (e.g., food, a shock). When one cue  $q_1$  is paired with outcome  $o$  on each trial, the resulting  $q_1$ - $o$  association is stronger than  $q_1$ - $o$  when two simultaneous cues  $\{q_1, q_2\}$  predict  $o$  during training; thus,  $q_2$  is said to *overshadow*  $q_1$  (Pavlov, 1927). A



reasonable way to explain overshadowing is that attention is split between the two cues, and thus the associations  $q_1-o$  and  $q_2-o$  grow more slowly than when  $q_1$  appears alone. Attention is also used to explain the *blocking* effect (Kamin, 1968), which can be induced using a design with two training stages. In the early stage, cue  $q_1$  is repeatedly paired with outcome  $o$ , and in the late stage  $q_1$  and  $q_2$  appear jointly preceding  $o$ . The association between  $q_2$  and  $o$  is found to be much weaker than when only the late stage occurs. Thus  $q_2$  has been blocked by  $q_1$ 's earlier association with  $o$ —much like mutual exclusivity (ME) states that learners will not map a second label ( $q_2$ ) to a known object ( $o$ ). Learning models, updating knowledge trial-to-trial, account for blocking using selective attention to  $q_1$ : since  $q_1$  already predicts  $o$ , there is no need to strengthen  $q_2-o$  (e.g., Rescorla & Wagner, 1972; Pearce & Hall, 1980). Is blocking found in word-learning experiments? Can ME be thought of as blocking? As it happens, two cross-situational word-learning studies can be seen to address these questions.

### Cross-situational Word Learning

A key challenge in early word learning is to deal with the referential uncertainty intrinsic to complex scenes and utterances. Recent research has focused on how regularities in the co-occurrence of words and objects in the world can significantly reduce referential ambiguity across situations. Statistical word learning relies on two assumptions: 1) that spoken words are often relevant to the current situation, and 2) that learners can remember to some degree the co-occurrence of multiple words and objects in a scene. Thus, as the same words and objects are observed in different situations across time, people can learn the correct word-object mappings.

In adult cross-situational learning studies (e.g., Yu & Smith, 2007), participants are asked to learn the meaning of alien words from a series of training trials, each of which contains a few spoken words and a few objects. Although each word refers to a particular onscreen object, the intended referent is not indicated in any way, leaving meanings ambiguous on individual trials. Ichinco, Frank, and Saxe (2009) used a cross-situational word-learning task in which learners are first exposed to 1-to-1 pairings on a series of trials with four word-object pairs per trial. In the late stage, after people had presumably learned some of the mappings, a fifth object (or word, in another condition) began to consistently co-occur with one of the early words (or objects). The result was little learning of the association between old word (or object) and new object (or word) association, consistent with ME. However, this design can be seen to closely match a blocking design (see Table 1), with a few notable differences.

First, it is unclear whether words should be construed as cues and objects as outcomes, or the reverse—an issue we will return to. Second, a cross-situational trial has multiple outcomes, unlike associative learning paradigms. Finally, no trial-to-trial feedback is given, but the learner may generate it on the basis of the preceding training trials. We contend that none of these differences are a fundamental problem

with seeing cross-situational learning as associative learning. Indeed, if anything the learning problem in the real world is more like cross-situational learning: with a multitude of stimuli that may simultaneously serve as either cues or outcomes for as many other stimuli, learners attempt to associate correlated stimuli.

Training Stage	Ichinco et al., 2009	Kamin, 1968
Early	$\{w_1, w_x, w_y, w_z\} - \{o_1, o_x, o_y, o_z\}$	$q_1-o$
Late	$\{w_1, w_x, w_x, w_z\} - \{o_1, o_2, o_x, o_y, o_z\}$	$\{q_1, q_2\}-o$

Table 1: Comparison of the blocking paradigm (right) with a cross-situational word learning paradigm (left). In both paradigms, the late-stage stimulus ( $q_2 / o_2$ ) is blocked from becoming associated with the outcome ( $o / w_1$ ), despite consistent co-occurrence in the late stage.

Thus, learners in the Ichinco et al. study may not learn the extra association ( $w_1-o_2$ ) because attention remains focused on strengthening the still-present early-trained association ( $w_1-o_1$ ). This attentional account is equivalent to the popular account for blocking, and is corroborated by an earlier result that defies ME: Yurovsky and Yu (2008) used a two-stage cross-situational design much like Ichinco et al., but in the late stage when adding a new stimulus to an existing association, removed the old object (or word). Faced with a word ( $w_1$ ) they have associated with  $o_1$ , but now seeing  $o_2$  without  $o_1$  repeatedly, people learned the association, but also retained  $w_1-o_1$  at test. Yurovsky & Yu's learners cast about for a new associate, unblocked by the presence of an old associate to attend to—unlike in Ichinco et al.'s study. In summary, by establishing an analogy of cross-situational learning as a complex associative learning paradigm, we found that two cross-situational studies can be explained with a domain-general selective attention mechanism, without recourse to a language-specific constraint such as ME. To further examine the role of attention in cross-situational learning, we do a word learning experiment using a design that in associative learning yields the interesting order effect of highlighting.

### Experiment: Highlighting

Like blocking, highlighting is a learning order effect that has been attributed to selective attention (Medin & Edelson, 1988; Kruschke, 1996). In an early stage of training, a cues *PE* (Perfect Early) and *I* (Imperfect) jointly appear on each trial, followed by outcome *E* (Early). In a late stage, cue *I* appears with *PL* (Perfect Late), followed by outcome *L*. Thus, *I* imperfectly predicts both outcomes, having first predicted *E*, and later *L*. On the other hand, *PE* perfectly predicts *E*, and symmetrically, *PL* perfectly predicts *L*. As depicted in Figure 1, learners show an order effect: *PE* and *I* both become associated with *E* in the early stage, and then *PL* becomes more strongly linked with *L* while *I-PL* languishes. This is presumably because attention is shifted away from *I*, since it already predicts *E* in the early stage.

Formerly known as the inverse-base rate effect (note that *I* is twice as frequent as *PE* or *PL*), Kruschke (2009) presented a study with balanced frequency of the early and late training stages and still found highlighting, lending further credence to the attention account.

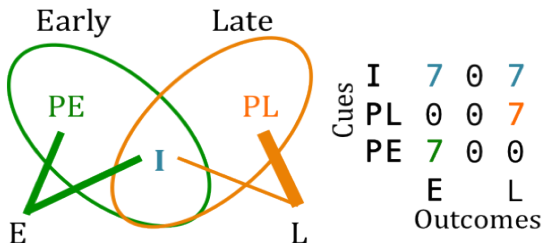


Figure 1: The co-occurrences of cues and outcomes in the highlighting design (right), and the estimated strength of associations between each cue and outcome (left), shown by the thickness of the lines.

Following a similar design, we use the Experiment to ask 1) whether highlighting occurs in a cross-situational framework with no explicit feedback on each trial, and 2) if words are cues and objects are outcomes, vice versa, or if they are interchangeable. As shown in Figure 2, this is done by making the cues in a highlighting design correspond to either words or objects, resulting in 2 words (cues) and 1 object (outcome) per trial, or 2 objects (cues) and 1 word (outcome) displayed per trial. Seeing a highlighting effect in one condition and not the other may suggest one correspondence over the other, whereas highlighting in both conditions suggests that words and objects can act as either cues or outcomes. Finally, finding no highlighting would suggest that domain-specific mechanisms may be at work in word learning.

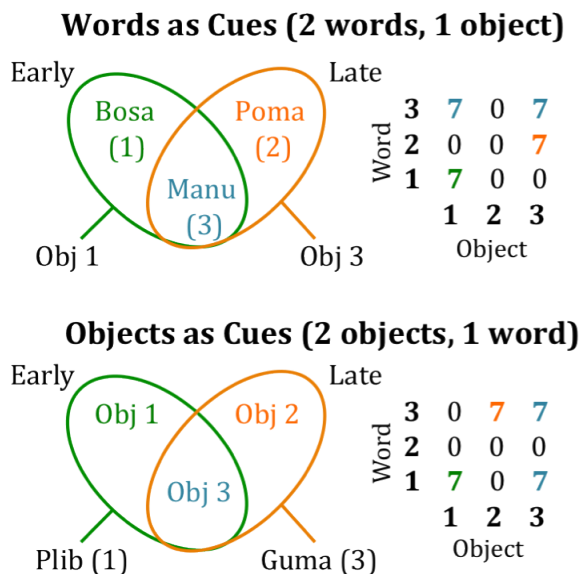


Figure 2: Highlighting designs in the Experiment (left), with word x object co-occurrence matrices (right). In the words as cues condition, 2 words and 1 object were given on each trial (top), while 2 objects and 1 word were given in the objects as cues condition (bottom).

## Subjects

Participants were 67 undergraduates at Indiana University who received course credit for participating. None had previously participated in cross-situational experiments.

## Stimuli & Procedure

Twelve pseudowords and 12 objects were randomly drawn from larger sets of stimuli, randomly paired, and split between the two conditions. The pseudowords (words) are phonotactically-probable in English (e.g., “bosa”), and were spoken by a monotone, synthetic female voice. The objects were photographs and drawings of uncommon objects (e.g., sculptures, specialty tools). Each training trial in the **words as cues** condition consisted of one object and two spoken words, while training trials in the **objects as cues** condition had two objects visible while one word was spoken. In both conditions, the object(s) remained visible for the duration of the trial. Each trial began with 2s of silence before the first 1s word was heard. In the words as cues condition, the second word was played after 1s of silence. In both conditions, the last word was followed by 3s of silence. In total, each trial in the objects as cues condition lasted 6s and trials in the words as cues condition lasted 7s.

Training for each condition consisted of 28 trials. The highlighting structures shown in Figure 2 were replicated within each condition: in words-as-cues, people heard six words and saw four objects, while in objects-as-cues, people heard four words and saw six objects. Knowledge was assessed after the completion of each condition using 6AFC testing: learners were asked to choose the best object for each of the six words. That is, we are probing the conditional probability objects, given a word. Note that in words-as-cues, two of the six objects available at test had not been seen during training, while in objects-as-cues, two words were never heard. These were not removed to keep the conditions symmetric, and in case systematic response deviations were found. Words were tested in random order. Note that the test in the words as cues condition corresponds most directly to associative learning testing: participants are given a cue (word) and asked to predict the outcome (object). In the objects as cues condition, we are actually asking learners to choose the best cue (object) when given an outcome (word). Participants completed both conditions in counterbalanced order.

## Results & Discussion

Figure 3 displays the conditional probabilities of choosing each object<sup>1</sup>, given each word, and the corresponding estimated relative strengths of each word-object association. The results in both conditions exhibit all the characteristics of highlighting: cue *I* is more strongly linked to *E* than *L*, and although *PE-E* and *PL-L* are both quite strong, *PL-L* is stronger. In the words as cues condition, object *o*<sub>1</sub> (*E*) was

<sup>1</sup> As noted before, there were two highlighting replications in each condition, so there were six objects available at test. Here we have collapsed the two replications for ease of presentation, and left out incorrect responses (e.g., choosing *o*<sub>4</sub>, *o*<sub>5</sub>, or *o*<sub>6</sub> for *w*<sub>1</sub>, *w*<sub>2</sub>, or *w*<sub>3</sub>). The mean response probability for these cells is .08.



chosen significantly more than  $o_3$  ( $L$ ; .51 vs. .25) for word  $w_3$  ( $I$ ;  $\chi^2(1, N=79) = 9.23, p < .01$ ). In the objects as cues condition,  $o_3$  ( $I$ ) was chosen significantly more often for  $w_1$  ( $E$ ) than  $w_3$  ( $L$ ; .28 vs. .16;  $\chi^2(1, N=73) = 6.04, p = .01$ ). Thus, the early association of  $I$  with  $E$  kept  $I$  from becoming strongly associated with  $L$ —much like a mutual exclusivity constraint would keep people from associating a second word with an already-labeled object. Given words as cues,  $L$  ( $o_3$ ) was chosen more often for  $PL$  ( $w_2$ ) than  $E$  ( $o_1$ ) was chosen for  $PE$  ( $o_1$ ; .82 vs. .69), though the difference was not significant ( $\chi^2(1, N=157) = 1.08, p = .30$ ). Similarly, given objects as cues,  $PL$  ( $o_2$ ) was chosen more often for  $L$  ( $w_3$ ) than  $PE$  ( $o_1$ ) was chosen for  $E$  ( $w_1$ ; .71 vs. .60), but again the difference was not significant ( $\chi^2(1, N=215) = 1.68, p = .20$ ). Despite not being statistically significant<sup>2</sup>, these conditional response rates match a highlighting result in both cases:  $PL$ - $L$  is learned faster (stronger) because little attention is given to  $I$ - $L$ , as cue  $I$  is already associated with  $E$ . In terms of word learning, this is much like the novel name-nameless category principle (N3C; Golinkoff et al., 1994): given a new object (or word— $PL$ ), it is reasonable to associate this with a new word (or object— $L$ ), rather than a word (or object— $PE$ ) with an already-known associate ( $E$ ).

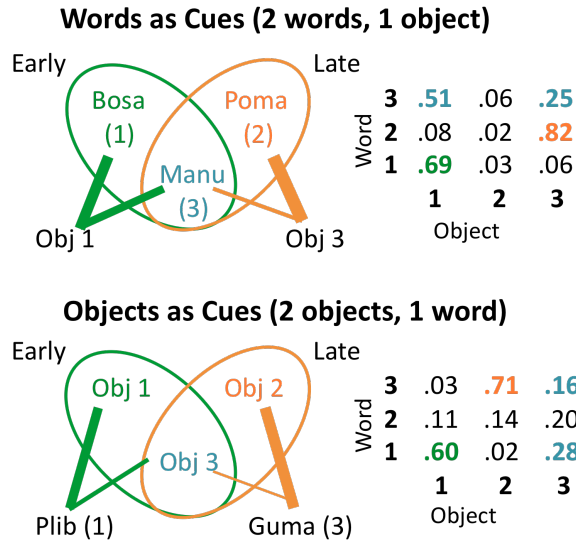


Figure 3: Collated response probabilities ( $p(o|w)$ ) for the two conditions in the Experiment (right). Both conditions show evidence of highlighting, with estimated association strengths shown by thickness of cue-outcome links (left).

In summary, the Experiment shows that highlighting can take place in a cross-situational word learning context, both with objects as cues and words as outcomes, and vice versa. The selective attention account of highlighting holds that the early association of  $PE$  and  $I$  with  $E$  reduces attention to the

later co-occurrence of  $I$  with  $L$ , thereby leaving  $PL$ - $L$  to gain more attention (i.e., strength). We contend that this domain-general account explains word-learning behavior not only in this Experiment, but in many situations that have motivated verbal theories of language-specific constraints. In the next section, we introduce a version of a recent associative model of word learning that shows what sort of attentional biases can account for highlighting—and word learning.

## Model

Familiarity and novelty are among the simplest ways to be aware of one's knowledge state about stimuli, and both biases have been observed in infants—inferred from their influence on attention (for an overview, see Hunter & Ames, 1988). Kachergis, Yu, and Shiffrin (2012) introduced an associative model with these biases, and showed that it accounts for fast mapping in adults, as well as gradual relaxation of ME with further training. The model assumes that word-object pairings on each trial compete for attention (i.e., associative strength). Attention is preferentially given to word-object pairings that are already associated by previous co-occurrence. Such a mechanism naturally exhibits blocking, since after the early association of  $q_1$  with  $o$ , it will continue to strengthen  $q_1$ - $o$  in the late stage, barely attending  $q_2$ - $o$ . However, the model's bias for familiar pairings competes with a bias to attend to stimuli with no strong associate (e.g., a novel stimulus). This bias can help explain behaviors covered by language-learning principles such as contrast and N3C. We describe the model below, and show how it accounts for highlighting using competing attention for familiar pairings and uncertain stimuli.

Formally, let  $M$  be an  $m$  word  $\times$   $m$  object association matrix that is arbitrarily large (here,  $m=100$ ). Cell  $M_{w,o}$  is the strength of association between word  $w$  and object  $o$ . Strengths are subject to forgetting (i.e., general decay) but are augmented by viewing the particular stimuli. Before the first trial,  $M$  has no information: each cell is set to  $1/m$ . On each training trial  $t$ , a subset  $S$  of words and objects appear.

Association strengths are allowed to decay, and on each new trial a fixed amount of associative weight,  $\chi$ , is distributed among the associations between words and objects, and added to the strengths. The rule used to distribute  $\chi$  (i.e., attention) balances a preference for attending to unknown stimuli with a preference for strengthening already-strong associations. When a word and referent are repeated, extra attention (i.e.,  $\chi$ ) is given to this pair—a bias for prior knowledge. Pairs of stimuli with no or weak associates also attract attention, whereas pairings between uncertain objects and known words, or vice versa, do not attract much attention. To capture stimulus uncertainty, we allocate strength using entropy ( $H$ ), a measure of uncertainty that is 0 when the outcome of a variable is certain (e.g., a word appears with one object, and has never appeared with any other object), and maximal ( $\log_2 n$ ) when all of the  $n$  possible object (or word) associations are equally likely (e.g., when a stimulus has not been observed before, or if a stimulus were to appear with

<sup>2</sup> Testing the relative strength of  $PE$ - $E$  and  $PL$ - $L$  would ideally be done with a trial that presents both cue  $PE$  and  $PL$ , and asks learners which outcome is preferred. However, a test of this sort is difficult to do in a paradigm with spoken words, and we instead chose to match previous word learning paradigms for consistency.

every other stimulus equally). In the model, on each trial the entropy of each word (and object) is calculated from the normalized row (column) vector of associations for that word (object),  $p(M_{w,\cdot})$ , as follows:

$$H(M_{w,\cdot}) = - \sum_{i=1}^n p(M_{w,i}) \cdot \log(p(M_{w,i}))$$

The update rule for adjusting and allocating strengths for the stimuli presented on a trial is:

$$M_{w,o} = \alpha M_{w,o} + \frac{\chi \cdot e^{\lambda \cdot (H(w) + H(o))} \cdot M_{w,o}}{\sum_{w \in S} \sum_{o \in S} e^{\lambda \cdot (H(w) + H(o))} \cdot M_{w,o}}$$

In this equation,  $\alpha$  is a parameter governing forgetting,  $\chi$  is the weight being distributed, and  $\lambda$  is a scaling parameter governing differential weighting of uncertainty and prior knowledge (familiarity). As  $\lambda$  increases, the weight of uncertainty (i.e., the exponentiated entropy term, which includes both the word and object's association entropies) increases relative to familiarity. The denominator normalizes the numerator so that exactly  $\chi$  associative weight is distributed among the potential associations on the trial. For stimuli not on a trial, only forgetting operates. After training, for each word the model's choice probabilities on  $k$  alternative objects is determined by the softmax choice rule (Bridle, 1990):

$$p(o|w) = \frac{\exp(\phi M_{w,o})}{\sum_k \exp(\phi M_{w,o_k})}$$

where  $\phi$  is a scaling parameter that determines the level of discrimination the model shows:  $\phi$  values above 1 amplify small differences in association weights.

The model was trained on the same 28 trials of word-object co-occurrences experienced by participants in the two conditions, and the four parameters were fit to minimize the discrepancy between the model's predicted response rates and the 36 human choice proportions for each condition. In the words as cues condition, the best-fitting parameters ( $\chi=.11$ ,  $\lambda=.46$ ,  $\alpha=1$ ,  $\phi=6.16$ ) achieved an  $R^2$  of .984 (MSE=9.5e-4). In the objects as cues condition, the best-fitting parameters ( $\chi=.12$ ,  $\lambda=.37$ ,  $\alpha=1$ ,  $\phi=6.16$ ) achieved an  $R^2$  of .884 (MSE=.0044). Both fits are quite good, and the best-fitting parameters are close in value. With  $\alpha=1$ , forgetting was not operating; perhaps memory is not taxed by such a small number of words and objects—cross-situational studies typically have more than a dozen pairs. With  $\phi=6.16$ , the model showed good discrimination at test.

Shown in Figure 4, the model's response proportions are close to the data and fit qualitatively well, showing the highlighting effect in both conditions. How does the model do this? In the first stage of words-as-cues, when  $w_1$  (*PE*) and  $w_3$  (*I*) co-occur with  $o_1$  (*E*), attention is split between the associations  $w_1-o_1$  and  $w_3-o_1$ , and the uncertainty about all three stimuli drops as knowledge grows. Moving to the second stage, when  $w_2$  (*PL*) and  $w_3$  (*I*) appear with  $o_3$  (*L*),  $w_2-o_3$  demands more attention than  $w_3-o_3$  because  $w_3$  has lower uncertainty from early training, while  $w_2$  is novel and has no associates. During the second stage,  $w_2-o_3$  thus gets

more attention than  $w_3-o_3$ , and becomes relatively stronger than the early  $w_1-o_1$  association. Thus, using competing biases for uncertain stimuli and familiar associations, the model mimics the highlighting effect shown by people.

#### Words as Cues (2 words, 1 object)

	Model				Human			
Word	3	.53	.05	.26	3	.51	.06	.25
	2	.04	.04	.81	2	.08	.02	.82
	1	.68	.06	.06	1	.69	.03	.06
		1	2	3		1	2	3
	Object							

#### Objects as Cues (2 objects, 1 word)

		Model					Human		
Word	3	.03	.72	.20	3	.03	.71	.16	
	2	.17	.17	.20	2	.11	.14	.20	
	1	.44	.03	.40	1	.60	.02	.28	
		1	2	3		1	2	3	
		Object							

Figure 4: Human response probabilities (right) and model response probabilities (left) closely match in the words as cues condition (top), and match well with objects as cues, showing highlighting in both cases.

Intriguingly, the model shows an asymmetry between conditions that is less striking in humans. With objects as cues, when given  $w_1$  (*E*), the model shows less bias towards  $o_1$  (*PE*) than humans do: people choose  $o_1$  twice as often as  $o_3$  (*I*), whereas the model chooses  $o_1$  only a bit more than  $o_3$ . Humans may show a stronger bias for  $o_1$  (*PE*) because they have retrospectively decreased the association between  $o_3$  and  $w_1$  once  $o_3$  began appearing with  $w_3$ . Another possibility is that people use uncertainty at test:  $o_1$  (*PE*) has lower entropy than  $o_3$  (*I*) since it only occurred with  $w_1$ . With words as cues, both objects have equal entropy. This asymmetry deserves future study, and may yet leave room for language-specific constraints.

### General Discussion

We have presented an analogy between cross-situational word learning and associative learning, shown how a study of the former (Ichinco et al., 2009) is a blocking design, and suggested the result is straightforwardly explained with the same domain-general attention mechanism. As evidence that attention creates order effects in word learning, we found highlighting—an associative learning effect ascribed to attention (Medin & Edelson, 1988; Kruschke, 1996)—in a cross-situational word learning experiment. Moreover, we showed that an associative word-learning model with competing attentional biases for familiarity and uncertainty (Kachergis, Yu, & Shiffrin, 2012) accounts for these results.

By linking word learning to associative learning, as suggested by Smith (2000), we may find that the plurality of overlapping language-specific constraints (e.g., ME, N3C, contrast, and fill-the-gap) are unnecessary to explain many

language learning behaviors. Instead, we predict that a more parsimonious explanation will emerge, built upon a foundation of domain-general mechanisms. Language-specific principles and constraints may yet exist, but we should first see how far more universal mechanisms take us.

Moreover, note that this bridge between domains is two-way: the present study used what was originally a word-learning model to explain highlighting. Although our model's attentional account is similar to the account given by other learning models (for an overview see Kruschke, 2011), other models do not use competing uncertainty and familiarity biases to shift attention. Instead, many models use a measure of prediction error to determine the rate of association change (e.g., Rescorla & Wagner, 1972; Pearce & Hall, 1980). In language, objects produce words in speakers ("Watch out—snake!"), but words predict objects for listeners. For language learners, we have shown that both directions of training produce a highlighting effect, captured by our model's symmetric associations and simple biases without generating predictions. These mechanisms, based on some of the simplest cues of knowledge state, may also fare well in other associative learning paradigms—in and out of a word-learning context.

Thus, future work in both domains can benefit from an exchange of ideas to uncover commonalities and differences, and to flesh out and refine verbal theories. We hope that others will find it enlightening to explore the link between associative learning, language acquisition, and other domains.

### Acknowledgments

Thanks to John K. Kruschke and Stephen Denton for helpful discussions, to Daniel Yurovsky for comments, and to Patrick LaFree, Jennifer Lee, and Kim Mullen for data collection.

### References

- Bloom, P. (2000). *How children learn the meaning of words*. Cambridge, MA: MIT Press.
- Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fogelman-Soulié & J. Héroult (Eds.), *Neurocomputing: Algorithms, architectures and applications* (pp. 227-236). New York: Springer-Verlag.
- Childers, J. B., & Tomasello, M. (2003). Children extend both words and non-verbal actions to novel exemplars. *Developmental Science*, 6(2), 185-190.
- Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1(1), 1-55.
- Golinkoff, R. M., Mervis, C. V., & Hirsh-Pasek, K. (1994). Early object labels: The case for a developmental lexical principles framework. *Journal of Child Language*, 21, 125-155.
- Hunter, M. & Ames, E. (1988). A multifactor model of infant preferences for novel and familiar stimuli. In Rovee-Collier, C. & Libsitt, L. (Eds.) *Advances in Infancy Research*, 5 (pp. 69-95). Stamford, CT: Ablex.
- Ichincio, D., Frank, M.C., & Saxe, R. (2009). Cross-situational word learning respects mutual exclusivity. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of CogSci 31* (pp. 749-754).
- Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). An associative model of adaptive inference for learning word-referent mappings. *Psychonomic Bulletin & Review*.
- Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M.R. Jones (Ed.), *Miami Symposium on the Prediction of Behavior, 1967: Aversive Stimulation*. Coral Gables, FL: University of Miami Press (pp. 9-31).
- Kaminski, J., Call, J., & Fischer, J. (2004). Word Learning in a Domestic Dog: Evidence for "Fast Mapping." *Science*, 304(5677), 1682-1683.
- Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22, 3-26.
- Kruschke, J. K. (2009). Highlighting: A canonical experiment. In B. Ross (Ed.), *The Psychology of Learning and Motivation*, 51, 153-185.
- Kruschke, J. K. (2011). Models of attentional learning. In: E.M. Pothos and A.J. Wills (Eds.), *Formal Approaches in Categorization*, pp.120-152. Cambridge University Press.
- Markman, E. M. (1992). *Constraints on word learning: Speculations about their nature, origins and domain specificity*. In M. R. Gunnar, & M. P. Maratsos (Eds.), *Modularity and constraints in language and cognition: The Minnesota symposium on child psychology* (pp. 59-101). Hillsdale, NJ: Erlbaum.
- Markman, E.M. & Wachtel, G.F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20, 121-157.
- Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psych: General*, 117, 68-85.
- Mervis, C.B. & Bertrand, J. (1994). Acquisition of the Novel Name - Nameless Category (N3C) principle. *Child Development*, 65, 1646-1662.
- Moher, M., Feigenson, L. & Halberda, J. (2010). A one-to-one bias and fast mapping support preschoolers learning about faces and voices. *Cognitive Science*, 1-33.
- Pavlov, I. P. (1927). *Conditioned Reflexes*. London: Oxford University Press.
- Pearce, J.M. & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, 87(6), 532-552.
- Pilley, J. W. & Reid, A. K. (2011). Border collie comprehends object names as verbal referents. *Behavioural Processes*, 86, 184-195.
- Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The Effects of Feature-Label-Order and their implications for symbolic learning. *Cognitive Science*, 34(7), 909-957.
- Rescorla, R.A., Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black, W.F. Prokasy (Eds.) *Classical Conditioning II: Current Research and Theory*. New York: Appleton Century Crofts, pp. 64-99.
- Smith, L. B. (2000). Learning how to learn words: An associative crane. In *Becoming a Word Learner*. New York: Oxford University Press.
- Yu, C. & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18, 414-420.
- Yurovsky, D. & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. *Proceedings of CogSci 30* (pp. 715-720). Austin, TX: Cognitive Science Society.

# Testing a Distinctiveness Explanation of the Primacy Effect in Free Recall Using Event-Related Potentials

Siri-Maria Kamp (SnKamp@Mail.Usf.Edu)

Glen R. Forester (GForester@Mail.Usf.Edu)

Anthony R. Murphy (ARMurphy@Mail.Usf.Edu)

Ty Brumback (TBrumbac@Mail.Usf.Edu)

Emanuel Donchin (Donchin@Mail.Usf.Edu)

Department of Psychology, University of South Florida  
4202 E Fowler Ave, PCD 4118G, Tampa, FL 33620-7200 USA

## Abstract

The primacy effect in free recall is commonly attributed to more frequent rehearsals for stimuli in the first few serial positions. Using event-related potentials (ERPs), we investigated whether the first list position also provides a distinctive feature to the stimulus, which enhances its encoding and aids retrieval on a recall test. The amplitude of the P300 elicited by stimuli that deviate physically or semantically from their context has previously been shown to correlate with the probability of later recall when participants use rote rehearsal. We reasoned that if the temporal distinctiveness of the first item in a list contributes to its enhanced recall, such a P300 subsequent memory effect should be present for this item as well. Participants studied and immediately recalled lists of 15 words including one physically deviant “isolate” while their EEG was recorded. We quantified P300 amplitude by a principal component analysis, and applied a correction for inter-trial latency jitter. The first words in a list and isolates were better recalled than regular words in the middle list positions, and the P300 elicited by these words was correlated with subsequent recall. Regular words in the middle list positions, as well as words in the second list position, did not show such a P300 subsequent memory effect. These results support a distinctiveness-based explanation of the primacy effect in free recall.

**Keywords:** Primacy effect; event-related potentials; free recall; P300; subsequent memory effect.

## Introduction

We tested the hypothesis that items in the first serial position of a study list are distinctive, which accounts for the primacy effect in free recall. We used the P300 event-related potential (ERP) as an index of distinctiveness. Prior studies have shown that under rote rehearsal the amplitude of the P300 elicited by physically or semantically deviant study items is correlated with later recall success (Fabiani & Donchin, 1995; Fabiani, Karis & Donchin, 1990; Karis, Fabiani & Donchin, 1984). We investigated whether this same effect is also observed for the first list item, which would support a distinctiveness-based explanation of the primacy effect in free recall.

The term “primacy effect” refers to the increased probability of free recall of the first few, compared to the middle items within a study list. The most influential explanation for this effect attributes it to more frequent

rehearsals of primacy items (e.g., Rundus, 1971). However, some data indicate that part of the recall enhancement for the first item (at least when it is also the first item retrieved) cannot be explained by rehearsal frequency (e.g., Howard & Kahana, 2002). The first item is also at a unique list position that may make this item stand out. Therefore, the distinctiveness of the first item may contribute to its greater probability of recall success (e.g., Brown, Neath & Chater, 2007).

One difficulty in testing distinctiveness-based explanations of behavioral phenomena lies in the fact that “distinctiveness” refers to a subjective experience rather than a physical property of an object (Hunt, 2006). Therefore, it is essential to find a measure of distinctiveness that is independent of the enhanced recall. To this end, a neural index of perceived distinctiveness is the P300 component of the ERP (Sutton, Braren, Zubrin & John, 1965), which peaks between 300 and 700 ms after the presentation of stimuli that are rare, unexpected, as well as task-relevant (for a review see Donchin, 1981). The *context updating model* (Donchin, 1981; Donchin & Coles, 1988) associates P300 amplitude with the degree to which novel information conflicts with expectations derived from a mental schema; information thus registered as unexpected, or distinctive, is then incorporated to update the schema. Since this process occurs interactively with information in long-term memory, this theory closely relates the P300 to memory processes.

To the extent that P300 indexes distinctiveness, the results of Ritter, Vaughan and Costa (1968) support the idea that stimuli at the beginning of a sequence are distinctive when their onset is unpredictable. They showed that, in addition to physically deviant stimuli, only the first stimulus in a monotonous sequence elicited a P300. By contrast, the second stimulus and all subsequent stimuli that did not stand out from the sequence, did not elicit a P300.

Items that stand out from their study list are more likely to be recalled than non-distinctive items (Von Restorff, 1933). Several studies have shown that when participants use rote rehearsal, physically distinctive words that elicit larger P300 amplitudes are more likely to be recalled on a later free recall test (Fabiani & Donchin, 1995; Fabiani et al. 1990; Karis et al. 1984; Otten & Donchin, 2000). Since the variance in P300 amplitude and the variance in recall

probability are correlated, the enhancing effects of distinctiveness on recall can be indexed by this correlation.

The design used in these studies, as well as in the present study, is known as the *subsequent memory paradigm*. Participants view several study items while their brain activity is recorded and later complete a memory test. Then, their brain activity is sorted according to the degree to which items were subsequently remembered (the *subsequent memory effect*, for a review see Paller & Wagner, 2002).

Previous studies examining primacy subsequent memory effects using the P300 have yielded inconsistent results. Azizian and Polich (2007) reported that P300 amplitude elicited by words in the initial list positions was correlated with recall; however the authors collapsed ERPs across the first three list positions – a disadvantage from the standpoint of our hypothesis that P300 amplitude will be correlated with recall only for position 1. By contrast, Wiswede, Ruesseler and Muentz (2007) reported that although the first word elicited a P300, there was a P300 subsequent memory effect only for the *final* study words. A problem with this study is that only 11 out of 18 participants showed a primacy effect; a subsequent memory effect may have been obtained if the behavioral effect was more reliable. A third study found only a small primacy effect and no P300 subsequent memory effects for either primacy or recency positions (Rushby, Barry & Johnstone, 2002). This study, however, averaged over 5 consecutive list positions, again preventing conclusions about the first item only.

All three studies ignored the possibility that P300 latency may have varied between trials and participants - as their broad ERP waveforms suggest. Since such *latency jitter* can reduce average ERP amplitudes, a correction allows for more meaningful comparisons between conditions (e.g., Spencer, Abad, & Donchin, 2000). A further shortcoming is that all studies used mean- or peak amplitude measures to quantify the P300, which are not able to disentangle overlapping ERP components. Finally, no study included a manipulation known to elicit a P300, such as the isolation of a word by its font size; such a manipulation would allow for a direct comparison of subsequent memory effects for the first list position and isolates.

Some support for the idea that primacy items may show a P300 subsequent memory effect comes from an fMRI study in which the first list items elicited stronger activity in brain areas known to generate the P300 (the temporoparietal junction) when these words were later successfully retrieved in an associative recall test (Sommer, et al., 2006). However, fMRI has a lower temporal resolution than ERPs and the design differed from typical free recall studies, so our distinctiveness hypothesis remains to be tested.

We addressed the shortcomings of the prior ERP studies by including a physical “isolate” in each list, by applying a principal component analysis (PCA), and by correcting for latency jitter. We hypothesized that, similar to the isolates, words in list position 1 are distinctive and therefore elicit a P300, which will be larger for those words that are later successfully recalled, compared to forgotten items.

## Methods

For this study we combined data from two experiments, each employing 20 critical lists (i.e., lists of interest for the present analysis) randomly interspersed with other list types. Each list was presented as part of a study-recall sequence. Critical lists consisted of 15 words, including one physically deviant word (see below). In one study, lists that varied in word frequency ( $n=23$ ) were randomly interspersed with the critical lists; in the other study, word lists of varied emotional content were employed ( $n=22$ ). Here, we only report data from the critical list type. A comparison of the recall- and ERP data ensured that there were no differences between the samples.

**Participants.** Forty five college students participated in exchange for course credit ( $n=33$ ) or \$7 per hour ( $n=12$ ). The data from 14 participants were excluded from the analysis due to excessive artifacts in their EEG<sup>1</sup>, and one participant was excluded due to non-compliance with the instructions. The remaining 22 female and 8 male participants were between 18 and 45 years old ( $M=22.57$ ). All participants gave written informed consent, and all procedures were approved by the institutional review board.

**Stimuli.** Each study list contained 15 words, presented one at a time in white 16 pt font of Arial Unicode style, on a black screen. Stimuli included emotionally neutral nouns, verbs and adjectives with a word frequency of 11-50 per million according to Francis & Kucera (1982), and were between 3 and 8 letters long. The composition of each list and the order of words within a list were randomized, and no word was presented to the same participant twice. Words were presented for 250ms, followed by a fixation cross for 2s. Each critical list contained an “isolate” in a larger font size (24 pt), which was randomly placed between serial positions 6 and 10. After the last word of each list, a grey triangle appeared indicating the start of the recall phase.

**Procedure.** The experiment consisted of two sessions, each up to 2 1/2 hours long. Over the course of the two sessions, participants studied a total of 70 (exp 1) or 80 (exp 2) word lists, including 20 critical lists. The first session also contained 2 practice lists. After the preparations for the EEG recording, participants were seated at a distance of 60cm from the computer screen and instructed to memorize the words using rote rehearsal. After each list, participants wrote down every word they remembered in any order. Recall lasted at least 45s, but participants were allowed to write down words for as long as they wished. Participants initiated the start of the next list with a button press and breaks were allowed after sets of 4 lists. After the second session participants were debriefed about their encoding strategies.

---

<sup>1</sup> The high number of participants excluded due to artifacts was due to frequent movement artifacts at the beginning of the lists, possibly because participants were still getting comfortable.

**EEG Recording and Analysis.** The EEG was recorded with a 128 channel Electrical Geodesics, Inc. (EGI) system, digitized with a sampling rate of 250 Hz and referenced to electrode Cz. For all off-line analysis we used NetStation (EGI) software, the EEGLab toolbox (Delorme & Makeig, 2004), J. Dien's EP toolkit (Dien, 2010), as well as self-written MATLAB scripts. The data were low-pass filtered at 20 Hz and segmented into epochs of 400 ms before to 2000 ms after word onset. Segments were corrected for eye blink artifacts using independent component analysis. Segments still containing artifacts were excluded and the data were mathematically re-referenced to linked mastoids and baseline corrected for a time window of 400 ms before the stimulus. We computed ERPs separately for regular words in serial positions 6-10 (henceforth referred to as "standards"), words in a larger font size ("isolates"), and for words in serial positions 1, 2, and 3.

**Principal Component Analysis (PCA).** To quantify ERP amplitudes, we conducted a spatio-temporal PCA (Spencer, Dien & Donchin, 1999). This approach has been widely used as temporal PCA (Donchin, 1966; Donchin & Heffley, 1978) in the analysis of ERP data. With the advent of dense electrode array EEG recordings, spatial PCA was developed to identify ERP component's spatial distributions (Spencer, Dien & Donchin, 1999); this is then followed by a temporal PCA to identify time courses. The PCA approach allows parsing of the ERP into components and yields measures of component amplitudes (factor scores) that can be used for testing hypotheses.

Submitted to the PCA were the ERP averages of isolates, standards (list positions 6-10), and words in positions 1, 2, and 3. We rotated 15 factors, as identified by a scree test (Cattell, 1966), using the Promax rotation method (e.g., Dien, Beal & Berg, 2005). The factor score coefficients of the PCA factor corresponding to the P300 were then applied to each EEG trial to calculate "virtual ERPs" (factor scores plotted across the time points; Spencer et al., 1999).

Since the broad peaks of the virtual ERPs indicated that P300 latency varied between trials, they were corrected for latency-jitter using a cross-correlation technique (see Gratton, Kramer, Coles & Donchin, 1989, for a review of this and other jitter correction techniques). The grand average P300 virtual ERP was used as a template, which was cross-correlated with every trial. The point of maximal cross correlation was then used to determine a lag to shift this trial, with the restriction that the P300 peak had to lie within 450 and 750ms after word onset. Each trial was baseline corrected again for 200ms, and the average over the latency corrected trials was computed for each word type and recalled and not recalled words. Since isolates and words in positions 1-2 had low trial numbers and since the number of trials included in an ERP average can affect ERP amplitudes, we matched the recalled and not recalled categories for trial number by randomly selecting the same number of trials. This resulted in an average of 5 trials contributing to the recalled- and the not recalled isolates,

and an average of 4.8 and 4.6 trials for the recalled- and not recalled words in positions 1 and 2, respectively.

Finally, we quantified P300 amplitude by applying a temporal PCA on the jitter-corrected virtual ERPs to obtain a single factor score for each participant, word type, and for recalled and not recalled words. In the temporal PCA we rotated 15 factors with the Promax technique.

**Statistical Analysis.** Since the data violated the assumption of sphericity necessary for repeated measures ANOVA, we conducted a MANOVA on the recall rates for words in position 1, standards and isolates; as well as a 2x4 MANOVA on the P300 amplitudes (as quantified by the factor scores) testing for differences between word types (isolates, standards, position 1 and position 2), and recalled and not recalled words. Words from position 3 were not included in the statistical analysis since the ERPs showed the same pattern as position 2, and since one participant had no artifact-free trials for the "position 3/recalled" category.

## Results

**Behavioral Data.** The debriefing confirmed that most participants had used a rote memorization strategy. This was supported by the serial position curve (figure 1), which showed the typical shape, with a primacy effect for the first three to four serial positions and a recency effect over the last five serial positions.

Recall rates differed between words at serial position 1, isolates, and standards [Wilk's Lambda=.14,  $F(2,28)=87.07$ ,  $p<.01$ ]. Paired samples t-tests showed that recall for words in list position 1 was superior to recall for both standards,  $t(29)=8.46$ ,  $p<.01$ , and isolates,  $t(29)=2.41$ ,  $p<.05$ , while isolates were recalled with a higher probability than standards,  $t(29)=8.18$ ,  $p<.01$ . All but 4 participants showed superior recall for words in position 1 and all but 3 participants (a different set than the aforementioned 4) showed superior recall for isolates, compared to standards.

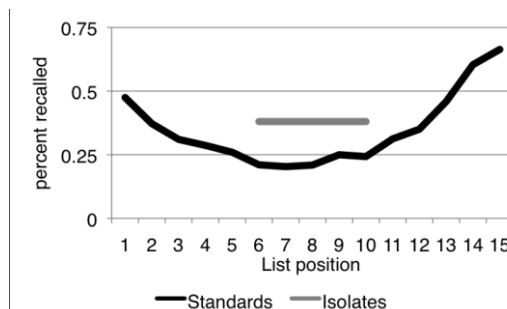
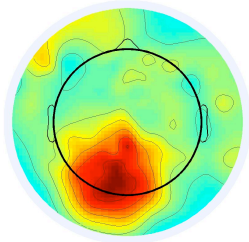


Figure 1. Serial position curve for regular-sized words ("standards"), and isolates. Note: percent recalled for isolates is averaged across positions 6-10.

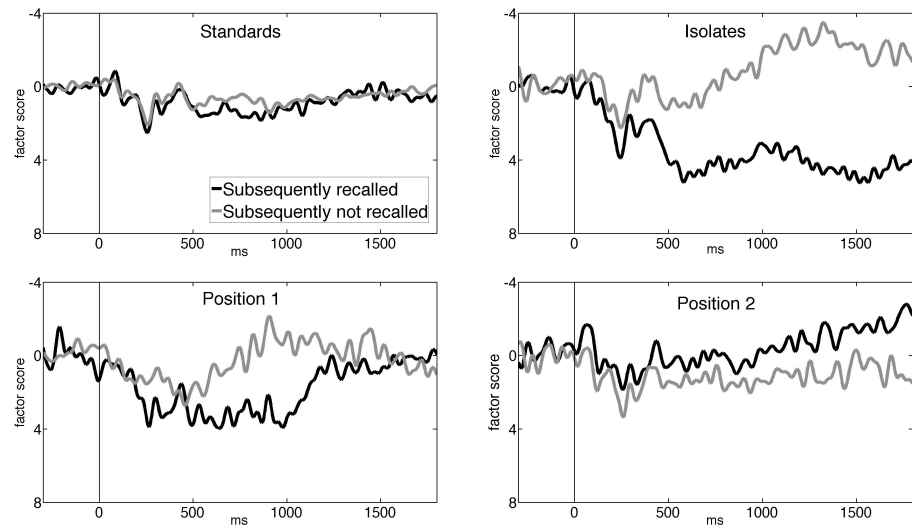
**Event-Related Potentials.** The spatial PCA solution accounted for a total of 85% of the variance in the data. Based on its parietal distribution, the fourth spatial factor was identified as the P300. Figure 2a displays the loadings of this factor, which accounted for 7% of the total variance.



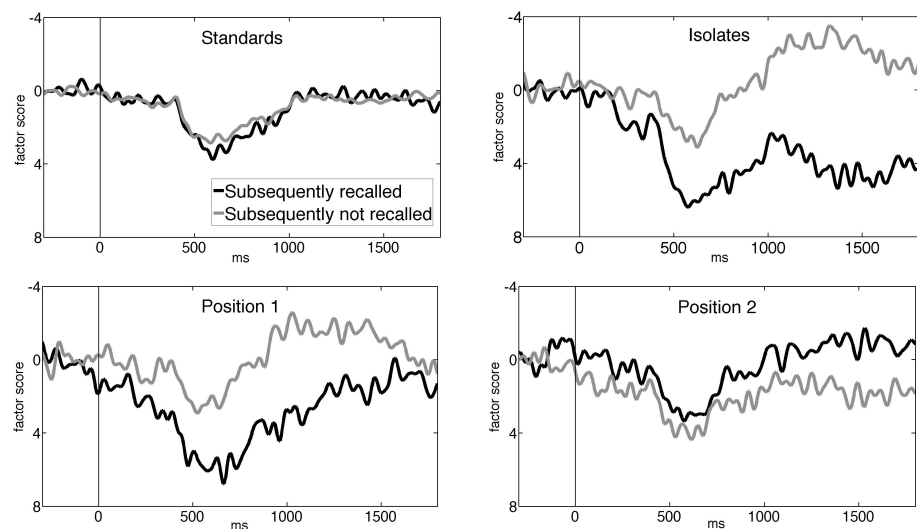
a. Spatial factor loadings



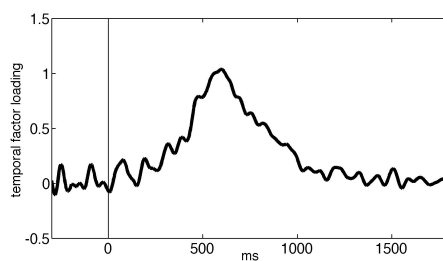
b. Raw virtual ERPs (not jitter-corrected)



c. Jitter-corrected virtual ERPs



d. Temporal factor loadings



e. Spatio-temporal factor scores

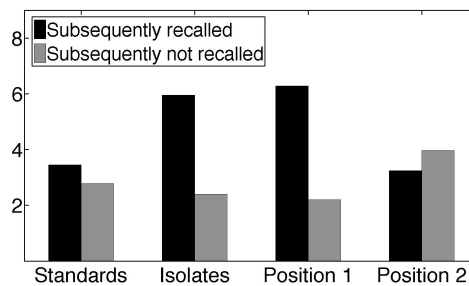


Figure 2. P300 PCA factor. a. Spatial factor 4; loadings over all 128 electrodes. b. Virtual ERPs for standards, isolates, and words in serial position 1 and 2 by subsequent recall. c. Latency-jitter corrected virtual ERPs. d. Temporal factor 2; loadings over all time points. e. Factor scores, indexing P300 amplitude, by word type and subsequent recall.



Figure 2b shows the raw virtual ERPs of the P300 factor by word type and subsequent recall. The P300 peaked between 500 and 700 ms after word onset. The broad peaks for isolates and words in position 1 strongly suggest the presence of latency jitter. The latency jitter corrected virtual ERPs are displayed in figure 2c. A comparison of figures 2b and 2c indicates that the jitter correction was successful, producing narrower peaks and higher amplitudes. Note that it is not unusual that even standards and words in position 2 now exhibit a small positivity, since the algorithm picks the point of maximal cross-correlation, and thus tends to bias the jitter-corrected average towards the template. It is also important to note that the direction of the difference between recalled and not recalled words was not changed by the jitter correction for any word type.

The second temporal factor, which accounted for 14% of the variance of the P300 virtual ERPs, lined up with the P300 peak and was therefore identified as the temporal P300 factor. Temporal factor loadings are displayed in figure 2c. Figure 2d shows the mean spatio-temporal factor scores, as a measure of P300 amplitude, for the four word types and for recalled and not recalled words.

The 2x4 MANOVA revealed that P300 amplitude differed between later recalled and not recalled items, Wilk's Lambda=.66,  $F(1,29)=14.77$ ,  $p<.01$ , which was qualified by an interaction between word type and recall success, Wilk's Lambda=.64,  $F(3,27)=4.98$ ,  $p<.01$ . There was no main effect for word type, Wilk's Lambda=.93,  $F(3,27)=63$ , *ns*. Critically, subsequent paired samples *t*-tests revealed a significant difference in P300 amplitude between recalled and not recalled words for isolates,  $t(29)=2.28$ ,  $p<.05$ , and words in position 1,  $t(29)=3.09$ ,  $p<.01$ , but not for standards,  $t(29)=1.21$ , *ns*, or words in position 2,  $t(29)=-.8$ , *ns*.<sup>1</sup>

To test whether overall, isolates and words in position 1 elicited a larger P300 than the other word types, we conducted planned comparisons between the combined P300 amplitude values of isolates and position 1; and the combination of standards and position 2, separately for recalled and not recalled words. Although for both recalled and not recalled words, isolates and words in position 1 elicited larger P300 amplitudes than the other word types, the difference only reached significance for the recalled words,  $t(29)=2.7$ ,  $p<.05$ .

## Discussion

We found a correlation of P300 amplitude with subsequent recall for isolates, replicating prior studies (e.g., Karis et al., 1984). Critically, the analogous subsequent memory effect

was evident for items in the first list position. Words in the middle- and the second list positions, by contrast, did not show this pattern. Since the P300 subsequent memory effect indexes the enhancing effects of item distinctiveness on recall, our results support the hypothesis that the first serial position provides a distinctive feature to the stimulus, thus enhancing encoding and aiding later retrieval.

Although Azizian and Polich (2007) reported a P300 subsequent memory effect for the first list positions, our study provides stronger support for the distinctiveness hypothesis of the primacy effect. First, our data indicate that the P300 subsequent memory effect is only present for the first word, suggesting that temporal distinctiveness does not extend to later serial positions. Second, by using PCA we were able to disentangle the P300 from other overlapping components. Furthermore, our latency-jitter correction insured that (1) differences between item types were not due to differences in P300 latency variability and (2) any true differences were not obscured by latency jitter, as may have been true in Wiswede et al. (2007) and Rushby et al. (2002). Finally, we were able to directly compare the subsequent memory effects for physical isolates and words in position 1, and these showed remarkable similarities (figure 2c).

Our distinctiveness explanation does not contradict the well-supported idea that rehearsal frequency accounts for the primacy effect (e.g., Rundus, 1971). Indeed, items at the first list position showed higher recall than the isolates, suggesting that item distinctiveness may not be the only factor enhancing recall of the first item. Recall was also enhanced for positions 2 and 3 (figure 1), which did not show a P300 subsequent memory effect. Therefore, we suggest that the temporal distinctiveness of the first item adds to the recall advantage by enhancing encoding and/or providing a distinctive retrieval cue. Further studies are necessary to investigate whether the effects of rehearsal frequency and distinctiveness are additive or synergistic.

Note that the P300 only indexes distinctiveness to the extent that the participant registers the distinctive feature at the time of stimulus encounter. It cannot index other conceptualizations of distinctiveness, such as distinctiveness of the first item due to the relatively early output position during recall (cf., Brown et al., 2007).

We did not have enough trials in the "position 15/not recalled" category to conduct a subsequent memory analysis for the recency positions. However, the last list item may also be perceived as distinctive, and therefore future studies should focus on such an analysis. Finally, an analysis of the relationship between individual differences in P300 amplitude and the behavioral effects was beyond the scope of this paper, but will be investigated in the future.

In conclusion, our study provides psychophysiological evidence for the hypothesis that the primacy effect in free recall is in part due to the enhancing effect of the first item's distinctiveness on recall. Our analysis focused only on the P300, but future studies will also be focused on the interaction of serial position effects with frontal slow wave subsequent memory effects, which are thought to index

<sup>1</sup> A supplementary analysis on the mean amplitudes of the raw ERPs between 500 to 700 ms at two parietal electrodes revealed the same patterns of results, with the exception that the subsequent memory effect for position 1 only approached significance ( $p=.11$ ). This may be due a decreased power for this comparison due to latency jitter.

working memory processes that support between-item elaborative encoding (e.g., Fabiani & Donchin, 1995).

### Acknowledgments

This work was in part funded by a USF Graduate School multidisciplinary challenge grant and a USF Signature Research Doctoral Fellowship awarded to S. Kamp.

### References

- Azizian, A., & Polich, J. (2007). Evidence for Attentional Gradient in the Serial Position Memory Curve from Event-Related Potentials. *Journal of Cognitive Neuroscience*, 19(12), 2071-2081.
- Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, 114(3), 539-576.
- Cattell, R. B. (1966). The Scree Test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245-276.
- Delorme, A., & Makeig, S. (2004). EEGLab: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134, 9-21.
- Dien, J. (2010). The ERP PCA toolkit: An open source program for advanced statistical analysis of event-related potential data. *Journal of Neuroscience Methods*, 187(1), 138-145.
- Dien, J., Beal, D. J., & Berg, P. (2005). Optimizing principal component analysis of event-related potentials: Matrix type, factor loading weighting, extraction and rotations. *Clinical Neurophysiology*, 116, 1808-1825.
- Donchin, E. (1966). A multivariate approach to the analysis of average evoked potentials. *IEEE Transactions on Bio-Medical Engineering*, BME-13, 131-139.
- Donchin, E. (1981). Surprise! ... Surprise? *Psychophysiology*, 18(5), 493-513.
- Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral Brain Science*, 11, 357-374.
- Donchin, E., & Heffley, E. (1978). Multivariate analysis of event-related potential data: A tutorial review. In D. Otto (Ed.), *Multidisciplinary perspectives in event-related brain potential research* (555-572). Washington, D.C.: U.S. Government Printing Office.
- Fabiani, M., & Donchin, E. (1995). Encoding processes and memory organization: a model of the von Restorff effect. *J Exp Psychol Learn Mem Cogn*, 21(1), 224 - 240.
- Fabiani, M., Karis, D., & Donchin, E. (1990). Effects of mnemonic strategy manipulation in a Von Restorff paradigm. *Electroencephalogr Clin Neurophysiol*, 75(2), 22 - 35.
- Francis, W., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston, MA: Houghton Mifflin.
- Gratton, G., Kramer, A. F., Coles, M. G. H., & Donchin (1989). Simulation studies of latency measures of the event-related brain potential. *Psychophysiology*, 26(2), 233-248.
- Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3), 269-299.
- Hunt, R. R. (2006). The concept of distinctiveness in memory research. In: R. R. Hunt, & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 3-26). New York: Oxford University Press
- Karis, D., Fabiani, M., & Donchin, E. (1984). "P300" and memory: Individual differences in the von Restorff effect. *Cognitive Psychology*, 16(2), 177 - 216.
- Otten, L. J., & Donchin, E. (2000). Relationship between P300 amplitude and subsequent recall for distinctive events: Dependence on type of distinctiveness attribute. *Psychophysiology*, 37, 644-661.
- Paller, K. A., & Wagner, A. D. (2002). Observing the transformation of experience into memory. *TRENDS in Cognitive Sciences*, 6(2), 93-102.
- Ritter, W., Vaughan, H. G., & Costa, L. D. (1968). Orienting and habituation to auditory stimuli: A study of short term changes in average evoked responses. *Electroencephalography and clinical Neurophysiology*, 25, 550-556.
- Rundus, D. (1971). Analysis of rehearsal procedures in free recall. *Journal of Experimental Psychology*, 89(1), 63-77.
- Rushby, J. A., Barry, R. J., & Johnstone, S. J. (2002). Event-related potential correlates of serial-position effects during an elaborative memory test. *International Journal of Psychophysiology*, 46(1), 13-27.
- Sommer, T., Rose, M., & Buechel, C. (2006). Dissociable parietal systems for primacy and subsequent memory effects. *Neurobiology of Learning and Memory*, 85(3), 243-251.
- Spencer, K. M., Abad, E. V., & Donchin, E. (2000). On the search for the neurophysiological manifestation of recollective experience. *Psychophysiology*, 37, 494-506.
- Spencer, K. M., Dien, J., & Donchin, E. (1999). A componential analysis of the ERP elicited by novel events using a dense electrode array. *Psychophysiology*, 36, 409-414.
- Squires, K., Wickens, C., Squires, N., & Donchin, E. (1976). The effect of stimulus sequence on the waveform of the cortical event-related potential. *Science*, 193(4258), 1142-1146. doi: 10.1126/science.959831
- Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-Potential Correlates of Stimulus Uncertainty. *Science*, 150(3700), 1187-1188.
- Von Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschung*, 18, 299-342.
- Wickens, C., Kramer, A., Vanasse, L., & Donchin (1983). Performance of concurrent tasks: A psychophysiological analysis of the reciprocity of information-processing resources. *Science*, 221(4615), 1080-1082.
- Wiswede, D., Rüsseler, J., & Münte, T. F. (2007). Serial position effects in free memory recall -- an ERP study. *Biological Psychology*, 75, 185-193.

# Modeling Learning of Relational Abstractions via Structural Alignment

Subu Kandaswamy (subu@u.northwestern.edu)

Kenneth D. Forbus (forbus@northwestern.edu)

Qualitative Reasoning Group, Northwestern University, 2133 Sheridan Road  
Evanston, IL 60201 USA

## Abstract

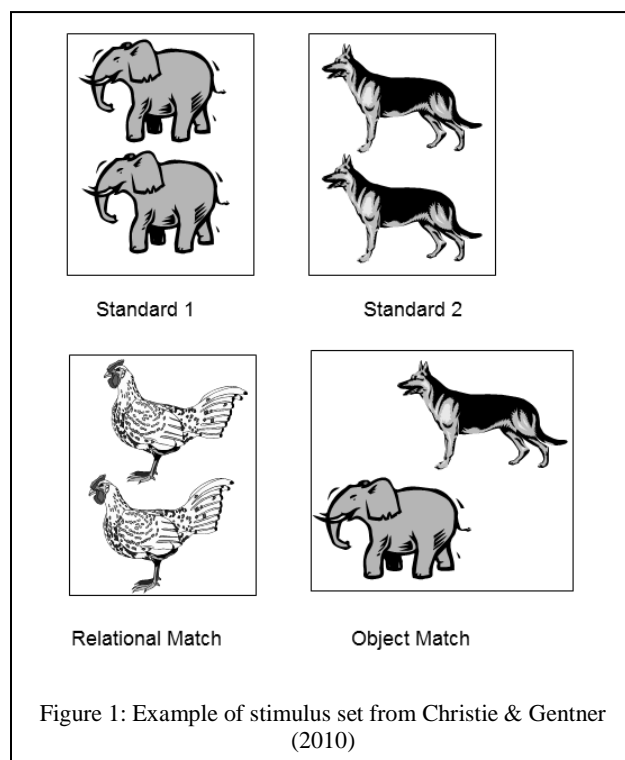
Learning abstract relationships is an essential capability in human intelligence. Christie & Gentner (2010) argued that comparison plays a crucial role in such learning. Structural alignment highlights the shared relational structure between compared examples, thereby making it more salient and accessible for subsequent use. They showed that 3-4 year old children who compared examples in a word-extension task showed higher sensitivity to relational structure. This paper shows how a slight extension to an existing analogical model of word learning (Lockwood et al 2008) can be used to simulate their experiments. This provides another source of evidence for comparison as a mechanism for learning relational abstractions.

## Introduction

Our ability to abstract and reason with relations between objects is an essential part of our intelligence. As children, we acquire a variety of relations, including spatial relations such as *above*, and *on*, and functional relations like *edible* and *dangerous*. How children acquire and use such relational abstractions is an important question in cognitive development. Gentner (2003) has argued that comparison promotes learning new relational abstractions. The idea is that structural alignment highlights common structure, which becomes more salient and available for subsequent use.

One line of evidence for this theory comes from an experiment by Christie & Gentner (2010). To show that children (ages 3-4) were learning new relations, they used novel spatial relational categories in a word extension task, as illustrated in Figure 1. Here the relationship might be characterized as “An animal above another identical animal”. In the Solo condition, children were shown a single standard (here, Standard 1) and told it was a novel noun (e.g. “Look, this is a jiggy! Can you say jiggy?”). In the Comparison condition, children were invited to compare two examples (e.g. “Can you see why these are both jiggies?” when presenting Standard 1 and Standard 2 simultaneously). In both conditions, children were then presented with a forced-choice task, where they had to choose which one of the alternatives is a jiggy (e.g. “Which one of these is a jiggy?” when presented with the relational match and object match cards). Children in the Solo condition preferred the object match, while those in the Comparison condition chose relational matches twice as often as object matches. This provides evidence that comparison can lead to learning new relational abstractions. In a second experiment, a third condition, Sequential, was added, where children saw two standards serially, to test

whether or not simple exposure to more examples was sufficient to promote learning. They found significant differences between Sequential and Comparison, and between Solo and Comparison, but the difference between Sequential and Solo were not significant. This provides additional evidence that it is comparison specifically that is promoting learning.



This paper shows that this phenomena falls out of computational models of analogical generalization already proposed for word learning. We start by summarizing the necessary background, including the models of analogical matching and generalization used and how we use sketch understanding to reduce tailorability by producing portions of the input representations automatically. Next we describe an extension to a similarity-based word learning model (Lockwood et al 2008) that enables it to model this task. Then we describe two simulation experiments that demonstrate that this model is capable of exhibiting behavior consistent with that described in Christie & Gentner (2010), including sensitivity analyses to shed light on why it does so. After discussing related work, we close with a discussion of future work.

## Background

Our model is based on Gentner’s (1983) structure-mapping theory of analogy and similarity. In structure-mapping, comparison involves a *base* and *target*, both structured, relational representations. Comparison results in one or more *mappings*, which contain a set of *correspondences* that describe how the elements in the structured representations align, a *score* that indicates the overall structural quality of the match, and possibly *candidate inferences* representing knowledge that could be projected from base to target (as well as from target to base). Our computational model of comparison is the *Structure-Mapping Engine*, SME (Falkenhainer et al 1989; Forbus et al 1994). Here SME is used both as a component in our model of analogical generalization (described below) and in making the decision in the forced-choice task. The score is normalized to be between zero and one, by dividing it by the score obtained for the maximum self-mapping of base and target.

Analogical generalization is modeled via SAGE (*Sequential Analogical Generalization Engine*), an extension of SEQL (Kuehne et al 2000) which incorporates probabilities and analogical retrieval. Information about concepts is stored in *generalization contexts* (Friedman & Forbus, 2008). Each generalization context maintains a set of examples of that concept, plus generalizations concerning it. Examples are provided incrementally. For each new example, the most similar prior examples and generalizations are retrieved via a model of analogical reminding (MAC/FAC, Forbus et al 1995). The retrieved items are compared, via SME, with the new example. For each comparison, if the score of the best mapping is over a threshold (the *assimilation threshold*), the compared items are assimilated into a generalization – a new one in the case of two examples, or an updated version of the existing generalization in the case of an example and a generalization. The assimilation process keeps the common structure of the mapping, replacing non-identical entities with abstract place-holders. Associated with each fact in generalizations is a probability, based on the number of times a statement aligning with it appears in an example (Halstead & Forbus, 2005). For example, in a generalization about swans, the fact that swans are birds might have a probability of 1.0, while the probability that their color is white might be 0.999 while the probability that their color is black might be 0.001. Non-overlapping facts are kept, albeit given a low probability (i.e.,  $1/N$ , where  $N$  is the number of examples assimilated into that generalization). Facts whose probability drops below the *probability cutoff* are removed from the generalization. SAGE is the central component in our word-learning model, as explained below.

Tailorability is an important problem in cognitive simulation. To reduce tailorability, we use automatically constructed visual and spatial representations. These representations are computed by CogSketch (Forbus et al 2011), an open-domain sketch understanding system. CogSketch uses models of visual and spatial processing to

compute qualitative relationships from digital ink. For example, it automatically computes topological relationships (e.g. inside, touching) and relative positions (e.g. above, right of) for the entities in a sketch. It also includes a model of mental rotation, which uses SME to first do a qualitative shape match which then guides a quantitative match (Lovett et al 2009). This enables it to compute relationships such as *sameShape*, *reflectedOnXAxis*, and so on. Conceptual information can be introduced by adding attribute information to entities in the sketch. For example, the top entities in Standard 1 of Figure 1 might be described as identical elephants, one positioned above the other. The attributes are derived from a large, independently-developed knowledge base<sup>1</sup>. The relationships automatically computed by CogSketch, along with the conceptual attributes provided for an entity, provide the inputs for our simulation. Moreover, CogSketch provides a mechanism for dividing a sketch into *subsketches*, which is what we use to combine all of the elements of a stimulus set onto the same sketch, for convenience. Figure 2 provides an example of a sketched stimulus set.

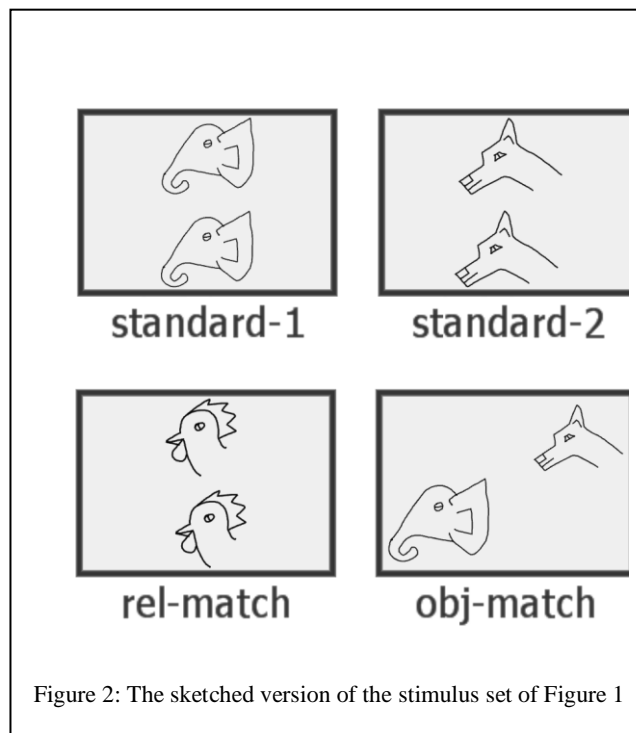


Figure 2: The sketched version of the stimulus set of Figure 1

<sup>1</sup> OpenCyc, see [www.opencyc.org](http://www.opencyc.org)

## Word Learning via Analogical Generalization

We model the learning of words as follows. For each word, there is a generalization context. Every time the word is used, an appropriate subset of the world is encoded to capture information about what that word denotes, and is provided to the generalization context for that word as an example. The generalizations constructed by SAGE can be considered as the meanings for the words. Note that such meanings can be probabilistic, since SAGE computes frequency information for every statement in the generalization. The ability to track multiple generalizations provides a mechanism for handling multiple senses of a word. The ability to store unassimilated examples provides a means of handling edge cases, and helps provide noise immunity in the face of changes in the underlying distribution of examples of a concept.

This account has been used to successfully model spatial propositions of contact in English and in Dutch (Lockwood et al 2008). It makes no commitment to the particulars of encoding, because this is a complex issue, especially since evidence from studies of novice/expert differences suggest that encoding strategies evolve with learning (Chi et al 1981). When using CogSketch as a source of stimuli, we use an entire subsketch as the relevant material to encode.

## Simulation Experiments

Now let us see how this model can be used to simulate the experiments of (Christie & Gentner, 2010). We begin with their Experiment 1.

### Simulation Experiment 1

Recall that Experiment 1 used two conditions to show children the new concept, followed by a forced-choice task. We model these as follows:

- **Forced-choice task:** Each of the choices is used with the generalization context for the word to retrieve the most similar generalization or example for that choice. The choice whose similarity score is highest constitutes the decision. For simplicity, We start with an empty generalization context for every novel word used.
- **Solo Condition:** The single example is added to the generalization context for the word.
- **Comparison Condition:** The two examples are added to the generalization context, but since the experimenter has asserted that they are both examples of the concept, we assume that the child is more likely to assimilate them into a generalization, which is modeled by lowering the assimilation threshold from its default of 0.8 to 0.1. We also assume that the probability cutoff is 0.6, so that facts which do not appear in the shared structure will be eliminated from the generalization.

The original experiment used 8 stimulus sets. We encoded 8 sketches of animals, using CogSketch. Each element of the stimulus set (e.g. Standard 1, Standard 2, etc.) was drawn as a separate subsketch. CogSketch’s default encoding methods were used, plus an additional query to ascertain if any of the entities in a subsketch had the same shape as any other, and if so, what transformation held between them (where no transformation implies the relationship `sameShape`). Moreover, filters were used to automatically remove three types of information: Redundant information (e.g. `given (rightOf B A), (leftOf A B)` is redundant), irrelevant information (e.g., global estimates of glyph size like `MediumSizeGlyph`), and bookkeeping information (e.g. relationships describing timestamps). The table below shows the final encoding for the sketches stimulus set (Figure 2) and the resultant generalization.

Table 1: Encoding for the sample sketch.

Standard-1	Standard-2
<pre>(sameShapes Object-99                   Object-420) (above Object-99         Object-420) (isa Object-420 Elephant) (isa Object-99 Elephant)</pre>	<pre>(sameShapes Object-104                   Object-425) (above Object-104         Object-425) (isa Object-425 Dog) (isa Object-104 Dog)</pre>
Generalization for “jiggy”	
<pre>(above (GenEntFn 1 0 jiggy) (GenEntFn 0 0 jiggy)) (sameShapes (GenEntFn 1 0 jiggy) (GenEntFn 0 0 jiggy))</pre>	

An interesting open parameter concerns the number of conceptual attributes that children might be encoding. While we suspect that a large number of attributes would be encoded<sup>2</sup>, we do not know of data that provides specific estimates. Consequently, we perform a sensitivity analysis by running the simulation while varying the number of conceptual attributes to ascertain their impact on the results. Specifically, we varied the number of attributes from zero to nine. We assumed that encoding is reasonably uniform, i.e. that the same attributes would always be computed for identical objects. For simplicity, we further assumed that the set of attributes computed for one entity had no overlap with the set of attributes computed for another entity whose shape is different. Given these assumptions, we used synthetic attributes (e.g. `Uniquetandard-1MtAttribute8`) for convenience.

Figure 3 shows the results. From the data, we can see that the model chose the relational match 100% of the time for the Comparison condition. This is qualitatively consistent with the behavior of participants in the Comparison condition, where participants chose the relational match around 60% of the time. We believe that the lack of object matches in this simulation condition are due to the use of completely independent attributes for each entity type in the

<sup>2</sup> See the Specificity Conjecture (Forbus & Gentner 1989).

stimuli sets. Since they are independent, no attributes are left in the generalization after assimilation. The more overlapping attributes there are, the more likely an object match is to become possible.

Returning to Figure 3, in the Solo condition, as the number of attributes rises, the proportion of object matches rises (i.e., the proportion of relational matches falls). Again, this provides a good qualitative fit for the results of (Christie & Gentner 2010) Experiment 1. Since attributes are more salient to children, due to lack of relevant domain knowledge (Ratterman & Gentner, 1998), it is reasonable to assume that they would encode more attributes than relations, which is compatible with the simulation results.

Recall that we assume that the probability cutoff is set high enough that non-overlapping information is immediately filtered out. (Since these are novel concepts, there can be at most two examples in any generalization, and hence the probability of any fact not in the overlap would be 0.5, which is less than the 0.6 threshold.) Would adding in probabilistic information improve the fit of the model to human data? To determine this, we tried changing the probability cutoff to its usual default of 0.2. This leads to all attributes remaining in the generalization, which results in the score for the object match being boosted so high that it always wins over the relational match, regardless of the experimental condition used. This suggests that when children are invited to compare, they do indeed restrict themselves to keeping exactly the overlapping structure.

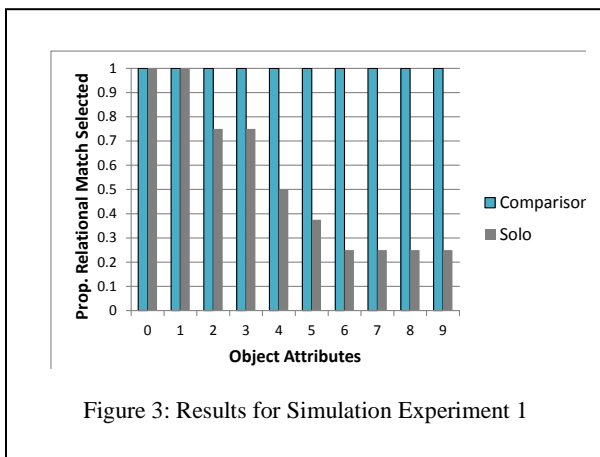


Figure 3: Results for Simulation Experiment 1

## Simulation Experiment 2

Experiment 2 in (Christie & Gentner 2010) actually consists of two experiments. Both involved a new condition, the Sequential condition, designed to rule out non-comparison explanations. In Experiment 2a, fillers, in the form of pictures of familiar objects, were interposed between the serial presentation of the standards. No invitation to compare was issued. In Experiment 2b, no fillers were used, and the Solo and Comparison conditions from Experiment 1 were added, by way of replication. In our

model, fillers would be added to some other generalization context, thus 2a and 2b look identical from the perspective of our model. We model the new condition as follows:

- **Sequential Condition:** The two examples are added to the generalization context, but with the default assimilation threshold 0.8.

Again we varied the number of conceptual attributes, in the same way as in Simulation Experiment 1.

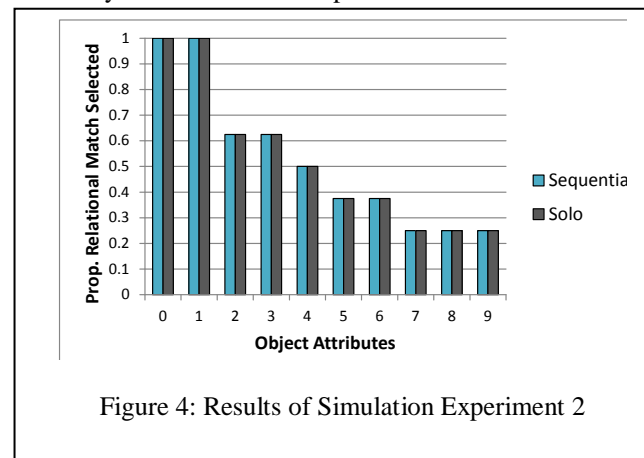


Figure 4: Results of Simulation Experiment 2

Figure 4 illustrates the results. As anticipated, the results for the Sequential condition are identical to the results the model generates for the Solo condition. This is because of the model does not generalize the two standards, and hence the choices will be compared to the exemplars in the generalization context. This makes the results of the

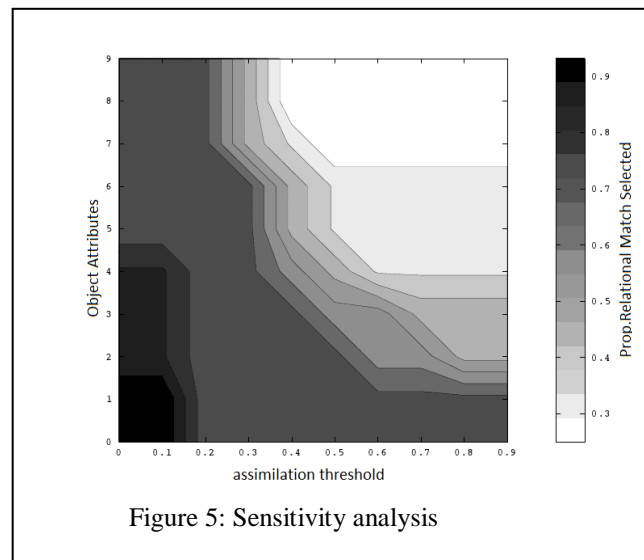


Figure 5: Sensitivity analysis

Comparison condition be the same as the Solo condition.

We know of no psychological evidence that would provide constraints on the value of the assimilation threshold. Consequently, we performed a sensitivity analysis by varying the assimilation threshold between 0.1 and 0.9, while varying the number of attributes from zero to nine. Figure 5 illustrates the results. The region marked as



black indicates a high proportion of relational match choices and then the contour fades down gradually.

The slope of the contour indicates that the model readily generalizes the standards when both the assimilation threshold and the number of object attributes are low. This can be interpreted as follows. A low assimilation threshold corresponds to a higher willingness to accept the standards as belonging to the same category, which fits the assumptions of our model. A low number of object attributes indicates a leaner encoding i.e. not enough attention was paid to the object, or it may be unfamiliar. This is a second possible explanation for why some children chose the relational match for the Sequential condition.

## Related Work

There have been several prior computational models of word learning. For example, Siskind (1996) developed an algorithm for cross-situational learning of word/meaning mappings. He used synthetic conceptual representations and lexicons to examine its scaling properties and noise immunity. Our use of arbitrary predicates is similar to his use of synthetic conceptual representations, but our visual representations are grounded in prior cognitive science research. It is an open question whether Siskind's algorithm would work on realistic conceptual representations, and similarly, it is an open question as to whether our word learning algorithm can scale to the size of vocabularies that his does. Another model, described in (Roy and Pentland 2002) uses speech and vision signals as input, to tackle the problem of how children learn to segment these perceptual streams while at the same time learning word meanings. A particularly novel aspect of their approach was modeling a corpus of infant-directed speech they gathered, to ensure their inputs were naturalistic. Our use of sketch understanding is motivated by the hypothesis that it forces us to incorporate high-level vision, while factoring out most of the complexities of signal processing. The relatively crude visual processing techniques used in Roy & Pentland's system, compared to mammalian vision systems, suggests that theirs, too, is an approximation, albeit a more signal-rich version than ours. Neither of these models, nor any other word learning model that we are aware of, has tackled the role of comparison in learning relational abstractions.

## Discussion

We have shown that a model of word learning based on analogical generalization, using automatically encoded sketches augmented by conceptual information, can simulate the behavior found in (Christie & Gentner 2010). The invitation to compare, we argue, leads the child to aggressively attempt to form a generalization between the two new exemplars, as modeled by lowering the assimilation threshold and only keeping overlapping structure. This finding is robust across a wide range of choices for the number of object attributes. Moreover, serial presentation to the model, as with humans, does not lead to

relational learning, as measured by responses in the forced-choice task.

There are several lines of future work that suggest themselves. First, we intend to explore if the model can handle closely related phenomena (e.g. Gentner & Namy 1999, who used a similar experimental paradigm but with pre-existing concepts instead of novel concepts). Second, we plan on exploiting more of CogSketch's automatic encoding capabilities, by using it to automatically decompose object-level spatial descriptions into edge-level representations. The sketches in the stimuli will be represented by a set of constituent edges, their attributes and relations that hold between them (e.g. (`isa edge2 StraightEdge`) (`edgesParallel edge2 edge4`)). For example, a square can be segmented into four constituent edges. These more detailed spatial representations will contain more shared attributes and relations and hence would naturally introduce more overlap between entities. This would provide a test of our hypothesis that such overlap is responsible for participants in the Comparison condition sometimes choosing the object match. Finally, we plan to extend this model to explore how object labels promote uniform relational encoding and re-representation (Gentner, 2010).

## Acknowledgments

This research was supported by the Spatial Intelligence and Learning Center (SILC), an NSF Science of Learning Center, SBE-1041707, and by a grant from the Socio-Cognitive Architectures Program of the Office of Naval Research.

## References

- Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5: 121-152.
- Christie, S. & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, 11 (3). 356-373.
- Falkenhainer, B., Forbus, K. and Gentner, D. (1989). The Structure Mapping Engine: Algorithm and examples. *Artificial Intelligence*, 41, 1-63.
- Forbus, K., Ferguson, R. and Gentner, D. (1994). Incremental structure-mapping. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, August.
- Forbus, K., Gentner, D., and Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19, 141-205.
- Forbus, K. and Gentner, D. (1989). Structural evaluation of analogies: What counts? *Proceedings of the Cognitive Science Society*.
- Forbus, K., Usher, J., Lovett, A., and Wetzel, J. (2011). CogSketch: Sketch understanding for Cognitive Science Research and for Education. *Topics in Cognitive Science*. pp 1-19.



Friedman, S. & Forbus, K. (2008). Learning Causal Models via Progressive Alignment & Qualitative Modeling: A Simulation. *Proceedings of the 30th Annual Conference of the Cognitive Science Society (CogSci)*.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.

Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp.195-235). Cambridge, MA: MIT Press.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science* 34,752-775.

Halstead, D. and Forbus, K. (2005). Transforming between Propositions and Features: Bridging the Gap. *Proceedings of AAAI-2005*. Pittsburgh, PA.

Kuehne, S., Forbus, K., Gentner, D. and Quinn, B. (2000). SEQL: Category learning as progressive abstraction using structure mapping. *Proceedings of CogSci 2000*, August.

Lockwood, K., Lovett, A., and Forbus, K. (2008). Automatic Classification of Containment and Support Spatial Relations in English and Dutch. *Proceedings of Spatial Cognition*.

Lovett, A., Tomai, E., Forbus, K. & Usher, J. (2009). Solving geometric analogy problems through two-stage analogical mapping. *Cognitive Science*, 33(7), 1192-1231.

Ratterman, M., & Gentner, D. (1998) More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task. *Cognitive Development*, 13, 453-478.

Roy, D. and Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26,113-146.

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition* 61,39-91.

# From Hands to Minds: Gestures Promote Action Understanding

**Seokmin Kang (sk2587@columbia.edu)**

Teachers College, Columbia University  
New York, NY 10027 USA

**Barbara Tversky (btversky@stanford.edu)**

Columbia Teachers College  
New York, NY 10027 USA  
Stanford University  
Stanford, CA 94305 USA

**John B. Black (jbb21@columbia.edu)**

Teachers College, Columbia University  
New York, NY 10027 USA

## Abstract

Understanding dynamic concepts is more difficult than understanding static ones. The present study showed that understanding dynamic concepts can be enhanced by gestures that convey action. Participants learned how an engine worked from one of two videos, with identical verbal scripts and identical diagrams. One video was accompanied by gestures showing the structure of the system; the other was accompanied by gestures showing the actions of the system. Both groups learned the basics of the system. Participants who saw the action gestures depicted more dynamic information in their visual explanations of the system and included more dynamic information in their verbal explanations of the system. Because they are inherently dynamic, gestures appear to be especially suited for conveying dynamic information.

**Keywords:** gesture; diagram; complex systems; knowledge construction

## Knowledge in the hands

When people explain something, they typically use gestures as well as speech. Gestures can carry information that is redundant with speech, reinforcing the message by presenting information in two modalities. Importantly, gestures sometimes carry information that is not carried in speech (e. g., Bavelas, 1994; Church & Goldin-Meadow, 1986; Perry, Church, & Goldin-Meadow, 1988). In some cases, speech refers listeners to gesture, as in “turn this way,” but in other cases, there is no cuing of the gestures. Nevertheless, the information carried solely in gesture can reveal the thought of speakers and affect the thought of both those who make gestures and those who watch them (e.g., Beattie & Shovelton, 1999; Chu & Kita, 2011; Goldin-Meadow, Cook, & Mitchell, 2009; Goldin-Meadow, Kim, & Singer, 1999; Hegarty, Mayer, Kriz, & Keehner, 2005; Kessel & Tversky, 2006; Singer & Goldin-Meadow, 2005; McGregor, Rohlfing, Bean, & Marschner, 2009; Ping & Goldin-Meadow, 2008; Schwartz & Black, 1996; Thompson, Driscoll, & Markson, 1998; Valenzano, Alibali, & Klatzky, 2003).

It is primarily iconic and deictic gestures that reveal the thought of those who make them and affect the thought of those who make them or observe them. Deictic gestures

point to places or things in the world or in a virtual world. Iconic gestures show what something looks like or acts like (e.g., McNeill, 1992; Goldin-Meadow, 2003). Together, these kinds of gestures can carry rich semantic content. A train of integrated deictic and iconic gestures can be used on virtual stages to create detailed models of situations, such as environments (e. g., Emmorey, Tversky, & Taylor, 2000) and actions, such as how a lock works (e. g., Engle, 1998). Are such gestures successful in communicating knowledge as well as in representing it?

## Knowledge on the page

As such, sequences of organized gestures can serve much like diagrams. In fact, many kinds of gestures can be mapped to kinds of diagrammatic features; that is, they carry the same meanings (Tversky, Heiser, Lee, & Daniel, 2009). Diagrams have some advantages over gestures as a means of representing knowledge. Diagrams have permanence, so they can be inspected and reinspected. Because they are external and persist, they do not need to be kept in mind, so the mind is free to use the diagram as a basis for reorganization, for inference, and for discovery. Diagrams use elements and spatial relations on a page to represent elements and relations that are actually spatial, as in maps or architectural plans, or that are metaphorically spatial, as in the periodic table or organization charts (e. g., Tversky, 2011; Tversky, et al., 2009). Gestures, like language, are external, but lack permanence. A series of gestures used to create a model of a situation requires working memory to create, understand, and remember, and can tax working memory. On the other hand, diagrams are static, so it can be challenging to convey action, change, and process in diagrams. Typically, arrows are used, but they can be ambiguous (e. g., Heiser & Tversky 2006; Tversky, 2011; Tversky, Heiser, MacKenzie, Lozano, & Morrison, 2007). Gestures are by nature dynamic, so they can portray action, if schematically (e. g., Kita & Özyürek, 2003; Rizzolatti & Arbib, 1998). In fact, when gestures are used with diagrams in explanations, diagrams are often used to convey structure, and gestures to portray action (e. g., Engle, 1998).

## Complex systems

Many explanations, in conversational as well as learning situations, are of complex systems, scientific, mechanical, social, athletic, or political. Complex systems typically have elements--actors or agents or object--that have properties and structure, social or geographic or other relations. They also have action or behavior: the actors or agents or objects act or are acted on in some sort of systematic ways usually associated with their properties and their relationships or structure. Many complex systems, from traffic patterns to election procedures, from spread of disease to workings of the nervous system, from the operation of an engine to a court of law, can be explained, especially when accompanied by deictic and iconic gestures. They can also be diagrammed, and, as noted, diagrams readily portray the structural relations of agents, actors, and objects, but do not easily portray the action or behavior of systems. Yet, it is the action of a system and its outcomes that is hardest for novices to comprehend (e. g., Hmelo-Silver and Pfeffer, 2004). Making inferences about action or function separates novices and experts across domains (e.g., Suwa & Tversky, 1997). Here we ask whether gestures showing action can promote understanding of the behavior of complex systems.

To ask whether iconic gestures that convey action can promote understanding of explanations of complex systems, we compared explanations that were identical except for gesture. One explanation was accompanied by gestures that portrayed action, and a control explanation used gestures to convey the form and structure of the parts of the system. Students viewed one of two videos of explanations of the operation of a four-stroke engine, the typical engine in an automobile. The language of the explanations was identical, and each explanation was based on a diagram of the structure of the engine superimposed to the front and side of the explainer. Because enactive gestures can convey action directly and information about action is more difficult, we were especially interested to know if gestures conveying action help students comprehend action information.

Performance was assessed in several ways: by questions about structure and action, by diagrams, by visual explanations, and by live explanations of the systems by the students. The questions could be answered solely on the basis of the language of the explanations and served partly as a manipulations check. Hence, if students who view action gestures have a better understanding of the action of the system than those who viewed structure gestures, they should be more likely to include action information, in their diagrams, and they should be more likely to deliver action information and use action gestures themselves in their later explanations to new learners.

## Method

**Participants** 59 (15 male) university students ranging in age from 20 to 36 with an average age of 26 ( $SD = 3.50$ ), participated in the study. They were all native English speakers and did not have prior knowledge of the system to be learned.

**Materials** We created two videos explaining how a four-stroke engine works. The videos were identical in language and number of gestures but differed in kinds of gesture. A diagram typical of those in science and engineering showing the labeled parts and configuration of the system was superimposed in front and to the side of the explainer. The explanations began with an introduction overviewing the structure using deictic gestures. The core portion of the explanation was a step-by-step explanation of the processes comprising the workings of the system. The final portion of the explanation explained how the process caused the car's wheels to rotate. Because the diagram showing the structure was always in view and because the introduction to both explanations overviewed the system structure, the gestures emphasizing structure served as a control and were not expected to affect performance on the questions.

For the core portion of the explanation, in the action video, the explainer used only gestures that portrayed the action of each part, always in the same location, so no structural information was provided. In the control structure video, the explainer used only gestures that pointed to the location of the parts of the system and showed the shape of each part as the process was explained. The accompanying verbal script explained both the locations of the parts and the actions of the parts identically. Figure 1 shows snapshots of two instructional videos.

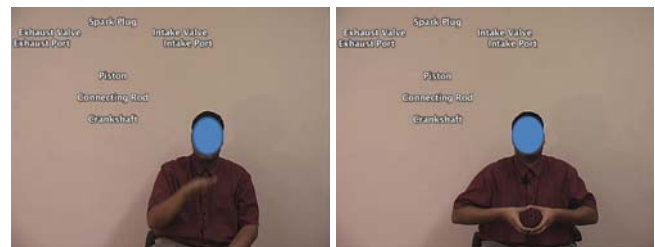


Figure 1. Still shots from the *action* (left) and *structure* (right) videos showing the superimposed diagrams.

The information in the script was categorized as structure or action, and gestures appropriate for each were devised. For the action gesture video, the explainer showed the rotational motion of the crankshaft, the direction of the piston's movement, the flow of fuel and air, the movement of the intake and exhaust valves, and so on with his hands. The action gestures were performed in the same place off the diagram, avoiding any positional information.

For the structure gesture video, the explainer used his hand(s) successively to show the shapes of the crankshaft, piston and cylinder, and showed the positions of the piston, crankshaft, spark plug, intake port, intake valve, exhaust port, exhaust valve, and mixture of fuel and air.

To eliminate any biasing effects of lexical stress (Heuven, 1988; Field, 2005), the speaker practiced the script several times, making sure to stress the actions and the parts for both videos.

**Posttests** The verbal posttest was based on the information in the script with 20 recognition questions, 16 True/False, and 4 multiple-choice questions. Of the 16 True/False

questions, 8 queried action and 8 queried structure. Action questions referred to movement, or causal relations of the parts and their consequences. Structure questions referred to shapes and positions of the parts of the system. Four multiple-choice questions queried general knowledge. The questions were presented in random order. Because the verbal posttest was based entirely on the verbal script, differences dependent on viewed gesture were not expected. The test served as a manipulation check.

The second posttest was a diagramming task. Participants were asked to diagram a visual explanation of how a four stroke engine works based on what they learned from the video. Finally, participants made a video to explain the workings of the four-stroke engine to a peer. It was expected that participants who viewed the videos with action gestures would include more action information in the latter two less-constrained measures.

**Procedure** Participants were seated at a table with a laptop computer with a 15.4 in screen. They were randomly assigned to either the action gesture or the structure gesture video group. The participants were then told: “Today, your job is to watch a video of how a four stroke engine works four times<sup>1</sup> in a row and explain the concept in the video to a peer coming later. However, since you are not directly explaining a concept, your explanation will be videotaped and showed later either to him or her. He or she will learn about the concept from your explanation.” Participants were not allowed to take notes or to pause or stop the video. The experimenter left the room while participants watched the video. After watching the video, participants were given the verbal and diagrammatic posttests, and then made a video explaining the system to a peer. The video camera was set opposite the participant 3 meters away. Participants were allowed to spend as much time as they wanted.

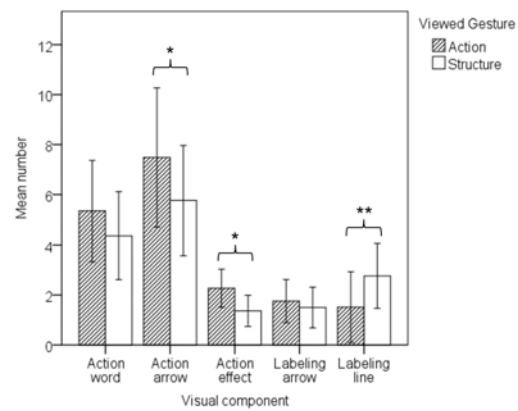
## Results

**Verbal Posttest** As expected, the type of gesture viewed yielded no differences in performance on action ( $p = .08$ ), structure ( $p = .85$ ) or general ( $p = .92$ ) questions, nor were there interactions between gesture viewed and question type,  $F(1, 114) = 1.70$ ,  $p = .20$ . However, in within group comparisons, those who viewed action gestures performed better on action questions than on structure questions,  $t(28) = 3.56$ ,  $p < .01$ ,  $d = 0.82$ . There were no differences between action and structure questions for those who viewed gestures conveying structure ( $p = .11$ ).

**Diagram Posttest** Two coders coded the diagrams for action components. The reliability for action words was  $Kappa = .56$  ( $p < .001$ ), for action arrows,  $Kappa = .63$  ( $p < .001$ ), for action effects,  $Kappa = .65$  ( $p < .001$ ), for labeling arrows,  $Kappa = .60$  ( $p < .001$ ), and for labeling lines,  $Kappa = .73$  ( $p < .001$ ). Action effects were depictions of actions, such as explosions. The means of the visual components by type of viewed gesture appear in Figure 2.

<sup>1</sup> A pilot study had revealed that two viewings were insufficient to achieve above chance performance.

The means were compared using Poisson regression analysis with the assumption that the conditional means equal the conditional variances.



\*  $p < .05$ , \*\*  $p < .01$

Figure 2. Mean number of visual components in diagrams. Error bars represent standard errors of the means.

Overall, those who viewed action gestures used more visual components than those who saw structure gestures ( $p < .05$ ). In addition, those who viewed action gestures produced more action arrows ( $p < .05$ ) and action effects ( $p < .05$ ) and labeled fewer lines ( $p < .01$ ) than those who saw structure gestures. Labeled lines typically linked names and parts; that is, structural information. Thus, for the diagrams, those who saw action gestures included more information about action and those who saw structure gestures included more information about structure, showing that the viewed gestures affected viewers' comprehension and later production.

**Explanations to a peer** Recall that after learning the system, participants made videos explaining the four-stroke engine to novices. Will those who saw action gestures use more of them in their own explanations? A gesture unit was defined as “the period of time between successive rests of the limbs (McNeill, 1992).” If the hands did not return to a resting position between two gestures, the boundary was defined by a pause in motion and an obvious change in shape or trajectory. If participants used both hands simultaneously to describe one object, concept, or part, it was regarded as one gesture. If participants used one hand to describe an object, a concept, or a part and the other hand a different concept, the gestures were coded as two different gestures.

For this study, only gestures conveying action or structure were coded. Action gestures were defined as showing the action of a part or process of a system. Structure gestures were defined as showing the location or static properties, notably shape, of objects or parts of the system. Inter-rater reliability was assessed on randomly selected 240 subsets (18%) of the data by a second coder who was trained and blind to the experimental design. Agreement for identifying gestures was 87.8% and for categorizing gestures was 99.6%.

For the speech analysis, the participants' verbal descriptions were segmented into propositions (following

Heiser & Tversky, 2006). The information units were coded as *action*, *structure*, or *other*. Propositions that contained action such as movement of each part within a cylinder were coded as *action* information, for example, “...that byproduct is pushed back up through the exhaust valve...”. Propositions that contained ‘is-a’ or ‘has-a’ were coded as *structure* information unless they referred to action, for example, “...then it has an exhaust valve”. *Other* information included greetings, such as “Good evening,” introductory information such as “I’m going to explain how a four stroke engine works,” and meta-comments such as “...let me tell you a little bit more about each stage...”

**Gesture analysis**<sup>2</sup> The average explanation time was 177.14 sec ( $SD = 56.84$ ) for the action group and 152.34 sec ( $SD = 55.94$ ) for the structure group (ns.  $p = .10$ ). There were a total of 1306 gestures: 754 by those who had viewed action gestures, 552 by those who had viewed structure gestures (ns.  $p = .13$ ). The means of action and structure gestures produced by participants who viewed action and structure videos are shown in Figure 3.

There was an interaction between type of gesture viewed and type of gesture produced,  $F(1,112) = 8.58, p = .004 < .01, \delta = .84$ . In within group comparison by paired sample t-test, even though participants in both groups delivered more action gestures than structure gestures, the action group ( $t(28) = 7.15, p < .0001, d = 1.49, r = .60$ ) reliably used more action gestures, when compared to the structure group ( $t(28) = 2.88, p = .008 < .01, d = 0.58, r = .28$ ).

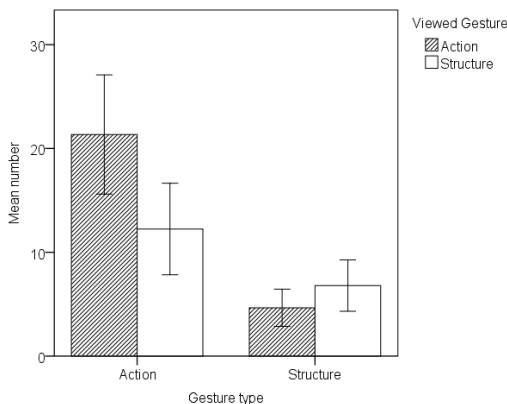


Figure 3. Mean number of type of produced gesture by type of viewed gesture. Error bars represent standard errors of the means.

It is possible that the number and the pattern of gestures differed by length of explanation. Although there were no significant differences in the overall gesture use, explanations in the action group were longer. Consequently, the next section presents more detailed analyses of the results by gesture rate, that is, gestures per minute.

<sup>2</sup> One participant’s explanation was not recorded because of malfunctioning of a video camera. Therefore, 58 participants’ videos were analyzed.

The same pattern of gesture use was observed. The action group used 6.89 ( $SD = 4.13$ ) action gestures per minute and 1.45 ( $SD = 1.35$ ) structure gesture per minute. The structure group used 4.62 ( $SD = 3.49$ ) action gestures per minute and 2.51 ( $SD = 2.38$ ) structure gesture per minute.

In group comparison, there was an interaction such that the action group used more action gestures and the structure group used relatively more structure gestures,  $F(1,112) = 8.83, p = .004, < .01, \delta = .84$ . In within group comparison, when compared to the structure group ( $t(28) = 3.08, p = .005 < .01, d = 0.71, r = .33$ ), the action group ( $t(28) = 7.95, p < .0001, d = 1.77, r = .66$ ) reliably used more action gestures than structure gestures.

**Speech analysis** The participants delivered a total of 2550 information units in their speech. Among them, 1607 conveyed *action* information, 737 *structure* information, and 206 *other* information. Those who saw action gestures delivered a total of 1425 information units. Among them, 929 conveyed action, 387 conveyed structure, and 109 conveyed other information. Those who saw structure gestures delivered a total of 1125 information units, 678 conveying action, 350 conveying structure and, 97 conveying other information. Figure 4 shows mean number of information units delivered by two groups.

Overall, those who viewed action gestures ( $M = 49.14, SD = 20.81$ ) delivered more information units than those who viewed structure gestures ( $M = 38.79, SD = 17.22$ ),  $F(1,56) = 4.25, MSE = 364.75, p = .044 < .05$ . In addition, those who viewed action gestures delivered more action information than the structure group,  $F(1,56) = 6.87, MSE = 158.03, p = .01 < .05$ . There were no differences in the quantity of structural information ( $p = .52$ ) or other information ( $p = .66$ ), but there was an interaction between kind of information and kind of viewed gesture ( $p = .02 < .05$ ). Post hac tests (Tukey HSD) showed that more action information was given than structural information ( $p < .001$ ) and more structure information than other information ( $p < .001$ ).

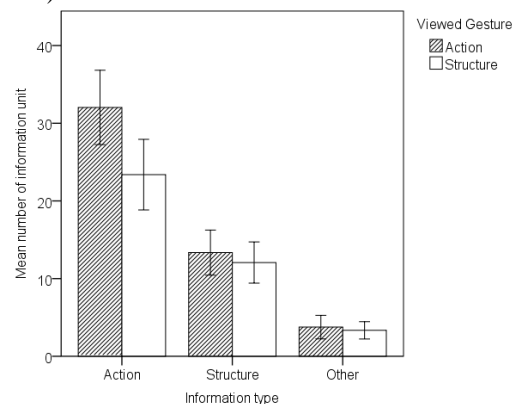


Figure 4. Mean kinds of information units by viewed gestures. Error bars represent standard errors of the means.

**Proportion of information types in speech** Although there were no group differences in explanation time, the group

who had viewed action gestures took more time and delivered more information units than the group who viewed structure gestures. To take that into account, percentages of information types were analyzed and appear in Figure 5.

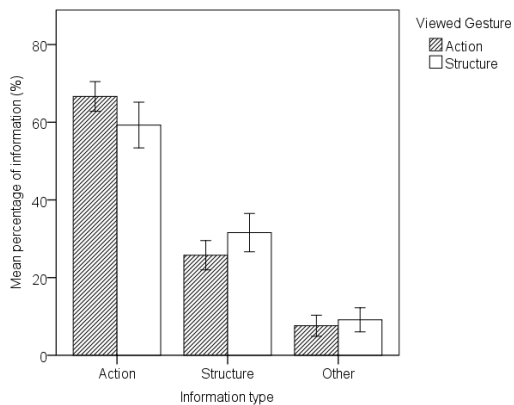


Figure 5. Mean percentage of kinds of information units by viewed gesture. Error bars represent standard errors of the means.

For those who viewed action gestures, action information accounted for an average of 66.62% ( $SD = 10.30$ ), structure information accounted for an average of 25.76% ( $SD = 10.13$ ), and other information accounted for an average of 7.61% ( $SD = 7.30$ ). For those who viewed structure gestures, 59.28% ( $SD = 15.89$ ) was action information, 31.59% ( $SD = 13.28$ ) was structure information, and 9.14% ( $SD = 8.34$ ) was other information. There was an interaction between group and information type,  $F(2,168) = 5.16$ ,  $MSE = 126.74$ ,  $p = .007 < .01$ . Those who viewed action gestures delivered relatively more action information than those who viewed structure gestures and those who viewed structure delivered relatively more structure information than those who viewed action group. Thus, in their own explanations, those who had viewed action gestures produced both more verbal information about action and showed more action in their gestures. Similarly, those who had viewed structure gestures used more structure gestures and included proportionately more verbal structure information than those who had viewed action gestures.

### Discussion

Understanding the behavior of complex systems is challenging (e. g., Hmelo-Silver & Pfeffer, 2004). Actions are not apparent in static diagrams, and the nature of actions often has to be imagined from purely symbolic language. Animations are typically too complex and too fleeting to be comprehended (e. g., Tversky, Morrison, & Betrancourt, 2002) and are not part of most natural settings for explanations. There is abundant evidence that gestures provide a rich source of information, including information about structure and process (e. g., Beattie et al., 1999; Becvar, Hollan, & Hutchins, 2008). Here we asked if gestures can successfully transmit dynamic information, over and above verbal and diagrammatic explanations,

simply and abstractly at a pace that allows comprehension. We taught a complex system, the operation of a four-stroke engine, to novices under two conditions. One group saw action gestures that conveyed the behaviors of the parts of the system; the other group saw structure gestures that conveyed static qualities of the parts of the system and their structure. Both groups heard exactly the same explanation and saw the same structure diagram of the parts of the system. The verbal explanation was sufficient to convey the basics of the structure and the dynamics of the engine. A number of posttests were administered: a verbal test based on the verbal explanation, a visual explanation task, and a videotaped explanation of the system to new novices.

The verbal memory test showed that both groups adequately learned the essentials of the structure and the operation of the system. However, the diagramming and the explanation tasks revealed substantial differences in the understanding of the behavior of the systems; the group who had viewed the action gestures appeared to have a deeper understanding of the behavior of the system than the group who had viewed the structure gestures. In the visual explanation task, those who had seen action gestures depicted more specific actions of the system than the group who had viewed the structure gestures. Furthermore, the group who had viewed the action gestures used more action gestures in their videoed explanations than the group who had viewed the structure gestures. Although the increase in the number of action gestures in explanations might be attributable at least in part to imitation of what they had viewed, the increase in number of depictions of specific actions cannot. The depictions of action must come from a deeper understanding of the specific chain of behaviors of the system. Moreover, many of the gestures used differed from those viewed.

The effects of the viewing the gestures that conveyed the structure of the system were weaker but evident both in diagrams and in explanations. The structure of the system was apparent from the diagram that was displayed during the viewed explanation, and the structure of the system was described in the verbal portion of the explanation. Furthermore, the structural information is easier than the behavioral because it was apparent in the diagram.

In both groups, gestures conveying action far outnumbered gestures conveying structure, suggesting that participants regarded the behavior of the system as paramount and regarded gesture as a good means for conveying action, over and above language.

Discourse in the wild, including explanations, is an integrated combination of word, gesture, and props, elements in the world (such as a diagram) or in a virtual world that can be continuously referred to during the course of the discourse. Each, word, gesture, prop, plays roles, sometimes overlapping, sometimes complementary. Understandably, actions, even miniature schematic ones as those in gestures, appear to be especially effective for conveying action, another example of cognitive congruence (e. g., Tversky, et al., 2002).



**Acknowledgments.** We are grateful to National Science Foundation HHC 0905417, IIS-0725223, IIS-0855995, and REC 0440103 for partial support.

## References

- Bavelas, J. B. (1994). Gestures as part of speech: Methodological implications. *Research on Language and Social Interaction*, 27, 201-221.
- Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica*, 123, 1-30.
- Becvar, A., Hollan, J., & Hutchins, E. (2008). Representational gestures as cognitive artifacts for developing theories in a scientific laboratory. In M. S. Ackerman, C. A. Halverson, T. Erickson, & W. A. Kellogg (Eds.) *Resources, Co-Evolution and Artifacts: Theory in CSCW* (pp. 117-143). London, England: Springer-Verlag.
- Chu, M. & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology*, 140, 102-116.
- Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition*, 23, 43-71.
- Emmorey, K., Tversky, B., & Taylor, H. (2000). Using space to describe space: Perspective in speech, sign, and gesture. *Journal of Spatial Cognition and Computation*, 2, 157-180.
- Engle, R. A. (1998). Not channels but composite signals: Speech, gesture, diagrams, and object demonstrations are integrated in multimodal explanations. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 321-326). Mahwah, NJ: Erlbaum.
- Field, J. (2005). Intelligibility and the Listener: The Role of Lexical Stress. *TESOL Quarterly*, 39, 399-423.
- Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge: Belknap Press.
- Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. A. (2009). Gesturing gives children new ideas about math. *Psychological Science*, 20, 267-272.
- Goldin-Meadow, S., Kim, S., & Singer, M. (1999). What the teacher's hands tell the student's mind about math. *Journal of Educational Psychology*, 91, 720-730.
- Hmelo-Silver, C. E., & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science*, 1, 127-138.
- Hegarty, M., Mayer, S., Kriz, S., & Keehner, M. (2005). The role of gestures in mental animation. *Spatial Cognition and Computation*, 5, 333-356.
- Heiser, J., & Tversky, B. (2006). Arrows in comprehending and producing mechanical diagrams. *Cognitive Science*, 30, 581-592.
- Heuven, V. J. van (1988). Effects of stress and accent on the human recognition of word fragments in spoken context: gating and shadowing. *Proceedings of the 7th FASE/Speech-88 Symposium*, edited by W.A. Ainsworth and J.N. Holmes, 811-818. Edinburgh: Institute of Acoustics.
- Kessell, A. M. and Tversky, B. (2006). Gestures for thinking and explaining. *Proceedings of the meetings of the Cognitive Science Society*.
- Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Cognition*, 48, 16-32.
- McGregor, K. K., Rohlfing, K. J., Bean, A., & Marschner, E. (2009). *Journal of Child Language*, 36, 807-828.
- McNeill, D. (1992). *Hand and mind*. Chicago: University of Chicago Press.
- Perry, M., Church, R. B. & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development*, 3, 359-400.
- Ping, R., & Goldin-Meadow, S. (2008). Hands in the air: Using ungrounded iconic gestures to teach children conservation of quantity. *Developmental Psychology*, 44, 1277-1287.
- Rizzolatti, G., & Arbib, M. A. (1998) Language within our grasp, *Trends in Neurosciences*, 21, 188-194
- Schwartz, D. L., & Black, J. B. (1996). Shuttling between depictive models and abstract rules. *Cognitive Science*, 20, 457-497.
- Singer, M. A., & Goldin-Meadow, S. (2005). Children learn when their teachers' gestures and speech differ. *Psychological Science*, 16, 85-89.
- Suwa, M., & Tversky, B. (1997). What architects and students perceive in their sketches: A protocol analysis. *Design Studies*, 18, 385-403.
- Thompson, L. A., Driscoll, D., & Markson, L. (1998). Memory for visual-spoken language in children and adults. *Journal of Nonverbal Behavior*, 22, 167-187.
- Tversky, B. (2011). Visualizations of thought. *Topics in Cognitive Science*, 3, 499-535.
- Tversky, B., Heiser, J., Lee, P., & Daniel, M.P. (2009). Explanations in gesture, diagram, and word. In Coventry, K. R., Tenbrink, T., & Bateman, J. (Eds.), *Spatial Language and Dialogue*. Oxford: Oxford University Press.
- Tversky, B., Heiser, J., MacKenzie, R., Lozano, S., and Morrison, J. B. (2007). Enriching animations. In R. Lowe and W. Schnotz, *Learning with animation: Research implications for design*. NY: Cambridge University Press.
- Tversky, B., Morrison, J. B. & Betrancourt, M (2002). Animation: Can it facilitate? *International Journal of Human Computer Studies. International Journal of Human Computer Studies*, 57, 247-262.
- Valenzeno, L., Alibali, M. W., & Klatzky, R. (2003). Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology*, 28, 187-204.



# An experimental investigation of consistency of explanation and graph representation

Nana Kanzaki (kanzaki@cog.human.nagoya-u.ac.jp)  
Graduate School of Information Science, Nagoya University, Japan

Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)  
Graduate School of Information Science, Nagoya University, Japan

## Abstract

Many previous studies of graph comprehension have confirmed that information gleaned from a graph is greatly influenced by its representation. When explaining data with a graph, writers/researchers must generate graphs whose representation is consistent with the explanation contents. In the current study, we defined those who engage in academic activities using graphs on a daily basis as expert graph users and investigated whether they and undergraduates (non-experts) can adaptively generate a consistent graph with explanations from the viewpoint of the consistency of the contents and graph representation. Experiment 1 indicated that expert graph users adaptively generate a graph whose structure is consistent with the explanation contents. On the other hand, Experiment 2 suggests that undergraduates cannot do so. But in Experiment 3 undergraduates were supported by selecting graphs from provided candidates, but there was a limited concordance between the type of explanation and graph representation.

**Keywords:** Diagrammatic representation; Graph; Explanation.

## Introduction

Many previous studies confirmed that using diagrams is effective for understanding information. In a pioneering psychological paper, Larkin & Simon (1987) theoretically demonstrated the efficacy of using diagrams while solving a problem and suggested that diagrammatic representation simplifies access to relative information more than sentential representation does, and transforms the cognitive processes into more efficient ones.

Norman (1991) pointed out that task difficulties depend on such visual representation as figures and tables and promote problem solving performance. He also suggested that a cognitively consistent correspondence between internal representation and external world is crucial.

In this study, we experimentally investigated the consistency of the contents of the explanations and the representations of graphs for the explanation. Kosslyn (2006), who marshaled psychological findings about graph design, argued that graph representation must be examined, especially based on the human cognitive system for effectively conveying information. And also, Hegarty (2011) mentioned the importance of cognitive science to design visual-spatial displays.

Graphs, which are pictures that convey logical relationships among numbers for a specific purpose, include line graphs, bar graphs, step graphs, and pie charts. Specific

examples of graph usage are demonstrated in the *Publication Manual of the American Psychological Association* (2001).

Experimental studies of graph comprehension confirmed that the information from a graph is greatly influenced by its representation. Graphs can be represented in various forms. Different representations generated from an identical data set elicit different interpretations of the graphs. For example, in studies of inferences from bar and line graphs, viewers are more likely to describe x-y trends when viewing line graphs than bar graphs (Zacks & Tversky, 1999; Shah, Mayer, & Hegarty, 1999). Peebles and Cheng (2003) suggested that the comprehension time of certain information differs depending on the graph structure.

Although many studies on graph comprehension have been conducted, graph generation is also an important issue. In the current study, we deal with the issue of arrangements of variables on graphs when generating them. Shah and Carpenter (1995) confirmed that x-y trends were comprehended easily, although z-y trends were comprehended with difficulty using three-variable line graphs (e.g., Carpenter & Shah, 1998; Kanzaki & Miwa, 2011). The reason may be that visual chunks are constructed for every line in a line graph, and the graph is interpreted based on each chunk.

Consider the case indicated in Table 1 where a line graph is depicted from the data that consist of two independent Variables, A and B, and one dependent variable, Variable C. Two types of graphs are considered. One is a graph in which Variable A is put on the x-axis and Variable B on the z-label (Figure 1(a)), and on the other Variable B is put on the x-axis and Variable A on the z-label (Figure 1(b)). According to Larkin & Simon (1998), these two graphs are informationally equivalent but computationally different; they are equivalent because they are constructed from identical data sets, but they are different since different chunks are constructed in each of the graphs and the information may be read in different ways. When explaining

Table 1: Variable C vs. Variables A and B.

		Variable B		
		10	30	50
Variable A	10	20	50	80
	30	50	50	50
	50	80	50	20

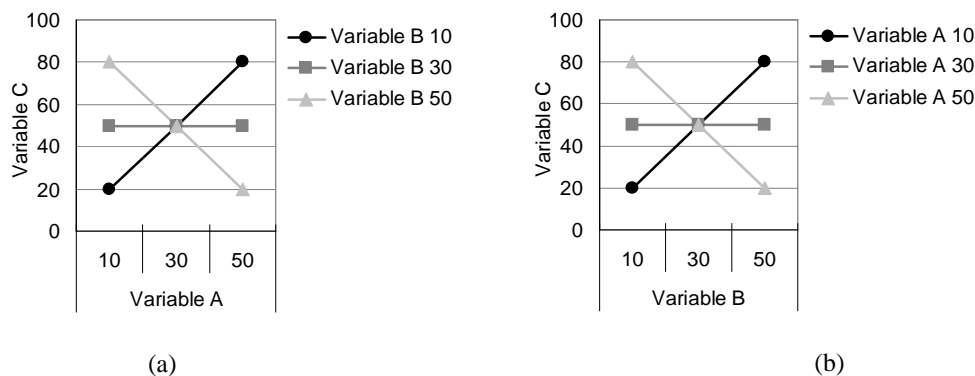


Figure 1: Graphs made from Table 1.

the data of Table 1, one of two alternative graphs must be generated whose structure is consistent with the explanation.

Consider a more specific situation in which we must adjust independent Variable B with an uncontrollable change of Variable A to intentionally control the quantity of dependent Variable C. When explaining how to manage Variable B in this situation, the explanation is classified into two types from the viewpoint of the representation of Variable A. In one explanation, Variable A is described discretely. In this classification, the explanation is generated for each level of Variable A, such as “when Variable A is at ..., you should set Variable B to ...” In the other explanation, Variable A is described continuously. In this classification, the explanation reflects whether its value is larger than a specific value, such as “when Variable A is larger than ..., set Variable B to ...”, or from the viewpoint of a continuously changing variable A, such as “according to the increase of Variable A.”

When two independent variables are placed on a line graph, one variable is commonly put on the x-axis and the other on the z-label. In such a situation, it is contemplated that the variable on the x-axis is regarded as a continuous variable, and the one on the z-label is regarded as a discrete variable because the x-axis is represented as a continuous factor and the z-label as a discrete factor.

Kosslyn (2006) noted that “the continuous rise and fall of a line is psychologically compatible with the continuous nature of an interval scale. . . Time, temperature, and amount of money are measured using an interval scale.” Additionally, Kanzaki & Miwa (2011) suggested that an explanation based on each line increases the comprehension of a line graph because a chunk of each line is generated (e.g., Carpenter & Shah, 1998; Shah & Carpenter, 1995). These studies indicate that a variable on the z-label in a line graph is regarded as a discrete variable.

The above investigations propose a hypothesis, where in a normative graph, an independent variable is put on the x-axis when the variable is regarded as a continuous variable in the explanation; in contrast, an independent variable is put on the z-label when it is regarded as a discrete variable.

If this hypothesis is correct, we predict the following when we must explain how to adjust independent Variable B, with the uncontrollable change of independent Variable A to control the quantity of dependent Variable C:

1. Participants who treated Variable A as a continuous variable in their explanation put it on the x-axis (Figure 1(a)).

2. Participants who treated Variable A as a discrete variable in their explanation put it on the z-label (Figure 1(b)).

In the current study, expert graph users participated in Experiment 1, whom we defined as those who daily engage in academic activities using graphs. Our first objective is to confirm whether such experts adaptively generate graphs that are consistent. We expected them to do so because they have much experience giving presentations with graphs and reading them in research papers.

Our second aim is to perform a similar experiment with undergraduates as novices. Undergraduates who have not received systematic training in statistics participated in our second and third experiments. We propose a hypothesis that they may have trouble generating consistent graphs when they are required to adaptively generate graphs based on understanding such highly abstract mathematical concepts as continuousness and discreteness.

Recently, various types of software have been developed for making graphs. User can automatically generate them by simply choosing some properties. In this situation, users select a graph rather than generate one. The third objective is to examine whether undergraduates can select a consistent graph when they are presented alternative candidates of consistent graphs.

## Experiment 1

Experiment 1 investigated whether people who use graphs daily can adaptively generate consistent graphs when constructing an explanation with them.

We set a situation in which either “air temperature” or “humidity” was adjusted to promote the growth of “newly discovered mushrooms.” Two dependent variables, air

temperature and humidity, can be treated either as continuous or discrete. An independent variable is the “amount of mushroom growth.” The table used in our experiments is the same as Table 1, but Variables A, B, and C were replaced with specific factors: humidity, air temperature, and amount of growth. The shape of the line graph generated from the table is also the same as in Figure 1. These two graphs’ shapes were controlled to be the same in order that the ease of constructing explanation should not be affected by the shapes of the graphs.

In our experiments, we set two situations for a particular explanation context. In one situation, participants explained how to promote mushroom growth by adjusting the humidity, where air temperature was not controllable. In the other situation, they did so by adjusting the air temperature, where the humidity was not controllable. These two situations were counter-balanced in the experimental procedure.

## Method

**Participants** The participants in Experiment 1 were either university associate professors or doctoral students in experimental psychology. 22 researchers participated as experts. 17 had Ph. Ds. All participants had published one or more peer-reviewed academic journal papers.

**Procedure** The experiment was performed individually or in small groups. The participants were presented the table shown in Table 1 and given the following instructions:

“A new kind of mushroom was recently discovered whose growth greatly depends on air temperature and humidity. Its growth at specific temperatures and humidities is shown in this table.”

Half of the participants were given a situation where the amount of growth was controlled by adjusting the humidity, but the temperature was not controllable. For these participants, the following instructions were given:

“You are a salesperson of mushroom seedlings. Your customers can adjust the humidity in their mushroom greenhouses, but they cannot adjust the temperature. Explain how to grow the mushrooms by adjusting the humidity with uncontrollable changes of temperature. Use a line graph in your explanation.”

For counter-balance manipulation, the other half was given a situation where the humidity could not be adjusted, and temperature was replaced by humidity in the instructions.

They wrote their explanation in ten minutes and then drew a graph on an experimental sheet shown in Figure 2 in five minutes, labeling the x-axis and the z-legend by themselves.

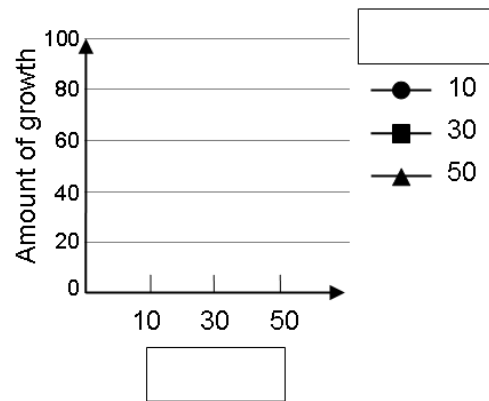


Figure 2: Graph format used in Experiments 1 and 2.

**Classifying generated graphs** The generated graphs were classified by the placement of an unadjustable variable in the graphs. The following were the classification criteria:

- (1) *X-axis unadjustable* graph: an unadjustable variable is put on the x-axis
- (2) *Z-legend unadjustable* graph: an unadjustable variable is put on the z-legend

**Classifying explanations** The participant explanations were classified depending on whether the unadjustable variable was described as a continuous or a discrete variable. The following were the classification criteria:

- (1) *Continuous explanation*: an unadjustable variable is described continuously. The following are example explanations in this category for a situation where the temperature is not adjustable: “When the temperature is above 30°C,” “according to the increase of temperature,” and so on.
- (2) *Discrete explanation*: an unadjustable variable is described discretely. Example explanations in the same situation include, “When the temperature is at 10°C,” “when the temperature is low,” and so on.

When both types of descriptions appeared in an explanation, the classification was made based on the description that was part of the conclusion. Such descriptions were usually seen in the last part of the explanation.

## Results and discussions

The participants were grouped depending on whether they generated *continuous* or *discrete explanations*. Those who generated *continuous explanations* were classified as the *continuous explanation* group, and those who generated *discrete explanations* as the *discrete explanation* group. Ten of the 22 participants were categorized in the *continuous explanation* group and the other twelve in the *discrete explanation* group.

Figure 3 shows the proportions of the graphs classified into each category in the two groups.

To examine whether the structure of the generated graphs was influenced by the described explanation, a two difference in the distribution ( $\chi^2(1, N=22) = 6.42, p < .05$ ).

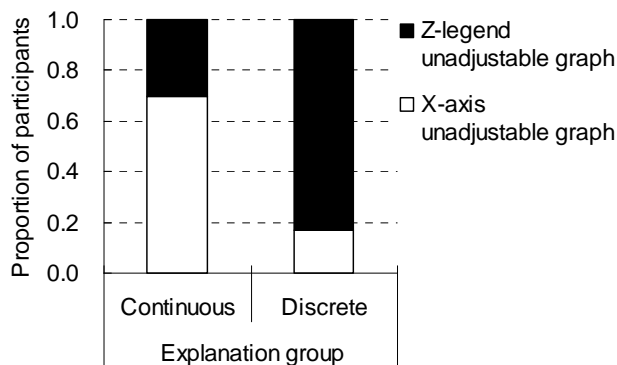


Figure 3: Proportions of participants who generated *x-axis* or *z-legend unadjustable* graphs in Experiment 1.

(*continuous* and *discrete* explanation groups) x two (*x-axis* and *z-legend* generated graphs) chi-square test was performed on their distribution. There was a significant A residual analysis shows that both the numbers of the participants who generated *x-axis unadjustable* graphs in the *continuous explanation* and *z-legend unadjustable* graphs in the *discrete explanation* group were greater than the expected values (the residual value was 2.53). On the other hand, both the numbers of participants who generated *z-legend unadjustable* graphs in the *continuous explanation* group and *x-axis unadjustable* graphs in the *discrete explanation* group were less than the expected values (the residual value was -2.53).

This result suggests a tendency to generate specific line graphs in which an unadjustable variable was put on the x-axis when it was regarded as a continuous variable in an explanation. On the other hand, there was a tendency to generate graphs in which an unadjustable variable was put on the z-legend when it was regarded as a discrete variable. This result implies that the participants, who use graphs on a daily basis, adaptively generate graphs whose structures are consistent with their explanations.

## Experiment 2

In Experiment 1, we confirmed that expert graph users adaptively generate graphs whose structures are consistent with their explanations. In Experiment 2, we performed the same investigation with undergraduates who have little experience of making graphs when explaining something.

### Method

In the following, descriptions are omitted about the same procedures as in Experiment 1.

**Participants** 44 undergraduate Liberal Arts majors who had not completed a course in statistics participated. Half were given a situation where the growth was controlled by adjusting the humidity, but the temperature was not controllable, and for counter-balance manipulation, the

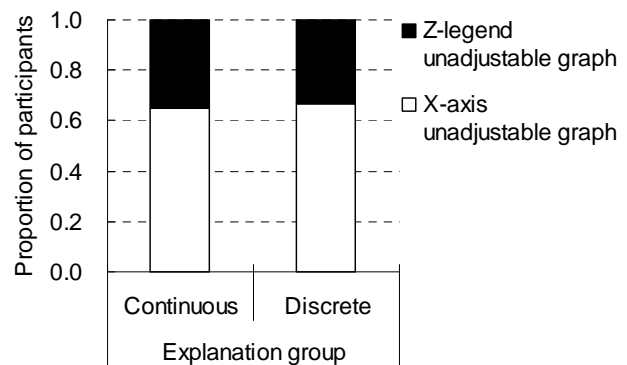


Figure 4: Proportions of participants who generated *x-axis* or *z-legend unadjustable* graphs in Experiment 2.

other half was given a situation where the growth was controlled by adjusting the temperature, but the humidity was not controllable.

**Procedure** The procedure of Experiment 2 was the same as that of Experiment 1.

### Results and discussions

Seventeen of the 44 participants were categorized in the *continuous explanation* group and the other 27 in the *discrete explanation* group.

Figure 4 shows the proportions of graphs classified into each category in the two groups.

To examine whether the structure of the generated graphs was influenced by the described explanation, a two (*continuous* and *discrete* explanation groups) x two (*x-axis* and *z-legend* generated graphs) chi-square test was performed on the distribution of the generated graphs. There was no significant difference in the distribution ( $\chi^2(1, N=44) = 0.02, ns$ ). This result suggests that undergraduates cannot generate consistent graphs.

## Experiment 3

In Experiment 2, we confirmed that undergraduates did not necessarily make graphs whose structure is consistent with their explanations. In Experiment 2, the participants generated graphs by themselves. But in Experiment 3, we gave them two candidates of consistent graphs and let them select one to investigate whether they could adaptively select a consistent graph from two alternatives.

### Method

In the following, descriptions are omitted about the same procedures as in Experiment 2.

**Participants** 57 undergraduate Liberal Arts majors who had not completed a statistics course participated. 29 were given a situation where the growth was controlled by adjusting the humidity, but the temperature was not controllable, and for the counter-balance manipulation, the other 28 were given a

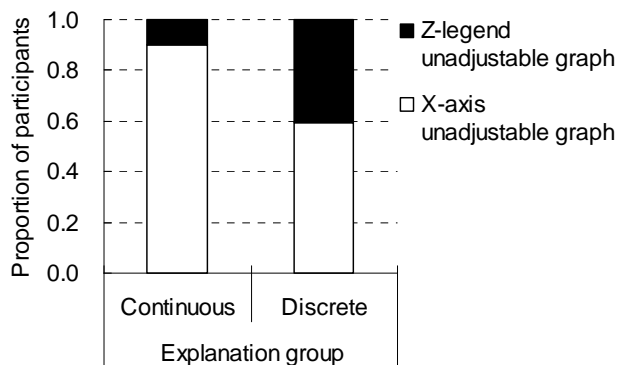


Figure 5: Proportions of participants who selected x-axis or z-legend unadjustable graphs in Experiment 3.

situation where the growth was controlled by adjusting the air temperature, but the humidity was not controllable.

**Procedure** The experiment was performed as part of their class assignments. The participants wrote a script for the given situation and presented the table used in Experiments 1 and 2. They were also presented two types of line graphs made from the table, such as those in Figure 1, and required to select a suitable one.

## Results and discussions

Five participants were excluded from analysis because they failed to follow the instructions. 20 were categorized in the *continuous explanation* group and the other 32 in the *discrete explanation* group.

Figure 5 shows the proportions of graphs classified into each category in the two groups.

To examine whether the structure of the selected graphs was influenced by the described explanation, a two (*continuous* and *discrete* explanation groups) x two (*x-axis* and *z-legend* selected graphs) chi-square test was performed on their distribution. There was a significant difference ( $\chi^2(1, N=52) = 5.62, p < .05$ ). A residual analysis shows that both the numbers of the participants who selected *x-axis unadjustable* graphs in the *continuous explanation* group and *z-legend unadjustable* graphs in the *discrete explanation* group were greater than the expected values (the residual value was 1.96). On the other hand, both the numbers of the participants who selected *z-legend unadjustable* graphs in the *continuous explanation* group and *x-axis unadjustable* graphs in the *discrete explanation* group were less than the expected values (the residual value was -1.96).

This result implies that undergraduates tended to adaptively select consistent graphs if alternatives were presented to them. However, in Experiment 3, the proportions of participants who selected *x-axis unadjustable* graphs were greater than those in the result of the expert graph users in Experiment 1. We discuss this point below.

## Discussion and conclusions

In this study, we investigated whether expert graph users and undergraduates make consistent graphs from given data to construct explanations. The result of Experiment 1 confirmed that expert graph users successfully generated graphs in relation to their explanations. On the other hand, Experiment 2 showed that undergraduates failed to indicate this tendency.

We discuss the results in the light of our hypothesis proposed in the introduction by analyzing the proportions of each type of graphs in each explanation group.

In the result of Experiment 1 (Figure 3) in the *discrete explanation* group, the proportion of *z-legend unadjustable* graphs was significantly larger than that of the *x-axis unadjustable* graphs ( $p = .019$ , one-tailed Fisher's exact tests). This means that the clearly greater use of a discrete factor in the graph was due to the influence of the greater use of discrete expressions in the explanation. In other words, there was a consistency between the explanation and the graph. On the other hand, in the *continuous explanation* group, although the proportion of *x-axis unadjustable* graphs was larger than that of the *z-legend unadjustable* graphs, there was no significant difference in their proportions ( $p = .172$ , one-tailed Fisher's exact tests). Three of the ten experts in the *continuous explanation* group put the variable described as a continuous one on the z-legend, contrary to our expectation. When analyzing their explanations, two of the three generated explanations based on the slopes of the lines. For example, in the humidity unadjustable situation, "when the humidity is lower than 30%, the mushroom growth proportionally increases with the air temperature because the line slope is positive." Since a line slope is described based on each label of the z-legend, and a line slope is represented continuously in a line graph, they put the continuous variable on the z-legend. The above investigation supports that, as a whole, expert graph users adaptively generated consistent graphs both in the *continuous* and *discrete explanation* groups.

Next, we discuss the graph generation by undergraduates in Experiment 2 (Figure 4). To generate consistent graphs, we must examine what should be represented based on deep consideration of explanations (Kosslyn, 2006). Our result implies that undergraduates tend to use graphs without such consideration.

On the other hand, when selecting a graph in Experiment 3, undergraduates were given choices. Such given choices might enable them to consider the consistency of their explanation and graph representation.

As Norman (1992) noted: "to think of the problem of designing something that people will find understandable and easy to use as the same problem as writing something that other people will understand and find easy to read." The process of generating a graph for conveying information resembles the process of writing. A cognitive writing model proposed by Hayes & Flower (1980) consists of *planning*, *translating* (text production), and *reviewing*. Similarly, the process of generating a graph also involves *planning*, in

which people discuss how to present information, *translating* (depicting), in which people depict a graph, and *reviewing*, in which people review whether the graph is appropriate for their purpose. Studies on writing point out the troubles of the *planning* phase in novices: e.g., not considering situations and objectives for explanation (Flower & Hayes, 1980) and tending to ignore planning (Carey, Flower, Hayes, Schriver, & Haas, 1989).

When selecting a graph from the candidates in Experiment 3, the undergraduates were only required to *review* a graph in the light of explanation without *planning* and *translating*. This may explain why more consistent graphs were selected in Experiment 3 than in Experiment 2.

In Experiment 3 (Figure 5), although we confirmed that the proportions of each type of graphs adaptively changed depending on the explanation contents, as a general tendency, the proportion of *x-axis unadjustable* graphs exceeded that of the *z-legend unadjustable* graphs. This tendency was also shown in Experiment 2 (Figure 4). These results suggest that undergraduates tended to put an unadjustable variable on the x-axis without considering their explanations.

In this study, we defined the normative consistency of explanation and graph representation from the viewpoint of the variable's continuousness and discreteness. Scrutiny is needed to confirm whether such graph consistency actually promotes understanding. Such investigation remains important future work.

Finally, even if the automatic generation of graphs with spreadsheet software simplifies their utilization, this study's results indicate that undergraduates still have trouble selecting consistent graphs. On the other hand, presenting candidate graphs improved the selection of consistent graphs, implying that presenting undergraduates with variations of possible graphs and having them consider the relation of what they wish to explain and the candidate graphs may be an effective method in tutoring graph construction.

## References

- American Psychological Association (2001). *Publication Manual of the American Psychological Association*. 5th ed., Washington, DC; American Psychological Association.
- Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology-Applied*, **4**, 75-100.
- Carey, L.J., Flower, L., Hayes, J. R., Schriver, K. A., & Haas, C. (1989). Differences in writers' initial task representations. Technical Report No. 35. (ERIC Document Reproduction Service No. ED310403)
- Hayes, J. & Flower, L. (1980). Identifying the organization of writing processes. In L. Gregg, & E. Steinberg (Eds.), *Cognitive processes in writing*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Hegarty, M. (2011). The cognitive science of visual-spatial displays: implications for design. *Topics in cognitive science*, **3**, 446-474.
- Kanzaki, N., & Miwa, K. (2011). Experimental investigation of effects of representations and contexts on comprehension and generation of line graphs. *Proceedings of 33rd annual conference of the cognitive science society (CogSci 2011)*, 2196-2201.
- Kosslyn, S. M. (2006). *Graph Design for the Eye And Mind*. New York: Oxford University Press.
- Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, **11**, 65-99.
- Norman, D. (1991). Cognitive artifacts. In J. M. Carroll (Ed.), *Designing interaction: Psychology at the human-computer interface*. Cambridge: Cambridge University Press, 17-38.
- Norman, D. (1992). *Turn signals are the facial expressions of automobiles*. Cambridge: Perseus Books.
- Peebles, D., & Cheng, P. C. H. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. *Human Factors*, **45**, 28-45.
- Shah, P., & Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology-General*, **124**, 43-61.
- Shah, P., Mayer, R. E., & Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology*, **91**, 690-702.
- Zacks, J., & Tversky, B. (1999). Bar and lines: A study of graphic communication. *Memory & Cognition*, **27**, 1073-1079.

# The Influence of Virtual Agents' Gender and Rapport on Enhancing Math Performance

Bilge Karacora<sup>1</sup> ([bilge.karacora@stud.uni-due.de](mailto:bilge.karacora@stud.uni-due.de)), Morteza Dehghani<sup>2</sup> ([morteza@ict.usc.edu](mailto:morteza@ict.usc.edu)),  
Nicole Krämer-Mertens<sup>1</sup> ([nicole.kraemer@uni-due.de](mailto:nicole.kraemer@uni-due.de)), Jonathan Gratch<sup>2</sup> ([gratch@ict.usc.edu](mailto:gratch@ict.usc.edu))

<sup>1</sup>Department of Social Psychology, University of Duisburg-Essen,  
Forsthausweg 2, 47057 Duisburg, Germany

<sup>2</sup>Institute for Creative Technologies, University of Southern California,  
12015 Waterfront Dr., Playa Vista, CA 90094-2536, USA

## Abstract

The purpose of the present research is to investigate whether virtual agents can help enhance participants' performance, effort and motivation in mathematics. We hypothesize that a minimal amount behavioral realism induced by display of rapport is necessary for any social effects to occur in human-computer interaction. Further, we examine whether social facilitation effects occur depending on the gender of the participants and the interacting virtual agents. In a 2x2 between subjects design, participants interacted with a male or female virtual agent that either displayed rapport or no rapport. Our results confirm that gender plays a role when interacting with virtual agents that are capable of establishing rapport. Participants' performance and effort were significantly enhanced when interacting with an agent of opposite gender that displayed rapport. Our results have implications on designing agents for education and training purposes.

**Keywords:** social facilitation, STEM, rapport, virtual agents

## Introduction

There is considerable interest in factors that enhance science and math performance. Recently, there has been an upsurge of interest in educational technology that exploits social and motivational factors that enhance math performance in general, and reduce gender inequality in particular (Kim, 2004; Baylor & Ryu, 2003). This work builds on the phenomena that people often treat computers as social actors. Therefore, psychological factors that improve people's performance in traditional face-to-face settings can be simulated by technologies in form of virtual learning companions or virtual instructors. In this paper, we seek to address two related goals. First, we aim to show that certain social psychological phenomena can enhance math performance in a human-computer setting. Specifically, we show that a form of social facilitation can improve performance on standardized math tests. Second, we seek to provide further evidence that people do treat computers as social actors and help elucidate the design principles that foster this effect. We specifically demonstrate that virtual agents must possess a minimum level of behavioral realism to achieve any social effects.

Rapport has been shown as an effective way to create behavioral realism in virtual agents. In social psychology, rapport is described as the establishment of a positive relationship among interaction partners by rapidly detecting and responding to each other's nonverbal behavior (Gratch et al., 2007a). This includes displaying behaviors that indicate positive emotions (such as head nods and smiles), showing mutual attentiveness (such as mutual gaze) and certain coordination behaviors (such as postural mimicry and synchronized movement) (Tickle-Degnan & Rosenthal, 1990). Niewiadomski et al. (2010) reports that when an agent displays appropriate and socially adapted emotional expressions, he is perceived as more human-like than an agent that shows human expressions which are inappropriate or not socially adapted. Garau et al. (2005) conducted a study showing that only participants who interacted with an agent that was responsive to their movements, experienced a sense of personal contact with the agent which influenced them to behave more socially considerate as opposed to interacting with a static or moving but unresponsive agent. This indicates that rapport is an important feature in order for the agent to be perceived as human-like and for any social effects, such as social facilitation, to occur.

Previous research on social facilitation/inhibition illustrates how the presence of others affects an individuals' task performance either positively or negatively (Guerin & Innes, 1982; Zanjong, 1965; Sanders, Baron & Moore, 1978). Whether or not similar facilitation/inhibition effects occur in presence of virtual agents has been subject to several studies. Rickenberg and Reeves (2000) found that tasks are facilitated or inhibited by the "social" presence of a virtual agent. A study by Zambaka et al. (2004) indicates that when asked to perform a task, participants reacted similarly to the presence of a virtual agent as they would have in the presence of another human. A follow-up study by Zambaka et al. (2007) demonstrates that the presence of a virtual agent inhibits the performance of participants on a mathematical task. The limitation of this study was that the sample consisted of only female participants being confronted with an agent of matching gender. Hayes et al. (2010) found a similar decrease in performance with regard



to male participants interacting with an agent of the same gender. However, the study demonstrated that when male participants interacted with an agent of the opposite gender their performance improved. The authors argued that participants experienced a stronger feeling of “being in the room with the agent” when interacting with an agent of opposite gender. A post-hoc explanation for Hayes et al.’s findings is that the improvement may have been caused of social facilitation effects.

## Experiment

Our goal in this paper is to investigate whether a virtual agent can motivate participants and help to improve their performance in a mathematical task. For this purpose, we examine whether social facilitation effects occur when participants interact with virtual agents and how gender and rapport influence these effects. We extend the previous findings of Hayes et al. (2010) by including both, male and female participants, in an experiment in which they interact with a virtual agent of either matching or opposing gender.

We hypothesize that interacting with a human-like rapport agent of the opposite gender facilitates participants’ performance on mathematical tasks. First, we assume that social facilitation effects will cause participants’ to show more motivation, invest more effort and achieve a higher improvement of performance when interacting with an agent of the opposite gender. Second, we further expect such social facilitation effects only to occur when rapport is displayed by the agent. The reason is that rapport has been shown to be necessary for an agent to be perceived as a (human-like) social entity, which is required for social effects to occur.

In order to test these assumptions, we designed an experiment in which we manipulated virtual agents’ gender and rapport behavior. We recruited male and female participants and had them perform two mathematical tasks, one before interacting with an agent and one during the interaction. Each participant was either confronted with an agent of matching gender or of opposing gender. The agents used in the rapport condition were capable of showing appropriate positive responses such as head nods and smiles in reaction to the participants’ verbal and nonverbal behavior. The agents in the no-rapport condition only show minimal unresponsive movements such as breathing or blinking. We explored the effects of our experimental conditions by comparing the participants’ performance before and during the interaction with the agent. We also investigated participants’ motivation and increase of effort to solve the math problems.

## Participants

We recruited seventy-four participants (58.1 % females), from the greater Los Angeles area. Their age ranged from 18 to 34 years with an average age of 23.64 (SD=3.97). 16.2% of participants had high school education, 78.4% collage education and 5,6% went to graduate school. Participants were recruited by responding to recruitment

posters posted on craigslist.com and were paid \$30. The experiment took about 60 minutes.

## Design

We used a 2x2 full factorial between subjects design, with the first variable being the gender of the agent matching the gender of the participant (gender match/gender no match) and the second variable being whether or not the agent displayed rapport (rapport/no rapport). Participants were randomly assigned to one of the four conditions.

Improvement in performance, motivation and increase in effort were measured as dependent variables. We calculated participants’ performance improvement by the difference in their performance before and during the interaction with the agent. For this purpose, we subtracted the number of math problems they solved correctly in the second task from the number they solved correctly in the first task. To measure participants’ motivational state with regards to the mathematical tasks we used the Situational Motivation Scale (SIMS) by Guay et al. (2000). By subtracting the number of solved math problems in the second task from the number solved in the first task, we calculated the increase in effort. This variable was interpreted as an additional indicator for their situational motivation.

### Sample of three questions in the first task:

- A) The child care center charges \$11 an hour plus a daily \$3 drop-off fee. How many hours of childcare did Robert pay for if he dropped his son off 3 days last week and paid \$130 at the end of the week?
- B) A rental store charges a fine of 25 percent of the usual full-day rental rate for each hour that an item is late. Ryan rented a rototiller for one day for \$40. He returned it 3 hours late. How much was his total bill for the rental and the fine?
- C) What is the value of the expression  $3 + 4 \times 6 - 2(5 \times 8)$ ?

### Sample of equivalent questions in the second task:

- A) To do a 12-page report, a word processor charges \$26.50. His fee includes a \$2.50 delivery charge. How much does he charge per page?
- B) Mr. Smith rented a car for 7 days. The car rental charges 500 \$ per week and a fine of 15% for each day the car is returned late plus a handling fee of 15 \$. He returned the car 2 days late. How much did he have to pay in the end?
- C) Evaluate the following expression:  $4(12 - 9)2$ .

Figure 1: Sample of math problems

## Mathematical Tasks

Two mathematical tasks were presented to the participants. Each task consisted of a set of 24 math problems comparable to original GRE and SAT math items (for samples of these questions please refer to Figure 1). The math problems were selected out of a larger set of questions pretested with regard to their difficulty. By pretesting the problems we made sure that both tasks have approximate levels of difficulty and require the same sets of skills. Also, tasks had enough number of questions to prevent ceiling effects due to participants finishing all the math problems in less than ten minutes. To avoid an improvement in performance in the second task due to simple learning/practice, the math problems were modified with regard to their wording and surface features. This way the first and the second task appeared distinct from each other while they still each required the same set of skills.

## Rapport Agent

Participants interacted with a female or male virtual agent with a human-like appearance. Four different characters were used: two male and two female (Figure 2) to control for possible effects of particular agents. We used the Rapport Agent developed by Gratch et al. (2006). To create rapport with the participant, the agent displayed positive listening behaviors (such as nodding and smiling) that correspond to the verbal and nonverbal behavior of a human speaker. Previous studies of the rapport agent have shown that it is highly capable of creating the experience of rapport comparable with a face-to-face condition (Gratch et al.,

2006, 2007a, 2007b). To produce listening behaviors, the Rapport Agent first collects and analyzes audiovisual features from the speaker's voice (silence, speech) and upper-body movements (head nod, smile, eye gaze) in real time. This happens via a microphone and a Videre Design Small Vision System stereo camera, which was placed in front of the participants to capture their movements. Watson, an image-based tracking library developed by Morency (2005), uses images captured by the stereo camera to track the participant's head position and orientation. Acoustic features are derived from properties of the pitch and intensity of the speech signal using a signal processing package, LAUN (Gratch et al., 2006). The Rapport Agent displays behaviors that show that the animated character is "alive" (eye blinking, breathing), and listening behaviors such as posture shifts and head nods automatically triggered by the system corresponding to participants' verbal and nonverbal behavior. This allows the agent to provide contingent feedback while the speaker is speaking by following a response model (Huang et al., 2011) to decide which behavioral response would be most appropriate (such as head nod or smile). The different animations are converted into Behavior Markup Language (BML) (Kopp et al., 2006), send to an action scheduler (to determine the duration of each animation) and passed on to Smartbody, an animation system that blends the different animations naturally into each other (Thiebaux & Marsella, 2007). The commercial game engine Gamebryo then renders the animations and displays them to the user.



Figure 2: Agents used in the experiment

## Procedure

After an explanation of the study and obtaining consent, participants were led to a private room where they completed the experiment individually. Participants were seated at a desk with two monitors that were positioned next to each other. They were then instructed to work on a mathematical task for ten minutes, which was presented as a computer-based survey on one of the monitors. The second monitor was turned off at this point. To minimize self-presentation concerns, the anonymity of task performance results and the non-competitiveness of the task were emphasized. The experimenter left the room for the duration of the working period. Participants' answers were not reviewed by the experimenter and they were not given any performance feedback. Next, the experimenter provided detailed (verbal and written) instructions on the following interaction with the virtual agent. It was emphasized that the agent is a computer program and that participants will be alone during their interaction period. The agent was then launched and displayed on the second monitor. The experimenter would leave the room before the interaction started and then would control the virtual agent's speech over a separate computer in a different location, without the participants' knowledge. The agent's nonverbal behaviors were automated by the system according to the condition (rapport/no rapport) as described above. First, the agent asked how the participants estimated their performance on a 5-point Likert-scale (*very poor* to *very good*). Next, they were asked to rate how difficult they thought the math problems were on another 5-point Likert-scale (*easy* to *hard*). This was followed by the second task period that was part of the interaction. The agent would instruct the participants to load the task on the other monitor and work on it for a time period of ten minutes. During that time the agent reminded the participants twice of the time remaining (5 minutes and 1 minute left) and also let them know when time was up. Afterwards, the agent asked the participants to estimate their performance with regard to the second task on the same scale. At the last part of the interaction, the agent interviewed the participants concerning their experiences while working on the tasks and their attitudes towards

mathematics in general. Then, the agent would announce the end of the interaction and would disappear from the monitor. Finally, situational motivation and demographic variables were measured in a post survey without the virtual agent visible or the experimenter present. Subsequently, participants were debriefed. During debriefing it was made sure that participants had not been aware during the experiment that the experimenter or any other human was involved and/or had any part in the interaction (for an overview of the study flow, see Figure 3).

## Results

We first verified that agent appearance did not affect the results. As anticipated, there were no significant differences between agents with the same gender but different appearances. Therefore, the data was collapsed for further analysis.

There was an overall improvement of performance between the first task when the agent was absent ( $M = 3.86$ ,  $SD = 2.72$ ) and the second task when the agent was present ( $M = 5.49$ ,  $SD = 3.33$ ,  $t(73) = -7.29$ ,  $p = 0.001$ ). Also, self-evaluation of participants increased from the first task ( $M = 2.93$ ,  $SD = 1.00$ ) to the second task ( $M = 3.15$ ,  $SD = 1.08$ ,  $t(73) = -2.44$ ,  $p = 0.017$ ). Moreover, Participants showed significantly more effort by attempting to solve more math problems in the second task, when the agent was present, ( $M = 10.01$ ,  $SD = 3.67$ ) compared to the first task ( $M = 8.41$ ,  $SD = 3.49$ ;  $t(73) = -4.70$ ,  $p = 0.001$ ).

A 2 X 2 ANOVA, with the first factor being the display of rapport by the agent (rapport/no rapport) and the second factor being the gender condition (matching gender/opposing gender) showed a main effect of rapport on the improvement in performance ( $F(71) = 4.96$ ,  $p = 0.029$ ) (Figure 4). When the agent displayed rapport ( $M = 2.09$ ,  $SD = 1.96$ ), participants showed a significantly higher improvement in performance than without rapport ( $M = 1.21$ ,  $SD = 1.79$ ;  $t(72) = 2.02$ ,  $p = 0.047$ ). Specifically, in the opposing gender condition, there was a significant difference between the rapport ( $M = 2.56$ ,  $SD = 1.59$ ) and no-rapport conditions ( $M = 0.87$ ,  $SD = 1.92$ ), with

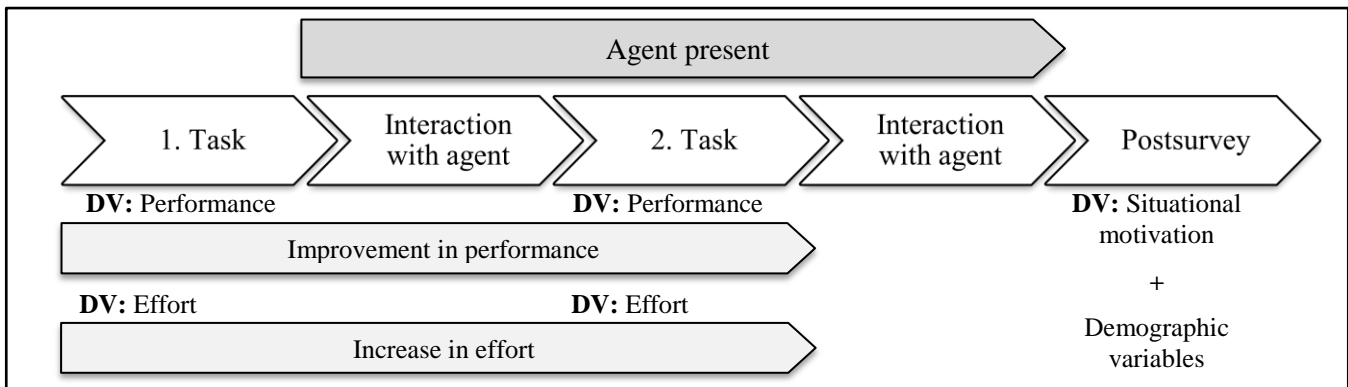


Figure 3: Study Flow

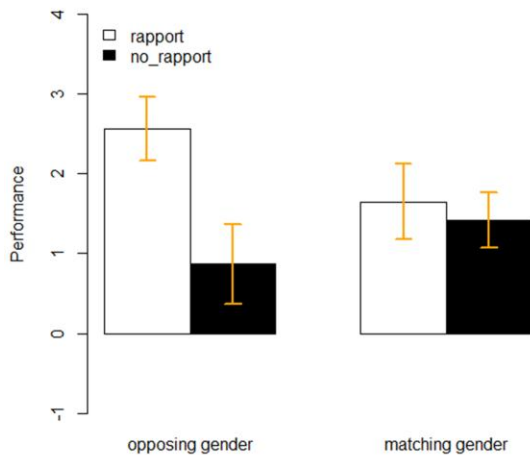


Figure 4: Performance Improvement

significantly higher performance in the rapport condition ( $t(29) = 2.68, p = 0.012$ ). The difference in improvement in the no-match condition did not reach significance.

We expected performance improvement to occur due to an increase of effort and motivation, thus we used one-tailed t-tests for these two variables. In the opposing gender condition participants invested more effort when rapport was displayed ( $M = 2.25, SD = 3.17$ ) compared to no displayed rapport ( $M = 0.27, SD = 3.37, t(29) = 1.69, p = 0.051$ ). In the opposing gender condition, there was a significant trend in increase of motivation between the rapport ( $M = 63.50, SD = 8.16$ ) and no rapport ( $M = 56.13, SD = 12.78$ ) condition ( $t(29) = 1.93, p = .032$ ). Within the matching gender condition the display of rapport did not show any influence on the dependent variables.

## Discussion

We expected that facilitation effects to occur when participants interact with a virtual agent of opposite gender and that this will only be the case if the agent displays rapport. The results of this study support these hypotheses. When rapport was displayed, participants' improvement was higher when they interacted with an agent of opposite gender than with an agent of matching gender. The same patterns were found for effort. This indicates that a social facilitation effect only occurs under a certain gender condition, i.e. opposing gender, and when rapport is displayed by the agent. Participants' performance improved most when they interacted with a virtual agent of opposite gender that displayed rapport and it improved least with an agent of the opposite gender who did not display rapport. This indicates that rapport has an effect on participants' improvement in performance only when the agents' gender

does not match their own. Research in social psychology has shown that establishing rapport between people and their instructors in face-to face interactions increases desirable outcomes such as motivation and improvement in task success (Granitz et al., 2009; Thomas et al., 1982). Our results show that rapport has a similar positive effect on performance in human-computer-interaction.

Overall, interacting with a virtual agent significantly enhanced participants' performance and effort. Our findings indicate that interacting with virtual agents of opposite gender that are capable of displaying rapport behavior improves participants' performance on mathematical tasks most. It supported participants' motivation and increased their effort to perform well by attempting to solve a higher number of math problems. We hope to further examine whether and how gender differences play a role in this interaction.

## Conclusion

In summary, contributions of this work are three-fold. First, the study adds to literature on human-computer-interaction with regard to virtual agents by showing that the agent's gender and rapport are both key factors for achieving desirable outcomes such as motivation, effort and performance with regard to mathematical tasks. Virtual Agents which are capable of establishing rapport can contribute to improve people's performance in mathematics, specifically with regard to standardized math tests. This observation may support the development of useful and effective applications in mathematical education and training, such as virtual instructors, tutors or learning companions. Second, it shows that social facilitation effects occur when interacting with virtual agents of opposing gender. Finally, this work makes a methodological contribution to the fields of experimental psychology and human-computer-interaction.

## Acknowledgments

This research was supported by the National Scientific Foundation under grant # IIS-0916858. The first author received support by the PROMOS Program of the German Academic Exchange Service (DAAD).

## References

- Baylor, A. L. & Ryu, J. (2003). Does the presence of image and animation enhance pedagogical agent persona? *Journal of Educational Computing Research*, 28(4), 373-395.
- Garau, M., Slater, M., Pertaub, D. P. & Razzaque, S. (2005). The responses of people to virtual humans in an immersive virtual environment. *Presence: Teleoperators and Virtual Environments*, 14, 104-116.

- Granitz, N. A., Koernig, S. K., & Harich, K. R. (2009). Now it's personal: Antecedents and outcomes of rapport between business faculty and their students. *Journal of Marketing Education*, 31(1), 52-65.
- Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R. J., et al. (2006). Virtual Rapport. *6th International Conference on Intelligent Virtual Agents*. Marina del Rey, CA: Springer.
- Gratch, J., Wang, N., Gerten, J., Fast, E., & Duffy, R. (2007a). Creating Rapport with Virtual Agents. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé (Eds.), *International Conference on Intelligent Virtual Agents 2007. LNAI 4722* (pp. 125-138). Paris, France: Springer Verlag Berlin Heidelberg.
- Gratch, J., Wang, N., Okhmatovskaia, A., Lamothe, F., Morales, M., & Morency, L. P. (2007b). Can virtual humans be more engaging than real ones? In J. Jacko (Ed.), *Human-Computer-Interaction, Part III, HCII 2007. LNCS 4552*, 286-297. Beijing, China: Springer Verlag Berlin-Heidelberg.
- Guay, F., Vallerand, R. J. & Blanchard, C. (2000). On the Assessment of Situational Intrinsic and Extrinsic Motivation: The Situational Motivation Scale (SIMS). *Motivation and Emotion*, 24 (3), 175-213.
- Guerin, B. & Innes, J. M. (1982). Social facilitation and social monitoring: A look at Zanjone's mere presence hypothesis. *British Journal of Social Psychology*, 2, 7-18.
- Hayes A. L., Ulinski, A.C. & Hodges, L. F. (2010). That Avatar Is Looking at Me! Social Inhibition in Virtual Worlds. In: Allbeck, J. et al. (Eds.). *International Conference on Intelligent Virtual Agents 2010, LNAI 6356*, 454-467.
- Huang, L., Morency, L. P. & Gratch, J. (2011). Virtual Rapport 2.0. In: *International Conference on Intelligent Virtual Agents*.
- Kim, Y. (2004). *Pedagogical agents as learning companions: The effects of agent affect and gender on learning, self-efficacy, and agent persona*. Tallahassee, FL: United State University.
- Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., Thrisson, K. R. & Vilhjmsson, H. (2006). Towards a common framework for multimodal generation: the behavior markup language. In: *International Conference on Intelligent Virtual Agents*, 21-23.
- Morency, L. P., Sidner, C., & Darrell, T. (2005). Towards context-based visual feedback Recognition for Embodied Agents. *Proceedings of the Symposium on conversational Informatics for Supporting Social Intelligence and Interaction* (pp. 69-72). Hatfield, UK: AISB.
- Rickenberg, R. & Reeves, B. (2000). The Effect of Animated Characters on Anxiety, Task Performance, and Evaluations of User Interfaces. *Letters of Computer-Human-Interaction 2000*, 49-56.
- Sanders, G. S., Baron, R. & Moore, D. L. (1978). Distraction and social comparison as mediators of social facilitation effects. *Journal of Experimental Social Psychology*, 14, 291-303.
- Thiebaux, T. & Marsella, S. (2007). Smartbody: Behavior realization for embodied conversational agents. In: *7th International Conference on Intelligent Virtual Agents*.
- Thomas, D., Ribitch, F. & Freie, J. (1982). The relationship between psychological identification with instructors and student ratings of college courses. *Instructional Science*, 11(2), 139-154.
- Tickle-Degnan L. & Rosenthal R. (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1(4), 285-293.
- Zanbaka, C., Ulinski, A., Goolkasian, P. & Hodges, L. F. (2004). Effects of Virtual Human Presence on Task Performance. In: *Proceedings of the International Conference of Additive Technologies 2004*, 174-181.
- Zanbaka C., Ulinski, A., Goolkasian, P. & Hodges, L. F. (2007): Social responses to virtual humans: implications on future interface design. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2007*, San Jose, CA, USA, April 28 - May 03, 1561-1570. ACM, New York.
- Zanjone, R.B. (1965): Social Facilitation. *Science*, 149, 269-274.

# A tripartite trans-modal relationship among sounds, shapes and emotions: A case of abrupt modulation

Shigeto Kawahara (kawahara@rci.rutgers.edu)

Linguistics Department,  
Rutgers University, New Brunswick NJ 08901, USA

Kazuko Shinohara (k-shino@cc.tuat.ac.jp)

Institute of Engineering, Tokyo University of Agriculture & Technology,  
2-24-16 Nakacho, Koganei, Tokyo 184-8588, JAPAN

## Abstract

The current project is a case study—and an extension—of the traditional investigation into sound symbolism (Hinton et al., 1994). Several studies have shown that certain sounds evoke images of particular shapes; for example, oral stop consonants are often associated with angular shapes, whereas sonorants (nasals, liquids, and glides) are associated with round shapes (Berlin, 2006; Köhler, 1947). Berlin (2006) attributes these associations to the similarities between abrupt acoustic amplitude modulation of stop consonants and abrupt change of the directions of lines, i.e., abrupt visual changes. In this study, we extend the stop-angular sound symbolic relation to the domain of emotions. Stops not only evoke the images of angular shapes, but are also associated with emotions that involve abrupt onsets. We further show that angular shapes themselves are associated with such emotions. Our three experiments thus establish a tripartite trans-modal symbolic relationship among three domains of cognition (sounds, shapes, and emotions). As an additional general implication, we argue that our experimental results support acoustic, rather than articulatory, bases of sound symbolism.

**Keywords:** sound symbolism; stop consonants; sonorants; angularity; emotions; abrupt modulations; modularity; synesthesia

## Introduction

A prevalent assumption in modern linguistics is the autonomy of semantics (meanings) from sounds. One of the Saussurian principles of languages suggests that there is no inherent connection between meanings and sounds. For example, there is no inherent reason why what we call [k<sup>h</sup>æt] is called in such a way. In fact, different languages call that animal by different names. Therefore, the argument goes, the sound-meaning relationship must be arbitrary. Saussure in fact raises this principle of arbitrariness as a first principle of natural languages (de Saussure, 1960).

However, evidence for an opposing view—that there is some inherent connection between sounds and meanings—is also available, and proponents of this view date back to at least Cratylus in Plato (Harris & Taylor, 1989). Especially since the seminal experimental work by Sapir (1929), a substantial body of experimental work shows that sounds are often associated with particular meanings. This relationship, often referred to as sound symbolism, is usually not absolute, and comes with some lexical

exceptions. For example, Sapir found that English speakers tend to associate [a] (back and low vowels) with big objects and [i] (high front vowels) with small objects; however, there are lexical items such as *big* that go against this trend. Since Sapir's work, the tendency to associate lower and backer vowels with bigger objects have been found in many other languages (e.g. Shinohara & Kawahara (2012) among many others).

In short, the previous work on sound symbolism in natural languages has established that we can at least identify stochastic tendencies toward some connection between sounds and meanings (despite the fact that some lexical items may not strictly follow such connections) (Hinton et al., 1994).

There is thus little doubt, albeit perhaps in non-categorical ways, that there are some associations between sounds and meanings. Moreover, these associations between sounds and meanings tend to make phonetic sense. For example, [a] is often considered to be larger than [i] in many different languages, and this association arises because [a] has wider opening of the mouth than [i] (Sapir, 1929; Shinohara & Kawahara, 2012). Alternatively, viewed from a (psycho)acoustic perspective, [a] involves lower second resonant frequency (F2) than [i], which would imply a larger resonator (Ohala, 1994). See also Shinohara & Kawahara (2012) for the comparison of these two theories of sound symbolism.

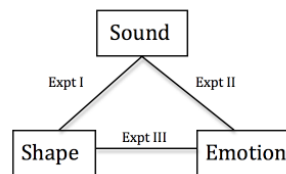


Figure 1: The roadmap of this paper.

In the current project, we focus on the meanings of oral stop consonants (like [p, t, k]), which acoustically involve abrupt amplitude changes, as opposed to sonorants, which involve gradual amplitude changes (nasals, liquids, glides, like [n, l, r, j]) (see section 2 for the illustration of these phonetic concepts.). We show that these acoustic characteristics of stops are associated with angular shapes



by native speakers of English (Experiment I). Furthermore, Experiment II shows that the acoustic characteristics of stops are also associated with emotion types with abrupt onsets, as opposed to those that involve gradual onsets. The final experiment goes beyond the traditional sound symbolism studies and shows that there is a direct connection between the emotion types with abrupt onsets and angular shapes. The last two experiments were motivated by a recent finding with Japanese speakers that particular types of emotions can be associated with particular sounds and shapes (Shinohara et al., 2011). Overall, our experiments show a tripartite relationship between three domains of cognition (auditory sounds, visual shapes, and emotion types). The overall conclusion and the roadmap of this paper are illustrated in Figure 1.

### Background: Phonetics of stops and sonorants

Oral stops are those sounds that are made with complete oral occlusion (and without the leakage of air through the nasal cavity), which results in rise of the intraoral air pressure (Ohala, 1983). The acoustic consequence of the rise in intraoral air pressure, upon the release of the stop occlusion, is abrupt bursts. Stop consonants thus involve a burst with abrupt amplitude changes after the release of the oral closure. Figure 2 shows a waveform of the 0.05 sec (=50 ms) interval centering around a stop burst of [t]. It represents amplitude changes on the y-axis across time on the x-axis. The transition from a closure phase (oral occlusion) into a burst is rather abrupt, and the burst itself involves abrupt amplitude changes.

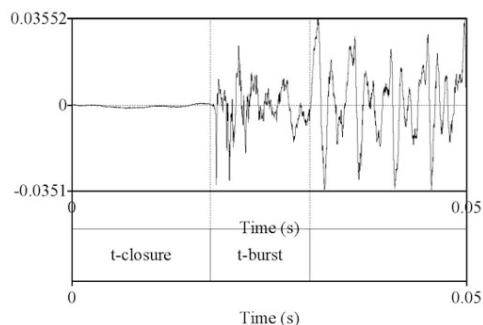


Figure 2: A waveform illustrating a closure phase and a burst of [t]. It shows amplitude changes within a 0.05 sec interval centering around a stop burst. A silent closure is followed by a burst with an abrupt onset.

Sonorants, on the other hand, include a class of sounds consisting of nasals ([m, n]), liquids ([l, r]), and glides ([w, j]) ([j] is the sound that is often represented with “y” in English orthography, as in *young*). In contrast to obstruents, sonorants do not involve rise in the intraoral air pressure because their aperture is wide enough to allow spontaneous vibration of vocal folds (Chomsky & Halle, 1968). Sonorants are thus characterized by energies with gradual changes, and their boundaries with respect to surrounding vowels are gradual. Figure 3 illustrates 0.05 sec intervals of

transitions from a vowel to a nasal [n] and to a glide [j]. As observed in the figure, the transitions from the vowel to sonorants are blurry, and the sonorants themselves are characterized by gradual amplitude changes.

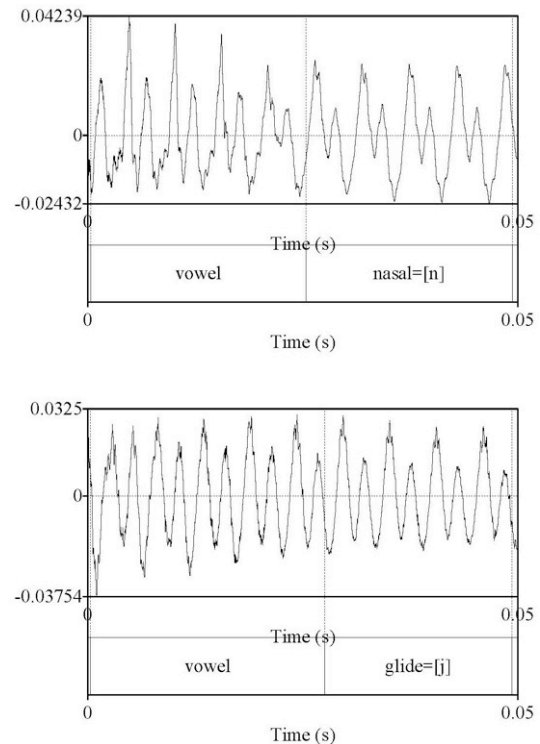


Figure 3: Wave forms of [n] and [j], illustrating 0.05 sec intervals including transitions from the preceding vowels to the sonorants. The boundaries between the vowels and the sonorants are blurry. The consonants themselves involve energies with gradual change.

### Experiment I: Sounds-Shapes

As observed in Figures 2 and 3, stops acoustically involve abrupt amplitude changes whereas sonorants involve gradual amplitude changes. Previous experiments have shown that speakers map these acoustic characteristics to the visual domain, considering stops to be more angular than sonorants. Köhler presented two types of figures, one with an angular shape and one with a round shape (see Figure 4), with two sound stimuli, *takete* and *maluma*. The result was that people often associate *takete* with the angular shape and *maluma* with the round shape. Berlin combined this observation and Ohala’s theory of sound symbolism (Ohala, 1994) to investigate animal nomenclature patterns (Berlin, 2006). He suggests [p. 34] that “[a]n angular, sharp, long-legged, streamlined bodied rail ought to show a preference for voiceless consonants, especially voiceless stops, while the rounded, short-legged tinamou should not favour these sounds.” (see also Ramachandran & Hubbard (2001) and footnote 2 for other sound distinctions that may yield the images of angularity and roundness.) Experiment I replicates these findings by testing whether English speakers



map the acoustic characteristics of stops and those of sonorants to a visual domain. For this purpose, we auditorily presented stimuli with stops and those with sonorants together with pairs of angular shapes and round shapes, and asked them to match each stimulus sound with either an angular shape or a round shape.

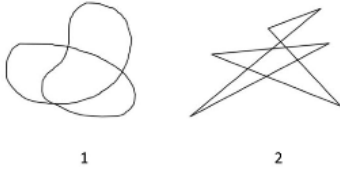


Figure 4: Our reproduction of a pair of shapes used by Köhler (1929/1947). The pair was also used in Experiments I and III.

## Method

Experiment I expanded on the previous results (Köhler, 1947; *et seq.*), and tested the connection between stops and angular shapes. The experiment used many stimulus pairs to test the generality of this connection. Furthermore, to avoid the effects of orthography, we used auditory stimuli.

**Stimuli** The stimuli were all disyllabic CVCV nonce words (i.e. nonexisting words in English). In one condition, both syllables contained stop onsets; in the other condition, both syllables contained sonorant onsets.<sup>1</sup> The vowel quality was controlled between the two conditions: the first vowels were [a, e, ɪ, o, u], and the second vowels were [ə, i] (10 vowel combinations). Two items were included for each vowel combination. The stimulus list is provided in Table 1.

Two native English speakers (one female, one male) pronounced all the stimuli three times in a sound-attenuated booth, at the sampling frequency of 44.1k Hz. To control for potential effects of F0 contour on the listeners' judgments about the images of the stimuli's shapes, the recorded tokens were acoustically resynthesized with a uniform falling contour from the first vowel to the second vowel. For the female speaker, F0 of the first syllable was adjusted to 300 Hz, and F0 of the second vowel to 200 Hz, with linear interpolation in between. For the male speaker, the two F0 parameters used were 150 Hz and 100 Hz, again with linear interpolation. Also, to control for the potential effects of amplitude, peak amplitude of all the stimulus files was modified to 0.7 by using Praat (Boersma & Weenink, 1999-2012). Together with 40 nonce words consisting of stops and 40 nonce words consisting of sonorants, seven different pairs of shapes, each pair consisting of an angular shape and a round shape, were prepared, as exemplified in Figure 5 (the experiment also included the pair of shapes similar to

Köhler's, shown in Figure 4). The experiment thus had a total of 560 stimuli (80 auditory stimuli \* 7 figure pairs).

Table 1: The stimulus list for Experiments I and II.

	Stop condition	Sonorant condition
a-ə	[tagə] [bakə]	[jamə] [ralə]
e-ə	[depə] [tekə]	[wejə] [rewə]
ɪ-ə	[kɪbə] [tɪbə]	[jimə] [wijə]
o-ə	[døkə] [dopə]	[jorə] [nojə]
u-ə	[dukə] [pukə]	[munə] [mujə]
a-i	[kabi] [tadi]	[maji] [jawi]
e-i	[tegi] [tepi]	[reni] [jewi]
ɪ-i	[tɪpi] [tɪgi]	[jini] [nɪwi]
o-i	[boki] [pobi]	[joli] [woji]
u-i	[buki] [gugi]	[wuni] [luri]

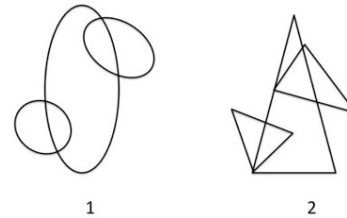


Figure 5: A sample pair of visual cues used in Experiments I and III.

**Procedure** For each trial, the participants were presented with a pair of objects, one angular and one round, immediately followed by a stimulus sound. They were then asked to choose a shape that better matched each auditory stimulus. The maximum time for the participant to respond to each trial was 3000 ms; if they did not respond within this time limit, that trial was skipped. The inter-trial interval was 250 ms. The visual and audio stimuli were presented using Superlab ver. 4.0 (Cedrus). All the participants wore high quality headphones (Sennheiser HD 280 Pro), and registered their responses using an RB-730 response box (Cedrus).

The experiment started with a practice block in order for the participants to familiarize themselves with the procedure. To avoid loss of attention due to exhaustion, the main session was organized into two blocks. The first block contained the combination of all the auditory stimuli with four pairs of shapes; the second block presented the rest of the three visual pairs. The two blocks were separated by a self-timed break. The order of trials within each block was randomized per each participant by Superlab.

<sup>1</sup> The current stimuli do not include fricatives, which also involve frication with abrupt amplitude changes. Testing the visual images associated with fricatives awaits further experimentation.

**Participants** Seventeen native speakers of English participated in the experiment in a sound attenuated room. They were all undergraduate students at Rutgers University, and received extra credit for linguistics classes.

**Statistics** Since the responses were categorical (angular or not), a logistic linear mixed model regression was run in which the dependent variable was the angular response and the independent variables were the difference between stops and sonorants as a fixed factor and subject as a random factor. All statistical analyses in this paper were performed using R.

## Results and discussion

Figure 6 presents the percentages of angular responses for the stop condition and the sonorant condition.<sup>2</sup> As observed in Figure 6, English listeners associated angular shapes much more frequently with the stop stimuli than with the sonorant stimuli. This difference between the two conditions is statistically significant ( $z = 35.00$ ,  $p < .001$ ). We thus conclude that stops, which involve bursts with abrupt amplitude changes, are associated with angular shapes, and that sonorants, which involve energies with gradual changes, are associated with round shapes.<sup>3</sup>

## Experiment II: Sounds-Emotions

Experiment II extended the results of Experiment I to the domain of emotions. We tested whether acoustic abrupt changes of stop consonants are projected on the domain of emotion types. We tested a pair of two emotions like “shocked” and “sad”, the former of which involves an abrupt onset; i.e. those types of emotions that start abruptly. The prediction is that, if there is a trans-modal relationship between sounds and emotion types, stops are more likely to be associated with emotions with abrupt onsets (Shinohara et al., 2011).

<sup>2</sup> A signal detection analysis would have been an alternative, which would tease apart sensitivity from bias (Macmillan and Creelman, 2005). We report percent correct analyses throughout this paper for the ease of interpretation.

<sup>3</sup> One remaining question to be addressed in future research is the effect of place of articulation and voicing. Ramachandran & Hubbard (2001) show that [kiki] is considered to be more angular than [bouba]—both of these nonce words contain stops. It seems that, in addition to the differences in vowels, [k] is more angular than [b]. Furthermore, this difference seems to arise because [b] is both labial and voiced. Labial sounds may be associated with round images because they involve movement of lips; voiced stops may be more round than voiceless stops, because their bursts are usually weaker, and voiced stops involve voicing during closure (intervocally), which consists of gradual amplitude changes (Berlin, 2006; Ohala, 1983). Testing the effect of place of articulation and voicing (and vowels, for that matter) requires future experimentation. See also Jakobson (1978) for discussion on the effect of the acute/grave distinction on the images of sharpness.

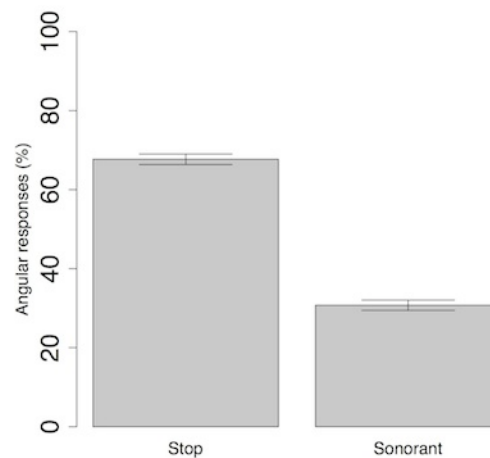


Figure 6: The percentages of angular responses in Experiment I. The error bars are 95% confidence intervals.

## Method

The auditory stimuli were identical to those used in Experiment I. For emotion types, we used a pair of negative emotions (“shocked” vs. “sad”) and a pair of non-negative emotions (“surprised” vs. “happy”).

In this experiment, the participants were told that they would be hearing words of a language that they do not know. After each auditory stimulus was presented, they were presented with one of the two pairs of words in English orthography (“shocked” vs. “sad” or “surprised” vs. “happy”). They were instructed to choose the meaning that better matches the auditory stimuli. Other details of the experiment were identical to Experiment I, except that there was no short break because the experiment was much shorter (80 auditory stimuli \* 4 emotions = 320 stimuli). Experiment II was conducted right after Experiment I after a short break with the same participants (seventeen native speakers of English). A logistic regression was run on abrupt responses, with the difference between stops and sonorants and the two types of pairs (negative vs. non-negative) and their interaction as fixed factors and subject as a random factor.

## Results and discussion

Figure 7 shows “abrupt responses” (“shocked” for the negative pair and “surprised” for the non-negative pair) for each condition. We again observe that English listeners associated stops with those emotions with abrupt onsets. The difference between stops and sonorants was significant ( $z = -7.80$ ,  $p < .001$ ). The difference between the two types of pairs was also significant ( $z = 2.13$ ,  $p < .05$ ), because abrupt responses were generally higher for the negative pair of emotions. The interaction term, however, was not significant ( $z = -1.53$ , *n.s.*), showing the difference between the stop condition and the sonorant condition was consistent between the two pairs. Experiment II thus shows that stops are not only associated with angular shapes, but also with emotion types that involve abrupt onsets. This finding, together with Shinohara et al. (2011), as far as we know,

adds a new type of sound symbolism to the sound symbolism literature.

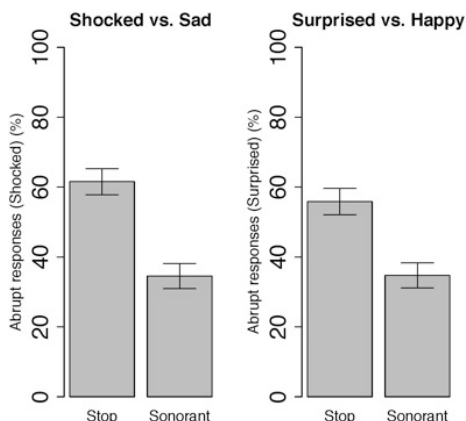


Figure 7: The percentages of abrupt responses in Experiment II.

### Experiment III: Shapes-Emotions

The final experiment tested the relationship between emotion types and shapes, the question being whether those emotion types with abrupt onsets are associated with angular shapes. The previous two experiments established that these two are both associated with stops, and the final experiment addressed whether emotions and shapes are directly related. The prediction from the previous two experiments is that English speakers would associate angular shapes with those emotions that involve abrupt onsets.

### Method

The stimuli were 16 pairs of angular and round shapes, as exemplified in Figure 5.<sup>4</sup> In this experiment, the participants were instructed to be an assistant of Steven Spielberg, a film-director. They were told that in his new movie, the setting is an extraterrestrial planet where people communicate using visual symbols rather than sounds. The participants were presented with a pair of visual cues, one with a round shape and the other with an angular shape (see Figure 5), and were asked to choose which one better matches a particular meaning (“shocked” or “sad”, “surprised” or “happy”). Experiment III was conducted as an online questionnaire survey, as it did not involve auditory stimuli. The experiment was created and distributed using surveymonkey, and the participants were recruited on Psychology on the Net, an online forum for psychology experiments. 37 native speakers of English voluntarily participated in the experiment. No compensation was offered for this experiment.

<sup>4</sup> Since Experiment III tested the combination of only two pairs of emotions (see below), it allowed us to use more pairs of visual stimuli than Experiment II.

## Results and discussion

Figure 8 represents the percentages of how often the angular shapes were associated with each emotion in Experiment III.

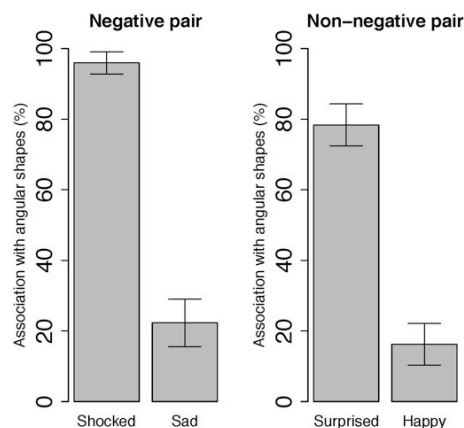


Figure 8: The percentages of how often the angular shapes were associated with each emotion in Experiment III.

We observe that the predictions are borne out: angular shapes were associated more frequently with “shocked” and “surprised” than with “sad” and “happy”, i.e. those emotions that involve abrupt onsets. Statistically, the difference between the two types of emotions (those with abrupt onsets vs. those without) was significant ( $z = 9.57, p < .001$ ). There was no overall difference between the negative pair (“shocked” vs. “sad”) and the non-negative pair (“surprised” vs. “happy”) in terms of angular responses ( $z = -1.21, n.s.$ ). The interaction was significant ( $z = -2.66, p < .01$ ), because the difference between the two emotions was more pronounced in the negative pair than in the non-negative pair. Since the interaction was significant, simple analyses were run separately for the negative pair and non-negative pair. They revealed that the difference within each pair was significant ( $z = 9.50, p < .001$  for the negative pair, and  $z = 9.35, p < .001$  for the non-negative pair).

### General discussion

To summarize, Experiment I has shown that English speakers associate oral stop consonants with angular shapes, supporting the previous work on this sound symbolic relationship. Experiment II shows that stops are also associated with emotion types that involve abrupt onsets. This sound-emotion connection, to the best of our knowledge, is new and adds a new instance of a sound symbolic relation to the literature. More generally, the results from these two experiments lend further support to the existence of sound symbolism, a general idea that there are particular sound-meaning relationships. Experiment III, going beyond traditional sound symbolic studies, has shown that angular shapes are associated with those types of emotions that involve abrupt onsets. Taken together, our three experiments establish a tripartite trans-modal symbolic relationship between three domains of cognition (auditory

sounds, visual shapes, and emotions), among those that involve abrupt modulation, as summarized in Figure 1.

In addition to establishing this tripartite trans-modal relationship among stops, angular shapes, and emotions with abrupt onsets, we suspect that the results of Experiment I have one implication for a general debate about sound symbolism: the debate concerning whether sound symbolism is based on articulation or acoustics. It seems plausible to assume that the image of angular shapes comes from the bursts of stops; i.e. it makes acoustic sense. Acoustically, the stop bursts with abrupt amplitude changes look “spiky” if we track the amplitude changes of stop bursts across time, as we illustrated in Figure 2. By contrast, if we track amplitude changes of sonorants across time, they look “roundish”, as in Figure 3. The association between stops and angular shapes and the one between sonorants and round shapes can be considered as projection of the acoustic characteristics of sounds to the visual domain.

On the other hand, an articulation-based explanation of the current results seem difficult, because there is nothing in the articulation of stops that is angular. In fact, the only superlaryngeal articulatory difference between [t] and [n] is opening of velum in [n], and it is not immediately clear why the opening of the velum can be associated with round shapes.<sup>5</sup>

Independent of our speculation about the basis of sound symbolic relationships, our results show that characteristics of sounds can be projected onto the domain of emotions. We have further shown in Experiment III that such a trans-modal relationship directly holds between the domain of visions and emotions. We hope that further research will address how different modalities of cognitions are linked to one another. We suspect that this line of research may go in tandem with general research on synesthesia (Ramachandran & Hubbard, 2001), as our results show that there may be tighter relationships between different modalities of cognitions than previously assumed. Finally, to strength our claim about the trans-modal relationship, it would be desirable to test the claim with more instances of

emotions, and with a wider range of experimental paradigms. We leave this task for future research.

## Acknowledgments

We thank the research assistants at the Rutgers Phonetics Laboratory for their assistance with this project. For comments on earlier draft of this paper, we thank Kimi Akita, Young Ah Do, Shinsuke Nakajima, Seunghun Lee and two anonymous reviewers.

## References

- Boersma, P. & Weenink, D. (1999–2012). Praat: Doing phonetics by computer. Software.
- Berlin, B. (2006). The first congress of ethnozoological nomenclature. *Journal of Royal Anthropological Institution*, 23–44.
- Chomsky, N. & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper and Row.
- de Saussure, F. (1916/1960). *Cours de linguistique générale*[Course in general linguistics]. Paris: Payot.
- Harris, R. & Taylor, T. J. (1989). *Landmark in linguistic thoughts*. London & New York: Routledge.
- Hinton, L., Nichols, J., & Ohala, J. (1994). *Sound Symbolism*. Cambridge : Cambridge University Press.
- Jakobson, R. (1978). *Six Lectures on Sound and Meaning*. Cambridge : MIT Press.
- Köhler, W. (1929/1947). *Gestalt Psychology*. New York: Liveright.
- Macmillan, N. & Creelman, D. (2005). *Detection Theory: A User's Guide. 2nd Edition*. Mahwah: Lawrence Erlbaum Associate Publishers.
- Ohala, J. J. (1983). The origin of sound patterns in vocal tract constraints. In MacNeilage, P., (Ed.), *The Production of Speech* (pp. 189-216). New York: Springer-Verlag.
- Ohala, J. J. (1994). The frequency codes underlies the sound symbolic use of voice pitch. In Hinton, L., Nichols, J., & Ohala, J. J. (Eds.), *Sound Symbolism* (pp. 325-347). Cambridge: Cambridge University Press.
- Ramachandran, V. & Hubbard, E. M. (2001). Synesthesia—a window into perception, thought, and language. *Journal of Consciousness Studies*, 8(12):3–34.
- Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology*, 12:225–239.
- Shinohara, K. & Kawahara, S. (2012). A cross-linguistic study of sound symbolism: The images of size. In *Proceedings of BLS 36*. Berkeley: BLS
- Shinohara, K., Natsume, F., & Matsunaka, Y. (2011). Sound-shape-emotion iconicity in visual psychomimes in Japanese. A talk presented at the Eighth International Symposium on Iconicity in Language and Literature, Linnaeus University, Vaxjo, Sweden.

<sup>5</sup> It is of course possible that an articulatory basis exists for the stop-angular sound symbolic relationship, which we are simply unaware of. However, a follow-up study of the current study provided further evidence for the acoustic basis of the angularity image. The follow-up study presented English listeners with non-speech sounds like sine waves and square waves. They were told that these sounds were used by an extraterrestrial language, and asked to judge whether sine waves and square waves mean “abrupt” or “gradual”. The results show that square waves, which acoustically involve more abrupt change, are indeed judged to mean “abrupt” more often. The result shows that English speakers can map abrupt acoustic change of non-speech to semantic meaning of abruptness, which is presumably the basis of the results of Experiment I. Since the non-speech sounds used in this follow-up experiment are unlike human sounds, whose articulatory origin cannot even be speculated by listeners, the results support the acoustic basis of the images of abruptness.

# Negating compound sentences

Sangeet Khemlani (khemlani@aic.nrl.navy.mil)  
Navy Center for Applied Research in Artificial Intelligence  
Naval Research Laboratory, Washington, DC 20375 USA

Isabel Orenes (iorenes@ull.es)  
Departamento de Psicología Cognitiva  
Universidad de la Laguna, Tenerife, Spain

P.N. Johnson-Laird (phil@princeton.edu)  
Department of Psychology  
Princeton University, Princeton, NJ, USA

## Abstract

How do reasoners negate compound sentences, such as conjunctions of the form *A and B* and disjunctions of the form *A or B or both*? A theory based on mental models posits that reasoners negate each clause independently, and enumerate the various possibilities consistent with the negation. It makes a novel prediction: negations of conjunctions should be more difficult to comprehend than negations of disjunctions. Two experiments corroborate the prediction. Experiment 1 tested participants' ability to comprehend sentential negations by giving them assertions of the form: *Bob denied that he wore a yellow shirt and he wore blue pants on Tuesday*. Participants selected the clothing options that Bob possibly wore on Tuesday. Experiment 2 gave participants descriptions such as *Bob loves Mary or Mary loves John or both*, and they were required to formulate a denial by completing a sentence that started with "No, ...". In both studies, participants' responses were more accurate for denials of disjunctions than denials of conjunctions.

**Keywords:** enumerative negation, sentential negation, conjunctions, disjunctions, and mental models

## Introduction

Consider the following two sentences:

- 1a. It's not the case that *both* Pat loves Viv *and* Viv loves Pat.
- 1b. It's not the case that Pat loves Viv *or* Viv loves Pat, *or both*.

How do people understand and reason about negated compound sentences like the negated conjunction (1a) and the negated disjunction (1b) above? Which one of the sentences is easier to process? Naïve individuals, i.e., those with no background in logic, syntax, or semantics, should have difficulty understanding the logical negation of multiple clause assertions, particularly when those assertions are complex (Clark, 1974; Clark & Chase, 1972; Hoosain, 1973). Nevertheless, people use negations frequently in everyday reasoning. Indeed, many of the world's languages contain a linguistic construction geared towards negating a disjunction, *neither A nor B*, where *A* and *B* stand for any clauses. Similar constructions exist in many other languages (Gazdar & Pullum, 1976) including

German, Malay, and Portuguese. And in logic, the negations of conjunctions and disjunctions, i.e., the NAND and the NOR connectives, can be used to derive every other logical connective. Negations therefore have powerful semantic implications, and they're used often in daily life, but individuals probably do not comprehend the full logical implications of a complex negated assertion. So, how do reasoners cope with sentential negations? In the present paper, we show that the theory of mental models can account for how individuals interpret such negations.

## Mental models and enumerative negation

The theory of mental models – the “model” theory for short – posits that individuals use the meaning of an assertion and any relevant knowledge to envisage what is possible (Johnson-Laird, 1983), and that they interpret negations by considering the various possibilities to which the negations refer (Khemlani, Orenes, & Johnson-Laird, 2012). Consider the examples above. When individuals represent the sentential negation of a conjunction, such as (1a), they need to consider the three separate possibilities that render it true. That is, the negation is true when a) neither loves the other; b) Pat doesn't love Viv but Viv loves Pat; or c) Pat loves Viv but Viv doesn't love Pat. In contrast, the sentential negation of the disjunction is consistent with only one possibility: neither loves the other. Assertion 1a above has the grammatical form:

2. It's not the case that both P and V.

where *P* stands for *Pat loves Viv*, and *V* stands for *Viv loves Pat*. According to the model theory, the core interpretation of negation and conjunction, (1a) refers to the following mental models:

¬ P	¬ V
¬ P	V
P	¬ V

where ‘¬’ denotes the symbol for negation. And (1b) refers to only the first of these models.

How do individuals construct the models for the assertions above? If individuals had prior knowledge of De

Morgan's laws for interrelating conjunctions and disjunctions, then they would not need to build models, and could simply apply the laws to infer the correct negation. Naïve individuals are unlikely to have mastered De Morgan's laws, however, and so the model theory postulates a more plausible hypothesis. The theory assumes that individuals think about discrete possibilities, where a possibility consists of a conjunction of individuals, their properties, and the relations among them. In the diagram above, the three rows refer to models of three separate possibilities consistent with the negation of the conjunction. To interpret the negation of a multiple-clause assertion, such as (1a), individuals envisage these models separately: they make a series of independent negations of individual clauses *P* and *V*. Hence, with *It is not the case that both P and V*, individuals begin with the possibility in which the negation is applied to each clause: *not-P* and *not-V*. This possibility is not consistent with the original affirmative assertion, *P* and *V*, and so they realize that it is one possibility in which the negation holds. At this point, some reasoners may stop and consider only this initial possibility in which both clauses are negated. However, if individuals go further, they apply the negation to only one of the clauses, e.g., *not-P* and *V*. Once they do, they can detect that it too is inconsistent with the original affirmative and accordingly a possibility consistent with the negation. Likewise, they may grasp that *P* and *not-V* is also a possibility that renders the negation true. Finally, reasoners need to consider the case, *P* and *V*. The possibility is consistent with the unnegated conjunction, and it is therefore inconsistent with the negation of the conjunction.

The general procedure, which we refer to as *enumerative negation*, is to construct a series of models of conjunctive possibilities for any sort of complex compound assertion. It starts with negations of both clauses, and checks whether the resulting possibility is consistent with the unnegated assertion. It then negates each clause, and accepts only those possibilities that are not consistent with the unnegated assertion. Finally, it affirms both main clauses. In each case, if a model is consistent with the unnegated assertion, it is rejected; otherwise, it is accepted as consistent with the negation. This hypothesis applies to all connectives between main clauses, but it is recursive so that it can cope with clauses within clauses. To be right for the right reasons depends on completing the full sequence of all possible conjunctions based on the two clauses.

There is an important rider to enumerative negation: individuals are likely to fail to construct the full sequence of models, which is difficult and time-consuming to envisage. Hence, they should be more likely to respond correctly if they are asked to evaluate given possibilities. In sum, naïve individuals should formulate the denial of compound assertions with multiple main clauses by envisaging, one at a time, the various sorts of possibility in which the denial holds. The order of constructing the models is unlikely to be constant, but it should usually begin with the negations of both clauses.

The model theory of negation makes a novel, and perhaps counterintuitive prediction. In the case, of affirmative assertions, conjunctions are easier to understand than disjunctions, but this difference should switch in the case of their negations. A conjunction has a single model; an inclusive disjunction has multiple models. But, the negation of a conjunction has multiple models; and the negation of an inclusive disjunction has one model. The relation is complementary. The prediction presupposes that the greater the number of mental models of various sorts of compound assertions, the harder it should be to understand them. The effect is easy to understand in the case of compound assertions such as conjunctions and disjunctions. Two atomic propositions and their respective negations yield four possible models:

A	B
A	¬ B
¬ A	B
¬ A	¬ B

A conjunction of the form, *A and B*, refers to only one of these models, but an inclusive disjunction of the form, *A or B or both*, refers to the first three of them. Hence, the conjunction should be easier to understand than the disjunction. In contrast, the negation of the conjunction, *not both A and B*, refers to the three models that are the complement of the model of the original conjunction, *A and B*, whereas the negation of the disjunction, *not (A or B)*, refers to the one model that is the complement of the three models of the original disjunction, *A or B or both*. This predicted interaction hinges, of course, on the theory that individuals construct mental models of assertions, and on the core meaning of negation. Theories in which models of possibilities play no part are unlikely to make the prediction (cf., e.g., Braine & O'Brien, 1998; Rips, 1994).

To test this prediction, we carried out two experiments examining the negation of conjunctions and disjunctions. In both studies, the participants had to deal with denials instead of negations, because pilot studies showed that naïve reasoners don't understand what it means to "negate" a sentence. The studies also examined the denials of conditional *if-then* assertions. The theory predicts that conditionals should be complicated to deny. On the one hand, denials of conditionals should be easier to comprehend than denials of conjunctions because individuals are likely to reduce the scope of the negation to the subordinate *then*-clause (the consequent). On the other hand, the correct negation of the conditional, *A and not-B*, is unlikely to be the first model that reasoners enumerate, so it should be difficult. Thus, the theory predicts that denials of conditionals should be an intermediate case, i.e., not as difficult to understand as denials of conjunctions but more difficult to understand than denials of disjunctions. The results of both studies corroborated the predictions of the model theory.



## Experiment 1:

### Understanding sentential negations

Experiment 1 tested the enumerative negation hypothesis for the task of listing what is possible given affirmations and denials of three sorts of statement: *A and B*, *A or B or both*, and *if A then B*. Conditionals are complicated. Their affirmations should yield two or three possibilities depending on whether participants make a biconditional, (e.g., If and only if A then B) or a regular conditional interpretation. Their negations, however, should either reduce the scope of the negation to the main clause, *If A then not-B*, or else be the correct response, *A and not-B*.

We carried out various preliminary studies, both online and face-to-face, which showed that the task was difficult. For example, when we asked participants to list what was impossible given a sentential negation, their performance was almost at chance. As a result of these initial studies, we settled on a task in which participants judged whichever of four cases: *A and B*, *A and not-B*, *not-A and B*, *not-A and not-B*, was “possible” given a statement. The statements, in turn, were either affirmations or denials of the three sorts of assertion.

### Method

*Participants.* 22 adult native-English speaking participants were recruited through an online system, Mechanical Turk, hosted by Amazon.com that allows people to volunteer for experiments for monetary compensation.

*Design and materials.* Participants acted as their own controls and selected the possible instances of three affirmations (based on *and*, *or*, and *if*) and the possible instances of their three denials. The sentences were presented as a block of affirmations and a block of denials in a counterbalanced order. The actual sentences concerned the color of the clothes of various individuals, who affirmed or denied what they wore on a particular day, e.g.,

*Bob [asserted/denied] that he wore a yellow shirt [and/or] he wore blue pants on [Monday/Tuesday/...].*

*Bob [asserted/denied] that if he wore a red shirt then he wore pink pants on [Monday/Tuesday/...].*

We used adverbial phrases, such as “on Tuesday”, to convey that the statement was about what a person wore on a particular occasion. For the preceding example, the participants indicated whichever of the following cases they judged to be possible given the statement:

*Bob wore a yellow shirt and he wore blue pants.*

*Bob wore a yellow shirt and he wore non-blue pants.*

*Bob wore a non-yellow shirt and he wore blue pants.*

*Bob wore a non-yellow shirt and he wore non-blue pants.*

The participants were told to select all the cases that they judged to be possible for each sentence. The order of

presentation of the four cases was counterbalanced over the trials.

### Results and discussion

No reliable difference occurred in the accuracy of the responses in the two blocks, and so we pooled the data for subsequent analyses. The predicted interaction between polarity and the connectives (conjunctions and disjunctions) was reliable. For affirmations, conjunctions yielded 86% correct responses and disjunctions yielded 68% correct responses; whereas for denials, conjunctions yielded 18% correct responses and disjunctions yielded 89% correct responses (Wilcoxon test,  $z = 3.47$ ,  $p < .0005$ ). Denials of conjunctions were very difficult: the participants’ mainly judged *not-A and not-B* alone as possible (45%), and 14 out of the 22 participants thought of only one possibility, whether right or wrong (Binomial  $p < .005$ , given a prior probability of .33).

The data for the conditionals also corroborated the model theory. Their affirmations yielded 45% conditional interpretations, 18% biconditional interpretations, and 27% interpretations equivalent to conjunctions – a phenomenon that occurs in judgments of probability (Giroto & Johnson-Laird, 2004; Johnson-Laird, Byrne, & Giroto, 2009), and which suggests a regression to a more child-like interpretation in a difficult task (see Barrouillet, Grosset, & Lecas, 2000). The denials of conditionals fell mainly into the two predicted categories: an interpretation that reduced the scope of the negation, *if A then not-B* (59%, see Khemlani et al., 2012, for an elaboration of this effect) or the correct response, *A and not-B* (14%). No one selected the correct possibilities for the denial of a biconditional despite the fact that this interpretation occurred in the affirmation.

The task called for the participants to understand affirmative and negative statements and to evaluate explicit possibilities in relation to them. When connectives interrelate main clauses, the model theory predicts the interaction with polarity: conjunctions are easier than disjunctions when they are affirmed, but their relative difficulty is reversed when they are denied. Conditionals also yield the predicted but unusual pattern of judgments: many individuals take the denial of a conditional, *if A then not-B*, to hold in some of the same possibilities as its affirmation, *if A then B*. Since this interpretation yields only a contrary to the affirmed conditional, such “small scope” interpretations are predictable, but erroneous.

When individuals have to formulate a denial of an assertion, their task is to map their models of the possibilities into a conclusion. Hence, the task should be easier in case their starting point is only one model as in the case of a denial of a disjunction than in case it is several models as in the case of a denial of a conjunction. In this way, the enumerative negation hypothesis yields predictions about the formulation of negative statements. Experiment 2 tested these predictions.



## Experiment 2:

### Formulating sentential negations

The previous study examined participants' understanding of denials; Experiment 2 examined their formulation of denials. A preliminary study showed that when individuals are asked to "negate" a conditional, they tended to negate both of its clauses: they did so on 69% of trials. This result suggests that the task of "negating" a compound sentence is unfamiliar to naïve individuals. The present experiment, like the one before it, was accordingly framed in terms of the semantic task of "denial". The participants had to formulate denials of three sorts of sentence:

- conjunctions, *A and B*;
- inclusive disjunctions, *A or B or both*;
- conditionals, *If A then B*;

The enumerative negation hypothesis predicts that individuals should construct a set of conjunctive models and retain those that are inconsistent with the statement. It follows that the participants should tend to be most accurate in denying inclusive disjunctions, because the first conjunction that they are likely to consider, *not-A and not-B*, is the one and only correct denial. They should be less accurate with conditionals, because they are likely to have to construct more than one conjunction before they encounter the correct denial: *A and not-B*. And another factor of greater importance may intervene. Individuals may reduce the scope of the negation, and this process is likely to apply to conditionals too. Hence, some individuals should assert *if A then not-B* as the denial of the affirmative conditional. Finally, the participants should tend to be least accurate with conjunctions, because their correct denial depends on enumerating three models of possibilities: *not-A and not-B*, *not-A and B*, and *A and not-B*. These possibilities are equivalent to the inclusive disjunction: *not-A or not-B*, but this realization is likely to be beyond anyone who does not know De Morgan's laws.

### Method

*Participants, design, and procedure.* 21 native English-speaking participants were recruited through the same participant pool as in Experiment 1. They acted as their own controls and had to formulate denials of six conjunctions, six disjunctions, and six conditionals, all of which were expressed in English, and which were presented to each participant in a different random order. They were instructed to deny the statements by formulating a complete sentence that began with the word, *No*, as a preface to their denial, and the sentence could be of any length. Each clause in the statements to be denied contained two noun phrases based on proper nouns, a transitive verb, and one co-reference, e.g., "If Bob loves Mary then Mary hates Julie." The materials were constructed so that no proper name or transitive verb occurred more than once in the experiment.

## Results and discussion

Table 1 presents the percentages of the various sorts of denial. The participants corroborated the predicted trend: they made correct denials for 67% of inclusive disjunctions (*not-A and not-B*), 28% of conditionals (*A and not-B*), and 0% of conjunctions (*not-A or not-B*, or the list of three conjunctive possibilities). The predicted trend was highly reliable (Page's  $L = 281.5$ ,  $z = 4.55$ ,  $p < .00001$ ). The conditionals also elicited 34% of denials of the form: *If A then not-B*, which is consistent with the hypothesis that reasoners reduce the scope of the negation to make it easier to comprehend. The participants making this response tended to be different from those who made the correct denials: 7 out of the 21 participants responded *if A then not B* on half or more of the trials, and 10 out of the 21 participants responded *A and not B* on half or more of the trials. The difference between these two post-hoc groups in the frequency with which they responded *if A then not B* was highly reliable (Mann-Whitney test,  $z = 3.50$ ,  $p < .0001$ ). In accord with the enumerative negation hypothesis, when participants had to deny statements, they were most accurate in denying inclusive disjunctions and least accurate in denying conjunctions. The model theory predicts this result, but it is contrary to Rips's PSYCOP theory (1994, p. 113), which makes the opposite prediction based on its formal rules for De Morgan's laws:  $\neg(A \& B)$ , therefore,  $\neg A \vee \neg B$ ; and  $\neg(A \vee B)$ , therefore,  $\neg A \& \neg B$ . For rules that work forwards from premise to conclusion, a single step yields the inference from the negation of a conjunction, whereas three steps based on different rules are needed for the inference from the negation of a disjunction.

Table 1: The percentages of the different denials of disjunctions, conditionals, and conjunctions in Experiment 2, where the balances of responses in each column were different miscellaneous errors that occurred on less than 10% of trials.

The structure of the participants' denials	Type of assertion to be denied		
	Disjunctions: <i>A or B or both.</i>	Conditionals: <i>If A then B.</i>	Conjunctions: <i>A and B.</i>
<i>No, not A and not B.</i>	67	9	66
<i>No, A and not B.</i>	2	28	9
<i>No, if A then not B.</i>	0	34	0
<i>No, not A.</i>	11	3	8
<i>No, not B.</i>	2	21	6

In sum, the model theory may be unique in its prediction that negated conjunctions should be more difficult than negated disjunctions, and the data from Experiment 2 corroborate the hypothesis.

## General Discussion

Two experiments showed that participants find negated conjunctions more difficult to understand and to formulate than negated disjunctions, whereas previous research has established that affirmative conjunctions are easier to understand than affirmative disjunctions (García-Madruga, Moreno, Carriedo, Gutiérrez, & Johnson-Laird, 2001). The data therefore revealed a novel interaction between the grammatical form of a sentence and its polarity, and they corroborated a theory of negation based on mental models (Khemlani et al., 2012). The theory posits that individuals do not know the negations corresponding to the different sentential connectives. They have to construct them on an ad hoc basis, so they consider a sequence of conjunctive models of possibilities, checking that they render the corresponding affirmative assertion false. This *enumerative negation* hypothesis predicts that individuals should find it easy to comprehend and formulate denials of inclusive disjunctions of the form *A or B or both*, because the first model that individuals should consider is the only true negation of the disjunction: *not-A and not-B*. In contrast, the hypothesis predicts that a conjunction, *A and B*, should be difficult to deny, because its denial is equivalent to *not-A or not-B or neither*, and so individuals need to envisage fully explicit models of three separate possibilities.

Denials of conditionals with the structure *if A then B* are an intermediate case. They should be easier to comprehend than denials of conjunctions but harder to comprehend than denials of disjunctions. The correct negation of the conditional, *A and not-B*, should be more difficult to envisage because, according to the enumerative negation hypothesis, this model is unlikely to be the first one that comes to mind. And their denials should also be susceptible to a reduction of scope, because *if* introduces a subordinate clause, whereas neither of the other sorts of compound contains a subordinate clause. Hence, some individuals should deny a conditional by using another conditional: *if A then not-B*.

When individuals had to understand affirmations and denials in Experiment 1, their evaluations of what was possible corroborated the model theory's predicted interaction. For affirmations, they found it easier to understand conjunctions than disjunctions, but for denials, they found it easier to understand disjunctions than conjunctions. The affirmation of a conjunction yields one possibility, and the affirmation of a disjunction yields three possibilities. In contrast, the denial of a conjunction requires an inference of three possibilities, and the denial of a disjunction requires an inference of only one possibility. The inferential aspect of this task for negatives may explain why it is so much harder than merely listing the three possibilities corresponding to an inclusive disjunction. Experiment 2 corroborated the interaction. Both experiments also revealed the occurrence of two sorts of negation of conditionals, as did a study by Handley and colleagues (Handley, Evans, & Thompson, 2006). These authors argue that the negation of a conditional, *if A then B*,

should be *if A then not-B*. This view, however, has a major drawback: it no longer treats negations as contradicting corresponding affirmatives. Likewise, it offers no principled explanation of why some individuals do take *A and not-B* to be the denial of a conditional, or why most people take it to falsify a conditional too (Espino & Byrne, 2011; Evans, Newstead, & Byrne, 1993; Johnson-Laird & Tridgell, 1972).

## Acknowledgements

This research was supported by a National Science Foundation Graduate Research Fellowship to the first author, and by National Science Foundation Grant No. SES 0844851 to the third author to study deductive and probabilistic reasoning. We are grateful to Jay Atlas, Jeremy Boyd, Herb Clark, Alan Garnham, Sam Glucksberg, Adele Goldberg, Geoff Goodwin, Jennifer Heil, Olivia Kang, Philipp Koralus, Mark Liberman, Max Lotstein, Anna Liu, Paula Rubio, Carlos Santamaría, and Elizabeth Sucuyan for their helpful suggestions and criticisms.

## References

- Barrouillet, P., Grosset, N., & Lecas, J.F. (2000). Conditional reasoning by mental models: chronometric and developmental evidence. *Cognition*, 75, 237-266.
- Braine, M.D.S., & O'Brien, D.P., Eds. (1998). *Mental logic*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Clark, H.H. (1974). Semantics and comprehension. In T.A. Sebeok (Ed.) *Current trends in linguistics, Vol. 12: Linguistics and adjacent arts* (pp. 1291-1428). The Hague: Mouton.
- Clark, H.H., & Chase, W.G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, 3, 472-517.
- Espino, O., & Byrne, R.M.J. (2011). It is not the case that if you understand a conditional you know how to negate it. *Journal of Cognitive Psychology*, in press.
- Evans, J.St.B.T., Newstead, S.E., & Byrne, R.M.J. (1993). *Human reasoning: The psychology of deduction*. Hillsdale, NJ: Erlbaum.
- García-Madruga, J.A., Moreno, S., Carriedo, N., Gutiérrez, F., & Johnson-Laird, P.N. (2001). Are conjunctive inferences easier than disjunctive inferences? A comparison of rules and models. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 54, 613-632.
- Gazdar, G., & Pullum, G. K. (1976). Truth-functional connectives in natural language. *Papers from the Twelfth Regional Meeting of the Chicago Linguistic Society*, 220-234.
- Giroto, V., & Johnson-Laird, P.N. (2004). The probability of conditionals. *Psychologia*, 47, 207-225.
- Handley, S.J., Evans, J. St. B.T., & Thompson, V.A. (2006). The negated conditional: A litmus test for the suppositional conditional? *Journal of Experimental Psychology: Learning, Memory, Cognition*, 32, 559-569.

- Hoosain, R. (1973). The processing of negation. *Journal of Verbal Learning and Verbal Behavior*, 12, 618-626.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.
- Johnson-Laird, P.N., Byrne, R.M.J., & Girotto, V. (2009). The mental model theory of conditionals: A reply to Guy Politzer. *Topoi*, 28, 75-80.
- Johnson-Laird, P. N., & Tridgell, J. (1972). When negation is easier than affirmation. *Quarterly Journal of Experimental Psychology*, 24, 87-91.
- Khemlani, S., Orenes, I., & Johnson-Laird, P.N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, in press.
- Rips, L.J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

# mReactr: A computational theory of deductive reasoning

Sangeet Khemlani and J. Gregory Trafton  
khemlani@aic.nrl.navy.mil, trafton@itd.nrl.navy.mil

Navy Center for Applied Research in Artificial Intelligence  
Naval Research Laboratory, Washington, DC 20375 USA

## Abstract

The mReactr system is a computational implementation of the mental model theory of reasoning (Johnson-Laird, 1983) that is embedded within the ACT-R cognitive architecture (Anderson, 1990). We show how the memory-handling mechanisms of the architecture can be leveraged to store and handle discrete representations of possibilities, i.e., mental models, efficiently. Namely, the iconic representation of a mental model can be distributed, in which each component of a model is represented by a “chunk” in ACT-R’s declarative memory. Those chunks can be merged to create *minimal* mental models, i.e., reduced representations that do not contain redundant information. Minimal models can then be modified and inspected rapidly.

We describe three separate versions of the mReactr software that minimize models at different stages of the system’s inferential processes. Only one of the versions provides an acceptable model of data from an immediate inference task. The resulting system suggests that reasoners minimize mental models only when they initiate deliberative mental processes such as a search for alternative models.

**Keywords:** reasoning, mental models, immediate inferences, mReactr, ACT-R

## Introduction

People regularly make complex deductive inferences. For instance, if you know that *none of the lawyers in the room are men*, you might refrain from asking any of the men in the room for legal advice. If so, you have made an “immediate” inference from a single premise:

1. None of the lawyers are men.  
Therefore, none of the men are lawyers.

The inference is *valid* because its conclusion must be true given that its premise is true (Jeffrey, 1981, p. 1). You likely followed up the deductive inference above with an inductive inference:

2. None of the men are lawyers.  
Therefore, they do not possess legal knowledge.

The second inference is inductive – the conclusion is not necessary given the truth of the premise.

How do reasoners make deductive and inductive inferences like the ones above? One prominent answer is that they construct mental simulations of the things they already know or believe. They then manipulate those simulations to obtain information they did not have at the outset. The idea that reasoning depends on building

simulations, or mental models, is the fundamental intuition behind the mental model theory of reasoning (Johnson-Laird, 1983). In the present paper, we outline the theory and address one of its major limitations, namely its inability to explain how models are stored and manipulated in memory. We describe a computational implementation of the theory that is embedded within the ACT-R cognitive architecture (Anderson, Bothell, Byrne, Douglass, Lebiere, & Qin, 2004), and we show how the memory-handling mechanisms of the architecture can be leveraged to store and handle mental models efficiently.

## Reasoning and mental models

The “model” theory of reasoning proposes that when individuals comprehend discourse, they construct mental models of the possibilities consistent with the meaning of the discourse (Johnson-Laird, 2006). The theory depends on three main principles: 1) Individuals use a representation of the meaning of a premise and their knowledge to construct mental models of the various possibilities to which the premises refer. 2) The structure of a model corresponds to the structure of what it represents (see Peirce, 1931-1958, Vol. 4), and so mental models are iconic insofar as possible. 3) The more models a reasoner has to keep in mind, the harder an inference is. On a model-based account, a conclusion is *necessary* if it holds in all the models of the premises and *possible* if it holds in at least one model of the premises.

mReasoner (Khemlani, Lotstein, & Johnson-Laird, under review) is a unified computational implementation of the mental model theory of reasoning. It implements two interoperating systems for reasoning:

- a) An intuitive system (system 1) for building an initial mental model and drawing rapid inferences from that model
- b) A deliberative system (system 2) for more powerful recursive processes that search for alternative models. This system can manipulate and update the initial model created in system 1, and it can modify conclusions

The system is akin to dual-process models of reasoning (see, e.g., Evans, 2003, 2007, 2008; Johnson-Laird, 1983, Ch. 6; Kahneman, 2011; Sloman, 1996; Stanovich, 1999; Verschueren, Schaeken, & d’Ydewalle, 2005). Below, we describe the various processes that each system implements.

# The intuitive system

*Model building.* The system builds an initial model from the meaning of a premise, and it updates that initial model if additional premises occur. The system begins by building a model with a small, arbitrary set of individuals. For example, the model of *some of the artists are bohemians* is built by first constructing a set of artists:

artist  
artist  
artist  
artist

In the diagram above, each row represents an individual with the property of being an artist, and so the model as a whole represents a finite number of individuals. Mental models are representations of real individuals, not letters or words, which we use here for convenience. The meaning of the assertion *some of the artists are bohemians* provides instructions for the system to add additional properties, namely the property of being bohemian. The model is updated accordingly:

artist	bohemian
artist	bohemian
artist	
artist	
	bohemian

The model therefore represents a set of individuals, some of whom are both artists and bohemians, some of whom are just artists, and one who is just a bohemian. Once a premise has been represented, the system can assess whether the given conclusion is true in the initial model.

*Assessing initial conclusions.* When reasoners have to assess a given conclusion, the system inspects the initial model to verify that the given conclusion holds or does not hold. For instance, suppose that reasoners are asked to decide whether it is possible that *all bohemians are artists* given the previous premise. From the model above, the system initially responds in the negative, i.e., the putative conclusion is impossible. The process is simple, and the response is rapid. However, it is incorrect: the system's ability to assess and generate initial conclusions is fallible. For instance, one can indeed show that *all of the bohemians are artists* is possible. To revise its initial conclusion, the system needs to find an alternative model in which both the premise and conclusion hold. We turn to the second system to explain how such a model is found.

# The deliberative system

*Searching for alternative models.* In the preceding section, we focused on how the system assesses conclusions based on an initial model. However, the conclusions it draws can be invalid. System 2 attempts to revise initial conclusions by searching for alternative models. To do so, it uses three

separate operations: *adding* properties to individuals, *breaking* one individual into two separate individuals, and *moving* properties from individual to another (see Khemlani, Lotstein, & Johnson-Laird, under review). The operations correspond to those that naïve participants spontaneously adopt when they reason about syllogisms (as evidenced by their manipulations of external models, see Bucciarelli & Johnson-Laird, 1999). Consider our example above. After an individual represents the initial model and provides an initial conclusion that is false, it can modify that conclusion by adding properties to the initial model. If the system can successfully create a model in which *some of the artists are bohemians* and *all of the artists are bohemians* are both true, then it can conclude that it is possible, but not necessary that all of the artists are bohemians. By adding properties, the system finds such a model:

artist	bohemian
artist	bohemian
artist	bohemian
artist	bohemian
artist	bohemian

The new model, which contains individuals who are all artists and bohemians at the same time, refutes the conclusion that it is impossible that *all the bohemians are artists*. However, the search for alternative models places a considerable tax on working memory. Until now, the limitations of the model theory have prevented it from characterizing the cost of holding models in memory.

# Limitations of the model theory

The model theory and its unified implementation explain many aspects of how people make inferences. The theory provides an explanation of how discourse is mapped to high-level representations. It accounts for why some reasoning problems are hard and others are easy (Khemlani & Johnson-Laird, 2012). It provides working algorithms for how individuals assess whether a given conclusion is possible, necessary, or consistent with a given set of premises. And the model theory as a whole can explain deductive, inductive, and abductive inferences (Johnson-Laird, 2006). As such, it represents a unified theory of reasoning.

The theory is limited by design, however, in that most of its predictions are qualitative. For instance, it can explain that an inference that requires a reasoner to hold one model in working memory should be easier than an inference that requires three models in memory, but it cannot explain or predict the degree of the difficulty. Is the former inference twice as easy or thrice as easy as the latter? And how long should each inference take? The computational model is silent on these matters, because it specifies only those algorithms that are pertinent to how individuals make inferences. It ignores other aspects of cognition, such as how models are stored in working memory and how they are retrieved. To overcome these limitations, we

implemented the theory in the ACT-R cognitive architecture, and we describe the resulting hybrid system below. The framework, which we call *mReactr* (mReasoner + ACT-R), imbues the model theory with a more robust account of how models are represented and manipulated. It also stands as a novel application of the ACT-R system, which has had only limited success in accounting for behavior on high-level deductive tasks (e.g., Emond, 2003, and Ragni, Fangmeier, & Brüssow, 2010).

### mReactr: Mental models in memory

The ACT-R cognitive architecture is a modular computational theory of human cognition (Anderson et al., 2004). It is a collection of interoperating modules that store and retrieve information relevant to a particular task. The central control system, called the *procedural* module, directs the way the system accesses capacity-limited buffers. The system also contains a *declarative* module for storing knowledge of facts and procedures. Facts are stored in structures called *chunks*, and procedures are represented by *productions*, i.e., condition-action pairings. The productions direct the procedural model to monitor the buffers for the existence of certain sorts of chunks, and if a chunk appears in a buffer in the manner that a production expects, the relevant action will be initiated. Each chunk has an associated level of activation. If the chunk’s activation is low, ACT-R will take longer to retrieve it, but if it is high, it will be retrieved quickly. Accordingly, the system automatically calculates the time it takes to trigger productions, modify goals, retrieve chunks, and clear buffers.

The architecture efficiently manages chunks in declarative memory. In particular, if it detects that two chunks are identical in every respect, it merges those chunks into one chunk. The merged chunk will then have a higher activation than either individual chunk. This “chunk-merging” feature of the system is particularly important for how mental models are handled.

The mReactr system is an implementation of mental model theory in ACT-R. The system can build initial models and assess putative conclusions (system 1) and likewise it can modify those models to search for alternative models (system 2). It stores models in declarative memory by assigning each individual to a separate chunk. Thus, the system will store the model of *all the artists are bohemians* as five separate chunks:

artist	bohemian	(chunk 1)
artist	bohemian	(chunk 2)
artist		(chunk 3)
artist		(chunk 4)
	bohemian	(chunk 5)

The system therefore represents the model in a distributed fashion, as a collection of chunks with similar properties. However, several of the separate chunks are identical to one another, and so ACT-R will try to merge those chunks

automatically, to produce just a condensed version of the model:

artist	bohemian	(chunk 1')
artist		(chunk 3')
	bohemian	(chunk 5')

By merging the chunks, the underlying architecture automatically produces a *minimal* mental model, i.e., a model that only retains information about the different *types* of individuals. The process of minimizing mental models is not something that is built into mental model theory as yet; the basic mechanisms of memory management within ACT-R provide a way to efficiently store and retrieve models. But, is there any evidence that reasoners minimize models? And if so, do they minimize models at the outset, or at a later stage of processing? To answer both of these questions, we compared mReactr’s accuracy and latency predictions against data from a recent reasoning experiment.

### An assessment of the model

We assessed whether the mReactr system could model that data from a recent study on so-called “immediate” deductive inferences akin to our introductory example above (1). Psychologists have investigated immediate inferences for many years (e.g., Begg & Harris, 1982; Newstead & Griggs, 1983; Wilkins, 1928), but have yet to resolve how logically untrained individuals make them. The inferences are based on singly-quantified assertions in four different *moods* of the premise:

- All the Xs are Ys
- Some of the Xs are Ys
- None of the Xs are Ys
- Some of the Xs are not Ys

and 8 different sorts of conclusion (4 moods by 2 *figures*, i.e., arrangements of terms ‘X’ and ‘Y’). Therefore, there are 32 possible immediate inference problems based on these premises. A typical problem looks like this:

Suppose that some of the students are Virginians.  
Is it possible that all of the Virginians are students?

Immediate inferences were chosen because the model theory and mReactr distinguish between the relative difficulties of three sorts of immediate inference: a) zero-model inferences, b) one-model inferences, and c) multiple model inferences.

Zero-model inferences are those in which the conclusion is identical to the premise, and so individuals needn’t even build a model to be able to solve the problem. For instance, consider the following problem:

All the aldermen are barters.  
Is it possible that all the aldermen are barters?

Reasoners should realize that the answer is true immediately; however, they should nevertheless need to extract the meanings from the assertions, and they need to establish a set of subgoals in order to infer a conclusion.

One-model inferences are those in which the conclusion holds in the initial model of the premise, and so individuals can rapidly determine that an assertion is possible. For example:

All the aldermen are barters.

Is it possible that some of the barters are aldermen?

Reasoners have to construct the meanings of the assertions, use them to build a model, and evaluate the truth of the conclusion in the model.

When the conclusion fails to hold in the initial model, but does hold in an alternative to it, then participants have to search for that alternative model. We refer to such problems as multiple-model inferences. For instance:

All of the aldermen are barters.

Is it possible that some of the barters are not aldermen?

For multiple-model inferences, mReactr predicts that reasoners extract the meaning of the assertion and build an initial model, but their initial model suggests an erroneous evaluation of whether or not the conclusion is possible. To obtain a correct evaluation, reasoners have to modify their initial model to produce an alternative model. The theory therefore predicts that zero-model inferences should be easier than one-model inferences, and one-model inferences should be easier than multiple-model inferences. Likewise, mReactr provides precise latency predictions for how long zero-, one-, and multiple-model inferences should take.

We used mReactr to simulate an experiment conducted by Khemlani, Lotstein, & Johnson-Laird (in revision). In the study, the participants carried out all 32 problems (4 sorts of premise x 8 sorts of conclusion), and they responded “yes” or “no” to a conclusion about a possible conclusion to each problem. The contents of the problems were based on nouns referring to common occupations. The instructions stated that the task was to respond to questions about a series of assertions concerning what was possible given the truth of the assertion.

## Simulation

Our goals in simulating immediate inference data were two-fold. First, we sought to test the fidelity of the mReactr system as an instantiation of the model theory. We restricted our simulation to valid immediate inferences, i.e., 22 of the 32 problems. The theory distinguishes between three sorts of problem, and so mReactr should reflect the same distinction. A failure of the computational model to capture those data indicates a poor implementation of the model theory. We retained all of the default values of the ACT-R architecture, except we increased the architecture’s default

tracking ability so that it could track 10 individual chunks (i.e., the :declarative-num-finsts parameter).

Second, we attempted to examine whether mReactr could fit the data better when it actively engaged in minimizing models by merging chunks. We created three separate versions of mReactr:

- 1) no chunk-merging version
- 2) system 1 chunk-merging version
- 3) system 2 chunk-merging version

In the *no chunk-merging* version, chunks were kept separate and ACT-R’s automated chunk-merging capability was disabled. In the *system 1 chunk-merging* version, chunks were merged before the system engaged in any inferential processes. And in the *system 2 chunk-merging* version, chunks were kept separate in the initial model. They were merged only when mReactr initiated a search for alternative models. The best performing version of the theory can help establish whether and when models should be minimized.

## Results and discussion

The results of the experiment corroborated the theory’s predictions of difficulty (Khemlani et al., in revision), and they yielded the following trend: reasoners were 98% correct for zero-model problems, 85% correct for one-model problems, and 71% correct for multiple-model problems (Page’s trend test,  $L = 340.0$ ,  $z = 3.88$ ,  $p < .0001$ ). Immediate inferences are relatively easy to deduce, nevertheless participants exhibit predictable patterns of difficulty. The mean latencies also corroborated the predicted trend: 4.30 s for zero-model problems, 5.17 s for one-model problems, and 5.41 s for multiple-model problems (Page’s trend test,  $L = 336.0$ ,  $z = 3.33$ ,  $p < .0005$ ).

Figure 1 illustrates the empirical latencies and the predicted latencies from the different versions of mReactr. As the figure shows, the system yielded the closest match to

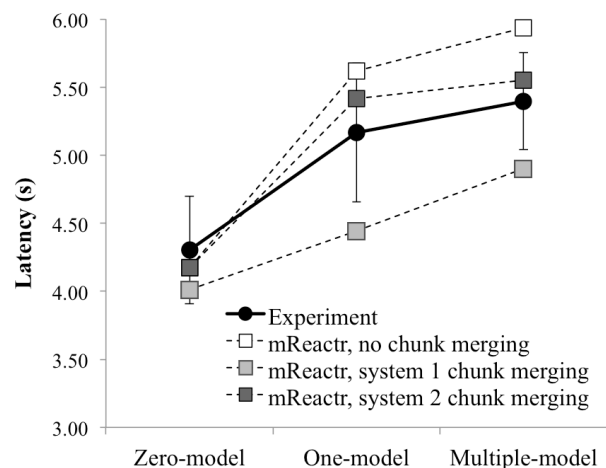


Figure 1: Participants’ mean latencies (in s) to solve zero-, one-, and multiple-model problems, and the latencies predicted by the three separate versions of mReactr.



mReactr version	Model fits			
	R <sup>2</sup>	RMSE	Goodness of fit	
			D	p
<i>a) By problem type</i>				
No chunk-merging	.99	.40	.67	.60
System 1 chunk-merging	.94	.54	.67	.60
System 2 chunk-merging	.99	.18	.67	.60
<i>b) By immediate inference</i>				
No chunk-merging	.45	.70	.41	.05
System 1 chunk-merging	.23	.86	.50	.008
System 2 chunk-merging	.45	.57	.18	.86

Table 1: Model fits for the three versions of mReactr by problem type (zero-, one-, and multiple-model problems) and by the 22 valid immediate inferences. Note: a lack of significance for the Kolmogorov-Smirnov D statistic indicates a good fit.

the data when chunk-merging was initiated at a later stage of processing, i.e., the *system 2 chunk-merging* version ( $R^2 = .99$ ,  $RMSE = .18$ ). When chunk-merging was disabled in the *no chunk-merging* version, the system did well, but it took too long to search for alternative models, ( $R^2 = .99$ ,  $RMSE = .40$ ). In the *system 1 chunk-merging* version, mReactr performed faster than participants tend to perform, yielding a poorer fit of the data ( $R^2 = .94$ ,  $RMSE = .54$ ).

Across all three simulations, the system negatively correlated with participants' accuracy ( $r$ 's  $< -.90$ ,  $p$ 's  $< .0001$ ). Likewise, the simulations fit the latencies well. Table 1a gives the model fits for the three separate versions of the system across the three types of problems as a whole, as well as across the 22 different problems separately.

We ran a separate set of analyses to examine how the three versions of the system modeled the 22 valid immediate inferences separately (see Table 1b). This set of analyses would by definition yield poorer model fits as a result of the inherent variation between different problems, and so any significant correlation can be construed as support for the theory. The analysis replicated and elaborated upon the aggregated results. The system fit the data moderately well with chunk-merging turned off, but its RMSE was relatively high ( $R^2 = .45$ ,  $RMSE = .70$ ), and a Kolmogorov-Smirnov goodness of fit analysis indicated that the system exhibited reliably different distributional properties than that of the experiment ( $D = .41$ ,  $p = .05$ ). Likewise, the system provided a relatively poor fit of the data when models were minimized at the outset ( $R^2 = .23$ ,  $RMSE = .86$ ) and so mReactr produced results that came from a separate distribution (Kolmogorov-Smirnov test,  $D = .50$ ,  $p = .008$ ). Only when models were minimized before the system searched for alternative models did the system fit the data well ( $R^2 = .45$ ,  $RMSE = .57$ ), and the goodness-of-fit analysis indicated a close match between mReactr and the data (Kolmogorov-Smirnov test,  $D = .18$ ,  $p = .86$ ).

The results of the simulations showed that across all three version of mReactr, the system successfully implemented

the model theory's predictions of difficulty, and it distinguished between zero-, one-, and multiple-model problems. However, the system performed best only when it initiated chunk-merging before it began a search for alternative models. The results have important implications for an overlooked process in the psychology of reasoning: representational minimization.

## General Discussion

The computational theory, mReactr, is system implemented in the ACT-R cognitive architecture that simulates the construction of mental models in order to draw immediate inferences from singly-quantified premises. The cognitive architecture comes equipped with the ability to manage its declarative memory efficiently, namely by merging identical chunks. mReactr repurposes this chunk-merging functionality to produce minimal mental models at a particular stage of inference. At the outset, mReactr uses the same collection of iconic representations as is specified in the model theory. However, the full representation is ephemeral, and it lasts only until the system starts to modify the model. If and until the system initiates a search for alternative models, it minimizes the model. This process maps onto the psychological strategy of abstracting over the different sorts of individuals.

The theory predicts that individuals should be faster and more accurate when an inference can be drawn from an identity in the meanings of the assertions, i.e., when they do not need to consult a mental model. They should be next fastest and accurate when an inference can be drawn from the initial model constructed in system 1. And they should be slowest and least accurate when an inference can be drawn only from the discovery of an alternative model constructed in system 2. These rank-order predictions were borne out in the data from an experiment that tested all 22 valid inferences about possible conclusions in the set of 32 inferences.

The system we describe is limited, however, and it can be improved to yield a more fine-grained processing account of the data. We suggest two separate ways of proceeding. One way to improve the fit of the system is to make the system sensitive to the direction in which it scans models. For instance, if reasoners read a particular premise, e.g., *all artists are bohemians*, they may be biased to scan the model in the opposite directions by considering bohemians before artists. This *figural* bias is widely documented in syllogistic reasoning (Khemlani & Johnson-Laird, 2012) and it is likely to make a difference when reasoning about immediate inferences as well.

Another way to improve the system's overall performance is to consider the process of model minimization as something that may or may not occur depending on strategy and individual differences (Bucciarelli & Johnson-Laird, 1999). Some reasoners may be more likely to minimize their models, and others might prefer to keep the full model representation in mind.

In sum, model minimization is an important way in which individuals can optimize the storage and retrieval of mental models. It is embodied in the computational system of deductive reasoning that we developed.

### Acknowledgments

This research was funded by a National Research Council Research Associateship awarded to SK and ONR Grant #s N0001412WX30002 and N0001411WX20516 awarded to JGT. We are also grateful to Len Breslow, Magda Bugajska, Hua Gao, Tony Harrison, Cathy Haught, Laura Hiatt, Phil Johnson-Laird, Gorka Navarrete, Marco Ragni, and Tobias Sonntag for their helpful comments.

### References

- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review* 111, 1036-1060.
- Begg, I., & Harris, G. (1982). On the interpretation of syllogisms. *Journal of Verbal Learning and Verbal Behavior*, 21, 595-620.
- Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, 23, 247-303.
- Emond, B. (2003). Cognitive representations and processes in syllogistic reasoning: existential graphs and cognitive modelling. *Psychologica*, 32, 311-340.
- Evans, J.St.B.T. (2003). In two minds: Dual process accounts of reasoning. *Trends in Cognitive Sciences*, 7, 454-459.
- Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove, UK: Psychology Press.
- Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology*, 59, 255-278.
- Jeffrey, R. (1981). *Formal logic: Its scope and limits* (2nd Ed). New York: McGraw-Hill.
- Johnson-Laird, P.N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge: Cambridge University Press.
- Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. N. (2006). *How we reason*. Oxford, UK: Oxford University Press.
- Johnson-Laird, P.N., & Byrne, R.M.J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Strauss, Giroux.
- Khemlani, S., & Johnson-Laird, P.N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, in press.
- Khemlani, S., Lotstein, M., & Johnson-Laird, P.N. (in revision). Immediate inferences from quantified assertions. Manuscript in revision.
- Khemlani, S., Lotstein, M., & Johnson-Laird, P.N. (under review). A unified theory of syllogistic reasoning. Manuscript under submission.
- Newstead, S.E., & Griggs, R.A. (1983). Drawing inferences from quantified statements: A study of the square of opposition. *Journal of Verbal Learning and Verbal Behavior*, 22, 535-546.
- Peirce, C.S. (1931-1958). *Collected papers of Charles Sanders Peirce*. 8 vols. Hartshorne, C., Weiss, P., & Burks, A. (Eds.) Cambridge, MA: Harvard University Press.
- Ragni, M., Fangmeier, T., & Brüssow, S. (2010). Deductive spatial reasoning: From neurological evidence to a cognitive model. In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th International Conference on Cognitive Modeling* (pp. 193-198). Philadelphia, PA: Drexel University.
- Sloman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3-22.
- Stanovich, K.E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
- Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. *Thinking & Reasoning*, 11, 278-293.
- Wilkins, M. C. (1928). The effect of changed material on the ability to do formal syllogistic reasoning. *Archives of Psychology*, 16, No. 102.

# The Specificity of Online Variation in Speech Production

Christo Kirov(kirov@cogsci.jhu.edu)

Department of Cognitive Science, 3400 N. Charles Street  
Baltimore, MD 21218 USA

Colin Wilson (colin@cogsci.jhu.edu)

Department of Cognitive Science, 3400 N. Charles Street  
Baltimore, MD 21218 USA

## Abstract

Phonetic variation is sensitive to lexical properties of words, such as frequency and neighborhood density, as well as contextual properties, such as predictability. Previous studies of lexically-induced variation have observed that both vowels and consonants are phonetically enhanced in words from dense neighborhoods, and have suggested that this effect is modulated only by the number and frequency of the neighbors. To determine whether contextual variation is driven by cognitive processes similar to those underlying lexical enhancement, three experiments examined the effect of contextually salient neighbors on the phonetic realization of vowels and initial consonant aspiration. Enhancement was found only for consonants, and only when the neighbor differed from the target word in a single feature. Unlike lexical effects, contextually-driven phonetic enhancement reflects a highly specific competition among words, a finding that can be rationalized in terms of the utility of speaker effort within a Bayesian model of word communication.

**Keywords:** Speech production; lexical competition; communication; Bayesian modeling

## Introduction

Competition among alternatives, and the need to resolve competition efficiently and correctly, are pervasive in speech perception and speech production (e.g., Luce & Pisoni 1998, Marslen-Wilson & Zwitserlood 1989, Dell & Gordon 2003). Listeners must determine the speaker's intended message as rapidly as possible given an inherently ambiguous signal. In speech production, words and sublexical units that are partially consistent with the intended message compete for realization at multiple levels of processing. A number of studies have examined how such competitive processes are reflected in the fine-grained phonetic realization of speech sounds.

The number and relative frequency of phonologically-similar words in the lexicon (lexical neighbors) are known to affect phonetic realization. We refer to such affects as *offline* because they appear to depend on relatively static lexical relationships among words rather than dynamic contextual factors. Researchers have found that “hard” words, those with low lexical frequency and many high frequency neighbors, tend to be phonetically enhanced relative to “easy” words with high frequency and few neighbors. Hard words beginning with aspirated consonants have longer aspiration (as measured by voice onset time, VOT) than easy words (Goldinger & Summers, 1989). They are pronounced with an expanded vowel space (Munson & Solomon, 2004; Wright, 2003), and also show increased vowel nasalization and vowel-to-vowel coarticulation (Scarborough, 2004).

Interestingly, offline phonetic enhancement effects seem to be rather general: they appear to depend only on the (frequency-weighted) density of a word's neighborhood, not on the precise phonological relationships between the word and its neighbors. For example, Scarborough (2004) found that words that were particularly confusable by their nasal consonant (i.e., had one or more lexical neighbors that differed in the position of the nasal) did not show greater vowel nasalization than words that were not similarly confusable. Words like *stem*, with minimal pair neighbor *step*, showed similar levels of nasalization as words like *plank*, with no nasal-differing neighbors in the lexicon. Similarly, Goldinger & Summers (1989) found more VOT enhancement for voiceless-initial words from dense neighborhoods than those from sparse neighborhoods, even though both sets of words had exactly one minimal pair lexical neighbor that began with a voiced sound. The generality of offline phonetic enhancement suggests that it is driven by competition among entire lexical items, not among sublexical units.

Unlike offline effects, *online* effects on phonetic realization by definition depend on the context in which a word is uttered, such as the discourse topic, transitional probabilities conditioned on preceding material, and other contextually-salient words. A classic online effect is the Lombard Reflex, a set of vocal changes that include increases in amplitude and pitch that occur when speakers attempt to talk over noise (Lau, 2008; Zhao & Jurafsky, 2009). More recently, a number of corpus-based studies have found that the contextual predictability of speech elements, including phonemes and syllables, is inversely related to their length. Less predictable elements tend to be longer (e.g., Cohen-Priva & Jurafsky 2008, Aylett & Turk 2004, van Son & Pols 2003).

This paper aims to expand our understanding of online phonetic enhancement effects, looking not just at predictability effects but at how a word's phonological neighborhood in context — the sound structure of contextually salient competitors — affects phonetic realization along several dimensions of possible hyperarticulation. This will provide further insight into how competition between similar words plays out during speech production. In previous work, Baese-Berke & Goldrick (2009) found that VOT is lengthened when a voiceless-stop initial word is pronounced in the context of a voiced-initial neighbor (in comparison to the context of a phonologically unrelated filler word). For example, *cot* shows increased initial VOT in the context of *got*, but not in

the context of *fan*.

Baese-Berke & Goldrick additionally put forward the claim that this online enhancement of VOT, and perhaps online enhancement effects in general, have the same underlying cognitive mechanism as the offline enhancement effects reviewed earlier. If this hypothesis were true, we would expect offline and online effects to be empirically parallel. In particular, we would expect to find an online analogue of every offline effect. Baese-Berke & Goldrick’s VOT enhancement effect mirrors that found offline by Goldinger & Summers (1989), providing partial support of the hypothesis. However, to our knowledge researchers have yet to investigate online analogues of other offline effects, including vowel space expansion and vowel nasalization.

Furthermore, if offline and online phonetic variation are driven by the same processes of cognitive competition, we would expect the generality of offline effects to be found in online enhancement as well. Just as offline VOT enhancement does not seem to be modulated by the specific phonological structure of a word’s lexical neighbors, online VOT enhancement should not be affected by the sound structure of the words that have become salient in the speech discourse. That is, any type of neighbor that is active in the speech context should induce online enhancement. Baese-Berke & Goldrick (2009) investigated only contextually salient neighbors of one kind, namely those differing in the voicing of the initial consonant, and consequently the results of that study cannot determine whether online competition is general or specific.

We examined this issue in three experiments, and found that online phonetic enhancements differ from offline effects in two significant respects. Most importantly, online effects appear to be sensitive to the phonological properties of words in the local discourse. Only competitors that have particular phonological relations with the target word — relations defined by word position and segmental makeup — induce online hyperarticulation. We show that these results are predicted if speakers expend the effort involved in phonetic enhancement only when that could contribute to listeners’ recognition accuracy, assuming a Bayesian model of word recognition.

## Experiments

All experiments used an experimental paradigm adapted from Baese-Berke & Goldrick (2009). The goal behind the paradigm is to simulate a situation where a speaker must accurately communicate a word to a listener even though contextually salient competitor words provide opportunities for miscommunication. The paradigm involves two participants, one playing the role of speaker and the other the role of listener. Each participant sits at a separate computer terminal, which is not visible to the other participant. In each trial of the experiment, two or more words appear on both screens: a target word along with competitor words that are sometimes neighbors of the target. After approximately 1000ms,

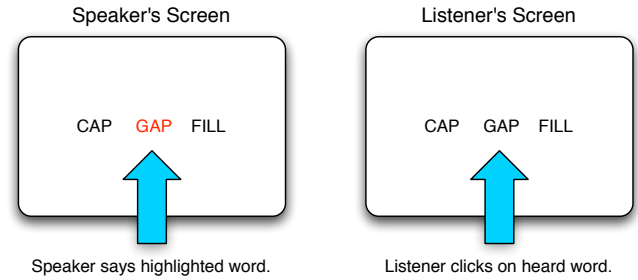


Figure 1: Experimental paradigm.

the target word becomes highlighted on the speaker’s screen, who then produces it aloud. At this point, the listener clicks the word that was heard — the same word produced by the speaker, if communication is successful. The speaker’s pronunciation of the target is recorded and analyzed acoustically after the experiment. The experimental setup is illustrated in Figure 1. This paradigm has the advantage of being able to precisely control a target word’s “context” (the neighbors that appear on-screen with it) and including motivation for the speakers to communicate clearly, as they are made aware if the listener fails to select the target word.

### Experiment 1: Online Vowel Space Expansion

If online effects have the same underlying cause as offline effects, we expect to find an online vowel expansion effect mirroring the offline effect found (e.g., Munson & Solomon 2004). To test this hypothesis, we presented target words in the context of neighbors that differed from the target only in the vowel position. A condition where the same targets were presented with unrelated filler words was used as a baseline for comparison.

Table 1: Table of conditions for Experiment 1.

Target	Vowel	Filler
CAT	KIT	DOLL

Following previous work on vowel space dispersion, the dependent acoustic variable measured was the Euclidean distance of each target vowel from the center of each subject’s vowel space (defined as the subject’s mean F1 and F2 formant values). Participants (N=18) produced each of 16 target words in each condition. Target position on screen was counterbalanced across speakers. Order of conditions for a given target was also counterbalanced to avoid confounds with repetition effects. Results were analyzed using linear mixed results regression (*lme4*) in R, with condition as a fixed effect and subject and target item as random effects. Results are summarized in Table 2; p-values were obtained using Markov Chain Monte Carlo (*pMCMC*). There was no significant effect of onscreen neighbor on vowel space dispersion in comparison to onscreen filler words, suggesting that there is

no online analogue to offline vowel space dispersion effects. This provides evidence that online and offline effects do not share the same underlying cause.

Table 2: Experiment 1: Statistical results.

Condition	Coeff	SE	<i>t</i>	<i>p</i>
Vowel Neighbor	1.707	10.696	0.160	< 0.8732

## Experiment 2: Positional Specificity of Online VOT Enhancement

Baese-Berke & Goldrick (2009) found VOT lengthening in the initial segment of a target word presented with an on-screen neighbor differing in the voicing of its initial segment. This experiment tested to see if any kind of neighbor can induce this enhancement effect, as might be expected if online effects lack specificity in the way that offline effects do. Target words were presented in the context of neighbors that differed only in onset (a replication of the Baese-Berke & Goldrick study using voice-differing neighbors), vowel, or coda positions. Different neighbor types were matched for frequency (pairwise paired t-tests, all  $p > 0.3$ ).

Table 3: Table of conditions for Experiment 2.

Target	Onset	Vowel	Coda	Filler
CAP	GAP	CUP	CAT	DOLL

Participants (N=24) produced each of 48 target words twice in one of the four conditions, so that each word appeared in all conditions every four subjects. Results are shown in Figure 2 and Table 4. Only onset-differing neighbors appear to cause a significant VOT enhancement effect over fillers. This result suggests that online enhancement effects depend at least on position-level sublexical processing and are thus more *specific* than offline effects.

Table 4: Experiment 2: Statistical results.

Condition	Coeff	SE	<i>t</i>	<i>p</i>
Onset Neighbor	2.07000	0.80555	2.570	< 0.0102*
Vowel Neighbor	0.03449	0.80555	0.043	< 0.9659
Coda Neighbor	0.54367	0.80555	0.675	< 0.4998

Interestingly, the effects found seem to be limited to the first production of each target word. Second productions show no VOT difference across conditions, suggesting a strong effect of repetition in this experiment. Furthermore, as shown in Figure 3, the effects found are limited to cases when the target word begins with /p/ or /t/. This may be due to a ceiling effect associated with the /k/-initial targets used in the experiment, as /k/-initial words are known to have long base VOTs that participants may find it difficult to lengthen further.

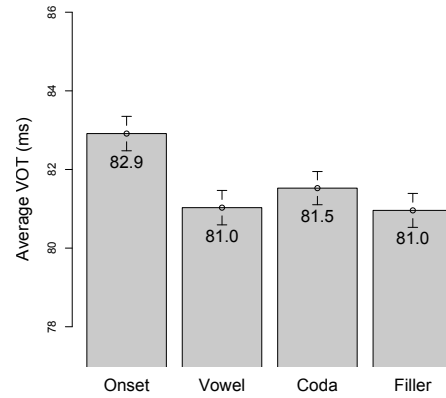


Figure 2: Experiment 2: Comparison of mean VOT across experimental conditions.

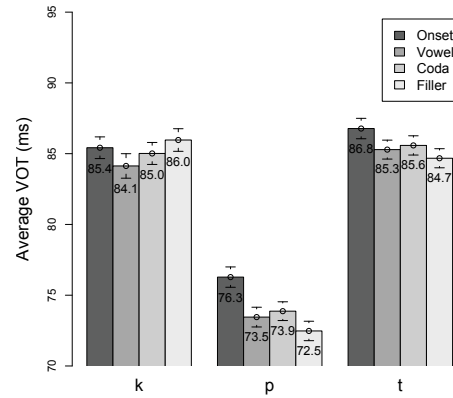


Figure 3: Experiment 2: VOT broken down by target onset phoneme and condition.

## Experiment 3: Featural Specificity of Online VOT Enhancement

The goal of this follow-up experiment, consisting of two subexperiments, was to determine if online VOT enhancement involves an even lower level of sublexical processing. In particular, we tested to see if only certain kinds of onset neighbors can induce VOT enhancement. In the first subexperiment, we looked for an enhancement effect in the context of place-differing neighbors. Different neighbor types were matched for frequency (pairwise paired t-test,  $p > 0.8$ ).

Table 5: Table of conditions for Experiment 3A.

Target	Voice	Place	Filler
CAP	GAP	TAP	DOLL

Participants (N=22) produced each of 33 target words

twice in one of the three conditions. Results are shown in Figures 4 and 5 and Table 6. There is a significant VOT enhancement effect of place neighbors, and the effect is consistent across /p/, /t/, and /k/-initial targets.

Table 6: Experiment 3A: Statistical results.

Condition	Coeff	SE	<i>t</i>	<i>p</i>
Voice Neighbor	2.1361	0.9329	2.290	< 0.0222*
Place Neighbor	1.8506	0.9333	1.983	< 0.0476*

It is interesting that the VOT of /p/ lengthens in the context of /k/ and /t/, given that /k/ and /t/ tend to have longer average VOT than /p/, and thus VOT lengthening might make /p/ initial words more similar to their competitors. However, aspiration also contains spectral cues for place of articulation (e.g., labial /p/ vs. coronal /t/ or dorsal /k/; (Suchato & Punyabukkana, 2005)), and lengthening VOT may strengthen these cues.

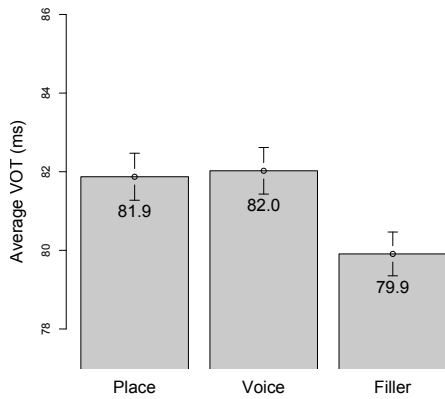


Figure 4: Experiment 3A: Comparison of mean VOT across experimental conditions.

In the second subexperiment we looked for an effect of neighbors differing in the manner of the onset. We attempted to choose neighbors that differed minimally from the targets with respect to manner, but were constrained by the phoneme inventory of English. The /p/-initial targets were paired with /f/-initial neighbors, which differ in manner and a minor place feature (labial vs. labiodental); /t/-initial targets were paired with /s/-initial neighbors, which differ in manner and stridency; and /k/-initial neighbors were paired with /h/-initial neighbors, which differ in both manner and place. Different neighbor types were matched for frequency (pairwise paired t-test,  $p > 0.8$ ).

Participants (N=22) produced each of 36 target words twice in one of the three counterbalanced conditions. Results are shown in Figures 6 and 7 and Table 8. There appears to be no overall significant effect of manner neighbors on VOT enhancement. However, the breakdown of the results

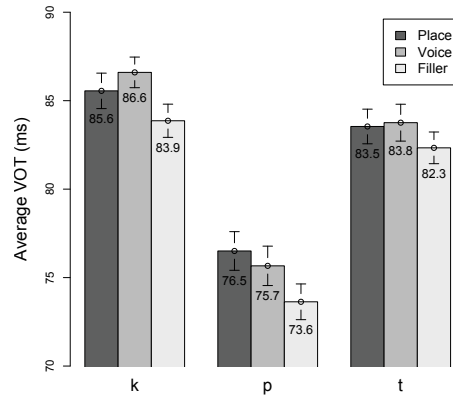


Figure 5: Experiment 3A: VOT broken down by target onset phoneme and condition.

Table 7: Table of conditions for Experiment 3B.

Target	Voice	Manner	Filler
PUN	BUN	FUN	DOLL
KILT	GUILT	HILT	DOLL
TEEM	DEEM	SEEM	DOLL

by target onset (Figure 7) indicates that there is an enhancement effect for /p/ onsets in the context of /f/ initial neighbors ( $p < 0.0225$ ). Since /p/ is likely more similar to /f/ than /k/ is to /h/ (differing in a major place feature) or /t/ is to /s/ (differing in stridency)<sup>1</sup>, it may be that online VOT enhancement may only occur in the context of neighbors that are sufficiently similar to the target word — about one major phonological feature away.<sup>2</sup>

Table 8: Experiment 3B: Statistical results.

Condition	Coeff	SE	<i>t</i>	<i>p</i>
Voice Neighbor	3.2236	0.9293	3.469	< 0.0005
Manner Neighbor	1.4489	0.9293	1.559	< 0.1192

## Explaining Online Variation as Listener-Orientation: Modeling Speech Perception

It has been hypothesized that language is designed to facilitate effective communication between speakers and listeners

<sup>1</sup> Although /p/ and /f/ tend to pattern together as a natural class more often than /k/ and /h/ or /t/ and /s/, the effects found might not be due to their apparent similarity; instead, they may be a property of /p/-initial targets. It is difficult to disentangle this question using English stimuli, since another stop/fricative pair as similar as /p/ and /f/ does not exist in the phoneme inventory.

<sup>2</sup> All of the place neighbors in experiment 3A differ from the target in just one place feature.



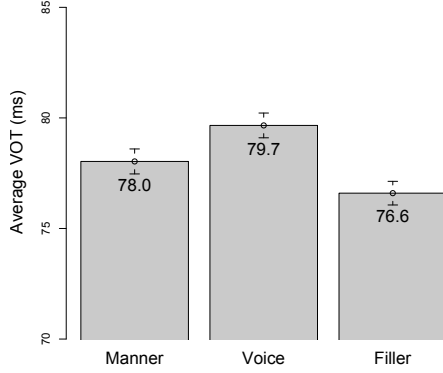


Figure 6: Experiment 3B: Comparison of mean VOT across experimental conditions.

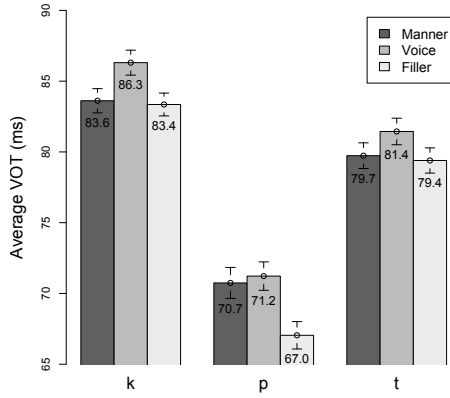


Figure 7: Experiment 3B: VOT broken down by target onset phoneme and condition.

(e.g., Lindblom 1990, Genzel & Charniak 2002, Levy 2006, Jaeger 2010, Frank 2008). This suggests that speakers may alter their speech in an effort to aid the recognition accuracy and speed of listeners.

This section presents a preliminary computational model of speech perception intended to help explain the specificity of online hyperarticulatory effects by allowing a comparison of the efficacy of different speech modifications in aiding listeners. The model is an extension of Norris’s Shortlist B (Norris & McQueen, 2008). It assumes that word recognition is a Bayesian process (Norris & McQueen, 2008; Feldman, Morgan, & Griffiths, 2009) that modifies a posterior distribution over possible input words as an acoustic signal unfolds. The word with maximum posterior probability after a certain amount of input is recognized. The ratio of the recognized word’s probability to that of its competitors determines how robust the match is — how likely it is to remain error-free

given noisier input. It is this ratio of posterior probabilities that expresses the concept of competition between alternatives in the model.

The posterior probability of each possible input word ( $W_t$ ) is equivalent to the likelihood that the word generated the signal seen so far ( $S$ ), multiplied by the prior probability of the word, divided by the total probability of the signal being generated by any word:

$$P(W_t|S) = \frac{P(S|W_t)P(W_t)}{\sum_i P(S|W_i)P(W_i)}$$

The likelihood function  $P(S|W_t)$  is equal to the likelihood that some prefix of the sequence of phonemes that make up the word ( $PP_t$ ) generated the signal seen so far:

$$P(S|W_t) = \sum_i P(S|PP_{ti})$$

Unlike Shortlist B, the present model does not assume a fixed amount of time per phoneme (see also Scharenborg, 2009).  $P(S|PP_t)$  can be broken down into a sum over possible segmentations ( $SS$ ) of the signal into phonemes.

$$P(S|PP_t) = \sum_i P(SS_i)$$

The likelihood that a particular phoneme generated a portion of the signal in a segmentation is a function specified for the model, and is intended to be empirically realistic (e.g., the likelihood that a voiced phoneme like /g/ generated a large amount of aspiration is low).

As an example, the model can be used to simulate the results found in Experiment 2. It receives input incrementally in 5ms frames, each containing one feature: C for closure, A for aspiration, and V for vowel (e.g. *gap* would be represented as [CAA...C]). In addition, the model only needs to distinguish between a target word and its on-screen competitor, resulting in a collapsed prior distribution over words (e.g.  $P(\text{cap}) = P(\text{gap}) = 0.5$ ). Simulation using this model indicate that when the target word is *cap* and it is pronounced in the context of *gap*, the ratio of posterior probabilities monotonically improves in its favor as VOT increases.

Overall, the model has certain formal properties that make it suitable for explaining the experimental results presented. First, the posterior odds in favor of a target word can’t be improved by hyperarticulating those parts of the word that are identical to its competitors. Doing this would equally increase the likelihood that the signal was generated by the target and its competitor. Second, the improvement in posterior odds gained by hyperarticulating the differing parts of the target is minimal if the target and its competitor are not sufficiently similar to each other, since in that case the likelihood that the competitor generated the signal,  $P(\text{signal}|\text{competitor})$ , remains near zero throughout the recognition process, and consequently so does the competitor’s posterior probability. In other words, if the target and competitor are different enough it would take a signal that is



very unlikely to have been generated by the target for it to be even slightly likely to have been generated by the competitor. Together, these two properties predict the specificity of online effects found experimentally. Speakers only seem to hyperarticulate when there is sufficient utility gained from the extra effort (Lindblom, 1990).

## Conclusions and Future Research

In summary, the experiments presented here indicate that offline and online effects of phonetic enhancement may *not* share the same underlying mechanisms. Not all offline effects appear to have online analogues, as evidenced by the apparent lack of a significant online vowel space expansion effect. In addition, online effects appear to be more *specific* than offline effects. Online enhancement can be caused only by neighbors in the speech context that are minimally different from the target word (differing by approximately one phonological feature). These findings are compatible with a system in which speakers attempt to aid listener comprehension, with a Bayesian model of word recognition indicating which speech changes are helpful and which aren't.

However, the latter finding opens the possibility that vowels may indeed be subject to online hyperarticulation in principle, but that Experiment 1 did not include competitors that were similar enough to induce this effect. In particular, the vowel neighbors used in the experiment were not controlled to be minimally different from the vowels in the target words (in terms of backness and height features). Experiments 2 and 3 suggest that minimal difference is essential for inducing online effects, and future experiments will explore whether online vowel enhancement can be induced by minimally-different vowel neighbors.

## References

- Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1), 31-56.
- Cohen-Priva, U., & Jurafsky, D. (2008, April). Phone information content influences phone duration. In *Conference on prosody and language processing*.
- Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (p. 9-39). Mouton, New York.
- Feldman, N. H., Morgan, J. L., & Griffiths, T. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4), 752-782.
- Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *CogSci* (Vol. 30, p. 939-944).
- Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the association for computational linguistics (ACL)* (p. 199-206). Philadelphia.
- Goldinger, S., & Summers, W. V. (1989). Lexical neighborhoods in speech production: A first report. In *Research on Speech Perception Progress Report* (p. 331-342). Bloomington.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61, 23-62.
- Lau, P. (2008). *The Lombard Effect as a communicative phenomenon* (Tech. Rep.). UC Berkeley.
- Levy, R., & Jaeger, T. F. (2006). Speakers optimize information density through syntactic reduction. In *NIPS* (Vol. 19, p. 849-856).
- Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W. Hardcastle & A. Maschall (Eds.), *Speech Production and Speech Modeling* (p. 403-439). Kluwer Academic Publishers.
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19, 1-36.
- Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 576-585.
- Munson, B., & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research*, 47, 1048-1058.
- Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115(2), 357-395.
- Scarborough, R. A. (2004). *Coarticulation structure and the lexicon*. Unpublished doctoral dissertation, UCLA.
- Scharenborg, O. (2009). Using durational cues in a computational model of spoken-word recognition. In *INTER-SPEECH* (p. 1675-1678).
- Son, R. van, & Pols, L. C. (2003). How efficient is speech? In *ICPhS* (Vol. 25, p. 171-184).
- Suchato, A., & Punyabukkana, P. (2005). Factors in classification of stop consonant place of articulation. In *INTER-SPEECH* (p. 2969-2972).
- Wright, R. (2003). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI* (p. 75-87). Cambridge University Press.
- Zhao, Y., & Jurafsky, D. (2009). The effect of lexical frequency and Lombard Reflex on tone hyperarticulation. *Journal of Phonetics*, 37(2), 231-247.

## Running it through the body

David Kirsh (kirsh@ucsd.edu)

Dept of Cognitive Science  
University of California, San Diego

### Abstract

Video data from three large captures of choreographic dance making was analyzed to determine if there is a difference between participant knowledge – the knowledge an agent acquires by being the cause of an action – and observer knowledge – the knowledge an observer acquires through close attention to someone else’s performance. The idea that there might be no difference has been challenged by recent findings about the action observation network and tacitly challenged by certain tenets in enactive perception. We explored why a choreographer ‘riff’s’ when appropriating and evaluating the movements of his dancers. By recruiting his body to help him cognize he is able to understand the possibilities of movement better than observation. He acquires participant knowledge.

**Keywords:** embodied cognition; thinking; choreography.

There is a tacit assumption in situated cognition that performing an action yields a type of knowledge – *participant knowledge* – that is irreducible to knowledge acquired by observing someone else performing the same action – *observer knowledge*. A violinist acquires more knowledge by playing a piece than by listening to someone else. He is embedded more profoundly in the situation. A dancer is able to understand something qualitatively different about a dance phrase by dancing it. Just watching leaves something out.

I present data, from a major study on dance creation that supports this intuitive claim: in dance, using one’s body to explore a dance idea is a better way to understand the idea than watching someone else explore it. This may seem so obvious as to need no justification. But, there is extensive neurophysiological evidence of a close link between action observation and action execution [Viviani, 02; Wilson & Knoblich, 05]. Whenever we observe another person’s goal directed actions we re-enact or mimic that actor’s movements by covertly behaving as if we are performing the action ourselves. [Sebanz, N., & Shiffrar, 07]. Activating the motor resonance system may be comparable to actual performance [Rizzolati & Sinigaglia, 07; Agnew et al, 07; Aglioti et al.].

The idea of a covert action was introduced by Jeannerod [94] to describe the subliminal activation of the motor system by “[intended] actions that will eventually be executed, [and] also [by] imagining actions, recognizing tools, *learning by observation*, or even understanding the behavior of other people”. [Jeannerod 01, emphasis added] Covert action involves motor planning, just as overt action does; and perhaps it involves some level of motor preparation, though probably less than overt action. The real difference is that covert action does not activate muscular control. Yet, the activity in the covert system is nonetheless so strong that

even just watching an action can be as powerful a learning experience as performing the action oneself. [Cross et al., 09]

If it is true that the motor system is activated through observation as powerfully as suggested it is worth asking whether actual execution is required for action understanding and deep processing of action. Maybe observation is indeed enough. What extra does one get by adding overt movement over and beyond what one gets from mentally simulating the same movement covertly?

Exploring this ‘extra’, at least in the dance context, is the topic of this essay. I begin by clarifying what is meant by participant knowledge. I then explain the methodology we used for studying participant knowledge. That section is followed by a presentation of empirical results – observed regularities about when and how a choreographer *runs dance phrases through his body* in order to understand a phrase better – to deep process it. The paper closes with a discussion of the central ideas of bodily mediation, and enactive thinking. Jointly these last two ideas form the basis for an explanation of why authoring an action can lead to a more profound understanding than observing it.

### What is participant knowledge?

If there is something special about participant knowledge, then on those occasions when a violinist – say, Yitzhak Perlman – physically plays a musical piece, he will connect with the piece in a way that is special because he is the one playing. The same holds true for dancers: they will connect in a special way when they take to the floor and perform.

As intuitive as this is it runs contrary to a theory even more general than motor resonance: the theory of internalization. According to Vygotsky [78], whenever Perlman is listening, his internalization of what it is like to play mediates his listening. He will engage violin music as if playing it because, as a result of constant practice, he has internalized the performance mode of interacting with music so deeply that he doesn’t need a physical instrument to participate in music making. He has an inner violin and plays it when hearing others play.

An analogy is private talk. It starts as real talk outside, it is shaped socially by norms and practice, it is internalized and then it is available inside [Vygotsky 86]. After that, we can keep our mouth shut and think entirely in our heads.

In this sense, acquiring mastery of an instrument leads to the internalization of a principle of organization – a way of perceiving the world that comes from having mastered externally a highly structured form of interaction. It is artifact-mediated cognition, but without the artifact. If you play enough, you eventually can simulate playing without an

instrument in your hands. Thus, Perlman understands the meaning structure of a musical piece because of acquired knowledge of music, because he has internalized the way a violinist would approach the music, and because, while listening or watching, his resonant system simulates playing it. [Cisek & Kalaska 04] His vast experience and prior practice grounds his perception sufficiently for him to realize the musical possibilities at each moment.

Despite the allure of motor neuron theories, and despite the importance of recognizing that humans internalize principles of organization I think intuition is against Vygotsky and motor resonance here. Neither dancer nor violinist can mediate their encounter of dance and music to quite the same depth with and without their instrument. Physical performance matters. Whatever Perlman may know about a piece through watching and listening, and it is considerable, he cannot know all that he would feel or register were he playing the piece personally. The difference – the remainder – lies in what it means to be situated and to be an agent; it depends on being the prime cause.

Part of the ‘extra’ that using a physical instrument provides an agent is a consequence of how working with a physical instrument causally shapes what a performing agent understands about the possibilities of a situation. This extra includes a phenomenological sense of freedom and responsibility. By being the person who is creating the music, a violinist has a responsibility to succeed, and during his performance he or she has access to a set of performance specific concepts and experiences unavailable to an observer, even a violinist observer. These concepts are ad hoc [Barsalou 10], situated [Greeno 89, Kirsh 09] and embodied [Barsalou 08], and they permit the agent to project a future that is conceptually and experientially richer than the future projected by an observer. They provide the performer with a framing of decisions at each point: how long to hold a note, how to attack it, its mood and emotionality. These differences are not reducible to the specifics of what it is like to play the violin – to move the bow on a string, to hold the violin under the chin. Those are practical elements that might bear on the moment-by-moment musical decisions that must be made, but many of these mechanical aspects of working with a violin are irrelevant to the performer’s conception of the musicality of the piece. The extra elements of knowledge conferred by participation concern the music itself. Agency is a special mode of making contact with that. The result is that in probing music with his violin, Perlman is able to discover something about the music he himself would miss were he just to listen. He needs the violinistic encounter with the musical composition to activate some of those concepts and sensory experiences. At least that’s the story.

This is a complicated and remarkably strong thesis, one that I believe lurks at the soul of the frameworks of situated and embodied cognition. To my knowledge it has not been closely considered.

## Method

The data for this research comes from two extensive case studies in which we captured the making of new choreographic work created by the celebrated choreographer Wayne McGregor and his contemporary dance company Random Dance. In the course of three periods – the first two occurring in two three-week periods (winter and fall of 2009), the third in a six-week period (fall of 2010) – all the face-to-face encounters between choreographer and dancers, (about 5 hrs/day) as well as all practice sessions involving the dancers and the associate choreographer Odette Hughes, were recorded by six high definition video-cameras. Over thirty 60-90 minute interviews were recorded between the choreographer and author and also with the dancers individually or in small groups. All notebooks, brain storming stimuli and real time notes were collected. Several experiments on marking, mental simulation, imagery ability and movement memory were carried out. Each case study yielded about 20TB of video. All videos had to be transcoded, collated and organized – altogether a massive process that required the help of several teams of students too numerous to thank individually. [Kirsh 10]

Once all materials were organized, work sessions were identified and cursorily annotated. Specific phenomena were then identified for intensive examination. We discuss here our observations and analyses of a process we call ‘riffing’ – an activity the choreographer regularly performs in which he tries out ideas by dancing them himself.

The detailed coding of riffing was performed by three college seniors long involved in this project. Each coder worked on separate days in the corpus and intercoder reliability was measured on 10% of the material done in common yielding .77 using Krippendorff’s alpha measure.

**Riffing off-of-others, the phenomenon studied.** When the dancers we studied are working on an assigned choreographic task, or when working on a duet, trio or quartet, we regularly observed that the choreographer, WM, would observe them closely, and then, if the dancers were to do something interesting or untoward, he would try out their movement himself. He would physically sketch the movement, appropriate what he likes, and then work on the phrase himself, substantially modifying it before sharing it. We call this activity *riffing off-of-others*. Superficially, it is the equivalent of playing a musical piece himself.

When asked in interview why he riffs off-of-others WM said he does it “to feel the moves”, and also “to redo them with [his] own signature”, “to ensure that they are authentic” or to test if they are “consistent with [his] artistic style and the integrity of the piece as a whole”. Executing the movement also lets him see its possibilities, its ability to “support invention”, its potential fertility. We cannot confirm these views on the basis of videographic observation because much of the interest of a movement for WM, he reports, lies precisely in its physical or dynamic novelty, something he recognizes in the movement that is quite different from previous movements he has worked on, owing

perhaps to weight, balance, force, resistance or other attributes that are kinesthetically meaningful but almost impossible to discern visually. This is a key point.

## Empirical Results

Because we have no access to our subject's motor encoding through imaging or otherwise, our empirical study (the non-interview part) involved reviewing nearly a thousand episodes of rifting and measuring about 15% of them. Our goal was to observe when the choreographer riffs, how faithful his riffs are to the target movement he is sketching, how he modifies the movements, and then what he does with these modified movements.

We found that rifting off-of-others follows a common pattern: 1) the choreographer watches a dancer or small group develop a movement idea; 2) he personally sketches or 'marks' their movement, though he also adds or subtracts from their idea by prepending, appending or deleting components in the first pass; 3) he runs through (i.e. he riffs) several more times, each occasion adding, subtracting, or altering more of the phrase as he initially sketched and modified it; 4) he then works with the dancers to share the new idea. The process is very collaborative, though not quite a dialogue, for the dancers do not attend to what WM is doing when he is rifting – they are busy dancing themselves – and WM himself does not seem much concerned to get the dancers' movement exactly right. He does not stop, look again, practice. Instead, he watches, physically sketches and remakes his own versions, all in relatively high speed.

This kind of physical sketching and rifting seems a way for him to pick up ideas he did not originate, and then play with them. He runs someone else's movement through his own body because it is not enough for him to see what others are doing; he needs to appropriate the full structure of the movement to explore how it might be developed, continued. In short, rifting is a way he thinks with his body. He wants 'agentive' knowledge.

Before we look at the data supporting this view, it is worth commenting on how this practice departs from the case where a violinist plays a piece to understand it rather than listening to another violinist playing it. In the musical case, both soloist and listener share a common musical specification: the score. Playing is a better way of appropriating the score. In dance, and most especially in creative dance, there is no prior score and no real-time capture used to 'freeze' a movement. WM never uses a score (or video, though the dancers sometimes do later in the process); and the company makes no effort to transcribe their movements in a dance notation, such as Laban. Understanding must happen on the dance floor and in real time. More importantly, the kind of understanding the choreographer is after is dynamic. He needs to deep process the movement to see its potential. But this does not always mean recreating it exactly. In interviews the choreographer says he wants to appropriate the movements his dancers make. The curious thing is why he does not feel compelled to duplicate their movements more precisely.

**Data.** The data shown in Table 1 are derived from studying the first Make session of the first day of creating a dance made in 2009. Ten sequences of rifting off-of-others were found in this one session. By studying them frame-by-frame we were able to measure the time in seconds of the mean duration of the referent movements – the dancer movements – that WM chose to riff off of, and the timing of his subsequent activity.

Rifting Off-of-Others (measurements in seconds)							
	Referent Move by dancer	Gap 1	Riff 1	Gap 2	Riff 2	Gap 3	Riff 3
Mean Sketch Duration	2.7 secs		2.8s		1.7s		2.4s
Mean gap		0.7s		20.2s		28.1s	
Mean Fidelity	100%		50%		29%		25%
Mean WM-added content			4.1s		3.1s		2.8s
Total			6.9s		4.8s		5.2s

**Table 1. Rifting off-of-others**

Looking at the columns, Gap 1 measures the time between the moment when a dancer performs a movement and the moment WM sketches it. Gaps 2 and 3 measure the time between subsequent riffs. WM-added content is the material he adds that is not found in the referent movement or in his sketch of that movement. It was surprisingly easy to recognize the referent material even though WM's sketch was not perfectly similar to the referent. Our interest was to determine how much of WM's movement was derivative, based directly on a referent, and how much of WM's movement was his own authored content.

In a typical riff, WM observes a referent move he likes and watches it a few times before sketching it in real time, immediately after the next time he sees it. As can be seen in table 1 this delay between seeing the referent and sketching it (after having seen it at least once before) is less than a second. After his first riff he then waits about twenty seconds, either watching other dancers or just standing pensively off stage. He then riffs again, which we call Riff 2; there is a gap again, on average 28 seconds, and then he makes a final riff, Riff 3.

Looking at the values in table 1 we see that on average his first riff is only 50% faithful to the referent. To determine fidelity we graded the quality of a riff along the dimensions of technicality, memory, timing and dynamics, the same dimensions we used in our marking experiment. (See this issue [Kirsh et al] and Kirsh [forthcoming]). Each dimension has four ordinal values: A, B, C, D. Overall fidelity was defined as the averaged score on all three dimensions. To calculate the mean and then return a letter grade for fidelity we converted letters to a percentage in a linear fashion (A=100% faithful, B=75%, ...).

Given his skill at real-time sketching WM's low fidelity suggests that his first riff is more selective than realistic sketching. It may mean that he is interested in appropriating only certain aspects of what another dancer is doing. In this

first riff, we found, further, that on average he adds more than twice as much of his own content to the material he appropriates. After a gap of 20 secs he seems to tighten up the movement by reducing the duration of both derived and WM-made content. Following another delay of 28 secs he increases the duration of the movement. He now has a phrase that contains only 25% of the original 2.7 secs movement he took, making that sketching look less like appropriation and more like inspiration; his own contribution is about the same length. It is this new movement, totaling on average 5.2 secs, that he eventually shows to the dancers in this or a later session.

What does this tell us? First, Table 1 shows that we were wrong in a conjecture we had made. We had assumed that riffs would unfold as a quick sequence of increasingly faithful sketches. WM would fully appropriate the referent phrase before his modifications and divergent sketching. This is typical of the way dancers sketch, when mastering the phrases of others. But it was not the case for WM. On average, WM will riff once, with only moderate fidelity to the referent, and then begin to truncate, add or modify the phrase. Even in his first riff he usually adds more of his own content than the phrase he appropriates.

Second, it suggests that his concern is with only certain aspects of a movement. The obvious analogy is with sketching on paper. An artist inspired by Soutine may sketch one of Soutine's paintings or drawings, hoping for ideas. But the sketch, much like WM's, is rarely faithful to the original and the creativity seems to lie in how the artist departs from the original.

Let us look at the process of choreographic sketching more closely to see if it may illuminate the nature of how physical movement acts as a mediating structure for thought.

**Sketching in Dance** is the process of copying in real-time the movements of another dancer – the referent. The referent dances, produces a target phrase, and the sketcher does her best to duplicate the target phrase herself. Inevitably there are stylistic differences, and most of all, differences owing to variances in dancers' height, weight, body form, strength, and gender. Sketching in this sense is an early shot at mastering a movement the way the referent does it. It is not to be confused with an artist's sketching, or a musician's sketching where often there is no referent – no touchstone of correctness.

If the sketching process in dance follows a normal pattern of structural approximation then the first sketch will be coarse, capturing essential elements of the referent such as emotion, general shape, gross dynamics, key positions, and occasionally sporadic details that they notice or like. What then follows is a trajectory of practice, a sequence of improvements and modifications to the original sketch to improve verisimilitude. The process is remarkably fast for professionals and a phenomenon worthy of study in its own right. After a minute or two, a talented dancer will stop watching the referent and practice on her own.

The majority of sketching we observed among the Random Dancers was real-time sketching. Each time the

choreographer makes a new phrase on one person (or a small group) the rest of the company is expected to learn the movements too. This is expeditious because when crafting phrases for duets and trios it is easiest to 'make' on a referent duo or trio on the assumption that the others, who typically were already doubled or tripled up (usually at WM'S explicit direction) would learn the phrase in their own duo or trio. In this dance company, moreover, the choreographer would sometimes swap dancers, putting a different dancer in the target role in the final piece, so dancers were expected to learn virtually all phrases.

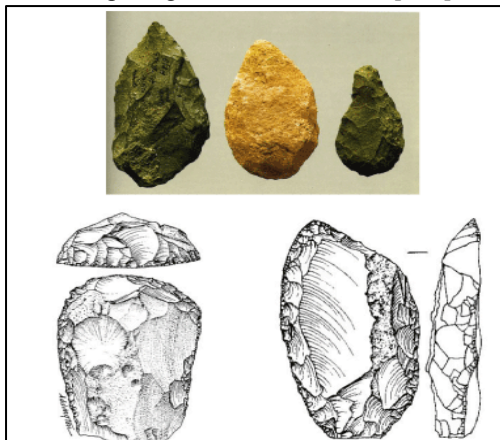
Sketching in dance is a topic of interest because of what it can teach us about how the body is used to manage attention. In my view, a major function of a mediating artifact is to regulate attention and activate priming. A hammer helps us drive nails into wood. It is a purely physical, non-cognitive artifact. [Norman, 91] But when it is in our hand it also coordinates a complex pattern of movement and attention shifts – sensori-motor adaptations and interactions. Some of these are below conscious threshold and involuntary (e.g. grip in mid-swing). Others are fully conscious, but often they too are almost involuntary. For instance, when a hammer glances off a nail imagine how little control we have in seeing where it lands. Our eyes are drawn to it. It is this pattern of action and attention that is hard to duplicate without a physical thing driving it. The physical artifact mediates our knowledge of hammering. It plays an essential role in organizing our hammer'ish interaction with things.

To return to our violinist, we would expect that Perlman can partially simulate his pattern of attention to a musical piece when not physically playing his Stradivarius. He has his inner violin, with all its interaction-organizing principles. Numerous behavioral and imaging studies suggest that when humans mentally rehearse a familiar action they execute some of the same neural operations used during overt motor performance. [Jeannerod 01] When listening, Perlman would have no problem imagining himself playing. And he would engage similar neural and cognitive operations. [Munzart 2009]. But there will be many involuntary, fast paced adjustments to playing for real that would demand his attention physically that simply do not arise during simulated playing, where there can be no direct sensory feedback from the environment. In short, his simulation of playing is at a lower resolution than actual playing.

The special role that a sketch, as mediating structure, may play in dance cognition can be appreciated by analogy with sketching in paleoarchaeology, where Lithic sketching is used to solve a hard problem: distinguishing human made from natural stones in lithic eras.

In figure 1 there is a picture of three stones, any one of them may be from the Paleolithic era, and below them are some drawings made by an expert sketcher following the principles of lithic illustration. According to Addington [86] and Lopes [06] the best method to tell whether a given Paleolithic stone is a cutting tool is to sketch the stone. Not just any form of sketching will do. There is an expert mode of sketching for Paleolithic objects codified in a set of

principles of ‘lithic illustration’. Good archaeological illustrators will draw a lithic stone to reveal the physical ‘problematic’ the tool cutter faced. They will show the “scale; the pattern, sequence, direction, and force of blows to the stone; the bulb and platform of percussion; areas of retouch, snapping, and truncation; areas of grinding, battering, or abrasion; fractures caused by heating; the effects of materials; and pitting and sickle sheen.” [ibid] Potentially



**Figure 1.** Lithic sketches are drawings of stones made according to the strict principles of lithic illustration. The stones in the top picture are either human made – artifactual – or they are nature made and not shaped by humans.

confusing features of the stone such as embedded fossils, variegated coloration, patina, seams, banding, and crystallization are left out of the drawing.

The implication is that expert illustrators, when practicing their craft, are forced to scrutinize stones in a special way. They coordinate hand and eye to interactively probe the stone to reveal knapping related features. The need to draw certain lines drives perceptual inquiry. Attention must be managed, and arguably, without the need to sketch, without the presence of an external structure that the illustrator is creating, attention would not be managed adequately. Of course, this is an exaggeration. Illustrators have professional vision [Goodwin 93] and so can see elements of what they *would* draw without actually drawing. But in drawing, the process of making lines and ensuring they are spaced revealingly, is itself a process that simulates knapping. Using a pencil to draw a curve is physically related to using a knapping stone to flake a chip off a stone. It physically simulates knapping. So, the drawing process can help the illustrator walk through the history of the axehead’s making. The drawing is an external representation, but the process of making this representation is a powerful method for structuring attention. It helps the illustrator figure out what an artifact is by studying ‘the details of its making’ (ibid).

The analogy to riffing should be clear. By riffing, the choreographer is forced to direct his attention to the central aspects of a movement. By running the phrase through his own body WM gets to feel its dynamics, balance, gravity, internal shape. Not as seen in a mirror, but as felt through

movement – he experiences ‘the details of its making’ including the many body decisions the dancer made.

We turn now to what riffing teaches us about the power of using the body to help think about dance; how being the agent of movement offers privileged knowledge of dance.

## Discussion

It is no surprise that dancers physically sketch, explore or probe movement ideas by using their bodies. The question at issue is why, when our choreographer sees an interesting movement performed by one of his dancers, he bothers to mimic it? Given his capacity as a super-expert he ought to be able to attend to enough aspects of the movement by observation alone, or perhaps by inner simulation, making external movement unnecessary. One would suppose that he can think through the possibilities of a movement well enough in his imagination. Observation of the movement ought to give him an adequately precise ‘perceptual blueprint’ [Hodges et al, 07] that he can then imaginatively work with.

Support for the idea of imaginative simulation being sufficient can be found in the idea of enactive perception. [Noe 05] On an enactive account of visual perception, an observer should see the counterfactual futures in the present. He should phenomenologically experience possible ways a phrase may be continued. In this case that would mean anticipating a dancer’s possible movements just before they were made, then saccading, moving the eyes, head, trunk and attending closely to see which of the movements that might have been made do in fact occur, and then revising perceptual expectations appropriately. The enactive process happens during perception, but it grounds an understanding of the movement process that encompasses more than what was literally seen and supports imaginative replay and exploration. [Thomas 99]. It supports projection [Kirsh, 12].

A second reason overt action might be superfluous is that humans have the capacity to improve motor performance by observation alone, without concurrent physical practice, [Torriero et al 08, ibid]. The fact that there are older choreographers (notably Merce Cunningham) who continued making noteworthy pieces after drastically reducing their physical exploration [Nolan, 12] offers further support that physical practice is not necessary for grasping the choreographic potential of a movement; observation and mental simulation may be enough.

For our forty-year old choreographer that is *not* what we found. He regularly runs possible steps and phrases through his own body, and he seems to rely on that process as part of his choreographic practice.

I suggest we view Riffing as a type of enactive thinking. It is not just a way of better activating what vision can supply. It constitutes a more interactive probing. Interaction requires more than simply changing one’s eye, head, trunk and body position to observe; it involves changing the object of inquiry. It is an intervention.

Thus in reply to the question how can thought be partly constituted by bodily movement I have a few answers.



First, since bodily movement is by definition part of the action-perception system it can be harnessed reliably as part of a simulation process as well as ‘mental’ simulation inside cortex. If internal simulation is good enough to ground thought, then why not regard the act of materializing the target process an even better source of grounded thought? Moreover, if nature plays a role in simulation the progression of states will be more detailed and reliable than mentally projecting, imagining, or simulating the next state, which is more error prone. So dancing a phrase ought to be a better way of grasping the possibilities of a phrase than simulating it. And perceiving possibilities is a lot of what understanding is. This leads to the second reply.

Badets et al., [06], showed that physical practice is better than mere observation for learning new movements. This may not always be the case with simple and even moderately complex movements [Cross et al, 09]. Presumably in those cases where physical practice surpasses prolonged action observation something extra is getting in. What is it? In Badets [op cit] the extra is detailed behavioral expertise and its neural basis. But with respect to thought, and not just skilled movement, the extra that comes from overt bodily involvement is an enhanced conceptualization of what the phrase is, a better grasp of what makes a performance *true to the phrase*. In simple dance phrases there is little to grasp or deeply conceptualize. But for complex, choreographically rich phrases this can be a real issue. It means being able to judge when two dancers with different genders, backgrounds and bodies have mastered the phrase ‘correctly’.

Riffing falls into this enhanced conceptualization category because when WM executes a phrase he is making decisions at each moment; he is ‘thinking’ about it. He reports looking for possible lines of development, for novelty, for discovering a point in movement space that is uncharted. In lithic illustration you also feel the decisions: why here, and not there? In dance, part of conceptualizing possibilities means understanding key dynamics like the effect of gravity, balance, force, and bodily tension. These arise through physical interaction and are highly sensitive to momentary physical factors. Observation alone cannot expose these elements. Without agency, intervention and physical engagement, human knowledge is different. Angels can never understand dance as humans can.

**Acknowledgements:** Richard Caballero, David Mazur, Gina Bello, Dafne Muntanyola, Cogs 160 class, WM | Random Dance. Funding from NSF: IIS-1002736 gratefully acknowledged.

## References

- Addington, L. R. (1986) *Lithic Illustration: Drawing Flaked Stone Artifacts for Publication*. Univ of Chicago Press
- Aglioti et al. Aglioti, S. M., Cesari, P., Romani, M., & Urgesi, C. (2008). Action anticipation and motor resonance in elite basketball players. *Nature Neuroscience*, 11(9), 1109–1116.
- Agnew, ZK and Bhakoo, KK and Puri, BK (2007). The human mirror system: A theory of mind reading. *Brain Research Reviews*, 54 (2) 286 - 293
- Badets, A., Blandin Y., Shea C.H. (2006) Intention in motor learning through observation. *Q J Exp Psychol*. 59:377-386.
- Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59, 617–64
- Barsalou, L.W. Ad hoc categories. (2010). In P.C. Hogan (Ed.), *The Cambridge encyclopedia of the language sciences* (pp. 87-88). New York: Cambridge Univ Press.
- Cisek P, Kalaska J F. (2004). Neural correlates of mental rehearsal in dorsal premotor cortex. *Nature*. 431:993--99
- Cross, Emily, et al. (2009) Sensitivity of the Action Observation Network to Physical and Observational Learning, *Cerebral Cortex*;19:315—326.
- Goodwin, Charles (1994) 'Professional Vision', *American Anthropologist* 96(3): 606-33
- Greeno, J. G. (1989). "A perspective on thinking". *American Psychologist* 44: 134-141
- Hodges, N.J., Williams, A.M., Hayes, S.J., Breslin, G. (2007) What is modelled during observational learning? *J Sports Sci* 25:531-545.
- Jeannerod, M. Neural simulation of action: a unifying mechanism for motor cognition. *Neuroimage* 14: 103–109
- Kirsh D., (2010). Thinking with the Body, in (eds) S. Ohlsson R. Catrambone, *Proc of the 32nd Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society. Pp 2864-2869.
- Kirsh, D. (2009). Problem Solving and Situated Cognition. In Philip Robbins & M. Aydede (eds.), *The Cambridge Handbook of Situated Cognition*. Cambridge Univ. Press.
- Kirsh D., (2012). When doing the Wrong Thing is Right. This issue.
- Lopes D: Drawing in the Social Sciences: Lithic Illustration. <http://www.interdisciplines.org/artcognition/papers/7>
- Munzert J, Lorey B, Zentgraf K (2009) Cognitive motor processes: the role of motor imagery in the study of motor representations. *Brain Res Rev* 60:306–326.
- Noë, A. (2005), *Action in Perception*. MIT Press.
- Nolan, C. (2012) *Leonardo Electronic Almanac*, vol:17, 2
- Norman, Donald A. (1991): Cognitive artifacts. In: Carroll, John M. (ed.). "Designing Interaction: Psychology at the Human-Computer Interface". Cambridge Univ Press pp.17-38
- Rizzolatti G., Sinigaglia, C., (2007). *Mirrors In The Brain: How Our Minds Share Actions And Emotions*. Oxford Univ. Press
- Rizzolatti G, Craighero L. (2004). The mirror-neuron system. *Annu Rev Neurosci*. 27:169-192
- Sebanz, N., & Shiffrar, M. (2007). Bodily bonds: Effects of Social Context on Ideomotor Movements. In Haggard, P. Rosetti, Y. & Kawato M. (eds), *Sensorimotor Foundations of Higher Cognition. Attention and Performance, XXII*. Oxford Univ Press.
- Thomas, N.J.T. (1999). Are Theories of Imagery Theories of Imagination? An Active Perception Approach to Conscious Mental Content. *Cognitive Science* 23. 207–245
- Torriero S, Oliveri M, Koch G, Caltagirone C, Petrosini L. (2007). The what and how of observational learning. *J Cogn Neurosci*. 19: 1656--1663
- Viviani, P. (2002) Motor competence in the perception of dynamic events: a tutorial. In Prinz, W. and Hommel, B. (eds), *Common Mechanisms in Perception and Action. Attention and Performance XIX*, pp. 406–442. OUP, New York.
- Vygotsky, L. (1930/1978) *Mind in Society*, Harvard University Press, Cambridge, MA.
- Vygotsky, L. S. (1986). *Thought and language* (A. Kozulin, Trans.). Cambridge, MA: The MIT Press
- Wilson, M. and Knoblich, G. (2005) The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, 131, 460–473



# A belief-updating model of adaptation and cue combination in syntactic comprehension

**Dave F. Kleinschmidt<sup>1</sup>, Alex B. Fine<sup>1</sup>, and T. Florian Jaeger<sup>1,2</sup>**

{dkleinschmidt, afin, fjeager} @ bcs.rochester.edu

<sup>1</sup>Department of Brain and Cognitive Sciences and <sup>2</sup>Department of Computer Science,  
University of Rochester, Rochester, NY, 14607 USA

## Abstract

We develop and evaluate a preliminary belief-updating model which links intermediate-term (i.e., over several days) syntactic adaptation to the joint statistics of syntactic structures and lexical cues to those structures. This model shows how subjects differentially depend on different cues to syntactic structure following changes in the reliability of those cues, as shown by Fine and Jaeger (2011). By relating syntactic adaptation and cue combination to rational inference under uncertainty, this work links learning and adaptation in sentence processing with adaptation in speech perception and non-linguistic domains.

**Keywords:** sentence processing, adaptation, Bayesian modeling, cue combination, rational analysis

## Introduction

Humans must maintain a stable representation of the environment despite the fact that available sensory input changes across time: for example, over the course of a day, we recognize and grasp objects in a variety of lighting conditions; we execute accurate motor commands despite changes in our own motor systems due to fatigue, over-cafeination, etc.; and during linguistic communication, we process rapidly unfolding acoustic information that varies from talker to talker.

Variability within each of these different modalities changes the correlation between cues—whether visual, haptic, or linguistic—and the things in the world we wish to make inferences about *based on* those cues. How do our brains make use of these cues in spite of variability in the environment? One possibility, suggested by research across a number of domains, is that humans deal with variability in the environment by *adapting to* changes in the statistical properties of the environment (for examples from vision, motor planning, and speech perception, see respectively: Blakemore & Campbell, 1969; Koerding, Tenenbaum, & Shadmehr, 2007; Norris, McQueen, & Cutler, 2003).

While most work on adaptation has been concentrated in perception, the question of whether adaptation is operative in higher level cognition has recently received more attention, particularly in language processing research. For instance, a number of researchers have shown that, when given sufficient experience with a structure initially judged to be ungrammatical, listeners come to subsequently comprehend (Luka & Barsalou, 2005), generalize, and even produce (Kaschak & Glenberg, 2004) that structure. Similarly, recent work has shown that we fine-tune our expectations about which syntactic structures are likely to occur in a given context based on recent experience (Thothathiri & Snedeker, 2008; Farmer, Fine, & Jaeger, 2011).

Thus, behavioral evidence seems to suggest that adaptation, qualitatively speaking, is a very general feature of perception and cognition. A question that arises from all of this previous work is whether adaptation observed across all of these domains can be modeled within a single framework. The goal of this paper is to take a step in this direction. In particular, we model adaptation in syntactic comprehension in terms of Bayesian belief update. Modeling syntactic adaptation in a Bayesian framework is appealing because the same basic computational approach has been successfully pursued in a variety of perceptual and motor domains (e.g. Koerding et al., 2007) and, more recently, in speech perception (Kleinschmidt & Jaeger, 2011; Sonderegger & Yu, 2010).

Moreover, Bayesian belief update is ideally suited to explicitly model the fact that syntactic comprehension involves the combination of multiple cues. This offers the advantage of suggesting a single computational framework for adaptation and cue combination, since Bayesian approaches to cue combination have been successful in a number of domains including visually-guided grasping, audio-visual cue combination, and the weighting of cues to phonetic category. In particular, Bayesian approaches to cue combination in perception have provided a formal means of capturing the fact that humans are able to weight multiple cues (e.g., multiple cues to object depth, such as shading and texture) according to how *reliable* those cues are. We return to the relationship between adaptation and cue combination in the discussion.

The goal of the current study is to ask whether a rational model of adaptation—implemented in the form of Bayesian belief update—can account for behavioral evidence for adaptation in the syntactic domain. Here we model behavioral data originally reported in Fine and Jaeger (2011), which concerns how subjects adjust their expectations about different syntactic structures conditioned on lexical information. Specifically, we exploit temporary syntactic ambiguities as a window onto syntactic expectations. In sentences such as (1), the syntactic assignment of the noun phrase *the judge* is temporarily ambiguous, since it can be parsed as either the subject of a sentence complement (SC) clause, as in (1a), or as the the direct object (DO) of *acknowledged* (as in 1b).

- (1) The lawyer acknowledged the judge ...  
                 disambiguation  
     a. ... {had been} unfair to the defendant.  
     b. ... in the black robe.

The sentence is disambiguated towards the latter reading at

*had been*. A great deal of previous work suggests that reading times at *had been* are a function of subjects' *expectations* about which syntactic structure is likely to occur, based on previous cues in the sentence, such as the verb, the combination of the verb and post-verbal noun phrase (e.g., Garnsey, Pearlmutter, Myers, & Lotocky, 1997), and whether or not the complementizer *that* occurs after the verb (e.g., *The lawyer acknowledged that the judge had been unfair to the defendant*). More recent work has explicitly quantified syntactic expectations in probabilistic terms (Hale, 2001; Levy, 2008). In other words, reading times at the disambiguating region (*had been*) provide information about subjects' subjective beliefs about the relative probability of the SC vs. the DO structures: If reading times are high, this indicates that subjects had assigned a relatively *low* probability to the SC structure; if reading times are low, then subjects had likely assigned a relatively *high* probability to the SC structure.

We model changes in reading times at the point of disambiguation as a consequence of syntactic expectation adaptation. Assuming that reading times in the disambiguating region in sentences such as (1a) reflect subjects' beliefs about the relative probabilities of different syntactic structures, we can interpret *changes* in reading times as changes in subjects' beliefs about the distribution of syntactic structures (at least in the context of the experiment, a point to which we return in the discussion). Syntactic adaptation, construed as the incremental adjustment of the subject's representation of a probability distribution over linguistic events, can therefore be naturally modeled in terms of Bayesian belief update.

In the following section, we briefly describe the behavioral data we set out to model. Next we present a Bayesian belief update model of this behavioral data, and assess the quality of the model's fit to the behavioral data. We conclude by summarizing the model's results and providing a discussion of the implications of this modeling work for our understanding of adaptation and cue combination.

## Methods and Summary of Behavioral Results

The data we use to test the hypothesis that syntactic adaptation can be understood in terms of incremental Bayesian belief update comes from (Fine & Jaeger, 2011). We briefly describe the design of that experiment.

### Experimental Procedure

In a between-subjects, multi-visit self-paced reading experiment, (Fine & Jaeger, 2011) investigated whether comprehenders update their estimates of the probability of the syntactic structures in (1) conditioned on the verb used in the sentence and the presence of the complementizer *that*. The All-SC group received evidence that SC-taking verbs always occur in sentences like (1a), while the 50-50 group was exposed to a 50/50 mix of SC (1a) and DO (1b) structures. For both groups, *that* occurred in 50% of all SC sentences. The experiment consisted of a pre-exposure session, three exposure sessions over at least 6 days, and a final post-exposure session at least 2 days after the last exposure session (cf. Wells,

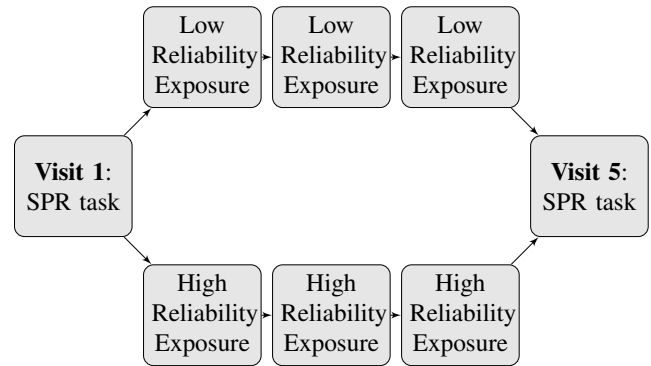


Figure 1: Schematic illustration of experimental procedure

Christiansen, Race, Acheson, & MacDonald, 2009). Subjects in both groups completed identical self-paced reading tasks in the first and final visits. A schematic representation of the experimental procedure is given in Figure 1. Given the design of the experiment, if reading behavior changes from visit 1 to visit 5 differentially across the groups, we can attribute this difference to the distributional properties of the exposure phase within each group.

This experiment allows us to ask two questions. First and foremost, do comprehenders track the distribution of syntactic structures in a given environment? That is, do comprehenders *adapt* to the statistical properties of novel linguistic situations? Second, given the distributional properties of that linguistic environment, do comprehenders *combine* multiple cues to syntactic structure in a way that is rational (i.e., by weighting cues according to their reliability; cf. Bates & MacWhinney, 1987; Anderson, 1990)? Specifically, for subjects in the 50-50 group, all verbs participating in the ambiguity in (1) become—in the experimental context—equally likely to occur with an SC or a DO. Thus the verb becomes, for this group, a very *unreliable* cue to syntactic structure. Qualitatively, according to rational models of cue combination, this should lead subjects in the 50-50 group to depend more on the complementizer *that* in the post-exposure reading task relative to subjects in the All-SC group. That is, the presence of the complementizer should more strongly influence reading times for the 50-50 group. In a regression framework, we therefore predicted a 3-way interaction between time (pre- vs. post-exposure), ambiguity (presence vs. absence of the complementizer *that*), and group (All-SC vs. 50-50). This three-way interaction was significant ( $\beta = 2.4, SE = .9, p < .05$ ), and is visualized in Figure 2 (light bars). This figure shows the decrease in ambiguity effect from pre-test to post-test as a function of training: there is a greater decrease after high-reliability, All-SC training, where subjects do not need to rely on the complementizer as much.

### Modeling framework

In constructing a belief updating model of syntactic adaptation and cue combination, there are two main considerations.

First, how can syntactic expectations be quantified, and second, how are those expectations related to the cues present in the linguistic environment and updated based on linguistic experience? In our model, syntactic expectations are quantified as discrete probability distributions over syntactic structures. In this case, the relevant syntactic structures are possible completions of sentences like (1), which we assume are limited to sentence complement completions (SC) and direct object completions, etc. (DO).

Syntactic expectations are related to relevant cues and in turn to linguistic experience via the conditional probability. SC completions are more common for some verbs than others, and are more common when the complementizer *that* is present. This dependence is captured by the conditional probability of  $p(S = \text{SC}|T, V)$ , where  $T$  indicates the presence or absence of *that* and  $V$  indicates the identity of the main verb of the sentence. The conditional probability of  $S = \text{SC}$  is closely related to the joint probability via the base probability of the various possible combinations of cue values:  $p(S, T, V) = p(S|T, V)p(T, V)$ .

We can model this joint distribution of syntactic structures and cues via a multinomial distribution. A multinomial distribution assigns a probability to a group of observations, each of which is, in our case, a triplet of the form  $S = i, T = j, V = k$ , each with a known probability of  $p(S = i, T = j, V = k) = \theta_{i,j,k}$ . The likelihood of making a group of observations  $X$ , where each outcome occurs  $n_{i,j,k}$  times, is

$$p(X|\theta) \propto \prod_{i,j,k} \theta_{i,j,k}^{n_{i,j,k}}$$

This provides a way of capturing syntactic expectations when the probability of each outcome is known with certainty. However, if the subject is truly certain about the probability of each outcome, then no adaptation should occur. Thus, in order to capture adaptation, or change in expectations, we must capture *uncertain* beliefs about such expectations, via a prior distribution over the probabilities  $\theta_{i,j,k}$ . The most natural choice is the conjugate prior for multinomial probabilities, the Dirichlet distribution:

$$p(\theta) \propto \prod_{i,j,k} \theta_{i,j,k}^{\alpha_{i,j,k}-1}$$

The primary advantage of using this prior distribution is that, after making observations  $X$ , the posterior is also Dirichlet, with parameters  $\alpha_{i,j,k} + n_{i,j,k}$ :

$$p(\theta|X) \propto p(X|\theta)p(\theta) \propto \prod_{i,j,k} \theta_{i,j,k}^{\alpha_{i,j,k}+n_{i,j,k}-1}$$

The parameters ( $\alpha_{i,j,k}$ ) of the Dirichlet prior can thus be interpreted as the number of times each outcome was observed in prior experience. Intuitively, this can be seen just by looking at the equations for the prior and likelihood and noting that  $\alpha$  and  $n$  appear in the same places.

Under this model, the conditional probability of SC, given specific  $V = v$  and  $T = t$  is

$$p(\text{SC}|V = v, T = t, \theta) = \theta_{\text{SC}|v,t} = \frac{\theta_{\text{SC},v,t}}{\theta_{\text{SC},v,t} + \theta_{\text{DO},v,t}}$$

It can be shown that  $\theta_{\text{SC}|v,t}$  (and by definition  $\theta_{\text{DO}|v,t} = 1 - \theta_{\text{SC}|v,t}$ ) follows a Beta distribution with parameters ( $\alpha_{\text{SC},v,t}, \alpha_{\text{DO},v,t}$ ). Marginalizing over the distribution of  $\theta_{i,j,k}$ , the conditional probability of SC given verb  $v$  and complementizer cue  $t$  is:

$$p(\text{SC}|v,t) = \int p(\text{SC}|v,t,\theta)p(\theta|\alpha)d\theta = \frac{\alpha_{\text{SC},v,t}}{\alpha_{\text{SC},v,t} + \alpha_{\text{DO},v,t}}$$

This conditional probability is the major predictor of syntactic expectation and associated comprehension difficulty.

In order to make quantitative predictions from this general framework, it is necessary to specify the parameters of the prior distribution ( $\alpha_{i,j,k}$ ) and likelihood function ( $n_{i,j,k}$ ), and to relate the model output (conditional probability) to the behavioral measure (reading times). These are addressed in the next sections.

### Determining the likelihood

The parameters of the likelihood function are the counts  $n_{i,j,k}$  of how often each unique combination of syntactic structure  $S = i$ , verb  $V = j$ , and complementizer presence/absence  $T = k$  was observed during training, and were set to the counts of the training phase.

### Determining the prior

The prior parameters are the pseudo-counts  $\alpha_{i,j,k}$  which are proportional to the joint prior probabilities. These probabilities are estimated based on a combination of corpus and norming data. The joint probability of syntactic structures, verbs, and complementizer presence  $p(S, T, V)$  can be factored as

$$p(S, T, V) = p(T|V, S)p(S|V)p(V)$$

The verb frequency  $p(V)$  is estimated from the British National Corpus, while the SC-bias of each verb (probability of SC completion)  $p(S|V)$  and *that*-bias (probability of complementizer presence for SC completions of each verb)  $p(T|V, S)$  are estimated based on a sentence-completion norming study (Garney et al., 1997).

Together, these factors provide an estimate of the relative prior frequency of each outcome, and thus the relative magnitudes of the  $\alpha_{i,j,k}$  Dirichlet prior parameters, but say nothing about their absolute magnitude. The absolute magnitude  $A = \sum_{i,j,k} \alpha_{i,j,k}$ , corresponds to the degree of confidence in the prior beliefs: the higher  $A$ , the more the distribution over the modeled probabilities  $\theta_{i,j,k}$  is peaked around the estimated prior frequency, and the less new observations will influence these beliefs. Since there is no way to determine the strength of the prior beliefs *a priori*, we treat  $A$  as a free parameter, which controls the degree of adaptation but does not change

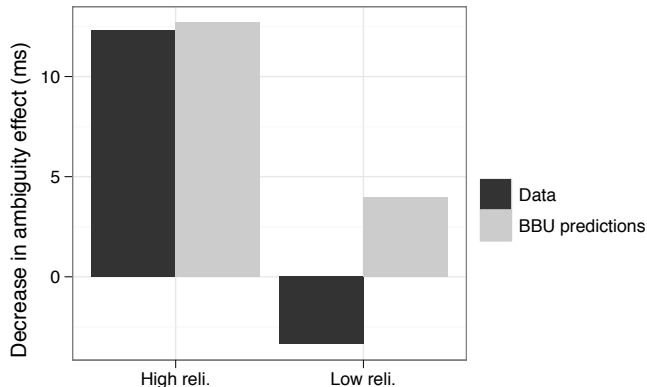


Figure 2: Behavioral results from Fine and Jaeger (2011) (light bars) and corresponding model predictions (dark bars), showing differential effect of high- and low-reliability training on ambiguity effect. Bars show decrease in ambiguity effect (difference in reading times for *that*-present vs. *that*-absent sentences) from pre-test to post-test.

its shape. This is the only free parameter of the model simulations reported here.<sup>1</sup>

## Analysis

To evaluate the predictions of the model against the behavioral data of Fine and Jaeger (2011), we regressed the negative log conditional probability against length-corrected reading times. This measure is known as *surprisal*, and has been shown to be a good predictor of reading times in syntactic comprehension (Levy, 2008; Hale, 2001; Fine, Qian, Jaeger, & Jacobs, 2010).

However, there are many factors which influence reading times, of which syntactic expectation may be just one. This measure explicitly removes the influence of verb frequency and *that*-bias, which independently predict reading times for SCs. Also, reading times decrease in self-paced reading tasks just because subjects become better at “pushing the button”, an effect which will confound any difference in reading times between pre- and post-test due to adaptation.

In order to control for these confounding effects and evaluate their relationship with our model’s predictions about adaptation of syntactic expectations, we fit an increasingly complex series of linear mixed-effects regression models. For each regression model, we compared the baseline, with only the “standard” suite of predictors, with the belief-updating model which additionally includes the surprisal of each item as a predictor. Table 1 shows the factors that were included in each model.

<sup>1</sup>The probability of *that* occurring as a non-complementizer,  $p(T = \text{that} | V, S = \text{DO})$  cannot be determined from the same norming study, and in corpora it varies dramatically between spoken and written English. For the simulations reported here it was fixed at 0.00005, based on the very low but non-zero value observed in the Wall Street Journal corpus. This does not dramatically change the predictions or the best-fitting pseudocount  $A$ .

Model (adds)	Deviance	$\Delta\text{Dev.}$	df	$p(\chi^2)$
(1 Subj)+ (1 Verb)	28066	195	2	<0.0001
Verb Freq.	28059	195	2	<0.0001
Time (pre/post)	27452	25	2	<0.0001
That presence	27448	15	2	0.0006
Verb SC Bias	27438	18	2	0.0001
Training Group	27432	15	2	0.0005
(interactions)	27216	7	2	0.03

Table 1: Results of linear mixed-effects regression analysis of belief-updating model predictions of self-paced reading times. Each row reports the goodness of fit of a model with belief-updating-predicted surprisal and all of the fixed and random effects listed in its row and the preceding rows (that is, the third model includes random intercepts for subjects and verbs, and a fixed effect for verb log-frequency). In the final row, the baseline model has all fixed effects and their interactions, except for the predicted surprisal, which does not interact with any other effects. The deviance (negative two times model log likelihood) is reported for each model, along with the improvement in deviance over the corresponding baseline model, the additional degrees of freedom, and the result of a  $\chi^2$  test.

We found the overall best-fitting parameter values by fitting the whole series of belief-updating regression models using a range of parameter values ( $A = 10^{-3}$  to  $10^4$ ). The parameter value with the best mean normalized goodness-of-fit (across the various regression models) was used for the results below. We compared both  $r^2$  and deviance as measures of regression goodness-of-fit; both measures produce similar relative goodness-of-fit values but we chose to use deviance since it suffers less from ceiling effects in the most complex (and best-fitting) models. The best fitting prior pseudocount, used to generate the predictions evaluated below, was  $A = 2.7$ .

## Results

Qualitatively, the belief-updating model predicts the three-way interaction between group, time, and ambiguity (presence or absence of *that*), Figure 2. The degree to which subjects rely on the complementizer as a cue to SC continuations—i.e., the strength of the effect of the complementizer on RTs—can be measured by the difference in reading times between complementizer-present and -absent sentences. The model predicts (dark bars), as is observed in the data (light bars), that ambiguity effects should decrease more after high-reliability training (All-SC group) and decrease less (if at all) after low-reliability training (50-50 group).

The results from the regression analysis of the belief updating model predictions show that the model predictions generally provide a good fit for reading times in the disambiguating region. First, the model predictions alone (with random intercepts for subject and verb) account for 17% of the variance in reading times. This effect cannot be reduced to any of the other controls we evaluated (verb frequency, time, presence of

*that*, verb SC-bias, and training condition group): the belief-updating predictions still significantly improve the fit of the model even after controlling for all of these fixed effects and all of their interactions ( $\chi^2(2) = 7, p < 0.03$ ).

Of all of these control predictors, time (pre- vs. post-training) has a notably large effect, and Wells et al. (2009) observed a similar global speed-up between pre- and post-test, independent of effects on the form of the test sentences and the type of training the subject received. This speed-up reflects both increased familiarity with the self-paced reading task (demonstrated by the fact that when the Time predictor is added, the deviance accounted for by the belief-updating predictions is reduced) and the effects of simply having seen more SC structures than in typical written English (captured by the belief-updating predictions after some SC exposure; Fine et al., 2010).

Finally, the value of the prior confidence pseudocount parameter  $A$  which best fits the data corresponds to an effective sample size of 2.7 observations for the prior beliefs. This value is very low, but is consistent with results from other belief-updating models of rapid syntactic adaptation and phonetic adaptation (Fine et al., 2010; Kleinschmidt & Jaeger, 2011) and with the larger idea that comprehenders weight prior evidence less in novel situations where rapid adaptation is likely required. Higher values correspond to less adaptation, and produce a worse fit, but interestingly *lower* values produce a worse fit as well. If the good fit of this model were simply due to the qualitative pattern of cue reliability in the exposure statistics, then lower values of  $A$ , which result in post-test reading times which better approximate the exposure statistics, should produce better fits, which is not the case. This supports the idea that post-test reading times reflect a combination of prior knowledge and recent experience.

## Discussion and conclusion

In this paper, we formally characterize syntactic adaptation as the incremental updating of a probability distribution over syntactic structures. We showed that such a model provides a good fit of behavioral data from a previously published study of syntactic adaptation (Fine & Jaeger, 2011). This model is a first step and leaves much open for future work. Because of how it tracks the co-occurrence statistics of cues and syntactic structures, it does not make meaningful predictions on a trial-by-trial basis for how the overall greater prevalence of SC structures in the experiment changes syntactic expectations for verbs not encountered during training (which influences reading times as well, Fine et al., 2010). Such on-line generalization is not in principle beyond the scope of the type of model presented here, but is omitted in favor of presenting a simple model which demonstrates the connections between adaptation, cue-combination, and statistical learning in syntactic comprehension.

Independent of the details of the particular model presented in this paper, characterizing syntactic adaptation in terms of Bayesian belief update is appealing for at least two rea-

sons. First, by modeling syntactic adaptation as incremental Bayesian belief update, we provide a natural, formal connection between previous work on probabilistic models of expectation-based processing (e.g., Hale, 2001; Levy, 2008) and behavioral work on syntactic adaptation (or syntactic priming; e.g., (e.g., Thothathiri & Snedeker, 2008). Second, using this modeling approach has allowed us to take a step towards providing a single computational framework for adaptation phenomena in language processing, since the same approach has been successfully applied in phonetic adaptation (e.g., Kleinschmidt & Jaeger, 2011). Providing such a “common language” is an important step since this provides a way of bridging insights from multiple strands of psycholinguistic research that have previously been pursued in isolation from each other, notably syntactic priming in comprehension (e.g., Thothathiri & Snedeker, 2008) and perceptual adaptation in speech (e.g., Norris et al., 2003).

As we briefly mentioned in the introduction, the model reported here provides a way of formally describing both adaptation and cue combination. Bayesian models of perception have consistently suggested that, when multiple cues are available in a given task, the perceptual system *weights* those cues according to how *reliable* they are, or, more formally, how narrow or wide the variance is over inferences made *based on* those cues (Jacobs, 2002). In the exposure phase of the study modeled here, the reliability of the verb as a cue to syntactic structure is very high in the All-SC group, but very low in the 50-50 group; on the other hand, the complementizer *that* is a consistently good cue across both groups. Our model qualitatively captures the behavioral result that comprehenders in the 50-50 group come to rely *more* on the complementizer *that* in the post- relative to the pre-exposure phase, compared to the All-SC group (see Figure 2). Significantly, this result comes out of the model as a natural consequence of tracking the joint distribution over syntactic structures (DOs vs. SCs) and syntactic cues (complementizers, verbs). The model here therefore suggests a very close relationship between adaptation and cue combination, and provides a formal account for the classic observation that cues are weighted differentially according to their reliability in language acquisition and language processing (Bates & MacWhinney, 1987).

In general, the approach here is conceptually compatible with a sentence processing research emphasizing the role of experience and learning in language comprehension (e.g., MacDonald, 1999). Bayesian models provide a formal framework for capturing the assumption—shared by many experience-based accounts of processing—that comprehenders monitor and constantly integrate new evidence from the input in order to maintain accurate linguistic expectations, and to process language more efficiently (cf., Smith & Levy, 2008).

The results reported here raise a number of interesting questions that we intend to pursue going forward. First, we employ the same modeling framework and find results that

are generally consistent with the modeling results reported in Fine et al. (2010). However, important differences in the experimental design between the two studies leave many questions open. Most significantly, the experiment in Fine et al. (2010) observed changes in reading behavior over a much shorter period of time (one half-hour experimental session) than the current study, which lasted several days. The modeling framework employed here could be extended to examine possible qualitative differences in adaptation over very different time courses, paralleling Bayesian accounts of the time course of adaptation in speech perception (Kleinschmidt & Jaeger, 2011).

Finally, the experiment and model reported here leave open the question of how much the changes in expectations which constitute adaptation *generalize* to novel situations (i.e., did the adaptation effects observed here persist, and influence the way subjects processed language outside the lab?). Rational models of linguistic adaptation generally predict that the extent to which comprehenders generalize adapted expectations should depend on their prior beliefs about the degree of similarity between different situations. This question has been addressed behaviorally in phonetic adaptation (Kraljic & Samuel, 2006, 2007) but remains virtually unexplored in other domains of language processing, and has not been quantitatively modeled. Answering the question of generalization is therefore a high priority for future work on adaptation.

### Acknowledgements

This research was partially funded by NSF Graduate Research Fellowships to DK and ABF and NSF Grant BCS-0844472 as well as an Alfred P. Sloan Research Fellowship to TFJ.

### References

- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Bates, E., & MacWhinney, B. (1987). Competition, Variation, and Language Learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157–194). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Blakemore, C., & Campbell, F. (1969). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *The Journal of Physiology*, 203(1), 237–260.
- Farmer, T. A., Fine, A. B., & Jaeger, T. F. (2011). Implicit Context-Specific Learning Leads to Rapid Shifts in Syntactic Expectations. In *The 33rd annual meeting of the cognitive science society (cogsci11)*. Boston, MA.
- Fine, A. B., & Jaeger, T. F. (2011). Language comprehension is sensitive to changes in the reliability of lexical cues. In *The 33rd annual meeting of the cognitive science society (cogsci11)*. Boston, MA.
- Fine, A. B., Qian, T., Jaeger, T. F., & Jacobs, R. A. (2010). Syntactic adaptation in language comprehension. In *Acl workshop on cognitive modeling and computational linguistics* (pp. 18–26).
- Garnsey, S., Pearlmutter, N., Myers, E., & Lotocky, M. (1997). The Contributions of Verb Bias and Plausibility to the Comprehension of Temporarily Ambiguous Sentences. *Journal of Memory and Language*, 37(37), 58–93.
- Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 NAACL 01*, 2, 1–8.
- Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in cognitive sciences*, 6(8), 345–350.
- Kaschak, M. P., & Glenberg, A. M. (2004). This construction needs learned. *Journal of experimental psychology. General*, 133(3), 450–67.
- Kleinschmidt, D., & Jaeger, T. F. (2011). A Bayesian belief updating model of phonetic recalibration and selective adaptation. In *2nd acl workshop on cognitive modeling and computational linguistics*.
- Koerding, K. P., Tenenbaum, J. B., & Shadmehr, R. (2007). The dynamics of memory as a consequence of optimal adaptation to a changing body. *Nature Neuroscience*, 10(6), 779–86.
- Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic bulletin & review*, 13(2), 262–8.
- Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, 56(1), 1–15.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–77.
- Luka, B., & Barsalou, L. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language*, 52(3), 436–459.
- MacDonald, M. C. (1999). Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. In *The emergence of language* (pp. 177–196). Mahwah, NJ: Erlbaum.
- Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47(2), 204–238.
- Smith, N., & Levy, R. (2008). Optimal processing times in reading: A formal model and empirical investigation. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 595–600). Austin, TX.
- Sonderegger, M., & Yu, A. (2010). A rational account of perceptual compensation for coarticulation. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 375–380).
- Thothathiri, M., & Snedeker, J. (2008). Give and take: syntactic priming during spoken language comprehension. *Cognition*, 108(1), 51–68.
- Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: statistical learning and relative clause comprehension. *Cognitive psychology*, 58(2), 250–71.

# A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation

Dave Kleinschmidt<sup>1</sup> and T. Florian Jaeger<sup>1,2</sup>

{dkleinschmidt, fjeager} @ bcs.rochester.edu

<sup>1</sup>Department of Brain and Cognitive Sciences and <sup>2</sup>Department of Computer Science,  
University of Rochester, Rochester, NY, 14607 USA

## Abstract

We have previously proposed that incremental belief updating can provide a unified account of the effect of cumulative exposure on phonetic recalibration and selective adaptation (Kleinschmidt & Jaeger, 2011). This model predicts that these are not two distinct phenomena but rather two points on a continuum. We investigate that prediction here using adaptor stimuli intermediate between those which induce recalibration and selective adaptation, and find that the quantitative predictions of the model fit the data well. We also demonstrate that with the proper controls, Mechanical Turk provides a suitable online platform for speech perception experiments.

**Keywords:** Phonetic adaptation; Speech perception; Recalibration; Selective adaptation; Bayesian models; Mechanical Turk

## Introduction

Language is fraught with variability, and such variability is a particular problem in the mapping from acoustic cues to phonetic categories. Some of this variability is simply random noise, from imprecision in the production and perception systems, but much of it is systematic, reflecting real differences in the way that individual speakers realize phonetic categories.

One of the ways in which the human speech perception system deals with such systematic variability is through rapid recalibration of phonetic categories in response to novel accents (Bradlow & Bent, 2008; Kraljic, Brennan, & Samuel, 2008) and unusual pronunciations of particular phonetic categories (Norris, McQueen, & Cutler, 2003; Vroomen, Linden, Gelder, & Bertelson, 2007). Through repeated exposure to a sound which is acoustically ambiguous between, say, /b/ and /d/, but which is disambiguated via visual or lexical information as a /b/, the listener's categorization behavior will change: they will classify more items on a /b/-to-/d/ continuum as /b/.

Such recalibration is often contrasted with selective phonetic adaptation, where repeated exposure to an *prototypical* /b/ results in exactly the opposite effect on the listener's categorization behavior: they will classify fewer items on the continuum as /b/ (Eimas & Corbit, 1973; Samuel, 1986; Vroomen, Linden, Keetels, Gelder, & Bertelson, 2004; Vroomen et al., 2007). Selective adaptation and recalibration are generally held to be qualitatively distinct, for a number of reasons.

First, recalibration results from repeated exposure to an acoustically ambiguous stimulus, while selective adaptation relies on repeated exposure to an unambiguous stimulus, and

is in fact stronger for more prototypical stimuli (Miller, Conine, Schermer, & Kluender, 1983). Second, despite the fact that both come from repeated exposure to a phonetic category member, they have opposite effects on that category.

Third, selective adaptation and recalibration are thought to reflect different levels of processing. On the one hand, selective adaptation is thought to generally reflect low-level, perceptual processing (Samuel, 1986). On the other, phonetic recalibration (or “perceptual learning for speech”, Norris et al., 2003) seems to reflect reorganization of the phonetic categories themselves (Clarke-Davidson, Luce, & Sawusch, 2008; Maye, Aslin, & Tanenhaus, 2008), is long-lasting (Eisner & McQueen, 2006), and in some cases generalizes to new speakers and phonetic contrasts (Kraljic & Samuel, 2006).

Fourth, the amount of cumulative exposure has a dramatically different effect on selective adaptation and recalibration (Vroomen et al., 2007). Selective adaptation grows stronger with more exposure, while recalibration becomes weaker after an initial rise, to the extent that it is eventually extinguished (Figure 2, left).

Despite these differences, selective adaptation and recalibration have something in common: they are reactions to unusual distributions of acoustic-phonetic cues. Listeners are exquisitely sensitive to such distributional information (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Maye, Werker, & Gerken, 2002), and we propose that incremental belief updating might provide a unified account of a range of phonetic adaptation phenomena, possibly including selective adaptation (Kleinschmidt & Jaeger, 2011).

This approach suggests that at least in some cases, selective adaptation and perceptual recalibration are better understood as part of a continuum of phonetic adaptation effects, rather than as qualitatively distinct processes, which raises the question of what sort of adaptation effect intermediate adaptors will produce. In this paper, we investigate the quantitative predictions about such intermediate adaptors made by our belief-updating model.

First, we describe how incremental belief updating provides a unifying account of the effects of cumulative exposure on recalibration and selective adaptation (Kleinschmidt & Jaeger, 2011). This model is used to derive quantitative predictions for the effect of cumulative exposure to audiovisual speech stimuli on a continuum between acoustically ambiguous and prototypical. These predictions are tested against human behavior. Finally, we conclude with a discus-



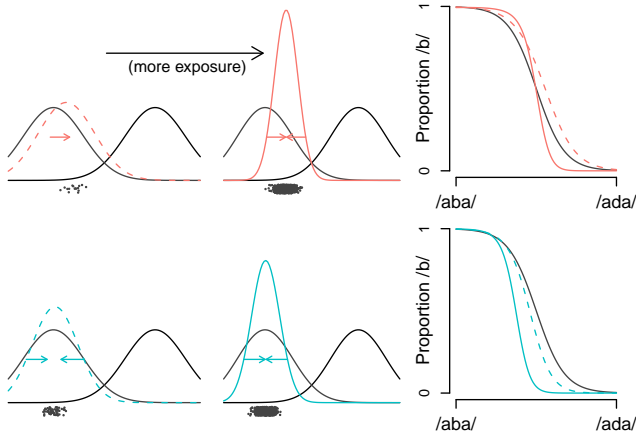


Figure 1: Qualitative predictions of the Bayesian belief updating model for phonetic recalibration (top) and selective adaptation (bottom). Left panels show changes in mean/variance after a small number of adaptor trials, and middle panels show the effect of further exposure. Note that while the adaptor for recalibration is acoustically ambiguous, it is disambiguated (by means of, e.g., a visual /aba/), and thus the percept is biased towards the /b/ prototype. Right panels show corresponding identification curves.

sion of future directions for both the computational modeling and behavioral work.

## Model and predictions

Speech perception can be considered a process of statistical inference: given a set of acoustic-phonetic cues, what is the most likely intended phonetic category (Sonderegger & Yu, 2010; Feldman, Griffiths, & Morgan, 2009; Kleinschmidt & Jaeger, 2011)? This inference critically depends on accurate beliefs about the distributions of acoustic-phonetic cues for each category. If the listener’s expectations about what a /b/ sounds like do not match what the speaker produces then the listener is at risk of mistaking the speaker’s intended phonetic category. Thus, a rational listener must update their expectations based on experience.

Figure 1 shows how such incremental belief updating explains the qualitative differences between recalibration (top) and selective adaptation (bottom) based on the respective distributions of acoustic cues (Kleinschmidt & Jaeger, 2011). Phonetic categories are naturally “wide”, with random variability inherent in perception and production processes. Selective adaptation effects are captured by the fact that in the prototypical-/b/ condition, the unusually tight clustering of acoustic cues from the repeated adaptor stimulus causes the /b/ category to shrink, resulting in fewer /b/ responses to test stimuli from the middle of the continuum.

For recalibration, the initial rise in /b/ responses after exposure to the acoustically ambiguous (but visually or lexically disambiguated) /b/ adaptor is due to a shift in the *mean* of the /b/ category, towards the middle of the continuum. With

more exposure, however, the tight clustering again causes the /b/ category to contract. Because the acoustically ambiguous adaptor is biased towards the prototypical /b/ by the disambiguating (visual or lexical) cue (by a McGurk effect or perceptual magnet effect, etc.), the shrinking of the /b/ category means that, like with selective adaptation, it pulls away from the middle of the continuum, resulting in an eventual decrease in /b/ responses.

This belief-updating model treats the behavioral phenomena of selective adaptation and recalibration as two points on a continuum of phonetic adaptation effects, which is determined by the statistics of the adaptor stimuli. One way of describing these statistics is by the mean and the variance of the adaptor stimuli, and our model makes quantitative predictions about the whole range of adaptors, from prototypical adaptors (which produce selective adaptation) to auditorily ambiguous adaptors (which produce phonetic recalibration), as well as intermediate adaptors, where the adaptor is neither fully auditorily ambiguous nor prototypical.

Specifically, phonetic adaptation with such intermediate audio-visual adaptors should be intermediate in time course and in direction. One simple way of characterizing the time course of phonetic adaptation is the number of exposures at which the direction of the effect of /b/ exposure switches from more /b/ classification to less /b/. In Vroomen et al. (2007, replotted in Figure 2, left), this “switch point” for the ambiguous condition is essentially the end of one full block: /b/-classification is greater after ambiguous /b/ (solid lines) than after ambiguous /d/ exposure (dashed lines) until the very end of the block, at 256 trials, where the two become indistinguishable again. For the prototypical condition, there is less /b/ classification after /b/ exposure from nearly the first adaptor repetition, and so can we say that the switch point is at the beginning of the block. For intermediate adaptors, this switch point will be between the point for fully ambiguous and fully prototypical adaptors, closer to the ambiguous switch point for more ambiguous adaptors and closer to the unambiguous point for less ambiguous adaptors (Figure 3, black curves with arrows showing switch points).

## Model specification

We refer the reader to Kleinschmidt and Jaeger (2011) for full details of the model. This model treats phonetic categories as Normal distributions over acoustic-phonetic cues,  $p(x_i|c_i) \sim \mathcal{N}(\mu_{c_i}, \sigma_{c_i}^2)$ . The listener’s beliefs about these categories are captured by Normal-Gamma probability distributions over the category means and variances,  $p(\mu_c, \sigma_c^2)$ . After exposure to adapting stimuli  $X$  from category  $c$ , the listener’s beliefs are updated via Bayes rule,  $p(\mu_c, \sigma_c^2|X, c) \propto p(X|\mu_c, \sigma_c^2)p(\mu_c, \sigma_c^2)$ . This finds the best compromise between the listener’s prior beliefs and the mean and variance that is most consistent with the data.

The probability of a /b/ response to a stimulus  $x$ ,  $p(b|x)$ , depends on the likelihood of  $x$  under each category  $p(x|b)$  and

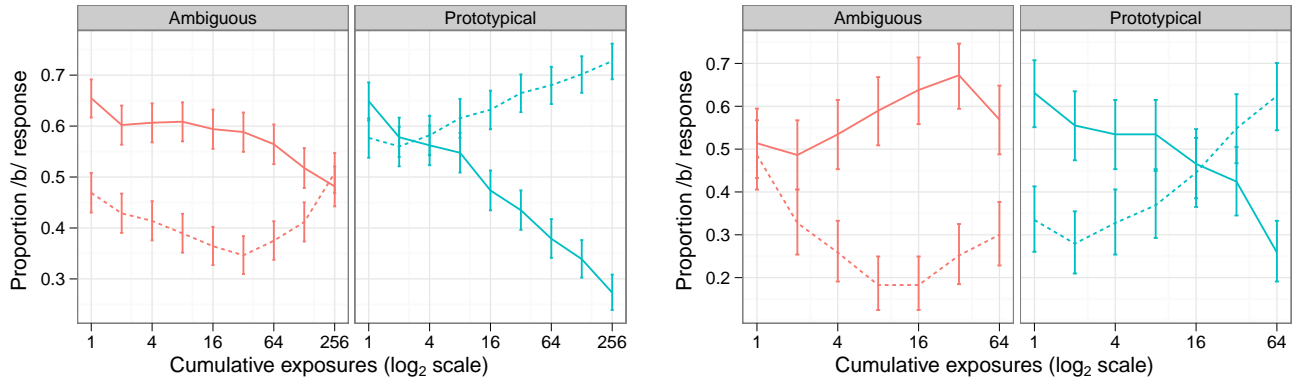


Figure 2: Results from Vroomen et al. (2007), left, and Experiment 1, right: More /b/-classification after ambiguous /b/ exposure (solid lines, left panels) than /d/ exposure (dashed lines). After prototypical-/b/ exposure (solid lines, right panels), less /b/-labeling.

$p(x|d)$ , as well as the prior probability of /b/ and /d/:

$$p(b|x) = \frac{p(x|b)p(b)}{p(x|b)p(b) + p(x|d)p(d)}$$

The predictions of the model depend not only on the mean and variance of the adaptor stimuli (the likelihood function) but also the prior beliefs about the category means and variances. While each subject's most likely means and variances can be determined based on their pre-test classification data, the confidence in these beliefs must be treated as free parameters (Kleinschmidt & Jaeger, 2011).

Here we incorporate a few minor improvements over our original model. Instead of fitting the after-effect measure used by Vroomen et al. (2007) (the difference between /b/ and /d/ exposure), here the model is fit to the /b/-classification rate curves for each individual condition. This requires accounting for asymmetry in the prior probability of /aba/ vs. /ada/, which was fixed to  $p(b) = 0.7$ , based on the English biphone probabilities of /aba/ and /ada/ (Vitevitch & Luce, 2004).<sup>1</sup> The free parameters are fit via MCMC sampling (rather than a grid-search) based on the binomial likelihood of the responses (using a weakly-informative prior).

We next describe experiments which test the predictions of this model. These are extensions of the paradigm of Vroomen et al. (2007), which we have adapted to run on a novel web-based platform. The first experiment replicates Vroomen et al. (2007) to validate this platform, and the second investigates the effect of cumulative exposure on adaptation to intermediate adaptors.

## Experiment 1

Our first experiment is a replication of Vroomen et al. (2007). This is important both to verify the findings of Vroomen et

al. (2007) with English speakers but also to verify our web-based platform for running audio-visual phonetic adaptation studies.

## Participants

24 participants were recruited and run via Amazon's Mechanical Turk service. An additional 4 subjects were excluded by the criteria described below.

## Procedure

Our procedure is adapted from Vroomen et al. (2007), and our stimuli are identical in order to maximize similarity to the original study for purposes of replication (we thank Jean Vroomen for graciously providing them). These stimuli consist of a 9-item /aba/-to-/ada/ continuum synthesized by manipulating F2 locus in equal steps on a log-frequency scale. Audio-visual adaptor stimuli were created by combining a video of a speaker producing /aba/ or /ada/ with one continuum item, according to the condition.

Subjects first did a pre-test and calibration phase, where they classified items from the audio continuum as /b/ or /d/. The point of maximum ambiguity was found via logistic regression for each subject. Subjects who mis-classified any of the two most prototypical stimuli on each end (less than 70% accuracy) were excluded, as were any subjects whose maximally ambiguous stimulus was not one of the three middle continuum items 4, 5, or 6. Three subjects (11%) were excluded by these criteria and replaced.

During the main phase, in each of four blocks subjects saw 64 repetitions of an adaptor stimulus. This stimulus was visually either /b/ or /d/, and auditorily either a *prototypical*, unambiguous rendition of the visual category (continuum item 1 or 9, respectively), or auditorily *ambiguous* (that subject's most ambiguous continuum item as determined during pre-test). These four blocks were presented in latin-square order across subjects. Audio-only test stimuli were interspersed throughout each block in sets of six (two repetitions of each

<sup>1</sup>When the prior probability of /b/ vs. /d/ is treated as a free parameter, the best-fitting value is 63% chance of /b/, very close to the corpus-derived value.

of the three most ambiguous stimuli) after 1, 2, 4, 8, 16, 32, and 64 exposures.

To ensure that subjects were paying attention to the videos, catch trials were randomly interspersed, where a small white dot flashed for one frame below the nose of the speaker. Subjects who failed to respond to at least 80% of these trials (or missed more than 50% in one block) were excluded; only one subject was excluded by these criteria and replaced. Besides this subject, accuracy was 95% (compared to 93% in Vroomen et al., 2007).

## Results

Because subjects were not run in the laboratory, there is some uncertainty about the conditions under which our subjects are doing the task. As one control for this, we compared their performance on the pre-test calibration task with the performance of the subjects in Vroomen et al. (2007) using a logistic mixed-effects regression model, with predictors for stimulus and group (web vs. lab) and with the maximum random effects structure justified by the data. Web and lab subjects did not differ in the slope of their category boundaries ( $\beta = 0.07$ ,  $SE = 0.14$ ,  $p = 0.60$ ). The category boundary (point of maximum ambiguity) was on average about one half to one continuum step closer to the /b/ end of the continuum compared to the lab subjects ( $\beta = -1.15$ ,  $SE = 0.27$ ,  $p < 10^{-4}$ ).

Our results from the main phase replicate Vroomen et al. (2007) as well (Figure 2). In the prototypical (selective adaptation) condition, subjects classify fewer test items as /b/ after /b/ exposure, and this effect grows stronger with further exposure. In the ambiguous (recalibration) condition they classify *more* as /b/ after /b/ exposure, but this effect eventually begins to weaken. Because our subjects only saw 64 exposure trials in each block, versus 256 in Vroomen et al. (2007), recalibration does not completely dissipate by the end of the block, but it does begin to weaken.

The most notable difference between our findings and those of Vroomen et al. (2007) is that in the prototypical, selective adaptation condition of our experiment, /b/-exposure initially produces *more* /b/-classification (as in the ambiguous, recalibration condition). There are two possible explanations for this. First, the “prototypical” stimuli we use correspond to natural productions of /aba/ and /ada/ by Dutch speakers, and may not be fully prototypical to English listeners, which, as we will see in Experiment 2, can produce recalibration-like effects with small amounts of exposure. Second, adaptation could carry over from one block to the next, producing baseline shifts. Such carry-over effects are observed in the data of Vroomen et al. (2007) as well, but are effectively confined to earlier blocks of their experiment (each of their subjects did a total of 16 blocks, versus only four here) and thus do not show up as strongly in the averaged data. In addition, at the end of the ambiguous (recalibration) blocks there is a bigger difference between /b/ and /d/ exposure in our data than in Vroomen’s et al., which would magnify any carry-over effects on the prototypical blocks (which are preceded two-thirds of the time by an ambiguous block).

## Experiment 2

The second experiment tests the predictions of our belief-updating model about the effect of cumulative exposure to adaptor stimuli intermediate between the prototypical and auditorily ambiguous stimuli used in the previous experiment. This also serves as another replication of Vroomen et al. (2007) using our web-based approach.

## Participants

A total of 126 participants were recruited, run, and paid through Amazon’s Mechanical Turk service as in Experiment 1. Six (5%) were excluded for poor catch trial performance, eight (6%) for misclassifying too many prototypical pre-test stimuli, and 18 (14%) for unusual category boundaries (as above).<sup>2</sup> After these exclusions, 94 subjects remained.

## Procedure

The procedure was similar to that from Experiment 1, except that there were four ambiguity conditions rather than two, and they were distributed between subjects rather than within. In addition to the prototypical (selective adaptation) and ambiguous (recalibration) conditions from Experiment 1, there were two intermediate conditions. In the intermediate-ambiguous condition, the audio component was one step towards the visual category prototype from the maximally ambiguous stimulus, and in the intermediate-prototypical condition it was two steps.

Each subject performed two blocks: one /aba/ and one /ada/ of one of the four ambiguity conditions, in order to minimize the effects of carry-over between blocks. Each block was 128 exposure trials, with test trials as in Experiment 1.

## Results

Figure 3 shows the results from all four conditions. As predicted by the belief-updating model, as the auditory component of the adaptor is less and less ambiguous, the switch point between recalibration-like and selective adaptation-like behavior moves closer and closer to the beginning of the block (Figure 4).

The quantitative predictions of the model are also a good match for the observed data. The predictions of the model are shown in black. These predictions correspond to the free parameters which best fit the ambiguous and unambiguous conditions. As in Kleinschmidt and Jaeger (2011), the model predictions provide a good fit for the behavior in the ambiguous and unambiguous blocks ( $r^2 = 0.67$ ). More importantly, without refitting, the model predictions match just as well on the two intermediate conditions ( $r^2 = 0.66$ ). That is, the model does not merely fit but quantitatively *predicts* the time course of adaptation to the intermediate adaptors.

<sup>2</sup>The large number of subjects excluded for unusual category boundaries could be attributed either to variability in listening equipment and environment, or to the fact that our listeners are English speakers but the continuum was generated based on Dutch productions of /aba/ and /ada/.

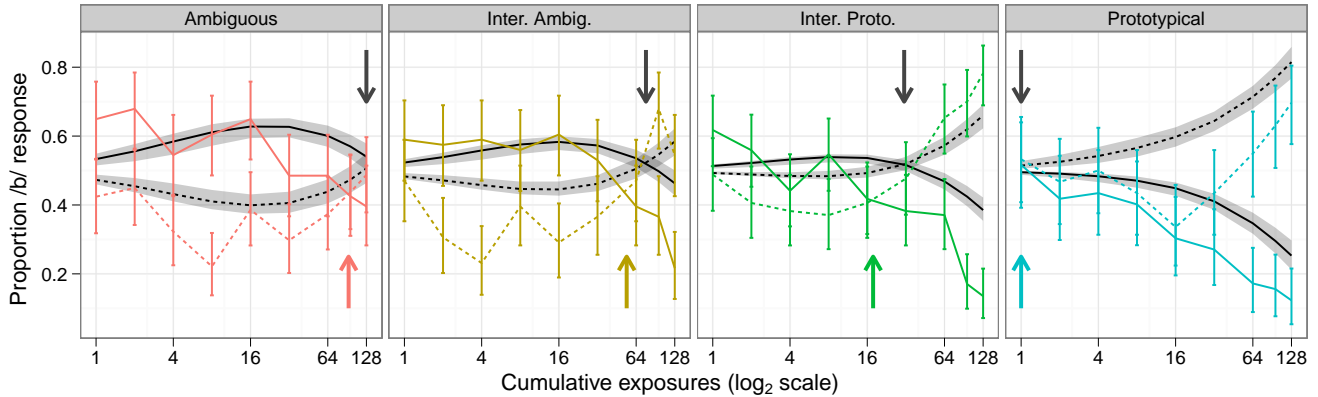


Figure 3: Model predictions and results from each subject’s first block. Left to right shows fully ambiguous, intermediate-ambiguous, intermediate-prototypical, and prototypical. Solid lines are for visual-b exposures and dashed are visual-d. Colored lines are average /b/ response rate, and error bars are 95% confidence intervals on the mean. Black lines are model predictions based on best-fitting parameters to only ambiguous and unambiguous blocks, and shaded regions are 95% confidence intervals based on MCMC sampling of model free parameters with a weakly informative prior.

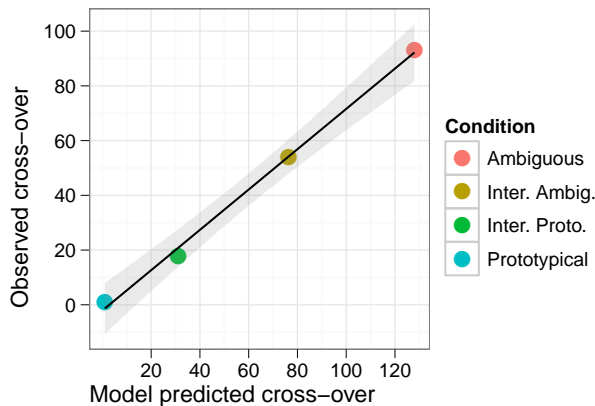


Figure 4: Predicted and actual recalibration-to-selective-adaptation switch points. The x-coordinate of each point is the predicted switch point (black curves/arrows in Figure 3), and the y-coordinate is the observed point (colored curves/arrows).

## Discussion and conclusion

We have proposed that incremental belief-updating can provide a unified account of selective adaptation and phonetic recalibration, and have tested a model based on that principle against existing data on the effect of cumulative exposure on these two effects from Dutch listeners (Kleinschmidt & Jaeger, 2011). In this study we have first replicated this data (Vroomen et al., 2007) with American English listeners, using a novel, web-based platform, and second, extended the original study to test a key prediction of the model that existing data does not address.

It is generally argued that selective adaptation and recal-

ibration are qualitatively distinct effects (e.g., Vroomen et al., 2004). However, our model suggests that they both result from adaptation to changes in the statistics of the linguistic environment, and can be understood as two points on a continuum of phonetic adaptation phenomena. The data we present here shows that intermediate audio-visual adaptors produce adaptation which is intermediate in time course and direction in a way that is quantitatively consistent with the model predictions.

There is, of course, a large body of work which highlights the differences between selective adaptation and recalibration, and this study addresses only one such difference. We cannot, and do not, on the basis of this study claim that selective adaptation and recalibration *are*, in a mechanistic sense, the result of the same process. Rather, our approach offers some insight into these seemingly distinct adaptation phenomena by considering how the two behavioral patterns are related to the statistics of the linguistic environment.

Consider the alternative explanation of our results in terms of separate processes of selective adaptation (a decrease in /b/ responses after /b/ exposure) and recalibration (an increase in /b/ responses after /b/ exposure). Selective adaptation is stronger for more prototypical adaptors (Samuel, 1982; Miller et al., 1983), and so more the more selective adaptation-like time course of adaptation with increasingly prototypical adaptors might be explained simply by an interpolation between the time courses of recalibration and selective adaptation. However, this explanation leaves unanswered the question of why selective adaptation should be involved at all in recalibration, especially considering the fact that selective adaptation has been claimed to only be a function of the *audio* component of an AV stimulus (Saldaña & Rosenblum, 1994) and there is little-to-no selective adaptation observed for boundary stimuli (Miller et al., 1983).

Treating selective adaptation and recalibration as rational reactions to changes in the linguistic environment makes the intuitive idea about interpolating between the two time courses precise and quantitative, but more importantly *why*. As the listener encounters different speakers and different listening conditions, their linguistic environment is constantly changing, often in unpredictable ways. The listener needs to be able to cope with such changes and so must update their beliefs (and behavior) in response to their recent experience. However, while the reasons for such adaptation may be the same, the mechanisms by which it occurs could differ, with recalibration resulting, mechanistically, from higher-level language processing and selective adaptation occurring at a lower, perceptual level (Samuel, 1986).

The major goal of this work is to formulate a learning model for phonetic adaptation and use that model to guide experimental investigation of the whole range of phonetic adaptation phenomena. While this program emphasizes the similarities between different sorts of adaptation, it also aims to uncover even finer-grained empirical distinctions, in part to further the use of adaptation as a window on language processing, generally.

The model we have evaluated here is a first step and makes many simplifying assumptions. Still, it is a Bayesian model which, via data-determined priors and known processes—like sensitivity to category variance and the McGurk effect—provides a unified theoretical explanation for a puzzling interaction between the behavioral effects of selective adaptation and phonetic recalibration. More importantly, it makes fine-grained and testable predictions about the time course of behavior, some of which we have verified here.

Future work will focus on exploring the relationship between this rational model and boundedly rational implementations like exemplar models and particle filters. We will also, via structured, hierarchical extensions of this simple model, focus on resolving the apparent conflict between rapid adaptation to new speakers and the long-term stability of the linguistic system. Finally, and most importantly, we will use the predictions of these models to continue to systematically probe the space of phonetic adaptation behavior in order to inform and constrain the modeling work.

## References

- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707–29.
- Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception & Psychophysics*, 70(4), 604–618.
- Clayards, M. A., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. a. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3), 804–9.
- Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 4(1), 99–109.
- Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, 119(4), 1950–3.
- Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, 116(4), 752–82.
- Kleinschmidt, D., & Jaeger, T. F. (2011). A Bayesian belief updating model of phonetic recalibration and selective adaptation. In *2nd acl workshop on cognitive modeling and computational linguistics*.
- Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: dialects, idiolects, and speech processing. *Cognition*, 107(1), 54–81.
- Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic bulletin & review*, 13(2), 262–8.
- Maye, J., Aslin, R., & Tanenhaus, M. (2008). The Weckud Wetch of the Wast: Lexical Adaptation to a Novel Accent. *Cognitive Science: A Multidisciplinary Journal*, 32(3), 543–562.
- Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3), B101–11.
- Miller, J. L., Connine, C. M., Schermer, T. M., & Kluender, K. R. (1983). A possible auditory basis for internal structure of phonetic categories. *The Journal of the Acoustical Society of America*, 73(6), 2124–33.
- Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47(2), 204–238.
- Saldaña, H. M., & Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *The Journal of the Acoustical Society of America*, 95(6), 3658–61.
- Samuel, A. G. (1982). Phonetic prototypes. *Perception & psychophysics*, 31(4), 307–14.
- Samuel, A. G. (1986). Red herring detectors and speech perception: in defense of selective adaptation. *Cognitive psychology*, 18(4), 452–99.
- Sonderegger, M., & Yu, A. (2010). A rational account of perceptual compensation for coarticulation. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 375–380).
- Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, 36(3), 481–7.
- Vrooomen, J., Linden, S. van, Gelder, B. de, & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia*, 45(3), 572–7.
- Vrooomen, J., Linden, S. van, Keetels, M., Gelder, B. de, & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication*, 44(1-4), 55–61.

# ERP Responses to Violations in Japanese Verb Conjugation Patterns

**Yuki Kobayashi (ykobayashi@ecs.c.u-tokyo.ac.jp)**

Center for Evolutionary Cognitive Sciences, The University of Tokyo  
3-8-1 Komaba, Meguro-ku, Tokyo, JAPAN

**Yoko Sugioka (sugioka@sfc.keio.ac.jp)**

Faculty of Economics, Keio University  
4-1-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa, JAPAN

**Takane Ito (ito@boz.c.u-tokyo.ac.jp)**

Graduate School of Arts and Sciences, The University of Tokyo  
3-8-1 Komaba, Meguro-ku, Tokyo, JAPAN

## Abstract

ERP (Event Related Potential) measurement using the violation paradigm of Japanese verb conjugation was conducted in order to investigate the mental and neural mechanisms involved in the processing of different conjugation patterns. A LAN-like component followed by a P600 was elicited for the anomaly of using a tense-bearing form with the negative ending, while only P600 was observed for the anomaly of using an infinitive form in the same environment. The non-application of morpho-phonological changes of verb roots (“onbin”) yielded an N400 component and a P600. The P600 components observed in all types of errors reflect the cost of processing morphological and/or syntactic anomalies, while the difference in the negativities suggest that two different mechanisms of rule-based computation and lexical memory are involved in the processing of Japanese verb conjugation.

**Keywords:** verb conjugation; inflection; N400; LAN; P600; Dual Mechanism Model; rule; memory

## Introduction

The aim of the present study is to elucidate the mental and neural mechanisms involved in the word-level language processing, exploiting the technique of ERP (Event Related Potential) measurement. More specifically, we investigated the processing of Japanese verb conjugation by recording ERP responses to different types of errors in the conjugation patterns. In doing so, we addressed the question of whether verb inflection involves more than one mechanism of processing in a language typologically different from European languages.

There has been heated debate since 1980’s concerning the mental mechanisms involved in word-level processing, with the focus on inflectional morphology in European languages. On the one hand, Dual Mechanism Model (Pinker 1999, Ullman 2001, among others) argues that two qualitatively different mechanisms, namely, rule-based computation and associative memory, are involved; English “regular” *-ed* past forms, for instance, are dealt with by the former, while “irregular” forms like *sing/sang* are by the latter. On the other hand, Single Mechanism Model (Joanisse and Seidenberg 1999, among others) contends that one and the

same mechanism can deal with both “regular” and “irregular” inflection.

Few studies have been conducted on Japanese in this context: a notable exception is Hagiwara et al. (1999), who argue for the Dual-Mechanism processing in Japanese derivational morphology. The present study places its focus on the processing of Japanese inflectional morphology, which has so far attracted little attention (cf. Vance 1991, Yu et al. 2011).

## ERP Components Related to Language Processing

Three components, N400, LAN, and P600 are known to be related to language processing.

The N400 is a negativity which peaks at around 400 ms after the onset of stimuli with wide, often posterior-centered distribution (Kutas and Hillyard 1980). The component is known to reflect semantic or pragmatic anomaly. It is likely to be related to the search of lexical memory as well, since its amplitude is known to reflect the frequency of the stimulus word (Kutas & Federmeier 2000), and it is observed in word-specific argument structure violation, namely, anomaly with respect to lexical information (Friederici & Frisch 2000, Friederici and Meyer 2004).

The LAN (Left Anterior Negativity) component is also a negativity observed at around 300-500 ms after the onset of stimuli: it is distinguished from the N400 in its distribution, which is limited to the left anterior region. This component is known to reflect morpho-syntactic anomalies like agreement errors (Coulson, King and Kutas 1998, among others).

The P600 is a positive component observed at around 600 ms after the onset of stimuli. Both anterior and posterior distributions have been reported. This component has been claimed to reflect the process of reanalysis or repair in face of morpho-syntactic or syntactic violations of various types (Osterhout and Holcomb 1992 and Hagoort et al. 1993, among others).

ERP studies on regular and irregular inflection have yielded somewhat varied results. The general tendency is that inappropriately attached or omitted regular inflectional suffixes (as in *bringed* instead of *brought*, or in *wip* instead



of *wipped*) tend to elicit a LAN, while modifications of irregular inflection (as in *pept* instead of *peeped*) tend to yield an N400-like component (Newman et al. 2007).

## Conjugation of Japanese verbs

Japanese verb roots can be divided into two types in terms of their conjugation patterns. The verb roots ending with vowels (/e/ or /i/) take various inflectional endings without any phonological change on the root.

(1) vowel-ending root: *tabe* ‘eat’

- a. non-past      *tabe-r-u*
- b. negative      *tabe-nai*<sup>1</sup>
- c. infinitive    *tabe* (-*masu* ‘polite form’, -*owar* ‘finish’)
- d. past          *tabe-ta*
- e. continuative *tabe-te*

The infinitive (ren’yoo) form (1c) takes the polite ending -*masu* and various aspectual verbs. When the endings start with a vowel, a consonant /r/ is inserted (1a).

In contrast, with consonant-ending roots, a vowel /a/ is inserted in the negative form and /i/ in the infinitive form (2b,c). In addition, the root-final consonants go through morpho-phonological changes called “onbin” in traditional Japanese grammar. Past tense ending -*ta* and continuative -*te* takes infinitive forms in both vowel-ending and consonant-ending verbs, but “onbin” takes place in consonant-ending verbs, as shown in (2d,e).

(2) consonant-ending root: *shaber* ‘chat’

- a. non-past      *shaber-u*
- b. negative      *shaber-a-nai*
- c. infinitive    *shaber-i* (-*masu* ‘polite’, -*owar* ‘finish’)
- d. past          *shabe?-ta*
- e. continuative *shabe?-te*

The onbin forms are conditioned by the final consonants of the roots, as summarized below.

(3) morpho-phonological changes (onbin)

- a. r,t,w→? (glottal stop):  
    *shaber* ‘chat’ / *shabe?-ta*; *kat* ‘win’ / *ka?-ta*
- b. k,g→i: *kak* ‘write’ / *kai-ta*; *kag* ‘smell’ / *kai-da*
- c. b,m,n→n: *tob* ‘fly’ / *ton-da*; *yom* ‘read’ / *yon-da*

As illustrated in (3b,c), the initial consonant /t/ of the ending is voiced after /b, m, n, g/. The roots ending with /s/ do not undergo “onbin” (*tas* ‘add’ /*tas-i-ta*).

It should be noted here that these morpho-phonological changes occur only with /t/-initial inflectional endings, but not with /t/-initial derivational suffixes. For instance, an agent nominal suffix -*te* does not trigger similar changes on the root: *kaki-te* ‘writer’ cf. *kai-te* ‘write (continuative)’ (Tagawa 2008). This confirms our assumption that these are not purely phonological changes, but are morpho-phonologically conditioned.

## Stimuli and Predictions

We focused on two aspects of Japanese conjugation. First, various vowels are added to consonant-ending verb roots as shown in (2a)-(2c): /a/ in Neg-form selecting a negative-ending -*nai* (2b), and /i/ in an infinitive form (2c), while the tense-marking morpheme /u/ yields a non-past form (2a), constituting minimal triplets with the same number of morae. We constructed our stimuli by adding the negative ending -*nai* to such triplets, resulting in one well-formed negative conjugation (4a) and two different illicit forms (4b,c).

- (4) a. *shaber-a-nai* (Neg-form + -*nai*)
- b. \**shaber-i-nai* (infinitive form + -*nai*)
- c. \**shaber-u-nai* (non-past form + -*nai*)

Insertion of these vowels to yield the forms (2a-c) is perfectly regular with consonant-ending verb roots, and hence we can hypothesize that these involve rule-based computation.

It should also be noted that the two types of illicit forms have different types of anomaly. (4b) contains a simple morphological ill-formedness, where a wrong non-tensed form (i.e., infinitive instead of Neg-form) is chosen. (4c), on the other hand, involves a phrase-structure violation, where the tense morpheme (non-past -*u*) is added before the negative ending, yielding the ungrammatical phrase structure where Tense is adjoined to V below the NEG node: [[[ *shaber* ]<sub>V</sub> -*u* ]<sub>T</sub> -*nai* ]<sub>NEG</sub> (the correct configuration would be: [[[ ]<sub>V</sub> ]<sub>NEG</sub> ]<sub>T</sub> ).

Consideration of the nature of unacceptability of these forms, together with the nature of ERP components surveyed above, leads us to predict that these illicit forms (4b, c) will elicit computation-related components (LAN and/or P600) when compared to the well-formed counterparts. Also, it can be expected that the two different illicit forms exhibit some difference in ERP responses.

The second aspect we focus on is onbin-forms exemplified in (2d,e). Although onbin-forms are determined by each root-final consonant as described in (3), there are some exceptions: the past form of *ik* ‘go’ is not *ii-ta*, but *i?-ta*, and the past form of *tow* ‘ask’ is not *to?-ta*, but *tow-ta*. It is also reported by Vance (1991) that native speakers experience difficulty in producing the past form of a novel verb. These facts suggest that onbin forms are lexically memorized. We constructed illicit forms by replacing onbin forms with the forms without onbin, namely infinitive forms (root+/i/), as shown in (5b). These forms can be predicted to elicit a memory-related ERP component N400 compared to the well-formed forms (5a).

- (5) a. *shabe?-ta* / *ka-i-ta* / *ton-da*
- b. \**shaber-i-ta* / \**kak-i-ta* / \**tob-i-ta*

## Method

**Participants** A total of 21 (15 males and 6 females) Japanese right-handed undergraduate students at the University of Tokyo participated in the experiment.

<sup>1</sup> The ending -*nai* in (1b) is the non-past form of the negative ending, which we will represent as one element for expository purposes.



**Stimulus Sentences** The target sentences were created on 162 consonant-ending verbs. For 90 verbs, we constructed illicit forms with violation of negative conjugations, and for 72 verbs, we constructed illicit forms with violation of onbin forms (See “Stimuli and Predictions”). Three experimental lists were created according to a Latin square design, so that each list contained 30 sentences in each of the three negative conjugation error conditions and 36 sentences in each of onbin error conditions. A total of 252 sentences, 162 target and 90 filler sentences, were presented. Half of these sentences were well-formed. All sentences used in this experiment had the structure [NP-adjunct-NP-V-X]. An additional element X, nominal + copula or Auxiliary predicate, was added after the verb where we manipulated the conjugation, so that the sentences do not end with the critical word.

#### (6) Negative conjugation error

Zyuumin-wa danti-de otiba-o  
residents-TOP housing.complex-in fallen.leaves-ACC  
(a) moyas-a-nai (Neg-form+ nai)  
burn-NEG  
(b) \*moyas-i-nai (infinitive form+nai)  
(c) \*moyas-u-nai (non-past form+nai)  
kisoku-da.  
rule-COP  
‘Residents are not allowed to burn fallen leaves in the site of the housing complex’

#### (7) Onbin error in past form

Kazoku-wa ima-de syasin-o  
family-TOP living.room-in pictures-ACC  
(a) to?-ta (onbin form +past)  
take-PAST  
(b) \*tor-i-ta (infinitive form+past)  
rasii.  
seem  
‘It seems that the family took pictures in the living room.’

**Procedures** Electroencephalogram (EEG) signals were recorded while the participants read to themselves the stimulus sentences shown automatically on the PC screen phrase by phrase. They were asked to refrain from blinking their eyes or moving their bodies until the end of the sentences. Each sentence had 5 phrases, and each phrase appeared on the screen for 600 ms with a 200 ms blank between each phrase. The critical word, i.e. the verb in all conditions, is the fourth phrase in the stimulus sentences. Following the presentation of a sentence, participants were instructed to make a grammaticality judgment (yes/no decision) by clicking a computer mouse. The 252 sentences were divided into 3 blocks, and within the blocks, sentences were randomized. The participants took a short break between each block.

EEG signals were recorded from 64 Ag/AgCl electrodes mounted in an elastic cap (Quikcap, NeuroScan) according to the International 10–20 system. To control the

participants’ horizontal and vertical eye movements, a bipolar electroencephalogram (EOG) was also recorded using four electrodes. All the EEG and EOG channels were digitized at a 250 Hz sampling rate using a Neuroscan Synamp2s amplifier with a band-pass between DC and 70 Hz. Recordings were referenced to the electrode located between Cz and CPz and then re-referenced offline to the average of the left and right mastoids. Electrode impedance was kept below 10 kOhm. ERP averages were computed with a 100 ms baseline and an 800 ms ERP time window. In the ERP analysis, 9.19% of the trials were rejected because of eye blinks or movement artifacts (the EOG rejection criterion was 70  $\mu$ V).

## Results

### Data analysis

Participants showing high rates of artifacts or error responses to yes/no questions (greater than 20% in the two types of trials combined) were excluded from the analysis. Six participants were excluded and the data from the remaining 15 participants (11 males and 4 females) were analyzed.

Since the visual inspection revealed a negativity with the focus in the left temporal area for (6c) compared with (6a), the following two regions of interest (ROIs) were used in the ANOVAS for the left-lateralized negativity: left-temporal (T7, C5, CP5, and P7) and right-temporal (TP8, C6, CP6, and P8). For statistical analyses of the N400 effect, four regions of interest (ROIs) are derived by crossing two factors; hemisphere (HEMI: left vs. right) and region (REGION: anterior vs. posterior). The ROIs are defined as follows: left-anterior (F5, F3, F1, FC5, FC1, C5, and C1); left-posterior (CP5, CP1, P5, P1, PO7, PO5, and PO3); right-anterior (F2, F4, F6, FC2, FC6, C2, and C6); and right-posterior (CP6, CP2, P6, P2, PO8, PO6, and PO4).

For statistical analyses of the P600, four regions of interest (ROIs) are derived by crossing two factors; hemisphere (HEMI: left vs. right) and region (REGION: anterior vs. posterior). The ROIs are defined as follows: left-anterior (F5, F3, F1, FC5, FC3, FC1, C5, C3, and C1); left-posterior (CP5, CP3, CP1, P5, P3, P1, PO7, PO5, and PO3); right-anterior (F2, F4, F6, FC2, FC4, FC6, C2, C4, and C6); and right-posterior (CP6, CP4, CP2, P6, P4, P2, PO8, PO6, and PO4).

### Negative conjugation error

Figure 1 shows the grand average ERPs for the critical verb, in correct, infinitive, and non-past form conditions.

**correct vs. illicit (infinitive+nai)** As shown in Figure 1, a positivity around 500–800 ms after the onset was elicited by the infinitive form, in comparison with the correct form. The latency and the distribution suggest that it is a P600 component, which is supported by the statistical analysis as follows: The analyses for the time window 500–800 ms revealed a significant main effect of condition,

$F(1,14)=21.98, p<.001$ . The infinitive form condition was more positive going than the correct. No other ERP components besides the P600 were observed in the comparison of the correct form and the infinitive form conditions ( $ps>.05$ ).

**correct vs. illicit (non-past+nai)** The non-past form condition, when compared with the correct form condition, elicited a negativity in the left temporal region at the time range of 300-400 ms and a positivity around 500-800 ms. The analyses for the time window 300-400 ms revealed a significant interaction between HEMI and the correct form vs. the non-past form,  $F(1,14)=4.54, p<.05$ . The non-past form condition was more negative going than the correct form condition in the left-temporal sites ( $F(1,14)=3.76, p=.066$ ), but not in the right-temporal sites. The latency and the distribution of the positivity suggest that it is a P600 component, which is supported by the statistical analysis as follows: The analyses for the time window 500-800 ms revealed a significant main effect of condition,  $F(1,14)=25.80, p<.001$ . The non-past form condition was more positive going than the correct form.

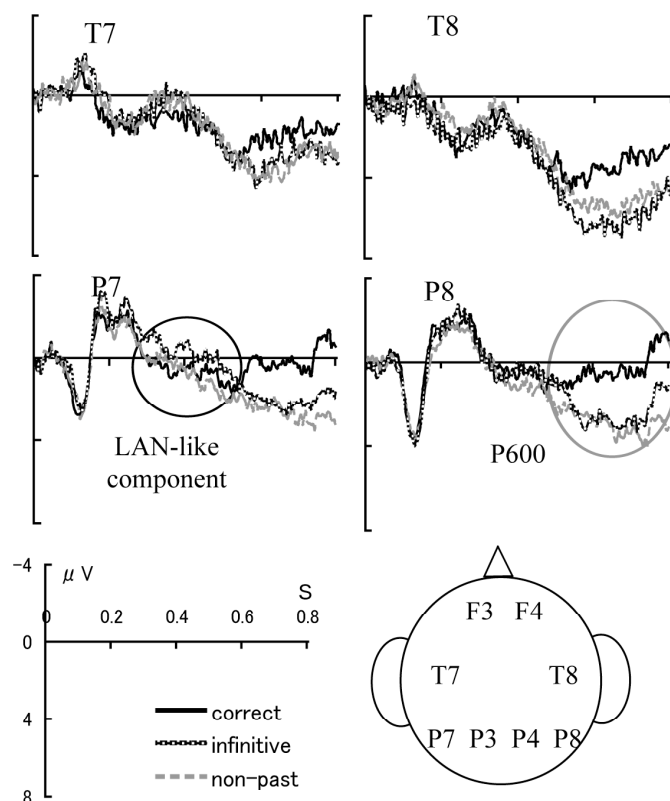


Figure 1: Grand average ERPs at selected electrodes at the position of the target verb (onset at the vertical bar) for the correct form vs. the infinitive form vs. the non-past form. Negativity is plotted upwards.

## Onbin error

Figure 2 shows the grand average ERPs for the critical verb, in the onbin and the infinitive form conditions. The illicit infinitive form condition, when compared with the correct onbin form condition, elicited a negativity at the time range of 200-500 ms with a focus in posterior sites, and a positivity around 600-800 ms. The latency and the distribution suggest that the negativity is an N400 component, which is supported by the statistical analysis as follows: The analyses for the time window 200-500 ms revealed a marginally significant interaction between REGION and the onbin form vs. the infinitive form,  $F(1,14)=3.47, p=.08$  and a significant main effect of condition,  $F(1,14)=5.17, p<.05$ . The infinitive form condition was more negative going than the onbin form condition in the posterior sites ( $F(1,14)=8.36, p<.01$ ), but not in the anterior sites. The latency and the distribution of the positivity suggest that it is a P600 component, which is supported by the statistical analysis as follows: The analyses for the time window 600-800 ms revealed a significant main effect of condition,  $F(1,14)=35.36, p<.001$ . The infinitive form condition was more positive going than the onbin.

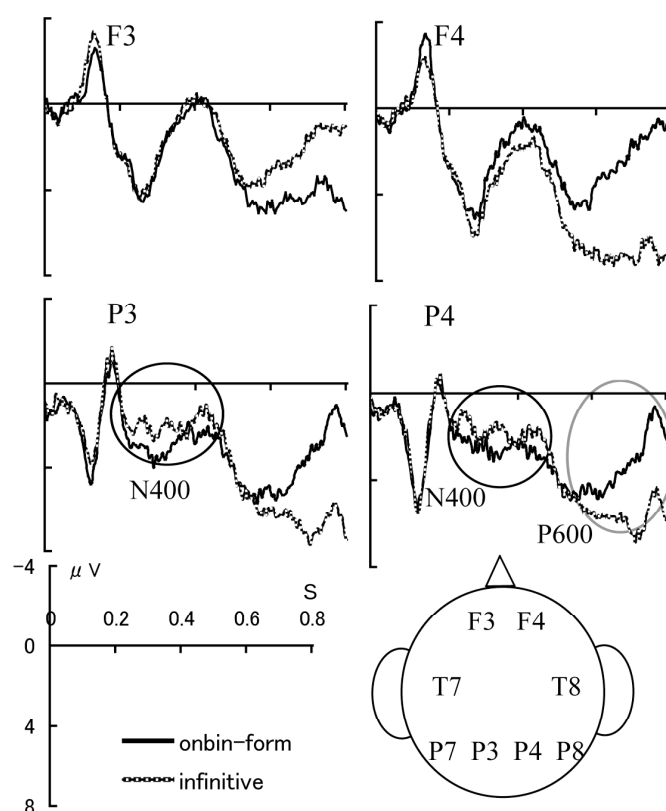


Figure 2: Grand average ERPs at selected electrodes at the position of the target verb (onset at the vertical bar) for the onbin-form vs. the infinitive form. Negativity is plotted upwards.

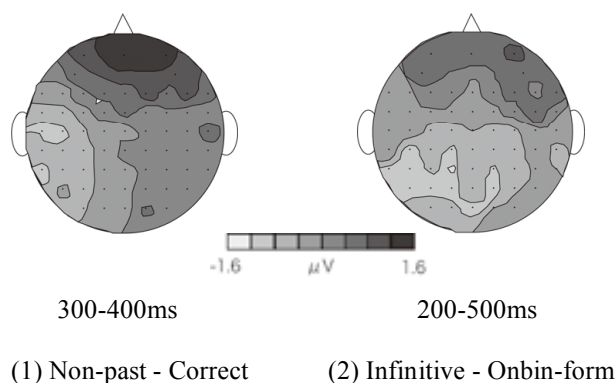


Figure 3. The topographical potential maps display the distribution of the negativities, (1) for the illicit non-past form as compared to the correct form and (2) for the illicit infinitive form as compared to the correct onbin form in the time window used for the statistical analysis. Lighter shading indicates more negative potential differences.

## Discussion

The above results are compatible with what our hypotheses predict: both conditions in negative conjugation errors elicited a P600, while a negative component was observed only in the non-past+*nai* condition, but not in the infinitive+*nai*. The onbin errors elicited an N400 component followed by a P600. We will discuss these results, focusing on what each component can be interpreted to reflect.

### Negative components

First, the negative component observed in negative conjugation errors (6c) and the one observed in onbin errors (7b) were clearly different in their distribution, and hence can be judged to be different components. Secondly, the negativity in negative conjugation errors was observed only for the non-past+*nai* condition, and not for the infinitive+*nai* condition.

The negativity elicited by the illicit infinitive form without onbin (7b), although not so robust in its amplitude, was judged to be an N400, given its distribution with centro-parietal focus and no hemispheric lateralization (Figure 3 (2)). This suggests that detecting this anomaly involved accessing of lexical memory. As has been discussed in the literature (e.g. Kutas and Federmeier 2000), difficulty in lexical access is one of the factors reflected in N400. Thus, this result is in accordance with our hypothesis that onbin does not involve computation by rule but the onbin forms for each verb root are memorized in the lexicon, hence application or non-application of onbin for a specific verb root with the past tense ending *-ta* must be checked against the lexicon.

The negativity elicited by the illicit non-past forms (6c) in comparison with the well-formed control (6a) has clearly different distribution from the N400 component, in that it was limited to the left hemisphere, as supported by the significant difference between the two hemispheres (Figure 3

(1)). Even though its distribution is also different from the classic LAN component in that its focus is in the temporal sites, it is similar in distribution to a left-lateralized negative component reported as a LAN by Rodriguez-Fornells et al. (2001) for Catalan overregularization of stem formation rule. Thus, it is not implausible to consider the negativity observed in our non-past condition as a LAN-like component. This component can be interpreted as reflecting the parser's detection of a phrase-structure violation of (6c), where the tense morpheme is placed below the Neg node. Thus the illicit non-past forms exhibited the biphasic pattern of the LAN-like negativity followed by a P600, which is in accordance with the literature reporting a LAN-P600 pattern for phrase structure violations (Friederici and Meyer 2004).

It is consistent with our prediction that there was a difference concerning negative components between the illicit infinitive forms (6b) and the illicit non-past forms (6c) in negative conjugation errors. It calls for some explanation, however, that a LAN-like component was not observed in (6b). In previous studies on morpho-syntactic anomalies, anterior negative components (LAN or AN) followed by a P600 are reported for agreement errors or case violations, which require syntactic computation of subject-verb or verb-object relation (e.g., Osterhout and Mobley 1995, Coulson et al. 1998). These errors are similar in nature to our non-past condition (6c). In contrast, our infinitive condition (6b) involves a purely morphological error, as mentioned above in "Stimuli and Predictions". Thus, even though our hypothesis holds that (6b) as well as (6c) involves rule-based computation, the violation in (6b) is, in a sense, much simpler, and hence it is conceivable that the cost of detecting the ill-formedness is too weak to elicit a statistically significant LAN-like component. In this vein, it is significant that the illicit infinitive forms did not elicit an N400 component, which supports our contention that the detection of the anomaly in (6b) does not involve lexical memory.

### Positive components

We observed late positive components, which can be judged to be a P600, in all the three illicit conditions (infinitive+*nai* (6b), non-past+*nai* (6c), and infinitive+*ta* (7b)) when compared to the well-formed counterparts. As discussed above, we contend that different mechanisms are involved in the processing of the negative conjugation and onbin forms. And yet, a P600 was observed across all three illicit conditions, which suggests that this component reflects the cost of dealing with conjugation errors, irrespective of their nature, namely, whether purely morphological (6b), morpho-syntactic (6c), or morpho-phonological (7b), and whether rule-based (6b,c) or memory-based (7b).

## Concluding Remarks

These findings taken together suggest the following points on the processing of Japanese verb conjugation. First, conjugation of Japanese roots with a specific vowel for each ending involves rule-based computation. On the other hand,

the morpho-phonological change (onbin) that takes place on the final consonant of some subclasses of the verb roots in the environment of certain ending forms such as past and continuative requires lexical memory. It can be thus concluded that Japanese verb conjugation involves two different mental mechanisms, rule-based computation and lexical memory, depending on the type of conjugation and particular set of endings. Hence our findings are consistent with the Dual Mechanism Model.

As mentioned in “ERP Components Related to Language Processing”, the ERP responses to regular/irregular inflection so far reported are rather varied, especially on negative components. While a LAN is observed for over-application of regular morphology in studies on English and German for instance (Morris and Holcomb 2005, Penke et al. 1997), it is also reported in Morris and Holcomb that the over-regularized irregulars (*bringed*) elicited an N400 when presented in a word-list format (in contrast to a LAN observed in a sentential context). Similarly, Gross et al. (1998) reports an N400-like negativity for overregularized Italian irregulars in a word-list format, where difference between stem-based inflection and affixal inflection is suggested to be relevant. In other words, a number of factors including methods of stimuli presentation and differences in inflectional systems between languages seem relevant, and hence more investigation is obviously needed. Our study has shown that different inflectional processes within one language can involve different mental mechanisms and thus induce different ERP components in the violation paradigm, indicating the importance of studying languages typologically different from European languages.

### Acknowledgements

The research reported here is supported in part by Grant-in-Aid for Scientific Research (B) #20320069 from the Japan Society for the Promotion of Science, the Center for Evolutionary Cognitive Sciences at the University of Tokyo, and the Japanese Lexicon Project at National Institute for Japanese Language and Linguistics.

### References

- Coulson, S., J. W. King, & M. Kutas (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes*, 13, 21-58.
- Friederici, A.D., & Frisch, S. (2000). Verb argument structure processing: The role of verb-specific and argument-specific information. *Journal of Memory and Language*, 43, 476-507.
- Friederici, A. D., & Meyer, M. (2004). The brain knows the difference: Two types of grammatical violations. *Brain Research*, 1000, 72-77.
- Gross, M., Say, T., Kleingers, M., Clahsen, H., & Münte, T. F. (1998). Human brain potentials to violations in morphologically complex Italian words. *Neuroscience Letters*, 241, 83-86.
- Hagiwara, H., Sugioka, Y., Ito, T., Kawamura, M., & Shiota, J. (1999). Neurolinguistic evidence for rule-based nominal suffixation. *Language*, 75, 739-763.
- Hagoort, P., Brown, C. & Groothusen, J. (1993). The syntactic positive shift as an ERP measure of syntactic processing. *Language and Cognitive Processes*, 8, 439-484.
- Joanisse, M. F. & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences* 96: 7592-7597.
- Kutas, M., & Federmeier, K.D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, 4, 463-470.
- Kutas, M., & Hillyard, S.A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207, 203-205.
- Morris, J. & Holcomb, P.J. (2005). ERPs to violations of inflectional verb morphology. *Cognitive Brain Research*, 25, 963-981.
- Newman, A.J., Pancheva, R., Waligura, D.L., & Neville, H.J., and Ullman, M.T. (2007). An ERP study of regular and irregular past tense inflection. *NeuroImage*, 34, 435-445.
- Osterhout, L., & Holcomb, P.J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31, 785-806.
- Osterhout, L., & Mobley, L.A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language*, 34, 739-773.
- Penke, M., Weyerts, H., Gross, M., Zander, E., Münte, T. F., & Clahsen, H. (1997). How the brain processes complex words: An event-related potential study of German verb inflections. *Cognitive Brain Research*, 6, 37-52.
- Pinker, S. (1999). *Words and Rules: The ingredients of language*. New York: Harper Perennial.
- Rodriguez-Fornells, A., Clahsen, H., Lleó, C., Zaake, W., & Münte, T. F. (2001). Event-related brain responses to morphological violations in Catalan. *Cognitive Brain Research*, 11, 47-58.
- Tagawa, T. (2008). Bunsan-keitairon niyoru doosi no katuyoo to gokeisei no kenkyuu [A study on verb conjugation and word formation by Distributed Morphology]. Doctoral dissertation submitted to Tsukuba University.
- Ullman, M.T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience*, 2, 717-726.
- Vance, T. (1991). A new experimental study of Japanese verbal morphology, *Journal of Japanese Linguistics*, 13, 145-156.
- Yu, Q., Deng, Y., & Sakai, H. (2011). Processing Japanese verb morphology by native Japanese speakers: An ERP study. IEICE Technical Report TL 2011-14, 37-42.

# Event Segmentation of Agent Interactions: Comparing the Whole with Its Parts

Bryan L. Koenig<sup>1,2</sup> (koenigbl@ihpc.a-star.edu.sg)

David Pautler<sup>1</sup> (pautlerd@ihpc.a-star.edu.sg)

Jonathan S. Herberg<sup>1</sup> (herbergjs@ihpc.a-star.edu.sg)

Kum Seong Wan<sup>1</sup> (kswan@ihpc.a-star.edu.sg)

Brian Monroe<sup>1,2</sup> (monroebm@ihpc.a-star.edu.sg)

Edwin Wirawan<sup>1</sup> (wirawane@ihpc.a-star.edu.sg)

<sup>1</sup>Institute of High Performance Computing, A\*STAR

<sup>2</sup>National University of Singapore, Singapore

## Abstract

How do observers perceptually organize the events of individual agents when observing interactions among them? Do they readily perceive all events? Do they selectively perceive some events but not others? Do they see events overlooked by observers focusing on the individual agents? To explore these questions, participants viewed the Heider and Simmel (1944) animation, which shows three moving figures and elicits strong impressions of interacting agents. Participants in the default condition segmented the animation into meaningful events. Those in focus conditions did likewise, but focusing on one of the figures. Results indicate that participants in the default condition disregarded many events identified in the focus conditions, but identified only one event missed by focus-condition participants. These findings suggest that observers of interactions do not encode all events or gain additional insight by “seeing the big picture”; rather, they selectively perceive some events at the cost of overlooking others.

**Keywords:** Social perception; event segmentation; unit formation; perspective taking; movement cues; animation.

## Introduction

Much research has centered on how observers perceive and understand the actions of individual agents. In a typical experimental task, participants view an activity such as washing dishes and mark when, in their judgment, a meaningful event ends and another begins. These marked time points are referred to as *breakpoints*, and participants substantially agree about their placement (e.g., Zacks, Tversky, & Iyer, 2001). Breakpoint placement is not an arbitrary consequence of the segmentation task. Brain activity selectively occurs during mere observation at the same time points that participants identify as breakpoints during a subsequent segmentation task (Zacks, Swallow, Vettel, & McAvoy, 2006). These findings suggest that event segmentation reflects the perceptual structure of events.

Other research on how observers perceptually segment the events of animations of *multiple* interacting figures

finds that motion cues are strongly correlated with breakpoints (Hard, Tversky, & Lang, 2006). Indeed, simple computational models can use seven motion cues to identify agentic motives in animations of two moving figures with accuracy levels similar to those of human observers (Blythe, Todd, & Miller, 1999). Absolute and relational motion cues can be quite complicated in such stimuli, which sometimes have three figures moving in relation to one another and the background context. We wondered how human observers incorporate information from each agent when viewing stimuli of multiple agents interacting. In order to address this question, we had participants in a *default condition* segment into meaningful events a multiple-figure animation previously shown to elicit compelling perceptions of social interactions (Heider & Simmel, 1944). Other participants in *focus conditions* segmented the same animation but only for events meaningful for the figure specific to the condition. We compared participants’ segmentation patterns across conditions to test whether observers by default process all events relevant for all agents. Alternatively, perhaps by default observers systematically miss events that are important for some agents or notice events that would be overlooked when focusing on any one agent. We now consider evidence that suggests each of these is plausible.

Research in which participants segment perceived events for animations of moving figures suggests that movement features such as sudden acceleration elicit the perception of event boundaries (Hard, et al., 2006; Zacks, 2004). For such stimuli, participants even indicate breakpoints at similar time points when viewing multiple-agent animations both forward and backward, and these time points strongly correlate with objective movement features (Hard, et al., 2006). Such findings provide support for the notion that observers incorporate all movement cues when perceiving events, which we label the *objective movement hypothesis*. According to this hypothesis, observers perceive ongoing interactions in terms of all movement

cues from all figures. This would lead to the prediction that participants in the default condition should generate a segmentation that includes all, and no more, of the breakpoints identified across all focus conditions.

Zacks and Tversky (2001) note that for both event segmentation and object perception observers must organize parts into wholes. Heider (1958) suggested that an analogous process – unit formation – occurs during social perception wherein multiple agents, such as those interacting, are perceived as a perceptual unit. If observers perceptually group agents, the action of one agent (e.g., hitting) might be subsumed as a lower-level component of a larger schematic interaction (e.g., fighting), and thereby be disregarded. The *unit formation hypothesis* thus postulates that when observers parse ongoing events in terms of interactions they sometimes neglect agents' individual actions. This hypothesis predicts that participants in the default condition, when viewing a highly salient interaction, will sometimes miss events identified by participants in the focus conditions when the events pertain to a figure that *is part of* that interaction.

Heider (1958) also suggested that unit formation might result in observers attending to the interaction unit as the focal point of perception (the “figure”), with agents outside the interaction receiving little attention (the “ground”). The *distraction hypothesis* therefore postulates that observers attending to an interaction will sometimes fail to notice events meaningful to agents who are not part of the interaction. This hypothesis predicts that participants in the default condition, when viewing a highly salient interaction, will sometimes miss events identified in the focus conditions when the events pertain to a figure that *is not part of* that interaction.

Other research has shown that perspective-taking can interfere with an objective evaluation of events. For example, in one study people from each side of a conflict perceived a media report of an event as biased against their own side (Vallone, Ross, & Lepper, 1985). Similarly, fans of each team in an important football game, when viewing identical tapes of the game, reported events that diverged from each other (Hastorf & Cantril, 1954). This research is complemented by experiments suggesting that adults sometimes have difficulty taking another's perspective even though all of the relevant information is available to them (Keysar, Lin, & Barr, 2003). To our knowledge, perspective-taking has not been evaluated using animated figures. Nonetheless, we embody these ideas in the *perspective-taking hypothesis*, which postulates that focusing on the perspective of one figure interferes with the overall perception of events. It seems plausible that events noted by participants in the default condition but not in any focus condition could be the result of difficulty

participants experience as they attempt to focus on or take the perspective of one figure. This hypothesis predicts that events perceived by participants in the default condition sometimes will be missed by participants in all of the focus conditions.

## The Current Study

We designed a study to test these four hypotheses. In it participants segmented the Heider and Simmel animation (1944) into meaningful events. The animation shows a house-like rectangle and three moving figures: a large triangle, a small triangle, and a circle (*T*, *t*, & *c*). Observers tend to describe the animation as a bully attacking two innocent passersby (Heider & Simmel, 1944). By comparing the default segmentation with those provided for focal agents, we evaluate how observers normally view the animation: whether perception of the whole is equal to, greater than, or less than the sum of its parts.

## Method

**Participants** Participants were 74 female and 51 male American workers on Amazon Mechanical Turk, a crowdsourcing marketplace service, who received US\$ 0.90 for participating. Their mean age was 32.50 years (*SD* = 11.47).

**Stimulus Animations** For the experimental task we modified a version of the Heider and Simmel (1944) animation downloaded on 3 September 2010 from Carnegie Mellon University at [http://anthropomorphism.org/img/Heider\\_Flash.swf](http://anthropomorphism.org/img/Heider_Flash.swf). We increased its frames per second rate from 10 to 30 to make it smoother, keeping it 74 seconds long. For the practice task we created a Flash version of the hide-and-seek animation used in Hard, et al. (2006). It is 84 seconds long and depicts two squares and a circle moving as if they were playing hide and seek in their environment, which has wall-like lines. Both animations were prepared in Adobe Flash and were monochrome, 550 pixels wide by 400 pixels high, but viewed dimensions depended on participants' monitor displays.

**Design** Participants were randomly assigned to one of four conditions, which differed only in their segmentation instructions. In the *default condition* (*n* = 39) participants segmented the animation with no instructions to focus on any specific figure. Thus, their segmentations were potentially based on events for all of the figures. In each of the three focus conditions, the instructions were to segment the animation with a focus on the events for a specific figure, that is, for the big triangle (*n* = 30), the little triangle (*n* = 20), or the circle (*n* = 29).

**Procedure** Participants completed the study over the internet. They provided consent and demographic information and then read instructions indicating that (a) they would see two short animations depicting geometric figures in motion, (b) while watching the animation they should press the spacebar whenever one meaningful event ended and another began, and (c) they would briefly describe each animation after watching it. The instructions also displayed pictures of the figures. Participants then watched and concurrently segmented the practice animation (i.e., hide and seek). Once it ended, participants described what happened in each one-second interval bin for which they had pressed the spacebar. While providing descriptions, participants could review the animation but they could not add or remove markers. The instructions clarified that participants should try to provide their initial impression of the animation (at the time of the spacebar press). Once participants submitted descriptions for each marker, they could continue to the experimental task.

The procedure for the experimental task was the same as that for the practice task, except for different segmentation instructions in the focus conditions. All participants viewed the Heider and Simmel (1944) animation and instructions that indicated that the animation has three moving geometric figures, a large triangle, a small triangle, and a circle, with pictures of all figures. In the default condition, the segmentation instructions were the same as for the practice animation. The circle-focus condition included the following additional instructions: “However, this time do so only for the circle. That is, press the spacebar whenever, for the circle, a meaningful event ends and another begins.” Similar instructions were added to the focus conditions for the big triangle and the small triangle. A blue arrow also pointed at the picture of the focal figure. Participants segmented the animation according to their condition’s instructions, provided descriptions as in the practice task, logged their MTurk ID into the system, and were debriefed.

## Results

If a participant pressed the spacebar during a one-second bin we counted that bin as containing a breakpoint (for similar approaches see, e.g., Hard, et al., 2006; Massad, Hubbard, & Newtonson, 1979). Since only *T* is visible at the start and end of the experiment animation, all results exclude the first 4 and last 12 bins, leaving 58 bins for analysis. We also excluded the data of five participants who indicated one or zero breakpoints while doing the experimental task and that of two who indicated no breakpoints in the practice task.

**Overall Segmentation** We evaluated whether focusing on one figure affected the number of one-second bins participants marked as containing a breakpoint. The objective movement hypothesis predicts that the default

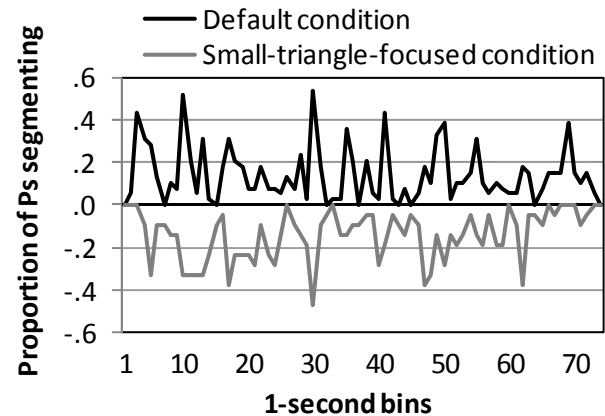


Figure 1. Proportions of participants who selected each one-second interval bin as containing a breakpoint in the default and *t*-focused conditions. The proportions for the *t*-focused condition are shown as negative to facilitate visual comparison.

condition should contain the breakpoints present in all three focus conditions. If this is the case, participants in the default condition should indicate more breakpoints than participants in the focus conditions (assuming segmentation patterns differed across focus conditions, and correlations shown below indicate they did). A one-way between-subjects ANOVA on the number of breakpoints did not indicate a difference across conditions ( $M = 8.96$ ,  $SD = 5.11$ ),  $F(3, 114) = 1.05$ ,  $p = .373$ ,  $\eta_p^2 = .03$ . This lack of a difference cannot be explained by a ceiling effect; participants on average indicated only 15.45% of bins as including breakpoints. Instead, these results suggest, counter to the objective movement hypothesis, that participants in the default condition did not segment based on the objective movements of all three figures. Rather, these results suggest that participants in the default condition may have neglected some events noticed in the focus conditions, consistent with the unit formation and distraction hypotheses. Because focus-condition participants may have noticed some events missed in the default condition, but missed others noticed in the default condition, these results provide no clear evidence regarding the perspective-taking hypothesis.

**Agreement about Event Timing** To gain further insight into how participants perceptually integrate information of individual agents when perceiving them interact, we compared across conditions which one-second interval bins participants tended to indicate as containing breakpoints. For these analyses, we calculated for each 58 one-second bin the proportion of participants in each condition who indicated that the bin contained a breakpoint. Figure 1 presents the segmentation histogram for the default condition and, for comparison, the *t*-focused condition. We correlated the segmentation histogram of the default



Table 1. Correlation coefficients across conditions among segmentation histograms (proportions of participants who indicated a breakpoint in each one-second bin).

Condition	Default	Big T.	Small T.	Circle
Default	1	.61*	.48*	.53*
Big T.	-	1	.41*	.22
Small T.	-	-	1	.38*

\*  $p < .01$ . Note: T. = triangle.

condition with that of each focus condition (see Table 1). All three Pearson correlation coefficients were positive and significant; however, all fell short of the lower bound of the 95% confidence interval of the estimated correlation coefficient for participants within the default condition (estimated using bootstrap aggregation, i.e., sampling 2000 groups of  $n = 20$  from the 39 default condition participants then calculating 58 bin means for each group and pairing groups to calculate 1000 correlation coefficients across bins; mean  $r = .78$ , median  $r = .79$ , 95% CI: .62, .90). This range was estimated to specify the approximate optimal correlation that we could expect between the default histogram and that of any other condition. The finding that the focus condition histograms were less than optimally correlated with the default condition histogram suggests that our manipulation was successful in that participants in the focus conditions were not simply responding to events involving any figure, but instead when focusing on one figure they perceived events differently.

The differences in magnitude across the three correlation coefficients were not significant (maximum  $z = 1.14$ ,  $p = .254$ , using the method suggested by Meng, Rosenthal, & Rubin, 1992), but their relative magnitudes suggest that default-condition participants' segmentation may have been influenced most by  $T$ , then  $c$ , and the least by  $t$ . To further evaluate this possibility, we did a multiple regression analysis wherein we predicted the segmentation histogram from the default condition by the segmentation histograms from all three focus conditions. The results suggest that  $T$  provided a large unique contribution to the perception of the animation,  $b = .47$ ,  $p < .001$ , 95% CI = .27, .67,  $c$  had a medium sized unique contribution,  $b = .31$ ,  $p < .001$ , 95% CI = .15, .47, but  $t$  did not contribute any information above and beyond that provided by  $T$  and  $c$ ,  $b = .16$ ,  $p = .194$ , 95% CI = -.08, .40 (the intercept did not differ significantly from zero,  $b = -.01$ ,  $p = .746$ , 95% CI = -.05, .04). This regression's multiple correlation coefficient (multiple- $R = .74$ ,  $F(3,57) = 22.10$ ,  $p < .001$ ) fell within the 95% CI for the estimated correlation coefficient within the default condition, suggesting that this regression performed near optimally. This suggests that participants in the focus conditions indicated many if not most of the breakpoints identified in the default condition; that is, these analyses provided no support for the prediction of the perspective-

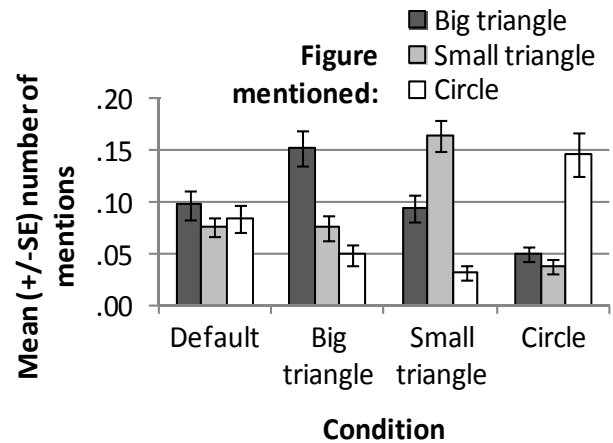


Figure 2. Mean (+/- SE) number of times each participant mentioned the figures, by condition.

taking hypothesis that events would be missed if focusing on a specific figure. We note that exploratory regression analyses including all two-way interactions and the three-way interaction indicated no additional significant predictors,  $ps > .226$ .

**Analyses of Breakpoint Descriptions** We coded which figures participants mentioned in their breakpoint descriptions. As Figure 2 shows, participants in the focus conditions mentioned their focal figure significantly more than participants in any other condition, minimum  $F(3,114) > 10.84$ ,  $p < .001$ ,  $\eta_p^2 = .22$ . This suggests that our manipulation was successful. Participants in the default condition did not mention the three figures an equal number of times,  $F(2, 76) = 8.32$ ,  $p = .001$ ,  $\eta_p^2 = .18$ . Uncorrected post hoc tests indicated that default-condition participants mentioned  $T$  more than  $t$ ,  $t(38) = 5.86$ ,  $p < .001$ ; they mentioned  $T$  marginally more than  $c$ ,  $t(38) = 2.02$ ,  $p = .051$ ; but they mentioned  $c$  and  $t$  about equally,  $t(38) = 0.16$ ,  $p = .124$ . These findings suggest that  $T$  was the most salient figure for the default-condition participants.

**Disagreement across Conditions** We also calculated for each bin a difference score equal to the proportion of participants in the default condition identifying the bin as containing a breakpoint minus the maximum such proportion among the three focus conditions (see Figure 3). Most strikingly, nearly all difference scores were negative. Indeed, a binomial test ( $p = .005$ ) indicated fewer positive scores than expected by chance given a .25 probability that each of the four conditions would have the largest proportion. Our earlier analysis indicated no significant difference in the number of breakpoints across conditions, whereas the current finding indicates that, when comparing bin-by-bin, for almost all bins participants in one of the focus conditions were more likely to indicate a breakpoint

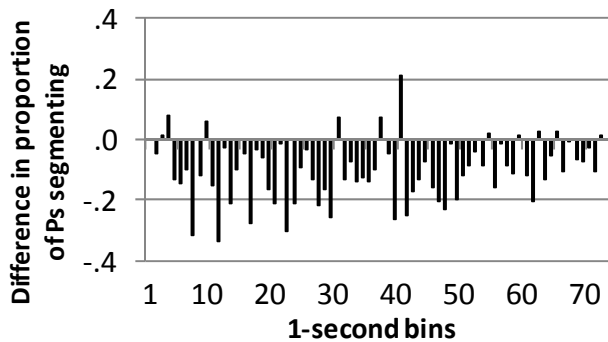


Figure 3. Difference in proportion in the number of participants indicating a breakpoint for each one-second interval bin in the default condition with the maximum proportion doing so from any of the focus conditions.

than participants in the default condition. This suggests that for observers in the default condition some events had reduced salience as compared to observers who focused on a particular figure, consistent with the unit formation and distraction hypotheses. By contrast, almost all events that are salient in the default condition remain so in the focus condition for at least one of the figures, contradicting the perspective-taking hypothesis. Indeed, setting a .2 difference in proportions as a cutoff to identify clear-cut differences in event salience, only one bin has a more salient event in the default condition as compared to any of the focus conditions. By contrast, 15 bins had events that were more salient in at least one of the focus conditions as compared to the default condition (see Table 2). Corroborating our previous findings that for default-condition participants the figures were of unequal salience, the three figures were not equally likely to be part of a reduced-salience event: *T* was in 3 such events, *c* in 5, and *t* in 8, multinomial test,  $p = .017$ .

Why did these 15 events have reduced salience for participants in the default condition? To determine the likely source of the reduced salience for each of these events we examined which figures were mentioned in participants' event descriptions across conditions. For convenience, we use the term "overlooked figure" to refer to the figure whose event had reduced salience in the default condition. Let us first consider which figures are mentioned in descriptions from the focus condition for the overlooked figure. If they predominantly mentioned the overlooked figure and one other figure, this suggests that the overlooked figure appeared to interact with the other figure and therefore that the reduced salience was due to unit formation. If these descriptions instead predominantly mentioned only the overlooked figure, but descriptions from the other focus conditions did not mention the overlooked figure but did mention the other two figures, this suggests that the other figures had a salient interaction that distracted default participants from noticing the

Table 2. Reduced-salience events, that is, events for which the discrepancy in the proportion of participants who indicated a breakpoint was greater by at least .2 in any focus condition (boldface) as compared to the default condition. The table reports the proportions by condition, the coded explanation for the difference (source: dist. = distraction, unit. = unit formation), and the ongoing overall event.

Bin	Proportion P. by condition				Source	Ongoing overall event
	Def.	<i>T</i>	<i>t</i>	<i>c</i>		
8	.10	<b>.42</b>	.14	.03	dist.	<i>t</i> & <i>c</i> arrive
12	.05	<b>.39</b>	<b>.33</b>	.07	unit.	<i>t</i> & <i>T</i> start fight
14	.03	.10	<b>.24</b>	.03	unit.	<i>t</i> & <i>T</i> fight
17	.31	.39	.38	<b>.59</b>	dist.	<i>t</i> & <i>T</i> fight
21	.08	.13	<b>.29</b>	.08	unit.	<i>t</i> & <i>T</i> fight
23	.08	.16	.24	<b>.38</b>	dist.	<i>t</i> & <i>T</i> fight
24	.08	.16	<b>.29</b>	.10	unit.	<i>t</i> & <i>T</i> fight
28	.23	.10	.14	<b>.45</b>	dist.	<i>t</i> & <i>T</i> fight
30	.54	.23	.48	<b>.79</b>	dist.	<i>t</i> & <i>T</i> fight
40	.03	.13	<b>.29</b>	.00	dist.	<i>c</i> & <i>T</i> fight
42	.03	.13	.05	<b>.28</b>	unit.	<i>c</i> & <i>T</i> fight
47	.18	.19	<b>.38</b>	.31	both	<i>t</i> joins <i>c</i> & <i>T</i>
48	.10	.19	<b>.33</b>	.21	both	<i>t</i> joins <i>c</i> & <i>T</i>
50	.38	<b>.58</b>	.29	.28	unit.	<i>T</i> fights <i>c</i> & <i>t</i>
62	.18	.06	<b>.38</b>	.28	unit.	<i>t</i> & <i>c</i> leave

Note: Def. = Default condition

overlooked figure. However, if descriptions from the focus condition for the overlooked figure predominantly mentioned the overlooked figure, but the other conditions had very few descriptions, we looked to recent bins for clarification. If descriptions from recent bins across conditions suggested a two-figure interaction, we coded the reduced salience as due to unit formation if the overlooked figure was part of that interaction or as due to distraction otherwise. Finally, two reduced-salience events had characteristics suggesting both unit formation and distraction. That is, descriptions from the default condition and the focus condition for the overlooked figure predominantly mentioned the overlooked figure and one other figure, suggesting the overlooked figure was interacting with the other figure and was thereby perceived as a unit with it. On the other hand, descriptions from the other two focus conditions predominantly mentioned both the other two figures but not the overlooked figure, suggesting their interaction distracted default condition participants from seeing the overlooked figure. Using these criteria, from 58 bins we clearly associated 7 with unit formation, 6 with distraction, and 2 with characteristics of both. These findings provide fairly direct support for the unit formation hypothesis and the distraction hypothesis.

## General Discussion

We investigated how observers perceptually organize the events that are meaningful for individual agents while observing them interacting with one another. Our results support the notion that observers selectively perceive some events and not others. Moreover, our results suggest that this sometimes occurs because observers link individual agents into larger perceptual units with the consequence that the salience of events at the unit-level sometimes dominates the salience of events for agents within the unit. At other times observers appear to have missed events important for one agent because their attention was focused on more salient interactions between other agents. These findings provide some empirical support for Heider's (1958) idea of unit formation. They also replicate the findings of Massad, et al. (1979) that observers selectively perceive some events and disregard others, although in their study selective perception resulted from pre-information about what would happen in the Heider and Simmel animation. Previous research has noted the importance of movement features for event segmentation, but our findings suggest that observers do not normally perceive interaction events in terms of all movement features for all agents. This finding suggests a caveat to the idea that observers use all motion information. Future research will be required to more fully determine when and why observers fail to incorporate some motion features. Our results suggest that participants were able to focus on one figure as instructed, and that doing so resulted in little difficulty due to perspective-taking. Indeed, to the extent to which participants in the focus conditions engaged in perspective-taking, our findings suggest this is not always a difficult process (cf. Keysar, et al., 2003), but can occur in a rather effortless fashion. In fact, the results indicate a greater cost, in terms of missed events, of engaging in the default rather than an agent-focused perspective.

The current study compared how observers segment events when focusing holistically on all agents to how they do so when focusing on individual agents. A future study could more completely separate information regarding single agents from the whole by removing all other figures from the animation, leaving a single, isolated figure. We could then compare segmentations of the animated solo figures with those of default and focus conditions. This would allow us to evaluate the relative importance for perceptual segmentation of context-free movement cues versus relational movement cues. Such a comparison might provide a more direct test of the unit formation hypothesis (Heider, 1958). We also note that participants in the default condition may have perceived events that they failed to report, but why they would do so more than participants in the focus conditions is unclear. We leave such problems

and further questions about the perception of interactions to future research.

## Acknowledgments

We thank Barbara Tversky and Bridgette Hard for providing their hide and seek animation. We thank Barbara Tversky and three anonymous reviewers for insightful comments on earlier drafts of this paper. We also thank Swati Gupta and William Chandra Tjhi for assistance with statistical analyses.

## References

- Blythe, P. W., Todd, P. M., & Miller, G. F. (1999). How motion reveals intention: Categorizing social interactions. In G. Gigerenzer, P. M. Todd, and the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 257–286). New York: Oxford University Press.
- Hard, B. M., Tversky, B. & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory & Cognition*, 34, 1221–1235.
- Hastorf, A. H., & Cantril, H. (1954). They saw a game: A case study. *The Journal of Abnormal Psychology*, 49, 129–134.
- Heider, F. (1958). *The psychology of interpersonal relations*. Wiley: New York.
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, 57, 243–249.
- Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind usage in adults. *Cognition*, 89, 25–41.
- Massad, C. M., Hubbard, M., & Newton, D. (1979). Selective perception of events. *Journal of Experimental Social Psychology*, 15, 513–532.
- Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, 111, 172–175.
- Vallone, R. P., Ross, L., & Lepper, M. R. (1985). The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of Personality and Social Psychology*, 49, 577–585.
- Zacks, J. M. (2004). Using movement and intention to understand simple events. *Cognitive Science*, 28, 979–1008.
- Zacks, J. M., Swallow, K. M., Vettel, J. M., & McAvoy, M. P. (2006). Visual motion and the neural correlates of event perception. *Brain Research*, 1076, 150–62.
- Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, 127, 3–21.

# Thinking in Patterns: using multi-voxel pattern analyses to find neural correlates of moral judgment in neurotypical and ASD populations

Jorie Koster-Hale<sup>1</sup>, James Dungan<sup>2</sup>, Rebecca Saxe<sup>1</sup>, Liane Young<sup>2</sup>

<sup>1</sup>Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge MA 02139

<sup>2</sup>Department of Psychology, Boston College, Chestnut Hill, MA 02467

## Abstract

Intentional harms are typically judged to be less forgivable than accidental harms. This difference depends on mental state reasoning (i.e., reasoning about beliefs and intentions), supported by a group of brain regions, the ‘theory of mind’ network. Prior research has found that (i) interfering with activity in this network can shift moral judgments away from reliance on mental state information, and (ii) high-functioning individuals with Autism Spectrum Disorder (ASD) rely significantly less on mental state information to make moral judgments than matched neurotypical (NT) participants. Across three experiments, we find using multi-voxel pattern analysis (MVPA) that, in NT adults, (i) one key region of the ToM network, the RTPJ, shows reliable and distinct spatial patterns of responses across voxels for intentional versus accidental harms, and (ii) individual differences in this neural pattern predict individual differences in moral judgment. By contrast, (iii) in ASD adults, the difference between intentional and accidental harms is not encoded in the voxel-wise pattern in the RTPJ or any other region, and (iv) higher symptom severity scores are predictive of diminished pattern discriminability. We conclude that MVPA can detect features of mental state representations and that these features are behaviorally and clinically relevant.

**Keywords:** morality, harms, theory of mind, autism, fMRI, multivoxel pattern analysis (MVPA)

## Introduction

Intentional harms are usually judged to be morally worse than the same harms caused by accident (Cushman, 2008; Knobe, 2005; Malle & Knobe, 1997; Piaget, 1965; Singer et al, 2004; Young & Saxe, 2011). The capacity to distinguish between intentional and accidental harms depends on the capacity to represent another person’s mental states, a cognitive function associated with a specific and selective group of brain regions (the ‘theory of mind network’). Prior research has revealed that moral judgments of harmful actions depend on one region in particular, the right temporo-parietal junction (RTPJ). For example, individual differences in moral judgments of accidental harms are correlated with RTPJ activity (Young & Saxe, 2009), and disrupting RTPJ activity interferes with these judgments (Young et al, 2010).

Recent evidence suggests that moral judgments may provide a sensitive measure of enduring impairments in ToM in high-functioning individuals with Autism Spectrum Disorders (ASD). Children with ASD are disproportionately impaired on tasks that require them to consider the beliefs and intentions of other people (Peterson et al, 2005; Baron-Cohen, 1995). Although children with ASD distinguish between moral and conventional transgressions (Blair et al, 1996), and between good actions and bad actions (Leslie et al, 2006), they are delayed in using information about innocent

intentions to forgive accidents (Grant et al, 2005). In a recent study, Moran et al. (2011) found that high-functioning adults with ASD show the same pattern, delivering less forgiveness and more blame for accidents than neurotypical (NT) adults.

These findings suggest that the RTPJ should encode the difference between accidental and intentional harm. Puzzlingly, however, we find that the average mean signal – a standard way of measuring neural involvement in a task – in RTPJ and the other theory of mind regions is not different for intentional versus accidental harmful actions in NT participants. One possibility is that this key dimension can be detected, not in the magnitude of response across a brain region, but in the pattern of responses across voxels.

A complementary approach to traditional neuroimaging analyses (which rely on average response magnitude) is to look at the pattern of activity across voxels within a region, using a technique called multi-voxel pattern analysis. If a different tasks, stimulus categories, or stimulus features are processed by (partially) different subpopulations of neurons within a brain region, the difference may not be detectable in the region’s average response, but may nevertheless produce systematic and distinct patterns of activity across neighboring voxels within the region (Normal et al, 2006; Haynes et al, 2006; Kriegeskorte & Bandetti, 2007). A key advantage of this technique is that these patterns can be used to ‘decode’ information from the neural response not otherwise detectable in the overall magnitude (e.g., object category in ventral temporal regions, Haxby et al 2001). Thus MVPA can reveal how stimulus categories are processed within a functional region (Peelen et al, 2006; Haynes & Rees, 2006).

Given the importance of intentions for moral judgments of accidental and intentional harms, we predicted that one or more brain regions in the ToM network would explicitly encode this feature of others’ mental states, in neurotypical (NT) adults. That is, we predicted that (i) while participants read about a wide range of harmful acts, we would be able to decode whether the described harm was intentional or accidental based on the spatial pattern of activity within ToM brain regions. We tested this prediction in three experiments. We also investigated (ii) whether the robustness of the spatial pattern information within individuals would predict those individuals’ moral judgments, and (iii) whether high-functioning adults with ASD, who make atypical moral judgments of accidental harms, would show correspondingly atypical patterns of neural activity.

## Methods

### Participants

**Experiment 1:** Sixteen right-handed members of the MIT community (aged 18-50, 7 women). **Experiment 2:** Eighteen

right-handed college undergraduate students (aged 18-25 years, 8 women). **Experiment 3:** Fourteen right-handed college undergraduate students (aged 18-25 years, 8 women).

**Experiment 4:** Twelve individuals diagnosed with Autism Spectrum Disorder (aged 25-43 years, 2 women). Participants were recruited via advertisements placed with the Asperger's Association of New England. All participants were prescreened using the Autism Quotient questionnaire (AQ; Baron-Cohen et al., 2001). ASD participants then underwent both the Autism Diagnostic Observation Schedule (ADOS) (Lord et al 2000, 2002) and impression by a clinician trained in both ADOS administration and diagnosis of ASD. All ASD participants received a diagnosis of an Autism Spectrum Disorder based on their social ADOS score ( $6.2 \pm 0.6$ ), communication ADOS score ( $3.5 \pm 0.4$ ), and total ADOS score ( $9.6 \pm 0.8$ ) and on clinical impression based upon the diagnostic criteria of the DSM-IV (APA, 2000). The matched NT (Exp. 1) and ASD (Exp. 4) groups did not differ in age (NT (mean  $\pm$  SEM)= $27.1 \pm 2.3$ ; ASD= $31.8 \pm 2.1$ ;  $t(28)=1.4$ ,  $p > 0.17$ ) or IQ [NT:  $118.1 \pm 2.8$ ; ASD:  $121.0 \pm 3.8$ ;  $t(27)=0.60$ ,  $p > 0.55$ ]. No participant had higher language than social scores, suggesting no specific language deficits.

All participants participated for payment, were native English speakers, had normal or corrected-to-normal vision and gave written informed consent in accordance with the requirements of Institutional Review Board at MIT. Data from Experiments 2 and 3 have previously been published, analyzing the magnitude but not the pattern of response in each region, in Young et al. (2008) and Young & Saxe (2009).

### fMRI Protocol and Task

**Experiment 1 & 4:** Participants were scanned while reading 60 stories: 12 intentional harm violations, 12 accidental harm violations, 24 stories with other types of moral violations, and 12 neutral scenarios. Stories were presented in the second person, using present tense, and displayed in four cumulative segments: 1. Background (6s), 2. Action (4s), 3. Outcome (4s), 4. Intent: Good (Accidental Harm) or Bad (Intentional Harm) (4s). In the scanner, after each story, participants made moral judgments of the action from "not at all morally wrong" (1) to "very morally wrong" (4), using a button press. For example stimuli, see [http://mit.edu/jorie/www/CogSci2012/Koster-Hale\\_CogSci2012\\_supp.pdf](http://mit.edu/jorie/www/CogSci2012/Koster-Hale_CogSci2012_supp.pdf).

Stories were presented in a pseudorandom order; condition order was counterbalanced across runs and subjects, and no condition was immediately repeated. Participants never saw both intentional and accidental versions of the same scenario. Word count was matched across conditions. Ten stories were presented in each 5.5 min run; the total experiment, six runs, lasted 33.2 min. Rest blocks of 10 s were interleaved between each story. Stories were projected onto a screen via Matlab 5.0 running on an Apple MacBook Pro in 40-point white font.

**Experiments 2 & 3:** Participants were scanned while reading 48 stories. Experiment 2 included 12 intentional harms, 12 accidental harms, and 24 non-harm stories. All harms were physical harms, resulting in someone's death. Stories were presented in cumulative segments: 1. Background (6s) 2.

Foreshadow (6s, only in Experiment 2), 3. Intent: Good (Accidental Harm) or Bad (Intentional Harm) (6 s), 4. Outcome (6s). Half of the stories in each run were presented with foreshadow before intent; the order was reversed in the other half. After each story, participants made moral judgments of the action on a 3-point scale, from "forbidden" (1) to "permissible" (3), using a button press.

Experiment 3 included 8 intentional harms, 8 accidental harms, and 32 other non-harm stories. Participants delivered a non-moral judgment, answering a true/false question about the content of the final sentence.

Stories were counter balanced and matched as in Experiment 1. Rest blocks of 14 s were interleaved between each story. Stories were projected onto a screen via Matlab 5.0 running on an Apple G4 laptop in 24-point white font.

**Theory of Mind Localizer task:** In all four experiments, participants also saw 4 runs of a theory of mind localizer task, contrasting stories requiring inferences about mental state representations (e.g., thoughts, beliefs) versus physical representations (e.g., maps, signs, photographs), which are similar in their meta-representational and logical complexity but differ in whether the reader is building a representation of someone else's mental state. See Saxe & Kanwisher (2003) and Dodell-Feder et al (2010) for further discussion; stimuli and presentation from Saxe & Kanwisher 2003, Exp. 2.

### Acquisition and Preprocessing

fMRI data were collected in a 3T Siemens scanner at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, using a 12-channel head coil. Using standard echoplanar imaging procedures, we acquired blood oxygen level dependent (BOLD) data in 26 near axial slices using  $3 \times 3 \times 4$  mm voxels (TR=2 s, TE=40 ms, flip angle=90°). To allow for steady state magnetization, the first 4 seconds of each run were excluded.

Data processing and analysis were performed using SPM8 (Experiments 1 & 4) and SPM2 (Experiments 2 & 3) and custom software. The data were motion corrected, realigned, normalized onto a common brain space (MNI template), spatially smoothed using a Gaussian filter (FWHM 5 mm kernel) and subjected to a high-pass filter (128 Hz).

### fMRI Analysis

All fMRI data were modeled using a boxcar regressor, convolved with a standard hemodynamic response function (HRF). The general linear model was used to analyze the BOLD data from each subject, as a function of condition. The model included nuisance covariates for run effects, global mean signal, and an intercept term. A slow event-related design was used. An event was defined as a single story, the event onset was defined by the onset of text on screen, and offset as the end of the story presentation.

**Functional Localizer: Individual ROIs** Based on prior research, functional regions of interest (ROIs) were defined in right and left temporo-parietal junction (RTPJ, LTPJ), medial precuneus (PC), and dorsal medial prefrontal cortex (DMPFC), for each participant. Using a pre-defined

hypothesis space for each ROI, each subject's contrast image (Belief > Photo) was masked and the peak voxel that occurred in a cluster of 10 or more voxels significant at  $p < 0.001$  was selected. All voxels contiguous with the peak voxel, individually significant at  $p < 0.001$ , within a 9mm radius, were defined as the ROI.

**Group ROIs:** To compare NT participants from Experiment 1 and ASD participants from Experiment 4, we also identified independent group-level ROIs, using data from a previous set of theory of mind localizers ( $n=477$  NT participants, 260 women). Pattern analyses for both groups were then conducted within these group ROIs. Selecting the same voxels across participants and group ensured that any differences found between the NT and ASD group are due to differences in the pattern itself, rather than any differences in the ROI selection method, across populations.

**Within-ROI Magnitude Analysis:** We measured the response to each condition in each ROI. The percent signal change (PSC) relative to baseline was calculated for each time point in each condition, averaging across all voxels in the ROI and across all blocks in the condition, where  $PSC(t) = 100 \times (\text{average BOLD magnitude for condition } (t) - \text{average BOLD magnitude for fixation}) / \text{average BOLD magnitude for fixation}$ . We averaged the PSC across the entire presentation – offset 6s from presentation time to account for hemodynamic lag – to get a single PSC for each condition, in each ROI, in each participant (Poldrack, 2006).

**Within-ROI Pattern Analysis:** In all experiments, we conducted within-ROI pattern analyses. Following Haxby et al. (2001), each participant's data were divided into even and odd runs ('partitions') and then the mean response (beta value) of every voxel in the ROI was calculated for each condition. The "pattern" of response was the vector of beta values across voxels within the ROI. To determine the within-condition correlation, the pattern in one (e.g., even) partition was compared to the pattern for the same condition in the opposite (e.g., odd) partition; to determine the across-condition correlations the pattern was compared to the opposite condition, across partitions.

For each individual, an index of classification was calculated for each condition pair as the z-scored within-condition correlation minus the z-scored across-condition correlation. A region successfully classified a category of stimuli if, across individuals, the within-condition correlation was higher than the across-condition correlation, using a Student's T complementary cumulative distribution function.

## Results

### Localizer

Replicating many studies using a similar functional localizer task (e.g., Saxe & Kanwisher, 2003), we localized four theory of mind brain regions showing greater activation for false belief stories compared to false photograph stories in the majority of participants (uncorrected,  $p < 0.001$ ,  $k > 10$ ): **Exp 1-3 (NT):** RTPJ 46/48, LTPJ, 44/48, PC 47/48, DMPFC

41/68; **Exp 4 (ASD):** RTPJ (12/12 participants), LTPJ (12/12), PC (11/12) and DMPFC (5/12).

### Behavioral Results

**Experiment 1:** Participants judged intentional harms ( $3.31 \pm 0.10$ ) to be worse than accidental harms ( $1.62 \pm 0.11$ ;  $t(14) = 15.1$ ,  $p < 0.0001$ ), both of which were judged to be worse than neutral stories ( $1.03 \pm .02$ ,  $t(14) = 18.1$ ,  $p < 0.001$ ).

**Experiment 2:** Replicating the results in Experiment 1, participants judged intentional harms ( $2.9 \pm .03$ ) to be worse than accidental harms ( $1.9 \pm .11$ ;  $t(12) = 8.24$ ,  $p < 0.0001$ ).

**Experiment 4:** Behavioral data were available for only 7 participants with ASD (remaining data were lost due to a coding error and due to theft of experimental equipment). When making moral judgments, ASD participants, like NT participants from Experiment 1, ASD participants judged intentional harms ( $3.5 \pm 0.12$ ) to be worse than accidental harms ( $1.85 \pm 0.21$ ;  $t(6) = 8.9$ ;  $p < 0.0001$ ), both of which were judged to be worse than neutral stories ( $1.09 \pm .04$ ;  $t(6) = 13.9$ ;  $p < 0.0001$ ).

**Group Comparison:** A mixed effects ANOVA crossing Group (NT in Exp 1, ASD in Exp 4) by Condition (neutral, accidental, intentional) yielded a main effect of condition and no interaction ( $F(2,36) = 284.4$ ,  $p < 0.0001$ ). Post-hoc t-tests revealed that ASD adults assign more blame for accidental harms than NT adults ( $t(19) = 1.7$ ,  $p < 0.05$ ), but there was no difference between NT and ASD judgments of intentional harms (Moran et al., 2011).

### fMRI - Magnitude Analysis

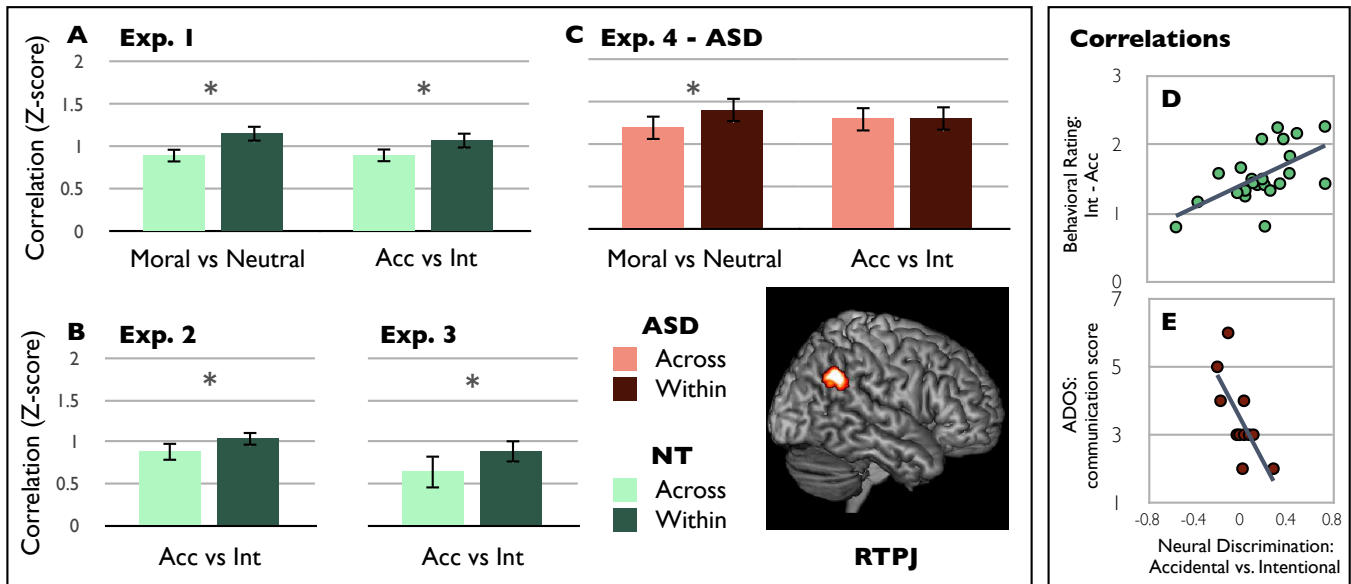
**Experiment 1-4** In all three experiments with NT adults, and in the final experiment with ASD adults, all four ROIs (RTPJ, LTPJ, PC, DMPFC) showed a higher BOLD response for moral violations than for neutral acts. However, none of the regions showed a significant difference between accidental and intentional harms (all  $p > .2$ ).

### fMRI - Pattern Analysis

**Experiment 1 Moral vs. Neutral:** Multi-voxel pattern analyses revealed reliably distinct patterns of neural activity for moral violations versus neutral acts in all four ROIs: RTPJ, LTPJ, PC, DMPFC. The pattern generated by stories within a category (moral or neutral) were more correlated with other stories in the same category compared to stories in the opposite category (RTPJ: across condition correlation = 0.90 (.1), within condition correlation = 1.16(.1),  $t(15) = 2.6$ ,  $p = 0.02$ ; LTPJ: across = 1.5(.07), within = 1.6(.06),  $t(14) = 2.3$ ,  $p = 0.019$ ; PC: across = 0.86(0.13) within = 1.2(0.12),  $t(14) = 3.1$ ,  $p = 0.005$ ; DMPFC: across = 1.1(0.13), within = 1.2(0.11),  $t(12) = 1.9$ ,  $p = 0.04$ ). Group ROIs yielded the same results.

**Accidental vs. Intentional:** In only the RTPJ, the pattern of response distinguished between accidental and intentional harms (across condition correlation = 0.91(.1), within condition correlation = 1.08(.11),  $t(15) = 2.6$ ,  $p = 0.01$ ). No other regions showed distinct patterns of response to intentional versus accidental harms (all correlation differences  $< 0.1$ , all  $p > 0.1$ ). Group ROIs yielded the same results, (**Figure 1-A**).





**Figure 1:** (A) MVPA results from Experiment 1: neurotypical adults ( $n=16$ ) show pattern discrimination for moral vs. neutral actions, and for accidental vs. intentional harms in RTPJ, (B) a finding replicated across Experiments 2 and 3 ( $n=18,14$ ). (C) Adults with ASD ( $n=12$ ) show neutral discrimination of moral vs. neutral actions in RTPJ, but not of intentional vs. accidental harms. (D) Individual differences in pattern classification predict individual differences in behavior in NT adults ( $n=23$ ). (E) In ASD adults ( $n=12$ ), symptom severity is negatively correlated with neural discrimination of accidental and intentional harms.

**Experiments 2 & 3:** Experiments 2 and 3 replicate Experiment 1. In RTPJ, but no other region, MVPA analyses revealed that accidental harm and intentional harm elicited reliably distinct neural patterns (Exp 2 RTPJ: across condition correlation=0.76(.13), within=1.1(.11),  $t(15)=2.6$ ,  $p=0.01$ ; Exp 3 RTPJ: across=0.65(.21), within=0.89(.14),  $t(13)=1.8$ ,  $p=0.04$ ; all other regions: correlation differences  $<0.1$ ,  $p>0.1$ ).

**Combining Experiments 1, 2, & 3:** Pooling the data across all three experiments allowed us to increase our power to detect results in neural regions beyond RTPJ. Again, MVPA analyses revealed distinct neural patterns for accidental and intentional harms in RTPJ (RTPJ: across=0.78(0.09), within=1.0(0.08),  $t(45)=3.9$ ,  $p=0.0002$ ) and no other region (all differences  $<0.1$ ,  $p>0.1$ , Figure 2); repeated measures ANOVAs crossing Region by Pattern yielded a significant interaction between RTPJ and each of the three other regions (all  $F > 8$ , all  $p < 0.007$ ), and no interactions between any of the other regions (however, see Smith et al., 2011 for caution in interpreting differences in discriminability across regions).

**Behavioral and Neural Correlation:** In Experiments 1 and 2, NT participants provided moral judgments of each scenario in the scanner, allowing us to determine whether behavioral responses were related to the spatial pattern of the neural response in RTPJ or any other region. For each participant, we calculated the difference between moral judgments for intentional versus accidental harms. We tested whether this difference score was correlated, across participants, with the index of classification in each region (intentional vs. accidental, within-condition correlation minus across-condition correlation). In both experiments we found that only in the RTPJ the difference between intentional and accidental harms in individuals' moral judgments correlated

with the neural classification index (Exp 1:  $r^2(12)=0.38$ ;  $p=0.03$ ; Exp 2:  $r^2(11)=0.40$ ;  $p=0.03$ ). The correlation was also significant after combining the data from both experiments ( $r^2(23)=0.35$ ,  $p=0.003$ ; **Figure 1-D**).

**Experiment 4 (ASD) Moral vs. Neutral:** As in NT controls, pattern analyses revealed a separation in the pattern of response for moral violations and neutral scenarios in ASD adults. Using individual ROIs, significant discrimination was found in RTPJ and LTPJ (RTPJ: across=1.2(0.18), within=1.4(0.16),  $t(11)=3.2$ ,  $p=0.005$ ; LTPJ: across=1.3(0.12), within=1.5(0.11),  $t(11)=2.3$ ,  $p=0.02$ ; see Figure 1); there was a trend in the same direction in the PC. DMPFC was found in only 5 of 12 individuals so we did not have sufficient power to test for pattern discriminability. However, using group ROIs, we found significant discrimination in DMPFC as well (across=1(0.12), within=1.2(0.10),  $t(11)=3$ ,  $p=0.007$ ).

**Accidental vs. Intentional:** Pattern analysis within both group and individual ROIs revealed no pattern discrimination between accidental and intentional harms in any ToM region, in participants with ASD (all differences  $<0.1$ ,  $p>0.1$ ).

**Group Comparison Moral vs. Neutral:** A Group (ASD, NT)  $\times$  Pattern (within, across) ANOVA revealed that NT and ASD participants show equally robust neural discrimination in response to moral violations versus neutral actions in their RTPJ, with a main effect of Pattern ( $F(2,51)=12.4$ ,  $p=.002$ ), no effect of Group ( $F(1,51)=1.07$ ,  $p=.3$ ), and no interaction ( $F(2,51)=.13$ ,  $p=.7$ ).

**Accidental vs. Intentional:** In contrast, NT participants discriminated between accidental and intentional harms to a greater extent than ASD participants, reflected in a significant



Group x Pattern interaction ( $F(2,51) = 5.1, p = .03$ ), and no main effects (**Figure 1-C**).

**Symptom Severity and Neural Correlation:** In ASD participants, we found no significant correlation between neural pattern and behavior in any region. However, we found, in RTPJ and no other region, a significant inverse correlation with symptom severity: individuals with higher ADOS (Lord et al, 2000, 2002) scores showed less neural discriminability between intentional and accidental harms ( $r^2(12)=0.51, p=0.01$ , **Figure 1-E**). No symptom severity score correlated with moral versus neutral discrimination.

## Discussion and Conclusion

### Moral Judgments: Neurotypical adults

A central aim of this study was to ask whether the difference between accidental and intentional harms could be decoded from the pattern of response within theory of mind brain regions. Across three experiments with neurotypical (NT) adults, using different stimuli, paradigms, and participants, we found converging results: stories about intentional versus accidental harms elicited spatially distinct patterns of response within the right temporo-parietal junction (RTPJ). Moreover, this neural response mirrored behavioral judgments: individuals who showed more distinct patterns in the RTPJ also made a larger distinction between intentional and accidental harms in their moral judgments.

The convergence across experiments provides strong evidence that intentional and accidental harms can be discriminated, using MVPA, in RTPJ. Designed to test a series of separate questions, the three experiments differed in the story content, voice of the narrative (2nd or 3rd person), the order of information provided, the length of the stories, the number of stories per condition, and the participants' explicit task. Perhaps most importantly, the information indicating that the harmful action was accidental or intentional was provided by different cues. In Experiment 1, the same mental state content (e.g., your cousin's allergy to peanuts) was described as known or unknown (e.g., "you had no idea" vs. "you definitely knew"). By contrast, in Experiments 2 and 3, sentences with the same syntax and mental state verbs were used to describe beliefs with different content (e.g., "Steve believes the ground beef is safe / rotten"). Nevertheless, the spatial pattern of response was reliable and distinct for intentional versus accidental harms, and only in the right TPJ. The generalizability of the pattern discriminability indicates that, rather than being driven by specific stimulus features or task demands, the discriminable neural patterns reflect an underlying distinction in the representation of accidental and intentional harms.

In Experiments 1 and 2, participants made moral judgments in the scanner. In both experiments, individuals differed in the amount of blame they assign to accidental harm, some weighing intent more strongly (and thus were more forgiving) and some weighing outcome more strongly (and thus were more condemning), (Young & Saxe, 2009). These individual differences in moral judgment were predicted by individual

differences in pattern discriminability in the RTPJ. While Experiment 1 used a blameworthiness scale ("How much blame should you get?") and Experiment 2 used a permissibility scale ("How permissible was Steve's action?"), we found the same result in both studies: individuals who showed more sensitivity to the dimension of intent in their neural pattern – those who processed accidental and intentional harms most differently in their RTPJ – were also those who showed the most forgiveness to characters who accidentally harmed someone.

As in prior work, the average magnitude of RTPJ response did not distinguish between intentional and accidental harms. This observation fits in a larger pattern emerging in the literature: while the RTPJ is selective for the cognitive process of mental state reasoning – and not, for example, generic attentional processes (Scholz et al., 2008; Young et al., 2010b; see also Decety and Lamm, 2007) – the average RTPJ response is unaffected by changes in the specific features of mental states, such as whether beliefs are true or false (Jenkins and Mitchell, 2009; Young, et al, 2010b), justified or unjustified (Young et al., 2010c), positively or negatively valenced (Kliemann et al., 2008), plausible or crazy (Young et al., 2010b), "constrained" or "open-ended" (Jenkins and Mitchell, 2009), attributed to friends or enemies (Bruneau & Saxe in press), or first-order or higher-order (Koster-Hale & Saxe, 2011).

These findings left open the question of whether and how the RTPJ or any other neural substrate encoded specific mental state features, like the dimension of intent. A key contribution of the current study then is to reveal that the dimension of intent is encoded in the voxel-wise pattern of the RTPJ, and specifically for the evaluation of harm.

### Moral Judgments: ASD adults

A group of high functioning adults with Autism Spectrum Disorders showed a different response profile: the response of the RTPJ showed a reliable spatial pattern of response across moral stories, compared to neutral stories, but did not distinguish between intentional and accidental harms. Moreover, we found that symptom severity on the ADOS was predictive of decreased neural discrimination in RTPJ: those individuals with more severe diagnoses showed a less distinct neural response to accidental and intentional harms. Thus, the neural pattern mirrored the behavioral performance previously observed in participants with ASD (Moran et al 2011): compared to NT controls, participants with ASD judged accidental and intentional harms to be more similar in moral permissibility (though note that, in the current sample, the group by condition interaction was not replicated, likely due to lack of power).

One possible mechanism of reduced pattern information in ASD might be more noisy or heterogeneous neural responses. However, both the strong discrimination between moral violations and neutral stories, and the high overall pattern correlations speak against this alternative. Rather ASD participants seem to show a less sensitive neural response: accidental and intentional harms appear to be processed by the same neural sub-populations within the RTPJ.

Multivoxel pattern analysis may therefore be a successful way of measuring behaviorally-relevant neural differences in ASD. Note that due to the demands of the task and scanning environment, the ASD participants in this study (as in previous task-oriented neuroimaging studies) are extremely high functioning, which may limit the generalizability of the results to lower-functioning individuals. Nevertheless, the individuals in the current study do have disproportionate difficulties with social interaction and communication, and the current results may provide a window on the neural mechanism underlying these difficulties.

## Conclusion

In summary, MVPA allows us to determine (i) that features of mental state representations that are not observable in the mean neural signal, including the behaviorally relevant difference between accidental and intentional harms, are encoded in the *pattern* of neural activity; (ii) that these mental state features elicit both stable and distinct patterns of neural activity in RTPJ, a region implicated in mental state reasoning; (iii) that individual differences in neural discrimination predict individual differences in moral judgment; and (iv) that atypical behavioral patterns in ASD are reflected in atypical neural patterns, which (v) are more atypical with increasing symptom severity.

## Acknowledgments

This material is based upon work supported by the Simons Foundation, the National Science Foundation under Grant 095518, a John Merck Scholars Grant, and a National Science Foundation Graduate Research Fellowship, Grant 0645960.

## References

- Baron-Cohen S. (1995) Mindblindness: an essay on autism & theory of mind. MIT Press, Cambridge MA.
- Blair J. (1996) Brief Report: Morality in the Autistic Child. *Journal of Autism & Developmental Disorders*, 26(5):571-579.
- Castelli F, Frith C, Happé F, & Frith U. (2002) Autism, Asperger syndrome & brain mechanisms for the attribution of mental states to animated shapes. *Brain*, 125:1839-49.
- Cushman, F. (2008). Crime & Punishment: Distinguishing the roles of causal & intentional analysis in moral judgment. *Cognition*.
- Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *The Neuroscientist*, 13, 580-593.
- Dodell-Feder, D., Koster-Hale, J, Bedny M, & Saxe, RR. (2011), fMRI item analysis in a theory of mind task, *NeuroImage*.
- Grant CM, Boucher J, Riggs KJ, & Grayson A. (2005). Moral understanding in children with autism. *Autism* 9(3):317-331.
- Lord C, Rutter M, DiLavore PC, Risi S (2002) Autism Diagnostic Observation Schedule (Western Psychological Services, Los Angeles)
- Happé F, Ronald A, & Plomin R. (2006) Time to give up on a single explanation for autism. *Nature Neuroscience* 9(10):1218-1220.
- Haxby JV, Gobbini MI, Furey ML, Ishai, A, Schouten JL, & Pietrini P. (2001). Distributed & Overlapping Representations of Faces & Objects in Ventral Temporal Cortex. *Science* 293(5539):2425-30.
- Haynes, JD & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7), pp.523-534.
- Jenkins, A. C., & Mitchell, J. P. (2009). Mentalizing under uncertainty: Dissociated neural responses to ambiguous & unambiguous mental state inferences. *Cerebral Cortex*, 20(2), 404-410.
- Kennedy DP & Courchesne E. (2008). Functional abnormalities of the default network during self- & other-reflection in autism. *Social Cognitive & Affective Neuroscience* 3(2): 177-190.
- Koster-Hale, J & Saxe, R. R. (2010). theory of mind brain regions are sensitive to the content, not the structural complexity, of belief attributions. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.
- Knobe, J. (2005). theory of mind & Moral Cognition. *Trends in Cognitive Sciences*, 9, 357-359.
- Kriegeskorte N & Bandettini P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*.
- Leslie A, Mallon R, & DiCorcia J. (2006). Transgressors, victims, & cry babies: Is basic moral judgment spared in autism? *Social Neuroscience* 1(3-4):270-283.
- Lombardo MV, Chakrabarti B, Bullmore ET, MRC AIMS Consortium, & Baron-Cohen S. (2011). Specialization of right temporo-parietal junction for mentalizing & its relation to social impairments in autism. *NeuroImage* 56(3):1832-8.
- Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental & Social Psychology*, 33, 101-121.
- Mason RA, Williams DL, Kana RK, Minshew N, & Just MA. (2008). theory of mind disruption & recruitment of the right hemisphere during narrative comprehension in autism. *Neuropsychologia* 46(1).
- Moran J, Young L, Saxe R, Lee SM, O'Young D, Mavros P, & Gabrieli J. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *In PNAS* 108(7):2688-92.
- Peelen, M.V., Wiggett, A.J. & Downing, P.E., (2006). Patterns of fMRI Activity Dissociate Overlapping Functional Brain Areas that Respond to Biological Motion. *Neuron*, 49(6), pp.815-822.
- Peterson C, Wellman H, & Liu D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Development* 76(2):502-517.
- Piaget, J. (1965). The moral judgment of the child. New York: Free Press.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage*, 19(4), 1835-1842.
- Scholz, J., Triantafyllous, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. (2009). Distinct regions of the right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS ONE*, 4(3), 1-7.
- Silani G, Bird G, Brindley R, Singer T, Frith C, & Frith U. (2007) Levels of emotional awareness & autism: An fMRI study. *Social Neuroscience* 3(2):97-112.
- Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain responses to the acquired moral status of faces. *Neuron*.
- Tesink CMJY, Buitelaar JK, Petersson KM, van der Gaag RJ, Kan CC, Tendolkar I, & Hagoort P. (2009). Neural correlates of pragmatic language comprehension in autism spectrum disorders. *Brain* 132(7).
- Wang AT, Lee SS, Sigman M, & Dapretto M. (2006). Neural basis of irony comprehension in children with autism: the role of prosody & context. *Brain* 129(4):932-43.
- Young, L., Saxe, R. (2008). The neural basis of belief encoding & integration in moral judgment. *NeuroImage*, 40, 1912-1920
- Young, L., & Saxe, R. (2009). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience* 21(7), 1396-1405.
- Young, L., & Saxe, R. (2009). Innocent intentions: a correlation between forgiveness for accidental harm & neural activity. *Neuropsychologia*.
- Young, L., Camprodon, J., Hauser, M., Pascual-Leone, A., Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *PNAS*, 107, 6753-6758
- Young, L., Dodell-Feder, D., & Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An FMRI investigation of attention & theory of mind. *Neuropsychologia*.
- Young, L., Nichols, S., Saxe, R. (2010). Investigating the Neural & Cognitive Basis of Moral Luck: It's Not What You Do but What you Know. *Review of Philosophy & Psychology*, 1, 333-349.
- Young, L., Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, 120, 202-214.

# A Unified Model of Categorical Effects in Consonant and Vowel Perception

**Yakov Kronrod** (yakov@umd.edu)  
**Emily Coppess** (ercoppess@umd.edu)  
**Naomi H. Feldman** (nhf@umd.edu)

Department of Linguistics, 1401 Marie Mount Hall  
University of Maryland, College Park, MD 20742

## Abstract

Consonants and vowels differ in the extent to which they are perceived categorically. We use a Bayesian model of speech perception to explore factors that might cause this difference. Simulations show that perception of vowels, fricatives, and stop consonants can all be captured under a single model in which listeners use their knowledge of phonetic categories to infer the sound that a speaker intended. This suggests that the differences in the way we perceive vowels and consonants, when viewed at the computational level, can be explained as parametric variation within a single framework.

**Keywords:** perceptual magnet effect; categorical perception; Bayesian modeling; computational linguistics

Phonetic categories influence perception of speech sounds, with stimuli belonging to different categories being easier to discriminate than stimuli from a single category (Lieberman, Harris, Hoffman, & Griffith, 1957; Kuhl, 1991). However, different types of sounds differ in the degree to which they are perceived categorically. At one end of the spectrum, perception of stop consonants is strongly categorical. Discrimination is little better than would be expected if listeners used only category assignments to distinguish sounds, and between-category differences are extremely pronounced (Lieberman et al., 1957). At the other end of the spectrum, vowel perception is much more continuous, with some even arguing that vowels display no categorical effects at all (Fry, Abramson, Eimas, & Liberman, 1962).

Researchers have used various mechanisms underlying speech perception to explain these differences. For example, differences have been proposed to stem from the way each type of sound is stored in memory (Pisoni, 1973) and to be related to innate auditory discontinuities that seem to influence stop consonant perception (Pisoni, 1977; Eimas, Siqueland, Jusczyk, & Vigorito, 1971). However, the qualitative similarity of categorical effects in consonants and vowels suggests that in some ways these are also instances of the same phenomenon. This raises the possibility that perceptual differences among different classes of sounds are quantitative rather than qualitative.

In this paper we explore these similarities and differences at Marr's (1982) computational level, looking at the optimal solution to the problem of inferring speakers' intended productions from the available acoustic information. We adapt a Bayesian model from Feldman, Griffiths, and Morgan (2009), in which listeners use their knowledge of phonetic categories to recover the sound a speaker intended. We show that an extended version of this model can account for perceptual data from stop consonants and fricatives as well as vowels. This

suggests that differences in the degree to which vowels and consonants are perceived categorically can be explained as parametric variations in a single underlying model.

Our paper is organized as follows. First, we review evidence concerning categorical effects in consonants and vowels, giving an overview of the types of explanations that have been proposed to account for these data. We then describe the model from Feldman et al. (2009) in detail, focusing on their results for vowel perception. In the subsequent section, we present simulations testing our extended model on two types of consonants, stop consonants and fricatives, to determine whether a model built for vowels can also account for patterns in consonant perception. We conclude by summarizing our findings and discussing implications for theories of speech perception.

## Categorical Effects in Speech Perception

Categorical perception in stop consonants was first described by Liberman et al. (1957) as consisting of a sharp change in the identification function between different consonants, as well as a peak in the discrimination function at the location of the identification boundary. The authors proposed a model in which participants used only category assignments to determine whether sounds were the same or different. If the sounds belonged to different categories, then participants would respond *different*; otherwise, they would respond *same*. By examining participants' identification functions, Liberman et al. could use this model to predict the extent to which participants should be able to discriminate each pair of sounds. Participants' actual discrimination performance exceeded the model's predictions only by a small amount, and the authors took this as evidence of a strong categorical component in stop consonant perception. Liberman et al.'s experiment focused on stop consonants that differed by place of articulation, but similar findings have been obtained along the voice onset time (VOT) dimension as well (Wood, 1976).

Descriptions of categorical effects in vowels have focused primarily on the *perceptual magnet effect* (Kuhl, 1991). This effect was originally proposed as a within-category phenomenon, characterized by sounds near category centers being more difficult to discriminate than sounds near category edges, with an accompanying correlation between goodness ratings and discriminability. There is disagreement over whether categorical perception and the perceptual magnet effect are separate phenomena or different variants of the same phenomenon (e.g. Lotto, Kluender, & Holt, 1998). Some characteristics of the perceptual magnet effect are similar to

pure categorical effects, such as reduced discriminability near category centers. Data from Iverson and Kuhl (2000) suggested that discrimination peaks near category boundaries are separable from correlations of discrimination and goodness ratings, whereas more recent studies have found that these two effects cooccur (Tomaschek, Truckenbrodt, & Hertrich, 2011). Regardless of terminology, however, categorical effects in vowel perception are much weaker than those found in consonant perception.

In addition to stop consonants and vowels, it is natural to consider categorical perception of another major class of speech sounds, fricatives. In this paper, we consider categorical perception of sibilant fricatives. There has been some disagreement over the degree of categorical perception in fricatives in previous research. Repp (1981) showed that fricatives follow patterns similar to the categorical perception found for stop consonants. However, in the same study, a subset of participants seemed to have perception that was more continuous, which Repp attributed to a choice between two processing strategies, acoustic and phonetic. Others have found that fricative perception is much less categorical than stop consonants (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Healy & Repp, 1982; Repp, 1984). Another more recent study showed identification patterns consistent with categorical perception together with a neural signature indicative of something like perceptual warping near category centers (Lago, Kronrod, Scharinger, & Idsardi, 2010).

These data set up a continuum ranging from nearly completely categorical perception of stop consonants to much more continuous perception of vowels, with fricatives falling somewhere in between. However, this continuum is not as clear cut as it may seem, as neural and behavioral evidence suggests that listeners attend to phonetic detail when perceiving stop consonants (Pisoni & Lazarus, 1974; Blumstein, Myers, & Rissman, 2005), and the degree of categorical perception in both consonants and vowels can be influenced by task-related factors (Pisoni, 1975; Repp, Healy, & Crowder, 1979). Nevertheless, the differences between consonant and vowel perception are robust. In what follows, we use a model to account for the variability in these effects within a single framework, identifying aspects of category structure that might contribute to differences in categorical effects across consonants and vowels.

## Model Overview

Our model is an extension of the model from Feldman et al. (2009). The model assumes that listeners are trying to recover phonetic detail about the speaker's intended production as well as category information. It differs from traditional models of categorical perception in that it recognizes two different sources of within-category variability: meaningful variability (also referred to as category variance) and noise variance. The category variance  $\sigma_c^2$  is assumed to arise from processes that yield information useful to listeners, such as coarticulatory effects that allow listeners to predict the iden-

tity of upcoming sounds (Gow, 2001). Once a speaker selects a target production,  $T$ , from the category, there is assumed to be additional articulatory, acoustic, and perceptual noise  $\sigma_s^2$  that further distorts this sound. This process results in a speech sound  $S$  that is heard by listeners.

Listeners are trying to infer the target production through the noisy speech signal. To do this, listeners can use their knowledge that speakers tend to produce sounds near category centers. Hence, they rely both on category knowledge and on acoustic cues to recover the phonetic detail of a speaker's target production. If listeners encounter little noise and the category allows a large amount of meaningful variability (e.g., coarticulation), then listeners attend more to acoustic detail in perceiving sounds; however, in situations with high noise and low meaningful variability, they rely more on their knowledge of phonetic categories. This relationship between category variance and noise plays an important role in determining the degree to which perception is biased toward category centers, and it thus has the potential to account for differences in the degree to which vowels and consonants are perceived categorically.

Feldman et al.'s (2009) original model relied on a simplifying assumption that all categories considered by a listener have equal category variance. While this assumption might be adequate for vowels, other sound categories do not necessarily reflect this simplification, particularly voiced and voiceless stop consonants which have been shown to have substantial differences in VOT variance (Lisker & Abramson, 1964). Hence, we extend the original model proposed by Feldman et al. (2009) to allow for unequal category variances. This section gives an overview of our extended model; full derivations are omitted due to space limitations, but are parallel to those in Feldman et al.

The model assumes phonetic categories are Gaussian distributions of sounds along the relevant auditory dimensions, so that a speaker's target production is normally distributed around the category mean,  $T|c \sim N(\mu_c, \sigma_c^2)$ . Noise in the speech signal causes the stimulus heard by listeners to be normally distributed around the target production,  $S|T \sim N(T, \sigma_s^2)$ . We can integrate over all possible target productions  $T$  to get an expression relating the perceived sound directly to the underlying categories,

$$S|c \sim N(\mu_c, \sigma_c^2 + \sigma_s^2) \quad (1)$$

In identification tasks, listeners recover a category given the sound  $S$ , which corresponds to computing the posterior distribution over category membership  $p(c|S)$  in the model. They can compute this by applying Bayes' rule,

$$p(c|S) = \frac{p(S|c)p(c)}{\sum_c p(S|c)p(c)} \quad (2)$$

If we limit ourselves to two categories but relax the assumption that these have equal category variances, we need two means ( $\mu_{c_1}$  and  $\mu_{c_2}$ ) and category variance parameters ( $\sigma_{c_1}$  and  $\sigma_{c_2}$ ). We derive the identification function by substituting

Simulation	Means		Variances			Category:Noise Variance Ratio
	$\mu_{c_1}$	$\mu_{c_2}$	$\sigma_{c_1}^2$	$\sigma_{c_2}^2$	$\sigma_S^2$	
Vowels (Equal Variance)	F1=224 Hz F2=2413 Hz	F1=423 Hz F2=1936 Hz	5,873	5,873 (Mels)	878	6.69
Stop Consonants (Unequal Variance)	60 ms	-0.3 ms	253.9	14 (ms)	82.3	/p/: 3.09, /b/: 0.17
Fricatives (Unequal Variance)	19.0 Barks	15.99 Barks	0.5992	0.5772 (Barks)	0.3098	/s/: 1.93, /f/: 1.86

Table 1: Best fitting model parameters for vowels (Feldman et al., 2009), stop consonants, and fricatives.

Equation 1 into Equation 2 and following a parallel derivation to that in Appendix B from Feldman et al. (2009), yielding

$$p(c_1|S) = \frac{1}{1 + \sqrt{\frac{\sigma_1^2}{\sigma_2^2}} \times \exp \frac{(\sigma_2^2 - \sigma_1^2)S^2 + 2(\mu_{c_2}\sigma_1^2 - \mu_{c_1}\sigma_2^2)S + (\mu_{c_1}^2\sigma_2^2 - \mu_{c_2}^2\sigma_1^2)}{2\sigma_1^2\sigma_2^2}} \quad (3)$$

where  $\sigma_1^2 = \sigma_{c_1}^2 + \sigma_S^2$  and  $\sigma_2^2 = \sigma_{c_2}^2 + \sigma_S^2$ .

The model assumes that listeners recover the phonetic detail of a speaker’s target production in addition to category information when perceiving a speech sound, and that they use this information when performing a discrimination task. Perceiving phonetic detail corresponds to computing the posterior distribution on target productions,  $p(T|S)$ . Applying Bayes’ rule, where the prior  $p(T)$  is a mixture of Gaussians and the likelihood  $p(S|T)$  is Gaussian, we obtain a posterior distribution whose form is a mixture of Gaussians and whose mean is

$$E[T|S] = \sum_c p(c|S) \frac{\sigma_c^2 S + \sigma_S^2 \mu_c}{\sigma_c^2 + \sigma_S^2} \quad (4)$$

(see Feldman et al., 2009 for a full derivation). Each category makes a contribution to this posterior mean with magnitude proportional to the posterior probability of the sound belonging to that category,  $p(c|S)$ . The specific contribution of each category is to bias perception toward the category mean. The strength of the bias is controlled by the relationship between parameters  $\sigma_c^2$  and  $\sigma_S^2$ , which represent the amount of meaningful variability and the amount of noise. Notice that the contribution of the category mean,  $\mu_c$ , is weighted by the noise variance,  $\sigma_S^2$ . This means that when there is more noise, listeners will rely more on their underlying knowledge of the categories. In contrast, the acoustic information,  $S$ , is weighed by the meaningful variance parameter,  $\sigma_c^2$ , such that when there is a lot of meaningful variability in the underlying category, listeners will pay more attention to the acoustic data. It is this relationship that will be critical to modeling differences in perception between different categories of sounds.

Feldman et al. (2009) applied their model to vowel perception (continuum from /e/ to /i/), obtaining a close fit to the multidimensional scaling data from Iverson and Kuhl (1995) (Figure 1). However, in analyzing their own data from an AX

discrimination experiment, the patterns they found suggested that multidimensional scaling was distorting the perceptual patterns, and that the noise parameter needed to capture experimental data directly was much lower than they initially found.<sup>1</sup> Thus, the “Vowels (Equal Variance)” section of Table 1 shows the values they derived on the basis of their experimental data. As might be expected for relatively continuous vowel perception, these parameters showed high meaningful category variance relative to noise variance, indicating that the bias toward category centers was small. We use these parameters as a baseline for comparison in our consonant simulations.

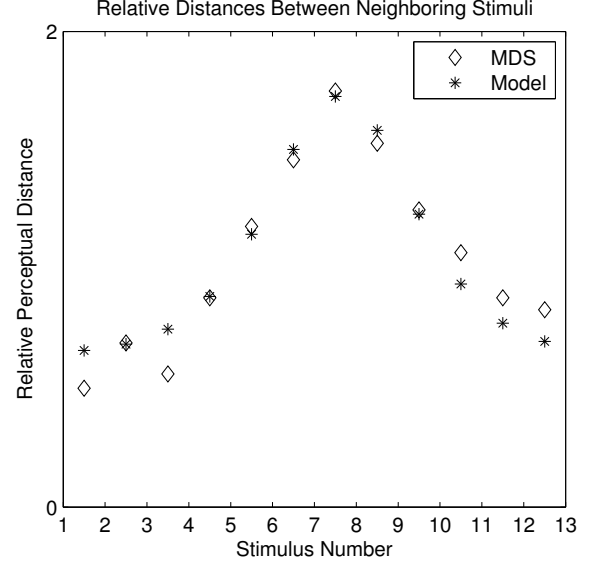


Figure 1: Figure from Feldman et al. (2009) showing inter-stimulus distances from Iverson and Kuhl’s (1995) multidimensional scaling solution and the fitted model.

<sup>1</sup>In our simulations below, we select data that use the distance measure  $d'$  rather than multidimensional scaling data in order to avoid this type of discrepancy.

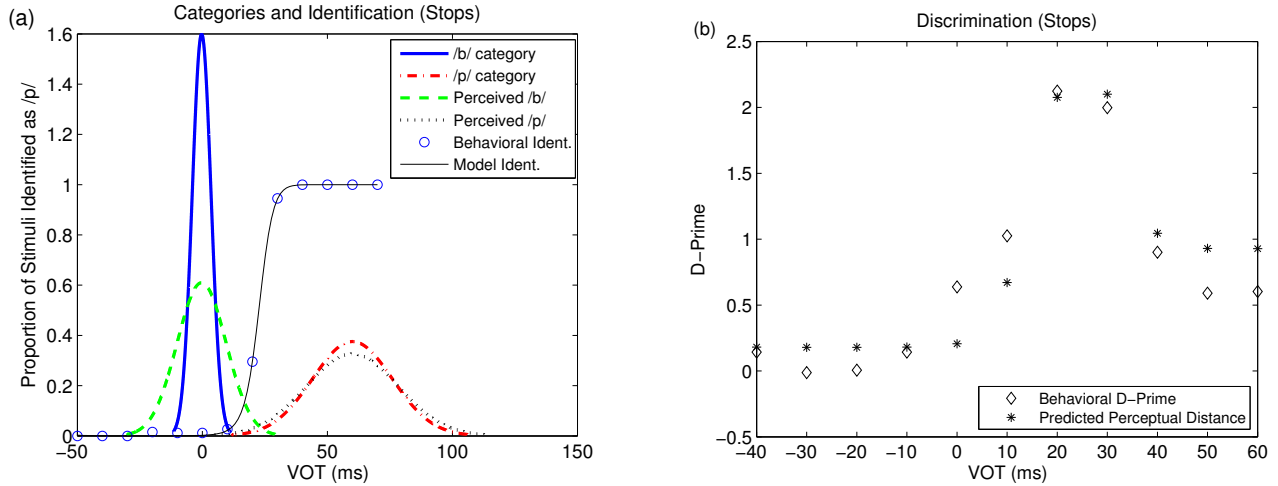


Figure 2: Stop consonants: (a) Underlying categories, perceived distributions, and identification curve in the model, together with behavioral identification data; (b) Interstimulus distances predicted by the model, together with behavioral  $d'$  data.

## Simulations

We applied this model to data from stop consonants and sibilant fricatives, deriving parameters on the basis of experimental data in order to determine whether categorical effects in each type of sound can be explained as the result of optimally inferring the phonetic detail of a speaker’s target production. Examining the resulting parameters then allows us to assess the degree to which those parameters are adequate with regard to existing data, as well as examine the relationship between the two variance parameters and the degree of observed perceptual warping. Following Feldman et al. (2009), our general strategy for fitting our parameters was as follows:

1. Set  $\mu_{c_1}$  on the basis of production data.
2. Determine  $\mu_{c_2}$ ,  $\sigma_1^2$ , and  $\sigma_2^2$  from identification data using Equation 3.
3. Determine the ratio of category variances,  $\sigma_{c_1}^2$  and  $\sigma_{c_2}^2$ , to noise,  $\sigma_s^2$ , by fitting acoustic differences between percepts,  $E[T|S]$ , in the model (Equation 4) to a distance measure such as  $d'$ .

We need to set one of the means in order to obtain a single identifiable set of parameters. The model is then fit to identification data, allowing us to derive the other mean as well as both sums of variances (one corresponding to each category). Note that the only parameter being fit directly to the discrimination data is the ratio of meaningful category variance to noise variance, which is the parameter of interest for examining the degree of bias toward category centers exhibited by each class of sounds. In effect, the discrimination data provide a general test of the model’s fit to behavioral data from each class of sounds.

## Stop Consonants

We first consider behavioral data from identification and discrimination experiments on stop consonants, which have been found to exhibit much stronger categorical effects in perception than vowels. Under our account, this difference might stem from low category variance relative to noise variance, such that listeners rely more on category information. If we are able to account for stop consonant perception with our model, then that would suggest that it may not be necessary to posit qualitative differences in the types of computations performed by listeners when perceiving consonants and vowels. It is not obvious that our model should be able to explain stop consonant data, however, as other factors such as innate phonetic boundaries (Eimas et al., 1971) or auditory discontinuities (Pisoni, 1977) might retain their influence on stop consonant perception even after phonetic learning is complete.

For this simulation we used identification and discrimination data from Wood (1976), who examined perception of /p/ and /b/ along a voice onset time (VOT) continuum. The continuum consisted of synthetic stimuli ranging from -50 to +70 ms VOT. A forced identification task as well as both a 10-ms and 20-ms difference AX discrimination task were administered. We used 20-ms discrimination data for our simulations. On the basis of data from Lisker and Abramson (1964), we set  $\mu_{/p/}$  at 60 ms VOT and derived the remaining parameters from the identification and discrimination data. The identification fit produced an estimated value of -0.3 ms for the mean  $\mu_{/b/}$ , which was a close match to production data found in Lisker and Abramson (1964). The full set of parameters is found in section “Stop Consonants (Unequal Variance)” of Table 1, and the resulting identification curve and category distributions are shown in Figure 2(a). The fit between model and data is very close: the model is even able to predict the reduced within-category discriminability of voiced stops compared to voiceless stops that is observed in the empirical data.

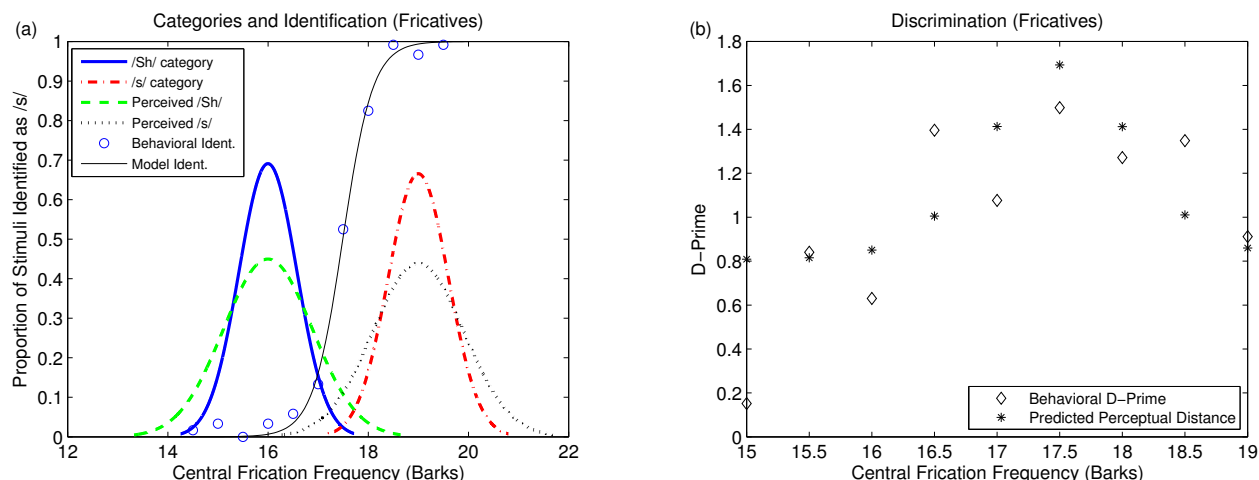


Figure 3: Sibilant fricatives: (a) Underlying categories, perceived distributions, and identification curve in the model, together with behavioral identification data; (b) Interstimulus distances predicted by the model, together with behavioral  $d'$  data.

This can be seen in Figure 2(b), where the perceptual distance between stimuli toward the left side of the continuum is lower than that toward the right side of the continuum.

As predicted, the ratio of category variance to speech signal noise was lower than that obtained for vowels for both categories of stop consonants, voiced and voiceless. These findings suggest that stop consonants have less meaningful within-category variance relative to noise variance than vowels, leading the listener to rely on prior category knowledge in inferring the speakers' target production. This causes a greater pull toward category centers and hence stronger categorical perception. Perceptual bias is particularly strong in voiced stops due to their low category variance.

### Sibilant Fricatives

The previous simulation indicates that the model provides a good account for stop consonants as well as vowels. We next apply our model to sibilant fricatives. Sibilant fricatives are obstruents (like stop consonants), but their characteristic noise components at higher frication frequencies show some similarity to the higher formant structures of vowels. As discussed above, there has been conflicting evidence on the strength of categorical effects they exhibit. These factors make fricatives an interesting modeling target.

For this simulation we used identification and discrimination data along the /s/-/ʃ/ continuum from Lago et al. (2010). The continuum consisted of 11 tokens with central frication frequencies varying from 14.5 to 19.5 Barks. A forced identification task as well as a 2-step AX discrimination task were administered to 12 participants. For our model, we fixed the value of  $\mu_{/s/}$  to 19.0 Barks based on natural productions by an adult male participant and derived values for the remaining parameters by fitting the model to behavioral identification and discrimination data. The resulting parameter values are given in the “Fricatives (Unequal Variance)” entry in Table 1.

Figure 3(a) shows the identification data and the identification curve used in our model, together with the underlying and perceived category distributions that correspond to the parameters used in our simulation. Figure 3(b) compares the model predictions to the observed discrimination measures. The fit is not perfect, due in part to noisy data from the original experiment, but both data and model show the peak in discrimination at the same location as the inflection point in the identification data.

Given that fricatives tend to be perceived more categorically than vowels but less so than consonants, we might expect the category variance to noise ratio to be smaller for fricatives than for vowels, leading to a larger perceptual bias toward category centers, but larger than that for stop consonants, indicating more attention focused on acoustic cues. As predicted, the ratios for the sibilant fricatives are reduced compared to the parameters estimated for vowels (1.93 and 1.86 compared to 6.69). Additionally, we see that they are close to the ratio for the voiceless stops but much higher than that of the voiced consonants, suggesting that they may be closer to the stop consonant end of the spectrum in terms of their degree of bias toward category centers.

### Discussion

This paper used a Bayesian model to investigate the relationship between categorical effects in consonant and vowel perception. Our results suggest that these effects can be explained at Marr's (1982) computational level by the same underlying principles: Listeners use their knowledge of phonetic categories to optimally infer a speaker's target production through a noisy speech signal, and this causes their perception to be biased toward category centers. The model accounts for differences in the strength of categorical effects by assigning consonants less meaningful variability, compared with noise variance, than vowels.



Our analysis is reminiscent of an analysis pursued by Pisoni (1973), Liberman et al. (1967), and others. Their account proposes that speech perception incorporates a phonetic mode of perception, i.e., categorical perception, and an auditory mode of perception, i.e., continuous perception. Pisoni (1973) argued that differences between consonant discriminability and vowel discriminability could be accounted for by assuming that listeners have less access to auditory short-term memory when hearing consonant stimuli than when hearing vowel stimuli. This distinction between phonetic and acoustic modes of perception is parallel to the weighted average in Equation 4, where acoustic information is weighted by the category variance and the category mean is weighted by the noise variance. When noise variance dominates over category variance, listeners rely more on the category mean rather than acoustic information (i.e. phonetic mode). Otherwise, acoustic information receives more weight and within-category discriminability increases (i.e. auditory mode). Looking at ratios of category variance to noise variance across consonants and vowels we see that for vowels category variance exerts much more influence than noise variance and therefore listeners' perception is drawn less towards the category center, causing within-category discriminability to increase (i.e. continuous perception). For consonants the ratio is smaller, coinciding with a decrease in within-category discriminability (i.e. categorical perception).

Our findings suggest that perception of three types of sounds – vowels, stop consonants, and fricatives – adheres to the same abstract computational principles. Importantly, however, the idea that listeners are performing the same computation at an abstract level does not necessarily mean that the underlying mechanisms are identical. Our analysis simply suggests that perception of each type of sound has been optimized to allow listeners to recover the sound intended by a speaker. Bias toward category centers may be implemented differently across different classes of sounds, and separate mechanisms are almost certainly necessary for extracting the various cues we have used as input to our model (formant frequencies for vowels; voice onset time for stop consonants; and central frication frequencies for fricatives). In future work we hope to explore these issues by considering perception of a fourth class of speech sounds, nasals, and by linking our computational approach with descriptions of sound perception at the algorithmic and implementational levels.

**Acknowledgments** We thank Sol Lago, Mathias Scharinger, and Bill Idsardi for sharing behavioral data from their experiments with fricatives. We also thank the Computational Psycholinguistics group, the PFNA Sounds group, and the Language Science Lunch group for valuable discussion and feedback. Finally, we thank Bill Idsardi and four anonymous reviewers for their insightful commentary and suggestions on earlier versions of this paper. This work was supported in part by NSF IGERT grant DGE-0801465.

## References

Blumstein, S. E., Myers, E. B., & Rissman, J. (2005). The perception of voice onset time: An fMRI investigation of phonetic category structure. *Journal of Cognitive Neuroscience*, 17(9), 1353-1366.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, 171(3968), 303-306.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4), 752-782.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech*, 5, 171-189.

Gow, D. W. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language*, 45, 133-159.

Healy, A. F., & Repp, B. H. (1982). Context independence and phonetic mediation in categorical perception. *Journal of Experimental Psychology: Human Perception and Performance*, 8(1), 68-80.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*, 97(1), 553-562.

Iverson, P., & Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception and Psychophysics*, 62(4), 874-886.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, 50(2), 93-107.

Lago, S., Kronrod, Y., Scharinger, M., & Idsardi, B. (2010). *Categorical perception of [s] and [sh]: An MMN study*. Neurobiology of Language Conference. San Diego, CA.

Liberman, A. M., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74, 431-461.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5), 358-368.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20, 384-422.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1998). Depolarizing the perceptual magnet effect. *Journal of the Acoustical Society of America*, 103(6), 3648-3655.

Marr, D. (1982). *Vision: A computational investigation in the human representation of visual information*. San Francisco: Freeman & Co.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, 13(2), 253-260.

Pisoni, D. B. (1975). Auditory short-term memory and vowel perception. *Memory and Cognition*, 3(1), 7-18.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception. *Journal of the Acoustical Society of America*, 61(5), 1352-1361.

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, 55(2), 328-333.

Repp, B. H. (1981). Two strategies in fricative discrimination. *Perception and Psychophysics*, 30(3), 217-227.

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. *Speech and Language: Advances in Basic Research and Practice*, 10, 243-335.

Repp, B. H., Healy, A. F., & Crowder, R. G. (1979). Categories and context in the perception of isolated steady-state vowels. *Journal of Experimental Psychology: Human Perception and Performance*, 5(1), 129-145.

Tomaschek, F., Truckenbrodt, H., & Hertrich, I. (2011). Processing german vowel quantity: Categorical perception or perceptual magnet effect? *Proceedings of the 17th International Conference of Phonetic Sciences*, 2002-2005.

Wood, C. C. (1976). Discriminability, response bias, and phoneme categories in discrimination of voice onset time. *Journal of the Acoustic Society of America*, 60(6), 1381-1389.

# Early and Repeated Exposure to Examples Improves Creative Work

**Chinmay Kulkarni**

Stanford University HCI Group  
Stanford, CA 94305-9035 USA  
chinmay@cs.stanford.edu

**Steven P. Dow**

Carnegie Mellon, HCI Institute  
Pittsburgh, PA 15213  
spdown@cs.cmu.edu

**Scott R Klemmer**

Stanford University HCI Group  
Stanford, CA 94305-9035 USA  
srk@cs.stanford.edu

## Abstract

This article presents the results of an online creativity experiment ( $N = 81$ ) that examines the effect of example timing on creative output. In the between-subjects experiment, participants drew animals to inhabit an alien Earth-like planet while being exposed to examples early, late, or repeatedly during the experiment. We find that exposure to examples increases conformity. Early exposure to examples improves creativity (measured by the number of common and novel features in drawings, and subjective ratings by independent raters). Repeated exposure to examples interspersed with prototyping leads to even better results. However, late exposure to examples increases conformity, but does not improve creativity.

## Introduction

Examples are considered “a cornerstone of creative practice” (Herring et al., 2009). Leveraging examples of prior work is an established technique in design (Buxton & Buxton, 2007), and many design programs encourage students to use examples of existing designs (Schön, 1985). However, the strategies employed by designers to seek and use examples is largely ad-hoc (Newman & Landay, 2000).

Frequently, these strategies differ in timing—in an informal survey we conducted among designers around Stanford University, one respondent described inspirational examples as “huge parts of my initial steps. I need to know as much as I can about the topic before I feel comfortable moving forward.” In contrast, another said that “I don’t do this [look at examples] at the very beginning because it gets your mind stuck in one way of thinking.” This fear of conformity was echoed by other participants, and one went on to say that he looked for inspiration only when “facing a creative block.”

These different strategies suggest that examples may modify the creative process differently depending on the point in the design process at which they are presented. This leads to the practical question: what are the tradeoffs of looking at examples earlier or later in the design process? Furthermore, even if there is an “ideal” time to view examples, some designers feel ubiquitous information access and their own “thirst for knowledge” bombards them constantly with examples (Herring et al., 2009). How does this repeated exposure to examples affect the creative process?

This article presents the results of an online creativity experiment we conducted on Amazon Mechanical Turk. Participants in the experiment generated drawings of alien creatures as a creative task. The pervasive use of sketches to develop and communicate conceptual designs in the creative fields (Suwa & Tversky, 1997), and the use of similar tasks in prior work Ward (1994) inspired the choice of the drawing task. Focusing on drawings of alien figures makes this task

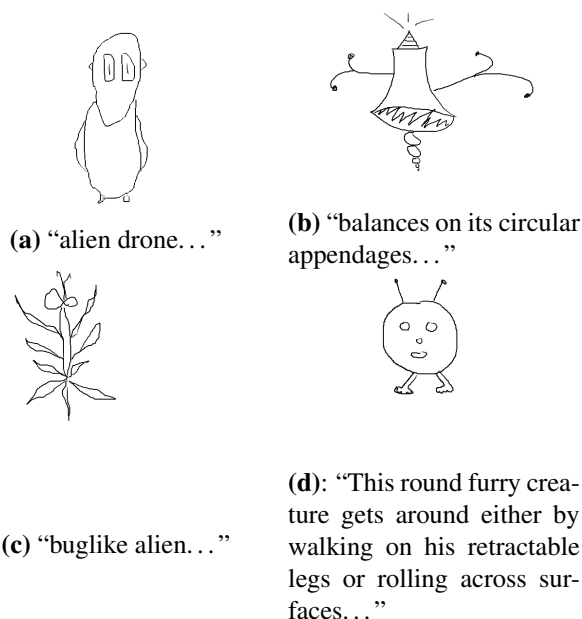


Figure 1: Sampling of drawings created in our experiment, with excerpts of participant-provided descriptions

readily accessible to non-designers (see Figure 1 for a sampling of drawings created by participants). Participants were randomly assigned to one of four conditions: examples early, examples late, examples early and late, or a control condition without examples. This study’s creativity measures were the number of uncommon and novel features in the drawings and Likert-scale ratings by condition-blind raters. Conformity was measured by the number of critical features (features that were directly copied from examples).

This paper’s experimental results suggest that while exposure to examples increases conformity, such exposure early in the creative process improves the creativity in the output, while later exposure provides no such benefit. Furthermore, exposure to examples followed by prototyping and subsequent re-exposure to the same examples improved creative output even more. This finding may allay some fears of example bombardment. Lastly, in our experiment, participants exposed to examples created fewer drawings, so these example driven quality improvements may come at the cost of a lower quantity of creative work.

## Related work

**Examples:** Bringing existing solutions to mind is crucial for creative generation (Smith et al., 1993). The Structured Imag-

ination theory by Ward (1994) describes creativity as a multi-step process: in the *recall* step, people bring to mind existing solutions and constructs. Then, in the *modification* step, these constructs are altered in novel ways. Similar analogical processes are found in other areas of cognition such as analysis and learning (Gentner & Colhoun, 2010).

Designers often incorporate features from examples directly into their work (Marsh & Bower, 1993, “inadvertent plagiarism”); but examples also “ultimately alter the nature of the creative product” in more subtle ways (Marsh et al., 1996). Lee et al. (2010) found that designing with examples generally improves the quality of creative work. These findings have also led to tools for discovering, storing and retrieving examples (Kerne et al., 2008; Ritchie et al., 2011).

The current work is an extension of Marsh et al. (1996) (itself an extension of Smith et al., 1993), so we describe Marsh et al.’s experiment in more detail. In their experiment, participants generated drawings of non-Earth-like creatures to inhabit an alien planet similar to Earth. In the example conditions, experimenters provided participants example drawings of aliens at the start of the experiment. Example drawings all had certain attributes, or *critical* features, in common—four legs, antennae and a tail. The proportion of these critical features incorporated into participants’ own drawings was used as a measure of conformity. The proportion of other, non-critical, features was used as a measure of creativity. These non-critical features were classified as either *novel* (not commonly found on animals, such as speakers or propellers), *uncommon* (such as a pouch or tentacles), or *common* (such as a nose, mouth or two legs).

Participants exposed to examples incorporated more critical features in their drawings, but not at the expense of novel and uncommon features. Instead, their drawings contained fewer *common* features. This suggests that while examples increase conformity by increasing activation of critical features, they do not block retrieval of original ideas (such as novel and uncommon features).

We use Marsh et al.’s feature-based evaluation metric, and extend their work by examining how the example timing affects creative output. In addition, we study the effects of repeated exposure to examples in the creative process.

**Research methods** Our experiment uses a task (drawing sketches of alien figures) that has previously been employed to study creativity in a context of no prior training (Marsh & Bower, 1993; Ward, 1994). Drawing tasks have also been demonstrated to be appropriate for online experiments (Yu & Nickerson, 2011).

This experiment was run on Amazon Mechanical Turk ([www.mturk.com](http://www.mturk.com)), a web-based crowdsourcing platform. This platform has been used for experiments on affect and creativity (Lewis et al., 2011). Mechanical Turk workers have also been employed to provide perception responses (Heer & Bostock, 2010), objective labels (Deng et al., 2009; Snow et al., 2008), and subjective ratings (Dow et al., 2011).

	Task before first session	Task before second session
<b>Condition</b>		
<b>Control</b>	Think	Think
<b>Early</b>	Examples	Think
<b>Late</b>	Think	Examples
<b>Repeated</b>	Examples	Examples

Table 1: Experimental conditions

## Experiment

Our experiment had two goals. First, we wanted to see if exposure to examples at the start of a creative process leads to a different quality of creative output in contrast to exposure when the creative process is underway. Second, we wanted to investigate the role of repeated exposure to examples.

Our initial hypothesis was that exposure to examples later in the creative process would have the same creative benefits but lower conformity than exposure at the start. This hypothesis was motivated by Weisberg (1999), who observed that creative failures are more often explained by the absence of relevant information than the presence of irrelevant information. Furthermore, the presence of one’s own ideas would inhibit the adoption of sub-optimal ideas from late exposure to examples (mirroring the intuitions of some designers).

In the case of repeated exposure, the activation account would predict that, similar to showing more examples at once, this would result in greater degree of conformity due to higher activation of features present in examples.

## Participants

We solicited US-resident participants on Mechanical Turk with a compensation of US\$1.00. 81 participants responded (27 male, 54 female; median age 34). All participants reported a high-school diploma or a higher degree. This between-subjects experiment randomly assigned participants to one of four conditions.

## Procedure

The experiment comprised two drawing sessions, each lasting 7 minutes. Participants were asked to create as many drawings as they could during the drawing session. To encourage this (and discourage participants from spending time perfecting only a few drawings), the experimental platform included a clear-canvas tool but no line-eraser tool (Figure 2).

Each session was preceded by a condition-specific task in which participants were either exposed to examples, or asked to think about the aliens they planned to draw in the next session (Table 1).

At the start of the experiment, all participants saw a Web page with instructions adapted from Marsh et al. (1996) to account for two drawing sessions and a break (see [hci.stanford.edu/example/aliens](http://hci.stanford.edu/example/aliens) for actual prompts used).

For the **Example** task, participants were shown three example alien drawings for 90 seconds (see Figure 4). We used drawings from (Marsh et al., 1996), p. 672. Using the prompt

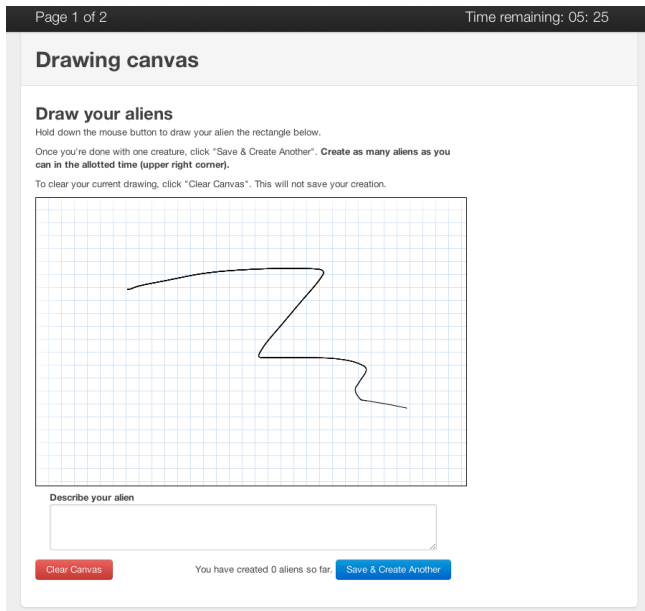


Figure 2: Drawing canvas with time remaining (top right), and an option to clear the canvas (bottom left, red)

of Marsh et al. (1996) (and Smith et al., 1993), participants were instructed that examples were only shown to help them create their original creations, and that we did not want them to copy the examples in any aspect. For the **Think** task, participants were asked to “think about aliens” they planned to draw in the next session for 90 seconds. In the Repeated Examples condition, participants saw the same three examples before both drawing sessions.

After the second drawing session, participants filled out a survey that covered demographics, artistic interest and ability and the thought-process they followed while drawing.

### Labeling features in drawings

Participants generated a total of 543 drawings. Each drawing was labelled with the features it incorporated from the feature set of Marsh et al. (1996) (Appendix). Drawings were annotated on Mechanical Turk, since the features were well-defined. All workers were US resident and at least 18 years of age, and were compensated US\$0.50 for the task. Workers who participated in the experiment were disallowed from the annotation task (and vice-versa). All annotators were blind to experimental condition.

Workers were trained using a drawing from a pilot participant (Figure 3). Then, each worker annotated a set of seven randomly assigned drawings. Workers also rated how creative they found the drawing on a 7-point Likert scale (each annotator saw at least one drawing from each condition). Lastly, annotators could flag offensive (or non-alien) drawings. Upon review, 34 flagged drawings were discarded by the authors. Each drawing was annotated by two workers. Disputes in annotation were resolved by the authors.

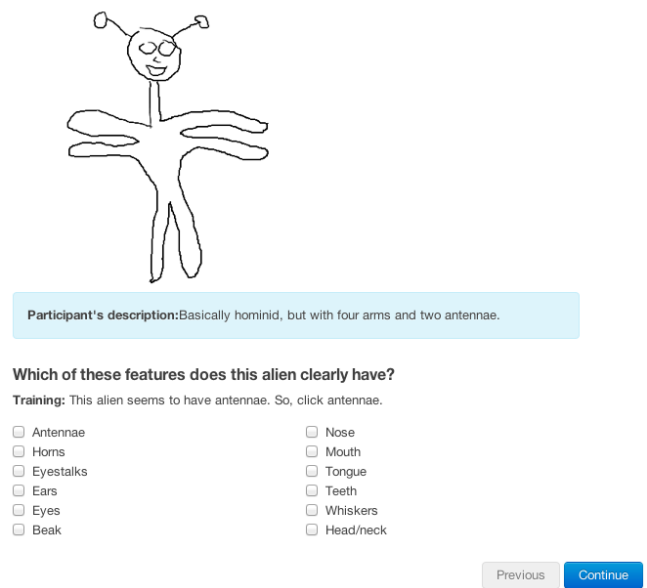


Figure 3: Training interface for annotators. The training interface shows what features to label (“click antennae”). The actual annotation is performed on an identical list of features.

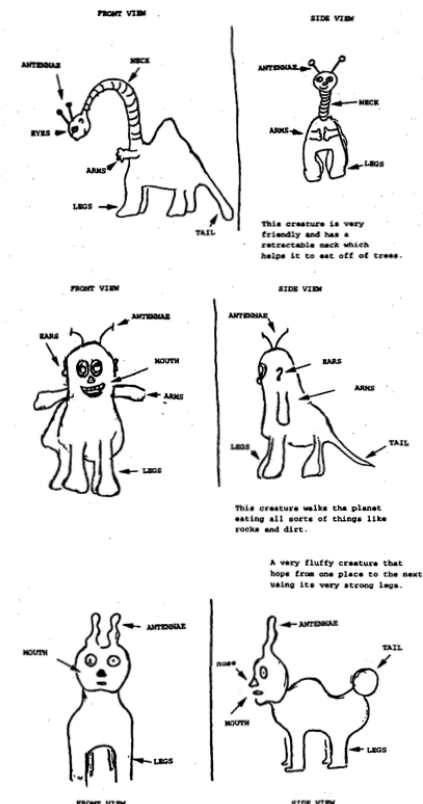


Figure 4: Example drawings provided to participants. All examples contain critical features—four legs, antennae, and a tail.

Condition	Critical	Common	Uncommon	Novel	Total	Drawings per session	Likert Rating
Control	0.39	4.21	0.95	0.47	6.03	4.00	3.71
Early	<b>0.57</b>	3.91	<i>1.15</i>	0.40	6.04	<b>3.00</b>	<b>4.10</b>
Late	<b>0.52</b>	3.82	0.78	0.45	5.57	<b>3.68</b>	3.43
Repeated	<b>0.64</b>	4.20	<b>1.21</b>	0.54	6.60	<b>3.00</b>	<b>4.22</b>

Table 2: Table of means. Means that differed from control at  $p < 0.05$  are **bold**, those marginally significant ( $p < 0.1$ ) are in *italics* (p-values from the post-hoc analysis using mixed models, see section Results).

## Results

We analyze data using a mixed-effects linear model. Since participants drew multiple drawings per drawing session, unless noted, we consider the participant as a random effect with a fixed intercept; and experimental condition, drawing session (first or second), and an interaction term as fixed-effects in all our analysis below. Reported p-values are from a Monte-Carlo (MCMC) simulation (Baayen et al., 2008).

### Examples increase conformity

Following Smith et al. (1993), conformity was measured as the number of critical features incorporated per drawing. Without controlling for the drawing session, examples shown at the start of the experiment increased the number of critical features that were incorporated into drawings ( $t(507) = 2.06$ ,  $p < 0.05$ ), consistent with results from (Smith et al., 1993; Marsh et al., 1996). Participants in the Late Examples condition show higher conformity in the second drawing session (*i.e.* post-exposure) [ $t(419) = 1.83$ ,  $p = 0.07$ ].

### Early exposure increases uncommon features

The number of uncommon features per drawing increased in the Early Examples condition ( $t(419) = 1.61$ ,  $p = 0.06$ ), and in the Repeated Examples condition ( $t(419) = 1.72$ ,  $p < 0.05$ ), but not in the Late Examples condition ( $t(419) = -0.45$ ,  $p = 0.649$ ) (Figure 5). The number of novel features did not vary significantly across condition. Participants in the Late exposure condition created drawings with marginally fewer common features ( $t(419) = -1.33$ ,  $p = 0.09$ ) and fewer total number of features ( $t(419) = -1.30$ ,  $p = 0.09$ ).

### Early and Repeated exposure leads to higher subjective ratings

Annotators rated drawings in the Early Examples and the Repeated Examples conditions higher ( $t = 2.24$ ,  $p < 0.05$  and  $t = 2.65$ ,  $p < 0.01$ , respectively). Intra-class correlation amongst raters (average, random raters) was 0.54 ( $F(508, 508) = 2.2$ ,  $p < 0.001$ ).

### Examples reduce number of drawings

Unlike Marsh et al. (1996), participants created fewer drawings per session in all example conditions<sup>1</sup> [Early:  $t(149) = -2.50$ ,  $p < 0.05$ ; Late:  $t(149) = -2.14$ ,  $p < 0.05$ , Repeated:

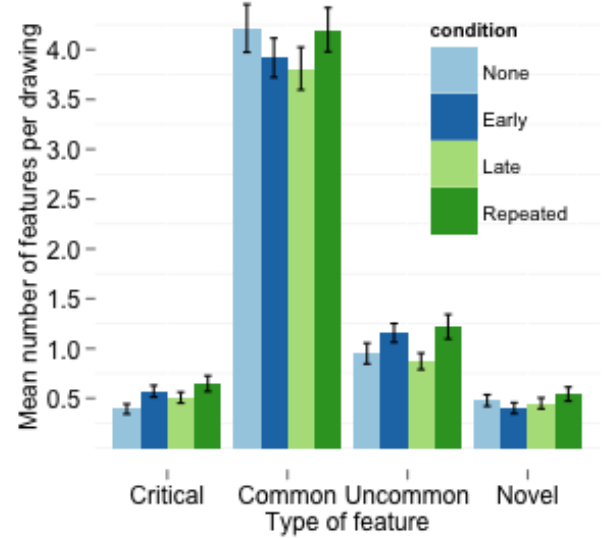


Figure 5: Participants in early and repeated exposure conditions included more uncommon features compared to Late exposure/control conditions.

$t(149) = -2.63$ ,  $p < 0.05$ ] (Figure 6). Participants in the Late Examples condition created fewer drawings after exposure to examples ( $\mu_{before} = 4.10$ ,  $\mu_{after} = 3.13$ ,  $t(149) = 1.91$ ,  $p_{interaction} < 0.05$ ).

## Discussion

### Example timing affects creative output

These results suggest that exposure to examples at any time increases conformity. However, early exposure increases the number of uncommon features and subjective ratings of creativity, while late exposure provides no such benefits. This runs counter to both our initial hypothesis and the intuitions of many designers who delay looking at examples in an effort to reduce fixation and think “out of the box” (Jansson & Smith, 1991).

One possible explanation for these effects is that early exposure to examples aids the designer in understanding the scope of acceptable solutions to a problem, and helps form an initial representation of the creative concept (Heit, 1992). Prototyping results in subsequent abstraction and refinement of the initial representation (Lim et al., 2008). Without initial exposure to examples, the refined representation may dif-

<sup>1</sup>Since the number of drawings is not a repeated measure, analysis uses a fixed-effects model with interaction, the experimental condition and the type of session being independent variables.

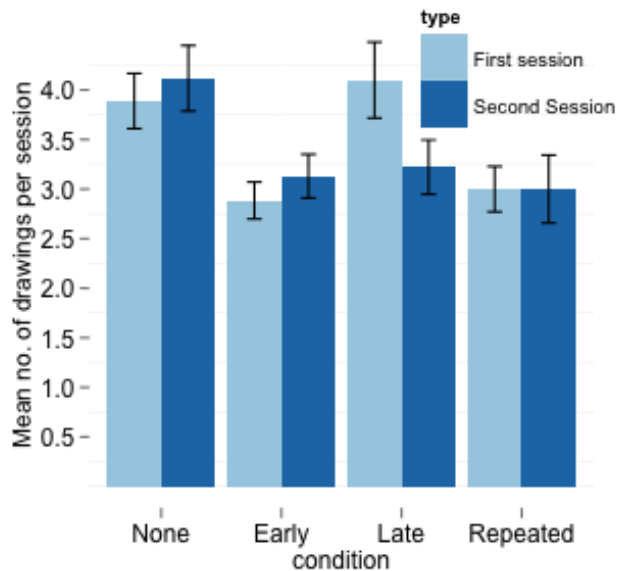


Figure 6: Participants drew fewer drawings when examples were shown (in the Late examples condition, participants drew fewer drawings in the second session).

fer widely from the one embodied in examples, which would make it harder to map concepts from the example to one's own representation. When exposure is only for a short duration (90s in our experiment), it is possible that only concepts with high enough activation, such as critical features in our experiment, are transferred (motivated by Boroditsky, 2007).

Another counter-intuitive experimental result is that repeated exposure to the same examples led to higher creative quality. This may also be explained by a seeding-and-transfer account. Initial exposure to examples prevents the refined representation formed by prototyping from diverging greatly from the one embodied in the examples. This refined yet similar representation would then allow the designer to learn different concepts on re-exposure to the same example.

In essence, the crucial ingredient that allows repeated exposure to improve creativity might be the prototyping that occurs between exposures.

### Why did examples yield fewer drawings?

Examples play a dual role in design— first, they inspire different solutions and ways of thinking. Second, they help form expectations about what characteristics a solution needs to have (Herring et al., 2009). The decrease in the number of drawings created may be due to this second role. Seeing examples may have signaled a higher threshold for “acceptable” drawings, resulting in participants spending more time on each drawing, and creating fewer drawings overall.

Our data suggest that this expectation-setting role has a different behavior than the inspirational role. While the number of drawings created decreased nearly uniformly post-exposure, changes in creativity measures (uncommon features and subjective ratings) were non-uniform. Therefore, while examples may set expectations any time they are pre-

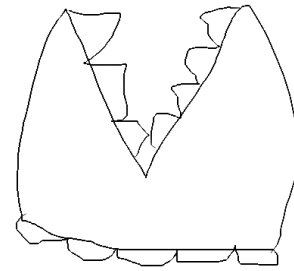


Figure 7: Participant-provided description: “An ambush predator that does move very much, but lures prey into its mouth using scent to make them think there is food there. It only occasionally shifts using the pads on its bottom, which can also suck up nutrients from the ground or water for emergencies.” Rated highly creative by our raters, this drawing has no novel or uncommon visual features, and uses a non-visual feature (scent).

sented (including late in the design process), their inspirational value may be time-dependent.

### Multiple Measures of Creativity

Results from both the feature-counting measure of creativity from prior work and the Likert-scale ratings provided by annotators are largely consistent. While the Likert ratings are subjective, they better capture the creativity in some drawings that combine common or critical features in a novel way, or use a non-visual feature (for *e.g.* see Figure 7). Using both together provides a better characterization of creativity.

### Conclusions and future work

This work demonstrates the benefits of early and repeated exposure to examples on creative work. In addition, it suggests that conformity may be the price one pays for these gains, regardless of when examples are seen. Hopefully, these results will encourage designers to seek examples early and often in the design process, when they are most useful.

This experiment also demonstrates a replication (and extension) of creativity studies in an online crowdsourced environment. Crowdsourced experiments often offer a lower cost, have a faster time to completion, and provide access to wider populations (Heer & Bostock, 2010). This paper's experiment and the labeling tasks took one week on Amazon Mechanical Turk. This was possible because the labeling scheme from Marsh et al. (1996) provided this study with a clear taxonomy of features that could be easily labeled by non-experts. We suggest crowdsourcing as a viable platform both for experiments that do not need participants with specialized skills or background (or modification per participant) and for analysis/labeling tasks that easily verifiable.

This work also raises a number of questions. First, the results of the repeated-exposure experimental condition indicate that the processes of prototyping and learning from examples may be intertwined in a creative task. Further empir-

ical studies could characterize the precise nature of this interaction. Second, this work shows that repeated exposure to examples is beneficial. How does the frequency of (or interval between) such exposures affect this result? Third, designers often spend years acquiring skills and specific domain knowledge. How do such skills and knowledge affect their interaction with examples? Furthermore, similar to cross-cultural effects of prototyping (Kim & Hinds, 2012), are effects of examples different in different cultures? Finally, how can the results of this work inform the design of tools that support creative work?

### Acknowledgements

We thank the Hasso Plattner Design Thinking Research Program for supporting this work.

### References

- Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4), 390–412.
- Boroditsky, L. (2007). Comparison and the development of knowledge. *Cognition*, 102(1), 118–128.
- Buxton, B., & Buxton, W. (2007). *Sketching user experiences: getting the design right and the right design*. Morgan Kaufmann.
- Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conf on* (pp. 248–255).
- Dow, S., Fortuna, J., Schwartz, D., Altringer, B., Schwartz, D., & Klemmer, S. (2011). Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results. In *Proc of CHI: ACM conf. on human factors in computing systems* (pp. 2807–2816).
- Gentner, D., & Colhoun, J. (2010). Analogical processes in human thinking and learning. *Towards a Theory of Thinking*, 35–48.
- Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proc of CHI: ACM conf. on human factors in computing systems* (pp. 203–212).
- Heit, E. (1992). Categorization using chains of examples. *Cognitive Psychology*, 24(3), 341–380.
- Herring, S., Chang, C., Krantzler, J., & Bailey, B. (2009). Getting inspired!: understanding how and why examples are used in creative design practice. In *Proc of CHI: ACM conf. on human factors in computing systems* (pp. 87–96).
- Jansson, D., & Smith, S. (1991). Design fixation. *Design Studies*, 12(1), 3–11.
- Kerne, A., Koh, E., Smith, S., Webb, A., & Dworaczyk, B. (2008). combinformation: Mixed-initiative composition of image and text surrogates promotes information discovery. *ACM Trans on Information Systems (TOIS)*, 27(1), 5.
- Kim, H., & Hinds, P. (2012). Harmony vs. disruption: The effect of iterative prototyping on teams creative processes and outcomes in the west and the east. In *Proc. icic: International conf. on intercultural collaboration*. ACM.
- Lee, B., Srivastava, S., Kumar, R., Brafman, R., & Klemmer, S. (2010). Designing with interactive example galleries. In *Proc of CHI: ACM conf. on human factors in computing systems* (pp. 2257–2266).
- Lewis, S., Dontcheva, M., & Gerber, E. (2011). Affective computational priming and creativity. In *Proc of CHI: ACM conf. on human factors in computing systems* (pp. 735–744).
- Lim, Y., Stolterman, E., & Tenenberg, J. (2008). The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 15(2), 7.
- Marsh, R., & Bower, G. (1993). Eliciting cryptomnesia: Unconscious plagiarism in a puzzle task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(3), 673.
- Marsh, R., Landau, J., & Hicks, J. (1996). How examples may (and may not) constrain creativity. *Memory & Cognition*.
- Newman, M., & Landay, J. (2000). Sitemaps, storyboards, and specifications: a sketch of web site design practice. In *Proc of DIS: ACM conf. on designing interactive systems* (pp. 263–274).
- Ritchie, D., Kejriwal, A., & Klemmer, S. (2011). d. tour: style-based exploration of design example galleries. In *Proc. of UIST: ACM symposium on user interface software and technology* (pp. 165–174).
- Schön, D. (1985). *The design studio: An exploration of its traditions and potentials*. RIBA Publications for RIBA Building Industry Trust London.
- Smith, S., Ward, T., & Schumacher, J. (1993). Constraining effects of examples in a creative generation task. *Memory & Cognition*, 21(6), 837–845.
- Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proc of the conf. on empirical methods in natural language processing* (pp. 254–263).
- Suwa, M., & Tversky, B. (1997). What do architects and students perceive in their design sketches? a protocol analysis. *Design studies*, 18(4), 385–403.
- Ward, T. (1994). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology*, 27(1), 1–40.
- Weisberg, R. (1999). Creativity and knowledge: A challenge to theories. *Handbook of creativity*, 226.
- Yu, L., & Nickerson, J. (2011). Cooks or cobblers?: crowd creativity through combination. In *Proc of CHI: ACM conf. on human factors in computing systems* (pp. 1393–1402).



# Role of Error Monitoring Mechanisms in Attribution of Sense of Self-Agency

**NeerajKumar, Jaison A. Manjaly (neeraj, jmanjaly@iitgn.ac.in)**

Cognitive Science Programme, Indian Institute of Technology Gandhinagar

VGEC Campus, Chandkheda, Ahmedabad-382424, INDIA

**K. P. Miyapuram(krishna.miyapuram@cantab.net)**

Center for Mind/Brain Sciences, Università degli Studi di Trento

Via delle Regole 101, 38060 Mattarello (TN), ITALY

## Abstract

Sense of agency refers to the sense of authorship of a given action. While the phenomenon seems too obvious to demand further investigation, pathological conditions such as delusions of control suggest the requirement of further investigation into the phenomenon of sense of agency. The traditional view point regarding the role of intention in sense of agency is complemented by computational models of motor control. Accordingly, we hypothesized and tested the role of error monitoring mechanisms in the sense of agency by manipulating the feedback given independent of the responses of the participant while performing a Flanker task. The results point out the potential role of error monitoring mechanisms by modulating the forward model predictions to experience a sense of agency for unintended actions.

**Keywords:** sense of agency; motor control; forward model; error monitoring mechanisms; Flanker task; error feedback; action intention.

## Introduction

In day-to-day life, we encounter various sensorimotor events, in which sensory perception and action are intertwined. A chain of events including intention to act, movement preparation, generating motor commands and sensory feedback are part of the underlying components of our sensorimotor experience (Haggard et al., 2002). The sense of agency for a given action refers to the sense that an agent has that s/he is the author of that action (Pacherie, 2007b). In other words, the sense of agency is a pre-reflective experience which enables the sense of authorship of one's own thoughts and actions. The sense of agency also critically contributes to a sense of self in terms of experiential immediacy (Tsakiris & Haggard, 2005). Though actions are found to be accompanied by a sense of agency, not much is known in terms of its underlying mechanism or set of processes responsible for this experience.

Different prevailing views have explained regarding the mediators between sense of agency and action (see David et al., 2008 for a review). One view proposed by Haggard et al. (2002) is that the intention of an agent, i.e. "intentional binding" has an important role on the sense of agency (Haggard et al., 2003; Tsakiris & Haggard, 2003), that contributes significantly to action awareness. Pacherie (2007a, b), on the other hand, claims that the sense of agency contains not only an experience of intentional causation, but a sense of initiation and control. Disturbance

in any phase can cause disruption in the sense of agency. Yet another view is based on established models of motor control (Wolpert, 1997) as explained below.

Computational models of motor control (Blakemore et al., 2001, 2002) have an alternative suggestion about the mechanisms responsible for the sense of agency. The computational view is that the sensorimotor loop consists of an inverse model, which identifies the motor commands required to achieve a certain desired state and a forward model, which predicts the sensory consequences of motor actions. These models are represented within the central nervous system in the form of internal models. According to Frith (1992), internal forward model is principally responsible for the sense of agency because it generates an efference copy, which predicts the sensory consequences of motor commands in advance. Predicted sensory information is matched against subsequent sensory information. If predicted and sensed information match, then the sensory events are self-generated, and the subject will experience sense of agency for those events. If there is mismatch, then the sensory information registers it as an external event, and therefore the sense of agency is absent. This model has been used to explain the perceptual attenuation of self-generated stimuli (Blakemore et al., 2000), and pathological experiences, such as delusions of control found in Schizophrenia. For example, Blakemore et al. (2002) have suggested that the misattribution of action shown in patients experiencing delusions of control can be explained by a deficit in the internal forward model (Frith, 1992).

Simulation theory of agency suggests that in understanding or predicting other's action we use our own experiences to simulate those of others (Goldman, 1989). Sebanz et al. (2005) found that subjects within the autism spectrum did not show deficits in representing another person's action but exhibited mentalizing deficits. David et al. (2007) showed autistic subjects show deficits in perspective taking which has been explicitly linked to simulation (Gallese & Goldman, 1998; Langdon & Coltheart, 2001). Children with autism have also shown reduced error monitoring (Vlamings et al., 2008) and altered cerebellar feedback projection (Catani et al., 2008).

Another explanation for experienced agency for unintentional actions is suggested by recent studies on error

monitoring mechanisms (Yordanova et al., 2004; Van Schie et al., 2004). It is well known that after an erroneous action is selected, internal monitoring mechanisms give the feedback that one has committed an error. Such error signals are based on the detection of a conflict that occurred while choosing between several action alternatives, rather than on the comparison between the predicted and actual consequences of a specific action selected for execution. Agency for erroneous actions could be experienced because an error-monitoring signal is used to readjust the system. The readjustment could serve as a direct indication of agency or it could influence post-hoc evaluations of performed actions (Knoblich & Natalie, 2005).

Experiments by Sato and Yasuda (2005) suggest that motor prediction contributes to the experience of agency. Their findings show that agency is experienced not only for intended, but also for erroneous/unintended actions. This result supports the view that the experience of agency depends on the discrepancy between predicted and actual sensory consequences regardless of whether an action was intended or unintended (Fournier et al., 1998; Sato & Yasuda, 2005).

Present study aims to investigate the role of error monitoring mechanism in the attribution of sense of agency. We propose that error monitoring mechanisms can update the ‘forward model’ efferent prediction to match the actual sensory outcome. The activation of error monitoring mechanisms can cause the online or real time alteration in predictions by forward model. This modulation in prediction can occur before assimilating the actual sensory outcome which can influence the sense of agency. We hypothesized that the feedback should modulate the sense of agency, more particularly when the feedback is inconsistent with the actual responses of the participant. We argue that, this online modulation of forward model prediction through error monitoring mechanisms can explain the sense of agency in unintentional action.

## Materials and Methods

### Participants

15 undergraduate students (Mean age = 19.4 years) have participated in the study. All participants were right-handed having normal or corrected to normal vision. Participants provided informed consent and were paid for participation.

### Design

We have used some of the designs proposed by Sato & Yasuda (2005) to examine the role of error monitoring mechanism and to explore the possibility of modulation of the forward model predictions. Participants performed an Eriksen Flanker task (Eriksen & Eriksen, 1974), which is a forced choice reaction time task in which the target letter is flanked by distracter letters (either congruent or incongruent

letter). When participants are asked to respond quickly to the flanker array, they tend to make errors and have low reliability on their responses. To activate error monitoring mechanisms, we have given an immediate feedback after the response which could be right (correct feedback) or wrong (incorrect feedback). If wrong feedback can alter the attribution of agency (refer table 1) then it might be possible that error monitoring mechanisms are capable to modulate the efferent predictions of forward model before assimilating the actual sensory outcome.

Thus, the experiment consisted of two within-subjects factors (a) Type of sensory outcome (Congruent tone or incongruent tone with prediction), and (b) type of feedback (Wrong feedback & Right feedback). Notably the error feedback was manipulated independent of the actual response.

Table 1: Prediction of sense of agency in different conditions.

Response → Condition ↓	Correct	Incorrect
Wrong Feedback-Congruent Tone	No	Yes
Wrong Feedback-Incongruent Tone	Yes	No
Right Feedback-Congruent Tone	Yes	No
Right Feedback-Incongruent Tone	No	Yes

### Procedure

Upon entering the laboratory, participants were seated in front of a computer screen with a pair of headphones. Prior to the experiment, participants performed 300 learning trials. On each trial, 1000ms after fixation onset, the target stimulus (i.e., “H” or “N”) was presented for 250ms on the center of the screen. Participants were told to press the left button with the left index finger as quickly and accurately as possible whenever an “H” appeared on the center of the screen and the right button with the right index finger whenever an “N” appeared on the screen. After each button press, a certain tone was immediately presented for 200ms through in-ear headphones: a 600 Hz tone or a 1000 Hz tone. The assignment of stimuli and tones to buttons was consistent for each participant and counterbalanced across participants. Participants were explicitly told that each button pressing would evoke a certain tone. Tones were identical in duration and sound pressure throughout the experiment.

In the main experiment, participants performed the Flanker task for 200 trials. Each trial started with the onset of centrally presented fixation sign. After the 1000ms of fixation onset, a five-letter array (i.e., HHHHH, NNNNN, HHNHH, or NNHNN) was presented for 250ms. Participants were instructed to respond to one of the two target letter (central H or N) with one finger and to the other letter with the other finger as quickly as possible and not to correct their responses even if they made errors. The assignment of responding finger to target letter was the

same as the learning session. After the response was made, an immediate feedback was provided on the screen for 200ms which could be right (congruent with response) or wrong (incongruent with response). After 200ms of offset of feedback a tone was presented through headphone for 200ms either congruent or incongruent with prediction. Then participants were asked for their rating regarding the sense of agency using a question “I was the one who produced the tone”. The responses could be one of the three options in the form of ‘Yes’, ‘No’ and “Maybe”. To prevent demand effects and any other possible biases in responses such as motor preparation, the question was randomly alternated with a second question “I was the one who was listening to the tone”. Further, the options Yes and No were counterbalanced across trials. The experiment was designed using Psychophysics toolbox (Brainard, 1997) in MATLAB (Mathworks Inc.) (See figure 1).

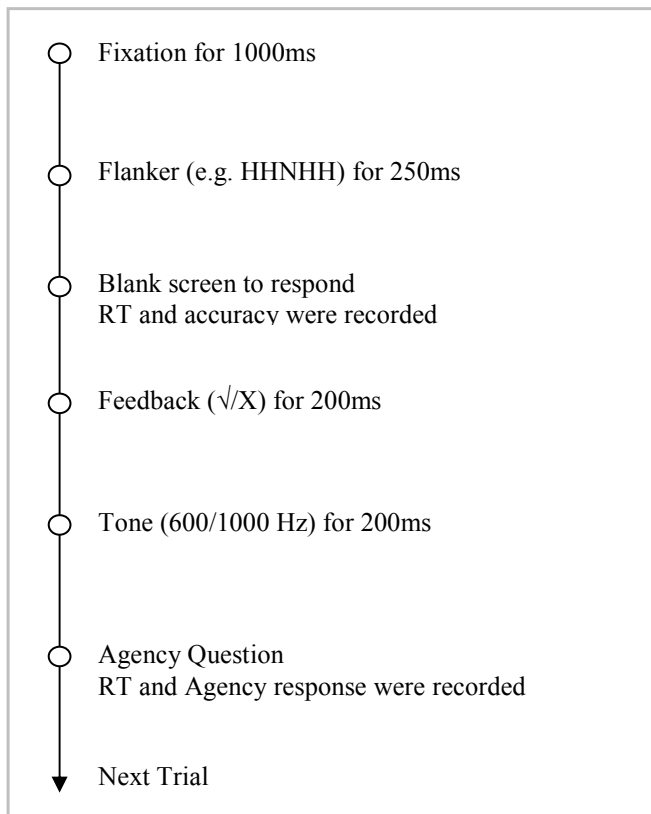


Figure 1: Trial procedure for an experimental trial

## Data Analysis

We present the results for the sense of agency, as this is the main hypothesis of our experiment. We have also analyzed the reaction time data to check our manipulation effect of flanker task. For quantitative analysis we have transformed the ratings into numerical values. Now ‘Yes’ is represented as 1, ‘No’ is represented as 0 and ‘Maybe’ lies in between as 0.5. These ratings of sense of agency were analyzed using 2x2 repeated measures ANOVA separately for correct and incorrect response trials (see results).

## Results

To check the manipulation effect of flanker task, we have analyzed the reaction time to target letter for correct and incorrect trials in both practice and experimental session. We have found that in practice session participants were significantly faster [ $t(14) = 2.97, p = 0.01$ ] in incorrect responses [Mean = 0.32 Sec, SD = 0.15] than correct responses [Mean = 0.53 Sec, SD = 0.18]. In experimental session, participants were significantly slower [ $t(14) = 3.34, p > 0.01$ ] when they made incorrect responses [Mean = 1.3 Sec, SD = 0.5] in comparison to correct responses [Mean = 0.75 Sec, SD = 0.06]. These results suggest that participants face internal conflict between various alternatives of actions in experimental session (target letter was flanked by congruent/incongruent letters) which in turn delayed the response and end up in incorrect action.

The rating scores on sense of agency were analyzed separately for correct-response trials and incorrect-response trials using repeated measure analysis of variance with two factors: Tone congruency (2 levels – Congruent & Incongruent Tone) X Type of feedback (2 levels – Wrong & Right Feedback). For correct responses, this analysis revealed the main effect of tone congruency [ $F(1,14)=12.86, p < 0.01$ ], but there was no significant main effect of feedback [ $F(1,14)=1.02, p=0.32$ ]. More crucially, we have found significant interaction between tone congruency and type of feedback [ $F(1,14)=353.0, p < 0.01$ ]. Further post-hoc analysis revealed that under congruent tone condition sense of self-agency was significantly reduced ( $p < 0.01$ ) in wrong feedback condition (Mean=0.25, SD=0.08) in compare to Right feedback condition (Mean=0.81, SD=0.04). It also revealed that sense of self-agency was significantly increased ( $p < 0.01$ ) under wrong feedback condition (Mean=0.68, SD=0.10) than in right feedback condition (Mean=0.18, SD=0.15) when tone was incongruent with prediction (See figure 2).

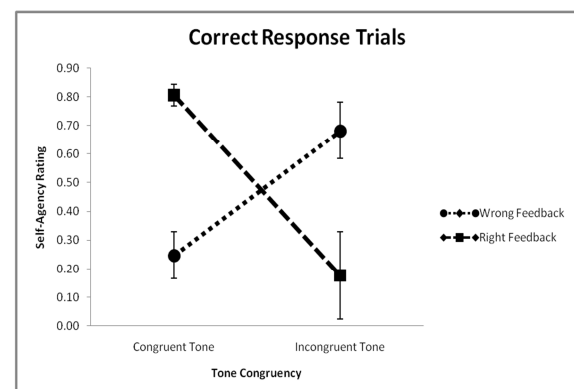


Figure 2: Sense of self-agency for correct responses

For incorrect response, repeated measure analysis of variance revealed that neither tone congruency [ $F(1,14)=0.12, p=0.73$ ] nor feedback [ $F(1,14)=2.6, p=0.12$ ] had a significant main effect on rating of sense of agency.

But there was significant interaction between tone congruency and type of feedback [ $F(1,14)=35.40$ ,  $p<0.01$ ]. Post-hoc analysis revealed that in the Congruent tone condition sense of self-agency was significantly increased in wrong feedback condition (Mean=0.63, SD=0.41) from sense of self-agency in right feedback condition (Mean=0.24, SD=0.16). It also shown that under the Incongruent tone condition, sense of self-agency was reduced when wrong feedback (Mean=0.11, SD=0.12) was given instead of right feedback (Mean=0.75, SD=0.31) (See figure 3).

Post-hoc analysis also has shown the magnitude of manipulation effect. For Correct responses, reduction of sense of self-agency in Wrong feedback – Congruent tone condition is up to level of self-agency in Right feedback – Incongruent tone ( $p=0.15$ ), but manipulation effect is not that much strong in Wrong feedback – Incongruent tone condition because sense of self agency in this condition is significantly lesser from self-agency in Right feedback – Congruent tone condition ( $p<0.01$ ).

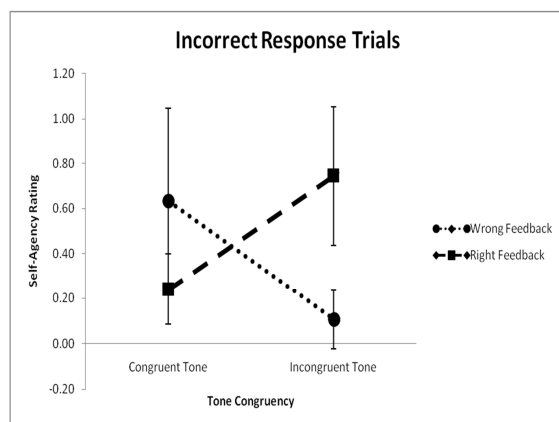


Figure 3: Sense of self-agency for incorrect responses.

These results point towards the potential role of interaction between the feedback and tone congruency on sense of agency.

## Discussion

In this study, we manipulated the effect of error monitoring mechanisms on sense of agency. Participants performed a Flanker task and the error feedback was manipulated orthogonal to the actual responses. Accordingly, we found a differential but consistent behavior between the correct and incorrect responses. For both the correct and incorrect responses, a significant interaction was observed for tone congruency and type of feedback. Consistent with our hypothesis, when participant made a correct choice but the feedback was falsely incorrect, the sense of self-agency increased. In contrast, when participant made an incorrect choice and the feedback was falsely correct, the sense of self-agency decreased. However, a

main effect for tone congruency was observed only for correct but not for incorrect responses. These results point out the role of error monitoring mechanisms in attribution of sense of agency.

Previous studies have suggested (Blakemore et al., 2001, 2002) the profound role of forward model prediction in the attribution of agency to self or an external agent. Forward model predictions are based on motor program which in turn are based on our intentions guided by motor planning. However, forward model of attribution of agency is not able to explain the self-agency attribution for unintended actions – actions which are not congruent with the intentions or motor plan. This incongruency in action could occur because of multiple causes such as (a) action slips due to internal noise (Wolpert & Ghahramani, 2000) in the transformation from motor signal to actual action, (b) the conflict in choosing an action between various alternatives as in a forced choice task. Unintentional action through internal noise or conflict activates the error monitoring mechanisms, and this activation of error monitoring mechanism can play a significant role in attribution of agency by modulating the forward model predictions in real time.

We have found that by activating error monitoring mechanisms by external feedback after the action (but before the actual sensory outcome) alters the attribution of agency. For the correct responses, when forward model prediction was congruent with actual sensory consequences, wrong/falsely incorrect feedback (a cross sign) after the action causes the attribution of agency to an external agent. Similarly, when prediction and actual sensory consequences were not congruent, participants tend to attribute the agency to themselves when their actions were followed by wrong/falsely correct feedback. In contrast, when participant made an incorrect response and the feedback was wrong/falsely correct, the sense of self-agency get decreased when consequence of the action is also incongruent with the actual sensory consequences. Similarly, when prediction and actual sensory consequence were congruent, sense of self-agency get increased when actions were followed by wrong/falsely correct feedback.

Previous experiments on error monitoring mechanisms show that cerebellum is involved in error monitoring and optimizing the system (Kawato & Gomi, 1992; Dreher & Grafman, 2002; Menon et al., 2001). Cerebellum is also responsible for predicting the sensory consequences of action (Blakemore et al., 2001). Synofzik et al. (2008) showed that in a motor learning task cerebellum updates predictions about the visual consequences of one's behavior. These findings suggest that error monitoring mechanisms play more profound role in generating sense of agency than just influencing the post-hoc evaluation or serving as direct indicator of agency. Our findings are also in support with the previous findings that cerebellum updates the

predictions of forward model. Since, the sensory consequences of actions vary as a result of changes of the effector's efficacy, internal predictions need to be updated continuously, our findings suggest that cerebellum as an error monitoring mechanism serves as a device which updates predictions of forward model in real time before assimilating any actual sensory consequences.

Alternatively, our results can also be explained by simulation approach to sense of agency. In this line, alteration in attribution of agency can be the result of activation of error monitoring mechanisms which can be used as a signal to navigate within shared representations. Our results also suggest that misattribution of agency in schizophrenia may not be based on imprecise predictions (Synofzik et al., 2010); they misattribute the agency might be because of failure in real time update in predictions by forward model.

These results of alteration in attribution of agency by activating error monitoring mechanisms suggest that forward model predictions are being modulated in real time as soon as they come to know that performed action was not correct. This modulation in prediction before actual sensory consequences helps them to experience a sense of agency for unintended/erroneous actions.

## Acknowledgments

This research was funded by Neotia Foundation, Kolkata & Indian Institute of Technology Gandhinagar.

## References

- Blakemore, S. J., Frith, C. D., & Wolpert, D. M. (2001). The cerebellum is involved in predicting the sensory consequences of action. *Neuroreport*, 12(9), 1879–1884.
- Blakemore, S. J., Smith, J., Steel, R., Johnstone, C. E., & Frith, C. D. (2000). The perception of self-produced sensory stimuli in patients with auditory hallucinations and passivity experiences: Evidence for a breakdown in self-monitoring. *Psychological Medicine*, 30(5), 1131–1139.
- Blakemore, S. J., Wolpert, D. M., & Frith, C. D. (2002). Abnormalities in the awareness of action. *Trends in Cognitive Sciences*, 6(6), 237–242.
- Brainard, D.H. (1997). The Psychophysics Toolbox, *Spatial Vision*, 10, 443–446.
- Catani, M., Jones, D.K., Daly, E., Embiricos, N., Deeley, Q., Pugliese, L., et al. (2008). Altered cerebellar feedback projections in Asperger syndrome. *Neuroimage*, 41(4), 1184–1191.
- David, N., Newen, A., & Vogeley, K. (2008). The "sense of agency" and its underlying cognitive and neural mechanisms. *Conscious Cogn*, 17, 2:523–34.
- David, N., Gawronski, A., Santos, N.S., Huff, W., Lehnhardt, F. G., Newen, A., et al. (2007). Dissociation between key processes of social cognition in autism: impaired mentalizing but intact sense of agency. *Journal of Autism and Developmental Disorders*, 38(4), 593–605.
- Drehr, J., & Grafman, J. (2002). The role of the cerebellum and basal ganglia in timing and error prediction. *European Journal of Neuroscience*, 16, 1609–1619.
- Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a non-search task. *Perception and Psychophysics*, 16, 143–149.
- Fournier, P., & Jeannerod, M. (1998). Limited conscious monitoring of motor performance in normal subjects. *Neuropsychologia*, 36(11), 1133–1140.
- Frith, C. D. (1992). *The cognitive neuropsychology of schizophrenia*. Hove, U.K.: Lawrence Erlbaum.
- Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2(2), 493–501.
- Goldman, A. I. (1989). Interpretation psychologized. *Mind & Language*, 4, 161–185.
- Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, 5(4), 382–385.
- Haggard, P., & Clark, S. (2003). Intentional action: Conscious experience and neural prediction. *Consciousness and Cognition*, 12, 695–707.
- Kawato, M., & Gomi, H. (1992). A computational model of four regions of the cerebellum based on feedback-error learning. *Biological cybernetics*, 68, 95–103.
- Knoblich, G., & Natalie S. (2005). Agency in the face of errors. *Trends in cognitive science*, 9(6), 259–261..
- Langdon, R., & Coltheart, M. (2001). Visual perspective-taking and schizotypy: Evidence for a simulation-based account of mentalizing in normal adults. *Cognition*, 82(1), 1–26.
- Menon, V., Adelman, N.E., White, C.D., Glover, G.H., & Reiss, A.L. (2001). Error-related brain activation during a Go/NoGo response inhibition task. *Human Brain Mapping*, 12, 131–143.
- Pacherie, E. (2007a). Towards a dynamic theory of intentions. In S. Pockett, W.P. Blanks, & S. Gallagher (Eds.), *Does consciousness cause behavior? An investigation of the nature of volition*. Cambridge, MA: MIT Press.
- Pacherie, E. (2007b). The sense of control and sense of agency. *Psyche*, 13(1).
- Sato, A., & Yasuda, A. (2005). Illusion of sense of self-agency: Discrepancy between the predicted and actual sensory consequences of actions modulates the sense of self-agency, but not the sense of self-ownership. *Cognition*, 94(3), 241–255.
- Sebanz, N., Knoblich, G., Stumpf, L., & Prinz, W. (2005). Far from action-blind: Representation of others' actions in individuals with autism. *Cognitive Neuropsychology*, 22, 433–454.
- Synofzik, M., Lindner, A., Their, P. (2008). The cerebellum updates predictions about the visual consequences of one's behavior. *Current Biology*, 18, 814–818.

- Synofzik, M., Their, P., Leube, D.T., Schlotterbeck, P., Lindner, A. (2010). Misattributions of agency in schizophrenia are based on imprecise predictions about the sensory consequences of one's actions. *Brain*, 133, 262-271.
- Tsakiris, M., & Haggard, P. (2003). Awareness of somatic events associated with a voluntary action. *Experimental Brain Research*, 149(4), 439–446.
- Tsakiris, M., & Haggard, P. (2005). Experimenting with the acting self. *Cognitive Neuropsychology*, 22, 387–407.
- Tsakiris, M., Haggard, P., Franck, N., Mainy, N., & Sirigu, A. (2005). A specific role for efferent information in self-recognition. *Cognition*, 96(3), 215–231.
- Van Schie, H.T. et al. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience*, 7, 549–554.
- Vlamings, P.H.J.M., Jonkman, L.M., Hoeksma, M.R., Engeland, H.V., Kemner, C. (2008). Reduced error monitoring in children with autism spectrum disorder: an ERP study. *European Journal of Neuroscience*, 28(2), 399–406.
- Wolpert, D. M. (1997). Computational approaches to motor control. *Trends in cognitive science*, 1(6), 209-16.
- Wolpert, D. M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, 3, 1212–1217.
- Yordanova, J., Falkenstein, M., Hohnsbein, J., Kuleva, V. (2004). Parallel systems of error processing in the brain. *Neuroimage*, 22, 590–602.

# Pragmatic interpretation of contrastive prosody: *It looks like* speech adaptation

Chigusa Kurumada

kurumada@stanford.edu

Dept. of Linguistics

Stanford University

Meredith Brown

mbrown@bcs.rochester.edu

Dept. of Brain and Cognitive Sciences

University of Rochester

Michael K. Tanenhaus

mtan@bcs.rochester.edu

Dept. of Brain and Cognitive Sciences

University of Rochester

## Abstract

Drawing on insights from recent work on phonetic adaptation, we examined how listeners interpret prosodic cues to two opposing pragmatic meanings of the phrase “It looks like an X” (e.g., “It looks like a zebra (and it is one)” and “It LOOKS like a zebra (but its actually not)”. After establishing that different prosodic contours map onto these meanings (Experiment 1), we demonstrated that prosodic interpretation is shifted by inclusion of another alternative (Experiment 2); the reliability a speaker’s use of prosody to signal pragmatic alternatives (Experiment 3); and most importantly by the distribution of cue values along a continua (Experiment 4). We conclude that listeners derive linguistically meaningful categories from highly variable prosodic cues through rational inference about assumptions that are shared in the conversational context and adaptation to distributional characteristics of prosodic cues.

**Keywords:** Prosody, contrastive accent, Gricean pragmatics, speech adaptation, rational inference

## Introduction

In a famous scene in the movie *Taxi Driver*, Robert DeNiros character repeatedly utters, *You talkin’ to me*. As he changes pitch contours, the intended meaning of the utterance shifts from a question to a challenge. As the example illustrates, prosody carries information about a speakers intentions. However, the acoustic features of prosodic alternatives, as well as the mappings between prosodic patterns and intended meanings vary considerably across speakers. For example, although rising boundary tones can distinguish between questions and assertions in many languages, many questions in fact do not end with a rising boundary tone. Also speakers who use “up-talk” often end assertions with a rising boundary tone. Likewise, a pitch accent preceded by an initial drop (fall-rise: often annotated as L+H\* in the ToBI convention (Silverman et al., 1992)) can signal the presence of contrast. However, characterizing the acoustic properties that signal this contrast in natural speech is far from straightforward. L+H\* and a simple rise (H\*) have overlapping interpretations that are highly dependent on utterance context (Itô & Speer, 2008; Watson, Tanenhaus, & Gunlogson, 2008).

How, then, do listeners navigate the lack of invariance in prosodic cues to pragmatic meaning? We propose that listeners solve the variability problem for prosody in the same way as they solve the variability problem for phonetic features, namely by *adaptation*. Just as prosodic contours vary according to both random and systematic factors, phonetic features of speech contain massive variability, which presents a challenge to listeners who are to derive discrete phonemic categories. It has been suggested that the speech perception system deals with the lack of invariance in two ways. One is to store separate, speaker-, group-, and context-specific representations of tokens from the same categories (Goldinger,

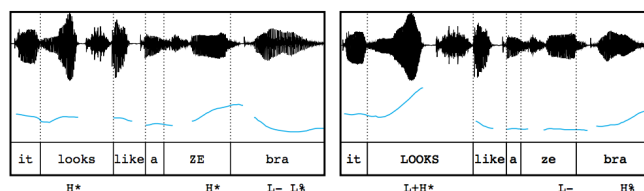


Figure 1: Waveforms (top) and pitch contours (bottom) of the utterance “It looks like a zebra”. The affirmative interpretation *It is a zebra* is typically conveyed by the pattern on the left, while the negative interpretation *It is not a zebra* is conveyed by the pattern on the right.

1998). The other is to adapt phonetic categories to the distributional characteristics of the acoustic input. For example, Clayards, Tanenhaus, Aslin, and Jacobs (2008) provided evidence that listeners adapt their perceptual categories according to the mean and the variance of a bimodal distribution along a VOT continuum (e.g. /b/-/p/).

Importantly, the way listeners integrate different kinds of information in speech perception is compatible with the hypothesis that they make rational inferences (Kleinschmidt & Jaeger, 2011). Listeners can weight different kinds of information according to their reliability, adjust phonetic categories based on more reliable information, and ignore deviation from the expected patterns when there is an ad-hoc source for the unfamiliar pronunciations (Kraljic, Brennan, & Samuel, 2008).

The current study draws on these insights to explore the hypothesis that listeners navigate prosodic heterogeneity by adapting their interpretations through rational inference. We focus on the construction “It looks like an X”, which can evoke different pragmatic meanings depending on its prosodic realization. A canonical accent placement (as illustrated in Figure 1, left panel, henceforth *noun-focus prosody*) typically elicits an affirmative interpretation (e.g. *It looks like a zebra and I think it is one*). When the verb “looks” is lengthened and emphasized with a contrastive accent (L+H\*) and the utterance ends with a L-H% boundary tone (Figure 1, right, *verb-focus prosody*), it can trigger a negative interpretation (e.g. *It LOOKS like a zebra but its actually not one*; see also Dennison & Schafer, 2010).

We explored the adaptation hypothesis in four rating experiments. Experiment 1 established that listeners systematically derive different pragmatic interpretations based on noun- and verb-focus prosodic contours. Experiments 2 and 3 demonstrated that pragmatic interpretations are systematic-



cally modulated by speaker-specific use of particular prosodic contours in different linguistic contexts and the reliability with which a speaker signals pragmatic contrasts prosodically. In Experiment 4, we exposed listeners to affirmative- and negative-interpretation tokens with different distributions of constituent duration and fundamental frequency ( $f_0$ ) values, sampled from a continuum of noun- and verb-focus prosodic contours. Consistent with the adaptation hypothesis, listeners' judgments shifted according to the distributional properties of the input. Taken together, our results provide novel evidence that listeners make optimal use of speaker and context-specific information to derive pragmatic meaning from contrastive prosody.

## Experiment 1

We elicited listeners' interpretations of "It looks like an X" in two types of rating tasks to establish that the proposed prosodic contours result in different pragmatic inferences.

### Methods

**Participants** We used an online crowd-sourcing platform (Amazon's Mechanical Turk) for the experiment. We posted 65 separate HITs (Human Intelligence Tasks: experimental tasks for participants to work on) and received 63 HITs from distinct individuals. Participants were all self-reported native speakers of American English. They received \$0.80 for completing the task and the mean task duration was 11 minutes.

**Stimuli** 24 imageable high-frequency nouns were embedded in the sentence frame "It looks like an X". Two tokens of each item with noun-focus and verb-focus prosodic patterns were recorded by a native speaker of American English.

**Procedure** Participants were first presented with a cover story in which a school teacher described animals and objects in an encyclopedia with pictures that were not directly accessible to his students. In response to a question from a child about what he saw on the page, the teacher said, "It looks like an X" (e.g., *It looks like a zebra*). The participants' task was to judge whether the teacher was referring to a picture of the target noun (e.g., a zebra) or something else.

For each item, participants first heard one of the two target prosodic patterns, and rated how likely it is that the teacher is looking at a picture of the target noun or a picture of something else. Likelihood was indicated using a 100-point scale (0 = something else, 100 = a picture of the noun referent). Next, they heard the same item produced with the other prosodic pattern and answered a 3-alternative forced choice question: If the teacher had used the second intonation pattern, the likelihood of the picture depicting an exemplar of X (e.g. a zebra) would be (a) greater, (b) the same, or (c) less.

### Results and Discussion

Figure 2a plots responses in the 100-point-scale rating task. Participants rated items with noun-focus prosody higher than those produced with the verb-focus prosody, indicating that they were more likely to derive the affirmative interpretation

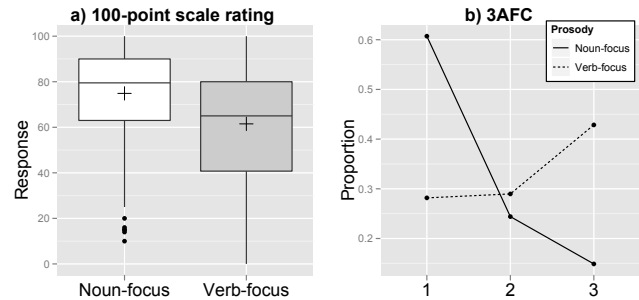


Figure 2: a) The likelihood estimation based on a 100-point scale by prosodic patterns. The crosses within the boxes indicate the mean values. b) Proportions of the responses given in the 3AFC questions [1 = MORE likely to be an X, 2 = no difference, 3 = LESS likely to be an X. The solid line and the dotted line represent the responses based on the noun-focus prosody and the verb-focus prosody, respectively.

(i.e., it is an X) based on the noun-focus prosody. A mixed effects regression analysis (Gelman & Hill, 2006) confirmed that the difference was statistically significant after controlling random effects of subjects and items ( $\beta = -13.38$ ,  $p < .001$ ). Notice, however, that mean values for both prosodic patterns (indicated by the crosses in Figure 2-a) were above 50%. Judgments were thus strongly biased towards the affirmative interpretation.

Figure 2b plots responses in the 3-alternative forced choice task. The difference between the prosodic patterns was significant ( $\beta = .41$ ,  $p < .001$ ): When participants first heard the verb-focus prosody, the noun-focus prosody was rated as more likely to refer to a denotation of the noun (61%) compared to 28% for the verb-focus when it followed the noun-focus prosody.

Overall, participants made distinct pragmatic interpretations based on the two prosodic patterns. However, the judgments were far from categorical and strongly biased towards the affirmative reading. Taking this as our point of departure, we evaluated the adaptation hypothesis by manipulating factors which we predicted would shift listeners' judgments.

## Experiment 2

The adaptation hypothesis posits that noun-focus and verb-focus prosodic contours are not directly mapped onto two distinct pragmatic meanings. Rather, these pragmatic interpretations are obtained through inference based on linguistic and non-linguistic information shared in a particular discourse context. Grice (1989) proposed that utterance meanings are derived through comparison among potential expressions that could have been used in the same situation. We hypothesized that "It looks like an X" would elicit more negative interpretations when the set of sentences produced by the speaker also included a stronger statement (e.g., *It is an X*), on the grounds that the speaker would have simply used this stronger statement to express the affirmative meaning.

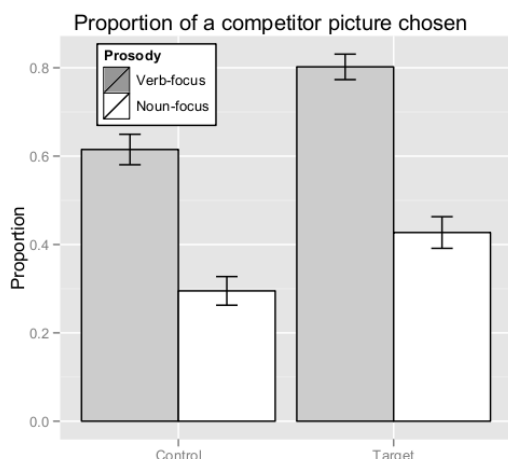


Figure 3: Proportions of competitor pictures chosen in the target and the control conditions in Experiment 2. Error bars represent standard error of the mean.

## Methods

**Participants** We posted 50 separate HITs and received 48 HITs from distinct individuals.

**Stimuli** An additional 24 stimuli in the form of “It is an X” (e.g. “It is a zebra”) were recorded by the same speaker as in Experiment 1. Two lists were created based on the items from Experiment 1 and these newly added items. In the *control condition*, all items were in the form of “It looks like an X”: 12 with verb-focus prosody and 12 with noun-focus prosody. In the *target condition*, 8 of the 12 noun-focus items were replaced by tokens of “It is an X”. Each participant was randomly assigned to the control or test condition.

**Procedure** Participants were presented with the same cover story as that in Experiment 1 and instructed to select an intended referent of an X out of two pictures: a *target picture* (e.g. a zebra) and a *competitor* (e.g. an okapi, a four-legged animal with black and white stripes only on its legs) after listening to each sentence. Participants completed 24 consecutive trials with no feedback.

## Results and Discussion

Analyses were conducted on the items that were common to the two conditions (i.e., we excluded responses to “It is an X” in the target condition and the corresponding sentences in the control condition). Figure 3 illustrates the proportion of a competitor picture chosen in each condition. A mixed logit regression (full factorial; maximum random effects (Jaeger, 2008)) confirmed that there were significant main effects of prosody ( $\beta = 1.9$ ,  $p < .0001$ ) and conditions (target vs. control,  $\beta = 1.18$ ,  $p = .061$ ) as well as a significant interaction term between them ( $\beta = 0.83$ ,  $p < .05$ ), with verb-focus prosody eliciting more competitor responses in both conditions.

Crucially, when the stronger statements were added, participants were more likely to select a competitor picture (indicating a negative interpretation) for both the noun-focus and verb-focus contours. This suggests that the pragmatic interpretation of prosody is not solely determined by the acoustic characteristics of utterances, consistent with the predictions of the adaptation hypothesis. Expectations based on context- and speaker-specific knowledge modulate pragmatic interpretations of contrastive prosody.

## Experiment 3

The adaptation hypothesis predicts that listeners are more likely to adapt to cues that are used reliably and systematically. Experiment 3 manipulates the overall reliability with which particular prosodic contours are associated with particular pragmatic meanings.

## Methods

**Participants** We posted 80 separate HITs and received 76 HITs from distinct individuals.

**Stimuli** 26 items of “It looks like an X” (16 training and 10 test items) were recorded in each of the two target prosodic patterns. For each of the training items, two continuation phrases were recorded to disambiguate the intended meaning. One continuation supported the affirmative interpretation (e.g., “It looks like a zebra because it has black and white stripes all over its body”). The other pattern supplied feedback confirming the negative interpretation (e.g., “It looks like a zebra but it’s not; it has stripes only on its legs”).

**Procedure** Participants were presented with a cover story in which a mother and a child were naming animals and objects in a picture-book. The exposure phase included 16 trials in which participants heard the mother tell the child, “It looks like an X”. Their task was to choose the likely referent of N between a target and competitor picture (e.g. a zebra and an okapi). They then heard a continuation phrase disambiguating the intended referent.

Each participant was randomly assigned to one of two conditions. In the *reliable-speaker condition*, items with noun-focus and verb-focus prosody (8 items each) were invariably associated with affirmative and negative continuations, respectively. In the *unreliable-speaker condition*, half of the items with noun-focus prosody were followed by negative continuations and half of the items with verb-focus prosody were followed by affirmative continuations.

The test phase was identical across conditions. In this phase, participants made 10 additional judgments in the same format without any feedback. For each item, they were also asked to rate confidence in their judgment on a 7-point scale.

## Results and Discussion

As illustrated in Figure 4, the verb-focus prosody systematically biased judgments towards competitor pictures ( $\beta = 1.78$ ,  $p < .0001$ ) in both conditions. Crucially, we also found a significant negative interaction term between prosody and

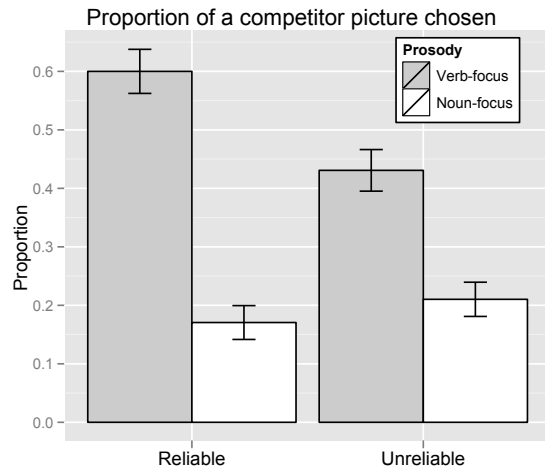


Figure 4: Proportions of competitor pictures chosen in the target and the control conditions in Experiment 3. Error bars represent standard error of the mean.

speaker reliability: In the unreliable-speaker condition, the effects of prosody on judgments, and particularly on negative interpretations of contrastive prosody, were weaker overall ( $\beta = -1.07$ ,  $p < .004$ ). That is, participants down-weighted the contrastive accent as a cue to a competitor object after exposure to a speaker who did not use the cue systematically.

Confidence rating data exhibited the same pattern. Confidence ratings were lower overall for utterances with verb-focus prosody ( $\beta = 0.3$ ,  $p < .001$ ), whereas speaker reliability did not significantly affect confidence ratings ( $p > .2$ ). We also found a significant negative interaction between these two factors ( $\beta = -.2$ ,  $p < .0001$ ). When exposed to an unreliable speaker, participants gave responses based on verb-focus prosody with diminished degree of confidence.

In sum, participants take into account prosodic features idiosyncratic to a particular speaker when deriving pragmatic meanings from prosodic contours. They down-weight prosodic information when it is an unreliable cue to intended meaning.

## Experiment 4

Experiment 4 was designed to provide a stronger test of one of the central assumptions of the adaptation hypothesis. If listeners are sensitive to the probabilistic nature of prosodic cues in the input, they should adapt their pragmatic interpretations according to the distribution of tokens in the input. Using resynthesized 12-step continua between noun-focus and verb-focus prosodic contours, we examined how different distributions of prosodic patterns in an initial exposure phase affect listeners' interpretation of utterances in the test phase.

## Methods

**Participants** We posted 360 separate HITs and received complete responses from 324 individuals.

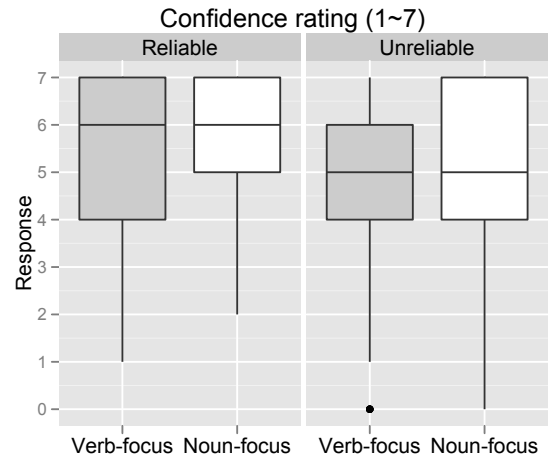


Figure 5: The responses in the confidence rating questions in Experiment 3.

**Stimuli** The stimuli created for Experiment 3 were divided into six regions corresponding to each of the four initial words (i.e. ① it ② looks ③ like ④ a) and the portions of the final word associated with each of the two tonal targets (i.e. the H\* and L-L% in the noun-focus contour and the L- and H% in the verb-focus contour). The turning point in the f0 contour within the final word was used to delineate the final two regions (e.g., ⑤ zeb ⑥ ra, as illustrated in Figure 1). The f0 of each region was sampled at 20 equally spaced time points, and measures from each time point were aggregated across items to derive mean f0 contours for noun-focus and verb-focus utterances, following (Isaacs & Watson, 2010). Likewise, the durations of each region were averaged across items by contour type. Twelve-step continua for each item were derived from these mean f0 contours and durations by interpolating between values within each region and then manipulating the F0 and duration of each recording to match the interpolated values using the pitch-synchronous overlap-and-add algorithm implemented in Praat (Moulines & Charpentier, 1990; Boersma & Weenink, 2008).

These items (12 steps \* 26 words) were normed by 50 people using the same 2AFC picture-selection paradigm as in Experiment 3 without feedback. The results of this norming study are summarized in Table 1. Items from most of the continuum steps were more likely to elicit affirmative responses. The items at Step 10 were judged to be the most ambiguous as to the pragmatic interpretations. Based on these norming responses, we postulated that the prosodic cue values for the Noun-focus and the Verb-focus stimuli form distributions with different means and variance, as schematically shown in Figure 6a. The distribution of cue values for the affirmative interpretation (solid line) has a mean value close to the midrange of the continuum and has relatively high variance. In contrast, the distribution for the negative interpretation (dashed line) is considered to have a mean value closer to

Table 1: Proportion of a target picture chosen at each step in the norming study.

Steps	1	2	3	4	5	6	7
Percentages (%)	84	79	80	80	75	77	78
	8	9	10	11	12		
	60	60	48	37	37		

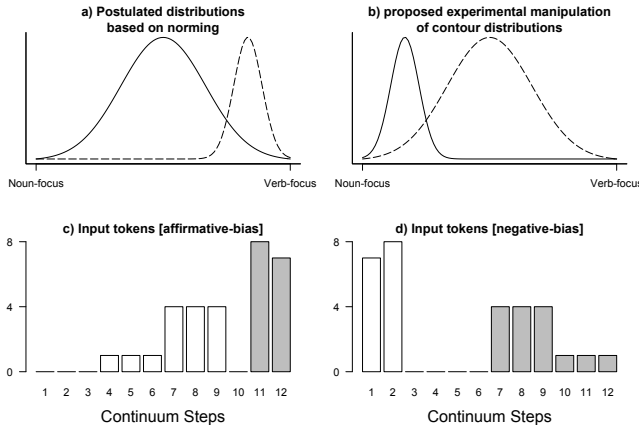


Figure 6: a) A schematic representation of distributions of prosodic cue values postulated based on the norming data. b) Proposed experimental manipulation of contour distributions. c) and d) Input frequencies of tokens sampled from each step of the continuum in the training phase of the affirmative-bias in the negative bias conditions. X-axis: continuum steps. Y-axis: Token frequencies of input utterances. Tokens indicated as white bars were disambiguated as affirmative interpretation and those indicated as shaded bars were disambiguated as negative interpretation.

the higher end of the continuum, with relatively low variance.

Based on these assumptions, we created two sets of exposure items for the current experiment. In the **affirmative-bias condition**, the distributions of the exposure items were meant to approximate the distributions observed in the norming study. Participants heard items sampled from Steps 4-9 with affirmative continuations and items from Steps 11 and 12 with negative continuations (Figure 6c). On the other hand, in the **negative-bias condition**, we tried to reverse this pattern and provided input in the distribution patterns, which are schematically illustrated in Figure 6b. In this condition, listeners heard items from Steps 1 and 2 with affirmative continuations and items from Steps 4-9 with negative continuations. Notice that in either of the conditions participants heard the same number of items from Steps 7-9 while they are identified as items from different categories (Figure 6d).

The adaptation hypothesis predicts that exposure to these affirmative-bias and negative-bias distributions should result in recalibration of the categorization function. Figure 7a plots the proportions of target pictures chosen at each step along the continuum in the norming study. We hypothesized that

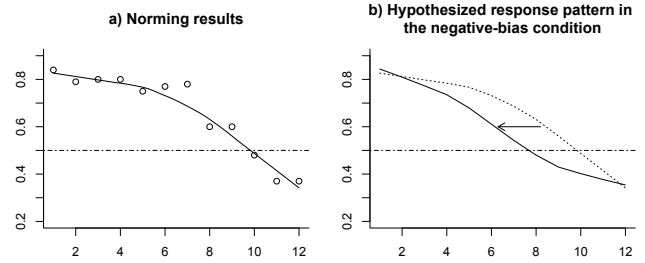


Figure 7: a) Proportions of target pictures chosen (affirmative interpretation) in the norming study. X-axis: Continuum steps (1 = prototypical noun-focus prosody, 12 = prototypical verb-focus prosody). Solid line represents lowest smoothing and dashed line indicates where the stimuli elicit most ambiguous responses (50% chance of a target picture chosen); b) A hypothesized pattern of category recalibration in the negative-bias condition in Experiment 4.

participants' categorization functions would shift towards the negative interpretation, as illustrated in Figure 7b.

**Procedure** The procedure of the experiment was nearly identical to Experiment 3. Participants were exposed to tokens of "It looks like an X" and selected either a target picture or a competitor picture as the more likely referent. In the exposure phase, they heard 30 items: 15 with affirmative continuations and 15 with negative continuations. The distribution of items sampled from a 12-step continuum differed across conditions, as illustrated in Figure 6c and 6d. In the test phase, which was identical across conditions, participants completed 12 trials in the same format without feedback.

## Results and Discussion

Responses are plotted in Figure 8. In the affirmative-bias condition, as predicted, participants' judgments did not deviate from those in the norming study. Items from most of the steps were associated with the affirmative interpretation and the items from Step 10 was judged to be most ambiguous as to their pragmatic meanings.

In the negative-bias condition, however, a wider variety of items were assigned the negative interpretation (i.e., *it is not an X*). As shown in Figure 8, the proportion of affirmative interpretations drops to 50% around Step 7. Items that had been normed to be highly ambiguous (i.e., those from Steps 10) were more reliably assigned the negative interpretation. These results lend strong support for the adaptation hypothesis: Pragmatic interpretation of contrastive prosody does not result from an invariant mapping between sound and meanings. Depending on the patterns of input, participants can rapidly and flexibly adjust their classification criteria so that they can make optimal use of the incoming input.

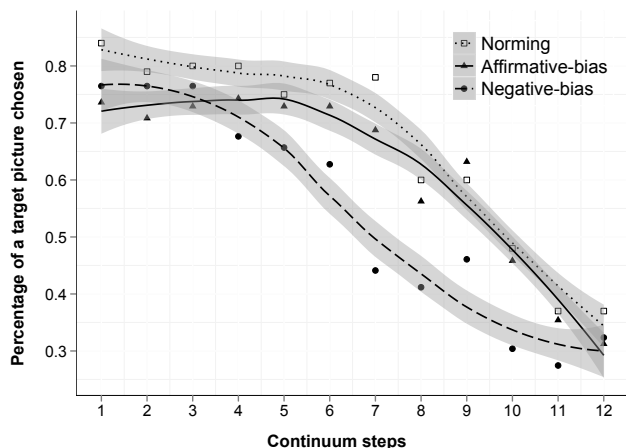


Figure 8: Percentages of target pictures chosen in the test phase [Experiment 4]. Dotted, solid, and dashed lines represent responses from the norming, the affirmative-bias, and the negative-bias conditions, respectively. X-axis plots the continuum steps. Step 1: prototypical noun-focus prosody; Step 12: prototypical verb-focus prosody.

## General Discussion

Our results provide novel evidence about how listeners navigate variability in prosodic information to make inferences about an intended meaning of an utterance. We first confirmed that listeners derive affirmative and negative interpretations for “It looks like an X” based on two distinct prosodic patterns (noun-focus and verb-focus contours). The pragmatic inference associated with the contrastive interpretation, however, is affected by the range of linguistic contrasts provided in the context: Introducing a stronger statement (“It is an X”) increases the proportion of contrastive implicatures for the “It looks like an X” construction. Listeners also rapidly adjust to speaker-specific use patterns of prosodic cues to pragmatic meanings. Finally, listeners optimize pragmatic interpretations based on probabilistic distributions of prosodic cue values. After hearing 30 tokens of input, listeners adjusted their classification criteria to reflect the properties of the distribution to which they were exposed.

These results provide strong support for the adaptation hypothesis: Listeners take into account the reliability of the mapping between prosodic patterns and intended meanings for a particular speaker when evaluating whether a prosodic contour provides evidence for a contrastive inference. Also, in order to effectively process noisy input, listeners integrate multiple acoustic cues as well as information idiosyncratic to a particular context. In future research we plan to extend our approach to a wider range of constructions and prosodic contours in order to further evaluate the hypothesis that adaptation to prosody can be understood within a rational inference framework.

## Acknowledgments

Thanks to Eve V. Clark, Christine Gunlogson and Sarah Bibyk for valuable discussion.

## References

- Boersma, P., & Weenink, D. (2008). *Praat: Doing phonetics by computer (version 5.0.26) [computer program]*. (Retrieved June 16, 2008, from <http://www.praat.org/>)
- Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3), 804–809.
- Dennison, H. Y., & Schafer, A. J. (2010). Online construction of implicature through contrastive prosody. In *Speech prosody 2010 conference*.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.
- Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, 105(2), 251–279.
- Grice, H. P. (1989). *Studies in the way of words* (Vol. 65) (No. 251). Harvard University Press.
- Isaacs, A., & Watson, D. (2010). Accent detection is a slippery slope: Direction and rate of f0 change drives listeners comprehension. *Language Cognitive Processes*, 25(7), 1178–1200.
- Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, 58(2), 541–573.
- Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446.
- Kleinschmidt, D., & Jaeger, T. F. (2011, June). A Bayesian belief updating model of phonetic recalibration and selective adaptation. In *Acl workshop on cognitive modeling and computational linguistics*. Portland, OR.
- Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: dialects, idiolects, and speech processing. *Cognition*, 107(1), 54–81.
- Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6), 453–467.
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., et al. (1992). ToBI: A standard for labeling English prosody. In *International conf. on spoken language processing* (Vol. 2, pp. 867–870). Banff.
- Watson, D., Tanenhaus, M., & Gunlogson, C. (2008). Interpreting pitch accents in online comprehension: H\* vs. L+H\*. *Cognitive Science A Multidisciplinary Journal*, 32(7), 1232–1244.

# A Graph-Oriented Approach to Measuring Expertise- Detecting Structural Differences between Experts and Intermediates

Andreas Lachner (andreas.lachner@ezw.uni-freiburg.de)

Johannes Gurlitt (johannes.gurlitt@ezw.uni-freiburg.de)

Matthias Nückles (matthias.nueckles@ezw.uni-freiburg.de)

Department for Educational Science, University of Freiburg  
Freiburg, Germany

## Abstract

For designing effective and tailored instruction, valid instruments that measure the level of expertise are necessary. We propose a graph-oriented approach for in-depth analyses of knowledge structures. Therefore, four measures of integration and encapsulation of knowledge structures were validated in an experiment. Participants (six experts and six intermediate students) recalled and explained the symptoms and laboratory data of a medical case description in the domain of cardiology. The results showed that the graph-oriented measures were more discriminative towards expertise-related differences than classic measures. Thus, our graph-oriented measures offer a more adequate and a more fine-grained analysis of knowledge structures.

**Keywords:** expertise, graph theory, knowledge encapsulation, knowledge integration.

## Introduction

“Experts are made, not born” (Schraw, 2009). This citation nicely illustrates that the way from a novice to an expert can be characterized as a bumpy road of deliberate practice and effort (Ericsson, 2006). For supporting novices in developing their skills and knowledge, good and accurate shock absorbers, such as effective instructional explanations, are necessary. Thus, a deep understanding of expertise and its unique differences to novices’ knowledge structures as a target state of novices’ development is crucial for designing effective instruction (Nückles, Wittwer, & Renkl, 2005). Cognitive science provides a comprehensive picture about the patterns of knowledge structures that constitute expertise: the main findings suggest that experts primarily differ from novices in the nature of their knowledge structure; more specifically in the extent to which their domain knowledge is integrated and compiled. Knowledge integration can be described as principled knowledge, which is characterized as coherent and well-integrated domain knowledge (Chi, Feltovich, & Glaser, 1981). More specifically, novices tend to organize their knowledge around literal, superficial features, while experts organize their knowledge around abstract principles lying underneath the superficial features. These abstract principles allow for the integration of obviously divergent concepts and subcomponents into a coherent, tightly connected schema. As novices and intermediates do not come up with these abstract principles, they have more difficulties in recognizing patterns that fit together, as, for example, it is easier for a medical doctor to ascribe divergent symptoms

like a loss of vision and a collapsing pulse to bacterial endocarditis, whereas novices and intermediates would not be able to intuitively ascribe these symptoms to a specific disease and would rather tend to “store” such details unconnectedly in long-term memory (Schmidt & Rikers, 2007).

The second distinctive feature of experts’ knowledge is the degree of compilation. Knowledge compilation refers to the process by which persons transform declarative knowledge into productions and automate these productions by combination of these productions to larger units (Anderson, 1981). For instance, compared to intermediates, in order to medicate a flu, a medical doctor does not need to reason on the detail-level of pathophysiology, but rather operates on the macro-level of automated clinical knowledge, like “if the patient has symptoms A,B,C, then she has...“, which allows the expert to omit reasoning steps when solving routine tasks (Koedinger & Anderson, 1990). But how does this compilation change the expert’s knowledge structure? Boshuizen and Schmidt (1992) suggested that knowledge compilation resulted in the “subsumption of lower level, detailed propositions under higher level [...] propositions.” This reorganization of the knowledge structure is called knowledge encapsulation. For example, Rikers, Schmidt, and Boshuizen (2002) found that experts’ knowledge structures were less detailed and they contained more encapsulated concepts compared to intermediates. In sum, developing expertise can be illustrated by progressing through a number of transitory stages that are characterized by the degree of integration and the degree of encapsulation of the knowledge structure (Schmidt & Rikers, 2007). To support learning, accurate and valid assessment strategies of these stages of knowledge structures are needed as a prerequisite for the design of effective and tailored instruction.

## Measurement of Knowledge Integration and Knowledge Encapsulation

To have experts and intermediates elicit their knowledge, we used the classic procedure by Patel and Groen (1986) that consists of the following elements: 1) Participants are provided with a medical case description, 2) they accomplish a free recall task of the medical case description, 3) explain the underlying processes that cause the disease, and 4) provide a diagnosis for the case description. Whereas the free recall protocol allows for an insight into a

participant's problem representation, the explanation captures the conceptual understanding of underlying patterns and the logical and semantic relations of the subject domain (Chi, 2006). For the analysis, the recall protocols and explanations were segmented into propositions, consisting of one relation and an ordered set of two concepts, containing the elementary idea units of the referring text base (Kintsch, 1988).

### Classic indicators of knowledge encapsulation and knowledge integration

Based on the propositional segmentation of the recall protocols and the explanations, Rikers et al. (2002) used three different measures of knowledge encapsulation and knowledge integration. Knowledge encapsulation was measured by the number of *high-level inferences* a participant made during the recall of the case description. A high-level inference is a statement that compiles several reasoning steps into one statement. Therefore, each statement in the recall protocol was coded as a 1) literal, 2) paraphrased, 3) low-level inferred, or 4) high-level inferred proposition of the case description. Low-level inferences were based just on one statement in the case description, whereas high-level inferences merged several propositions of the case description into one inferential proposition. Consider the following propositional segmentation of a case description and a fictitious participant's recall (cf. figure 1). In this case, the participant merged seven propositions of the case description to one proposition and solely recalled that the man has endocarditis. Therefore, the proposition was coded as high-level inference.

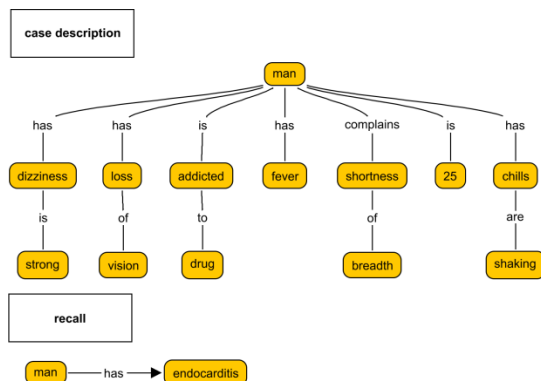


Figure 1: Example for the coding of high-level inferences

The second indicator, *encapsulated concepts*, was measured by the number of matching concepts between the participant's explanation and a reference model that included encapsulated concepts of the subject domain. For instance, the match between the graphs in figure two and three would have four matching encapsulated concepts. Thus, the participant would have used four encapsulated concepts.

For the knowledge integration, Rikers et al. used the number of mentioned *concepts* and *relations* in the explanations. For instance, the graph depicted in Figure 2

would have a detailedness index of four concepts and three relations.

### Limitations of the classic indicators

Although the measure *high-level inferences* provides a fine-grained analysis on the level of inferences in the recall protocols, the analyses of the participants' explanations imply some shortcomings: The measurement of the use of encapsulated concepts and detailedness were solely based on the computation of frequencies of concepts and relations. However, structural dependencies, like inter-relations between concepts, were not investigated and thus lack validity. In order to properly measure knowledge encapsulation, what must be demonstrated is the reorganization of the knowledge structure, more precisely how participants subsume their knowledge of details under higher-level concepts (Boshuizen & Schmidt, 1992). In a similar vein, measuring knowledge integration requires both structural indicators for the connectedness and the fragmentation of the knowledge structures. Therefore, an in-depth analysis of structural properties is necessary for validly measuring knowledge integration and knowledge encapsulation.

### A Graph-Oriented Approach for Measuring Knowledge Integration and Knowledge Encapsulation

The purpose of this paper was to improve existing measures in order to increase the reliability and validity of knowledge encapsulation and knowledge integration measures. Therefore, to capture key latent variables of knowledge encapsulation and knowledge integration, we developed four measures that were strictly based on graph theory (Sowa & Shapiro, 2006). The analysis of knowledge structures with graph-oriented measures has two main methodological advantages. First, they are capable of directly tracking structural differences, which heightens the validity of our methodology. More precisely, with graph-oriented measures, subsumption and integration processes can be captured in the graphical structure. Second, due to the mathematical formalization of knowledge encapsulation and knowledge integration, our graph-oriented measures could easily be automated, which increases objectivity and efficiency.

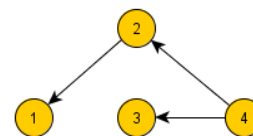


Figure 2: Example for a conceptual graph

Mathematically, an explanation segmented into single propositions can be interpreted as a directed simple graph (Sowa, 2006). A graph  $G$  is an abstract representation of a finite set of nodes  $V$  that are connected by edges  $E$ , mathematically as  $G = (V, E)$ . Nodes represent concepts, whereas edges represent relations between the concepts.



## Knowledge Integration

For the analysis of knowledge integration, we used two different measures. *Connectedness* was computed by the proportion of the sum of edges  $e$  and the sum of nodes  $v$ , formally as:

$$\frac{\sum_{i=1}^n e_i}{\sum_{i=1}^n v_i}$$

This expression describes the average relatedness of concept to concept and can take values between 0 and 1, where 0 represents a non-connected graph and 1 means that all concepts are directly related to each other. Figure 2 shows an example of a graph consisting of three relations and four concepts. The connectedness for the example graph would be .75.

The second indicator for integration is *fragmentation* of the knowledge structure. Fragmentation was computed as the number of isolated knowledge units. A knowledge unit is represented as a disconnected component in a graph, indicating a subgraph that is not connected to the rest of the graph. Formally, we define fragmentation as the number of components  $C_n$  which are subsets of the graph  $G$ , where each node  $v \in V$  has no edge connection to the set of nodes  $v$  of the complement of  $G \setminus C_n$  (Sowa & Shapiro, 2006). Our example graph would have a fragmentation index of 1, because there are no disconnected subcomponents in the graph.

## Knowledge Encapsulation

For the analysis of the encapsulation of the knowledge structures, we used two different measures. The *omission of concepts* is an indicator of how many inferential steps a participant skips while explaining a phenomenon. The more encapsulated a knowledge structure is, the more concepts a participant omits. For the identification of the inferential steps, a reference model is needed. This reference model must include all causal relations to sufficiently understand the phenomenon under investigation. Thus, the reference model depicts an accurate causal representation of the phenomenon. The omission of concepts is computed as the number of concepts that are in the set of the reference model ( $rm$ ), but not in the participant's model ( $pm$ ), formally as:

$$rm \setminus pm = \{x \in rm \mid x \notin pm\}$$

The more omission a participant made the less accurate was her explanation. Figure 3 shows an example for a reference model. Located in the reference model are the concepts 5, 6 and 7 that do not appear in the participants' model in figure 2. In this case, we would have an omission of 3, because in this case the participant would have omitted three concepts in her explanation.

A second indicator concerning knowledge encapsulation is the length of the *inference path*. It describes the shortest path between the most distinct concepts and is an indicator for the average length of inferences in the experts' explanation (Dijkstra, 1959). It is computed as the shortest

distance between the most distant nodes and can take values from 1 to  $N$ . A low index in the inference path indicates high encapsulation, whereas high  $N$  indicates a very detailed description of the phenomenon. In our example, the most distant nodes would be *Node 6* and *Node 7*, and the shortest path would include 4 edges; therefore the inference path would be 4.

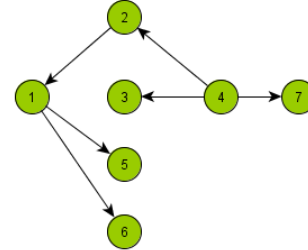


Figure 3: Example for a reference model

## Research Questions and Hypotheses

The main purpose of the experiment was to test, if our graph-oriented measures were more sensitive when investigating expertise-related differences of knowledge encapsulation and knowledge integration compared to the classic measures (Boshuizen & Schmidt, 1992; Patel & Groen, 1986; Rikers, Schmidt, & Boshuizen, 2002). In more detail, we examined, if our graph-oriented measures were more capable to detect differences concerning knowledge encapsulation and knowledge integration compared to the classic measures. Furthermore, we validated our graph-oriented measures with the classic measures. For this purpose, we, analogically to previous experiments, asked cardiology experts and intermediate medical students to recall and explain the signs and symptoms of a clinical case description of a fictitious patient who had bacterial endocarditis, taken from Patel and Groen (1986).

Based on these theoretical considerations, we addressed the following research questions.

### Predictions Regarding Structural Differences between Experts and Intermediates

For the classic indicators, in accordance with Rikers et al. (2002), we hypothesized that experts would make more *high-level inferences* and would use more *encapsulated concepts* compared to intermediates. For knowledge integration, analogically to Rikers et al., we assumed that experts' explanations would be less *detailed* (less concepts and less relations) compared to intermediates.

For the graph-oriented measures, we expected the following effects: Generally, we assumed that experts would subsume specific concepts under encapsulated concepts, which would result in *shorter inference paths* and more *omissions of concepts* in their explanations compared to intermediates. With regard to knowledge integration, we expected that experts' knowledge structures should be more *tightly connected* and less *fragmented* compared to intermediates (Chi et al., 1981).

## Predictions Regarding the Validity of the Graph-Oriented Measures

In order to test concurrent validity of the graph-oriented measures, we examined whether the classic measures were correlated with the graph-oriented measures. More importantly, we tested whether our graph-oriented measures would be able to discriminate between experts' and intermediates' explanations. More specifically, we hypothesized that our graph-oriented measures would be more discriminatory with regard to differences related to expertise, because they would be better able to measure structural differences.

## Method

### Participants

Six experts and six intermediates participated in the experiment. Experts were recruited from a German cardiology hospital. All were medical specialists who had a mean work experience of 19.5 years and board certification in their specialty of cardiology. They were, on average, 49.75 years old ( $SD = 6.24$ ). Intermediates were advanced medical students in the clinical block of their study program. They were on average 25.83 years old ( $SD = 1.72$ ). Their average number of semesters in the medical program was 10.83 semesters ( $SD = 1.17$ ) and they had attended at least one special course in cardiology.

### Design

A quasi-experimental between subjects design was used, with expertise as the independent variable. Dependent variables encompassed measures of knowledge integration and knowledge encapsulation.

### Materials

The materials were merged into one booklet, containing a demographic questionnaire about the participants' age, prior knowledge and experience in the area of cardiology. The main component was a clinical case description of a fictitious patient who had bacterial endocarditis (an inflammation of the inner layer of the heart). This description was used in several previous studies (Patel & Groen, 1986; Rikers et al., 2002). The clinical case description included context information, central findings of laboratory data, and descriptions of symptoms. Furthermore, we included two blank sheets for the recall task and the explanation.

### Procedure

The entire experiment lasted approximately 40 minutes. First, the participants completed the demographic questionnaire (5 minutes). Second, they read the case description (5 minutes). Third, in the recall task, participants wrote down everything they could remember (5min). Fourth, the participants provided an explanation for the signs and symptoms of the case description, in full

sentences (20 minutes). They were asked to write an intelligible and comprehensive explanation. Fifth, participants provided a diagnosis and suggested possible therapies (5 minutes).

## Analysis of the Knowledge Structures

We used our graph-oriented measures, described above, to examine differences between experts and intermediates regarding their knowledge structures. For the cross-validation of our graph-oriented measures, we used the classic measures by Rikers et al. (2002) as well. To heighten reliability, we implemented a computer program which automatically calculated all mathematically formalized measures for knowledge integration and knowledge encapsulation, except for the number of high level inferences. Latter was coded by two independent raters that were blind to the experimental conditions. Interrater agreement as determined by Cohen's Kappa was very good ( $\kappa = .77$ ) and differences were resolved by discussions.

## Results

There were no significant differences between experts and intermediates regarding the number of propositions in the recall protocol,  $F(1, 10) = 3.44$ ,  $p = .09$ , partial  $\eta^2 = .26$ , and the number of propositions in the explanations,  $F(1, 10) = 3.01$ ,  $p = .11$ , partial  $\eta^2 = .24$ . Furthermore, as all of our participants were knowledgeable in the domain of cardiology, all participants correctly diagnosed that the patient had bacterial endocarditis, and proposed broad antibiotic mediation as first treatment. The means and standard deviations for all the dependent measures as well as for the propositions can be seen in table 1.

## Predictions Regarding Structural Differences between Experts and Intermediates

### Classic indicators

Concerning *knowledge integration*, our analyses showed that intermediates' explanations contained more *concepts*,  $F(1, 10) = 6.23$ ,  $p = .03$ , partial  $\eta^2 = .38$ , but did not significantly differ with regard to the number of *relations*,  $F(1, 10) = 3.04$ ,  $p = .11$ , partial  $\eta^2 = .23$ . Concerning *knowledge encapsulation*, there was no significant difference between experts and intermediates regarding the number of high-level inferences,  $F(1, 10) = 2.43$ ,  $p = .15$ , partial  $\eta^2 = .20$  and in the use of encapsulated concepts,  $F(1, 10) = .06$ ,  $p = .82$ , partial  $\eta^2 = .01$ .

### Graph-Oriented Measures

With regard to *knowledge integration*, intermediates' knowledge structures were significantly more fragmented than experts' knowledge structures,  $F(1, 10) = 6.58$ ,  $p = .03$ , partial  $\eta^2 = .40$ . Concerning connectedness, there was no significant difference between experts and intermediates,  $F(1, 10) = 2.83$ ,  $p = .12$ , partial  $\eta^2 = .22$ .

With regard to *knowledge encapsulation*, experts' inference paths were significantly shorter than those of intermediates,  $F(1, 10) = 4.40$ ,  $p = .05$ , partial  $\eta^2 = .33$ . Furthermore, experts omitted more relevant concepts in

their explanations compared to intermediates,  $F(1, 10) = 7.50$ ,  $p = .02$ , partial  $\eta^2 = .43$ .

Table 1: Means, standard deviations and effect sizes for knowledge integration and encapsulation.

Variables	Intermediates	Experts	$\eta^2$
Propositions RC <sup>a</sup>	33.33 (5.99)	27.50 (4.85)	.26
Propositions EX <sup>b</sup>	44.67 (8.59)	34.50 (11.31)	.24
<i>Classic measures of knowledge integration</i>			
Concepts	50.83 (9.45)	35.83 (11.29)	<b>.38</b>
Relations	44.50 (8.46)	34.33 (11.52)	.23
<i>Graph-oriented measures of knowledge integration</i>			
Connectedness	.88 (.05)	.96 (.11)	.22
Fragmentation	6.5 (1.98)	3.5 (2.07)	<b>.40</b>
<i>Classic measures of knowledge encapsulation</i>			
High-level	1.17 (1.17)	5.33 (6.44)	.20
Inferences			
Encapsulated concepts	7.50 (1.87)	7.17 (2.93)	.01
<i>Graph-oriented measures of knowledge encapsulation</i>			
Inference path	10.83 (3.06)	7.17 (2.64)	<b>.33</b>
Omission of concepts	2 (1.10)	4 (1.41)	<b>.43</b>

Note. Differences with  $p < .05$  are in boldface.

<sup>a</sup> mean number of propositions in the recall protocols.

<sup>b</sup> mean number of propositions in the explanations.

## Predictions regarding the Validity of the Graph-Oriented Measures

Table 2 presents the correlations between the classic and the graph-oriented measures. With regard to *knowledge integration*, we found high correlations between the classic measure of detailedness and the graph-oriented measure of fragmentation.

For *knowledge encapsulation*, we found high correlations between the classic measure of high-level inferences and the graph-oriented measure of inference path. Correlations between high-level inferences and omission of concepts were not significant. As the effect sizes (table 1) indicated, the best indicator of knowledge integration was our fragmentation measure; for knowledge encapsulation, our omission of concepts measure.

To test if our graph-oriented measures discriminated better between experts and intermediates as compared to the classic measures, we conducted a discriminant analysis (step-wise). All variables both of the classic and the graph-oriented measures were entered into the analysis. The method of minimizing Wilks' lambda was used for inclusion of the variable, and the criterion F to enter was set to 4. The stepwise discriminant heuristic selected as relevant predictors *omission of concepts* and *fragmentation*, canonical  $R^2 = .62$ , which significantly discriminated all the cases into the expert's and intermediate's condition,  $\Lambda = .39$ ,  $\chi^2(2) = 8.58$ ,  $p = .01$ . The discriminant heuristic solely selected graph-oriented measures, but none of the classic measures was selected.

Table 2: Correlations of the dependent measures ( $N=12$ )

	1	2	3	4	5	6	7
<i>Classic measures</i>							
1 Concepts	-						
2 Relations	<b>.95</b>	-					
3 High-level Inferences	-.43	-.46	-				
4 Encapsulated Concepts	.24	.05	.25	-			
<i>Graph-oriented measures</i>							
5 Connectedness	-.35	-.04	-.20	<b>-.65</b>	-		
6 Fragmentation	<b>.64</b>	.41	-.06	.55	<b>-.78</b>	-	
7 Inference Path	<b>.68</b>	<b>.63</b>	<b>-.62</b>	.30	-.18	.45	-
8 Omission of concepts	-.54	-.49	.14	-.19	.24	-.34	-.26

Note. Correlations with  $p < .05$  are in boldface.

## Discussion

In this paper, we proposed four graph-oriented indicators for measuring knowledge integration and knowledge encapsulation. Based on graph theory, these indicators allow for an in-depth analysis of the structure of knowledge integration and encapsulation. The results from our study can be summarized as follows:

Overall, our results showed the validity of our graph-oriented measures with regard to knowledge encapsulation and knowledge integration by detecting structural differences between experts' and intermediates' knowledge structures.

For the classic measures by Rikers et al. (2002), the only statistically significant expertise-related difference occurred in regard to the number of concepts, indicating that experts' explanations were less detailed than intermediates' explanations. For the other classic knowledge encapsulation measures, that is, the number of high-level inferences and the number of encapsulated concepts, no significant differences between experts and intermediates were found. However, we concede that our sample of experts and intermediates was very small. Thus, given the considerable effect sizes for the classic measures, it can be assumed that with a larger sample size, those differences would have also reached statistical significance.

Our graph-oriented measures of encapsulation, namely the omission of concepts and the length of the inference path, significantly differed between experts and intermediates. Additionally, for knowledge integration, we found significant differences with regard to the fragmentation of the explanations, indicating that experts' explanations were less fragmented (i.e. more integrated) than intermediates' explanations. However, connectedness did not differ significantly between experts and intermediates. Generally, the largest effects in our study resulted for the graph-oriented indicators: analyses showed that the most discriminative indicator of knowledge integration was fragmentation. Similarly, regarding knowledge encapsulation, the omission of concepts was the most discriminative predictor. Hence, the graph-oriented

indicators were more sensitive towards differences between experts and intermediates, that is, the graph-oriented indicators were better able to discriminate between experts and intermediates. Further evidence for the validity of our graph-oriented measures can be found in the high correlations between the classic measures by Rikers et al. (2002) and our graph-oriented measures. They seem to measure the construct of knowledge encapsulation and knowledge integration related to the classic measures, but due to the granularity of the graph-oriented measures in a more sensitive and discriminative way.

Despite the promising results of our experiment, there are also limitations and open questions that need to be addressed. One limitation refers to the small sample size in the experiment. Although we showed that the graph-oriented measures were more discriminative compared to the classic measures by Rikers et al., the small sample size limited test power and therefore results should be interpreted with caution. As our experts were cardiologists with around 20 years of work experience, it proved to be difficult to convince a large number of them to participate in our study. Additionally, in using only one task, namely to explain the reasons of bacterial endocarditis, the scope of our experiment was rather restricted. Therefore, additional tasks should be included to map a more integrated representation of the domain of cardiology. Apart from the scope, it should also be acknowledged that assessing participants' knowledge structures by analyzing written recall protocols and written explanations is a rather indirect measure of participants' knowledge structure. Therefore, it should be examined whether our results can also be replicated using a more direct elicitation technique, such as think-aloud protocols. Beside these methodological issues, there remains the question, if the graph-oriented indicators are able to model the development of expertise. Therefore, novices should be included in future studies.

In conclusion, we see our methodology as a promising starting point for future research. The results showed that our graph-oriented indicators are well suited to detect differences between different expertise levels concerning the encapsulation and integration of knowledge structures. Graph-oriented indicators proved to be more sensitive and therefore more valid measures of structural differences, compared to the classic measures that solely rely on frequencies of concepts and high-level inferences. Likewise, due to the formalization of the measures, they can be easily automated, which heightens objectivity and reliability and offers a more efficient way of measuring knowledge structures. It is up to further research to explore these possibilities.

## References

Anderson, J. R. (1981). *Cognitive skills and their acquisition*. Hillsdale, NJ: Erlbaum.

Boshuizen, H. P., & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts,

intermediates and novices. *Cognitive Science*, 16, 153-184. doi: 10.1207/s15516709cog1602\_1

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2), 121-152. doi: 10.1207/s15516709cog0502\_2.

Chi, M.T.H. (2006). Methods to assess the representations of experts' and novices' knowledge. In K.A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), *Cambridge Handbook of Expertise and Expert Performance*. (p. 167-184), Cambridge University Press.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1), 269-271. doi:10.1007/BF01386390

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge: Cambridge University Press.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2), 163-182. doi:10.1037/0033-295X.95.2.163

Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science*, 14(4), 511-550. doi:16/0364-0213(90)90008-K

Nückles, M., Wittwer, J., & Renkl, A. (2005). Information about a layperson's knowledge supports experts in giving effective and efficient online advice to laypersons. *Journal of Experimental Psychology: Applied*, 11(4), 219-236. doi:10.1037/1076-898X.11.4.219

Patel, V. L., & Groen, G. J. (1986). Knowledge based solution strategies in medical reasoning. *Cognitive Science*, 10(1), 91-116. doi:10.1207/s15516709cog1001\_4

Rikers, R. M. J. P., Schmidt, H. G., & Boshuizen, H. P. A. (2002). On the constraints of encapsulated knowledge: Clinical case representations by medical experts and subexperts. *Cognition and Instruction*, 20(1), 27-45. doi:10.1207/S1532690XCI2001\_2

Schmidt, H. G., & Rikers, R. M. J. P. (2007). How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical Education*, 41(12), 1133-1139. doi:10.1111/j.1365-2923.2007.02915.x

Schraw, G. (2009). Knowledge: structures and processes. In P.A. Alexander & P.H. Winne (Eds.), *Handbook of educational psychology* (245-265). NY: Routledge.

Sowa, J. E., & Shapiro, S. C. (2006). *Knowledge representation: logical, philosophical, and computational foundations*. CA: Brooks/Cole.

# Concept learning as motor program induction: A large-scale empirical study

**Brenden M. Lake**

Department of Brain and Cognitive Sciences  
Massachusetts Institute of Technology

**Ruslan Salakhutdinov**

Department of Statistics  
University of Toronto

**Joshua B. Tenenbaum**

Department of Brain and Cognitive Sciences  
Massachusetts Institute of Technology

## Abstract

Human concept learning is particularly impressive in two respects: the internal structure of concepts can be representationally rich, and yet the very same concepts can also be learned from just a few examples. Several decades of research have dramatically advanced our understanding of these two aspects of concepts. While the richness and speed of concept learning are most often studied in isolation, the power of human concepts may be best explained through their synthesis. This paper presents a large-scale empirical study of one-shot concept learning, suggesting that rich generative knowledge in the form of a motor program can be induced from just a single example of a novel concept. Participants were asked to draw novel handwritten characters given a reference form, and we recorded the motor data used for production. Multiple drawings of the same character not only produced visually similar drawings, but they also showed a striking correspondence in their strokes, as measured by their number, shape, order, and direction. This suggests that participants can infer a rich motor-based concept from a single example. We also show that the motor programs induced by individual subjects provide a powerful basis for one-shot classification, yielding far higher accuracy than state-of-the-art pattern recognition methods based on just the visual form.

**Keywords:** concept learning; one-shot learning; structured representations; program induction

The power of human thought derives from the power of our concepts. With the concept “car,” we can classify or even imagine new instances, infer missing or occluded parts, parse an object into its main components (wheels, windows, etc.), reason about a familiar thing in an unfamiliar situation (a car underwater), and even create new compositions of concepts (a car-plane). These abilities to generalize flexibly, to go beyond the data given, suggest that human concepts must be representationally rich. Yet it is remarkable how little data is required to learn a new concept. From just one or a handful of examples, a child can learn a new word and use it appropriately (Carey & Bartlett, 1978; Markman, 1989; Bloom, 2000; Xu & Tenenbaum, 2007). Likewise, after seeing a single “Segway” or “iPad,” an adult can grasp the meaning of the word, an ability called “one-shot learning.” A central challenge is thus to explain these two remarkable capacities: what kinds of representations can support such flexible generalizations, and what kinds of learning mechanisms can acquire a new concept so quickly? The greater puzzle is putting them together: how can such flexible representations be learned from only one or a few examples?

Over the last couple of decades, the cognitive science of concepts has divided into different traditions, focused largely on either the richness of concepts or on learning from sparse data. In contrast to the simple representations popular in early cognitive models (e.g., prototypes; Rosch, Simpson, & Miller, 1976) or conventional machine learning (e.g., support vector machines), one tradition has worked to develop

more structured representations that can generalize in deeper and more flexible ways. Concepts have been characterized in terms of “intuitive theories,” which are mental explanations that underly a concept (e.g., Murphy & Medin, 1985), or “structural description” models, which are compositional representations based on parts and relations (e.g., Winston, 1975; Hummel & Biederman, 1992). In the latter framework, the concept “Segway” might be represented as two wheels *connected* by a platform, which *supports* a motor, etc. Most recently, research in AI and cognitive science has emphasized rich *generative* representations. Concepts like “house” can vary in both the number and configuration of their parts (windows, doors, balconies, etc.), much like the variable syntactic structure of language. This has lead researchers to model objects and scenes using generative grammars (Wang et al., 2006; Savova, Jakel, & Tenenbaum, 2009; Zhu, Chen, & Yuille, 2009) or programs (Stuhlmüller, Tenenbaum, & Goodman, 2010).

A different tradition has focused more on rapid learning and less on conceptual richness. People can acquire a concept from as little as one positive example, contrasting with early work in psychology and standard machine learning that has focused on learning from many positive and negative examples. Bayesian analyses have shown how one-shot learning can be explained with appropriately constrained hypothesis spaces and priors (Shepard, 1987; Tenenbaum & Griffiths, 2001), but where do these constraints come from? For simple prototype-based representations of concepts, rapid generalization can occur by just sharpening particular dimensions or features, as described in theories of attentional learning (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002) and overhypotheses in hierarchical Bayesian models (Kemp, Perfors, & Tenenbaum, 2007). From this perspective, prior experience with various object concepts may highlight the most relevant dimensions for whole classes of concepts, like the “shape bias” in learning object names (as opposed to a “color” or “material bias”). It is also possible to learn new features over the course of learning the concepts (Schyns, Goldstone, & Thibaut, 1998), and recent work has combined dimensional sharpening with sophisticated methods for feature learning (Salakhutdinov, Tenenbaum, & Torralba, 2011).

Despite these different avenues of progress, we are still far from a satisfying unified account. The models that explain how people learn to perform one-shot learning are restricted to the simplest prototype- or feature-based representations; they have not been developed for more sophisticated representations of concepts such as structural descriptions, grammars, or programs. There are also reasons to suspect that these richer representations would be difficult if not impos-

sible to learn from very sparse data. In linguistics, for instance, grammar induction is typically studied in the limit as the number of examples goes to infinity; why should we expect learning a grammar that describes instances of houses, or cars, to be possible from just one example? Theoretical arguments (e.g., the bias/variance tradeoff; Geman, Bienenstock, & Doursat, 1992) imply that representationally rich concepts should generally require more data to learn, not less. The work of Winston (1975) and Lovett, Dehghani, and Forbus (2007) might be the closest to human-level concept learning, where they learned relational schemata for simplified notions of “arches,” “houses,” “stoves,” and “fireplaces” from short sequences of examples. But a fully human-like, one-shot learning ability was beyond their scope.

Even with these gaps in our understanding, we believe that the power of human concepts will be best explained by bringing these two traditions together. By doing so, we hope to explore the extent to which people can learn representationally rich concepts from very sparse data, and we also hope to explain this ability in computational terms. These are the long-term goals of our work. Here we take a first step with a large-scale empirical study of one-shot concept learning, using a domain of handwritten characters from the world’s alphabets (see Figure 1). These objects are not nearly as complex as many object concepts such as “house,” “dog,” or “Segway,” but they still offer a vast number of novel, high-dimensional, and cognitively natural categories with important relational structure. They are much richer than the highly oversimplified artificial stimuli used in previous laboratory studies of one-shot learning (Feldman, 1997; Kemp & Jern, 2009). Yet characters are still simple enough for us to hope that tractable computational models can represent all the structure people see in them – unlike natural images.

What is the right structural representation for these simple visual concepts? The generative process for any handwritten character is a motor program, which is a set of instructions, in the mind of the drawer, that can be sent to the motor effectors such as an arm or a hand. These programs are complex compositions of pen strokes (the “parts” or the “sub-routines” of the program), which might vary in their number, order, and style across drawers. Despite these various degrees of freedom, human drawing is noted for its regularity, which has been likened to a “grammar of action” (Goodnow & Levine, 1973). Thus it seems fruitful to explore a generative approach based on motor programs, especially since people have the generative capacity for drawing. There are also well-developed, feature-based alternatives from psychology (Grainger, Rey, & Dufau, 2008) and machine learning, especially “deep learning” models which have achieved some of the best results on handwritten digit classification (0, 1, 2, ..., 9) (e.g., Salakhutdinov & Hinton, 2009). Thus it will be important to compare multiple computational approaches, with the goal of better understanding the psychological processes and also improving one-shot learning in machines.

To begin exploring these questions, we ran a large-scale

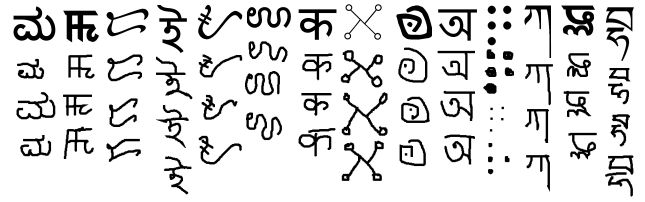


Figure 1: The top row shows example characters from our dataset, in the original printed form. Below are three example drawings from participants.

online study where participants drew novel character concepts after seeing just a single example. We refer to this task as “one-shot category production,” drawing inspiration from numerous studies that have used the generation of category exemplars as a window into conceptual representation (e.g., Battig & Montague, 1969; Rosch et al., 1976; Feldman, 1997). We see one-shot category production as a special case of “one-shot learning,” which includes classification and other types of generalization from just one example. Our large-scale study produced about 32,000 images of characters across a set of 1,600 concepts, and the on-line drawing trajectories were recorded for each image. From the production data, we analyzed the extent to which people can infer a robust motor program representation from a single example. We also compared humans and multiple computational approaches on a one-shot classification task, using methods based on either the motor data or just the visual forms.

## Category production experiment

The 1,600 character concepts were collected from 50 alphabets, including current or historic scripts (e.g., Bengali, Sanskrit, and Tagalog) and invented scripts for purposes like sci-fi novels. The characters were taken from [www.omniglot.com](http://www.omniglot.com) in printed fonts, and several originals and their subsequently drawn images are shown in Fig. 1. This dataset was previously used to compare models of one-shot classification (Lake, Salakhutdinov, Gross, & Tenenbaum, 2011).

The drawing experiment was run through Amazon Mechanical Turk, and participants were asked to draw at least one entire alphabet. For each template image, they were asked to “draw each character as accurately as you can.” An alphabet’s printed characters were displayed in rows on a webpage, with an associated drawing pad below each image. Participants could draw by holding down a mouse button and moving the mouse, and we also included “forward,” “back,” and “clear” buttons. Some participants made minor image adjustments with small mouse movements, and we tried to mitigate this inconsistency by excluding strokes that were very short in time and space from the analysis.

## The structure of the motor programs

When people perceive a new character, in what sense do they infer a new concept? While this mental representation might be just a bundle of features, the concept might also include richer structure in the space of motor programs. To investigate this possibility, we analyzed how multiple drawers produced a particular concept during the drawing task. We rea-



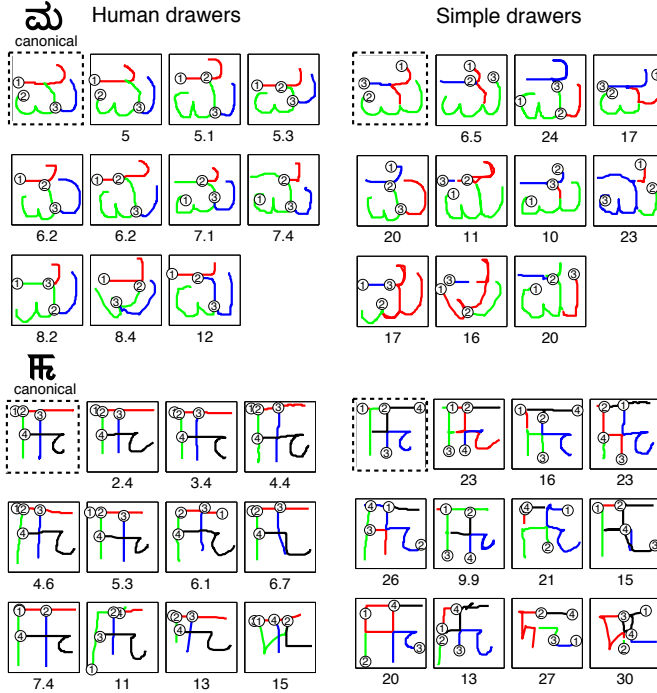


Figure 2: For two concepts (out of the 1600 total), each box shows the motor data produced by human drawers (left) or simple drawers (right). “Canonical” drawers are in the dotted boxes, and their distances (Eq. 1) to the other examples are the numbers below each frame. Stroke color shows correspondence to the canonical, circles indicate the beginning of each stroke, and numbers inside circles denote stroke order.

soned that in order to do this task, participants must infer a novel motor program, which will be reflected in the time course of drawing. Consistency in the structure of these drawings would provide evidence for two interlinked claims: people seem to grasp the same underlying concept from one example, and this concept includes a highly structured generative program. To measure consistency for a particular character, we quantitatively analyzed the number, shape, direction, and order of the parts (strokes) in the motor data.

### The number of parts

This analysis (and subsequent ones) used just 20 of the alphabets in the dataset, excluding the six most common as determined by Google hits. The remaining alphabets were needed to train the alternative models in the later classification experiment. The simplest statistic to analyze was the number of parts. For each character, we investigated whether the drawers clustered around a common number of parts (the mode number across participants). Aggregating across each drawing in the dataset, the histogram in Fig. 3A (red) shows the absolute difference between the actual number of strokes and the mode number of strokes from all of the drawings of that character. Although this distribution is guaranteed to peak at zero, a strikingly large percentage of drawers used exactly the modal number (66%). As a control, a null dataset was created by replacing each number of strokes by a uniform draw (1 to 6 here, but other values are similar). This distribution was not nearly as peaked around the mode (Fig. 3A blue).

### The shape of the parts

The parse of a character into parts (strokes) is at the core of each drawing. When people look at a new concept, do they perceive the same parts? This is difficult to analyze, since the number and length of the strokes can differ between images. A similarity measure should also be invariant to the order and direction of the strokes. Despite these challenges, we found that it was possible to analyze consistency in the shape of the strokes, and we discuss our method in the section below.

**Shape-based distance in motor space.** Since most drawers (66%) used the modal number of strokes, we restrict this and subsequent analyses to only these modal drawings. With this simplification, the strokes in two images can be matched in correspondence (one-to-one and onto). Our approach also matches the sub-structure within two strokes, finding an alignment between the points in the two trajectories (onto but not one-to-one). Given an optimal matching at both levels, the overall shape distance is roughly the mean distance between all of the aligned trajectory points. Before computing distance, characters were also transformed to be translation and scale invariant.<sup>1</sup> Examples of the distance are illustrated in Fig. 2, where the number below each drawing is the distance to the drawing in the dotted box.

The details of the distance measure are as follows. Consider two drawings  $S_1, \dots, S_k$  and  $R_1, \dots, R_k$  with  $k$  strokes each. Each stroke is a sequence of positions  $S_i = [S_{i1}, \dots, S_{in}]$  with arbitrary length, where  $S_{ij} \in \mathbb{R}^2$ . The overall distance between the characters is defined as

$$\min_{\pi} \frac{1}{k} \sum_{i=1}^k \min [dtw(S_i, R_{\pi(i)}), dtw(S_i, F(R_{\pi(i)}))], \quad (1)$$

where  $\pi(\cdot)$  is a permutation on the stroke indices  $1, \dots, k$  (a bijective function from the set  $\{1, \dots, k\}$  to  $\{1, \dots, k\}$ ), and the flip function  $F(S_i) = [S_{in}, \dots, S_{i1}]$  reverses the stroke direction to provide direction invariance. The distance  $dtw(\cdot, \cdot)$  between two trajectories is calculated by Dynamic Time Warping (DTW; Sakoe & Chiba, 1978), which fits a non-linear warp such that each point in one trajectory is aligned with a point in the other. The DTW distance is then the mean Euclidean distance across all pairs of aligned points.

**The simple drawer model.** Upon visual inspection of the stroke matches  $\pi(\cdot)$  chosen by the outer minimization in Eq. 1, there is a striking consistency across drawers in the inferred parts for a character. We show two characters in Fig. 2, where color denotes the stroke matches (left panels). The plots for the entire dataset are available online.<sup>2</sup> While this qualitative correspondence may reflect richly structured motor processes, there could be a more simplistic explanation. The consistency could be a consequence of selection bias, since we selected drawers that used the modal number of strokes,

<sup>1</sup> This transformation subtracts the center of gravity and rescales, such that the range of the largest dimension is 105.

<sup>2</sup> <http://web.mit.edu/brenden/www/consistency.html>



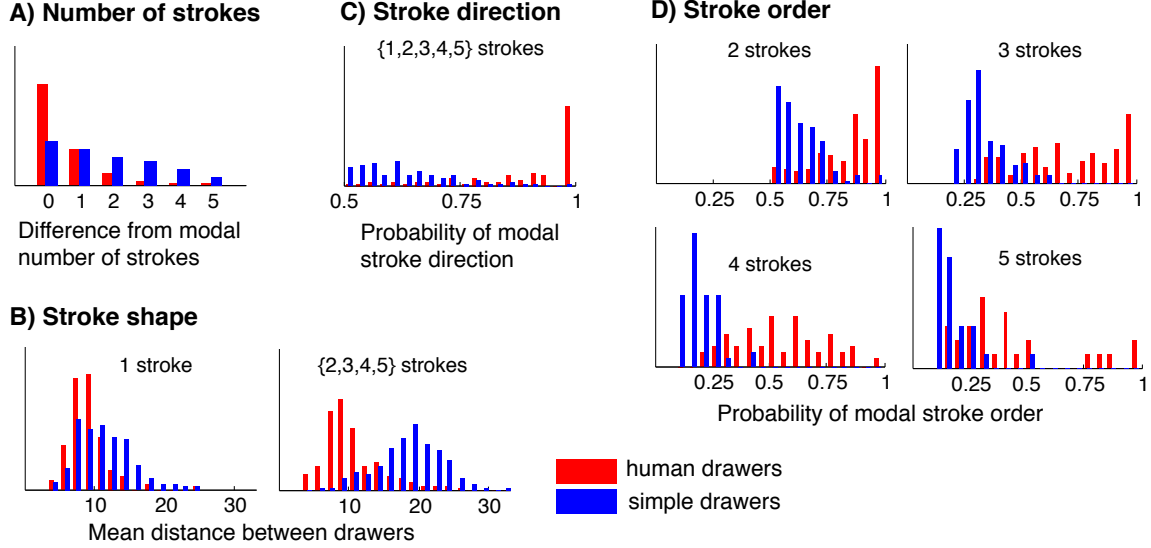


Figure 3: A histogram analysis of the consistency in the motor data, comparing human drawers (red) with a parallel dataset of simple drawers (blue) designed as a null hypothesis. Humans are strikingly consistent across a range of statistics compared to the simple model. As labeled, some histograms pool data from characters with different numbers of strokes (e.g.,  $\{2,3\}$  includes 2- and 3-stroke characters).

and there will be fewer degrees of freedom available to a  $k$ -stroke drawer for any given  $k$ . In the special case of  $k$  disjoint segments (like in Braille), there may only be one production option. To explore the degrees of freedom and to provide a baseline for the observed consistency, we devised a “simple drawer” model that is likely to mimic human drawers when the space is highly constrained, but otherwise it more freely explores the potential motor space.

The simple drawer is given access to the same set of points as a real drawer traversed in the motor data, but without the sequential information. It then tries to draw the same character as efficiently as possible using the same number of strokes. It must visit every point exactly once, while minimizing the distance traveled while ink is flowing. Given a real drawing with strokes  $S_1, \dots, S_k$ , the simple drawer’s interpretation  $Q_1, \dots, Q_k$  is defined by the problem

$$\operatorname{argmin}_{Q_1, \dots, Q_k} \sum_{i=1}^k \sum_{j=1}^{|Q_i|-1} \|Q_{ij} - Q_{i(j+1)}\|_2, \quad (2)$$

where  $|\cdot|$  is the number of points in the stroke sequene, and  $\|\cdot\|_2$  is Euclidean distance. Each point  $S_{ij}$  in the original drawing is equal to exactly one point  $Q_{ab}$  in the new drawing. This formulation encourages smooth strokes, but it also leads to creative parses (Fig. 2 right panels), in part because there are multiple optima. A drawback of the model is that it sometimes draws paths where no ink exists. To reduce this problem, the simpler drawer is not allowed to travel large distances between adjacent points, where the upper bound is the maximal adjacent distance in the corresponding real drawing. For optimization, we can reformulate the problem as the well-known traveling salesman problem (TSP) by inserting  $k$  cost-free “points” to indicate the stroke breaks. Inspired by efficient approximate solvers for the TSP

problem, we optimized using simulated annealing with alternating Metropolis-Hasting node swaps and Gibbs sampling (Rubinstein & Kroese, 2008).

**Results.** The simple drawer was used to re-sketch each image, creating an entire parallel dataset for comparative analysis. The shape-based consistency of a character is the mean distance (Eq. 1) between each pair of drawings of that character. Fig. 3B shows histograms of this consistency measure for the human drawers (red) and the simple drawers (blue). The aggregate histogram (right) for characters with two to five strokes shows a large difference in the consistency of the parts. The histogram for characters with one stroke (left) shows a closer correspondence between participants and the simple drawer, due to the limited degrees of freedom.<sup>3</sup> These results suggest that people inferred motor programs that were based on a characteristic set of strokes.

### The direction of the parts

Do different drawers infer the same stroke directions? For each character, a single canonical drawer was chosen to minimize the sum shape-based distance across all other drawers of that character (Eq. 1). Example canonical drawers are shown in the dotted boxes in Fig. 2 (left). For each person’s drawing compared to the canonical drawing, the chosen value of the inner minimization in Eq. 1 indicates whether each stroke, or that stroke in reverse direction ( $F(\cdot)$ ), is a better match to the corresponding stroke in the canonical drawer. Aggregating across each stroke in the dataset, Fig. 3C (red) displays the proportion of times the modal stroke direction was picked, using the canonical drawer as the reference point. The dataset

<sup>3</sup>Some single stroke characters can still be drawn in a number of ways, such as choosing the starting location of an “O.” People tend to start at the top, while the simpler drawer is agnostic.

of simple drawers (blue) provides a direction-agnostic baseline. By comparison, people’s inferred programs clearly have preferred directions.

### The order of the parts

Is stroke order also consistent across drawers? As in the analysis of direction, and the canonical drawers were used as the reference points, from which stroke order was defined. For any person’s drawing compared to the canonical drawing of that character, the chosen permutation  $\pi(\cdot)$  from the outer minimization in Eq. 1 defines a relative ordering of the strokes. Aggregating across each drawing, Fig. 3D shows the proportion of times the modal stroke order was picked. Like the other statistics, stroke order was also highly consistent across characters. Unsurprisingly, consistency was less pronounced as the number strokes increased.

### One-shot classification

The previous analyses suggest that people can infer rich motor-based concepts from just a single example. If the same perceptual inferences occurred in the context of categorization, would these representations prove useful for one-shot classification? We investigated this question by using the motor data to classify characters by type, based on the shape-based distance measure (Eq. 1). The model received 20 random characters with just one example each. Test examples (2 per class) were classified as the best fitting category. All 20 categories used the same (modal) number of strokes. This classification task was repeated 20 times with different characters, and the mean percent correct is shown in Fig. 4.

We used several baselines for comparison. The simplest method picked the closest image in pixel space, using Euclidean distance. We also tested Deep Boltzmann Machines (DBMs; Salakhutdinov & Hinton, 2009) as a representative feature-based model. DBMs learn a hierarchy of distributed feature representations for the raw pixels, without using a priori knowledge about the geometry of images. DBMs have obtained state-of-the-art performance on handwritten digit recognition when trained with thousands of digits, and we pre-trained it on the 30 alphabets that were not used for classification (about 19,000 images). For one-shot classification, new items were represented in feature space and classified based on cosine similarity across all hidden layers (two with 1000 units each). We also tested a model that infers latent stroke-like parts from the raw images (Lake et al., 2011), as well as the simple drawing model, which uses the same motor data but without the strong structural consistency.

Performance was measured across a range of different numbers of strokes (Fig. 4). Chance performance is 5% correct, and pixel distance performed at 20% correct on average (“pixels” in Fig. 4). Next was the DBM at 37% (“features”), the inferred parts model at 48%, and then the simple drawer at 50%. The real stroke data was far better than all of the other methods, with an average performance of 83% correct. We also tried to include stroke order and direction information in the classification cost function, but performance did

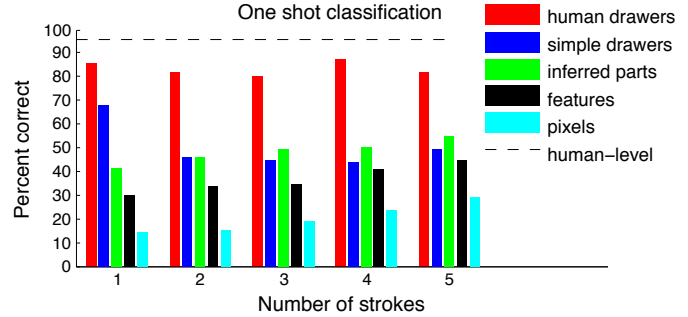


Figure 4: Classification performance based on one example of 20 different characters. Test instances were compared to each class, and the best match was selected.

not improve significantly. Finally, human one-shot classification performance was 96%, as measured behaviorally in a 20-way classification task (“human-level” in Fig. 4; see footnote for experimental setup).<sup>4</sup> Overall, the motor data was by far the most effective means for one-shot classification.

### Discussion

Our category production experiment produced over 32,000 images of handwritten characters. Each of the roughly 1,600 characters was drawn by 20 different participants, and we found a strong correspondence in the structure of their inferred motor programs. On the whole, the number, shape, order, and direction of the parts (strokes) was highly consistent across participants. Also the motor data provided a powerful basis for one-shot classification. These results suggest that when people look at a new character, they can infer a richly structured motor program. This motor program is capable of both synthesizing new examples and classifying new instances with high discriminative accuracy.

How can these motor programs be learned from just one example? This ability clearly depends on prior experience, but how does this translate into constraints on the formation of these programs? There are various possibilities. Prior knowledge might come in the form of shared sub-programs or shared strokes, like our preliminary model in Lake et al. (2011). From their general writing and drawing experience, people might learn sub-routines like “circles, diagonal lines, or S-shapes,” and then they could parse novel characters into this rich set of parts. But prior knowledge could come in many other forms, including more general constraints and biases (learned or otherwise) in human drawing and motor capabilities. Researchers have found a number of rules that usefully characterize drawing: start drawing at the top-left, draw horizontal strokes left-to-right, draw vertical strokes top-to-bottom, and minimize the number of strokes (Goodnow & Levine, 1973; Van Sommers, 1984). In a preliminary analysis, we have observed strong versions of these effects in our dataset of natural alphabets. Thus, it is possible that

<sup>4</sup>This study was run on Amazon Mechanical Turk with 15 participants and 50 trials. Each trial consisted of a single test image, and participants were asked to pick one of the 20 other images that looked the most similar. This was the same task that the models performed, except that characters with different numbers of strokes were intermixed and a different set of alphabets was used.

some of the richness of these newly acquired concepts (including shape, direction, and order) is a consequence of relatively simple, low-level principles. But it is also unclear how these directives should be combined when they conflict, or how they might interact with other forms of prior knowledge. Computational models are well-suited to help answer these questions, and we hope that future work will clarify how prior knowledge can support such rapid program induction.

Finally, although our work has focused on handwritten characters, we expect that similar phenomena and computational accounts are relevant more broadly. Characters share interesting structure with other kinds of symbols used for communication, including spoken words and gestures. Characters are produced by a sequence of strokes, and likewise, spoken words are produced as a sequence of phonemes. Characters, spoken words, and gestures are also “embodied,” since the mind and body can both generate and perceive concepts in these domains (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Freyd, 1983). All of these concepts must also be learnable from one or a few examples, in the context of efficient communication and social learning. One-shot program induction may also be possible in learning very different kinds of natural concepts, such as trees or ferns that have distinctive branching patterns and unique leaf shapes. One-shot learning could be possible here for a different reason: not because of the strong priors imposed by motor constraints or previous learning, but because a single example is highly complex and contains extensive repeated structure. We hope that future work will explore a fuller range of rich representations for concepts, while explaining how these same concepts can be learned from just one or a few examples.

**Acknowledgements** We gratefully acknowledge Jason Gross for developing the experimental interface and for collecting the data. We also thank John McCoy for helpful comments, and Dan Ellis for making his DTW code available.

## References

- Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph*, 80(3).
- Bloom, P. (2000). *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.
- Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, 15, 17–29.
- Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology*, 41, 145–170.
- Freyd, J. (1983). Representing the dynamics of a static form. *Memory and Cognition*, 11(4), 342–346.
- Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4, 1–58.
- Goodnow, J. J., & Levine, R. A. (1973). “The Grammar of Action”: Sequence and syntax in children’s copying. *Cognitive Psychology*, 4(1), 82–98.
- Grainger, J., Rey, A., & Dufau, S. (2008). Letter perception: from pixels to pandemonium. *Trends in Cognitive Sciences*, 12(10), 381–387.
- Hummel, J. E., & Biederman, I. (1992, July). Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99(3), 480–517.
- Kemp, C., & Jern, A. (2009). Abstraction and relational learning. In *Advances in Neural Information Processing Systems* 22.
- Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning over-hypotheses with hierarchical Bayesian models. *Developmental Science*, 10(3), 307–321.
- Lake, B. M., Salakhutdinov, R., Gross, J., & Tenenbaum, J. B. (2011). One shot learning of simple visual concepts. In *Proceedings of the 33rd Annual Cognitive Science Conference*.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431–461.
- Lovett, A., Dehghani, M., & Forbus, K. (2007). Incremental Learning of Perceptual Categories for Open-Domain Sketch Recognition Kenneth Forbus Comparisons and Generalization. In *Proceedings of the International Joint Conference on Artificial Intelligence*.
- Markman, E. M. (1989). *Categorization and Naming in Children*. Cambridge, MA: MIT Press.
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92(3), 289–316.
- Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, 2(4), 491–502.
- Rubinstein, R. Y., & Kroese, D. P. (2008). *Simulation and the Monte Carlo method* (Second ed.). Hoboken, New Jersey: John Wiley & Sons.
- Sakoe, H., & Chiba, S. (1978, February). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1), 43–49.
- Salakhutdinov, R., & Hinton, G. E. (2009). Deep Boltzmann Machines. In *12th International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- Salakhutdinov, R., Tenenbaum, J. B., & Torralba, A. (2011). Hierarchical deep models for one-shot learning. In *Neural Information Processing Systems (NIPS)*.
- Savova, V., Jakel, F., & Tenenbaum, J. B. (2009). Grammar-based object representations in a scene parsing task. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.
- Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, 21, 1–54.
- Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, 237(4820), 1317–1323.
- Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object Name Learning Provides On-the-Job Training for Attention. *Psychological Science*, 13, 13–19.
- Stuhlmüller, A., Tenenbaum, J. B., & Goodman, N. D. (2010). Learning Structured Generative Concepts. In *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society*.
- Tenenbaum, J. B., & Griffiths, T. L. (2001, August). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4), 629–40.
- Van Sommers, P. (1984). *Drawing and Cognition*. Cambridge University Press.
- Wang, W., Pollak, I., Wong, T.-S., Bouman, C., Harper, M., & Siskind, J. (2006, October). Hierarchical Stochastic Image Grammars for Classification and Segmentation. *IEEE Transactions on Image Processing*, 15(10), 3033–3052.
- Winston, P. H. (1975). Learning structural descriptions from examples. In P. H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill.
- Xu, F., & Tenenbaum, J. B. (2007). Word Learning as Bayesian Inference. *Psychological Review*, 114(2), 245–272.
- Zhu, L., Chen, Y., & Yuille, A. (2009, January). Unsupervised learning of Probabilistic Grammar-Markov Models for object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1), 114–28.

# How many kinds of reasoning?

## Inference, probability, and natural language semantics

Daniel Lassiter, Noah D. Goodman

Department of Psychology, Stanford University  
{danlassiter, ngoodman} @ stanford.edu

### Abstract

Previous research (Heit & Rotello, 2010; Rips, 2001; Rotello & Heit, 2009) has suggested that differences between inductive and deductive reasoning cannot be explained by probabilistic theories, and instead support two-process accounts of reasoning. We provide a probabilistic model that predicts the observed non-linearities and makes quantitative predictions about responses as a function of argument strength. Predictions were tested using a novel experimental paradigm that elicits the previously-reported response patterns with a minimal manipulation, changing only one word between conditions. We also found a good fit with quantitative model predictions, indicating that a probabilistic theory of reasoning can account in a clear and parsimonious way for qualitative and quantitative data previously argued to falsify them. We also relate our model to recent work in linguistics, arguing that careful attention to the semantics of language used to pose reasoning problems will sharpen the questions asked in the psychology of reasoning.

**Keywords:** Reasoning, induction, deduction, probabilistic model, formal semantics.

Suppose that you have learned a new biological fact about mammals: whales and dogs both use enzyme B-32 to digest their food. Is it now *necessary* that horses do the same? Is it *plausible*, *possible*, or *more likely than not*? Expressions of this type—known as *epistemic modals* in linguistics—have played a crucial role in recent work that argues for a sharp qualitative distinction between inductive and deductive modes of reasoning. In the paradigm introduced by Rips (2001) and extended by Heit and Rotello (2010); Rotello and Heit (2009), participants are divided into two conditions and are either asked to judge whether a conclusion is “necessary” assuming that some premises are true, or whether it is “plausible”. The former is identified with the deductive mode of reasoning, and the latter with the inductive mode.

These authors asked participants in both conditions to evaluate a variety of logically valid and logically invalid arguments. An example invalid argument might be “Cows have sesamoid bones; Mice have sesamoid bones; therefore, Horses have sesamoid bones”. An example valid argument might be “Mammals have sesamoid bones; therefore, horses have sesamoid bones.” They found that there was a non-linear relationship between the endorsement rates of arguments depending on condition: participants in both conditions generally endorsed logically valid arguments, but participants in the deductive condition were much less likely to endorse invalid arguments than those in the inductive condition. These results are interpreted as a challenge to theories of reasoning which rely on a single dimension of argument strength and interpret deductive validity as simply the upper extreme of this dimension (Harman, 1999; Johnson-Laird, 1994; Oscherson, Smith, Wilkie, Lopez, & Shafir, 1990). In particu-

lar, Rips and Heit & Rotello argue that non-linearities cannot be accounted for by probabilistic theories of reasoning, which identify the strength of an argument with the conditional probability of the conclusion given the premises (Heit, 1998; Kemp & Tenenbaum, 2009; Oaksford & Chater, 2007; Tenenbaum, Griffiths, & Kemp, 2006). On the other hand, they claim that the results are consistent with two-process theories of reasoning (Evans & Over, 1996).

We argue that the manipulation involving “necessary” and “plausible” hinges not on a qualitative distinction between two reasoning processes, but rather on facts about the semantics of these words which can be modeled using a single underlying scale of argument strength—conditional probability. We propose a semantically motivated model of reasoning with epistemic concepts which predicts non-linear response patterns depending on the choice of modal similar to those observed in previous work, and makes detailed quantitative predictions about response patterns in invalid arguments.

We test the claim that the modal word is the crucial factor using a new paradigm that isolates its effects. Our arguments had the same form as the examples above, except that we placed the modal word of interest in the conclusion:

**Premise 1:** Cows have sesamoid bones.

**Premise 2:** Mice have sesamoid bones.

**Conclusion:** It is {plausible/necessary/possible/likely/probable/certain} that horses have sesamoid bones.

We will refer to configurations such as “It is plausible/possible/etc. that *C*” as a **modal frame**. If varying the modal frame gives rise to a non-linear pattern of responses similar to the one found in previous work, this would indicate that an explanation of these results should be framed in terms of the meaning of these modal words.

Together, the model and experimental evidence indicate that the negative conclusions of previous work regarding one-dimensional theories of argument strength are not warranted: it is possible to explain non-linear response patterns with a probabilistic account of argument strength.

### Previous Work

Rips (2001) conducted a reasoning experiment designed to investigate the traditional distinction between deductive and inductive reasoning. Participants in two groups were asked to judge arguments either according to whether the conclusion was *necessary* (assuming that the premises were true) or whether it was *plausible*. Most participants in both conditions accepted logically valid arguments and rejected invalid arguments whose conclusion was not causally consistent with the

premises, such as “Car X strikes a wall, so Car X speeds up”. However, participants differed by condition in whether they rejected non-valid arguments which were causally consistent with the premises: those in the inductive condition generally accepted arguments such as “Car X strikes a wall, so Car X slows down”, while those in the deductive condition did not. Rips argued that this result falsifies theories of reasoning in which argument strength is a one-dimensional quantity such as conditional probability: “[i]f participants base all forms of argument evaluation on the position of the argument on a single psychological dimension, then induction and deduction judgments should increase or decrease together” (p.133).<sup>1</sup>

Heit and Rotello (2010); Rotello and Heit (2009) extended Rips’ paradigm in a number of ways. Their core finding was that  $d'$ , a standard measure of sensitivity in Signal Detection Theory (SDT), was significantly higher in the deductive condition across a variety of arguments types and manipulations.  $d'$  is defined as  $z(H) - z(F)$ , the difference between the  $z$ -scored hit rate  $H$  and false alarm rate  $F$  (Macmillan & Creelman, 2005). This difference means that participants in the inductive condition were more sensitive to argument validity than participants in the deductive condition (see Table 1).

Table 1: Acceptance rates and  $d'$  in Experiment (1a) of Rotello and Heit (2009) (three-premise arguments only).

	Deduction	Induction
Acceptance, valid	.94	.95
Acceptance, invalid	.06	.17
Sensitivity ( $d'$ )	3.31	2.56

Differential sensitivity indicates that the difference between conditions is not simply a shift in response criterion in the presence of two equal-variance Gaussians representing signal and noise. Thus we cannot fit a one-dimensional SDT model to such results. In accord with Rips, Rotello and Heit (2009) argue that the non-linear relationship between validity and condition is a challenge to probabilistic theories of reasoning. They argue that the results are better captured by a two-dimensional SDT model with possibly orthogonal dimensions of inductive strength and deductive validity, in which the response criterion can vary in two dimensions.<sup>2</sup>

<sup>1</sup>Rips (2001) also found a crossover effect in which participants in the “necessary” condition were slightly more likely to endorse valid arguments that were inconsistent with causal knowledge than participants in the “plausible” condition, but the reverse was true for arguments consistent with causal knowledge. Heit & Rotello’s work did not find any analogous effect, nor did we in our experiment reported below, and we do not consider it further. However, it is possible that the effect is real and attributable to one of the various detailed differences in materials and instructions between the experiments.

<sup>2</sup>Heit and Rotello (2010); Rotello and Heit (2009) also introduced a number of further manipulations involving e.g. constrained response time, number of premises, and readability which we will not discuss in detail for reasons of space. See the concluding section for a brief consideration of two of these manipulations, however.

## A Probabilistic Model

In contrast to Rips and Rotello & Heit, we do not see these results as strong support for a two-process theory. Instead, we will argue that these results are also compatible with a probabilistic account of inductive reasoning once the semantics of the modal words used in the experiment is taken into account. In this section we propose a model of the relationship between the probability on the one hand and epistemic concepts such as certainty, necessity, and plausibility on the other. The latter are treated as non-linear functions of the former determined by a single parameter per item. This model, inspired by recent work in formal semantics, predicts non-linear response patterns and variation in  $d'$  depending on the modal used, and makes a number of fine-grained predictions that we will later evaluate against the results of an experiment.

Our probabilistic model of reasoning with epistemic modals is intended to capture the following intuitions. A maximally strong conclusion  $C$  remains maximally strong whether you ask if it is *possible*, *plausible*, *likely* or *necessary*; a maximally weak conclusion remains maximally weak under the same conditions; but there is much more flexibility around the middle of the probability scale depending on which question is asked. If  $C$  has a probability of .4, it presumably will count as *possible* and perhaps as *plausible*, but it would not seem to be *likely* and surely not *necessary* or *certain*. Thus the effect of an epistemic modal on a conditional probability should be a transformation that preserves the minimum value 0 and the maximum value 1.

Perhaps the simplest way to capture the behavior just described is to suppose that each modal  $M \in \{\text{possible, plausible, likely, probable, certain, necessary}\}$  is associated with a parameter  $\alpha_M \in \mathbb{R}^+$  which, in combination with the conditional probability  $pr(C|P)$  of conclusion  $C$  given premises  $P$ , determines the probability of *It is M that C* given  $P$ . We propose that  $\alpha_M$  relates these two probabilities by a power-law:

$$pr(\text{It is } M \text{ that } C|P) = pr(C|P)^{\alpha_M} \quad (1)$$

We model arguments with no modal as also being governed by some  $\alpha_M$  as in (1) (rather than directly reflecting the conditional probability of the conclusion given the premises).

We assume that participants’ responses to a modalized question  $Q$  are governed by (1) plus a noise parameter  $\epsilon$  (interpreted as the proportion of trials in which participants choose a response at random).

$$pr(\text{“yes”}|P, Q = \text{Is it } M \text{ that } C?) = pr(C|P)^{\alpha_M} (1 - \epsilon) + \frac{\epsilon}{2} \quad (2)$$

Depending on  $\alpha_M$  we get a variety of possible curves, a few of which are sketched in figure 1b (setting  $\epsilon = .1$  for illustrative purposes). This model captures the behavior just described: in the limits  $\alpha_M$  does not influence the response probability, but there is significant variation in response probabilities in the middle range depending on  $\alpha_M$ . This feature leads to a prediction that the choice of modal will have less influence on response rates for arguments with very high strength (e.g., valid arguments) than those with intermediate strength.

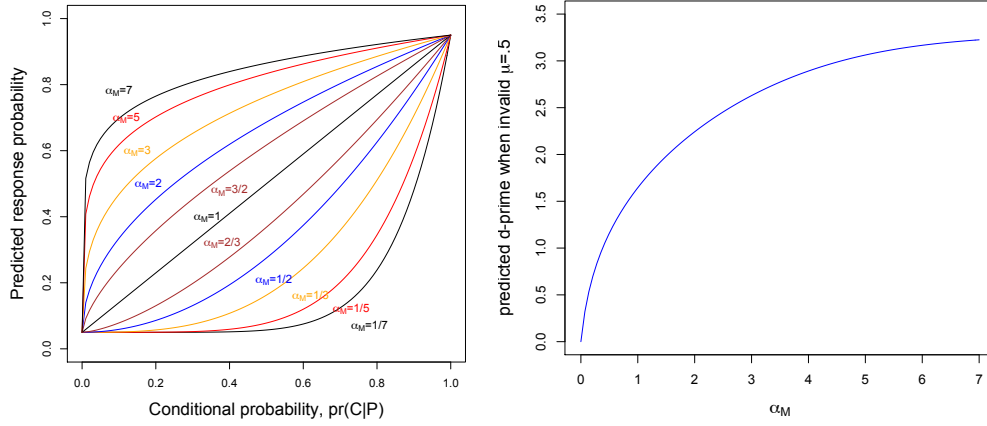


Figure 1: (a): Predicted response probability for various settings of  $\alpha_M$ . (b): Example of relation between  $\alpha_M$  and  $d'$ .

We also predict that  $d'$  will vary depending on  $\alpha_M$ . Suppose for illustration that the mean conditional probability of logically invalid arguments in some sample is .5, and that the conditional probability of valid arguments is 1. The  $d'$  statistic estimated from this data should then be (on average)

$$\begin{aligned} d' &= z(1^{\alpha_M}(1-\epsilon) + \epsilon/2) - z(.5^{\alpha_M}(1-\epsilon) + \epsilon/2) \\ &= z(1-\epsilon/2) - z(.5^{\alpha_M}(1-\epsilon) + \epsilon/2) \end{aligned} \quad (3)$$

If  $\epsilon = .1$ , we expect the observed  $d'$  to be related to  $\alpha_M$  as in figure 1b. This illustrates the fact that the value of the  $d'$  statistic is not predicted to be constant in our probabilistic model, but should depend on the choice of  $M$ . Thus, a model with one dimension of argument strength (conditional probability) is able to predict non-linearities of the type previously claimed to be problematic for probabilistic accounts.

The model also makes strong quantitative predictions about the relationship between different modal frames. That is, we predict a systematic (though non-linear) relationship between the response rates to the same argument in the modal frames *It is  $M_1$  that  $C$*  and *It is  $M_2$  that  $C$* . If  $M_1$  is associated with parameter  $\alpha_1$  and  $M_2$  with  $\alpha_2$ , for any argument with premises  $P$  and conclusion  $C$  there is some positive  $r$  such that

$$\begin{aligned} pr(Is\ it\ M_1\ that\ C|P) &= pr(C|P)^{\alpha_{M_1}} \\ &= pr(C|P)^{(r \times \alpha_{M_2})} \\ &= pr(Is\ it\ M_2\ that\ C|P)^r \end{aligned} \quad (4)$$

The prediction that every pair of modals should be related by a power-law allows us to evaluate model fit using a variety of arguments which, although not logically valid, vary widely in intuitive strength. It also shows that our model predicts that the strength of any two arguments should be related monotonically (though non-linearly) across modal frames.

## Experiment

Our experiment tested the hypothesis that non-linear response patterns can be attributed to the semantics of the modal expressions used. The main innovation was to manipulate the choice of modal  $M$  within the stimulus sentence:

*Non-valid:*

Cows have enzyme X.  
Seals have enzyme X.  
So, it is  $M$  that horses  
have enzyme X.

*Valid:*

Horses have enzyme X.  
Cows have enzyme X.  
So, it is  $M$  that horses  
have enzyme X.

This minimal manipulation allowed us to isolate the effect of the modal frame on acceptance rates both for valid vs. invalid arguments and for invalid arguments of varying strengths. We also used a larger set of epistemic modal expressions than were used in previous work, including *possible*, *plausible*, *likely*, *probable*, *necessary*, and *certain*.

## Methods

**Participants** 507 participants were recruited using Amazon’s Mechanical Turk platform, with a restriction to participants located in the United States. They were compensated for their participation.

**Materials** Participants saw 21 valid and invalid arguments with two premises and a conclusion. 18 of these included one of the 6 modal words listed above (3 for each modal), and 3 did not include a modal word. The properties used in the arguments were chosen from a list of unfamiliar biological and pseudo-biological properties. In every case, the animal in the conclusion was “horses” (Osherson et al., 1990). The arguments seen were randomly selected from a total of 63 arguments tested. 36 were non-valid arguments, in which the premise animals were chosen from the set {cows, chimps, gorillas, mice, squirrels, elephants, seals, rhinos, dolphins}. 27 were logically valid arguments with one of two forms, following (Rotello & Heit, 2009): one premise involved a mammal and the other was an *identity* premise (stating that horses have the property) or an *inclusion* premise (stating that mammals or animals have the property).

**Procedure** Participants were instructed to answer each question according to whether they agreed with the conclusion, assuming that the premises were true. For each question participants were asked to select “agree” or “disagree” and to give a confidence rating on a five-point scale.



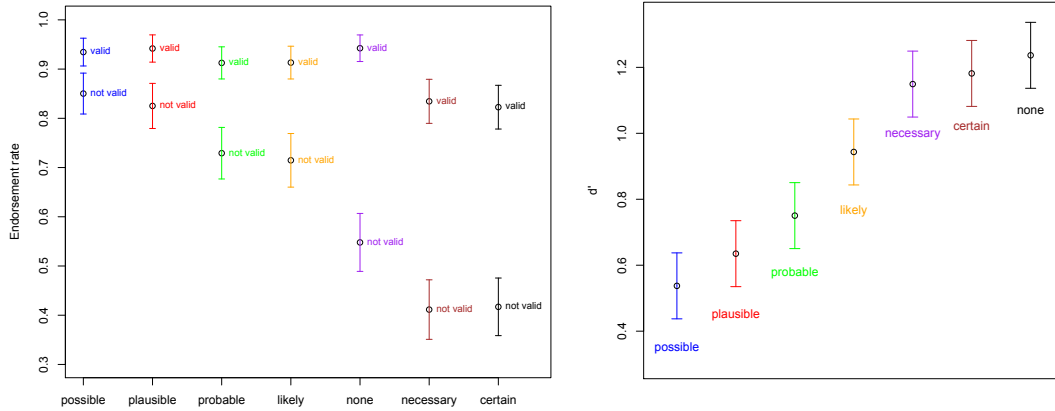


Figure 2: (a): Valid and invalid endorsement rates by modal frame with 95% CIs; (b):  $d'$  by modal frame with 95% CIs.

## Results

We analyzed data from the 485 participants who reported that their native language was English, for 10,185 total judgments. No systematic differences emerged between identity and inclusion arguments, and we group them together as valid.

Our results replicated the crucial finding of (Heit & Rotello, 2010; Rotello & Heit, 2009) showing that sensitivity to argument validity is greater when participants are asked to judge whether a conclusion is “necessary” than when they are asked whether it is “plausible” (Table 2). The difference in  $d'$  values is significant at the  $p < .05$  level.

Table 2: Acceptance rates and  $d'$  in our experiment (*plausible* and *necessary* only), with 95% confidence intervals.

	<i>Necessary</i>	<i>Plausible</i>
Acceptance, valid	.82	.94
Acceptance, invalid	.41	.82
Sensitivity ( $d'$ )	1.15 ± .19	0.63 ± .27

Comparing Table 1 with Table 2, there are two clear differences between our results and those of Rotello and Heit (2009): our participants rejected valid arguments with *necessary* more often than with *plausible*, and they accepted invalid arguments at much higher rates. These factors contributed to lower  $d'$  in both conditions. The first difference suggests that participants in our experiment judged some logically valid arguments as strong but not maximally strong. The second is plausibly due to the use of different materials in our experiment: Rotello & Heit used the same predicate “have property  $X$ ” in all arguments, using the variable  $X$  in the stimuli and instructing participants to treat property  $X$  as a novel biological property. The use of more natural biological predicates in our experiment may have encouraged participants to have greater confidence in their guesses, particularly if they had general background knowledge about e.g. enzymes and bones.

The fact that  $d'$  differed significantly for “necessary” and “plausible” suggests that our within-participants manipulation successfully captured the core features of between-participants manipulations in previous work. That is, the dif-

ference previously reported can be elicited by a difference of a single modal word, and so appears to be triggered by semantic properties of the modal words.

**Assessing Model Fit** As pointed out in introducing the model, our account predicts three general trends. First, it predicts that acceptance rates should vary less across valid arguments than across invalid arguments (Figure 1a). This is indeed what we find (Figure 2a). Second, it predicts the possibility of a continuous gradient in  $d'$  values (Figure 1b). Our results confirm this prediction as well (Figure 2b). Third, it predicts that the acceptance rate of argument  $A$  in any two modal frames  $M_1$  and  $M_2$  should be related by a power-law. This is the quantitative prediction that we now set out to test.

We fit the model to our results using the acceptance rates of each argument in the no-modal condition as a baseline. Note that we do not believe that acceptance rates in this condition are an estimate of the true probability of the conclusion given the premises. However, our model predicts that the choice of baseline condition should not affect predictions (cf. equation 4). In accord with this prediction we found no systematic effect on  $R^2$  by varying the choice of baseline. The primary effect of the choice of baseline condition, then, is that estimates of  $\alpha_M$  given for the other conditions are up to multiplication by  $\alpha_0$ , the parameter which determines the probability of arguments in the baseline condition.

Figure 3 plots the endorsement rate of each argument in the unmodalized condition against the endorsement rate for the same argument in various modal frames. We calculated the best-fit  $\alpha_M$  for each condition. For each graph in Figure 3 the curve determined by equation (2) is superimposed, with the overall best-fit noise parameter  $\epsilon = .12$ . As the  $R^2$  values in Figure 3 and Table 2 show, this model captures much of the variance in the data, but not all of it.

In order to discern whether the remaining variance was due to systematic factors that our model does not capture, we performed a split-half reliability test, randomly dividing the data into equal-sized halves 10,000 times and testing the correlation between the two halves on the same measure that



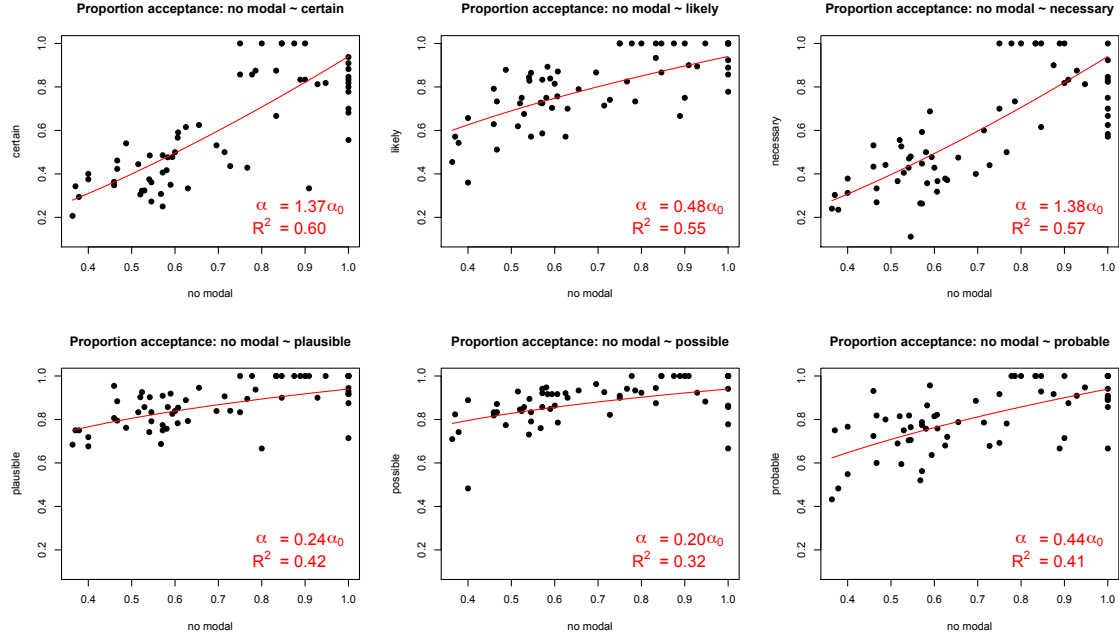


Figure 3: Endorsement rates by argument and modal frame. Curves represent model predictions using (2) and the best-fit  $\alpha$ .

was used to fit our model: acceptance rate by argument and modal. The model correlations were overall very close to the split-half correlations (Table 3). This suggests that the model captures most of the true structure in the data, though a good deal of noise remains that we cannot expect to explain.

Recall that the model predicts a consistent order in the endorsement rates of arguments across modal frames (see equation 4). In order to test this prediction we considered two tests, a permutation test and a model-based simulation. The average value of Spearman’s rank-order correlation for all pairwise comparisons between modal frames in our experiment was .53. A 10,000-sample permutation test revealed that this correlation was highly significant,  $p < .0001$ .

For the model-based test, we simulated 1,000 data sets with the same dimensions as our data using the best-fit model predictions as binomial proportions for each argument and modal frame. Equation 4 entails that all rank-order correlation in the model predictions are 1, and so this test gives us an opportunity to observe what correlations we should expect to see in a data set of this size if the prediction of a perfect underlying rank-order correlation is correct. The result was consistent with the observed rank-order correlation of .53, with 95% of the model-generated correlations falling between .49 and .64 (mean = .57). This result suggests that the model’s

monotonicity predictions fit the data well, providing further support for our claim that argument strength is based on a single scale which is manipulated by modal expressions.

## Relation to Formal Semantics

Epistemic modality has been the subject of much investigation in formal semantics (Kratzer, 1991; Egan & Weatherson, 2011). This paper has been concerned with a subtype of epistemic modals whose grammatical features indicate that they are *gradable adjectives*. This means that, like other gradable adjectives, they can be used in comparative and degree modification structures: for example, conclusion  $C$  might be *more plausible* than  $C'$  given some evidence, or  $C$  might be *very likely* or *almost necessary*. This corresponds to the fact that person  $x$  can be *taller* than person  $y$ , and a glass of water might be *very large* or *almost full*. Gradable expressions are generally treated in formal semantics as functions which map objects to points on a *scale* and compare them to *threshold values* (Kennedy, 2007). Recent empirical and formal work on gradable epistemic modals suggests that a scalar analysis is appropriate for them as well, and that probability is a good choice for the underlying scale (Lassiter, 2010; Yalcin, 2010).

Although our model is superficially different from linguistic models, there is in fact a straightforward translation between the two. Linguistic models assume that the location of the threshold is typically uncertain, a fact which is closely related to the problem of vagueness. If the uncertain location of the threshold for gradable expressions is modeled probabilistically as in (Frazee & Beaver, 2010; Lassiter, 2011; Schmidt, Goodman, Barner, & Tenenbaum, 2009), our model can be seen as describing the cumulative distribution of this threshold. Letting  $p_M(\theta)$  be the probability density of the unknown threshold associated with modal  $M$ , the cumulative distribu-

Table 3: Model correlations vs. split-half reliability results.

Modal	model $R^2$	mean split-half $R^2$
<i>certain</i>	.60	.66
<i>likely</i>	.55	.57
<i>necessary</i>	.57	.70
<i>plausible</i>	.42	.37
<i>possible</i>	.32	.34
<i>probable</i>	.41	.47

tion is given by (5), which — plugging in (1) — gives us (6).

$$pr(It\ is\ M\ that\ C|P) = \int_0^{pr(C|P)} p_M(\theta) d\theta \quad (5)$$

$$pr(C|P)^{\alpha_M} = \int_0^{pr(C|P)} p_M(\theta) d\theta \quad (6)$$

Using the fundamental theorem of calculus we can derive from equation (6) a simple formula linking  $p_M(\theta)$  to  $\alpha_M$ .

$$p_M(\theta) = \alpha_M \theta^{\alpha_M-1}, \quad 0 \leq \theta \leq 1. \quad (7)$$

This relationship allows us to interpret our model as an implementation of scalar semantics for gradable epistemic modals closely related to recent work in formal semantics.

## Conclusion

We have shown that non-linearities in responses to valid and invalid arguments can be explained by a simple probabilistic model, and thus are not evidence against a probabilistic account of reasoning. Our experimental manipulation involved the difference of a single word (an epistemic modal), and uncovered a gradation of non-linearities as a function of the specific modal used. Our model predicts a particular functional form for these differences, a power law in conditional probability, which we found in our data. This shows that it is possible to account for different patterns of results in reasoning experiments without assuming that everyday reasoning makes use of two (or more) qualitatively different types of reasoning. Rather, our model utilizes a single type of reasoning — probabilistic inference — together with a number of different but related linguistic mechanisms for talking about the results of inferential processes. This indicates that a one-dimensional theory of argument strength, coupled with an explicit formal semantics for epistemic modals, can account for a variety of patterns of reasoning in a parsimonious and explanatory way. This does not rule out a qualitative distinction between inductive and deductive reasoning, but it does call into question previous efforts to show that such a distinction is necessary to account for everyday human reasoning, suggesting instead that a single process may underlie both.

The probabilistic approach also suggests accounts of several related phenomena. For instance, the fact that argument length tends to affect plausibility judgments more than necessity judgments (Rotello & Heit, 2009) may be attributable to the fact that, in a probabilistic theory, we expect that adding premises of the type used in these experiments will usually increase the probability of the conclusion given the premises (Heit, 1998). The non-linearities predicted by equation 2 lead us to expect that the same change in probability will have different effects depending on the modal used.

Our probabilistic theory leaves open the psychological process by which people evaluate arguments. One possibility is that people are only able to *sample* possible worlds in accord with the distribution implied by the premises (Vul, Goodman, Griffiths, & Tenenbaum, 2009), and evaluate the truth of the

conclusion in these sampled worlds. If people take several samples and respond “yes” when the conclusion is true in each sampled world, we recover a power law. If the average number of samples depends on the modal, we recover the probabilistic model described above. For instance, we would posit that people tend to take more samples to evaluate “necessary” conclusions than “plausible” conclusions. This process-level implementation predicts that under time pressure, when people can take fewer samples, “necessary” would begin to look more like “plausible”. Indeed, this is exactly the finding of Heit and Rotello (2010).

We have illustrated an approach to reasoning based on an overall probabilistic view of inference, together with careful attention to natural language semantics. We believe that this approach will prove fruitful in studying a wide variety of phenomena related to human reasoning.

## Acknowledgements

We thank Evan Heit for helpful discussion of an earlier draft. This work was supported by a John S. McDonnell Foundation Scholar Award and ONR grant N00014-09-1-0124.

## References

- Egan, A., & Weatherson, B. (2011). *Epistemic modality*. OUP.
- Evans, J., & Over, D. (1996). *Rationality and reasoning*. Psychology Press.
- Frazee, J., & Beaver, D. (2010). Vagueness is rational under uncertainty. *Proceedings of the 17th Amsterdam Colloquium*.
- Harman, G. (1999). *Reasoning, meaning, and mind*. OUP.
- Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. *Rational models of cognition*, 248–274.
- Heit, E., & Rotello, C. (2010). Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(3).
- Johnson-Laird, P. (1994). Mental models and probabilistic thinking. *Cognition*, 50(1–3), 189–209.
- Kemp, C., & Tenenbaum, J. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, 116(1), 20–58.
- Kennedy, C. (2007). Vagueness and grammar. *Ling. & Phil.*, 30(1).
- Kratzer, A. (1991). Modality. In A. von Stechow & D. Wunderlich (Eds.), *Semantics: Int'l hbk. of contemp. research*. de Gruyter.
- Lassiter, D. (2010). Gradable epistemic modals, probability, and scale structure. In *Semantics and linguistic theory (SALT) 20*.
- Lassiter, D. (2011). Vagueness as probabilistic linguistic knowledge. In Nouwen et al. (ed.), *Vagueness in communication*. Springer.
- Macmillan, N., & Creelman, C. (2005). *Detection theory: A user's guide*. Lawrence Erlbaum.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.
- Osherson, D., Smith, E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psych. Review*, 97(2), 185.
- Rips, L. (2001). Two kinds of reasoning. *Psych. Science*, 12(2).
- Rotello, C., & Heit, E. (2009). Modeling the effects of argument length and validity on inductive and deductive reasoning. *J. Exp. Psych.: Learning, Memory, and Cognition*, 35(5).
- Schmidt, L. A., Goodman, N. D., Barner, D., & Tenenbaum, J. B. (2009). How tall is Tall? Compositionality, statistics, and gradable adjectives. In *Proc. 31st annual CogSci*.
- Tenenbaum, J., Griffiths, T., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7), 309–318.
- Vul, E., Goodman, N., Griffiths, T., & Tenenbaum, J. (2009). One and done? Optimal decisions from very few samples. In *Proc. 31st annual CogSci*.
- Yalcin, S. (2010). Probability Operators. *Phil. Compass*, 5(11).

# A Behavioral Investigation of Dimensionality Reduction

Joshua M. Lewis

josh@cogsci.ucsd.edu

Department of Cognitive Science  
University of California, San Diego

Laurens van der Maaten

lvdmaaten@gmail.com

Pattern Recognition & Bio-informatics Lab  
Delft University of Technology

Virginia R. de Sa

desa@cogsci.ucsd.edu

Department of Cognitive Science  
University of California, San Diego

## Abstract

A cornucopia of dimensionality reduction techniques have emerged over the past decade, leaving data analysts with a wide variety of choices for reducing their data. Means of evaluating and comparing low-dimensional embeddings useful for visualization, however, are very limited. When proposing a new technique it is common to simply show rival embeddings side-by-side and let human judgment determine which embedding is superior. This study investigates whether such human embedding evaluations are reliable, i.e., whether humans tend to agree on the quality of an embedding. We also investigate what types of embedding structures humans appreciate *a priori*. Our results reveal that, although experts are reasonably consistent in their evaluation of embeddings, novices generally disagree on the quality of an embedding. We discuss the impact of this result on the way dimensionality reduction researchers should present their results, and on applicability of dimensionality reduction outside of machine learning.

**Keywords:** dimensionality reduction; unsupervised machine learning; psychophysics

## Introduction

There is an evaluative vacuum in the dimensionality reduction literature. In many other unsupervised machine learning fields, such as density modeling, evaluation may be performed by measuring likelihoods of held-out test data. Alternatively, in domains such as topic modeling, human computation (Ahn, Maurer, McMillen, Abraham, & Blum, 2008) resources such as Amazon’s Mechanical Turk may be employed to exploit the fact that humans are phenoms in evaluating semantic structure (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009). Human evaluations have also been used to assess image segmentation techniques (Martin, Fowlkes, Tal, & Malik, 2001). The field of dimensionality reduction, however, lacks a standard evaluation measure (Venna, Peltonen, Nybo, Aidos, & Kaski, 2010), and is not as obvious a target for human intuition. Two or three dimensional embeddings can be visualized as scatter plots, but on what intuitive basis can we judge a 200 to 2-dimensional reduction to be good? In addition, Gestalt effects or simple rotations may bias human evaluations of scatter plots. Nevertheless, with no broadly agreed upon embedding quality measure (though a few have been proposed, see below), human judgment is often explicitly and implicitly solicited in the literature. The most common form of this solicitation consists of placing a scatter plot of the preferred embedding next to those of rival embeddings and inviting the reader to conclude that the preferred embedding is superior (e.g., (Maaten & Hinton, 2008)). If one is interested in applying a dimensionality reduction algorithm to visualize a dataset, is this a valid way to select from the

wide range of techniques?<sup>1</sup> To answer this question, we need to evaluate whether humans are good at evaluating embeddings. As there is no external authority we can appeal to, this is a daunting task. However, it is relatively easy to find out whether human data analysts are at least consistent in their evaluations, which is the first aim of this study. Consistency, across individuals and across a wide range of inputs, is a reasonable prerequisite for evaluation.

Beyond investigating whether human data analysts are consistent when they evaluate embeddings, the second aim of this study is to investigate what humans are doing when they evaluate embeddings. Such information could be useful for determining whether humans are appropriate for an evaluation task with a known structure (e.g. if they naturally prefer embedding characteristics appropriate to the structure), or for developing techniques that are tailored towards producing results that humans will find helpful (e.g. algorithms that selectively emphasize informative data structure). We can to some extent infer human strategies from the algorithms humans prefer, but we can also investigate those strategies by correlating embedding characteristics with human evaluations.

Motivated by the two aims described above, we solicit embedding quality judgments from both novice and expert subjects in an effort to determine whether they are consistent in their ratings, and which embedding characteristics they find appealing. For the novice subjects, we manipulate dataset knowledge—half read a description and see samples from each dataset, and half do not. We hypothesize that providing dataset information will increase consistency, as it should if the evaluative process is principled. The study consists of two experiments. The first presents subjects with a selection of embeddings derived from nine distinct dimensionality reduction algorithms; the second uses embeddings from a single algorithm with several different parameter settings for a more controlled comparison between “clustered” and “gradual” embeddings.

## Dimensionality reduction techniques

Dimensionality reduction techniques can be subdivided into several categories: linear or non-linear, convex or non-convex, parametric or non-parametric, etc. (Lee & Verleyesen, 2007). Whilst many new techniques have been proposed over the last decade, data analysts still often resort to linear, convex, parametric techniques such as PCA to visualize their

<sup>1</sup>Moreover, one should note that dimensionality reduction comprises only a small part of the “visualization zoo” (Heer, Bostock, & Ogievetsky, 2010).

data. Non-linear, convex, non-parametric manifold learners such as Isomap (Tenenbaum, Silva, & Langford, 2000), LLE (Roweis & Saul, 2000), and MVU (Weinberger, Sha, Zhu, & Saul, 2007) are also frequently used for visualization purposes (Jain & Saul, 2004; Lewis, Hull, Weinberger, & Saul, 2008; Mahecha, Martínez, Lischeid, & Beck, 2007), even though it is unclear whether these techniques produce superior results (Maaten & Hinton, 2008).

As described in the introduction, one of the key problems of dimensionality reduction is that it lacks a widely agreed upon evaluation measure (Venna et al., 2010). In fact, it is very unlikely that there will ever be such an evaluation measure, as it would imply the existence of a *free lunch* (Wolpert, 1996): the “optimal” dimensionality reduction technique would be the technique that optimizes the measure. Also, there is a lot of debate within the field on what a good objective for dimensionality reduction is: for instance, a latent variable model approach to dimensionality reduction suggests one should focus on preserving *global* data structure (Lawrence, 2005), whereas a manifold learning viewpoint deems preservation of *local* data structure more important (Roweis & Saul, 2000). The lack of an evaluation measure and the ongoing debate within the field motivate the use of a more anthropocentric approach.

In our study, we investigate nine dimensionality reduction techniques, selected for their importance in the literature: (1) PCA, (2) projection pursuit, (3) random projection, (4) Sammon mapping, (5) Isomap, (6) MVU, (7) LLE, (8) Laplacian Eigenmaps, and (9) t-SNE. PCA and projection pursuit find a subspace of the original data space that has some desired characteristic. For PCA, this subspace is the one that maximizes the variance of the projected data. For projection pursuit (Friedman & Tukey, 1974), the subspace maximizes the non-Gaussianity of the projected data. Random projections are linear mappings that mostly preserve pairwise distances in the data due to the Johnson-Lindenstrauss lemma (Bingham & Mannila, 2001). Sammon mapping constructs an embedding that minimizes a weighted sum of squared pairwise distance errors, where distances are weighted in inverse proportion to their magnitude (Sammon, 1969). Isomap constructs an embedding by performing classical scaling on a geodesic distance matrix that is obtained by running a shortest-path algorithm on the nearest neighbor graph of the data (Tenenbaum et al., 2000). MVU learns an embedding that maximizes data variance, while preserving the pairwise distances between each data point and its  $k$  nearest neighbors, by solving a semidefinite program (Weinberger et al., 2007). LLE constructs an embedding that preserves local data structure by minimizing a sum of squared errors between each map point and its reconstruction from its  $k$  nearest neighbors in the original data (Roweis & Saul, 2000). Laplacian Eigenmaps try to minimize the squared Euclidean distances between each points corresponding to its  $k$  nearest neighbors in the original data, while enforcing a covariance constraint on the solution (Belkin & Niyogi, 2002). t-SNE embeds points

by minimizing the divergence between two distributions over pairs of points, in which the probability of picking a pair of points decreases with their pairwise distance (Maaten & Hinton, 2008).

## Experimental setup

We perform two experiments with our human subjects. The first experiment uses stimuli generated from the dimensionality reduction algorithms described above to determine whether humans are consistent in their evaluations when the embeddings are fairly distinct (the first aim of the study). The second experiment uses stimuli that are all generated by t-SNE, but with different parameter settings that affect how clustered the resulting embedding appears. This helps us determine what type of structure humans generally prefer in embeddings (the second aim of our study).

### Experiment 1

In the first experiment, we divided subjects into (1) an expert group with detailed knowledge of dimensionality reduction and information on the underlying datasets presented in written and pictorial form, (2) a novice group with no knowledge of dimensionality reduction and no information on the visualized data, and (3) a group of similar novices but with the same information on the underlying datasets given to the experts. The dataset information we presented to groups 1 and 3 constituted of an intuitive description of the data, as well as images representing the underlying data (e.g., the Swiss roll, or handwritten character images).

Thirty one undergraduate human subjects were recruited for this study as the novice group, 16 female and 15 male, with an average age of 19.1 years. None of the novice subjects had any specific knowledge of dimensionality reduction techniques. Our expert group consisted of five subjects—three graduate students and two faculty members. The expert subjects were drawn from the same institution and represent two different departments. Amongst the five expert subjects there are four distinct academic backgrounds at the graduate level. The informed novice group had 15 subjects and the uninformed novice group 16. We generated two-dimensional point-light stimuli (see Figure 1 for a visualization of all the stimuli) by running the nine dimensionality reduction techniques discussed in Section on seven different high-dimensional datasets, comprising a variety of domains. We ran each technique for a reasonable range of parameter settings, and we selected the embedding that was best in terms of the trustworthiness<sup>2</sup> (Venna & Kaski, 2006) for presentation to the subjects.

Each trial consisted of nine different embeddings of the same dataset arranged randomly per trial in a  $3 \times 3$  grid. The datasets were shown as scatter plots with white points on a black background to reduce brightness-related eye fatigue. For novice subjects, trials were organized into three blocks

<sup>2</sup>The trustworthiness measures the ratio of  $k$  nearest neighbors in the data that is still among the  $k$  nearest neighbors in the maps.

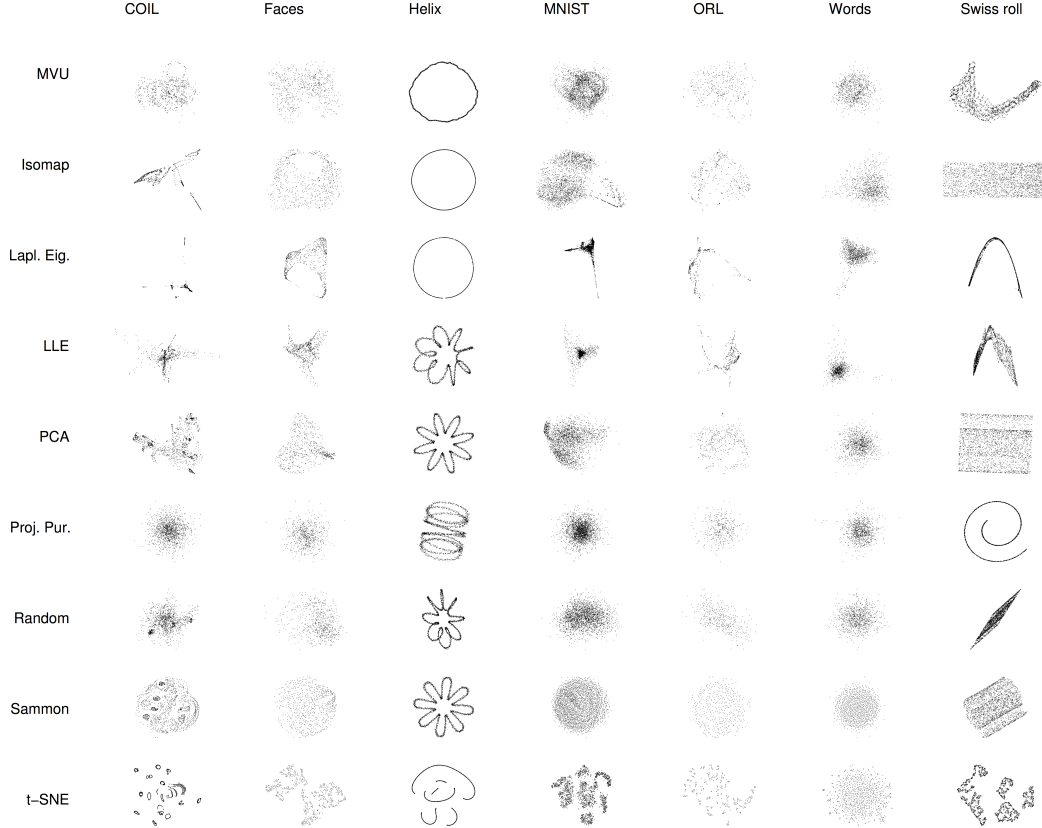


Figure 1: All stimuli from experiment 1. Methods are in rows; datasets are in columns.

of seven, where each dataset appeared once per block and the order of the datasets within each block was randomized. Expert subjects were tested on one block. We instructed subjects to choose the two most useful displays and the one least useful display from the nine available on every trial. After describing what a scatter plot is and emphasizing that each set of nine plots is a different perspective on the same dataset, we gave subjects the following instructions: *For each trial, please examine all the scatter plots and choose the two that you find most useful and the one that you find least useful. The task in the second part of this experiment will be much faster and easier if you choose useful scatter plots. Do the best you can to choose useful plots based on whatever criteria you deem appropriate.* We intentionally left the task ambiguous so as not to bias subjects towards particular evaluation criteria<sup>3</sup>, and the fictitious “second part” of the experiment existed solely for increasing subject motivation.

We analyzed our novice subjects for internal consistency of their positive and negative ratings across blocks and found that even our least consistent subject was more consistent than expected by chance. Hence, we did not exclude any subjects due to internal inconsistency. To analyze consistency across subjects (the first aim of this study) we use Fleiss’  $\kappa$  (Fleiss,

1971) and include neutral responses. Fleiss’  $\kappa$  measures the deviation between observed agreement and the agreement attributable to chance given the relative frequency of ratings, and normalizes for the number of raters. Neutral ratings are twice as frequent as non-neutral, and positive ratings are twice as frequent as negative ratings, so the compensation for relative frequency in Fleiss’  $\kappa$  makes it well-suited to our data.

We also measured the following six characteristics of our embedding stimuli: (1) variance, (2) skewness, (3) kurtosis, (4) clusteredness, (5) visual span, and (6) Gaussianity. The variance, skewness, and kurtosis were measured per dimension in a map that was scale-normalized such that  $\mathbf{y}_i \in [0, 1]^d$  (preserving the aspect ratio of the maps), and averaged over the  $d$  dimensions of the map. We measured clusteredness by constructing  $k$ -nearest neighbor graphs in the map with  $k = 3, \dots, 12$ , and measuring the maximum clustering coefficient of the resulting graphs (Watts & Strogatz, 1998). The clustering coefficient computes the ratio of connections between the adjacent vertices of map point  $i$ , averaged over all map points. The visual span of each map was measured by fitting a Parzen kernel density estimator with Gaussian kernels on the map (the variance  $\sigma$  of the Gaussians was optimized on a small validation set). Subsequently, we measure the ratio of the map surface that has a density of at least 10% of the max-

<sup>3</sup>For instance, defining a classification task would bias subjects to embeddings that show separated clusters.

imum density of the density estimate. The Gaussianity of the maps was determined by averaging the results of Lilliefors tests (Lilliefors, 1967) performed on 5,000 one-dimensional random projections of the map<sup>4</sup>. We analyze the relationships between novice informed ratings, novice uninformed ratings, expert ratings, and the six quantitative measures by calculating the Pearson’s correlation coefficient  $\rho$  between ratings and measures (after normalization within trial).

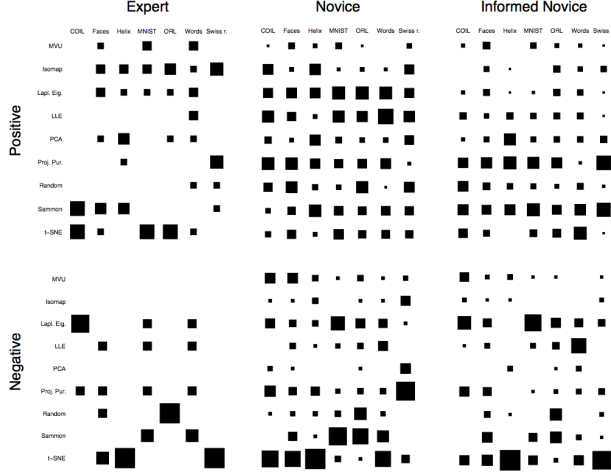


Figure 2: Human responses to the embeddings in experiment 1. Positive responses in the first row, negative in the second row. Experts (left), novices (center) and informed novices (right) by column. Algorithm and dataset ordering are the same as in Figure 1 within each block.

## Experiment 2

The second experiment was run directly following experiment 1 on the same subject pool using the same methods, save stimulus design. In experiment 2, the nine stimuli in each trial are obtained by running t-SNE with nine different degrees of freedom  $v$  (viz.  $v = 0.5, 0.8, 1.0, 1.2, 1.5, 2.0, 2.5, 3.0, 4.0$ ) on the seven datasets. The degrees of freedom in t-SNE determine to what extent the visualizations are “clustered” (Maaten, 2009). Sample stimuli are shown in Figure 3.

## Results

### Experiment 1

For the first experiment, the Fleiss’ kappa consistency measure  $\kappa$  for experts is 0.39, for uninformed novices is  $-0.28$ , and for informed novices is  $-0.40$ . Fleiss’ kappa  $\kappa$  ranges from  $-1$  to  $+1$ , with  $-1$  representing complete disagreement,  $+1$  representing complete agreement and  $0$  representing the amount of agreement expected by chance. Though there is no standard significance test for Fleiss’ kappa, based on the Landis and Koch scale (Landis & Koch, 1977), experts exhibited fair to moderate agreement, while both groups of novices

<sup>4</sup>Note that if the distribution of points in the embedding is Gaussian, the point distribution in each of the random projections has to be Gaussian as well.

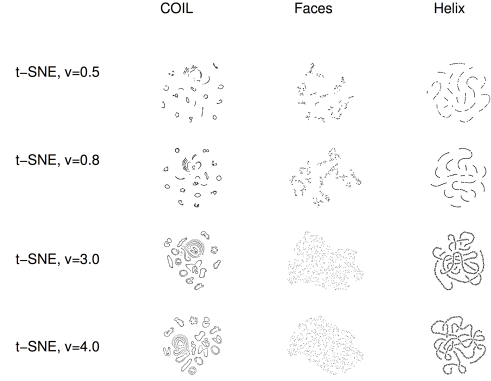


Figure 3: Sample stimuli from experiment 2. Parameter values are in rows; datasets are in columns.

exhibited poor agreement. Hence, the consistency measures reveal that, whereas experts tend to agree with each other on the quality of an embedding, novices strongly disagree with each other in their evaluations (they disagree more than if the evaluation was done randomly). Surprisingly, novices who received information on the underlying data disagree more strongly with each other than novices who had no information about the underlying data (counter to our hypothesis but interpretable, see below).

Table 1: Correlation coefficients between human responses and dataset characteristics. Text in bold if  $p < .0036$  after Bonferroni correction for  $n = 14$  comparisons per subject group and  $\alpha = .05$ .

$\rho$	Lilliefors	Skewness	Kurtosis	Variance	Visual Span	Clusteredness	Trustworthiness
Ex. Pos.	.26	-.01	-.19	.34	.17	.22	<b>.41</b>
Ex. Neg.	-.08	.17	.19	-.14	-.17	.08	-.08
Nov. Pos.	.07	-.03	<b>.50</b>	-.18	-.29	.01	-.08
Nov. Neg.	.00	.17	-.10	.22	.10	-.03	.24
Inf. Pos.	-.02	-.16	-.10	-.11	<b>.44</b>	<b>-.45</b>	-.09
Inf. Neg.	.03	.31	.19	.10	-.19	.20	.19

In Figure 2, we depict the raw ratings (averaged over each group) as a collection of Hinton diagrams. In the figure, a large square indicates that a relatively large number of subjects gave a positive or negative evaluation of the embedding of the corresponding dataset, constructed by the corresponding technique. The top row of diagrams represent positive responses and the bottom negative, so if subjects are in disagreement about a stimulus, there will be a large box in its corresponding location in both rows. The diagrams reveal that informed novices exploit dataset knowledge in specific



instances to differ significantly from uninformed novices. For example, the t-SNE embedding of the Swiss roll dataset (a relatively clustered embedding) is rated much more negatively by novices when they know that the data are gradual. Experts tend to rate t-SNE positively or negatively depending on the dataset and show a relatively consistent rating for Isomap. Informed novices consistently rated Sammon mapping and projection pursuit positively while generally rating manifold learners such as Isomap and LLE negatively. Uninformed novices are all over the map with the exception of (like all other subjects) rating MVU as not notable in either a positive or negative sense.

Table 1 shows correlation coefficients between the six embedding characteristics and the evaluations by the three human groups. We also present the correlation between the evaluations and the trustworthiness, which gives an indication of the quality of the embedding (in terms of local neighborhood preservations). The significant correlations are in bold type, after a Bonferroni correction for multiple comparisons (14 comparisons per subject group,  $\alpha = .05$ ). Notably, expert positive ratings are the only ratings that correlate significantly and in the correct direction with trustworthiness. Another correlation that stands out is visual span: it appears to play a substantial role in informed novice ratings (they apparently surmise an embedding should fill up the available space), whereas it plays little role in expert ratings.

## Experiment 2

For the second experiment, the consistency measure  $\kappa$  for experts is 0.35, for uninformed novices is  $-0.32$ , and for informed novices is  $-0.26$ . The results of the second experiment thus reveal a similar trend: experts have fair agreement on the quality of embeddings, whereas novices give ratings have poor agreement. The ratings reveal that, whereas experts selectively rate more clustered or more continuous embeddings positively depending on the dataset, novices overwhelmingly rate the more clustered embeddings as negative. On the other hand, for positive ratings the novices tend to choose embeddings at either end of the spectrum while avoiding the moderate values of  $v$ . Moderate values of  $v$  might be avoided since subjects want to classify displays closest to the prototypical clustered or graded display (Rosch, 1975). Using the same set of correlations from Experiment 1 we find that experts ratings do not strongly correlate with any of the characteristics (including trustworthiness), but both groups of novices show a correlation between negative ratings and those stimuli with low kurtosis and high clusteredness.

## Discussion

In both experiments, experts show themselves to be more consistent than chance and much more consistent than novices in either condition. This is reassuring, and indicates that the idea of having experts evaluating embeddings is not flawed to begin with. In the first experiment, novice subjects actually get less consistent with each other if they are informed. While this seems troubling at first, it actually makes

some sense after closer consideration. Comparing the Hinton diagrams between novices and informed novices, one can plainly see that informed novices are converging on a smaller selection of techniques for both positive and negative ratings. The issue for the informed novices, however, is that they are not sure whether these stimuli should be rated as positive or negative. As a result, there is often energy for the same cell in both diagrams. Since the base rate of positive and negative ratings is low compared to the neutral ratings, the  $\kappa$  consistency measure interprets this as substantial disagreement and thus the negative numbers. Importantly, the informed novice  $\kappa$  is further from chance level than the novice  $\kappa$ . In Experiment 2, uninformed novices actually differ more from chance but the effect is about half the size, and experts remain consistent.

Expert ratings are laudable in that they correlate in the correct direction with trustworthiness and have a context-dependent appreciation of clusteredness. Both novice groups rate clusteredness negatively regardless of context and are more influenced by elementary characteristics such as visual span. The substantial difference in strategy between novices and experts indicates that novices could really benefit from training on the task of evaluating embeddings (unlike evaluating results from topic modeling, image segmentation, or object recognition).

## Conclusion

With respect to the first aim of our study (determining whether humans are consistent in rating embeddings), we conclude that dimensionality reduction experts are indeed reasonably consistent judges of embedding quality. This supports the practice of soliciting expert judgment for embedding evaluations, as nowadays is common in the literature on dimensionality reduction. However, we also conclude that novices are very inconsistent with one another in terms of their rating of an embedding, even when they have detailed information on the dataset the embedding is visualizing. In fact, novices even correlate negatively with a measure of embedding quality.

With respect to the second aim of our study (determining what types of structure humans appreciate in embeddings), we conclude that humans do not appear to have overwhelmingly strong *a priori* preferences, although novices appear to appreciate gradual embeddings that span a large portion of the space. Experts can adapt their preference for gradual vs. clustered depending on the dataset.

Overall, our results discourage free-form solicitation of human computation approaches à la (Chang et al., 2009) and (Martin et al., 2001) to the evaluation of dimensionality reduction techniques. Moreover, the novices' lack of consistency lends worry to the prospect of naïve dimensionality reduction-based analysis. Most data analysts seeking to apply dimensionality reduction are not very familiar with the field. As a result, they are likely to be susceptible to the favorable visualizations presented in many dimensionality re-



duction papers. To ensure that dimensionality reduction techniques are applied wisely, authors should strive to explicate the dataset characteristics that favor their algorithms (e.g., t-SNE is useful if the data is expected to have cluster structure, Isomap if the data lie on a convex manifold). Authors could also cover usage scenarios appropriate to their algorithm (e.g., if a researcher is interested only in visualizing points that are most different then PCA would suffice and other techniques would be overkill), including guidelines for interpreting the relationship between the high and low dimensional spaces (sometimes this relationship will be very clear, as in PCA; other times, as in MVU, there is not a clear relationship). In addition, data analysts should be encouraged to use sanity checks such as the trustworthiness measure in order to prevent them from basing analysis on interesting, but flawed, embeddings.

### Acknowledgments

This work is funded by NSF Grant #SES-0963071, Divvy: Robust and Interactive Cluster Analysis (PI Virginia de Sa).

### References

- Ahn, L. von, Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, 321(5895), 1465-1468.
- Belkin, M., & Niyogi, P. (2002). Laplacian Eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems* (Vol. 14, pp. 585-591). Cambridge, MA, USA: The MIT Press.
- Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the 7<sup>th</sup> acm sigkdd* (pp. 245-250).
- Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (Vol. 21).
- Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378-382.
- Friedman, J., & Tukey, J. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 23, 881-890.
- Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *ACM Queue*, 8(5).
- Jain, V., & Saul, L. (2004). Exploratory analysis and visualization of speech and music by locally linear embedding. In *Proceedings of the international conference of speech, acoustics, and signal processing* (Vol. 3, pp. 984-987).
- Landis, J. R., & Koch, G. G. (1977, March). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.
- Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6(Nov), 1783-1816.
- Lee, J., & Verleysen, M. (2007). *Nonlinear dimensionality reduction*. New York, NY, USA: Springer.
- Lewis, J., Hull, P. M., Weinberger, K., & Saul, L. (2008). Mapping uncharted waters: exploratory analysis, visualization, and clustering of oceanographic data. In *Proceedings of the international conference on machine learning and applications* (pp. 388-395).
- Lilliefors, H. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62, 399-402.
- Maaten, L. van der. (2009). Learning a parametric embedding by preserving local structure. In *Proceedings of the 12<sup>th</sup> international conference on artificial intelligence and statistics* (pp. 384-391).
- Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov), 2431-2456.
- Mahecha, M., Martínez, A., Lischied, G., & Beck, E. (2007). Nonlinear dimensionality reduction: Alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. *Ecological Informatics*, 2, 138-149.
- Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001, July). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the 8<sup>th</sup> international conference on computer vision* (Vol. 2, pp. 416-423).
- Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, 7(4), 532 - 547.
- Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by Locally Linear Embedding. *Science*, 290(5500), 2323-2326.
- Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5), 401-409.
- Tenenbaum, J., Silva, V. de, & Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2319-2323.
- Venna, J., & Kaski, S. (2006). Visualizing gene interaction graphs with local multidimensional scaling. In *Proceedings of the 14<sup>th</sup> european symposium on artificial neural networks* (pp. 557-562).
- Venna, J., Peltonen, J., Nybo, K., Aidos, H., & Kaski, S. (2010). Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11(Feb), 451-490.
- Watts, D., & Strogatz, S. (1998). Collective dynamics of small-world networks. *Nature*, 393, 440-442.
- Weinberger, K., Sha, F., Zhu, Q., & Saul, L. (2007). Graph Laplacian regularization for large-scale semidefinite programming. In *Advances in neural information processing systems* (Vol. 19).
- Wolpert, D. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8, 1341-1390.

# Modeling Melodic Perception as Relational Learning Using a Symbolic-Connectionist Architecture (DORA)

**Ahnate Lim (ahnate@hawaii.edu)**

Department of Psychology, University of Hawaii at Manoa  
2530 Dole Street, Honolulu, HI 96822 USA

**Leonidas A. A. Doulas (leonidas@hawaii.edu)**

Department of Psychology, University of Hawaii at Manoa  
2530 Dole Street, Honolulu, HI 96822 USA

**Scott Sinnott (ssinnott@hawaii.edu)**

Department of Psychology, University of Hawaii at Manoa  
2530 Dole Street, Honolulu, HI 96822 USA

## Abstract

Like many other cognitive processes, the perception of music involves processes and structural considerations that are highly relational in nature. To date, no physiologically plausible model has been used to simulate and explain how infants perceive melodic content. Here we used DORA (*Discovery Of Relations by Analogy*; Doulas et al., 2008), a domain-general symbolic-connectionist model of relational learning, to simulate melodic perception and categorization by infants (Chang & Trehub, 1977; Trehub et al, 1984), and to provide an account of the mechanism for melodic processing in infants. Given four input semantic features for each note in the melodic stimuli sequence (two of which could be internally obtained from the other two via a comparator), DORA's performance matched the behavioral data from the infant studies. Furthermore, the ability of our model to simulate infants' behavior is evidence that structured representations of relational musical properties can be bootstrapped from unstructured feature representations.

**Keywords:** Melodic perception; relative pitch; relational learning; symbolic connectionist; DORA.

## Introduction

While there are many defining characteristics of music (e.g., harmony, rhythm, timbre, pitch, etc.), one of the most fundamental and salient aspects is the melody. Indeed, simple melodies were likely the earliest form of music to have been created and transmitted, and have been (and still are) prevalent in all documented cultures past and present (Sachs, 1943).

Simple melodies consist of discrete units or notes, with each note characterized by a pitch, or fundamental frequency. Importantly, there are several ways in which the pitch sequence of a melody can then be encoded and stored. The two most well documented forms of encoding are *absolute* or *relative* pitch. Absolute pitch encodes and stores a melody using the fundamental frequencies of each pitch, while relative pitch (or intervallic) encodes the melody in terms of the *relations* (or specific frequency differences) between each note. Notably, processing melodies in terms of the relative pitch information (or intervallic patterns) is considered to be the strategy most humans use to

characterize and store familiar melodies (Attneave & Olson, 1971). Another characteristic upon which melodies can be categorized, however, is according to the contour (general shape, or sequence of up and down movements in frequencies from note to note). Given the existence of these various characteristics, there has been considerable research and speculation on the extent to which these categorizations contribute to a listener's mental representation of a melody, and how they may interact.

## Relative pitch and melodic contour

The properties of relative pitch (or intervallic patterns) are most commonly used in long term musical storage and recall (Page, 1994). For example, when listening to the melody of a song, such as *Happy Birthday*, what makes the song immediately recognizable is the unique intervals between each of the notes in the song. That is, the song is recognizable whether it initially starts on a low or high note due to the unique intervallic pattern between all subsequent notes. There is much evidence on the use of relative pitch information in adults through both behavioral studies (Dowling, 1978, 1984, 1988) as well as neuroimaging studies (Fujioka, Trainor, Ross, Kakigi, & Pantev, 2004; Trainor, McDonald, & Alain, 2002).

It is worth noting that while a melody with an identical contour to *Happy Birthday*, but with a different intervallic sequence would sound like a completely different tune, it would still have the same general "shape", or up and down pattern. Although the intervallic pattern may be the most overtly salient feature of a melody, studies have shown that human adults are also sensitive to absolute pitch and melodic contour in the short term (Bartlett & Dowling, 1980; Dowling, 1978). And while there is evidence that infants may also be sensitive to intervallic information (Trehub, Bull, & Thorpe, 1984), numerous experiments with infants suggest that they may primarily encode melodies using contour information (for review, see Trehub, 2001; Trehub, Trainor, & Unyk, 1993).

Even though intervallic and contour properties may characteristically differ in the type of information they

carry, what is perhaps more important is the fact that the nature of the information they carry are both fundamentally relational in nature. That is, this information depends on the relationship (whether it is the precise intervallic distances or the general contour shape) between each pitch, and not on the actual pitch frequencies themselves. And it is within this capacity that melodic perception can be said to share a cornerstone property with many higher-level cognitive tasks (and arguably certain “lower” level processes such as pattern recognition as well).

### Relational processing

The ability to explicitly represent and reason about relational properties has been proposed as a fundamental mechanism underlying a wide range of cognitive phenomenon, including analogy-making (Gentner, 1983; Gick & Holyoak, 1980; Holyoak & Thagard, 1995), language (Kim, Pinker, Prince, & Prasada, 1991), detection of perceptual similarities (Medin, Goldstone, & Gentner, 1993), and the application of rules in novel situations (Lovett & Anderson, 2005). Given that melodic processing appears to require extracting relational information from melodies, it is reasonable to assume that the same mechanisms used in other relational tasks might also operate when processing musical information. That is, common to both of the main approaches used by adults and infants (intervallic and contour) to encode melodic information is that the underlying structure of the melody is represented as the relations between the individual notes. The strength of relational reasoning is in the ability to reason about the roles that objects play rather than the literal features of those objects (see Doumas, Hummel, & Sandhofer, 2008). Similarly, the ability to recognize a melody (or its shape) rests on appreciating the relationship between the pitches, and not the specific frequencies of each note. To evaluate the similarity between relational reasoning and music processing, we modeled melodic perception using a neurally plausible domain-general model of relational cognition.

### The LISA/DORA models

LISA (*Learning and Inference with Schemas and Analogies*; Hummel & Holyoak, 1997, 2003) is a symbolic-connectionist model of analogy and relational reasoning. DORA (*Discovery Of Relations by Analogy*; Doumas et al., 2008) is an extension of LISA that learns structured (i.e., symbolic) representations of relations from unstructured inputs. That is, DORA provides an account of how the structured relational representations LISA uses to perform relational reasoning can be learned from examples. At present, DORA accounts for over 30 phenomena from the literature on relational learning, and cognitive development, and as it learns representations of relations it develops into LISA and can simulate the additional 40+ phenomena in relational thinking for which LISA accounts for (e.g., Doumas et al., 2008). In the following, we provide a very brief description of the LISA/DORA models (for full details, see Hummel & Holyoak, 1997, 2003; Doumas et al., 2008).

**LISAese Representations** In LISA (and DORA *after* it has gone through learning), relational structures are represented by a hierarchy of distributed and localist codes (see Figure 1). At the bottom, “semantic” units (small circles in Figure 1) represent the features of objects and roles in a distributed fashion. At the next level, these distributed representations are connected to localist units (POs) representing individual predicates (or roles) and objects (triangles and larger circles in Figure 1). Localist role-binding units (RBs; rectangles in Figure 1) link object and role units into specific role-filler bindings. At the top of the hierarchy, localist P units (ovals in Figure 1) link RBs into whole relational propositions.

Relational structures (or propositions) are divided into two mutually exclusive sets: a driver and recipient(s). In LISA/DORA, the sequence of firing events is controlled by the driver. Specifically, one (or at most three) proposition(s) in the driver become(s) active (i.e., enter working memory). When a proposition enters working memory, role-filler bindings must be represented dynamically on the units that maintain role-filler independence (i.e., POs and semantic units) to allow for reusability of units and preservation of similarity across different bindings (Hummel & Holyoak, 1997). In LISA, binding information is carried by synchrony of firing (with roles firing simultaneously with their fillers). In DORA, binding information is carried by systematic asynchrony of firing, with bound role-filler pairs firing in direct sequence (for details, see Doumas et al., 2008).<sup>1</sup>

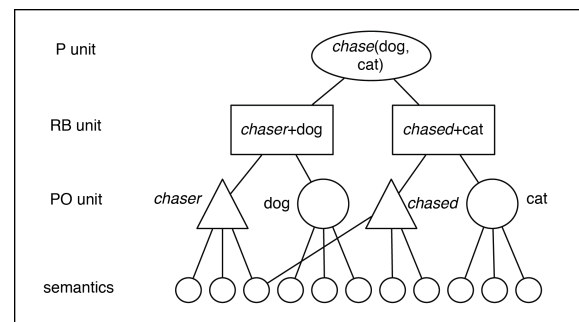


Figure 1. LISA/DORA representation of the proposition, *chase* (dog, cat).

**Relational Learning** In broadest strokes, DORA learns structured representations by comparing objects to isolate their shared properties and to represent these shared properties as explicit structures. More specifically, DORA starts with simple feature-vector representations of objects (i.e., a node connected to set of features describing that object; large and small circles from Figure 1). When DORA compares one object to another, corresponding elements (i.e., shared features) of the two representations fire simultaneously. Any semantic features common to both

<sup>1</sup> Asynchrony-based binding allows role and filler to be coded by the same pool of semantic units, which allows DORA to learn representations of relations from representations of objects (Doumas et al., 2008).

objects receive twice as much input and thus become roughly twice as active as features connected to one but not the other. By recruiting a new PO unit and learning connections between that unit and active semantics via Hebbian learning (wherein the strength of connections is a function of the units' activation), DORA learns stronger connections between the new PO unit and more active semantic units. The new PO thus becomes an explicit representation of the featural overlap of the compared objects and can act as a single place predicate, taking other object representations as arguments to form role-filler pairs (see Doumas et al., 2008). Applied iteratively, this process allows DORA to learn structured explicit single-place predicate representations of any properties compared objects may share. Comparison also allows DORA to learn representations of multi-place relations by linking sets of constituent role-filler pairs into relational structures (i.e., to learn the *chases* relation by linking together representations of the roles *chaser* and *chased*; see Doumas et al., 2008 for details).

**Mapping** For the purpose of analogical mapping, LISA/DORA learns *mapping connections* between units coactive of the same type in the driver and recipient (e.g., between PO units in the driver and PO units in the recipient). These connections grow whenever corresponding units in the driver and recipient are active simultaneously. They permit LISA to learn the correspondences between matching structures in separate analogs. They also permit correspondences learned early in mapping to influence the correspondences learned later.

## Methods

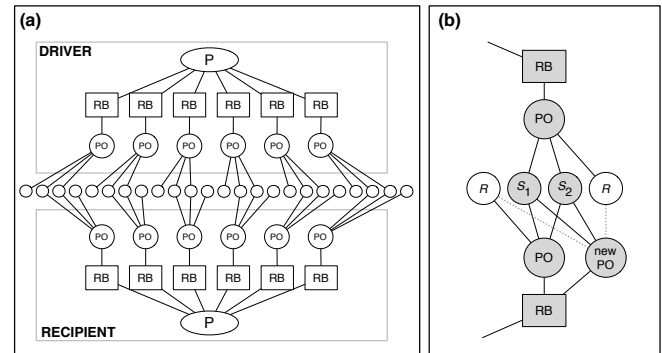
In this section we describe two infant studies (Chang & Trehub, 1977; Trehub et al., 1984), followed by the details and outcomes of DORA's simulation.

### Task 1 description

In an experiment by Chang and Trehub (1977), infants (4.5 to 6 months of age) were tested on their ability to recognize melodies based on either the absolute pitch frequencies or relational properties extracted from these pitches. This between-group experiment was conducted with a set of 15 habituating trials, followed by four novel dishabituation trials, while the infants' heart rates were monitored throughout to determine their expectation and recognition levels for the novel stimuli. The habituation stimuli consisted of randomly constructed six note melodic patterns. The dishabituation stimuli varied depending on which of two groups the infants were in.

Crucially, in the "transposed" group, the novel test stimulus consisted of the same melody transposed to a different key. The novel melody retained the relational information between the individual notes (intervallic sequence), but none of the featural information (specific frequencies) of the individual notes. In the control group, the novel melody was a scrambled version of the original melody. The individual notes' featural characteristics (pitch

frequencies) were retained, while the relational characteristics between the notes were not. Thus, comparing performance between the "transposed" and control groups would indicate whether infants were processing the melodies based on the individual frequencies, or extracting the relational information between notes.



Figures 2. a) The note sequences in the driver and recipient are compared with each other. b) After mapping notes in the sequence, DORA learns a new PO unit from the featural overlap of the mapped notes.  $S_n$  are semantic units, and  $R$  are random units (only two each are shown here).

### Simulation 1

To simulate the training portion of the study, we created a "melody" consisting of 6 object PO units—one PO for each note (see Figure 2a). Each note PO unit was attached to four random semantic units (chosen from a pool of 100 features), one semantic indicating the note's place in the stimuli sequence (1-6), one semantic describing the note's specific frequency (between f1 and f24), one semantic for whether the note was higher or lower than the previous note (the first PO in the sequence was not connected to such a semantic), and one semantic describing the note's distance (i.e., frequency difference) from the previous note. The information the semantic units carried was based on features which infants have been shown to be capable of extracting from melodies to greater or lesser extents. For instance, infants have been shown to be sensitive to sequential order (Thorpe & Trehub, 1989; Thorpe, Trehub, Morrongiello, & Bull, 1988), are sensitive to and can discriminate absolute pitch information under certain conditions (Lynch, Eilers, Oller, & Urbano, 1990; Trehub, Cohen, Thorpe, & Morrongiello, 1986), can process contour information (Trehub et al., 1984; Trehub et al., 1993), and are also sensitive to intervallic differences (Schellenberg & Trehub, 1996a, 1996b). Importantly, the semantic values specifying frequency direction and frequency difference can be generated from the raw frequency values using the comparator mechanism described in Doumas et al. (2008) and adopted from the JIM model of object recognition (Hummel & Biederman, 1992). Finally, each PO was attached to an RB unit, and all the RBs attached to a single P unit, representing that the notes all belonged to a single sequence.

We allowed DORA to compare each note sequence to the previously experienced note sequence, map the two sequences, and learn new predicate POs using the predication learning algorithm described above. The sequence of firing of the PO units in the driver was the same as the order of the notes in the melodic sequence (i.e., the first note in the sequence fired first, the second note second, and so on). More specifically, DORA represented the current note sequence in the driver and the previous note sequence in the recipient (see Figure 2a). Next DORA attempted to map the sequences. Finally, DORA learned new PO units using these mappings (Figure 2b). DORA stored the results of learning in memory.

In previous studies, DORA has successfully been used to simulate frontal lobe maturation by adjusting the level of lateral inhibition between units in the recipient (e.g., Dumas, Morrison, & Richland, 2009, 2010; Morrison, Dumas, & Richland, 2006, 2011). Reflecting the fact that we are simulating infants we used a highly reduced lateral inhibition parameter of 0.5.

After training, DORA's LTM consisted of the 15 sequences of notes it had learned during training. In addition, we created 50 additional sequences of between 2 and 8 notes to serve as distractors in memory (following the assumption that other melodic sequences may have been learned by the infant). In each distracter note sequence each PO from the same sequence was attached to a single RB unit with all RBs from the same sequence attached to a single P unit, indicating all the notes belonged to a single sequence (as with the training items). Each PO was attached to 4 random features as well as one semantic indicating the note and another semantic indicating the interval (frequency difference) from the previous note.

To simulate the test, we created two melodies, each consisting of 6 PO units, each representing a single note. Each note PO unit was attached to four random semantics, one semantic unit describing the note's place in the sequence (1-6), one semantic describing the frequency, one describing frequency difference from the previous note, and another describing the direction of the difference (just as for the melodies created in the training condition). Additionally, as originally conducted by Chang and Trehub (1977), we created a transposed melody that consisted of the same sequence of notes as the training melodies, but in a different key. The control melody consisted of the exact same POs used in training, but in a scrambled order.

We put the test melody (transposed or control) in the driver, and allowed DORA to attempt to retrieve an item from LTM, and attempt to map it to the melody in the driver. If DORA successfully mapped the new melody to one of the sequences it had learned during training, this implied that DORA recognized the new melody. Otherwise, DORA was taken to be surprised by the new melody.

We ran 200 simulations (100 transposed and 100 control), each consisting of 15 training and one test trial (the exact same number of training and test trials used in the original study). DORA's performance was a close qualitative match

to the data from Chang and Trehub's (1977) study. Just like infants in the transposed condition, when presented with transposed melodies, DORA was much more frequently unsurprised (77 of 100 trials). On the other hand, for control melodies, DORA was unsurprised much less frequently (31 of 100 trials). These results indicate, that DORA, like the infants in the original study, could detect and extract regularity in melodic sequences, and generalize that regularity to novel keys.

## Task 2 description

For the second simulation, we used a study by Trehub, Bull, and Thorpe (1984). This study was conducted on infants 8 to 11 months of age, and used a broader range of melodic stimuli to examine the extent to which infants process intervallic, contour, octave transpositions, and range information from melodies. Although the first task (Chang & Trehub, 1977) demonstrated that infants used relational information to categorize melodies, the design did not specifically differentiate between intervallic and contour relations (it is possible that infants could have used either strategy to categorize the melodies).

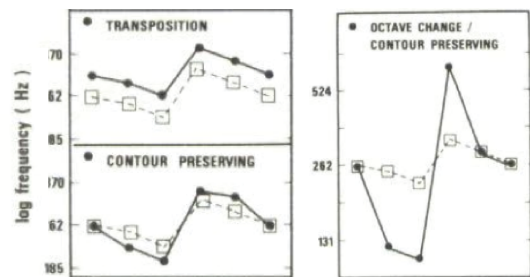


Figure 3. Three types of transformations applied to test melodies in Trehub, Bull, & Thorpe (1984).

Although a similar paradigm to Chang and Trehub's (1977) study was used by Trehub et al. (1984), the two studies differed in two important aspects. First, the training and testing methodology was different. Whereas the first experiment used a habituation/dishabituation training paradigm and monitored heart-rate during testing, the second experiment used a training procedure that habituated infants to a melodic pattern and also trained them to respond with head turns to melodies that differed in melodic contour and range. Infants were then tested for subsequent discriminations of novel stimuli by monitoring head turns. Secondly, although training and testing stimuli also consisted of six note melodic patterns, several additional melodic properties were examined Trehub et al.'s testing condition. In addition to the transposed melody (as used in Chang and Trehub, 1977), the testing conditions included *contour preserving* and *contour violating* conditions in order to test for octave and frequency range sensitivity. The *contour preserving* condition (see Figure 3) allowed the researchers to test whether infants categorize melodies based on intervallic or contour properties. That is, it was

assumed that if infants recognized only the transposed condition and not the contour-preserving condition, then that would be taken as evidence that they processed the melodies based on intervals. On the other hand, if they recognized both, the most parsimonious explanation would be that they were processing the melodies based on contour alone. Additionally, the *octave change* conditions tested whether infants' were also sensitive to larger changes in intervallic patterns. Crucially, it was found that infants did not discriminate either the transposition or contour preserving melodies, but discriminated the *octave change* melodies.<sup>2</sup> In summary, Trehub et al. found that infants could categorize melodies by contour properties, but were also sensitive to the magnitude of the contour, and therefore could discriminate larger intervals (outside of the general original range) from the smaller intervals of melodies that occurred within the original melodic range.

## Simulation 2

Although there were methodological differences between Task 1 and 2, we simulated Trehub et al. (1984) using the same basic procedure as in Simulation 1. Fundamentally, Task 2 used the same approach by exposing infants to a standard melody and then subsequently observing how they would perceive and categorize novel test stimuli. Accordingly, we created a set of training patterns and trained the model as in Simulation 1.

To test the model, we created transposition melodies just as in Simulation 1. In addition, we created two kinds of contour preserving melodies. Close contour preserving melodies were similar to the training melodies, but with frequencies within 2 units of the training trials. So for example, if the first three notes of the training stimulus were: frequency2, frequency6, frequency8, the first three notes of the test pattern would be frequency2  $\pm$ 2, frequency6  $\pm$ 2, frequency8  $\pm$ 2 (under the constraint that the direction of the note was maintained across training and test patterns—e.g., if the second note of the training melody was higher than the first, the second note of the test melody was also higher than the first). Similarly, the far contour preserving melodies were created in exactly the same manner, but with each note  $\pm$ 6 from the original.

The results followed the same qualitative pattern observed in Trehub et al. (1984). As in the previous simulation, DORA successfully matched transposed melodies. Importantly, DORA also successfully matched close contour preserving melodies the majority of the time (74 of 100 trials), and was surprised on far contour preserving melodies more frequently (63 of 100 trials). In other words, like the infants in Trehub et al.'s study, DORA was sensitive to contour preservation, but under conditions when the contour was preserved but coupled with large changes in

frequency, DORA was more likely to categorize the melody as being different or produce a surprise reaction.

## Discussion

To our awareness, this is the first time a general model of relational cognition has been used to simulate melodic perception, and the results subsequently compared to existing behavioral data from infants.<sup>3</sup> We view these first steps as a very simple beginning, and hope to expand the complexity of the model and the range of future simulations.

The results of both simulations were a good match to their behavioral counterparts, and supported our hypothesis that relational processing might play an important role in music perception. In the first simulation, DORA performed similar to infants in extracting the relational properties of transposed melodies, and also in failing to recognize the scrambled melody. In the second simulation, both DORA and infants categorized the melodies based primarily on relational information of the melodic contours. Furthermore, DORA's ability to discriminate large contour distortions (far contour) in Simulation 2 suggests that infants may be sensitive to certain intervallic properties.

While these simulations provide insights into some of the mechanisms that infants may use when categorizing music, we hope to determine through future studies, when and how children begin to learn the more typically defining feature of melodies: the intervallic sequences, or relative pitch relations between notes. Crucially, this study corroborates existing evidence that infants as young as four months are sensitive to relational features of music and appear to reason about these relational features in a structure sensitive manner (i.e., generalizing relational properties to novel inputs). Another important question that future simulations and studies should attempt to answer is whether this widespread ability to discriminate intervallic sequences in adults is innate, or in fact a learned ability.

Lastly, DORA is currently the only model that learns complex structured relations and that can subsequently "grow up" to reason like an adult (Doumas & Hummel, 2005). Accordingly, we hope to determine through future simulations whether DORA can perhaps also grow up to "appreciate" (or even compose) music like an adult.

## References

- Attneave, F., & Olson, R. K. (1971). Pitch as a medium: A new approach to psychophysical scaling. *The American journal of psychology*, 84, 147-166.

<sup>2</sup> There was no evidence that infants processed octave shifts as musical pitches with closely related harmonic properties (as adults generally do), but rather that they only processed them as large shifts in frequency (see Trehub et al., 1984).

<sup>3</sup> We also simulated Task 1 using an Elman neural network with similar inputs to what was given to DORA. With a hidden layer of 24 neurons, the network learned the training melody sequence to a MSE of .01 within 15 iterations, however, after training it failed to systematically generalize to contour preserving melodies even when the intervallic pattern was consistent (performing at chance when predicting the contour). Moreover, this failure occurred even while the Elman network was provided a clean input set (unlike DORA, which was "handicapped" with distractor sets in memory and a highly reduced lateral inhibition parameter).



- Bartlett, J. C., & Dowling, W. J. (1980). Recognition of transposed melodies: A key-distance effect in developmental perspective. *Journal of Experimental Psychology: Human Perception and Performance*, 6(3), 501.
- Chang, H. W., & Trehub, S. E. (1977). Auditory processing of relational information by young infants. *Journal of Experimental Child Psychology*, 24(2), 324-331.
- Doumas, L. A. A., & Hummel, J. E. (2005). *A symbolic-connectionist model of relation discovery*. Paper presented at the 23rd Annual Conference of the Cognitive Science Society.
- Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, 115(1), 1.
- Doumas, L. A. A., Morrison, R. G., & Richland, L. E. (2009). The Development of Analogy: Working Memory in Relational Learning and Mapping.
- Doumas, L. A. A., Morrison, R. G., & Richland, L. E. (2010). *Differences in development of analogy across cultures: a computational account*. Paper presented at the 32nd Annual Conference of the Cognitive Science Society.
- Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, 85(4), 341.
- Dowling, W. J. (1984). Assimilation and tonal structure: Comment on Castellano, Bharucha, and Krumhansl. *Journal of Experimental Psychology*, 113(3), 417-420.
- Dowling, W. J. (1988). Tonal structure and children's early learning of music. In J. Sloboda (Ed.), *Generative Processes in Music*. Oxford: Oxford University Press.
- Fujioka, T., Trainor, L. J., Ross, B., Kakigi, R., & Pantev, C. (2004). Musical training enhances automatic encoding of melodic contour and interval structure. *Journal of Cognitive Neuroscience*, 16(6), 1010-1021.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2), 155-170.
- Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12(3), 306-355.
- Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.
- Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99(3), 480.
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3), 427.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110(2), 220.
- Kim, J. J., Pinker, S., Prince, A., & Prasada, S. (1991). Why no mere mortal has ever flown out to center field. *Cognitive science*, 15(2), 173-218.
- Lovett, M. C., & Anderson, J. R. (2005). Thinking as a production system. In K. J. Holyoak & R. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 401-429). New York: Cambridge University Press.
- Lynch, M. P., Eilers, R. E., Oller, D. K., & Urbano, R. C. (1990). Innateness, experience, and music perception. *Psychological Science*, 1(4), 272.
- Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100(2), 254.
- Morrison, R. G., Doumas, L. A. A., & Richland, L. E. (2006). *The development of analogical reasoning in children: A computational account*. Paper presented at the 28th Annual Conference of the Cognitive Science Society.
- Morrison, R. G., Doumas, L. A. A., & Richland, L. E. (2011). A computational account of children's analogical reasoning: balancing inhibitory control in working memory and relational representation. *Developmental Science*.
- Page, M. P. A. (1994). Modelling the perception of musical sequences with self-organizing neural networks. *Connection Science*, 6(2-3), 223-246.
- Sachs, C. (1943). *Rise of music in the ancient world east and west*. W.W. Norton.
- Schellenberg, E. G., & Trehub, S. E. (1996a). Children's discrimination of melodic intervals. *Developmental Psychology*, 32(6), 1039.
- Schellenberg, E. G., & Trehub, S. E. (1996b). Natural musical intervals: Evidence from infant listeners. *Psychological Science*, 7(5), 272.
- Thorpe, L. A., & Trehub, S. E. (1989). Duration illusion and auditory grouping in infancy. *Developmental Psychology*, 25(1), 122.
- Thorpe, L. A., Trehub, S. E., Morrongiello, B. A., & Bull, D. (1988). Perceptual grouping by infants and preschool children. *Developmental Psychology*, 24(4), 484.
- Trainor, L. J., McDonald, K. L., & Alain, C. (2002). Automatic and controlled processing of melodic contour and interval information measured by electrical brain activity. *Journal of Cognitive Neuroscience*, 14(3), 430.
- Trehub, S. E. (2001). Musical predispositions in infancy. *Annals of the New York Academy of Sciences*, 930(1), 1.
- Trehub, S. E., Bull, D., & Thorpe, L. A. (1984). Infants' perception of melodies: The role of melodic contour. *Child Development*, 55(3), 821-830.
- Trehub, S. E., Cohen, A. J., Thorpe, L. A., & Morrongiello, B. A. (1986). Development of the perception of musical relations: Semitone and diatonic structure. *Journal of Experimental Psychology: Human Perception and Performance*, 12(3), 295.
- Trehub, S. E., Trainor, L. J., & Unyk, A. M. (1993). Music and Speech Processing in the First Year of Life. *Advances in Child Development and Behavior* (Vol. 24, pp. 1-35). New York: Academic Press.



# The Persisting Benefits of Using Multiple-Choice Tests as Learning Events

**Jeri L. Little (jerilittle@wustl.edu)**

Department of Psychology, Washington University in St. Louis  
Campus Box 1125, One Brookings Drive, St. Louis, MO 63130-4899

**Elizabeth Ligon Bjork (elbjork@psych.ucla.edu)**

Department of Psychology, University of California, Los Angeles  
1285 Franz Hall, Box 951563, Los Angeles, CA 90095-1563

## Abstract

Taking a test tends to improve the retention of the tested information. Additionally, taking a test often influences the later retention of non-tested information, provided such information is related to the tested information in a specific manner. To illustrate, recent research has demonstrated that multiple-choice tests containing competitive alternatives can improve retention of both tested and non-tested information pertaining to such incorrect alternatives at least over a short delay. The present research investigated whether such improvements in retention would persist with a delay more likely to occur in educational contexts (i.e., 48 hr). Taking an initial multiple-choice test improved retention more than a comparable cued-recall test—for both previously tested and related information—and over both short and long delays. Moreover, misinformation effects seen for the multiple-choice test at the short delay were reduced. These results thus have important implications for the use of multiple-choice tests as learning opportunities.

**Keywords:** Testing; test effects; multiple-choice; retrieval-induced forgetting; retrieval-induced facilitation

## Introduction

Taking a test does more than assess one's knowledge: It can also improve one's long-term retention of the tested information. Not all tests, however, are equally beneficial in this manner. For example, more open-ended tests (e.g., cued-recall), in general, have been shown to improve long-term retention more than multiple-choice tests (see meta-analyses by Anderson & Biddle, 1975; Hamaker, 1986). Moreover, some research has shown that taking multiple-choice tests can actually harm later performance on open-ended tests by exposing the test takers to misinformation in the form of incorrect, but plausible, answer choices, with the consequence of such information sometimes being intruded as incorrect responses to later cued-recall questions—findings referred to as misinformation effects (e.g., Roediger & Marsh, 2005). Thus, multiple-choice tests are often accused of not only being less effective for learning than are more open-ended types of tests, but also to risk negative misinformation effects, bringing their use as learning tools in educational contexts into question.

Although, as indicated, multiple-choice tests can produce misinformation effects on later open-ended tests, research, in general, has shown that the positive effects of multiple-choice testing (as compared to no testing) outweigh any such negative effects they engender. Furthermore, recent

research has indicated that multiple-choice testing may have a benefit for later overall performance that is not produced by the taking of prior cued-recall tests: namely, not only improving the retention of the information tested but also that of non-tested, but related, information—an outcome that would seem to be particularly desirable in educational contexts. When, for example, instructors give quizzes or practice tests to be followed later in the course by a more comprehensive exam, they are likely to ask questions about related information instead of (or in addition to) questions about the specific information tested earlier. Thus, it seems critical that the use of prior testing in educational contexts should help to improve the retention of both types of information.

With respect to the question of how retrieving some information affects the later retention of non-tested related information, previous research has demonstrated that the effects can be negative—that is, answering a cued-recall question can lead to impaired recall of competitive related information on a later test, a finding referred to as retrieval-induced forgetting (Anderson, Bjork, & Bjork, 1994). Even in cases in which retrieval-induced forgetting does not occur, however, the retention of non-tested related information is rarely facilitated as a consequence of a cued-recall test, especially when such non-tested information has a competitive relationship with the tested information. (For exceptions to this finding when tested and non-tested related questions were created to be facilitative, see Chan, McDermott, & Roediger, 2006, and Chan, 2009.)

In contrast to such findings for cued-recall tests, however, Little, Bjork, Bjork, and Angello (in press) recently demonstrated that—in addition to improving retention of previously tested information—taking an initial multiple-choice test, as compared to taking an initial cued-recall test, can improve retention of related information when the answers to the competitive related questions have occurred as incorrect alternatives in the initial multiple-choice test. Although these findings have clear implications for educational practice, particularly how an instructor might construct a practice exam, the procedure used by Little et al. employed a delay that is considered to lack educational realism (e.g., 5 min). Thus, a primary goal of the present research was to assess whether the demonstrated retention benefits (particularly for competitive related information) occurring as a consequence of taking multiple-choice

tests—compared to cued-recall tests—would persist at more educationally realistic delays.

### Retention of Non-tested Related Information

Although the effect of testing on the retention of both tested and related information was thoroughly explored in the 1960s and 70s (e.g., see literature on the use of adjunct questions, Anderson & Biddle, 1975; Frase, 1971; Hamaker, 1986; Rothkopf, 1970), interest in this topic has recently been renewed, not only because of the implications for educational practice, but also, in light of the finding of retrieval-induced forgetting (Anderson et al., 1994), to gain a better understanding of the circumstances under which an initial test might improve or impair retention of related information (e.g., Carroll, Campbell-Ratcliffe, Murnane, & Perfect, 2007; Chan et al., 2006; Chan, 2009; Little, Storm, & Bjork, 2011).

The research by Little et al. (in press), in which they directly compared the effects on later recall of taking an initial cued-recall test versus a multiple-choice test following study of a passage, provides some insight into this issue. Because retrieval-induced forgetting seems to depend upon the occurrence of a competitive relationship between tested and non-tested related information (see, e.g., Storm, 2010), Little et al. specifically constructed related question pairs so that one question in the pair would have a competitive relationship with the other question in the pair. For example, given a passage about Saturn, one question would ask the participant “how long it takes Saturn to revolve around the Sun” to which the correct answer is “30 Earth years,” and the other question would ask “how long it takes Saturn to rotate on its axis” to which the correct answer is “10 Earth hours.” On the initial multiple-choice tests, the answer to the related question appeared as one of the incorrect alternatives (e.g., *How long does it take Saturn to rotate on its axis? a. 10 Earth hours, b. 88 Earth days, c. 176 Earth days, d. 30 Earth years*).

Little et al. hypothesized that the presence of competitive alternatives as answer choices on the multiple-choice question would induce students to think deeply about the question and alternatives—perhaps not only recalling information about why the correct alternative is right, but also why the incorrect alternatives are wrong—with this spontaneous recall of information pertaining to incorrect alternatives serving as a learning event for that information and potentially improving its future recall. In contrast, they hypothesized that a cued-recall test would not afford the opportunity for such beneficial processing of related competitive information; thereby, perhaps making it subject to retrieval-induced forgetting. And, indeed, their findings supported this type of analysis: Whereas the taking of an initial cued-recall test following study of the passage led to impaired later recall of non-tested competitive information (Exp. 1), the taking of a multiple-choice test after reading the passage not only spared such information from forgetting (Exp. 1), but facilitated its retention (Exp. 2).

Although one might wonder if the facilitated retention of related information observed by Little et al. (in press) might have occurred owing simply to the exposure of those answers as incorrect alternatives during the initial multiple-choice test, a study by Little and Bjork (2010) provides evidence that the benefit for the retention of related information would not have occurred for this reason. Little and Bjork manipulated whether the incorrect alternatives were competitive or non-competitive for a given question, with the expectation that alternatives would need to be competitive for facilitation to occur. To illustrate, students answered questions (e.g., *Which outer planet was discovered by mathematical predictions rather than direct observation? Answer: Neptune*) with competitive alternatives (e.g., *Neptune, Uranus, Saturn*) or with non-competitive alternatives (e.g., *Neptune, Mercury, Mars*), and they were better able to answer a later related question (for which *Uranus* or *Mercury*, respectively, were the correct answers) when the answer to that related question had served as a competitive alternative versus as a non-competitive alternative on the initial multiple-choice test. Thus, simple prior exposure to the incorrect alternatives cannot explain the observed results: Competitive alternatives seem to trigger the retrieval processes that support the improved retention of information pertaining to those incorrect alternatives.

The results of the experiment by Little and Bjork (2010), although theoretically informative, do not address the question of concern in the present research: namely, whether such effects persist over educationally realistic durations. On the one hand, the pattern of their results suggests that the presence of competitive alternatives on the initial multiple-choice test induces learners to engage in both deep processing of the choices and spontaneous recall of information pertaining to those choices, both of which might promote long-term retention. On the other hand, the pattern of results observed by Little et al. (in press, Exp. 1) might not persist. Their observation of a benefit of multiple-choice testing over cued-recall testing might reflect, in part, a temporary impairment in the recall of information related to the cued-recall questions, as other research has suggested that such retrieval-induced forgetting does not persist (Chan, 2009; MacLeod & Macrae, 2001). Chan, for example, found that although forgetting occurred as a consequence of an initial cued-recall test when the retention interval was short (i.e., 20 min), forgetting was not observed when the interval was long (i.e., 24 hrs). Furthermore, when forgetting did not occur at a short delay, facilitation emerged at a longer delay. Thus, it is possible that Little et al.’s observed pattern of a benefit of multiple-choice testing over cued-recall testing would be eliminated when retention is assessed at a delay longer than 5-20 minutes.

### Retention of Previously Tested Information

In addition to assessing the effect of initial testing on the later recall of related information, we were also interested in its effect on previously tested information. Specifically, we

also wanted to assess whether the benefit of multiple-choice testing compared to cued-recall testing observed by Little et al. (Exp. 1) would persist over a long delay. If, as has been suggested (e.g., Foos & Fischer, 1988), cued-recall tests involve deeper processing than do multiple-choice tests, then such an outcome would seem unlikely; that is, improved retention of the tested information would seem more likely to persist as a consequence of a cued-recall test than as a consequence of a multiple-choice test (see, e.g., Roediger & Karpicke, 2006), which might lead to a reversal in the direction of the effect.

### **Misinformation: A Negative Consequence of Multiple-choice Testing**

Finally, multiple-choice tests can result in misinformation effects (Roediger & Marsh, 2005) and seem particularly prone to do so when feedback is not provided (Butler & Roediger, 2008). Thus, another goal of the present research was to assess the influence of a longer retention interval on the prevalence of misinformation effects.

### **The Present Experiment**

In summary, Little et al. (in press) claimed that multiple-choice tests should be exonerated—at least from the criticisms involving their use in learning, but they used a retention interval that has limited application for educational purposes. In the present experiment, we assessed whether their observed benefits of taking a multiple-choice test over taking a cued-recall test or no test at all would persist over a longer retention interval. Additionally, we assessed the occurrence of misinformation at these two retention intervals.

## **Method**

### **Participants and Design**

Seventy-two undergraduates at the University of California, Los Angeles, who were all recruited for a two-session study, participated for either partial course credit or payment. Each participant read three passages, followed immediately by two initial tests: a cued-recall test for one passage and a multiple-choice test for another passage (the remaining passage was not tested initially and served as the control condition). After a delay (either 5-min or 48-hr), all participants took a final test that included previously tested questions and previously non-tested related questions (pertaining to multiple-choice questions for one passage and cued-recall questions for another passage), and non-tested questions from the non-tested passage.

### **Materials**

The same materials as used by Little and Bjork (2011, Exp. 3) were employed: Three passages of approximately 800 words each about Saturn, Yellowstone National Park, and stimulant drugs with ten pairs of multiple-choice questions for each passage. The two multiple-choice questions in each pair were semantically related in that both tested the

same topic (e.g., geysers) and had the same four alternatives (e.g., *Old Faithful*, *Steamboat Geyser*, *Castle Geyser*, and *Daisy Geyser*), but different correct answers (e.g., *What is the tallest geyser in Yellowstone National Park?* Answer: *Steamboat Geyser*; and, *What is the oldest geyser in Yellowstone National Park?* Answer: *Castle Geyser*). For each passage, the questions in each pair were randomly assigned to a different 10-item set. Cued-recall questions were the same as the multiple-choice ones, but without alternatives provided.

### **Procedure**

All participants were given 6 min to read each of the three passages in immediate succession, being told to spend the full time reading and studying. After this initial study phase, they were given a 10-item multiple-choice test for one passage and a 10-item cued-recall test for another passage (with each test containing all the items in one of the question sets for that passage). Questions were shown on the computer screen one at a time for 24 s, and participants typed in their responses, being instructed to spend the full time thinking about the question and their answer. Each question was answered twice (as is common in studies of retrieval-induced forgetting) such that participants answered each of the ten questions once, and then answered each question again, but in a different order.

After taking the initial multiple-choice and cued-recall tests, all participants engaged in a non-verbal 5-min distractor task of maze completion. Then, participants randomly assigned to the 5-min delay condition immediately took the final test; those randomly assigned to the 48-hr delay condition were excused, with the instruction to return 48 hrs later.

On the final test, participants answered 60 cued-recall questions, with the questions presented one at a time on the computer screen. Of the 20 final-test questions associated with the passage that received an initial cued-recall test, half were identical to the 10 cued-recall questions initially asked about that tested passage, while the other half consisted of the related questions that had not been presented on the initial cued-recall test (i.e., the remaining questions from the set of 10-paired questions constructed for that passage). Of the 20 final-test questions associated with the passage that received an initial multiple-choice test, half were identical to the 10 multiple-choice questions initially asked about the tested passage (except they were now asked as cued-recall questions without alternatives presented), while the other half consisted of the related questions that had not been presented on the initial multiple-choice test (i.e., the remaining questions from the set of 10-paired questions constructed for that passage). Of the 20 final-test questions related to the non-tested passage, all were previously non-tested and served as baseline control items.

As the later recall of non-tested related information was most crucial to our research questions, the non-tested questions from the tested passages were tested in the first half of the test, along with half of the non-tested control

questions, to which their performance would be compared. Previously tested questions were tested in the second half of the test, along with the remaining non-tested control questions, to which their performance would be compared. The order in which the passages were studied, the passages assigned to the different testing conditions, the order in which the tests were administered, and which 10-item question set was presented during the initial tests were counterbalanced across participants.

## Results

### Initial Test Performance

Participants in the 5-min and 48-hr delay conditions correctly answered 63% ( $SD = 18\%$ ) and 66% ( $SD = 19\%$ ) of the questions on the initial multiple-choice test, respectively. Participants in the 5-min and 48-hr delay conditions correctly answered 40% ( $SD = 21\%$ ) and 39% ( $SD = 18\%$ ) of the questions on the initial cued-recall test, respectively. As would be expected, initial test performance did not differ between the two retention-interval conditions.

### Final Test Performance for Previously Tested Information

Final-test performance for previously tested information and comparable control information is presented in Figure 1. As indicated there, taking an initial test improved performance for those previously tested items on the later test, as compared to not taking an initial test, regardless of the type of initial test taken (cued-recall or multiple-choice) or the delay from initial test to the final test (5 min or 48 hr).

Interestingly, although performance appeared to be lower at a 48-hr delay than at a 5-min delay for questions initially presented in a multiple-choice test and for control questions (i.e., questions not previously tested), no forgetting appeared to occur for those items initially presented in a cued-recall test, suggesting a possible interaction. A  $3 \times 2$  mixed-model

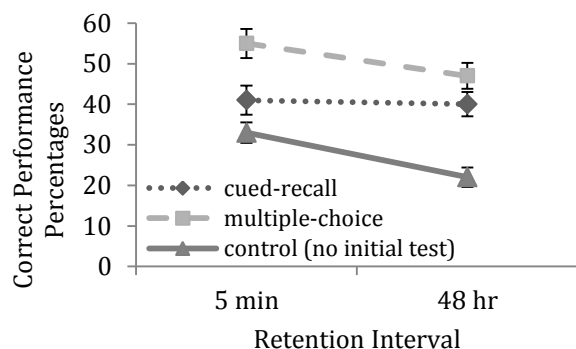


Figure 1: Correct performance percentages for previously tested information (and comparable non-tested control information) as a consequence of the testing condition (cued-recall, multiple-choice, or no-test control) and the retention interval between initial and final test (5 min or 48 hr). Error bars represent  $\pm 1$  SE.

ANOVA, however, did not reveal this apparent interaction between testing condition (cued-recall vs. multiple-choice vs. control) and delay (5 min vs. 48 hr), to be reliable,  $F(2, 69) = 1.97, p > .05$ . A significant main effect of testing condition, however, was obtained,  $F(2, 69) = 43.73, \eta^2 = 0.56, p < .01$ . Most importantly, taking a multiple-choice test was better for retention of previously tested information than was taking a cued-recall test,  $t(71) = 4.32, d = 0.51, p < .01$ .

That performance for questions previously tested with a cued-recall test did not appear to differ when assessed at a 5-min delay versus a 48-hr delay is noteworthy, indicating the effectiveness of cued-recall tests for the retention of tested information and consistent with findings that items successfully retrieved at short delays tend to remain accessible at longer delays (e.g., Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011).

### Final Test Performance for Non-tested Related Information

Results relevant to our primary objective of testing whether the finding that taking a multiple-choice test improves retention of non-tested information more than does taking a cued-recall test with a short retention interval between initial and final tests would replicate, and, if so, to assess whether such benefits would persist with a longer retention interval are shown in Figure 2. As can be seen there, information that was related to questions on an initial multiple-choice test appeared to be better recalled than was information related to questions on an initial cued-recall test or information from the non-tested control passage, both at a 5-min delay and at a 48-hr delay.

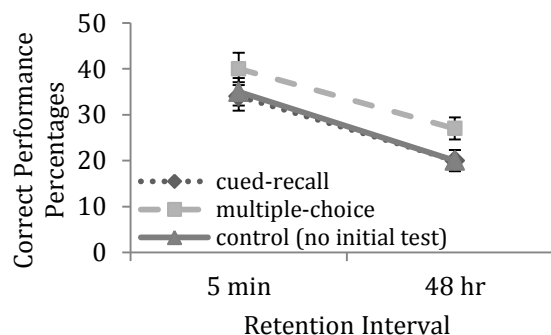


Figure 2: Correct performance percentages for non-tested related information (and comparable non-tested control information) as a consequence of the testing condition (cued-recall, multiple-choice, or no-test control) and the retention interval between the initial and final tests (5 min or 48 hours). Error bars represent  $\pm 1$  SE.

Again, a  $3 \times 2$  mixed-model ANOVA did not reveal an interaction between testing condition and delay,  $F(2, 69) = 0.12, p > .05$ , but did reveal a main effect of testing condition,  $F(2, 69) = 5.04, \eta^2 = 0.13, p < .01$ . Overall, taking a multiple-choice test was better for the retention of

related information than was either taking a cued-recall test or not taking a test (control condition),  $t(71) = 2.72$ ,  $d = 0.33$ ,  $p < .01$  and  $t(71) = 2.82$ ,  $d = 0.33$ ,  $p < .01$ , respectively. Importantly, however, no difference in performance for the cued-recall condition versus the control condition was observed, indicating that although taking a cued-recall test did not impair the later recall of related information in the present study, neither did it help it compared to not taking a test at all. Taking a multiple-choice test did, however, and this benefit occurred regardless of the retention interval.

### **Intrusions of Incorrect Information**

As previously mentioned, multiple-choice tests can result in misinformation effects (Roediger & Marsh, 2005; Marsh, Roediger, Bjork, & Bjork, 2007) thought to be a drawback to multiple choice testing as compared to more open-ended types of testing, and, indeed, participants in the present study made such intrusions on the final test in the multiple-choice condition. Because, however, all of the alternatives had appeared in the passages, participants also made such intrusions in the control and cued-recall conditions.

In the present study, there appeared to be more intrusions as a consequence of taking a multiple-choice test ( $M = 24\%$ ,  $SE = 3\%$ ) than as a consequence of taking a cued-recall test ( $M = 12\%$ ,  $SE = 2\%$ ) or having no initial test ( $M = 14\%$ ,  $SE = 2\%$ ) when the final test occurred at a 5-min delay. At the 48-hr delay, however, intrusions as a consequence of taking a multiple-choice test ( $M = 17\%$ ,  $SE = 2\%$ ) appeared to be reduced while intrusions occurring as a consequence of taking a cued-recall test ( $M = 13\%$ ,  $SE = 2\%$ ) or having no test ( $M = 14\%$ ,  $SE = 2\%$ ) did not. Looking specifically at the interaction between the two test types and the delay, a  $2 \times 2$  mixed-model ANOVA revealed an interaction,  $F(1, 70) = 4.24$ ,  $p < .05$ , with an independent samples  $t$ -test demonstrating that multiple-choice alternative intrusions were significantly lower at the 48-hr delay than at the 5-min delay,  $t(70) = 2.14$ ,  $d = 0.50$ ,  $p < .05$ . In considering these results, one should note that while correct performance in the control condition was lower at the 48-hr delay than at the 5-min delay, intrusion rates at the two delays were not different, suggesting that forgetting rates of correct and incorrect responses are not necessarily related. For the multiple-choice condition, however, both correct responses and intrusions were reduced at the 48-hr delay compared to the 5-min delay. Interestingly, in the multiple-choice condition, correct responses and intrusions were negatively correlated,  $r(72) = -.30$ ,  $p < .05$ , suggesting that the more correct answers one recalled, the fewer intrusions one made, in both delay conditions.

### **Discussion**

Taking a multiple-choice test shortly after study appears to have some particularly positive consequences for learning. Doing so not only increases retention of both tested and non-tested related information compared to taking a cued-recall test or no test, but such benefits also persist over a

retention interval that is, by most standards, considered educationally realistic. In the present study, we also found that even misinformation effects, a frequently cited negative consequence of taking a multiple-choice test, become less of a problem at a longer retention interval. Thus, at a longer delay from initial to final test, multiple-choice testing still had the benefits seen at a shorter delay, but with less cost.

The present results can be contrasted with much past research demonstrating that cued-recall tests are better for retention than multiple-choice tests (e.g., Hamaker, 1986). Our finding of a greater benefit for multiple-choice testing than cued-recall testing might have occurred owing to the relatively low performance for cued-recall questions on our initial test: Although participants were unlikely to forget answers to questions that they answered correctly on the initial cued-recall test, the majority of responses on the initial test were incorrect and thus unlikely to be answered correctly later. In contrast, more of the multiple-choice questions were answered correctly on the initial test. Thus, although answering a given multiple-choice question might be less powerful for long-term retention than answering a given cued-recall question (Foos & Fisher, 1988), the greater number of correct responses on the initial multiple-choice test would result in more items having the potential to be recalled correctly later. It should be noted, however, that our use of highly competitive alternatives might also have resulted in our multiple-choice questions being more powerful for retention than those typically used in past research. Even so, perhaps providing feedback, a common practice in educational contexts, would lead to a reversal in our observed effect because feedback provides a second opportunity for the correct answers to incorrectly answered cued-recall questions to be learned (e.g., Kang, McDermott, & Roediger, 2007). Importantly, however, Little et al. (in press) also demonstrated that—although providing feedback did improve performance for information initially tested with a cued-recall test more than for information initially tested with a multiple-choice test, a reversal in final test performance was not realized. Furthermore, performance for related information was not affected as a consequence of whether or not feedback was provided.

Additionally, in the present study, we did not find evidence of retrieval-induced forgetting in the cued-recall condition, which is interesting, particularly at the short delay and given the competitive nature of our question pairs. Although retrieval-induced forgetting has been demonstrated with a variety of materials (see, e.g., Bjork, Bjork, & MacLeod, 2006; Storm, 2011), its occurrence with educational text materials is less consistent, with some studies observing retrieval-induced forgetting (e.g., Carroll et al., 2007; Little et al., 2011) and others not (e.g., Chan et al., 2006). Competition during the initial test, coherence of the to-be-learned materials, and the delay between initial and final tests have been pointed to as predictors of whether or not the effect will occur, but the exact specifications of the boundary conditions for the occurrence, or lack thereof, of retrieval-induced forgetting remain to be determined.

## Concluding Comments

Multiple-choice tests are widely used in educational contexts and elsewhere, but their use—either as instruments of assessment or learning—is frequently disparaged. Although the present research does not directly speak to their potential value as tools of assessment, it does speak to their use as tools for learning. And, in that respect, as tools to reinforce knowledge, for the long as well as the short term, they seem quite useful when constructed with competitive alternatives and, perhaps, uniquely so with respect to the increased learning of competitive related as well as actually tested information.

## Acknowledgments

A collaborative research grant from the James S. McDonnell Foundation supported this research.

## References

- Anderson, M. C., Bjork, R. A., Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 1063-1087.
- Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. H. Bower (Ed). *The Psychology of Learning and Motivation* (Vol.9). New York, NY, US: Academic Press.
- Bjork, E. L., Bjork, R. A., & MacLeod, M. D. (2006). Types and consequences of forgetting: Intended and unintended. In L-G Nilsson and N. Ohta (Eds.), *Memory and Society: Psychological Perspectives*. Psychology Press: Hove and New York.
- Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, 36, 604-616.
- Carroll, M., Campbell-Ratcliffe, J., Murnane, H., & Perfect T. (2007). Retrieval-induced forgetting in educational contexts: Monitoring, expertise, text integration, and test format. *European Journal of Cognitive Psychology*, 19, 580-606.
- Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, 61, 153-170.
- Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, 135, 553-571.
- Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology*, 80, 179-183.
- Frase, L. T. (1971). Effect of incentive variables and type of adjunct questions upon text learning. *Journal of Educational Psychology*, 62, 371-375.
- Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 801-812.
- Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research*, 56, 212-242.
- Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, 19, 528-558.
- Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, 65, 85-97.
- Little, J. L., & Bjork, E. L. (2010). Multiple-choice testing can improve the retention of non-tested related information. In S. Ohisson & R. Catrbone (Eds.) *Proceedings of the 32<sup>nd</sup> Annual Conference of the Cognitive Science Society* (pp. 1535-1540). Austin, TX: Cognitive Science Society.
- Little, J. L., & Bjork, E. L. (2011). Pretesting with multiple-choice questions facilitates learning. In L. Carlson, C. Hölscher & T. F. Shipley. (Eds.) *Proceedings of the 33<sup>rd</sup> Annual Conference of the Cognitive Science Society* (pp. 294-296). Austin, TX: Cognitive Science Society.
- Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (in press). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*.
- Little, J. L., Storm, B. C., Bjork, E. L. (2011). The Costs and Benefits of Testing Text Materials. *Memory*, 19, 346-359.
- Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, 14, 194-199.
- MacLeod, M. D., & Macrae, C. N. (2001). Gone but not forgotten: The transient nature of retrieval induced forgetting. *Psychological Science*, 18, 29-34.
- Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181-210.
- Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 1155-1159.
- Rothkopf, E. Z. (1970). The concept of mathemagenic activities. *Review of Educational Research*, 40, 325-336.
- Storm, B.C. (2011). Retrieval-induced forgetting and the resolution of competition. In A. S. Benjamin (Ed.), *Successful Remembering and Successful Forgetting: A Festschrift in honor of Robert A. Bjork*. New York, NY: Psychology Press.

# The perception of simplified and traditional Chinese characters in the eye of simplified and traditional Chinese readers

Tianyin Liu ([kanalty@hku.hk](mailto:kanalty@hku.hk))  
Janet Hui-wen Hsiao ([jhsiao@hku.hk](mailto:jhsiao@hku.hk))

Department of Psychology, University of Hong Kong  
604 Knowles Building, Pokfulam Road, Hong Kong SAR

## Abstract

Expertise in Chinese character recognition is marked by analytic/reduced holistic processing (Hsiao & Cottrell, 2009), which depends mainly on readers' writing rather than reading experience (Tso, Au, & Hsiao, 2011). Here we examined whether simplified and traditional Chinese readers process characters differently in terms of holistic processing. When processing characters that are distinctive in the simplified and traditional scripts, we found that simplified Chinese readers were more analytic than traditional Chinese readers in perceiving simplified characters; this effect depended on their writing rather than reading/copying performance. In contrast, the two groups did not differ in holistic processing of traditional characters, regardless of their performance difference in writing/reading traditional characters. When processing characters that are shared in the two scripts, simplified Chinese readers were also more analytic than traditional Chinese readers. These results suggest that simplified Chinese readers may have developed better analytic processing skills than traditional Chinese readers from experiences with simplified characters, and these skills are transferrable to the processing of shared and even traditional characters.

**Keywords:** Chinese character recognition, holistic processing, reading, writing

## Introduction

Chinese characters are the basic writing units in Chinese orthography. They consist of strokes packed into a square configuration, in contrast to words in most alphabetic languages, which are linear combinations of letters. It has been suggested that Chinese characters have many visual similarities with faces (McCleery et al., 2008). For example, both faces and Chinese characters have a homogenous shape, are recognized at the individual level, and learnt in an upright orientation. However, expertise in face recognition is marked by holistic processing, that is, to perceive features of an object as a whole instead of various parts (Gauthier & Tanaka, 2002); in contrast, expertise in recognizing Chinese characters is marked by reduced holistic processing (Hsiao & Cottrell, 2009). This effect may be due to expert Chinese readers' knowledge about Chinese orthography. Chinese characters are composed of strokes, which combine to form over a thousand different stroke patterns in Chinese orthography (Hsiao & Shillcock, 2006), and stroke patterns are the smallest functional units in Chinese character recognition (Chen, Allport, & Marshall, 1996). For expert Chinese readers, when recognizing Chinese characters, they may be more sensitive to the internal constituent

components and have better ability to ignore some unimportant configural information for recognition, such as exact distances between features (Ge, Wang, McCleery, & Lee, 2006) compared with novices (Ho, Ng, & Ng., 2003; Hsiao & Cottrell, 2009). Consequently, expert readers may process Chinese characters less holistically than novices.

There are currently two Chinese writing systems in use in Chinese speaking regions, namely simplified and traditional Chinese. The simplified script was created during the writing reform initiated by the central government of the People's Republic of China in the 1960s for easing the learning process. Today the majority of Chinese speaking regions including Mainland China, Singapore, and Malaysia use the simplified script, while Hong Kong and Taiwan continue to use the traditional script. The simplification process did not apply to all characters; among the most frequently used 3,500 characters, around 40% were simplified, which have approximately 22.5% fewer strokes than the traditional counterparts (Gao & Kao, 2002); the remaining 60% remained the same in two scripts.

The effects of simplifying the script have aroused some discussion since last decade. For instance, while simplified characters were designed to ease the learning process, many researchers (e.g., Chen, 1999) believe that pure reduction of strokes without standardization of principles may make the characters harder to learn: on one hand, reducing the stroke numbers may make the characters more legible and easier to write for beginners; on the other hand, up to a certain point, characters may become less distinguishable due to high visual similarity among one another as readers' lexicon size expands (Chen, 1999). Consistent with this speculation, McBride-Chang et al. (2005) recently found that visual skills of children who learned simplified characters were significantly better than those of Hong Kong children who learned the traditional script. Peng, Winett, and Wang's (2010) ERP data were also consistent with the finding that simplified character readers have greater visual discrimination skills than traditional character readers in perceiving Chinese characters.

Thus, it is possible that the simplification has significantly increased visual similarity among characters in the Chinese lexicon. Simplified characters may differ from one another in fewer strokes than traditional characters. As the ability to identify these diagnostic features is important for recognition (e.g., Oliva & Schyns, 1997), reading simplified characters may involve more analytic/reduced holistic processing than reading traditional characters. Here we aim



to examine whether native traditional and simplified Chinese readers process Chinese characters differently due to the differences in their scripts. We first examine their perception of characters in either the simplified or the traditional form; we predict that simplified Chinese readers will perceive simplified characters less holistically than traditional readers due to their expertise with the simplified script, and vice versa in the perception of traditional characters. In addition to the characters that have different visual forms in the two scripts, around 60% of the most frequently used characters have the same form in the two scripts, and these shared characters provide us a unique opportunity to test the transfer effect of reading simplified/traditional characters. Because both reader groups are experts in reading shared characters, if the two groups differ in the way they perceive/process the shared characters, it will suggest a transfer effect from their experience with the simplified or traditional scripts.

Tso, Hsiao, and Au (2011) recently examined how writing experience influences holistic processing in Chinese character recognition. They recruited proficient Chinese readers who were skilled in both reading and writing (Writers), and those who had limited writing experience regardless of their proficient reading ability (Limited-Writers). They found that Writers perceived Chinese characters less holistically than Limited-Writers, and holistic processing effect was dependent on writing rather than reading performance. Although simplified Chinese readers may still be able to read traditional characters through their similarity with simplified characters or context information, they generally do not know how to write them (and vice versa for traditional Chinese readers). Thus, similar to Limited-Writers, they may perceive characters in their unfamiliar script more holistically, and this effect may depend on their writing rather than reading performance. To verify this hypothesis, we also measure participants' reading and writing performance in the two scripts and examine whether their (reduced) holistic processing can be predicted by these measures. This study is also the first to investigate holistic processing effects in the two Chinese scripts across two groups of readers.

## Methods

### Participants

24 native simplified Chinese readers (5 males and 19 females) from Mainland China and 24 native traditional Chinese readers (9 males and 15 females) from Hong Kong participated in the study. They were all skilled writers in their own script: all Mainland participants had passed the Chinese test of National Entrance Examination to college, and all Hong Kong participants had passed the Chinese test of Hong Kong Advanced Level Examination. They were all students at University of Hong Kong; all simplified Chinese readers had resided in Hong Kong for less than 1 years (average length of stay was 9.35 months) by the time they were recruited and had limited exposure to traditional

Chinese before<sup>1</sup>. Both groups had similar education background (average years of education, Mainland = 15.46,  $SE = .37$ ; Hong Kong = 15.38,  $SE = .44$ ) and similar age (Mainland average = 22.25,  $SE = .65$ ; Hong Kong average = 22.42,  $SE = .81$ ). All of them had normal or corrected-to-normal vision and were right-handed as measured by the Edinburgh Handedness Inventory (Oldfield, 1971).

### Holistic processing

The complete composite paradigm was used to examine holistic processing effects (Gauthier & Bukach, 2007). The experiment procedure was adopted from Hsiao and Cottrell (2009; Fig. 1). In each trial, participants were presented with a pair of Chinese characters simultaneously, and told to attend to only half of each character and judge whether they were the same or different. In congruent trials, the attended and irrelevant halves of the characters led to the same response (i.e., both were the same or different); in incongruent trials, they led to different responses. The level of holistic processing was assessed by the performance difference between the congruent and incongruent trials.

**Materials** The materials consisted of 480 pairs of low to medium frequency (Kwan, 2001) Chinese characters in Ming font, divided equally into three script types: 160 pairs were simplified characters; 160 pairs were the corresponding traditional version of the simplified characters, i.e., having same meanings and pronunciations and differing in orthography; the remaining 160 pairs were characters shared between the two scripts (i.e., shared characters). Characters of different script types were matched in relative character frequency, and the traditional characters were significantly more complex than the simplified ones ( $t(159) = 6.17, p < .01$ ). In each script type, half of the characters had a top-bottom (TB) configuration, and the other half were left-right (LR) structured<sup>2</sup>, and two groups were matched in complexity and frequency. The 80 character pairs in each script type and character configuration combination were further divided into the four conditions in the complete composite paradigm, with 20 pairs in each condition shown in Fig. 1a. Each character could be divided into two components, horizontally for TB and vertically for LR characters. In either character configuration condition, the attended halves were matched across congruent and incongruent trials (see Fig. 1a for an illustration), and character frequency and visual complexity were also matched across congruent and incongruent trials.

**Design** The design had three within-subject variables: congruency (congruent vs. incongruent), character

<sup>1</sup> Note that the official written languages used in Hong Kong are English and traditional Chinese, and the official language for instruction at University of Hong Kong is English.

<sup>2</sup> We used both TB and LR characters to counterbalance possible influence from character structure; the LR structure is the most dominant structure in Chinese orthography, followed by the TB structure (see, e.g., Hsiao & Shillcock, 2006).

configuration (TB vs. LR), and script type (simplified vs. traditional vs. shared); and a between-subject variable: group (simplified vs. traditional Chinese readers). The dependent variable was discrimination sensitivity measured by  $A'$ , a bias-free nonparametric measure of sensitivity<sup>3</sup>.

**Procedure** All characters were shown in low contrast to avoid ceiling effects. In each trial, participants were presented with a central fixation cross for 1000 ms, followed by a symbol indicating which half of the character (top or bottom for TB characters; left or right for LR characters) they should attend to. They were then presented with a pair of characters above and below the initial fixation respectively for 500 ms, followed by a mask (Fig. 1b). Both characters were about 2.5° of visual angle away from the center, each occupying around 1.5° of visual angle. Participants performed a same-or-different judgment task as quickly and accurately as possible with a response box; their accuracy was collected. There were six blocks of test; each block had 80 trials; characters with different configurations or in different script types were presented in different blocks. The sequence of blocks was counterbalanced across participants. A practice session with characters not used in the materials was given before the test.

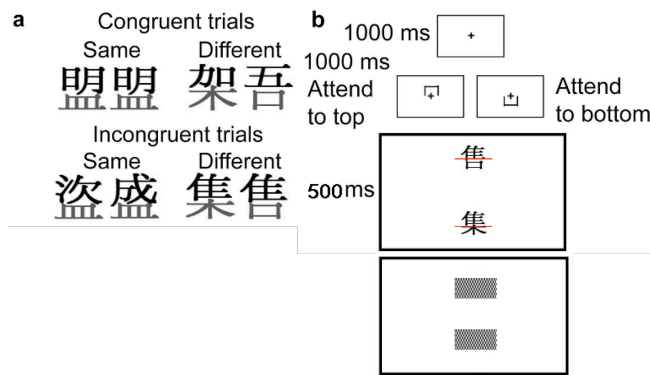


Fig. 1: Illustration of stimulus pairs in the complete composite paradigm (a) and trial sequences (b). In (a), the attended components are the bottom halves shaded in grey.

### Reading & Writing Performance

Tasks assessing participants' reading and writing abilities were adopted from Tso et al. (2011). Participants' reading ability was assessed by a character naming task, in which they named the characters in their mother tongue (i.e., Putonghua for Mainland China participants and Cantonese for Hong Kong participants). Their writing ability was assessed by a character copying and a word dictation task.

**Character Naming** The materials consisted of 120 Chinese characters, half with a TB structure and the other half with a

LR configuration. In either configuration, 20 characters were simplified characters, 20 were the corresponding traditional version of the simplified characters, and the remaining 20 were shared characters; these characters were not used in the holistic processing task. All characters were of medium to high frequency (Kwan, 2001); they were matched in relative frequency across the script types. The traditional characters had significantly more strokes than the simplified ones ( $t(39) = 10.92, p < .01$ ). In each trial, after a 500 ms central fixation, participants were presented with a character occupying 1.5° of visual angle at the center of the screen, and they were asked to read it out in front of a microphone. Upon their response, the screen turned blank and the experimenter pressed buttons on a response box to record the accuracy and initiate the next trial.

**Character Copying** Participants copied 60 characters (20 shared, 20 simplified, and 20 traditional) as quickly and as accurately as possible. The characters were randomly selected from those used in the character naming task; half of them had a TB structure whereas the other half had a LR configuration. All characters were of medium to high frequency and were matched in relative frequency across script types. The traditional characters had significantly more strokes than their simplified counterparts ( $t(19) = 8.26, p < .01$ ). In each trial, after a 500 ms central fixation, participants were shown a stimulus at the center of the screen, occupying around 1.5° of visual angle, and were asked to copy it as quickly and as accurately as possible. After they finished copying, they pressed a button on a response box to signal completion and the screen turned blank. Then the experimenter pressed buttons on a response box to record the accuracy and to initiate the next trial. 60 stimuli were presented in a random order in one block.

**Word Dictation Task** 40 characters (20 shared and 20 traditional/simplified) were selected from the character naming task. Each character was concatenated with a second character to compose a two-character word, and these words were used here. Words instead of characters were used to avoid the ambiguity due to the many homophone characters in Chinese. All words were of medium to high frequency (Taiwan Ministry of Education, 1997) and were matched in relative word frequency across different script types. Participants listened to the words presented in a female voice in their native language respectively, i.e., Cantonese for Hong Kong participants and Mandarin for Mainland participants. The audio recordings of the words were presented by a computer in a random order, and participants wrote down each word in their own script first and then in the other script, even if they thought the characters were the same in the two scripts. If they did not know how to write a character, they indicated it by putting a cross on the space. In each trial, after the words "get ready" presented on the screen for 500 ms, participants were presented with a stimulus; they then pressed buttons on a response box to indicate whether they knew how to write it or not. After

<sup>3</sup>  $A' = .5 + \left[ \text{sign}(H - F) \frac{(H - F)^2 + |H - F|}{4\max(H, F) - 4HF} \right]$ , where  $H$  and  $F$  present hit rate and false alarm rate respectively.

writing the word in both scripts, they pressed a button to indicate completion and start the next trial. Their accuracy of writing the first character of each word was assessed (to match the character naming task).

## Results

### Reading and Writing Performances

**Reading Performance** Participants' character naming performance was summarized in Table 1. Repeated measures ANOVA revealed a main effect of script type in accuracy ( $F(2, 92) = 5.40, p < .01$ ) and response time (RT) ( $F(2, 92) = 5.60, p < .01$ ); and an interaction between group and script type in accuracy ( $F(2, 92) = 14.66, p < .01$ ) and RT ( $F(2, 92) = 15.70, p < .01$ ). Simplified Chinese readers were more accurate in naming shared ( $t(46) = 1.51, p < .05$ ; although the difference was only 1%) and simplified characters ( $t(46) = 2.87, p < .05$ ) than traditional Chinese readers, and traditional Chinese readers were more accurate in naming traditional characters ( $t(46) = 3.33, p < .01$ ; these differences were not significant in RT).

Task	Script	Simplified Chinese Readers		Traditional Chinese Readers		Comparison between two groups	
		Acc	RT (ms)	Acc	RT (ms)	Acc	RT
Namin	Shared	1.00	251.05	0.99	327.12	*	
	Simplified	1.00	251.85	0.96	351.36	*	
	Traditional	0.97	286.79	0.99	326.36	*	
Copyin	Shared	0.96	7531.16	0.96	6201.31		
	Simplified	0.98	4815.76	0.94	5834.29	*	
	Traditional	0.65	12251.36	0.94	6173.10	**	**
Dictatio	Shared	0.99		0.97		*	
	Simplified	1.00		0.66		**	
	Traditional	0.20		0.99		**	

Table 1: Summary of participants reading and writing performance  
\*  $p < .05$ , \*\*  $p < .01$

**Writing Performance** Participants' writing performance was summarized in Table 1. In character copying, repeated-measures ANOVA revealed a main effect of script type in both accuracy ( $F(2, 92) = 93.40, p < .01$ ) and RT ( $F(2, 92) = 135.28, p < .01$ ), and an interaction between group and script type in accuracy ( $F(2, 92) = 81.76, p < .01$ ) and RT ( $F(2, 92) = 115.38, p < .01$ ). Traditional Chinese readers were faster ( $t(46) = -7.58, p < .01$ ) and more accurate ( $t(46) = 9.58, p < .01$ ) in copying traditional characters, but less accurate in copying simplified characters ( $t(46) = -3.58, p < .05$ ) than simplified Chinese readers. In contrast, the two groups did not differ in the accuracy or RT of copying shared characters; this suggests that both group had similar level of copying skills in shared characters. In the dictation task, a main effect of script type ( $F(2, 92) = 59.15, p < .01$ ) and an interaction between script type and group ( $F(2, 92) = 171.96, p < .01$ ) were found. Simplified Chinese readers were more accurate in recalling and writing shared ( $t(46) = 2.41, p < .05$ ) and simplified characters ( $t(46) = 6.13, p < .01$ ), but were less accurate in recalling and writing traditional characters ( $t(43) = 21.86, p < .01$ ) than traditional Chinese readers.

### Holistic Processing

In A', repeated measures ANOVA revealed main effects of congruency ( $F(1, 46) = 70.40, p < .01$ ), and character configuration ( $F(1, 46) = 33.79, p < .01$ ). Participants did better in congruent ( $M = .98, SE = .00$ ) than incongruent trials ( $M = .94, SE = .01$ ), and in processing LR ( $M = .97, SE = .02$ ) than TB ( $M = .95, SE = .02$ ) characters. A significant interaction between congruency and group ( $F(1, 46) = 6.60, p < .05$ ) indicated that traditional Chinese readers were more holistic than simplified Chinese readers in general. There was also a three-way interaction between script type, congruency, and group ( $F(2, 92) = 5.027, p < .05$ ), suggesting that the interaction between congruency and group was different across the three script types. To investigate how the two groups differed in processing different script types, we first contrasted their difference in processing simplified vs. traditional characters; we then compared their behavior in processing shared characters to examine how their experience in processing simplified/traditional characters influenced their perception of shared characters.

**Simplified vs. Traditional characters** Repeated-measure ANOVAs revealed main effects of congruency ( $F(1, 46) = 65.55, p < .01$ ) and character configuration ( $F(1, 46) = 33.26, p < .01$ ). There was an interaction between congruency and group ( $F(1, 46) = 5.30, p < .05$ ): simplified Chinese readers perceived both characters less holistically overall; and a three-way interaction between script type, congruency, and group ( $F(1, 46) = 5.11, p < .05$ ), suggesting the interaction between congruency and group was different between the two scripts.

When we examined their performance in processing the two scripts separately, in processing simplified characters, as predicted, the interaction between congruency and group was significant ( $F(1, 46) = 5.74, p < .05$ ): simplified Chinese readers processed simplified characters less holistically than traditional Chinese readers (Fig. 2a), possibly due to their expertise with simplified characters. Nevertheless, in processing traditional characters, there was no interaction between group and congruency ( $p = .76$ ; Fig. 2b). This suggests that the two groups processed traditional characters with a similar level of holistic processing, regardless of their performance difference in reading and writing traditional characters.

Since the two groups differed in some reading and writing performance measures in processing simplified characters (Table 1), to examine whether the difference in holistic processing of simplified characters was dependent on these measures, we put them as covariates in separate ANCOVA tests. Participants' reading performance in traditional and shared characters could hardly explain the holistic processing difference, because the interaction was still significant if we put their shared character reading accuracy ( $F(1, 46) = 5.62, p < .05$ ) or RT ( $F(1, 46) = 6.08, p < .05$ ), or traditional character reading RT ( $F(1, 46) = 5.66, p < .05$ ) as a covariate separately, and it was marginal when

traditional character reading accuracy was put as a covariate ( $F(1, 46) = 3.62, p = .06$ ). Similarly, when putting either their simplified character naming accuracy ( $F(1, 46) = 4.01, p < .05$ ) or simplified character copying RT ( $F(1, 46) = 5.99, p < .05$ ) as covariates the interaction between congruency and group was still significant. When we put their simplified character reading RT ( $F(1, 46) = 3.23, p = .08$ ) or copying accuracy ( $F(1, 46) = 2.71, p = .11$ ), the interaction between congruency and group became marginal. Only when we put simplified character dictation accuracy ( $F(1, 46) = .627, p = .43$ ) as a covariate did the interaction become insignificant. Furthermore, participants' difference between congruent and incongruent trials (the measure of holistic processing) in processing simplified Chinese characters was significantly correlated with their simplified character dictation ( $r = -.39, p < .05$ ) but not with reading or copying performances. These results were consistent with Tso et al.'s (2011) finding that the reduced holistic processing effect in expert Chinese character processing may depend more on writing rather than reading or copying performance.

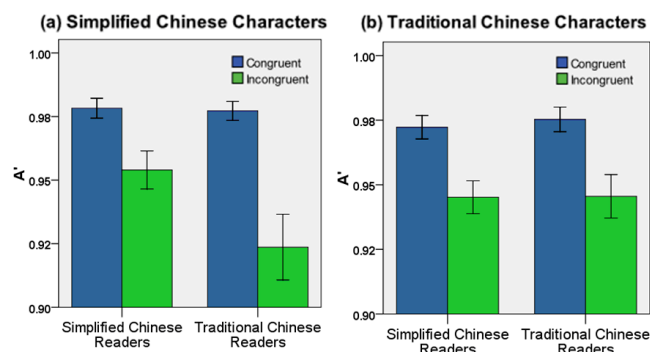


Fig. 2: Performance of simplified and traditional Chinese readers in the holistic processing task with (a) simplified Chinese characters and (b) traditional Chinese characters.

**Shared characters** A main effect of congruency was found ( $F(1, 46) = 54.96, p < .01$ ). There was an interaction between congruency and group ( $F(1, 46) = 5.01, p < .05$ ): simplified Chinese readers were less holistic than traditional Chinese readers (Fig. 3), even though these characters were shared in the two scripts. Since the two groups also differed in some reading/writing performance measures (Table 1), to examine whether the difference in holistic processing was dependent on these measures, they were put as covariates in separate ANCOVA tests. We found that the interaction between congruency and group was still significant when putting shared character naming accuracy ( $F(1, 46) = 4.72, p < .05$ ) or RT ( $F(1, 46) = 5.25, p < .05$ ), copying accuracy ( $F(1, 46) = 4.97, p < .05$ ), or dictation accuracy ( $F(1, 46) = 4.74, p < .05$ ) as a covariate. When putting shared character copying RT ( $F(1, 46) = 3.44, p = .07$ ), as a covariate, the interaction became marginally significant. However, when putting simplified character copying accuracy as a covariate, the interaction became very marginal ( $F(1, 46) = 2.52, p = .12$ ), and when putting simplified character dictation accuracy as a covariate ( $F(1, 46) = .60, p = .43$ ), the

interaction became insignificant. Also, participants' shared character holistic processing significantly correlated with simplified character dictation ( $r = -.36, p < .05$ ) but not with any other reading/writing performance measures. These results suggested that the difference in holistic processing of shared characters could not be completely accounted for by their performance difference in reading/writing shared characters; rather, it was dependent on their writing performance of simplified characters. Thus, the holistic processing difference was likely due to a transfer effect from simplified Chinese readers' expertise with simplified characters.

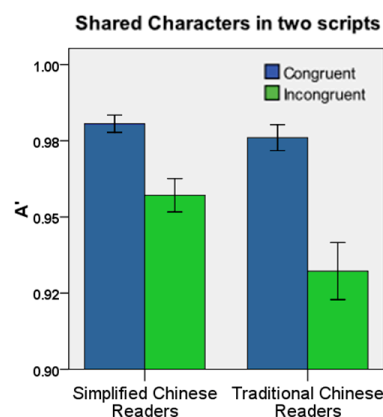


Figure 3: Performance of simplified and traditional Chinese readers in the holistic processing task with shared characters.

## Discussion

Here we examined whether simplified and traditional Chinese readers processed Chinese characters differently in terms of holistic processing, and whether their writing and reading performance measures could uniquely predict these differences. We found that when processing simplified characters, simplified Chinese readers were less holistic than traditional Chinese readers, and the difference was dependent on their word dictation performance rather than reading or copying performances. This finding is consistent with Tso et al.'s (2001) study, which showed a close relationship between writing experience and reduced holistic processing in Chinese character recognition. In contrast, although simplified Chinese readers performed much worse in both reading and writing traditional characters than traditional Chinese readers, their performance in holistic processing of traditional characters did not differ from traditional Chinese readers. This effect may be because processing simplified characters generally requires more analytic processing due to higher visual similarity among characters compared with traditional characters. Thus, simplified Chinese readers may have developed a better analytic processing skill in reading Chinese characters in general, and it could be more easily transferred to reading traditional characters compared with the generalization from traditional to simplified characters in traditional Chinese readers. This speculation is consistent with the recent finding that simplified Chinese readers have

better visual skills than traditional Chinese readers (e.g., McBride-Chang et al., 2005; Peng et al., 2010).

When processing characters that are the same in the two scripts (shared characters), simplified Chinese readers were also less holistic than traditional Chinese readers, even though both groups were experts in processing shared characters. Although simplified Chinese readers had better dictation performance in shared characters, further analysis showed that this difference in holistic processing was not dependent on their writing abilities of shared characters, but on their writing abilities of simplified characters. These findings further support our hypothesis that recognizing simplified characters requires more analytic processing than recognizing traditional characters, and this enhanced analytic processing skill is transferrable to the processing of characters that are shared in both scripts, and even to traditional characters.

Note however that the enhanced analytic processing in the simplified Chinese readers compared with the traditional Chinese readers here may also be accounted for by the difference in Chinese teaching method adopted in Mainland China and Hong Kong. In Hong Kong, children learn to read and write Chinese mainly through rote repetition, whereas in Mainland, children are taught explicitly about character components. As the result, simplified Chinese readers in Mainland may be generally more sensitive to the internal constituent components of characters than traditional Chinese readers in Hong Kong (McBride-Chang et al., 2005). To rule out the influence from the teaching methods in accounting for the current results, future work will examine whether similar effects can be found in traditional Chinese readers in Taiwan, where Children are also taught about character components explicitly.

In conclusion, here we show that expertise in reading and writing simplified Chinese characters equips readers with better analytic processing skills that are transferrable to the processing of shared and even traditional characters.

### Acknowledgments

We are grateful to the Research Grant Council of Hong Kong (project code: HKU 745210H to J.H. Hsiao).

### References

- Chen, P. (1999). *Modern Chinese: History and Sociolinguistics*. Cambridge, UK: Cambridge University Press.
- Chen, Y. P., Allport, D. A., & Marshall, J. C. (1996). What are the functional orthographic units in Chinese word recognition: The stroke or the stroke pattern? *Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, 49, 1024-1043.
- Gauthier, I., & Tanaka, J. W. (2002). Configural and holistic face processing: The Whole story. *Journal of Vision*, 2 (7), 616.
- Gao, D.-G., & Kao, H. S. R. (2002). Psycho-geometric analysis of commonly used Chinese characters. In H. S. R. Kao, C.-K. Leong, & D.-G. Gao (Eds.), *Cognitive*

- neuroscience studies of the Chinese language* (pp. 195–206). Hong Kong: Hong Kong University Press.
- Ge, L., Wang, Z., McCleery, J. P., & Lee, K. (2006). Activation of face expertise and the inversion effect. *Psychological Science*, 17, 12-16.
- Ho, C. S. H., Ng, T. T., & Ng, W. K., (2003). A ‘radical’ approach to reading development in Chinese: The role of semantic radicals and phonetic radicals. *Journal of Literacy Research*, 35 (3), 849-878.
- Hsiao, J. H., & Cottrell, G. W. (2009). Not all visual expertise is holistic, but it may be leftist: The case of Chinese character recognition. *Psychological Science*, 20, 455-463.
- Hsiao, J. H., & Shillcock, R. (2006). Analysis of a Chinese phonetic compound database: Implications for orthographic processing. *Journal of Psycholinguistic Research*, 35, 405-426.
- Kwan, T. W. (2001). Hong Kong, Mainland China & Taiwan: Chinese character frequency-A trans-regional, diachronic survey. *Character Frequency*. Retrieved October 30, 2011 from <http://arts.cuhk.edu.hk/Lexis/chifreq/>
- McBride-Chang, C., Chow, B. W. Y., Zhong, Y., Burgess, S., & Hayward, W. (2005). Chinese character acquisition and visual skills in two Chinese scripts. *Reading and Writing*, 18, 99-128.
- McCleery, J. P., Zhang, L., Ge, L., Wang, Z., Christiansen, E. M., Lee, K., & Cottrell, G. W. (2008). The roles of visual expertise and visual input in the face inversion effect: Behavioral and neurocomputational evidence. *Vision Research*, 48, 703-715.
- Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, 34, 72–107.
- Peng, G., Minett, J. W., & Wang, W. S. -Y. (2010). Cultural background influences the liminal perception of Chinese characters: An ERP study. *Journal of Psycholinguistics*, 23, 416-426.
- Taiwan Ministry of Education. (1997). 八十六年常用語詞調查報告書. Retrieved October 30, 2011 from [http://www.edu.tw/files/site\\_content/M0001/86news/ch2.html?open](http://www.edu.tw/files/site_content/M0001/86news/ch2.html?open)
- Tso, R. V. Y., Au, T. K., & Hsiao, J. H. (2011). The Influence of Writing Experiences on Holistic Processing in Chinese Character Recognition. In L. Carlson, C. Hoelscher, & T.F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1442-1447). Austin, TX: Cognitive Science Society.



# The Chinese Route Argument: Predicting the Longitude and Latitude of Cities in China and the Middle East Using Statistical Linguistic Frequencies

**Max M. Louwerse (mlouwerse@memphis.edu)**

Department of Psychology / Institute for Intelligent Systems, University of Memphis  
365 Innovation Drive, Memphis, TN 38152 USA

**Sterling Hutchinson (schtchns@memphis.edu)**

Department of Psychology / Institute for Intelligent Systems, University of Memphis  
365 Innovation Drive, Memphis, TN 38152 USA

**Zhiqiang Cai (zcaai@memphis.edu)**

Institute for Intelligent Systems, University of Memphis  
365 Innovation Drive, Memphis, TN 38152 USA

## Abstract

The Chinese Room argument describes a thought experiment that suggests that for symbols to become meaningful, they must be grounded in perceptual experiences. Embodied cognition theorists frequently use this argument to claim that cognition requires perceptual simulation. We shed light on the symbol grounding problem by arguing that the structure of natural language provides language users with cues that allow them to bootstrap meaning from non-grounded symbolic co-occurrences, such that the statistical linguistic structure can bootstrap meaning with minimal grounding. Two studies show that co-occurrences of both Chinese and Arabic city names can reliably predict their longitude and latitude in China and in the Middle East. Using the statistical linguistic technique Latent Semantic Analysis, similarity ratings were obtained for Chinese city names (Study 1) and for Arabic city names (Study 2). Multidimensional scaling (MDS) coordinates of these similarity ratings correlated with the actual longitude and latitude of these cities, showing that cities that are located together share similar semantic contexts. These results suggest that the Chinese Room argument might be substituted with a Chinese Route argument: statistical linguistic frequencies of word co-occurrences provide language users with implicit cues about how to form perceptual representations.

**Keywords:** symbol grounding problem; geography; spatial cognition; latent semantic analysis; symbolic cognition; embodied cognition

## Chinese Room Argument

A monolingual English speaker sits in a room; all he has is a Chinese newspaper. Even though there is a wealth of Chinese language at his disposal, few would argue that he understands Chinese. In fact, even if he can successfully find a specific Chinese word, e.g., 上海, in his newspaper, and the collocations of that word, say, 北京, and 香港, there is little evidence he knows the meaning of those words. This Chinese room argument has been used by Searle (1980) to illustrate the symbol grounding problem in cognition (Harnad, 1990), which questions a computational account of meaning acquisition.

Many cognitive scientists place a strong theoretical emphasis upon how symbols become grounded (Barsalou, 1999; Glenberg, 1997; Harnad, 1990; Pulvermüller, 1999; Searle, 1980). These researchers express an increased concern regarding symbolic representations of meaning, and do not endorse analogies between computational and human approaches towards deriving meaning (Pecher & Zwaan, 2005; Semin & Smith, 2008). Embodiment theorists state that meaning cannot lie within arbitrary amodal symbol systems; instead, meaning extraction continuously involves the activation of perceptual experiences. Indeed, learning Chinese as a monolingual English speaker with only a Chinese-Chinese dictionary would lead to a symbolic merry-go-round (Harnad, 1990). Computationally translating symbols into other symbols is however what computer models do, and computational simulations are therefore fundamentally different from human cognitive processes (Glenberg & Robertson, 2001). One computational model of meaning extraction that is frequently used is latent semantic analysis (LSA; Landauer, McNamara, Dennis, & Kintsch, 2007). LSA establishes meaning representation based on co-occurrences of words in same contexts. The words *Beijing* and *Shanghai* therefore have a high similarity in (computational) meaning. However, because *Beijing* is not grounded in perceptual experiences, it does not have human-like meaning (Glenberg & Robertson, 2001). For instance, when humans process the word *Beijing* they might see a map of China and are able to 'see' the city in the northeast of the country. Estimating the location of *Beijing* using a corpus-based computational model would seem impossible, because of the Chinese Room argument. If the location of *Beijing* were to be estimated using amodal symbol systems, explicit spatial cues, such as prepositions and cardinal directions are needed (e.g., *Beijing is north of Shanghai*). Such spatial cues would lead to a combinatorial explosion with each city being added (*Beijing is north of Shanghai, Chongqing is southwest of Beijing, Chongqing is west of Shanghai, etc.*).

Returning to the Chinese room with the monolingual English speaker described earlier, how could our English speaker possibly extract geographical locations from Chinese newspapers without seeing a map of China and seeing the cities marked on the map? Louwerse and Zwaan (2009) concluded that language encodes geographical information. Louwerse and Zwaan took the 50 largest cities in the United States and computed their co-occurrence frequencies in the New York Times, Wall Street Journal, and Los Angeles Post. None of these newspaper corpora necessarily described the spatial locations of the American cities. Yet, using LSA, Louwerse and Zwaan (2009) were able to estimate the longitude and latitude of the 50 cities using statistical linguistic frequencies for each of the three corpora (for a detailed description of the method being used, see below). Moreover, the population size of these cities could be estimated using frequency in the newspapers. The computational estimates were on par with human performance. The findings in this study showed that statistical linguistic frequencies can be used to estimate the location and the size of cities. Determining the semantic associations between cities in a corpus allows for estimating physical distance between cities. In fact, a heuristic like this might be used during cognitive map construction (Goldstein & Gigerenzer, 2002).

Harnad (1990) suggested that ungrounded symbolic representations can inherit meaning from grounded words related to them; we similarly propose that with minimal grounding of some symbols, the meaning of all symbols can be bootstrapped. If language encodes geographical information, we will be at least one step closer towards a bootstrapping solution. That is, the vacuum of the Chinese room now becomes an opportunity for a Chinese route: with very limited symbol grounding, the native speaker of English can bootstrap the geography of China.

### Latent Semantic Analysis

To test the possibility of a Chinese Route Argument, i.e., the Chinese language encodes geographical locations in China, we used LSA on a sample of texts segmented into paragraphs as input. Mathematical transformations created a large term-document matrix from the input. For example, if there are  $m$  terms in  $n$  paragraphs, a matrix of  $A = (f_{ij} \times G(j) \times L(i, j))_{m \times n}$  is obtained. The value of  $f_{ij}$  is a function of the integer that represents the number of times term  $i$  appears in document  $j$ ,  $L(i, j)$  is a local weighting of term  $i$  in document  $j$ , and  $G(j)$  is the global weighting for term  $j$ . Such a weighting function is used to differentially treat terms and documents to reflect knowledge that is beyond the collection of the documents. As in most LSA studies (Dumais, 2007; Martin & Berry, 2007), we used natural log as the local weight and log entropy as the global weight in the current analyses. The large matrix of  $A$  has, however, much redundant information. Singular Value Decomposition decomposes the matrix  $A$  into three matrices  $A = U \Sigma V^T$ ; where  $U$  is an  $m$  by  $m$  square matrix and  $V$  is an  $n$  by  $n$  square matrix, with  $\Sigma$  being an  $m$  by  $n$  diagonal matrix

with singular values on the diagonal. Removing dimensions corresponding to smaller singular values and keeping the dimensions corresponding to larger singular values reduces the representation of each word to a low dimensional vector. Although the new representation for the words (the reduced  $U$  matrix) is no longer orthogonal, each word now becomes a weighted vector on 300 dimensions, with only the most important dimensions that correspond to larger singular values being kept. The number of dimensions can be determined *ad hoc*, but we followed the trend set by most LSA studies and used 300 factors (Landauer & Dumais, 1997). The semantic relationship between words can be estimated by taking the cosine between two vectors. With LSA the semantic relatedness is not only determined by the relation between words, but also by the words that accompany a word (Landauer & Dumais, 1997).

Two studies each tested two hypotheses: 1) Cities that are located together are debated together. That is, cities that are in close geographical proximity are also in close proximity in the text, so that language structure itself provides cues to derive perceptual-semantic information. 2) Cities that are populated more are debated more. That is, larger cities are talked about more, so that city word frequency provides cues about the importance of the city. In Study 1 we tested these hypotheses with city names in China in a Chinese text corpus, in Study 2 with city names in the Middle East in an Arabic corpus.

### Study 1: China

In Study 1 we used a Chinese corpus collected online, consisting of 4 of the most popular classic fiction books, 29 popular modern fiction books, 26 history books, 49 philosophy books, 34 economy books, 15 politics books, and 8 military books. These books provided 86MB of text in 14768 documents (paragraphs) and 47,226 word types. In terms of text size, 33.4% texts are in history, 24.6% in philosophy, 10.4% in economics, 9.3% in modern fiction, 9.3% in politics, 7% in military and 6% in classic fiction. Note that the texts did not explicitly describe geographical relations between Chinese cities, and that the corpus was very heterogeneous.

The standard procedure was used when creating the LSA space, whereby each word was a weighted vector on 300 dimensions. The 50 largest cities in China were selected, and their latitude and longitude were determined using census data. All cities had a population size of more than one million ( $M = 2,393,188$ ,  $SD = 2,340,707.88$ ) (Table 1).

Cosine values were computed for each of the city pairs. Two cities resulted in cosine errors and were removed from the analysis, resulting in a 48 x 48 cosine matrix. This matrix was submitted to an MDS analysis using the ALSCAL algorithm. A Euclidean distance measure transformed the semantic similarities into dissimilarities, such that the higher the value, the longer the distance. Default MDS criteria were used with an S-stress convergence of .001, a minimum stress value of .005, and a maximum of 30 iterations.



We chose a low-dimensionality to rule out over-fitting the data. The fitting on a two-dimensional scale was moderate, with a Stress value = .33 and an  $R^2 = .59$ . The LSA estimated coordinates of the 48 cities were compared with the actual coordinates of the cities.

The loadings of the 48 cities on the two dimensions generated by the MDS analysis correlated with the longitude and latitude of the cities, latitude – dimension 1,  $r = .64$ ,  $p < .001$ ,  $n = 48$ ; longitude – dimension 2,  $r = .33$ ,  $p = .02$ ,  $n = 48$ .

To do justice to the geometry of the 2D variables, we used bi-dimensional regression analyses to compare the computational estimates with the actual coordinates of the 50 cities. Tobler (1964) and Friedman & Kohler (2004) introduced bi-dimensional regressions in order to compute the mapping of any two planes under consideration. Whereas in a uni-dimensional regression each data point is shifted by intercept and slope, each actual and predicted

value of the dependent variable are presented by a point in space, whereby vectors represent intercept and slope.

A bi-dimensional regression yielded a significant correlation between the LSA estimates and the actual city coordinates,  $r = .57$ ,  $p < .001$ ,  $n = 48$ . This result supported the hypothesis that Chinese cities that are located together in China are debated together in the Chinese language.

The question can be raised whether the bi-dimensional regressions not always yield significant correlations. To answer this question we conducted 1000 Monte Carlo simulations on the 48  $x$  and  $y$  pairs. The average bi-dimensional regression of these simulations yielded no significant result, average  $r = .13$  ( $SD = .06$ ),  $p = .37$ ,  $n = 48$ .

In addition, we tested the hypothesis that cities that are populated more are debated more by comparing the frequency of the 50 cities in the Chinese corpus with their actual population size. A Pearson correlation was significant,  $r = .47$ ,  $p < .001$ ,  $n = 50$ .

Table 1: Chinese Cities

City	Lat.	Long.	Dim.1	Dim.2	City	Lat.	Long.	Dim.1	Dim.2
上海	31.23	121.40	-0.38	-1.42	苏州	31.30	120.60	-0.83	-1.34
北京	39.93	116.40	-0.13	-1.24	汕头	23.37	116.60	-1.04	-0.69
重庆	29.57	106.50	-1.07	0.81	荣成	23.54	116.30	1.05	-0.17
西安	34.27	108.90	-0.92	1.28	兰州	36.05	103.60	-0.68	1.06
武汉	30.58	114.20	-1.25	-0.93	合肥	31.85	117.20	-0.27	1.40
成都	30.67	104.00	-1.24	0.45	抚顺	41.87	123.80	1.28	-0.50
天津	39.13	117.20	0.28	-1.26	洛阳	34.68	112.40	0.83	1.32
沈阳	41.80	123.40	1.96	-0.64	邯郸	36.58	114.40	0.88	1.09
哈尔滨	45.75	126.60	1.36	-0.19	包头	40.60	110.00	-0.98	0.68
南京	32.05	118.70	-0.47	-1.48	香港	22.27	114.10	-0.68	0.64
广州	23.12	113.20	-0.63	-1.31	苏州	34.27	117.10	-0.84	-1.31
太原	37.87	112.50	0.16	1.40	深圳	22.53	114.10	-0.74	-0.01
长春	43.87	125.30	1.98	-0.55	福州	26.08	119.30	-1.09	-0.99
石家庄	38.05	114.40	1.63	0.07	无锡	31.58	120.30	-0.35	1.57
长沙	28.20	112.90	-1.30	-0.30	淮南	32.63	116.90	-1.16	0.46
济南	36.67	117.00	0.86	-0.91	贵阳	26.58	106.70	-1.27	-0.57
大连	38.92	121.60	0.56	-0.89	鞍山	41.12	122.90	1.61	-0.55
吉林	43.85	126.50	1.70	0.11	保定	38.87	115.40	-0.17	0.99
南昌	28.68	115.80	-1.02	-0.94	咸阳	34.37	108.70	-0.01	1.42
郑州	34.75	113.60	1.25	0.28	昆明	25.05	102.70	0.01	-0.83
九龙	22.32	114.10	-0.04	1.73	大同	40.08	113.30	0.66	0.86
杭州	30.25	120.10	-0.92	-1.18	本溪	41.33	123.70	1.82	-0.30
青岛	36.07	120.30	1.03	-0.23	淮北	33.95	116.70	-0.75	1.08
唐山	39.62	118.10	0.37	1.53	常州	31.78	119.90	-1.06	0.50

## Study 2: Middle East

In order to determine whether the findings could be generalized beyond China and the Chinese language, we

used a different language (Arabic) and a different geography (the Middle East) in the second study.

An LSA space was created using an Arabic corpus collected online, consisting of books and news on history (49%), fiction (42%), politics (3%), philosophy (2%), economy (1%) and other unknown types of texts (3%). The total size of the corpus was 71.8 MB, including 27,937 paragraphs and 147,535 word types. Again, the texts did not specifically discuss the geography of the Middle East. Instead, the corpus covered many topics and was, again, very heterogeneous in nature.

Similar to the previous analysis, 50 of the largest cities across the Arabic speaking countries in the Middle East were selected ( $M = 1,304,154$ ,  $SD = 1,451,579$ ). These cities were located in Egypt, Iraq, Jordan, Kuwait, Lebanon, Oman, Syria, United Arab Emirates, and Yemen (Table 2).

Some countries were not included because the Arabic notation of cities for those countries was unavailable (Saudi Arabia, Ethiopia, Eritrea, Somalia, Djibouti).

As in the Chinese analysis, geographical location (longitude and latitude) as well as population size for these 50 cities were determined. A 50 x 50 cosine matrix was submitted to an MDS ALSCAL analysis, and the MDS coordinates were compared with the actual coordinates.

Again, the fitting on a two-dimensional scale was moderate, with Stress = .35,  $R^2 = .69$ . The LSA estimated

coordinates of the 50 cities were compared with the actual coordinates of the cities.

The loadings of the 50 cities on the two dimensions generated by the MDS analysis correlated with the longitude and latitude of the cities, latitude – dimension 1,  $r = .41$ ,  $p < .001$ ,  $n = 50$ ; longitude – dimension 2,  $r = .57$ ,  $p < .001$ ,  $n = 50$ .

A bi-dimensional regression also yielded a significant correlation between the LSA estimates and the actual city coordinates ( $r = .53$ ,  $p < .001$ ,  $n = 50$ ). These results again supported the hypothesis that cities in the Middle East that share geographical context, share textual context (cities that are located together are debated together).

As in Study 1, we ran 1000 Monte Carlo simulations to rule out the possibility that the significant bi-dimensional regressions could be obtained from an accidental pairing of the estimates. The average regression coefficient again ruled out that the findings could be obtained by chance, average  $r = .13$  ( $SD = .07$ ),  $p = .37$ ,  $n = 50$ .

Finally, as before, we compared the frequency of the 50 cities in the Arabic corpus with their actual population size. A Pearson correlation was significant ( $r = .61$ ,  $p < .001$ ,  $n = 50$ ), providing evidence for the hypothesis that cities in the Middle East that have a higher population, are talked about more frequently.

Table 2: Middle Eastern Cities

	Lat.	Long.	Dim1.	Dim.2		Lat.	Long.	Dim. 1	Dim. 2
تهران	35.67	51.43	0.2169	1.5093	كرمانشاه	34.38	47.06	-0.075	2.0242
بغداد	33.33	44.44	0.6485	-1.2606	السليمانية	35.56	45.43	1.0587	-1.0911
الرياض	24.65	46.77	-0.9637	-0.9586	اروميه	37.53	45.00	0.3161	1.2087
جدة	21.50	39.17	-1.1027	-0.7991	زاهدان	29.50	60.83	0.4181	1.1594
الموصل	36.34	43.14	1.007	-1.0505	رشت	37.30	49.63	0.3014	1.2215
مشهد	36.27	59.57	-0.2017	1.0937	الطائف	21.26	40.38	-1.044	-0.7242
كابل	34.53	69.17	1.2753	-0.0876	كرمان	30.30	57.08	-0.04	2.0397
بيروت	33.89	35.50	-0.6608	-0.9994	حماة	35.15	36.73	-1.3177	-0.7392
البصرة	30.53	47.82	0.0922	-1.4058	الحلة	32.48	44.43	1.034	-0.7171
حلب	36.23	37.17	-1.2136	-0.7176	تبوك	28.39	36.57	-1.1946	-0.7448
اصفهان	32.68	51.68	-0.4703	1.6961	كربلاء	32.61	44.03	1.0672	-1.0522
دمشق	33.50	36.32	-0.7743	-1.0202	همدان	34.77	48.58	-0.0679	2.01
كرج	35.80	50.97	0.2241	1.2555	العمارة	31.84	47.15	1.0166	-0.8805
مكة	21.43	39.82	-1.0349	-0.7787	الزرقاء	32.07	36.10	1.0995	-0.6294
تبريز	38.08	46.30	-0.9509	0.7826	اراك	34.08	49.70	0.578	0.9256
شيراز	29.63	52.57	-0.8681	0.8734	الديوانية	31.99	44.93	0.973	-0.9436
اريل	36.18	44.01	1.2833	-0.3946	خميس مشيط	18.31	42.73	-0.9944	-0.6579
عمان	31.95	35.93	-0.5422	-1.022	يزد	31.92	54.37	-0.0563	1.81
المدينة	24.48	39.59	-0.7522	-1.0729	بريده	26.37	43.97	-1.1197	-0.2557
اهواز	31.28	48.72	0.1934	1.3174	اردبيل	38.25	48.30	-0.0501	2.002
قم	34.65	50.95	-0.1317	1.2444	بغقوبة	33.74	44.65	1.1115	-0.6968
الدمام	26.43	50.10	-1.1988	-0.7561	بندر عباس	27.25	56.25	0.6984	0.9249
حمص	34.73	36.72	-1.3472	-0.7446	هرات	34.35	62.18	1.2645	-0.043
كركوك	35.47	44.39	1.0518	-1.1014	اسلام شهر	35.54	51.20	1.233	0.0334
النجف	32.00	44.34	1.0731	-1.0775	اللاذقية	35.54	35.78	-1.0628	-0.7091

## General Discussion

This study showed that statistical linguistic frequencies can be used to estimate the location and population size of cities. In the first study we estimated the location and size of cities in China using Chinese text, in the second the location and size of cities in the Middle East using Arabic texts. These findings show that the results reported by Louwerse, Cai, Hu, Ventura, & Jeuniaux (2006) for France, those reported by Louwerse and Zwaan (2009) for the United States, and by Louwerse and Benesh (in press) for (the fictional) Middle Earth can be extended to China and the Middle East. Moreover, the current study has demonstrated that geographical locations are not only encoded in English (Louwerse et al., 2006; Louwerse & Benesh, in press; Louwerse & Zwaan, 2009), but also in Chinese and Arabic.

There are several questions that should be addressed with regards to the findings reported in this study. First, we should address the question whether the findings reported in this study should be attributed to LSA or to statistical linguistic frequencies. Louwerse and Zwaan (2009) addressed this question by demonstrating that geographical locations of cities in the United States could be predicted using higher-order co-occurrences (using LSA), but also by first-order co-occurrences. Louwerse (2011), however, pointed out that for first-order co-occurrences a corpus needs to be approximately 25,000 times larger than a corpus that is the appropriate size for an LSA analysis. A second question concerns the explicit spatial cues potentially present in the corpus. After all, the argument could be made that the Chinese and Arabic corpora we used for our semantic spaces consisted of explicit spatial cues (e.g., cardinal directions, prepositions) that explained our findings rather than implicit semantic relationships. This seems extremely unlikely for two reasons. First, the LSA algorithm shows minimal sensitivity to explicit cues because it uses higher-order co-occurrences (see Landauer, McNamara, Dennis, & Kintsch, 2007). Secondly, the corpora were so diverse in nature that the results can better be explained by statistical linguistic frequencies than by the specifics of the texts.

We began this paper with the Chinese Room Argument, which suggests that meaning cannot be extracted from symbols unless a referent is perceptually activated (Searle, 1980). Even though this study did not compare the computational results with experimental data (see Louwerse and Zwaan, 2009 and Louwerse and Benesh, in press, for such a comparison), it does provide some insight in the Chinese Room argument. The current study puts forward that, with minimal grounding of some symbols (city names), the meaning of all symbols (city names) can be bootstrapped, because of the organization of the symbolic network. The language system has many built-in regularities that are utilized during cognitive processing (Louwerse, 2011; Louwerse & Jeuniaux, 2010). To illustrate this further, the current study has shown that if a language user knows the location of the city 乌鲁木齐, and knows only that the other Chinese words are Chinese city names, the

language user can bootstrap the geographical locations of these other cities on a country map of China. Moreover, they can make estimates about the size of each city, because frequency correlates with population size.

Obviously, we do not deny the essence of the symbol grounding problem: the language user must ground at least one symbol and must also have partial meaning with regards to the other words (i.e., know that they are city names). Moreover, the geographical estimates are relative estimates, rather than a specific longitude and latitude. However, findings like these do challenge an extreme view of symbol grounding that dismisses the possibility of statistical linguistic frequencies playing a significant role in cognition. Experimental evidence has shown that statistical linguistic frequencies often explain experimental findings better than perceptual simulations account do (Louwerse, 2008; Louwerse, 2011), yet whether humans rely more on statistical linguistic frequencies or perceptual simulations depends on at least the cognitive task and the stimulus (Louwerse & Jeuniaux, 2010). We therefore advocate the pursuit of a unified account in which both statistical linguistic frequencies and perceptual simulation help establishing meaning. In line of this research agenda, this study has shown that with minimal grounding the symbolic vacuum of the Chinese Room can become a guiding Chinese route.

## Acknowledgments

This research was in part supported by grant NSF BCS-0904909. The usual exculpations apply.

## References

- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-660.
- Dumais, S. T. (2007). LSA and information retrieval: Getting back to basics. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.
- Friedman, A., & Kohler, B. (2003). Bidimensional regression: A method for assessing the configural similarity of cognitive maps and other two-dimensional data. *Psychological Methods*, 8, 468-491.
- Glenberg, A. M. (1997). What memory is for: Creating meaning in the service of action. *Behavioral and Brain Sciences*, 20, 41-50.
- Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory & Language*, 43, 379-401.
- Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109, 75-90.
- Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42, 335-346.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of

- acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
- Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.
- Louwerse, M. M. (2008). Embodied relations are encoded in language. *Psychonomic Bulletin & Review*, 15, 838-844.
- Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *TopiCS in Cognitive Science*, 3, 273-302.
- Louwerse, M. M. & Benesh, N. (in press). Representing spatial structure through maps and language: Lord of the Rings encodes the spatial structure of Middle Earth. *Cognitive Science*.
- Louwerse, M. M., Cai, Z., Hu, X., Ventura, M., & Jeuniaux, P. (2006). Cognitively inspired natural-language based knowledge representations: Further explorations of Latent Semantic Analysis. *International Journal of Artificial Intelligence Tools*, 15, 1021-1039.
- Louwerse, M. M., & Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. *Cognition*, 114, 96-104.
- Louwerse, M.M. & Zwaan, R.A. (2009). Language encodes geographical information. *Cognitive Science*, 33, 51-73.
- Martin, D. I., & Berry, M. W. (2007). Mathematical foundations behind latent semantic analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.
- Pecher, D., & Zwaan, R. A. (Eds.) (2005). *Grounding cognition: The role of perception and action in memory, language, and thinking*. New York, NY: Cambridge University Press.
- Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, 22, 253-270.
- Searle, J.R. (1980). Minds, brains, and programs. *Behavioral & Brain Sciences*, 3, 417- 424.
- Semin, G. R. & Smith, E. R. (Eds.) (2008). *Embodied grounding: Social, cognitive, affective, and neuroscientific approaches*. New York, NY: Cambridge University Press.
- Tobler, W. R. (1964). Bidimensional regression. *Geographical Analysis*, 26, 187-212.

# Modeling Multiple Strategies for Solving Geometric Analogy Problems

Andrew Lovett (andrew-lovett@cs.northwestern.edu)

Kenneth Forbus (forbus@northwestern.edu)

Qualitative Reasoning Group, EECS Department, 2133 Sheridan Road  
Evanston, IL 60208USA

## Abstract

We present an improved computational model for performing geometric analogy. The model combines two previously modeled strategies and makes explicit claims about when people will use one strategy or the other. We compare the model to human performance on a classic problem set. The model's strategy shifts, along with working memory load, account for most of the variance in human reaction times.

**Keywords:** geometric analogy; visual problem-solving; structure-mapping

## Introduction

Visual problem-solving has long been a popular tool for evaluating people's cognitive abilities (Raven, Raven, & Court, 1998; Dehaene et al., 2006). Problem-solving tasks frequently involve a sequence of images (e.g., Figure 1). Individuals must compare the images, identifying some pattern across them. They must then apply this pattern, finding the answer that best completes (or violates) it.

Visual problem-solving depends on a comparison process for identifying commonalities and differences in images. We have previously argued that structure mapping (Gentner, 1983), a theory of analogical comparison, may also explain concrete visual comparison in humans (Markman & Gentner, 1996; Lovett et al., 2009a; Sagi, Gentner, & Lovett, in press). According to structure mapping, people compare stimuli by aligning the common relational structure in symbolic, qualitative representations. We have posited that structure mapping may play a ubiquitous role, identifying commonalities and differences and estimating similarity. Based on this hypothesis, we have built models of three visual problem-solving tasks: geometric analogy (Lovett et al., 2009b), Raven's Progressive Matrices (Lovett, Forbus, & Usher, 2010), and the oddity task (Lovett & Forbus, 2011a).

Here, we complement our model of visual comparison with a model of *visual inference*. Visual inference explains how individuals apply a set of differences to one image to create a novel image representation. It plays a key role in tasks such as geometric analogy (Figure 1), where participants are asked "A is to B as C is to..." We show how this leads to a new model for geometric analogy. Rather than assuming participants always infer the correct answer, the model makes explicit claims about when visual inference will succeed and when it will fail, requiring a shift to an alternate strategy. The model's strategic shifts correlate well with human reaction times on a classic geometric analogy problem set (Evans, 1968).

In the following section, we provide some background on the geometric analogy task. We then present our computational model, which utilizes two strategies for performing the task. Afterwards, we compare the model against human performance and discuss the results. We close with related work and conclusions.

## Background

Geometric analogy involves comparing images to identify differences. However, researchers disagree on which comparisons people make. The debate encompasses two competing strategies. The first involves inserting each possible answer into the analogy to evaluate it. Consider Figure 1A. The strategy proceeds as follows:

- 1) Compare A and B to get  $\Delta(A,B)$ , the differences between A and B. Here there is a change from two overlapping objects to one object inside the other.

- 2) For each possible answer  $i$ , compare C to  $i$  to get  $\Delta(C,i)$ , the differences between C and that answer. Then, perform a second-order comparison: measure the similarity of  $\Delta(A,B)$  to  $\Delta(C,i)$ . Whichever answer produces the most similar set of differences is chosen. Here answer 3 produces an identical set of differences to  $\Delta(A,B)$ , so it is chosen.

The second strategy solves for the answer directly:

- I) Compare A and B to get  $\Delta(A,B)$ .

- II) Compare A and C to get the corresponding objects. In Figure 1A, the large leftmost shapes correspond, and the small rightmost shapes correspond.

- III) Apply the differences in  $\Delta(A,B)$  to the corresponding objects in C to infer  $D'$ , a representation of the answer. Here the small rectangle in C would move inside of the larger shape. Pick the answer most similar to  $D'$ .

Mulholland, Pellegrino, and Glaser (1980) call the first strategy *infer-infer-compare* and the second strategy *infer-map-apply*. However, this assumes that different processes are used to compare images in steps 1), 2), and II). We believe structure-mapping can determine differences, identify correspondences, and measure similarity. Therefore, we instead call the strategies *second-order comparison* and *visual inference*.

Sternber (1977) argued that people use visual inference to perform geometric analogy. However, Mulholland, Pellegrino, and Glaser (1980) found evidence that second-order comparison was being used. Bethell-Fox, Lohman, and Snow (1984) suggested that individuals may adjust their strategy, depending on problem difficulty. Their eye-tracking data demonstrated that people typically use visual inference, solving directly for the answer. However, as problems become more difficult, people may abandon this

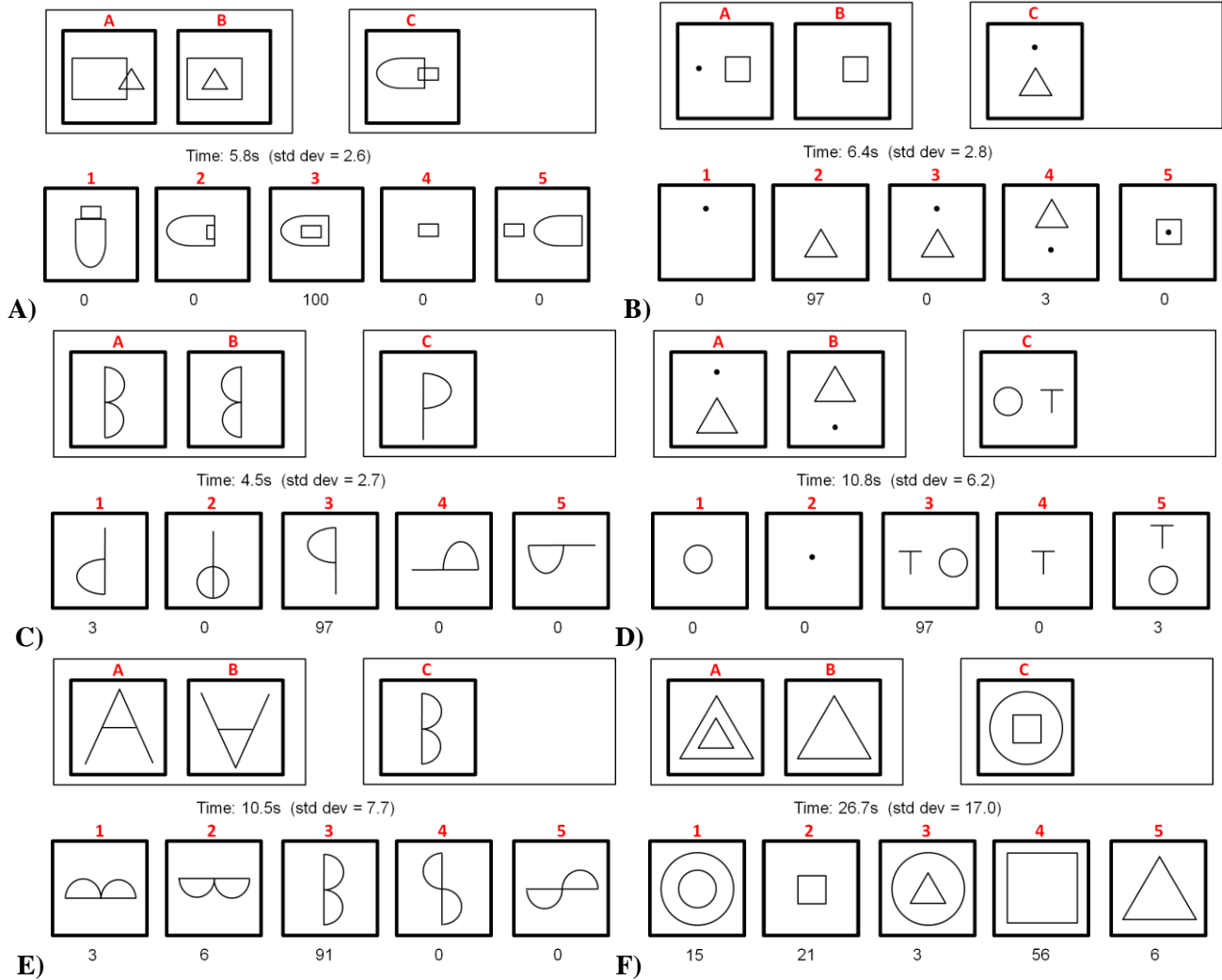


Figure 1: Geometric analogy problems with average response times and percent selecting each answer (Lovett et al., 2009b).

approach, instead trying out each possible answer in the analogy.

One might conclude that people use visual inference when problems are easy and second-order comparison when problems are relatively difficult. However, this is not the full story. Sternberg (1977) found that participants used visual inference on multiple choice problems, where they had to consider several possible answers. Mulholland, Pellegrino, and Glaser (1980) found that participants used second-order comparison on true/false problems, where they saw a completed analogy and simply judged whether it was correct. The Sternberg problems appear more difficult—at the very least, they put more load on working memory, as there are multiple answers to consider. Why, then, would people use second-order comparison on the easier Mulholland et al. problems?

We believe the answer lies in the algorithmic complexity of the two strategies. In the analysis below, we compute each strategy's complexity by counting the number of comparisons necessary to solve a problem. We concede that this may be a simplification; for example, second-order comparisons may require more time and effort than others.

However, our model predicts that a common process (structure-mapping) is used for all comparisons. Thus the number of times this process repeats should provide a reasonable approximation of a strategy's complexity.

Consider second-order comparison. If there are  $n$  possible answers, then the number of comparisons is  $1$  (A to B) +  $n$  (C to each answer) +  $n$  ( $\Delta(A,B)$  to  $\Delta(C,n)$  for each answer) =  $2n + 1$ .

Now consider visual inference. The number is  $1$  (A to B) +  $1$  (A to C) +  $n$  (D' to each answer) =  $n + 2$ . However, this strategy also requires a non-comparison operation: inferring D' by applying  $\Delta(A,B)$  to the corresponding objects in C.

Suppose we have a true/false problem. Then the number of answers  $n = 1$ . The number of comparisons is 3 for both strategies. In this case, one might prefer second-order comparison, as it doesn't require inferring D'. This explains why Mulholland et al. found that participants used second-order comparison.

Suppose we have a multiple-choice problem. The number of answers  $n > 1$ . Now, there will be fewer comparisons for visual inference, and this should be the preferred strategy, as Sternberg found. However, if a problem is particularly

complex, participants may be unable to perform the inference operation. In this case, they may fall back on the second-order comparison strategy, which requires only comparison operations.

Our model, described below, attempts to solve problems via visual inference. When it fails, either because it cannot perform the inference or because the inferred image doesn't match any of the answers, the model reverts to second-order comparison. Thus, the model can help explain why some problems take longer to solve than others: certain problems require a shift from the default visual inference strategy to a less efficient second-order comparison strategy.

## Model

Our model depends on three core processes: perception, visual comparison, and visual inference. Below, we describe each process and then show how the model combines these processes to implement different problem-solving strategies.

### Perception

Our model uses the CogSketch sketch understanding system (Forbus et al., 2011) to generate a qualitative representation for each image. Qualitative, or *categorical*, representations are abstract, describing features like relative position (**right of**) or relative orientation (**parallel**), rather than exact numerical sizes and orientations. There is abundant psychological evidence that people are sensitive to such features (e.g., Huttenlocher, Hedges, & Duncan, 1991; Rosielle & Cooper, 2001).

CogSketch performs sketch understanding, rather than full vision. It requires users to draw separate line drawings of each object in a visual scene (e.g., the rectangle and triangle in image A of Figure 1A). Given these objects, CogSketch automatically computes spatial relations between objects and attributes for individual objects.

Our model takes the process a step further. It can automatically segment an object into edges and build an edge-level representation, describing qualitative spatial relations between the edges. Alternatively, it can group several objects together based on similarity to build a group-level representation. See Lovett and Forbus (2011b) for details on this process, as well as the full vocabulary of qualitative terms at the edge, object, and group levels.

### Visual Comparison

We model visual comparison using the Structure-Mapping Engine (SME) (Falkenhainer, Forbus, & Gentner, 1989), a computational model based on Gentner's (1983) structure-mapping theory. Given two cases described in predicate calculus, it computes a mapping between them by aligning their common relational structure. SME is biased to prefer aligning deep structure. For example, at the edge level, a first-order relation might indicate that there is a convex corner between two edges. A second-order relation might indicate that two convex corners are adjacent. These higher-order relations, which take other relations as their arguments, receive priority during mapping.

SME computes up to 3 mappings between the compared cases. Each mapping contains: A) a similarity score based on the breadth and depth of aligned structure; B) correspondences between the entities and expressions in the two cases; and C) *candidate inferences* (CIs), inferences based on expressions in one case that failed to align with the other. For example, consider the A/B comparison in Figure 1A. A forward CI (A->B) would indicate that the two shapes no longer partially overlap. A backward CI (B->A) would indicate that one shape no longer contains the other. CIs are useful for identifying differences between the cases.

The model compares two images via the following steps:

- 1) Compare the qualitative image representations using SME. This produces a mapping indicating the corresponding objects, commonalities, and differences.

- 2) For each pair of corresponding objects, compare the objects' shapes to identify a shape transformation. This is done by comparing the objects' edge-level representations to get corresponding edges, and then using those correspondences to compute a transformation, such as a rotation or reflection (see Lovett & Forbus, 2011b).

Sometimes there are multiple valid transformations. For example, in Figure 1C, there is both a rotation and a reflection between the 'B' shapes. In such cases, the model picks the simplest transformation, according to the following rankings: identity, reflection, rotation. Objects can also change scale, becoming larger or smaller.

- 3) Compute the similarity between images. This is primarily based on SME's similarity score, but it is updated according to the shape comparisons: if two objects are identical, the images will be rated more similar.

- 4) Compute a qualitative, structural representation of the differences between the images. This describes differences of the following types:

- A) Spatial relation addition/removals, based on the CIs.

- B) Reversals of spatial relations. This is a special case of A) where two objects swap places in a relation. In Figure 1D, the dot and triangle swap places in an **above** relation.

- C) Object additions/removals, where objects are added or removed between images (e.g., Figure 1B).

- D) Object transformations, where there is a shape transformation between corresponding objects.

### Visual Inference

The visual inference operation applies a set of differences to an image to produce a novel image representation. In geometric analogy, the A/B differences are applied to C to produce D', a representation of the answer image. Consider Figure 1A. Inference proceeds as follows:

- 1) Compare image A to image C to get the corresponding objects.

- 2) Apply the A/B differences to the corresponding objects in C to produce a new qualitative representation:

- A) Add or remove spatial relations.

- B) Reverse the arguments of spatial relations.

- C) Add or remove objects. If an object is added, create a new object in CogSketch, basing it off some existing object.



If an object is removed, remove any spatial relations referring to that object.

D) For all other objects, apply the appropriate shape transformation to the object in C to create a new object for D'. This might mean leaving the shape unchanged (identity), rotating it, reflecting it, scaling it, etc.

E) Compute shape attributes for all the newly created objects, and add them to the D' representation.

Note that D' contains: a) a qualitative, structural image description; and b) a set of concrete, quantitative objects. Thus, it contains just enough to support visual comparisons between D' and other images. However, D' is not a concrete image: the model lacks exact, quantitative locations for each object.

There are several ways that visual inference can fail:

A) When a spatial relation cannot be added because the objects it describes are not found in C (there are no corresponding objects).

B) When a spatial relation cannot be reversed. In Figure 1D, there is a reversal of **above** in the A/B differences. However, there is no **above** in C to reverse.

C) When an object cannot be removed or transformed because the object is not found in C.

D) When transforming an object doesn't produce the desired effect. On Figure 1E, the model reflects the 'B' shape over the x-axis. However, when it compares the result to the original 'B' shape, they appear identical (recall that identity is ranked before reflection). The model treats this as a failure to transform.

The model is focused on generation: adding expressions to C's representation to produce D'. Thus, visual inference fails when a spatial relation cannot be added to C or reversed in C, but it does *not* fail when a spatial relation cannot be removed from C. For example, in Figure 1B, the A/B differences include removing a **rightOf** relation. There is no such relation in C to be removed. Visual inference succeeds here, whereas it fails in 1D, where there is no **above** relation in C to be reversed. Thus, the model explains why 1D is a harder problem (compare the reaction times).

## Geometric Analogy

Our new geometric analogy model solves problems via two strategies: visual inference and second-order comparison. For visual inference, the model compares A and B to get  $\Delta(A,B)$ , the differences between them. It applies  $\Delta(A,B)$  to C to get D', a representation of the answer image. It compares D' to each possible answer. If an answer is sufficiently similar, it selects that answer.

For second-order comparison, the model compares A and B to get  $\Delta(A,B)$ . For each answer *i*, it compares C and *i* to get  $\Delta(C,i)$  and then compares  $\Delta(C,i)$  to  $\Delta(A,B)$  (again, using SME). If an answer's  $\Delta(C,i)$  is sufficiently similar to  $\Delta(A,B)$ , it selects that answer.

In each case, an answer is sufficiently close if either a) SME detects no differences; or b) the SME similarity score lies above a *similarity threshold*. We use a similarity threshold of 0.8 (where 1.0 is a perfect match). However, a

sensitivity analysis shows that our results would be the same for values ranging from .67 to .87. If multiple answers tie for the best score, this is treated as a failure.

Note that when SME compares  $\Delta$ 's for second-order comparison, it is possible to find a perfect match even for non-identical  $\Delta$ 's. SME supports tiered identity (Falkenhainer, 1990), where non-identical predicates can align when they are members of a common category. For example, in Figure 1D,  $\Delta(A,B)$  and  $\Delta(C,3)$  each involve reversal of a positional relation (**above** and **rightOf**). Thus, Figure 1D is not solvable by visual inference, but it is easily solvable by second-order comparison.

**Strategic Shifts** The model first attempts to solve a problem via visual inference. This can fail in two ways: either the inference operation may fail (Figures 1D, 1E), or D' may fail to match any of the answers. For example, in Figure 1F, the A/B differences show the inner shape being removed. The model applies these differences to C to infer an image with a large circle, which matches none of the answers.

If visual inference fails, the model reverts to second-order comparison. When the model utilizes this strategy, it must make two other strategic decisions: the comparison mode when comparing A to B, and the comparison mode when comparing C to each answer *i*. The comparison modes are:

A) Normal: Images are compared as described above.

B) Reflection: Instead of preferring identity during shape comparison, the model prefers reflection. In Figure 1E, the C/3 comparison will find a y-axis reflection between the 'B' shapes, instead of treating them as identical.

C) Rotation: Instead of preferring identity during shape comparison, the model prefers rotation.

D) Alternate: The model looks for an alternate mapping between the images. In Figure 1F, an alternate A/B mapping aligns the small triangle in A with the large triangle in B.

The model only considers an alternate mapping when SME finds more than one mapping between the images. It only considers the Reflection/Rotation modes when the images each contain a single object, allowing the model to focus on different ways of comparing that one object.

The model independently varies the mapping mode for A/B and C/*i* comparisons, beginning with Normal for each. It terminates when it identifies a sufficient answer. If no such answer is found, it picks the highest-scoring answer.

## Experiment

We evaluated our model on 20 geometric analogy problems from Evans (1968). We recreated each problem in PowerPoint and then imported the problems into CogSketch. This required us to manually segment each problem into images (image A, image B, etc) and segment each image into objects (each object was drawn as a separate shape in PowerPoint). Beyond this, the model automatically segmented each object into edges and generated representations at the edge and object levels—this problem set did not contain any groups of objects.

Our prior behavioral study (Lovett et al., 2009b) provides data on human performance which our simulation models, so we summarize it next.

### Behavioral Study

The Evans problems were shown to 34 adult participants. They were given a description of the geometric analogy task followed by two simple example problems (without feedback) before they saw the 20 problems. Both the ordering of the problems and the ordering of the five possible answers were randomized across participants.<sup>1</sup>

Before each problem, participants clicked on a fixation point in the screen’s center to indicate readiness. After the problem was presented, participants clicked on the picture that best completed the analogy. Participants were instructed to be as quick as possible without sacrificing accuracy. The two measures of interest were the answer chosen and the reaction time, i.e., the time taken to solve the problem.

**Results** The results show a high degree of consistency across participants. All participants chose the same answer for 9 of the 20 problems, while over 90% chose the same answer for 7 additional problems. The greatest disagreement was on Figure 1F, where only 56% chose the same answer. Henceforth, we refer to the answer chosen by the majority as the *preferred answer*. In reporting and analyzing reaction times (including Figure 1), we consider only responses with the preferred answer, filtering out minority responses.

### Simulation & Analysis

The model chose the preferred answer on all 20 problems. This indicates that our approach—qualitative representation, comparison via structure mapping, and visual inference—is sufficient for matching human performance on the task.

We next asked whether people take longer to solve problems where our model must make a strategy shift. We coded each problem for three factors: Alt-Strategy, Alt-Mapping, and Alt-Transform. Alt-Strategy indicates that our model reverts to second-order comparison to solve the problem. Alt-Mapping indicates that it uses the Alternate image mapping mode. Alt-Transform indicates that it uses the Reflection or Rotation mapping modes. We group these mapping modes together, as our model uses the same mechanism for computing both transformation types.

We also coded each problem for working memory load. Previous research has shown that geometric analogy problems get harder as either the number of elements or the number of transformations increases (Mulholland, Pellegrino, & Glaser, 1980; Bethell-Fox, Lohman, & Snow, 1984). Mulholland et al. found that this effect was non-linear: there was a higher cost when the numbers of both

elements and transformations increased. They suggested this was because at some point the problem exceeds people’s working memory capacity, requiring a shift in strategy.

We coded for working memory load by counting the number of elements in  $\Delta(A,B)$ , the differences between images A and B. This is a key representation for both visual inference and second-order comparison. Because Mulholland et al. found a non-linear effect of working memory, we discounted the first two elements. Thus, if  $\Delta(A,B)$  was one or two, the WM Load was coded as zero.

We ran a linear regression to identify the effect of the above factors on human reaction times. Table 1 shows the results. Overall, this model achieves an  $R^2$  of .95 (.93 adjusted), meaning it explains almost all the variance in human reaction times. The grayed cells indicate which factors made a significant contribution to the model ( $p < .01$ ). The intercept of 6.4 indicates that the easiest problems took around 6.4 s, while the various factors increased the time to complete a problem.

Note that with correlations, extreme values can result in an overestimation of the explained variance (the  $R^2$  value). In this case, participants took far longer to solve the two problems requiring the Alt-Mapping shift (e.g., Figure 1F). If we remove these data points and rerun the analysis, Alt-Mapping ceases to be a factor, and  $R^2$  drops to .80 (.76 adjusted). Thus, even discounting these difficult problems, the regression explains most of the variance in performance.

Table 1. Linear model for human reaction times on geometric analogy (grayed cells are significant factors).

Intercept	WM Load	Alt-Strategy	Alt-Transform	Alt-Mapping
6.4 s	5.7 s	4.4 s	- 0.7 s	10.5 s

The only factor that did not contribute significantly was Alt-Transform. Alt-Transform refers to problems like Figure 1E, where the model must switch to a Reflection mode to identify a reflection between the identical ‘B’ shapes. The analysis suggests there is no increased cost for Alt-Transform problems. However, this does not mean such problems are easy; they are difficult in that the model must make the Alt-Strategy shift to solve them, changing to second-order comparison. Once this strategy shift has been made, there is no additional cost for the Alt-Transform shift.

### Related Work

Evans’ ANALOGY (1968) was the first computational model of analogy. A ground-breaking system, it solved the same 20 geometric analogy problems as our model using second-order comparison. However, its brute-force comparison processes do not align well with human cognition (see Lovett et al., 2009b for a discussion).

Our own previous model (Lovett et al., 2009b) also solved problems via second-order comparison. The present approach builds on that model by implementing visual inference as a complementary strategy. The previous model explained .56 of the human variance on the Evans problems,

<sup>1</sup> Due to experimenter error, some participants received the same random orderings. As many as five received one ordering, but on average only 1.5 received the same ordering. When we randomly selected one instance of each ordering, the participant number dropped to 22, and the pattern of results remained the same.

whereas the current model explains .95 of the variance. However, the previous analysis did not consider multiple factors or filter out reaction times for minority responses.

Several other approaches have utilized visual inference strategies, but these suffer from important limitations. Some (Schwering et al., 2009; O'Donoghue, Bohan, & Keane, 2006) use hand-coded symbolic inputs, rather than automatically generating representations. This means the models are unable to reason about quantitative spatial information, e.g., shape transformations. Others (Ragni, Schleipen, & Steffenhagen, 2007) are unclear on their comparison processes. Finally, because these models have not been systematically evaluated on a pre-existing problem set, it is unclear how well they match human performance.

## Conclusions

We believe our model is the first to combine two established problem-solving strategies: visual inference and second-order comparison. Beyond utilizing both strategies, the model makes explicit claims about when people will abandon visual inference and fall back on second-order comparison. Our analysis shows that these claims help explain human reaction times on the 20 Evans problems: people take longer to solve problems where the model reverts to second-order comparison, and they take even longer when the model must find an alternate mapping.

Importantly, our two problem-solving strategies are not unique to geometric analogy. We recently (Lovett & Forbus, in prep) integrated these strategies into a new model of Raven's Progressive Matrices, a more complex task that is popularly used to evaluate general intelligence. As that model and the present model show, successful problem-solving requires flexibly moving between different comparison strategies. These models, along with our oddity task model (Lovett & Forbus, 2011a), also demonstrate the utility of structural alignment across qualitative representations. In the future, we plan to evaluate the generality of our approach on new problem-solving tasks.

## Acknowledgments

This work was supported by NSF SLC Grant SBE-0541957, the Spatial Intelligence and Learning Center (SILC).

## References

- Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, 8, 205-238.
- Dehaene, S., Izard, V., Pica, P., & Spelke, E. (2006). Core knowledge of geometry in an Amazonian indigene group. *Science*, 311, 381-384.
- Evans, T. (1968). A program for the solution of geometric-analogy intelligence test questions. In M. Minsky (Ed.), *Semantic Information Processing*. Cambridge: MIT Press.
- Falkenhainer, B. (1990). Analogical interpretation in context. *Proceedings of CogSci '90*.
- Falkenhainer, B., Forbus, K., & Gentner, D. (1989). The structure mapping engine: Algorithm and examples. *Artificial Intelligence*, 41, 1-63.
- Forbus, K., Usher, J., Lovett, A., Lockwood, K., & Wetzel, J. (2011). CogSketch: Sketch understanding for cognitive science research and for education. *Topics in Cognitive Science*, 3(4), 648-666.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.
- Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, 98(3), 352-376.
- Lovett, A., & Forbus, K. (in preparation). A model of strategic comparison during visual problem-solving.
- Lovett, A., & Forbus, K. (2011a). Cultural commonalities and differences in spatial problem-solving: A computational analysis. *Cognition*, 121(2), 281-287.
- Lovett, A., & Forbus, K. (2011b). Organizing and representing space for visual problem-solving. *Proceedings of Qualitative Reasoning Workshop*.
- Lovett, A., Forbus, K., & Usher, J. (2010). A structure-mapping model of Raven's Progressive Matrices. *Proceedings of CogSci '10*.
- Lovett, A., Gentner, D., Forbus, K., & Sagi, E. (2009a). Using analogical mapping to simulate time-course phenomena in perceptual similarity. *Cognitive Systems Research* 10(3), 216-228.
- Lovett, A., Tomai, E., Forbus, K., & Usher, J. (2009b). Solving geometric analogy problems through two-stage analogical mapping. *Cognitive Science*, 33(7), 1192-1231.
- Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory & Cognition*, 24(2), 235-249.
- Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, 12, 252-284.
- O'Donoghue, D. P., Bohan, A., & Keane, M. T. (2006). Seeing things - Inventive reasoning with geometric analogies and topographic maps. *New Generation Computing*, 24(3), 267-288.
- Ragni, M., Schleipen, S., & Steffenhagen, F. (2007). Solving proportional analogies: A computational model. Workshop on Analogies: Integrating Multiple Cognitive Abilities, CogSci '07.
- Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Oxford: Oxford Psychologists Press.
- Rosielle, L. J., & Cooper, E. E. (2001). Categorical perception of relative orientation in visual object recognition. *Memory & Cognition*, 29(1), 68-82.
- Sagi, E., Gentner, D., & Lovett, A. (in press). What difference reveals about similarity. *Cognitive Science*.
- Schwering, A., Gust, H., Kühnberger, K., & Krumnack, U. (2009). Solving geometric proportional analogies with the analogy model HDTP. *Proceedings of CogSci 2009*.
- Sternberg, R. J. (1977). *Intelligence, Information Processing and Analogical Reasoning*. Hillsdale, NJ: Erlbaum.

# A unified theory of counterfactual reasoning

Christopher G. Lucas

cglucas@cmu.edu

Department of Psychology  
Carnegie Mellon University

Charles Kemp

ckemp@cmu.edu

Department of Psychology  
Carnegie Mellon University

## Abstract

A successful theory of causal reasoning should be able to account for inferences about counterfactual scenarios. Pearl (2000) has developed a formal account of causal reasoning that has been highly influential but that suffers from at least two limitations as an account of counterfactual reasoning: it does not distinguish between counterfactual observations and counterfactual interventions, and it does not accommodate backtracking counterfactuals. We present an extension of Pearl's account that overcomes both limitations. Our model provides a unified treatment of counterfactual interventions and backtracking counterfactuals, and we show that it accounts for data collected by Sloman and Lagnado (2005) and Rips (2010).

In addition to reasoning about actual states of affairs, humans find it natural to reason about what might have been. A doctor may ask “if Alice had not been treated with the experimental drug, would she have survived?” and a parent might tell a child that “if you had been paying attention, you wouldn't have gotten hurt.” Researchers from several disciplines have developed formal models of counterfactual reasoning, and recent empirical studies have evaluated the psychological merits of some of these models (Rips, 2010; Dehghani, Iliev, & Kaufmann, 2012). This paper describes a new model of counterfactual reasoning and evaluates it using data sets from the psychological literature.

The problems that we consider can be illustrated using a causal chain over three variables (Figure 1a). For example, suppose that  $A$ ,  $B$ , and  $C$  are variables that indicate whether three transponders are active. Transponder  $A$  is active about half of the time, and whenever it is active it tends to activate  $B$ , which in turn tends to activate  $C$ . Suppose that we observe on a certain occasion that all three transponders are active. We can now ask counterfactual questions such as “if  $B$  had not been active, would  $C$  have been active?”

The formal approach that we present is inspired by the work of Pearl (2000), who developed a model of counterfactual reasoning that we refer to as the *modifiable structural model*, or MSM for short. The MSM assumes that the causal system in question is a *functional causal model*, where *exogenous* variables are introduced if necessary so that the variables of primary interest are deterministic functions of their parents. For example, the system in Figure 1a may be represented more precisely by adding exogenous variables  $U_A$ ,  $U_B$  and  $U_C$  such that  $U_A$  determines whether or not node  $A$  is active, and  $U_B$  and  $U_C$  capture factors such as atmospheric conditions that determine whether the links in the chain operate successfully. Suppose now that  $A$ ,  $B$  and  $C$  are all observed to be active, and that we want to know whether  $C$  would be active if  $B$  were not active. The MSM addresses this question by using the observations in the actual world to update

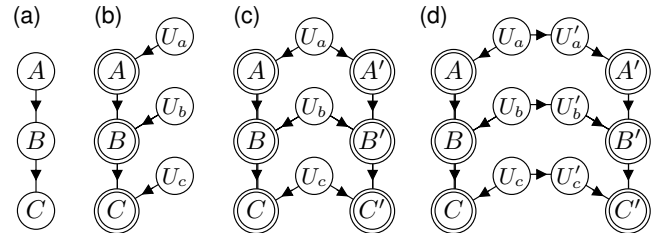


Figure 1: (a) A causal chain in which  $A$  causes  $B$ , which causes  $C$ . (b) A functional causal model that captures the causal chain in (a). Exogenous variables  $U_A$ ,  $U_B$  and  $U_C$  have been added, and nodes with double edges are deterministic functions of their parents. (c) A twin network where  $A'$ ,  $B'$  and  $C'$  represent counterfactual values of  $A$ ,  $B$ ,  $C$ . (d) An augmented twin network which allows for the possibility that the exogenous variables take different values in the counterfactual scenario.

prior beliefs about the status of the exogenous variables, then *modifying* the resulting causal model to reflect a counterfactual intervention where  $B$  is forced to be inactive. Inferences about any other variables can be computed using the modified causal model—for example, if  $B$  were inactive, then  $C$  would probably also be inactive. Several psychological studies of counterfactual reasoning have evaluated the predictions of the MSM and have found that some counterfactual inferences do appear to be treated as inferences about counterfactual interventions (Sloman & Lagnado, 2005; Kemp, Shafto, & Tenenbaum, 2012).

The approach we present builds on the key ideas behind the MSM, and we refer to it as the *doubly-modifiable structural model* or DMSM for short. Like the MSM, the DMSM works with causal systems that are represented using functional causal models and allows these systems to be modified via counterfactual interventions. In addition, however, the DMSM permits a second kind of modification where exogenous variables are altered not because of a counterfactual intervention, but simply because the counterfactual world might have turned out differently from the real world. An important consequence of this difference is that the DMSM alone accounts for backtracking counterfactuals. For example, suppose again that all three transponders in  $A$  were observed to be active, and we are asked to decide whether  $A$  would be active if  $B$  were inactive. The DMSM allows for the possibility that variables upstream of  $B$  might explain the counterfactual premise that  $B$  is inactive, and thus predicts that  $A$  is likely to be inactive. The MSM, however, can only reason about

the downstream consequences of a counterfactual intervention that renders  $B$  inactive.

The inability of the MSM to deal with backtracking counterfactuals is widely acknowledged, and Hiddleston (2005) has developed a *minimal networks* model that overcomes this limitation. Rips (2010) has recently evaluated the psychological merits of Hiddleston’s account, and reports that Hiddleston’s model can account for human inferences about backtracking counterfactuals when supplemented with additional psychological assumptions. Although this variant of Hiddleston’s approach accounts relatively well for the data collected by Rips, the minimal networks model is not well suited for reasoning about counterfactual interventions, and does not account for empirical data suggesting that counterfactual inferences are sometimes treated as inferences about counterfactual interventions.

At present, then, the psychological literature on counterfactual reasoning is fragmented. The MSM provides an elegant account of reasoning about counterfactual interventions but does not account for inferences about backtracking counterfactuals. The minimal networks model can handle backtracking counterfactuals, but does not give a clear account of inferences about counterfactual interventions. In contrast, the DMSM accommodates both counterfactual interventions and backtracking counterfactuals, and we will show that it accounts for previously-published experiments that explore both kinds of inferences. As we discuss towards the end of the paper, the DMSM is not a complete account of counterfactual reasoning, but we believe that it comes closer to this goal than any previous model.

### The Modifiable Structural Model (MSM)

The MSM was introduced informally above, and we now describe how the predictions of this model can be computed by constructing and manipulating a twin network (Pearl, 2000). The first step is to specify a functional causal model such as the example in Figure 1b that captures the causal system under consideration. Functional causal models are described in detail by Pearl (2000), but for our purposes, their most important feature is that they represent noise or randomness using unobserved exogenous variables, rather than inherently stochastic relationships. This functional model is converted into the twin network in Figure 1c by adding nodes  $A'$ ,  $B'$  and  $C'$  that represent counterfactual versions of  $A$ ,  $B$ , and  $C$ . The counterfactual nodes  $A'$ ,  $B'$  and  $C'$  and the original nodes  $A$ ,  $B$ , and  $C$  have the same exogenous variables as parents, which captures the idea that the causal mechanisms in the counterfactual world are identical to the causal mechanisms in the actual world. Given the twin network, a counterfactual premise can be captured using an intervention that fixes the value of one of the counterfactual variables. For example, suppose again that  $A$ ,  $B$  and  $C$  are all active and we are asked about a scenario where  $B$  is inactive. The counterfactual premise is captured using graph manipulation to modify the twin network. In other words, we set  $B'$  to 0, and remove all arrows between  $B'$  and its parents to reflect the fact that  $B'$

was fixed by an intervention instead of being brought about by  $U_b$  and  $A'$ . We can now use the manipulated twin network to compute predictions about the other counterfactual variables. Because  $A$  must have been caused by  $U_A$  and  $U_A$  also causes  $A'$ , the MSM infers that  $A'$  is active. Because  $B'$  is inactive, the MSM infers that  $C'$  is also inactive.

Two aspects of the MSM are worth emphasizing for comparison with the DMSM described in the next section. First, the MSM handles all counterfactual queries by reasoning about counterfactual interventions. The model therefore does not distinguish between counterfactual interventions (“imagine that someone had disabled transponder B”) and counterfactual observations (“imagine that you had observed that B was inactive”). Second, the MSM cannot make inferences about backtracking counterfactuals. If asked to imagine that  $B$  were inactive, the MSM fixes the status of  $B'$  by means of an intervention and therefore cannot reason about upstream variables such as  $A'$  which may explain the inactivity of  $B'$ .

### The Doubly-Modifiable Structural Model (DMSM)

Just as the MSM can be characterized in terms of computations over a twin network, the DMSM can be characterized in terms of computations over an *augmented twin network*. The augmented twin network for the three element chain is shown in Figure 1d. Note that the network includes nodes for counterfactual versions of the exogenous variables  $U_A$ ,  $U_B$  and  $U_C$  in addition to nodes for counterfactual versions of  $A$ ,  $B$  and  $C$ . The value of each counterfactual exogenous variable is either copied across from the corresponding real-world variable or generated from the same distribution as the corresponding real-world variable. More precisely, if  $P_i(\cdot)$  is the prior distribution on exogenous variable  $U_i$ , the value of  $U'_i$  is drawn from the distribution

$$P(U'_i|U_i) = s\delta(U_i) + (1-s)P_i(U_i)$$

where  $\delta(U_i)$  is a delta distribution that takes value 0 at every point except  $U'_i = U_i$  and  $s$  is a stability parameter where  $0 \leq s \leq 1$ . If  $s = 1$ , then the exogenous variables are perfectly stable, which means that  $U'_i = U_i$  for all  $i$  and that the DMSM is equivalent to the MSM<sup>1</sup>. If  $s = 0$ , then the exogenous variables are maximally unstable, and the values of  $U'_i$  and  $U_i$  are independently drawn from the distribution  $P_i(\cdot)$ . We will refer to this special case as the USM, or “unattached structural model” because setting  $s = 0$  decouples the counterfactual nodes from the actual nodes, meaning that the model effectively discards all observations of the actual world.

We propose that people are sensitive to both the true state of the world and base rate information, and therefore hypothesize that the judgments of most individuals reflect stability values between 0 and 1. A second hypothesis is that some individuals always use  $s = 0$  and others always use  $s = 1$ . A third hypothesis is that each individual uses  $s = 0$  in some

<sup>1</sup>A stability of 1 for counterfactual observations can lead to mutually incompatible or impossible states, so we assume that all counterfactual premises are treated as interventions when  $s = 1$ .

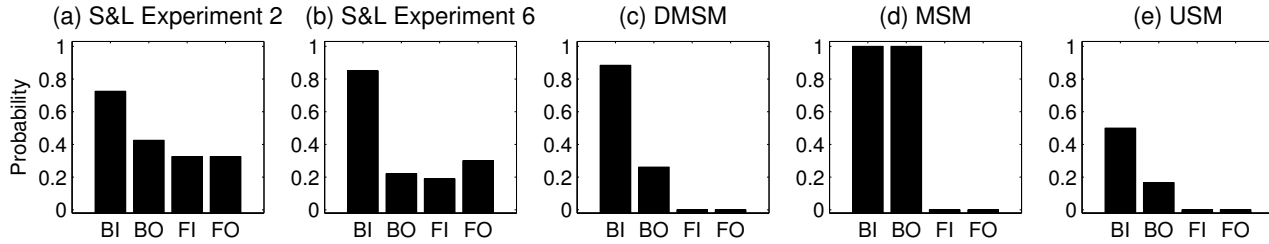


Figure 2: Human judgments and predictions of the DMSM, MSM and USM models for Sloman and Lagnado’s Experiments 2 and 6. The four bars in each plot show inferences about backtracking (B) or forward (F) counterfactuals that were either interventional (I) or observational (O).

contexts and  $s = 1$  in other contexts. We return to these hypotheses later and describe some preliminary evidence that supports the first hypothesis. Different individuals may use different settings of  $s$ , but for all analyses we set  $s = 0.77$  which is the value that maximizes model performance across the entire set of studies reported by Rips (2010).

The augmented twin network can be used to address two kinds of counterfactual queries. Queries about counterfactual *interventions* are addressed by manipulating the network in the standard way—for example, counterfactual interventions on  $B$  are captured by fixing the value of  $B'$  and removing all arrows between  $B'$  and its parents. Queries about counterfactual *observations* are carried out by reasoning over the unmanipulated network. For example, if  $A$ ,  $B$  and  $C$  are observed to be active and we want to reason about a case where  $B$  is observed to be inactive, we set  $A = B = C = 1$  and  $B' = 0$  and can subsequently compute the posterior distribution induced on any other node. For instance, if the stability parameter is less than one then the posterior distribution  $P(A' | A = B = C = 1, B' = 0)$  will indicate that  $A'$  is relatively likely to take value 0.

### Counterfactual interventions vs. observations

A key conceptual difference between the MSM and the DMSM is that the DMSM allows for counterfactual observations and counterfactual interventions, but the MSM treats all counterfactual queries in terms of interventions. To the best of our knowledge, two experiments reported by Sloman and Lagnado (2005) are the only psychological studies that contrast counterfactual observations and counterfactual interventions. This section compares the predictions of the DMSM with the results of these experiments.

Experiment 2 of Sloman and Lagnado (2005) considers a three node chain where  $A$  causes  $B$  and  $B$  causes  $C$ . The experiment included three different cover stories—one scenario involved a rocket ship, and a second involved a causal chain where smoking causes cancer which causes hospitalization. The third involved abstract events  $A$ ,  $B$ , and  $C$  and participants were told that “when  $A$  happens, it causes  $B$  most of the time” and “when  $B$  happens, it causes  $C$  most of the time.” In all cases, participants were told that  $A$  and  $C$  happened and were asked to make counterfactual inferences about a situation where  $B$  did not happen. The counterfactual questions

asked about both counterfactual interventions and counterfactual observations. In the abstract scenario, the backwards intervention (BI) question stated that “someone intervened directly on  $B$ , preventing it from happening,” and asked participants to rate the probability that  $A$  would have happened. The forward intervention (FI) question was similar except that it asked participants to rate the probability that  $C$  would have happened. The backwards and forwards observation questions (BO and FO) asked participants to rate the probability that  $A$  and  $C$  “would have happened if we observed that  $B$  did not happen.”

Average human responses are shown in Figure 2a. Responses were originally provided on a 1 to 5 scale, but we map them to probabilities for comparison with model predictions. Following Sloman and Lagnado (2005), responses are collapsed across the three different cover stories. Both the intervention and observation questions produce the forward inference that  $C$  is unlikely to occur. The two kinds of questions, however, lead to different backward inferences about  $A$ . Participants tend to infer that  $A$  would still have occurred if a counterfactual intervention had prevented  $B$ , but find it less likely that  $A$  would have occurred if  $B$  had been *observed* not to occur.

Experiment 6 of Sloman and Lagnado (2005) is similar in structure but involves a two node chain rather than a three node chain. The cover story described a rocket ship with two components where “movement of component  $A$  causes component  $B$  to move.” Participants were informed that both components are moving and asked to reason about counterfactual cases where either  $A$  or  $B$  was not moving. The counterfactual intervention questions were of the form “suppose component  $A$  were prevented from moving, would component  $B$  still be moving?” The counterfactual observation questions were of the form “suppose component  $A$  were observed to be not moving, would component  $B$  still be moving?”

The proportion of participants who responded “yes” to each question is shown in Figure 2b. As for Experiment 2, forward inferences about  $B$  given  $A$  are similar regardless of whether  $A$  is prevented from moving or simply observed not to move. Backward inferences about  $A$  given  $B$  again reveal a difference between counterfactual observations and counterfactual interventions. As for Experiment 2, participants tend

to infer that  $A$  would still be moving if  $B$  were prevented from moving, but are less likely to infer that  $A$  would be moving if  $B$  were observed not to move.

We generated model predictions for the two experiments by making the simplest possible assumptions about the parameters in each causal structure. The base rate for node  $A$  was set to 0.5, and the strength of each causal link was set to 0.8 to capture the fact that causes produce their effects “most of the time.” To keep the analysis simple we assumed that nodes  $B$  and  $C$  had no background causes. Given these assumptions, predictions of the DMSM and the MSM are shown in Figures 2c and 2d. The DMSM accounts for the result that counterfactual interventions and counterfactual observations are treated differently. The MSM accounts for human responses to the intervention questions, but makes identical predictions about responses to the observation questions.

Although the DMSM performs better than the MSM, the quantitative predictions of the DMSM depart from human inferences in some cases. For example, humans give non-zero responses to the forward questions in Experiment 6, but the DMSM infers that component  $B$  is definitely not moving if component  $A$  is not moving. Including a background cause of  $B$  would allow the DMSM to match the human responses to the forward questions more closely.

The MSM can be viewed as a special case of the DMSM where the stability parameter  $s$  is equal to 1, and the USM model in Figure 2e is the special case where  $s = 0$ . The USM distinguishes between counterfactual observations and interventions, but its inferences in the BI case are not shaped by the observation that event  $A$  occurred in the real world. As a result, the model falls back on the baseline probability that  $A$  occurs, and does not account for the human inference that  $A$  probably occurred in the counterfactual scenario. Note, however, that these data do not permit us to distinguish between the DMSM and a mixture of MSM and USM strategies. The analysis in the next section will partially address this issue.

## Backtracking counterfactuals

The previous section suggested that the DMSM improves on the MSM by distinguishing between counterfactual interventions and counterfactual observations. One important consequence of this distinction is that the DMSM alone is able to handle backtracking counterfactuals, or queries where a reasoner must think about causes that might be responsible for a counterfactual premise. Rips (2010) has carried out an extensive psychological study of backtracking counterfactuals, and this section argues that the DMSM accounts for Rips’ data about as well as the minimal network model that he advocates. Dehghani et al. (2012) have also developed a theory that handles backtracking counterfactuals, and have presented some data in support of their theory. They do not describe a fully-specified computational model, but we were able to implement a model that we believe is consistent with their core assumptions. This model, however, did not account for Rips’ data as well as the minimal network model, and we there-

fore focus here on comparing the DMSM with the minimal network model.

Experiment 3 in Rips (2010) asked participants to reason about four causal systems shown at the top of Figure 3. Each system includes components  $L$  and  $H$  which cause component  $C$  to operate. The systems in Figure 3a include two cases where the operation of  $C$  is *jointly caused*: arcs between edges in Figure 3a indicate that  $L$  and  $H$  operate together to cause  $C$  to operate. The remaining systems are cases where the operation of  $C$  can be *separately caused* by either  $L$  or  $H$ . Two of the systems include probabilistic causal relationships shown as dashed arrows. For example, the probabilistic jointly caused system was described as a system where component  $L$ ’s operating and component  $H$ ’s operating together usually cause component  $C$  to operate. The remaining two systems include deterministic causal relationships. For each of the four systems, base rates for causes  $L$  and  $H$  were provided. Participants were told that  $L$  operates 5% of the time, and that  $H$  operates 95% of the time.

In each case participants were told that components  $L$ ,  $H$ , and  $C$  “are all operating.” They then responded to the question “if component  $C$  were not operating, would component  $L$  be operating?” and answered a similar question with respect to component  $H$ . The red points in Figure 3b show the proportion of participants who said yes to each question. For the deterministic separately caused system, most participants inferred that neither  $L$  nor  $H$  would be operating in the counterfactual scenario. For all remaining systems, participants tended to infer that the cause with higher base rate would be operating.

Predictions for four models are shown in Figure 3 using black lines. To generate these predictions, we again assumed that the probabilistic causal relationships had a strength of 0.8. The experimental materials did not explicitly specify whether each counterfactual scenario involved an intervention or an observation, and the predictions for the DMSM and the USM are based on counterfactual observations.

The DMSM predicts that average responses for the deterministic separately caused system should be low, and accounts for the base rate effects observed for all other systems. In contrast, the MSM predicts that the answer to all eight questions should be yes. Since the counterfactual premise is treated as an intervention, the model infers that  $L$  and  $H$  are operating in each counterfactual scenario. The USM makes predictions that are fairly close to the predictions of the DMSM, but inferences about the deterministic separately caused system reveal one important difference. Because the rare cause  $L$  was observed to operate in the actual world, the DMSM assigns non-negligible probability to the conclusion that  $L$  is operating in the counterfactual scenario. The USM ignores all information about the actual world, and therefore generates a much lower probability. Of these two models, only the DMSM successfully predicts that the probability of  $L$ ’s operating is higher for the probabilistic separately caused system than for any other system.



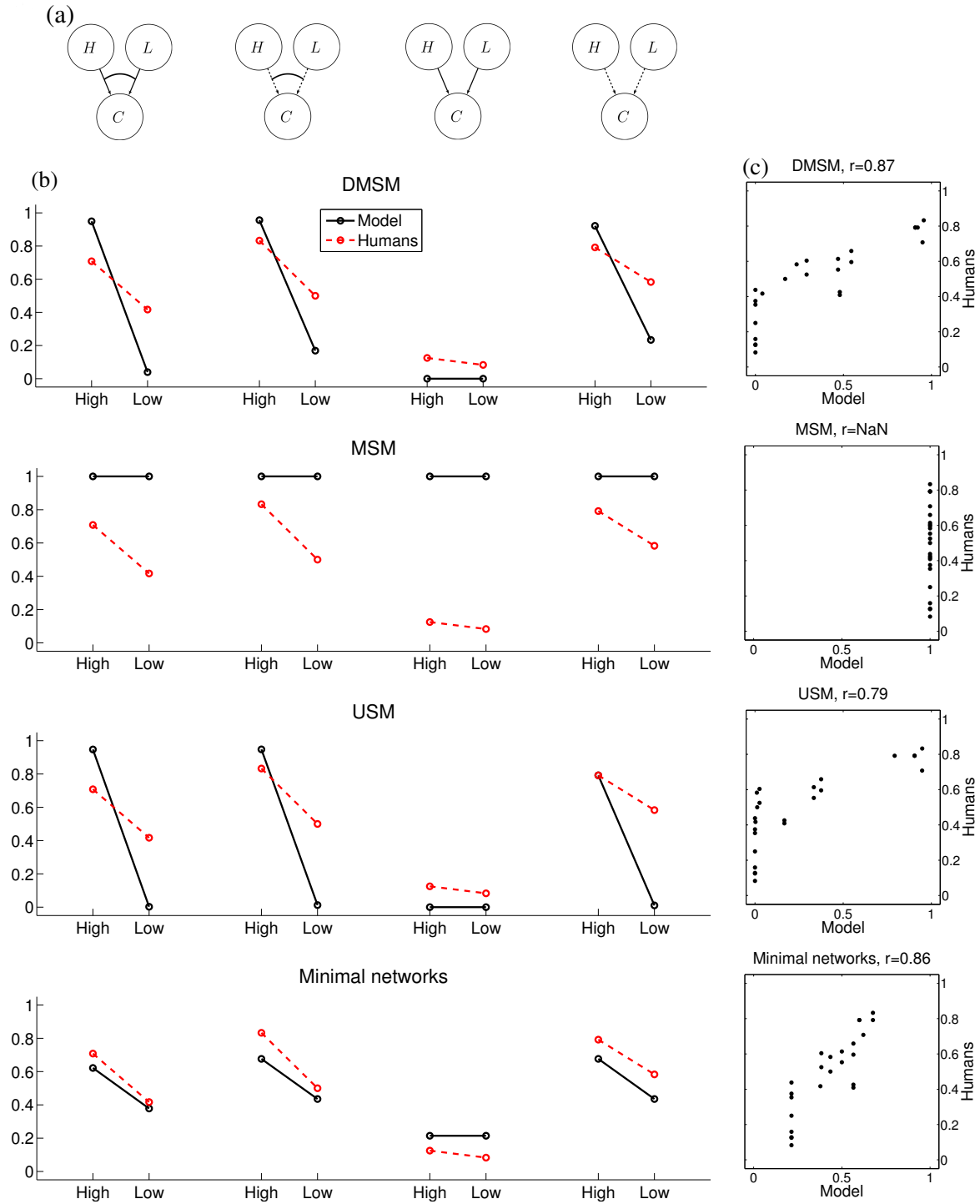


Figure 3: Model predictions and human responses for the experiments of Rips (2010). (a) Causal systems used in Rips's Experiment 3. Dotted lines represent probabilistic relationships and solid lines represent deterministic relationships. The edges linked by arcs represent AND relationships, where both causes must be active for the effect to occur, while un-linked edges represent OR relationships. Base rates differed between variables, with  $H$  or "High" variables occurring with probability .95 and  $L$  or "low" variables occurring with probability 0.05. (b) Results for counterfactual queries about the systems in (a). Proportions of "yes" judgments (human data) and probabilities (model predictions) are grouped according to the causal systems in (a). (c) Model predictions versus human judgments for all conditions across all four experiments in Rips (2010).

Because the MSM makes constant predictions across all of the different cases, an account in which individuals adopt either an MSM or USM strategy yields the same correlation with the data as the USM. Consequently, the DMSM fits the data better than accounts where some individuals always use  $s = 0$  and others always use  $s = 1$ , or where each individual tosses a coin to set  $s = 0$  or  $s = 1$  on a question-by-question basis. Even so, our data do not decisively show that most individuals are characterized by intermediate values of  $s$ , and future empirical work is needed to address this question.

The final plot in Figure 3b shows predictions for the minimal network model described by Rips. Unlike the three models considered thus far, the minimal networks approach works with causal models such as Figure 1a instead of functional models such as Figure 1b. Given observations of the actual world (e.g.  $L = H = C = 1$ ) and a counterfactual premise (e.g.  $C = 0$ ), the minimal networks approach assumes that the counterfactual scenario contains a minimal set of breaks, or cases where variables differ from their actual values while their immediate causes do not.<sup>2</sup> In Hiddleston's (2005) original version of minimal networks theory, counterfactual queries receive affirmative answers only if they are true in *all* minimal networks. This version of the theory accounts poorly for the data in Figure 3, and the predictions shown there are based on a variant of the original theory that incorporates two additional assumptions suggested by Rips. First, if multiple configurations are minimal then participants are assumed to respond based on just one of these networks. With probability  $\theta_1$  participants sample one minimal network at random, and with probability  $1 - \theta_1$  minimal networks are sampled according to their prior probabilities. Second, Rips proposes that with probability  $\theta_2$ , a participant will ignore the evidence that they see and pick an answer at random. We set the parameters  $\theta_1$  and  $\theta_2$  to values that maximize the correlation between model predictions and human data.

Figure 3b shows that the minimal networks model accounts well for the data, and produces quantitative predictions that are superior to the DMSM. Note, however, that the DMSM has one free parameter and the minimal networks model has two. The DMSM can be adjusted in the same way as the minimal network model to allow for the fact that some participants responded randomly, and making this modification will bring the quantitative predictions of the DMSM into closer correspondence with human judgments.

We have focused so far on Experiment 3 of Rips (2010), but the scatterplots in Figure 3c summarize the performance of the four models across all four of Rips' experiments. The two best performing models are the DMSM and the minimal networks model. The DMSM therefore accounts for Rips' data as well as his own model despite requiring fewer free parameters. The DMSM performs better than the USM and the mixed-strategy account, but recall that the DMSM has one free parameter and the USM has no free parameters. Addi-

tional studies are therefore needed to confirm that sensitivity to observations about the actual world is critical when modeling human inferences about backtracking counterfactuals.

## Discussion

We presented a model of counterfactual reasoning that accounts for inferences about both counterfactual interventions and backtracking counterfactuals. Our approach is closely related to the modifiable structural model developed by Pearl and inherits the ability of this model to reason about counterfactual interventions. Our model, however, differs from the MSM in one critical respect: we allow for the fact that exogenous causal variables may take counterfactual values. We showed that this difference between the models allows the DMSM but not the MSM to account for Rips' experimental study of backtracking counterfactuals.

Although we believe that the DMSM is a step towards a unified theory of counterfactual reasoning, there are important theoretical and empirical questions that still need to be addressed. The DMSM accommodates both counterfactual observations and counterfactual interventions, but additional work is needed to characterize the conditions under which a generic counterfactual premise is interpreted as an observation or an intervention. A second direction for future work is to draw a sharper contrast between the DMSM and accounts that combine the predictions of the MSM and the USM using a weighted average, and to better understand individual differences in counterfactual reasoning.

**Acknowledgments.** This work was supported by the James S. McDonnell Foundation Causal Learning Collaborative Initiative and by NSF award CDI-0835797.

## References

- Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language*, 27(1), 55-85.
- Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs*, 39(4), 632-657.
- Kemp, C., Shafto, P., & Tenenbaum, J. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, 64(1), 35-73.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.
- Rips, L. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, 34(2), 175-221.
- Sloman, S., & Lagnado, D. (2005). Do we 'do'? *Cognitive Science*, 29(1), 5-39.

<sup>2</sup>Minimality of breaks is determined by set inclusion rather than counts—see Rips (2010) and Hiddleston (2005) for details.

# Superspace extrapolation reveals inductive biases in function learning

**Christopher G. Lucas**

cglucas@cmu.edu  
Department of Psychology  
Carnegie Mellon University

**Douglas Sterling**

douglas.sterling@bccn-berlin.de  
Bernstein Center for Computational  
Neuroscience, Berlin

**Charles Kemp**

ckemp@cmu.edu  
Department of Psychology  
Carnegie Mellon University

## Abstract

We introduce a new approach for exploring how humans learn and represent functional relationships based on limited observations. We focus on a problem called *superspace extrapolation*, where learners observe training examples drawn from an  $n$ -dimensional space and must extrapolate to an  $n + 1$ -dimensional superspace of the training examples. Many existing psychological models predict that superspace extrapolation should be fundamentally underdetermined, but we show that humans are able to extrapolate both linear and non-linear functions under these conditions. We also show that a Bayesian model can account for our results given a hypothesis space that includes families of simple functional relationships.

## Introduction

People regularly face situations where they must reason about functions defined over continuous variables. For example, consider a truck driver who wants to predict how quickly his truck can accelerate based on the mass of his cargo. If the driver has transported similar masses in the past, he can generate an accurate prediction by recalling the accelerations observed on these previous occasions. The real test of whether and how he has learned the function is how he extrapolates from past examples and makes predictions about loads that are much lighter or heavier than those he has seen in the past. Figure 1a, for example, shows how a learner might use linear extrapolation to generalize on the basis of two examples.

In any function learning setting, extrapolation judgments are shaped by the examples observed and the assumptions or *inductive bias* that the human brings to the problem. Minimizing the information carried by the training examples makes the role of the inductive bias especially apparent. Here we explore how humans learn functions from impoverished training data, and focus in particular on the problem of *superspace extrapolation*. Given training examples that fall within an  $n$ -dimensional space, we study how learners extrapolate to an  $n + 1$ -dimensional superspace that encloses the training examples. If the underlying function is one-dimensional, superspace extrapolation requires the learner to generalize on the basis of a single training example (Figure 1b). We focus on the corresponding problem in two dimensions, where the learner observes training examples drawn from a one-dimensional space and must generalize to the full two-dimensional space (Figures 1c-f).

Superspace extrapolation is an interesting problem in its own right, but also provides a way to distinguish between competing accounts of function learning. The psychological literature on function learning includes two prominent approaches that we will call the rule-based approach and the

similarity-based approach. The rule-based approach proposes that humans rely on a set of parametric functions that have explicit mental representations, including linear functions, polynomial functions, and others (Carroll, 1963; Brehmer, 1974; Koh & Meyer, 1991; Koh, 1993; Bott & Heit, 2004). The similarity-based approach proposes that humans remember specific examples encountered during training, and make predictions about test points based on similarity to the training points (Busemeyer, Myung, & McDaniel, 1993; Kelley & Busemeyer, 2008). Similarity-based approaches have traditionally struggled to account for extrapolation, and superspace extrapolation is especially challenging for these approaches. We show that humans are able to learn several different functions in a superspace extrapolation paradigm, which supports the idea that people can formulate and use explicit representations of both linear and nonlinear functions.

The hybrid approach to function learning proposes that humans can make both rule-based and similarity-based inferences. We show that this approach can account for our data by evaluating a hybrid model that builds on the Gaussian process account of Griffiths, Lucas, Williams, and Kalish (2009). Other models of function learning are prominent in the literature, and here we mention two representative examples. The Population of Linear Experts (POLE) model (Kalish, Lewandowsky, & Kruschke, 2004) proposes that humans learn functions that are piecewise linear (in the 1D case) or piecewise planar (in the 2D case). Since the training examples in a superspace extrapolation task are collinear, any given piecewise planar function belongs to an infinite family of piecewise planar functions that make very different extrapolation predictions but fit the training examples equally well. For example, Figures 1c and 1d show two different extrapolation functions that account perfectly for the same set of training examples. As a result, models that rely exclusively on linear extrapolation suggest that superspace extrapolation problems are fundamentally underdetermined and are unlikely to lead to consistent patterns of human responses. The Sigma model (Juslin, Karlsson, & Olsson, 2008) is an alternative approach which proposes that humans can acquire explicit representations of linear functions, but that knowledge about non-linear functions is “carried implicitly by memory for exemplars.” We show that people are successfully able to extrapolate non-linear functions in a superspace extrapolation paradigm, which suggests that the rule-based component of a hybrid approach should include room for non-linear functions.

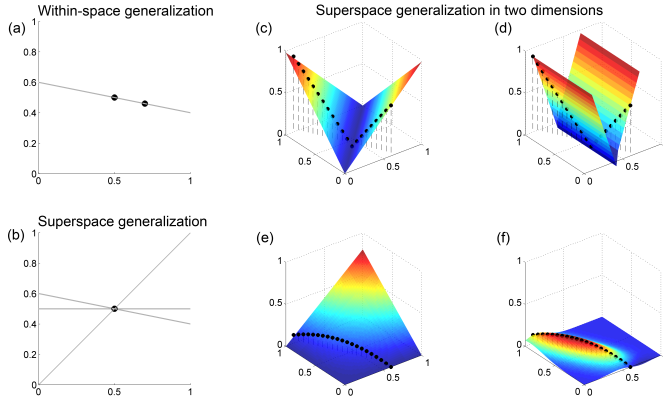


Figure 1: Examples of superspace extrapolation in one and two dimensions. Black dots are training points. (a) Standard function learning with one cue dimension; (b) Extrapolating from cues in a zero-dimensional subspace; (c) Superspace extrapolation in two dimensions, where  $f(x, y) = |x - y|$ ; (d) A second example of superspace extrapolation applied to  $|x - y|$ , assuming a difference piecewise linear function; (e) Superspace extrapolation where  $f(x, y) = xy$ ; (f) A second example using  $xy$ , with similarity-based extrapolation.

## Experiment

We developed a behavioral experiment with two goals in mind. The first and most basic goal is to find out whether superspace extrapolation is possible at all. Expecting participants to make generalizations about a function given a single training point seems unreasonable (Figure 1b), and it is possible that participants will find the two dimensional version of superspace extrapolation equally underdetermined. If superspace extrapolation turns out to be possible, our second goal is to understand how this kind of extrapolation is achieved. In particular, we aimed for a task that could address whether participants use explicit rules to make inferences that go beyond linear and similarity-based extrapolation.

We hypothesized that participants could learn a range of two-dimensional functions and chose to focus on five specific functions that are relatively simple and qualitatively different from one another. These functions are shown in Table 1 and plotted in Figure 2. Note that the family of functions includes both linear and non-linear functions. Consider one such function, the absolute difference function plotted in Figure 1c. Suppose that a learner observes the training points shown in black, which happen to fall along a line. There are many possible ways to extrapolate from the training points to the entire space—for example, Figure 1d shows an extrapolation to an axis-aligned function that is especially simple in the sense that it is invariant with respect to one of the dimensions. If people extrapolate by fitting a piecewise linear (i.e. planar) function to the training points, then there seems to be no reason to prefer the extrapolation in Figure 1c to Figure 1d or the infinitely many alternative extrapolations that fit two planes to the training points. On the other hand, if  $|x - y|$  is in human

learners' representational toolkit, we might predict that their extrapolations would resemble Figure 1c.

If extrapolation in cases like Figure 1c does depend on explicit rules, then extrapolations might vary dramatically if the positions of the training points are rotated. For example, the function  $f(x, y) = |x - y|$  turns into the function  $|x - 1 + y|$  when rotated by  $\pi/2$  around the line  $(0.5, 0.5, t)$  which passes through  $(0.5, 0.5)$  in the  $xy$ -plane and is perpendicular to the  $z$ -axis. It seems plausible that participants rely on a hypothesis space of rules that can accommodate the original but not the rotated function. We therefore compare each extrapolation problem to a rotated variant where the training points are rotated around the line  $(0.5, 0.5, t)$ , and predict that participants will be able to learn the unrotated but not the rotated version of each function. Linear extrapolation is equally possible in the rotated and unrotated cases, and a linear extrapolation account therefore predicts no qualitative difference between these two versions of the problem. Many similarity-based approaches also predict that the rotated and unrotated versions should lead to similar results, since similarity metrics (e.g. Euclidean distance) are often rotation-invariant.

## Methods

**Participants.** 33 participants were recruited from Carnegie Mellon's participant pool and the local community and received course credit or ten dollars for participating.

**Materials.** Cues were presented using adjacent horizontal bars and participants made predictions by adjusting a third horizontal bar centered under the midpoint between the cue bars. Each bar had a bounding box, so the range of valid values—which we denote with  $[0, 1]$  for simplicity—was evident to participants. No numerical information was provided about any of the variables. Feedback presentations took the form of a green bar overlaid on the prediction bar.

**Procedure.** Participants were told that they would have to learn several cause-effect relationships through trial-and-error. Each participant was presented with the five distinct functions listed in Table 1 in random order, in either *rotated* or *unrotated* form. For a given unrotated function  $f(x, y)$  and a rotation angle  $\theta_r$ , we define a rotated function  $g(x, y) = f(x', y')$  where  $(x', y')$  is the result of rotating  $(x, y)$  around the point  $(0.5, 0.5)$  by  $\theta_r$ . Table 1 contains explicit definitions of the unrotated and rotated versions of all functions. For each function, participants saw a training phase followed by a test phase. Both phases consisted of a series of trials in which participants were presented with cues  $(x, y)$  and asked to predict  $f(x, y)$ .

The training phase included 40 randomly-ordered examples that fell along a single line. Specifically, training examples fell at equal intervals along a line segment with length 0.9 centered at  $(0.5, 0.5)$ , making an angle of  $\theta_t$  (see Table 1) with the  $x$ -axis. After each training prediction, participants who gave guesses within 0.04 of the true value moved to the next example point, while inaccurate guesses were followed by feedback in which the correct value of  $f(x, y)$  was presented and participants had to adjust their prediction to match

Name	Unrot. $f(\cdot)$	$\theta_r$	Rot. $f(\cdot)$	$\theta_l$
Projection	$x$	$\frac{1}{2}\pi$	$1 - y$	$\frac{1}{8}\pi$
Average	$\frac{1}{2}(x + y)$	$-\frac{1}{2}\pi$	$\frac{1}{2}(x + 1 - y)$	$\frac{3}{8}\pi$
Product	$xy$	$\frac{1}{2}\pi$	$x(1 - y)$	$\frac{5}{8}\pi$
Difference	$ x - y $	$\frac{1}{2}\pi$	$ x + y - 1 $	$\frac{5}{8}\pi$
Max	$\max(x, y)$	$-\frac{1}{2}\pi$	$\max(x, 1 - y)$	$\frac{7}{8}\pi$

Table 1: List of functions that participants learned.  $\theta_r$  refers to the relative angles of the original and rotated functions, and  $\theta_l$  denotes the angle of the original line.

that value in order to continue.

In the subsequent test phase, participants received no feedback and were presented with 10 equidistant points along the original training line, 10 within-space points that fell beyond the extrema of the original line, and 36 superspace extrapolation points in a uniform 6-by-6 grid over the  $[0, 1] \times [0, 1]$  range. After each test phase, participants were prompted to describe what they thought the function was before moving on to the next function. The bars corresponding to variables  $x$  and  $y$  were selected randomly.

## Experimental Results

We excluded one participant who did not attempt to learn the functions, indicated by a mean absolute error exceeding 0.25. Ten of the 32 participants who remained did not complete all of the functions in the allotted hour, but each version (rotated or unrotated) of each function was completed by at least 11 participants. The side on which  $x$  and  $y$  were presented had no significant influence on performance, so the two orientations were grouped together.

The five panels labeled (iii) in Figure 2 show average human responses for the five unrotated functions. The black dots show responses for the training points, and the surfaces show responses for the extrapolation points. In all cases, participants were able to learn the function values for the training points, and their extrapolation judgments were qualitatively similar in all cases to the true functions. Note that superspace extrapolation was possible even for the three non-linear functions in the set. The panels labeled (iv) in Figure 2 show average responses for the rotated functions. Participants appeared to learn the rotated projection function, but extrapolation judgments for the four other rotated functions appear qualitatively different from the true functions. Table 1 shows that the rotated version of the projection function is  $f(x, y) = 1 - y$ . Recall that the cues were presented using sliders on horizontal bars, and that the value of each cue corresponds visually to the proportion of the bar to the left of the slider. The rotated projection function can be learned by paying attention to the complement of the  $y$  cue, or the proportion of the  $y$ -bar to the right of the slider. Although responses for the rotated projection function suggested that participants

Function	$\text{MAE}_u$	$\text{MAE}_r$	p
$x$	0.039	0.055	0.51
$(x + y)/2$	0.071	0.151	0.00088
$xy$	0.092	0.140	0.042
$ x - y $	0.123	0.247	0.0015
$\max(x, y)$	0.046	0.130	0.0077

Table 2: Mean absolute error for test points in learning rotated  $\text{MAE}_r$  and unrotated functions ( $\text{MAE}_u$ ). p-values were obtained using a two-tailed permutation test, using 200,000 samples per test.

are sensitive to complements in some cases, responses for the remaining rotated functions suggest that participants find it difficult to learn simple functions defined in terms of complements.

The descriptions provided by individual participants indicated that many had acquired explicit representations of the unrotated functions. Five examples of these descriptions are: “effect was identical to cause B” (projection); “roughly the average of the two causes” (average); “fraction multiplication” (product); “difference of the causes” (difference); “the larger of the two values” (max). Responses for the rotated functions sometimes indicated complex hypotheses, but more often indicated confusion or uncertainty about the nature of the function. These descriptions indicate that some individuals had clearly learned the functions. To further explore responses at the individual level, we looked at the extent to which individual participants’ judgments fit the true functions versus several alternatives, for both unrotated and rotated functions, shown in Figure 3. The space of candidate functions included all true unrotated and rotated functions, along with a set of simple alternatives shown in the caption to Figure 3. The alternatives include a function that captures complete uncertainty ( $f(x, y) = 0.5$ ), floor and ceiling responses ( $f(x, y) = 1$  and  $f(x, y) = 0$ ), and some simple linear combinations of  $x$  and  $y$ . For all of the unrotated functions, most participants’ extrapolation judgments were best fit by the true function. For the rotated versions, the modal judgment only matched the true function for rotations of  $f(x, y) = x$  and  $f(x, y) = \max(x, y)$ .

Additional evidence that individual participants often learned the true functions relatively well is provided by examining the mean absolute error with respect to the true function. Performance for the projection function was near-ceiling for both unrotated and rotated versions, but in all other cases participants had lower mean absolute error for the unrotated functions than the rotated functions. Table 2 shows mean absolute error for the extrapolation points, and a similar pattern held for the training points, with significantly better performance in the unrotated cases for  $f(x, y) \in \{x, |x - y|, xy, \max(x, y)\}$  at  $\alpha = 0.05$ .

Previous experiments have explored the relative learnability of one-dimensional functions, and our results provide some initial evidence about a learnability ordering for two dimensional functions. The results in Table 2 suggest that the

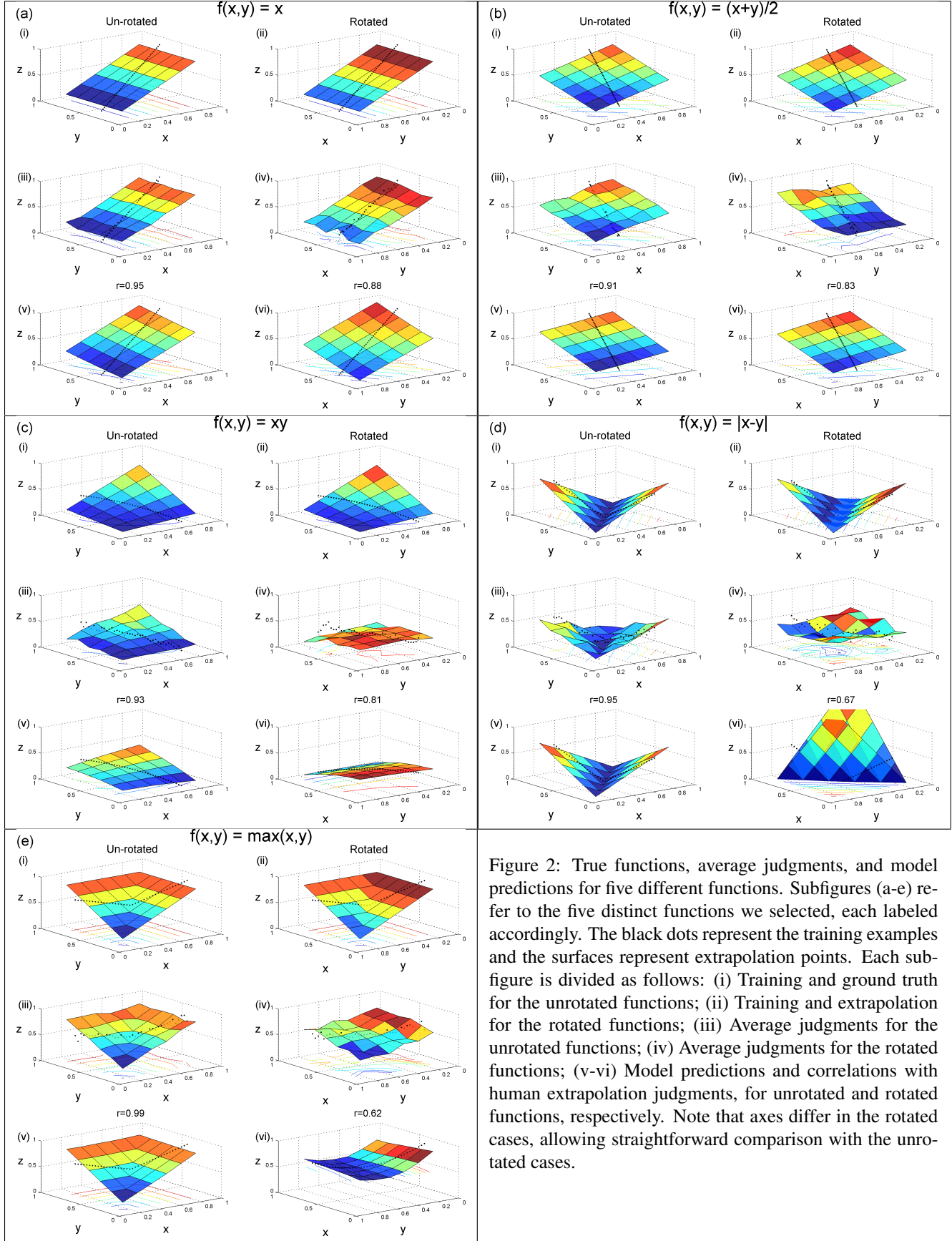


Figure 2: True functions, average judgments, and model predictions for five different functions. Subfigures (a-e) refer to the five distinct functions we selected, each labeled accordingly. The black dots represent the training examples and the surfaces represent extrapolation points. Each subfigure is divided as follows: (i) Training and ground truth for the unrotated functions; (ii) Training and extrapolation for the rotated functions; (iii) Average judgments for the unrotated functions; (iv) Average judgments for the rotated functions; (v-vi) Model predictions and correlations with human extrapolation judgments, for unrotated and rotated functions, respectively. Note that axes differ in the rotated cases, allowing straightforward comparison with the unrotated cases.



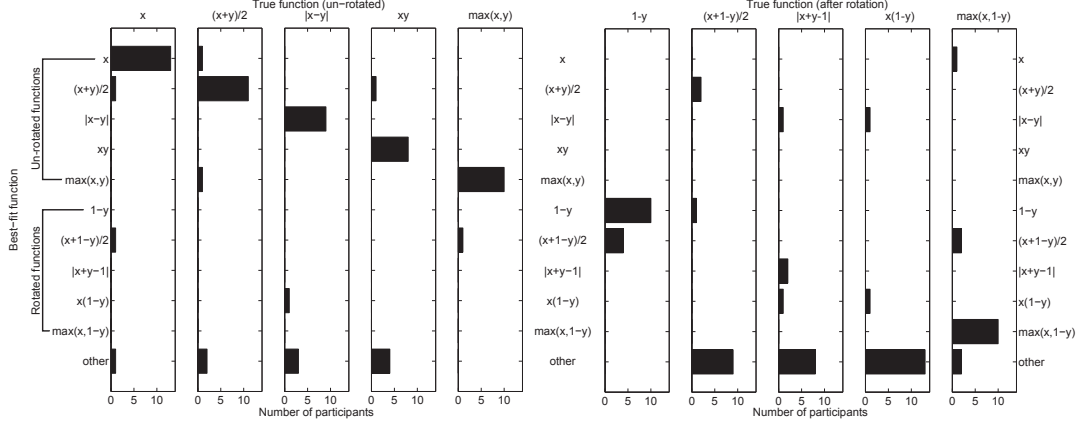


Figure 3: Trained functions versus functions that best fit participants’ judgments. The *other* group includes  $f(x, y) \in \{1, 0.5, 0, 0.5x, 0.5y, 0.75x + 0.25y, 0.25x + 0.75y\}$ . Best-fitting functions were those that minimized mean squared error.

projection and maximum functions are easiest to learn, and that the product and difference functions are most difficult to learn.

Taken together, our data provide strong evidence that humans are capable of superspace extrapolation, and can learn both linear and non-linear functions under this paradigm. The lower performance for the rotated functions suggests that the space of learnable functions is restricted. Both results are compatible with the idea that people can acquire explicit representations of rules, but raise challenges for approaches that focus on similarity-based computations alone. Our data, however, are compatible with a hybrid approach, and the next section describes one such approach that accounts for our data relatively well.

### Modeling superspace extrapolation

The hybrid approach to function learning is motivated by the idea that humans readily learn certain rules but fall back on similarity-based computations when no simple rule is consistent with the observed examples. The rule-based component of this approach can potentially explain how humans carry out superspace extrapolation when learning the unrotated functions in our experiment, and the similarity-based component may help to explain responses for the rotated functions.

To demonstrate that the hybrid approach can account for our data, we developed a computational model which assumes that humans make use of a hypothesis space that contains several families of functions. Some of these function families correspond to simple rules, and others are more generic and include all smooth functions. The first column of Table 3 shows one such hypothesis space that includes linear, quadratic, difference, maximum and product functions, along with one generic family of smooth functions. Given this hypothesis space and a set of training examples, extrapolation judgments can be made by using the posterior distribution over the space of functions.

The model we implemented builds on the hybrid approach of Griffiths et al. (2009), which uses Gaussian processes to

Function family	Prior	Mean function
$\beta[1 \ x \ y]^T$	0.5	$\mu_0 = 0, \mu_1 = \mu_2 = 1$
$\beta[1 \ x \ y]^T$	0.4	$\mu_0 = 1, \mu_1 = \mu_2 = -1$
$\beta[1 \ x \ y \ x^2 \ y^2]^T$	0.09	$\mu_0 = \mu_1 = \mu_2 = \mu_3 = 0$
$\beta_1  x - y $	0.001	$\mu_1 = 1$
$\beta_1 \max(x, y)$	0.001	$\mu_1 = 1$
$\beta_1 xy$	0.001	$\mu_1 = 1$
Smooth functions	0.01	$f(x, y) = 0$

Table 3: Hypothesis space captured by the Gaussian process model. Un-normalized prior probabilities are given for each function family for readability. For the first five families, coefficients  $\beta_i$  are distributed normally around  $\mu_i$  with a common variance for each coefficient. The difference, maximum, and product families are not described by Griffiths et al. (2009), but the prior probabilities on all remaining families and the  $\mu_i$  values for these families are drawn from Griffiths et al. (2009).

capture both rule-based and similarity-based function learning. As originally presented, this model takes kernel functions that express linear and quadratic rules as well as a standard similarity-based kernel function for which the covariance between any two points  $\mathbf{x}$  and  $\mathbf{x}'$  is  $K(\mathbf{x}, \mathbf{x}') = \theta_1 \exp(-\frac{1}{\theta_2^2} \|\mathbf{x} - \mathbf{x}'\|^2)$ , where  $\theta_1$  and  $\theta_2$  determine the smoothness of the function. Intuitively, this last kernel expresses the assumption that functions are locally smooth, and was used to produce the extrapolations in Figure 1f. The model generates predictions by integrating over all possible functions for all function types, integrating out all applicable parameters. For a more detailed description, see Griffiths et al. (2009).

Our extension to the original model was to add a kernel capturing each of the three non-linear rules in our experiment, which are equivalent to Bayesian regression models of the form  $\beta|x - y|$ ,  $\beta\max(x, y)$ , and  $\beta xy$ , where  $\beta$  is a coefficient distributed normally around one. We assigned each new kernel a prior probability of 0.001, or one tenth that of the least-



probable kernel in the original model, before renormalizing kernel probabilities. See Table 3 for a summary of all of the kernels in the model and their corresponding probabilities.

Model predictions are shown in Figures 2e and f, along with correlations with human extrapolation judgments. In most cases the predictions of this model closely matched participants' superspace extrapolations for both unrotated and rotated functions. The latter result is the more striking of the two, as the predictions arise from averaging over several kernels rather than choosing suitable ones in advance.

The one major discrepancy between model predictions and human judgments occurs for the rotated version of  $|x - y|$ , where the extrapolation judgments predicted by the model are substantially more extreme than the human responses. This result is driven by the fact that the family of difference functions in Table 3 can perfectly account for the rotated training points if  $\beta_1$  takes a value larger than 1. Unlike the model, humans may be unable to learn weighted versions of the difference function in Table 3, which could be captured by setting the coefficient  $\beta_1$  for this family to 1. The model represents the simplest possible extension of the Gaussian process account of Griffiths et al. (2009), but adjusting the priors on the coefficients may result in a more accurate model of human learning.

### Alternative models

Several recent models that address extrapolation in function learning and multiple cue judgment, including POLE (Kalish et al., 2004), EXAM (DeLosh, Busemeyer, & McDaniel, 1997), and Sigma (Juslin et al., 2008), all suggest that humans extrapolate according to linear functions. In their present forms, none of these models appear to account for our results. We fit the POLE model to our data and found that extrapolations were consistently piecewise linear in one cue dimension, and invariant to the other, taking a form like that in Figure 1c. This approach to superspace extrapolation seems plausible *a priori* but it does not reflect the behavior of our participants. The EXAM model makes extrapolation predictions using the nearest past examples to a new point, implying that peoples' judgments are invariant to the rotation of a given function, which is inconsistent with our data. Finally, the Sigma model (Juslin et al., 2008) proposes that humans can acquire explicit representations of linear functions but that extrapolation of non-linear functions relies on similarity-based generalization. The Sigma model is therefore inconsistent with our finding that people were able to learn several non-linear functions.

### Conclusion

We introduced the problem of superspace extrapolation, which provides a new way to explore the inductive biases that people bring to the task of function learning. Our data suggest that these inductive biases include a toolkit of linear and non-linear rules that can be compared against the available data. Our results challenge several popular accounts of function learning, but we showed that they are compatible with a

hybrid approach to function learning that accommodates both explicit rules and similarity-based inferences.

Superspace extrapolation requires learners to go beyond the available data in a fundamental way, and other problems where humans make inferences based on limited data have also provided important evidence about human inductive biases (Shepard, 1994; Feldman, 1997). Psychologists sometimes study what can be learned from textual corpora and other massive data sets, but exploring what humans learn from highly constrained data sets can be equally valuable.

**Acknowledgments.** This work was supported by the James S. McDonnell Foundation Causal Learning Collaborative Initiative and by NSF award CDI-0835797.

### References

- Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(1).
- Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Decision Processes*, 11, 1-27.
- Busemeyer, J., Myung, I. J., & McDaniel, M. (1993). Cue competition effects: Theoretical implications for adaptive network learning models. *Psychological Science*, 4(3), 196.
- Carroll, J. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua.
- DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non of abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 968-986.
- Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology*, 41, 145-170.
- Griffiths, T. L., Lucas, C. G., Williams, J. J., & Kalish, M. L. (2009). Modeling human function learning with Gaussian processes. *Advances in Neural Information Processing Systems*, 21.
- Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, 106(1), 259-298.
- Kalish, M., Lewandowsky, S., & Kruschke, J. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, 111, 1072-1099.
- Kelley, H., & Busemeyer, J. (2008). A comparison of models for learning how to dynamically integrate multiple cues in order to forecast continuous criteria. *Journal of Mathematical Psychology*, 52(4), 218-240.
- Koh, K. (1993). Induction of combination rules in two-dimensional function learning. *Memory & cognition*, 21(5), 573-590.
- Koh, K., & Meyer, D. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(5), 811-836.
- Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin and Review*, 1, 2-29.

# Does the utility of information influence sampling behavior?

Doug Markant (doug.markant@nyu.edu)  
Todd M. Gureckis (todd.gureckis@nyu.edu)  
New York University  
Department of Psychology  
6 Washington Place, New York, NY 10003 USA

## Abstract

A critical aspect of human cognition is the ability to actively query the environment for information. One important (but often overlooked) factor in the decision to gather information is the cost associated with accessing different sources of information. Using a simple sequential information search task, we explore the degree to which human learners are sensitive to variations in the amount of utility related to different potential observations. Across two experiments we find greater support for the idea that people gather information to reduce their uncertainty about the current state of the environment (a “disinterested”, or cost-insensitive, sampling strategy). Implications for theories of rational information collection are discussed.

**Keywords:** information sampling, active learning, information access costs

## Introduction

From controlling the movement of our eyes to determining which sources of news to consult, judging the quality of alternative sources of information is a critical part of adaptive behavior. Research exploring how people make information gathering (or “sampling”) decisions has shown that people can discern subtle differences in the potential information value of various aspects of the environment. For example, measurements of eye movements during object categorization show that people preferentially allocate attention to object features that are most useful for making subsequent classification decisions (c.f., Nelson & Cottrell, 2007; Rehder & Hoffman, 2005).

One aspect that typically complicates the analysis of information sampling behavior is that information rarely comes for free. All natural tasks involve information access costs, even if the only cost is the time required to gather information (Fu, 2011). In addition, different pieces of information may be more useful depending on how one will be tested in the future. Optimal search behavior must weigh the costs of collecting particular bits of information against the benefit it is expected to convey (Edwards, 1965; Juni, Gureckis, & Maloney, 2011; Tversky & Edwards, 1966), a point frequently made in research on animal foraging (Stephens & Krebs, 1986).

Despite its importance, previous work on information sampling has often failed to test whether people take into account costs related to different sources of information. For example, Nelson (2005) provides a comprehensive review of various ways an optimal Bayesian agent might value potential information sources in the absence of task-specific costs (see also Nelson et al., 2010). One conclusion from this line of work is that people make information search decisions that are consistent with normative measures of information value (many of which often make similar predictions). For example, Nelson et al. (2010) studied information sampling in a diagnostic

reasoning task where the predictions of these measures could be readily distinguished. Learners who could query different stimulus features before making classification decisions were found to prefer to learn about features that maximized *probability gain*, a measure of how well a potential observation is expected to improve classification accuracy.

In these studies, however, costs were not explicitly manipulated or controlled. Taking costs into account can alter the optimal strategy in a given task, but it is unclear whether people adjust their behavior in a similar way. The goal of the present paper is to explore the impact of costs on sampling decisions. We begin by evaluating two alternative objectives that people may adopt when deciding what information to gather. Like the models reviewed by Nelson (2005), the first ignores the implications of task-specific costs and casts information sampling strictly in terms of uncertainty reduction (i.e., information gain). The second approach balances the costs and expected benefits of information in the context of the task. We then describe the results of two experiments that manipulate the concordance between these two approaches, in one case creating an environment where the goals of minimizing uncertainty and maximizing utility predict different patterns of information sampling. Our results show that people tend to value information (in terms of the number of hypotheses ruled out by a new observation) over situation-specific costs and benefits. The implications of these results for theories of information sampling are discussed.

## The rational analysis of information sampling: comparing “interested” and “disinterested” search

How should a rational agent make information sampling decisions? Existing proposals fall into two broad categories which, borrowing from Chater, Crocker, and Pickering (1998), we call “interested” and “disinterested.” Unlike the distinctions explored by Nelson (2005), these two proposals differ significantly in terms of the overall goal of information sampling.

**Interested (or cost sensitive) sampling** The first approach represents a decision-theoretic approach to information sampling. In particular, the agent considers the cost for collecting a piece of evidence and weighs this against the expected benefit it should convey with respect to the goals of the task. For example, a car shopper might decide if the possible savings available from obtaining information contained in a vehicle history report is worth the cost of the report. Similarly, preferentially fixating the features of an object that are diagnostic of its category membership may be a cost-sensitive strategy under the

assumption that additional fixations cost time and the number of fixations needed to reach a correct decision should be minimized. Sampling in this case is “interested” in that information acquisition is focused on some purpose or goal beside acquiring the information itself. In many ways, “interested” sampling is a fully rational strategy and this formulation is often adopted in economics research.

**Disinterested (or cost insensitive) sampling** The second approach values information to the degree that it reduces our uncertainty about the world. Chater et al. liken this to basic research where the goal is learning without regard for the ultimate utility of this knowledge for society. In their words, “inquiry is valuable for its own sake, because it leads to knowledge” (Chater et al., p.4). Disinterested inquiry can be conveniently expressed as actions which have a high probability of reducing the Shannon entropy over the agent’s beliefs (Lindley, 1956; Mackay, 1992). Critically, disinterested inquiry doesn’t depend on the costs associated with collecting information or using it to make subsequent decisions.

The basic premise of the experiments reported in this paper is that these two strategies or ways of valuing information can be dissociated on the basis of observed choice behavior. We gave participants a simple, intuitive information search task where they were asked to make sequences of decisions to reduce their uncertainty about a hidden target. Mathematical models instantiating each of the two theories just described are then fit to the choice patterns of individual subjects. This fitting procedure (common in the reinforcement learning literature) allows us to evaluate which of the approaches we have described gives the best account of participants’ choices.

### The Experimental Task: The Shape Search Game

Participants in our task are presented with a 10 by 10 grid that contains non-overlapping hidden shapes made up of individual grid cells. The hidden targets are randomly drawn from a set that is known to the participant. There are two phases in each game: a *sampling phase* and a *painting phase*. In the sampling phase, the player learns about the form and location of the hidden shapes by choosing squares in the grid to uncover. On each trial, they make an observation at one location, revealing either part of a hidden shape or an empty square. When they think they know the location and form of all the shapes they can stop sampling and enter the painting phase. In the painting phase, the player is tested for their knowledge of the shapes by “painting” any remaining squares they believe belong to one of the shapes in the appropriate color.

The player is penalized one point for every observation made in the sampling phase and two points for every error committed in the painting phase (e.g., failing to fill in a square that belongs to a shape). These costs promote efficient information search in two ways. First, the observation cost discourages sampling in locations whose contents can be inferred from evidence that has already been uncovered. Second, it encourages continued sampling while there is still uncertainty about the hidden shapes, since painting errors are more costly

### Experiment 1: Rectangle Search

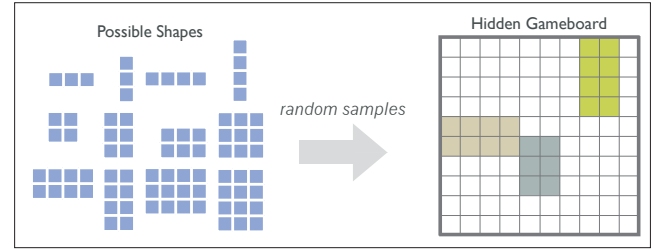


Figure 1: The generative process underlying the shape search game. A fixed set of possible shapes is specified. A hidden gameboard is created by sampling from this set of shapes and randomly arranging the targets in the grid. During the sampling phase the participants clicks on grid locations to reveal their contents. In the painting phase, the subject draws in the remaining squares using the mouse and is rewarded for accuracy.

than observations. The goal of the game is to finish with the lowest score possible, which is achieved by learning the most about the hidden shapes in the fewest number of observations.

Based on past work with this task (Gureckis & Markant, 2009), we have found that the overall objective of the game is easily understood by the participants. A critical feature of the task (which we exploit in our second experiment) is that it allows for arbitrarily defined targets (i.e., the target shapes may be composed of any configuration of squares) that can be manipulated to vary the complexity of the task.

### Formal task analysis

In order to evaluate both “interested” and “disinterested” information search in the task, we compare the search behavior of subjects to that of a rational learner who updates their beliefs about the gameboard in an optimal way. Formally, players make a sequence of observations in order to learn the hidden gameboard,  $g_h \in G$ , where  $G$  is the universe of legal gameboards. Each individual gameboard is defined by an arrangement of  $N$  non-overlapping shapes  $\{r_1, r_2, \dots, r_N\}$  with unique labels  $\{l_1, l_2, \dots, l_N\}$ , and each shape consists of a set of squares such that  $o_{ij} \in r_n$  (where  $i$  and  $j$  index the  $x$ - and  $y$ -coordinates of the square).

On each trial the player makes an observation  $o_{ij}$  and receives feedback in the form of a label  $l_n$ , where  $l_0$  indicates the observed square is empty,  $l_1$  means it contains part of shape  $r_1$ , and so on. Since each square in the grid is deterministically assigned to either one or zero shapes, we assume that the likelihood of a particular observation  $o_{ij}$  belonging to one of the shapes (i.e.,  $o_{ij} = l_n$  for  $n > 0$ ) for a particular gameboard  $g$  is deterministic (i.e.,  $p(o_{ij} = l_n | g) = 1$  if the location falls within a target and 0 otherwise).

The prior belief about the likelihood of each individual gameboard is represented by  $p(g)$ . In our experiments, participants were instructed that the shapes were chosen at random and that all legal gameboards were equally likely (i.e.,  $p(g) = 1/|G|$  for all  $g$ , a uniform prior).

Bayes rule can be used to compute the posterior belief

about the identity of the hidden gameboard and to predict the marginal probability of any point in the grid having any particular label  $l_n$  (this is a very straightforward Bayesian approach to the problem, see Gureckis and Markant, 2009).

**Interested (cost-sensitive) Sampling** The objective of the game is to minimize the number of points accumulated, where each individual observation costs  $C_{obs}$  points and each error during painting costs  $C_{error}$  points. Given these constraints, we can quantify the value of observations with respect to the overall goal of minimizing total costs. We assume that the likelihood of labeling a point  $o_{ij}$  with label  $l_n$  during the painting phase is simply the marginal probability  $p(o_{ij} = l_n | \mathcal{B})$ , and the cost associated with that action is tied to the uncertainty about its label when the sampling phase ends (e.g., if  $p(o_{ij} = l_n | \mathcal{B}) = 1$ , the true label is known with certainty and there is no chance of committing an error during painting)<sup>1</sup>. On each trial, the total expected cost  $EC(\mathcal{B})$  of ending the sampling phase and entering the painting phase is defined as:

$$EC(\mathcal{B}) = C_{error} \cdot \sum_i \sum_j \sum_n p(o_{ij} = l_n | \mathcal{B}) \cdot [1 - p(o_{ij} = l_n | \mathcal{B})] \quad (1)$$

For a new observation and its observed outcome ( $o_{ij} = l_n$ ) we then calculate the resulting cost *savings*, or reduction in expected costs:

$$S(\mathcal{B}, o_{ij} = l_n) = EC(\mathcal{B}) - EC(\mathcal{B} | o_{ij} = l_n) \quad (2)$$

The savings achieved from feedback is offset by the cost of making the observation ( $C_{obs}$ ). To account for uncertainty about the true outcome we find the *expected savings* by weighting the savings for each outcome by its likelihood of occurring:

$$ES(\mathcal{B}, o_{ij}) = -C_{obs} + \sum_n p(o_{ij} = l_n | \mathcal{B}) S(\mathcal{B}, o_{ij} = l_n) \quad (3)$$

For each trial,  $ES(\mathcal{B}, o_{ij})$  is calculated for all  $o_{ij}$ , giving a distribution of the expected saving for remaining observations in the grid. An ideal learner maximizes this value by choosing the location  $o_{ij}$  with the highest  $ES(o_{ij})$  on each trial.

**Disinterested (cost insensitive) Sampling** A “disinterested” sampling norm values observations according to their effect on the learner’s beliefs without account for task-specific costs and benefits. This captures the intuition that observations that produce a large change in the agent’s beliefs tend to be more useful than observations that have little or no effect (i.e., nothing new is learned). In our approach this was modeled using *information gain*, which values an observations according to the expected reduction of uncertainty about the hidden gameboard. This uncertainty can be quantified by the Shannon entropy measured over the current belief distribution ( $H(\mathcal{B})$ ).

Entropy is maximized when all hypothesized gameboards are equally likely (as with our initial uniform prior), and minimized when there is only one possible hypothesis. For a given

<sup>1</sup>For shorthand,  $\mathcal{B}$  represents a vector of probabilities  $p(g | o_{ij} = l_n)$ , for all  $g \in G$ , that represents the full posterior distribution over the entire space of gameboards. This is the agents current “belief distribution” about which gameboard is the current target.

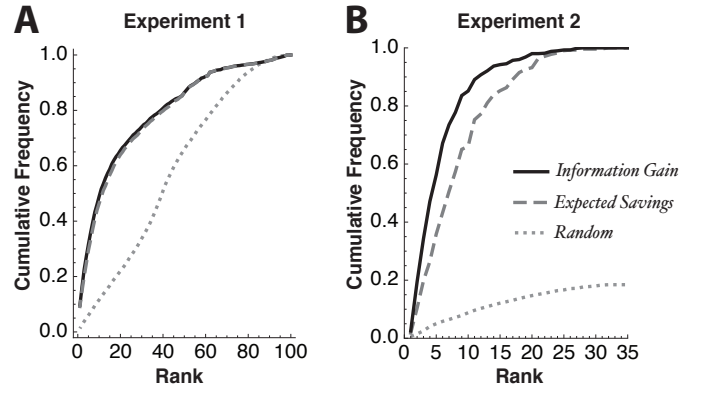


Figure 2: Cumulative frequency of ranks assigned to participants’ samples. A: In Experiment 1, the average rank of participants’ choices is higher than expected from a random sampling strategy, but there is no difference between rankings assigned by the two models. B: In Experiment 2, participants’ samples are more highly ranked according to EIG than ES.

observation and its observed outcome ( $o_{ij} = l_n$ ) we can calculate information gain as:

$$I(\mathcal{B}, o_{ij} = l_n) = H(\mathcal{B}) - H(\mathcal{B} | o_{ij} = l_n) \quad (4)$$

To account for uncertainty about the true outcome of an observation, information gain for each possible outcome is weighted by the predicted probability of that outcome occurring, giving the *expected information gain* for an observation  $o_{ij}$ :

$$EIG(\mathcal{B}, o_{ij}) = \sum_n p(o_{ij} = l_n | \mathcal{B}) I(\mathcal{B}, o_{ij} = l_n) \quad (5)$$

As above, on each trial we compute  $EIG(\mathcal{B}, o_{ij})$  for all locations in the grid, and assume that the optimal model chooses the location with the highest value on each trial<sup>2</sup>.

In applying each model to human choice data, the model is “yoked” to the decisions of the player. On each trial, the models assign a value (either  $EIG$  or  $ES$ ) to each point in the grid. These utilities can be used to compute choice probability of various grid locations. After revealing what the subject actually chose on a given trial, the model updates its posterior beliefs about the current gameboard configuration. These new beliefs then feed into new predictions about the utility of choosing each grid location. The process ends when the participant ends the game.

## Experiment 1: Rectangle Search

The first experiment re-analyses a previously published result which introduced explicit task-specific costs (Gureckis &

<sup>2</sup>It is important to note that this represents a “greedy” policy that chooses the best observation available on any given trial, but this may not reflect the globally optimal solution. The current framework could be extended to account for how participants might estimate the value of sequences of observations. However, due to the computational complexity of finding this solution given the large number of potential observations on any trial, for the present studies we focus our analysis on the greedy model.

Markant, 2009). Six participants played a series of games in which they searched for three rectangular shapes, randomly drawn from the set shown in Figure 1. The set of shapes was displayed on screen throughout the game. Participants were instructed that the three shapes in each gameboard were non-overlapping and were shown a large number of examples gameboard configurations prior to the experiment.

Each observation made by a participant during the sampling phase was ranked according to the predictions of both models (the median rank was used when multiple observations had equal value). Overall, the results show that people consistently sampled points that were assigned a high value by both models, with approximately 50% of their samples falling within the top 10 ranked observations available to them (see Figure 2A). In this experiment, however, the hypothesis set that was used (rectangular shapes of varying shape and size) led to highly similar predictions for both information gain and expected savings, precluding a test of whether people were sensitive to the costs in the task.

It is important to consider why the predictions of the two models converged in this case. As discussed previously, a cost-sensitive learner should value observations that have higher utility—that is, those that will reduce the likelihood of committing errors in the painting phase by the greatest amount. Intuitively, this implies that learning about bigger shapes is especially useful, since it will allow one to correctly label a greater number of squares. This idea is illustrated in Figure 3A for a simple hypothesis set made up of three rectangles in different locations. While observing a “hit” on any shape determines the true hypothesis (middle column), observing a “miss” (righthand column) has different utilities depending on the size of the shape it rules out. For example, ruling out the smallest shape (top row) leaves uncertainty about how to label eight other squares, whereas ruling out the largest shape (bottom row) leaves uncertainty about only four.

While the shape set used in Experiment 1 contained a range of sizes, the fact that the hypotheses were “nested” (i.e., the largest shapes overlapped with a set of progressively smaller ones) meant that learning about larger shapes also tended to rule out many hypotheses. As a result, the predicted choice values according to both models were highly similar. For our second experiment we created an alternative hypothesis space in which there were clearer differences in both the size of alternative hypotheses and the choices that were related to shapes of different size, leading to a greater number of potential observations where the predictions of the two models diverged.

## Experiment 2: Letter Search

In Experiment 2, we simplified the task to involve searching for a single target in the grid. For the hypothesis space we created a set of simple “letter” shapes (see Figure 3B). There were two types of games: L/D games, where the hidden letter could be an L or D, and C/U games, where the hidden letter could be a C or U. In each game a single point belonging to the hidden shape was revealed before the participant began sampling.

These modifications resulted in a much less complex hypothesis space (e.g., only 15 hypotheses were possible at the beginning of an L/D game, and 14 for C/U games). Importantly, because the two shapes involved in each type of game differed in area (for example, the ‘D’ shape contained a greater number of filled squares than the ‘L’), the predictions of information gain and expected savings diverged in the task.

## Methods

**Participants** Sixteen NYU undergraduates completed the study for course credit or for a \$10 payment. The experiment was presented on a standard Macintosh computer.

**Materials** A gameboard consisted of one of the four letter shapes seen in Figure 3 placed in any location on the grid. All possible gameboards were generated, resulting in 56 gameboards for each letter, or 112 for each game combination (L/D or C/U). For L/D games, 15 gameboards were randomly selected from the pool of L and D gameboards. This was repeated for C/U games. The resulting total of 30 unique games were used for all participants. The order of games played was randomized for each person.

**Procedure** Many aspects of the design were identical to Experiment 1 (described in Gureckis and Markant, 2009), so we highlight differences below. In Experiment 2 we sought to reduce any reliance on the visual display for recalling the specific shapes being used. Participants began the experiment with two training phases to memorize the four shapes. In the first, a letter cue (e.g., the character ‘L’ in a standard computer font) was presented at the top of the screen along with its corresponding shape, which appeared inside a 4x3 grid. The participant was asked to copy the same shape onto an empty 4x3 grid below. This was done twice for each letter (L, D, C, U) in randomized order. In the second training phase, they were presented only with the letter cue and an empty 4x3 grid, and asked to fill in the correct letter shape from memory. This was repeated three times for each letter in random order. In order to progress from one training trial to the next, the participant was required to successfully reproduce the correct shape. Training was followed by on-screen instructions which were modified to reflect the new hypothesis spaces.

Before the sampling phase began, a 2-second cue was displayed on the screen that indicated the type of game about to be played (the characters “L D” or “C U”). This cue was also displayed on the right side of the display during the game. This ensured that participants were aware of the shapes that were possible but that they had to use their memory of the actual shapes to guide their observations.

Sampling and painting phases proceeded in the same way as Experiment 1. The final score was then displayed, including how many points were the result of sampling and how many were the result of painting errors. The lowest score obtained by the participant in any game so far was shown to provide motivation and a means to evaluate their performance over time. Each participant played 30 games at their own pace, resulting in a total of 480 games collected.

## Results

**Sample rank** On each sampling trial, the ideal models were used to compute the expected information gain and expected savings for all remaining observations available. A participant’s decisions were ranked according to EIG and ES (if multiple observations had the same value, the median rank was used). The relative frequency of each sample rank was computed for each participant across all trials, and averaged across participants (see Figure 2B). The rank frequency shows that participants’ choices were more highly ranked on average according to EIG than ES. Participants’ samples were ranked within the top 5 observations according to EIG on approximately 57% of trials, whereas according to ES only 35% of samples fell in the same range.



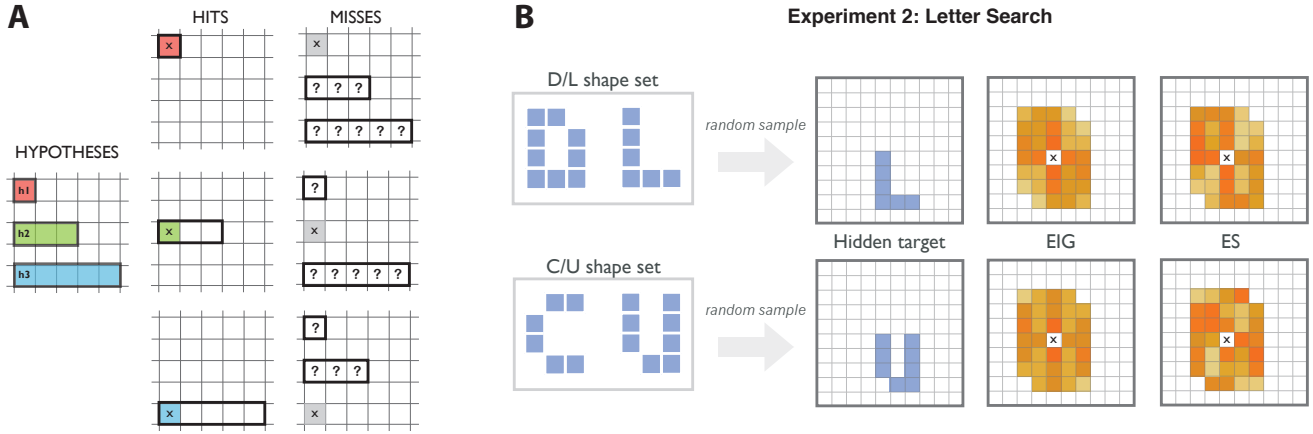


Figure 3: **A:** Illustration of the divergence between information gain and expected savings. **Left:** A hypothesis space is comprised of three possible rectangles,  $h_1$ ,  $h_2$ , and  $h_3$ . **Middle:** A hit on any one rectangle leads to the same number of hypotheses being ruled out, and no uncertainty about the label of any square in the grid. **Right:** A miss in any of the three locations rules out a single hypothesis, but the *predictive utility* of the sample differs based on its location. The labels for 8 squares are uncertain following a miss that rules out rectangle  $h_1$ , 6 squares following a miss ruling out  $h_2$ , and 4 squares following a miss ruling out  $h_3$ . **B:** Gameboard design and example model predictions in Experiment 2. Two types of games were possible: L/D games, where the hidden shape could be an L or D, and C/U games, where the hidden shape could be a C or U. Predicted value distributions for EIG and ES are shown for the first sampling trial in each kind of game, with a darker value indicating a higher value according to the model.

**Model fits** We next computed the likelihood of participants' decisions under the two alternative models. For each trial, the value of available observations was transformed into choice probabilities using the softmax function:

$$P(o_{ij}) = \frac{e^{\beta \cdot V(o_{ij})}}{\sum_{x,y} e^{\beta \cdot V(o_{xy})}} \quad (6)$$

The parameter  $\beta$  was fit on an individual basis for each model by maximizing the log-likelihood summed across all observations made by a participant. In all cases, EIG provided a better fit to participants' data than ES (Table 1).

**Stopping decisions** Our final analysis focused on participant's decisions to stop sampling. While EIG and ES make the same prediction as to when sampling should stop<sup>3</sup>, we were interested in whether people showed any sensitivity to the cost of collecting information. If people uncovered more squares than necessary it would suggest a failure to account for the cost of new observations (either in terms of ES or EIG). We classified each game according to whether the person decided to stop sampling before the trial predicted by the model ("early"), on the same trial ("optimal"), or after that trial ("late"). On average, participants ended sampling early ( $M = 0.46$ ,  $SD = 0.14$ ) or on the optimal trial ( $M = 0.50$ ,  $SD = 0.14$ ) on a similar proportion of games. In contrast, participants oversampled very rarely ( $M = 0.04$ ,  $SD = 0.01$ ).

<sup>3</sup>This convergence was due to the cost structure we used, in which the penalty for stopping before the hidden target was known was greater than the cost of an additional sample. Increasing the cost of sampling relative to the cost of errors would lead to ES predicting earlier stopping decisions than EIG.

## Discussion

The results of our second experiment show that people performed well compared to the ideal searcher model, frequently choosing highly ranked observations and consistently performing better than expected by a random search strategy. In addition, participants rarely gathered more information than necessary, which is consistent with prior work showing that people are sensitive to costs incurred by oversampling (Fu & Gray, 2006). Most importantly, we found that participants' sampling choices were better described by information gain than a cost-sensitive utility measure (expected savings).

Prior studies of human information collection have focused to a large extent on "disinterested" accounts of sampling decisions, showing that people are sensitive to the amount of information conveyed by different sources of information (Nelson, McKenzie, Cottrell, & Sejnowski, 2010). However, this line of work has often failed to consider whether people account for variations in task-specific utility when making sampling decisions. In our task, we introduced penalties for information access and uncertainty that altered the optimal sampling strategy. By manipulating the set of hypotheses in Exp. 2, we showed how sampling based on an information-maximizing strategy can be distinguished from cost-sensitive sampling. With respect to the distinction we began this paper with, our results suggest that people (at least in this task) prefer to gather information according to a "disinterested" measure of value.

Notably, differences in value between choices were not directly observable by the participant (as opposed to when some observations are more costly or difficult to make than others). Whether a given observation was valued differently according to EIG and ES depended upon the set of hypotheses remaining, and this divergence could change from trial to trial in response to new data. As a result, establishing which model pro-

vided a better account required fitting them to participants' decisions across a variety of choice contexts. This highlights a feature of our approach in that we could evaluate different sampling strategies using a set of highly variable choice sequences. Through the use of a well-defined hypothesis set and explicit cost structure, the "shape-search" task provides a useful framework for studying how information search decisions and task demands interact over the course of learning.

Of course, one potential caveat of the current study is that (due to computational demands) we evaluated a greedy decision policy such that the predicted value of a new observation does not take into account the consequence or utility of subsequent actions. It is possible that fully accounting for sequential dependencies in the search problem may alter the optimal utilities computed by the model. However, one might reasonably question if the computational demands of such a multi-step planning process are within reach of human reasoners. In addition, it is unclear that accounting for multi-step planning strategies would alter the choice utilities in a way that would bias for or against the results we report. Also note that comparing Exp. 1 and 2 illustrates that expected saving is not always at an advantage (i.e., in some choice environments the models become less distinguishable).

## Conclusion

So, why might human reasoners preferentially adopt "disinterested" sampling over "interested" sampling? One possibility is that sampling based on information gain (or other "disinterested" norms) may reflect a general purpose strategy that is useful in a variety of contexts. In particular, information gain can still be computed even when the cost of uncertainty (i.e., not knowing which hypothesis is true at the end of sampling) is difficult to predict. In addition, in many natural environments it may be consistent with the predictions of a cost-sensitive utility function, as illustrated by our first experiment. At the very least, our results highlight the need to understand the kinds of problems that lead people to adapt to task-specific costs in lieu of a general-purpose, "disinterested" approach to information search.

Table 1: Model fits

Subj	$\beta_{EIG}$	$\beta_{ES}$	-LLH(EIG)	-LLH(ES)
1	17.544	0.95	215.05	334.65
2	5.18	0.88	391.77	426.18
3	5.93	0.77	326.45	388.79
4	6.96	0.75	325.81	403.61
5	10.15	0.89	241.04	324.30
6	12.34	1.06	250.17	329.72
7	9.64	1.20	276.05	339.62
8	7.63	0.98	308.63	373.03
9	5.36	0.90	354.49	387.43
10	7.79	0.91	315.92	387.56
11	5.85	0.85	340.22	388.05
12	6.39	0.92	328.24	373.22
13	6.97	1.03	319.96	362.27
14	7.02	0.86	283.32	341.79
15	8.94	0.99	293.38	371.22
16	10.0	0.76	293.65	403.22

## Acknowledgments

The authors would like to thank Dan Navarro, Jonathan Nelson, and the other reviewers for their helpful comments. This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

## References

- Chater, N., Crocker, M., & Pickering, M. (1998). The rational analysis of inquiry: the case of parsing. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 441–468). Oxford University Press.
- Edwards, W. (1965). Optimal strategies for seeking information: models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology*, 2, 312–329.
- Fu, W., & Gray, W. (2006). Suboptimal tradeoffs in information seeking. *Cognitive Psychology*, 52(3), 195–242.
- Fu, W. (2011). A dynamic context model of interactive behavior. *Cognitive Science*, 35(5), 874–904.
- Gureckis, T., & Markant, D. (2009). Active learning strategies in a spatial concept learning game. In N. Taatgen & H. van Rijn (Eds.), *Proc of the 31st annual conference of the cognitive science society* (pp. 3145–3150). Cognitive Science Society. Austin, TX.
- Juni, M., Gureckis, T., & Maloney, L. (2011). Don't stop 'till you get enough: adaptive information sampling in a visuomotor estimation task. In L. Carlson, C. Hölscher & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society*. Cognitive Science Society. Austin, TX.
- Lindley, D. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 986–1005.
- Mackay, D. (1992). Information-based objective functions for active data selection. *Neural Computation*, 4, 590–604.
- Nelson, J. (2005). Finding useful questions: on bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, 114(3), 677.
- Nelson, J., & Cottrell, G. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing*, 70(13–15), 2256–2272.
- Nelson, J., McKenzie, C., Cottrell, G., & Sejnowski, T. (2010). Experience matters: information acquisition optimizing probability gain. *Psychological Science*, 21(7), 960–969.
- Rehder, B., & Hoffman, A. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, 51, 1–41.
- Stephens, D., & Krebs, J. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.
- Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, 71(5), 680–683.



# One piece at a time: Learning complex rules through self-directed sampling

Doug Markant (doug.markant@nyu.edu)  
Todd M. Gureckis (todd.gureckis@nyu.edu)  
New York University  
Department of Psychology  
6 Washington Place, New York, NY 10003 USA

## Abstract

Self-directed information sampling—the ability to collect information that one expects to be useful—has been shown to improve the efficiency of concept acquisition for both human and machine learners. However, little is known about how people decide which information is worth learning about. In this study, we examine self-directed learning in a relatively complex rule learning task that gave participants the ability to “design and test” stimuli they wanted to learn about. On a subset of trials we recorded participants’ uncertainty about how to classify the item they had just designed. Analyses of these uncertainty judgments show that people prefer gathering information about items that help refine one rule at a time (i.e., those that fall close to a pairwise category “margin”) rather than items that have the highest overall uncertainty across all relevant hypotheses or rules. Our results give new insight into how people gather information to test currently entertained hypotheses in complex problem solving tasks.

**Keywords:** self-directed learning, categorization, active learning, information search, rule learning

## Introduction

A cornerstone of many educational philosophies is that people learn more effectively when they direct or control their own learning experiences (Bruner, 1961). While there are many ways that control might influence learning, an important factor is the ability to actively gather information that one considers potentially useful while avoiding data that is potentially redundant, a behavior referred to as *self-directed sampling* (Gureckis & Markant, in revision).

One recent study directly examined the interaction of self-directed information sampling and learning (Markant & Gureckis, 2010, in revision). In this study, people attempted to learn simple dichotomous categories of objects that varied along two perceptual dimensions (circles that differed in size and the orientation of a central line segment, see Figure 1). In contrast to traditional categorization training procedures, we allowed participants to “design” stimuli that they wanted to learn more about on each trial. Like a child asking their parent to label an unfamiliar object, self-directed “designing” or “sampling” allows the learner to focus on information they want rather than be limited by the flow of passive experience.

The major finding from this study was that for simple unidimensional rules, self-directed learners acquired the correct category rule faster than “passive” participants who were provided samples from an experimenter-defined distribution. In addition, self-directed learners out-performed a set of “yoked” learners who viewed the same examples but did not get to make information sampling decisions themselves (consistent with studies of causal learning with similar yoked comparisons, Lagnado and Sloman, 2004; Sobel and Kushnir, 2006).

## How do people make information sampling decisions?

In light of evidence that self-directed sampling can speed learning, it is important to understand *how* people decide what data to collect. Given a potential observation, what information do people rely on to decide if it will be useful?

One aspect that may help explain a person’s decision to sample an item is their uncertainty in how to classify it (or more generally, their uncertainty about the outcome of any test performed on the item). Intuitively, a self-directed learner should direct their attention to items that are high in uncertainty while ignoring items that can already be confidently classified or predicted. Consistent with this strategy, the pattern of stimuli sampled by self-directed learners in our previous study (see Figure 1) revealed that participants systematically directed their samples toward the category boundary as the task progressed. Intuitively, the learner is mostly likely to be uncertain about these items (e.g., most of the errors in classification occur near the category boundary).

In the current study, we examine how subjective uncertainty in how to classify an item can be used to predict whether or not it is sampled. We begin by presenting three psychologically motivated proposals for how sampling decisions relate to judgments of uncertainty, and then test these models in a new experiment that extends the “self-directed” classification learning paradigm used in Markant and Gureckis (2010). Our results highlight the need for models of sampling behavior that go beyond monolithic measures of information value to consider how people collect and use data during the sequential learning of concepts.

## Three models for relating uncertainty and information sampling decisions

The following sections lay out three possible ways in which uncertainty might guide information sampling decisions.

### Model 1: Sampling to reduce global uncertainty

Prior work on how people gather information has often focused on diagnostic reasoning problems in which the learner is given a set of alternatives (e.g., different diseases) and asked to sample observable features (e.g., symptoms) in order to identify the true diagnosis (Nelson, McKenzie, Cottrell, & Sejnowski, 2010; Skov & Sherman, 1986; Trope & Bassok, 1982). From a computational perspective, various authors have proposed *sampling norms* that attempt to predict information sampling decisions based on a learner’s representation of the task (Nelson, 2005; Nelson et al., 2010; Oaksford & Chater,

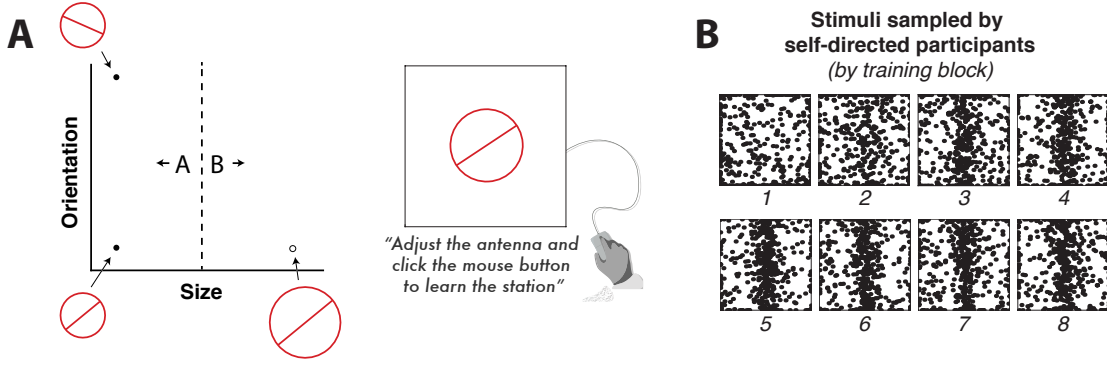


Figure 1: A: Abstract stimulus space used in Markant and Gureckis (2010, in review) and which is adapted for the current study. Stimuli were circles which varied in size and orientation of a central diagonal. In Markant and Gureckis, these objects were assigned to one of two categories (“A” or “B”). Participants “designed” a stimulus they wanted to learn about using the mouse. Clicking the mouse button reveals the category membership of the item. B: The pattern of sampling behavior observed by self-directed learners in Markant and Gureckis (2010) across eight training blocks. Each dot represents a single stimulus which was selected by a participant. In the first block, participants distributed their samples widely over the entire stimulus space but then gradually focused their choices on the region surrounding the category boundary.

1994). Much of this work has focused on what we will call “prospective” models (e.g., *probability gain*, *information gain*, etc.) that estimate the expected drop in uncertainty that will result from making an observation (taking into account all possible outcomes). While in many contexts these models make similar predictions, Nelson et al. (2010) designed a diagnostic reasoning task which found that participants’ choices were best fit by *probability gain*, which values a potential observation according to how much it increases the chance of classifying an item correctly.

In the context of learning a classification rule, this approach is consistent with a preference for sampling items that the learner is least certain about how to classify. Assume that for a given stimulus  $x = (f_1 \dots f_d)$  with  $d$  observable features the learner represents the probability that  $x$  is a member of each possible category  $y$  in the distribution  $P(y|x)$ . We can then define the *least certain* measure as:

$$LC(x) = 1 - \max(P(y|x)) \quad (1)$$

for all stimuli  $x$ . Note that there are alternative norms that make similar predictions to *least certain*, such as using the Shannon entropy of the marginal distribution to calculate uncertainty (see Settles, 2009 for a review). Regardless of the particular form, the important property of this approach is that the most valuable observation is always an item that is considered equally likely to belong to all possible categories. In general, choosing items which maximize  $LC(x)$  should convey the greatest amount of information to the learner.

### Model 2: Isolating individual rule components through margin sampling

While focusing on items that are the most “globally” uncertain or unpredictable seems intuitively useful, there is reason to expect that it may not be the sampling strategy humans use, particularly when learning in complex, multivariate environments. One natural strategy, not captured by *least cer-*

*tain*, might be to decompose a complex task into a series of simpler problems. For example, when multiple features may be related to an outcome, a learner might choose to hold one feature constant while varying the other across multiple samples (Rottman & Keil, 2012). Such a strategy is related to the “control of variables” strategy which is essential to scientific reasoning. Isolating variables often helps people to more efficiently search the space of potential hypotheses (Klahr & Dunbar, 1988) and is a key part of “learning to learn” about complex concepts (Kuhn & Dean, 2005). In an experiment similar to that of Markant and Gureckis (2010), Avrahami et al. (1997) had participants choose samples to teach a partner about a single-dimensional rule and found that the “teachers” frequently used a strategy of isolating individual features. Moreover, their students learned better from this strategy than when given items that were closest to the category boundary.

We can formalize the strategy of focusing on separate components in a sampling model that values uncertainty about any individual boundary between only two categories. *Label margin* predicts that the learner will prefer instances for which the likelihood of any two categories is similar, independent of any other categories. For example, when there are three categories and the marginal distribution is defined as  $P(y|x) = (p_1, p_2, p_3)$ :

$$LM(x) = 1 - \min[|p_1 - p_2|, |p_1 - p_3|, |p_2 - p_3|] \quad (2)$$

Critically, label margin does *not* preferentially select items for which the learner is globally uncertain. Instead, by this approach, a learning problem is decomposed into simpler problems and items are selected which are expected to resolve uncertainty about the sub-components.

### Model 3: Seeking confirmation

While the previous two models propose that people search for uncertain items, previous work on hypothesis testing suggests that people may prefer items that they already know how to

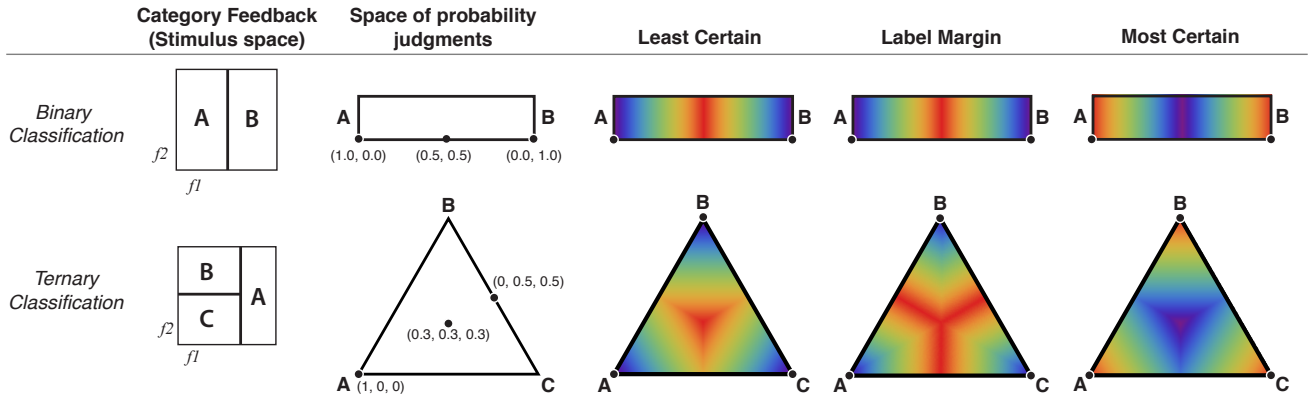


Figure 2: Comparing predictions of sampling norms (red = more highly valued choices, blue = less valued choices) **Top:** For a binary classification problem, a new observation in stimulus space will correspond to a location on the probability judgment scale, where the leftmost point reflects confidence that the observation will be classified “A” and the rightmost point reflects confidence it will be classified “B”. For the binary problem, the predictions of *least certain* and *label margin* are identical. **Bottom:** In a ternary classification problem, an item in stimulus space will correspond to a location in the 3-category simplex depending on the learner’s uncertainty. Here the predictions of *least certain* and *label margin* diverge, allowing us to test which of the two models better account for sampling behavior.

classify. For example, people have a well-documented bias toward seeking positive evidence of the hypothesis they are considering (Klayman & Ha, 1989; Wason, 1960), a strategy that in certain conditions is aligned with the goals of maximizing uncertainty reduction (Austerweil & Griffiths, 2011; Navarro & Perfors, 2011; Nelson & Movellan, 2001). To quantify this strategy, we define the *most certain* measure as:

$$MC(x) = \max(P(y|x)) \quad (3)$$

The predictions of this model directly contrast those of *least certain*, with the highest value assigned to items that can already be classified with confidence. One may also think of the *most certain* measure as instantiating confirmation bias—it shows a preference for items for which the learner has a strong prediction about the category label.

### Empirically distinguishing these alternatives

Given these various approaches, a key question is if they are distinguishable based on empirical data. The predictions of each model are shown for a binary classification problem (like the task used in Markant and Gureckis, 2010) in the top row of Figure 2. Each heatmap describes the value assigned to a potential observation depending on the learner’s uncertainty in how to classify it. For example, an item that can be confidently classified (e.g.,  $p(y|x) = (1, 0)$ ) would be assigned a high value by *most certain* and a low value by *least certain*. Note that for the binary classification problem, *least certain* and *label margin* make identical predictions about how items will be valued (i.e., items close to the center of the space are preferred), making it impossible to separately test these models. However, an interesting observation made by Settles (2009) is that the predictions of these models strongly diverge when considering more complex categorization tasks. For example, in a ternary classification task (see bottom row of Figure 2), *label margin* assigns the maximum value to any items for which one

category is highly unlikely but the learner is uncertain about the other two (shown in Figure 2 by the high predicted value along the radial axes of the simplex, including the midpoints of each edge). In short, this model predicts that samples are likely to be allocated close to any boundary (i.e., “margin”) between two categories. In contrast, *least certain* predicts sampling close to the *junction* of the category boundaries, where all three classes are likely.

### Overview of the current study

The design of our experiment capitalizes on the distinction described in the previous section by extending the paradigm used in Markant and Gureckis (2010) to a ternary classification problem. In the experiment participants collect data by sampling new instances and receiving feedback about their category membership. As shown above, using the ternary classification problem allows us to separate the predictions of the three sampling models, two of which were confounded in our previous design. In order to obtain an estimate of the learner’s uncertainty at any point in time, on a subset of sampling trials participants report how likely they believe the instance they created will belong to each of the three categories (before receiving feedback). The goal of our analysis is to use these subjective judgments to test which model provides the best account of their sampling decisions.

## Experiment

**Participants** Fifty-seven undergraduates at New York University participated in the study for course credit. The experiment was run on standard Macintosh computers in a single 1-hour session.

**Stimuli** Stimuli were defined by a two-dimensional continuous-valued feature space corresponding to the size (radius) of a circle and the angle of a central diameter. These feature values were mapped to a limited range of orientations and sizes on the display. Orientation could vary over only 150 degrees, ensuring that a full rotation of the stimulus was not possible. The two halves of the central diameter were given different colors, further reducing the perceptual similarity of stimuli at the two extremes of the orientation dimension.

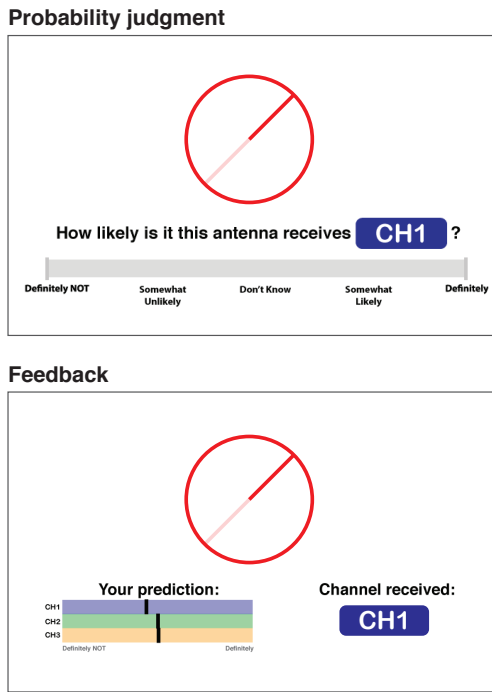


Figure 3: **Top:** Probability judgments were entered by clicking on a scale for each of the three categories (CH1, CH2, and CH3). **Bottom:** Probability judgments were displayed alongside the category label during feedback.

The minimum radius and orientation were randomized so that the boundary between the categories corresponded to a unique boundary in perceptual space for each participant. A total of 256 stimuli were sampled from a uniform grid over the feature space and used as test items for all participants, presented in random order.

The category label associated with each stimulus was deterministically defined by a conjunctive, ternary classification rule of the form shown in Figure 2. In addition to the structure that is shown, three more rules were created through different rotations (90, 180, and 270 degrees) of the same boundaries in stimulus space. Each participant was randomly assigned to one of the four rules.

**Procedure** Participants were instructed that the stimuli in the experiment were television “loop antennas” and that each unique antenna received one of three channels (CH1, CH2, or CH3). Their goal was to learn the difference between the three types of antennas so that they could correctly classify new antennas during the test blocks. Participants were told that the experiment would end when they correctly classified 20 consecutive test items. If a participant failed to reach that goal the experiment ended after 32 blocks or at the end of an hour (whichever occurred first).

**Training Trials.** Participants “designed” antennas by adjusting the size and orientation and receiving feedback about which channel was received. They were instructed that they should design antennas they thought were useful and that would help them to predict the TV channel for other designs they had not yet tested.

Each trial began with the presentation of a randomly generated antenna. Participants then adjusted the size and orientation by moving the mouse from left to right while holding either the ‘Z’ or ‘X’ key, respectively. Only one dimension could be changed at a time, but participants could make any number of changes and were self-paced. When the stimulus was the desired size and orientation, they pressed the mouse button to reveal the channel received, displayed above the stimulus for 4 seconds.

**Probability judgments.** Half of the training trials in each block were randomly selected to include probability judgments. On these trials, after participants had designed an antenna but before the category label was shown, they judged the likelihood that the antenna would receive each of the three channels using a sequence of rating scales (shown in Figure 3). The three scales were presented independently such that only one was visible at a time. When each scale appeared, the participant clicked on a location in the scale according to their belief that the antenna they had designed would receive that channel. A response was required for each scale, and there was no time limit for entering the response. The initial position of the mouse cursor within each scale was randomized, allowing us to record whether responses were influenced by the starting position. After probability judgments were recorded, they were displayed alongside the category label for the same duration as in regular training trials. This allowed the participant to evaluate the accuracy of their prediction given the true category label.

**Test Trials.** Each block of 6 training trials was followed by 8 test trials. On each test trial, a single item was presented in the center of the display and the participant classified the item according to the channel they believed it was most likely to receive. A response was required to complete the trial, and participants responded at their own pace. No feedback was provided on individual test trials. At the end of each block participants were told their accuracy during the block they just completed, as well as the number of consecutive correct responses.

## Results

Three participants were excluded from analysis for failing to complete the task, leaving  $N = 54$ . Thirty-one people reached the goal of correctly classifying 20 items in a row. However, there were a number of additional people who achieved similarly high rates of accuracy. For each subject we computed a moving average of their classification accuracy with a window of 3 blocks, and found 43 people for whom this average exceeded 83% at any point in the experiment (i.e., they correctly classified 20 of 24 items within any three consecutive blocks).

**Probability judgments.** On half of participants’ sampling trials they judged how likely the stimulus they selected belonged to each of the three categories, resulting in three values between 0 and 1. In order to verify that participants were not simply responding based on the position of the cursor, for each rating we measured the difference between the initial (random) position and the participant’s response. One participant was excluded from further analysis because the majority of their ratings (82%) did not change by more than 0.01% from the initial values (for the remaining subjects, the average proportion of samples that met the same condition was  $M = 0.04$ ,  $SD = 0.05$ ).

**Fitting the alternative sampling models.** Our first goal was to assess the overall fit of the three sampling models to each participant’s full set of probability judgments. For each model we computed the normalizing constant necessary to define the probability density function. Each triplet of ratings was normalized so that they summed to one. We then calculated the log-likelihood of each judgment made by a participant and summed across all trials to get an overall score for each model.

Classifying participants according to the model with the highest log-likelihood, we found that 3 people were best-fit by *least certain*, 25 people were best-fit by *label margin*, and the



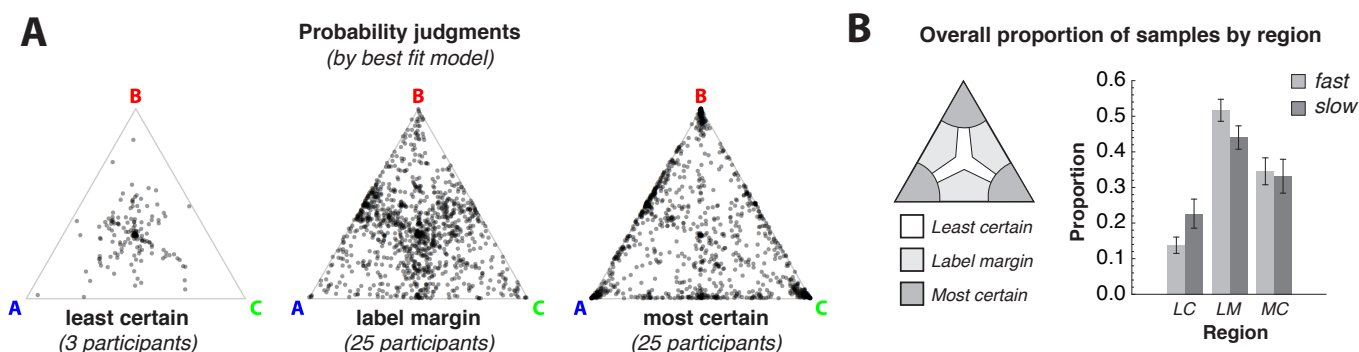


Figure 4: A: Probability judgments are plotted using the 3-category simplex for participants best-fit by each of the three models (see Figure 2 for reference). Each point represents a single judgment after normalization. B: Each judgment was classified according to the model assigning it the highest likelihood, effectively dividing the probability space into three regions. Participants were divided into two groups based on the number blocks they required to complete the task (“fast” and “slow”), and the relative frequency of sampling in each region is shown at right.

remaining 25 people were best-fit by *most certain*. Judgments made by participants, separated by the best-fitting model, are plotted using the 3-category simplex in Figure 4A, with each point a single sample chosen by a participant. A higher density of points reflects an increased tendency (as a group) to sample stimuli in a given region of probability space. Upon visual inspection, the overall pattern for each group corresponds to the predictions of the best-fitting model (Figure 2).

**Relating sampling decisions to learning efficiency.** We next tested whether a participant’s overall success at learning the target concept was related to the sampling behavior reflected in their probability judgments. Our approach was to divide participants into two groups based on the number of blocks they required to complete the task. We performed a median split on the number of blocks (median = 16) to create a group of “fast” ( $N = 26$ ) and “slow” ( $N = 27$ ) learners. With respect to overall model fits, however, there was no difference in the proportion of participants best fit by each model between groups (fast learners:  $N_{LC} = 1$ ,  $N_{LM} = 12$ ,  $N_{MC} = 13$ ; slow learners:  $N_{LC} = 2$ ,  $N_{LM} = 13$ ,  $N_{MC} = 12$ ).

While overall model fits provide a measure of each participant’s sampling behavior in general, inspection of the data showed that most subjects had relatively mixed strategies. For example, a participant best-fit by *most certain* may have made a number of judgments consistent with *label margin*. Given this heterogeneity, we classified individual probability judgments according to the model that assigned it the highest likelihood, effectively dividing the probability space into three regions corresponding to each model (Figure 4B). Multinomial logistic regression was used to test for differences between the relative frequency with which fast and slow learners sampled in each of the three regions. Overall frequency differed significantly between the two groups ( $\chi^2 = 24.7$ ,  $df = 2$ ,  $p < .001$ ). Post-hoc tests showed that fast learners sampled somewhat more frequently in the *label margin* region ( $t(51) = 1.68$ ,  $p = .09$ ) and less frequently in the *least certain* region overall ( $t(51) = -1.91$ ,  $p = 0.06$ ), suggesting that this pattern of sampling was related to success in the task.

## Discussion

Theories of rational information acquisition propose that the decision to make an observation is related to the amount of information it conveys (Nelson, 2005; Oaksford & Chater, 1994). Sampling norms such as *probability gain* prospectively evaluate the change in uncertainty that is expected to occur following an observation, and a rational learner should choose the data that maximizes that measure. Our results illustrate the relative inadequacy of these models when applied to even a basic rule learning task. Very few of our participants were best fit assuming they preferentially selected observations they were least certain about. In addition, the heterogeneity of participants’ sampling strategies is a noteworthy finding. For example, about 20% of samples in the first 4 blocks were “confirmatory” (i.e., data that the learner could already classify with relative confidence), and overall there was no difference in the frequency of this kind of sampling between fast and slow learners. Confirmatory sampling could serve a number of purposes, including helping to organize the representation of a rule in mind (Mathy & Feldman, 2009) or to facilitate comparisons between successive observations, but further work is required to understand its exact role in this task.

**Margin sampling vs. information maximization** A second way in which participants’ behavior diverged from the “rational” prediction was their preference for samples that fell along the category margins over items that offered information about all three categories (i.e., those located at the junction of the boundaries). From the perspective of an ideal observer (i.e., a model that can represent the full set of possible hypotheses and use Bayesian inference to update its beliefs), the most efficient strategy is to maximize the amount of information contained in each observation. Sampling at the category margins should only decrease the efficiency of learning since it will tend to rule out a smaller number of hypotheses, which raises the question of why this kind of behavior was so common in our task.

In our discussion we motivated the margin sampling model by noting that people might decompose a complex prob-

lem into simpler pieces. The use of such a strategy may reflect a participant's limited ability to simultaneously represent all possible alternatives and to remember prior observations. Thus, margin sampling may reflect an adaptation whereby people isolate individual components to learn in succession. Separately testing the role of different features is an important part of scientific thinking in general (Klahr & Dunbar, 1988; Kuhn & Dean, 2005), particularly when intervention is necessary to remove the effect of confounding variables. Importantly, our results do not reveal the particular cognitive processes underlying participants' decisions, but merely provide a descriptive account of their overall behavior. Nonetheless, they provide a useful constraint for theories of information sampling, particularly when applied to more complex tasks that involve sequential learning and memory demands.

**Measuring subjective uncertainty.** It is important to consider that the probability judgments we collected provide an incomplete picture of participants' uncertainty over the course of the task. Although we found some evidence that fast and slow learners differed in the kinds of samples they collected, because uncertainty judgments were assessed on only half of sampling trials it is difficult to draw strong conclusions about the impact of those samples on classification accuracy. Moreover, we cannot be sure that the judgments reported by participants accurately reflected their subjective belief since there were no costs for failing to report accurately<sup>1</sup>. These issues arise whenever considering sampling models based on a learner's subjective uncertainty rather than objective measures of value such as *information gain*, and as such present an important challenge to be addressed in future work.

## Conclusion

Past accounts of information sampling have suggested that a single normative model might account for people's decisions across many learning problems, and that people tend to seek out data that lead to the greatest reduction in uncertainty. In contrast, we found little evidence of a single sampling norm that was consistently applied across individuals. Instead, participants' sampling choices seem to reflect ongoing aspects of constructive problem solving. Our approach highlights the need for theories of self-directed learning to move beyond individual measures of information value to capture interactions with task demands and cognitive constraints.

## Acknowledgments

The authors wish to thank the reviewers for their helpful comments. This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding

<sup>1</sup>Two aspects of our procedure may lessen the impact of this concern. First, we were able to measure a failure to respond by randomly initializing the cursor position before each rating, and found that people changed the position in about 95% of ratings. Second, because judgments were displayed alongside the category label feedback, participants may have been encouraged to respond accurately in order to facilitate processing of that feedback

any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

## References

- Austerweil, J., & Griffiths, T. (2011). Seeking Confirmation Is Rational for Deterministic Hypotheses. *Cognitive Science*, 35, 499–526.
- Avrahami, J., Kareev, Y., Bogot, Y., Caspi, R., Dunaevsky, S., & Lerner, S. (1997). Teaching by examples: Implications for the process of category acquisition. *The Quarterly Journal of Experimental Psychology Section A*, 50(3), 586–606.
- Bruner, J. (1961). The act of discovery. *Harvard Educational Review*, 31(1), 21–32.
- Gureckis, T., & Markant, D. (in revision). A cognitive and computational perspective on self-directed learning. *Perspectives in Psychological Science*.
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, 12, 1–48.
- Klayman, J., & Ha, Y. (1989). Hypothesis testing in rule discovery: strategy, structure, and content. *Journal of Experimental Psychology: Learning*, 15(4), 596–604.
- Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science*, 16(11), 866.
- Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(4), 856–876.
- Markant, D., & Gureckis, T. (2010). Category Learning Through Active Sampling. In S. Ohlsson & R. Catrambone (Eds.). Austin, TX: Cognitive Science Society.
- Markant, D., & Gureckis, T. (in revision). The impact of self-directed information sampling on learning. *Journal of Experimental Psychology: General*.
- Mathy, F., & Feldman, J. (2009). A rule-based presentation order facilitates category learning. *Psychonomic bulletin & review*, 16(6), 1050–1057.
- Navarro, D., & Perfors, A. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological review*, 118(1), 120.
- Nelson, J. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, 114(3), 677.
- Nelson, J., & Movellan, J. (2001). Active inference in concept learning. *Advances in Neural Information Processing Systems*, 45–51.
- Nelson, J., McKenzie, C., Cottrell, G., & Sejnowski, T. (2010). Experience Matters: Information acquisition optimizes probability gain. *Psychological Science*, 21(7), 960.
- Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101(4), 608–630.
- Rottman, B., & Keil, F. (2012). Causal structure learning over time: observations and interventions. *Cognitive Psychology*, 64(1), 93–125.
- Settles, B. (2009). *Active Learning Literature Survey* (tech. rep. No. 1648).
- Skov, R., & Sherman, S. (1986). Information-gathering processes: diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology*, 22(2), 93–121.
- Sobel, D., & Kushnir, T. (2006). The importance of decision making in causal learning from interventions. *Memory and Cognition*, 34(2), 411.
- Trope, Y., & Bassok, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology*, 43(1), 22–34.
- Wason, P. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology*, 12(3), 129–140.

# Shallow learning as a pathway for successful learning both for tutors and tutees

**Noboru Matsuda** (mazda@cs.cmu.edu), **Evelyn Yarzebinski** (eey2@cs.cmu.edu),

**Victoria Keiser** (keiser@cs.cmu.edu), **Rohan Raizada** (rohan@cs.cmu.edu),

**William W. Cohen** (wcohen@cs.cmu.edu)

School of Computer Science, Carnegie Mellon University

5000 Forbes Avenue, Pittsburgh, PA 15213, USA

**Gabriel Stylianides** (gabriel.stylianides@education.ox.ac.uk)

Department of Education, University of Oxford

15 Norham Gardens, Oxford, OX2 6PY, UK

**Kenneth R. Koedinger** (krk@cs.cmu.edu)

School of Computer Science, Carnegie Mellon University

5000 Forbes Avenue, Pittsburgh, PA 15213, USA

## Abstract

SimStudent is a computational model of learning with its cognitive fidelity of learning being demonstrated especially in the way it makes human-like errors. Using SimStudent as a teachable agent in an interactive peer-learning environment, we have investigated how tutee (i.e., SimStudent) learning affected tutor (i.e., human student) learning. In this paper, we are particularly interested in how tutees' shallow learning affects tutor learning. We are also interested in how the errors that the tutee makes affect tutor learning. The results show that teaching SimStudent on a fixed set of problems makes students easy to tutor SimStudent, which in turn helps students learn, but is likely to allow SimStudent to commit shallow learning, which is harmful for tutor learning. It is thus crucial to let the student detect SimStudent's shallow learning and extend teaching until SimStudent and the student achieve satisfactory competence.

**Keywords:** Learning by teaching; teachable agent; SimStudent; shallow learning; learning from errors.

## Introduction

Studying the effect of learning by teaching through the use of teachable-agent technology is a rapidly growing research field. There have been a number of teachable agents used in empirical classroom studies, for example, Betty's Brain (Biswas, Leelawong, Schwartz, Vye, & Vanderbilt, 2005) and TAAG (Pareto, Arvemo, Dahl, Haake, & Gulz, 2011).

Researchers have explored different aspects of the effect of tutor learning, including learning meta-cognitive skills for self-regulated learning (Biswas, Jeong, Kinnebrew, Sulcer, & Roscoe, 2010), the protégé effect (Chase, Chin, Oppizzo, & Schwartz, 2009), the adaptive assistance (Walker, Rummel, & Koedinger, 2009), and the effect of self-explanation (Matsuda, Keiser, et al., 2012).

The teachable agent we have developed is called SimStudent. SimStudent is a machine-learning agent that learns procedural problem-solving skills from examples (Matsuda, Cohen, Sewall, Lacerda, & Koedinger, 2008). SimStudent can be interactively tutored (aka, learning from tutored problem-solving), and has been integrated into an on-line,

game-like learning environment, called APLUS (Artificial Peer Learning environment Using SimStudent). The current version of APLUS allows students to learn Algebra equations by teaching SimStudent. Using APLUS, we have conducted a number of classroom studies to advance cognitive and social theories of tutor learning (Matsuda, Keiser, et al., 2012; Matsuda et al., 2011).

The goal of this paper is to investigate the relationship between tutee- and tutor-learning. As previous empirical studies show (e.g., Cohen, 1994), peer tutoring is known to be beneficial both for tutors and tutees. We thus hypothesize that there must be a strong correlation between SimStudent's and human students' learning. We are particularly interested in how a tutee's shallow learning affects tutor learning. When tutoring, the tutor might fail to detect the tutee's shallow learning by observing the tutee's satisfactory performance at the surface level without actually probing for underlying deep understanding of the domain knowledge. However, if there is actually a symbiotic relationship between tutee and tutor learning, then the tutee's shallow learning should be detrimental to tutor learning.

We are also interested in studying how tutee errors help not only tutee but also tutor learning. In a previous experiment, we studied a theoretical account of the impact of corrective feedback on SimStudent's learning (Matsuda, et al., 2008). We found that committing errors and receiving explicit corrective feedback facilitates tutee learning. On the other hand, it is also known that (human) students learn by explaining erroneous worked-out examples (Grosse & Renkl, 2006; Siegler, 2002). Therefore, tutee errors would also help tutors learn when tutors explain errors committed by tutees. The cognitive fidelity of SimStudent has been demonstrated especially in the way it makes human-like induction errors to learn incorrect skills and hence makes human-like errors when solving problems (Matsuda, Lee, Cohen, & Koedinger, 2009). Therefore, using SimStudent to understand how tutee errors affect tutor learning would be a valid research methodology.



To test the above hypotheses, we conducted a secondary data analysis using the data we collected from our previous classroom studies in which we tested the effect of APLUS.

In the remaining sections, we will first briefly introduce an overview of SimStudent and APLUS in enough detail to understand the research questions and hypotheses. We then describe the data we analyze and the classroom studies from which the data were collected. Finally, we show results followed by a discussion.

## Learning by Teaching SimStudent

### SimStudent

SimStudent is a machine-learning agent that learns procedural skills from examples. When serving as a teachable agent, SimStudent commits to guided problem solving. That is, SimStudent attempts to solve problems given by the student, suggesting one step at a time by applying a learned production. SimStudent asks the student about the correctness of the suggestions. If the student provides negative feedback, SimStudent may attempt to provide an alternative suggestion. When SimStudent has no suggestion that receives positive feedback from the student, then it asks the human student to demonstrate the step as a hint.

The student's feedback and hints become examples that SimStudent generalizes using domain specific background knowledge. As a result, SimStudent generates hypotheses about how to solve problems in the form of *production rules*. SimStudent uses a hybrid learning algorithm that involves (1) inductive logic programming to learn when to apply a production rule, (2) a version space to learn upon what to focus attention, and (3) an iterative-deepening depth-first search to learn how to change the problem state.

SimStudent occasionally prompts students to explain their tutoring actions by asking "why" questions. Such questions include (1) the reason for selecting a particular problem to solve, (2) the reason for an incorrect suggestion, and (3) the reason for the student's demonstration.

### APLUS

APLUS has a Tutoring Interface on which the student and SimStudent collaboratively solve problems. To pose a problem, the student enters an equation into the first row of the Tutoring Interface. As SimStudent makes suggestions for each step, they are placed into the Tutoring Interface. When SimStudent requires a hint, the student demonstrates the next correct step in the Tutoring Interface.

In the regular version of APLUS, the goal of the student is to tutor SimStudent well enough to pass the Quiz. At any time while tutoring, the student may click on the [Quiz] button. SimStudent's productions learned thus far are applied to a set of Quiz problems, and the summary of the results appears in a separate window showing the correctness of the steps suggested by SimStudent. See (Matsuda, et al., 2011) for more details about APLUS.

There have been two versions of APLUS implemented so far, and each version was used in different classroom studies

(see Section "Classroom Studies" for details about the classroom studies). The two versions differ in the structure of the Quiz. In the earlier version, the Quiz problems were fixed. SimStudent took the exact same set of Quiz problems each time it was quizzed. In the later version, the Quiz problems were randomly generated based on a fixed problem *type*. That is, the coefficients and constants were randomly generated each time SimStudent was quizzed.

### How does SimStudent commit Shallow Learning?

In APLUS, one potential pit-fall that may induce SimStudent's shallow learning is the usage and structure of the Quiz. In the earlier study (called the Self-Explanation Study), since the problems in the Quiz were fixed, students could have focused on tutoring only those fixed problems. SimStudent's learning might have been "shallow," or overly specific to solve only those problems, which could have also led human students to "shallow" learning.

On the other hand, the problems in the Quiz for the later study (called the Game Show Study) were randomly generated each time SimStudent took a Quiz (although, they are always in the fixed *type*). Therefore, if SimStudent passes the Quiz in Game Show study, it is likely that SimStudent has learned a high quality set of productions – i.e., "deep" learning. In fact, there were 19 SimStudents that passed the Quiz in the Self-Explanation study, but no SimStudents passed the Quiz in the Game Show study.

An example would help to understand SimStudent's shallow learning. In one instance, SimStudent in the Self-Explanation Study learned to divide both sides of the equation in the form of  $Ax=B$ , where  $x$  is a variable,  $A$  is a coefficient, and  $B$  is a constant term. The production for division says "divide both sides by a chunk of digits before the variable." The "chunk of digits" by definition only perceives a number before the variable without a sign. This piece of background knowledge was designed to model human student's common induction errors (Matsuda, et al., 2009).

As a consequence, this SimStudent could solve equations  $Ax=B$  only when the coefficient  $A$  is a positive number. In the fixed set of the Quiz, this SimStudent learned to solve the equations in such a way that it always happened to have a positive coefficient on the last step, i.e.,  $Ax=B$  (or  $A=Bx$ ). However, even when the same productions were applied, the randomized Quiz problems sometimes produced negative coefficients when combining like terms or balancing the equation. Because of such an accidental transformation, this SimStudent was not able to pass the randomized quiz.

### Research Questions and Hypotheses

This paper addresses the following three research questions and hypotheses.

Q1: *How do tutee and tutor learning relate?* We first hypothesize that SimStudent's learning and human students' learning are correlated. To test this hypothesis, we quantify SimStudent's learning as the "quality" of productions learned by SimStudent. Human students' learning will be quantified using test scores.

Q2: *Is a tutee's shallow learning detrimental to tutor learning?* We hypothesize that letting SimStudents do shallow learning is harmful for tutor learning. To test this hypothesis, we will validate the production rules of SimStudents who passed the fixed Quiz to see if they can also pass the randomly generated Quiz. We will then examine if students who allowed their SimStudents to commit to shallow learning showed poor learning.

Q3: *How do tutee errors influence tutor learning?* We hypothesize that the effect of learning from erroneous examples would apply for tutor learning, that is, detecting SimStudent's errors correctly and explaining those errors would facilitate tutor learning.

## Method

### Sample

The analysis was done on the data we previously collected from two classroom “*in-vivo*” studies; the Self-Explanation Study (Matsuda, Keiser, et al., 2012) and the Game Show Study (Matsuda, Yarzebinski, et al., 2012). Both sets of data are available (upon request) on the large-scale educational database, DataShop (Koedinger et al., 2010), maintained by Pittsburgh Science of Learning Center.

The data include the outcome data and process data. The outcome data are test scores. Students took pre- and post-tests before and after tutoring SimStudent. The test consists of (1) ten equation-solving items, (2) twelve items to determine if a given operation is a logical next step for a given equation, and (3) five items to identify the incorrect step in a given incorrect solution. The pre- and post-tests are isomorphic.

The process data shows the interaction between individual students and SimStudent. It contains (among other things) problems tutored, feedback provided by the students (and their correctness), steps performed both by students and SimStudent (and their correctness), hints requested by SimStudent, and quiz attempts (and their results).

### Classroom Studies

Two classroom studies were conducted in the same school near Pittsburgh, PA, but for different Algebra I classes. The Self-Explanation (SE) study included 111 students from advanced 8th grade and regular & remedial 9th grade classes. The Game Show (GS) study included 141 students from advanced 7th and regular 8th grade classes.

Both studies were conducted as randomized control trials. There were three intervention days (a single class period per day) when students used APLUS. All students took pre- and post-tests before and after the intervention.

For the current study, only the data from the treatment condition in the SE Study and the control condition in the GS Study were used, because the students in those conditions used the same version of SimStudent and APLUS with the same goal for tutoring (i.e., passing the Quiz). As a result, there were 44 students in each condition who took both pre- and post-test and completed the intervention (meaning,

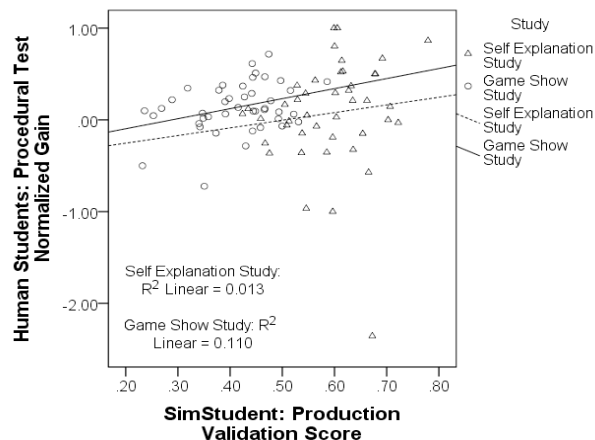


Figure 1: Scatter plot showing SimStudent's validation scores (X-axis) and students normalized gain (Y-axis).

they either attended all three days of the intervention or passed the Quiz sooner).

### Measures

In the following analysis, human students' learning is measured as the normalized gain of the test score, which is computed as  $(\text{post-test} - \text{pre-test}) / (1 - \text{pre-test})$ .

SimStudent's learning is measured as the accuracy of productions learned to solve equations. The productions were applied to a total of 30 equation problems taken from the actual tests that the human students took.<sup>1</sup> For each step in solving an equation, the correctness of each of the applicable productions (i.e., the conflict set) was judged by using the expert model of the Algebra Cognitive Tutor. The *step score* was then calculated as the ratio of correct production firing to all applicable productions. The step score is zero when there are no applicable productions. The *problem score* was computed by averaging the step scores across all steps. Finally, the *validation score* was computed as the average problem score for the 30 equation problems.

The quality of students' responses to SimStudent's “why” questions was also evaluated by three human coders. The coders categorized the student's responses into “deep” and “shallow” responses.

## Results

### Tutor-tutee Learning Correlation

*How does tutee learning correlate with tutor learning?* To answer this question, we first tested the correlation between SimStudent's learning gain and human students' learning gain.

Figure 1 shows the scatter plot with SimStudent's learning represented by the validation score (X-axis) and the human students' learning as the normalized gain on the test scores (Y-axis). Data points from the Self-Explanation

<sup>1</sup> In the classroom studies, there was also a delayed-test. Thus, the students took three tests each containing 10 equation problems.

Study and the Game Show Study are represented using circles and triangles, respectively.

There is a significant correlation between SimStudent's learning and human students' learning for the Game Show Study;  $r(43)=0.331, p<0.05$ .

The correlation between SimStudent's learning and human student's learning was not significant for the Self-Explanation Study;  $r(43)=0.115, p=0.463$ . This might be partly because of the large variance in the human student's learning;  $M=0.07, SD=0.59$ .

*Was there any difference in SimStudent's and human students' learning between the two studies?* The study (SE vs. GS) is a main effect for the SimStudent's validation score;  $t(86)=10.488, p<0.000$ . SimStudents in the Self Explanation Study learn better than the Game Show; mean validation score  $M_{SE}=0.59 (SD=0.08)$  vs.  $M_{GS}=0.41 (SD=0.08)$ .

There was, however, no study difference in the human students' learning; mean normalized gain,  $M_{SE}=0.07 (SD=0.59)$  vs.  $M_{GS}=0.14 (SD=0.27)$ ;  $t(59)=-0.644, p=0.522$ .

## Depth of Learning

The strong correlation between tutee and tutor learning indicates that when the tutee commits shallow learning (which by definition shows good behavior at the surface level without actual learning gain), then the tutor might not learn well.

As mentioned earlier, one potential pit-fall for SimStudent's shallow learning in the APLUS environment is the structure of the Quiz. The likelihood of shallow learning would become higher when the Quiz problems are fixed. To test if the fixed set of Quiz problems actually induced SimStudent's shallow learning, and, if so, whether SimStudent's shallow learning also induced human students' shallow learning, we analyzed both human students' and SimStudent's shallow learning.

To test SimStudent's shallow learning, we investigated if SimStudents in the SE study who passed the (fixed) Quiz could also pass the (randomly generated) Quiz used in the GS study. There were 19 SimStudents who passed the Quiz in the SE study (SE passing SimStudent). We first extracted productions learned by those 19 SE passing SimStudents from the process data. Ten sets of the GS study Quiz (each with eight problems) were randomly generated. For each SE passing SimStudent, we then applied productions for each problem in the ten sets of the Quiz from the GS study.

Figure 2 shows the results of the SE passing SimStudents taking the GS study Quiz. The table shows the number of SE passing SimStudents that passed at most the specified number of quiz sets (out of 10).

There were 8 SE passing SimStudents who passed one or more sets of GS quiz. Only 7 (37%) SE passing SimStu-

Maximum Num. Quizzed Passed	0	1	10
Num. of SimStudent	11	1	7

Figure 2: Result of the SE SimStudents who passed the SE Quiz (N=19) taking the GS Quiz

dents could pass two or more sets of GS Quiz. Interestingly, if SimStudent could pass two GS Quizzes, it could also pass all ten sets of GS Quizzes. To our surprise, 63% of SimStudents in the SE study passed the SE Quiz by committing "shallow" learning that was enough to pass the fixed set of Quiz problems.

*Was SE SimStudent's shallow learning detrimental for human students learning?* To see if human students actually committed shallow learning by quitting the tutoring session after seeing that their SimStudents passed the Quiz, we re-examined relationship between human students' learning and SimStudent's learning.

Figure 3 plots human student's plain test scores (Y-axis) and SimStudent's validation scores (X-axis). The figure only includes those 19 students who passed the Quiz in the SE Study. Data taken from a single human student are plotted as two dots connected with a vertical line. A large and a small dot show a human student's post- and post-test scores, respectively, both on the Y-axis.<sup>2</sup> The vertical line represents the pre- to post-test gain, with an upward line showing a positive gain and a downward line negative gain. SimStudent's learning (measures as the validation score) is shown as the position of the connected dots on the X-axis.

If there were students who committed shallow learning, then we should see the pair of dots connected with a relatively short upward line (i.e., a small positive gain) or a downward line (i.e., a negative gain) in the lower left corner. The human students with relatively short lines in the top area are likely to be ceiling students, and SimStudents in the right half are not likely to have committed shallow learning.

As can be seen in Figure 3, there are a group of human students who have relatively short or downward line at a relatively low pre-test score. They are the students who managed to have their SimStudents pass the Quiz, but the students themselves achieved very little learning gain.

## Impact of Tutee Error for Tutor Learning

*How do tutees' errors help tutor learning?* To answer this question, we probed the process data to quantify several tutoring activities related to error detection and correction, and tested their correlations with tutor learning.

On average, SimStudent made 3.3 errors per problem ( $SD=2.4$ ). The number of errors made by SimStudent per problem was not correlated with tutor learning;  $r(84)=-.012, p=0.92$ . The average probability for SimStudent making an error, which was computed as a ratio of incorrect suggestions to all suggestions per problem and averaged for all problems aggregated across all SimStudents, was not correlated with tutor learning;  $r(85)=-.087, p=0.429$ .

On average, human students correctly detected SimStudent's errors 2.3 times per problem ( $SD=1.9$ ). There was no correlation between the number of times human students

<sup>2</sup> Test scores can be negative, because a point was subtracted for a wrong selection on multiple-choice items.

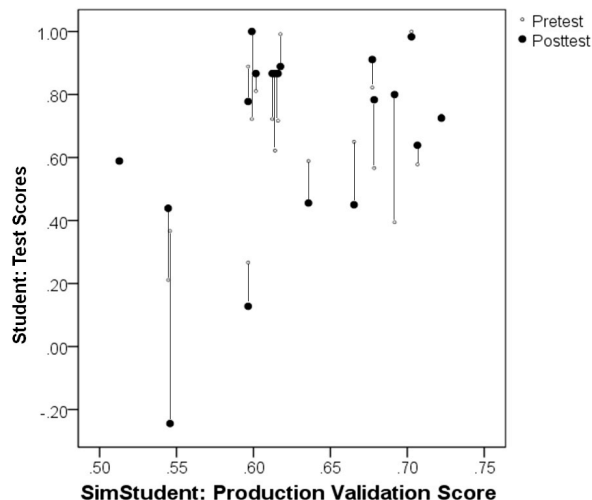


Figure 3: Relation between human students' test scores (Y-axis) and SimStudent's validation scores (X-axis). Large filled dots show post-test scores, whereas small circles show pre-test scores. Only students in the SE study who managed their SimStudents passing the Quiz are included.

correctly detected SimStudent's errors and tutor learning;  $r(84) = -0.178, p = 0.105$ .

The tutors' probability of correctly detecting tutee errors (PED) significantly predicts tutors' normalized learning gain (NRG). The regression coefficient in predicting NRG with PED was  $NRG = 0.17 * PED, p < 0.05$ . It is important to note that part of the test requires error detection within worked examples; success on the test therefore depends on this skill.

However, the NRG does not significantly predict the PED. It appears that tutor learning does not solely contribute to the ability of tutors to detect tutee errors. The regression coefficient in predicting PED with NRG was  $PED = 0.69 + 0.07 * NRG, p < 0.001$  and  $p = 0.17$ , respectively.

As explained earlier, when a student provided negative feedback during tutoring, SimStudent asked the student to explain why the step was incorrect. The theory of self-explanation on erroneous examples predicts that responding to SimStudent's questions about its errors facilitates tutor learning (e.g., Grosse & Renkl, 2007). This is actually the case in our study as well. The ratio of "deep" explanations to all explanations on the "why am I wrong?" type of questions (DXP) was also significantly predictive of normalized gain (NRG). The regression coefficient was  $NRG = 0.44 * DXP$  with  $p < 0.05$ .

The above observations suggest that having tutors correctly detect tutee errors and elaborately explain the error would likely facilitate tutor learning. The current APLUS does not provide explicit assistance for the students to ensure such a good tutoring behavior. As some of the previous studies demonstrated (Leelawong & Biswas, 2008; Walker, et al., 2009), integrating a meta-tutor that guides the human student tutoring into APLUS might, therefore, improve the efficacy of APLUS.

## Discussion

Shallow learning is an inevitable natural pathway for deep learning. When students engage in inductive learning, they usually search through a huge problem space with limited search heuristics that hardly avoid making errors. Indeed, there are very many different types of errors that students can make when doing induction (Matsuda, et al., 2009).

About 2/3 of SimStudents (12 out of 19) who passed the fixed Quiz in the SE Study were actually shallow learners (i.e., failed to pass the randomized Quiz in the GS study). On the other hand, there were no SimStudents who passed the Quiz in the GS study. Since there was a difference in the student's grade level for these two studies, the results must be interpreted with caution. Nonetheless, the GS Study data show a high correlation between SimStudent learning and human student learning. This means that the better SimStudent learns, the better the human student also learns. Therefore, a fixed set of quiz problems should work better than a randomized set of quiz problems, because it helps students tutor SimStudent, which makes a better SimStudent. In turn, this should lead to better tutor learning. In the SE Study, students who passed the Quiz used a higher percent of the failed Quiz problems for tutoring ( $M = 0.95, SD = 0.11$ ) than students who did not pass ( $M = 0.59, SD = 0.42$ );  $t(28) = -4.079, p = .000$ , suggesting that copying the failed Quiz problems helped students in managing to pass the Quiz.

On the other hand, the high correlation between SimStudent's and (human) students' learning also suggests that failing to detect SimStudent's shallow learning is likely to cause students' poor learning. Therefore, the students must detect SimStudent's shallow learning. Our data show that catching SimStudent's shallow learning is rather inexpensive – only two sets of Quiz problems are enough. Having SimStudent take two or more different sets of Quiz problems should help students detect SimStudent's shallow learning.

Our current data also show that tutors learn from errors that the tutee makes. The probability of correctly detecting tutee errors is significantly predictive of tutor learning. Also, the ratio of elaborated explanations to all explanations given to incorrect steps is significantly predictive of tutor learning.

In conclusion, our data suggest that the inevitable nature of inductive learning, i.e., the tutee's *intermediate* shallow learning and errors of commission, facilitate tutor (as well as tutee) learning. In a certain situation (such as APLUS), letting the tutee reach shallow learning might help the tutor manage to teach the tutee without too much of a teaching burden. However, it is crucial for the tutor to detect the tutee's shallow learning and continue teaching toward deeper understanding. Our data also suggest that errors that the tutee makes during tutoring are beneficial both for the tutor and tutee. For the tutee, corrective feedback expedites its learning. For the tutor, elaborated reflective explanations on tutee errors facilitate learning.

The above findings also suggest that weaving fixed and randomly generated sets of quiz problems should induce

optimal learning both for SimStudent and human students. One realization would be to provide a set of randomly generated Quiz problems and let SimStudent try the same (fixed) set of problems until SimStudent passes them, and then provide another set of randomly generated Quiz problems. As shown in Figure 2, passing only two sets of randomly generated Quizzes would be enough to ensure SimStudent's deep learning, which in turn prevents students from shallow learning.

### Acknowledgments

The research reported here was supported by National Science Foundation Award No. DRL-0910176 and the Institute of Education Sciences, U.S. Department of Education, through Grant R305A090519 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. This work is also supported in part by the Pittsburgh Science of Learning Center, which is funded by the National Science Foundation Award No. SBE-0836012.

### References

- Biswas, G., Jeong, H., Kinnebrew, J. S., Sulcer, B., & Roscoe, R. (2010). Measuring Self-Regulated Learning Skills through Social Interactions in a teachable Agent Environment. *Research and Practice in Technology Enhanced Learning*, 123-152.
- Biswas, G., Leelawong, K., Schwartz, D., Vye, N., & Vanderbilt, T. T. A. G. a. (2005). Learning by teaching: a new agent paradigm for educational software. *Journal of Applied Artificial Intelligence*, 19(3&4), 363-392.
- Chase, C., Chin, D., Oppezzo, M., & Schwartz, D. (2009). Teachable Agents and the Protégé Effect: Increasing the Effort Towards Learning. *Journal of Science Education and Technology*, 18(4), 334-352.
- Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research*, 64(1), 1-35.
- Grosse, C. S., & Renkl, A. (2006). Learning from worked examples: what happens if errors are included? . In J. E. P. Gerjets, R. Joiner & P. Kirschner (Eds.), *Instructional design for effective and enjoyable computer-supported learning*. Tübingen: Knowledge Media Research Center.
- Grosse, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction*, 17(6), 612-634.
- Koedinger, K. R., Baker, R. S. J. d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A Data Repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy & R. S. J. d. Baker (Eds.), *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press.
- Leelawong, K., & Biswas, G. (2008). Designing Learning by Teaching Agents: The Betty's Brain System. *International Journal of Artificial Intelligence in Education*, 18(3).
- Matsuda, N., Cohen, W. W., Sewall, J., Lacerda, G., & Koedinger, K. R. (2008). Why Tutoed Problem Solving may be better than Example Study: Theoretical Implications from a Simulated-Student Study. In B. P. Woolf, E. Aimeur, R. Nkambou & S. Lajoie (Eds.), *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 111-121). Heidelberg, Berlin: Springer.
- Matsuda, N., Keiser, V., Raizada, R., Yarzebinski, E., Watson, S., Stylianides, G. J., et al. (2012). Studying the Effect of Tutor Learning using a Teachable Agent that asks the Student Tutor for Explanations. In M. Sugimoto, V. Aleven, Y. S. Chee & B. F. Manjon (Eds.), *Proceedings of the International Conference on Digital Game and Intelligent Toy Enhanced Learning (DIGTEL 2012)* (pp. 25-32). Los Alamitos, CA: IEEE Computer Society.
- Matsuda, N., Lee, A., Cohen, W. W., & Koedinger, K. R. (2009). A Computational Model of How Learner Errors Arise from Weak Prior Knowledge. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 1288-1293). Austin, TX: Cognitive Science Society.
- Matsuda, N., Yarzebinski, E., Keiser, V., Raizada, R., Stylianides, G., Cohen, W. W., et al. (2011). Learning by Teaching SimStudent – An Initial Classroom Baseline Study comparing with Cognitive Tutor. In G. Biswas & S. Bull (Eds.), *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 213-221): Springer.
- Matsuda, N., Yarzebinski, E., Keiser, V., Raizada, R., Stylianides, G., & Koedinger, K. R. (2012). Motivational factors for learning by teaching: The effect of a competitive game show in a virtual peer-learning environment. In S. Cerri & W. Clancey (Eds.), *Proceedings of International Conference on Intelligent Tutoring Systems* (pp. 101-111). Heidelberg, Berlin: Springer-Verlag.
- Pareto, L., Arvemo, T., Dahl, Y., Haake, M., & Gulz, A. (2011). A Teachable-Agent Arithmetic Game's Effects on Mathematics Understanding, Attitude and Self-efficacy. In G. Biswas, S. Bull, J. Kay & A. Mitrovic (Eds.), *Proceedings of the International conference on Artificial Intelligence in Education* (pp. 247-255). Heidelberg, Berlin: Springer.
- Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31-58). New York, NY: Cambridge University Press.
- Walker, E., Rummel, N., & Koedinger, K. R. (2009). Integrating collaboration and intelligent tutoring data in the evaluation of a reciprocal peer tutoring environment. *Research and Practice in Technology Enhanced Learning*, 4(3).

# Going with TRACE beyond Infant Mispronunciation Studies: Lexical Networks and Phoneme Competition

Julien Mayor (julien.mayor@unige.ch)

FPSE, University of Geneva  
1211 Genève 4, Switzerland

Kim Plunkett (kim.plunkett@psy.ox.ac.uk)

Department of Experimental Psychology, University of Oxford  
Oxford OX1 3UD, United Kingdom

## Abstract

The TRACE model of speech perception (McClelland & Elman, 1986) is used to simulate graded sensitivity to mispronunciations of familiar words as reported by White and Morgan (2008). Our simulations predict that phoneme or lexical competition may be absent in the mental lexicons of the 19-month-old infants tested experimentally.

**Keywords:** Word learning; speech perception; language acquisition; inhibition

## Introduction

Research on infant spoken word recognition has made dramatic advances over the past two decades. Spurred on by the refinement of experimental techniques such as the familiarisation head turn preference procedure (Jusczyk & Aslin, 1995), the switch task (Stager & Werker, 1997) and the mispronunciation task (Swingley & Aslin, 2000), our understanding of *what* infants and young children know about the sounds of words, both familiar and newly learnt, has expanded incrementally. However, our appreciation of the representations and processes underlying early phono-lexical knowledge and *how* these develop is less advanced. Although these approaches offer important insights as to how infants and young children develop knowledge about the sounds of words, they do not provide a precise computational account of the representations and processes involved. In this paper, we describe our attempt to apply the TRACE model of word recognition (McClelland & Elman, 1986) to simulate aspects of spoken word recognition during infancy and early childhood.

TRACE was originally proposed as a model of adult spoken word recognition. In TRACE, spoken word recognition is modelled as an incremental process involving the elimination of competing candidates that are represented in the individual's mental lexicon. Various accounts have emphasised the role of cohort competitors (Marlsen-Wilson & Welsh, 1978) and phonological neighbours (Cutler, 1995; Goldinger, Luce, & Pisoni, 1989) in this competition. Allopenna, Magnuson, and Tanenhaus (1998) have argued that the TRACE model of speech perception provides a satisfactory accommodation for the role of cohorts and phonological neighbours in the resolution of the competitive process.

In TRACE, acoustic-phonetic features are mapped over time onto phoneme nodes that map onto lexical nodes,

with lexical-phonemic feedback and lateral inhibition at the phonemic and lexical levels (see Fig. 1).

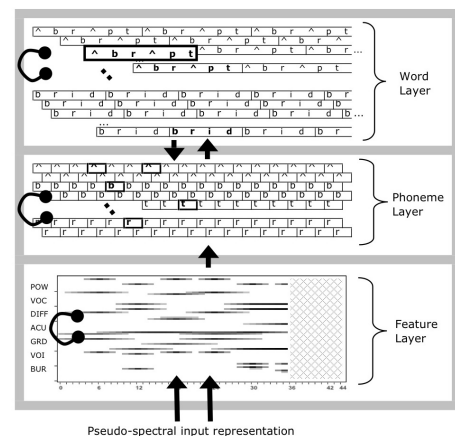


Figure 1: Schematic diagram of TRACE architecture. Drawing by Ted Strauss.

Allopenna et al. (1998) found that the time course of spoken word recognition indexed by eye movements in human participants can be modelled by such continuous mapping models: Adults were instructed to move one of four objects that were on a screen, while they were simultaneously monitored by an eye-tracker. Along with the referent, three competitors were displayed on screen; a cohort competitor (object starting with the same onset and vowel), a rhyme competitor and an unrelated competitor. Using the TRACE model, implementing a forced choice with Luce's choice rule (Luce, 1959), Allopenna et al. (1998) accurately reproduced the typical pattern of eye-gaze of the participants. For example, adults were likely to be distracted by both cohort and rhyme competitors in this task. TRACE also exhibits enhanced activation for these competitors resulting in enhanced levels of "eye fixations" when using Luce's choice rule. More recently, TRACE has been used to model adult's gradient sensitivity to within-category voice onset time manipulations in a visual world task (McMurray, Tanenhaus, & Aslin, 2009) and individual differences in online spoken word recognition, including individuals at risk for specific language impairment (McMurray, Samelson, Lee, & Bruce Tomblin, 2010). In both these applications, exploration of TRACE's parameter

space identified factors (phoneme inhibition and lexical decay, respectively) that might account for observed human performance.

We adopt the same approach and try and refine the potential architecture underlying infant word recognition by simulating with TRACE the finding that infants display a graded sensitivity to the severity of mispronunciations (White & Morgan, 2008).

### Graded sensitivity to the severity of mispronunciations; Implications

Infants show graded sensitivity to mispronunciations of familiar words, as a function of the severity of the mispronunciation. White and Morgan (2008) have shown that 19-month-olds show a graded response in their looking behaviour when presented, in an Inter-Modal Preferential Looking (IPL, Golinkoff, Hirsh-Pasek, Cauley, & Gordon, 1987) task, with a correct pronunciation, 1-feature, 2-feature or 3-feature mispronunciation of the onset consonant of a target word: Infants look longer at the target object when supplied with more accurate renditions of the target object's name. In their experiment, the two pictures corresponded to a target object and a novel object. In contrast to other mispronunciation experiments (Mani & Plunkett, 2007; Swingley & Aslin, 2000), the distracter image is name-unknown and thus does not represent a potential competing lexical entry as it is a novel image. White and Morgan (2008) argued that using a novel object as a distracter is important for demonstrating graded sensitivity as it offers the infant the opportunity to consider the mispronunciation as a label for the novel distracter. This possibility is not available to the infant when the distracter is a name-known object. On the basis of their experimental findings, White and Morgan (2008) argue that lexical processing in toddlers is affected by sub-segmental phonological detail. In this simulation, we examine the adaptations of the TRACE architecture that are needed to simulate the (White & Morgan, 2008) results, and explore the ramifications of these adaptations for interpreting their experimental findings.

### Method

We used jTRACE (Strauss, Harris, & Magnuson, 2007), a re-implementation of the TRACE model (McClelland & Elman, 1986). We created typical lexicons for 18 month olds by compiling British CDIs ((Hamilton, Plunkett, & Schafer, 2000), a British adaptation of the MacArthur-Bates CDI, (Fenson et al., 1993)) using words that are understood by at least 50% of the infants at 18 months of age. The lexicon is specified using data from 179 infants and count 131 words.

Recognition time for spoken words is affected not only by the number of phonological neighbours (Cutler, 1995), but also by their frequency (Goldinger et al., 1989). Therefore, we identified individual token frequencies, by extracting word frequencies on all tiers based on the Manchester corpora (Theakston, Lieven, Pine, & Rowland, 2001) from

the CHILDES database (MacWhinney, 1991), where 12 English children were recorded weekly from 20 to 36 months of age. Word frequencies used in the simulations are raw word counts on the whole corpora, converted to frequency per million. When implementing frequencies in the model, we follow the suggestions advocated by MacKay (1982) and implemented by Dahan, Magnuson, and Tanenhaus (2001), i.e., frequency modulates the connection weights associated with lexical units, using the same value for the scaling parameter (0.13) used in (Dahan et al., 2001). The modulation of frequency effects via phoneme-lexicon connection strengths is consistent with a learning basis for frequency (e.g., of the Hebbian type). In addition, Dahan et al. (2001) found this type of bottom-up connection strength implementation to have qualitative advantages over resting state and post-perceptual frequency manipulations.

Given the large size of the infant lexicon at 18 months of age, many of the phonemes needed to represent the different words were not encoded in the original TRACE model (McClelland & Elman, 1986) nor in its re-implementation (Strauss et al., 2007). Therefore, we added feature values for all phonemes used in the infant's lexicon<sup>1</sup>.

Correctly pronounced words and mispronounced words are presented to the model and activation levels of two competitors (the target and a distracter) are monitored. We adopt the same linking hypothesis as (Alloppenna et al., 1998) in order to map the activation levels to fixation durations. Activation of a word is the result of both its direct activation due to phonological overlap with the input and the result of competition with all other words that are activated with that same input. Only items that are on display are available as potential responses. Similarly to (Alloppenna et al., 1998), the activation levels  $a$  of the displayed items are then transformed into response strengths following (Luce, 1959). Given the high salience of the images, we assume that total looking time is split entirely between the target and distracter objects, enabling us to convert the response strengths into fixation durations using the Luce choice rule. The proportion of looking to the target at time  $t$  is given by:  $p_{target}(t) = \frac{e^{ka_{target}(t)}}{e^{ka_{target}(t)} + e^{ka_{distractor}(t)}}$  where  $k$  is a free parameter determining the amount of separation between units of different activations (value set to  $k = 2$ ). All other parameters used in jTRACE were set to their default values. Proportion of looking times to the target and distracters are reported as the average over 100 processing cycles starting with the onset of the pronounced word.

We used the stimuli described in Experiment 1 of (White & Morgan, 2008), reproduced in Table 1, with the exception of the word "cookie", which is not present in the British version of the CDI that we used to create the new jTRACE dictionaries. Since the distracter is name-unknown in the White and Morgan (2008) experiment, the activation level associated with the novel object on display is set to zero. It is noteworthy that, however, due to the application of Luce's

<sup>1</sup>Thanks to Ōiwi Parker-Jones for help in assigning feature values for phonemes not present in the original TRACE model.



rule, both images share some amount of the total looking time spent during each trial<sup>2</sup>.

Table 1: Correctly pronounced and mispronounced labels presented to infants in Experiment 1 of White & Morgan (2008). The unfamiliar words used by White and Morgan (2008) are not listed here because they do not compete for recognition in TRACE. The table also includes the cohort size as a function of pronunciation type for the stimuli used in White & Morgan (2008).

Correct pronunciations	Mispronunciations		
	1-feature	2-feature	3-feature
keys	teys	deys	zeys
book	dook	took	sook
bear	gear	tear	sear
foot	soot	zoot	goot
car	par	dar	zar
ball	gall	kall	sall
bird	gird	kird	sird
bottle	gottle	kottle	sottle
shoe	foe	voe	goe
cup	tup	bup	vup
hand	fand	zand	dand
Mean cohort size (SD)			
18.7 (12.1)	7.7 (7.2)	11.7 (9.9)	4.4 (2.5)

In this approach, mispronunciations cannot act as potential labels for the distracter image since the distracter image is name-unknown. The unfamiliar words used by White and Morgan (2008) do not belong to the lexicon, and therefore do not compete for recognition in TRACE. Simulations were run with the 18-month-lexicon to mimic the behavior of 19-month olds.

## Results

First, we ran simulations with jTRACE’s default parameters for the same stimuli used by White and Morgan (2008). The top panel of Figure 2 depicts the proportion of looking time associated with the target in the correct, 1-feature-, 2-feature- and 3-feature-mispronunciations. No graded sensitivity is observed as a function of the severity of mispronunciation. Since the metrics used by White and Morgan (2008) to derive the severity of mispronunciation may differ slightly from jTRACE’s, we also evaluate the impact of the severity of mispronunciations on the level of activation of the target words within jTRACE’s metrics. The bottom panel of Figure 2 depicts the reduction in activation level as a function of the magnitude of the mispronunciation (Euclidean distance between the two phonemes in jTRACE’s feature space) for all stimuli. The absence of any correlation suggests that activation levels of target words are not directly sensitive to the severity of mispronunciations, in contrast to White and Morgan (2008) findings.

Closer examination of the stimuli used by White and Morgan (2008) reveals that the number of cohort competitors in the typical lexicon of an 18-month old differs dramatically with mispronunciation type. Table 1 presents an analysis

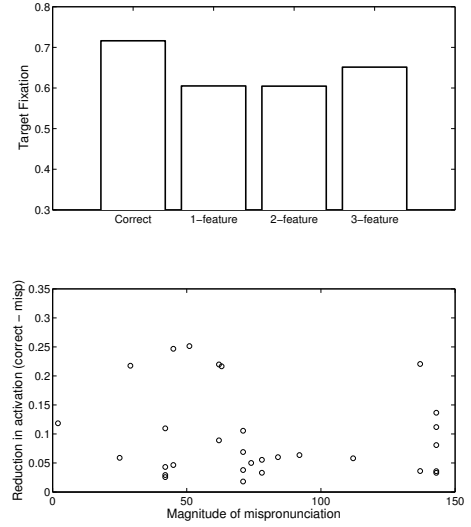


Figure 2: Top Panel: Simulation of White and Morgan (2008) with jTRACE’s default parameters. The unbalanced cohort sizes in each condition interferes with the bottom-up activation flow favoring graded sensitivity to the severity of the mispronunciations. In particular, looking times in the three-feature mispronunciation condition are longer than in the one- and two-feature mispronunciation conditions. Bottom panel: Mispronunciation effect (reduction in activation due to the mispronunciation) as a function of the magnitude of the mispronunciation in jTRACE’s feature space. No correlation is observed between looking times and the severity of mispronunciations.

of the cohort size associated with correct pronunciations and each mispronunciation type. It is apparent that 3-feature mispronunciations have far fewer cohort competitors than any of the other mispronunciation conditions. An item-analysis of variance of the number of cohort competitors across types of pronunciation yielded a main effect of pronunciation condition ( $F=5.53$ ,  $df=3$ ,  $p=0.0028$ ). Two feature mispronunciations have marginally more cohort competitors than 1-feature mispronunciations ( $t=1.34$ ,  $df=10$ ,  $p=0.21$ , n.s.), and more importantly, more than 3-feature mispronunciations ( $t=2.40$ ,  $df=10$ ,  $p=0.038$ ).

An important characteristic of TRACE is that it implements competition within the different layers of the network. As a consequence, cohort competitors impact the activation levels associated with a target word. A low number of cohort competitors leads to reduced inhibition which, in turn, leads to higher activation of the target word. For the stimuli used in (White & Morgan, 2008), we expect the cohort competition in TRACE to interfere with any mispronunciation effect. In particular, the low number of cohort competitors in the case of the 3-feature mispronunciation would lead to an *increase* in the activation of the target word, rather than to a *decrease*

<sup>2</sup> $p_{target}(t)$  being an exponential function of word activation.

in its activation level. Clearly, this outcome would be incommensurate with White & Morgan’s finding of a graded sensitivity to severity in the mispronunciation and explains why a graded sensitivity to the severity of mispronunciations was not observed with jTRACE’s default parameters. Therefore, we conducted a series of simulations so as to evaluate the impact of word-layer and phoneme-layer inhibition on sensitivity to mispronunciation.

First, we investigate the impact of reducing the level of lexical inhibition. Both theoretical and experimental considerations motivate this adaptation of TRACE: Lexical inhibition may be reduced in infancy due to the sparseness of the lexical space (Gaskell & Marslen-Wilson, 1997). Also, several recent experimental findings provide evidence that word to word interactions do not reach adult levels of competition before about 21 months of age. For example, Arias-Trejo and Plunkett (2009) and Styles and Plunkett (2009) used a semantic priming task with infants to demonstrate evidence for lexico-semantic networks in 21- and 24-month old infants. However, they failed to find evidence of semantic priming in 18-month olds. Arias-Trejo and Plunkett (2009) suggest that entries in the 18-month old lexicon may be best characterised in terms of *lexical islands* that are not in competition with each other because they are unconnected. More direct evidence is provided in a phonological priming task (Mani & Plunkett, 2011) conducted with 18- and 24-month old infants. Mani and Plunkett (2011) reported cohort effects in 24-month olds (less target looking for words from large cohorts than words from small cohorts) but no cohort effects for 18-month olds. It is likely that these age differences in cohort effects are driven by differences in the vocabulary sizes of the infants involved in the study, even though both age groups were tested on the same set of words. This set of findings, together with the findings from (Arias-Trejo & Plunkett, 2009) and (Styles & Plunkett, 2009), provide a convergent rationale for reducing lexical competition in the simulation of White & Morgan’s 19-month old infants.

The top panel of Figure 3 displays the proportion of target looking in jTRACE associated with the stimuli used by White and Morgan (2008) for correct, 1-feature, 2-feature and 3-feature mispronunciations when lexical inhibition is essentially turned off ( $C = 0.0001$ )<sup>3</sup>. A graded sensitivity to the severity of mispronunciations emerges, similar to the 19-month-olds tested by White and Morgan (2008). However, correlations between the reduction of activation levels associated with target words and the magnitude of the mispronunciations in TRACE’s feature space did not reach significance ( $p = 0.13$ , see bottom panel of Figure 3). For  $C \geq 0.001$ , cohort effects counteract the effect of mispronunciation such that the activity level associated with the 3-feature mispronunciations is higher than the activity level associated with the 2-feature mispronunciations.

<sup>3</sup>For comparison, the value commonly used to model adult sensitivities to mispronunciations is  $C = 0.03$  (see for example (Allopenna et al., 1998)) which means inhibition in the word layer is 300 times stronger than the value used here.

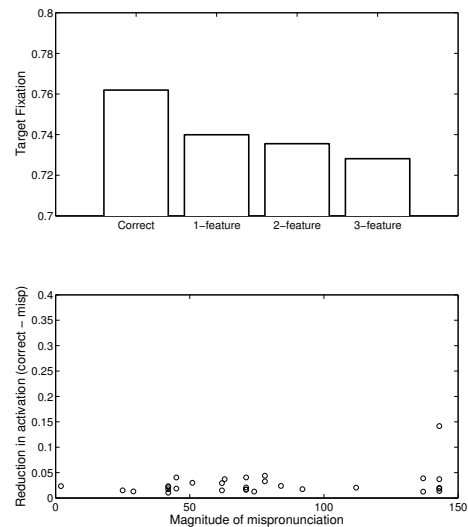


Figure 3: Top Panel. Simulation of White and Morgan (2008) with TRACE with reduced lexicon competition. Cohort effects are reduced and the bottom-up activation flow favoring graded sensitivity to the severity of the mispronunciations is not disrupted. Bottom panel. Mispronunciation effect (reduction in activation due to the mispronunciation) as a function of the magnitude of the mispronunciation in TRACE’s feature space. A weak, non-significant, correlation is observed between looking times and the severity of mispronunciations.

A second manipulation that may lead to a reduction in the influence of imbalanced cohort sizes when simulating White & Morgan’s findings is to reduce phoneme inhibition. McMurray et al. (2009) suggest that phoneme-level inhibition in TRACE is incompatible with recovery from “lexical garden-paths” initiated by ambiguous phonemes early in a word. We now consider the impact that the absence of phoneme-level inhibition may have on simulations of White & Morgan’s findings. The top panel of Figure 4 depicts the proportion of looking time at the target when correctly pronounced, and with three levels of mispronunciation severity, when phoneme level inhibition is eliminated in TRACE. A clear, graded reduction in activation level emerges as the number of feature changes increases. Furthermore, the bottom panel of Figure 4 indicates that, within TRACE’s feature metrics, a significant correlation ( $R = 0.753$ ,  $p = 1.56 \cdot 10^{-6}$ ) is present between the magnitude of the mispronunciation and its impact on activation levels. Cohort effects are effectively reduced and the bottom-up flow from the feature level to the lexical level, via the phoneme level, is not disrupted by cohort effects.

## Discussion

(White & Morgan, 2008) reported a graded sensitivity in 19-month old infants to the severity of the mispronunciation of

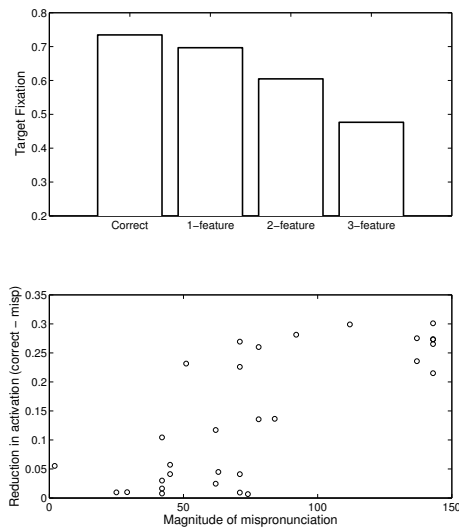


Figure 4: Top Panel: Simulation of White and Morgan’s (2008) findings in jTRACE with no phoneme inhibition. Cohort effects are reduced and the bottom-up activation flow favouring graded sensitivity to the severity of the mispronunciations is well established. Bottom panel: Mispronunciation effect (reduction in activation due to the mispronunciation) as a function of the magnitude of the mispronunciation in TRACE’s feature space. A strong, significant, correlation is observed between target preference and the severity of mispronunciations.

a target word and argued that this finding demonstrated fine-grained sensitivity at the sub-segmental level. The gradual decrease in looking time at the target object as the number of modified features increased was observed despite the fact that the number of cohort competitors for mispronunciations, as evaluated by an analysis of CDI reports, was smaller for the 3-feature mispronunciations than for the 2-feature mispronunciations. Competition between word activation levels in TRACE has an opposite effect on target word activation to mispronunciation severity for the stimuli used by White and Morgan (2008), leading to an apparent incompatibility between White & Morgan’s findings and the predictions of jTRACE. The fact that White and Morgan (2008) report that target looking decreased with mispronunciation severity suggests that either inhibition between competing words in the lexicon is not present (or extremely reduced) at 19 months of age (consistent with Mani and Plunkett (2011); Arias-Trejo and Plunkett (2009) or that phoneme-level inhibition should be removed (consistent with McMurray et al., 2009).

An alternative possibility is that the apparent asymmetry between cohort sizes used in (White & Morgan, 2008) is illusory. It is recognised that parental reports provide under-estimates of actual vocabulary sizes (Mayor & Plunkett, 2011). A proper estimate of vocabulary composition

may result in a more balanced lexicon structure, in turn reducing the impact of cohort imbalance disrupting the graded sensitivity to mispronunciation severity. However, an analysis of a dense recording at 30 months of age, the Haggerty corpus (Haggerty, 1929), revealed that /b/-onset words (89 words) are almost twice as numerous as /p/-onset words (48 words). Better descriptions of the lexical composition in infancy would no doubt help refine the distribution of cohort sizes associated with different onsets. However, they are unlikely to reveal an even profile in cohort sizes.

Taken individually, neither a reduction in lexical-level inhibition, the removal of phoneme-level inhibition, nor a finer-grained estimate of vocabulary composition in infancy can fully account for the graded sensitivity to mispronunciations described in (White & Morgan, 2008) while also capturing the findings that both onset consonant and medial-vowel mispronunciations lead to a reduction in target preferences reported by Mani and Plunkett (2007)<sup>4</sup>. A proper explanation of both phenomena will likely incorporate all of these explanations to a certain degree.

In an attempt to adjudicate between these different hypotheses, or to confirm the contribution of multiple contributing factors (reduction in overall inhibition and a slightly more balanced lexicon), one might ask whether 24-month olds would also display graded sensitivity to the severity of mispronunciations. Indeed, an important prediction of the TRACE simulation of White & Morgan’s (2008) results is that graded sensitivity to mispronunciation severity will be affected by cohort and neighbourhood effects if lexical competition is active. We justified switching off lexical competition in the model on the grounds that empirical studies have reported lexical island effects and lack of cohort effects with 18-month old infants (Arias-Trejo & Plunkett, 2009; Mani & Plunkett, 2011). However, these studies also report that lexical competition effects are apparent by 24-months of age. If the lexical-level inhibition hypothesis holds, we would predict, therefore, that when a task like White & Morgan’s (2008) study is conducted with 24-month-old infants, then the impact of severity of mispronunciation is likely to diminish when using the same stimuli.

It is noteworthy that the acceleration of rapid word learning, often dubbed the “vocabulary spurt” (Bloom, 1973), between 18 and 21 months of age coincides with the potential emergence of lexical competition. Of course, TRACE only implements lexical competition at a phonological level. Lexico-semantic competition, which is outside the purview of TRACE, may follow a different developmental trajectory and lead to different patterns of competition.

Finally, it should be noted that many simplifying assumptions were adopted in the simulations reported in this research. The dictionaries used in the simulations were created by assessing typical vocabularies as assessed by the Oxford CDI (Hamilton et al., 2000). However, individual differences in lexicon sizes and composition would lead to a

<sup>4</sup>A full analysis is reported in (Mayor & Plunkett, *Submitted*).

distribution of phonological sensitivities and looking patterns rather than a single uniform result in TRACE for a given age group. Moreover, the nonlinear impact of lexical competition in TRACE implies that a mean looking pattern based on a mean lexicon would not match the mean of looking patterns associated with different lexicon sizes. Fitting TRACE to individual lexicons rather than a standardised lexicon would provide yet another series of novel experimental predictions against which to evaluate the model.

### Acknowledgments

This work is supported by the Swiss National Science Foundation grant 131700 awarded to Julien Mayor and by the Economic and Social Research Council Grant RES-062-23-0194 awarded to Kim Plunkett.

### References

- Allopenna, P., Magnuson, J., & Tanenhaus, M. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, 38(4), 419–439.
- Arias-Trejo, N., & Plunkett, K. (2009). Lexical-semantic priming effects in infancy. *Philosophical Transactions of the Royal Society B*, 364, 3633–3647.
- Bloom, L. (1973). *One word at a time: The use of single word utterances*. The Hague: Mouton.
- Cutler, E. (1995). *Spoken-word recognition*. San Diego: Academic Press.
- Dahan, D., Magnuson, J., & Tanenhaus, M. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, 42(4), 317–367.
- Fenson, L., Dale, P., Reznick, S., Thal, D., Bates, E., Hartung, J., et al. (1993). *Macarthur communicative development inventories: User's guide and technical manual*. San Diego: Singular Press.
- Gaskell, M., & Marslen-Wilson, W. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and cognitive Processes*, 12(5-6), 613–656.
- Goldinger, S., Luce, P., & Pisoni, D. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, 28(5), 501–518.
- Golinkoff, R., Hirsh-Pasek, K., Cauley, K., & Gordon, L. (1987). The eyes have it: lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, 14, 23–46.
- Haggerty, L. (1929). What a two-and-one-half-year-old child said in one day. *Journal of Genetic Psychology*, 38, 75–100.
- Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a British communicative development inventory. *Journal of Child Language*, 27, 689–705.
- Jusczyk, P., & Aslin, R. N. (1995). Infant's detection of sound patterns of words in fluent speech. *Cognitive Psychology*, 29, 1–23.
- Luce, R. (1959). *Individual choice behavior*. Wiley New York.
- MacKay, D. (1982). The problems of flexibility, fluency, and speed-accuracy trade-off in skilled behavior. *Psychological Review*, 89(5), 483–506.
- MacWhinney, B. (1991). *The CHILDES project : Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Mani, N., & Plunkett, K. (2007). Phonological specificity of vowels and consonants in early lexical representations. *Journal of Memory and Language*, 57(2), 252–272.
- Mani, N., & Plunkett, K. (2011). Phonological priming and cohort effects in toddlers. *Cognition*.
- Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10, 29–63.
- Mayor, J., & Plunkett, K. (2011). A statistical estimate of infant and toddler vocabulary size from cdi analysis. *Developmental Science*, 14(4), 769–785.
- Mayor, J., & Plunkett, K. (Submitted). *Infant Word Recognition: Insights from TRACE Simulations*.
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1–86.
- McMurray, B., Samelson, V., Lee, S., & Bruce Tomblin, J. (2010). Individual differences in online spoken word recognition: Implications for sli. *Cognitive psychology*, 60(1), 1–39.
- McMurray, B., Tanenhaus, M., & Aslin, R. (2009). Within-category voicing affects recovery from. *Journal of memory and language*, 60(1), 65–91.
- Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than word learning tasks. *Nature*, 388, 381–382.
- Strauss, T., Harris, H., & Magnuson, J. (2007). jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behavior Research Methods*, 39(1), 19.
- Styles, S., & Plunkett, K. (2009). How do infants build a semantic system? *Language and Cognition*, 1, 1–24.
- Swingle, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, 76, 147–166.
- Theakston, A., Lieven, E., Pine, J., & Rowland, C. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*, 28(01), 127–152.
- White, K., & Morgan, J. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, 59, 114–132.

# Causal Status meets Coherence: The Explanatory Role of Causal Models in Categorization

Ralf Mayrhofer (rmayrho@uni-goettingen.de)

Anselm Rothe (anselm.rothe@stud.uni-goettingen.de)

Department of Psychology, University of Göttingen,  
Goßlerstraße 14, 37073 Göttingen, Germany

## Abstract

Research on causal-based categorization has found two competing effects: According to the causal-status hypothesis, people consider causally central features more than less central ones. In contrast, people often focus upon feature patterns that are coherent with the category's causal model (coherence hypothesis). Following up on the proposal that categorization can be seen as inference to the best explanation (e.g., Murphy & Medin, 1985), we propose that causal models might serve different explanatory roles. First, a causal model can serve as an explanation why the prototype of a category is as it is. Second, a causal model can also serve as an explanation why an exemplar might deviate from the prototype. In an experiment, we manipulated whether typical or atypical features were linked by causal mechanism. We found a causal-status effect in the first case and a coherence effect in the latter one, suggesting both are faces of the same coin.

**Keywords:** categorization; causal reasoning; causal status effect; coherence effect; explanation.

## Introduction

The question how people organize objects into categories and form abstract concepts about the world to make sense of it has puzzled philosophers for centuries. It is therefore not surprising that categorization has been an important topic in cognitive science since its beginnings. Early but nevertheless prominent accounts concentrated on the role of similarity between exemplars, or exemplars and category prototypes, or rules with respect to defining features of a category (e.g., Nosofsky, 1986; Rosch & Mervis, 1975; for an overview see Ashby & Maddox, 2005). In contrast, more recent accounts emphasize the role of abstract conceptual, mostly causal knowledge as an integral part of category representations (see Murphy & Medin, 1985; Rehder, 2010; Rehder & Hastie, 2001; Sloman, Love, & Ahn, 1998): People do not only know which features are typical for a category and which not. They often represent knowledge about how strongly and *why* features are correlated with each other within a category (Ahn, Marsh, Luhmann, & Lee, 2002; Murphy & Medin, 1985). For instance, people do not only know that birds typically have wings, can fly, and build nests on trees. People also know that birds build nests on trees because they can fly and that they can fly because they have wings.

This kind of causal knowledge underlying category concepts can be formalized in causal graphical models or Bayes nets (see Rehder, 2003a, 2003b; Waldmann, Holyoak, &

Fratianne, 1995). A causal Bayes net consists of nodes, which represent causally relevant variables (i.e., in case of categorization: the presence or absence of features or—more general—properties of objects), and arrows, which stand for counterfactual or statistical dependencies between these variables. The arrows are placeholders for underlying causal mechanisms (Pearl, 2000) and render the variables into causes and effects. Figure 1 shows an example of a common-cause network that relates a cause feature  $F_C$  to three effect features  $F_{E1}$ ,  $F_{E2}$ , and  $F_{E3}$ . The features of a category are usually coded such that the typical feature value is 1 (i.e., presence) and the atypical value is 0 (i.e., absence).

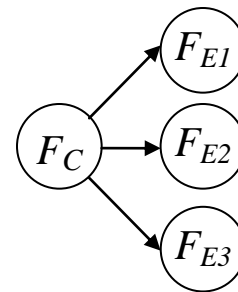


Figure 1: An example of a simple common-cause structure that connects a cause feature  $F_C$  with three effect features  $F_{E1}$ ,  $F_{E2}$ , and  $F_{E3}$ . Due to the causal relations, the state of each effect feature depends counterfactually or statistically upon the state of the cause feature.

Nowadays, it's quite uncontroversial that causal knowledge is an important part of people's concepts that underlie category representation (see Rehder, 2010, for a review). But it is still in controversial debate *how* causal knowledge affects the classification of objects.

In a typical causal-based categorization task people are introduced to a target category that possesses a set of mostly three or four features. In addition, it is pointed out how these features are causally related to each other due to some causal mechanisms that hold for the category (e.g., a common-cause model as shown in Figure 1). Then, participants are presented with several potential exemplars with the category's features being either present or absent. For each of the presented exemplars, membership ratings are obtained. The enduring controversy, then, spans around the question how the instructed causal model interacts with the presence and

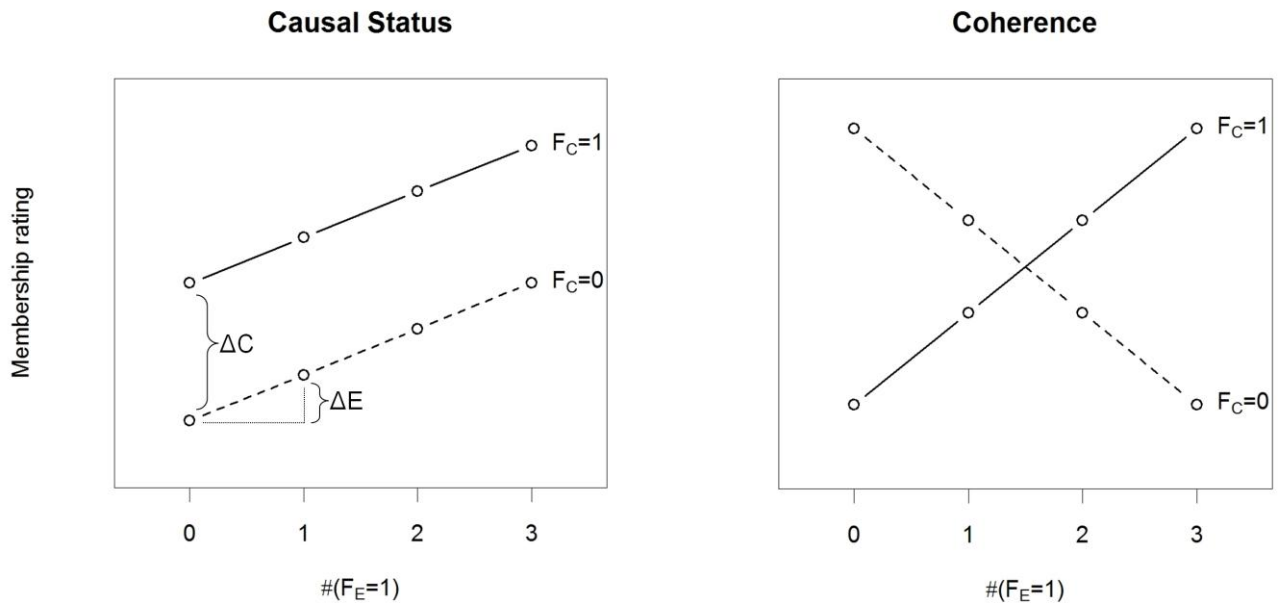


Figure 2. Predicted patterns for category membership ratings are shown according to (a) the *causal status hypothesis* and (b) the *coherence hypothesis*. The predicted ratings are computed for a category possessing four features that are connected as a common-cause model (as shown in Figure 1). The ratings depend upon the presence of the cause feature ( $F_C = 0$  vs.  $F_C = 1$ ; dashed vs. solid lines) and upon the number of effect features being present,  $\#(F_E = 1) = \{0, 1, 2, 3\}$ .

absence of features with respect to the membership ratings of the presented exemplars.

Some researchers propose that causal knowledge is an important determinant of individual feature weights in these judgments (e.g., Ahn, Kim, Lassaline, & Dennis, 2000; Marsh & Ahn, 2006). According to this account, the *causal status* of a feature matters. Others, however, emphasize the importance of feature configurations that are more or less *coherent* with the category's causal model (Rehder, 2003a, 2003b; Rehder & Hastie, 2001; Rehder & Kim, 2010). Although the effects are conceptually independent of each other (Rehder, 2010), both sides claim (and have—puzzlingly—shown empirically) that the other effect plays only a marginal, if any, role in human categorization.

In the following sections, we will describe the causal status hypothesis and the coherence hypothesis in more detail. Then, we will offer an account that—in our view—makes sense of the diverging evidence, and present an experiment that tests our claims.

### Causal status effect

According to the causal status hypothesis, features that are causally more central (i.e., that have more dependents in the causal model or that appear earlier in a causal chain) have greater influence in categorization decisions when other perceptual (e.g., salience) or statistical (e.g., cue validity) properties of the material are held constant or controlled for (e.g., Ahn et al., 2000). The causal status of a feature is, therefore, an important determinant of its decision weight.

For example, with respect to a common-cause model (see Figure 1), the presence vs. absence of the cause feature  $F_C$  should have more influence on the membership rating of an exemplar than the presence vs. absence of an effect feature. In Figure 2a, such an idealized causal status effect is shown: The membership ratings increase with the number of features being present; the increase, however, is higher for the cause feature ( $\Delta C$ ), than for effect features ( $\Delta E$ ).

Conceptually, the causal status effect has been linked to psychological essentialism (Ahn et al., 2000). Hence, people believe in things having essences that make them the objects they are. An essence, then, is the (unobservable) root cause for the surface features that can be observed in category members (Gelman & Wellman, 1991). Features that have a high causal status might be seen as most diagnostic for an object's essence and, therefore, category membership (Ahn et al., 2000; see also Rehder & Kim, 2010).

### Coherence effect

Whereas the causal status effect is defined with respect to the weight of individual features, the coherence effect arises from the impact of feature interactions (Rehder, 2010). According to the coherence hypothesis, exemplars whose feature configurations are most coherent with the category's causal model are seen as the best members. Causal models, therefore, provide us with information about which features should go together in an exemplar. Features that are connected by a causal link should be either both present, or both absent (Rehder, 2003a; Rehder, 2010). With respect to a

common-cause model (see Figure 1), for example, membership ratings should be an increasing function of the number of effects being present when the cause feature is present, but a decreasing function in its absence. Figure 2b shows such an idealized coherence effect. Membership ratings are expected to be highest when all features are either present or absent. In this case, all causal links are preserved (i.e., such an exemplar is most coherent). The worst (i.e., most incoherent) exemplars, in contrast, are those that preserve none of the links: The cause feature is present but all effect features are absent, or the cause feature is absent but all effect features are present (three violated links in both cases).

Coherence in the proposed manner, however, faces a problem when assessed intuitively in real world cases that pop into one's mind: It does not make sense. An animal that does not have wings, does not fly, and does not build nests on trees is not a bird, although the absence of these features is perfectly coherent with the causal model of the concept "bird". Marsh and Ahn (2006) therefore suggested that the coherence effect may be not more than an experimental artifact arising from the artificial material used by Rehder and colleagues (e.g., Rehder 2003a, 2003b; Rehder & Hastie, 2001). Nevertheless, we propose otherwise.

### Explanatory roles of causal models

So far, the whole debate has neglected the fact that causal models may play different roles in the representation of concepts that underlie categories. Causation in those models is implemented (or thought) in a way that causes, when present, have the power to bring about their effects, but are causally inactive when absent (Cheng, 1997; Rehder, 2003a). At first glance, this might not matter anyway: Usually, the presence of the cause goes together with the presence of its effects, as well as the absence of the cause goes together with the absence of its effects (Note, that this superficial symmetry is the basis for the coherence hypothesis). However, whereas the first fact is a direct consequence of causal mechanism, the latter is just an indirect "side effect" of it (Dowe, 2000, aptly mentioned that in this case the absence of the cause prevents the presence of its effect by omission, i.e., just by not causing it). Although this difference hasn't received much attention yet, we think it is crucial for understanding causal-based categorization.

Categorization can be seen as a kind of inference to the best explanation (Jameson & Gentner, 2008; Lombrozzo, 2009; Murphy & Medin, 1985; Rips, 1989), according to which causal models provide a system of explanatory links that tie the features of a category together. Therefore, exemplars whose configuration of features can be best explained in the light of the category's causal concept are rated as the best members. With respect to the causal analysis given above, we propose—in contrast to the coherence hypothesis—that only those feature combinations matter that are relevant with respect to the underlying causal mechanisms (e.g., when both are present, but not when both are absent), because it is the mechanism but not the regularity that has explanatory value (see Keil, 2006, for an overview).

From this point of view, we can at least differentiate two explanatory roles of causal models. First, when causal mechanisms are established in terms of typical features, the causal model serves as an explanation why the category's prototype or prototypical exemplar (i.e., all features present) is as it is. The bird example given above belongs to this type of explanation. With respect to a common-cause model (see Figure 1), we would expect a strong increase of membership ratings with more effect features being present when the cause feature is also present. In this case more and more explanatory links are served (i.e., the exemplar becomes more and more coherent with the provided explanation). But when the cause is absent, the effect features are conceptually unrelated to each other. Therefore, we would expect a much smaller increase when more and more effects are present. Since the presence vs. absence of the cause feature modulates the positive influence of the effect features, we expect membership ratings that—in the aggregate—exhibit a strong causal status effect.

Second, however, it is also possible to establish causal mechanisms with respect to atypical feature values (usually coded as absences). In this case, the causal model serves as an explanation for why a category member might deviate from the category prototype (e.g., fouling that makes an apple not looking like an apple anymore). When now presented with an exemplar that lacks all typical features, you would probably be much more willing to judge this exemplar a category member than in the bird example, because the causal model provided you with an explanation why this atypical exemplar deviates from the prototype. Thus, in case the causal model links atypical feature values, we expect a pattern that looks quite like the prediction of the coherence hypothesis (see Figure 2b). First related evidence for this proposal comes from Ahn, Novick, and Kim (2003): In their studies, participants judged persons who showed a set of abnormal characteristics (e.g., suffering from insomnia, memory deficits, and episodes of extreme anxiety) as more "normal" when provided with plausible causal relations between these abnormal characteristics compared to a condition in which no such links were provided.

To summarize, we believe that the diverging evidence found in the literature regarding the causal status and the coherence effect stems from the fact that causal models play different roles in categorization and that those studies might differ with respect to the explanatory role of the instructed causal model. In the next section we present an experiment that tests our claim.

### Experiment

To test our hypothesis we adapted the material used in the experiments of Rehder (2003a; in similar versions also used in Marsh & Ahn, 2006; Rehder, 2003b; Rehder & Kim, 2008, 2010, and others). Rehder presented subjects with instructions about several artificial categories (e.g., Kehoe Ants, Mya Stars) possessing four features that were linked in a common-cause model (see Figure 1). Features were introduced without giving precise base rate information



(e.g., “Some Kehoe Ants have blood that is very high in iron sulfate. Others have blood that has normal levels of iron sulfate.”, “Some Kehoe Ants have an immune system that is hyperactive. Others have a normal immune system.”)<sup>1</sup>. Then, causal mechanisms were introduced by plausible descriptions (e.g., “Blood high in iron sulfate causes a hyperactive immune system. The iron sulfate molecules are detected as foreign by the immune system, and the immune system is highly active as a result.”). After participants have learned the category, they had to rate all possible exemplars (all possible combinations of features being present or absent) regarding their category membership. In his studies, Rehder found evidence for the coherence effect (but see Marsh & Ahn, 2006, for a critical discussion).

In our experiment, we used the same procedure and same material. To manipulate the explanatory role of the instructed causal model, we explicitly instructed which of the feature values were described as *typical* for the category (e.g., hyperactive immune system or normal immune system). So, between conditions, the typicality of the feature values changed but the description of causal mechanisms remained constant for the same feature values. By that, however, we manipulated the explanatory role of the causal mechanisms (i.e., whether *typical* or *atypical* values were linked by mechanisms). Furthermore, we added a replication condition that was identical to Rehder (2003a), to ensure that our procedure (and translated material) yields the same findings.

## Method

**Participants** 96 students (62 women, mean age 22.4 years) from the University of Göttingen, Germany, participated in this experiment as part of a series of various unrelated computer-based experiments in our computer lab. Participants received either course credit or were paid €8 per hour.

**Material** Two categories used in Rehder (2003a) were translated into German: Kehoe Ants (a biological kind) and Mya Stars (a non-living natural kind).<sup>2</sup> Each category possessed four binary features. Depending on condition, each feature had a typical value (coded throughout this paper as “1”) and an atypical value (coded as “0”). For example, Kehoe Ants have an immune system that was either hyperactive or working normal. Which of the two values was described as typical depended on the experimental condition (e.g., in the *typical* condition, it was stated: “Typically, Kehoe Ants have an immune system that is hyperactive. A few have a normal immune system.”, in the *atypical* condition, it was stated: “Typically, Kehoe Ants have a normal immune system. A few have an immune system that is hyperactive.”).

Additionally, the features were causally linked in a common-cause network. Each causal relationship was described as one feature causing another in the same way Rehder

(2003a) did (see above). The description of causal mechanism was identical for all conditions.

**Procedure** Participants were randomly assigned to one of the two categories and to the *typical*, *atypical*, or *replication* condition. They completed the experiment individually on desktop computers. The experiment consisted of two phases, an instruction phase and a test phase.

In the instruction phase, we presented subjects with information about the category (Kehoe Ants, Mya Stars). Subjects were introduced to the four binary features and their typical values (depending on *typical* or *atypical* condition, respectively). Then, subjects were provided with information about how the features are causally connected. (As stated above, in all conditions the causal links were instructed between the same feature values. However, the typicality of these feature values and, therefore, the explanatory role of the causal model differed depending on condition.) In the *replication* condition the causal links were presented in the same way, but no information about the typicality of the values was given (as in Rehder, 2003a). The instructions were followed by a multiple choice test in which participants were required to demonstrate that they had learned all given information about the assigned category. In case of incorrect answers they had to reread the instructions and had to take the test again until they committed 0 errors.

In the test phase, subjects were presented sequentially with all 16 possible exemplars (all combinations of the four binary features) in two consecutive blocks. Order of exemplars was randomized in each block. For each exemplar, subjects were requested to give a category membership

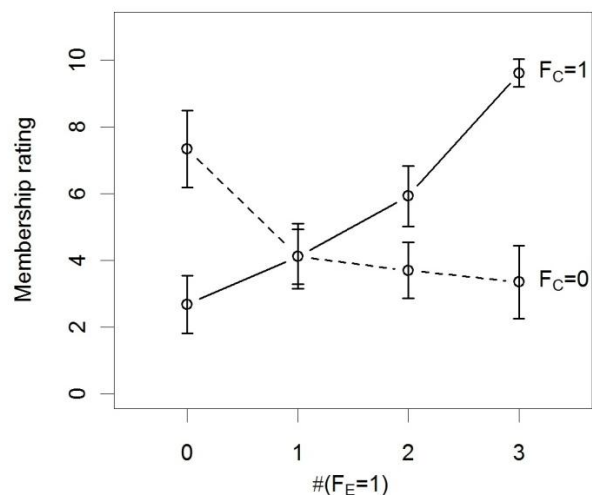


Figure 3. Results of the replication of the experiment of Rehder (2003a). Membership ratings of exemplars are shown with cause feature present ( $F_C=1$ ) vs. absent ( $F_C=0$ ). X-axis displays the number of effect features being present (i.e., having typical value). Error bars indicate 95% confidence intervals.

<sup>1</sup> Note that normal values were coded as “0” (= absence).

<sup>2</sup> Rehder (2003a) used six categories, but no differences for membership ratings were obtained. Therefore, we only used two randomly chosen categories.

rating on a scale from 0 (not a member at all) to 10 (definitely a member).

**Design** The membership ratings were aggregated for each subject across blocks and with respect to the cause feature being present ( $F_C = 0$  vs.  $F_C = 1$ ) and the number of effect features being present ( $\#F_E = 1 = \{0, 1, 2, \text{ or } 3\}$ ). This yielded a 3 (*typical*, *atypical*, *replication* condition)  $\times$  2 (category: Kehoe Ants vs. Mya Stars)  $\times$  2 ( $F_C = 0$  vs.  $F_C = 1$ )  $\times$  4 ( $\#F_E = 1 = \{0, 1, 2, \text{ or } 3\}$ ) ANOVA design with condition and category as between-subjects factors and the presence of the cause or effect features, respectively, as within-subjects factors and average membership rating as dependent variable.

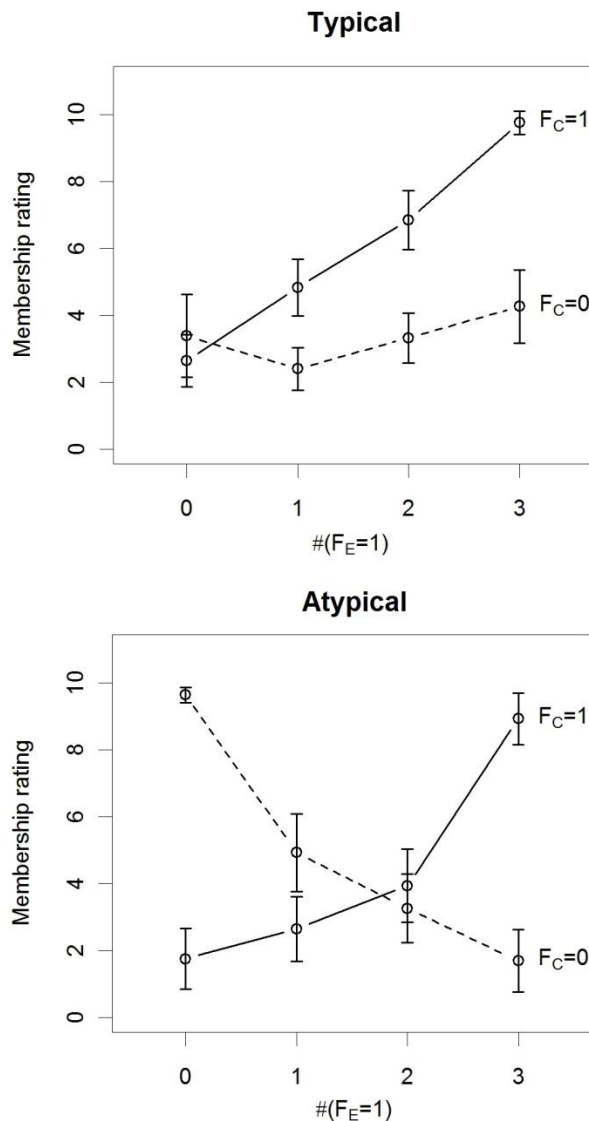


Figure 4. Average category membership ratings are shown for (a) *typical* condition and (b) *atypical* condition with cause feature present ( $F_C=1$ ) vs. absent ( $F_C=0$ ). X-axis displays the number of effect features being present (i.e., having typical value). Error bars indicate 95% confidence intervals.

## Results

**Category** In line with previous studies, the category (Kehoe ants, Mya stars) revealed no significant influence on membership ratings, neither as main effect nor in any interaction (all  $ps > .1$ ). Therefore, the factor is aggregated over in the following analyses.

**Replication** In Figure 3, the results of the replication condition are shown. When the cause feature was present ( $F_C = 1$ ) membership ratings increased with increasing number of effects being present. In contrast, when the cause feature was absent ( $F_C = 0$ ) membership ratings decreased, yielding a significant interaction ( $F_{3,99}=62.5, p < .001$ ). This replicates the findings of Rehder (2003a) and is in line with other studies using the same material.

**Explanatory Role** Figure 4 displays the aggregated membership ratings for the *typical* and *atypical* condition. In both *typical* and *atypical* conditions, subjects rated exemplars with a present cause feature ( $F_C = 1$ ) better members the more typical effect features were present. So, the exemplar with all features being present was rated very high (9.8 and 8.9 in *typical* and *atypical* condition, respectively) whereas the exemplar with all effects being absent was rated very low (2.6 and 1.7, respectively).

Exemplars, however, with the cause feature being absent ( $F_C = 0$ ) exhibit a significant interaction between conditions ( $F_{3,198}=28.2, p < .001$ ) (The three-way interaction was also significant,  $F_{3,198}=17.79, p < .001$ ). In *typical* condition, ratings' increase was only marginal significant ( $F_{3,102}=2.24, p=.088$ ). In *atypical* condition, subjects rated exemplars lower, the more effect features expressed the typical value. So, the exemplar with all effects being absent (and, therefore, all features being absent) was rated very high (9.6), whereas the exemplar with all effects being present was rated very low (1.7).

Thus, the *atypical* condition looks like the prototypical case of a coherence effect. In fact, individual influence of features (i.e., marginalized across the states of the other features) are negligible and even negative ( $\Delta C = -0.67, \Delta E = -0.14$ ). In contrast, the *typical* condition revealed a strong causal status effect ( $\Delta C = 2.83$  vs.  $\Delta E = 1.37$ ), as we have predicted.

## Discussion & Summary

It is widely accepted that causal knowledge is an important part of people's concepts that underlie category representations. Nevertheless, it is still quite controversial how causal knowledge affects membership ratings: Some researchers propose that causal knowledge determines the individual feature weights in categorization judgments (causal status hypothesis; see Ahn et al., 2000), whereas others, however, emphasize the role of feature combinations and whether those are coherent with the statistical regularities imposed by the category's causal model (coherence hypothesis; see e.g., Rehder, 2003a, 2003b; Rehder & Kim, 2010). We presented one possible solution to this puzzle: According to

our proposal, causal knowledge is important in categorization because it provides people with explanatory links such that they can make sense of presented exemplars. Thus, categorization is seen as inference to the best explanation (Murphy & Medin, 1985; Rips, 1989). And because the explanatory value of causal knowledge is engrained in people's beliefs about underlying mechanisms (and not statistical regularities), we derived at least two possible explanatory roles of causal models: First, a causal model can serve as explanation why a prototypical exemplar is as it is (e.g., why most birds can fly). Second, a causal model can also serve as explanation why a category member might deviate from the prototypical exemplar (e.g., why some birds cannot fly). Depending on which kind of causal model people have in mind for a given category we expect people to judge different exemplars as best and worst category members.

We presented an experiment in which we manipulated the explanatory role of the instructed causal knowledge directly, and we found huge differences in membership ratings. Interestingly, in the *typical* condition (i.e., typical feature values were linked by causal mechanisms) judgments exhibited a causal status effect. In contrast, in the *atypical* condition (i.e., atypical feature values were link by causal mechanisms) we found a strong coherence effect. Therefore, we believe that causal-status as well as coherence effects are both faces of the same coin.

### Acknowledgments

This research was supported by a research grant of the Deutsche Forschungsgemeinschaft (DFG Wa 621/20).

### References

- Ahn, W.-K., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, 41, 361-416.
- Ahn, W.-K., Marsh, J. K., Luhmann, C. C., & Lee, K. (2002). Effect of theory-based feature correlations on typicality judgments. *Memory & Cognition*, 30, 107-118.
- Ahn, W.-K., Novick, L. R., & Kim, N. S. (2003). Understanding behavior makes it more normal. *Psychonomic Bulletin & Review*, 10, 746-752.
- Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, 56, 149-178.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104, 367-405.
- Dowe, P. (2000). *Physical causation*. Cambridge, UK: Cambridge University Press.
- Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the nonobvious. *Cognition*, 38, 213-244.
- Jameson, J., & Gentner, D. (2008). Causal status and explanatory goodness in categorization. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 291-296).
- Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, 57, 227-254.
- Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?" *Cognition*, 110, 248-253.
- Marsh, J. K., & Ahn, W.-K. (2006). The role of causal status versus inter-feature links in feature weighting. *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 561-566).
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289-316.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification categorization relationship. *JEP: General*, 115, 39-57.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, MA: Cambridge University Press.
- Rehder, B. (2003a). A causal-model theory of conceptual representation and categorization. *JEP:LMC*, 29, 1141-1159.
- Rehder, B. (2003b). Categorization as causal reasoning. *Cognitive Science*, 27, 709-748.
- Rehder, B. (2010). Causal-based classification: A review. In B. Ross (Ed.), *The Psychology of Learning and Motivation* (52), 39-116.
- Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *JEP: General*, 130, 323-360.
- Rehder, B., & Kim, S. (2008). The role of coherence in causal-based categorization. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 285-290).
- Rehder, B., & Kim, S. (2010). Causal status and coherence in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 1171-1206.
- Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21-59). Cambridge, UK: Cambridge University Press.
- Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573-605.
- Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, 22, 189-228.
- Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal knowledge and the acquisition of category structure. *JEP: General*, 124, 181-206.

# Sparse category labels obstruct generalization of category membership

John V. McDonnell (john.mcdonnell@nyu.edu)

Carol A. Jew (carol.jew@nyu.edu)

Todd M. Gureckis (todd.gureckis@nyu.edu)

New York University, Department of Psychology

6 Washington Place, New York, NY 10003 USA

## Abstract

Studies of human category learning typically focus on situations where explicit category labels accompany each example (supervised learning) or on situations where people must infer category structure entirely from the distribution of unlabeled examples (unsupervised learning). However, real-world category learning likely involves a mixture of both types of learning (semi-supervised learning). Surprisingly, a number of recent findings suggest that people have difficulty learning in semi-supervised tasks. To further explore this issue, we devised a category learning task in which the distribution of labeled and unlabeled items suggested alternative organizations of a category. This design allowed us to determine whether learners combined information from both types of episodes via their patterns of generalization at test. In contrast with the prediction of many models, we find little evidence that unlabeled items influenced categorization behavior when labeled items were also present. **Keywords:** Semi-supervised category learning; rule induction; unsupervised learning

## Introduction

Category learning is a critical cognitive ability which is central to many aspects of cognition. As a result, considerable research over the last 50–60 years has explored the psychology of category learning using laboratory tasks. The majority of this work can be divided into two groups. Most research has focused on *supervised learning* tasks where corrective feedback or category labels are presented following or alongside each observation of a stimulus (e.g., Medin & Schaffer, 1978; Nosofsky, 1986). More recently, there has been an interest in *unsupervised learning*, wherein participants must organize examples in the absence of explicit instruction using the distributional properties of the stimuli (e.g., Clapper & Bower, 1994; Love, 2002; Pothos et al., 2011). However, neither of these situations adequately reflect the problem of real world category learning, in which feedback is not altogether absent nor always present, but is typically sparse and intermittent. Such tasks require learners to combine information from both labeled and unlabeled episodes. In machine learning, this problem is frequently studied under the name *semi-supervised learning* (for review, see Zhu, 2005).

Aside from offering a more ecologically relevant approach to the study of category learning, the study of semi-supervised learning has important implications for theories of human concept learning. Consider the problem of learning a concrete noun such as *horse*. One proposal is that word learning essentially links sound tokens (words) to already-acquired hypotheses or representations (Bloom, 2000; Gentner, 1982). Under this view, the label information from a teacher or parent about a single example horse must be integrated with the

child's pre-linguistic grouping of objects in their environment into classes.

A similar position is advocated by a number of influential theories of category learning which hold that supervised and unsupervised learning are subserved by a single underlying learning process (e.g., the rational model of categorization, Anderson, 1991; or the Supervised and Unsupervised STRatified Adaptive Incremental Network, abbreviated *SUSTAIN*, Love, Medin, & Gureckis, 2004). Such models naturally predict that semi-supervised learning should not only be possible, but may be the primary way in which people learn categories and their respective names.

## Can people acquire categories via semi-supervised learning?

Despite these arguments, recent empirical attempts to demonstrate semi-supervised category learning in the lab have met with mixed success. For example, Vandist, De Schryver, and Rosseel (2009) found that adding unlabeled training examples to a mostly supervised task offered no additional benefit beyond learning from only the supervised trials. However, the category structures they tested (known as *Information-Integration* categories) are typically difficult for people to learn even in fully unsupervised settings (Ashby, Queller, & Berretty, 1999), which may explain the limited impact that the unlabeled examples had.

On the other hand, Kalish, Rogers, Lang, and Zhu (2011) showed that after learning a simple category distinction on a single dimension from a small set of labeled examples, participants' estimate of the category boundary could be shifted by the presentation of a large number of unlabeled examples whose distribution was shifted compared to the labeled set (see also Lake & McClelland, 2011). While this study provides some evidence of semi-supervised learning, there remain alternative explanations of the effect. For example, since the central tendency of both categories are shifted in these studies it is unclear whether people are separately updating each category representation or responding to the global shift in the stimulus space.

Finally, Rogers, Kalish, Gibson, Harrison, and Zhu (2010) compared learning in a semi-supervised learning condition with a fully supervised condition. In this study, adding unlabeled items to a supervised category learning task caused faster learning only when trials were speeded. However, the question of whether people can integrate labeled and unlabeled training examples is logically separate from claims about

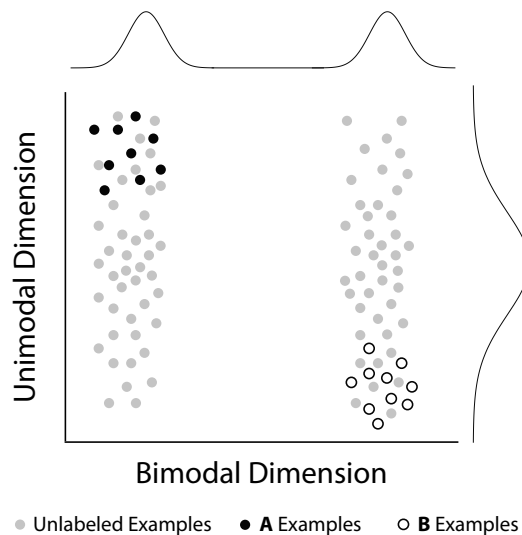


Figure 1: A schematic depiction of the design used in the experiment. Category stimuli varied along two continuous dimensions. The plot edges represent the marginal distribution of examples. Unlabeled examples fall in two columnar clusters, while two clusters of examples with labels A and B appear in the corners of the space. Taken alone, the distribution of labeled examples is ambiguous concerning how to generalize, since a rule on either dimension alone could explain the labels.

learning rates between tasks. For example, a participant's learning rate might vary based on features of the overall task context rather than the information conveyed by any subset of examples.

Collectively, these results tell a surprisingly unclear story. Despite decades of research on supervised and unsupervised learning with artificial stimuli, studies which have attempted to combine these two forms of learning fail to show robust and consistent effects. Some find limited evidence of semi-supervised learning while others fail to find any evidence at all. The goal of the present study is to attempt to revisit this issue with a novel design which may be more diagnostic of semi-supervised learning. As will be revealed shortly, our results add modest light to an already murky picture.

### Evaluating semi-supervised learning through patterns of generalization at test.

Our study (summarized abstractly in Figure 1) departs from the studies described above in a number of ways. In some previous work, the distributional properties of both labeled and unlabeled examples were identical (e.g., Vandist et al., 2009). In contrast, we manipulated the distribution of examples so that the distribution of unlabeled examples and the distribution of labeled examples suggested alternative organizations of the category. In particular, the labeled items alone were ambiguous about the basis for the category difference. However, the distribution of unlabeled examples suggested a clear organization of the categories along a single dimension. Our

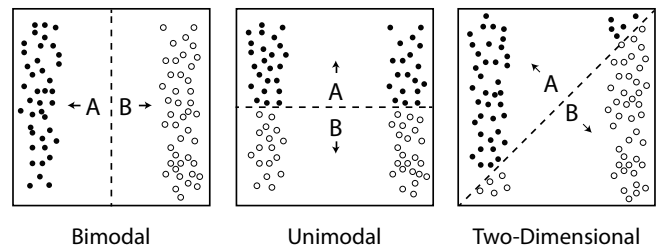


Figure 2: A range of possible category strategies consistent with the category in Figure 1. In the Bimodal strategy, the learner classifies all the items that fall in each clustered column with the label given to the labeled items within the column. In the Unimodal strategy, the learner divides the example along the unimodal dimension. This strategy is acceptable if the unlabeled examples are ignored. The 2D strategy is more complex (in the sense that it depends on attention to both stimulus features) but is also consistent with the labeled examples and inconsistent with the unlabeled distribution.

prediction was that if people combine information from both the labeled and unlabeled examples, they will generalize the label information according to the distribution implied by the unlabeled examples. This should be clearly captured in their patterns of generalization in a test phase (see Figure 2).

In addition to comparing semi-supervised learning to fully supervised learning we also include a second control condition assessing fully unsupervised learning. In fact, this condition of our study represents a conceptual replication of a previous study on unsupervised category learning by Zeithamova and Maddox (2009, henceforth referred to as Z&M). This served two purposes. First, it allows us to establish a baseline measure of behavior for both extremes of supervision. Second, this ensures that participants *can* learn the category from the distribution of unlabeled examples alone.

Finally, rather than test a single semi-supervised learning condition, we systematically explore the effect of the *number* of labeled examples on semi-supervised learning. Our design thus interpolates between fully unsupervised learning to fully supervised learning by changing the relative amount of labeled versus unlabeled information.

## The Experiment

In our experiment, we developed a cover story which provided a plausible explanation for why some category examples were unlabeled (but still came from the same category). The cover story asked participants to imagine that they were working in a tv repair shop in a town where people tuned special loop antennas to pick up one of only two possible channels (similar to Markant & Gureckis, 2010). Similar antennas tended to pick up similar channels. Although all the tvs were tuned to one of the two channels, many had broken tubes making it impossible to turn on and verify the channel. The participants' job over the course of the experiment was to determine how different settings of the antennas determine which channel the tv is tuned to pick up. They were reminded that learning about the antennas was possible even if the tv tube was broken.



The experiment was organized into two phases. The first was a training phase in which participants were shown various category members with and without labels (depending on condition). The second was a test phase in which participants were asked to classify novel examples. Decision bound models (Ashby, 1992) were fit to subjects' responses during the test phase in an attempt to infer the strategy they applied. We then analyzed the frequency by which different strategies were adopted as a function of condition.

## Methods

**Participants** 124 New York University undergraduates participated for course credit. Participants were randomly assigned to one of four possible conditions: Unlabeled ( $N = 33$ ), 10-Labeled ( $N = 31$ ), 40-Labeled ( $N = 30$ ), or 40-All-Labeled ( $N = 30$ ). Four participants, three in the Unlabeled Condition and one in the 10-Labeled Condition, were classified as responding randomly (see the results section) and were dropped from the analysis, leaving 30 participants in each condition.

**Materials** The objects to be categorized were line stimuli varying in their length and orientation. The stimulus properties (lengths and angles) of the antennas were chosen to be similar to those used by Z&M (2009). The range of possible angles was different for each subject, but it covered  $60^\circ$  and was constrained not to cross the vertical or horizontal axes. The range of lengths was always between 100 and 560 pixels. The line stimuli were attached to pictures of TV via a stem. Category label information was given by changing what was showing on the TV screen. For unlabeled examples, the screen took on the appearance of broken glass. Participants were told that these TVs were broken, but still tuned correctly to one of the two channels. When category label information was given, the letters CH1 or CH2 appeared on the screen, indicating that the TV was set to pick up one of the channels (see the top row of Table 1 for examples).

**Design** During the training phase, the TVs were drawn from two elongated distributions which were naturally separable along one of the stimulus dimensions (the *bimodal* dimension). For reasons of control, stimuli were sampled using a discrete binning method described in Figure 3. This differs slightly from Z&M (2009) who used bivariate normal distributions but was necessary to ensure tight control over the distributional properties of the stimuli, and in particular to ensure that the distributions of labeled items were unbiased with respect to the particular categorization strategies.

Our primary experimental manipulation was to alter the training that participants received in the task. Four training conditions ranging from completely unsupervised to completely supervised (with no unlabeled training items) were included (see Table 1 for a summary).

**Unlabeled Condition.** In this condition, all TVs were broken (i.e., unlabeled). This condition is a traditional unsupervised category learning task and a conceptual replication of the *intermixed* condition from Z&M (2009), Exp. 1A.

**10-Labeled Condition.** This condition was identical to the Unlabeled Condition except that ten of the items in the corners were presented along with category labels (i.e., the appropriate channel).

**40-Labeled Condition.** This condition was similar to the Unlabeled Condition and the 10-Labeled Condition except that all of the items in the corners (40 in total) were presented along with category labels.

**40-All-Labeled Condition.** In this condition, all antennas were labeled in the training phase, meaning that this condition was fully supervised. However, to hold other aspects of the task consistent with the other conditions, 240 broken TVs without antennas (sham trials) took the place of the unlabeled examples, giving participants in this condition the same number of training trials as those in other conditions, and similar temporal spacing between labeled items to participants in the 40-Labeled Condition.




	 Labeled	 Unlabeled	 Sham	Total
Unlabeled	0	280	0	280
10-Labeled	10	270	0	280
40-Labeled	40	240	0	280
40-All-Labeled	40	0	240	280

Table 1: Summary of the four training conditions. All participants viewed 280 items in the training condition. *Labeled* and *Unlabeled* here denotes a TV with an antenna, which were either working (labeled) or broken (unlabeled). A *sham* TV consisted of a broken TV set without an antenna (examples are provided along the top).

Regardless of condition, labeled items in the training phase always came from the corners of the space as depicted in Figure 3, which meant they were always non-diagnostic with respect to the best category rule.

The test phase was identical for all four groups and, following Z&M (2009), involved the presentation of 50 broken TVs sampled from the same distribution as used during training. Participants were asked to predict the channel based on the antenna setting.

All remaining arbitrary aspects of the design (e.g., which dimension served as the bimodal dimension) were counterbalanced across conditions.

**Procedure** The experiment was administered on standard Macintosh computers using an in-house data collection system written in Python<sup>1</sup>. Participants were tested over a single one-hour session.

The instructions emphasized that all of the antennas were in good working order and purposefully tuned either to CH1 or CH2. Participants were also told that, as a result, the antennas as a whole constituted two categories of items. Although many of the TVs were broken, being broken had nothing to do with the setting of the antenna or the potential to pick up one of the two channels. Broken TVs were missing some information, but were otherwise not different from the others. To confirm that they had understood the instructions, participants were given a brief quiz and misconceptions were addressed.

Next, participants observed 80 randomly generated antennas in quick succession (100ms each), giving information about the range of values for each of the two stimulus dimensions (angle and length).

On each trial of the training phase, participants viewed a new TV (which was broken, working, or a sham), and after 500ms were prompted to press the space bar to continue. After the button press, the stimulus remained on the screen for 500ms. Between trials the screen was blank for an inter-stimulus interval of 500ms.

The test phase consisted of 50 trials in which participants saw a broken TV drawn from the same distribution as the training trials. On each test trial, participants viewed a new broken TV and were asked to press a button on their keyboard to indicate whether they believed the antenna was tuned to CH1 or CH2. After each trial, a thank you message (along with the original stimulus) remained on the screen for 1000ms. No feedback was given. The next trial followed after 500ms.

## Results

**Accuracy Analysis** In our first analysis, we considered whether participants correctly applied the category labels in the unambiguous regions of the space (i.e., the corners) in the 10-Labeled, 40-Labeled, and 40-All-Labeled conditions. Responding was significantly above chance in all conditions: 10-

<sup>1</sup> Available at <http://www.pypsyexp.org>





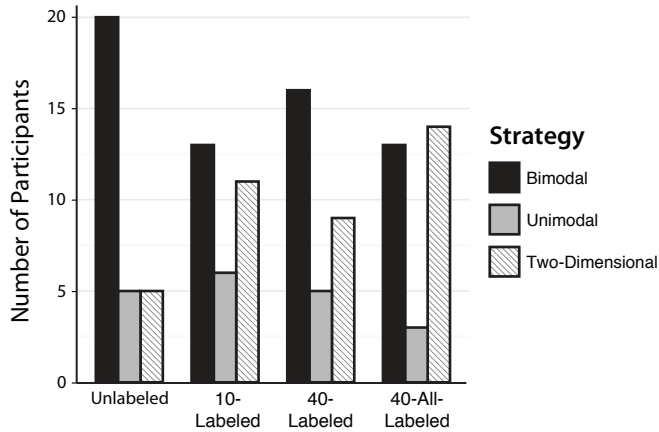


Figure 5: Histogram showing the trend in the number of subjects adopting each strategy across conditions. There is a general trend of an increase in the use of 2D rules in the presence of labeled items, as well as a drop in the use of a 1D rule on the bimodal dimension, but trends among the labeled conditions were weak.

the same bias can be evaluated in other conditions. Overall, the difference between the Unlabeled and 40-All-Labeled conditions was driven largely by the increase in the use of 2D rules in the 40-All-Labeled Condition. Note that while the distribution of labeled examples in the 40-All-Labeled Condition is logically consistent with all three strategies (Unimodal, Bimodal, 2D), a bias toward 2D rules is in line with the predictions of the optimal linear discriminant for the labeled examples.

Turning to the semi-supervised conditions, the distribution of strategies did not differ significantly between these two conditions (Fisher’s exact test,  $p = .79$ ). In addition, the proportion of participants using 2D rules did not vary between the conditions (Fisher’s exact test,  $p = .78$ ).

Combining the two semi-supervised conditions, we find an overall interaction between condition and the use of 2D rules (3 conditions  $\times$  2 strategies, Fisher’s exact test,  $p < .05$ ). However, the primary source of this effect seems to be the difference between the Unlabeled Condition and the other conditions. When we aggregate all the labeled conditions together, we see a greater use of 2D rules by the labeled conditions than the unlabeled condition (2 conditions  $\times$  2 strategies, Fisher’s exact test,  $p < .05$ ), while evidence for a parallel effect when aggregating the conditions had access to unlabeled training items together fell short of significance (2 conditions  $\times$  2 strategies, Fisher’s exact test,  $p = .07$ ). In summary, we found minimal evidence that the semi-supervised conditions were different from the all-labeled condition.

## Discussion

Semi-supervised learning is a bit like the Higgs boson in particle physics. It is believed to occur (e.g., to allow word learning) and is strongly suggested by theories of human category

learning (Anderson, 1991; Love et al., 2004), but has proven surprisingly difficult to observe. Our study represents yet another attempt to find laboratory support for this form of category learning. However, the patterns of generalization behavior exhibited at test during the two semi-supervised conditions most closely resembled the strategies of participants who learned in the fully supervised condition.

This result is striking for two reasons. First, unlike some of the previous work on semi-supervised learning, our experiment closely following existing protocols for studying unsupervised category learning in the literature in successfully replicating the results of Z&M (2009). In addition, given no other information participants in our study were willing to generalize according to the distribution of unlabeled examples. In the Unlabeled Condition, the most common strategy was to use a rule on the bimodal dimension. However, when labeled examples were included, participants responded similarly to the 40-All-Labeled Condition. This is the response pattern we would expect to see if subjects mostly failed to incorporate the unlabeled items into their representation of the category in the semi-supervised learning conditions. In this sense, our results join a growing chorus of studies which have failed to find semi-supervised learning except under very specific and limited circumstances (Gibson, Zhu, Rogers, Kalish, & Harrison, 2010; Rogers et al., 2010; Vandist et al., 2009).

In the following sections, we outline a number of possibilities about why semi-supervised learning has been so elusive in the lab.

**Noticing “gaps” in the input?** In our design, the distribution of labeled examples was systematically biased. One possibility is that learners eventually noticed the “gaps” in their input (i.e., that the labels only appeared with particular items) and thus inferred that these examples were somehow special or different. Such a hypothesis may be consistent with a rational learner who tries to determine which items should be clustered together (Griffiths, Sanborn, Canini, & Navarro, 2008). Under this view, an even smaller number of labeled examples (perhaps even one) may actually facilitate generalization (since the amount of data is enough to learn, but not enough to infer some systematic bias). We attempted to get at this issue by modulating the number of labeled training examples, and found no evidence of a trend. However, recent studies of one-shot learning suggest that often even a single labeled example can support robust generalization (Lake, Salakhutdinov, Gross, & Tenenbaum, 2011).

**Overweighting of labeled examples** Another interpretation, suggested by Zhu et al. (2010) and Lake and McClelland (2011), is that labeled items may simply be given more weight. Although this seems plausible, it would appear that the 10 labeled items in the 10-Labeled Condition outweighed the other 270 trials in the training phase, suggesting a weight for unlabeled items much lower than the previously reported estimate of around 40% (Lake & McClelland, 2011). Interestingly, subjects in our 10-Labeled Condition spent consider-

ably more time studying the labeled items, presumably raising their relative influence. One possibility is that the weight given to labeled items is actively adjusted by learners based on the task context.

**Pedagogical sampling** While assuming that labeled examples are given more weight might *describe* the lack of semi-supervised learning, it offers no specific proposal for why this should be the case. One possibility is that participants believed that the experimenter was providing information to teach them the category via the labeled examples (i.e., the labeled examples were pedagogically sampled). In this case, it may be reasonable to trust that the labeled items are particularly informative about the category distinction. For example, Shafto, Goodman, Gerstle, and Ladusaw (2010) have shown that adults adjust their inferences based on the intention of a speaker (either pedagogical or overheard). It is possible that participants in a lab-like setting often assume that training examples are presented pedagogically, causing them to downplay the relevance of unlabeled trials.

**Is an explicit prediction required?** A final hypothesis, there were minor differences between our task and previous work that may have influenced performance. For example, participants made observations and then simply pressed the space bar to acknowledge each item. By contrast, both Kalish et al. (2011) and Lake and McClelland (2011) asked participants to make a response on each trial. It is possible that making a response or prediction on each trial facilitates the integration of information across learning episodes. Consistent with this view is the fact that subjects 40-All-Labeled Condition showed evidence of learning from the items presented at test (where predictions were required)—recall that participants were slightly biased to respond according to the bimodal dimension, even though the only information about its bimodality was provided by the distribution of test examples. A similar effect may have carried over to the semi-supervised conditions as well.

Current work is exploring each of these possibilities. The hunt for semi-supervised learning continues.

## Acknowledgments

We thank Seth Madlon-Kay and Dylan Simon for helpful comments and discussion in the development of this project. TMG was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

## References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98, 409–429.  
 Ashby, F. G. (1992). Multidimensional Models of Categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449–483). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception and Psychophysics*.  
 Bloom, P. (2000). *How children learn the meaning of words*. Cambridge, MA: MIT Press.  
 Clapper, J. P., & Bower, G. H. (1994). Category Invention in Unsupervised Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 443–460.  
 Gentner, D. (1982). Why nouns are learned before verbs: linguistic relativity versus natural partitioning. In S. Kuczaj (Ed.), *Language development: language, cognition, and culture* (pp. 301–334). Hillsdale, NJ: Erlbaum.  
 Gibson, B., Zhu, X., Rogers, T., Kalish, C., & Harrison, J. (2010). Humans learn using manifolds, reluctantly. In *Advances in neural information processing systems* (Vol. 24).  
 Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as Nonparametric Bayesian Density Estimation.  
 Kalish, C. W., Rogers, T. T., Lang, J., & Zhu, X. (2011). Can semi-supervised learning explain incorrect beliefs about categories? *Cognition*, 1–13.  
 Lake, B., & McClelland, J. (2011). Estimating the strength of unlabeled information during semi-supervised learning. In L. Carlson, C. Hölscher & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.  
 Lake, B. M., Salakhutdinov, R., Gross, J., & Tenenbaum, J. B. (2011). One shot learning of simple visual concepts. In L. Carlson, C. Hölscher & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society*. Cognitive Science Society. Austin, TX.  
 Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, 9, 829–835.  
 Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, 111, 309–332.  
 Markant, D., & Gureckis, T. M. (2010). Category Learning Through Active Sampling. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.  
 Medin, D. L., & Schaffer, M. M. (1978). Context Theory of Classification Learning. *Psychological Review*, 85, 207–238.  
 Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57.  
 Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Hines, P., & McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition*, 121, 83–100.  
 Rogers, T. T., Kalish, C., Gibson, B. R., Harrison, J., & Zhu, X. (2010, May). Semi-supervised learning is observed in a speeded but not an unspeeded 2D categorization task. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.  
 Shafto, P., Goodman, N. D., Gerstle, B., & Ladusaw, F. (2010). Prior expectations in pedagogical situations. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.  
 Vandist, K., De Schryver, M., & Rosseel, Y. (2009). Semisupervised category learning: The impact of feedback in learning the information-integration task. *Attention*, 71, 328–341.  
 Zeithamova, D., & Maddox, W. (2009). Learning mode and exemplar sequencing in unsupervised category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 731.  
 Zhu, X. (2005). Semi-Supervised Learning Literature Survey. *Technical Report 1530*, 1–60.  
 Zhu, X., Gibson, B. R., Jun, K.-S., Rogers, T. T., Harrison, J., & Kalish, C. (2010). Cognitive Models of Test-Item Effects in Human Category Learning. In *The 27th international conference on machine learning (ICML)* (p. 158).

# Automated and Partner-Specific Factors Influencing Lexical Entrainment

Lisette Mol (l.mol@uvt.nl)

Rens Bogers (e.a.j.j.bogers@uvt.nl)

Tommy Bouwens (t.j.f.bouwens@uvt.nl)

Tilburg center for Cognition and Communication (TiCC), School of Humanities, Tilburg University  
P.O. Box 90135, NL-5000 LE Tilburg, The Netherlands

## Abstract

Both automated priming (Pickering & Garrod, 2004) and partner-specific adaptation (Brennan & Hanna, 2009) have been proposed to underlie lexical entrainment (the repetition of words across interlocutors). Since activation levels of infrequently used words are relatively low, the effect of automated priming is predicted to be weaker in L2- than in L1-conversations, leaving more room for deliberate partner-specificity (Costa, Pickering, & Sorace, 2008). We tested this prediction by means of a production experiment, in which we varied whether participants interacted in their L1 or L2, and whether they addressed the confederate who had introduced a certain reference or another addressee. We found that in their L2, participants repeated references more frequently when addressing the person who had introduced the reference. Yet we did not find this effect of partner-specificity in the L1 conditions. Therefore, our results support the proposed combination of the two accounts.

**Keywords:** Speech production; alignment; lexical entrainment; interactive-alignment account; conceptual pact

## Introduction

When people interact in dialogue, the referring expressions they use tend to converge (e.g. see Branigan, Pickering, Pearson, & McLean, 2010). For example, if one person refers to a landmark as 'the cathedral', it is likely that a conversation partner would refer to it as such as well, rather than saying 'the church'. This process is known as *lexical entrainment* (Garrod & Anderson, 1987). Different explanations have been proposed for this observation, either relying on automated processes (Pickering & Garrod, 2004) or on the more deliberate process of grounding (Brennan & Clark, 1996; Clark, 1996). In the current study, we test predictions made by each of these two accounts.

## Accounts of Lexical Entrainment

Lexical entrainment can be measured by comparing the ways in which people refer to objects, locations, times, actions, people, etc. If two interlocutors use the same expressions, they are said to be lexically entrained (Garrod & Anderson, 1987). Brennan and Clark (1996) describe three factors that influence reference production that are independent of a conversation's history: informativeness, lexical availability, and perceptual salience. If all speakers

choose their referring expressions based on these factors, they may end up choosing the same references as a result. However, most referents can be referred to in very many ways (Brennan & Clark, 1996). Therefore, these three factors alone are insufficient to explain the (frequent) occurrence of lexical entrainment. Rather, the course of a conversation needs to be taken into account too.

Brennan and Clark also propose four factors influencing reference production that are related to the conversation history: recency, frequency of use, provisionality and partner-specificity. The first two factors imply that speakers choose the reference that was used successfully for a certain referent most recently, taking into account the frequency of occurrence as well. That is, if a certain referring expression has repeatedly been used successfully (e.g. 'couch') and then an alternative reference is used successfully only once (e.g. 'sofa'), speakers may still revert to the more frequently used reference afterward. The factors recency and frequency of use together provide an explanation of lexical entrainment.

These two factors are compatible with a view that lexical entrainment is based on automated priming (Pickering & Garrod, 2004). In a (partly) connectionist model of cognition (Anderson & Lebiere, 1998; McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986), both frequency of use and recency affect the availability of chunks of information, such as lexical items. In line with such models, Pickering and Garrod (2004) propose that through the process of automated priming, interlocutors align their linguistic representations, which causes them to produce similar utterances. For example, when hearing a certain reference, linguistic representations are associated with it in the process of interpretation. As a result, the activations of these representations increase, which makes it more likely that they will be used in the process of language production. Therefore, a referent that has been heard frequently or recently is more likely to be produced. This automated account of lexical entrainment, known as the *interactive-alignment* account, predicts that lexical entrainment will occur whenever speakers are exposed to, and thus primed with, linguistic input.

Effects of provisionality and partner-specificity are not necessarily predicted by such an automated account. These

factors have to do with a more active view of dialogue, in which interlocutors are actively trying to establish common ground (Clark & Brennan, 1991). According to Brennan and Clark, "[w]hen speakers present a reference, they do so only provisionally, and they then work with their addressee to establish that it has been understood" (Brennan & Clark, 1996, p. 1484). Once this is established, interlocutors are said to have formed a *conceptual pact*, which they are likely to maintain. Since a conceptual pact results from an active process, conceptual pacts are established between specific interlocutors only (partner-specificity). Therefore, they cannot simply be transferred to new interaction partners. Rather, the process of grounding a referring expression will have to start anew with any new interlocutor. Therefore, in Brennan and Clark's account of lexical entrainment, entrainment is more likely to occur when interlocutors share a conversation history.

### Adaptation in L2 and L1

Consistent with the interactive-alignment account of lexical entrainment, Costa, Pickering and Sorace (2008) propose that lexical entrainment will be less pronounced when at least one of the interlocutors needs to speak a second language (L2) instead of the native language (L1). Since some lexical items may have been used very rarely by an L2 speaker, their basic level activation may be very low. Therefore, even though recency enhances this activation, it may still be too low for the lexical item to be repeated. This could lead to partially failed entrainment, as in the following example, taken from Costa et al. (2008, p. 538):

- (1) L2 speaker: I need a piece of paper with nothing on it  
 L1 speaker: A blank sheet of paper?  
 L2 speaker: Yeah, a blank piece of paper.

The more frequent word 'blank' is successfully entrained on, which can be explained as it reaching a level of activation that is sufficient for production. However, the infrequent word 'sheet' is not repeated, which may evidence a too low activation level, even after the recent occurrence. Thus, interlocutors are less likely to entrain on words they have encountered only infrequently, such as less familiar words in their L2.

Apart from the automated process of priming underlying lexical entrainment, which Costa et al. (2008) propose to be the default, they also recognize that speakers can make conscious decisions to either suppress the outcomes of this automated process, or to entrain in situations where automated priming does not facilitate entrainment. This leaves room for the factors proposed by Brennan and Clark (1996), such as partner-specificity. Bortfeld and Brennan (1997) for example, found that native speakers were more likely to entrain on the reference 'wheel' to refer to a tire when interacting with an L2-speaker, than when interacting with another L1-speaker. This seems to illustrate a conscious decision by the L1-speaker, based on their knowledge of their interlocutor's proficiency in the L2.

Thus, deliberate motivations for entrainment can also play a role in L1-L2 conversations. Costa et al. (2008) predict that there will always be some extent of lexical entrainment due to automated priming, but this extent will be larger when interlocutors have similar activation profiles for lexical items, such as in L1-L1 conversations, than when they have dissimilar activation profiles, such as in L1-L2 or L2-L2 conversations. Therefore, more deliberate ways of reaching alignment of representations may be more likely to come into play in L1-L2 and L2-L2 conversations.

### Present Study

We are not aware of any empirical study that tested the predictions made by Costa et al. (2008). Therefore, in this study we aim to test the predictions previously laid out, as well as the predictions made by Brennan and Clark's theory on conceptual pacts (Brennan & Clark, 1996). To do so, we measure the degree of lexical entrainment when speakers are talking to an interlocutor who just addressed them (No Switch condition) and when they need to switch to a different addressee (Switch condition). In addition, we assess whether there is a difference in the extent to which this factor affects lexical entrainment when interlocutors communicate in their native language (Dutch, L1-L1), as compared to when they communicate in their second language (English, L2-L2).

In order to keep track of who adapts to whom, we use a controlled experiment, in which participants interact with a confederate. The confederate introduces certain references in one speech turn. We measure the degree to which these references are repeated by the participant in the subsequent speech turn.

The prediction following from grounding theory is that participants are more likely to (partially) repeat a reference when interacting with the partner who introduced the reference than when switching partners. Even though interactivity is limited in our study, participants can decide to accept the reference (provisionally) introduced by the confederate, introduce an adaptation of it, or introduce a completely new reference. It is expected that participants are less likely to produce a completely new reference for a given referent, when interacting with a partner who already introduced a reference for that referent.

The interactive-alignment account does not necessarily predict such a difference between the Switch and No Switch conditions, since whether a referent is reproduced solely depends on automated priming and the linguistic input does not differ across these settings. However, the characteristics of the person who introduced the reference may serve as a prime as well, such that an automated account may also predict a (slight) difference, with more lexical entrainment in the No Switch condition (e.g. Horton & Gerrig, 2005).

In both the L1 (Dutch) and L2 (English) conditions, the confederate uses scripted route directions, with references that were designed in advance. In the L2 conditions, these

references include lexical items that participants may have come across only infrequently. Therefore, the account by Costa et al. (2008) predicts that the automated process of priming plays a more important role in our L1 conditions than in our L2 conditions. Since the effect of automated priming is predicted to be weaker in the L2 conditions, the effect of deliberate partner-specificity is predicted to be stronger there.

## Method

### Participants

Forty-eight (14 male) native Dutch first-year students from Tilburg University participated in our study as part of their curriculum. They were aged between 18 and 35 years old ( $M = 21.94$ ,  $SD = 3.59$ ). In the Netherlands, formal teaching in (British) English starts around the age of twelve. Hence, all participants knew English as a second language, but they were not raised with it.

### Design

We used a 2 x 2 between participants design with two factors: Partner (levels: No Switch, Switch) and Language (levels: L1/Dutch, L2/English). As dependent variable, we used the mean number of (partial) repetitions of the confederate's descriptions of certain landmarks. Two confederates were counterbalanced across the different conditions. Both confederates were male, native speakers of Dutch, from the same (linguistic) region, similar in age (22, 24) and overall appearance (e.g. length, body type, dress, haircut, complexion and hair color).

### Task and Material

A confederate and a participant took turns describing two routes to each other, which were depicted on bird's-eye view maps of a city (see Figure 1). While one interlocutor described a route as depicted on their map, the other interlocutor tried to draw this route on a map that was identical, except that the route was not depicted on it yet. Interlocutors were instructed not to interrupt one another during the route descriptions.

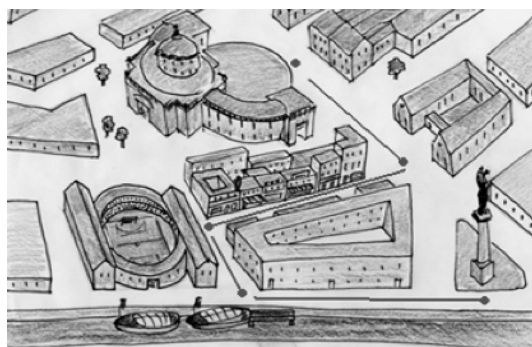


Figure 1: Example of a map used in the experiment.

### Procedure

Participants were randomly assigned to the No Switch or Switch conditions, as well as to the L1 (Dutch, L1-L1) or L2 (English, L2-L2) conditions. In the L1 conditions, the entire experiment was carried out in Dutch. In the L2 conditions, the routes had to be described in English, but instructions were given in Dutch.

In the Switch conditions, two participants came to the lab, where two confederates showed up as participants as well. The experimenter took the four 'participants' to a room and handed out written instructions to the participants and confederates, who were allowed to ask clarification questions. Once all was clear, one confederate and one participant were selected by the experimenter to start the experiment in another room. In each room, the participant and confederate were seated across from each other, with a table in between that had a low visual barrier on it, such as to keep the listening interlocutor from already seeing the route. A camera was positioned on one end of the room, which was used to make recordings of the interaction.

In each of the two rooms, a confederate started the task by sequentially describing two routes to a participant, who was to draw these routes on corresponding maps. Once two routes were described, the two participants were asked to switch rooms, while the two confederates stayed in the same room (this procedure was explained in the written instructions). Each participant then described two routes to the confederate whom they had not yet interacted with. After these two descriptions, the participants switched rooms again and the cycle repeated once. That is, the participants heard two more routes from the confederate they first interacted with and described two more routes to the other confederate. Thus, in the Switch conditions, participants never described routes to the confederate who had given them route descriptions as well, but always to the other confederate.

In the No Switch conditions, there was only one confederate and one participant at a time. Both the participant and the confederate switched rooms after the confederate had described two routes and then the participants described two routes in the other room, to the same confederate. The interlocutors then switched rooms again and the cycle was repeated once. The switching of rooms was the same as in the Switch condition, to ensure equal intervals between hearing and describing routes. In the No Switch conditions, participants described routes to the confederate who had also given them route descriptions.

After all routes had been described, the participants and confederates filled out a questionnaire, including questions on their native and second language. In the English conditions, the questionnaire included the participant's high school grade in English and a short 'fill in the gaps' test on English. Finally, participants were asked consent for the use of their data for research and educational purposes. All participants consented to their data being used in this study.

Table 1: Examples of landmark descriptions given by the confederates.

Landmark Descriptions	
Dutch:	English:
historisch stadspaleis	historical city palace
hoge symmetrische fontein	tall symmetrical fountain
kleurrijke tramstation	colorful tram station
arenavormige voetbalstadion	arena-shaped soccer stadium
kleine intieme terras	small intimate terrace

Since all our participants were students in the same year, we did not inform them on the use of confederates before the study was completed, to ensure they would not know of this in advance. Participants could freely withdraw from the study at any point during and after the experiment.

### Stimuli and Coding

Each route description given by a confederate included three unique critical references to landmarks; see Table 1 for some examples. Since word boundaries sometimes differ between English and Dutch, these references each consisted of three content units, for example 'historical city palace' or the Dutch equivalent: 'historisch stadspaleis'. The confederate started by describing two routes on two different maps. The participant was then to describe two different routes that were depicted on the same maps that the confederate's routes had been on. This allowed for the three landmarks that the confederate had described to be along the participant's route as well, such that partial or full repetitions of the confederate's references could occur. To reduce the possibility of participants accidentally using the same references as the confederate, that is, due to factors independent of the conversation history, the confederate used rather specific wordings (Table 1).

We first transcribed participants' descriptions from the videos. We then coded any literal repetition of content units originally contained in the confederate's critical references. Because full repetitions (3 content units) of references were rare, we report the total number of partial (1 or 2 content units) and full repetitions of the confederate's references. Since true dialogue was not possible between interlocutors, a conceptual pact may not have been fully established yet after the confederate had described the landmarks once. Therefore, participants may still shorten, elaborate on or adapt the reference (provisionally) introduced by the confederate. By counting each reference that contained a literal repetition of at least one content unit, these adaptations were mostly included in our measure of entrainment, while avoiding subjective coding. Note that although the confederate described twelve landmarks, it is possible for a participant to repeat a single landmark more

Table 2: Mean (SD) number of (partial) repetitions of the confederate's landmark descriptions by participants.

Language condition	Partner condition	N	Number of Repetitions:
L1: Dutch	No Switch	11	9.72 (2.24)
	Switch	14	10.00 (2.94)
	Total	25	9.88 (2.60)
L2: English	No Switch	11	11.27 (3.26)
	Switch	12	8.42 (2.39)
	Total	23	9.78 (3.13)
Total	No Switch	22	10.50 (2.84)
	Switch	26	9.27 (2.76)

than once. Hence, a participant could produce more than twelve repetitions.

### Results

We performed an ANOVA with Partner (levels: No Switch, Switch) and Language (levels: L1, L2) as independent factors. In line with the predictions, analysis of participants' (partial) repetitions of the confederate's descriptions, as listed in Table 2, showed a marginally significant interaction between the two factors, such that the difference between the No Switch and Switch conditions was larger in the L2 conditions than in the L1 conditions,  $F(1, 44) = 3.86$ ,  $p = .06$ ,  $\eta_p^2 = .08$ . We did not find a main effect of Partner,  $F(1, 44) = 2.63$ ,  $p = .11$ ,  $\eta_p^2 = .06$ , or Language,  $F(1, 44) < 1$ ,  $p = .98$  on the number of (partial) repetitions.

Posthoc analyses by means of independent-samples T-tests did not reveal an effect of Partner in the L1 conditions,  $t(23) = .26$ ,  $p = .80$ . In the L2 conditions however, there was an effect of Partner,  $t(21) = 2.41$ ,  $p < .05$ ,  $\omega^2 = .17$ , such that participants repeated more of the confederate's references if they were interacting with the same partner ( $M = 11.27$ ,  $SD = 3.26$ ) than when they had to switch ( $M = 8.42$ ,  $SD = 2.39$ ).

Given our design, it is important that participants in the L2 (English) conditions were equally proficient in English in the No Switch and Switch condition. Analysis of the score participants obtained on the short English post-test did not reveal a difference between the English No Switch ( $M = 7.00$ ,  $SD = .89$ ) and the English Switch condition ( $M = 6.83$ ,  $SD = 1.19$ ),  $t(21) = .38$ ,  $p = .71$ . Similarly, we did not find a difference in the reported high school grade in English between the English No Switch ( $M = 7.55$ ,  $SD = .96$ ) and the English Switch condition ( $M = 7.08$ ,  $SD = 1.12$ ),  $t(21) = 1.05$ ,  $p = .30$ . (These high school grades are on a 1 to 10 scale, with 6 being sufficient and 10 being exceptional.) Since there seems to be a numerical difference in the high school grades between the English Switch and the English No Switch condition, we also performed an ANOVA on the data from the English conditions with high-school grade as a covariate. Similar as before, analysis of participants' repetitions of the confederate's descriptions revealed an effect of Partner,  $F(1, 23) = 5.05$ ,  $p < .05$ ,  $\eta_p^2 = .20$ , such

that in the English conditions, participants repeated the confederate's reference more often when they interacted with the same partner, than when they had to switch. We did not find an effect of high school grade in this analysis,  $F(1, 23) < 1, p = .84$ .

Similar results were obtained when analyzing at most one repetition per landmark, and also when separate variables were computed for each half of the experiment (the order of the routes was counterbalanced across the first and second half of the experiment).

## Discussion

Our results confirm the prediction that the effect of partner-specificity is stronger for non-native speakers than for native speakers (Costa, et al., 2008). When participants were asked to communicate in a foreign language (English), they more often repeated parts of a confederate's referring expressions when they were talking to the confederate who introduced the reference, compared to when they were addressing a third person. When speaking in their native language (Dutch), the difference between addressing the same or a different partner was not as large. This finding seems to support both grounding theory (Brennan & Clark, 1996; Clark & Brennan, 1991) and the interactive-alignment account (Pickering & Garrod, 2004).

In the English conditions, we found that when participants did not need to switch conversation partners in between hearing and giving route descriptions, they repeated the confederate's references to a greater extent than when they had to switch partners. This goes well with the prediction that conceptual pacts are partner-specific (Brennan & Clark, 1996). Even though our paradigm did not allow for interlocutors to freely interact in arriving at conceptual pacts, repeating part of a previously (provisionally) introduced reference can be seen as a first step in forming a conceptual pact. We found that in the L2 conditions, participants produced such a repetition more often to the person who had introduced the reference than to a person who had no knowledge of this conversation history. Therefore, these results support the account of lexical entrainment based on grounding theory (Clark, 1996; Clark & Brennan, 1991).

The prediction by Costa et al. (2008) that automated priming will be more prominent when interlocutors interact in their native language than when they interact in a foreign language also seems to be supported by our data. Our results show that the effect of partner-specificity, which can be thought of as a more deliberate factor, was stronger when participants were asked to interact in a foreign language as compared to when they interacted in their native language, exactly as predicted. The explanation offered by Costa et al., that the weaker effect of automated priming in L1-L2 and L2-L2 conversations would allow more room to such deliberate factors seems very plausible. Because the lexical items from the native language are very familiar, they are

easily primed by recent uses. This process does not depend on whom is being addressed. Therefore, no effect of partner-specificity is predicted in the L1 conditions, which is in line with our findings. Yet when using the L2, the activation levels of less frequently used lexical items may be too low for these items to be primed by a single recent use. Therefore, the effect of automated priming is weaker in L2 conversations. Speakers may therefore make a more deliberate decision on whether to entrain on a given reference or not in the L2 conditions, which could very well depend on whom they are interacting with. The stronger effect of partner-specificity that we found in the L2 conditions thus indirectly supports the interactive-alignment account.

Can our findings be accounted for without the interactive-alignment account? One can think of many reasons why we did not find an effect of partner-specificity in the Dutch (L1) conditions. For example, there were only few turns and participants were not free to engage in dialogue. However, these factors equally apply to the English (L2) conditions, in which we did find an effect of partner-specificity. This also holds for the argument that in the second half of the experiment, there may be some conversation history in the Switch conditions. (Critical landmarks from the first half of the experiment did not occur in the second half.) Thus, from grounding theory alone, it is hard to explain why we did not find an effect of partner-specificity in the L1 conditions, whereas we did find this effect in the L2 conditions.

A possible explanation of our results in terms of a deliberate process is that in the L2 conditions, participants did not repeat the confederate's references as frequently when they had to switch partners, because they were less certain of their new interaction partner's proficiency in English. Therefore, they may have used more common references instead. At the same time, when interacting with the same partner, the already introduced reference was most likely to be understood. This explanation fits the numerical pattern in our data, as the larger effect of partner-specificity in the L2 conditions seems to result both from there being more repetitions in the No Switch condition and from there being fewer repetitions in the Switch condition (see Table 2). However, although to a lesser extent, similar deliberate considerations apply to the L1 conditions. Moreover, in the L1 condition in which participants had to switch partners, there seems little reason for participants to repeat the reference introduced by the confederate at all (which they did to the same extent as in the No Switch condition), other than because of automated factors. Therefore, deliberate partner-specificity alone does not convincingly explain all of our results.

Can our findings be accounted for without deliberate partner-specificity? The interactive-alignment account predicts that interlocutors will show less lexical entrainment in their L2, but it does not predict an effect of either addressing the person who introduced the reference or



another person. That is, a main effect of the factor Language would be predicted, rather than an interaction between the factors Language and Partner. Horton and Gerrig's *association account* states that conversation partners automatically form an association between a given expression and a conversation partner (e.g. Horton & Gerrig, 2005). Hence, the conversation partner may serve as a prime as well, leading to more entrainment when interacting with the same partner as compared to when switching partners. This could provide an automated account of partner-specificity. Yet importantly, if partner-specificity were fully automated, there is no reason to expect its effect to be stronger in the L2 than in the L1 conditions. That is, a main effect of the factor Partner would be predicted, rather than an interaction between Partner and Language. Yet although we did not find an effect of partner-specificity when interlocutors interacted in their L1 (Dutch), we did find that in their L2 (English), participants showed more lexical entrainment when they kept interacting with the same partner than when they had to switch. This interaction effect cannot be accounted for by an automated account alone. Therefore, a combination of both the interactive alignment account (Pickering & Garrod, 2004) and more deliberate partner-specificity (Brennan & Clark, 1996), as proposed by Costa et al. (2008), explains our results best. Moreover, this account *predicted* the results that we found.

Our study illustrates that automated and deliberate accounts of adaptation in dialogue are compatible (also see Brennan & Hanna, 2009; Costa, et al., 2008; Pickering & Garrod, 2004). In future work, it would be interesting to further explore what factors influence the extent to which automated and deliberate factors come into play, as well as to assess which factors are more automated and which are more deliberate, given a certain setting. Next to the theoretical merit, this could provide insight into how and when to facilitate effective communication.

## Conclusion

Our results on lexical entrainment support both the interactive-alignment account (Pickering & Garrod, 2004) and the account of partner-specific conceptual pacts (Brennan & Clark, 1996). However, neither of these two theories alone predicted the pattern of results that we found. When interacting in their second language, speakers were shown to entrain more on previously heard references when interacting with the person who introduced these references, than when interacting with another partner. When speakers interacted in their native language, this effect did not reach significance, evidencing an interaction between the factors of whether or not speakers used their native language and whether or not speakers switched conversation partners. This interaction was previously predicted, yet not tested, by Costa et al. (2008), who combined the two theories and predicted that the effect of automated priming on lexical

entrainment would be stronger when interlocutors interact in their native language, compared to when at least one of them uses a non-native language. Therefore, when interlocutors interact in their non-native language, the effect of more deliberate factors, such as partner-specificity, will be (relatively) stronger. This is exactly what we found. Our findings thus support the view that both automated priming and deliberate partner-specific adaptation influence the degree of lexical entrainment between interlocutors.

## Acknowledgements

We gratefully acknowledge all speakers for allowing us to analyze their data, we thank the anonymous reviewers for their insightful suggestions and we thank Nathalie Bastiaansen for drawing the maps used in this experiment.

## References

- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Manwah, NJ: Erlbaum.
- Bortfeld, H., & Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23, 119-147.
- Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between humans and computers. *Journal of Pragmatics*, 42, 2355-2368.
- Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology-Learning Memory and Cognition*, 22(6), 1482-1493.
- Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialogue. *Topics in Cognitive Science (Special Issue on Joint Action)*, 1, 274-291.
- Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.
- Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. Levine & S. D. Teasley (Eds.), *Perspectives on socially shared cognition*. Washington, DC: APA.
- Costa, A., Pickering, M. J., & Sorace, A. (2008). Alignment in second language dialogue. *Language and Cognitive Processes*, 23(4), 528-556.
- Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27, 181-218.
- Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, 96, 127-142.
- McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume II*. Cambridge, MA: MIT Press.
- Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2), 169-225.
- Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume I*. Cambridge, MA: MIT Press.

# Gesture Structure Affects Syntactic Structure in Speech

**Lisette Mol (l.mol@uvt.nl)**

Tilburg center for Cognition and Communication (TiCC), School of Humanities, Tilburg University  
P.O. Box 90135, NL-5000 LE Tilburg, The Netherlands

**Sotaro Kita (s.kita@bham.ac.uk)**

University of Birmingham, School of Psychology, Birmingham B15 2TT, UK

## Abstract

Different functions have been proposed for the hand gestures speakers spontaneously produce while speaking. The Information Packaging Hypothesis (Kita, 2000) states that gestures can structure rich spatio-motoric information into packages suitable for speaking. It therefore predicts that how information is divided over different gestures affects how it is divided over different processing units in speech: clauses. We indeed found that if participants were asked to express the manner and path of a motion in one gesture, they were also more likely to conflate this information into one clause in speech, whereas if they were asked to produce separate gestures, they were more likely to express manner and path in separate clauses too. These results support the view that there are speaker-internal motivations for gesture production. They confirm predictions made by the Information Packaging Hypothesis, which the Lexical Retrieval Hypothesis and the Image Activation Hypothesis do not make.

**Keywords:** Gesture; Speech; Production; Motion Event

## Introduction

When speaking, most people tend to produce hand gestures that are closely synchronized with their speech semantically (e.g. McNeill, 2005), temporally (e.g. Chui, 2005), and structurally (e.g. Kita & Özyürek, 2003). Because of this careful coordination, it is generally assumed that the processes of speech and gesture production are somehow related. Yet what is the exact role of gesture production in relation to speech production?

## Gesture and Speech Production

In this paper, we focus on representational hand gestures (McNeill, 1992). Representational gestures either depict action, motion or shape ("iconic gestures") or indicate a location or direction ("deictic gesture"). Much evidence has been gathered in support of a theory that (representational) gestures, like speech, are part of a speaker's communicative effort (Kendon, 2004). In line with this view, Growth Point Theory (McNeill, 2005; McNeill & Duncan, 2010) starts from the observation that gesture and speech co-express idea units, each using a different form of semiosis. While gesture employs a global/synthetic form of representation, speech is expressed in an analytic/ combinatoric form. It is assumed that gesture and speech production share a common origin: the growth point. From this origin, a bimodal utterance develops from the interplay of imagistic

and linguistic processing. Thus, in this view, gesture and speech are two outcomes of a single process.

In addition to the line of thought that gestures are intended communicatively, it has also been proposed that there are speaker-internal motivations for gesture production. Some propose that gesture production facilitates cognitive processes in general, by lightening cognitive load (Goldin-Meadow, Nusbaum, Kelly & Wagner, 2001). Others propose that gesture production facilitates a specific process in speech production. In this article, we focus on the latter class of proposals. There are three prominent proposals in the literature: The Lexical Retrieval Hypothesis, the Image Activation Hypothesis, and the Information Packaging Hypothesis.

The Lexical Retrieval Hypothesis (LRH) states that gestures facilitate the retrieval of lexical items from the mental lexicon (Krauss, Chen, & Gottesman, 2000). In this view, gesture production is based on spatial imagery in working memory. Rather than there being an interplay between the processes of gesture and speech production, the execution of a gesture is thought to activate spatio-dynamic features, which in turn activate conceptual information. Through cross-modal priming, this aids the retrieval of lexical items. Thus, gesture production precedes speech formulation entirely.

The Image Activation Hypothesis (IAH) states that gesturing serves to keep an image (Freedman, 1977) or certain spatial features (De Ruiter, 1998) activated while they are encoded by the process of speech formulation.

The Information Packaging Hypothesis (Kita, 2000) critically differs from the Lexical Retrieval Hypothesis and the Image Activation Hypothesis in its assumptions on the for-speaker motivations of gesture production, and on the interplay between gesture and speech production. Rather than simply activating information or maintaining the activation of spatial information, gesture production is thought to structure information, and to package it into units that are suitable for the speech formulation process.

Like Growth Point Theory, the Information Packaging Hypothesis (IPH) assumes that different forms of processing underlie gesture and speech. It is proposed that gesture is based on spatio-motoric processing and speech on analytic processing. The IPH assumes that "[s]patio-motoric thinking, which underlies representational gestures, helps speaking by providing an alternative informational

organization that is not readily accessible to analytic thinking, the default way of organizing information in speaking" (Kita, 2000, p. 163). Furthermore, it assumes that "[s]patio-motoric thinking and analytic thinking have ready access to different sets of informational organizations" (Kita, 2000, p. 163). The representations in these two modes of thinking are thought to be coordinated online during language production, such that they tend to converge.

The Information Packaging Hypothesis is implemented in the Interface Model (Kita & Özyürek, 2003), which adds gesture production components to Levelt's (1989) model of speech production. In the Interface model, what needs to be expressed is determined in the communication planner. This module informs the action generator, where gestural contents are determined, as well as the message generator, where the preverbal message is determined. Importantly, the action and the message generator are bi-directionally linked to each other, allowing for gesture and speech to coordinate their contents during language production. Lastly, the message generator is also linked bi-directionally to the formulator, which converts a preverbal message into an utterance, through accessing the mental lexicon and retrieving and processing morphological, syntactic and phonological information. This way, the formulator can pass on information to the action generator, via the message generator. Thus, the *bidirectional* link between the action and message generator allows for gesture and speech production to be coordinated semantically and structurally. It is assumed that the processes of speech and gesture generation constantly exchange information, which is transformed from one format into another, such that the content of both modules tends to converge.

What is the evidence for this convergence of information packaging in speech and gesture? The evidence for the speech-to-gesture influence comes from studies of motion event descriptions (Kita, 2000; Kita & Özyürek, 2003; Özyürek, Kita, Allen, Furman, & Brown, 2005). It was found that whether speakers used a single or multiple clauses to verbally describe the manner and path of a motion tended to match whether they expressed manner and path in a single or in separate gestures. For example, "he rolled down the hill" would likely be accompanied by a gesture in which the hand describes circular motions as it is moved down diagonally, while "he rolled, as he went down the hill" would be more likely to be accompanied by one gesture illustrating the rolling and another gesture illustrating the downward path. In one study on English speakers (Kita, et al., 2007), different clausal structures in speech were elicited by varying whether a manner was incidentally or causally related to the path of a motion in the stimulus animations. Different clause structures lead to the predicted different patterns of packaging of manner and path in gestures.

The evidence for the gesture-to-speech influence comes from studies in which the availability of gestures was manipulated (Alibali & Kita, 2010; Alibali, Spencer, Knox,

& Kita, 2011). It was found that the availability of gestures changed the type of information encoded in speech. More specifically, when gestures are produced, speakers tended to encode spatial information that gestures readily had access to. To date, there is no study that manipulated gesture structure (as opposed to gesture availability) to examine its influence on speech production.

## Present Study

We will test the prediction made by the IPH (but not by the LRH and IAH), that the information structure underlying gesture production can influence the information structure underlying speech. In doing so, we will use the task of describing motion events, as in the study described above (Kita, et al., 2007). Yet how can we measure the analytic representations underlying speech production?

Bock and Cutting (1992) propose that syntactic processing units comprise of (finite) clauses. Using a procedure to elicit verb agreement errors, they found that these errors occurred more frequently when the head noun and its verb were separated by a phrase (e.g. "The claim about the stolen babies was rejected", p. 104) than when they were separated by a relative clause (e.g. "The claim that wolves were stealing babies was rejected", p. 104). Assuming a hierarchical processing structure, they explain this as that the more information is introduced within a single processing unit, the more sources of interference there are between similar, concurrently active elements. In this example, when the head noun and the local noun are part of a single clause, their numbers are more likely to interfere than when they are part of separate clauses. Thus, within a clause, elements are more likely to interfere than across clauses. This supports the notion that clauses are the units of syntactic processing. Following Bock and Cutting (1992), we will assume finite clauses to be indicative of the processing units underlying speech.

We instructed participants to either conflate the manner and path of a motion into a single gesture (Conflated condition), or to produce one gesture for the manner and a separate gesture for the path of the motion (Separate condition) and observed the syntactic packaging of manner and path in speech. Since the IPH assumes that the processing units underlying speech and gesture are coordinated, it predicts this manipulation will affect the clausal structure of speech, such that conflated gestures tend to go with single, conflated clauses, whereas separate gestures tend to go with separate clauses for manner and path.

The Image Activation Hypothesis may not make specific predictions as to the difference between the conflated vs. separate gesture condition. Both the conflated gesture and separate gesture condition should equally boost the activation of the imagery of manner and path. More crucially, the hypothesis does not propose any mechanism as to how linguistic expressions are influenced by gesture

production, aside from the assumption that more strongly activated imagery leads to better quality descriptions.

The Lexical Retrieval Hypothesis may predict an effect of gesture production on speech production. Yet this effect is different from the effect predicted by the IPH. Rather than the clausal structure of speech being affected by the way speakers gesture, it may predict that speakers who use different gestures would activate different lemmas and would thus use different words.

## Method

### Participants

Twenty-one native Dutch first-year students (4 male) from Tilburg University participated in our study as part of their curriculum. They were aged between 18 and 24 years old ( $M = 21.24$ ,  $SD = 1.61$ ). Four participants were left-handed. The number of male and left-handed participants was equal in the two conditions.

### Material

The ten stimulus clips that our participants described were from a set of animated cartoons known as 'the Tomato Man movies' (Özyürek, Kita, & Allen, 2001). Each clip consisted of an initial entry event, followed by a target event in which one of the two figures completes a motion along a certain path and in a certain manner, and finally a closing event (see Figure 1).

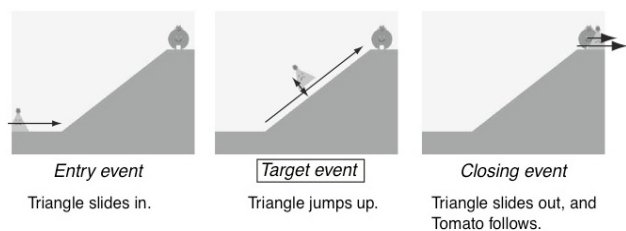


Figure 1: Example of a stimulus clip, taken from (Kita, et al., 2007).

### Procedure

Participants came to the lab and were randomly assigned to one of the two gesture conditions. They first received a written instruction, after which they were allowed to pose clarification questions. The instruction explained that the participant was to watch animated cartoons in which a cartoon figure sometimes conducted a motion involving both a certain manner and a certain path of movement. In the Conflated condition, participants were asked to produce a hand gesture with their description of such motions, such that 'the gesture illustrates both the path and manner of movement at the same time'. In the Separate condition, participants were requested to produce two different gestures with their description of the motion, with 'one

gesture illustrating the path, while the other gesture shows the manner of movement'. Note that participants were asked to *gesture* differently only. Otherwise, the instructions were exactly the same in both conditions.

Participants were seated next to a table with a laptop on it, which showed a Powerpoint presentation. Upon pressing a button, an animated cartoon played twice, after which the participant was to describe it to the experimenter, who was seated across from the participant. Behind the experimenter was a camera, capturing the participant. The first clip was a practice clip. If needed, the experimenter gave additional instructions on how to do the task, by asking specific questions (e.g. "And how exactly did the figure go around the tree?") or by describing how the gestures produced by the participants differed from the gestures requested. The participant then proceeded to watch and describe the ten stimulus clips. Afterward, the participant filled out a short questionnaire, which asked for participants' native language and whether they were left or right handed. Lastly, they filled out a consent form. All participants permitted their data to be used for research and educational purposes.

### Coding and Analysis

All recordings were analyzed using Elan (Max Planck Institute for Psycholinguistics; Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006). For each description of a target event, speech was transcribed into finite clauses and gestures were coded. We used finite verbs to decide on clause boundaries. Each clause contained one conjugated verb. Hence, the first two examples in Table 1 (next page) were both coded as a single clause. Few utterances contained a praedicativum, as in (1). *Al springend* is linked both to the verb and to the noun. Its status is therefore not entirely clear. It can be thought of as an adjective, a verb or an adverb (Jansen & Lentz, 2002). Because of this and since these cases were few, we excluded them from our analyses.

- (1) "Al springend gaat hij omhoog."  
while in a jumping manner goes he up.  
"(While) jumping he goes up."

For each clause and each gesture, it was determined whether it contained information on manner, on path, or on both. Verbal descriptions in which manner and path were solely expressed in a single clause were coded as *conflated*. Descriptions in which there were separate clauses for manner and path were coded as *separate*, even if the description also contained a clause in which the two were conflated. Descriptions in which either manner or path was described in a separate clause, yet not both, and in which there was a conflated clause as well were coded as *mixed*. Table 1 provides some of participants' utterances and our coding. Gestural descriptions were coded analogously, for whether manner and path were conflated into a single gesture, expressed in separate gestures or a mixture of both.

Table 1: Examples of utterances and our coding. Coded clause boundaries are indicated by forward slashes.

Utterance:	Label:
"het rode rondje draait dan zo omhoog" /the red circle (+diminutive) turns then like this upwards/ the red circle then <i>twists upwards</i> like this	conflated
"en hij gaat zo springend omhoog" /and he goes like this in a jumping manner upwards/ and he <i>goes jumping upwards</i> like this	conflated
"en die draait terwijl die naar boven gaat" /and that (pronoun feminine/masculine) turns/ while that to up goes/ and he <i>turns</i> as he <i>goes up</i>	separate: (manner, path)
"en de driehoek gaat omhoog en dat doet ie door uh te springen" /and the triangle goes up/ and that does he by uh to jump/ and the triangle <i>goes up</i> and that he does by uh <i>jumping</i>	separate: (path, manner)
"de appel die rolt en hij rolt van links naar rechts omhoog" /the apple that rolls/ and he rolls from left to right upwards/ the apple <i>rolls</i> and it <i>rolls upwards</i> from left to right	mixed: (manner, conflated)

Occasionally, participants failed to describe both the manner and the path of a motion in gesture and speech (8 cases). Also, participants sometimes failed to comply with the instructions, only producing separate gestures for manner and path in the Conflated condition (3 cases), or only producing a conflated gesture in the Separate condition (24 cases). These cases were discarded from our analyses. The more frequent problems in the Separate condition may indicate that conflated gestures (along with conflated speech) are more common among Dutch speakers, similarly to English speakers (Kita & Özyürek, 2003). The remaining dataset contained 175 descriptions (verbal and gestural) by 21 participants. To ensure fair comparison between participants and between conditions, despite an unequal number of descriptions, we computed the proportion of verbal expressions of a certain type (conflated, separate, mixed) for each participant, rather than using the raw counts. The means of both conditions were compared using independent-samples T-tests. When Levene's test for equality of variances was significant, we report the adjusted statistics.

## Results

Participants produced greater proportions of verbal descriptions in which manner and path were conflated into a single clause when they were asked to produce a single conflated gesture for manner and path ( $M = .88$ ,  $SD = .18$ ),

than when they were asked to express manner and path in two separate gestures ( $M = .19$ ,  $SD = .31$ ),  $t(19) = 6.26$ ,  $p < .001$ . The reverse pattern was found for descriptions in which manner and path were expressed in separate clauses (see Table 2). Similar results were obtained when descriptions on which participants failed to comply with the instruction were included. We did not find any effects of gender, or left or right-handedness.

## Discussion

Our results show that the way people gesture influences the way they speak. When asked to divide manner and path information over two gestures, participants were more likely to also use two clauses for manner and path. These results are in line with the Information Packaging Hypothesis (Kita, 2000). Being required to separate manner and path into two gestures forced speakers to think about the event in a certain way spatially. That is, it requires them to separately focus on the path of the motion and the manner of the motion sequentially, as each unit of information is the base of one unit in gesture. This differs critically from the conflated condition, in which participants were asked to conflate manner and path into a single gesture, which calls for spatially processing the motion as a whole, that is, a single unit of information. The Information Packaging Hypothesis predicts that the analytic processing units underlying speech tend to converge with the spatio-motoric processing units

Table 2: The effects of gesture condition (Conflated vs. Separate) on the mean proportions (SD) of verbal descriptions in which manner and path were expressed in a single clause (conflated), separate clauses (separate) or a combination (mixed).

Verbal descriptions	Conflated gesture (N=10)	Separate gestures (N=11)	Statistics
proportion conflated	.88 (.18)	.19 (.31)	$t(19) = 6.25$ , $p < .001$
proportion separate	.09 (.11)	.63 (.31)	$t(12.7) = 5.46$ , $p < .001$
proportion mixed	.03 (.07)	.18 (.24)	$t(19) = 1.85$ , $p = .08$

underlying gesture. We have taken clauses to be a measure of the analytic processing units (Bock & Cutting, 1992). When gesture production forced participants to spatially process the motion as a whole, manner and path were more frequently conflated into a single clause, reflecting one analytic processing unit. Yet when gesture forced participants to process the manner and the path of the motion separately, they more frequently expressed manner and path in two separate clauses, reflecting two units in analytic processing. This supports the prediction made by the Information Packaging Hypothesis, that the processing units underlying speech can be adapted to the processing units underlying gesture.

Our results also support the Interface Model of gesture and speech production (Kita & Özyürek, 2003). Specifically, they confirm that the link between the action generator and the message generator is bidirectional in nature. Earlier work had already shown that the constraints a language imposes on what information can be linguistically expressed within a clause affect gesture production (Kita & Özyürek, 2003) and that the structure of speech could affect the structure of gesture (Kita, et al., 2007; Özyürek, et al., 2005). Our current study shows that when gesture formulation is constrained, this affects speech formulation as well, exactly as the model would predict.

Since gesture is generally assumed to be less conventionalized than speech, it may not be as straightforward to see in what kind of naturalistic situations gesture would impose constraints on speech formulation. However, there is a growing body of evidence that speakers adapt their gestures to one another (Holler & Wilkin, 2011; Kimbara, 2008; Mol, Krahmer, Maes, & Swerts, 2012; Parrill & Kimbara, 2006). When a gesture shape or a structure in gesture is imitated from another speaker, this may in turn influence the speech formulation process, potentially causing speech to converge across interlocutors as a result. Also, there can be cultural and pragmatic constraints on gesture (Enfield, Kita, & De Ruiter, 2007; Kita & Essegbey, 2001). More importantly though than gesture imposing constraints on how information can be expressed, it can open up new possibilities of organizing information, by supporting spatio-motoric thinking (Chu & Kita, 2008; Kita, 2000). Our results confirm that speech production can benefit from gesture this way. This supports theories in which gesture results from speaker-internal motivations.

Can the current findings on clause structure be accounted for by the Lexical Retrieval Hypothesis? The Lexical Retrieval Hypothesis may be supported if the manipulation of gestures caused different choices of manner verbs in the two conditions, which in turn lead to different clause structures. Though, in principle, any Dutch manner verb can be used in both clausal structures, we examined the manner verbs in the two conditions. We included all inflections of manner verbs, such as in (2), as well as manner adverbs, as in (3).

(2) *hij rolt zo omhoog*  
he rolls like this upwards  
"he rolls up like this"

(3) *hij gaat rollend van de helling af*  
he goes in a rolling manner from the hill off  
"he goes rolling off the hill"

The numbers of manner (ad)verbs used in the two conditions were highly correlated,  $R(11) = .90, p < .001$ , see Table 3. This indicates that the compositions of (ad)verbs used in the two conditions were very similar. Thus, there is no support for the idea that gesture affected clause structures via different choices of manner (ad)verbs.

Can the current findings be accounted for by the Image Activation Hypothesis? According to this hypothesis, gestures boost the activation level of the imagery that is intended to be communicated. This hypothesis does not specify the relationship between clause structures and imagery; thus, gesture's effect on clause structure cannot be accounted for.

## Conclusion

Our results demonstrate that gesture production can influence speech production. Specifically, the way information was divided over individual gestures affected the way information was divided into clauses. This supports the Information Packaging Hypothesis (Kita, 2000), in which gesture production serves to organize rich spatio-motoric information into packages suitable for speaking, and the spatio-motoric processing units underlying gesture production are coordinated with the analytic processing units underlying speech production. Therefore, these results also support the view that there are speaker-internal motivations for gesture production.

Table 3: Manner (ad)verbs used in each condition.

Lemma:	Translation:	Number of occurrences:	
		Conflated	Separate
draaien	turn	34	34
springen	jump	13	23
rollen	roll	12	16
stuiteren	bounce	11	3
cirkelen	circle	8	0
huppen/hoppen	hop	5	6
twisten	twist	0	1
tuimelen	tumble	0	3
koprollen	rollover	0	1
kantelen	topple	0	1
buitelen	tumble	0	1
Total:		83	89

## Acknowledgements

We gratefully acknowledge all participants for allowing us to analyze their data, we thank Pieter Spronck for helping with the analysis of word frequencies and Maria Mos, Jorrig Vogels and Joost Schilperoord for sharing their knowledge on clause structures in Dutch. We thank the anonymous reviewers for their comments and suggestions.

## References

- Alibali, M. W., & Kita, S. (2010). Gesture highlights perceptually present information for speakers. *Gesture*, 10(1), 3-28.
- Alibali, M. W., Spencer, R. C., Knox, L., & Kita, S. (2011). Spontaneous Gestures Influence Strategy Choices in Problem Solving. *Psychological Science*, 22(9), 1138-1144.
- Bock, K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, 31, 99-127.
- Chu, M., & Kita, S. (2008). Spontaneous gestures during mental rotation tasks: Insights into the microdevelopment of the motor strategy. *Journal of Experimental Psychology: General*, 137(4), 706-723.
- Chui, K. (2005). Temporal patterning of speech and iconic gestures in conversational discourse. *Journal of Pragmatics*, 37(6), 871-887.
- De Ruiter, J. P. (1998). Gesture and Speech Production. Unpublished Doctoral Dissertation. University of Nijmegen.
- Enfield, N. J., Kita, S., & De Ruiter, J. P. (2007). Primary and secondary pragmatic functions of pointing gestures. *Journal of Pragmatics*, 39, 1722-1741.
- Freedman, N. (1977). Hands, words, and mind: on the structuralization of body movements during discourse and the capacity for verbal representation. In N. Freedman & S. Grand (Eds.), *Communicative Structures and Psychic Structures: A Psychoanalytic Approach*. New York and London: Plenum Press.
- Holler, J., & Wilkin, K. (2011). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, 35, 133-153.
- Jansen, F., & Lentz, L. R. (2002). Braad dikwijls bedruipende. Twee manieren om gelijktijdige handelingen aan te duiden: deelwoordconstructies en onderconstructies. *Neerlandistiek.nl*, 6(2), 1-26.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.
- Kimbara, I. (2008). Gesture form convergence in joint description. *Journal of Nonverbal Behavior*, 32(2), 123-131.
- Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and Gesture*. Cambridge: Cambridge University Press.
- Kita, S., & Essegbey, J. (2001). Pointing left in Ghana: How a taboo on the use of the left hand influences gestural practice. *Gesture*, 1, 73-94.
- Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 47, 16-32.
- Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes*, 22(8), 1212-1236.
- Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and gesture*. New York: Cambridge University Press.
- Levelt, W. J. M. (1989). *Speaking*. Cambridge, MA: MIT Press.
- Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. <http://www.lat-mpi.eu/tools/elan/>
- McNeill, D. (1992). *Hand and Mind: What gestures reveal about thought*. Chicago and London: The University of Chicago Press.
- McNeill, D. (2005). *Gesture and Thought*. Chicago and London: University of Chicago Press.
- McNeill, D., & Duncan, S. (2010). Gesture and growth points in language disorders. In J. Guendouzi, F. Loncke & M. J. Williams (Eds.), *The handbook of psycholinguistic and cognitive processes*. New York, London: Psychology Press.
- Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language*, 66(1), 249-264.
- Özyürek, A., Kita, S., & Allen, S. (2001). Tomato Man movies: Stimulus kit designed to elicit manner, path and causal constructions in motion events with regard to speech and gestures. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics, Language and Cognition group.
- Özyürek, A., Kita, S., Allen, S., Furman, R., & Brown, A. (2005). How does linguistic framing of events influence co-speech gestures? Insights from cross-linguistic variations and similarities. *Gesture*, 5(1), 215-237.
- Parrill, F., & Kimbara, I. (2006). Seeing and hearing double: The influence of mimicry in speech and gesture on observers. *Journal of Nonverbal Behavior*, 30(4), 157-166.
- Wittenburg, P., Brugman, H., Russel, H., Klassmann, A., & Sloetjes, H. (2006). *ELAN: a Professional Framework for Multimodality Research*. Paper presented at the LREC 2006, Fifth International Conference on Language Resources and Evaluation.



# Modeling Millisecond Time Interval Estimation in Space Fortress Game

Jungaa Moon (jungaam@andrew.cmu.edu)

John R. Anderson (ja+@cmu.edu)

Department of Psychology, Carnegie Mellon University  
Pittsburgh, PA 15213 USA

## Abstract

We investigated sources of the asymmetric bias found in estimation of a time interval (250–400 ms) embedded in the Space Fortress task (Donchin, 1989). Two hypotheses to explain this bias were tested in a behavioral experiment: 1) contamination from a different time interval representation, and 2) pressure to complete the task in time. Participants alternated between producing the target interval and producing either a shorter or a longer interval while the total time allowed for the task was manipulated. The results showed that the target interval estimate was significantly influenced by both manipulations. The effects were captured by incorporating the timing model of Taatgen and Van Rijn (2011) into the ACT-R model for Space Fortress (Bothell, 2010). Time estimation performed in a dynamic task requires understanding the influence of external temporal tasks as well as the procedural demands of performing multiple tasks under time pressure.

**Keywords:** Time estimation; cognitive model; multitasking.

## Introduction

Time interval estimation underlies various skills such as motor control (Ivry, Spencer, Zelaznik, & Diedrichsen, 2002), musical performance (Jones, 1990), and speech processing (Schirmer, 2004). Millisecond-to-second interval timing is critical in real-time dynamic tasks that require adaptive responses to the changing environment. For instance, when driving it is necessary to estimate how long one can attend to a navigator before switching back to attending to the road and driving control (Salvucci, Taatgen, & Kushleyeva, 2006).

Time estimation can be studied under various paradigms (Zakay, 1990). Participants can be asked to retrospectively generate verbal estimation of an interval, to judge whether a presented interval is the same length as a target interval, or reproduce a target interval. Studies using the reproduction paradigm typically show response distributions that are 1) centered at the real-time criteria, 2) symmetrical, and 3) have a standard deviations that increase in proportion to the mean interval (e.g. Rakitin, Gibbon, Penney, Malapani, Hinton, & Meck, 1998).

In most studies under those paradigms, time estimation is often an isolated task performed in a static environment. It is the primary task on which participants focus, even when a secondary task is given for various purposes (Fortin, Rousseau, Bourque, Kirouac, 1993; Rakitin, et al., 1998). However, one may wonder to what extent the time estimation performed in those paradigms reflects the time estimation that people usually perform in various

multitasking situations. As in the driving example, time estimation is often an implicit secondary task that one performs to coordinate primary tasks. In addition, people sometimes need to estimate multiple time intervals concurrently (e.g., cooking breakfast). It seems plausible that time estimation in those circumstances will exhibit properties not seen when it is performed as an isolated task in a static environment. We investigated this question in the Space Fortress task (Donchin, 1989), a video game that simulates real-time complex tasks performed in dynamic environments (e.g., piloting an aircraft).

## Time Interval Estimation in Space Fortress Task

The goal of the Space Fortress task (Figure 1) is to maximize the total scores by navigating a ship in a frictionless space, destroying a fortress multiple times and handling mines while protecting the ship from the fortress and mines. The participant navigates the ship by rotating left or right (A/D keys) or thrusting (W key) to make it fly within an area enclosed by two hexagons. A fortress stationed in the center rotates like a turret, tracking the ship's trajectory and firing shells at it. The participant has to shoot the fortress with a missile (spacebar) at least ten times and then make a rapid double-shot to destroy it.

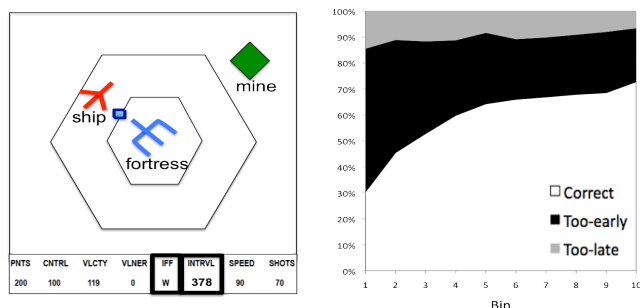


Figure 1: Schematic representation of the Space Fortress task (left) and performance in the IFF tapping task (right).

The mine task, which is the focus of the current study, consists of a series of activities in a specific order. At the beginning of the game, the participant is presented with a screen with three alphabetic letters ('foe letters') and asked to remember them. During the game, a mine appears at a random location on the screen 5 seconds after the destruction of the previous mine and starts pursuing the ship with the intent of crashing into the ship. When a mine appears, the participant has to check a letter that appears in

the IFF box in the bottom panel (see Figure 1). The mine is a foe if the letter matches one of the foe letters; otherwise, it is a friend. Mine identification is a version of the Sternberg memory-scanning task (Sternberg, 1966). If the mine is a foe, one has to perform an Identify Friend/Foe (IFF) tapping task, which involves tapping the J key twice with a 250-400 ms interval ('IFF interval') between the two key presses. Once a correct interval has been generated, the mine can be destroyed by aiming the ship at the mine and firing a missile. A missile can be fired even after a wrong IFF interval, but the missile can destroy the mine only after a correct IFF interval. If the mine is a friend, then the IFF tapping task should not be performed and the mine can be destroyed by a missile shot. If all steps are completed successfully before the mine reaches the ship, then the mine is destroyed and points are earned. Otherwise, the mine eventually collides with the ship and points are lost.

As a time interval estimation task, the IFF tapping task has three notable characteristics. First, it is a prospective time estimation task. Participants are initially told the target interval in written instructions, and then produce the interval whenever a foe mine appears during a game. Immediately after each attempt, the produced interval is displayed as feedback (e.g., "378") in the INTRVL box in the bottom panel. Second, both the initiation and the termination of the interval are under the control of participants. Finally, and most importantly, it is performed not as an isolated task but as part of a real-time complex task. The game requires time-sharing multiple tasks such as navigating the ship while dealing with the fortress and the mines. Even within the mine task, a series of activities precede (checking the letter and determining the mine's identity) and follow (aiming the ship and firing a missile) the IFF tapping task, all of which need to be completed within a brief period of time, usually 2-3 seconds.

A study previously conducted in our laboratory revealed an interesting pattern of performance in the IFF tapping task. Figure 1 (right) displays the percentage of responses within each of three categories: correct (the produced interval was between 250-400 ms), too-early (<250 ms), and too-late (>400 ms) responses. The figure shows the average percentages from 100 participants over 300 attempts (30 attempts per bin). Participants improved with practice, as indicated by the percentage of correct responses reaching almost 70% accuracy by the end. More notable is the error pattern, with participants making too-early responses more often than too-late responses.

This too-early bias deviates from the roughly symmetrical responses observed in time interval estimation studies (e.g. Rakitin, et al., 1998). We suspected two factors might be responsible for the too-early bias. The first possibility is that estimating a shorter time interval contaminated performance in the target interval. In the Space Fortress task, the fortress task involves shooting a fast double-shot (<250 ms interval). Studies (Grondin, 2005; Jones & Wearden, 2004; Taatgen & Van Rijn, 2011) suggest that representations of different time intervals are not independent of each other. Participants

in Taatgen & Van Rijn (2011) study alternated between producing a short interval and a long interval. When the feedback criterion for the long interval was shifted unbeknownst to the participants, not only did the estimate of the long interval change, but the estimate of the short interval also changed. Thus, estimating the shorter interval for the fortress task might have influenced estimating the target interval for the mine task.

A second possibility is that participants might be more likely to commit too-early errors as less time is allowed for the mine task. Note that the mine task consists of multiple demanding activities that are in competition with each other for the limited length of time available for the mine task. One might hypothesize that participants adjust the IFF interval based on their estimation of time remaining to fire a missile before the mine crashes into the ship.

## The IFF Tapping Experiment

We tested those two hypotheses in a within-subjects design by manipulating 1) the speed of tapping (fast/slow) alternated with the IFF (intermediate) tapping, and 2) the distance between ship and mine (short/long) at mine onset. We created three types of games: fast-tap, slow-tap, and intermediate-tap-only games. Those games were a simplified version of the original Pygame Space Fortress task (Destefano, 2010).

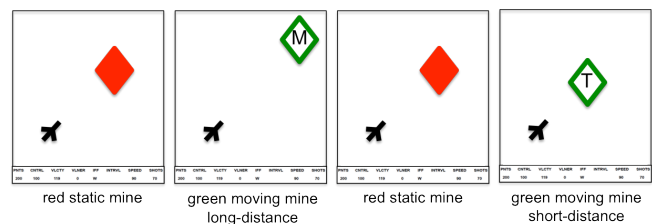


Figure 2: Sample sequence of trials in the fast-tap game.

Figure 2 shows a sample sequence of trials in the fast-tap game. The game had a static ship fixed at the bottom left of the screen always correctly aimed toward the mine that appeared from the other side. During the game, 8 red static and 8 green moving mines appeared in a strictly alternating order. For a red static mine, participants simply had to produce a fast (<250 ms) double-tap (spacebar). In the following trial, a green mine containing a letter appeared and approached the ship. For the green moving mine, participants had to 1) check the letter and determine its identity, 2) produce the IFF interval using an appropriate key (F key for friend and J key for foe), and 3) fire a missile (space bar). If all steps were successfully completed, the mine was destroyed. If any of the three steps were missed or performed incorrectly, the mine became invincible and eventually destroyed the ship. The slow-tap games were identical to the fast-tap games except that they had blue static mines (instead of red static mines) for which participants produced a slow (400-650 ms) double-tap. The distance manipulation was applied to the green moving

mines in the fast-tap and slow-tap games. The distance between ship and mine at the moment of mine onset was randomized to be either short (283 pixels, corresponding to 1.86 s until mine collision) or long (566 pixels, 3.73 s). The intermediate-tap-only games were intended to test whether the too-early bias would still be present when participants produced the target interval without the demands of the mine task and without estimating different time intervals. In each trial, when a letter (either F or J) appeared in the center of the screen, participants simply produced the IFF interval using the corresponding key. Each intermediate-tap-only game had eight trials.

Twenty participants (5 males, mean age: 19 yrs) from Carnegie Mellon University participated in the experiment. The experiment consisted of 12 blocks of games. Each block had one intermediate-tap-only game, one fast-tap game, and one slow-tap game in a randomized order.

### Behavioral Results

Figure 3 (left) presents the IFF tapping performance in intermediate-tap-only games over 12 blocks. Participants overall performed very well (mean accuracy: 86%). Importantly, the too-early bias was not present confirming our prediction. Participants committed too-early and too-late errors with roughly equal frequencies in the first block, but they quickly reduced their too-early errors. Thus, there was a small too-late bias on later trials, which may reflect a floor effect on the shortest intervals participants could produce. Figure 3 (right) shows that the mean produced IFF interval fell within the targeted 250-400 ms range and did not fluctuate much over blocks.

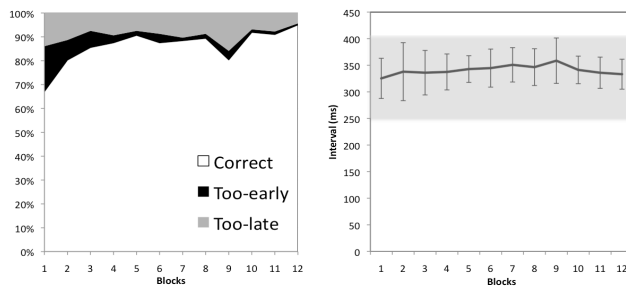


Figure 3: IFF tapping performance in intermediate-tap-only games: percentages of correct/too-early/too-late responses (left) and mean produced IFF intervals (right).

The results from the fast-tap and slow-tap games confirmed both the contamination and the distance hypotheses. Figure 4 displays the performance in the IFF tapping task in the four conditions defined by crossing the tap speed (fast/slow) and the distance (short/long) manipulations: fast-short, fast-long, slow-short, and slow-long. The percentage of correct responses increased over practice in all conditions. In all conditions the too-early responses dominated the first couple of blocks, but afterwards the bias stabilized at a lower level. The largest too-early bias was present in the fast-short condition,

whereas the smallest too-early bias (and the largest too-late bias) was found in the slow-long condition. Note that the fast-short condition best reflects the original Space Fortress game in which participants handle both mines (IFF taps) and the fortress (fast-taps), and have only a short time for the mine task (short-distance).

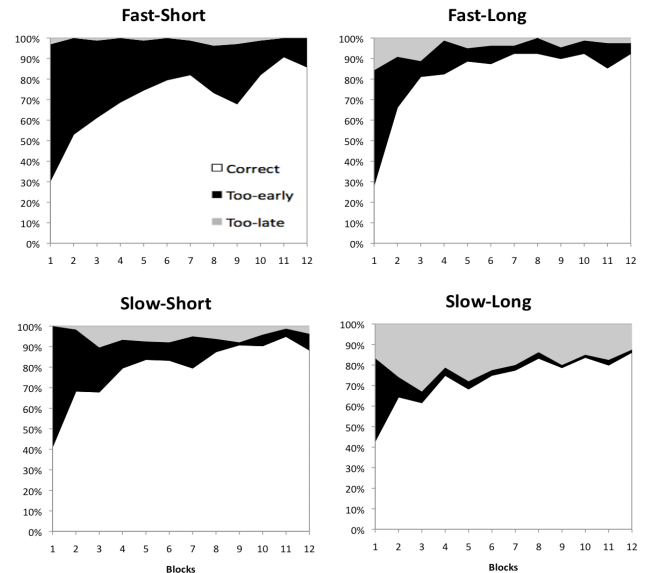


Figure 4: IFF tapping performance in fast/slow-tap games.

A repeated measures analysis of variance was performed with tap speed (fast/slow), distance (short/long), and practice (early: block 1-6 vs. late: block 7-12) as within-subjects factors and the mean produced IFF interval as the dependent measure. There were significant interactions between tap speed and distance ( $F(1,19) = 10.23, p < 0.01$ ), tap speed and practice ( $F(1,19) = 5.13, p < .05$ ), and distance and practice ( $F(1,19) = 11.62, p < 0.01$ ). The interaction between tap speed and distance reflects the larger distance effect in the slow-tap condition compared with the fast-tap condition. The interactions between tap speed and practice, and distance and practice reflect that those effects were larger in earlier blocks than in later blocks. The three-way interaction was not significant.

### The ACT-R Model

We developed a simulation model of our time estimation task, incorporating ideas from the Taatgen and Van Rijn (2011) timing model into a task model based on the ACT-R model for Space Fortress (Bothell, 2010). The model<sup>1</sup> was implemented in the ACT-R architecture (Anderson, Bothell, Byrne, Douglass, Lebiere, & Qin, 2004), which allows us to simulate all aspects of the task, not just the timing component. In ACT-R, time estimation is achieved through the processing in the temporal module (Taatgen, Van Rijn,

<sup>1</sup> Model parameters: :rt 1.0, :lf 1.1, :ans 0.385, :mp 2.25, :time-master-start-increment 0.011, :time-multi 1.1, :time-noise 0.0015.

& Anderson, 2007) and its interaction with the rest of the system. The temporal module, based on the internal clock model (Matell & Meck, 2000), assumes a pacemaker keeps accumulating pulses as time progresses. The production system can access the current pulse value through the temporal module's buffer and compare it with a criterion (e.g., a value retrieved from memory) to determine if the target interval has elapsed.

The model uses an instance-based approach to learn the required tapping times. When the model produces a time interval (e.g., 15 pulses) and observes its outcome (e.g., too-early), the specific instance of that experience is stored in declarative memory as a chunk. This chunk can be retrieved later to serve as a basis for deciding how long to wait the next time the model has to produce the interval. As such chunks are added to memory, the speed of retrieval increases and the accuracy of the retrieved result improves (similar to Logan's 1988 instance theory)

Figure 5 displays the series of steps in which the model performs the IFF tapping task. When a mine appears, the model attends the letter and determines its identity by retrieving the letter from memory. The model then starts retrieving a criterion value for the IFF interval. The retrieval of the criterion value is based on the blending mechanism discussed in the next section. If blending is successful, the model uses the blended result as the criterion. If blending fails, the model uses a default value. Once the criterion is determined, the model issues the first tap and starts incrementing the pulse value in the temporal buffer. When the current pulse value exceeds the criterion, the model issues the second tap that terminates the interval. The model then taps the spacebar to fire a missile. After completing both IFF tapping and the missile firing, the model attends the feedback<sup>2</sup>, evaluates the outcome (too-early, correct, or too-late), and assigns a feedbackshift value (positive for too-early, zero for correct, and negative for too-late) so that the criterion could be appropriately adjusted in the next trial.

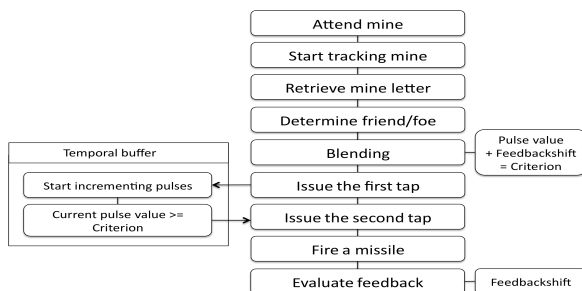


Figure 5: The ACT-R model of the IFF tapping task.

<sup>2</sup> According to our data, approximately 90% of the wrong IFF intervals were followed by a missile firing. We interpret this as indicating that participants tended to execute the entire sequence of key presses as a unit rather than interrupting the sequence after the IFF tapping to attend to feedback.

## Blending

The ACT-R blending mechanism (Lebiere, Gonzalez, & Martin, 2007) was adopted to model the contamination from representations of different time intervals. In the standard retrieval mechanism of ACT-R, a retrieval request results in retrieval of a single chunk with the highest activation that exceeds the retrieval threshold. Blending is an alternative mechanism that allows retrieval of a weighted aggregation of all candidate chunks available in memory. Each candidate chunk is given a different weight based on how recently the chunk has been created and how closely it matches the retrieval request.

Figure 6 illustrates how contamination occurs during the blending in the fast-tap game in which the model alternates between the intermediate-tap and the fast-tap. When the blending request is made for pulse value (the value that was previously used for the 'intermediate-tap' and its outcome was 'correct'), blending is performed for candidate chunks that perfectly match the request (e.g., interval44 with 'intermediate-tap' type and 'correct' outcome) and imperfectly matching chunks (e.g., interval45 with 'fast-tap' type and 'too-late' outcome), with the latter penalized according to their degree of mismatch with the blending request<sup>3</sup>. Due to the contribution of fast-tap chunks (e.g. interval45), the final pulse value (15.551) is smaller than it is supposed to be had only intermediate-tap chunks contribute to blending. The model performed blending separately for pulse value and feedbackshift value, then used the sum of those two ( $15.982=15.661+0.321$ ) as the criterion.

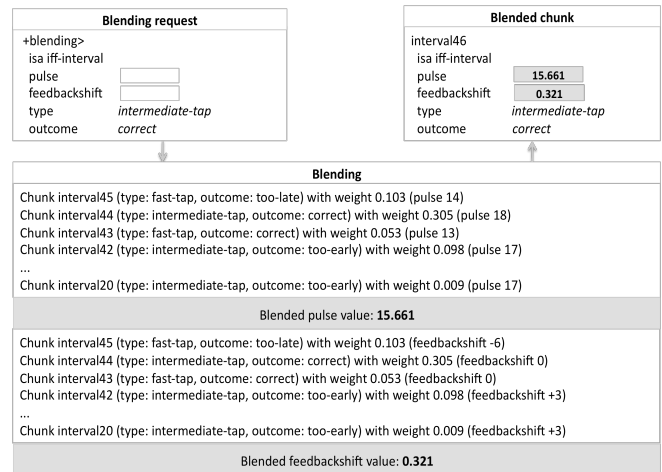


Figure 6: Blending for intermediate-tap.

## Modeling the Distance effect

The model has a production rule that issues the second IFF tap when the current pulse value is greater than or equal to the criterion. We modeled the distance effect by adding an additional 'emergency' production for the second IFF tap.

<sup>3</sup> Partial matching was enabled for the tap type (fast/intermediate/slow) and outcome (too-early/correct/too-late).

During the trial, the model tracks the mine's trajectory by updating the visual-location buffer with the mine's current location. The emergency production specifies a threshold value in pixels that forces the model to issue the second tap such that it will have enough time remaining to fire a missile before it hits the ship. The model ignores the pulse value in the temporal buffer when this production fires.

## Model Results

Contrary to human, we found that the model does not show the burst of too-early responses in early blocks. This is not surprising because the model starts out with a perfect representation of the task instructions, whereas participants have to work out any misunderstandings. Thus, participants show many more start up errors (e.g., no response). Since our goal is not to model this skill learning, we decided to focus on modeling the stable effects in the last 8 blocks, where participants and the model have both mastered the basic task requirements. Figure 7 offers comparisons of human and model performance in the last 8 blocks based on 100 model runs. The model successfully captures not only the lack of a too-early bias in the intermediate-tap-only condition, but also the distance and contamination effects in the other conditions. The overall correlation between model and participants equals .992.

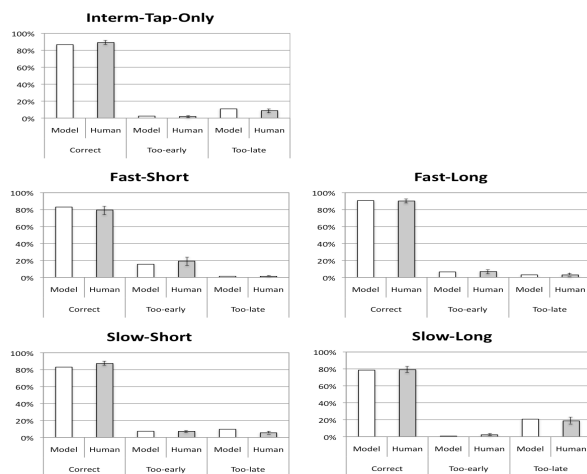


Figure 7: The model fit.

## Discussion

Two factors appear to be responsible for the too-early bias in time estimation that occurs in the context of a dynamic task. First, the representation of the shorter or longer interval shifted the representation of the intermediate interval, supporting the claim (Taatgen & Van Rijn, 2011) that more than a single experience determines the representation of the target interval. The blending mechanism of ACT-R offers quantitative descriptions of interference among time interval representations in declarative memory. Second, the time remaining until the end of the task influenced time interval production. It shows

that time estimation can be sensitive to one's knowledge of what is about to happen, consistent with animal literature (e.g., Church, Miller, Meek, and Gibbon, 1991). Our model captures this by having a procedural rule that overrides the outcome of the internal temporal estimate based on its perceptual processing of the environment.

Regardless of the conditions, participants showed a strong too-early bias in the early blocks (see Figure 4). There are a number of possible explanations for this result. First, participants were likely learning how to speed up other aspects of the task besides the blending process across blocks. In early blocks, these other processes might have been so slow as to increase use of the emergency rule. Second, participants might not even have been trying to time the target interval in the early blocks; instead they may have just practiced the sequence of responses in the task and focused on time estimation only when they had become proficient at responding. The third possible explanation is arousal. Studies (e.g., Penton-Voak, Edwards, Percival, & Wearden, 1996), suggest that arousal can affect the subjective duration of intervals by speeding up the rate at which a temporal pacemaker produces pulses.

The clear contrast between performance in the intermediate-tap-only condition and the other conditions demonstrates that time estimation performed in a dynamic task can exhibit properties different from when it is performed as an isolated task in static environments. External temporal or nontemporal tasks can influence production of the target time interval not only when those tasks are performed concurrently (Van Rijn & Taatgen, 2008) but also when performed in the same context in an alternating order (Taatgen & Van Rijn, 2011). Those results emphasize the virtue of modeling time estimation in the integrated cognitive architecture of ACT-R. The critical aspect of our model is not just the module's internal temporal processing, but also the contributions of the declarative and procedural components. This integrated approach of modeling time estimation in cognitive architecture can be especially useful in understanding time estimation in multitasking situations.

One possible application of our results concerns the training of time estimation tasks. In skill acquisition literature, two instructional strategies, part-task training (e.g., Frederiksen & White, 1989) and integrated training (e.g., Gopher, Weil, & Bareket, 1994), have been compared. The contrast between the intermediate-tap-only condition and the other conditions in our study suggests that a greater training emphasis can be directed to integrating timing with other subtasks (whole-task approach) rather than drilling on timing alone (part-task approach). Good performance in the intermediate-tap-only condition did not transfer to good intermediate timing in the more complex games.

Human factors researchers have studied timing performance and patterns of timing errors in dynamic multitasking situations (e.g., Rantanen & Xu, 2001). Similar to those studies, another potential application of our results regards to improving safety and reducing errors in time-



critical multitasking situations (e.g., traffic environments). Identifying patterns of timing errors and investigating the underlying causes may suggest changes in work procedures. For instance, a timing-critical task can be separated from other tasks that involve less critical timing, putting it under lower time pressure.

The time estimation mechanism in ACT-R has successfully captured time estimation performance in dual-task conditions (Taatgen et al., 2007) as well as in dynamic multitasking situations (Salvucci, et al., 2006). We explored millisecond-level time estimation embedded in a complex real-time task that imposes especially high perceptual-motor demands. The model provided an integrated account of why time estimation performed in this context exhibited different properties than when it was performed in an isolated context. This study further supports the need to model time estimation in the broader context of cognition as we attempt to expand our understanding of human temporal cognition in the domain of complex skills.

### Acknowledgments

This work was supported by ONR grant N00014-09-1-0402 to Wayne Gray & John Anderson.

### References

- Anderson, J. R., Bothell, D., Byrne, M., Douglass, D., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review*, *111*, 1036–1060.
- Bothell, D. (2010). *Modeling Space Fortress: CMU Effort* [PowerPoint slides]. Retrieved from <http://act-r.psy.cmu.edu/workshops/workshop-2010/schedule.html>
- Church, R. M., Miller, K. D., Meek, W. H., & Gibbon, J. (1991). Symmetrical and asymmetrical sources of variance in temporal generalization. *Animal Learning and Behavior*, *19*, 207–214.
- Destefano, M. (2010). *The mechanics of multitasking: The choreography of perception, action, and cognition over 7.05 orders of magnitude*. Unpublished doctoral dissertation, Rensselaer Polytechnic Institute, Troy, NY.
- Donchin, E. (1989). The learning strategies project. *Acta Psychologica*, *71*, 1–15.
- Fortin, C., Rousseau, R., Bourque, P., & Kirouac, E. (1993). Time estimation and concurrent nontemporal processing: Specific interference from short-term-memory demands. *Perception & Psychophysics*, *53*, 536–548.
- Frederiksen, J. R., & White, B. Y. (1989). An approach to training based upon principled task decomposition. *Acta Psychologica*, *71*, 89–146.
- Gopher, D., Weil, M., & Bareket, T. (1994). Transfer of skill from a computer game trainer to flight. *Human Factors*, *36*, 387–405.
- Grondin, S. (2005). Overloading temporal memory. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 869–879.
- Ivry, R., Spencer, R. M., Zelaznik, H. N., & Diedrichsen, J. (2002). The cerebellum and event timing. In S.M. Hightstein and W.T. Thach (Eds.), *The cerebellum: Recent developments in cerebellar research*. *Annals of the New York Academy of Sciences*. New York: New York Academy of Sciences.
- Jones, L. A., & Wearden, J. H. (2004). Double standards: Memory loading in temporal reference memory. *Quarterly Journal of Experimental Psychology*, *57B*, 55–77.
- Jones, M. R. (1990). Musical events and models of musical time. In R. Block (Ed.), *Cognitive models of psychological time*. Hillsdale, NJ: Lawrence Erlbaum.
- Lebiere, C., Gonzalez, C. & Martin, M.K. (2007). Instance-based decision making model of repeated binary choice. *Proceedings of the 8th International Conference on Cognitive Modeling* (pp. 77–82). Ann Arbor, MI.
- Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.
- Matell, M. S., & Meck, W. H. (2000). Neuropsychological mechanisms of interval timing behavior. *BioEssays*, *22*, 94–103.
- Penton-Voak, I. S., Edwards, H., Percival, A., Wearden, J. H. (1996). Speeding up an internal clock in humans? Effects of click trains on subjective duration. *Journal of Experimental Psychology: Animal Behavior Processes*, *22*, 307–320.
- Rakitin, B. C., Gibbon, J., Penney, T. B., Malapani, C., Hinton, S. C., & Meck, W. H. (1998). Scalar expectancy theory and peak-interval timing in humans. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*, 15–33.
- Rantanen, E. M. & Xu, X. (2001). Human performance in timing of discrete actions. *Proceedings of the 45<sup>th</sup> Annual Meeting of the Human Factors and Ergonomics Society* (pp. 527–531). Santa Monica, CA.
- Salvucci, D., Taatgen, N., & Kushleyeva, Y. (2006). Learning when to switch tasks in a dynamic multitasking environment. *Proceedings of the Seventh International Conference on Cognitive Modeling* (pp. 268–273). Trieste, Italy.
- Schirmer, A. (2004). Timing speech: A review of lesion and neuroimaging findings. *Cognitive Brain Research*, *21*, 269–287.
- Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*(3736), 652–654.
- Taatgen, N.A. & Van Rijn, H. (2011). Trace of times past: Representations of temporal intervals in memory. *Memory & Cognition*, *39*, 1546–1560.
- Taatgen, N., Van Rijn, H., & Anderson, J. R. (2007). An integrated theory of prospective time interval estimation: The role of cognition, attention, and learning. *Psychological Review*, *114*, 577–598.
- Van Rijn, H., & Taatgen, N. A. (2008). Timing of multiple overlapping intervals: How many clocks do we have? *Acta Psychologica*, *129*, 365–375.
- Zakay, D. (1990). The evasive art of subjective time measurement: Some methodological dilemmas. In R. A. Block (Ed.), *Cognitive models of psychological time*. Hillsdale, NJ: Erlbaum.

# The Role of Gesture in Second Language Learning: Communication, Acquisition, & Retention

**Laura M. Morett (morett@pitt.edu)**

Department of Psychology, University of Pittsburgh  
210 S. Bouquet Street, Pittsburgh, PA 15260 USA

**Raymond W. Gibbs (gibbs@ucsc.edu)**

Department of Psychology, University of California, Santa Cruz  
1156 High Street, Santa Cruz, CA 95064 USA

**Brian MacWhinney (macw@cmu.edu)**

Department of Psychology, Carnegie Mellon University  
5000 Forbes Ave., Pittsburgh, PA 15213 USA

## Abstract

Previous research has provided evidence that second language (L2) learners use gesture to enhance spoken communication with interlocutors, and that gesture facilitates L2 word learning. The current study investigates how L2 learners use gesture to communicate in conversational settings, and whether their gesturing in these settings facilitates L2 acquisition beyond the immediate context. The results reveal that L2 learners produce more iconic gestures when their interlocutor is visible, and that gesture production predicts their recall for novel words introduced in conversation. As such, they indicate that conversational gesturing enhances acquisition of the target language more broadly, facilitating L2 communication, acquisition, and retention.

**Keywords:** Second language acquisition, word learning, gesture, mental imagery, embodied cognition

## Introduction

Language acquisition, like other aspects of human communication, is profoundly multimodal. Children learning how to speak often use their hands and bodies to express themselves, even before they can speak (Acredolo & Goodwyn, 1988). Once children begin speaking, they begin producing gestures concurrently with speech to help them express themselves (Iverson & Goldin-Meadow, 2005). However, it is less well-known how gesture facilitates second language (L2) learning by adults. Although previous research has provided evidence that isolated gestures can enhance L2 word learning in instructional settings (Allen, 1995; Kelly, McDevitt, & Esch, 2009; Tellier, 2008), it is unclear whether L2 learners' discourse comprehension benefits from conversational gestures. Furthermore, it is unclear whether L2 learners' gestures facilitate communication with interlocutors in the target language, or whether they benefit only the speaker by providing a method of "thinking out loud." The present research addresses these questions by examining how interlocutor visibility and gestural priming affect gesture use among L2 learners. Moreover, this research examines whether enhanced real-time comprehension transfers to other domains, benefiting L2 learning more broadly.

## How Gesture Affects L2 Communication

To date, little research has addressed how gesture affects communication between L2 learners and their interlocutors in real-time conversation. Work examining the question of whether speakers use gesture in a communicative sense in general has provided evidence that speakers modify their gesture qualitatively and quantitatively when interlocutors are visible. In particular, this work has revealed that speakers produce gestures that are more iconic, elaborate, and larger when they are speaking to an interlocutor who is present and visible, as opposed to an interlocutor who is present but occluded or an interlocutor who is on the phone (Alibali, Heath, & Myers, 2001; Bavelas, Gerwing, Sutton, & Prevost, 2008). These results suggest that speakers do indeed use gesture as a communicative device.

If speakers use gesture to facilitate communication, then it follows that they should also modify their gestures when conversing with L2 learners, who often encounter difficulty comprehending L2 speech. Indeed, related research has shown that native speakers hyperarticulate vowels when speaking to adult L2 learners, demonstrating that they are aware of L2 learners' comprehension-related needs (Uther, Knoll, & Burnham, 2007). In one of the few studies to examine native speakers' gesture in the presence of L2 learners, native speakers of English produced more deictic and iconic gestures when retelling a narrative to L2 English learners than to English-speaking interlocutors (Adams, 1998). This finding suggests that speakers rely on gesture as a communicative medium to a greater degree with L2 learners than with native interlocutors.

There is also evidence that the gestures that L2 learners produce when speaking the target language differ qualitatively from the gestures that they produce when speaking their native language, as well as from gestures produced by native speakers of the target language. One such difference is that gestures accompanying L2 learners' utterances in the target language tend to be over-explicit. For example, L2 learners are much more likely to produce iconic gestures when referring to entities that they have previously mentioned when speaking the target language



than their native language (Gullberg, 2003). This tendency mirrors their use of nouns rather than pronouns to refer to previously mentioned entities in the target language. Other work has provided evidence that the syntax and semantics of L2 learners' native language influences the gestures that they produce when speaking the target language. As a result, L2 learners' gestures differ in subtle ways from the gestures of native speakers of the target language, such as in the way that path and manner are expressed (Stam, 2006; Yoshioka & Kellerman, 2006). Nevertheless, it is unclear whether native speakers of the target language notice these differences in L2 learners' gesture, and to date, there is no conclusive evidence that they hinder communication between L2 learners and native interlocutors.

In fact, despite the nuanced ways in which L2 learners' gesture may differ from the gesture of native speakers, there is evidence that gesture *facilitates* communication between L2 learners and native interlocutors. To this end, one study revealed that L2 learners frequently gestured when they were unable to complete an utterance in Japanese, prompting Japanese interlocutors to suggest semantically appropriate completions (Mori & Hayashi, 2006). Along similar lines, a case study of gesture use between a L2 English learner and his tutor revealed that the learner often gestured when searching for words, cuing the tutor to complete the learner's utterances while simultaneously producing similar gestures (McCafferty, 2002). Taken together, these findings reveal that both L2 learners and native speakers use gesture to convey information that their interlocutors may not be able to comprehend via speech, serving as an alternative channel of symbolic communication. Furthermore, these findings suggest that both L2 learners and native interlocutors can understand and use information conveyed via gesture to rectify incomprehensible or incomplete utterances, thereby facilitating communication in the target language.

### How Gesture Affects L2 Acquisition and Retention

Given the evidence that L2 learners can use gesture to facilitate their production and comprehension of the target language in conversational settings, it follows that gesture may facilitate not only communication, but also *acquisition* and *retention* of the target language. Indeed, research has provided evidence that gesture can enhance the acquisition and recall of L2 lexical items. One study showed that first-semester L2 students learned French figurative expressions accompanied by representational gestures more effectively than expressions that were presented without gesture (Allen, 1995). Similarly, individuals unfamiliar with Japanese recalled the meanings of Japanese verbs over the course of a week more accurately when they were presented with representative iconic gestures than when they were presented as speech only, without gesture (Kelly et al., 2009). Finally, during instruction spanning 4 weeks, children learning English recalled more words from this language when they enacted representative iconic gestures than when they learned the word meanings via images.

Taken together, these results indicate that iconic gesture may help beginning L2 learners to associate L2 lexical items with their meanings, thereby facilitating L2 word learning.

Despite the preponderance of evidence that gesture facilitates L2 lexical acquisition, research examining the effect of gesture on other aspects of L2 acquisition has been rather sparse and has produced inconsistent findings. The results of several studies have failed to show evidence that beat (simple rhythmic), deictic (pointing), or iconic gestures can enhance the perception of novel L2 phonological contrasts (Hirata & Kelly, 2010; Jesse & Mitterer, 2011; Kelly & Lee, 2011). To date, no published research has investigated whether gesture can facilitate the acquisition of morphosyntax by L2 learners. On a broader level, some work (Lazaraton, 2004; Sueyoshi & Hardison, 2005) has provided evidence that gesture facilitates L2 learners' comprehension of instructional lectures given in the target language. However, it is unclear from this work what types of gestures enhance listening comprehension, as well as which aspects of comprehension they affect.

Although previous research has not directly investigated the cognitive processes responsible for gesture's impact on various aspects of L2 acquisition, it is possible to make some inferences about them by considering the foci and design of the studies discussed above. First, gesture's facilitation of the acquisition of lexical items but not phoneme perception suggests that the benefits of gesture in L2 learning may be localized to individual stimuli, rather than generalizable to related members of a category. This interpretation is particularly plausible in light of the fact that most of the research investigating the effect of gesture on L2 acquisition has been conducted with novice L2 learners, who lack an understanding of how various elements of the target language are related. Second, the observation that representative iconic gestures—but not non-representative iconic gestures or other types of gestures—enhance L2 lexical acquisition suggests that isomorphism between gestures and the visuospatial properties of referents is necessary for gesture to benefit L2 word learning. Third, the more pronounced effect of gesture enactment than gesture viewing on L2 lexical acquisition suggests that the engagement of embodied action may also play a key role in explaining gesture's facilitation of L2 word learning.

## Methods

### Participants

Sixty undergraduates were recruited in pairs from the participant pool at UCSC. All recruited individuals were fluent English speakers<sup>1</sup> and confirmed that they had no knowledge of Hungarian prior to the experiment. Additionally, all recruited individuals had normal hearing and normal or corrected-to-normal vision.

<sup>1</sup> Participants were not required to be native English speakers, given that the English glosses of the signs were common words that should be comprehensible to non-native undergraduates, who must be sufficiently proficient to understand academic English.

Table 1: Hungarian and English words used in study.

Hungarian	English	Hungarian	English
<i>Betegség</i>	Illness	<i>Unott</i>	Bored
<i>Kalapács</i>	Hammer	<i>Varrni</i>	To sew
<i>Kulcs</i>	Key	<i>Öltözet</i>	Clothing
<i>Löni</i>	To shoot	<i>Leforgatni</i>	To record
<i>Mászni</i>	To climb	<i>Csomó</i>	Knot
<i>Megütni</i>	To hit	<i>Hosszú</i>	Long
<i>Órá</i>	Watch	<i>Bajusz</i>	Moustache
<i>Öröm</i>	Joy	<i>Testgyakorlás</i>	Sports
<i>Seprű</i>	Broom	<i>Kezbasiteni</i>	To deliver
<i>Tréfa</i>	Joke	<i>Kefe</i>	Brush

### Stimuli

Twenty Hungarian words were selected for use in this study (see Table 1). Prior to this research, 15 English speakers who did not participate in this study were asked to rate the concreteness, imageability, and meaningfulness of the English glosses of 80 candidate words, and to gesture in a way that represented the meaning of each gloss. The 20 words with the most consistent responses from each of the three categories listed above were selected for inclusion in the study. Videos of iconic gestures were created by recording a fluent Hungarian-English bilingual saying these words in each language while enacting the gestures produced by the most participants. In order to control for possible vocal iconicity, audio of the pronunciation of Hungarian and English words was extracted from the iconic gesture videos and was played during presentation of text words in the non-gesture presentation condition.

### Procedure

In the learning phase of the experiment, participants were randomly assigned via coin flip to be either the teacher or the learner, and to be in either the visible or occluded condition. Participants were seated on either side of a table that was either unobstructed or was divided by a large, opaque cardboard screen. The teacher was seated in front of an iMac G4 with a 20 in. screen, on which the stimuli were presented. Prior to beginning the experiment, participants were told that the focus of the study was to examine teaching and learning strategies for L2 vocabulary. To this end, the teacher was told that he or she would learn the meanings of twenty Hungarian words one-by-one. The teacher was instructed to teach each word to their partner “however [they] think the learner will learn them best,” with the only restriction being that both the teacher and learner must remain seated at the table during the entire learning phase. The learner was told that he or she would be tested after the learning phase to determine how well he or she had learned the words. After ensuring that the participants understood the instructions, the experimenter left the room.

Although participants were informed that their speech would be audio recorded during the learning phase for later analysis, they were not told that they would also be video

recorded while learning the words. A video camera hidden behind a one-way mirror oriented perpendicular to the table was focused on participants during the learning phase of the experiment, providing a 180° view of them (see Figure 1). The screen of the computer on which stimuli were presented was oriented at a 180° angle from the camera, in order to ensure that the experimenters were blind to the presentation modality of the words. Participants’ interactions were never heard or seen by anyone other than the experimenters.

During trials of the learning phase, a Hungarian word was presented to the teacher for 2500 ms., and after a 1000 ms. interval, its gloss was presented for 2500 ms. Words in both languages were presented to the teacher as speech over headphones connected to the computer. Words were presented visually via either video of a Hungarian-English bilingual saying them while concurrently producing an iconic gesture representing their meaning, or via text displayed on screen; word presentation modality was varied within participants. Following a 2000 ms. interval, the words were repeated once in this sequence, and then a screen instructing the teacher to teach the word to the learner was displayed until the teacher pressed a button to indicate readiness to proceed to the next trial. After all 20 words had been presented, participants were told to summon the experimenter so that the learner could be tested.

In the test phase, the teacher was informed that he or she would also be tested to gauge how well he or she had learned the words. Both participants were placed in separate rooms for this part of the experiment. During test trials, each Hungarian word that participants had learned was presented as text and speech. Participants responded by saying the corresponding English word or by saying “skip” if they could not remember it. After having completed the test phase, participants were debriefed and informed of the actual purpose of the experiment, and were given the opportunity to have their recordings destroyed. All declined.

### Coding

Participants’ speech was transcribed verbatim, and all gestures were identified. Gestures were classified as one of three types: iconic (handshape and/or motion resembles referent attributes), beat (non-iconic emphasizing), or deictic (pointing). Individual gestures were distinguished from one another by a change in hand shape or motion. For



Figure 1: Screenshot of video captured during learning phase of participants assigned to occluded condition.

Table 2: Average amount of speech and target words produced during learning task, by participant and condition.

PP	Measure	Visible	Occluded
Teach	Total speech	633.30 (305.89)	856.83 (486.37)
	Target words	103.70 (14.10)	94.54 (22.04)
Learn	Total speech	352.00 (214.97)	584.67 (370.83)
	Target words	110.10 (18.89)	106.35 (33.43)

example, an extension of the hand forward with the palm up produced concurrently with the word “to deliver” was coded as one iconic gesture. If a similar movement occurred as the first motion of a sequence in which the fingers of the hand closed and the hand moved in a lateral turning motion (for key), this entire sequence was coded as one iconic gesture.

## Results

### Learning task

In order to investigate the relationship between speech and concurrent gestures produced during the learning task, we first examined the quantity and content of participants’ speech. A univariate ANOVA conducted on total amount of speech produced in the learning task revealed main effects of participant role,  $F(1, 59) = 6.22, p = .02, \eta_p^2 = .14$ , and visibility,  $F(1, 59) = 4.23, p = .05, \eta_p^2 = .10$ . However, the interaction between these factors failed to reach significance. Bonferroni-corrected post hoc tests showed that teachers spoke significantly more than learners, ( $p = .02$ ), and that occluded interlocutors spoke more than visible interlocutors ( $p = .05$ ; see Table 2). Nevertheless, all participants repeated the target words with comparable frequency across the visible and occluded conditions, suggesting that differences in speech were caused by verbal elaboration rather than repetition of target words.

Prior to analysis, the distribution of gesture data was examined to determine whether it was normal. Shapiro-Wilk tests revealed that none of the distributions of any of the gesture types for either participant were normal. As can be seen from Table 3, this was primarily due to the skewness and kurtosis of the data. Because analysis of variance is robust against violations of normality (Hays, 1973), it was used as the primary method of analysis. However, non-parametric tests were also conducted to ensure the validity of the results obtained using parametric statistics.

Before conducting the main analyses, it was necessary to determine whether the number of gestures produced by

teachers and learners differed reliably for pairs assigned to the visible and occluded conditions. Independent-samples  $t$  tests revealed that the difference between the number of iconic gestures produced by teachers and learners whose interlocutor was visible was marginally significant,  $t(29) = 1.92, p = .07, d = .86$  (this was confirmed by a Mann-Whitney  $U$  test,  $U(29) = 22.50, p = .04$ ). No other significant differences were found between teachers’ and learners’ production of iconic, beat, or deictic gestures. Based on the results of these preliminary analyses, teachers’ and learners’ iconic gestures were analyzed separately, but beat and deictic gestures were collapsed across participants.

To determine whether interlocutor visibility and word presentation mode affected iconic gesture production, iconic gesture data was submitted to two repeated measures ANOVAs (one for the teacher and one for the learner). The analysis of teachers’ iconic gesture showed a main effect of word presentation mode,  $F(1, 29) = 21.33, p < .001, \eta_p^2 = .50$ , and interlocutor visibility,  $F(1, 29) = 6.95, p = .02, \eta_p^2 = .25$ ; however, the interaction between these two factors failed to reach significance (see Figure 2). The effect of word presentation mode on teachers’ iconic gesture was confirmed by a Wilcoxon signed-rank test,  $Z(29) = 3.61, p < .001$ , and the effect of visibility was confirmed by a Mann-Whitney  $U$  test,  $U(29) = 23.50, p = .02$ . Bonferroni-corrected post hoc tests revealed that teachers produced a significantly greater number of iconic gestures when conveying the meanings of words that they had learned via iconic gesture than words that they had learned via text ( $p < .001$ ), and that they produced more iconic gestures when communicating with visible interlocutors than occluded interlocutors ( $p = .02$ ). The analysis of learners’ iconic gesture failed to show a main effect of either presentation mode or interlocutor visibility; similarly, the results of a Wilcoxon signed-rank test and a Mann-Whitney  $U$  test failed to reach significance. However, in this case, the interaction between word presentation mode and interlocutor visibility reached significance,  $F(1, 29) = 4.71, p < .04, \eta_p^2 = .18$ ; see Figure 2. Taken together, these results provide evidence that the acquisition of novel word meanings via iconic gestures and the presence of a visible interlocutor cause L2 learners to increase their production of iconic gestures, thereby facilitating communication.

To determine whether interlocutor visibility and word presentation mode affected participants’ production of beat and deictic gestures, data for each of these gesture types was examined using a repeated measures ANOVA.

Table 3: Descriptive statistics and normality test results for gesture types, by participant role.

Gesture	Teacher			Learner		
	Iconic	Beat	Deictic	Iconic	Beat	Deictic
Mean (SD)	11 (10.17)	20.17 (26.63)	3.96 (4.25)	5.13 (7.61)	18.09 (37.10)	1.96 (5.10)
Range	0-40	0-121	0-15	0-129	0-177	0-24
Skewness	1.24	2.87	0.86	2.01	3.88	4.00
Kurtosis	1.62	9.44	0.17	3.83	16.71	17.38
Shapiro-Wilk	.89, $p = .01$	.66, $p < .001$	.85, $p = .003$	.71, $p < .001$	.50, $p < .001$	.43, $p < .001$

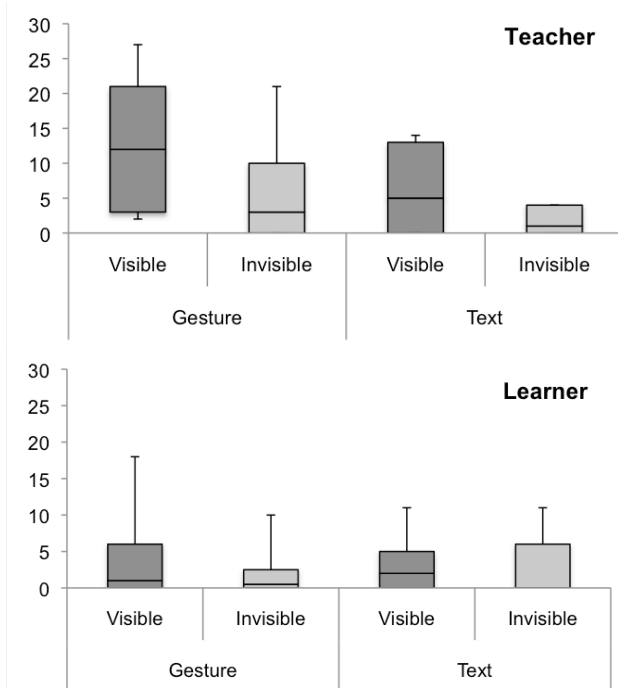


Figure 2: Boxplots showing number of iconic gestures produced by teachers and learners, by word presentation mode and interlocutor visibility.

This analysis revealed no significant effects of either interlocutor visibility or presentation mode, indicating that participants produced beat and deictic gestures with similar frequency, regardless of these factors.

## L2 Word Recall

L2 word recall was quantified using a binary coding scheme for each word (1 = correct, 0 = incorrect/skipped), which was summed across words to yield a total number of words recalled out of twenty for each participant. Prior to investigating the gesture data, we first examined whether total amount of speech and target word repetitions predicted word recall accuracy.

Three separate linear regression analyses revealed that the amount of speech produced by teachers and learners significantly predicted their word recall, and that the amount of teachers' speech also predicted learners' recall.

However, additional analogous analyses failed to show that repetitions of target words by teachers and learners during the learning phase predicted their own word recall at test, or that teachers' target word repetitions during learning predicted learners' recall at test (see Table 4).

In order to examine whether the viewing of iconic gestures during the learning phase affected L2 word recall for the teacher, word recall data was submitted to a paired-samples *t* test using word presentation modality as a within-participants factor. This analysis showed that teachers recalled words with comparable accuracy regardless of whether the words had been presented with iconic gestures or text in the learning phase,  $t(29) = 1.08$ ,  $p = .29$ ,  $d = .20$ . In order to examine whether gesture viewing during the learning phase affected word recall for the learner, learners' word recall scores were regressed on number of iconic gestures produced by teachers during the learning phase. This analysis showed that teachers' iconic gestures failed to predict learners' word recall accuracy,  $b = -.28$ ,  $t(29) = 1.04$ ,  $p = .31$ . Together, these results indicate that simply viewing iconic gestures during L2 word learning is insufficient to promote word recall.

Finally, in order to determine whether gesture enactment affected L2 word learning, teachers' and learners' word recall scores were regressed on the number of gestures (iconic, beat, and deictic) that they produced during the learning phase. This analysis revealed that teachers' beat and deictic gestures positively predicted their recall of L2 words, but that their iconic gestures negatively predicted L2 word recall. Moreover, this analysis failed to reveal that learners' gestures predicted their word recall at test (see Table 4).

## Discussion

The current study examined the impact of gesture production on communication and acquisition of a novel second language. The results showed that participants produced more iconic gestures when the meanings of L2 words are conveyed to them via gesture, and when they are communicating with a visible interlocutor. This indicates that viewing iconic gestures primes the production of iconic gestures. More importantly, however, it demonstrates that interlocutors use iconic gesture to facilitate communication in a novel second language, corroborating the findings of other studies of

Table 4: Word recall, as predicted by teachers' and learners' speech and iconic, beat, and deictic gesture production.

Predictor	Dependent variable	Teacher			Learner		
		<i>b</i>	<i>SE b</i>	$\beta$	<i>b</i>	<i>SE b</i>	$\beta$
Speech	Constant	1.95	1.21		4.02	1.52	
	Total speech	.008	.001	-.77	.006	.003	.46
	Word repetitions	-.04	.03	-.29	.02	.02	.20
Gesture	Constant	5.84	.97		5.03	1.21	
	Iconic gestures	-.18	.08	-.45	.24	.17	.44
	Beat gestures	.08	.03	.49	.07	.08	.23
	Deictic gestures	.62	.21	.63	-.09	.27	-.11

L2 learners' conversational interactions (Adams, 1998; McCafferty, 2002; Mori & Hayashi, 2006). Furthermore, the results of the current study revealed that participants produced more gestures when their interlocutor was visible, but more speech when their interlocutor was occluded. This dissociation indicates that interlocutors are aware of the communicative properties of each modality, and use them appropriately when conveying L2 words.

More interestingly, however, the results of the current study provide evidence that participants' use of beat and deictic gesture during conversational interactions enhanced their L2 acquisition beyond the immediate communicative context, whereas their use of iconic gestures did not produce this effect. Considered as a whole, these results suggest that gesture effectively supplements information conveyed via speech, but cannot replace it. Moreover, the lack of facilitation of gesture viewing for both teachers and learners is inconsistent with work showing that mere exposure to gesture facilitates L2 word learning (Allen, 1995; Kelly et al., 2009), but is consistent with work showing that gesture enactment enhances L2 word learning more effectively than gesture viewing (Tellier, 2008). Future research should investigate the circumstances under which gesture enactment and viewing differentially benefit L2 acquisition, further clarifying its role in the communication, acquisition, and retention of a novel second language.

### Acknowledgements

This research was supported by a National Defense Science and Engineering Graduate Fellowship and the Perlino Award to Laura M. Morett. The authors thank Eve LeBarton and Jana Iverson for helpful discussion.

### References

- Acredolo, L., & Goodwyn, S. (1988). Symbolic gesturing in normal infants. *Child Development*, 59(2), 450–466.
- Adams, T. W. (1998). *Gesture in foreigner talk*. University of Pennsylvania, Philadelphia, PA. Retrieved from <http://repository.upenn.edu/dissertations/AAI9829850/>
- Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, 44(2). doi:10.1006/jmla.2000.2752
- Allen, L. Q. (1995). The effects of emblematic gestures on the development and access of mental representations of French expressions. *The Modern Language Journal*, 79(4), 521–529.
- Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58(2), 495–520. doi:10.1016/j.jml.2007.02.004
- Gullberg, M. (2003). Gestures, referents, and anaphoric linkage in learner varieties. *Information structure and the dynamics of language acquisition*. John Benjamins Publishing Company.
- Hays, W. L. (1973). *Statistics for the social sciences* (Vol. 410). Holt, Rinehart and Winston New York.
- Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research*, 53(2), 298.
- Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, 16(5), 367–371. doi:10.1111/j.0956-7976.2005.01542.x
- Jesse, A., & Mitterer, H. (2011). Pointing Gestures do not Influence the Perception of Lexical Stress.
- Kelly, S. D., & Lee, A. L. (2011). When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high.
- Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, 24, 313–334. doi:10.1080/01690960802365567
- Lazaraton, A. (2004). Gesture and speech in the vocabulary explanations of one ESL teacher: A microanalytic inquiry. *Language Learning*, 54(1), 79–117. doi:10.1111/j.1467-9922.2004.00249.x
- McCafferty, S. G. (2002). Gesture and creating zones of proximal development for second language learning. *The Modern Language Journal*, 86(2), 192–203. doi:10.1111/1540-4781.00144
- Mori, J., & Hayashi, M. (2006). The achievement of intersubjectivity through embodied completions: A study of interactions between first and second language speakers. *Applied Linguistics*, 27(2), 195–219. doi:10.1093/applin/aml014
- Stam, G. (2006). Thinking for speaking about motion: L1 and L2 speech and gesture. *IRAL - International Review of Applied Linguistics in Language Teaching*, 44(2), 145–171. doi:10.1515/IRAL.2006.006
- Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, 55(4), 661–699. doi:10.1111/j.0023-8333.2005.00320.x
- Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, 8, 219–235.
- Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, 49(1), 2–7. doi:10.1016/j.specom.2006.10.003
- Yoshioka, K., & Kellerman, E. (2006). Gestural introduction of ground reference in L2 narrative discourse. *IRAL - International Review of Applied Linguistics in Language Teaching*, 44(2), 173–195. doi:10.1515/IRAL.2006.007

# The Role of Imitation in Generating a Shared Communication System

Junya Morita, Takeshi Konno, Takashi Hashimoto

{j-morita,t-konno,hash}@jaist.ac.jp

School of Knowledge Science, Japan Advanced Institute of Science and Technology

1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

## Abstract

What types of learning or reasoning are involved in forming a new communication system? To answer this question, this paper presents a computational model for forming a new communication system. The model was developed with ACT-R (Adaptive Control of Thought-Rational). In the model, two agents autonomously assign their roles to themselves. Agents also possess general learning mechanisms implemented in ACT-R. By incorporating imitative learning into these general learning mechanisms, this paper studies the role of imitation in the process of forming a new communication system. Finally, we compared the proposed model against a human experiment. The results of the simulation indicate that through imitation, after a short period of interaction, an isomorphic system is created. The result of the simulation also suggests the existence of imitation in the process of forming a human communication system.

**Keywords:** Communication; Imitation; ACT-R

## Introduction

People try to communicate with others even when they do not have a common language. They also understand intentions of others through repeated interactions. Apparently, humans have the ability to develop a new communication system where only a few common ground rules are shared in advance. How can a new communication system be developed? What types of learning or reasoning are involved in this process? Addressing these questions will contribute not only to understand the origins of our communication but also to predict changes in our communication in this era of globalization.

Some researchers have examined these questions by designing communication environments in the laboratory (for a review Scott-Phillips & Kirby, 2010). For example, Galantucci (2005) conducted an experiment to observe the formation of communication systems in which a pair of participants communicated through a medium that restricted the use of standard communication means such as utterances and letters. He observed the process of forming a new communication system, and discussed that implicit information was conveyed through routine behavior, and reported that a temporal order of messages was built into communication systems. However, this study cannot answer the questions above because he did not identify the cognitive mechanism involved in this process.

The present study focuses on imitation as the type of reasoning in forming a new communication system. Imitation has been investigated in various fields of cognitive science (Barnes & Thagard, 1997; Gergely, Bekkering, & Király, 2002; Tomasello, 1999; Thagard, 2001). For example, Tomasello (1999) described the role of imitation in language acquisition by the infant/child. He especially argued









that a type of imitation called “role-reversal”, in which the child aligns herself with the adult speaker, is essential for producing a communicative symbols.

However, there is an important difference between language acquisition and forming a new communication system. In language acquisition, there is a clear distinction between a learner and an instructor (or a demonstrator). On the other hand, a new communication system usually emerges from a situation where no predefined roles exist. Although several studies concerning imitation exist, only few studies have dealt with this situation.

We argue that the model-based approach is the best method to explore the role of imitation in an interactive situation. From this perspective, we present a computational model, in which two agents autonomously assign roles to themselves. Agents also possess general learning mechanisms such as reinforcement learning and instance-based learning to form a new communication system. By incorporating imitative learning into these general learning mechanisms, we investigate the role of imitation in the process of forming a new communication system. Furthermore, we compare the constructed model against a human experiment. The results of this comparison reveal the cognitive mechanism involved in the formation of a human communication system. Before presenting our model, we will provide an overview of our previous experiment.

## Experiment

The present study simulates the experiment reported in Konno, Morita, and Hashimoto (in press), where we modified and used a coordination game taken from Galantucci (2005). As in Galantucci’s study, the game environment contained two characters, each controlled by a player, and four intercommunicating rooms. The game was composed of several repeated rounds. At the beginning of a round, characters were randomly placed in two different rooms. Players were unaware of the location of each other and aimed to bring their characters to the same room. The characters could not move to rooms that were located diagonally. Owing to this constraint, players need to communicate before moving their characters.

Figure 1 presents the flow of each round consisting of three steps: step 1 for exchanging messages; step 2 for moving characters; and step 3 for confirming the result of the movement. Among these steps, step 1 is the most crucial for the success of this task. In this step, the two players construct their own messages composed of two figures such as “”. Using six available figures: , , , , , and . The meanings and usage of the figures were not shared among



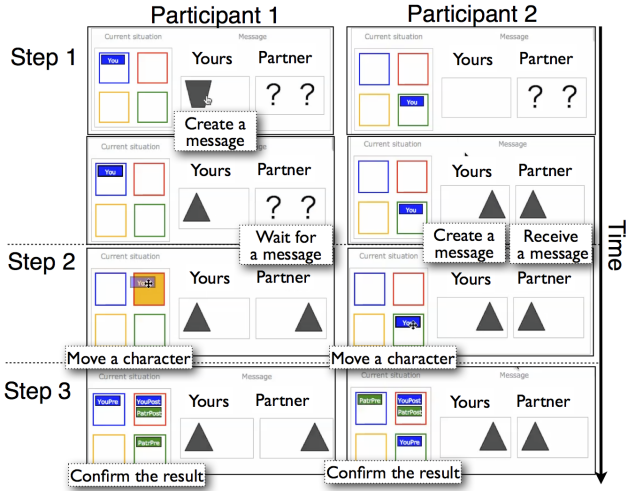


Figure 1: A round of the coordination game consists of three steps. In step 1, to create a message, participants select figures by clicking the segments indicated by “Yours”. Each time a participant clicks a segment, a figure appears in the order of , , , , and . In step 2, a character (blue boxes indicated by “You”) is moved by drag-and-drop. In step 3, the result of the movement are shown to the participants. Blue boxes (“You-Pre” and “You-Post”) and green boxes (“Pat-Pre” and “Pat-Post”) represent a movement made by the participant and the partner, respectively. In the case of this figure, they succeeded in moving to the same room (the upper-right room).

the participants in advance. Each player could send only one message per round, but they could take turns in exchanging messages. A message sent by the first sender instantly appeared on the screen of the other player. The second sender could compose her/his message after observing the message of her/his partner (see the participant 2 in Figure 1). By this turn-taking approach (role-settings), the first sender could transmit her/his current room location, and the second sender could transmit the destination while taking into account the current room location of her/his partner. Importantly, the participants were not assigned their roles by the experimenter. They were required to assign their roles by themselves.

The experimental procedure consisted of one trial session and three test sessions. In the trial session, the participants (21 pairs) attempted to develop a communication system within one hour time limit. When characters moved to the same room, players received two points, otherwise they lost one point, but the scores did not drop below zero. The trial session was terminated when the score reached 50 points.

Test sessions were conducted subsequently. The  $T_{NM}$  (Test with No Message exchanges) did not allow message exchanges. In the  $T_{SM}$  (Test with Simultaneous Message exchanges), messages were displayed on the screen of each player after both players had sent their messages. Thus, taking turns in sending messages was prevented in this test session. The  $T_{IM}$  (Test with Immediate Message exchanges) was conducted under the same conditions as in the trial session. Each test had 12 rounds that contained all possible room combinations for two characters. The order of appearances was set at random.

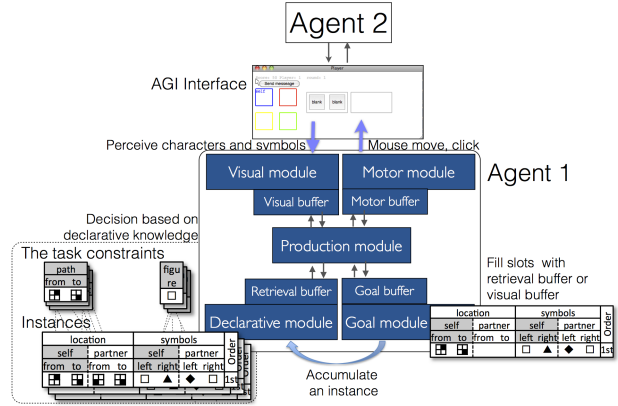


Figure 2: Schema of the model.

The results of the experiment confirmed the success of forming a shared symbol communication system, indicating the differences of the performance between the three tests. The detailed results are shown in a latter section.

## Model

### Architecture

The task presented in the previous section requires symbolic learning for constructing a new symbol system. In addition, according to Galantucci (2005)’s report, implicit learning, which is not present in symbolic systems, possibly plays a role in this task. In this study, we construct a model using ACT-R (Adaptive Control of Thought-Rational: Anderson, 2007), which integrates symbolic and subsymbolic learning mechanisms.

ACT-R is composed of several independent modules. The modules used in this study are presented in Figure 2. Except for the production module, each module has a buffer to temporarily store information called a chunk (a set of slot-value pairs). The production module integrates the other modules using production rules, which consist of a condition-action pair that is used in sequence with other productions to perform a task. The conditions and actions in production rules are specified along with buffer contents of each module.

In our model, two independent agents interact through a simulated task environment developed in the ACT-R graphical user interface (AGI). AGI provides screens that hold visual information as chunks. In this study, the locations of the characters and messages associated with each agent are displayed on the screen. An agent’s visual module searches for a character and stores its location (room) into a visual buffer. The visual buffer also stores the symbols that compose a message, attending to the screen locations where the figures appear. The simulated task environment also provides a virtual mouse to change the figures and move the characters on the screen.

Visual information stored in the visual buffer are transferred to the goal buffer through the production module. The goal buffer holds the goal of the current task and other task-related information. Specifically, our model has nine slots for



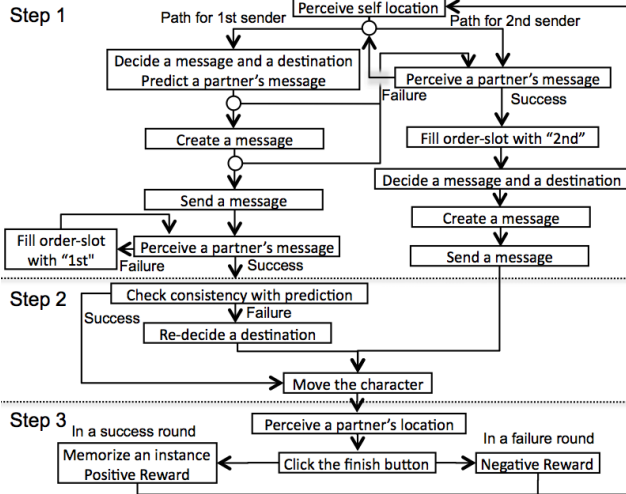


Figure 3: Process of the model. Circles indicate decisions based on conflict resolution.

the goal buffer: four slots for storing room locations (initial (from)-destination (to)  $\times$  self-partner), four slots for storing the symbols (left-right  $\times$  self-partner), and a slot for encoding the order for exchanging messages.

The declarative module stores past states of the goal buffer as instances. It also stores chunks representing task constraints such as path information indicating a room the characters can move to (e.g., *from* *to* *isa\_path*) or figures the agent can use to construct a message (e.g., *isa\_figure*). An agent uses these chunks (i.e., declarative knowledge) to choose its destination and construct a message.

## Process of the model

**Overview** We prepared 169 productions that construct the process presented in Figure 3. This process is divided into three steps just as in the original experiment (Figure 1).

In addition, the operation of taking turns to send a message is autonomously managed by this process. There are two paths in this process. The left path is for the first sender and the right path is for the second sender. The choice of path is made by conflict resolution, which is a comparison of two conflicting productions with noise added utilities. In each phase of the path of the first sender, there is a conflict (indicated by circles) between keeping the path of the first sender and changing to the path of the second sender. If in any of these the agent selects the path of the second sender, the agent tries to perceive the message of her/his partner from the screen. When the agent obtains the message of her/his partner, s/he realizes that s/he is the second sender (fills the order slot with "2nd"). Otherwise, s/he resolves a conflict by waiting for the message of her/his partner and changing to the path of the first sender. This conflict loop continues until one of the agents sends a message.

**Decision making** In step 1, regardless of the contents of the order slot, both agents make decisions about their destinations

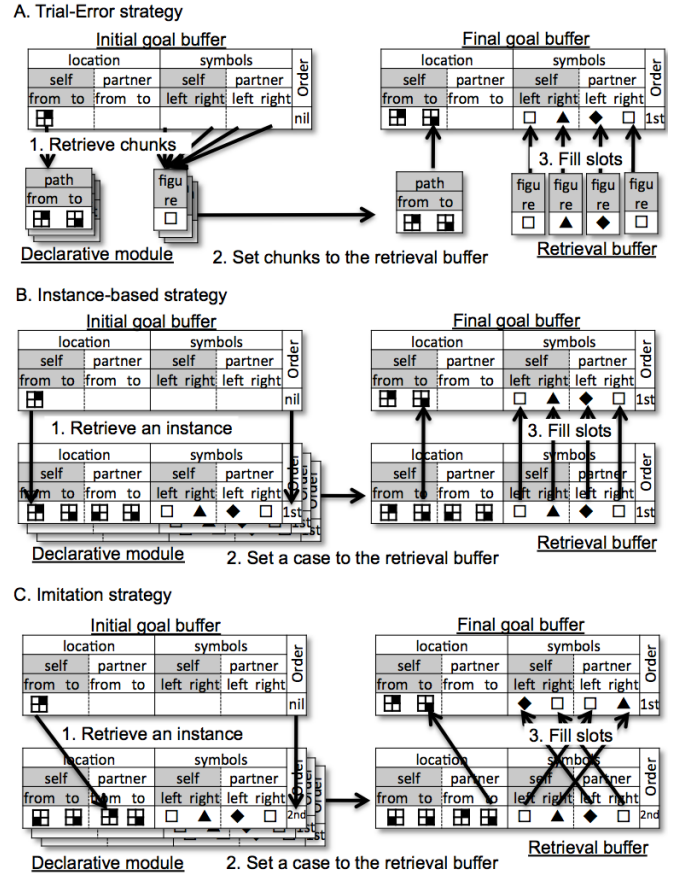


Figure 4: Three types of decision strategies.

and their messages. Concurrently, the first sender predicts the message that s/he will receive from her/his partner. The predicted message is checked against the message received in step 2. When the received message is inconsistent with the predicted message, the agent makes a new decision about her/his destination.

Summarizing, there are three situations where agents make decisions. In these situations, agents apply one of the three decision strategies shown in Figure 4<sup>1</sup>. Every decision strategy begins by retrieving chunks from the declarative module by using the current goal buffer as a cue. In the trial-error strategy, chunks concerning task constraints (chunks representing a path and symbols) are retrieved, and are used to fill in the blank goal slots. In the instance-based strategy, the agent retrieves an instance that is consistent with the current goal buffer. The retrieved instance is used to fill slots concerning a destination and symbols. The imitation strategy also uses an instance, but the roles of an agent and her/his partner are reversed when retrieving and filling slots.

The implementation of the trial-error and instance-based strategies follow a past study on decision making using ACT-

<sup>1</sup>Figure 4 explains each strategy by using an example of the first sender in step 1. In the case of the second sender in step 1, a partner's message is added to the retrieval cue. The first sender in step 2 uses a message as a cue to retrieve an instance and make a new decision about the destination.

R (Lebiere, Gonzalez, & Martin, 2007; Reitter & Lebiere, 2011). The imitation strategy is constructed according to the role-reversal imitation described by Tomasello (1999).

**Learning** The decision making strategy implemented in this model changes through a learning process. This model uses the standard symbolic and subsymbolic learning mechanisms of ACT-R. Symbolic learning includes instance-based learning and production compilation. Subsymbolic learning includes utility learning and activation updating.

Instance-based learning and utility learning occur at the end of each round (see Figure 3). In a success round (two characters meet in the same room), the agent stores a state of the goal buffer into the declarative module. Conversely, in a failure round (two characters move to different rooms), the agents do not store an instance. Regardless of the results of the movement, the utilities, which are used in conflict resolution, are updated by the following formula.

$$U_i(n) = U_i(n-1) + \alpha[R_i(n) - U_i(n-1)] \quad (1)$$

where  $\alpha$  is the learning rate and  $R_i(n)$  is the reward value given to production  $i$  at time  $n$ . In a success round, productions used in the round receive positive rewards ( $R_i(n) = 10$ ). Otherwise, productions used in the round receive negative rewards ( $R_i(n) = 0$ ).

As instances are accumulated, the chance to retrieve an instance increases, but utility learning does not directly affect decision making. In each decision strategy, there are no conflicting productions. Instead, each decision strategy needs to select declarative chunks because a single state of the goal buffer usually matches several chunks. The selection of chunks is controlled by the activation values of the chunks<sup>2</sup>. In ACT-R, an activation value is updated by the following formula.

$$B_i = \ln\left(\sum_{j=1}^n t_j^{-d}\right) + \beta_i \quad (2)$$

where  $n$  is the number of presentations of chunk  $i$ ,  $t_j$  is the time since the  $j$ th presentation, and  $d$  is the decay parameter<sup>3</sup>. This value is determined by the frequency and recency of a particular chunk. Therefore, an agent usually retrieves a chunk that has been frequently observed or retrieved in the past round.

Even though utility learning does not directly affect decision strategies, there is a possible effect that occurs through production compilation. Production compilation is a mechanism that creates a new production integrating sequentially firing two productions. It typically occurs when the first production requests a retrieval and the second harvests it. The resulting production is specialized to include the retrieved information. In other words, the declarative knowledge is proceduralized into the production. Because compiled productions receive rewards, it is possible to change behavior through production compilation and utility learning.

<sup>2</sup>The simulation uses a summation of this base-level activation and the spreading activation values

<sup>3</sup>In this study, we use  $d = 0.50$  and  $\beta = 0.00$

Table 1: The performance in the trial session. The numbers in parentheses indicate standard deviation.

	Data	Trial-error	Instance	Imitation
Success rates	0.66	0.00	1.00	1.00
Round	48.42 (13.36)	NA (NA)	72.08 (16.95)	54.50 (15.93)

## Simulation

### Simulation conditions

To explore the cognitive process behind forming a communication system, we set up the following three models controlling the decision strategies presented in Figure 4.

- **Trial-Error model:** This model does not have a decision strategy other than the trial-error strategy. The agent tries to construct a communication system from subsymbolic learning and production compilation.
- **Instance model:** This model, in addition to the trial-error strategy, also has the instance-based strategy. The agent first tries the instance-based strategy. If the instance-based strategy fails, the agent chooses her/his destination and message based on the trial-error strategy.
- **Imitation model:** This model extends the instance model by adding the imitation strategy. The agent first tries to choose her/his destination and message using the instance-based strategy. If the agent fails to retrieve an instance, the imitation strategy is applied. When all other decision strategies fail, the agent uses the trial-error strategy.

By comparing the imitation model with the other two models, we can identify the role of imitation in forming a shared communication system. We also identify the required learning mechanism from the difference between the trial-error model and the other models. Furthermore, by comparing the experimental data, we reveal the features of human communication systems.

In this simulation, each model runs 100 times. In each run, the model continued the trial session for 3,600 sec<sup>4</sup> or until the scores reached 50 points. Following the trial session, the model was engaged in three test sessions similar to the experiment presented in the section 2.

## Results

**Performance of trial session** Table 1 shows the proportion of runs/pairs whose scores reached 50 points, which is a termination condition for the trial session. It also presents the numbers of rounds required to reach the termination condition. All runs with the trial-error models failed to form a communication system whereas all runs with the instance and imitation models succeeded in completing the session. Even though there were some pairs that did not reach the termination condition, the number of rounds required to complete the session in the experiment (data) was smaller than that in the instance and imitation models. Compared to the instance model, the imitation model finished the session in less number of rounds, and the difference in the number of rounds

<sup>4</sup>We used the simulation time estimated by ACT-R.

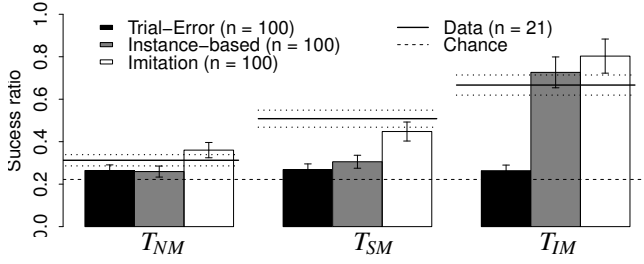


Figure 5: Performance of three test sessions. Solid lines in the graph indicate the results of the experiment. Error bars and dashed thin lines represent the standard error of means. The dashed thick line indicates chance-level performance (2/9).

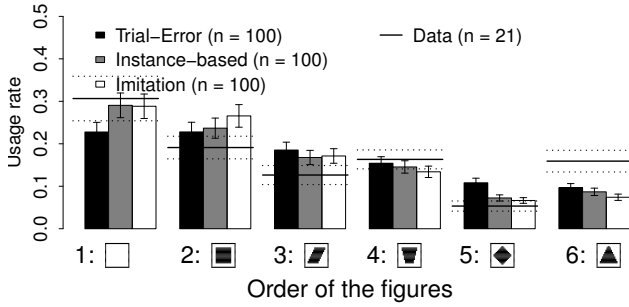


Figure 6: Rate of using each symbol. Solid lines in the graph indicate the data obtained in the experiment. Error bars and dashed thin lines represent the standard error of means.

between the experimental data and the imitation model is quite small (i.e., 48.42 and 54.50 respectively, compared with 72.08 in the instance model).

**Performance of tests sessions** Figure 5 presents the results of the three tests. The experimental data reveals significant differences between the three tests. The imitation and the instance models reproduced these differences well. From the difference between  $T_{NM}$  and  $T_{IM}$ , we confirmed that the pairs/models formed effective symbol communication systems. In addition, from the difference between  $T_{SM}$  and  $T_{IM}$ , we confirmed that the pairs/models take turns in transmitting messages. In contrast to the other models/data, we did not observe the differences of the three tests in the trial-error model.

**Messages constructed in the trial session** To examine the messages created in each model/experiment, we first computed the frequency of using each symbol in the trial session. From Figure 6, we can observe the decrease in the frequencies with respect to the order of the figures (see the caption of Figure 1). Using the symbols listed in the earlier order is considered as an adaptive behavior that reduces the time to construct a message. The three models reproduced this behavior. We postulate that this behavior is derived from the activation updating mechanism. As shown in formula 2, activation values are influenced by the frequencies of observing/retrieving chunks. Hence, symbols listed in the earlier order usually receive higher activation values.

We examined this distribution for each participant/agent

Table 2: Indices of sharing when using symbols.

	Data	Trial-error	Instance	Imitation
Single	0.80 (0.25)	0.92 (0.04)	0.76 (0.14)	0.94 (0.07)
Combination	0.63 (0.31)	0.52 (0.19)	0.28 (0.19)	0.84 (0.11)

Table 3: Bias of using symbols.

	Data	Trial-error	Instance	Imitation
Single	0.48 (0.22)	0.07 (0.02)	0.31 (0.12)	0.33 (0.17)
Combination	1.81 (0.48)	0.43 (0.07)	1.43 (0.26)	1.57 (0.30)

and computed symbol sharing indices within pairs. The indices were computed as the dot-product of two unit-vectors that consisted of the frequencies of using each symbol. These indices were computed not only for single symbols (e.g.,  $\blacksquare$ ,  $\blacklozenge$ ) but also for combinations of symbols (e.g.,  $\blacksquare\blacklozenge$ ).

Table 2 shows the computed indices of symbol sharing. The imitation model has the highest indices for both the single symbol and combination of symbols. Conversely, the indices in the instance model are the lowest. This suggests that an agent in the imitation model created an isomorphic symbol system with a partner. The values obtained in the experiment are between those of the instance and imitation model, implying that imitation certainly occurred in the experiment.

The other explanation of the high indices of sharing symbols is that agents/participants chose symbols according to a uniform-distribution. To exclude this possibility, we computed a bias index using the geometric mean of two Kullback-Leibler divergences of the probability distributions  $P$  and  $E$ .

$$B = \sqrt{D_{KL}(P_1||E) \cdot D_{KL}(P_2||E)},$$

$$D_{KL}(P_i||E) = \sum_{n=1}^N P_i(n) \log \frac{P_i(n)}{E}, \quad (3)$$

where  $P_i$  is the probability distribution of the use of symbols of agent/participant  $i$ ,  $E$  is the uniform distribution, and  $N$  is the number of bins of the probability distributions<sup>5</sup>. If the distribution  $P$  deviated from the uniform distribution, the index would increase.

The results are presented in Table 3. The trial-error model, which had the high symbol sharing indices, has the lowest values for both the single symbol and combination of symbols, indicating that the choice made in the trial-error model was less biased. Contrary to the trial-error model, the other two models and the experiment were biased towards using specific symbols, indicating that they communicated with a small number of symbols.

## Discussion and Conclusions

This study constructed a model that forms a new communication system through interactive coordination. To date, many

<sup>5</sup>For the distribution of using a single symbol,  $E = 1/6$  and  $N = 6$ . For the distribution of using a combination of symbols,  $E = 1/36$  and  $N = 36$

models for language evolution have been developed (for a review Steels, 2011). In addition, there exists a research work that uses ACT-R to simulate experiments of forming a communication system (Reitter & Lebiere, 2011). However, these researchers did not deal with a situation with spontaneous turn-taking or role-setting operations. Most of the previous models assign roles to agents, such as director or matcher, using simulation parameters.

Unlike the previous studies, we dealt with a situation where roles are autonomously assigned. In such a situation, agents efficiently develop a shared communication system through a two-way imitation, which uses a single instance in two ways by reversing roles. It reduces trial-errors, and results in a faster forming process as presented in Table 1. In addition, differently from the instance model, the imitation model create an isomorphic symbol system as indicated by Table 2. Summarizing these results, we conclude that with few interactions, imitation can create an isomorphic symbol system.

Moreover, the comparison of the models with the experiment revealed the existence of imitation in the formation of a human communication system. Table 2 showed the apparent difference in the sharing index between the instance model and the experimental data. This result is consistent with previous laboratory experiments, in which participants used similar symbols in pairs through interactions (Fay, Garrod, Roberts, & Swoboda, 2010; Garrod, Fay, Lee, & Oberlander, 2007).

An additional finding was observed from the trial-error model. Unlike the imitation and the instance models, the trial-error model failed to construct a communication system. All models possess a common subsymbolic learning that updates activations of chunks and utilities of compiled productions. Therefore, this difference indicates an advantage of symbolic learning. Our results suggest that subsymbolic learning by itself have a limitation in constructing a communication system. Rather, subsymbolic learning are used to adapt to the structure of the environment as shown in Figure 6. We consider that this adaptivity provides a scaffold for forming a new symbolic communication system (Konno et al., in press).

There are several limitations in this study. We could examine neither the syntax (combination rules of symbols) nor the symbol-meanings mappings. These important features of human communication systems, namely language, are difficult to be captured by simple statistics. We need analysis methods to understand how a new language emerges in an experiment or computer simulation.

To extend our model to more complex situations, the goal buffer of the model should be hierarchically decomposed. The current design of the goal buffer cannot be applied to a complex situation where many symbols are combined. Imitation strategy also needs to include more detailed processes. We believe that the theory of analogical reasoning (Gentner, 1983) can be applied to a model of a complex imitation strategy. There is some previous research that point out that imitation is a type of analogical mapping (Barnes & Thagard,

1997; Thagard, 2001). It is a big challenge for the broad cognitive science community to examine how analogical mapping is changed through symbolic and subsymbolic learning when forming a new language.

## Acknowledgment

This work was supported by a Grant-in-Aid for Scientific Research on Innovative Areas “The study on the neural dynamics for understanding communication in terms of complex hetero systems (No.4103)” (21120011) of The Ministry of Education, Culture, Sports, Science, and Technology, Japan.

## References

- Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.
- Barnes, A., & Thagard, P. (1997). Empathy and analogy. *Dialogue: Canadian Philosophical Review*, 705–720.
- Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, 34(3), 351–386.
- Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, 29(5), 737–767.
- Garrod, S., Fay, N., Lee, J., & Oberlander, J. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, 31, 961–987.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155–170.
- Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, 415, 755.
- Konno, T., Morita, J., & Hashimoto, T. (in press). Symbol communication systems integrate implicit information in coordination tasks. In *Advances in cognitive neurodynamics (iii)*. Springer.
- Lebiere, C., Gonzalez, C., & Martin, M. (2007). Instance-based decision making model of repeated binary choice. *Proceedings of the 8th International Conference on Cognitive Modeling*.
- Reitter, D., & Lebiere, C. (2011). How groups develop a specialized domain vocabulary: A cognitive multi-agent model. *Cognitive Systems Research*, 12, 175–185.
- Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, 14(9), 411–417.
- Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews*, 8(4), 339–356.
- Thagard, P. (2001). Emotional analogies and analogical inference. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), (pp. 335–362). *The analogical mind Perspectives from cognitive science*.
- Tomasello, M. (1999). *The cultural origins of human cognition*. Harvard University Press.

# Force Dynamics as a Basis for Moral Intuitions

Jonas Nagel (jnagel1@uni-goettingen.de)

Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)

Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany

## Abstract

People seamlessly generate moral intuitions about a wide range of events they observe, but to date the cognitive processes underlying this competency are poorly understood. We propose that our moral intuitions are grounded in force-dynamic intuitions. We show how the evaluation of entities engaged in schematized interactions can be predicted from specific force-dynamic properties of those interactions, and we point out how these evaluative tendencies relate to our moral norm of not interfering with others' interests.

**Keywords:** moral judgment; intuition; force dynamics

## A New Theory of Moral Intuitions

Recent moral psychology views intuitions as important determinants of our moral judgments. Haidt (2001) defined moral intuitions as “the sudden appearance in consciousness of a moral judgment, including an affective valence (good-bad, like-dislike), without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion” (p. 818). Intuitions are thus mainly defined in contrast to deliberative reasoning.

However, to date there is no worked-out theory of the automatic processes by which our moral intuitions are formed. How do we solve this computational task? Which observed events elicit which specific moral intuitions? In what format are these events represented, and how is this representation automatically integrated with pre-existing evaluative standards? We propose that the semantic category of *force dynamics* (Talmy, 1988) provides a cognitive structure that might serve as a representational basis in this task.

## The Semantic Category of Force Dynamics

Talmy (1988) described force dynamics as a semantic category of how entities interact with respect to force. When two entities interact in the world, our language assigns to them the two thematic roles of *patient* (*P*) and *agent* (*A*). (Talmy uses the terms *agonist* and *antagonist*.) *P*, the focal entity, is perceived to have an intrinsic force tendency either towards rest or towards motion. In force-dynamic interactions, *P* finds itself in opposition to *A*, another entity with the opposed force tendency. *A* is mainly thought of in terms of the effects it has on the resultant force manifested by *P* as outcome of the interaction.

*P*'s resultant force depends on the relative strength of the two opposing forces. If *P*'s force is stronger than *A*'s, then *P*'s resultant force equals its intrinsic tendency. This constellation is expressed in familiar words in our natural language, such as *despite* or *although*. In “The flame [*P*] kept burning despite the wind [*A*] blowing at it,” *P*

manifests its intrinsic tendency (to burn) in spite of the opposing force exerted by *A*.

If *A*'s force is stronger than *P*'s, then *P*'s resultant force is opposed to its intrinsic tendency. There are many words in natural language describing variants of this basic constellation (e.g., *make*, *cause*, or *prevent*). In the sentence “The wind [*A*] made the flame [*P*] go out,” *P* does not manifest its intrinsic tendency (to burn) but the opposite (to go out). Note that both example sentences “are about” *P*, while *A* is mainly relevant in terms of the effect it has on *P*.

Talmy (1988) argues that force dynamics is a fundamental semantic category, profoundly structuring our cognitions in a variety of domains. Whether we deal with the physical, the (intra-)psychic, or the social world, the same basic force-dynamic concepts pervade our language and thought. Thus, force dynamics is conceived as a domain-independent representation underlying intuitions in various domains, not only in the physical domain. Actually it is interesting to see that when force dynamics is applied to physical tasks, the resulting intuitions about forces and intrinsic tendencies seem to be more compatible with our understanding of actions than with Newtonian physics. For example, the intuition that causes have stronger forces than effects is inconsistent with Newtonian physics but seems to be grounded in sensory-motor representations of our actions (White, 2009). Analyzing social interactions in terms of force dynamics is thus not a reduction of the social to the physical domain. Force dynamics is better viewed as a domain-independent *abstract* conceptual framework.

## Force Dynamics as a Basis for Moral Intuitions

The category of force dynamics combines causal and teleological aspects and is abstract enough to be naturally applicable across physical and social domains. These properties make it a promising candidate to serve in the process of enriching representations of observed events with a basic evaluative aspect.

Imagine observing the following event: Jack shoves Jones. In a first step, the observer could abstract the force-dynamic pattern instantiated by this event. This would include assigning *A*- and *P*-roles to the entities involved in the interaction, determining *P*'s intrinsic tendency and resultant force, and comparing the latter two. In this example, this would yield a representation of *A* (Jack) forcing *P* (Jones) to deviate from his intrinsic tendency (toward rest) into a different resultant force (motion). This is an instance of *onset causing of motion* (Talmy, 1988).

In a second step, this abstract representation could be automatically subjected to default normative principles formulated on the same level of abstraction. We stipulate

the existence of a *noninterference principle (NIP)*: By default, *patients should be allowed to manifest their intrinsic tendencies*. This substantive assumption (which has itself a force-dynamic structure) can be motivated with reference to the negative prima facie duty not to interfere with others' interests, which seems to be a fundamental moral norm at least in Western cultures. In *onset causing of motion*, this abstract principle is violated. *P* changes its tendency because of *A*'s impingement.

Finally, the fact that *A* is identified as causing a violation of the principle leads people to evaluate *A* negatively relative to *P*. This negative evaluation is then applied to the concrete observed instance of *A* (in this case, to Jack).

In the current research we focus on basic scenarios with only two protagonists (*A*, *P*). Obviously, there are many more complex instances of *onset causing* patterns in which *A* might eventually be evaluated positively. Imagine you receive the additional information that Jack shoved Jones *out of harm's way*. Such more complex constellations will not be treated here, but it seems that our theory can in principle be extended to handle them as well (e.g., Jack could be evaluated positively for *preventing* another agent [the harm] to violate the NIP by means of intervening on *P*).

In what follows, we will provide initial evidence that such default evaluations are in fact assigned on the abstract level of decontextualized force-dynamic representations. To this end, we had experimental subjects evaluate the movements of two abstract shapes engaged in simple interactions.

### Force Dynamics in Abstract Animated Displays

Displays of moving objects allow for a non-verbal presentation of decontextualized force-dynamic interactions. In the absence of linguistic cues, we first need to explicate the criteria according to which we assume our subjects to abstract force-dynamic patterns from our visual displays (i.e., the first step in the process outlined above).

We instantiated the *onset causing of motion* pattern with a version of the well-known *launching event* (Michotte, 1963; see Fig. 1). A stationary Object Y is situated in the center of the stage. After a moment, another Object X enters the scene from the side and approaches Object Y on a straight line and at constant speed. At the moment of contact, Object X stops and Object Y immediately starts moving as if it was continuing on X's trajectory.

According to our theory, subjects first need to assign agent and patient roles to these interacting entities. We argue that the extremely impoverished nature of this display leaves but three cues to make this assignment: (a) pre-collision movement relative to the position of the other entity; (b) causing change of state in the other entity; and (c) appearing on the scene after the other entity. According to Dowty (1991), the first two cues increase the likelihood that a given entity is assigned the agent role in an interaction. Concerning the third cue, the entity appearing first will likely be seen as the focal entity the display "is about" (i.e., the patient); the second entity should therefore be regarded as agent affecting this focal entity. We assign the agent role

to an entity if it embodies more of these three cues than the alternative entity. In the launching event, Object X's behavior is consistent with all three cues (a+, b+, c+), while Object Y only causes Object X to stop (b+), but does not move initially (a-) and is first on the screen (c-). Object X is therefore assigned the agent role, while Object Y is the patient. There is ample evidence that this analysis is in line with people's qualitative experience of launching events. For example, X is seen as *exerting force on* Y, whereas Y is perceived to merely *exhibit resistance against* X (White, 2009).

Next, subjects need to infer *P*'s intrinsic tendency and resultant force. We assume that in the absence of further contextual cues indicating the presence of external forces, *P*'s pre-collision movement will be regarded as its intrinsic tendency. The identification of *P*'s resultant force with its post-collision movement seems unproblematic.

Finally, *P*'s intrinsic tendency and resultant force need to be compared in order to determine the force-dynamic pattern and to decide whether the NIP was violated, as would be indicated by a change of *P*'s movement as a result of the collision event.

### Hypothesis

With all force-dynamic concepts operationalized, we now turn to the specifics of the hypothesis we tested in the present experiment. We predict that in the *Launch* case described above, which instantiates *onset causing of motion*, subjects asked to evaluate the movements of both entities on a negative/positive dimension will evaluate Object X more negatively than Object Y. We contrast this case with a *Blocked* case which is identical to *Launch* except that Object Y does not start moving on X's trajectory after the collision, so that the interaction ends with both entities at rest in the center of the screen. In this case, X has two agentic cues (a+, b-, c+) while Y has only one (a-, b+, c-). Object Y is thus still the patient, and its intrinsic tendency (rest) is identical to its resultant force (rest). This corresponds to a *despite* pattern in Talmy's (1988) terminology. The NIP is not violated here, so we do *not* expect *A* (Object X) to be rated negatively relative to *P* (Object Y) in this case. Across the cases, we expect *A* to be rated more negatively in *Launch* than in *Blocked*.

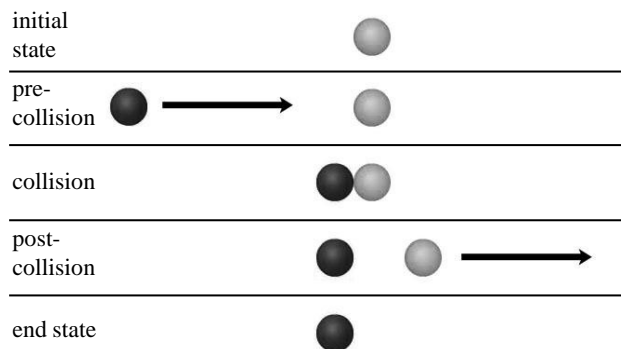


Figure 1: Animation of the launching event. Dark sphere = *A*; light sphere = *P*. See text for details.

## Experiment

We divide the presentation of the experiment into two parts. In the first part, we test the hypothesis just explicated. The second part will deal with an additional aspect. However, the data for both parts were gathered within one and the same experiment and from the same subjects. Therefore, we begin by describing the general procedure for the whole experiment before we discuss the specific materials and the results for both parts separately.

### Participants and Procedure

31 undergraduates of the University of Göttingen (23 female, mean age 22 years) participated in a computer-based experiment containing 27 trials in random order.<sup>1</sup> Each trial consisted of three consecutive screens. The first screen displayed the trial number and a button with which the subjects could start the animation. The second screen contained a looped display of one of 27 animations. Each animation started with a fixation cross displayed for one second in the center of the stage. Then the initial state was presented for one second, before the force dynamic interaction began to unfold. The interaction always consisted of a blue sphere and a green sphere moving in specific ways on the same straight horizontal trajectory. After the interaction was finished, the stationary end state remained on the screen for one second. As an example, Figure 1 illustrates how we implemented the *Launch* case described above. Each condition was instantiated in two animations, counterbalanced for color assignment and direction of movement. Subjects watched each animation as often as they wished before they proceeded to the third screen, where two identical 7-point rating scales ranging from “-3 (negative)” to “+3 (positive)” were presented above one another. Each referred to one of the two entities from the animation. The question wording was “How do you evaluate the movement of the blue/green figure?”

### Part 1: Interference with Intrinsic Tendency

**Design and Material** In this part we tested whether people’s intuitive evaluations of the entities in *Launch* and *Blocked* can be predicted from the underlying force dynamics in combination with the noninterference principle (NIP). In the initial state of all animations, *P* was displayed at rest in the center of the stage. The force-dynamic interaction always began with *A* entering from one side and reaching *P* within one second on a straight horizontal trajectory at constant speed. We then manipulated *A*’s and *P*’s post-collision movement which could either be stationary (0), movement continuing *A*’s initial trajectory at half of *A*’s initial speed (1), or movement on the same trajectory at *A*’s full initial speed (2). Both *A* and *P* could display all three movements, with the constraint that *A* could not be faster than *P* after collision because this would imply the objects going through each other. Thus, the

manipulation yielded six conditions which are displayed in Table 1. Conditions 1 and 3 are of main theoretical interest because they manifest the *Blocked* and *Launch* cases to which our main hypothesis refers.

Table 1 also lists several resulting properties of the interactions displayed in each condition. *A\_change* and *P\_change* indicate whether *A* and *P* change their overt tendency in the course of the interaction. *P\_change* is the most important variable for our purpose. Given the pre-collision constellation (i.e., *A* in motion, *P* at rest), as soon as *P* changes its tendency, the case is an instance of *onset causing of motion*, and the NIP is violated. Only if *P* stays stationary, the case becomes an instance of *despite* where the principle is not violated. The remaining properties are further implications of the entities’ post-collision movements. Concordance indicates whether *A* and *P* have a concordant tendency (either of rest or of movement in the same direction) after the collision. Contact indicates whether *A* and *P* remain in direct physical contact after the collision. Finally, Resistance indicates whether *P* displays resistance by not overtaking the pre-collision tendency of *A* in a one-to-one manner. As can be seen in Table 1, this property is not identical to *P\_change*.

Concordance, Contact, and Resistance are listed because they are perfectly confounded with *P\_change* across the two cases of main interest, 1 and 3. Any change in ratings between 1 and 3 could thus just as well be caused by these properties. The remaining four conditions serve to isolate *P\_change* from these confounds.

Table 1: Design of Part 1

Cond	Speed		Resulting properties				
	<i>A</i>	<i>P</i>	<i>A_ch</i>	<i>P_ch</i>	Conc	Cont	Res
1	0	0	1	0	1	1	1
2	0	1	1	1	0	0	1
3	0	2	1	1	0	0	0
4	1	1	1	1	1	1	1
5	1	2	1	1	1	0	0
6	2	2	0	1	1	1	0

*Note.* Cond = condition, Speed = post-collision speed, *A/P\_ch* = *A/P\_change*, Conc = Concordance, Cont = Contact, Res = Resistance. See text for further explanations.

**Specific predictions** Three specific predictions follow directly from our hypothesis. (i) *A* will be rated more negatively than *P* in condition 3 (*Launch*). (ii) *A* will *not* be rated more negatively than *P* in condition 1 (*Blocked*). (iii) *A* will be rated more negatively in 3 than in 1.

The remaining conditions serve to separate the manipulation of the force-dynamic pattern from the properties of Concordance, Contact, and Resistance. If a concordant post-collision tendency is responsible for more positive ratings for *A* in 1 compared to 3, *A*-ratings should also be more positive in 4, 5, and 6. If sustained physical contact is to be made responsible, *A*-ratings should be more positive in 4 and 6. Finally, if the display of resistance by *P* (in not adopting *A*’s pre-collision tendency) is to be made

<sup>1</sup> Three of these 27 trials tested a third hypothesis which is not reported here due to space constraints.



responsible, A-ratings should be more positive in 2 and 4. We predict that none of these alternatives will be the case. Instead, we expect (iv) all control cases to be treated like 3 since they all conform to the *onset causing of motion* pattern. This result would support our hypothesis that the evaluative ratings in 1 (*Blocked*) and 3 (*Launch*) are in fact a function of the underlying force-dynamic pattern as indicated by the variable  $P\_change$ .

**Results and Discussion** The descriptive results are displayed in Figure 2. A global 6 (Condition: 1 to 6)  $\times$  2 (Entity: A vs. P) repeated-measures ANOVA yielded a main effect for Entity ( $F_{1,30} = 9.57$ ;  $p < .01$ ,  $\eta_p^2 = .24$ ), indicating that, across conditions, A was rated more negatively than P. More importantly, the Condition  $\times$  Entity interaction term was significant ( $F_{5,150} = 9.69$ ;  $p < .001$ ,  $\eta_p^2 = .24$ ), showing that A and P were treated differently across conditions. We now turn to the contrasts testing our specific predictions.

(i): In 3 (*Launch*), A-ratings were lower than the P-ratings ( $t_{30} = -3.68$ ,  $p < .001$ ,  $d = -.66$ ). A is thus rated more negatively than P in the launching event as paradigmatic instance of the *onset causing of motion* pattern violating the NIP.

(ii): In 1 (*Blocked*), A-ratings were *higher* than P-ratings ( $t_{30} = 3.63$ ,  $p < .01$ ,  $d = .65$ ). Thus, it seems that A's negative evaluation disappears with an underlying *despite* pattern in which the NIP is not violated. The significant drop in P-ratings was not expected because our predictions only concerned A-ratings. One post-hoc explanation for this phenomenon might be that the agent-patient distinction is not as clear cut in this case as in the *Launch* case (i.e., the agentic cues are distributed more evenly across both entities). Maybe some participants interpreted P as agent due to its capacity to cause change in A, turning the interaction into an *onset causing of rest* pattern (Talmy, 1988) in which P (now the agent) violates the NIP by forcing A (now the patient) to deviate from its intrinsic tendency to motion into a resultant state of rest.

(iii): A-ratings in 1 were higher than those in 3 ( $t_{30} = 3.27$ ,  $p < .01$ ,  $d = .59$ ). Our hypothesis is thus confirmed by comparisons between entities within cases and across cases with different underlying force-dynamic patterns.

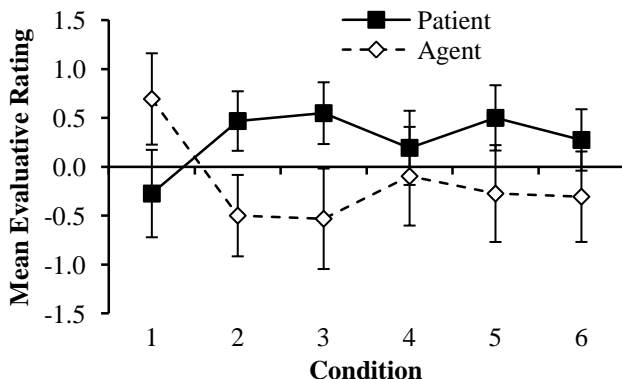


Figure 2: Results of Part 1. Error Bars = 95% CI.

(iv): The significant increase in A-ratings only occurred in 1, while, as predicted, the four control conditions generally behaved like 3 (*Launch*). The only exception is condition 4, where the A-ratings were not significantly lower than the P-ratings (although the descriptive trend still holds,  $t_{30} = -1.31$ ). Furthermore, across all six conditions, there seems to be a trend for concordant post-collision movement (i.e., 1, 4, 5, and 6) to yield slightly higher A-ratings than discordant post-collision movement ( $t_{30} = 2.26$ ,  $p < .05$ ,  $d = .41$ ). This might indicate that Concordance is used as an additional cue for the evaluation of A. However, note that this effect is driven mainly by the selective increase of A-ratings in 1 (*Blocked*).

In sum, these findings demonstrate that in clear-cut cases of *onset causing of motion* such as 3 (*Launch*), agents are evaluated more negatively than patients. This is not the case in *despite* cases in which the NIP is not violated. Together, these findings show that people's evaluations of movements are sensitive to underlying force-dynamic patterns. Entities are by default evaluated negatively if they cause other entities to deviate from their intrinsic tendency. This result is consistent with the moral norm in our society to not force others into states in which they would not enter on their own.

## Part 2: Prior Concordance

In this part we will test whether the default evaluations we have discovered are robust enough to be consistently applied to cases in which P is initially not at rest but rather in motion. Imagine a moving P colliding with a faster-moving A, changing the speed and/or direction of its movement after the collision. According to our criteria, P's intrinsic tendency would be to move in exactly the manner manifested prior to collision (including direction and speed parameters). A should thus be identified as causing P to deviate from its intrinsic tendency which constitutes a violation of the noninterference principle (NIP).

Crucially, this should be the case regardless of whether P and A exhibit concordant or discordant pre-collision movement. The direction of forces is not represented in Talmy's (1988) framework so that two entities moving on the same trajectory at different speed are still conceptualized as being in opposition (contrary to Wolff, 2007; see General Discussion). Note that without reference to Talmy's framework of opposing forces it seems a priori plausible that the concordant and discordant cases will be conceptualized differently. Specifically, in the concordant case it may seem as if the faster A *enhances* the slower P in its tendency which could result in A being evaluated positively for "helping" P. According to our theory, however, this should not be the case. If people's default evaluations correspond to Talmy's framework and the NIP, then they should be insensitive to prior concordance: They should evaluate an A making P go faster into the direction of its initial movement similarly to an A making P go into the opposite direction of its initial movement. Both cases violate the NIP.

**Design and Material** The initial state of all animations was an empty stage. After one second, *P* entered the stage from one side at a constant speed on a straight horizontal trajectory, reaching the center of the stage after two seconds. One second after *P*'s appearance, *A* entered the stage at twice the speed of *P*, either from the same side (concordant condition, C+) or from the opposite side (discordant condition, C-). Consequently, in both conditions the collision of *A* and *P* took place in the center of the screen, one second after *A*'s appearance. After the collision, both entities moved in the direction of *A*'s initial movement in all conditions. This implies that in C+, *P* continued in the direction of its initial movement, while in C- it reversed the direction of movement. Similar as in Part 1, we manipulated the post-collision speed of both entities, which could be the initial speed of *P* (1) or *A* (2). Again, *P* had to be at least as fast as *A* after the collision. This yielded three Speed conditions crossed with the two Concordance conditions, resulting in the six experimental conditions summarized in Table 2. Concerning the definition of agentic cues, we refined the criterion of pre-collision movement (a, see above) to *faster* pre-collision movement relative to the other entity. As in Part 1, *A* exhibits at least one more cue for agency than *P* in all conditions.

Table 2: Design of Part 2

Cond	Conc	Speed		Resulting Properties		Cont	Res
		A	P	A_ch	P_ch		
C+1	1	1	1	1	0	1	1
C+2	1	1	2	1	1	0	0
C+3	1	2	2	0	1	1	0
C-1	0	1	1	1	1	1	1
C-2	0	1	2	1	1	0	0
C-3	0	2	2	0	1	1	0

Note. Cond = Condition, Conc = pre-collision Concordance, Speed = post-collision speed, A/P\_ch = A/P\_change, Cont = Contact, Res = Resistance.

**Specific Predictions** (i) Straightforward predictions arise from our model for the evaluation of *A* in all three C- cases. The reversal of *P*'s direction of movement caused by *A* is a clear violation of the NIP which should lead subjects to evaluate *A* negatively relative to *P*.

(ii) Case C+1 is an instantiation of *despite* in which *P* continues manifesting its intrinsic tendency after the collision. *A* should not be evaluated negatively relative to *P* since the NIP is not violated.

(iii) The crucial new conditions are C+2 and C+3. Here, subjects encounter a violation of the NIP preceded by concordant pre-collision movement. *P* is thus merely caused to deviate from its intrinsic (slow) speed, but not to deviate from its intrinsic direction. This could in principle lead subjects to conceptualize *A* as helping *P* to advance faster on its path. However, our theory predicts that subjects will still evaluate *A* negatively for causing a violation of the NIP. The *A*-ratings should also not differ from the *A*-ratings in the C- conditions.

**Results and Discussion** The descriptive results are displayed in Figure 3. A global 2 (Concordance: C+ vs. C-)  $\times$  3 (Speed: 1 to 3)  $\times$  2 (Entity: *A* vs. *P*) repeated-measures ANOVA yielded a main effect for Concordance ( $F_{1,30} = 19.10$ ;  $p < .001$ ,  $\eta_p^2 = .39$ ), indicating that both entities were generally rated more negatively in C- than in C+. Again there was a main effect for Entity ( $F_{1,30} = 28.90$ ;  $p < .001$ ,  $\eta_p^2 = .49$ ), indicating that *A* was generally rated more negatively than *P*. Finally, the Speed  $\times$  Entity and the Speed  $\times$  Concordance interaction terms were significant ( $F_{1,30} = 7.08$ ;  $p < .01$ ,  $\eta_p^2 = .19$ , and  $F_{1,30} = 4.07$ ;  $p < .05$ ,  $\eta_p^2 = .12$ , respectively), showing that post-collision movements of *A* and *P* affected *A*- and *P*-ratings differentially, and that they also had different effects depending on the prior concordance of both entities.

(i) *A*-ratings in the three C- conditions are lower than the respective *P*-ratings ( $t_{30} = -4.02$ ,  $p < .001$ ,  $d = -.72$ ), replicating the result of Part 1 that *A* receives negative evaluations when it clearly violates the NIP.

(ii) *A*-ratings in C+1 are not different from the *P*-ratings in the same condition ( $t_{30} = -.97$ ,  $p = .34$ ). As expected, *A* is again not evaluated negatively if it does not interfere with *P*'s intrinsic tendency.

(iii) *A*-ratings in C+2 and C+3 are lower than the respective *P*-ratings ( $t_{30} = -5.81$ ,  $p < .001$ ,  $d = -1.04$ ), while they do not differ significantly from the *A*-ratings in the C- conditions ( $t_{30} = 1.44$ ,  $p = .16$ ). Both results indicate that our subjects evaluated the concordant cases according to the same principles of opposing force-dynamics that they used in discordant cases, as predicted by our model. As soon as *A* violated the NIP, it was evaluated negatively, regardless of whether *P* was forced to reverse direction of movement or merely to continue faster on its original trajectory.

## General Discussion

In this paper, we have argued that the semantic category of force dynamics (Talmy, 1988) provides a cognitive structure underlying our moral intuitions. We provided evidence that various evaluations of content-free interacting entities can be predicted from force-dynamic properties in combination with a single normative principle (NIP) that

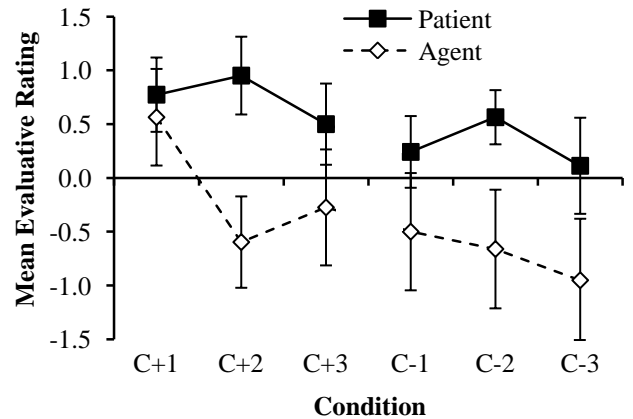


Figure 3: Results of Part 2. Error Bars = 95% CI.

expresses our *prima facie* moral norm not to interfere with others' interests. We thus propose that force dynamics might be part of the missing link between the apprehension of a situation and the automatic generation of a moral intuition. Observed events activate an abstract representation of the force-dynamic pattern which they instantiate. This representation is subjected to a basic normative principle, yielding default evaluative tendencies that are automatically applied to the participants of the observed interaction. We do not claim that the output of this process already represents a full-blown moral intuition. The automatically generated resulting representation could constitute a building block serving as input for higher-level processes (e.g., contextual analysis, inferences from other cues, background knowledge, application of exceptions, etc.) that eventually lead to rich, conscious moral intuitions.

Of course, the present study is only an encouraging first step in our research endeavor. So far we have only demonstrated an association between some force-dynamic patterns and explicit evaluations under maximally impoverished context conditions. It needs yet to be shown that the observed force-dynamic interactions also spontaneously elicit basic evaluations when no explicit test questions are given that request moral evaluations.

It may be seen as problematic that our model does not differentiate between animate and inanimate entities, contrary to many other theories in the field (e.g., Carey, 2009). Instead, our proposal is that force dynamic intuitions underlie event representation across domains as an abstract common code. If we are correct that basic evaluations are automatically elicited on this level of abstraction, this implies that observing one billiard ball launching another should elicit the same evaluative tendencies as observing Jack pushing Jones. Note, for example, that our displays contained no cues to animacy (such as self-propelled motion), and yet evaluations consistent with our model were observed. The postulation of a common code eliciting basic intuitions in both physical and social domains does not rule out that people use additional cues to differentiate between animate and inanimate entities (see Hamlin & Wynn, 2011, for evidence with infants). Force dynamics does not postulate that our representations of physics and psychology are *exhaustively* characterized as interplay of interacting forces. Additional semantic knowledge may of course enrich the force dynamic representation.

A related concern is that our model does not seem to capture all moral intuitions. A force-dynamic analysis of Jack lying to Jones, for example, will probably be less straightforward. We are aware that the practice of our moral judgment is very intricate and involves more considerations than those touched upon here. Our claim is thus not to provide a comprehensive theory of our moral intuitions. However, note that the range of intuitions our approach *does* potentially capture seems remarkable given its simplicity. The abstract nature of force dynamics makes it applicable to heterogeneous morally relevant events (e.g., dictators *oppressing* their people, people *resisting* temptations, etc.).

Another limitation of our approach is that it only predicts evaluations of agents. Yet, patient ratings also varied across our experimental conditions, sometimes independently from agent ratings. This might suggest that additional inferences are drawn from our stimuli which are not captured by our model.

Result (iii) of Part 2 suggests that the default conceptualization of force-related interactions is one of antagonism. It is likely that this default can quite easily be overridden if additional contextual cues are available that activate concepts of cooperation. Wolff (2007), for example, investigated cases in which *P* initially approaches a specific end state and *A* exerts a concordant force on *P*, resulting in *P* reaching the end state more quickly. Such displays reliably elicited concepts of *enable* or *help* in which *A* would presumably receive positive evaluations. We would predict that if a salient end state was provided in our displays, the default conceptualization of *P*'s intrinsic tendency might be replaced by a higher-level goal-directed conceptualization in which *P*'s intrinsic tendency would be *to reach the end state*. Once this more abstract intrinsic tendency would be attributed to *P*, the NIP would no longer be violated. Future studies will need to test these and related predictions for more complex structures with more than two protagonists.

## Acknowledgements

This research was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG WA 621/21-1), and by the Courant Research Centre 'Evolution of Social Behaviour', University of Göttingen (funded by the German Initiative of Excellence). We thank Laila Drummond Nauck and Julia Wiecha for helping with stimulus design and data collection.

## References

- Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.
- Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67, 547-619.
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814-834.
- Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, 26, 30-39.
- Michotte, A. E. (1963). *The perception of causality*. New York: Basic Books. (Original work published 1946)
- Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, 12, 49-100.
- White, P. A. (2009). Perception of forces exerted by objects in collision events. *Psychological Review*, 116, 580-601.
- Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, 136, 82-111.

# Preservation of the Initial Analysis in Absence of Pragmatic Inference with Japanese Relative Clause Sentences

Chie Nakamura (arumakan@nak.ics.keio.ac.jp)

JSPS Research Fellow / Graduate School of Science and Technology, Keio University

3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, Kanagawa, Japan

Manabu Arai (m-arai@phiz.c.u-tokyo.ac.jp)

JSPS Research Fellow / Department of Language and Information Sciences, The University of Tokyo

3-8-1 Komaba, Meguro-ku, Tokyo, Japan

## Abstract

Previous studies reported that language comprehenders tend to preserve the initial incorrect analysis with temporarily ambiguous sentences following structural reanalysis (Christianson et al., 2001; van Gompel et al., 2006). One possible criticism is that the sentences tested in previous studies allow comprehenders to pragmatically infer that the initial misanalysis may be true. It is thus unclear whether the tendency can still be observed where such inferences are not possible. The current study therefore tested the relative clause sentences in Japanese, which are temporarily ambiguous between the main clause and relative clause analysis. Crucially, our sentences differed from those in the past studies in that the correct interpretation following reanalysis makes an interpretation for the initial analysis pragmatically incompatible. The results demonstrated that an interpretation for the initial analysis persists even without pragmatic inference and that such incomplete syntactic representations are most likely due to excessive processing load.

**Keywords:** good-enough representation; garden-path sentence; self-paced reading; eye-tracking; reading

## Introduction

It is known that structurally ambiguous sentences such as *While the man hunted the deer ran into the woods* cause readers to experience difficulty at the point where the structure is disambiguated (i.e., *ran*), which indicates that they initially adopted an incorrect analysis (i.e., the transitive analysis). Past research showed that in processing this type of ambiguous sentences, language comprehenders tend to preserve an interpretation of the initial incorrect analysis even after the structure was fully reanalyzed and they often end up with incomplete, or so-called *good-enough* sentence representations (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Ferreira, Bailey & Ferraro, 2002). In Christianson et al. (2001), for instance, participants read sentences like (1) and were next shown comprehension questions like (2). Their results showed that participants often answered incorrectly by responding “yes”. Christianson et al. (2001) argued that the reanalysis of ambiguous sentence structures is not always complete and the initial incorrect analysis often remains preserved and not fully suppressed.

- (1) While Anna dressed the baby spit up on the bed.  
(2) Did Anna dress the baby?

However, in their study as well as their follow-up studies (e.g., Christianson, Williams, Zacks, & Ferreira, 2006; Ferreira et al., 2002 for the review), it is arguable that the initial incorrect analysis, even though it was suppressed following reanalysis, may have been re-activated by processing the question sentences. Another criticism for their argument of incomplete representations is an influence of pragmatic inference. For example, even with their items with reflexive verbs such as (1), it is still possible to infer that the baby spit up while being dressed by Anna even though such an interpretation is not syntactically licensed. The first issue was addressed by van Gompel, Pickering, Pearson, and Jacob (2006), who reported an effect of syntactic priming for the initial incorrect analysis. In their study, participants first read a prime sentence that was either structurally ambiguous (3a) or unambiguous with a comma following the subordinate clause verb (3b). Next, they completed a sentence fragment such as *When the doctor was visiti....*

- (3a) While the man was visiting the children played outside.  
(3b) While the man was visiting, the children played outside.

Their results showed that participants produced more transitive sentences following (3a) than following (3b), providing evidence that the initial incorrect analysis remained activated. However again it is possible to infer that in the example (3a) the man was visiting the children who played outside. It is therefore unclear from previous studies that the initial incorrect analysis would still remain activated and its interpretation would persist even where such pragmatic inferences are not possible.

To address this issue, the current study tested Japanese relative clause structure such as (4).

- (4) *Akachan-ga nomimono-o koboshita joyuu-o jitto mitusmeta.*  
Subject [RC object RC verb] RC head adverb MC verb  
Baby-NOM [drink-ACC spilled] actress-ACC fixedly stared at  
'The baby stared fixedly at the actress who spilled the drink.'

In Japanese, relative clauses precede lexical heads without an overt complementizer and without any grammatical marking on the verb. This creates local syntactic ambiguity up to the first verb (*koboshita*, 'spilled'); on hearing the verb, the structure is ambiguous between a main clause (MC, henceforth) and a relative clause (RC). It is known that people initially analyze the first verb as a part of MC and construct a sentence representation of the sentence-initial

NP (*akachan*, ‘baby’) to be the agent-subject of the verb, as in (4a). Upon encountering another noun phrase (NP) following the verb (i.e., *joyuu*, ‘actress’), readers are forced to reanalyze for a correct syntactic structure that the initial verb phrase is a part of an embedded RC that modifies the RC-head NP (4b).

- (4a) [*Akachan-ga nomimono-o koboshita*]  
 (4b) *Akachan-ga [nomimono-o koboshita] joyuu-o*

Crucially, the correct interpretation following reanalysis with Japanese RC sentences makes the interpretation of the initial misanalysis pragmatically incompatible. For the example (4), it is impossible to infer that the baby spilled milk after readers correctly reanalyze the structure. This is because Japanese RC structure does not include an implicit argument and therefore there is no ambiguity about who did the spilling action after reanalysis. On the other hand, English sentences in previous studies always included an implicit argument for the verb, which created ambiguity about the direct object.

Using this structure, we manipulated semantic bias of the direct object noun within the RC. Past research demonstrated that readers integrate non-structural information without delay in processing similar ambiguous sentences (e.g., McRae, Spivey-Knowlton, & Tanenhaus, 1998; Garnsey, Pearlmuter, Myers, & Lotocky, 1997). McRae et al. (1998) showed that when a subject noun is plausible as patient but implausible as agent (e.g., *The crook arrested by the detective was guilty...*), readers were more likely to consider the infrequent RC structure on encountering the verb and the preposition *by* within the RC. Also, Garnsey et al. (1997) demonstrated that plausibility of post-verbal nouns as a direct object immediately affected the processing of temporary ambiguous sentence complement sentences (e.g., *The senior senator regretted the decision/reporter had ever...*). Importantly, the plausibility effect interacted with structural bias of individual verbs; when the verb was biased toward the sentence complement, there was no ambiguity cost irrespective of whether the post-verbal noun was plausible as a direct object or not. These results indicate that the verb introduces possible thematic roles within its event semantics and readers check the fit with those roles for each candidate phrase and assess the probabilities of possible structures. Importantly, as Garnsey et al. (1997) suggests, the verb bias appears to exert a stronger influence than the thematic-fit of postverbal elements in English.

On the other hand, in Japanese, the semantics of any arguments may affect parsing since the head does not appear until the end of a clause/sentence. This is rather likely as many studies now demonstrated evidence for pre-head structural analysis (Kamide, Altmann, & Haywood, 2003; Miyamoto, 2002). In fact, Inoue (2006) observed greater processing difficulty with similar RC sentences when the subject noun was biased toward the MC analysis than when it was not. In the present study, we adopted a similar manipulation on the RC direct object and examine whether the preservation of the initial incorrect analysis would be observed where there is no room for pragmatic inference

and, if it is, whether the difference in processing difficulty due to the manipulation of semantic bias would in any way be related to the tendency to preserve the initial analysis.

## Experiment 1

We conducted a moving window self-paced reading experiment with word-by-word presentation.

### Participants

Twenty-four native speakers of Japanese, recruited from the student community at the University of Tokyo, participated in the experiment in exchange for small remuneration.

### Materials

We created 24 sets of experimental items such as (5). We manipulated semantic bias of the direct object for the two alternative analyses at three levels; *MC-biased* (the RC object is biased toward the MC analysis, 5a), *Neutral* (the RC object is neutral toward either the MC and the RC analysis, 5b), and *RC-biased* (the RC object is biased toward the RC analysis, 5c).

#### (5a) *MC-biased*

*Akachan-ga miruku-o koboshita joyuu-o jitto mitusmeta.*  
 Baby-NOM [milk-ACC spilled] actress-ACC fixedly stared at  
 ‘The baby stared at the actress who spilled the milk.’

#### (5b) *Neutral*

*Akachan-ga nomimono-o koboshita joyuu-o jitto mitusmeta.*  
 ‘The baby stared at the actress who spilled the drink.’

#### (5c) *RC-biased*

*Akachan-ga shanpan-o koboshita joyuu-o jitto mitusmeta.*  
 ‘The baby stared at the actress who spilled the champagne.’

If the manipulation of semantic bias has an influence on processing the RC sentences before the RC head is encountered, the sentences in the MC-biased condition and the Neutral condition should initially be analyzed as a MC and exhibit processing difficulty at the disambiguating region. People may commit more strongly to the MC with the former condition than the latter condition because the MC analysis is highly plausible. In contrast, the sentences in the RC-biased condition may be initially analyzed as a RC or be reanalyzed easier because of the MC analysis is less plausible and the RC analysis is highly plausible.

### Design

Three experimental lists were created using a Latin square design, in which each experimental item appears only once in one condition in each list. Each list contained 72 fillers. The comprehension questions followed all the 24 experimental items and 72 fillers. All the questions following the experimental items inquired about the agent of the event denoted by the relative clauses with two options shown below the questions (e.g., *Who spilled the milk? BABY / ACTRESS* for (5a)).

## Procedure

Before each trial, participants saw a star that appeared on the left edge of the screen. They then pressed the space bar to reveal the first word in the sentence. They pressed the space bar to reveal the next word, following which the previous one was masked again. They continued doing so until they reached the end of the sentence. The experimental session began with two practice items.

## Results

**Reading Times** We removed all trials including a region with a reading time that was either extremely long (8000 ms or over) or extremely short (250 ms or under) (Sturt, Pickering, & Crocker, 1999). This resulted in the exclusion of 5.0% of the whole data. For the remaining data, all reading times over 2.5 standard deviations either side of the mean for each participant and each region were replaced with the cut-off value (Sturt, Pickering & Crocker, 1999). We analyzed the reading times using Linear Mixed Effects models (e.g., Baayen, 2008; Baayen, Davidson, & Bates, 2008), including Semantic Bias as a fixed effect with Number of Characters as a covariate. We also included participants and items as random effects. Furthermore, we checked whether the model improved its fit by adding random slopes for each participant and item with a forward-selection approach. Figure 1 shows the mean reading times in the disambiguating region; the RC head (e.g., actress-ACC) per condition.

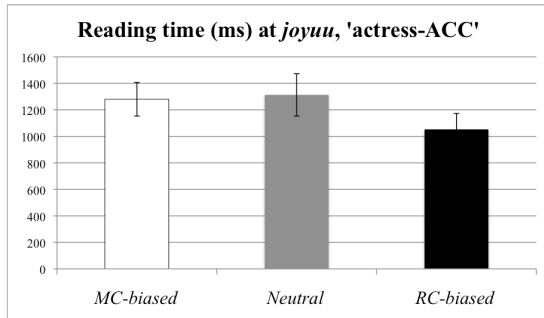


Figure 1: Mean reading times for each condition (Experiment 1). Error bars denote 95% confidence intervals.

The analysis on the reading time in the disambiguating region revealed significant differences between the conditions. In the RC-biased condition, the disambiguating region was read faster compared to the other two conditions ( $\beta = -217.9$ ,  $t = 2.94$ ,  $p < 0.01$  for RC biased vs. MC biased;  $\beta = -245.9$ ,  $t = 3.33$ ,  $p < 0.01$  for RC biased vs. Neutral). There was no difference between the MC-biased and the Neutral conditions ( $t < 1$ ).

**Comprehension Accuracy** Table 1 shows the percentage of correct answers for comprehension questions. We analyzed the log-odd of correct answers using LME models with a binomial function following the same model selection procedure as for the reading time analysis.

The results showed no difference between the conditions ( $z < 1$  for MC-biased vs. Neutral;  $\beta = 0.03$ ,  $z = 1.50$ ,  $p =$

0.13 for MC-biased vs. RC-biased,  $\beta = -1.49$ ,  $z = 1.66$ ,  $p = 0.10$  for RC-biased vs. Neutral).

Table 1: Percentage of Correct Answers for the Comprehension Questions.

MC-biased	96.1%
Neutral	95.6%
RC-biased	98.9%

## Experiment 2

In Experiment 1, we observed a difference in reading times due to bias of the RC direct object but failed to observe its effect on the responses to the comprehension questions. One possibility for this is that processing of the RC structure may have been relatively easy despite the difference in the reading times (i.e., a ceiling effect). In Experiment 2, we thus used sentences with the longer ambiguous region in an effort to increase the processing cost by leading participants to commit to the MC analysis for a prolonged period.

## Participants

Thirty native speakers of Japanese were recruited from the same population as in Experiment 1. None of them participated in the previous experiment.

## Materials, Design, and Procedure

The material and design were identical to those in Experiment 1, except that we added two adverbial phrases (underlined) to lengthen the ambiguous region in the relative clause as in (6).

(6a, b, c) MC-biased / Neutral / RC-biased

*Akachan-ga miruku / drink / shanpan-o te-buru-de hade-ni koboshita joyuu-o jitto mitusmeta.*

‘The baby stared at the actress who spilled the milk / drink / champagne wildly on the table.’

## Results

The reading time at the disambiguating region and the responses to the comprehension questions were analyzed following the same procedure as in Experiment 1.

**Reading Times** Figure 2 shows the mean reading time in the disambiguating region; the RC head (e.g., actress-ACC) for each condition.

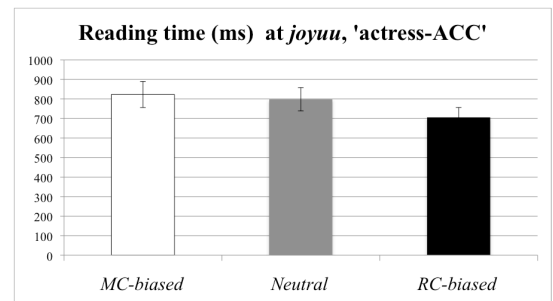


Figure 2: Mean reading times for each condition (Experiment 2). Error bars denote 95% confidence intervals.

We removed invalid trials with the same criteria as Experiment 1, which resulted in exclusion of 8.3% of the whole data. The results showed that the reading time in the disambiguating region was significantly longer in the MC-biased and Neutral conditions compared to that in the RC-biased condition ( $\beta = -108.78$ ,  $t = 3.63$ ,  $p < .001$  for RC-biased vs. MC-biased;  $\beta = -83.78$ ,  $t = 2.78$ ,  $p < 0.01$  for RC-biased vs. Neutral). There was no difference between the MC-biased and Neutral conditions ( $t < 1$ ).

**Comprehension Accuracy** Table 2 shows the percentage of correct answers for the comprehension questions in Experiment 2. The analysis revealed that participants answered more incorrectly in MC-biased condition than in RC-biased condition ( $\beta = -1.17$ ,  $z = 2.11$ ,  $p < 0.05$ ). No significant difference was observed between the Neutral and RC-biased conditions ( $\beta = 0.94$ ,  $z = 1.64$ ,  $p = 0.10$ ), or between the MC-biased and Neutral conditions ( $z < 1$ ). The results demonstrate that when the RC direct object was biased toward the MC and the RC was relatively long, participants tended to preserve the initial incorrect analysis compared to when it was biased toward the RC. The Neutral condition showed a similar pattern to the MC-biased condition, which may suggest that the difference in plausibility for the RC interpretation (i.e., champagne or drink for an actress) did not influence the persistence of the initial analysis. However, the effect was weaker and did not turn out to be fully significant.

Table 2: Percentage of Correct Answers for the Comprehension Questions

MC-biased	92.9%
Neutral	93.8%
RC-biased	97.1%

### Experiment 3

Experiment 2 revealed both the difference in reading time and that in question accuracy, suggesting that preservation of the initial analysis is related to the processing cost. However, the results from Experiment 1, which showed only the difference in reading time, are inconsistent with this. Since the semantic bias and the length of ambiguous region were not tested in a single experiment, it is not clear if the processing cost was any greater in Experiment 2 than 1. Thus, in the next experiment, we crossed the two factors to examine the interaction of the two manipulations. Also, since people could not make any regressions to earlier regions with the self-paced reading task and it is arguable that people may have adopted some task-specific strategy to deal with the structural ambiguity. The next experiment therefore examined eye-movements in normal reading of these sentences. In this experiment, we focus on the interaction between the semantic bias and clause length and its influence on processing cost. We did not include comprehension questions as unrestrained regressive eye-movements make it unlikely to find a subtle difference in response accuracy.

### Participants

Twenty-eight native speakers of Japanese were recruited from the same population as in experiment 1 and 2. None of them participated in the previous experiments.

### Materials and Design

We created 24 sets of experimental items such as (6) with a  $2 \times 2$  design (Semantic Bias [*MC-biased*, *RC-biased*])  $\times$  RC Length [*Short RC*, *Long RC*]). For the analysis, each sentence was divided into six regions as shown below, separated by vertical lines (|).

#### (6a) MC-biased + Short RC

Akachan-ga | miruku-o | koboshita | joyuu-o | jitto | mitusmeta.  
'The baby stared at the actress who spilled the milk.'

#### (6b) MC-biased + Long RC

Akachan-ga | miruku-o | te-buru-de hade-ni koboshita | joyuu-o | jitto | mitusmeta.  
'The baby stared at the actress who spilled the milk wildly on the table.'

#### (6c) RC-biased + Short RC

Akachan-ga | shanpan-o | koboshita | joyuu-o | jitto | mitusmeta.  
'The baby stared at the actress who spilled the champagne.'

#### (6d) RC-biased + Long RC

Akachan-ga | shanpan-o | te-buru-de hade-ni koboshita | joyuu-o | jitto | mitusmeta.  
'The baby stared at the actress who spilled the champagne wildly on the table.'

### Procedure

Four lists of items were created following Latin square design. Each list included 72 fillers and was presented in pseudo-random order. The eye-movements were recorded using Eye-link II (SR Research) at the sampling rate of 500 Hz. Participants first underwent a brief calibration procedure. Before each trial, participants were required to fixate a square box that appeared in the position of the first character of the sentences, which triggered the presentation of sentences. They pressed the space bar when they finished reading each sentence. Thirty-two comprehension questions were included to keep the participants focused.

### Results

We removed fixations that were either extremely long (1200 ms or over) or extremely short (80 ms or under). For statistical analysis, we focus on two eye-movement measures; first-pass reading times and second-pass reading times. *First-pass reading times* are the sum of the fixations in the region following the first entry in the region until the first fixation outside the region (either to the left or the right). *Second-pass reading times* are the sum of fixations made in a region after the region has already been exited to the right. The former measure did not include trials when the region was skipped (i.e., the value of zero) whereas the latter measure did. It is generally assumed that first-pass reading times reflect the early stage of processing and second-pass the late stage. Table 3 shows mean reading



times from Region 2 to Region 5 per condition. In the LME model, we included Semantic Bias (MC-biased vs. RC-biased) and RC Length (short RC vs. long RC) as well as the interaction of the two factors as fixed factors, and participants and items as random factors. Random slopes of the two fixed factors and of the interaction were included for participants and items.

**First-pass reading times** In Region 2, there was an effect of Semantic Bias ( $\beta = 29.2$ ,  $t = 3.64$ ,  $p < 0.001$ ). Participants read this region slower with the sentences in the RC-biased than in the MC-biased condition. In Region 3, we found an effect of RC Length ( $\beta = 22.6$ ,  $t = 10.86$ ,  $p < 0.001$ ), which simply reflects that the longer region took longer to read. In Region 5, which is the *spill-over* region following the disambiguating phrase, there was an interaction between the two factors ( $\beta = 19.2$ ,  $t = 3.21$ ,  $p < 0.01$ ). Further analysis on the effect of Semantic Bias for each level of RC Length showed that the two simple effects were in the opposite direction and were both marginally significant ( $\beta = -19.2$ ,  $t = 1.93$ ,  $p = 0.06$  for Short RC condition;  $\beta = 18.7$ ,  $t = 9.93$ ,  $p = 0.07$  for Long RC condition); Participants tended to read this region faster in the MC-biased than in the RC-biased condition when the RC was long, but they did slower in RC-biased condition than in MC-biased condition when the RC was short. In fact, the mean fast-pass time in the MC-biased and long RC condition was shortest across conditions and this appears *prima facie* at odds with our prediction. However, an additional analysis on the regression-out rate (the probability of regressive eye-movements) revealed that participants made the highest rate of regressive eye-movements in this region in MC-biased + Long RC

condition (0.42). The analysis with the LME model showed an effect of Semantic Bias ( $\beta = 0.5$ ,  $z = 4.59$ ,  $p < 0.001$ ) showing that there were more regressions in the Long RC condition than in the Short RC condition. There also was a marginally significant effect of Semantic Bias ( $\beta = -0.2$ ,  $z = 1.88$ ,  $p = 0.06$ ), showing that there was more regressions in the MC-biased condition than in the RC-biased condition. This indicates that the shortest first-pass time in the MC-biased + Long RC condition was not a reflection of reduced processing difficulty but on the contrary it reflected the excessive processing difficulty that forced participants to immediately regress to the previous region for reanalysis.

**Second-pass reading times** In Region 2, there was a main effect of RC Length ( $\beta = 61.2$ ,  $t = 4.65$ ,  $p < 0.001$ ). This suggests that in this region participants experienced greater cost for reanalysis when the RC was long than it was short. From Region 3 to Region 5, there were effects of Semantic Bias ( $\beta = -136.9$ ,  $t = 3.41$ ,  $p < 0.01$  in Region 3;  $\beta = -37.9$ ,  $t = 3.22$ ,  $p < 0.01$  in Region 4;  $\beta = -44.2$ ,  $t = 3.60$ ,  $p < 0.001$  in Region 5) and those of RC Length although that in Region 5 was marginal ( $\beta = 455.1$ ,  $t = 8.67$ ,  $p < 0.001$  in Region 3;  $\beta = 36.0$ ,  $t = 3.32$ ,  $p < 0.05$  in Region 4;  $\beta = -41.2$ ,  $t = 3.46$ ,  $p = 0.07$  in Region 5). This suggests that in these regions participants experienced greater cost for reanalysis when the RC direct object was biased toward the MC than when it was toward the RC and also did so when the RC was long than when it was short. Importantly, there was an interaction between the two factors in Region 3 ( $\beta = -86.8$ ,  $t = 2.35$ ,  $p < 0.05$ ). Further analysis revealed that the effect of Semantic Bias was larger when the RC was long ( $\beta = -224.9$ ,  $t = 3.63$ ,  $p < 0.001$ ) compared to when it was short ( $\beta = -50.9$ ,  $t = 2.41$ ,  $p < 0.05$ ).

Table 3: Mean reading times for first-pass and second-pass.

	Region 2 (milk/champagne-ACC)	Region 3* (spilled (wildly on the table))	Region 4 (actress-ACC)	Region 5 (fixedly)
<i>First-pass reading time</i>				
MC-biased + short relative clause	283	260 (56)	262	316
MC-biased + long relative clause	288	701 (54)	265	275
RC-biased + short relative clause	330	283 (62)	283	278
RC-biased + long relative clause	356	735 (57)	265	313
<i>Second-pass reading time</i>				
MC-biased + short relative clause	337	353 (78)	253	239
MC-biased + long relative clause	464	1462 (116)	329	331
RC-biased + short relative clause	281	251 (54)	182	159
RC-biased + long relative clause	403	1006 (79)	249	244

\*Region 3 differs in the number of words across conditions, so that the reading time per character is provided in brackets.

## General Discussion

Experiment 1 showed that participants experienced greater difficulty at the disambiguating information when the direct object in the relative clause was biased toward the main clause compared to when it was toward the relative clause. However, there was no effect of the manipulation on the

responses to comprehension questions. Experiment 2 showed the same pattern of results for reading times when the relative clause was lengthened. Importantly, the results from the comprehension questions showed the effect of semantic biases; participants responded less accurately when the direct object was biased toward the main clause compared to when it was toward the relative clause. In

Experiment 3, the results of eye-tracking reading data showed participants had more difficulty for reanalysis when the direct object was biased toward the main clause than when it was toward the relative clause. It also showed that the reanalysis cost was greater when the relative clause was longer than when it was short. Importantly, the effect of semantic biases was larger when the relative clause was long than when it was short. The results taken together provided evidence that the initial incorrect analysis persisted even in absence of pragmatic inference and that it is related to how much people commit to the initial analysis and how much difficulty they experience for reanalysis. We argue that our results are consistent with previous studies in English which showed that the more committed readers were to the initial analysis, the more difficulty they experienced in reanalysis (i.e., “digging-in” effect, Tabor & Hutchins, 2004; see also Ferreira & Henderson, 1991; Ferreira & Henderson, 1998; Frazier & Clifton, 1998; Warner & Glass, 1987).

In a previous study, there was no clear evidence for the effect of phrase/clause length on the persistence of the initial incorrect structure (Experiment 1 in van Gompel et al., 2006) and also for that of the manipulation of plausibility (Experiment 2 in the same study). One possible reason for this discrepancy is that these two factors were tested independently. Yet, another possibility is that this is due to a qualitative difference in how these ambiguous sentences were processed between English and Japanese. With English, it has been shown that readers adopt an inappropriate transitive analysis even when the post-verbal noun phrase is semantically inappropriate as a direct object (Pickering & Traxler, 1998). It is likely that this is at least partly due to head-driven parsing; speakers of English check the fit as a direct object for a post-verbal noun phrase regardless of the verb type (except unaccusative verbs; see Staub, 2007). On the other hand, in the head-final Japanese, it is possible that the plausibility of arguments influences the pre-head structural analysis independently of the verb. That is, in some trials under the RC-bias condition, participants may have not adopted the main clause analysis even as an initial analysis and this may have resulted in the higher accuracy to the comprehension questions in Experiment 1 and 2 and in the less reanalysis cost compared to the MC-bias condition in Experiment 3.

To summarize, the current study provided evidence that comprehenders tend to preserve the initial analysis even when the sentence structure does not permit pragmatic inferences. Our finding of the effects of semantic biases and clause length revealed that such a tendency is related to the degree of processing difficulty that reflects how much people committed to the misanalysis. The results of eye-tracking data showed that participants indeed experienced excessive processing cost with the sentences when both the semantic bias and clause length encourage the main clause analysis. The current study also provided evidence for pre-head processing in Japanese and also demonstrated that the persistence of initial misanalysis that has been reported in a head-initial language such as English occurs in typologically different head-final language.

## Acknowledgments

We would like to thank Edson T. Miyamoto and Yuki Hirose for their help on Experiment 3.

## References

- Baayen, R. H. (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.
- Christianson, K., Hollingworth, A., Halliwell, J., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42, 368-407.
- Christianson, K., Williams, C. C., Zacks, R. T., & Ferreira, F. (2006). Younger and older adults' "good-enough" interpretations of garden-path sentences. *Discourse Processes*, 42, 205-238.
- Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11, 11-15.
- Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, 30, 725-745.
- Ferreira, F., & Henderson, J. M. (1998). Syntactic reanalysis, thematic processing, and sentence comprehension. In J.D. Fodor & F. Ferreira (Eds.), *Reanalysis in sentence processing* (pp. 73-100). Dordrecht, The Netherlands: Kluwer Academic.
- Frazier, L., & Clifton, C. (1998). Sentence reanalysis and visibility. In J.D. Fodor & F. Ferreira (Eds.), *Reanalysis in sentence processing* (pp. 143-176). Dordrecht, The Netherlands: Kluwer Academic.
- Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37, 58-93.
- Inoue, M. (2006). Ambiguity resolution or retention in comprehending Japanese sentences. *Cognitive Studies*, 13, 353-368.
- Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49, 133-156.
- McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38, 283-312.
- Miyamoto, E. T. 2002. Case markers as clause boundary inducers in Japanese. *Journal of Psycholinguistic Research*, 31, 307-347.
- Pickering, M. J., Traxler, M. J. (1998). Plausibility and recovery from garden paths: an eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 940-961.
- Staub, A. (2007). The parser doesn't ignore intransitivity, after all. *Journal of Experimental Psychology-Learning Memory and Cognition*, 33, 550-569.
- Sturt, P., Pickering, M. J., & Crocker, M. W. (1999). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, 40, 136-150.
- Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: Digging in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 431-450.
- Warner, J., & Glass, A. L. (1987). Context and distance-to-disambiguation effects in ambiguity resolution: Evidence from grammaticality judgments of garden path sentences. *Journal of Memory and Language*, 26, 714-738.

# The Role of the Amygdala in the Process of Humor Appreciation

**Tagiru Nakamura (tagiru@sfc.keio.ac.jp)**<sup>1)</sup>

<sup>1)</sup> Faculty of Environment and Information Studies, Keio University,  
5322 Endo, Fujisawa, Kanagawa 252-8520, Japan.

**Tomoko Matsui (matsui@u-gakugei.ac.jp)**<sup>2)</sup>

<sup>2)</sup> Center for Research in International Education, Tokyo Gakugei University,  
4-1-1 Nukuikitamachi, Koganei, Tokyo 184-8501, Japan.

**Akira Utsumi (utsumi@inf.uec.ac.jp)**<sup>3)</sup>

<sup>3)</sup> Department of Informatics, Graduate School of Informatics and Engineering, The University of Electro-Communications,  
1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan

**Mika Yamazaki (mika@u-fukui.ac.jp)**<sup>4,6)</sup>

**Kai Makita (kai@nips.ac.jp)**<sup>4,5)</sup>

**Hiroki C. Tanabe (htanabe@nips.ac.jp)**<sup>4,5)</sup>

**Norihiro Sadato (sadato@nips.ac.jp)**<sup>4,5)</sup>

<sup>4)</sup> Department of Cerebral Research, Division of Cerebral Integration, National Institute for Physiological Sciences (NIPS),  
38 Nishigonaka, Myodaiji, Okazaki, Aichi 444-8585, Japan

<sup>5)</sup> Department of Physiological Sciences, School of Life Science, The Graduate University for Advanced Studies,  
Shonan Village, Hayama, Kanagawa 240-0193, Japan

<sup>6)</sup> Research Center for Child Mental Development, Graduate School of Medical Sciences, University of Fukui,  
23-3 Shimoaizuki, Matsuoka, Eiheiji, Fukui 910-1193, Japan

## Abstract

The purpose of this study was to investigate the neuropsychological process and timing of appreciating humor upon reading a set of Japanese riddles which were made up of 4 distinct phases: “Given A (1st phase), I’d say B (2nd phase). Do you know why? (3rd phase) It is because X (4th phase).” To investigate how the brain responds when an individual finds the answer to the riddle (“It is because X”) humorous, this study used the fMRI method. We found that when a participant judged the answer to the riddle humorous, the bilateral amygdalae had been significantly activated at the 4th phase. When the participant judged the answer to the riddle banal (non-humorous), by contrast, the left amygdala was found to be significantly activated earlier at the 3rd phase. Therefore, we argue that in both cases, activation of the amygdala is related to the detection of optimal relevance.

**Keywords:** amygdala; detection of optimal relevance; fMRI; process of humor appreciation; Relevance Theory.

## Introduction

Timing control is essential for a type of Japanese riddle in the following example, which has the fixed form of “Given A (target word), I’d say B (response word). Do you know why? (question) It is because X (answer)” (“A *to kakete*, B *to toku*, *sono kokoro wa*, X.”).

“Given ‘the savings,’ I’d say ‘a smile of my wife’. Do you know why? It is because if they disappear, I will be in trouble.”

In this sequence of utterances, two seemingly unrelated items were mentioned by the speaker and the hearer is expected to reason how on earth they are connected. The process of mentally searching for the connection and discovering it finally on the basis of the rationale provided by the speaker yield humorous effects. The sequence has four distinctive phases: the 1st phase introduces the target concept (“Given A”); the 2nd phase introduces the response concept (“I’d say B”); the 3rd phase asks the hearer if he knows the rationale for making connection between the two concepts (“Do you know why?”); and the 4th phase provides the rationale for the supposed connection (“It is because X”). In this functional magnetic resonance imaging (fMRI) study, we presented a set of such sequences to participants, and then asked them to judge whether or not the given riddles are humorous after they heard the rationale.

## Incongruity resolution in humor appreciation

According to a standard humor appreciation model (Suls, 1972; Wyer & Collins, 1992; Yus, 2003; Martin, 2006), incongruities in the content of the utterance must be identified and resolved for it to be humorous. The incongruities are typically caused by violation of a set of expectations stored in “mental schemas” which are “formed on the basis of past experience with objects, scenes, or events and consists of a set of (usually unconscious) expectations about what things look like and/or the order in which they occur” (Mandler, 1979, p. 263). Thus, in humor appreciation, the simultaneous activation of two incompatible schemas is essential (Wyer & Collins, 1992).

In the Japanese riddle, it is structurally controlled. During the 1st and the 2nd phase, two different schemas corresponding to a target word and a response word are likely to be activated. Two incompatible schemas are simultaneously activated between the 2nd and the 4th phase. When incongruities between these schemas are noticed, the hearer of the riddle naturally looks for a novel and interesting way to resolve them. When a rationale for connecting the two seemingly unrelated concepts is provided at the 4th phase, the incongruities get resolved. At that point, the two existing incompatible schemas and a new schema corresponding to the rationale given are likely to be activated simultaneously. The success of the humor relies upon this resolution being inaccessible before it is given, yet obvious with hindsight when given. If either the resolution fails to be logically coherent or is instead too obvious, the riddle is judged banal.

According to Relevance Theory (Matsui, 2000; Sperber & Wilson, 1995; Yus, 2003; Utsumi, 2005), the motivation for appreciating humor comes from search for “relevance,” which is a property of inputs to cognitive processes and is determined by the balance between the cognitive effect and processing effort. An input has cognitive effect if it significantly improves a cognitive environment of an individual. In order to process the input, though, some processing effort is required. Relevance Theory has two fundamental principles. The first, or the cognitive principle of relevance, states that “human cognition tends to be geared to the maximization of relevance.” The second, or the communicative principle of relevance, states that “every act of ostensive communication communicates a presumption of its own optimal relevance” (Sperber & Wilson, 1995, p. 260). The first principle of relevance predicts that an individual has an innate tendency to process an input to yield maximum cognitive effect with least possible processing effort. The second principle predicts that as a hearer, an individual automatically expects that the speaker will provide information which yields enough cognitive effect in return for the processing cost.

A relevance-theoretic view of interpretation of the Japanese riddles is based on the second principle. When encountering the Japanese riddles, an individual, who expects that it would provide enough cognitive effect with least possible processing effort, automatically tries to resolve the incongruities between the two seemingly unrelated schemas corresponding to the target and the response words. When the individual is finally given the rationale for connecting the two unrelated concepts at the 4th phase, he is likely to feel “Aha!” At that point, he is satisfied with the interpretation and stops processing it.

Based on this principle, it was predicted that, if the answer to the riddle was finally judged humorous, there would be little relevance-based understanding in the 3rd phase because no cognitive effect was yet achieved at that time point, and the intensity of relevance-based understanding would go up in the 4th phase because enough cognitive effect was finally achieved then to repay the large

processing effort spent during the 3rd phase. It was also predicted that, if a riddle was finally judged banal (non-humorous), relevance-based understanding would peak at the 3rd phase since the individual already found a rationale satisfying optimal relevance. Since the riddle was in this case resolved already by the 3rd phase, it was predicted that there would be no further relevance-based understanding in the 4th phase.

### **The amygdala as a detector of optimal relevance**

Previous lesion and neuroimaging studies have shown that the amygdala is involved in an evaluation of motivationally relevant events (Sander, Grafman, & Zalla, 2003; Zald, 2003; Bach et al., 2008). Several studies have demonstrated the key role of the amygdala in negative emotion, but a few studies have suggested a corresponding role for the amygdala in both positive and negative emotion (Hamann & Mao, 2002; Burgdorf & Panksepp, 2006). So, the view of the amygdala as a relevance detector has been proposed (Sander et al., 2003; Zald, 2003), and many neuroimaging studies have supported the hypothesis (Bach et al., 2008; Ousdal et al., 2008; Herbert et al., 2009; Bach et al., 2011). In this article, we take the hypothesis a step further and propose that the amygdala extends its function in humans as a relevance detector in ostensive communication. Thus, it was predicted that the amygdala would be activated if relevance-based understanding occurred. It was also predicted that the amygdala would be deactivated if no such relevance-based processing occurred.

Previous neuroimaging humor studies report the activation of subcortical structures including the amygdalae during humor appreciation (Mobbs, Greicius, Adabel-Azim, Menon, & Reiss, 2003; Moran, Wig, Adams, Janata, & Kelley, 2004; Bartolo, Benuzzi, Nocetti, Baraldi, & Nichelli, 2006; Wild, Rodden et al., 2006; Watson, Matthews, & Allman, 2007; Bekinschtein, Davis, Rodd, & Owen, 2011; Kohn, Kellermann, Gur, Schneider, & Habel, 2011). We propose that activation of amygdala in humor appreciation can be interpreted as the result of detecting the optimal relevance in humorous utterances — the “Aha” reaction. Other main common activation areas of these humor studies include the left hemisphere of the cerebral cortex: an area around the left fusiform gyrus or the left temporo-occipital junction (Brodmann’s area [BA] 37) ) to detect incongruity in humorous utterances (Mobbs et al., 2003); the left posterior middle temporal gyrus (BA 21) for semantic comprehension of humor (Moran et al., 2004); the left inferior frontal gyrus including Broca’s area (BA 44/45) to resolve the incongruity or ambiguity in humorous utterances (Mobbs et al., 2003; Moran et al., 2004; Bekinschtein et al., 2011); and the medial frontal cortex (Goel & Dolan, 2001; Mobbs et al., 2003; Kohn et al., 2011).

In this study, we used the fMRI method to investigate the relationship between humor appreciation and its time course that composed two factors: the presence/absence of humor (humorous vs. non-humorous) and the expecting/shown rationale (the 3rd phase vs. the 4th phase). We predicted that

when a participant explicitly judged the answer to the riddle humorous at the end, its relevance would be detected implicitly at the 4th phase. In addition, it was predicted that when the participant judged the answer to the riddle non-humorous at the end because it was banal, its relevance would be detected earlier at the 3rd phase.

## Methods

### Participants

Twenty participants (10 females and 10 males; mean age, 23.3 years; range, 18–37 years) were recruited as paid volunteers for the fMRI experiment. All participants had normal/corrected-to-normal visual acuity, and were right-handed according to the Edinburgh handedness inventory (Oldfield, 1971), and no history of neurological/psychiatric illness. They were educated higher than high-school graduates. Written informed consent in order to take part in this study was obtained following procedures approved by the Ethical Committee of the National Institute for Physiological Sciences, Japan.

### Preparation of task materials

The riddles had the following structure: “Given A, I’d say B. Do you know why? It is because X.” The topics for the riddles were obtained from an article (Nakamura, 2009) and Internet searches by Google (<http://www.google.com>) in order to create candidates for strongly humorous riddles. For example, “Given ‘savings,’ I’d say ‘a smile of my wife,’ Do you know why? It is because if they disappear, I will be in trouble.” Then, we altered the response word of each riddle in order to create candidates for weakly humorous version. For example, “Given ‘savings,’ I’d say ‘a credit card,’ Do you know why? It is because if they disappear, I will be in trouble.”

In order to select well-controlled pairs of riddles, 8 normal volunteers (4 females and 4 males; mean age, 27.0 years; range, 22–44 years) participated in a pilot study. We selected a paired riddle if one of the pair was evaluated humorous by over one third of participants and the other one was evaluated non-humorous by over one third of participants. We obtained 24 topics from the article and created 24 pairs of riddles, but 8 were removed after the pilot study. We also obtained 20 topics from the Internet searches and created 20 pairs of riddles, but 2 were rejected. So, finally, we selected 34 pairs of riddles and used them in the fMRI session.

### fMRI procedures

Prior to the fMRI session, the participants received detailed instructions and an explanation of the task procedure, and were trained with training stimuli that were not used during the fMRI session. All stimuli were prepared and presented using Presentation 14.8 software (Neurobehavioral Systems, Albany, CA) running on a personal computer. Using a liquid crystal display (LCD) projector, the visual stimuli

were projected onto a half-transparent viewing screen located behind the head coil of the magnetic resonance imaging (MRI) scanner. Participants viewed the stimuli via a mirror attached to the head coil. The sentence stimuli were written in Japanese and presented with white letters on a black background. The maximum visual angle was  $7.8^\circ$  (horizontal)  $\times$   $0.9^\circ$  (vertical).

In each trial, the 1st phase was presented on the screen for 1.5 sec followed by a cross-hair for 1.25 sec, then the 2nd phase appeared for 2 sec followed by a cross-hair for 1.25 sec, then the 3rd phase appeared for 0.75 sec followed by a cross-hair for 1.75 sec, after that the 4th phase was presented for 3.5 sec followed by a cross-hair for 2 sec. The length of each phase was adjusted corresponding to the maximum length of presented stimuli because timing control was important for the Japanese riddles, yet the length of the 3rd phase was shortened and the time between the 3rd and the 4th phase was lengthened because of reducing the value of correlation coefficients of regressors in imaging data analyses. Then the participant was required to judge whether or not the riddle is humorous and to press the button after the question mark “?”, which was presented for 1 sec followed by a cross-hair for 5 sec, was presented.

We used an event-related design to minimize habituation and learning effects. The 34 paired riddles were presented in a pseudorandom order. The paired riddles were presented in different sessions. In total, two sessions, each with 17 candidates of humorous riddles and 17 candidates of non-humorous riddles, were run, and the session order was counterbalanced across participants.

All images were acquired using a 3-Tesla MR scanner (Allegra; Siemens, Erlangen, Germany). For functional imaging during the sessions, an ascending T2\*-weighted gradient-echo echo-planar imaging (EPI) procedure was used to produce 34 continuous 4-mm thick transaxial slices covering the entire cerebrum and cerebellum (time repetition [TR], 2000 ms; time echo [TE], 30 ms; flip angle,  $85^\circ$ ; field of view [FoV], 192 mm;  $64 \times 64$  matrix; voxel dimensions,  $3.0 \times 3.0 \times 4.0$  mm). Oblique scanning was used to exclude the eyeballs from the images. Each session consisted of a continuous series of 354 volume acquisitions with a total duration of 11 min 48 sec. For anatomical imaging, T1-weighted magnetization-prepared rapid-acquisition gradient-echo (MP-RAGE) images were also obtained (TR, 2500 ms; TE, 4.38 ms; flip angle,  $8^\circ$ ; FoV, 230 mm; 1 slab; number of slices per slab, 192; voxel dimensions,  $0.9 \times 0.9 \times 1.0$  mm).

After the fMRI session, the participant was asked to select the best reason for the judgment among the given possibilities (Nakamura, 2009) in order to remove the reasons in which we were not interested. The given options for judging non-humorous were “non-understandable,” “banal,” “objectionable” and “too serious.” The given options for judging humorous were “sympathetic,” “convincing,” “to-the-point” and “close to banal.” The total duration of the experiment was under 60 min, including the acquisition of the structural MR images and these responses.

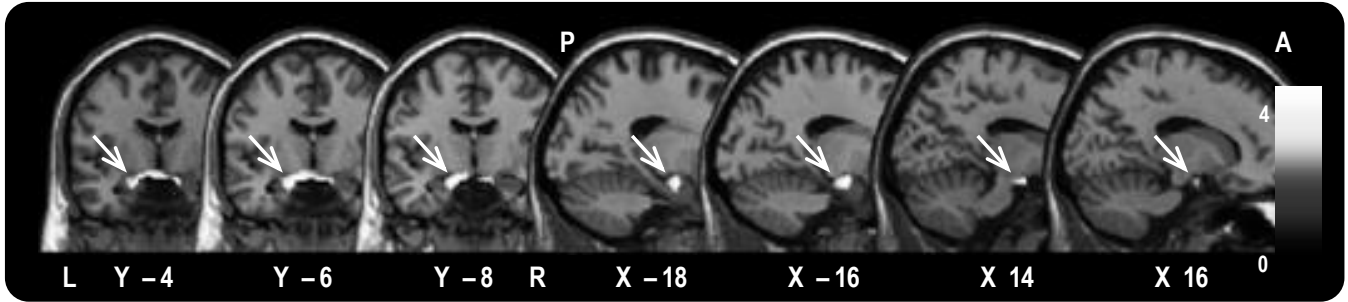


Figure 1: The amygdala as a detector of optimal relevance (an interaction area of the  $2 \times 2$  factorial design).

## Data analysis

**Performance** In this study, we assumed that the judgment about whether or not a riddle is humorous was based on several different reasons, which we separated out using our experimental design. We created paired riddles which contained humorous version and non-humorous version, and we called convincing-banal pairs if one of the pair was judged humorous because it was convincing by a participant and the other of the pair was judged non-humorous because it was banal by the same participant. Only riddles of convincing-banal pairs were included in the following analyses. Other reasons were rejected because they could contain negative values or ambiguous meanings.

**Imaging data** The preprocessing of the imaging data was performed as follows. The first 6 EPI volumes of each session were eliminated to allow for the stabilization of the magnetization, and the remaining 348 EPI volumes per session (a total of 696 EPI volumes per participant for two sessions) were used for the analyses. The data were analyzed using Statistical Parametric Mapping 8 (SPM8; Wellcome Department of Imaging Neuroscience, London, UK; Friston, Ashburner, Kiebel, Nichols, & Penny, 2007). The EPI volumes were realigned to correct for head motion, and corrected for differences in slice timing within each volume. Then, the whole-head MP-RAGE image volume was coregistered with the mean image volume of the EPI volumes, and segmented into the gray and the white matter volumes using the Montreal Neurological Institute T1 image template. The normalizing parameters of the segmentation were applied to the all EPI volumes. After that, the EPI volumes were spatially smoothed in three dimensions using an 8 mm full-width half-maximum Gaussian kernel.

The signal time course for each participant was modeled with a general linear model. Regressors of interest (trial effects) of the 10 conditions (12e, 3e, 4e, 12a, 3a, 4a, 12z, 3z, 4z, and J) were generated using a box-car function convolved with a hemodynamic-response function (Abbreviation: 12, the 1st and the 2nd phase; 3, the 3rd phase; 4, the 4th phase; J, the judgment phase; e, judging humorous (i.e., convincing of convincing-banal pairs); a, judging non-humorous (i.e., banal of the pairs); z, judging as other six reasons in which we were not interested).

The weighted sum of the parameter estimates in the individual analyses constituted the contrast images, which were used for the group analysis with a random effects-model to make population-level inferences regarding the task-related activation. The contrast images obtained by the individual analyses represent the normalized task-related increment of the MR signal of each participant. In total, the data from 15 participants (other 5 contained no convincing-banal pairs) and four different contrasts (3e, 4e, 3a, and 4a) were incorporated into the  $2$  (humor effect)  $\times$   $2$  (its time course effect) factorial design (Friston et al., 2007). As a result of the factorial design, a significant interaction was found in the bilateral amygdalae, which we defined as regions of interest and used as an explicit mask. In order to show the difference of activated timing of the Amygdala, four different contrasts (3e vs. 3a, 3a vs. 3e, 4e vs. 4a, and 4a vs. 4e) were used in the one-sample t-tests with the explicit mask image. In order to check the activation of each phase, six different contrasts against the rest condition (12e, 3e, 4e, 12a, 3a, and 4a) were used in the one-sample t-tests.

The statistical threshold was set at  $p < .05$  with a correction for multiple comparisons at the cluster level.

## Results

### Behavioral performance

During the fMRI experiment, 734 riddles (54.0%) were judged humorous, while 626 riddles (46.0%) were judged non-humorous. After the fMRI session, 231 riddles (17.0%) were judged convincing, while 243 riddles (17.9%) were judged banal. Then, the 43 paired riddles (6.32%) were judged as convincing-banal pairs, which were used in the following imaging data analysis (a random-effects model).

### Group analysis with a random-effects model

In the  $2 \times 2$  factorial design, no significant main effect was found, but a significant interaction was found in the bilateral amygdalae (see Figure 1 for a spread cluster and Table 1 for peak-levels). In the interaction area, the left amygdala was significantly activated in the 4th phase before the riddle was judged humorous and also in the 3rd phase before the riddle was considered banal (non-humorous), while the right amygdala was significantly activated in the 4th phase before the riddle was judged humorous (see Table 2).

Table 1: The amygdala as a detector of optimal relevance (an interaction area of the  $2 \times 2$  factorial design).

Cluster Size	Peak $p$	Z	Coordinates			Location
			x	y	z	
227	.046	4.72	-18	-6	-20	L Amygdala
	.032	4.81	-14	-8	-16	L Amygdala

Table 2: Difference of activated timing of the Amygdala (results of the one-sample t-test in the interaction area).

Cluster		Z	Coordinates			Location
<i>p</i>	Size		x	y	z	
<b>humorous vs. banal (non-humorous) in the 4th phase</b>						
< .001	117	4.43	-18	-6	-18	L Amygdala
.014	17	3.79	16	-4	-20	R Amygdala
.041	2	3.21	2	-2	-12	R Hypothalamus
<b>banal (non-humorous) vs. humorous in the 3rd phase</b>						
.001	57	4.01	-14	-8	-16	L Amygdala
.036	4	3.20	6	-6	-16	R Hypothalamus

When the riddle was to be judged humorous in the end, significant activations were found mainly in the left hemisphere of the cerebral cortex: (12e) the bilateral fusiform gyri (BA 37), the bilateral inferior occipital gyri (BA 17–19), the left middle frontal gyrus (BA 9), and the left middle temporal gyrus (BA 21); (3e) the left inferior frontal gyrus (BA 47/45); (4e) the left posterior rostral medial frontal cortex (prMFC; BA 8/6), the left inferior frontal gyrus (BA 9/45), the left fusiform gyrus (BA 37), and the left inferior occipital gyrus (BA 18/19). When the riddle was to be judged banal (non-humorous), on the other hand, significant activations were found mainly in the left hemisphere and left amygdala: (12a) the bilateral inferior occipital gyri (BA 18–19); (3a) the left amygdala; (4a) the left prMFC (BA 8/6) including the left anterior cingulate cortex (BA 32), the left inferior frontal gyrus (BA 9/45), and the bilateral inferior occipital gyri (BA 18/19).

## Discussion

### Performance

There are clear individual differences in the way humor is appreciated. In this study, we were interested in the amygdala detecting the optimal relevance during humor appreciation, and we used convincing-banal pairs in order to strictly cancel out the effects in which we were not interested. In the 3rd and the 4th phase of a paired riddle, same participants read same stimuli, so only effect of the 2nd phase, which was modeled out in the  $2 \times 2$  factorial design, was remained. Thus, a significance level in the following analysis can be reached with data of only 15 participants.

### Neural activation

The present study revealed that the amygdalae are specifically involved in relevance-based understanding, showing that the amygdalae have a key role in appreciating

a riddle humorous through the time course of humor appreciation. We were also able to reproduce the other activations recorded in other studies.

**Relevance-related activation of amygdala** On the basis of a significant interaction in the  $2 \times 2$  factorial design, we argue that the amygdala is a candidate for relevance-based processing. The left amygdala was significantly activated in the 4th phase when the riddle was to be judged as humorous. It was also activated in the 3rd phase when the riddle was to be judged banal (non-humorous). Thus, we claim that the left amygdala functions as a detector of optimal relevance during humor appreciation. The right amygdala was significantly activated only in the 4th phase when the riddle was to be judged humorous, and hence we argue that the right amygdala is a kind of positive value detector.

While previous studies have suggested that the bilateral or the left amygdala activation involves in positive emotion (Bartolo et al., 2006; Bekinschtein et al., 2011; Mobbs et al., 2003; Moran et al., 2004; Watson et al., 2007), our results suggested that the left amygdala activation involves in relevance detection because it was found not only in the 4th phase of judging humorous (i.e., positive emotion) but also in the 3rd phase of judging non-humorous (i.e., non-positive or negative emotion). If same brain parts are activated in both opposite values, the distinction between positive and negative emotion could not be critical feature, but we argue that the common feature of positive and negative emotion, that is, relevance for the organism is more important.

**Humor-related activation of left hemisphere** Our results also reproduced previous neuroimaging studies of humor appreciation. The activation of the left fusiform gyrus in (12e) and (4e) is interpreted as incongruity detection of humor (Mobbs et al., 2003). The activation of the left middle frontal gyrus in (12e) and the left inferior frontal gyrus in (4e) are interpreted as incongruity or ambiguity resolution of humor (Mobbs et al., 2003; Moran et al., 2004). So, the activation of these areas corresponds to the incongruity resolution a la Suls (1972). The activation of the left middle temporal gyrus in (12e) is interpreted as semantic comprehension of humor (Moran et al., 2004).

Considering that the prMFC has been suggested to represent and update the value of possible future actions such as response selection (Amodio & Frith, 2006), the activation of the left prMFC in (4e) and (4a) is interpreted as being involved in judging whether or not a riddle is humorous. The left inferior frontal gyrus is also activated in both (4e) and (4a), so it might be also involved in the selection. Considering the participants of this study looked at written riddles, the activation of the inferior occipital gyrus is interpreted as seeing effects. So, our results correspond well to previous neuroimaging studies.

## Conclusions

Our results highlight the neural substrates for the detection of optimal relevance. When a participant explicitly judges



the answer to the riddle humorous, the relevance of the answer has been detected implicitly as the activation of the amygdala at the 4th phase. On the other hand, when the participant judges the answer to the riddle non-humorous because it is banal, its relevance has been detected as the activation of the amygdala earlier at the 3rd phase. Therefore, we argue that the amygdala is a detector of optimal relevance in the process of humor appreciation.

## Acknowledgments

This study was partly supported by Scientific Research on Innovative Areas grant #22101007 (H.C.T.) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (MEXT), and Challenging Exploratory Research grant #23650224 (H.C.T.), Grant-in-Aid for Scientific Research (S) #21220005 (N.S.), (A) #2124013 (H.C.T., N.S.), and (B) #20330136 (T.M., T.N.) from the Japan Society for the Promotion of Science. A part of this study represents the results of the “Development of biomarker candidates for social behavior” and “Integrated research on neuropsychiatric disorders” projects carried out under the Strategic Research Program for Brain Science by MEXT.

## References

- Amodio DM, & Frith CD. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, 7, 268–277.
- Bach DR, Grandjean D, Sander D, Herdener M, Strik WK, & Seifritz E. (2008). The effect of appraisal level on processing of emotional prosody in meaningless speech. *NeuroImage*, 42, 919–927.
- Bach DR, Talmi D, Hurlmann R, Patin A, & Dolan RJ. (2011). Automatic relevance detection in the absence of a functional amygdala. *Neuropsychologia*, 49, 1302–1305.
- Bartolo A, Benuzzi F, Nocetti L, Baraldi P, & Nichelli P. (2006). Humor Comprehension and appreciation: An fMRI study. *Journal of Cognitive Neuroscience*, 18, 1789–1798.
- Bekinschtein TA, Davis MH, Rodd JM, & Owen AM. (2011). Why clowns taste funny: The relationship between humor and semantic ambiguity. *The Journal of Neuroscience*, 31, 9665–9671.
- Burgdorf J, & Panksepp J. (2006). The neurobiology of positive emotions. *Neuroscience and Biobehavioral Reviews*, 30, 173–187.
- Friston KJ, Ashburner J, Kiebel SJ, Nichols TE, & Penny WD. (2007). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. London: Elsevier.
- Goel V, & Dolan RJ. (2001). The functional anatomy of humor: Segregating cognitive and affective components. *Nature Neuroscience*, 4, 237–238.
- Hamann S, & Mao H. (2002). Positive and negative emotional verbal stimuli elicit activity in the left amygdala. *Neuro Report*, 13, 15–19.
- Herbert C, Ethofer T, Anders S, Junghofer M, Wildgruber D, Grodd W, & Kissler J. (2009). Amygdala activation during reading of emotional adjectives: An advantage for pleasant content. *Scan*, 4, 35–49.
- Kohn N, Kellermann T, Gur RC, Schneider F, & Habel U. (2011). Gender differences in the neural correlates of humor processing: Implications for different processing modes. *Neuropsychologia*, 49, 888–897.
- Mandler JM. (1979). Categorical & schematic organization in memory. In Puff CR (Ed.), *Memory organization and structure* (pp. 259–299). New York: Academic Press.
- Martin RA. (2006). *The Psychology of Humor: An Integrative Approach*. London: Elsevier Academic Press.
- Matsui, T. (2000). *Bridging and Relevance*. John Benjamins.
- Mobbs D, Greicius MD, Adbel-Azim E, Menon V, & Reiss AL. (2003). Humor Modulates the Mesolimbic Reward Centers. *Neuron*, 40, 1041–1048.
- Moran JM, Wig GS, Adams RB Jr, Janata P, & Kelley WM. (2004). Neural correlates of humor detection and appreciation. *NeuroImage*, 21, 1055–1060.
- Nakamura T. (2009). The mechanism of sensing interestingness in metaphorical expressions. *The Japanese Journal of Psychology*, 80, 1–8.
- Oldfield RC. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, 9, 97–113.
- Ousdal OT, Jensen J, Server A, Hariri AR, Nakstad PH, & Andreassen OA. (2008). The human amygdala is involved in general behavioral relevance detection: Evidence from an event-related functional magnetic resonance imaging Go-NoGo task. *Neuroscience*, 156, 450–455.
- Suls JM. (1972). A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In Goldstein JH and McGhee PE (Eds.), *The psychology of humor: Theoretical perspectives and empirical issues* (pp. 81–100). New York: Academic Press.
- Sander D, Grafman J, & Zalla T. (2003). The human amygdala: An evolved system for relevance detection. *Reviews in the Neurosciences*, 14, 303–316.
- Sperber D, & Wilson D. (1995). *Relevance: Communication and Cognition* (2nd Ed.). Oxford: Blackwell.
- Utsumi A. (2005). The role of feature emergence in metaphor appreciation. *Metaphor and Symbol*, 20, 151–172.
- Watson KK, Matthews BJ, & Allman JM. (2007). Brain Activation during Sight Gags and Language-Dependent Humor. *Cerebral Cortex*, 17, 314–324.
- Wild B, Rodden FA, Rapp A, Erb M, Grodd W, & Ruch W. (2006). Humor and smiling: Cortical regions selective for cognitive, affective, and volitional components. *Neurology*, 66, 887–893.
- Wyer RS, & Collins JE. (1992). A theory of humor elicitation. *Psychological Review*, 99, 663–688.
- Yus F. (2003). Humor and the search for relevance. *Journal of Pragmatics*, 35, 1295–1331.
- Zald DH. (2003). The human amygdala and the emotional evaluation of sensory stimuli. *Brain Research Reviews*, 41, 88–123.

# The Footbridge Dilemma Reflects More Utilitarian Thinking Than The Trolley Dilemma: Effect Of Number Of Victims In Moral Dilemmas

Kuninori Nakamura (nakamura.kuninori@gmail.com)

Graduate School of Decision Science & Technology,

Tokyo Institute of Technology

2-12-1, Ohkayama, Meguro-Ku, Tokyo 152-8552, Japan

## Abstract

Previous studies on moral judgment have assumed that the trolley and footbridge dilemmas (Thomson, 1985) reflect utilitarian and deontologist thinking, respectively. However, on the basis of the “intervention myopia” hypothesis (Waldmann & Dieterich, 2007) and recent findings in analyses of moral dilemmas (Nakamura, 2011), the current study led a somewhat paradoxical prediction: An effect of the manipulation of the number of victims, considered a utilitarian aspect of moral dilemmas, is larger in the footbridge dilemma than in the trolley dilemma. In order to test this prediction, two experimental studies were conducted in which the number of victims in the trolley and footbridge dilemmas were manipulated. Results of the two studies consistently showed an interaction between the dilemma type and the number of victims, thereby indicating that the manipulation of the utilitarian aspect of moral dilemmas has more effect on the footbridge dilemma, which is believed to reflect deontologist thinking.

**Keywords:** trolley dilemma, footbridge dilemma, utilitarian, deontologist

## Introduction

Is it permissible to sacrifice fewer lives to save more? This is a central question in the debate between utilitarianism and deontology. Utilitarians (e.g., Bentham, 1789; 1948) argue that it is indeed permissible because saving more lives results in greater utility for society than saving fewer ones, whereas deontologists (e.g., Kant, 1965) argue that it is not permissible because life is an ultimate right that should not be violated, irrespective of the number benefit yielded by its sacrifice. This debate has drawn the attention of various researchers who have proposed a number of solutions (see e.g., Singer, 1979; Thomson, 1986; Greene & Haidt, 2002; Mikhail, 2009).

The philosophical debate between utilitarians and deontologists concerns the normative theory of moral judgment. However, psychologists are interested in the descriptive aspect of moral judgment: are people utilitarian or deontologist? The answer to this question is, surprisingly, “it depends on the context.” A discrepancy between the trolley and footbridge dilemmas (Thomson, 1985) clearly demonstrates the context dependency in moral judgment. The trolley dilemma can be described in the following manner: A runaway trolley is headed for five people who will be killed if it proceeds on its current course. The only way to save them is to hit a switch that will turn the trolley onto an alternate set of tracks where it will kill one person instead of five. Should one turn the trolley in order to save

five people at the expense of one? Most people answer yes to this dilemma. Then, consider a similar problem, the footbridge dilemma. As before, a trolley threatens to kill five people. You are standing next to a large stranger on a footbridge that spans the tracks, in between the oncoming trolley and the five people. In this scenario, the only way to save the five people is to push this stranger off the bridge, onto the tracks below. He will die if you do this, but his body will stop the trolley from reaching the others. Should one save the five others by pushing this stranger to his death? To this question, most people answer no (with regard to precise data, see Greene & Haidt, 2002).

With regard to the dominant responses in these dilemmas, people appear to be utilitarians when solving the trolley dilemma and deontologists when solving the footbridge dilemma. In the former dilemma, a person’s choice appears to depend on the number of workmen to be saved, whereas people make much of the right of the man on the bridge in the latter dilemma. Thus, it has been considered that the trolley dilemma reflects the utilitarian way of thinking, whereas the footbridge dilemma reflects the deontologist way of thinking (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Waldmann & Dieterich, 2007). With regard to this discrepancy in the dominant responses between the trolley and footbridge dilemmas, various theoretical explanations have been proposed such as the dual process theory (Greene et al., 2001), moral grammar theory (Hauser, 2007; Mikhail, 2009), or causal decision theory (Waldmann & Dieterich, 2007).

Although theorists of moral reasoning differ in terms of how to explain the discrepancy between the trolley and footbridge dilemmas, they consistently receive correspondence between the two dilemmas and philosophical way of thinking without any doubt. For example, Greene et al. (2001) said that the dominant response in the footbridge dilemma could be justified in a Kantian (deontologist) vein, but this justification has trouble when considering the trolley dilemma (Greene et al., 2001, p2106). Hauser (2006, p113–4) explained the philosophical implication of the trolley and footbridge dilemmas in terms of whether the utilitarian calculation can justify dominant responses in these dilemmas. Waldmann and Dieterich (2007) also argued that throwing the switch in the trolley dilemma is in line with the utilitarian view, whereas the footbridge dilemma is in line with the deontologist perspective (Waldmann & Dieterich, 2007, p247–8). Although there are differing theoretical explanations for the discrepancy between the trolley and footbridge dilemmas, the theorists in moral reasoning research have no doubt in

the assumption that the trolley dilemma reflects utilitarian thinking, and the footbridge dilemma reflects deontologist thinking.

However, current research proposes another interpretation of the difference between the two moral dilemmas: the footbridge dilemma reflects more utilitarian thinking than the trolley dilemma. Although this hypothesis apparently sounds strange when considering the presumption in related studies, it can be naturally derived from a theoretical explanation (Waldmann & Dieterich, 2007) and recent empirical findings (Nakamura, 2011) in moral dilemmas. In what follows, we explain this interpretation more precisely.

Waldmann and Dieterich (2007) proposed the “intervention myopia hypothesis”, which insists that moral intuitions are influenced by the locus of the intervention in the underlying causal model, and an attentional focus on the victims is highlighted by the intervention, leading to the neglect of other victims located in the background. More specifically, it treats the trolley dilemma as the intervening agent (trolley) and the footbridge dilemma as the intervening potential patient (victim). Thus, attentional focus on the one victim becomes stronger in the footbridge dilemma than in the trolley dilemma, thereby resulting in deontologist judgments being more likely in the former dilemma. Based on this theoretical explanation for moral dilemmas, Waldmann and Dieterich (2007) performed a series of experiments in which the focus of intervention was manipulated. For example, they compared how participants’ moral judgment would differ between “throwing a bomb on the man” and “throwing a man on a bomb” to save many people. In terms of their hypothesis, the former action corresponds to the agent intervention, and the latter action corresponds to the patient intervention. Results of their experiment consistently supported their hypothesis and revealed that sacrificing one victim to save more victims is more permitted in the agent intervention than in the patient intervention.

One crucial aspect of Waldmann and Dieterich’s (2007) hypothesis is that it considers the difference between the trolley and footbridge dilemma as that of attentional focus to causal structure. As stated above, according to this hypothesis, people make more of the patient when considering the footbridge dilemma compared to the trolley dilemma; they do not permit the sacrificing of a man in the footbridge dilemma as they do in the trolley dilemma. This explanation is intuitively natural and appears to match the dominant responses in these two dilemmas.

At the same time, the explanation in terms of attentional focus also leads to an interesting prediction. Many studies (e.g., Slovic, Griffin, & Tversky, 1990; Tversky & Koheler, 1994; Tversky, Sattath, & Slovic, 1988; also see Fischer & Hawkins, 1993) have demonstrated that a given attribute or element carries more weight for decision making when it becomes prominent. Although these studies are varied in their research subjects, such as preference reversal (Slovic et al., 1990; Tversky et al., 1988) or

probability judgment (Tversky & Koheler, 1994), these studies consistently assume that an attribute or element looms larger when it receives attention and its impact on judgment becomes stronger than when it does not receive attention. If so, the following prediction can be drawn from Waldmann and Dieterich’s (2007) hypothesis; *people are more sensitive to a difference in the number of victims in the footbridge dilemma than in the trolley dilemma because the victims are paid more attention in the footbridge dilemma than in the trolley dilemma*. More specifically, the effect of the manipulation of the number of victims is larger in the footbridge dilemma than in the trolley dilemma because the victims in the former dilemma loom larger than in the latter dilemma. As you see, this prediction can be derived from the existing explanation very naturally. However, it contradicts the dominant view that matches the trolley and footbridge dilemmas to utilitarian and deontologist thinking (e.g., Greene et al., 2001).

Although the above prediction appears to be paradoxical, Nakamura (2011) also demonstrates that the footbridge dilemma surely reflects utilitarian thinking more so than the trolley dilemma. He required participants to answer 62 types of moral dilemmas used in Greene et al. (2001) and analyzed the correlation structure of participants’ judgments using factor analysis and structural equation modeling. The results demonstrated that the moral dilemmas used in Greene et al. (2001) can be explained by four factors: rationality, life-dilemma, risk averse, and efficiency (see Nakamura, 2011). Among these four factors, the risk-averse factor contributed to the difference between the trolley and footbridge dilemmas. This factor mainly comprises problems similar to Asian disease problems (e.g., Tversky & Kahneman, 1984) that require participants to consider a trade-off between probability and outcome (“90% chance of causing no deaths at all and has a 10% chance of causing 1000 deaths or an 88% chance of causing no deaths and a 12% chance of causing 10 deaths”). This factor can be interpreted as the calculation of expected value for each alternative that can be thought of as a utilitarian aspect of moral dilemmas. Surprisingly, a result of structural equation modeling in Nakamura (2011) demonstrated that the risk-averse factor had a significant effect on the footbridge dilemma but not on the trolley dilemma, which is in accordance with the entailment of the intervention myopia hypothesis stated above. That is, results of the multivariate analysis that deals with the correlation structure among the moral dilemmas also indicate a relationship between utilitarian thinking and the footbridge dilemma.

The above discussion consistently suggests that the footbridge dilemma reflects utilitarian thinking more so than the trolley dilemma. Considering that previous studies, including both psychological and philosophical ones, have positioned these dilemmas as symbols for utilitarian and deontologist thinking (e.g., Foot, 1978; Greene et al., 2001; Thomson, 1985), this implication is very important because it contradicts the prevailing view of these two dilemmas. Additionally, the suggestion of the above discussion is

drawn from a natural deduction of the existing theoretical approach (Waldmann & Dieterich, 2007) that also supports this prevailing view. Thus, an exploration of the utilitarian aspect of these two dilemmas leads to a clarification of the meaning of “utilitarian” and “deontologist.”

The purpose of this study is to test the hypothesis that the footbridge dilemma is more related to utilitarian thinking than the trolley dilemma. In order to accomplish this, the current research emphasizes the number of victims. Previous studies combined the trolley and footbridge dilemmas with utilitarian and deontologist thinking in terms of whether participants make much of the number of people to be saved or one man’s right to live. In this vein, the number of victims can be considered a utilitarian aspect of the moral dilemmas. The following two studies manipulated the number of victims in both the trolley and footbridge dilemmas and examined its effect on these two dilemmas.

### Study 1

Study 1 aimed to investigate how a manipulation of the number of victims in the trolley and footbridge dilemmas would work under a standard experimental procedure. Many studies on moral dilemmas (e.g., Greene et al., 2001; Hauser, Cushman, Young, Jin, & Mikhail, 2009) have adopted a forced choice paradigm in which participants are required to choose whether sacrificing a victim to save more people would be permissible or not and have analyzed the percentage of participants who answered yes. Thus, Study 1 examined how the percentage of participants would change according to the number of victims in the trolley and footbridge dilemmas.

### Participants and design

Two hundred eighty-five undergraduates who were naïve to the dilemma tasks participated in Study 1 for course credit. We prepared six types of scenarios in which the story (trolley or footbridge) and number of victims (one/two/five) were manipulated, and each participant received one of the six types of scenarios randomly. As a result, in the trolley dilemma condition, the number of participants who were assigned to the one-, two-, and five-victim condition were 49, 49, and 53, respectively. In the footbridge dilemma condition, the number of participants was 43, 58, and 53 (one-, two-, and five-victim condition, respectively).

### Materials and procedure

Participants received a booklet and before they read the booklet, they were told that the study was about moral dilemmas. The instructions on the first page stated that the task was to read descriptions of a situation and to consider an act described in the scene. The second page presented the scenario and included a response format requiring participants to indicate whether the act (“turn the trolley”/“push the man”) was morally permissible.

The following was the first paragraph in both the trolley and footbridge dilemma conditions:

There is an emergency where a trolley runs out of control. Although a driver tries to stop the trolley, it does not appear to stop. Unfortunately, the trolley is rushing toward ten workmen. If the trolley does not stop, it will surely kill all ten workmen.

After this paragraph, participants in the trolley dilemma condition were shown the following scenario:

There is another railway of the trolley, and if you hit the switch, the trolley changes its course, and the ten workmen will be saved. However, there is #\_\_ workman (or men) on the other course, and if you hit the switch, this workman (or men) will surely be killed by the runaway trolley. Is it permissible to hit the switch to save the ten workmen?

After the first paragraph, participants in the footbridge dilemma condition were shown the following scenario:

There is a footbridge on the course of the trolley, and #\_\_ man (or men) standing on this footbridge. If you throw the man on the railway, the trolley will stop because the man’s body becomes a barrier, and the ten workmen will be saved. However, the man (or men) on the footbridge will be killed. Is it permissible to push the man on the footbridge to save the ten workmen?

The blanks shown in the above texts were replaced by numbers (one, two, or five), depending on the conditions of the number of the victims. The descriptions of these two dilemmas comprised only of sentences and did not employ any pictures. All the participants finished the tasks within 10 minutes.

### Results and discussion

Figure 1 depicts the percentage of participants who believed that sacrificing fewer to save more was “morally permissible” in each condition. As this graph demonstrates, the effect of the number of the people on moral judgment differs between the trolley and footbridge dilemmas. In the footbridge dilemma, acceptability for the death of one person to save ten people decreases as the number of lives sacrificed increases, whereas in the trolley dilemma, the percentage of participants who permitted sacrificing the few remains constant. A logistic multiple regression analysis showed that an effect of the number of victims,  $B = -0.25$ , Wald (1) = 5.47,  $p = 0.02$  and interaction between the number of victims and the type of dilemma,  $B = -0.33$ , Wald (1) = 4.33,  $p = 0.37$  were significant, thereby indicating that whether manipulation of the number of people to be sacrificed would affect moral judgment depends on the context of the dilemma. Multiple comparisons among the conditions in the number of the victims indicate significant differences in the footbridge dilemma, (chi-square test:  $P_s < .01$ ), but not in the trolley dilemma ( $P_s > .10$ ). Thus, results of Study 1 supported our prediction that the

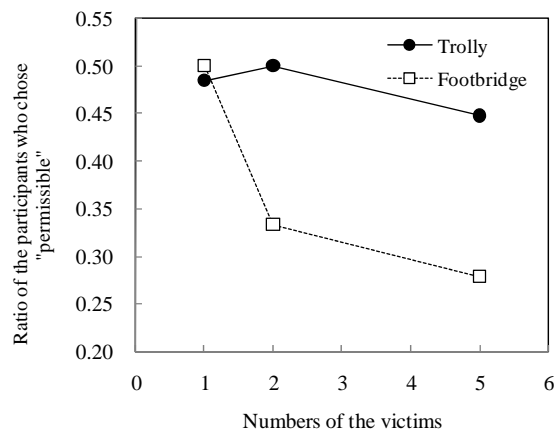


Figure 1 Results of Study 1

footbridge dilemma reflects utilitarian thinking more so than the trolley dilemma.

One noteworthy result of Study 1 is that the difference between the trolley and footbridge dilemmas is not significant when one person must be sacrificed to save ten ( $P > .10$ ). As far as I know, this is the first example demonstrating that the ratio of participants who chose “permissible” in the footbridge dilemma was equal to that of the trolley dilemma. One plausible reason for this result is that the current research differs in terms of the number of people to be saved and the number of victims. Most previous studies on the trolley and footbridge dilemmas (e.g., Greene et al., 2001; Mikhail, 2009; Waldmann & Dieterich, 2007) used five and one as the number of people to be saved and number of victims, respectively. In contrast to these studies, in order to examine the utilitarian aspect of the moral dilemma, the current study used ten as the number of people to be saved, and three values (one, two, or five) for the number of victims. It is possible that the number of victims affect the difference between the trolley and footbridge dilemmas, although it requires further study. Study 2 will also address this problem in the results and discussion section.

## Study 2

Study 2 aimed to replicate the findings in Study 1 under a condition where the following two modifications in the experimental procedure were added. First, Study 2 manipulated the number of victims as a within factor to examine whether the results in Study 1 were due to a reflection of individual differences. Second, Study 2 adopted the number of people to be saved as a dependent variable. Although the current research aimed to examine the utilitarian aspects of the moral dilemmas, the permissibility judgment used in Study 1 appears to be somewhat different from utilitarian calculation because “permissibility” sounds like a subjective impression. Thus, it is not certain whether the same trends would be found if participants are required to make a utilitarian calculation for

permissibility to sacrifice few to save more. In order to examine this, Study 2 required participants to estimate the number of people to be saved that seemed to be equal to sacrificing fewer people.

## Participants and design

Fifty undergraduates participated in Study 2 for course credit and were randomly assigned to one of two conditions: the trolley dilemma condition or the footbridge dilemma condition. All the experimental materials and response formats were given in the form of a booklet. The first page contained the same instructions as Study 1, and the second page described the scenario and included a response format. In both the trolley dilemma conditions, participants read the following:

There is an emergency where a trolley runs out of control. Although a driver tries to stop the trolley, it does not appear to stop. Unfortunately, the trolley is rushing toward some workmen. If the trolley does not stop, it will surely kill all the workmen.

There is another railway for the trolley, and if you hit the switch, the trolley changes its course, and the ten workmen will be saved. However, there are *some* workmen on the other course, and if you hit the switch, the workmen on the other railway will surely be killed by the runaway trolley.

As shown, these instructions are almost the same as those of Study 1, except that the number of victims and people to be saved were not stated explicitly. In the footbridge condition, the first paragraph shown to participants was the same as that in the trolley dilemma condition, but the second paragraph was as given below:

There is a footbridge on the path of the trolley, and some men are standing on this footbridge. If you throw the men on the railway, the trolley will stop because the men’s bodies become a barrier, and the workmen will be saved. However, the men on the footbridge will be killed. Is it permissible to push the men on the footbridge to save the ten workmen?

After reading the above texts, participants in both conditions were required to answer the following question under three conditions, where the number of victims were one, two, or five:

Consider that the number of workmen on the other railway (the footbridge) is #\_\_. How many people do you think are enough to justify hitting the switch (pushing them)? Insert a number in the blank.

The blank shown in the response format was replaced by numbers (one, two, or five). All participants finished within 20 minutes.

## Results and discussion

The results of Study 2 depicted in Figure 2 indicate that the slope of the function between the number of victims and the dependent variable is steeper in the footbridge condition than in the trolley condition. A 2 (type of dilemma: trolley/footbridge) by 3 (number of people to be sacrificed: 1/2/5) ANOVA demonstrated significant main effects of the two factors: type of dilemma,  $F(1, 48) = 7.22, p < .01$ ; number of victims,  $F(2, 47) = 16.57, p < .01$ ; and interaction  $F(2, 96) = 6.60, p < .01$ . Analyses of the simple main effect by Ryan's method indicated that the simple effect of the type of dilemma was significant only in the five-person condition,  $F(1, 144) = 18.68, p < .01$ , and the simple main effect of the number of people was significant only in the footbridge condition,  $F(2, 96) = 22.05, p < .01$ . These results replicate the findings of Study 1, which suggest that the effect of the number of victims is stronger in the footbridge dilemma than in the trolley dilemma, thereby supporting our position that the footbridge dilemma reflects the utilitarian aspect of the moral dilemma more than the trolley dilemma. Additionally, Study 2 also failed to find the difference between the trolley and footbridge dilemmas when the number of victims is small. These results replicate the pattern found in Study 1.

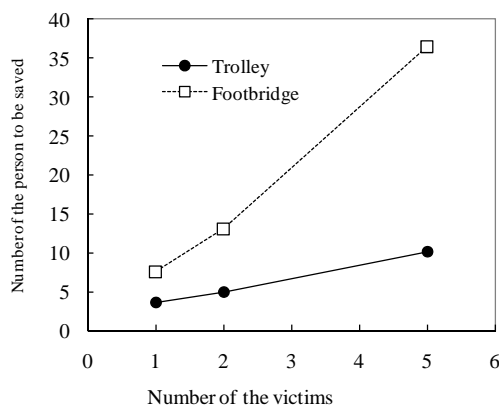


Figure 2 Results of Study 2

## General discussion

The results of the two studies consistently demonstrated that the footbridge dilemma was more sensitive to the manipulation of the number of people to be sacrificed than the trolley dilemma. Study 1 demonstrated that as the number of victims increased, the ratio of participants who permitted sacrificing the few in the footbridge dilemma decreased, whereas those in the trolley dilemma did not change. Study 2 asked participants the number of people to be saved to permit a sacrifice and found that the effect of the number of victims was larger in the footbridge dilemma than in the trolley dilemma.

The current results provide the following four theoretical implications. First, this article demonstrates a connection

between the footbridge dilemma and the utilitarian aspect of the moral dilemma. Previous studies (e.g., Greene et al., 2001; Hauser, 2007; Waldmann & Dieterich, 2007) have assumed that the trolley dilemma reflects utilitarian thinking, whereas the footbridge dilemma reflects deontologist thinking. These studies mainly draw this assumption from a pattern of the dominant responses of the trolley and footbridge dilemmas. In contrast to these studies, the current research took into account sensitivity to the manipulation of the utilitarian aspect of the moral dilemma and derived a contradictory view. As far as we know, this is the first example that challenges the prevailing view: “utilitarian” trolley and “deontologist” footbridge. In addition, the current results reveal the reason why people consider pushing the man not permissible in order to save the five workmen in the original footbridge dilemma. It is not because they think the man's right to live should not be violated; rather, people think that five people are not enough to sacrifice one person.

Second, the current results are in line with the perspective that causal structure might be key to understanding moral reasoning (Waldmann & Dieterich, 2007). According to Waldmann and Dieterich's (2007) view, intervention to causal path in moral dilemmas plays an important role for moral reasoning because it changes attentional weights on agent and patient. The hypothesis tested in the current study is naturally derived from this explanation because attention is believed to affect sensitivity to the attribute that it focuses on (e.g., Tversky & Koheler, 1994; Tversky et al., 1988). The results of the two studies consistently support the hypothesis, thus indirectly confirming Waldmann and Dieterich's (2007) proposition. Additionally, the current results also support Nakamura's (2011) implication that the footbridge dilemma is considered more consequential than the trolley dilemma because the risk-averse factor solely affected the footbridge dilemma.

Third, the current research also suggests that moral reasoning processes may be easily influenced by the number of victims. Both Studies 1 and 2 failed to find a difference between the trolley and footbridge dilemmas that have been replicated robustly (e.g., Greene et al., 2001; Greene & Haidt, 2002). The main difference between previous studies and the current study is the number of victims and people to be saved. With regard to the number of people to be saved by sacrificing fewer people, this research used “ten” in Study 1, whereas previous studies used “five.” Study 2 left the number of people to be saved blank and required participants to provide a number that they would be willing to sacrifice. In this vein, there is a possibility that this difference in the numbers used in the scenario might produce a discrepancy in the results between the current research and previous studies. Although this possibility is an issue for future examination, it would be useful to explore the relationship between numerical value and moral judgment.

Fourth, the current research indicates the importance of exploring not only the dominant response but also sensitivity to manipulation when investigating the moral dilemma. The current research focused on the effect of the number of victims and succeeded to derive a somewhat different conclusion by identifying an interaction between the number of victims and types of dilemma. This result might provide an important implication to a methodology of experimental philosophy (e.g., Knobe, 2004, 2007; Knobe & Nichols, 2009). Experimental philosophy attempts to solve philosophical issues not by speculation but by empirical investigation. In doing so, experimental philosophy mainly deals with average responses in moral reasoning problems, as previous studies on moral dilemmas have done (e.g., Knobe, 2003). However, empirical data are not limited to the average. Correlation among the problems (e.g., Nakamura, 2011) or sensitivity to independent variables can also provide interesting information in understanding the nature of a moral issue. Concern for the data analysis method would benefit experimental philosophy and produce results that are applicable to philosophical issues.

This discussion also leads to an examination of “utilitarian” and “deontologist” thinking. The proposition that the trolley dilemma reflects the utilitarian thinking has its basis on the dominant responses to this dilemma, whereas the current research has its basis on examination to the sensitivity to the number of the victims in moral reasoning. Then, some reader may consider a following question; which is the more plausible evidence to determine the utilitarian thinking? I think this question is a fundamental one for experimental philosophy. That is, one more message of the current research is that empirical studies on the moral reasoning should consider not only the meaning of the moral dilemma but also *how* to interpret empirical evidence in the dilemma. As far as I know, there is no study that considers this problem, and I hope this research would be a first step to this problem.

## References

- Bentham, J. (1948/1789). *An Introduction to the Principles of Morals and Legislation*. Halfner Press, New York.
- Fischer, G. W., & Hawkins, S. A. (1993). Strategy compatibility, scale compatibility, and the prominence effect. *Journal of Experimental Psychology: Human Perception and Performance*, 19, 580–597.
- Foot, P. (1978). *The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices*. Oxford: Basil Blackwell.
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293, 2105–2108.
- Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Science*, 12, 571–523.
- Hauser, M. (2006). *Moral Minds: How Nature Designed a Universal Sense of Right and Wrong*. Harper Collins/Ecco, NY.
- Hauser, M., Cushman, F., Young, L., Jin, R. K., & Mikhail, J. (2009). A dissociation between moral judgment and its justification. *Mind & Language*.
- Kant, I. (1959). *Foundation of the Metaphysics of Morals*. (Lewis White Beck, Trans.) Indianapolis: Bobbs-Merrill. (Original work published 1785).
- Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, 16, 309–324.
- Knobe, J. (2004). What is experimental philosophy? *The Philosophers' Magazine*, 28.
- Knobe, J. (2007). Experimental philosophy and philosophical significance, *Philosophical Explorations*, 10, 119–122.
- Knobe, J., & Nichols, S. (2008). *Experimental Philosophy*. Oxford University Press, USA.
- Mikhail, J. (2009). Moral grammar and intuitive jurisprudence: A formal model of unconscious moral and legal knowledge. *Psychology of Learning and Motivation*, 50, 27–100.
- Nakamura, K. (2011). A closer look at the moral dilemma: Exploration of the latent structure and meaning of “emotional” and “rational.” *Proceedings of the Thirty-third Annual Conference of the Cognitive Science Society*, 1084–1089.
- Singer, P. (1979). *Practical Ethics*. Cambridge University Press.
- Slovic, P., Griffin, D., & Tversky, A. (1990). Compatibility effects in to pay for public goods. In R. M. Hogarth (Ed.), *Insights in Decision Making: Theory and Applications*. Chicago: Univ. of Chicago.
- Thomson, J. J. (1985). The trolley problem. In J.M. Fischer & M. Ravizza (Eds.), *Ethics: Problems and Principles* (pp. 67–76). Harcourt Brace Jovanovich, Fort Worth, TX.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453–458.
- Tversky, A., & Koheler, D. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101, 547–567.
- Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review*, 95, 371–384.
- Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, 18, 247–253.



# Anticipating changes: Adaptation and extrapolation in category learning

Daniel J. Navarro ([daniel.navarro@adelaide.edu.au](mailto:daniel.navarro@adelaide.edu.au))

School of Psychology, University of Adelaide, SA 5005, Australia

Amy Perfors ([amy.perfors@adelaide.edu.au](mailto:amy.perfors@adelaide.edu.au))

School of Psychology, University of Adelaide, SA 5005, Australia

## Abstract

Our world is a dynamic one: many kinds of objects and events change markedly over time. Despite this, most theories about concepts and categories are either insensitive to time-based variation, or treat people's sensitivity to change as a result of process-level characteristics (like memory limits, captured by weighting more recent items more highly) that produce irrational order effects during learning. In this paper we use two experiments and nine computational models to explore how people learn in a changing environment. We find, first, that people adapt to change during a category learning task; and, second, that this adaptation stems not only from weighting more recent items more highly, but also from forming sensible anticipations about the nature of the change.

**Keywords:** categorization, change detection, concepts, dynamics, time dependence, order effects

At no two moments in time are we presented with the same world. Objects move, plants and animals are born and die, friends come and go, the sun rises and sets, and so on. More abstractly, while some of the rules that describe our world – like physical laws – are invariant over the course of our everyday experience, others – like legal rules – are not. Given some appropriate time scale, certain characteristics of an entity or class of entities can change; moreover, they may tend to change in *systematic* ways. Temperature varies cyclically as a function of time of day and time of year; with a gradual rise over the last century. As another example, the features that describe phones have changed considerably over recent decades: not only do modern phones perform many new functions, they are also physically smaller, sleeker, and smoother. People are aware of this temporal structure and adapt their predictions to it: if asked to predict the temperature six months from today, people will give quite different answers than if asked to predict the temperature tomorrow. We do not modify predictions in an *ad hoc* or senseless fashion; instead, we appear to be attuned to particular details of the nature of the dynamic variation in category structure.

One consequence of recognizing the changeable nature of categories is that the time at which observations are made matters. A machine with a rotary dial could very well be a typical phone if it were observed in 1980, but this is much less plausible if the observation dates from 2010. These changes over time introduce strong *order effects* into the classification problem. Order effects in categorization are well-studied, but people's sensitivity to order is generally thought to result from weighting more recent items more highly (generally as a result of poor memory) or inefficient learning rules (e.g., Kruschke, 2006; Sakamoto, Jones, & Love, 2008). Irrespective

of whether these process limitations arise from the use of *ad hoc* (Anderson, 1990) or rationally motivated (Sanborn, Griffiths, & Navarro, 2010) computation strategies, the shared assumption is that people should *not* be sensitive to order information when learning new categories. While this is certainly true in some cases, in other situations order sensitivity might actually be a sensible adaptation. Especially when a category is changing in systematic ways (e.g., phones are getting steadily more complex), a sensible learner should be able to extrapolate about the future in a way that is sensitive to that systematicity. The central question of this paper is whether people can perform this extrapolation, and whether it can be separated from an order sensitivity that arises from simply weighting more recent items more highly.

The structure of this paper is as follows. The first half explores the advantages of being sensitive to order in category learning. This is important because the literature tends to focus on the negative consequences of order sensitivity – namely, irrational order effects when the environment is in fact stationary. However, we present computational modeling highlighting the fact that when the world is actually changing, being sensitive to that change constitutes an enormous advantage. In the second part of the paper, we investigate the nature of human order sensitivity. To what extent does it reflect imperfect memory, and to what extent does it reflect sensible adaptation and extrapolation about systematic change? Experimental and computational results<sup>1</sup> demonstrate that both factors play an important role.

## The importance of order sensitivity

We begin our exploration into the importance of order sensitivity with a prototype model for categorization, which in its standard form is not sensitive to order. In this model, people are assumed to construct a single representation of the prototype, or the central tendency of the category. In its most usual form, if the learner has seen a series of  $t$  items, then the prototype  $\mu_t$  is found by taking the arithmetic mean of the mental representations of the category members:

$$\mu_t = \frac{1}{t} \sum_{i=1}^t x_i \quad (1)$$

where  $x_i$  denotes the mental representation (e.g., co-ordinates in psychological space) of the  $i$ -th category member. If the

<sup>1</sup>Because the paper involves a large number of models (9 in total), we outline the structure of the models only in general terms here: precise specifications are available in the supplementary materials found at [www.compcogscilab.com/dan/](http://www.compcogscilab.com/dan/).

distribution of the category members is normal, then the learner is assumed to keep track the standard deviation of the  $t$  category members around the prototype  $\sigma_t$ . The probability that a new item belongs to the category is proportional to the probability that the category distribution assigns to that item. As noted by several authors (e.g. Ashby & Alfonso-Reese, 1995) this model has an interpretation as a rational model when the category members satisfy certain assumptions – including the assumption that the world does not change.

It is precisely because the world is assumed not to change that all category members are assigned equal weight in Equation 1, regardless of when they were observed. As a consequence, the model predictions on trial  $t$  do not depend on the order in which the preceding items were observed. This order invariance is characteristic of most probabilistic models for category learning, whether they be prototype, exemplar, or cluster-based (see Griffiths, Sanborn, Canini, & Navarro, 2008). As noted by a number of authors (e.g. Sakamoto et al., 2008; Navarro & Perfors, 2009) such models perform poorly as models of human performance when the data presented to people contains strong order effects. To address this, it is commonplace to consider an updating rule that is more sensitive to the changing structure of the data. The general idea is to assume that when the learner encounters a new category member, he or she moves the prototype a little closer to the new category member. This has the effect of weighting new items more than older items. The simplest version of this rule is as follows: if  $x_t$  refers to the feature value(s) for the new item, then the new prototype location at time  $t + 1$  is

$$\mu_t = \phi\mu_{t-1} + (1 - \phi)x_t. \quad (2)$$

One advantage of this rule is that it is simple and easy to implement. Viewed from a computational perspective, it corresponds to an “exponential weighting” scheme, in which the most recent observations are weighted most highly, where the weight decays exponentially as a function of time.<sup>2</sup> In that sense, it is quite similar to the exponential generalization gradients that are typically used in category learning models (Nosofsky, 1984), but applied across time rather than psychological space. Additionally, this weighting function is similar to the shape of the retention function in memory research.<sup>3</sup>

The key feature of this and similar rules is that the learner is constantly adapting the prototype, never converging to a single true value. Because of this, the model is quite sensitive to the recent trends in the data. Learning rules of this form capture human order effects in simple choice tasks (e.g., Yu & Cohen, 2009; Gökyaydin, Ma-Wyatt, Navarro, & Perfors, 2011). Moreover, such rules make sense as a near-optimal inference scheme when there is a constant probability that “something changes”, but the nature of the change is not predictable or systematic (e.g., Yu & Cohen, 2009; Brown & Steyvers, 2009). In the next section, we illustrate how important this sensitivity can be.

<sup>2</sup>The exact equation is  $\mu_t = (1 - e^{-\tau}) \sum_{i=1}^t e^{-\tau(t-i)} x_i + e^{-\tau} \mu_0$ , where  $\tau = -\ln \phi$ , and  $\mu_0$  is the initial location of the prototype. The derivation is included in the supplementary materials, but since Equation 2 is the simpler and more interpretable version, we use it throughout the paper rather than this more explicit exponential form.

<sup>3</sup>Although retention curves are generally have heavier tails than exponentials (e.g. Rubin, Hinton, & Wenzel, 1999), this would likely make little difference to a typical category learning experiment.

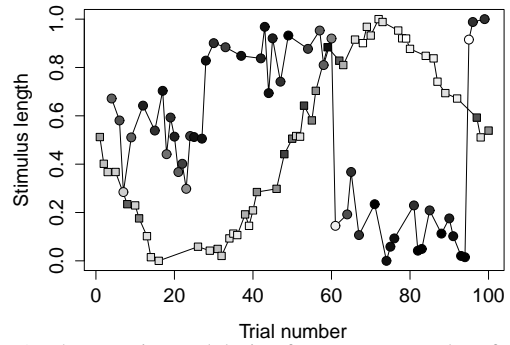


Figure 1: The experimental design from Navarro and Perfors (2009). Stimuli varied along a single dimension (line length). Circles denote items belonging to one category, and squares refer to the other category. The two categories consist of the exact same set of 100 stimuli, but are presented in a different order. The shading summarises the human responses: the darker the shade, the more likely people were to select the “circle” category (see Figure 2 for histograms showing the probability of a correct response). Note the large “jump” around trial 60, and the fact that humans quickly adapt to it.

## Modelling adaptation to change

The typical view of “rationality” in a category learning experiment assumes that the world is stationary. Under this assumption, any sensitivity to the order in which items are observed is irrational. However, this view is sometimes expressed even when using experimental designs that *violate* the stationarity assumption. A good example of this is the paper by Sakamoto et al. (2008), which presented results from a simple supervised categorization task in which one category possessed a strong time dependency. In this task, an order sensitive model accounted for human classification decisions better than an order insensitive one. Despite the fact that is sensible to be sensitive to order such a situation, this was argued to be evidence *against* the rationality of human behavior rather than evidence in favor of it.

To explore the extent to which people are sensitive to categories that change over time, Navarro and Perfors (2009) conducted a categorization experiment in which the *only* information that distinguished the categories was the order in which items were presented (the design is illustrated in Figure 1). Human performance was well above chance: participants correctly classified the stimuli on 76% of trials. However, models were not fitted to this data, making it somewhat unclear exactly how badly an order insensitive model would perform on the task, and the extent to which order sensitive models would improve matters.

To address these question, here we fit<sup>4</sup> six different models to that data. Three of the models are not sensitive to order. This includes two variations of a standard prototype model, one assuming that both category distributions have equal variance (i.e., the same value of  $\sigma$ ) and the other allowing unequal variances, and a standard exemplar model (Nosofsky, 1984). The other three models are analogous, but each use an exponential weighting scheme to assign more importance to recent

<sup>4</sup>Throughout the paper we use the ordinary least-squares (OLS) method to estimate model parameters: that is, we minimize the sum squared deviation between model predictions and human classifications, though we exclude the first 10 trials since human and model predictions are both quite variable initially.

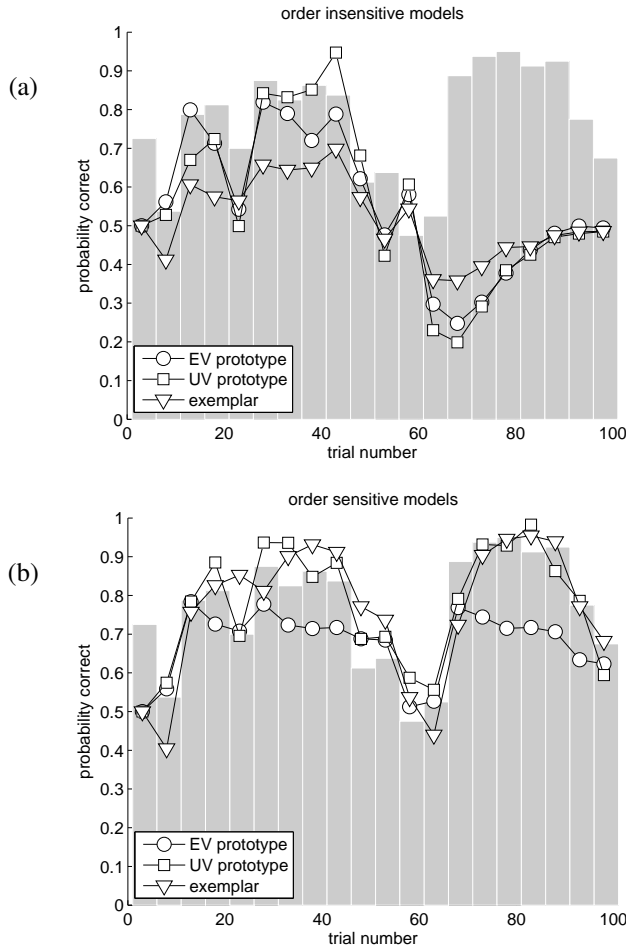


Figure 2: Graphical depiction of the performance of prototype (EV = equal variance, UV = unequal variance) and exemplar models on the classification task, for both the standard order-insensitive versions (panel a) and the order-sensitive versions (panel b). The grey bars plot human performance, and for reasons of visual clarity all plots are averaged over 5 trial blocks. Notice the catastrophic prediction failure for the order insensitive models during the second half of the experiment.

items: this is equivalent to Equation 2 for the prototype models, and is slightly more complex for the exemplar model. The formal details for all models are included in the supplementary materials. None of the models have more than two free parameters in this task.

The important finding here, shown in Figure 2a, is that none of the order insensitive models can mimic human behavior, nor can they achieve good classification performance on the data. No order insensitive model can correctly classify the stimuli at a rate higher than 55%, regardless of what parameter values are chosen. Moreover, none of the models can account for more than 5% of the variance in human responses.<sup>5</sup> The basic problem is evident in Figure 2a: when the task changes at trial 60, no order insensitive model can

<sup>5</sup>We measure this using the Variance Accounted For (VAF) statistic, which is equivalent to an  $R^2$  statistic in regression, but does not estimate an intercept and slope: the model predictions are compared *directly* to human performance.

adapt to it in the way humans do. Viewed as *psychological* models, the inability to describe human behavior is a serious issue. In addition, the inability to achieve an acceptable classification performance makes it clear that none of these models provides a good *rational* standard for the task either.

In stark contrast, all three of the order sensitive models perform reasonably well, as shown in Figure 2b. When fit to the human responses, the exemplar model and unequal-variance prototype model are nearly indistinguishable, accounting for 79% and 80% of the variance respectively. The equal variance prototype model performs slightly worse, accounting for 72% of the variance. This is not surprising: as Figure 1 illustrates, the categories are not equally variable. In any case, the important point is that all three order sensitive models are massively superior to any of the order insensitive models.

This improvement in psychological performance is matched by an improved classification rate. If the parameters are chosen to maximize the classification accuracy rather than to maximize the variance accounted for in human performance, the exemplar model and unequal variance prototype models can both exceed human performance (89% and 82% accuracy), while the equal variance prototype model is slightly worse (68% compared to the 76% of humans). While it would be incorrect to suggest that any of these models provides an exact normative standard for this task, since the data have a rather more complex structure than a simple exponential weighting mechanism suggests, it is clear from the superior classification performance that the order sensitive models are much closer to the required standard than the usual order insensitive models.

We have shown so far that order sensitivity is important in understanding human category learning: standard order insensitive models fail catastrophically when the world changes, both as pure classifiers as well as models of human classification. By contrast, models that use a simple exponential weighting method do surprisingly well in capturing human order sensitivity. But is that the *entire* explanation for how humans respond to time dependence in the input? One possibility is that it is, and that people change over time because they have poor memories and are most likely to recall and use more recent information. The other possibility is that when the changes are systematic, people actively notice and adapt to them, allowing them to extrapolate in a sensible way when making predictions about new items. In the rest of this paper we describe experimental and computational work showing that both possibilities appear to be true: human performance is affected by memory limitations as well as the ability to adapt in a sensible way.

## Learning to predict changes

The central question going forward is whether people's sensitivity to change is the result of memory limitations or to a sensible adaptation to a changing world. To some extent, these explanations are not in conflict: memory is sensitive to the need probability of the information to be recalled, and recent information is more likely to be useful than older information when the world changes (Anderson & Schooler, 1991). Nevertheless, it is possible to make distinctions between the two. For instance, when the world changes in *systematic* ways, it is not usually an optimal strategy to just reweight old informa-

Table 1: An overview of the five models used in the experiment. All models are equal-variance prototype models, but differ in the manner in which the category prototype is estimated. The first column gives the label for the model, the second column indicates whether the model used exponential weighting or not, and the third column indicates what method the model used to anticipate what future changes would occur.

Model	Weighting	Extrapolation
STANDARD	constant	none
RECENCY	exponential	none
RECENCY+BIAS	exponential	constant
REGRESSION	constant	regression
RECENCY+REGRESSION	exponential	regression

tion. Instead, the right strategy is to learn to anticipate those changes, and guess how the world is going to change before the change actually happens. In this section we present a category learning experiment in which people were presented with fairly obvious and systematic temporal changes, and investigate how well people learned to anticipate those changes.

## Method

**Participants** 59 participants were recruited through a mailing list whose members consist primarily of current and former undergraduate psychology students, and paid \$10/hour for their time. The median age was 23, and the participants were predominantly (63%) female.

**Materials & Procedure** The learning task was a standard supervised classification experiment, performed on a computer. Stimuli were little cartoon objects (“floaters”), which were displayed floating above a horizontal line (“the ground”). The height of the floater was the only respect in which the stimuli varied from each other.

On each of 100 trials participants were shown a single floater and asked to predict whether it would flash red or blue. After making their prediction, they would receive feedback for two seconds while the floater flashed the appropriate color. As Figure 3 illustrates, as the experiment progressed, all the stimuli shown to people tended to rise, regardless of which category they belonged to. In the figure, circles correspond to items that belong to the “high” category and crosses correspond to times that belong to the “low” category. Assignment of flash color (red or blue) to category (high or low) was randomized across participants.

The classification rule was such that, if  $x_t$  denotes the height of the stimulus on trial  $t$ , then the optimal response is to select the response option corresponding to the high category if  $x_t > t$ . Such a rule, shown as the solid line in Figure 3, achieves 100% accuracy on the task. However, because most stimuli tend to lie quite close to the classification boundary, the task is relatively difficult even though the general trend is clear. Consistent with this, participants during informal discussions indicated that they detected the upward trend early in the experiment, but still found the task to be quite challenging.

**Models** We fit five models, outlined in Table 1, to the experimental data. For this task a prototype model provides a good description of the category representation, and since the two categories are in fact equal variance we can use the equal

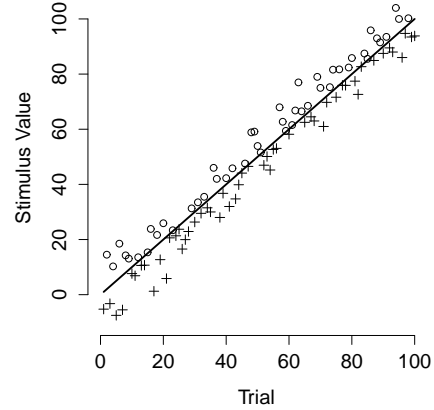


Figure 3: The experimental design. Circles denote stimuli belonging to the “high” category and crosses denote stimuli belonging to the “low” category. Although the classification rule changes over time – that is, the classification boundary is constantly rising – it does so in a regular fashion. The scale on the vertical axis is normalized so that the average “rise” from one trial to the next is 1 unit; on screen, this corresponded to an average rise of approximately 2mm per trial. Note that although it is logically possible to correctly classify all items, in practice the task is quite difficult, since most stimuli lie close to the boundary.

variance prototype model as our basic model.<sup>6</sup> The only differences among the five models lies in how they calculate the location of the prototype  $\mu_t$ ; we systematically vary effects of memory (i.e., how previous items are weighted) as well as anticipation (i.e., what extrapolation method, if any, is used to anticipate future items).

The first two models, STANDARD and RECENCY, do no extrapolation at all: they are identical to the two equal-variance prototype models from in the previous section. The STANDARD model is the order-insensitive model with constant weights as in Equation 1 whereas the recency model captures order sensitivity via the exponential weighting scheme in Equation 2. A third model, which we call RECENCY+BIAS, incorporates this exponential weighting scheme, but also incorporates a component that captures participant adaptation to the experimental design. It does so by shifting the prototype upwards by a constant amount  $\beta$ . In other words, this model not only *follows* where the data are moving (using the exponential weighting scheme), but it *extrapolates* to guess where the data are moving to.<sup>7</sup>

The final two models explore the extent to which people use a different and more optimal extrapolation method. These models estimate an explicit linear regression model for each category prototype, using the trial number as the predictor (i.e.,  $\mu_t = at + b$ ), re-estimating the regression equation after each trial. This is not quite an ideal observer model for the task, but it is extremely close: if we were to allow it to respond deterministically (via the Luce (1963) choice rule, for

<sup>6</sup>We expect that the same pattern would be observed if we used exemplar models or unequal variance prototype models; nothing in the design is specific to the equal variance prototype model.

<sup>7</sup>Note that this is not quite identical to the usual manner in which bias parameters are used in category learning models. The bias in this case alters the underlying category representation (the prototype) rather than the response bias, although in this experimental design the effect is not dissimilar.

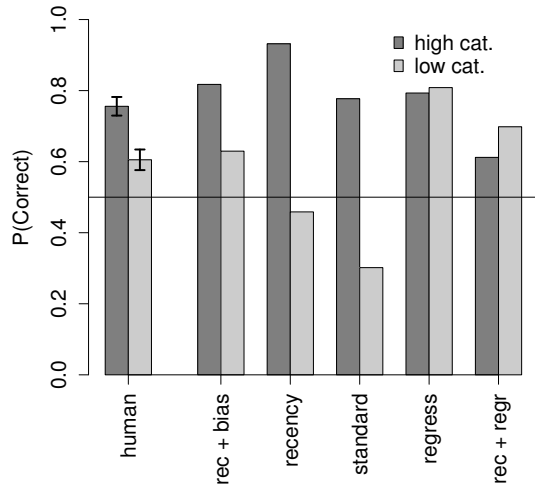


Figure 4: Overall performance for humans and models. The qualitatively important characteristics are that humans perform above chance on both categories, but perform better for the high category. This qualitative pattern is captured only by the RECENCY + BIAS model, which extrapolates about the future but also weights more recent items more highly.

instance), it would be able to perfectly classify all items. The fourth model, REGRESSION, uses exactly this linear regression model, while the fifth (RECENCY+REGRESSION) adds an exponential weighting scheme in order to attach more importance to recent observations when estimating the regression equation. In effect, this model assumes that the learner applies a nonparametric regression model (in this case lowess regression; Cleveland & Devlin, 1988) to estimate the location of the prototype, rather than a linear regression model.

The formal specification for all models can be found in the supplementary materials. For now it is sufficient to note that the only free parameters in the models are the  $\phi$  parameter that governs the amount of weighting for the three models that use exponential weighting (Equation 2) and an additional bias parameter  $\beta$  for the RECENCY+BIAS model.<sup>8</sup> The category variances are not free parameters, as these are estimated by the models, though somewhat differently in each case.<sup>9</sup>

## Results

Human performance was significantly above chance for both categories: 76% of the “high” category items and 61% of the “low” category items were classified correctly. Using

<sup>8</sup>Because it turns out that the model with the best performance is also the one with the most parameters, we checked that model complexity was not the source of its superior performance. This was done by calculating the Bayes factors for all models, using a uniform prior over  $\phi$  (which varies between 0 and 1) for the three models that use exponential weighting and a uniform prior over  $\beta$  across the range of 0 to 10. The RECENCY+BIAS model was easily the preferred model: a numerically estimated Bayes factor produced an odds ratio of about  $10^7$  to 1 favoring it over the next best model.

<sup>9</sup>Note that including this as a free parameter would make no qualitative difference to the model predictions: all of the characteristics of the data that we focus on depend only on the classification boundary of the model, which does not depend on the variance at all.

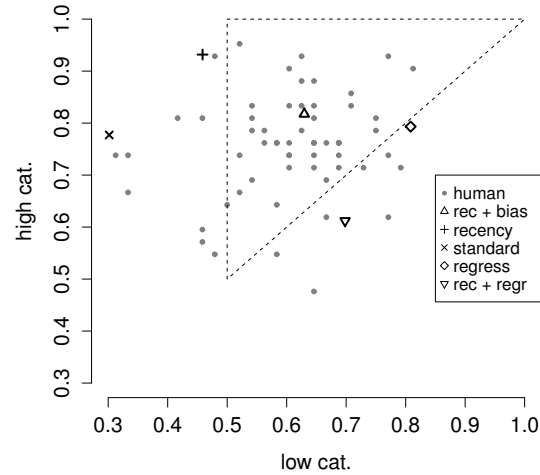


Figure 5: Individual participant data. Each dot represents one participant: the horizontal co-ordinate shows their classification performance on items that belonged to the lower category, and the vertical co-ordinate plots the performance on the higher category. The other markers show the model predictions, as indicated in the legend. The dotted triangular region corresponds to all possible patterns in which classification performance is above chance for both categories, but with the high category showing better performance. The majority of human participants fall within the triangle, as does the RECENCY+BIAS model.

each participant as a distinct data point, a single sample  $t$ -test against chance performance of 50% correct produced significant results for both the high category ( $t_{58} = 19.6$ ,  $p < .001$ ) and the low one ( $t_{58} = 7.3$ ,  $p < .001$ ). The difference between the two categories was also significant: a paired samples  $t$ -test showed that people perform significantly better on the high category than the low one ( $t_{58} = 8.5$ ,  $p < .001$ ). This is illustrated in Figure 4, which also shows the performance of all five models. Only the RECENCY+BIAS model reproduces the human pattern. Of the other four, two of them perform below chance for the low category, and the other two are unable to correctly capture the asymmetry between the categories.

This effect is even more obvious when we plot the data from the individual subjects. Figure 5 shows the classification performance for each of the participants for both the low category (horizontal axis) and the high category (vertical axis). Most human participants fall within the triangular region, in which the high category is classified better but performance is above chance for both categories. Only the RECENCY+BIAS model falls into the same region.

In addition to looking at the data at an individual subjects level, it is useful to examine model performance at an individual stimulus level. That is, how well does each model capture the trial-by-trial changes in performance? The RECENCY+BIAS model captures 39% of the variance at the item level, compared to the 13% of the variance for the RECENCY+REGRESSION model. No other model can explain any of the variability in the human data at the item level.

Finally, since the quantitative evidence for the RECENCY+BIAS model is so strong, it is important to see if the parameter estimates are sensible. The estimated

recency parameter was  $\phi = 0.25$ , meaning that participants weighted each new observation  $x_t$  about one-quarter as strongly as the old prototype  $\mu_{t-1}$  when updating the prototype. The bias parameter of  $\beta = 4.41$  appears to be a substantial upward shift, but as Figure 4 illustrates, it is not quite large enough to overcome the “lag” induced by a reliance on old observations (as per the simpler RECENCY model). In other words, this model still lags a little behind the data.

## Discussion

The experiments and computational modelling in this paper provide evidence that people adapt to changing categories in the world. This adaptation occurs not only because people weight more recent items more highly, but also because they form anticipations about the nature of the change. This work is somewhat preliminary, however, insofar as there are a number of extensions that are important to pursue. Most notable among these are the need to consider other patterns of change besides simple linear trends, and the need to extend the modelling framework to incorporate “relative judgment” approaches to categorization (Stewart, Brown, & Chater, 2002) which encode stimuli in terms of their relationship to recent items. We are currently pursuing both of these avenues.

Nevertheless, the results are interesting even without these extensions. The fact that world changes with time, and our conceptual representations must change with it, carries a number of important corollaries. In situations where the world is not constant over time, then models of category learning that assume a stationary environment are not sensible choices as rational standards for behavior. Indeed, such models do not turn out to be good models for human behavior. This does not invalidate them as good models in the contexts for which they were originally developed, but it sharply limits the generalizations one should draw from them. It is incorrect to conclude that a “rational” model that relies on an assumption of stationarity will apply to a world that changes over time and still remain a rational model in that context. Humans adapt to change, and as illustrated in the second part of the paper, if the structure of the temporal change is regular enough, we can and do learn to anticipate those changes before they occur.

That said, the best model for human performance in the task is not the ideal observer model (REGRESSION). It is instead the one that takes the simple prototype model with recency weighting (Equation 2) and makes the smallest change required to be suitable to the task. The performance of this model suggests that the adaptations that people make to anticipate systemic changes are conservative: they make smaller adjustments than the data warrant, strictly speaking (Phillips & Edwards, 1966). Conservatism itself may be a sensible adaptation (Navon, 1978), but this is not in our view the major point. The key goal is to learn what kinds of changes people are able to detect and anticipate, and how that helps us adapt to a changeable world.

Viewed from this perspective, it is notable that we were led to these results by considering the underlying inference problem that the learner has to solve: category learning in a dynamic world. In that sense, our approach offers a computational analysis, and the models that had an ex-

trapolative component (RECENCY+BIAS, REGRESSION, and RECENCY+REGRESSION) are all inspired by that observation. However, our approach also has much in common with process-level analyses that take seriously the possibility that people must operate within the constraints imposed by memory limitations. Our work helps to bridge the two levels, bringing us closer to a united explanation of how people learn categories within a changing world.

## Acknowledgements

This research was supported by ARC grant DP110104949. Salary support for DJN was provided by ARC grant DP0773794.

## References

- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2, 396-408.
- Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39, 216-233.
- Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, 58, 49-67.
- Cleveland, W. S., & Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596-610.
- Gökaydin, D., Ma-Wyatt, A., Navarro, D. J., & Perfors, A. (2011). Humans use different statistics for sequence analysis depending on the task. In *Proceedings of the 33rd annual conference of the cognitive science society* (p. 543-548). Austin, TX: Cognitive Science Society.
- Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (p. 303-328). Oxford: Oxford University Press.
- Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, 113, 677-699.
- Luce, R. D. (1963). Detection and recognition. In *Handbook of mathematical psychology* (p. 103-189). New York: Wiley.
- Navarro, D. J., & Perfors, A. (2009). Learning time-varying categories. In *Proceedings of the 31st annual conference of the cognitive science society* (p. 412-424). Austin, TX: Cognitive Science Society.
- Navon, D. (1978). The importance of being conservative: Some reflections on human Bayesian behavior. *British Journal of Mathematical and Statistical Psychology*, 31, 33-48.
- Nosofsky, R. N. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10, 104-114.
- Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, 72, 346-357.
- Rubin, D. C., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 25, 1161-1176.
- Sakamoto, Y., Jones, M., & Love, B. C. (2008). Putting the psychology back into psychological models: Mechanistic versus rational approaches. *Memory and Cognition*, 36, 1057-1065.
- Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117, 1144-1167.
- Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 28, 3-11.
- Yu, A. J., & Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior? *Advances in Neural Information Processing Systems*, 21.

# Language-induced Biases on Human Sequential Learning

Luca Onnis (lucao@hawaii.edu)

Department of Second Language Studies & Center for Second Language Research  
University of Hawai'i at Manoa, 1890 East-West Road, Honolulu, Hawaii 96822 USA

Erik Thiessen (thiessen@andrew.cmu.edu)

Department of Psychology  
Carnegie Mellon University, Baker Hall, Pittsburgh, Pennsylvania 15213 USA

## Abstract

What are the effects of experience on subsequent learning? We explored the effects of language-specific word order knowledge on the acquisition of sequential conditional information. Korean and English adults were engaged in a sequence learning task involving three different sets of stimuli: auditory linguistic (nonsense syllables), visual non-linguistic (nonsense shapes), and auditory non-linguistic (pure tones). The forward and backward probabilities between adjacent elements generated two equally probable and orthogonal perceptual parses of the elements, such that any significant preference at test must be due to either general cognitive biases, or prior language-induced biases. We found that language modulated parsing preferences with the linguistic stimuli only. Intriguingly, these preferences are congruent with the dominant *word order* patterns of each language, as corroborated by corpus analyses. These findings suggest that mechanisms of statistical sequential learning are implicated in language, and experience with language may affect cognitive processes and later learning.

**Keywords:** corpus analyses; experience-dependent learning; implicit learning; prediction; retrodiction; second language acquisition; sensitive periods; sequential learning; statistical learning; transitional probabilities; word order; linguistic typology.

## Introduction

Statistical information has been argued to be an important cue to linguistic structure. For example, sounds within a word are more predictable than sounds across boundaries, which may help infants discover words in fluent speech. Because this type of statistical information is present in all languages, statistical information may be a particularly important cue early in development, one that can be used without requiring prior experience with the native language (e.g., Thiessen & Saffran, 2003).

But while statistical learning may be a universal cue to linguistic structure, it is also the case that the statistical structure across languages differs. If statistical learning fails to adapt to these differences, it is unlikely to be an optimal learning strategy. While much research has examined how statistical learning helps learners adapt to the structure of their native language (e.g., Maye, Werker, & Gerken, 2002; Thiessen & Saffran, 2007), it is unknown whether statistical learning itself adapts to the characteristics of the native language. In this series of experiments, we ask whether experience with language alters the kinds of statistical

regularities that learners detect in ways that are consistent with the predominant statistical structure in their native language.

## Prediction and retrodiction in sequential learning

Recent studies have shown that learners can exploit both predictive and retrodictive relations, operationalized as forward and backward transitional probabilities respectively. For instance, Jones & Pashler (2007) showed participants sequences of shapes governed by probabilistic relations, and then asked them to choose which shape reliably came after a probe shape (prediction test) or before a probe shape (retrodiction test). In experiments where forward and backward probabilities were made informative, they found that both prediction and retrodiction were used effectively for recalling memories. In a similar experiment using a continuous sequence of nonsense syllables, Perruchet & Desaulty (2008) found that participants perceived word boundaries based on backward transitional probabilities as well as forward probabilities equally well. Likewise, Pelucchi, Hay, & Saffran (2009) provided evidence that infants can track backward statistics in speech.

The three studies above tested cases in which forward and backward probabilities were never in conflict. Rather, each cue was made maximally informative in a given experiment, while the other was made uninformative. Yet in naturalistic circumstances, prediction and retrodiction may need to be effectively combined, as when learning the word order of a language. In this respect, a comparison between English and Korean seems particularly appropriate because several word order relations of English are reversed in Korean.

## Prediction and retrodiction in natural languages: Typology and word order tendencies.

In English, the head elements in a phrase come first, while in Korean the head follows the phrase (e.g., '[Door-OBJECT] [close-IMPERATIVE]' = '[You close] [the door]', where square brackets indicate phrase groupings). The English sentence "I saw him go there" is glossed as "I him there go saw". Likewise "Give me the ball" is glossed as "Ball me give"; "Let's go get some food" is glossed as "Food get go let's". Thus, frequent constructions such as transitives, imperatives, and exortatives in English have a reversed word-order in Korean.

English is also prepositional ('to school'), while Korean is postpositional ('school to'). We conjectured that since the



linear word order relations in English and Korean are often specular, different sets of expectancies for predictive and retrodictive dependencies may emerge during learning each specific language: for example, in English the predictive probability of the noun *school* following the preposition *to* ( $p(\text{school}|\text{to})$ ) is lower than the retrodictive probability of *to* preceding *school* ( $p(\text{to}|\text{school})$ ), and vice-versa in Korean.

To corroborate these intuitions, we first conducted large-scale corpus studies of the two languages. Then an ambiguous artificial grammar containing different conflicting cues was presented to Korean and American speakers in a sequence learning task. This grammar is unlearnable from surface statistical regularities alone unless previous biases that favor predictive and retrodictive cues are in place in the system before learners start the experiment. Thus, the grammar was used as a litmus test for assessing potential prior biases on learning. To establish whether the biases should be attributable to experience with language or general sequential biases, we tested sequential learning across speakers of the two languages with opposite word order, and in three different modalities: auditory linguistic (speech), visual non-linguistic (abstract shapes), and auditory non-linguistic (pure tones).

We reasoned that if sequential learning mechanisms are directly involved in language acquisition and processing, as it has been put forth, these mechanisms should show language-specific biases effects when adults are engaged in sequence-learning tasks with speech-like stimuli. We further conjectured that if the bias is due to language experience – and not to some more general temporal processing bias – adult participants engaging in the same sequence task using non-linguistic stimuli (visual shapes and auditory tones), should behave consistently *irrespective* of their language background. Another possibility is that sequential learning mechanisms are shared among perceptual modalities and exhibit inherent a priori biases for sequences of stimuli, for example for predictive relations. In this latter case, we would expect a consistent pattern of preference across languages and modalities. Finally, there may be patterns of preference consistent across languages but differing by modality, in which case any effect may be attributable to modality-specific biases.

## Corpus Analyses

We quantified the hypothesis that word order tendencies in Korean and English generate *opposite* patterns of predictive and retrodictive conditional probabilities that signal phrase cohesiveness and syntactic information.

### Corpora

For English we used the SUSANNE Corpus, consisting of 130,000 words of published American English annotated with part-of-speech (POS) and syntactic information (Treebank)<sup>1</sup>. For Korean we sampled the freely available Sejong Corpus, with a syntactically annotated

subcomponent containing 800,000 words<sup>2</sup>. For each sentence in the corpus, we derived unigram and bigram frequency counts as well as forward and backward transitional probability statistics between any two words. Ngram frequencies in English were sampled from the Google Ngram database for the year 2000 (~4 million unigrams and 60 million bigrams). Korean ngram token frequencies were summed over three different corpora: Sejong, HC Korean (55 million unigrams), and KAIST (70 million unigrams) in order to obtain reliable frequency counts. Finally, for each word pair in a sentence we derived the level of syntactic boundary inherent in the syntactic annotation.

## Corpus Measures of phrase cohesiveness

**Independent Variable I: Ngram frequencies** Because several psycholinguistic studies have shown that humans are sensitive to the logarithm of event frequencies as opposed to raw frequencies, it is customary to consider log-frequencies as opposed to raw frequencies. The logfrequency of a sequence of two words (logBigram) can be taken as an approximation of phrase cohesiveness, following Tremblay et al., 2011. The logfrequency of each individual word can also be useful in predicting headedness, as higher frequency words tend to be heads of phrase constituents (Gervain et al., 2008).

**Independent Variable II: Conditional probability** Another way to measure how likely two words are to occur together is to look at a word and estimate what words are likely to follow it. The likelihood of a given word following is the forward probability of the word pair. For example, for the sequence ‘in Sapporo’:

$$\text{fwdTP}(\text{Sapporo}|\text{in}) = \frac{\text{freq}(\text{in\_Sapporo})}{\text{freq}(\text{in})}$$

The calculation can also be computed in the opposite direction. That is, examine a word and estimate what words are likely to precede it. The likelihood of a given word preceding is known as the backward probability between the two words.

$$\text{bckTP}(\text{in}|\text{Sapporo}) = \frac{\text{freq}(\text{in\_Sapporo})}{\text{freq}(\text{Sapporo})}$$

For example, suppose the word “in” occurs 2853 times in the corpus, but the word “Sapporo” occurs only 9 times and the sequence of words “in Sapporo” occurs 3 times. Since the word “in” occurs 2853 times, and only 3 of those times with the word Sapporo, this pair of words has a very low forward probability (3/2853). However, if we examine the pair from the opposite direction, we see that three out of the nine times the word “Sapporo” appears, it is preceded by the word “in.” Thus, the backward probability is 3/9, or .33.

**Dependent Variable: Phrase structure cohesiveness** To estimate the informativeness of transitional probabilities and frequencies in parsing at the constituent level, we followed Johnson (1965): Sentences from the tree-tagged Susanne corpus and Sejong corpus were divided up into phrasal constituents. For every word pair transition considered

<sup>1</sup> Available at <http://www.grsampson.net/Resources.html>

<sup>2</sup> Available at <http://rocker.snu.ac.kr:8080/search>

linearly from left to right, it is possible to rank-order the level at which a constituent node for that transition occurs. For example, in “[[[The house] [across [the street]]] [is burning]]” the highest node is at transition 5 (*street\_is*), followed in rank by transition 2 (*house\_across*), then transition 3 (*across\_the*). Finally, transitions 1, 4, and 6 are tied on the same rank. Using the syntactically annotated corpora, every bigram in each sentence can be assigned a syntactic rank, following the example above.

Typological studies of the world languages have uncovered important correlations in the linear order of the constituents in different subdomains. For instance, the constituent order of a clause (the relative order of subject, object, and verb); the order of modifiers (adjectives, numerals, demonstratives, possessives, and adjuncts) in a noun phrase; and the order of adverbials and adpositions. Most languages appear to have a preferred word order that is usually most frequent. This ordering of constituents is often represented as a tree where branches can be divided into other minor branches, which may also branch in turn. English is often described as a right-branching language, because it tends to place dependents after the head words. Adjectives follow nouns, direct objects follow verbs, and adpositions are prepositional. This type of branching is also known as head-first order. Left-branching languages, like Korean and Japanese, exhibit the opposite tendency, that is, they tend to place the head element of a phrase to the left. Objects appear to the left of verbs, sentences appear to the left of subordinating conjunctions and noun phrases appear to the left of prepositions (which, for this reason, are often called postpositions in these languages). Since postpositions come after the noun in left-branching languages, our example phrase, “in Sapporo,” would actually be in the opposite order, “Sapporo in.”

We were interested in assessing whether regularities in the word order of English and Korean engender language-specific probabilistic expectations between sequences of adjacent words. For example, considering the sequence “in Sapporo” the forward probability is expected to be low, arguably because many words can follow “in” (Rome, New York, summer, me, the, lovely, etc.). Conversely, the backward probability should be high, because only a few words are expected before “Sapporo” (to, in). Thus a pattern of “forward low-backward high” probability (*LoHi* for short) is expected to run as a sentence unfolds in English. If we express this combined pattern in mathematical terms as the difference between the forward and the backward probability (TPdiff), for any word pair in a sentence we should expect a positive larger TPdiff to indicate a more cohesive phrase unit in the language. Thus, the TP difference could be taken to predict the level of syntactic constituency between any word pair in the syntactically tagged corpora. Notably, for Korean the pattern of transition probabilities is expected to be reversed. For “Sapporo in”, the forward probability should be high relative to the backward probability. Thus, *HiLo* patterns are expected to run as a sentence unfolds in Korean. Using the same

differential measure (TPdiff) between forward and backward probability, this time we can expect a large negative TPdiff to predict more cohesive phrase units in the syntactically-tagged Korean corpus. These predictions “by example” are by no means granted for the whole language. In the syntactic literature it has long been noted that the right-branching/left-branching dichotomy may not hold for an entire language, and in the case of English it is not fully consistent even at the phrasal level (for instance for the word ordering within a Noun Phrase, see Cook and Newson 2007). Thus it is important to evaluate whether these probabilistic biases are significantly and robustly correlated with word order across the two language corpora.

## Results

Ordinal logistic regressions were run to predict the syntactic tree level between any two members of word pairs (word1, word2, e.g., “in Sapporo”) in a sentence. The syntactic tree level was obtained from the syntactic parsing provided in the Susanne and Sejong corpora (henceforth English and Korean corpus respectively). Therefore, the tree level (henceforth *tree*) was the dependent variable to be predicted by the regression models. The following predictors were considered in the following order: log frequency of each bigram, log-frequency of first word, log-frequency of second word, forward probability, backward probability. Because node levels above 6 were very infrequent in both corpora, we considered the first six node levels, accounting for 99.5% of bigrams in the English corpus (109,861 bigrams entered in the analyses) and for 98.1% of bigrams in the Korean corpus (22,382 bigrams entered).

**Model fit** For each corpus, ordinal logistic regression models with different complexity were fitted and compared for goodness of fit. The null model contained no predictors, then increasingly complex models added logBigram, LogFrequency of first word, LogFrequency of second word, Forward Probability, and Backward Probability as predictors. Analyses of deviance between each increasingly complex model indicated that including all predictors except LogFrequency of second word increased the fit of each regression model significantly with respect to the previous less complex model by reducing deviance. This result held for both corpora. Thus, in the following analyses the log-frequency of the second word was excluded as a predictor. All other variables contributed significantly to predict the level of syntactic node between any two adjacent words in a sentence in the corpus. Using the *lrm* function in R we were also able to assess the goodness of fit of the models. As the p-values of the G test statistics are 0 in both language models, the null hypothesis can be rejected that there is no overall significant relationship between the dependent variable *tree* and the independent variables. The predictive ability of the model can also be measured using C, an index of concordance between the predicted probability and the observed response. (if C=0.5 the predictors are random, when it is 1, prediction is perfect). Since C=0.71 for English and C=0.64 for the Korean corpus, we have confidence that

both models have moderate to strong predictive capacity. Somer's  $D_{xy}$  is a rank correlation between predicted probabilities and observed responses. It ranges between 0 (randomness) and 1 (perfect prediction). Since  $D_{xy}=0.43$  (English) and  $D_{xy}=0.30$  (Korean) we have again confidence that both language models have moderate predictive capacity. Kendall's Tau-a rank correlations also assess the correlations between all predicted probabilities and observed response. Here  $\text{Tau-a}=0.26$  (English) and  $\text{Tau-a}=0.16$  (Korean).

**English Corpus** We were particularly interested in the *coefficient sign* of the independent variables across the two corpora. For English, the coefficients for *logBigram* were consistently negative, indicating that the more frequent bigrams are associated with tighter phrase boundaries. Thus, the frequency of a bigram can be used to partially predict phrase cohesiveness, in accord with the psycholinguistic evidence obtained by Tremblay et al. (2011). Most importantly for the hypothesis being tested in this study, lower forward probabilities and concurrently higher backward probabilities (a *LoHi* pattern) were associated with higher phrase cohesiveness, as indicated by a positively valued coefficient for *TPdiff*. Remember that low syntactic levels indicate that the word pair tends to be occurring within the same phrase, or across a transition that is at a lower level up the syntactic tree. In addition, higher frequency of first words was associated with more phrasal cohesiveness, in accord with Gervain et al. (2008).

**Korean Corpus** *LogBigram* frequency was negatively associated with syntactic depth, indicating again that more cohesive phrases tend to be more frequent. Crucially, the coefficient for *TPdiff* was now positive (and reversed with respect to English), indicated that higher forward probabilities and lower backward probabilities (a *HiLo* pattern) were associated with higher phrase cohesiveness.

**Summary** When comparing English and Korean, the patterns of probability that support syntactic parsing are clearly reversed in the two languages, as originally predicted. In particular, phrase cohesiveness correlates with a *LoHi* pattern of transition probabilities in English, and with a *HiLo* pattern in Korean. Below we ask whether these language-specific patterns of probabilities are a source of experience-induced bias when learners group novel stimuli in a sequence learning task.

## Experiment 1

The corpus analyses above provide a rationale for our main hypothesis, namely that the predictive regularities most consistently experienced in one's native language, may impose processing biases on human sequential learning (Table 2). A speech-synthesized stream of syllables was constructed so that two mutually exclusive sets of syllable groupings could possibly be extracted, according to either a bias for a *LoHi* probability pattern (as in English 'to school', Table 2), or a *HiLo* probability pattern (as in Korean 'school to', Table 2). Because the two sets were equally frequent (*HiLo* grouping,  $M=59.2$ ,  $SD=2.9$ ; *LoHi* grouping,  $M=59.3$ ,

$SD=3.2$ ; difference *ns*), a consistent preference for either of them would be indicative of a statistical learning bias developed prior to the experiment.

## Method

**Participants** Thirty-seven English monolingual and 36 native Korean students participated. Korean participants were enrolled in graduate programs at the University of Hawaii, and their scores in the TOEFL test of English as Second Language was high ( $M=252.14$ , out of 300,  $SD=16$ ).

**Stimuli** A seamless 5-min speech sequence of 8 synthesized syllables (*bu, fu, ra, ti, she, zi, ge, ni*) was generated following the forward and backward transitional probabilities in Table 3, with 80 ms for consonants and 260 ms for vowels, and 5 s fade-in and out. We used the Italian diphone set in order to engage participants in a foreign language learning task for both groups. The Italian phonemes we chose had equivalent phonemic realizations in English and Korean, and all syllable sequences were phonotactically legal. No participant knew Italian. No cue to word boundary was present other than the patterns of predictive and retrodictive dependencies (Table 1). Importantly, whenever forward probability was low between any two adjacent syllables, ( $\text{fwdTP}(zi|she=.33)$ ), backward probability was high ( $\text{backTP}(she|zi=1)$ ), and vice versa (Table 1). At test, two groupings corresponding to a pattern of *HiLo* probability and *LoHi* probability were pitted one against the other in a forced-choice task. None of the possible groupings was an actual word in either language. The six test pair trials were presented in random sequential order, while the order within a pair was counterbalanced by repeating each test pair twice.

**Procedure** Participants in each language group were randomly assigned either to the experimental condition that included Training and Test or to a control condition that included the Test phase only (18 English native speakers, 21 Korean native speakers). Participants in the experimental condition listened to the training stream for 5 minutes. Test consisted in 12 two forced-choice task trials between pairs of *LoHi* and *HiLo* groupings. For each pair they were asked to choose which sound sequence formed a grouping in the language they had just heard. All instructions were administered in the native language of participants.

Table 1. The forward [square brackets] and backward (no brackets) transition probabilities associated with any two adjacent stimuli in the training sequence of the three experiments (in Experiments 2 and 3 the syllables were replaced by abstract shapes and tones respectively). For example, given *bu* only *fu* could follow ( $\text{fwd-TP}(fu|bu)=[1]$ ), while given *fu* there was a .33 probability that *bu* preceded it ( $\text{back-TP}(bu|fu)=(.33)$ ). Last three rows:

A sample of the training sequence of Experiment 1. Perceptual grouping boundaries could emerge either when the forward transitional probability between adjacent syllables was high and the backward probability was low (*HiLo* groupings), or viceversa (*LoHi* groupings).

From	To							
	bu	fu	ra	Ti	she	zi	ge	ni
bu	0	[1]	0	0	0	1	0	0
fu	.33	0	.33	.33	[.33]	0	[.33]	[.33]
ra	0	[1]	0	0	0	1	0	0
ti	0	[1]	0	0	0	1	0	0
she	0	1	0	0	0	1	0	0
zi	[.33]	0	[.33]	[.33]	[.33]	0	.33	.33
ge	0	1	0	0	0	[1]	0	0
ni	0	1	0	0	0	[1]	0	0
Training ... fushezirafunizitifugezibu ...								
HiLo groupings ... fushe zira funi ziti fuge zibu ...								
LoHi groupings ...fu shezi rafu nizi tifu gezi ...								

## Results

In all three experiments presented participants responses were coded in terms of proportion endorsements for HiLo groupings. Consequently, low endorsement rates for HiLo indicate preferences for LoHi groupings. A 2 (Language: Korean, English) x 2 (Condition: Experimental, Control) ANOVA revealed a main effect of Language ( $F(1,72)=11.22$ ,  $p<0.01$ ), and a Language by Condition interaction ( $F(1,72)=10.67$ ,  $p<0.01$ ). In particular, English native speakers exposed to training consistently preferred the LoHi groupings 62% of times ( $SD=14\%$ ), which was reliably different than chance ( $p < .001$ ). Conversely, Korean native speakers exposed to training preferred HiLo groupings 59% of times ( $SD=13.6\%$ ), which was reliably different than chance ( $p < .025$ ). Thus, English and Korean participants attended to the transitional probabilities that were most predictive of the canonical word order of their native language, as predicted by the corpus analyses. When presented with the Test items alone, no preference emerged for either groupings above or below change (English,  $t(17) = 0.44$ ,  $p = 0.67$ ; Korean,  $t(20) = 1.03$ ,  $p = 0.32$ ), ensuring that the bias in the experimental condition was not due to inherent preferences for certain sound combinations of the test items.

## Experiment 2

In order to further ascertain that the different scores in Experiment 1 were due to language-specific biases, in Experiment 2 we tested whether learning biases would arise when the same miniature grammar was implemented with non-linguistic stimuli in the visual modality. As discussed previously, sensitivity to forward and backward probabilities has been demonstrated in non-language domains, including visual processing (Fiser & Aslin, 2002; Jones & Pashler, 2007). However, the two cues were not pitted against each other in those experiments, but rather one or the other was maximally informative in the input. Here, at any stimulus transition the two cues are equally informative, but pitted against each other. Therefore we expected one of two scenarios. A null result would obtain if learners as a group weigh each cues equally, as indeed Jones & Pashler (2007)'s data suggest. Alternatively, if a priori visual preferences are attested in participants' responses, we expect them to be general visual sequential processing

biases not influenced by language experience. Thus, regardless of scenario, we expected no differential preferences based on the language of our participants.

## Method

**Participants** 15 new English and 15 Korean native speakers from the same population as Experiment 1 participated.

**Stimuli** A continuous sequence was generated that had exactly the same structure and length as Experiment 1, with the only exception that the 8 synthesized syllables were now replaced by 8 abstract shapes. Shapes appeared in succession on the screen for 340ms. The sequence had the same statistical properties as the language in Experiment 1.

**Procedure** The same learning and test procedure as Experiment 1 applied. At test participants received a two forced-choice task between pairs of LoHi and HiLo shapes. For each pair they were asked to choose which one formed a grouping in the sequence they had just seen. All instructions were administered in the native language of participants.

## Results

A one-way ANOVA revealed no significant main effect of Language ( $F(1,28)=0.025$ ,  $p=0.87$ ). Moreover, mean test items endorsements did not differ from chance (Korean,  $M = 0.51$ ,  $SD=0.17$ ,  $t(14) = 0.26$ ,  $p = 0.8$ ; English,  $M = 0.51$ ,  $SD=0.16$ ,  $t(13) = 0.03$ ,  $p = 0.97$ ).

## Experiment 3

The results of Experiment 2 indicate that the difference in directional preference between English and Korean speakers does not extend to visual sequences. This is consistent with the hypothesis that what drives the difference between language speakers in Experiment 1 is their experience with language. However, prior research indicates that there are important differences between processing visual stimuli and audio stimuli, due in part to the more transient nature of audio information (e.g., Conway & Christiansen, 2005). As such, it may be the case that the differences between English and Korean speakers are not specific to language, but rather arise from more general differences in auditory processing. To assess this possibility, we created a tonal analog of the input from Experiment 1.

## Method

**Participants** Both English monolinguals ( $N=15$ ) and Korean/English bilingual ( $N=15$ ) who reported Korean as their dominant language participated in this experiment. All English monolinguals were undergraduates at Carnegie Mellon University, as were six of the Korean English bilinguals. The other 9 bilingual participants were recruited via advertising in Pittsburgh churches.

**Stimuli** Each of the syllables in the language used in Experiment 1 was replaced by a unique 330 ms tone (bu = A4, ti = B, ge = C#, zi = D, fu = E, ni = F#, she = G#, ra = A5) in the key of A major. The resulting tonal sequence thus had an identical statistical structure as in Experiment 1.

**Procedure** Participants listened to the tone sequence over headphones. Next, participants were given 12 forced choice questions and asked to indicate, on a response sheet, which of two items sounded “more like” the tone sequence they had just heard. On each of the 12 questions, a tone item with a high-forward, low-backward transitional probability was paired with an item with low-forward, high-backward transitional probability.

## Results

A one-way ANOVA revealed no effect of Language ( $F(1,28)=0.025$ ,  $p=0.87$ ). Both English ( $M = 0.56$ ,  $SD = 0.09$ ,  $t(14) = 2.6$ ,  $p = .01$ ) and Korean-English bilinguals ( $M = 0.58$ ,  $SD = 1.14$ ,  $t(14) = 2.1$ ,  $p = .05$ ) selected test items with high forward transitional probabilities (HiLo items) at a rate above chance. The fact that they performed equivalently in this experiment is consistent with the hypothesis that the differences between these two populations in Experiment 1 are language-specific, and strengthen the claim that they arise from linguistic experience. Unlike adults’ lack of preference for shape test items in Experiment 2, participants did have a consistent preference for test items with high forward transitional probabilities, perhaps because the preference for forward-going items is a domain-general auditory preference, similar to the Iambic-Trochaic Law (e.g., Hay & Diehl, 2007). This preference may be early-developing, and then modified by linguistic experience: strengthened for Korean learners, and contravened by English learners. Alternatively, experience with music may inculcate a bias in both English and Korean listeners; on this account, musical experience is more consistent cross-culturally than linguistic experience.

## Discussion

We tested the hypothesis that adult English and Korean speakers come to the lab having already developed opposite statistical preferences for parsing continuous speech. The results of Experiment 1 supported that hypothesis: where English speakers preferred items with high backward probabilities, Korean speakers preferred items with high forward probabilities. Experiments 2 and 3 suggest that this preference is limited to linguistic materials. This limitation is consistent with the possibility that the preference arises from experience with language. Corpus analyses of English and Korean are consistent with this possibility, as the predominant word order of both languages mirrors the direction preference of English and Korean speakers.

Our findings have implications for understanding processes of second language acquisition. Our Korean participants were advanced second language speakers of English, and were enrolled in graduate programs in the United States. Their sensitivity to the learning bias opposite to that of the English participants suggests that they implicitly used their solidified L1’s learning biases when learning the novel artificial language. Because patterns that conform to those initially learned are further promoted, interference can be most severe with the learning of patterns

that do not conform to those initially learned. We propose that lower-order kinds of transfer involving basic sequential processing biases are at play in second language acquisition, and may have a ripple effect on encoding higher-order processes such as word order structure.

The results suggest that statistical learning changes throughout development by adapting to the characteristics of the native language. This opens many avenues for subsequent research, including understanding the mechanisms and developmental time course through which experience with native language alters subsequent learning.

## Reference

- Conway, C.M., & Christiansen, M.H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 24-39.
- Cook, V.J., & Newson, M. (2007). *Chomsky’s universal grammar: An introduction* (3rd ed.). Wiley-Blackwell.
- Fiser, J., & Aslin, R.N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, 99, 15822-15826.
- Gervain, J. Nespor, M., Mazuka, R., Horie, R., Mehler J. (2008) Bootstrapping word order in prelexical infants: a Japanese-Italian cross-linguistic study. *Cognitive Psychology*, 57(1), 56-74.
- Hay, J.F., & Diehl, R.L. (2007). Perception of rhythmic grouping: Testing the Iambic/Trochaic Law. *Perception and Psychophysics*, 69, 113-122.
- Johnson, N.F. (1965) The psychological reality of phrase structure rules. *Journal of Verbal Learning and Verbal Behavior*, 4, 469-475.
- Jones, J., & Pashler, H. (2007). Is the mind inherently forward looking? Comparing prediction and retrodiction. *Psychonomic Bulletin and Review*, 14(2), 295-300.
- Maye, J., Werker, J. F. & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3), B101- B111.
- Pelucchi, B., Hay, J.F., Saffran, J.R. (2009). Learning in reverse: Eight-month-old infants track backwards transitional probabilities. *Cognition*, 113, 244-247.
- Perruchet, P., & Desauty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, 36, 1299-1305.
- Thiessen E.D., & Saffran, J.R. (2007). Learning to learn: Infants’ acquisition of stress-based strategies for word segmentation. *Language Learning & Development*, 3, 73-100.
- Thiessen, E.D., & Saffran, J.R. (2003). When cues collide: Use of statistical and stress cues to word boundaries by 7- and 9- month-old infants. *Developmental Psychology*, 39, 706-716.
- Tremblay, A., Derwing, B., Libben, G. and Westbury, C. (2011), Processing Advantages of Lexical Bundles: Evidence From Self-Paced Reading and Sentence Recall Tasks. *Language Learning*, 61: 569–613.

# Semantic Coherence Facilitates Distributional Learning of Word Meanings

Long Ouyang, Lera Boroditsky, and Michael C. Frank

{louyang, lera, mcfrank}@stanford.edu

Department of Psychology

Stanford University

## Abstract

Computational work has suggested that one could, in principle, learn aspects of word meaning simply from patterns of co-occurrence between words. The extent to which humans can do this distributional learning is an open question – artificial language learning experiments have yielded mixed results, prompting suggestions that distributional cues must be correlated with other cues, such as phonological regularities, for successful learning. We conducted a large-scale experiment comparing how distributional learning is afforded by two different types of cues – phonological regularities and semantic coherence. We found that semantic coherence more strongly facilitates distributional learning than onset-consonant phonological regularities.

**Keywords:** word learning; distributional learning; semantic coherence

## Introduction

How do people learn what words mean? According to the distributional hypothesis (Harris, 1951; Firth, 1957), patterns of word co-occurrence are a powerful source of information about word meaning. It may be possible to acquire some facets of word meaning by simply keeping track of which linguistic contexts words appear in, even without access to any physical or social cues. Computational models have lent quantitative support for the distributional hypothesis. For example, the Latent Semantic Analysis model of Landauer & Dumais (1997) clustered words according to their patterns of occurrence across documents and was able to closely match human performance on synonym tests. The success of early models like LSA led to a proliferation of models that use distributional information to learn word meaning (see Riordan & Jones, 2010 for an overview) as well as other linguistic properties such as syntactic category (e.g., Redington, Chater, & Finch, 1998).

We know that distributional learning is computationally possible, but the evidence of human capacity for distributional learning is mixed. Some research has found that people learn co-occurrence statistics and use them to inform categorization (Mintz, 2002; Gerken, Wilson, & Lewis, 2005; Reeder, Newport, & Aslin, 2010), but other work has reached the opposite conclusion (Braine, 1966; Smith, 1966; Frank & Gibson, 2009). A holistic reading of the literature suggests that learning is more likely to occur when additional cues are correlated with distributional properties. Below, we review two representative findings using the *MNPQ* artificial language learning paradigm (Braine, 1966; Smith, 1966).

The *MNPQ* grammar contains four categories of words – *M*, *N*, *P*, and *Q* – and sentences take one of two forms: *MN* and *PQ*. Early investigations (Braine, 1966; Smith, 1966)

found that subjects recall “grammatical” *MN* and *PQ* sentences but also tend to recall ungrammatical *MQ* and *PN* sentences, suggesting that they learn position regularities (that *M/P* come first and *N/Q* come second) but not co-occurrence regularities (that *M* co-occurs with *N* but not *Q* and that *P* occurs with *Q* but not *N*). *MNPQ* learning has been an important paradigm because it provides an empirical means to consider purely distributional learning of word categories.

Braine (1987) investigated the effect of correlating co-occurrence regularities with natural gender. Subjects acquired an artificial language by learning to name pictures of referents. In the experimental condition, all pictures of men were labeled by *M* words and all pictures of women were labeled by *P* words. Learning of the co-occurrence regularities was significantly higher in the experimental condition than in a control condition where natural gender was not predictive of *M/P* membership. Though Braine’s experiment combined distributional cues with natural gender, he suggested that phonological cues might better serve real-world language learners. For instance, Spanish and Italian speakers might learn grammatical gender categories by taking advantage of the fact that feminine nouns often end with *-a*, while masculine nouns often end with *-o*. Recently, this suggestion received attention in the work of Lany and Saffran (2010), who found that 22-month old infants learned *MNPQ* co-occurrence regularities when they were aligned with the number of syllables in a word – that is, when *N* words were disyllabic and *Q* words were monosyllabic, but *not* when the number of syllables was not predictive of *N/Q* membership.

The defining feature of the artificial language learning paradigm is that, at the outset of the experiment, subjects do not know the meanings of any of the words. Yet, this condition generally does not hold true for real language learners, who typically know the meanings of some words but not others. It is plausible that the presence of known words facilitates distributional learning. If this is true, how effective are such semantic cues, and how do they compare to (e.g.) phonological cues? Here, we report the initial results of the first large-scale study comparing the effectiveness of various correlated cues on distributional learning. We presented subjects with an *MNPQ* language where sentences took the form “*M* and *N*” or “*P* and *Q*”. We hypothesized that distributional learning would be afforded given a certain level of *semantic coherence*, where all *M* and *P* words were familiar and adhered to some semantic organization (e.g., *M* = animal words, *P* = vehicle words). For instance, hearing the four sentences “cat and dax”, “cat and ziv”, “car and wug”, and “car and pif” might allow learners to infer that daxes and zivs belong to the

same category, as both words co-occur with “cat”, and that wugs and pifs belong to the same category, as both words co-occur with “car”.

In Experiment 1, we tested whether semantic coherence facilitated distributional learning. In Experiments 2 and 3, we compared semantic coherence to phonological coherence and semantic incoherence (the presence of known words that *do not* adhere to some semantic organization).

## Experiment 1: Semantic Coherence

We parametrically varied two independent variables: (1) the amount of exposure to the language and (2) semantic coherence – the fraction of known words that adhered to a taxonomic organization ( $M$  = animal words,  $P$  = vehicle words). We then applied three measures of  $MNPQ$  learning – sentence memory, similarity rating, and a referent assignment task.

### Method

**Participants** 654 Amazon Mechanical Turk (MTurk) workers participated in the study. Using MTurk’s worker qualifications we limited participation to workers located in the US and with a previous HIT approval rate  $\geq 90\%$ . We chose MTurk workers because the number of experimental conditions required a large number of subjects.

**Materials** Sentences took the form “ $M$  and  $N$ ” or “ $P$  and  $Q$ ” (see Figure 1). We generated the actual lexical items randomly for each subject.  $N$ ’s and  $Q$ ’s were always novel nonsense words and were drawn without replacement from {*moke*, *thite*, *jiv*, *pif*, *dex*, *wug*}.  $M$ ’s and  $P$ ’s could be either novel or familiar. Novel  $M$ ’s were drawn from {*feeb*, *bim*, *lup*} and novel  $P$ ’s were drawn from {*zabe*, *vap*, *chuv*}. Familiar  $M$ ’s and  $P$ ’s obeyed a taxonomic organization – familiar  $M$ ’s were drawn from {*hamster*, *cat*, *dog*} (animal words) and familiar  $P$ ’s were drawn from {*car*, *bus*, *truck*} (vehicle words). To create the audio files, we input the sentences as “X. and. Y.” (e.g., “*car*. and. *chuv*.”, including periods) into an American English text-to-speech engine using a female voice. The periods between words introduced substantial pauses ranging in length from 150 to 300 ms; piloting revealed that without pauses, it was difficult for participants to distinguish the words. Sentences using only monosyllabic words were around 2 seconds long. Sentences using the sole disyllabic word, *hamster*, were around 3 seconds long.

The referent assignment task involved visual referents. For the context words, we used 128x128 pixel images of a cat, dog, hamster, car, bus, and truck. For the target words, we used 100x100 pixel images of a horse, rabbit, sheep, bear, goldfish, mouse, boat, van, train, motorcycle, plane, and bicycle.

**Design and Procedure** We parametrically varied coherence – the number of familiar  $M$  and  $P$  words. The language for a subject contained either 0/3, 1/3, 2/3, or 3/3 familiar  $M$  and  $P$  words each. We also varied the amount of exposure to the language – subjects heard either 56, 126, 196, or 392 sentences. Before starting the experiment, we asked subjects

Figure 1: The  $MNPQ$  language. Underlined sentences were withheld from exposure in Experiments 1-3.

<u><math>m_1 n_1</math></u>	$m_1 n_2$	$m_1 n_3$	<u><math>p_1 q_1</math></u>	$p_1 q_2$	$p_1 q_3$
$m_2 n_1$	<u><math>m_2 n_2</math></u>	$m_2 n_3$	$p_2 q_1$	<u><math>p_2 q_2</math></u>	$p_2 q_3$
$m_3 n_1$	$m_3 n_2$	<u><math>m_3 n_3</math></u>	$p_3 q_1$	$p_3 q_2$	<u><math>p_3 q_3</math></u>

to turn on their speakers and click a button, which played a spoken English word. Subjects were required to correctly type the word to continue. The experiment had four phases – exposure, similarity, memory, and referent assign. Below, we detail these phases (for purposes of exposition, we have switched the order of memory and similarity).

**Exposure.** Subjects listened to sentences from the language. We withheld six sentences from exposure (see Figure 1), yielding 14 unique sentences in the exposure set. Each sentence was heard either 4, 9, 14, or 28 times, giving 56, 126, 196, or 392 total trials. We presented the sentences in random order subject to the constraint that there were no repeated words between consecutive trials (pilot testing suggested that repeated words substantially afforded learning). To encourage compliance, subjects had to click a button to hear each sentence.

**Memory.** Subjects listened to sentences and judged on a 5 point scale how confident they were that they had previously heard the sentence during exposure. We tested four types of sentences – *familiar* sentences heard during exposure, *withheld* sentences not heard during exposure but conforming to the  $MNPQ$  structure, *cross-category* sentences of the form  $MQ$  and  $PN$ , and *position-violation* sentences of the form  $MM$ ,  $NN$ ,  $PP$ , and  $QQ$ . Sentences were presented in random order such that there were no repeated words between consecutive trials. In two catch trials, instead of a sentence from the  $MNPQ$  language, we played a non-repeatable audio instruction to press a specific response button. If subjects learned the  $MN$  and  $PQ$  co-occurrence relationships, then we expected that they would rate novel grammatical sentences respecting the co-occurrence relationships as more familiar than the cross-category sentences violating the co-occurrence relationships.

**Similarity.** For each pair of words in the union of  $N$  and  $Q$ , we asked subjects to rate on a 5 point scale how similar they believed the two words to be in meaning. This resulted in within-category judgments (e.g.,  $n_1$  vs.  $n_2$ ) and cross-category judgments (e.g.,  $n_1$  vs.  $q_1$ ). We presented the pairs in a fixed pseudorandom order containing no repeated words between consecutive trials. Though exposure was entirely auditory, for convenience, we presented these similarity questions as text (e.g., “How similar are **pif** and **thite**?”); to facilitate mapping between visual and spoken word forms, the speaker button next to each word played the spoken word when clicked. In two catch trials, subjects were asked to press the response button corresponding to the solution of a simple arithmetic problem. If subjects learned the  $MN$  and  $PQ$



co-occurrence relationships *and* used these relationships as a basis for lexical categorization, then we expected that within-category pairs of words would be rated as more similar than cross-category pairs.

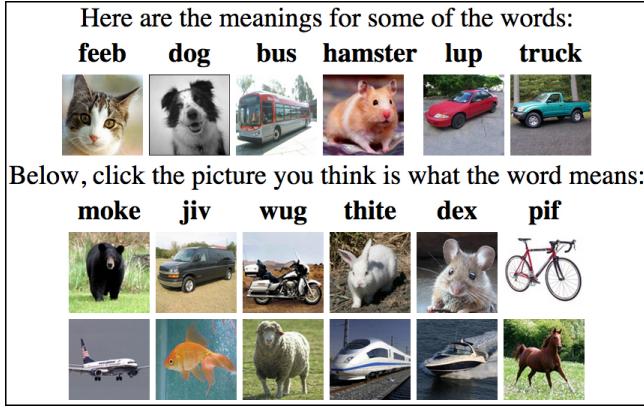


Figure 2: The referent assignment task.

*Referent assignment.* Subjects made 2AFC referent assignments for the  $N$  and  $Q$  words (see Figure 2). At the top of the screen, we displayed the  $M$  and  $P$  words in random order. Under each word, we showed an image of an associated referent. The referents corresponded to the familiar pools for  $M$  and  $P$  words: CAT, DOG, HAMSTER, CAR, BUS, and TRUCK. Familiar words were always associated with the obvious referents (e.g., “dog” was always paired with an image of a dog). Below the “seeded” word meanings, we displayed a row containing the  $N$  and  $Q$  words. Under each word, we displayed a 2AFC referent choice between an animal (the “correct” choice for  $N$  words) and vehicle words (the “correct” choice for  $Q$  words); subjects made a choice by clicking on one of the two pictures. If subjects learned the  $MN$  and  $PQ$  co-occurrence relationships *and* used them to form nascent lexical categories *and* used these lexical categories as a basis for inferences about word meaning, then we expected that referent assignment scores would reflect a tendency to choose on the basis of the taxonomic categories of the co-occurring words (e.g.,  $N$ ’s should be animals because they co-occur with  $M$ ’s, which are known to be animals).

## Results and Discussion

We excluded the 47 subjects who did not correctly answer all of the catch trials. Results are shown in Figure 3. Next, for each dependent measure – memory, similarity, and meaning – we defined a within-subject score representing the sensitivity to the co-occurrence regularities in the language. Memory score was the difference in mean ratings between novel withheld sentences (e.g.,  $m_1-n_1$ ) and novel category violation sentences (e.g.,  $m_1-q_1$ ). Similarity score was the difference between mean ratings of within-category (e.g.,  $N-N$ ) and cross-category (e.g.  $N-Q$ ) ratings. Referent assignment score was the total number of correct choices in the referent assignment task. We analyzed two aspects of the data. First, we were in-

terested in main effects of coherence on score. Second, as one hallmark of statistical learning is sensitivity to the amount of evidence observed, we were interested in the relationship between amount of exposure and score. Accordingly, we looked for exposure  $\times$  coherence interactions. A significant interaction would indicate a difference in how *efficiently* the statistical learning process makes use of evidence at different coherence levels. For all scores, we coded coherence as a categorical variable and analyzed the data using an interactive regression model: score  $\sim$  exposures  $\times$  condition. To examine the differences between the different coherence levels, we used Helmert contrasts analyzing (i) the difference between the 1/3 and 0/3 conditions, (ii) the difference between the 2/3 condition and the 0/3 and 1/3 conditions combined, and (iii) the difference between the 3/3 condition and the 0/3, 1/3, and 2/3 conditions combined. Results of these analyses are shown in Table 1.

Table 1: Regression models

Regressor	$\beta$	$t$	$p$
Memory			
Exposures	< 0.001	2.67	< 0.01*
Condition: 1/3 – (0/3)	-0.003	-0.03	0.97
Condition: 2/3 – (0/3,1/3)	0.101	1.78	0.074
Condition: 3/3 – (0/3,1/3,2/3)	0.154	3.62	< 0.01*
E $\times$ C: 1/3 – (0/3)	< 0.001	0.17	0.86
E $\times$ C: 2/3 – (0/3,1/3)	> -0.001	-0.20	0.83
E $\times$ C: 3/3 – (0/3,1/3,2/3)	< 0.001	2.76	< 0.01*
Similarity			
Exposures	< 0.001	1.82	0.06
Condition: 1/3 – (0/3)	-0.039	-0.36	0.71
Condition: 2/3 – (0/3,1/3)	0.075	1.33	0.18
Condition: 3/3 – (0/3,1/3,2/3)	0.097	2.30	0.02*
E $\times$ C: 1/3 – (0/3)	< 0.001	0.53	0.59
E $\times$ C: 2/3 – (0/3,1/3)	< 0.001	0.42	0.67
E $\times$ C: 3/3 – (0/3,1/3,2/3)	< 0.001	2.66	< 0.02*
Referent assignment			
Exposures	0.001	3.16	< 0.01*
Condition: 1/3 – (0/3)	0.153	1.01	0.31
Condition: 2/3 – (0/3,1/3)	0.183	2.26	0.02*
Condition: 3/3 – (0/3,1/3,2/3)	0.121	2.00	0.04*
E $\times$ C: 1/3 – (0/3)	> -0.001	-0.29	0.77
E $\times$ C: 2/3 – (0/3,1/3)	> -0.001	-0.89	0.37
E $\times$ C: 3/3 – (0/3,1/3,2/3)	< 0.001	1.29	0.19

**Memory** There were significant main effects of exposure and condition, with scores in the 3/3 condition being significantly higher than in the other conditions combined. Additionally, there was a significant exposure  $\times$  condition interaction; the effect of exposures on score was significantly higher in 3/3 than in the other conditions combined, suggesting greater efficiency of statistical learning in 3/3. Thus, more coherent linguistic input (1) bolstered memory for the  $MN$  and  $PQ$  co-occurrence regularities and (2) increased the efficiency of the statistical learning process responsible for learning those regularities, at least in the 3/3 condition.

**Similarity** There was a significant main effect of condition, with scores in 3/3 being significantly higher than in the other conditions combined. Additionally, there was a significant exposure  $\times$  condition interaction; the effect of exposures on

score was significantly higher in 3/3 than in the other conditions combined. Thus, more coherent linguistic input (1) increased the distinction between within-category and cross-category pairs of words and (2) increased the efficiency of the statistical learning process involved in making such distinctions, at least in the 3/3 condition.

**Referent assignment** There were significant main effects of exposure and condition. 2/3 scores were significantly higher than 0/3 and 1/3 scores combined and 3/3 scores were significantly higher than the rest of the scores combined. None of the interaction terms reached significance, indicating that the amount of exposure to the language and greater coherence independently increased the ability to assign *N* and *Q* words to the correct referents. We also computed this model using coherence as a continuous variable; this continuous regressor significantly predicted increases in score,  $\beta = 0.29$ ,  $t(650) = 3.06$ ,  $p < 0.005$ , indicating that *parametrically* increasing coherence resulted in *parametric* increases in referent assignment score.

To summarize, in Experiment 1, we found that higher coherence (1) increased ability to distinguish novel grammatical sentences from sentences violating co-occurrence regularities, (2) sharpened sensitivity to lexical category boundaries related to the co-occurrence regularities, and (3) increased inductive bias in associating words with referents. How does coherence bring about these effects? Frank & Gibson (2009) have shown that *MNPQ* learning can be bolstered by easing working memory demands. Furthermore, there is evidence that novel words tax the memory system more, as they are encoded in terms of smaller phonological units (Treiman & Danis, 1988), so it is conceivable that the presence of semantically coherent known words reduced memory demands and thus improved *MNPQ* learning. We tested for this possibility in our data using mediation analyses. In particular, we tested whether memory scores mediated the effect of coherence on either (1) similarity scores or (2) referent assignment scores. In both cases, we found partial mediation. After controlling for memory, the regression coefficient relating coherence and similarity decreased significantly from 0.28 to 0.12, Sobel  $z = 7.74$ ,  $p < 0.05$ ; this reduced value was significantly greater than zero,  $t(657) = 3.60$ ,  $p < 0.0005$ , indicating partial mediation. After controlling for memory, the regression coefficient relating coherence and referent assignment score decreased significantly from 0.31 to 0.19, Sobel  $z = 5.19$ ,  $p < 0.05$ ; this reduced value was significantly greater than zero,  $t(651) = 3.67$ ,  $p < 0.0005$ , again indicating partial mediation. Thus, improved memory can explain some, but not all, of the increase in similarity and referent assignment scores due to semantic coherence.

Given this result, it is natural to ask whether known words *per se* can sufficiently ease memory demands so as to facilitate *MNPQ* learning, or whether they must have semantic coherence; we consider this question in Experiment 3. A different, though not mutually exclusive, possibility is that any relatively salient type of coherence, such as phonological co-

herence, is sufficient to facilitate distributional learning. We consider this possibility in Experiment 2.

## Experiment 2: Phonological Coherence

As noted previously, Lany and Saffran (2010) found evidence of successful *MNPQ* learning when co-occurrence regularities were perfectly correlated with a phonological property – the number of syllables in an *N/Q* word. We sought to compare phonological coherence with semantic coherence. Thus, in Experiment 2, we measured the result of a phonological manipulation in which all *M* words began with “r” and all *P* words began with “z”.

### Method

**Participants** 157 MTurk workers participated in the study.

**Design** The method was similar to that of Experiment 1. *M*’s were {*rull*, *rudge*, *ruck*} and *P*’s were {*zof*, *zerm*, *zabe*}.

### Results and Discussion

Results are shown in Figure 4. We compared the phonological condition results with the 0/3 and 3/3 conditions of Experiment 1 using a regression model with Helmert contrasts analyzing (i) the difference between the 0/3 and phonological conditions and (ii) the difference between the 3/3 condition and the 0/3 and phonological conditions combined.

**Memory** Phonological scores were not significantly different from 0/3 scores,  $t(466) = 0.96$ ,  $p > 0.05$  and both combined were significantly lower than 3/3 scores,  $\beta = 0.2$ ,  $t(466) = 3.44$ ,  $p < 0.001$ . Phonological efficiency was not significantly different from 0/3 efficiency,  $t(466) = -0.17$ ,  $p > 0.05$  and both combined were significantly lower than 3/3 efficiency,  $\beta = 0.0007$ ,  $t(466) = 2.76$ ,  $p < 0.01$ .

**Similarity** Phonological scores were not significantly different from 0/3 scores,  $t(466) = -0.16$ ,  $p > 0.05$ , and both combined were significantly lower than 3/3 scores,  $\beta = 0.14$ ,  $t(466) = 2.54$ ,  $p < 0.05$ . Phonological efficiency was not significantly different from 0/3 efficiency,  $t(466) = -0.29$ ,  $p > 0.05$ , and both combined were significantly lower than 3/3 efficiency,  $\beta = 0.0008$ ,  $t(466) = 3.2$ ,  $p < 0.005$ .

**Referent assignment** Phonological scores were not significantly different from 0/3 scores,  $t(466) = 1.47$ ,  $p > 0.05$ , and both combined were significantly less than 3/3 scores,  $\beta = 0.2$ ,  $t(466) = 2.35$ ,  $p < 0.05$ . There were no differences in efficiency (recall that this was also the case in Experiment 1). In terms of facilitating acquisition of the *MN* and *PQ* co-occurrence regularities, the phonological manipulation was indistinguishable from the 0/3 condition, and hence was markedly less effective than semantic coherence at the 3/3 level. It must be noted that the phonological regularity we introduced was the onset consonant (*r*- words versus *z*- words) applied to *M/P* words, whereas Lany and Saffran (2010) used syllable length (monosyllabic versus disyllabic words) applied to *N/Q* words, making direct comparison difficult. Presently, we have established that a particular kind

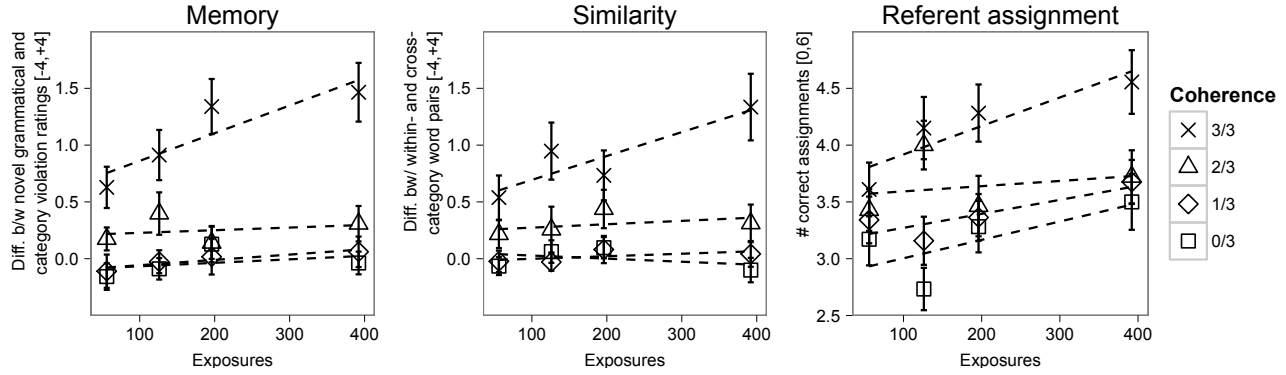


Figure 3: Experiment 1 results. Each plot shows data for one measure (memory, similarity, meaning) in Experiment 1. Dots show condition means, while dashed lines show the best-fitting linear trend.

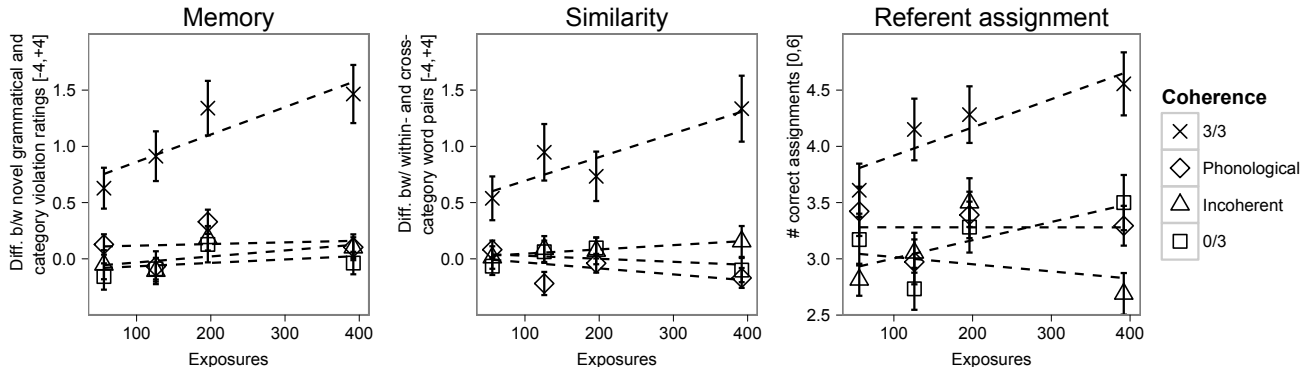


Figure 4: Experiment 2 and 3 results. Each plot shows data for one measure (memory, similarity, meaning) in Experiments 2 and 3. Dots show condition means, while dashed lines show the best-fitting linear trend.

of phonological regularity (onset consonant of a co-occurring word) is a far weaker correlated cue than semantic coherence.

### Experiment 3: Semantic Incoherence

The *M* and *P* words in Experiment 1 were all known and had semantic coherence. In Experiment 3, we explored whether coherence is necessary for facilitation of distributional learning, or whether the mere presence of known words is sufficient – that is, whether a semantically *incoherent* language facilitates distributional learning.

### Method

**Participants** 151 MTurk workers participated in the study.

**Design and Procedure** The method was similar to that of Experiment 2. The *M* and *P* words were known but did not adhere to any clear semantic organization. The specific *M* and *P* words were drawn from the pool {*shelf*, *glove*, *rain*, *leash*, *card*, *ball*}. In the referent assignment task, these known words were paired with images of the obvious referents, e.g., *card* with a picture of card.

### Results and Discussion

**Memory** Incoherent scores were not significantly different from 0/3 scores,  $t(460) = 0.056$ ,  $p > 0.05$  and both combined were significantly lower than 3/3 scores,  $\beta = 0.23$ ,  $t(460) = 3.95$ ,  $p < 0.0001$ . Incoherent efficiency was not significantly different from 0/3 efficiency,  $t(460) = 0.25$ ,  $p > 0.05$  and both combined were significantly lower than 3/3 efficiency,  $\beta = 0.0006$ ,  $t(460) = 2.55$ ,  $p < 0.05$ .

**Similarity** Incoherent scores were not significantly different from 0/3 scores,  $t(460) = -0.22$ ,  $p > 0.05$ , and both combined were significantly lower than 3/3 scores,  $\beta = 0.15$ ,  $t(460) = 2.47$ ,  $p < 0.05$ . Incoherent efficiency was not significantly different from 0/3 efficiency,  $t(460) = 0.723$ ,  $p > 0.05$ , and both combined were significantly lower than 3/3 efficiency,  $\beta = 0.0006$ ,  $t(460) = 2.54$ ,  $p < 0.05$ .

**Referent assignment** Incoherent scores were not significantly different from 0/3 scores,  $t(460) = 0.82$ ,  $p > 0.05$ , and both combined were significantly less than 3/3 scores,  $\beta = 0.23$ ,  $t(460) = 2.83$ ,  $p < 0.01$ . There were no differences in efficiency (recall that this was also the case in Experiments 1 and 2). The familiar but semantically incoherent linguistic

input appeared to have provided no benefit compared to the novel words of the 0/3 condition, suggesting that the presence of known words by itself does not aid distributional learning.

## General discussion

We have conducted the first large-scale systematic investigation of the effects of exposure and various correlated cues (semantic and phonological coherence) on distributional learning of word meanings. We have shown that semantic coherence aids distributional learning in the *MNPQ* regime far more than phonological (onset consonant) coherence. Additionally, our experiments indicate that *coherence* is necessary – semantically incoherent linguistic input provided virtually no benefit. Additionally, we showed that coherence works in part by alleviating memory limitations, though our data suggests that there may be aspects of distributional word learning not bottlenecked by memory resources.

We conjecture that word learners may be using semantic coherence to infer the topic of discourse and that this gist topic meaning influences the representations of co-occurring novel words (cf. the topic-learning models of Griffiths, Steyvers, & Tenenbaum, 2007). It may be through this process that people know *tort* to be a legal term and *transducer* to be an engineering term, despite not knowing the precise meanings of these words. Under this account, learning would be easier for words that occur in contexts high in semantic information and coherence. Thus, we would expect learning to have a “contiguous” character (faster learning for words occurring in familiar contexts than for words occurring in less familiar contexts), a possibility we plan to test in future work.

Our experiments highlight a limitation of artificial language learning paradigms. Researchers using *entirely* artificial languages may be severely limiting the power of distributional learning mechanisms, which our experiments show to be greatly enhanced by the presence of known words that adhere to some semantic organization.

Our work on distributional learning of semantic properties is in agreement with the extant literature on distributional learning of syntactic properties (viz. grammatical gender). Results from these studies (e.g., Brooks et al., 1993; Frigo & McDonald, 1998) indicate that children and adults fail to learn *MNPQ* categories without correlated cues, but they can learn given (e.g.,) correlated phonological markers. We believe that correlated cues – be they semantic, phonological, or otherwise – serve a common purpose: to reduce the space of possible categories.

## Acknowledgments

We thank Paul Thibodeau and Jay McClelland for helpful discussions.

## References

Braine, M. D. S. (1966). Learning the positions of words relative to a marker element. *Journal of Experimental Psychology*, 72(4), 532–540.

- Braine, M. D. S. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. In *Mechanisms of language acquisition* (pp. 65–87). Hillsdale, NJ: Lawrence Erlbaum.
- Brooks, P. J., Braine, M. D. S., Catalano, L., Brody, R. E., & Sudhalter, V. (1993, February). Acquisition of Gender-like Noun Subclasses in an Artificial Language: The Contribution of Phonological Markers to Learning. *Journal of Memory and Language*, 32(1), 76–95.
- Firth, J. R. (1957). A Synopsis of Linguistic Theory 1930–1955. In *Studies in linguistic analysis* (pp. 1–32). Oxford: Blackwell.
- Frank, M., & Gibson, E. (2011). Overcoming memory limitations in rule learning. *Language Learning and Development*, 7(2), 130–148.
- Frigo, L., & McDonald, J. L. (1998, August). Properties of Phonological Markers That Affect the Acquisition of Gender-Like Subclasses. *Journal of Memory and Language*, 39(2), 218–245.
- Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*.
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, 114(2), 211–244.
- Harris, Z. S. (1951). *Methods in structural linguistics*. Chicago: University of Chicago Press.
- Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological review*, 104(2), 211–240.
- Lany, J., & Saffran, J. (2010). From Statistics to Meaning: Infants’ Acquisition of Lexical Categories. *Psychological Science*, 21(2), 284.
- Mintz, T. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition*, 30(5), 678.
- Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4), 425–469.
- Reeder, P., Newport, E., & Aslin, R. (2010). Novel Words in Novel Contexts: The Role of Distributional Information in Form-class Category Learning. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.
- Riordan, B., & Jones, M. N. (2010). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2), 303–345.
- Smith, K. H. (1966). Grammatical intrusions in the recall of structured letter pairs: Mediated transfer or position learning? *Journal of Experimental Psychology*, 72(4), 580–588.
- Treiman, R., & Danis, C. (1988). Short-term memory errors for spoken syllables are affected by the linguistic structure of the syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1), 145–152.

# Grounding spatial language in non-linguistic cognition: Evidence for universal *and* relative spatial semantics in thought

Michael Pacer\* (mpacer@berkeley.edu)

Alexandra Carstensen\* (abc@berkeley.edu)

Department of Psychology, University of California at Berkeley  
Berkeley, CA 94720 USA

Terry Regier (terry.regier@berkeley.edu)

Department of Linguistics and Cognitive Science Program, University of California at Berkeley  
Berkeley, CA 94720 USA

## Abstract

The categories named by spatial terms vary considerably across languages. It is often proposed that underlying this variation is a universal set of primitive spatial concepts that are combined differently in different languages. Despite the inherently cognitive assumptions of this proposal, such spatial primitives have generally been inferred in a top-down manner from linguistic data. Here we show that comparable spatial primitives can be inferred bottom-up from non-linguistic pile-sorting of spatial stimuli by speakers of English, Dutch, and Chichewa. We demonstrate that primitives obtained in this fashion explain meaningful cross-linguistic variation in spatial categories better than primitives designed by hand for that purpose, and reflect both universal and language-specific spatial semantics.

**Keywords:** Language and thought; spatial cognition; semantic primitives; semantic universals; linguistic relativity.

## Spatial language and semantic primitives

Languages categorize spatial relations differently, and the significance of this cross-language variation for spatial cognition is a topic of ongoing debate (e.g. Bowerman & Pederson, 1992; Feist, 2000; Hespos & Spelke, 2004; Khetarpal et al., 2010; Levinson & Meira, 2003; Majid et al., 2004; see also e.g. Boroditsky & Gaby, 2010 on spatial structuring of non-spatial domains). This debate has traditionally pitted two views against each other. On the one hand, some have argued (e.g. Majid et al., 2004) that language structures spatial cognition, such that cross-language differences in spatial categorization cause underlying cognitive differences in speakers of those languages. On the other hand, others have suggested (e.g. Levinson & Meira, 2003) that the cross-language variation may reflect different partitions of a universal underlying conceptual representation.

An influential version of the universalist view holds that a set of *semantic primitives* (e.g. Wierzbicka, 1996) is universally available to human cognition, and that spatial categories in different languages can be obtained by composing such spatial semantic primitives in different ways. Feist (2000) proposed a set of spatial attributes characterizing cross-linguistic uses of spatial relations

similar to those expressed by *in* and *on* in English. She demonstrated that these primitives could be conjoined to form the linguistic spatial categories observed in diverse languages, accounting for both universal similarities and variation across languages. Xu and Kemp (2010) have since expanded on Feist's attributes, constructing their own set of universal primitives and demonstrating that conjunctions of these primitives can be used to describe a wide range of variation in the ways that languages partition the semantic space of spatial relations.

Primitive-based accounts have been demonstrated to characterize spatial terms across languages, capturing distinctions in both the intensional meanings and extensional uses of spatial words. However, the central issue in this debate is about cognition—not language. It is unclear whether these primitives accurately characterize the structure of thought as well.

Despite the assumption in the literature that semantic primitives are universal components of spatial cognition, these proposed units of thought have been developed and tested exclusively on the basis of linguistic data. Primitives of spatial cognition are typically inferred top-down from observations of cross-language variation in spatial terms, and evaluated on their ability to explain that variation. Importantly, this is generally done without direct reference to non-linguistic cognition. Consequently, we know little about how the semantic primitives of spatial language relate to spatial cognition.

While the primitives derived from language could plausibly account for variation in spatial cognition, they clearly suggest a subsequent inquiry: would it be possible to infer spatial primitives more directly from measures of nonlinguistic cognition? If so, would these cognitive primitives similarly account for the varying semantic systems across languages? Would they reflect language-specific influences as well?

## Spatial primitives in language and cognition

To determine whether proposals of semantic primitives are supported by direct evidence from nonlinguistic cognition, we would ideally want to obtain both cognitive and linguistic data from speakers of differing languages, extract primitives from the cognitive data of each group of

---

\*The first two authors contributed equally to this work.

speakers, and test the ability of these cognitive primitives to explain the linguistic spatial systems of their own and—most importantly—other languages. We could then compare the descriptive ability of these cognitive primitives to primitives previously proposed on the basis of cross-language data. If cognition-derived primitives from one group of speakers can account *well* for the spatial language of a different group this would provide support for the universalist account of semantic primitives.

However, it is not transparently obvious how to evaluate the performance of a set of primitives. One approach would be to compare their performance to that of previously attested linguistic primitives. However, comparing multiple representations raises the problem of accounting for the flexibility or representational power of differing representational systems. Correspondingly, primitives that accurately capture spatial semantics should explain language well *without* explaining random noise well. This gives us a criterion for testing cognitive primitives; if these primitives perform equally or better than language-derived primitives on a measure that accounts for representational power, then the cognitive primitives are doing well at characterizing *meaningful variation* in spatial semantic systems. Thus, if cognitive features derived from one set of speakers could account for another language's spatial system in such a way, this would support the universalist account.

In addition to testing support for this universal view, it is important to note that our examination is sensitive to linguistic relativity as well. If cognition-derived primitives from a group of speakers tend to explain the language of those speakers better than other languages, this would additionally provide support for linguistic relativity. For this reason, primitive-based proposals need not presume a universalist account. However, a fully relative account would also be difficult to support, as it requires that linguistic data is *always* best explained by cognitive data from speakers of the same language, as similarities in cognition across languages are theoretically limited to the extent that those languages overlap.

To examine the ideas described above, we obtain cognitive and language data through behavioral and linguistic tasks, respectively, in which speakers of various languages partition a set of spatial scenes into disjoint subsets through sorting and naming of the depicted spatial relations. We then infer primitives from the cognitive data in a neutral way, using an unsupervised, bottom-up statistical approach (additive clustering; see Lee, 2002). We evaluate whether these cognition-derived primitives support the semantic primitives account of spatial cognition and language by assessing the ability of cognitive primitives to explain variation in spatial language across a sample of diverse languages, in comparison to language-derived spatial primitives proposed in the literature. Finally, we address general implications for universal and relative views of language and thought, as well as suggestions specific to semantic primitives in spatial cognition.

To preview the results, we find that our spatial primitives

derived from cognitive data (1) explain semantic variation in language better than proposed primitives designed by hand for that purpose, (2) support both universal and relative views on spatial cognition, and (3) express generally coherent, but variably intuitive, semantic components of spatial relations.

## Methods

In order to compare primitives derived from cognition to those inferred from language in the literature, we drew on existing nonlinguistic cognitive data on spatial relations as a source for our primitives. We also incorporated existing spatial naming data, which the primitives attempt to explain. Here we briefly describe the prior collection of the cognitive and linguistic data by Khetarpal et al. (2009, 2010) and by Carstensen (2011). We then explain our process for inferring primitives from these cognitive data, and our procedures for testing the adequacy of these primitives in accounting for linguistic data.

**Participants.** A total of 24 native English speakers (Khetarpal et al., 2010), 24 native Dutch speakers (Khetarpal et al., 2009), and 38 native Chichewa speakers (Carstensen, 2011) took part in both the nonlinguistic and linguistic tasks, administered in their native languages and home countries of the United States, the Netherlands, and Malawi, respectively.

**Cognitive spatial task.** In each of the three studies from which we draw data (Khetarpal et al., 2009; 2010; Carstensen, 2011), participants sorted the 71 scenes in the Topological Relations Picture Series (TRPS; Bowerman & Pederson, 1992; see Figure 1 for examples) into piles based on the spatial relation depicted in each scene. Each scene showed an orange figure object positioned relative to a black ground object and participants were instructed to group the scenes into piles based on this spatial relation, such that the relation was similar for all cards in a given pile. Participants were informed that they could make as few or as many piles as they chose, rearrange their piles as they felt necessary, and could take as much time as they wanted.

**Linguistic spatial task.** In these previous studies (Khetarpal et al., 2009; 2010; Carstensen, 2011), after completing the sorting task, participants were asked to name the spatial relation depicted on each card. Labels picking out the target and ground objects were supplied in the participant's native language and the participant filled in the blank between these labels to complete a sentence specifying the figure's location in relation to the ground.

**Attested linguistic spatial categories.** In the linguistic spatial task, participants supplied terms or short phrases characterizing each spatial relation. Previous studies sanitized these data to collapse over responses that differed only in components without spatial meaning (e.g. variations in verb tense), leaving 88 unique spatial phrases supplied in English, 29 in Dutch, and 70 in Chichewa. For each phrase in every language, we recorded all scenes that the phrase was applied to at least once. These linguistic categories are



used as a target below to evaluate the ability of our primitives in describing categories in language. The attested linguistic categories also provide a standard for groupings of spatial scenes that are coherent and articulable in human language.

**Similarity from partitions of spatial scenes.** We use additive clustering (Lee, 2002) to infer cognitive primitives from nonlinguistic partitions (i.e., pile sorts) of spatial relations. Because this method operates over similarity matrices, we first create similarity matrices for each language based on the frequency with which speakers of that language co-sorted each pair of scenes.

These matrices reflect how often any two scenes were placed in the same pile by speakers of a given language. To create the matrix for language  $L$  for every pair of scenes  $(i, j)$ , we calculate the similarity value  $s_{ij}$  at row  $i$  and column  $j$  by counting the number of times each speaker of  $L$  placed scenes  $i$  and  $j$  into the same pile, and dividing this by  $|L|$ , the sampled number of speakers of language  $L$ .

**Additive clustering to derive primitives.** The spatial primitives proposed to underlie thought and compose categories in language have traditionally been designed by hand. In order to infer primitives from different languages in an unbiased and language-neutral way (e.g. unaffected by the researcher’s native language), we create primitives using an unsupervised clustering algorithm—*stochastic-optimized additive clustering* (Lee, 2002). This algorithm does not require that we assume a particular number of primitives and it has been used in the past to extract meaningful primitives from linguistic semantic partitions (Lee, 2002).<sup>1</sup>

The algorithm approximates  $s_{ij}$  with  $\hat{s}_{ij}$  for all  $i$  and  $j$ , minimizing  $\sum_{i < j} (s_{ij} - \hat{s}_{ij})^2$  under certain assumptions on how  $\hat{s}_{ij}$  is obtained. Specifically, we assume that objects possess a set of  $n$  underlying features, each of which is shared by a subset of the objects, and we assume that each feature has an associated positive weight or salience. The estimated similarity value for two items,  $i$  and  $j$ , is thus the sum of the weights of the features that those two items share (after scaling the weights to be between 0 and 1). That is, let  $\mathbf{1}_k(i)$  be the indicator function for feature  $k$  with weight  $w_k$  (i.e., it has value 1 if its argument has feature  $k$  and 0 otherwise), then  $\hat{s}_{ij} = \sum_{k=1}^n w_k \mathbf{1}_k(i) \mathbf{1}_k(j)$ .

Stochastic-optimized additive clustering uses a stochastic search which grows the set of primitives until the variance explained by adding further primitives fails to outweigh complexity afforded by adding those primitives. Through this process, we generate the features that we treat as cognitive primitives underlying the spatial scenes in the TRPS. That is, each primitive is the set of images defined

by a features indicator function (e.g., see Figure 1 below).

We apply this algorithm to the co-sorting matrices from all three languages, producing a set of primitives based on nonlinguistic cognitive data from speakers of each language. Because the algorithm is stochastic, running it multiple times will return different sets of primitives. To sample the space of potential primitive sets, we create 10 primitive sets for each language’s co-sorting matrix, making for 30 primitive sets in all.

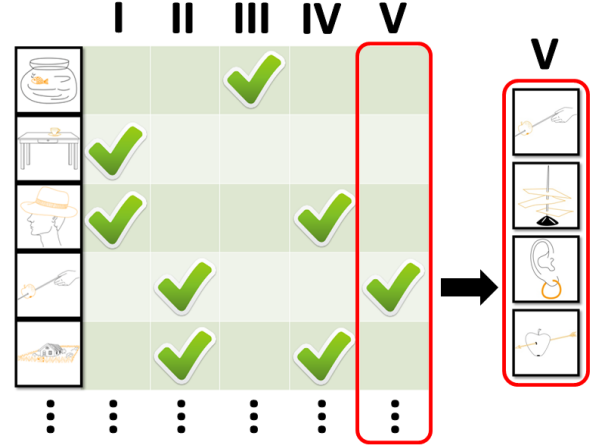


Figure 1: Additive clustering is used to produce a set of primitives which are clusters of “similar” images. Portions of five actual primitives derived from English speakers’ spatial scene sorting are presented above. The pop-out of primitive V shows all four of the 71 TRPS scenes that compose this particular primitive, which appears to characterize spatial relations that involve figure or ground piercing.

**Conjoined primitive sets.** A common assumption is that spatial primitives can be conjoined (e.g., “ $x$  is supported and in contact with  $y$ ”) to produce the spatial categories named by spatial terms in different languages (Feist, 2000; Xu & Kemp, 2010). Accordingly, we examine how the primitives we infer from pile-sort data can be conjoined to explain variation in spatial language across cultures.

We create and denote our conjoined primitive sets as follows. The base primitive set for a language is designated  $F(0)$ . Then,  $F(1)$  is the set of primitives formed by the union of  $F(0)$  and all conjunctions of two primitives from  $F(0)$ . Similarly, we create  $F(2)$  by including all the primitives in  $F(1)$  as well as all conjunctions of three primitives. Finally, we create  $F(3)$  from the union of  $F(2)$  and all conjunctions of four primitives.

**Defining distance.** In order to assess how well cognitive primitives account for spatial language, we need to determine their fit by defining a metric for the distance between sets of binary vectors with the same number of dimensions,  $d$ . First, we define a distance metric between pairs of binary vectors (e.g. primitives and linguistic categories) as being the city-block distance between those two vectors,  $f_1$  and  $f_2$ , which both have  $d$  dimensions (e.g. the 71 TRPS scenes, presence indicated by 1 and absence 0):

<sup>1</sup> Stochastic-optimized additive clustering requires a “precision” parameter describing how precise our similarity measures are, and a parameter that describes how much explanatory power is needed to warrant greater primitive set complexity. Because we used co-sorting as a proxy for similarity, we chose to use the least precise conventionally used value (i.e., 0.15) and the default complexity/simplicity trade-off value of six (Lee, 2002).



$$\text{dist}(f_1, f_2) = \sum_{i=1}^d |f_1(i) - f_2(i)|$$

This counts the scenes present in only one of  $f_1$  and  $f_2$ .

**Best-match analysis.** Though we have defined a distance measure between individual vectors, this does not explain how we measure the distance between *sets* of primitives and categories. Primitives are intended to be composed together to describe variation in linguistic categories across languages. Thus, we would want to create a distance metric that captures good performance on that measure across all primitives in a set.

Suppose that there is a set of primitives,  $P$ , and a set of linguistic categories,  $L$ . While the primitives and linguistic categories that make up these sets have very different interpretations and origins, they share a formal structure. That is, they are both binary vectors over the full set of images. This means that we can use the vector to vector distance as a measure of distance between an individual primitive and a single linguistic category.

Because we are attempting to explain a linguistic categorization system, one reasonable measure for the distance between  $P$  and  $L$  is to take the sum of the distance between the *best matching primitive* in  $P$  for every category in  $L$ . There are no constraints on how often a primitive may be used to explain linguistic categories; thus, this criterion will maximize the explanatory capabilities of  $P$  for  $L$ . Furthermore, each of the 10 primitive sets derived from a language’s pile-sort data will produce a best-match distance with a language  $L$ . We will consider only the value for the run with the lowest distance, since, arguably, by this criterion that run is the best run for explaining  $L$ .

**Primitives in the literature as a benchmark.** Because the description of linguistic spatial categories in terms of proposed universal primitives is a well-visited topic, previous proposed primitives provide a natural benchmark against which to test our cognitive primitives. Xu and Kemp (2010) describe a set of 19 primitives (e.g. “contact”) drawn from the wider literature and define the 71 TRPS images in terms of these primitives. To obtain definitions of scenes in terms of their primitives, Xu and Kemp asked three individuals to state whether each primitive applied to a given scene and assigned primitives to scenes based on majority vote. Using the same 19 primitives and 71 TRPS images, we replicated this procedure with three participants ( $\kappa = .91$ ) to obtain a set of primitives comparable to those described in Xu and Kemp (2010).

We considered both this primitive set and an expanded version consisting of the original 19 primitives together with negated versions of these primitives (i.e. the opposite, complementary set of scenes; e.g. the set of things that are ‘not in contact’) when appropriate, as determined by Xu and Kemp (2010). We found that in all cases the primitives with negation outperformed the simpler set<sup>2</sup>, and we therefore only report the performance of this expanded set. We use this set as a benchmark for evaluating the performance of

our primitives, as these primitives from the literature are hand-designed and generally considered to characterize semantic content across languages.

## Results & Discussion

The universalist account of semantic primitives holds that a set of conceptual primitives is universally available to human cognition, and spatial categories in varying languages can be created from different compositions of such spatial semantic primitives. To assess whether this view is supported directly by evidence from cognition, we derived sets of cognitive primitives from speakers of a sample of three diverse languages. We tested the ability of these cognitive primitives to explain variation in spatial language against previous proposals designed and tested on their ability to characterize such cross-linguistic data.

After creating 10 sets of primitives per language from the nonlinguistic pile-sorting of English, Dutch, and Chichewa speakers, we identified the best-scoring set of primitives from each language, making for three base sets of cognitive primitives. The fourth base set considered was the best-performing set of spatial primitives (with negations) from the literature (specifically from Xu and Kemp, 2010). From each base set, we derived a sequence of increasingly complex sets of features, by allowing increasing numbers of primitives to be conjoined together, as described above. We then recorded the distance of each primitive set in each sequence to the linguistic spatial systems of English, Dutch, and Chichewa, as a measure of how closely each primitive set characterizes variation in these languages.

Figure 2 shows that the distance scores for all four primitive base sets improve (i.e. provide a closer fit to the linguistic data) with the addition of conjunctions, affirming Xu and Kemp’s (2010) finding that conjunctions of primitives (linguistic in their case, cognitive in ours) can indeed provide for closer approximations to the categories in language. (Note, however, that because each further level of conjoined primitives contains the previous level, decrement was impossible and improvement very likely.) Although our primitives were derived from cognition and not hand-designed, like the language-derived primitives, their performance is generally comparable to these previously proposed primitives, substantially improving with the inclusion of the first level of conjunctions, but rapidly tapering off as more features are conjoined.

Notably, our primitives consistently outperform those from the literature in the base case or with pairs of intersections, revealing that they themselves are closer to the attested linguistic categories. At greater depths (i.e. with more conjunctions, and thus at the cost of representational complexity), the primitives from the literature are able to more closely approximate the linguistic categories of one of

<sup>2</sup> Interestingly, while Xu & Kemp (2010) did not find improved performance due to the inclusion of negative primitives, we found that in our comparison, including negated primitives offered large improvements over just the simple set of primitives.

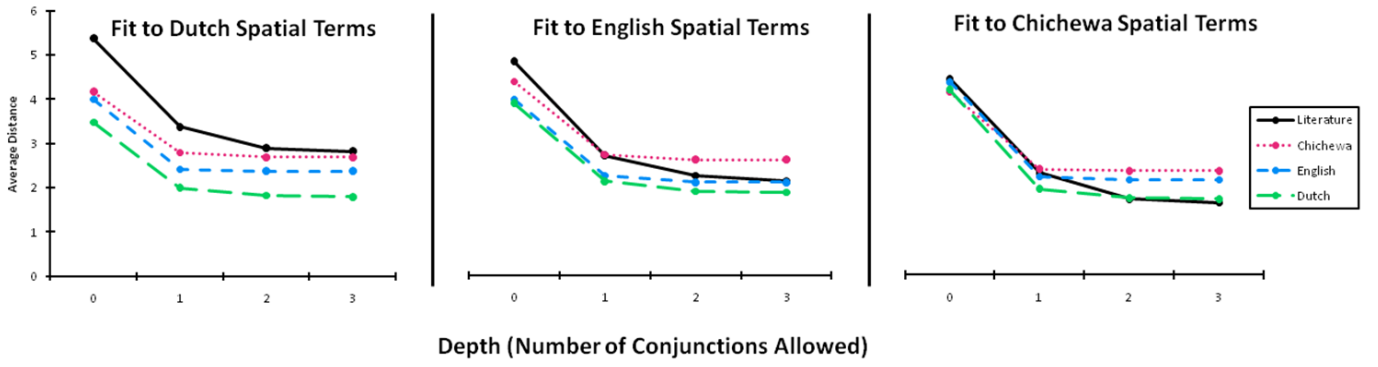


Figure 2: Average distance between literature-derived, Dutch, English, and Chichewa pile-sort-derived primitives, and the Dutch, English, and Chichewa spatial categorization systems.

our target languages (Chichewa) than are primitives derived from pile-sorting by speakers of any of the languages we considered. However for the other two target languages (English and Dutch), at least some of the sets of pile-sort derived primitives outperform those from the literature even at greater depths.

However, as previously discussed, representational power is an important mediating factor in the ability of these differing representational systems to account for cross-language variation. Thus, primitives that accurately capture spatial semantics should explain language well *without* explaining random noise well. To determine the amount of *meaningful variation* in spatial semantic systems that each feature set captures, we must correct for its ability to capture meaningless variation. We indexed this by creating 10 randomly permuted versions of each test language, where the size and structure of “linguistic terms” was preserved, but the specific spatial scenes included in each term were randomly swapped. We then measured the average fit of each feature set to these new nonsense “languages,” and corrected for varying representational power by subtracting the average distance from the feature set to a real language from the average distance between that feature set and the 10 permutations generated from that real language.

From this analysis, presented in Figure 3, it’s evident that the cognition-derived features explain semantic variation in language better than proposed primitives designed by hand for that purpose, in that they characterize considerably more of the *meaningful variation* in language, relative to nonsense variation. It is also apparent that these data provide support for the universal semantic primitives

account: cognition-derived primitives from *all* groups of speakers can account well for the spatial language of the other groups, relative to the comparison feature set hand-designed from cross-linguistic data. Simultaneously, we find that cognition-derived primitives from a given group of speakers *tend* to explain the language of those speakers better than other languages, providing support for accounts of linguistic relativity—although this is not always the case, suggesting some compromise between relative and universal forces in shaping these cognitive primitives.

**Semantic coherence.** Previously proposed spatial primitives were intended to capture cross-language variation, but were also intuitively designed to correspond to meaningful and easily describable semantic components that might underlie spatial cognition. A possible disadvantage of inferring primitives in an unsupervised manner (e.g. by additive clustering) is that this method may propose primitives that lack obviously meaningful interpretations.

Thus, having established that primitives derived from non-linguistic cognitive data can indeed be used to explain cross-cultural variation in linguistic spatial systems, we wished to also examine whether these primitives represent a similarly coherent grouping of spatial semantics. Here, we refer to the distance measures between the primitives themselves (i.e. at depth 0) and our attested categories in language, as an index of semantic coherence.

First, we find that 7.97% of categories in language correspond near perfectly with individual cognitive primitives, in that they either exactly match or have no more than one different scene (distance of 0 or 1). In comparison,

### Semantic Explanatory Power (for Actual Terms - Permutations)

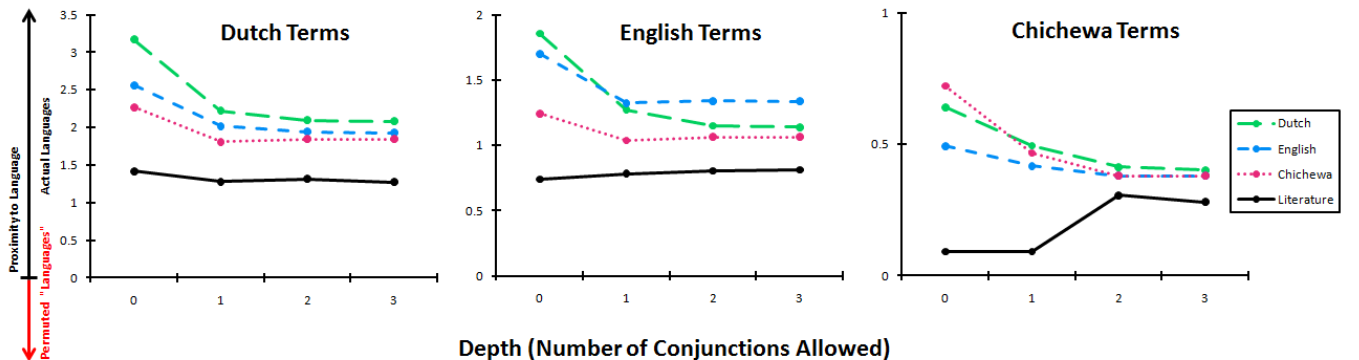


Figure 3: Relative proximity of literature-derived, Dutch, English, and Chichewa pile-sort-derived primitives to the Dutch, English, and Chichewa spatial categorization systems, compared to permuted versions, as a correction for representational power. All primitive sets are closer to real languages; the degree to which they are is the degree to which they can be taken to characterize *meaningful variation* in spatial semantics, rather than flexibly over-fitting noisy data.

the primitives from the literature match language categories near perfectly 5.82% of the time.

Second, the success of cognitive primitives in picking out articulable and coherent components of spatial relations is also apparent from a subjective evaluation of the primitives themselves. Figure 4 illustrates this point with two typical examples of actual cognitive primitives derived from English and Chichewa, which appear to be composed of spatial relations involving full or partial encirclement. While the primitives differ somewhat between languages, both express relatively clear and coherent spatial meanings.

Our analyses suggest the pile-sort-derived primitives represent semantically coherent, articulable components of spatial relations. In fact, these primitives match attested categories in language to a degree comparable with primitives designed by hand and surpass the hand-designed primitives in doing so when representational power is corrected for.

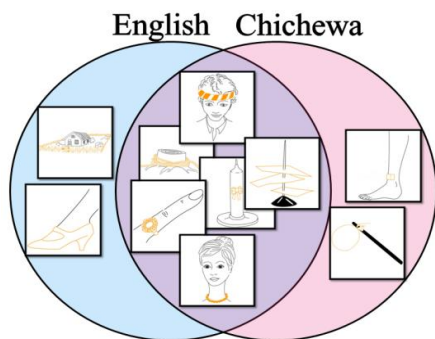


Figure 4: Example primitives derived from English and Chichewa speakers' cognitive data using additive clustering.

## Conclusions & Future Directions

We have shown that spatial primitives derived from non-linguistic pile-sort data account well for spatial terms across three languages. These primitives perform similarly to or better than hand-designed primitives from the literature. Furthermore, despite the unsupervised procedure used to derive them, these primitives reflect relatively coherent, articulable components of spatial cognition.

The present analyses suggest bottom-up inference may be a suitable method for generating spatial primitives. Further, the success of nonlinguistic cognitive data in explaining linguistic variation in spatial semantics supports the argument that universal primitives not only can be used to compose linguistic categories (as demonstrated previously), but may also be able to accurately characterize non-linguistic cognition. As an index of non-linguistic cognition, these primitives provide support for both universal and relative views on spatial cognition in showing that the cognitive primitives derived from one group of speakers can well account for the spatial language in another group, although nevertheless, these cognitive primitives do tend to more closely reflect the language of the speakers from whom they were derived, suggesting a simultaneous role of linguistically relative forces on spatial cognition.

Many questions remain open, suggesting directions for further research. Xu and Kemp (2010) found that allowing weighted primitives gave greater expressive capability to their model, and all distance metrics here were binary. Adapting our approach to weighted primitives, then, could result in improved fits overall, and could alter the general conclusions reached. Additionally, we did have to apply weak parametric assumptions to obtain our primitives, and thus an approach relying on Bayesian non-parametric methods would be beneficial—especially if it could work directly from the pile-sort partition data rather than over similarity matrices derived from the partition data. Finally, we have shown results for only three languages, two of which are from the same family and are closely related within that family. It would be informative to assess these ideas against a broader range of languages.

## Acknowledgments

We thank Asifa Majid and Naveen Khetarpal for kindly sharing their data. We also thank an anonymous reviewer for suggesting the permutation analysis. This work was supported by a Berkeley Fellowship awarded to MP as well as the NSF grant SBE-0541957, to TR and the Spatial Intelligence and Learning Center (SILC).

## References

- Boroditsky, L. & Gaby, A. (2010). Remembrances of times East: *Psychological Science*, 21(11), 1635-1639.
- Bowerman, M. & Pederson, E. (1992). Cross-linguistic studies of spatial semantic organization. In *Annual Report of the Max Planck Institute for Psycholinguistics 1992* (pp. 53-56).
- Carstensen, A. (2011). Universals and variation in spatial language and cognition: Evidence from Chichewa. Undergraduate thesis, University of California, Berkeley.
- Feist, M. I. (2000). On in and on: An investigation into the linguistic encoding of spatial scenes. Doctoral dissertation, Northwestern University.
- Hespos, S. J. & Spelke, E. S. (2004). Conceptual precursors to language. *Nature*, 430, 453 - 456.
- Khetarpal, N., Majid, A., Malt, B., Sloman, S., and Regier, T. (2010). Similarity judgments reflect both language and cross-language tendencies: Evidence from two semantic domains. In S. Ohlsson and R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.
- Khetarpal, N., Majid, A., & Regier, T. (2009). Spatial terms reflect near-optimal spatial categories. In N. Taatgen et al. (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.
- Lee, M.D. (2002). Generating additive clustering models with limited stochastic complexity. *Journal of Classification*, 19, 69-85.
- Levinson, S. C. & Meira, S. (2003). Natural concepts in the spatial topological domain—adpositional meanings in crosslinguistic perspective. *Language*, 79, 485-516.
- Majid, A., Bowerman, M., Kita, S., Haun, D., & Levinson, S. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, 8, 108-114.
- Wierzbicka, A. (1996) *Semantics: Primes and universals*. Oxford University Press.
- Xu, Y., & Kemp, C. (2010). Constructing spatial concepts from universal primitives. In S. Ohlsson and R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

# Elements of a rational framework for continuous-time causal induction

Michael Pacer (mpacer@berkeley.edu)

Thomas L. Griffiths (tom\_griffiths@berkeley.edu)

Department of Psychology, University of California, Berkeley, Berkeley, CA 94720

## Abstract

Temporal information plays a major role in human causal inference. We present a rational framework for causal induction from events that take place in continuous time. We define a set of desiderata for such a framework and outline a strategy for satisfying these desiderata using continuous-time stochastic processes. We develop two specific models within this framework, illustrating how it can be used to capture both generative and preventative causal relationships as well as delays between cause and effect. We evaluate one model through a new behavioral experiment, and the other through a comparison to existing data.

## Introduction

Causal induction plays a key role in human cognition, allowing people to identify the causal relationships that structure their environment. Recent work in cognitive science has resulted in many successful models of how people infer causal relationships from contingency data (Anderson, 1990; Anderson & Sheu, 1995; Cheng, 1997; Griffiths & Tenenbaum, 2005) and events that unfold in discrete time (Wasserman, 1990; Greville & Buehner, 2007). However, relatively few models have explored events that occur in continuous time. And yet, people regularly and easily reason about causal phenomena that evolve in continuous time (Michotte, 1963; Griffiths & Tenenbaum, 2009). Our understanding of causal inference would thus benefit from a framework capable of explaining human continuous-time causal inferences.

In this paper, we address this challenge by undertaking a rational analysis of continuous-time causal induction, in the spirit of Anderson (1990) and Marr (1982). We formalize the abstract problem posed by continuous-time causal inference, identifying a set of desiderata that a solution to this problem needs to incorporate. We then outline a framework that satisfies these desiderata, based on rational statistical inference over continuous-time stochastic processes. Our framework makes it possible to define both generative and preventative causes that unfold in continuous time, and to take into account delays between causes and effects.

With this framework in hand, we present two case studies from experimental psychology on human continuous-time causal inference. The first case study involves a novel experiment based on an experiment conducted by Griffiths and Tenenbaum (2005), allowing us to show how our framework can be used to infer whether a cause prevents events from occurring. The second case study is a re-analysis of an experiment on the effects of temporal information on human causal inference that was originally conducted by Greville and Buehner (2007). This second case study demonstrates the value of being able to use delay distributions to characterize how the effect of a cause changes over time.

We begin with a brief overview of previous work on the role of time in human causal inference, focusing on Griffiths and Tenenbaum (2005) and Greville and Buehner (2007). We then lay out a set of desiderata for a computational framework for human continuous-time causal inference. We go on to describe formally how we implement these desiderata in our proposed framework. Following this, we apply the framework to our two case studies, evaluating models that use preventative causes and delay distributions. Finally, we conclude and suggest directions for future work.

## Continuous-time causal induction

Studying the role of time in causal induction has a long history in cognitive science. One of the earliest established findings in the study of human causal inference is our ability to perceive causal relations in collisions, which is highly dependent on precise timing (Michotte, 1963). More recently, Buehner and colleagues have been very active in studying causal inference as it interacts with temporal information (e.g., Buehner & May, 2003; Greville & Buehner, 2007). However, while studies of causal induction often present events to participants in continuous time, they are typically analyzed using the discrete trial structure which the researchers used to design the stimuli (e.g., Anderson & Sheu, 1995; Wasserman, 1990).

Nonetheless, it may be enlightening to treat events as if they occurred over continuous time. Models considered by Griffiths and Tenenbaum (2005, 2009) take this approach, treating events that occur in continuous time as existing in a continuous dimension, or analyzing summaries of events as if they had occurred during a continuous interval. Even stimuli that are explicitly designed to convey information in discrete time (e.g., Greville & Buehner, 2007) can be analyzed in terms of continuous time by integrating over the time intervals. In the remainder of this section we summarize results from two studies on causal induction with temporal information, providing context for our later analyses.

### Causal induction from rates

Griffiths and Tenenbaum (2005) showed that people are capable of reasoning about causes that increase the rate at which events occur over continuous time, and their judgments are in close accordance with the predictions of a computational model engaging in continuous-time causal inference. In their experiments, participants observed a series of results that they were told came from physics experiments studying whether different electrical fields cause different radioactive compounds to emit more or fewer particles (the compound always released particles at some rate). For each “experi-

ment”, participants were told how many particles were emitted during one minute when the electrical field was on and one minute when the field was off. Participants then indicated the degree to which they endorse the claim that the field caused the compounds to emit more particles on a scale of 0 (the field definitely does not cause the compound to decay) to 100 (the field definitely does cause the compound to decay).

### Causal induction from tabular displays

Greville and Buehner (2007) demonstrated that the temporal distribution of event occurrences will alter people’s causal judgments, even if the relative frequencies of the occurrence of the effect in the presence or absence of a cause is held constant. Their purpose was to show that “temporal regularity” influences people’s judgments above and beyond mere contingency information. Their experiments used a tabular format to display events that unfolded over five days (split up into five segments of one day each), reporting in which day events occurred. This discretization allows the use of traditional models of causal inference which infer causes on the basis of contingency information.

In each condition of Greville and Buehner’s experiments, participants were shown two groups of 40 bacterial cultures, one group which was exposed to radiation and one which was not. Participants were shown (in tabular format) on which of 5 days each batch of bacteria died (if they died). Participants were asked to rate the effect of the radiation on a scale of –100 to 100 where –100 meant that the treatment was very effective at killing the bacteria, while 100 meant that the treatment was very effective at preventing the bacteria from dying (a rating of 0 meant that the treatment had no effect).

Greville and Buehner asked each participant about 18 pairs of tables, which differed in the frequency and distribution of times of death. In particular, Greville and Buehner varied the number of cultures dead by day five and the distribution over the times at which the bacteria died. They first fixed the number of deaths that would occur in each table. In all conditions, the time distribution for the bacteria not exposed to radiation was such that each of the deaths occurred with equal probability in any of the five days. However, for the bacteria exposed to radiation there were three time-of-death distributions: “strong contiguity”, in which bacteria death was more likely in the first few days after the radiation treatment; “weak contiguity”, in which bacteria died more often later in five day period; and “random”, in which bacteria death was uniformly distributed among the five days. Contingency information was held constant while varying contiguity. The results of the experiments showed that temporal information dramatically affects human causal inference.

### Defining desiderata for the framework

The studies discussed in the previous section illustrate some of the great variety in the phenomena to be considered by a framework for continuous-time causal induction. Thus, it will be helpful to identify the most vital features for allowing the

framework to capture a wide class of these cases. The following sections detail an important set of these properties.

**Intervention.** The framework should be capable of considering interventions in the sense meant in causal graphical models (Pearl, 2000). That is, an intervened upon node is said to be rendered independent of its parent nodes.

**Instantaneous and interval causes.** The framework should include both causes that exist instantaneously as well as over intervals of time.

**Generative and preventative causal relations.** It is vitally important when modeling human causal inference to distinguish between causes that generate effects and causes that prevent effects (Griffiths & Tenenbaum, 2005, 2009). People make dramatically different predictions based on which type of relationship they are looking for. Thus, we would want the framework to be capable of doing the same. In discrete time, Griffiths and Tenenbaum (2005) used the Noisy-OR and Noisy-ANDNOT logic gates to represent a cause that generates or prevents effects with reference to a background rate of the effects’ occurrence. Because these discrete time parameterizations will not hold in continuous time, we will have to redefine what we mean by a generative and a preventative relation for continuous time.

**Delay distributions.** In most models of causation that work in discrete time or over trials in which events occur simultaneously, a cause can only influence an effect if and only if that cause is present. This is undesirable if we are to develop a framework for continuous-time causal inference. Not only would it be useful to track how a cause’s influence changes over time, instantaneous events occur for only an infinitesimal period of time. Thus, in order for such events to have any effect on other variables they must be able to exert influence even after they are no longer present. Thus, we will need to characterize *delay distributions*, which define how a cause’s influence on its effects changes over time.<sup>1</sup>

### A framework based on Poisson processes

To form a rational framework encompassing these desiderata, we draw from the wide literature in statistics and computer science on continuous-time stochastic processes. In particular we pay attention to one class of continuous-time stochastic processes: Poisson processes. Poisson processes provide an excellent starting ground for generalizing causal graphical models (and hence intervention) as they define a series of independent random variables indexed over continuous time, being the continuous analogue of the independent Bernoulli events that take place on a series of discrete trials in many causal graphical models (Griffiths, 2005).

In its simplest sense, a Poisson process is a stochastic process (i.e., a series of random variables) that defines the rate at which instantaneous events occur over continuous time. That rate is determined by the rate function  $\lambda(t)$ . If a set

<sup>1</sup>This notion of change over time is not meant to capture that described in Rottman and Ahn (2009) where the change occurs over successive presentations of the cause, but change associated with temporal distance to one presentation of the cause.

of events are produced by a Poisson process, the probability that a certain number of events ( $k$ ) occurred in a time interval  $[t_0, t_1]$ ,  $0 \leq t_0 < t_1$  is,

$$P[(N(t_1) - N(t_0)) = k] = \frac{e^{-\lambda_{t_0, t_1}} (\lambda_{t_0, t_1})^k}{k!},$$

where  $\lambda_{t_0, t_1} = \int_{t_0}^{t_1} \lambda(t) dt$ .

The rate function defines the distribution of waiting times between events. For example, the waiting time before the first event ( $\tau_1$ ) is distributed  $P(\tau_1 = t) = \lambda(t)e^{-\lambda_0 t}$ .

Poisson processes have several desirable properties. If you have two independent Poisson processes with rates  $\lambda_1(t)$  and  $\lambda_2(t)$ , you can take the union of their event sets and this produces another Poisson process with rate  $\lambda_1(t) + \lambda_2(t)$ . This is a “superposition” of Poisson processes. Now, suppose the existence of some Poisson process  $PP_0$ , which has rate  $\lambda_0(t)$ . Suppose also that you have another function with the same support on  $t$  called  $\pi_1(t)$ , the range of which is a subset of  $[0, 1]$ . Then, for an event produced by  $PP_0$ , cancel that event (i.e., treat it as if it had never occurred) with probability  $\pi_1(t)$ . This procedure is called “thinning” the Poisson process  $PP_0$ , and the resultant Poisson process has rate  $\lambda(t)(1 - \pi(t))$ .

Poisson processes have been used to model aspects of continuous-time causal induction, including both causes that occur instantaneously and over intervals (Griffiths, 2005; Griffiths & Tenenbaum, 2005, 2009). However, this work has focused on generative causes and only explored a limited class of delay distributions. In the remainder of the paper, we show that this framework can address more of these desiderata. First, we define preventative causes within the Poisson process framework, evaluating the resulting model through a new behavioral experiment based on Griffiths and Tenenbaum (2005). We then introduce an extremely general approach to handling delay distributions, which we evaluate using the results of Greville and Buehner (2007).

## Generative and preventative causes

The properties of the Poisson process – specifically invariance of the form of the stochastic process under the superposition and thinning transformations – can be used to characterize generative and preventative causal relations. Suppose that there are  $i$  generative causes ( $\{C_i\}$ ) and  $j$  preventative causes ( $\{C_j\}$ ), and they exist over intervals of time. That is  $\forall C_a \in \{C_i\} \cup \{C_j\}, \exists T_a(C_a = 1) \subset \mathcal{T}$  where  $\mathcal{T}$  is the set of all non-measure-zero time intervals and  $T_a(C_a = 1)$  is the set of intervals during which  $C_a$  occurs. Let the Poisson process  $PP_0$  be a background rate of effect occurrence with an unknown time-invariant rate function  $\lambda_0 > 0$ . Causes assert their influence by altering the base-rate of the effect.

Generative causes will superpose themselves onto the background process, thereby increasing the rate of effect occurrence. That is, we can think of a generative cause  $C_i$  as producing a series of effects on its own, thereby inducing

a Poisson process  $PP_i$  with parameter  $\lambda_i(t)$ , where we assume that the cause only exhibits a non-zero effect when it is present (i.e.,  $t \in T_i(C_i = 1)$ ). That is, when  $C_i$  is present, the rate will be  $\lambda_0 + \lambda_i$ , and otherwise the rate will be  $\lambda_0$ . This is equivalent to a continuous-time version of the Noisy-OR logic gate, used in models of discrete-time causal inference (see Griffiths, 2005; Simma et al., 2008).

We will assume preventative causes will thin all Poisson processes that generate effects including both the background and generative processes. A preventative cause  $C_j$  will have thinning parameter  $\pi_j$  which affects the generative processes if and only if cause is present. Thus, if  $\lambda_{\text{total}}(t)$  is the total rate, when  $C_j$  is absent, the rate be  $\lambda_{\text{total}}(t)$ , but when  $C_j$  is present the rate will become  $\lambda_{\text{total}}(t)(1 - \pi_j)$ . This is equivalent to a continuous-time version of the Noisy-ANDNOT logic gate, which in the discrete-time setting defines the probability that an event will be canceled when the cause is present.

In the case where there are many causes, we will presume that they are independent and thus can be composed with one another, such that you will have a summation of the rates for the generative causes and a product of  $1 -$  the thinning parameters for the preventative causes. The rate function for a case with background rate  $\lambda_0$  and  $i$  generative causes and  $j$  preventative causes is defined as

$$\lambda(t) = \left( \lambda_0 + \sum_i \lambda_i \int_{T' \in T_i(C_i=1)} \delta(t, T') dT' \right) \prod_j (1 - \pi_j \int_{T' \in T_j(C_j=1)} \delta(t, T') dT'). \quad (1)$$

where  $\delta(\cdot, \cdot)$  is the Dirac delta function, which has an infinite spike where the two arguments agree and 0 elsewhere. Note, this does not include a prior for the causes; i.e., we treat these causes as continuous-time interventions.

The situation used as a cover story by Griffiths and Tenenbaum (2005) – determining whether electrical fields change the rate at which radioactive compounds emit particles – involves a system that can be analyzed using this model, where it has one effect (particle emissions) with a background rate and (possibly) one generative cause (the electrical field,  $C_i$ ). Griffiths and Tenenbaum presented participants with information summarizing the number of effect occurrences (particle emissions) that occurred during one minute with the cause on and one minute with the cause off. For each compound, participants rated on a scale of 0 (the electric field definitely does not cause the compound to decay) to 100 (the electric field definitely does cause the compound to decay) their belief regarding whether  $C_i$  was indeed a cause.

To model the participants’ predictions, Griffiths and Tenenbaum (2005) treated the problem as one of model selection between a graphical model  $G_0$  where the cause had no effect (i.e.,  $\lambda_i(t) = 0, \forall t$ ) and a graphical model  $G_1$  where the cause did have an effect (i.e.,  $\lambda_i(t) > 0, \exists t$ ). They parameterized  $G_0$  and  $G_1$  as we have above, as Poisson processes with different rate functions, where generative causes are treated



as we have treated them above. The quantity used to predict human judgments, termed “Causal Support”, was the log likelihood ratio in favor of  $G_1$ , integrating over the values of all of the parameters of the Poisson process. This model performed well at predicting the mean judgments of the participants, with a scaled correlation of  $r = .978, \alpha = .35$ .<sup>2</sup> Other models considered by Griffiths and Tenenbaum (2005) also performed well, with the raw difference in rates  $\Delta R$  (Anderson & Sheu, 1995) giving  $r = .899, \alpha = .05$ , a variant on the Power-PC theory (Cheng, 1997) giving  $r = .845, \alpha = .06$ , and a modified  $\chi^2$  score giving  $r = .980, \alpha = .01$ .

Griffiths and Tenenbaum (2005) considered only generative causes, creating the opportunity to use the same paradigm to evaluate whether the treatment of preventative causes outlined above is effective. We ran a new experiment to address this question. Considering only one preventative cause, we used nearly identical materials to Griffiths and Tenenbaum (2005), only changing the word “increases” to “decreases” and using the following  $(N(c^-), N(c^+))$  pairs (where  $N(c^-)$  and  $N(c^+)$  are the number of particles that were emitted during the minute when, respectively, the cause was absent and was present): (52, 2), (60, 10), (100, 50), (12, 2), (20, 10), (60, 50), (4, 2), (12, 10), (52, 50).

We recruited 18 participants through Amazon Mechanical Turk to participate in our study online. We asked each participant to make the following judgment about each of the nine cases: “Does this field decrease the rate at which this compound emits particles?” Participants responded on a scale ranging from 0 (the electrical field definitely does not decrease the rate of particle emissions) to 100 (the electrical field definitely does decrease the rate of particle emissions).

Following Griffiths and Tenenbaum (2005) we modeled this task as a model selection problem between two graphs  $G_0$  where the cause has no effect and  $G_1$  where the cause has a (preventative) effect on the rate of particle emissions. We used the model defined in Equation 1 with one potential preventative cause with parameter  $\pi_1$  and a background rate  $\lambda_0$  to define the likelihood functions for  $G_0$  and  $G_1$ . We assumed that in  $G_0$ ,  $\pi_1$  is constrained to be equal to 0. To obtain the log likelihood ratio, we need to provide likelihoods in terms of the graphical models (i.e.,  $P(D|G_0)$  and  $P(D|G_1)$  for the observed data  $D$ ). However, as they stand, the Poisson processes associated with these graphical models assume that the parameters  $\lambda_0$  and  $\pi_1$  are known, which is not the case. We thus need to define prior distributions over these parameters. With defined prior distributions, we can use Monte Carlo integration to obtain our marginal likelihoods, corresponding to the probability of the data given just the graphical model. We defined the prior for  $\pi_1$  as  $U(0, 1)$ , i.e., uniformly distributed in the interval  $[0, 1]$ . Griffiths and Tenenbaum (2005) used an improper prior for  $\lambda_0$ , with  $\lambda_0 \sim \frac{1}{\lambda_0}$ . We approximated the previously used prior by sampling  $v_0 \sim$

<sup>2</sup>As is usual in these studies, the authors scaled their model’s values with the non-linear transformation  $y = \text{sign}(x)|x|^\alpha$  where  $\alpha$  is chosen to maximize the linear correlation  $r$ .

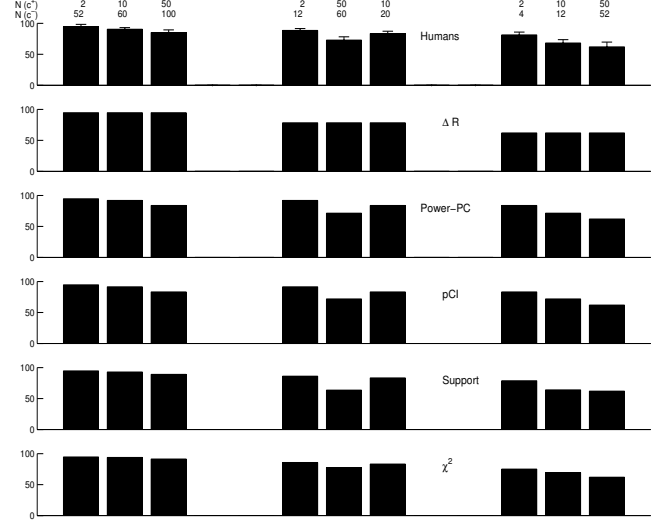


Figure 1: Preventative data for particle emissions: human responses and scaled model predictions. Support is the model that results from our framework.

$U(\log(10^{-6}), \log(10^6))$  and letting  $\lambda_0 = e^{v_0}$ .<sup>3</sup>

Using the log likelihood ratio in favor of the hypothesis that  $\pi_1 \sim U(0, 1)$  ( $G_1$ ) over  $\pi_1 = 0$  ( $G_0$ ) as our predictor of mean human judgments, we see a high scaled correlation with the results of the experiment: Causal Support gives  $r = 0.963, \alpha = 0.23$  (see Figure 1). We also evaluated the models tested by Griffiths and Tenenbaum (2005), which showed similarly high performance,  $\Delta R$  :  $r = 0.780, \alpha = 1.95 \times 10^{-4}$ ; Power PC:  $r = 0.986, \alpha = 0.45$  ; and  $\chi^2$  :  $r = 0.942, \alpha = 1.95 \times 10^{-4}$ . Our purpose is not to claim that the model we have defined is the best model of human inference, but to demonstrate that the assumptions we have made about handling preventative causes in our framework are reasonable. Future work will hopefully clarify whether this model outperforms the other models in cases where their predictions diverge more dramatically.

## Delay distributions

We will now describe how we implement delay distributions in our framework and apply the resultant model to modeling the results of Greville and Buehner (2007) – i.e., the bacteria death studies. We assume that generative and preventative causes have the same representation as above. We will assume that delay functions define what proportion of a cause’s influence remains an arbitrary amount of time after it occurs, where a base parameter defines the maximum influence of the cause.

Let  $f_i^g(\cdot, \cdot; \gamma_i)$  indicate the delay function with unknown parameters  $\gamma_i$  for generative cause  $C_i$ , and  $f_j^p(\cdot, \cdot; \theta_j)$  indicate the delay function with unknown parameters  $\theta_j$  for preventative cause  $C_j$ . Let the set  $\{t_k\}$  be the set of times that instantaneous cause  $C_k$  occurs and  $\{[t_{l,0}, t_{l,1}]\}$  be the set of

<sup>3</sup>To see the approximation, note that  $v_0 = \log(\lambda)$  and  $v'_0 = \frac{1}{\lambda}$  and use a change of variables to find  $f_{\lambda_0}(\cdot)$ .



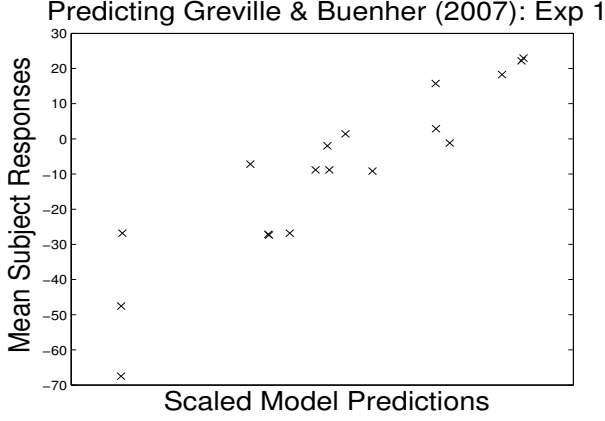


Figure 2: Model predictions for Greville and Buehner (2007), Experiment 1.

times over which interval cause  $C_l$  occurs. We set  $\lambda_0$  to indicate the underlying rate of an effects' occurrence, and  $\lambda_i$  and  $\pi_j$  to indicate the maximum value of a cause's influence.

The effects of delay distributions can be accommodated by defining a Poisson process with rate

$$\lambda(t) = \left( \lambda_0 + \sum_{C_i^g} \left( \sum_{t' \in \{t_i\}} \lambda_i f_i^g(t, t'; \gamma_i) + \sum_{[t'_0, t'_1] \in \{[t_{i,0}, t_{i,1}]\}} \lambda_i f_i^g(t, [t'_0, t'_1]; \gamma_i) \right) \prod_{C_j^p} \left( \prod_{t'' \in \{t_j\}} (1 - \pi_j f_j^p(t, t''; \theta_j)) \prod_{[t'_0, t'_1] \in \{[t_{j,0}, t_{j,1}]\}} (1 - \pi_j f_j^p(t, [t'_0, t'_1]; \theta_j)) \right) \right)$$

where  $f(t, [t_0, t_1]; \cdot)$  is the convolution of  $f(t, x; \cdot)$  with the boxcar function on  $[t_0, t_1]$  (i.e., the function that takes the value 1 for all  $x \in [t_0, t_1]$  and 0 otherwise). This allows us to keep the expressivity needed to capture our first findings, while allowing greater generality in the types of delay distributions applicable to interval causes.

Modeling the studies in Greville and Buehner (2007) requires further formal specification. Because events in these experiments were deaths they happen only once. As such, we only consider a bacterium to have died on the first arrival in a Poisson process defining the rate of death (i.e.,  $p(\tau_1 = t) = \lambda(t)e^{-\lambda_0 t}$ , where  $\lambda_{a,b} = \int_a^b \lambda(t)dt$ ). But, we do not know the precise time at which the bacterium died, merely the day on which it died. Therefore, the likelihood that bacterium  $i$  died on day  $t_{i,1}$  is  $\int_{t_{i,0}}^{t_{i,1}} \lambda(t) e^{-\lambda_0 t} dt = e^{-\lambda_0 t_{i,0}} - e^{-\lambda_0 t_{i,1}}$ , where  $t_{i,0}$  is the day before  $t_{i,1}$ . Finally, we model the 80 bacterial cultures in each condition as 80 conditionally independent Poisson processes given an underlying graph, i.e.,  $p(D|G) = \prod_{i=1}^{80} \int_{t_{i,0}}^{t_{i,1}} \lambda(t) e^{-\lambda_0 t} dt$ .

Because Greville and Buehner (2007) asked participants to respond on a scale of -100 (the radiation definitely causes

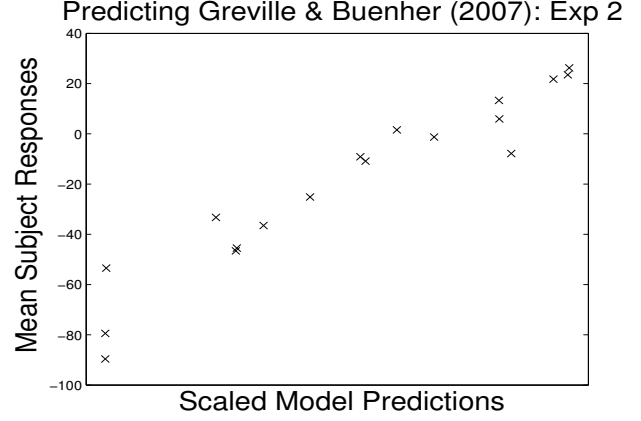


Figure 3: Model predictions for Greville and Buehner (2007), Experiment 2.

death) to 0 (the radiation has no effect) to 100 (the radiation definitely prevents death), we have effectively three graphs to choose from:  $G_g$ , the generative graph (where  $\pi_1 = 0$  and  $\lambda_1 \in \mathbb{R}^+$ );  $G_p$ , the preventative graph (where  $\lambda_1 = 0$  and  $\pi_1 \in [0, 1]$ ); and  $G_0$  the null graph (where  $\pi_1 = \lambda_1 = 0$ ). As in Griffiths and Tenenbaum (2009), we modeled participants' mean responses in each condition as  $P(G_p|D) - P(G_g|D)$ , assuming all three graphs are a priori equally likely.

We assumed a scaled exponential decay function with parameter  $\phi_1$  is used for both generative and preventative causes (i.e.,  $f_i(t, t'; \gamma_i) = f_j(t, t'; \theta_j) = e^{-\phi_1(t-t')}$  where  $t'$  is the time that a cause occurs). Because the radiation is only applied to the bacteria once at the beginning of the five days, for  $G_g$  and  $G_p$ , the only occurrence of the cause is instantaneous and appears at  $t = 0$ . Thus, for the generative graph,

$$\lambda_{0,t} = \int_0^t \lambda_0 + \lambda_1 f_1(s, 0; \gamma_1) ds = t\lambda_0 + \frac{\lambda_1}{\phi_1}(1 - e^{-t\phi_1}),$$

and for the preventative graph,

$$\lambda_{0,t} = t\lambda_0(1 - \frac{\pi_1}{\phi_1}(1 - e^{-t\phi_1}))$$

Similar to before, as a prior for  $\lambda_0$  we used  $v_0 \sim U(\log(10^{-1}), \log(10^1))$  and set  $\lambda_0 = e^{v_0}$ . The remaining priors were defined as  $\pi_1 \sim U(0, 1)$ ,  $\lambda_1 \sim \Gamma(1, \lambda_0)$ , and  $\phi_1 \sim \Gamma(1, \lambda_0)$ , where the priors are defined in terms of  $\lambda_0$  such that they inherit the scale defined by  $\lambda_0$ .

Using Monte Carlo integration, we calculated our model's judgments  $P(G_p|D) - P(G_g|D)$  for the data in Greville and Buehner (2007). Because the experiments used slightly different methods we evaluated our model predictions separately for each experiment but concurrently for all 18 conditions within each experiment. Our model has a scaled correlation of  $r = .910$  ( $\alpha = 2.74$ ) with mean participant responses in Experiment 1 and a scaled correlation of  $r = .957$  ( $\alpha = 1.72$ ) with mean participant responses in Experiment 2. Since submitting this paper, we learned of another model which outperforms our own – namely that described in Buehner (2006).

This was used to analyze the same data and had an excellent linear fit for the two experiments,  $r = .97$  and  $.953$  (Buehner, 2006). In these cases, our models make very similar predictions thus, we will need to explore more complex experimental scenarios (e.g., trials with multiple exposures to the cause at different times). This would put predictions from these models in starker contrast.

## Conclusions and future directions

Continuous-time causal induction is so pervasive that we often go about not even noticing that we are engaging in it. The richness of temporal information surely aids people as they infer causes in their everyday life. Here we have developed a rational framework that makes use of that same wealth of information. The framework is based on an extension to causal graphical models to include continuous-time stochastic processes – specifically Poisson processes. We demonstrate in two case studies that our framework is capable of accurately predicting human judgments in tasks that require reasoning about preventative causes and reasoning about delay functions. This extends previous work on Poisson processes as rational models of continuous-time causal induction.

Continuous-time stochastic processes are a very rich class of mathematical objects, and we expect that the formal framework we have outlined will grow more powerful as further tools are added to it. Fortunately, there are currently many tools being crafted. Our hope is to develop inference algorithms that allow our framework to consider large, complex networks of causal variables and the relations between them, in the vein of Pearl (2000) and Simma and Jordan (2010). Additionally, it will add an additional layer of generality to develop an account for how instantaneous events can alter the states of events that occur over intervals of time. Currently we take the parameters of causes as fixed at all times – it is merely the influence of the cause on the effect that wanes. However, people are very capable of reasoning about causal relations that change their form over time (Rottman & Ahn, 2009), and as such explaining data of that sort may be essential for capturing the full range of human causal reasoning.

Developing a rational framework for continuous-time causal induction could potentially provide insight into other phenomena of human causal judgment. One of the advantages of taking a Bayesian approach to causal induction is its ability to form strong inferences from very small amounts of data. This feature may be essential in explaining why perceptual causality (where one makes a causal inference from a single piece of data) is extremely sensitive to subtle differences in timing (Michotte, 1963; Newman, Choi, Wynn, & Scholl, 2008). Finally, one of the central motivations for studying time in causation is a pervasive belief that the timing of events can unveil the direction of the underlying relationship (i.e., what causes what; Rottman & Keil, 2012). If people do believe this (even implicitly), then characterizing the role of continuous-time in causation would be absolutely necessary if we are to understand the full extent of the human

mind's capacity and propensity for causal inference.

Of course, all of this belies the fact that there are many phenomena on human causal reasoning that have yet to be studied. What any computational-level framework offers is the ability to develop new questions out of the formal principles that originally drove the design – even if those questions did not exist when the framework was formulated. If such an event occurs in the near future, it would be a pleasant thought to think that it could have been the effect of the work presented here; but whether that will occur – only time will tell.

**Acknowledgments.** This work was supported by a Berkeley Fellowship awarded to MP and grant number FA-9550-10-1-0232 from the Air Force Office of Scientific Research.

## References

- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R., & Sheu, C.-F. (1995). Causal inferences as perceptual judgments. *Memory & Cognition*, 23, 510-524.
- Buehner, M. (2006). A causal power approach to learning with rates. In *Proceedings of the 28th annual conference of the cognitive science society*.
- Buehner, M., & May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *The Quarterly Journal of Experimental Psychology Section A*, 56(5), 865-890.
- Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104, 367-405.
- Greville, W., & Buehner, M. (2007). The influence of temporal distributions on causal induction from tabular data. *Memory & Cognition*, 35, 444-453.
- Griffiths, T. L. (2005). *Causes, coincidences, and theories*. Unpublished doctoral dissertation, Stanford University.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51, 354-384.
- Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological review*, 116(4), 661.
- Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.
- Michotte, A. (1963). *The perception of causality*. New York: Basic Books.
- Newman, G., Choi, H., Wynn, K., & Scholl, B. (2008). The origins of causal perception: Evidence from postdictive processing in infancy. *Cognitive psychology*, 57(3), 262-291.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.
- Rottman, B. M., & Ahn, W. (2009). Causal learning about tolerance and sensitization. *Psychonomic Bulletin and Review*, 16(6), 1043-1049.
- Rottman, B. M., & Keil, F. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, 64(1), 93-125.
- Simma, A., Goldszmidt, M., MacCormick, J., Barham, P., Black, R., Isaacs, R., et al. (2008). Ct-nor: representing and reasoning about events in continuous time.
- Simma, A., & Jordan, M. (2010). Modeling events with cascades of poisson processes. In *International conference on uncertainty in artificial intelligence*.
- Wasserman, E. A. (1990). Detecting response-outcome relations: Toward an understanding of the causal texture of the environment. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, p. 27-82). San Diego, CA: Academic Press.

# Musicians are better at learning non-native sound contrasts even in non-tonal languages

Amy Perfors (amy.perfors@adelaide.edu.au)  
Jia Hoong Ong (jia.h.ong@student.adelaide.edu.au)  
School of Psychology, University of Adelaide, Australia

## Abstract

It is very difficult for adults to perceive phonetic contrasts in their non-native language. In this study we explored the effects of phonetic training for different populations of people (musicians and non-musicians) and with different kinds of phoneme contrast (timing-based, like the Hindi /g/-/k/ contrast, and pitch-based, like the Mandarin /i/-/i/ tonal contrast). We found that musicians had superior perception for both contrasts, not just the pitch-based one. For both phonemes, training had little to no effect. We consider the implications of this for first and second language acquisition. **Keywords:** phonetic learning; music perception; language acquisition;

## Introduction

Second language learning is a difficult task for a variety of reasons. Adults have difficulty with many aspects of language acquisition, including language processing (Clahsen & Felser, 2006) and certain aspects of syntax (e.g., Birdsong, 2006), but limitations in phonetic perception relative to infants and young children are especially strong and well-documented (e.g., Werker & Lalonde, 1988; Kuhl, 2004; Maye, Weiss, & Aslin, 2008). Phonological deficits can be found not just in perception, but in production and processing as well (e.g., Flege, 1995; Sebastián-Gallés & Soto-Faraco, 1999). Such deficits are sometimes thought to have cascading effects onto other aspects of language (Perani, 2005; Werker & Yeung, 2005; Perfors & Dunbar, 2010).

One striking aspect of adults' poor phonetic perception is that it is quite difficult to overcome it through training. There are various training regimes for teaching adults to learn a phonetic contrast that does not exist in their native language. Some rely on implicit learning of the phonemic categories based on distributional information (Maye & Gerken, 2001, 2002; Shea & Curtin, 2005; Hayes-Harb, 2007), while in others some form of feedback is given (e.g., Jamieson & Morosan, 1989; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002; Wang, Jongman, & Soreno, 2003; Golestani & Zatorre, 2004; Wayland & Li, 2008; Bradlow, 2008). These training regimes have rarely been compared directly, and are often used for different kinds of phonemes and with different goals. It is therefore still unclear precisely to what extent different kinds of training are effective and why.

Given the importance of phonetic perception, understanding why phonetic training works (to the extent it does) and how it can be improved (to the extent it doesn't) is a matter of some importance. One of the ways to explore this is by investigating the effectiveness of training on different kinds of phonemes as well as in different populations of people. Expanding this exploration is one of the central goals

of this paper. We perform two main manipulations, comparing the effectiveness of the same implicit distributional training method on different populations (musicians and non-musicians) as well as different phonemes (tonal and timing-based). We find that musicians show improved phonetic perception on all phonemes, and that any effects of training are smaller than these population differences. The implications of these findings for language acquisition and representation more broadly are considered in the discussion.

## Different phonemes, different populations

**Tonal vs timing-based phonemes.** A phoneme is the smallest unit in a language that forms a meaningful contrast, like the /b/ in *bat* and the /p/ in *pat*. Many phonemes are distinguished from another based on timing. For instance, one of the differences<sup>1</sup> between the English “g” and “k” sounds is the presence of voicing (i.e., the vibration of the vocal cords). English “g” and “k” differ in their voice onset time, or VOT, which refers to the time at which voicing begins and the vocal cords begin to vibrate. For the English “g” sound, voicing is immediate as soon as the tongue leaves the roof of the mouth; for “k”, there is a time delay between the release of the stop closure and the vibration of the vocal cords.

Hindi makes a further timing-based distinction that does not exist in English. The Hindi /g/ and /k/ differ according to the presence of *pre-voicing*, which is the occurrence of voicing during the silent interval during which the vocal tract is blocked. Because this distinction does not occur in English, both sound like a “g” to a native English speaker. It is possible to train English speakers to hear this distinction, although they are far below native-speaker proficiency even after the best training. One of the simplest techniques, though not the most common for adults, is implicit distributional training (Hayes-Harb, 2007; Maye & Gerken, 2000; Maye et al., 2008). In this type of training, described in more detail later, participants hear a bimodal distribution of phonemes whose peaks are centered around the two phonemes to be learned. Distributional training has been used to teach adults to distinguish these Hindi phonemes given as little as 10 minutes of exposure (Maye & Gerken, 2000; Perfors & Dunbar, 2010).

Another kind of phoneme occurs in tonal languages like Mandarin, which uses pitch to convey meaning. In such a language, the meaning of the syllable changes when it is spoken in a different tone. For instance, in Mandarin, *ma* in the high level Tone 1 ([mā]) and the rising Tone 2 ([má]) means *mother*

<sup>1</sup>The other difference between these two phonemes is aspiration, which refers to the presence of a puff of air after making a sound (/g/ is not aspirated but /k/ is). We do not consider aspiration here.

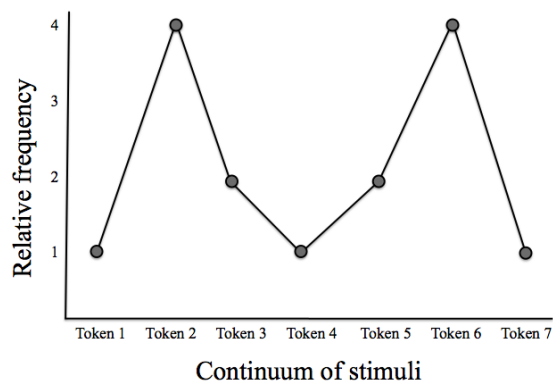


Figure 1: Distribution of stimuli used in phonetic training, defined along a continuum based on voice onset time in the HINDI condition and pitch in the MANDARIN condition. Tokens 2 and 6 occurred four times as often as tokens 1 and 7.

and *hemp*, respectively. People who are not native speakers of tonal languages process tone differently (Gandour, 1983). As with Hindi phonemes, there is evidence that non-native speakers can be trained to perceive lexical tones (Wang et al., 2003; Wayland, Herrera, & Kaan, 2010; Wong & Perrachione, 2007). However, these training programs generally take longer (days to weeks) and are more intensive than implicit distributional training, with participants being given explicit feedback and sometimes visual help (e.g., pitch graphs).

To our knowledge, no studies have explored the effectiveness of implicit distributional training on lexical tones. Our first goal in this study is therefore to compare the effects of distributional training (and baseline perception) of a Hindi timing-based contrast with a Mandarin tone-based contrast. Are both phoneme types equally easy for native English speakers to perceive? Does implicit training work better or worse with one kind of phoneme?

**Musicians vs non-musicians.** It is well-known that musicians consistently show a superior ability to learn lexical tones (Alexander, Wong, & Bradlow, 2005; Wong & Perrachione, 2007; Wayland et al., 2010). In many ways, this is no surprise: because both music and tonal languages involve pitch, extensive musical training (or superior auditory abilities) may result in increased sensitivity to pitch-related cues.

Much less is known about whether musicianship facilitates the learning of non-native contrasts that are not defined by pitch. While there are few studies investigating this issue, early evidence suggests that it might. For instance, Slevc and Miyake (2006) found that musical ability predicted Japanese speakers' ability to discriminate and produce the English /r/-/l/ contrast, and Sadakata, van der Zanden, and Sekiyama (2010) found that Japanese musicians were better than non-musicians at distinguishing the Dutch vowel /u/. There is also some evidence suggesting that musicians have higher brain-stem plasticity not just for musical stimuli, but for speech stimuli as well (Musacchia, Sams, Skoe, & Kraus, 2007).

For all these reasons, it seems reasonable to think that people with musical training might have superior perception of

timing-based phonetic contrasts as well as tones. However, to our knowledge this question has not been investigated before. Our second goal is therefore to compare the performance of musicians and non-musicians on the Hindi /g/-/k/ contrast. Are musicians better at perceiving that contrast as well as a Mandarin tonal contrast? How much does musicianship help (if it does) in either case? Are musicians more or less responsive to distributional training than non-musicians?

## Goals of the study

This paper addresses two main questions. First, we are interested in comparing performance on two different kinds of phonetic contrast within people given the exact same implicit distributional training. Is one easier than the other? Does training have more of an effect for one than another? Second, we are interested in comparing different populations of adults in their ability to perceive these two kinds of contrasts: namely, musicians and non-musicians. Do musicians have an advantage in perceiving timing-based contrasts as well as pitch-based ones? Does training have more or less of an effect on them than non-musicians?

## Method

There were two phases in this experiment, a training phase and a testing phase. Participants were randomly allocated to either a HINDI or MANDARIN condition. During the training, the participants were exposed to a distribution of sounds from the appropriate language for their condition. All participants participated in the same testing phase, which included two common tests of phoneme discrimination. The testing phase included stimuli from both Mandarin and Hindi, as well as a control set of English phonemes to ensure that they were paying attention. This design allows participants from each condition to serve as each other's control. For instance, the performance of participants in the HINDI condition on Mandarin stimuli reflects performance on those stimuli without having been trained on the Mandarin contrast.

**Participants.** 96 native English-speaking adults from the University of Adelaide and surrounding community participated in the experiment. 48 were classified as musicians according to criteria adapted from Wong and Perrachione (2007) and later slightly loosened in order to recruit enough participants. Musicianship in this study was defined as having had at least five continuous years of formal musical training, starting before the age of 15. The mean duration of musical training was 12.75 years for the musicians and 0.89 years for the non-musicians. There was no significant difference in duration of musical training between the HINDI and MANDARIN conditions ( $t(46) = 0.81, p = 0.4245$ , two-tailed).

**Training.** Participants in the HINDI condition were trained on a distribution of seven stimuli that differed according to voice onset time, while those in the MANDARIN condition were trained on a distribution of seven stimuli that differed according to pitch. The stimuli are described in detail below, but the training procedure is the same in all conditions. As

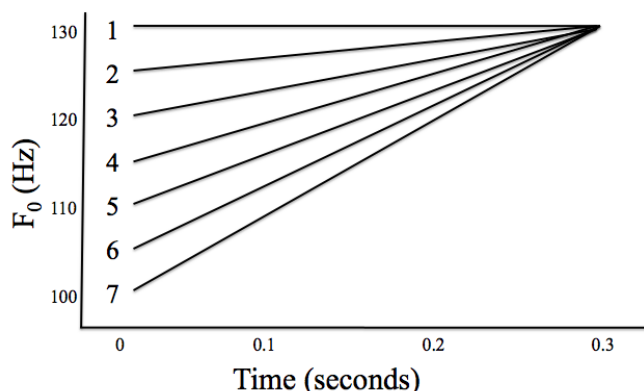


Figure 2: A seven-step continuum of the Mandarin syllable [ɿ]. Each step differed according to its fundamental frequency ( $f_0$ ) contour. (Figure adapted from Xu et al., 2006)

in Maye and Gerken (2000), we presented subjects with a bimodal distribution of these phonemes, as illustrated in Figure 1; thus, some tokens (e.g., 2 and 6) occurred four times as often as others (e.g., 1 and 7). Participants heard a total of 900 tokens presented in random order and separated by 250 ms each, for a total of approximately 10 minutes of exposure. During stimulus presentation the participants were told not to speak or read, but also that they need not consciously concentrate on the sounds. To alleviate boredom, they were allowed to doodle while listening.

**Testing.** After the training, the participants were presented with two standard discrimination tests, ABX and same/different (S/D); the order of the tests was randomized for each participant. In the ABX test, the participants heard three sounds separated by 1s and were asked whether the third (“X”) was the same as the first (“A”) or the second (“B”). In the S/D test, they heard a pair of sounds separated by 500ms and were asked whether the sounds were the same or not. Performance is given by the percentage of correct answers, and chance is 50% for both kinds of tests. Only tokens 1 and 7 were used in testing, since they most resemble the natural sounds of the language. Both tests consisted of 40 trials presented in random order, with 8 control stimuli corresponding to an English contrast, 16 stimuli corresponding to the Hindi /g/-/k/ contrast, and 16 corresponding to the Mandarin /ɿ/-/i/ contrast. The stimuli are described in more detail below.

**Stimuli.** Each of the two conditions were trained on distributions of stimuli taken from their respective languages – the HINDI condition on a contrast defined by timing, and the MANDARIN condition defined by pitch.

**HINDI.** The contrast used in this study was the unaspirated velar plosive voiced/voiceless contrast (/g/-/k/), which occurs in Hindi but not in English (both phonemes sound like a “g” to an English speaker). The /g/ and /k/ phonemes differ in terms of voice-onset time (VOT), such that /g/ contains a pre-voicing component while /k/ does not. It is therefore possible to gradually convert /g/ tokens into /k/ by successively removing parts of the pre-voicing component. Our training stimuli consisted of the Hindi syllable pairs [gɪ]-[kɪ], constructed by

recording a male native Hindi speaker saying [gɪ] and systematically removing the pre-voicing component using Praat phonetics software. This yields a continuum of seven stimuli from [gɪ] to [kɪ], separated by an average of 19ms in VOT from each other, and identical except for the pre-voicing.

Half of the 16 test trials used the [gɪ]-[kɪ] stimuli, with order of presentation of each and the side of the correct response counterbalanced. The other half of the test trials (also fully counterbalanced) consisted of the same contrast spoken in a different vowel context ([ga]-[ka]) and recorded by a female native Hindi speaker. The continuum of [ga] to [ka] was constructed in the same way as [gɪ] to [kɪ], although only tokens 1 and 7 were used during the test trials. For space reasons, we will report on overall performance among all of the test trials rather than on the two kinds of test trials individually.

**MANDARIN.** Participants were trained on a continuum bridging two tones, one of which is high level (Tone 1) and one of which is rising (Tone 2), as obtained from Xu, Gandour, and Francis (2006) and illustrated in Figure 2. The training stimuli consisted of a continuum between a vowel in Tone 1 ([ɿ]) and the same vowel in Tone 2 ([i]). A male native Mandarin speaker produced the syllable [ɿ], and its fundamental frequency ( $f_0$ ) contour was systematically altered to synthesize tokens on the continuum to Tone 2. This resulted in a seven-equal step continuum, as shown in Figure 2. All tokens have the same offset frequency (130Hz) and were normalized for duration and amplitude. As documented in Xu et al. (2006), three native Mandarin speakers judged tokens 1 and 7 to be good exemplars of the [ɿ] and [i] syllables.

Analogously to the Hindi stimuli, half of the 16 Mandarin test trials used the [ɿ]-[i] stimuli, with the order and side fully counterbalanced. The other half of the test trials consisted of the same tonal contrast spoken by a female Mandarin speaker with a different vowel ([ā] to [á]), constructed by the same method as the [ɿ]-[i] stimuli. As with the Hindi stimuli, we will report on overall performance among all of the test trials rather than on the two kinds of test trials individually.

**CONTROL.** In order to make sure that participants were attending to and understood the task, we included 8 trials of control stimuli during each of the two tests. These corresponded to a phonemic contrast they could already recognize: the dental plosive aspirated/unaspirated voiced/voiceless contrast (/d/-/tʰ/), which sound like “d” and “t” respectively to a native English speaker). Because the /d/-/tʰ/ contrast also exists in Hindi, the phonemes were recorded by the same male Hindi speaker as before.

## Results

Our study used two phoneme discrimination tests, the ABX and the S/D. Performance on these two tests was comparable: there was no significant difference in overall percent correct between discrimination tests (paired-sample  $t(383) = 1.33, p = 0.184$ ), with a mean performance of 70.5% (SD=20.4) in the ABX tests and a mean performance of 69.2% (SD=18.8) in the S/D tests. Moreover, the scores on

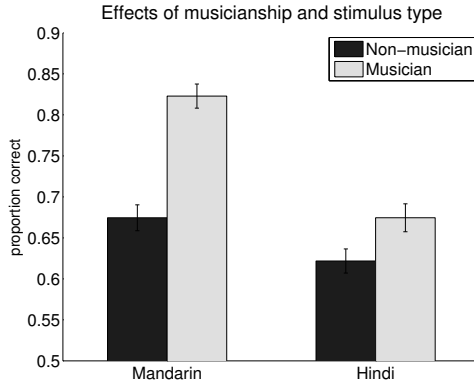


Figure 3: Overall accuracy by musicianship and stimulus type. Musicians showed superior performance to non-musicians, and performance was also higher on Mandarin than Hindi stimuli. Moreover, there was an interaction indicating that being a musician helped relatively more for Mandarin than Hindi stimuli.

the two tests were correlated ( $r = 0.50, p < 0.0001$ ). In all of the subsequent analyses we therefore collapse performance on the two tests into one overall accuracy score.

Our first question is whether discrimination performance is different on the Hindi and the Mandarin contrasts: is one easier than the other? Is there a differential effect of musicianship on each? We address this by considering performance on the stimuli for each language, collapsing (for now) any effects of training. As Figure 3 shows, musicians performed better than non-musicians for both types of contrast (two-way ANOVA,  $F(1,380) = 41.5, p < 0.0001$ ) and performance was higher on the Mandarin stimuli ( $F(1,380) = 41.5, p < 0.0001$ ). Moreover, there was an interaction, indicating that being a musician helped more for Mandarin stimuli than for Hindi stimuli ( $F(1,380) = 9.39, p = 0.002$ ).

How did the effects of musicianship play out within each stimulus type, and how did that interact with training? Were musicians or non-musicians helped more by training? Were the effects different depending upon the nature of the contrast in question? To address these issues we evaluate performance on the Hindi and Mandarin stimuli separately.

**Hindi stimuli.** Figure 4 shows overall accuracy on the Hindi test stimuli by musicianship and training. Participants were considered to have been trained on the stimuli if they were in the HINDI condition and untrained if they were in the MANDARIN condition. Musicians performed significantly better than non-musicians (two-way ANOVA,  $F(1,188) = 5.53, p = 0.019$ ), but there was no significant effect of training ( $F(1,188) = 1.71, p = 0.193$ ) and no interaction ( $F(1,188) = 0.049, p = 0.156$ ).

**Mandarin stimuli.** Figure 5 shows overall accuracy on the Mandarin test stimuli by musicianship and training (where people in the HINDI condition were considered to be untrained on the Mandarin stimuli). Here too, musicians performed significantly better than non-musicians (two-way ANOVA,  $F(1,188) = 46.8, p < 0.0001$ ). As before, there was no significant effect of training ( $F(1,188) = 0.001, p = 0.810$ ) and no interaction ( $F(1,188) = 0.71, p = 0.402$ ).

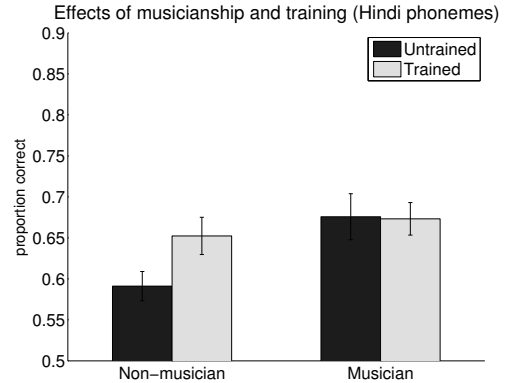


Figure 4: Overall accuracy on the Hindi phonemes by musicianship and training. Musicians did significantly better than non-musicians. Training did not significantly improve performance, although the trend among non-musicians approached significance.

## Discussion

Our overall findings show a strong effect of musicianship on phonetic perception: musicians had superior perception for both timing-based and pitch-based non-native contrasts (although the effect was much stronger for pitch-based contrasts). Interestingly, there was no effect of distributional training. Here we consider some of the implications of these findings for first and second language acquisition in general.

### Why did musicians perform better?

These results are consistent with the extensive literature documenting the fact that musicians have superior performance in linguistic tasks involving lexical tones (Alexander et al., 2005; Wong & Perrachione, 2007; Wayland et al., 2010). However, there is relatively little prior work showing that musicians have an advantage for non-tone-based phonemes, and none that we know of that investigates phonemes defined by differences in voice onset time (Slevc & Miyake, 2006; Sadakata et al., 2010). This suggests that whatever advantage musicians enjoy is not limited to differences in pitch perception, even though the timing-based advantages are smaller. Indeed, we even found a slight difference in performance between musicians and non-musicians on the control stimuli; although both groups performed extremely well (98.7% accuracy for the musicians, 95.2% accuracy for the non-musicians), the difference between the groups was significant ( $t(94) = 2.94, p = 0.004$ ). This is somewhat surprising, but it is true that even the control sounds were potentially confusable than more distinct phonemes would have been, and the difference is small in magnitude.

What is the root of the musician advantage? Consistent with their slightly better performance even on the control trials, one possibility is that musicians simply have a “better ear” in general – that is, they have superior auditory processing abilities overall. While this possibility is consistent with existing research (Schön, Magne, & Besson, 2004; Wong, Skoe, Russo, Dees, & Kraus, 2007; Musacchia et al., 2007), it does not really answer the question: *why* do they have superior abilities? Does musical training itself improve such

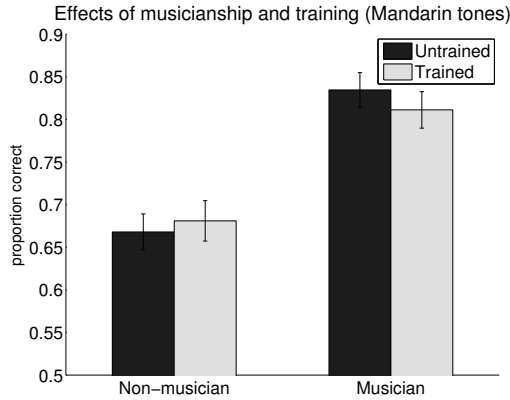


Figure 5: Overall accuracy on the Mandarin phonemes by musicianship and training. Musicians did significantly better than non-musicians, and training had no effect on performance.

abilities, or is it simply that people with better auditory processing become musicians in the first place?

This is a hard question to firmly disentangle, but our data provides one way to address it. We find that although the total duration of musical training is correlated with overall accuracy ( $r = 0.306, p = 0.002$ ), this effect is carried by the presence of non-musicians in the sample; there is no effect of duration of musical experience among the musicians only ( $r = 0.012, p = 0.932$ ). This occurs despite the fact that our sample of musicians was fairly diverse, ranging from people with 5 years to 45 years of training ( $M = 12.75, SD = 7.7$ ). It implies that perhaps at least part of the difference between musical and non-musical populations may be due to non-training-related differences in auditory perception. That said, since our participants played a wide range of instruments and were involved with music at different intensities – and any effects of duration are confounded with age – this is at best suggestive and should be interpreted with caution.

Could the better performance by the musicians be due to a motivational difference? While this might explain why the musicians performed better even on the control trials, it seems unlikely. The musicians were not told explicitly that they were being compared to non-musicians; they were only told that we were interested in how musicians learn language in general. It is unclear why they would be more motivated in such a scenario. Furthermore, our findings are consistent with the large amount of previous work showing superior phonetic processing in musicians (Alexander et al., 2005; Slevc & Miyake, 2006; Sadakata et al., 2010; Wayland et al., 2010). That said, in follow-up work we plan to investigate this possibility by recruiting visual artists, who would have the same motivational advantages (if any) that come from being recruited as a member of a special group, but who we would not expect to have superior auditory perceptual abilities.

### Why did training have no effect?

In addition to the musician advantage, our other main finding was that distributional training had no effect on the dis-

crimination of phonemes in our experiment.<sup>2</sup> Although a few previous experiments (Maye & Gerken, 2000; Hayes-Harb, 2007; Perfors & Dunbar, 2010) have found effects of distributional training for Hindi contrasts among non-musicians, these effects were small. Indeed, most training regimes for adults last far longer than 10 minutes and give reinforcement of some type, precisely because it is far more effective to do so (Jamieson & Morosan, 1989; McCandliss et al., 2002; Golestani & Zatorre, 2004; Bradlow, 2008).

The lack of training effect among musicians and for the Mandarin stimuli may have occurred for similar reasons. However, it also may be that implicit distributional training is not an effective means for teaching adults to discriminate pitch-based contrasts like the Mandarin /i/-/i/ distinction. This seems more plausible given the fact that even within the non-musicians – for whom there was non-significant trend of training for the Hindi contrast – there was no effect of training on the Mandarin contrast. But why would distributional training be more effective for one kind of contrast than another? We know that it is *possible* to train non-native speakers to perceive this kind of tonal contrast (Wang et al., 2003; Wayland et al., 2010; Wong & Perrachione, 2007). However, the training programs that have been successful have been far more intensive and explicit than ours was. We can only speculate, but perhaps this sort of instruction is necessary for people to understand the pitch distinctions they should be listening for. In contrast, distributional training may naturally focus people's attention on timing by playing one phoneme after another in rapid succession. That said, since distributional training had no statistically significant effect in any case, the simplest conclusion is that it was insufficient regardless of the nature of the contrast.

A related possibility is that our participants were already performing near their ceiling – that is, near the peak of what would be possible for them without decades of experience distinguishing the sounds in question. Perhaps the benefits of training for non-musicians in Hindi found in previous studies and implied here come from making people aware of the more obvious cues that can be used to distinguish the sounds. Due to their superior auditory skills, musicians may already be aware of those cues; and pitch differences are blatant enough that even non-musicians can hear some of the differences between /i/ and /i/. Of course, the difference between musicians and non-musicians, and between Hindi and Mandarin phonemes, also implies that if there is a ceiling effect, there are probably multiple different ceilings rather than only one.

Is there a ceiling effect? Given the widely-documented difficulty of training adults to recognize non-native contrasts,

<sup>2</sup>We did find that if we performed a post-hoc analysis on only non-musicians in the Hindi contrast, which is the only condition directly corresponding to the previous literature (Maye & Gerken, 2000; Hayes-Harb, 2007; Perfors & Dunbar, 2010), the training effect was significant ( $t(94) = 2.12, p = 0.037$ ). This suggests that had that been the only condition studied – analogous to the previous literature – it would have reached significance. However, since this analysis in our study followed an omnibus ANOVA with no main effect or interaction, it is not statistically appropriate to apply here.



this is certainly possible. After all, native speakers have decades of experience distinguishing between those sounds, and it would be quite surprising if these vast differences in exposure could be eliminated with a small amount of training, even among especially capable subjects like musicians. The notion of a ceiling effect for training is also consistent with the Native Language Neural Commitment hypothesis, which suggests that early experience results in changes in the brain that encode the phonetic contrasts of one's native language (Kuhl, 2004). Because of these neural changes, learning non-native contrasts as an adult is therefore extremely difficult. This would also explain why training made no difference among any of the musicians, who may have been already performing near the limit possible for brains that grew up dedicated to hearing other kinds of contrasts. If there is a ceiling, it has unfortunate implications for second language acquisition. Perhaps it is intrinsically limited by poor phonetic perception abilities, at least without years and years of exposure to the new language. That said, we must remind ourselves that the training approach used in this study was quite simple and short compared to other, more intensive methods, which might very well have more of an effect.

In many ways, this study raises more questions than it answers. Why did distributional training have no effect, particularly for musicians and Mandarin contrasts? What is the root of the musician advantage, and why does it extend to include timing-based contrasts as well as pitch-based ones? These questions are still open, but this work is an important step toward understanding the roots of phonetic perception and its relationship to both first and second language learning.

## Acknowledgments

Thank you to Natalie May, Angela Vause, and Daniel Carabellese for recruiting participants and running the experiments, and to Dan Navarro and the CLCL lab for useful discussions. We especially thank Dr. Jackson Gandour for generously sharing his Mandarin stimuli. AP was supported by ARC grant DE120102378.

## References

- Alexander, J., Wong, P., & Bradlow, A. (2005). Lexical tone perception in musicians and nonmusicians. In *9th European Conference on Speech Comm. and Tech.* Lisbon.
- Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Language Learning*, 56(1), 9–49.
- Bradlow, A. (2008). Training non-native language sound patterns. In J. Hansen Edwards & M. Zampini (Eds.), *Phonology and second language acquisition* (p. 287–308). Benjamins.
- Clahsen, H., & Felser, C. (2006). How native-like is non-native language processing? *Trends in Cognitive Sciences*, 10(12), 564–570.
- Flege, J. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 229–273). Timonium, MD: York Press.
- Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics*, 11, 149–175.
- Golestani, N., & Zatorre, R. (2004). Learning new sounds of speech: Reallocation of neural substrates. *NeuroImage*, 21, 494–506.
- Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, 23(1), 65–94.
- Jamieson, D., & Morosan, D. (1989). Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques. *Canadian Journal of Psychology*, 43(1), 88–96.
- Kuhl, P. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, 5, 831–843.
- Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. In *24th Annual Meeting of the Boston University Conference on Language Development*.
- Maye, J., & Gerken, L. (2001). Learning phonemes: How far can the input take us? In *25th Annual Meeting of the Boston University Conference on Language Development*.
- Maye, J., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82, B101–B111.
- Maye, J., Weiss, D., & Aslin, R. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, 11(1), 122–134.
- McCandliss, B., Fiez, J., Protopapas, A., Conway, M., & McClelland, J. (2002). Success and failure in teaching the [r]–[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience*, 2(2), 89–108.
- Musacchia, G., Sams, M., Skoe, E., & Kraus, N. (2007). Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. *Proceedings of the National Academy of Sciences*, 104(40), 15894–15898.
- Perani, D. (2005). The neural basis of language talent in bilinguals. *Trends in Cognitive Sciences*, 9(5), 211–213.
- Perfors, A., & Dunbar, D. (2010). Phonetic training makes word learning easier. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.
- Sadakata, M., van der Zanden, L., & Sekiyama, K. (2010). Influence of musical training on perception of l2 speech. In *11th Annual Conference of the International Speech Communication Association*. Chiba.
- Schön, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology*, 41, 341–349.
- Sebastián-Gallés, N., & Soto-Faraco, S. (1999). Online processing of native and non-native phonemic contrasts in early bilinguals. *Cognition*, 72, 111–123.
- Shea, C., & Curtin, S. (2005). Learning allophones from the input. In *29th Annual Meeting of the Boston University of the Conference on Language Development*.
- Slevc, L., & Miyake, A. (2006). Individual differences in second language proficiency: Does musical ability matter? *Psychological Science*, 17(8), 675–681.
- Wang, Y., Jongman, A., & Soreno, J. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustical Society of America*, 113(2), 1033–1043.
- Wayland, R., Herrera, E., & Kaan, E. (2010). Effects of musical experience and training on pitch contour perception. *Journal of Phonetics*, 36, 250–267.
- Wayland, R., & Li, B. (2008). Effects of two training procedures in cross-language perception of tones. *Journal of Phonetics*, 36, 250–267.
- Werker, J., & Lalonde, C. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology*, 24(5), 672–683.
- Werker, J., & Yeung, H. (2005). Infant speech perception bootstraps word learning. *Trends in Cognitive Sciences*, 9(11), 519–527.
- Wong, P., & Perrachione, T. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, 565–585.
- Wong, P., Skoe, E., Russo, N., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience*, 10, 420–422.
- Xu, Y., Gandour, J., & Francis, A. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *Journal of Acoustical Society of America*, 120(2), 1063–1074.

# Probability matching vs over-regularization in language: Participant behavior depends on their interpretation of the task

Amy Perfors (amy.perfors@adelaide.edu.au)  
School of Psychology, University of Adelaide, Australia

## Abstract

In a variety of domains, children have been observed to over-regularize inconsistent input, while adults are more likely to “probability match” to any inconsistency. Many explanations for this have been offered, usually relating to cognitive differences between children and adults. Here we explore an additional possibility: that differences in the social assumptions participants bring to the experiment can drive differences in over-regularization behavior. We explore this in the domain of language, where assumptions about error and communicative purpose might have a large effect. Indeed, we find that participants who experience less pressure to be “correct” and who have more reason to believe that any inconsistencies do not correspond to an underlying regularity do over-regularize more. Implications for language acquisition in children and adults are discussed.

**Keywords:** over-regularization; statistical learning; probability matching; language acquisition

## Introduction

In a variety of situations, humans given probabilistic input will tend to *probability match* – that is, they respond differentially in a way that is proportional to those probabilities. In learning theory, this occurs when people (or animals) choose a stimulus proportional to the relative number of times it has been reinforced (e.g., Herrnstein, 1961, 1970; Baum, 1979; Pierce & Epling, 1983; Wearden, 1983). In decision making, this occurs when people are asked to predict the next item in a sequence (e.g., a card drawn from a deck, or a flashing light) and respond by choosing proportionally to the frequency of that item in the past (e.g., Castellan, 1974; Shanks, Tunney, & McCarthy, 2002; Vulkan, 2000). And in language learning, this occurs when people given linguistic input that varies inconsistently (such as an affix or particle occurring only 60% of the time, for no apparent reason) and they produce that particle proportional to its frequency in the input (e.g., Hudson Kam & Newport, 2005, 2009).

Although these cases vary widely in many details, what is interesting in all of them is that an overmatching or maximizing strategy may often be the more sensible one. This strategy, also called over-regularization in the language literature, involves producing or responding to the most frequent item closer to 100% of the time, rather than in a way that is proportional to its probability or frequency. Such a strategy is more sensible for different reasons in different domains. One receives more reinforcement if one always chooses the more frequently-reinforced stimulus; one makes more successful predictions if one always chooses the most frequent item; and one minimizes the burden on the listener as well as the chance of miscommunication by removing linguistic variability when it serves no purpose. Given this, why do people probability match?

One clue may come from the literature on children’s behavior. Although children are less well-studied than that of adults, there is some evidence that preschool-aged children may overmatch or over-regularize more in a decision making or reinforcement learning context (Jones & Liverant, 1960; Derks & Paclisanu, 1967). This is consistent with the small amount of work on children in a linguistic domain suggesting that they are more likely than adults to over-regularize there as well (Hudson Kam & Newport, 2005, 2009). These findings are somewhat limited, since they are based on relatively few studies (especially in the area of language) and generally involve statistical rather than absolute differences: that is, more children over-regularize than adults, but some still do not. Nevertheless, they raise the intriguing possibility that whatever factor causes adults to probability match may play less of a role in preschool-aged children.

What might that factor be? A common hypothesis is that it is related to a cognitive change occurring between childhood and adulthood. One possibility is that preschool children have poorer metacognitive control – that is, they find it difficult to inhibit a previous response (Jones & Liverant, 1960; Weir, 1964), have a harder time monitoring and responding to conflict in their representations (Ramscar & Gitcho, 2007), and/or are more insensitive to the reward structure (Stevenson & Hoving, 1964). Another is that children’s poorer memory and/or processing abilities might result in over-regularization (Hudson Kam & Newport, 2005, 2009).

Although there is support for many of these possibilities, it is also possible that the difference arises, at least in part, for a more prosaic reason: adults and children may be given slightly different tasks, or have different interpretations of the same task. It is natural for minor task differences to arise as a natural by-product of the effort to make the same experiment apply to widely varying age groups. There were indeed small differences in all of the experiments in which differential over-regularization is observed (e.g. Jones & Liverant, 1960; Derks & Paclisanu, 1967; Hudson Kam & Newport, 2005, 2009). However, we are more interested here in the kinds of differences in interpretation that might arise even with precisely the same methodology. That is, different groups (like children and adults) might have different interpretations or different assumptions that they bring to the *exact same* task. As explained below, this may be particularly an issue in linguistic tasks, which are the focus of this work.

This paper explores one main question: is human behavior in a language-learning domain affected by changes in the goal or assumptions underlying the task? It is known that adults can be pushed away from probability matching by

varying the delay between changing a response and reinforcement (Baum, 1975), punishing people for some responses (Bradshaw, Szabadi, Bevan, & Ruddle, 1979), making one response easier (Bradshaw, Ruddle, & Szabadi, 1981), making rewards especially enticing (Shanks et al., 2002), or offering extensive corrective feedback (Shanks et al., 2002). However, none of these possibilities are relevant when the task is linguistic, since language learners generally receive little to no direct feedback or reinforcement, both in the real world (Pinker, 1989) and in this kind of task (e.g., Hudson Kam & Newport, 2005, 2009; Perfors & Burns, 2010; Perfors, 2011).

What goals or assumptions might adults and older children have that younger children do not, which cause the former to probability match but the latter to over-regularize? We hypothesize that adults and older children may feel a strong sense that there is a “right” answer, and a concomitant pressure to be pressure to be “correct” that younger children do not. We noticed in previous studies run in our lab (Perfors & Burns, 2010; Perfors, 2011) that adults reported a strong intuition that there must be some underlying reason behind the inconsistency in the linguistic input they received; for instance, one participant confided that they thought all of the “shiny things” were linguistically marked in the same way. In reality, there were no regularities between the inconsistent items and anything else in the experiment, but the participants did not know that and often tried to produce language in accordance with the regularities they thought they had observed.

Why did our adult subjects have such a strong intuition that there were regularities in the input? One possibility is that this is simply an intuition that all people typically bring to *any* language-learning scenario. It is not an unreasonable assumption; although some phenomena in language are truly arbitrary, much variation does occur for a reason. If people truly do always come to language learning tasks with this assumption, then we would expect adults to show the same behavior regardless of where they thought the language came from or what their goal in the learning task was. Conversely, this intuition may have been caused by, or at least exacerbated by, characteristics of the situation: being in an official lab, presented with stimuli that are clearly designed and non-accidental, and asked to learn about those stimuli all create the strong impression that there actually is some regularity there to learn. By contrast, real linguistic input contains many errors and inconsistencies that arise from the fact that it is produced on the fly, for communication, by real people. If young children are either more blind to the social pressures inherent in a lab-based experiment or more likely to interpret underlying irregularities in the input as errors, then this might be responsible for at least some of the observed difference in over-regularization between children and adults. After all, it is sensible to over-regularize inconsistencies if they do not hide some underlying regularity that you will be judged for missing.

A full test of our hypothesis would require us to manipulate children’s beliefs about the nature of the input they are

receiving and experiment they are in. This is very difficult to do, and it is even more difficult to evaluate whether it has been done successfully. Alternatively, we can test the assumptions underlying our hypothesis by investigating adults. If adults respond to task characteristics that remove a pressure for generating “correct” responses by over-regularizing more, then this offers some support for the idea that at least part of the reason for the different behavior of children and adults might relate to different assumptions about the task.

Therefore, in this paper we explore the hypothesis that adult over-regularization behavior can be changed by changing the pressure for generating “correct” responses. This pressure is manipulated in two ways. First, we vary the cover story to change the assumptions people make about how likely the data is to reflect an underlying regularity of some sort. Second, we vary the goal of the task to emphasize or de-emphasize effective communication. Consistent with our hypothesis, we find that when the pressure for correct communication is reduced, adults over-regularize more. In the discussion we consider the implications of these results for experimental work on over-regularization, probability learning, and language learning more broadly.

## Experiment

**Participants.** 52 adults<sup>1</sup> were recruited from the University of Adelaide and surrounding community and were paid \$10 for their time. Participants were divided randomly into one of two conditions, HIGH PRESSURE and NO PRESSURE (described below). One person in the HIGH PRESSURE condition suffered a computer error causing a failure to save data, and two participants in the NO PRESSURE condition were excluded from the analysis for typing gibberish.<sup>2</sup> This left 25 participants in the HIGH PRESSURE condition and 24 in the NO PRESSURE condition.

**Procedure.** The standard task, which was the same in both conditions, involved a word learning task originally modelled after Hudson Kam and Newport (2009) and Perfors (2011). The original Hudson Kam and Newport (2009) taught a language containing many words taught over multiple days, but the key element for our purposes was that in this language, units (which they called determiners) covaried with the nouns in an inconsistent fashion: participants heard the **main** determiner only 60% of the time. Participants were asked to provide the noun and determiner associated with a scene and sentence and the frequency with which each determiner was produced after each noun was noted.

As in Perfors (2011), we removed extraneous elements of the task so as to focus on the aspect involved in producing the inconsistent units. Our language consists of words composed from 10 stems, all one-syllable consonant-vowel-consonant nonsense words mapped to images representing common ob-

<sup>1</sup>We ran 52 rather than the more round 50 because the HIGH PRESSURE condition required pairs of participants (an even number of people). We matched that number in the NO PRESSURE condition.

<sup>2</sup>This was reflected in their accuracy score, which was over four standard deviations below the mean accuracy for either condition.

jects. Each stem was attached to a by a one-syllable affix: the **main** affix occurred 60% of the time, and each of the four **noise** affixes occurred 10% of the time.<sup>3</sup> The main difference between this study and the previous ones is that while they were entirely auditory in both their presentation of stimuli and the modality of the response, in this study everything was written. This was necessary in order to make the experimental manipulation possible and believable, as explained below. Words were presented with no space between the stem and the affix: thus, participants would see words like PIMUT or JAFIG. They were not told that each of the words was composed from smaller units. The specific image-label mapping and choice of **main** affix was randomized for each participant.

As in Perfors (2011), the task consisted of a total of 200 trials of image-label pairs. On each trial, an image appeared on the computer screen and at the same time the person saw a label written in all capitals below it: for instance, they might see a picture of a baby and read YOKOM. People went to the next trial by clicking a next button. Learning was tested with 20 questions every 100 trials, for a total of 40 test questions. At each test, the participant was presented with an image and asked to enter the label for it. No feedback was given.

**Conditions.** The goal of this experiment was to explore the possibility that adult over-regularization behavior can be affected by changing the pressure for generating “correct” responses. We therefore constructed two conditions, one designed to increase this pressure as much as possible, and one designed to decrease it as much as possible.

**HIGH PRESSURE.** In this condition, we tried to increase the pressure to be correct by pairing each participant with another person who was in the lab at the same time. Each person was informed that the goal of this experiment was to learn a new language, and then successfully use it to communicate with the other person. They were asked to imagine they were scientists who had just discovered a community speaking this language, and they had gotten an informant to label a series of pictures for them. They were to read these labels, and then they would be tested on how well they had learned them by having to fill in the labels for new pictures. Participants sat at different computers and did the standard task individually, but at the end of the standard task each person was given the labels the other person generated during their test questions, and asked to match each of those labels with the correct image. The participants were told that they would get paid proportionally to how many of this final set of questions both of them got right. This created a great deal of social pressure to learn the language correctly, since not only was each individuals’ payment dependent on it, so was their partner’s.

It is important to note that this manipulation *in itself* does not favor either over-regularization or probability matching. Since the affixes did not correlate to the images at all, people could get 100% correct on the test regardless of what they

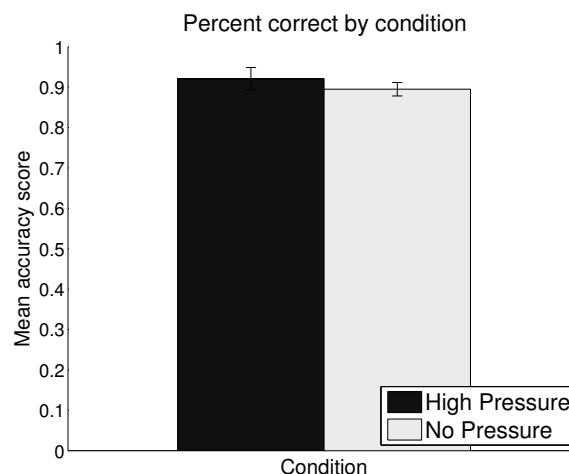


Figure 1: Accuracy in mapping the stems to the correct item. Both conditions showed high accuracy, and there is no significant difference between conditions. This suggests that any differences in over-regularization probably do not occur because participants were paying different amounts of attention to the stems and affixes.

did with the affixes, as long as the stems were matched to the correct image. Thus, any effect on over-regularization is due to increasing the sense that there is a “right answer” as well as the social pressure to find that right answer – *not* because participants could make more money using one strategy.

**NO PRESSURE.** In this condition, the pressure for being “correct” was reduced in two ways: first, by not pairing participants with a partner and making their payment dependent on performance in any way; and second, by changing the cover story so that people were less likely to believe there was an absolutely correct answer. The cover story in this condition was that we were studying how languages change when multiple people learn it. Thus, a previous participant had learned some words in a fake language, and during the course of the learning they were presented with images and asked to label them with those words. Current participants were told that they were being given the labels that had been generated by the previous participant, and that those labels might be kind of strange if the previous person made errors or did something weird. The participants were asked to just do their best to learn the language, and to provide labels that would then be given to the next participant.

A critical element of this design is that the standard task was *exactly the same* across conditions. All of the data we analyze here is from that standard task (since the partner testing in the HIGH PRESSURE condition was only there to determine payment, and not relevant to our research question). The only difference between conditions that could affect the data we analyze is what participants thought their goal was and how they thought the stimuli were generated.

## Results

The main question of interest is how much participants in each condition over-regularized by producing the **main** affix (or any single affix) more than 60% of the time. However,

<sup>3</sup>Stems were: dut, sil, zeg, mab, yok, pim, ren, jaf, wux, and cov and the affixes were: om, ep, ad, ig, and ut. Objects used were: babies, balls, beds, birds, books, cars, cats, cups, dogs, and shoes.

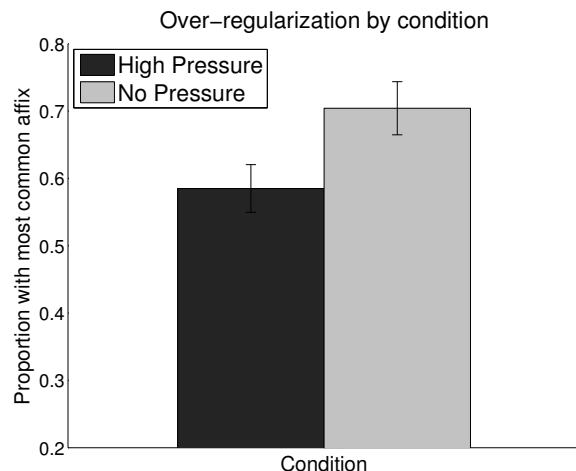


Figure 2: Proportion of time the most common affix is produced in each condition (after having occurred 60% of the time in the input). Participants in the NO PRESSURE condition over-regularized significantly more than did participants in the HIGH PRESSURE condition.

it is first necessary to determine whether any differences in over-regularization are associated with differences in overall performance or attention between conditions.

**Accuracy.** We can calculate an accuracy score for each condition, based on the percentage of stems that are generated in response to the appropriate item. A stem is counted as correct if it is no more than one letter different from the correct stem; thus, *cav* would be an acceptable variant of *cov*, but *div* would not be.<sup>4</sup> As Figure 1 demonstrates, people in both conditions were highly accurate, and there was also no significant difference between conditions ( $t(47) = 0.807, p = 0.424$ , two-tailed). This suggests that any differences in over-regularization did not result from people in the two groups paying different amounts of attention to learning the stems or affixes.

**Over-regularization.** The main question is if people show different levels of over-regularization in each of the two conditions. To determine this, we calculated the percentage of time that the most commonly-used affix was used by each participant. (For almost all participants, the most commonly-used affix was the **main** one; however, we did not want to presume that it always would be). As with the *accuracy* score, two affixes counted as the same if they differed by no more than one letter.<sup>5</sup>

Figure 2 shows the average percentage of trials in which participants produced the most common affix. It is evident that participants in the HIGH PRESSURE condition approximately probability match, while participants in the NO PRES-

<sup>4</sup>All analyses were also performed with a definition requiring the stem to match exactly; results were qualitatively identical.

<sup>5</sup>We also performed all possible combinations of two additional analyses. In the first, we used a more stringent definition of sameness, such that the affixes had to be identical to be counted. In the second, we considered only the subset of affixes for which the stem had been applied correctly. In all of these cases, participants in the NO PRESSURE condition over-regularized significantly more than participants in the HIGH PRESSURE condition.

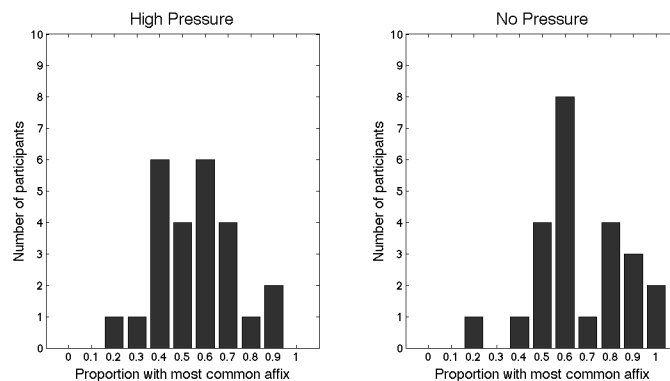


Figure 3: Histogram of the number of participants according to how often they produced the most common affix. It appears that in the HIGH PRESSURE condition, there is a unimodal distribution of responses centered near 60%. By contrast, the NO PRESSURE condition appears to have a bimodal distribution of one group that over-regularizes strongly and another group that probability matches.

SURE condition over-regularize significantly more ( $t(47) = -2.250, p = 0.029$ , two-tailed). This sort of population-level analysis can be misleading, however. It is possible that 60% of the people in the HIGH PRESSURE condition produce the most common affix 100% of the time and 40% of the people never produce any affix more than once. Individual performance can also be revealing about to what extent the conditions resulted in different strategies among the participants.

Figure 3 therefore shows the histograms of each of the individual participants. Although such histograms are inevitably somewhat noisy and we should be careful about overinterpreting, they suggest two things. First, the number of participants who over-regularize in an extreme way (by producing the most common affix 80% of the time or more) is noticeably higher in the NO PRESSURE condition. In other words, these population differences actually do reflect the presence of more people who over-regularize strongly. The second interesting thing is that the distribution of participants in the HIGH PRESSURE condition is approximately unimodal: most people produce the most common affix between 40% and 70% of the time, and there are few outliers on either extreme. By contrast, the distribution in the NO PRESSURE condition looks more bimodal: there is one group of people who over-regularize strongly, and another set who do not. We will address some of the implications of this in the next section.

## Discussion

Overall, this work indicates that adults over-regularize inconsistent linguistic input more often when placed in a situation in which the pressure for correct responding is reduced and they have reason to believe that the inconsistency does not correspond to an underlying regularity. Indeed, what strategy is sensible changes depending on whether one believes that the affixes covary according to some hidden regularity, or covary arbitrarily. If there is a regularity, it is sensible to try to find and match it; if not, it is sensible to remove the inconsistency. The NO PRESSURE condition, by emphasizing that the

input came from a previous participant who might have made errors, strongly implied that any variability was random or accidental. In contrast, the HIGH PRESSURE condition, by giving a cover story implying that the participant was a scientist whose job it was to learn the language, and especially by making payment contingent on successful “communication” of that language, created the strong implication that there was some regularity there to learn. It is interesting that increasing communicative pressure did not result in more removal of inconsistency in the language; again, this is probably because the participants thought that the inconsistency reflected an important regularity, even if they weren’t sure what it was.

Our two conditions confound two effects: one is about the different assumptions participants might make about how likely the data is to contain errors, and the other is about how much pressure they feel to match some “correct” standard. We did not try to disentangle these things here because it would be very difficult, since simply knowing that the data might contain errors would decrease the pressure to correctly learn that data. The main point of this research is that differences in the impressions one brings to a task – whatever they may be – can change the degree of over-regularization.

Other than the experimental manipulations, the main implementational difference between this work and previous work is that both the input and the responses were written rather than auditory. However, since *both* conditions were written, the non-auditory nature of the experiment could not explain the difference between the conditions.

What does it mean that only some of the participants adopted an over-regularization strategy in the NO PRESSURE condition? One possibility is that those who did not adopt one still assumed that the inconsistencies may have been associated (perhaps more noisily) with regularities in the data. After all, it is a rather strange pattern of errors to have 10% each of four other kinds of affixes, but no other misspellings or incorrectly used stems. One way to investigate this possibility would be to design an experiment in which the affixes occur less often or there are additional errors in the input. However, since it is known that adults over-regularize more frequently when the inconsistency is less frequent (Hudson Kam & Newport, 2009) and any introduction of other errors would have to also occur in the HIGH PRESSURE condition in order to make the tasks equal, it is not clear that such a manipulation would be very illuminating. Still, this is a possibility we will consider for future work.

A potential limitation of this work is that the differences between the HIGH PRESSURE and NO PRESSURE conditions, although statistically significant no matter how the data is analyzed, are not large. In particular, even in the NO PRESSURE condition, not as many adult participants over-regularized as did children in Hudson Kam and Newport (2005) and Hudson Kam and Newport (2009). For this reason one of the first steps in future work will be to replicate these findings. That said, our results might underestimate the magnitude of the true effect to the extent that our participants didn’t believe the

“error” cover story but more of the children thought or assumed that the language they heard had some errors. Another quite likely possibility is that not all of the difference between children and adults is reducible to their different assumptions. Finally, any effect of naturally different assumptions between children and adults on the same task in Hudson Kam and Newport (2005) and Hudson Kam and Newport (2009) may have been exacerbated by subtle differences in their experimental design. In particular, while both adults and children learned a language called “Sillyspeak”, only the children were told they were learning from someone who did not know the language themselves. It seems plausible that the children were far more likely to conclude that their input contained a lot more error, especially when combined with the very different social assumptions about the nature of laboratory studies that each group may have had.

What do these results mean for research showing that children over-regularize in non-linguistic situations as well as linguistic ones, as in Derks and Paclisanu (1967) and Jones and Liverant (1960)? Any application to non-linguistic domains must be made extremely cautiously, since many of the differences between the HIGH PRESSURE and NO PRESSURE conditions do not translate to a non-linguistic context. For instance, there is no obvious analogue of errors in a deck of cards or reinforcement pattern. In addition, these experiments do not have the learner/teacher dynamic that the linguistic ones do, which might be associated with very different patterns of assumptions. That said, children may still feel less pressure than adults to be “correct” or not look “stupid” to the scientists, so we cannot rule out the possibility that this plays a role even in non-linguistic domains.

Within the area of language, much of the interest in children’s over-regularization arises because children and adults differ in their propensity to over-regularize outside of the lab as well as in it. Deaf children exposed to the inconsistent sign language of hearing parents will over-regularize that language and produce regular grammatical forms (Singleton & Newport, 2004), but adult language learners are known to produce highly variable, inconsistent utterances, even after years of experience with the language and after their grammars have stabilized (Johnson, Shenkman, Newport, & Medin, 1996). If over-regularization in children in these experiments is driven by differences in the assumptions they bring to the task, how do we explain these differences in real life?

In answer, we can only speculate. However, several possibilities present themselves. One is that children’s over-regularization in real life is driven by some of the same factors that we are calling task demands here: that is, perhaps children just assume that more of their input is irregular or full of errors, or they are less bothered by trying to be “correct” or not look like an idiot and therefore focus on simply communicating clearly. Another possibility is that the different assumptions manipulated in the tasks here do not explain all of the differences between children and adults that are found in laboratory experiments. Still another possibility is that the

kind of over-regularization measured in these experimental tasks does not map cleanly onto the over-regularization differences observed in natural language. At a minimum, one of these processes occurs in hours or days, and one takes years. Moreover, children may not bring the same assumptions to real language learning as to this kind of experimental task. Even beyond that, most variation in natural language is consistent in some way (e.g., Chambers, Trudgill, & Schilling-Estes, 2003). Thus, any child/adult differences in learning languages from native speakers<sup>6</sup> may not be traceable to the kinds of over-regularization differences found in these experiments. Thus, the implications for language acquisition must be speculative at this point, and we cannot say at this point for sure how to reconcile these findings with the language acquisition literature. It is important to note, however, that there are many ways they could be reconciled. Pursuing this is a project for future work.

This work may additionally have implications for adult second language acquisition, since it demonstrates that adults may change strategies in response to the nature of the explicit instructions they receive. This too is a project for future work.

The bottom line from this paper is that it is important to be cautious about drawing strong conclusions from existing laboratory studies to differences in how children and adults over-regularize when learning natural language. At the very least, the story is probably complex. For instance, some work shows that infants probability match in a looking time experiment (Davis, Newport, & Aslin, 2011), and even some of the original work in non-linguistic domains showed that children over-regularized the same amount as adults (Weir, 1964) or flipped strategies differentially depending on the nature of the reward (Stevenson & Hoving, 1964). The picture is therefore currently somewhat murky, even as regards the extent to which – and at what ages – children tend to over-regularize more than people of other ages.

In sum, this paper offers some reason to believe that over-regularization behavior can be driven by different assumptions about the goal of the experimental task and the origin of the data. If we are to understand what drives differences between adult and child language learning, we need to determine the extent to which the experimental findings in the literature stem from differences in such assumptions. It is also critical to further explore the extent to which these assumption explain differences in the learning of natural languages. This work is the first step along that path.

## Acknowledgments

Thank you to Natalie May, Angela Vause, and Daniel Carabellese for recruiting participants and running the experiments. Thanks to them as well as Dan Navarro and Irina Baetu for useful discussions. This research was supported by ARC grants DE120102378 and DP110104949.

<sup>6</sup>Pidgins and creoles would be a different story, of course, because there the input might indeed be truly inconsistent. That said, there are many additional complexities in that case as well; for instance, adults receive input from different people than children, and have different goals in their communications in the kind of setting that causes pidgin languages to emerge.

## References

- Baum, W. (1975). Time allocation in human vigilance. *Journal of the Experimental Analysis of Behavior*, 23, 45–53.
- Baum, W. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the Experimental Analysis of Behavior*, 32, 269–281.
- Bradshaw, C., Ruddle, H., & Szabadi, E. (1981). Studies of concurrent performances in humans. In C. Bradshaw, E. Szabadi, & C. Lowe (Eds.), *Quantification of steady-state operant behaviour*. Amsterdam: North Holland Biomedical Press.
- Bradshaw, C., Szabadi, E., Bevan, P., & Ruddle, H. (1979). The effects of punishment on free-operant choice behavior in humans. *Journal of the Experimental Analysis of Behavior*, 31, 71–81.
- Castellan, N. J. (1974). The effect of different types of feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance*, 11, 44–64.
- Chambers, J., Trudgill, P., & Schilling-Estes, N. (2003). *The handbook of language variation and change*. Blackwell.
- Davis, S., Newport, E., & Aslin, R. (2011). Probability-matching in 10-month-old infants. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.
- Derks, P., & Paclisanu, M. (1967). Simple strategies in binary prediction by children and adults. *Jn. Exp. Psych.*, 73(2), 278–285.
- Herrnstein, R. (1961). Relative and absolute strength of responses as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, 4, 267–272.
- Herrnstein, R. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, 13, 243–266.
- Hudson Kam, C., & Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Jn. of Exp. Psych.: Lng., Mem., & Cog.*, 35(3), 815–821.
- Hudson Kam, C., & Newport, E. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Lang. Lng. & Dev.*, 1(2), 151–195.
- Hudson Kam, C., & Newport, E. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, 59, 30–66.
- Johnson, J., Shenkman, K., Newport, E., & Medin, D. (1996). Indeterminacy in the grammar of adult language learners. *Journal of Memory and Language*, 35, 335–352.
- Jones, M., & Liverant, S. (1960). Effects of age differences on choice behavior. *Child Development*, 31, 673–680.
- Perfors, A. (2011, .). Memory limitations alone do not lead to over-regularization: An experimental and computational investigation. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.
- Perfors, A., & Burns, N. (2010). Adult language learners under cognitive load do not over-regularize like children. In *Proc. 32nd Annual Conf. of the Cognitive Science Society* (p. 2524–2529).
- Pierce, W. D., & Epling, W. F. (1983). Choice, matching, and human behavior: A review of the literature. *The Behavior Analyst*, 6, 57–76.
- Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.
- Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends in Cognitive Sciences*, 11(7), 274–279.
- Shanks, D., Tunney, R., & McCarthy, J. (2002). A re-examination of probability matching and rational choice. *Jn. of Behavioral Decision Making*, 15, 233–250.
- Singleton, J., & Newport, E. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, 49, 370–407.
- Stevenson, H., & Hoving, K. (1964). Probability learning as a function of age and incentive. *Journal of Experimental Child Psychology*, 1, 64–70.
- Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, 14, 101–118.
- Wearden, J. (1983). Undermatching and overmatching as deviations from the matching law. *Journal of the Experimental Analysis of Behavior*, 40, 332–340.
- Weir, M. (1964). Developmental changes in problem-solving strategies. *Psychological Review*, 71, 473–490.



# An Operational Model of Joint Attention - Timing of Gaze Patterns in Interactions between Humans and a Virtual Human

Nadine Pfeiffer-Lessmann (nlessman@techfak.uni-bielefeld.de)

Thies Pfeiffer (tpfeiffe@techfak.uni-bielefeld.de)

Ipke Wachsmuth (ipke@techfak.uni-bielefeld.de)

Artificial Intelligence Group, Faculty of Technology, Bielefeld University, Bielefeld, Germany

## Abstract

Joint attention has been identified as a foundational skill in human-human interaction. If virtual humans are to engage in joint attention, they have to meet the expectations of their human interaction partner and provide interactional signals in a natural way. This requires operational models of joint attention with precise information on natural gaze timing. We substantiate our model of the joint attention process by studying human-agent interactions in immersive virtual reality and present results on the timing of referential gaze during the initiation of joint attention.

**Keywords:** joint attention; virtual humans; social interaction

## Introduction

Attention has been characterized as an increased awareness (Brinck, 2003) and intentionally directed perception (Tomasello, Carpenter, Call, Behne, & Moll, 2005) and is judged to be crucial for goal-directed behavior. Joint attention builds on attentional processes and has been identified to be a foundational skill in communication and interaction. The term joint attention is often used confusably with shared attention. We follow Kaplan and Hafner (2006) and Tomasello et al. (2005) in using the term joint attention for the phenomenon which presupposes a higher level of interactivity requiring intentional behavior and an awareness of the interaction partner. Joint attention can be defined as simultaneously allocating attention to a target as a consequence of attending to each other's attentional states (Deak, Fasel, & Movellan, 2001). In contrast, we see shared attention (as well as shared gaze) as the state in which interactants are just perceiving the same object simultaneously without further constraints concerning their mental states or their interaction history.

Mundy and Newell (2007) differentiate joint attention behaviors into two categories: *responses* to the bids of others and spontaneous *initiations*. Responding to joint attention refers to the ability to follow the direction of gaze and gestures of others in order to share a reference. On the other hand, to initiate joint attention humans use gestures and eye contact to direct the attention of others to objects, events, and to themselves.

For joint attention, interlocutors have to deliberately focus on the same target while being mutually aware of sharing their focus of attention (Tomasello et al., 2005; Hobson, 2005). To this end, respond and feedback behaviors are necessary. Tasker and Schmidt (2008) argue that to establish joint attention a sequence of behaviors is required which has to meet certain time constraints.

We constructed an operational model of joint attention (Pfeiffer-Lessmann & Wachsmuth, 2009) for our virtual human Max (Lessmann, Kopp, & Wachsmuth, 2006) to create a more natural and effective interaction partner. The model covers four phases: the initiate-act (1), the respond-act (2), the feedback phase (3), and the focus-state (4). However, for Max to appear believable and to use the same behavior patterns in the phases as humans do, investigations on time-frames, human expectations and insights on how humans actually perceive his behavior are indispensable. The topic of concrete reaction and duration times of feedback behaviors during the joint attention process has to our knowledge not been discussed in the area of human-computer interaction yet. The time-frames and expectations of humans for natural interactions are central subject of this paper.

In the section to follow, we provide an overview on related work covering research on joint attention in human-human interaction and in the area of technical systems. In the subsequent "Model" section, a brief summary of our own definition of joint attention is provided. Next, we present a study in immersive virtual reality concerning the exact timing of the first phase, the initiate-act, of our joint attention model. Thereafter, results are discussed and the paper ends with our conclusions and future work.

## Related Work

Staudte and Crocker (2011) raise the question whether joint-attention-like behavior is unique to human-human interaction or whether such behaviors can play a similar role in human-robot interaction. They conclude that their own findings suggest that humans treat artificial interaction partners similar to humans and that it is therefore valid to investigate joint attention in settings with artificial agents.

These artificial agents can consist, on the one hand, of robots (Deak et al., 2001; Imai, Ono, & Ishiguro, 2003; Breazeal et al., 2004; Doniec, Sun, & Scassellati, 2006; Nagai, Asada, & Hosoda, 2006; Yu, Schermerhorn, & Scheutz, 2012; Huang & Thomaz, 2011; Staudte & Crocker, 2011) and, on the other hand, of virtual humans (Peters, Asteriadis, & Karpouzis, 2009; Zhang, Fricker, & Yu, 2010; Bailly, Raidt, & Elisei, 2010).

Kaplan and Hafner (2006) point out that research in robotics concentrates only on partial and isolated elements of joint attention (e.g. gaze following, simultaneous looking or simple coordinated behavior) covering solely the surface of the process but not addressing the deeper, more cognitive

aspects of the problem. The same authors stress that no system achieved true joint attention between a robot and a human or between two robots according to their definition yet. This appears to be still the case, however progress has been made with respect to investigating joint attention behaviors.

A number of researchers in cognitive science and cognitive robotics use developmental insights as a basis for modeling joint attention showing how a robot can acquire joint attention behaviors by supervised and unsupervised learning (Deak et al., 2001; Nagai et al., 2006; Doniec et al., 2006). However, the aspect of intentionality and explicit representation of the other's mental state are not accounted for in these approaches.

Another area of research investigates the impact of artificial agents' joint attention behavior on humans. Here, real interaction scenarios can be distinguished from humans rating video material. Huang and Thomaz (2011) argue that video-based experiments offer the advantage of studying humans' perception of joint attention behaviors without dealing with technical challenges of identifying the humans' behaviors. According to Staudte and Crocker (2011), it has been shown that video-based scenarios without true interaction yield similar results to live-scenarios and can therefore provide valuable insights into humans' perceptions and opinions.

Huang and Thomaz (2011) use videos to investigate humans' judgements of robots initiating and ensuring joint attention behavior. Their results suggest that humans overall preferred robots showing joint attention behavior. Staudte and Crocker (2011) also follow a video-based approach; they conclude that participants robustly follow the robot's gaze and use it to anticipate upcoming referents. Bailly et al. (2010) try to quantify the impact of deictic gaze patterns of their agent. They explicitly instructed participants not to take the agent's behavior into account, but the participants were drastically influenced by the agent's gaze patterns anyway.

In a real interaction scenario, Peters et al. (2009) study how human participants perceive the virtual agent's simple shared attention behavior of non-verbal cuing and how subtle changes of this behavior affect the gaze-following of human participants. Huang and Thomaz (2011) investigate the respond-act of their robot and the resulting impact on a human-robot collaborative task using a task-based metric. They find that the robot responding to referential foci significantly outperforms the one staying focused on the human.

The robot of Breazeal et al. (2004) keeps a representation of its current focus of attention calculated by saliency values. Additionally, it monitors the human participant's focus of attention. It is thereby able to notice when both interactants focus on the same object simultaneously. However, the robot appears to miss feedback mechanisms on a higher level of interactivity covering intentional behavior and the awareness of the interaction partners in the joint attention process.

Many researchers investigating the impact of artificial agents which show joint attention behaviors do not account for the necessary time courses. As an exception, Yu et al. (2012) try to investigate the exact time course of multi-modal

interaction patterns occurring naturally as part of joint attention processes.

In our own approach, we let human participants engage with our interactive virtual agent Max in an immersive virtual environment. As with human-robotic and human-human interactions, the interactants thus share the same three-dimensional environment and reciprocal interactions are possible. However, other than today's robotic systems, the virtual agent has (more than) human-like reaction times and is controlled by a cognitive architecture which goes from basic activation processes up to concepts of epistemic modal logics to model mutual beliefs which are essential for joint attention.

For human-human interactions, Tasker and Schmidt (2008) postulate a time frame of 5 s for the addressee of an initiate-act to respond appropriately. According to them, the duration of the respond-act has to last for at least 3 s in order to establish evidence that the partner's attention has been captured. The focus-state of joint attention has to last for a minimum of 3 s, too. This duration is in accordance to the results found by Vaughan et al. (2003) for maintaining focus on the same object in an episode of joint engagement.

Mueller-Tomfelde (2007) takes a closer look at the research literature to figure out appropriate time scales for referential actions. Since an initiate-act or respond-act could be characterized as such, his results should be highly relevant for natural time-scales of joint attention behaviors. He argues that since a pointing action includes cognitive aspects, it is more than a basic movement-primitive and thus more than a basic physical act constrained by the nature of cognitive operations at a time period of about a  $\frac{1}{3}$  of a second. Therefore, Mueller-Tomfelde (2007) expects an appropriate temporal scale of referential primitives to be greater than 300 ms while being less than the temporal scale of actions of a higher cognitive level with a temporal time window of 2-3 s.

## Model of Joint Attention

The model of joint attention presented here is in agreement with the model of Tasker and Schmidt (2008), except that we do not adopt their time constraints for joint attention. Our model also meets the requirements of Kaplan and Hafner (2006), for a longer discussion see Pfeiffer-Lessmann and Wachsmuth (2009). However, our model differs in that we do not require interactants to perform a certain sequence of behaviors. Instead we define the effects of joint attention behaviors on their mental states. Thereby, different behaviors can be performed counting as joint attention behaviors. However, to realize a natural interaction partner, we are now investigating valid joint attention behaviors performed by humans to be implemented in our artificial agent.

We define four phases characterized by the mental states of the interactants (see Figure 1). In order to engage in joint attention, the interaction partners need to have a certain kind of psychological engagement with each other, which can be described as involving a species of perception as well as a species of emotional responsiveness (Hobson, 2005). This

can be defined as the precondition for joint attention. To establish joint attention, certain behaviors leading to certain mental states need to take place. The first phase can be described as the *initiation-phase*; one of the interactants performs an initiate-act, which the other interactant can recognize. The second phase can be described as the *respond-phase*. Now the addressee of the initiate-act needs to perform a respond-act. The third phase is characterized as the *feedback-phase*; the interactants affirm that they have recognized the interaction attempts of their interaction partners. The forth and last phase consists of the *focus-phase*; now both interactants focus on the object of attention and are aware of the joint attention state (see also Pfeiffer-Lessmann and Wachsmuth (2009) for a formalized definition of the required mental state for joint attention).

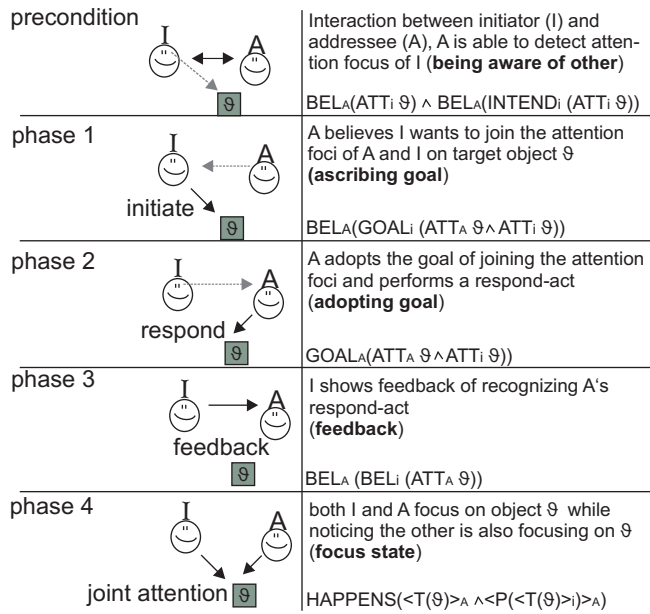


Figure 1: Phases of the joint attention process - Initiator (I) and addressee (A) wear hats according to their roles.

The model of Huang and Thomaz (2011) with five steps shares many features with our model except that we define to-be-aware of the interactant as a prerequisite and not a step in the joint attention process and that they concentrate in their step 4 on verifying the response of the addressee whereas we lay more emphasis on the required feedback mechanisms between both interactants.

### Study on the Timing During the Initiate-Act

While the review of related work has brought up timing data on the phases 2 to 4 of our model, little has been found on the internal timing of events during the initiate-act. With the following study, we address the question on the timing of the initiator's referential act in which she first introduces the target of the joint attention process. Additionally, we investigate acceptable response times of the addressee for a referential

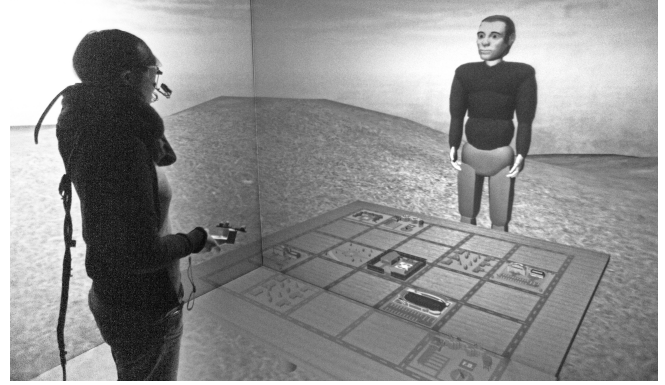


Figure 2: In the study, the human participant faces the virtual agent Max in a fully immersive virtual reality environment. Between the two interlocutors is a table with ten objects from a city planning scenario, which serve as reference objects. The eye gaze and the head movements of the human participant are tracked and the line of gaze onto the objects in the virtual environment is computed in real-time.

act to be considered successful. As a first step, we thereby focus on referential acts via eye gaze.

**Scenario** We investigate joint attention in a cooperative interaction scenario with the virtual human Max, where the human interlocutor meets the agent face-to-face in 3D virtual reality (see Figure 2). The human's body movements and gaze are picked up by infrared cameras and an eye tracker (Pfeiffer, 2011). This enables Max to follow the human's head movements and gaze in real-time, the two aspects of human joint attention behavior considered in this study.

### Participants

Altogether data from 20 participants (10 women, 10 men) has been collected. All participants were students or employees of Bielefeld University. The age of the participants was between 21 and 45 years, with a mean of 28 years and a SD of 5.17.

### Method

The participants were invited to our lab and given a brief introduction to the study. At this time, they filled out a short questionnaire and read the written instructions for the tasks. After that, they were equipped with the tracked stereo glasses required for the immersive virtual reality setup. For controlling the experiment, they were given a Wii Remote to step through the trials. After the participants had entered the virtual environment, they had time to get accustomed to the scenario. Finally, the eye-tracking system was calibrated and the participants repeated verbally the procedure of the study. After all questions had been answered, the trials started.

The two possible roles of an interlocutor (initiator or addressee), are reflected by the study design: two blocks I and A are repeated, where the human participant is the initiator

in block I and the addressee in block A. The blocks were repeated three times, for the first ten participants in the order IAIAIA, for the second ten in the order AIAIAI. The tasks within each block are described below. The ten items in block I and block A had a pre-randomized sequence, which was static between participants but different for the first, second and third presentation of the block.

After all blocks were completed, the participants were debriefed. Before departing, all participants received a recompense for taking part in the experiment.

### I: Dwell Time of Referential Gaze Produced by Initiator

The aim of block I is gathering data about the typical dwell time of the referential gaze act of an interlocutor when attempting an initiate-act. During the initiate-act, the interlocutor focuses on the target object for a certain amount of time  $\alpha_r$  ( $r$  for reference) until she checks back by focusing on the face of the interaction partner for time  $\beta_r$ . The total duration of the initiate-act is  $\alpha_r + \beta_r + 2\epsilon_r$ , with  $\epsilon_r$  being the very short time needed to shift the gaze focus.

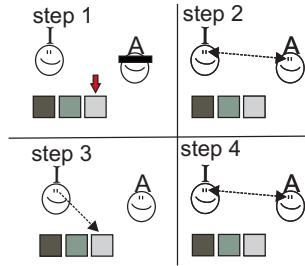


Figure 3: The sequence of steps for one item in block I. I=initiator (human) and A=addressee (agent)

The interaction scenario described above, with a human interlocutor addressing the virtual agent Max, provides the frame for this task. Max plays the role of an interlocutor, while the participant is instructed to perform an initiate-act for one of the objects located on a virtual table between the participant and Max (see Figure 2). In an orientation phase prior to each initiate-act, Max gets blindfolded and the next target object is highlighted with a red arrow (see Figure 3, step 1). This design has been chosen to make explicitly clear that Max has no prior knowledge of the target object. Once the participant has located the new target object, she has to return her gaze to Max and press a button to start the interaction. This removes the arrow as well as the blindfold of Max. Max then gives a short verbal phrase to provide the context of the joint attention act and the human participant can start her initiate-act. The participant is instructed to use gaze only to try to direct the attention of Max towards the given target object (Figure 3, step 2). She should start while focusing on Max' face, then attempt an initiate-act by focusing at the target object as long as she feels is needed (Figure 3, step 3, while we collect data on  $\alpha_r$ ). She should then interrupt focusing on the target object and check back at Max' face (Fig-

ure 3, step 4). Finally, she should press a button as soon as she feels that Max should have reacted by then (while we collect data on  $\beta_r$ , the expected maximum response time). Because at this point in time we do not want Max to influence the participant's timing behavior, Max does not show any reaction in response to the participant's attempts. The whole procedure is repeated for the remaining objects, until all ten objects have been covered.

### A: Dwell Time of Referential Gaze Accepted by Addressee

In the second part of the study we reverse the roles of the human interlocutor and the virtual agent Max.

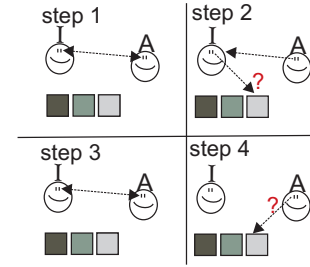


Figure 4: The sequence of steps for one item in block A. I=initiator (agent) and A=addressee (human)

Now, it is Max who is performing initiate-acts to achieve joint attention and the human interlocutor observes and evaluates these attempts. During the initiate-act, Max will stop focusing on the interlocutor and move his gaze focus to the target object for an amount of time from a predefined set ranging from 600 ms to 3000 ms in steps of 600 ms (Figure 4, step 2). These values have been selected to comprise typical non-communicative gaze durations and the findings on mean durations from the literature. Max will then focus back at the participant's face (Figure 4, step 3). The participant is asked to watch Max' gaze. Once Max focuses back at the participant, she has to decide whether Max had intended her to follow his gaze. If she decides so, she has to press a button and gaze at the target object (Figure 4, step 4). If not, she has to do nothing. After five seconds, Max will automatically continue with the next item.

During the interaction, one measurement is made. By pressing the button the human participant ascribes Max to have performed a valid initiate-act. We then count the dwell time used by Max from the given set as an acceptable dwell time for an initiate-act,  $\alpha_a$  ( $a$  for acceptance).

## Results

In a post-study questionnaire (seven-point Likert scales (1-7), median score is given here), the participants reported that they felt present in the virtual environment (score 5) and experienced the agent as being even more present (score 6). The naturalness of the communication with the agent, however, was rated 3 (SD 2). The participants also were able to fully concentrate on the task (score 6) and were not hindered by the devices. Overall, they enjoyed the experience in the virtual

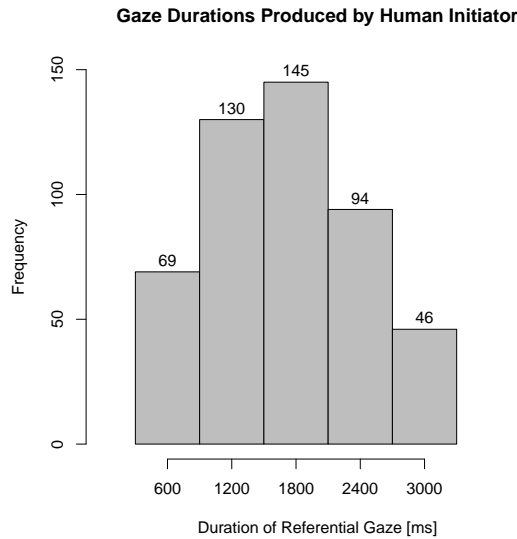


Figure 5: Dwell times of referential gaze during initiate-acts produced by the human participants in study part I.

reality (score 5) and had no difficulties with the task (score 3 rating the difficulty).

### I: Dwell Time of Produced Referential Gaze

During block I, 560 initiate-acts were recorded. Overall, the mean dwell time of referential eye gaze ( $\alpha_r$ ) was 1896.82 ms (SD 963.46 ms) and the median was 1796 ms. A histogram of the durations of the referential gaze is depicted in Figure 5. During the orientation phase when the participants had to identify and remember the target object, the mean dwell time of eye gaze was 1559.58 ms (SD 1029.24 ms) and the median was 1390.5 ms. The dwell time during search was significantly shorter than the dwell time of referential gaze (t-Test results in  $t=5.91$  with  $p=0.001$  by 545 DoF, confidence interval 215.75 ms to 430.42 ms).

Overall, the mean duration until the human participant expected a feedback after the production of a referential eye gaze towards the target object ( $\beta_r$ ) was 2556.07 ms (SD 1721.06 ms) and the median was 2247.5 ms.

### A: Dwell Time of Accepted Referential Gaze

In block A, Max produced altogether 600 initiate-acts with referential gaze of different durations (600 ms to 3000 ms in 600 ms steps). The task of the human participant was to decide, whether she accepts the gazing behavior as being intentional in that Max wanted to guide her attention to the target object. The dwell time of accepted referential gaze of the five discrete levels is  $\alpha_a$  with a median of 1800 ms. The histogram of the accepted dwell times is depicted in Figure 6.

A chi-squared test comparing the accepted dwell times  $\alpha_a$  in block A and the dwell times  $\alpha_r$  for referential gaze used by the participants in block I (discretized to the discrete values

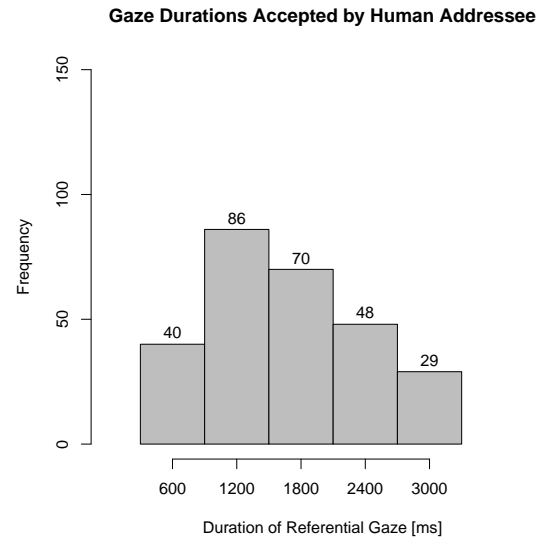


Figure 6: Dwell times of referential gaze during initiate-acts produced by Max in study part A, which have been accepted by the human participant as being intentional.

used in block A) shows no significant differences ( $p=0.22$ , see also Figure 5 and Figure 6).

### Discussion

For the presented study we created an immersive virtual environment and let the participants engage in joint attention with a virtual agent to have a realistic but highly controlled experimental setup to run our studies on cognitive models of joint attention. The feedback from the participants regarding their own experience of presence and the presence of the virtual agent renders this approach a success.

With this advanced setup, we aimed at substantiating our knowledge about the timing of referential gaze within the initiate-act. In block I, we found a mean dwell time of referential gaze towards the target object  $\alpha_r$  of about 1897 ms. We take the significant differences of the  $\alpha_r$  from the dwell time on the same target objects during the orientation phase (when the target objects are shown to the participants) as a confirmation of the different nature of gaze use in search and in referential gaze. This also shows that the design of the study is plausible to the participants regarding the different interaction states (orientation phase vs. dialog). Using these timing patterns, Max will learn to arbitrate between gaze search and referential gaze in the future.

If roles are switched and the initiate-act is performed by the virtual agent Max, we found that participants accepted the same kind of gaze patterns as natural as they themselves performed when they had the initiative. This substantiates our findings and at the same time emphasizes the high acceptance of the virtual agent Max as an interaction partner.

A respond-act of the addressee to an initiate-act of the human participant was expected before 2556 ms. This is well

below the 5 s time frame postulated by Tasker and Schmidt (2008) based on human-human interactions. However, as our study focused on the dwell times during referential gaze and Max by design only showed a response when triggered, it was difficult for the participants to decide this threshold. A more thorough investigation of this threshold should use a more complex scenario, were, e.g., Max produces responds with different delays, similar to the design in block A.

## Conclusion

The high acceptance of Max as an interaction partner with human-like capabilities and the comparability of our findings in human-machine interaction with those found in human-human interaction motivate us to follow this line of research further. The 1.9 s dwell time of the referential gaze act is compatible with related findings in human-human interaction. In next steps, we would substantiate our model of joint attention by incrementally increasing the complexity of the interaction scenario until the full process of joint attention can be simulated in real-time in a more natural scenario. This would also allow us to directly compare joint attention behaviors between human-human and human-agent interactions.

Although autonomous behaviors of Max were reduced to a minimum in our controlled setup his naturalness of communication was already rated 3. We believe this rating will increase significantly when he shows his full range of communicative and joint attention behaviors.

## Acknowledgments

This research is supported by the Deutsche Forschungsgemeinschaft in the SFB 673 Alignment in Communication.

## References

- Bailly, G., Raidt, S., & Elisei, F. (2010). Gaze, conversational agents and face-to-face communication. *Speech Communication*, 52(6), 598–612.
- Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., et al. (2004). Humanoid robots as cooperative partners for people. *Int. J. of Humanoid Robots*, 1–34.
- Brinck, I. (2003). The objects of attention. In *Proc. of ESPP2003, Torino* (pp. 1–4).
- Deak, G. O., Fasel, I., & Movellan, J. (2001). The emergence of shared attention: Using robots to test developmental theories. In *Proc. of the First Intl. Workshop on Epigenetic Robotics, Lund University Cognitive Studies*, 85 (p. 95-104).
- Doniec, M. W., Sun, G., & Scassellati, B. (2006). Active learning of joint attention. In *Proc. of 2006 IEEE-RAS int. conf. on humanoid robots (Humanoids 2006)*.
- Hobson, R. P. (2005). What Puts the Jointness into Joint Attention? In N. Eilan, C. Hoerl, T. McCormack, & J. Roessler (Eds.), *Joint attention: communication and other minds* (p. 185-204). Oxford University Press.
- Huang, C.-M., & Thomaz, A. (2011). Effects of responding to, initiating and ensuring joint attention in human-robot interaction. In *RO-MAN 2011 IEEE* (pp. 65–71).
- Imai, M., Ono, T., & Ishiguro, H. (2003). Physical Relation and Expression: Joint Attention for HumanRobot Interaction. *IEEE Transactions on Industrial Electronics*, 50(4), 636-643.
- Kaplan, F., & Hafner, V. (2006). The challenges of joint attention. *Interaction Studies*, 7(2), 135-169.
- Lessmann, N., Kopp, S., & Wachsmuth, I. (2006). Situated interaction with a virtual human - perception, action, and cognition. In G. Rickheit & I. Wachsmuth (Eds.), *Situated Communication* (p. 287-323). Berlin: Mouton de Gruyter.
- Mueller-Tomfelde, C. (2007). Dwell-based pointing in applications of human computer interaction. In *Proc. of the 11th Int. Conf. on Human-Computer Interaction (INTERACT 2007)* (pp. 560–573). Springer Verlag.
- Mundy, P., & Newell, L. (2007). Attention, joint attention, and social cognition. *Current directions in psychological science*, 16, 269–274.
- Nagai, Y., Asada, M., & Hosoda, K. (2006). Learning for joint attention helped by functional development. *Advanced Robotics*, 20(10), 1165–1181.
- Peters, C., Asteriadis, S., & Karpouzis, K. (2009). Investigating shared attention with a virtual agent using a gaze-based interface. *Journal on Multimodal User Interfaces, Kluwer Academic Publishers*, 3(1-2), 119–130.
- Pfeiffer, T. (2011). *Understanding multimodal deixis with gaze and gesture in conversational interfaces*. Aachen, Germany: Shaker Verlag.
- Pfeiffer-Lessmann, N., & Wachsmuth, I. (2009). Formalizing joint attention in cooperative interaction with a virtual human. In B. Mertsching, M. Hund, & Z. Aziz (Eds.), *KI 2009: Advances in Artificial Intelligence* (pp. 540–547). Springer Verlag.
- Staudte, M., & Crocker, M. W. (2011). Investigating joint attention mechanisms through spoken human-robot interaction. *Cognition*, 120(2), 268 – 291.
- Tasker, S. L., & Schmidt, L. A. (2008). The dual usage problem in the explanations of joint attention and children's socioemotional development: A reconceptualization. *Developmental Review*, 28(3), 263–288.
- Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28, 675-691.
- Vaughan, A., Mundy, P., Block, J., Burnette, C., Delgado, C., & Gomez, Y. (2003). Child, caregiver, and temperament contributions to infant joint attention. *Infancy*, 6(6), 603–616.
- Yu, C., Schermerhorn, P., & Scheutz, M. (2012). Adaptive eye gaze patterns in interactions with human and artificial agents. *ACM Trans. Interact. Intell. Syst.*, 1(2), 13:1–13:25.
- Zhang, H., Fricker, D., & Yu, C. (2010). A multimodal real-time platform for studying human-avatar interactions. In *Proc. of the 10th Int. Conf. on Intelligent Virtual Agents* (pp. 49–56). Springer-Verlag.

# Manipulating Manner: Semantic Representations of Human Locomotion Verbs in English and German

Katherine S. Phelps (Katherine.Phelps@Colorado.Edu)

Department of Linguistics, Hellems 290, 295 UCB  
Boulder, CO 80309 USA

Steve Duman (Steven.Duman@Colorado.Edu)

Department of Linguistics, Hellems 290, 295 UCB  
Boulder, CO 80309 USA

## Abstract

It has been argued that real-world structure constrains the semantic representations of verbs, resulting in cross-linguistic convergence of naming patterns for motion events. This study explores the nature of this real-world structure by manipulating individual features of human locomotion in video stimuli and comparing the responses of English and German speakers in an elicitation task. We show that individual features influence naming patterns and that languages encode these features differently. Furthermore, the semantic representations of several German motion verbs sharply contrast with their English equivalents.

**Keywords:** semantic representations; naming; concepts; cross-linguistic diversity; motion verbs; locomotion

## Introduction

Languages divide the world in different ways. Moreover, the boundaries between semantic categories within a particular language are not necessarily fixed. These two factors contribute to a complicated picture in any cross-linguistic comparison of naming patterns. Still, such research has yielded strong evidence of convergent naming patterns across languages in domains such as color (Berlin & Kay, 1969; Kay et al., 1997), emotion (Eckman, 1972), body terms (Majid, Enfield & van Staden, 2006) and events (Majid, Boster, & Bowerman, 2008; Malt et al., 2008). While cognitive biases shared by humans may result in similar construals, humans are also sensitive to salient discontinuities in the world. This real-world structure constrains naming patterns.

Malt et al. (2008) show that structure in the world has a strong influence on the naming patterns of motion events. In a cross-linguistic study in which participants were asked to describe human locomotion, the researchers demonstrate that Dutch, English, Japanese and Spanish speakers uniformly mark a biomechanical distinction between ‘walking’ and ‘running’ gaits when naming these events. However, gait is a cluster of co-occurring features and Malt et al.’s (2008) data do not indicate which of these features are encoded by motion verbs. Also, their study is limited to four languages and should be augmented with data from more languages.

The research we report here manipulates cadence independently from other gait features and shows that it is the latter that influence the category boundary between ‘walking’ and ‘running’ terms, but cadence influences naming on either side of the boundary. Second, it incorporates naming patterns from a German dialect that suggest Malt et al.’s (2008) claim may be too strong. Though some German verbs of human locomotion do encode the biomechanical distinction between ‘walking’ and ‘running’, the term ‘laufen’ (often translated as both *walk* and *run*) does not. This runs contrary to prior predictions. Moreover, the extent to which cadence influences naming may be language-dependent, as some German verbs appear to be less sensitive to manipulations of cadence than their English counterparts.

The present study therefore contributes to research of event categorization in two ways. First, it adds to our understanding of the semantic representations underlying motion verbs by providing a concise picture of the gait features speakers attend to when naming human locomotion. Second, it compares naming patterns of human locomotion in English and German, revealing unexpected patterns not present in previous cross-linguistic comparisons.

## Naming Human Locomotion Events

Continuous human locomotion is particularly interesting due to its biomechanical complexity; it is composed of many co-occurring features. These biomechanical features include but are not limited to stride length, knee bend, elbow bend, and cadence, i.e., the number of steps per unit of time (Kiss, Kocis & Knoll, 2004). Combined, these features can be described as a person’s gait, or their manner of motion. A speaker may draw on several of these gait features when naming a human locomotion event. Importantly, at a particular speed there is a dramatic switch between the clusters of features often categorized as a ‘walking’ gait—a pendulum-type body motion where at least one foot stays on the ground at all times—and a ‘running’ gait—characterized by more elastic, springing movement (Alexander, 1992).

Malt et al. (2008) demonstrate how this real-world structure—namely the dramatic shift in gait—informs the



semantic representations of motion verbs in Dutch, English, Japanese, and Spanish. While viewing stimuli of a woman on a treadmill at varying speed settings and inclines, participants were asked to fill in the blank in the sentence: “What is the woman doing? She is \_\_\_\_.” The striking finding of this study was the uniformity in responses with regard to the 4.5 to 5.5 mph treadmill settings. For each language, ‘walking’ terms always appeared from 4.5 mph and slower (and never over 4.5 mph) whereas ‘running’ terms always appeared from 5.5 mph and faster (and never under 5.5 mph). As mentioned above, this distinction marks an important gait difference. The authors argue that this cross-linguistic convergence is the result of structure in the world exerting strong influence over naming patterns.

This cross-linguistic convergence does not appear to the same extent on either side of the 4.5/5.5 mph boundary. In English, for example, much more within-language variation for lexemes such as ‘jog’ and ‘run’ was found, where use of the latter increases with an increase in treadmill speed (between 5.5 and 8.5 mph), but it is never used 100% of the time.

While the authors acknowledge that there are many features to which speakers may attend, they admit that “the data do not tell us exactly what cues our participants were responding to” (Malt et al. 2008, p. 239). Through a small manipulation of the video stimuli in Study 1 and the addition of a German dialect in Study 2, we demonstrate the detailed nature of speakers’ semantic representations of gait terms and show how individual features, particularly cadence, can be a driving force behind naming patterns. We suggest that naming on either side of the 4.5/5.5 mph boundary is quite sensitive to cadence. Thus, individual features may significantly affect naming patterns in some circumstances (on either side of the 4.5/5.5 mph boundary) and not others (at the boundary marked by other gait elements).

## Study 1: English

The first study had two primary goals. The first was to replicate the English findings of Malt et al. (2008). For this reason, we created stimuli as similar as possible to their original human locomotion study. The second goal was to explore the nature of the semantic representations that may underlie naming patterns in terms of relevant features encoded by motion verbs. This required a manipulation of the video stimuli so as to manipulate cadence while controlling other gait features.

### Stimuli

Stimuli consisted of 21 videos of a college student on a treadmill at varying treadmill settings (1 mph increments from 2.5 to 8.5 mph), in three different playback conditions. Seven of the videos were unmanipulated and shown at Normal Playback. Using Final Cut Express video editing software, the remaining 14 videos were digitally manipulated to be in either ‘slow motion’ or ‘fast motion’. Seven videos were manipulated to Slow Playback, or 20%

slower than Normal Playback. The remaining 7 videos were manipulated to Fast Playback, or 20% faster than Normal Playback.<sup>1</sup>

The Slow and Fast Playback conditions are the critical manipulation in this study. In the Normal Playback condition, all features of human locomotion are coordinated. Digital manipulation disrupts this coordination by altering cadence, i.e., the number of steps per unit of time, while controlling other gait elements. (We recognize that cadence is a sub-parameter of gait, but for the purposes of this study we refer to cadence as separate from gait, where the latter remains a collection of co-occurring features such as stride length, knee bend, elbow bend, etc.). For example, with the 6.5 mph Treadmill Setting at Normal Playback, gait (stride length, knee bend, elbow bend, etc.) and cadence are in sync. In the Slow Playback, all of these elements except for cadence remain constant. The stride length, knee bend, and elbow bend are all identical to the Normal Playback condition. However, the cadence is different. There are fewer steps in the same amount of time. In the Fast Playback condition, there are more stride revolutions than the Normal Playback condition. In other words, this manipulation allows us to ‘mismatch’ cadence with other gait elements.

### Methods

Stimuli were shown to 30 native English-speaking undergraduates at the University of Colorado at Boulder. All undergraduates were monolingual in English with limited experience in a foreign language. Videos were randomized to prevent order biases from previous videos and were displayed using the online survey system Qualtrics. The videos were mixed with 8 distracters featuring the same actor on a treadmill engaging in activities such as crawling and skipping.

Upon presentation of each video, participants were asked to respond to the following question: “What is the man doing? He is \_\_\_\_.” Participants were asked to use as few words as possible when describing the motion, but using more than one word was allowed. Moreover, they were instructed to repeat any word they used as many times as they liked. All participants viewed all videos and the data from all participants was included in the final analysis.

### Results

All responses were grouped based on the head verb. Responses such as ‘running’ and ‘running quickly’ were all grouped as ‘running’. The ‘other’ category includes responses that appeared infrequently (five or fewer times)

---

<sup>1</sup>Prior to the study, naturalness ratings for manipulated videos were obtained. Nine undergraduates at the University of Colorado at Boulder were shown video stimuli and asked to answer the question “How natural is this motion?” by providing a rating on a scale from 1 (not natural) to 5 (very natural). The mean naturalness rating for manipulated videos was 2.8, indicating that while the manipulations were noticeable, they were not unnatural.

within a Treadmill Setting, such as ‘meandering’ or ‘moseying’.

The results of the first study reveal two important findings. First, the Normal Playback condition replicates the results of Malt et al.’s (2008) study. ‘Walking’ terms are used from 2.5 to 4.5 mph and ‘running’ terms are used from 5.5 to 8.5 mph.

Second, playback condition is shown to influence naming patterns on either side of the 4.5/5.5 mph boundary, such that some videos are named differently depending on playback condition. The overall effect of playback condition was confirmed by a binomial test ( $p < .005$ ).<sup>2</sup> Across all Treadmill Settings (2.5 to 8.5 mph), playback speed influences naming patterns. For example, at the Treadmill Setting of 6.5 mph, the term ‘jogging’ is preferred by 83% of the participants for the Slow Playback condition, 40% for Normal Playback, and only 5% for Fast Playback. A comprehensive view of the data can be seen in Figure 1.

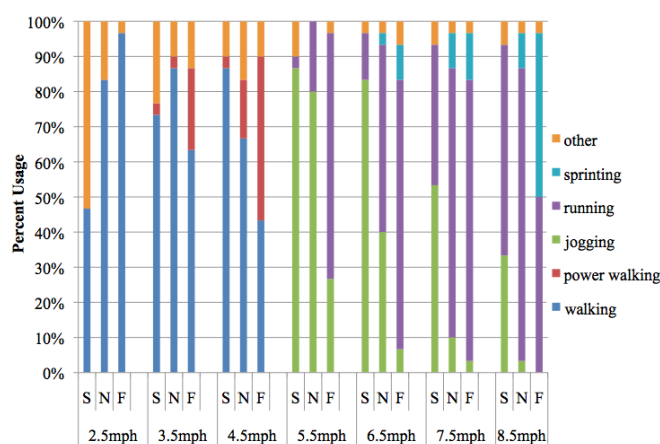


Figure 1: The English data from Study 1. Treadmill Settings from 2.5 mph to 8.5 mph and Playback Conditions are Slow (S), Normal (N), and Fast (F).

This provides a very clear picture of the structure that informs English lexemes: cadence seems to be a critical feature in many terms of human locomotion. Even when other gait features remain constant, a change in cadence can result in a change of the most common lexeme for that

<sup>2</sup> To conduct the binomial test, 12 undergraduates from CU Boulder were asked to provide a speed ranking of the lexemes from Study 1 (e.g., ‘running’ was rated as faster than ‘jogging’ by the majority of participants). These speed rankings were then compared to the data in Study 1. Across Treadmill Settings, videos in Slow Playback were more often paired with lexemes that were rated slower than the most common lexemes in Normal Playback (e.g., ‘jogging’ < ‘running’), while videos in Fast Playback were more often paired with lexemes that were rated faster than the most common lexemes in Normal Playback (e.g., ‘running’ < ‘sprinting’). The binomial test compared the number of times the lexeme changed in the predicted direction (e.g. ‘jogging’ < ‘running’ < ‘sprinting’) to the total number times there was a lexeme change due to playback condition.

event. Indeed, the manipulation causes additional semantic categories to appear such as ‘power walking’<sup>3</sup>, which is the most common response for 4.5 mph, but only in Fast Playback. At the very slowest cadence (2.5 mph, Slow Playback), participants are compelled to use a term other than ‘walking’, though there is less agreement as to what that term should be. This accounts for the large ‘other’ category here, consisting of words such as ‘meandering’, ‘sauntering’, and ‘moseying’.

Despite their attendance to the playback manipulation, participants did not use the same lexical item to refer to stimuli on either side of the 4.5/5.5 mph boundary. Rather, at this boundary, other gait features seem to be more critical than cadence. If cadence alone were a determining feature, we might expect to see ‘running’ terms applied to 4.5 mph in Fast Playback, but this is not the case. Increased cadence in this condition did not ‘override’ the category boundary, nor did decreased cadence in the 5.5 mph Slow Playback condition. In the English data, this is without exception.

Study 1 teases apart cadence from other gait elements and shows that a change in cadence influences naming patterns on either side of the biomechanical boundary. While Malt et al. (2008) indicate that strong structure in the world influences naming patterns, they are agnostic in terms of which features are attended to. Our results indicate that there is clearly ample structure to which speakers can selectively attend, and that a single feature can play a central or peripheral role in driving naming patterns.

## Study 2: German

By manipulating cadence while controlling other gait elements, Study 1 provides a concise picture of what external structural elements of human locomotion influence naming patterns. It also opens up the possibility that speakers of other languages will draw upon these structural features differently than English speakers. To explore this possibility, Study 2 replicates Study 1 in a German dialect.

### Stimuli

Study 2 used the same stimuli as Study 1.

### Methods

The videos were shown to 28 speakers of a Bavarian dialect of German known as Rieserisch. This dialect is spoken in the Ries area, the capital of which is the town of Nördlingen. Rieserisch is closely related to the more common German dialect of Schwäbisch, spoken primarily in the state of Baden-Württemberg (Schmidt, 1898). Though there are several important differences between the grammar and lexicon of Rieserisch, Schwäbisch, and standard German, specific contrasts between semantic

<sup>3</sup> ‘Power walking’ was not considered a modified form of ‘walking’, but rather a compound lexical term, in part because ‘power’ in this case does not pattern with other adverbial modifiers of ‘walking’, e.g., ‘quickly’, which can occur pre- or post-verbally.

representations of human locomotion verbs in these dialects remain largely unexplored. It is possible that Study 2's results could extend to speakers of more standard German, but this hypothesis requires further investigation.

Speakers ranged in age from 17 to 48. Participants viewed the videos on their own computers through the use of the survey system Qualtrics. Data from 4 speakers were discarded due to incompleteness. Of the remaining 24, all but 2 claimed to have relatively good knowledge of English. An additional 13 claimed knowledge of a third language, and 4 claimed knowledge of a fourth. Therefore, only 2 or the 24 participants could be described as monolingual. All, however, identified themselves as native speakers of the Rieserisch dialect.

Upon presentation of each video, participants were asked to respond to the following question: "Was macht der Mann? Er \_\_\_\_." (*What is the man doing? He is \_\_\_\_.*). Again, participants were asked to use as few words as possible when describing the motion, but using more than one was allowed. They were instructed to repeat any word they used as many times as they liked. All participants viewed all videos.

# Results

Again, all responses were grouped based on the head verb. The 'other' category includes responses that appeared infrequently (five or fewer times within a Treadmill Setting), such as 'spazieren' (*stroll*) and 'bummeln' (*saunter*).

To begin, use of the term 'laufen' gives rise to four noteworthy observations. First, contrary to Malt et al.'s (2008) predictions, the term 'laufen'—translated as both *walk* and *run*—was used to refer to stimuli on either side of the 4.5/5.5 mph boundary (see Figure 2). Close analysis indicates that 6 speakers used 'laufen' across the boundary, 14 used 'laufen' but did not cross the boundary, and 4 speakers did not use the lexeme at all. Therefore, usage of 'laufen' across the 4.5/5.5 mph boundary does not seem to be idiosyncratic or limited to one speaker. Second, the use of 'laufen' seems to be most common at 5.5 mph.

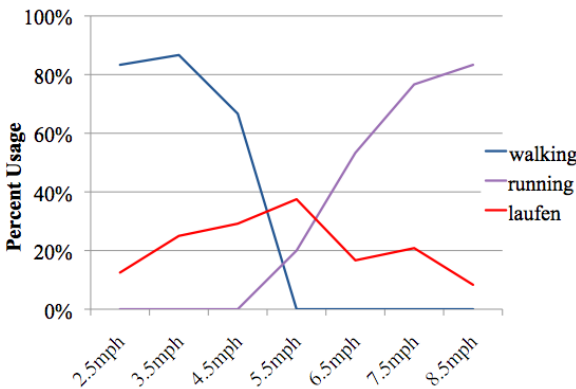


Figure 2: Comparison of English 'walking' and 'running' with German 'laufen' in Normal Playback.

Third, 'laufen' is used at every Treadmill Setting. While never the most frequent term in any given condition, 'laufen' is used with high frequency overall, equal to that of terms such as 'gehen' and 'joggen'. Fourth, 'laufen' does not appear to be affected by the cadence manipulation. Hypotheses concerning these observations will be addressed in the General Discussion.

The effect of playback condition for verbs other than 'laufen' was confirmed by a binomial test ( $p < .05$ ), indicating that a change in cadence affected the choice of lexeme.<sup>4</sup> The verb 'gehen' (translated as *go* or *walk*) seems to behave differently than English 'walk'. At 2.5 mph, the change in cadence did little to change naming patterns (as seen in Figure 3). The same is true for 4.5 mph: where English speakers designate a category of 'power walking' for 4.5 mph in the Fast Playback condition, German speakers do not seem to agree on a motion lexeme in this same condition. This suggests that 'gehen' may not encode cadence in the same way 'walk' does, and therefore its underlying representation may be qualitatively different.

Cadence effects for other verbs were similar to their English counterparts. The verbs 'joggen' (*jog*), 'rennen' (*run*), and 'sprinten' (*sprint*) were sensitive to the change in cadence. For example, at 6.5 mph, 'joggen' was used by 58% of the participants in the Slow Playback condition, 42% of participants in Normal Playback, and 17% in Fast Playback.

With the exception of 'laufen', German naming patterns align with those in other languages that mark the biomechanical distinction between 'walking' and 'running'. The term 'gehen' only appears at 4.5 mph and slower; terms such as 'joggen' and 'rennen' only appear from 5.5 to 8.5 mph. As in English, the cadence manipulation did not cause speakers to break this boundary.

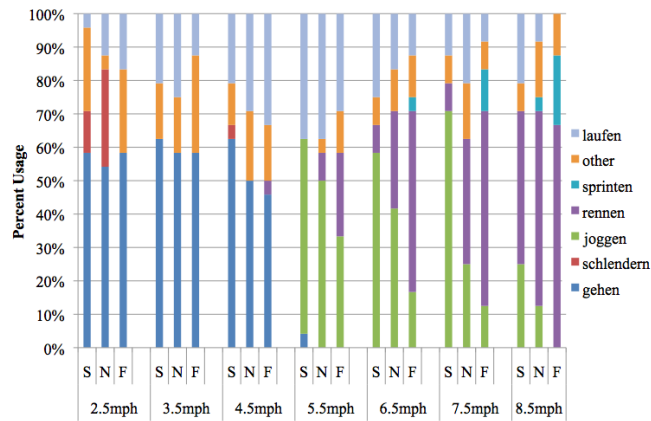


Figure 3: The German data from Study 2. Treadmill Settings from 2.5 mph to 8.5 mph and Playback Conditions are Slow (S), Normal (N), and Fast (F).

<sup>4</sup> Lexical rankings for the binomial test were provided by 3 additional native Rieserisch speakers who did not contribute to the elicitation task in Study 2.

## General Discussion

Both Study 1 and Study 2 bring relevant observations to bear on the nature of the real-world structure that informs semantic representations of human locomotion. They also help support and inform the findings of previous studies. We show that cadence is a structural feature to which English and German speakers attend and that this salience is reflected in naming patterns. Presumably, underlying concepts of these verbs will also highlight cadence in this way.

Both studies show the relative roles of cadence and other gait elements in naming patterns of continuous human locomotion. First, previous hypotheses that the biomechanical distinction between a ‘walking’ and ‘running’ gait (the 4.5/5.5 mph boundary) is the primary structure influencing speakers’ lexeme choices are strongly confirmed. The manipulation of Playback (and thus cadence) did not cause any speakers to use a ‘walking’ term from 5.5 mph and higher or a ‘running’ term from 4.5 mph and lower. Clearly, the biomechanical gait distinction is a critical aspect of speakers’ semantic representations of these human locomotion verbs at this point in the continuum of motion.

However, results also indicate that some English and German lexemes are extremely sensitive to the manipulation of cadence. In other words, digitally manipulating playback causes people to change the lexeme they use to describe the event (with respect to Normal Playback). This has important implications. First, it demonstrates that, though the 4.5/5.5 mph biomechanical distinction is of great importance, so too are cues of cadence. Moreover, the relative weight given to these features when naming depends on where individual events occur in the continuum of motion. At the 4.5/5.5 mph boundary, motion verbs are distinguished by gait features rather than cadence. On either side of this boundary, however, cadence drives naming patterns to some extent. This is not to say that *only* cadence drives naming patterns in these regions. Rather, it seems that the conjunction between cadence and gait elements (e.g., stride length, knee bend, elbow bend, etc.) is important in the encoding of motion verbs. That is, the semantic space is multidimensional, with cadence (perhaps perceived as speed) set against other gait elements. The combinatory nature of these dimensions gives rise to particular semantic categories, e.g., ‘jog’ is the conjunction between self-propelled, bounce-and-recoil gaits at medium cadence (see Malt et al., 2011 for similar treatments of this multidimensional space). Furthermore, contextual dimensions are undoubtedly critical (see Labov, 1973). For example, people may also attend to the weight of the person (under, average, or overweight), where the locomotion is taking place (indoors, outdoors, on a treadmill), or what they are wearing (casual or sports clothes, etc.). Therefore, subsequent studies in this domain should take into account the complicated nature of semantic representations. Rather than assume a priori that certain structure in the world will determine naming patterns, we suggest that descriptions of

semantic representations should proceed by induction, testing how the possible dimensions of semantic space may be encoded in a given language.

The German verb ‘laufen’, a difficult term for English speakers, is of particular note. It is commonly translated as both *walk* and *run*. However, as can be seen in Study 2, this translation is not accurate. There is little overlap in terms of the semantic space encoded by English ‘walk’ and ‘run’ in comparison to ‘laufen’. Therefore, direct translation is problematic or even impossible.

Though such a verb is not present in their data, Malt et al. (2008, p. 239) suggest that it may be “possible that some [languages] do not have separate words for walking and running gaits.” While German clearly does have words that mark the biomechanical distinction, ‘laufen’ is a frequent verb that crosses the 4.5/5.5 mph boundary. In fact, ‘laufen’ seems to be used most often at these treadmill settings, perhaps indicating a grouping of 4.5 and 5.5 mph to the exclusion of speeds such as 2.5 and 8.5 mph. This is in contrast to previous predictions regarding such a grouping.

One possible interpretation of this finding is that ‘laufen’ is a term of general motion. There are at least two reasons why this explanation is not likely. First, use of ‘laufen’ is used most frequently at 5.5mph. If it were a term of general motion, then it should be distributed evenly across all treadmill settings. Second, ‘laufen’ features similar metaphorical extensions as the English manner verb ‘run’, not of general motion verbs such as ‘go’. This is demonstrated in (3) and (4):

(3) Die Maschine läuft (*The machine is running*).

(4) Das Wasser läuft (*The water is running*).

These reasons are compelling evidence to dismiss the characterization of ‘laufen’ as a verb of general motion. A second response is that ‘laufen’ is not a manner verb at all. Instead, it is aspectual, meaning “put into motion without delay” (Cadiot et al. 2006, p. 182). Again, the metaphorical extensions above argue against this treatment, as ‘laufen’ is used to denote a continuous state, rather than indicating the placement of the event in time. Second, participants only viewed videos in which motion had already begun. The lack of transition from a non-motion state to a motion state in the videos runs counter to this aspectual reading. In other words, the videos do not indicate that the subject was ‘put’ into motion.

We propose, instead, that ‘laufen’ is a specific manner verb of continuous human locomotion that simply draws upon different structure than verbs in English, Dutch, Japanese and Spanish. However, due to the lack of response to the cadence manipulation, we are unable to posit which features in particular figure prominently in its semantic representation.

These results also have important implications with regard to so-called ‘manner’ and ‘path’ languages (Talmy, 1985). With these studies, we show that ‘manner of motion’ is not an unanalyzable primitive, as has been assumed in the

Linguistics literature (Slobin, 1996; Talmy, 1985) and the Psychology literature (Gennari et al., 2002; Papafragou & Selimis, 2010). Rather, ‘manner’ has a fine-grained structure, and the current studies are an attempt to tease apart this structure (following efforts of e.g., Ikegami, 1969).

## Conclusion

As close follow-ups to previous studies of motion verbs, the two studies presented here have confirmed prior findings while adding to our understanding of the cues to which speakers may attend when naming human locomotion. By manipulating cadence while controlling for other gait elements, Studies 1 and 2 demonstrate that cadence is a critical feature driving naming patterns in addition to the biomechanical distinction between ‘walking’ and ‘running’ gaits. The addition of German data provided unexpected results in the form of the lexeme ‘laufen’, which does not follow the trends found in previous literature. This finding highlights the importance of including as many languages as possible when analyzing semantic representations.

This study has concentrated on the semantic representations of human locomotion verbs. We believe, however, that these semantic representations may have important implications for underlying conceptual representations, as defined by Levinson (1997). It is possible that the differences in how people talk about human locomotion events may also influence how they think about these events. Therefore, this work can inform current research (e.g., Malt et al., 2011) on the relationship between semantic representations and underlying concepts.

## Acknowledgments

This work was supported by a student research grant from the Institute of Cognitive Science at The University of Colorado at Boulder. We thank Dr. Bhuvana Narasimhan, Dr. Gary McClelland, Jill Duffield, Dr. Les Sikos, Michael Thomas, Alison Hilger, Shaw Ketels and David Harper for support, advisement and discussion. German data collection was made possible by our contact in Germany, Brigitte Ulbricht. We would also like to thank the anonymous reviewers of the previous draft who pointed us to important research related to the current study.

## References

- Alexander, R. M. (1992). *The human machine*. New York: Columbia University Press.
- Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.
- Cadiot, P., Lebas, F., & Visetti, Y. (2006). The semantics of motion verbs: Action, space and qualia. In M. Hickman & S. Robert (Eds.), *Space in Languages: Linguistic Systems and Cognitive Categories*. Amsterdam & Philadelphia: John Benjamins Publishing Company.
- Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), *Nebraska symposium on motivation, Vol. 19*. Lincoln: University of Nebraska Press.
- Gennari, S., Sloman, S. A., Malt, B. C., & Fitch, W. T. (2002). Motion events in language and cognition. *Cognition*, 83, 49-79.
- Ikegami, Y. (1969). *The Semiological structure of the English verbs of motion*. New Haven: Yale University Press.
- Kay, P., Berlin, B., Maffi, L., & Merrifield, W. (1997). Color naming across languages. In C.L. Hardin & L. Maffi (Eds.), *Color categories in thought and language*. Cambridge: Cambridge University Press.
- Kiss, R., Kocsis, L., & Knoll, Z. (2004). Joint kinematics and spatial-temporal parameters of gait measured by an ultrasound-based system. *Medical Engineering & Physics*, 26, 611-620.
- Labov, W. (1973). The Boundaries of words and their meanings. In C-J. N. Bailey & R. Shuy (Eds.), *New ways of analyzing variation in English*. Washington, D. C.: Georgetown University Press.
- Levinson, S. C. (1997). From outer to inner space: Linguistic categories and non-linguistic thinking. In J. Nuyts & E. Pederson (Eds.), *Language and Conceptualization*. Cambridge: Cambridge University Press.
- Majid, A., Boster, J., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition*, 109, 235-250.
- Majid, A., Enfield, N. J., & van Staden, M. (Eds.). (2006). Parts of the body: Cross-linguistic categorization. *Language Sciences*, 28 (2-3) [Special issue].
- Malt, B. C., Ameel, E., Gennari, S., Imai, M., Saji, N., & Majid, A. (2011). Do words reveal concepts? In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 519-524). Austin, TX: Cognitive Science Society.
- Malt, B. C., Gennari, S., Imai, M., Ameel, E., & Tsuda, N. (2008). Talking about walking: Biomechanics and the language of locomotion. *Psychological Science*, 19, 232-241.
- Papafragou, A., & Selimis, S. (2010). Event categorization and language: A Cross-linguistic study of motion. *Language and Cognitive Processes*, 25, 224-260.
- Schmidt, F. G. G. (1898). *Die Rieser Mundart*. Munich: J Lindauersche Buchhandlung.
- Slobin, D. (1996). Two ways of travel: Verbs of motion in English and Spanish. In M. Shibatani & S. Thompson (Eds.), *Grammatical constructions: Their forms and meaning*. Oxford: Carendon Press.
- Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description: Vol. 3. Grammatical categories and the lexicon*. Cambridge: Cambridge University Press.

# “Less is More” in Bayesian word segmentation: When cognitively plausible learners outperform the ideal

Lawrence Phillips (lawphill@uci.edu)  
Department of Cognitive Sciences, 2235  
SBSG Irvine, CA 92697 USA

Lisa Pearl (lpearl@uci.edu)  
Department of Cognitive Sciences, 2314  
SBSG Irvine, CA 92697 USA

## Abstract

Purely statistical models have accounted for infants’ early ability to segment words out of fluent speech, with Bayesian models performing best (Goldwater et al. 2009). Yet these models often incorporate unlikely assumptions, such as infants having unlimited processing and memory resources and knowing the full inventory of phonemes in their native language. Following Pearl, et al. (2011), we explore the impact of these assumptions on Bayesian learners by utilizing syllables as the basic unit of representation. We find a significant “Less is More” effect (Pearl et al 2011; Newport 1990) where memory and processing constraints appear to help, rather than hinder, performance. Further, this effect is more robust than earlier results and we suggest this is due a relaxing of the assumption of phonemic knowledge, demonstrating the importance of basic assumptions such as unit of representation. We argue that more cognitively plausible assumptions improve our understanding of language acquisition.

**Keywords:** language acquisition; Bayesian modeling; cognitively plausible learning; less is more; statistical learning; word segmentation

## Introduction

Knowledge of words plays a crucial role in language acquisition but requires a child to identify words out of fluent speech. Children seem to accomplish this word segmentation very early (~7.5 months (Jusczyk & Aslin 1995; Echols et al. 1997; Jusczyk et al., 1993a)), and therefore many strategies have been proposed for this early success. One popular explanation for initial language learning relies purely on distributional information, rather than language-specific biases. This idea is bolstered by findings that infants keep track of the statistical regularities in speech (Saffran et al. 1996), and because languages vary greatly in their cues to word boundaries which would weaken the use of language specific knowledge. One very successful, purely distributional, learning approach uses Bayesian inference (Goldwater, Griffith & Johnson 2009 (GGJ), Pearl, Goldwater & Steyvers 2011 (PGS)). However, these Bayesian models incorporate modeling assumptions that are unlikely to be true. Both have assumed that the basic unit of representation available to the infant is the phoneme. We will argue from experimental evidence that syllables (or syllable-like representations) are a more natural representation for infants at this stage of acquisition. In

addition, GGJ conducted an ideal learner analysis, which assumes unlimited processing and memory resources for the learner. PGS investigated the impact of this assumption, finding a limited “Less is More” effect (Newport 1990) where cognitive resource limitations help, rather than hinder, some Bayesian learners. We examine the effect of the phoneme assumption in addition to these cognitive resource assumptions. We find not only that syllable-based Bayesian learners can do well at word segmentation but also a much more robust “Less is More” effect in our constrained Bayesian learners. This suggests that the unit of representation for models of language acquisition plays a crucial role. Here, using more cognitively plausible assumptions showcases a surprising learning effect, the “Less is More” effect that has been hypothesized to explain language acquisition success in children.

## The syllables as the representational unit

The first evidence that infants possess categorical representations of syllabic units appears at 3 months: Eimas (1999) finds that infants have categorical representations of syllables whereas infants at this age have no categorical representation of phonemes. Since word segmentation first occurs around 7.5 months (Jusczyk & Aslin 1995), it is likely that infants have robust access to syllables at this age. In contrast, knowledge of phonemes does not occur until approximately 10 months (Werker & Tees 1984) making it unlikely the learner has adult knowledge of their native language phonemes during the initial stages of word segmentation. Although it is possible that word segmentation and phoneme learning bootstrap from one another, we consider a more conservative approach which assumes infants only have access to syllabic information.

While the success of previous statistical word segmentation models is heartening, how dependent is their success on the assumption of the phoneme as a representational unit? With this question in mind, we modify existing phoneme-based statistical models of word segmentation that use Bayesian inference (GGJ, PGS) to operate over syllables. All of our modified Bayesian learners treat syllables as atomic units in the same way phonemes are thought of as atomic units. This mimics the performance of infants who are able to discriminate between syllables such as /ba/, /bu/, and /lu/, but who are unable to



recognize the phonemic similarity between /ba/ and /bu/ which does not exist between /ba/ and /lu/ (Jusczyk & Derrah 1987).

Utilizing syllables alleviates the learning problem somewhat because it reduces the number of potential boundary positions (e.g., a baby has three syllables but five phonemes). However, a potential sparse data problem then surfaces: A model operating over English phonemes must track statistics over approximately 40 units; a model operating over English syllables must track statistics over approximately 4000 units, while using less data than a phoneme-based model since there are fewer syllable tokens than phoneme tokens. This increases the statistical difficulty of the task tremendously. Additionally, because syllables are treated as atomic, almost all phonotactic information about English is lost in the model. Although previous work (e.g. Gambell & Yang 2006) shows that heuristic syllable-based models can perform quite well, it is unclear a priori whether a distributional learner with phonemes or syllables will produce better results for Bayesian word segmentation, due to the tradeoffs just mentioned.

In changing our unit of representation, we attempt to create a more psychologically faithful model of word segmentation. To foreshadow our results, we show that successful Bayesian word segmentation does not depend on the phoneme assumption. Moreover, by utilizing a more cognitively plausible unit of representation, we find a much more robust “Less is More” effect. The success of our models demonstrates the effectiveness of this purely statistical approach. Replicating and extending results from PGS concerning the surprising utility of processing constraints for Bayesian word segmentation. This suggests that the task of word segmentation may be structured to be more easily learned with strong memory limitations, such as those that infants have. Moreover, Bayesian models may be on the right track with respect to the kind of strategies infants are using during early word segmentation, since the learners demonstrate this “Less is More” behavior, and infants are thought to as well. In addition, the fact that this pattern of results was only hinted at by the phoneme-based models of PGS means that the unit of representation for models of language acquisition has a strong, non-trivial effect on the results found.

## Methods

### Corpus

We test our syllable-based models using English child-directed speech from the Pearl-Brent corpus (CHILDES: MacWhinney, 2000). This modification of the Brent corpus contains 100 hours of child-directed speech from 16 mother-child pairs. We restrict ourselves, however, to child-directed utterances before 9 months of age, leaving 28,391 utterances (3.4 words per utterance, 10.4 phonemes per utterance, 4.2 syllables per utterance, on average).

While there are many ways to syllabify a corpus automatically, we opted for a two-stage approach. First,

where we have human judgments of syllabification we used them; second, when not, we automatically syllabify our corpus in a language-independent way. We take human judgments of syllabification from the MRC Psycholinguistic Database (Wilson 1988), but not all words in the Pearl-Brent corpus have syllabifications in the MRC dictionary. To solve this problem we used the Maximum-Onset Principle to syllabify all remaining words. This principle states that the onset of any syllable should be as large as possible while still remaining a valid word-initial cluster. We use this principle out of convenience for the kind of syllabification that infants might possess. Given a lack of experimental evidence as to the exact nature of infant syllabification, this representation is likely only an approximation. Approximately 25% of lexical items were syllabified automatically. Only 3.6% of human judgments on our items differ from automatic syllabification. Each unique syllable is then treated as a single, indivisible unit losing all sub-syllabic phonetic (and phonotactic) information.

### Models

Bayesian models are well suited to questions of language acquisition because they explicitly distinguish between the learner’s pre-existing beliefs (prior) and how the learner evaluates incoming data (likelihood), using Bayes’ theorem:

$$P(h|d) \propto P(d|h)P(h)$$

The Bayesian learners we use are those of GGJ as well as the constrained learners of PGS. All learners are based on the same underlying hierarchical Bayesian models developed by GGJ. The first of these models assumes independence between words (a *unigram* assumption) while the second assumes words depend only on the word before them (a *bigram* assumption). To encode these assumptions into the model, GGJ use a *Dirichlet Process* (Ferguson, 1973), which supposes that the observed sequence of words  $w_1 \dots w_n$  is generated sequentially using a probabilistic generative process. In the unigram case, the identity of the  $i$ th word is chosen according to:

$$P(w_i = w | w_1 \dots w_{i-1}) = \frac{n_{i-1}(w) + \alpha P_0(w)}{i-1 + \alpha} \quad (1)$$

where  $n_{i-1}(w)$  is the number of times  $w$  appears in the previous  $i-1$  words,  $\alpha$  is a free parameter of the model, and  $P_0$  is a *base distribution* specifying the probability that a novel word will consist of the phonemes  $x_1 \dots x_m$ :

$$P(w = x_1 \dots x_m) = \prod_{j=1}^m P(x_j) \quad (2)$$

In the bigram case, a *hierarchical Dirichlet Process* (Teh et al. 2006) is used. This model additionally tracks the frequencies of two-word sequences and is defined as in:

$$P(w_i = w | w_{i-1} = w', w_1 \dots w_{i-2}) = \frac{n_{i-1}(w', w) + \beta P_1(w)}{n_{i-1}(w') + \beta} \quad (3)$$



$$P_1(w_i = w) = \frac{b_{i-1}(w) + \gamma P_0(w)}{b_{i-1} + \gamma} \quad (4)$$

where  $n_{i-1}(w', w)$  is the number of times the bigram  $(w', w)$  has occurred in the first  $i - 1$  words,  $b_{i-1}(w)$  is the number of times  $w$  has occurred as the second word of a bigram,  $b_{i-1}$  is the total number of bigrams, and  $\beta$  and  $\gamma$  are free model parameters.

In both the unigram and bigram case, this generative model implicitly incorporates preferences for smaller lexicons by preferring words that appear frequently (due to (1) and (3)) as well as shorter words in the lexicon (due to (2) and (4)). The ideal learner based on this model is fit using Gibbs sampling (Geman & Geman 1984), run over the entire corpus, sampling every potential word boundary 20,000 times. GGJ found that their bigram ideal learner performed better than their unigram ideal learner, so we begin by examining this distinction in our syllable-based Bayesian learners. In addition, we will consider the constrained learners that PGS investigated—incorporating processing and memory constraints.

The Dynamic Programming Maximization (DPM) learner incorporates a basic processing limitation: linguistic processing occurs online rather than in batch after a period of data collection. Thus, the DPM learner processes one utterance at a time, rather than processing the entire corpus at once. This learner uses the Viterbi algorithm to converge on the optimal word segmentation for the current utterance, conditioned on the utterances seen so far. In all other aspects, the DPM learner is essentially identical to the Ideal model: it has perfect memory for previous utterances and unlimited processing resources.

The Dynamic Programming Sampling (DPS) learner is similar to the DPM learner in processing utterances incrementally, but is additionally motivated by the idea that infants, and human beings in general, are not ideally rational. This could mean that infants do not *always* select the best segmentation. Instead, infants select segmentations probabilistically. So, they will often choose the best segmentation but occasionally choose less likely alternatives, based on the likelihood of the various segmentation alternatives. To implement this, the DPS learner uses the Forward algorithm to compute the likelihood of all possible segmentations and then chooses a segmentation based on the calculated distribution.

The Decayed Markov Chain Monte Carlo (DMCMC) learner also processes data incrementally, but uses a DMCMC algorithm (Marthi et al. 2002) to implement a memory constraint. This learner is similar to the original GGJ ideal learner in that it uses Gibbs sampling. However, the DMCMC learner does not sample all boundaries; instead, it samples some number  $s$  of previous boundaries using the decayed function  $b_a^{-d}$  to select the boundary to sample, where  $b_a$  is the number of potential boundary locations between  $b$  and the end of the current utterance  $a$  and  $d$  is the decay rate. Thus, the further  $b$  is from the end of the current utterance, the less likely it is to be sampled.

Additionally, larger values of  $d$  indicate a stricter memory constraint. All our results here use a set, non-optimized value for  $d$  of 1.5, which was chosen to implement a heavy memory constraint. Having sampled a set of boundaries, the DMCMC learner can then update its beliefs about those boundaries and subsequently update its lexicon.<sup>1</sup> Because of the decay function, the DMCMC's sampling is biased towards boundaries in recently seen utterances and thus the DMCMC learner implements a recency effect.

In addition to comparing our syllable-based learners against the original phoneme-based learners, we also compare our learners against other syllable-based learners. The first baseline is the Transitional Probability (TP) model based on Gambell & Yang (2006), which calculates TPs over syllables and places boundaries at all local minima. Our second baseline is a “Syllable=Word” learner which simply assumes that all syllables are words (a strategy that can be very useful in languages containing many monosyllabic words, like English).

## Results

We measure our results in terms of precision, recall and F-score, where precision is defined as (5) and recall is defined as (6):

$$Precision = \frac{\# correct}{\# guessed} \quad (5)$$

$$Recall = \frac{\# correct}{\# actual} \quad (6)$$

F-score is the harmonic mean of the two:

$$F - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (7)$$

Precision and recall are considered jointly, through the harmonic mean, because it is possible for learners to succeed on one measure while failing on the other. For instance, a learner that posits only a single boundary scores 100% on precision if that boundary is correct. In comparison, the same learner will have just over 0% recall. Similarly, a learner could posit boundaries at every position, producing a 100% recall with very low precision because many of the boundaries were false. As the F-score balances these two measures, a high F-score indicates the learner is succeeding at both precision and recall. We can make these measurements over individual word tokens, word boundaries, and lexical items.

In order to prevent overfitting, we train each learner on 90% of the corpus and then test the learner on the remaining 10%. This train-test validation was done five times for each learner. Given the probabilistic nature of our learners, all

<sup>1</sup> All DMCMC learners sample  $s=20,000$  boundaries per utterance. According to PGS, this works out to approximately 89% less processing than the original ideal learner in GGJ, which samples every boundary 20,000 times.

results presented here are averaged over the five iterations to ensure the validity of each learner’s performance.

Table 1 shows the F-score for word tokens over all of the syllable-based learners. First, we observe that, in all cases, the Bayesian bigram learners outperform their unigram equivalents. In the unigram case, all constrained learners (DPM, DPS, DMCMC) significantly outperform the ideal learner; in contrast, in the bigram case this is true for the DMCMC learners only. This indicates that constrained learning helps generally if statistics cannot be tracked across words. However, if bigram statistics can be tracked, a memory constraint is only beneficial for the DMCMC strategy. Additionally, all learners outperform the TP baseline learner and all bigram learners outperform the Syl = Word baseline.

	Unigram	Bigram
Ideal	53.12	77.06
DPM	<b>58.76</b>	75.08
DPS	<b>63.68</b>	77.77
DMCMC	<b>55.12</b>	<b>86.26</b>
TP	43.98	
Syl=Word	72.41	

Table 1. Word token F-scores across all syllable-based models. Constrained Bayesian learners that significantly outperform their ideal counterpart ( $p < .05$ ) are in bold.

	Syl-U	Phon-U	Syl-B	Phon-B
Ideal	53.1	<b>54.8</b>	<b>77.1</b>	71.5
DPM	58.8	<b>65.9</b>	<b>75.1</b>	69.4
DPS	<b>63.7</b>	58.5	<b>77.8</b>	39.8
DMCMC	55.1	<b>67.8</b>	<b>86.3</b>	73.0

Table 2. Token F-scores for syllable-based (Syl) vs. phoneme-based (Phon) models, comparing Unigram (U) and bigram (B) learners. Learners in bold outperform their baseline counterparts.

Clearly our syllable-based learners perform well, but are syllables a better unit of representation than phonemes for this task? Table 2 compares our syllable-based learners with the original phoneme-based models of PGS. We see that in the unigram case, phoneme-based learners outperform their syllable-based counterparts, except in the case of the DPS learner. In the bigram case, however, all syllable-based models outperform their phoneme-based equivalents. This suggests that the bigram assumption is crucial to a syllable-based learner. We speculate that this is due to an additional source of information that the bigram learner has access to. In particular, because the unigram learner assumes that words are independent of one another, the TPs between syllables are the only source of boundary information. Because there are roughly 4000 syllables, there will often be cases where a problem of sparse data arises. In contrast, the bigram learner has access to the boundary information

inherent in word bigrams, in addition to TPs. These word bigrams may help supplement the sparseness of the TP data.

A desired behavior for all learners is undersegmentation since children are known to undersegment the input they receive (Peters 1983). All of our Bayesian learners exhibit this behavior. This can be seen by comparing the values of boundary precision and recall. High boundary precision (indicating the boundaries are often correct) but low recall (indicating not enough boundaries are put in) indicates general undersegmentation, whereas high boundary recall (indicating a lot of boundaries are put in) but low boundary precision (indicating the boundaries are not often correct) indicates oversegmentation. Although this trend of undersegmentation exists for both unigram and bigram learners, we present data only on our bigram learners since the results are qualitatively similar. Table 3 shows the boundary precision and recall for all Bayesian bigram and comparison learners. The Syl=Word baseline learner tends to oversegment, so although it performs much better than the TP learner and the Bayesian unigram learners, its error pattern does not match what we expect from infants. In contrast, all of our Bayesian learners are producing more undersegmentations than oversegmentations. Table 4 presents sample segmentation errors from the Ideal and DMCMC bigram learners.

	Boundary Precision	Boundary Recall
Ideal	<b>96.50</b>	80.45
DPM	<b>96.49</b>	76.21
DPS	<b>95.78</b>	79.72
DMCMC	<b>94.11</b>	91.57
TP	<b>90.00</b>	53.14
Syl = Word	76.26	<b>100</b>

Table 3. Boundary precision and recall for all bigram Bayesian and comparison learners.

Bigram Ideal	Bigram DMCMC
<i>putit</i> away	put it away
<i>Iloveyou</i>	I love you
<i>Let’ssee</i> what that <i>feltlike</i>	Let’s see what that <i>feltlike</i>
<i>If youdon’t</i> like it	<i>Ifyou</i> don’t like it

Table 4. Example output from Bigram Ideal and DMCMC learners. Undersegmentation is marked in italics.

To explain the difference between our ideal and constrained learner results, we can examine the token and lexicon item recall scores, as shown in Table 5. We observe that both the DMCMC learners identify fewer word types than their ideal learner counterparts, as shown by their comparatively low lexicon recall scores. The token recall score for the DMCMC learners, however, is higher than their ideal learner counterparts. Since this requires the DMCMC learners to identify more word tokens from a smaller stock of lexical items, it can be inferred that these

DMCMC learners are identifying more frequently occurring words than the ideal learners.

	Token Recall	Lexicon Recall
Uni-Ideal	44.96	73.44
Uni-DMCMC	48.09	68.9
Bi-Ideal	72.47	79.69
Bi-DMCMC	85.43	76.84

Table 5. Token and lexicon recall for the Ideal and DMCMC learners. Lower lexicon recall with higher token recall implies that the DMCMC learners identify more frequently occurring words.

## Discussion

Our results support two broad findings. First, we find that memory-constrained learners outperform their “ideal” equivalents, which we take as support for the “Less is More” hypothesis (Newport 1990). In particular, limited cognitive resources, rather than hurting learner, seem to help word segmentation. Second, because this effect was obscured in the phoneme-based learners of PGS, we argue that the unit of representation posited by a model of language acquisition has a crucial impact on the results found. In particular, making more cognitively plausible assumptions may yield answers to the puzzling behaviors we observe—namely, that children, who are more cognitively limited than adults, nonetheless are far more successful at language acquisition.

What exactly is causing the “Less is More” effect here? Perhaps it is due to the properties of online vs. batch unsupervised probabilistic learning algorithms. Liang & Klein (2009) show that for unsupervised models using Expectation-Maximization, online models not only converge more quickly than batch models, but, also in cases as varied as word segmentation, part-of-speech induction and document classification, can actually outperform their batch equivalents. However, this explanation fails to account for our results in two ways: (a) the most direct online equivalent of our batch model (the DPM learner) actually performs worse than the Ideal model, and (b) this does not explain the performance boost caused by sub-optimal segmentation (the DPS learner).

Perhaps the answer lies in the kinds of words these models identify. We find, as in table 5, that our ideal bigram learner segments 72.5% of the words in the input, building a lexicon that contains 80% of the actual word-types it encounters. Yet we find that a learner with memory constraints (the DMCMC learner) can successfully segment 85% of the words in the input, although this makes up only 76.8% of the word-types encountered. This suggests that while an ideal learner identifies more lexical items, the memory-constrained learner identifies more *frequent* lexical items. Not only is this true in both the unigram and bigram syllable-based learners, but it is also true of the equivalent phoneme-based learners of PGS. The robustness of this

phenomenon suggests that, irrespective of the representational unit, memory-constrained learners are biased towards identifying more commonly occurring units, a potentially useful bias in language acquisition.

In effect, this strategy in word segmentation may help in learning the *important* things. Although this has been hypothesized by the literature on “Less is More” in artificial language learning (Kersten & Earles 2001; Cochran et al. 1999), we are unaware of experimental support for why constrained processing helps in real language acquisition. The fact that we can help to explain, from a computational perspective, why “Less is More” is beneficial highlights a very major contribution computational modeling can make to developmental research more generally.

For our claim regarding the impact of the unit of representation, we can compare the syllable-based learner results with those of phoneme-based learners. Table 2 highlights a number of crucial distinctions. First, and most basically, syllable-based learners perform well, and in the bigram case better than phoneme-based learners. This suggests that the tradeoff between number of potential boundaries and number of potential transitional probabilities works out in favor of the syllable-based learner. This underscores the utility of a Bayesian inference strategy for the initial stages of word segmentation – without access to phonotactics, stress, acoustic cues, or innate linguistic knowledge, a learner can be very successful at segmenting words from fluent speech.

Still, there is a major difference in the performance of the sub-optimal (DPS) learner – the syllable-based DPS learner has comparable performance to the Ideal learner while its phoneme-based equivalent suffers greatly. We speculate that this is due to the number of potential segmentations the phoneme-based learner considers, compared to the syllable-based learner since the DPS learner chooses a segmentation probabilistically, the phoneme-based learner may be more easily led astray in the initial stages of segmentation, and never recover. In addition, we also notice a strengthening of the “Less is More” effect in the syllable-based learner, compared to its phoneme-based counterpart (Ideal vs. DMCMC). By making more realistic assumptions about the learner’s unit of representation, we also create a learner that exhibits the kind of behavior that infants show. This highlights one benefit of pursuing more cognitively plausible computational models, as opposed to models that are more idealized.

In that vein, there are a number of areas where we could improve the existing syllable-based Bayesian learners. First, some segmental cue information is likely available to infants such as phonotactics or articulatory cues. Similarly, suprasegmental cues such as primary stress are known to affect infant word segmentation (Jusczyk et al. 1999) and there is evidence that stressed and unstressed syllables are represented separately in infants (Pelucchi, Hay, & Saffran 2009). Finally, the exact form which infants use to represent syllables is unclear. While it is our view that syllabification must be learned by infants, we make no attempt here to

explain by what means this occurs. When one looks cross-linguistically, languages treat syllabification in very different ways. In addition, languages vary significantly on the number of syllable types they have – languages such as English number their unique syllables in the thousands, while some languages, like Japanese, have very few unique syllables. To ensure that our pattern of results is truly representative of word segmentation *generally* and not just in English, syllable-based word segmentation models must be tested across multiple languages.

In conclusion, this study highlights the benefits of using empirical research from psychology to inform decisions on how to model language acquisition: not only can we identify the strategies that are likely to be used by children, but we may also discover potential explanations for existing, sometimes puzzling, observations about child language acquisition, as with the “Less is More” hypothesis.

### Acknowledgments

We would like to thank Caroline Wagenaar and James White for their help on syllabifying the Pearl-Brent corpus. In addition, we are very grateful to Robert Daland, Constantine Lignos, and the audiences at the Psycho-Computational Models of Human Language Acquisition 2012 and the Linguistic Society of America meeting in 2012 for their helpful comments.

### References

- Brent, M.R. & Siskind, J.M. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81, 31-44.
- Cochran, B., McDonald, J. & Parault, S. 1999. Too smart for their own good: The disadvantage of superior processing capacity for adult language learners. *Journal of Memory and Language*, 41, 30-58.
- Echols, C.H., Crowhurst, M.J. & Childers, J.B. 1997. The perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language*, 36, 202-225.
- Eimas, P.D. 1999. Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustical Society of America*, 105(3), 1901-1911.
- Ferguson, T. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 209-230.
- Frank, M. C., Goodman, N. D., & Tenenbaum, J. 2009. Using speakers’ referential intentions to model early cross situational word learning. *Psychological Science*, 20, 579-585.
- Gambell, T. & Yang, C. 2006. Word Segmentation: Quick but not dirty. Manuscript. New Haven: Yale University
- Geman S. & Geman D. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.
- Goldwater, S., Griffiths, T. & Johnson, M. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112(1), 21-54.
- Jusczyk, P.W. & Derrah, C. 1987. Representation of speech sounds by young infants. *Developmental Psychology*, 23(5), 648-654.
- Jusczyk, P.W., Cutler, A. & Redanz, N.J. 1993a. Infants’ preference for the predominant stress patterns of English words. *Child Development*, 64(3), 675-687.
- Jusczyk, P.W., Friederici, A.D., Wessels, J.M.I., Svenkerud, V.Y. & Jusczyk, A.M. 1993b. Infants’ sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, 32, 402-420.
- Jusczyk, P.W., Luce, P.A. & Charles-Luce, J. 1994. Infants’ sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33, 630-645.
- Jusczyk, P.W., Houston, D.M. & Newsome, M. 1999. The beginnings of word segmentation in English learning infants. *Cognitive Psychology*, 39, 159-207.
- Kersten, A.W. & Earles, J.L. 2001. Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language*, 44, 250-273.
- Liang, P. & Klein, D. 2009. Online EM for unsupervised models. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, 611-619.
- MacWhinney, B. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Marthi, B., Pasula, H., Russell, S. & Peres, Y., et al. 2002. Decayed MCMC filtering. In *Proceedings of 18<sup>th</sup> UAI* 319-326.
- Newport, E. 1990. Maturation constraints on language learning. *Cognitive Science*, 14, 11-28.
- Pearl, L., Goldwater, S., & Steyvers, M. 2011. Online Learning Mechanisms for Bayesian Models of Word Segmentation, *Research on Language and Computation*, special issue on computational models of language acquisition. DOI 10.1007/s11168-011-9074-5.
- Pelucchi, B., Hay, J., & Saffran, J. 2009. Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, 113, 244-247.
- Peters, A. 1983. *The Units of Language Acquisition*, *Monographs in Applied Psycholinguistics*, New York: Cambridge University Press.
- Saffran, J.R., Aslin, R.N. & Newport, E.L. 1996. Statistical learning by 8-Month-Old Infants. *Science*, 274, 1926-1928.
- Teh, Y., Jordan, M., Beal, M., & Blei, D. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 1566-1581.
- Wilson, M.D. 1988. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2, *Behavioral Research Methods, Instruments and Computers*, 20 6-11.
- Werker, J.F. & Tees, R.C. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development*, 7, 49-63.
- Xu, F. & Tenenbaum, J.B. 2007. Word learning as Bayesian inference. *Psychological Review*, 114(2), 245-272.

# Categorical compositionality continued (further): A category theory explanation for the systematicity of recursive cognitive capacities

Steven Phillips (steve@ni.aist.go.jp)

Mathematical Neuroinformatics Group, National Institute of Advanced Industrial Science and Technology (AIST),  
Tsukuba, Ibaraki 305-8568 JAPAN

William H. Wilson (billw@cse.unsw.edu.au)

School of Computer Science and Engineering, The University of New South Wales,  
Sydney, New South Wales, 2052 AUSTRALIA

## Abstract

Human cognitive capacity includes recursively definable concepts, which are prevalent in domains involving lists, numbers, and languages. Cognitive science currently lacks a satisfactory explanation for the systematic nature of recursive cognitive capacities. The category-theoretic constructs of initial  $F$ -algebra, catamorphism, and their duals, final coalgebra and anamorphism provide a formal, systematic treatment of recursion in computer science. Here, we use this formalism to explain the systematicity of recursive cognitive capacities without ad hoc assumptions (i.e., why the capacity for some recursive cognitive abilities implies the capacity for certain others, to the same explanatory standard used in our account of systematicity for non-recursive capacities). The presence/absence of an initial algebra/final coalgebra implies the presence/absence of all systematically related recursive capacities in that domain. This formulation also clarifies the theoretical relationship between recursive cognitive capacities. In particular, the link between number and language does not depend on recursion, as such, but on the underlying functor on which the group of recursive capacities is based. Thus, many species (and infants) can employ recursive processes without having a full-blown capacity for number and language.<sup>1</sup>

**Keywords:** systematicity; category theory; endofunctor,  $F$ -(co)algebra; (final) initial algebra; universal construction; catamorphism; anamorphism; classicism; connectionism

## Introduction

Many cognitive domains include recursively definable concepts (i.e., concepts defined with reference to themselves), such as domains involving lists, or languages. In card games, for example, a deck of cards can be defined (recursively) as a top card (perhaps turned face up to reveal its value) and a (remaining) deck of cards. To include finite decks, the definition has an alternative clause specifying an empty deck; that is, a deck is either empty, or contains a top card and a (smaller) deck. Operations on recursively defined concepts may also be defined recursively. For example, removing jokers from a deck of cards can be defined (recursively) as removing the top card if it is a joker and then removing jokers from the remaining deck of cards. Given that you don't find people who can remove the jokers from a hand of seven cards without being able to remove jokers from a deck of fifty-three, recursion-related capacities are further instances (see below) of the systematic nature of human cognition.

Systematicity is a property of human cognitive architecture (i.e., the basic processes and modes of composition that

together afford cognition) whereby cognitive capacity is organized around groups of related abilities. A standard example since the original formulation of the problem (Fodor & Pylyshyn, 1988) has been that you don't find people with the capacity to infer John as the lover from the statement *John loves Mary* without having the capacity to infer Mary as the lover from the related statement *Mary loves John*. An instance of systematicity is when a cognizer has cognitive capacity  $c_1$  if and only if the cognizer has cognitive capacity  $c_2$  (see McLaughlin, 2009). So, e.g., systematicity is evident where one has the capacity to remove the jokers if and only if one has the capacity to remove the aces (assuming, of course, one has the capacity to identify jokers and aces).

The classical explanation for systematicity has two components: (1) combinatorial syntactically structured representations; and (2) processes that are sensitive to (i.e., compatible with) those syntactic structures. In a classical cognitive architecture, mental representations of constituent entities (e.g., *John*, *Mary*) are tokened (instantiated) whenever the mental representations of their complex hosts (e.g., *John loves Mary*) are tokened, with the meaning of a complex host representation obtained (recursively) from the meaning assigned to its constituent mental representations and their syntactic relationships. By analogy to language, this form of mental representation is called a *language of thought* (LoT) (Fodor, 1975).

The three aspects of systematicity, i.e., *systematicity of representation*, *systematicity of inference*, and *compositionality of representation* (Fodor & Pylyshyn, 1988), can often be derived from classical cognitive architectures, because the same component processes are often used for each and every member of a group of systematically-related capacities. For instance, a classical system has the capacity to represent *John loves Mary* if and only if the system has a capacity to represent *Mary loves John* when the common component process is something like a production rule:  $S \rightarrow \text{Agent loves Patient}$  (where *John* and *Mary* are produced from *Agent* and *Patient* by other production rules)—systematicity of representation. Likewise, a classical system has the capacity to infer John as the lover in *John loves Mary* if and only if it has the capacity to infer Mary as the lover in *Mary loves John* given a common process that is sensitive to the syntactic structure whereby the lover constituent is represented by the first token—systematicity of inference. And, again, the capacity

<sup>1</sup>This paper is a short version of Phillips and Wilson (2012).

to assign the semantic content of John being the lover of Mary to the representation *John loves Mary* if and only if there is the capacity to assign the semantic content of Mary being the lover of John to the representation *Mary loves John* derives from the tokening principle (above) mediating classical representations and processes: the process for juxtaposing tokens (symbols) *John*, *loves*, and *Mary* to form *John loves Mary* with corresponding semantic content is the same process that is used to form *Mary loves John* with corresponding content.

Classical compositionality would seem to provide an elegant explanation for systematicity with regard to recursive capacities, even though it fails to provide a full account of systematicity generally (Aizawa, 2003).<sup>2</sup> For recursive definitions, like the deck of cards, one self-referencing rule typically covers all cases (bar the terminating case, such as the empty deck). For example, removing jokers from a single hand, or an entire deck invokes the same component process. The two tasks only differ in the number of recursive steps.

### Classical, but not systematically recursive

However, the classical explanation with regard to recursive capacities still suffers the same general problem that it suffers for non-recursive capacities. To illustrate, suppose one card game requires removing the lowest value card in the hand dealt, while another card game requires removing the highest value card. Suppose the following recursive procedure, *lowest*, for identifying the lowest valued card in a deck of cards (containing at least one card):

$$\begin{aligned} \text{lowest}(c : cs) &= \text{lower}(c, \text{lowest}(cs)) \\ \text{lowest}(c : []) &= c \end{aligned}$$

where a deck of cards  $c : cs$  is represented by a recursively defined list with  $c$  as the top card and  $cs$  as the remaining deck,  $[]$  is the empty deck, and *lower* returns the lower of two cards. Suppose, also, the following classical non-recursive procedure, *highest*, for identifying the highest valued card:

```
highest(cs) = (i, high) ← (0, undefined)
while i < n do
    (i, high) ← (i + 1, higher(high, csi))
return high
```

where deck  $cs$  is represented by an array of  $n$  cards with position indexed by  $i$  (i.e.,  $cs_i$  is the  $i$ th card), *high* maintains a representation of the (currently) highest card, *higher* returns the higher of two cards (*undefined* is some value guaranteed to be lower than any card), and  $\leftarrow$  indicates variable-value assignment. Clearly, the two procedures do not share

<sup>2</sup>Classical theory fails to provide a complete explanation because one can construct syntactically compositional systems that support some but not all members of a group of systematically-related cognitive capacities. Additional (*ad hoc*) assumptions are needed to derive only those classical cognitive architectures that support systematicity (Aizawa, 2003). This problem echoes the one originally raised against connectionism as a *theory* of cognitive architecture (Fodor & Pylyshyn, 1988; Fodor & McLaughlin, 1990).

any component processes, and so do not provide a basis for systematicity, even though systematicity could be supported when both tasks are implemented in either the first style only, or the second style only. Notice that we are not unfairly stressing classical theory by apportioning capacity at the level of constituents—systematicity concerns “molecular”, not “atomic” capacities (Fodor & Pylyshyn, 1988). Rather, given constituent capacities *lower* and *higher*, classical theory admits two independent compositional forms, as the example illustrates.

In this paper, we extend our category theory explanation to recursive capacities using universal constructions called an *initial F-algebra* and a *final F-coalgebra*, which have been extensively developed in computer science as a theoretical basis for recursive computations (Manes & Arbib, 1986). Our previous work (Phillips & Wilson, 2010, 2011) dealt with non-recursive domains using a kind of universal construction called *adjoint functors*—a *functor* is a way relating categories, which can be viewed as a way of constructing objects and morphisms from one category based on those in another. The current work uses *endofunctors*, which relate categories to themselves, hence their relevance to recursion.

### Category theory: *F*-(co)algebras

A category theory treatment of recursion starts with the concept of an *F-algebra* constructed on an endofunctor  $F$ .<sup>3</sup>

**Definition (*F*-algebra).** For an endofunctor  $F : \mathbf{C} \rightarrow \mathbf{C}$ , an *F-algebra* is a pair  $(A, \alpha)$ , where  $A$  is an object and  $\alpha : F(A) \rightarrow A$  is a morphism in  $\mathbf{C}$ .

For example, if  $F(A) = A \times A$ , then the addition operator  $+: A \times A \rightarrow A$  is an *F-algebra*.

**Definition (*F*-algebra homomorphism).** An *F-algebra homomorphism*  $h : (A, \alpha) \rightarrow (B, \beta)$  is a morphism  $h : A \rightarrow B$  (in  $\mathbf{C}$ ) such that the following diagram *commutes*:

$$\begin{array}{ccc} F(A) & \xrightarrow{\alpha} & A \\ F(h) \downarrow & & \downarrow h \\ F(B) & \xrightarrow{\beta} & B \end{array} \quad (1)$$

That is,  $h \circ \alpha = \beta \circ F(h)$ .

**Definition (*F*-algebra category).** For endofunctor  $F : \mathbf{C} \rightarrow \mathbf{C}$ , an *F-algebra category*  $\mathbf{Alg}(F)$  has *F-algebras*  $(A, \alpha)$  for objects, and *F-algebra homomorphisms*  $h : (A, \alpha) \rightarrow (B, \beta)$  for morphisms.

**Definition (*Initial algebra*).** An *initial F-algebra*  $(A, in)$ , also called *initial algebra*, is an *initial object* in the category

<sup>3</sup>We omit definitions of *category*, *functor*, and *initial object* (see Phillips & Wilson, 2010, 2011, for introductions tailored to the systematicity problem—short versions appear in CogSci10/CogSci11 proceedings). An example category is **Set**, whose *objects* are sets, *morphisms* are functions, and *composition* of morphisms is function composition. Functors are like generalized functions, but map morphisms to morphisms as well as objects to objects. An *endofunctor* is a functor whose domain and codomain are the same category.

of  $F$ -algebras  $\mathbf{Alg}(F)$ . I.e., there exists a unique  $F$ -algebra homomorphism from  $(A, in)$  to every  $F$ -algebra in  $\mathbf{Alg}(F)$ .

**Definition (Catamorphism).** A *catamorphism*  $h : (A, in) \rightarrow (B, \beta)$  is the unique  $F$ -algebra homomorphism from initial  $F$ -algebra  $(A, in)$  to  $F$ -algebra  $(B, \beta)$ . That is,  $h \circ in = \beta \circ F(h)$ , and the uniquely specified  $h$  for each such  $\beta$  is denoted  $\text{cata } \beta$  (i.e.,  $h = \text{cata } \beta$ ). In Diagram 1, replace  $\alpha$  with  $in$  and  $h$  with  $\text{cata } \beta$  for a commutative diagram indicating a catamorphism.

Hence the importance of initial algebras to the systematicity of recursive capacities: every algebra (process) factors through an initial algebra in a category  $\mathbf{Alg}(F)$  that has one.

Duals:  $F$ -algebra, initial algebra, and catamorphism have dual constructs called *F-coalgebra*, *final coalgebra*, and *anamorphism* (respectively), which we shall also use.

### Systematicity: List-related capacities

Returning to the example raised as a problem for classical theory: a common task is to select the smallest or largest item in a collection. Systematicity, in this case, means that if one has the capacity to distinguish the relative sizes of items, and one has the capacity to identify the smallest item in a list, then one also has the capacity to identify the largest item in a list. (Notation:  $1_A$  is the identity morphism for  $A$ ;  $1_v$  is a constant function returning  $v$ .) List-related capacities are constructed from the functor  $F_A : S \mapsto A \times S, f \mapsto 1_A + 1_A \times f$ . ( $\phi : x \mapsto y$  means  $\phi(x) = y$ .) The algebras are the pairs  $(S, [1_v, f])$ . An initial algebra for lists is the pair  $(L, [\text{empty}, \text{cons}])$ , where  $L$  is a set of lists, and  $[\text{empty}, \text{cons}] : 1 + A \times L \rightarrow L$  is the list-constructing morphism, consisting of the constant function  $\text{empty} : 1 \rightarrow L$  for constructing the empty list  $[]$ , and the function  $\text{cons} : A \times L \rightarrow L; (a, l) \mapsto a \cdot l$  for constructing the list with element  $a \in A$  prepended  $(\cdot)$  to list  $l \in L$ . Here  $1 + A \times L$  is the disjoint union of a 1-element set with the cartesian product of  $A$  and  $L$ . If, for example,  $A$  is the natural numbers  $\mathbb{N}$ , then  $L$  is the set of all finite natural number lists. Catamorphisms from this initial algebra have the form  $\text{foldL}[1_v, f] : L \rightarrow S$ , where  $\text{foldL}[1_v, f] :$

$$\begin{aligned} [] &\mapsto v \\ (a, l) &\mapsto \text{foldL}[1_v, f](l) \end{aligned}$$

The catamorphism for identifying the smallest number is  $\text{foldL}[1_\infty, \text{min}]$ , where  $\text{min} : (x, y) \mapsto x$ , if  $x \leq y$ , else  $y$  returns the smaller of two items, indicated in commutative diagram

$$\begin{array}{ccc} 1 + \mathbb{N} \times L & \xrightarrow{[\text{empty}, \text{cons}]} & L \\ \downarrow 1_1 + 1_{\mathbb{N}} \times \text{foldL}[1_\infty, \text{min}] & & \downarrow \text{foldL}[1_\infty, \text{min}] \\ 1 + \mathbb{N} \times \mathbb{N} & \xrightarrow{[1_\infty, \text{min}]} & \mathbb{N} \end{array} \quad (2)$$

E.g.,  $\text{foldL}[1_\infty, \text{min}](2, 1, 3) = \text{min}(2, \text{min}(1, \text{min}(3, \infty))) = 1$ . By replacing  $\text{min}$  in Diagram 2 with  $\text{max} : (x, y) \mapsto x$ , if  $x \geq y$ , else  $y$ , and  $\infty$  with 0 (or,  $-\infty$  for lists of integers or reals), we have the catamorphism that corresponds to identifying the largest number. For example,  $\text{foldL}[1_0, \text{max}](2, 1, 3) =$

$\text{max}(2, \text{max}(1, \text{max}(3, 0))) = 3$ . Since the two computations have the morphism  $[\text{empty}, \text{cons}]$  as the common component, this arrangement accounts for systematicity with respect to these capacities. Since the catamorphisms are uniquely determined, we have an account of systematicity without further (*ad hoc*) assumptions.

### Systematicity: language-related capacities

In this domain, we use an artificial grammar (for arithmetic expressions) to illustrate our explanation for systematicity with regard to language-related capacities. Artificial grammars are often used, because their forms are more easily adapted to the question at hand. Up to this point, we have addressed systematicity with respect to inference, e.g., why the capacity to infer the smallest list item is systematically related to the capacity to infer the largest list item—*systematicity of inference*. This aspect of systematicity assumes that the cognitive system also has the capacity to systematically represent the entities from which such inferences are made—*systematicity of representation*. Here, we also provide a category theory explanation for systematicity of representation, using the closely related, dual notion of an  $F$ -coalgebra.

### Arithmetic expressions: systematicity of inference

The example in this section is based on Hutton (1998), but adapted to model cognitive capacity for evaluating numerical expressions. The category of  $F$ -algebras that includes language-related capacities is constructed from the functor  $F_A : S \mapsto A + S \times S, f \mapsto 1_A + f \times f$ . The  $F$ -algebras for the category  $\mathbf{Alg}(F_A)$  are the pairs  $(S, [f, g])$ , where  $[f, g] : A + S \times S \rightarrow S$ ,  $f : A \rightarrow S$  is a unary function, and  $g : S \times S \rightarrow S$  is a binary function. An initial algebra in this category is  $(T, [\text{leaf}, \text{branch}])$ , where  $T$  is the set of trees of type  $A$ ,  $[\text{leaf}, \text{branch}] : A + T \times T \rightarrow T$ ,  $\text{leaf} : A \rightarrow T; a \mapsto \langle a \rangle$  returns a tree consisting of a single leaf  $a \in A$ , and  $\text{branch} : T \times T \rightarrow T; (l, r) \mapsto \langle l, r \rangle$  returns a tree consisting of a left branch  $l$  and a right branch  $r$ , where  $l, r \in T$ . For example, a binary tree of numbers  $\langle \langle 1 \rangle, \langle \langle 2 \rangle, \langle 3 \rangle \rangle \rangle$  has a leaf 1 as its left branch, and a tree, with left leaf 2 and a right leaf 3, as its right branch. A catamorphism from initial algebra  $(T, [\text{leaf}, \text{branch}])$  to an arbitrary  $F$ -algebra  $(S, [f, g])$  in  $\mathbf{Alg}(F_A)$  is the recursive function  $\text{foldT}$  (i.e., fold for trees), defined as follows. The (higher-order) function  $\text{foldT}$  takes a unary function  $f : A \rightarrow S$  and a binary function  $g : S \times S \rightarrow S$  and returns the recursive function  $\text{foldT}[f, g] : T \rightarrow S$ , where

$$\begin{aligned} \langle a \rangle &\mapsto f(a) \\ \langle l, r \rangle &\mapsto g(\text{foldT}[f, g](l), \text{foldT}[f, g](r)) \end{aligned}$$

and  $T$  is a set of trees of type  $A$ , indicated in diagram

$$\begin{array}{ccc} A + T \times T & \xrightarrow{[\text{leaf}, \text{branch}]} & T \\ \downarrow 1_A + \text{foldT}[f, g] \times \text{foldT}[f, g] & & \downarrow \text{foldT}[f, g] \\ A + S \times S & \xrightarrow{[f, g]} & S \end{array} \quad (3)$$



Suppose participants are given arithmetic expressions involving a particular operator, say, *addition*, e.g.,  $(1 + 2) + (2 + 3)$ , which they are required to evaluate. Given that participants can correctly evaluate such expressions, there is a host of other capacities that are also afforded provided that they have some other basic knowledge. For example, given knowledge of another binary operator, say, *subtraction*, participants can also evaluate the related expression  $(4 - 2) - (2 - 1)$  as 1. The specific catamorphism for the addition case is given by replacing  $A$  in Diagram 3 with the set of numbers  $N$ ,  $f$  with identity morphism  $1_N$ , and  $g$  with addition  $+$ . For the case of *subtraction*, the binary operator  $(+)$  for addition is replaced with  $(-)$  in Diagram 3. Hence, the second task is computed as  $\text{fold}T[1_N, (-)]$ . The universal construction common to these two capacities is the morphism  $[leaf, branch]$ . So, the explanation for systematicity is essentially the same as the explanations we provided for list- and number-related capacities, albeit based on a different underlying functor—the capacities for evaluating expressions involving addition and subtraction contain  $[leaf, branch]$  as the common factor.

### Arithmetic expr.: systematicity of representation

Evaluating trees is an example of *systematicity of inference* (Fodor & Pylyshyn, 1988; Aizawa, 2003; Phillips & Wilson, 2011). However, such expressions are not given to the cognitive system in tree-form. Typically, such trees are assumed to be constructed from an input (list of characters) by another process. The input may take on several different formats: e.g., numeric/symbolic, as in “1+(2+3)”, or word form, as in *one plus (two plus three)*, which correspond to the same tree. Again, these two forms are systematically related: one has the capacity to represent the expression “1+(2+3)” if and only if one has the capacity to represent the expression *one plus (two plus three)* assuming, of course, a person knows that *one*, *two* and *three* denote the same things as 1, 2 and 3 (respectively), and *plus* denotes the same thing as  $+$ . This form of systematicity is called *systematicity of representation* (Fodor & Pylyshyn, 1988; Aizawa, 2003; Phillips & Wilson, 2011). In this section, we show how systematicity of representation is addressed using coalgebras.

Constructing trees from lists is achieved by a dual construction called an *F-coalgebra* (Hutton, 1998). The explanation for systematicity in this case proceeds in a “dual” manner: i.e., every morphism in a category of *F-coalgebras* with a terminal (dual to initial) object, called a *final coalgebra* (dual to initial algebra) is composed of a unique *anamorphism* (dual to catamorphism) and a common final coalgebra.

Final coalgebras derive from their dual definitions of initial algebras in the category of *F-algebras*  $\mathbf{Alg}(F_A)$  on the functor  $F_A : \mathbf{Set} \rightarrow \mathbf{Set}; S \mapsto A + S \times S, f \mapsto 1_A + f \times f$ . A final coalgebra in this category is  $(T, (p_{\langle \rangle} \rightarrow \text{fmleaf}, \text{fmbranch}))$ , where conditional  $p_{\langle \rangle} \rightarrow \text{fmleaf}, \text{fmbranch}$  consists of a condition  $p_{\langle \rangle} : T \rightarrow \mathbf{Bool}$  that tests whether  $t \in T$  is a leaf (i.e.,  $t = \langle a \rangle, a \in A$ ), or a branch (i.e.,  $t = \langle l, r \rangle, l, r \in T$ ), and associates functions  $\text{fmleaf} : T \rightarrow A, \langle a \rangle \mapsto a$ , for retrieving a value from a leaf, and  $\text{fmbranch} : T \rightarrow T \times T, \langle l, r \rangle \mapsto (l, r)$ ,

for retrieving a pair of left and right subtrees from a branch. The dual category  $\mathbf{CoAlg}(F_A)$  has *F-coalgebras*  $(S, (p \rightarrow f, g))$  as objects, and *F-coalgebra homomorphisms* as morphisms. The anamorphism associated with this final coalgebra is called *unfoldT* (i.e., unfold for trees), defined recursively as  $\text{unfold}T(p \rightarrow f, g) : S \rightarrow T$

$$\begin{aligned} s &\mapsto \langle f(s) \rangle && \text{if } p(s) \\ s &\mapsto \langle \text{unfold}T p? (p_1 \circ g(s)), \text{unfold}T p? (p_2 \circ g(s)) \rangle && \neg p(s) \end{aligned}$$

where  $p?$  abbreviates  $(p \rightarrow f, g)$ . The final coalgebra and anamorphism are indicated in commutative diagram

$$\begin{array}{ccc} S & \xrightarrow{p \rightarrow f, g} & A + S \times S \\ \text{unfold}T p? \downarrow & & \downarrow 1_A + \text{unfold}T p? \times \text{unfold}T p? \\ T & \xrightarrow{p_{\langle \rangle} \rightarrow \text{fmleaf}, \text{fmbranch}} & A + T \times T \end{array} \quad (4)$$

Diagram 4 indicates the general form of the anamorphism from which we specify a particular  $p?$  (i.e.,  $p \rightarrow f, g$ ) for our domain of arithmetic expressions. That is, we need to define the test function  $p : L \rightarrow \mathbf{Bool}$ , where  $\mathbf{Bool} = \{\text{True}, \text{False}\}$  that determines whether an expression indicates a simple (value) or complex expression, and associated functions  $f : L \rightarrow N$  and  $g : L \rightarrow L \times L$  for transforming simple and complex expressions into numbers and expression pairs (respectively).

Specifications of  $f$  and  $g$  (in Diagram 4) are obtained from case analysis. Examples of simple expressions, which indicate values, are: “1”, “(2)”, and “((3))”, i.e., any well-formed expression that does not contain the “+” character. A complex expression is any well-formed expression that is not simple. So,  $p$  is the function  $\text{isVal} : l \mapsto “+” \notin l$ . Since  $f$  is associated with  $p(l)$  being true, we require a function to convert a string into a (internal) representation for the corresponding number, i.e.,  $f$  is the function  $\text{str2num} : L \rightarrow N$ . Finally, we need a function  $g$  for complex expressions. Examples of complex expressions include: “1+2”, “1+(2+3)”, “(1+2)+3”, “(1+2)+(3+4)”, and so on. The purpose of  $g$  is to split an expression into two subexpressions, one corresponding to the left branch of the tree, and the other to the right branch. That is,  $g$  must split the expression at the topmost operator into two subexpressions containing the strings before and after the “+” symbol, after stripping off the outer brackets. Identifying the split point is also determined by case analysis: Basically, the split point is the first instance of “+” in the absence of an unmatched right bracket “)”. So, one simply maintains a counter, starting from 0 (i.e., no unmatched brackets, or top level), which is incremented/decremented on every occurrence of a left/right bracket, when read from left to right. So,  $g$  is the function  $\text{split} : L \rightarrow L \times L$ . Thus, the function for parsing expressions into trees is the anamorphism  $\text{unfold}T(\text{isVal} \rightarrow \text{str2num}, \text{split})$ .

Systematicity of representation (in this example, constructing trees) is obtained in the same way as systematicity of inference (“destructing” trees). To represent the same tree from

the expressions in word form, one simply replaces the argument  $isVal \rightarrow str2num, split$  as appropriate. For example, the function  $str2num$  is replaced with, say,  $word2num$  which converts numbers in word form (e.g., “one”, “two”, etc.) to their corresponding internal representation of number. In any case, the resulting anamorphism factors through the same universal morphism, i.e.,  $p_{\langle \rangle} \rightarrow fmleaf, fmbranch$  from Diagram 4.

Given initial algebra  $in : F(A) \rightarrow A$  in a category  $\mathbf{Alg}(F)$ , the corresponding final coalgebra  $fin : A \rightarrow F(A)$  is guaranteed to exist, because  $F(A) \cong A$ , so  $in$  has as inverse  $fin$ . Thus, further (*ad hoc*) assumptions are not required to guarantee a correspondence between expressions and evaluations since they are indivisibly bound by the (final) initial (co)algebra. By contrast, classical theory assumes that the processes for constructing syntactically compositional representations and the processes for systematically transforming those representations correspond (Phillips & Wilson, 2011).

## Discussion

Our explanation in regard to recursive domains employs the same category theory construct (i.e. universal construction) as our previous explanations for (quasi-)systematicity in regard to non-recursive domains (Phillips & Wilson, 2010, 2011), albeit with different kinds of functors: here, for recursive domains, the universal constructions involved endofunctors (i.e., where the domain and codomain are the same category), whereas for non-recursive domains, the universal constructions involved adjoint functors (which are reciprocating, though not necessarily inverse, functorial maps between categories that are not necessarily the same. Every composition of left and right adjoints is an endofunctor, but not every endofunctor can be decomposed into a pair of adjoint functors. So, having some (primitive) form of systematicity over a recursive domain does not imply having systematicity for non-recursive domains. Nor, for that matter, does having the systematicity property for one recursive domain (e.g., numbers) imply the having the systematicity property for another recursive domain (e.g., lists), when the universal constructions involve functors not related by a natural isomorphism (Manes & Arbib, 1986)—this distinction also applies to non-recursive domains. This functorial distinction has implications for comparative and developmental psychology (discussed later).

## Limitations

Our theory may be incomplete at two points: one point is where competence meets performance, such as when supposed systematically related capacities span memory or cognitive complexity limits. The other point is where systematic cognition meets non-systematic cognition: not all cognition is regarded as systematic; idioms (e.g., *John kicked the bucket*—i.e., he died—is not systematic with *Mary kicked the bucket [with her foot]*) are a paradigm (Fodor & Pylyshyn, 1988). We discuss our theory in the context of both cases.

Competence versus performance: In the case of lists where the morphism  $f$  is not associative (e.g., subtraction), comput-

ing with a right-fold version of list fold means keeping all list items in memory (if presented once only), so systematicity would not extend beyond lists of more than a few items. Such cases are generally not regarded as evidence against the systematicity property—human cognition is *ceteris paribus* (e.g., memory requirements being the same) largely systematic (see McLaughlin, 2009). Nonetheless, a more complete theory will address both aspects of cognition. Category theory may also provide independent principles for performance, since cognitive development-related limits in children were identified with the arity of the (co)product underlying the task (Phillips, Wilson, & Halford, 2009): e.g., the ability of children older than the median age of five years to perform transitive inference and class inclusion in the more difficult condition versus children younger than five was related to (co)product arity, i.e., binary versus unary (co)products. Note that here, too, the difference in “complexity” of the endofunctors for number (no/unary product of functors, not given), list (binary product of constant and identity functors) and tree (binary product of two identity functors). However, performance related differences are beyond the scope of our theory as it currently stands.

Systematic versus non-systematic cognition: Category theory also provides a principled means for joining two cognitive (sub)systems via (co)products of categories, where one category models systematic cognitive capacity and the other non-systematic capacity, and (say) the coproduct category models both. However, as Aizawa (2003) explains, the required explanatory standard for hybrid theories is higher, because one must also explain why/when component theories are invoked. A possible reason is efficiency. A primitive form of addition is supported (systematically) by the category of  $F$ -algebras that included number-related capacities via  $foldN$  (not presented here), where the number of iterations is proportional to the size of the addends. The time needed to add numbers can be reduced by memorizing the addition table for small numbers, which is what children are taught to do. However, addition via memorized associations is not a systematic process: one can memorize part of a table without memorizing the other part; this is an analog of the phrase-book example in language (Fodor & Pylyshyn, 1988). Utility may drive the cognitive system to employ a faster, but non-systematic process, but it is also outside the scope of our current theory.

## Perspective

At the core of our category theory explanation for systematic recursive capacity is a special pair of dual constructions: an (final) initial (co)algebra in a category of (co)algebras on a polynomial functor  $F$ . Although one can reverse the direction of any collection of arrows to form a dual, such duals may not exist in the category of interest (e.g., some categories have initial objects but no final objects). Yet, for categories of (co)algebras on a polynomial functor (final) initial (co)algebras are guaranteed to exist (Manes & Arbib, 1986), and an initial algebra  $in : F(A) \rightarrow A$  is guaranteed to have an inverse  $fin : A \rightarrow F(A)$ , because the component objects are

isomorphic (i.e.,  $A \cong F(A)$ ), which constitutes a final coalgebra for the domains we have investigated. So, the systematic relationship between representation and inference is guaranteed without further (*ad hoc*) assumptions, in contrast to the classical explanation where the link between the two is just assumed (Phillips & Wilson, 2011). Notice that this dual relationship between systematicity of representation and systematicity of inference is more general (and more useful) than an inverse. In the arithmetic expressions example, lists were represented as trees (systematicity of representation), but trees were evaluated as numbers (systematicity of inference). This form of duality goes beyond the simple inverse relationship between sentence recognition and generation found in parsing/production rules in a classical approach to language.

The capacity for recursion has been a contentious issue in the broader interests of cognitive science, which includes comparative and developmental psychology. Some argue that recursion is specific to humans and depends on language (Hauser, Chomsky, & Fitch, 2002); more particularly, a fully inductive (recursive) basis for number is specific to adults and distinct from infants' non-inductive basis (Rips, Bloomfield, & Asmuth, 2008). In contrast, others claim a human language-like capacity for recursion in songbirds (Gentner, Fenn, Margoliash, & Nusbaum, 2006) (but, see Corballis, 2007), and that adult understanding of number (in its fully induced form) is founded on a more primitive infant conception (Carey, 2009). See also Gelman and Butterworth (2005), for a review of the debate over the link between number and language. Our category theory treatment of recursive cognitive capacities provides a different perspective on this issue: specifically, the particular systematic capacities for recursion depend on the underlying functor, not a general capacity for recursion, as such. In particular, one can have a basic recursive capacity for number without having a full-blown capacity for language, because the functor underlying recursive number-related capacities does not provide a systematic basis for recursive language-related capacities, though by our account language-related recursive capacities afford number-related recursive capacities. Analysis of the songbird evidence (Gentner et al., 2006) for supposed center-embedded recursion suggested that these birds were using a simple *counting* strategy (Corballis, 2007), which accords with our  $F$ -(co)algebraic basis for recursion in cognition, where simple counting involves a fold for numbers, not trees. Thus, other species (and infants) can have elementary recursive capacities without a full-blown capacity for number and language as available in adult humans.

The classicist's approach to cognitive architecture is fundamentally limited not in advocating syntax, but in placing syntax at the foundation of their theory. Given the often *ad hoc* and idiosyncratic choices that go into programming language design, computer scientists in recent decades have turned to category theory for a deeper syntax-free understanding of the principles of computation. Cognitive science, as couched within the framework of computationalism, can likewise do

better than lay foundations on the shifting sands of syntax.

## Acknowledgments

This work was supported by a Japanese Society for the Promotion of Science Grant-in-aid (22300092).

## References

- Aizawa, K. (2003). *The systematicity arguments*. New York: Kluwer Academic.
- Carey, S. (2009). *The origins of concepts*. New York, NY: Oxford University Press.
- Corballis, M. C. (2007). Recursion, language, and starlings. *Cognitive Science*, 31, 697–704.
- Fodor, J. A. (1975). *The language of thought*. New York, NY: Crowell.
- Fodor, J. A., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, 35, 183–204.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3–71.
- Gelman, R., & Butterworth, B. (2005). Number and language: how are they related? *Trends in Cognitive Sciences*, 9(1), 6–10.
- Gentner, T. Q., Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature*, 440(7088), 1204–1207.
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, 298(5598), 1569–1579.
- Hutton, G. (1998). Fold and unfold for program semantics. In *Proceedings of the 3rd ACM SIGPLAN International Conference on Functional Programming*.
- Manes, E. G., & Arbib, M. A. (1986). *Algebraic approaches to program semantics*. New York, NY: Springer-Verlag.
- McLaughlin, B. P. (2009). Systematicity redux. *Synthese*, 170, 251–274.
- Phillips, S., & Wilson, W. H. (2010). Categorical compositionality: A category theory explanation for the systematicity of human cognition. *PLoS Computational Biology*, 6(7), e1000858.
- Phillips, S., & Wilson, W. H. (2011). Categorical compositionality II: Universal constructions and a general theory of (quasi-)systematicity in human cognition. *PLoS Computational Biology*, 7(8), e1002102.
- Phillips, S., & Wilson, W. H. (2012). Categorical compositionality III: F-(co)algebras and the systematicity of recursive capacities in human cognition. *PLoS ONE*, 7(4), e35028.
- Phillips, S., Wilson, W. H., & Halford, G. S. (2009). What do Transitive Inference and Class Inclusion have in common? Categorical (co)products and cognitive development. *PLoS Computational Biology*, 5(12), e1000599.
- Rips, L. J., Bloomfield, A., & Asmuth, J. (2008). From numerical concepts to concepts of number. *Behavioral and Brain Sciences*, 31(6), 623–687.

# Modeling Concept Activation in Working Memory during Online Sentence Processing

Patrick Plummer (pplummer@ucsd.edu)  
Hsueh-Cheng Wang (hchengwang@gmail.com)  
Yuhtsuen Tzeng (ttcytt@ccu.edu.tw)  
Marc Pomplun (marc@cs.umb.edu)  
Keith Rayner (krayner@ucsd.edu)

Department of Psychology, University of California at San Diego, USA

Department of Computer Science, University of Massachusetts at Boston,  
100 Morrissey Boulevard, Boston, MA 02125 USA

Center for Teacher Education & Institute of Curriculum, National Chung Cheng University, Taiwan

## Abstract

There have been several computational alternatives to the cloze task (Taylor, 1953) intended to approximate word predictability effects on eye movements during reading. In this study, we implement a computational model that instantiates each content word in a sentence as an input that activates semantic concepts in working memory. The predictability of a word is then determined by the extent to which its corresponding semantic representation is associated with the network of concepts already active in working memory from the preceding context. The computation of concept activation is based on a connectionist model (Landscape model, see van den Broek, 2010). Latent semantic analysis (LSA) is used to establish connections between words and simulate the long-term semantic associations among concepts (Landauer & Dumais, 1997). This model provides a means of investigating how language comprehension and eye movement behavior are affected by the activation of concepts in working memory.

**Keywords:** eye movements; reading; word predictability; latent semantic analysis; Landscape model.

## Introduction

It has been well-established that eye movement behavior is affected by lexical variables such as frequency and predictability (Rayner, 1998; 2009). As such, the eye movement record provides an indication of language processing as it unfolds during normal reading. Rayner and Well (1996; see also Ehrlich & Rayner, 1981) found that the predictability of target words had a strong influence on eye movements during reading. In their experiment, subjects fixated unpredictable target words longer than either highly or moderately predictable target words; highly predictable words were also skipped more often than moderately predictable or unpredictable target words.

Accordingly, in the E-Z Reader model (Pollatsek, Reichle, & Rayner, 2006; Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Rayner, & Pollatsek, 1999; 2003), word predictability within a given sentence context is considered in both first stage processing (i.e., L1, including identification of orthographic form and a familiarity check) and second stage processing (i.e., L2, including

identification of phonological/semantic form and completion of lexical access). The model also maintains that the predictability effect is stronger in L2 than in L1.

Estimates of word predictability are typically derived from a modified cloze task procedure (Taylor, 1953) in which subjects are asked to guess the identity of a word when given the prior sentence context. Most reading studies utilize the cloze task to establish or confirm word predictability manipulations. These experiments use target words that differ substantially in cloze value (the probability with which subjects select the word), often with probabilities of .70 to .90 for highly predictable words and less than .10 for “low” predictability words. As an alternative to necessarily subjective cloze responses, several computational methods have been successfully utilized to approximate degrees of contextual constraint and predict the influence on eye movements during reading; including, transitional probabilities (McDonald and Shillcock, 2003; but see Frisson, Rayner, & Pickering, 2005), surprisal (Boston, Hale, Kliegl, Patil, and Vasissth, 2008; Levy, 2008), conditional co-occurrence probability (Ong and Kliegl, 2008). Additionally, Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) was used by Pynte, New, and Kennedy (2008) as well as Wang, Chen, Ko, Pomplun, and Rayner (2010), who reported that eye movement behavior during first-pass reading on content words could be predicted using LSA. McDonald and Shillcock (2003) and Wang et al. (2010) used the transitional probability (corpus-based statistical likelihood of encountering a word given the preceding or subsequent word) to categorize predictability conditions; both proposing that predictability effects could be accounted for using only the content word preceding a target. One limitation of these objective measures could be that prior context, before the immediately preceding lexical item, may affect processing of a word in many instances. Wang et al. (2010) also used *all* concepts in the preceding sentence context to compute contextual constraint for targets using the standard weighting from LSA. However, without a clearer understanding of working memory constraints

during comprehension it is difficult to make regarding semantic constraint.

The predictability of a given word can, in large part, be conceptualized as the degree to which the semantic concept represented by the word is associated with the context of preceding lexical items. By treating incoming lexical items as semantic concepts that interactively influence working memory processes, prior context for a word can be represented as inputs which influence the activation of associated concepts and have the potential to facilitate or inhibit the processing of upcoming words. As a result, the higher the activation of a concept when it is encountered, the more processing of the concept is facilitated. Importantly, individuals can allocate their processing attention to only a finite number of linguistic items at a given moment. Thus, any model of language processing and working memory must establish limits to the number of lexical-semantic concepts that can be simultaneously active and exert an appreciable influence on the processing of upcoming lexical inputs.

## A Connectionist Model for Sentence Reading

This study proposes a computational model to monitor the activations of concepts in working memory. The computation of concept activation is derived from a connectionist model (the Landscape model, see van den Broek, 2010). The current model is not connectionist in the sense of having distributed semantic representations; rather, words are represented as localized semantic "concepts" with weighted connections to a network of additional concepts. The semantic connections among concepts in the simulation are computed using LSA cosine values based on the default 300 dimension semantic space, "general reading up to 1st year college", available at the LSA@CU Boulder website (<http://lsa.colorado.edu/>). LSA represents word meaning and computes associations by applying a linear algebra method, singular value decomposition (SVD), to a large corpus of text (see Landauer & Dumais, 1997).

The Landscape model is a connectionist approach to instantiating comprehension using psychologically plausible algorithms that can potentially be used to model several aspects of text comprehension (see van den Broek, 2010; Tzeng, van den Broek, Kendeou, & Lee, 2005). The architecture of the conventional Landscape model assumes that as a reader proceeds through a text in reading cycles (with each cycle roughly corresponding to the reading of a new sentence), concepts fluctuate in activation as a function of four sources of information: the current processing cycle, the preceding cycle, the current episodic text representation, and reader's background knowledge. With the reading of each cycle, particular concepts are activated and added as nodes to the episodic memory representation of the text. If a concept is already part of the text representation and is reactivated, its trace is strengthened. Furthermore, co-

activation of concepts leads to the establishment (or strengthening) of connections between those concepts. The resulting network representation influences subsequent activation patterns. This phenomenon is called the *cohort effect*. These cyclical and dynamically fluctuating activations lead to the gradual emergence of an episodic memory representation and discourse model of the text, in which textual propositions and inferences are connected via semantic relations (such as causal and referential links). Thus, the model captures the fluctuations of concepts during reading (Linderholm, Virtue, Tzeng, van den Broek, 2004), as well as readers' memory representation of text (Tzeng, 2007). As such, this model has prescribed mechanisms that can link the iterative and reciprocal relations between fluctuations of activations and the episodic text representation. However, there are necessary differences with regard to how readers generate and update active discourse representations for the comprehension of an individual sentence, compared to the processing of a longer narrative or expository text. For the comprehension of an individual sentence, a reader must primarily rely on establishing connections between relevant concepts in working memory and pre-existing long-term semantic representations. For a longer text, on the other hand, readers are often able to take advantage of more extensive and detailed context and presumably a more enriched discourse model. Thus, the current computational approach adapts the Landscape Model to a connectionist framework more suitable for capturing sentence reading. The current model utilizes LSA in order to represent pre-existing connections between semantic representations stored in long-term memory (i.e., background or world knowledge).

In the current model, as with the Landscape Model, text inputs are represented by an *input matrix* and each is indexed as a *Mention* (concepts being read from the text). The conventional Landscape model also defines other sources of activation including *Referential* (for building referential coherence), *Causal*, and *Enabling* (for the causal explanation of the current statement), but those activations are as of yet, not implemented here. The input matrix for example sentence: "*The knight uses his sword to fight the dragon*" is shown in Table 1.

Table 1: Input matrix for the *Knight* example.

cycle	knight	use	sword	fight	Dragon
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1

Initially, the sentence is segmented into component concepts: "knight", "use", "sword", "fight", and "dragon"; as, currently, only content words are considered as concepts. The model assumes that each word is fixated and processed

sequentially. In each cycle, the concept of *Mention* receives 1 unit of activation. In addition to the sequential activation of concepts, the influence of semantic knowledge and pre-existing lexical associations between concepts is established using LSA corpus-learned associations. Table 2 presents the *connection matrix* for the example sentence. The values are always between -1 and 1, but are rarely below 0 because of LSA’s high-dimensional space.

Table 2: Connection matrix for the example.

.	knight	use	sword	fight	dragon
knight	1	.01	.64	.15	.28
use	.01	1	.03	-.02	.06
sword	.64	.03	1	.20	.40
fight	.15	-.20	.20	1	.13
dragon	.28	.06	.40	.13	1

The activation values for each concept are represented in an  $m \times n$  *activation matrix*, where  $m$  represents the number of concepts in the sentence and  $n$  represents the number of *cycles*. Each column in the matrix thus represents the status of each concept. The *activation matrix* takes each column of the *input matrix* as raw input and processes it row by row. In our model, the activation during the current reading cycle is defined by Equation (1):

$$A_i^{cycle} = \sum_{j=1}^m \delta A_j^{cycle-1} \sigma(S_{ij}) + \sum_{j=i}^m input_i^{cycle} \sigma(S_{ij}) \quad (1)$$

$A_i^{cycle}$  is the activation of concept  $i$  during the current cycle. Starting from the summation ( $\Sigma$ ) term in Equation (1), for all activated concepts in the previous reading cycle, each activation value is multiplied by a transformation function  $\sigma$  of connection strength ( $S_{ij}$ ) and by the cohort activation parameter  $\delta$ .  $S_{ij}$  is the strength of the relation from concept  $j$  to  $i$ . For the current cycle,  $input_i^{cycle}$  is the activation of concept  $i$  in the *input matrix*. The sum of the net inputs for these  $m$  concepts is multiplied by the transformation function  $\sigma$  of connection strength ( $S_{ij}$ ).

The conventional Landscape model uses a sigmoid function as the transformation function  $\sigma$  to control the possible linear growth of spreading of activation and limit the effect of cohort activation to those strongly related to the concept. Since  $S_{ij}$  is usually between 0 and 1, a linear function with absolute value is used in this model. The value of the cohort activation parameter,  $\delta$ , directly determines the amount of cohort activation and in the future can be used to mimic individual differences in the spreading of activation. Our model assumes that for any concept, its cohort activation can never exceed its input activation. For this reason the model will take the largest of the input and cohort activation values, and *Mention* is the maximum activation a concept can receive. Furthermore, the system parameter- *Activation Threshold* sets any activation below a set threshold to zero.

The working memory constraint is implemented by a parameter *WMC* (*Working Memory Capacity*). When the actual sum of activation exceeds the value of WMC, the activation of each concept is scaled down using Equation (2):

$$A_i^{cycle} = A_{i,Actual}^{cycle} \cdot \frac{WMC}{\sum_{i=1}^m A_{i,Actual}^{cycle}} \quad (2)$$

For the example sentence, the activation matrix is shown in Figure 1. For the 1<sup>st</sup> cycle (in which “*knight*” is processed during first-pass reading), the activation of “*knight*” is 1, from the *Mention* input. There is no cohort effect for the first reading cycle since no previous cycle exists. The activations for “*use*”, “*sword*”, “*fight*”, and “*dragon*” are established by multiplying their connections, .01, .64, .15, and .28 respectively, and the input of “*knight*” (1). The activation of “*use*” does not reach the threshold (set to 0.1) and as a result receives an activation of 0. For the 2<sup>nd</sup> cycle when “*use*” is being processed, the activation of each concept is calculated according to Equation (1). Figure 1 illustrates that the activation of “*dragon*” stays relatively high from cycle 1 to cycle 4 because of relatively strong connections to “*knight*”, “*sword*”, and “*fight*.” Conversely, the activation of “*use*” decreases from cycle 2 to 5 because of relatively weak connections to “*sword*”, “*fight*”, and “*dragon*” (less than .06).

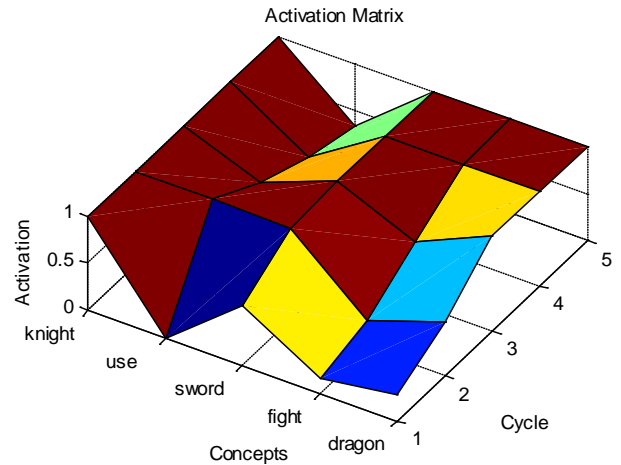


Figure 1: The “landscape” of the activation matrix for the *Knight* example.

The conventional Landscape model updates the connection strengths in its episodic memory using a learning algorithm in order to adjust active discourse representations for the comprehension of a longer text. In this study, we assume that the background knowledge (represented by the connection matrix) is not altered during sentence reading.

In summary, by assuming (a) that words in a sentence are read and processed sequentially, and (b) long-term memory

representations (i.e., background knowledge) are not affected during comprehension; we propose a computational model of sentence processing which takes advantage of an existing discourse comprehension model designed to take into account contextual effects. The proposed model allows us to examine several factors that affect sentence comprehension; namely, (1) semantic activation in working memory, (2) background knowledge, and (3) working memory capacity. To assess the model's ability to reflect linguistic processing we will compare its performance to the cloze task.

## Experiment: Reanalysis of Previous Data

The key objective for this implementation is to disambiguate high from low semantic constraint in sentence contexts. Another objective of this implementation is to demonstrate that the LS model surpasses previously utilized methods as an alternative to the cloze task. In order to demonstrate that the proposed computational model is capable of matching cloze results more accurately than previous approaches, i.e., Wang et al. (2010), we re-analyzed the materials in Gollan, Slattery, Goldenberg, Van Assche, Duyck, and Rayner (2011), in which predictable/unpredictable target words were determined by a norming cloze task. We estimated predictability of a target word by (1) the previous content word, (2) all words in prior context, and (3) the estimates of the proposed connectionist model in this endeavor. We expect that our model can outperform other predictors on differentiating high- and low-constraint contexts and generate higher correlation to cloze values.

**Participants.** Twenty undergraduate students at the University of California, San Diego, participated. All of them were native speakers of English.

**Materials.** There were 90 target words; all target words were embedded in either a high-constraint (HC) or low-constraint (LC) sentence. For example, “*the hockey player moved on the ice on his \_\_\_\_\_*” (S1) was considered HC while “*The little girl was very happy when she unwrapped her brand new \_\_\_\_\_*” (S2) was LC for the target “skates”. The target words in HC context were generated 87% of the time, whereas the ones in LC context were generated less than 3% of the time.

**Procedure.** Participants were presented with the sentences up to the target words, and asked to provide one-word continuations for each sentence.

**Analysis.** The first estimate of predictability of each target word was derived by extracting the LSA connection weight to the previous content word (PreCont) for each target, e.g., the previous content word of S1 is “ice,” while the one of S2 is “new.” The second approach computed the LSA cosine value using all words in the previous context (AllW). The final estimate was derived from Landscape

model of sentence processing described above in the previous section (LS). We manually segmented the sentence into concepts and removed function words such as “a”, “the”, “in”, etc., for instance, “hockey / player / moved / ice” for S1. The parameters of our model were set as following:  $\delta = .7$ , *Mention* = 1, *Activation Threshold* = .1, and *WMC* = 7. The averages and standard deviations of Cloze, PreCont, AllW, and LS for HC and LC are described in Table 4.

Table 4. The averages and standard deviations (in parentheses) of Cloze, PreCont, AllW, and LS for HC and LC conditions.

	Cloze	PreCont	AllW	LS
HC	.87 (.13)	.17 (.16)	.21 (.16)	.66 (.29)
LC	.03 (.03)	.05 (.11)	.04 (.07)	.13 (.20)

## Results

As shown in Figure 2, an operating characteristic (ROC) analysis demonstrates that the area under the curves (AUC) of Cloze, PreCont, AllW, and LS are 1, .70, .87, .91, respectively. The LS model obtains a higher AUC than AllW or PreCont. Furthermore, a correlation analysis demonstrates that the Pearson correlations between Cloze and PreCont, AllW, and LS are .39, .56, and .70, respectively.

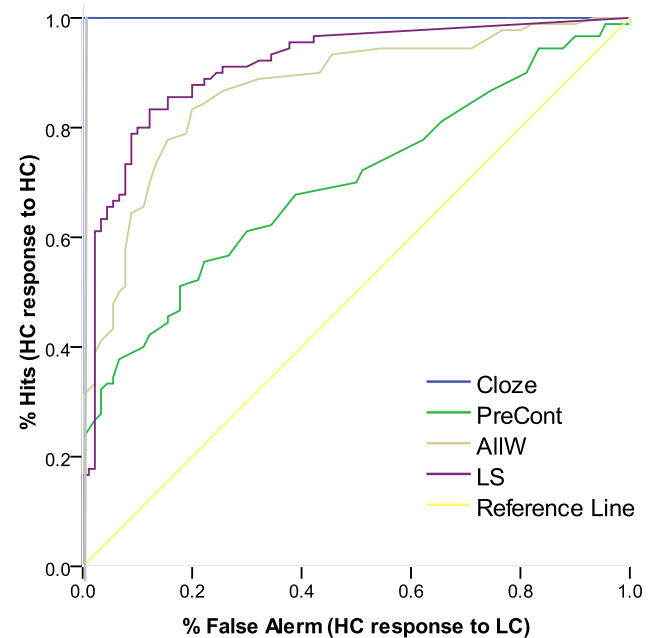


Figure 2: ROC curves for Cloze, PreCont, AllW, and LS.

The results suggest that the LS model can simulate much of the linguistic processing subjects perform when



producing cloze responses (and presumably during normal reading). The current objective is not to match cloze probabilities per se, but to successfully demonstrate the model's ability to differentiate highly constrained and unconstrained sentence contexts as well as the conventionally used cloze task. The LS model also demonstrates superiority over objective measures that utilize only the prior content word or LSA connections between content words exclusively.

## Discussion

The current implementation of the model has demonstrated that it is an effective measure of contextual constraint; in that it differentiates high and low-constraint sentence contexts better than previously employed alternatives to the cloze task. Furthermore, model activations for target words correlate with cloze responses more highly than previous objective methods of measuring contextual constraint. We believe this is an initial step toward the ultimate objective of representing both the fluctuating activation of lexical-semantic concepts in working memory during online sentence processing and how the processing of upcoming words can be facilitated by prior context. Discourse-mediated spreading activation across lexical-semantic representations has been proposed as an appreciable source of predictability effects during reading (Morris, 1994; Pynte et al., 2008; Traxler, Foss, Seely, Kaup, & Morris, 2000). Thus, modeling the process whereby linguistic inputs activate concepts in long-term memory and continuously influence working memory operations during sentence comprehension is an important endeavor in psycholinguistics.

As shown by the comparison to standard cloze responses, the current model can be used to reliably derive predictability of word  $n$  given the preceding context. The model generates a specific level of activation for word  $n$ , assuming that each word in the preceding context has been identified and all associated concepts have been engaged in working memory. As demonstrated above, this predicted level of activation correlates to cloze probabilities for a target word ( $n$ ).

Critically, the LS Model is able to reliably differentiate high and low constraint sentence fragments. Moreover, when using the LS model, in many cases the level of activation for word  $n$  will provide a more psychologically realistic measure of word processing difficulty when compared to cloze proportions, especially in neutral or unconstrained contexts. For instance, referencing cloze scores alone, there is no distinction between words that are plausible, yet not highly-predictable, and those that are completely implausible or anomalous given the preceding sentence context. In fact, it is quite feasible for plausible target words in unconstrained sentence frames to receive cloze probabilities at or around zero; however, low cloze probabilities are not necessarily indicative of potential processing difficulty. The manner in which the cloze task is conventionally used produces binary measurements (to the

extent that non-target responses are ignored). In this way, the current computational model may produce a more accurate representation than cloze scores with regard to indexing online word processing difficulty. This is particularly true for low constraint sentence frames. As such, the next logical step is to assess the LS model's goodness-of-fit to reading times and other eye movement data.

By modifying the framework of the conventional Landscape Model to reduce the size of text segments being processed during a reading cycle and situating activated concepts within limited working memory resources, we have attempted a psychologically plausible computational model of semantic effects on sentence comprehension. Crucially, the fluctuating activation of within sentence concepts is not determined merely by summing its cumulative activation across all preceding words; rather, the interactive and co-dependent influence of the prior sequence of words determined the extent to which the prior sentence context results in activation for a particular lexical-semantic concept.

The model is also a useful tool for investigating the number of semantic entities that are generally active in working memory. As well as the upper limits for the number of lexical-semantic entities simultaneously activated. Computationally examination of working memory limitations during reading could provide insight into what linguistic constructions are likely to elicit processing difficulty, result in longer fixation times, and lead to inter-word regressions during sentence reading. Model outputs can also be used to make predictions as to which concepts are likely to maintain relatively high levels of activation in working memory.

While among the most sophisticated computational frameworks in the field of cognitive science, current models of eye movement control during reading do not focus on how prior words render specific words predictable. The more well-developed models of oculomotor behavior and language comprehension represent the predictability of a given word in a sentence using only its cloze probability (Engbert, Nuthmann, Richter & Kliegl, 2005; Reichle et al., 1999; 2003). Our model successfully attempts to represent the cognitive processes that are sensitive to semantic constraint. Future implementations of the LS model will be capable of more thoroughly examining aspects of language processing and eye movement behavior. The *connection matrix* in the LS model can operationalize a variety of linguistic characteristics stored represented in long-term memory. Semantically-based connection weights can be modified to accommodate mediation by parafoveal preview information. In addition, the *connection matrix* could be modified to capture morphological, orthographic, and phonological relationships between lexical items. Currently, the LS model is a computational alternative to the cloze that is sensitive to both strong and subtle changes in contextual semantic constraint; as shown by the reasonable activation of plausible words in low constraint sentence

frames. Ultimately, the model will be expanded in an effort to achieve more comprehensive measurement of lexical-semantic predictability as it affects reading behavior.

## Acknowledgments

Preparation of the article was conducted when the first author held a predoctoral fellowship on Grant T32DC00041 from the Center for Research in Language. The work was also supported by Grant R01EY021802 from the National Eye Institute, USA, to Marc Pomplun and by Grant 97-2410-H-194-090-MY2 from the National Science Council, Taiwan, to Yuhtsuen Tzeng.

## References

- Boston, M. F., Hale, J., Kliegl, R., Patil, U. & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1):1, 1-12.
- Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Sciences*, 41, 391-407.
- Ehrlich, S.F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20, 641-655.
- Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, 112, 777-813.
- Gollan T. H., Slattery T. J., Goldenberg D., Van Assche E., Duyck W., Rayner K. (2011). Frequency Drives Lexical Access in Reading but Not in Speaking: The Frequency-Lag Hypothesis. *Journal of Experimental Psychology: General*, 140, 2, 186-209
- Jones, M. N. & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114, 1-37.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
- Landauer, T. K., McNamara, D. S., Dennis S., & Kintsch W. (2007). *Handbook of Latent Semantic Analysis*, Lawrence Erlbaum Associates.
- Levy, R. (2008). Expectation based syntactic comprehension. *Cognition* 106, 1126- 1177.
- Linderholm, T., Virtue, S., Tzeng, Y., & van den Broek, P. W. (2004). Fluctuations in the Availability of Information during Reading: Capturing Cognitive Processes using the Landscape Model. *Discourse Processes*, 37(2), 165-186.
- McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, 14, 648-652.
- Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. *Journal of experimental psychology: Learning, Memory, & Cognition*, 20, 92-103.
- Ong, J. K. Y. & Kliegl, R. (2008). Conditional co-occurrence probability acts like frequency in predicting fixation durations. *Journal of Eye Movement Research*, 2(1):3, 1-7
- Pynte, J., New, B. & Kennedy, A. (2008). A multiple regression analysis of syntactic and semantic influences in reading normal text. *Journal of Eye Movement Research*, 2(1):4, 1-11
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372-422.
- Rayner, K. The Thirty Fifth Sir Frederick Bartlett Lecture: Eye movements and attention during reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62, 1457-1506.
- Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, 3, 504-509.
- Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, 105, 125-157.
- Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, 26, 445-476.
- Reichle, E. D., Rayner, K., & Pollatsek, A. (1999). Eye movement control in reading: Accounting for initial fixation locations and refixations within the E-Z Reader model. *Vision Research*, 39, 4403-4411.
- Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30, 415-433.
- Traxler, M. J., Foss, D. J., Seely, R. E., Kaup, B., & Morris, R. K. (2000). Priming in sentence processing: Intralexical spreading activation, schemas, and situation models. *Journal of Psycholinguistic Research*, 29, 581-594.
- Tzeng, Y. (2007). Memory of narrative texts: How parts of Landscape model work. *Chinese Journal of Psychology*, 49 (3), 1-25.
- Tzeng, Y., van den Broek, P., Kendeou, P., & Lee, C. (2005). The computational implementation of the landscape model: Modeling inferential processes and memory representations of text comprehension. *Behavior Research Methods*, 37(2), 277-286.
- van den Broek, P. (2010). Using texts in science education: cognitive processes and knowledge representation. *Science*, 328, 453.
- Wang, H. C., Pomplun, M., Ko, H. W., Chen M. L., & Rayner, K. (2010). Estimating the effect of word predictability on eye movements in Chinese reading using latent semantic analysis and transitional probability, *Quarterly Journal of Experimental Psychology*, 63, 1374-1386.

# Exploring the Role of Representation in Models of Grammatical Category Acquisition

Ting Qian (tqian@bcs.rochester.edu)<sup>1</sup>

Patricia A. Reeder (preeder@bcs.rochester.edu)<sup>1</sup>

Richard N. Aslin (aslin@cvs.rochester.edu)<sup>1</sup>

Josh B. Tenenbaum (jbt@mit.edu)<sup>2</sup>

Elissa L. Newport (newport@bcs.rochester.edu)<sup>1</sup>

<sup>1</sup>Department of Brain & Cognitive Sciences, University of Rochester

<sup>2</sup>Department of Brain & Cognitive Sciences, MIT

## Abstract

One major aspect of successful language acquisition is the ability to generalize from properties of experienced items to novel items. We present a computational study of artificial language learning, where the generalization patterns of three generative models are compared to those of human learners across 10 experiments. Results suggest that an explicit representation of word categories is the best model for capturing the generalization patterns of human learners across a wide range of learning environments. We discuss the representational assumptions implied by these models.

## Introduction

Learning the grammar of a language consists of at least two important tasks. First, learners must discover the cues in the linguistic input that are useful for constructing the grammar of the language. Second, learners must represent their knowledge of the grammar in a form that makes it possible to assess the grammaticality of future input. With an appropriate representation of the grammar, learners can generalize from properties of the small set of experienced items to predicted properties of novel items. This ability for generalization is crucial for language acquisition, as the input for learning is naturally limited. Such generalization should extend to only the novel items that are actually licensed by the language, no more (over-generalization) and no less (under-generalization).

Previous research has offered several hypotheses regarding the cues that learners use and the representations of grammar they form. In the realm of syntactic category acquisition, one hypothesis is that the categories (but not their contents) are innately specified prior to receiving any linguistic input, with the assignment of words to categories accomplished with minimal exposure (e.g. McNeill, 1966). On this view, both the cues and the representations are predefined and independent of linguistic input. A contrasting view states that grammatical categories are learned, though different hypotheses appeal to the importance of different cues or cue combinations during the learning process (such as semantic cues, e.g., Bowerman, 1973). Within this class of non-nativist hypotheses, several studies have suggested that distributional cues may be sufficient for extracting the grammar of the input language (e.g., Braine, 1987; Maratsos & Chalkley, 1980; Mintz et al., 2002). Distributional cues are defined over patterns in the linguistic input, such as token frequencies, co-occurrence statistics, and latent structural dependencies be-

tween linguistic elements. Although studies have shown that human learners and computational models can successfully learn grammatical categories when only these cues are available, the question of representation still remains poorly understood. How do learners represent the knowledge of previously encountered linguistic items in order to generalize to novel ones?

The aim of the present work is to ask what types of representations are used by human learners in an artificial grammar learning (AGL) task that includes many of the distributional properties of spoken language. We focus on how learners induce grammatical categories and assign words to them. Our approach involves computational modeling, comparing the simulated learning outcome of three different models, each of which makes a different assumption about how learners represent the learned grammar. We assess the models by comparing the generalization patterns of each model and those of human learners. Our experimental data come from our previous findings across 10 AGL experiments (Reeder et al., in review; Schuler et al., in prep). In the next section, we first provide a brief summary of these results. Importantly, the goal of our modeling work is not to mirror every detail of human behavior in AGL experiments: to do so, one must consider psychological variables such as memory and attention, which are currently not included in our models. Instead, we are interested in exploring the representational assumptions that human learners have adopted in our experiments.

## Background on Behavioral Results

The behavioral data come from a series of 10 experiments with adult participants in which we created an artificial grammar with the structure (Q)AXB(R). Each letter represents a category of nonsense words. Q and R words served as optional categories that made sentences of the language vary in length from 3 to 5 words and made words of the language observe patterning in terms of relative order but not fixed position. The sizes of the categories varied across experiments, leading to different numbers of possible sentences in the language. For ease of presentation, we will number the experiments. In Experiments 1-4 (Reeder et al., 2009), there were 108 possible sentences that could be created from this grammar; in Experiment 5 (Reeder et al., 2009), there were 576 possible sentences; in Experiments 6-10 (Reeder et al., 2010;

Schuler et al., in prep), there were 144 possible sentences.

Participants in these experiments were first exposed to a carefully selected subset of the possible sentences of the grammar. The exposure strings were chosen to test whether specific distributional cues enabled learners to form a category of lexical items and generalize to novel words, or to allow exceptions that maintain lexical specificity. In particular, different experiments tested learners sensitivity to the *contexts* of individual words and their individual frequencies, the *sparsity* of sampling the language, the *overlaps* among contexts across words, the non-overlap of contexts (or *systematic gaps* in information), and the size of the exposure set. In each experiment, a portion of the possible strings was withheld in order to create different kinds of “gaps” in the input to participants.

After exposure, subjects completed a grammaticality rating task, where they rated strings on a scale from 1-5, with larger values indicating higher grammaticality. The test was comprised of three types of test strings: familiar AXB sentences (presented during exposure), novel AXB sentences (withheld from the exposure set), and ungrammatical strings that violated the AXB word order (i.e., “A1X1A2” or “A1A2B3”). Importantly, in order to understand how learners generalized from training sentences and the type of knowledge representations suggested by their generalization behaviors, we varied the way the presentation set and the gaps occurred in each experimental condition, as summarized in Table 1. In Experiments 1-2, we varied the sparseness of sampling the language, but learners heard all AX and XB bigrams. In Experiments 3-4, we varied the overlap of contexts across X words: each X was heard with only 2 of the 3 As and Bs. Experiments 6-9 included a new X word (called “X4”) that appeared in only one sentence frame in the training subset. The purpose was to test whether learners would generalize to X4 as one of the X words (and therefore able to occur in all X-word contexts), despite its own extremely limited exposure and minimal overlap with the other X-word contexts. Experiment 5 created subcategories in the language, with distinct occurrence privileges for X words and contexts words: half of the X words only occurred with half of the A words and half of the B words, while the remaining X words occurred with the remaining As and Bs.

In experiments 1-9, the bigram statistics were carefully balanced: all grammatical bigrams were presented equally often (with the exception of the X4 bigrams in Experiments 6-9). Under this balanced design, one possible strategy for judging grammaticality could be simply to keep track of bigram statistics. To examine this, we ran Experiment 10, where the bigram statistics were not balanced.

By definition, the generative grammar used in all these experiments is the same: (Q)AXB(R). However, our distributional manipulations across all of these experiments led human subjects under certain circumstances to restrict generalization to be maximally compatible with the input, while in other circumstances to generalize to the full grammar.

Table 1: Descriptions of the sampling bias in each experiment

Experiment	Sampling bias
Expt 1	Uniformly Distributed Gaps, Dense Sampling (1/3 withheld): Every X-word heard with every A- and B-word
Expt 2	Uniformly Distributed Gaps, Sparse Sampling (2/3 withheld): Every X-word still heard with every A- and B-word
Expt 3	Systematic Gaps, Sparse Sampling: Each X-word heard with a subset of possible A- and B-words
Expt 4	Extended Exposure to Systematic Gaps: Same as Experiment 3, but exposure was tripled
Expt 5	Subcategorization: Gaps were inserted such that a clear divide segregated X-words and contexts words into two subcategories
Expts 6-9	Same as Experiments 1-4, but included a very minimally overlapping X-word (X4); X4 seen in just one sentence frame in each condition
Expt 10	Same as Experiment 3, but bigram statistics are not balanced because words varied in frequency

Learners rated novel grammatical sentences as high as familiar grammatical strings in Experiments 1 and 2, showing a strong tendency to generalize across the words within a category. In Experiments 3 and 10, where a systematically-gapped training set was presented (balanced or not), learners became more conservative and treated novel grammatical sentences as somewhat less grammatical than familiar ones (but still more grammatical than ungrammatical ones). Generalization was further reduced in Experiment 4, when the exposure to systematic gaps was increased. In the subcategorization experiment (Experiment 5), learners did not fully generalize across the gaps created by the subcategory structure, indicating that they used the distributional information to learn that there were two subcategories within the X category. Lastly, the results of Experiment 6 showed that when learners were given a dense sampling of a language with almost complete overlap of contexts for several words in the X category, learners generalized a novel word (X4) to the full range of grammatical contexts of the other X-words, even when they heard X4 in only one of those contexts. When contexts were more sparse (Experiment 7) and there were significantly more systematic gaps in the input (Experiments 8 & 9), learners did not fully transfer their knowledge of X-category structure to the minimally overlapping X4 word. In all experiments, ungrammatical sentences were rated significantly lower than any novel grammatical test string.

## Models

We use a generative model-based framework to develop our three models. The structures of these generative models make explicit the assumptions about knowledge representations. The goal of our modeling effort is to understand what elements must be included in the representation of the QAXBR grammar so that the models’ generalization behavior will be

most compatible with human behavior across all 10 experiments. The answer to this question is related to the types of distributional cues that human learners attend to in the experiments. For the models reported in this paper, we make the simplifying assumption that learners only attend to local bigram information in the input, the bare minimum to capture the sequential dependencies within QAXBR sentences (although our models can easily be extended to use other distributional cues). A successful model should assign high probabilities to grammatical sentences and low probabilities to ungrammatical ones. Crucially, a successful model should assign probabilities to novel grammatical sentences that match the ratings of human learners.

### Word Bigram Model

The first model is the *word bigram model*: the probability of a sentence is simply the product of the probabilities of its ordered word pairs, where the probability of each word  $w_i$  is conditioned on the preceding word  $w_{i-1}$ :

$$p(s) = \prod_{w_i \in s} p(w_i | w_{i-1}) \quad (1)$$

Equation (1) can be interpreted to suggest that a word bigram model represents the grammar with a set of multinomial distributions. Each distribution specifies the probabilities that a word will be followed by any other words in the vocabulary. The parameters of these distributions are typically estimated from training data with maximum likelihood estimation (MLE). However, the standard MLE algorithm is insensitive to sample size, which is a crucial variable of interest in several experiments. When comparing Experiments 3 and 4, for example, our subjects exhibited different generalization patterns as a result of the change in the amount of exposure to the same set of training data (i.e., a change in sample size only). Therefore, we adopt a Bayesian approach that is sensitive to sample size. The fully derived form for estimating the probability of a word is:

$$p(w_i | w_{i-1}, \text{all previous bigrams}) = \frac{n_{w_{i-1}, w_i} + \beta}{\sum_k n_{w_{i-1}, w_k} + v\beta} \quad (2)$$

where  $v$  is the vocabulary size,  $n_{w_{i-1}, w_i}$  is the frequency of bigram  $(w_{i-1}, w_i)$ , and  $\beta$  is a free parameter. The  $\beta$  parameter determines whether certain parameter settings of the multinomial distributions are favored. Here, we report results with  $\beta$  set to 1, which is a non-biased prior.

**Simulation Procedure** In each experiment, the word bigram model first estimates its model parameters according to the training sentences. In experiments where the length of exposure is a predictor of interest (Expts. 3, 4, 8 & 9), we duplicate the training data to simulate the effect of extended exposure. Unlike human subjects, however, the model is given information regarding the size of the vocabulary, and does not have memory limitations.

### Word Bigram Mixture Model

The word bigram model implies that there is one single representation that corresponds to the grammaticality of a sentence. Natural languages, however, are usually more flexible: a sentence can have many different types of grammaticality (or ungrammaticality), such as Noun-Verb agreement, as well as lexical restrictions, such as transitive/intransitive verbs. We address this problem by developing a word bigram mixture model, where multiple patterns of grammaticality can be modeled simultaneously. Each component in the mixture is a word bigram model. A grammatical sentence is generated from a component grammar, which is in turn generated from a stochastic process (the model can be viewed as a Dirichlet process mixture model; Ferguson, 1973). We can describe the process of generating a sentence  $s$  in two steps:

- (a)  $p(s \text{ is generated by an existing component } k) = \frac{n_k}{n + \alpha}$   
 (b)  $p(s \text{ is generated by a new component}) = \frac{\alpha}{n + \alpha}$
- If (a),  $p(s = w_1, \dots, w_m) = p(w_1, \dots, w_m | \mathcal{B}_k)$   
 If (b),  $p(s = w_1, \dots, w_m) = p(w_1, \dots, w_m | \mathcal{B}_{new})$

where  $n_k$  is the number of sentences that have been generated as instances of component grammar  $k$ ,  $n$  is the number of sentences that have already been generated,  $\alpha$  is a free parameter of the model (a larger  $\alpha$  leading to more new clusters), and  $\mathcal{B}$  refers to the parameters of a component bigram model (as described in the previous section). Combining these two steps, the probability that  $s$  will be generated by a word bigram mixture model is

$$p(s = w_1, \dots, w_n) = \frac{\sum_k p(s | \mathcal{B}_k) n_k + p(s | \mathcal{B}_{new}) \alpha}{n + \alpha} \quad (3)$$

**Simulation Procedures** Equation (3) describes a generative model, with which we can assess the probability that a sentence is generated by an existing representation of the grammar. However, the learner faces the opposite problem: they must infer the representation given the observed sentences. We used the Gibbs sampling method to infer these parameters (the exact details are not described due to space limits). We run the model on the training data used in the experiments. The first 500 samples of each run are discarded (which may be biased towards initial values). Due to the small scale of our artificial language, the sampler converges quickly, well within the discarded 500 samples. Each of the remaining posterior samples is considered as a candidate representation of the grammar. For experiments with longer exposure, we also run the sampler longer to approximate the effect. The average probability that a sentence is generated by these posterior representations is taken as a measure of the grammaticality of the sentence.

### Category Bigram Mixture Model

A notable feature of the two models presented so far is the lack of explicit representation for grammatical categories.

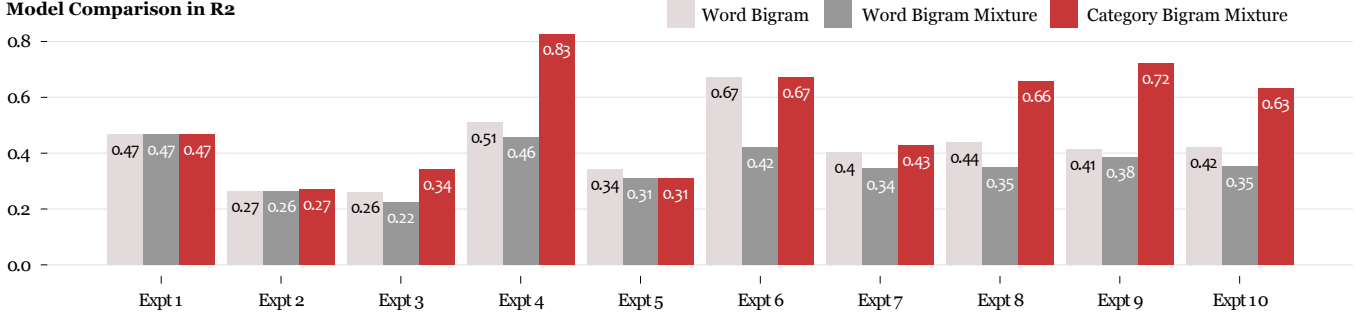


Figure 1: Grammaticality predictions made by the category bigram mixture model best approximate subject ratings in most experiments.  $R^2$  is calculated by regressing subject ratings against model predictions.

In both models, bigram statistics are based on word tokens. However, a crucial component of language acquisition involves organizing words into grammatical categories and discovering relations between them. To investigate whether human learners were in fact organizing the words into categories, we also developed the category bigram mixture model. The category bigram mixture model preserves the notion of multiple component grammars and introduces a bigram-based word category discovery process nested within each component grammar. In other words, the component grammars in the category bigram mixture model are themselves infinite mixtures of bigram models on categories. Therefore, generating a sentence under the grammar is a two-step process with the second step containing another two-step process:

- (a)  $p(s \text{ is generated by an existing component } k) = \frac{n_k}{n+\alpha}$   
 (b)  $p(s \text{ is generated by a new component}) = \frac{\alpha}{n+\alpha}$
- If (a), for the category of each word,  $c_i$ :
  - (i)  $p(c_{i-1}, c_i \text{ belongs to existing bigram } l) = \frac{n_l^k}{n^k + \alpha_0}$
  - (ii)  $p(c_{i-1}, c_i \text{ is novel}) = \frac{\alpha_0}{n^k + \alpha_0}$
  - If (i),  $p(w_i) \sim \text{MultiNomial}(c_i | c_{i-1})$
  - If (ii),  $p(w_i) \sim \text{MultiNomial}(c_{\text{new}} | c_{i-1})$
- If (b), for the category of each word,  $c_i$ :
  - (i)  $p(c_{i-1}, c_i \text{ belongs to existing bigram } l) = \frac{n_l^{\text{new}}}{n^k + \alpha_0}$
  - (ii)  $p(c_{i-1}, c_i \text{ is novel}) = \frac{\alpha_0}{n^k + \alpha_0}$
  - If (i),  $p(w_i) \sim \text{MultiNomial}(c_i | c_{i-1})$
  - If (ii),  $p(w_i) \sim \text{MultiNomial}(c_{\text{new}} | c_{i-1})$

where in the top-level process,  $n_k$  is the number of sentences that have been generated by component grammar  $k$ ,  $n$  is the total number of generated sentences,  $\alpha$  is the free parameter (as in the word mixture model) influencing the tendency of creating more component grammars. For clarity, we write  $c_{i-1}, c_i$  as the bigram label that each bigram  $l$  is associated with in the nested process:  $n_l^k$  is the frequency of category bigram  $l$  in component grammar  $k$ ;  $n_k$  is the total number of category bigrams in component grammar  $k$ , and  $\alpha_0$  is a free

parameter (a larger value leading to more category bigrams). Finally, the probability that each word is generated from its category  $c_l$ , conditioned on the category of its preceding word  $c_{i-1}$ , is modeled as a multinomial distribution.

**Relation to other models** The problem of discovering categories for word tokens in a language is analogous to the problem of part-of-speech tagging in computational linguistics, which has been under active research for several decades. Our category bigram mixture model is most similar to the Bayesian unsupervised tagging algorithm developed by Goldwater & Griffiths (2007). While our approach is not fully Bayesian (in the sense that hyper-parameters are treated as free parameters), it has the flexibility of discovering multiple part-of-speech sequence patterns (i.e. component grammars) and creating as many part-of-speech tags as needed (due to the nested Dirichlet Process).

**Simulation Procedure** As in the case of the word bigram mixture model, Gibbs sampling is applied to the inference problem to find samples of the posterior distribution. Each of the remaining posterior samples is considered as a candidate representation of the grammar under the category bigram mixture model. The average probability that a test sentence will be generated by these representations is taken as a measure of the grammaticality of the sentence.

## Results and Discussion

Model predictions are in the format of probability estimates. A higher probability estimate means that a sentence is more grammatical. The quality of model predictions is determined by examining how well they correlate with subject ratings. To ensure that subject ratings are maximally comparable with model predictions, we transformed discrete ratings into z-scores within each subject and experiment, so that the ratings of subjects with consistent biases (consistently high or consistently low ratings) were normalized. We computed the  $R^2$  metric for each group using a linear regression where model predictions were used to predict subject ratings. A model with a high  $R^2$  indicates that the particular model explains a significant amount of variance in subject ratings (see Fig 1). Overall, the category bigram mixture model best captures hu-

man behavior across all 10 experiments combined ( $R^2$  Word Bigram = 0.4,  $R^2$  Word Bigram Mixture = 0.35,  $R^2$  Category Bigram Mixture = 0.47).

The general advantage of the category bigram mixture model suggests that our human learners may have acquired a representation of an X category, and not just a set of simple word co-occurrences. In X4-related experiments (Expts. 6-9), we asked whether learners could extend their knowledge of a target category to a very infrequently presented word for which they only had minimal context information. We found that there was a point in learning where hearing just one context for the minimally overlapping X4 word was enough to generalize full category privileges for that word (Expt 6). Simple word co-occurrence and bigram counts will not achieve this outcome. The category bigram mixture model, however, has the appropriate representation for supporting such a learning outcome. Indeed, in Experiment 6, X4 gets assigned to the same category as all other X-words, thus enabling the generalization to novel X4 sentences (the effect of X4 sentences on overall  $R^2$  is reduced by the extremely small number of X4 sentences in the testing phase).

### Limitations of the category bigram mixture model

While the category bigram mixture model best approximates human generalization patterns across the 10 experiments, it does no better than the other two models in capturing human performance in the subcategorization experiment. Indeed, the two mixture models acquire the subcategory structure, but fit human performance no better than the simplest word bigram model. This paradoxical result is due to the experimental design: all bigrams in the training subset conform to the subcategory boundaries and are presented equally often. At test, novel subcategory-conforming items are rated as high as familiar ones because they contain only bigrams that have been presented (thus indistinguishable from familiar ones). Test strings violating the subcategory structure are rated low by the word bigram model simply because they contain one or two bigrams which are never seen in the training data. The balanced presentation of all within-subcategory bigrams enables the word bigram model to distinguish between subcategory conforming and violation items without learning the existence of two subcategories. As a result, even though the two mixture models successfully discover the existence of two subcategories, the additional advantage of such discoveries is minimal.

The category bigram mixture model also tends to overgeneralize in experiments with systematic gaps. This is most clearly demonstrated in Experiments 4 and 9 (see Fig 2). In those experiments, subjects were exposed to a language with frequent systematic gaps in the input. Human learners gave novel grammatical sentences a significantly lower rating than familiar grammatical strings, especially when the training materials were presented multiple times. We view this restriction of generalization as a rational behavior that prevents human learners from over-generalizing when systematic and persistent gaps occur in the input.

### Normalized Grammaticality Rating (y-axis)

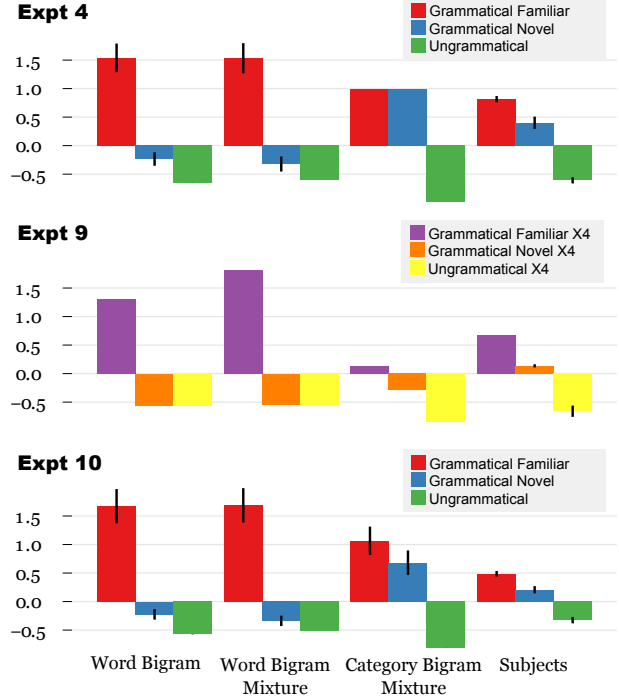


Figure 2: Z-score normalized grammaticality ratings by models and human subjects in 3 experiments where the Category Bigram Mixture model had highest  $R^2$ . Error bars = SE. Regarding the difference between familiar and novel ratings, Expt 4 shows overgeneralization by the category bigram mixture model, Expt 9 shows overly conservative behavior of the category bigram mixture model, and Expt 10 shows the category bigram mixture model best capturing human behavior, despite systematic gaps and variable bigram frequencies.

In the word-based models, the reduced generalization is captured because an increase in the probabilities of observed word bigrams necessarily leads to a decrease in the probabilities of unobserved ones, thus producing restricted generalization. However, this effect is much weaker in the category bigram mixture model, where repeated exposure to training sentences only strengthens the category bigram dependencies. When a novel grammatical sentence is presented to this model, its category bigrams have been observed many times during training and the sentence will receive a relatively high rating as a result (despite this,  $R^2$  is relatively higher in the category bigram mixture model because its predictions are qualitatively closer to subject ratings; see Fig 2). This is an indication that learners have slightly different constraints on learning and/or a slightly different strategy from the category bigram mixture model. We are currently exploring other possible models that build on the idea of an underlying category representation, but incorporate learning constraints that more closely mimic human learning (e.g. incremental models of learning) and lead to the construction of a more restrictive grammar that is still compatible with the input.



This pattern can be contrasted with model performance on Expt 10, where not all grammatical bigrams are seen equally often during exposure. Results from this experiment make clear that the category bigram mixture model is the most robust to manipulations of the bigram distribution (see Fig 2). The other two models rate grammatical novel sentences almost as low as ungrammatical sentences, since novel grammatical sentences contain novel and low-frequency bigrams. By definition, these are less grammatical to the word-based models due to having a lower probability. The category mixture model, on the other hand, is not negatively influenced by the unbalanced design due to the abstraction of word categories.

## General Discussion

Across 10 experiments, we compared the grammaticality predictions of three different models to human subject ratings. Our primary interest was to find the representational elements of the grammar that are most compatible with the generalization behaviors displayed by the learners. Generalization depends on the ability to abstract over categories, which is fundamental to linguistic productivity. A number of researchers have asked whether there is adequate distributional information in the input to form linguistic categories. Previous work uses hierarchical clustering and a computational learning mechanism to attempt to deduce grammatical categories from corpora of child directed speech based solely on distributional analyses of the input (e.g. Mintz et al., 2002; Redington et al., 1998). These models have been able to use co-occurrence statistics among words to achieve relatively good categorization performance for frequent target words, indicating the utility of these types of distributional cues for categorization.

The behavioral experiments that this work is built upon suggest that the patterning of word tokens in a substantial corpus of linguistic input appears to be sufficient to extract the underlying structural categories in a natural language, given an appropriately capable learner. Our modeling results have further explicated the representational assumptions for extracting the knowledge of a grammar. Of the three models, two models are based on simple word bigrams collected from training data. While word bigrams are useful for capturing the lexical dependencies of the grammar, they cannot explain how human learners could generalize from experienced examples to novel items, especially when the prior experience is minimal (i.e., Expts 6-9). Such rapid and automatic generalization behavior calls for a richer representation, in which the grammar of the artificial language is organized around potential categories of vocabulary words. The category bigram mixture model introduces the notion of categories as a representational assumption, which led to model predictions that better approximated the behavior of human learners in almost all experiments. A limitation of the category bigram mixture model, however, is that it overgeneralizes compared to human performance. A fourth type of model could add a

generative component that asks how likely it is that a string is absent given a random sampling process. If that probability is low, then it would penalizes the probability even further by downweighting it in the grammar. We are exploring this direction, in conjunction with other mixture models that may more closely mirror the constrained learning environment that human learners face during natural grammatical category acquisition.

## Acknowledgments

This research was supported by NIH Grants HD037082 to RNA and DC00167 to ELN, and by an ONR Grant to D. Bavelier at the University of Rochester.

## References

- Bowerman, M. (1973). Structural relationships in childrens utterances: Syntactic or semantic? In T. Moore (Ed.), *Cognitive development and the acquisition of language*. Harvard University Press.
- Braine, M. (1987). What is learned in acquiring word classes a step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (p. 65-87). Lawrence Erlbaum Associates.
- Ferguson, S. (1973). A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 209-230.
- Goldwater, S., & Griffiths, T. (2007). A fully bayesian approach to unsupervised part-of-speech tagging. In *ACL*.
- Maratsos, M., & Chalkley, M. A. (1980). The internal language of childrens syntax: The ontogenesis and representation of syntactic categories. In K. Nelson (Ed.), *Childrens language* (Vol. 2). Gardner Press.
- McNeill, D. (1966). Developmental psycholinguistics. In *The genesis of language: A psycholinguistics approach* (p. 69-73). MIT Press.
- Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26, 393-425.
- Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22, 435-469.
- Reeder, P. A., Newport, E. L., & Aslin, R. N. (2009). The role of distributional information in linguistic category formation. In *CogSci 2009* (p. 2564-2569).
- Reeder, P. A., Newport, E. L., & Aslin, R. N. (2010). Novel words in novel contexts: The role of distributional information in form-class category learning. In *CogSci 2010* (p. 2063-2068).
- Reeder, P. A., Newport, E. L., & Aslin, R. N. (in review). *From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes*.
- Schuler, K., Reeder, P. A., Newport, E. L., & Aslin, R. N. (in prep). *The effects of uneven frequency information in linguistic category formation*.

# Acoustic analysis supports the existence of a single distributional learning mechanism in structural rule learning from an artificial language

**Okko Räsänen (okko.rasanen@aalto.fi)**

Department of Signal Processing and Acoustics, Aalto University,  
Otakaari 5 A, FI-00076 Aalto FINLAND

**Heikki Rasilo (heikki.rasilo@aalto.fi)**

Department of Signal Processing and Acoustics, Aalto University,  
Otakaari 5 A, FI-00076 Aalto FINLAND

## Abstract

Research on artificial language acquisition has shown that insertion of short subliminal gaps to a continuous stream of speech has a notable effect on how human listeners interpret speech tokens constructed from syllabic constituents of the language. It has been argued that the observed results cannot be explained by a single statistical learning mechanism. On the other hand, computational simulations have shown that as long as the gaps are treated as structurally significant units of the language, a single distributional learning model can explain the behavioral results. However, the reason why the subliminal gaps interfere with processing of language at a linguistic level is currently unknown. In the current work, we concentrate on analyzing distributional properties of purely acoustic representations of speech, showing that a system performing unsupervised learning of transition probabilities between short-term acoustic events can replicate the main behavioral findings without a priori linguistic knowledge.

**Keywords:** language acquisition; pattern discovery; distributional learning; acoustic analysis; lexical learning

## Introduction

There is an ongoing debate regarding the degree that distributional learning mechanisms can explain aspects of language acquisition from speech, and the degree that rule-based mental processes are required in the task (e.g., Endress & Bonatti, 2007; Laakso & Calvo, 2011; Peña et al. 2002). Experimental studies with human test subjects have shown that both infants and adults are able to learn statistical regularities in continuously spoken artificial languages and use these regularities to segment speech into word-like units (e.g., Peña et al. 2002; Saffran, Aslin & Newport, 1996). Based on these findings, it has been suggested that the listeners may be using transitional probabilities (TPs) between speech units such as phones or syllables in order to discover statistically regular segments of speech (e.g., Saffran et al., 1996). Computational simulations have also verified that the TPs between signal events can be used to discover word-like units from continuous speech, and that these units do not necessarily need to be linguistic or phonetic in nature (Räsänen, 2011).

Of especial interest is the degree that distributional learning can explain the learning of non-adjacent dependencies in a language. In earlier work, the learning of non-adjacent dependencies has been studied using an

artificial nonsense language consisting of three-syllabic CVCVCV words with the middle syllable being always randomly selected from a pool of “fillers”, but the first and last syllable occurring always together (hence a “*high-probability word*”). It has been found out that when human listeners are familiarized with a continuous stream of such language without gaps between the high-probability words, and then later tested for preference between three-syllabic words that have different TPs between the syllables in terms of the familiarization stream, the listeners seem to prefer words that have occurred with higher internal TPs in the familiarization stream (Endress & Bonatti, 2007; Peña et al. 2002). However, introduction of 25 ms subliminal segments of silence between the high-probability words in the familiarization stream leads to a notable change in the learning outcome: the listeners start to prefer word forms that do not necessarily have the highest TPs across all syllables in the word. Instead, the preferred words may contain partially novel surface form but have dependencies between syllables that can be explained by abstract rules that are also valid for the words in the familiarization stream (Endress & Bonatti, 2007; Peña et al. 2002).

The above finding is somewhat unexpected from the perspective of distributional learning at a linguistic level. The learning results between continuous and gapped familiarization streams should not differ as long as the perceived linguistic units and their ordering in the two conditions do not differ either. The result is also counterintuitive due to the fact that the gaps are tiny in duration in comparison to the other relevant signal segments such as syllables, and since CV-syllable based languages already contain natural silences associated with closures of plosives (e.g., word “#pura#ki”, where # denotes a closure).

Peña et al. (2002) and Endress and Bonatti (2007) suggest that the additional silent gaps provide direct (but unconscious) cues to the segmentation of words from speech, freeing computational resources to structural learning of rule-like relations between constituents of the words. On the contrary, the absence of the gaps necessitates that the segmentation has to be first learned from the data (Endress & Bonatti, 2007; but see also discussion in Laakso & Calvo, 2011). It is therefore argued that the change in learning outcomes after introduction of the gaps provides evidence for non-distributional learning of structural relations between syllabic units (Bonatti & Endress, 2007).

However, a possible auditory processing mechanism for differentiating gaps associated with segmental cues and, e.g., the intra-word gaps related to closures of plosives has not been described in the existing work.

Lately, Laakso and Calvo (2011) have shown that the experimental results of Peña et al. (2002) and Endress and Bonatti (2007) *can* actually be modeled with a single distributional connectionist model when the silent gaps are represented as equally significant units as the consciously perceived syllables. As long as Occam's razor is concerned, the distributional model of Laakso and Calvo (2011) provides a more coherent and simple explanation for the observed data instead of resorting to the more than one mechanisms (MOM) hypothesis of Peña et al. (2002) and Endress and Bonatti (2007). However, the model of Laakso and Calvo also has a shortcoming: it does not explain how the short subliminal gaps end up with an equally large role as the syllabic units in the distributional learning process.

The goal of the current work is to study the distributional learning hypothesis in the context of the artificial language of Peña et al. (2002) by focusing on the analysis of recurring acoustic patterns in a speech stream. Unlike earlier work, we study TPs of short-term acoustic events instead of linguistically or phonetically motivated units such as syllables or segments. This provides a novel perspective to the learning problem by assuming that the listeners may not be directly analyzing the speech stream as a sequence of linguistic units, but may treat the language-learning task as a generic auditory patterning problem. Still, the current approach does not exclude the possibility that the listeners can extract basic recurring units such as syllables or segments from the acoustic speech stream and perceive these as linguistically significant units. We simply show that the behavioral results of Peña et al. (2002) and Endress and Bonatti (2007) can be explained with a single distributional learning mechanism that performs pattern discovery at the level of acoustic signal instead of assuming TP analysis of segments or syllables.

### Motivation for Acoustic Learning

There are multiple reasons to assume that the listeners may utilize generic acoustic patterning instead of purely linguistic coding of input during perception of an artificial language. First of all, test subject preferences towards specific test probe types are typically only slightly above chance level even for extended familiarization periods (Peña et al., 2002; Endress & Bonatti, 2007). If the learning would be based on categorically perceived segments or syllables, one could expect more robust preference for one probe type over another due to the systematically different overall TPs or learned rules for the tokens. Also, the initial preference for specific probe types degrades over longer familiarization periods, suggesting that the low-level distributional properties of the speech stream interfere with the processing of the abstract generalizations. Finally, the introduction of subliminal gaps introduces notable qualitative changes to the learning outcomes. Since these gaps are clearly not

serving any explicit linguistic function but still affect the learning results, it can be taken as evidence that the acoustic level perception, including temporal relationships of acoustic patterns, may play an important role in the process.

Why distributional analysis at the acoustic level would then lead to different results than analysis on the segmental or syllabic level? The major difference comes from temporal relationships between sound events. At the syllabic level, the relevant units and their distances from each other are well defined. Therefore the TP statistics also become well defined after a small number of word occurrences in different contexts. At the acoustic level, a syllable is not perceived as a categorical unit with a well-defined duration, but as a constantly evolving spectrotemporal trajectory that has very low predictability over larger temporal distances. This means that the typical acoustic level dependencies are limited to a time scale much shorter than the tri-syllabic words in the artificial language of Peña et al. (2002). Therefore the acoustic TP analysis must also pay attention to dependencies at a very fine temporal resolution, potentially increasing the relative role of temporal asynchronies caused by the introduction of silent gaps to the familiarization stream.

### Material

The speech material for the experiments was reproduced from the work of Peña et al. (2002). In this material, the familiarization stream of the artificial language consists of three CV-syllable words of form  $A_iXC_i$  so that each word starts with one of three possible syllables  $A_i$  ( $i \in \{1,2,3\}$ ). Importantly, the first syllable always uniquely determines the last syllable  $C_i$  of the word (i.e.,  $P(C_i|A_i) = 1, \forall i$ ) so that there are also three different possibilities for end syllables. Finally, the medial syllable, or *filler*, is chosen randomly from a set of three CV syllables. In total this produces three word templates “pu ... ki”, “be ... ga”, and “ta ... du” where one of the following three fillers are used in the medial position: “li”, “ra” or “fo”.

Based on Endress and Bonatti (2007), four types of probes were used during testing: 1) *words*, i.e., tri-syllable constructs that correspond directly to the ones used in the familiarization (e.g.,  $A_iXC_i$ ), 2) *part-words*, where the sequential order of syllables was from the familiarization data but the word straddles a word boundary (e.g.,  $XC_iA_j$ ), therefore having a smaller overall TPs across the word, 3) *rule words* of form  $A_iX'C_i$ , where the  $X'$  is familiar from the training but has never occurred in the word-medial position, and 4) *class words* of form  $A_iXC_j$  ( $i \neq j$ ) so that all  $A_i$ ,  $X$ , and  $C_j$  are familiar from the familiarization data but the  $A_i$  and  $C_j$  have never occurred in the same word (see Endress & Bonatti, 2007, for detailed word lists).

The familiarization data and test probes were synthesized into speech signals using a Kelly-Lochbaum model based articulatory synthesizer of Rasilo, Räsänen and Laine (in preparation) using articulatory positions of Finnish vowels as targets for the vowel sounds. Sampling rate of the signals was set to 16000 Hz and fundamental frequency of the

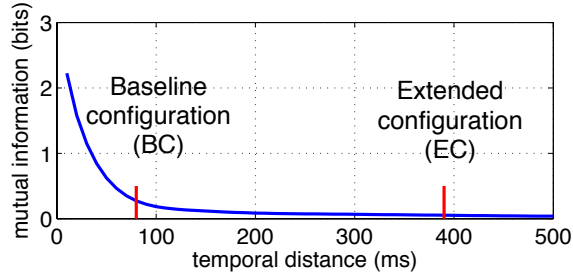


Figure 1: Temporal dependencies of acoustic events measured from continuous English speech. The two learning parameter configurations BC and EC are also shown.

speaker was set to 120 Hz. In order to create familiarization data, all words in a training epoch (one occurrence of each word) were concatenated into one long string before synthesis so that the coarticulatory effects were consistent for both intra-word and across-word transitions. In addition to the continuous stream, the gapped familiarization stream of Peña et al. was also created by inserting silent segments of 25 ms between the words. It was also confirmed perceptually that the perception of the gaps was subliminal and no other audible artifacts were introduced to the signals.

## Methods

### Preprocessing

The goal of the preprocessing was to convert synthesized speech signals into sequences of automatically discovered discrete acoustic events for further statistical modeling. This was achieved by extracting Mel-frequency cepstral features (MFCCs) from the signals using a window length of 25 ms and a step size of 10 ms (see, Appendix B in Räsänen 2011a for detailed description). A total of 12 coefficients + energy were used. A random subset of 10000 MFCC vectors from the familiarization data set was then clustered into 64 clusters using the standard k-means algorithm. The obtained cluster centroids were treated as prototypes for the corresponding clusters (“atomic acoustic events”) and each cluster was assigned with a unique integer label  $i \in [1, 2, \dots, 64]$ . Finally, all MFCCs vectors were vector quantized (VQ) by representing the original feature frames with labels corresponding to the nearest cluster centroids for the given frame. This led to a signal representation where the synthesized speech was represented as a sequence of discrete elements, each element being one of the 64 possible choices and one element occurring every 10 ms.

### Discovery of Acoustic Patterns

In order to learn distributional patterns from the artificial speech data, a statistical learning mechanism is needed. In the current work, we utilized the unsupervised word learning model of Räsänen (2011) that has been shown to be able to discover recurring word patterns from real continuous speech. This algorithm will be referred to as the *unsupervised distributional learning algorithm* (UDLA).

The basic principle of the UDLA is to study the TPs between the atomic acoustic events (VQ indices) in order to discover multiple segments of speech that share similar local TP distributions. Unlike typical distributional analysis of syllabic, phonemic, or orthographic units (e.g., Saffran, 1996), UDLA analyzes TPs between short-term acoustic events at several temporal distances (lags) in parallel so that dependencies between non-adjacent acoustic events also become modeled. When recognizing novel patterns, statistical support from all lags is combined in order to provide a uniform and noise robust estimate of familiarity of the pattern. Instead of modeling global TPs, UDLA creates a separate TP model for each novel pattern discovered from the data, where a novel pattern is defined as a sequence of acoustic events whose TPs do not match any of the previously learned patterns.

From the perspective of pattern discovery, it is beneficial to study temporal dependencies up to approximately 200 ms in case of continuous speech. This is because the statistical dependencies between acoustic events diminish to a non-existent level at larger temporal distances and provide no further support for pattern discovery (Räsänen & Laine, 2012). This temporal scale also corresponds to the typical signal integration times measured in human auditory perception in the context of loudness perception or forward masking of speech sounds, suggesting that the integration times in human hearing are matched to the typical temporal structure of acoustic speech signals. As an example, Figure 1 shows the statistical dependencies of short-term acoustic events as a function of temporal distance for continuous English speech measured in terms of mutual information function (MIF; Li, 1990). As can be observed from the figure, majority of the dependencies at the acoustic level are limited to temporal distances shorter than 100 ms.

Since the amount of statistical information diminishes at longer distances, one can hypothesize that the human hearing system would be adapted to process temporal dependencies at such timescale where, on average, dependencies do exist. Therefore, in *baseline configuration* (BC), we use UDLA in a mode in which dependencies are modeled up to 80 ms, capturing approximately 90 % of the statistical dependencies in terms of MIF (Fig. 1). However, we also measure UDLA behavior in the artificial language learning task using TP modeling up to 390 ms. This configuration will be referred to as *extended configuration* (EC). In terms of the current experiments, this means that the TPs were studied at lags  $\mathbf{k} = \{1, 2, \dots, 8\}$  for BC and at lags  $\mathbf{k} = \{1, 3, 5, \dots, 39\}$  for EC, corresponding to the modeling of acoustic dependencies at temporal distances of 10 ms – 80 ms and 10 ms – 390 ms, respectively.

The hypothesis was that, if acoustic and non-linguistic patterning can explain the results of the experiment of Peña et al. (2002), and if human hearing is actually specialized for learning dependencies according the curve shown in Fig. 1, the learning outcomes in the baseline configuration should have better correspondence to the behavioral results than the extended condition. On the other hand, the extended

configuration should show higher preference for part words than class or rule words due to the diminishing role of the gaps in terms of dependencies across all temporal distances.

**Training Phase** The learning process in UDLA proceeds as follows (see also Räsänen, 2011): the sequential discrete familiarization stream  $X$  is analyzed in windows of length  $L_r$  elements and window step size  $L_s$ . For each window position, the TPs between all elements  $a_i$  and  $a_j$  in the window are modeled in parallel for lags  $\mathbf{k} = \{k_1, k_2, \dots, k_K\}$ . For the TPs in the first window position, the first statistical model  $c_1$  is created by storing all transitions at all lags to a transition probability matrix. In the model, the probability of a transition from element  $a_i$  to  $a_j$  at lag  $k$  is defined as

$$P_c^S(a_j | a_i, k) = F_c(a_i, a_j | k) / \sum_{j=1}^{N_A} F_c(a_i, a_j | k) \quad (1)$$

where  $F_c(a_i, a_j | k)$  is the frequency of ordered pairs  $[a_i a_j]$  at distance  $k$  in the context of model  $c$ .

When the window is moved incrementally across the input sequence, all previously learned models are used to recognize the contents of the current window position. First, activation  $A_c(t)$  of each model  $c$  at each moment of time  $t$  is computed by calculating the mean of the TPs over all  $\mathbf{k}$ :

$$A_c(t) = \frac{1}{K} \sum_{k=1}^K P_c^S(X[t] | X[t-k], k) \quad (2)$$

The cumulative activation of each model is then calculated over the window and normalized by the window length:

$$A_c^{cum}(T) = \frac{1}{L_r} \sum_{x=T}^{T+L_r-1} A_c(t[x]) \quad (3)$$

where  $T$  denotes the window position. Now if activation  $A_c^{cum}$  of the most activated model  $c_M$  exceeds a pre-defined familiarity threshold  $t_r$ , the transition frequencies in the current window of analysis  $X_T$  are used to update the statistics of the model  $c_M$  according to Eq. (1). Otherwise, a new model  $c_N$  is created from the window contents using the Eq. (1). This process is repeated for the entire training data set, producing a set of models that incrementally increase their selectivity towards specific structures in the speech signal.

After the familiarization is complete, the learned models are normalized according to

$$P_c(a_j | a_i, k) = P_c^S(a_j | a_i, k) / \sum_{m=1}^{N_C} P_m^S(a_j | a_i, k) - \frac{1}{N_C} \quad (4)$$

where  $N_C$  is the total number of models learned. This changes the nature of the statistics so that now  $P_c$  describes how likely the given transition from  $a_j$  to  $a_i$  occurs in case of pattern  $c$  instead of any other pattern (i.e., *classification* task). The  $1/N_C$  term forces the total activation across all models to zero at all times, ensuring that the total activation level of the system does not increase with increasing number of learned models. Note that the learning process is purely incremental and requires the storage of the previous inputs only up to maximum lag  $K$  (i.e., 80 or 390 ms).

**Recognition Phase** During the testing phase, the test probes were pre-processed into discrete VQ sequences similarly to the familiarization data. Then the instantaneous activation of each model  $c$  at time  $t$  given input probe  $X$  was measured according to

$$A_c(t) = \frac{1}{K} \sum_{k=1}^K P_c(X[t] | X[t-k], k) \quad (5)$$

The total activation induced by the probe was then computed as

$$A_{tot} = \arg_{t,c} \max(A_c(t) | \forall t, c) \quad (6)$$

In other words, the total activation caused by the probe  $X$  was obtained as the maximum instantaneous activation<sup>1</sup> in the pool of all pattern models  $c$ .

## Experiments

In the experiments, UDLA was first used to discover recurring acoustic patterns from the familiarization stream, and then to recognize novel test probes using the learned models. During each test round, the system was shown one token from each of the four possible probe classes and the overall activation caused by each token was measured. A total of 600 probe quartets were generated by randomly sampling one token from each probe class for each quartet.

In all experiments, the UDLA model was run with a familiarity threshold of  $t_r = 0.16$  and window step size  $L_s = 50$  ms (5 frames). The analysis window length was set to  $L_r = 200$  ms and  $L_r = 600$  ms for baseline and extended conditions, respectively, so that multiple transitions at maximal lags would fit to the analysis window. These parameters led to the learning of  $N_C = 26-33$  acoustic patterns depending on the familiarization type (continuous vs. segmented), modeling conditions (baseline vs. extended), and on the duration of the familiarization. Since the number of learned patterns exceeded the number of unique syllables (nine), the system had learned multiple context-sensitive variants of syllable-like units.

Figure 1 shows the mean activation levels of the four different probe types (words, part words, rule words and class words) as a function of familiarization duration for segmented (top) and continuous (bottom) familiarization stream in the baseline condition with temporal dependency modeling up to 80 ms. As can be observed, the insertion of 25 ms gaps between tri-syllable words in the familiarization stream is sufficient to induce a change of preference from part words to rule words and class words. This is in line with the behavioral results of Peña et al. (2002) and Endress and Bonatti (2007) who found out that the use of subliminal

<sup>1</sup> The decoding criterion of probabilities was compared across numerous different possibilities, including, e.g., total activation of all models across the entire probe, temporally integrated maximum activation, and number of models exceeding a pre-defined threshold in activation. However, unlike the used approach in Eq. (6), none of the other criteria were able to replicate the main findings of Peña et al. (2002) and Endress & Bonatti (2007).

gaps in the familiarization stream causes a change of preference from part words to rule words and class words at short familiarization periods.

However, when the TPs between acoustic events are measured beyond the typical dependencies in speech signals, the situation changes notably. Figure 3 shows the mean activation levels of the probes in the extended condition where temporal dependencies are modeled up to 390 ms. Despite the fact that the only difference to the earlier simulation is the distance up to which TPs are measured, there is no sign of difference between the continuous and segmented familiarization streams.

Based on the mean probe activities, it seems that the distributional learning of acoustic patterns without any a priori or intervening linguistic component can explain the experimental results of Peña et al. (2002) and Endress and Bonatti (2007), but only if it is assumed that the system is able to learn acoustic dependencies up to a limited temporal distance defined by typical structure in continuous speech. If the dependency modeling is extended up to much longer delays, the UDLA model is no longer able to replicate the behavioral findings.

In addition to computing overall activations, pair-wise comparisons of probe activities were carried out for all possible probe pairs in the test set in order to simulate behavior in a forced-choice task similar to the one used with human experiments.

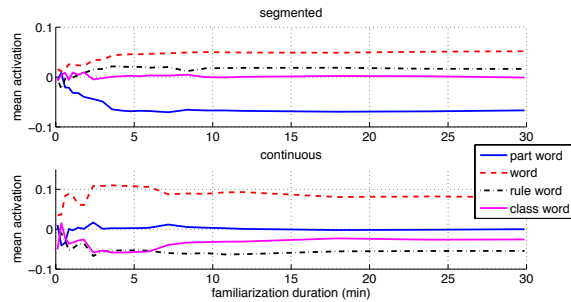


Figure 2: The mean activation levels of the four different probe types in *baseline condition* for segmented stream (top) and for continuous stream (bottom). Only relative mean activations of the probes are shown (zero mean).

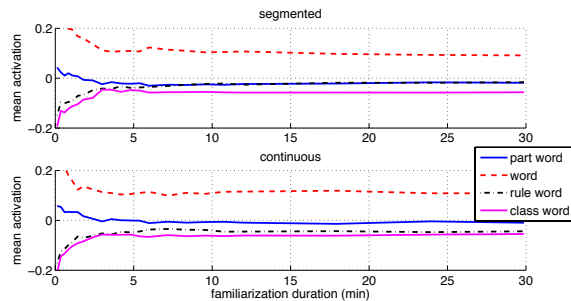


Figure 3: The mean activation levels of the four different probe types in *extended condition* for segmented stream (top) and for continuous stream (bottom).

More specifically, the relative probabilities of the tokens in each pair were compared separately across all 600 test cases in the baseline configuration. For each pair, a binary flag was used to denote a response for the probe that had the higher activation. Then the distribution of responses was tested against the null hypothesis that the model shows no preference for either probe type (*t*-test). Table 1 illustrates the results from the statistical analysis.

It is evident that the segmented familiarization stream leads to a preference order of *words* > *rule words* > *class words* > *part words* at short familiarization durations. On the other hand, continuous stream leads to order of *words* > *part words* > *rule words* and *class words*. This is largely in line with the results of Laakso and Calvo (2011), confirming that a single distributional learning mechanism can explain the change of preference between the two conditions. However, the previous studies do not always report statistically significant order of preference between all probe types (Laakso & Calvo, 2011), whereas the current simulations show statistically significant order of preference for all learning conditions except for the continuous familiarization stream of 3 minutes. This can be largely explained by the fact that the deterministic nature of UDLA leads to a consistent response pattern across multiple trials even for minor statistical biases between the probe types. In contrast, responses of human test subjects contain additional sources of variation (e.g., fatigue) and are based on a limited number of test trials, possibly rendering minor differences in probe familiarity invisible to statistical analysis.

## Discussion

In Peña et al. (2002) and Endress and Bonatti (2007) it was found that adult test subjects, when familiarized with 10 minutes of continuous stream of speech from an artificial language, prefer words over part words and show no preference between class words, part words and rule words. However, when subliminal gaps were introduced between words in the familiarization stream, the participants started to prefer class words and rule words over part words. Based on these findings, Peña et al. (2002) put forward the MOM hypothesis that the learning of a language might consist of several different processes: a distributional process responsible for discovery of statistically significant patterns and a separate mechanism responsible for modeling of structural relation between the discovered patterns. Endress and Bonatti (2007) provided further support to the MOM hypothesis by failing to replicate the behavioral findings of Peña et al. when modeling the learning task with a distributional system (a recurrent neural network or RNN).

Lately, Laakso and Calvo (2011) showed that RNNs can replicate the main behavioral findings of Peña et al. when the modeling parameters are properly set up, and when the silent gaps between syllables are modeled as separate units with equal importance to syllabic units. Their results undermine the argument for the necessity of multiple mechanisms of learning in this specific context. However, Laakso and Calvo limited their analysis to purely linguistic



Table 1: Pair-wise preference for the four different types of test probes with *segmented* (left) and *continuous* (right) familiarization streams. W stands for word, PW for part word, C for class word and R for rule word.

	segmented			continuous		
	preference	%	<i>p</i>	preference	%	<i>p</i>
3 min	W over PW	82.1	0.0000	W over PW	77.8	0.0000
	R over PW	74.0	0.0000	PW over R	57.1	0.0005
	C over PW	70.5	0.0000	PW over C	64.0	0.0000
	W over R	58.8	0.0000	W over R	89.5	0.0000
	W over C	68.3	0.0000	W over C	90.0	0.0000
	R over C	56.9	0.0032	No pref. R and C	51.4	0.5156
10 min	W over PW	78.4	0.0000	W over PW	71.2	0.0000
	R over PW	70.0	0.0000	PW over R	60.1	0.0000
	C over PW	69.1	0.0000	PW over C	57.0	0.0006
	W over R	68.4	0.0000	W over R	83.4	0.0000
	W over C	74.9	0.0000	W over C	79.0	0.0000
	No pref. R and C	55.9	0.0113	C over R	59.7	0.0000

level, assuming that the learner perceives artificial language as a sequence of syllabic units and silences even though the silences were not consciously perceived by the participants.

Current work studied the hypothesis that the findings of Pena et al. could be based on generic distributional learning at the acoustic level instead of using linguistic level representations. More specifically, we analyzed TPs of short-term acoustic events that were extracted from speech in purely unsupervised manner. Notably, we were able to replicate the behavioral findings related to the change of preference across familiarization conditions by using the UDLA model of word learning from continuous speech, but only when the TP analysis of acoustic events was limited to a temporal window matching to the temporal dependencies of normal continuous speech (Räsänen & Laine, 2012).

If this constraint is violated by exceeding the temporal scale of modeling to several hundreds of milliseconds, the model systematically prefers words over part words, and part words over class words or rule words also in case of segmented familiarization stream. The change of model behavior is driven by the fact that the synthesized speech lacks the acoustic variability and lexical complexity of normal speech, and therefore unnaturally strong long-distance dependencies exist in the speech tokens. By modeling the TPs at increasingly long distances, the relative statistical contribution of the short-term gaps between the words in the segmented condition become too small to affect the preference of word tokens in the testing phase.

This suggests that if human responses in the task are based on acoustic level patterning, it may be the case that the human auditory system is not able to capture dependencies at extended temporal distances. This is closely related to the study of Newport and Aslin (2004) who found that adult listeners are unable to learn dependencies between non-adjacent syllables whereas dependencies between non-adjacent segments (either vowels or consonants) were readily learned when familiarized with continuous stream of artificial language. The inability to learn non-adjacent

syllabic dependencies could be also explained by the finite length temporal integration in the auditory processing. Segmental dependencies with an interleaved random segment in between could be readily captured by a system modeling statistical dependencies up to, e.g., 150 ms, but dependencies across multiple syllables may simply be too distant to be captured by such short-term analysis.

Note that the inability to capture acoustic dependencies at longer temporal distances does not imply that long-range linguistic dependencies would not exist or could not be captured by a distributional learning mechanism. It is well known that such dependencies do exist. However, the huge variability and dimensionality of the acoustic space strongly points towards the necessity of an intermediate representation upon which further analysis and learning can take place. Given the current knowledge of human speech perception, it is early to say whether these units are phones, syllables, morphemes or something else (see Räsänen, 2011), and whether the computations are distributional or structural in nature. The current study does not exclude the possibility that the human listeners are directly utilizing syllable level TPs in the artificial language learning task, but simply shows that the TP analysis at the acoustic level can also explain behavioral observations to a large degree.

## Acknowledgements

This research was financially supported by Nokia NRC.

## References

- Endress, A. D., & Bonatti, L. L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, 105(2), 247-299.
- Laakso, A., & Calvo, P. (2011). How Many Mechanisms Are Needed to Analyze Speech? A Connectionist Simulation of Structural Rule Learning in Artificial Language Acquisition. *Cognitive Science*, 35, 1243-1281.
- Li, W. (1990). Mutual Information Functions versus Correlation Functions. *J. Statistical Physics*, 60, 823-837.
- Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48, 127-162.
- Peña, M., Bonatti, L. L., Nespore, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, 298(5593), 604-607.
- Rasilo, H., Räsänen, O., & Laine, U. (In preparation). An approach to language acquisition of a virtual child: learning based on feedback and imitation by caregiver.
- Räsänen, O. (2011). A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events. *Cognition*, 120, 149-176.
- Räsänen, O., & Laine, U. (2012). A method for noise-robust context-aware pattern discovery and recognition from categorical sequences. *Pattern Recognition*, 45, 606-616.
- Saffran, J., Aslin, R., & Newport, E. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, 274, 1926-1928.



# Optimally Designing Games for Cognitive Science Research

Anna N. Rafferty (rafferty@cs.berkeley.edu)

Matei Zaharia (matei@cs.berkeley.edu)

Computer Science Division, University of California, Berkeley, CA 94720 USA

Thomas L. Griffiths (tom\_griffiths@berkeley.edu)

Department of Psychology, University of California, Berkeley, CA 94720 USA

## Abstract

Collecting cognitive science data using games has the potential to be a powerful tool for recruiting participants and increasing their motivation. However, designing games that provide useful data is a difficult task that often requires significant trial and error. In this work, we consider how to apply ideas from optimal experiment design to designing games for cognitive science experiments. We use Markov decision processes to model players' actions within a game, and then make inferences about the parameters of a cognitive model from these actions. We present a general framework for finding games with high expected information gain based on this approach. We apply this framework to Boolean concept learning, inferring the difficulty of Boolean concepts from participants' behavior. We show that using games with higher expected information gain allows us to make this inference more efficiently.

**Keywords:** optimal experiment design; Markov decision process; computer games; concept learning

## Introduction

Computer games have become increasingly popular tools for gathering psychological data and for educational purposes (e.g., Michael & Chen, 2005; Von Ahn, 2006; Siorpaes & Hepp, 2008; Klopfer, 2008), providing a way to recruit large numbers of motivated participants. However, creating games that actually result in useful data requires significant engineering, normally based on trial and error. In this paper we propose a method for automating the process of designing games, using ideas from optimal experiment design.

The key problem in designing games for cognitive science research is finding a game that provides as much information as possible about the research question being addressed. For traditional experiments, the field of optimal experiment design seeks to choose the design that will give the most information about the dependent variable (Atkinson, Donev, & Tobias, 2007). We adapt this method to identify the game that will give the most information about the parameters of a cognitive model. By automating the process of game design, we limit the trial and error necessary to find a game that will provide useful data, while still reaping the benefits of using games rather than traditional experiments.

Adapting optimal experiment design methods to game design requires predicting people's behavior within games, which may differ from behavior in traditional experiments. For instance, in a categorization experiment a participant's response to a stimulus may roughly correspond to whether she believes the stimulus is in the category. However, in a game this relationship is complicated by competing incentives. For example, available actions in a game may be contingent on

the success of previous actions, leading to complex strategies. We propose using Markov decision processes (MDPs) to predict people's actions in games. MDPs incorporate the current and future benefit of an action, and thus allow us to take into account the incentive structures and rules of a particular game. By combining MDPs with ideas from optimal experiment design, we create a framework for finding the game that will provide the highest expected information gain.

The plan of the paper is as follows. The next section provides background on optimal experiment design and MDPs. We then show how to combine these ideas in a framework for optimal game design. The remainder of the paper applies this general framework to the specific case of learning Boolean concepts, illustrating the benefits of optimal game design. We introduce a novel concept learning game, and use our approach to optimize the game parameters. Two behavioral experiments show that our optimized game results in more efficient estimation of the difficulty of learning different kinds of Boolean concepts, and that the actual amount of information obtained from players is positively correlated with the expected information gain of the game.

## Background

The optimal game design framework we propose relies on ideas from Bayesian experiment design and Markov decision processes, which we will introduce in turn.

### Bayesian Experiment Design

Bayesian experiment design, a subfield of optimal experiment design, seeks to choose the experiment that will maximize the expected information gain about a parameter  $\theta$  (Atkinson et al., 2007; Chaloner & Verdinelli, 1995). In cognitive science, this procedure and its variations have been used to design more informative experiments that allow for clearer discrimination between alternative hypotheses (Myung & Pitt, 2009). Throughout this paper, let  $\xi$  be an experiment (or game) design and  $y$  be the data collected in the experiment. Then the Bayesian experimental design procedure is as follows:

$$\begin{aligned} \text{maximize } U(\xi) &= \int p(y|\xi)U(y, \xi)dy \\ \text{where } p(y|\xi) &= \int p(y|\xi, \theta)p(\theta)d\theta \\ \text{and } U(y, \xi) &= \int (H(p(\theta|y, \xi)) - H(p(\theta)))d\theta, \quad (1) \end{aligned}$$

where  $H(p)$  is the Shannon entropy of a probability distribution  $p$ , defined as  $H(p) = \int p(x)\log(p(x))dx$ . Thus, the

procedure maximizes the expected utility of an experiment  $\xi$ , defined as the information gain over all outcomes  $y$  weighted by their probabilities  $p(y|\xi)$  under the current prior.

### Markov Decision Processes

The Bayesian experiment design procedure uses  $p(\theta|y, \xi)$  to calculate the information gain from an experiment. This quantity represents the impact that the data  $y$  collected from experiment  $\xi$  have on the parameter  $\theta$ . In a game, the data  $y$  are a series of actions, and to calculate  $p(\theta|y, \xi)$ , we must interpret how  $\theta$  affects those actions. Via Bayes' rule, we know  $p(\theta|y, \xi) \propto p(y|\theta, \xi)p(\theta)$ . We thus want to calculate  $p(y|\theta, \xi)$ , the probability of taking actions  $y$  given a particular value for  $\theta$  and a game  $\xi$ . To do so, we turn to Markov decision processes (MDPs), which provide a natural way to model sequential actions. MDPs and reinforcement learning have been used previously in game design for predicting player actions and adapting game difficulty (Erev & Roth, 1998; Andrade, Ramalho, Santana, & Corruble, 2005; Tan & Cheng, 2009).

MDPs describe the relation between an agent's actions and the state of the world and provide a framework for defining the value of taking one action versus another (see Sutton & Barto, 1998). Formally, an MDP is a tuple  $\langle S, A, T, R, \gamma \rangle$ , where  $S$  is the set of possible states and  $A$  is the set of actions that the agent may take. The transition model  $T$  gives the probability  $p(s'|s, a)$  that the state will change to  $s'$  given that the current state is  $s$  and the agent takes action  $a$ . The reward model  $R(s, a, s')$  describes the probability of receiving a reward  $r \in \mathbb{R}$  given that action  $a$  is taken in state  $s$  and the resulting state is  $s'$ . Finally, the discount factor  $\gamma$  represents the relative value of immediate versus future rewards. The value of taking action  $a$  in state  $s$  is defined as the expected sum of discounted rewards and is known as the Q-value:

$$Q(s, a) = \sum_{s'} p(s'|s, a) \left( R(s, a, s') + \gamma \sum_{a' \in A} p(a'|s') Q(s', a') \right), \quad (2)$$

where  $p(a'|s')$  is the probability that an agent will take action  $a'$  in state  $s'$  and is defined by the agent's policy  $\pi$ . We assume that people's actions can be modeled as a Boltzmann policy, as in Baker, Saxe, and Tenenbaum (2009):

$$p(a|s) \propto \exp(\beta Q(s, a)), \quad (3)$$

where higher values of  $\beta$  mean the agent is more likely to choose the best action, while  $\beta = 0$  results in random actions.

### Optimal Game Design

We can now define a procedure for optimal game design, identifying the game with maximum expected information gain for  $\theta$ . We assume there is an existing game design with parameters to adjust, corresponding to point values, locations of items, or any other factor that can be varied. To apply Bayesian experiment design to choosing a game, we define the utility of a game  $\xi$  as the expectation of information gain over the true value of  $\theta$  and the actions chosen by the players:

$$U(\xi) = E_{p(\theta, \mathbf{a})} [H(p(\theta)) - H(p(\theta|\mathbf{a}, \xi))], \quad (4)$$

where  $\mathbf{a}$  is the set of action vectors for all players. The expectation is approximated by sampling  $\theta$  from the prior  $p(\theta)$ , and then simulating players' actions given  $\theta$  by calculating the Q-values for the MDP and sampling from Equation 3.

The remaining quantity in Equation 4 is  $p(\theta|\mathbf{a}, \xi)$ . Intuitively, this quantity connects actions taken in the game with the parameter of the cognitive model that we seek to infer,  $\theta$ . For a game to yield useful information, it must be the case that people will take different actions for different values of  $\theta$ . Concretely, we expect that players' beliefs about the reward model and the transition model may differ based on  $\theta$ . For instance, in a categorization task with two objects  $A$  and  $B$ ,  $\theta$  might determine the probability that  $A$  is a positive instance and  $B$  is a negative instance of the category. If taking a particular action leads to positive rewards only when a positive instance is observed, then we would expect that the value of  $\theta$  is large if many players take that action when observing  $A$ .

The process of inferring  $\theta$  from actions assumes that each  $\theta$  corresponds to a particular MDP. If this is the case, we can calculate a distribution over values of  $\theta$  based on the observed sequences of actions  $\mathbf{a}$  of all players in the game  $\xi$ :

$$p(\theta|\mathbf{a}, \xi) \propto p(\theta)p(\mathbf{a}|\theta, \xi) \quad (5)$$

$$= p(\theta)p(\mathbf{a}|\text{MDP}_\theta, \xi) \quad (6)$$

$$= p(\theta) \prod_i p(\mathbf{a}_i|\text{MDP}_\theta, \xi), \quad (7)$$

where  $\mathbf{a}_i$  is the vector of actions taken by player  $i$  and  $\text{MDP}_\theta$  is the MDP derived for the game based on the parameter  $\theta$ . Calculating this distribution can be done exactly if there is a fixed set of possible  $\theta$  or by using Markov chain Monte Carlo (MCMC) methods if the set of  $\theta$  is large or infinite (see Gilks, Richardson, & Spiegelhalter, 1996).

Now that we have defined  $p(\theta|\mathbf{a}, \xi)$ , we can use this to find the utility of a game. Equation 4 shows that this calculation follows simply if we can calculate the entropy of the inferred distribution. In the case of a fixed set of possible  $\theta$ ,  $H(p(\theta|\mathbf{a}, \xi))$  can be calculated directly. If MCMC is used, one must first infer a known distribution from the samples and then take the entropy of that distribution. For example, if  $\theta$  is a multinomial and  $p(\theta)$  is a Dirichlet distribution, one might infer the most likely Dirichlet distribution from the samples and find the entropy of that distribution.

We have now shown how to (approximately) calculate  $U(\xi)$ . To complete the procedure for optimal game design, any optimization algorithm that can search through the space of games is sufficient. Maximizing over possible games is unlikely to have a closed form solution, but stochastic search methods can be used to find an approximate solution to the maximum utility game. For example, one might use simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983). This method allows optimization of both discrete and continuous parameters, where neighboring states of current game are formed by perturbations of the parameters to be optimized.

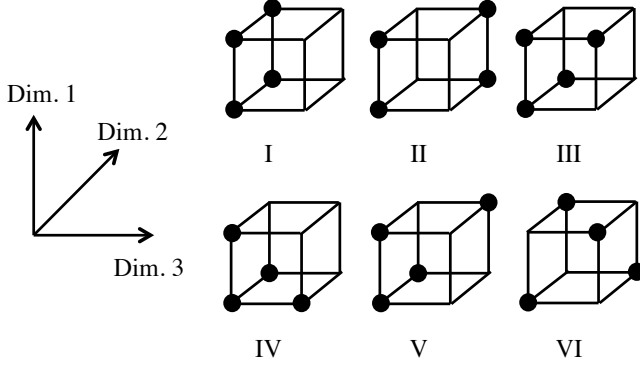


Figure 1: Boolean concept structures. In each structure, eight objects differing in three binary dimensions are grouped into two categories of four elements. Each object is represented as a corner of a cube based on its combination of features, and the objects chosen for one category in each problem type are represented by dots.

## Optimal Games for Boolean Concept Learning

We have described a general framework for automatically finding games that are potentially highly informative about model parameters. To test this framework, we applied it to a particular question: What is the relative difficulty of various Boolean concept structures? This question has been studied in past work (e.g. Shepard, Hovland, & Jenkins, 1961; Griffiths, Christian, & Kalish, 2008), so we can compare our results to those produced using more traditional methods. We first describe Boolean concept learning, and then turn to the game we created and the application of optimal game design.

### Boolean Concepts

In Boolean concept learning, one must learn how to categorize objects that differ along several binary dimensions. We focus on the Boolean concepts explored in Shepard et al. (1961). In these concepts, there are three feature dimensions, resulting in  $2^3$  possible objects, and each concept contains four objects. This results in a total of 70 concepts with six distinct structures, as shown in Figure 1. Shepard et al. (1961) found that the six concept structures differed in learning difficulty, with a partial ordering from easiest to most difficult of  $I > II > \{III, IV, V\} > VI$ . Similar results were observed in later work (Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Feldman, 2000) although the position of Type VI in the ordering can vary (Griffiths et al., 2008).

To model learning of Boolean concepts, we assume learners' beliefs about the correct concept  $h$  can be captured by Bayes' rule (Griffiths et al., 2008):

$$p(h|\mathbf{d}) \propto p(h)p(\mathbf{d}|h) \quad (8)$$

$$= p(h) \prod_{d \in \mathbf{d}} p(d|h), \quad (9)$$

where each  $d \in \mathbf{d}$  is an observed stimulus and its classification, and observations are independent given the category.

The likelihood  $p(d|h)$  is then a simple indicator function:

$$p(d|h) \propto \begin{cases} 1 & \text{if } h \vdash d \\ 0 & \text{otherwise} \end{cases}, \quad (10)$$

where  $h \vdash d$  if the stimulus classification represented by  $d$  matches the classification of that stimulus in hypothesis  $h$ . We seek to infer the prior  $p(h)$ , which represents the difficulty of learning different concepts and thus gives an implicit ordering on structure difficulty. In our earlier terminology,  $\theta$  is a prior distribution on concepts  $p(h)$ . For simplicity, we assume all concepts with the same structure have the same prior probability, so  $\theta$  is a 6-dimensional multinomial.

### Corridor Challenge

To teach people Boolean concepts we created the game Corridor Challenge, which requires learning Boolean concepts to achieve a high score. Corridor Challenge places the player in a corridor of islands, some of which contain a treasure chest, joined by bridges (Figure 2).<sup>1</sup> The islands form a linear chain and the bridges can be crossed only once, so players cannot return to previous chests. Some chests contain treasure, while others contain traps; opening a chest with treasure increases the player's score and energy, while opening a chest with a trap decreases these values. Each chest has a symbol indicating whether it is a trap; symbols differ along three binary dimensions and are categorized as a trap based on one of the Boolean concepts. Players are shown a record of the symbols from opened chests and their meanings (see the right hand side of Figure 2). Players are told to earn the highest score possible without running out of energy, which is depleted by moving to a new island or opening a trapped chest. When a player runs out of energy, the level is lost and she cannot explore the rest of the level; surviving a level earns the player 250 points. Corridor Challenge games may consist of several levels. Each level is a new corridor with different chests, but the same symbols are used and they retain the same meaning as on the previous level. At the start of each level, the player's energy is restored, but points are retained from level to level.

### Optimizing Corridor Challenge

Applying optimal game design to Corridor Challenge requires specifying the parameters to optimize in the search for the optimal game, formulating the game as an MDP, and specifying the model for how the player's prior on concepts ( $\theta$ ) relates to the MDP parameters. The structure of Corridor Challenge allows for many variants that may differ in the expected information gain. To maximize expected information gain while keeping playing time relatively constant we limited the game to two levels, with five islands per level. We then used optimal game design to select the number of points gained for opening a treasure chest, points lost for opening a trap chest, the energy lost when moving, the symbols that

<sup>1</sup>Corridor Challenge uses freely available graphics from <http://www.lostgarden.com/2007/05/dancs-miraculously-flexible-game.html>



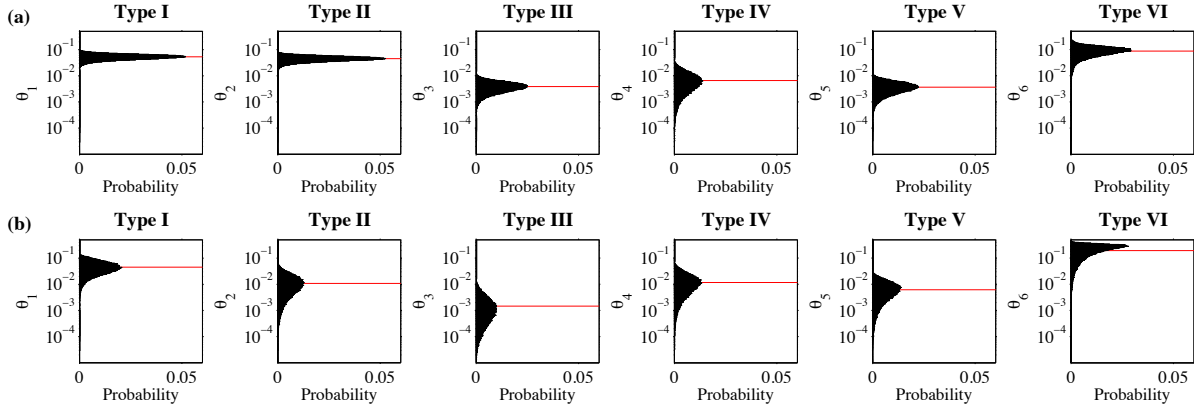


Figure 3: Results of Experiment 1, in the form of posterior distributions on concept difficulty from participants’ responses in (a) the optimized game and (b) the random game; red lines indicate the mean of each distribution. Each panel shows the distribution over the inferred difficulty of a concept with the given structure (Types I-VI), as reflected by its prior probability in the Bayesian model. Concepts with higher prior probability are easier to learn. Note the logarithmic scale.

high information gain had true concepts with different structures. While the information gain found for any given game is approximate, since we estimated the expectation over possible priors with only 35 samples, this was sufficient to separate poor games from relatively good games; we explore this relationship further in Experiment 2.

## Experiment Methods

**Participants.** A total of 50 participants were recruited online and received a small amount of money for their time.

**Stimuli.** Participants played Corridor Challenge with parameters set based either on an optimized game (expected information gain of 3.4 bits) or on a random game (expected information gain of 0.6 bits). The symbols differed along the dimensions of shape, color, and pattern.

**Procedure.** Half of the participants were randomly assigned to each game design, and played the game in a web browser. The participants were shown text describing the structure of the game, and then played several practice games to familiarize them with interface. The first practice game simply had chests labeled “Good” and “Bad”; the next three games used Boolean concepts of increasing difficulty based on previous work. All practice games used different symbols from one another and from the final game. Practice games used the point and energy values from the game chosen for their condition (i.e., the random game or the game found by the search) in order to make players aware of these values, but the symbols in the practice games were identical across conditions. The final game differed by condition. After completing the final game, participants were asked to rate how fun and how difficult the game was, both on 7-point Likert scales. Additionally, they were shown the stimuli and categorization information that they observed during the final game, and asked to classify the remaining stimuli from the game that were not observed.

## Results

Figure 3 shows the inferred distribution over the prior probability of each concept ( $\theta_i$ ) based on participants’ actions

for the optimized game and the random game; if a concept has higher prior probability, it will be easier to learn. These distributions were obtained via MCMC using a Metropolis-Hastings algorithm on both the prior and the noise parameter  $\beta$ . Results show samples generated from five chains with 100,000 samples each; the first 10% of samples from each chain were removed for burn-in.

Qualitatively, the distributions inferred from the optimized game appear more tightly concentrated than those from the random game; this is confirmed by the actual information gain, which was 3.30 bits for the optimized game and 1.62 bits for the random game. This implies that we could halve the number of participants by running the optimized game.

For both games, the ordering of the mean prior probabilities of each type, shown by red lines in Figure 3, is the same as that found in previous work, except for Type VI. Our inferred distributions for Type VI placed significant probability on many values, suggesting that we simply did not gain much information about its actual difficulty. We do infer that Type VI is easier than Types III, IV, or V, which has a precedent in the results of Griffiths et al. (2008).

## Experiment 2: Estimating Information Gain

To verify the relationship between actual and expected information gain, we conducted a second experiment in which players played games with a range of information gains. In order to isolate the impact of the symbols on the chests and the true concept we fixed the point structures to those found for the optimized game in Experiment 1 and conducted new searches over the remaining variables. We then selected games from the search paths that had varying expected information gains, demonstrating that even without changing the incentive structure a range of information gains was possible.

## Methods

**Participants.** A total of 175 participants were recruited online and received the same payment as in Experiment 1.

**Stimuli.** Participants played one of seven new games.

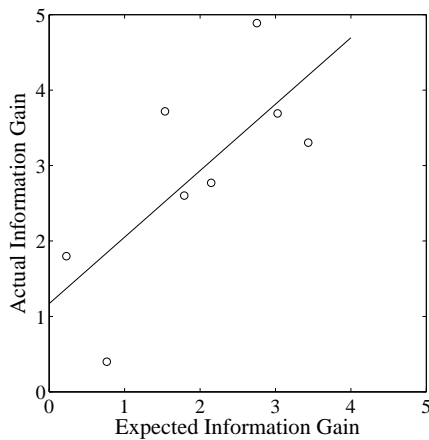


Figure 4: Results of Experiment 2, showing expected versus actual information gains ( $r = 0.72$ ). Each circle represents a game, and the least-squares regression line is shown.

**Procedure.** Procedure matched Experiment 1.

## Results

We compared the actual and expected information gains for the seven new games and the optimized game from Experiment 1, all of which used the same point structure. As shown in Figure 4, we found that expected and actual information gain were positively correlated ( $r(6) = 0.72$ ,  $p < 0.05$ ). This demonstrates that the design of the game does influence how much information we can infer from human players' actions, and that this gain is predicted by our estimates.

## Conclusion

We have presented a general framework for the optimal design of games for cognitive science experiments that adapts ideas from Bayesian experiment design. Our methodology hinges on using Markov decision processes to infer cognitively relevant information from players' actions. By combining this model with a stochastic search method, we were able to find appreciably more informative games for Boolean concept learning. Our experimental results demonstrate that using this framework to infer concept difficulty gives similar results to prior work, and that the expected information gain of a game predicts the actual information gain when the game is played by real players. One of the chief advantages of this framework is its applicability to a wide variety of scenarios, from creating other types of games that examine different psychological questions to designing optimally informative assessments within educational software.

There are a number of ways in which this initial test of a framework for designing games could be expanded. In this work, we ignored the value of information for computational tractability. However, people may consider how their future knowledge will be affected by their current actions, leading to deviations from our model's predictions; compar-

ing the fit of a partially observable Markov decision process model to the fit of an MDP model would help to determine whether this is an issue. In the optimal game design procedure, one might also want to explore other metrics than entropy for measuring a game's expected utility. For instance, one might use a loss metric that considers the distance of samples from the true value of  $\theta$ . Finally, it would be interesting to explore whether there are advantages beyond motivation for using games rather than traditional experiments. While there remain many areas for future exploration, this work gives a starting point for designing highly informative games and gives experimental support that these games can provide meaningful data for cognitive science.

**Acknowledgements.** This work was supported by a DoD NDSEG Fellowship to ANR, a Google Ph.D. Fellowship to MZ, and NSF grant number IIS-0845410 to TLG.

## References

- Andrade, G., Ramalho, G., Santana, H., & Corruble, V. (2005). Extending reinforcement learning to provide dynamic game balancing. In *Proceedings of the Workshop on Reasoning, Representation, and Learning in Computer Games, 19th IJCAI* (pp. 7–12).
- Atkinson, A., Donev, A., & Tobias, R. (2007). *Optimum experimental designs, with SAS* (Vol. 34). New York: Oxford University Press.
- Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113(3), 329–349.
- Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10(3), 273–304.
- Erev, I., & Roth, A. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, 88(4), 848–881.
- Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, 407, 630–633.
- Gilks, W., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Suffolk, UK: Chapman and Hall.
- Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, 32, 68–107.
- Kirkpatrick, S., Gelatt, C., & Vecchi, M. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671–680.
- Klopfer, E. (2008). *Augmented learning: Research and design of mobile educational games*. Cambridge, MA: The MIT Press.
- Michael, D., & Chen, S. (2005). *Serious games: Games that educate, train, and inform*. Boston, MA: Thomson Course Technology.
- Myung, J., & Pitt, M. (2009). Optimal experimental design for model discrimination. *Psychological Review*, 116(3), 499.
- Nosofsky, R. M., Gluck, M., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, 22, 352–369.
- Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, 75. (13, Whole No. 517)
- Siorpaes, K., & Hepp, M. (2008). Games with a purpose for the semantic web. *Intelligent Systems*, 23(3), 50–60.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.
- Tan, C., & Cheng, H. (2009). IMPLANT: An integrated MDP and POMDP learning agent for adaptive games. In *Proceedings of The Artificial Intelligence and Interactive Digital Entertainment Conference* (pp. 94–99).
- Von Ahn, L. (2006). Games with a purpose. *Computer*, 39(6), 92–94.

# Cognitive Workload and the Motor Component of Visual Attention

Jason Ralph, Wayne D. Gray, & Michael J. Schoelles  
Rensselaer Polytechnic Institute

## Abstract

Outside the laboratory, the ability to control visual input during multiple task performance by controlling *where the eyes look and when* is an obvious component of multiple task performance. However, inside the laboratory researchers either obviate the control of the eyes by presenting information from one task at a time or are oblivious to the need for *just-in-time* control of the motor component of visual attention. We investigate the effects of cognitive workload on eye movements in a paradigm that controls the demand on the eyes as an input channel while increasing workload by increasing the demand on working memory. Despite constant visual demands, we find that fixations become more scattered with increasing working memory load.

**Keywords:** dual mechanisms of control, cognitive control, cognitive workload

## Introduction

Our ability to switch among multiple tasks has been the subject of extensive research for many decades (Allport, Styles, & Hsieh, 1994; Altmann & Gray, 2008; Broadbent, 1952; Cherry, 1953). People generally exhibit a multitask effect in which they are slower and commit more errors when performing multiple concurrent tasks than a sequential series of the same tasks. Most theories of multitasking propose a resource capacity explanation for this multi-task effect. These theories, including Wicken's multiple resource theory (MRT) (Wickens, 2002; Wickens & Colcombe, 2007), and Salvucci and Taatgen's threaded cognition (Salvucci & Taatgen, 2008; Salvucci & Taatgen, 2011), propose a variety of resources which must be shared by concurrent tasks. They argue that when the capacity of one of these resources is exhausted, the cognitive system must wait for the bottleneck to clear, which causes slower task execution.

MRT identifies input modalities (visual or auditory), response modalities (motor, vocal), and cognition as resources responsible for multi-task effects (Wickens, 2002; Wickens, 1992; Wickens, Goh, Helleberg, Horrey, & Talleur, 2003). For instance, if two tasks require visual processing, performance will be slower because the capacity of the visual processor is limited. MRT has not been instantiated in computational form but its assumptions appear compatible with other capacity theories such that multiple tasks can be performed without interference until some capacity limit is reached (e.g., see Just, Carpenter, & Hemphill, 1996a, 1996b).

Threaded cognition is a model based procedural theory

of task switching, which is implemented within the ACT-R cognitive architecture (Anderson et al., 2004). In this theory, "concurrent multitasking emerges from the interaction of autonomous process threads in conjunction with a straightforward mechanism for resource acquisition and conflict resolution," (Salvucci & Taatgen, 2008, p. 102). Although we see Threaded Cognition as a major advance in the modeling of complex cognition, at present it focuses on the control of task switching per se, not on the control of tasks.

As detailed by the Dual Mechanisms of Control (DMC) theory (Braver, Gray, & Burgess, 2007; Braver, 2012), the brain's ability to prepare ahead of time (proactive control) is limited. Proactive control requires that the strategy and information for a task be kept active. Hence, proactive control places strategies such as the verbal-articulatory loop (Baddeley, 2012) and updating, shifting, and inhibition (Miyake et al., 2000) under control of the PFC and subject to limits in PFC processing. The importance of proactive control depends on the nature of the task being performed. In situations with a limited number of stimuli and responses (i.e., most experimental psychology tasks), then proactive control seems best. However, in tasks that require the subject to respond to one of a number of different stimuli in a number of different ways (e.g., driving on the freeway during rush hour) then reactive processing seems required. Cognitive control is recruited by a mix of proactive and reactive influences, and understanding how this mix changes based on task demands is key to understanding performance of multiple tasks.

We hypothesize that differences between single-task and multi-task performance depend on how the brain manages the demands on proactive control. A major limiting factor in multitasking is our ability to distribute our attention to perform multiple concurrent tasks. In visual tasks, the capacity of the eyes to focus on one part of the visual scene at a time is a factor constraining our performance. Many tasks require the ability to control eye movements to capture required information from the world, or to maintain sustained focal attention.

The ability to control our eyes during task performance is an overlooked but critical variety of control. Unless we assume that pointing-the-eyes at potentially informative areas of visual interest is somehow both automatic and effortless, then the need to control gaze to optimize task performance adds yet another burden to our control mechanisms. Hence, it may be possible to determine when a switch in



control mode has occurred by examining patterns of eye fixations. We hypothesize that effective proactive control results in long, constant fixations on the area of the visual scene related to a task. As the demands on proactive control increase, eye movements should become more scattered, evidenced by short fixation durations spread across a wider area. After presenting our paradigm and results, we return to speculative discussion of this issue in our Discussion section.

## Previous Research

In prior research, we found that performance in a dual-task paradigm was affected by requiring subjects to look in different locations for information needed for different tasks (Ralph, Gray, & Schoelles, 2010). Subjects performed a continuous visual tracking task while concurrently performing an n-back style memory task. (The paradigm will be explained in detail in the methods section below.) Consistent with an MRT prediction, subjects who received information aurally outperformed those who received it visually. However, we could not determine whether auditory instruction relieved the cognitive demands of the task, or simply the visual demands.

The DMC account implies that proactive control of behavior relies on a common set of PFC mechanisms. Viewing the eye as something that needs to be controlled, we hypothesize that it is the extra control of the eye that contributes to the auditory vs visual tasks differences, not necessarily the differences in auditory vs visual processing (as implied by MRT).

To test this hypothesis, the current study compares two visual conditions which share the same visual demands, but differ in the cognitive (working memory) requirements. This differs from most MRT research, which typically increases cognitive load by increasing the difficulty of visual tracking (Wickens et al., 2003). In the current study, differences in eye movement behavior and performance cannot be attributed to increased demands on the eye as these demands are identical across conditions. Thus the focus of the current study is on how increased demands for proactive control affect eye movement strategies and performance in the absence of additional demands for the control of the eye.

## The Study

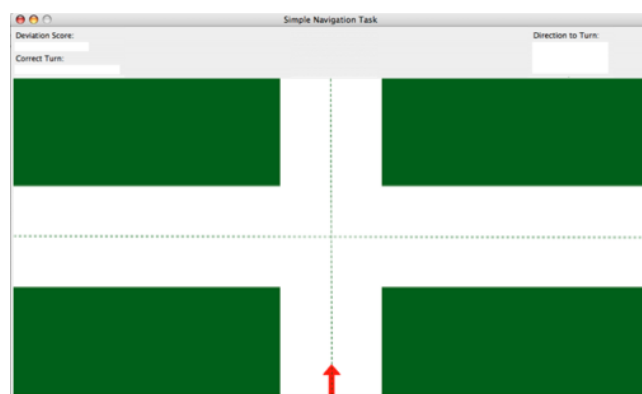
The NavBack paradigm was designed to collect detailed empirical data in a task at the approximate complexity of those used by Wickens in testing MRT (e.g., Martin-Emerson & Wickens, 1992, 1997). The NavBack task combines a tracking task with a working memory task. The tracking task requires the subject to keep an arrow centered as it *jitters* (i.e., moves randomly) from side to side. Concurrently, subjects perform a continuous working memory span task, which is similar to the n-back memory task (Gevins & Smith, 2003; Jaeggi et al., 2010; McEvoy, Smith, & Gevins, 1998). In the high workload conditions, subjects must maintain a

list of instructions in memory of how to *turn* in the next three intersections (e.g., “left”, “right”, or “forward”). After each intersection the subject has to delete the just completed instruction and add a new instruction to the end of his mental list. In the low workload condition, only one instruction must be maintained.

## Methods

### Subjects

22 undergraduate students of Rensselaer Polytechnic Institute (mean age = 19) volunteered to participate in this study for course credit. Eleven subjects were randomly assigned to each of two conditions: *High Memory* or *Low Memory*.



*Figure 1.* Screenshot of the NavBack Paradigm. Subjects must keep the arrow in the center of the road as they receive turn instructions in the box at the upper right corner. The arrow remained at the bottom of the screen, and the “road” scrolled downward, simulating forward movement. Subjects could turn when the arrow was in the intersection.

### Apparatus and Materials

The experiment was run on an Apple Mac Mini computer (running Mac-OS 10.4) at a 1024x768 screen resolution. Eye fixation data were collected using an LC Technologies tracker at a 120 Hz sampling rate. A chin-rest was used to stabilize head movements and ensure a fixed viewing distance of 60 cm. The NavBack software is a custom application implemented in Lispworks 5.1.

### Design

**Between-Subject Condition: Memory Load.** Each memory load condition received a new direction (the turn to make in the future) while traveling through a “city block”. For the *High Memory* condition, the new instruction specified how to turn three intersections in the future. In the *Low Memory* condition, the new instruction specified how to turn at the next intersection. Hence, the memory load for the High vs Low Memory conditions was three versus one items.

### Within-Subject Condition: Instruction Presentation

**Time.** All subjects received the new turn instruction *early*, *middle*, or *late* during their *travel* through a city block. The instruction appeared for 2-s beginning either 1-s, 3.4-s, or 5-s after the arrow exited the intersection and entered the next city block. On any given episode cycle, whether the instruction was presented early, middle, or late was determined randomly.

### Procedure

Each subject completed a 2-min practice session to familiarize them with the demands of the task, followed by eight 5-min experimental blocks. Each 5-min block consisted of a continuous series of episode cycles. Each cycle began when the tip of the arrow left an intersection and entered the next city block (city blocks are the green areas in Figure 1). At one of three randomly chosen times a new instruction appeared in the direction box on the upper right of the screen (see Figure 1). Travel time through each city block was 6-s.

Although, once in the intersection, subjects could turn at any time, minimizing the jitter score required the subject to turn at the exact center of the intersection. The animation for the turn added 1,500 ms to the time spent in the intersection when subjects made left or right turns. During each episode cycle subjects had to do two related tasks: the jitter task and the turn direction task. The jitter task is a visual-motor task requiring constant attention. The turn-direction task requires monitoring for the appearance of a new turn direction while performing the jitter task. Depending on condition subjects needed to hold either one or three turns in memory. Each new episode cycle required them to update the list of items held in memory.

### Jitter Task: Visual-Motor

Subjects were instructed to keep the arrow in the center of the road (on the dotted line in Figure 1) as the arrow *jittered* actively from side to side, every 200 ms, based on a pseudo-random function. Subjects corrected the arrow's horizontal position by pressing the *a* and *d* keys on a standard keyboard. Their goal was to keep the arrow as close to the center of the lane as possible. The arrow's position at the beginning of each city block was determined by the timing of the previous turn. If the previous turn (left or right) was initiated at the exact vertical center of the intersection, then the arrow began the next city block in the center of the lane. If the turn was initiated early or late, it began the next city block deviated from the center by an amount proportional to the distance from the turn point to the center of the intersection. (Ss were not instructed on this aspect of the task.) The computer logged the absolute value of the number of pixels deviated from the center every 200 ms. We refer to this value as "jitter score".

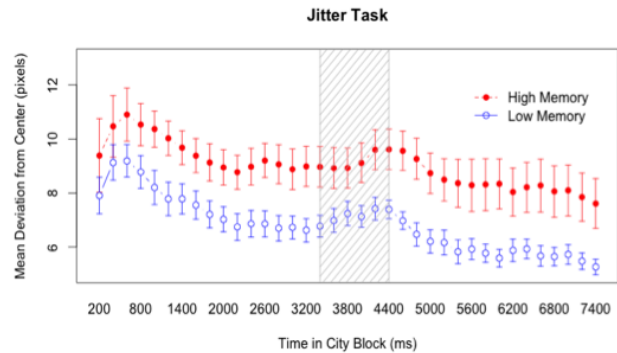


Figure 2. Mean jitter scores for middle instruction city blocks. The gray rectangle denotes when the instruction was on-screen.

### Turn Direction Task: Working Memory Updating

Concurrent with the Jitter Task, subjects were required to keep either one or three turns in memory (e.g., left, right, forward). At the beginning of a 5-min set, high memory subjects were presented with the initial three turns that were to be made in the first, second, and third intersection. After the set began, new turn directions appeared at one of the three times described earlier. The direction presentation time was randomly chosen for each city block.

As discussed previously, success in the high memory condition essentially required subjects to rehearse, update, and maintain a list of three instructions. Subjects had to mentally delete the instruction for the just completed turn and to append a new instruction at the end of their mental list. Subjects in the low memory group had to remember only the most recently presented direction. Feedback on the correctness of the most recent turn was available at the top left of the screen (see Figure 1).

### Results

**Jitter Scores.** An analysis of variance was performed on jitter scores by memory load and instruction time and revealed main effects of memory load  $F(1,20)=7.96$   $p<.02$ , and instruction time  $F(2,40) = 17.31$   $p<.001$ . High memory subjects ( $M=10.42$  pixels  $SD=2.03$ ) had higher mean jitter scores than low memory subjects ( $M=8.38$  pixels,  $SD=1.05$ ). Figure 2 shows the mean jitter scores throughout an average city block (for middle instruction times). Low memory subjects outperformed high memory subjects throughout the city block.

**Turning Task.** The turn results were as expected with the Low Workload group more accurate (94.64%) than the High Workload group (85.31%). These differences were significant ( $F(1,20) = 12.57$   $p<.01$ ) but will not be discussed further in this short report.

**Fixation Locations.** Analysis of the eye data was performed using the areas of interest displayed in Figure 3. Only data recorded while the arrow was within a city block is included (e.g. not in an intersection). The data yielded several differences in fixations between the high vs low memory conditions. Figure 4 shows the proportion of time spent looking at the arrow, direction box, and road areas. High memory subjects spend less time fixated on the arrow and more time looking at the direction box throughout the city block. This difference is most pronounced before the instruction appears (compare Figure 5 top and bottom for those times before the appearance of the instruction), suggesting that low memory subjects employed a more economical strategy to monitor the direction box for the appearance of the arrow. As Figures 4 and 5 demonstrate, compared to low memory subjects, high memory ones devoted proportionately more time looking at the direction box and less time looking at the arrow<sup>1</sup>.

The fixation pattern for low memory subjects likely reflects a proactive strategy designed to maximize time spent fixating on the arrow while maintaining the ability to monitor for instructions. It is a strategy that high memory subjects were either unwilling or unable to pursue.

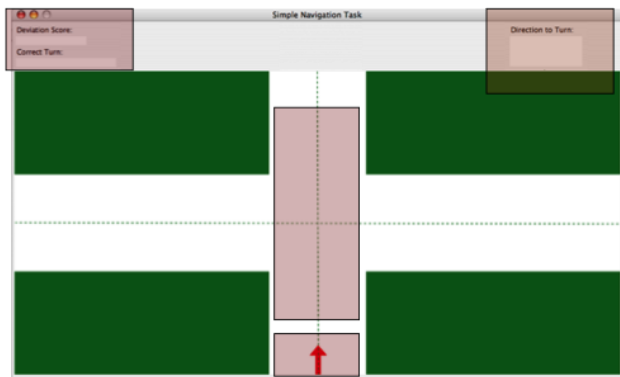


Figure 3. Areas of Interest used for the analysis of eye fixations. Shaded areas represent the arrow, direction box (upper right), feedback (upper left), and road areas.

#### Eye Movements Reveal Task Priority Differences with Workload.

Figures 4 and 5 limit the data they present to the four task relevant areas shown in Figure 3. We find it both interesting and important that the two conditions spent different proportions of time gazing at screen areas relevant to different parts of the task. The low workload condition spent more time on jitter related areas (e.g. arrow), and the high memory condition spent more time looking at areas relevant to the turning task (e.g direction box, road).

We further examined the degree of task focus by conducting an analysis of the variability of eye fixations throughout the screen area for the two conditions (see Figure 6). For this analysis, we split the screen into 30x30 pixel boxes and computed the standard deviation of the number of fixations

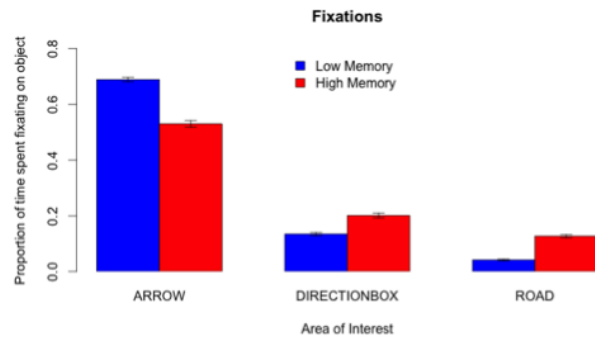


Figure 4. Time spent fixating on the arrow, direction box and road areas, respectively. High Memory subjects spent more time looking at the road and direction box.

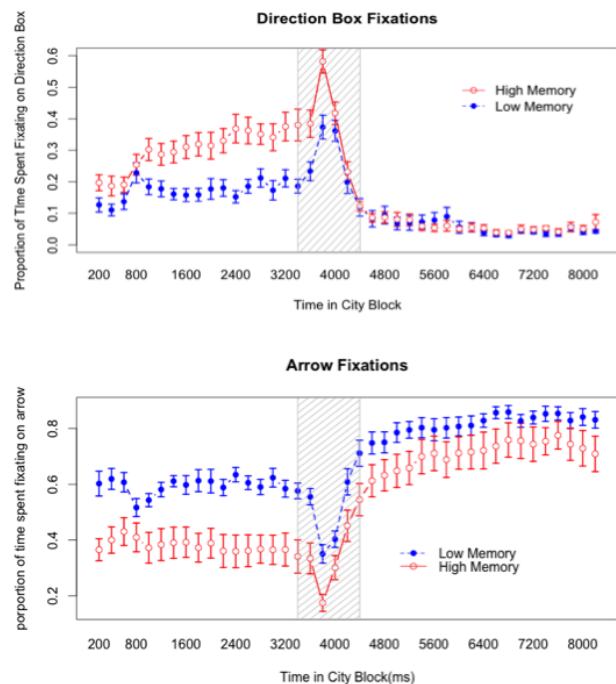
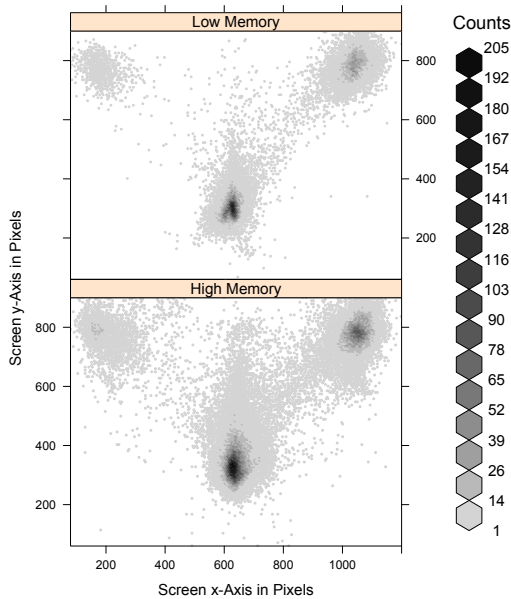


Figure 5. Proportion of time spent fixating on the direction box (top) and arrow (bottom) for middle instruction city blocks. The gray area represents the time period when the instruction is on-screen.

per box for each subject. By this measure, increasing standard deviations show that subjects focus more on some of the pixel boxes than others. The lower the standard deviation, the more evenly spread the fixations are across all pixel boxes. An analysis of variance performed on this measure shows that the fixations of high memory subjects were significantly

<sup>1</sup>Less than 4% of fixations for *high memory* and 3% of fixations for *low memory* subjects were to the *feedback* area with large variations between-Ss within the same group. A one-way ANOVA on that area was not significant.

more evenly distributed (i.e., more scattered) than those of low memory subjects  $F(1,20)=8.31$   $p<.01$ . But, when separated into the relevant areas of interest, individual ANOVAs showed that high memory subjects were more scattered in the *arrow* area  $F(1,20) = 4.18$   $p<.05$ , but not in the *direction box* or *road* areas.



**Figure 6.** Density plot of eye fixations on screen for Low Memory (top) and High Memory (bottom) groups. Scale on the right shows the density count for the number of fixations centered at each pixel of the screen. Note that despite the same nominal task requirements for where-to-look-when the High Memory group scattered their fixations more across the entire screen than did the Low Memory group.

## Discussion

As the jitter and memory tasks do not share any obvious resource conflicts, it is unlikely that resource based theories of multitasking (such as MRT) would predict the effects of memory load on jitter task performance. Likewise, as there are only two tasks being interleaved, it is also unlikely that Threaded Cognition would predict differences in how they are controlled. Yet our results show a significant difference in *inattention* to task relevant areas between conditions.

A general theory of cognitive control must explain the control required to switch among multiple tasks as well as the control required to perform each single task. For many tasks, a bridge between these two types of control may lie in understanding the role of eye movements. In our study, low memory subjects performed better on the jitter task than high memory subjects throughout the average city block. In addition, our microanalysis revealed that visual attention on

the arrow is a very important factor in predicting jitter performance. Jitter performance in both conditions suffered in the time period before instructions were presented (as per Figure 2), as visual attention had to be split between the arrow and direction box.

Eye fixation data suggested major differences in the allocation of control caused by memory load. Low memory subjects were able to remain fixated on the arrow for much longer periods of time than high memory subjects. Fixations to the direction box increased until an instruction appeared for high memory subjects, but remained relatively constant for low memory subjects<sup>2</sup>.

The demands on the proactive working memory system had an impact on the fixation strategy employed. The data suggest that high memory fixations reflect a different proactive strategy meant to focus on the turn instruction at the cost of a loss of focus on the arrow. Despite this difference in strategy, our variability measure showed that memory load also caused the high memory condition to lack focus on the arrow *while they were looking at it*. It can be argued that Increased memory load indirectly affected jitter performance by causing a lack of task focus. Task that require constant focus for optimal performance (like our jitter task) will be particularly susceptible to this effect.

In currently planned studies, we hypothesize that these results are not a special case, but rather a demonstration of how cognitive control affects the use of all the brain's resources. We can no longer assume, as Threaded Cognition does, that task performance is only subject to the control required to switch tasks. We must also consider the effects that overall control demands have on how we perform each task. Recent research points to differences in the use of proactive control in affecting behavior in many tasks, including the attentional blink (Taatgen, Juvina, Schipper, Borst, & Martens, 2009), AX-CPT (Braver et al., 2007), and n-back (Szmales, Verbruggen, Vandierendonck, & Kemps, 2011) tasks. Our future research will focus on the mode of control (e.g. proactive/reactive) to help us explain these complex multitask effects that do not fit into a strict resource (MRT) framework.

## Acknowledgements

Research reported in this paper has been supported, in part, by grant N000141010019 to Wayne Gray from the Office of Naval Research, Dr. Ray Perez, Project Officer. Julia Van Cleve and Matt Nebel assisted in data collection.

<sup>2</sup>We remind the reader that the data presented in Figure 5 is for those trials on which Ss received their instruction in the middle of the city block. These data are representative of the early and late conditions in that those conditions also show increased fixations on the Direction Box until the instruction is presented for both Low and High load conditions with High > Low.

## References

- Allport, A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: exploring the dynamic control of tasks. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance iv* (pp. 421–452). Cambridge, MA: MIT Press.
- Altmann, E. M., & Gray, W. D. (2008, July). An integrated model of cognitive control in task switching. *Psychological review*, 115(3), 602–39. doi:10.1037/0033-295X.115.3.602
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglas, S., Lebiere, C., & Quin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036–1060.
- Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual Review of Psychology*, 63(1), 1–29.
- Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: dual mechanisms of control. In C. Conway, M. J. K. Jarrold & a. N. T. A. Miyake (Eds.), *Variation in working memory* (pp. 76–106). Oxford University Press.
- Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in Cognitive Sciences*, 16(2), 106–113. doi:10.1016/j.tics.2011.12.010
- Broadbent, D. E. (1952). Speaking and listening simultaneously. *Journal of Experimental Psychology*, 43(4), 267–273. doi:10.1037/h0058014
- Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5), 975–979. doi:10.1121/1.1907229
- Gevens, A., & Smith, M. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science*, 4(1), 113–131.
- Jaeggi, S. M., Studer-Luethi, B., Buschkuhl, M., Su, Y.-F., Jonides, J., & Perrig, W. J. (2010, November). The relationship between n-back performance and matrix reasoning — implications for training and transfer. *Intelligence*, 38(6), 625–635.
- Just, M. A., Carpenter, P. A., & Hemphill, D. D. (1996a). Constraints on processing capacity: architectural or implementational? In D. Steier & T. M. Mitchell (Eds.), *Mind matters: a tribute to allen newell* (pp. 141–178). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Just, M. A., Carpenter, P. A., & Keller, T. A. (1996b). The capacity theory of comprehension: new frontiers of evidence and arguments. *Psychological Review*, 103(4), 773–780.
- Martin-Emerson, R., & Wickens, C. D. (1992). The vertical visual field and implications for the head-up display. In *Proceedings of the 36th annual meeting of the human factors and ergonomics society* (pp. 1408–1412). Santa Monica, CA: Human Factors and Ergonomics Society.
- Martin-Emerson, R., & Wickens, C. D. (1997). Superimposition, symbology, visual attention, and the head-up display. *Human Factors*, 39(4), 581–601.
- McEvoy, L. K., Smith, M., & Gevins, A. (1998). Dynamic cortical networks of verbal and spatial working memory: effects of memory load and task practice. *Cereb. Cortex*, 7(8), 563–574.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cognitive Psychology*, 41(1), 49–100.
- Ralph, J., Gray, W., & Schoelles, M. (2010). Squeezing the balloon: analyzing the unpredictable effects of cognitive workload. In *Proceedings of the human factors and ergonomics society 54th annual meeting*: (pp. 299–303).
- Salvucci, D. D., & Taatgen, N. a. (2008, January). Threaded cognition: an integrated theory of concurrent multitasking. *Psychological review*, 115(1), 101–30.
- Salvucci, D. D., & Taatgen, N. a. (2011, April). Toward a Unified View of Cognitive Control. *Topics in Cognitive Science*, 3(2), 227–230.
- Szmalc, A., Verbruggen, F., Vandierendonck, A., & Kemps, E. (2011, February). Control of interference during working memory updating. *Journal of experimental psychology. Human perception and performance*, 37(1), 137–51.
- Taatgen, N. a., Juvina, I., Schipper, M., Borst, J. P., & Martens, S. (2009, August). Too much control can hurt: a threaded cognition model of the attentional blink. *Cognitive psychology*, 59(1), 1–29.
- Wickens, C. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3(2), 159–177.
- Wickens, C., & Colcombe, A. (2007). Dual-Task Performance Consequences of Imperfect Alerting Associated With a Cockpit Display of Traffic Information. *Human Factors*, 49(5), 839–850.
- Wickens, C. D. (1992). *Engineering psychology and human performance* (Second). New York: HarperCollins.
- Wickens, C. D., Goh, J., Helleberg, J., Horrey, W. J., & Talleur, D. a. (2003, January). Attentional models of multitask pilot performance using advanced display technology. *Human factors*, 45(3), 360–80.



# Order effects in diagnostic reasoning with four candidate hypotheses

**Felix G. Rebitschek (felix.rebitschek@uni-greifswald.de)**  
University of Greifswald, Department of Psychology

**Agnes Scholz (agnes.scholz@psychologie.tu-chemnitz.de)**  
Chemnitz University of Technology, Department of Psychology

**Franziska Bocklisch (franziska.bocklisch@psychologie.tu-chemnitz.de)**  
Chemnitz University of Technology, Department of Psychology

**Josef F. Krems (josef.krems@phil.tu-chemnitz.de)**  
Chemnitz University of Technology, Department of Psychology

**Georg Jahn (georg.jahn@uni-greifswald.de)**  
University of Greifswald, Department of Psychology

## Abstract

Sequentially observed symptoms in diagnostic reasoning have to be integrated to arrive at a final diagnosis. In our experiments employing quasi-medical problems, four sequentially presented symptoms were consistent with multiple diagnostic hypotheses. We tested whether symptom order creates biases in symptom evaluation. Early symptoms induced a bias towards the initial hypothesis even though an alternative hypothesis was equally supported. In two experiments, stepwise ratings were prompted to explicitly highlight alternative hypotheses. Explicit highlighting eliminated the bias towards the initial hypothesis if only two hypotheses competed, but the bias remained if more than two hypotheses were associated with symptoms in a sequence. Our results are consistent with process models of information integration that specify how early information can frame the processing of later information. Extending previous results obtained with fewer contending hypotheses, we show limits in impartially considering more than two hypotheses.

**Keywords:** Order Effects; Diagnostic Reasoning; Multiple Candidate Hypotheses; Construction Integration Theory

## Introduction

When humans explain observations in their environment, they apply knowledge about possible causes and the effects that each cause can bring about. Explaining observed symptoms by a diagnosis that specifies the most probable cause can be difficult for symptoms which are ambiguous and thus consistent with multiple diagnoses or inconsistent and hard to subsume under a single diagnosis (Johnson & Krems, 2001). Imagine the sequential integration of symptoms in medical diagnosis in its simplest form: You, a physician, become aware of a symptom pointing towards different possible diseases your new patient might have caught. Bit by bit you take notice of a second, third and a fourth symptom. Some of them are unspecific, others strengthen your belief in a diagnosis and weaken

alternatives, but none is decisive by itself. The order in which symptoms are encountered can influence the final diagnosis because the initial diagnostic hypotheses may affect how the subsequent symptoms are weighed and integrated (e.g., Chapman, Bergus & Elstein, 1996). If symptoms are observed in sequence, the initially encountered symptoms trigger diagnostic hypotheses (Mehlhorn, Taatgen, Lebiere, & Krems, 2011).

Sequential symptom processing towards the initial hypothesis demonstrates a confirmation bias (Nickerson, 1998), which would be overcome if all alternative diagnostic hypotheses could be considered in parallel. In previous studies such impartial symptom integration sometimes succeeded for two alternative diagnoses (McKenzie, 1998), but doubts have been raised whether more than two alternative diagnoses can be considered impartially in parallel (Dougherty & Hunter, 2003).

According to normative Bayesian information integration, the order of symptom presentation should not matter. Symptom patterns equally supportive of two alternative diagnoses should produce equal proportions of these diagnoses. However, already updating of a single hypothesis can be biased by the order, in which pieces of evidence are encountered (Wang, Johnson, & Zhang, 2006). Hogarth and Einhorn (1992) specified circumstances under which normative updating of a single belief is possible and no order effects should occur (e.g. stepwise simple evaluation of short and consistent sequences). Yet, models of sequential information integration including the belief adjustment model of Hogarth and Einhorn (1992) typically postulate a disproportionately large influence of early encountered information resulting in a *primacy effect*.

When multiple diagnostic hypotheses compete, such a strong influence of early information can take the form of a bias towards the diagnosis that is most strongly supported by the first symptom. The memory dynamic resulting in a confirmation bias in sequential symptom integration can be described in terms of the construction-integration theory of text comprehension by Kintsch (1998) (Baumann, Mehlhorn, & Bocklisch, 2007). After observing the first symptom, the first construction-integration cycle results in high activation of the candidate hypothesis most strongly supported by the first symptom. Subsequent construction-integration cycles start from this state. Thus, initial symptoms and preliminary hypotheses frame the processing of later symptoms.

Recently, HyGene (Thomas, Dougherty, Spenger, & Harbison, 2008), which models memory processes in hypothesis generation for a specified set of symptoms has been extended to capture effects of sequential symptom processing in detail (Lange, Thomas, & Davelaar, 2012). The activations of symptom representations compete in working memory as symptoms are sequentially encountered, however, framing of symptom processing by preliminary hypotheses is not yet implemented in HyGene.

Our main goal was to study framing of symptom processing by preliminary hypotheses and its effects on final diagnoses in diagnostic reasoning with multiple candidate hypotheses. So, the reported experiments examine sequential diagnostic reasoning with four candidate diagnoses. Differing from previous studies (Koehler, 1991), we will not set a single hypothesis, whose probability has to be rated. Instead, participants have to choose among four candidate hypotheses. We determine effects of symptom order by evaluating proportions of final diagnoses for ambiguous symptom sequences, which equally support alternative diagnoses.

Whereas framing by preliminary hypotheses should bias towards initial hypotheses, increasing the saliency of alternative diagnoses should decrease biased symptom processing. We examine both explicit and implicit highlighting of alternative diagnoses. Alternative diagnoses were explicitly highlighted by asking participants to rate the current support for each possible diagnosis after each symptom. This procedure constantly reminds participants of the competing hypotheses.

Implicit highlighting of alternative diagnoses was attempted by presenting inconsistent symptom sequences. Symptoms inconsistent with the initial hypothesis could increase the salience of diagnostic alternatives. In terms of support theory (Tversky & Koehler, 1994), symptoms inconsistent with the focal hypothesis that strongly suggest

specific alternatives could unpack the complement of the focal hypothesis into specified alternative diagnoses.

## Experiments

Participants were told that they should evaluate symptoms of workers in a chemical plant to determine which of four chemicals had most likely affected each worker. In all four experiments, the diagnostic reasoning tasks referred to the same knowledge about symptoms and causes, which was acquired in a learning phase. Firstly, participants learned which symptoms belonged to which symptom classes and subsequently, with which probability each of the four chemicals caused symptoms from a symptom class (see Table 1).

Each diagnostic reasoning trial consisted of four sequentially presented symptoms after which participants had to respond with a diagnosis. Diagnostic symptoms pointing more strongly to one chemical also pointed weakly to a second chemical. For example, an “Ab”-symptom would point strongly to A and weakly to B. In addition, there were unspecific symptoms, which were caused with equal probability by all four chemicals. These were denoted with “x”. Thus, an Ab-x-Ba-x symptom sequence could induce A as the initial hypothesis but was ambiguous because it contains equal support for A and B. Such a sequence is ambiguous, but it is still consistent because all symptoms are consistent with both A and B.

In Experiment 1, we presented such ambiguous symptom sequences (AB) together with sequences that more strongly supported A (AAB) or B (ABB). The A-diagnosis was strongly supported by the first symptom, which should result in a higher proportion of A- than B-diagnoses for ambiguous AB-items (primacy order effect). In Experiment 2, participants rated the current support for each of the four alternative hypotheses after each symptom. This explicit highlighting of alternative diagnoses should reduce order effects. In Experiment 3, the procedure was identical to Experiment 1, however, inconsistent symptom sequences such as Cd-Ab-x-Ba were presented that may implicitly highlight alternative diagnoses. Finally, in Experiment 4 the inconsistent symptom sequences were presented as in Experiment 2 with ratings of all alternative diagnoses after each symptom to highlight alternative diagnoses explicitly as well.

## Method

**Participants** Forty (28 female; mean age 23.6, SD = 2.8) undergraduate students from the University of Greifswald and 39 (30 female; mean age 22.1, SD = 2.7) undergraduate



students from Chemnitz University of Technology took part in experiments 1 and 2.

Forty (32 female; mean age 21.5, SD = 2.2) undergraduate students from the University of Greifswald and 39 (26 female; mean age 23.5, SD = 3.2) undergraduate students from Chemnitz University of Technology took part in experiments 3 and 4.

**Material** The four alternative diagnoses were introduced as chemicals that cause symptoms when they affect workers. Each chemical caused symptoms from one symptom class (e.g. Eyes) “almost always” (see Table 1). These were symptoms with a strong causal link to the respective chemical. In addition, each chemical caused symptoms from a second symptom class “occasionally”. These were weak symptoms for the respective chemical. As shown in Table 1, there were two pairs of chemicals. Within a pair, strong and weak symptoms did overlap. For example, the symptom class “Eyes” was strong for R and weak for B, “Respiration” was strong for “B” and weak for “R”. Furthermore, there were two unspecific symptom classes that each chemical could cause “occasionally”.

Each symptom class contained two symptoms. For example, the “Eyes”-symptoms were “Tears” and “Eyelid swelling”. The symptom sequences presented in the diagnostic reasoning trials consisted of four symptoms. Table 2 shows the item types and the symptom orders that they subsume. We constructed each symptom order with each chemical in the “A”-role and each possible assignment of symptoms. For example, if “W” was the “A”-chemical, and “K” was the “B”-chemical for “Ab-x-Ba-x”, one possible symptom assignment would be “Rash-Sting-Paralysis-Swoon”.

Table 1: Domain knowledge participants had to acquire at the beginning

Group	Chem.	Strong symptoms concerning	Weak symptoms concerning	Unspecific symptoms concerning
Gasi-form	R	Eyes	Respiration	Circulatory problems, Pain
Gasi-form	B	Respiration	Eyes	Circulatory problems, Pain
Fluid	W	Skin	Neurolog.	Circulatory problems, Pain
Fluid	K	Neurolog.	Skin	Circulatory problems, Pain

*Note.* The original materials were in German.

In each experiment ambiguous AB-items were presented. In Experiments 1 and 2, the additional item types were AAB and ABB. AB, AAB and ABB item types contain Ab- and Ba-symptoms which both could have been caused by A or B. AB items thus are ambiguous, but they are not inconsistent. In Experiments 3 and 4, the additional item types were CAB and ABC subsuming inconsistent symptom sequences (see Table 2). Inconsistent sequences confronted participants with a “Cd”-symptom that could not be caused by A or B and that was strong for C and weak for D. CAB and ABC items are inconsistent and they are ambiguous with regard to A and B.

**Procedure** In all four experiments, participants were first introduced to the cover story and then acquired knowledge about the chemicals and symptom classes. They studied a table of symptom classes and symptoms and were tested until they could assign symptoms to symptom classes with 100% accuracy. Then they studied a table similar to Table 1 and were tested until they could assign the correct chemical or the correct set of chemicals to a symptom-frequency combination with 100% accuracy. Then, the diagnostic reasoning task was explained and participants were told that the symptoms to be diagnosed were caused by exactly one of the four chemicals.

In each diagnostic reasoning trial in Experiments 1 and 3, four symptoms were presented serially in the center of the screen. Each symptom was shown for 2 s followed by a fixation cross shown for 1 s. After the fourth symptom, participants were prompted to enter one of the four chemicals as their final diagnosis. Then, they were asked to rate their confidence from 1 (very unsure) to 7 (very sure). In Experiments 2 and 4, the trial procedure was similar

except that after each symptom participants rated for each chemical how likely it had caused the symptoms seen so far on a scale from 0 to 100. These ratings are not reported in the present paper. We just consider the effect that this procedure had on the final diagnosis.

Table 2: Orders of symptoms related to first (A) and second (B) respectively third (C) and fourth (D) chemicals; included x stands for unspecific symptoms

Experiment	Item type	Order
1 and 2	AAB Consistent	Ab-Ab-x-Ba
		Ab-Ab-Ba-x
		Ab-x-Ab-Ba
1 and 2	ABB Consistent	Ab-x-Ba-Ba
		Ab-Ba-Ba-x
		Ab-Ba-x-Ba
		Ab-Ba-x-Ba
1 and 2	AB Consistent	Ab-x-x-Ba
		Ab-x-Ba-x
		Ab-Ba-x-x
3 and 4	AB Consistent	x-Ab-Ba-x
		x-Ab-x-Ba
		x-x-Ab-Ba
3 and 4	CAB Inconsistent	Cd-Ab-Ba-x
		Cd-Ab-x-Ba
		Cd-x-Ab-Ba
3 and 4	ABC Inconsistent	Ab-Ba-Cd-x
		Ab-Ba-x-Cd
		Ab-x-Ba-Cd

In each experiment, each participant was presented with each of nine symptom orders (see Table 2) with each of the four chemicals in the A-role resulting in 36 trials in total. The assignment of symptoms to symptom orders was chosen randomly and the trials were presented in randomized order. In addition, four training trials were presented in each experiment.

## Results

**Experiments 1 and 2** Mean proportions of final diagnoses are shown in the top half of Table 3 separated by item type. In both Experiment 1 and Experiment 2, the proportion of A-diagnoses decreased from AAB to AB to ABB items reflecting the decrease in relative support of A. Within-subjects contrasts confirmed this decrease in the proportion of A-diagnoses by significant linear trends,  $F(1, 39) = 230.84, p < .001, \eta^2 = .86$ , and  $F(1, 38) = 474.09, p < .001, \eta^2 = .93$ , respectively.

Focusing on ambiguous AB-items, equal proportions of A- and B-diagnoses each about 50% would be expected normatively. In Experiment 1, there was a clear bias towards A-diagnoses compared with B-diagnoses for AB-items, confirmed by a paired t-test,  $t(39) = 4.54, p < .001, d = 0.72$ . Thus, we obtained a clear primacy order effect for

ambiguous AB-items in Experiment 1, whereas in Experiment 2, the proportion of A-diagnoses did not deviate from the proportion of B-diagnoses,  $t(38) = -0.10, p = .924$ . Mean confidence ratings are shown in the top half of Table 4. Space limitations preclude a detailed analysis but it is apparent that confidence was reduced for the ambiguous AB-items.

Table 3: Means of proportions of diagnoses

Exp.	Item type	A (SD)	B (SD)	C (SD)	D (SD)
1	AAB	.91 (.15)	.09 (.15)		
	AB	.65 (.20)	.35 (.20)		
	ABB	.14 (.21)	.86 (.21)		
2	AAB	.83 (.16)	.17 (.16)		
	AB	.50 (.21)	.50 (.21)		
	ABB	.09 (.10)	.91 (.10)		
3	AB	.62 (.26)	.28 (.19)		
	ABC	.60 (.25)	.22 (.17)	.12 (.12)	.06 (.10)
	CAB	.48 (.25)	.20 (.19)	.26 (.20)	.06 (.08)
4	AB	.50 (.21)	.44 (.19)		
	ABC	.45 (.22)	.29 (.15)	.21 (.18)	.06 (.07)
	CAB	.42 (.22)	.39 (.17)	.14 (.12)	.05 (.08)

*Note.* Proportions for AB items in Experiments 3 and 4 do not sum to 1 because proportions of wrong C and D diagnoses are omitted from the table.

**Experiments 3 and 4** In Experiments 3 and 4, the ambiguous item type AB and inconsistent item types ABC and CAB were presented. Note that with respect to A and B, the item types ABC and CAB contain equal support as well. Thus, normatively equal proportions of A- and B-diagnoses should be elicited by all three item types. Mean proportions of final diagnoses are shown in the bottom half of Table 3.

For AB-items the results are similar to Experiments 1 and 2. Without explicit highlighting of diagnostic alternatives in Experiment 3, there was a clear bias towards A-diagnoses compared with B-diagnoses (primacy order effect) confirmed by a paired t-test,  $t(39) = 4.94, p < .001, d = 0.78$ , whereas with explicit highlighting in Experiment 4, the proportion of A-diagnoses did not deviate from the proportion of B-diagnoses,  $t(38) = 1.02, p = .315$ .

For ABC and CAB items, the leading strong symptom took effect in Experiment 3. The proportion of A-diagnoses was higher for ABC than for CAB items,  $t(39) = 3.42, p = .001, d = 0.47$ , and the proportion of C-diagnoses was higher for CAB than for ABC items,  $t(39) = 4.12, p < .001, d = 0.84$ . Nonetheless, A-diagnoses were more frequent than

C-diagnoses for both item types reflecting the superior support by a strong and a weak symptom as opposed to a single strong symptom. Despite equal support for A and B, A-diagnoses were also more frequent than B-diagnoses for both item types suggesting a primacy order effect in ABC and a similar order effect in CAB, in which the Ab-symptom can frame the integration of the later Ba symptom.

In Experiment 4, there was hardly any effect of the leading symptom on A- and C-diagnoses for ABC and CAB. The proportion of A-diagnoses was comparable for ABC and CAB,  $t(38) = 0.65$ ;  $p = .520$ , and the proportion of C-diagnoses was even lower for CAB than for ABC items,  $t(38) = -2.02$ ;  $p = .050$ ;  $d = -0.46$ .

Table 4: Means of confidence ratings of related diagnoses

E. Item type	A (SD)	B (SD)	C (SD)	D (SD)
1 AAB	5.66 (0.96)	4.08 (1.80)		
AB	3.77 (1.19)	3.69 (1.29)		
ABB	4.63 (1.11)	5.47 (1.14)		
2 AAB	5.10 (1.07)	4.06 (1.34)		
AB	3.49 (0.94)	3.38 (1.06)		
ABB	3.54 (1.16)	5.16 (0.95)		
3 AB	4.05 (1.45)	3.36 (1.52)		
ABC	3.52 (1.45)	3.08 (1.40)	2.83 (1.51)	2.33 (1.35)
CAB	3.41 (1.38)	3.23 (1.41)	3.06 (1.52)	2.29 (1.24)
4 AB	4.00 (1.75)	3.80 (1.43)		
ABC	3.64 (1.66)	3.32 (1.38)	3.32 (1.81)	3.08 (1.60)
CAB	3.92 (1.57)	3.39 (1.47)	2.94 (1.47)	3.25 (1.75)

On top of a decreased primacy order effect, which increased A-diagnoses compared to B diagnoses for ABC items, there was a stronger influence of the last diagnostic symptom than in Experiment 3. The proportion of B-diagnoses was higher for CAB than ABC,  $t(38) = -3.02$ ,  $p = .004$ ,  $d = -0.63$ , and the proportion of C-diagnoses was higher for ABC than CAB. B-proportions in ABC could be reduced simply because of a primacy order effect favoring A. The difference in C-proportions, however, is not open to such an alternative explanation and suggests an increased influence of the last diagnostic symptom in Experiment 4.

Mean confidence ratings are presented in the bottom half of Table 4 and show that inconsistent symptom sequences reduced confidence compared to ambiguous AB-items.

## Discussion

We have investigated effects of symptom order in a diagnostic reasoning task with four candidate hypotheses. Ambiguous symptom sequences (AB-items) equally supporting two alternative diagnoses revealed a clear primacy order effect if participants only responded with a final diagnosis (Experiments 1 and 3). Consistent with the processing assumptions of construction-integration theory, the initial hypothesis suggested by the first symptom framed the integration of subsequent symptoms. An equally supported alternative diagnosis was therefore chosen less often. This order effect is in line with the notion that alternative hypotheses are typically not considered impartially in parallel. Instead, symptom processing proceeds with respect to a focal hypothesis if subsequent symptoms are consistent.

Inconsistent symptoms were not an effective means to highlight alternative diagnoses in Experiment 3. There was still a considerable primacy order effect favoring A over B in ABC- and CAB-items despite equal support. Explicit highlighting of alternative diagnoses, however, was effective. In Experiments 2 and 4, participants rated the current likelihood of each candidate hypothesis after each symptom and thus were led to consider alternative diagnoses. This eliminated the primacy order effect for AB-items. As noted in previous studies, impartial consideration of two alternative diagnoses in parallel can succeed under favorable conditions.

Yet, eliciting ratings of all alternative diagnoses after each symptom did not eliminate the primacy order effect if an inconsistent symptom pattern added a third and presumably even a fourth candidate hypothesis to the set of contenders (ABC- and CAB-items in Experiment 4). In these cases, we did not only observe an advantage for the alternative that was supported by a strong symptom before its equally supported rival (A before B), but also an effect of the last strong symptom. Forcing the participants to consider the current support for all alternatives after this last strong symptom before the final diagnosis resulted in a recency effect. The proportion of final diagnoses was increased for the alternative most strongly supported by the last strong symptom for the inconsistent items ABC and CAB in Experiment 4. Our results are consistent with process models of information integration that specify how early information can frame the processing of later information (Kintsch, 1998, Baumann et al., 2007). They are also consistent with descriptive models predicting order effects in belief updating, hypothesis testing, classification, and judgment and decision making (Hogarth & Einhorn 1992, Koehler, White, & Grondin, 2003) and are a further instance

of confirmation bias (Nickerson, 1998). Studies with more than two contending alternatives are rare. Here, we have shown that the number of relevant contenders matters. The primacy order effect was overcome with two competing alternatives by explicit highlighting. With inconsistent items, more than two hypotheses had to be considered. Constrained by working-memory capacity unpacking of the set of alternatives was incomplete and rather the most likely alternatives were taken into consideration (Dougherty & Hunter, 2003).

Our results may not generalize to instances of diagnostic reasoning in everyday life, in which symptoms can be evaluated more thoroughly without time pressure and search for further information is possible. However, there are situations, in which incoming information has to be processed quickly. For example, physicians evaluating case histories are influenced by early emerging hypotheses (e.g., Kostopoulou, Mousoulis, & Delaney, 2009). The difficulties in considering more than two contenders impartially in the present experiments clearly illustrate the limits in diagnostic reasoning with multiple alternative diagnoses in similar situations.

### Acknowledgments

This research was supported by German Research Foundation (DFG) Grants JA 1761/7-1 and KR 1057/17-1.

### References

- Baumann, M., Mehlhorn, K., & Bocklisch, F. (2007). The activation of hypotheses during abductive reasoning. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Chapman, G. P., Bergus, G. R., & Elstein, A. S. (1996). Order of information affects clinical judgment. *Journal of Behavioral Decision Making*, 9, 201-211.
- Dougherty, M. R. P., & Hunter, J. E. (2003). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, 113, 263-282.
- Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24, 1-55.
- Johnson, T., & Krems, J. F. (2001). Use of Current Explanations in Multicausal Abductive Reasoning. *Cognitive Science*, 25, 903-939.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.
- Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin*, 110, 499-519.
- Koehler, D. J., White, C. M., & Grondin, R. (2003). An evidential support accumulation model of subjective probability. *Cognitive Psychology*, 46, 152-197.
- Kostopoulou, O., Mousoulis, C., Delaney, B. C. (2009). Information search and information distortion in the diagnosis of an ambiguous presentation. *Judgment and Decision Making*, 4(5), 408-418.
- Lange, N. D., Thomas, R. P., & Davelaar, E. J. (2012). Data acquisition dynamics and hypothesis generation. In N. Rußwinkel, U. Drewitz, & H. van Rijn (Eds.), *Proceedings of the 11th International Conference on Cognitive Modeling* (pp. 31-36). Berlin: Universitätsverlag der TU Berlin.
- McKenzie, C. R. M. (1998). Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 771-792.
- Mehlhorn, K., Taatgen, N. A., Lebiere, C., Krems, J. F. (2011). Memory Activation and the Availability of Explanations in Sequential Diagnostic Reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 37, 1391-1411.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2, 175-220.
- Thomas, R. P., Dougherty, M. R., Sprenger, A., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, 115, 155-185.
- Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101, 547-567.
- Wang, H., Johnson, T. R., & Zhang, J. (2006). The order effect in human abductive reasoning: An empirical and computational study. *Journal of Experimental and Theoretical Artificial Intelligence*, 18 (2), 215-247.

# Gaze cues in complex, real-world scenes direct the attention of high-functioning adults with autism

Elizabeth Redcay (redcay@umd.edu)  
Daniel O'Young (droyoung@umd.edu)  
L. Robert Slevc (slevc@umd.edu)

Department of Psychology, University of Maryland  
College Park, MD 20742 USA

Penelope L. Mavros (lmavros@mit.edu)  
John D. Gabrieli (gabrieli@mit.edu)  
Pawan Sinha (sinha@mit.edu)

Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology  
Cambridge, MA 02139 USA

## Abstract

**Abstract** Autism is characterized by atypical use of social communicative cues, such as another person's gaze or point. Despite these real-world difficulties, experimental manipulations often reveal minimal group differences. One factor that may contribute to this failure to find differences is the use of oversimplified and decontextualized social stimuli to examine these behaviors (e.g., a solo floating face). In the current study, we examined whether typical individuals and those with high-functioning autism spectrum disorder would use subtle gaze cues embedded in a natural, real-world scene to aid them in a change detection task using flicker presentation. Each scene contained three changes, and in some pictures, one of those changes was in the direction of gaze of the people in the scene. Even though neither group was aware that gaze cues aided in identification of the change, both groups showed a robust effect of gaze cues in change detection, such that changes in the line of gaze were detected first in a scene more often than those not in the line of gaze. These data illustrate typical use of gaze cues in high-functioning adults with ASD even in the context of a complex, naturalistic scene. Our findings suggest that attention to gaze cues may not be at the root of difficulties with joint attention in adults with autism

**Keywords:** change blindness; joint attention; autism spectrum disorder.

## Introduction

Imagine we are taking a walk in a park and a bird flies overhead. I point to it and look at you, then you look at the bird and you tell me it's a magpie. These simple instances of joint attention, or the intentional coordination of one's own attention with another on an object or topic, occur often in everyday social interactions. Joint attention interactions begin in infancy and provide a robust social learning tool for those who engage in them (e.g., Baldwin et al., 1996; Tomasello, Carpenter, Call, Behne, et al., 2005). Atypical joint attention is a hallmark characteristic in individuals with autism spectrum disorders (ASD) (e.g., Charman, 2003; Mundy & Newell, 2007, but see Gernsbacher, Stevenson, Khandakar, & Goldsmith, 2008) and early joint attention abilities are correlated with atypical language and

social development. While joint attention refers to both the initiation of a bid for joint attention and the response, the focus of the current paper is on processes that may underlie atypical responding. Atypical responding to joint attention could be due to general differences in visual attention (e.g., Brenner, Turner, & Müller, 2007), social attention specifically, for example reduced attention to people, faces, and eyes (Klin, Jones, Schultz, Volkmar, et al., 2002; Pelphrey et al., 2002), and/or a failure to direct one's own attention based on cues from others, to name a few. These various aspects of joint attention behavior have been investigated in a number of different experimental paradigms, however evidence for difficulties with these processes in individuals with autism has been mixed - at least in experimental contexts.

Social attention in autism has been investigated in several ways including the collection of eye-tracking data during viewing of videos, scenes, or pictures of faces and the examination of response patterns during a change blindness paradigm<sup>1</sup>. The first eye-tracking studies to examine whether social attention was atypical in ASD revealed reduced attention to people, and in particular to their eyes, during video viewing (Klin et al., 2002; Pelphrey et al., 2002). While more recent evidence also supports this atypical focus of attention on people within a scene, and on eyes within a face (e.g., Jones, Carr, & Klin, 2008; Dalton et al., 2005; Spezio, Adolphs, Hurley, & Piven, 2007), other evidence does not (e.g., Bar-Haim, Shulman, Lamy, & Reuveni, 2006). Recent studies have capitalized on change blindness methods in order to ask whether attention in a complex scene is prioritized to social aspects of the scene. These studies find that visual attention is prioritized for social agents (New, Schultz, Wolf, Niehaus et al., 2010; Fletcher-Watson, Leekam, Benson, et al., 2009) and eyes (Fletcher-Watson, Leekam, Findlay, & Stanton, 2008) in typical individuals and those with ASD. Thus, among a

---

<sup>1</sup> In a change blindness paradigm participants must identify small change(s) between two images through comparing images side-by-side or through flicker presentation.

high-functioning group of participants with ASD, there is mixed support for problems with social attention, suggesting that either this may not be a source of atypical joint attention (at least by adulthood) or the experimental paradigms used are failing to capture the difficulties.

In a joint attention context social attention is necessary but not sufficient to coordinate attention with another. If I just looked at you pointing to the magpie but didn't look at the magpie I would not have engaged in joint attention with you nor would I have learned the name for that type of bird. Indeed, individuals must use the other person's shift in attention (e.g., through gaze) to redirect their own attention to achieve joint attention. Thus, one hypothesis is that atypical use of gaze, rather than social attention alone, may underlie atypical joint attention. Researchers have examined the effect of gaze cueing on attention in experimental tasks and again find mixed to weak support for an atypical response in autism. However, these studies have been largely devoid of social context. These studies have mostly been conducted through the use of a modified Posner cueing task in which participants are told to push the left or right button when they detect a target on the left or right side of the screen, respectively. A face at the center of the screen shifts gaze just prior to appearance of the target to a location that is either congruent or incongruent with the target location. Greater reaction time for the incongruent than congruent trials suggests reflexive use of the gaze cue information to redirect attention even though participants are told the gaze cue is irrelevant to the task. In a review of 12 experiments, 8 revealed no differences between autism and control groups with this method (review, Nation & Penny, 2008), suggesting reflexive orienting of attention to gaze cues in ASD is fairly typical. As mentioned above, this failure to find a difference could be due to the lack of social context in the stimuli. A study using eye-tracking methods examined attention to objects in a person's line of gaze in a real-world scene. While individuals with autism reliably followed the person's gaze to the objects, they looked less at the objects of attention than controls (Freeth, Chapman, Ropar, & Mitchell, 2010). This study illustrates typical gaze following in ASD individuals, but suggests atypical use of gaze information (i.e., less looking at the objects). One limitation of the study, though, is that the gaze cues were so salient in the scenes that the study may not be tapping into more subtle, real-world difficulties in attention to and use of gaze information to orient attention to relevant aspects in ones environment.

In sum, studies of social attention using eye-tracking or change blindness methods do not examine the effect of gaze on attention reorienting, while gaze cueing paradigms that do examine the effect of gaze have previously used oversimplified stimuli with minimal social context or real-world validity. To overcome these limitations and examine whether the object of another person's attention is prioritized in a complex visual scene, we used a change blindness paradigm in which gaze cues were subtly embedded in a complex, natural scene.

## Methods

### Participants

Eleven adults with autism spectrum disorder (ASD) ( $29.6 \pm 4.2$  years, 8 male) and fifteen typically developing adults (TD) ( $24.4 \pm 5.2$  years, 9 male) participated in this study. All participants gave written informed consent and received monetary compensation for their participation. Two TD subjects were excluded from subsequent analysis, one for having been previously diagnosed with depression and another for an ophthalmologic developmental disorder. Participants entered the study with a diagnosis of autism or Asperger's from their community healthcare provider but in order to confirm a research diagnosis of autism spectrum disorder, an autism behavioral therapist (P.L.M.) who was certified on the Autism Diagnostic Observation Schedule (ADOS) – Module 4 (Lord et al., 2000) administered and scored the ADOS. A psychiatrist confirmed a diagnosis on the spectrum by viewing tapes of the ADOS interview. Of the 11 participants, 8 met criteria for autism (i.e., a combined social-communication score of 10 or higher), the other three 3 met criteria for spectrum (i.e., a combined social-communication score of 7 or greater). Social-communication scores ranged from 7-14 with a mean of 10 (stdev 2.24).

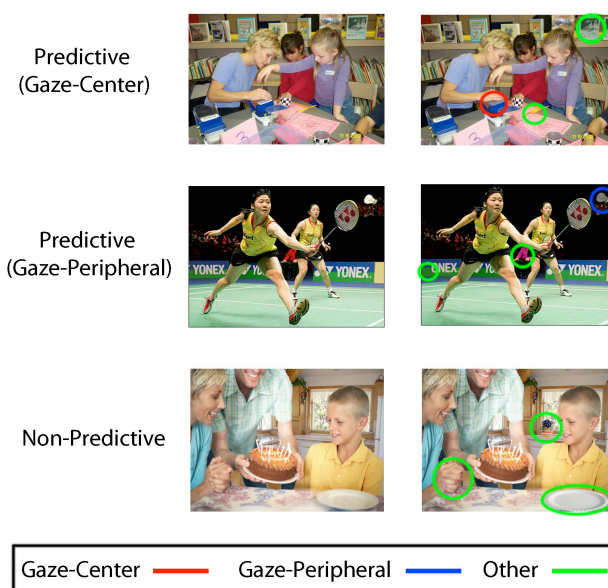


Figure 1: Examples of the picture types (Predictive or Non-Predictive) and change types (gaze-center, gaze-peripheral, non-gaze center, non-gaze peripheral).

### Stimuli

Twenty-nine color photographs of natural scenes were used to create the stimuli. Each photograph included between one and three people with visible faces and several objects. The photographs were edited using Adobe Photoshop. For each photograph, three changes were made to each picture.



Attempts were made to balance the changes in the center and periphery across pictures. Center regions were defined as the area within an oval that is half of the height and width of the entire picture, while peripheral regions consisted of any areas outside of the center regions. Twenty of the photographs had a change that was within the line of direct gaze of the person(s) in the photograph. These twenty photos are labeled 'Predictive' because one of the changes can be predicted by the gaze direction. The other nine photos ('Nonpredictive') were included as fillers so that participants were unlikely to notice the predictive nature of the gaze changes.

### Stimulus ratings

Eight new participants not informed of the purpose of the current study rated each change within the set of predictive pictures as 'center' or 'periphery' based on the definition given above. Additionally, these participants judged whether the change occurred in the foreground or background of the scene. Two changes could not be categorized as center or periphery (50% ratings for each). These two changes were within the same picture so that picture was removed from all further analyses. Thus, out of 19 predictive pictures, 11 changes were rated as in the center and in the line of gaze (gaze-center), 8 were in the periphery and in the line of gaze (gaze-periphery), 15 were in the center and not in the line of gaze (non-gaze center), 23 were in the periphery and not in the line of gaze (non-gaze periphery). All gaze changes were rated as in the foreground of the picture whereas non-gaze changes were a mix of foreground and background, thus the grounding (foreground vs. background) of changes was included as a covariate in the statistical models (below).

### Visual controls

Photographs were analyzed for visual saliency using the saliency toolbox (Walther & Koch, 2006). This toolbox identifies the most salient 'objects' (or locations) in a scene. Saliency was determined based on a combination of intensity, orientation, and color maps. When choosing regions in which to make a change, we attempted to avoid these most salient regions. In order to compare whether saliency values from the pixels that contained a change differed between the change types, a saliency map was generated which contained a saliency value at each pixel. Regions of interest were created for each change region in each picture. Mean saliency value was extracted from within each change region using Matlab. Another potential low-level feature that could have influenced change detection was size of the change. The size of the change was determined by counting the number of pixels in the regions of interest for each change. Two one-way ANOVA's were conducted to examine the effect of saliency and size on change category (gaze center, gaze periphery, non-gaze center, non-gaze periphery). No significant difference in size of the change types was found ( $F(3)=.576, p<.63$ ) but saliency did show an effect of change category

( $F(3)=4.19, p<.01$ ) with changes in the center and line of gaze showing higher saliency values than changes in the periphery. Pairwise contrasts were examined using Tukey's HSD ( $\alpha<.05$ ), and no differences in saliency were found between nongaze-center changes, gaze-center changes, and gaze-periphery changes or between gaze-periphery and non-gaze periphery changes. Saliency and change size were both included as covariates in the model (below).

### Task

The experiment was conducted on an iMac computer with stimuli presented on an auxiliary 19" monitor using Matlab Version 7.7 (R2008b) with Psychophysics Toolbox Version 3. Participants received written and verbal instructions that they would be viewing a series of picture pairs and that these pairs would be presented in rapid alternation and would be identical except for three changes, such as changes in the color or the presence of an object. Participants were asked to identify these changes as quickly as possible and to respond via a button press. Participants were first given a practice task consisting of two trials that were identical to the actual task to ensure that subjects fully understood the task. Each trial began when the subject indicated that they were ready by pressing the mouse button. The subject would then be presented with the original photograph for 1000ms, a blank screen for 200ms, then the altered image for 1000ms, and another blank screen for 200ms. This sequence was looped until the subject indicated via a button press that a change was observed between the photographs. The button press brought up a blank screen. The subjects were required to give a verbal description of the change and to point on the blank screen in the area where the change was observed. The experimenter would record the change on the computer. Accuracy was determined online by the experimenter. When the subject indicated they were ready, this process would continue with the same photograph sequence. The subsequent trial would begin with a new stimulus once all three changes were observed, or until a time limit of 3 minutes was reached. Stimulus order was randomized for each subject.

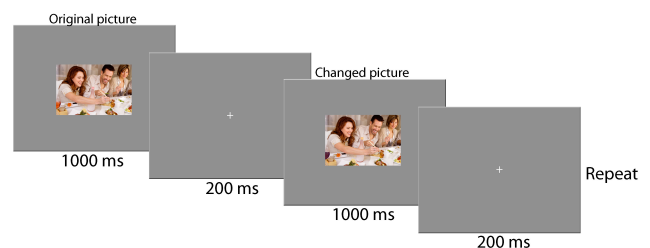


Figure 2: Timeline of one trial

### Statistical analyses

Separate analyses were conducted on total number of changes detected and on the first change detected in each picture using logistic mixed-effects models as implemented in the lme4 package (Bates, Maechler, & Bolker, 2011) in



the R software package (version 2.14.1; R Development Core Team, 2011). *Gaze* (change in the line of gaze or not), *location* (center or peripheral change) and *group* (ASD or control) were treated as fixed effects using orthogonal contrast coding, and participants and items were treated as crossed random effects, with the maximal random effects structure supported by the data. In addition, change *saliency*, *size*, and *grounding* (foreground or background) were included as centered fixed effect covariates.

Mixed effects models are useful for data such as these because they do not require aggregation across items nor do they require fully balanced data (Jaeger, 2008). Log-transformed response times were analyzed with similar (non-logistic) models. For readability and for purposes of graphical presentation, values are described as proportions (calculated as means of participant means) rather than as log-odds ratios and as raw response times (after excluding times exceeding two standard deviations from each participant's mean response time) rather than log transformed times.

## Results

### Gaze cues facilitate change detection

Participants first detected 53.5% of the changes in the line of gaze compared to only 25.3% of the changes not in the line of gaze; this 28.2% difference was reflected in a significant main effect of gaze (see Figure 3a;  $b = 1.46$ ,  $SE = 0.47$ ,  $z = 3.09$ ,  $p < .01$ ). Unsurprisingly, participants were (marginally significantly) more likely to first detect changes that were larger ( $b = 0.30$ ,  $SE = 0.02$ ,  $z = 1.69$ ,  $p < .10$ ). Although participants first detected numerically more changes in the periphery than in the center (37% vs. 32%), this difference did not reach significance ( $b = 0.64$ ,  $SE = 0.39$ ,  $z = 1.64$ ,  $n.s.$ ). There was, however, a marginally significant effect of location when the covariates of size, saliency, and grounding were not included in the model ( $b = 0.90$ ,  $SE = 0.54$ ,  $z = 1.68$ ,  $p < .10$ ); this likely reflects the fact that center changes were nearly always also foreground changes with higher saliency values than changes in the periphery. (Note that the model does not show concerning levels of collinearity; all variance inflation factors  $< 1.7$ ).

There is evidence that changes in the foreground of pictures are easier to detect than changes in the background (Rensink, O'Regan, & Clark, 1997), however *grounding* did not explain significant variance in change detection or response times, suggesting that these differences were relatively minor.

Analysis of detection times to the first change (Figure 3b) revealed a similar pattern of results.<sup>2</sup> A main effect of gaze showed that participants detected changes in the line of gaze an average of 2.3 seconds faster than changes not in the line of gaze ( $b = -0.30$ ,  $SE = 0.13$ ,  $t = -2.37$ ), a main effect of saliency showed faster detection of more visually salient

changes ( $b = -0.40$ ,  $SE = 0.17$ ,  $t = -2.30$ ), and a marginally significant effect of group ( $b = -0.33$ ,  $SE = 0.18$ ,  $t = -1.82$ ) showed that control participants detected their first change an average of 2.4 seconds faster than ASD participants.

### No effect of group on identification of gaze changes

No other effects or interactions reached significance (all  $ps > .1$ ); in particular, the effect of gaze did not differ for ASD and control participants (29.0% vs 27.6%, respectively).

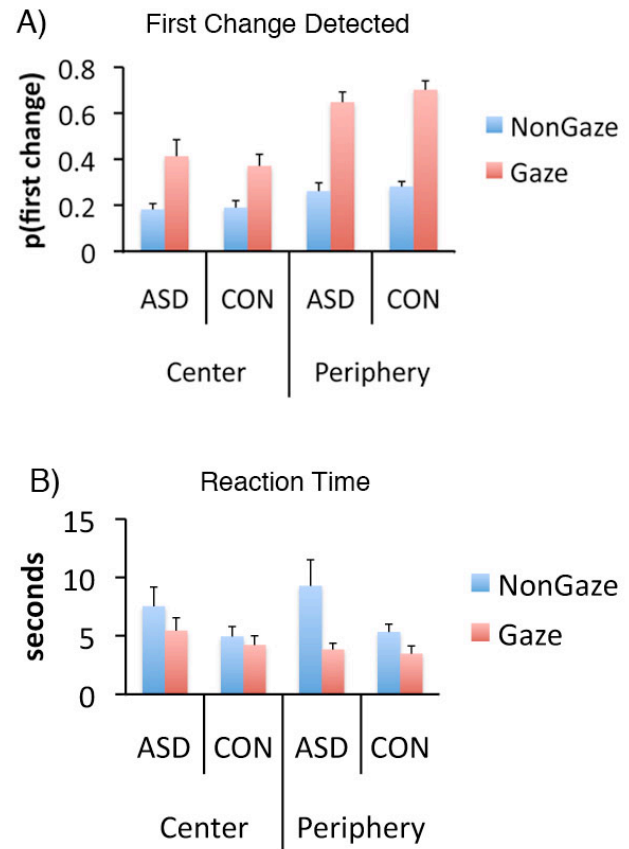


Figure 3: Effect of gaze on first change detected (A) and reaction time to detect change (B).

### Total number of changes detected by group

Overall, ASD participants detected fewer changes than did controls, however this was largely due to the performance of one ASD participant who missed 35% of the changes (compared to an average miss-rate of 5.3% for the rest of the ASD group and 2.8% for the control group). With this participant excluded, ASD participants did not detect significantly fewer changes than did controls (94.7% vs. 97.2% detected;  $b = -0.71$ ,  $SE = 0.75$ ,  $z = -0.95$ ,  $p > .10$ ), although a marginally significant gaze by group interaction ( $b = -2.08$ ,  $SE = 1.22$ ,  $z = -1.71$ ,  $p = .09$ ) showed that ASD participants were more likely to miss changes that were not in the line of gaze (missing 6.6% non-gaze changes vs. 2.6%

<sup>2</sup> The lme4 package cannot currently estimate  $p$  values for  $t$ -statistics in models with random slopes, however common practice is to assume that  $t$  values greater than 2 are statistically significant.

of gaze changes) whereas controls showed a smaller difference, numerically in the other direction (2.4% missed non-gaze vs. 3.6% missed gaze changes).<sup>3</sup> ASD participants also took longer to detect all three changes (when all changes were, in fact, detected) than did control participants (an average of 51.5 vs. 40.1 seconds, respectively;  $b = -0.16$ ,  $SE = 0.05$ ,  $t = -3.00$ ).

## Discussion

### Subtle gaze cues robustly affect visual attention

A robust effect of gaze on change detection was found in both the typical and ASD groups. This effect cannot be explained by low-level visual effects like size or saliency or grounding (foreground vs. background) of the change because the effect of gaze on first change detected and on time to detect the change were significant even with these variables as covariates.

Gaze cueing effects in a change detection paradigm have been demonstrated in typical adults previously (Langton, O'Donnell, Riby, & Ballantyne, 2006) but the gaze cues in those stimuli were highly salient (e.g., one person in the picture at a desk with body facing forward and some objects around him). Our findings provide a stronger test for the role of gaze in object detection through the presentation of multiple changes within a complex scene with multiple people and subtle gaze cues. In fact the cues were so subtle that in a post-test debriefing session, participants reported that they were unaware that direction of gaze provided any information about the location of the changes. Thus, these findings provide strong support that the object of someone's attention is prioritized during visual search – even without explicit awareness of the utility of gaze cues in this context.

### No effect of group on change detection

Surprisingly, no effect of group was found. In fact, there was a trend for individuals with ASD to show a greater effect of gaze on change detection, especially for changes in the periphery. While we predicted an effect of group, this lack of a difference is consistent with several other change blindness studies suggesting normal prioritization of attention to social agents in ASD (Fletcher-Watson et al., 2008; New et al., 2010). Our findings extend previous work by showing normal prioritization of attention to the object of another person's attention in ASD even when the gaze cues were subtly embedded into the scene. These findings suggest that attention to (and use of) gaze information in high-functioning adults with autism may not be impaired and thus are not a determinant of atypical social interactions in adults.

An alternative possibility is that these (and other) findings of a failure to find group differences in offline, experimental contexts suggest that difficulties in social attention and gaze processing are due to inherent difficulties with real-time

social interactions which are not captured by these offline tasks. The unpredictability and fast-pace of live interactions may add sufficient challenges to cause an otherwise capable system of social attention to break down. This conclusion remains speculative however without concurrent data from participants in the context of a real-time face-to-face interaction.

Finally, other factors may have contributed to this lack of a group difference. The participants in this study were high-functioning adults with autism and Asperger's disorder. Thus, their level of social impairment is by definition less severe than the majority of individuals with autism. Second, they have had years to develop strategies for success in social interactions, such as directing attention to the face and eyes. Indeed, atypical gaze processing may be greatest early in life (e.g., Klin, Jones, Schultz, & Volkmar, 2003; Leekam, Hunnisett, & Moore, 1998) but through treatment and compensatory strategies these differences may be minimized by adulthood.

## Conclusions and Future Directions

These data show that normal use of gaze information in high-functioning adults with autism is not restricted to visually simplistic scenes; instead, in complex, real-world scenes with subtle embedding of gaze cues individuals with ASD show normal attentional prioritization to objects in the line of another's gaze. An important question for future research will be to determine when and how adults with autism do show atypical use of gaze information (e.g., see review by Klin, Jones, Schultz, & Volkmar, 2003). Atypical use of gaze information may be the most robust in real-time face-to-face contexts that are unpredictable and quickly changing. However, limited research has been done to investigate spontaneous difficulties with joint attention in adults with autism. This work will be critical to understand the specific difficulties with social interaction experienced throughout the lifespan in autism. For example, factors that underlie atypical social interactions in a toddler may be very different than those of an adult. Understanding how joint social attention is atypical will allow for more targeted interventions and treatments for adults with autism, an area of relatively less investigation (e.g., Autism Speaks Strategic Plan, 2009).

## Acknowledgments

The authors thank Kang Won Choi for discussions on the design of the study. We gratefully acknowledge Rashida Callendar, Joy Ekuta, and Eugenia Luo for assistance with stimuli creation and editing and Ariana Mann for assistance with stimuli creation and pilot testing. This work was supported by funding from the Simons Foundation Autism Research Initiative (SFARI) awarded to J.G. and P.S. and a postdoctoral fellowship awarded from the National Institute of Child Health and Human Development awarded to E.R.

<sup>3</sup> All other analyses are reported over data from all participants, but note that the pattern of results does not change appreciably if this participant is excluded.

## References

- Baldwin, D. A., Markman, E. M., Bill, B., Desjardins, R. N., Irwin, J. M., & Tidball, G. (1996). Infants' Reliance on a Social Criterion for Establishing Word-Object Relations. *Child development*, 67(6), 3135–3153. Wiley Online Library.
- Bar-Haim, Y., Shulman, C., Lamy, D., & Reuveni, A. (2006). Attention to eyes and mouth in high-functioning children with autism. *Journal of autism and developmental disorders*, 36(1), 131-7.
- Bates, D. M., Maechler, M., & Bolker, B. (2011). Linear mixed-effects models using S4 classes. *R A Language and Environment for Statistical Computing*. R package version 0.9975-0.
- Brenner, L. a, Turner, K. C., & Müller, R.-A. (2007). Eye movement and visual search: are there elementary abnormalities in autism? *Journal of autism and developmental disorders*, 37(7), 1289-309.
- Charman, T. (2003). Why is joint attention a pivotal skill in autism? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 358(1430), 315-24.
- Dalton, K. M., Nacewicz, B. M., Johnstone, T., Schaefer, H. S., Gernsbacher, M. A., Goldsmith, H H, Alexander, A. L., et al. (2005). Gaze fixation and the neural circuitry of face processing in autism. *Nature neuroscience*, 8(4), 519-26. doi:10.1038/nn1421
- Fletcher-Watson, S, Leekam, S R, Benson, V., Frank, M. C., & Findlay, J M. (2009). Eye-movements reveal attention to social information in autism spectrum disorder. *Neuropsychologia*, 47(1), 248-57.
- Fletcher-Watson, Sue, Leekam, Susan R, Findlay, John M, & Stanton, E. C. (2008). Brief report: young adults with autism spectrum disorder show normal attention to eye-gaze information-evidence from a new change blindness paradigm. *Journal of autism and developmental disorders*, 38(9), 1785-90.
- Freeth, M., Chapman, P., Ropar, D., & Mitchell, P. (2010). Do gaze cues in complex scenes capture and direct the attention of high functioning adolescents with ASD? Evidence from eye-tracking. *Journal of autism and developmental disorders*, 40(5), 534-47.
- Gernsbacher, M. A., Stevenson, J. L., Khandakar, S., & Goldsmith, H. Hill. (2008). Why Does Joint Attention Look Atypical in Autism? *Child Development Perspectives*, 2(1), 38-45.
- Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of memory and language*, 59(4), 434-446. Elsevier Inc. Jones, W., Carr, K., & Klin, A. (2008). Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. *Archives of general psychiatry*, 65(8), 946-54.
- Klin, A., Jones, W., Schultz, R., & Volkmar, F. (2003). The enactive mind, or from actions to cognition: lessons from autism. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 358(1430), 345-60.
- Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of general psychiatry*, 59(9), 809-16.
- Langton, S. R. H., O'Donnell, C., Riby, D. M., & Ballantyne, C. J. (2006). Gaze cues influence the allocation of attention in natural scene viewing. *Quarterly journal of experimental psychology* (2006), 59(12), 2056-64.
- Leekam, S R, Hunnisett, E., & Moore, C. (1998). Targets and cues: gaze-following in children with autism. *Journal of child psychology and psychiatry*, 39(7), 951-962.
- Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, a, et al. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, 30(3), 205-23.
- Mundy, P., & Newell, L. (2007). Attention, Joint Attention, and Social Cognition. *Current directions in psychological science : a journal of the American Psychological Society*, 16(5), 269-274.
- Nation, K., & Penny, S. (2008). Sensitivity to eye gaze in autism: is it normal? Is it automatic? Is it social? *Development and psychopathology*, 20(1), 79-97.
- New, J. J., Schultz, R. T., Wolf, J., Niehaus, J. L., Klin, A., German, T. C., & Scholl, B. J. (2010). The scope of social attention deficits in autism: prioritized orienting to people and animals in static natural scenes. *Neuropsychologia*, 48(1), 51-9.
- Pelphrey, K. a, Sasson, N. J., Reznick, J. S., Paul, G., Goldman, B. D., & Piven, J. (2002). Visual scanning of faces in autism. *Journal of autism and developmental disorders*, 32(4), 249-61.
- Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To See or not to See: The Need for Attention to Perceive Changes in Scenes. *Psychological Science*, 8(5), 368-373. SAGE Publications.
- Spezio, M. L., Adolphs, R., Hurley, R. S. E., & Piven, J. (2007). Abnormal use of facial information in high-functioning autism. *Journal of Autism and Developmental Disorders*, 37(5), 929-939. Springer.
- Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *The Behavioral and brain sciences*, 28(5), 675-91; discussion 691-735.
- Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural networks : the official journal of the International Neural Network Society*, 19(9), 1395-407.

# Examining the Representation and Understanding of Large Magnitudes Using the Hierarchical Alignment model of Analogical Reasoning

**Ilyse Resnick (ilyse.resnick@temple.edu)**

Department of Psychology, 1701 North 13th Street  
Philadelphia, PA 19122 USA

**Thomas F. Shipley (thomas.shipley@temple.edu)**

Department of Psychology, 1701 North 13th Street  
Philadelphia, PA 19122 USA

**Nora Newcombe (nora.newcombe@temple.edu)**

Department of Psychology, 1701 North 13th Street  
Philadelphia, PA 19122 USA

**Christine Massey (massey@seas.upenn.edu)**

Institute for Research in Cognitive Science, 3401 Walnut St  
Philadelphia, PA 19104 USA

**Theodore Wills (twills@temple.edu)**

College of Education, 1301 W. Cecil B. Moore Avenue  
Philadelphia, PA 19122 USA

## Abstract

Understanding scale is fundamental in science education, but scale comprehension is difficult. One reason difficulties may arise is a disconnect between the linear scale of magnitude and how scale information is cognitively represented. An intervention was designed to foster a linear representation of magnitude, based on the theory that people represent magnitude information in a hierarchically organized structure. The intervention extends principles from the progressive alignment model of analogical reasoning to include hierarchical alignment. Half the students in an undergraduate introductory-level geology class were given multiple opportunities to progressively align time to a constant spatial scale in a linear representation, and locate all previous scales relative to the current scale. The other half of the class received the same content and practice aligning time to space. The intervention group demonstrated a more accurate sense of the relative durations of geological events and a reduction in the magnitude of temporal location errors relative to the control group. These findings suggest that the hierarchical and progressive alignment of geologic time is an effective way to reduce magnitude-based errors in understanding geologic time. These findings are consistent with the category adjustment model, and suggest commonalities between number and time magnitude representation. Educational implications are discussed.

**Keywords:** Scale; Hierarchical alignment; Progressive alignment; analogy

## Introduction

Having a strong conceptual understanding of scale and the relationships between scales is essential for scientific literacy (Tretter, et al., 2006). Fundamental concepts in

many disciplines require understanding scales outside those familiar to human experience. For example, research on geologic time, the atom, the size of the universe, and nanotechnology is all based on phenomena occurring at scales that cannot be directly perceived. Being able to understand important current social issues, such as the U.S. deficit, population growth, and global warming also require an understanding of magnitudes outside of direct human experience. Given the importance of understanding scale, it should be no surprise that the new NRC *Framework for K-12 Science Education* (NRC, 2011) and the *Benchmarks for Science Literacy* (AAAS, 1993) have both identified “size and scale” as fundamental and a unifying theme in science education. “Size and scale” was also identified as one of the “big ideas” at recent nanoscience and education national workshops (Swarat, et al., 2010).

Unfortunately, people consistently have trouble understanding and comparing values of very small or large magnitudes (e.g. Jones, et al., 2008; Libarkin, et al., 2005; Tretter, et al., 2006; Swarat, et al., 2010). Undergraduate students, even those in STEM majors, have difficulty mastering concepts of size and scale (Drane et al., 2008). Size and scale has been described as a critical barrier to learning and higher-level understanding (Hawkins, 1978). While people are more accurate at ranking relative sizes, they struggle assigning, comprehending, and comparing absolute sizes, especially at extreme scales (Jones, et al., 2008; Tretter, et al., 2006). For example, while most students are able to place major geologic events in the correct order, they fail to demonstrate an understanding of the magnitude of time between these events (Libarkin, Kurdziel, & Anderson, 2007).

Difficulties processing extreme sizes and scales may stem from how magnitude information is cognitively represented. Magnitudes at extreme scales are unfamiliar. Activation of representations of unfamiliar magnitudes is less automatic than of familiar values (Kadosh & Walsh, 2009). For example, people possess a weaker association between magnitude and number words for larger quantities than for smaller more familiar quantities (Sullivan & Barner, 2010).

Unfamiliarity with the magnitude and content information associated with extreme scales may lead to the large conceptual categories held by novices (Trend, 2000; Tretter, et al., 2006). While experts working with extreme scales are characterized as having a “detailed, secure, sophisticated, and well developed” mental framework, novices’ mental frameworks are found to be “scant, insecure, and nebulous” (Trend, 2000). For example, even in-service science teachers who teach geologic time represent the roughly 14 billion years of geologic events as only three conceptual categories: extremely ancient, moderately ancient, and less ancient (Trend, 2000). Conceptual boundaries are defined by consistent estimations of events near each other, creating the boundary, and increased variation of estimations of events within each conceptual category across participants.

Huttenlocher and colleagues’ (1988) category adjustment model offers an account for this pattern of estimations. The category adjustment model applies to both objects and events. It suggests that 1D, 2D, and 3D magnitudes, such as location, distance, and duration, are stored as a hierarchical combination of metric and categorical information. A person retrieves needed information at the level required by a question, as well as the boundaries of any associated higher-level units (Huttenlocher, et al., 1988). For example, remembering that dinosaurs first appeared in the Triassic Period implicitly contains information that dinosaurs also first appeared during the Mesozoic Era, which is a larger division that includes the Triassic period.

However, in the absence of exact information, people use boundaries of other objects/events to help make estimations. Variation in estimation, therefore, occurs because of imprecision of boundaries (Shipley & Zacks, 2008; Zacks & Tversky, 2001). As people use object/event boundaries to help make estimations, the more imprecise or the larger the boundaries, the more variation one could expect to find (Huttenlocher, et al., 1988; Shipley & Zacks, 2008). With no information at a lower level, estimations must default to a higher level. Thus, if a student cannot recall which period dinosaurs first appeared, but can recall it happened in the Mesozoic Era, their estimation will range 180 million years, spanning all of the periods that comprise the Mesozoic Era.

The placement of object/event boundaries will systematically distort estimations in predictable ways. Subjective experience of magnitude is influenced by the number of boundaries the person can recall; the more boundaries a person can recall the greater the subjective magnitude, and the converse for recollection of a smaller number of boundaries (Block, 1990). In line with the category adjustment model, when people hold relatively few

conceptual categories, such as at extreme scales, they should underestimate magnitudes. For example, elementary to graduate-level students estimated objects as too small at large scales and as too large at small scales (Tretter, Jones, & Minogue, 2006).

Additionally, because change is usually perceptually salient, and thus plays a role in object/event comprehension and memory, at points of unpredictability humans are more likely to attend to information to permit more accurate future predictions (Shipley & Zacks, 2008). Subsequently, people tend to remember objects/event boundaries by attending to them (Speer, Zacks, & Reynolds, 2007), and recall those objects/events at boundaries more clearly than those in between (Zacks & Tversky, 2001). Therefore, regions sparsely populated with objects/events will tend to elicit more variation in estimation of location and an underestimation of magnitude.

The more organizational structure a person has for the material in memory, the better their recall (Mandler, 1967). Where people have more conceptual categories, perhaps arising from personal experience with the scale, they are more accurate when making judgments. For example, most adults (e.g. Dehaene & Marques, 2002; Dehaene, et al., 2008) and children (Booth & Seigler, 2008) are able to use a proportional linear number line to make estimations for smaller, more familiar numbers, but they fail to do so with larger or unfamiliar numbers. While there currently is a debate about the nature of people’s mental representation of size and scale (e.g. logarithmic, power, scalar variability, or segmented linear model), it is clear that there are compressive effects on people’s estimation of size and scale as magnitude increases to unfamiliar scales. The variation of people’s estimations of quantity increases as a function of the magnitude of the judgment (Dehaene, 2003). A consequence of a compressed number line is that, as magnitudes become less familiar, values will become less discriminable (the distance effect). The distance effect can be seen in slower response times when people make judgments about larger numbers compared to making judgments about smaller numbers (Dehaene, et al., 2008).

If the representation of scale information drives student difficulties in learning about size and scale, then learning interventions designed to address scale representation more directly should improve learning. Effectively teaching reasoning about unfamiliar scale magnitudes should require an intervention that fosters a linear representation of magnitude, populated with boundary information at that scale. An intervention was designed based on the theory that people represent magnitude information in a hierarchically organized structure.

The intervention tested in this study is based on the progressive alignment model of analogical reasoning (Kotovsky & Gentner, 1996). The progressive alignment model has been shown to foster a linear representation of number magnitude (Thompson & Opfer, 2010). The progressive alignment model advocates the comparison of two similar items. The more commonalities that exist

between these items, and the more these commonalities are highlighted, the more salient corresponding relations will be. Comparing two similar items then helps extend the analogy to unfamiliar items (Gentner & Namy, 2006). Furthermore, the act of performing comparisons may change the original mental representations, increasing uniformity between the two representations. Thus, the process of alignment may make higher-order relational similarities more salient. Recognition of higher-order relational commonalities may promote making similar higher-order connections with subsequent unfamiliar items (Kotovsky and Gentner, 1996). The progressive alignment of scales may alleviate the conceptual dissimilarity between human scales and extreme scales by providing greater structural alignment across changes of scale.

The current study uses the Geologic Time Scale, extending from present day back 4.6 billion years. Novices have trouble understanding geologic time, demonstrating a pattern of errors consistent with a hierarchically organized representation of temporal magnitude (e.g. Libarkin, et al., 2005; Trend, 2000). Novices' estimations of when geologic events occurred may differ from the correct magnitude by as much as five orders of magnitude (Catley & Novick, 2008).

A commonly employed classroom exercise to teach students about the magnitude of geological time is to have them align time to a spatial representation. The current study builds on the use of space as an analogy for time. We note that using space to represent time is particularly important in geoscience education because geologically relevant temporal information is often stored in spatial arrays (e.g., as sequences of layers in a sedimentary deposit). In line with the progressive alignment model, the current intervention gives students multiple opportunities to align time to space in a linear representation, progressing from small familiar scales to geological scales. While the amount of time varies, the amount of space remains constant: students align increasing amounts of time to one meter.

Importantly, the current intervention extends the principles of the progressive alignment model to include the hierarchical organization of all previous scales. Every time students align a new temporal scale to space, they locate all previous scales relative to the current scale. This hierarchical organization highlights how each temporal scale is related to the others, helping to populate each scale with boundary information by providing internal structure of magnitude relations within event boundaries.

## Methods

### Participants

Participants consisted of 58 (control group) and 49 (experimental group) students enrolled in an undergraduate introductory-level geoscience course at a major university located in an urban setting. The demographics of participants were consistent with those of a large urban American university.

The geoscience course consisted of twice weekly lectures and a laboratory period. All lectures were given by the same faculty member; the students were divided into different sections for the laboratory period. One TA covered four sections and two TAs covered two sections each. The intervention was conducted by the first author (as a guest lecturer) in the laboratory sections as part of the standard stratigraphy lab. Experimental (intervention) and control conditions were evenly distributed across the TAs to control for instructor-based differences.

**Intervention Design** In the hierarchical alignment intervention, students aligned time to space beginning with a familiar personal time scale, then worked through different historic and geologic timelines, up to the full Geologic Time Scale. For each timeline, students were required to indicate the timeline's length, locate specific events, and locate where all previous timelines would begin on the current timeline (see Fig. 1).

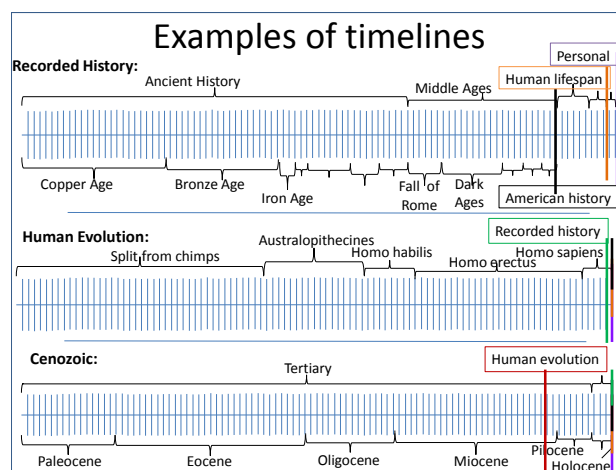


Figure 1: Example of three time lines in the hierarchical alignment intervention. Note that for each time line, all previous time lines are located.

**Procedure** In a two-hour laboratory session, the experimental group participated in the hierarchical alignment intervention (1.5 hr) after a shortened stratigraphy lab (30 min). The control group completed the full stratigraphy lab (2 hr). During the stratigraphy lab, students learned about the age and distribution of rock types and the types of environments in which those rocks are formed by making and examining stratigraphy columns. Importantly, stratigraphy columns involve aligning geologic temporal information to space. Thus, the intervention and control groups both received practice aligning geologic time to space and exposure to magnitude information. The only difference between the intervention and control groups is the way in which the magnitude information was presented (hierarchically or conventionally). Both the intervention and control groups received further instruction on the Geologic Time Scale and concepts explicitly related to geologic time



(i.e. two fossil labs) prior to completing the outcome measures. The fossil labs include identifying fossils from different divisions in time.

**Measures** All students completed outcome measures one month after the stratigraphy lab as part of a laboratory exam. There were two items that assessed understanding of geologic time magnitude. One item came from the Geoscience Concept Inventory, which is a valid and reliable instrument measuring a range of geoscience concept knowledge (Libarkin, et al., 2005). For this item, students were presented with five time lines that had the same geologic events in different locations. Four of the time lines represented common misconceptions students have (response option A. life occurred when Earth formed, B. humans and dinosaurs coexisted, C. dinosaurs appeared much earlier than they did, E. all life formed at the beginning of Earth's history), and one time line showed the events in the correct relative locations (D). Students were asked to choose the most correct time line. Two of the incorrect response options (A/B) reflected relatively small magnitude errors (i.e. they are wrong on the scale of millions of years) and the other two incorrect response options (C/E) reflected relatively large magnitude errors (i.e. they are wrong on the scale of billions of years).

The second item is a new measure of geologic time developed for use with middle school students (Barghaus & Porter, 2010). This item is a multiple-choice item that required students to identify which duration-based statement was true using a conventional diagram of the Geologic Time Scale. The correct choice is the statement: *The Proterozoic Eon lasted much longer than the Phanerozoic Eon*. While numerical information is provided in the diagram, the correct choice may not be obvious to novices in the standard diagram because the spatial intervals of the eons do not proportionally correspond to their temporal lengths. This type of compressed representation is how the Geologic Time Scale is typically depicted. In past work the most commonly chosen incorrect response was a statement that is consistent with the visible spatial intervals (*The Phanerozoic lasted much longer than the Proterozoic*).

A third test item served as a control for other potential group differences (e.g. motivation). This item is a knowledge-based question, asking when mammals were the dominant land animal. This item did not require an understanding of magnitude.

## Results

On the Geoscience Concept Inventory item the intervention group was significantly less likely to make large-magnitude errors than the control group ( $\chi^2(1) = 6.08, p = .01$ ), although both groups were just as likely to choose D, the correct option ( $p > .05$ ). The intervention group was significantly less likely than the control group to choose C, the most common error ( $\chi^2(1) = 7.35, p = .01$ ).

The intervention group was more accurate than the control group on the Geologic Time Scale Diagram item

( $\chi^2(1) = 3.99, p = .05$ ). The groups did not differ significantly on the knowledge-based test item, which did not require an understanding of magnitude ( $p > .05$ ).

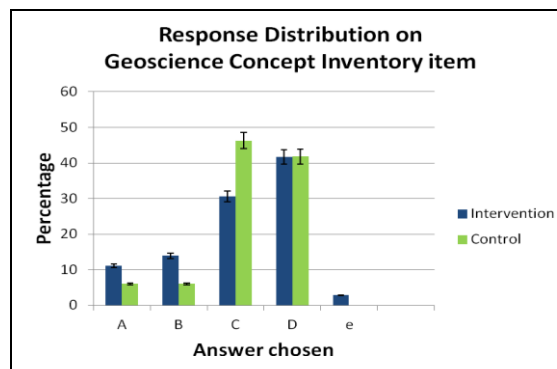


Figure 2: Distribution of student responses to Geoscience Concept Inventory item on geologic time. Response option “D” is the correct answer. Incorrect response options reflect common misconceptions: (A) life occurred when Earth formed, (B) humans and dinosaurs coexisted, (C) dinosaurs appeared much earlier than they did, (E) all life formed at the beginning of Earth's history).

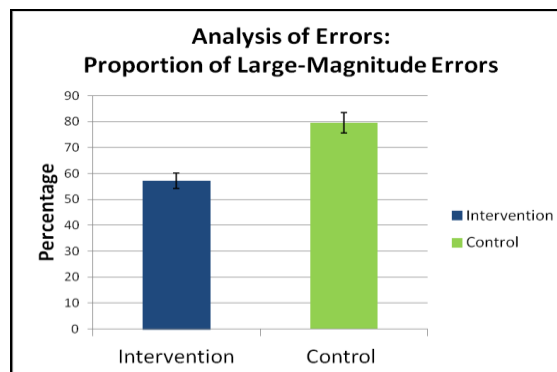


Figure 3: Percentage of students making large-magnitude errors. The incorrect response options from the Geoscience Concept Inventory item on geologic time were broken up into small-magnitude errors and large-magnitude errors

## Discussion

The current study found that the hierarchical and progressive alignment of geologic time is an effective way to reduce magnitude-based errors in understanding geologic time. The intervention group demonstrated a more accurate sense of the relative durations of geological events and a reduction in the magnitude of temporal location errors relative to the control group. Importantly, the intervention and control groups did not differ significantly on the knowledge-based item, indicating that the intervention affected understanding the magnitude of geologic time, and did not provide additional content or increase effort or motivation in the intervention group. These findings were attained one month later, suggesting a durable effect.



That the intervention, aimed at fostering a linear representation of geologic time, was successful at reducing magnitude errors suggests that mental representation of magnitude influences understanding of the Geologic Time Scale. Specifically, the increased accuracy on the Geologic Time Scale diagram item and the pattern of errors on the Geoscience Concept Inventory item are consistent with developing a more linear scale of time at large magnitudes. One limitation of the current study is the limited number of measurements assessing geologic time understanding. We are developing new assessments of scale representation that evaluate magnitude understanding in abstract numerical domains as well as spatial and temporal content domains. Such assessments will be important for further development of our understanding the nature of the role of analogical mapping in representing magnitude and scale.

The category adjustment model can be applied to any type of magnitude (e.g. space and time). Currently, there are competing accounts of the nature of the representation of magnitude information. Some researchers advocate a generalized mapping of more/less relations across dimensions (Walsh, 2003), while other researchers maintain separate or asymmetrical representations (e.g., Agrillo, Ranpura, & Butterworth, 2010). Our finding that progressive alignment helps foster a linear representation of magnitude for time, and Thompson & Opfer's (2010) work on numbers indicates that there may be commonalities between number and time magnitude representation. Hierarchical alignment could serve as a valuable technique in future research on the nature of representations of different types of magnitude information.

There are other factors besides those studied here that may contribute to the representation and comprehension of size and scale. Measurement, estimation, perspective, and proportional reasoning may all play some role in understanding size and scale (Jones & Taylor, 2009). For example, proportional reasoning is correlated with students' ability to order objects and assign correct sizes to objects (Jones, et al., 2007). Being able to conceptualize a new unit from existing units (unitizing), and then use that new unit to make comparisons or calculations, are particularly important aspects of proportional reasoning related to understanding size and scale (Lamon, 1994). Further research is needed to examine how these other factors contribute to the understanding of size and scale.

The finding from the current study has clear educational implications. While analogy is one of the most commonly used pedagogical practices (Libarkin, et al., 2007), students still continue to demonstrate difficulties in understanding size and scale information (e.g. Jones, et al., 2008; Libarkin, et al., 2005; Tretter, et al., 2006; Swarat, et al., 2010). It is possible for analogies to fail to bring about conceptual change (Brown & Salter, 2010; Duit, 1991), and even mislead students' understanding of a concept, making misconceptions hard to identify and resolve (Brown & Salter, 2010; Duit, 1991). Two obstacles faced when using analogy in representing scale information include failure of

alignment and unrelated salient features (Gentner, 1983). The hierarchical alignment intervention is specifically designed to control for these issues by keeping everything aligned except for magnitude information. Thus, the hierarchical alignment intervention may be a more effective teaching tool than current practices employing single analogical mapping exercises.

Thus far the research discussed has described representations of large unfamiliar whole numbers compared with relatively smaller more familiar whole numbers. While there has been little research examining scaling and ordering of numbers less than the integer *one*, it is hypothesized that as magnitudes become unfamiliar at both large and small scales, the representation of those magnitudes will become "fuzzy and indistinct" (Tretter, et al., 2006). For example, people also hold a limited number of conceptual categories of extremely small scales (e.g. things we can see versus things we cannot) (Jones, et al., 2008; Tretter, et al., 2006). The increased variation within categories and little to no variation across categories should result in difficulty discriminating among objects/events that are very small. However, in the case of small scales, there is an added complication with the hierarchical alignment intervention: familiar scales are not able to be hierarchically organized within the unfamiliar scales as one progresses towards smaller and smaller scales, as was the case when progressing towards larger and larger scales. A solution may be cycling back up to larger level (or levels) for each smaller scale. In any event, the implications of this complication in cognitive representations and pedagogical practices should be examined in future research.

## Acknowledgments

This research was funded in part from National Science Foundation Grant SBE-0541957 and National Science Foundation Grant SBE-1041707 that both support the Spatial Intelligence and Learning Center (SILC).

## References

- Agrillo, C., Ranpura, A., & Butterworth, B. (2010) 'Time and numerosity estimation are independent: Behavioral evidence for two different systems using a conflict paradigm', *Cognitive Neuroscience*
- American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy*. New York: Oxford University Press.
- Barghaus, K. & Porter, A. C. (2010, April). *Building aligned assessments for middle school science teachers and students*. Paper presented at the annual meeting of the American Educational Research Association., Denver, CO.
- Block, R. A. (1990). Models of psychological time. In R. A. Block (Ed.), *Cognitive models of psychological time*. Lawrence Erlbaum: Hillsdale, NJ.
- Booth, J. & Siegler, R. (2008). Numerical magnitude representations influence arithmetic learning. *Child Development*, 79(4)

- Brown, S. & Salter, S. (2010). Analogies in science and science teaching. *Advanced Physiological Education*, 34
- Dehaene, S. (2003). The neural basis of the Weber–Fechner law: a logarithmic mental number line. *Trends in Cognitive Sciences*, 7(4)
- Dehaene, S., Izard, V., Spelke, E., & Pica P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science*, 320
- Dehaene, S., & Marques, J. F. (2002). Cognitive Neuroscience: Scalar variability in price estimation and the cognitive consequences of switching to the euro. *The Quarterly Journal of Experimental Psychology*, 55(3)
- Drane, D., Swarat, S., Hersam, M., Light, G., & Mason, T. (2008). An evaluation of the efficacy and transferability of a nanoscience module. *Journal of Nano Education*.
- Duit, R. (1991). On the role of analogies and metaphors in learning science. *Science Education*, 30
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7
- Gentner, D. & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science*, 15(6)
- Hawkins, D. (1978), Critical barriers to science learning, *Outlook*, 29
- Huttenlocher, J., Hedges, L., & Prohaska, V. (1988). Hierarchical organization in ordered domains: Estimating the dates of events. *Psychological Review*, 95: 471–484.
- Jones, M. G. & Taylor, A. R. (2009). Developing a sense of scale: Looking backward. *Journal of Research in Science Teaching*, 46(4)
- Jones, M. G., Taylor, A. R., & Broadwell, B. (2009). Concepts of scale held by students with visual impairment. *Journal of Research in Science Teaching*, 46(5)
- Jones, M. G., Taylor, A., Minogue, J., Broadwell, B., Wiebe, E. & Carter, G. (2007). Understanding scale: Powers of ten. *Journal of Science Education and Technology*, 16(2)
- Jones, M. G., Tretter, T., Taylor, A., & Oppewal, T. (2008). Experienced and novice teachers' concepts of spatial scale. *International Journal of Science Education*, 30(3)
- Kadosh, R.C. & Walsh, V. (2009). Numerical representation in the parietal lobes: Abstract or not abstract? *Behavioral and Brain Sciences*, 32
- Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, 67
- Lamon, S. (1994). Ratio and proportion: Cognitive foundations in unitizing and norming. In G. J. Harel Confrey (Ed.), *The development of multiplicative reasoning in the learning of mathematics*. Albany, NY: State University of New York Press.
- Libarkin, J.C., Anderson, S.W., Dahl, J., Beilfuss, M., & Boone, W. (2005). Qualitative analysis of college students' ideas about the Earth: Interviews and open-ended questionnaires. *Journal of Geoscience Education*, 53(1)
- Libarkin, J.C., Kurdziel, J.P. & Anderson, S.W. (2007). College student conceptions of geological time and the disconnect between ordering and scale. *Journal of Geoscience Education*, 55
- Mandler, G. (1967). Organization and memory. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory*, 1: 328–372. New York: Academic Press.
- National Research Council. (2011). *A Framework for K-12 Science Education*. Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, DBASSE. Washington, DC: The National Academies Press.
- Resnick, I., Shipley, T.F., Newcombe, N., Massey, C., Wills, T. (2011, October). Progressive Alignment of Geologic Time. Talk presented at 2011 Geological Society of America Annual Meeting, Minneapolis, MN.
- Shipley, T. F. & Zacks, J., 2008, *Understanding events: From perception to action*. New York, NY, Oxford University Press.
- Speer, N. K., Zacks, J. M., & Reynolds, J. R., 2007, Human brain activity time-locked to narrative event boundaries: *Psychological Science*, 18
- Sullivan, J. & Barner, D (2010). Mapping number words to approximate magnitudes: associative learning or structure mapping? 32nd Annual Meeting of the Cognitive Science Society.
- Swarat, S., Light, G., Park, E.-J., & Drane, D. (2010). A typology of undergraduate students' conceptions of size and scale: Identifying and characterizing conceptual variation. *Journal of Research in Science Teaching*.
- Thompson, C., & Opfer, J. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development*, 81(6)
- Trend, R.D. (2009). The power of deep time in geoscience education: linking 'interest', 'threshold concepts' and 'self-determination theory'. *Studia Universitatis Babeş-Bolyai, Geologia*, 54(1)
- Trend, R.D. (2001). Deep Time Framework: a preliminary study of UK primary teachers' conceptions of geological time and perceptions of geoscience. *Journal of Research in Science Teaching*, 38 (2)
- Tretter, T. R., Jones, M. G., Andre, T., Negishi, A., & Minogue, J. (2006). Conceptual boundaries and distances: Students' and experts' concepts of the scale of scientific phenomena. *Journal of Research in Science Teaching*, 43(3)
- Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity, *TRENDS in Cognitive Sciences*, 7(11)
- Wheeling Jesuit University. (2004). Geologic time activity. Wheeling Jesuit University/NASA-supported Classroom of the Future.
- Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, 127

# The Development of Joint Belief-Desire Inferences

Hilary L. Richardson<sup>1</sup> (hrich@mit.edu), Chris L. Baker<sup>1</sup> (clbaker@mit.edu),  
Joshua B. Tenenbaum<sup>1</sup> (jbt@mit.edu), and Rebecca R. Saxe<sup>1,2</sup> (saxe@mit.edu)

Department of Brain and Cognitive Sciences<sup>1</sup> and McGovern Institute for Brain Research<sup>2</sup>

Massachusetts Institute of Technology

Cambridge, MA 02139 USA

## Abstract

Human beings infer complex mental states given very little information—a facial expression, a sarcastic tone, or even a simple behavior. Previous work suggests that adults make joint belief and desire inferences based on an actor’s path, and that these inferences are well-explained by a Bayesian framework (Baker, Saxe, & Tenenbaum, 2011). We investigate the development of this ability by assessing mental state inferences made by children ages 3-6 after watching a short movie. Our results suggest that young children spontaneously make inductive inferences about desires or preferences, and that the ability to infer belief from behavior develops between ages 3-6, and possibly throughout later childhood. We formulate three computational models that capture the developmental shift between non-representational and representational theory of mind, and show that these models capture qualitative patterns in the children’s data.

**Keywords:** theory of mind, false-belief task, Bayesian inference, cognitive development

## Introduction

As we move about the world, our actions are the observable manifestation of unobservable intentions: we act to fulfill our hopes and desires in accordance with our beliefs. Adults understand this intuitively. When a girl exclaims “I’m starving—I’m craving a piece of fruit!” and begins to search extensively for an apple in the kitchen even though a pear is in plain sight, adults can infer that the girl wants to eat a fruit, that she has a preference for apples over pears, and that she has a reasonable degree of belief that there is an apple in the kitchen. Our explanation of the girl’s action in terms of inferred beliefs and desires relies on a Theory of Mind (ToM): we understand that agents have a working representation of the world that may or may not reflect reality, that this representation is influenced by perceptual access and priors, and that this representation is the basis for subsequent behavior (Gopnik & Wellman, 1992).

If the girl believes that apples are in the fruit basket, we confidently predict she will look for one there, even if we know that the apples are actually in the cupboard. This ability is assessed by the famous “False-Belief task<sup>1</sup>” (Wimmer & Perner, 1983), on which children typically transition from failure to success between the ages of three and five (Wellman, Cross, & Watson, 2001). Most prior

work studying the development of representational ToM has used versions of the False-Belief task to ask children to predict an agent’s behavior, given the agent’s previously established beliefs and desires. In contrast, there has been less work studying how children solve the *inverse problem*: inferring beliefs and preferences from an observed behavior. Given the girl’s extensive search for a fruit in the kitchen, how do we simultaneously infer her beliefs and preferences? Both kinds of judgments—predicting actions given beliefs and desires, and inferring beliefs and desires given actions—tap similar ToM reasoning abilities. This paper aims to test whether the development of the ability to make ToM inferences parallels the transition to understanding false beliefs, and to provide evidence for a formal account of the knowledge supporting both ToM abilities.

The ability to solve this inverse problem is analogous to solving one equation with two unknowns; our natural ability to consider context, weigh in with priors, and make rational inferences enables us to come up with a good guess on questions we would otherwise not be able to answer. Studying this ability in the social domain illustrates the power of ToM to go “beyond the data” and infer multiple implicit mental states from just one observed action. Prior work by Baker et al. (2011) presented adult participants with an inverse mental-state-inference task and showed that adult mental state inferences are well-explained within a rational probabilistic inference framework. Here, we use an analogous paradigm to measure spontaneous mental state inferences made by children 3-6 years of age and assess which observed behaviors prompt mental state inference. By doing so, we are measuring children’s expectation that all parts of an action should have a sufficient explanation in terms of mental states.

If this inferential ability develops in parallel with the ability to predict behavior given a mental state, we would expect a similar shift in performance between the ages of three to five on inverse problems that require mental state inference. On the other hand, it is possible that the ability to infer mental states from sparse information develops later in life. This process not only requires the ability to take the perspective of another and maintain multiple representations of the world, but it also requires that the viewer spontaneously seeks to understand observed actions in terms of underlying beliefs and desires.

In our experiment, children watched a short 3D animation of a hungry bunny navigating a world to find and eat one of three different fruits. The bunny can take one of three paths: (1) pass the nearest, visible fruit to check around a wall to choose the fruit there (2) take a direct path to the nearest,

<sup>1</sup> “Explicit” vs. “implicit”: Surprisingly, infants succeed in looking-time paradigms tapping analogous notions of perceptual access and false belief representation (Onishi & Baillargeon, 2005).

visible fruit (3) take the longest path, passing the nearest fruit to check around the wall (in sight of the farther fruit), and then turning back to choose the near fruit. These three paths suggest different orders and degrees of preference for the three fruits, according to both our intuition and three computational models. In particular, the third path can only be fully interpreted as rational if children infer that the bunny was “looking for” the missing fruit. We analyze how many children inferred the bunny’s fruit preference, as a function of age and condition (i.e., the bunny’s path). In doing so, we believe we are assessing a sophisticated aspect of ToM reasoning: the ability to make rich inverse inferences from limited data.

## Methods

### Participants

143 children were recruited from a local children’s museum to participate in the study. Out of these, 103 were included in the final sample (70 females, 3-4yo group:  $M=4.02$ ,  $SD=0.62$ ; 5-6yo group:  $M=5.79$ ,  $SD=0.56$ ). 40 participants were excluded; 28 for answering at least one of four memory and control questions incorrectly, 5 for parent or sibling interference, 4 for not answering all of the questions, and 3 for experimenter error.

55 adults were recruited from an MIT human subjects listserv. Out of these, 54 were included in the final sample (30 females,  $M=25.95$ ,  $SD=6.14$ ). 1 participant was excluded for not following task instructions.

### Stimuli and Design

3D animated movies were created using Alice 2.2 programming software (<http://www.alice.org>). Stimuli are available at: <http://saxelab.mit.edu/bunny/>. Each participant watched one movie two times followed by a short ending.

**Movie Introduction** Each movie begins with a bunny standing on a green platform. A brown wall divides the platform—this wall reaches above the bunny’s eye level, obstructing his view of the other side. There is a tree with three different fruit on the bunny’s side of the wall. The fruit varied in shape and color (yellow, red, and orange), and position of the fruit was counterbalanced across children. At the beginning of the movie, the bunny waves and says “Hello!” He then turns to the tree, points at the three fruits (ambiguously), and says, “I’m hungry, I want that one!” He attempts to reach the fruit, and eventually succeeds in knocking all three fruits off the tree simultaneously. While the bunny is still facing the tree, the three fruit fall down and roll away—one lands in plain sight of the bunny (Fruit 3), but the other two roll to the other side of the wall. One of these fruits stays on the other side of the wall (Fruit 2), and the other (Fruit 1) falls off of the platform. In the movies viewed by children, Fruit 1 rolled out of sight. The bunny turns to get his fruit, sees Fruit 3, and takes one of three paths: Check Stay (CS), No Check (NC), or Check Turn (CT) (see below for detailed description).

This introduction was created to allow viewers to understand the bunny’s initial belief state about the world. Before he embarks on one of three paths to find a fruit, participants are provided with evidence that the bunny knows that the three fruit exist, and knows that they rolled away. They are also prompted to understand that the bunny is hungry and has a preference for one fruit over the others. Finally, participants know that the bunny has initial perceptual access to Fruit 3, and some degree of belief that the other fruits might be on the other side of the wall. These priors were similarly built into our computational models.

Between the two movie viewings, participants saw a black screen with the words “Let’s watch one more time!” for 3 seconds. The experimenter read this text out loud to child participants at this time.

**Paths** The bunny’s paths were designed to evoke inferences about the bunny’s preferences and beliefs. Below we describe each path, and note in italics the inferences that each path elicits. While we believe these inferences are intuitive, we also characterize them through three formal computational models (see Computational Models section).

**Check Stay (CS)** The bunny passes Fruit 3, walks to the other side of the wall, and chooses Fruit 2 (see Fig. 1). *Predicted Inferences: all age groups will infer that Fruit 2 is the bunny’s favorite fruit. Passing Fruit 3 is the strongest evidence from which to infer that Fruit 3 is the least favorite fruit. Fruit 1 is underspecified.*

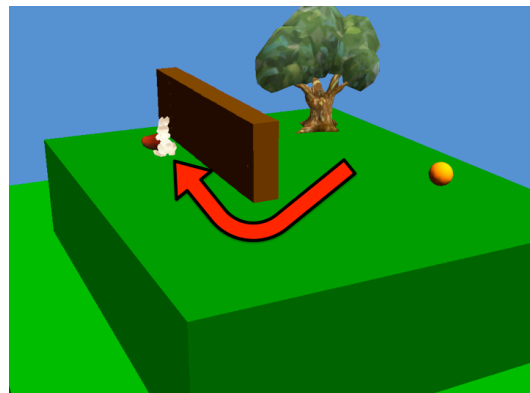


Figure 1: Final frame of Check Stay Condition (red arrow shows path).

**No Check (NC)** The bunny walks directly to Fruit 3 (see Fig. 2). *Predicted Inferences: all age groups will infer that Fruit 3 is the favorite fruit. There is no distinguishing evidence for preference order between Fruit 1 and Fruit 2, so the bunny’s least favorite fruit remains ambiguous.*

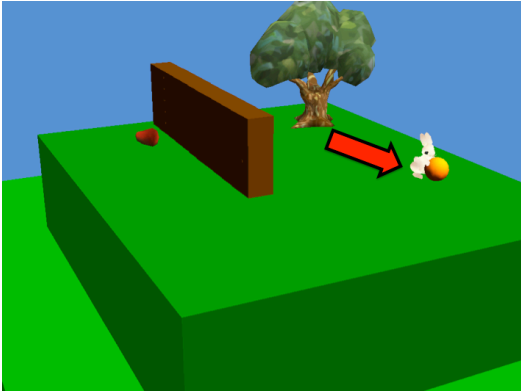


Figure 2: Final frame of No Check Condition (red arrow shows path).

**Check Turn (CT)** The bunny passes Fruit 3, walks to the other side of the wall, sees Fruit 2, and goes back to Fruit 3 (see Fig. 3). *Predicted Inferences: Participants who attribute initial uncertainty about the locations of the occluded fruits to the bunny will infer the correct preference order: the bunny’s favorite fruit is Fruit 1, and his least favorite fruit is Fruit 2. Participants who do not consider the bunny’s beliefs will infer that the bunny’s favorite fruit is Fruit 3 and his least favorite fruit is Fruit 2.*

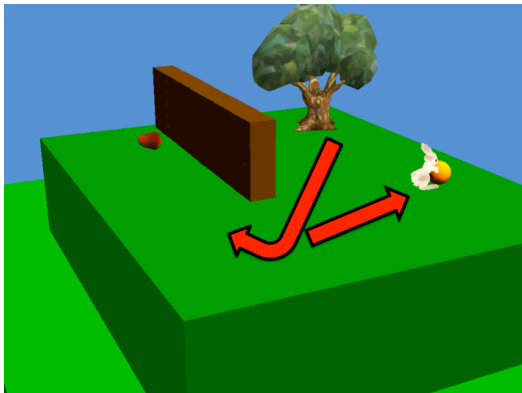


Figure 3: Final frame of Check Turn Condition (red arrow shows path).

**Movie Ending** After children viewed both the introduction and path twice, they viewed a short ending. This ending shows all three fruit roll within sight and reach of the bunny, making it clear that this time, the bunny can choose any of the fruits. All participants were shown a laminated picture of the final frame of this ending, which shows the bunny equidistant from all three fruit. This picture enabled children to respond to questions by pointing, and was placed in front of adults as they filled out the short questionnaire.

All participants sat approximately 12-18” from the movie, which was played on a 15” MacBook Pro. Each participant watched two trials of one condition for a total watching time of 3:11 minutes (CT condition), 2:48 minutes (CS condition), or 2:25 minutes (NC condition).

Adult participants were asked to rank how much the bunny liked each of the fruits on a Likert scale from 1-7 (Not Much-Very Much). They were also asked to explain their rankings (“How do you know?”). All responses were recorded by the participant using pencil and paper.

Child participants were asked two control questions, four test questions, and two memory check questions after watching the movie. The first control question asked the child to identify each fruit by pointing to it on the laminated picture; the order of fruits was counterbalanced across children. The second control question asked, “At the beginning of the movie, how many fruit did the bunny see in the tree?” Children were then asked “Which fruit do you think is the bunny’s favorite?” and “How do you know?” Next, children were asked “Which fruit do you think is the bunny’s least favorite—which one does he not like?” and “How do you know?” Finally, children were asked “During the movie, which fruit fell off and rolled away?” and “During the movie, which fruit did the bunny pick?” The control questions provided reasonable exclusion criteria and enabled us to confidently report that children were engaged by the movie and remembered its content, while the test questions provided us with sufficient information regarding preference understanding (which fruit was favorite, neutral, and least favorite). Most children responded to the favorite and least favorite questions by pointing to the laminated picture (rather than responding verbally). Child testing sessions were recorded using a video camera. Parents were usually in the testing room with the child participants, but were asked to remain quiet during the experiment.

### Computational Models

We model children and adults’ mental state inferences as a kind of probabilistic causal reasoning. One way to approach inverse problems is through Bayesian inference, which describes the process by which people generate and test hypotheses based on their expectations and evidence. The probability of a hypothesized mental state  $H$ , given observed evidence  $E$ , denoted  $P(H|E)$ , depends on  $P(E|H)$ , the likelihood of the evidence, given the hypothesis, and  $P(H)$ , the prior probability that the hypothesis is true, according to Bayes’ rule:

$$P(H|E) \propto P(E|H)P(H).$$

Posed in the context of this study: given the path that the bunny takes and the fruit that he chooses ( $E$ ), what were his initial beliefs and desires ( $H$ )?

To capture the development of the ability to represent and infer mental states, we formulated three models: Outcome-Based (OB), Desire Theorist (DT), and Bayesian ToM (BToM). The Outcome-Based model assigns full preference to the fruit that the bunny chooses; it is not influenced by the bunny’s beliefs or the path that he takes. The DT model is based on the Copy Theorist model of Goodman, et al. (2006). DT is a non-mentalistic model that represents how an agent’s desires and the world state—but not the agent’s beliefs—cause its actions, via the principle of rational action (Gergely et al., 1995). The DT model can infer



straightforward desires that do not depend on the actor's belief state. We expect this model to explain mental state inferences that derive mainly from which fruit the bunny chose. The third model is based on the BToM framework proposed by Baker et al. (2011), which models a mentalistic theory. BToM incorporates a principle of rational belief: the assumption that agents maintain beliefs about the world that are consistent with physical laws and depend on their perceptual access to the world (i.e., visual access to different fruits). This enables simultaneous inference of agents' beliefs and desires, given observations of their behavior.

The DT and BToM models assume that the bunny's actions are guided by his degree of desire (subjective value) for each fruit, which trades off against his expected costs to reach them. The bunny incurs costs for effort (quantified by the number of steps taken per path: 1.5 steps for NC, 2.75 for CT, and 3 for CS) and time (an additional cost of .25 for checking around the wall). The bunny's favorite fruit is assigned a value of 15, the second-favorite fruit a value of 5, and the least favorite fruit a value of 1. This desire scale reflects a strong preference for one fruit and is calibrated to the spatial scale of the environment; changing these approximate values does not alter the trends observed. The OB model also assumes that the bunny's actions are guided by desire, but does not assign incorporate costs or rewards; the chosen fruit is considered the bunny's favorite, and the two unchosen fruits are equally least favorite.

The BToM model attributes initial uncertainty to the bunny about the locations of the two non-visible fruit: both, one, or neither of the two fruits may be available behind the wall. In the BToM model, the bunny's beliefs are updated when he moves to the other side of the wall. The Desire Theorist model, on the other hand, assumes the bunny's actions depend only on the true world state. For BToM, costs incurred to check around the wall are rational if the bunny desires one of the two non-visible fruits and believes it could be there; in the DT model, there is no explanation of why the bunny incurred these costs.

This key difference is most evident in the Check Turn condition, which allows us to assess whether three/four-year olds, five/six-year-olds, and adults incorporate information about uncertainty, planning, and belief updating into their desire inferences. If so, their responses should be better predicted by the BToM model than the DT model. This difference is comparable to the performance shift on the False-Belief task between children who are three and five years old; three-year-old children refer to salient outcomes, actions, and desires, while five-year-old children take beliefs into consideration.

## Data Analysis

Child responses to the four control and memory questions, two preference questions, and two explanation questions were transcribed from the recording of the testing session. From these data we recorded the proportion of participants in each age group (3-4yo, 5-6yo, and adults) who reported each fruit (1,2, and 3) as "favorite" and "least favorite" in

each condition (CS, NC, CT). We binarized adult participant data; the fruit that was assigned the highest number on the Likert scale was coded as the favorite, and the fruit assigned the lowest number was coded as the least favorite. Whereas all participants picked one fruit as the favorite for each condition, participants often reported two fruits as least favorite. In this case, each fruit received half the weight assigned to a single favorite or least favorite fruit (.5).

We used logistic regression to test for main effects of age (a continuous variable), condition (CS, NC, CT), and age by condition interactions on favorite and least favorite fruit choice. We used a Bonferroni correction ( $n=3$ ) for multiple comparisons.

To compare behavioral judgments to the three models, we separated participants into three groups: younger children (age 3-4 years,  $n=56$ ), older children (age 5-6 years,  $n=46$ ), and adults ( $n=54$ ). We calculated the probability of choosing each fruit (3) as favorite or least favorite (2) in each condition (3). We compared the resulting 18 values for each group to the corresponding predictions from each model, using a Pearson's correlation.

## Results

### Check Stay (CS) Condition

This condition is a good measure of spontaneous understanding of preference; the bunny picks Fruit 2 after explicitly passing Fruit 3. We expected that children of all ages as well as adults would correctly identify Fruit 2 as the favorite, as predicted by all three models. We found a positive main effect of the CS condition on choosing Fruit 2 as the favorite fruit ( $p<0.001$ ), and a negative effect of the CS condition on choosing Fruit 2 as the least favorite fruit ( $p=0.043$ ). There were no significant effects of age on fruit choice in this condition. Fig. 4 shows that participant judgments were well predicted by the DT and BToM models across age.

### No Check (NC) Condition

This condition is an interesting measure of how participants understand preference when there is less evidence available. The bunny approaches the only visible (and closest) fruit, Fruit 3, providing weaker evidence for his preference; his choice may reflect efficiency or lack of options, rather than a strong preference. Again, all groups successfully picked Fruit 3 as the most likely favorite. We found a positive main effect of the NC behavior on choosing Fruit 3 as the favorite ( $p<0.001$ ), and a negative main effect of the NC behavior on choosing Fruit 3 as the least favorite ( $p=0.020$ ). There was no significant difference between age groups.

We were specifically interested in whether observers were sensitive to the weaker evidence in NC compared to CS paths; if so, Fruit 1 should be more likely to be chosen as least favorite in the NC condition than in the CS condition (because in the CS condition, the bunny explicitly avoided Fruit 3). Only the BToM model showed this qualitative pattern. Although this difference was not significant in any age group individually, combining across age groups did

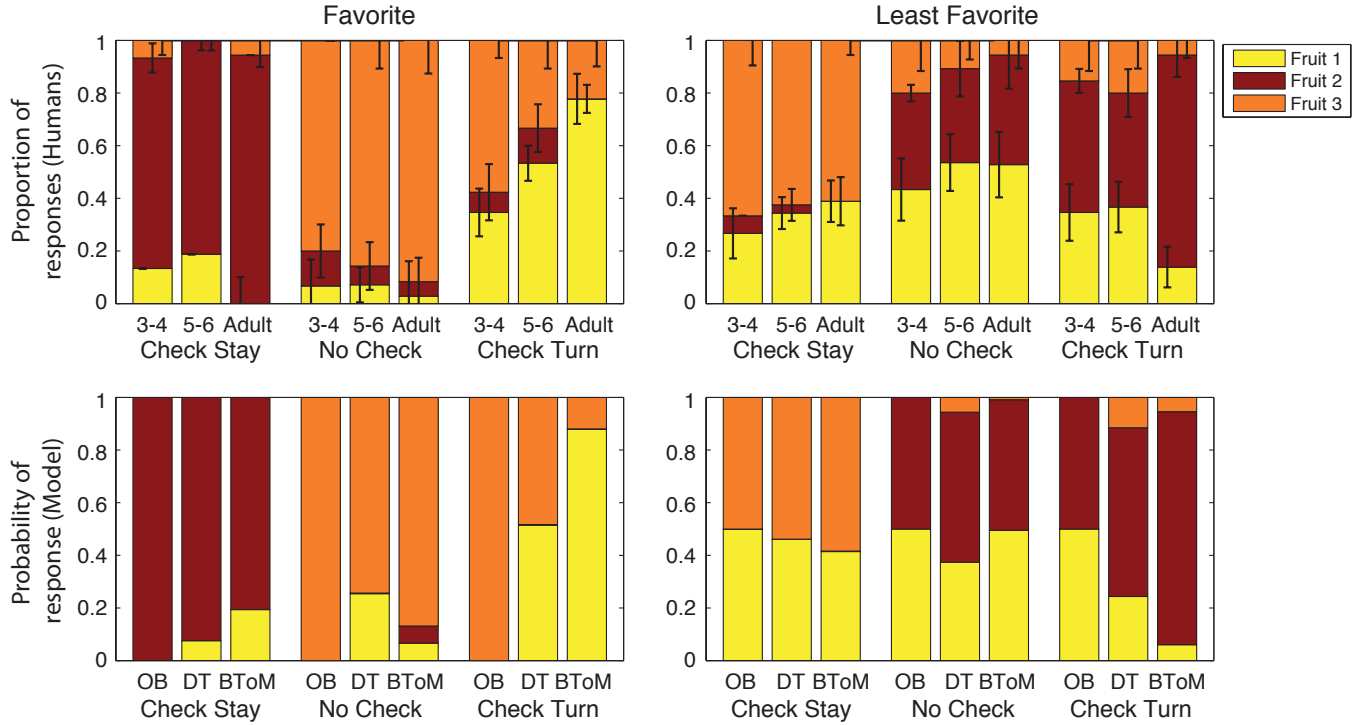


Figure 4: Results of “Favorite” and “Least Favorite” judgments across conditions and age groups. The bottom row compares the results of the OB, DT, and BToM models to human judgments, matching the qualitative developmental shift between younger and older participants.

reveal the predicted pattern (25/47 chose Fruit 1 as least favorite in NC, compared to 16/49 in CS,  $p < 0.05$ , Fisher’s exact test).

### Check Turn (CT) Condition

This condition tests spontaneous inference of both the beliefs and preferences of the agent. It requires participants to consider the agent’s actions in terms of preferences (which fruit did he really want?) and counterfactuals (what was he looking for behind the wall, that he could have seen but did not?) rather than outcome (which fruit did he pick?). Given the well-documented shift from failure to success on False-Belief tasks, we expected 3-4 year olds to perform significantly worse than 5-6 year olds and adults on the “favorite fruit” question.

We found a positive age by condition interaction on choosing Fruit 1 as the favorite fruit ( $p = 0.029$ ) and a significant negative age by condition effect on choosing Fruit 1 as least favorite ( $p = 0.021$ ). These results show that as participants got older, their judgments slowly came to better resemble the predictions of the BToM model. The salient fact that the bunny chose Fruit 3 was difficult for younger participants to ignore, as suggested by the OB and DT model predictions. As a result, there was also a positive main effect of the CT condition on choosing Fruit 3 as the favorite ( $p = 0.001$ ) and a negative main effect of condition on choosing Fruit 3 as least favorite ( $p = 0.020$ ).

### Model Comparison

Combining across the three conditions, we compared the responses of each age group to the three models. Judgments made by 3-4 year old children were most strongly correlated with the predictions of the OB model ( $r = .902$ ), followed by the DT model ( $r = .898$ ), and least with the BToM model ( $r = .719$ ). In contrast, judgments made by 5-6 year old children correlated most strongly with predictions made by the DT model ( $r = .898$ ) followed by the BToM model ( $r = .812$ ), and correlated least with the OB model ( $r = .786$ ). Adult preference predictions were most strongly correlated with the BToM model ( $r = .943$ ), followed by the DT model ( $r = .920$ ), and least correlated with the OB model ( $r = .678$ ). These data support the idea that across development, people increasingly incorporate spontaneously attributed beliefs and desires into their understanding of agents’ behavior.

### Discussion

In this study we used a novel ToM task to examine the development of the ability to spontaneously attribute mental states in order to understand an agent’s observed actions, i.e., to solve inverse social inference problems. Our results suggest that this ability emerges gradually, with performance continuing to improve after age 5-6 years.

In particular, on the critical Check Turn condition a sufficient explanation of the bunny’s behavior requires children to infer that the bunny checked behind the wall because he was looking for (and preferred) the missing Fruit 1. This inference depends on recognizing that the bunny



initially didn't know which fruit was behind the wall, and that he couldn't have been looking for Fruit 2, because he subsequently turned back to Fruit 3 (which he initially bypassed). Observers would only make this complex inference if they sought a sufficient rational explanation of the bunny's whole path. Most 3-4 year old children effectively ignored the "checking" path, inferring that the bunny prefers whichever fruit he chooses in the end. This behavior mimicked the prediction of our Outcome-Based and Desire Theorist models. Five and six year olds were somewhat more likely to take the checking path into consideration, recognizing Fruit 1 as the favorite, and adults were even more likely to do so. This created a gradual increase with age in the match between participants' choices and the predictions of the Bayesian Theory of Mind model.

One limitation of the current experiment is that we did not test children older than 5-6 years, so we cannot say whether children's performance on these tasks reaches adult levels by age 7-9 years, or whether there is extended development through adolescence. Another limitation is that children's choices for the "least favorite" fruit were frequently ambiguous (both fruits not identified as favorite were often considered least favorite). We are currently working on an extension in which children provide a full ordering of the bunny's preference for all three fruits.

The current results are amenable to multiple interpretations. First, children may be becoming more sophisticated and adept at thinking about other minds. Simultaneously inferring a belief and a desire may require a more robust ToM capacity than using beliefs and desires to predict actions. If so, these results may be related to the observation that children's ToM brain regions also become increasingly selective after age 5-6 years (Gweon, Dodell-Feder, Bedny, & Saxe, in press). Another possibility is that over development, children become more committed to the idea that agent's actions are rational, and require sufficient rational explanations. Thus, younger children may be less likely to view others' actions as efficient paths toward their goals, and so the bunny's deviant path may not seem to require any explanation. As children expect more efficiency from others, the deviant path may become more salient, and demand an explanation. Finally, a third possibility is that children's ability to focus on and interpret the bunny's path, in addition to his (highly salient) final position, depends not on ToM development per se, but on unmasking of prior competence by the development of executive function (Carlson, Moses, & Breton, 2002; Baillargeon, Scott, & He, 2010).

Another key question is how infants will perform on this task. Recent evidence from looking time suggests that infants have a nascent ToM, and can update representations of agents' beliefs and desires given their perceptual access, in change-in-location and appearance-reality tasks (for a review, see Baillargeon et al., 2010). We believe that if infants make simultaneous inferences about beliefs and desires in the current paradigm, it would be particularly

strong evidence for a mentalistic account of infant ToM abilities.

In sum, the current study is an initial step toward clarifying how we develop the ability to make joint belief-desire inferences in order to understand other minds. It contributes to the current literature on ToM development, suggesting that children develop this ability in parallel with other mentalistic reasoning abilities between the ages of three and five, and could serve as a launching point for future work studying rich social inferences made by infants.

**Acknowledgements** The authors were supported by the Packard Foundation. We thank members of the Saxe and Schulz labs for thoughtful discussion and feedback and Mika Asaba for help with data collection. Thanks also to the Children's Museum of Boston and participating families.

## References

- Baillargeon, R., Scott, R.M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, 14, 110-118.
- Baker, C.L., Saxe, R.R., & Tenenbaum, J.B. (2011). Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. In *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society* (pp. 2469-2474).
- Carlson, S.M., Moses, L.J., & Breton, C. (2002). How specific is the relation between executive functioning and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development*, 11, 73-92.
- Gergely, G., Nadasdy, Z., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, 56, 165-193.
- Goodman, N.D., Baker, C.L., Bonawitz, E.B., Mansinghka, V.K., Gopnik, A., Wellman, H., Schulz, L., & Tenenbaum, J.B. (2006). Intuitive Theories of Mind: A Rational Approach to False Belief. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 1382-1387).
- Gopnik, A. and Wellman, H. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, 7, 145-171.
- Gweon, H., Dodell-Feder, D., Bedny, M., and Saxe, R. (in press). Theory of Mind performance in children correlates with functional specialization of brain regions recruited for thinking about thoughts. *Child Development*.
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719), 255-258.
- Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development* 72, 655-684.
- Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103-128.

# Expectations About the Temporal Structure of the World Result in the Attentional Blink and Repetition Blindness

Cory A. Rieth & Edward Vul (crieth, evul@ucsd.edu)

Department of Psychology, 9500 Gilman Dr. # 109  
La Jolla, CA 92093-109 USA

## Abstract

We consider how repetition blindness and the attentional blink might arise from prior assumptions about the occurrence of task-relevant states of the world. Repetition blindness and the attentional blink are behavioral deficits in the identification of items during rapid serial visual presentation at varying delays after identifying the first “target.” Here we propose that both of these effects are explained by rational inference given prior expectations about the timing of task-relevant world transitions. While such expectations would be helpful in the natural world, they may result in unanticipated biases in laboratory settings. We show that a rational model using prior expectations of the timing of task-relevant information captures the basic repetition blindness and the attentional blink effects, and also the specific distributions of errors made during the attentional blink.

**Keywords:** Attentional blink; Repetition blindness; Attention; Computational Modeling; Bayesian Models

## Introduction

Repetition blindness and the attentional blink refer to common failures in identifying stimuli shortly after an important event. Both phenomena are typically studied in rapid serial visual presentation (RSVP) tasks, in which many stimuli (often letters or digits) are displayed in quick succession. In the task people are asked to pick out some specific “targets” from the RSVP stream (for instance, letters among digits, or letters cued by a ring). After the first target ( $T_1$ ), a second target ( $T_2$ ) appearing within a few stimuli of the first tends to be misreported – this is called the “attentional blink” (AB; Figure 1, top; Raymond, Shapiro, & Arnell, 1992; Weichselgartner, 1987). If an identical target is presented twice without much time intervening, people tend to not notice the repetition, yielding “repetition blindness” (RB; e.g., the ‘E’ in the bottom section of Figure 1; Kanwisher, 1987).

Although both RB and AB have many surface similarities, they follow a qualitatively different time-course as a function of the “stimulus onset asynchrony” (or lag) – the time between the onset of  $T_1$  and the onset of  $T_2$ . The attentional blink exhibits lag-one sparing – if  $T_2$  appears immediately after  $T_1$ , detection is often not impaired; in contrast, the RB deficit is most pronounced at such short delays. Moreover, when manipulated within one experiment, these effects have independent, dissociable effects (Chun, 1997). Several accounts of RB and AB postulate resource limitations (Shapiro, Raymond, & Arnell, 1997) in individuating items into unique tokens and binding them to type identities (Kanwisher, 1987; Bowman & Wyble, 2007). Moreover, the time-course of these effects has been the target of several computational process models that utilize specific mechanistic dynamics of attention

and memory (Bowman & Wyble, 2007; Olivers & Meeter, 2008).

Here we do not aim to supplant these models, but only to provide a rational account of these dynamics, to explain why the visual system may exhibit AB and RB. We remain agnostic to the particular processes and implementations, and instead seek an understanding of the computational principles that may be in play. We develop a probabilistic model of target identification by *co-occurrence detection*, which aims to infer which item occurred simultaneously with the cue, given temporal uncertainty and expectations about world dynamics. First we describe the derivation of the model based on assumed temporal statistics, then illustrate its application to the basic RB and AB paradigms. Finally, we will apply this model to AB response distributions.

## Attentional Blink



## Repetition Blindness

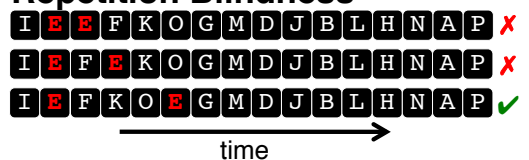


Figure 1: Repetition blindness and attentional blink paradigms. Each row of letters represents a particular type of trial where the letters appear in rapid succession. Repetition blindness arises when a target repeats, and participants indicate whether it appeared once or twice: performance is worse at short inter-target intervals when the two targets are identical compared to when they are different. The attentional blink occurs when the task requires identifying two designated targets: Participants can report items in quick succession, as in the top attentional blink row (“lag-one sparing”); however, when the second target appears a few items later, participants are less accurate.

## RSVP Identification Model

We assume that people identify targets in an RSVP task via the probability that a given item co-occurred with the cue. Under this assumption, the observer’s task is to infer the letter identity ( $\mathcal{L}$ ) of the second target ( $T_2$ ), by marginalizing over all possible stimuli used in the experiment ( $x_i$ ) that may have coincided with the cue. The probability that the cue

co-occurred with item  $i$  at time  $t$  is the product of the probability that the item ( $\mathbb{X}_i(t)$ ) and cue ( $\mathbb{C}(t)$ ) were present at that time. This quantity is integrated over time to obtain an unnormalized probability of item  $i$  co-occurring with the cue. The core of our model is an assumption that participants use prior expectations about the times when salient or task-relevant events are likely to occur ( $P(\text{any}|t)$ ). The probability of co-occurrence at a given time is thus weighted by the probability the stimulus at time  $t$  will be relevant for the task. Thus, the participant's subjective probability distribution over possible identities of the second target is given by:

$$P(Y(T_2) = \mathcal{L} | S, \Omega) \propto \sum_i P(x_i = \mathcal{L}) \int_{t=0}^{\infty} \mathbb{C}(t) \mathbb{X}_i(t) P(\text{any}|t), \quad (1)$$

where  $P(x_i = \mathcal{L})$  is 1 when  $x_i$  is  $\mathcal{L}$ , and 0 otherwise. To distinguish between the letter identity, or *type*, and a particular instance of a letter, or the *token*, we use  $T_2$  to denote the token of the second target, and  $Y(T_2)$  to denote its type.<sup>1</sup>  $S$  refers to the stimulus sequence of the trial, including the particulars of which items occurred at which points in time, and when the cue occurred.  $\Omega$  includes the set of parameters which our model uses to describe subjective uncertainty and expectations about the task. The following sections describe the parameters used to calculate  $\mathbb{C}(t)$ ,  $\mathbb{X}_i(t)$ , and  $P(\text{any}|t)$ .

### Perceptual Likelihood

In RSVP tasks, participants often misreport item identities (Vul & Rich, 2010; Vul, Hanus, & Kanwisher, 2009) and their presentation order (Wyble, Potter, Bowman, & Nieuwenstein, 2011), indicating that even without RB and AB, there is considerable uncertainty about the relative timing of cues and items. The cue-function ( $\mathbb{C}(t)$ ) and the item-function ( $\mathbb{X}_i(t)$ ), account for this temporal uncertainty: they represent the subjective probability that these stimuli were presented at a given point in time. Both of these functions are obtained by convolving some temporal uncertainty kernel with the physical time series of stimulus presentation. We use the Student's t-distribution as the uncertainty kernel.<sup>2</sup>

Thus, the cue-function ( $\mathbb{C}(t)$ ) and the item-function ( $\mathbb{X}_i(t)$ ), are defined as a boxcar function – representing the physical presence of the stimulus – convolved with a t-distribution – representing perceptual uncertainty about the precise time of stimulus onset and offset:

$$\mathbb{C}(t) = P(\text{cue on}|t) = \text{boxcar}(t | \text{onset}, \text{offset}) \otimes \text{Student}(t | \sigma_c, v_c),$$

<sup>1</sup>In general,  $P(Y(x_i) = \mathcal{L})$  could capture perceptual uncertainty about stimulus identity (to reflect that some letters are more confusable than others), but here we ignore this complication, and assume no pairwise letter confusion, by assigning probability 1 to the actual identity of  $x_i$ , and 0 to all other alternatives.

<sup>2</sup>Our qualitative results do not much depend on the specific functional form of the temporal uncertainty kernel – simple Gaussian distributions are sufficient – but our numerical results are more consistent with human behavior when using a heavy-tailed distribution, like the t-distribution.

where ( $\otimes$ ) is the convolution operator, the boxcar function is 1 between the onset and offset of the cue, and 0 otherwise, and  $\text{Student}(t | \sigma_c, v_c)$  is a scaled Student's t-distribution with mean 0,  $v_c$  degrees of freedom, and standard deviation  $\sigma_c$ . This convolution operation effectively captures uncertainty about the exact onset and offset of the cue. Similarly, the item function can be written as:

$$\mathbb{X}_i(t) = P(x_i | t) = \text{boxcar}(t | \text{onset}, \text{offset}) \otimes \text{Student}(t | \sigma_x, v_x).$$

In the simulations presented later  $v_x = v_c = 2$ ,  $\sigma_x = 30$ , and  $\sigma_c = 75$ . The larger uncertainty for the cue reflects the fact that the cue in a RSVP stream is a rare and unpredictable occurrence relative to the steady stream of items.

### Prior for Task-Relevant Events

Both RB and AB are conditioned on correct detection/identification of the first target ( $T_1$ ) in an RSVP stream. Therefore, we consider expectations about whether any task-relevant stimulus is likely to be visible at time  $t$  after the onset of the  $T_1$ . We write this as  $P(\text{any}|t)$ . Since  $T_1$  is, by definition, task-relevant,  $P(\text{any}|t)$  decomposes into the mutually exclusive probabilities that old task-relevant information is still visible ( $P(\text{old}|t)$ ) and that new task-relevant information has appeared ( $P(\text{new}|t)$ ):

$$P(\text{any}|t) = P(\text{old} \cup \text{new}|t) = P(\text{old}|t) + P(\text{new}|t). \quad (2)$$

To obtain these time-varying probability functions, we must assume a transition distribution for the world, that is: what is the probability distribution of the interval between changes in the world? We write this distribution as  $P(R_1 = t)$ , the probability that the first transition ( $R_1$ ) occurred at time  $t$ . This distribution reflects implicit beliefs about the temporal structure of the world (we will later discuss different choices of this transition distribution). Figure 2 illustrates the derivation of the different elements of  $P(\text{any}|t)$ , from  $P(R_1 = t)$ . Based on this transition distribution, the probability that an old task-relevant stimulus is still visible at time  $t$  is given by the probability that the first transition has not yet happened – that is, the probability that the first transition will happen at a time later than  $t$ :

$$P(\text{old}|t) = P(R_1 > t) = 1 - \int_{t'=0}^t P(R_1 = t') \quad (3)$$

To find the probability that a new task-relevant stimulus has appeared by a given point, we must take into account that not every transition in the world produces a task-relevant stimulus. We define the parameter  $\theta$  as the probability that a transition yields a task-relevant stimulus, and define  $P(\text{new}|t)$  as the sum of a convolution series of transitions. The probability that the  $n$ th transition will occur at time  $t$  ( $P(R_n = t)$ ) is computed as the  $n$ th convolution power of the transition dis-

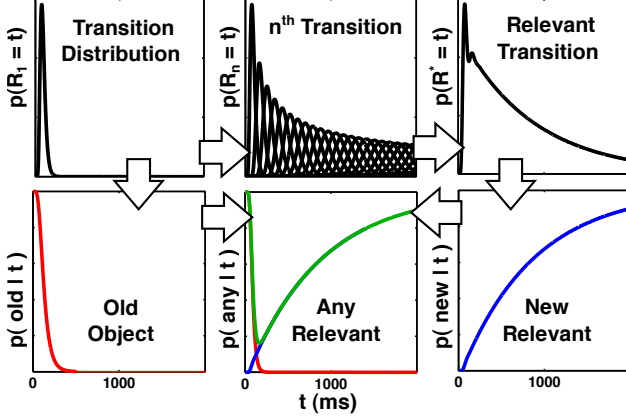


Figure 2: Construction of the prior probability of any task-relevant information at time  $t$  from an assumed transition distribution between world states (see text).

tribution (where  $\otimes$  is the convolution operator):

$$P(R_n = t) = P(R_1 = t) \otimes^n = \underbrace{P(R_1 = t) \otimes P(R_1 = t) \otimes \dots \otimes P(R_1 = t)}_n. \quad (4)$$

Since each transition yields a task-relevant stimulus with probability  $\theta$ , the probability that the first task-relevant stimulus ( $R^*$ ) appears at time  $t$  is given by the probability of any number of non-task-relevant transitions occurring followed by a relevant observation (in other words, the sum of the geometrically weighted convolution series of transitions):<sup>3</sup>

$$P(R^* = t) = \sum_{n=1}^{\infty} \theta(1 - \theta)^{n-1} P(R_n = t) \quad (5)$$

The probability that a new item has appeared at some point before time  $t$  – that is,  $P(\text{new}|t)$  – can be calculated by evaluating the cumulative distribution of  $P(R^* = t)$ :

$$P(\text{new}|t) = P(R^* \leq t) = \int_{t'=0}^t P(R^* = t') \quad (6)$$

### Choice of Transition Distribution

What might be a reasonable form of  $P(R_1 = t)$ ? If we consider world changes to be the result of saccades,  $P(R_1 = t)$  would be a distribution of inter-saccadic intervals, which is well-approximated by a log-normal distribution (Wang, Freeman, Merriam, Hasson, & Heeger, 2012). Alternatively, assuming transitions in the world follow a Poisson process will yield an exponential distribution of transition times. Another approach is to consider transition times to be normally distributed around the rate of the RSVP task itself. This would be the case if subjects were noisily calibrated to the RSVP stream used, which has a regular structure. The insets of Figure 3 illustrate these three possibilities. The larger graphs

<sup>3</sup>In practice, we evaluate this infinite sum numerically by truncating higher order elements of the series (dropping  $ns$  for which the probability of having not yet transitioned to a task-relevant state  $((1 - \theta)^{n-1})$  is less than .001).

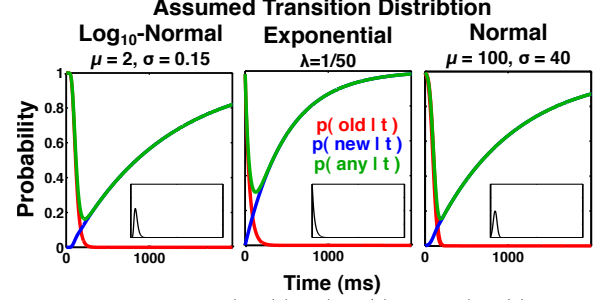


Figure 3: The resulting  $P(\text{old}|t)$ ,  $P(\text{new}|t)$ , and  $P(\text{any}|t)$  from several theoretically-motivated possible transition distributions. The probability density of each assumed transition distribution is illustrated in the insets.

show, in green,  $P(\text{any}|t)$  from each of these options. These resulting priors on a task-relevant stimulus being visible (in green) are similar for all three possible transition distribution functions. At short delays from the first target, the probability of a task-relevant object is initially high, since the first transition is unlikely to have yet occurred. However, after the first transition, with  $\theta < 1$ , a new object is not certain to be task-relevant; thus the probability that a new task-relevant item has appeared remains low. With more time, the probability of observing a novel task-relevant item increases. This combination results in the initial dip and later recovery of the prior on task-relevance – this shape of  $p(\text{any}|t)$  is the backbone of our model.

Throughout the results sections, we use the log<sub>10</sub>-normal distribution with  $\mu = 100$  and  $\sigma = .15$ , and  $\theta = .1$ . A higher  $\theta$  results in a smaller dip in the resulting prior, and a smaller RB or AB effect; decreasing  $\mu$  results in a faster time-course, while increasing  $\mu$  results in a slower time-course.<sup>4</sup>

### Modeling Repetition Blindness

In an RB paradigm, participants must decide whether a given stimulus was presented once or twice. In other words, they must decide if there were two tokens of the same type, or just one token (e.g., one or two “E”s in Figure 1). In our notation, the participants challenge is to determine if the two identified targets correspond to only one token ( $T_1 = T_2$ ). This posterior belief is given by:

$$P(T_1 = T_2 | S, \Omega) \propto \begin{cases} \Upsilon(T_1) = \Upsilon(T_2) & \int_{t=0}^{\infty} \mathbb{C}(t) \frac{P(\text{old}|t)}{P(\text{old}|t) + P(\text{new}|t)} \\ \Upsilon(T_1) \neq \Upsilon(T_2) & 0 \end{cases} \quad (7)$$

In words: if the types of  $T_1$  and  $T_2$  are different, then they must be different tokens; if their types are the same, then the

<sup>4</sup>To the best of our knowledge, any transition distribution defined on the support of  $(0, \infty)$  will produce qualitative results similar to RB and AB. This is because  $P(\text{new}|t)$  is defined as the cumulative distribution of a series of convolutions of  $P(R_1 = t)$ , making the prior robust to the specific choice. The time-course of the RB/AB effects will be determined by the median of the transition distribution. In our case, anything with a median of about 100 seems to yield a qualitatively appropriate time course.

probability of them being the same token is given by the probability that the participant perceived the cue before a transition from  $T_1$  was expected.

We assume that participants adopt a soft-max response strategy (effectively sampling from this distribution); thus the expected frequency of correct detections of a repetition is  $P(T_1 \neq T_2 | S, \Omega) = 1 - P(T_1 = T_2 | S, \Omega)$ . The left graph of Figure 4 shows simulation results for a generic RB task with 40 ms targets and 40 ms gaps between items. Overall the qualitative pattern of data is correct. There is no “lag-one sparing”, and the probability of a correct response increases with greater delays.

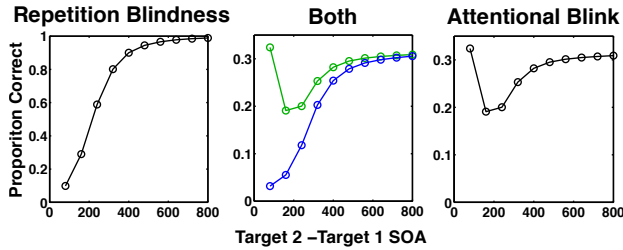


Figure 4: Model results for the RB, AB, and combination tasks. RB accuracy corresponds to the probability that the target item is a new token. AB accuracy is the probability that the cue is inferred to coincide with the second target item. The middle plot shows the combination of both effects, as in Chun, (1997).

### Modeling the Attentional Blink

In the attentional blink task, the participant needs to determine the identity of the stimulus co-occurring with the cue. Therefore the posterior distribution that the correct  $\mathcal{L}$  was  $P(Y(T_2))$  is given directly by Equation 1. Responses from the posterior are again considered to follow a soft-max rule. That is, the relative frequency of reporting the item  $\mathcal{L}$  as the second target is  $P(Y(T_2) = \mathcal{L} | S, \Omega)$ . The right panel in Figure 4 shows that the time-course of  $T_2$  accuracy matches the typical AB effect: Accuracy for target detection just after  $T_1$  is high (lag-one sparing), and then drops off, recovering after a few hundred milliseconds. The center panel illustrates simultaneous AB and RB effects using the same task and model structure, as has been observed behaviorally (Chun, 1997). Here we can account for both effects as a natural consequence of temporal uncertainty and expectations. While capturing of the qualitative patterns of the AB and RB results is interesting, there are many models that can fit this data (Bowman & Wyble, 2007; Olivers & Meeter, 2008). We now turn to the specific error distributions in the attentional blink to test finer-grained predictions of our model.

### Error Distributions in the Attentional Blink

Our model has the fidelity required to make not only predictions of changes in accuracy as a function of inter-stimulus interval in the attentional blink, but also the error distributions – when  $T_2$  is misreported, which items are reported instead? Vul, Nieuwenstein, and Kanwisher (2008) presented subjects with an RSVP stream containing all 26 English letters at 12 items/sec. Two letters were cued as targets by flashes of an

annulus around the stream (as illustrated in Fig 1). Participants were asked to report the identities of the letters that appeared simultaneously with the cues, in order. Since each english letter appeared only once in the RSVP stream, the authors could identify the exact serial position where the reported letter occurred on each trial; thus they could determine exactly which items, relative to  $T_2$ , tended to be reported in its place (Figure 5). The attentional blink can be decomposed into three changes in error distributions that follow different time-courses (Figure 6): (1) At  $T_1 - T_2$  lags of 100 to 500 ms, participants tend to make more random guesses: their responses are less likely to come from a window around  $T_2$ ; (2) At lags between 50 and 400 ms, responses tend to be more highly dispersed around  $T_2$ : participants are more likely to report letters several items away from  $T_2$ ; (3) At lags shorter than 300 ms, items *preceding*  $T_2$  tend to be misreported in its place, but after 400 ms (and lasting to lags as long as several seconds) errors instead come from items *following*  $T_2$ . We now test whether these three effects and their unique time-courses can also be accounted for by our model based on expectations about temporal structure of the environment.

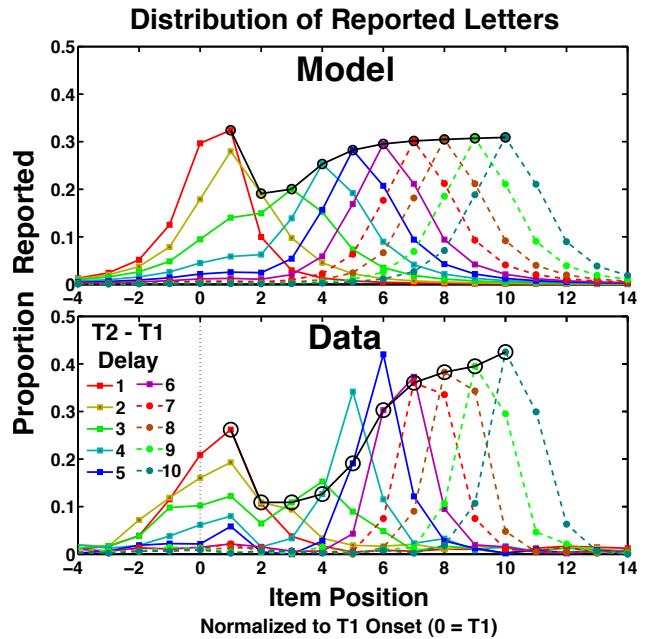


Figure 5: Response distributions for cues appearing at various positions. Behavioral data are replotted from (Vul et al., 2008). Each curve is the distribution of responses from a second cue appearing at a particular SOA relative to the first target. The dark curves show the correct responses that make up the basic AB effect.

### Error Distributions in the Model

To model error responses, we calculate the model’s prediction about the probability of each possible letter identity  $\mathcal{L}$  being reported (following Equation 1). These error distributions are presented in Figure 5. Each colored curve plots the distribution of responses for a particular delay between  $T_1$  and  $T_2$  as a function of the reported item position relative to  $T_1$ . During the attentional blink, the distribution of responses is distorted by the prior about when task-relevant items occur.

The most notable effect is the reduction in accuracy itself: the items corresponding to  $T_2$  itself are less likely to be reported at  $T_1 - T_2$  lags between 150 to 500 ms. Critically, not only does the model match the accuracy time-course of human data, but it also captures the changes in error distributions. At short delays between  $T_1$  and  $T_2$ ,  $T_2$  errors tend to come from items preceding  $T_2$ , and this pattern reverses at longer delays such that items following the second target are more common intrusions instead. Finally, our model also demonstrates increased variability in which items are reported during the attentional blink. These changes in error distributions are quantified identically to the behavioral data (Vul et al., 2008), and presented in Figure 6.

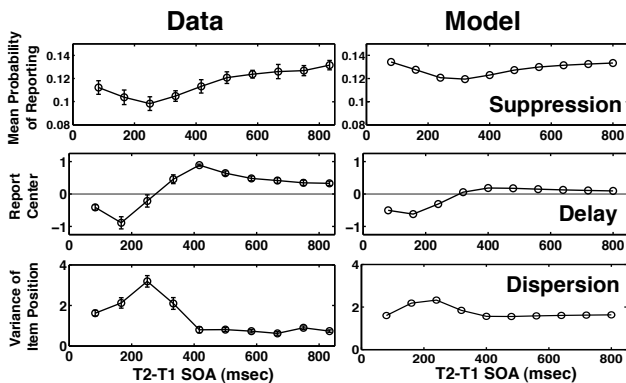


Figure 6: Summary statistics of  $T_2$  response distributions during the attentional blink, as a function of  $T_1 - T_2$  delay (stimulus onset asynchrony; SOA). Each statistic is computed within a seven item window centered on the position of the second target. Top row: “Suppression” refers to a decrease in the average probability of reports (y-axis) of items within the seven item window around  $T_2$ . Middle row: “Delay” refers to the tendency to report items preceding  $T_2$  at short SOAs, and items following  $T_2$  at long SOAs; the y-axis is the mean position (relative to  $T_2$ ) of reported items within the window. Bottom row: “Dispersion” refers to the increased variance of reported positions within the window around  $T_2$  at SOAs around 150 to 300 msec.

## Discussion

We proposed an account of perception in RSVP paradigms based on probabilistic inference about the co-occurrence of cues and items under expectations about the transitions of interesting items in the world. Our model captures patterns of accuracy in RB and AB, as well as the specific error distributions in AB. The core of this account is the premise that not every world transition brings an important item into view, and a prior constructed based on this premise produces qualitatively similar results regardless of the specific choice of transition distribution.

Why would participants use this prior, since it seems to result in performance deficits? We can only speculate that performance deficits in laboratory RSVP paradigms do not reflect the benefits that such a prior could confer to people in a more natural environment. Recent work has suggested a similar account of inhibition of return (Posner & Cohen, 1984) and immediate priming (Huber, 2008), by demonstrating that the time-courses of these effects are sensitive to learned tem-

poral properties of the environment (Rieth, 2012).

Our account captures a number of other RSVP phenomena. First, because the prior is defined with respect to time, not serial position, the dynamics of lag-one sparing and the detriment to  $T_2$  reports should be constant in time, as has been experimentally shown (Wyble & Bowman, 2005; Vul et al., 2008). Second, because the relative delay of  $T_2$  reports modulates  $T_2$  accuracy, pre-cueing should improve  $T_2$  accuracy, as is the case (Nieuwenstein, Chun, van der Lubbe, Hooge, 2005; Vul et al, 2008). Third, because we assume that perception of particular targets are driven by a *sampling* process (Vul, Hanus, & Kanwisher, 2009), perception of  $T_2$  accuracy will be all-or-none, which appears to be the case (Sergent & Dehaene, 2004). Finally, since we explicitly consider the transition probability from one event to another, our account is consistent with results that the attentional blink facilitates item individuation (Wyble, Bowman, & Nieuwenstein, 2009).

A number of other findings that are not adequately captured by our account provide future challenges. Because we only consider perception of  $T_2$  contingent on  $T_1$  identification, we cannot account for “spreading the sparing” – the finding that multiple cues presented during the “lag-one sparing” interval extend the sparing window to later items (Olivers, van der Stigchel, & Hulleman, 2007) – or whole report errors and order confusion more generally (Wyble, et al., 2009; Nieuwenstein & Potter, 2006). Another challenge is the independence of the *types* of items and the assumed dynamics of the world; because of this independence, our account cannot capture the fact that sometimes words, and sometimes letters, may be treated as the relevant items of interest (Kanwisher & Potter, 1990). Furthermore, additional assumptions would be needed for our account to be consistent with findings that missed second targets in the attentional blink produce semantic priming (Shapiro, Driver, Ward, & Sorensen, 1997). Additionally, there are some aspects of human performance for which the model is quantitatively off. Overall, suppression is smaller in the model (compare the graphs in the top row of Figure 6). For longer SOAs there is less delay in the model, slightly more dispersion, and a less extreme peak in dispersion. Furthermore, in the data – but not the model – the modal response for  $T_2$  at SOAs of four to six items is actually an incorrect answer. Finally, in the AB effect, accuracy for the item just after the first cue is too high relative to other positions.

Despite these challenges, we are encouraged that our rational inference account is consistent with classic type-token theories (Kanwisher, 1987; Chun, 1997): the discrepancy between identifying, and individuating, items in an RSVP stream seems to be of central importance to combining and distinguishing between repetition blindness and the attentional blink, and our account formalizes this intuition in probabilistic inference. Furthermore, the mechanistic dynamics of attention postulated in process models of these phenomena match our prior about task transitions (Bowman & Wyble, 2007; Olivers & Meeter, 2008). We hope that future work



might provide a synthesis of these accounts.

One promising direction for future work is to extend our model to include the perception of  $T_1$ . Because our objective was to demonstrate that RB and AB effects can result from rational use of expectations about world transitions, our current account is conditioned on the identification of the first target; however, a complete model of RSVP perception must include  $T_1$ . This extension might be achieved via an online inference process to account for changing beliefs over the time course of a trial, capitalizing on particle filter approximate inference algorithms (Doucet, Freitas, & Murphy, 2000) and expected duration hidden Markov models (Rabiner, 1989). With an online inference process the model could capture whole report paradigms (Nieuwenstein & Potter, 2006) and might explain the effect of “spreading the sparing” (Olivers, et al., 2007). Moreover, such an extension might help connect our computational level (Marr, 1982) account to previous models of the attentional blink (Bowman & Wyble, 2007; Olivers & Meeter, 2008) by postulating approximate inference algorithms, and their possible neural implementations (Fiser, Berkes, Orbán, & Lengyel, 2010; Vul & Pashler, 2008; Moreno-Bote, Knill, & Pouget, 2011).

## Conclusions

Our results suggest that the attentional blink and repetition blindness are both consequences of rational inference about cue-item co-occurrence given a prior about the rate of transitions to task-relevant items. This framework accounts for the attentional blink and repetition blindness effects as well as the error distributions of in the attentional blink paradigm. The difference between these tasks under our framework supports the intuition that repetition blindness and the attentional blink are two sides of the same coin: both are consequence of rational inference under identical assumptions, but with the observer asked to identify types, or distinguish tokens.

**Acknowledgments:** EV and CR were supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

## References

Bowman, H., & Wyble, B. (2007). The simultaneous type, serial token model of temporal attention and working memory. *Psychological Review*, 114(1), 38-70.

Chun, M. M. (1997). Types and tokens in visual processing: A double dissociation between the attentional blink and repetition blindness. *Journal of Experimental Psychology: Human Perception and Performance*, 23(3), 738-755.

Doucet, A., Freitas, N. D., & Murphy, K. (2000). Rao-Blackwellised particle filtering for dynamic Bayesian networks. *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence*, 176-183.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, 14(3), 119-130.

Huber, D. E. (2008). Immediate priming and cognitive aftereffects. *Journal of Experimental Psychology: General*, 137(2), 324-347.

Kanwisher, N. (1987). Repetition Blindness: Type recognition without token individuation. *Cognition*, 27(2), 117-143.

Kanwisher, N. G., & Potter, M. C. (1990). Repetition blindness: levels of processing. *Journal of experimental psychology. Human Perception and Performance*, 16(1), 30-47.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Freeman.

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, 108(30), 12491-12496.

Nieuwenstein, M. R., Chun, M. M., van der Lubbe, R. H. J., & Hooze, I. T. C. (2005). Delayed Attentional Engagement in the Attentional Blink. *Journal of Experimental Psychology: Human Perception and Performance*, 31(6), 1463-1475.

Nieuwenstein, M. R., & Potter, M. C. (2006). Temporal Limits of Selection and Memory Encoding A Comparison of Whole Versus Partial Report in Rapid Serial Visual Presentation. *Psychological Science*, 17(6), 471-475.

Olivers, C. N. L., & Meeter, M. (2008). A boost and bounce theory of temporal attention. *Psychological Review*, 115(4), 836-863.

Olivers, C. N. L., van der Stigchel, S., & Hulleman, J. (2007). Spreading the sparing: Against a limited-capacity account of the attentional blink. *Psychological Research*, 71(2), 126-139.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. *Attention and performance X: Control of language processes*, 32, 531-556.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257-286.

Raymond, J., Shapiro, K. L., & Arnell, K. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink?. *Journal of Experimental Psychology. Human Perception and Performance*, 18(3), 849-860.

Rieth, C. A. (2012). Adaptations of temporal dynamics: Faces, places, and words. Doctoral dissertation, UCSD

Shapiro, K. L., Raymond, J., & Arnell, K. (1997). The attentional blink. *Trends in Cognitive Sciences*, 1(8), 291-296.

Sergeant, C., & Dehaene, S. (2004). Is Consciousness a Gradual Phenomenon? Evidence for an All-or-None Bifurcation During the Attentional Blink. *Psychological Science*, 15(11), 720-728.

Shapiro, K., Driver, J., Ward, R., & Sorensen, R. E. (1997). Priming from the Attentional Blink: A Failure to Extract Visual Tokens but Not Visual Types. *Psychological Science*, 8(2), 95-100.

Vul, E., & Pashler, H. (2008). Measuring the Crowd Within. *Psychological Science*, 19(7), 645-647.

Vul, E., & Rich, A. N. (2010). Independent sampling of features enables conscious perception of bound objects. *Psychological Science*, 21(8), 1168-1175.

Vul, E., Hanus, D., & Kanwisher, N. (2009). Attention as inference: Selection is probabilistic; responses are all-or-none samples. *Journal of Experimental Psychology: General*, 138(4), 546-560.

Vul, E., Nieuwenstein, M., & Kanwisher, N. (2008). Temporal selection is suppressed, delayed, and diffused during the attentional blink. *Psychological Science*, 19(1), 55-61.

Wang, H., Freeman, J., Merriam, E., Hasson, U., & Heeger, D. J. (2012). Temporal eye movement strategies during naturalistic viewing. *Journal of Vision*, 12, 1-27.

Weichselgartner, E. (1987). Dynamics of automatic and controlled visual attention. *Science*, 238(4828), 778-780.

Wyble, B., & Bowman, H. (2005). Computational and experimental evaluation of the attentional blink: Testing the simultaneous type serial token model. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, 2371-2376.

Wyble, B., Bowman, H., & Nieuwenstein, M. R. (2009). The attentional blink provides episodic distinctiveness: Sparing at a cost. *Journal of Experimental Psychology. Human Perception and Performance*, 35(3), 787-807.

Wyble, B., Potter, M. C., Bowman, H., & Nieuwenstein, M. (2011). Attentional episodes in visual perception. *Journal of Experimental Psychology. General*, 140(3), 488-505.



# Relating Activity Contexts to Early Word Learning in Dense Longitudinal Data

Brandon C. Roy

The Media Laboratory  
Massachusetts Institute of Technology  
bcroy@media.mit.edu

Michael C. Frank

Department of Psychology  
Stanford University  
mcfrank@stanford.edu

Deb Roy

The Media Laboratory  
Massachusetts Institute of Technology  
dkroy@media.mit.edu

## Abstract

Early word learning is contingent on linguistic input, but a child's linguistic experience is also embedded in the larger, natural structure of everyday life at home. We investigate the activity structure of life in the home of one young child, and link this structure to the child's early word learning. Our analysis is based on the dense, naturalistic, longitudinal corpus collected for the Human Speechome Project. To study activity structure, we apply probabilistic topic modeling techniques to the corpus. The emergent topics capture not only linguistic structure, but also spatial and temporal regularities indicative of coherent activity contexts. We consider the child's word learning with respect to caregiver word usage frequency and word distributions across activity contexts. We find that frequency and consistency of use across context are predictive of age of acquisition. Words that are used more frequently and in more contextually constrained settings are learned earlier, suggesting that activity contexts may be an important aspect of the child's natural learning environment and worthy of further study.

**Keywords:** Language acquisition; word learning; non-linguistic context; topic modeling.

## Introduction

Children's early word learning is a remarkable achievement, the result of powerful learning processes unfolding in the natural setting of a child's first years of life. Cultural and individual variability in children's early environments has led researchers to question the contributions of the child's innate faculties relative to the role of the environment. But to the extent that children are *learning* language, the environment must provide appropriate conditions for learnability: There must be some consistent underlying structure for learning mechanisms to build upon.

In lexical development in particular, the linguistic environment—what words a child hears, and how often—provides essential input for the young learner. Yet the child's natural environment consists of other dimensions in addition to language: spatial, physical and social dimensions, to name a few. Learners are exposed to their input in the rich, multimodal domain of everyday experience. In this work, we begin to investigate the activity structure of day-to-day life and its contributions to early word learning. Based on the idea that words and referents are more predictable in sufficiently constrained situations, we hypothesize that words associated with a limited range of recurrent activities will tend to be learned earlier. That is to say, consistent lin-

guistic input across a narrower range of activities poses a simpler learning problem.

The effect of overall linguistic input on lexical development was investigated by Huttenlocher, Haight, Bryk, Seltzer, and Lyons (1991). They were the first to document positive correlation between the quantity of child-directed speech and a child's vocabulary size and growth rate. For individual words, increased frequency of use was also tied to earlier acquisition of those words; our own (Roy, Frank, & Roy, 2009) and other (Goodman, Dale, & Li, 2008) findings replicate this pattern. In addition to frequency, words presented in single word utterances (Brent & Siskind, 2001) and with prosodic stress (Echols & Newport, 1992; Vosoughi, Roy, Frank, & Roy, 2010) are also acquired earlier.

In addition to studying linguistic input, work in cross-situational word learning has investigated how words can be linked to referents through their consistent co-occurrence across a range of situations. In the face of referential uncertainty, a learner sensitive to the statistics of which words and referents co-occur can learn correct word-referent pairings (Yu & Smith, 2007). But the idea of learning by gradually accumulating word-referent co-occurrences was challenged by Medina, Snedeker, Trueswell, and Gleitman (2011), on the grounds that the sheer number of possible pairings in everyday experience, coupled with memory limitations, leads to an intractable learning problem. Their data suggest a different learning strategy based on early binding between words and referents, with errors corrected through natural processes of forgetting.

While the natural environment is complex, it does provide structure notably absent from many laboratory-based word learning experiments. Bruner (1985) emphasized the importance of naturally occurring, predictable *formats* of interaction that support communication. To study the role of formats in language acquisition, Bruner moved his research into the "clutter of life at home" via naturalistic, observational methods. One format that Bruner studied was the game of "peek-a-boo", a recurring, rule-bound activity that occurs across a wide developmental period. Language works in concert with the game to help reveal the meaning of words.

With Bruner's formats in mind, the goal of the present study is to investigate the activity structure of a child's first years of life, how the child's linguistic input links

to these activities, and how such language in context relates to vocabulary growth. Bruner’s formats are complex, with deep rule-governed structure and social roles, patterns that recur over time during the child’s early life. They are difficult to study in detail, especially since they must be observed and deconstructed from longitudinal observations of natural behavior. To avoid this difficulty, we study a simplified representation of formats: consistent *activity contexts*.

We operationalize the idea of an activity context using data mining and machine learning techniques, applied to the multimodal, dense longitudinal recordings collected for the Human Speechome Project (Roy et al., 2006). We apply Latent Dirichlet Allocation models (Blei, Ng, & Jordan, 2003) to the transcribed speech in the Speechome Corpus, obtaining a set of “topics” that connect groups of related words. Inspection of these topics along linguistic, spatial, and temporal dimensions demonstrates that many correspond to coherent, everyday activity contexts such as **mealtime**, **diaper-change**, and so on. We then consider the child’s vocabulary growth relative to both the standard input frequency and measures of a word’s diversity across activity contexts.

## The Human Speechome Corpus

The Human Speechome Project (HSP) (Roy et al., 2006) was launched in 2005 to study early language development through analysis of audio and video recordings of the first three years of one child’s life. The house of one of the authors (DR, who had a newborn child), was outfitted with eleven omnidirectional cameras, fourteen microphones, and a custom recording system designed for large-scale audio/video recording. The cameras and microphones, embedded in the ceilings, provided near complete coverage of the house while remaining unobtrusive, and the practice of simply turning the system on in the morning and leaving it on all day facilitated adoption of the system and helped to minimize observer effects. The nature of this project required extreme sensitivity to the family’s privacy: They had full control over recordings and the ability to “back-delete” recordings if an embarrassing moment was captured. Audio was recorded using boundary-layer microphones which yield high quality audio, even for whispered speech. Video was recorded at approximately 1 megapixel, 15 frames per second, using high dynamic-range cameras for the wide range of lighting conditions. On average, the family recorded 10 hours per day, from the child’s birth to age three. Altogether, the recordings span roughly 120,000 hours of audio and 90,000 hours of video, capturing an estimated 70% of the child’s waking hours.

## Data Annotation

To date, the focus of our annotation and analysis has been on the subset of data spanning the child’s 9-24

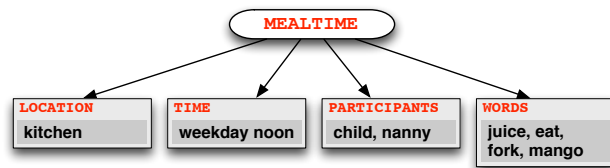


Figure 1: A schematic illustrating how four different dimensions of observable data can depend on a common latent activity context. Viewed as a generative model, the activity context **mealtime** gives rise to the four kinds of observed data.

month age range. For this subset, our long-term goal has been to transcribe *all* speech both heard and produced by the child, but this is a significant challenge using traditional transcription methods. To address this, we have developed BlitzScribe, a new tool for fast, semi-automatic speech transcription (Roy & Roy, 2009). The BlitzScribe system processes raw, unstructured audio and automatically finds speech, segments it into manageable segments, and presents those candidate speech segments to a human transcriber in a simplified user interface. We then use a fully automatic speaker ID system to identify the speaker in an utterance.

Human annotators label which room the child is in (and whether he is awake) over the course of the day. This step ensures that what is transcribed is effectively “child available speech” (CAS), or speech that could be considered linguistic input. Although many studies focus on child-directed speech (CDS) for input-uptake analysis, CDS is much more difficult to obtain at a large scale than CAS. Using BlitzScribe, we have transcribed more than 80% of the CAS audio collected in the 9-24 month age range, which we refer to as the Speechome Corpus. Currently we have transcribed approximately 8 million words, and when fully transcribed we expect the corpus to consist of about 10 million words. However, since some post-processing is required for the latest transcripts, the work described here uses an earlier version of the corpus consisting of approximately 5 million words.

## The Child’s Lexicon

The density and coverage of the Speechome Corpus enables a detailed look at lexical development, including both caregiver speech and the child’s vocabulary over time. In earlier work (Roy et al., 2009), using a smaller version of the corpus, we identified a *word birth* as the first productive use of a word by the child in our transcripts. For our purposes, this served as the age of acquisition (AoA) of each word in the child’s lexicon. We repeated this procedure using the current, larger corpus and identified a large set of candidate word births. We then manually reviewed each of these, removing morphological variations like plurals, dropping invalid word

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ic	ok	ball	i	ba	duck	fish	spider	he	bye	butterfli	book	chew	water	la	cow
cream	put	it	thei	diaper	quack	e	twinkl	like	gaga	crunch	cat	eat	okai	baa	moo
you	come	push	just	poo	bear	farm	piggi	wa	ya	two	sam	mango	no	wool	sun
want	pant	go	know	beep	doggi	turtl	woof	he'	hi	four	fox	it	abar	sheep	moon
mango	your	catch	to	chang	giraff	jellyfish	meow	him	ee	three	read	chees	milk	[noise]	done
what	on	bounc	but	bath	dog	jiggli	star	yeah	NANNY	five	hat	juic	merrili	sir	diddl
shake	go	throw	that	grape	camel	sea	oink	i	rock	ladybug	chip	yum	row	[crying]	all
babi	let'	kick	year	pee	ruff	jelli	neigh	it	wibbl	tigger	knox	pea	tea	[babbling]	jump
do	it	get	like	tama	bird	jenni	bitsi	but	bye-by	bee	ham	bite	cup	[laugh]	mulberri
drum	shower	basketbal	dollar	crayon	eleph	fishi	itsi	just	cradl	pooh	beetl	more	cooki	dame	frederick

Figure 2: The top 10 words (orthographic substitutions are due to the stemming process) for 16 of the top 25 topics.

births, and adjusting birth dates. We identified 670 unique forms in the child’s lexicon at 24 months.<sup>1</sup>

### Activity Contexts

The detailed record of development contained in the Speechome Corpus includes the child’s first words up to multiword utterances. But in addition, the basic routines of daily life are also captured, providing a backdrop for early development. What activities does a child participate in during his first years, and how can they be found in a large, unstructured collection of recordings?

Our approach is to view an *activity context* as a hidden or *latent* variable that explains a set of observable data. An activity such as **mealtime** typically takes place in the kitchen, around noon or in the early evening and involves the whole family, with the speakers often uttering food- and eating-related words. A particular combination of observed *time*, *location*, *words* and *participants* may be best explained by the **mealtime** activity context, illustrated by Figure 1. Thus, an activity context is a latent variable identified by observations across modalities. We wish to identify a set of latent activity contexts from these observables across the entire Speechome Corpus.

Automatic methods for inferring latent variables have been successfully used in data mining applications like document modeling. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is one such technique, which finds a set of latent “topics” that best capture the thematic content of a collection of documents. In LDA, each document is represented as an unordered collection (“bag”) of words; the inferred topics are modelled as distributions over words. Topics group related words together and documents are represented as sparse mixtures of topics. Often, a human can interpret and label the topics simply by inspecting the topic words. As a first exploration of activity structure in the Speechome corpus, we

apply LDA directly to transcripts. We then assess the relationship between LDA topics and activity contexts using data from time and location.

### Applying LDA to the Speechome Corpus

To apply LDA to the Speechome Corpus, we partitioned the transcripts into “documents” using a sliding window procedure. Beginning at the 9-month mark we advanced a 10 minute window over the corpus, shifting the window forward by 10 minutes up to the 24 month mark. All transcribed speech in a window was output as a document for processing by LDA, skipping empty time windows that didn’t contain speech, resulting in 13,672 documents. After some experimentation, we found stemming to be a useful preprocessing step, normalizing word forms to a common root using the Porter stemmer (Porter et al., 1980), and only accepting those words occurring in more than five documents (and occurring more than five times in the corpus.) This yielded a vocabulary of 6,583 unique word types.

In the case of standard LDA, the number of topics to produce is a parameter of the algorithm, and we found 25 topics to be a manageable number while still producing coherent topics. Extensions to LDA such as Hierarchical Dirichlet Processes (Teh, Jordan, Beal, & Blei, 2006) can automatically select the number of topics, and informal experiments with this method also resulted in 20 – 30 topics. To interpret the resultant topics, a common starting point is to review the top words in each topic. We ranked words using the method in (Blei & Lafferty, 2009), which roughly measures the informativeness of the word for the topic relative to the other topics (Figure 2).

### From Topics to Activities

Do topics capture activities? We investigate two methods to make the link: via correlations in time and space, and via human-annotated activities.

**Activities in time and space** LDA outputs topic mixture weights for each document; since documents also have spatial and temporal attributes, we can exploit this to measure how topics are distributed in time and space. Each topic’s time distribution was calculated by weighting the time of day of each document by the topic’s con-

<sup>1</sup>We did not annotate or study *receptive* AoA, which is often documented in diary studies but is much more difficult with a large corpus. Identifying word births is challenging in its own right, since the child’s word form may differ from the adult form. In describing the diary study of her daughter’s early lexical development, Dromi (1987) reviews these and other challenges. In our case, the original audio, video, and access to caregivers were all helpful resources.

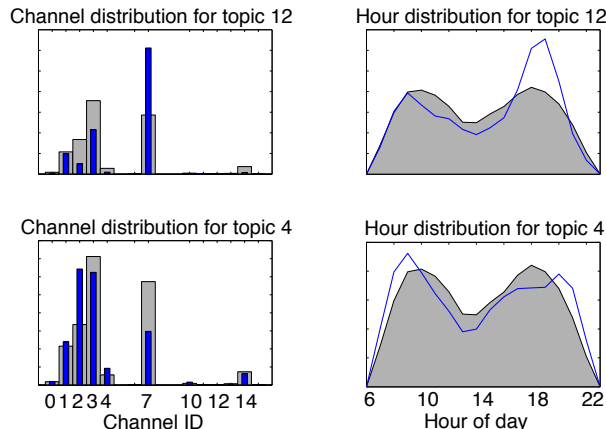


Figure 3: Spatial and temporal distributions for topics 12 and 4. The gray background graphs show overall averages, the blue foreground graphs the “conditional” distributions for the topic (distributions reweighted by the selected topic activity.) The channels of interest here are 7 (kitchen) and 2 (baby bedroom.)

tribution to that document. The spatial distribution of topics across rooms of the house was calculated similarly.

Figure 3 shows the temporal and spatial distributions for topics 12 and 4, relative to the average temporal and spatial distributions. When topic 12 is active, we see that recording channel 7 (the kitchen) is well above average, and the temporal distribution peaks at about 7pm. Inspecting the words in topic 12 shows that it captures food and eating related terms. So, topic 12 appears to be a *mealtime* activity context, or perhaps is more specific to *dinnertime*. Topic 4 is most active, relative to the average, in the early morning and late evening, and in channel 2, the baby’s bedroom. This topic appears to capture the *diaper-change* activity. Thus, at least a subset of topics appear to follow coherent spatial and temporal distributions.

**Human annotated activities** In concert with our efforts to automatically identify activity contexts, we are also manually annotating activities. Using BlitzScribe, annotators now transcribe assignments spanning 15 minutes of “house time,” then list the activities that took place. When we began this annotation project, we gave little instruction to transcribers, asking them to make up their own activity tags as necessary. Nevertheless, we found consistency in the activities that emerged. After conventionalizing tag names, we obtained roughly 30 activities for around 300 annotated assignments. These annotations can provide another means for validating LDA topics as proxies for activity contexts.

To test for relationships between LDA topics and activity contexts, we examined the correlations between individual topics and the human-annotated ac-

Table 1: Coefficients on a multilevel linear regression model predicting age of acquisition (months) on the basis of log frequency in child-available speech, topic entropy of the word, and their interaction.

Predictor	Coefficient	Std. Err.	<i>t</i> -value
Intercept	18.49	0.28	65.83
Log frequency	-0.83	0.12	-7.08
Topic entropy	0.54	0.10	5.44
Log freq $\times$ entropy	0.06	0.13	0.48

tivities. While these correlations remain speculative due to the sparsity of the human-annotated activities, several significant correlations emerged. In the case of the *diaper-change* activity, for example, only topic 4 was significantly correlated (with words like “diaper,” “poo,” and “change” highly active). In the case of *eating*, topics 12, 16 and 0 are significantly positively correlated, with 12 being the strongest (e.g. “chew,” “eat,” and “mango”). For *reading*, a number of topics were active, including 5, 6, 10, and 11, all of which contained words related to different books that were read to the child. And for *crying*, topic 17 (e.g. “daddy,” “blanket,” and “ssh”) was most active. In summary, although at present human annotation of activities is too limited to provide full coverage, the relationships between activities and topics makes us optimistic that our topics are capturing at least some aspects of the varying activity contexts in the child’s environment.

## Word Learning

If LDA topics act as a proxy for activity contexts, then we should be able to use them to test a primary hypothesis of interest: that words that appear in consistent activity contexts are learned relatively earlier than those that appear across a range of contexts. Said another way, words with high *topic entropy*—that do not appear consistently in one or a small set of topics—should be produced later by the child.

We used multilevel linear regression (Gelman & Hill, 2007) to predict age of acquisition (AoA, in months) on the basis of word frequency and topic entropy. AoA measures are described above. For word frequency, we measured the total number of utterances of a target word in our sample up to the age of acquisition of the word, normalized by the number of days of transcripts up until that time to allow these measurements to be compared for words with different AoA.<sup>2</sup> For topic entropy,

<sup>2</sup>We measure only up until the acquisition of the word to avoid a confound: the child’s production of a word could change the adult use of the word. Note that this change, the exclusion of words from the topic model, and several other minor changes make regression coefficients for frequency slightly lower compared with our previous work.

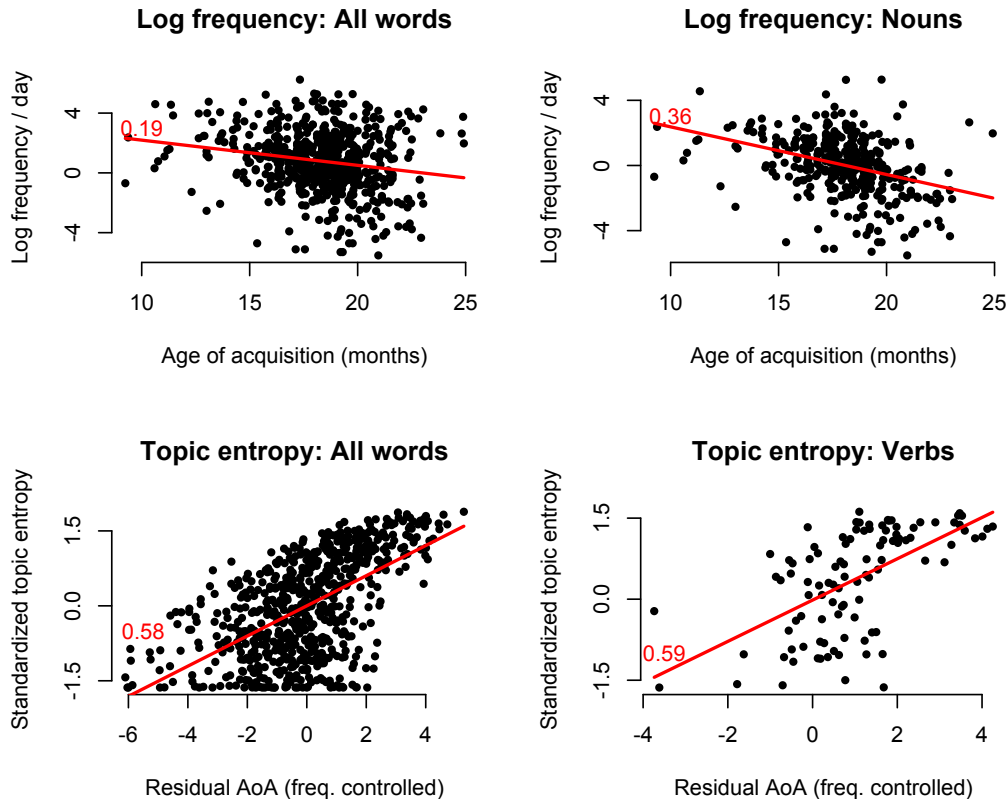


Figure 4: Each plot shows age of acquisition in months for individual words plotted by a predictor. (top) Standardized log word frequency per day (prior to the child’s first production of the word) plotted by age of acquisition for words. Left plot shows all words, right plot shows nouns. (bottom) Standardized topic entropy plotted by residual age of acquisition for words. Left plot shows all words, right plot shows verbs. Red lines show best fitting linear function; numbers indicate correlation coefficients.

we computed the entropy of the word’s weight distribution over the 25 different topics found by the LDA model. We z-scored the units for both of these predictors to be able to compare coefficients for each.

Our model included fixed effects for topic entropy, frequency, and their interaction. The model also included random terms for each syntactic category, and its frequency and topic entropy, and their interaction. Coefficients are shown in Table 1. Coefficient weights can be interpreted as months in AoA per standard deviation in log frequency or topic entropy. We assessed reliability for individual coefficients by testing whether they increased model likelihood. Log frequency had a large negative effect on AoA: more frequent words were learned earlier ( $p < .005$ ). Topic entropy had almost as large an effect: words in more constrained activity contexts (lower entropy) were learned earlier as well ( $p < .005$ ). Their interaction did not significantly increase model fit ( $p = .69$ ).

Figure 4 shows the relationships described by this model. Without regressing out frequency, topic entropy is relatively uncorrelated with age of acquisition. When both terms are entered in a model, however, the effect is

much larger. Figure 4 displays this conditional relationship by plotting topic entropy by residual AoA (controlling frequency).

Topic entropy and part of speech are likely correlated: closed class words like “if” are likely to occur in every topic (topic entropy of 1.3 SDs above mean), while nouns like “pasta” only appear frequently in one context (1.6 SDs below). A key part of this analysis was the use of multilevel models to control for part-of-speech effects. Without including random effect terms for part-of-speech, the interaction between frequency and topic entropy was large, probably because topic entropy and frequency for closed class words is high. Adding the random effects terms eliminated this interaction, however.

To summarize this analysis: we found that the consistency of the contexts within which words appeared was almost as strong a predictor of age of acquisition as pure frequency.

## Discussion and Future Work

Early word learning is a product of powerful learning mechanisms coupled with the rich experience of early childhood. Linguistic input is of critical importance to

lexical development, but it is situated in the larger structure of daily life. The importance of social activity structures was emphasized by Bruner (1985), yet large-scale, quantitative study of their effect on language acquisition has proven difficult. To address this, we used document modeling techniques to operationalize activity contexts. We found evidence that many of the resultant topics captured coherent, interpretable patterns of linguistic, temporal and spatial activity. These activity contexts then provided a useful source of information in modeling lexical acquisition: we found that more contextually focused words were learned earlier.

In future work, we plan to add location, participants, and time to models of latent activity contexts. An interesting question for these extensions is whether some contexts are of more value than others for general word learning or for learning particular words. In addition, a study of episodes of a particular activity may help build intuitions about how activities develop and change over time, and how this progression relates to the child's development.

Our study here represents a first step towards a more complete model of lexical acquisition, one that incorporates elements of social and physical context. In a similar vein, Miller (2011) and Shaw (2011) studied the spatial distribution of language in the Speechome Corpus. Miller (2011) found that more spatially localized words correlated with earlier AoA, noting that many of the most salient locations were directly interpretable in terms of the activities known to take place at those locations. Our work builds on this intuition, targeting activities via their linguistic manifestations. While both of these methods are at best proxies for as-yet-unseen structures, our hope is that by continuing to develop methods for identifying activity contexts, we can gain some insight into the crucial role these social structures play in early language learning.

## Acknowledgments

Many thanks to our team of annotators, and to Cybelle Smith and anonymous reviewers for helpful comments.

## References

- Blei, D. M., & Lafferty, J. (2009). Topic models. In A. Srivastava & M. Sahami (Eds.), *Text mining: Classification, clustering, and applications* (pp. 71–93). Chapman & Hall.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003, January). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Brent, M., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81, 33–44.
- Bruner, J. (1985). The role of interaction formats in language acquisition. In J. P. Forgas (Ed.), *Language and social situations* (pp. 31–46). Springer-Verlag.
- Dromi, E. (1987). *Early lexical development*. Cambridge University Press.
- Echols, C., & Newport, E. (1992). The role of stress and position in determining first words. *Language acquisition*, 2.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (Vol. 648). Cambridge University Press New York.
- Goodman, J., Dale, P., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, 35, 515–531.
- Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, 27.
- Medina, T., Snedeker, J., Trueswell, J., & Gleitman, L. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22), 9014.
- Miller, M. (2011). *Semantic Spaces: Behavior, Language and Word Learning in the Human Speechome Corpus*. Unpublished master's thesis, Massachusetts Institute of Technology.
- Porter, M., et al. (1980). *An algorithm for suffix stripping*. Program.
- Roy, B. C., Frank, M. C., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. In *Proceedings of the 31st Annual Cognitive Science Conference*.
- Roy, B. C., & Roy, D. (2009). Fast transcription of unstructured audio recordings. In *Proceedings of Interspeech*. Brighton, England.
- Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., et al. (2006). The Human Speechome Project. In *Proceedings of the 28th Annual Cognitive Science Conference* (pp. 2059–2064). Mahwah, NJ: Lawrence Erlbaum.
- Shaw, G. (2011). *A taxonomy of situated language in natural contexts*. Unpublished master's thesis, Massachusetts Institute of Technology.
- Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476), 1566–1581.
- Vosoughi, S., Roy, B. C., Frank, M. C., & Roy, D. (2010). Effects of caregiver prosody on child language acquisition. In *Fifth international conference on speech prosody*. Chicago, IL.
- Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5), 414.

# Modeling Cognition: How Fiction Relates to Fact

**Anna-Mari Rusanen (anna-mari.rusanen@helsinki.fi)**

Philosophy of Science Group/  
Department of Philosophy, History, Art and Culture Studies, PO BOX 24  
00014 University of Helsinki, FINLAND

**Otto Lappi (otto.lappi@helsinki.fi)**

Cognitive Science  
Institute of Behavioural Sciences, PO BOX 9  
00014 University of Helsinki, FINLAND

## Abstract

The increasing use of computational modeling and simulation methods offers interesting epistemic and theoretical challenges for the philosophy of science. One of the main questions discussed in the philosophical literature relates to the explanatory role of false, unrealistic and sometimes even fictional models. In this paper we argue that (i) some fictional models can offer explanations known as structural model explanations, and (ii) at least some variants of realism, such as the information semantic account of scientific models, can consistently hold that this subset of fictional models are explanatory.

**Keywords:** Models; fictional models; explanation; information semantics

## Introduction

For a philosopher of science interested in the philosophical issues of modeling, cognitive science is a wonderful source of case studies. Cognitive science utilizes modeling in a unique way, both methodologically and theoretically. The increasing use of computational modeling and simulation methods offers interesting methodological challenges for scientists, but also philosophers of science find many things of interest in the theoretical and epistemic status of modeling methods.

One of the main questions discussed in the philosophical literature relates to the explanatory role of models. A growing number of philosophers have proposed that explanation of the behavior and capacities of complex systems (such as those found in the cognitive, biological and neurosciences) does not typically involve natural laws, but specific models of particular mechanisms (Bechtel and Richardson, 1993; Craver, 2006, 2007; Machamer, Darden, and Craver, 2000). It has also been argued that this mechanistic account of explanation could be extended to cover explanations in cognitive science (Kaplan & Craver, 2011, Sun, 2008) and computer sciences, as well as computational neuroscience (for instance, Piccinini, 2007).

According to this account, to explain a phenomenon is to construct a model of how a causal mechanism - a hierarchical system composed of component parts, their

properties and their causal relations - gives rise to or produces the phenomenon. Constructing an explanatory mechanistic model involves mapping elements of a mechanistic model to the system of interest, so that the elements of the model correspond to identifiable constituent parts with the appropriate organization and causal powers to sustain that organization.

The mechanistic account of explanation is a typical example of the realist interpretation of scientific models. According to realism, a model explains the behavior of a target system, if and only if it is a correct account of the target's behavior underlying observed phenomena - i.e. the model must correspond to, depict or represent the target system in a sufficiently correct way. In addition, many current realist accounts require that the target systems are actual or real - i.e. have causal power to generate observable phenomena and data.

However, models are always more or less abstract, simplified and idealized descriptions of their real world target systems. Target systems are just too complicated to be studied in a full fidelity, and thus all kinds of assumptions are made to reduce the complexity of a model. Thus most (if not all) models used in science are unrealistic. Often models are nevertheless considered useful, even if they are known to be false, and they are known to contain assumptions that are not even approximately true, but highly idealized. For this reason, it has been argued that this feature of modeling seriously undermines the realist interpretation of models. If all models involve unrealistic elements, how is it possible that they could correspond, depict or describe the real world target system in a correct or truthful way? If they do not, where does their explanatory force come from?

Sometimes models involve assumptions about fictional entities and processes that are known not to exist in the real world. These fictional models describe systems that (i) do not exist in the real world or (ii) have elements that do not exist in the real world. Obvious examples of fictional models in cognitive science are for instance the models of



artificial intelligence<sup>1</sup>. These hypothetical (at the time of their conception) systems offer an example of modeling, which starts with explicitly fictitious entities – non-existing, imaginary cognitive systems – and then converts these fictions into the fact-like platforms from which further research can be done.

Although there is a growing consensus among philosophers that fictions have a legitimate role to play in science, traditionally those philosophers who endorse realism have denied that fictional models could explain. For a realist, the main obstacle to admitting that also fictional model can be explanatory is that it is difficult to imagine how a model without an existing target system could be explanatory. Instead, fictional models have been treated as, for example, only tools for generating and calculating predictions.

However, it seems to us that for example the models offered by AI are more than mere “tools for prediction”. That these fictional models can be converted into real working systems does require that they get some principles of cognition “right” (a fictional AI model that is completely *unrealistic* has little hope of successful implementation). And they do seem to work more like blueprints or instructions for a design than simply devices for predictions. Not only do they offer structural information about the constitution of a model system, but they also restrict and guide the construction process itself – and not in an arbitrary manner.

For this reason it seems intuitively plausible to think that these models do hold a potential to represent or explain something about cognitive systems. In this paper, we argue that (i) these models explain by showing how the structure of a model limits what sorts of objects, properties, states, or behaviors are admissible within that model, and they offer explanations known as structural model explanations (Bokulich, 2008, 2009). In addition, we argue that (ii) at least some variants of realism, such as the information semantic account of scientific models, can consistently hold that this subset of fictional models are explanatory. However, (iii) whether or not a fictional model can be explanatory, depends also on the relationship of the fictional models and the real world<sup>2</sup>.

## Requirements for Realistic Interpretation

Scientific models can be interpreted *realistically* and *instrumentally*. The instrumental interpretation, roughly, holds that scientific models are *instruments for generating predictions about the system's behavior*. According to the instrumentalist interpretation, the models are only used to

produce predictions, and the question whether or not models are realistic or unrealistic, does not arise.

In contrast to instrumentalism, the realist interpretation of explanatory models holds that entities/processes the model posits actually exist and that there is an objective relationship between the model and its target system. In addition, many current realist accounts of explanatory models require that the target system must be actual or real. Following Salmon, many philosophers of science agree that to explain a phenomenon is to describe the actually existing/causal processes that constitute the phenomenon (Salmon, 1984). According to this view, only descriptions of actually existing genuinely causal processes can lead to scientific explanations<sup>3</sup> (Salmon, 1984).

**Aboutness.** A realist typically thinks that the relationship between a model and its target, the *aboutness* of scientific models, is essential to their explanatory power. A scientific model can *explain* the behavior of a target system if and only if it depicts, describes or represents the target system.

For example, the mechanist account of explanation is a typical variant of realism in this sense. According to it explanation involves constructing a model of such mechanisms that correctly depicts or describes the causal interactions among their parts that enable them to produce the phenomena under various conditions. The relationship between a model and its target is seen, for example, as “one of approximate similarity” (Glennan, 2000) in a way, where “the behavior of the system in nature is described (to varying degrees of approximation) by the model's behavioral description and the internal structure of the system is described (again to varying degrees of approximation) by the model's mechanical description”, or as “correspondence” (Craver), because (emphasis added) “in successful explanatory models... (a) the variables in the model *correspond to* components, activities, properties and organizational features *of the target mechanism* that produces, maintains, or underlies the phenomenon, and (b) the... dependencies posited among these variables in the model *correspond to* the... causal relations among the components of the target mechanism” (Kaplan & Craver, 2011; see also Craver, 2006).

Following Giere (1988), both Glennan (2000) and Craver & co (2006, Kaplan & Craver, 2011) seem to think that the relationship – “the aboutness” – is some kind of similarity or correspondence relation. But as Quine (1969) pointed out, similarity is a vague notion, and correspondence is actually not much better. In order to specify these concepts, many philosophers of science have appealed various “morphisms” – iso-, partial- or homomorphisms. However, all these various morphisms, similarities and correspondences are usually discussed only intuitively and philosophically problematic<sup>4</sup>. They have been criticized on logical and

<sup>1</sup> The aim of AI is not only to help us investigate the existing cognitive systems by simulating them, but also to help us to investigate and understand the structure of cognitive systems by producing or to creating artificially new kind of cognitive systems.

<sup>2</sup> Not all fictional models are explanatory, and some of them are useful as calculational devices, other as prototheories and some are useful in generating predictions etc., but not explanation.

<sup>3</sup> For example, the ontic version of mechanistic explanation requires that the mechanistic organization of the target system causally produces the phenomenon to be explained.

<sup>4</sup> There is a huge debate on this issue in philosophy of modeling. See for instance, Suárez, 2003; Frigg, 2006.

substantial grounds<sup>5</sup>. One particularly difficult problem for these accounts is the problem of relevance: a model typically cannot be perfectly “similar” or “isomorphic” with respect to every entity and every relation in the target system, since almost any target system is too complex. This implies that models should be “sufficiently isomorphic” or “sufficiently similar” to the target “in the relevant respects”. However, it is quite tricky to characterize “sufficiency” and “relevance” in a non-circular and precise manner<sup>6</sup>.

There are also some other variants of realism in which the relationship between a model and its target is defined in a different way. According to the information semantic account<sup>7</sup>, models depict, describe, represent or are about their target systems, if and only if there is an appropriate information-content relationship between a model system and its target. The information connection is implemented in a model building process, in which data about the world is incorporated to the model i.e. the information relationship between a model system and its target is implemented in empirical data, which carries the information about the target system into the model.

**The actual existence of target systems.** As the variants of realism typically do, also the information semantic account requires the target systems must be actual or real. Because in information semantics “carrying information” is understood in terms of statistical dependence (Shannon, 1948; Usher, 2001), and statistical dependence is usually understood in causal terms, target systems must be “actual” or “real”: they must have *causal power* to produce the data, which carries the information which is incorporated into the model during the model building process.

To summarize, the realist interpretation of explanatory models requires that an explanatory model (i) represents, depicts or describes the explanatorily relevant features of target systems in a “correct” way, and that (ii) target systems are real and actual. Some variants of realism, such as the information semantic account, require also that (iii) the target systems have appropriate causal properties.

However, most, if not all, models used in science are unrealistic, and sometimes even known to be strictly false because of the abstraction, simplification and idealization and the postulation of fictional entities that goes on in the model building process. This raises the epistemic problem for realism: if models include all these kinds of sources of false assumptions, how can they be explanatory? How, exactly, are these models are used to gain information or knowledge about the real world phenomena?

### The different ways to make a model unrealistic

In practice, when scientists present a model, they offer a model system<sup>8</sup> which is a description of a hypothetical

system, and this model system is, or is thought to be, a hypothetical representation of a real world target system. These model systems are always more or less unrealistic descriptions of their real world target systems, because target systems are too complicated to be studied in every detail. This is typically motivated, rhetorically, on pragmatic grounds. If all the parameters were included in models, they would become too complicated to be understandable, tractable or useful. As McClelland (2009) puts it, “the more detail we incorporate, the harder the model is to understand.”

There are at least four different ways to make models unrealistic: abstraction, simplification, idealization and fictionalization. In practice, this distinction between these types is not always entirely clear. Moreover, these classes are not exclusive. Models can, and often are, abstracted and idealized at the same time<sup>9</sup>. For example, a Turing Machine can be seen as an abstraction of real computations<sup>10</sup>, because it neglects many computationally irrelevant features of computational systems, such as the material basis of the implementing mechanisms. If a Turing Machine is also defined to have properties that are not implemented in any real computational system (unlimited memory, unlimited processing time), it is assumed to never break down and so on, it also involves idealization.

*Abstraction* and *simplification* can be considered to be species of information reduction. Roughly speaking, an abstract model is the result of the process of abstraction, in which information about domain-external factors is disregarded<sup>11</sup> (e.g. a model of a ball rolling on an inclined plane may abstract away the color of the ball, which is not in the domain of Newtonian dynamics). A simplified model is a model, in which some domain-internal factors are given a simplified description (e.g. the ball may be considered a perfect sphere, with the center of gravity in perfectly in the middle). Although abstracted or simplified models do not describe *all* the factors, they describe correctly or approximately certain features of their target systems. Abstracted and simplified models can thus be genuinely explanatory, if they accurately depict the relevant properties of their target systems. These models tell us how phenomena behave in a simpler world than our own<sup>12</sup>, or these models can work as surrogate systems<sup>13</sup> for understanding, how fundamental properties of a system

<sup>9</sup> See Thomson-Jones (2005) for the distinction between abstraction and idealization.

<sup>10</sup> See Piccinini, 2007.

<sup>11</sup> An abstract model is sometimes described as a model, in which *only some factors* or only some of potentially many factors of target system are included into a model system. A simplified model is a model, in which *only some factors of the potentially many factors that are relevant* to the behavior of a target system are included into a model or some factors are given a simplified description.

<sup>12</sup> This idea is explicated by Stephan Hartmann, but the authors did not find the article, in which this idea was presented.

<sup>13</sup> The term “surrogate” is borrowed from Uskali Mäki (2009).

<sup>5</sup> See Suárez, 2003; Frigg, 2006; See also Rusanen and Lappi, 2011.

<sup>6</sup> Suárez, 2003; Frigg, 2006, see also Rusanen & Lappi, 2011.

<sup>7</sup> Rusanen & Lappi, 2011.

<sup>8</sup> The term “model system” is from Frigg (2006, 2010).

generate or produce the certain phenomenon of interest by helping scientists to formulate correct what if- inferences.

However, philosophers of science disagree on the question, whether or not abstracted and simplified models have more explanatory power than non-abstracted models. There are at least two different views about the explanatory power of abstracted models. One of them is the so-called “traditional view”<sup>14</sup>, according to which the more exact, more detailed, more complete and more realistic the model is, better it is. The most explanatory model is the model, which offers the complete description of the phenomenon of interest. The non-traditional view maintains that in some cases the abstracted model can explain better the dominant and significant features of the target system, because it isolates and emphasize the crucial elements in a tractable way.

The third way to make a model unrealistic is *idealization*. In idealization one is not only excluding parameters. Instead, idealization involves distorting theories or models, because at least one of the parameters of the target system is represented in a way that makes the model false. For example, if a model in cognitive science is used to analyze the processing of a perceptual system scientists may stipulate that all processing is described in the model as linear and strictly feedforward<sup>15</sup>, even if in reality the processing would be non-linear and have backforwarding properties.

The fourth way to make models unrealistic is *fictionalization*. *Wide fictionalism*<sup>16</sup> states that idealization and abstraction are subspecies of fiction (Suárez, 2009). However, there are reasons to argue that models that are only simplified, idealized and abstracted representations of real entities (*about* which they make counterfactual claims) should be distinguished from those which refer to fictional entities which do not actually exist. Logically speaking, there is a difference in kind between a representation of real entity and a representation of an imaginary entity (Russell, 1905, Suárez, 2009)<sup>17</sup>.

*Narrow fictionalism* takes it that only those models which involve or describe explicitly fictional or imaginary entities, systems and situations that do not actually exist in real world, are fictional (Suárez, 2009). Such models do not only involve idealization, simplification or abstraction, but they are, and also known to be false, for a further reason: because they describe fictional or imaginary entities, systems and situations that do not actually exist in real world.

---

<sup>14</sup> The terms “traditional view” and “non-traditional” are from Batterman, 2009.

<sup>15</sup> See, for example, McLelland (2009) for a detailed analysis of simplification and idealization of this sort.

<sup>16</sup> About the difference between wide and narrow fictionalism, see Suarez 2009.

<sup>17</sup> As Suárez (2009) has proposed, one way to describe the distinction is to emphasize the difference between “fictional” and “fictive” representations. A fictional representation is a representation of a non-existing entity, and a fictive representation is an inaccurate representation of a real entity (Suárez, 2009).

*Completely fictional models* refer to model systems that do not actually exist in the real world *and do not have any real components*. It is difficult to find a genuine example of a completely fictional model in natural or behavioral sciences, because in these sciences most (and probably all) fictional models include at least some real world elements at some level of analysis.

Actually, fictional models are typically only *partially fictional models*. Often partially fictional scientific models are combinations of real and fictional components. Sometimes these models refer to *real* target systems, but they are fictional, because they include some components or system level descriptions that are taken to represent non-existing entities. For example, the frequency components in the wavelet analysis of EEG components, which are used to explain the synchronization properties of neuron population in neurosciences are typically interpreted as non-existing entities.

Some of partially fictional models consist of realistic constituents, but the combination of constituents is known to be unreal. Some of these model systems may describe systems that are in principle physically possible, or sometimes they are physically impossible, because they violate natural laws etc. Typically these models are used to test the possible behavior of a complex system by creating all kinds of what-if simulations. An example of a model of this sort is the model of xDNA<sup>18</sup>. All of the components of model can be given a real world interpretation, but the combination of these components, xDNA, is unrealistic.

The study of artificial intelligence offers another example of modeling of this sort. For example, if a cognitive scientist wanted to build synthetic brains, she might end up building a model system that does not mimic or simulate any existing brains. Although the design or the computational layout of the artificial brains was novel, the model system might involve elements, which refer to real world entities, such as cells, cell organs, transmitters, or depending on the material implementation, silicon chips, batteries and so on. In addition, if the model system would then be implemented in a concrete way, then a model system of a fictitious entity would have been converted into an actual model organism.

Although the current study of artificial intelligence is not developed to that point, are already existing artificial cognitive systems examples of modeling, which starts explicitly fictitious, non-existing entities – the imaginary cognitive systems - and then convert these fictions into *the fact-like platforms* from which further research can be done. Because these fictional models can be converted into fact-like platforms, fictional models seem to work more like blueprints or instructions for a design than simply devices for predictions. Not only do they offer *structural* information about the constitution of a model system, but they also restrict and guide the construction process itself. For this reason it seems intuitively to think

---

<sup>18</sup> This example is from Michael Weisberg’s presentation in Helsinki in May 2009. Actually, as also Weisberg mentioned in his talk, there is no such a model in biology.

that these models do explain. They seem to explain by showing how the structure of a model limits what sorts of objects, properties, states, or behaviors are admissible within that model. They also show that whatever the system can do is in fact a consequence of that structure and they also offer information about how the converted system will behave *before* it has been converted

## Do Fictional Models Explain?

Alisa Bokulich (2008, 2009) has recently developed an interesting account of scientific explanation, called “model explanations” to describe the sort of explanation that is being offered by fictional models.

Bokulich (2008;2009) characterizes these model explanations as follows: First, the explanans of explanatory fictions must make a reference to scientific model, which involves some idealization and/or fictionalization. Second, that model is taken to explain the explanandum by showing that the pattern of counter-factual dependence in the model mirrors the relevant respects of counterfactual dependences in the target system. Following Woodward (2003), in Bokulich’s account this pattern of counterfactual dependence can be explicated in terms of “what if things have been different- questions”<sup>19</sup>. That is, the explanation must enable us to see what sort of difference it would have made to *explanandum* if the factors cited in *explanans* had been different in various possible ways. The third feature of model explanations is that they must specify what the domain of applicability of the model is and show that the phenomenon in a real world to be explained falls within that domain. If model explanations are characterized in this way, one subspecies of model explanations are structural model explanations (Bokulich, 2009).

A structural model explanation is one in which the the explanandum is explained by showing how the structure of a model limits what sorts of objects, properties, states, or behaviors are admissible within that model, and then showing that the explanandum is in fact a consequence of that structure. A structural model explanation is thus an explanation, in which the explanandum exhibit a pattern of counterfactual dependence on the elements represented in the model, and this dependence is a consequence of the structural features of the model.

It seems to us that *those partially fictional models that are combinations of realistic parts offer structural model explanations* as Bokulich proposes. For example, a model of artificial cognitive system characterizes why certain cognitive processes are possible for a certain kind of cognitive architecture, or how a possible computational structure of certain type architecture will limit its possible cognitive and computational capacities, before the system is actually implemented.

However, Bokulich’s characterization may be a bit too broad. For this reason, we’d like to add one crucial requirement for model explanations. In order to count as genuinely explanatory a model explanation must also be *credible*. Characterizing the credibility is, of course, a challenging task, and there are different suggestions in the literature. For example, according to Sugden (2000) models are artificial “worlds” i.e. “surrogate systems”, and their epistemic dimension is based on inductive extrapolation from these artificial worlds to the real world. In Sugden’s account the relationship between models and the world can be evaluated in terms of similarity; more similar to the real world a model is judged to be, more credible it is. However, similarity alone is usually not sufficient for establishing credibility<sup>20</sup>.

For this reason, credibility considerations must be based on more fundamental claims. According to the information semantic account (Rusanen & Lappi, 2011) the credibility of a model explanation requires that there is, or it is at least possible to imagine, a causally implemented information relationship between a model and its target and a credible data gathering method for that particular model. Information semantic account requires that there is an appropriate information relation between a model and its target. For this reason, if a model is credible, there must exist or it must be possible to imagine a causally implemented information relationship between a model and its target. Because of this, from an information semantic view, completely fictional models, arbitrary models or models, which have only unrealistic constituents, are not explanatory. Instead, only such partially fictional models can offer model explanations, in which the constituents of a model system are realistic. These constituents should (at least to certain extent) refer to/ carry information about real world elements and this information relationship could be, in principle, implemented in data.

So, the final problem is: How is it possible that a fictional model can carry information about the real world? As philosophers of cognitive science know, a structurally similar problem, the problem of uninstantiated properties, plagued the early versions of information semantics in philosophy of mind. In a nutshell, the problem was the following: If A does not carry information about B, it is not a representation of B. So, if B is a non-existing, uninstantiated entity or a property, A cannot carry information about it, and thus A is not, strictly speaking, a representation of B. However, there are still terms, such as unicorns or pegasuses, which clearly have semantic content, even if their referents do not exist. We can attribute properties to these non-existing entities; we can make thought experiments on them and so on. So, if these terms have no existing targets, how do these terms have their semantic properties? What is the basis for the meaning of these terms?

Jerry Fodor proposed (1991) one possible solution for this problem. On his view, these terms, such as “pegasus”, could be seen as complex terms, which can be decomposed into its

<sup>19</sup> While in Woodward’s manipulationist construal explanations are restricted purely to causal explanations, Bokulich adds that not all scientific explanations must be causal.

<sup>20</sup> Kuorikoski and Lehtinen (2009) make a similar point.

constituents (a horse and wings), and these constituents refer to/carry information about the real world components. So, these terms can have a meaning, because the constituents of the terms can carry information about the real world.

Partially fictional models can be treated in a same way. They are complex constructions, which can be decomposed into constituents. If these constituents refer to/carry information about the real world elements, they are realistic. For that reason these models are not completely fictional, although the complex composition of constituents would be fictional. Because the constituents of partially fictional models carry information about/refer to real world elements, these models may indeed offer structural model explanations.

## Concluding Remarks

There are at least four different ways – simplification, abstraction, idealization and fictionalization – to make models unrealistic, not all of them make models false in a way that is problematic for the realist. Even if in practice the difference between these types is not always clear, they should be treated separately. They have different implications for the explanatory power of a model. For example, although simplified or abstracted models do not describe all the factors or all the relevant factors of target systems, they describe certain some features of their target systems. Depending on their degree of truthlikeness they can be more or less explanatory. In this paper we argued that also fictional models may explain by showing how the structure of a model limits what sorts of objects, properties, states, or behaviors are admissible within that model, and they offer explanations known as structural model explanations (Bokulich, 2008, 2009). However, whether or not a fictional model is explanatory, depends also on the relationship of the fictional model and the real world. Only such partially fictional models that have constituents, which can carry information/refer to real world elements, can be explanatory. Completely fictional models, arbitrary models, or models with only unrealistic constituents do not explain.

## Acknowledgements

The members of POS and TINT, the audience of EPSA 2011 and the anonymous referees for commenting an earlier draft of this paper.

## References

- Ankeny, Rachel (2009). "Model Organisms as Fictions". In M. Suarez (Ed.) *Fictions in Science: Philosophical Essays on Modeling and Idealization* (Routledge, pp. 158-178.)
- Batterman, Robert. (2009). Idealization and Modeling. *Synthese* 169 (3):427 - 446.
- Bechtel, William & Richardson, Robert (1993). *Discovering Complexity, Decomposition and Localization as Strategies in Scientific Research*. New Jersey: Princeton University Press.
- Bokulich, Alisa (2008). "How Scientific Models Can Explain", *Synthese* 180 (1): 33-45 (2011).
- Bokulich, Alisa. (2009). "Explanatory Fictions", in M. Suarez (Ed.) *Fictions in Science: Philosophical Essays on Modeling and Idealization* (Routledge, 2008: 91-109).
- Craver, Carl (2006). When mechanistic models explain. *Synthese*, 153: 355–376.
- Dretske, Fred. (1981). *Knowledge and the Flow of Information*. Cambridge, Massachusetts: MIT Press.
- Fodor, Jerry A. (1991). Modal Argument for Narrow Content. *Journal of Philosophy* 88 (1):5-26.
- Frigg, Roman. (2006). "Scientific Representation and the Semantic View of Theories." *Theoria* 55: 49-65
- Frigg, Roman (2010). "Models and Fiction". *Synthese* 172(2), 2010, 251-268
- Giere, Ronald. (1988). *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.
- Glennan, Stuart. (2000). "A Model of Models". Unpublished.
- Kaplan, D.M. and Craver, C.F. (2011) "The Explanatory Force of Dynamical Models" *Philosophy of Science* 78 (4): 601-627
- Kuorikoski, Jaakko & Lehtinen, Aki (2009): Incredible World, Credible Results. *Erkenntnis* 70: 119-131.
- Mäki, Uskali (2009). "MISSing the world: Models as isolations and credible surrogate systems", *Erkenntnis*, vol. 70, no.1, 29-43
- Piccinini, G. (2007). Computing mechanisms. *Philosophy of Science*, 74: 501–526.
- Rusanen, Anna-Mari & Lappi, Otto (2011). "Information Semantic account of Scientific Models". In H.W. de Regt, S. Hartmann and S. Okasha (eds) "EPSA Philosophy of Science: Amsterdam 2009", Springer.
- Russell, Bertrand. (1905). "On denoting."
- Salmon, Wesley. (1984). *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.
- Suárez, Mauricio (2003). "Scientific Representation: Against Similarity and Isomorphism." *International Studies in the Philosophy of Science*, 17: 3, October 2003, pp. 225-244.
- Suárez, Mauricio. (2009b). "Fictions in Scientific Practice", in M. Suarez (Ed.) *Fictions in Science: Philosophical Essays on Modeling and Idealization* (Routledge, pp. 1-15.).
- Sugden, Robert (2000): "Credible Worlds: The Status of Theoretical Models in Economics", *Journal of Economic Methodology*, vol. 7, no. 1, pp. 169-201.
- Thomson-Jones, Martin. (2005). Idealization and Abstraction: A Framework. In Jones, M. & Cartwright, N. (eds). *Idealization XII: Correcting the Model-Idealization and Abstraction in the Sciences* (Poznań Studies in the Philosophy of the Sciences and the Humanities 86) Amsterdam/New York, NY, 2005.
- Usher, Marius. (2001). "A Statistical Referential Theory of Content: Using Information Theory to account for Misrepresentation". *Mind & Language* 16:311-334.

# Role of Kolmogorov Complexity on Interest in Moral Dilemma Stories

Antoine Saillenfest, Jean-Louis Dessalles

({antoine.saillenfest, jean-louis.dessalles}@telecom-paristech.fr)

Telecom ParisTech, INFRES, 46 Rue Barrault, 75013, Paris

## Abstract

Several studies have highlighted the combined role of emotions and reasoning in the determination of judgments about morality. Here we explore the influence of Kolmogorov complexity in the determination, not only of moral judgment, but also of the associated narrative interest. We designed an experiment to test the predictions of our complexity-based model when applied to moral dilemmas. It confirms that judgments about interest and morality may be explained in part by discrepancies in complexity. This preliminary study suggests that cognitive computations are involved in decision-making about emotional outcomes.

**Keywords:** Kolmogorov complexity; moral dilemma; moral judgment; narrative interest; emotion.

## Introduction

Humans devote a considerable amount of time to producing narratives. Spontaneous conversational narratives constitute a large amount of our conversational time (Norrick, 2000) and fictional narratives constitute a large part of human productions (*e.g.* novels, movies, video games). Modeling narrative interest and emotional impact in narratives, especially in fictional narratives, is of major importance, both scientifically and economically, as there are a variety of potential applications (*e.g.* serious games, film industry, video games). The selection of events that people consider relevant to tell is a not yet fully understood process. Only a small proportion of our experiences passes the selection. Moral dilemma belong to the situations that make good stories.

Previous studies have pointed out that emotional intensity and complexity drop have a decisive influence on narrative interest. The aim of this article is to explore the role of complexity change in the determination of morality and interest in moral dilemma.

## Morality judgments in dilemma

The number of studies on morality and emotion grew steadily in the 1980s and 1990s, and even more during the last decade. Various disciplines now investigate human morality and the interplay of emotion and reason in moral judgment and decision-making.

Moral psychology initially focused on reasoning. During the 1950s and 1960s, mental models and information processing were the preferred framework in psychology. Kohlberg (1958) proposed a six-stage developmental model of moral reasoning which, he thought, drives moral judgment. In the 1980s, however, the idea that moral emotions also play a role has been highlighted. New findings in evolutionary psychology and in primatology pointed to the crucial role of a specific set of emotions. This "affective revolution" has been reinforced during the two last decades. Recent evidence suggests

that moral judgment is more a matter of emotion and affective intuition than of deliberate reasoning (Haidt & Hersh, 2001). Emotion now plays a central role in moral psychology research (Haidt, 2007). Evidence that emotions guide moral judgments comes from brain imagery (Moll et al., 2002; Greene et al., 2004; Moll & Oliveira-Souza, 2007; Koenigs et al., 2007; Decety, Michalska & Kinzler, 2011), philosophy (Roeser, 2006), and psychology (Wheatley & Haidt, 2005; Valdesolo & DeSteno, 2006; Schnall et al., 2008).

Recently, new findings from several areas of cognitive neuroscience have suggested that emotions and reasoning both matter, but that automatic emotional processes tend to dominate (Greene & Haidt, 2002; Greene et al., 2004).

Various elements of particular importance to our study, such as social consensus, proximity (the feeling of nearness), the magnitude of consequences or the probability of effect have been shown to affect our judgment in moral dilemma (Jones, 1991). It has also been shown that people judge permissible to harm people as a side effect but not as a means (Cushman, Young & Hauser, 2006).

## Interest in narratives

Both reasoning and emotions seem to control the intensity of narrative interest. It has been observed that human cognition is sensitive to complexity, in the sense of Kolmogorov (*i.e.* the length of the minimal determination of a situation) (Chater, 1999; Chater & Vitényi, 2003). Simplicity Theory (Dessalles, 2008a; see also [www.simplicitytheory.org](http://www.simplicitytheory.org)) highlighted the role of unexpectedness in the selection of interesting events: a situation is unexpected if it is more complex to produce than to determine. This means that the generation complexity  $C_w$  of an unexpected event is higher than the complexity  $C_d$  of its determination.

$C_w$  measures the minimum quantity of information that must be given for the "world" (as the observer knows it) to make the situation happen. It evaluates the size of the minimal explication of the situation.  $C_d$  measures the quantity of information needed by the observer to describe the situation unambiguously. It evaluates the size of the minimum description of the situation.

Unexpectedness  $U$  is the difference between the generation complexity and the determination complexity ( $U = C_w - C_d$ ). The study of unexpectedness makes good predictions about which parameters control narrative interest in situations such as fortuitous encounters, atypical events, coincidences or rare events (Dessalles, 2008b; see also [www.simplicitytheory.org](http://www.simplicitytheory.org)).

Interest is also a matter of emotions. Emotional intensity plays a crucial role in the selection of narrated events (Rimé,

2005). In what follows, we will present a model in which emotional intensity results from the combination of the emotional category of the event (whether it is about ten deaths or merely about a ten-Euro loss) with unexpectedness. Then we will present an experiment designed to test the model. Finally, we will discuss the validity and the generality of this approach based on complexity.

### A complexity-based model of narrative interest and of moral judgment

The model proposed in this article intends to show that variations of Kolmogorov complexity may contribute to explain both moral judgment and interest in moral dilemma stories.

Any outcome  $i$  comes with an hypothetical emotional intensity  $E_h^i$  ( $> 0$ ) attached to it. This value does not take the valence (positive or negative) of the emotion into account, but only the standard magnitude of the corresponding class of events (sometimes considered to result from social consensus) (e.g. the death of child is supposed to be emotionally more intense than the death of an adult person, all things being equal) (Jones, 1991; Bleske-Rechek et al. 2010).

According to Simplicity Theory, the emotional intensity attached to an event is the sum of the hypothetical emotional intensity and of unexpectedness:  $E = E_h + U$ . Since  $U$  depends on the complexity of the persons involved in the event (with a minus sign), the definition of  $E$  reflects the fact that one is more affected if the victim of an accident is a close acquaintance or a celebrity (as close acquaintances or celebrities require less information to be determined).

When considering the narrative value of an event  $s$  which is the outcome of an action  $a$ , the computation of  $E$  is performed *ex post*:  $E$  is derived from  $E_h$  and  $U$ . In *ex post* calculus,  $C_w(s) = C_w(s|a) + C_w(a)$  in which  $C_w(s|a)$  measures the amount of information the world needs to produce  $s$  from  $a$  (see Figure 1). To judge the moral value of  $a$ , the computation is *ex ante*:  $s$  is evaluated from its emotional intensity  $E$ , from the causal unexpectedness  $U(s|a) = C_w(s|a) - C_d(s|a)$  and from the unexpectedness of action  $a$ .  $C_d(s|a)$  measures what is still to be determined about  $s$  once  $a$  is known. Then the moral evaluation  $E_h^a$  of  $a$  is  $E_h^a(s) = E(s) - U(s|a) - U(a)$ .

We introduce the notion of responsibility:  $R^a(s) = C_w(s) - C_w(s|a)$ . The more  $a$  makes  $s$  easy (resp. hard) to produce, the more  $R^a(s)$  increases (resp. decreases) and the more (resp. less) the actor is judged responsible for  $s$ . We also introduce the notion of targetting:  $T^a(s) = C_d(s) - C_d(s|a)$  which evaluates the contribution of  $a$  in the description of  $s$ .  $T_a(s) = C_d(s)$  means that  $s$  is fully described by  $a$ , the outcome  $s$  is targeted by the actor that does  $a$ . Lastly, we introduce the notion of inadvertence  $F^a = U(a)$  which measures how unexpected the action is. If  $F^a$  is large, then  $a$  has been done inadvertently. Eventually,  $E_h^a = R^a - T^a - F^a$ .

In this paper, we only consider premeditated actions ( $F^a = 0$ ) that do fully describe the outcomes ( $C_d(s|a) = 0$ ). Therefore:

$$E_h^a(s) = E(s) - C_w(s|a)$$

We define the emotional gain  $\Delta E_a$  of a moral dilemma as the difference between emotional intensities for the desired consequences and undesired consequences (see Figure 1).  $\Delta E_a$  estimates how satisfying the consequences of an action appear (note that  $\Delta E_a$  can be negative or positive). The model leads us to the following predictions:

1. The narrative interest  $I$  of a situation increases with its unexpectedness  $U$  of its hypothetical emotional intensity  $E_h$ .

$$I = E_h + U$$

2. The moral judgment  $MJ$ , in the case of an action, increases with the emotional gain  $\Delta E_a$ .

$$MJ = \Delta E_a$$

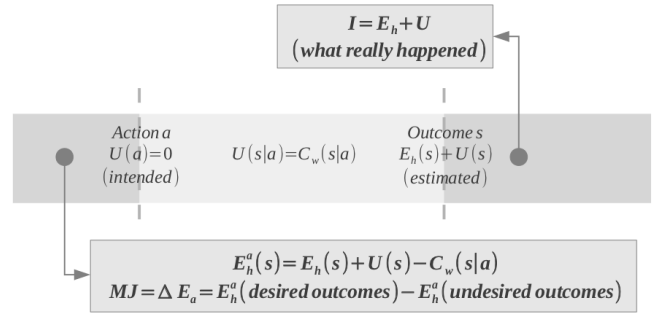


Figure 1: Complexity and Emotional intensity in a cause-to-consequence schema

The model leads to the following particular expectations:

1. A character's actions that have negative outcomes will be less morally approved if they are more direct, because their causal complexity is smaller.
2. The complexification of a causal chain will increase the narrative interest of an event, because it increases the generation complexity of the consequences, which thus appear more unexpected.
3. An action will appear more interesting, but will be less approved, if its negative consequences are simpler. In particular, an action that provokes the death of relatives or family members raises more interest (after the fact) but will be less approved.
4. Unexpected events that alter the normal course of a causal chain will have a positive influence on narrative interest.

### Experiment

Participants were asked to take the perspective of a reader and to evaluate how alternative endings of a moral dilemma story would be globally perceived by other readers on two aspects: narrative interest and moral approval of the character's action. Participants therefore were not supposed to engage their own judgment.

Participants had to read the following story (original in French):



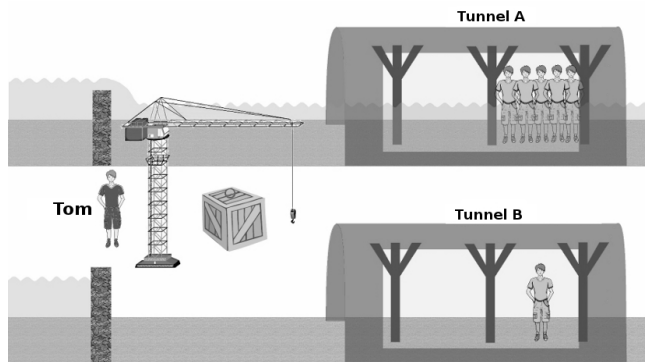


Figure 2: Illustration for the *flooded mine* dilemma

*Tom is a miner who left Scotland, his native country, to go to work in a mine in Argentina. Here is what happened to him two years ago, while he was watching the mining site upstream. The river flooded suddenly. It was flooding one of the two tunnels of the mine. Tom knew that there were five people in this tunnel (tunnel A). In the other tunnel (tunnel B), there was only one person. The water level was rising fast, and the current in the tunnel A was increasing. The trapped persons were going to drown. Tom knew that by interrupting the current in one tunnel, one would flood the other one and provoke the death of people in it. The current cannot be interrupted in both tunnels at the same time. Tom was the only one who could act. He stood at the entrance of the two tunnels, near a crane and a heavy and voluminous box.*

This *flooded mine* dilemma (Figure 2) is largely inspired from the classical trolley problem (Thomson, 1985). The context has been augmented to allow a larger variety of alternative endings.

### Method, Participants and Materials

A total of 64 individuals (aged from 19 to 65 y.o., mean 26.11 (std. dev. 7.72), 26 females) participated to the test. The study was conducted online. Participants, mainly engineering students, were recruited via social networks and billposting.

They were asked to read the flooded mine story carefully. For each of the 4 phases of the test, several alternative endings were proposed, in which actions, causal links and consequences were varied. Alternatives were presented in the same order to all participants. Using the numbering defined later in this paper, the original orders for phases 1 to 4 were 1-2-3, 3-1-4-2, 2-4-1-3 and 1-4-2-3.

For each alternative, participants were asked to answer the two following questions on a 10-point scale:

1. "According to you, will the readers of the story approve Tom's actions?" (-5: "Disapprove", 5: "Approve")
2. "According to you, will the readers of the story find the alternative interesting?" (-5: "Not Interesting", 5: "Interesting")

We omitted the zero from this scale to force participants to choose between approval or disapproval (resp. interesting or not interesting). The answering times (with standard deviation) for the 4 phases were 3'32"(1'12"), 3'12"(1'36"), 1'36"(1'17") and 2'17"(1'31"). We manually checked the answer files for individuals who provided random or uncomplete results.

The different phases of the test successively explore the role that Kolmogorov complexity plays both on interest and on moral judgment.

### Phase 1

Former studies have shown that harming actions are more likely to be judged moral if their consequences are more indirect (Cushman, Young, & Hauser, 2006). In phase 1 of the test, we tried to reproduce this result and explore how causal complexity also affects narrative interest. We also investigated the role of the complexity of the action.

#### Alternative endings of phase 1

1. Tom pushed the box in front of the entrance of the tunnel A in order to interrupt the current in this tunnel.
2. Tom got in the crane, grabbed the box with the crane's hook, brought the box above the entrance of the tunnel A and dropped it in the middle of the current to stop the current in the tunnel.
3. Tom broke the dam of tunnel B to flood this tunnel immediately.

It interrupted the current in the tunnel A, the tunnel B was flooded.

Five persons were saved, one person died by drowning

**Results** There is a main effect of intention (F-test:  $F(1, 190) = 60.87, p < 0.0001$ ) on morality but no significant effect on interest ( $F(1, 190) = 0.0045, p = 0.95$ ). A pairwise comparison of the two alternatives involving the box revealed no effect of the way the box is carried in the middle of the current on morality ( $p = 1$ ) and interest ( $p = 1$ ) (see Figure 3). In this phase, we could replicate a classical result of the trolley problem, in which people approve actions that lead to harming a victim as a side-effect but not as a means.

The main result of this first phase is that situations which are less approved by participants are not necessarily more interesting. Elements such as harming someone as a side effect or as a means only affect moral judgment. In this test, the emotional intensity of consequences is not manipulated. In alternatives 1 and 2, the causal chain between Tom's action and its harming consequence is more complex than in alternative 3; our model predicts that alternative 3 will be less approved than alternatives 1 and 2. Since all consequences are equally (un)expected, our model predicts no effect on interest.

More generally, our model predicts that actions are more likely to be approved if their positive effects are more direct and if their negative effects are more indirect.

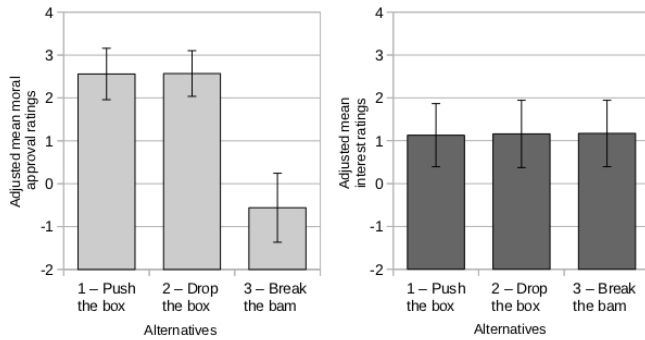


Figure 3: Results of phase 1: mean approval (left panel) and interest (right panel) ratings for three alternatives in which the proposed harmful actions vary in their intentional status (more or less direct action on the victim). Error bars indicate the 95% confidence interval.

## Phase 2

Phase 2 explores how the *length* of the causal chain of events between a causal action and its consequences affects moral judgments and narrative interest.

### Alternative endings of phase 2

Tom pulled the box in front of tunnel A.

1. As he expected, it stayed across the current because of its weight.
2. It was carried by the current and, as Tom expected, it was stopped by the struts in the tunnel.
3. It was carried by the current, hit the struts in the tunnel and, as Tom expected, some struts got broken and part of the ceiling at the entrance of the tunnel collapsed.
4. It was carried by the current and hit the struts in the tunnel; beams fell down from the ceiling; they were also carried by the current and were stopped by other struts. As Tom expected, it formed a new dam.

It interrupted the current in tunnel A. The five persons were saved, but the tunnel B got flooded and one person died by drowning

**Results** There is a main effect of the length of the causal chain of events on both morality ( $F(3, 252) = 3.01, p = 0.03$ ) and interest ( $F(3, 252) = 3.29, p = 0.02$ ) (see Figure 4)

In this phase, only causal generation complexity is manipulated. Since unexpectedness  $U$  in  $I$  is an increasing function of this complexity, longer deterministic chain of events make the outcomes more unexpected.  $\Delta E_a$  is a decreasing function of generation complexity for desired consequences. Our model correctly predicts that the outcomes will appear more interesting and that the actions will be less approved.

Jones (1991) used the expression *probability of effect* to refer to the probability that a harming event will occur. We suggest that the term of probability is not adapted, because in

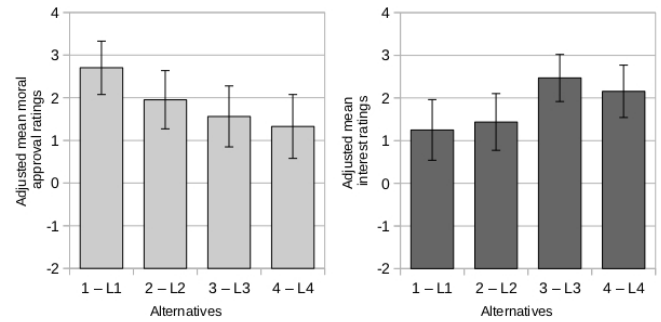


Figure 4: Results of phase 2: mean approval (left panel) and interest (right panel) ratings for four alternatives in which the proposed causal chain of events vary in their length ( $L1 < L2 < L3 < L4$ ). Error bars indicate the 95% confidence interval.

many cases, the length of the chain of events is what is relevant, even if it is deterministic. The notion of unexpectedness correctly captures this phenomenon.

## Phase 3

Previous studies on morality (Bleske-Rechek et al., 2010) have shown that the identity of victims may affect our moral judgment. This phase explores how the identity of the single victim in tunnel B influences both moral judgment and narrative interest.

### Alternative endings of phase 3

Tom knew that [#1] was in tunnel B. Tom pushed the box in front of the entrance of tunnel A. It interrupted the current in tunnel A, tunnel B was flooded. Five persons were saved, [#2] died.

[#1]/[#2] were :

1. someone / one person
2. one of his friends / Tom's friend
3. his own cousin / Tom's cousin
4. a 10-years old child / the child

**Results** There is a main effect of the identity of the victim (undefined person, cousin, friend and 10-year old child) on morality ( $F(3, 252) = 11.79, p < 0.05$ ) and interest ( $F(3, 252) = 14.26, p < 0.05$ ) ratings. A series of pairwise contrasts clarifies the nature of this interaction. The presence of an undefined person elicited significantly higher moral approval ratings (*undefined person* vs. *friend*, *cousin*, *child*, respectively:  $F(1, 126) = 11.32, 14.07$  and  $38.25$ ;  $p = 0.001, 0.0003$  and  $< 0.0001$ ) and significantly lower interest ratings (*undefined person* vs. *friend*, *cousin*, *child*, respectively:  $F(1, 126) = 34.37, 18.32$  and  $24.87$ ;  $p < 0.0001, < 0.0001$  and  $< 0.0001$ ) (see Figure 5).

Alternatives in which victims are less complex to describe (Tom's cousin and Tom's friend) are more interesting than the ones involving some undefined victim. As suggested by our

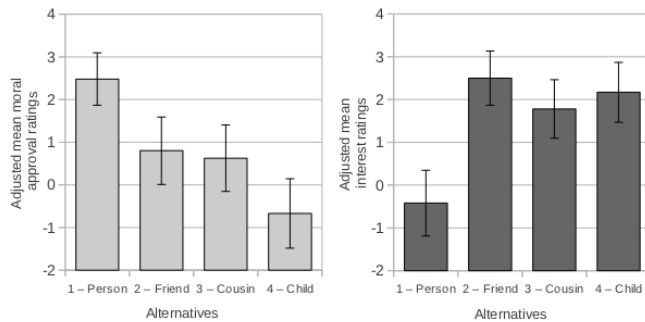


Figure 5: Results of phase 3: mean approval (left panel) and interest (right panel) ratings for four alternatives in which the victims vary in their identity. Error bars indicate the 95% confidence interval.

model, the unexpectedness increases and the emotional gain decreases when undesired consequences are simpler. The model correctly predicts that alternatives 2 and 3 will be less approved and appear more interesting than alternative 1.

The fact that the victim is a child increases the hypothetical emotional intensity. Our model correctly predicts that harming a child will be less approved but more interesting.

Our model agrees with results of Bleske-Rechek et al. (2010) and with some aspects of Jones' "components of moral intensity" such as proximity (defined as the feeling of nearness the moral agent has for the victim) and the social consensus (defined as a degree of social agreement that a proposed act is evil) (Jones, 1991).

#### Phase 4

Two elements are manipulated in phase 4: the more or less direct action of Tom and the presence or absence of unexpectedness in the course of events.

##### Alternatives

1. Tom broke the dam of tunnel B to flood this tunnel immediately. What happened was not expected. The fragments of the dam formed a new dam at the entrance of tunnel B. It was not enough to stop the current in tunnel A. Five persons died, one person was saved.
2. Tom broke the dam of tunnel B to flood this tunnel immediately. It interrupted the current in tunnel A, the tunnel B was flooded. Five persons were saved, one person died.
3. Tom pushed the box in front of the entrance of the tunnel A in order to interrupt the current in this tunnel. What happened was not expected: the box was not big enough to stop the current in tunnel A. Five persons died, one person was saved.
4. Tom pushed the box in front of the entrance of tunnel A in order to interrupt the current in this tunnel. It interrupted the current in tunnel A. Tunnel B was flooded. Five persons were saved, one person died.

**Results** There is a main effect of unexpectedness in the course of events for both morality ratings and interest rat-

ings (*non unexpectedness* vs. *unexpectedness*, respectively :  $F(1, 252) = 14.66$  and  $9.30$ ;  $p = 0.0002$  and  $0.003$ ). As in the phase 1, the action of Tom has significant effect on approval ratings (*push the box* vs. *break the dam*,  $F(1, 252) = 16.03$ ,  $p < 0.0001$ ) but no significant effect on interest ratings ( $F(1, 252) = 1.83$ ,  $p = 0.18$ ) (see Figure 6).

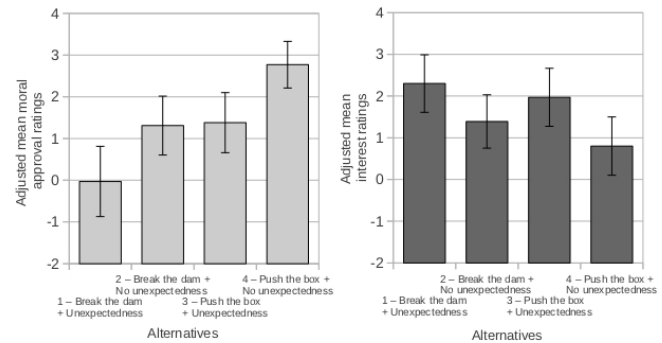


Figure 6: Results of phase 4: mean approval (left panel) and interest (right panel) ratings for four alternatives in which in which the proposed harmful actions vary in their intentional status and the courses of events vary in the presence or non presence of unexpectedness. Error bars indicate the 95% confidence interval.

These results are consistent with the observations of the previous phases. They show that variations of Tom's action only affect moral judgment. They also show that a combined increase of unexpectedness and of hypothetical emotional intensity positively influences interest but negatively affects our moral judgment about Tom's action.

This test does not only confirm previous results. It also confirms the prediction that the combination *breaking the dam + unexpected outcome* will be the least approved alternative (and the most interesting one), whereas the combination *pushing the box + deterministic result* will be the most approved alternative (and the least interesting one). However, the model is not accurate enough to predict which one among the alternatives 2 and 3 will be more approved or more interesting. Such ranking would require precise estimates of complexity values, which depend on observers' knowledge and personal history.

## Discussion

The aim of this study was to present and test a complexity-based model for narrative interest and moral judgment in moral dilemma stories. The point was to explore the role played by Kolmogorov complexity. The model, based on Simplicity Theory, intends to offer a purely cognitive account of some judgements about morality and interest in moral dilemma stories, without using *ad hoc* hypothesis about moral or immoral classes of actions. This preliminary and exploratory study has shown that the model makes correct predictions and may explain classical results about moral judgements in terms of complexity. It intends to point out the dif-

ferent but complementary roles of complexity and emotional intensity in moral judgment and narrative interest.

Our moral judgments depend on the estimated emotional gain of the dilemma, which depends not only on the emotion attached to outcomes, but also on the complexity of the causal links that lead to them. A positive gain, for example an action which saves five but kills one, may be less morally approved if it appears that this positive consequences are more complex to produce than negative ones. The gain would also be considered as uncertain or unexpected. This contributes to explain why humans are highly sensitive to actions that would jeopardize the lives of relatives, friends, or family members.

Emotional intensity also plays an important role. This may be related to Haidt's work (Haidt & Hersh, 2001) that point out the role of automatic emotional processes. Due to cultural or societal consensus, some situations appears more emotionally intense than others.

Our model makes good qualitative predictions. In future work, we will explore how quantitative parameters that influence complexity (e.g. distance in space or time) affect the emotional intensity attached to outcomes. The identification and control of these quantitative parameters open the way to a variety of potential applications, for example in the domain of decision-making. Several factors that are spontaneously attributed to the emotional component of decision-making may be reinterpreted as complexity-based computations. Such an account, if valid, would not only be relevant from a scientific perspective. It would also be potentially useful to help decision-makers evaluate or anticipate certain decisions in which emotions are supposed to play a major role.

## References

- Bleske-Rechek, A., Nelson, L., Baker, J., Remiker, M., & Brandt, S. (2010). Evolution and the trolley problem : people save five over one unless the one is young, genetically related, or a romantic partner. *Journal of Social, Evolutionary, and Cultural Psychology*, 4(3), 115–127.
- Chater, N. (1999). The search for simplicity: A fundamental cognitive principle ? *The Quarterly Journal of Experimental Psychology*, 52A(2), 273–302.
- Chater, N., & Vitányi, P. (2003). Simplicity : a unifying principle in cognitive science ? *TRENDS in Cognitive Sciences*, 7(1), 19–22.
- Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment. *Psychological Science*, 17(12), 1082–1089.
- Decety, J., Michalska, K. J., & Kinzler, K. D. (2011). The developmental neuroscience of moral sensitivity. *Emotion Review*, 3(3), 305–307.
- Dessalles, J.-L. (2008a). Coincidences and the encounter problem: A formal account. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30<sup>th</sup> annual conference of the cognitive science society (cogsci 2008)* (pp. 2134–2139).
- Dessalles, J.-L. (2008b). *La pertinence et ses origines cognitives : nouvelles théories*. Hermes-Science Publications.
- Greene, J. D., & Haidt, J. (2002). How (and where) does moral judgment work ? *Trends in Cognitive Sciences*, 6(12), 517–523.
- Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2), 389–400.
- Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 316, 998–1002.
- Haidt, J., & Hersh, M. A. (2001). Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology*, 31(1), 191–221.
- Jones, T. M. (1991). Ethical decision making by individuals in organizations: An issue-contingent model. *The Academy of Management Review*, 16(2), 366–395.
- Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446, 908–911.
- Kohlberg, L. (1958). *The development of modes of moral thinking and choice in the years 10 to 16*. Unpublished doctoral dissertation, University of Chicago. (Unpublished)
- Moll, J., & Oliveira-Souza, R. de. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences*, 11(8), 319–321.
- Moll, J., Oliveira-Souza, R. de, Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J., Andreiuolo, P. A., et al. (2002). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *The Journal of Neuroscience*, 22(7), 2730–2736.
- Norrick, N. R. (2000). *Conversational narrative: Storytelling in everyday talk*. John Benjamins Publishing Company.
- Rimé, B. (2005). *Le partage social des émotions*. Presses Universitaires de France.
- Roeser, S. (2006). The role of emotions in judging the moral acceptability of risks. *Safety Science*, 44(8), 689–700.
- Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, 34(8), 1096–1109.
- Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, 94(6), 1395–1415.
- Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, 17(6), 476–477.
- Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, 16(10), 780–784.

# Using Concept Map to Evaluate Learning by Searching

**Hitomi Saito (hsaito@aecc.aichi-edu.ac.jp)**

Aichi University of Education, 1 Hirosawa, Igaya, Kariya, Aichi, 448-8542, Japan

**Yuka Egusa (yuka@nier.go.jp)**

National Institute for Educational Policy Research, 3-2-2 Kasumigaseki, Chiyoda-ku, Tokyo, 100-8951, Japan

**Masao Takaku (TAKAKU.Masao@nims.go.jp)**

National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki, 305-0047, Japan

**Makiko Miwa (miwamaki@ouj.ac.jp)**

The Open University of Japan, 2-11 Wakaba, Mihama, Chiba, 261-8586, Japan

**Noriko Kando (kando@nii.ac.jp)**

National Institute of Informatics, 2-1-2 Hitotsubashi Chiyoda-ku, Tokyo, 101-8430, Japan

## Abstract

We defined learning while searching and browsing on the Web as "learning by searching." We used concept map to evaluate how users' knowledge structure changed as a result of their learning by searching. The influence of different topics and different scenarios was also investigated. In the experiment, participants were divided into divergent and convergent scenario groups. They were asked to assume the role of a journal editor and search for information to be presented at an editorial meeting. They performed two tasks within the topics of environment and travel. We compared the two concept maps drawn by the participants before and after the search and analyzed the web pages browsed by the participants during the tasks. The results showed that the participants' knowledge representation dynamically changed through learning as a result of searching, and the topics and scenarios influenced their processes and changed their conceptual structures before and after searching.

**Keywords:** Concept Map; Knowledge Structure; Information Seeking on the Web; Learning by Searching

## Introduction

Searching on the Web is more than just a tool for information retrieval; it has used a tool for investigating, learning, and decision making. For instance, when people buy a new digital camera, they often search and browse many digital camera sites. During these activities, they learn and acquire knowledge about digital cameras (i.e., function, structure, price, design, and size). We define this learning through searching and browsing on the Web as "Learning by Searching."

Learning by searching is considered a type of discovery learning. Discovery learning is inquiry-based, constructivist learning. The learner discover facts and relationships and new truths on their own (Bruner, 1967). Discovery learning facilitates the learner's spontaneous motivation and encourages discovery, as compared with learning by teacher instruction only. However, discovery learning is difficult for learners and research suggests that the effects of discovery learning are influenced by learning processes and the learner's skill or motivation. Learning by searching is considered to have the same benefits and problems as the discovery learning, but the factors that affect learning by searching have not been investigated. In this study, we focus on search topics and problem

contexts, and investigate how people acquire knowledge during their search and what influence search topics and situations have on their knowledge.

We used concept maps to evaluate the knowledge acquired by the users and how their knowledge structure changed as a result of their searching for information on the Web. The concept map is a graphical representation that allows people to explicitly represent their knowledge (Novak & Gowin, 1984). The concept map consists of concept words, arrows that connect the concept words, and linking words (link labels on the arrows.)

Concept maps have been used as measures to assess learners' knowledge and understanding. Meagher (2009) reported that the graph structures of concept maps became more complex from the first class in a course through the final exam. Rebich and Gautier (2005) also showed that the total number of useful items on post-course maps increased, while the total number of weak items and misconceptions decreased. Concept maps are also used to compare expert knowledge structures with novice knowledge structures. Chi, Feltovich, and Glaser (1981) compared the categories used by experts and novices and their explanations about the problems to examine the relationship between categorization and representation of physics problems. They drew concept maps to compare expert explanations with novice explanations.

Some research on information seeking behaviors has used concept maps as a means of measuring the change in a learner's knowledge. Vakkari and Pennanen (2003) explored how a student's conceptual structure is related to search tactics and search success. They reported that, between the beginning and end of the overall task, different features of the student's conceptual structures were connected to search success in terms of the useful documents they found. Cole, Lin, Leide, Large, and Beheshti (2007) focused on how students' mental model diagrams for a topic were represented in an early exploration stage of the information-seeking process and they suggested a 12-category classification schema of the mental models. Our previous researches also have studied



Figure 1: Concept map drawn by a participant for the environment topic in the divergent scenario.

how a user's concept map differs before and after a search (Egusa et al., 2010). A comparative analysis of the pre- and post-search concept maps indicated that users significantly changed their knowledge structures on a topic through learning by searching. However, there has been little research using concept maps as a measurement of how the user's knowledge has changed through the search process. Our final goal is to explore what and how people learn through searching and how their search processes influenced their learning.

## Methods

### Experimental Design

In the experiment, we focused on the influences of the topics (environment and travel) and scenarios (divergent and convergent). The participants were assigned to a factorial experiment that included two topics as within-subjects factors and two scenarios as between-subjects factors. The two within-subjects factors were counter-balanced.

### Participants

Thirty-two undergraduate students aged 20 to 23 years participated in this study (sixteen male and sixteen female). Participants

were recruited from various departments and universities in the Tokyo area. We selected participants who didn't have much experience and knowledge on the topics based on their responses to a pretest questionnaire. They were divided into two scenarios equivalent as to age, sex and major.

### Topics

The participants were instructed to assume the role of a magazine editor and to gather information on the Web in preparation for a regular magazine series on environmental and travel topics. The environment topic required the participants to introduce various environmental issues, while the travel topic asked them to present various destinations for one-day trips from Tokyo.

### Scenarios

There were two scenarios, divergent and convergent. In the divergent scenario, the participants were required to gather web pages for a series of articles to be a regular feature of the magazine. In the convergent scenario, they were required to gather pages for a single article of the regular feature. We prepared tasks for each scenario of the two topics.

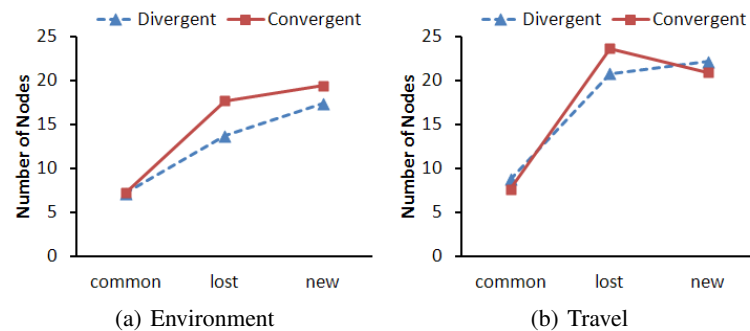


Figure 2: The average number of common, lost, and new nodes in each task.

The instruction sets for the environment topic using the divergent scenario and the travel topic using the convergent scenario are shown in Appendix A and Appendix B, respectively.

## Procedures

The participants answered questions about their experience using web search engines and the Internet on a search-experience questionnaire. They were instructed on how to create the concept maps and were given time to practice. They then received their task instructions and drew a concept map for the assigned topic with a 10-minute time limit. A blank sheet of paper with a single node of the topic, either environmental issues or one-day trips from Tokyo, was provided for drawing the concept map.

After drawing the concept map, the participants conducted a 15-minute search task. While performing the search task, the log data were recorded using a plug-in software developed by our group. After completing each search task, the participants were required to again draw a concept map about the assigned topic and to answer questions about their prior knowledge of the task, their interest in the topic, the difficulty of the task, and the difficulty gathering information. They were also asked to provide comments on the task. They then repeated the task for the other topic, from the instruction stage to answering the questionnaire.

After the two main tasks were completed, the participants answered questions comparing the two tasks and the changes in their knowledge after completing the task, and then were asked to comment on how they felt about the changes between the two concept maps, i.e., before and after the web search.

In the final session, the participants were asked to check whether the same concept could be found on both concept maps and, if such corresponding concepts were found, they were assigned the same number.

## Results

We analyzed the two (before and after search) concept maps drawn by the participants to examine whether their knowledge representation of the topic changed. We also analyzed the web pages browsed by the participants during the tasks to

investigate the relationship between the participants' conceptual changes and their search behaviors.

## Numbers of Common, Lost, and New Nodes

Figure 1 shows the concept map drawn by a participant for the environment topic in the divergent scenario. The node enclosed in a double line show the center node. The nodes enclosed in a dotted line with the same number show that the participant checked these nodes as having the same meaning in the final session.

We defined three types of change between the participants' pre and post-search concept maps. The nodes that participants identified as having the same meaning in the pre and post-search maps were defined as common nodes, nodes existing only in the pre-search map were defined as lost nodes, and nodes first appearing in the post-search map were defined as new nodes. We then analyzed the number of common nodes, new nodes, and lost nodes. For instance, Figure 1(a) has four common nodes, 13 lost nodes, and 22 new nodes.

Figure 2 shows the average number of common, lost, and new nodes in each scenario of the two tasks. A 3-way mixed ANOVA (analysis of variance) with scenario as a between-subjects factor and topic and type of change as within-subjects factors revealed that there is significant interaction between topic and type of change ( $F(2,60)=6.56, p < .01$ ). The number of lost and new nodes in the travel topic were more than those in the environment topic (lost:  $F(2,60)=21.98, p < .01$ ; new:  $F(2,60)=8.81, p < .01$ ). There also were differences among the three types of change. In the environment topic, the number of new nodes was more than the other two types of nodes and the number of common nodes was less than the other two types of nodes ( $F(2,60)=41.21, p < .01$ ). In the travel topic, the number of common nodes was less than the other two types of nodes ( $F(2,60)=49.93, p < .01$ ).

In total, there were few common nodes and a relatively large number of lost and new nodes. These results suggest that the concept maps changed greatly after the web searches were conducted.



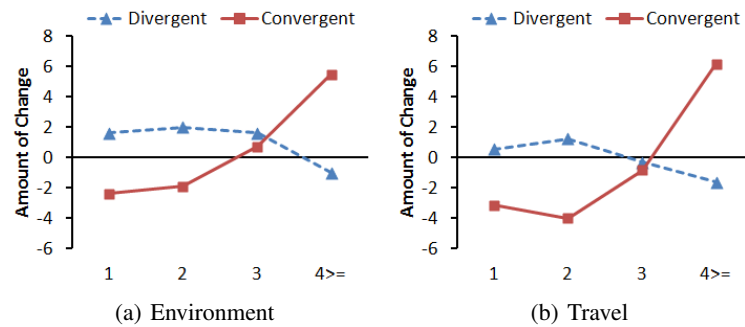


Figure 3: The differences from pre to post-search in the number of nodes at each distance.

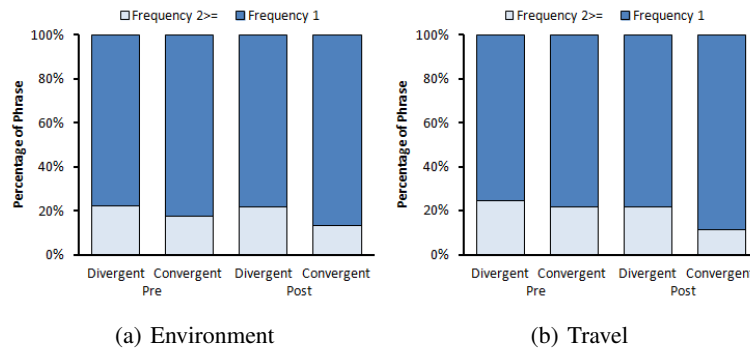


Figure 4: The percentage of phrases with a frequency of 1 or more than 2 for each topic.

### Number of Nodes at Each Distance from the Center Node

Regardless of the scenarios and topics, the participants' maps changed dynamically before and after searching. Next, we examined whether or not there were any differences in the position of each node in the map between scenarios and topics.

To analyze differences in the position of each node, we defined distance in each node. Distance in each node is measured by the number of arrows following the center. Nodes that were linked to more than two nodes and had more than two distances were counted at each distance. We counted the number of nodes at each distance from the center node, that is, the node placed at the center of the concept map. Nodes at distance 4 or higher were counted as in the same category. Moreover, to clear the differences between scenarios and topics, we calculated the amount of change for each distance from the pre to post-search maps by subtracting the number of nodes at each distance in the post-search map from those in the pre-search map. If the number of nodes in the post-search map at distance  $n$  was more than that in the pre-search map, the amount of change at distance  $n$  was considered a positive value.

Figure 3 shows the amount of change at distances 1, 2, 3, and 4 or more for the two topics in each scenario. A 3-way mixed ANOVA with topic and distance as within-subjects factors and the scenario as a between-subjects fac-

tor revealed significant interaction between scenario and distance ( $F(3,90)=14.40, p < .01$ ). The amount of change at distances 1 and 2 in the divergent scenario occurred more than those in the convergent scenario (distance 1:  $F(3,90)=27.90, p < .01$ ; distance 2:  $F(3,90)=14.23, p < .01$ ). Differences were also found for each distance in the convergent scenario. The amount of change at distance 4 or more was more than the other distances ( $F(3,90)=16.80, p < .01$ ).

### Frequency of Phrases Used in Nodes

The position of the nodes in the map were different for the two scenarios. However, it is unclear whether or not these differences were also found in the phrasing used by the participants. We therefore compared the changes in the phrases used in the maps between the two scenarios and the two topics.

Before making our analysis, we extracted phrases included in nodes from the participants' concept map. These were phrases in both the pre and post-search maps put together with respect to each scenario and topic. Because these phrases were spelled in several different ways, we adjusted the differences based on the rules listed below. Finally, we counted the frequency of the phrases.

- phrases expressed in different forms, such as "東京", "とうきょう", and "トウキョウ" (These phrases all mean "Tokyo.").

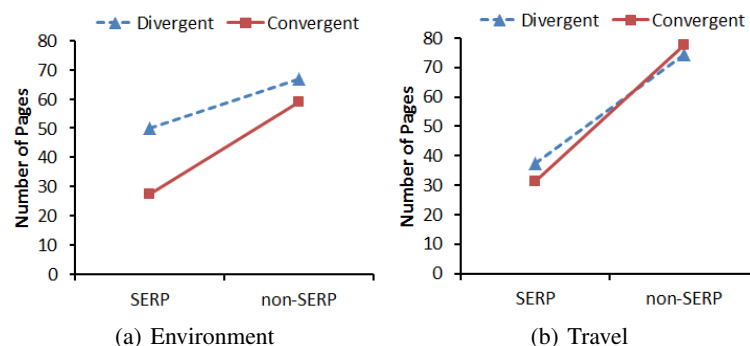


Figure 5: The number of SERP and nonSERP in each topic.

- phrases expressed with or without a particle, as in “海面上昇” and “海面上昇”.
- phrases expressed in abbreviated form or not, such as “eco” and “ecology.”
- phrases with typographical errors.

Figure 4 shows the percentage of phrases used once and more than twice in each task. On the whole, 80 % of phrases were used only once. The results of the  $\chi^2$  test revealed significant differences between the two scenarios in the post-search map (Environment:  $\chi^2(1)=6.810$ ,  $p < .01$ ; Travel:  $\chi^2(1)=13.181$ ,  $p < .01$ ). The number of phrases occurring once in the convergent scenario was greater than those in the divergent scenario for the two topics and the number of phrases with a frequency of more than two in the convergent scenario was less than those in the divergent scenario for the two topics. There were no differences in the pre-search map.

### Number of Web Pages Browsed

Finally, we examined the relationship between the participants' conceptual changes and their search behaviors. We focused on the web pages browsed by the participants during the task because their conceptual changes would be most likely influenced by them. We extracted the web page URLs from the browser logs. These URLs were classified by a logging tool as a search engine result page (SERP) or some other page (nonSERP).

Figure 5 shows the average number of SERPs and nonSERPs for each topic. A 3-way mixed ANOVA with topic and page type as within-subjects factors and scenario as a between-subjects factor revealed a significant interaction between topic and scenario ( $F(1,30)=7.08$ ,  $p < .05$ ). In the environment topic, participants in the divergent scenario browsed more SERPs and nonSERPs than participants in the convergent scenario ( $F(1,30)=10.27$ ,  $p < .01$ ). In the convergent scenario, participants browsed more SERPs and nonSERPs for the travel topic than the environment topic ( $F(1,30)=9.39$ ,  $p < .01$ ). A significant interaction was also found between topic and page type ( $F(1,30)=15.13$ ,  $p < .01$ ). The participants browsed more nonSERPs for the travel topic than the environment topic ( $F(1,30)=12.55$ ,  $p < .01$ ). The number of SERPs

was more than those of nonSERPs for the two topics (Environment:  $F(1,30)=18.16$ ,  $p < .01$ ; Travel:  $F(1,30)=42.20$ ,  $p < .01$ ).

### Discussion

We defined learning during searching and browsing on the Web as “learning by searching.” We investigated how searchers' conceptual knowledge of topics changed by comparing their before search and after search concept maps. The influence of different topics and different scenarios was also investigated. In the experiment, participants were divided into divergent and convergent scenario groups. They were asked to assume the role of a journal editor and search for information to be presented at an editorial meeting. They also performed two tasks under the topics of environment and travel.

We analyzed the before and after search concept maps drawn by the participants to examine whether their knowledge representation of the topic changed during the search. The results show that their concept maps dynamically changed after the search regardless of the topics or scenarios. Egusa et al. (2010) also reported the same results. These results indicate that searching and browsing web pages strongly influences a searcher's knowledge structure of a topic. Why did their conceptual structure change so dramatically? One likely reason is that the participants searched and browsed web pages that suited their interests, which allows them to easily build new knowledge onto their existing knowledge base. An alternate reason is that they had this knowledge already and they reconstructed their knowledge through learning by searching. However, these changes might be temporary and forgotten with the passage of time. We need to further examine the continuousness of learning by searching.

Results also demonstrated the influence of topics and scenarios. The number of lost and new nodes in the travel topic was greater than those in the environment topic. These results indicate that there are differences in the knowledge structures for the topics and that the concept map can represent these differences. Cole et al. (2007) also found the differences between the concept map drawn by history students and those by psychology students.

Differences between the scenarios were found in the node distances and phrase frequencies. In the divergent scenario, nodes placed near the center increased and nodes placed far from the center decreased. On the other hand, nodes placed far from the center increased in the convergent scenario. As can be seen from the examples in Figure 1, the node concepts changed from general to specific as distance from the center node increased. Therefore, the participants in the convergent scenario acquire specific knowledge while the participants in the divergent scenario acquire general knowledge. A similar tendency was seen in the analysis of the frequency of phrases used in the nodes. In the convergent scenario, the phrases used once increased and the phrases used more than twice decreased.

We also analyzed the web pages browsed by the participants during the tasks to investigate the relationship between the participants' conceptual changes and their search behaviors. Differences between the scenarios were only seen in the environment topic. The number of SERPs and nonSERPs in the divergent scenario was greater than those in the convergent scenario. This indicates that the participants in the divergent scenario searched and browsed quickly to gather a wide scope of information about the environment topic, while the participants in the convergent scenario browsed web pages more minutely to gather detailed information about the specific theme of the environment topic. We presume that these web page results relate to the above concept map results. However, with the travel topic, we could not find differences between the scenarios. We must analyze other process data to explicitly capture the relationship between the searchers' search processes and the changes in their knowledge representation.

Finally, we discuss on the significance of our study. In this study, we compared different scenarios about the same topic. We found that these differences of scenario influenced breadth and depth of searching and quality of knowledge acquisition. In the discovery learning, it is also important to explicit relationship between how students explored and what they learned for helping students learn and discover effectively.

## References

- Bruner, S. J. (1967). *On knowing: Essays for the left hand*. Cambridge: Harvard University Press.
- Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152.
- Cole, C., Lin, Y., Leide, J., Large, A., & Beheshti, J. (2007). A classification of mental models of undergraduates seeking information for a course essay in history and psychology: preliminary investigations into aligning their mental models with online thesauri. *Journal of the American Society for Information Science and Technology*, 58, 2092-2104.
- Egusa, Y., Saito, H., Takaku, M., Terai, H., Makiko, M., & Kando, N. (2010). Using a concept map to evaluate exploratory search. In *Proceedings of the third symposium on information interaction in context* (pp. 175-184). New York, NY: ACM.
- Meagher, T. (2009). Looking inside a student's mind: can an analysis of student concept maps measure changes in environmental literacy? *Electronic Journal of Science Education*, 13, 1-28.
- Novak, D. J., & Gowin, B. D. (1984). *Learning how to learn*. New York, NY: Cambridge University Press.
- Rebich, S., & Gautier, C. (2005). Concept mapping to reveal prior knowledge and conceptual change in a mock summit course on global climate change. *Journal of Geoscience Education*, 53, 355-365.
- Vakkari, P., & Pennanen, M. (2003). Students' conceptual structure, search process, and outcome while preparing a research proposal: a longitudinal case study. *Journal of the American Society for Information Science and Technology*, 54, 759-770.

## Appendix

### A. Instructions for environment in divergent scenario

Please carry out this task as if you were in the following situation: You are working as an editor for a news magazine published by a national newspaper. Your chief has asked you to gather information to introduce various environmental issues for a regular series that will begin with the next issue. The chief will hold a meeting with other managers from several departments to discuss this regular series.

In the meeting, the chief would like to present an overall picture of each article for the series. For that purpose, you must gather various pieces of information that are sufficient for discussing which topic should be selected for each article of the series, rather than gathering detailed information for just a single article in the series.

Please choose the most interesting topics for the readers. You have just 15 minutes to search the Web and find the needed information. If you find a useful web page for the topic, add it to the bookmarks.

### B. Instructions for travel in convergent scenario

You are working as an editor for a travel magazine published by a major publishing company. Your chief has asked you to gather information to introduce various one-day trips for a regular series that will begin with the next issue. The chief will hold a meeting with other managers from several departments to discuss this regular series.

In the meeting, the chief would like to present the specific contents for one article of the series. For that purpose, you must gather detailed information for a single article of the series rather than information for several trips.

Please choose the most interesting topics for the readers. You have just 15 minutes to search the Web and find the needed information. If you find a useful web page for the topic, add it to the bookmarks.

# Towards a Cognitive Science of Literary Style: Perspective-Taking in Processing Omniscient versus Objective Voice

Manami Sato (msato@hiroshima-u.ac.jp)

Hiromu Sakai (hsakai@hiroshima-u.ac.jp)

Graduate School of Education, Hiroshima University  
1-1-1 Kagamiyama, Higashi-Hiroshima City, Japan 739-8524

Jennifer Wu (jjw012@ucsd.edu)

Benjamin K. Bergen (bkbergen@cogsci.ucsd.edu)

Department of Cognitive Science, University of California, San Diego  
9500 Gilman Dr., La Jolla, CA 92093-0515

## Abstract

What are the consequences of narrative style for the cognitive operations that comprehenders perform? Third person narratives can adopt different voices. Omniscient voice has access to the mental states of characters, while objective voice only describes how characters would appear to an observer. It's currently unknown what cognitive consequences different voices have for people processing third person language. We hypothesize that in building representations of described scenes, omniscient voice may make comprehenders more likely to adopt the internal perspectives of characters than objective voice. We tested this prediction in a narrative-image matching study. Participants read short passages describing a third person character in either omniscient or objective voice. They then saw an image that either depicted the described scene or not, and which depicted the event from the perspective of the character or not. Their task was to decide as quickly as possible whether the image matched the narrative. In cases where the narrative and image matched, participants were significantly faster to indicate the correct decision when the narrative voice and the image perspective matched—that is, an image from the character's perspective after an omniscient narration or an image from a different perspective after an objective narration. This finding provides the first evidence that narrative voice affects the perspective from which comprehenders represent described scenes.

**Keywords:** language processing; perspective; narrative voice; literary style; grammatical person; mental simulation

## Introduction

Understanding the processes that underlie language comprehension is among the primary concerns of cognitive science, and justifiably so. Language is pervasively, and uniquely, human. The study of the cognitive processes underlying language comprehension has, also justifiably, begun by focusing on how people process words or sentences in isolation. Yet this is not language's natural state in the wild. We mostly interact with words and sentences embedded in context—social, physical, and linguistic. And one—though not the only—context in which words and sentences appear is in narrative. Across cultures, humans recount and process accounts of sequences of events, whether purportedly fictive or factive. We're all

story-tellers, from marketing directors to kindergarteners to shamans. The outstanding question for cognitive science is how we go about understanding narratives, with all their stylistic peculiarities. What cognitive operations do we perform to go from a story to understanding?

## Narrative Voice

A key feature of every narrative is that it is told using a particular (though possibly variable) narrative voice. Here, by way of illustration, is an example of two different narrative voices that appear in the same text. Ernest Hemingway's *The Old Man and the Sea* features an eponymous old man. Throughout the narrative, we read different kinds of descriptions of him. Early on, we read:

*The old man was thin and gaunt with deep wrinkles in the back of his neck. The brown blotches of the benevolent skin cancer the sun brings from its reflection on the tropic sea were on his cheeks. The blotches ran well down the sides of his face and his hands had the deep-creased scars from handling heavy fish on the cords. But none of these scars were fresh. They were as old as erosions in a fishless desert.*

Compare this with a passage that follows, as we're getting to know the old man a little better:

*He was asleep in a short time and he dreamed of Africa when he was a boy and the long golden beaches and the white beaches, so white they hurt your eyes, and the high capes and the great brown mountains. He lived along that coast now every night and in his dreams he heard the surf roar and saw the native boats come riding through it. He smelled the tar and oakum of the deck as he slept and he smelled the smell of Africa that the land breeze brought at morning.*

Although these two passages both describe the old man in the third person, they differ in terms of their narrative voice. While the first describes properties of the old man that can be viewed by an outside observer, the second omnisciently enters the old man's mind, so that the narrator is able to

recount aspects of the old man's mental life that would only be known to him.

For cognitive scientists, the first question is what the consequences of narrative voice choices are for comprehenders. Do we process omniscient voice differently from objective voice? And if so, in what way? Yet, to date, we know of no work addressing narrative voice and how it affects language processing.

## Perspective in Language Processing

One thing that's quite clear from recent work on language processing is that comprehenders construct detailed mental representations of scenes that they read or hear about. These are variously described in different literatures as *situation models* (Zwaan & Radvansky, 1998) or *mental simulations* (Barsalou, 1999). It also seems clear that these mental representations are often constructed from a particular perspective within the described scene. Different features seem to affect the perspective a comprehender will adopt (D'Argembeau, Comblain, & Van der Linden, 2002; Frank & Gilovich, 1989; Nigro & Neisser, 1983; Robinson & Swanson, 1993), but these at the very least include the types of actions the narrative describes (Borghi, Glenberg, & Kaschak, 2004) and the grammatical person of the narrative (for instance, 2<sup>nd</sup> versus 3<sup>rd</sup> person; Brunyé et al., 2009).

The effects of grammatical person on mental representations are particularly relevant to voice—in fact, deciding whether a narrative should use 1<sup>st</sup> person (*I*) or 3<sup>rd</sup> person (*he*), is a dimension of voice. There's been some work (Brunyé et al., 2009) showing that when people process 3<sup>rd</sup> person language, they are more likely to mentally represent the described scenario from the viewpoint of an outside observer (they adopt an *external* perspective) than from that of a character (an *internal* perspective). By way of comparison, 2<sup>nd</sup> person language (about *you*) is more likely to induce an *internal* perspective.

And yet, not all third person narratives are alike, as the Hemingway passages illustrate. There's evidence that the more a comprehender identifies with a character, the more likely he or she is to adopt that character's perspective when mentally reconstructing the described scene (Libby & Eibach, 2002; Libby, Eibach, & Gilovich, 2005).

So voice might make a difference. In cases where the narrator omnisciently describes a character's mental states, it could well be that this draws the comprehender into adopting that character's perspective in mental representations of the described scene. By contrast, objective voice, which describes characters as they would be viewed externally, might be more likely to induce external perspectives in mental representations of described scenes.

This reasoning leads to two key questions. First of all, is it ever the case that third person language can systematically lead comprehenders to adopt an internal perspective, rather than the external perspective that has previously been shown to predominate with third person language? And second, if it is, what is it about certain third person narratives that leads comprehenders to adopt an internal perspective?

We pursued both of these questions through an experiment looking at one property of narratives—narrative voice—that presents itself as a viable candidate for engaging internal perspectives in comprehenders. This is a first step in applying tools used to address comprehension processes to the stylistic details of narrative—a step in the direction of a cognitive science of literary style.

## Method

In order to investigate whether third person language using different narrative modes—objective or omniscient—induces mental representations of events from different perspectives in comprehenders, we adapted a method first used by Brunyé et al. (2009). We began by creating pairs of four-sentence narratives in English; one member of each pair used omniscient voice and the other used objective voice. Each pair of narratives concluded with the same fourth sentence. Native speakers of English read one narrative from each pair, and then saw a picture that either depicted the scene described in the narrative or not. The participant's task was to decide whether the depicted event could be part of the preceding story. We were only interested in those trials in which the image did depict the scene. These images depicted the scene from either an internal perspective (as if the reader were the character described in the narrative) or an external perspective (as if the reader were an outside observer of the action performed by the character). We predicted that if readers were more likely to adopt an internal perspective when reading omniscient voice narratives, this should make them faster to indicate their judgments about the internal perspective images following omniscient voice narratives, and conversely, if they were more likely to adopt an external perspective while processing objective voice narratives, then objective voice should make them faster when confirming matching external perspective images.

## Participants

Fifty-eight native speakers of English who were undergraduate students at the University of California, San Diego participated in this study in exchange for course credit. All participants reported normal or corrected-to-normal hearing and vision.

## Language Materials

We created twenty-four pairs of narratives. Within each pair, the two narratives were made up of three sentences that differed, but they both ended with a fourth sentence that was the same. Each narrative was entirely in the third person, and used one of the two narrative modes: omniscient or objective. Omniscient narratives included information about the mental states of the protagonist, while objective narratives only described externally visible features of the protagonist. The final sentence in both conditions described an event where the protagonist manipulated the given object with her hand (e.g., she threw away, grabbed, peeled off, or picked up the object). In addition to these twenty-four

critical narratives, which were paired with matching pictures, as described below, we also created twenty-four filler stories, half in the omniscient and half in the objective mode. These were paired with non-matching pictures. Here is a sample pair of critical narratives:

(1a) Third person omniscient narrative

She was very uncomfortable because her hands felt sticky and there was still clay under her nails from her ceramics class.

She desperately wanted to wash her hands, but could not see a sink anywhere.

She could feel the clay drying even more and eyed the small towel on the table.

She picked up the hand towel.

(1b) Third person objective narrative

She appeared out of breath when she rushed into the room.

She looked down at the table, where there was a hand towel.

Her hands were covered with clay, and she glanced back and forth between her clay-covered hands and the towel.

She picked up the hand towel.

In both narrative modes, the number of third person pronouns *she* was matched (mean: 4.8 for omniscient, 4.4 for objective), and the total number of words used for each set of items was similar (mean: 52.5 for omniscient and 56.3 for objective).

## Image Materials

The experiment used forty-eight critical pictures (twenty-four internal and twenty-four external perspectives) and twenty-four filler pictures (twelve internal and twelve external perspectives). For each set of critical sentences, corresponding pictures were taken to create a set that depicted the event from the internal perspective (i.e., a protagonist's or performer's viewpoint) as shown in Figure 1(a), and from the external perspective (i.e., an outside observer's viewpoint) as shown in Figure 1(b). Images were photographs taken using a tripod to ensure that all internal and all external images were taken from the same angle.

(a) Internal perspective



(b) External perspective



Figure 1: Internal versus external perspective images

## Procedure

Participants were tested individually. The experiment began with a set of four practice trials, followed by the experimental session, which was composed of twenty-four criticals (requiring “yes” responses) randomly mixed with twenty-four fillers (requiring “no” responses). In the experimental session, each participant viewed twelve critical and twelve filler item-sets in the omniscient narrative mode, and twelve criticals and twelve fillers in the objective narrative mode.

Each trial began with a fixation cross for 500 ms, followed by the first sentence in the middle of the screen. Participants pressed the spacebar as soon as they finished reading the sentence, at which point it was replaced on the screen by the next sentence. After the fourth and final sentence, participants saw another fixation cross for 500 ms, followed by a picture depicting, from either an internal or external perspective, an image that either was or was not part of the scene described in the story that they had just read. Participants then indicated if the pictured event was mentioned in the prior set of sentences, as quickly and accurately as possible, by pressing a button (“1” for “yes” or “a” for “no”).

Participants were asked to answer “yes” when the depicted scene was part of the prior story. No instructions were given regarding the different perspectives that images used or the different narrative voices, so as not to draw attention to these dimensions of the manipulation.

To ensure that participants paid equal attention to each of the four sentences in the narratives, every trial was followed by a comprehension question (after picture verification) that addressed one of the four sentences in the set, in equal proportions. We recorded the responses (i.e., “yes” or “no”) and measured the reaction times for picture verification and responses to comprehension questions.

The two independent variables, Narrative Mode (Omniscient, Objective) and Picture Perspective (Internal, External) were fully crossed and manipulated within participants. The four experimental conditions produced by crossing these two variables were equally assigned to four lists in a Latin-square design, resulting in six experimental items in each condition for each participant. Likewise, the twelve omniscient and twelve objective filler stories were followed by half internal, half external perspective pictures that depicted objects unrelated to the preceding scene.

## Predictions

We predicted that if third person omniscient narratives lead participants to project themselves into the protagonist and accept an internal perspective, participants should respond faster to internal perspective pictures, whose perspective matched that evoked by the preceding story, than to external ones, which mismatched. Conversely, if third person objective narratives drive participants to adopt an outside observer's perspective that has clear mental distance from the protagonist, it should facilitate responses to external perspective pictures.

## Results

Three participants were excluded for being left-handed, and an additional three participants were excluded due to low accuracy (below 80%) to picture verification or question comprehension. One item was excluded for its low accuracy rate (below 80%). The image, of sushi, may have been problematic because it depicted a type of sushi not typically available in the United States, which might have confused participants. Extremely slow responses (those over 4000 ms), incorrect responses to picture verifications and/or to comprehension questions, and responses that were more than 2.5*sd* above or below the mean response time for each participant were removed. This resulted in eliminating 11.8% of the data (3.0% exclusion due to incorrect picture verification, 4.8% due to inaccurate response to comprehension questions, and 4.0% due to the outliers).

Two-way Repeated-Measures ANOVAs revealed no significant main effect of Narrative Mode ( $F_1(1,51) = 1.0$ ,  $p = 0.3$ ,  $\eta^2_p = 0.02$ ;  $F_2(1,22) = 1.4$ ,  $p = 0.3$ ,  $\eta^2_p = 0.06$ ). Picture Perspective produced a non-significant main effect in the subject analysis ( $F_1(1,51) = 1.3$ ,  $p = 0.3$ ,  $\eta^2_p = 0.03$ ) but reached a significant effect in the item analysis ( $F_2(1,22) = 6.9$ ,  $p = 0.02$ ,  $\eta^2_p = 0.24$ ). However, as we predicted, Narrative Mode and Picture Perspective produced significant interaction effects ( $F_1(1,51) = 6.6$ ,  $p = 0.01$ ,  $\eta^2_p = 0.12$ ;  $F_2(1,22) = 5.5$ ,  $p = 0.03$ ,  $\eta^2_p = 0.2$ ) (Figure 2).

Planned pairwise *t*-tests showed that external-perspective pictures were verified significantly faster after participants read the third person sentences framed in the objective mode than after their counterparts in the omniscient mode (mean RTs for external-perspective pictures: 1510 ms after omniscient narratives, 1385 ms after objective narratives;  $t_1 = 2.3$ ,  $p = 0.03$ ;  $t_2 = 2.1$ ,  $p = 0.049$ ). The converse was true as well: internal perspectives were verified numerically faster after reading sentences formulated in the omniscient mode than after reading sentences in the objective mode. However, the differences did not reach significance (mean RTs for internal-perspective pictures: 1392 ms after omniscient narratives, 1432 ms after objective narratives;  $t_1 = 0.8$ ,  $p = 0.4$ ;  $t_2 = 0.4$ ,  $p = 0.7$ ). The most robust difference was found in the picture verification time after participants read the omniscient narratives (mean RTs after reading omniscient narratives: 1510 ms for external-perspective pictures, 1392 ms for internal-perspective pictures;  $t_1 = 2.6$ ,  $p = 0.01$ ;  $t_2 = 3.8$ ,  $p = 0.001$ ), while only a numerical difference was found after participants read the objective narratives (mean RTs after reading objective narratives: 1385 ms for external-perspective pictures, 1432 ms for internal-perspective pictures;  $t_1 = 1.1$ ,  $p = 0.28$ ;  $t_2 = 0.1$ ,  $p = 0.9$ ).

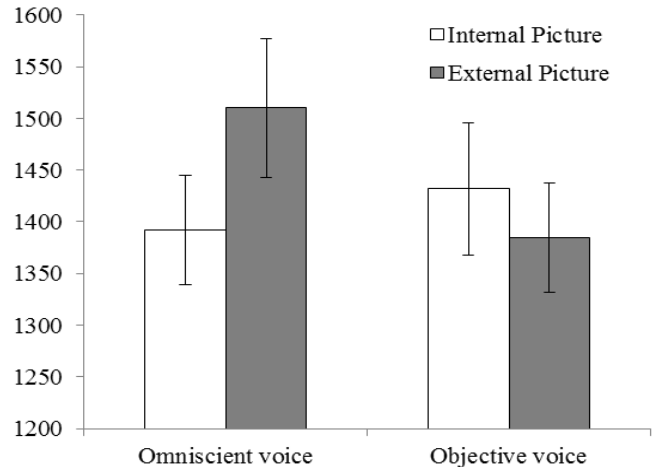


Figure 1: Mean RTs for picture verification, demonstrating an interaction effect between Narrative Mode (Omniscient vs. Objective) and Picture Perspective (Internal vs. External). Error bars indicate standard error.

In order to assess whether there could be a speed-accuracy tradeoff present in these data, we conducted an error analysis. We ran Repeated-Measures ANOVAs with Accuracy as the dependent measure and Narrative Mode and Picture Perspective as the independent measures. These revealed no significant main effects of Narrative Mode ( $F_s < 1$ ) or Picture Perspective ( $F_s < 1$ ), nor any interaction effect ( $F_s < 1$ ).

We also conducted post-hoc analyses to determine whether perspective adoption effects differed according to participants' sex; as the narratives all described a female protagonist, it's possible that female participants were more likely to adopt an internal perspective on the described scenes. We had unbalanced numbers of male and female participants (male = 15, female = 37). Three-way Repeated-Measures ANOVAs with Reaction Time as the dependent measure and Narrative Mode, Picture Perspective, and Participant's Sex as the independent variables did not show a significant three-way interaction of Gender by Narrative Mode by Picture Perspective ( $F_1(1,50) = 2.4$ ,  $p = 0.1$ ,  $\eta^2_p = 0.05$ ). This might indicate that perspective adoption is not significantly affected by participants' gender.

## Discussion

We investigated the notion that language comprehenders construct mental representations of described scenes by adopting particular perspectives, perspectives that, by hypothesis, might be affected by the narrative voice used. Previous work (Brunyé et al., 2009) has shown that third person language tends to elicit an external perspective on described events, but we found that it did not do so across the board. Rather, as predicted, the adopted perspective was modulated by the type of narrative voice. Omniscient voice made comprehenders significantly more likely to adopt an internal perspective, while objective voice made them quantitatively more likely to adopt an external perspective.



The asymmetry between the strong effect of omniscient voice and the relatively weaker one of objective voice may relate to the fact that all narratives and images in the experiment described or depicted to the same female protagonist. It's possible that there was a cumulative effect over the course of the experiment whereby participants came to identify with the protagonist. This might have resulted in third person objective language, which according to previous results should have facilitated external images, instead driving internal mental representations of described events.

In general, these results we observed are compatible with previous work showing that language comprehenders not only construct detailed mental representations of described scenes, but do so from a particular perspective. Critically, this study adds to the existing literature on perspective by showing that person is not the only linguistic factor that can push the comprehender's adopted perspective around—narrative voice appears to have a similar effect. The perspective a comprehender adopts in constructing mental representations of described scenes is the flexible product of stylistic aspects of the narration itself. It is already known that the same person does not always lead to the same perspective in comprehension. Previous research conducted by Brunyé et al. (2009) shows that first person *I* in an isolated sentence is more likely to induce an internal perspective, but when it's embedded in a richer discourse context, it tends to evoke an external perspective. Our findings show a similar flexibility for third person language, but whereas it is the presence or absence of a back-story that has been shown to modulate perspective during first person language processing, we found that the style of the narrative context itself modulates perspective in third person language processing.

Our specific finding is that when the protagonist's internal states are described, as in a third person omniscient narrative, readers are more likely to adopt the visual perspective of the relevant character. This may be related to empathetic projection in which a reader engages with that character (Carr et al., 2003; Decety & Sommerville, 2003; Lamm, Batson, & Decety, 2007; Perrine & Decety, 2004). Reading descriptions of the mental and emotional states of a character might lead comprehenders to identify with and imagine themselves as that character. As a result, omniscient narration might not only lead to measurable differences in the visual perspective comprehenders adopt, but also influence the extent to which they adopt the affective perspective of a character—the extent to which they recreate, while reading, the emotions that a character might experience (Havas, Glenberg, & Rinck, 2007).

Our results may also be related to resonant projection—observers are more likely to find a situation resonant when they share action-relevant characteristics with the actor, such as a viewpoint in hand-object interacting events (Bruzzo, Borghi, & Ghirlanda, 2008), similar motor competence (Calvo-Merino et al., 2006), or relevant motor knowledge or expertise (Calvo-Merino et al., 2010).

Readers do not adopt a fixed perspective that is evoked by language, and it might be that the modulation of perspective by narrative voice works similarly to these other factors that influence projection.

## Conclusion

In sum, the results we've reported here add to the existing literature on perspective adoption in language comprehension by showing that narrative style has an impact on whether comprehenders view a described scene from an internal or an external perspective. In some circumstances, third person language can be just as effective at transporting the comprehender into the described experience of a character as second person language can. Research like this shows the promise of using empirical cognitive science methods to explore the effects of literary style. Readers are, after all, humans, and reading is, after all, a cognitive behavior. Understanding the effects of cognitive style is well within our reach.

## Acknowledgments

This research was supported by Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research (A) “Typological Variation in Human Language and Plasticity of Neurocognitive Systems (PI: Hiromu Sakai, #23242020)” and Grant-in-Aid for Scientific Research (B) “Neurocognitive Basis for Language Learning through the Processing of Input and Output (PI: Hiromu Sakai, #20320060).”

## References

- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577–660.
- Borghi, A. M., Glenberg, A. M., & Kaschak, M. P. (2004). Putting words in perspective. *Memory & Cognition*, 32, 863–873.
- Brunyé, T. T., Ditman, T., Mahoney, C. R., Augustyn, J. S., & Taylor, H. A. (2009). When you and I share perspectives: Pronouns modulate perspective-taking during narrative comprehension. *Psychological Science*, 20, 27–32.
- Bruzzo, A., Borghi, M. A., & Ghirlanda, S. (2008). Hand-object interaction in perspective. *Neuroscience Letters*, 441, 61–65.
- Calvo-Merino, B., Ehrenberg, S., Leung, D., & Haggard, P. (2010). Experts see it all: Configural effects in action observation. *Psychological Research*, 74, 400–406.
- Calvo-Merino, B., Grezes, J., Glaser, E. D., Passingham, E. R., & Haggard, P. (2006). Seeing or doing? Influence of visual and motor familiarity in action observation. *Current Biology*, 16, 1905–1910.
- Carr, L., Iacoboni, M., Dubeau, M. C., Mazziotta, J. C., & Lenzi, G. L. (2003). Neural mechanisms of empathy in humans: A relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences*, 100, 6055–6060.

- Sciences of the United States of America*, 100(9), 5497–5502.
- D'Argembeau, A., Comblain, C., & Van der Linden, M. (2002). Phenomenal characteristics of autobiographical memories for positive, negative, and neutral events. *Applied Cognitive Psychology*, 17(3), 281–294.
- Decety, J., & Sommerville, J. A. (2003). Shared representations between self and other: A social cognitive neuroscience view. *Trends in Cognitive Science*, 7(12), 527–523.
- Frank, M., & Gilovich, T. (1989). Effect of memory perspective on retrospective causal attributions. *Journal of Personality and Social Psychology*, 57(3), 399–403.
- Havas, D. A., Glenberg, A. M., & Rinck, M. (2007). Emotion simulation during language comprehension. *Psychonomic Bulletin & Review*, 14(3), 436–441.
- Lamm, C., Batson, C. D., & Decety, J. (2007). The neural substrate of human empathy: Effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience*, 19(1), 42–58.
- Libby, L. K., & Eibach, R. P. (2002). Looking back in time: Self-concept change affects visual perspective in autobiographical memory. *Journal of Personality and Social Psychology*, 82, 167–179.
- Libby, L. K., Eibach, R. P., & Gilovich, T. (2005). Here's looking at me: The effect of memory perspective on assessments of personal change. *Journal of Personality and Social Psychology*, 88, 50–62.
- Nigro, G., & Neisser, U. (1983). Point of view in personal memories. *Cognitive Psychology*, 15, 467–482.
- Perrine, R., & Decety, J. (2004). How would *you* feel versus how do you think *she* would feel? A neuroimaging study of perspective-taking with social emotions. *Journal of Cognitive Neuroscience*, 16(6), 988–999.
- Robinson, J., & Swanson, K. (1993). Field and observer modes of remembering. *Memory*, 1(3), 169–184.
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123, 162–185.

# Beyond one's own understanding: How text comprehensibility affects laypeople's decision about scientific claims

**Lisa Scharrer (lisa.scharrer@uni-muenster.de)**

Institute for Psychology, University of Münster, Fliednerstrasse 21,  
48149 Münster, Germany

**M. Anne Britt (britt@niu.edu)**

Psychology Department, Northern Illinois University, 363 Psychology-Math Building,  
DeKalb, IL60115, USA

**Marc Stadtler (marc.stadtler@uni-muenster.de)**

Institute for Psychology, University of Münster, Fliednerstrasse 21,  
48149 Münster, Germany

**Rainer Bromme (bromme@uni-muenster.de)**

Institute for Psychology, University of Münster, Fliednerstrasse 21,  
48149 Münster, Germany

## Abstract

The study investigated whether the facilitating effect of high text comprehensibility on lay recipients' inclination to rely on their own decision about scientific claims is mitigated if the presented information is controversial. Moreover, it was assessed whether the impact of both information features is mediated by perceptions of topic complexity. Lay readers read medical text information varying in comprehensibility and controversiality and indicated their agreement strength and confidence with contained claims. Results revealed that participants' reliance on their own agreement decision was stronger after reading comprehensible than incomprehensible texts, but this difference was larger in case of uncontroversial than controversial information. However, these effects were not mediated by perceptions of topic complexity.

**Keywords:** knowledge evaluation; expertise; text comprehensibility; controversiality

## Introduction

Laypeople often find themselves in situations where they have to come to a judgment about the veracity of science-based knowledge claims, e.g. when they need to decide whether to undergo a certain medical treatment. The ease of accessing information on the Web has eliminated problems regarding the availability of science knowledge that might act as a basis for such decisions. The great challenge lies instead in the evaluation of accessed information and the knowledge claims they contain regarding their acceptability (Mason, Ariasi & Boldrin, 2011). The reason why this evaluation presents such a challenge is that we all are laypeople with regards to most science domains and lack the background knowledge and specialized training usually required for making well-founded decisions about knowledge claims. Consequently, laypeople are generally not in a position to evaluate claims by themselves but rather depend on experts for decision support (Bromme, Kienhues & Porsch, 2010; Keil, 2008). Hence, when confronted with

the necessity to decide about a science-based claim, laypeople can choose between two major courses of action: They can either rely on their own judgment despite their lay status or they can outsource the decision to an expert. However, given their own epistemic limitations, relying on experts is most often the appropriate and sensible strategy.

If laypeople nevertheless rely on their own claim decision, this should be reflected in their inclination to be easily persuaded by provided claim-related information that contains apparently sound arguments. Only if they see themselves in a position to decide about information quality should recipients make strong judgments in favor of a claim. Conversely, they should be more hesitant to strongly agree if they do not feel ready to decide. In addition, laypeople's willingness to rely on their own judgments should be reflected in high confidence in their claim decision, i.e. strong trust in their own judgment and a weak desire to hand over the decision to an expert (Scharrer, Bromme, Britt & Stadtler, 2012).

## Text Features influence Laypeople's Reliance on their own Decisions

Previous research indicates that laypeople's readiness to rely on their own decision about the veracity of science-based claims can be influenced by features of the text information they read about the topic.

**Information Comprehensibility** One such text feature is the comprehensibility of contained information. In popular scientific reports addressed to the lay public, information is frequently simplified in order to make it understandable for the audience (Wagner, Elejabarrieta & Lahnsteiner, 1995). Such simplification can be achieved through enhancement of comprehensibility, e.g. by translating technical terms. But while an increase in comprehensibility facilitates laypeople's ability to follow what is being said, it may also

raise their inclination to rely on their own decisions. A layperson's experience of easily understanding information about a scientific topic may induce the impression that they are not in need of expert support to reach a decision about related information. Previous findings support this assumption by showing that science-based information simplified by means of increased comprehensibility leads laypeople to more strongly and confidently agree with claims than incomprehensible information (Eagly, 1974; Scharrer et al., 2012).

Hence, while it may be worthwhile to simplify information for the sake of allowing the lay public to gain insight into topics important for their daily lives, such simplification also comprises the risk of making readers underestimate their own dependence on experts for decision advice. The question arises whether and how it is possible to inform laypeople about scientific topics without tempting them to overconfidently rely on their own derived decisions, i.e. whether there are further information characteristics that can prevent the observed comprehensibility effect.

**Information Controversiality** A textual feature that might counteract the influence of comprehensibility on laypeople's reliance on their own decisions is information controversiality. Due to the evolving and discursive nature of scientific knowledge production, views on a particular phenomenon held by different scientists are frequently conflicting. Confronting lay recipients with the controversiality of a particular topic might alert them that judging related claims is generally a highly demanding task, independently from whether related text information is easy or difficult to comprehend. The notion that encountering controversies might decrease recipients' persuasion is in accordance with previous findings showing readers to be more hesitant to agree with consistently supported than with controversial claims (Kienhues, Stadtler & Bromme, 2011; Yaniv, Choshen-Hillel & Milyavsky, 2009).

However, as of yet it has not been investigated whether the impact of information controversiality on recipients' persuasion interacts with the influence of information comprehensibility and thus can prevent the persuasive effect of high comprehensibility. Furthermore, it is still unclear which cognitive processes underlie the impact of comprehensibility and controversiality on recipients' reliance on their own decisions.

### **Possible Mechanism underlying the Influence of Comprehensibility and Controversiality**

On a theoretical level it is conceivable that the effects of both comprehensibility and controversiality on laypeople's decision readiness are mediated by recipient's assessment of the epistemic complexity of the topic. If laypeople encounter simplified texts tailored for them to comprehend, their understanding of the *text* may mislead them to consider the *subject matter* itself as less complicated than it really is (cf. Goldman & Bisanz, 2002; Scharrer et al., 2012). Consequently, they may overestimate their actual insight

into the topic and their ability to appropriately evaluate the provided information (cf. Keehner & Fischer, 2011). Likewise, laypeople may explain information controversiality with high epistemic complexity of the topic. As a result of encountering a conflict, laypeople might become attentive to the possibility that scientists can have opposing views on the same phenomena, and hence that a given claim might be valid only under specific circumstances. Laypeople may then become aware that reading the provided information does not equip them with sufficient knowledge to confidently decide about related claims.

However, the described mediating role of perceived topic complexity would require that laypeople elaborate on the epistemic demands of the problem at hand relative to their own epistemic capabilities to determine whether or not they can decide. Conversely, it is also possible that lay recipients do not engage in such elaborate reflections but rather base their judgment on intuitive or affective reactions. In light of such alternative possibilities it remains an open empirical question whether the influence of comprehensibility and controversiality on laypeople's decision behavior are indeed mediated by perceived epistemic topic complexity.

### **The Present Study**

The present study set out to pursue two goals. Firstly, we aimed to assess whether laypeople's increased reliance on their own decision after reading comprehensible compared to incomprehensible information is reduced if this information contains conflicting rather than consistent positions. Secondly, given the above considerations about the possible mediating role of perceived topic complexity, we aimed to gain insight into the processes by which comprehensibility and controversiality exert their influence on laypeople's decision behavior. For this purpose, we presented lay readers with argumentative text information that varied in comprehensibility and controversiality.

With regards to the first goal, we expected that comprehensible information would lead laypeople to rely more on their own decision than incomprehensible information if information is uncontroversial. However, we assumed that this comprehensibility effect is mitigated or even prevented if the received information is controversial. Specifically, we hypothesized that laypeople agree more strongly with a claim when reading comprehensible than incomprehensible information but that this difference is greater in case of uncontroversial than controversial information (H1). Furthermore, we expected laypeople's decision confidence to be analogously influenced by comprehensibility and controversiality. Decision confidence should be reflected in laypeople's respective preferences of three decision strategies: Strategy A to decide by oneself based on one's knowledge after reading the information should indicate high decision confidence, Strategy B to decide by oneself but only after obtaining further content information should indicate intermediate confidence, and Strategy C to leave the decision to an expert should indicate

low decision confidence. Hence, we hypothesized that laypeople find Strategy A more preferable after reading comprehensible than incomprehensible texts, but that this effect should be more pronounced in case of uncontroversial than controversial information (H2a). Similarly, Strategy B should be more popular after reading comprehensible than incomprehensible texts, but this difference should be greater in case of uncontroversial than controversial information (H2b). Finally, incomprehensible texts should lead to a greater preference of Strategy C than comprehensible texts but this effect should be stronger if the information is uncontroversial than controversial (H2c).

As to the second goal, we considered it possible that laypeople's impression of epistemic topic complexity mediates the influence of comprehensibility and controversiality on laypeople's persuasion strength and confidence. Comprehensible and uncontroversial information might facilitate the impression of the subject matter being simple and straightforward, i.e. decrease perceptions of epistemic complexity. As a result, laypeople may regard themselves able to appropriately understand the topic and may thus be inclined to rely on their own claim decisions. However, we would assume no mediation effect of topic complexity if laypeople's determination of their own decision readiness does not depend on elaborations of epistemic demands. Since both possibilities are conceivable, we formulated the following exploratory research question: Is the influence of comprehensibility and controversiality on claim agreement and agreement confidence mediated by perceived epistemic topic complexity?

## Method

The experiment was conducted using a 2x2 mixed design with the within-participant factor information comprehensibility (comprehensible vs. incomprehensible) and the between-participant factor controversiality (controversial vs. uncontroversial). In each condition, participants were presented with texts about a medical issue that were either comprehensible or incomprehensible and controversial or uncontroversial, respectively. Eighty-eight undergraduates of various majors at a German university participated in the experiment (54 female; mean age = 22.81 years,  $SD = 4.39$ ). To ensure participants' lay status, students of medicine, biology or related subjects were excluded from participation. Moreover, the medical issues addressed in the texts were fictitious in nature to ensure that participants were unable to make informed decisions about the contents based on their everyday knowledge and had no strong prior attitudes about the topics.

## Materials

Participants read a document (Document 1) containing a particular medical knowledge claim (e.g. "Bouchard arthrosis is caused by a deficiency of purinerase") followed by an explanation that described the mechanisms underlying the proposed claim and supporting empirical evidence. In addition, a second document (Document 2) was presented

consisting of a claim supported by empirical evidence. Depending on the controversiality condition, Document 2 either contained a claim that was in conflict with Document 1 (e.g. "A lack of purinerase is not among the causes of Bouchard arthrosis") or a claim that did not render the Document 1 information controversial (e.g. "A persistent lack of folic acid is not among the causes of Bouchard arthrosis"). The information contained in the documents was furthermore comprehensible or incomprehensible. In the incomprehensible conditions, both texts contained a large number of unexplained technical terms, whereas in the comprehensible conditions, technical jargon was translated into words that should be understandable for laypeople (e.g. "articulations" was translated to "joints"). All documents were comparable in length ( $M = 142.67$  words,  $SD = 44.66$ ).

Before reading a document pair, participants were confronted with a framing scenario in which a fictitious friend was described as having a medical problem and, due to their insecurity about the correctness of a particular problem-related claim, asked the participant for advice. The claim in question was the same that was later on stated and supported in Document 1. Participants were then presented with both text documents which they had allegedly found during an Internet search and which were described as being authored by different sources.

## Dependent Measures<sup>1</sup>

**Manipulation Check** In order to verify that comprehensibility had been manipulated as intended, participants rated each document for perceived comprehensibility on a Likert-scale from 1 ("very incomprehensible") to 7 ("very comprehensible"). Before providing their ratings, participants were given a definition of what we meant by comprehensibility. This definition described information as comprehensible when readers perceive the contents as clear and feel able to discriminate essential from less important parts and to evaluate information consistency.

**Claim Agreement** We assessed the extent to which participants were persuaded by measuring their agreement with the Document 1 claim after reading both documents. For this purpose, participants indicated their agreement on a 1 ("I don't agree at all") to 7 ("I totally agree") Likert-scale.

**Confidence in the Agreement Decision** Participants' confidence in their ability to decide about the claim was indicated by their agreement with three statements. Each statement reflects a strategy which individuals might use to come to a claim decision. Participants provided their agreement with these strategies on three separate 1 to 7 Likert-scales (1: "don't agree", 7: "strongly agree"):

<sup>1</sup> In addition to the above variables, further measures were collected. However, due to space constraints we only report the analysis of the presently listed measures.

(A) Trust in own agreement decision based on present knowledge: Following this strategy means participants felt ready to decide based on their knowledge after reading the documents. Preference for this strategy was measured by strength of agreement with the statement: “Based on my present knowledge about the topic, I am confident to decide whether it is correct that [Document 1 claim inserted]”.

(B) Trust in own decision based on further information: This strategy indicates that participants felt principally able to decide, but only after obtaining further topic information. Preference for this strategy was measured through agreement with the statement: “I want to obtain further information about [topic] which I then use to decide myself whether it is correct that [Document 1 claim inserted]”.

(C) Desire to consult an expert: Following this strategy means that participants wished to leave the decision to an expert. Preference for this strategy was measured by agreement with the statement: “I want to obtain information about experts in the field in order to identify a particular competent and credible expert. I would then consult this expert and rely on their judgment as to whether it is correct that [Document 1 claim inserted]”.

**Perceived Epistemic Complexity** Participants’ perception of epistemic topic complexity was assessed with six adjective pairs in a 7-point scale semantic differential format (very uncomplex—very complex, very multifaceted—very single-faceted, very unscientific—very scientific, very easy—very difficult, very uncomplicated—very complicated, very difficult to comprehend—very easy to comprehend). Two additional distracter pairs were presented to decrease transparency of the measurement intent (very unimportant—very important, very boring—very entertaining). Exploratory factor analyses (ML-extraction, oblimin rotation) showed the target items to load on one common factor in both comprehensibility conditions (comprehensible:  $KMO = .88$ ;  $\chi^2(15) = 350.71$ ,  $p < .001$ ; 70.43% explained variance; incomprehensible:  $KMO = .86$ ;  $\chi^2(15) = 230.52$ ,  $p < .001$ ; 58.88% explained variance). To determine a score of perceived topic complexity, the arithmetic mean of the target items was calculated. Internal consistency of the items was satisfactory in both comprehensibility conditions (comprehensible: Cronbach’s  $\alpha = .91$ , incomprehensible: Cronbach’s  $\alpha = .86$ ).

## Procedure

Participants received a paper booklet containing the text materials and scales for collecting the dependent measures. The booklet first presented a scenario describing the fictitious friend’s problem. Participants then read the document pair and provided their measures of claim agreement and confidence in their agreement decision. This was repeated four times, so that each participant read four document pairs in total, two in each condition. Afterwards, readers were asked to provide their ratings of comprehensibility for each document they had read and to evaluate the complexity of each addressed topic. Finally,

participants completed a demographic questionnaire before being debriefed about the fictitious nature of the text contents.

## Results

The means and standard deviations of the collected measures per experimental condition are shown in Table 1.

### Manipulation Check

A mixed ANOVA on comprehensibility ratings with document (Document 1 vs. Document 2) and comprehensibility (comprehensible vs. incomprehensible) as within-participant factors and controversiality (controversial vs. uncontroversial) as between-participant factor verified that texts designed as comprehensible were perceived as more comprehensible than texts designed as incomprehensible,  $F(1,85) = 648.11$ ,  $p < .001$ , part.  $\eta^2 = .88$ . Moreover, Document 2 was overall rated more comprehensible than Document 1  $F(1,85) = 22.54$ ,  $p < .001$ , part.  $\eta^2 = .21$ . The other main and interaction effects did not reach significance, all  $F(1,85) < 2.91$ , *ns*.

### Claim Agreement

Claim agreement scores were analyzed using a mixed ANOVA with the within-participant factor comprehensibility and the between-participant factor controversiality. Results showed that laypeople agreed more strongly with the Document 1 claim when the supporting information was comprehensible than when it was incomprehensible,  $F(1,86) = 9.34$ ,  $p < .05$ , part.  $\eta^2 = .10$ . Moreover, agreement was higher in the uncontroversial than the controversial condition, although this difference was only marginally significant,  $F(1,86) = 3.41$ ,  $p = .07$ , part.  $\eta^2 = .04$ . In line with H1 a significant interaction of comprehensibility and controversiality ( $F(1,86) = 4.89$ ,  $p < .05$ , part.  $\eta^2 = .05$ ) indicated that only in the uncontroversial condition did participants agree more strongly with claims from comprehensible than incomprehensible texts,  $t(43) = 3.52$ ,  $p < .01$ . In contrast, there was no difference in agreement between comprehensible and incomprehensible texts in the controversial condition,  $t(43) = .64$ , *ns*.

### Confidence in the Agreement Decision

To test H2a-c, we conducted separate mixed ANOVAs for each decision strategy using comprehensibility as within- and controversiality as between-participant factor.

**(1) Trust in own agreement decision based on present knowledge** This strategy was more popular in the comprehensible than in the incomprehensible conditions,  $F(1,86) = 51.08$ ,  $p < .001$ , part.  $\eta^2 = .37$ , as well as in the uncontroversial compared to the controversial conditions,  $F(1,86) = 23.74$ ,  $p < .001$ , part.  $\eta^2 = .22$ . Furthermore, and in line with H2a, there was a significant comprehensibility\*controversiality interaction,  $F(1,86) = 6.43$ ,  $p < .05$ , part.  $\eta^2 = .07$ . This was due to the difference

Table 1: Means and standard deviations (in brackets) of the collected measures as a function of comprehensibility and controversiality.

Condition	Comprehen- sibility		Claim- agreement	Decision based on present knowledge	Decision based on further info.	Decision through expert advice	Epistemic complexity
	Doc.1	Doc.2					
Compr./ uncontr.	5.80 (1.13)	5.93 (1.08)	4.53 (1.05)	3.30 (1.66)	5.16 (1.57)	5.16 (1.87)	4.38 (1.03)
Incompr./ uncontr.	2.05 (1.17)	2.38 (1.34)	3.90 (1.04)	2.10 (1.31)	5.25 (1.64)	5.41 (1.65)	5.46 (.86)
Compr./ contr.	5.57 (1.00)	5.97 (1.05)	3.98 (.94)	1.90 (.94)	4.99 (1.70)	5.47 (1.80)	4.22 (.92)
Incompr./ contr.	1.78 (.80)	2.37 (1.30)	3.88 (.621)	1.33 (.61)	4.82 (1.91)	5.73 (1.73)	5.47 (.78)

between both comprehensibility conditions being larger when the documents were uncontroversial than when they were controversial.

### (2) Trust in own decision based on further information

With regards to the popularity of this strategy results revealed no significant main or interaction effects (all  $F(1,86) < 1.26$ , *ns*), hence providing no support for H2b.

**(3) Desire to consult an expert** This strategy was shown to be more popular in the incomprehensible than in the comprehensible conditions,  $F(1,86) = 5.24$ ,  $p < .05$ , part.  $\eta^2 = .06$ . However, contrary to H2c, controversiality had no influence on participants' willingness to ask an expert, all further  $F(1,86) < .97$ , *ns*.

### Perceived Epistemic Complexity

To answer the research question whether the influence of comprehensibility and controversiality on claim agreement and agreement confidence was mediated by perceived epistemic topic complexity, we conducted separate mediator analyses for each combination of independent variable, mediator variable and dependent variable following the approach suggested by Baron and Kenny (1986) and Judd, Kenny, and McClelland (2001). Results showed that of the three preconditions necessary for a variable to act as a mediator (1. the independent variable affects the assumed mediator 2. the independent variable affects the dependent variable, 3. the mediator affects the dependent variable) the first two were fulfilled for most combinations, but the third precondition was fulfilled in no case (all  $F(2,41) < 3.23$ , *ns*). Hence, complexity perceptions do not appear to mediate the observed comprehensibility and controversiality effects.

## Discussion

By presenting medical laypeople with texts varying in comprehensibility and controversiality, the present experiment assessed whether the facilitating influence of high comprehensibility on recipients' reliance on their own decision about scientific claims is mitigated by information

controversiality. Moreover, the study was aimed to gain insight into the possible process through which both text features affect laypeople's decision behavior by examining the role of perceived epistemic topic complexity as a potential mediator.

The results revealed that in line with our expectations and previous research (Eagly, 1974; Scharrer et al., 2012), lay recipients agreed more strongly with claims from comprehensible than incomprehensible texts; however, this difference occurred only when the information was uncontroversial. In case of controversial information, laypeople's claim agreement was not affected by text comprehensibility. It seems that encountering controversial information makes laypeople more cautious to agree with a claim even if the information is easy to comprehend.

Moreover, both comprehensibility and controversiality affected laypeople's decision confidence; however this impact was manifested differently on the three decision strategies. Similar to the state of affairs regarding claim agreement, the strategy to decide based only on one's present knowledge was more popular after reading comprehensible than incomprehensible texts, but this influence of comprehensibility was diminished when information was controversial. In contrast, none of the text features had an effect on the strategy to decide based on further information and only comprehensibility had an impact on the strategy to ask an expert, with high comprehensibility decreasing the desire for expert advice. We assume that if information is incomprehensible, laypeople are generally willing to consult an expert regardless of controversiality. In case of comprehensible controversial texts, laypeople feel possibly more encouraged to determine which of both conflicting position is correct due to their comprehension success, for instance by seeking out further information. This might explain why their willingness to ask an expert in this condition is not higher than in case of comprehensible uncontroversial texts.

It is noteworthy that even when participants received comprehensible and uncontroversial texts, ratings of their willingness to ask an expert did not average below 5 on a scale from 1 to 7 (7 indicating a strong willingness).



However, the observed influence of comprehensibility on desire for expert advice suggests that a too strong simplification of scientific contents may mislead lay recipients to underestimate their dependence on experts.

As to our second goal, to get an insight into the possible process through which comprehensibility and controversiality affect laypeople's reliance on their own decisions, we found that the influence of neither information feature is mediated by perceived topic complexity. It seems that laypeople do not base their judgment of whether or not to rely on their own decisions on reflections about epistemic complexity. Perhaps the experience of fluently comprehending information simply triggers positive affective reactions, which then translate to more favorable evaluations in general (Schwarz, 2004). However, this can so far only be assumed, and the exact mechanisms that underlie the influences of comprehensibility and controversiality remain subject to further empirical clarification.

In sum, the present findings largely confirm our assumption of a combined influence of comprehensibility and controversiality on laypeople's reliance on their own decisions. The results indicate that the facilitating influence of comprehensibility on claim agreement found in previous research is more pronounced if the information is uncontroversial but seems to be reduced or even prevented in case of controversial information. As to laypeople's confidence in their claim decision, the results are less conclusive. While it appears that controversiality has a moderating effect on the influence of comprehensibility regarding laypeople's trust in their own decision based on current knowledge, this influence does not translate to their preference of the decision strategy to consult an expert. Finally, we found that perceived epistemic topic complexity does not appear to act as a mediating factor of either the observed comprehensibility or controversiality influence.

The results have practical implications by informing about how scientific information can be optimally communicated to the lay public. Although simplified science reports have the advantage to provide lay recipients insight into scientific issues, it is possible that based on such easy-to-comprehend information laypeople become overly ready to make own decisions, in spite of their lack of background knowledge and training. The present results suggest that when scientific findings are communicated to laypeople, the inclusion of information about related controversies can serve to weaken the comprehensibility effect. As a result, laypeople are more likely to be prevented from readily relying on their own claim evaluations and might rather turn to a pertinent expert for support.

### Acknowledgments

We are grateful to Jasmin Hettinger, Stephanie Sievers and Nikolai Wystrychowski for their help with data collection. This research was supported by the Deutsche Forschungsgemeinschaft (DFG), grant BR 1126/6-1.

### References

- Baron, R. & Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.
- Bromme, R., Kienhues, D. & Porsch, T. (2010). Who knows what and who can we believe? Epistemological beliefs are beliefs about knowledge (mostly) attained from others. In L. Bendixen & F. Feucht (Eds.), *Personal epistemology in the classroom: Theory, research, and implications for practice*, Cambridge: University Press.
- Eagly, A.H. (1974). Comprehensibility of persuasive arguments as a determinant of opinion change. *Journal of Personality and Social Psychology*, 29, 758-773.
- Goldman, S.R., & Bisanz, G.L. (2002). Toward functional analysis of scientific genres: Implications for understanding and learning processes. In J. Otero, J.A. Leon, & A.C. Graesser (Eds.), *The psychology of science text comprehension*, Mahwah NJ: Erlbaum.
- Judd, C., Kenny, D. & McClelland, G. (2001). Estimating and testing mediation and moderation in within-participant designs. *Psychological Methods*, 6, 115-134.
- Keehner, M. & Fischer, M.H. (2011). Naive realism in public perceptions of neuroimages. *Nature Reviews Neuroscience*, 12, 118.
- Keil, F. (2008). Getting to the truth: Grounding incomplete knowledge. *Brooklyn Law Review*, 73, 1035-1052.
- Kienhues, D., Stadtler, M. & Bromme, R. (2011). Dealing with conflicting or consistent medical information on the Web: When expert information breeds laypersons' doubts about experts. *Learning and Instruction*, 21, 193-204.
- Mason, L., Ariasi, N. & Boldrin, A. (2011). Epistemic beliefs in action: Spontaneous reflections about knowledge and knowing during online information searching and their influence on learning. *Learning and Instruction*, 21, 137-151.
- Scharrer, L., Bromme, R., Britt, M.A. & Stadtler, M. (2012). The seduction of easiness: How science depictions influence laypeople's reliance on their own evaluation of scientific information. *Learning and Instruction*. doi:10.1016/j.learninstruc.2011.11.004
- Schwarz, N. (2004). Meta-cognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology*, 14, 332-348.
- Wagner, W., Elejabarrieta, F. & Lahnsteiner, I. (1995). How the sperm dominates the ovum – Objectification by metaphor in the social representation of conception. *European Journal of Social Psychology*, 25, 671-688.
- Yaniv, I., Choshen-Hillel, S. & Milyavsky, M. (2009). Spurious consensus and opinion revision: Why might people be more confident in their less accurate judgments? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 558-563.

# Subjective Confidence of Acoustic and Phonemic Representations During Speech Perception

Jordan Schoenherr (psychophysics.lab@gmail.com)

John Logan (john\_logan@carleton.ca)

Amy Winchester (awinches@connect.carleton.ca)

Department of Psychology, Carleton University  
1125 Colonel By Drive, Ottawa, ON K1S5B6 Canada

## Abstract

Although many speech perception studies have suggested that long-term memory representations of phonemes induce categorical perception along an acoustic continuum (e.g., voice-onset time; VOT) when identifying speech sounds, other studies have suggested that acoustic information is preserved and that graded responses can be observed in within-category comparisons. Using subjective confidence reports, we present findings that support the use of both acoustic and phonemic cues during speech perception. Replicating earlier findings, we observed evidence for two well-defined phoneme categories along the voice-onset time continuum. Additionally, we also observed overconfidence in responses suggesting that the explicit representation of phonemes differs from the representations used to make identification and discrimination responses. Taken together with results from other studies, our findings support the claim that listeners can access both phonemic and acoustic representations, with explicit knowledge of the former but not the latter.

**Keywords:** speech perception, category boundaries, confidence processing

## Introduction

Listeners presented stimuli from a continuum of sounds varying in an acoustic cue such as voice-onset time respond as though they perceive sharp discontinuities or category boundaries when tested in identification and discrimination tasks. On their own, however, measures of identification (ID) and discrimination (e.g., AX) do not indicate the extent to which participants are aware of these categories or their boundaries. In order to examine subjective awareness of categories and category boundaries, we used post-decisional confidence reports after the ID and discrimination responses to examine how response certainty varies across the continuum.

## Categorical Perception of Speech Sounds

Results from speech perception studies were originally interpreted as indicating speech sounds are perceived as members of discrete phonemic categories (e.g., Liberman, Harris, Hoffman, Griffith, 1957). In such studies an acoustic cue such as VOT is varied to generate a continuum of stimuli (e.g., Pisoni, 1973; for a review, see Raphael, 2005). Participants then identify items using two categories such as /ba/ and /pa/. Importantly, the resulting ID functions show a well-defined category boundary that partitions the continuum into two phonemic categories. Moreover, in the accompanying discrimination task, participants exhibit greater accuracy when discriminating between speech

sounds from two phonemic categories relative to those from the same phonemic category even when the absolute acoustic differences are equivalent. Collectively, these ID and discrimination results constitute the phenomenon of categorical perception.

Categorical perception is a robust phenomenon. However listeners also retain some within-category acoustic information (Pisoni & Tash, 1974) and can show graded responses within a phonemic category (e.g., McMurray, Tanenhaus, Aslin, & Spivey, 2003; Miller, J. L., & Volaitis, 1989). Using an AX task, Pisoni (1973) found that when two speech sounds were presented with a long inter-stimulus interval (ISI), participants perceived could not discriminate between stimuli. When stimuli were presented with shorter ISIs, however, participants' accuracy at making within-category comparisons improved (cf. Werker & Logan, 1985). Pisoni (1973) used these findings to suggest a two-stage model for speech perception with an initial stage that processes acoustic information and a second stage that retrieves categorical cues from long-term memory (see also Pisoni & Tash, 1974). Further support for multiple representations also comes from studies of so-called sine-wave speech wherein a participant will identify the stimuli as either noise or speech depending on their prior expectations (e.g., Remez, Rubin, Pisoni, & Carrell, 1981; Davis & Johnsruide, 2007; for other top-down effects, see also Davis, Johnsruide, Hervais-Adelman, Taylor, & McGettigan, 2005).

One possible account of these findings is that participants only have explicit access to phonemic information and are insensitive to acoustic properties of the stimuli that could be used to parse the continuum in an alternative manner. If an effective measure of explicit knowledge about category structure can be obtained that can be contrasted with performance on perceptual tasks, we should be capable of determining the extent to which listeners are aware of stimulus properties beyond explicit phonemic categories.

## Subjective Awareness and Confidence Reports

Whether participants can maintain a representation and yet have little or no awareness of it is a controversial issue. For instance, in a typical experiment assessing such awareness, participants perform a task and indicate how certain they are in their response on a subjective probability scale (e.g., with 50% representing a guess and 100% representing complete certainty). In these tasks participants'

confidence reports typically deviate from their actual performance, i.e., they are miscalibrated. *Calibration* assesses the difference between the subjective probability of an event (confidence level) and the observed probability of a correct response (proportion correct; for formulae, see Baranski & Petrusic, 1994). In this way, calibration represents the extent to which participants are aware of their primary decision on a trial-to-trial basis. When assessing average confidence and accuracy in a given condition, participants are typically either *overconfident* [confidence > p(cor)] or *underconfident* [confidence < p(cor)]. These systematic deviations have been argued to be evidence for both implicit and explicit representations of knowledge (e.g., Dienes & Berry, 1997). For instance, if overconfidence is observed, this suggests an explicit representation that is less accurate than the implicit representation used to classify stimuli. Confirming this, recent studies do find overconfidence bias in perceptual tasks and no overconfidence bias in conceptual tasks thought to involve information stored in long-term memory (e.g., Kvidera, & Koustaal, 2008). Such effects suggest two representations, with moderate calibration suggesting at least some explicit awareness but with overconfidence suggesting a well-defined explicit representation that does not reflect the actual representation used to discriminate and classify stimuli. If a multiple-representation account of speech perception is correct (e.g., Pisoni, 1973) then a confidence report methodology might be capable of assessing these different representations.

An important concern related to the existence of multiple sources of information is how confidence reports are generated. Traditional approaches to confidence processing have been agnostic about the nature of the representations used to make the primary decision and report confidence. Early models of confidence assumed a decisional-locus (e.g., Ferrel & McGooney, 1980; Gigerenzer, Hoffrage, & Kleinbolting, 1991; Pleskac & Busemeyer, 2010) wherein confidence reports are based solely on information used by the primary decision process thereby requiring no additional processing, a post-decisional locus wherein confidence is computed following the primary decision (e.g., Audley, 1960; Vickers & Packer, 1980), or an alterable locus wherein confidence processing can occur during or after the primary decision depending on speed or accuracy stress (Baranski & Petrusic, 1998). For instance, in a study conducted by Baranski and Petrusic (2001) participants were given blocks of trials wherein they were required to simply make a decision or make a decision followed by a post-decisional confidence report. They found that response latencies for the primary decision were significantly longer when confidence was required relative to a no confidence condition indicating an additional set of operations was required to compute confidence. More recently, Schoenherr, Leth-Steensen, and Petrusic (2010) found that information that is nondiagnostic of the primary decision can create variations in confidence reports independently of accuracy. Applied to phonemic

categorization, if acoustic information is available from a perceptual process and phonemic representations are available from the activation of long-term memory representations, then both sources of information should influence confidence reports. Substantial differences in the patterns observed between accuracy and confidence would suggest the existence of acoustic and phonemic representations.

### Present Study

In order to assess whether participants have explicit knowledge of acoustic information and phonemic category boundaries, the present study compares confidence reports to performance in ID and discrimination (AX) tasks. In the ID task, awareness of acoustic cues would be evidenced if certainty increases as a function of the distance from the category boundary. This would suggest that the ID function is merely an artifact of the requirement that participants use only two labels to categorize stimuli when they have in fact encoded and stored (temporarily) acoustic information. If, on the other hand, there is no systematic deviation of confidence and ID performance, then it seems reasonable to conclude that participants are only using phonemic information to identify stimuli. Alone, however, ID performance might not be capable of differentiating.

In order to determine whether participants have access to both acoustic and phonemic representations, we replicated Pisoni and Tash's (1974) paradigm wherein response times in the AX task were used to suggest different levels of processing. In addition, the present study also used postdecisional confidence reports. Again, deviations between performance and confidence reports would suggest that two representations are used to classify and discriminate speech sounds. If participants only perceive speech sounds as exemplars of discrete phonemic categories, then they should be reasonably well-calibrated on a trial-to-trial basis and exhibit little or no over-/underconfidence bias. If participants exhibit poor calibration and overconfidence in the AX task, then this suggests that despite the availability of acoustic properties within the implicit system the explicit representation is phonemic. More specifically, such a finding suggests that participants have an explicit representation of the phonemic category but also maintain graded acoustic information from stimuli along a continuum. Given the intuitive saliency of the phoneme, we assume that category boundaries are an explicit representation but that some acoustic information must remain available (e.g., Pisoni, 1973).

### Experiment

The goals of this experiment were threefold. First, we sought to validate the use of subjective confidence reports in a speech perception task by comparing a confidence and no confidence condition. Second, using subjective calibration measures, we examined whether participants had explicit awareness of the phonemic category boundary. Third, we examined whether this awareness was task-dependent by using both ID and AX tasks.

## Participants

Listeners were 15 Carleton University students who received course credit for their participation. All participants reported normal hearing and no speech pathologies.

## Materials

Using the paradigm developed by Pisoni and Tash (1974) participants were presented with /b/ and /p/ stimuli that varied along the VOT continuum. Seven speech stimuli corresponding to 0 to 60 ms VOT, originally synthesized by Lisker and Abramson (1967), were obtained from the Haskins Laboratories website (HL, 2011). The sounds were originally recorded on reel-to-reel tape and later converted into AIFF format. Stimuli were pre-processed using a DC offset correction to eliminate high frequency noises present in the AIFF versions and converted into WAV files. These stimuli were used in both the ID and AX tasks.

## Procedure

Trials in the ID task had one or two components depending upon block. In both blocks of trials participants reported whether the stimulus was a /b/ or /p/ using the 'V' or 'N' key, respectively. For one block participants also rated the confidence they had in their ID responses using a 6-point scale, with 50% representing a guess and 100% representing certainty. Participants completed a total of 180 trials in each block of the ID task.

Trials in the AX task also had one or two components depending on block. In both blocks of trials participants decided whether two stimuli separated by a 250 ms ISI were the same or different, using the 'D' and 'K' key, respectively. Replicating Pisoni and Tash (1974), stimulus pairs differed in either 0-, 2-, 4- or 6-steps and were either selected from the same phonemic category or different phonemic categories. Three replications of the eight within-category comparisons and four replications of the six between-category comparisons were presented in a block of 48 trials. For one block of AX trials participants also provided a confidence rating of their AX decision using the same scale described above.

Half of the participants performed the ID task first whereas the other half performed the AX task first. Half of the blocks of trials required participants to provide confidence reports whereas the other half only required participants to complete the ID and AX tasks alone. Presentation of confidence and no confidence blocks was counterbalanced as were the responses keys for the AX task. The experiment required approximately 30 minutes to complete. Stimuli were presented via headphones using PsychoPy software (Peirce, 2007).

## Results

In the following analyses we use two sets of assumptions. Following Pisoni and Tash (1974), we assume that stimuli 1-3 are assigned to the /ba/ category whereas stimuli 4-7 are assigned to the /pa/ category. From this we derive measures of accuracy. In the AX task we additionally assume that accuracy is determined by acoustic properties when making paired comparisons.

Calibration was computed by obtaining the average differences between proportion correct for each confidence category with calibration scores range from 0.0 (perfect calibration) and 1.0 (perfect miscalibration). Notably, calibration scores above 0.10 are rare. Under-/Overconfidence was computed by taking the difference between mean confidence and mean accuracy for each condition, with positive values representing overconfidence and negative values representing underconfidence (for further details see, Baranski & Petrusic, 1994).

For all ID and AX analyses, repeated measures ANOVAs were conducted using dependent variables for the primary decision and confidence reports. Greenhouse Geisser adjusted values are reported with unadjusted degrees of freedom. Bonferroni pairwise comparisons were also performed as a post-hoc test.

## Identification Task

**Proportion Identification.** Figure 1 shows the mean response frequency for each category label in the ID task in the confidence report condition. Participants clearly identified two discrete categories for /ba/ and /pa/, respectively, with a category boundary situated between +20 and +30 ms VOT. This pattern replicates the findings obtained by Pisoni and Tash (1974) as well as other studies (e.g., Experiment 1 in McMurray et al., 2003).

The proportion of correct ID responses was analyzed for each VOT stimulus and whether a confidence report was provided or not. The only significant finding observed was the location of the stimuli along the VOT continuum,  $F(6,84) = 6.394$ ,  $MSE = .019$ ,  $p = .02$ ,  $\eta^2 = .314$ . The absence of a main effect or interaction of confidence reports is important as it suggests that the addition of confidence reports did not significantly affect ID performance thereby permitting a straightforward interpretation of the remaining results.

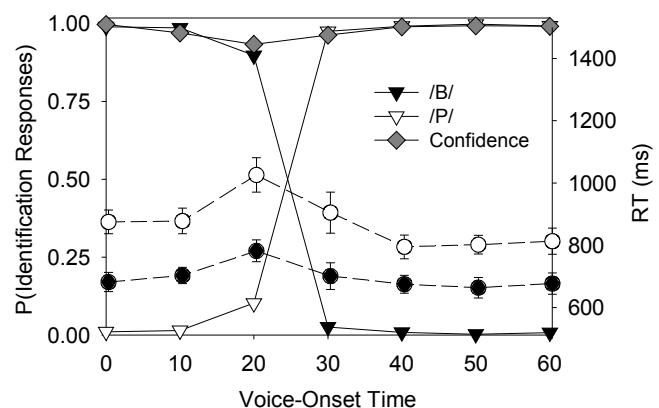


Figure 1. Mean identification functions, response times for confidence (unfilled circles) and no confidence (filled circles) conditions and mean confidence across VOT continuum. Identification function uses performance in confidence condition to allow comparison with mean confidence.

**Mean Confidence.** Figure 1 also demonstrates the effect of confidence measures. Like ID accuracy, we found that subjective confidence varied along the VOT continuum,  $F(1,14) = 6.55$ ,  $MSE = 44.11$ ,  $p = .008$ ,  $\eta^2 = .319$ . Pairwise comparisons revealed that this effect arose from the difference in confidence between stimuli located at 20 and 30 ms VOT ( $p = .035$ ), which corresponds to the stimuli adjacent to the category boundary.

**Calibration Indices.** An analysis of subjective calibration revealed only a marginally significant difference across the VOT continuum,  $F(6,84) = 3.401$ ,  $MSE = .013$ ,  $p = .085$ ,  $\eta^2 = .195$ . This suggests that the greatest difference between subjective awareness and performance occurs for the 20 ms VOT stimulus. Our comparison of over/underconfidence bias did not reveal any significant effects,  $F(6,84) = 1.948$ ,  $MSE = .035$ ,  $p = .183$ ,  $\eta^2 = .122$ . Together, these findings suggest that participants are only explicitly aware of the phonemic representation.

Table 1. Mean dependent measures for “Same” responses to Within-Category Paired Comparisons

	/ba/		/pa/	
	A-A	A-a	A-A	A-a
Mean RT	756 ms	818 ms	693 ms	737 ms
% Error	4.4	11.1	1.1	1.1
Conf	97.44	98.67	99.67	99.56

**Identification Response Time.** Analysis of our RT data also replicated Pisoni and Tash’s (1974) findings (see Figure 1). Specifically, we observed significant changes in RT across the VOT continuum,  $F(6,84) = 8.323$ ,  $MSE = .030$ ,  $p < .001$ ,  $\eta^2 = .373$ . We found that responses to the stimulus located at 20 ms VOT were longer than those for stimuli at 0, 10, 50, and 60 ms VOT. Moreover, replicating Baranski and Petrusic (2001), we also observed a significant main effect of confidence condition,  $F(6,66) = 4.701$ ,  $MSE = .041$ ,  $p = .021$ ,  $\eta^2 = .572$ . Participants took longer to identify stimuli when confidence reports were required ( $M = 871$  ms) relative to the no confidence condition ( $M = 698$  ms).

#### Discrimination Task

**Proportion “Same” Responses.** AX responses were assessed based on a category criterion such that ‘same’ responses were considered correct for within category comparisons and incorrect for between category comparisons. Table 1 provides a point of comparison with Pisoni and Tash (1974), wherein AA and Aa represent, acoustically and phonemically identical stimuli, respectively. Table 2 provides mean dependent measures for within (AA, Aa) and between (AB through AB”) for 2-step through 6-step, respectively) category comparisons.

Using a category criterion, we observed an interaction between phoneme category (/ba/ v. /pa/) and the comparison type (within v. between),  $F(1,13) = 13.421$ ,  $MSE = .004$ ,  $p = .003$ ,  $\eta^2 = .508$ . Participants were far more accurate in making within-category comparisons from the /pa/ category relative to all others (see Table 2).

Table 2. Mean dependent measures for correct and incorrect responses for within- and between-category comparisons collapsed across confidence condition.

	/ba/		/pa/	
	AA/Aa	All AB	AA/Aa	All AB
Mean RT	753 ms	809 ms	693 ms	734 ms
% Error	7.6	5.6	1.1	7.2
% Conf	96.6	96.4	99.6	97.3

**Mean Confidence.** Our analysis of mean confidence revealed a pattern similar to that of the accuracy analysis. We found an a marginally significant interaction of phoneme category and comparison type,  $F(1,13) = 4.589$ ,  $MSE = 3.8393$ ,  $p = .052$ ,  $\eta^2 = .261$ . The only significant effect was for phonemic category,  $F(1,13) = 5.895$ ,  $MSE = 9.627$ ,  $p = .030$ ,  $\eta^2 = .312$ . Participants expressed more confidence when responding to stimuli from the /pa/ category ( $M = 98.456$ ) relative to those from the /ba/ category ( $M = 96.507$ ). Taken together with the results of the ID task, this suggests that the representation of the /pa/ category is more well-defined than the /ba/ category for these VOT stimuli.

**Calibration Indices.** As with the mean confidence analysis, the interaction of phoneme category and comparison type was only marginally significant when using category coding,  $F(1,13) = 3.613$ ,  $MSE = .0004$ ,  $p = .080$ ,  $\eta^2 = .217$ . However, when responses are scored with acoustic coding we find significant miscalibration,  $F(4,52) = 776.8$ ,  $MSE = .019$ ,  $p < .001$ ,  $\eta^2 = .984$ . Like the ID task, we did not observe any overconfidence bias in the AX task when category coding of accuracy was used, all  $F$ s  $< 2.5$ . Again this suggests that participants access a phonemic representation to make their confidence decision. Confirming this, significant overconfidence was observed when acoustic coding of accuracy was used to compute overconfidence,  $F(4,52) = 1709$ ,  $MSE = .007$ ,  $p < .001$ ,  $\eta^2 = .992$ . Post-hoc paired comparisons on these means (see Table 3) revealed that AA pairs differed from all other pairs ( $p < .001$ ) but no other differences were observed.

Table 3. Mean calibration and overconfidence for comparison type using acoustic coding for response accuracy.

	Comparison Type				
	AA	Aa	AB	AB’	AB”
OU	.013	.976	.950	.962	.994
CAL	.008	.959	.913	.936	.988

**Discrimination Response Time.** Following Pisoni and Tash (1974), response latencies across comparison pairs were also analyzed. Only correct response latencies were analyzed (i.e., “Same” responses for within-category comparisons and “Different” responses for between-category comparisons). Replicating their findings, the type of comparison affected response latencies,  $F(4,52) = 5.976$ ,  $MSE = .088$ ,  $p = .007$ ,  $\eta^2 = .315$ . Importantly, an interaction was also observed between the confidence condition and comparison type,  $F(1,13) = 9.072$ ,  $MSE = .029$ ,  $p = .01$ ,  $\eta^2$

= .315. As Figure 2 indicates, participants were fastest when responding to acoustically similar stimuli and slowest to compare stimuli between categories separated by small steps along the VOT continuum. The additional requirement of confidence increased response latencies for acoustically dissimilar pairs.

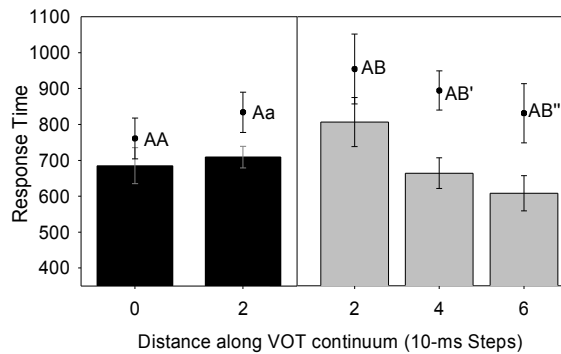


Figure 2. Discrimination response time for paired-comparisons within and between phonemic categories for no confidence (bars) and confidence (points) responses. Error bars represent standard error of the mean.

As in the ID task, we also observed a main effect of confidence condition,  $F(1,13) = 4.92$ ,  $MSE = .352$ ,  $p = .045$ ,  $\eta^2 = .275$ . More time was required to discriminate stimuli when confidence was required ( $M = 861$  ms) relative to the no confidence condition ( $M = 696$  ms)<sup>1</sup>.

### Discussion

Results generally replicated those observed by Pisoni and Tash (1974): participants perceived two distinct phonemic categories along the VOT continuum. In addition, we also found that stimuli could be discriminated with greater speed and accuracy when they were selected from two contrasting phoneme categories rather than within a category. However, although our results suggest participants represented stimuli as phonemic categories, the /ba/ category was not as well defined as the /pa/ category. This was evidenced in a shallower portion of slope between stimuli at 10 and 20 ms VOT in the ID function and the reduced accuracy and confidence when comparing acoustically dissimilar stimuli within that category.

Of equal importance, the experiment also replicated findings in the confidence processing literature: participants were faster when they made ID and AX decisions alone compared to when they were additionally required to report confidence post-decisionally (e.g., Baranski & Petrusic, 2001). Such a finding suggests that confidence processing requires a secondary set of operations to generate a confidence report. Importantly, however, the requirement of

a confidence report did not appear to adversely affect ID or AX performance.

Confidence performance complimented ID and AX performance: participants expressed the most uncertainty in the /ba/ category, and more specifically in the stimulus at the category boundary (20 ms VOT). An absence of overconfidence bias and excellent calibration in the category analysis of both ID and AX tasks suggests that participants' explicit knowledge of phoneme categories guides their classification. Supporting this interpretation, when we reanalyze discrimination accuracy in terms of acoustic properties we found that mean indices of overconfidence and miscalibration were at ceiling for all stimuli other than AA pairs, indicating an inability to discriminate acoustically dissimilar pairs within the same phonemic category.

### Conclusions

The present study replicated findings in both the speech perception and confidence processing literatures. Participants' responses indicated categorical perception of acoustically dissimilar stimuli along a voicing continuum and confidence reports required additional time to process. Moreover, confidence reports revealed that participants do not appear to be aware of acoustic information used to activate phoneme representations under the presentation conditions. Although the present study used only one ISI, follow-up studies will vary ISI in an AX task to further differentiate subjective awareness from performance. Moreover, phoneme categories that participants might have less familiarity with (e.g., Pisoni et al., 1982) should demonstrate larger differences in overconfidence and calibration. In short, our findings suggest that confidence reports can be used along with other measures (see also McMurray et al., 2003; Miller & Volaitis, 1989) to assess metalinguistic awareness in the context of speech perception.

Several caveats remain. First, VOT represents one among many physical cues that have been implicated in speech perception. In as much as the processing of speech in a natural environment requires multiple cues, the findings of the present study might be limited to this continuum. One possibility is that with a greater number of cues subjective certainty might increase. This concern about the limits of using synthesized speech has been a recurrent theme in speech perception research (Raphael, 2005). Second, space limits prevent inclusion of an analysis of individual differences, such as working memory capacity and individual ID functions. Finally, studies will need to assess whether these findings generalize to non-speech sounds that share similar properties such as the relative onset of two tones (Pisoni, 1977) or whether overconfidence is limited to only speech sounds. Despite these caveats, the results of the current work suggest that the application of a confidence report methodology holds considerable promise in clarifying the nature of the representations used for speech perception and how these representations are accessed. Calibration can be used to assess whether one or multiple acoustic cues are

<sup>1</sup> An analysis of all (correct and incorrect) responses also revealed the effect of comparison type,  $F(4,52) = 7.729$ ,  $MSE = .037$ ,  $p < .001$ ,  $\eta^2 = .373$ , and the requirement of confidence,  $F(1,13) = 5.07$ ,  $MSE = .247$ ,  $p = .042$ ,  $\eta^2 = .281$ . Post hoc paired comparisons revealed that AA differed from both Aa and AB (all  $ps < .035$ ).

used whereas under-/overconfidence suggests the extent to which phonemic and acoustic information is available to listeners.

## References

- Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgements. *Perception & Psychophysics*, 55, 412-428.
- Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 929-945.
- Baranski, J. V., & Petrusic, W. M. (2001). Testing the architectures of the decision-confidence relation. *Canadian Journal of Experimental Psychology*, 55, 195-206.
- Davis, M.H., Johnsruide, I.S., Hervais-Adelman, A., Taylor, K. & McGettigan, C.M. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, 134, 222-241.
- Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin & Review*, 4, 3-23.
- Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98, 506-528.
- Haskins Laboratories (2011). Abramson/Lisker VOT Stimuli. Retrieved 01/12/2011. From <http://www.haskins.yale.edu/featured/demo-liskabram/index.html/>.
- Kvidera, S., & Koustaal, W. (2008). Confidence and decision type under matched stimulus conditions: overconfidence in perceptual but not conceptual decisions. *Journal of Behavioral Decision Making*, 21, 253-281.
- Lieberman, A.M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, 358-368.
- Lisker, L., & Abramson, A. S. (1967). The voicing dimension: Some experiments in comparative phonetics. *Proceedings of the 6th International Congress of Phonetic Sciences*. Prague: Academia.
- McMurray, B., Tanenhaus, M. K., Aslin, R. N., & Spivey, M. J. (2003). Probabilistic constraint satisfaction at the lexical/phonetic interface: Evidence for gradient effects of within-category VOT on lexical access. *Journal of Psycholinguistic Research*, 32, 77-97.
- Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, 46, 505-512.
- Peirce, J. W. (2007) PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods*, 162, 8-13.
- Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, 13, 253-260.
- Pisoni, D. B. (1977) Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, 67, 1352-1361.
- Pisoni, D. B., & Tash, J. B. (1974) Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, 15, 285-290.
- Raphael, L. J. (2005). Acoustic cues to the perception of segmental phonemes. In D.B. Pisoni & R.E. Remez (Eds.), *The handbook of speech perception* (pp. 182-206). Oxford: Blackwell.
- Remez, R.E., Rubin, P.E., Pisoni, D.B., Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science*, 212, 947-949.
- Vickers, D., & Packer, J. S. (1982). Effects of alternating set for speed or accuracy on response time, accuracy, and confidence in a unidimensional discrimination task. *Acta Psychologica*, 50, 179-197.
- Werker, J. F., & Logan, J. S. (1985). Cross-language evidence for three-factors in speech perception. *Perception & Psychophysics*, 37, 35-44.



# Enough is enough: Inductive sufficiency guides learners' ratings of informant helpfulness

**Patrick Shafto** (p.shafto@louisville.edu)  
University of Louisville

**Hyowon Gweon** (hyora@mit.edu)  
Massachusetts Institute of Technology

**Chris Fargen** (cmfarg01@louisville.edu)  
University of Louisville

**Laura Schulz** (lschulz@mit.edu)  
Massachusetts Institute of Technology

## Abstract

Much of what we learn, we learn from others. What guides learners' choice of informants? Research suggests that learners resist informants who provide incorrect information or insufficient information for accurate inference. Here we propose that learners' choices of informants are rationally guided by the extent to which evidence supports accurate inference, rather than the sheer amount of evidence provided. Extending recent research formalizing pedagogical reasoning, we propose a computational model of efficient teaching. We present an experiment on adults testing three different hypotheses about learners' preferred level of the amount of data. The results suggest that learners care about the inductive sufficiency of evidence, rather than the amount of evidence provided. We conclude by discussing the implications of these findings for cognition and cognitive development.

**Keywords:** Trust; Pedagogical reasoning; Bayesian model.

People face a seemingly intractable problem in learning about the world. There is an endless amount of information to learn, but relatively limited time to acquire information. Fortunately, learners are surrounded by other agents who can help them learn. However, although some people may be valuable sources of information, not all are, and learners must decide whom to ask for information. What governs learners' choices of informants?

Most prior research about learners' sensitivity to the reliability of informants has been conducted with children. Koenig and Harris (2005) found that by four years of age, children track whether informants have been correct or incorrect in the past and use this to guide their future choices of informants. Moreover, children are sensitive to parametric variations in informants' accuracy (Pasquini, Corriveau, Koenig, & Harris, 2007). Additionally, children use information about group consensus to select informants (Corriveau, Fusaro, & Harris, 2009). These results suggest that by four years of age, children can use diverse cues to establish the reliability of informants.

However, there is good reason to believe that reliability is not the only factor that influences children's epistemic trust. A recent study by Gweon, Pelton, and Schulz (2011) suggests that children not only expect teachers to provide accurate data, but also expect teachers to provide inductively sufficient data. Children gave lower ratings to a teacher who showed one function of a toy that actually had multiple functions, than to a teacher who gave the same demonstration on a toy that actually had just the one function.

Indeed, one advantage of social learning is that it reduces the amount of data required for accurate inference by allowing the learner to make inductive inferences from small amounts of data. How much evidence is enough?

We hypothesize that people do not simply use the sheer quantity of data to decide how helpful a teacher is, but instead consider the extent to which the data provided supports accurate inductive inferences. If a learner's goal is simply to acquire as much data as possible, people should always prefer a teacher who offers more data. However, if the learner's goal is specifically to acquire as much data as necessary for accurate inference, then two teachers can be considered equally helpful, even if one of them provides much less data overall. Consider the toys in Figure 1a; the toys have a number of knobs, which when pressed may or may not cause exciting effects. As a learner, you may be curious to know how many of the knobs cause an effect. You may also have past experience with toys like this, and this experience might generate expectations about how many knobs are likely to work. For instance, you may know that just a few knobs (e.g., two on average) cause effects and the rest do not (independent of the total number of knobs). If you were to learn about one of the toys, would you choose a demonstrator who exhaustively pressed all of the knobs, or one who pressed a few working knobs and stopped? (See Figure 1b.)

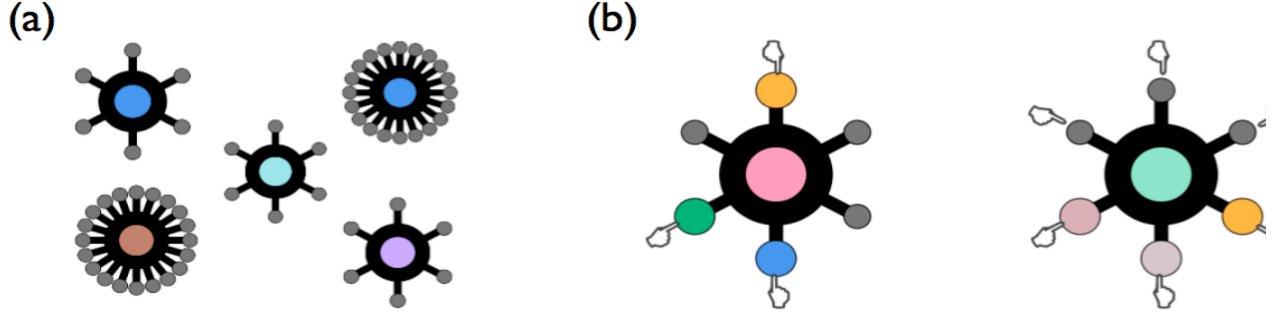


Figure 1: Figure illustrating the toys used in the experiment, and possible demonstrations. (a) Toys could have 6 or 20 knobs, which when pressed may (or may not) lead the toy to create a sound and the button to change color. (b) Two possible sets of demonstrations. In the first case, only three knobs are pressed, all of which elicit effects. In the second case, all of the knobs are pressed, only three of which elicit effects.

Clearly, the exhaustive demonstration would provide the most data; indeed, such a demonstration would be deductively sufficient to infer the number of working knobs. If the only goal is to maximize the overall amount of data, the learner should always prefer more demonstrations to fewer demonstrations.

By contrast, if the learner is sensitive to the cost of providing data and cares most about minimizing this cost, then the learner should penalize teachers who provide data beyond what is necessary to make a reasonable guess. Rather than thinking that a teacher who provides exhaustive evidence is being helpful (e.g., by demonstrating that the remaining knobs are in fact inert and thereby marginally reducing the uncertainty), a learner who has a strong bias against incurring the cost of additional demonstrations should resist demonstrations that are consistent with the learner's beliefs. In this case, learners should prefer informants who provide fewer demonstrations.

However, if as we hypothesize, learners value data that leads to accurate inferences, learners may be satisfied with seeing just a few working knobs (since the learner can infer that informant omitted superfluous demonstrations of the inert knobs), but also may be happy to see additional demonstrations that provide maximal certainty about the toy. Such a flexible trade-off between efficiency and certainty would lead to a preference for inductively sufficient demonstrations.

Building off of recent research formalizing data selection in teaching situations, we introduce a computational model of efficient teaching. The model captures these three hypotheses about learners' choices of informants: that learners strongly prefer informants who offer as much data as possible; that learners are very sensitive to the cost of data and thus prefer teachers who offer as little data as possible, or finally, that learners care that the data supports accurate induction but are happy to acquire additional data as well. In our experiment, we ask adult participants to choose between informants who provide data that is always true but varies in quantity and informativeness. We conclude by contrasting our research with previous findings

and discussing the implications for cognition, cognitive development and education.

### A computational model of efficient teaching

To formalize what constitutes sufficient data, we must consider which data should be chosen and the degree to which the data increase the learners' certainty relative to the added cost of the demonstration. To do so, we adopt a Bayesian learning perspective, building off Shafto and Goodman's (2008) research formalizing teaching and learning in pedagogical settings.

In Bayesian learning, the goal for the learner is to infer the probabilities of different hypotheses,  $h$ , given data,  $d$ . The degree to which the learner believes a hypothesis after observing the data---the learner's posterior beliefs---are denoted  $P(h|d)$ . According to Bayes' theorem, posterior beliefs are determined by the product of the learner's prior beliefs in the hypothesis,  $P(h)$ , and the probability of sampling the data assuming the hypothesis is true,  $P(d|h)$ .

Standard approaches to learning typically assume that data are sampled randomly. However, in pedagogical contexts in which the informant is knowledgeable and the informant's goal is to help the learner infer the true hypothesis, the data are not randomly sampled, but purposefully selected. Shafto and Goodman (2008) formalized teaching and learning in such pedagogical situations. The key differences between pedagogical and random sampling are that the teacher is assumed to be knowledgeable and helpful in her choice of data, and the learner believes that the teacher is knowledgeable and helpful. Teaching is formalized as choosing data that tend to maximize the learner's probability of inferring the correct hypothesis,  $P_T(d|h) \propto P_L(h|d)$ , where the subscripts T and L indicate teacher and learner, respectively. Learners update their beliefs using the knowledge that teachers are choosing data purposefully,

$$P_L(h|d) \propto P_T(d|h)P_L(h). \quad (1)$$

That is, the key difference between Equation 1 and standard approaches to learning is that in Equation 1 the learner

updates her beliefs based on the assumption that the teacher chooses data to help the learner infer the correct hypothesis.

Here we propose that teachers, in addition to choosing data that is helpful, may also consider the *degree* to which additional data increase the learner's certainty. Similarly, learners may vary in how much data they expect the teacher to provide. To capture this difference, we introduce prior probabilities of choosing data. The pedagogical model can be extended to reflect this fact by introducing a term,  $P(d)$ , in the teacher's choice of data,

$$P_T(d|h) \propto P_L(h|d)P(d). \quad (2)$$

Data are assumed to have a cost, and differences in the cost of data capture the three hypotheses of interest.

The probability of data  $P(d)$  depends on two factors: the number of total demonstrations,  $n$ , and the cost of an individual demonstration,  $c$ . Intuitively, an informant may be biased toward presenting more data, less data, or may be unbiased; the learner may accordingly have different expectations of the informant. These three possibilities correspond to three qualitatively different cost parameters in our model. If the learner expects the teacher to demonstrate as much data as possible then providing fewer demonstrations incurs a higher cost. If the learner prefers to minimize the number of demonstrations, then given a choice between more data and less data, less data is more probable; the total cost of a set of demonstrations increases with the number of demonstrations. If the learner is willing to accept any evidence that it is inductively sufficient, then any amount of data is equiprobable; the total costs are constant, independent of the quantity of data. To capture these possibilities, we formalize the prior probability of the data as

$$P(d) \propto e^{-cn}. \quad (3)$$

A negative value of  $c$  corresponds to an expectation of more data; a positive value of  $c$  corresponds to an expectation of less data;  $c=0$  corresponds to equiprobable data.

These hypotheses generate different predictions about learners' choices of informants. In the following experiment, we investigate how learners evaluate informants. Participants are asked to make a choice between two informants, which we model using a log-likelihood ratio. To test the hypotheses, we treat the cost as a free parameter and fit it to the behavioral data. If the best-fitting cost parameter is less than 0, then this would support the hypothesis that learners prefer as much data as possible; if the best-fitting parameter is greater than 0, this would support the hypothesis that learners prefer to minimize data; if the best-fitting parameter is approximately 0, then this supports inductive sufficiency, the hypothesis that learners want enough data to make a confident inference but are happy to accept additional data.

### Experiment: Who is the better teacher?

To investigate learners' choices of informants, we conducted an experiment in which two informants provided demonstrations on different toys (as in Figure 1). The

experiment included two conditions. In one condition, either one, two or three knobs worked on the toys (Consistent condition); in the other condition, between 1/6 and 1/2 of the knobs worked on the toys (Proportional condition). These two conditions allowed us to generate cases in which the model makes different predictions about identical sets of evidence. (See Model predictions below.)

Each demonstration consisted of informants pressing knobs that either elicited effects or were inert. Trials varied in the number of demonstrations, as well as their composition. Additionally, toys varied in the number of total knobs (either 6 or 20). Participants used a sliding scale to indicate which informant was relatively more helpful.

This design allows us to assess the correlation between the model predictions and people's choices. In addition, we can investigate specific cases where the three accounts generate contrasting predictions.

## Method

**Participants.** Forty-four University of Louisville undergraduates (22 per condition) participated in exchange for partial or extra credit.

**Materials.** Participants saw a series of novel toys, described as wugs or daxes on a computer screen. The toys had either 6 or 20 knobs extending from a central sphere (see Figure 1). Clicking on a knob caused a change in its color and size. Only some knobs elicited the effect when clicked; the rest were inert.

**Design.** Participants were randomly assigned to one of two conditions: Consistent or Proportional. In both conditions, participants first interacted with three six-knob-toys and then with three twenty-knob toys to learn how many knobs worked on average. In the Consistent condition, the six-knob toys had one, two or three working knobs, as did the twenty-knob toys. In the Proportional condition, the six-knob toys again had one, two, and three working knobs but the twenty-knob toys had 3, 7, and 10 working knobs.

**Procedure.** Participants were seated at Mac Pro Desktop computers. The experiment proceeded in two phases: training and testing. In the training phase, participants learned how many knobs tended to work by interacting with the toys, as described in the Design section. At the end of the training phase, participants indicated how many knobs worked on average for the six-knob and twenty-knob toys. These questions were checks to ensure that the training phase was successful in inducing appropriate prior beliefs about the toys. Participants were then given feedback about the actual average number of working knobs.

During the testing phase, participants were presented with a series of pairs of unique informants each performing demonstrations on a unique toy. The screen was split in half; in each half of the screen there was an informant with a toy. In the Consistent condition, the pairs were generated by

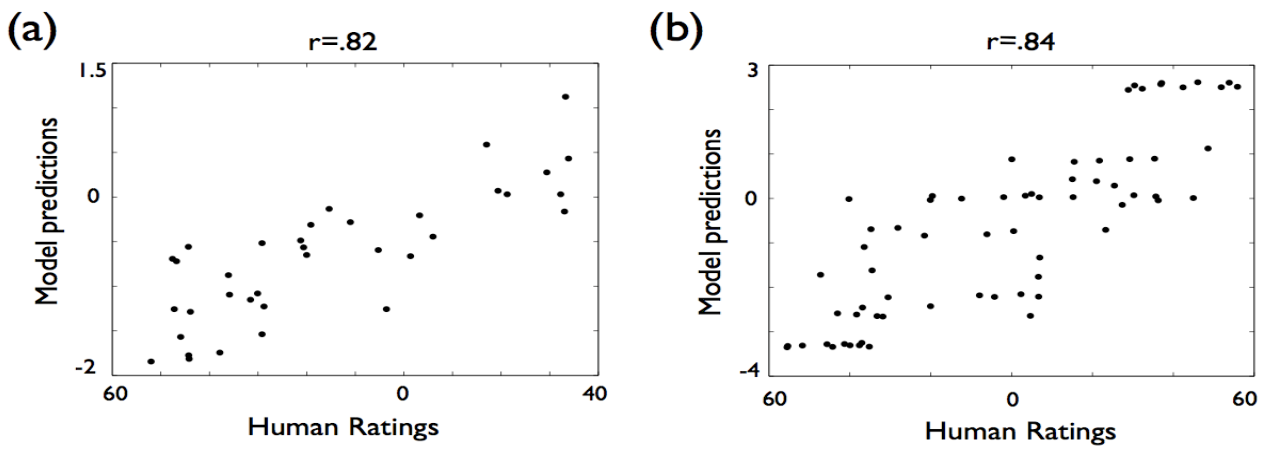


Figure 2: Correlations between human judgments in the (a) Consistent and (b) Proportional conditions. Overall, there is a strong correlation between the model predictions and people's judgments.

comparing all possible pairs of the following demonstrations. For the six-knob-toys, there were four kinds of demonstrations: 3+ 3- (show 3 working knobs and 3 inert knobs), 3+ (show 3 working knobs), 3- (show 3 inert knobs), 2+ 2- (show two working knobs and two inert knobs). For twenty-knob toys, there were five kinds of demonstrations: 3+3-, 3+, 3-, 2+2-, 10-, resulting in 36 questions total. In the Proportional condition, the pairs were generated from the following demonstrations 3+3-, 3+, 3-, 2+2-, 10+, 10-, 7+7- and 10+10- for the twenty-knob toys, resulting in 66 total questions. To ensure that participants remembered the prior knowledge established during training, they were reminded of the average number of working knobs every ten questions.

The start of each demonstration was indicated by the appearance of a hand symbol pointing to a knob, which proceeded around the toy clockwise. The order of positive and negative examples and the locations of working knobs on the toy were determined randomly. After observing the demonstrations by the informant on the left, participants watched the informant on the right. After both sets of demonstrations, participants indicated which informant they judged as 'more helpful' using a slider that appeared in the middle, below the two toys. The order of pairs was randomized, as were the sides on which each informant appeared. After completing all of the questions for their condition, participants were debriefed and thanked.

**Modeling.** The prior knowledge,  $P(h)$ , was set based on the demonstrations that participants observed. For simplicity, trials were assumed to be independent. For each condition, the number of working knobs (out of total number of knobs on the toy) were entered to a Beta-Binomial model with uniform parameters and the resulting distribution was the prior for the experimental judgments. This was performed separately for the six- and twenty-knob toys for each condition.

To find the parameter that best fits people's judgments, we performed a grid search over the values from -2 to 2 in increments of .02.

## Results & Discussion

As an initial test of the model, we assessed the correlation between people's judgments and the model predictions for

the 36 questions in the Consistent condition and the 66 questions in the Proportional condition. To do so, we fit the cost parameter separately to each of the two sets of data. The best-fitting value for the consistent condition was .02 and for the proportional condition was 0. These resulted in robust correlations between the model predictions and human judgments; they were  $r=.82$  and  $r=.84$  for the Consistent and Proportional conditions, respectively.

We can compare our model to a number of alternative proposals in which people's judgments might be explained by attention to more superficial aspects of the stimuli. Specifically, we investigated whether people's judgments were consistent with choosing based on the number of knobs on the toy, the number of positive examples demonstrated, the number of negative examples demonstrated, the number of knobs pressed, the percent of positive examples (out of the total knobs), and the percent of negative examples (out of the total number of knobs).

Our model provided significantly better fits to people's judgments than (a) number of knobs (Consistent:  $r=-.47$ ,  $z=6.77$ ,  $p<.0001$ ; Proportional:  $r=-.47$ ,  $z=9.72$ ,  $p<.0001$ ), (b) the number of knobs pressed (Consistent:  $r=.27$ ,  $z=3.57$ ,  $p<.001$ ; Proportional:  $r=.58$ ,  $z=3.14$ ,  $p<.01$ ), and (c) the number of negative examples demonstrated (Consistent:  $r=-.14$ ,  $z=5.27$ ,  $p<.0001$ ; Proportional:  $r=.01$ ,  $z=6.80$ ,  $p<.0001$ ), (d) the percent of negative examples (Consistent:  $r=.13$ ,  $z=4.17$ ,  $p<.0001$ ; Proportional:  $r=.14$ ,  $z=6.06$ ,  $p<.0001$ ).

However, the correlation between people's judgments and (e) the number of positive examples demonstrated were not significantly different from our model (Consistent:  $r=.66$ ,  $z=1.48$ ,  $p=.14$ ; Proportional:  $r=.75$ ,  $z=1.39$ ,  $p=.16$ ) and (f) the percent of positive examples (Consistent:  $r=.77$ ,  $z=-.55$ ,  $p=.58$ ; Proportional:  $r=.88$ ,  $z=-.87$ ,  $p=.38$ ).

To further investigate the degree to which our model and the remaining alternatives (number of positive examples and percent of positive examples) fit the data, we turn to analyses of individuals' judgments. For our model, we fit the parameter to each individual's judgments as described above. We correlated the predictions of our model, the number of positive examples, and the percent of positive examples with individual participants' judgments. Our model (Consistent:  $M=.52$ ; Proportional:  $M=.57$ ) predicted people's judgments better than the number of positive examples (Consistent:  $M=.29$ ,  $t(42)=2.7$ ,  $p<.01$ ; Proportional:  $M=.38$ ,  $t(42)=2.84$ ,  $p<.01$ , by one-tailed t-test)

and the percent of positive examples (Consistent:  $M=.33$ ,  $t(42)=2.27$ ,  $p<.05$ ; Proportional:  $M=.46$ ,  $t(42)=1.67$ ,  $p=.05$ , by one-tailed t-test). These results suggest that our model provides a better explanation of people's behavior than these alternatives.

Next, we turn to the amount of data learners expect. Recall that expecting the informant to provide as much data as possible is indicated by parameter values much greater than 0, expecting the informant to provide as little data as possible is indicated by parameter values much less than 0, and expecting inductively sufficient data is indicated by parameter values near 0. For the group data, the best fitting parameters were 0 for the Consistent condition, and .02 for the Proportional condition. These parameters are most consistent with the hypothesis that learners expect data that suffices for accurate inference but exact no penalty for additional data.

The three different accounts make opposite qualitative predictions for subsets of the questions. To explore these differences we contrast the three hypotheses using parameter values of 2, -2, and 0 respectively.

The hypothesis that learners expect as much evidence as possible and the hypothesis that learners expect inductively sufficient evidence but are happy with more evidence make opposite predictions for five questions in both the Consistent and Proportional conditions. For example, in the Consistent condition, the preference for maximal evidence predicts that 2+2- should be preferred to 3+ (because there's a total of four demonstrations versus only three); whereas inductive sufficiency predicts the opposite preference.

With so few questions, it is not surprising that, although people's judgments tended toward a preference for inductive sufficiency, there were not statistically significant differences in participants' responses to these questions (Consistent:  $M=11.36$ ,  $t(4)=1.81$ ,  $p=.14$ ; Proportional:  $M=4.66$ ,  $t(4)=-.38$ ,  $p=.72$ ).

To separate the predictions of a preference for maximal data and inductive sufficiency, we identified the ten questions on which the predictions differed the most in each condition. We standardized the predictions of each model and chose the questions that had the largest absolute difference in predictions.

In the Consistent condition, the question with the largest difference compared a six-knob-toy with 3- and a twenty-knob-toy with 3+. Inductive sufficiency predicts a strong preference for the 3+ demonstration on the twenty-knob toy (because the learner is certain that three knobs work on the twenty-knob toy but uncertain whether 1, 2, or 3 knobs work the six-knob-toy) whereas a preference for maximal data predicts a strong preference for the 3- demonstration on the six-knob-toy (because the learner has evidence for half of the knobs on the six-knob-toy but only 3 of 20 for the twenty-knob-toy). That is, the learner might prefer the greater confidence afforded by the demonstration of the working knobs or might prefer a greater relative number of demonstrations (3- out of 6). Over the ten questions, there was a stronger correlation between inductive sufficiency and

people's judgments,  $r=.85$ , than between the preference for maximal evidence and people's judgments,  $r=-.20$ ,  $z=2.73$ ,  $p<.01$ .

In the Proportional condition, the question with the largest difference compared 10+ versus 10- on twenty-knob-toys. Inductive sufficiency predicts a strong preference for 10+, whereas the preference for maximizing demonstrations predicts a slight preference for 10+. Over the 10 questions, there was a stronger correlation between inductive sufficiency and people's judgments,  $r=.90$ , than between maximizing demonstrations and people's judgments,  $r=.29$ ,  $z=2.2$ ,  $p<.05$ .

The hypothesis that learners expect as little evidence as possible and the hypothesis that learners expect inductively sufficient evidence make opposite predictions for 13 questions in the Consistent condition and 18 questions in the Proportional condition. For example, in the Consistent condition, a preference for minimal demonstrations predicts that 2+2- is preferred to 3+3- whereas inductive sufficiency predicts the opposite.

People's responses on these questions were coded as positive if consistent with the predictions of inductive sufficiency and negative if they were consistent with a preference for less data. People's judgments in the Consistent condition were in agreement with the predictions of inductive sufficiency both on average,  $M=32.6$ ,  $t(12)=8.91$ ,  $p<.0001$ , and in every individual case. Similarly, people's judgments in the Proportional condition were overwhelmingly in accord with the predictions of inductive sufficiency,  $M=34.63$ ,  $t(17)=10.27$ ,  $p<.0001$ .

## General Discussion

We have proposed that learners' choice of informants is guided primarily by the degree to which evidence supports accurate inference. We presented a computational model that differentiates among the hypotheses that learners choose informants who provide as much data as possible, informants who minimize the amount of data provided, and informants who provide at least enough data to support accurate induction. The results show that people's behavior is best explained by inductive sufficiency.

Note that providing maximal data can, in simple cases, lead to deductive certainty. For finite, well-defined sets of possibilities (like those tested here), exhaustive demonstrations eliminate uncertainty. However, our results show that even on relatively small, well-defined learning problems, learners do not simply prefer informants who provide maximal amounts of data. In fact, people are just as likely to endorse much smaller sets of data, as long as the data provided suffices for accurate inductive inference. This suggests that learners are sensitive to the trade-off between the benefit of increased certainty from acquiring more data and the cost of acquiring more data; this sensitivity enables learners to decide how much data is 'enough'.

We did not find evidence for a simple preference for less data. A particularly interesting example is that people did not prefer someone who provides three working knobs over

someone who provides exhaustive evidence, despite that the amount of data were twice as much in the latter. This may be due to features of our experimental design. There was relatively little reason to avoid additional demonstrations (among other things, each additional knob only took a second to press). Furthermore, our dependent measure asked learners to rate the helpfulness of the informant and learners have little reason to consider an exhaustive informant unhelpful. Finally, the additional demonstrations genuinely reduced some uncertainty: each toy was unique and the training only provided a few examples to establish the base rate of the effective knobs. Had the learners been more certain about the number of working knobs, they might have shown a stronger bias against exhaustive evidence.

We have presented evidence that learners' choice of informants is not solely guided the sheer amount of information; learners do not merely maximize the amount of data they can observe, nor do they minimize it. Learners use their inductive certainty to decide when enough is enough.

### Acknowledgments

This research was partially supported by a James S. McDonnell foundation subaward to P.S.

### References

- Corriveau, K. H., Fusaro, M., & Harris, P. L. (2009). Going with the flow: Preschoolers prefer non-dissenters as informants. *Psychological Science*, 20, 372–377.
- Corriveau, K. H., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and past accuracy. *Developmental Science*, 12, 426–437.
- Gweon, H., Pelton, H., & Schulz, L. E. (2011). Adults and school-aged children accurately evaluate sins of omission in pedagogical contexts. In *Proceedings of the 33rd annual conference of the cognitive science society*.
- Koenig, M., & Harris, P. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, 76, 1261–1277.
- Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, 112, 367–380.
- Mischel, W., Shoda, Y., & Rodriguez, M. (1989). Delay of gratification in children. *Science*, 244, 933–938.
- Pasquini, E. S., Corriveau, K. H., Koenig, M. A., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, 43, 1216–1226.
- Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*, 15, 436–447.
- Shafto, P., & Goodman, N. D. (2008). Teaching games: Statistical sampling assumptions for pedagogical situations. In *Proceedings of the 30th annual conference of the Cognitive Science Society*.

# Investigating the Locus of the Word Frequency Effect in Spoken Word Recognition

Cynthia S. Q. Siew (cynsiewsq@gmail.com)

Melvin J. Yap (melvin@nus.edu.sg)

Winston D. Goh (psygohw@nus.edu.sg)

Department of Psychology, National University of Singapore, 9 Arts Link,  
Singapore 117570, Singapore

## Abstract

The present study aims to isolate the locus of the frequency effect within the spoken word recognition architecture. By applying the additive factors logic (Sternberg, 1969) to an auditory lexical decision task where both word frequency and stimulus quality were factorially manipulated, the reaction time data can be analyzed to study processing stages along the time course of spoken word recognition, and determine if frequency has an early or late locus. A significant underadditive interaction of frequency and stimulus quality was obtained. Surprisingly, the typically robust frequency effect was not reliable for words of low stimulus quality. This finding suggests that word frequency influences a relatively late stage in the spoken word recognition process. Implications for extant models of spoken word recognition are discussed.

**Keywords:** Spoken word recognition; word frequency effects; stimulus quality effects; additive factors logic; auditory lexical decision.

## Introduction

Determining whether word frequency has an early or late locus has profound theoretical implications for models of spoken word recognition (SWR). While it is well established that frequently occurring words are recognized faster than less frequently occurring words (Goldinger, 1996), what is less obvious is where the *locus* of the frequency effect lies within the word recognition process. Specifically, does frequency influence word recognition at an early stage, as the speech signal begins to unfold, or does frequency influence word recognition at a later stage in the form of a bias? Models of SWR can easily account for the frequency effect, but they do not necessarily agree on the locus of the frequency effect due to varying assumptions and architectures. Hence one way to test the validity of these models is to isolate the locus of the frequency effect.

Several researchers have investigated this issue by employing a variety of experimental techniques and methodologies. Generally, studies which used traditional behavioral experiments (e.g., lexical decision and word identification) have demonstrated that word frequency has a late locus that occurs after lexical processes are complete (Broadbent, 1967; Connine, Mullennix, Shernoff & Yelen, 1990; Luce & Pisoni, 1998). On the other hand, recent studies employing eyetracking technology (e.g., Dahan, Magnuson & Tanenhaus, 2001) and novel behavioral applications of the parallel refractory period paradigm (Cleland, Gaskell, Quinlan & Tamminen, 2006) concluded that word frequency exerts early and facilitatory effects on

word recognition. With overwhelming evidence supporting both sides of the debate, the question as to whether word frequency affects spoken word recognition at an early or late stage continues.

The present study aims to isolate the locus of the frequency effect in spoken word recognition by making use of the additive factors logic to investigate this particular research question. The additive factors logic (Sternberg, 1969) is widely used by cognitive psychologists to interpret RT data in factorial experiments and study the stages of processing in a number of research topics (e.g., Stanovich & Pachella, 1977), as the logic can be easily applied in the study a wide array of research topics, including psycholinguistics (e.g., Yap & Balota, 2007).

## Additive Factors Logic

According to the additive factors logic (Sternberg, 1969), when two factors affect theoretically determined independent stages in the information processing stream, it should result in *additivity* in mean RTs (i.e., two main effects for each factor, but no interaction). This is represented in the top part of Figure 1, where Factor A affects processing at only Stage 1 and Factor B affects processing at only Stage 2. On the other hand, if the two factors affect the same stage in the information processing stream, this results in a statistical interaction (more precisely, an *overadditive* interaction where the effect of one factor is larger on the “slower” level of the second factor). This is depicted in the bottom part of Figure 1, where both Factor A and Factor B affect processing at a common Stage X.

How can the incorporation of an additional variable, stimulus quality within the auditory lexical decision task, allow us to isolate the locus of the word frequency effect? How can the additive factors logic be used to help us make specific hypotheses about the pattern of results for RT data? In contrast to the lack of consensus with regards to the locus of the frequency effect, few would question the notion that stimulus quality has an early locus of influence in the word recognition process. In fact, a major assumption of most SWR models (e.g., TRACE) involves a process which converts physical, acoustic input into phonemic information (McClelland & Elman, 1986). This necessarily implies that degraded input must be normalized at a relatively early point in the word recognition process.

Hence, if we assume that stimulus quality affects an early stage in the word recognition process, then Factor A corresponds to stimulus quality and Factor B corresponds to



frequency, as shown in Figure 2. Hence, additivity (i.e., main effects of frequency and stimulus quality, but no interaction) indicates that stimulus quality and frequency have independent loci of influence, and this further implies that frequency affects a later stage (one that occurs after stimulus quality; as shown in the upper section of Figure 2). An overadditive interaction where the frequency effect is greater for words of low stimulus quality as compared to high stimulus quality would indicate that these two variables influence at least one stage in common, and it follows that frequency has an early locus of influence (as shown in the bottom of Figure 2).

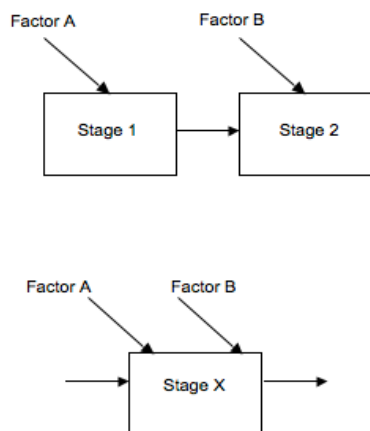


Figure 1. Sternberg's (1969) Additive Factors Logic

In fact, within the area of visual word recognition, some studies have employed the additive factors logic to investigate the joint effects of stimulus quality and frequency in lexical decision (Borowsky & Besner, 1993; Yap & Balota, 2007). Researchers have consistently found that frequency and stimulus quality have *additive* effects in visual lexical decision. This finding is best accommodated within a two-stage model where stimulus quality influences an early stage and frequency influences the second stage (Borowsky & Besner, 1993), which implies that processing at earlier stages is not necessarily frequency-sensitive.

It is also interesting to note that, despite extensive research involving perceptual identification and auditory lexical decision paradigms, researchers almost universally study the effect of stimulus quality on identification accuracy, but not on response latencies. The study of the effects of stimulus quality on spoken word recognition has been largely limited to perceptual and tone identification experiments (Broadbent, 1967; Hawkins & Stevens, 1950; Luce & Pisoni, 1998; Savin, 1963). To our knowledge, stimulus quality has never been directly manipulated as an independent variable in auditory lexical decision, and the joint effects of frequency and stimulus quality have not been previously studied in any auditory word recognition task. Hence, another objective of the present study is to address these gaps in the literature.

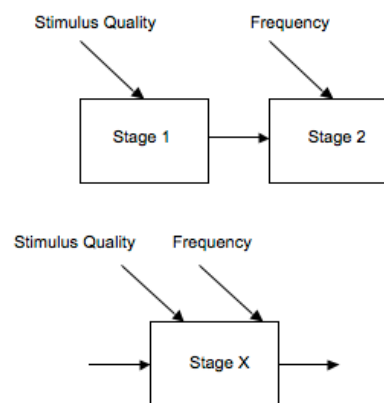


Figure 2. Hypothetical diagrammatic representation of the loci of stimulus quality and frequency effects

## Method

### Participants

Eighty National University of Singapore undergraduates participated in this study for course credit. Participants' first language was English, and they had no previous reported history of speech or hearing disorders.

### Design

A 2 (word frequency: high, low)  $\times$  2 (stimulus quality: clear, degraded) mixed-design was used. The within-participants independent variable was word frequency and the between-participants independent variable was stimulus quality. Stimulus quality was manipulated as a between-participants variable to minimize possible carry-over effects that may occur in a fully within-participants design (Poulton, 1982). The dependent variables were reaction time (RT) and accuracy.

### Stimuli

Table 1 shows a summary of the descriptive statistics for word and nonword stimuli. 58 high frequency and 58 low frequency English words were selected as stimuli. Using LogFreqHal values generated from the English Lexicon Project (ELP; Balota *et al.*, 2007), the difference between the high and low frequency conditions was reliable,  $F(1,114) = 329.72$ ,  $MSe = 273.29$ ,  $p < .001$ . High and low frequency words were also matched on number of phonemes, number of syllables, phonological neighborhood density, familiarity rating, uniqueness point, and word duration. A one-way between-items ANOVA showed that for all lexical characteristics,  $F_s < 1$ .

116 nonwords were constructed and matched with words on number of phonemes, number of syllables, duration and baseword phonological neighborhood density (an estimation of the nonword density based on the neighborhood density of its closest sounding word). The difference between words and nonwords on each of those variables was not significant, all  $F_s < 1$ . All stimuli were spoken by a linguistically trained female speaker and digitally recorded in 16-bit mono, 44.1kHz, .wav format.

Table 1: Descriptive statistics of word and nonword stimuli.

Lexical Characteristics	Word Stimuli				Nonword Stimuli	
	High Frequency		Low Frequency		<i>M</i>	<i>SD</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Number of phonemes	3.95	0.98	4.12	1.04	4.03	1.01
Number of syllables	1.48	0.60	1.43	0.53	1.46	0.57
Phonological neighbourhood density (Baseword phonological density for nonwords)	6.98	6.80	7.31	7.69	7.15	7.23
Duration	582	91	590	81	586	86
Familiarity rating	6.99	0.07	6.98	0.06		
LogFreqHal	11.37	0.75	8.30	1.05		
Uniqueness Point	4.33	0.87	4.26	0.79		

**Degrading auditory stimuli.** White noise was used to degrade the spoken stimuli, in accordance with past perceptual identification experiments (Broadbent, 1967; Savin, 1963). All degraded trials were presented at SNR +10dB, with white noise at 70dB and target stimuli at 80dB.

**Phonemic distributions.** Various studies have shown that white noise has differential masking effects on different phonemes (Horii, House & Hughes, 1970; Pisoni, 1996). Following Chan and Vitevitch (2007), chi-square analyses were conducted on the onset consonants, vowels and fricatives of all word stimuli to ensure that no single phoneme was overrepresented among them. The phonetic transcriptions of each word were obtained from the ELP, and subsequent chi-square analyses were not significant.

## Procedure

Participants were tested on individual PCs in groups no larger than five. Forty subjects were assigned to the clear condition (without noise) and forty participants were assigned to the degraded condition (with noise). Stimuli were binaurally played through BeyerDynamic DT150 headphones, and E-prime 1.2 software and the PST serial response box (Schneider, Eschman & Zuccolotto, 2002) were used for stimuli presentation and data collection. Participants were instructed to listen to the stimuli carefully and decide, as quickly and accurately as possible, whether the token was a word or a nonword, using the right- and left-most buttons respectively on the response box. Prior to the actual experiment, participants were given 20 practice trials which were not included in the subsequent analyses. For degraded trials, white noise was played for 100ms before the stimulus was presented and continued until 100ms after stimulus offset. Once a response was made, 500ms elapsed before the initiation of another trial. Latencies were measured from the onset of the stimulus until button press. There were a total of 232 experimental trials and participants were allowed a short break after every 58 trials.

## Results

For the reaction time data, only correct word trials with RTs more than 200ms and less than 3000ms were included in

the analyses. Trials with RTs less than 200ms were excluded, and trials with RTs more than 3000ms were substituted with 3000ms and included in the analysis. This reduces the amount of data excluded and ensures that extreme scores are preserved while reducing their impact (e.g., Marian, Blumenfeld & Boukrina, 2008). Following which, the overall mean and *SD* of each participant's RT was calculated and trials with latencies that were 3 *SD*s above or below each participant's mean RT were removed. These trimming criteria resulted in the removal of 10.2% of all word trials.

The average RTs and Accuracy across the 4 conditions are summarized in Table 2. A two-way mixed-design ANOVA was conducted on the RT and accuracy data, by participants and items.

Table 2: Mean RTs (ms) and accuracy (proportion)

	RT		Accuracy	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Clear trials				
High frequency	890	64	0.96	0.03
Low frequency	908	65	0.92	0.04
Frequency effect	18		0.04	
Degraded trials				
High frequency	997	130	0.90	0.05
Low frequency	1002	129	0.87	0.07
Frequency effect	5		0.03	
Interaction	13		0.01	
Nonwords	1237	260	0.76	0.15

## Reaction Time

A reliable main effect of stimulus quality,  $F_p(1,78) = 19.56$ ,  $MSe = 400656.27$ ,  $p < .001$ ;  $F_i(1,114) = 263.64$ ,  $MSe = 617419.13$ ,  $p < .001$ , was found for both participant and item analyses. Across high and low frequency words, participants were slower at recognizing words presented with noise ( $M = 999$ ,  $SD = 129$ ) than for words presented in the clear ( $M = 899$ ,  $SD = 64$ ). The main effect of frequency was significant by participants,  $F_p(1,78) = 13.91$ ,  $MSe = 5029.26$ ,  $p < .001$ , but not by items,  $F_i < 1$ . Across both conditions of stimulus quality, response latencies for high frequency words ( $M = 944$ ,  $SD = 115$ ) were significantly

faster than response latencies for low frequency words ( $M = 955$ ,  $SD = 112$ ).

The Frequency  $\times$  Stimulus Quality interaction was significant by participants,  $F_p(1,78) = 4.20$ ,  $MSe = 1517.33$ ,  $p < .05$ , but not by item analyses,  $F_i(1,114) = 2.14$ ,  $MSe = 5019.49$ ,  $ns$ .

Tests of the simple main effect of frequency was significant in the clear condition,  $F(1,78) = 16.69$ ,  $MSe = 6035.73$ ,  $p < .001$ , but not in the degraded condition,  $F(1,78) = 1.41$ ,  $MSe = 510.86$ ,  $ns$ . Participants recognized clear high frequency words ( $M = 890$ ,  $SD = 64$ ) more quickly than clear low frequency words ( $M = 908$ ,  $SD = 65$ ). However, for degraded words, participants did not differ on their response latencies for high ( $M = 997$ ,  $SD = 130$ ) and low frequency words ( $M = 1002$ ,  $SD = 129$ ). There was an 18 ms frequency effect at the clear condition, but this was abolished at the degraded level. Tests of the simple main effect of stimulus quality was significant for both high frequency,  $F(1,78) = 21.57$ ,  $MSe = 225743.00$ ,  $p < .001$ , and low frequency conditions,  $F(1,78) = 16.99$ ,  $MSe = 176430.60$ ,  $p < .001$ . Among high frequency words, participants were slower to recognize degraded words ( $M = 997$ ,  $SD = 130$ ) than clear words ( $M = 890$ ,  $SD = 64$ ). Among low frequency words, participants were also slower to recognize degraded words ( $M = 1002$ ,  $SD = 129$ ) than clear words ( $M = 907$ ,  $SD = 65$ ).

## Accuracy

A reliable main effect of stimulus quality was also found for both participants and items,  $F_p(1,78) = 37.52$ ,  $MSe = .14$ ,  $p < .001$ ;  $F_i(1,114) = 34.68$ ,  $MSe = .20$ ,  $p < .001$ . Participants were more accurate at recognizing high and low frequency words presented in the clear ( $M = 0.94$ ,  $SD = 0.04$ ) than in the degraded condition ( $M = 0.88$ ,  $SD = 0.06$ ). The frequency effect was reliable by participants,  $F_p(1,78) = 30.38$ ,  $MSe = .04$ ,  $p < .001$ , and by items,  $F_i(1,114) = 4.09$ ,  $MSe = .06$ ,  $p < .05$ . Across both levels of stimulus quality, accuracy rates for high frequency words were higher ( $M = 0.93$ ,  $SD = 0.05$ ) than for low frequency words ( $M = 0.90$ ,  $SD = 0.06$ ). No interaction was observed for frequency and stimulus quality in both analyses by participants and by items,  $F_p(1,78) = 2.18$ ,  $MSe = .003$ ,  $ns$ ;  $F_i < 1$ .

## Discussion

In the present study, the joint effects of stimulus quality and word frequency are characterized by an underadditive interaction, as the frequency effect for words of high stimulus quality was reliable but not for words of low stimulus quality. This finding may be considered counterintuitive because additive factors logic does not *a priori* predict underadditivity between two factors. According to additive factors, a statistical interaction is indicative of both variables influencing at least one stage in common in the processing architecture. This interpretation was based on an *overadditive* interaction (Sternberg, 1969),

where the effect of one factor is larger at the “slower” level of the second factor. However, the interaction observed here was an *underadditive* one, where the effect of one factor is smaller, instead of larger, at the “slower” level of the second factor.

Consider Figure 3 below, where stimulus quality influences Stage 1 and frequency influences Stage 2. For clear words, word recognition proceeds from Stage 1 to Stage 2. For degraded words, degradation could have slowed processing to the extent that the optional Stage 2 is not initiated. Note that if we assume Stage 2 to be optional and is presumably not necessary for word recognition, then word recognition can still take place without engaging this frequency-sensitive stage. This is consistent with the hypothesis that stimulus quality and frequency influence *separate* processing stages. According to this interpretation, word frequency has a *late* locus of influence that occurs after that of stimulus quality.

It is interesting to note that the underadditive interaction between frequency and stimulus quality parallels the findings of previous studies which have studied the joint effects of frequency and neighborhood density in auditory lexical decision (Goh, Suarez, Yap, & Tan, 2009; Luce & Pisoni, 1998; Metsala, 1997). These studies found that frequency and neighborhood density interact underadditively, and frequency effects are attenuated for words belonging to dense neighborhoods. This appears to correspond with our present finding that the frequency effect was not reliable for degraded words, as both results indicate that frequency effects are *smaller* when word processing is *slowed* down, either via degradation or neighborhood density effects.

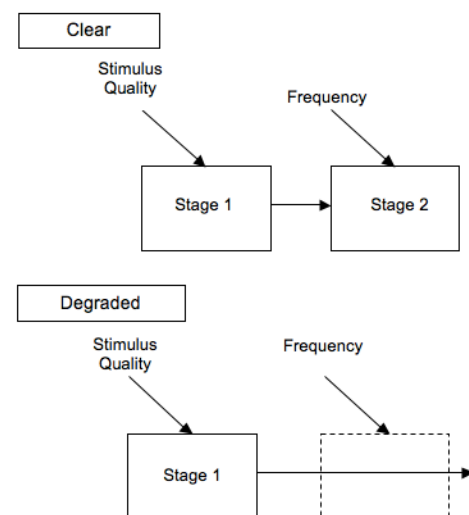


Figure 3. Diagrammatic representation of the underadditive effects of frequency and stimulus in a two-stage model

To some extent, degraded words are analogous to words belonging to dense neighborhoods. Words in dense

neighborhoods have several phonological neighbors, which are defined as words differing from the target word on at least one phoneme in any position (Yates, 2005). The acoustic-phonetic patterns of these words are likely to be more confusable because there are several words sharing similar patterns (Luce & Large, 2001). Hence, words with several neighbors tend to activate many more word units than words with fewer neighbors. This results in more competition among word units which inhibits word recognition performance for high density words (Luce & Pisoni, 1998).

Therefore, it is possible that introducing degradation to word stimuli similarly increases the level of competition among word units. Due to increased acoustic ambiguity, a degraded stimulus can be potentially matched to a large number of words in the lexicon and this leads to a large number of potential word candidates being activated and subsequently competing for recognition. In general, it appears that the ambiguity of the acoustic input (due to either exogenous noise or the perception of the acoustic input as possibly corresponding to several word units) ultimately leads to an increase in competition among activated word units, slowing processing to the point that any biasing effects of word frequency are not observed.

### Implications for Models of SWR

The present study is of theoretical importance because the results can impose additional constraints on speech recognition models. In this section the implications of the underadditive interaction for extant models of SWR are briefly reviewed.

To account for the finding that frequency effects are attenuated in certain tasks (e.g., Connine *et al.* 1990), NAM posits that frequency effects are non-obligatory and that it is possible for word recognition to occur *without* involving the later, frequency-sensitive stage (Luce & Pisoni, 1998). NAM conceptualizes the frequency effect as a decision bias that occurs later in the word recognition process, and it appears that this bias can be “turned off” depending on the task demands and conditions. Therefore, NAM is able to accommodate the present finding because in this model word recognition can still occur with limited or no processing at Stage 2 (see Figure 3), and this explains why a reliable frequency effect was not obtained for degraded words.

Other models of SWR are unable to accommodate the present finding as easily. In the TRACE model of speech perception, (McClelland & Elman, 1986), since frequency and stimulus quality both influence an architecture that allows bidirectional flow of information between processing levels, it should predict an overadditive interaction as this is analogous to two variables influencing a common stage (the influences of word frequency or stimulus quality are not independent of each other). In order to accommodate the underadditive interaction, we speculate that the model needs to allow for a flexible word processing

system that can reduce the influence of frequency when the acoustic input is compromised such that bottom-up flow of perceptual evidence is considerably slowed down.

The results are also inconsistent with the predictions of another major model of SWR - Shortlist B. According to this model, optimal listeners rely more on prior probabilities to compute conditional probabilities (using Bayes Theorem) when ambiguity of the speech input is high (Norris & McQueen, 2008). Since prior probabilities of words are approximated to word frequency and adding white noise to spoken stimuli increases the perceptual uncertainty of the speech input, the model predicts that the frequency effect should be larger for degraded words compared to clear words (Norris & McQueen, 2008). However, in this study, the frequency effect was abolished for degraded words, which seems to imply that listeners actually rely *less* on prior probabilities under increased perceptual uncertainty.

In summary, to account for the non-reliable frequency effect in the degraded condition, we proposed that this was due to degradation inducing a high level of competition among word candidates, such that the frequency-sensitive stage is not invoked over the course of spoken word recognition. Therefore, the finding of an underadditive, rather than overadditive, interaction between stimulus quality and frequency can be accommodated by a two-stage model where the second, frequency-sensitive stage is not mandatory for word recognition. This further suggests that these variables influence *separate* stages in the word recognition process, and by extension, that word frequency has a late locus of influence occurring after that of stimulus quality.

### Notes

As suggested by a reviewer, we conducted mix-effects modelling on our data using R (R Development Core Team, 2011). A linear mixed effects model was fitted to the RT data from the experiment, using the lme4 package (Bates *et al.*, 2012); *p*-values for fixed effects were computed using the languageR package (Baayen, 2012). The main effects of stimulus quality and frequency, and the interaction between the two factors were treated as fixed effects, while participants and items were treated as random variables. Our results revealed a significant main effect of stimulus quality ( $p < .001$ ) and no effect of frequency. These were qualified by a marginally significant stimulus quality by frequency interaction,  $p = .070$ .

### Acknowledgements

This work was supported by Research Support Scheme C-581-000-222-091 to WDG. We thank Eileen Soh for data collection assistance.

### References

Baayen, R. H. (2012). languageR: Data Sets and Functions with “Analyzing Linguistic Data: A Practical

- Introduction to Statistics,” R Package Version 1.4. Vienna: R Foundation for Statistical Computing.
- Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445-459.
- Bates, D. M., Maechler, M., & Dai, B. (2012). lme4: Linear Mixed-Effect Models Using S4 Classes, R Package Version 0.999375-42. Vienna: R Foundation for Statistical Computing.
- Borowsky, R., & Besner, D. (1993). Visual word recognition: A multistage activation model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 813-840.
- Broadbent, D. E. (1967). Word-frequency effect and response bias. *Psychological Review*, 74, 1-15.
- Chan, K. Y., & Vitevitch, M. S. (2007). The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 1934-1949.
- Cleland, A. A., Gaskell, M. G., Quinlan, P. T., & Tamminen, J. (2006). Frequency effects in spoken and visual word recognition: Evidence from dual-task methodologies. *Journal of Experimental Psychology: Human Perception and Performance*, 32, 104-119.
- Connine, C. M., Mullennix, J., Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 1084-1096.
- Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, 42, 317-367.
- Goh, W. D., Suarez, L., Yap, M. J., & Tan, S. H. (2009). Distributional analyses in auditory lexical decision: Neighborhood density and word-frequency effects. *Psychonomic Bulletin & Review*, 16, 882-887.
- Goldinger, S. D. (1996). Auditory lexical decision. *Language and Cognitive Processes*, 11, 559-567.
- Hawkins, J. E., & Stevens, S. S. (1950). The masking of pure tones and of speech by white noise. *Journal of the Acoustical Society of America*, 22, 6-13.
- Horii, Y., House, A. S., & Hughes, G. W. (1970). A masking noise with speech-envelope characteristics for studying intelligibility. *Journal of the Acoustical Society of America*, 49, 1849-1856.
- Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, 16, 565-581.
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear & Hearing*, 19, 1-36.
- Marian, V., Blumenfeld, H. K., & Boukrina, O. V. (2008). Sensitivity to phonological similarity within and across languages. *Journal of Psycholinguistic Research*, 37, 141-170.
- McClelland, J. L., & Elman, J. L. (1986). The TRACE Model of Speech Perception. *Cognitive Psychology*, 18, 1-86.
- Metsala, J. L. (1997). An examination of word frequency and neighborhood density in the development of spoken-word recognition. *Memory & Cognition*, 25, 47-56.
- Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115, 357-395.
- Pisoni, D. B. (1996). Word identification in noise. *Language and Cognitive Processes*, 11, 681-687.
- Poulton, E. C. (1982). Influential companions: Effects of one strategy on another in the within-subjects designs of cognitive psychology. *Psychological Bulletin*, 91, 673-690.
- R Development Core Team. (2011). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.
- Savin, H. B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America*, 35, 200-206.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2001). *E-prime user's guide*. Pittsburgh: Psychology Software Tools, Inc.
- Stanovich, K. E., & Pachella, R. G. (1977). Encoding, stimulus-response compatibility, and stages of processing. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 411-421.
- Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, 30, 276-315.
- Yap, M. J., & Balota, D. A. (2007). Additive and interactive effects on response time distributions in visual word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 274-296.
- Yates, M. (2005). Phonological neighbors speed visual word processing: Evidence from multiple tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 1385-1397.

# Zero anaphora and object reference in Japanese child-directed speech

**Cybelle Smith**

cybelle@umd.edu

Department of Linguistics  
University of Maryland

**Michael C. Frank**

mcf Frank@stanford.edu

Department of Psychology  
Stanford University

## Abstract

To learn the meanings of words, children must connect referents in the world around them with the sounds they hear. One proposed mechanism for this process is cross-situational word learning: tracking associations between words and objects across time. We consider the problem of anaphora for a cross-situational word learner: after an object has been introduced it is unlikely to be named in every succeeding reference to it. This problem is particularly pronounced in Japanese, which uses “zero anaphora,” where pronouns can be omitted from utterances. We analyze a corpus of Japanese mothers talking to children about sets of objects, originally recorded by Fernald and Morikawa (1993). Overall rates of anaphora were much higher for Japanese mothers compared with English mothers. Zero anaphora was primarily used when the discourse topic was already established, suggesting that a discourse-finding strategy may be important for word learning in Japanese. In addition, unexpectedly, due to the existence of zero anaphora as a common referential strategy, pronouns were more likely to be used when the topic was new than when it was given (reversing common results for English).

**Keywords:** Child-directed speech; Japanese; zero anaphora; discourse analysis; language acquisition.

## Introduction

From the perspective of a scientist, early word learning seems a difficult problem. Although infants quickly learn to pair sound sequences in their caregivers’ speech with concepts and entities in their environment, it is still uncertain how they associate words and concepts. One proposed mechanism for this process is cross-situational word learning: tracking associations between words and objects across time (Siskind, 1996; Yu & Smith, 2007). Though possible for even very young children in simple contexts (Smith & Yu, 2008) and in principle feasible for large lexicons (Blythe, Smith, & Smith, 2010), the effectiveness of pure cross-situational learning in natural contexts is still unknown.

One issue for cross-situational learning in the natural learning environment is *anaphora*: the use of shortened—or even absent—expressions that refer back to a previously-named entity. If learners are keeping track of associations between words and objects, the tendency towards anaphoric reference should cut down considerably on these associations. Although anaphora is pervasive in language use, object names are typically repeated frequently in the speech of English-speaking mothers to their children (Fernald & Morikawa, 1993). This repetition has allowed models of cross-situational word learning to succeed in establishing word-object mappings even in small natural datasets (Yu & Ballard, 2007; Frank, Goodman, & Tenenbaum, 2009).

Repetition of object labels is not nearly as prevalent in some other languages, however. In the same study that established the presence of repetition in English mothers’ speech,

Fernald and Morikawa (1993) noted that Japanese mothers used far fewer noun labels and the labels they used were generally more diverse, including onomatopoeia and diminutive forms as well as the prototypical labels used by American mothers. In addition, Japanese, unlike English, is a pro-drop language, meaning that the subject and object of verbs may be omitted. This omission is known as “zero anaphora.” For example, when an English speaker might say “the dog barks,” a Japanese speaker might say only “barks.” One study suggests that zero anaphora in Japanese may lead to increased difficulty in early verb learning (Rispoli, 1995), but to our knowledge, no work has examined the direct relationship between zero anaphora and object reference, in Japanese or any other language.

What effect does the varied use of noun-labels and anaphora (especially zero anaphora) have for Japanese word learners? Under a pure cross-situational analysis, the sparse mappings between words and objects in this language might be very difficult to overcome. If objects that are being talked about often go unnamed in Japanese, a pure cross-situational learner might be more likely to learn, for example, “bark” rather than “dog” for the concept of a dog. If, however, word learners do not treat utterances as independent entities, but instead resolve reference within a *topical discourse*—a set of utterances about a particular topic—the problems posed by changing object labels and zero anaphora might be mitigated. A learner could figure out what topic was being talked about and then assume that future utterances refer to this topic, even if it was not named.

Recent work has taken up the suggestion that children could potentially aggregate information about word meanings—as well as other knowledge about a particular object referent—not just across sentences but also across these topical discourse units (Rohde & Frank, under review; Frank, Tenenbaum, & Fernald, in press). On this kind of view, a first utterance establishes the topic (in the cases we examine, often a simple object referent), and then future utterances contribute new information (Clark, 1996). For such a learner, zero anaphora might not be as problematic if the discourse topic were already known.

The current study examined zero anaphora in Japanese from this perspective. We conducted a reanalysis and annotation of Japanese infant- and child-directed speech from the Fernald and Morikawa (1993) study, focusing on anaphora. We asked when zero anaphora was used within topical discourses, in comparison with object naming and the use of other pronouns. We found that although overall rates of anaphora were much higher for Japanese mothers compared

with English mothers, zero anaphora was primarily used when the discourse topic is already established. Also, we found a trade-off between zero anaphora and pronominal anaphora that caused pronouns to be more likely when the discourse topic was newly established. This sensitivity of zero and pronominal anaphora to the discourse topic suggests that a discourse-finding strategy may be even more important for Japanese-learning children than it is for English learners.

## Methods

### Corpus Materials

Our data consisted of a set of transcribed videos of object-centered play between mothers and children in their homes, from a study by Fernald and Morikawa (1993). While the original corpus contained both American and Japanese mother-child pairs, the current analysis focuses primarily on the Japanese mothers (American data was analyzed in Frank et al., in press). Discourse from 29 Japanese mother-child pairs with audio and video data was analyzed. The infants were divided into three age groups: 5-6 months (N=9, 5 males), 11-14 months (N=10, 5 males), and 18-21 months (N=10, 5 males).

Prior to recording, mother and child played comfortably together with the child's toys. Next, the child's own toys were removed and the video recording began. During the video the mothers were asked to play with the child using three standardized pairs of toys: dog and pig, car and truck, and brush and box. The mother was asked to play with the toys "as she normally would." The toys were introduced one pair at a time and removed before introduction of the following pair. The ordering of whether the dog and pig were introduced first or the car and truck were introduced first was counterbalanced across trials, but the brush and box were always introduced last (and only for the older two groups).

Towards the end of the play session, mothers of children from the two older age groups were asked to hide the toys and get the child to retrieve them using words alone. Because this scenario might affect how the mother referred to the objects (indeed it was inserted in order to elicit object names), we only considered utterances prior to this "hiding game." We also excluded utterances with sound and audio issues (167), and those spoken by the mother to the experimenter. In total, 8852 utterances taken from 6 hours and 51 minutes of video were analyzed in the current study.

### Conventions for Annotating Object Reference

A native Japanese speaker first divided the mothers' speech into "utterances," or segments of speech separated by pauses, on the basis of prosodic and syntactic cues. Most utterances ranged from a single word to a complete sentence; complete sentences were usually not counted as multiple utterances unless there was a pause or interruption of the speaker's turn.

Next, using the video and transcript data, a native Japanese speaker annotated, for each utterance spoken by the mother

Table 1: Descriptive statistics for each file in the FM Corpus. Utts = utterances, Length = length in minutes and seconds.

Age Grp	Code	Gend	Age	Utts	Length
6mos	J29	M	6	208	10:59
	J30	M	5	158	10:25
	J31	F	6	299	12:51
	J32	M	6	203	12:57
	J33	F	5	342	11:28
	J34	M	5	322	11:51
	J35	F	5	334	13:17
	J36	M	6	127	10:51
	J37	F	5	289	11:33
12mos	J2	M	14	325	19:50
	J3	F	13	346	15:53
	J7	F	11	594	19:36
	J8	M	11	364	16:35
	J11	M	12	285	15:18
	J13B	M	12	269	11:49
	J18	M	13	331	17:40
	J20	F	13	241	14:43
	J23	F	11	384	12:51
18mos	J26	F	12	280	12:40
	J4	F	18	354	18:11
	J5	F	18	186	10:50
	J9	M	18	268	16:29
	J10	F	20	297	13:51
	J12B	F	18	385	15:54
	J16	M	21	330	12:47
	J19	M	21	410	15:02
	J24	F	19	295	15:50
	J27	M	19	325	14:47
	J28	M	18	301	14:36

to the infant, what object or objects (if any) were being referred to. An object was considered to be referred to if A) the mother said the name of the object or B) the mother used a pronoun that the annotator judged to refer to the object. Although other toys and objects were occasionally referred to in the corpus, in the following analysis we will only examine references made to the six toys that were standardized across participating dyads (dog, pig, car, truck, brush, and box).

In Japanese, baby words for toys are often derived from onomatopoeia. We counted misnomers and onomatopoeia as references to the toy under clearly referential circumstances. For misnomers, this was when e.g. "moomoo" ("moo-moo," meaning "cow") was used in reference to the pig. We also counted alternative labels, such as "omocha" ("toy") and "nuigurumi" ("stuffed animal"), as references. However, we did not count misnomers and alternative labels as cases of "object naming," which we defined in a more restricted sense, described below. We counted onomatopoeia-derived noun phrases as references to objects, but not onomatopoeia that was used to describe sounds or actions. Our annotator made judgments about when the mother was using phrases such as "wanwan" as a noun and when she was using them to indicate sounds or actions that the toy was making. Thus, "wan-chan" ("Mr. woof") and, in some cases, "wanwan" ("woof-woof"), were coded as referring to nouns.

In Japanese, objects are frequently referred to without use of an explicit noun or pronoun. As noted above, grammat-



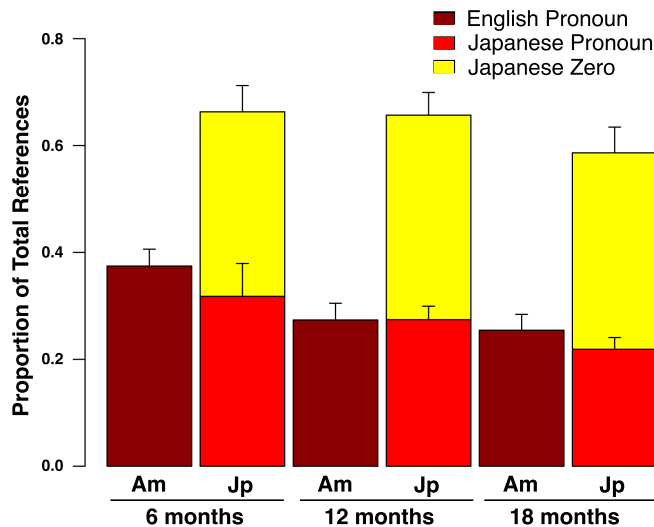


Figure 1: Proportion of total references that used pronominal or zero anaphora for American (Am) and Japanese (Jp) dyads. Error bars show standard error of the mean across dyads.

ical subjects and objects are frequently omitted in Japanese, where in English they would be marked using a pronoun. We included implicit subjects and objects as intended referents, but gave them a special marking. Whenever the subject or object of a verb or the noun modified by an adjectival phrase was omitted, we counted it as a case of implicit reference to the omitted noun, or “zero anaphora.” We did not count objects as being referred to if they were a missing instrument of action, the possessor of an object that was explicitly mentioned, or if they had a spatial or some other relationship to the objects that were mentioned. While each of these might sometimes qualify as zero anaphora, our annotation scheme treated them conservatively (hence, zero anaphora counts would if anything be higher under a revised scheme).

We further categorized non-zero references to objects into two types: those that used a referential pronoun and those that named the object. We counted a reference as “naming” the object when: 1. the utterance was marked as referring to the object by our annotator and not marked as being a case of zero anaphora, 2. the utterance contained one of a list of character strings found in words for the referred object (including common misspellings), and 3. the English gloss (created by a second native speaker for the original Fernald & Morikawa, 1993 study) contained one of a list of possible glosses for words for the referred object. We counted a pronoun as being used referentially when: 1. the utterance was marked as referring to the object by our annotator and not marked as being a case of zero anaphora, 2. the utterance contained one of a list of pronouns, and 3. the reference had not already been marked as object naming.

## Results

Mothers talked about each toy in alternating bouts of utterances. They frequently used onomatopoeia, and engaged in “social routines” (Fernald & Morikawa, 1993), such as re-

questing objects from the child and saying thank you. A plot of references to the six toys over a single video is presented in Figure 2, as a representative discourse structure.

We report three main sets of analyses. First, simple univariate counts of pronominal and zero anaphora. Second, analyses of transitions between different kinds of reference. Third, changes in use of zero anaphora across the corpus.

## Anaphora Use Across Languages

Our first analysis counted pronominal and zero anaphora proportions in the corpus at each age. We normalized these counts by the total number of object references that were identified (including all anaphoric references)<sup>1</sup>. These data are shown in Figure 1, along with English data on pronoun use in object references from Rohde and Frank (under review).

Several trends are apparent in these data. First, pronoun use is approximately equivalent between English and Japanese speakers. Speakers of both languages use pronouns approximately one-third of the time. Second, Japanese use of zero anaphora constitutes an additional third of *all* object references. Unlike the English-speaking mothers, Japanese-speaking mothers were using anaphora more often than not to refer to the toys that they were playing with. Finally, pronoun use declined with the age of the children in our sample. This trend was somewhat modest compared with the large cross-linguistic differences, but was nevertheless significant in a simple linear regression predicting pronoun use by age in months ( $\beta = -.005$ ,  $p < .001$ ). However, this age difference likely reflects the fact that videos were shorter for the 6 month old group than the other two groups. As will be discussed below, pronoun use declined substantially over the discourse for all three age groups, and so proportion of pronoun use in the 6 month old group might have been inflated by the shorter video—and hence shorter discourse—length.

## Discourse Continuity in Japanese

We next examined how likely each mother was to refer to the same object in two consecutive utterances. This analysis was used by Frank et al. (in press) as a first-pass indicator that discourse references to objects were relatively continuous. A high probability of repeated reference to an object suggests that a learner who “smoothed” their guesses about reference across time would be relatively successful. If they didn’t know what a particular utterance was referring to, they could just guess that the referent was the same as in the previous utterances in the discourse.

**Continuity of Reference** We calculated transition probabilities between referential and non-referential utterances of

<sup>1</sup>In the Japanese data, 281 out of 3642 references (7.7%) were marked as “ambiguous” as to whether the toy was referred to or not. Only non-ambiguous cases were used in the first analysis.

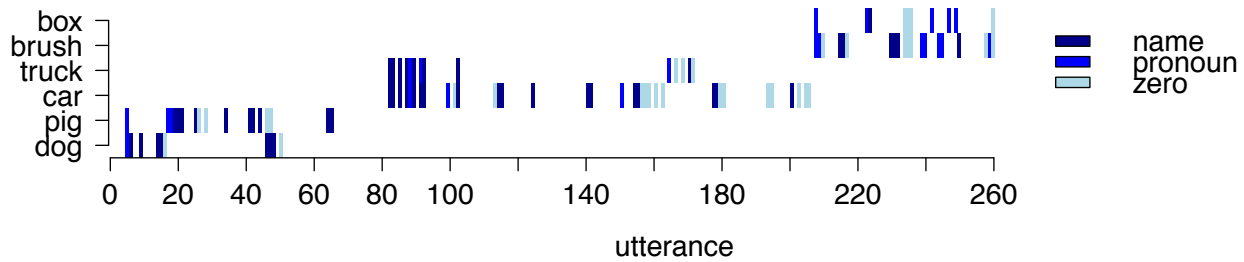


Figure 2: References to the six main toys in a sample video. Toys are plotted on the Y axis and blue lines signal that the toy was referred to in a particular utterance. name = object naming, zero = zero anaphora.

Table 2: Transition probabilities between referential and non-referential utterances of various types.  $P(Y|X)$  refers to the probability that in the current utterance an object is referred to using Y given that in the previous utterance it was referred to using X. zero = zero anaphora, name = object naming, nonref = nonreferential utterance.

X	P(name X)	P(zero X)	P(pronoun X)	P(other X)	P(nonref X)
name	.25	.12	.06	.02	.55
zero	.04	.28	.05	.01	.62
pronoun	.10	.15	.19	.03	.53
other	.05	.06	.11	.08	.70

various types.<sup>2</sup> As in our previous work, transition probabilities were first calculated for each of the six toys, then averaged together, weighted by the number of times each toy was referred to in the video (or referred to using a particular reference type, as appropriate).

Mothers varied considerably in how likely they were to refer to an object in two consecutive utterances. The probability of referring to an object in the next utterance given it was referred to in the current utterance ranged between .20 and .57 for each mother (mean = .43). Probability of repeated reference remained relatively stable when calculated for the utterances in each age group, although it tended to rise with the age of the child, from .39 at 6 mos, to .44 at 12 mos, and .47 at 18 mos. Overall these levels were somewhat lower than those for English-speaking dyads.

When we removed utterances that consisted of backchanneling—checking for a response from the child—such as “nn,” “n?,” “a,” and “hai hai hai” (“yes, yes, yes”), this slightly increased the probability of consecutive reference (mean = .49), and decreased variance between mothers. This analysis suggests that the higher tendency of Japanese mothers to use backchanneling, combined with the lower overall frequency of object reference, may account for the difference between Japanese and English-speaking dyads.

**Transitions Between Reference Types** We next examined transitions between referential and non-referential utterances of various types, including zero anaphora, pronouns, and object naming. The goal of this analysis was to understand the directionality of zero anaphora use in discourse. If zero

anaphora is used more after naming, then transition probabilities should be asymmetric:  $p(\text{zero}|\text{name})$  should be higher than  $p(\text{name}|\text{zero})$ .

Results are summarized in Table 2. The largest trend was for referential utterances to be followed by non-referential utterances, indicating (as above) that Japanese discourses contained more non-referential speech overall than in English. Nevertheless, there were still distinct trends in which referential strategies were used earlier in discourses. Zero anaphora after object naming was three times as likely as object naming after zero anaphora. Zero anaphora after pronominal anaphora was three times as likely as pronominal anaphora after zero anaphora. Surprisingly, pronouns were more likely to be used before object naming than after it, unlike in English (Ariel, 1990; Gundel, Hedberg, & Zacharski, 1993).

## Changes in Anaphora Use Across the Discourse

**Zero Anaphora Over Time** In our third analysis, we examined how often mothers used zero anaphora at different points in the discourse. For each utterance that referred to one of the six main toys, we calculated the number of times the object had been mentioned prior to that utterance. We call this the “number of previous references” (NPR) for the utterance. If an utterance refers to the dog for the fifth time in a video, that utterance has an NPR of 4. Next, we collapsed all references with a particular NPR across mothers and toy referents, and calculated the proportion of the time the references used zero anaphora.

The proportion of time that mothers referred to an object using zero anaphora increased with the log of the NPR at all ages (Figure 3). We created a mixed effects model of zero anaphora usage, with age group and NPR as fixed effects and mother as a random effect with random intercept and slope with respect to NPR. Because of the very large number of observations, we used the  $z$  approximation to esti-

<sup>2</sup>We calculated transition probabilities between utterances both counting and discounting cases where the referent was ambiguous. However, none of the transition probabilities besides those for the “other” category changed by more than .02 when the ambiguous cases were counted as non-reference, and so we count them as non-reference in the current analysis.

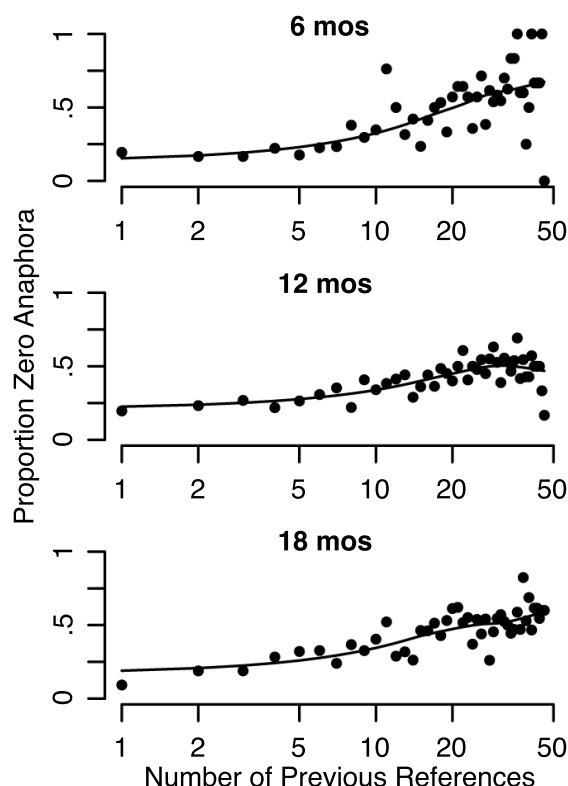


Figure 3: Proportion of total references that use zero anaphora as a function of the number of previous references to the object (NPR), plotted separately by age group. NPR depicted on a logarithmic scale, NPRs with fewer than 10 data points were dropped. A loess curve was fitted to each plot.

mate significance for individual coefficients. We found that zero anaphora use increased significantly with NPR ( $\beta = .049, p < .001$ ), while age group had no effect ( $\beta = -.001, ns$ ). We modified the model to include the interaction term of age group and NPR as a fixed effect, and found that the resulting coefficient was nonsignificant ( $\beta = -.002, ns$ ).

We also examined whether mothers were more likely to use zero anaphora when referring to some toys than others. We created a mixed effects model with age group, NPR and toy as fixed effects and mother as a random effect with random slope and intercept with respect to NPR. There was no significant effect of any of the toys with the exception of the box, which had significantly higher zero anaphora use compared to all other toys ( $\beta$  values ranged between  $-.79$  and  $-1.08, p < .001$  for each of the other toy coefficients when box was used as the referent). This is likely due to the fact that mothers would repeatedly ask their children to “open” and “close” the box without specifying a direct object. We find this an intriguing hint that some objects may be referred to via associated verbs, especially when they provide salient action affordances.

**Zero Anaphora vs. Other Reference** Our final analyses examined the proportion of zero anaphora use as compared to the proportion of pronominal anaphora and object naming (Figure 4). As the log of the number of previous references increases, zero anaphora use increases substantially

and pronominal anaphora decreases substantially. However, object naming only decreased slightly at higher NPRs. For the first 20 times the mother referred to an object, the proportion of the time that she named the object stayed relatively flat. In other words, zero anaphora increases over time at the expense of pronominal anaphora, rather than object naming.

We also extracted discourses of consecutive references to the same object, and examined zero anaphora, pronominal anaphora and object naming at each “utterance position” in the discourse. The first utterance of a discourse has an utterance position of 1, the second has an utterance position of 2, etc. We examined a total of 791 discourses of three or more consecutive references to an object. Once again, the proportion of zero anaphora use increased with the log of the utterance position, and the proportion of pronominal anaphora decreased. For the first six utterance positions, the sum of the proportion of zero anaphora use and pronominal anaphora use was flat. (Over 90% percent of the discourses we examined were six or fewer utterances long). This analysis again indicates that the proportion of zero anaphora rises at the expense of pronominal anaphora before having any effect on the proportion of references that name the object.

## General Discussion

We examined patterns of anaphora use in child-directed speech by Japanese mothers of infants aged 6, 12 and 18 months. We found that Japanese mothers used far more total anaphora than English-speaking mothers, although pronominal reference was about equally likely in Japanese and English. We examined the transition probabilities between zero anaphora, referential pronouns, and object naming, finding that pronouns are more likely to occur earlier in the discourse, while zero anaphora is more likely to follow both pronouns and object naming. Finally, we assessed how object naming and anaphora use evolve over the discourse, using two measures: the number of previous references to an object and the utterance position in discourses of consecutive reference. Zero anaphora use rapidly accelerated over the discourse at the expense of pronouns, while object naming persisted at a steady rate that only gradually declined later on.

These findings give insight into how object reference varies between languages like English, which have a two-tiered system of nominal and pronominal reference, and those like Japanese that have a third possibility: zero anaphora. In English, both pronoun use and elision are more likely when the referent is given information (MacWhinney & Bates, 1978). From this one might expect that both pronominal and zero anaphora use would increase with the givenness of the referent in Japanese.

However, the distribution of pronoun types differs drastically between these languages: both personal and demonstrative pronouns are frequent in English, while in Japanese, most pronouns are demonstrative, and zero anaphora is used in place of personal pronouns most of the time. According to Gundel’s hierarchy of givenness (Gundel et al., 1993), per-

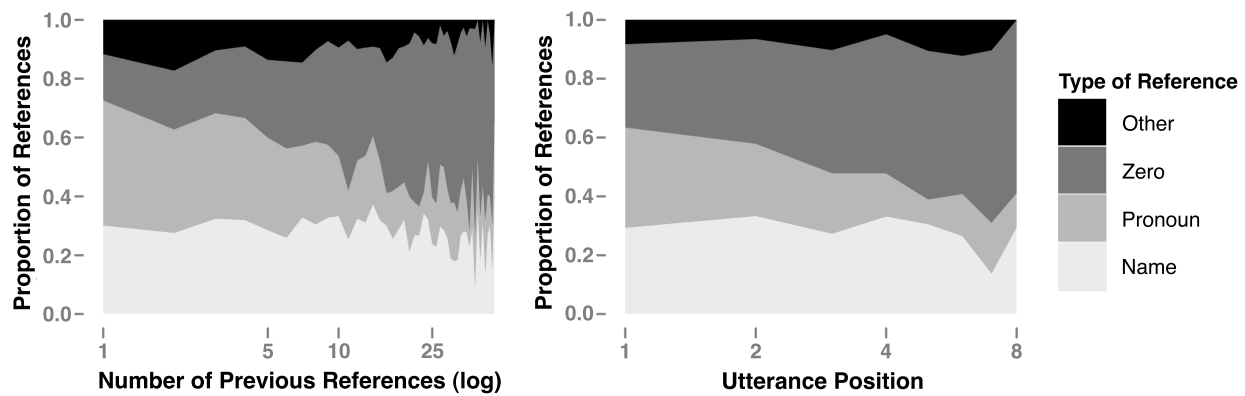


Figure 4: Proportion of total references that involve object naming, zero anaphora, or referential pronouns, as a function of the number of previous references to the object (NPR) and utterance position. NPR and utterance position depicted on a logarithmic scale, NPRs and utterance positions with fewer than 10 data points were dropped. Utterance position calculated using all discourses of three or more consecutive references to an object.

sonal pronouns and zero anaphora are used only when the referent is in focus, but demonstratives can also be used when it is out of focus but in working memory. By Grice's maxims, this leads to demonstrative pronouns being used more frequently when there is a topic shift, or when the referent has been introduced non-linguistically (Gundel et al., 1993; Gundel, Hedberg, & Zacharski, 2004). In our data, Japanese pronouns were more likely to be used in reference to an object the first time it was referred to than at any other point in the discourse. This suggests that in Japanese child directed speech pronoun use may signal topic change (e.g. "look at this!"), rather than topic continuity.

As suggested originally by Fernald and Morikawa (1993), pure word-object mapping in Japanese might be a very hard problem: Only about a third of references name the object, while the other two-thirds make use of pronouns and zero anaphora. Moreover, references to objects in Japanese (as in English) are not evenly distributed in discourse. But if children have some sense of what the current topic of discourse is—what is given, and what is new—this problem might be somewhat alleviated. A Japanese-learning infant could infer the topic of conversation and then assume that future comments, whether using names or anaphora, referred to that topic. Thus, this study underscores the importance of topical discourse in early word learning, suggesting that tracking the topic of conversation across utterances may be even more crucial to word learning success in pro-drop languages.

### Acknowledgments

Special thanks to Akiko Knott for her invaluable annotation work, and to Stephan Meylan for advice on using the mechanical turk interface to rekey our transcripts.

### References

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London: Routledge.  
 Blythe, R., Smith, K., & Smith, A. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, 34(4), 620–642.

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.  
 Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development*, 64(3), 637–656.  
 Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5), 578.  
 Frank, M., Tenenbaum, J., & Fernald, A. (in press). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language, Learning, and Development*.  
 Gundel, J., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69, 274–307.  
 Gundel, J., Hedberg, N., & Zacharski, R. (2004, September). Demonstrative pronouns in natural discourse. In *Proceedings of daarc-2004*. Sao Miguel, Portugal.  
 MacWhinney, B., & Bates, E. (1978). Sentential devices for conveying givenness and newness: A cross-cultural developmental study. *Journal of Verbal Learning and Verbal Behavior*, 17, 539–58.  
 Rispoli, M. (1995). Missing arguments and the acquisition of predicate meanings. *Beyond names for things: Young childrens acquisition of verbs*, 331–352.  
 Rohde, H., & Frank, M. (under review). Markers of topical discourse in child-directed speech.  
 Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2), 39–91.  
 Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558–1568.  
 Yu, C., & Ballard, D. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15), 2149–2165.  
 Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5), 414.

# Sources of uncertainty in intuitive physics

Kevin A Smith (k2smith@ucsd.edu) and Edward Vul (evul@ucsd.edu)

University of California, San Diego  
Department of Psychology, 9500 Gilman Dr.  
La Jolla, CA 92093 USA

## Abstract

Recent work suggests that people predict how objects interact in a manner consistent with Newtonian physics, but with additional uncertainty. However, the sources of uncertainty have not been examined. Here we measure perceptual noise in initial conditions and stochasticity in the physical model used to make predictions. Participants predicted the trajectory of a moving object through occluded motion and bounces, and we compared their behavior to an ideal observer model. We found that human judgments cannot be captured by simple heuristics, and must incorporate noisy dynamics. Moreover, these judgments are biased in a way consistent with a prior expectation on object destinations, suggesting that people use simple expectations about outcomes to compensate for uncertainty about their physical models.

**Keywords:** intuitive physics, stochastic simulation, uncertainty, probabilistic inference

## Introduction

Predicting how the world will unfold is key to our survival and ability to function on a daily basis. When we throw a ball, cross a busy street, or catch a pen about to fall off of a desk, we must foresee the future physical state of the world to plan our actions. The cognitive mechanisms that help us make these predictions have been termed ‘intuitive physics’ models.

Although human performance in physical prediction tasks tends to approximate real-world (Newtonian) physics, it does not match exactly: people make systematic prediction errors. While this has been taken as evidence that human models of intuitive physics are non-Newtonian (e.g., McCloskey, 1983), more recently human behavior has been explained by intuitive Newtonian physics models under uncertainty. On this account, human predictions deviate from Newtonian mechanics because of stochastic error – uncertainty about the initial positions or velocity of objects propagates through the non-linear physical model and causes variability and bias in final judgments. For instance, human predictions about the stability of a tower of blocks or the most likely direction for that tower to fall are consistent with a purely Newtonian model of physics with a small amount of uncertainty in the initial positions of the constituent blocks (Hamrick, Battaglia, & Tenenbaum, 2011). Similar models of physics with perceptual noise have been used to explain relative mass judgments in collisions (Sanborn, Mansinghka, & Griffiths, 2009) and infants’ expectations for object movement (Téglás et al., 2011).

There are numerous ways in which uncertainty can be introduced into intuitive physical reasoning. We broadly classify these into two categories: perceptual uncertainty and uncertainty about dynamics. Perceptual uncertainty

arises because initial measurements of the location and velocity of objects is imperfect; this initial noise propagates through the model. Uncertainty about dynamics reflects noise in the physical model itself. Real object movement and collisions are perfectly deterministic only in an idealized system; in the world, objects can deviate from their ideal path because of multiple, unknowable interactions with the environment (e.g., a ball rolling across gravel will not move in a straight line). Stochastic dynamics could thus reflect such environmental uncertainty.

Our goal is to disentangle the influence of initial noisy percepts and noisy physics on human predictions of object dynamics. We compared human behavior in a simple physical prediction task to a stochastic physics model with parameters reflecting the different types of uncertainty.

## Stochastic Physics Model

We designed a model to replicate stochastic physics in a simple environment: a ball bouncing around a two-dimensional box. We based this model on idealized mechanics, but also incorporated the two sources of uncertainty: we added noise to the initial position and velocity to capture perceptual uncertainty, while dynamic uncertainty was captured by jitter in object movement over time, and variability in bounce angles.

## Uncertainty Parameters

The model was based on a simple two-dimensional physics engine customized to add our sources of uncertainty. As physical uncertainty goes to zero, this model reduces to laws from idealized mechanics: the ball would continue to move in a straight line at a constant velocity until it hit a wall, at which point it would bounce elastically and with angle of incidence equal to the angle of exit. Uncertainty was captured using four parameters, two for the perceptual error, and two for the stochastic error:

**Perceptual Uncertainty** At the start of the simulation, the ball’s position and velocity were based on where the ball would be in a perfectly deterministic simulation, but with noise added. Position was perturbed by isotropic two-dimensional Gaussian noise parameterized by standard deviation,  $\sigma_p$ . Noise in velocity direction was captured in a von Mises (circular normal) distribution on direction of motion, parameterized by concentration (inverse variance)  $\kappa_v$ . We did not consider uncertainty in the speed of the ball, as this would only affect the timing of the ball’s movement but not its destination, which is the prediction we aim to capture.

**Dynamic Uncertainty** Noise was added during the simulation in two ways. First, at each time step (1000/sec), the direction of the ball was ‘jittered’ by adjusting its direction using a von Mises distribution with the concentration parameter  $\kappa_m$ . In addition, noise was added during each bounce by assuming that the angle the ball bounced off of the wall was defined by a von Mises distribution centered on the angle of incidence with a concentration parameter  $\kappa_b$ .

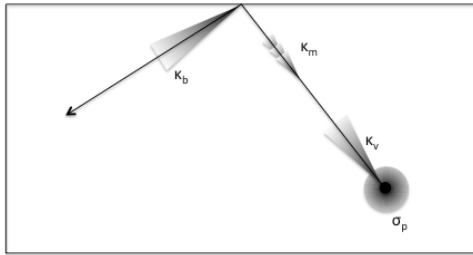


Figure 1: Sources of uncertainty in the stochastic physics model

## Experiment

We aimed to test model predictions against human data and to estimate uncertainty parameters in intuitive dynamics. In this experiment, subjects predicted the trajectory of a ball in a two-dimensional environment on a computer screen. This was done in a ‘Pong’ game where participants tried to catch the ball with a paddle. Crucially, we occluded the latter part of the ball’s movement, so that successful prediction of the final position required the mental simulation of the object trajectory. We could estimate the final position predicted by our stochastic physics model with different parameters, and thus compare human behavior to model predictions under varying types and degrees of uncertainty.

In this experiment we parametrically varied both the distance the ball would travel<sup>1</sup> and the number of bounces off of walls while occluded. If intuitive dynamics models are deterministic, then the number of bounces will have no effect on human predictions. The distance manipulation was designed to tease apart the contributions of perceptual uncertainty about velocity and dynamic velocity noise.

## Methods

52 UCSD undergraduates (with normal or corrected vision) participated in the experiment for course credit.

Subjects used a computer mouse to control the vertical position of an on-screen ‘paddle’ to catch a moving ball. The ball moved according to the deterministic physics underlying the stochastic physics model. Both the paddle and the ball were confined to a 1200 by 900 pixel area in the center of the screen. Each trial began with a display of only the paddle, which subjects could move up and down. The

paddle was 100 pixels in height and was centered on the vertical position of the mouse before each trial. A mouse click triggered the start of a trial. A ball would then appear on the screen, moving at a constant velocity of 600 pixels/second. After the ball moved 400 pixels, a grey rectangle would occlude the portion of the screen containing the ball (Figure 2). During this period, the ball would continue to move, bouncing perfectly elastically off of the edges of the field, but would not be visible. Once the subjects caught the ball with the paddle, or the ball broke the plane of the paddle, the trial would end and the occluder would be removed, showing whether (and by how far) the subject missed the ball. Upon clicking the mouse, the screen would clear and reset for the next trial. The number of balls caught by the subject was always displayed in the upper right corner as a motivation to perform well.

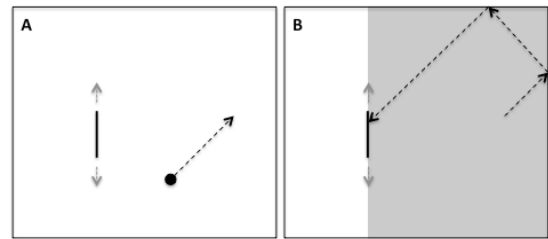


Figure 2: Diagram of a trial. (A) The ball moves unoccluded in a straight line. (B) Once the field is occluded, the ball continues to move until caught or it passes the paddle plane.

Subjects were given 648 trials throughout the experiment. These 648 trials were identical for all subjects, but presented in a randomized order. Each trial had a particular ball trajectory, generated by one of nine conditions. The nine trajectory conditions crossed the distance the ball travelled while occluded (600, 800, or 1000 pixels) with the number of bounces (0, 1, or 2); there were 72 trials of each condition. The specific path for each trial was generated prior to the experiment subject to the constraints of the condition and the constraint that the final position was not in the top 20% or bottom 20% of the enclosed area to avoid bias due to positioning the paddle at the ends of the screen.

Before starting the experiment, subjects were given seven trials without the occluder to demonstrate how the ball would move, then six practice trials with the occluder.

For each trial, we recorded the position of the midpoint of the paddle once the ball was caught or moved past the paddle. From this measure we could calculate, for each trial, (a) the average expected position of the ball, and (b) the variance of predictions around that expectation.

## Subject Performance

**Accuracy** Subjects caught the ball on 43.8% of all trials. Individual subject accuracies varied between 25.6% and 63.7% (chance was 11%). Accuracy also varied by trial condition: subjects were most accurate in the shortest, no

<sup>1</sup> Because the ball always moved at a constant velocity, distance was proportional to duration of occlusion.

bounce condition (69%) and least accurate in the longest, two-bounce condition (32%).

Accuracy improved slightly over time, increasing from 42.7% in the first half of trials to 44.9% on the second half ( $\chi^2(1) = 15.9$ ,  $p < 0.001$ ). However, because this was a small effect, and because in a logistic model predicting accuracy, trial order did not interact with either distance ( $\chi^2(2) = 0.72$ ,  $p = 0.70$ ) or number of bounces ( $\chi^2(2) = 4.18$ ,  $p = 0.12$ ), we do not try to account for this change.

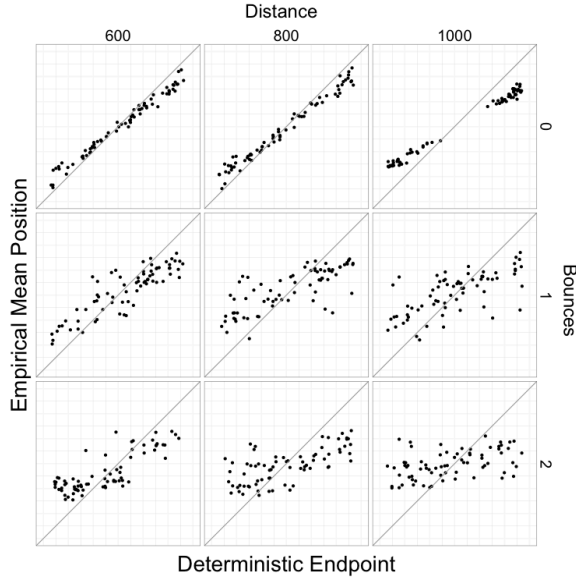


Figure 3: Mean predicted paddle position versus path endpoint using deterministic physics as a function of trial condition. Each point represents a separate trial.

**Expected Positions** In addition to decreasing accuracy, subjects also showed increasing bias in average predictions as the distance or number of bounces increased. The mean final position of the paddle for each trial shifted towards the center as compared to the final ball position (see Figure 3).<sup>2</sup> The magnitude of this bias toward the center of the screen increased as either distance or number of bounces increased.

Table 1: Percent of distance ‘shifted’ from actual end ball position towards center by trial condition

		Distance		
		600	800	1000
Bounce	0	24%	44%	53%
	1	23%	60%	70%
	2	41%	63%	84%

<sup>2</sup> There was low lag-one autocorrelation between the position of responses and prior responses (0.11) that did not vary by condition, suggesting that this mean shifting was not driven by subjects leaving the paddle in the same position as their prior trial.

**Variance of Responses** The variability of subjects’ responses around the mean also increased with distance and bounces, but only up to a ceiling - well below the maximum possible spread - once subjects had to take into account even one bounce.

Table 2: Average standard deviation (in pixels) of responses within a trial by condition

		Distance		
		600	800	1000
Bounce	0	65	76	94
	1	111	115	114
	2	115	111	121

## Model Application

The coarse results suggest that prediction error and variability increases with distance or number of bounces. But they do not indicate which sources of uncertainty contribute to intuitive physics predictions, nor do they explain why some trials within the same condition produce greater bias and variability than others.

We aimed to tease these factors via our model of stochastic physics. By finding the set of uncertainty parameters that best fits the empirical data, we can compare the relative contribution of the perceptual uncertainty parameters to the dynamic uncertainty parameters. A good model should capture trial-level differences in subjects’ performance, and explain trial difficulty based on the interplay of different sources of uncertainty.

## Simulation

We replicated the experimental task in the stochastic physics model, simulating the same 648 trials. To mirror this task, each simulation started at the point of occlusion (when subjects could no longer visually track the ball and must predict its path) and ended when the simulated ball crossed the plane of the paddle. On each simulation, we measured the position of the simulated ball along that plane. Because there is no analytic form of the probability distribution over possible trajectories, we simulated each trial 500 times, thus estimating the predictive distribution for each trial via sampling.

No set of uncertainty parameters produced mean estimates of the final position of the ball that were systematically shifted toward the center like the empirical data; as long as Newtonian physics underlies the model, it will in general simulate trajectories that are centered around the actual endpoint, regardless of the uncertainty parameters chosen. Since the magnitude of the center bias scaled with distance and number of bounces, we suspected that subjects were incorporating a prior on final position, producing a center bias proportional to the uncertainty in their physics-



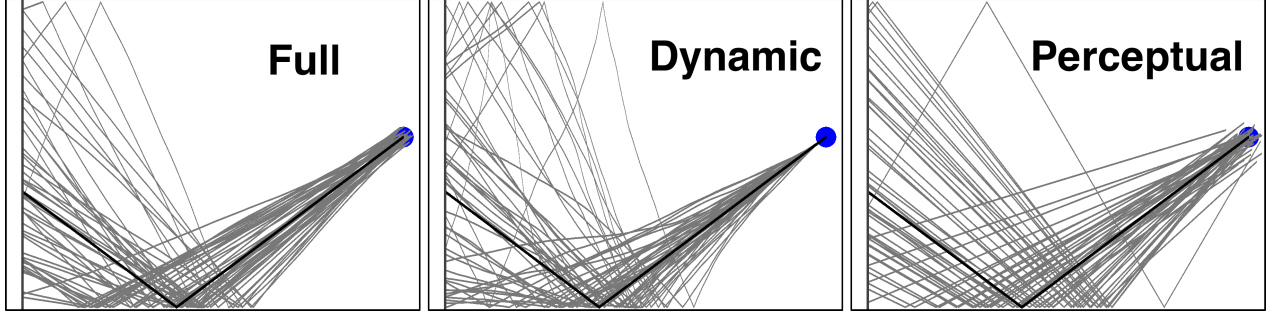


Figure 4: Sample simulation paths for one trial with each model. The grey lines represent individual simulations, the black line represents deterministic simulation. There is no initial uncertainty in the dynamic model, but it builds quickly over time, resulting in wavy paths. The initial position and velocity vary significantly in the perceptual model, but once started, the simulation unfolds deterministically. The full model uses both types of uncertainty and so has more certainty in starting positions than the perceptual model and straighter paths than the dynamic model.

based predictions.<sup>3</sup> People therefore appear to incorporate prior expectations with their intuitive physics models.

We treated this bias as a simple Gaussian prior on the final ball position centered on the middle of the screen, with standard deviation as a free parameter ( $\sigma_{\text{prior}}$ ). One value of this parameter was used for all trials and conditions.

The final distribution of predictions for each trial was calculated by combining the center-prior with the distribution of predicted positions simulated by the stochastic physics engine. We treated the distribution of predicted positions as a Gaussian and calculated their mean and standard deviation. We could then calculate the mean and standard deviation of the posterior distribution using Bayesian cue combination (e.g., Ernst & Banks, 2002):

$$\sigma_{\text{post}}^2 = \left( \frac{1}{\sigma_{\text{prior}}^2} + \frac{1}{\sigma_{\text{sim}}^2} \right)^{-1}, \mu_{\text{post}} = \left( \frac{x_{\text{center}}}{\sigma_{\text{prior}}^2} + \frac{\mu_{\text{sim}}}{\sigma_{\text{sim}}^2} \right) * \sigma_{\text{post}}^2$$

Using these equations, trials with greater simulation variance will be more affected by the prior, and will shift further towards the screen center. Thus, the model can account for the center-bias in a manner sensitive to prediction uncertainty.

We found the maximum likelihood parameters to fit three quarters of the data (with an equal number of trials from each of the distance by bounce conditions).<sup>4</sup> We also fit two other models: one with only perceptual uncertainty and prior parameters, and one with only dynamic uncertainty and prior parameters. We compared these models based on the likelihood of the 25% of the remaining (cross-validation) data.

## Model Results

**Model Comparison** The stochastic physics model was designed to tease apart how various sources of uncertainty

contribute to intuitive physics. Thus we compared the model with both dynamic and perceptual uncertainty to the two nested models with either dynamic or perceptual uncertainty parameters alone to determine which sets of parameters were necessary to best explain the data.

In addition, we tested how well any of the stochastic models captures human behavior by comparing them to a simple regression model with different parameters for each condition. The regression model assumes that people will provide the correct answer, plus some error that varies by condition without regard to individual trial details. We assume that the average reported position will have some variance (estimated independently for each condition), and some bias towards the center (estimated by regressing the average reported position against the deterministic end positions within each condition) – in other words, the regression model is a non-physical error model. This model can capture the gross ‘shift’ in expected position that was observed in the data in each condition (see Figure 3), but does not treat the shift as an inference done independently on each trial. The spread in responses was assumed to be constant within each condition, and was set at the average empirical standard deviation from that condition. Like the stochastic models, this model was fit on three-quarters of the trials and tested on the remaining data.

Table 3: Model prediction of left-out data

Model	$\Delta\text{LLH}$
Full	2,588
Dynamic	2,568
Perceptual	2,197
Regression	2,326
Oracle	3,259

Table 3 shows cross-validation likelihood for the four models. All log-likelihoods are shown as improvement over a baseline assuming that all data came from a single Gaussian. In addition, we included an ‘oracle’ model that knows the mean and standard deviation of responses for each trial – this serves as the plausible upper limit on how well different models might do. The full stochastic model

<sup>3</sup> The actual endpoint of the ball was uniformly distributed within the space of allowable endpoints. Therefore, this prior is unlikely to have been learned from the experiment.

<sup>4</sup> Numerical optimization techniques can find local minima, so we used multiple starting points and grid search across 1,600 sets of parameters to ensure we were finding the global minimum.

does best, followed closely by a model including only dynamic noise. Both the perceptual noise model and the non-physical model perform worse by many orders of magnitude.

The dynamic model performed nearly as well as the full model for two reasons. First, the parameter representing error in the initial position ( $\sigma_p$ ) was set to a small value in the full model and explained very little of the variance in simulations. Second, much of the noise in initial velocity direction can be captured by increasing dynamic velocity noise, and so we cannot say whether any initial velocity noise is required. The model with only perceptual noise did quite poorly because subjects' performance changed with each additional bounce, and thus human performance cannot be captured without dynamic uncertainty.

**Trial-Level Simulations** Human predictions about individual trials within the same distance-by-bounce condition varied significantly: some had much larger variations in responses or greater shifts toward the center than others. These differences arose from trajectory characteristics other than total distance traveled or number of bounces. For instance, it is harder to predict the end position of a ball that bounces in a corner or balls that approach the paddle at a steep angle. If the stochastic physics model is capturing characteristics of intuitive physics, then it should capture this within-condition variability as well.

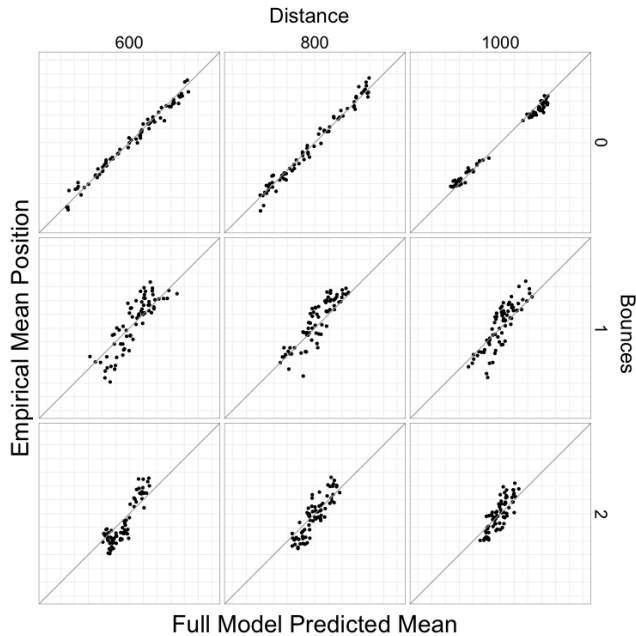


Figure 5: Full model vs. empirical mean position by condition. Each point is a separate trial.

The full stochastic model fit the variation in mean paddle position across trials well ( $r=0.93$ ), and slightly better than the predictions of the regression model ( $r=0.90$ ). However, the difference between models is highlighted when

considering individual conditions: although both models account for the mean position in the no-bounce conditions, only the full model continues to perform well as bounces and distance are added (see Table 4).

Table 4: Correlation between model and empirical by-trial means within condition

		Full			Regression		
		Distance			Distance		
		600	800	1000	600	800	1000
Bounce	0	.99	.99	.99	.99	.99	.99
	1	.86	.88	.85	.88	.77	.68
	2	.89	.87	.82	.82	.68	.45

The standard deviation of predictions from the full stochastic model was well correlated with the standard deviation of subjects' responses across trials ( $r=0.79$ , see Figure 6), albeit with a tendency to overestimate. Moreover, the stochastic physical model also captures the variability across trials within each distance-by-bounce condition (Table 5). Together, these results indicate that human uncertainty about final outcomes accumulates in a manner qualitatively similar to that predicted by a stochastic physical model.

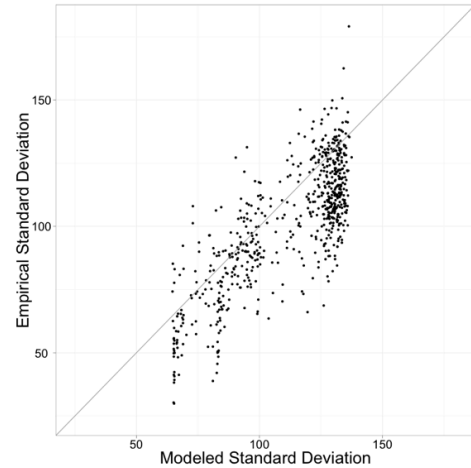


Figure 6: Full model vs. empirical standard deviation by trial

Table 5: Correlation between full model and empirical by-trial standard deviations within condition

		Distance		
		600	800	1000
Bounce	0	.54	.43	.17
	1	.53	.44	.30
	2	.14	.16	.17

In the experimental data, the amount of mean-shifting for each trial is related to the variance of the observations from that trial (Spearman's  $\rho = 0.30$ ), suggesting that people hedge their guesses towards the middle more as the amount

of uncertainty increases. A center-prior captures this behavior by causing more reliance on the prior when there is a wider distribution of model simulations. This has the effect of shifting guesses more towards the center when physical simulations are more uncertain. The stochastic physics model captures this phenomenon by predicting trial-level differences in uncertainty, and is thus better able to describe variation in human responses across trials than a constant mean-shift for each condition (see Figure 5).

## General Discussion

We found that human performance on a physical prediction task is captured by a model of stochastic physics with a prior expectation about the final position of objects. Furthermore, we found that bias and variability of human predictions are driven by uncertainty about the dynamics: people use stochastic, rather than deterministic, physics to make predictions. This result supports recent findings that people predict object dynamics using unbiased intuitive physics models (e.g., Hamrick, et al., 2011), and suggests two refinements to this view. First, the internal physics models themselves must be stochastic rather than rely solely on perceptual uncertainty to demonstrate non-determinism. Second, people do not directly use predictions from their physical models, but combine them with simple priors to produce rich behaviors.

Though we found that dynamic uncertainty contributes substantially to predictions in this task, we do not know how people might adjust this uncertainty based on task demands. In this experiment, the ball was easy to see (low perceptual uncertainty) and the background was uniform (suggesting less perturbation during movement). Lower contrast between object and background might cause greater perceptual uncertainty; likewise, backgrounds suggesting a rough surface might cause people to introduce more stochastic movement error into their simulations. An interesting direction for future work is to explore how people adjust the uncertainty within their intuitive physics models to account for different expectations about the world.

We also found that people modulate their physical predictions via prior expectations about the outcomes. Although these expectations could arise in many ways, here we were able to capture human behavior well by using a simple expectation about the final position: despite there being no evidence for it within the experiment, people believed that the ball was more likely to end up in the center of the screen. We made a simplifying assumption that this was a prior expectation about final location; it is possible that this is an approximation of other sorts of priors (e.g., objects tend to travel in a more horizontal direction). More research is required to understand exactly what these prior expectations are, how they develop, and under what conditions they become integrated into models of intuitive physics. Regardless of the prior used, we think that this might reflect a more general strategy that people may adopt to account for their uncertainty in their internal physical

model itself: by adjusting model predictions via a simple prior on outcomes, behavior will be more robust to errors in the simulation model. A similar process may suggest a means for combining model-based and model-free predictions (Gläscher, Daw, Dayan, & O'Doherty, 2010): learning simple expectations about the world is a good hedge against model error.

Our models predicted systematically larger variances than those we observed. This may be due to our simplistic choice of the shape of the prior. Gaussian cue combination of the prior and simulated distributions produces dependence between variance and mean-shift: a greater mean-shift arises only from greater variance. Thus to best fit the predicted means, using a Gaussian prior required a biased variance estimate. Further work is required to understand the priors people actually hold (e.g., Stocker & Simoncelli, 2006) to refine the models that people use to simulate the world.

This work supports the hypothesis that intuitive physics models can be built upon a Newtonian framework. Moreover, these models are not deterministic, but incorporate sources of dynamic uncertainty. Furthermore, people do not trust these models entirely, but combine their predictions with simple expectations about the outcome itself. Though just a first step, this provides a framework for disentangling and understanding the various components of intuitive physics models.

## Acknowledgments

This work was supported by BIAL Foundation grant to EV

## References

- Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429-433.
- Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*, 585-595.
- Hamrick, J., Battaglia, P., & Tenenbaum, J. (2011). *Internal physics models guide probabilistic judgments about object dynamics*. Paper presented at the Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, Boston, MA.
- McCloskey, M. (1983). Intuitive Physics. *Scientific American*, *248*(4), 122-130.
- Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2009). *A Bayesian framework for modeling intuitive dynamics*. Paper presented at the Proceedings of the 31st Annual Conference of the Cognitive Science Society, Amsterdam.
- Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*(4), 578-585.
- Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*, 1054-1059.

# Change detection under autocorrelation

Maarten Speekenbrink (m.speekenbrink@ucl.ac.uk), Matthew A. Twyman (m.twyman@ucl.ac.uk)

Nigel Harvey (n.harvey@ucl.ac.uk)

Cognitive, Perceptual and Brain Sciences, University College London  
Gower Street, London WC1E 6BT, England

## Abstract

Judgmental detection of changes in time series is an ubiquitous task. Previous research has shown that human observers are often relatively poor at detecting change, especially when the series are serially dependent (autocorrelated). We present two experiments in which participants were asked to judge the occurrence of changes in time series with varying levels of autocorrelation. Results show that autocorrelation increases the difficulty of discriminating change from no change, and that observers respond to this increased difficulty by biasing their decisions towards change. This results in increased false alarm rates, while leaving hit rates relatively intact. We present a rational (Bayesian) model of change detection and compare it to two heuristic models that ignore autocorrelation in the series. Participants appeared to rely on a simple heuristic, where they first visually match a change function to a series, and then determine whether the putative change exceeds the variability in the data.

**Keywords:** change detection; judgment; forecasting

## Introduction

Detecting changes in time series is a surprisingly ubiquitous task. Doctors and therapists monitor diagnostic indicators for signs of disease onset and for evidence that a prescribed treatment is effective; farmers monitor soil conditions to decide whether additional irrigation is necessary; local authorities monitor river levels for increased likelihood of flooding; probation officers monitor probationers' behaviour for evidence of return to crime; financiers monitor data, such as exchange rates, for signs of trend reversal. Many other examples could be given. As with forecasting and control tasks, monitoring tasks may be tackled by formal statistical methods, by using judgment alone, or by using some combination of these two approaches. The method most favoured depends to a large extent on the domain. Typically, implementation of and training in formal methods consume more resources (time, money, effort) but the investment may be worthwhile if those methods have considerable benefits over judgment in terms of accuracy. Thus, it would be useful to know just how good human judgment is relative to formal methods.

There are many formal statistical methods for detecting change in time series (e.g., Albert & Chib, 1993; Carlin, Gelfand, & Smith, 1992; Hamilton, 1990). This variety is partly because some approaches represent the event producing the regime change as deterministic whereas others represent it as a random variable and partly because, whichever of these approaches is adopted, there is still some debate about how best to estimate the likelihood that a change has occurred.

In contrast, there has been very little research into judgmental assessment of regime change. Originally, behavioural

psychologists working within the Skinnerian tradition used judgment (visual inference) to assess whether a manipulation changed some aspect of an animal's behaviour represented as a time series. They argued that this is a conservative approach because only large effects can be detected (e.g., Baer, 1977). Their claims were not directly tested. However, when behaviour analysts later used the same approach to assess human patients, there was concern that the shorter pre-treatment baselines in the series impaired visual inference. As a result, some experiments were carried out to investigate how accurately people can detect change.

## Judgmental change detection and autocorrelation

Jones, Weinrott, and Vaught (1978) found that people were poor at detecting change in real series: inter-rater reliability of judgments was low at .39 and average miss and false alarm rates were 48% and 33%, respectively. Sequential dependence (autocorrelation) in series increased false alarm rates. This study used interrupted time series analysis as the gold standard for establishing whether there was a real change in the series. However, series were so short that this statistical approach would have lacked power. People may have been able to detect changes that the statistical analysis could not: if so, their performance may not have been as bad as it appeared to be. To circumvent this problem, Matyas and Greenwood (1990) simulated series with known levels of random noise and first-order autocorrelation. However, they still found that false alarm rates (typically over 40%) were much higher than miss rates (typically about 10%), especially when data were autocorrelated. They concluded that judgment is not as conservative as behaviour analysts assumed.

The increase in false alarm rates under positive autocorrelation is problematic. In single-subject research, where visual assessment of change is still the dominant method (Brossart, Parker, Olson, & Mahadevan, 2006), there is positive autocorrelation in the large majority of series (Busk & Marascuilo, 1988). Why does autocorrelation impair change detection? Consider a time series  $y_{1:T} = (y_1, \dots, y_T)$  which follows an  $r$ -th order autoregressive process

$$y_t = \mu_t + \sum_{k=1}^r \alpha_k (y_{t-k} - \mu_{t-k}) + \varepsilon_t \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \quad (1)$$

This process implies a serial dependence between successive time points such that when a previous value  $y_{t-k}$  is above the mean  $\mu_{t-k}$ , a later value  $y_t$  is more likely to also be above the mean (for  $\alpha_k > 0$ , positive autocorrelation), or more likely to be below the mean ( $\alpha_k < 0$ , i.e., negative autocorrelation).

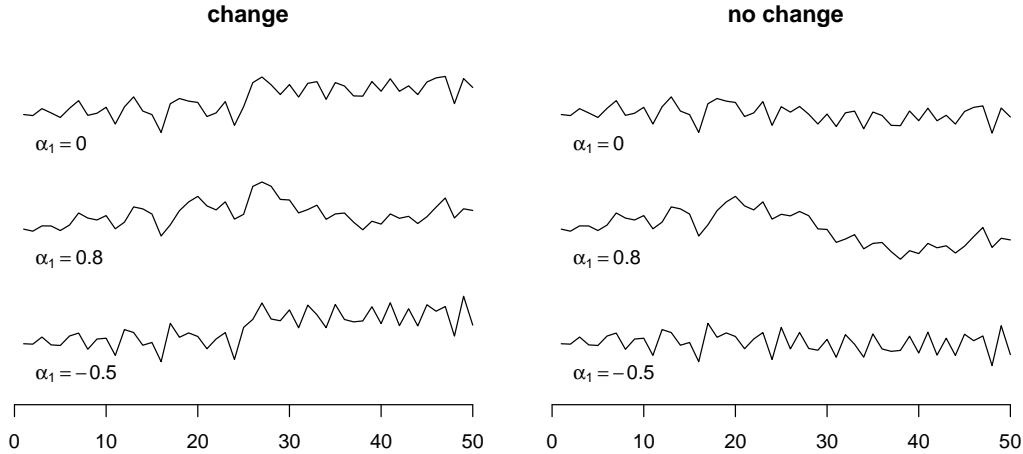


Figure 1: Examples of time series with a change and no change, under different levels of autocorrelation.

We allow for the possibility of an abrupt change in baseline value

$$\mu_t = \mu_0 + \delta \mathbb{I}_{t \geq t^*} \quad \delta \sim N(\mu_\delta, \sigma_\delta^2) \quad (2)$$

where  $t^*$  is the change point and  $\mathbb{I}_{t \geq t^*}$  is an indicator function with value 1 when  $t \geq t^*$  and value 0 otherwise. This is the process used in the experiments reported below. Figure 1 shows examples of time series produced by a first-order ( $r = 1$ ) autoregressive process. Note that each series is created from identical noise values  $\varepsilon_t$ ; the only difference is in the value of the autocorrelation, and whether there is a change in the baseline (after time point 25 in the series on the left) or not. As can be seen in these examples, positive autocorrelation ( $\alpha_1 = 0.8$ ) tends to make series “smoother”, which can make abrupt changes less apparent. Also, when there is no change, positive autocorrelation can increase false alarms, because a sudden large noise perturbation will tend to persist, giving the appearance of a change in mean. In contrast, negative autocorrelation ( $\alpha_1 = -0.5$ ) tends to make series more “jagged”, which can impair change detection by increasing the apparent noisiness of the series, even though the actual error variance is unaffected.

### Cognitive processes underlying change detection

Research on judgmental change detection has been mainly empirical. Little attention has been given to the cognitive processes underlying performance in the task. An exception is the work of Wampold and Furlong (1981), who argued that people may use one of two types of heuristic.

First, people may assess a putative change in the data relative to the overall variability that is present. A putative change is judged to be real if its magnitude exceeds the natural variability in the data by some criterial amount. This is a global assessment: all the presented data are taken into account. According to this model, positive autocorrelation increases false alarms because it is ignored when the natural variability in the data is estimated. This results in an underestimation of the

variability and, hence, over-estimation of the likelihood of a change. In contrast, negative autocorrelation would result in an over-estimation of the variability, and hence an underestimation of the likelihood of a change.

Alternatively, people may monitor the data for large absolute changes and ignore the natural point-to-point variability. This approach assumes that they access internal prototypes of possible changes and classify series into those with and without a change by matching them against these prototypes. In other words, they make local assessments: candidate changes are judged in isolation from the rest of the series.

Steyvers and Brown (2005) proposed that people may rely on a heuristic similar to the latter<sup>1</sup> when detecting changes in times series while making sequential predictions about the next datum in the series. They showed that this heuristic closely followed an optimal Bayesian model, which also fitted participants responses well. In later work, Brown and Steyvers (2009) proposed that people generally use Bayesian inference in these problems, although they may work from incorrect assumptions and use limited samples to approximate the full Bayesian analysis.

### Present study

Previous research suggests that human detection of change in time series is impaired by the presence of autocorrelation. In the two experiments presented here, we sought to replicate this finding with a range of (positive) autocorrelations (Experiment 1) and a second-order autoregressive process (Experiment 2). By formulating different models and fitting these to participants’ judgments, we sought to uncover the cognitive processes underlying change detection in graphically

<sup>1</sup>But not exactly the same. According to Steyvers and Brown (2005), a change will be detected whenever the distance between a prediction and the actual value exceeds a criterial amount. However, the distance was not measured on the scale of the outcome, but rather as the log likelihood ratio of the outcome given a change vs no change.

presented time series. Uncovering the strategies people use when monitoring series for regime change is a first step in determining how to improve judgmental change detection.

## Experiment 1

The objective of Experiment 1 was to assess the accuracy of judgmental change detection under levels of no ( $\alpha_1 = 0$ ), medium ( $\alpha_1 = .4$ ), and high ( $\alpha_1 = .8$ ) autocorrelation.

### Method

**Participants** Fifty participants (23 male) were recruited from the UCL subject pool and paid £5 for their time. The mean age was 26.06.

**Task** The change detection task consisted of 60 trials. On each trial, participants were presented a graph depicting a time series and asked to indicate whether the series contained a change or not. After this, they indicated their confidence in their response.

In Experiment 1, the time series were created by a first-order autoregressive process (i.e., setting  $r = 1$  in Equation 1). The autocorrelation was varied within participants. For each level of autocorrelation,  $\alpha_1 = \{0, .4, .8\}$ , there were 10 series with, and 10 series without a change. Each change  $\delta$  was randomly drawn from a Gaussian distribution with mean  $\mu_\delta = 8$  and variance  $\sigma_\delta^2 = 9$ . The initial baseline value was set at  $\mu_0 = 50$  and the variance of the noise was set at  $\sigma_\epsilon^2 = 5$ .

**Procedure** Participants took the role of a trainee flood engineer with the task of monitoring water levels for risk of flooding. Participants were told a risk of flooding consisted of a persistent increase in water level, but that the level would fluctuate regardless of whether there was a risk of flooding or not. For 60 different locations, they would monitor the water level over a 50 hour period, and participants were informed that a flood risk could occur anywhere between hour 11 and hour 40 (i.e.,  $t^* \in \{11, \dots, 40\}$ ). Finally, participants were instructed that there was a flood risk for half of the locations.

### Results

The main detection results are given in Table 1. Autocorrelation did not affect hit rates,  $F(2, 98) = 1.278$ ,  $p = .283$ , but increased false alarms,  $F(2, 98) = 27.913$ ,  $p < .001$ . This indicates that, as autocorrelation increased, participants adjusted their criterion to detect changes. This was confirmed in a signal detection analysis. Increased levels of autocorrelation reduced the discrimination ( $d'$ ),  $F(2, 98) = 11.915$ ,  $p < .001$ . Contrast analysis showed that this effect was mainly due to a linear trend,  $F(1, 49) = 22.79$ ,  $p < .001$ ; the quadratic trend was not significant,  $F(1, 49) = 1.37$ ,  $p = .25$ . In addition, increased autocorrelation reduced the centered decision criteria ( $C$ ),  $F(2, 98) = 20.56$ ,  $p < .001$ . Contrast analysis showed that this effect was mainly due to a linear trend,  $F(1, 49) = 32.58$ ,  $p < .001$ ; the quadratic trend was not significant,  $F(1, 49) = 0.79$ ,  $p = .38$ . This indicates that autocorrelation increased the difficulty of the task and participants

Table 1: Mean hit (H) and false alarm (FA) rates, discrimination ( $d'$ ) and (centered) criterion ( $C$ ) parameters. Values in parentheses are standard deviations.

$\alpha_1$	$\alpha_2$	H	FA	$d'$	$C$
Experiment 1					
0		.72 (.18)	.06 (.15)	2.06 (.69)	-0.40 (.12)
0.4		.76 (.14)	.11 (.19)	1.90 (.65)	-0.44 (.12)
0.8		.77 (.15)	.24 (.20)	1.48 (.77)	-0.50 (.11)
Experiment 2					
0.5	0.3	.68 (.24)	.14 (.15)	1.65 (.86)	-0.42 (.14)
0.5	0.0	.74 (.22)	.10 (.15)	2.01 (.85)	-0.43 (.12)
0.5	-0.3	.70 (.23)	.07 (.16)	2.00 (.81)	-0.40 (.14)
-0.5	0.3	.72 (.22)	.07 (.14)	2.07 (.89)	-0.41 (.12)
-0.5	0.0	.72 (.23)	.03 (.07)	2.18 (.67)	-0.39 (.12)
-0.5	-0.3	.68 (.24)	.03 (.09)	2.09 (.80)	-0.37 (.12)

responded by lowering the criterion to detect a change (biasing decisions towards changes), resulting in increased false alarms.

Average confidence levels for the different trial types and autocorrelation levels are depicted in Figure 2. We analysed the confidence ratings with a linear mixed effects model, including random intercepts for each participant, as well as random slopes for the autocorrelation and contrast codes for whether a decision was correct and whether it was a “change” (vs a “no change”) decision. This showed that confidence was higher for correct (hits and correct rejections) than incorrect decisions (misses and false alarms),  $F(1, 2946) = 216.4$ ,  $p < .001$ . In addition, confidence was generally lower when participants responded change (hits and false alarms) compared to no change (correct rejections and misses),  $F(1, 2946) = 11.19$ ,  $p < .001$ . A significant interaction between these two factors shows that confidence was more strongly related to correctness when people judged there was a change than when people judged there was no change,  $F(1, 2946) = 4.68$ ,  $p = .031$ . Finally, confidence decreased as the level of autocorrelation increased,  $F(1, 2946) = 16.83$ ,  $p < .001$ .

To summarize the results, increasing autocorrelation resulted in poorer discrimination between series with and those without a change. This increased difficulty was also reflected in participants’ confidence in their judgments. Participants appeared to respond to the increased difficulty by relaxing their decision criteria in favour of detecting change, resulting in increased false alarm rates.

## Experiment 2

The objective of Experiment 2 was to investigate change detection in a second-order autoregressive process. Depending on the autocorrelation values, second-order autoregressive processes can show complex periodic patterns (e.g., Gottman, 1981). In particular, when  $\alpha_1^2 + 4\alpha_2 < 0$ , the spectral density functions show broad peaks across a band of mid-range

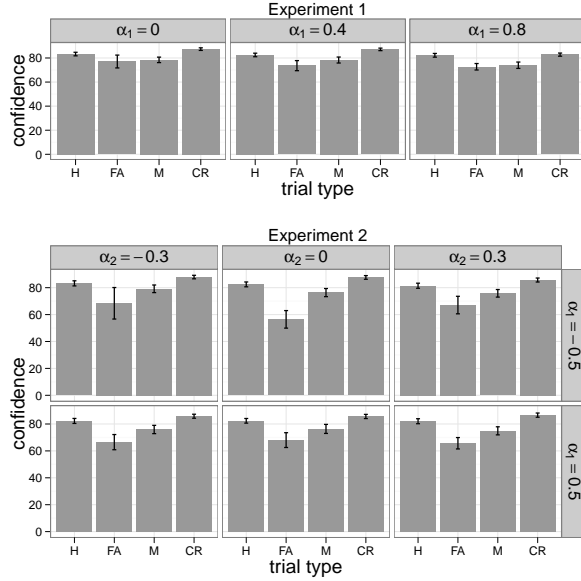


Figure 2: Mean confidence levels by autoregression level and trial type (H = hit, FA = false alarm, M = miss, CR = correct rejection). Error bars represent 95% confidence intervals.

frequencies, and the time series is nondeterministically periodic. Such periodic behaviour may appear as a change in mean, increasing false alarms. To investigate this possibility, Experiment 2 included series with and without such periodic behaviour.

## Method

**Participants** 70 students (18 male) participated in the experiment for course credit. The mean age was 18.97 ( $SD = 1.12$ ).

**Task** The task was identical to that in Experiment 1. However, the time-series were now produced by a second-order autoregressive process (setting  $r = 2$  in Equation 1), with autocorrelation parameters  $\alpha_1 = \{-0.5, 0.5\}$  and  $\alpha_2 = \{-0.3, 0, 0.3\}$ . There were 10 series for each combination of the autocorrelation parameters, of which five did and five did not contain a change. Series were presented in random order. Series with  $\alpha_1 = .5$  and  $\alpha_2 = -.3$ , or  $\alpha_1 = -.5$  and  $\alpha_2 = -.3$ , were likely to show periodic trends. If this impairs detection ability, we would expect to find a difference between these series and those generated with other combinations of autocorrelation parameters.

## Results

The main detection results can be found in Table 1. As in Experiment 1, autocorrelation affected false alarm rates,  $F(1, 68) = 56.67$ ,  $p < .001$  and  $F(2, 136) = 24.99$ ,  $p < .001$ , for  $\alpha_1$  and  $\alpha_2$  respectively, but not hit rates,  $F(1, 68) = 0.03$ ,  $p = .87$  and  $F(2, 136) = 1.42$ ,  $p = .25$ . Signal detection analysis showed that both autocorrelation parameters affected

discrimination ability ( $d'$ ),  $F(1, 69) = 9.81$ ,  $p = .003$ , and  $F(2, 138) = 3.50$ ,  $p = .033$ , for  $\alpha_1$  and  $\alpha_2$  respectively; the interaction was not significant,  $F(2, 138) = 1.79$ ,  $p = .17$ . Discrimination was generally better for  $\alpha_1 = -.5$  than  $\alpha_1 = 0.5$ . Discrimination was worse for  $\alpha_2 = 0.3$  compared to the other two values,  $F(1, 69) = 5.77$ ,  $p = .019$ , while there was no difference between  $\alpha_2 = 0$  and  $\alpha_2 = -0.3$ ,  $F(1, 69) = 0.52$ ,  $p = .47$ . Both autocorrelation parameters also affected the centered decision criteria ( $C$ ),  $F(1, 69) = 7.52$ ,  $p = .008$ , and  $F(2, 138) = 3.17$ ,  $p = .045$ . Decision criteria were more biased towards change for positive compared to negative autocorrelations (for  $\alpha_2$ , there was a linear trend,  $F(1, 69) = 5.78$ ,  $p = .019$ , but no quadratic trend,  $F(1, 69) = 0.81$ ,  $p = .37$ ).

Average confidence levels for the different trial types and autocorrelation levels are depicted in Figure 2. Analysis with a linear mixed effects model replicated the results of Experiment 1. Confidence was higher for correct (hits and correct rejections) than incorrect decisions (misses and false alarms),  $F(1, 4124) = 226.14$ ,  $p < .001$ , and lower when participants responded change (hits and false alarms) compared to no change (correct rejections and misses),  $F(1, 4124) = 44.80$ ,  $p < .001$ . Confidence was more strongly related to correctness when people judged there was a change than when people judged there was no change,  $F(1, 4124) = 8.1$ ,  $p = .005$ . Confidence was higher for negative than positive first-order autocorrelation ( $\alpha_1$ ),  $F(1, 4124) = 5.69$ ,  $p = .017$ , and this difference was larger when the second-order autocorrelation was negative rather than positive,  $F(1, 4124) = 5.62$ ,  $p = .018$ .

The results of this experiment replicate those of Experiment 1 with a second-order autoregressive process. Positive (first- and second-order) autocorrelation resulted in poorer discrimination between series with and series without a change and this increased difficulty was reflected in participants' confidence in their judgments. Participants responded to the increased difficulty by relaxing their decision criteria, increasing false alarm rates. As the interaction between the autocorrelations was not significant, we found no evidence that periodic trends in the time series affected detection ability.

## A Bayesian change detection model

To compare participants' judgments against a gold standard, we used a Bayesian model to detect changes in time series as defined by Equations 1 and 2. By taking the autocorrelation into account, this model is expected to perform well, although the relatively short length of the series may limit its performance.

In our analysis, change detection is based on the relative evidence for a model  $M_1$ , which incorporates the possibility of a change, over a model  $M_2$ , which does not allow for change. The measure of relative evidence is the Bayes Factor

$$BF = \frac{p(y_{1:T}|M_1)}{p(y_{1:T}|M_2)} \quad (3)$$



where  $p(y_{1:T}|M_j)$  the marginal likelihood

$$p(y_{1:T}|M_j) = \int p(y_{1:T}, \theta_j | M_j) d\theta_j \quad (4)$$

integrating over the parameters  $\theta_j$  of model  $j$ . For  $M_1$ , the parameters are  $\theta_1 = \{\mu_0, \delta, \alpha_1, \alpha_2, \sigma_\epsilon, t^*\}$ . The parameters of model  $M_2$  exclude  $\delta$  and  $t^*$ . Recall that participants were informed there could be only one change point in each series, and that possible change points could be anywhere between time points 11 and 40. In model 1, the posterior distribution of the change point  $t^* \in \{11, \dots, 40\}$ , conditional on the other parameters, can then be expressed relatively simply as

$$p(t^* | y_{1:T}, \theta_{1,-t^*}, M_1) \propto p(y_{1:T} | t^*, \theta_{1,-t^*}, M_1) p(t^* | M_1) \quad (5)$$

where  $\theta_{1,-t^*}$  denotes the parameter vector excluding  $t^*$ , and  $p(t^* | M_1)$  the prior distribution over the change points  $t^*$ , which we took to be uniform. For  $\mu_0$  and  $\delta$ , we used truncated Normal distributions with means of 50 and 0 respectively, and variance  $10^5$ . Both distributions were restricted to the range between 0 and 100. For  $\mu_0$ , this reflects the range of the time series on the graphs. For  $\delta$ , this reflects that changes can only be positive. For  $\sigma_\epsilon$ , an inverse Gamma distribution was used. For  $\alpha_1$  and  $\alpha_2$ , we used truncated Normal distributions (centered on 0) restricting the range such that the process is stationary<sup>2</sup>. Posterior distributions for the parameters can be efficiently estimated by Gibbs sampling (for computational details, see Albert & Chib, 1993). Bayes Factors were estimated from the Gibbs sampler using the technique of Chib (1995).

We computed the Bayes Factor for each of the time series in the two experiments. Using the simple criterion of  $BF > 1$  to detect a change, the hit rate of the model was 92.3%, but the false alarm rate was rather high at 15.8%. Inspection of the parameter estimates showed that false alarms were generally associated with relatively small changes (50% of the posterior means of  $\delta$  were smaller than 2.31) occurring relatively late in the series (50% of the posterior modes of  $p(t^*)$  were larger than  $t = 33$ ). False alarm rates increased with positive autocorrelation. This suggests that even while explicitly accounting for autocorrelation, the Bayesian model is not immune to illusory changes produced by autocorrelation.

## Modelling human change detection

To link the Bayesian and heuristic models to participants' responses  $R_{ik}$ , we assume that, for a time-series  $k$ , each method of change detection  $j$  provides a signal  $v_{jk}$ , which is corrupted by noise  $e_{ijk} \sim N(0, \sigma_{ij})$ , and that participants judge there to be a change when the noisy signal exceeds a criterion  $c_{ij}$ . As a result, the probability of a change judgment, for participant  $i$  judging series  $k$  with model  $j$ , can be written as

$$P(R_{ik} = \text{change} | j) = 1 - \Phi\left(\frac{v_{jk} - c_{ij}}{\sigma_{ij}}\right) \quad (6)$$

<sup>2</sup>For a first-order autoregressive process, that means that  $|\alpha_1| < 1$ . For a second-order autoregressive process, the requirements are that  $\alpha_1 + \alpha_2 < 1$ ,  $\alpha_2 - \alpha_1 < 1$ ,  $|\alpha_2| < 1$ .

Table 2: Model fits (AIC) and numbers of participants best fitted according to the AIC ( $n_{\text{best}}$ ).

	Experiment 1		Experiment 2	
	AIC	$n_{\text{best}}$	AIC	$n_{\text{best}}$
Bayes	2891	3	3728	9
CRV	2280	47	2846	61
LAC	3512	0	5314	0

where  $\Phi$  denotes the cumulative Normal distribution. For the Bayesian model, we assume that

$$v_{\text{Bayes},k} = \log BF \quad (7)$$

i.e., that decisions depend on the logarithm of the Bayes Factor (Equation 3).

In addition to the Bayesian model, we formalised the heuristics of Wampold and Furlong (1981) described in the introduction. We'll refer to these as the Change-Relative-to-Variability (CRV) and the Largest-Absolute-Change (LAC) heuristic. The CRV heuristic compares a putative change to the overall variation in the series. In doing so, we assume people first visually fit a step function to the series, in such a way as to minimize the noise. The step function represents the mean water level before and after the change-point, and with that the timing and size of the putative change in water level. This putative change is compared to the deviations of the series around the step function. More formally, we assume the (increasing) step function, defined by the initial level  $m_0$ , change point  $t'$ , and the increase  $d$  after the change, is determined by minimizing the sum of squared error (SSE):

$$t', m_0, d = \arg \min_{t', m_0, d} \sum_{t=1}^T (Y_t - m_0 - d \mathbb{I}_{t \geq t'})^2 \quad (8)$$

The putative change  $d$  is then compared to an estimate  $s$  of the standard deviation (derived from the sum of squared errors). The value used for decisions, according to this heuristic, is then simply

$$v_{\text{CRV},k} = d - s \quad (9)$$

According to the second heuristic, a change is determined solely by comparing deviations between time points to a pre-existing "prototype". For this heuristic, we therefore assume the signal consists of the maximum deviation between consecutive time points

$$v_{\text{LAC},k} = \max_t |y_t - y_{t-1}| \quad (10)$$

We fitted the models in two ways. First, we fitted each model to the whole group of participants using a generalized linear mixed effects model, including random decision criteria  $c_{ij}$  and dispersion parameters  $\sigma_{ij}$ . In addition, we fitted each model to each participant separately. The results of both analyses (Table 2) were in agreement: the model that best described participants' responses was the Change-Relative-to-Variability (CRV) heuristic, followed by the Bayesian model.

The Largest-Absolute-Change (LAC) heuristic fitted none of the participants best.

## Discussion

We presented two experiments on human change detection in autocorrelated time series. In agreement with Matyas and Greenwood (1990), we found little evidence of conservatism. As the level of autocorrelation increased, participants maintained a similar hit rate, but increased their false alarm rate, indicating a relaxing of decision criteria such that more changes are (erroneously) detected. For second-order autoregressive processes, detection was most impaired when both autocorrelations were positive. There was no evidence that periodic trends resulting from particular autocorrelation levels affected detection performance.

Most participants responded in accordance with a simple change detection heuristic, where an underlying change-in-mean function is (visually) fitted to a noisy series and it is determined whether the putative change exceeds the natural variability in the series.<sup>3</sup> For the graphically presented time series used here, this strategy seems plausible. And as this strategy closely matches the performance of the Bayesian analysis which accounts for autocorrelation in the series, it is not a bad strategy either – at least not for the types of series studied here. Participants were explicitly told the range of time points over which a change could occur, as well as that there could only be one change in each series. For more complex problems with multiple change points or changes in other parameters than the baseline, the Bayesian and heuristic analyses are more likely to diverge.

We found no evidence that people relied on (absolute) differences between values on successive time points, a heuristic suggested by e.g. Steyvers and Brown (2005). An important difference between the present study and the latter one is that our participants were presented with complete time series, while Steyvers and Brown used an online prediction task in which participants viewed each datum sequentially and thus had to rely on memory. It is likely that change detection strategies will differ between online and offline detection tasks. In online tasks, the information available for detection is constrained by (working) memory capacity, as well as by the fact that only previous data can be used to judge a change at the current datum. In this case, strategies that rely on small samples, such as an absolute change heuristic, seem more plausible than strategies which use all the data, such as the Bayesian model and Change-Relative-to-Variability heuristic as implemented here.

Autocorrelation clearly impeded detection performance. Further research is required to assess the extent to which people can learn to “see through” autocorrelation. In the present experiments, participants did not receive feedback about their detection performance. It is possible that, after extensive

training, people can learn to distinguish between real changes and those that are merely apparent due to autocorrelation. In domains such as risk assessment, the finding that autocorrelation increases false alarms, but does not decrease hit rates, may provide some comfort. However, when assessing treatment effectiveness, a more cautious approach may be called for.

## Acknowledgements

This work was supported by the Economic and Social Research Council (ESRC), grant RES-062-23-2735.

## References

- Albert, J. H., & Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics*, 11, 1–15.
- Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis*, 10, 167–172.
- Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification*, 30, 531–563.
- Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive psychology*, 58, 49–67.
- Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment*, 10, 229–242.
- Carlin, B. P., Gelfand, A. E., & Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Journal of the Royal Statistical Society. Series C (Applied statistics)*, 41, 389–405.
- Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90, 1313–1321.
- Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientist*. Cambridge: Cambridge University Press.
- Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45, 39–70.
- Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis*, 10, 151–166.
- Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single case time-series: Effects of variability, serial dependence and magnitude of intervention effects. *Journal of Applied Behavior Analysis*, 23, 341–351.
- Steyvers, M., & Brown, S. D. (2005). Prediction and change detection. In *Advances in neural information processing systems*, 18 (pp. 1281–1288).
- Wampold, B., & Furlong, M. (1981). The heuristics of visual inference. *Behavioral Assessment*, 3, 79–82.

<sup>3</sup>In the present experiment, the variability actually seemed to have little effect on responses and a version without variability fitted just as well. However, as differences in variability between the series were relatively small, further research is required.

# Using listener gaze to augment speech generation in a virtual 3D environment

Maria Staudte  
Saarland University

Alexander Koller  
University of Potsdam

Konstantina Garoufi  
University of Potsdam

Matthew Crocker  
Saarland University

## Abstract

Listeners tend to gaze at objects to which they resolve referring expressions. We show that this remains true even when these objects are presented in a virtual 3D environment in which listeners can move freely. We further show that an automated speech generation system that uses eyetracking information to monitor listener's understanding of referring expressions outperforms comparable systems that do not draw on listener gaze.

## Introduction

In situated spoken interaction, there is evidence that the gaze of interlocutors can augment both language comprehension and production processes. For example, speaker gaze to objects that are about to be mentioned (Griffin & Bock, 2000) has been shown to benefit listener comprehension by directing listener gaze to the intended visual referents (Hanna & Brennan, 2007; Staudte & Crocker, 2011; Kreysa & Knoeferle, 2011). Even when speaker gaze is not visible to the listener, however, listeners are known to rapidly attend to mentioned objects (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). This gaze behavior on the part of listeners potentially provides speakers with useful feedback regarding the communicative success of their utterances: By monitoring listener gaze to objects in the environment, the speaker can determine whether or not a referring expression (RE) they have just produced was correctly understood or not, and potentially use this information to adjust subsequent production.

In this paper we investigate the hypothesis that speaker use of listener gaze can potentially enhance interaction, even when situated in complex and dynamic scenes that simulate physical environments. In order to examine this hypothesis in a controlled and consistent manner, we monitor listener performance in the context of a computer system that generates spoken instructions to direct the listener through a 3D virtual environment with the goal of finding a trophy. Successful completion of the task requires listeners to press specific buttons. Our experiment manipulated whether or not the computer system could follow up its original RE with feedback based on the listener's gaze or movement behavior, with the aim of shedding light on the following two questions:

- Do listener eye movements provide a consistent and useful indication of referential understanding, on a per-utterance basis, and when embedded in a dynamic and complex, goal-driven scenario?
- What effect does gaze-based feedback have on listeners' (gaze-)behavior and does it increase the more general effectiveness of an interaction?

We show that the listeners' eye movements are a reliable predictor of referential understanding in our virtual environ-

ments. A natural language generation (NLG) system, that exploited this information to provide direct feedback, communicated its intended referent to the listener more effectively than similar systems that did not draw on listener gaze. Gaze-based feedback was further shown to increase listener attention to potential target objects in a scene, indicating a generally more focused and task-oriented listener behavior. This system is, to our knowledge, the first NLG system that adjusts its referring expressions to listener gaze.

## Related work

Previous research has shown that listeners align with speakers by visually attending to mentioned objects (Tanenhaus et al., 1995) and, if possible, to what the speaker attends to (Richardson & Dale, 2005; Hanna & Brennan, 2007; Staudte & Crocker, 2011). Little is known, however, about speaker adaptation to the listener's (gaze) behavior, in particular when this occurs in dynamic and goal-oriented situations. Typically, Visual World experiments have used simple and static visual scenes and disembodied utterances and have analyzed the recorded listener gaze off-line (e.g., Altmann & Kamide, 1999; Knoeferle, Crocker, Pickering, & Scheepers, 2005). Although studies involving an embodied speaker inherently include some dynamics in their stimuli, this is normally constrained to speaker head and eye movements (Hanna & Brennan, 2007; Staudte & Crocker, 2011). Besides simplifying the physical environment to a static visual scene, none of these approaches can capture the reciprocal nature of interaction. That is, they do not take into account that the listeners' eye movements may, as a signal of referential understanding to the speaker, change the speaker's behavior and utterances on-line and, as such, affect the listener again.

One study that emphasized interactive communication in a dynamic environment was conducted by Clark and Krych (2004). In this experiment, two partners assembled Lego models: The directing participant advised the building participant on how to achieve that goal. It was manipulated whether or not the director could see the builder's workspace and, thus, use the builder's visual attention as feedback for directions. Clark and Krych found, for instance, that the visibility of the listener's workspace led to significantly more deictic expressions by the speaker and to shorter task completion times. However, the experimental setting introduced large variability in the dependent and independent variables, making controlled manipulation and fine-grained observations difficult. In fact, we are not aware of any previous work that has successfully integrated features of natural environments—realistic, complex and dynamic scenes in which the visual salience of objects can change as a result of the listener's moves in the environment—with the reciprocal

nature of listener-speaker-adaptation while also being able to carefully control and measure relevant behavioral data.

Recently, researchers have examined eye gaze of speakers and listeners in the scenes of Tangram puzzle simulations on computer screens (Kuriyama et al., 2011; Iida, Yasuhara, & Tokunaga, 2011). In these experiments, eye gaze features are found to be useful for a machine learning model of reference resolution. However, this setting is restricted in its dynamics, as it does not embed the objects into physical scenes or involve any updates to the spatial and visual context of the objects in the scenes. In contrast, by *generating* REs and asking the subjects to resolve them, rather than resolving human-produced REs itself, the system we propose here can provide more control over the language that is used in the interaction.

Computational models of gaze behavior are frequently implemented in embodied conversational agents as part of non-verbal behavior that aims at improving the human-computer interaction (see e.g. Foster, 2007). Such agents do not typically employ listener gaze tracking for the generation of appropriate REs, though. One work that focuses on situated RE generation is Denis (2010), which takes the visual focus of objects into account for the gradual discrimination of referents from distractors in a series of utterances. However, visual focus in Denis’ work is modeled by visibility of objects on screen rather than eye gaze. To our knowledge, there exists no prior RE generation algorithm that is informed directly by listener gaze.

Finally, gaze as a modality of interaction has been investigated in virtual reality games before, e.g. by Hülsmann, Dankert, and Pfeiffer (2011). However, most such settings do not use language as a further modality. One virtual game-like setting which focuses on language is the recent Challenge on Generating Instructions in Virtual Environments (GIVE; Koller et al., 2010), which evaluates NLG systems that produce natural-language instructions in virtual environments. In this work we use the freely available open-source software infrastructure provided by GIVE<sup>1</sup> to set up our experiment.

## Methods

In the GIVE setting (Koller et al., 2010; Striegnitz et al., 2011), a human user can move about freely in a virtual indoor environment featuring several interconnected corridors and rooms. A 3D view of the environment is displayed on a computer screen as in Fig. 1, and the user can walk forward and backward, and turn left and right, using the cursor keys. They can also navigate to buttons and, once they have approached them closely enough, click on them with the mouse to press them. In Fig. 1 the object currently under inspection by the user is the rightmost button on the wall, marked with a large white circle. The trace of the fixation’s coordinates is rendered by smaller white circles. These gaze markings do not appear on the user’s screen during the experiment.

The user interacts with an NLG system in the context of a treasure-hunt game, where the user’s task is to find a trophy



Figure 1: A screenshot of one of the virtual 3D environments.

hidden in a wall safe. They must press certain buttons in the correct sequence in order to open the safe; since they do not have prior knowledge of which buttons to press, they rely on instructions and REs generated by the NLG system in order to carry out the task. A room may contain several buttons other than the *target*, which is the button that the user must press next. These other buttons are called *distractors* and are there to make the RE resolution task more challenging. Rooms also contain a number of landmark objects, such as chairs and plants, which cannot be interacted with, but may be used in REs to nearby targets. For our experiment we use three different virtual environments designed by Gargett, Garoufi, Koller, and Striegnitz (2010), which differ in what objects they contain and where they are located.

## Generation systems

We implemented three different NLG systems for generating instructions in these virtual environments. All systems generate navigation instructions, which guide the user to a specific location, as well as object manipulation instructions such as “press the blue button” containing REs such as “the blue button”. The generated instructions are converted to speech by the MARY text-to-speech system (Schöder & Trouvain, 2003) and presented via loudspeaker. At any point, the user may press the ‘H’ key on their keyboard to indicate that they are confused. This will cause the NLG system to generate a clarified instruction. All three systems operate on the same codebase for the generation of simple yet effective navigation instructions (e.g. “go through the doorway”), but differ in their RE generation strategies.

Our baseline system generates REs that are optimized for being easy for the listener to understand, according to a corpus-based model of understandability (Garoufi & Koller, 2011). Crucially, this system does not monitor whether the listener understood an RE. It never gives any (positive or negative) feedback, and will only generate a follow-up RE if the user either asks for help (‘H’ key) or presses the wrong button. Therefore we call this system the *no-feedback* system.

The *movement* system extends the no-feedback system by

<sup>1</sup><http://www.give-challenge.org/research>

monitoring the user's movements in the game after it has uttered an RE, and attempting to predict whether they will press the button it described or not. This system does nothing until only a single button in the current room is visible to the user; then it tracks the user's distance from this button, where "distance" is a weighted sum of walking distance to the button and the angle the user must turn to face the button. If, after hearing the RE, the user has decreased the distance by more than a given threshold, the system concludes that the hearer has resolved the RE as this button. If it is the button the system intended to refer to, it utters the positive feedback "yes, that one!" For incorrect buttons, it utters the negative feedback "no, not that one."

Finally, the *eyetracking* generation system attempts to predict whether the user will press the correct button or not by monitoring their gaze. At intervals of approximately 15 ms, the system samples the (x,y) position on the screen that the user is looking at. It then resolves this (x,y) screen position to an object in the 3D scene. If the user fixates the same object for more than 300 ms, the system counts this as an inspection of that object; interruptions of the inspection of less than 150 ms are ignored. Once it has detected an inspection to a button in the room, the eyetracking system generates positive or negative feedback utterances in exactly the same way as the movement system does.

The system maps the screen positions reported by the eyetracker to 3D objects as follows: When the 3D engine renders the 3D scene onto the 2D screen, it assumes a certain position of the "camera" in the 3D environment; this roughly corresponds to the position of the user's eyes. For each object that is currently visible, the system computes its bounding box, i.e. the smallest box that completely contains the object. It determines the minimum angle  $\alpha$  between the ray from the camera position to some corner of the bounding box and the ray from the camera position to the center of the bounding box. Intuitively,  $\alpha$  represents the size of the object on the screen. The system also determines the angle  $\beta$  between the ray from the camera position to the (x,y) position in the screen plane reported by the eyetracker and the center of the bounding box. Small values of  $\beta$  represent situations in which the user looks directly at the center of an object. An object is a candidate for being fixated if one of  $\beta/\alpha$  or  $\beta - \alpha$  is below a certain threshold. Among all candidates (if there are any), the system then finally chooses the object with the smallest  $\beta$ .

Both the movement-based and the eyetracking-based model withhold their feedback until a first full description of the referent (a *first-mention RE*) was spoken. Additionally, they only provide feedback on newly approached or inspected buttons and will not repeat this feedback unless the listener has approached or inspected another button in the meantime. We call the time between the onset of the first-mention RE and the next button press in a scene, the *critical time region*.

## Participants

Thirty-one students, enrolled at Saarland University, were paid to take part in this study (12 females). All reported

their English skills as fluent, and all were able to complete the tasks. Their mean age was 27.6.

## Task and procedure

A faceLAB eyetracking system<sup>2</sup> remotely monitored participants' eye movements on a 24-inch monitor. Before the experiment, participants received written instructions that described the task and explained that they would be given instructions by an NLG system. They were encouraged to request additional help anytime they felt that the instructions were not sufficient (by pressing the 'H' key).

The eye-tracker was calibrated using a nine-point fixation stimulus. We disguised the importance of gaze from the participants by telling them that we videotaped them and that the camera needed calibration. Participants then started with a short practice session to familiarize themselves with the game controls and to clarify remaining questions, before playing three full games (each with a different virtual environment and generation system). The order of games was alternated according to the Latin square design. Finally, each participant received a questionnaire which aimed to assess whether participants noticed that they were eye-tracked and that one of the generation systems made use of that. The entire experiment lasted approximately 30 minutes.

## Analysis

Firstly, we determined whether the participant pressed the correct button (without having to ask for help by pressing the 'H'-key) by comparing each button the participant pressed with the target referent of the most recent first-mention RE. REs that did not lead to a button press (e.g. because the participant navigated away to another room, causing the system to switch to navigation instructions) were considered unsuccessful. This served as a dependent variable but also as a means for subdividing data according to un-/successful trial completion. Secondly, inspections recorded on a button in the player's room, i.e., on the target or a distractor, during the critical time region were registered in all conditions (not just the eyetracking NLG system) and analyzed as a main dependent variable. Further total trial time, i.e., the time taken from the onset of an RE to the button press, as well as the onset time of system feedback (when provided) were recorded. Finally, we considered the frequency with which participants asked for help by pressing the 'H' key as a measure of confusion.

To control for external factors, we discarded individual scenes in which the systems rephrased their first-mention REs (e.g. by adding further attributes), as well as a few scenes which the participants had to go through a second time due to technical glitches. To remove errors in eyetracker calibration, we included interactions with the eyetracking NLG system in the analysis only when we were able to record inspections (to the referent or any distractor) in at least 80% of all referential scenes. This filtered out 9 interactions out of the 93 we collected.

<sup>2</sup><http://www.seeingmachines.com/product/facelab>

Inferential statistics on this data were carried out using mixed-effect models from the lme4 package in R (Baayen, Davidson, & Bates, 2008). Specifically, we used logistic regression for modeling binary data such as referential success rates, Poisson regression for count variables (e.g., ‘H’-key strokes) and linear regression for inspection durations. Further, main effects and interactions were determined through model reduction, which assesses the contribution of a predictor or interaction to a fitting model by running a  $\chi^2$ -comparison between models with and without the particular predictor(s).

## Results

The post-task questionnaires, revealed no differences in participants’ preferences for any particular NLG system. Similar numbers of participants chose each of the systems on questions such as “which system did you prefer”. When asked for differences between the systems in free-form questions, no subject mentioned eye gaze. We take this to mean that the participants did not realize they were being eyetracked.

### Eye movements

We recorded and analyzed inspections to target and distractor buttons in all conditions. Mean inspection durations during the critical time region (reference onset until button press) were correlated with the success in pressing the correct button and are provided in Table 1.

To investigate our first hypothesis, namely that listener eye movements provide a consistent and useful indication of referential understanding even when embedded in a dynamic, complex and goal-driven scenario, we first consider our baseline condition, the no-feedback system, separately: Model reduction revealed that both inspection duration on the target and inspection duration on the distractors indeed predict success ( $\chi^2(1) = 28.87, p < .001$  and  $\chi^2(1) = 96.24, p < .001$ , respectively). While target inspection duration positively predicts success (Coeff. = 0.00110, SE = 0.00024, Wald’s  $Z = 4.53, p < .001$ ), distractor inspections negatively predict success (Coeff. =  $-0.00178$ , SE = 0.00027, Wald’s  $Z = -6.71, p < .001$ ).

Further, to assess the influence of gaze-based feedback back on listeners’ gaze behavior, we investigated whether the type of system used for generating REs did in fact influence inspection durations (as given in Table 1). We fitted models to target inspection duration and distractor inspection duration using system as predictor, for successful and unsuccessful scenes separately. Model reductions revealed a main effect of system (target:  $\chi^2(2) = 12.79, p < .01$ , distractor:  $\chi^2(2) = 47.10, p < .001$ ) on both inspection variables, but only in successful scenes. That is, with the eyetracking-based feedback system, participants inspected both the target and distractor buttons longer than with the other two systems. An average trial also lasted longer with this system than with the no-feedback system. In unsuccessful scenes no significant differences between inspection durations were observed.

Table 1: Mean inspection durations for target and distractor buttons and the total trial time in milliseconds, for successful and unsuccessful button presses separately. (ET = eyetracking-based system, MOV = movement-based system, NO = no-feedback system.) Differences to ET are significant at: \*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$ , #  $p < 0.1$ .

System (# Trials)	Target	Distractor	Trial Total
Successful:			
<b>ET</b> (182)	2111.6	720.5	8096
<b>MOV</b> (258)	1493.8**	260.5***	7418
<b>NO</b> (237)	1492.0***	185.7***	6877**
Unsuccessful:			
<b>ET</b> (16)	752.1	3378.9	10892
<b>MOV</b> (37)	602.6	2113.1	10343
<b>NO</b> (47)	619.5	1891.7	9130

However, this is most likely due to the low amount of unsuccessful scenes.

Finally, we considered only cases in which feedback was indeed given in order to more precisely assess the influence of effective feedback (types) on participant inspections during reference resolution. Table 2 shows this data further subdivided into scenes with initially positive feedback and scenes with initially negative feedback (the eyetracking system is used as intercept for comparisons between both systems). This is to explore the effect of positive and confirming feedback given by each system and the possibly different effect of negative feedback which unspecifically re-directs the participant to other buttons. We observed that positive feedback of both systems leads to a similar increase of target and decrease of distractor inspections (cf. before and after columns in Table 2). However, eyetracking-based feedback was given earlier (Coeff. = 573.6, SE = 240.2,  $t = 2.39, p(\text{MCMC}) < 0.05$ ) and led to overall longer inspections of the target *and* distractor buttons relative to the trial duration. That is, participants spent significantly more time of a trial (34.1%) looking at potential target buttons than with movement-based feedback (25.5%, Coeff. =  $-0.0552$ , SE = 0.0178,  $t = -3.11, p(\text{mcmc}) < 0.01$ ). This effect was even larger with negative feedback where the difference in feedback onset was even greater (Coeff. = 1237.8, SE = 378.1,  $t = 3.27, p(\text{MCMC}) < 0.01$ ) and the relative button inspection time was also longer (Coeff. =  $-0.1818$ , SE = 0.0283,  $t = -6.43, p(\text{MCMC}) < 0.001$ ). Possibly because of this large difference in feedback onset, we also found (marginally) longer inspections to the buttons after feedback onset.

### Interaction Effectiveness

To evaluate our second hypothesis, namely that gaze-based feedback potentially sustains a more effective interaction than other or no feedback, we considered several indicators for in-

Table 2: Mean values for initial positive and negative feedback separately: inspection durations for target and distractor buttons (before and after feedback onset), feedback onset times, total trial durations, proportion of time spent fixating buttons during trials, and referential success rates. Differences to ET are significant at: \*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$ , #  $p < 0.1$ .

	Target		Distractor		Feedback Onset	Trial Total	Button Fix. Proportion	Success
	Before	After	Before	After				
Positive Feedback:								
ET	513	1389	111	67	4115	6511	34.1	97.6
MOV	465	1123	196	30	4688*	7051*	25.5**	97.0
Negative Feedback:								
ET	109	2155	733	1596	3987	11888	39.5	84.0
MOV	120	926***	484#	802#	5225**	11319	20.1***	68.0*

teraction effectiveness. As a first measure, we looked at the frequency with which participants pressed the ‘H’ key to indicate their confusion. The overall average of ‘H’ keystrokes per game was 1.14 for the eyetracking generation system, 1.77 for the movement system was employed, and 2.26 for the no-feedback system. A model fitted to the key stroke distribution per system shows significant differences both between the eyetracking and the no-feedback system (Coeff. = 0.703, SE = 0.233, Wald’s  $Z = 3.012$ ,  $p < .01$ ) and between the eyetracking and the movement-based system (Coeff. = 0.475, SE = 0.241, Wald’s  $Z = 1.967$ ,  $p = .05$ ).

A second measure of interaction quality is the ratio of all REs that the participants resolved correctly. Mean success rates for trials with feedback only are further reported in the final column of Table 2. Logistic mixed-effects models revealed a significant difference in success rates (Coeff. =  $-0.918$ , SE = 0.461, Wald’s  $Z = -1.990$ ,  $p < .05$ ) for negative feedback while the success rates were similar for positive feedback. Additionally, total trial time is significantly shortened by positive (but not negative) eyetracking-based feedback (Coeff. = 713.7, SE = 311.4,  $t = 2.29$ ,  $p < .05$ ). Thus, when positive feedback was given, the eyetracking system had shorter trial times (along with earlier feedback), while having similar success rates as the movement system. Conversely, negative feedback led to similar trial times but with higher success rates by the eyetracking system.

## Discussion

Concerning our first hypothesis—that gaze reflects online referential understanding even in dynamic 3D environments—we find that participants indeed tend to rapidly fixate the object described by the system. Appropriate feedback by the eyetracking system, in turn, elicits longer inspection durations on potential targets, showing more focused, task-oriented listener attention.

This positive finding is further supported by the perfor-

mance of the eyetracking NLG system, which outperforms the no-feedback baseline on listener confusion and on RE success rate. If gaze was not a reliable indicator of RE interpretation, this system would frequently give misleading feedback and therefore perform worse. Together with the finding that positive gaze-based feedback leads to shorter trial times than positive movement-based feedback, while negative gaze-based feedback leads to better success rates than negative movement-based feedback, this confirms our second hypothesis. That is, the eyetracking system (positively) influences interaction effectiveness.

One observation from the games in the experiment is that listeners tend to rapidly look back and forth between different buttons when they are confused. However, it needs to be still worked out, how to interpret such signals more generally. A further issue is that all objects in the 3D world shift on the screen when the user turns or moves in the virtual environment. The user’s eyes will typically follow the object they are currently inspecting, but lag behind until the screen comes to a stop again. One topic for future work would be to remove such noise from the eyetracking signal.

Finally, the negative feedback our systems gave was very unspecific (“no, not that one”, even when there were other distractors) and given earlier and numerically also more frequently by the eyetracking system. This could explain the different effects of positive and negative feedback on inspection behavior and the time-accuracy trade-off for each system: Longer trial times but better success rates for negative gaze-based (compared to movement-based) feedback. We used negative feedback to keep the experimental situation more controlled but the performance of the feedback systems could possibly be improved by giving more specific feedback (“no, the BLUE button”). Another avenue for future research is to examine whether listener gaze could also be useful for other NLG or dialog tasks apart from RE generation.



## Conclusion

We reported on an experiment in which an NLG system used listener gaze to track the listener's understanding of REs and provide positive or negative feedback when needed. This shows that listener gaze provides consistent and useful feedback about the listener's interpretation process, and that NLG systems can be improved by tracking this interpretation process in real time.

These findings have consequences both for psycholinguistics and for computational linguistics. On the psycholinguistic side, they open the way for eyetracking experiments that are set in a more natural and dynamic, and importantly, truly interactive, environment than traditional Visual World experiments. On the computational side, they offer a testbed for interactive NLG and dialogue systems; even though eyetracking devices are not yet commonplace as computer peripherals they can still allow us to implement and test theories of how to effectively track the comprehension process of the user.

## Acknowledgments

The research reported of in this paper was partly supported by the "Multimodal Computing and Interaction" Cluster of Excellence at Saarland University. We thank Irena Dotcheva for help with data collection as well as Alexandre Denis and Christoph Clodo for software support.

## References

- Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264.
- Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.
- Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1), 62–81.
- Denis, A. (2010). Generating referring expressions with reference domain theory. In *Proceedings of the 6th International Natural Language Generation Conference*.
- Foster, M. E. (2007). Enhancing human-computer interaction with embodied conversational agents. In *Proceedings of HCI International 2007*.
- Gargett, A., Garoufi, K., Koller, A., & Striegnitz, K. (2010). The GIVE-2 Corpus of Giving Instructions in Virtual Environments. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*.
- Garoufi, K., & Koller, A. (2011). The Potsdam NLG systems at the GIVE-2.5 Challenge. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation (ENLG)*.
- Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11, 274–279.
- Hanna, J., & Brennan, S. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57, 596–615.
- Hülsmann, F., Dankert, T., & Pfeiffer, T. (2011). Comparing gaze-based and manual interaction in a fast-paced gaming task in virtual reality. In *Virtuelle & Erweiterte Realität, 8. Workshop der GI-Fachgruppe VR/AR*.
- Iida, R., Yasuhara, M., & Tokunaga, T. (2011). Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In *Proceedings of 5th International Joint Conference on Natural Language Processing*.
- Knoeferle, P., Crocker, M. W., Pickering, M., & Scheepers, C. (2005). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition*, 95, 95–127.
- Koller, A., Striegnitz, K., Byron, D., Cassell, J., Dale, R., Moore, J., et al. (2010). The First Challenge on Generating Instructions in Virtual Environments. In E. Krahmer & M. Theune (Eds.), *Empirical Methods in Natural Language Generation* (pp. 337–361). Springer.
- Kreysa, H., & Knoeferle, P. (2011). Peripheral speaker gaze facilitates spoken language comprehension: syntactic structuring and thematic role assignment in German. In B. Kokinov, A. Karmiloff-Smith, & N. Nersessian (Eds.), *Proceedings of the European Conference on Cognitive Science 2011*.
- Kuriyama, N., Terai, A., Yasuhara, M., Tokunaga, T., Yamagishi, K., & Kusumi, T. (2011). Gaze matching of referring expressions in collaborative problem solving. In *Proceedings of International Workshop on Dual Eye Tracking in CSCW (DUET)*.
- Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, 29(6), 1045–1060.
- Schröder, M., & Trouvain, J. (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *International Journal of Speech Technology*, 6, 365–377.
- Staudte, M., & Crocker, M. W. (2011). Investigating joint attention mechanisms through human-robot interaction. *Cognition*, 120(2), 268–291.
- Striegnitz, K., Denis, A., Gargett, A., Garoufi, K., Koller, A., & Theune, M. (2011). Report on the Second Second Challenge on Generating Instructions in Virtual Environments (GIVE-2.5). In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*.
- Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632–1634.

# The effect of metabolic loading on statistical learning

**David Stevens (david.stevens@sydney.edu.au)**

Discipline of Exercise and Sport Science, East Street,  
Lidcombe, N.S.W., 2141, Australia.

**Joanne Arciuli (joanne.arciuli@sydney.edu.au)**

Discipline of Speech Pathology, East Street,  
Lidcombe, N.S.W., 2141, Australia.

**David I. Anderson (danders@sfsu.edu)**

Department of Kinesiology, 1600 Holloway Avenue,  
San Francisco, CA 94132, U.S.A.

**A. Mark Williams (m.williams@ljmu.ac.uk)**

School of Sport and Exercise Science, 15-21 Webster Street,  
Liverpool, L3 2ET, U.K.

## Abstract

We investigated whether concurrent exercise would affect statistical learning (SL). During familiarization, participants were exposed to pictures that appeared sequentially, in a seemingly random fashion. In fact, the pictures were grouped into triplets. In the surprise test phase, participants identified triplets they had seen during familiarization. There were three groups: a group that performed familiarization seated on an exercise bike (CON), a group that performed familiarization while engaged in resistance free cycling (RF), and a group that performed familiarization while cycling at 60% of maximum effort (EX). The CON group correctly identified 72% of triplets in the test phase. The RF and EX groups correctly identified 61% and 55%, respectively. Only the CON group demonstrated performance that was significantly greater than chance. The RF group only just failed to demonstrate significant SL. Thus, concurrent exercise can suppress SL. Work is underway to determine the mechanism by which such suppression occurs.

**Keywords:** Statistical learning; exercise; metabolic loading.

## Introduction

Statistical learning (SL) refers to the ability to detect statistical regularities implicitly and use this information to guide related behavior. Since the seminal research conducted by Saffran, Aslin and Newport (1996) demonstrating that 8-month old infants are capable of learning syllables embedded in an auditorily presented sequence, this area of research has increased rapidly. SL has now been shown to operate on sequential auditorily presented stimuli (Aslin, Saffran & Newport, 1998; Saffran et al., 1996; Saffran et al., 1997) and sequential visually presented stimuli (Arciuli & Simpson, 2011; Fiser & Aslin, 2001; Turk-Browne & Scholl, 2009). It has been shown to operate on adjacent dependencies (Lany & Gómez, 2008; Newport & Aslin, 2004) and non-adjacent dependencies (Creel, Newport & Aslin, 2004; Gebhart, Newport & Aslin, 2009). SL operates on spatial as well as temporal regularities (Turk-Browne & Scholl, 2009). Recently,

researchers have linked a capacity for SL with key cognitive activities such as spoken language proficiency (Conway, Karpicke & Pisoni, 2007; Evans, Saffran & Robe-Torres, 2009) and reading (Arciuli & Simpson, in press).

In short, SL is a robust process that develops early in life and is linked with a wide range of cognitive and perceptual activities. As such, it is important to understand whether SL can be enhanced or suppressed. While there have been a handful of studies investigating brain activity during SL (Turk-Browne, Scholl, Chun & Johnson, 2009; Turk-Browne, Scholl, Johnson & Chun, 2010), little is known about the physiological processes that underpin the capacity for SL. In this study, we sought to learn more about these physiological underpinnings by examining the effects of metabolic loading (i.e., exercise) on SL.

It is well known that exercise can affect brain activity and cognitive function; however, these effects can operate in different ways. Cognition appears to be enhanced over the longer term as a result of *ongoing* exercise (Hill, Storandt & Malley, 1993; Ratey & Loehr, 2011). Researchers studying the behaviors of animals have demonstrated exercise-related increases in brain-derived neurotrophic factors as well as increasing levels of neural plasticity of brain tissue (for a comprehensive review, see Hertzog, Kramer, Wilson & Lindenberger, 2009).

There is conflicting evidence related to cognition during *acute* exercise, with several researchers finding an enhancement (Chmura, Nazar, Kaciuba-Uscilko, 1994; Yagi, Coburn, Estes & Arruda, 1999), others finding suppression (Audiffren, Brisswalter, Brandet & Bosquet, 1998; Dietrich & Sparling, 2004), and some finding no difference (Davranche et al., 2006; Dietrich & Sparling, 2004; McMorris et al, 2003;). Yagi et al. (2006) suggested that neural activation in the brain is enhanced by exercise, whereas Dietrich and Sparling (2004) suggested that exercise causes function in the prefrontal lobe to become depressed during exercise; hence, they found suppression in some cognitive tests, but no effect in others.

Importantly, existing research has used tasks of explicit cognitive processing to assess performance. The task that has been most commonly utilized is the choice reaction time task (e.g. Audiffren et al., 1998; McMorris et al., 2003). However, a range of cognitive tasks have been used. For example, Dietrich and Sparling (2004) used a mathematics test (to test pre-frontal function) and a vocabulary test (which they suggested was a test of non pre-frontal function). We are not aware of research on the effects of exercise on implicit cognitive processing. Specifically, to date, no researchers have examined the impact of exercise on the brain's capacity to learn statistical regularities.

The primary aim of this experiment was to examine how metabolic loading affects concomitant SL. The exploratory nature of this research meant that directional hypotheses could not be easily formulated. It seemed likely that concomitant exercise might suppress SL in comparison with a non-exercising control group; however, it was important to control for the possibility that some kind of dual-task loading, rather than the exercise itself, might also affect SL. Thus, we included an additional condition where participants performed the SL task while engaging in resistance-free cycling.

## Methods

### Subjects

A total of 24 participants (age  $24 \pm 3.3$ ; 14 females,  $\dot{V}O_{2\max} = 47.8 (\pm 4.9) \text{ ml.kg}^{-1}.\text{min}^{-1}$ , no known neurological or physical problems) were recruited from the University of Sydney population. Ethics approval was granted by the institution.

### Statistical learning task

Participants undertook the embedded triplet learning task created by Arciuli and Simpson (2011; 2012). The task is comprised of a *familiarization* and then a surprise *test* phase, controlled by Eprime presentation software (v2.0, Psychology Software Tools, PA, U.S.A.). Stimuli were eighteen cartoon-like figures sourced from the website <http://www.clipartconnection.com/en/>. Six were used exclusively for instruction and practice. The remaining twelve appeared only in the familiarization and test phases. None resembled real-world animals, people or popular cartoon characters. The twelve stimuli used for the familiarization and test phases can be found in the Appendix sections of Arciuli and Simpson (2011; 2012). These twelve experimental stimuli were divided into four groups of three (four base triplets), hereafter referred to as *ABC*, *DEF*, *GHI* and *JKL* (see Appendix 1).

The familiarization phase consisted of a continuous stream of stimuli, with each cartoon character shown in isolation in the centre of the display against a white background. Each was visible for 800msec with an inter-stimulus-interval (ISI) of 200msec. Each base triplet was

selected for inclusion 24 times each (resulting in a total of 96 triplets). For six of these 24 instances, one of the cartoon characters was presented twice in a row in order to provide a cover task (detection of repeated characters). Detection of these repeated characters was the cover task during familiarization. This cover task ensured that participants paid attention to the familiarization stream because participants were required to watch the screen and press a button whenever they saw a repeated character. Repetitions were counterbalanced among and within each triplet. So, for example, the repetitions for base triplet *ABC* meant there were two occurrences of the sequence *AABC*, two occurrences of the sequence *ABBC*, and two occurrences of *ABCC* (along with 18 occurrences of the sequence *ABC*). This procedure meant that for each base triplet the strict triplet structure was violated (e.g., *ABBC*) on two of twenty-four occasions. The repetition was done in this way (all three items in a triplet used for repetition) to ensure that the repetitions did not inadvertently cue the participants to the existence of the triplet boundaries. The familiarization phase consisted of 312 individual characters, with each of the 12 characters appearing 26 times each. The order of the triplets within the familiarization phase was randomized. The sole restriction was that the same base triplet could not appear consecutively (e.g., *ABCABC*).

In both of the previous studies reported by Arciuli and Simpson the familiarization phase was followed immediately by the surprise test phase. However, in the present study, the surprise test phase was given to participants after a time delay of 5 minutes. This was necessary because participants required time to move from the bike to a desk where they performed the surprise test phase.

The surprise test phase consisted of 64 trials with each trial containing two triplets: one of the four base triplets and one of four impossible triplets. These impossible triplets never appeared in the familiarization stream and each was created by taking one character from three different base triplets (e.g., *AEI*, *DHL*, *GKC* and *JBF*). The stimuli in the test trials were presented individually with the same duration and ISI as was used in the familiarization stream. A 1,000 msec gap separated the two triplets in each test trial. After all six characters had been presented participants were prompted to identify which of the two triplets had appeared previously (during familiarization). This procedure constituted a 2-alternative forced-choice task (2AFC). The presentation order of base triplets and impossible triplets was counterbalanced. Across the 64 test trials each base triplet and each impossible triplet was seen 16 times, and each individual character was seen 32 times. Participants received a different random order for the test trials.

### Procedure

Participants performed a  $\dot{V}O_{2\max}$  test on a Lode Cycle ergometer. The  $\dot{V}O_{2\max}$  test required participants to cycle until fatigued. They started at a low intensity and it was

increased every minute until either the participant chose to stop or the investigator ended the test (for safety or diagnostic reasons). Heart rate was collected with a heart rate monitor (Polar) and  $\dot{V}O_2$  were collected throughout the test using an electronic metabolic cart (MedGraphics) (Thompson, Gordon & Pescatello, 2009).

Participants were randomly allocated into 3 groups, with 8 participants in each; a control group (CON) who performed *familiarization* while seated, a resistance free group (RF) who performed familiarization while free pedaling on the cycle, and the exercise group (EX), who performed the familiarization whilst cycling at their own 60%  $\dot{V}O_{2\max}$  to ensure participants in this group were exercising at equivalent levels of intensity (i.e., relative to their personal  $\dot{V}O_{2\max}$ ). This power output was chosen because it is a commonly used power output in studies of exercise and conscious cognition (Arcelin & Brisswalter, 1999; Pesce, Capranica, Tessitore & Figura, 2002; Pontifex & Hillman, 2006).

All participants underwent the familiarization phase while seated on the cycle so that all had the same level of postural discomfort. The Lode Cycle ergometer had the function of allowing a power output to be set which would be automatically maintained even when cadence changed. This was particularly important for the EX as it ensured participants could maintain attention on the SL task without having to focus on maintaining a certain cadence. For the cover task during familiarization, participants were asked to respond by pressing a button that was placed on the handlebars. This button was connected to an E-prime response box which was not fixed to the handle bar, so the participant could choose their preferred hand and move the button into a position for comfort. Each participant in the EX group cycled at their respective 60%  $\dot{V}O_{2\max}$  values for the duration of the familiarization phase only. These participants cycled for 5 minutes prior to the commencement of the familiarization, during which time metabolic data was collected. Metabolic recording equipment was removed before beginning the familiarization phase so as to not discomfort the participants. Instead, continuous heart rate monitoring was used during familiarization. Upon completion of the familiarization phase, all groups were given a 5 minute rest period before the testing phase. For all participants the surprise *test* phase was performed at a desk while seated on a chair.

### Data Analysis

A participant was considered to have learnt about the embedded regularities in the familiarization phase if they identified more than 50% of the triplets (above chance level). Thus, one-sample t-tests were used to determine whether there was significant SL for each group. This was the same data analysis technique used by Arciuli and Simpson (2011; 2012). Overall averages were then compared across the three groups using a one-way ANOVA

and follow-up t-tests. An  $\alpha$  level of .05 was set for the experiment.

## Results

First, detection of repeated aliens during familiarization (the cover task) was examined. These data are reported in Table 1. Participants appeared to be paying similar, high levels of attention during familiarization across each of the three experimental conditions. A one-way ANOVA was conducted on the accuracy of detecting repeated characters during the *familiarization phase*. No significant difference was found ( $F_{(2,21)} < 1$ ), across the three groups.

Table 1: Accuracy of responding during *Familiarization* ( $\pm$ SD).

Group	Number of repeated characters identified
CON	97.4% ( $\pm$ 2.2%)
RF	98.4% ( $\pm$ 3.1%)
EX	99.5% ( $\pm$ 1.5%)

The overall degree of SL for each group during the surprise test phase is displayed in Table 2. One-sample t-tests revealed that only the CON group showed significant SL by demonstrating a level of performance that was significantly greater than chance (i.e., greater than 50%),  $t_{(7)} = 4.175$ ,  $p < .05$ . The RF group just failed to demonstrate statistical learning,  $t_{(7)} = 2.311$ ,  $p = .054$  whereas the EX group did not demonstrate statistical learning,  $t_{(7)} = 1.050$ ,  $p = .329$ .

Table 2: Degree of SL during *Test* phase ( $\pm$ SD).

Group	Percentage triplets identified
CON	72% ( $\pm$ 14%)
RF	61% ( $\pm$ 14%)
EX	55% ( $\pm$ 13%)

A one-way ANOVA demonstrated that there was no significant difference amongst these means  $F_{(2,21)} = 3.011$ ,  $p = .071$ ; however, given the marginal significance, post-hoc t-tests were conducted. Results demonstrated that the only significant difference was between the CON and EX groups,  $t_{(14)} = 2.337$  ( $p = .035$ ).

## Discussion

The results reported here provide several insights into statistical learning (SL). First, significant SL was observed in the CON group. This finding adds to a growing body of evidence indicating that humans can implicitly learn regularities embedded in input, even while undertaking a cover task. A novel finding was that metabolic loading appears to impact upon participants' ability to detect these statistical regularities. This effect occurs even during a relatively short bout of exercise. The RF group showed a higher degree of SL compared with the EX group but less

learning compared with the CON group. SL in the RF group just failed to reach significance, potentially due to low statistical power. The latter finding may suggest a graded pattern of performance whereby the mental effort associated with dual-task requirements (cycling resistance-free while also performing the SL task) affects SL. We interpret this data to demonstrate that even without any metabolic interference on the brain, the task of simply moving the legs, which in itself is a seemingly automatic task, appears to reduce the level of SL. This reduction appears to be amplified when metabolic loading is introduced.

Given the uncertainty surrounding the mechanisms underpinning explicit cognitive function during acute exercise, and given the possible differences between explicit and implicit cognition, it is difficult to know what caused the reduction in SL we observed in our EX group.

As exercise only lasted for around 15 minutes, at what is classed 'moderate intensity', and given that all participants had eaten and were hydrated, it is highly unlikely that metabolic factors such as dehydration or low blood glucose would have contributed to the results. It seems more likely that some form of psychological or neurological interference induced by exercise may have suppressed SL.

From a psychological standpoint, a similar phenomenon was demonstrated by Audiffren et al. (2008) where reaction time significantly increased during exercise compared to rest. This phenomenon has since been referred to as resource allocation competition. Pontifex and Hillman (2007) demonstrated that EEG activation is reduced when cycling, reflecting an 'inefficiency' of resource allocation. Specifically, they demonstrated reduced N1 amplitude, which has been shown to be part of the visual discrimination resource component (Vogel & Luck, 2000). Turk-Browne et al. (2009) conducted an fMRI study where participants performed a similar statistical visual learning task to the one presented here and found that activation in the occipital regions was increased (along with significant SL). If cycling suppresses some function in the visual cortex, sensitivity to visually presented statistical regularities may be affected. We aim to conduct the same experiment using an auditorily presented version of this SL task to further investigate this hypothesis.

There is another possibility, related to a more general effect of acute exercise on the brain. A review by Williamson, Fadel and Mitchell (2006) highlighted that there is a high neuroelectrical resource demand on the brain to maintain central control over metabolic reactions, such as vascular control for blood pressure, despite these processes being autonomic. Pontifex and Hillman (2007) suggested that additional neuroelectric demands further increase the load associated with dual tasking within the brain. Given that the RF group had the added demand of the resistance free cycling, and the EX group had the even greater demand of cycling plus central metabolic control, this may help to explain why there appeared to a graded pattern of results.

These findings open up many avenues for further research and as such we are currently undertaking several additional studies. To investigate whether exercise causes changes in both implicit and explicit tasks measures of cognitive performance, we will use the same exercise paradigm whilst participants undertake an explicit working memory task, such as digit span. We are also planning to compare the effects of exercise on implicit versus explicit versions of the embedded triplet task. In addition, we are exploring the effect of different exercise intensities so that we can determine at what metabolic level impairments in begin to SL occur. As mentioned, we will also examine whether exercise affects SL similarly regardless of whether stimuli are presented visually or auditorily. It is also interesting to ponder what would happen if we were to reverse the timing of the exercise in order to study how metabolic loading during the *test* phase affects answers after completing the familiarization phase at rest.

## Acknowledgments

The first author is in receipt of an Australian Postgraduate Award scholarship. Additional funding was provided by the University of Sydney Postgraduate Research Support Scheme. Thanks goes to Ian Simpson, PhD., for his help with E-Prime.

## References

- Arcelin, R. & Brisswalter, J. (1999). Performance stability in simultaneous tasks of pedaling and reaction time. *Perceptual and Motor Skills*, 88, 1193-1199.
- Arciuli, J. & Simpson, I. C. (2011). Statistical learning in typically developing children: the role of age and speed of stimulus presentation. *Developmental Science*, 14, 464-473.
- Arciuli, J., & Simpson, I. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science*, 36, 286-304.
- Aslin, R. N., Saffran, J. R. & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9, 321-324.
- Audiffren, M., Brisswalter, J., Brandet, J. P. & Bosquet, L. (1998). The relation of exercise intensity to attention deficit: Analysis of a cycling task. *Science & Sports*, 13, 81-83.
- Brisswalter, J., Arcelin, R., Audiffren, M. & Delignieres, D. (1997). Influence of physical exercise on simple reaction time: effect of physical fitness. *Perceptual and Motor Skills*, 85, 1019-1027.
- Chmura, J., Nazar, K. & Kaciuba-Uscilko, H. (1994). Choice reaction times during graded exercise in relation to blood lactate and plasma catecholamine thresholds. *International Journal of Sports Medicine*, 15, 172-176.
- Conway, C. M., Karpicke, J. & Pisoni, D. B. (2007). Contribution of implicit sequence learning to spoken language processing: Some preliminary findings with hearing adults. *Journal of Deaf Studies and Deaf Education*, 12, 317-334.

- Creel, S.C., Newport, E. L. & Aslin, R. N. (2004). Distant melodies: statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30, 1119-1130.
- Davranche, K., Audiffren, M. & Denjean, A. (2006). A distributional analysis of the effect of physical exercise on choice reaction time task. *Journal of Sports Sciences*, 24, 323-329.
- Evans, J. L., Saffran, J. R. & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language and Hearing Research*, 52, 321-335.
- Fiser, J. & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 12, 499-504.
- Griffin, E. W., Mullally, S., Foley, C., Warmington, S. A., O'Mara, S. M. & Kelly, A. M. (2011). Aerobic exercise improves hippocampal function and increases in BDNF in the serum of young adult males. *Physiology and Behavior*, 104, 934-941.
- Hayes, N. A. & Broadbent, D. E. (1988). Two modes of learning for interactive tasks. *Cognition*, 28, 479-488.
- Hertzog, C., Kramer, A. F., Wilson, R. S. & Lindenberger, U. (2009). Enrichment effects of adult cognitive development. *Psychological Science in the Public Interest*, 9, 1-65.
- Hill, R. D., Storandt, M. & Malley, M. (1993). The impact of long-term exercise training on psychological function in older adults. *Journals of Gerontology*, 48, P12-P17.
- Hunt, R. H. & Aslin, R. N. (2001). Statistical learning in a serial reaction time task: access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General*, 130, 685-680.
- Lany, J. & Gómez, R. L. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science*, 19, 1247-1252.
- Misyak, J. B., Christiansen, M. H. & Tomblin, J. B. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science*, 2, 138-153.
- Newport, E. L. & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48, 127-162.
- O'Callaghan, R. M., Ohle, R. & Kely, A. M. (2007). The effects of forced exercise on hippocampal plasticity in the rat: A comparison of LTP, spatial- and non-spatial learning. *Behavioural Brain Research*, 176, 362-366.
- Perruchet, P. & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *TRENDS in Cognitive Sciences*, 10, 233-238.
- Pesce, C., Capranica, L., Tessitore, A. & Figura, F. (2002). Effects of a sub-maximal physical load on the orienting and focusing of visual attention. *Journal of Human Movement Studies*, 42, 401-420.
- Pontifex, M. B. & Hillman, C. H. (2006). Neuroelectric measurement of cognition during aerobic exercise. *Methods*, 45, 271-278.
- Pontifex, M. B. & Hillman, C. H. (2007). Neuroelectric and behavioral indices of interference control during acute exercise. *Clinical Neurophysiology*, 118, 570-580.
- Ratey, J. J. & Loehr, J. E. (2011). The positive impact of physical activity on cognition during adulthood: A review of underlying mechanisms, evidence and recommendations. *Reviews in Neurosciences*, 22, 171-185.
- Rathus, J. H., Reber, A. S., Manza, L. & Kushner, M. (1994). Implicit and explicit learning: differential effects of affective states. *Perceptual and Motor Skills*, 79, 93-133.
- Saffran, J. R., Aslin, R. N. & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Nature*, 374, 1926-1928.
- Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A. & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, 8, 101-105.
- Thompson, W. (Ed.). (2009). *ACSM's Guidelines for Exercise Testing and Prescription* (8<sup>th</sup> Ed.). Baltimore, MD: Lippincott Williams & Wilkins.
- Turk-Browne, N. B., Jungé, J. A. & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, 134, 552-564.
- Turk-Browne, N. B. & Scholl, B. J. (2009). Flexible visual statistical learning: Transfer across space and time. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 195-202.
- Turk-Brown, N. B., Scholl, B. J., Chun, M. M. & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience*, 21, 1934-1945.
- Turk-Brown, N. B., Scholl, B. J., Johnson, M. K. & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of Neuroscience*, 30, 11177-11187.
- Vogel, E. K. & Luck, S. J. (2000). The visual N1 component as an index of a discrimination task. *Psychophysiology*, 37, 190-203.
- Williamson, J. W., Fadel, P. J. & Mitchell, J. H. (2006). New insights into central cardiovascular control during exercise in humans: a central command update. *Experimental Physiology*, 91, 51-58.
- Yagi, T., Coburn, K. L., Estes, K. M. & Arruda, J. E. (1999). Effects of aerobic exercise and gender on visual and auditory P300, reaction time and accuracy. *European Journal of Applied Physiology and Occupational Physiology*, 80, 402-408.
- Zoladz, J. A. & Pilc, A. (2010). The effect of physical activity on the brain derived neurotrophic factor: from animal to human studies. *Journal of Physiology and Pharmacology*, 61, 533-541.

# Spaun: A Perception-Cognition-Action Model Using Spiking Neurons

Terrence C. Stewart (tcstewar@uwaterloo.ca)

Feng-Xuan Choo (fchoo@uwaterloo.ca)

Chris Eliasmith (celiasmith@uwaterloo.ca)

Centre for Theoretical Neuroscience, University of Waterloo,  
Waterloo, ON, N2L 3G1

## Abstract

We present a large-scale cognitive neural model called Spaun (Semantic Pointer Architecture: Unified Network), and show simulation results on 6 tasks (digit recognition, tracing from memory, serial working memory, question answering, addition by counting, and symbolic pattern completion). The model consists of 2.3 million spiking neurons whose neural properties, organization, and connectivity match that of the mammalian brain. Input consists of images of handwritten and typed numbers and symbols, and output is the motion of a 2 degree-of-freedom arm that writes the model's responses. Tasks can be presented in any order, with no "rewiring" of the brain for each task. Instead, the model is capable of internal cognitive control (via the basal ganglia), selectively routing information throughout the brain and recruiting different cortical components as needed for each task.

**Keywords:** Neural engineering; cognitive architecture; spiking neurons; cognitive control; whole-brain systems

## Introduction

In a forthcoming book, Eliasmith (2012) details a neural architecture for biological cognition called the *semantic pointer architecture* (SPA). This architecture, based on the Neural Engineering Framework (Eliasmith & Anderson, 2003), uses groups of spiking neurons to form distributed representations of high-dimensional vectors, which can in turn encode symbol-like tree structures. Synaptic connections between groups of neurons compute particular functions on those vectors, allowing high-level cognitive algorithms to be implemented in detailed spiking neuron models.

In this paper, we present an overview of the Semantic Pointer Architecture: Unified Network (Spaun) model and discuss its behaviour on six different tasks. We demonstrate that this biologically plausible spiking neuron model has the following features:

**Task Flexibility:** No changes are made to the model between tasks. Visual input indicates which task to do next.

**Motor Plans:** Model output provides a motor plan for a simple 2-joint arm, giving hand-written digits as responses.

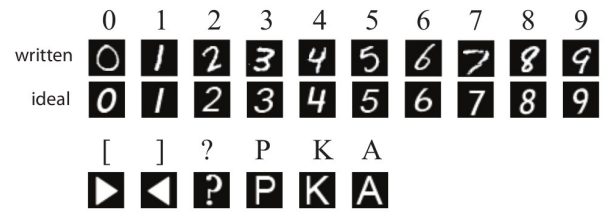
**Visual Memory:** Even after an input has been recognized and classified as a particular symbol, details of the original image can still be recovered and used.

**Compositionality:** Multiple items can be represented and reliably bound together, allowing for the creation and manipulation of symbol-tree-like structures.

**Symbolic Induction:** Language-like patterns in visual input can be discovered after only a few presentations, and used to guide subsequent responses.

Input to the model consists of idealized and hand-written digits and symbols (Figure 1a). These images are given as a 28x28 grid of pixels. This input domain has the advantage of including significantly variable, real-world input, while also providing a reasonably limited semantics that the model must reason about. Output from the model consists of the motion of a 2-degree-of-freedom arm. The neural model generates a sequence of target locations which directly drive the controller for the arm. This provides the model with its own handwriting output (Figure 1b).

### a) Input



### b) Output

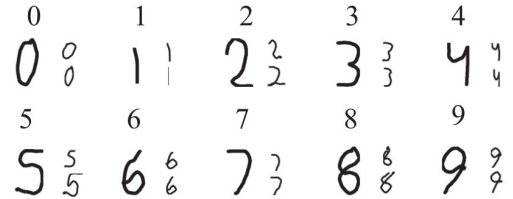


Figure 1: Example visual input and arm-movement output from Spaun. Input digits are from the MNIST database, and input symbols are used to inform the model of details of the current task. The model produces output by controlling a 2-joint arm, and variation in the internal representations produces the variation in its output hand-writing.

We can think of Spaun as having a single, fixed eye and a single 2-joint arm. The eye does not move, but instead the experimenter changes the image falling on it by showing different inputs over time, with each input shown for 150ms, followed by 150ms of blank background. To begin a specific task, Spaun is shown the letter "A" followed by a number between zero and seven. The subsequent input is then interpreted by the model in the context of the specified task and processed accordingly, resulting in arm movements that provide Spaun's response. All internal processing is performed using spiking neurons, with neural properties and connectivity consistent with the mammalian brain.



## Neural Engineering Framework

While complete details on the construction of Spaun are available elsewhere (Eliasmith, 2012; <<http://nengo.ca>>), it is based on our continuing work on the Neural Engineering Framework (NEF; Eliasmith & Anderson, 2003). The NEF is a generic method for converting high-level algorithms into realistic spiking neuron models.

Two basic principles of the NEF are that a) groups of neurons form distributed representations of vectors, and b) connections between groups of neurons specify a computation to be performed on those vectors. Importantly, the NEF provides a method for analytically solving for the synaptic connection weights that will efficiently compute any given function.

While the NEF supports any type of neural model, for Spaun we use Leaky Integrate-and-Fire (LIF) neurons. The various properties of the LIF model (refractory period, capacitance, resistance, post-synaptic time constant, etc.) are set to be consistent with known neurophysiological results for the various brain regions modelled.

To represent a vector using a group of neurons, the NEF generalizes the idea of preferred direction vectors. Each neuron in a group has a randomly chosen vector  $\mathbf{e}$  for which it will fire most strongly. In particular, the amount of current  $J$  flowing into the neuron is the dot product of the preferred vector  $\mathbf{e}$  with the represented value  $\mathbf{x}$ , times the neuron's gain  $\alpha$ , plus the background current  $J_{bias}$  (Eq. 1).

While Eq. 1 lets us convert  $\mathbf{x}$  into neural activity, we can also do the opposite by computing  $\mathbf{d}$  via Eq. 2. This produces a set of linear decoding weights that can be multiplied by the activity of each neuron in the group. The result is the optimal least-squares linear estimate of  $\mathbf{x}$ . Thus, given a spiking pattern we can estimate what value is currently represented by those neurons.

Most crucially, we can use  $\mathbf{d}$  to calculate the synaptic connection weights that will compute particular operations. To compute a linear operation where one group of neurons represents  $\mathbf{x}$  and a second group should represent  $\mathbf{M}\mathbf{x}$ , where  $\mathbf{M}$  is an arbitrary matrix, we set the connection weights between neuron  $i$  in the first group and neuron  $j$  in the second group to  $\omega_{ij}$  as per Eq. 3. For non-linear operations, we need to compute a new set of  $\mathbf{d}$  values via Eq. 4.

$$J = \alpha \mathbf{e} \cdot \mathbf{x} + J_{bias} \quad (1)$$

$$\mathbf{d} = \Gamma^{-1} \mathbf{Y} \quad \Gamma_{ij} = \int a_i a_j dx \quad \mathbf{Y}_j = \int a_j \mathbf{x} dx \quad (2)$$

$$\omega_{ij} = \alpha_j \mathbf{e}_j \cdot \mathbf{M} \mathbf{d}_i \quad (3)$$

$$\mathbf{d}^{f(\mathbf{x})} = \Gamma^{-1} \mathbf{Y} \quad \Gamma_{ij} = \int a_i a_j dx \quad \mathbf{Y}_j = \int a_j f(\mathbf{x}) dx \quad (4)$$

This approach allows us to convert a high-level algorithm written in terms of vectors and computations on those vectors into a detailed spiking neuron model. Importantly, this approach works for recurrent connections as well. For example, we can implement memory by connecting a group of neurons back to itself with connections weights determined by Eq. 3 where  $\mathbf{M}$  is the identity matrix. If that group of neurons is currently representing  $\mathbf{x}$ , then given no external input it will drive itself to keep representing  $\mathbf{x}$ , thus storing information over time.

## Spaun

The Semantic Pointer Architecture: Unified Network model consists of multiple modules, depicted in Figure 2. These modules are considered to be cortical and subcortical areas that implement different operations. All components consist of LIF neurons connected via synaptic weights (Eq. 3), but each area computes a different set of functions.

To perform a particular task, information must be selectively routed between cortical areas, as each task uses a different subset of the components. This is achieved through an action selection system modelled after the mammalian basal ganglia and thalamus. We have previously shown that this model matches the anatomy and timing behaviour of the basal ganglia (Stewart, Choo, & Eliasmith, 2010) and provides enough flexibility to perform planning and problem solving in our model of the Tower of Hanoi task (Stewart & Eliasmith, 2011).

The model presented here is the first use of this neural action selection system with multiple tasks and detailed perceptual-motor systems. The action selection system is a neural production system, allowing us to write rules of the form “if cortical area  $X_1$  matches the vector  $\mathbf{a}$  and cortical area  $X_2$  matches the vector  $\mathbf{b}$ , then send vector  $\mathbf{c}$  to area  $X_3$  and route the vector from area  $X_4$  to area  $X_5$ ”. These rules are implemented by converting the rules into functions, applying Eq. 3, and using the resulting synaptic connection weights between the cortex, basal ganglia, and thalamus.

Importantly, the set of rules is fixed across all the tasks, giving a single, unified model. All input to Spaun is through its perceptual system, and all behavioral output from its motor system. The representational repertoire, background knowledge, cognitive mechanisms, neural mechanisms, etc. remain untouched while the system performs any of the tasks in any order.

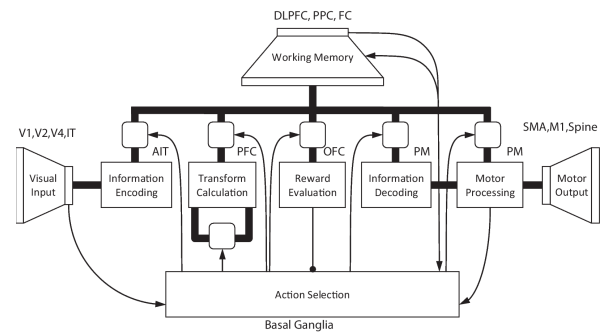


Figure 2: The Spaun architecture

The Spaun model (Figure 2) consists of three hierarchies, an action selection mechanism, and five subsystems. The first hierarchy is the visual system, which compresses an input image into a high-level abstract representation of that input. We adapt Hinton's (2010) Deep Belief Network to use LIF spiking neurons (via the NEF) and use it to compress a 28x28 image into a 50-dimensional vector we refer to as a *semantic pointer*: it is semantic because the high-level representation maintains similarity relationships from the image space; and it is a pointer because, as we will

see, the system can recover the original information from the compressed form. Similarly, Spaun includes a motor hierarchy which dereferences an output semantic pointer representing a number into a motor plan to drive a two degree-of-freedom arm (DeWolf, 2010). A third internal hierarchy (discussed in more detail below in the *serial working memory* section) forms a working memory capable of binding and unbinding arbitrary semantic pointers, providing the compositionality that is crucial for complex cognition. The working memory component also provides stable representations of intermediate task states, task subgoals, and context. Anatomically, these functions cover large portions of prefrontal and parietal cortex.

The five subsystems, from left to right in Figure 2, are used to: 1) map the visual hierarchy output to a conceptual representation as needed (**information encoding**); 2) extract relations between input elements (**transformation calculation**); 3) evaluate the reward associated with the input (**reward evaluation**); 4) map output items to a motor semantic pointer (**information decoding**); and 5) control motor timing (**motor processing**). Several of the subsystems and hierarchies consist of multiple components needed to perform the identified functions. For instance, the working memory subsystem includes eight distinct memory systems, each of which can store semantic pointers. Overall, the model uses 2,341,212 spiking leaky integrate-and-fire (LIF) neurons. Additional details necessary to fully reimplement the model, and a downloadable version of the model can be found at <<http://nengo.ca>>.

## Digit recognition

The simplest task for Spaun is digit recognition. We present the input sequence **A1[X?**, where **X** is a randomly chosen hand-written digit from the MNIST database. Each symbol is shown for 150ms, with 150ms between symbols. To perform this task, Spaun includes synaptic connections in the action selection system to implement the following algorithm:

- If visual input matches **A**, store **?** in the *state* area of the working memory. (This tells the system it is about to start a new task.)
- If **?** matches *state*, route the output of the visual encoding system to the state area of working memory. (This identifies and remembers which task is to be performed.)
- If the visual input is **1**, activate the routing between the information encoding system and the general-purpose working memory. (This tells Spaun to store whatever digits appear next.)
- If the visual input is **?** and *state* is 1, route the pattern in working memory to the motor system.

Responses from this model for different inputs **X** are shown in Figure 3. Recognition accuracy is 94%, which compares well to humans on this task (~98%; Chaaban and Scheessele, 2007). The model is thus capable of correctly categorizing over a wide range of input variability.

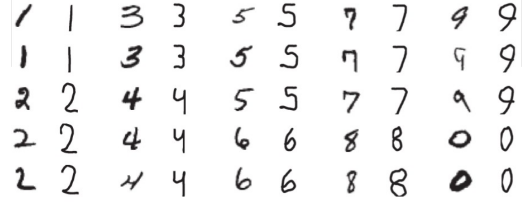


Figure 3: Input-output pairs for 20 different inputs. Each input (on the left) is correctly recognized by the model, which produces the output (on the right) via motor control.

## Tracing from memory

While the previous task showed that the model can treat very different stimuli as tokens of the same type, we also want the model to be able to be sensitive to the variations within a type. To demonstrate this ability, we ask the model to do digit recognition, but to draw its response *in the same style as the original input*.

This is implemented by defining, for each digit, five different motor control sequences that draw five visually distinct versions of that digit. If the input pattern is **X** and the motor sequence is **Y**, then we can define the tracing function  $f(X)=Y$ . We build a neural connection between the vision system and the motor system via Eq. 3, and allow it to be selectively controlled via the action selection in the basal ganglia. If we now present new input  $X_{\text{new}}$  that was not among the 5 original inputs, the resulting  $Y_{\text{new}}$  value will be the model's linear extrapolation of what motor sequence would be appropriate for that novel input pattern.

To use this system, we use the sequence **A0[X?** and add a single rule to the basal ganglia action selection system:

- If the visual input is **?** and *state* is 0, route the pattern in working memory to the motor system via the tracing function.

This rule, combined with the previous ones for digit recognition, result in the behaviour shown in Figure 4. Note that it is capable of drawing 2's both with and without loops, 6's where the loops join in different locations, and generally following the slanting of the digits. It is not a perfect reconstruction of the original input, but it does demonstrate that while the internal neural representation of all 2's are very similar (as shown by Spaun's success in the previous task), there are still variations in the representation due to different visual features, and those variations can be used to successfully drive behaviour.

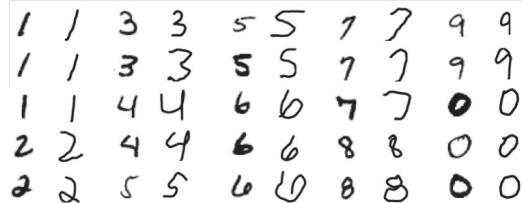


Figure 4: Input-output pairs for tracing from memory. The inputs (on the left) are recognized by the model, and then recreated from memory based on representational similarity to five previously known example pairs for each digit.

## Serial Working Memory

For this task, Spaun is given a list of numbers and must repeat them back, in order. The algorithm used here is based on our previous work (Choo & Eliasmith, 2010) on a special-purpose serial memory model.

The NEF gives us a method for representing vectors via spiking neurons. The Semantic Pointer Architecture approach maps symbols onto particular vectors, so one vector might represent ONE while another vector represents TWO. Each vector, when represented, would produce a distinct firing pattern in that population of neurons.

To represent an ordered sequence, we cannot simply add the vectors together, since we could not distinguish ONE+TWO from TWO+ONE. Instead, we can perform *binding* by creating a new vector from two different vectors. We start by introducing new vectors for different positions: P1, P2, P3, etc. We then store the sequence ONE, TWO by representing the vector  $\text{ONE} \otimes \text{P1} + \text{TWO} \otimes \text{P2}$ . Many mathematical operations can be used for  $\otimes$ ; we choose circular convolution since it is easy to accurately implement in spiking neurons using the NEF. This approach to representation using vectors is generally known as a Vector Symbolic Architecture (Gayler, 2003), and has been shown to scale well to adult-level vocabulary and grammar.

The action selection rules for this task are based on the previous one, with the addition of the **]** marker to indicate the end of a sequence. Figure 5 shows not only the input and output of the model, but also the ongoing spiking behaviour in various areas during the execution of the task.

Figure 5 also shows a method for interpreting what is currently being represented in a particular cortical area. The second working memory line shows how similar the current pattern in the working memory is to various different ideal patterns<sup>1</sup>. In particular, after the presentation of the final digit ( $t=2$  seconds), the value being represented is similar to  $\text{FOUR} \otimes \text{P1}$ ,  $\text{THREE} \otimes \text{P2}$ ,  $\text{TWO} \otimes \text{P3}$ , and  $\text{SIX} \otimes \text{P4}$ . In other words, this one group of neurons is capable of storing any arbitrary sequence of digits. As the sequence gets longer, accuracy decreases, and both primacy and recency effects are seen (Choo & Eliasmith, 2010).

## Question Answering

In addition to simply repeating a sequence, Spaun is capable of answering questions about the sequence. In particular, it can identify which digit is at a given location, and it can identify the location of a given digit.

This is accomplished by adjusting the transformation which takes the contents of working memory and routes it to the motor area. Given the vector  $S = \text{FIVE} \otimes \text{P1} + \text{SIX} \otimes \text{P2}$ , we can find the digit in position 1 by computing  $S \otimes \text{P1}$ , where  $\otimes$  is circular correlation, since  $S \otimes \text{P1} \approx \text{FIVE}$ . The accuracy of this approximation is dependent on the length of the sequence, the number of neurons used, and the dimensionality of the vectors.

<sup>1</sup> Formally, this is the dot product between the ideal vector for that symbol and the decoded value found using Eq. 2.

To implement this task, the sequence is presented in the same manner as the serial working memory task. We then present the symbol P for a query based on position, or the symbol K for a query based on the kind of digit to look for in the list. The action selection rules route the appropriate transformation vector to the working memory area, providing the required information to the motor system.

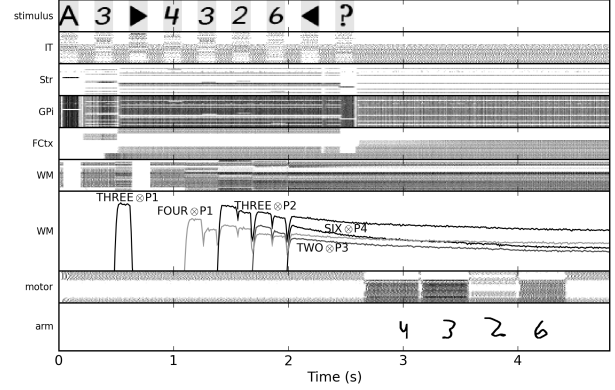


Figure 5: Serial recall of the sequence 4,3,2,6. Infero-temporal cortex (IT) holds the compressed representation of the visual input. Striatum (Str) activity determines how good a match each rule is to the current state. Globus pallidus internus (GPi) performs action selection, inhibiting all but the current best-matching rule. Frontal cortex (FCtx) holds task information, and working memory (WM) stores the list as a single vector:

$$\text{FOUR} \otimes \text{P1} + \text{THREE} \otimes \text{P2} + \text{TWO} \otimes \text{P3} + \text{SIX} \otimes \text{P4}.$$

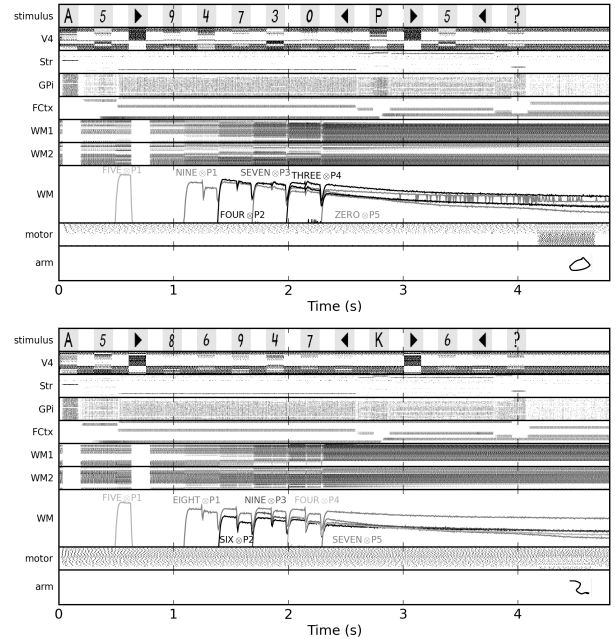


Figure 6: Answering questions about a list. The first case presents the list 9,4,7,3,0 and asks what is position 5. The model correctly answers 0. The second case presents the list 8,6,9,4,7 asks where the 6 can be found. The model correctly answers that it is in location 2.

## Addition by Counting

For the fifth task, we show that Spaun is capable of performing sequences of internal actions, where there are a multiple steps to go through before producing a final output. This is demonstrated here by performing mental addition via counting. That is, to compute  $4+3$ , the model must go through the steps of counting 4, 5, 6, and 7, entirely internally, and then finally producing the output 7.

Spaun achieves this by having multiple, general-purpose working memories. We use the first of these (WM1; the same neurons that stored the list and the recognized numbers in the previous task) to store the current value. A second group of neurons (WM2) stores the number of counting steps that are needed, and a third group (WM3) stores how many steps have been made.

Figure 7 shows Spaun performing this task over time for the specific case of  $4+3$ . Importantly, the model produces accurate results for any single-digit addition. Furthermore, Spaun exhibits the expected linear relationship between subvocal counting and response times, as seen in human subjects (Cordes et al., 2001). That is, each counting step requires  $419 \pm 10$ ms, which is within the empirical range of  $344 \pm 135$ ms for subvocal counting (Landauer, 1962).

Successful counting demonstrates that flexible action selection is effectively incorporated into Spaun. It also shows that the model has an understanding of order relations over numbers, and can exploit that knowledge to produce appropriate responses.

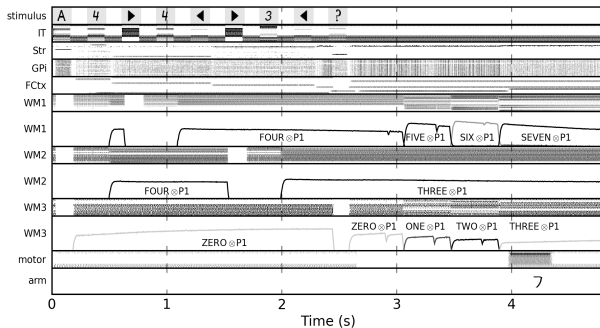


Figure 7: Adding  $4+3$  by mentally counting 4, 5, 6, 7 (WM1). WM2 keeps track of the fact that it should stop counting when it reaches the third step, and WM3 keeps track of what step it is at (0, 1, 2, 3).

## Pattern Completion

Finally, we show that Spaun is capable of quickly identifying and responding to patterns in its input via inductive learning. This specifically targets a type of task that has long been held to be problematic for connectionist approaches: the ability to rapidly create and bind symbolic variables (e.g. Marcus, 2001; Jackendoff, 2002). The following example (from Hadley, 2009) shows a pattern completion task that humans are readily able to solve given only a few items. They are told that if they hear “biffle biffle rose zarple”, the correct response is “rose zarple”.

After a three such examples, they must generalize to a new case:

### Training Set

- Input: *Biffle biffle rose zarple*. Output: *rose zarple*.
- Input: *Biffle biffle frog zarple*. Output: *frog zarple*.
- Input: *Biffle biffle dog zarple*. Output: *dog zarple*.

### Test Case

- Input: *Biffle biffle quoggie zarple*. Output: ?

Hadley suggests that this task requires rapid variable creation because the second last item in the list can take on any form, but human cognizers can nevertheless identify the overall syntactic structure and identify “quoggie zarple” as the appropriate response. So it seems that a variable has been created, which can receive any particular content, and that will not disrupt generalization performance.

Figure 8 shows Spaun’s behavior on a stimulus sequence with the same structure as that proposed by Hadley. Importantly, this is done extremely quickly ( $\sim 2$  seconds, consistent with human performance), and *without changing neural connection weights*. In other words, there is no learning rule; Spaun is able to learn to complete this pattern without any neural rewiring.

To achieve this result, we use a simplification of our earlier work with a neural model capable of performing Raven’s Matrices (Rasmussen & Eliasmith, 2011). We store the representation of the first list in one area of working memory (WM1) and a representation of the second list in another area (WM2). Since we are using the semantic pointer method of representing lists, these lists are encoded as two high-dimensional vectors ( $V1$  and  $V2$ ). We can then compute the transformation  $T$  which takes the first vector ( $V2=V1 \otimes T$ , so  $T=V2 \oslash V1$ ). We implement this in Spaun by adding a cortical area which computes the transformation between two working memory components, and adding rules to the basal ganglia to route this information appropriately when performing this task. As more examples are given, the value  $T$  is built up as the average over all the examples, improving accuracy.

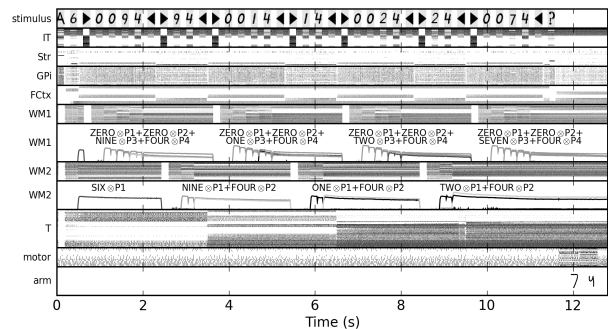


Figure 8: Pattern completion solving Hadley’ rapid variable creation problem. The input consists of pairs of lists. After seeing  $0094 \rightarrow 94$ ,  $0014 \rightarrow 14$ , and  $0024 \rightarrow 24$ , it correctly concludes that given  $0074$ , it can complete the pattern by outputting  $74$ .

## Discussion

The basic components of the model presented here are not new; we have previously published spiking neuron models capable of exhibiting list memory (Choo & Eliasmith, 2010), pattern completion (Rasmussen & Eliasmith, 2011), action selection (Stewart, Choo, & Eliasmith, 2010), and sequential reasoning (Stewart & Eliasmith, 2011). However, the work presented here is the first demonstration of these capabilities in a single, unified model. There are no adjustments made to the model between tasks, and indeed the model can seamlessly go from one task to the next.

Furthermore, we feel that an important feature of this cognitive model is that it includes the entire system: visual perception, cognition, and motor action. The neural representations used throughout the model are the same, as are the underlying computational principles, and methods of mapping to neural spikes. As a result, this single model has neuron responses in visual areas that match known visual responses, as well as neuron responses and circuitry in basal ganglia that match known responses and anatomical properties of basal ganglia, as well as behaviorally accurate working memory limitations, as well as the ability to perform human like induction, and so on. Spaun is thus both physically and conceptually unified.

It should be noted that the current set of tasks Spaun can perform are in a constrained semantic space – that of lists of numbers. However, the basic principle of using high-dimensional vectors that can be bound together (i.e. semantic pointers) generalizes to more complex domains.

Furthermore, the model's architecture is not tightly tied to the set of tasks being implemented. That is, rather than having particular components to perform each task, the components presented here provide generic cognitive capacities, and any given task can recruit these components as needed. For example, the Pattern Completion task requires use of a component that can find the transformation that relates the information in two different areas of working memory. This cortical component would also be useful for performing other tasks, such as a Raven's Matrix task (Rasmussen & Eliasmith, 2011).

For the model presented here, all synaptic connection weights between neurons are analytically derived, rather than having them be learned, as in traditional neural network models. While this demonstrates that our model is capable of learning without connection weight changes (as in the pattern completion task), it leaves open the question of how these connections are learned in the real brain. While we do not have a complete developmental story for the various cortical components, we have developed a dopamine-based reinforcement learning system (Stewart, Bekolay, & Eliasmith, 2012) that has been integrated with Spaun in an n-arm bandit task, but the results are not presented here due to space limitations. This system allows Spaun to learn the connections between the cortical components and the basal ganglia, allowing the model to learn to recruit different components for different tasks.

Spaun presents a detailed spiking neural model capable of visual recognition, cognitive control, working memory, symbolic manipulation, and producing hand-written motor outputs. This sort of model is required for connecting high-level cognitive theory and behavioural data to the biological constraints available from neuroscience.

## References

- Chaaban, I., & Scheessele, M. R. (2007). Human performance on the USPS database. *Technical Report*, Indiana University South Bend.
- Choo, F., Eliasmith, C. (2010). A Spiking Neuron Model of Serial-Order Recall. In Richard Cattrambone & Stellan Ohlsson (Eds.), *32<sup>nd</sup> Annual Conference of the Cognitive Science Society*. Portland, OR: Cognitive Science Society.
- Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review*, 8(4), 698–707.
- DeWolf, T. (2010). *NOCH: A framework for biologically plausible models of neural motor control*. Masters Thesis. University of Waterloo, Waterloo.
- Eliasmith, C. (2012). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press, New York, NY.
- Eliasmith, C. & Anderson, C. (2003). *Neural Engineering: Computation, representation, and dynamics in neurobiological systems*. Cambridge: MIT Press.
- Gayler, R. (2003). Vector Symbolic Architectures Answer Jackendoff's Challenges for Cognitive Neuroscience, in Slezak, P. (ed). *Int. Conference on Cognitive Science*, Sydney: University of New South Wales, 133–138.
- Hadley, R. F. (2009). The problem of rapid variable creation. *Neural computation*, 21(2), 510–32.
- Hinton, G.E. (2010). Learning to represent visual input. *Phil. Trans. Roy. Soc. B*, 365, 177–184.
- Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.
- Landauer, T. (1962). Rate of implicit speech. *Perceptual and Motor Skills*, 1, 646.
- Marcus, G. F. (2001). *The algebraic mind*. MIT Press, Cambridge, MA.
- Rasmussen, D., Eliasmith, C. (2011). A neural model of rule generation in inductive reasoning. *Topics in Cognitive Science*, 3(1), 140–153.
- Stewart, T.C., Bekolay, T., Eliasmith, C. (2012). Learning to select actions with spiking neurons in the basal ganglia. *Frontiers in Decision Neuroscience*. 6.
- Stewart, T.C., Choo, X., and Eliasmith, C. (2010). Dynamic Behaviour of a Spiking Model of Action Selection in the Basal Ganglia. *10<sup>th</sup> Int. Conf. on Cognitive Modeling*.
- Stewart, T.C., Eliasmith, C. (2011). Neural Cognitive Modelling: A Biologically Constrained Spiking Neuron Model of the Tower of Hanoi Task. In L. Carlson, C. Haelscher, & T. Shipley (Eds.), *33<sup>rd</sup> Annual Conference of the Cognitive Science Society*.

# Perception of Randomness: Subjective Probability of Alternation

Yanlong Sun (Yanlong.Sun@uth.tmc.edu)

Hongbin Wang (Hongbin.Wang@uth.tmc.edu)

School of Biomedical Informatics, University of Texas Health Science Center  
Houston, TX 77030 USA

## Abstract

We present a statistical account for the subjective probability of alternation in people's perception of randomness. By examining the *spatio-temporal distances* between pattern events, specifically, the *frequency* and *delay* of binary patterns in a Markov chain, we obtain some normative measures to calibrate people's expectation of randomness. We suggest that it can be fruitful to study subjective randomness in the context of human object representation and perception of time and space.

**Keywords:** subjective randomness; probability of alternation; waiting time; perception of time and space.

## Introduction

Much is known that subjective randomness—people's intuitive judgment on how an event or a series of events appears random—systematically deviates from the stochastic randomness described by normative probability theories. Among many statistics describing such discrepancy, the *probability of alternation* ( $p_A$ ) has been the most extensively studied in psychology literature (e.g., Budescu, 1987; Falk & Konold, 1997; Kahneman & Tversky, 1972; Kareev, 1992; Lopes & Oden, 1987; Nickerson, 2002; Sanderson, 2011). In a binary sequence generated by independent and stationary Bernoulli trials (for example, repeatedly tossing a fair coin),  $p_A$  can be defined as the probability that the outcome of any single event is different from the preceding one. If the process is truly random (e.g., the same fair coin is being tossed independently), the probability of alternation has the expected value  $p_A = .5$ . However, reviewed by Falk and Konold (1997, Table 1, p.304), in almost all of the studies with the tasks of recognizing or generating randomness, the modal *subjective probability of alternation* was approximately .60. That is, people tend to perceive sequences with  $p_A \approx .60$  as the most random and generate random sequences with  $p_A \approx .60$ .

One particular reason that the probability of alternation  $p_A$  receives a great deal of attention in the studies on subjective randomness is that it is highly correlated with many other sequential statistics, such as the runs test and serial correlation. Together, these statistics cover a variety of empirical phenomena, for example, the perception of streaks in basketball shooting (Burns, 2004; Gilovich, Vallone, & Tversky, 1985; Sun & Wang, 2010b), the recency effect (Ayton & Fischer, 2004), the working memory capacity and detection of covariances in short sequences (Kareev, 1992), and the encoding of subjective complexity (Falk & Konold, 1997; Falk, Falk, & Ayton, 2009). (For a recent review, see Oskarsson, Van Boven, McClelland, & Hastie, 2009).

To explain the biased probability of alternation in subjective randomness, Falk and Konold (1997) developed a “difficulty predictor” (DP) as a measure of “subjective complexity”.

Based on the concept that random sequences are irreducibly complex (i.e., algorithmic complexity, Kolmogorov, 1965), Falk and Konold propose that people's sense of randomness is not based on the deviations from the equiprobability of patterns of the same length (i.e., “ $n$ -grams”), rather, it may be based on the difficulty level when people attempt to memorize or copy a sequence by its minimal description. Given any binary sequence, the difficulty predictor is defined by adding twice the number of alternating runs to the number of pure runs. For example, the following sequence is partitioned into 5 segments, where pure runs (streaks) are underlined and alternating runs are double underlined:

H H T H T H T H T H T T T H H T H T H T H

Thus, the DP score for this particular sequence is  $1 + 2 + 1 + 1 + 2 = 7$ . By this measure, a perfect streak would be perceived as the most nonrandom because of its lowest DP score (i.e., the easiest to remember). In contrast, sequences with more alternating runs—hence greater  $p_A$ —are more difficult to encode therefore tend to be perceived as more random.

Overall, it has been demonstrated that DP correlates remarkably well with participants' ratings of randomness, memorization time, assessed difficulty of memorization, and copying difficulty. And, the mean ratings of randomness show the classic preference for over-alternating sequences (Falk et al., 2009; Falk & Konold, 1997). However, Griffiths and Tenenbaum (2003) point out that DP remains a subjective measure since its objective counterpart, algorithmic complexity, is not computable. Instead, Griffiths and colleagues propose to use Bayesian inferences to account for subjective randomness (Griffiths & Tenenbaum, 2001, 2003; Hsu, Griffiths, & Schreiber, 2010). By this account, the subjective randomness of a particular sequence  $X$  is defined as

$$\text{random}(X) = \log \frac{p(X|\text{random})}{p(X|\text{regular})} \quad (1)$$

Then, the problem of judging randomness can be reduced to comparing two probabilities—whether the sequence is produced by a random process (e.g., independent and stationary Bernoulli trials), or, by a process with some regularities. To specify  $p(X|\text{regular})$ , Griffiths and Tenenbaum (2003) develop a hidden Markov model that makes transitions between hidden states depending on whether a motif is to be repeated or altered, where a motif is a short pattern such as H, T, HT, or, TH. By maximizing  $p(X|\text{regular})$ , they obtain a set of parameters that provide a better fit to the mean randomness ratings reported by Falk and Konold (1997).

The difficulty predictor and the Bayesian account have the



advantage to test against specific encoding strategies. However, DP has some counterintuitive properties. For example, in Figure 1, sequence (a) may appear to be more random than sequence (b), but the former actually has a lower DP score ( $DP = 5$ ) than the latter ( $DP = 6$ ). The Bayesian approach can fix this problem by adding more motifs of various length, but at the cost of computational complexity—to include all motifs of length 4, the hidden Markov model will have 22 motifs and 72 states (Griffiths & Tenenbaum, 2003).

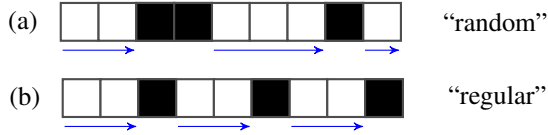


Figure 1: “Fast detection” of regularities. Without exactly counting alternating runs or calculating probabilities, it would be easily discernible that sequence (a) appears more “random” than sequence (b). The regularity in sequence (b) can be detected by the equal distances between patterns (for example, between the filled squares or the interruptions of unfilled squares). In addition, this example shows some of the counterintuitive properties of the difficulty predictor in that sequence (a) has a lower DP score than sequence (b).

Perhaps more interestingly, Figure 1 also prompts a speculation: whether the judgment of randomness can be reached at *before* any effort of encoding or estimating the probabilities of the observed sequences. For example, the regularity in Figure 1(b) might be quickly spotted by the equal distances between patterns. Such a speculation has actually led us to consider the recent advances in the investigations on *perception of time and space*. For instance, it has been posited that complex achievements such as mathematics and geometry, which are uniquely human in their full linguistic and symbolic realization, rest nevertheless on a set of core knowledge systems driven by the representations of object, space, time, and number, and these representations may have an early developmental origin shared by human infants as well as animals (e.g., Dehaene & Brannon, 2010; Spelke & Kinzler, 2007; Spelke, Lee, & Izard, 2010). Applied to the research on subjective randomness, it would be plausible to hypothesize that when people attempt to judge randomness (or detect regularities), the processing of spatio-temporal distances between observations and patterns is the primitive driving force, *before* any encoding effort of memorizing, copying the observed sequences or comparing the probabilities of specific processes.

### Spatio-Temporal Distances between Patterns

In the present paper, we propose to utilize the spatio-temporal principles in object representation and human perception of time and space (e.g., Spelke & Kinzler, 2007) to study subjective randomness. Our approach is to first examine the spatial and temporal distributions of pattern events produced by random processes then match them to the psychological

spatio-temporal distances in people’s perception of randomness. To study the spatio-temporal distances between events, we focus on two sets of statistics, namely, given a random or a regular process, *how often* or *how likely* an event or a series of events would occur; And, from the start of an observation, *when* or *where* the events of interest would be encountered.

Apparently, *how often*, *when*, and *where* are different statistical properties and may bear different psychological relevancies. To set ideas, consider a simple case of coin tossing. If we tossed a coin three times and got three heads in a row as HHH (H = heads and T = tails), many of us might start getting suspicious about the fairness of the coin. But we would think it not at all noteworthy if the three tosses resulted in the pattern THH. The apparent randomness (or nonrandomness) cannot be explained by the frequency of encounters since the probability of obtaining either pattern in their exact orders is precisely the same,  $(\frac{1}{2})^3 = \frac{1}{8}$  (i.e., the equiprobability of the “*n*-grams”, Falk & Konold, 1997). However, less is known that there is a set of statistical properties that may very well explain why people consider a streak pattern rare and remarkable. When a fair coin is tossed repeatedly, it takes on average 8 tosses to observe the first occurrence of THH, but it takes on average 14 tosses to observe the first occurrence of HHH. Moreover, when both patterns are monitored simultaneously in one global sequence, the odds are 7 to 1 that one is more likely to first encounter THH than to first encounter HHH. That is, despite equal probabilities of occurrences, the time it takes to first encounter HHH is significantly “delayed” than that of THH.

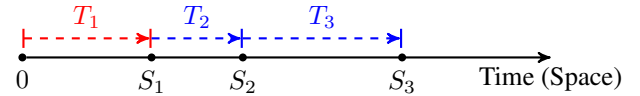


Figure 2: Spatio-temporal intervals between encounters of random events. An observation starts *from scratch* at Time = 0.  $T_1$  is the first arrival time and its expected value  $E[T^*]$  is called *waiting time*.  $T_2, T_3, \dots$ , are the interarrival times between successive occurrences of the events, and their expected value  $E[T]$  is called *mean time*.

For formal definitions, we record the time (or location) when an event occurs at  $S_1, S_2, S_3, \dots$ , counting from the very beginning of the process (Figure 2). Then,  $T_1 = S_1$  is the first arrival time with an expected value  $E[T_1] = E[T^*]$  called *waiting time*.  $T_2 = S_2 - S_1, T_3 = S_3 - S_2, \dots$ , are the interarrival times between successive occurrences of the events, and their expected value  $E[T]$  is called *mean time*.<sup>1</sup> It can be shown that the mean time of a pattern is in effect a measure of *frequency* as the inverse of probability of occurrence, and the waiting time is a measure of *delay* in that a pattern’s expected first arrival time may be longer but not shorter than

<sup>1</sup>Note that since the first arrival time may have a different distribution than later interarrival times, we use  $T^*$  to denote the first arrival time and  $T$  to denote interarrival times.



its mean time (Ross, 2007).

Consider an irreducible Markov chain  $\{X_n, n \geq 0\}$  with transition probabilities  $P_{i,j}$  and stationary probabilities  $\pi_i$ ,  $i \geq 0$ . Then, for pattern  $(i_1, i_2, \dots, i_k)$ , its mean time (i.e., inverse of the pattern's frequency) is the mean number of transitions between successive visits to the pattern,

$$E[T_k] = \frac{1}{\pi_{i_1} P_{i_1, i_2} \dots P_{i_{k-1}, i_k}} \quad (2)$$

For the pattern's waiting time, we first consider whether a successive arrival of the pattern can "reuse" any of the elements from its previous arrival. For example, in sequence THHH, pattern HH has occurred twice and its second arrival has reused the last element from the first arrival. An overlap index  $s$  is defined as the maximum number of elements at the end of the pattern that can be used as the beginning part of the next arrival,

$$s = \max \{j < k : (i_{k-j+1}, \dots, i_k) = (i_1, \dots, i_j)\} \quad (3)$$

For example,  $s_{HT} = 0$  and  $s_{HH} = 1$ . Let  $\mu(a, b)$  denote the mean number of transitions for the Markov chain to enter state  $b$  from state  $a$ . If the pattern has no overlap,  $s = 0$ , its waiting time is,

$$E[T_k^*] = \mu(a, i_1) - \mu(i_k, i_1) + E[T_k] \quad (4)$$

If the pattern has an overlap  $s > 0$ , we first consider  $E[T_s^*]$ , the waiting time for a shorter sub-pattern  $(i_1, i_2, \dots, i_s)$ , which is consisted of the first or the last  $s$  elements in pattern  $(i_1, i_2, \dots, i_k)$ . Then,

$$E[T_k^*] = E[T_s^*] + E[T_k] \quad (5)$$

By recursively applying Equation (5) until we reach the shortest sub-pattern with no overlap, we can obtain the waiting time for the original pattern. Comparing Equations (4) and (5), we can see that when looking for the first arrival of a pattern, if the pattern has an overlap  $s > 0$ , anything that goes wrong after the first  $s$  elements will make the counting process start from scratch. In other words, a pattern's *waiting time* can be *delayed* by the pattern's overlapping property. In contrast, Equation (2) shows that a pattern's *mean time* or *frequency* is not affected by the overlapping property.

### Frequency and Delay by $p_A$

To generate binary patterns, we can use a Markov chain with two states H and T, where  $P_{H,T} = P_{T,H} = p_A$ , and  $P_{H,H} = P_{T,T} = 1 - p_A$ . This Markov chain is equivalent to the models used by Lopes and Oden (1987), where  $p_A < .5$  represents the tendency of repetition, and  $p_A > .5$  represents the tendency of alternation.

Assuming that the initial state is equally likely to be in either H or T, from equation (2), we have

$$E[T_{HT}] = E[T_{TH}] = \frac{2}{p_A} \quad (6)$$

$$E[T_{HH}] = E[T_{TT}] = \frac{2}{1 - p_A} \quad (7)$$

From Equations (4) and (5), we have

$$E[T_{HT}^*] = E[T_{TH}^*] = 1 + \frac{3}{2p_A} \quad (8)$$

$$E[T_{HH}^*] = E[T_{TT}^*] = \frac{-2p_A^2 + 5p_A + 1}{2p_A(1 - p_A)} \quad (9)$$

Figure 3 plots the mean time and waiting time for patterns of length 2 as the functions of probability of alternation  $p_A$ . When  $p_A = .5$ , we have a case of independent and stationary Bernoulli trials. We first note that the mean time is the same for all patterns of the same length. For example,  $E[T_{HT}] = E[T_{HH}] = 4$ . However, the waiting time can be different depending on the pattern's overlapping property. For example,  $s_{HT} = 0$ ,  $E[T_{HT}^*] = 4$ , and,  $s_{HH} = 1$ ,  $E[T_{HH}^*] = 6$ . That is, the waiting time is longer for the shortest streak patterns HH or TT than for the shortest alternating pattern HT or TH. Solving the equality between Equations (8) and (9), we obtain  $p_A = \frac{1}{3}$ . Thus, as long as  $p_A > \frac{1}{3}$ , we have  $E[T_{HH}^*] > E[T_{TH}^*]$ .

Moreover, let  $\text{Var}(T)$  denote the variance of the interarrival times, it can be shown that patterns may differ substantially in how evenly they are distributed over time (or space), for

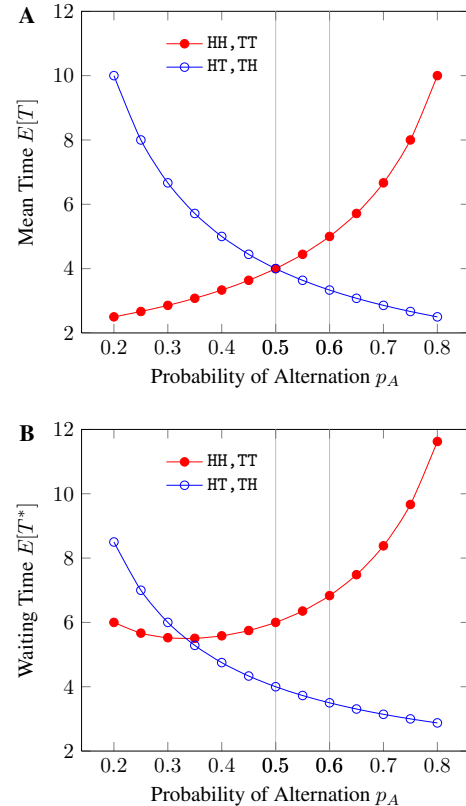


Figure 3: Mean time (A) and waiting time (B) as the functions of probability of alternation  $p_A$ . When  $p_A = .5$ , the process is equivalent to independent and stationary Bernoulli trials.

example,  $\text{Var}(T_{\text{TH}}) = 4$ , and  $\text{Var}(T_{\text{HH}}) = 20$ . In addition, the waiting time is highly correlated with the variance of interarrival times since both values are extended by a pattern's overlap tendency (see Table 1). (For the calculation of variances, see Sun & Wang, 2010a.)

Table 1: Mean and variance of the first arrival time ( $T^*$ ) and interarrival times ( $T$ ) for binary patterns in independent Bernoulli trials ( $p_A = 0.5$ ) when tossing a fair coin. Reciprocal patterns are listed only once, for example, HH is equivalent to TT, and, HT is equivalent to TH.

Patterns	$s$	$E[T]$	$\text{Var}(T)$	$E[T^*]$	$\text{Var}(T^*)$
H	0	2	2	2	2
HT	0	4	4	4	4
HH	1	4	20	6	22
HHT	0	8	24	8	24
HTT	0	8	24	8	24
HTH	1	8	56	10	58
HHH	2	8	120	14	142

## Psychological Implications

The Markov model described above may provide quantitative measures to calibrate subjective randomness, particularly regarding the seemingly miscalibrated subjective probability of alternation  $p_A$ . Different from previous studies that focus on the frequency of patterns, here we examine the spatio-temporal distances between pattern events that cover both frequency and delay (including variances), either from the very beginning of the process (waiting time), or between successive occurrences of the pattern given that the pattern has occurred before (mean time).

It should be noted that so far our analyses are limited to short patterns. This consideration is based on the empirical findings that people are sensitive to even the shortest patterns of length 2. For example, it has been reported that participants would report “a streak is occurring” beginning at the third repeating event (Carlson & Shu, 2007). In an fMRI study, Huettel, Mack, and McCarthy (2002) show that a distributed set of regions in prefrontal cortex are exquisitely sensitive to the presence and the termination of streak patterns even when pattern length was only 2, despite the fact that participants were informed of the random order of the sequences.

It appears that the waiting time statistics fit well to one of the most influential accounts for subjective randomness, the “representativeness heuristic” (e.g., Gilovich et al., 1985; Kahneman & Tversky, 1972). By this account, people expect smaller sequences to resemble the balanced distributions in the long run, such that a streak of heads would seem to be rare and remarkable as it would not be representative of the process of random coin tossing. This would have explained the biased subjective probability of alternation in that people tend to expect fewer and shorter streaks than would be mathematically probable when observing sequences produced by a random

process, and, they avoid repetitions of the same elements when instructed to generate such sequences (e.g., Falk & Konold, 1997; Wagenaar, 1972).

Despite its plausibility, the representativeness account has been criticized for the lack of definition (Ayton & Fischer, 2004; Falk & Konold, 1997; Gigerenzer, 1996). Nevertheless, it has been proposed that the waiting time statistics may provide quantitative explanations to this account (Hahn & Warren, 2009; Sun, Tweney, & Wang, 2010a; Sun & Wang, 2010a, 2010b). Specifically, people judge the frequency of an event on the basis of how it is representative of the underlying population or process (representativeness), and how easily an example can be brought to mind (availability). When people think of a truly random process (by actually tossing a coin or conducting a mental experiment), a streak pattern—even at its shortest length of 2 (e.g., HH in tossing a coin)—may be perceived as the most nonrepresentative and the most unavailable. Compared with other patterns of the same length, a streak is the most delayed in its first arrival and has the largest variance of interarrival times thus the most uneven or clustered distribution over time (see Table 1). Note that these particular properties are not limited to independent Bernoulli trials where  $p_A = .5$ . For example, Figure 3B shows that as long as  $p_A > \frac{1}{3}$ , the waiting time for the streak pattern HH will be longer than that of non-streak pattern HT.

Moreover, there has been direct evidence suggesting that people may at least have an approximate sense of the waiting time, namely, the delayed occurrence of streak patterns. Oppenheimer and Monin (2009) report that when participants were asked to estimate the number of coin flips before the occurrence of a pattern of length 5, they believed that a sequence of coin flips was nearly twice as long before a streak (mean estimate = 16.2) than when there was no streak (mean estimate = 8.7). Applying Equations (4) and (5), we can show that for patterns of length 5, the waiting time for a streak is 62 tosses, and the average waiting time for non-streak patterns is approximately 34.3 tosses: the former is nearly twice as long as the latter.

In the light of the delayed first arrival for streak patterns, we speculate that people's expectation of the probability of alternation as  $p_A \approx 0.6$  might in effect have been driven by their experiences of pattern events as random patterns unfold in time and space. For example, we can reconstruct the task of generating randomness with the Markov chain described above. Since at any given moment, participants face the choice of either repeating or reversing the current outcome (an H or a T), the generation process is equivalent to the process of choosing patterns of length 2—either a streak pattern (HH or TT) or a non-streak pattern (HT or TH). Then,  $p_A = .6$  means that 60% of the time participants choose a non-streak pattern, indicating a false belief that in tossing a fair coin independently (i.e.,  $p_A = .5$ ), streak patterns should occur less frequently than non-streak patterns. Such belief can be formulated as a ratio of pattern mean times. From Equations (6) and (7), when

$$p_A = .6,$$

$$E [T_{HH,TT}] : E [T_{HT,TH}] = p_A : (1 - p_A) = 3 : 2$$

Then, comparing the waiting time in the process of independent coin tossing yields exactly the same ratio, where  $p_A = .5$ ,

$$E [T_{HH,TT}^*] : E [T_{HT,TH}^*] = 6 : 4 = 3 : 2$$

That is, measured by the mean time, participants have failed the task of producing randomness (i.e., the independence property where  $p_A = .5$ ) as if they have falsely believed that patterns HH and TT would occur less frequently than patterns HT and TH. Quantitatively, this comparison indicates that the observed bias might have stemmed from participants' expectation of waiting time from a truly random process, since the mean time does not distinguish any patterns when  $p_A = .5$  (e.g., see Figure 3A).

## Discussion

The development of waiting time statistics appears to be promising and may have potential significance in explaining a range of human cognitive functions (e.g., Oppenheimer & Monin, 2009; Sun et al., 2010a; Sun, Tweney, & Wang, 2010b; Sun & Wang, 2010b, 2010a, 2011). Specifically, we argue that it can be fruitful to study subjective randomness in the context of *human perception of time and space*. And, the *frequency* and *delay* of pattern events, rather than individual events (e.g., a single coin toss), may be the key statistics and theoretical constructs that underlie human perception of randomness.

It has been posited that by exposing to the various environmental statistics, human mind may have evolved an accurate sense of randomness but may fail to reveal it by the standard of a particular measuring device (e.g., Pinker, 1997). Given that the waiting time and the variance of interarrival times can be substantially different for patterns with the same mean time (e.g., Table 1 and Figure 3), one may logically assume that these statistics may play a critical role in shaping people's perception and judgment of randomness. Unfortunately, in the long lasting investigations on subjective randomness, the mean time of patterns serves as the sole normative measure of randomness. Despite various forms of experimental tasks (e.g., randomness generation or recognition, probabilistic predictions) and statistical methods (e.g., runs test, serial correlation, Bayesian inferences), existing studies have been focusing on the discrepancies between subjective responses and the probabilities of the occurrences of random patterns. Nevertheless, the absence of waiting time statistics in the investigations on subjective randomness may be due to its late and still ongoing development in statistical research (e.g., Pozdnyakov, 2008; Ross, 2007), and, remain fairly novel to the audience in psychology (c.f., Hahn & Warren, 2009; Konold, 1995; Nickerson, 2007; Sun et al., 2010a, 2010b).

More significantly, recent advances in the behavioral and neurological sciences on human cognitive achievements all point to the role of the perception of time and space. It has

been proposed that complex achievements such as mathematics and geometry, which are uniquely human in their full linguistic and symbolic realization, rest nevertheless on a set of core knowledge systems that are driven by the representations of object, space, time and number (Dehaene & Brannon, 2010; Spelke & Kinzler, 2007; Spelke et al., 2010). And, these representations may have a common perceptual metric in the form of a *mental number line* (Burr & Morrone, 2011; Dehaene, Piazza, Pinel, & Cohen, 2003) and have an early developmental origin shared by human infants as well as other animals (de Hevia & Spelke, 2010; Haun, Jordan, Vallortigara, & Clayton, 2010; Hubbard, Piazza, Pinel, & Dehaene, 2005). Applied to the research on subjective randomness, it would be plausible to hypothesize that when people attempt to judge randomness (or detect regularities), the processing of spatio-temporal distances between observations and patterns is the primitive driving force, *before* any encoding effort for memorizing, copying, or assessing the probability of pattern occurrences (e.g., see Figure 1).

Nonetheless, we need to collect more empirical evidence to investigate whether and how human cognition is sensitive to the statistics of random patterns, for example, via experiments that manipulate the overlapping and delay properties of pattern events then measure the psychological responses. Moreover, a theoretical breakthrough would also require us to firmly demonstrate the psychological relevance of the pattern time statistics and the spatio-temporal principles in object representation and human perception of time and space, in order to develop a mental calculus of how these constructs work together.

## Acknowledgments

Supported by the Office of Naval Research (ONR) grant number N00014-08-1-0042, and Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract number D10PC20021. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

## References

- Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition*, 32(8), 1369–1378.
- Budescu, D. V. (1987). A Markov model for generation of random binary sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 13(1), 25–39.
- Burns, B. D. (2004). Heuristics as beliefs and as behaviors: The adaptiveness of the "hot hand". *Cognitive Psychology*, 48(3), 295–331.
- Burr, D. C., & Morrone, M. C. (2011). Spatiotopic coding and remapping in humans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1564), 504–515.

- Carlson, K. A., & Shu, S. B. (2007). The rule of three: How the third event signals the emergence of a streak. *Organizational Behavior and Human Decision Processes*, 104(1), 113–121.
- de Hevia, M. D., & Spelke, E. S. (2010). Number-space mapping in human infants. *Psychological Science*, 21(5), 653–660.
- Dehaene, S., & Brannon, E. M. (2010). Space, time, and number: A Kantian research program. *Trends in Cognitive Sciences*, 14(12), 517–519.
- Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, 20(3–6), 487–506.
- Falk, R., Falk, R., & Ayton, P. (2009). Subjective patterns of randomness and choice: Some consequences of collective responses. *Journal of Experimental Psychology: Human Perception and Performance*, 35(1), 203–224.
- Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, 104(2), 301–318.
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, 103(3), 592–596.
- Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17(3), 295–314.
- Griffiths, T. L., & Tenenbaum, J. B. (2001). Randomness and coincidences: Reconciling intuition and probability theory. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 398–403). Mahwah, NJ: Lawrence Erlbaum Associates.
- Griffiths, T. L., & Tenenbaum, J. B. (2003). Probability, algorithmic complexity, and subjective randomness. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th annual conference of the cognitive science society* (pp. 480–485). Cognitive Science Society.
- Hahn, U., & Warren, P. A. (2009). Perceptions of randomness: Why three heads are better than four. *Psychological Review*, 116(2), 454–461.
- Haun, D. B. M., Jordan, F. M., Vallortigara, G., & Clayton, N. S. (2010). Origins of spatial, temporal and numerical cognition: Insights from comparative psychology. *Trends in Cognitive Sciences*, 14(12), 552–560.
- Hsu, A., Griffiths, T. L., & Schreiber, E. (2010). Subjective randomness and natural scene statistics. *Psychonomic Bulletin & Review*, 17(5), 624–629.
- Hubbard, E. M., Piazza, M., Pinel, P., & Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nature Reviews Neuroscience*, 6(6), 435–448.
- Huetzel, S. A., Mack, P. B., & McCarthy, G. (2002). Perceiving patterns in random series: Dynamic processing of sequence in prefrontal cortex. *Nature Neuroscience*, 5(5), 485–490.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3), 430–454.
- Kareev, Y. (1992). Not that bad after all: Generation of random sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 18(4), 1189–1194.
- Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1, 1–7.
- Konold, C. (1995). Confessions of a coin flipper and would-be instructor. *The American Statistician*, 49(2), 203–209.
- Lopes, L. L., & Oden, G. C. (1987). Distinguishing between random and nonrandom events. *Journal of Experimental Psychology: Learning Memory and Cognition*, 13(3), 392–400.
- Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review*, 109(2), 330–357.
- Nickerson, R. S. (2007). Penney Ante: Counterintuitive probabilities in coin tossing. *The UMAP Journal*, 28(4), 503–532.
- Oppenheimer, D. M., & Monin, B. (2009). The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making*, 4(5), 326–334.
- Oskarsson, A. T., Van Boven, L., McClelland, G. H., & Hastie, R. (2009). What's next? Judging sequences of binary events. *Psychological Bulletin*, 135(2), 262–285.
- Pinker, S. (1997). *How the mind works*. New York: Norton.
- Pozdnyakov, V. (2008). On occurrence of patterns in Markov chains: Method of gambling teams. *Statistics & Probability Letters*, 78(16), 2762–2767.
- Ross, S. M. (2007). *Introduction of probability models* (9th ed.). San Diego, CA: Academic Press.
- Sanderson, Y. B. (2011). Color charts, esthetics, and subjective randomness. *Cognitive Science*.
- Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, 10(1), 89–96.
- Spelke, E. S., Lee, S. A., & Izard, V. (2010). Beyond core knowledge: Natural geometry. *Cognitive Science*, 34(5), 863–884.
- Sun, Y., Tweney, R. D., & Wang, H. (2010a). Occurrence and nonoccurrence of random sequences: Comment on Hahn and Warren (2009). *Psychological Review*, 117(2), 697–703.
- Sun, Y., Tweney, R. D., & Wang, H. (2010b). Postscript: Untangling the gambler's fallacy. *Psychological Review*, 117(2), 704–705.
- Sun, Y., & Wang, H. (2010a). Gambler's fallacy, hot hand belief, and time of patterns. *Judgment and Decision Making*, 5(2), 124–132.
- Sun, Y., & Wang, H. (2010b). Perception of randomness: On the time of streaks. *Cognitive Psychology*, 61(4), 333–342.
- Sun, Y., & Wang, H. (2011). Probability theory and perception of randomness: Bridging “ought” and “is”. *Behavioral and Brain Sciences*, 34(5), 271–272.
- Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, 77(1), 65–72.

# Time Course of Inhibitory Control During Analogical Reasoning: An Event-Related Potential Approach

**Brian M. Sweis (bsweis@luc.edu)**

Departments of Biology and Psychology  
Loyola University Chicago  
1032 W Sheridan Rd  
Chicago, IL 60626 USA

**Krishna L. Bharani (kbharani@luc.edu)**

Department of Psychology  
Loyola University Chicago  
1032 W Sheridan Rd  
Chicago, IL 60626 USA

**Robert G. Morrison (rmorrison@luc.edu)**

Department of Psychology, Neuroscience Institute  
Loyola University Chicago  
1032 W Sheridan Rd  
Chicago, IL 60626 USA

## Abstract

Inhibitory control is an important aspect of analogical reasoning critically dependent on prefrontal cortex. We used a novel visual analogy paradigm with scalp electroencephalography (EEG) to explore several ways the brain uses inhibitory control to perform analogy. Previous studies have suggested that inhibitory control helps to manage working memory, so we used a separate task to measure individual differences in working-memory span to help us interpret differences in inhibitory control during reasoning. We found evidence that low working-memory span individuals likely lacked the necessary inhibitory control to keep unattended relations from entering visuospatial working memory early in processing. We also found that a late frontal event-related potential sensitive to relational distraction was differentially modulated in high and low working memory span individuals. These findings provide additional evidence for the importance of inhibitory control during analogical processing.

**Keywords:** analogy, working memory, inhibitory control, EEG, ERP

## Introduction

Studies involving children (Richland, Chan, Morrison, & Au, 2010; Richland, Morrison, & Holyoak, 2006; Thibaut, French, & Vezneva, 2010a, 2010b), younger adults (Cho, Holyoak, & Cannon, 2007; Cho et al., 2010), older adults (Viskontas et al., 2004), and patients with damage to prefrontal cortex (Krawczyk et al., 2008; Morrison et al., 2004) have all provided evidence that inhibitory control in working memory (WM) is an important aspect of both visual and verbal analogical reasoning. Neuroimaging

studies of analogical reasoning have implicated areas in prefrontal cortex (PFC) as critical for semantic retrieval during analogy (Bunge, Wendelken, Badre, & Wagner, 2005), avoiding distraction from non-goal related relational information (Cho et al., 2010), and performing analogical mapping and similar types of relational integration (Bunge, Helskog, & Wendelken, 2009; Cho et al., 2010; Green et al., 2010; Morrison, Nikitin, Bharani, & Doumas 2012). Computational accounts of these data (Doumas, Morrison, & Richland, under review; Knowlton, Morrison, Hummel, & Holyoak, 2012; Morrison et al., 2004; Morrison, Doumas, & Richland, 2011; Viskontas et al., 2004) suggest that inhibitory control is central to the processes of semantic retrieval and analogical mapping; however, there is little direct experimental evidence for mechanisms by which inhibitory control is recruited during analogical reasoning.

WM has been thought to play a critical role during analogical reasoning (Halford, 1992; Morrison, 2005). Domain-specific as well as central-executive WM dual tasks interfere with analogical processing (Morrison, Truong, & Holyoak, 2001; Waltz, Lau, Grewal, & Holyoak, 2000). Likewise, individual differences in working-memory span (see Conway et al., 2005) are frequently related to matrix reasoning performance (e.g., Kane & Engle, 2002). In an effort to understand how inhibitory control may be involved in WM processing, Vogel, McCollough, and Machizawa (2005) asked participants with high and low WM span to perform a simple delayed match-to-sample WM task while their brain activity was observed using scalp electroencephalography (EEG). Vogel et al. identified a Continuous Negative Variation (CNV) event-related

potential (ERP) during the delay period in the task that correlated with the number of items the participants were required to hold in WM. Interestingly, when participants were asked to remember two items and ignore two others, the ERPs of high-WM span participants resembled those for two item trials, while those of low-WM span participants resembled those for four items. Thus, high-WM span individuals appear to be better at managing their WM using inhibitory control to suppress goal-irrelevant information. Shimamura (2000) has argued that his type of dynamic filtering appears to be a fundamental function of PFC.

Based on Learning and Inference with Schema and Analogy (LISA; Hummel & Holyoak, 1997, 2003), a neurally-plausible model of analogical reasoning, we have previously argued that inhibitory control is necessary throughout analogical processing (Morrison et al., 2004; Morrison, Dumas, & Richland, 2011; Viskontas et al., 2004). Specifically, inhibition plays a central role in (a) LISA's manipulation of relations in WM, (b) its ability to select items for placement into WM, (c) its ability to discover analogical mappings. Thus, we anticipate that inhibitory control will be evident in analogical reasoning, and may be modulated by the WM span of participants.

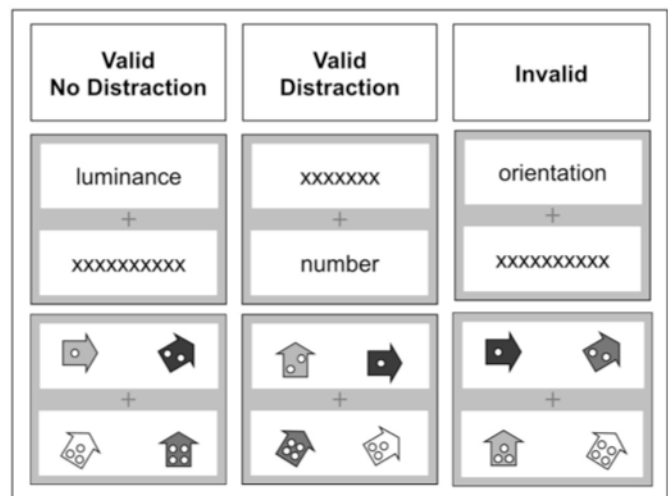
In an effort to explore two possible roles for inhibitory control during analogical processing, we developed an experimental paradigm for use with EEG (see Figure 1). On each trial, participants were cued to solve a visual analogy based on one of three abstract relations present in the stimuli. Critically, on some valid trials one of the unattended relations was not congruent. Participants need to ignore this relation to arrive at the correct solution. Thus, the task is similar to the version of the People Pieces analogy task we had developed for behavioral (Cho, Holyoak, & Cannon, 2007; Viskontas et al., 2004) and neuroimaging studies (Cho et al., 2010) except that participants were considering asymmetric relations as opposed to same-different relations. The task was also similar to Vogel et al. (2005), in that participants had a goal-relevant WM load (i.e., the to-be-attended-to relation) in the presence of potentially distracting information.

We had two central hypotheses. First, people lower in WM-span would be less efficient than higher WM-span individuals in keeping goal-irrelevant information out of visuospatial WM. Thus, we expected to see a more negative CNV (Vogel et al., 2005) in occipitoparietal regions in low WM-span individuals, indicating that they were storing more information in their WM than higher WM-span individuals who were efficiently filtering the goal-irrelevant information.

Second, we predicted that relationally distracting trials require the engagement of inhibitory control and thus should engage areas in inferior frontal cortex (e.g., Cho et al., 2010) to protect the analogical mapping process from goal-irrelevant information.

## Method

Participants verified visual analogies constructed from shapes that possessed three varying properties (luminance, orientation, and number; see Figure 1). On each trial, participants were cued to attend to *only one* of the relations formed by the three properties. Participants decided whether the relation in the top pair was the same or different than the relation in the bottom pair. There were three conditions in this experiment: *Valid–No Distraction*, *Valid–Distraction*, and *Invalid*. In the Valid–No Distraction condition, all three relations were congruent between the top and bottom pair, with the correct answer being “yes” to indicate that the problem as cued was a valid analogy. In the Valid–Distraction condition, the cued property had congruent relations as in the previous condition; however, one of the two unattended relations was incongruent, thus creating a response conflict between the attended and unattended relation. Invalid trials were just like Valid–Distraction stimuli, except that participants were asked to attend to the relation that did not map.



*Figure 1:* Participants saw analogy problems in one of three conditions. In “Valid–No Distraction” problems, participants mapped based on a single relation (e.g., luminance—the shapes get darker from left to right), while the other relations (e.g., orientation and number) could also be successfully mapped between pairs. In “Valid–Distraction” problems, participants once again were only required to map based on one relation (e.g., number—the number of dots in the shape decreased from left to right); however, one of the other relations present did not successfully map between pairs (e.g., luminosity in the source decreased from left to right, while it increased from left to right in the target). “Invalid” trials were like “Valid–Distraction” trials; however, participants were to map based on the invalid relation (e.g., orientation—the arrowhead of the shape rotated counterclockwise in the source, but clockwise in the target).

## Participants

Twenty-nine undergraduate students from Loyola University Chicago participated in the experiment. Of the 29 participants three were omitted from the analysis because of poor EEG recording quality. The remaining 26 participants were divided based on median WM-span into two equally sized groups. The low-WM span group ( $M = 33$ ,  $SEM = 2$ ) had a WM span smaller than the high WM-span group ( $M = 61$ ,  $SEM = 2$ ;  $t(24) = 7.5$ ,  $p < .001$ ).

Participants gave informed consent to take part in the study. The Loyola University Chicago Institutional Review Board approved all recruitment methods and procedures.

## Materials

Each analogy problem consisted of two pairs of geometric shapes (see Figure 1). Each shape had one of four levels of three parametrically manipulated properties: luminance, orientation, and number. Shapes were combined into pairs to create order relations with respect to the three properties. For instance, pairs of shapes could be increasingly bright or dark (luminance); rotate clockwise or counter clockwise (orientation); and increase or decrease in *number*. In any given problem a relation in the source (i.e., top pair) could either match or mismatch the corresponding relation in the target (i.e., bottom pair). A set of 144 unique stimuli was generated, 72 of which contained pairs of shapes with all congruent relations (used for Valid-No Distraction trials). The remaining 72 stimuli were divided into thirds, with each third having one mismatching relation in one of the three properties. For the problems containing a mismatching relation, if the participant was cued to attend to the mismatching relation the trial was Invalid, but if the participant was cued to attend to one of the matching relations the trial type was Valid-Distraction.

## EEG Recording

Scalp electroencephalography signal (EEG) was recorded from each participant using a 38-channel Biosemi Active2 EEG system. 32 electrodes were located at standard 10/20 locations in a nylon-elastic cap. Two electrodes were placed on the left and right mastoid bones for subsequent digital re-referencing. To expand the coverage of EEG monitoring, we placed four electrodes on the face on the inferior and lateral aspects of the eye orbit. These electrodes were used to expand PFC electrode coverage and for ocular artifact correction and rejection. Unfiltered EEG was re-referenced to an average of the two mastoid electrodes and a 0.01 Hz high-pass filter was applied after recording. A band-stop filter from 59 to 61 Hz was also applied to the raw EEG to remove any AC electrical contamination. EEG signal was corrected for ocular artifacts using a spatial PCA filter corrected for the average noise level in the signal according a method available in EMSE (Source Signal Imaging, San Diego CA). Signal was further cleaned via a  $\pm 100\mu V$

rejection criterion. Included participants have fewer than 15% of trials rejected due to EEG artifacts.

## Procedure

After a participant was fitted with the EEG cap and electrodes, he or she sat in a soundproof chamber equipped with a 21-inch CRT monitor and an electronic response box controlled by a program written in e-Prime 2.0. The participant was positioned so that their head was 100cm from the monitor. The stimulus was adjusted to 4 degrees of visual angle. The participant then received task instructions followed by 24 practice trials with feedback. After completing these trials, the participant was asked if they had any questions, and then was reminded to respond as quickly and as accurately as possible and to blink only after a response was made.

Each trial began with a randomly jittered fixation screen that lasted 500 to 1000 ms. Then, the name of one of the three properties appeared near the fixation point (see Figure 1), also for 500 to 1000 ms, before it disappeared (for 500 to 1000 ms) and was replaced by the stimulus shapes, which remained visible until a button press was made. There were no systematic difference in any of these jittered times between conditions. The entire experiment consisted of 216 trials, and accuracy and response times (RT) were measured. Participants completed four blocks of 54 trials, with conditions and stimuli randomized within and across blocks. One-minute breaks were given between blocks, during which cumulative mean accuracy and RT were reported to the participant.

## WM Span

After completing the visual analogy task, participants completed a 15-20 min operation span WM task (Conway et al., 2005). On each trial, participants were asked to verify a simple mental arithmetic problem and then were to remember a letter. Trials were from 2 to 7 problems/letters long. At the conclusion of a trial, participants were presented with an array of letters and were to click the letters in the sequence they were presented. WM span was defined as the total number of letters correctly remembered in the presented order. All participants performed the math problems at 85% correct or better.

## Results

### Behavioral Results

Because yes-valid/no-invalid responses were used, we report accuracy using  $d$ -prime<sup>1</sup>. Participants were less accurate in the Valid-Distraction ( $M=3.0$ ,  $SEM=.13$ ) than the Valid-No Distraction ( $M=3.2$ ,  $SEM=.16$ ) condition  $F(1,24) = 17$ ,  $p < .001$ ,  $\eta_p^2=.4$ ); however, there was no

<sup>1</sup> Hit rates of 1 were replaced with .99 and hit rates of 0 were replaced with .01 for purposes of calculating  $d$ -prime.

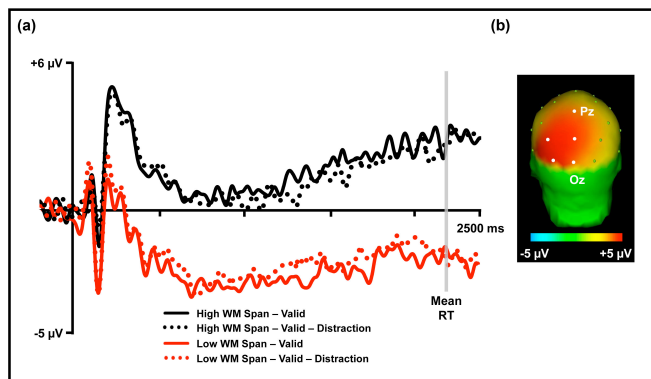


difference in RT between Valid-Distractor ( $M=2.2s$ ;  $SEM=.08$ ) and Valid-No Distractor ( $M=2.2s$ ;  $SEM=.09$ ;  $F(1,24) = .001$ ,  $ns$ ,  $\eta_p^2<.001$ ). Additionally, there was no interaction with WM-span group for either accuracy ( $F(1,24) = .46$ ,  $p = .5$ ,  $\eta_p^2=.02$ ) or RT ( $F(1,24) = 1.3$ ,  $p=.3$ ,  $\eta_p^2=.05$ ). Thus, we saw an effect of relational distraction even in these relational simple problems; however, this effect appeared not to be moderated by WM span.

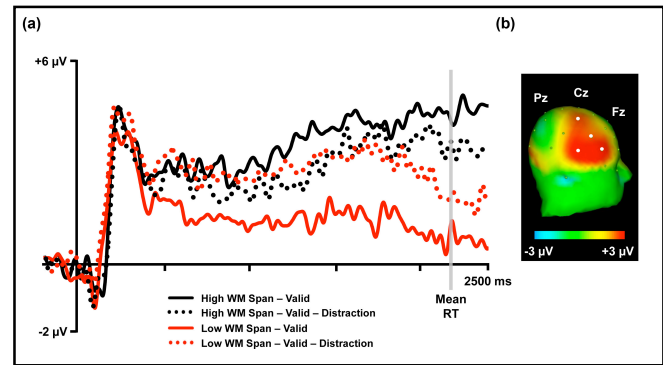
## EEG Results

**Individual Differences in WM** Our first predication was that WM-span would be reflected in the CNV (Vogel et al., 2005) in occipitoparietal regions. We believed this would result from low WM-span individuals' inferior ability to ignore goal-irrelevant information during analogical processing, similar to the effect observed by Vogel et al. (2005) during a delay period in a delay match-to-sample task. Consistent with this prediction we found that participants in the low WM-span group showed a more negative CNV (see Figure 2) from just after the occipital N1 wave (180ms) all the way through the end of task processing (1700ms;  $F(1,24) = 7.1$ ,  $p = .01$ ,  $\eta_p^2=.2$ ). This effect was not modulated by relational distraction ( $F(1,24) = .2$ ,  $p = .7$ ,  $\eta_p^2 = .008$ ).

**Effect of Relational Distraction** Our second prediction involved the role of PFC in managing distraction during analogical mapping. Using a similar analogy task with fMRI, Cho and colleagues (2010) had previously shown areas in middle and inferior frontal gyri were sensitive to relational distraction. We further hypothesized that this effect would be late in processing, coincident with analogical mapping (Morrison et al., 2012). We did not find a main effect of relational distraction (see Figure 3;  $F(1,24) = .6$ ,  $p = .4$ ,  $\eta_p^2=.024$ ); however, we did find an area in right



**Figure 2:** Modulation of the neural correlates of visuospatial working memory during analogical reasoning by WM-span. (a) A CNV ERP was more negative for the low than the high WM-span group in occipitoparietal electrodes (indicated by white dots in Figure 2b). (b) Map showing high minus low WM-span subtraction topography from 180 to 1700ms post-stimulus.



**Figure 3:** Modulation of the neural correlates of relational distraction by WM-span. (a) The Valid-No-Distractor ERP for low WM-span participants was significantly less positive than either the Valid-Distractor ERP for the same group, or either condition ERP for the high WM-span group in right frontal electrodes (indicated by white dots in Figure 3b). (b) Map showing double subtraction topography from 1000 to 1700ms post-stimulus.

PFC consistent with Cho et al. (2010) which showed an interaction between WM-span and relational distraction ( $F(1,24) = 7.1$ ,  $p = .01$ ,  $\eta_p^2=.2$ ). While the high WM-span group did not show a difference between the Valid-Distractor and Valid-No Distractor conditions ( $F(1,12) = 5.1$ ,  $p = .3$ ,  $\eta_p^2=.07$ ), the low WM-span group did ( $F(1,24) = 1.0$ ,  $p = .01$ ,  $\eta_p^2=.2$ ). Specifically, both conditions for the high WM-span group showed similar levels to the Valid Distractor condition for the low WM-span group, while the Valid-No Distractor condition for the low WM-span group was reliably less positive.

## Discussion

In many real-world problem-solving situations, one may have to choose between multiple common relations for use as the source for an analogy. For instance, a molecular biologist may want to favor repeat sequences instead of common codons as the basis for an analogy between two genes. Attending to codons when one is looking for repeat sequences may be misleading, thus good attention to just the chosen relation (e.g., a repeat sequence) and inhibition of the irrelevant information (e.g., a specific codon) for the situation is most efficient. However, faced with a new situation, codons may now be the better relation to use, so the system must be flexible to solve the problem at hand. In the paradigm used in this study, reasoners were sometimes asked to favor the relation based on one stimulus characteristic over another as the basis for their analogy. However, their focus needed to be flexible because what was critical on one trial may be misleading on the next. Using this paradigm we demonstrated two ways in which inhibitory control can influence analogical processing.

First, as in Vogel and colleague's (2005) demonstration using EEG with a delayed match to sample task, we found

EEG evidence that low WM-span individuals allowed more information to enter visuospatial WM than high WM-span individuals. It is likely that this difference resulted from low-WM span individuals considering goal-irrelevant relations regardless of condition. This occurred early in the reasoning time course, beginning just after the first signals of spatial attention (i.e., the occipital N1), and continued throughout the trial time course. While, this did not result in a difference in behavioral performance on the task, one can imagine that in a more difficult task (e.g., at higher levels of relational complexity), this inefficient gating of WM may have behavioral consequences.

Secondly, we found sensitivity to relational distraction as measured by a subtraction between Valid-No Distraction and Valid-Distraction trials in a right frontal ERP was modulated also by WM-span. It appears that high WM-span individuals frequently engage this area of the brain during later stages of analogical mapping, while low WM-span individuals engage it more in the face of distraction. We believe this aspect of inhibitory control is likely distinct from the previous result. If high WM-span individuals are better at using early inhibitory control to gate WM from non-goal related information one might hypothesize that they would require the later PFC mechanisms less; however, it appears that high WM-span individuals use it consistently, and on average more than low WM-span individuals. In contrast, low WM-span individuals, could certainly make use of the PFC mechanism for Valid-No Distraction trials to focus on just the relevant relations; however, they don't, using it only when actual conflict is detected. Thus, the use of PFC seems to be more reactive in the case of low WM-span individuals, while it is more proactive with high WM-span individuals. Thus, this appears to be a second neural mechanism that favors high WM-span individuals.

This result is also consistent with the results from a previous fMRI study using a very similar task that identified areas in bilateral inferior frontal gyrus (IFG) as being more active during analogy in the presence of relational distraction (Cho et al., 2010). The EEG topography shown in Figure 3 is consistent with the activation reported by Cho and colleagues and also many other studies investigating the role of inhibitory control in WM (e.g., Goel et al., 2000; Goel & Dolan, 2003; Prado & Noveck, 2007; De Neys et al., 2008). Also, as in Cho and colleague's study the topography resulting from the distraction contrast appears to be at least partially distinct from the frontopolar area previously identified as being associated with analogical mapping via both EEG (Morrison et al., 2012) and fMRI (Green et al., 2010) methods. Future investigations will need to focus on how frontopolar PFC and IFG may interact in the service of analogical reasoning.

Several previous behavioral studies have shown evidence of the importance of inhibitory control during analogical reasoning in the face of distraction (e.g., Cho et al., 2007; Krawczyk et al., 2008; Morrison et al., 2004; Viskontas et al. 2004); however, these studies typically only found

reliable effects when distraction was present in more relationally complex problems.<sup>2</sup> In the present study we show the engagement of inhibitory control for even simple one-relation analogy problems. However, this effect was apparent only in the ERP results, and accuracy, not RT.

So what exactly does inhibitory control do during analogical reasoning? Given our results it is quite likely that the answer is not a unitary one. Likewise, in the LISA model of analogical reasoning (Hummel & Holyoak, 1997, 2003; Knowlton et al., 2012), the function of inhibitory control may be multifaceted, and may differ across the time course of processing analogies. Future neuroimaging studies will be driven by precise computational accounts of the neural mechanisms underlying analogical processing (e.g., Knowlton et al., 2012; Morrison et al., 2012) and will likely require EEG time-frequency analysis techniques to appreciate the temporal dynamics of the neural circuits responsible for analogical reasoning.

## Acknowledgments

The authors thank Keith Holyoak and Rebecca Silton for comments on an earlier draft of this paper. The Carbon Undergraduate Research Fellowship Program at Loyola University Chicago (BMS, KLB, RGM), the American Federation of Aging Research/Rosalinde and Arthur Gilbert Foundation (RGM), the Illinois Department of Public Health (RGM), and the Loyola University Chicago Deans of Arts and Sciences and the Graduate School (RGM) provided generous support.

## References

- Bunge, S. A., Helskog, E. H., & Wendelken, C. (2009). Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *NeuroImage*, 46(1), 338-342. doi:10.1016/j.neuroimage.2009.01.064
- Bunge, S. A., Wendelken, C., Badre, D., & Wagner, A. D. (2005). Analogical reasoning and prefrontal cortex: Evidence for separable retrieval and integration mechanisms. *Cerebral Cortex*, 15(3), 239-249. doi:10.1093/cercor/bhh126
- Cho, S., Holyoak, K. J., & Cannon, T. D. (2007). Analogical reasoning in working memory: Resources shared among relational integration, interference resolution, and maintenance. *Memory & Cognition*, 35(6), 1445-1455.
- Cho, S., Moody, T. D., Fernandino, L., Mumford, J. A., Poldrack, R. A., Cannon, T. D., . . . Holyoak, K. J. (2010). Common and dissociable prefrontal loci associated with component mechanisms of analogical reasoning. *Cerebral Cortex*, 20(3), 524-533. doi:10.1093/cercor/bhp121

<sup>2</sup> One notable exception is distraction effects during analogical reasoning in young children (Richland et al., 2006; 2010).

- Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, 12, 769-786.
- De Neys W, Vartanian O, Goel V. 2008. Smarter than we think: when our brains detect that we are biased. *Psychological Science*, 19, 483-489.
- Doumas, L.A.A., Morrison, R.G., & Richland, L.E. (2012). *Individual differences in relational learning and analogical reasoning: A computational approach*. Manuscript submitted for publication (copy on file with author).
- Goel V, Buchel C, Frith C, Dolan RR. 2000. Dissociation of mechanisms underlying syllogistic reasoning. *Neuroimage*, 12, 504-515.
- Goel V, Dolan RJ. 2003. Explaining modulation of reasoning by belief. *Cognition*, 87, B11-B22.
- Green, A., Kraemer, D.J.M., Fugelsang, J., Gray, J.R., & Dunbar, K. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex*, 10, 70-76.
- Halford, G.S. (1992) Analogical reasoning and conceptual complexity in cognitive development. *Human Development*, 35, (4), 193-217.
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104, 427-466.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110, 220-264.
- Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, 9(4), 637-671.
- Knowlton, B.J., Morrison, R.G., Hummel, J.E., & Holyoak, K.J. (2012). *A Neurocomputational System for Relational Reasoning*. Manuscript submitted for publication (copy on file with author).
- Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, J., Chow, T. W., Mendez, M. F., Miller, B. L., & Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia*, 46, 2020-2032.
- Morrison, R.G. (2005). Thinking in working memory. In K.J. Holyoak & R.G. Morrison (Eds.), *Cambridge Handbook of Thinking and Reasoning*. New York, NY: Cambridge University Press.
- Morrison, R.G., Doumas, L.A.A., & Richland, L.E. (2011). A computational account of children's analogical reasoning: Balancing inhibitory control in working memory and relational representation. *Developmental Science*, 14(3), 516-529. doi:10.1111/j.1467-7687.2010.00999.x
- Morrison, R.G., Krawczyk, D., Holyoak, K.J., Hummel, J.E., Chow, T., Miller, B., & Knowlton, B.J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, 16, 260-271. doi:10.1162/089892904322984553
- Morrison, R.G., Nikitin, V., & Bharani, K.L. (2012). *Neurocorrelates of analogical mapping: An event-related potential approach*. Manuscript submitted for publication (copy on file with author).
- Morrison, R.G., Holyoak, K.J., & Truong, B. (2001). Working memory modularity in analogical reasoning. *Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society* (pp. 663-668). Mahwah, NJ: Erlbaum.
- Prado J, Noveck IA. 2007. Overcoming perceptual features in logical reasoning: a parametric functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience*, 19, 642-657.
- Richland L.E., Chan, T-K., Morrison, R.G., & Au, T.K-F. (2010). Young children's analogical reasoning across cultures: Similarities and differences. *Journal of Experimental Child Psychology*, 105, 146-153.
- Richland, L.E., Morrison, R.G., & Holyoak, K.J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, 94, 249-273.
- Shimamura, A. P.(2000). The role of the prefrontal cortex in dynamic filtering. *Psychobiology*, 28, 207-218.
- Thibaut, J., French, R., & Vezneva, M. (2010a). Cognitive load and semantic analogies: Searching semantic space. *Psychonomic Bulletin & Review*, 17(4), 569-574. doi:10.3758/PBR.17.4.569
- Thibaut, J., French, R., & Vezneva, M. (2010b). The development of analogy making in children: Cognitive load and executive functions. *Journal of Experimental Child Psychology*, 106(1), 1-19. doi:10.1016/j.jecp.2010.01.001
- Viskontas, I. V., Morrison, R. G., Holyoak, K. J., Hummel, J. E., & Knowlton, B. J. (2004). Relational integration, inhibition, and analogical reasoning in older adults. *Psychology and Aging*, 19(4), 581-591.
- Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, 438, 500-503.
- Waltz, J. A., Lau, A., Grewal, S. K., & Holyoak, K. J. (2000). The role of working memory in analogical mapping. *Memory & Cognition*, 28, 1205-1212.

# Effects of explicit knowledge on transfer of visuomotor sequence learning

**Kanji Tanaka (kanji@fennel.rcast.u-tokyo.ac.jp)**

<sup>1</sup>Research Center for Advanced Science and Technology, The University of Tokyo

<sup>2</sup>Japan Society for the Promotion of Science

**Katsumi Watanabe (kw@fennel.rcast.u-tokyo.ac.jp)**

<sup>1</sup>Research Center for Advanced Science and Technology, The University of Tokyo

## Abstract

Skilled, sequential movements can be acquired explicitly or implicitly. In the present study, we examined the effects of explicit knowledge obtained through instruction or spontaneous detection on transfer of visuomotor sequence learning. In the first session, participants learned a visuomotor sequence by trial and error. In subsequent sessions, the sequence was changed according to specific rules. Some participants received explicit instruction about which specific rules changed, while the others did not. Knowledge of changes via explicit instruction led to slower performance with fewer errors; the sluggishness persisted even in the last phase of transfer learning. On the other hand, knowledge discovered independently by the participants produced slower performance in the initial phase of learning with fewer errors, but their performance speed eventually reached the same level as that of the unaware participants. These results suggest that explicit knowledge may help to reduce errors in the initial phase of visuomotor sequence learning but may interfere with increasing speed, particularly when the knowledge is given rather than found.

**Keywords:** Sequence learning; Explicit knowledge; Transfer; Interference

## Introduction

Skilled sequential movements, such as typing on a keyboard, dialing a phone number, or playing the piano, have key roles in motor behavior in daily life. Many studies have examined how people acquire or improve such sequential behaviors. One of the most common ways to investigate the acquisition of sequential learning is called the Serial Reaction Time (SRT) task (Nissen & Bullemer, 1987). In this task, participants press one of four aligned buttons that are associated with visual stimuli at different positions. In several experimental blocks, a specific sequence was repeated or partly repeated, but participants were not informed of the repetition and did not notice the structure of the sequence (because the sequence was long enough; for example, see Reed & Johnson, 1994; Reber & Squire, 1988). Nevertheless, a reduction in response time occurs without awareness of the sequence structure. In another paradigm, participants first learn a sequence of finger movements and then performance is measured in terms of speed and accuracy (i.e., this paradigm focuses on improvement in performance rather than acquisition) (Karni et al., 1998; Walker et al., 2003).

Explicit knowledge (e.g., whether or not participants notice a repeated sequence and whether or not they were instructed about the sequence explicitly) likely leads to various changes in performance and learning processes in terms of speed and accuracy. In general, implicit learning in the SRT task facilitates response even when participants do not notice a specific sequence (Nissen & Bullemer, 1987). Curran and Keele (1993) assigned participants into three groups: the intentional, more aware, and less aware groups. In the intentional group, the participants were instructed about the repeated sequence before completing the task. The authors defined the participants who were aware of the repeated sequence as the more-aware group and those who were not aware of the sequence as the less-aware group. Participants in that study conducted a single task first, then a dual task. In the single-task blocks, participants were required to press certain buttons when an X appeared on the screen. In the dual-task blocks, a tone sounded several times within the 200-ms stimulus-response interval of the primary task, and the tones were composed of low and high pitches. The participants were instructed to count the number of high-pitched tones while ignoring the low-pitched tones. The results showed that all three groups differed in terms of performance time in the single task: explicit knowledge conveyed intentionally before starting the task led to faster performance during the task. However, in the dual task, the groups did not differ in terms of performance times. They interpreted the results as signifying that awareness affects sequence learning only when attention is fully available; when attention was divided, additional knowledge gained by the aware and intentional groups was not conveyed to the non-attentional mechanism. Similarly, in a previous study (Moisello et al., 2011), participants performed visuomotor sequence learning of a finger opposition task, and response time (i.e., the interval between stimulus presentation and the onset of the corresponding touch) and touch duration (i.e., the contact time between thumb and another finger) were measured for each finger opposition movement of the sequence. The results indicated that sequence learning induced a double-faced effect on motor performance: the participants who were instructed explicitly about a sequence decreased their reaction times, but increased their touch duration, which was regarded as the combination of a sensory phase and a motor preparation phase. Thus, whether

explicit knowledge enhances or interferes with performance in sequential learning is still under debate.

In the present study, we examined the effects of explicit knowledge on transfer of visuomotor sequence learning by using another sequential learning paradigm employing instruction or spontaneous detection. The task we employed was a sequential button press task that was originally devised for monkeys as participants (Hikosaka et al., 1995; Rand et al., 1998) and subsequently used on humans (Hikosaka et al., 1995, 1996, 2002; Sakai et al., 1998, 2003) and with which performance improvements in terms of speed and accuracy can be measured separately (Hikosaka et al., 1995, 1996, 2002; Sakai et al., 1998, 2003; Watanabe et al., 2006, 2010). In the first session, participants were required to complete a visuomotor sequence learning task without any instruction (i.e., by trial and error). In the subsequent sessions, the sequence was altered according to a specific rule; either the first and the third responses were switched (reversed) or the first and the second responses were switched (partially reversed). Some participants were explicitly instructed about the specific rule changes before performing the task, while the others did not receive explicit instructions. We also examined whether participants who spontaneously noticed the specific rules without instruction used their knowledge to change their performance.

## Method

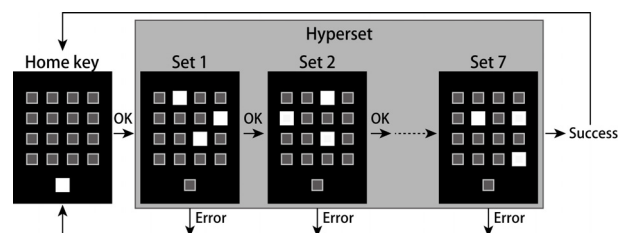
### Participants

Thirty-seven right-handed participants (23 male, 14 female; mean age = 21.02 years, standard deviation = 2.39) participated in the experiment. They were divided into two groups: the explicit-instruction group (14 participants; 8 male, 6 female) and the no-instruction group (23 participants; 15 male, 8 female). All participants had normal or corrected-to-normal visual acuity, normal motor functioning, and were naïve to the purpose of the study. All participants gave written informed consent prior to participation. All procedures were conducted in accordance with the Declaration of Helsinki.

### Procedure

The participants performed a sequential button press task, which we call “3 × 7 task.” We used essentially the same experimental paradigm as previous studies (e.g., Sakai et al., 2003; Watanabe et al., 2006, 2010). Sixteen LED buttons (10 mm × 10 mm each) were mounted on a panel in a 4 × 4 matrix and were separated from each other by 8-mm spaces. At the bottom of the panel was another LED button, which was used as the “home” key. The participant used his/her right index finger to press the buttons. The home key was turned on at the beginning of each trial. When the participant pressed the home key for 500 ms, 3 out of the 16 target LEDs turned on simultaneously, which we called the “set.” The participant was required to press the illuminated buttons in the correct order, which he/she was required to

discover by trial and error. Upon success, the LEDs turned off one by one and a different triplet of LEDs was illuminated; the participant was again required to discover the correct order and press the buttons accordingly. When the participant pressed an incorrect button, all the LED buttons were briefly illuminated, a beep sounded, and then the trial was aborted. The participant then had to restart the trial by pressing the home key. A total of seven sets, which we call the “hyperset,” were presented in a fixed order for trial completion. A trial was considered successful when the participant completed an entire hyperset (seven sets). The same hyperset was repeated until the participant completed it successfully for 20 trials (called a “block”). Participants were asked to perform the task as quickly and as accurately as possible. We prepared three hypersets, called “original,” “reversed,” and “partially reversed.” The original hyperset was randomly generated for each participant. For each set, the triplet of buttons were defined [1][2][3] in the to-be-pressed order. In other hypersets, the spatial configurations of the sets were not changed, but the sequences of correct button presses were changed. In the reversed hypersets, the participants needed to press the buttons in the order [3][2][1] (i.e., the first and the third buttons were switched). In the partially reversed hypersets, they needed to press the buttons in the order [2][1][3] (i.e., the first and second buttons were switched).

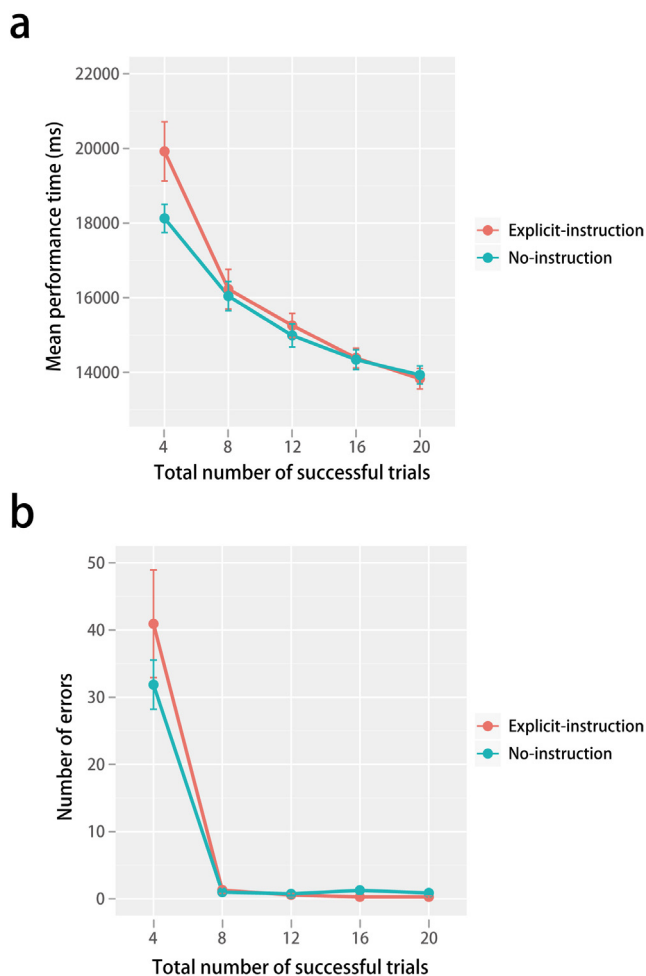


**Figure 1.** Schematic flow of the experiment. Participants were instructed to learn the correct order of button presses by trial and error. The LED buttons were square-shaped (10 mm × 10 mm) and 8 mm apart.

The participants conducted one session using the original hypersets, followed by two sessions using the reversed and the partially reversed hypersets. The order of the final two test sessions was counterbalanced across participants. The three sessions were separated by a 5-minute break. In the explicit-instruction group, the rules of the reversed and the partially reversed orderings were explicitly told to participants before the beginning of each session. In the no-instruction group, they were told that the sets were randomly assigned. For the no-instruction group, the participants were asked whether they had noticed anything peculiar after each session. If the answer was yes, then they were questioned further and asked if they had noticed the rules of changes in the hypersets. Then, participants who noticed the rules were defined as the aware group, and those who did not notice the rules were defined as the unaware group. Hence, we eventually had three groups: the explicit, aware, and unaware groups.

## Data Analysis

We used two measures to assess the accuracy and speed of performance in each block. As a measure of accuracy, we counted the number of errors before correctly completing each trial. To evaluate speed, we measured the time that elapsed from the moment the home key was pressed to the moment the third button of the final (7th) set was pressed for each successful trial. Similar parameters were employed in previous studies and proved to be useful (Hikosaka et al., 1999, 2002; Watanabe et al., 2006, 2010). We calculated average performance using each hyperset by assessing five sections of cumulative successes (i.e., 1st to 4th, 5th to 8th, 9th to 12th, 13th to 16th, and 17th to 20th). For the reversed and partially reversed hypersets, we normalized performance time by using each participant's performance in the last section (i.e., mean performance from the 17th to 20th trials) in which the original hypersets were used as the baseline. Mean performance times of the five sections using the reversed and partially reversed hypersets were subtracted from performance times at baseline.



**Figure 2.** Performance changes in the first block with the learning hyperset. Error bars show the standard error of the means. (a) Average performance time for successful trials.

(b) Average number of errors before the successful completion of each trial.

## Results

### Learning session with original hyperset

A significant decrease was found in both the accuracy (the number of completion failures) and speed measures (averaged completion time for successful trials) in the first session with the original hyperset irrespective of the group (Figure 2), indicating that learning did occur [ANOVA;  $F(4, 140) = 51.44$ ,  $p < 0.0001$ ; for both measures]. The accuracy measure decreased rapidly in the first few completed trials while the speed measure decreased more gradually. These results are in accordance with those of previous studies (Hikosaka et al., 1995, 1996, 2002; Sakai et al., 1998, 2003; Watanabe et al., 2006).

### Comparison between explicit-instruction and no-instruction groups

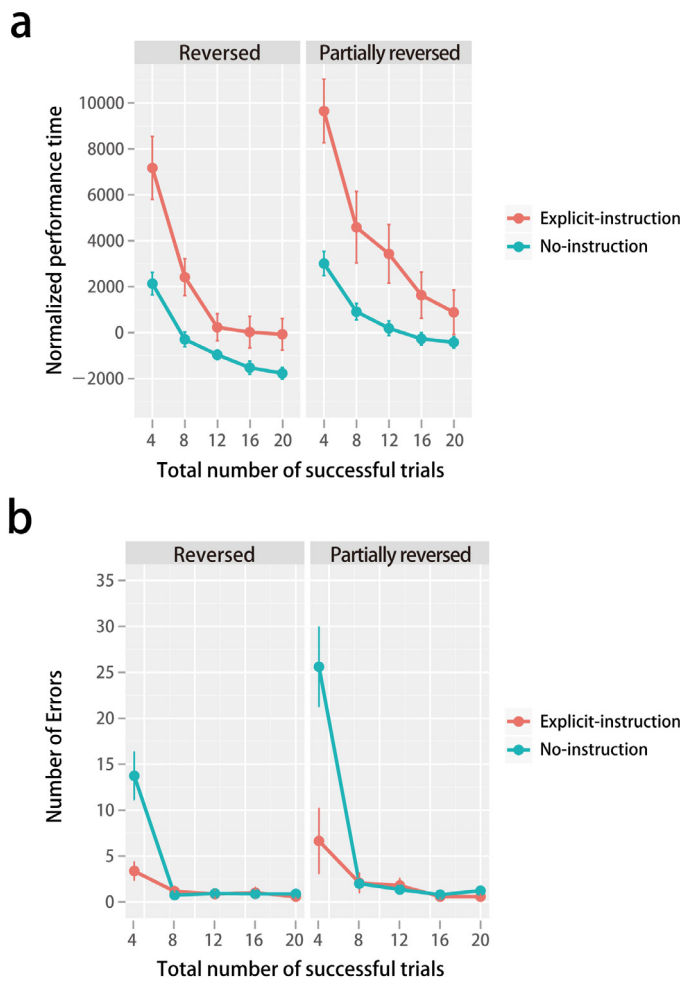
First, we examined the effects of explicit instruction on the  $3 \times 7$  task. Explicit instruction slowed down performance. The partially reversed hypersets also slowed performance. A three-way ANOVA revealed significant main effects of instruction ( $F(1, 35) = 13.84$ ,  $p < 0.001$ ; explicit > no-instruction), hyperset ( $F(1, 35) = 40.78$ ,  $p < 0.0001$ ; partially reversed > reversed), and trial section ( $F(4, 140) = 145.85$ ,  $p < 0.0001$ ; post hoc, Shaffer's method, 1st > 2nd > 3rd > 4th > 5th). A significant interaction was also found between instruction and trial section ( $F(4, 140) = 30.34$ ,  $p < 0.001$ ).

On the other hand, explicit instruction had a positive effect on the accuracy measure, but only for the first sections using the reversed and partially reversed hypersets. A three-way ANOVA revealed significant main effects of instruction ( $F(1, 35) = 8.35$ ,  $p < 0.01$ ; no-instruction > explicit), hyperset ( $F(1, 35) = 12.44$ ,  $p < 0.01$ ; partially reversed > reversed), and trial section ( $F(4, 140) = 29.15$ ,  $p < 0.001$ ; post hoc, Shaffer's method, 1st > all other sections). A significant interaction was also found between instruction and trial section ( $F(4, 140) = 12.33$ ,  $p < 0.001$ ).

### Comparison between aware and unaware groups

Among 23 participants in the no-instruction group, 8 participants spontaneously noticed the specific rules of both the reversed and the partially reversed hypersets during the experiment; hence, they were classified as the "aware" group. The other 15 participants did not notice either of the rules during the experiment and were classified as the "unaware" group.

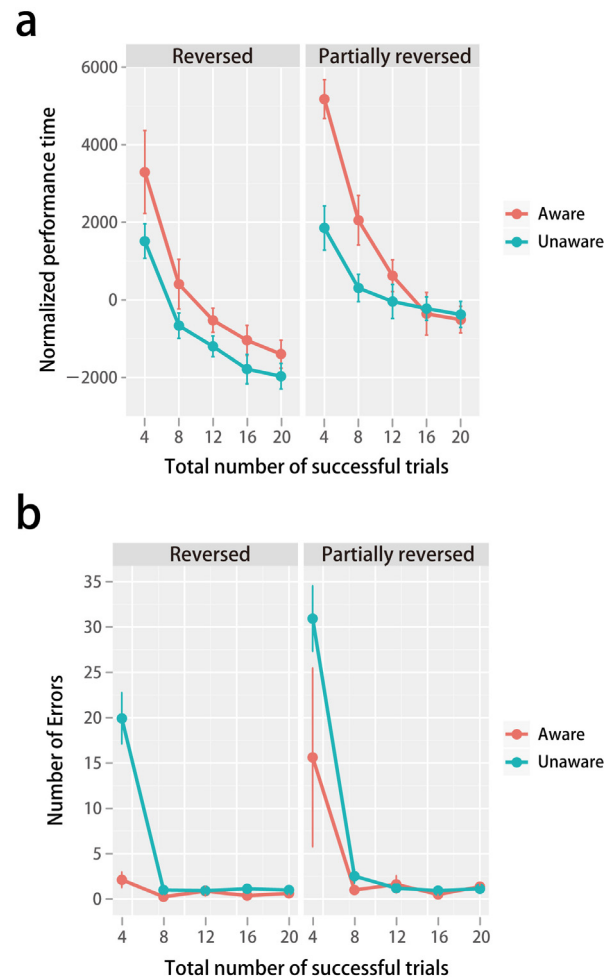




**Figure 3.** Performance of the explicit and no-instruction groups in the second and third sessions using the reversed and partially reversed hypersets. Error bars show the standard error of the means. (a) Average normalized performance times. (b) Average number of errors before the successful completion of each trial.

Figure 4a shows the mean normalized performance times of the aware and unaware groups. The aware group exhibited slower normalized performance times when using both the reversed and partially reversed hypersets. A three-way ANOVA revealed significant main effects of instruction ( $F(1, 21) = 4.99, p < 0.05$ ; aware > unaware), hyperset ( $F(1, 21) = 16.83, p < 0.001$ ; partially reversed > reversed), and trial section ( $F(4, 84) = 116.51, p < 0.0001$ ; post hoc, Shaffer's method, 1st > 2nd > 3rd > 4th = 5th). A significant interaction between instruction and trial section was also revealed ( $F(4, 84) = 10.34, p < 0.0001$ ). The number of errors observed was higher in the aware group's first section compared to the unaware group's first section (Figure 4b). A three-way ANOVA revealed significant main effects of instruction ( $F(1, 21) = 8.43, p < 0.01$ ; unaware > aware), hyperset ( $F(1, 21) = 13.49, p < 0.01$ ; partially reversed > reversed), and trial section ( $F(4, 84) = 35.20, p <$

$0.001$ ; post hoc, Shaffer's method, 1st > all other sections). A significant interaction between instruction and trial section was also found ( $F(4, 84) = 8.87, p < 0.001$ ).



**Figure 4.** Performance of the aware and unaware groups in the second and third sessions using the reversed and partially reversed hypersets. Error bars show the standard error of the means. (a) Average normalized performance times. (b) Average number of errors before the successful completion of each trial.

## Discussion

In the present study, we examined the effects of the acquisition of explicit knowledge through instruction or spontaneous detection on transfer of visuomotor sequence learning. Knowledge of changes via explicit instruction led to slower performance with fewer errors, and this sluggishness persisted even in the last phase of learning. On the other hand, knowledge discovered spontaneously by participants produced slower performance in the initial phase of learning with fewer errors, but performance speed tended to peak at the same level as that of the unaware participants (using the partially reversed hyperset). These



results suggest that explicit knowledge may help to reduce errors in the initial phase but may interfere with increasing speed, particularly when the knowledge is given rather than found.

The participants who spontaneously noticed the presence of the rules showed slower performance than those who did not notice the rules in the initial phase of the session, which they performed with fewer errors. The lack of influence of explicit knowledge on the later stages of learning is consistent with the two-loop model of visuomotor sequence learning (Nakahara et al., 2001) and with previous studies (Watanabe et al., 2006). Watanabe et al. (2006) examined the influence of explicit knowledge of stimulus configuration (workspace) in visuomotor sequence learning, and their experimental paradigm was essentially the same as that of the present study. After the first session (i.e., learning a specific visuomotor sequence by trial and error), the workspace was rotated for the second session without notifying the participants. It is noteworthy that participants who noticed the rules of rotation did not improve in terms of performance time, though they were able to use their explicit knowledge of the rotation. In the task employed in the present study, explicit knowledge of the sequence is critical for performing and proceeding through the task (as in other learning paradigms that involve explicit instructions of sequences; Jueptner et al., 1997a, 1997b; Karni et al., 1995). For other types of procedural learning, including rotary movement pursuit and mirror tracing, explicit knowledge has little effect on the accuracy and/or speed of task performance (e.g., Heindel et al., 1989). Differences in the necessity of explicit knowledge may explain this discrepancy between results.

Slower performance by participants who received explicit instruction might appear contradictory with the results of previous studies. In the SRT task, explicit knowledge given before the task could lead to faster performance during the task (Curran & Keele, 1993). One possible interpretation is that the role of explicit knowledge may differ in different paradigms of sequential learning because spatial sequence is learned by trial and error in the  $3 \times 7$  task, whereas spatial sequence is defined in a stimulus-driven way in the SRT task (Curran & Keele, 1993; Nissen & Bullemer, 1987; Willingham et al., 1989). Another possible interpretation is that the effects of explicit knowledge may depend on the demands of the task. Curran and Keele (1993) showed that explicit knowledge of a to-be-learned sequence led to faster performance in a single task but not in a dual task, and they implied that explicit knowledge facilitated sequence learning only when attention was fully available. In the present study, the participants were required to complete a task without instruction first, and then the participants in the explicit-instruction group conducted the task with explicit knowledge of the rule changes. In other words, the participants were required to maintain the prior order and the instruction to reverse the original hyperset, whereas participants in the SRT task were required only to retain information about which button to press. This difference

also might be related to the capacity limit of working memory. Previous work showed that if an individual is asked to hold words in working memory and to judge whether a probe word was one of the retained words, response time increased with memory set size (e.g., McElree & Doshier, 1989; Sternberg, 1969). In the present study, the performance speed of the explicit-instruction group did not reach an equal level to that of the no-instruction group even in the final phase. Thus, explicit knowledge given by another person thoroughly hindered performance speed for the duration of the experiment. As for individual differences in hindrance, participants who have high working memory capacities might not be influenced, and vice versa. Clarification of this proposal would require further investigation.

The present findings can be exemplified by a more familiar hypothetical case. Assume that you are dialing a phone number that you know well. In this case, you can dial it quickly. If you are asked to dial a new phone number that is actually the reverse of the well-known phone number, and you do not notice this fact, you will become able to dial it fast. If you notice that the new phone number is the reverse of the well-known phone number on your own, it would slow your learning, but you will eventually be able to dial the new number quickly. However, if you are explicitly asked to dial the reverse of the well-known phone number, you will not be able to dial the reversed phone number as quickly as the original phone number. It is noteworthy that in the final phase, the performance speed in the explicit-instruction group did not reach the same level as that in the no-instruction group. This could be because attentional resources need to be partly devoted to holding explicit knowledge, which reduces the efficacy of learning. However, further investigation is warranted in order to elucidate how explicit knowledge interferes with learning.

## Acknowledgments

This work was supported by Grant-in-Aid for JSPS Fellows and Japan Science and Technology Agency.

## References

- Curran, T., Keele, S. W. (1993). Attentional and nonattentional forms of sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 189-202.
- Heindel, W. C., Salmon, D. P., Shults, C. W., Walicke, P. A., & Butters, N. (1989). Neuropsychological evidence for multiple implicit memory systems: a comparison of Alzheimer's, Huntington's, and Parkinson's disease patients. *Journal of Neuroscience*, 9, 582-587.
- Hikosaka, O., Rand, M. K., Miyachi, S., & Miyashita, K. (1995). Learning of sequential movements in the monkey: process of learning and retention of memory. *Journal of Neurophysiology*, 74, 1652-1661.
- Hikosaka, O., Sakai, K., Miyauchi, S., Takino, R., Sasaki, Y., & Putz, B. (1996). Activation of human

- presupplementary motor area in learning of sequential procedures: a functional MRI study. *Journal of Neurophysiology*, 76, 617-621.
- Hikosaka, O., Nakamura, K., Sakai, K., Nakahara, H. (2002). Central mechanisms of motor skill learning. *Current Opinion in Neurobiology*, 12, 217-222.
- Jueptner, M., Stephan, K., Frith, C., Brooks, D., Frackowiak, R., & Passingham, R. (1997a). The anatomy of motor learning. I. Frontal cortex and attention to action. *Journal of Neurophysiology*, 77, 1313-1324.
- Jueptner, M., Stephan, K., Frith, C., Brooks, D., Frackowiak, R., & Passingham, R. (1997b). The anatomy of motor learning. II. Subcortical structures and learning by trial and error. *Journal of Neurophysiology*, 3, 1325-1337.
- Kami, A., Meyer, G., Jezzard, P., Adams, M. M., Turner, R., & Ungerleider, L. G. (1995). Functional MRI evidence for adult motor cortex plasticity during motor skill learning. *Nature*, 377, 155-158.
- McElree, B., & Doshier, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General*, 118, 346-373.
- Moisello, C., Avanzino, L., Tacchino, A., Ruggeri, P., Ghilardi, M. F., & Bove, M. (2011). Motor sequence learning: acquisition of explicit knowledge is concomitant to changes in motor strategy of finger opposition movements. *Brain research bulletin*, 85(3-4), 104-108.
- Nakahara, H., Doya, K., Hikosaka, O. (2001). Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences: a computational approach. *Journal of Cognitive Neuroscience*, 13, 626-647.
- Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: evidence from performance measures. *Cognitive Psychology*, 19, 1-32.
- Rand, M. K., Hikosaka, O., Miyachi, S., Lu, X., & Miyashita, K. (1998). Characteristics of a long-term procedural skill in the monkey. *Experimental Brain Research*, 118, 293-297.
- Reber, P. J., & Squire, L. R. (1998). Encapsulation of implicit and explicit memory in sequence learning. *Journal of Cognitive Neuroscience*, 11, 248-263.
- Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 585-594.
- Sakai, K., Hikosaka, O., Miyachi, S., Takino, R., Sasaki, Y., & Putz, B. (1998). Transition of brain activation from frontal to parietal areas in visuomotor sequence learning. *Journal of Neuroscience*, 18, 1877-1840.
- Sakai, K., Kitaguchi, K., & Hikosaka, O. (2003). Chunking during human visuomotor learning. *Experimental Brain Research*, 152, 229-242.
- Sternberg, S. (1969). High-speed scanning in human memory. *Science*, 153, 652-654.
- Watanabe, K., Ikeda, H., Hikosaka, O. (2006). Effects of explicit knowledge of workspace rotation in visuomotor sequence learning. *Experimental Brain Research*, 174, 673-678.
- Watanabe, K., Ikeda, H., Miyao, M. (2010). Learning efficacy of explicit visuomotor sequences in children with attention-deficit/hyperactivity disorder and Asperger syndrome. *Experimental Brain Research*, 203, 233-239.
- Willingham, D. B., Nissen, M. J., Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 1047-1060.
- Walker, M.P., Brakefield, T., Hobson, J. A., Stickgold, R. (2003). Dissociable stages of human memory consolidation and reconsolidation. *Nature*, 425, 616-620.

# Spontaneous body movements in spatial cognition

Sergiu TCACI POPESCU (sergiusergiu@gmail.com) and Mark WEXLER (mark.wexler@gmail.com)

CNRS & Université Paris Descartes  
Laboratoire Psychologie de la Perception  
45, rue des Saints-Pères  
75006, Paris, France

## Abstract

People often perform spontaneous body movements during spatial tasks. How are these spontaneous gestures related to spatial problem-solving? We measured spontaneous movements during a perspective-taking task inspired by map reading. Analyzing the motion data to isolate its rotation and translation components in specific geometric relation to the task, we found out that most participants executed spontaneous miniature rotations of the head that were significantly related to the main task parameter. These head rotations were as if participants were trying to align themselves with the orientation on the map, but with tiny amplitudes, typically below 1% of the actual movements. Our results are consistent with a model of sensorimotor prediction driving spatial reasoning. The efference copy of planned movements triggers this prediction mechanism. The movements themselves may then be mostly inhibited; the spontaneous gestures that we measure are the visible traces of these planned but inhibited actions.

**Keywords:** spatial cognition; motor action; sensorimotor prediction; embodied cognition; mental simulation

## Introduction

### Motor activity in spatial tasks

People often perform spontaneous body movements during spatial tasks such as giving complex directions or orienting themselves on maps. Spontaneous gestures in spatial tasks have been studied by Chu & Kita (2011), who showed that their participants spontaneously produced hand gestures while performing a mental rotation task. Motor activity can also trigger mechanisms that simulate the outcome of an action (see Wolpert & Flanagan, 2001, for a review of sensorimotor prediction) and thus infer otherwise unavailable information. For instance, Wexler, Kosslyn, & Berthoz (1998) and Wohlschläger & Wohlschläger (1998) showed that unseen manual rotations improved performance in mental rotation tasks when the mental and manual rotations were in the same direction, and interfered with mental rotation when the two were in opposite directions. The execution of at least some of the visuo-spatial tasks mentioned above includes a motor component that can either improve task performance or interfere with it. This conclusion is supported by the findings of neuroimaging studies (Kosslyn, Ganis, & Thompson, 2001).

### Spatial perspective-taking (SPT)

Spatial perspective-taking occurs when one adopts a viewpoint different from one's physical viewpoint. SPT is more difficult when the imagined perspective differs

from the actual (physical) one by a rotation than by a translation (Rieser, 1989). Performance after an imagined rotation depends on the absolute magnitude of the rotation angle between the actual and the imagined perspective and shows the typical and robust angular disparity effect: the bigger the angle of rotation to the imagined perspective, the lower the performance. More importantly, when people are allowed to move to the location of their imagined or novel perspective, even in absence of visual and auditory cues, performance after perspective rotations is greatly facilitated and may even attain the baseline level.

Spatial updating seems simple and automatic if a person were to perform the full rotations that he or she imagines. The updating is therefore driven by a sensorimotor prediction mechanism, and this mechanism is activated by motor plans or efference copies of the motor command (Wolpert & Flanagan, 2001). The planned action itself could be wholly or partly inhibited further downstream in the motor system. If spontaneous movements are a visible reflection of such simulated but inhibited actions, they should be correlated in some geometrically specific way with the mental task being performed. To determine if this is so was the major goal of our study.

## Methods

24 unpaid participants took part in the experiment. The motion tracking data of 5 participants did not attain our inclusion criterion (see below) and were discarded. We therefore performed all analyses on the data of the remaining 19 participants (8 women, mean age 33.8, standard deviation (SD) 7.1 years).

The participants were told to watch on a computer display a simple map depicting the crossing of two streets. (see Fig. 1). The participants' task was to answer as quickly and accurately as possible if, at the intersection, they needed to turn left or right in order to reach the (red) dot.

The stimuli were parametrized by two variables: the *deviation angle* (see Fig. 1), and the *corner angle* (not shown in Fig. 1). We take the upward orientation as our "zero" because pilot results showed that it is easiest to perform the task when one's initial imagined orientation is upwards. Deviation angles are taken as positive counterclockwise and negative clockwise. The second independent variable, the corner angle, is the angle between the two streets on the map. It was used to mask the similarity between the trials with the same value of the deviation angle.

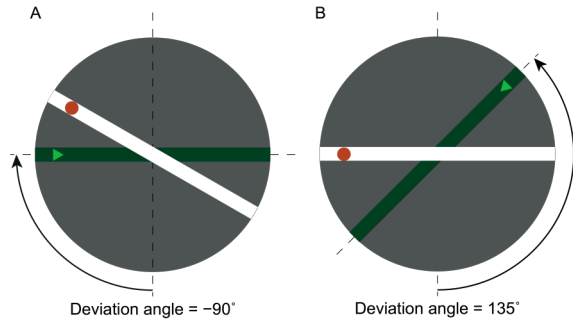


Figure 1: Two examples of stimuli (the dashed lines, the angle arrow and text were not part of the stimuli). Every stimulus represents the crossing of two streets. Participants imagined being on the darker (green) street, at the position of the triangle, and facing the intersection. We call this orientation the *imagined orientation*. The task was to decide if at the intersection one should turn left or right in order to reach the (red) dot on the other street. We call the angle between the 6 o'clock (or upwards-facing) direction and the imagined orientation the *deviation angle*. (A) An example stimulus with deviation angle of  $-90$  deg. (B) An example stimulus with deviation angle of  $+135$  deg.

Participants were seated at about 60 cm from a computer display on which the stimulus was displayed. They used the left and right shift keys on a keyboard to answer respectively “left” and “right” with the corresponding hand. Each trial began with the display of a central fixation red cross. After 0.5 seconds, the stimulus was displayed until participant’s answer. We recorded both the response and the response time (RT). The experimental session lasted for about 40 minutes and included 10 practice trials and 500 experimental trials. Every one hundred trials were followed by a pause; its duration did not exceed 5 minutes.

Participants’ head and shoulder movements were recorded using a CODA cx1 scan unit of a Codamotion optical motion tracking system (Charnwood Dynamics Ltd., UK); we used three sensors for each body part. The system recorded the spatial coordinates of each of the six sensors at a sampling rate of 200 Hz.

We used a within-participant factorial design. The main independent variable, the deviation angle, had 8 levels ( $0^\circ$ ,  $\pm 45^\circ$ ,  $\pm 90^\circ$ ,  $\pm 135^\circ$ ,  $180^\circ$ ). The second independent variable, the corner angle, had 10 levels ( $\pm 30^\circ$ ,  $\pm 60^\circ$ ,  $\pm 90^\circ$ ,  $\pm 120^\circ$ ,  $\pm 150^\circ$ ). Five repetitions were set for each condition except for deviation angle angles of  $0^\circ$  and  $180^\circ$ , for which 10 repetitions were set. Trials were presented in random order.

A trial was considered valid if all sensor values were available for at least half of its duration and only subjects with at least 50% of valid trials were included in the motion analysis. Only data from correctly answered valid trials with a RT that did not exceed the mean RT plus 3 SD were included in the analyses. A rectangular moving average filter of 20 samples (0.1 s) was applied in order to smooth the motion data. The

trials with a deviation angle of  $180$  deg were excluded from the analysis of the geometrical properties of rotations and translations as the sign of the angle cannot be used to discriminate the direction of rotations or translations.

We used the distance travelled by a body part (by summing the absolute Euclidean distances between all successive samples of a sensor) as a first measure of motion. If participants did not move more for higher values of deviation angles, our hypothesis would be invalidated from the start. We selected the maximum path length among the three sensors for each body part as the representative value of its motion extent. Since the path length is always positive, we posited a simple regression model of the path length on the absolute values of deviation angle:  $P_i = a + |\theta_i| b$ , where  $P_i$  is the maximum path length of the three motion sensors on trial  $i$  (expressed in mm),  $\theta_i$  the deviation angle on that trial,  $b$  a regression coefficient, and  $a$  a constant term. If the slope is found to be positive, we can proceed to a more specific analysis, which consists in decomposing the motion in its translational and rotational components and analyzing their geometrical specificity in relation to the signed values of deviation angles.

For the sake of the detailed motion analysis, we assume that both the head and shoulders undergo rigid motion in space—a combination of rotation and translation. We extracted the rotation and translation, using an optimization algorithm. We first calculated the relative vectors between the three sensors, which isolates the rotation component of the rigid motion. Our algorithm then searched through the (three-dimensional) space of rotations, finding the rotation that most closely matched the final relative vectors. We calculated the translation separately by performing vector subtraction between centers of mass of the three sensors for the head and shoulders.

For each sample of sensor positions provided by the motion tracker, we computed participants’ head and shoulder rotations (axes and angles) and translations with respect to their initial orientation and position, respectively. We then selected the maximum values of rotation and translation reached during the trial. We could not predict the axis about which the spontaneous rotations take place. We therefore posited the following simple linear regression model, in terms of axis-angle rotation vectors (indicated in boldface) for the relation between spontaneous movements and task variables:  $\mathbf{R}_i = \theta_i \mathbf{r}$ , where  $\mathbf{R}_i$  is the maximum rotation—of either the head or the shoulders—on trial  $i$  (expressed in the axis-angle vector representation),  $\theta_i$  the deviation angle on that trial, and  $\mathbf{r}$  a triplet of regression coefficients. Thus, the vector  $\mathbf{r}$  represents the rotation (again, as a vector in axis-angle space) that the participant would perform for deviation angle  $\theta$  equal to 1 deg. We decomposed this vector into its axis-angle components:  $\mathbf{r} = z \hat{\mathbf{a}}$ , where its length or norm,  $z$ , is a regression coefficient that we will call the *spontaneous rotation coefficient*, and its direction,  $\hat{\mathbf{a}}$ , the unit vector that corresponds to the axis of rotation. Our regression therefore yields both the

spontaneous rotation coefficient and the axis of spontaneous rotation.

We posited a similar model for translations:  $\mathbf{T}_i = \theta_i w \hat{\mathbf{u}}$ , where  $\mathbf{T}_i$  is the maximum translation vector of the head or shoulders, and  $\hat{\mathbf{u}}$  is a unit vector indicating the direction of translation, and  $w$  the spontaneous translation coefficient.

To calculate statistical confidence intervals of these spontaneous motion coefficients, we performed a bootstrap. For each bootstrap resample  $j$ , we calculated the rotation vector  $\mathbf{r}^{(j)}$  [or the translation  $w^{(j)} \mathbf{t}^{(j)}$ ]. We then calculated a 95% confidence ellipsoid for these points. If the origin fell outside this ellipsoid, then the regression was said to yield a coefficient statistically different from zero. We used the geometric mean of the ellipsoid semi-axes as a measure of standard error of the spontaneous motion coefficients.

## Results

### Response times and error rates

Overall, the mean RT on raw unfiltered data was  $1.17 \pm 0.38$  s ( $\pm$  between-subject SD). Increasing the deviation angle lowers performance, increasing the RT, as shown in Fig. 2.

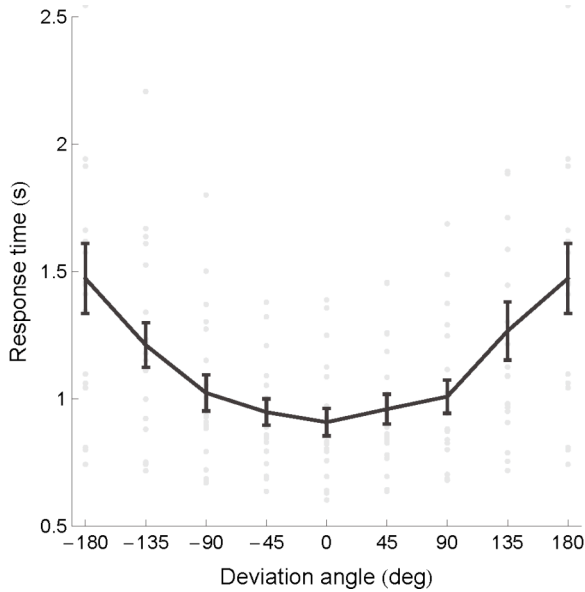


Figure 2: RT as a function of the deviation angle. Gray dots represent individual participants' data, the black curve the mean, and error bars between-subject standard errors. The data for deviation angles  $\pm 180^\circ$  is shown twice. Several outlying datapoints are not shown.

The mean reaction time was submitted to a repeated-measures ANOVA with the factors: sign of deviation angle (2 levels), absolute value of deviation angle (6 levels, excluding 0 and 180 deg), sign of corner angle (2 levels) and absolute value of corner angle (5 levels). The analysis revealed a significant main effect of absolute value of the deviation angle ( $F_{2,36} = 29.1$ ,  $p < 0.001$ , Huynh-Feldt corrected), a significant main effect

of the sign of corner angle ( $F_{1,18} = 16.9$ ,  $p < 0.001$ ), a significant main effect of the absolute value of corner angle ( $F_{3,4,60,6} = 8.3$ ,  $p < 0.001$ ) and a significant third-order interaction between the sign of deviation angle, the sign of the corner angle and the absolute value of corner angle ( $F_{2,8,50,9} = 4.3$ ,  $p < 0.01$ ). The main effect of the sign of the deviation angle was not statistically significant nor were the other interactions.

To quantify the relation between the deviation angle and the RT, we calculated the slopes of the linear regression of the RT on the absolute values of the deviation angle for every participant. (Since the sign of the deviation angle had no effect on the response times, we collapsed data for positive and negative deviation angles.) All individual slopes were positive and statistically significant (bootstrap with  $10^4$  resamples,  $p < 0.05$ ); the mean slope was  $3.09 \pm 2.30$  ms/deg ( $\pm$  between-subject SD). In other words, mean RT increased by 3.09 ms for every additional degree of deviation angle. The plot of RT versus deviation angle (Fig. 2) has a noticeably curvilinear shape, with the RT slope seemingly higher for deviation angles above 90 deg. The mean slope for deviation angles between 0 and 90 deg was  $1.20 \pm 0.89$  ms/deg, whereas between 90 and 180 degrees it was  $5.05 \pm 3.82$  ms/deg. This difference between slopes for small and large deviation angles was statistically significant (paired  $t_{18} = 5.41$ ,  $p < 0.0001$ ) and showed that RTs increased faster (more than 4 times faster, according to the means) as a function of deviation angle above 90 deg.

The median error rate was  $1.2 \pm 0.6\%$  ( $\pm$  between-subject median absolute deviations). Overall, the error rate was very low: the task was seemingly well understood by our participants and easy to perform. The analyses of the relation between error rates and deviation angles lead to similar findings as the ones of the RT and are not provided here.

### Spontaneous body movements and their relation to the task

**Analysis of Path Length** As a first analysis of the relation between task performance and body movements, we wanted to see if there was a relationship between the extent of spontaneous motion and the deviation angle. As a measure of motion extent, we used the length of the path traveled in space. Fig. 3 shows the mean path lengths as a function of the absolute deviation angle.

The mean path length across participants and deviation angles is  $13.1 \pm 10.2$  mm ( $\pm$  between-subject SD) for the head and  $10.3 \pm 6.9$  mm for the shoulders. The slope of the linear regression (including a constant term, see Methods) of path lengths on the absolute deviation angles provides an indication on the relation between the movements and the deviation angle: if positive, it would indicate that the participants move more in trials with higher deviation angles. For head movements, 17/19 (89%) regression slopes were positive and 13/19 (68%) were significantly so (bootstrap with  $10^4$  resamples,  $p < 0.05$ ); the mean slope

was  $0.038 \pm 0.058$  mm/deg ( $\pm$  between-subject SD). For the shoulders, 17/19 (89%) regression slopes were positive and 12/19 (63%) were significantly so (bootstrap with  $10^4$  resamples,  $p < 0.05$ ); the mean slope was  $0.026 \pm 0.041$  mm/deg.

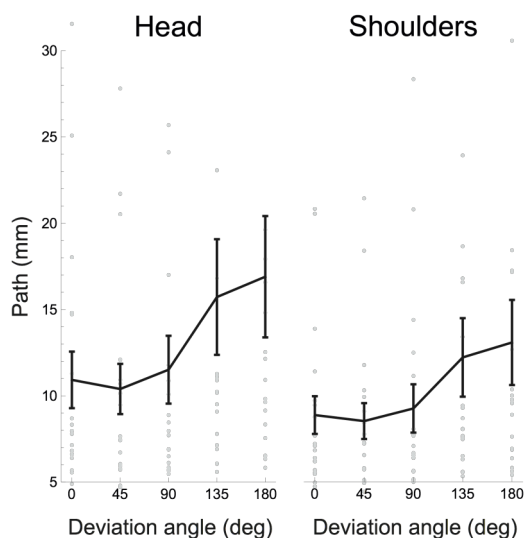


Figure 3: Mean path length traveled by the head and shoulders as a function of deviation angle. Gray dots represent individual participants' data, the black curve the mean, and error bars between-subject standard errors. Several outlying datapoints are not shown.

This analysis of path lengths shows that for most participants there was a relationship between the *absolute extent* of spontaneous movements and the absolute value of the principal task parameter, the deviation angle. Because the movements of the head and shoulders were nearly rigid, for further analysis we decomposed them into the two components of rigid motion, rotations and translations.

**Analysis of Absolute Amplitude of Rotations** As stated in the Methods, for each trial we calculated the maximal rotation of the head and shoulders with respect to their initial orientations at the start of the trial. We represented these rotations as 3D vectors using the axis-angle representation, in which the length of the vector is the angle of rotation and its direction the axis.

To begin with, we analyzed only the angles of the maximal rotations. As in the preceding analysis, we wished to test whether this measure of absolute magnitude of rotation was correlated with task difficulty, i.e., the absolute value of deviation angle. Fig. 4 shows the mean maximal rotation magnitude as a function of the absolute deviation angle.

The overall mean rotation amplitude is  $1.57 \pm 0.5$  deg ( $\pm$  between-subject SD) for the head and  $0.78 \pm 0.17$  deg for the shoulders. Some of the spontaneous rotations were not specifically related to the main task parameter, as shown by the presence of rotations even when deviation angle is zero.

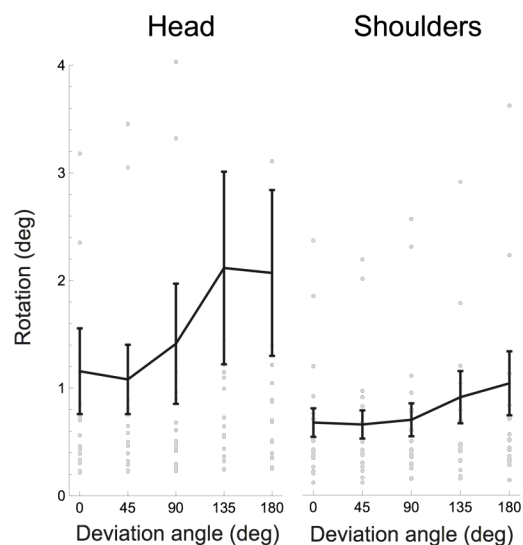


Figure 4: Absolute rotation amplitude, for the head and shoulders, as a function of the absolute deviation angle. Gray dots represent individual participants' data, the black curve the mean, and error bars between-subject standard errors. Several outlying datapoints are not shown.

We performed a linear regression (including a constant term) of the rotation amplitude versus absolute deviation angles to quantify the relation between the rotations and the deviation angles. We found out that 17/19 (89%) regression slopes were positive for both head and shoulder rotations and 10/19 (53%) for the head and 7/19 (37%) for the shoulders were significantly so (bootstrap with  $10^4$  resamples,  $p < 0.05$ ); the mean slope was  $0.006 \pm 0.013$  ( $\pm$  between-subject standard deviations) for the head and  $0.002 \pm 0.005$  for the shoulders.

The analysis of absolute rotation angles shows that there was a relationship between the *absolute amplitudes* of spontaneous rotations and the deviation angle. It doesn't tell us, however, if this relationship was geometrically specific. Did participants spontaneously move in one direction for the positive deviation angles and in the opposite direction for the negative ones?

**Directional Analysis of Rotations** To answer the question above, we performed a linear regression of the full axis-angle rotation vectors (i.e., including the direction of rotations in addition to their amplitudes) on the deviation angle—rather than just its absolute value—of the corresponding trial. We call the slopes of these linear regressions the spontaneous rotation coefficients (see Methods for details).

For head rotations, the mean spontaneous rotation coefficient is  $0.007 \pm 0.018$  ( $\pm$  between-subject SD); the median coefficient is  $0.001 \pm 0.0009$  ( $\pm$  between-subject median absolute deviations). For the shoulders, the mean coefficient is  $0.001 \pm 0.002$ ; the median coefficient is  $0.0005 \pm 0.0002$ . The interpretation of these parameters, for example for head spontaneous



rotations, is as follows: on the average, participants rotated their head by 0.7% (or 0.1%, if we use the medians) of the deviation angle. Contrary to the preceding analyses, we have extracted the directionally-specific component of the spontaneous rotations: rotations that are in opposite directions for clockwise and counterclockwise deviation angles. The axis of these rotations varies from one participant to the next; we will return to the question of axes below.

To test whether these correlations were statistically significant, we stepped back to our original regression model,  $\mathbf{R}_i = \theta_i \mathbf{r}$  (recall that the spontaneous rotation coefficients are the lengths of the regression vectors  $\mathbf{r}$ ), and used the regression vectors  $\mathbf{r}$  for significance analysis. We performed a bootstrap resampling ( $10^5$  resamples) of the vectors  $\mathbf{r}$  and calculated the 95% confidence ellipsoid of these vectors (see Methods).

First of all, an omnibus regression analysis, including all data sets of all participants at once, shows a statistically significant spontaneous rotation coefficient of 0.0068 for the head and of 0.0006 for the shoulders. Second, although the individual spontaneous rotation coefficients were small (all but two were smaller than 1%), in case of head rotations 15/19 (79%) participants had a statistically significant linear relationship between maximum rotation and deviation angle. In case of shoulder rotations, on the other hand, only 4/19 (21%) participants had significant fits to the model. Given that only a few participants executed significant shoulder rotations, we carried out the rest of rotation analyses only for head movements.

**Analysis of Rotation Axes** Along with the spontaneous rotation coefficients, our analysis also yielded an axis of rotation for each subject, separately for the head and the shoulders. Fig. 5 shows these axes, as unit vectors (the vector  $\hat{\mathbf{a}}$  in our regression model), for the head rotations of the 15 participants whose regression analyses yielded significant results. The meaning of each of these vectors is as follows: it is the axis that maximizes the correlation between a participant's rotations and the corresponding values of the deviation angle.

Fig. 5 also shows the mean head rotation axis over all of these participants, equal to  $(+0.13, -0.65, +0.75)$ . The axes of the fifteen participants are rather tightly clustered around this mean; the mean difference between the individual axes and the mean axis is only 24 deg. The largest contributions to this mean rotation axis come from the Z and Y axes. The signs of the components in this vector mean that for *positive* values of the deviation angle, participants tended to carry out rotations about the *positive* Z axis (head turned to the left, as seen from above, see Fig. 6 B) and the *negative* Y axis (head inclined to the left, as seen from behind, see Fig. 6 D); for trials in which the deviation angle was negative, on the other hand, the rotations tended to be in the opposite direction. We will return to the significance of these axes of rotation in the Discussion.

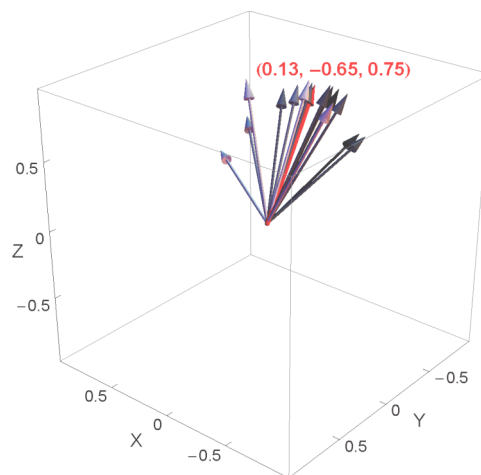


Figure 5: Individual head rotation axes represented in the space of rotations, shown in gray, as well as their mean, shown in red. The individual axes are shown in shades of gray that correspond to the value of the spontaneous rotation coefficient (the darker the arrow, the higher the corresponding coefficient). The three euclidean coordinates represent the mean axis. Given our regression model, the axes change to opposite directions for negative deviation angles.

**Analysis of Translations** The statistical analysis here is the same as for rotations. The mean spontaneous translation coefficients for the head and shoulders are respectively  $0.012 \pm 0.017$  ( $\pm$  between-subject SD) and  $0.008 \pm 0.016$ . The translations of 7 (37%) and 6 (32%) out of 19 participants, for respectively the head and shoulders, were significantly correlated to deviation angle (bootstrap with  $10^4$  resampled datasets,  $p < 0.05$ ).

## Discussion

When asked for directions some people execute spontaneous incipient body movements. If a geometrical relation were found between represented spatial self-displacements and co-occurring incipient spontaneous body movements, it would be indicative of the implication of motor processes (motor plans or efference copies) in our spatial task and consistent with the activation of a sensorimotor prediction mechanism in solving spatial updating problems. Based on findings of the studies on spatial perspective-taking, we focused on the study of the imagined rotations and the angular disparity effect.

We devised a spatial updating task (see Methods and Fig. 1). In addition to behavioral data, we measured the spontaneous movements of our participants. To our knowledge, spontaneous movements have not been quantified so far in a spatial updating task.

Our behavioral results replicate the studied angular disparity effect on task performance (see Introduction).

We found that 15 out of 19 (79%) participants executed spontaneous head rotations related to the task parameters (if we include translations, 17 out of 19 participants (89%) executed a statistically significant motion)—in spite of the ease of the task, as shown by low error rates. These rotations were very small in



amplitude (typically below 2 deg). In most of the participants, the movements were too small to be seen, but could nevertheless be measured with the motion tracker, and their relationship to the task parameters shown using our analysis. Indeed, these miniature head rotations were reliably correlated to the deviation angle, but much smaller (typically less than 1% of the deviation angle).

The geometrically specific correlation between spontaneous head rotations and the deviation angle has two aspects. First, larger deviation angles (corresponding to more difficult trials) led to larger rotations. Second, opposite deviation angles led to head rotations in opposite directions about a specific rotation axis, that we calculated using our linear model in rotation space.

The mean axis of rotation, averaged across participants, has a main vertical Z-axis component (i.e. a head turn, see Fig. 6 A, B) and a strong but lesser front-back Y-axis component (Fig. 6 C, D): the resulting head movement is thus a horizontal rotation of the head with an important tilt component. These head rotations are as if participants were trying to align themselves with the imagined orientation on the map. In the case of the front-back Y-axis, this alignment is in the image plane; in the case of the vertical Z-axis, it is as if the participants back-projected the vertical image onto the ground plane, and then tried to align themselves with the imagined orientation in this projection.

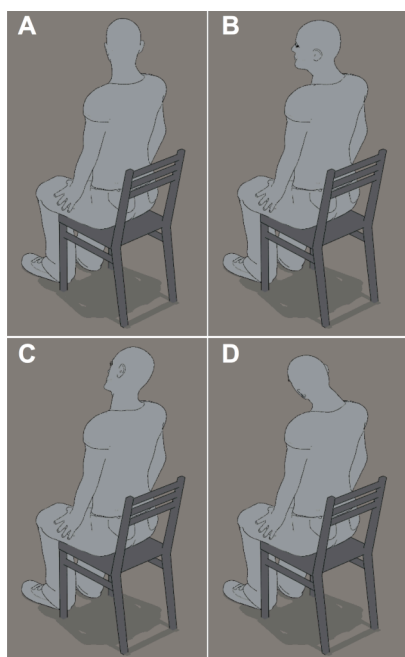


Fig. 6: The main component axes of average spontaneous head rotations. (A, B) Head turn about vertical Z axis. (C, D) Head tilt about naso-occipital Y axis.

Our findings on spontaneous head rotations are consistent with a motor contribution to spatial-updating task performance and with our action inhibition

hypothesis as the characteristics of the spontaneous movements are geometrically consistent with those of actual rotations in the ground plane or image plane that would be required to bring the participant into alignment with the required initial orientation. We may speculate on several types of contribution. The premotor cortex could prepare an actual movement, which would lead to two separate processes: an anticipation process that predicts the outcome of the action (i.e., the map with the you-are-here street aligned with the participant's vertical axis) from an efference copy of the motor command, and the execution of the overt motor action, which would be inhibited at early stages (earlier for some participants than for others).

Alternatively, the implication of the motor system may be *epiphenomenal*, related to concurrent cognitive processes but not causally so.

We cannot at this stage answer the question of causality of the spontaneous movements. To settle this argument, we need a new experimental setting contrasting a condition in which movement is allowed or facilitated with another one where movement is restrained. It will allow us to measure the impact of each condition on the task performance and shed more light on the causality of the motor processes in mental spatial updating tasks.

## Acknowledgements

We are grateful to Thales Group for support of part of this research.

## References

- Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of experimental psychology: General*, 140(1), 102-16.
- Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3(11), 419-429.
- Kosslyn, S M, Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews. Neuroscience*, 2(9), 635-642.
- Wang, R. F., Crowell, J. a, Simons, D. J., Irwin, D. E., Kramer, A. F., Ambinder, M. S., Thomas, L. E., et al. (2006). Spatial updating relies on an egocentric representation of space: effects of the number of objects. *Psychonomic Bulletin & Review*, 13(2), 281-286.
- Wexler, M., Kosslyn, S. M., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition*, 68(1), 77-94.
- Wohlschläger, Andreas, & Wohlschläger, Astrid. (1998). Mental and manual rotation. *Journal of Experimental Psychology: Human Perception and Performance*, 24(2), 397-412.
- Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Current Biology*, 11(18), 729-732.

# Auditory Saliency Using Natural Statistics

Tomoki Tsuchida (ttsuchida@ucsd.edu)

Garrison W. Cottrell (gary@ucsd.edu)

Department of Computer Science and Engineering  
9500 Gilman Drive, Mail Code 0404  
La Jolla CA 92093-0404 USA

## Abstract

In contrast to the wealth of saliency models in the vision literature, there is a relative paucity of models exploring auditory saliency. In this work, we integrate the approaches of (Kayser, Petkov, Lippert, & Logothetis, 2005) and (Zhang, Tong, Marks, Shan, & Cottrell, 2008) and propose a model of auditory saliency. The model combines the statistics of natural soundscapes and the recent past of the input signal to predict the saliency of an auditory stimulus in the frequency domain. To evaluate the model output, a simple behavioral experiment was performed. Results show the auditory saliency maps calculated by the model to be in excellent accord with human judgments of saliency.

**Keywords:** attention; saliency map; audition; auditory perception; environmental statistics

## Introduction

In general, attention plays a very important role in the survival of an organism, by separating behaviorally relevant signals from irrelevant ones. One approach to understanding how attention functions in the brain is to consider the “saliency map” over the sensory input space, which may determine subsequent motor control targets or selectively modulate perceptual contrast thresholds. The brain’s putative saliency maps can be thought of as interest operators that organisms use to enhance or filter sensory signals.

Many visual saliency models have been investigated, but relatively little attention has been paid to modeling auditory saliency. However, since the fundamental necessity for perceptual modulation remains the same regardless of modality, the principles of visual saliency models should apply equally well to auditory saliency with appropriate sensory input features. Two representative visual saliency models are the center-surround contrast model (Itti, Koch, & Niebur, 2002) and the SUN (Saliency Using Natural Statistics) model (Zhang et al., 2008). Itti et al.’s model is neurally-inspired, with the response of many feature maps (e.g., orientation, motion, color) combined to create a salience map. The SUN model uses a single feature map learned using Independent Components Analysis (ICA) of natural images, and the salience at any point is based on the rarity of the feature responses at that point - novelty attracts attention. Here, rarity is based on statistics taken from natural images, so the model assumes experience is necessary to represent novelty.

Previous works that apply the visual saliency paradigm to the auditory domain include the models of (Kayser et al., 2005) and (Kalinli & Narayanan, 2007). Both adapt the visual saliency model of (Itti et al., 2002) to the auditory domain

by using spectrographic images as inputs to the model. Although this is a reasonable approach, these models fail to capture several important aspects of the auditory modality. First, this approach treats time as simply another dimension within the spectrographic representation of the sound. Even though these models utilize asymmetric temporal filters, the resulting saliency map at each time point is contaminated by information from the future. Second, spectrographic features are not the most realistic representations of human auditory sensations, since the cochlea exhibits complex nonlinear responses to sound signals (Lyon, Katsiamis, & Drakakis, 2010). Finally, Itti et al.’s model determines the saliency values from the current input signal, with no contribution from the lifetime experience of the organism. This makes it impossible for the model to account for potential perceptual differences induced by differences in individual experience.

## The Auditory Saliency Model

In this work, we propose the Auditory Saliency Using Natural statistics model (ASUN) as an extension of the SUN model. The extension involves (1) using realistic auditory features instead of visual ones, and (2) combining long-term statistics (as in SUN) with short-term, temporally local statistics. Although the SUN model has both a top-down, task-based component and a bottom-up, environmentally driven component, here we restrict ourselves to just the bottom-up portion of SUN. SUN defines the bottom-up saliency of point  $x$  in the image at time  $t$  as:

$$s_x(t) \propto -\log P(F_x = f_x) \quad (1)$$

Here,  $f$  is a vector of feature values, whose probability is computed based on prior experience. This is also known as the “self-information” of the features, and conveys that rare feature values will attract attention. In the SUN model, this probability is based on the lifetime experience of the organism, meaning that the organism already knows when feature values are common and when they are rare. Assuming the primary purpose of attention is to separate remarkable events from the humdrum, it is logical to equate the rarity of the event with the saliency of it. For example, a loud bang may be salient not only because of its physical energy content, but also because of its relative rarity in the soundscape. An organism living under constant noise may not find an explosion to be as salient as another organism acclimated to a quieter environment.

For features, SUN uses ICA features learned from natural images, following Barlow’s efficient coding hypothesis (Barlow, 1961). This provides a normative and principled rationale for the model design. While ICA features are not completely independent, they justify the assumption that the features are independent of one another, making the computation of the joint probability of the features at a point computationally simple. This is the goal of efficient coding: By extracting independent features, the statistics of the visual world can be efficiently represented. Although the saliency filters used in Kayser et al.’s model have biophysical underpinnings, exact shape parameters of the filters cannot be determined in a principled manner. More importantly, their model does not explain *why* the attention filters should be the way they are. In contrast, by using filters based on the efficient coding hypothesis, the SUN and ASUN models make no such assumptions; the basic feature transformation used (Gammatone filters) reasonably approximate the filters learned by the efficient encoding of natural sounds (Lewicki, 2002), and the distributions of filter responses are learned from the environment as well. Assuming that the attentional mechanism is modulated by a lifetime of auditory experience is neurologically plausible, as evidenced by the experience-induced plasticity in the auditory cortex (Jääskeläinen, Ahveninen, Bellevue, Raji, & Sams, 2007).

Here, we extend this model to quickly adapt to recent events by utilizing the statistics of the recent past of the signal (the “local statistics”) as well as the lifetime statistics. Denoting the feature responses of the signal at time  $t$  as  $F_t$ , saliency at  $t$  can be defined as the rarity in relation to the recent past (from the input signal) as well as to the long-term past beyond suitably chosen delay  $k$ :

$$s(t) \propto -\log P(F_t = f_t | \underbrace{F_{t-1}, \dots, F_{t-k}}_{\text{recent past}}, \underbrace{F_{t-k-1}, \dots}_{\text{long past}})$$

In this paper, we simply define  $t - k$  as the onset of the test stimulus. Under the simplifying assumption of independence between the lifetime and local statistics, this becomes

$$\begin{aligned} s(t) &\propto -\log P(F_t = f_t | F_{t-1}, \dots, F_{t-k}) \\ &\quad -\log P(F_t = f_t | F_{t-k-1}, \dots) \\ &= s_{\text{local}}(t) + s_{\text{lifetime}}(t), \end{aligned}$$

where  $s_{\text{local}}(t)$  and  $s_{\text{lifetime}}(t)$  are the saliency values calculated from the local and lifetime statistics, respectively. By using the local statistics at different timescales, the model can simulate various adaptation and memory effects as well. In particular, adaptation effects emerge as the direct consequence of dynamic information accrual, which effectively suppresses the saliency of repeated stimuli as time proceeds. With such local adaptation effects, the model behaves similarly to the Bayesian Surprise model (Baldi & Itti, 2006), but with asymptotic prior distributions provided by lifetime experience.

## Feature Transformations

A model of auditory attention necessarily relies upon a model of peripheral auditory processing. The simplest approach to modeling the cochlear transduction is to use the spectrogram of the sound, as was done in (Kayser et al., 2005). More physiologically plausible simulations of the cochlear processing require the use of more sophisticated transformations, such as Meddis’ inner hair cell model (Meddis, 1986). However, the realism of the model comes at a computational cost, and the complexity of the feature model must be balanced against the benefit. Given these considerations, the following feature transformations were applied to the audio signals in the ASUN model:

1. At the first stage, input audio signals (sampled at 16 kHz) are converted to cochleagrams by applying a 64-channel Gammatone filterbank (from 200 to 8000 Hz.) Response power of the filters are clipped to 50dB, smoothed by convolving with a Hanning window of 1 msec and downsampled to 1 kHz. This yields a 64-dimensional frequency decomposition of the input signal.
2. At the second stage, this representation is further divided into 20 frequency bands comprised of 7 dimensions each (with 4 overlapping dimensions,) and time-frequency patches are produced using a sliding window of 8 samples (effective temporal extent of 8 msec). This yields 20 bands of  $7 \times 8 = 56$ -dimensional representation of 8 msec patches.
3. Finally, for each of the four sound collections (described below), a separate Principal Components Analysis (PCA) is calculated for each of the 20 bands separately. Retaining 85% of the variance reduces the 56 dimensions to 2 or 3 for each band.

This set of transformations yield a relatively low-dimensional representation without sacrificing biological plausibility. The result of these transformations at each time point,  $f_t$ , provides input for subsequent processing. Figure 1 illustrates this feature transformation pipeline.

## Density Estimation Method

In order to calculate the self-information described in equation 1, the probability of feature occurrences  $P(F = f_t)$  must be estimated. Depending on the auditory experience of the organism, this probability distribution may vary. To assess the effect of different types of lifetime auditory experiences, 1200 seconds worth of sound samples were randomly drawn from each of the following audio collections to obtain empirical distributions:

1. “Environmental”: collection of environmental sounds, such as glass shattering, breaking twigs and rain sounds obtained from a variety of sources. This ensemble is expected to contain many short, impact-related sounds.

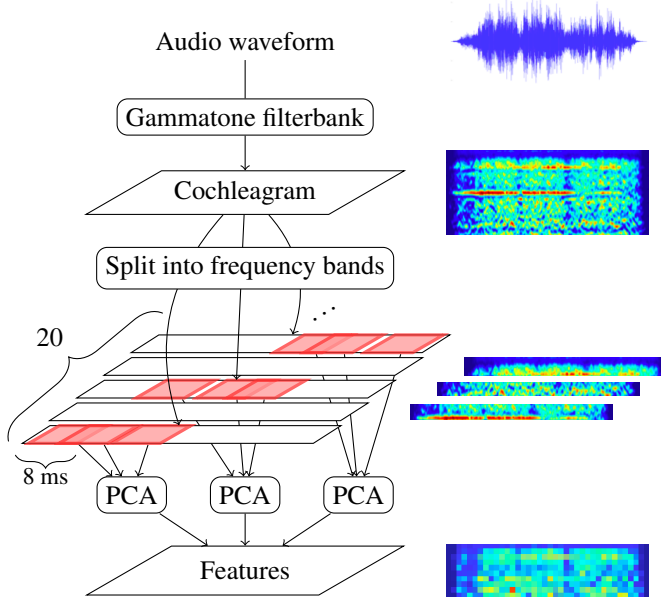


Figure 1: Schematics for the feature transformation pipeline. Input signals are first converted to smoothed cochleagram. This is separated into 20 bands of 8 msec patches. The dimensions of each band are reduced using PCA.

2. “Animal”: collection of animal vocalizations in tropical forests from (Emmons, Whitney, & Ross, 1997). Most of the vocalizations are relatively long and repetitious.
3. “Speech”: collection of spoken English sentences from the TIMIT corpus (Garofolo et al., 1993). This is similar to the animal vocalizations, but possibly with less tonal variety.
4. “Urban”: this is a collection of sounds recorded from a city (van den Berg, 2010), containing long segments of urban noises (such as vehicles and birds), with a limited amount of vocal sounds.

In the case of natural images, ICA filter responses follow the generalized Gaussian distribution (Zhang et al., 2008). However, the auditory feature responses from the sound collections did not resemble any parameterized distributions. Consequently, a Gaussian mixture model with 10 components was used to fit the empirical distributions for each band from each of the collections. Figure 2 shows examples of density model fits against empirical distributions. The distributions from each collection represent the lifetime statistics portion of ASUN model, and each corresponds to a model of saliency for an organism living under the influence of that particular auditory environment.

The local statistics of the input signal were estimated using the same method: at each time step  $t$  of the input signal, the probability distribution of the input signal from 0 to  $t - 1$  was estimated. For computational reasons, the re-estimation of the local statistics were computed every 250 msec. Unfortunately, this leads to a discontinuity in the local probability

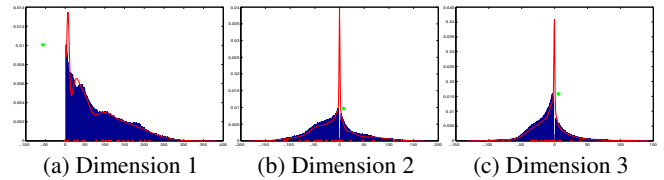


Figure 2: Gaussian mixture model fits (red) against the empirical distribution of feature values (blue). The mixture model is used to estimate  $P(F_t = f_t | F_{t-k-1}, \dots)$ .

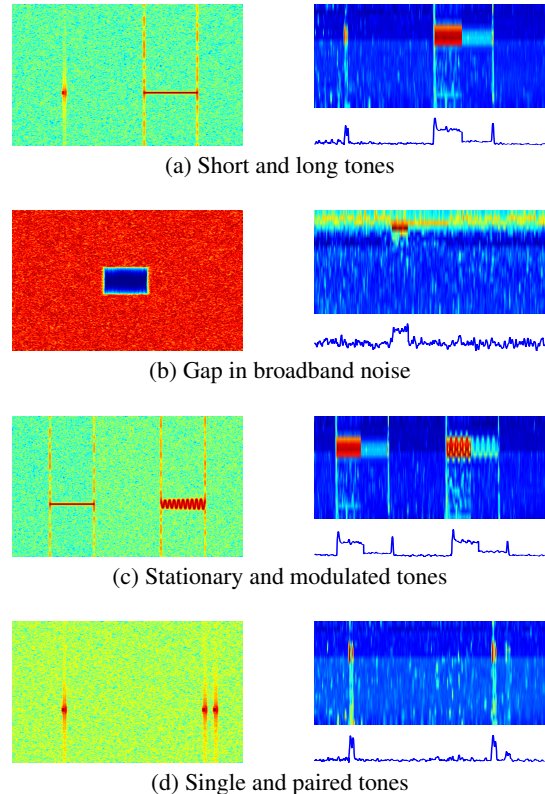


Figure 3: Spectrograms and saliency maps for simple stimuli. Left columns are the spectrograms of the stimuli, and right columns are the saliency maps (top) and saliency values summed over frequency axis (bottom). Due to the nonlinear cochleogram transform, the y-axes of the two plots are not aligned. (a) Between short and long tones, the long tone is more salient. (b) Silence in a broadband noise is salient compared to the surrounding noise. (c) Amplitude-modulated tones are slightly more salient than stationary tones. (d) In a sequence of closely spaced tones, the second tone is less salient.

distribution every 250 msec. This will be improved in future work, where we plan to apply continually varying mixture models to eliminate such transitions.

## Qualitative Assessments

In (Kayser et al., 2005), the auditory saliency model reproduces basic properties of auditory scene perception described

in (Cusack & Carlyon, 2003). Figure 3 shows the saliency maps of the ASUN model using the “Environmental” lifetime statistics. These examples demonstrate that the model is capable of reproducing basic auditory salience phenomena.

## Human Ratings of Saliency

In order to test the validity of the model in a more quantitative manner, a human rating experiment similar to (Kayser et al., 2005) was performed. In this experiment, seven subjects were asked to pick the more “interesting” of two stimuli. The goal of the experiment was to obtain an empirical rating of the “interestingness” of various audio stimuli, which we conjecture is monotonically related to the saliency. By presenting the same set of stimuli to the saliency models, we can also calculate which of the sounds are predicted to be salient. We assume that the correct model of saliency will have a high degree of correlation with the human ratings of saliency obtained this way.

## Materials

Audio snippets were created from a royalty-free sound collection (SoundEffectPack.com, 2011), which contains a variety of audio samples from artificial and natural scenes. In order to normalize the volume across samples, each sample was divided by the square root of the arithmetic mean of the squares of the waveform (RMS). To create snippets used in the experiment, each sample was divided into 1.2-second snippets, and the edges were smoothed by a Tukey window with 500 ms of tapering both sides. Snippets containing less than 10% of the power of a reference sinusoidal signal were removed in order to filter out silent snippets.

From this collection, 50 high-saliency, 50 low-saliency and 50 large-difference snippets were chosen for the experiments. The first two groups contained snippets for which the Kayser and ASUN models agreed on high (or low) saliency. Snippets in the last group were chosen by virtue of producing highest *disagreements* in the predicted saliency values between Kayser and ASUN models.

With these snippets, 75 trial pairs were constructed as follows:

- (1) *High saliency difference* trials (50): Each pair consists of one snippet from the high-saliency and another from the low-saliency groups.
- (2) *High model discrimination* trials (25): Both snippets were drawn from the large-difference group uniformly.

We expected both models to perform well on *high saliency difference* trials but to produce a performance disparity on the *high model discrimination* trials.

## Procedure

In each trial, each subject was presented with one second of white noise (loudness-adjusted using the same method as above) followed immediately by binaural presentation of a pair of target stimuli. The subject would then respond with

the left or right key to indicate which stimuli sounded “more interesting” (2AFC.) Each experiment block consisted of 160 such trials: 75 pairings balanced with left-right reversal, plus 10 catch trials in which a single stimulus was presented to one side. Each subject participated in a single block of the experiment within a single experimental session.

## Model Predictions

To obtain the model predictions, the same trial stimuli (including the preceding noise mask) were input to the models to produce saliency map outputs. To reduce border effects, 10% buffers were added to the beginning and end of the stimuli and removed after saliency map calculation. The portion of the saliency map that corresponded to the noise mask were also removed from peak calculations.

In (Kayser et al., 2005), saliency maps for each stimuli pair were converted to scores by comparing the peak saliency values. It is unclear what the best procedure is to extract a single salience score from a two-dimensional map of salience scores over time. Following (Kayser et al., 2005), we also chose the peak salience over the snippet. To make predictions, the score for the left stimulus was subtracted from that of the right stimulus in each trial pair. This yielded values between  $-1$  and  $1$ , which were then correlated against the actual choices subjects made ( $-1$  for the left and  $1$  for the right.)

Seven different candidate models were evaluated in this experiment. (1) The *chance* model outputs  $-1$  or  $1$  randomly. This model serves as the baseline against which to measure the chance performance of other models. (2) The *intensity* model outputs the Gammatone filter response intensity. This model simply reflects the distribution of intensity within the sound sample. (3) The *Kayser* model uses the saliency map described in (Kayser et al., 2005). Finally, *ASUN* models with different lifetime statistics were evaluated separately: (4) “Environmental” sounds, (5) “Animal” sounds, (6) “Speech” sounds, and (7) “Urban” sounds.

## Results

To quantify the correspondence between the model prediction and the human judgments of saliency, Pearson product-moment correlation coefficients (PMCC) were calculated between the model predictions and human rating judgment results ( $N=7$ ) across all 75 trials. All subjects responded correctly to the catch trials, demonstrating that they were paying attention to the task. Figure 4 shows the correlation coefficient values for the ASUN models for each type of dataset from which lifetime statistics were learned. The correlation between the ASUN model predictions and the human subjects ( $M = 0.3262, SD = 0.0635$ ) was higher than the correlation of the Kayser model predictions ( $M = 0.0362, SD = 0.0683$ ). The result shows that the ASUN model family predicted the human ratings of saliency better than the Kayser model ( $t(6) = 7.963, p < 0.01$ .)

To evaluate the model performance in context, across-subject correlation was also calculated. Since the models

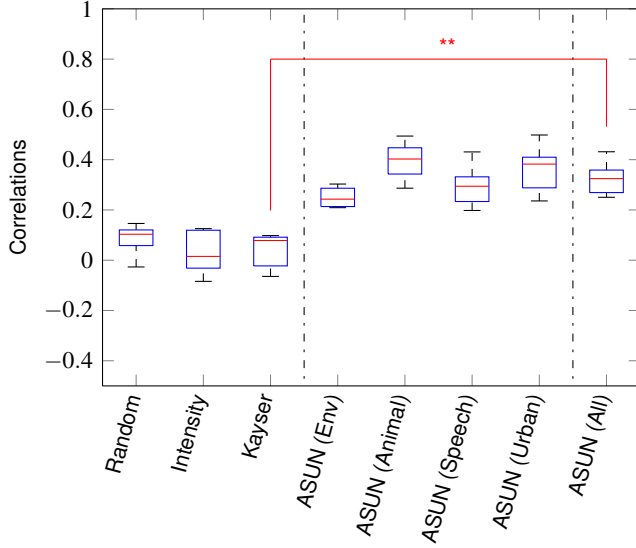


Figure 4: Correlation coefficient between various models and human ratings of saliency (N=7.) ASUN models correlated with the human ratings of saliency significantly better than the Kayser model.

are not fit to individual subjects, this value provides the ceiling for any model predictions. Because three of the seven subjects went through the same trial pairs in the same order, these trials were used to calculate the across-subject correlation value, and the model responses. Figure 5 shows the correlation values including the across-subject correlation. The result shows that the difference between the across-subject correlations ( $M = 0.6556, SD = 0.0544$ ) and the ASUN model predictions ( $M = 0.4831, SD = 0.0432$ ) was significant ( $t(2) = 16.9242, p = 0.0035$ ), indicating that the models do not yet predict saliency at the subject-consensus level. Nevertheless, the ASUN model correlations were still significantly higher than the Kayser model ( $M = 0.1951, SD = 0.0815$ ) at ( $t(2) = -9.855, p = 0.0101$ ).

The performance for the Kayser model in this experiment was notably worse than what was reported in (Kayser et al., 2005). There are several possible explanations for this. First, the audio samples presented in this experiment were roughly normalized for the perceived loudness. This implies that a saliency model that derives saliency values from the loudness measure in large part may not perform well in this experiment. Indeed, the intensity model does not predict the result above chance ( $t(6) = 0.66, p = 0.528$ ). Although the Kayser model does combine information other than the intensity image alone, it is possible that the predictive power of the model is produced largely by loudness information.

Second, as described previously, some of the trial pairs were chosen intentionally to produce maximal difference between the Kayser and ASUN models, and this produced the large performance disparity. Figure 6 support this hypothesis: in the *high saliency difference* trials, both models performed

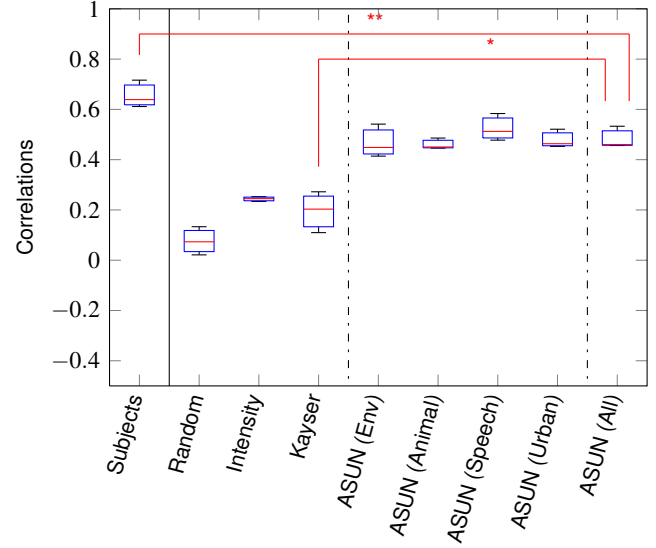


Figure 5: Correlation coefficient between various models and human ratings of saliency. A subset of data for which the same trial pairs were presented was analyzed (N=3). Across-subject performance was estimated using the correlation coefficients for all possible pairs from the three subjects.

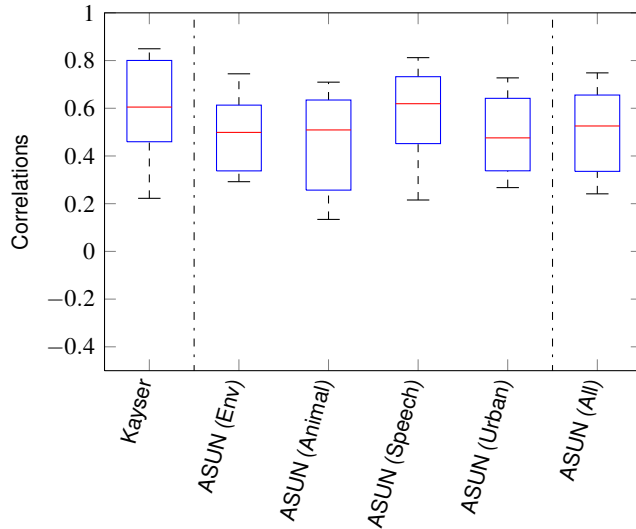
equally well ( $t(6) = 0.3763, p = 0.7091$ .) In contrast, in *high model discrimination* trials, ASUN models performed significantly better than the Kayser model ( $t(6) = 17.31, p < 0.01$ .) Note that the *high model discrimination* group was not picked based on the absolute value (or “confidence”) of the model predictions, but rather solely on the large difference between the two model predictions. This implies the procedure itself does not favor one model or the other, nor does it guarantee performance disparity on average. Nevertheless, the result shows that the ASUN models perform better than the Kayser model in those trials, suggesting the performance disparity may be explained in large part from those trials.

## Discussion

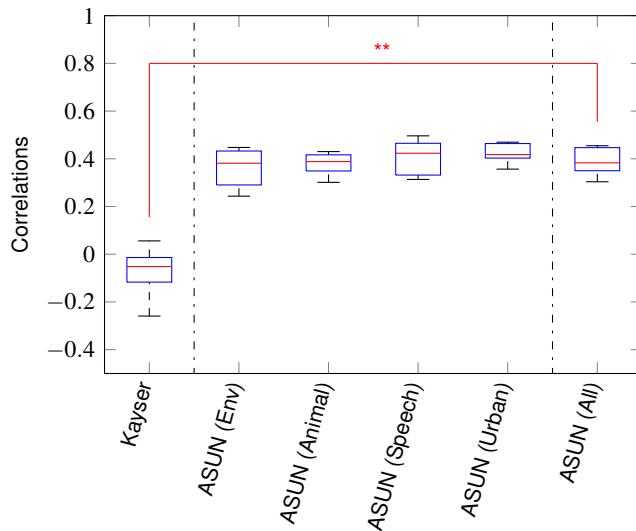
In this work, we demonstrated that a model of auditory saliency based on the lifetime statistics of natural sounds is feasible. For simple tone signals, auditory saliency maps calculated by the ASUN model qualitatively reproduce phenomena reported in the psychophysical literature. For more complicated audio signals, assessing the validity of the saliency map is difficult. However, we have shown that the relative magnitudes of the saliency map peaks correlate with human ratings of saliency. The result was robust across different training sound collections, which suggest a certain commonality in the statistical structure of naturally produced sounds.

There are aspects of the saliency model that may be improved to better model human physiology. For example, there is ample evidence of temporal integration at multiple timescales in human auditory processing (Poeppel, 2003). This indicates that the feature responses of the input signal





(a) High saliency difference trials



(b) High model discrimination trials

Figure 6: Correlation coefficients for the subsets of trials. (a) For *High saliency difference* trials, both Kayser and ASUN models show high correlation to human rating of saliency, and there are no significant differences between them. (b) For *High model discrimination* trials, ASUN models show significantly higher correlation with human ratings of saliency compared to the Kayser model.

may be better modeled by multiple parallel streams of inputs, each convolved with exponentially decaying kernels of varying timescales. This may be especially important for calculating saliency of longer signals, such as music and spoken phrases. In order to accommodate higher-level statistical structure, the model can be stacked in a hierarchical manner as well, with appropriate feature functions at each level. These expansions will provide insights into the nature of attentional modulations in human auditory processing.

## Acknowledgments

We thank Dr. Christoph Kayser for kindly providing us with the MATLAB implementation of his model. We also thank Cottrell lab members, especially Christopher Kanan, for insightful feedback. This work was supported in part by NSF grant #SBE-0542013 to the Temporal Dynamics of Learning Center.

## References

- Baldi, P., & Itti, L. (2006). Bayesian Surprise Attracts Human Attention. In *Nips 2005* (pp. 547–554).
- Barlow, H. B. (1961). Possible Principles Underlying the Transformations of Sensory Messages. *Sensory Communication*, 217–234.
- Cusack, R., & Carlyon, R. (2003). Perceptual asymmetries in audition. *J Exp Psychol Human Percept Perf*, 29(3), 713–725.
- Emmons, L. H., Whitney, B. M., & Ross, D. L. (1997). *Sounds of the neotropical rainforest mammals*. Audio CD.
- Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., et al. (1993). *Timit acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium, Philadelphia.
- Itti, L., Koch, C., & Niebur, E. (2002). A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 20(11), 1254–1259.
- Jääskeläinen, I. P., Ahveninen, J., Belliveau, J. W., Raij, T., & Sams, M. (2007). Short-term plasticity in auditory cognition. *Trends Neurosci*, 30(12), 653–661.
- Kalinli, O., & Narayanan, S. (2007). A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In *Interspeech 2007* (pp. 1941–1944). Antwerp, Belgium.
- Kayser, C., Petkov, C., Lippert, M., & Logothetis, N. (2005). Current biology; mechanisms for allocating auditory attention: An auditory saliency map. , 15(21), 1943–1947.
- Lewicki, M. S. (2002). Efficient coding of natural sounds. *nature neurosci*, 5(4), 356–363.
- Lyon, R. F., Katsiamis, A. G., & Drakakis, E. M. (2010). History and future of auditory filter models. In *Iscas* (pp. 3809–3812). IEEE.
- Meddis, R. (1986). Simulation of mechanical to neural transduction in the auditory receptor. *JASA*, 79(3), 702–711.
- Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication*, 41(1), 245–255.
- SoundEffectPack.com. (2011). *3000 sound effect pack*. Retrieved 2011-03-31, from [tinyurl.com/7f4z2wo](http://tinyurl.com/7f4z2wo)
- van den Berg, H. (2010). *Urban and nature sounds*. Retrieved 2011-02-27, from <http://tinyurl.com/89mr6dh>
- Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7), 1-20.



# A Multi-Measure Analysis of Context Effects in Multi-Attribute Decision Making: Examining the Similarity, Attraction, and Compromise Effects

**Takashi Tsuzuki** (tsuzuki@rikkyo.ac.jp)

Department of Psychology, Rikkyo University,  
1-2-26 Kitano, Niiza, Saitama 352-8558 Japan

**Jerome R. Busemeyer** (jbusemey@indiana.edu)

Department of Psychology and Brain Sciences, Indiana University  
1101 E. Tenth Street, Bloomington, IN 47405 USA

## Abstract

The similarity, attraction, and compromise effects warrant specific investigation in multi-attribute decision making. To examine these effects concurrently, we assigned 145 undergraduates to three context effect conditions. They were requested to solve 20 hypothetical purchase problems that had three alternatives described along two attribute dimensions. We measured their choices, confidence ratings, and response times. We found that adding the third alternative had significant effects for choice proportions and confidence ratings in all three conditions. The attraction effect was more prominent than the other two effects with regard to choice proportions. The compromise effect condition yielded low confidence ratings and long response times, although the choice proportion was high for the third alternative. These results indicate that the mutual relationship among choice proportions, confidence ratings, and reaction times requires theoretical investigation.

**Keywords:** decision making; choice; context effects; similarity effect; attraction effect; compromise effect

## Introduction

Theories of rational decision making suggest that choice is intrinsically determined by the utilities of the individual alternatives and thus unaffected by the relationships among the alternatives in the choice context. However, many studies have found violations of this tenet (Busemeyer, Barkan, & Chaturvedi, 2007; Tsetsos, Usher, & Chater, 2010). Three much-studied findings regarding such context-dependent choice effects warrant specific attention since they constitute violations of axioms fundamental to rational choice. The present paper collectively addresses these effects because they share important commonalities and can be explained using a unified framework. These findings include the attraction, similarity, and compromise effects.

These effects occur with the addition of a third alternative (decoy) to the two-alternative choice set (Roe, Busemeyer, & Townsend, 2001; Tsetsos et al., 2010; Tsuzuki & Guo, 2004; Usher & McClelland, 2004). Consistent with established research, the present paper examines these effects in a two-attribute form (see Figure 1). The alternatives that constitute the core two-alternative set are

commonly referred to as the target and the competitor. The target and the competitor form a trade-off—one is better than the other on one attribute, but worse on the other attribute. The third alternative is then added to this core set.

Depending on the relative position of the third alternative with respect to the target, three types of phenomena are likely to occur. Two arise when the third alternative is more similar to the target than it is to the competitor. However, if a trade-off exists between the third alternative and the target, the choice probability of the target decreases relative to that of the competitor. This is called the similarity effect (Brenner, Rottenstreich, & Sood, 1999; Tversky, 1972). In contrast, if the third alternative is inferior to the target on all attributes, the choice probability of the target should increase relative to that of the competitor. This is called the attraction effect (Hedgcock & Rao, 2009; Huber, Payne, & Puto, 1982).

The third phenomenon occurs when the third alternative rests between the target and the competitor, in which case the third alternative, now constituting a compromise between the core items, would be chosen most often. This is called the compromise effect (Mourali, Böckenholt, & Laroche, 2007; Simonson, 1989). All three of these phenomena constitute a violation of the axioms of rational choice.

Numerous explanations have been provided for each of the three kinds of decoy effects (Simonson & Tversky, 1992; Tversky, 1972; Tversky & Simonson, 1993). However, Roe et al. (2001) were the first to explain all three within a single framework that was implemented in a connectionist model derived from a previous stochastic mathematical theory (Busemeyer & Townsend, 1993; Tsuzuki, Kawahara, & Kusumi, 2002). Their model (the multi-alternative decision field theory, MDFT) accounts for these findings specifically with the aid of variable lateral inhibition, which is due to similarity relations among alternatives and the momentary shifting of attention from one attribute to another.

Tsetsos et al. (2010, p. 1280) remarked that “before we start, we note that these effects (the similarity, attraction, and compromise effects) were so far obtained in different studies, so until a study reports all three effects with the same

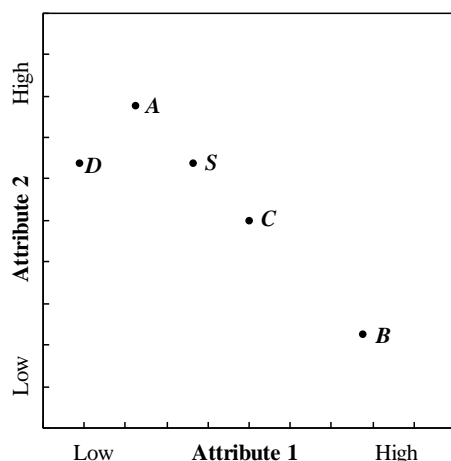


Figure 1: A summary of the phenomena simulated. The letters *S*, *D*, and *C* stand for the third alternatives for the similarity effect, attraction effect, and compromise effect, respectively.

materials, procedures, and subjects, there is the possibility that more freedom exists if parameters (noise) can be modified for various decoy effects.” Therefore, empirical studies are needed to test if the three effects can be replicated using the same experimental design and materials.

The present experiment focused on the functioning of all three major context effects in multi-attribute, multi-alternative decision-making processes using the same materials and procedures. This was done using valid choice sets (hypothetical purchase problems) based on preliminary research. In the three-choice session, participants made a selection, gave confidence ratings for the alternatives, and were measured for response times in 20 different choice sets.

## Method

### Participants and Design

One hundred and forty-five university undergraduates participated in this experiment, for which they received course credit. The basic design variables were (a) the type of the third alternative (corresponding to the similarity, attraction, or compromise effect), which was manipulated between subjects and (b) the type of the alternative (target, competitor, or a third alternative). The participants were randomly assigned to the between-subjects conditions. For the similarity, attraction, and compromise effect condition, 48, 49, and 48 undergraduates were assigned, respectively. The presentation of choice sets was quasi-randomized for each participant in each condition.

### Materials and Apparatus

Based on the stimuli of previous studies (Okuda, 2003; Pettibone & Wedell, 2000; Wedell & Pettibone, 1996), we conducted four preliminary surveys and subsequently

developed 20 choice sets (see Appendix). Each set contained alternatives from a single type of consumer product or service and consisted of two core alternatives (the target and the competitor) and a third alternative that was described on two dimensions (see Figure 1). Across the 20 choice sets, the average choice proportion for the target vs. the competitor was 51.32 vs. 48.27; these two proportions were not significantly different ( $n = 77$ ,  $\chi^2 = 0.16$ ,  $df = 1$ ).

For the similarity effect condition, the third alternative was created by lowering the value of the target on one of its dimensions by one-fourth of the difference between the target and the competitor and by raising the value of the other dimension of the target by one-fourth of the difference between the target and the competitor. For the attraction effect condition, the third alternative was created by lowering the values of the target on both its dimensions by one-fourth of the difference between the target and the competitor. Finally, for the compromise effect condition, the third alternative was created by lowering the value of the target on one of its dimensions by half of the difference between the target and the competitor and by raising the value of the other dimension of the target by half of the difference between the target and the competitor. All materials and instructions were presented using personal computers.



Figure 2: A screen image of the task used in the experiment with English translation (attraction effect condition).

## Procedure

**A session with three alternatives.** Participants were informed that they would be presented with many sets of three alternatives (the target, the competitor, and the third alternative), and that they would need to indicate their preference for each set. Each choice set was represented by three alternatives, each constructed using two values of differing dimensions (see Figure 2). The arrangement of the alternatives and dimensions on the screen was quasi-randomized in each trial. Choice sets were presented

on the screen and remained on the screen until the preference choice was made. The reaction time between the start of the presentation of the choice set and the choice response was measured using a personal computer. Following this, a confidence rating of the choice was provided based on a 9-point scale.

**A session with two alternatives.** Participants performed a similar experimental session using two alternatives (the target and the competitor).

## Results

### Binary Choice Session

In the binary choice session, the average choice proportions for the target vs. the competitor were 51.42 vs. 48.58 in the similarity effect condition, 48.61 vs. 51.39 in the attraction effect condition, and 51.42 vs. 48.58 in the compromise effect condition. The two choice proportions were not significantly different in any of the conditions ( $\chi^2 = 0.54$ ,  $df = 1$ ;  $\chi^2 = 0.53$ ,  $df = 1$ ;  $\chi^2 = 0.01$ ,  $df = 1$ , in the similarity, attraction, and compromise effect conditions, respectively). These results confirm the equivalence of the binary choice sets in this experiment as the baseline data.

### Choice Proportion in the Three-Choice Session

The arcsin transformed choice proportions were analyzed by two-way ANOVA (3 [type of context]  $\times$  3 [type of alternative]) with repeated measures (see Figure 3; Greer and Dunlap [1997] demonstrated that ANOVAs are applicable for the ipsative measures). Context type was a between-subjects factor and alternative type was a within-subjects factor.

The main effects of context type,  $F(2, 142) = 48.88$ ,  $p < .001$ , and alternative type,  $F(2, 284) = 55.45$ ,  $p < .001$ , and the interaction of the two factors,  $F(4, 284) = 80.33$ ,  $p < .001$  were significant. The simple main effects of alternative type were significant in the similarity, attraction, and compromise effect conditions ( $F(2, 426) = 5.74$ ,  $p < .01$ ;  $F(2, 426) = 129.20$ ,  $p < .001$ ;  $F(2, 426) = 105.68$ ,  $p < .001$ , respectively).

A multiple comparison (Tukey's *WSD* test) was performed on the three conditions of alternative type. In the similarity effect condition, the proportion of the competitor was significantly higher than that of the target and third alternative (both  $ps < .05$ ). In the attraction effect condition, the proportion of the target was significantly higher than those of the competitor and third alternative (both  $ps < .01$ ). Furthermore, the proportion of the competitor choice was significantly higher than that of the third alternative,  $p < .01$ . In the compromise effect condition, the proportion of the third alternative was significantly higher than that of the competitor,  $p < .05$ . Overall, these results indicate that the three kinds of decoy effects were replicated in the choice proportions for each of the three context effect conditions.

### Confidence Rating in the Three-Choice Session

The confidence rating scores were analyzed using a two-way mixed model ANOVA (3 [type of context]  $\times$  3 [type of alternative]) with repeated measures (see Figure 4). The main effects of context type,  $F(2, 153.44) = 3.65$ ,  $p < .05$ , and alternative type,  $F(2, 215.91) = 15.60$ ,  $p < .001$ , and the interaction of the two factors,  $F(4, 205.99) = 7.63$ ,  $p < .001$ , were significant. The simple main effects of alternative type were found to be significant in the similarity, attraction, and compromise effect conditions ( $F(2, 2.99) = 15.23$ ,  $p < .001$ ;  $F(2, 29.83) = 4.49$ ,  $p < .05$ ;  $F(2, 61.11) = 26.02$ ,  $p < .001$ ).

A multiple comparison was performed on the three conditions of alternative type. In the similarity effect condition, the confidence rating for the third alternative was significantly lower than those of the target and the competitor (both  $ps < .001$ ). In the attraction effect condition, the confidence rating of the target was significantly higher than that of the competitor,  $p < .05$ . In the compromise effect condition, the confidence rating of the third alternative was significantly lower than those of the target and competitor (both  $ps < .001$ ).

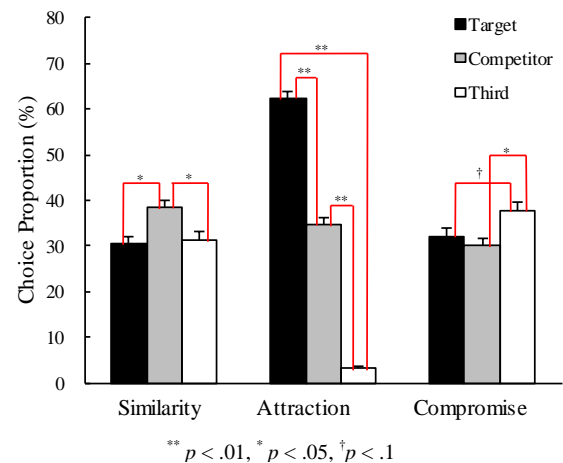


Figure 3: Mean choice proportions (%) of three alternatives in three context conditions. Error bars show standard errors.

In the similarity and attraction effect conditions, the confidence rating scores were largely consistent with the choice proportions. However, in the compromise effect condition, the confidence rating scores were reversed in magnitude relative to the choice proportions.

### Reaction Time in the Three-Choice Session

Choice latencies more than 2 *SD* above the mean for each subject were classified as errors and excluded from the RT analysis. The log-transformed choice latencies were analyzed in a one-way repeated-measures ANOVA (three-alternative types) in each of the three context conditions (see Figure 5).

Although the compromise effect condition yielded a

significant main effect,  $F(2, 93.87) = 4.58, p < .05$ , no such significant effects were found for the similarity or attraction conditions,  $F(2, 60.42) = 0.25$  and  $F(2, 73.88) = 2.17$ , respectively. For the compromise effect condition, a multiple comparison performed on the alternative type indicated that the decision time of the third alternative was significantly longer than those of the target and competitor, (both  $ps < .05$ ). These results are consistent with those of the confidence ratings.

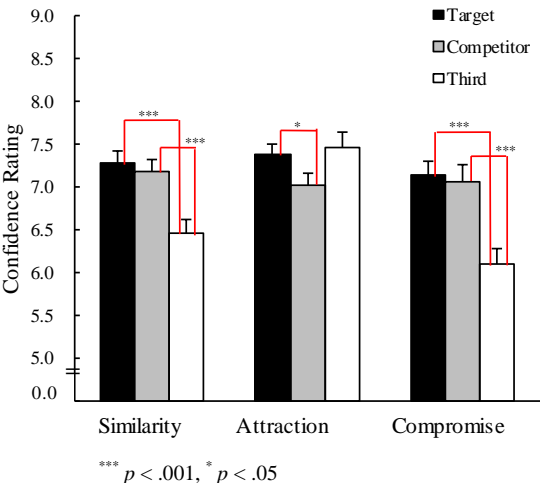


Figure 4: Mean confidence ratings of three alternatives in three context conditions. Error bars show standard errors.

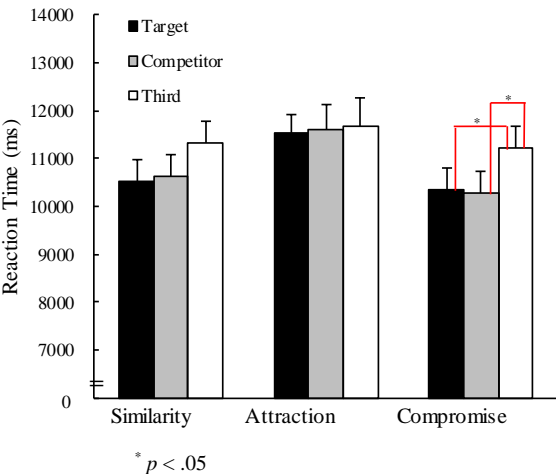


Figure 5: Mean response time (ms) of three alternatives in three context conditions. Error bars show standard errors.

### Discussion

Subsequent to the integrated account of three decoy effects by the MDFT model (Roe et al., 2001), Guo and Holyoak (2002) proposed a connectionist model that accounts for the attraction and similarity effects; this model

was also based on inter-alternative similarity. According to this model, the decision process is divided into two stages in which the two most similar alternatives (i.e., the target and the third alternative) are compared first, followed by the incorporation of the competitor.

Despite its explanatory simplicity and consistency with some established experimental data, the two-stage model appears to be oversimplified for the purpose of describing human behavior. Studies have demonstrated that in multi-alternative choice tasks similar to those of the similarity, attraction, and compromise effects mentioned above, (1) people momentarily shift their attention across pairwise comparisons and (2) similar pairs are compared more frequently than dissimilar pairs (Russo & Rosen, 1975; Satomura, Nakamura, & Sato, 1997).

Based on the data collected from these studies, Tsuzuki and Guo (2004) proposed a stochastic comparison-grouping model in which all possible types of comparisons are performed momentarily using differential frequencies (Figures 6, 7). In addition, while Guo and Holyoak's model uses a mathematical conversion to estimate choice probabilities from the results of only one simulation, Tsuzuki and Guo's model runs a large number of simulations in order to represent decisions across individuals, thereby directly estimating choice probabilities (Table 1).

In contrast to this research, Usher and McClelland (2004) offered an alternative to previous models that account for the three major context effects simultaneously. Their model, the leaky competing accumulator (LCA), shares many of the same principles of the MDFT model but makes different assumptions about loss aversion and the non-linear activation function (Busemeyer, Townsend, Diederich, & Barkan, 2005).

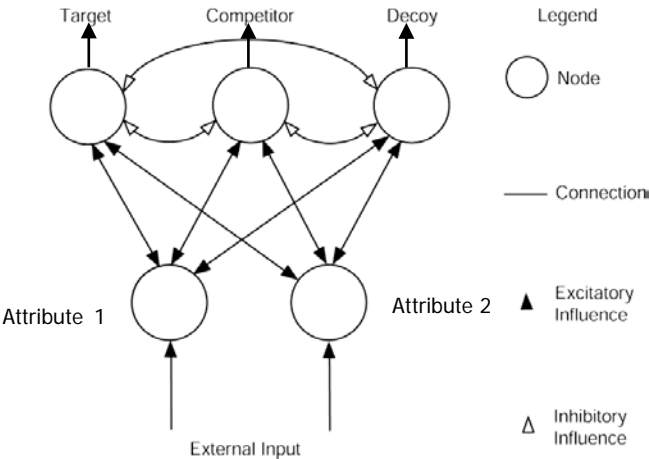


Figure 6: The architecture of the model (Tsuzuki & Guo, 2004). External Input represents the motivational and attentional sources that drive the decision process.

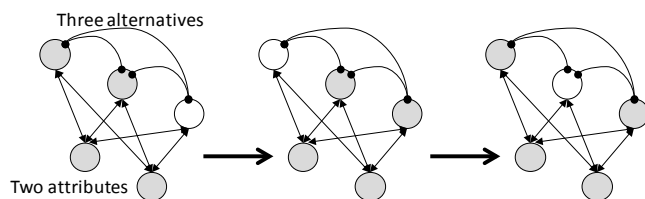


Figure 7: The time-series image of the dynamic fluctuation of the stochastic comparison (Tsuzuki & Guo, 2004). The dark color of the node reflects high node activation.

Table 1: Simulation results as choice probability (estimated from 10,000 simulations; Tsuzuki & Guo, 2004).

Choice scenarios	Choice probability		
	Target	Competitor	Decoy
Binary choice	0.504	0.496	-----
Attraction effect	0.587	0.366	0.048
Similarity effect	0.278	0.397	0.326
Compromise effect	0.213	0.219	0.568

In the present experiment, significant effects of manipulating the third alternative with respect to the similarity, attraction, and compromise effects were found for choice proportions, confidence ratings, and reaction times. Specifically, we found significant effects for choice proportions and confidence ratings in all three of these context-effect conditions, with partially significant effects in response time.

Furthermore, the attraction effect was more prominent than the other two effects with regard to choice proportions. The compromise effect condition yielded low confidence ratings and long response times, although the choice proportion of the third alternative was high. One possibility is that for participants in the compromise effect condition, one kind of selection effect happens for the participant confidence rating for the third alternative in the context of a trade-off or conflict with regard to the evaluation of both attributes. In order to further test this conjecture, we have begun experiments to study the role of eye movements in multi-attribute, multi-alternative processes (Tsuzuki, Shirai, Ohta, Matsui, & Honma, 2008).

Our experimental results support not only our stochastic comparison-grouping model but also the other major models of multi-attribute, multi-alternative choice processes. These results indicate that the relationship between choice proportions and confidence ratings requires theoretical investigation (Pleskac & Busemeyer, 2010), and also suggest that further examination of process-tracing data is needed to

determine the mechanisms underlying these three effects (Schulte-Mecklenbeck, Kühberger, & Ranyard, 2011; Willemsen, Böckenholt, & Johnson, 2011).

## Acknowledgments

We are grateful for comments from James L. McClelland. We are also appreciative of Hiroshi Matsui for his assistance in data collection.

## References

- Brenner, L., Rottenstreich, Y., & Sood, S. (1999). Comparison, grouping, and preference. *Psychological Science*, 10, 225–229.
- Busemeyer, J. R., Barkan, R., Mehta, S., & Chaturvedi, A. (2007). Context effects and models of preferential choice: Implications for consumer behavior. *Marketing Theory*, 7, 39–58.
- Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100, 432–459.
- Busemeyer, J. R., Townsend, J. T., Diederich, A., & Barkan, R. (2005). Contrast effects or loss aversion? Comment on Usher and McClelland (2004). *Psychological Review*, 112, 253–255.
- Greer, T., & Dunlap, W. P. (1997). Analysis of variance with ipsative measures. *Psychological Methods*, 2, 200–207.
- Guo, F. Y., & Holyoak, K. J. (2002). Understanding similarity in choice behavior: A connectionist model. *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 393–398). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hedgcock, W., & Rao, A. R. (2009). Trade-off aversion as an explanation for the attraction effect: A functional magnetic resonance imaging study. *Journal of Marketing Research*, 46, 1–13.
- Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, 9, 90–98.
- Mourali, M., Böckenholt, U., & Laroche, M. (2007). Compromise and attraction effects under prevention and promotion motivations. *Journal of Consumer Research*, 34, 234–247.
- Okuda, H. (2003). Context effects in decision making: Attraction, phantom, and plurality effects. *Japanese Journal of Social Psychology*, 18, 147–155.
- Pettibone, J. C., & Wedell, D. H. (2000). Examining models of nondominated decoy effects across judgment and choice. *Organizational Behavior and Human Decision Processes*, 81, 300–328.
- Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of confidence, choice, and response time. *Psychological Review*, 117, 864–901.

- Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, 108, 370–392.
- Russo, J. E., & Rosen, L. D. (1975). An eye fixation analysis of multialternative choice. *Memory & Cognition*, 3, 267–276.
- Satomura, T., Nakamura, H., & Sato, E. (1997). Consumers' attitude toward price (4): Experiments of reference price. *Distribution Information*, 5, 18–24.
- Schulte-Mecklenbeck, A., Kühberger, A., & Ranyard, R. (Eds.). (2011). *A handbook of process tracing methods for decision research: A critical review and user's guide*. New York: Taylor & Francis.
- Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects, *Journal of Consumer Research*, 16, 158–174.
- Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, 29, 281–295.
- Tsetos, K., Usher, M., & Chater, N. (2010). Preference reversal in multiattribute choice. *Psychological Review*, 117, 1275–1293.
- Tsuzuki, T., & Guo, F. Y. (2004). A stochastic comparison-grouping model of multialternative choice: Explaining decoy effects. *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society*. (pp. 1351–1356). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Tsuzuki, T., Kawahara, T., & Kusumi, T. (2002). Connectionist modeling of higher-level cognitive processes. *Japanese Journal of Psychology*, 72, 541–555.
- Tsuzuki, T., Shirai, T., Ohta, A., Matsui, H., & Honma, M. (2008). An eye-tracking analysis of context effects in multi-attribute, multi-alternative decision making: Examining the attraction effect and the compromise effect. *Abstract of the XXIX International Congress of Psychology, Berlin*, PS-Mon-pm/113.
- Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, 79, 281–299.
- Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management Science*, 39, 1179–1189.
- Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, 111, 757–769.
- Wedell, D. H. & Pettibone, J. C. (1996). Using judgements to understand decoy effects in choice. *Organizational Behavior and Human Decision Processes*, 67, 326–344.
- Willemssen, M. C., Böckenholt, U. & Johnson, E. J. (2011). Choice by value encoding and value construction: processes of loss aversion. *Journal of Experimental Psychology: General*, 140, 303–324.

## Appendix

Binary choice sets used in the experiment: 20 consumer products or services and their two attributes

Consumer product or service	Two attributes
Cell-phone	Number of distinctive functions Weight (g)
Electronic dictionary	Types of useful dictionaries Weight (g)
MP3 Player	Recording capacity (Number of tunes) Weight (g)
Digital watch	Quality of design (1–100) Price (thousand ¥)
Notebook computer	Screen size (inch) Weight (g)
LCD TV	Screen size (inch) Price (thousand ¥)
HDD DVD Recorder	Video recording time (hour) Price (thousand ¥)
Digital camera	Image quality (megapixels) Weight (g)
Video camcorder	Image quality (megapixels) Weight (g)
Component stereo	Sound quality (1–100) Price (thousand ¥)
Sport shoes	Quality of design (1–100) Price (thousand ¥)
School bag	Quality of design (1–100) Weight (g)
Single sofa	Comfort in seating (1–100) Price (thousand ¥)
City bike	Quality of design (1–100) Price (thousand ¥)
Gas scooter	Quality of design (1–100) Gas mileage (km per liter)
Rented apartment	Monthly rent (thousand ¥) Walking distance from the station to the apartment (min)
Fitness club	Repletion of equipment (1–100) Time taken to reach the fitness club from home (min)
Hair saloon	Magazine's rating of skill (1–100) Time taken to reach the saloon from home (min)
Restaurant	Magazine's rating of skill (1–100) Time taken to reach the restaurant from school (min)
Part-timer at eating and drinking place	Hourly wage (¥) Time taken to reach from home to that place (min)

# Mental Arithmetic Efficiency: Interactivity and Individual Differences

Frédéric Vallée-Tourangeau

Department of Psychology, Kingston University  
Kingston-upon-Thames UNITED KINGDOM KT1 2EE  
f.vallee-tourangeau@kingston.ac.uk

## Abstract

Thinking efficiency as a function of interactivity was examined in a mental arithmetic task. Participants carried out single-digit additions, involving either 7 or 11 numbers, as fast and as accurately as possible. They completed the sums in blocks, five from the 'easy' set first, and five from the 'hard' set second. These sets were interpolated among a series of other tasks that measured numeracy, working memory capacity, visuo-spatial processing speed and attention switching, in such a way as to permit the presentation of the sets twice, once with each of the sums presented on a piece of paper and participants placing their hands flat on the table and once with the sums presented as a set of manipulable tokens. Efficiency was measured as the ratio of performance over time invested. A significant interaction between condition and difficulty was observed: Efficiency was slightly better in the static condition for easy sums but declined substantially relative to the interactive condition for hard sums. Regression analyses revealed that in the static condition 22% of the variance in efficiency for the harder sums was explained by numeracy and working memory capacity, but 45% by numeracy, working memory capacity and attention switching skills in the interactive condition. Verbal protocols revealed that paths to solution and arithmetic strategies were substantially transformed by the opportunity to manipulate tokens.

**Keywords:** Mental arithmetic, interactivity, efficiency, individual differences, distributed cognition

## Introduction

Mental arithmetic is clearly an important skill with many quotidian applications. It is the quintessential example of what Kahneman (2011) calls "slow thinking": "(a) deliberate, effortful, and orderly" (p. 20) mental process that can be slowed down by a working memory busy holding information about interim steps and selecting strategies to proceed closer to the result. To be sure, for very simple arithmetic problems, answers are retrieved rather than computed; but as problem complexity increases, performance is constrained by limited internal resources.

The role of working memory in mental arithmetic is clearly revealed with experiments employing a dual-task methodology: Performance is significantly impaired by concurrent tasks that tax different components of working memory (e.g., Logie, Gilhooly, & Wynn, 1994). There is

a substantial body of evidence that implicates working memory deficits and poor maths performance in primary school children (e.g., McLean & Hitch, 1999). In adults, the impact of maths anxiety (Ashcraft & Kirk, 2001) and test pressure (DeCaro, Rotar, Kendra, & Beilock, 2010) is explained in terms of the rehearsal and retrieval of performance related thoughts and memories that limit the working memory resources that can be committed to solving the problem.

## Interacting with External Resources

When confronted with internal resource limitations, reasoners naturally mine their surrounding physical space for additional resources. "Artifacts saturate everyday environments" (Kirsh, 2009a, p. 284) and they are routinely recruited to supplement and augment internal cognitive resources. Within such an extended cognitive system (Wilson & Clark, 2009) internal and external resources are coupled by actions, producing a dynamic distributed problem representation. As a result, performance may surpass a level of accuracy and efficiency achievable on the basis of resources internal to the reasoner alone.

Recent experiments on insight and non-insight problem solving reveal how interactivity transforms performance. For example, release from mental set in Luchins's well known volume measurement problems is significantly facilitated when participants interact with actual jars with water (Vallée-Tourangeau, Euden, & Hearn, 2011). Additionally, insight in matchstick algebra problems is substantially enhanced when participants solve these problems with actual matchstick-like objects that permit the physical re-arrangement of the problem representation (Weller, Villejoubert, & Vallée-Tourangeau, 2011). Performance is facilitated by the affordances offered by a modifiable problem representation. In the case of mental set, the physically available resources are more easily perceived as offering simpler and less costly solutions (in terms of pouring and transposing) and help defuse mental set. As for matchstick algebra, the physical movement of a matchstick transforms the presentation of the problem which anchors new mental projections of potential solutions that in turn can be reified by additional physical modification. Insight is thus better driven by a concrete and explicit project-create-project cycle (Kirsh, 2009b).



**Mental Arithmetic.** As for mental arithmetic, recruiting artifacts, such as pen and paper, substantially augments performance largely because working memory content is nearly completely off-loaded onto the external environment. In this case, potential working memory limitations can be compensated by externalising the algorithmic process. While measures of working memory processing capacity may well be correlated with unaided mental arithmetic performance, these correlations would likely disappear when the process is completely externalised (or indeed delegated to a computational device). Thus, examining the role of interactivity in mental arithmetic may more fruitfully proceed in a cognitive system where reasoners cannot record subtotals and remainders in arriving at a solution, but still can interact and modify a physical problem representation (Neth & Payne, 2011).

The experiment reported here examined the impact of interactivity on mental arithmetic. Participants completed simple additions involving single-digit numbers. These additions were carried out for sets of 7 or 11 numbers. Thus one of the independent variables was problem difficulty. The second independent variable was interactivity. In one condition, participants completed the sums by looking at the set of numbers with hands down on the table in front of them. In a second condition, the sums were presented as a set of movable tokens: Participants were free to manipulate and re-arrange the tokens to arrive at a solution. Engineering an extended cognitive system such as the one created through interacting with number tokens may augment performance and enhance reasoning efficiency. The shaping and re-shaping of the physical representation of the problem may encourage and cue different paths to solution and different arithmetic strategies. Limited internal resources in the absence of interactivity may constrain the manner with which participants arrive at a solution.

Measuring efficiency involves assessing the benefits accrued as a function of cost or resources invested. An index of efficiency was calculated as the ratio of performance accuracy –proportion of correct answers– over the proportion of time invested to solve the problem out of the maximum time the slowest participants required to solve the task (Hoffman & Schraw, 2010, refer to such a measure as a likelihood model). Efficiency might be improved in an interactive context because some aspect of executive control is governed, guided and constrained by the shifting physical representation of the problem, freeing internal resources to ensure arithmetic accuracy. In other words, fewer resources are devoted to rehearsing subtotals or identifying and re-identifying the numbers to be added with a dynamic configuration of the sum to complete, enabling participants to devise more creative and efficient ways to solve the problems.

Finally, individual differences in terms of skills and

working memory processing capacity were measured and correlated with performance in the different experimental conditions. Patterns of correlations can help understand more precisely how coupling of internal and external resources lead to better performance. Importantly, the experiment employed a repeated-measures design. Thus the same participants completed the easy and hard sets in both the static and interactive conditions: Between-subjects variance could not explain differences in performance across the experimental conditions.

## Method

### Participants

Forty two university undergraduates (35 females, overall mean age = 21.8,  $SD = 6.8$ ) received course credit for their participation. Three additional participants (all females, mean age 23.0) were later recruited to provide verbal protocols while they performed the easy and hard sums in both conditions.

### Material and Measures

**Numeracy.** Numeracy was measured using the subjective numeracy scale developed by Fagerlin, Zikmund-Fisher, Ubel, Jankovic, Derry, and Smith (2007) which consists of eight questions (such as “how good are you at calculating a 15% tip”). Participants answer using a 7-point scale (1 = “not good at all” and 6 “extremely good”). An objective measure of arithmetic skill was designed by having participants complete as many simple problems (such as  $11 - 9 = ?$ ) as they could in 60 seconds.

**Visuo-spatial information processing speed.** The clerical checking subtest of the Beta III (Kellog & Norton, 1999) was used to measure visuo-spatial processing speed. In this test, participants must identify whether two symbols, figures or strings of digits are identical or not. The measure is the number of correct judgments out of a possible 55 in a 2-min period.

**Executive function: Shifting.** Attention switching skills were measured using the plus-minus task (Miyake, Friedman, Emerson, Witzki, & Howerter, 2000). Using three different series of 30 double-digit numbers, participants were instructed to add 3 to each in the first series, subtract 3 to each in the second series, and alternate between adding and subtracting 3 with the third series. The switching cost, measured in seconds, was the difference in completion time for the third series minus the average completion time for the first two.

**Working memory capacity.** Working memory capacity was assessed with a modified reading span test. Sentences in series ranging in number from 3 to 6 were presented on index cards to participants which they read aloud. At the

end of a series they were prompted to recall the last word of each sentence in that series. There were two different series for each sequence length for a total of 36 sentences. Working memory performance was measured as the total number of words recalled.

**Arithmetic Task.** Participants carried out single-digit additions, involving either 7 or 11 numbers (see Fig. 1), as fast and as accurately as possible. They completed the problems in blocks, five from the ‘easy’ set first, and five from the ‘hard’ set second. Performance was measured as the proportion of correct sums, the mean absolute deviation from the actual sums, the mean latency to announce a solution, and in terms of efficiency. Efficiency was measured as the ratio of addition accuracy (proportion correct sums) over time invested in the task. The latter was measured as the proportion of actual time to complete the sums divided by the maximum time needed to complete them in that condition; this maximum was determined by taking the average of the top quartile latencies. A ratio smaller than 1 meant that proportion accuracy was smaller than proportion time invested, indicating inefficient performance.

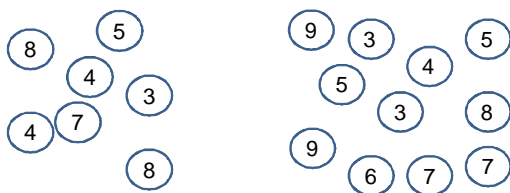


Figure 1: Examples of single-digit additions from the ‘easy’ set (7-digit additions) and the ‘hard’ set (11-digit additions). Participants performed 5 additions from both sets for a total of 10.

## Procedure

Participants first completed the 8-item subjective numeracy scale, followed by the objective arithmetic test, the clerical checking subtest from the Beta III, and the plus-minus task. They were then presented with the five additions from the ‘easy’ set. After a 2-min distractor task (a word search puzzle), participants were presented with the five additions from the ‘hard’ set. These two sets of sums were presented twice to the participants. For one presentation participants performed the additions with their hands on the table facing them (the static condition) and announced their answer out loud; for the second presentation, numbered tokens (2-cm in diameter) were used, and participants were encouraged to move the tokens about in helping them add the numbers (the interactive condition); as in the static condition, participants announced the solution for each problem out loud. While the hard set always followed the easy set, the order of condition (static, interactive) was counterbalanced across participants. With 10 different problems, involving 10

unique configurations of digits, and 90 digits across the two sets, it was unlikely that participants remembered the solution to each problem when presented a second time. Still, to prevent a direct retrieval of solutions during the second presentation, the participants completed the reading span test which lasted approximately 10 minutes. After this test of working memory, participants were presented with the 10 sums again (either in the interactive or static condition depending on which they had experienced first). Thus set size (with two levels) and interactivity (with two levels) were independent variables that were manipulated within subjects in a 2x2 repeated measures design. The experimental session lasted approximately 45 minutes.

## Results

The order of presentation of the interactivity conditions did not significantly influence performance on any of the dependent measures nor did set repetition: Performance on the first 10 sums was no different than performance on the second iteration of the same 10 problems within each experimental condition. Hence, order and repetition were not included in any of the analyses reported below.

### Percent Correct

The mean percent correct solutions for the easy and hard sums are plotted in the top left quadrant of Figure 2. Interactivity did not influence performance for the easy sums, but substantially enhanced performance for the hard sums. In a 2x2 repeated measures analysis of variance (ANOVA), the main effect of condition was significant,  $F(1, 41) = 6.58, p = .014$ , as were the main effect of difficulty,  $F(1, 41) = 20.9, p < .001$  and the interaction,  $F(1, 41) = 12.5, p = .001$ .

### Absolute Error

Non-interactive mental addition did not lead to larger absolute deviations from the correct solution for the easy set, but did for the hard set (see top right quadrant of Fig. 2). In a 2x2 repeated measures ANOVA, the main effect of interactivity was significant,  $F(1, 41) = 13.8, p = .001$ , as was the main effect of difficulty,  $F(1, 41) = 28.6, p < .001$ ; the more important pattern was the significant interaction between condition and difficulty,  $F(1, 41) = 28.9, p < .001$ .

### Latency to Solution

Set size had a large impact on solution latencies (see Fig 2. bottom left quadrant). Interactivity influenced latencies in an interesting manner: For the easy sums, interactivity slowed down participants (by nearly 2.5 s), but marginally reduced latencies (by .4 s) with the hard sums. In a 2x2 repeated measures ANOVA, the main effect of interactivity was not significant,  $F(1, 41) = 1.45, p = .236$ , but the main effect of difficulty was significant,  $F(1, 41) =$

182,  $p < .001$ , as was the interaction,  $F(1, 41) = 6.64$ ,  $p = .014$ .

## Efficiency

Participants were more efficient when solving the easy problems without the tokens (see bottom right quadrant of Fig. 2). Efficiency dropped marginally for the hard sums when participants could use the tokens, but dipped substantially without the tokens. In a 2x2 repeated measures ANOVA, the main effect of condition was not significant,  $F < 1$ , but the main effect of difficulty,  $F(1, 41) = 13.3$ ,  $p = .001$ , as well as the condition by difficulty interaction,  $F(1, 41) = 10.6$ ,  $p = .002$ , were significant.

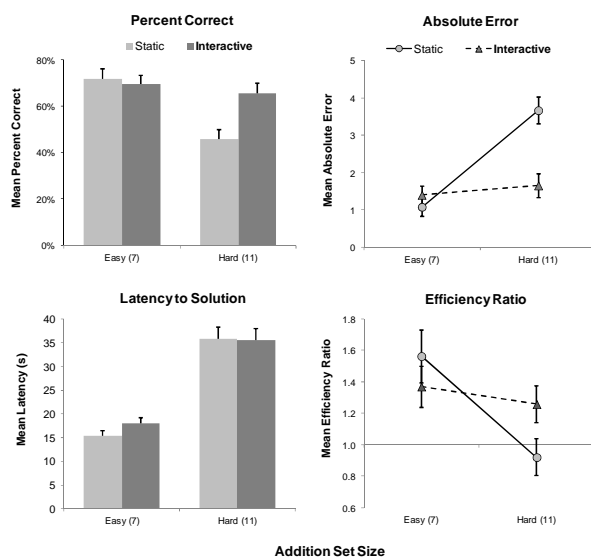


Figure 2: Mean percent correct additions in the static (light bars) and interactive (dark bars) condition (top left quadrant); mean absolute error per sum in the static (circles) and interactive (triangle) conditions (top right quadrant); mean latencies in the four conditions (bottom left quadrant); mean ratio of correct proportion over proportion of maximum time to complete problem (or efficiency ratio) in all four conditions (bottom right quadrant) as a function of set size. Error bars are standard error of the means.

**Predictors of efficiency.** To better understand the relative contribution of different internal resources to performance in the hard set of additions, initial analyses determined the nature of the correlations between efficiency and individual differences (see Table 1). The strongest correlations were observed in the interactive condition with objective numeracy,  $r(40) = .43$ ,  $p = .005$ , and attention switching,  $r(40) = -.39$ ,  $p = .01$ ; in the static condition, objective numeracy was significantly correlated with efficiency,  $r(40) = .32$ ,  $p = .04$ . A stepwise regression analysis for the static condition produced a significant model,  $F(2, 41) = 5.40$ ,  $p = .009$ , composed of objective numeracy ( $\beta = .400$ ) and reading span ( $\beta = .350$ ) that

explained 22% of the variance in efficiency. In the interactive condition, the analysis identified a significant model,  $F(3, 41) = 10.2$ ,  $p < .001$ , that explained 45% of the efficiency variance; the model included objective numeracy ( $\beta = .447$ ), reading span ( $\beta = .426$ ) and attention switching ( $\beta = -.398$ ).

Table 1: Correlation matrix involving individual differences in terms of subjective and objective numeracy, clerical checking, attention switching, reading span and the efficiency ratio in the static and interactive condition for the hard set (involving 11 single digit numbers);  $df = 40$ .

	1	2	3	4	5	6	7
	SBJ-N	OBJ-N	C-C	Att-S	Span	ER-S	ER-I
1	-	.47 **	.16	-.09	.16	.29	.38 *
2		-	.26	-.20	-.24	.32 *	.43 **
3			-	.04	-.03	.10	.16
4				-	.22	-.15	-.39 *
5					-	.26	.23
6						-	.60 **
7							-

Note: \*  $p < .05$  \*\*  $p < .01$ . SBJ-N = Subjective numeracy; OBJ-N = Objective numeracy (basic arithmetic skill); C-C = Clerical Checking; Att-S = Attention Switching; Span = Reading Span; ER-S = Efficiency in the static condition; ER-I = Efficiency in the interactive condition.

## Path to Solution and Strategies

In order to obtain a window onto the paths to solution and the strategies employed to chart these paths in both conditions, three additional participants completed the mental arithmetic tasks while verbalising their progress – the sessions were also videotaped. Inferential statistics could not be performed on data from such a small sample, but very clear differences in strategies emerged in the two conditions.

The simplest strategy, and in the static condition the one that taxes working memory the least, is to add the numbers in the order scanned, without seeking to group numbers to create more congenial sub-totals. Across the three participants, and over all problems, the sequential scan strategy was used exclusively 15 times in the static condition (or for 50% of the problems) and twice in the interactive condition. There were 26 instances of grouping numbers (mostly in pairs) on the path to solution in the static condition, but 75 instances of such groupings in the interactive condition. Congenial sub-totals (defined as  $\Sigma \text{MOD } 5 = 0$ ) on the path to solution was observed 28 times in the static condition but 53 times in the interactive condition. Figure 3 below illustrates the paths to solution and strategies employed by participant 44 for problem A, a 7-number addition. She clearly employed a sequential scanning strategy in the static condition, but was much more creative in the interactive condition, grouping numbers to produce convenient sub-totals to arrive at the solution.

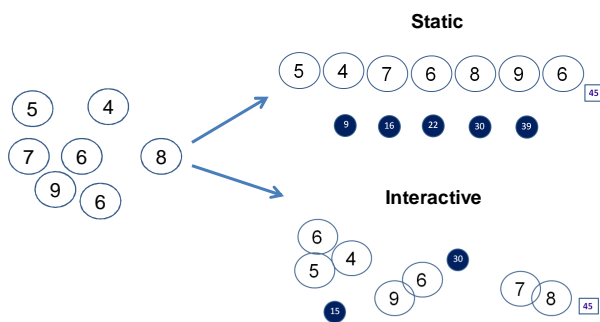


Figure 3: Path to solution and strategy employed for problem A (a 7-number problem) by participant 44 in the static and interactive condition.

## Discussion

This experiment examined mental arithmetic in conditions where participants only used their internal cognitive resources to complete easy and hard sums of single digit numbers or where they could couple their cognitive resources to modifiable external resources in completing the sums. The experiment employed a repeated measures design such that the same participants completed the arithmetic problems in both conditions, thus eliminating between condition variance due to between-subjects differences. This is a particularly important feature of this experiment because it ensured that whatever benefits were conveyed in the interactive condition, these could not be attributed to better or different internal resources brought to the task by a different group of participants.

Interactivity substantially enhanced performance in terms of accuracy and efficiency with the harder sums involving 11 single-digit numbers: Participants were more accurate and the wrong answers were closer to the actual sums in the interactive condition than in the non-interactive condition. Solution latencies offered a gauge of the effort invested to solve the additions. With the hard sets the mean latencies were nearly identical between conditions (35.9 s vs. 35.5 s in the static and interactive condition, respectively) but mean percent accuracy was 20% higher in the interactive condition. Hence, reasoning efficiency was substantially enhanced by allowing participants to couple and regulate their cognitive efforts with a continuous reconfiguration of the tokens in a manner that best served their goal. With the easier sums participants performed marginally better without manipulating tokens, relying solely on their internal resources. The degree to which the design of an extended cognitive system can augment performance is clearly relative to the degree of task difficulty and the cognitive ability of the reasoner (Webb & Vallée-Tourangeau, 2009).

Interactivity offered the opportunity to deploy more creative and efficient paths to solution, which was clearly beneficial for the harder sums. The improvement in performance and the greater efficiency in the interactive

condition was not simply a matter of off-loading content from working memory onto the environment. Rather, a shifting environment suggests different arithmetic paths and permits the identification of congenial interim sums that simplify the task and enhance efficiency. Thus the opportunity to interact with the tokens substantially transformed the nature of strategies employed and the paths to solution. Some of these paths might have been discovered strategically or accidentally by moving the tokens. Still, a dynamic physical presentation of the problem shouldered some of the executive functions freeing resources to better plan how to achieve the goal efficiently. These data support the conjecture that reasoners are better able to deploy arithmetic skills, and may be more receptive to learning new ones, in an environment that augments storage and processing capacity through the coupling of internal and external resources.

## Individual Differences

Profiling participants in terms of cognitive skills and capacities and then correlating these measures with indices of performance help identify the cognitive factors that drive mental arithmetic. This approach has been employed with some success to identify the skills and capacities implicated in insight and non-insight problem solving (Gilhooly & Fioratou, 2009). The resulting data inform the development of process models of performance in these problems. Such process models will likely differ for tasks that are purely reliant on internal cognitive resources in comparisons with tasks that afford a tighter reciprocal influence between cognition, perception and action.

In the static condition, basic arithmetic skills and working memory capacity explained 22% of the variance in efficiency for the harder sums. In the interactive condition, nearly 50% of the variance in efficiency was explained by a model composed of arithmetic skill, working memory capacity and attention switching. These findings suggest that participants with better arithmetic skills, larger working memory capacity and swifter attention switching abilities were more likely to benefit from interacting with tokens in arriving at a solution. In other words, the coupling of internal and external resources was more effectively deployed by participants with better internal resources. This pattern of results was also observed in a recent experiment that contrasted non-interactive and interactive version of Luchins's volume measurement problems: Participants scoring higher in fluid intelligence performed better with the interactive version of the task (Vallée-Tourangeau et al., 2011). Designing an interactive version of an otherwise non-interactive static problem solving task does not benefit every reasoner in the same way. Future research should also determine whether non-intellectual factors such as anxiety or self-efficacy mediate the impact of interactivity on problem solving performance.

The measure of working memory capacity explained unique variance in performance efficiency in the non-interactive context for the hard sums. Still, the correlation between performance and working memory capacity was modest. This finding suggests one of two things. The first is that the task may not have taxed working memory that much. Certainly the degree of absolute departure from the correct answers in the non-interactive condition suggests that participants rarely miscalculated sums by a substantial margin. Future research may thus more fruitfully contrast non-interactive and interactive conditions with a more challenging arithmetic task, either by using larger single-digit sets (e.g., sums including 15 or more numbers) or by using double-digit numbers. A better window onto the role of interactivity in supplementing working memory capacity might be proffered by a task that is more reliant on working memory when it is completed without interaction. Second, the exact composition of the complex span measure of working memory should include arithmetic material and operations. There is evidence to suggest that span and outcome measures are better correlated when they share a domain (DeStefano & Lefevre, 2004).

### Acknowledgments

I would like to thank Ellie McCourty, Angie Makri, Svetlana Stefanova, and Joakim Westh Wiencken for recruiting and running the participants; as well as Ken Gilhooly, David Gilmore, David Kirsh and Gaëlle Villejoubert for helpful comments on an earlier version of this manuscript. Financial support from the Kingston University Faculty of Arts and Social Sciences Research Capability Fund is gratefully acknowledged.

### References

- Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General*, 130, 224-237.
- DeCaro, M. S., Rotar, K. E., Kendra, M. S., & Beilock, S. L. (2010). Diagnosing and alleviating the impact of performance pressure on mathematical problem solving. *Quarterly Journal of Experimental Psychology*, 63, 1619-1630.
- DeStefano, D., & LeFevre, J.-A. (2004). The role of working memory in mental arithmetic. *European Journal of Cognitive Psychology*, 16, 353-386.
- Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the Subjective Numeracy Scale. *Medical Decision Making*, 27, 672-680.
- Gilhooly, K. J., & Fioratou, E. (2009). Executive functions in insight versus non-insight problem solving: An individual differences approach. *Thinking and Reasoning*, 15, 355-376.
- Hoffman, B., & Schraw, G. (2010). Conceptions of efficiency: Applications in learning and problem-solving. *Educational Psychologist*, 45, 1-14.
- Kahneman, D. (2011). *Thinking, fast and slow*. London: Allen Lane.
- Kellog, C. E., & Morton, N. W. (1999). *Beta III Manual*. The Psychological Corporation. A Harcourt Assessment Company.
- Kirsh, D. (2009a). Problem solving in situated cognition. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 264-306). Cambridge: Cambridge University Press.
- Kirsh, D. (2009b). Projection, problem space and anchoring. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2310-2315). Austin, TX: Cognitive Science Society.
- Logie, R. H., Gilhooly, K. J., & Wynn, V. (1994). Counting on working memory in arithmetic problem solving. *Memory & Cognition*, 22, 395-410.
- McLean, J. F., & Hitch, G. J. (1999). Working memory impairments in children with specific arithmetic learning difficulties. *Journal of Experimental Child Psychology*, 74, 240-260.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex frontal lobe tasks: A latent variable analysis. *Cognitive Psychology*, 41, 49-100.
- Neth, H., & Payne, S. J. (2011). Interactive coin addition: How hands can help us think. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society* (pp. 279-284). Austin, TX: Cognitive Science Society.
- Vallée-Tourangeau, F., Eudén, G., & Hearn, V. (2011). Einstellung defused: Interactivity and mental set. *Quarterly Journal of Experimental Psychology*, 64, 1889-1895.
- Webb, S., & Vallée-Tourangeau (2009). Interactive word production in dyslexic children. In N. Taatgen, H. van Rijn, J. Nerbonne & L. Schomaker (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (1436-1441). Austin, TX: Cognitive Science Society.
- Weller, A., Villejoubert, G., Vallée-Tourangeau, F. (2011). Interactive insight problem solving. *Thinking & Reasoning*, 17, 429-439.
- Wilson, R. A., & Clark, A. (2009). How to situate cognition: Letting nature take its course. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 55-77). Cambridge: Cambridge University Press

# Conceptual alignment in reference with artificial and human dialogue partners

Koen van Lierop (koenvanlierop@gmail.com)

Martijn Goudbeek (m.b.goudbeek@uvt.nl)

Emiel Krahmer (e.j.krahmer@uvt.nl)

Tilburg center for Cognition and Communication (TiCC), Faculty of Humanities, Tilburg University,  
PO Box 90153, 5000 LE Tilburg, The Netherlands

## Abstract

Previous work on reference in dialogue has shown that speakers adapt to the concepts that were used in earlier references during an interaction (such as orientation when a dialogue partner describes a chair as “the chair seen from the front”), even if these concepts are generally dispreferred. Here, we investigate to what extent it matters whether speakers interact with an artificial or a human dialogue partner (Study 1) and whether this adaptation indeed takes place at the conceptual level (Study 2). For Study 1 participants interacted either with a computer or with a human confederate and it was found that participants adapt in similar ways and just as much to a human dialogue partner as to a computer. Study 2 used a cross-language interaction paradigm, in which bilingual participants listened to English descriptions after which they had to refer in Dutch (thereby reducing the possibilities for lexical and syntactic alignment). The results showed that even with crosslinguistic prime-target pairings, participants aligned with the attributes used by their dialogue partner, providing further evidence for alignment at the conceptual level.

**Keywords:** referring expressions, alignment, human-computer interaction, conceptual alignment

## Introduction

During conversations, people continuously refer to other people, objects or events, for example using descriptions such as *the man with the beard* or *the blue chair*. Since such descriptions are so common (Poesio & Vierra, 1998), their underlying production process has drawn many researchers’ attention, both from a computational and from a psycholinguistic perspective. Much of this research focusses on the question of what makes people choose one possible way of referring to an object over another. Why do speakers include certain properties in their descriptions and others not?

Computational studies of reference often frame the production of referring expressions as a problem of choice where a (usually fixed) preference order determines the order in which attributes (such as color or orientation) are considered for inclusion in a referring expression. The Incremental Algorithm (Dale & Reiter, 1995), for instance, which is probably the most widely used algorithm in this field, operates according to this principle, assuming the existence of a domain-dependent preference order on the

relevant attributes, and first tries out preferred attributes before less preferred ones are considered, thereby modeling the intuition that speakers prefer certain attributes over others, partly based on findings of Pechmann (1989). For example, when referring to a chair, speakers are more likely to refer to its color (*the blue chair*) than to its orientation (*the chair facing left*) even though both may be successful in singling out a particular chair.

However, one could argue that the Incremental Algorithm is “addressee-blind” (Clark & Bangerter, 2004) in that it pays no attention to references that were produced earlier in an interaction (the same holds, incidentally, for the various alternatives that have been proposed to the Incremental Algorithm and which are surveyed in Krahmer and van Deemter, 2012). Indeed, it has been shown that speakers do take prior references into account during reference production. One study, for example, found that if one dialogue participant refers to a couch as a *sofa*, the next speaker is more likely to use the word *sofa* as well (Branigan, Pickering, Pearson, & McLean, 2010). This can be seen as a lexical form of “alignment” (Garrod & Pickering, 2004; Pickering & Garrod, 2004) between speaker and addressee. Pickering and Garrod argue that alignment may take place on all levels of interaction, and indeed it has been shown that participants also align, for example, their intonation patterns and syntactic structures.

Goudbeek and Krahmer (2010, 2012) have shown that a similar kind of adaptation can take place at the conceptual level of attributes. While the Incremental Algorithm and its ilk only predict the use of dispreferred attributes when preferred attributes alone are unable to uniquely identify a given target object, Goudbeek and Krahmer showed that participants do use dispreferred attributes when these were used earlier in an interaction. In particular, they found that when one dialogue partner used a dispreferred attribute to refer to an object, the other dialogue partner was more likely to use values of that attribute in subsequent references as well. For example, in the furniture domain used in their experiments, participants could always uniquely identify an object by using its color (e.g., *the green chair*<sup>1</sup>) or its orientation (*the chair seen from the front*). Of these two attributes, color is the preferred one, in the sense that

---

<sup>1</sup> Here and elsewhere we provide English translations of Dutch originals.

without prior context speakers are more likely to use color than orientation, as was firmly established independently for Dutch (Koolen, Gatt, Goudbeek, & Krahmer, 2011) and English (van Deemter, Gatt, van der Sluis, & Power, 2012). Yet, Goudbeek and Krahmer found that when the participants had been exposed to descriptions such as *the chair seen from the front*, they were themselves more likely to use the dispreferred orientation attribute in their own references, despite the fact that using color would have been perfectly sufficient. Goudbeek and Krahmer (2010, 2012) established these effects in two different referential domains (people and furniture).

Crucially, Goudbeek and Krahmer suggest that their findings cannot be explained in terms of lexical alignment: participants were primed, for instance, with *seen from the front*, while the target was *facing left*. They argue that instead they found evidence for what they call “conceptual alignment”, where participants align at the level of attributes (such as orientation) and not values (such as seen from the side). However, two potential criticisms could be levelled against this claim: first, in their experiments, participants interacted with a computer rather than with another participant, and it is conceivable that this influenced the conclusions; and second, given that primes and targets were always referred to using Dutch descriptions, the possibility that some, possibly indirect, form of lexical alignment influences the results cannot be discarded. In this paper, we address these two criticisms.

The first question we therefore ask is to what extent the results obtained with participants interacting with a computer, using a procedure in which participants had to repeatedly listen to a prerecorded description and respond with a description in front of a computer screen, are representative for human-human interaction. While using a computer-based paradigm offers several advantages in an experimental context (especially concerning controllability), Branigan et al. (2010) point out some dangers in drawing conclusions about alignment in human-human interactions from human-computer interactions (HCI). They indicate that alignment in the latter kind of interactions is mainly based on considerations of communicative success, the speaker’s model of the computer and what they think the computer might or might not be able to know. Such considerations can override strong linguistic preferences, which Branigan and colleagues interpret as signs of HCI not always being a reliable predictor of real interactions between humans. Branigan et al. (2010) also argue that alignment in HCI is potentially stronger than alignment in interactions between humans; humans, they reason, may have doubts about the communicative capabilities of computers, and hence might be more inclined to adapt to the computer. This suggests that the previous findings of Goudbeek and Krahmer could have overstated the amount of alignment in interactive referring expression generation.

On the other hand, one could also argue that the participants in these studies did not strictly engage in human-computer interaction, but rather could be argued to

be interacting with an “imaginary audience”. Various recent studies, including Ferreira, Slevc and Rogers (2005) and Van der Wege (2009), have shown that there generally are only small differences between referring for a real and an imagined audience. References produced by participants are not more precise when they are interacting with a real instead of an imaginary addressee (Van Der Wege, 2009) nor do participants avoid potential ambiguities in their references more when they are speaking to a real addressee (Ferreira et al., 2005).

Ultimately, however, this an empirical question, which we address in Study 1. In this study, we attempt to replicate the previous findings from Goudbeek and Krahmer with human dialogue partners. Here, participants took part in an otherwise identical interactive alignment paradigm, the only difference being that instead of with a computer, participants communicated with another person which (unbeknownst to them) was a confederate of the experimenter. The participants and the confederate took turns in referring to objects and identifying objects based on the descriptions produced by their dialogue partner. As in the original paradigm of Goudbeek and Krahmer (2010), participants could always use either a preferred or a dispreferred attribute to refer to an object, and depending on the condition, the confederate either included a preferred or a dispreferred attribute when referring to an object. We compute the amount of alignment, and compare it with the amount of alignment observed in the earlier, computer-based study.

The second question we address in this paper is to what extent the alignment observed in the referential tasks indeed occurs at the *conceptual* level. In Levelt’s model of speech production (Levelt, 1989; 1999), the conceptual level is the level at which the speaker decides which information to put into an utterance. In contrast to lexical alignment, where the use of *sofa* by one dialogue partner may trigger the switch from *couch* to *sofa* in the other partner, conceptual alignment refers to alignment with respect to the attribute and not necessarily with the value used by the speaker. In the previous experiments, there certainly was no direct relation between the attribute value of the prime and that of the target (and the prime and target were always separated by a pair of unrelated fillers). Nevertheless, it could be that attribute values occurring elsewhere in the experiment as primes somehow lexically primed values over longer distances, or that some other form of lexical or syntactic alignment played a role in one way or another.

To more directly test the claim of conceptual alignment, a crosslinguistic version of the interactive alignment paradigm was devised, inspired by earlier crosslinguistic priming experiments (e.g., Loebell, & Bock, 2003; Schoonbaert, Hartsuiker, & Pickering, 2007). In this experiment, described as Study 2, bilingual participants are exposed to primes in English (*the chair seen from the side*), and subsequently have to describe (after two fillers items) a target in Dutch (e.g., *de stoel van voren*; English (literally) “the chair from the front”). If Dutch participants align



equally frequent with English as with Dutch primes, this would be further, and arguably more compelling evidence for conceptual alignment, since lexical and syntactic priming are relatively less likely between two different languages, even when they are relatively similar as Dutch and English (Schoonbaert et al., 2007). Again, we compare the amount of alignment when primes are in Dutch (as in the original studies) with the amount of alignment when primes are in English.

## Study 1: Artificial vs. Human Partner

This first experiment tests whether the finding of Goudbeek and Krahmer (2010), that people align with their interaction partner by using a dispreferred target attribute in their referring expressions, also holds when the interaction partner is a real person rather than a computer.

### Method

**Participants** 29 Students (23 female) from Tilburg University participated in this experiment, either for partial course credit or a small payment. All participants were fluent speakers of the Dutch language, and had normal hearing and normal (or corrected to normal) vision. None had participated in one of the earlier studies of Goudbeek and Krahmer (2010; 2012).

**Materials** The stimulus material for this study consisted of a set of furniture images and images of people, which have frequently been used in previous research on the production of referring expressions (e.g., van Deemter et al., 2012) and which were also used in Goudbeek and Krahmer (2010; 2012)<sup>2</sup>. The target images were always one of five different furniture items, which varied in both color and orientation. An overview of the possible combinations is provided in Table 1.

The pictures from the people domain (all black and white photographs of famous mathematicians) served as filler images, and were included to distract participants' attention from the goal of this study. Target images were always presented together with two distractors from the same domain, and were shown to participants on two synchronized monitors, to ensure that the set up was identical for the confederate and participant, in order to not raise any suspicions about the confederate with the participants. Targets could always be distinguished both in terms of color (preferred) and in terms of orientation (dispreferred). That color information is preferred and orientation information is not was determined independently from corpus data for Dutch and English (van Deemter et al., 2012; Koolen et al., 2011), showing that speakers frequently

use color spontaneously and that they rarely use orientation when referring to the furniture items under study here.

Table 1: The attributes and their possible values.

Attribute	Possible values
type	chair, desk, fan, sofa, television
color	red, green, blue, grey, black
orientation	front, back, side

**Procedure** Participants took part in an interactive understanding and referring task, closely modelled on the paradigm presented in Goudbeek and Krahmer (2010). In this study, however, participants interacted with a female confederate (a student of the same age as the participants) instead of with a computer. The experiment consisted of two blocks, each featuring 30 trials<sup>3</sup>. During block one, the confederate systematically used preferred attributes when referring to furniture items (e.g., *the green chair*), during block two, the confederate systematically used dispreferred attributes (e.g., *the chair seen from the front*). The order of blocks was counterbalanced across participants. In every trial, participants first had to listen to the confederate describing a critical target (to which we refer as the prime), who referred to one of three pieces of furniture, which participants subsequently had to identify on their screen. After doing so, the next slide of images (three persons) appeared for both participant and confederate, and the participant had to describe a filler target from the person domain to the confederate, who identified it on her screen. Third, the confederate would describe a filler target (again from the people domain). In the fourth turn, the participant described a critical target that could be described with a preferred or dispreferred attribute (or both) to the confederate. Figure 1 shows an example of a critical trial.



Figure 1. Example of a critical trial. The target is indicated by a red border and can be distinguished both in terms of its color (blue) or its orientation (facing left).

The use of a confederate was motivated from the fact that participants were unlikely to systematically use dispreferred properties, which would hinder a direct comparison with the results of Goudbeek and Krahmer (2010; 2012). The confederate was instructed to engage in each interaction and ask questions when participants provided insufficient information to identify a target. After

<sup>2</sup>The pictures of furniture items were taken from the Object Databank, developed by Michael Tarr at Brown University and are freely distributed. URL: <http://titan.cog.brown.edu:8080/TarrLab/stimuli/objects/>

<sup>3</sup> The order of the trials within a block was the same for every participant

the experiment participants were debriefed. None suspected the other person to be a confederate.

## Results

The number of times participants aligned with the attribute they were primed with was used as the dependent variable. This includes cases in which participants used an overspecified referring expression, where both preferred and dispreferred attributes of the target were used by the participant, even though only one of them would suffice for the purpose of identification.

The analysis focuses on the proportional use of dispreferred attributes (orientation) when participants are exposed to dispreferred primes (note that when participants use a preferred prime, i.e., color, we cannot tell whether they used it because it is preferred or because it was primed). The results show that with dispreferred primes, the proportion of dispreferred attributes used by participants is considerable ( $M = 0.41$ ,  $SD = 0.46$ ). Contrary to the predictions of the Incremental Algorithm, the proportion of dispreferred attributes is significantly larger than zero;  $t(28) = 4.82$ ,  $p < .001$ .

To investigate whether it matters if participants were interacting with a computer or a real person, the data of Goudbeek and Krahmer (2010) was compared with the data from the current experiment. Figure 2 displays the proportion of alignment of the participants who had been interacting with a computer ( $M = 0.53$ ,  $SD = 0.43$ ) and those of who interacted with a confederate ( $M = 0.41$ ,  $SD = 0.46$ ). A one-way analysis of variance with interaction partner (computer versus confederate) as the independent variable and proportion of alignment as the dependent variable showed no significant difference between interacting with a computer and interaction with a human  $F(1,48) = 2.20$ ,  $p = .14$ ).

Study 1 revealed that participants have a strong tendency to align with their conversation partner; when their partner uses a dispreferred attribute (referring to a piece of furniture in terms of its orientation) they are more likely to do so themselves later on. Whether their conversation partner is a computer of person has no significant influence.

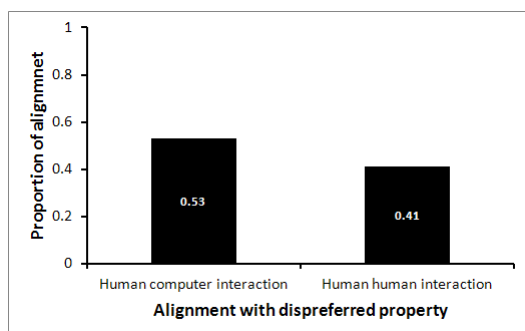


Figure 2. The proportion of alignment in human-computer and in human-human interaction

## Study 2: Crosslingual priming

The second study aims to find more conclusive evidence for the claim that the kind of alignment under discussion here takes place at the conceptual level, rather than the realization one. This is addressed using a cross-language priming experiment, where bilingual participants are not primed in their native Dutch, but in English. They do have to refer in Dutch, however, as in the previous experiment.

## Method

**Participants** 40 Students (31 female) participated in this experiment in exchange for partial course credit. All were unbalanced bilinguals, namely native Dutch speakers with formal instruction in English for 7 years or more. All had normal hearing and (corrected to normal) vision. None had participated in Study 1 or in any of the other studies described in Goudbeek and Krahmer (2010; 2012).

**Materials** The stimuli were identical to those described in Study 1, except that primes were now pre-recorded descriptions produced by a native English speaker (of roughly the same age as the participants), that referred to the objects (e.g., *the chair seen from the left*) with the same preferred and dispreferred attributes (colour and orientation, respectively) as before.

**Procedure** Before starting the experiment, participants were told they had to identify pre-recorded descriptions provided by an English speaker but had to describe the objects themselves in Dutch. In this experiment, following Goudbeek and Krahmer (2010), descriptions were once again produced by a computer, which was warranted by the findings of Study 1. The remainder of the procedure was identical to that of the previously described study.

## Results

As in Study 1, the number of times participants aligned with the dispreferred attribute when this attribute was used earlier in the interaction was used as the dependent measure. If this alignment indeed takes place on the conceptual level, it should not matter whether primes were in English or in Dutch. If, however, the alignment that participants' showed in Krahmer and Goudbeek (2010) was of a lexical or syntactic nature, we predict that people in this study will align not or to a lesser extent with the dispreferred primes, since the linguistic realization of the prime and the target differ considerably.

Figure 3 shows the use of color (the preferred attribute) and orientation (the dispreferred attribute) for each prime type. Clearly, when participants were primed with the dispreferred attribute orientation, they start using that attribute themselves more often (and the color-attribute less).

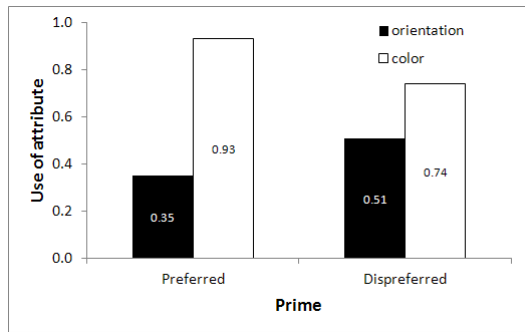


Figure 3. Participants' proportional use of the preferred and dispreferred attribute following preferred and dispreferred primes.

Statistical analysis of the data shows that despite the primes being in English, people still use the dispreferred attribute orientation significantly more than zero with  $t(39) = 7.29$ ,  $p < .001$ , contrary to the prediction of the Incremental Algorithm. Moreover, participants used the dispreferred target attribute more often when they had been primed with a dispreferred attribute than when they had been primed with a preferred attribute,  $F(39) = 10.92$ ,  $p < .005$ .

Furthermore, a comparison was made between the current data and the data that was collected through the experiment conducted by Goudbeek and Krahmer (2010) to test whether the language of the primes influenced the level of alignment. The results of this comparison are depicted in Figure 4. A statistical analysis showed no significant effect of the language of the primes on the amount of alignment with the dispreferred target attribute, with  $t(57) = 0.83$ ,  $p = 0.41$ . Given that the experimental set-up was exactly the same, apart from the manipulation of the primes (English here, Dutch in the earlier study), this shows that the language of the prime has no impact on the amount of alignment with the dispreferred attribute, which we take as further evidence for the claim that the kind of alignment observed here is conceptual rather than lexical.

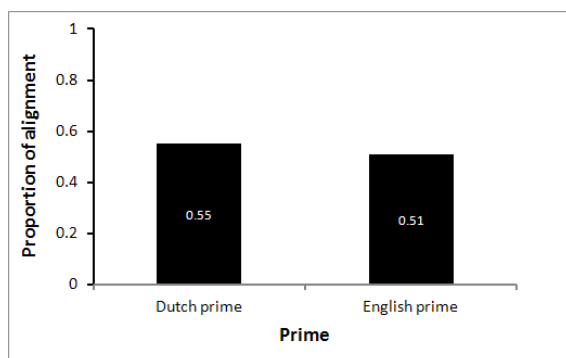


Figure 4. A comparison of the amount of alignment with the dispreferred attribute as a function of the language of the prime.

## Discussion

In this paper, we studied how speakers refer to objects in an interactive setting, and how this influences speakers' decision of which attributes to include in a description. In both studies, we found that speakers have a strong tendency to use dispreferred attributes in their descriptions when these were used earlier in the interaction, even though there was always the possibility to rely only on a preferred attribute. These results replicate the earlier findings of Goudbeek and Krahmer (2010; 2012), and extend them in two important directions.

The results of the first study show that alignment between humans is statistically indistinguishable from alignment between humans and computers. This indicates that the use of a pre-recorded conversation partner in the earlier studies did not influence the results, and is in line with the claims from Ferreira et al. (2005) and van der Wege (2009) that referring for an "imaginary audience" is similar to referring for a real audience. The results of the second study show that participants which are primed with dispreferred attributes in English, use these attributes to the same extent in their descriptions in Dutch as participants that were primed with dispreferred attributes in Dutch. This strongly suggests that the kind of alignment under study here is conceptual, rather than lexical or syntactic. What seems to be primed is a way to conceptualize or look at an object (in terms of orientation rather than color, for instance), rather than a specific property (such as facing left or being blue).

It is interesting to contrast the current findings with the predictions of state-of-the-art computational models for Referring Expression Generation (REG), including the Incremental Algorithm (Dale & Reiter, 1995) as well as more recent extensions and variations of these models. Generally, these models fail to account for the alignment results presented here; the Incremental Algorithm, for instance, predicts that a dispreferred attribute would never be used if a preferred attribute is sufficient to uniquely characterize a target object. The basic problem is that these algorithms treat the decision of which attributes to include in a description as a decision that can be made independent of context, and hence does not need to take into account the reference history, something which our data clearly contradict.

To make these algorithms suitable for the generation of referring expressions in interactive settings (as is required for many applications), they should become more sensitive to the preceding interaction and the references that were produced in it. For the Incremental Algorithm, this could be achieved, for example, by combining the (fixed, domain dependent) list of preferred attributes with a (flexible) list of "previously mentioned" attributes. The relative weighting of these two lists can be estimated based on data such as those presented here.

Gatt, Goudbeek and Krahmer (2011) go one step further, proposing a new model for alignment in reference production that integrates alignment and preference order

based attribute selection. Their model consists of two parallel processes: a preference-based search process based on the Incremental Algorithm, and an alignment-based process. These two processes run concurrently and compete to contribute attributes to a limited capacity working memory buffer that will produce the referring expression. Gatt et al. (2011) show that their model can account for the alignment findings well.

## Conclusion

When producing referring expressions in interactive settings, speakers have a strong tendency to re-use attributes that were used before, even if these attributes are otherwise dispreferred. The frequency with which dispreferred attributes are re-used does not depend on whether the interaction is with a computer or with another person, nor on whether the preceding primes were produced in a different language or not. Taken together, these results suggest that the alignment is automatic, and of a conceptual nature. Current state-of-the-art computational models of reference production fail to account for this, since they tend to be “addressee-blind” and mostly rely on domain-dependent preferences only.

## Acknowledgments

The research reported in this paper forms part of the VICI project “Bridging the gap between psycholinguistics and Computational linguistics: the case of referring expressions”, funded by the Netherlands Organization for Scientific Research (NWO grant 277-70-007). We thank Manon Yassa for serving as the confederate in Study 1 and Elsa Jonker for assistance in running Study 2.

## References

- Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9), 2355-2368.
- Clark, Herbert H., & Bangerter, A. (2004). Changing ideas about reference. In Ira A. Noveck and Dan Sperber, editors, *Experimental Pragmatics*. Palgrave Macmillan, Basingstoke, pages 25–49.
- Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2), 233-263.
- van Deemter, K., Gatt, A., van der Sluis, I., & Power, R. (2012). Generation of Referring Expressions: Assessing the Incremental Algorithm. *Cognitive Science*, in press.
- Ferreira, V. S., Slevc, L. R., & Rogers, E. S. (2005). How do speakers avoid ambiguous linguistic expressions? *Cognition*, 96(3), 263–284.
- Garrod, S. & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, 8, 8–11.
- Gatt, A., M. Goudbeek and E. Krahmer (2011). Attribute preference and priming in reference production: Experimental evidence and computational modeling. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society (CogSci 2011)*, July 20-23, Boston, Massachusetts, 2627-2632.
- Goudbeek, M., & Krahmer, E. (2010). Preferences versus adaptation during referring expression generation. *Proceedings of the 48th annual meeting of the Association for Computational Linguistics (ACL)*. July 2010, Uppsala, Sweden, 55-59.
- Goudbeek, M., & E. Krahmer (2012). Alignment in interactive reference production: Content planning, modifier ordering and referential overspecification. *Topics in Cognitive Science*, 4, 269-289.
- Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43, 3231-3250.
- Krahmer, E., & van Deemter, K. (2012). Computational Generation of Referring Expressions: A Survey. *Computational Linguistics*, 38(1), 173-218.
- Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levelt, W. (1999). Producing spoken language: A blueprint of the speaker. In C. Brown & P. Hagoort (red.), *The Neurocognition of Language* (p. 83-122). Oxford: Oxford University Press.
- Loebell, H., & Bock, K. (2003). Structural priming across languages. *Linguistics*, 41, 791-824
- Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27, 89–110.
- Pickering, M. & Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, 27, 169–226.
- Poesio, M., & Vieira R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24, 183–216.
- Schoonbaert, S., Hartsuiker, R., & Pickering, M.J. (2007). The representation of lexical and syntactic information in bilinguals: Evidence from syntactic priming. *Journal of Memory and Language*, 56, 153-171.
- van der Wege, M. M. (2009). Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, 60, 448-463.

# Neural Circuits for Any-Time Phrase Recognition with Applications in Cognitive Models and Human-Robot Interaction

Richard Veale<sup>1</sup> and Matthias Scheutz<sup>2</sup>

riveale@indiana.edu mscheutz@cs.tufts.edu

<sup>1</sup>Cognitive Science Program, Indiana University <sup>2</sup>Department of Computer Science, Tufts University

## Abstract

Humans are remarkably good at recognizing spoken language, even in very noisy environments. Yet, artificial speech recognizers do not reach human level performance, nor do they typically even attempt to model human speech processing. In this paper, we introduce a biologically plausible neural model of real-time spoken phrase recognition which shows how the time-varying spiking activity of neurons can be integrated into word tokens. We present a proof-of-concept implementation of the model, which shows promise both in terms of recognition accuracy as well as recognition speed. The model is also pragmatically useful to cognitive modelers who require robust any-time speech recognition for their models such as real-time models of human-robot interaction. We thus also present such an example of embedding our model in a larger cognitive model, along with offline analysis of its performance on a speech corpus.

**Keywords:** Liquid State Machine; Neural Network Model; Any-Time Speech Recognition

## Introduction

The mechanisms that convert physical signals such as light and sound into firing rates of neurons are well-studied. However, the way these signals subsequently influence behavior, especially cognitive behavior, is less well understood. In particular, the progression from continuous real-time input streams at the physical transducer level to high-level cognitive processes that abstract over many physical characteristics is challenging, so much so that classical cognitive architectures simply assume higher-level representations such as word tokens instead of modeling the processes that generate them. While these assumptions do not pose problems for disembodied models (e.g., of higher-level cognitive processes such as analogical reasoning in language), they are critical showstoppers for embodied situated models that depend on being implemented on robots that interact with their environments in real-time (e.g., in the context of human-robot interactions in natural language). In such models, sensory processing must be performed one way or another, and while it is sometimes possible to substitute engineering solutions for biologically plausible sensory modules (e.g., artificial speech recognizers instead of biologically plausible models of speech recognition), those substitutions often come at a price that assumptions have to be made about the nature of the interface between those parts of the model that are meant to be biologically plausible and those parts that function as proxies for yet-to-be-developed biologically plausible parts. Specifically, the sensory modules must be able to perform their function (e.g., recognizing words or objects) *at least* as well as humans, or else other components of the model have to account for errors in perceptual processing (e.g., word recognition errors).

Moreover, the sensory module must be able to perform its task and make its result available *at least as fast as* the corresponding module would in humans to be able to respect human timing (e.g., the human expectation to hear a verbal acknowledgment at the right time in response to an utterance). As a result, these two requirements often pre-empt the use of traditional computational methods that perform sensory-input-to-token conversion.

In this paper, we will address the problem of biologically plausible real-time sensory processing of speech signals with a two-fold goal: (1) to provide a biologically plausible neural model of human spoken phrase recognition, and (2) to provide a sensory module that can be embedded in classical cognitive architectures for the study of embodied situated models of natural language interactions. Specifically, we propose a new approach to robust speech recognition which gives continuous access to meaningful partial results, and returns word or phrase tokens that can be directly used by higher-level cognitive models. The model includes several parts of the early auditory processing system in humans: the cochlea (converting time to frequency domain), parts of the olivary complex (applying several filters to the cochlear-processed signal), and a sensory cortical area (comprising a recurrent spiking neural circuit to integrate the signal). Category separation is performed by “readout neurons” (one per category) that respond continuously to ongoing activity in the recurrent circuit based on the particular weighted projection they receive, as is customary for the employed *liquid-state machine* (LSM) neural model (Maass, Natschlager, & Markram, 2002). The weighted projection received by each readout neuron is determined before-hand based on offline training on a speech corpus. The main contributions are thus the implementation of the speech processing neural architecture and the method of converting the instantaneous output of readout neurons to the token-type output for use by subsequent cognitive processes. The paper starts by laying out the background and the problems to be solved by a model that has to convert time-varying instantaneous neural-readout behavior to discrete categories. The subsequent model section then describes the neural speech-encoding and integration parts of the model. Then an analysis of the performance of the model on subsets of a speech-corpus from human-human interaction experiments is presented together with links to videos in which the model is used as part of a larger cognitive model for situated embodied human-robot interaction experiments where speech utterances are used to control the robot’s ongoing behavior.

## Background

Liquid state machines have been proposed as neurologically plausible models of cortical microcircuits (Maass et al., 2002) which can be used for time-invariant categorization of continuous-time signals by way of simple linear “readout units”. The conversion from instantaneous readout activity to discrete tokens that accurately encode the temporally-extended category, however, is not a trivial task as the instantaneous activity of a readout neuron only reflects that the very recent activity (e.g., tens of milliseconds) of the recurrent circuit is similar to what its activity was *at some point* during its response to the category on which the readout neuron was trained. In other words, if a readout neuron  $R$  is trained to respond to an isolated category, e.g., a word  $W$ , and is firing vigorously to a stimulus  $S$  presented to the circuit, it is not clear which part of  $S$  is recognized by the readout as being similar to  $W$ . It is possible that the readout will fire (entirely by coincidence) in response to a 20 ms segment of  $S$  but remain silent otherwise (as that part of the speech signal bears resemblance to patterns that occur in typical speech signals for  $W$ , e.g., similar phones). Obviously, it would be premature in this case to conclude that  $S$  is an instance of the category  $W$  – after all,  $W$  might be typically 500 ms long, and parts of it will be similar to parts of many other words. This dissociation between recognizing matching parts of a word versus recognizing the whole word (or phrase, for that matter), is the first problem to be solved: the challenge here is to produce a mechanism for filtering out coincidental noise and choosing a single “winning” category from among all readout neurons that will have time-variant activations throughout the presentation of  $S$ .

The second problem to be solved is to determine *when* to select a winner (clearly a winner cannot be select at very small time intervals as this would be tantamount to recognizing new words all the time). Since a single stimulus can span thousands of time-steps at which neurons can fire, the model should return a single token exactly once per stimulus and only if the stimulus is one of the categories that it has been trained on. Even if there is a stimulus-length patch of noise, the model should not recognize it as being similar to the readout with the highest activation, but should not detect any word token at all (alternatively, it could detect a “noise token”).

Liquid state machine models have been previously applied to cases where the speech corpus was pre-processed into frequency channels that were guaranteed to have only one spike per word (coding onset, offset, or peak of that frequency band) (Maass et al., 2002). However, these assumptions are unrealistic and it is difficult if not impossible to produce this type of encoding in real-time from raw audio streams. Another approach addressed this encoding limitation and compared the performance of LSMs using different sound-coding front-ends (Lyon cochlear vs. MFCC) as well as different methods for converting the front-ends’ analog output into input spike trains for the LSM (Verstraeten,

Schrauwen, & Stroobandt, 2005). While both approaches performed category-token recognition well by their own metrics (i.e., the ratio of the number of correct readout spikes to total time points in (Maass et al., 2002) and the class with the most readout spikes in response to a given word file in (Verstraeten et al., 2005)), neither method is applicable to real-time speech recognition, since real continuous-time audio is not separated into “files” with a clear stimulus onset and offset. Rather, it is non-trivial to detect the onsets and offsets of real utterances from continuous speech streams, which are often full of non-word noises, variations in word pronunciation, or words on which the system has not been trained.

Hence, we developed novel and realistic neural implementations of onset/offset detectors to increase recognition performance and aid in utterance detection and classification. Specifically, the model is based on the approach of Smith and Faser (2004) who, inspired by Ghitza (1987), present a biologically inspired onset-detection regime using depressing synapses. This implementation via short-term plasticity (STP) synapses is justified based on the evidence provided by MacLeod, Horiuchi, and Carr (2007), who argue that not only synaptic depression, but also facilitation, can play a critical role in auditory processing. Finally, for auditory input signal processing we utilize the Lyon cochlear model (Lyon, 1982; Slaney, 1998) which effectively applies band-pass filters and transformations to sound waves to approximate the firing activity of a set of neural channels along the cochlea.

## Model Architecture and Implementation

Figure 1 is a visualization of the neural circuits implemented for speech recognition without the learned components (i.e., readout neurons and readout integrators). Here, we describe the components of the model together with all parameters used for the empirical evaluation presented later.

### Neural and Synaptic Models

The neural model uses Leaky Integrate-and-Fire (LIF) neurons, whose membrane potential  $V_m$ :

$$\frac{\partial V_m}{\partial t} = \frac{-(V_m - V_{rest}) + R_m \cdot (I_{bg} + I_{syn})}{\tau_m} \quad (1)$$

where  $R_m$  is the membrane resistance,  $I_{bg}$  the background current, and  $I_{syn}$  the total current impinging from afferent synapses.  $-V_m$  represents the leakage term, causing the membrane potential to decay exponentially with time constant  $\tau_m$ . When  $V_m$  reaches the threshold value  $V_{thresh}$ , the neuron “fires” and  $V_m$  is reset to  $V_{reset}$  and enters a refractory period during which it does not update.

The model uses static or dynamic synapses to connect neurons. A static synapse has a post-synaptic response (PSR) that decays exponentially with time constant  $\tau_{psr}$ . The dynamics of the post-synaptic response  $q_{psr}$  of a synapse is thus:

$$\frac{\partial q_{psr}}{\partial t} = \frac{-q_{psr}}{\tau_{syn}} \quad (2)$$



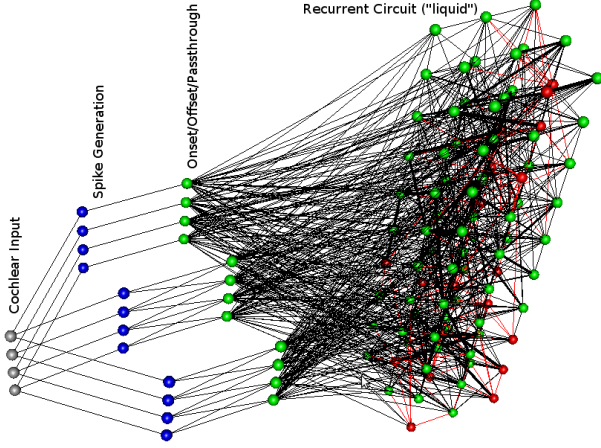


Figure 1: 3-D visualization the neural model described in this paper. The pictured circuit has only 4 input channels, and a  $3 \times 3 \times 10$  recurrent circuit. The actual circuit has 52 input channels and a  $5 \times 5 \times 15$  recurrent circuit. Readout neurons not shown (they would be on the right of the recurrent circuit, receiving input from it).

Synaptic dynamics (short-term plasticity, STP) are implemented following (Legenstein, Naeger, & Maass, 2005) using the UDF model. The arrival of a spike  $k$  after interspike interval  $\Delta_{k-1}$  induces an increase in the post-synaptic charge of amplitude  $A_k$ :

$$A_k = w \cdot u_k \cdot R_k \quad (3)$$

$$u_k = U + u_{k-1}(1 - U)e^{-\Delta_{k-1}/F} \quad (4)$$

$$R_k = 1 + (R_{k-1} - u_{k-1}R_{k-1} - 1)e^{-\Delta_{k-1}/D} \quad (5)$$

where  $w$  is the weight of the synapse (synaptic efficacy),  $u_k$  and  $R_k$  are hidden dynamic variables maintaining the facilitatory and depressionary tendencies of the short-term plasticity of the synapse, and  $U$ ,  $D$  and  $F$  are the parameters modulating synaptic use, time constant of depression (in seconds), and of facilitation. Initially,  $R_k = 1$  and  $u_k = U$ . Each spike contributes  $A_k$  to its PSR at the time it hits.

### Recurrent Circuit (Liquid)

The auditory recurrent circuit (“liquid”) is a  $15 \times 5 \times 5$  column of current-based leaky integrate-and-fire (LIF) neurons (for a total of 375 neurons; 20% are randomly chosen to be inhibitory).

For the neurons in the liquid,  $I_{bg} = 13.5$  mV uniformly and  $\tau_m = 30$  ms.  $V_{rest}$  is 0 mV. When the membrane potential of a neuron exceeds  $V_{thresh}$  (15.0 mV), the membrane potential is reset to  $V_{reset}$  (13.5 mV) and the neuron enters a refractory period during which its dynamics are frozen. For excitatory neurons this is 3 ms (inhibitory 2 ms).  $I_{syn}$  is equal to the difference of the post-synaptic responses (PSR) of excitatory afferent synapses and the PSRs of inhibitory afferent synapses.

The probability that a synapse exists between neurons at 3-D points  $a$  and  $b$  is  $C \cdot e^{(-D(a,b)/\lambda)^2}$ , where  $\lambda$  is a global parameter controlling the density of connections ( $= 2.0$ ),  $D(\cdot)$  is

the Euclidean distance function, and  $C$  is a parameter to modulate the probability of a synapse depending on properties of the connected neurons. In our case,  $C = 0.3$  if  $a$  is an excitatory neuron and  $b$  is an excitatory neuron (EE),  $C = 0.2$  for excitatory and inhibitory neurons (EI),  $C = 0.4$  for inhibitory and excitatory neurons (IE), and  $C = 0.1$  for two inhibitory neurons (II).

The parameters ( $U, D, F$ ) were selected for each synapse depending on the type of neurons that were connected and were drawn from a Gaussian distribution with means (0.5, 1.1 s, 0.05 s) for EE, (0.05, 0.125 s, 0.120 s) for EI, (0.25, 0.7 s, 0.02 s) for IE, and (0.32, 0.144 s, 0.06 s) for II (standard deviation 50% of the respective means). Negative results were redrawn from a uniform distribution between 0.001 of the mean and double the mean. The weights  $w$  of the synapses were drawn from Gamma distributions with means 30.0 (EE), 60.0 (EI), 19.0 (IE), and 19.0 (II); SD 100% of mean, with negative results redrawn from a uniform distribution as described above. In addition, a synaptic delay of 1.5 ms was implemented for EE synapses, 1.0 ms otherwise.

### Input Neurons

Raw audio streams (PCM 16 kHz) are converted into firing probabilities by a cochlear model (Slaney, 1998) which approximates the instantaneous firing activity of the auditory nerve at different points along the cochlea (“cochlear input”, gray neurons in Fig. 1). These probabilities are linearly scaled and injected as current into a set of “spike generating” LIF neurons (“spike generation”, blue neurons in Fig. 1). There are three differently parameterized classes of spike generating neurons (the three columns of neurons), for onset, offset, and passthrough. The onset spike generating neurons have a low (no firing without input) baseline firing rate ( $I_{bg} = 13.5$  mV), and receive strong positive input from the cochlear model in the form of the spike probability for that channel  $\times 20000$  nA (thus increasing activity when input is present in that channel). The offset spike generating neurons have a higher (firing without input) baseline rate affected by Gaussian noise  $I_{bg} = 13.5 + \Gamma(3.0, 0.05)$  nA, where  $\Gamma$  indicates a value drawn from a Gaussian distribution mean 4.0 and standard deviation of 0.05. The input from the corresponding cochlear channel is scaled by  $-20000$  nA, thus suppressing activity when input is present in that channel. The passthrough (direct) spike generation neurons have the same parameters as the onset spike generation neurons.  $V_{thresh} = 15.0$  mV for all these neurons.

The actual onset and offset detector neurons (green input neurons in Fig. 1) receive dynamic synapses from the surrounding three spike generation channels of their corresponding class of spike generating neurons. These synapses modulate the current injected into the post-synaptic neuron based on pre-synaptic firing activity. Large pre-synaptic activity will cause an initial facilitation, followed by a longer depression in the strength of injected post-synaptic current per action potential. Thus, they will inject strong current for the first few pre-synaptic spikes, followed by less current for a period thereafter. This, combined with the different baseline



firing rates of the spike generators, is what implements the onset/offset detectors. The passthrough neurons have quickly-recovering dynamic synapses and perform more like static synapses, but limit their firing rate to a slower rhythm.

$V_{thresh}$  for each sensitivity level of onset/offset/passthrough detector is:

$$V_{thresh} = V_{reset} + E_0 \cdot (D^i \cdot (c + 1)) \quad (6)$$

where  $E_0 = 1.0$  for onset/offset detectors and  $E_0 = 0.2$  for passthrough neurons.  $D = 1.414$ , with  $i$  from  $c = 0$  to  $c = N$  for each of the  $N$  sensitivities of onset/offset detectors ( $N = 1$  for the experiments, i.e. only one sensitivity level for onset/offset detectors). For the one passthrough level,  $D$  is scaled by a factor of 9.0.

For the dynamic synapses between the spike generation and the onset/offset/passthrough neurons, the UDF parameters are (0.5, 1.1, 0.05) and  $w = 3.0$  (onset), (0.5, 0.025, 0.5) and  $w = 9.0$  (offset), and (0.5, 0.025, 0.5) and  $w = 9.0$  (passthrough).

Input (offset/onset/passthrough) neurons synapse into a randomly selected 30% of circuit neurons via static synapses. The weight  $A$  of each of these input synapses is drawn from a Gamma ( $shape = 1$ ) distribution with mean  $A_{mean} = 18.0$  when the post-synaptic neuron is excitatory and  $A_{mean} = 9.0$  when it is inhibitory. Negative weights are set appropriately from a uniform distribution between  $0.001 \cdot A_{mean}$  and  $2 \cdot A_{mean}$ .

The cochlear model described in (Slaney, 1998)<sup>1</sup> was modified and updated to run in real time. Parameter defaults are retained (except for:  $breakf = 500$ ,  $qconst = 8.0$ ,  $stepfactor = 0.5$ ,  $sharpness = 5.0$ ,  $notchoffset = 1.5$ ,  $preemphfreq = 300$ ,  $taufactor = 3.0$ ) producing 52 output channels which encode on each simulation step the probability that a spike occurs in that channel on that time step.

## Readout Neurons and Phrase Integration

A final set of neurons (readout neurons) serve as classifiers ( $r_n$ , one for each category  $n$ ). They receive as input a weighted projection of the liquid's instantaneous firing activity (if spiked +1, otherwise -1), low-pass filtered to mimic the change in post-synaptic membrane potential had the readout been modeled as an LIF neuron (time constant 30 ms). A readout neuron is said to fire at a given time point if the sum of its inputs exceeds a threshold (the bias term determined by the linear regression below). The shape of the weighted projection is determined by supervised learning on a training corpus for each readout neuron independently. This is achieved by linear regression of the matrix of the liquid response to all stimuli (with an additional bias column which is always -1), with a supervisor vector which contains +1 for every time point during which input was from the target word class, and -1 otherwise.

Phrase Integration is performed by injecting 1.0 nA per  $r_n$  spike into a corresponding readout integration neuron  $i_n$  (with

membrane time constant  $\tau_m = 50$  ms). The readout integration neuron is considered to be active when  $V_m > 0.25$ , with no reset or refractory period. These readout integrators provides a more continuous picture of the readout activity that is robust to small recognition errors.

Utterance onset and offset detection (modeled as a neuron with membrane potential  $V_{utter}$ ) is performed by combining input from the onset/offset/passthrough input neurons with the current readout firing activity. Each spike of an onset/offset/passthrough neuron imparts directly 1.0/-0.5/0.7 mV to  $V_{utter}$ , which decays exponentially with  $\tau = 200$  ms. For a word onset to be detected,  $V_{utter} > 1.0$ , and the highest value of all  $n$ ,  $i_n > 0.25$ . An offset occurs when either of these variables falls below the threshold. On onset, an accumulator neuron  $a_l$  increases its voltage at a constant rate of 0.002 mV/ms, with a threshold of 0.3 mV. An utterance is only considered to be an instance of a category (i.e. not noise) if  $a_l$  is over threshold. The accumulator is reset to 0 mV at offset. Another set of accumulators  $a_n$  (one for each readout integrator  $i_n$ ) sums the value of its corresponding readout integrator  $i_n$  from the point an onset is detected, and are reset at offset.

When an utterance that meets all the above conditions is detected, its category is determined by dividing each readout accumulator by the length accumulator. If the highest value is greater than the  $i_n$  threshold (0.25), the utterance is classified as being an instance of the winning category  $n$ . A token (e.g. a textual representation of it, or a number) indicating that category is returned.

## Experimental Validation, Analysis and Results

The model was tested on part of the "CReST corpus" developed from human-human interactions in a search task (K. Eberhard, Nicholson, Kuebler, Gundersen, & Scheutz, 2010) (<http://www.cs.indiana.edu/~riveale/hricorpora.html>). The liquid was trained 10 times each on 9 audio samples each for each of ten phrase categories. In addition, it was "counter-trained" on a sample of recorded microphone noise (that portion of the supervisor vector for all categories was -1 for all time-points for the linear regression). The final two audio samples for each category were set aside for testing.

To test phrase recognition, a new audio stream was created by concatenating the 10 test samples which were presented to the model in real-time. The returned categories were verified to be recognized only once during the correct portion of the audio stream. The best liquid was able to correctly recognize all ten phrase categories. 50 liquids were randomly generated and trained. All were able to recognize at least 7 of the categories, with the exception of 3 liquids that only recognized 5 of them (recall that liquids are randomly generated). By observing the readout behavior during recognition, it was determined that the primary reason for failure was similarities in the categories (e.g. it was most likely to fail when phrases shared similar words, or words had large regions that were similar-sounding).

<sup>1</sup>Code from <http://www.slaney.org/malcolm/pubs.html>

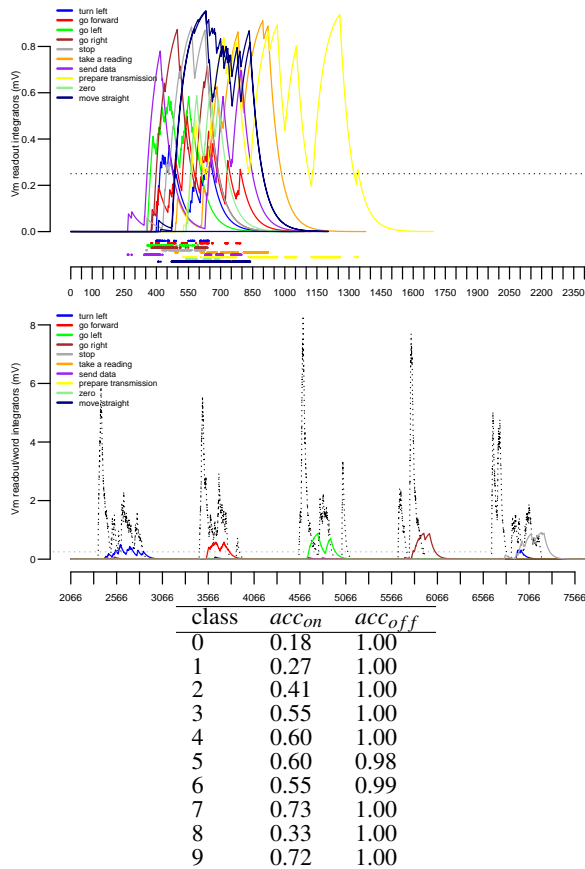


Figure 2: *Top*: Readout (spike) and integrator (lines) response of each trained category from the best liquid to a test case of its own class. *Center*: Readout integrator responses of 5 phrases uttered sequentially, along with scaled utterance detector. Horizontal line is threshold. Note each “correct” readout is active during its own category and relatively inactive otherwise. *Bottom*: Proportion of correctly spiking and correctly not spiking time points for each trained readout (corrected for large leading/trailing silence).

Figure 2 shows the time-line of phrase recognition over the course of the test speech stream in two different ways (including the raw spiking characteristics of the best liquid for the auditory corpora used). Note that even *during* an utterance, it can be predicted based on readout integrator levels whether the utterance stream is probably of a given class. This information can be used to prepare (pre-cache or “prime”) responses to that phrase.

The model has also recently been used for speech recognition embedded in a larger cognitive model of embodied situated human natural language interactions in human-robot interactions tasks, thus providing evidence for its utility in real environments (video at <http://www.cs.indiana.edu/~riveale/muridemo.html>). In this task, the model played the role of a speech recognizer that sends processed tokens representing whole phrases to a natural language parser and understanding system. This system processed the meaning of

the phrase, and passed it on to a cognitive architecture (planner, etc.) which was able to initiate actions (such as changing its own goal state, change its behavior, learn new actions such as door-pushing) based on the content of the phrases (Cantrell, Talamadupula, Schermerhorn, Benton, & Scheutz, 2012).

## Discussion

While previous research has presented methods for converting spike-encoded streams (generated offline from audio input) into liquid activity and, subsequently, liquid activity into instantaneous category readout firing, the problem of using readout firing activity to generate category-tokens of the kind used by most higher-level cognitive models has not been sufficiently addressed. Moreover, the method of encoding sound as spikes did not take into account the onset/offset-detection capabilities of the human auditory system, which contributes significantly to the robust recognition of shorter phonemes like consonants. The model presented in this paper addresses both shortcomings by processing sound in a realistic way up to the neural readout-level and being capable of returning correct word tokens based on readout activity at the right time. One limitation of the current model that must be addressed in the future is that some aspects of the token conversion (the conversion from raw spiking activity to phrase-tokens) are not guaranteed to work well in all speech situations. For example, the assumption that phrases will be preceded and followed by silence may not hold in situations where speech is very fast. In those situations recognizing phrases based on other information such as prosody could lead to better results (Christiansen, Allen, & Seidenberg, 1998).

In addition to the practical benefits of having a real-time model that shows promising performance on natural speech, the model makes important predictions about the mechanisms by which humans are capable of extracting discrete category information from a continuous real-time stream of sensory data. For example, the model proposed an explanation for how category priming or biasing effects come about, as the ongoing activity of each readout integrator can be viewed as the probability that the category it represents is currently present in the stimulus stream. This probability can be made to bias ongoing behavior even before the stimulus signal ends. To see this, imagine an experimental paradigm (e.g., the visual world paradigm (K. M. Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995)) in which an auditory cue (e.g., a color word) determines to which of two locations the subject should direct her eyes – it is well-known that humans in that case are capable of performing eye saccades shortly after the onset of the color word (even though they will not always perform them right away). For computational models of these human behaviors it means that a decision for eye movements has to be reached before the word stimulus ends. Specifically, if there are two category readouts, one for “Red” and one for “Blue”, and the first phoneme of “Red” is presented, the model could use the already high activity of the “Red”

category to bias its looking behavior even before the stimulus has ended and the category word “Red” is fully recognized.

This example demonstrates that there can be multiple routes from the “readout” level to other parts of the cognitive system that can influence and bias behavior (one being the route where recognized word tokens are passed on to higher syntactic and semantic processing areas, while another being the more direct, intermediate route that can bias looking behavior). The current model is silent about the exact nature of these routes, as this would require additional specifications of those higher-level cognitive components and behavior-generating components, an important direction for future work. The current model is also silent about the important top-down biases coming from various other parts of the human cognitive system such as the syntactic and semantic biases as well as other biasing information based on perceptual, dialogue, task, and goal information. These top-down biases are critically involved in human speech processing and contribute to the robustness of human speech recognition in noisy environments. However, we believe that the particular model architecture will directly allow for this kind of information integration by way of appropriate top-down connections to the readout units whose activations represent the dynamically changing hypotheses about recognized words. We are currently investigating extended versions of the model that include additional perceptual areas (e.g., as described in (Veale, Schermerhorn, & Scheutz, 2011)) to test the extent to which perceptual biases can influence recognition rates and behavior (such as eye saccades).

Different from our previous models (Scheutz, Eberhard, & Andronache, 2004) which only at a high level of abstraction resembled the human cognitive architecture and only modeled the human data qualitatively, the goal for these new models is to be fully realized in neural architectures and to model the human data quantitatively.

## Conclusions

In this paper, we presented a novel biologically plausible model for human speech recognition together with a proof-of-concept implementation and evaluation of the model. As part of the model, we introduced new methods for any-time phrase recognition based on the human early auditory system coupled with biologically plausible methods to produce word category tokens from a continuous auditory stream. We also introduced biologically plausible implementations of onset/offset detectors for word signals. Future work will include the already mentioned extensions by perceptual areas to be able to allow for biologically plausible quantitative neural models of human eye gaze behavior during reference resolution. In addition, we will also investigate mechanisms for improving noise-robustness with multiple sensitivity levels at the onset/offset stage and online learning of novel word categories.

## Acknowledgments

RV is an NSF IGERT and NSF Graduate Research Fellow.

## References

- Cantrell, R., Talamadupula, K., Schermerhorn, P., Benton, J., & Scheutz, S. K. M. (2012, March). Tell me when and why to do it! Run-time planner model updates via natural language instruction. In *Proceedings of the 2012 human-robot interaction conference* (p. forthcoming).
- Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13(2/3), 221-268.
- Eberhard, K., Nicholson, H., Kuebler, S., Gundersen, S., & Scheutz, M. (2010). The indiana cooperative remote search task (crest) corpus. In *Proceedings of Irec 2010: Language resources and evaluation conference*. Malta.
- Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24, 409-436.
- Ghitza, O. (1987). Auditory nerve representation criteria for speech analysis/synthesis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(6), 736-740.
- Legenstein, R., Naeger, C., & Maass, W. (2005). What can a neuron learn with spike-timing-dependent plasticity? *Neural Comput.*, 17(11), 2337-2382.
- Lyon, R. F. (1982, May). A computational model of filtering, detection, and compression in the cochlea. In *Acoustics, speech, and signal processing, ieee international conference on icassp '82* (p. 1282-1285).
- Maass, W., Natschlager, T., & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11), 2531-2560.
- MacLeod, K. M., Horiuchi, T. K., & Carr, C. E. (2007). A role for short-term synaptic facilitation and depression in the processing of intensity information in the auditory brain stem. *Journal of Neurophysiology*, 97, 2863-2874.
- Scheutz, M., Eberhard, K., & Andronache, V. (2004). A real-time robotic model of human reference resolution using visual constraints. *Connection Science Journal*, 16(3), 145-167.
- Slaney, M. (1998). *Lyon's cochlear model* (Tech. Rep. No. 13). Apple Computer Inc. Cupertino, Ca.
- Smith, L. S., & Faser, D. S. (2004). Robust sound onset detection using leaky integrate and fire neurons with depressing synapses. *IEEE Transactions on Neural Networks*, 15(5), 1125-1134.
- Veale, R., Schermerhorn, P., & Scheutz, M. (2011). Temporal, environmental, and social constraints of word-referent learning in young infants: A neurobotic model of multimodal habituation. *IEEE Transactions on Autonomous Mental Development*, 3(2), 129-145.
- Verstraeten, D., Schrauwen, B., & Stroobandt, D. (2005). Isolated word recognition using a liquid state machine. In *Esann* (p. 435-440).

# An Integrated Model of Associative and Reinforcement Learning

Vladislav D. Veksler (vdv718@gmail.com)

Christopher W. Myers (christopher.myers.29@us.af.mil)

Kevin A. Gluck (kevin.gluck@wpafb.af.mil)

Air Force Research Laboratory

Wright-Patterson AFB, USA

## Abstract

Any successful attempt at explaining and replicating the complexity and generality of human and animal learning will require the integration of a variety of learning mechanisms. Here we introduce a computational model which integrates associative learning and reinforcement learning. We contrast the integrated model with associative learning and reinforcement learning models in two simulation studies. The first simulation demonstrates performance advantages for the integrated model in an environment with a dynamic and diverse reward structure. The second simulation contrasts the performances of the three models in a classic latent learning experiment (Blodgett, 1929), demonstrating advantages for the integrated model in predicting and explaining the behavioral data.

**Keywords:** Associative Learning, Reinforcement Learning, Model Integration, Cognitive Modeling, Cognitive Systems, Latent Learning

## Introduction

Integration of computational cognitive models is critical for accelerating progress in the field of cognitive modeling (Gray, 2007a). By means of integrative approaches the field can begin to predict and explain the robustness and flexibility of human behavior in complex, uncertain, non-stationary environments (or *large worlds*, Binmore, 2009). Specifically, Choi & Ohlsson (2011) assert that the integration of *learning mechanisms* is essential to improving the predictability and explanatory power of cognitive models.

There is no denying that people are adaptive – for example our memories are retrieved based on their recency and frequency of use (Anderson & Schooler, 1991), visual search is adapted to the structure of the task environment (Shen, Reingold, & Pomplun, 2000; Myers & Gray, 2010), and problem solving strategies are adapted with increasing task experience (Siegler & Stern, 1998). People’s ability to adapt allows them to persist and thrive in large worlds. If we are to build cognitive models for large worlds, we have to endow them with human learning mechanisms. Hand-coded knowledge engineering results in brittle and expensive models, and is a method that does not scale well beyond simple laboratory environments (Gluck, 2010). Our hypothesis is that models may begin to demonstrate human-like flexibility and adaptivity in large worlds through the integration of multiple human learning mechanisms.

In the current paper we present an integrated model of associative and reinforcement learning, as it is evident that humans are capable of learning both the spatiotemporal contingencies and the reward structures of their environment

(Stevenson, 1954; Chun, 2000; Myers, Gray, & Sims, 2012). We demonstrate that model integration improves flexibility and adaptability, provides better predictions of behavioral data, and produces more efficient behavior in environments with diverse and dynamic reward structures when compared to each of the individual models.

In the following sections we first provide background on associative and reinforcement learning theories and models. Next we describe the integrated model. Finally, two simulations are presented. Simulation 1 contrasts the associative and reinforcement learning models with the integrated model in their ability to efficiently adjust to novel goals and diverse reward structures in a grid-navigation environment. Simulation 2 contrasts the associative and reinforcement learning models with the integrated model in their ability to predict behavioral data from a classic latent learning experiment.

## Reinforcement Learning

Reinforcement learning (RL) is a formal model of action selection where the utility of different actions is learned by attending to the reward structure of the environment. It has been used in a wide array of domains, from robotics (Peters, Vijayakumar, & Schaal, 2003) and artificial intelligence (Russell & Norvig, 1995) to cognitive architectures (Fu & Anderson, 2006; Nason & Laird, 2005) and cognitive neuroscience (Holroyd & Coles, 2002).

Generally speaking, RL works in a trial-and-error fashion – attempting various actions and recording the reward gained for those actions (for a review see Sutton & Barto, 1998). More formally, given the state that an agent is experiencing, the action most likely to be chosen is the one with the highest learned utility, plus or minus some exploratory noise. The utility of any given state-action pair, *SA*, in turn, is directly proportional to the value of the reward, that the agent receives after *SA* is executed. Hence, state-action pairs are *reinforced* when they result in a reward; and the likelihoods of their future selection are directly proportional to the values of the experienced rewards.

There are several variations on how utility is learned in RL (for an introduction, see Sutton & Barto, 1998). For example, Temporal Difference RL (TDRL), a version commonly used to model human behavior (Anderson, 2007; Holroyd & Coles, 2002), propagates the received reward to past actions. Reward is discounted as a function of time, so that actions taken just prior to the reward are strengthened more than earlier actions. In this way, TDRL reinforces a sequence of actions that lead to the reward, rather than just a single state-

action pair, helping to obtain a solution in a more efficient manner.

Some RL approaches take into account transitions between  $SA$ , the resultant reward  $R$ , and the utility of the next  $SA$  (SARSA). SARSA models update the utility of the state-action pair executed at time  $t$ ,  $SA(t)$ , by a function of the reward that follows it,  $R(t+1)$ , combined with a function of the utility of the state-action pair that follows it,  $SA(t+1)$ . SARSA models are not as efficient as TDRL, but are guaranteed to converge on an optimal solution.

The Model-based RL approach extends RL by learning the structure of the world beyond utilities. The term *model* in “model-based RL” refers to an agent’s internal representation of the environment, and an agent developed in this framework is capable of planning its route before execution. This is extremely useful when memory and decision cycles are less expensive than actions (e.g. robotics).

One of the limitations of RL as a complete model of human decision-making becomes apparent in environments where goals change. Imagine that on your way to work each day you pass a post office. One day you need to mail a letter. At this point, an RL agent would consider, “let’s try a random action, see how that works.” This is because, by definition, RL models make decisions based solely on the learned state-action utilities. If the goal changes, the utilities representing the reward structure from the initial goal become irrelevant at best, or subversive at worst. Humans and animals, of course, will employ their knowledge of the environment (e.g. that there is a post office on the way to work) to make better-than-chance decisions for achieving new goals (Stevenson, 1954; Tolman, 1948; Quartermain & Scott, 1960).

The SARSA and Model-based approaches are major steps toward more flexible behavior. The SARSA approach considers the state-action-state transitions when learning utilities, but stops short of learning these transitions. The Model-based RL approach learns such transitions, but employs them strictly to enable planning. The decision process during the planning stage, however, is still based on the learned utilities. Thus, when presented with a new goal a Model-based RL agent will still begin to plan its route by considering random actions.

## Associative Learning

Another class of decision models relies on associative learning. Associative learning (AL) models focus on acquiring the spatiotemporal contingencies of the environment and employing these in action-selection. The utility of any given choice is estimated as a function of previously experienced spatiotemporal proximity between this choice and the current goal. The advantage of this approach over RL is that the stored knowledge is goal-independent. Whenever a new goal is given, an AL model can employ its knowledge to make informed goal-directed decisions.

Voicu and Schmajuk (2002) implemented a computational model that learns the structure of the environment as a network of adjacent cells. Once a goal is introduced, reward signal spreads from the goal-cell through this network, such that the cells farther from the goal-cell receive less activation than those that are close. Goal-driven behavior in this model comprises moving towards the cells with the highest activa-

tion. Once this model memorizes the map of the environment, it does not need to learn the reward structure through trial-and-error; rather, the utility of each action-path is identified through spreading activation from the goal.

SNIF-ACT (Fu & Pirolli, 2007) is another model that employs associative rather than reward knowledge for action-selection. SNIF-ACT is a model of human information-seeking behavior on the World Wide Web. The World Wide Web is unpredictable in the sense that there is no way for any of its users to know what links they will encounter during web browsing. The utility of selecting a link in SNIF-ACT is not based on any prior reward, but rather on the semantic association of a link’s text to the current goal (i.e., *information scent*). This mechanism allows SNIF-ACT to make non-random decisions in novel situations based on associative knowledge.

A limitation of SNIF-ACT is that it does not learn the association strengths between links and goals, but rather imports these values from an external source. The Voicu & Schmajuk model learns association strengths in a psychologically implausible manner. The Goal-Proximity Decision-making model (GPD; Veksler, Gray, & Schoelles, 2009) mends this by employing the psychologically-plausible delta learning rule (Rescorla & Wagner, 1972; Widrow & Hoff, 1960) to update association strengths. Like the other two models, GPD then estimates the utility of a path based on its association strength to the current goal. Veksler, Gray, & Schoelles demonstrate that in an environment where goals continue to change, GPD is able to replicate human performance and RL cannot.

A limitation of AL models is that no reward information is learned. In this class of models decisions are based on explicitly specified goals. Associative learning does not help to understand a diverse reward structure, where some actions may result in less reward and some in greater reward. Hence, AL models cannot explain why an organism might learn to prefer actions leading to one goal-state over another.

## Integrating Associative and Reinforcement Learning

It is our opinion that AL and RL complement each other. As discussed above, RL models capture behavior based on a given reward structure. However, as agent goals change, so does the reward structure of the environment. Since RL fails to capture environmental contingencies beyond the original reward structure, it cannot predict the efficiency of human behavior in environments where goals tend to change. Contrariwise, AL models store the the spatiotemporal contingencies of the environment independent of the reward structure, and are more flexible in adapting to new goals. However, in ignoring the reward information in the environment, association-based models cannot capture people’s sensitivity to the value of reward. In this section we describe how the two learning approaches can be integrated to produce more flexible behavior in environments where the reward structure is both diverse and dynamic.

Given some agent state,  $S$ , and a possible action,  $A$ , RL models learn the utility,  $u$ , of the  $SA$  state-action pair as directly proportional to the reward that has been experienced after prior executions of  $SA$ , and, in models like TDRL, inversely proportional to the length of time between  $SA$  and the

reward in prior experience. AL models do not learn the utility of  $SA$ , but estimate it based on the strength of association,  $w$ , between  $SA$  and the current goal,  $G$ , where  $w$  is inversely proportional to the length of time (or distance) between  $SA$  and  $G$  in prior experience. From the perspective of what is stored in model memory, the RL models store the values of  $u$  for each state-action pair,  $SAu$ , and AL models store the values of  $w$  for each state-action-state transition,  $SAwS$ .

To integrate these two models, we propose that the association strength,  $w$  should continue to be recorded as in the AL models, whereas the utility  $u$  should be recorded for each state,  $S$ , rather than for each state-action pair  $SA$ . Thus, what will be stored in memory and used for action-selection in the integrated model is both  $w$  and  $u$  for each state-action-state transition,  $SAwSu$ . The strength of association,  $w$ , is useful as an estimate of the probability that a state might follow a given state-action pair and the length of time of this transition. The utility,  $u$ , is useful as an estimate of the reward probability/value to be received after a transition occurs.

The integrated model uses the delta learning rule to update both utilities and association strengths. For each previously executed state-action pair  $j$  and each new state  $i$ , the strength of association between  $j$  and  $i$ ,  $w_{ji}$ , at current time,  $n$ , is increased in the following manner:

$$\Delta w_{ji}(n) = \beta[a_i(n) - w_{ji}(n-1)] \quad (1)$$

where  $\beta$  is the learning rate parameter, and  $a_i$  is the activation of  $i$  ( $a_i = 1$  if  $i$  is present, else 0). The utility for each new state  $i$ ,  $u_i$ , at current time,  $n$ , is increased in the following manner:

$$\Delta u_i(n) = \alpha[r(n) - u_i(n-1)] \quad (2)$$

where  $\alpha$  is the learning rate parameter, and  $r(n)$  is the reward experienced at time  $n$ .

At each decision point, the utility of a given state-action pair,  $j$ , is calculated as follows:

$$U_j = \sum_{\forall i} (w_{ji} \times u_i \times \delta^t) + N \quad (3)$$

where  $\delta$  is a discount parameter ( $0 < \delta < 1$ ),  $t$  is the temporal distance between  $j$  and  $i$ , and  $N$  (exploratory noise) is a number drawn randomly from a normal distribution with a mean of zero and a standard deviation set to some parameter,  $\sigma$ .

In the following section the  $SAwSu$  model is examined in terms of efficiency and psychological validity within environments with diverse and dynamic reward structures.

## Simulations

The following subsections compare the integrated AL+RL model ( $SAwSu$ ) with AL-only and RL-only models. First, the models are evaluated based on the efficiency of finding rewarding states in a  $10 \times 10$  grid. Second, the models are evaluated based on the ability to match data from a classic latent learning experiment. A single value for the discount parameter ( $\epsilon = .9$ ) was used for both simulations.

### Simulation 1: Dynamic & Diverse Reward Structure

As pointed out in the Introduction, the strength of RL is in learning a diverse reward structure, where some actions may lead to greater reward than others; AL excels at learning the environmental structure independent of rewards, such that this knowledge may be applied in purposive behavior whenever new goals arise. However, large worlds are both diverse and dynamic. The following simulation was conducted to *highlight the conditions under which the RL and AL approaches begin to falter*, and how an integrated approach addresses these limitations.

A  $10 \times 10$  navigation grid was used, where a model's state was uniquely identified as one of the cells in the grid, and the model had four possible actions from each cell – to move north, south, east, or west<sup>1</sup>. If an illegal move was selected (i.e. a move that would take the model off the grid), the model's state was not changed. For each model run, a model was placed in a random cell on the grid. Each time the model reached a reward state, the model would again be placed in a random cell on the grid. Before the model began the task of locating a reward, a reward of 1.0 was placed in a random cell on the grid. After 4000 steps, the reward was cleared, and placed in a different random cell. Following the next 4000 steps (8,000 total steps), the reward was cleared, and rewards of 1.0 and 0.1 were placed in two randomly selected cells. Finally, after the next 4000 steps (12,000 total steps), the reward was cleared again, and replaced with rewards of 1.0 and 0.1 in two randomly selected cells.

The integrated AL+RL model ( $SAwSu$ ) was compared with Random-walk, AL (GPD), and two RL models – Temporal-Difference RL (TDRL), and Q-learning (Q-RL). As discussed above, TDRL is the version of RL most commonly used in modeling human/animal behavior. Q-RL is a popular SARSA model that is not as efficient as TDRL, but is guaranteed to converge on an optimal solution (Sutton & Barto, 1998). The results, averaged over 100 model runs for each model type, may be observed Figure 1.

AL+RL was the best overall model, averaging a total score of 1328.4 for the entire run in this environment, whereas AL (GPD), TDRL, Q-RL, and Random models scored 1107.8, 362.2, 237.1, and 66.1, respectively. Q-RL is guaranteed to converge to an optimal solution for any one reward structure in the environment, but it is too inefficient to find such a solution within the 4000 trials allotted in this task (though its efficiency improves after the first 8000 steps, where there is more than one goal-state).

TDRL, AL, and AL+RL produce indistinguishable performance until the first goal change (see Figure 1, first 4000 steps). However, once a new goal is presented, TDRL struggles to relearn the reward-structure of the environment, as all of the state-action utility values need to be relearned (these may be relearned faster if the exploratory noise was increased, but this would come at the expense of performance, even for the first goal). Q-RL struggles with the same issue, as both RL models drop to random-level performance once a new reward-structure is introduced. In contrast, AL and

<sup>1</sup>This is a standard simulation environment for performance examination of computational agents, as it aims to represent a generic problem space.

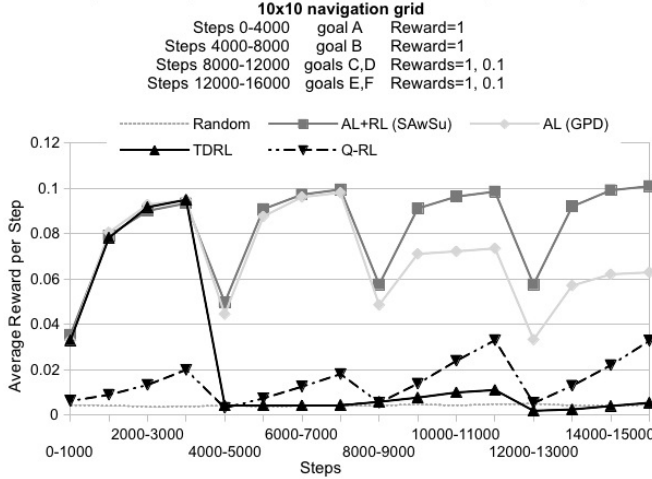


Figure 1: Simulation 1 results.

AL+RL can employ all of the associative knowledge that was gathered in the first 4000 steps, and apply it to achieving the new goal.

Where AL+RL begins to differ from AL is when the reward structure of the environment becomes more diverse. After 8000 steps, there are two rewarding states introduced into the environment, one of these having a high value (1.0) and the other having a low value (0.1). The AL+RL model learns the correct reward values of these states. AL, however, cannot distinguish between the two types of goals, as it records no information corresponding to varying reward values.

In summary, the integration of associative and reinforcement learning results in better performance than could be achieved by either model alone in an environment where the reward structure is dynamic and diverse. The AL+RL model displays more flexibility than RL in adapting to changing goals, and more flexibility than AL in adapting to a varying reward structure.

## Simulation 2: Blodgett, 1929

Latent learning is a classic behavioral paradigm that focuses on performance in an environment with a dynamic reward structure, and often involves a diverse reward structure. In this paradigm, after having spent some time in an environment, subjects are presented with some goal. Upon the introduction of the goal, subjects display a higher level of performance than would be expected if they had not spent any time in the environment prior to the goal introduction. This phenomenon is observable in children, adults, and animals (e.g. Quartermain & Scott, 1960; Stevenson, 1954; Tolman, 1948).

For example, Blodgett (1929) ran three groups of rats in a maze-learning experiment. One group (the control) was rewarded upon reaching the end of the maze on every trial (R1). The second group began receiving rewards on trial 3 (R3). The third group began receiving rewards on trial 7 (R7). Results demonstrate that subjects in groups R3 and R7 began to perform at the level of control subjects immediately upon the introduction of the reward, producing much steeper error-reduction slopes in these groups than that of R1 (see Figure 2, top-left panel). An associative learning model can predict this

phenomenon. Such a model would learn the structure of the maze and begin to employ its knowledge immediately once the reward is introduced.

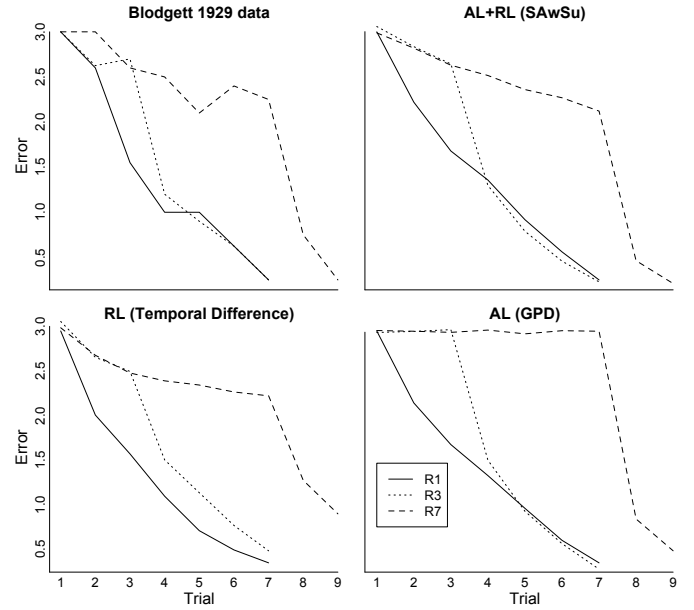


Figure 2: Maze Performance: Avg. Errors by Trial. Data adapted from Blodgett, 1929 (top-left) and simulation results from Reinforcement Learning (bottom-left), Associative Learning (bottom-right), and integrated (top-right) models.

Interestingly, groups R3 and R7 did not continue to display random-level performance until the introduction of the reward. Rather, these groups displayed a shallow error-reduction curve, indicating that there was at least some intention to complete the maze even in the “no-reward” trials (“low-reward” from hereon)<sup>2</sup>. An RL model can predict this phenomenon, producing a shallow learning curve for the “low-reward” trials (R3 until trial 3, R7 until trial 7), and a steeper learning curve for the high-reward trials (R1).

A model that integrates RL and AL should reproduce both (1) the better-than-random level of performance in groups R3 and R7 prior to the introduction of reward, and (2) the steep improvements in performance once this reward is introduced.

The integrated AL+RL model (SAwSu) was compared with AL (GPD) and RL (TDRL) models. A parameter search was performed, seeking the model parameters that produced the best fit (least sums of square differences) to data in the constant-reward (R1) and the “low-reward” (R7, trials 1-7) conditions. Three parameters were varied for each model: learning rate, amount of exploratory noise ( $\sigma$ ), and the perceived low-reward ( $LowR$ ) for finishing the maze on the “low-reward” trials. The learning rate parameter varied for RL was the utility-learning constant,  $\alpha$ , and for AL and AL+RL it was the associative-learning constant,  $\beta$  ( $\alpha$  remained unvaried for AL+RL at 1.0).

Once the best parameter values were found (RL:  $\alpha = .4, \sigma = .08, LowR = .15$ ; AL and AL+RL:  $\beta = .2, \sigma =$

<sup>2</sup>We interpret the shallow learning curves as resulting from a low reward, such as being taken out of the maze.



Table 1: Root Mean Square Difference to Blodgett, 1929.

Model	Best fit to data		Predicted	
	R1	R7 [trials 1-7]	R3	R7*
AL+RL	0.21	0.14	0.11	0.15
RL	0.26	0.17	0.19	0.32
AL	0.22	0.56	0.23	0.50

\*Only trials 8 and 9 are predicted.

.05,  $LowR = .15$ ), the full simulations were executed to get model predictions for R3 and for R7 after the introduction of reward (these conditions were not included during the parameter search). Results may be observed in Figure 2 and Table 1. As expected, AL and AL+RL produced steeper performance improvements than RL upon the introduction of the reward by the experimenter on trials 3 and 7. As expected, RL and AL+RL replicated the shallow error-reduction curves in trials 1-3 for condition R3 and 1-7 for condition R7, and AL did not.

AL+RL produced a better overall fit to data than did the other two models (see Table 1). The advantages become more apparent when we focus on the error-reduction after the introduction of reward. Figure 3 demonstrates model predictions for error reduction in the R3 group between trials 3 and 5, and the R7 group between trials 7 and 9. The AL model predicts too high a performance improvement (because the initial performance is underestimated), and the RL model predicts too low a performance improvement in these trials.

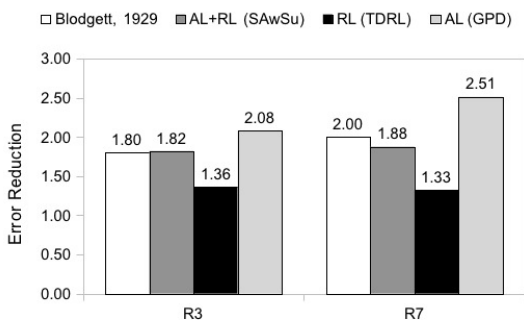


Figure 3: Error reduction after the introduction of reward in Blodgett, 1929.

## Summary and Discussion

In this paper we described how two learning mechanisms widely supported in the psychological literature, reinforcement and associative learning, may be integrated. In contrast with RL-only and AL-only models, the integrated model, SAwSu, was shown to produce more efficient, higher fidelity behavior in environments where the reward structure is both diverse and dynamic.

Gläscher, Daw, Dayan, and O'Doherty (2010) propose an alternative integration of AL and RL by including a supervisory mechanism that learns to arbitrate between AL and RL. This implementation seems less parsimonious than SAwSu –

it has three learning and three decision mechanisms, whereas SAwSu has two and one, respectively. Further comparison of the two approaches is warranted.

To the best of our knowledge there are no other computational frameworks that learn the reward structure and the spatiotemporal predictions of the environment, and employ both in the decision-making process. Frameworks that employ some form of Model-based planning (e.g. Daw, Niv, & Dayan, 2005; Sutton & Barto, 1998) include both AL and RL, but these tend to focus on the trade-off between planning in the head and acting in the world. Associative knowledge in this class of models is used to enable planning rather than to determine how a path of actions, whether in the head or in the world, is chosen.

The overall scarcity of decision models that employ AL and RL together is rather surprising given the long history of research on learning in experimental psychology, cognitive science, and artificial intelligence. Ohlsson (e.g. Choi & Ohlsson, 2011) has been promoting the integration of learning mechanisms, including AL and RL, and Alonso & Mondragón (2006) and Dickinson & Balleine (1993, 1994) call for AL+RL integration. None of these proposals, however, has been implemented as a computational model, and thus cannot be easily contrasted with the SAwSu implementation.

The Voicu & Schmajuk (2002) model mentioned in the Introduction, does employ AL in action-selection, and even considers variable utility of the goal state in the decision phase. However, the Voicu & Schmajuk model does not specify any way of actually learning state utilities.

Earlier versions of the ACT-R integrated cognitive architecture included both RL and AL (see Anderson, 1993; Anderson & Lebiere, 1998). However, according to Anderson (2001), the particular form of associative learning implemented in ACT-R turned out to be “disastrous,” and produced “all sorts of unwanted side effects” (p. 6). Thus, as it stands, the implementation of associative learning in ACT-R 6 has been reduced to a single equation that relates the fan effect to spreading activation. This limits AL to chunks that have a direct symbolic relationship, where associative strengths can only decrease as more knowledge enters the system and the “fan” of associations to each chunk increases.

The current effort to integrate AL and RL is in accord with the many calls for the integration of cognitive mechanisms within a unified computational framework (e.g. Gray, 2007b; Choi & Ohlsson, 2011). However, the current work presents the integration of only two learning mechanisms, addressing only some of the complexities of large worlds. In the pursuit of models that can produce persistent, adaptive, and flexible behavior in large worlds, it is required that we address how a model like SAwSu might be incorporated into a broader cognitive architecture such as ACT-R. Further integration of AL and RL with other cognitive mechanisms is the necessary next step for this research.

## Acknowledgements

This research was performed while the author held a National Research Council Research Associateship Award with the Air Force Research Laboratory's Cognitive Models and Agents Branch.

## References

- Alonso, E., & Mondragón, E. (2006). Associative Learning for Reinforcement Learning: where animal learning and machine learning meet. In *Proceedings of the 5th symposium on adaptive agents and multi-agent systems*.
- Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Anderson, J. R. (2001). Activation, Latency, and the Fan Effect. In *Eighth annual act-r workshop*. Pittsburgh, PA.
- Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Anderson, J. R., & Schooler, L. J. (1991). Reflections of the Environment in Memory. *Psychological Science*, 2(6), 396–408.
- Binmore, K. (2009). *Rational Decisions*. Princeton University Press.
- Blodgett, H. C. (1929). The effect of the introduction of reward upon the maze performance of rats. *University of California Publications in Psychology*.
- Choi, D., & Ohlsson, S. (2011). Effects of multiple learning mechanisms in a cognitive architecture. In *Proceedings of the thirty-third annual meeting of the cognitive science society*.
- Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, 4(5), 170–178.
- Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12), 1704–1711.
- Dickinson, A., & Balleine, B. (1993). Actions and responses: The dual psychology of behaviour. Spatial representation: Problems in philosophy and psychology. In N. Eilan, R. McCarthy, & B. Brewer (Eds.), *Spatial representation: Problems in philosophy and psychology*. (pp. 277–293). Oxford University Press.
- Dickinson, A., & Balleine, B. (1994, March). Motivational control of goal-directed action. *Animal Learning & Behavior*, 22(1), 1–18.
- Fu, W. T., & Anderson, J. R. (2006). From recurrent choice to skilled learning: A reinforcement learning model. *Journal of Experimental Psychology: General*, 135(2), 184–206.
- Fu, W. T., & Pirolli, P. (2007). SNIF-ACT: A Cognitive Model of User Navigation on the World Wide Web. *Human Computer Interaction*.
- Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4), 585–595.
- Gluck, K. (2010). Cognitive architectures for human factors in aviation. In E. Salas & D. Maurino (Eds.), *Human factors in aviation, 2nd edition* (pp. 375–400). New York, NY: Elsevier.
- Gray, W. D. (2007a). Composition and control of integrated cognitive systems. In W. D. Gray (Ed.), *Integrated models of cognitive systems*. New York: Oxford University Press.
- Gray, W. D. (Ed.). (2007b). *Integrated models of cognitive systems*. New York: Oxford University Press.
- Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109(4), 679–709.
- Myers, C. W., & Gray, W. D. (2010). Visual scan adaptation during repeated visual search. *Journal of Vision*, 8(10).
- Myers, C. W., Gray, W. D., & Sims, C. R. (2012). The insistence of vision: Why do people look at a salient stimulus when it signals target absence? *Visual Cognition*, 9(19), 1122–1157.
- Nason, S., & Laird, J. I. (2005). Soar-RL: Integrating reinforcement learning with Soar. *Cognitive Systems Research*, 6, 51–59.
- Peters, J., Vijayakumar, S., & Schaal, S. (2003). Reinforcement Learning for Humanoid Robotics. In *Humanoids2003, third ieee-ras international conference on humanoid robots, karlsruhe, germany, sept.29-30*.
- Quartermain, D., & Scott, T. H. (1960). Incidental learning in a simple task. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 14(3), 175–182.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In P. W. F. Black AH (Ed.), *Classical conditioning ii: Current research and theory* (pp. 64–99). New York: Appleton Century Crofts.
- Russell, S. J., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice Hall.
- Shen, J., Reingold, E. M., & Pomplun, M. (2000). Soar-RL: Integrating reinforcement learning with Soar. *Perception*, 29, 241–250.
- Siegler, R. S., & Stern, E. (1998). Conscious and unconscious strategy discoveries: A microgenetic analysis. , 127(4), 377–397.
- Stevenson, H. W. (1954). Latent Learning in Children. *Journal of Experimental Psychology*, 47(1), 17–21.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, Massachusetts: The MIT Press.
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4), 189–208.
- Veksler, V. D., Gray, W. D., & Schoelles, M. J. (2009). Goal-Proximity Decision Making: Who needs reward anyway? In *31st annual conference of the cognitive science society*.
- Voicu, H., & Schmajuk, N. (2002). Latent learning, shortcuts and detours: a computational model. *Behavioural Processes*, 59(2), 67–86.
- Widrow, B., & Hoff, M. (1960). Adaptive switching circuits. In *1960 ire wescon convention record* (pp. 96–104). New York: Institute of Radio Engineers.

# The Impact of Colour Difference and Colour Codability on Reference Production

Jette Viethen (h.a.e.viethen@uvt.nl)

Martijn Goudbeek (m.b.goudbeek@uvt.nl)

Emiel Krahmer (e.j.krahmer@uvt.nl)

Tilburg center for Cognition and Communication (TiCC)  
Tilburg University  
The Netherlands

## Abstract

It has often been observed that colour is a highly preferred attribute for use in distinguishing descriptions, that is, referring expressions with the purpose of identifying an object within a visual scene. However, most of these observations were based on visual displays containing only colours that were maximally different in hue and for which the language of experimentation possessed basic colour terms. The experiment described in this paper investigates the question whether people's preference for colour is reduced if the colour of the target referent is similar to that of the distractors. Because colours that look similar are often also harder to distinguish linguistically, we also examine the impact of the codability of the used colour values. The results of our experiment show that, while people are indeed less likely to use colour when the colours in the display are similar, this effect is entirely due to the difficulty in naming similar colours. When the colours of target and distractors are similar but can be named using different basic colour terms, no reduction in colour use was observed.

**Keywords:** reference production, language production, colour

## Introduction

Referring expressions are an essential part of communication. Whenever people engage in any type of discourse they use referring expressions to encode the entities that they are talking or writing about. Sometimes it suffices to use a pronoun to let the addressee know what is meant, but often a distinguishing description, a noun phrase differentiating the target referent from all other visually available distractor objects, is necessary. The production of such distinguishing descriptions has been a central theme for researchers both in psycholinguistic and in computational research on reference production. One particular question of interest is which attributes should be chosen for realisation in a distinguishing description, the problem of semantic content selection.

One of the most often made observations in psycholinguistic research regarding the choice of attributes for distinguishing descriptions is that people seem to favour colour over almost all other attributes when describing a target referent with the aim of identification (cf. Pechmann, 1989; Belke & Meyer, 2002; Sedivy, 2003; Brown-Schmidt & Tanenhaus, 2006; Arts, Maes, Noordman, & Jansen, 2011). This includes frequent redundant use of colour; cases in which the referring expression would be equally as distinguishing if colour was not mentioned. In some cases, people even use colour when all objects in a scene are of the same colour (Koolen, Goudbeek, & Krahmer, 2012).

However, as far as we know, all of this research was based on stimulus material using prototypical primary colours with clearly defined basic colour terms. In this paper, we investigate the question of whether people's preference for using the colour attribute diminishes or remains the same when the colour values in a visual scene are more similar to each other, and when no different basic colour terms exist for them.

Various researchers have argued that colour is preferred over, for example, size, in reference production, because it expresses absolute rather than relative information. In particular, Pechmann (1989) found in an early eye-tracking study that people usually begin to verbalise a description before they have fully scanned the scene. He found that a third of the descriptions in his data that contained both size and colour did not follow standard word order by mentioning colour before size (e.g., *the blue small car*).<sup>1</sup> He also noted that the first-mentioned attribute in overspecified descriptions was almost always colour, which often was ultimately not useful for the task of distinguishing the target referent from the visual context. He argued that both these observations might be due to the fact that colour is more easily cognisable than the other distinguishing features in his experiment because it can be perceived without having to compare the target referent to the other objects in the scene.

Belke and Meyer (2002) found similar overspecification effects for colour and size as Pechmann. They additionally provided eye-tracking evidence from a same-different judgement task for an account which credits this effect to differences in the way absolute and relative attributes are processed at a perceptual level. Based on experiments using the Stroop paradigm, Naor-Raz, Tarr, and Kersten (2003) even argued that an object's colour is an intrinsic component of the visual representation retained in long-term memory.

Another prominent source of evidence for people's preference for colour comes from corpus studies on purpose-built collections of referring expressions. The furniture section of the TUNA Corpus is a collection of human-produced distinguishing descriptions for furniture items differing in type, colour, size and orientation. In this corpus, colour is used redundantly more than three times as often as the other at-

<sup>1</sup>The standard word order is in this case identical for English and Dutch, the language of Pechmann's experiment.

tributes (Gatt, 2007, p. 82). In their recent experiments on semantic alignment in referring expressions, Goudbeek and Krahmer (2012) examined whether people can be primed to use a dispreferred attribute over a preferred one. Because they re-used the visual stimulus objects from the TUNA Corpus, they made the much higher frequency of colour over that of orientation in that corpus an underlying assumption in their experimental design.

A further corpus analysis by Viethen and Dale (2011), based on a large set of referring expressions for simple 3D scenes, also found that people mentioned the colour of the target object in a large proportion of the cases in which it was not necessary for identification. For size, on the other hand, their analysis found that its use depended highly on how well it distinguished the target from the other distractors, especially those of the same type as the target, pointing to a much more ‘utilitarian’ attitude towards size than towards colour. This is in line with findings from eye-tracking experiments which have shown that size is rarely used in situations where it adds no discriminatory power to the referring expression at all, while the same is not true for colour (Sedivy, 2003; Brown-Schmidt & Tanenhaus, 2006).

In light of this evidence, it is uncontroversial that colour plays a special role in referential communication. Yet, it must be noted that all of these results are based on stimuli with objects coloured in a small number of very different hues (red, blue, green, yellow, grey), sometimes even only black and white. In other words, the colour differences between the objects presented to participants were as large as possible.

No research exists using stimulus objects in similar colours. An intuitively plausible prediction is that the use of colour decreases as the similarity between the colours in the scene increases. This prediction follows also from Deutsch and Herrmann’s (1976) third postulate (p. 43). They show that in a situation with two identical objects that only differ in width and height, with a large difference in width and a small difference in height, people tend to use only the width attribute in a distinguishing description, and vice versa. Herrmann and Deutsch extrapolate from these findings that, in any situation in which more than one attribute can be used for identification, people will tend to use the one in which the objects differ most. If, on the other hand, the observed high rate of colour use in referential communication is indeed due to a smaller cognitive effort involved in mentioning it, as many other psycholinguists and computational linguists have argued, it should be unchanged in situations with colours that are not maximally different.

A confounding factor lies in the varying codability of different colour values. The more similar two colours are, the more likely it is that they fall within the range of the same basic colour term, such as *red*, *blue* or *yellow*, depending on the basic colour terms that exist in a given language.<sup>2</sup> In such

a case, more complex colour terms, such as *dark red* or *light blue* have to be constructed. It is conceivable that a colour value that is harder to encode is less likely to be verbalised.

Regarding the effect of colour difference, two conflicting hypotheses can be formulated:

1. The colour of an object is perceived independently from the colours of surrounding objects and gets included in distinguishing descriptions reflexively rather than based on a consideration of its usefulness. This hypothesis is in line with most claims in the literature and predicts that the extent of the difference between the target’s colour and that of the distractor objects has no impact on people’s reference behaviour.
2. The high use of colour is based to some extent on an assessment of the difference in colour between the target item and the distractors. Following from (Herrmann & Deutsch, 1976), a lower use of colour should be expected when the colours are similar than in situations where the colours are as different from each other as possible.

For the effect of colour codability our hypothesis is:

3. The codability of a target’s colour with respect to distractor colours effects the likelihood of it being used in a distinguishing description. Colours that can be named by a basic colour term are more likely to be included than those for which a complex term has to be used.

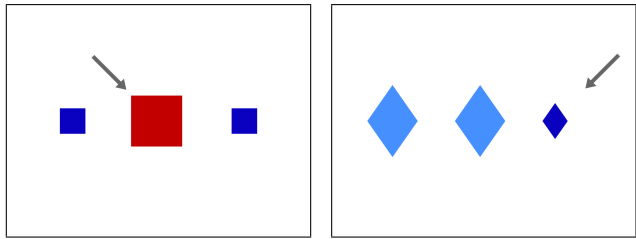
In the following, we describe an experiment designed to arbitrate between Hypotheses 1 and 2 and to test Hypothesis 3. Our results support the assumption of Hypothesis 3 that the use of colour is reduced when the codability of the colour value of an item is reduced, and advocate Hypothesis 1 over Hypothesis 2.

These results can inform ongoing research on developing computational models of reference production, as this work has begun to align its focus with that of psycholinguistic research. Researchers from the computational field are looking more and more for evidence about how humans solve the problem of content selection for reference production, in order to inform their models (cf. Dale & Reiter, 1995; Kelleher & Kruijff, 2006; Viethen & Dale, 2006; Deemter, Gatt, Sluis, & Power, 2012). One main reason for this move towards human-likeness as a criterion for task success of reference generation systems is the aim to create computational models that are in some sense cognitively plausible. The results of our experiment show that even computational models that are solely focussed on content selection for reference production need to pay more attention to the problem of lexical choice, as these two issues appear to be more closely intertwined than most existing models acknowledge.

## Experiment

The experiment took the form of a reference production task, in which participants were shown displays of simple geometric objects on a computer screen. They were asked to describe

<sup>2</sup>Which hues are grouped under the same basic colour terms differs for different languages, as they carve up the colour spectrum in different ways and at different granularities (Kay, Berlin, Maffi, Merrifield, & Cook, 2010).



(a) A hidiff item: one large red and two small blue squares. (b) A lodiff item: one small dark blue and two large light blue diamonds.

Figure 1: Example stimuli from the two colour-difference conditions.

one of the objects in such a way that an imaginary partner would be able to identify it.

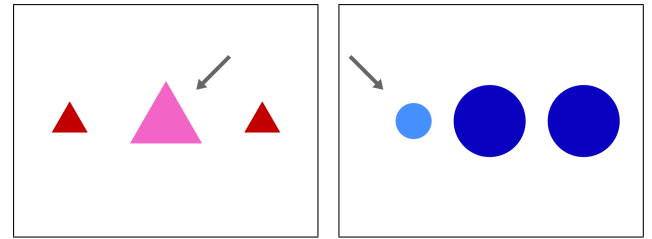
## Method

**Participants** 63 undergraduate students of Tilburg University took part in the experiment in return for course credit. 48 were female and 15 male. Their age ranged from 19 to 26 years ( $M = 20$  years and 10 months). They were all fluent speakers of Dutch, the language of the experiment.

**Materials and Design** Each participant was shown 32 critical items and 64 filler items. The critical trials consisted of simple scenes containing three two-dimensional geometrical figures: one intended referent and two distractor objects. In order to keep the design as simple as possible, the two distractor objects were identical. However, the target item differed in both colour and size from the two distractors, so that either of these two attributes was sufficient to fully distinguish it.

Our main manipulation concerned the difference in colour between the target and the distractors. In half of the trials this difference was large (hidiff condition), in the other half it was small (lodiff condition). Figure 1(a) shows a trial from the hidiff condition, and Figure 1(b) is an example from the lodiff condition.

As discussed above, the more similar two colours are, the less likely it is that they can be distinguished by basic colour terms. For example, the basic colour term *blue* is not sufficient to distinguish the target in Figure 1(b) from the distractors; instead, the complex colour term *dark blue* has to be used. This applies in Dutch in the same way as in English. To test the impact that the codability of different colour values might have on the content of referring expressions (see Hypothesis 3), we used a nested variable within the lodiff condition, by including two different hues: red and blue. For red hues, Dutch (just as English) possesses two different basic colour terms, even at a low difference, namely *rood* (red) and *roze* (pink). Thus, stimuli with red and pink objects, such as the one in Figure 2(a), form the hicode condition. For blue, the complex colour terms *donkerblauw* (dark blue) and *lichtblauw* (light blue) have to be used, resulting in a locode condition (an example stimulus is shown in Figure 2(b)). The lodiff items were equally divided between the hicode condi-



(a) A hicode item: one large pink and two small red triangles. (b) A locode item: one small light blue and two large dark blue circles.

Figure 2: Example stimuli from the two colour-codability conditions.

tion and the locode condition.

To determine the exact colour values to use we referred to the Hue Saturation Brightness (HSB) colour model. For the two dark colours we used the canonical values for blue ( $H = 245^\circ$ ) and red ( $H = 0^\circ$ ), 100% saturation, and a slightly lowered brightness (75%). For the lighter colours, we subtracted  $35^\circ$  from the original hue values, decreased the saturation and increased the brightness. We finetuned the values for the lighter colours based on a pretest, to ensure that people would agree on calling them *roze* and *lichtblauw*. This resulted in the HSB values ( $215^\circ$ , 70%, 100%) for light blue and ( $320^\circ$ , 58%, 95%) for pink.

To ensure that there were the same number of target objects in each of the four colours (red, pink, dark blue and light blue), half the items in the hidiff condition used red and dark blue objects, and the other half pink and light blue ones. The position of the target was balanced across items. Furthermore, each condition contained a balanced number of trials using each of the four object types.

The type of the distractor objects was always the same as that of the target, so that type was never distinguishing. However, the size of the distractors was different from that of the target object, in order to give the participants an alternative option to using colour. It would not make sense to measure the rate of colour use, if colour was the only distinguishing feature in some or all trials.

We aimed to keep the size difference between target and distractors constant across all trials. To this end, we defined the size of an object by the length of its longest internal distance (the diameter for a circle, the diagonal for a square, an edge for a triangle, and the vertical line in a diamond), rather than, for example, its area. The longest internal distance of the large objects was set to twice that of the small objects.

**Filler Items** We included two types of fillers, which were carefully designed to mislead the participants regarding the exact aims of the experiment.

The 32 geometrical fillers were similar to the critical stimuli in that they showed three geometrical objects, but they used type and pattern as distinguishing attributes. Colour and size were never fully distinguishing in the fillers, in order to avoid priming the use of these two attributes. The target was

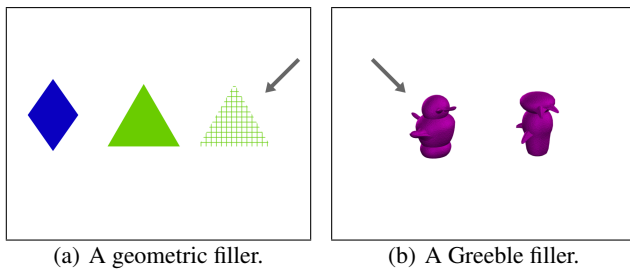


Figure 3: Two example filler items.

either striped or checkered so as not to prime the use of *solid* as a pattern which might then also show up in the critical trials where it was non-distinguishing. Half of the geometrical fillers were in black and white and in 9 of them the target was green, in order to distract from the small set of colours used in the critical trials. Again, the target referent's type and size was balanced across the whole set. Figure 3(a) shows an example of a geometric filler.

The 32 'Greeble' fillers each showed two novel 3D figures in purple.<sup>3</sup> We chose pairs such that the target object could always be distinguished from the distractor object by its main shape and the direction in which its protrusions were pointing. Because these objects are designed to be difficult to describe and look very different from the geometric items, we hoped they would prevent the participants from adopting a standard strategy for describing the geometric items. An example Greeble filler is shown in Figure 3(b). Debriefing revealed that the participants were not aware of the purpose of the experiment, and the majority of participants believed that the Greeble items were the critical stimuli of the experiment.

**Procedure** Two stimulus lists were created by producing one random ordering and then reversing it for the second list. Each critical stimulus was prepended with one geometrical and one Greeble filler item, which were chosen semi-randomly in a way such that the target was never in the same position in more than four items in a row. The item directly before each critical stimulus was always a Greeble filler to minimise any possibility of lexical or semantic priming from the geometrical filler responses to the critical responses.

The Dutch instructions told the participants that they would see a number of simple scenes on a computer screen. They were asked to verbally describe the object pointed at by an arrow to an imaginary partner without using position information. They had to complete the sentence *Klik nu op de/het ...* ('Now click on the ...') which was shown underneath each item. Their voice was recorded using a headset.

Before each item, a fixation cross was displayed for 1.5 seconds, then the stimulus item was shown for 4.5 seconds during which the participants had to give their response. We introduced this relatively short response time after finding in

<sup>3</sup>The Greebles are courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, <http://www.tarrlab.org/>.

Table 1: Count and proportion of responses containing colour.

condition	count	mean	stdev
hidiff (N=1008)	747	.74	.35
lodiff (N=1008)	681	.68	.36
lodiff-hicode (N=504)	376	.75	.36
lodiff-locode (N=504)	305	.61	.40

a pilot experiment that participants tended to exhaustively describe the whole scene.

## Results

**Coding of the Independent Variables** As main dependent measure, we analyse the proportion of colour use in the different conditions. We consider a description to contain colour, if a colour value is mentioned that is true of the target, independently of the distinguishingness of this value. For example, we consider the description in Example (1) for the target in Figure 2(b), where all three objects are blue, as a use of colour. As a secondary measure we also look at the use of size, in order to get an insight into the question whether colour gets mostly used redundantly.

(1) de kleine blauwe cirkel

(the small blue circle) [for the stimulus in Figure 2(b)]

The responses were transcribed and coded for use of colour and size by a Dutch native-speaker.

**Data Analysis** Table 1 displays the mean proportion and standard deviation of colour use in the hidiff and lodiff conditions as well as the two nested conditions under lodiff (hicode and locode). It shows that people were more likely to use colour when the colours in the stimulus scene were very different than when they were similar. However, it also shows that the mean proportion of colour use in the lodiff-hicode condition was very similar to that in the hidiff condition.

We conducted a within-participants analysis of variance (ANOVA) to compare the three meaningful conditions (hidiff, lodiff-hicode, lodiff-locode), which showed the differences between these conditions to be highly statistically significant ( $F(2, 124) = 19.9, p < .001, \eta^2 = .24$ ). A test of planned within-participant Contrasts confirmed that the difference between the hidiff and the lodiff condition was significant ( $F(1, 62) = 18.7, p < .001, \eta^2 = .23$ ); participants used colour more in the hidiff condition than in the lodiff condition. The same is the case for the effect of codability (locode vs. hicode conditions) ( $F(1, 62) = 20.3, p < .001, \eta^2 = .25$ ), confirming that people used colour more when the colours could be distinguished by basic colour terms. However, there was no statistically significant difference between lodiff-hicode and the hidiff condition ( $F(1, 62) < 1$ ).

For size, the opposite picture emerges. Table 2 shows that people were less likely to use size when the colour difference was high than when the colours were similar, and that people used size more often in situations in which the name of the

Table 2: Count and proportion of responses containing size.

condition	count	mean	stdev
hidiff (N=1008)	584	.58	.36
lodiff (N=1008)	679	.67	.31
lodiff-hicode (N=504)	282	.56	.36
lodiff-locode (N=504)	397	.79	.30

colour was difficult to encode. Again, the difference between the hidiff and the lodiff-hicode conditions does not appear very big.

The statistical analysis with tests of planned Contrasts revealed the same pattern as for colour use: the overall difference between hidiff, lodiff-hicode, and lodiff-locode is significant with an even bigger effect size ( $F(2, 124) = 39.7, p < .001, \eta^2 = .39$ ); as are the differences between hicode and overall locode ( $F(1, 62) = 19.3, p < .001, \eta^2 = .24$ ) and between hicode and locode-lodiff ( $F(1, 62) = 52.3, p < .001, \eta^2 = .48$ ). This means that people used size less often in the hidiff and the hicode conditions than in the locode condition. Again, there was no statistically significant difference between hidiff and lodiff-hicode ( $F(1, 62) = 1.1$ ).

## Discussion

The main observation from our results is that a smaller difference in colour alone does not result in a decrease in the use of colour in referring expressions. The apparent difference in colour use between the hidiff and lodiff conditions arises solely from the difficulty in coding the colour value in the locode condition. This lends support to Hypothesis 1, stating that people's preference for colour is independent from its value. It also confirms Hypothesis 3, which predicts that colours that are difficult to name because no distinguishing basic colour term is available, are less likely to be mentioned in a distinguishing description.

Interestingly, there were 99 distinguishing descriptions that contained a non-distinguishing colour value, such as in Example (1) above. All 99 of these cases occurred in the lodiff-locode condition. It is not surprising that no such cases occurred in the other conditions, because no basic colour terms exist that encompass both red and blue, red and pink, or pink and blue. However, the fact that almost a third of all colour terms used in the locode condition were non-distinguishing further supports the hypothesis that people often mention colour not for its discriminatory power but because it is easily available perceptually. By mentioning the basic, yet non-distinguishing, colour term *blauw* they can follow their preference for using colour but avoid the difficulty involved in retrieving and uttering a more complex colour term. This raises the question whether it is indeed the complexity of a colour term that stops people from using it or rather the fact that in our locode scenes the target's colour term (e.g. *lichtblauw* in Figure 2(b)) partly overlaps lexically with that applying to the distractors (*donkerblauw* in Figure 2(b)).

Furthermore, of the 37 descriptions in which a property was mentioned that was not true of the target object, only one

used a wrong colour (*pink* instead of *red*, and in this case the participant corrected themselves). This further strengthens the argument that colour naming is an inherently easier task than naming the size made by a number of researchers including (Pechmann, 1989; Belke & Meyer, 2002; Naor-Raz et al., 2003; Kelleher & Kruijff, 2006).

The rate at which people used size was inversely proportional to the use of colour. Of course, size had to be used in descriptions not including colour in an identification task with these two attributes as the only distinguishing features. However, this does not necessarily mean that it has to be omitted in cases in which colour was mentioned. Instead, the rate of size use might have stayed constant, indicating a relatively high rate of overspecification in the hidiff and hicode conditions. Two possible explanations for the difference in the use of size between the different conditions are conceivable. First, it might be the case that the choice to use size is influenced directly by the experimental variables. This might be due to the fact that the speaker has to scan the scene in order to determine the relative size of the target object. While scanning a locode scene he might notice the usefulness of colour—which according to Pechmann's (1989) and Belke and Meyer's (2002) incrementality accounts might already have been uttered at this stage—and decide whether to use size based on this information alone, independently of whether colour is actually mentioned or not. Second, the use of size might be impacted by the use of colour. People might make their choices about which attributes to use sequentially, one attribute at a time and the decision about size succeeds the decision about colour. So, once a speaker has decided not to mention colour, size has to be included in order to fulfil the referential task of identification. Further experimentation would be required to arbitrate between these two accounts.

**Consequences for Computational Modelling** The main assumption regarding the use of colour remains unchallenged by our results: colour is highly preferred by human speakers and should therefore feature highly in the output of computational referring expression generation systems that are aimed at producing human-like output. However, our results re-emphasise the importance of an issue which seems to have lost traction in the decades since (Dale & Reiter, 1995): that of lexical choice. Dale and Reiter's original algorithm included a *FindBestValue* function, acknowledging the fact that different level values exist for many attributes and that not all values are equally adequate in a given situation. However, their algorithm makes its decision about which attributes to include based on the most distinguishing value for an attribute, meaning that a colour value expressed by a more complex term, such as *light blue*, is more likely to be included for the colour attribute than a basic one, such as *blue*. This is of course not advocated by our data.

Our findings speak loudly against the separation of semantic content selection and lexical choice present in most recent computational approaches to referring expression generation. Computational reference production models with a claim to



human-likeness need to take into account how difficult it will be to realise each attribute lexically already when they make the decision about the use of this attribute. The results presented here clearly show that even highly preferred attributes such as colour should get included less often in situations in which they are hard to code, or that in some cases a less specific value should get used.

A second point emerging from our data is that the deterministic nature of most existing computational reference production models is clearly not in line with human reference behaviour. While we can observe increases or decreases in the use of certain attributes depending on different experimental variables, there always remains a large amount of variation. Therefore, REG systems that are serious about modelling human behaviour must begin to use probabilistic mechanisms in order to be able to capture the non-deterministic choices people make when they refer. A notable first move in this direction was made by Gatt, van Gompel, Krahmer, and van Deemter (2011).

### Conclusions

Previous research often took it for granted that colour is a highly preferred attribute in reference production, but so far a serious and systematic study of this has been lacking. Existing results were based on stimuli in maximally different primary colours; this paper is the first to investigate what happens if the stimulus colours are similar to each other. Our results suggest that the similarity between the colour of the target referent and that of any distractor objects indeed has little effect on the content people choose for a referring expression, supporting the view that colour gets chosen due to being perceivable with low cognitive effort.

However, we show that colours that can be encoded using a basic colour term, such as *blue*, are more likely to be mentioned than those for which a more complex term, such as *light blue*, has to be found in order to distinguish from, for example, dark blue distractors. Current computational models of reference production do not account for this result, as they usually separate the selection of semantic content and lexical choice into two distinct processes.

### Acknowledgments

The research reported in this paper forms part of the VICI project “Bridging the Gap between Psycholinguistics and Computational Linguistics: the Case of Referring Expressions”, funded by the Netherlands Organization for Scientific Research (NWO grant 277-70-007). We thank Elsa Jonkers for help with the transcription and annotation of the data.

### References

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Over-specification in written instruction. *Linguistics*, 49(3), 555–574.

Belke, E., & Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing time during same-different decisions. *European Journal of Cognitive Psychology*, 14(2), 237–266.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, 54, 592–609.

Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2), 233–263.

Deemter, K. van, Gatt, A., Sluis, I. van der, & Power, R. (2012). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, in press.

Gatt, A. (2007). *Generating Coherent Reference to Multiple Entities*. Unpublished doctoral dissertation, University of Aberdeen, UK.

Gatt, A., van Gompel, R., Krahmer, E., & van Deemter, K. (2011). Non-deterministic attribute selection in reference production. In *Proceedings of the Workshop on Production of Referring Expressions: Bridging the gap between empirical, computational and theoretical approaches to reference (PRE-CogSci 2011)*. Boston MA, USA.

Goudbeek, M., & Krahmer, E. (2012). Alignment in interactive reference production: Content planning, modifier ordering and referential overspecification. *Topics in Cognitive Science*, 4(2), 269–289.

Herrmann, T., & Deutsch, W. (1976). *Psychologie der Objektbenennung*. Bern: Verlag Hans Huber.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2010). *The World Color Survey*. Stanford CA, USA: CSLI Publications.

Kelleher, J., & Kruijff, G.-J. (2006). Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 1041–1048). Sydney, Australia.

Koolen, R., Goudbeek, M., & Krahmer, E. (2012). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, to appear.

Naor-Raz, G., Tarr, M. J., & Kersten, D. (2003). Is color an intrinsic property of object representation? *Perception*, 32(6), 667–680.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27(1), 89–110.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32(1), 3–23.

Viethen, J., & Dale, R. (2006). Algorithms for generating referring expressions: Do they do what people do? In *Proceedings of the 4th International Conference on Natural Language Generation* (pp. 63–70). Sydney, Australia.

Viethen, J., & Dale, R. (2011). GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. In *Proceedings of the Workshop on Using Corpora in Natural Language Generation (NLG): Language Generation and Evaluation (UCNLG+Eval)*. Edinburgh, UK.

**Bayesian Logic and Trial-by-Trial Learning**  
**Momme von Sydow (momme.von-sydow@psychologie.uni-heidelberg.de)**  
**Klaus Fiedler (klaus.fiedler@psychologie.uni-heidelberg.de)**  
University of Heidelberg, Department of Psychology, Hauptstr. 47-51,  
D-69117 Heidelberg, Germany

**Abstract**

Standard logic and probability theory are both beset with fundamental problems if used as adequacy criteria for relating logical propositions to learning data. We discuss the problems of exception, of sample size, and of inclusion. Bayesian pattern logic ('Bayesian logic' or BL for short) has been proposed as a possible rational resolution of these problems. BL can also be taken as psychological theory suggesting frequency-based conjunction fallacies (CFs) and a generalization of CFs to other logical inclusion fallacies. In this paper, this generalization is elaborated using trial-by-trial learning scenarios without memory load. In each trial participants have to provide a probability judgment. Apart from investigating logical probability judgments in this trial-by-trial context, it is explored whether under no memory load the propositional assessment of previous evidence has an influence on further probability judgments. The results generally support BL and cannot easily be explained by other theories of CFs.

**Keywords:** Conjunction fallacy, probability judgments, trial-by-trial learning, Bayesian induction, logical predication.

**Standard Logic and Probability Theory  
as Criteria for True Logical Propositions**

The relationship between general logical propositions (or sentences) and evidence is fundamental to both epistemology and psychology. We here investigate general predication of logical relationships between two dichotomous attributes (or predicates), like "ravens are black *and* they can fly" (with the conjunction 'and'). What would be an adequate justification for such a type of sentences?

Arising from an old tradition going back to Aristotle, modern formal logic uses truth table definitions for all 16 logical connectives. The truth table definition may be used as a deterministic criterion of truth for empirical relationships. With regard to a conjunctive predication, like "ravens (*R*) are black (*A*) and they can fly (*B*)" ( $A \wedge B | R$ ), the whole sentence is true (or, more correctly, 'not false') as long as one has observed only exemplars corresponding to true cells of a truth table (for the conjunction this is the '*a*-cell', ' $A \wedge B$ '). In contrast, the proposition would be falsified, if one observed a single case defined to be false (here: *b*-cell: ' $A \wedge \neg B$ '; *c*-cell: ' $\neg A \wedge B$ ', or *d*-cell: ' $\neg A \wedge \neg B$ ').

**Problem of Exceptions** Exceptions may not prove the rule, but in ordinary language exceptions are indeed regularly tolerated. This may reflect the deeper epistemological point that in the empirical world deterministic relationships are rather the exception than the rule. Actually, in philosophy of science it has been argued that strict falsificationism would absurdly imply that *all* important theories would be falsified. Even more so in normal language, as evident from our deterministic example, there exist exceptions: white (albino) ravens as well as ravens that

cannot fly. If exceptions are the rule for contingent, empirical relationships, it seems reasonable to replace the strict deterministic truth criteria of logic by a high-probability criterion (see Schurz, 2005):  $P(\text{black} \ \& \ \text{can fly} \mid \text{ravens}) > \Psi$ , with  $\Psi > .5$ . However, the following two problems beset a simple extensional probability criterion of truth as well as one based on standard formal logics.

**Problem of Sample Size** If we had to access the truth of "ravens are black and they can fly" without previous knowledge about ravens, either one confirmatory raven ( $A \ \& \ B$  case) or many cases both equally yielded the same extensional probability of 1 (the number of confirmative cases divided by all cases). In the latter case, however, a higher subjective probability of this sentence seems justified. Therefore, a kind of second order probability, a probability concerning probabilities, is needed, as introduced in the model.

**Problem of Inclusion** The extension (all cases falling into a set) of a subset can never be larger than that of a superset. Comparing conjunctions and inclusive disjunctions, it follows that  $P(\text{ravens are black AND they can fly}) \leq P(\text{ravens are black OR they can fly or both})$  [formally:  $P(A \wedge B | R) \leq P(A \vee B | R)$ ]. If we use extensional probabilities as truth criterion, the second sentence can therefore never be 'less true' than the first one. If one assumes at least some exceptions, the latter is even 'truer' in principle. Going one step further, the logical tautology, allowing for all values ("Ravens are black or not, and they can fly or not"), is *a priori* the extensionally most probable sentence [ $P(A \vee B | R) \leq P(A \text{ T } B | R)$  or even  $P(A \vee B | R) < P(A \text{ T } B | R)$ ]. Using standard (extensional) probabilities as truth criterion, one would therefore always have to choose tautologies as the most suitable hypothesis, regardless of the evidence and of the properties in question. In conclusion, if a truth criterion should be informative about the observable world, simple extensional probabilities in principle cannot provide a reasonable truth criterion.

**Bayesian Logic**

Bayesian pattern logic (or 'Bayesian logic', BL, for short) formulates a second order probability that given data may have been generated by noisy-logical patterns of probabilities. The model provides a technical, rational solution to the three mentioned problems and – in approximation – a potential psychological model of human induction of noisy-logical relationships as well. The model is part of a renaissance of Bayesian approaches in cognitive science (e.g., Chater, Tenenbaum, Yuille, 2006; Oaksford & Chater, 2007; Kruschke, 2008). The following sketch is meant to clarify the main idea of Bayesian logic (for more detail, see von Sydow, 2011).

The construction of the model starts with all 16 dyadic logical connectives known from standard propositional logic. The logical truth tables are taken as explanations that are distinguished from the data level. While standard logic makes no assumptions about probabilities of true classes in a truth table, Step 1 of BL formulates ideal explanations by assuming equi-probability of all true classes of a truth table. For instance, for the exclusive disjunction ( $X$  are *either*  $A$  or  $B$ , *but not both*) it is assumed that  $P(\text{b-cell}) = P(\text{c-cell}) = 1/2$  (for no noise,  $R = 0$ ). Thereby, 2 by 2 truth tables become 2 by 2 probability tables. Note, however, that such ideal explanations need not generate ideal data patterns. In Step 2 (cf. Fig. 1) the idea of exceptions is modeled by introducing possible levels of noise. For each possible level a uniform noise function is added to all four cells of probability table, followed by a normalization, so that the resulting sum of all four cells of a probability table adds up to unity. This results in a field of ideal (explanatory) noisy-logical patterns of probabilities, each with an additional second order probability:  $P(A \text{ connective } B, \text{ noise level } R | \text{data}) =: P(A \circ B, R | D)$ . Here flat priors for the connectives and noise levels are used for each new situation.

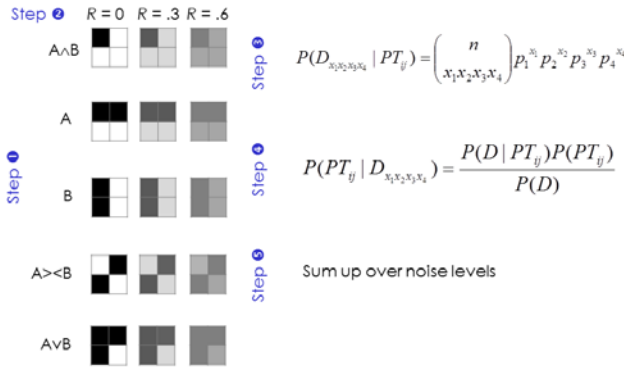


Figure 1: Sketch of the model of only five logical propositions and three noise levels (cf. text for details).

For the novel noisy-logical representation one can calculate the posterior probabilities for each probability table by combining some standard statistics. Given observed data about the co-occurrence of  $A$  and  $B$  (a 2 by 2 contingency matrix), one can calculate (Step 3) the likelihood of the data given a probability table,  $P(D | A \circ B, R)$ , by using the multinomial distribution, which here determines for each table of four probabilities how likely it produces the observed four frequencies. In Step 4, Bayes' theorem is used to transform the likelihood  $P(D | A \circ B, R)$  into a posterior  $P(A \circ B, R | D)$ . In a final step, one sums up the probabilities of a connector over all noise levels (here we modeled 11 equidistant levels from  $R = 0$  to 1). We obtain the requested posterior *pattern* probability,  $P_P(A \circ B | D)$ , clearly differing from the analogous *extensional* probability,  $P_E(A \circ B | D)$  (frequency of positive cases, divided by all cases).

## BL and the Conjunctions Fallacy Debate

One of the most heated and philosophically interesting psychological debates concerns the apparent inability of people to understand that conjunctions (for instance, "Linda is a Bank teller and a feminist") can never be (extensionally) more probable than their conjuncts (e.g., "Linda is a bank teller")—even for apparent feminists. This phenomenon has been called "conjunction fallacy" and first has been explained by the representativeness heuristic (Kahneman & Tversky, 1982). This heuristic, however, has been criticized as being formulated too imprecise (Gigerenzer, 1996; cf. Nilsson, Juslin, & Olsson, 2008).

There have been several other classes of explanations of CFs. One focusses on possible misunderstandings. "A and B" may actually be understood as "A or B" or to "if A then B" (Mellers, Hertwig, & Kahneman, 2001; Hertwig, Benz & Krauss, 2008). Moreover, "A" may be interpreted as "A but not B" instead of "A, whether B or not B" (Kahneman & Tversky, 1983; Hilton, 1995; cf. Sides, Osherson, Bonini, & Viale, 2002; Wedell & Moro, 2008). A second class of explanations considers different ways in which probabilities are introduced and how the probability question is posed. It has been shown that frequency presentations (Fiedler, 1988; Gigerenzer, 1996), rating formats (Sloman, Over, Slovak, & Stibel, 2003), and clear set inclusions (Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 1999; Sloman et al., 2003) often substantially reduced the portion of CFs. Although these factors often do play a role, BL under certain conditions (if one is concerned with alternative hypotheses about whole situations) has predicted CFs even when all these factors apply simultaneously (von Sydow, 2011a, b). A third class of explanations specifies quantitative conditions of CFs. Most prominently, it has been suggested that the requested probability  $P(A \wedge B | D)$  is replaced by the inverse probability  $P(D | A \wedge B)$ ; cf. Wolford, 1991; Fisk & Slattery, 2005), or by a measure of support, like  $P(A \wedge B | D) - P(A \wedge B)$  (support theory, cf. Sides, et al., 2002; Lagnado & Shanks, 2002; cf. Tentori, Bonini, & Osherson, 2004; Crupi, Fitelson, & Tentori, 2008), or several other measures, like signed summation, averaging, quantum logic, or rescaling (see Wedell & Moro, 2008; von Sydow, 2009).

BL provides a rational quantitative account of frequency-based a particular class of conjunction fallacies and made several novel predictions that cannot be explained by the previous models (von Sydow, 2011). One important aspect has been the generalization of the idea of CFs into a system of logical inclusion 'fallacies' (von Sydow, 2009).

## Experiment: Trial-By-Trial Induction of Logical Relationships

The primary goal of the reported experiment is to test aspects of the postulated system of frequency-based logical inclusion 'fallacies' in a trial-by-trial way. Whereas confirmatory results for this system have already been achieved, even using trial-by-trial *presentation* of items (von Sydow, 2011b; cf. Lagnado et al., 2001), we here additionally

investigate trial-by-trial *assessment of the dependent variable*: the selection of the most probable hypotheses after each new observation. To the best knowledge of the authors, this has never been investigated before in the CF debate.

A supplementary goal is to assess whether putting evidence into language in the course of trials may have an additional top-down effect on the successive evaluation of evidence. Here the *ways* how one obtains a final (fixed) pattern of evidence are varied, so that this may affect the predicted propositional representations. In one condition the finally predicted hypothesis is expected to appear most probable all along (homogeneous condition) and in another condition different hypotheses are predicted to appear more probable throughout the first learning trials (heterogeneous condition). In its current formulation BL, as a model of data-based induction, would not be able to account for such top-down effects. This is the case although BL goes beyond naïve probability, and leaves room also for subjective priors. As we think there are top-down effects for instance of categorization (Hagmayer, Meder, von Sydow, & Waldmann, 2011) or causal coherence (von Sydow, Hagmayer, Meder, & Waldmann, 2010), we think there may well be top-down-effects of mere verbalization. In this experiment, however, participants are provided with summary statistics, excluding memory effects. In such settings, also intended as base-line for future experiments, no such additional top-down effects are expected.

	Phase 1: Pattern phase	Phase 2: First trial-by-trial phase	Phase 3: Second trial-by-trial Phase
G1	For all groups G1 to G8	C1: AND, homogeneous	C3: ><, homogeneous
G2	Pattern 1    Pattern 2	C1: AND, homogeneous	C4: ><, heterogeneous
G3	15   2    5   6	C2: AND, heterogeneous	C3: ><, homogeneous
G4	2   3    2   2	C2: AND, heterogeneous	C4: ><, heterogeneous
G5	Pattern 3    Pattern 4	C3: ><, homogeneous	C1: AND, homogeneous
G6	0   2    4   1	C3: ><, homogeneous	C2: AND, heterogeneous
G7	0   1    0   0	C4: ><, heterogeneous	C1: AND, homogeneous
G8	Pattern 5    Pattern 6	C4: ><, heterogeneous	C2: AND, heterogeneous
	8   10    2   12		
	9   9    13   3		

Figure 2: Design (see main text for details).

The design involves three phases. All phases involve a selection of the most probable logical hypothesis given some evidence. In *Phase 1*, participants in all conditions are randomly presented with six patterns of evidence, each referring to a different situation (Fig. 2, Phase 1). First, this phase should replicate previous generalizations of BL (von Sydow, 2009, 2011b). Secondly, it investigates whether participants grasp the intended meaning of logical terms, and,

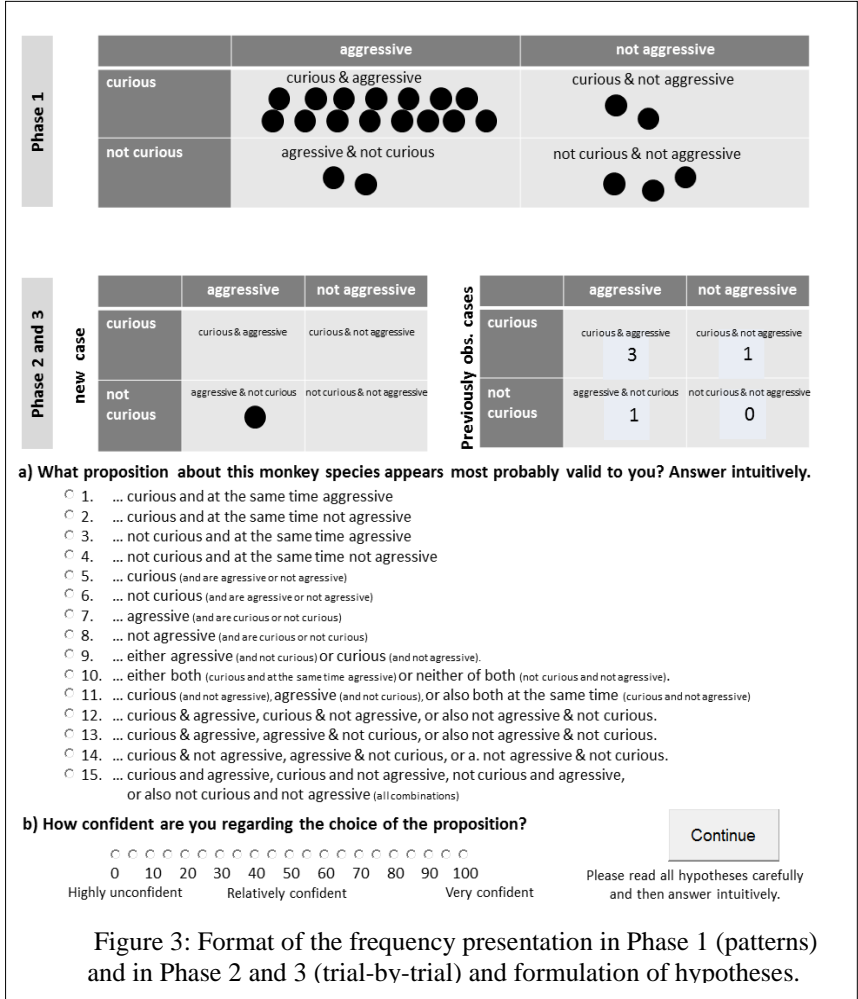


Figure 3: Format of the frequency presentation in Phase 1 (patterns) and in Phase 2 and 3 (trial-by-trial) and formulation of hypotheses.

thirdly, it excludes a deterministic understanding of the rules in the next phases by inducing a non-deterministic noise-prior (especially for few observed cases, priors may well affect the results).

Phase 2 and 3 are both trial-by-trial judgment tasks. BL predicts that various hypotheses should be selected to be most probable, each from an extensional perspective involving several logical inclusion fallacies. The sequences should end up either in an AND hypothesis (C1, C2) or an EITHER-OR hypothesis (C3, C4). Both hypothesis are extensionally less probable than the OR hypothesis or the tautology. Additionally, the order in which data is presented differs, investigating whether verbalization throughout learning affects the verbalization of identical final patterns (the probability judgments). As sketched, either a homogeneous condition (C1, C3) or a heterogeneous condition (C2, C4) is used. Finally, Phase 2 and 3 are identical, in order to assess whether the previous learning phase had an effect (as, e.g., suggested by support theory) and to find out whether participants increasingly make either extensional or BL selections.

## Material

130 participants of the University of Göttingen participated in the experiment. The participants were told about newly

discovered species of apes on a lonely island. They were in the role of ethologist concerned with statements the animals of a species are curious or not (here *A*) and whether they are aggressive or not (here *B*), as well as judging the relation of these properties.

In Phase 1 participants were concerned with six species of apes in randomized order. For each species they were shown a photo of an ape (e.g., “*P. calvus*”) with a text “The animals of this species are...”, leading to the main instructions (Fig. 3) and a contingency table summing up the observed features combinations (cf. Fig. 2, 3). For each species one had to select the most probable logical hypothesis and one had to provide a confidence rating (Fig. 3).

Phase 2 and 3 were concerned with trial-by-trial learning. Participants were randomly assigned to the eight conditions. Single events were symbolized by a circle flying to a place in the contingency table (Fig. 3, Phase 2/3, left table),

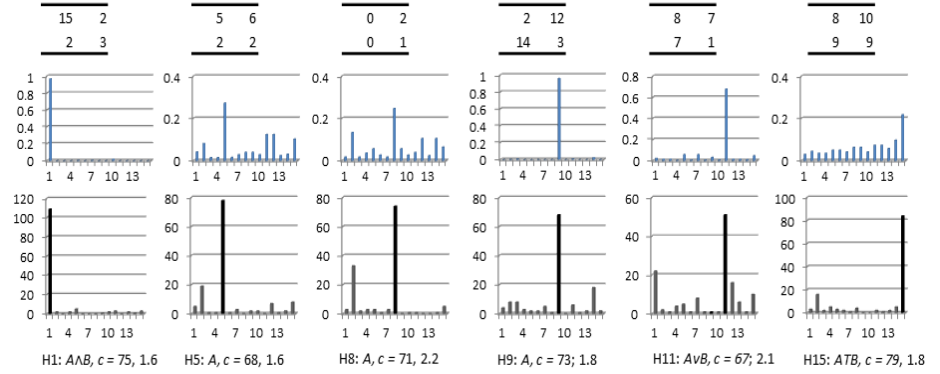


Figure 4: Six patterns of Phase 1 (first row), BL's pattern probabilities (second row) and the frequency of hypotheses (cf. Fig. 3) selected (third row).

followed by an update of a summary table (right table). Of the 18 trials the first nine are presented in Figure 5 and 6.

In all probability judgment tasks the formulations of the hypotheses were carefully chosen to rule out the plausible misunderstandings discussed in the CF debate. For instance, the conjunctions were formulated as “*A* and *at the same time B*” and the single conjuncts (the affirmations or negations) as “*A* (and are *B* or not *B*)” (Fig. 3).

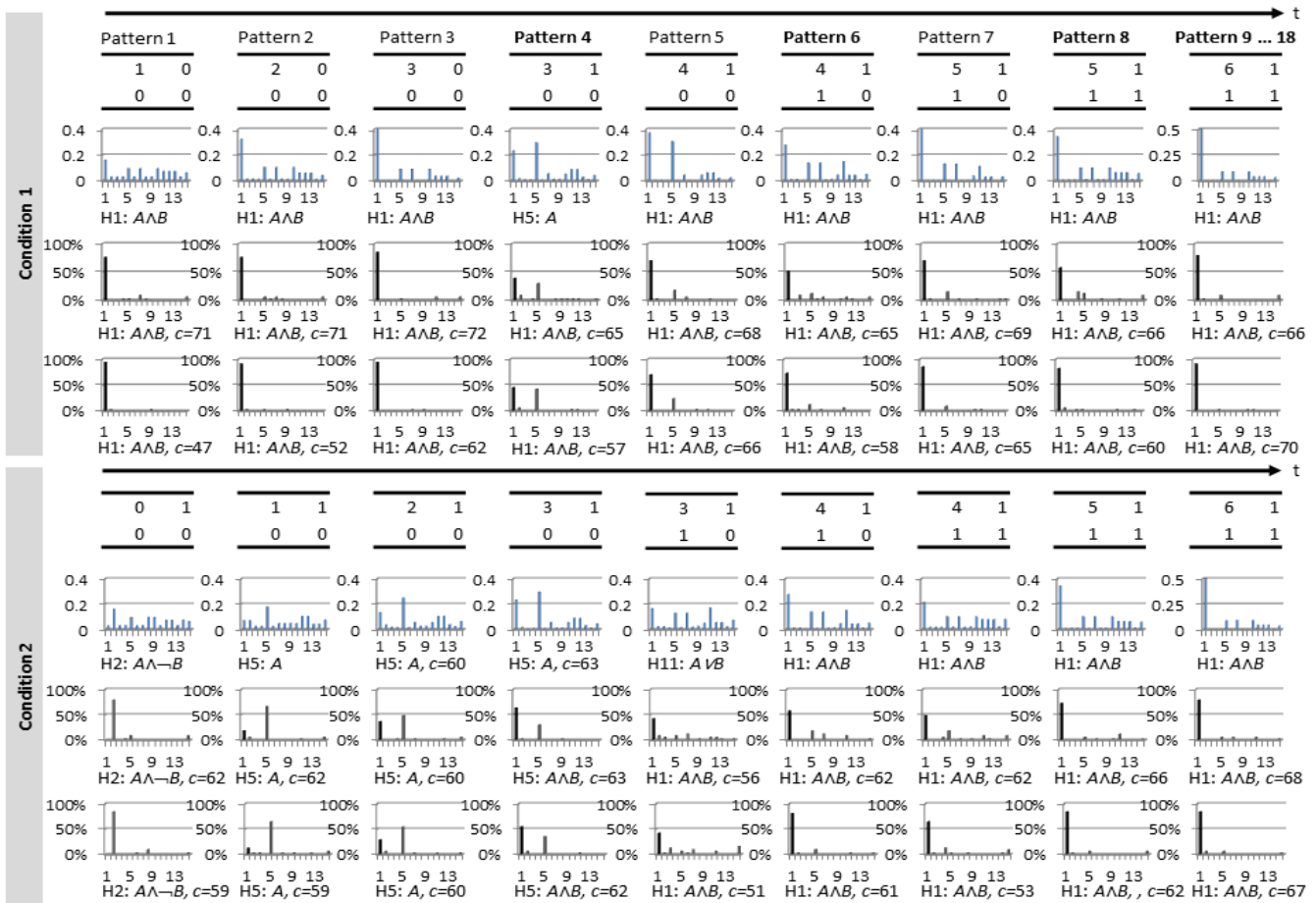


Figure 5: Patterns, predictions and results for Condition 1 and 2 (Phase 2 and 3). Within each condition, Row 1 shows the first nine shown patterns (Fig. 3, Phase 2 and 3, right). Row 2 depicts BL's pattern probabilities for 15 hypotheses (cf. Fig. 3). Row 3 and 4 show the portion of hypotheses actually selected to be most probable (in Phase 2 and 3).



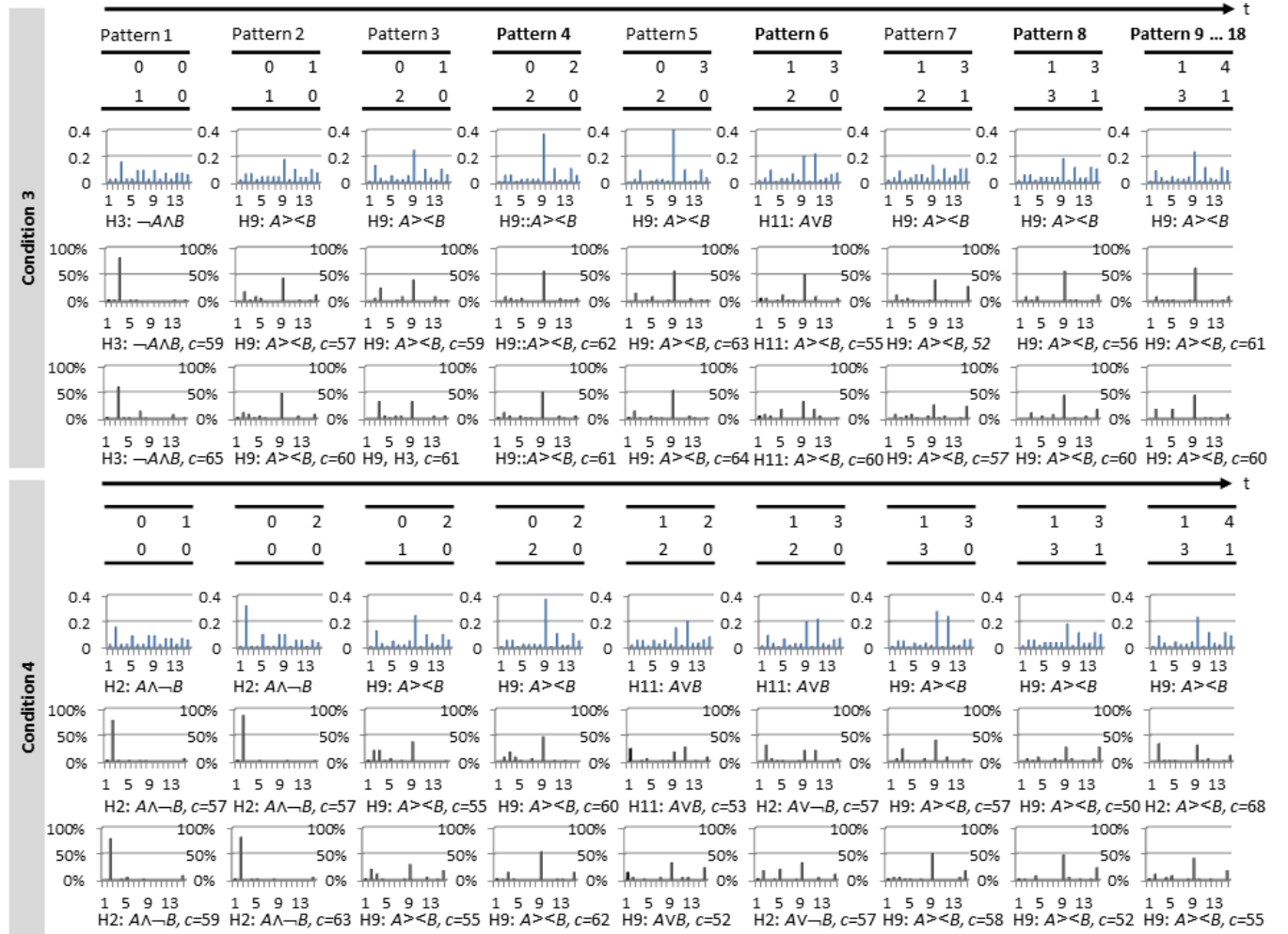


Figure 6: Patterns, predictions and results for Condition 2 and 3 (cf. Fig. 4 and main text for details).

## Results

Figure 4 shows the presented data patterns, the predicted pattern probabilities (BL), and the empirically found frequency of selected logical hypotheses for the six shown species of apes. Participants for each pattern actually selected the hypothesis that had the highest pattern probability,  $P_P(A \text{ o } B)$ ; from left to right: H1 ( $A$  and  $B$ ); H5 ( $A$ ); H8 (not  $B$ ); H9 (either  $A$  or  $B$ ), H11 ( $A$  or  $B$  or both), H15 (everything is possible). If one extended other theories so that they may predict these connectives, one would presumably not be able to explain the data (cf. von Sydow, 2010). For instance, the interesting support theory would make predictions for Pattern 6 [8, 10, 9, 9] based on the five other patterns (resulting in sum in [30, 29, 25, 10]). The highest support is suggested for the  $d$ -cell (H4) which is actually found only rarely. The strongest deviation from BL is observed in Pattern 5 where participants did *not only* select H11 but also H1. But this needs not to refer to an alternative strategy, but perhaps – and without elaborating this here – with a noise prior excluding deterministic patterns and causing the actual outcome (cf. von Sydow, 2011b).

With regard to Phase 2 and 3, Figures 5 and 6 show for all conditions the presented data sequence, the resulting BL predicted probabilities, and the actually observed frequencies of the selections of the most probable hypotheses. Even for the low trial numbers 1 to 9, reported here, the main selections are generally surprisingly in line with the pattern probabilities (presented without any fitting).

There were only small deviations. For instance, in Condition 1 only in Pattern 4 the predicted mode of answers (H5:  $A$ ) differed from the observed one (H1:  $A \wedge B$ ). However H1 has actually the second highest pattern probability and there may again have been a plausible influence of noise priors resulting from Phase 1 (lowering  $P(R = 0)$ ), which would actually increase  $P_P(H1)$ . This would likewise be coherent with Pattern 6 [4 1 0 0], where a surprisingly clear majority choose the AND-hypothesis (H1) and the extensional answer would be the  $A$ -hypothesis (H5).

The patterns that were kept identical in the corresponding homogeneous and heterogeneous conditions (the bold printed Patterns number 4, 6, 8, and 9) mostly corroborated the same results, suggesting that if memory effects are ruled out (as done here), no or only small effects of homogeneous

versus heterogeneous conditions are obtained. Furthermore, as predicted based on BL, the results were more pronounced for the conjunction conditions than for the exclusive disjunctions. Finally, the outcomes of Phase 2 and 3 did not differ much (or the results for BL even improve over time).

The confidence ratings varied less clearly than expected. One reason may be that this measure reflects not only, for instance,  $P_H(H \text{ most probable})/P_P(H \text{ second most probable})$ , but a general belief in a system of answers corresponding to BL or extensional probabilities. Furthermore, the ratings, averaged over all participants, may not be diagnostic, since they include ratings of unpredicted hypotheses (particularly relevant in C3 and C4). However, at least in the second trial-by-trial phase (Phase 3) participant's confidence ratings roughly corresponded to predictions derivable from BL: In C1 confidence increases from Pattern 1 to 3. In Pattern 9 the confidence is higher than in all previous patterns (despite more outliers). For Condition 3 and 4 the ratings show less differences, as is understandable based on pattern probabilities. Nonetheless, if one additionally takes a look at the next repeated nine trials, not reported here, Trial 18, for instance, confirmed a high confidence, leading to a median of 80 in C3 and 70 in C4. Hence, also the confidence ratings, at least in Phase 3, strongly reflect changes coherent with BL.

## Discussion

The results show correspondence with the predictions of BL also in trial-by-trial probability judgment tasks. Although other models of the CF have not been extended to all other connectives, it seems implausible that they could account for the findings (cf. von Sydow, 2009, 2011a). Without being able to discuss this here, some deviations (but clearly not all findings) may be coherent with a model that I have previously called pattern support, combining the pattern idea of BL with the idea of support. Overall, however, the results provide additional evidence for the predicted class of frequency-based CFs and for BL as a (computational level) psychological model for noisy-logical relationships.

Furthermore, as expected the results show no (or only a small) top-down effects of verbalization of hypotheses about the same situation (homogeneous vs. heterogeneous conditions). In the future it will be interesting to investigate identical settings without memory hooks (without summary statistics in Phase 2 and 3). Then verbalization may well effect represented exemplars (cf. von Sydow, 2011b). A further line of future research should be to investigate the role of noise priors on the selection of hypotheses.

## Acknowledgments

The first author has been the main author of this paper. We thank H. Wilke and J. Frisch for support. The first author is grateful for inspiring discussions at the Universities of Göttingen (M. R. Waldmann, R. Mayrhofer, Y. Hagmayer) and Heidelberg (group of K. Fiedler). This work was supported by the grant Sy 111/2-1 to M. von Sydow from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program *New Frameworks of Rationality* (SPP 1516).

## References

- Chater, N., Tenenbaum, J., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Science*, 10, 287-291.
- Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, Confirmation, and the Conjunction Fallacy – Theoretical Note. *Thinking and Reasoning*, 14(2), 182 - 199.
- Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, 50, 123-129.
- Gigerenzer, G. (1996). On narrow norms and vague heuristics. *Psychological Review*, 103, 592-596.
- Hagmayer, Y., Meder, B., von Sydow, M., & Waldmann, M. R. (2011). Category Transfer in Sequential Causal Learning: The Unbroken Mechanism Hypothesis. *Cognitive Science*, 35, 842-873.
- Hertwig, R., Benz, B., & Krauss, B. S. (2008). The conjunction fallacy and the many meanings of and. *Cognition*, 108, 740-753.
- Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psych. Bulletin*, 118, 248-271.
- Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S. & Caverni, J.-P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, 106, 62-88.
- Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, 36, 210-226.
- Lagnado, D. A. & Shanks, D. R. (2002). Probability judgment in hierarchical learning: A conflict between predictiveness and coherence. *Cognition*, 93, 81-112.
- Mellers, B. A., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, 12, 269-275.
- Nilsson, H., Juslin, P., & Olsson, H. (2008). Exemplars in the mist: The cognitive substrate of the representativeness heuristic. *Scandinavian Journal of Psychology*, 49, 201-212.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality. The probabilistic approach to human reasoning*. Oxford: Oxford University Press.
- Schurz, G. (2005). Non-monotonic reasoning from an evolutionary viewpoint: ontic, logical and cognitive foundations. *Synthese*, 146, 37-51.
- Sides, A., Osherson, D., Bonini N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory and Cognition*, 30, 191-198.
- Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions. *Organizational Behavior and Human Processes*, 91, 296-309.
- Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science*, 28, 467-477.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293-315.
- von Sydow, M. (2009). On a General Bayesian Pattern Logic of Frequency-Based Logical Inclusion Fallacies. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society* (pp. 248-253). Austin, TX: Cognitive Science Society.
- von Sydow, M. (2011 a). The Bayesian Logic of Frequency-Based Conjunction Fallacies. *Journal of Mathematical Psychology*, 55(2), 119-139.
- von Sydow, Momme (2011 b). Logical Inclusion Fallacies - Transfer of Logical Patterns and Noise (pp. 1-6). In: Kokinov, B., Karmiloff-Smith, A., Nersessian, N. J. (eds.). *European Perspectives on Cognitive Science*. New Bulgarian University Press.
- von Sydow, M., Hagmayer, Y., Meder, B. & Waldmann, M. (2010). How Causal Reasoning Can Bias Empirical Evidence. *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society* (pp. 2087-2092). Austin, TX: Cognitive Science Society.
- Wedell, D. H., & Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type. *Cognition*, 107, 105-136.



# Color word learning is a gradual inductive process

**Katie Wagner**

kgwagner@ucsd.edu  
Dept. of Psychology, UCSD

**Karen Dobkins**

kdobkins@ucsd.edu  
Dept. of Psychology, UCSD

**David Barner**

barner@ucsd.edu  
Dept. of Psychology, UCSD

## Abstract

Most current accounts of color word acquisition propose that the delay between children's first production of color words and adult-like understanding is due to problems abstracting color as a domain of meaning. Here we present evidence against this hypothesis, and show that, from the time children produce color words in a labeling task they use them to represent color. In Experiment 1, an analysis of early color word production errors finds that, before acquiring adult-like understanding, children make systematic hypotheses about color word meanings, which are best characterized as overextensions of adult meanings. Experiment 2 analyzes comprehension errors and finds that these overextensions result from overly broad categories, rather than a communicative strategy. These results indicate that the delay between production and adult-like understanding of color words is largely attributable to the problem of determining language-specific color boundaries.

**Keywords:** Concepts and Categories; Language Acquisition; Cognitive Development

## Introduction

Color words like *red*, *green*, and *blue* pose a difficult problem to children learning language. According to early reports from the turn of the 20th century, children did not acquire the meanings of color words until as late as 8 years of age. Recent reports suggest that children now acquire color words earlier around 3 or 4 years of age (possibly due to early education, see Shatz et al., 1996), but nevertheless struggle to learn them (e.g., Backscheider & Shatz, 1993; Sandhofer & Smith, 1999). The primary evidence of children's difficulty is that, similar to the domains of number (Wynn, 1990) and time (Shatz et al., 2010), children produce color words well before they use them with adult-like meanings. Also, it's often argued color word use is initially "haphazard and inconsistent" (p.70, Pitchford & Mullen, 2003). By most current accounts, this delay between production and adult-like understanding is caused by a difficulty abstracting color as a dimension of linguistic meaning (e.g., O'Hanlon & Roberson, 2006; Kowalksi & Zimiles, 2006; Sandhofer & Smith, 1999). Here we present evidence that children's initial use of color words is in fact systematic rather than haphazard, and that children have abstracted color by the time they begin using color words. We argue that the main source of children's delay is the problem of inferring category boundaries for color words.

Current accounts of color word acquisition typically assume that once children have conceptualized color as a domain relevant to word meaning, the mapping of color words to their target color categories proceeds quickly.

According to some, the moment of abstraction resembles a conceptual epiphany. For example, according to Franklin (2006), "Children seem to struggle with their first color word yet learn most of the other basic terms fairly rapidly over the next several months.... This seems to suggest that there is some kind of 'switch' for children's ability to learn and map color words correctly—" (p. 324). On this view, once children have mapped their first color word, mapping of other color words to adult-like meanings is relatively simple and fast (see Franklin, 2006; Soja, 1994).

The idea that the mapping process ought to be rapid comes from two main lines of research. First, in an often-cited study of color words in 110 languages, Kay and colleagues reported evidence for cross-linguistic universals in linguistic color categories (Kay et al., 2009) and argued that the number of color categories cross-linguistically is relatively small and constrained. Second, there is mounting evidence suggesting that pre-linguistic infants possess perceptual color categories very similar to those found in adults (e.g., Bornstein, Kessen & Weiskopf, 1976; Bornstein, 1976; Franklin et al., 2008; Franklin et al., 2005). In each case, the purported existence of constraints on language and perception have led researchers to conclude that color word learning is a simple mapping problem, whereby largely innate perceptual categories are associated with labels provided in language input.

Examples of this view are common in the literature, with important consequences for how color word learning is studied. For example, according to Pitchford and Mullen (2004), "Developmental studies have shown young children's perceptual colour space is organized in a similar manner to that of the adult... Thus, when children engage in the learning of colour terms, they already possess colour percepts on which colour concepts can be mapped." (p.53) The implication of such arguments is that, because color words can be mapped to pre-existing perceptual categories, the lag between production and adult-like understanding must not be due to the problem of determining boundaries. Instead, the delay must be due to the prior problem of identifying color as a domain of linguistic meaning.

There are good reasons, however, to believe that the acquisition of color words is not a simple mapping problem. Despite being restricted by universals of human perception, languages vary both in the number of color words they have (2 to 12) and how these words encode color (Kay et al, 2009). For example, some languages that have four basic color terms mark a boundary between red and yellow (e.g., Culina, spoken in Peru; Waorini, spoken in Ecuador) whereas others do not (e.g., Chácobo, spoken in Bolivia; Múra-Pirahã, spoken in Brazil; Kay et al.,

2009). Also, Russian, Korean and English have roughly the same number of basic color terms (11-12; Berlin & Kay, 1969), but divide the blue-green region of color space differently (e.g. Roberson et al., 2009; Winawer et al., 2007). In sum, while infants may perceive color like adults, the categories encoded by language are not fully determined by perception, suggesting that inductive learning plays a significant role in color word learning.

In the present study, we explored the idea that children acquire preliminary meanings for color words well before they converge on adult-like meanings, and thus abstract color much earlier than typically thought. On this hypothesis, the delay between color word production and adult-like understanding is mostly due to a gradual process of determining language-specific color word boundaries.

Past studies have typically failed to address the nature of the mapping problem because of how they characterized children's color word meanings. For example, researchers often classify children according to their knowledge of adult-like meanings – e.g., using *red* to label only red objects (e.g., Kowalski & Zimiles 2006; Soja, 1994; O'Hanlon & Roberson, 2006). In doing so, such studies may underestimate children's color word knowledge, and thus the point at which they first abstract color. Consistent with this concern, a number of studies have found that before children acquire all 11 adult-like color word meanings, they make errors that are systematic in nature (Pitchford & Mullen, 2003; Davies et al., 1998; Bartlett, 1977). For example, Pitchford and Mullen found that 3-year-olds often use their color words to incorrectly label hues adjacent to the target category (e.g., labeling orange as *red*). On the basis of this, they argued that pre-linguistic perceptual categories strongly constrain early color word meanings. However, proximity errors are not easily explained by this hypothesis. Instead, such errors most strongly support the existence of categories that are broader than those used by adults, and thus that are not acquired on the basis of pre-defined perceptual categories.

In the current study, we investigated the first meanings that children assign to color words by analyzing the errors they make in both language production and comprehension. Although some past studies have reported error data in early color word use (see above), here we present new evidence and analyses that directly address the nature of the delay between production and adult-like understanding. In Experiment 1, we present data from a color-labeling task sampled from a large group of children including a subset who have not yet acquired any adult-like meanings. This experiment finds that children make errors that are systematic in nature prior to acquiring any adult-like meanings. These data suggest that children in our study have abstracted color and possess partial knowledge of color words by at least the time they begin producing them. In Experiment 2, we corroborate these findings using a language comprehension task, and show that children's early overextensions of color words reflects overly broad meanings, rather than a communicative strategy.

## Experiment 1

### Methods

**Participants** 141 children (68 girls) participated. Children with a 25% chance or higher of protanopia or deuteranopia color deficiency (based on family history) were excluded (n=5). Children who made no errors (n=21), used only one color term (n=6) or did not cooperate (n=11) were also excluded. Data from the remaining 98 children (50 girls) were analyzed (mean age 3;0, range 22 to 61 months).

**Stimuli** Stimuli were constructed using 11 pieces of colored posterboard, which were chosen by a consensus of five experimenters as being prototypical of the 11 basic color terms in English (i.e., *red, orange, yellow, green, blue, purple, pink, brown, black & gray*). The posterboard was cut into a set of 11 fish shapes (Fish Task) and a set of 11 squares (Book Task). For the Fish Task the colored fish were glued to black foam and were presented on a black background. For the Book Task, the colored squares were glued onto black pages and covered with white flaps.

### Procedure

**Fish Task.** Each child was presented with a black box containing the 11 colored fish, placed color-side down. The experimenter (E) began by announcing, "My turn!" and randomly picking up one of the fish asking, "What color is it?" After the child responded, E placed the labeled fish on the table and told the child, "Your turn!", indicating that the child should pick up a fish. E and the child continued taking turns until each fish had been selected and labeled.

**Book Task.** Following the Fish Task, E presented the child with a book that contained the colored squares. For each page, the child lifted the flap that covered the color and E asked, "What color is it?" Colors were presented in the following order: orange, blue, yellow, pink, white, purple, gray, brown, green, red, black. When children did not respond, E repeated the question and gave the child another chance to respond. Trials with no response (103 trials, 4.7%) or with two responses (e.g., the child said both *blue* and *red*, 13 trials, 0.05%) were not analyzed.

### Results

**Color-Knowledge Groups** Children were separated into four groups based on the number of Basic Color Terms they used in an adult-like manner (e.g., using *red* consistently and exclusively for red stimuli).

Level 1: Adult-like knowledge of 0 color terms. Produced between 2 and 6 color terms during experiment (mean=3.1). Mean age of 2;5 (n=8, 1 girl).

Level 2: Adult-like knowledge of 1-3 color terms (mean=2.0). Produced between 3 and 9 color terms during experiment (mean=6.6). Mean age of 2;8 (n=16, 5 girls).

Level 3: Adult-like knowledge of 4-6 color terms (mean=5.1). Produced between 8 and 10 color terms during experiment (mean=9.1. Mean age of 3;2 (n=19, 9 girls).

Level 4: Adult-like knowledge of 7-9 color terms (mean=8.2). Produced between 9 and 12 terms during experiment (mean=10.3). Mean age of 3;2 (n=53, 34 girls). **Error Consistency Analysis.** Given that a child used an incorrect label for a particular stimulus color on one task (using *red* to label the orange stimulus on the Fish task), we asked how likely it was for the child to repeat the error on the other task (using *red* to label orange on the Book task). Using a binomial test, we asked whether the proportion of consistent trial pairs was greater than chance.

For this analysis, we excluded trial pairs in which the child labeled the stimulus correctly on both tasks (725 pairs). The remaining 275 pairs were classified as either consistent (the same incorrect label for a color on both tasks, 122 pairs) or inconsistent (153 pairs).

The probability of repeating any single label on two separate trials was defined as the square of the proportion of total trials a child used that label. For example, if a child used *red* on 6 of 22 trials, that child's probability of using *red* incorrectly on both the orange fish trial and the orange book trial was  $(6/22)^2$ . A child's overall chance probability of consistency was defined as the sum of the probability of repeating each label. Each color knowledge group's overall chance probability of consistency was defined as the average of the individual participant probabilities, weighted by the number of data points each individual contributed to the analysis. In other words:

$$p(\text{consistency}) = \sum_c \sum_j \left(\frac{l_j}{n}\right)^2 \left(\frac{i_c}{i}\right)$$

where  $i$  is the total number of stimulus pairs in which at least one label (either book or fish) was incorrect,  $i_c$  is the number of such incorrect pairs that each child,  $c$ , contributed to the analysis,  $l_j$  is the number of times a child produced each label  $j$  and  $n$  is the total number of responses a child produced. Note that by this definition, the chance probability of consistency appropriately decreases as a child adds more color words to his/her lexicon.

Averaged across the different Color-Knowledge groups, the proportion of incorrect trial pairs that was consistent (0.44) was greater than expected by chance (0.28), using a binomial test,  $p < 0.001$ . When this analysis was conducted separately for the different Color-Knowledge groups, rates of consistency were greater than chance for all groups except Level 1, see Table 1.

Table 1: Chance and Observed Rates for Exp 1 Analyses

	Consistency		Overextension		Proximity	
	Chc	Obs	Chc	Obs	Chc	Obs
Level 1	0.57	0.61	0.23	0.72*	0.18	0.26*
Level 2	0.27	0.41*	0.062	0.73*	0.25	0.44*
Level 3	0.14	0.31*	0.028	0.75*	0.33	0.66*
Level 4	0.11	0.43*	0.024	0.81*	0.54	0.82*
Total	0.28	0.44*	0.054	0.76*	0.27	0.52*

**Overextension Analysis** This analysis asked whether, in some cases, children's color errors were overextensions of

adult color categories. For example, a child may correctly know that *red* refers to red objects, yet have a broader meaning for *red* than adults and overextend it to orange and yellow objects. Given that a child used a label incorrectly for at least one trial (e.g., using *red* to label an orange or yellow stimulus), we asked whether they used that label consistently for its target color (i.e., using *red* to label both the red book and red fish stimulus). In other words, when a child used the word *red* to label orange and yellow, was this a case of overextension of the *red* category? For this analysis, if the child did not produce responses for the target hue on both tasks, that color word was not included in the analysis (19 incidences). Using a binomial test, we asked whether the proportion of errors that reflect overextensions was greater than chance.

As noted in the consistency analysis (above), the probability of repeating any single label on two separate trials is the square of the proportion of total trials a child used that label. In contrast to the consistency analysis in which consistent use of any incorrect label to any color stimulus was sufficient, in order for an incorrectly applied *red* label to be classified as an overextension a child must specifically use *red* (not any other color) in response to the red stimulus in both tasks. To calculate chance for this analysis, we first squared the base rates of every term a child used incorrectly (e.g., using *red* for purple) to calculate the probability that each of these incorrect terms would also be used on both trials containing the correct color stimulus (e.g., red fish and red book trial). We then took the mean of these probabilities to calculate the child's overall probability of overextension. To calculate the overall probability of overextension for each group, we calculated a group mean, weighted by how many labels each child used incorrectly. In other words:

$$p(\text{overextension}) = \sum_c \frac{i_c}{i} \frac{\sum_j \left(\frac{i_{cj}}{n}\right)^2}{i_c}$$

where  $i$  is the total number of labels that were used incorrectly at least once,  $i_c$  is the number of such incorrect labels that each child,  $c$ , contributed to the analysis,  $i_{cj}$  is the number of times a child produced each incorrect label  $j$ , and  $n$  is the total number of responses a child produced.

A very large proportion of children's errors – 0.76 – were overextensions, which was significantly greater than expected by chance (chance = 0.054), as measured by a binomial test ( $p < 0.001$ ). This high proportion indicates that if a child produced a color word, they were very likely to use it correctly when presented with its target hue. Thus, it suggests that most of children's errors were overextensions of color terms that were anchored to adult-like focal hues. Critically, rates of overextension were statistically greater than chance and above 0.72 for each Color Knowledge Group (all  $ps < 0.001$ , see Table 1), including children who had no adult-like color meanings (Level 1).

**Proximity Analysis** The overextension analysis indicates that before children acquire adult-like color word meanings, they use color words correctly for their target hues. However, the analyses described so far do not

address errors were made to proximal colors or to distant ones. An error was considered proximal if the stimulus color and the correct referent color for the misused label were from adjacent categories in Munsell Color Space (Long & Luke, 2001). Critically, although overextension to distant hues would be consistent with a gradual inductive process (e.g., since children's initial categories may be very large), the inclusion of proximal hues should be significantly more likely even in this case, since any category that includes red and yellow, for example, should also include intervening hues like orange.

Given that a child used a color label incorrectly, we asked whether the label and its referent stimulus were from perceptually adjacent categories. Using a binomial test, we asked whether the proportion of errors that were from proximal color categories was greater than expected by chance. Chance was defined using both the frequency with which children made errors for each stimulus and the frequency with which they used each label incorrectly. It was necessary to account for these base rates because some color words are proximal to a greater number of color categories than others. For example, red is considered proximal to orange, pink, purple, and brown, while gray is proximal to black and white. To determine chance, we calculated the probability of each label-stimulus error pair (the probability of using *red* for an orange stimulus) as equal to the product of the base rates. For example, if 20% (0.2) of errors were in response to an orange stimulus and 80% (0.8) of errors involved the label *red*, the probability of using *red* to label orange would be  $0.2 \times 0.8$ , or 0.16.

To compute the chance probability of proximal errors, we summed across the probability of all label-stimulus pairs that are classified as proximal. In other words:

$$p(\text{proximity}) = \sum_i \sum_j p(r|l_j \cap s_i) p(l_j|\text{incorrect}) p(s_i|\text{incorrect})$$

where,  $p(s_i|\text{incorrect})$  is the probability of a particular stimulus  $i$  given an incorrect response;  $p(l_j|\text{incorrect})$  is the probability of a particular elicited label  $j$  given an incorrect response to stimulus  $i$ ; and  $r$  is the probability of proximity. Note that  $p(r|l_j \cap s_i)$  is either 1 or 0 because a given label/stimulus pair is either proximal or not.

The proportion of total errors that were proximal was 0.52. This was significantly greater than the rate predicted by chance (0.27),  $p < 0.001$ . Rates of proximity were statistically greater than chance for all Color-Knowledge groups, including Level 1 children who had no adult-like color word meanings (see Table 1). Like the findings of the overextension analysis, this indicates that even before a child acquires the adult meanings of any color terms, they already have partial knowledge of some color words.

## Discussion

Experiment 1 examined children's color word production errors in early acquisition. Our results revealed that if children used a word in the study, they were very likely to have a systematic meaning for the word, despite the fact that these meanings were often non-adult-like in nature. Together, these results suggests that (1) children

have abstracted color around the time they begin using color words to label stimuli, and thus well before they acquire their adult-like meanings, and (2) they learn color words by making overly broad hypotheses about their meanings, and gradually narrowing these meanings as they acquire additional, contrasting words.

Several pieces of evidence support these conclusions. First, in our Error Consistency Analysis we found that all but the Level 1 children made highly consistent in their errors, demonstrating that these children were able to abstract color across different objects despite their other differences, and use this knowledge to formulate hypotheses about color word meanings. Although Level 1 errors were not consistent, these children were nonetheless highly systematic, as shown by our two other analyses. In our Overextension Analysis we found that, at all color-knowledge levels, a majority of children's errors were overextensions. This indicates that children have partial knowledge of the specific color properties denoted by a color word when they first begin producing it. Specifically, children appear to know the focal color denoted by the color words they use, though they frequently overextend these words. Finally, our Proximity Analysis revealed that the errors made by children at all levels were likely to be labels for perceptually similar colors.

In sum, the data from Experiment 1 demonstrate that children with adult-like understanding of no color words have nonetheless abstracted color. These data are consistent with the idea that children begin acquisition of color words by positing overly broad color categories and that these categories are gradually narrowed as children gain experience and acquire other color words that contrast in meaning. We refer to this as the "Broad Color Categories" hypothesis. Another possibility, however, is that overextension errors instead reflect a pragmatic strategy (for a similar discussion in the domain of nouns, see Clark 1978). For example, imagine a child who has an adult-like meaning for *red* but not for *orange*. When presented with an orange stimulus, the child may recognize that this color is not *red*, but use the word *red* to describe it nonetheless since no better word is available to them. We refer to this as the "Communicative Strategy" hypothesis.

One way of testing these hypotheses is to use a comprehension task, where the experimenter selects the label, thereby removing the possibility of communicative overextensions (Clark, 1978, Gelman et al. 1998). If a child has a broad meaning for *red* (that includes both red and orange), when asked for *red*, they should provide a stimulus that satisfies their meaning of *red* (e.g., either a red or orange one). By contrast, if the child has adult-like color categories, when asked for *red* they should always prefer a red stimulus over an orange one (even if they use *red* to label orange as a communicative strategy).

## Experiment 2

In Experiment 2 we presented children with stimuli identical to those used in the Fish task in Experiment 1 and

asked them to find fish of different colors and asked whether children made proximity errors. Note that by the Communicative Strategy hypothesis, proximal errors should not occur because responding should either be correct for known color words (above example, *red*) or random for unknown color words (above example, *orange*) because children do not possess broad categories (e.g., that include both red and orange). Proximal errors are only expected under the Broad Category Hypothesis, since it claims that children possess linguistic categories that include multiple adult categories (e.g. *red* including both red and orange).

## Methods

**Stimuli** Items were those of the Fish Task in Experiment 1.

**Participants** A total of 28 children (14 girls) participated. Children were screened for color deficiency via a family history questionnaire. Eight children were excluded because they made no errors. Data from the remaining 20 children (8 girls) were analyzed. These children ranged in age from 23 to 48 months (mean=2;10). Unlike in Experiment 1, we did not group children into different Color-Knowledge groups, for two reasons. First, children in this study were not asked to produce color labels, and we therefore could not test how many color terms they knew. Second, the number of subjects required to test the hypothesis of this experiment was relatively small, and thus there was insufficient power to analyze subgroups.

**Procedure** Children were presented with the fish stimuli placed color-side up and in a random configuration. In succession, the experimenter (E) asked the child to find a specific colored fish, "Give me a (*red*) fish. Can you put a (*red*) fish in my hand?" After the child handed a fish to E, it was returned to its place on the table (back with the other fish), and E requested the next color fish. Colors were requested in the order of Experiment 1: *red, brown, green, orange, white, blue, gray, pink, black, yellow, and purple*.

If the child did not respond on a particular trial (e.g., they got distracted), E repeated the question, giving the child an additional opportunity to respond. Trials with no response ( $n = 3$  trials) or on which two or more fish were provided ( $n = 1$  trial) were not included in the analysis.

## Results and Discussion

**Proximity Analysis.** Across all 216 trials, 79 trials (36%) were errors and were included in the analysis. The mean number of correct responses was 6.85 (range 0 to 10).

As in the proximity analysis of Experiment 1, we accounted for the base rate of errors that involved each color stimulus and the base rate of errors made to a particular request (e.g., *red*). Consistent with the results of Experiment 1, the proportion of errors that were proximal in the comprehension task was 0.58. This was significantly greater than the rate predicted by chance (0.30),  $p < 0.001$ . This suggests that the systematic production overextensions in Experiment 1 reflected broad color categories rather than a communication strategy.

## General Discussion

We tested the hypothesis that the delay between color word production and acquisition of adult-like meanings is due to the gradual construction of linguistic color categories, rather than the process of abstracting color as a domain of word meaning. Consistent with this idea, we found that if children used color words in our study, they typically used them in a meaningful and consistent way. When children made errors, the vast majority (75%) were overextensions of adult-like categories. Also, these errors were frequently to proximal hues. This was true for children at all levels of color word competence – even those who had no adult-like meanings. Further, the results of a comprehension task corroborated this finding, and indicated that overextension is not the product of a communicative strategy, but instead reflects broad linguistic color categories.

These results have important implications for our understanding of color word acquisition. First, contrary to past reports, the results suggest that children abstract color at the earliest stages of color word production, and that there is little, if any, lag between children's use of color words as labels and their construction of preliminary meanings for these words. Although abstraction may pose a problem to children early in acquisition, it is likely resolved by the time children begin using colors to label things in their environment. Second, our results suggest that the observed delay between production and adult-like understanding of color words is likely due to the problem of constructing language-specific category boundaries. Our data suggest that children begin acquisition by making overly broad inductive inferences regarding the scope of their color words, and that they gradually shrink their early categories as they gain experience with the words and acquire other color words that contrast in meaning.

These data are consistent with earlier data from Carey and Bartlett (1978), which are commonly cited as evidence for children's ability to "fast map" color words to their referents. Bartlett and Carey's fast mapping proposal, unlike some theories of color word learning that followed it, did not assume that learning color word meanings was fast, or that it was a simple mapping problem. Instead, they argued that fast mapping was a first step in the learning process, used to link labels to particular referents, and that acquiring the adult-like meanings of color words involves much more additional learning (see also Swingley, 2010; Clark, 1997). Consistent with this, many of the children in Carey and Bartlett's study used the novel word *chromium*, which was used by experimenters to refer to an olive-colored stimulus, to refer to perceptually similar colors (e.g., green, brown) and often did not converge on the intended narrower meaning for *chromium* even after many trials (also, see Bartlett, 1977; Pitchford & Mullen, 2003).

These results, like ours, suggest that color word learning is a gradual inductive process, and that children form interim meanings for their color words well before they attain adult-like understanding. However, because these

studies focused on samples of children who for the most part had acquired at least one adult-like color word, their data do not address the delay between onset of color word production and children's first adult-like color word meaning. In contrast, our study addressed this question using data from a wider range of children (including those who had not acquired any adult-like meanings of color words), and using a novel set of analyses that tested not only proximity errors but also consistency and overextension. Consequently, our study was able to show that children possess broad, overextended color categories early in acquisition, before they have acquired their first adult-like meaning, and perhaps even from the time they first begin producing color words.

This view of color word learning is consistent with findings in other domains of language and conceptual development. In the case of number, children quickly recognize that numerals form a class of words that contrast in meaning (Wynn, 1992; Brooks et al., under review), despite taking years to learn what these meanings are. Similarly, young children recognize that time words like minute, second, and hour form a lexical class, but take many years to acquire their individual meanings (Shatz et al., 2010). Finally, children produce emotion words from early in development, and understand that they belong to a class of words that describe human sentiment, but nonetheless take years to master their adult-like meanings, and form many interim hypotheses along the way (Widen & Russell, 2003). Our study suggests that the case study of color is not an exception to this general pattern, and that as in other cases that involve identifying abstract content, children begin formulating preliminary meanings early in acquisition, and take years to attain adult competence.

## References

- Backscheider, A. G. & Shatz, M. (1993). Children's acquisition of the lexical domain of color. In K. Beals et al. (eds), *What we think, what we mean, and how we say it, Papers from the parasession on the correspondence of conceptual, semantic and grammatical representations*, CLS 29, v. 2. Chicago: The Chicago Linguistic Society.
- Bartlett, E. J. (1978). The acquisition of the meaning of color terms: A study of lexical development. In R. N. Campbell & P. T. Smith (Eds.), *Recent Advances in the Psychology of Language* (pp. 89–108). NY: Plenum.
- Berlin, B. & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley, CA: University of California Press.
- Bornstein, M., Kessen, W., Weiskopf, S. (1976). Color vision and hue categorization in young human infants. *Exp Psychol Human*, 2, 115-129.
- Brooks, N., Audet, J., & Barner, D. (under review). Pragmatic inference, not semantic competence, guides 3-year-olds' interpretation of unknown number words.
- Carey, S. (2009). *The Origin of Concepts*. New York: Oxford University Press.
- Carey, S. & Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford Child Language Conference*, 15, 17-29.
- Clark, E. (1978). Strategies for communicating. *Child Dev*, 49, 953-959.
- Clark, E. (1997). Conceptual perspective and lexical choice in acquisition, *Cognition*, 64, 1-37.
- Davies, I., Corbett, G., McGurk, H. & MacDermid, C. (1998). A developmental study of the acquisition of Russian colour terms. *J Child Lang*, 25, 395-417.
- Franklin, A. (2006) Constraints on children's color term acquisition. *J Exp Child Psychol*, 94, 322-327.
- Gelman, S., Croft, W., Fu, P., Clausner, T. & Gottfried, G. (1998). Why is a pomegranate an apple? The role of shape, taxonomic relatedness, and prior lexical knowledge in children's overextensions of apple and dog, *J Child Lang*, 25, 267-291.
- Kay, P., Berlin, B., Maffi, L., Merrifield, W. R. & Cook, R. (2009). *The World Color Survey*. Palo Alto, CA: CSLI Press.
- Kowalski & Zimiles (2006). The relation between children's conceptual functioning with color and color term acquisition. *J Exp Child Psychol*, 94, 301-321.
- Long, J. & Luke, J. (2001). *The New Munsell Student Color Set, Second Edition*. Fairchild Books: New York.
- O'Hanlon, C. & Roberson, D. (2006). Learning in context: Linguistic and attentional constraints on children's color term learning. *J Exp Child Psychol*, 94, 275-300.
- Pitchford, N. & Mullen, K. (2003). The development of conceptual colour categories in pre-school children: Influence of perception on categorization. *Vis Cogn*, 10, 51-77.
- Roberson, D., Hanley, J.R. & Pak, H. (2009). Thresholds for color discrimination in English and Korean speakers. *Cognition*, 112, 482-487.
- Sandhofer, C. & Smith, L. (1999). Learning color words involves learning a system of mappings. *Dev Psychol*, 35, 668-679.
- Shatz, M., Behrend, D., Gelman, S., & Ebeling, K. (1996). Colour term knowledge in two-year-olds: evidence for early competence. *J Child Lang*, 23, 177-199.
- Shatz, M., Tare, M., Nguyen, S.P., Young, T. (2010). Acquiring non-object terms: The case for time words. *Journal of Cognition and Development*, 11, 16–36.
- Soja, N. (1994). Young children's concept of color and its relation to the acquisition of color words. *Child Dev*, 65, 918-937.
- Swingle, D. (2010). Fast Mapping and Slow Mapping in Children's Word Learning. *L, L & D*, 6, 179-183.
- Widen, S. & Russell, J. (2003). A closer look at preschoolers' freely produced labels for facial expressions, *Dev Psychol*, 39, 114-128.
- Winawer, J., Witthoft, N., Frank, M., Wu, L., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *PNAS*, 104, 7780-5.
- Wynn, K. (1990). Children's understanding of counting. *Cognition*, 36, 155-193.

# The Role of the Primary Effect in the Assessment of Intentionality and Morality

Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)

Alex Wiegmann (alex.wiegmann@psych.uni-goettingen.de)

Department of Psychology, University of Göttingen,  
Gosslerstr. 14, 37073 Göttingen, Germany

## Abstract

In moral dilemmas performing an action often leads to both a good primary and a bad secondary effect. In such cases, how do people judge whether the bad secondary effect was brought about intentionally, and how do they assess the moral value of the act leading to the secondary effect? Various theories have been proposed that either focus on the causal role or on the moral valence of the secondary effect as the primary determinants of intentionality and morality assessments. We present experiments which show that these theories have neglected a further important factor, the primary effect. A new theory is proposed that is based on the key assumption that people's judgments of intentionality and morality depend on the strength of assumed reasons the agent has for the primary and secondary effects.

Keywords: intentional action; doctrine of double effect; moral dilemmas; causal structure; trade-off.

## Introduction and Overview

How do people judge whether an effect was brought about intentionally? Since Joshua Knobe's seminal paper (2003) this question has gained increasing attention in recent cognitive science (see Waldmann, Nagel, & Wiegmann, 2012, for an overview). In the present paper we focus on the following question: If an agent in a moral dilemma performs an action that leads to two effects<sup>1</sup>, a good one, which is the primary goal of the agent, and a bad secondary one, how do subjects judge whether the bad effect was brought about intentionally?

Currently two very different theories are competing to answer this question. Knobe (2003) has proposed that intentionality attributions regarding the secondary effect depend on its moral value. If the secondary effect is morally bad, his theory predicts high intentionality ratings, whereas the ratings are lowered when the secondary act is good. The second class of theories focuses on the causal structure linking primary and secondary effects. If the secondary act is a means for achieving the primary one, then high intentionality ratings and low morality judgments are to be expected. When the secondary act is just a causal side effect, intentionality ratings are predicted to be lower, and morality judgments higher (see Mikhail, 2011).

We are going to propose a third account. Our main claim is that intentionality attributions are a function of the strength of the assumed reasons that can be attributed to the

agent for causing the primary and the secondary effect. These reasons are inferred on the basis of observable cues, with the inferences being influenced by both the causal structure of the scenario, and the trade-off between good and bad effects. Morality judgments are also influenced by the trade-off, although, based on previous research, we expect that moral judgments are influenced less by causal role than intentionality judgments (see Waldmann et al., 2012; Waldmann & Dieterich, 2007). We will outline these three theories in greater detail below, and then present two experiments testing them.

## Knobe's (2003) Side-Effect Effect

Theories of intentional action have gained a lot attention since Knobe (2003) had discovered that subjects rate the assumed intentionality regarding a secondary act higher when this effect is morally bad than when it is morally good. Consider Knobe's (2003) famous vignette:

"The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.' The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was harmed." (p. 191)

In a second version of this scenario, the word "harm" was replaced by "help". When subjects were asked whether they think the chairman intentionally harmed the environment, 82% answered in the affirmative. In contrast, in the help condition 23% said that the agent did bring about the good side effect intentionally. Knobe concluded that in judging whether the side effect (harming and helping the environment, respectively) was brought about intentionally, the (moral) value of the side effect is crucial. People seem considerably more willing to say that a side effect was brought about intentionally when they regard it as bad, than when they regard it as good.

In the present research we focus on moral dilemmas in which the secondary effect is kept invariant and constantly bad (killing one person). For such scenarios, Knobe's (2003) version of his theory entails two theoretical predictions: (1) As long as the secondary effect is invariant, equal degrees of intentionality ratings should be observed. Since in our cases, the secondary effect is bad, generally high ratings are to be expected. (2) The moral value of the secondary effect is an intrinsic property of the act leading to the bad outcome. For example, harming nature is bad, helping nature is good. Of course, Knobe's (2003) theory

---

<sup>1</sup> Following the literature we here use the term primary and secondary effect as referring to acts leading to a good or bad outcome (e.g., harming nature).



does not explicitly rule out that other factors, such as the primary effect, may play a role in the assessment of intentionality and moral value. However, the present version of the theory neglects the potential role of the primary effect, and solely focuses on the role of the secondary effect. This seems like a crucial oversight given that the primary effect is constitutive for assigning the other effect the role of being secondary. Moreover, it is far from clear whether this additional factor can be simply added to the present theory without changing key theoretical assumptions (see also General Discussion).

### The Doctrine of Double Effect

The Doctrine of Double Effect (DDE) is one of the oldest and best known moral principles (cf. Mikhail, 2011). There are several versions of the DDE, here is one by Timmons (2002): Whenever an action would produce at least one good effect and one bad or evil effect, then one is permitted to perform the act if and only if all of the following conditions are met:

1. The action in question, apart from its effects, must not be wrong.
2. The bad effect must not be intended by the agent. There are two principal ways in which an effect might be intended:
  - a) Any effect that is a chosen end of action is intended.
  - b) Any effect that is a means for bringing about some intended end is also intended.
3. The bad effect must not be “out of proportion” to the good effect.

To illustrate the DDE, let us apply it to two popular trolley cases (see Waldmann et al., 2012, for an overview of trolley research). In case one (“bystander”), a runaway trolley is threatening five people and the only possibility to rescue the five people is to re-direct the trolley onto another track where only one person would die. In case two (“push”), the initial situation is the same but this time the only possibility to rescue the five is to throw a heavy man from a bridge into the path of the trolley. The trolley would be stopped due to the weight of the heavy man and, not surprisingly, the heavy man would die in the collision. According to the DDE, redirecting the trolley might be permissible but throwing the heavy man from the bridge is not, since throwing the heavy man is a means for bringing about the intended end (2b).

Although the primary goal of the DDE is to offer a guide for the moral evaluation of moral dilemmas, the DDE also provides a criterion that tells us when an act or an effect is intended: Any effect that is a means for bringing about some intended end, is itself intended, whereas secondary effects that are only causal side effects are merely foreseen, but not intended. Cushman and Young (2010) have presented a series of studies in which they showed that means elicited higher intentionality ratings than side effects. This pattern holds for both moral dilemmas and isomorphic non-moral scenarios. Again, the primary effect does not figure in the predictions of intentionality. It does play a limited role in

permissibility judgments, however, but the DDE only states that the secondary bad effect must not be out of proportion to the good effect.

### Trade-Off of Lives in Moral Dilemmas

Whereas Knobe (2003) focuses on the moral value of the secondary effect in his predictions about intentionality ratings, the DDE focuses on its causal role (means vs. side effect). However, there is an additional neglected factor that is actually constitutive for the labeling of one of the effects as secondary: the primary effect. A typical feature of moral dilemmas is that they contain a trade-off between acts saving and killing people. For instance, in the standard version of the bystander dilemma, the act causes the death of one person and saves five persons. If you consider instead a version of this case in which acting kills one person but nobody is saved, it seems clear that causing the death of the one person was brought about intentionally because otherwise it is hard to explain why the agent should have intervened. However, if one person is killed and one is saved one might judge that saving the one person was the primary intended goal. If more lives are saved than killed it seems even more likely that subjects will view the good outcome as a mitigating reason for generating the bad side effect, and will therefore be less inclined to regard the bad effect as strongly intended. Thus, based on the assumption that intentionality ratings are influenced by the inferred reasons an agent might have for causing a bad secondary effect, the prediction can be derived that intentionality ratings should decrease the more reasons the agent has for causing the primary effect. In trolley dilemmas, this means that lower intentionality ratings with respect to the bad secondary effect should be expected, the more people are saved by the act. The number of saved people provides excellent reasons for acting, and allows the agent to dismiss the bad secondary effect as intentionally pursued.

The trade-off between primary and secondary effect may not only affect intentionality assessments but also moral evaluations. Again here our trade-off hypothesis makes predictions different from Knobe (2003) and the DDE. Whereas Knobe (2003) treats the badness of the secondary effect as an intrinsic feature of this effect, the DDE predicts that secondary effects should be judged generally worse when they constitute a means compared to a side effect. By contrast, our trade-off hypothesis claims that the judged badness of the secondary effect is not only based on intrinsic harmful features of the corresponding act, but also on the strength of reasons for accepting a bad secondary effect in light of a good primary goal. Thus, again we predict that moral evaluations will be sensitive to the relationship between the primary and secondary effects. Empirically this prediction is supported by research presenting trolley dilemmas with catastrophe conditions in which one person needs to be killed to save, for example, 1,000,000 (Nichols & Mallon, 2006; see also Bartels, 2008). Typically permissibility judgments rise with increasing

numbers of saved people. Intentionality judgments have not been studied in these experiments, however.

### The Role of Causal Structure

So far we have elaborated the role of the primary effect as a mitigating reason for accepting a bad side effect. However, reasons for acting are not only influenced by the size of the primary effect, but also by the causal relationship between primary and secondary effect. Following the DDE, we argue that secondary effects that play the role of means need to be intended more strongly than when they are in the position of side effects because agents should be aware of the fact that the secondary effect constitutes a necessary step on the path to the primary goal. Thus, causal structure also should determine the strength of reasons an agent has for causing a secondary effect.

### Combining Trade-off and Causal Structure

We have identified two sources of reasons an agent might have for acting in a moral dilemma in which a varying primary effect is pitted against an invariantly bad secondary effect: The causal role of the secondary effect and the trade-off between the primary and secondary effects. There are two possibilities how the two factors can be combined. One possibility would be to claim that the primary effect only matters when the secondary effect is a side effect, not when it is a necessary means for the primary effect. This pattern of interaction could be motivated by the assumption that means are necessary steps on the way to the more distant effects, and therefore need to be intended regardless of the quality of the distant (i.e., primary) effects. The other possibility might be that the strength of reasons for the primary goal affect intentionality ratings in both conditions, but to a stronger extent in the side-effect condition in which it is easier to imagine that people just foresee, but do not intend the secondary effect. In this case, we also expect an interaction but also a main effect that is driven by the strength of the primary goal.

The same alternatives also arise for moral judgments about the badness of bringing about the secondary effect (i.e., killing one person). Whereas Knobe (2003) predicts invariant ratings signifying the invariant badness of killing one person, the DDE predicts a main effect driven by the causal role of the act leading to death. By contrast, based on previous research we expect that mitigating factors arising from the evaluation of the goodness of the primary effect will also affect the moral assessment of the secondary effect. However, previous findings on moral judgments in trolley dilemmas cast doubt on the hypothesis that causal role plays the same role in moral judgments as in intentionality judgments. The empirical evidence rather points to other factors (i.e., aversiveness, attentional focus, directness) as the crucial factors affecting moral judgments whereas causal role vanishes as a factor once its confounds are controlled (see Waldmann et al., 2012). Thus, intentionality and moral judgments need not be driven by the same factors, as implied by the DDE.

## Experiment 1a

Experiment 1a focuses on intentionality attributions. To test our predictions we used two variants of trolley cases in which we manipulated the causal structure (means versus side effect) between subjects, and the number of lives saved (0, 1, 5, 100) within subjects. While the number of lives saved was manipulated across conditions, the secondary effect always involved killing one person.



Figure 1: Illustration of the bystander scenario (Experiment 1).

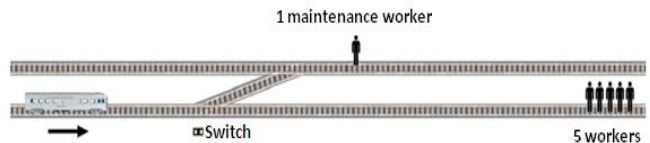


Figure 2: Illustration of the trap scenario (Experiment 1).

**Design, Materials, and Procedure** 140 subjects were recruited for a compensation of £0.50 via an online database located in the U.K.. Subjects were randomly assigned to the means or side-effect condition, and were then presented with the corresponding scenarios in which the number of lives saved was varied in a randomized order within subjects (0, 1, 5, 100 victims). It is important to note that the 0 condition only served as a control condition because in a condition in which one person was killed without anybody being saved, the means or side effect relations cannot be realized. This condition is expected to yield maximal intentionality and badness ratings because there are no mitigating reasons for the act.

The initial situation was identical in both conditions:

“On the test ground of a modern railroad property an unmanned speed train (that normally can be remote-controlled) is out of control due to a technical defect. This speed train is heading towards five railroad workers that are checking the tracks. Since these workers are wearing a novel type of hearing protection, they would not notice the speed train on time and hence would be run over by it. Peter, an employee of the rail track control center, recognizes the upcoming accident. However, it is not possible to stop the train on time anymore.”

Next the conditions, which were additionally illustrated (see Fig. 1, 2), varied (examples with 5 saved workers).

**Side-Effect Condition (“bystander”):**

“There is only one possibility to avoid the death of the five workers: Peter could push a button and thereby re-direct the speed train from the lower track onto a parallel upper track before it reaches the five workers on the lower track. On the upper parallel track the speed train would run into a worker maintaining the tracks. The maintenance worker on the upper track would lose his life due to the collision.”

#### Means Condition (“trap”):

“There is only one possibility to avoid the death of the five workers on the tracks: Peter could push a button that would open a trap door and thereby causing a maintenance worker on top of the bridge to fall on the tracks. The speed-train would collide with the maintenance worker and be stopped before it reaches the five workers on the track. The maintenance worker would lose his life due to the collision.”

Both scenarios ended as follows:

“Peter understands the situation and knows the consequences of the action just described. Peter decides to throw the switch and the maintenance worker dies.”

In the control condition (zero people saved) we used the same instructions mentioning re-directing of the train or the trap door but these acts were not motivated by saving anybody. After reading the scenario description, subjects were asked to judge whether Peter “caused the death of the maintenance worker intentionally” on a six-point Likert scale ranging from “certainly no” to “certainly yes.” On the last page, subjects were asked some demographic questions, and were given a simple logical question unrelated to the experiment to identify the subjects who did not pay attention to the task.

**Results and Discussion** Eight subjects were removed from the analyses because they failed to solve the logical question or completed the whole survey in less than a minute. The results for all scenarios are depicted in Figure 3.

We generally excluded the control condition in the ANOVAs of all experiments because here the difference between means and side effects could not be realized. Looking at the remaining three conditions, the intentionality ratings proved generally higher in the means than in the side-effect condition ( $F_{1, 130}=22.158, p<.0001$ ). Moreover, the analysis yielded a significant interaction,  $F_{2,260}=3.0911, p<.05$ .

More detailed planned comparisons showed that in both conditions intentionality ratings decreased with increasing numbers of lives saved. However, in the means condition the ratings were statistically equivalent in the three conditions in which the means relation could be realized (1, 5, 100). In contrast, in the three side-effect conditions (1, 5, 100) ratings dropped from 3.97 when one life was saved to 3.41 when one hundred lives were saved ( $p<0.001$ ). In both causal conditions, the control scenario in which nobody was saved by killing one received significantly higher ratings than the averaged ratings for the three other scenarios (means condition:  $p<0.01$ ; side-effect condition:  $p<0.0001$ ). Thus, in general the presence of a primary good effect lowers intentionality assessments regardless of causal status, but the influence of the quantitative size of the primary goal affects the side-effect condition more than the means condition.

Interestingly, subjects’ intentionality ratings for the cases in which 5 and 100 lives are saved did not differ. Possibly subjects are only sensitive to qualitative differences, that is, it only matters if fewer lives (0), just as much (1), or more

lives (5 and 100) are saved but not how many more are saved.

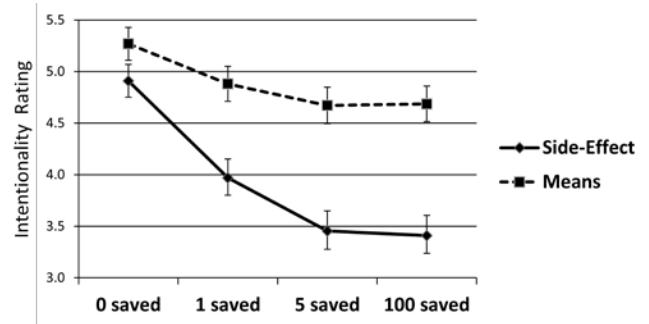


Figure 3: Results of Experiment 1a. Error bars indicate standard error of means.

### Experiment 1b

Experiment 1b uses the same conditions and manipulations as Experiment 1a. The only difference is that we asked a different test question with which we assessed the moral evaluation of the secondary effect: “How bad is Peter’s causing the one person’s death?” Subjects responded using a 6-point rating scale ranging from “not bad at all” to “very bad.”

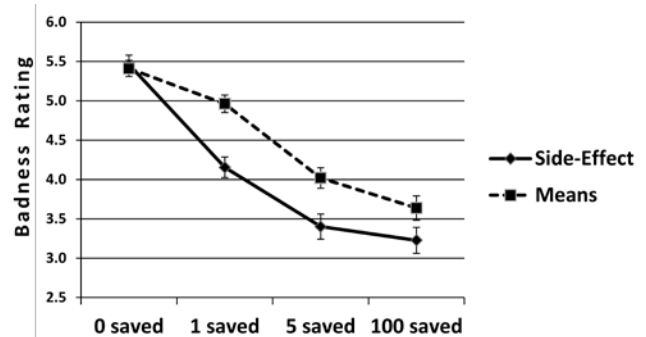


Figure 4: Results of Experiment 1b. Error bars indicate standard error of means.

**Results and Discussion** The results are based on 147 participants from the same online site used in the previous experiment (29 from the larger sample needed to be removed). Figure 4 shows the results. Excluding the control condition (0 people saved) which yielded uniformly high badness ratings, we found a main effect between the side-effect and means condition,  $F_{(1,145)}=7.43, p<.01$ , along with a main effect of the size of the primary effect,  $F_{(2,290)}=56.004, p<0.0001$ . Thus, in contrast to the predictions of the DDE, the primary effect affects moral evaluations. The main effect between side-effect and means conditions is consistent with the predictions of the DDE, but could also indicate differences in aversiveness of death (Waldmann & Dieterich, 2007). Experiment 2 will address this question.

## Experiment 2a

One might object that in Experiment 1a the act in the means condition is more aversive than the one in the side-effect condition (falling through a trap-door and getting hit by a train vs. just getting hit by a train). In fact, the moral judgment data revealed that the trap dilemma was generally assessed as more aversive than the bystander dilemma. Hence, the difference in intentionality ratings could be caused by this factor and might not be due to different causal structures. To counter such an objection we designed a new experiment with more similar scenarios. In Experiment 2a we focused on intentionality assessment, in Experiment 2b on morality judgment.

**Design, Materials, and Procedure** 142 subjects were recruited and compensated as in Experiment 1a. The same design as in Experiment 1a was used. Moreover, we used the same side-effect scenarios except that we placed the single worker inside a train to make his death appear less cruel. Additionally, we made the means scenario less violent by using a different, more technical mechanism. The key differences between the old and new means scenarios lie in the point of intervention and in the equipment the workers are wearing. In the new means scenario it is not the runaway train that is redirected but the train containing the one worker who is going to be killed. Furthermore, the workers in the means condition are wearing a security system that causes all running trains to stop when a worker's heart stops beating. The crucial paragraph that distinguishes the means and side-effect conditions is the following:

“There is only one possibility to avoid the death of the five workers: Every worker is wearing a security system that causes all running trains to stop when a worker's heart stops beating. Peter could throw the switch (by pushing a button), and thereby redirect the yellow train carrying one worker from the parallel upper track onto the main track. The speed-train would collide with this yellow train so that the one worker in this train would instantly lose his life in this accident. However, due to the security system the one worker in the yellow train is wearing, the speed train would stop before it reaches the five workers.”

The security vest was introduced to highlight the role of the single victim as a necessary means. Without the vest one might argue that the five are saved by the train, and the single worker's death is only a side effect. The procedure and test questions were otherwise identical with the ones used in Experiment 1a.

**Results and Discussion** 22 subjects were removed for the same reasons as in the previous experiments. The results for all scenarios are depicted in Figure 5 and based on 120 subjects. Again the intentionality ratings were generally higher in the means than in the side-effect condition,  $F_{1,118} = 4.51$ ;  $p < .05$  (the control condition was again excluded from the analyses). Furthermore, this main effect was moderated by a significant interaction,  $F_{2,236} = 5.23$ ;  $p < 0.01$ . Intentionality ratings only decreased with

increasing numbers of saved lives in the side-effect but not in the means condition. In the means condition, ratings decreased from 4.42 when no lives were saved to 4.31 when one hundred lives were saved ( $p=0.55$ ). In the side-effect condition, ratings dropped from 4.20 when no lives were saved to 3.64 when one hundred lives were saved ( $p < 0.01$ ). Again, in both conditions ratings for the control scenario were significantly higher than the weighted ratings for the remaining three scenarios (means condition:  $p < 0.01$ ; side-effect condition:  $p < 0.0001$ )

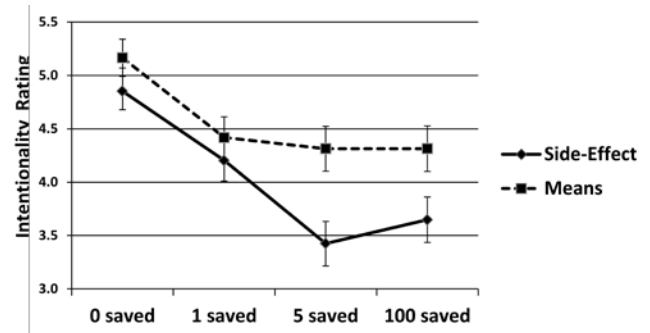


Figure 5: Results of Experiment 2a. Error bars indicate standard error of means.

As in Experiment 1a there was again no difference between the scenarios in which 5 and 100 lives are saved indicating that qualitative differences rather than quantitative differences might be crucial for ratings of intentionality.

## Experiment 2b

Experiment 2b uses the same conditions and manipulations as Experiment 2a along with the moral test question from Experiment 1b.

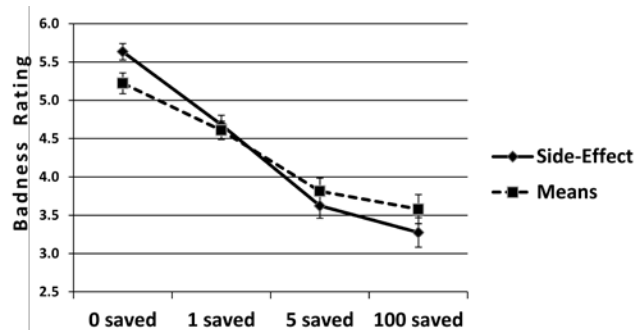


Figure 6: Results of Experiment 2b. Error bars indicate standard error of means.

**Results and Discussion** The results are based on 114 participants from the same online site used in the previous experiments (29 needed to be removed). Figure 6 shows the results, which are clear. We apparently managed to equate the moral aversiveness of harming the one worker. Thus,

our intentionality rating patterns are not moderated by different degrees of aversiveness across conditions. The only significant effect is the overall downward trend that is driven by the size of the primary effect. The more people are saved, the less bad the secondary effect was assessed,  $F_{2,224}=54, p<0.0001$ . These results refute the assumption of the DDE that there is a general moral difference between side-effect and means conditions beyond the typically confounded differences of aversiveness of death (see also Waldmann & Dieterich, 2007; Waldmann et al., 2012).

## General Discussion

We have proposed a new theory of intentionality attributions for moral dilemmas, according to which judgments of intentionality regarding a secondary effect depend on the strength of reasons for accepting the secondary effect in the pursuit of the primary goal. We showed that this assessment is a joint function of the causal role of the secondary effect, and the size of the mitigating reasons provided by the primary goal. The more people are saved by an act that also has a bad secondary effect, the better the secondary act is evaluated on a badness scale, and the less intentionality is attributed to the agent with respect to this effect. Whereas moral evaluations were only sensitive to the size of the mitigating reasons provided by the primary cause and the general aversiveness differences between the scenarios, we found reliable interactions between the size of the primary effect and the causal role of the secondary effect in the intentionality attributions. Means generally were rated lower than ends, but beyond this effect the size of the mitigating reasons did not further affect intentionality assessments. In contrast, in the side-effect conditions intentionality assessments were lowered by the size of the primary effect.

The findings provide important constraints for theories of intentionality attributions. They show, for example, that Knobe's (2003) theory which focuses on the moral value of the secondary effect is at least incomplete. We doubt that the role of the primary effect can simply be added to the present theory, however. To explain our findings the assumption needs to be made that the moral evaluation of the secondary effect is a function of a trade-off with the primary effect. Moreover, in Knobe's (2010) theories causal structure is not viewed as a determinant of intentionality attributions. Causal intuitions are rather, similar to intentionality assessments, conceived as being triggered by moral evaluations of the outcomes. Other theories that have been proposed as competitors to Knobe's (2003) account have also difficulties with our findings since they also only focus on the secondary effect while typically holding the primary effect constant.

Theories based on the DDE are also only partially consistent with our results. In contrast to the predictions of the DDE we showed that the causal structure is not the sole determinant of intentionality attributions but needs to be augmented by our trade-off factor. Moreover, our moral evaluation data contradict the predictions of the DDE. Once

the aversiveness of death differing between the means and side-effect conditions is equated, differences in moral judgments disappeared. This finding is consistent with previous results, and casts doubt on the adequacy of the DDE as a theory of moral evaluations (see Waldmann & Dieterich, 2007).

One general important result is therefore that intentionality and moral judgments, which are closely linked in the DDE, are not influenced by the same factors. Whereas both the primary effect and the causal role of the secondary effect influence intentionality assessments, only the former factor has an impact on moral evaluations, once aversiveness is controlled. This finding casts doubt on the often held assumption that intentionality and moral judgments are closely linked.

## Acknowledgments

This research was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG WA 621/21-1), and the Courant Research Centre "Evolution of Social Behaviour", University of Göttingen (funded by the German Initiative of Excellence).

## References

- Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, 108, 381-417.
- Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from non-moral psychological representations. *Cognitive Science*, 35, 1052-1075.
- Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63, 190-194.
- Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, 33, 315-329.
- Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind and Language*, 23, 165.
- Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York: Cambridge University Press.
- Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, 100, 530-542.
- Timmons, M. (2001). *Moral theory: An introduction*. Rowman & Littlefield Publishers, Inc.
- Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, 18, 247-253.
- Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 364-389). New York: Oxford University Press.

# Children's Causal Learning from Fiction: Assessing the Proximity Between Real and Fictional Worlds

Caren M. Walker<sup>1</sup>, Patricia A. Ganea<sup>2</sup>, & Alison Gopnik<sup>1</sup>

(caren.walker@berkeley.edu, patricia.ganea@utoronto.ca, gopnik@berkeley.edu)

<sup>1</sup>Psychology Department, University of California, Berkeley, 3210 Tolman Hall, Berkeley, CA 94720

<sup>2</sup>Human Development & Applied Psychology, University of Toronto, 56 Spadina Rd., Toronto, Canada M5R 2T1

## Abstract

Fictional information presents a unique challenge to the developing child. Children must learn when it is appropriate to transfer information from the fictional space to the real world and what contextual cues should be considered in this decision. The current research explores children's causal inferences between fictional representations and reality by examining their developing sensitivity to the proximity of the fictional world to the real world, and the effect of this judgment on their subsequent generalization of novel causal properties. By 3-years of age, children are able to evaluate the data that they receive from fictional stories in order to inform their generalization of novel story content to the real world. Additionally, as children develop, they become better able to discriminate between close (realistic) and far (fantastical) fictional worlds when assessing which stories are likely to provide relevant causal knowledge.

**Keywords:** causal inference; fiction; cognitive development; prior knowledge; representation

## The 'Reader's Dilemma'

Children's growing knowledge about the world comes from a variety of sources, including their exposure to fictional material. In fact, much of the unfamiliar information that children encounter appears in the context of stories and fantastical representations of the world. Children, like adults, therefore often encounter the "reader's dilemma": the need to compartmentalize fictional information to insulate real world knowledge from false facts, and the simultaneous need to incorporate this information due to its potential application to a host of real world topics (Gerrig & Prentice, 1999; Potts, St. John, & Kirson, 1989). There is substantial evidence in developmental psychology that indicates that the ability to distinguish reality from fiction develops significantly during the preschool years (e.g., DeLoache, Pierroutakos, Uttal, Rosengren, & Gottlieb, 1998; Flavell, Flavell, & Green, 1989; Woolley & Cox, 2007; Woolley & Wellman, 1990; Woolley & Van Reet, 2006). However, very little research has explored children's ability to learn causal information about the real world from their exposure to fictional material.

Fictional information presents a unique challenge to the developing child. Research has shown that the transfer of knowledge is generally facilitated by similarity between the context in which the information is learned, and the context in which it is to be applied (Catranbone & Holyoak, 1989; Spencer & Weisberg, 1986). However, many of the learning contexts that are created for young children act to *reduce*

this perceived similarity by presenting information in a fictional world that seamlessly interweaves fantasy and reality (Woolley & Cox, 2007). Even in explicitly pedagogical scenarios, teachers often embed their intended curriculum within a fantasy context. This decision is based on the assumption that fictional worlds are more engaging to the young child, and may therefore encourage increased sustained attention and learning of novel material (Harris, 2000; Renninger & Wozniak, 1985).

Previous research supports the proposal that fantasy contexts serve to improve children's performance on certain types of cognitive tasks, such as deductive and syllogistic reasoning and theory of mind (e.g., Dias & Harris, 1988; Dias, Roazzi, & Harris, 2005; Hawkins, Pea, Glick & Scribner, 1984; Richards & Sanderson, 1999; Lillard & Sobel, 1999; Sobel & Lillard, 2001). For example, according to Dias et al. (2005), placing an unfamiliar premise in a fantastical context – particularly when the premise directly contradicts a currently-held theory – allows children to override their natural empirical orientation, or bias to reason in line with their past experiences. It is unknown, however, how learning and generalization of novel causal information (which does not require the suspension of existing knowledge) is affected by the fantastical contexts of the fictional stories in which this information is embedded.

The ability to effectively process fictional information is dependent upon a variety of representational skills, including at least two major factors that are unique to learning from fictional material. The first includes the development of a mature concept about the boundary between the fictional and real world, as well as an understanding of what information is more appropriately quarantined to the fictional space. Second, it is necessary for children to develop an understanding of when it is appropriate to transfer information from the fictional to the real world, and what contextual cues should be considered in this decision. The current research explores the early development of each of these factors, and in particular, examines whether children's sensitivity to contextual cues in fictional worlds changes over the course of development.

## Children's Beliefs about Fictional Worlds

There is a growing literature in developmental psychology regarding when and how children distinguish between fantasy and reality. Methods for testing this distinction vary

greatly, and include assessing children's beliefs about their imaginary companions (e.g., Taylor, 1999), their beliefs in magic (e.g., Rosengren Kalish, Hickling, & Gelman, 1994), and directly asking children about their beliefs in familiar and novel fantastical entities (e.g., Clark, 1995; Woolley & Van Reet, 2006). There has also been some work specifically aimed at assessing children's beliefs about the reality status of fictional content in storybooks (e.g., Morison & Gardner, 1978; Wellman & Estes, 1986; Woolley & Cox, 2006). Taken together, this research has shown that children do distinguish fantasy characters from real ones, and that (depending on the particular method and the nature of the task) this ability matures between 3- and 6-years of age.

Related research also indicates that children differentiate the particular contexts in which they encounter information in storybooks from a relatively young age. For example, Woolley & Cox (2006; 2007) presented preschoolers with realistic, fantastical, or religious stories in a variety of contexts and found that while 3-year-olds were more likely to judge characters as real than were 4- and 5-year-olds, most children accurately judged all characters as not real for all story types. They also found that children made more claims that the realistic story events "could happen in real life" than they did for fantastical story events, which indicates that context matters in the formation of these judgments. While this work explores children's willingness to believe that the story events themselves could happen in real life, the authors do not consider whether children learn and apply the information presented in the storybook to real world scenarios.

According to a study conducted by Skolnick & Bloom (2006), children conceptualize multiple fictional worlds as separate from one another, and separate from the real world. Given children's tendency to quarantine fictional worlds from the real world, it is possible that children also consider the content of these worlds to be distinct, regardless of the assessed possibility of the events themselves.

## Learning from Stories

Despite the importance of understanding the distinction between fiction and reality, storybooks do often provide important opportunities for children to learn information about the real world that cannot be experienced directly (Ganea, Pickard, & DeLoache, 2007). There is currently a small, but growing number of researchers examining the development of children's ability to learn from picture books and the factors that affect their successful generalization of this newly learned information to the real world (e.g., Ganea, Pickard, & DeLoache, 2008; Ganea, Ma, & DeLoache, 2011; Simcock & Dooley, 2007; Walker, Walker, & Ganea, under review). To date, most of this work has focused primarily on transferring labels and simple concepts from a realistic or factual representation, rather than embedded in the context of a fictional story.

For example, Ganea, et al. (2008) demonstrated that 15- and 18-month-old infants are able to extend newly learned

labels both from picture books to real objects and from real objects to picture books. Additionally, they showed that performance was affected by the iconicity of the pictorial images, indicating that the nature of the represented content matters for transferring labels learned in the context of a picture book interaction. In related work, Ganea, et al. (2011) showed that 3- and 4-year-old children can also learn simple biological information about color camouflage in animals from a single picture book interaction, and apply this newly acquired knowledge to real world situations. These experiments indicate that from a very young age, children are able to incorporate factual information about the real world from minimal exposure to picture books, in certain highly realistic and pedagogical scenarios.

In an attempt to explore children's ability to learn from stories that include fantastical content, Richert, et al. (2009) conducted a series of studies looking at analogical reasoning from picture books to other stories and from picture books to the real world. In three experiments, 3½- to 5-year-old children were presented with analogical problems in the context of a short story which involved either real or fantasy characters. In the first experiment, children were tested on their ability to transfer a solution from a story about familiar fantasy characters to a story about realistic characters, and vice versa. In general, children were more likely to transfer the solution to the novel problem from the real source than from the fantasy source. In the second experiment, children were asked to generalize these same solutions to real world contexts (games that involved the manipulation of physical objects). Again, children were more likely to transfer the solution from the real source than from the fantasy source. Later, Richert and Smith (2011) replicated these findings using more complex stimuli that were introduced in a pedagogical context.

The results of these experiments indicate that the context of a story does affect children's ability to draw analogies between the story content and novel scenarios in the real world. One explanation for these results may be that children are sensitive to the *proximity* of the story world to reality, or the similarity of the causal structure of the fictional world to the real world.

## Assessing Proximity of the Fictional World

Fictional worlds that are closer in possibility space (i.e., have higher proximity) share more of the causal structure with the real world, while those that are further away (i.e., have lower proximity) share less. In line with this idea, research with adult participants has demonstrated that the perceived proximity of the fictional world to reality influences participants' decisions to import facts about the real world in making inferences about fictional environments (Skolnick & Goodstein, 2009).

To test this, Skolnick and Goodstein (2009) presented adult participants with three stories that varied in their similarity to reality. They found that participants were more likely to import true facts from the real world to the fictional worlds that were more similar in underlying causal



structure. Adult participants were also more likely to import facts that were considered to be more causally central to the representation of reality (e.g., mathematical facts) to all worlds (regardless of their proximity) than facts that are less central (e.g., conventional or contingent facts). Thus, adults infer that fictional worlds that are more similar to the real world, or closer in possibility space, should contain more facts that are true of the real world. It is currently unknown whether children also display this sensitivity to the distance that a story world lies from reality, and to what extent (if any) this sensitivity to world proximity would affect children’s learning from fictional representations. Examining these issues will inform us about the nature of the mechanisms that underlie learning from fictional material and contribute to a more complete understanding of how causal knowledge is acquired more broadly.

### Current Research

In the current research, we explore children’s causal inferences about fictional content and examine whether contextual information influences their subsequent generalization of novel causal properties to the real world. In particular, we examine whether the likelihood that children will generalize novel causal information varies based upon the perceived proximity of the fictional world (with far worlds generating a lower probability of generalization than close worlds), and whether sensitivity to the proximity of the fictional world changes over the course of development.

### Participants

One hundred and eight preschoolers participated in the study, including 36 3-year-olds ( $M = 43.7$ ,  $SD = 3.9$ , range = 37.2 – 48.0), 36 4-year-olds ( $M = 54.9$  months,  $SD = 3.2$ , range = 49.8 – 59.9), and 36 5-year-olds ( $M = 66.8$  months,  $SD = 2.8$ , range = 61.6 – 71.8). An approximately equal number of males and females were included at each age. Eight additional 3-year-olds and two 4-year-olds were tested, but excluded for failure on both memory questions or failure to complete the training for the sorting task. Although most children were from White, middle-class backgrounds, a range of ethnicities resembling the diversity of the local population was represented. All children were recruited from local preschools and museums.

### Materials

Two 13-page illustrated storybooks were constructed for the experiment. Both stories depicted human protagonists who go on a family camping trip. One version of the story (the close world) was realistic, including no explicit violations of reality (i.e., all events could have easily taken place in the real world), and the other version of the story (the far world) was fantastical, including major violations of reality. Both stories shared the same general structure and the same number and type of events, but varied in the degree of proximity to the real world (see Table 1 for a list of all major story events and Figure 1 for sample pages).

Table 1: Close World and Far World Story Events.

Close World Events	Far World Events
Drive in car	Fly with magic cape
Find a ladybug	Find a fairy
Climb a tree	Talk with a tree
Raining raindrops	Raining stickers
<b>Smell ‘Popple Flower’</b>	<b>Smell ‘Popple Flower’</b>
<b>Get Hiccups</b>	<b>Get Hiccups</b>
Swim in pond	Swim in chocolate pond

In both stories, a novel (plausible) causal relationship was embedded within context of the other events – smelling a ‘Popple Flower’ causes the protagonist to get the hiccups (see Figure 1). This causal relationship was identical across both versions of the story.



Figure 1: Sample pages from *close world* (top left) and *far world* (top right) versions and the target causal relationship as it appears in both storybooks (bottom).

For the sorting task, eight training cards and two sets of six story event cards were constructed. The eight training cards depicted illustrations of real and fantastical versions of events that did not appear in the story (e.g., a boy eating spaghetti vs. a boy eating lightening). The two sets of story event cards depicted each of the individual story events (see Table 1). One set was constructed for children in the *close world* condition and the other set was constructed for children in the *far world* condition. One of the six cards in each set was an identical depiction of the target causal relationship (i.e., a boy smelling a ‘Popple Flower’ and getting the hiccups).

For the generalization task, we used a 5 x 7 color photograph of a real flower that was similar in shape and color to the illustrated ‘Popple Flower’ in the stories.

### Procedure

In a between-subjects design, half of the children in each age group were randomly assigned to the *close world* or *far world* story condition. Children were tested individually, sitting next to the experimenter. The experimenter read one of the two books to the child, interacting naturally, pointing to illustrations, and asking questions in a manner that is typical of parent-child book interactions. The experimenter

introduced the story by saying, “This is a made-up story about a boy who goes on a camping trip.” While children were encouraged to engage with the content of the story, the experimenter provided no additional information over the course of the interaction.

**Memory Assessment.** Immediately after hearing the story, children were asked two memory questions intended to assess attention and recall. One question assessed recall of the novel causal relationship (“*What happened to the boy in the story when he smelled the Popple Flower?*”). The second question was open-ended, and intended to assess recall for other story events (“*What kinds of things did the boy do on his camping trip in the story?*”). Children were prompted to continue responding until they successfully recalled at least three story events. If children responded with fewer than three events, the experimenter would ask, “did anything else happen?” until the child could no longer recall any more story events. Children who failed both memory questions were excluded from the study.

**Sorting task.** Children were trained to sort picture cards into “real” and “pretend” piles. The eight training cards were presented, one at a time, and children were instructed to sort the cards into two piles: one pile for things that “can really happen” and one pile for things that “cannot really happen, and are just pretend.” This training was discontinued after children successfully sorted four cards in a row. Children who failed this training were excluded from analysis. Immediately following the training, children were asked to continue sorting, using the six test cards that depicted each of the events that took place in the story (including a card depicting the target causal relationship). As in the training, children were asked to sort each of the depicted events into the “real” pile or the “pretend” pile.

**Generalization task.** To assess generalization, children were presented with the target causal property that appeared in the story (*smelling ‘Popple Flowers’ causes hiccups*) in a real world context, and asked to judge whether this causal relationship would hold in the real world. To do so, the experimenter showed the child a realistic photograph depicting flowers that were similar in shape and color to the illustrated flowers in the story, saying, “On my way here today, I saw these. I didn’t know what kind of flowers they were, but I smelled them. What do you think happened to me, here in the real world? Do you think that I got the hiccups or that I did not get the hiccups?” The order of presentation of the potential outcomes was counterbalanced. The generalization question was presented in a forced choice format in order to eliminate a “yes” bias. Children received a score of “0” if they responded that the experimenter did not get the hiccups (no generalization) and a score of “1” if they responded that the experimenter did get the hiccups (generalization). The order of the sorting and generalization tasks was counterbalanced.

## Results

Nearly all children who were included in the final analysis answered both memory questions correctly (97% of 3-year-olds, 97% of 4-year-olds, and 100% of 5-year-olds). There was no difference found between conditions on the memory assessment,  $F(1, 106) = 1.86, p = .18$ , indicating that children in both conditions were equally able to recall the content of the story.

Analysis of sorting judgments indicates that children were also sensitive to the presence of fantastical and realistic content in the story that they heard. There were a total of five contextual story events (not including the target causal relationship). Children in the *close world* condition correctly sorted the majority of the realistic story events to the “real pile” ( $M = 4.43, SD = 0.93$ ), while children in the *far world* condition correctly sorted the majority of the fantastic story events to the “pretend pile” ( $M = 4.33, SD = 1.06$ ). While the purpose of the sorting task was to assess whether children were capable of identifying the story events as real or pretend, children were also asked to sort a story event card depicting the target causal relationship (i.e., a boy smelling a ‘Popple Flower’ and getting the hiccups), which served as an additional measure of generalization. Although this story event card was identical in both conditions, children in the *close world* condition were more likely to sort this story event in the “real pile” ( $M = .67, SD = .40$ ), while children in the *far world* condition more likely to sort this story event in the “pretend pile” ( $M = .72, SD = .45$ ), with a significant difference between conditions  $\chi^2(108, 1) = 9.69, p < .01$ .

Loglinear analysis and chi squares were conducted to assess differences in children’s responses on the generalization task at each age and for each condition. Results appear in Figure 2 below. Results of loglinear analysis demonstrate an effect of condition on generalization,  $\chi^2(108, 1) = 27.39, p < .001$ , indicating that children successfully differentiated between *close worlds* and *far worlds* when selectively generalizing novel causal information from the story to the real world.

Overall, children in the *close world* condition chose to generalize the target causal information to the real world scenario,  $\chi^2(54, 1) = 10.67, p < .01$ , with no difference between age groups,  $\chi^2(54, 2) = 0.45, p < .80$ , indicating that preschoolers are able to generalize novel causal information from realistic stories. Children in the *far world* condition made the opposite inference, with the majority of children choosing not to generalize the target causal information to the real world scenario,  $\chi^2(54, 1) = 14.42, p < .001$ . While 3-, 4-, and 5-year-olds all generalized more often from the *close world* than from the *far world* ( $\chi^2[36, 1] = 5.04, p < .05$ ;  $\chi^2[36, 1] = 7.80, p < .01$ ; and  $\chi^2[36, 1] = 14.57, p < .001$ , respectively), our results also provide evidence for a developmental change: children’s willingness to generalize novel causal information from the *far world* decreased (marginally) with age,  $\chi^2(54, 2) = 5.67, p = .059$ , with 3-year-olds more likely to generalize the target causal relationship (39%) than 4-year-olds (28%) and 5-year-olds

(6%). There was a significant difference between 3- and 5-year-olds' willingness to generalize from the *far world*,  $\chi^2(36, 1) = 5.79, p < .02$ .

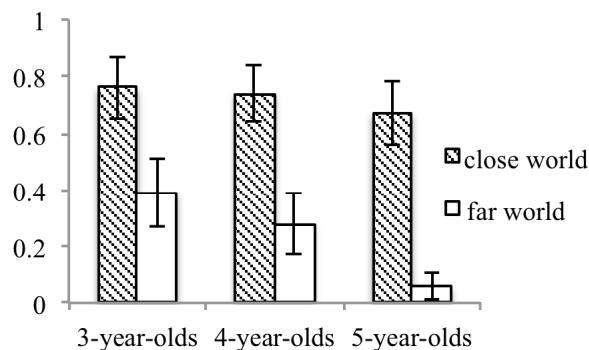


Figure 2. Percentage of 3-, 4-, and 5-year-old children who generalized the target causal relationship from the story to the real world in each condition.

## Discussion

The current study examined children's generalization of novel causal information from fictional representations to the real world. Findings provide evidence that preschool-aged children are sensitive to the proximity of the fictional world when selectively learning and applying novel causal information from stories. While children in both *close world* and *far world* conditions were able to remember the target causal relationship embedded in the story, the proximity of the fictional world to reality influenced their subsequent generalization of this novel information. These results demonstrate that children begin to differentiate between close and far fictional worlds from a very early age, and that this sensitivity undergoes a process of developmental change, increasing between 3- and 5-years.

How might this sensitivity to the proximity of fictional worlds develop over time? The development of this ability requires the learner to successfully integrate the information provided in the story with their prior knowledge and beliefs about the causal structure of the real world. However, little was previously known about children's use of prior knowledge when evaluating the applicability of information learned in fictional representations, and how their reliance on this prior knowledge may change over the course of early development.

Recent probabilistic accounts of learning (e.g., Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Schulz, Bonowitz, & Griffiths, 2007) may provide a natural framework for addressing these questions. According to these accounts, a learner's background knowledge and prior beliefs are productively integrated with new data when forming novel inferences about the causal structure of the world. As prior knowledge increases over the course of development, children will become better able to approximate the true causal structure of the real world. As a result, children's ability to use contextual cues from the

story to inform their decision to generalize novel information should improve over time. For example, if story events are determined to have a high probability of occurring in the real world, children should be more likely to generalize novel causal information learned in this context than in cases in which the story events are determined to have a low probability of occurring in the real world. Therefore, as their prior knowledge about the causal structure of real world increases, children become better able to evaluate the information that they receive from fictional contexts to inform and structure their own learning.

Similarly, children's developing sensitivity to the proximity between fictional worlds and reality may be mediated by their increasing prior knowledge about the nature of fantastical representations. Previous research has shown that children who score higher on fantasy orientation scales (i.e., children who have more experience engaging with fantasy worlds) are less likely to transfer solutions to analogical problems from fantastical stories to real world scenarios (Richert & Smith, 2011). In other words, those children with the greatest amount of prior knowledge about fantastical representations are the least likely to draw analogies between worlds. One explanation for these findings is that children with more experience with fantastical representations have developed an increased appreciation of the distinction between the causal structure of *far worlds* and reality, which may lead to the sophisticated strategy of quarantining causal information acquired from these fantastical contexts. The developmental change that we document in the current study provides evidence for each of these related proposals. Future research should further explore the particular type of prior knowledge – knowledge about the true causal structure of the real world, knowledge about the nature of fictional representations, or some combination of the two – that is most relevant to developing this sensitivity during early childhood.

In sum, our findings demonstrate that by 3-years of age, children are already able to evaluate the data that they receive from fictional stories in order to inform their generalization of novel story content to the real world. Additionally, as children develop, they become better able to discriminate between close and far fictional worlds when assessing which stories are likely to provide relevant causal knowledge about the real world. These findings have important implications for educational contexts that rely upon children's literature to present intended curriculum. Storybooks provide rich opportunities for children to learn about aspects of the world that are otherwise inaccessible to them. However, because children's selective learning from storybooks is at least partly contingent upon the perceived proximity between worlds, the presence of fantastical events may inadvertently undermine educational goals. By explicitly directing children to the generalizable information in fictional stories, adults may help young learners to negotiate the complex relationship between fictional worlds and reality.

## Acknowledgements

This research was funded in part by the McDonnell Foundation to A. Gopnik. The authors thank Ellen Winner for her contributions to the development of this research and Ngoc Nyugen for illustrating the storybooks. We also thank Rosie Aboody, Anna Akullien, Sierra Eisen, and Brynna Ledford for their assistance with data collection.

## References

- Catranbone, R. & Holyoke, K.J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15: 1147-1156.
- Clark, C.D. (1995). *Flights of fancy, leaps of faith: Children's myths in contemporary America*, University of Chicago, Chicago.
- DeLoache, J. S., Pierroutsakos, S. L., Uttal, D. H., Rosengre, K. S., & Gottlieb, A. (1998). "Grasping the nature of pictures," *Psychological Science*, 9 (3): 201-210.
- Dias, M.G. & Harris, P.L. (1988). The effect of make-believe play on deductive reasoning. *British Journal of Developmental Psychology*, 6: 207-221.
- Dias, M.G., Roazzi, A., & Harris, P.L. (2005). Reasoning from unfamiliar premises: A study with unschooled adults. *Psychological Science*, 16(7): 550-554.
- Flavell, J. H., Flavell, E. R., & Green, F. L. (1989). Young children's knowledge about the apparent-real and pretend-real distinctions. *Developmental Psychology*, 23, 816-822.
- Ganea, P. A, Pickard, M., & DeLoache, J.S. (2008). "Transfer between picture books and the real world by very young children," *Journal of Cognition and Development*, 9, 46-66.
- Ganea, P. A., Ma, L., DeLoache, J. (2011). Young children's learning and transfer of biological information from picture books to real animals. *Child Development*, 82(5): 1421-1433.
- Gerrig, R.J. & Prentice, D.A. (1991). "The representation of fictional information." *Psychological Science*, 2(5): 336-340.
- Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes Nets, *Psychological Review*, 111(1):3-32.
- Harris, P.L. (2000). *The work of the imagination*. Blackwell: Oxford, UK.
- Hawkins, J., Pea, R.D., Glick, J., and Scribner, S. (1984). Merds that laugh don't like mushrooms: Evidence for deductive reasoning by preschoolers. *Developmental Psychology*, 20:584-594.
- Lillard, A. & Sobel, D. (1999). Lion kings or puppies: The influence of fantasy on children's understanding of pretense. *Developmental Science*, 2: 75-80.
- Morison, P. & Gardner, H. (1978). Dragons and dinosaurs: The child's capacity to differentiate fantasy from reality. *Child Development*, 49 (3): 542-648.
- Nichols, S., & Stich, S. (2000). A cognitive theory of pretense. *Cognition*, 74: 115-147.
- Potts, G.R., St. John, M.F., & Kirkson, D. (1989). "Incorporating new information into existing world knowledge." *Cognitive Psychology*, 21:303-333.
- Richert, R. A., Shawber, A. B., Hoffman, R. I., & Taylor, M. (2009). "Learning from real and fantasy characters in preschool and kindergarten." *Cognition & Development*.
- Richert, R.A. & Smith, E.I. (2011). Preschoolers' quantifying of fictional stories. *Child Development*, 82(4): 1106-1119.
- Renninger, K.A. & Wozniak, R.G. (1985). Effect of interest on attentional shift, recognition, and recall in young children. *Developmental Psychology*, 21: 624-632.
- Richards, C.A. & Sanderson, J.A. (1999). The role of imagination in facilitating deductive reasoning in 2-, 3- and 4-year-olds. *Cognition*, 101: B9-B18.
- Schulz, L., Bonawitz, E., Griffiths, T. (2007). Can being scared make your tummy ache? Naïve theories, ambiguous evidence, and preschoolers' causal inferences, *Developmental Psychology*, 43(5): 1124-1139.
- Simcock, G. & Dooley, M. (2007). Generalization of learning from picture books to novel test conditions by 18- and 24-month-old children. *Developmental Psychology*, 43, 1568-1578.
- Skolnick, D. and Bloom, P. (2006). What does Batman think about SpongeBob? Children's understanding of the fantasy/fantasy distinction. *Cognition*, 101: B9-B18.
- Skolnick, D. & Goodstein, J. (2009). What belongs in a fictional world? *Journal of Cognition and Culture*, 9, 69-78.
- Sobel, D.M. & Lillard, A.S. (2001). The impact of fantasy and action on young children's understanding of pretense. *British Journal of Developmental Psychology*, 19,85-98.
- Spencer, R.M. & Weisberg, R.W. (1986). Context-dependent effects on analogical transfer during problem solving. *Memory & Cognition*, 14: 442-449.
- Walker, C.M., Walker, L., & Ganea, P. (2012). The role of symbol-based experience in learning and transfer from pictures: Evidence from Tanzania. Manuscript under review.
- Woolley, J.D. & Van Reet, J. (2006). Effects of context on judgments concerning the reality status of novel entities. *Child Development*, 77(6): 1778-1793.
- Woolley, J.D. & Cox, V. (2007). "Development of beliefs about storybook reality." *Developmental Science*, 10 (5): 681-693.

# Explaining Influences Children's Reliance on Evidence and Prior Knowledge in Causal Induction

Caren M. Walker, Joseph Jay Williams, Tania Lombrozo, & Alison Gopnik

(caren.walker@berkeley.edu, joseph\_williams@berkeley.edu, lombrozo@berkeley.edu, gopnik@berkeley.edu)  
Department of Psychology, University of California, Berkeley, 3210 Tolman Hall, Berkeley, CA 94720

## Abstract

In two studies, we examine how prompting 5- and 6-year-olds to explain observed outcomes influences causal learning. In Study 1, children were presented with data consistent with two causal regularities. Explainers outperformed controls in generalizing the regularity that accounted for more observations. In Study 2, this regularity was pitted against an alternative that accounted for fewer observations but was consistent with prior knowledge. Explainers were *less* likely than controls to generalize the regularity that accounted for more observations. These findings suggest that explaining drives children to favor causal regularities that they expect to generalize, where current observations and prior knowledge both provide cues.

**Keywords:** Explanation; cognitive development; causal reasoning; prior knowledge; generalization; abduction.

Since Piaget, researchers have regarded children's explanations as a window into cognitive development, revealing children's understanding of the world (e.g., Keil, 2006). More recently, however, the very process of seeking, generating, and evaluating explanations has additionally been proposed as a powerful mechanism for learning and generalization, scaffolding knowledge acquisition and contributing to theory change (e.g., Chi, DeLeeuw, Chiu, & LaVancher, 1994; Lombrozo, 2006; Wellman, 2011). Here we investigate the role of explanation in young children's causal learning, focusing on whether and how explaining influences the relative contributions of observed evidence and currently-held theories ("prior knowledge").

Discovering the underlying causal structure in the world is one of the major inductive problems that young learners face during development. By 5 years of age, children have already developed abstract, coherent representations of causal relationships in a variety of domains (e.g., Carey, 1985; Gelman & Wellman, 1991). The acquisition of this causal knowledge is supported by powerful learning mechanisms that allow children (and adults) to infer novel causal relationships from patterns of evidence and prior beliefs (e.g., Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Griffiths, Sobel, Tenenbaum, & Gopnik, 2011). For example, 5-year-old children can effectively track covariational evidence to identify a novel cause, but require stronger evidence to endorse a cause that conflicts with their prior beliefs than one consistent with those beliefs (Schulz, Bonawitz, & Griffiths, 2007). These findings reveal that children integrate current observations with prior beliefs in the process of learning and revising their causal knowledge. Here we propose that explaining may play a

role in this integration, by influencing the relative contributions of evidence and prior knowledge – that is, the extent to which learners revise their beliefs in light of their observations and prior commitments.

A first possibility, which we call the *evidence hypothesis*, is that explaining a set of observations directs learners to update their beliefs in light of those observations. As a result, explaining will lead learners to change their beliefs *more* than they would have in the absence of explaining. This might be expected, for example, if explanation boosts children's engagement with a task (Siegler, 2002), thereby making them more responsive to feedback and less likely to disregard relevant observations. Consistent with this possibility, a variety of studies have found that presenting young children with feedback is less effective as a means to changing their beliefs than having them explain as well (e.g., Wellman, 2011). More indirect support comes from the fact that explaining can direct children towards anomalous data, which conflicts with their prior beliefs and may therefore signal a need for belief revision (e.g., Legare, 2010; Legare, Gelman, & Wellman, 2010). In sum, the *evidence hypothesis* predicts that explaining should make children more responsive to observed data, leading to greater revision in current beliefs.

A second possibility, which we call the *prior knowledge hypothesis*, is that engaging in explanation invokes learners' currently-held theories, leading them to revise their beliefs *less* than they might have in the absence of explanation. In line with this hypothesis, several researchers have proposed that explaining encourages learners to accommodate what they are trying to explain in the context of what they already know (e.g., Chi et al., 1994; Lombrozo, 2006). Explaining could thus lead children to favor hypotheses consistent with prior knowledge, even when those hypotheses might not have been entertained or preferred on the basis of evidence alone. Kuhn and Katz (2009) further suggest that by encouraging learners to consider *why* something is the case, explaining can lead them to discount observed evidence altogether. From this perspective, learners who explain could be less inclined towards hypotheses that simply match current evidence, either because they weight prior beliefs more heavily or because they underweight current evidence.

A final possibility is that explaining does not affect the role of evidence or prior knowledge uniformly across contexts. Instead, explaining could change the criteria employed in evaluating evidence and prior knowledge in the course of learning, resulting in greater belief revision under some conditions and less belief revision under others. The version of this possibility that we explore here is the

*generalization hypothesis*, according to which generating explanations leads children to evaluate both evidence and prior knowledge as cues to which patterns in their observations are most likely to *generalize* to new cases. This proposal is motivated by the theory-theory of cognitive development (e.g., Carey, 1985; Gopnik & Wellman, 1994) as well as recent research concerning the role of explanation in adult learning (Williams & Lombrozo, 2010a, 2010b). According to the theory-theory, children construct intuitive theories that support explanation, prediction, and control, where theories are regarded as consisting in “causal-explanatory” knowledge. In order to effectively support prediction, successful theories must account for past observations as well as generalize to future observations. The process of generating explanations could therefore influence learning by directing children to construct the causal-explanatory beliefs that they judge most likely to generalize, and to consult both evidence and prior knowledge in making this assessment.

Insight into how explanation could contribute to this process comes from subsumption and unification theories of explanation from philosophy, according to which successful explanations demonstrate how what is being explained can be subsumed under a broad regularity, such as a natural law (e.g., Friedman, 1974; Kitcher, 1989). If explaining drives learners to understand current observations in terms of broadly-applicable regularities, it should facilitate the discovery and generalization of regularities that subsume the greatest number of cases. This idea has recently been developed as an empirical hypothesis concerning the role of explanation in learning, called the *subsumptive constraints account*, and is supported by studies with adults learning novel categories (Williams & Lombrozo, 2010a, 2010b).

To illustrate, consider Williams and Lombrozo (2010a), in which adults learned novel categories from observations and were prompted to either explain each observation’s category membership or to engage in a control task. The categories could be differentiated by a subtle regularity that accounted for all observations, or by a more salient regularity that accounted for only a subset of observations. Compared to participants in other conditions, those prompted to explain were more likely to discover the regularity with greater subsumptive scope (the one that accounted for more cases), and more recent evidence suggests that prior knowledge additionally serves to inform these assessments (Williams & Lombrozo, 2010b, under review). The *generalization hypothesis* proposes that explanation influences children’s causal learning in a similar way: by directing young learners to the causal hypotheses with the greatest subsumptive scope, which is assessed on the basis of both evidence and prior knowledge.

These candidate hypotheses are not intended as comprehensive accounts, but rather provide a useful framework for considering the possible mechanisms by which explanation could influence causal induction. In the two studies that follow, we manipulate both the observed evidence and children’s prior knowledge in order to test

these hypotheses and assess the role of explanation in children’s causal learning. Specifically, children learned about a novel causal system in which some objects were causally efficacious and others were not. Half of the participants were prompted to explain the causal outcomes during the training phase of the experiment, and the other half were prompted to describe these outcomes. In Study 1, the observed data suggested two candidate causal regularities that were equally consistent with prior knowledge. However, one of the two causal regularities accounted for more of the observed data. In Study 2, children were again presented with two candidate causal regularities that accounted for all or a subset of the data, but the regularity that accounted for fewer observations was more consistent with prior knowledge. In both studies, children were then asked to generate causal predictions in order to assess which causal regularity was discovered, preferred, and generalized to novel cases.

The three hypotheses outlined above generate different predictions across these two studies. The *evidence hypothesis* predicts that children will generalize according to the regularity that covers more of their observations in both studies. The *prior knowledge hypothesis* predicts that because explaining prompts children to rely upon their current theories, those who explain should respond no differently in Study 1, and be more likely to form generalizations in line with their prior knowledge in Study 2. The *generalization hypothesis* predicts that explainers should consider both the data and their prior knowledge to identify the regularity that is likely to apply most broadly, which could lead to opposite generalizations for Studies 1 and 2. Taken together, the results will thus help characterize the role of explanation in children’s causal learning.

## Study 1: Regularities in Observed Evidence

In Study 1, we present two groups of children with the task of learning a novel causal relationship from a series of observations, where one group is prompted to explain each observation (the *explain* condition) and the other to describe each observation (the *control* condition). The observed data suggest two causal regularities: a regularity that accounts for 100% of observations (the 100% regularity) and a regularity that accounts for 75% of observations (the 75% regularity). The evidence for candidate regularities is presented using a *blicket detector* paradigm (e.g., Gopnik & Sobel, 2000), in which children learn which objects have causal efficacy in activating a machine.

### Participants

Forty-two 5-year-old children ( $M = 64.2$  months;  $SD = 3.6$ , range: 59.9 – 72.7; 25 girls) were included, with 21 children randomly assigned to each condition. Children were recruited from local preschools and museums.

### Materials

The machine used in the training phase of both studies was a “blicket detector” – a box concealing a wireless doorbell.



When an object “activated” the machine, the doorbell was pressed remotely, producing a melody.

An illustration of the complete set of training blocks appears in Figure 1A. Eight 2” wooden blocks (four “Go” blocks and four “No-Go” blocks) were used during the training phase. A green plastic lego plate was affixed to the top and front of each block. Attached to each lego plate was a single, small rectangular lego of uniform size. The two causal regularities (100% and 75%) were represented by different-colored legos. For the four blocks that activated the machine (*Go* blocks), the 100% regularity was represented by a green lego and the 75% regularity was represented by a red lego. For the four blocks that did not activate the machine (*No-Go* blocks), the 100% regularity was represented by a yellow lego and the 75% regularity was represented by a white lego. For half of the children, the 100% regularity (the green/yellow lego) appeared on the top of the block and the 75% regularity (the red/white lego) appeared on the front of the block, and for half of the children, these positions were reversed.

For the testing phase, four additional blocks were used. These testing blocks were identical to the training blocks, but included only one of the four lego colors on each. Additionally, a cardboard “hiding box” was constructed, with four cut-out windows that were covered with black felt flaps. These windows were designed so the experimenter could place two blocks inside the hiding box and lift two flaps to show the participant only one of the two legos (on the top or front) of each block.

## Procedure

The experimenter introduced the machine, explaining that some things make the machine play music and some things do not. The child was then instructed to sort the blocks into two piles according to whether they activated the machine.

The experimenter placed each of the eight blocks on the machine, one at a time. After children observed each outcome, they were asked for a verbal response. In the *explain* condition, children were asked to explain the outcome: “Why did/didn’t this block make my machine play music?” In the *control* condition, children were asked to describe the outcome: “What happened to my machine when I put this block on it? Did it make music?” The order of presentation of the eight blocks was semi-random (see Figure 1A). All blocks remained visible and were grouped on the table throughout the training and test phases of the experiment to eliminate memory demands.

For the test phase, the machine was removed and the experimenter introduced the “hiding box,” saying: “This is my hiding box. I am going to put two new blocks into my hiding box, and lift these flaps so you can only see part of each block.” The child was told: “One of the blocks I put in my hiding box will make my machine play music, and one of the blocks will not. I want you to tell me which one you think *will/will not* make my machine play music.”

Each test question was designed to pit one potential causal regularity against the other. In the first set, the 100%

and 75% test items, the *Go* features were pit against the *No-Go* features for both the 100% regularity (green vs. yellow) and the 75% regularity (red vs. white). In the next set, the experimenter presented *conflict items*, in which the *Go* feature from the 100% regularity (green) was pit against the *Go* feature from the 75% regularity (red). Each time, children were asked to predict which block would make the machine play music. These questions were intended to present a conflict between the two potential causes to examine whether children would privilege one regularity over the other. Each of these test questions was repeated for a second time in which the experimenter asked the child to indicate which block *would not* make the machine play music. There was a total of six test questions in this format: four 100% and 75% test items (two green vs. yellow; two red vs. white) and two *conflict items* (two green vs. red).

Responses on the 100% and 75% test items were scored for accuracy, where accuracy reflected the correct selection of the *Go* block when asked for an item that would make the machine go, and the *No-Go* block when asked for an item that would not. Children received 1 point for selecting the correct response and 0 points otherwise. For the *conflict items*, we examined whether responses conformed to the 100% regularity (selecting the green lego over the red lego) or the 75% regularity (selecting the reverse), coding the former as “1” and the latter as “0.”<sup>1</sup>

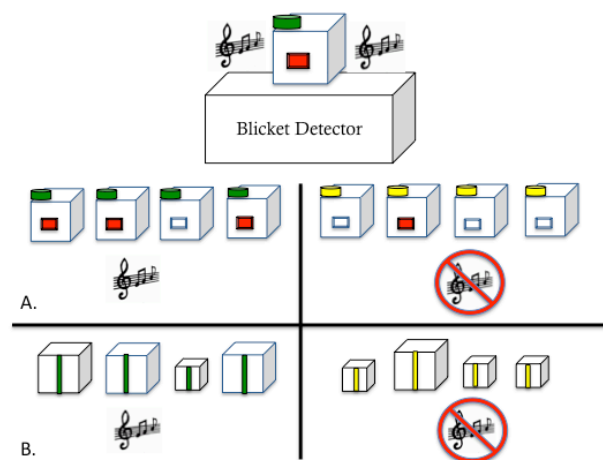


Figure 1: Procedure for Studies 1 (A) and 2 (B).

## Results

Whereas children in both conditions were able to learn the 100% and 75% regularities, those who were prompted to explain during the training phase were more likely than controls to privilege the 100% regularity over the 75% regularity when generalizing to novel cases (see Figure 2).

We begin by reporting the results of children’s

<sup>1</sup> Both studies additionally included certainty ratings and an explanation selection task with findings that mirror those that we report; we do not report these data in the interest of space.



performance on the 100% and 75% test items. A repeated measures ANOVA was conducted with each question type (100% and 75% test items) as the repeated measure and condition (*explain* vs. *control*) as a between-subjects variable. Analysis revealed no difference in children's performance on the two question types,  $F(1, 40) = 0.195$ ,  $p = .661$ , no difference between conditions,  $F(1, 40) = 1.84$ ,  $p = .182$ , and no interaction between question type and condition,  $F(1, 40) = 0.780$ ,  $p = .382$ . Children in both the *explain* condition ( $M = 0.87$ ,  $SD = 0.23$ ) and the *control* condition ( $M = 0.77$ ,  $SD = 0.22$ ) were able to learn the 100% and 75% regularities during the training phase and use this information when generalizing to novel blocks.

To analyze children's performance on the *conflict items*, a one-way ANOVA was conducted with condition (*explain* vs. *control*) as the between-subjects variable. There was a significant difference between conditions,  $F(1, 40) = 5.79$ ,  $p < .02$ : Children in the *explain* condition were more likely ( $M = .667$ ,  $SD = .33$ ) than those in the *control* condition ( $M = .405$ ,  $SD = .38$ ) to base their generalizations on the 100% regularity. Results support the hypothesis that prompting learners to explain helps them to discover and extend causal regularities that account for more of their observations, consistent with the *evidence* and *generalization* hypotheses.

## Study 2: Conflicting Prior Knowledge

Study 1 was designed to assess the role of current evidence in children's causal learning and generalizations. However, because the two candidate causes differed only in color (green vs. red), children had no *a priori* reason to privilege either cause. Study 2 therefore examines the influence of explanation in causal learning when a candidate cause that accounts for more observations is pitted against an alternative that is more consistent with prior knowledge.

Study 2 again presented two sets of causal regularities: a 100% regularity that accounted for all observations (block color), and a 75% regularity that accounted for most observations *and* presented a plausible causal mechanism (block size). Unlike color, relative size provides a plausible cause – larger objects are assumed to be more causally efficacious. This assumption was confirmed during pilot testing. We also included two age groups in Study 2 (5- and 6-year-olds). Because 6-year-olds have more experience with mechanical systems and have begun formal schooling, they could hold stronger prior beliefs in this domain.

If explaining principally drives learners to favor regularities from the data that they observe, then we would expect to replicate the results from Study 1, with explainers more likely to generalize according to the 100% regularity. If, however, explaining prompts learners to privilege causes that are more plausible in light of prior beliefs, then children prompted to explain should be less likely to generalize according to the 100% regularity, instead favoring the prior-knowledge consistent 75% regularity.

## Participants

Seventy-two children were included in Study 2, including

36 5-year-olds ( $M = 64.4$ ;  $SD = 3.8$ ; range = 60.1 – 71.7; 20 girls) and 36 6-year-olds ( $M = 78.3$ ;  $SD = 4.1$ ; range = 72.7 – 83.7; 18 girls). Eighteen 5-year-olds and 18 6-year-olds were randomly assigned to each condition. Children were recruited from preschools and museums.

## Materials

Study 2 used the same machine as in Study 1. Training stimuli consisted of four large, 3" wooden blocks and four small, 1" wooden blocks. A strip of colored electrical tape was affixed to each of the eight blocks. The four *Go* blocks had a green strip and the four *No-Go* blocks had a yellow strip. An illustration of the complete set of training blocks appears in Figure 1B. In place of the hiding box, test blocks were hidden in an opaque bag.

## Procedure

While the procedure for the training phase and test questions was similar to Study 1, there were some changes to the procedure for the testing phase. Rather than placing the test objects in the hiding box, the experimenter looked inside the opaque bag and described one feature for each of two new blocks, saying, for example, "I see a green one and I see a yellow one. Which one will make my machine play music, the green one or the yellow one?"

## Results

As in Study 1, children in both conditions were able to learn the 100% and 75% regularities. However, unlike in Study 1, children who were prompted to explain were significantly *less* likely than controls to privilege the 100% regularity over the 75% regularity when forming generalizations about novel cases, instead favoring the regularity consistent with prior knowledge (see Figure 2).

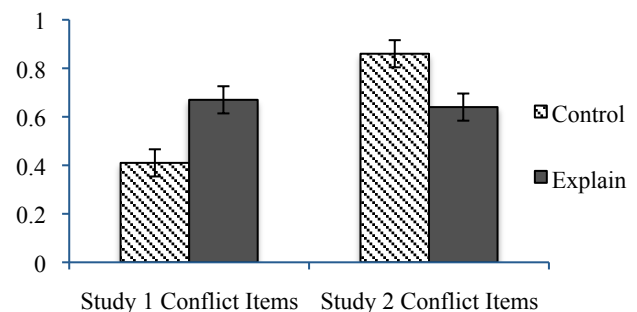


Figure 2: Mean proportion of responses consistent with the 100% regularity on the conflict items in Studies 1 and 2.

We first report results of children's performance on the 100% and 75% test items. A repeated measures ANOVA was conducted with question type (100% and 75% test items) as the repeated measure and both age (5- vs. 6-year-olds) and condition (*explain* vs. *control*) as the between-subjects variables. Analysis revealed a significant interaction between age and performance on the two question types,  $F(1, 68) = 4.46$ ,  $p < .05$ , with 6-year-olds

showing higher accuracy ( $M = .96$ ,  $SD = 0.18$ ) than 5-year-olds ( $M = .79$ ,  $SD = 0.33$ ) on the 75% test items. However, there was no main effect of condition,  $F(1, 68) = .73$ ,  $p = .40$ , and no interaction between question type and condition,  $F(1, 68) = 0.04$ ,  $p = .85$ . As in Study 1, children in both the *explain* condition ( $M = .93$ ,  $SD = .24$ ) and the *control* condition ( $M = .89$ ,  $SD = .30$ ) were able to learn the 100% and 75% regularities during the training phase, and use this information when generalizing to novel blocks.

To analyze children's performance on the *conflict items*, a univariate ANOVA was conducted with both age and condition as the between-subjects variables. As in Study 1, there was a significant difference between conditions on the *conflict items*,  $F(1, 68) = 6.46$ ,  $p < .02$ . However, results reveal a reversal of the findings. While children in both conditions responded in line with the 100% regularity more often than chance,  $p < .05$ , children in the *explain* condition were significantly *less* likely to do so ( $M = .64$ ,  $SD = .37$ ) than children in the *control* condition ( $M = .83$ ,  $SD = .27$ ). There was no difference in performance between 5- and 6-year-olds,  $F(1, 68) = .13$ ,  $p = .72$ , and no interaction between age and condition,  $F(1, 68) = 1.18$ ,  $p = .28$ .<sup>2</sup>

## General Discussion

In two studies, young children's attempts to generate explanations influenced which candidate cause they privileged when generalizing to novel cases. In Study 1, children were presented with data consistent with two causal regularities, where one accounted for more observations. Children who were prompted to explain were more likely than controls to generalize according to the regularity that accounted for more of the observed data. In Study 2, a regularity that accounted for all observations was pitted against an alternative that accounted for fewer observations but was consistent with children's prior knowledge about plausible causes. When presented with this conflict, children who were prompted to explain their observations were now *less* likely than controls to generalize the regularity that accounted for more of their observations, instead making judgments in line with prior knowledge.

Taken together, these studies shed light on the mechanisms by which explanation informs and constrains causal learning. Our findings challenge the *evidence hypothesis*, that explaining (always) makes children more responsive to evidence. If this were the case, explaining should have led children to base their judgments on the regularity that was most consistent with their current observations in both Studies 1 and 2. Our findings also challenge the *prior knowledge hypothesis*, according to which the (primary) role of explanation is to align new information with current theories. While children prompted

to explain were more likely to form a generalization in line with their prior knowledge in Study 2, children were also more likely to generalize according to their observations in Study 1, indicating that the *prior knowledge hypothesis* is limited at best, and false at worse. Instead, these findings best support the *generalization hypothesis*, that explanation prompts children to privilege observed regularities that they expect to generalize most broadly, with both observations and prior knowledge serving as cues to generalization.

The *generalization hypothesis* is broadly consistent with the "theory-theory" approach to cognitive development, and additionally provides some insight concerning how the process of theory construction and revision occurs. Within cognitive development, researchers have suggested that children's theory-like conceptual development is largely motivated by explanation, which acts as a mechanism for building abstract causal theories. In particular, Wellman and Liu (2007) propose that explanations make a particular occurrence understandable by placing it within the context of a larger, coherent framework. In so doing, explanation must accommodate both observations and prior beliefs. Relatedly, explanations have been shown to encourage adult learners to *subsume* the observation being explained under a general pattern that is expected to generalize to novel cases (Williams & Lombrozo, 2010a, 2010b). Our current findings not only provide additional support for these ideas by demonstrating that young children are also sensitive to such constraints, but additionally shed light on how explanations contribute to the formation of causal theories: Explaining helps negotiate the critical balance between prior beliefs and novel evidence, directing learners to regularities that will generalize to new cases.

Important questions for future research include precisely *how* explanation influences the evaluation of evidence and prior beliefs, and whether this influence results in judgments that are more or less closely aligned with the predictions of normative (Bayesian) models (e.g., Griffiths et al., 2011). For example, computational approaches to cognitive development have proposed that the formation of multiple levels of generalizations, or *overhypotheses*, enables learners to make principled abstractions from a class of observations, which then serves to inform future inferences about novel cases (e.g., Kemp, Perfors, & Tenenbaum, 2007). By prompting children to evaluate which candidate regularity generalizes most broadly, explanation could play a role in pushing children to consider higher-order inductive generalizations that support abstract knowledge.

The interpretation outlined thus far has focused primarily on the impact of prompts to explain on children's causal judgments. However, it is also worthwhile to consider the performance of children who were not prompted to explain. In particular, why didn't the control children in Study 2 spontaneously consult their prior knowledge in forming causal judgments? Williams and Lombrozo (under review) report similar findings in an adult population: Only learners who were prompted to explain during learning utilized informative labels in guiding their discovery of subtle

<sup>2</sup> Due to limited space, we do not report the results of qualitative analyses examining the content of children's explanations in Studies 1 and 2. These results are consistent with the quantitative data and provide additional support for our conclusions in the General Discussion.

patterns underlying novel categories. The authors suggest that explanation can guide learners to consult prior knowledge that would otherwise remain under-utilized.

The current work also suggests a number of future directions. For example, while the results reported here demonstrate the presence of an early effect of explanation on causal learning, questions remain regarding the development of this mechanism from infancy to adulthood. Additionally, future research should consider how these findings can be productively combined with previous research on the self-explanation effect (e.g., Chi, et al., 1994; Lombrozo, 2006; Siegler, 2002), as well as examining educational implications of the present findings.

In sum, these two studies provide evidence for the mechanisms by which explaining scaffolds causal learning in early childhood and serve to inform prevailing theories of learning. Learning by explaining challenges a simple data-driven view of knowledge acquisition in which children's learning is simply a function of observation and testimony. Instead, these findings provide evidence for more complex theories of learning in which processes such as explaining to oneself influence how data and current theories inform judgments. Understanding how engaging in explanation influences learning therefore contributes to a more complete understanding of how knowledge is acquired and revised.

### Acknowledgments

Thanks to Anna Akullien, Brynna Ledford, Sierra Eisen, and Rosie Aboody, the UC Berkeley Early Childhood Centers, Lawrence Hall of Science, and Habitot. Research supported by a McDonnell Foundation grant to A. Gopnik and NSF grant DRL-1056712 to T. Lombrozo.

### References

- Ahn, W.K., & Kalish, C. (2000). The role of mechanism beliefs in causal reasoning. In F. Keil & R.A. Wilson (Eds.), *Explanation and cognition*. (pp. 199-226). Cambridge, MA: MIT Press.
- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press/Bradford Books.
- Chi, M.T.H., de Leeuw, N., Chiu, M.H., LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439-477.
- Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy*, 71, 5-19.
- Gelman, S. A., & Wellman, H. M. (1991). Insides and essence: Early understandings of the non-obvious. *Cognition*, 38, 213-244.
- Gopnik, A. & Wellman, H.M. (1994). The "theory theory". In L. Hirschfield and S. Gelman (Eds.) *Domain specificity in culture and cognition*. New York: Cambridge University Press.
- Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111, 1-31.
- Gopnik, A. & Sobel, D. (2000). Detectingblickets: How Young children use information about novel causal powers in categorization and induction. *Child Development*, 71: 1205-1222.
- Griffiths, T. L., Sobel, D., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science*, 35(8): 1407-1455.
- Keil, F. (2006). Explanation and understanding. *Annual Review of Psychology*, 57: 227-254.
- Kemp, C., Perfors, A., & Tenenbaum, J.B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10, 307-321.
- Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher and W. Salmon (Eds.) *Scientific Explanation* (pp. 410-505). Minneapolis: University of Minnesota Press.
- Kuhn, D. & Katz, J. (2009). Are self-explanations always beneficial? *Journal of Exp. Child Psych*, 103, 386-394.
- Legare, C.H. (2011). Exploring explanation: Explaining inconsistent information guides hypothesis-testing behavior in young children. *Child Dev.*, 81, 929-944.
- Legare, C.H., Gelman, S.A., & Wellman, H.M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Dev.*, 81, 929-944.
- Lombrozo, T.L. (2006). The structure and function of explanations. *Trends in Cog. Sci.*, 10: 464-470.
- Schulz, L.E., Bonawitz, E.B., & Griffiths, T.L. (2007). Can being scared make your tummy ache? Naive theories, ambiguous evidence and preschoolers' causal inferences. *Developmental Psychology*, 43, 1124-1139.
- Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31-58). New York: Cambridge Univ.
- Wellman, H.M. (2011). Reinvigorating explanations for the study of early cognitive development. *Child Development Perspectives*, 5(1): 33-38.
- Wellman, H.M. & Liu, D. (2007). Causal reasoning as informed by the early development of explanations. In *Causal Learning: Psych, Philosophy, & Computation* (Eds.) L. Schulz & A. Gopnik, pp. 261-279.
- Williams, J. J., & Lombrozo, T. (2010a). The role of explanation in discovery and generalization: evidence from category learning. *Cognitive Science*, 34: 776-806.
- Williams, J. J., & Lombrozo, T. (2010b). Explanation constrains learning, and prior knowledge constrains explanation. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX.
- Williams, J. J., Lombrozo, T., & Rehder, B. (2010). Why does explaining help learning? Insight from an explanation impairment effect. *Proceedings of 32nd Annual Conference of the Cognitive Science Society*.
- Williams, J. J., & Lombrozo, T. (under review). Explanation constrains learning and prior knowledge constrains explanation.

# An Abductive Approach to Covert Interventions

**Hongbin Wang (Hongbin.Wang@uth.tmc.edu)**

School of Biomedical Informatics, University of Texas Health Science Center at Houston  
7000 Fannin Suite 600, Houston, TX 77030, USA

**Yanlong Sun (Yanlong.Sun@uth.tmc.edu)**

School of Biomedical Informatics, University of Texas Health Science Center at Houston  
7000 Fannin Suite 600, Houston, TX 77030, USA

## Abstract

We explore ways of covertly delivering interventions into the adversary decision cycles so as to effectively shape adversary decision-making and performance without inducing much suspicion. Recognizing that completely covert interventions, while most effective, are difficult to implement, we focus on a more general mode of covertness. Based on insights from human abductive reasoning, we propose a delivery scheme that contains interventions that may be noticeable but whose true meanings are hidden or distorted (e.g., the human operators do not easily attribute the interventions to malicious attacks). We evaluate, both theoretically and empirically, the effectiveness and robustness of this scheme in escaping detection and disrupting performance.

**Keywords:** Abduction, decision making, cybersecurity, intervention.

## General Formatting Instructions

A cyber attack is more damaging and harmful if it is stealthy and escaping detection. One critical challenge in cyberspace security is therefore to find ways to effectively detect hidden or covert attacks. One approach to meeting the challenge is to look at the other side of the coin and study how and why some attacks can be delivered covertly that induce no or minimal suspicion from the human operator. The results from this approach can then be used to design better countermeasures and improve security.

Here we focus on the concept of “covertness” in cyber attacks and intend to discover the theoretical essence and practical guidelines of implementing “covertness”. Presumably, a covert attack would be one that is completely hidden and not noticed by the targeted operator at all. In this sense, covertness can be implemented as slip of attention. Examples include attention blink, change blindness, and inhibition of return, to name a few. While a large body of evidence has confirmed that attention is a fragile cognitive function that can be manipulated and exploited for the purpose of implementing covertness, it has also been suggested that the attention-based approach is quite limited and difficult to apply in the real world situations.

There is at least another mode of covertness. In this mode, signs of the intervention are noticed by the human operator (therefore, the intervention is not completely hidden and escaping attention), however, the true meaning/significance of the intervention is disguised or distorted or hidden in such a way that they do not easily result in suspicion of

outside influence. This mode of covertness suggests new ways of implementing covertness.

Consider the following scenario: It is 12am and that John, an analyst, is working on a sensitive document on his computer and you have delivered a virus to his computer in order to take a peek. Ideally, you would like your operation is completely invisible to John, but unfortunately, one inevitable side effect of your virus is that John’s computer becomes slow, which *John eventually notices and starts to become suspicious*. Then John receives an alerting pop-out message informing him that the antivirus software on his computer has started scanning as scheduled and that so far no virus has been found. John now understands why his computer becomes slow, is relieved, and continues to work on his document, without realizing your peeking eyes.

Though hypothetical, this example highlights an important aspect of covertness, which has to do with an understanding of how a human operator reasons and explains unexpected observations and if and when the operator becomes suspicious given data. In the example, John becomes suspicious when he notices that his computer slows down, an often-inevitable indicator of attacks. But his suspicion fades away after the pop-out message, which is also delivered by the attacker with the goal to provide a better explanation for the slow-down so as to reduce John’s suspicion.

Instead of directly exploiting the low-level attentional function, this mode of covertness depends upon exploiting a higher-level human inference system called abduction. We argue that this mode of covertness is more general, more realistic and potentially more powerful.

## Abduction-based Covertness

Abduction was introduced by American philosopher Charles S. Peirce (1839-1914) as a form of human inference that is different from deduction and induction. According to Peirce, in abduction one infers causes from effects or explanations from observations (See Fann, 1970 for a general introduction to Peirce’s theory of abduction). The general form of abduction is shown below,

*A fact C is observed,*

*H can explain C;*

*Hence, H may be true.*

Here is a specific example of abductive inference, in the context of the hypothetical scenario above,

*The computer suddenly slows down,  
A malicious attack explains the slowdown;  
Hence, a malicious attack may be occurring.*

And here is another example,

*The computer suddenly slows down,  
Antivirus scanning explains the slowdown;  
Hence, nothing is wrong and just be patient.*

Charniak and McDermott (1985) characterize abduction as *modus ponens* turned backward (see also Brachman & Levesque, 2004). It is clear in abduction the conclusion does not necessarily follow the premises – in the above examples two different explanations are inferred to explain a same observation. However, according to Peirce, abduction is important in that it "is the only logical operation which introduces any new ideas; for induction does nothing but determine a value [to classify], and deduction merely evolves the necessary consequences of a pure hypothesis" (Peirce, 1931, v. 5, p. 171). Though inconclusive, the explanation inferred by abduction "is adopted for some reason, good or bad, and that reason, in being regarded as such, is regarded as lending the hypothesis some plausibility" (Peirce, 1931, v. 2, p. 511).

Modern researchers often regard abduction as a complex process of finding a best explanation for a set of observations (Josephson & Josephson, 1994; Paul, 1993; Thagard, 1992). Since "explaining" is such an inevitable aspect of human everyday activities, abductive reasoning is almost ubiquitous, ranging from hearing the thunder ("It's going to rain?"), seeing a falling maple leaf ("Autumn has come?"), to medical diagnosis (from symptoms to diseases) and scientific discovery (from data to knowledge and theories). In battlefields, commanders have to infer the enemy's motivations based on observations and intelligence and then take proper actions. In cyberspace security, operators have to infer if an attack has occurred given observations.

How do people do abduction? The Theory of Explanatory Coherence (TEC) is an influential theory of human abduction (Thagard, 1989, 1992). According to TEC, abduction is a parallel constraint satisfaction process in that all propositions, including explanations, evidence, and explanatory relations, form a network that constantly seeks harmony. An explanation should be accepted if it is coherent with all other propositions in the network, rejected if it is incoherent, and the best explanation for available observations is the one that enjoys the most explanatory coherence in the network. TEC proposes seven principles that establish explanatory relations among propositions and regulate the global coherence of an explanatory system: (1) *symmetry*: If  $P$  and  $Q$  cohere, then  $Q$  and  $P$  cohere; If  $P$  and  $Q$  incohere, then  $Q$  and  $P$  incohere. (2) *explanation*: If  $P_1...P_m$  explain  $Q$ , then  $P_1...P_m$  cohere with each other and with  $Q$  cohere, and the degree of coherence is inversely proportional to the number of propositions  $P_1...P_m$ . (3) *Analogy*: If  $P_1$  explains  $Q_1$ ,  $P_2$  explains  $Q_2$ ,  $P_1$  is analogous to  $P_2$ , and  $Q_1$  is analogous to  $Q_2$ , then  $P_1$  and  $P_2$  cohere, and  $Q_1$  and  $Q_2$  cohere. (4) *data priority*: Observations have a

degree of acceptability of their own. (5) *Contradiction*: If  $P$  contradicts  $Q$ , then  $P$  and  $Q$  incohere. (6) *competition*: If  $P$  and  $Q$  both explain a proposition, and if  $P$  and  $Q$  are not explanatorily connected, then  $P$  and  $Q$  incohere. (7) *acceptability*: The acceptability of a proposition  $P$  depends on its coherence with all the propositions in the system.

TEC has been computationally implemented in a connectionist system called Echo (Thagard, 1992). In Echo, propositions (both data and hypotheses) are represented by nodes. Coherence relations are represented by excitatory links and incoherence relations are represented by inhibitory links. Node activation represents the node's degree of coherence with all propositions in the network. The system updates itself based on parallel constraint satisfaction (Thagard, 1992). During this process, propositions that are incoherent die out and propositions that are coherent are strengthened. In the end, the most activated propositions represent the most plausible and coherent explanations. Echo has been extended to UEcho to incorporate more sophisticated handling of uncertainty (Wang, Johnson, & Zhang, 1998; Wang, Johnson, & Zhang, 2006).

TEC and UEcho capture several critical constraints in abduction, including explanatory breadth (the model prefers a hypothesis that explains more); simplicity (the model prefers a simpler hypothesis); being explained (the model prefers a hypothesis which itself is explained); data reliability (the credibility of an observation also depends on its coherence in the system); and analogy (analogous hypotheses are coherent). More important, however, they shed interesting new insights on human suspicion and implementing covertness. In this context, suspicion can be viewed as the degree of acceptance of an explanation such as "a malicious attack is occurring", and implementing covertness is not much more than to make the degree of acceptance of this explanation as low as possible.

Based on this reasoning, we hypothesize that effective covert interventions can be delivered in such a way that suspicion-bearing explanations (e.g., "a malicious attack is occurring") cannot become the best (winning) explanation given data. TEC has already offered several straightforward ways to do just this. For example, one way is to "explain away", which says that when delivering an intervention, deliver an explanation for that intervention as well so that the true meaning of observations can be shielded. This is exactly what happens in our previous hypothetical example. an attack is delivered, which caused John's computer to be slower. In anticipating this, a secondary message is delivered to John to "explain" to him that why his computer became slower. This new explanation "explained away" the John's observation and therefore reduced his suspicion – that is, the acceptance of "an attack is occurring" became low. Another example of abduction-based covertness derives directly from the data reliability principle – we can discredit those "suspicion-inducing" observations by introducing conflicting data ("unreliable data"). "Are you sure that your computer becomes slower?" By inducing new

data to promote John to cast his doubt, the suspicion level of “an attack is occurring” is reduced.

To a certain extent the attention-based covertness (i.e., delivering interventions that are invisible to the human attention) is a special case of this new abduction-based covertness. Since abduction starts with observations (that is, the data to be explained, e.g., “the fact C is observed”), completely invisible interventions suggest that suspicion-generating abduction will not even be starting in the first place. However, abduction-based covertness is more general in the sense that in case some suspicion-inducing observations become available, covertness can still be achieved if proper measures are taken so that the suspicion-bearing explanation will not become the most plausible one.

### Stealth and Disruption with IMPs

As a preliminary step toward evaluating the effectiveness of abduction-based covertness, we conducted a study to examine how a human operator digests unexpected interventions and adjusts his level of suspicion. The study utilized a so-called Interface Manipulation Protocol (IMP). The toolbox of IMP contained dozens of possible intervention types that could be delivered to the adversary computers to cause disruption with, for example, keyboard and mouse operations. We were interested in finding out the optimal chain of IMP delivery scheme (e.g., when to deliver what IMP for how long?) that causes maximal disruption with minimal suspicion.

### Method

#### Participants

Nine college students and graduate students in the Houston medical center area were paid to participate in the experiment.

#### Procedure

The experiment was programmed in E-Prime and conducted on a computer with a 20 inch LCD monitor. Subjects were instructed to type in sequences of random numbers as prompted (Figure 1). Table 1 shows three independent variables manipulated in the study, including the type of IMPs and the type of delivery themes.

Target  
Sequence: 2 5 9 8 ... 4 6  
Responses: 2 5 7 8 ...

Figure 1. Subjects were required to reproduce the target sequence. Errors were prompted in red color and need to be corrected with extra keystrokes. Errors could include “IMP errors” (produced deliberately by IMPs) and “genuine errors” (subjects’ own typos). In this example, the subject had mistyped the target character number “9” with number “7”, and subsequently typed number “8” before realizing the error.

At the beginning of each trial, subjects were first prompted with a sequence of 20 characters of random numbers, shown at the top of the computer screen. Then, they were to copy the entire sequence in exactly the same order, and their responses were shown one by one for each keystroke, in a separate line below the target sequence. If an error occurred, either by subjects’ own error (“genuine errors”) or by deliberate IMP interventions, the mismatched character would be shown in red. Subjects were instructed to immediately erase the error by using the backspace key. If subjects have skipped the error for several keystrokes, they had to erase all subsequently typed characters (including the correct ones). That is, correcting an error had to be done in a backward sequential order (similar to the situation of typing documents without the ability of adjusting the cursor position by mouse).

Table 1. Independent variables manipulated

<b>3 types of IMPs</b> <ul style="list-style-type: none"> <li>Non-responsive key (IMP1): when a key is pressed, nothing shows up, so the subject has to retype to correct;</li> <li>repetitive key (IMP2): when a key is pressed, the same character will show up twice. For example, typing “3” would result in “33” shown on the screen, so that the subject has to erase the extra character;</li> <li>altered key (IMP3): when a key is pressed, a randomly selected different character is shown (e.g., typing “3” and “4” shows up), so the subject has to erase the wrong character and retype.</li> </ul>
<b>4 types of delivery themes</b> <ul style="list-style-type: none"> <li>Pure: only one type of IMPs is delivered.</li> <li>Mixed: multiple types of IMPs are delivered.</li> <li>Clumped: IMPs are delivered consecutively.</li> <li>Dispersed: IMPs are delivered sparsely.</li> </ul>
<b>4 experimental conditions (combination of themes)</b> <ul style="list-style-type: none"> <li>PC: Pure-Clumped.</li> <li>PD: Pure-Dispersed.</li> <li>MC: Mixed-Clumped</li> <li>MD: Mixed-Dispersed</li> </ul>
<b>4 levels of IMP delivery rates</b> 10%, 20%, 30%, 40% of the target characters are affected by IMPs).

Three types of IMP interventions were silently delivered by hijacking the subject’s keyboard (see Table 1 and Figure 2). Each delivery of IMP intervention was designed to affect only one keystroke. For instance, by IMP1 (non-responsive key, Figure 2A), when the subjects typed any key on the keyboard (not necessarily matching the target character), the program would silently remove the keystroke such that no response character would be shown below the target character. Then, the IMP intervention would be temporarily



“disarmed” for this particular target character. Only when subjects pressed the same key for second time, the character would show up in the response line.

Four delivery themes were designed based on the mixture of IMP types and the temporal intervals between each IMP intervention (Table 1). In the “Pure” theme, only one IMP type was implemented for the target sequence. In the “Mixed” theme, all three types of IMPs were implemented. In the “Clumped” theme, IMPs were clustered together such that one IMP intervention could be immediately followed by another. In the “Dispersed” theme, IMPs were evenly distributed among the 20 target characters.

The delivery themes were grouped into 4 experimental conditions with each condition containing one particular combination of the mixture and temporal distribution (Table 1). For example, in the “PC” condition, only one type of IMP was delivered but in a clustered fashion. We also implemented 4 levels of IMP delivery rates, which were evenly distributed in each of the delivery themes.

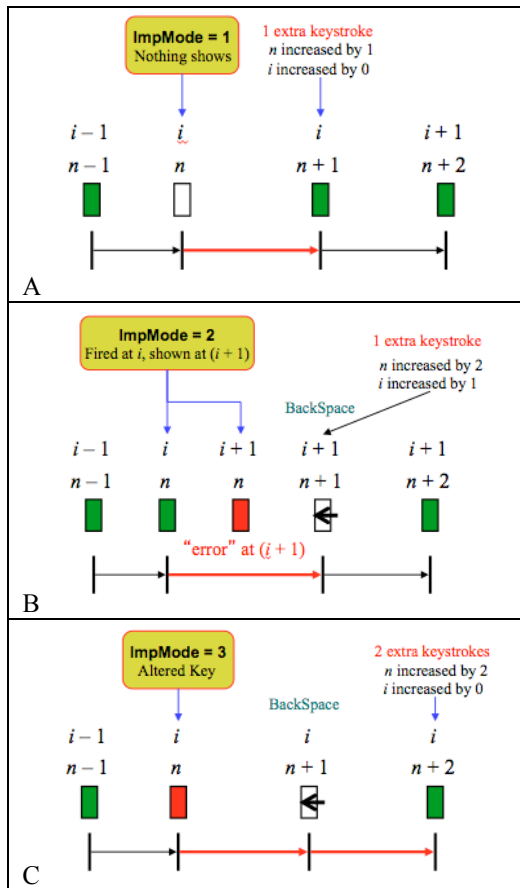


Figure 2. How types of IMPs (A: IMP1; B: IMP2; C: IMP3) affect dependent measures.  $i$  = position in the target sequence;  $n$  = number of keystrokes.

After practice trials, each subject completed 4 blocks of trials, with each block corresponding to one of the experimental conditions. The order of the conditions was randomized between subjects. Each block consisted of 20 trials, and each trial consisted of 2 target sequences. At the

end of each block, subjects were asked to evaluate the “reliability” of the input device on an 1-to-7 scale with “1” for the most “unreliable” and “7” for the most “reliable”. Subjects took a brief break before moving to the next block of trials.

There were two major dependent measures. Stealth (covertiness) was measured by the subjective evaluations of the reliability of the input device. Higher evaluation scores indicated higher tolerance of IMPs and therefore less suspicion. Disruption was measured by the number of extra keystrokes (“ExtraKS”) required to complete the sequence (excluding the extra keystrokes directly caused by IMPs). Higher scores of ExtraKS indicated more severe disruptions to the performance. The relation between IMPs and dependent measures is depicted in Figure 2.

## Results

One main result of the study is shown in Figure 3, which depicts the effect of delivery themes on stealth and disruption. Statistics show that in terms of stealth the mixed-clumped delivery (IMPs with mixed types are delivered continuously) is the best (mean evaluation scores = **4.36**, 3.94, 3.76, 4.01, with standard errors = 0.22, 0.31, 0.19, 0.25, for MC, MD, PC and PD, respectively). And in terms of disruption the pure-dispersed delivery (IMPs with the same type are delivered sparsely) is the best (mean disruption scores = 2.06, 3.08, 2.62, **3.50**, with standard errors = 0.59, 0.58, 0.47, 0.73, for MC, MD, PC and PD, respectively). Further analysis shows that if we combine the two dependent measures, the pure-dispersed delivery has the highest effectiveness score, as shown in Figure 4 (mean effectiveness scores = 0.43, 0.51, 0.49, **0.56**, with standard errors = 0.03, 0.02, 0.03, 0.02, for MC, MD, PC and PD, respectively). Overall, it seems that the pure-dispersed delivery is the most effective IMP delivery theme if the tradeoff between stealth and disruption is considered.

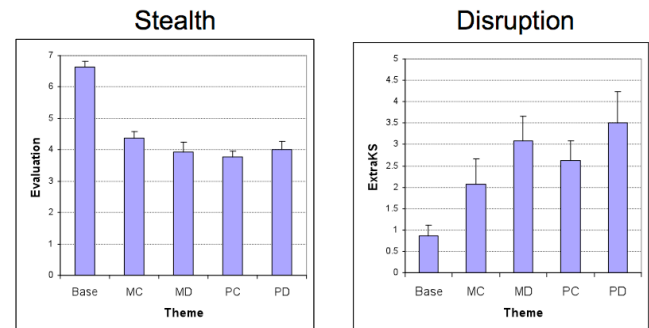


Figure 3. The effect of delivery themes (Base: no IMP was delivered; MC: IMPs were delivered in mixed-clumped fashion; MD: mixed-dispersed; PC: pure-clumped; PD: pure-dispersed).



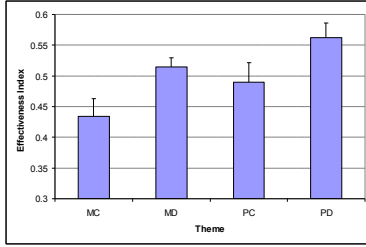


Figure 4. Effectiveness by themes. The effectiveness index is computed by adding the normalized stealth and disruption measures.

The effect of IMP types is shown in Figure 5. In terms of stealth, statistics show that  $IMP1 > IMP2$  (mean evaluation scores = 4.34, 3.70, with standard errors = 0.21, 0.23, for  $IMP1$  and  $IMP2$ , respectively,  $p < 0.05$ ) and  $IMP1 > IMP3$  (mean evaluation scores = 4.34, 3.62, with standard errors = 0.21, 0.31, for  $IMP1$  and  $IMP3$ , respectively,  $p < 0.05$ ). In terms of disruption, no significant difference is found (mean disruption scores = 2.63, 3.21, 3.34, with standard errors = 0.44, 0.68, 0.71, for  $IMP1$ ,  $IMP2$  and  $IMP3$ , respectively).

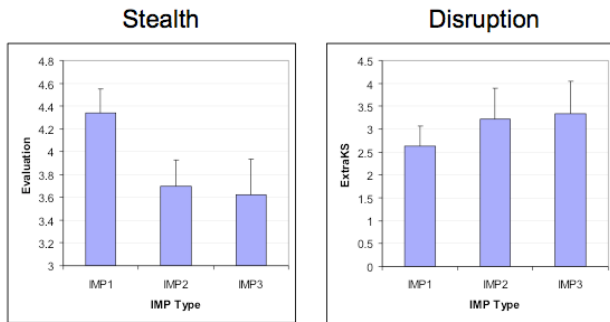


Figure 5. Effect of IMP types on stealth and disruption.

The effect of IMP delivery rate is shown in Figure 6. It is clear that with the rate increase the evaluation scores decrease (mean evaluation scores = 6.64, 5.12, 4.28, 3.58, 3.10, with standard errors = 0.17, 0.15, 0.23, 0.25, 0.28, for 0, 10%, 20%, 30% and 40%, respectively) and the disruption scores increases (mean disruption scores = 0.87, 1.77, 2.75, 2.97, 3.76, with standard errors = 0.25, 0.37, 0.50, 0.81, 0.59, for 0, 10%, 20%, 30% and 40%, respectively). A nonlinear regression supports the notion that rate increases led to more disruption and less stealth.

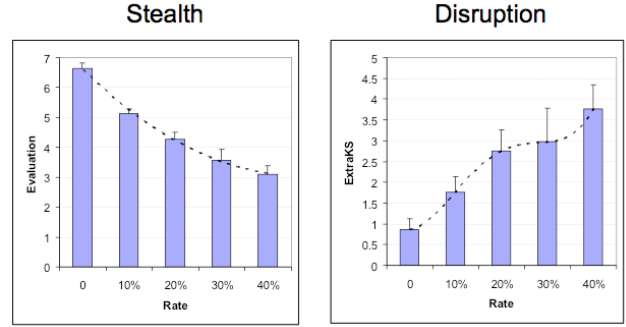


Figure 6. Effect of IMP delivery rates on stealth and disruption, with polynomial fitting curves.

## Summary and Discussion

In this article we explore ways of covertly delivering interventions into the adversary decision cycles so as to effectively shape adversary decision-making and performance without inducing much suspicion. The focus here is not on the delivery technology, which we assume can be achieved, but on the covertness. That is, how can we deliver interventions that do not induce significant suspicion and effectively shape operators' behavior?

Attention is often the first cognitive faculty explored in the attempt to understand covert interventions. On the one hand, there are hardly better ways to implement covertness than designing interventions that are invisible even to the adversary operator's attentional system. In this case, the interventions are completely hidden and therefore can potentially cause most and long-term damage. On the other hand, a large body of evidence in the field of psychology has shown that attention is a fragile function that is subject to exploitation and manipulation. A recent theoretical breakthrough of attention research is the notion that there exist different types of attention, each of which is subserved by different brain regions and is sensitive to different variables (Fan, McCandliss, Sommer, Raz, & Posner, 2002; Posner, 2004). Equipped with the taxonomy, it has been suggested that each type of attention could be subject to different exploitations for the purpose of covertness. Studies have been conducted to systematically examine the effect of parameter changes on inducing covertness and affecting performance (Sun, Wang, Zhang, & Smith, 2008; Wang & Fan, 2007; Wang, Liu, & Fan, 2012).

Recognizing the limitation of attention-based covertness in real world situations, in this article we propose to a more general approach to covertness. That is, instead of delivering completely hidden interventions, it is possible to deliver interventions that may be noticeable but whose true meanings are hidden or distorted. Consequentially the similar effect of covertness can be achieved. This approach, based on insights from human abductive reasoning rather than straightforward attentional manipulations, is easier to implement and potentially more powerful. However, the success of the approach would require a better understanding of adversary decision processes and more sophisticated delivery strategies. The study reported in this

article is a step towards developing and evaluating guidelines and schemes for such deliveries.

In the experiment we manipulated the type of interventions and the delivery themes. In particular, we distinguished pure vs mixed and clumped vs dispersed deliveries. We evaluated the effect of these manipulations on suspicion and performance. Our results support the general notion of abduction-based covertness. We show that covertness can be achieved even when interventions are detected as long as they are not properly explained. Our results demonstrate that different intervention types have different effectiveness. And more important, we show that pure-dispersed delivery scheme is more effective than the other delivery schemes, suggesting that when delivering interventions, to achieve effective stealth and disruption, try to keep the interventions dispersed and do not mix different types of interventions.

In sum, we demonstrate that abduction is a sound and insightful framework for understanding human reasoning in general and human suspicion in particular. Techniques such as “explaining away” and “data reliability” are powerful in manipulating suspicion and implementing covertness. Additional work is clearly needed for a deeper theoretical understanding of the underlying cognitive process and more comprehensive guidelines for covert intervention delivery in real-world situations.

### Acknowledgments

This work was partially supported by an Air Force Office of Scientific Research (AFOSR) grant (FA9550-07-1-0181), and an Office of Naval Research (ONR) grant (N00014-08-1-0042). We would like to thank Scott Thompson and Leanne Hirshfield for the IMP conceptualization.

### References

- Brachman, R., & Levesque, H. (2004). *Knowledge representation and reasoning*. San Francisco, CA: Morgan Kaufmann.
- Charniak, E., & McDermott, D. (1985). *Introduction to artificial intelligence*. Reading, MA: Addison-Wesley Publishing Company.
- Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, 14(3), 340-347.
- Fann, K. T. (1970). *Peirce's theory of abduction*. The Hague: Martinus Nijhoff.
- Josephson, J. R., & Josephson, S. G. (1994). *Abductive inference: Computation, Philosophy, Technology*. Cambridge, NY: Cambridge University Press.
- Paul, G. (1993). Approaches to abductive reasoning: An overview. *Artificial Intelligence Review*, 7, 109-152.
- Peirce, C. S. (1931). *Collected papers* (Vol. 1-6). Cambridge, MA: Harvard University Press.
- Posner, M. I. (Ed.). (2004). *Cognitive neuroscience of attention*. New York: Guilford Press.
- Sun, Y., Wang, H., Zhang, J., & Smith, J. W. (2008). Probabilistic judgment on a coarser scale. *Cognitive Systems Research*, 9(3), 161-172.
- Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12, 435-502.
- Thagard, P. (1992). *Conceptual revolutions*. Princeton, N.J.: Princeton University Press.
- Wang, H., & Fan, J. (2007). Human attentional networks: A connectionist model. *Journal of Cognitive Neuroscience*, 19(10), 1678-1689.
- Wang, H., Johnson, T. R., & Zhang, J. (1998). UEcho: A model of uncertainty management in human abductive reasoning. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 1113-1118). Hillsdale, NJ: Lawrence Erlbaum.
- Wang, H., Johnson, T. R., & Zhang, J. (2006). The order effect in human abductive reasoning: An empirical and computational study. *Journal of Experimental and Theoretical Artificial Intelligence*, 18(2), 215-247.
- Wang, H., Liu, X., & Fan, J. (2012). Symbolic and connectionist models of attention. In M. I. Posner (Ed.), *Cognitive Neuroscience of Attention* (pp. 47-56). New York: The Guilford Press.

# Choosing quantity over quality: syntax guides interpretive preferences for novel superlatives

Alexis Wellwood (wellwood@umd.edu)

Department of Linguistics  
1401 Marie Mount Hall  
College Park, MD 20742 USA

Justin Halberda (halberda@jhu.edu)

Department of Psychological and Brain Sciences  
3400 N. Charles Street  
Baltimore, MD 21218 USA

Darko Odic (darko.odic@jhu.edu)

Department of Psychological and Brain Sciences  
3400 N. Charles Street  
Baltimore, MD 21218 USA

Jeffrey Lidz (jlidz@umd.edu)

Department of Linguistics  
1401 Marie Mount Hall  
College Park, MD 20742 USA

## Abstract

Acquiring the correct meanings of number words (e.g., *seven*, *forty-two*) is challenging, as such words fail to describe salient properties of individuals or objects in their environment, referring rather to properties of *sets* of such objects or individuals. Understanding how children succeed in this task requires a precise understanding not only of the kinds of data children have available to them, but also of the character of the biases and expectations that they bring to the learning task. Previous research has revealed a critical role for language itself in how children acquire number word meanings, however attempts to pinpoint precisely the strong linguistic cues has proved challenging. We propose a novel “syntactic bootstrapping” hypothesis in which categorizing a novel word as a determiner leads to quantity-based interpretations. The results of a word learning task with 4 year olds indicates that this hypothesis is on the right track.

**Keywords:** Number, quantity, language acquisition, learning, determiners, adjectives, quantifiers, syntax.

## Words for quantities

While it is uncontroversial that young children necessarily make use linguistic and extralinguistic information when they set about learning the meaning of novel words, the idea that some pairing of “situation and sound” is *sufficient* has been repeatedly questioned (e.g., Landau & Gleitman 1985, Waxman & Lidz 2006). An especially difficult problem for any view that posits a simple mapping from a portion of experience to the meaning of a novel word has been the acquisition of number words (e.g., *five*, *sixty-seven*). This is particularly challenging as numbers refer to properties of *sets* of objects rather than to properties of any object in particular (Frege 1893; Bloom & Wynn, 1997). Understanding how number words are learned must be informed not only by a precise understanding of the kinds of data children have available to them, but also of the character of the biases and expectations they bring to the learning task.

In this paper, we consider the question of how children decide that a novel word describes numerosity and not some other salient property. To take a simplified example of the problem, consider the novel word *gleeb* in (1).

- (1) The *gleeb* cows are by the barn.

The novel word *gleeb* appearing in an adjectival position may describe any number of properties relevant to the cows, for

instance their color (e.g., they may be *blue*), texture (e.g., especially *soft*), size (e.g., for cows, quite *big*), or even their number (e.g., approximately *many*, or exactly *seven*). Under what circumstances would a child prefer to assume that number is the intended property? Finding such circumstances would aid researchers in finding the unique biases and expectations children must bring to the task in order to accurately acquire number words.

Research on the acquisition of exact number words suggests that language itself must provide critical support for the child to map new words onto such abstract meanings. Wynn (1992; see also Condry & Spelke 2008) found that children at 2;6, who do not yet understand the relationship between the words in the count list and exact cardinalities, nevertheless understand that the number words describe numerosities. This result is striking, as it takes children another full year to gain the knowledge that which exact quantities are intended (Wynn 1992, Carey 2009). Bloom and Wynn (1997) examined the distribution of the numerals in the CHILDES database of child-directed speech to determine what syntactic cues might prompt quantity-based interpretations. They proposed that the appearance of an item in the partitive frame (e.g., as *X in X of the cows*) was a strong cue to number word meaning. The plausibility of such a view is bolstered by the linguistics literature: partitivity has been said to signal to the semantic role of quantification (Jackendoff 1977).

This proposal was recently investigated by Syrett, Musolino and Gelman (2012). Conducting their own corpus study, they pointed out that a great variety of non-quantity-referring expressions occur in the partitive frame (2), so perhaps we should not expect it to be a strong cue to numerical meanings.

- (2) a. **Amount:** *all, two, seven, most, some*  
b. **Segment:** *back, front, edge, side, top*  
c. **Measure:** *mile, hour, pound, bucket*

Regardless, if the partitive were a strong cue, then a novel word embedded in the partitive should lead children to pick a quantity-based interpretation even when the environment supports both this and an alternative interpretation. That is, in a novel word learning task, the novel word *pim* appearing with

the partitive (as in *pim of the trains*) should be analyzed as referring to the quantity TWO but not the quality RED<sup>1</sup> when both interpretations were supported. Syrett et al found that the partitive predicted quantity-based judgments only in restricted cases,<sup>2</sup> casting doubt on the robustness of a “syntactic bootstrapping” account based on the partitive as a strong cue.

The puzzle raised by Wynn’s (1992) original finding remains. Indeed, it appears to raise the question whether it may be necessary to understand how children decide that novel words describe quantities *at all* before we can understand how they learn the meanings of words for *exact* numerosity (see also Barner, Chow & Yang 2009 for discussion). The numerals pattern with a larger class of expressions in natural language called *quantifiers* (i.e., the Amount terms in (2)), which share a similar syntactic distribution. If a child could figure out (as Bloom & Wynn suggested) that a certain bit of syntax corresponded in a stable fashion with the semantics of quantity, they might have a foothold on deciding a novel word referred to numerosity. To get a sense of the problem, consider the sentences in (3a) and (3b) with the novel word *gleebest* against the image in Figure 1.

- (3) a. The gleebest cows are by the barn.
- b. Gleebest of the cows are by the barn.

For adults, *gleebest* in (3a) could in principle describe something about the numerosity or some other property of the cows by the barn in contrast to those in the field. Indeed, the meaning one perceives is similar to that conveyed either by the familiar *most* or e.g. *spottiest*. However, if adults were exposed to the novel word in the syntactic context given in (3b), they would *never* suppose it to designate something about the spottiness of the cows by the barn, only their numerosity.

Adults, however, have had a lifetime of language experience. Under what conditions would a child still in the process of mastering their native tongue hypothesize that *gleebest* means MOST as opposed to SPOTTIEST? Would their pattern of preferences be the same if presented with either of the sentences in (3a) or (3b)?

While the evidence for the partitive frame (*...of the cows*) as a strong cue to quantity-based meanings is mixed, we think pairs of sentences like the above suggest that a stronger cue might be whether something occurs to the left of *X*. Of the classes of counterexamples provided by Syrett et al given in (2), we may note that only amount terms can appear without a determiner (*a* or *the*) on the left:

- (4) a. Two/most of the cows lowed.
- b. \* Back/side of the fridge is blue.
- c. \* Mile/hour of the race was hard.

<sup>1</sup>We follow custom in using italics for linguistic expressions and small caps as shorthand for their meanings.

<sup>2</sup>Only when it was used at test; when the partitive was used during training but not at test, children were at chance at picking the quantity interpretation.

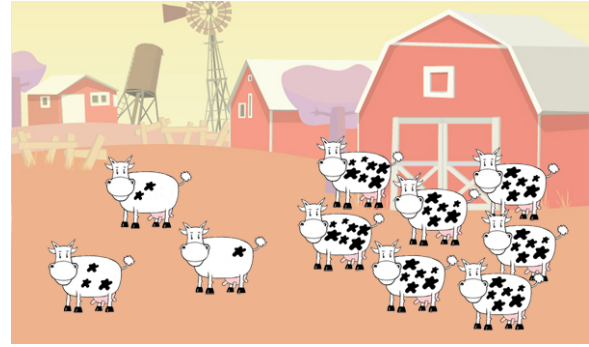


Figure 1: **The gleebest cows** are by the barn.

Such data illustrate an important linguistic generalization. Amount terms, or *quantifiers*, can occur in a privileged syntactic position (e.g., Barwise & Cooper 1981): that of determiners, instantiating the category D. Unlike the partitive frame, Ds have a stable syntax-semantics mapping: their interpretation only references quantities, never qualities, of individuals (van Benthem 1989, Gajewski 2002).<sup>3</sup> Observing this pattern leads us to a novel hypothesis: if a child categorizes a novel word as D, she will understand that word to have a quantity- rather than quality-based meaning.

A test of this would be to make both numerosity and spottiness salient, and test children’s preferences for interpreting a novel word across syntactic contexts. To construct such a test, we turn to superlatives. As we will see, superlatives (the result of combining a word like *heavy* with the morpheme *-est*) allow for a direct comparison of the hypothesis that syntactic category, and not partitivity, is a strong cue to positing quantity-based meanings.

Combining a word with a quality-based meaning like *heavy* with the morpheme *-est* allows the formation of expressions like *the heaviest animals*, with a meaning like THE ANIMALS THAT ARE HEAVIER THAN ANY OTHERS. Similarly, combining *many* with *-est*<sup>4</sup> gives *the most animals*, with a meaning like THE ANIMALS THAT ARE MORE NUMEROUS THAN ANY OTHERS. Importantly for our purposes, both of these types of superlatives surface in the position of an adjective (5a) (where *the* instantiates the syntactic category D), but only the quantity-based superlative *most* can appear bare on its left, instantiating D (contrast (5b) with (5c):

- (5) a. The heaviest/most animals are happy.
- b. Most of the animals are happy.
- c. \* Heaviest of the animals are happy.

<sup>3</sup>As a simple rule to determine which word in a string is D, take *X* in *X of the cows* to be D unless *the* precedes *X*. Since *the* cannot appear without an element to its right before *of* (cp. \**the of the cows*), it instantiates D whenever it is present. In *the most cows*, *the* instantiates D, but *most* instantiates D in *most of the cows*.

<sup>4</sup>It is widely assumed that *most* is the superlative of *many*, following Bresnan 1973; cf. Bobaljik 2007 who argues *most* is *more+est*.

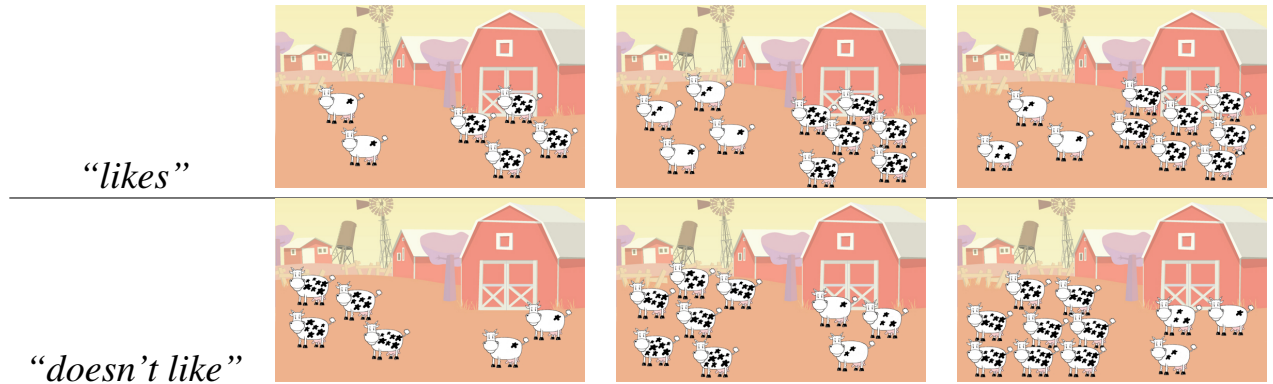


Figure 2: Ambiguous training cards.

This restriction can’t be conceptual: where we understand the sentence in (5b) to mean MORE THAN HALF OF THE ANIMALS BY NUMBER ARE HAPPY, by analogy we might have expected (5c) to mean MORE THAN HALF OF THE ANIMALS BY WEIGHT ARE HAPPY. To see what this would mean, consider a situation in which the only animals are a cow *C*, a lamb *L*, and a rabbit *R*. It is clear that (5b) is true if any two of the animals are happy. But (5c) requires more information: if *C* weighs 700kg, *L* weighs 35kg, and *R* weighs 8kg, we would know (5c) is true **only if *C* is happy**. Individuals and their particular properties matter for quality-based superlatives, where only set cardinality matters for *most*. While it is clear that no conceptual necessity rules out a determiner-like meaning for a quality-based adjective, why it is excluded remains a mystery.<sup>5</sup>

Lastly, regardless of whether they have quantity- or quality-based meanings, superlatives can appear in the partitive frame:

- (6) a. \* The spottiest of the cows were by the barn.
- b. \* The many of the cows were by the barn.
- c. The spottiest of the cows were by the barn.
- d. The most of the cows were by the barn.

In the next section, we put our hypothesis that syntactic category cues category meanings to the test in a novel word learning task. At the same time, we contrast this hypothesis with that suggesting partitivity as a strong cue.

### Testing superlatives

In the previous section, we hypothesized that representing a novel word as an instance of the category D was a strong cue to the learner that the word should be assigned a quantity-based meaning. An alternative was presented that suggested presence of the partitive frame alone was a strong cue. We test

<sup>5</sup>This is especially surprising, given recent proposals in the formal semantics literature that nothing much *semantically* distinguishes *most* from *spottiest* (Hackl 2009). Yet, it is difficult to see how appeal to numerosity would be possible in formulating a *syntactic* constraint to make sense of facts like (5b)-(5c).

these ideas by examining children’s preferences when embedding *gleebest* in a variety of syntactic contexts, using a variant of the Picky Puppet task (Waxman & Gelman 1986).

### Method

In this task, the experimenter first explains that the game is to sort cards according to whether a puppet likes them or not. The puppet is described as picky, but friendly enough to share the reasons for why he likes what he does. The experimenter explains the puppet’s criterion by showing preferred and dispreferred cards along with the sentence: “The puppet said he likes the cards where **target sentence**, but he doesn’t like the ones where it’s not true that **target sentence**”. The target sentence always contained the novel word *gleebest* (see Table 1). The experimenter explains that she doesn’t know what *gleebest* means, but was hoping the child could help her figure it out. In the Training phase, the child is shown 6 training cards (Figure 2), the ones the puppet had “already told” the experimenter it liked or didn’t like.

Table 1: **Target sentences:** “*The puppet likes the cards where DP are by the barn.*”

cond	DP	<i>the</i>	partitive
ADJ	<i>the gleebest cows</i>	✓	×
CON	<i>the gleebest of the cows</i>	✓	✓
DET	<i>gleebest of the cows</i>	×	✓

While the training cards are perfectly ambiguous (the group by the barn is both the most numerous and the most spotty), the test cards are perfectly *unambiguous*.<sup>6</sup> The same cards (in counterbalanced order) were presented to each participant. The form of the target sentence was our between-subjects factor, with *gleebest* appearing in adjectival (ADJ), confounded (CON), and determiner (DET) positions, so our conditions feature different combinations of presence/absence of *the* and the partitive, as schematized in Table 1.

<sup>6</sup>For our test cards, the ratio of the numerosities of the cows was inversely proportional to the ratio of the spots of the cows.



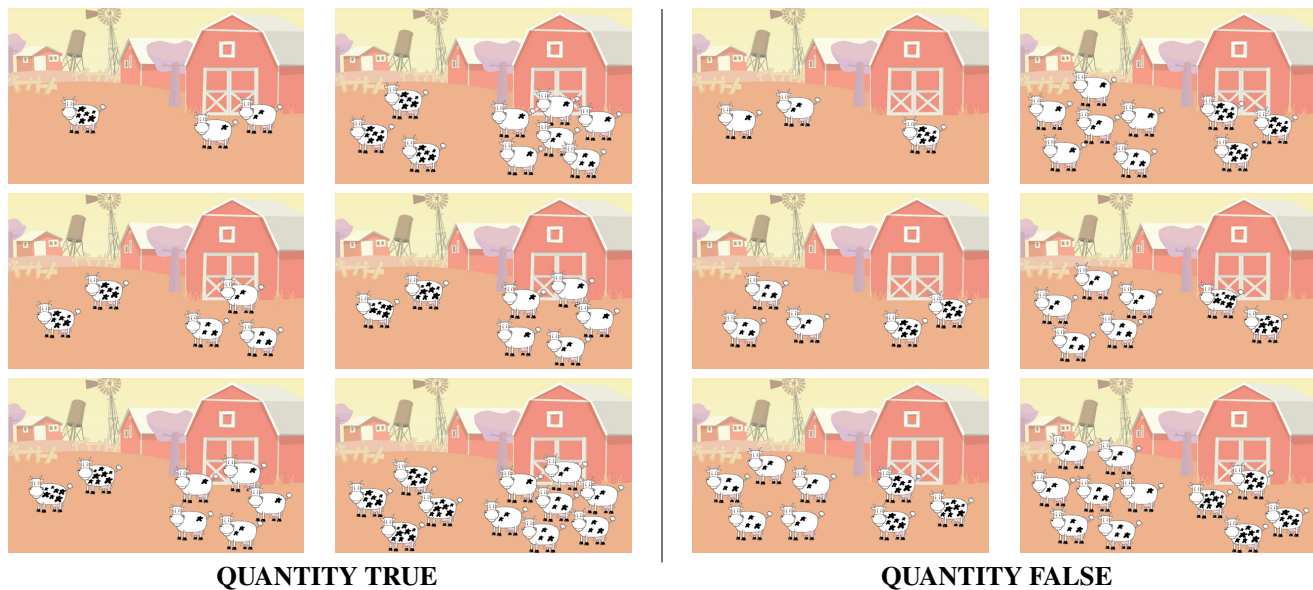


Figure 3: Unambiguous test cards.

At the beginning of the Test phase, the experimenter handed each test card to the child with the question “Do you think he likes this one?”. The child was to place each card below a green circle with a checkmark on it if the puppet likes it, and below a red circle with a black X if he doesn’t like it. At the end of the experiment, the child was probed as to what s/he thought *gleebest* meant, and responses were recorded.

We hypothesized that categorizing a novel word as D restricts a child’s hypothesis about the word’s meaning to a quantity-based interpretation. Another proposal was that the presence of the partitive frame itself was a strong cue to such interpretations. Thus the relevant hypotheses are schematized in Table 2 according to whether they predict a greater-than-chance quantity-based response (indicated by +).

Table 2: Predicted neutral (–) versus increased quantity-based responses (+).

Hypothesis	ADJ	CON	DET
Category as strong cue	–	–	+
Partitive as a strong cue	–	+	+
No bootstrapping	–	–	–

36 children participated (range 4;0-5;2, mean 4;7), recruited from families in the University of Maryland area. Each child was given a small gift for participating. Four additional children were tested and subsequently excluded—2 due to experimenter error, 1 due to presenting with a strong “yes” bias (i.e., the participant indicated the puppet “liked” 11/12 of the test cards), and 1 due to a strong “no”-bias (i.e., they said the puppet “didn’t like” 12/12 of the test cards). We measured the percentage of cards sorted consistent with a quantity-based interpretation.

## Results

Across our three conditions, responses were significantly different from chance (sign tests: ADJ  $p < 0.0001$ , CON  $p < 0.05$ , DET  $p < 0.0001$ ). These differences were in different directions, however. Children sorted cards consistent with a quantity-based interpretation in DET 72% of the time, compared to 29% in ADJ and 40% in CON. In addition, DET was significantly different from both ADJ (t-test,  $p < 0.0001$ ) and CON (t-test,  $p < 0.0001$ ). These results are presented graphically in Figure 4.

It is noteworthy that these results are not simply an averaging effect: 8 out of 12 of the children in DET sorted at least 9 out of 12 test cards consistent with a quantity-based interpretation, while only 2 out of 12 children did so in ADJ and 3 out of 12 in AMB.

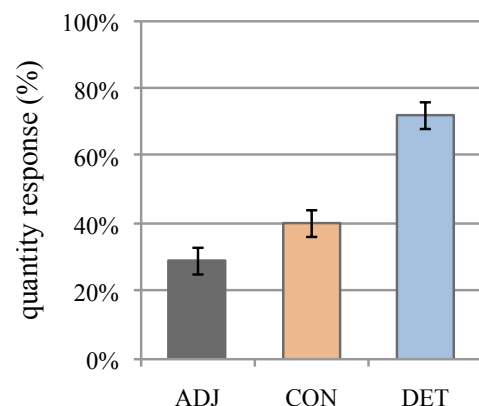


Figure 4: Percent quantity responses by condition.

As there were no differences between our conditions except for the syntactic context in which *gleebest* occurred, these results support the idea that syntax cues children into quantity-based meanings, with syntactic category playing a strong role. Partitivity, on the other hand, is a fairly weak cue: while there was a slight effect (CON had slightly higher quantity-based responses than ADJ,  $p < 0.05$ ), in neither of these conditions did children sort cards consistent with a quantity-based interpretation, in fact both conditions displayed *lower* than chance sorting of cards consistent with that interpretation. A table summarizing the predicted versus actual results is given in Table 3.

Table 3: **Predicted neutral (+) versus increased quantity response (+): prediction met (✓) versus not (×).**

Hypothesis	ADJ	CON	DET
Category as strong cue	– ✓	– ✓	+ ✓
Partitive as strong cue	– ✓	+ ×	+ ✓
No bootstrapping	– ✓	– ✓	– ×

Of the three hypotheses sketched, only syntactic category as a strong cue captures the results we found.

## Discussion

Our results show that a syntactic bootstrapping hypothesis for acquiring novel superlatives is supported. An additional hypothesis, that the presence of the partitive frame is a strong cue to quantity-meanings, was not supported. These results are important for a number of reasons. As observed in the introduction, choosing number as the relevant property from a set of available properties is potentially challenging for children. Our results highlight the role of the child's syntactic representations in narrowing her hypotheses about what matters when she tries to determine the meaning of a novel word, in particular the role of the syntactic category D as a strong cue to quantity-based meanings.

A different but related question that this work raises is the strength of the bias towards quality-based meanings in ADJ and CON. Given that children had no problem deciding that *gleebest* referred to numerosity in DET, we cannot assume some inability to reason about number. One might speculate that the bias is due to the child's distribution of known adjective (or superlative) meanings: since many more words in this category refer to object properties than set properties, the prior distribution of meanings biases her towards the former, absent syntactic cues to the contrary. Future work with younger children could examine the degree to which this bias emerges as a function of the size of their lexicons. The line of thought just outlined predicts that the youngest children would show less of a bias in this direction.

## Acknowledgments

This work was made possible by generous support from a Social Sciences and Humanities Research Council of Canada

doctoral award (#752-2010-0499) to Alexis Wellwood. The authors would especially like to thank Tim Hunter, Research Assistants Leah Whitehill and Jessica Lee, the University of Maryland's infant and preschool labs, the Center for Young Children, and the audience at the Linguistic Society of America's 2012 annual meeting.

## References

- Barner, D., Chow, K., & Yang, S.-J. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive Psychology*, 58, 195-219.
- Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4, 159-219.
- Benthem, J. van. (1989). Logical constants across types. *Notre Dame Journal of Formal Logic*, 3.
- Bloom, P., & Wynn, K. (1997). Linguistic cues in the acquisition of number words. *Journal of Child Language*, 24, 511-533.
- Bobaljik, J. D. (2007). *On comparative suppletion*. (University of Connecticut, m.s.)
- Bresnan, J. (1973). Syntax of the comparative clause construction in English. *Linguistic Inquiry*, 4(3), 275-343.
- Carey, S. (2009). Where our number concepts come from. *Journal of Philosophy*, 106(4), 220-254.
- Condry, K. F., & Spelke, E. S. (2008). The development of language and abstract concepts: The case of natural number. *Journal of Experimental Psychology*, 137, 22-38.
- Frege, G. (1893[1967]). *Grundgesetze der arithmetik, begriffsschriftlich abgeleitet (the basic laws of arithmetic)* (English translation in M. Furth (trans.) ed.). University of California: Berkeley.
- Gajewski, J. (2002). L-analycity in natural language. *Unpublished manuscript: MIT*.
- Hackl, M. (2009). On the grammar and processing of proportional quantifiers: *most* versus *more than half*. *Natural Language Semantics*, 17, 63-98.
- Jackendoff, R. (1977). *X' syntax*. Cambridge, Massachusetts: MIT Press.
- Landau, B., & Gleitman, L. (1985). *Language and experience: Evidence from the blind child*. Cambridge, Massachusetts: Harvard University Press.
- Syrett, K., Musolino, J., & Gelman, R. (2012). How can syntax support number word acquisition. *Language Learning and Development*, 8(146-176).
- Waxman, S., & Gelman, S. (1986). Preschoolers' use of superordinate relations in classification and language. *Cognitive Development*, 1, 139-156.
- Waxman, S., & Lidz, J. (2006). Early word learning. In D. Kuhn & R. Siegler (Eds.), (6th edition ed., Vol. 2, p. 299-335). Hoboken NJ: Wiley.
- Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology*, 24, 220-251.



# Expertise and the Wisdom of Crowds: Whose Judgments to Trust and When

Matthew B. Welsh (matthew.welsh@adelaide.edu.au)

University of Adelaide, North Tce  
Adelaide, SA 5005, Australia

## Abstract

The Wisdom of Crowds describes the fact that aggregating a group's estimate regarding unknown values is often a better strategy than selecting even an expert's opinion. The efficacy of this strategy, however, depends on biases being non-systematic and everyone being able to make a meaningful assessment. In situations where these conditions do not hold, expertise seems more likely to produce the best outcome. Amateurs and professional judgments are examined in a subjective domain – reviews of shows from an Arts festival – asking which group provides better information to the potential theatre-goer. In conclusion, while following the crowd produces good results, where a smaller number of reviews are available, taking expertise into account improves their usefulness and discrimination between shows.

**Keywords:** Expertise, Wisdom of Crowds, subjective judgment.

## Introduction

When making decisions between diverse options, we often do not have sufficient time or resources to conduct the sorts of thorough analyses recommended by decision analysts (see, e.g., Newendorp & Schuyler, 2000). Instead, we rely on simple rules to greatly reduce the complexity of our decision making while maintaining as much quality as possible (Gigerenzer & Todd, 1999). Perhaps the simplest such rule is: if someone recommends option A, then I will select option A.

This approach, of course, requires that you have some idea of whether or not you should trust the opinion of the person offering it, which is easy when it is a person you know but more difficult when you are forced to rely on the opinions of strangers – as is often the case.

As an example, consider a person's decisions regarding what to spend his/her entertainment budget on. While they could wait and hope that their friends will go to see all of the various shows that they were interested in, more often, they will have to rely on reviews from either professional reviewers or sites such as "Rotten Tomatoes" that aggregate amateur review data. In either case, the criteria on which the reviewers have provided their rating is generally unknown to the people using the information.

The question, then, is how to make the best use of the available information – from both professional and amateur reviewers – in order to make informed decisions about the quality of entertainment on offer.

## The Wisdom of Crowds

The wisdom of crowds describes a well-known effect first discussed by Galton (1907) and more recently repopularized

by Surowiecki (2004). The observation is simply that, when making decisions under uncertainty, the median or mean estimate of a crowd is often a better predictor than the estimate of a randomly chosen individual – even an expert.

This initially surprising observation results simply from the underlying mathematics of the problem. If any biases or errors in people's estimates are independent, then they will tend to be in random directions and thus, when averaged, will be removed. This has allowed researchers to demonstrate that even having the same individual make an estimate twice and averaging those values can produce better estimates – so long as some degree of independence can be established between the two estimates (Herzog & Hertwig, 2009; Vul & Pashler, 2008).

For the wisdom of crowds to work, therefore, one needs to be considering a domain in which biases in people's judgments are not systematically related to those of other people. If this condition is met, then one expects that averaging the judgments of a group regarding the quality of a particular show would provide a better estimate of how much you will enjoy it than relying on the advice of any single reviewer.

## Expertise

By comparison with the wisdom of crowds, expertise is a harder creature to pin down. While we all have an implicit understanding of what expertise is, actually defining it proves surprisingly difficult (see, e.g., Shanteau, 2002; Weiss, 2003) and people commonly confuse it with simple length of experience (Malhotra, Lee, & Khurana, 2005).

Despite this, given that we know there is such a thing as expertise and that people are employed on the basis of this to provide expert advice, it would seem reasonable for us to expect that this advice will be valuable – more valuable, at least, than a non-expert's judgment.

## Decision Criteria

An important question, which should be asked before continuing, relates to the decision criteria being used. This is important as, when we ask a question, we can only receive meaningful responses if the person understands and answers the question we have asked. In the case of reviews of entertainment, then, what is the question that is being asked?

The difficulty here is that expert and non-expert reviewers may be answering different questions. Experts might be answering the question – how much artistic merit does the show have? Non-experts, by comparison, may be answering the simpler question – how much did you enjoy the show. In both cases, the judgment is subjective and dependent on the

reviewers personal tastes but, in the first, it is also being judged against taught norms of quality.

A secondary concern is the fact that most reviews are undertaken on an absolute scale, whereas people are far more comfortable and more accurate making relative judgments (see, e.g., Stewart, Brown, & Chater, 2005; Stroop, 1932). Given this, we need to be cautious in interpreting what a reviewer may mean by any given review.

## This Study

In this study, reviews of entertainment will be analyzed in order to determine how a person could best use the available information to select a show to attend. It thus overlaps significant with problems such as the Netflix Prize (Bennett & Lanning, 2007) but is approached from a psychological rather than machine learning stance – that is, incorporating concepts such as expertise and considerations of *why* we have the data we do and how this should affect its use (for further discussions of this, second, point, see, Welsh & Navarro, 2011; Welsh, Navarro, & Begg, 2011).

## Method

The data sets selected for analysis consisted of reviews of acts performing at the 2011 Adelaide Fringe Festival – a large, “unjuried” Arts Festival held annually in Adelaide, Australia. Being an unjuried festival, any act is free to register to perform without being selected by the festival’s governing body. As such, the quality of performances is (presumably) more variable than would be observed in a juried festival where acts must convince the festival’s jury of their quality before registering.

Given this, selecting a quality show to attend from the hundreds (750 in 2011) on offer becomes a difficult task in the absence of reliable indicators of quality. To this end, two databases of reviews were acquired: first, the Adelaide Fringe’s summary of published, professional reviews from newspapers and news websites – labeled simply “Fringe” hereafter; and, second, the database from BankSA’s “Talkfringe” website which allows anyone to register and post reviews of any Fringe shows that they have seen.

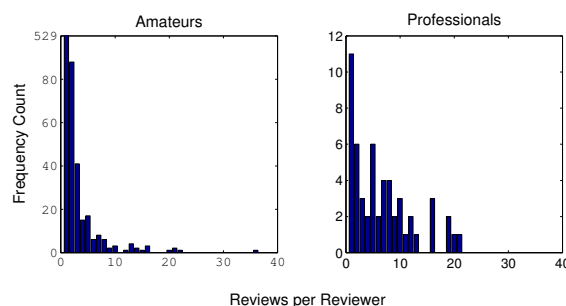
All of the Talkfringe reviews use the same 1 to 5 ‘Star’ rating system (with half stars). The professional reviews, however, were in a variety of formats. To maintain comparability, therefore, only professional reviews that used a 5-star rating system were included in the analyses.

## Data Characterization

The Fringe database records 365 reviews in the required 5-star format, made by 54 reviewers – an average of 6.8 reviews per reviewer. By contrast, the Talkfringe database contains 1436 reviews made by 731 reviewers. Figure 1 displays this information as a histogram of reviews per reviewer for the Amateurs (Talkfringe) and Professionals (Fringe) separately. Between the two databases, reviews were obtained for a total of 420 shows, with each being reviewed an average of 4.3 times.

Looking at Figure 1, one sees that both subplots seem to display similarly shaped distributions – a decay function of some type. The figure is, however, somewhat misleading as the y-axis of the Amateur subplot is displayed as if the highest count was 100 when, in fact, it was 529 (as indicated by the high value on the y-axis).

Figure 1. Histogram of number of reviews per reviewer by reviewer group. Note: Amateur y-axis is non-linear at top.



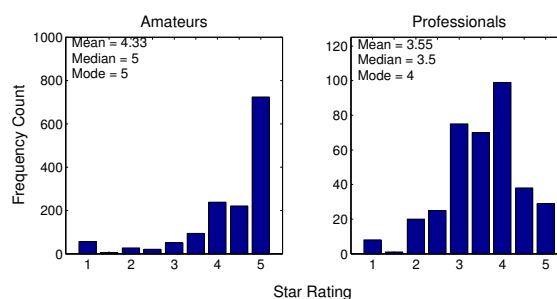
That is, while only a modest proportion (12/54) of the professionals reviewed only a single show, the majority of amateurs (529/735) did so.

## Results

### Indirect Comparisons

As an initial approach to the question of whose reviews should be trusted, the distributions of star-ratings within each database were compared. Figure 2 shows the histograms of this data.

Figure 2. Histogram of Star ratings by reviewer group.



Looking at Figure 2, one sees that the two distributions differ significantly from one another, as confirmed by an independent samples t-test,  $t(1799) = 13.9$ ,  $p < .001$ , Cohen’s  $d = 0.81$ . The Amateurs display something close to an exponential distribution of star-ratings, with a median and mode at 5 and a mean of 4.33, while the professionals display something closer to a Gaussian, with a mean and median around 3.5 and a mode at 4. This raises questions about the discriminability of Amateur reviews – that is, whether seeing a 5 star review from an amateur allows you to conclude anything meaningful about that show.

There are, however, alternate possible explanations for this pattern of responses. The first is that amateurs tend to be less discriminating in their tastes than the professional

and, thus, enjoy shows more. The second, however, is a selection effect – while professionals are told which shows to attend and write reviews of all of the shows that they attend, amateurs choose shows that they think they will like and are less likely to write a review unless motivated by particularly enjoying or disliking the show. Given that more popular shows attract greater audiences, and assuming a positive relationship between quality and popularity, this will tend to result in large numbers of high-star reviews for popular shows and relatively few reviews of any sort for less popular shows.

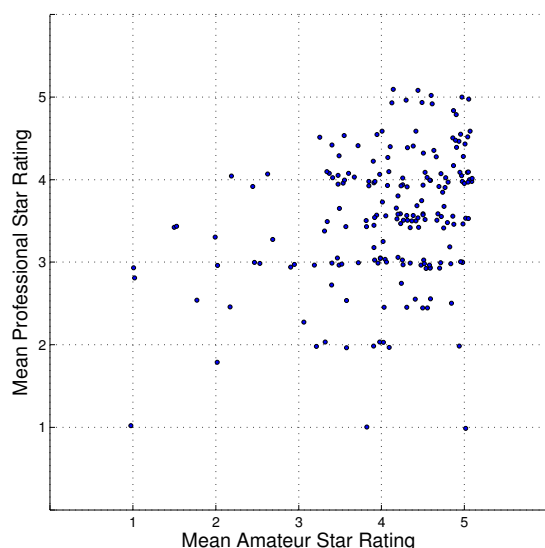
Based on this reasoning, one could assume that any show that has multiple, high-star reviews from amateur reviewers is likely to have been a popular show.

## Direct Comparisons

The above discussion considers only the distributions of star ratings, rather than those instances where we have reviews of the same show made by both amateur and professional reviewers. An examination of the two databases revealed that, of the 420 shows, 191 of these were ‘shared’; that is, had been reviewed by at least one member of each reviewer group.

Looking only at these ‘shared’ shows, the difference between the professional and amateur groups (3.59 versus 4.33) is almost exactly the same as for the full dataset (3.55 versus 4.33) and remains significant by a paired samples t-test,  $t(1231) = 11.2$ ,  $p < .001$ , Cohen’s  $d = 0.79$ .

Figure 3. Scatterplot of mean amateur versus mean professional review for all 191 ‘shared’ shows. NB – some jitter has been added to the points to reduce overlap and facilitate display.



Despite the removal of over 200 shows that lacked a rating from each group, a consideration of only the overlapping shows still contains the majority of the review data as these 191 shows attracted 1233 of the total 1801

reviews and a comparison of the distribution of star ratings within this group with that for the complete datasets shown in Figure 2 revealed no noticeable differences. Figure 3 plots the mean reviews provided by each group for each show against that calculated from the other group.

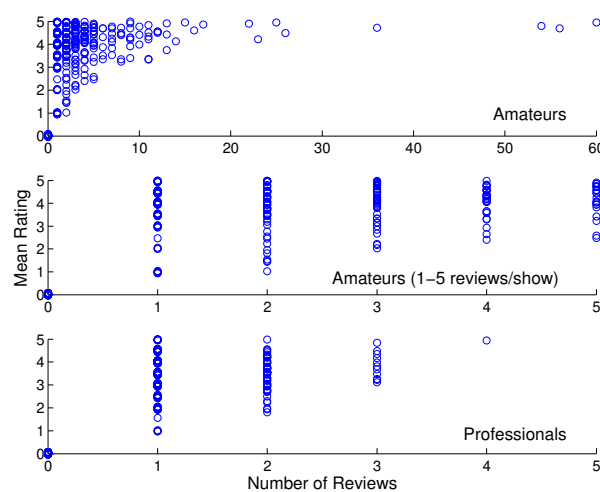
Looking at Figure 3, one can see that the relationship between the amateur and professional reviews is positive, but not particularly strong – confirmed by a correlation  $r(190) = 0.32$ ,  $p < .001$ , indicating significant disagreement between the two groups on the quality of shows.

A closer examination of the figure reveals that a partial explanation for the poor correlation may be restricted range – with relatively few datapoints in the lower left quadrant. Again, this is likely to reflect selection biases, with all type of reviewers more likely to attend and review popular shows – which, in turn, are likely to be of higher quality.

## Quality by Popularity

Given the data above, what can we say about how a person should go about selecting a show to see? As noted above, there is an assumption that higher quality shows are more likely to become more popular and that the number of reviews can be used as a proxy for popularity. This means that we can compare the star-ratings for shows of differing popularity to see how these variables interact. Figure 4, below, plots show star-ratings against number of reviews for all 420 shows contained in both databases.

Figure 4. Scatterplots of number of reviews (show popularity) versus mean rating (show quality) for Amateur and Professional reviewers. NB – some jitter has been added on the y-axis to facilitate display.



Looking at Figure 4, one sees that the mean ratings of shows that received low numbers of reviews vary quite significantly – indeed for shows with only one or two reviews, the mean ratings are fairly uniformly distributed across the 1-to-5 range.

For shows with higher numbers of reviews, however, one sees a striking pattern emerge – as the number of reviews increases, so does the *minimum* mean rating that that show

received. Comparing the bottom two subplots, one sees that this pattern emerges early in both the amateur and professional reviews; no show with 3 or more reviews averages less than a 2-star rating.

Looking across the top subplot of Figure 4, one can see this predictive power continues for higher numbers of reviews: no show with 6 or more reviews was rated lower than 3 star (on average); no show with 14 or more reviews was rated lower than 4 star (on average); and the 7 shows that were reviewed by 25 or more people all averaged at least 4.5 star reviews.

This would seem to confirm the prediction that popularity and quality are, in fact, linked and suggest that an appropriate strategy for selecting a quality show would be to select one that many people have reviewed – even without reading those reviews.

### Expert vs Non-Expert Reviews

A final question to be addressed is that of expertise. While we have, above, divided reviewers according to whether they are Professional or Amateurs – and assume that this reflects some difference in expertise (in reviewing shows) – the data afford us some scope to test this assumption.

Looking once more at Figure 4, for example, one can see a suggestive pattern in the comparison between the Amateur and Professional results – where the speed at which the predictive multiple reviews increases seems greater for the Professional. That is, having had multiple Professional reviewers attend a show may be a better indicator of quality than having had the same number of Amateurs review it.

A more important question, however, is whether we can establish that expert reviews are *better* than non-expert reviews. The difficulty, of course, is in determining how we measure the quality of a review – after the fact and in the absence of any objective standard. A simple wisdom of crowds approach would suggest that we use the median or mean review from all reviewers as the standard but this runs into the problem of non-discriminability in the amateur data where too many shows will all be rated 5-star.

There are, however, at least two methods of using the current data to shed light on the relative usefulness of professional and amateur reviews in selecting a good show.

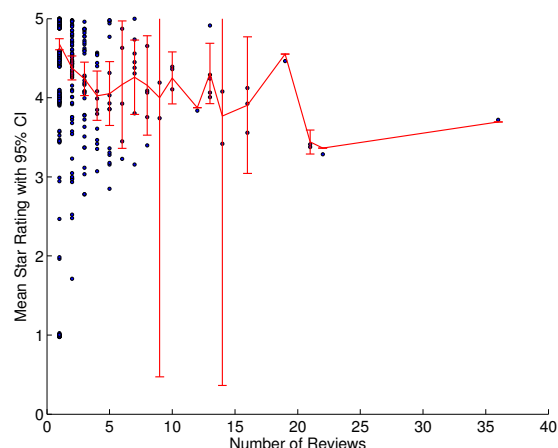
### Measuring the Expertise of Amateur Reviewers

The first of these involves a comparison of the differences within the two groups. For example, it seems a reasonable assumption that those Amateurs who review more shows become more expert in doing so. The same relationship, of course, is less likely to hold in the Professional reviewers as the assumption is that these people have significant previous experience that is not available to us through the data set; and which is likely to outweigh any effect of the relatively few reviews they made during this event. Given the above, it seems necessary to restrict this discussion to differences within the Amateur group.

What then are the differences between the more and less ‘expert’ amateurs – that is, between those who posted many

rather than few reviews. Figure 5 thus plots number of reviews per amateur reviewer against star ratings.

Figure 5. Scatterplot comparing number of reviews to mean star rating (amateurs only). ‘Jitter’ has been added to the data along the y-axis to prevent datapoints overlapping. The red line shows the overall mean for each group of reviewers.



Looking at Figure 5, one sees a trend as the number of reviews that a person has posted increases; specifically, as the number of reviews increases, the average review tends to decrease,  $r(729) = -0.20, p < .001$ .

This could be explained by a drop-off in the quality of shows – if everyone were seeing the same shows and there were only a small number of genuinely 5-star shows, for example. Given the number of shows involved, however, and how many of these received 5 star ratings from someone, this seems an unlikely explanation. Instead, it seems more likely that we have support for the idea that increased experience in reviewing (and, therefore, seeing more shows) changes the ratings that one is likely to give.

Suggestively, the most prolific reviewers in Figure 5 give average ratings that are more typical of Professional reviewers than the other Amateurs. That is, their mean ratings tend to be between 3 and 4 rather than 4 and 5.

The question remains, however, as to whether this reflects *better* reviews; and the problem is, of course, that as enjoyment of a show is highly subjective, it is possible that what is the *better* (i.e., more predictive) review differs between individuals.

On the basis of these results, for example, one might conclude that the more shows one is inclined to see, then the more similar one's own ratings will be to those of Professional reviewers. If so, then one should weight professional reviews more highly than amateur ones – or, where these are unavailable, downgrade ‘overly-enthusiastic’ amateur reviews.

### Consistency of Different Reviewers

A second consideration in what makes one review *better* than another is their reliability. That is, when two people have seen the same show, are they inclined to give the same rating? A comparison between the Amateurs and

Professionals on such a measure might allow one to have greater or lesser confidence in one group's ratings.

Within the Professional reviewers group, there were 70 shows that had been reviewed by at least 2 reviewers – which yielded a total of 97 pair-wise comparisons (due to some shows being rated by three or four reviewers). Thirty of these had exactly the same rating, with another 40 differing by only half a star. Overall, the average difference between ratings of the same show by professional reviewers was approximately half a star ( $M = 0.56$ ,  $SD = 0.52$ ).

The Amateur group, by comparison, had 228 shows with multiple reviewers, which resulted in 10,401 pair-wise comparisons. This number, however, is dominated by the relatively small number of very popular shows – those on which we see a ceiling effect resulting from the selection bias. The most popular show, for example, has 60 reviews, 58 of which are 5-star – with one 1-star and one 3-star review making up the numbers. This show contributes 1770 unique pair-wise comparisons – over a sixth of the total – and would thus, if included, overwhelm any effects of the inter-rater reliability more generally. To ensure comparability with the Professional results, therefore, only shows that had been reviewed by between 2 and 4 reviewers (the numbers observed in the professional sample) were included in the analyses. This resulted in the removal of 79 shows, leaving 149 and a total of 404 unique pair-wise comparisons.

Of these, 120 had exactly the same rating, 114 differed by half a star and 170 differed by 1 full star or more. The average difference between the amateur reviewers' ratings for these shows was 0.82 stars ( $SD = 0.87$ ), significantly higher than that observed in the Professional reviewers' ratings,  $t(499) = 2.83$   $p = .002$ .

## Discussion

The results paint a complex picture of the relationships between reviewer expertise and the use of aggregation strategies such as the wisdom of crowds for reviews from multiple sources.

Perhaps the single best predictor of show quality (i.e., how much people enjoyed the show) was the total number of reviews that the show had received – reinforcing the assumption that popularity and quality are linked. Note, however, that this is a distinct effect from the wisdom of crowds as the results suggest that we don't need to look at the ratings provided by reviewers at all. Instead, all we need to do is "follow the crowd" and they will lead us to good shows.

In cases without such overwhelming endorsement, however, we are forced to rely on the numerical ratings provided by the expert and amateur reviewers and can run into difficulties in determining what to do.

The first problem we observed in the data was the strong selection bias in the amateur data; because people tend only to pay to see shows that they expect to like, the distribution of star ratings gets shifted to the right – with more 5-star reviews. Added to this is the voluntary nature of amateur

reviews, which results in people only writing a review if they are motivated to do so – which, we suggest is most likely when they particularly like or dislike a show. This effect will, therefore, tend to push results even further towards the extremes and, given the effect described above, this will tend to push more people into the very high part of the rating range.

Thus we have a large number of reviews that are relatively uninformative – reflecting the fact that a person predisposed to like a particular show really liked it. A result of this is the lack of discrimination in the amateur data where, because so many reviews give 5-star ratings, it simply doesn't help us to make a decision regarding which of these shows we should attend and short-circuits attempts to use the wisdom of crowds based on median values – as we would end up comparing 5-stars with 5-stars.

A second (but related) concern is that the majority of amateur reviewers (529 of 731) wrote only a single review. Given what we know about people's inability to directly assess values, the use of relative preferences (e.g., converting the ratings to rankings) is a sound method for improving our understanding of what people's expressed preferences actually mean. With only one review per reviewer, however, we cannot meaningfully assess relative preferences.

By comparison, a professional reviewer, while exercising some choice over which shows to see will also have some dictated by their employers and will be asked to write a review of all of the shows that they see. They are, from our data, far more likely to see multiple shows, and have a less-skewed distribution of ratings. They were also, in the subset of shows with a relatively few reviewers, more often in agreement with one another than were the amateurs.

This means that, in relying on professional reviews, one is better able to discriminate between their preferences for those shows that they have seen and also can be more assured that their review is reliable – that is, that another professional reviewer would have a similar opinion.

An addendum to this is that the data support the idea that the difference between amateurs and professional is related to experience/expertise. Amateurs who reviewed larger numbers of shows gave ratings that were more like the professionals. This could suggest that people are, in fact, rating shows on a relative scale but that the single-review amateurs have fewer shows to compare with and thus the chance of the show being amongst the best they have seen is relatively greater. The professionals and high-rate amateurs, by comparison, have a great many shows to compare the current show to and thus the likelihood of it being judged exceptional (5-star) is relatively less.

## Caveats

In so subjective a domain, there are, of course, a number of caveats to consider in conjunction with the arguments made above. A primary one, of course, is that we have not made any attempt to look at the types of shows that different people have attended and rated. If we expect that different

people have different tastes in entertainment, then we could conduct a far more fine-toothed analysis of preferences.

This importance of this for the current findings, however, is that one might expect a difference in preferences between professional and amateur reviewers. For example, while purely speculative, it would seem entirely feasible that professional reviewers prefer more serious art whereas the amateurs prefer lighter, comedic events.

If this is the case, then one would have to take into account such between group differences when determining whose reviews should be taken into account when making a decision. That is, knowing that professionals reliably tend to rate a show highly may be of no help at all if it is a type of show that you do not enjoy.

A second caveat is that there has not, as yet, been any attempt to weight or rank the data, which would, as described earlier, be expected to improve the predictive power of ratings – from those reviewers who reviewed multiple shows at least. An appropriate application of such tools, however, requires a fundamental grasp on the nature of the data; a grasp that has been greatly strengthened by the exploratory approach taken here.

### Future Research

Given the findings and the caveats noted above, a number of directions for continuing the research suggest themselves. The first is to examine the data in finer detail, dividing shows according to type - to see whether specific reviewers can be identified as having preferences between these.

Data beyond the ratings could also be accessed – for example, using ticket sales to directly measure the popularity of a show rather than simply assuming that number of reviews is a reflection of popularity.

This additional information, used in conjunction with ranking and weighting algorithms, could then be used to generate predictive models for individuals based on the shows that they have seen and how much they enjoyed them and using one half of the data to predict the other – in a similar fashion to the Netflix recommendation algorithms developed as part of the Netflix Prize competition (Bennett & Lanning, 2007).

Finally, experimental work designed to directly measure selection biases in reviews could be conducted, building on the work herein. Similarly, such work could potentially distinguish between alternative judgment strategies – for example, if experts are attempting to provide ‘absolute’ quality judgments whereas amateurs are just indicated whether they like a show or not.

### Conclusions

Within a domain such as entertainment reviews, good decisions can be made by following the crowd – if not always using the wisdom of crowds, per se. Where choices need to be made between shows, however, amateur reviewers ratings tend to cluster too closely around the maximum rating – as a result of selection bias in both show choice and the decision to write a review.

In these cases, therefore, following the advice of more expert reviewers (i.e., professionals and experienced amateurs) seems more likely to provide discrimination as they display less selection bias in their shows seen, meaning that they tend to write reviews of a variety of shows and have clearly discriminable preferences between these.

### Acknowledgments

Thanks to ExxonMobil and Santos for their support of the CIBP group in the Australian School of Petroleum; to Michelle Read from the Adelaide Fringe Festival and Simon Evans at BankSA for their assistance in accessing the review databases; and to Dan Navarro, Anna Ma-Wyatt, Steve Begg and three reviewers for their comments.

### References

- Bennett, J., & Lanning, S. (2007). The Netflix Prize. Proceedings of KDD Cup and Workshop, San Jose, CA.
- Galton, F. (1907). Vox populi. *Nature*, 75, 450-451.
- Gigerenzer, G., & Todd, P. M. (Eds.). (1999). *Simple heuristics that make us smart*. Oxford, UK: Oxford University Press.
- Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: improving individual judgments with dialectical bootstrapping. *Psychological Science*, 20(2), 231-237.
- Malhotra, V., Lee, M. D., & Khurana, A. K. (2005). Domain experts influence decision quality: towards a robust method for their identification. *Journal of Petroleum Science and Engineering, Special Issue*.
- Newendorp, P. D., & Schuyler, J. (2000). *Decision Analysis for Petroleum Exploration*. Aurora, CO: Planning Press.
- Shanteau, J. (2002). Performance-based assessment of expertise: how to decide if someone is an expert or not. *European J. of Operational Research*, 136(2), 253-263.
- Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, 112(4), 881-911.
- Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology*, 15(5), 550-562.
- Surowiecki, J. (2004). *The Wisdom of Crowds*. New York, NY: Random House.
- Vul, E., & Pashler, H. (2008). Measuring the crowd within: probabilistic representations within individuals. *Psychological Science*, 19(7), 645-647.
- Weiss, D. J. (2003). Empirical assessment of expertise. *Human Factors*, 45(1), 104-116.
- Welsh, M. B., & Navarro, D. J. (in press). Seeing is believing: priors, trust and base rate neglect. *Organizational Behavior and Human Decision Processes*. Accepted April 6<sup>th</sup> 2012.
- Welsh, M. B., Navarro, D. J., & Begg, S. H. (2011). Number preference, precision and implicit confidence. In L. Carlson, C. Hölscher & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1521-1526) Austin, TX: CSS.



# Complex First? On the Priority of Nouns in Language Acquisition and Evolution

Markus Werning (markus.werning@rub.de)

Department of Philosophy II, Ruhr University Bochum  
44780 Bochum, Germany

## Abstract

The paper points to an apparent paradox in the science of language. It regards the semantics of nouns and consists of a set of together incompatible, but individually well confirmed propositions about the evolution and development of language, the semantics of word classes and the cortical realization of word meaning. Theoretical and empirical considerations support the view that the concepts expressed by concrete nouns are more complex and their neural realizations more widely distributed in cortex than those expressed by other word classes. For a cortically implemented syntax-semantics interface, the more widely distributed a concept's neural realization is, the more effort it takes to establish a link between the concept and its expression. If one assumes the principle that in ontogeny and phylogeny capabilities demanding more effort develop, respectively, evolve later than those demanding less effort, the empirical observation seems paradoxical that the meanings of concrete nouns, in ontogeny and phylogeny, are acquired earlier than those of other word classes.

**Keywords:** evolution of language, language acquisition, compositionality, modularity, complexity, frames, situated conceptualization

## Introduction

When one conjoins relatively well supported views on language acquisition and typology with frequently held views on the neural realization of meaning and some general principles of evolution and development, one seems to arrive at what I shall call the Complex-First Paradox. At its core is the question why concepts of substances, typically expressed by concrete nouns, seem to lexicalize ontogenetically and phylogenetically so early, even though they are apparently semantically far more complex than concepts that lexicalize later. The paradox consists of five propositions each of which seems plausible in its own right and is supported by empirical or theoretical reasons. The set of propositions – as is the nature of paradoxes – is apparently inconsistent, though, and thus points to an explanatory deficit in linguistic theory (Werning, 2008, 2010):

- (P1) The meanings of concrete nouns, in ontogeny and (probably) phylogeny, are acquired earlier than those of many – eventually even all – other word classes.
- (P2) The meanings of concrete nouns are substance concepts.
- (P3) Substance concepts are semantically more complex and their neural realizations more widely distributed in cortex than those expressed by the other word classes in question.

- (P4) For a cortically implemented syntax-semantics interface, the more widely distributed a concept's neural realization is, the more effort it takes to establish a link between the concept and some lexical expression thereof.
- (P5) In ontogeny and phylogeny, capabilities demanding more effort, all other things being equal, develop and, respectively, evolve later than those demanding less effort.

The paradox should be obvious now: Assume that the meanings of concrete nouns like *daddy*, *milk*, and *cat* are indeed semantically more complex or, to use another word, thicker than the meanings of other word classes, e.g., adjectives like *blue*, *big*, and *bold*. If one accepts that meanings are mental concepts, the view is illustrated as follows: The substance concept [milk] has not only perceptual components of various modalities like [white], [fluid], and [sweet], but also components that relate to affordances like [to drink]. The attributive concept [blue], in contrast, seems to be relatively thin: it does not decompose into distinct conceptual parts and seems to pertain to the visual domain only. Assume, furthermore, that conceptual complexity correlates with a wider distribution of the conceptual parts, respectively, their neural realizations in the cortex. One then expects the neural correlate of [milk] to pertain to visual, tactile, gustatory, and action-related regions. In contrast, the correlate of [blue] seems to be bound to the visual cortex. Following another of the assumptions, a word-to-meaning assignment ought to be more easily tractable for a cortically realized syntax-semantics interface if the neural correlate of the meaning is relatively local, rather than widely distributed. Consequently, the link between the adjective *blue* and the attributive concept [blue] should require less effort than the link between *milk* and [milk].

Take it as a quite general principle of evolution now that with regard to one and the same domain incrementally more complex capabilities *ceteris paribus* evolve later than simpler ones. There had to be feathers first, only then some reptile species could evolve wings. Vision could succeed in evolution only after light-detection had evolved. It seems to be a simple truth that lies behind it: *Natura non facit saltus*. There is an outright analogy in development: A child must have acquired the capability to hold a stick before it will be able to use a hammer. Children have to acquire simple closed syllables (CVC, e.g., *come*) before they are able to pronounce syllables with complex codas (CVCC, e.g., *cast*).

Given those assumptions, how can it be that the meaning of the noun *milk* ontogenetically and phylogenetically still is



acquired earlier than that of the adjective *blue*? Since the concept [milk] is semantically more complex than [blue], its neural correlate should be more widely distributed, the link between the concept and its expression should imply more effort, and thus ought to be established later in ontogeny and phylogeny. Rather than the empirical claim made by the first proposition, we should on the basis of the other four assumptions expect that the meanings of concrete nouns, in ontogeny and phylogeny, be acquired later than those of other word classes. In the paper I would like to press the paradox a little further by putting forward arguments for each of the five propositions and rejecting objections against them. Even though my résumé will be rather pessimistic, I will conclude with some more speculative remarks on a potential solution. The paper is primarily intended as an exposition of the paradox, rather than as its solution. Due to limited space and time I defer to Werning (2008, 2010) for a more elaborate discussion of each of the five propositions.

## Words and Concepts

The primary role for concepts is the integration of perception and action control. In order to survive in a world with a multitude of things, subjects must subsume them under concepts. Categorization allows the subject to recognize objects and events in the world as well as states of the body, to generate generalizations, and to preserve this information over time. Only thus goal directed interaction between one's body and the world is possible to the degree we observe it in many species.

With regard to humans, concepts are assigned a twofold explanatory role: (i) as content providers and (ii) as meaning providers. In their first role concepts provide contents to intentional states. In their second role concepts are identified with the meanings of linguistic expressions. Concepts are apt to fulfill the two roles because they are individuated as internal states of the system that essentially bear a causal-informational relation of co-variation to external contents (Fodor, 1992). This way, concepts may explain why intentional states are about things and why the meanings of expressions in a given context determine which things are referred to.

Intentional states include such diverse modes as perception, belief, desire, memory, expectation, imagination, emotion, and the will. Concepts provide the satisfaction conditions of intentional states, enter into inferential relations, and play a role in the causation of action. The twofold role of concepts suggests a view that intimately links meaning to intentionality. A unified approach of meaning and intentional content holds that the meaning of the sentence *There is milk in the bottle*, the perception of milk being white, the belief that milk is nutritious, and the desire to drink a glass of milk have one thing in common: they involve the concept [milk]. This at least follows if one assumes (i) the compositionality of linguistic meaning, and (ii) the compositionality of intentional content. The compositionality of meaning is the

principle that the meaning of a complex expression is in a syntax-dependent way determined by the meanings of its parts. It explains how the concept [milk] contributes to the meaning of the sentence *There is milk in the bottle*. The compositionality of content says that the content of a complex intentional state is in a structure-dependent way determined by the contents of its parts. It explains how the content of the concept [milk] determines the contents of the perceptual, doxastic, and volitional states just mentioned (for a discussion of the reasons for compositionality see Werning, 2005; Werning et al. eds., 2012).

In our context, the most important distinction in the domain of concepts is that between attributive concepts and substance concepts. Attributive concepts represent features of objects that are volatile in the sense that one and the same object can fall under different attributive concepts at different times: An object may, e.g., change its color, size, or speed, but still continues to exist. [blue] thus is a paradigmatic attributive concept.

Substance concepts, in contrast, are governed by the identity conditions of objects: A mug ceases to exist when it no longer falls under the substance concept [mug], say, because it has been shattered. Substance concepts serve to re-identify things over time in spite of their contingent changes of attributes and so allow us to gather, store and update information in a systematic and enduring way (Millikan, 1998). They are typically expressed by concrete nouns. Attributive concepts, in contrast, are typically expressed by adjectives or abstract nouns.

## Nouns and Adjectives

The paradox arises from the fact that substance concepts are ontogenetically and probably phylogenetically earlier lexicalized than attributive concepts. The great mass of children's earliest words are concrete nouns. During the so-called naming explosion, when children around 18 months of age first systematically organize their concepts by means of a lexicon, they preponderantly pair substance concepts with concrete nouns, whereas the assignment of adjectives and abstract nouns to the attributive concepts they express comes much later (Ingram, 1989). Some languages even don't have adjectives or just a closed set of them (Dixon, 1999), while the class of concrete nouns is arguably universal and always open. One may thus also argue that nouns in phylogeny are prior to adjectives. With respect to the typology of the earliest words, Barrett (1995) in a handbook article provides the following overview:

- 0th-100th word: high proportion of common nouns.
- 200th-...: proportion of common nouns decreases.
- 50th-100th word: proportion of verbs begins to increase.
- 400th-500th word: verb proportion continues to increase and finally begins to level out.
- 50th-100th word: proportion of adjectives begins to increase.

100th-500th word: proportion of adjectives continues to increase.

Even authors like Bloom (2000) who are more critical of the notion of a naming explosion concede that in the earliest phase of language development there is an “object bias”: A new word by default is interpreted as a name of an object (i.e., as a concrete noun). It needs some counterevidence for the child to realize that a word (an adjective or verb) expresses a property or an action, instead.

If the data are interpreted correctly, we can make the following inference: Since concrete nouns express substance concepts and prototypical adjectives express attributive concepts, and since concrete nouns are earlier acquired by the child than adjectives, it logically follows that substance concepts are ontogenetically earlier lexicalized than attributive concepts.

With respect to the claim on phylogeny, the evidence is more indirect and less compelling – hence the qualification “probably”. It is an undeniable fact that in all languages, in which the types of nouns and adjectives exist, there are more concrete nouns than adjectives (Dixon, 1999). Even in English (Givon, 1970) most adjectives are derived from either nouns or verbs, while there are only very few original adjectives. One can still defend the claim that the noun type is universal (Mithun, 2000). Even in languages like Iroquoian, which is sometimes said to have no nouns, there are at least very noun-like words. The adjective type, in contrast, clearly is not universal. If adjectives were phylogenetically earlier than concrete nouns, we should expect the situation with regard to universality be the other way round. In light of the available evidence, proposition P1 is hence relatively well supported, at least if one identifies the contrasting word class with the class of adjectives.

## The Structure of Meaning

One of the main controversies regarding the processing and neuro-cognitive implementation of meaning is whether the semantics of language is processed in a modular or non-modular way. According to modular approaches, the meanings of words and sentences are processed in an informationally largely encapsulated, autonomous, and amodal way (Clifton & Ferreira, 1987). Candidates for cortical correlates of semantic processes are often supposed to be localized in left temporal and partially frontal regions (Friederici, 2002). Regions typically associated with either perceptual or motor processes in this paradigm are typically not regarded as contributing to semantics.

Modular approaches towards perception, in turn, argue for informationally encapsulated, domain-specific and cognitively impenetrable modules for various perceptual tasks (Barrett & Kurzban, 2006, for review). Modularism with respect to semantics, perception, and perhaps other types of intentional states would thus be hardly compatible with the view that the same mental concept, respectively its neural correlate, is both a meaning provider for linguistic expressions and a content provider for various types of

intentional states. A manifold of concept tokens with the content of milk would thus be required: the concepts [milk]-in-meaning, [milk]-in-perception, [milk]-in-desire, etc. – eventually even [milk]-in-desires-to-drink, [milk]-in-desires-to-cook, etc. It is easy to imagine that such a view would quickly lead to an ontological explosion of concepts, at least, if concepts are supposed to exist in a realist manner.

Much more compatible with a realist attitude towards concepts and the methodological goal of ontological parsimony is the anti-modularist view of situated conceptualization (Barsalou, 2005). Here concepts are regarded as situated, i.e., largely based on sensori-motor schemata. The controversy between semantic modularism and semantic anti-modularism relates to the question whether some lexical concepts – i.e., concepts listed in the lexicon and thus expressed by single words – decompose into conceptual parts. Some authors believe that lexical concepts are altogether not decomposable (Fodor & Lepore, 1992). According to those so-called atomist positions, only concepts that are linguistically expressible by syntactically explicitly combined expressions can be complex. In neuroscience some researchers hold that substantial features like that of being an elephant or even features as specific as that of being Halle Berry are represented by highly specialized single neurons (Quiñero et al., 2005). Lexical atomism is a view semantic modularists can easily live with. For, if meanings are unstructured, it is completely unproblematic to conceive them as localizable elements in an encapsulated module. Proponents of a situated view of meaning, in contrast, will assume that at least some lexical meanings are structured so that parts of the meaning providing concepts may involve various sensori-motor schemata. Semantic anti-modularism seems to exclude lexical atomism.

Our propositions P3 and P4 seem hardly tenable for someone who shares the views of lexical atomism or semantic modularism. Proposition P3 saying that substance concepts expressed by nouns are semantically more complex than concepts expressed by other word classes immediately contradicts lexical atomism, according to which all lexical concepts have the same complexity, viz. zero. Proposition P4 seems to be empty if lexical atomism is true and largely unmotivated if semantic modularism holds. The proposition says that for a cortically implemented syntax-semantics interface, the more widely distributed a concept’s neural realization is, the more effort it takes to establish a link between the concept and its lexical expression. Now, if lexical atomism is true, there simply should not be any concepts with a widely distributed neural realization. For, how could this be the case if all lexical concepts are unstructured? If semantic modularism were to hold, the meanings even of words that are semantically complex – modularism does not entail atomism – would still be locally realized in the postulated semantics module. There would thus be no reason to assume that significantly more effort is needed to assign a word to its meaning, even if the expressed concept is complex.

Since the doctrines of lexical atomism and semantic modularism conflict with P3 and P4, the natural way to defend the two propositions is to argue against lexical atomism and semantic modularism. This is what I will do in the next section. I will outline a view of situated conceptualization which refutes atomism and modularism.

## Situated Conceptualization

In psychology, philosophy and linguistics various theories have been proposed to account for the decomposition of concepts. For the present purpose the choice of frame theory as a starting point seems most fruitful (Barsalou, 1992). Frame theory provides us with a universal account not only for categorization and its link to action-control, but also for the decomposition of concepts. Frames are recursive attribute-value structures. Attributes assign unique values to objects and thus describe functional relations. The values can be structured frames themselves. A frame is defined for a large domain of things and contains a fixed set of attributes (e.g., color, form) each of which allows for a number of different values (red, green, ... ; round, square, ...). The attributes in question are not constrained to perceptual modalities, but may as well involve attributes of motor affordances. Frames can be nested hierarchically and mutual constraints between attributes (e.g., between states of an object and actions directed to it) and between larger frames can be incorporated (see Figure 1).

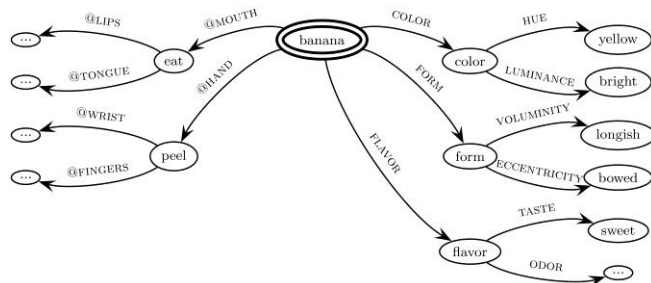


Figure 1. Hypothetical fragment of the frame for the concept [banana]. The substance concept to be decomposed is marked by a double-circle as the referring node of the frame. The labeled arrows denote attributes, the nodes their values. Nodes are themselves regarded as concepts and thus as conceptual parts of the central concept. Whereas, in English, feature attributes (shown on the right) are frequently lexicalized – their arguments typically enter possessive constructions like *The color of the banana is yellow* or *The banana has the color yellow* – affordance attributes (on the left) are rarely overtly expressed. Based on linguistic and neurobiological evidence, we assume that affordances often relate to body parts and hence use the convention “@ + body part”. Formally, attributes are mappings from domains of some type into domains of some other type. Petersen & Werning (2007) provide an explicit account of frames using a calculus of typed feature hierarchies and incorporating typicality effects.

For many attributes involved in perceptual processing one can anatomically identify cortical correlates. Those areas often exhibit a twofold topological structure and justify the notion of a feature map: (i) a receptor topology (e.g., retinotopy in vision, somatotopy in touch): neighboring regions of neurons code for neighboring regions of the receptive field; and (ii) a feature topology: neighboring regions of neurons code for similar features (See Figure 2). With regard to the monkey, more than 30 cortical areas forming feature maps are experimentally known for vision alone (Felleman & van Essen, 1991). Also affordance attributes seem to have cortical correlates, predominantly in the premotor cortex. The discovery of the so-called mirror neuron system (Rizzolatti, & Craighero, 2004, for review) may provide a basis to integrate affordances into frames.



Figure 2. Cortical realizations of frame attributes. Fragment (ca. 4mm<sup>2</sup>) of the neural feature map for the attribute orientation of cat V1 (adapted from Crair et. al., 1997). The arrows indicate the polar topology of the orientation values represented within each hypercolumn. Hypercolumns are arranged in a retinotopic topology.

The fact that values of different attributes may be instantiated by the same object, but are processed in distinct regions of cortex poses the problem of how this information is integrated in an object-specific way: the binding problem. A prominent and experimentally well supported solution postulates oscillatory neural synchronization as a mechanism of binding: Clusters of neurons that are indicative for different properties sometimes show synchronous oscillatory activity, but only when the properties indicated are instantiated by the same object in the perceptual field; otherwise they are firing asynchronously. Synchronous oscillation, thus, might be regarded to fulfill the task of binding together various property representations to form the representation of an object as having these properties (Singer, 1999). Using oscillatory networks as biologically motivated models, it could be demonstrated how the topological organization of information in the cortex, by mechanisms of synchronization, may yield a logically structured semantics of concepts (Werning, 2005). Oscillation functions play the role of object concepts. Clusters of feature sensitive neurons play the role of attributive concepts. Schnitzler et al. (2006) could experimentally demonstrate the essential role of neural synchronization for action control. This may justify

the extension of the synchrony-based neuro-frame approach from features to affordances.

Provided that a concept is completely decomposable into a fully specified frame and provided that neural maps for each attribute can be identified in the cortex, the degree to which the cortex represents an object as an instance of the concept is rendered by a general pattern of synchronizing neural activity distributed over neural clusters that correspond to the basic values of the frame. This pattern may be called the cortical fingerprint of the concept.

Support for the theory of neuro-frames also comes from a number of neuro-linguistic studies. Based on a review of neurobiological data, Pulvermüller (1999) suggests that neural assemblies that pertain to the sensori-motor cortices and are bound by neural synchronization play an important role in understanding the meanings of words. FMRI studies (Pulvermüller, 2005) regarding the understanding of verbs, e.g., hint to a differential top-down activation of motor and pre-motors areas. We know that the understanding of concrete nouns like *hammer*, for which not only features, but also affordances are salient, results in an activity distributed over the premotor and the visual cortex (Martin et. al. 1996). The hypothesis that words for substance concepts arouse more widely distributed activity than words for attributive concepts is, furthermore, supported by EEG studies (Rappelsberger et al., 2000).

From this and further evidence (reviewed by Martin, 2007, Werning, 2012) we may conclude that the correlates of substance concepts are highly distributed neural states. Substance concepts are thus not expected to be realized by single cells, or locally circumscribed regions of the cortex, but by cell assemblies that may pertain to highly distinct parts of the cortex and involve perception as well as motor areas. In contrast, the neural correlates of attributive concepts would be constrained to local cortical regions. The view that substance concepts decompose into complex frames and that their neural realizations are widely distributed in cortex contradicts the doctrines of atomism and modularism.

### **Evolution and Development of the Syntax-semantics Interface**

Another strategy to avoid the paradox is to limit the scope of the assumption P2 that the meanings of concrete nouns are substance concepts. One might advocate a meaning shift of a certain kind in nouns during development or evolution: Whereas for modern adults concrete nouns express substance concepts with a complex semantics, it might be that the child's usage of the noun *mama* only labels a salient person in his or her daily life or that, for an early human, the noun for water just expressed the affordance of being drinkable. It is indeed very likely that the concepts expressed by nouns change in development and evolution. [birth-giving] is not a conceptual part of [mama] for the two-year old as it is for us. Early humans did not represent water as molecularly complex. However, is it plausible that

nouns of young children and early humans do not at all express substance concepts with some decent, if only different, semantic complexity? How could the word *mama* in the child's language be a label for a particular person if the child were not able to recognize and treat that person as *mama* (in his/her sense)? To recognize and treat *mama* as *mama*, the child mentally represents a number of salient features and affordances. Otherwise we would have to withdraw to a rather unwarranted iconic theory of representation.

In the case of phylogeny, the challenge could also be phrased as follows: Was there a time when [water] was an attributive concept – for a simple affordance or feature? That substance concepts finally reduce to a single attributive concepts is the tenet of essentialism: If essentialism about conceptual representation is true, we represent a substance by an essential feature or affordance which the substance must never change. The problem is that for most everyday substances one can hardly find any cognitively plausible candidates for essences. Being H<sub>2</sub>O is essential for water, but is this how humans cognitively represent water? The alternative is to decompose a substance concept into a structure of feature and affordance concepts, none of which specifies an essential property, but only a typical one. Even though water prototypically is tasteless, there is salty water. Water can change its color, taste, aggregate state, etc., even though some values for each of those attributes are more typical than others. Water is also used in typical ways: for drinking, washing, swimming, but it can also be burned by magnesium torches.

There are, of course, lots of nouns in English that express single attributive concepts: abstract nouns. The large majority of them are morphologically derived or, at least, syntactically marked (compare *water* to *beverage*, *fluidity*, etc.). This indicates that nouns expressing single attributive concepts are evolutionarily rather late. There is thus little evidence that [water] in the early stages of language evolution ever was a semantically simple attributive concept, rather than a semantically complex substance concept as it is today. P2 holds also for the early stages of development and evolution.

The remaining option to attack the paradox seems to be the principle P5 that capabilities demanding more effort, *ceteris paribus*, develop and, respectively, evolve later than those demanding less effort. One might argue that the demand of effort is not the only, maybe not even the most important factor that determines evolutionary priority. One may point out that there is stronger evolutionary pressure to lexicalize concepts as complex as substance concepts than to lexicalize attributive concepts. It arguably is rather economic to lexicalize concepts for often recurring, highly specific entities of great survival value. Telling someone that there are bananas somewhere is not only shorter, but also more exact than telling someone that there are sweet, longish, bowed, bright, yellow things around that one may peel and eat. However, an appeal to greater selection pressure does not suffice to explain evolutionary priority:

To explain why proto-birds evolved wings, one has to appeal to some sort of evolutionary pressure to fly. If flying did not have a selective advantage for proto-birds, wings would not have evolved. Maybe proto-birds had to reach or leave trees quickly to escape predators. However, if proto-birds had not had feathers in the first place (maybe for cooling as some hypothesis goes), wings would not have evolved either. Even if selection pressure were maximal and flying the only way a certain reptile species could have survived, if the species did not have feathers and very wing-like forelimbs, it would have died out rather than evolve wings. In addition to evolutionary pressure any explanation of capabilities must appeal to some step-by-step evolution of mechanisms: from the more primitive to the more complex.

What we still have no answer for is the following question: How could a mechanism evolve that enables certain regions of cortex that are involved in representing a word (phonologically, syntactically, etc.) to address those regions of the sensori-motor cortices that represent the word's meaning, i.e., the concept it expresses. Given that semantically complex words are evolutionary prior such an interface must have had strong distributive capacities from the beginning.

## References

- Barrett, H., & Kurzban, R. (2006). Modularity in cognition. *Psychological Review*, 113, 628–47.
- Barrett, M. (1995). Early lexical development. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language* (pp. 362–92). Cambridge, MA: Blackwell.
- Barsalou, L. (1992). Frames, concepts, and conceptual fields. In A. Lehrer & E. Kittay (Eds.), *Frames, fields, and contrasts* (pp. 21–74). Hillsdale, NJ: Erlbaum.
- Barsalou, L. (2005). Situated conceptualization. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 619–50). St. Louis: Elsevier.
- Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.
- Boas, F., & Deloria, E. (1939). Dakota grammar. *Memoirs of the National Academy of Sciences*, 23(2).
- Clifton, C., & Ferreira, F. (1987). Modularity in sentence comprehension. In J. Garfield (Ed.), *Modularity in knowledge representation and natural-language understanding* (pp. 277–90). MIT Press.
- Crair, M., Ruthazer, E., Gillespie, D., & Stryker, M. (1997). Ocular dominance peaks at pinwheel center singularities of the orientation map in cat visual cortex. *Journal of Neurophysiology*, 77, 3381–5.
- Crowley, T. (1995). Inalienable possession in Paamese grammar. In H. Chappell & W. McGregor (Eds.), *The grammar of inalienability* (pp. 383–432). Berlin: de Gruyter.
- Dixon, R. (1999). Adjectives. In K. Brown, J. Miller, & R. Asher (Eds.), *Concise encyclopedia of grammatical categories* (pp. 1–8). Amsterdam: Elsevier.
- Felleman, D., & van Essen, D. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1, 1–47.
- Fodor, J. (1992). *A theory of content and other essays*. Cambridge, MA: MIT Press.
- Fodor, J., & Lepore, E. (1992). *Holism: A shopper's guide*. Oxford: Blackwell.
- Friederici, A. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, 6, 78–84.
- Givon, T. (1970). Notes on the semantic structure of English adjectives. *Language*, 46, 816–37.
- Ingram, D. (1989). *First language acquisition; method, description and explanation*. Cambridge: CUP.
- Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, 58, 25–45.
- Martin, A., Wiggs, C. L., Ungerleider, L., & Haxby, J. V. (1996). Neural correlates of category specific knowledge. *Nature*, 379, 649–52.
- Millikan, R. G. (1998). A common structure for concepts of individuals, stuffs and real kinds: More mama, more milk, and more mouse. *Behavioral and Brain Sciences*, 21, 55–100.
- Mithun, M. (2000). Noun and verb in Iroquoian languages: Multicategorisation from multiple criteria. In P. Vogel & B. Comrie (Eds.), *Approaches to the typology of word classes* (pp. 397–420). Berlin: de Gruyter.
- Petersen, W., & Werning, M. (2007). Conceptual fingerprints: Lexical decomposition by means of frames – a neuro-cognitive model. In U. Priss, S. Polovina, & R. Hill (Eds.), *Conceptual structures: Knowledge architectures for smart applications* (LNAI 4604) (pp. 415–28). Heidelberg: Springer.
- Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, 22, 253–79.
- Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6, 576–82.
- Quiñero, R., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single-neurons in the human brain. *Nature*, 435, 1102–7.
- Rappelsberger, P., Weiss, S., & Schack, B. (2000). Coherence and phase relations between EEG traces recorded from different locations. In R. Müller (Ed.), *Time and the brain* (pp. 297–330). Harwood.
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169–92.
- Schnitzler, A., Timmermann, L., & Gross, J. (2006). Physiological and pathological oscillatory networks in the human motor system. *Journal of Physiology*, 99, 3–7.
- Singer, W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron*, 24, 49–65.
- Werning, M. (2005). The Temporal Dimension of Thought: Cortical Foundations of Predicative Representation. *Synthese*, 146, 203–24.
- Werning, M. (2008). The complex first paradox why do semantically thick concepts so early lexicalize as nouns? *Interaction Studies*, 9(1), 67–83.
- Werning, M. (2010). Complex first? On the evolutionary and developmental priority of semantically thick words. *Philosophy of Science*, 77, 1096–1108.
- Werning, M. (2012). Non-symbolic Compositional Representation and Its Neuronal Foundation: Towards an Emulative Semantics. In Werning, M., Hinzen, W., & Machery, M. (Eds.), *The Oxford Handbook of Compositionality*. OUP, Oxford.
- Werning, M., Hinzen, W., Machery, E. (2012, eds.). *The Oxford Handbook of Compositionality*. Oxford: OUP.

# Order Effects in Moral Judgment

## Searching for an Explanation

Alex Wiegmann (alex.wiegmann@psych.uni-goettingen.de)

Department of Psychology, Gosslerstr. 14  
University of Göttingen, Germany

Yasmina Okan (yokan@ugr.es)

Department of Experimental Psychology, University of Granada  
Campus Universitario de la Cartuja s/n, 18071, Granada, Spain

### Abstract

Research on moral judgment has shown that the order in which dilemmas are presented to subjects often has a strong influence on their judgment. However, the psychological mechanisms underlying order effects are still opaque. In this paper we aimed to isolate the features that a scenario must exhibit in order to influence judgment of subsequent scenarios. For this enterprise, we identified several features from a scenario known to cause order effects, and tested which of these features are necessary to influence subsequent scenarios. Although we still do not have a full understanding of what causes order effects, we made significant progress towards this aim. In five experiments we ruled out some promising explanations such as order effects being driven by an emotional activation linked to the first scenario. Instead, we found order effects to depend on whether the scenario being influenced and its preceding scenario share rather subtle structural similarities.

**Keywords:** Order effects; moral judgment; trolley dilemma.

### Introduction

Imagine one group of subjects is presented with two moral dilemmas, A and B, one after the other. For each of these scenarios subjects have to make a judgment concerning which of two different hypothetical actions should be taken by the agent in each case. In both dilemmas, the life of people is at stake. Imagine a second group of subjects is presented with the same task, the only difference being the order in which the two scenarios are presented. From a normative perspective it seems clear that the order of presentation should not influence subjects' judgments. However, a number of studies have shown that the order of presentation actually influences judgments. Moreover, the impact of the order of presentation is often stronger than that of factors that are generally considered to influence moral judgments (e.g., the existence of physical contact with the potential victim in the scenarios; Wiegmann, Okan, & Nagel, 2012). Interestingly, not only lay people are susceptible to order effects but also professional philosophers (Schwitzgebel & Cushman, 2012). In the paper at hand we present five experiments aiming to identify the factors causing order effects. The question guiding the experiments that will be reported is: Which are the features of a scenario known to cause order effects that enable it to influence other scenarios?

Wiegmann et al. (2012) claimed that research on moral judgment pointed to a systematic pattern of order effects that had been previously overlooked: Only judgments of actions that are normally (i.e., if judged in isolation) regarded as morally acceptable are affected by the order of presentation, and this is only the case if the dilemma is immediately preceded by a dilemma in which the proposed action was not considered as morally acceptable. If there is such a constellation, judgments of actions normally regarded as morally acceptable can approach judgments of previous actions (i.e., they can be deemed as less acceptable).

In order to test this claim Wiegmann and colleagues presented two groups of subjects with five trolley dilemmas, one after another (see Table 1). In all cases a train out of control was heading towards three railroad workers. An action was described that could be conducted by an agent in the situation to save the workers. This action varied in each of the five scenarios. The ordering of the scenarios was based on the level of agreement with the proposed action in each case, according to independent judgments provided in an independent pilot study (see Table 1). Level of agreement was measured on a scale of 1 to 6, where 1 was "not at all", and 6 was "absolutely". While in one group the level of agreement with the proposed action steadily increased, it decreased in the other group.

Table 1: Summaries of the actions proposed in the five dilemmas.

Scenario	Proposed action
Push	Push a large person from a bridge to stop the train
Trap	Push a button that will open a trap door that will let a person on top of the bridge fall and stop the train
Redirect	Redirect a train with a person inside that is on a parallel track onto the main track to stop the train
Run Over	Redirect an empty train that is on a parallel track onto the main track to stop the train, running over a person that is on the connecting track
Standard	Press a switch that will redirect the train that is out of control to a parallel track where one person will be run over

Table 2: Mean ratings (standard deviations) of agreement and percentage of subjects disagreeing with the proposed action in the five scenarios when evaluated independently.

Measure	Scenario (each $n=20$ )				
	Push	Trap	Redirect	Run Over	Standard
Mean Rating	1.95	3.4	4.15	4.4	4.45
(SD)	(1.76)	(1.76)	(1.42)	(1.14)	(1.15)
% Disagreement	80	40	30	10	15

Note. % Disagreement is the percentage of subjects who gave a rating  $\leq 3$  on a scale ranging from 1 to 6.

According to the pattern of order effects outlined above, subjects' ratings for actions in the condition where the level of agreement was steadily decreasing (i.e., from Standard to Push; in the following called Least Aversive First, LAF) should not differ from ratings for the same actions when presented separately. The reason is that in such a constellation it is never the case that a judgment of an action normally (i.e., if judged in isolation) regarded as morally acceptable is preceded by an action that is normally regarded as morally unacceptable. In contrast, ratings for the actions in the last three scenarios in the Most Aversive First (MAF) condition (i.e., Redirect, Run Over and Standard) should decrease to the level of agreement of the preceding scenario (i.e., Trap), according to the pattern outlined above. That is, the low rating of Trap is assumed to reduce the level of agreement in Redirect, that in turn is assumed to decrease the rating for Run Over, that eventually decreases Standard's rating. This prediction was confirmed (see Figure 1). Unexpectedly, the ratings of the action in Trap were also affected by the ratings of the action in Push, although Trap is normally judged as unacceptable by a slim majority of subjects.

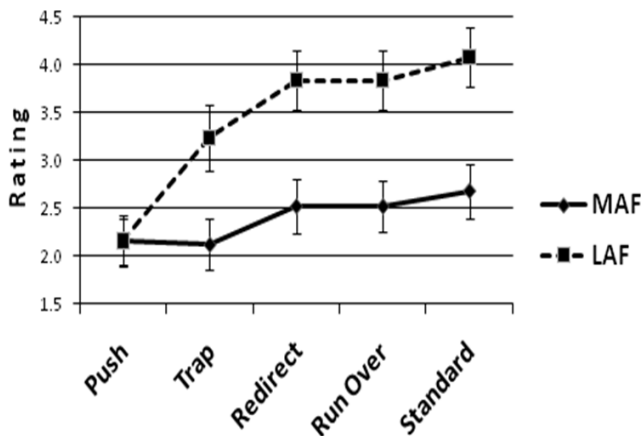


Figure 1: Mean ratings of agreement (1 stands for “not at all, 6 stands for “absolutely”) with the proposed action in the five scenarios when evaluated sequentially, as a function of the order of presentation. Error bars indicate SEM. MAF = Most Aversive First; LAF = Least Aversive First.

This finding motivated a closer look at the results at the level of individual participants. In particular, the data was explored treating the ratings as a set of binary choices made by each participant (i.e., treating ratings  $\leq 3$  as indication of disagreement and ratings  $\geq 4$  as indication of agreement with the proposed action). The following tendency was observed: A disagreement with an action was virtually always “transferred” to the judgment of the action in the next scenario. That is, an action receiving a positive rating when judged independently received lower ratings when presented as part of a sequence if the preceding action was rated negatively by the same participant. However, positive ratings did not affect the ratings of the next action (by changing them into positive ones) if this action was rated negatively in independent ratings. Reformulating the pattern this way allows order effects to occur not only for actions rated positively when judged independently, but also for actions rated negatively on average. It just has to be the case that the number of participants that disagree with the proposed action in a particular scenario is sufficiently higher than the number of participants that disagree with the action in the subsequent scenario. This excess of “disagreements” can be transferred to the next scenario and cause an order effect. On the flipside, an order effect might also occur when a particular dilemma is preceded by another one where the proposed action is judged positively. Again, it just has to be the case that the number of disagreements in the preceding scenario is sufficiently higher than in the following scenario.

Although the pattern outlined at the individual level fits the data and allows making accurate predictions, the psychological mechanisms underlying order effects are still opaque. The experiments reported below pinpointed some of the features of the Push scenario that could be affecting other scenarios, and tested the effect of each of them individually. Features include differences in an emotional activation associated with Push, the activation of moral principles (e.g., “do not kill”) and the trade-off of lives.

## Experiment 1

In this experiment we aimed to test the hypothesis that the order effect described can be explained in terms of an emotional activation linked to Push, which would affect judgments in subsequent scenarios. As Green and his collaborators have shown, dilemmas like Push are more likely to activate brain regions associated with emotional processing than dilemmas like Standard (Green, Sommerville, Nystrom, Darley, & Cohen, 2001). Thus, in the sequence of scenarios described above (MAF) subjects might first experience a negative emotion when they are presented with Push, and once this negative emotion is in place, it might lead subjects to judge all the actions proposed in the remaining scenarios as morally unacceptable (cf. Haidt, 2001, Prinz, 2007). If the activation of such negative emotion is sufficient to cause order effects, then the presentation of other aversive scenarios that elicit a



similar emotion should also affect the judgment for other less aversive dilemmas (e.g., Standard).

**Participants** 259 subjects were recruited for a compensation of £ 0.50 via an online database located in the U.K..

**Design, Materials, and Procedure** Subjects were randomly assigned to one of four conditions. In two conditions subjects first had to read an aversive story, and then they were asked to judge the proposed action in Standard. The story was different in each of these two conditions. The following two stories were used:

**Incest (Haidt, 2001):**

Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other.

**Starving Child (actual newspaper article):**

The 41-year-old man and 25-year-old woman, who met through a chat website, reportedly left their infant unattended while they went to internet cafes. They only occasionally dropped by to feed her powdered milk.

According to the Yonhap news agency, South Korean police said the couple had become obsessed with raising a virtual girl called Anima in the popular role-playing game Prius Online. The game, similar to Second Life, allows players to create another existence for themselves in a virtual world, including getting a job, interacting with other users and earning an extra avatar to nurture once they reach a certain level.

In the two remaining conditions subjects had to either judge Standard alone (to obtain a baseline rating), or Standard after having judged Push. Additionally, as the study was conducted online, at the end of the questionnaire subjects completed a simple logical task to identify those who did not pay sufficient attention to the task.

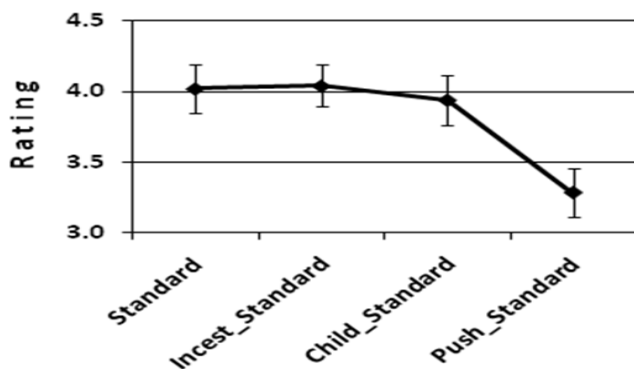


Figure 2: Mean Ratings for Standard as a function of the preceding scenario. Error bars indicate SEM.

**Results** Fifty subjects dropped out because they did not

answer the test question, failed to solve the logical question or went through the whole survey in less than 40 seconds.

The results for the remaining subjects are summarized in Figure 2. The mean rating for Standard when judged independently was 4.02 ( $SD=1.18$ ), while it was lower ( $M=3.28$ ,  $SD=1.39$ ) when it was preceded by Push (Push\_Standard), as predicted. A planned-contrast test confirmed this difference to be significant,  $F_{1,205}=9.68$ ,  $p<0.001$ . In contrast to this, reading and judging Newspaper or Incest did not have any effect on Standard (Standard vs. Incest\_Standard:  $F_{1,205}=0.01$ ,  $p>0.9$ ; Standard vs. Child\_Standard:  $F_{1,205}=0.10$ ,  $p>0.7$ ).

**Discussion** The findings obtained suggest that an emotional activation may not be sufficient to cause the kind of order effect described above. That is, judgments for Standard were not affected by the prior presentation of different scenarios that are likely to have elicited negative emotions. Further evidence for this idea comes from ongoing research conducted in Spain in which participants were presented with selected pictures of unpleasant affective valence and high arousal, before judging Standard. In line with the study described above, preliminary results revealed that the emotional priming did not affect judgments for Standard.

## Experiment 2a and 2b

The two following experiments test the hypothesis that the order effect described is related to differences in the activation of principles associated with each dilemma. In particular, Push could trigger the urge for subjects to justify their judgment, or the principle “Do not kill!”, while Standard may not. If the activation of such principle can account for the carry over effect of judgments when Push is presented first, it is reasonable to expect that forcing the activation of a principle relevant for Standard when this dilemma comes first (e.g., “Save the greater number”) should lead to an order effect in the opposite direction (i.e., people should be more likely to agree with the action proposed in subsequent scenarios).

### Experiment 2a

The rationale for this experiment was as follows: When subjects judge the action proposed in Push as morally unacceptable they articulate, so to say, a prohibition or imposing a ban. Actions which we prohibit are generally accompanied by a justification. For instance, one often has to justify or explain to kids why something is forbidden. In contrast, there are fewer situations where one has to explain why something is allowed. Justifications for allowed actions generally only happen when the actions were expected to be forbidden. In Push, the vast majority of subjects judge the proposed action as forbidden while in Standard they don't. Hence, it could be the case that subjects first judging Push have a stronger urge to justify their judgment. If the justification is “You are not allowed to kill innocent people” (or something similar), judgments for subsequent scenarios could accommodate this justification, explaining why all

proposed actions are regarded as not acceptable. In contrast, subjects starting with Standard might not have an urge to justify their initial judgment, explaining why subsequent judgments are not affected.

**Participants** 36 subjects were recruited from a student subject pool. Participants were compensated with course credits.

**Design, Materials, and Procedure** Subjects were presented with the LAF condition. However, in contrast to the original LAF condition described above, here participants were required to justify their judgment for Standard. Since participants were recruited from the same student pool (mainly psychology students) and this experiment was conducted only a few weeks after the one conducted by Wiegmann and colleagues (2012), the original LAF condition was used as a control condition.

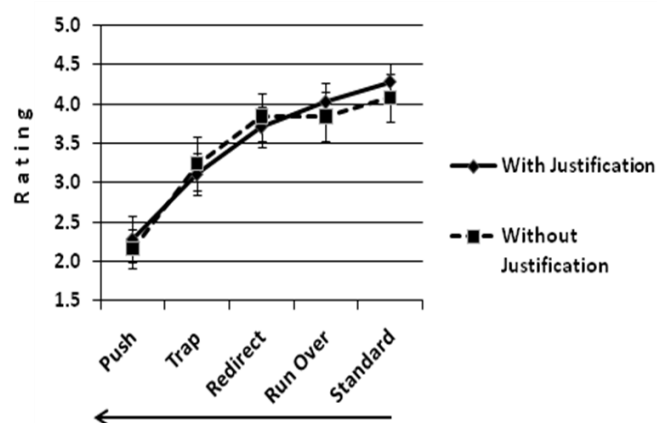


Figure 3: Mean Ratings for the five scenarios as a function of whether subjects had to justify their judgment for Standard. Arrow indicates order of presentation. Error bars indicate SEM.

**Results** One subject failed to answer all questions for unknown reasons. Results for the remaining subjects are summarized in Figure 3. As can be seen, being forced to justify the judgment for Standard did not influence judgments for the actions proposed in the following scenarios. An ANOVA with justification (required vs. not required) as a between-subjects factor and scenario as a within-subjects factor revealed that there was neither a main effect of justification ( $F_{1,58}=0.02, p=.88$ ) nor an interaction between justification and scenario ( $F_{4,232}=0.62, p=.65$ ).

## Experiment 2b

The rationale for this experiment was as follows: When judging Push the principle “Do not kill!” comes easily to mind because the proposed action in this dilemma is a paradigmatic case of killing a person. In contrast, Standard might not trigger such a clear principle. In philosophy, the permissibility to intervene in Standard is often justified by rather subtle principles like the Doctrine of Double Effect or

even more subtle principles (Kamm, 2007). In this experiment we sought to trigger a clear principle (saving the greater number) and test whether it would be carried over to the following scenarios, thereby affecting judgments.

**Participants** 63 subjects were recruited from a student subject pool. Participants were compensated with course credits.

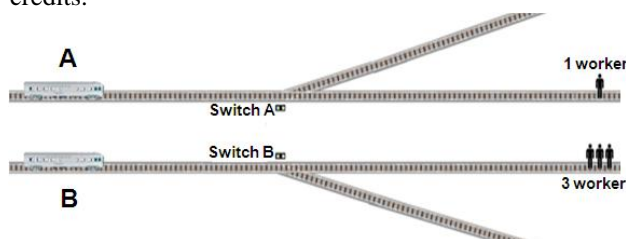


Figure 4: Illustration of the Rescue scenario

**Design, Materials, and Procedure** Two of the three conditions tested were LAF and MAF. The third condition included a new scenario called Rescue that was placed before Standard and was intended to trigger the principle “Save the greater number of lives” (see Figure 4). In this scenario a train is threatening one person and another train is threatening three persons. There is not enough time to throw the switch for both trains so that everyone is saved. We assumed that in such a case virtually everyone would agree to throw the switch that will save three persons rather than only one.

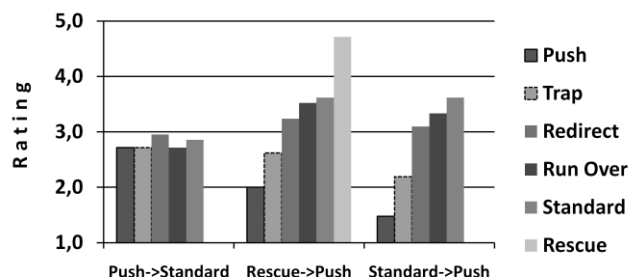


Figure 5: Mean ratings for the 5 (6) scenarios as a function of both the order of presentation and whether the Rescue scenario was present. For example, Push->Standard means subjects started with Push and ended with Standard.

**Results** The results are shown in Figure 5. As expected, a transference in judgments was observed for MAF, but not for LAF ( $F_{4,160}=11.37, p<.001$ ). However, introducing Rescue before Standard did not affect judgments for Standard, as evidenced by the absence of an interaction of condition and scenario between LAF and the new condition starting with Rescue ( $F_{4,160}=0.60, p=.67$ ).

**Discussion** The results of these two experiments can be interpreted in at least two ways. First, it could be the case that the order effect in MAF is neither based on a stronger urge to justify one’s judgment in Push (2a) nor due to Push triggering a clear principle (2b). Alternatively, the order effect in MAF could indeed be driven by a principle

triggered by Push which is carried over. However, the principle triggered by Standard or Rescue may not be, so to say, strong enough to override intuitions in other scenarios. In other words, while a principle like “Do not kill!” may be carried over once it has been triggered, a principle like “Save the greater number!” may not be potent enough to influence moral judgments in following scenarios (cf, e.g., Gert, 2007).

### Experiment 3

In this experiment we took a further step aiming to investigate which features of Push are necessary to cause an order effect. In particular, we examined the impact of the number of lives that are traded-off in Push.

**Participants** 343 subjects, each receiving £ 0.50, were recruited via an online database located in the U.K..

**Design, Materials, and Procedure** Subjects were randomly assigned to one of four conditions. In one condition Standard was judged alone (to obtain a baseline rating), while in a second condition Standard was judged after Push, involving the same number of potential victims as in previous experiments (Standard\_Push\_3). In the third and fourth conditions we manipulated the number of potential victims that would be saved by the intervention in Standard. In one condition nobody would be saved by pushing the person from the bridge (Standard\_Push\_0), while in the other condition a group of one hundred people would be saved (Standard\_Push\_100).

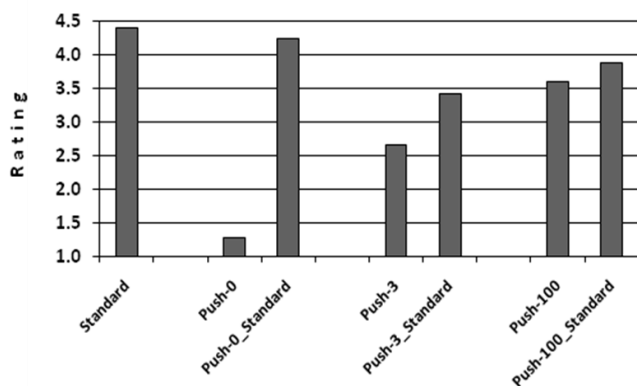


Figure 6: Mean ratings for Standard and its preceding scenario, i.e. Push-0 is the rating for Push if nobody is rescued by killing the one person and Push-0\_Standard is the rating for Standard if preceded by Push-0.

**Results** 96 subjects did not answer the test question, gave the wrong answer to the logical task, or took less than 40 seconds for completing the task.

The results for the remaining subjects are summarized in Figure 6. The baseline rating for Standard was the highest in descriptive terms ( $M=4.4$ ,  $SD=1.26$ ). The next highest rating for Standard was delivered if it was preceded by the version of Push where pushing the man from the bridge would not

save anyone (i.e., Standard\_Push\_0) ( $M=4.26$ ,  $SD=1.46$ ). The difference between Standard in this condition and the baseline rating was not significant,  $F<1$ . In contrast, ratings for Standard were lower in Standard\_Push\_3 and in Standard-Push-100 than for the baseline ( $F_{1,243}=18.14$ ,  $p<0.001$  and  $F_{1,243}=4.57$ ,  $p<0.05$ , respectively).

**Discussion** One might wonder why the rating for Standard is lowered to a lesser extent when it is preceded by Push involving saving 100 people than by Push involving saving three. However, the pattern outlined above can account for this finding. Recall that most of the time only negative ratings were transferred to the next scenario. Since the rating for Push\_100 ( $M=3.61$ ,  $SD=1.50$ ) was already high, it is not surprising that the rating for Standard\_Push\_100 was relatively high too. There were seemingly just not enough negative ratings for Push\_100 to be transferred and lower the rating of Standard\_Push\_100 to the same extent as for Standard\_Push\_3. Interestingly, the results show that Standard is not influenced by a version of Push in which killing the person does not save anyone. Since Push\_0 and Push\_3 only differ with regards to whether there is a trade-off of lives involved, we can infer that a dilemma must contain such trade-off to influence judgments for Standard.

### Experiment 4

The results of experiment 3 suggest that a scenario preceding Standard needs to contain a trade-off of lives in order to influence Standard. In this experiment we aimed to investigate whether Standard could be influenced by a preceding scenario similar to Push with regards to being aversive and containing a trade-off, but with a different cover story. Furthermore, we tested whether just reading (and not judging) Push or similar scenarios would be sufficient to influence Standard.

**Participants** 321 subjects were recruited for a compensation of £ 0.50 via an online database located in the U.K..

**Design, Materials, and Procedure** Subjects were randomly assigned to one of five conditions. As in previous experiments there was one condition to get a baseline for Standard and another where Standard was preceded by Push. In a third condition (Organ) Standard was preceded by a scenario in which a doctor can save three patients by transplanting organs from a healthy person into them.. The last two conditions, (Push\_readonly) and (Organ\_readonly) were identical to Push and Organ, respectively, except that subjects were not given the opportunity to judge the proposed action.

**Results** 43 subjects did not answer the test question, gave the wrong answer to the logical task, or took less than 40 seconds for completing the task.

The results for the remaining subjects are summarized in Figure 7. The baseline rating for Standard was the highest in

descriptive terms ( $M=4.35$ ,  $SD=1.25$ ). Again, the difference between Standard\_Push\_3 ( $M=3.37$ ,  $SD=1.37$ ) and the baseline rating was, as expected, significant ( $F_{1,282}=14.46$ ,  $p<0.001$ ). Interestingly, ratings for Standard when preceded by Organ were also significantly lower than the baseline ( $F_{1,282}=6.13$ ,  $p<0.05$ ). As Figure 7 shows, it did not make a difference whether subjects just read or also judged the action proposed in Push or Organ.

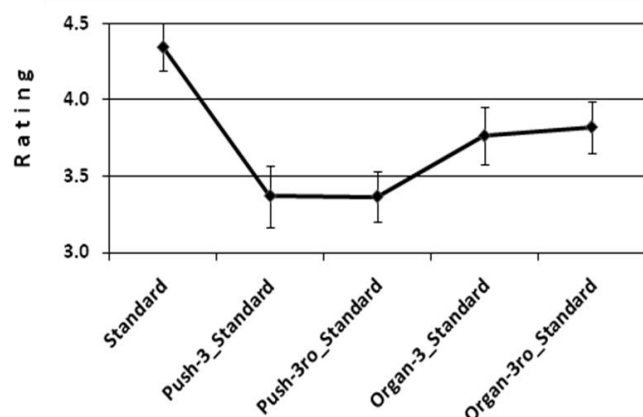


Figure 7: Mean Ratings for Standard as a function of the preceding scenario. Error bars indicate SEM. “ro” stands for “read only”.

**Discussion** The results showed that Standard can be influenced by a scenario with a different cover story than Push, but that also involves a trade-off of lives. This influence is not as strong as the influence of Push but it is still significant. Moreover, whether subjects only read a scenario or also judged it did not affect judgments.

### General Discussion

In this paper we sought to search for an explanation of order effects in moral judgment. The overarching question guiding our experiments was: Which features of the Push scenario are the ones that enable Push to influence other scenarios such as Standard? Although we still do not have a complete explanation, we think we have made some progress towards this aim.

Our findings suggest that the order effect is likely not be caused by negative emotions activated by Push. Presenting subjects with aversive stories which were likely to elicit such emotions did not have any effect on the judgment for Standard (Experiment 1).

Forcing participants to justify their judgment for Standard or saliently triggering the principle “Save as many lives as possible!” did not affect judgments for Standard either (Experiments 2a and 2b). As noted above, these results can be explained in two ways: First, it might be that the order effect is not driven by the activation of a principle for the Push scenario, which is then applied to subsequent scenarios. Second, it might be that the principle “Save lives” affects subsequent judgments only to a lesser extent than the

principle triggered in Push (presumably: “Do not kill!” or “Do not kill in order to save lives!”).

Interestingly, we found a feature of Push that seems necessary to influence a subsequent scenario: The trade-off of lives. A version of Push which did not involve such trade-off had no influence on judgments for Standard. Hence, apart from being judged more negatively than Standard (see Introduction) containing a trade-off of lives seems to be a necessary feature for a scenario to influence subsequent scenarios.

Interestingly, when we presented subjects with a scenario (Organ) similar to Push in that it contained a very aversive action and a trade-off of lives, but that differed with regards to the cover story, the agreement with the action proposed in Standard was also reduced. Future research should examine other similarities between Push and Organ which could be related to their potency to cause order effects.

In summary, our findings suggest that the features that a moral dilemma must exhibit in order to affect judgments in subsequent dilemmas (here, in Standard) are:

1. It must receive significantly more negative ratings than the following dilemma (see Introduction)
2. It must contain a trade-off of lives

Furthermore, if the preceding scenario exhibits these two features the influence on Standard becomes even stronger if the superficial features of the cover story resemble the ones in Standard.

### Acknowledgments

This research was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG WA 621/21-1), and the Courant Research Centre Evolution of Social Behaviour,, University of Göttingen (funded by the German Initiative of Excellence).

### References

- Gert, B. (2007). *Common morality: Deciding what to do*. New York: Oxford University Press.
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293, 2105–2108.
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814–834.
- Kamm, F. M. (2007). *Intricate ethics*. Oxford, England: Oxford University Press.
- Prinz, J. J. (2007). *The emotional construction of morals*. New York: Oxford University Press.
- Schwitzgebel, E., & Cushman, F. (in press). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind and Language*.
- Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology*,

# Explaining increases belief revision in the face of (many) anomalies

Joseph Jay Williams (joseph\_williams@berkeley.edu)

Caren M. Walker (caren.walker@berkeley.edu)

Tania Lombrozo (lombrozo@berkeley.edu)

Department of Psychology, 3410 Tolman Hall  
Berkeley, CA, 94709

## Abstract

How does explaining novel observations influence the extent to which learners revise beliefs in the face of anomalies – observations inconsistent with their beliefs? On one hand, explaining could recruit prior beliefs and *reduce* belief revision if learners “explain away” or discount anomalies. On the other hand, explaining could *promote* belief revision by encouraging learners to modify beliefs to better accommodate anomalies. We explore these possibilities in a statistical judgment task in which participants learned to rank students’ performance across courses by observing sample rankings. We manipulated whether participants were prompted to explain the rankings or to share their thoughts about them during study, and also the proportion of observations that were anomalous with respect to intuitive statistical misconceptions. Explaining promoted greater belief revision when anomalies were common, but had no effect when rare. In contrast, increasing the number of anomalies had no effect on belief revision without prompts to explain.

**Keywords:** explanation, self-explanation, learning, generalization, statistics, misconceptions, anomalies.

## Introduction

Human learning relies on the ability to use novel observations about the world to revise current beliefs. This raises basic questions about how observations impact beliefs, and in particular how different cognitive processes influence the process of belief revision. The current paper examines how *explaining* observations that are anomalous with respect to a learners’ current beliefs influences the nature of belief revision.

Previous research reveals that seeking and generating explanations can play an important role in learning and reasoning across a variety of task and domains, including learning novel categories (Williams & Lombrozo, 2010), inferring causal relationships (Koslowski, 1996), and generalizing the properties of people and objects from known to unknown cases (Rehder, 2009; Sloman, 1994). In addition, generating explanations can drive conceptual development in children (Wellman & Liu, 2006) and has been shown to have important pedagogical benefits in a variety of educational contexts (e.g. Fonseca & Chi, 2011).

Despite widespread appreciation for the impact of explaining on learning and belief revision, current accounts of explanation’s effects pose a challenging puzzle. On the one hand, explaining is frequently hypothesized to encourage learners to accommodate novel observations in

the context of their prior beliefs (Ahn, Brewer & Mooney, 1992; Chi et al, 1994; Lombrozo, 2006; Walker, Williams, Lombrozo & Gopnik, under review; Williams & Lombrozo, 2010). This suggests that explaining could lead learners to draw on a range of belief-preserving strategies in the face of anomalous observations (Chinn & Brewer, 1993; Kuhn, 1962; Koslowski, 1996; Lord, Lepper, & Ross, 1979). Learners who seek explanations could engage in less belief revision (relative to those who do not explain) if they “explain away” anomalies by discounting them as implausible, or reinterpret observations as consistent with or irrelevant to preferred theories.

On the other hand, explaining anomalies could prompt learners to reject their currently-held beliefs and construct new theories that better accommodate the anomalous observations. For example, explaining could ensure that anomalies are not simply ignored, increasing learners’ allocation of attention and processing time to these unanticipated observations (Legare, 2010; Siegler, 2002). In addition, explaining could encourage learners to seek and discover general patterns that go beyond prior beliefs to capture the anomalies being explained (Walker, Williams, Lombrozo & Gopnik, under review; Williams & Lombrozo, 2010).

A third possibility is that explaining has the potential to produce both effects. Whether explaining preserves or revises beliefs could depend on the nature of the evidence provided by the anomalies. For example, relative to other learning strategies, explaining could encourage learners to discount anomalous observations if they are infrequent or there is no alternative theory available to explain them. However, seeking explanations for more extensive inconsistencies could emphasize the limitations of current beliefs and guide learners to alternative theories, thus promoting more radical belief revision.

We investigate these possibilities in the context of a statistical judgment task in which participants learn a system for ranking students’ performance across different courses. Participants learn to rank by observing sample rankings, and they are either prompted to provide explanations for the rankings or to engage in free study as a control condition. Previous research suggests that participants’ prior beliefs will favor rankings on the basis of statistically problematic principles rather than one that is normatively defensible (e.g., Schwartz & Martin, 2003; Belenky & Nokes-Malach, in press). We can therefore manipulate how many of the sample rankings happen to be consistent with the non-

normative principles, and how many are anomalous and only accounted for by the correct principle. This allows us to examine how the process of belief revision is influenced by explaining the novel observations (the sample rankings), by the number of anomalous observations, and by any interaction between these factors.

## Experiment

Participants were instructed to learn how a university ranks their students across different courses by studying five examples of ranked pairs of students. Each example reported which of two students in two different courses was ranked higher, listing each student's (percentage) grade along with each course's mean grade, average deviation (the average absolute deviation), and minimum and maximum grade. We used average deviation as a measure of variability instead of standard deviation to follow past research in avoiding the need for formulas and using a concept more transparent to participants (Schwartz & Martin, 2003; Belenky & Nokes-Malach, in press).

All five examples were ranked according to a *relative-to-deviation* principle: the better student was the one that scored a greater number of average deviations above the mean (see Schwartz & Martin, 2003; Belenky & Nokes-Malach, in press). However, some of the examples were also consistent with three non-normative principles (e.g. the student with higher absolute scores is always ranked higher). These are described further in the *Materials & Procedure* section, below.

To probe how anomalous observations influenced belief revision, we manipulated how many of the five ranked examples were consistent (or anomalous) with respect to the non-normative principles. The relative-to-deviation principle always accounted for all five examples. In the *single anomaly* condition, four of the five example rankings also conformed to the non-normative principles, and there was just one anomaly that was inconsistent with them. In the *multiple anomalies* condition, there were four anomalies and only one of the five rankings was consistent with the non-normative principles.

To examine how explaining interacted with anomalies to impact learning, we also manipulated the extent to which participants engaged in explanation. In the *explain* condition we prompted participants to explain each ranked example. In the *free study* control condition, participants were free to use any study strategy, but were prompted to articulate their thoughts while studying each ranked example. Like explaining, this control condition involved paying attention to the details of the cases and articulating one's thinking in language.

As discussed in the introduction, engaging in explanation and varying the number of anomalies could impact belief revision in several ways. Explaining the anomalies could reduce the revision of prior beliefs and inhibit learning about the relative-to-deviation principle. This effect could be especially potent when there is only a single anomaly to the non-normative principles. On the other hand, explaining

could magnify the effects of anomalies in rejecting belief in the non-normative principles and instead encourage learners to induce and adopt the relative-to-deviation principle.

Finally, the effects of explaining could depend on whether the explained observations include a single anomaly or many. Explaining could have a large impact on belief revision in the context of multiple anomalies, but have no effect or even inhibit belief revision when only a single (and easily discounted) anomaly is present. Alternatively, explaining could boost the impact of a single anomaly that might otherwise be ignored, but have no effect (relative to control) when there are multiple anomalies that make the need for belief revision completely apparent. The design of our experiment allows us to differentiate these possibilities.

## Methods

### Participants

Participants were 275 adults recruited online through the Amazon Mechanical Turk marketplace and reimbursed for their time.<sup>1</sup>

### Materials & Procedure

The materials consisted of five examples of student pairs ranked by the university, ten unranked pre-test pairs, and ten unranked post-test pairs. The experiment involved introduction, pre-test, study, and post-test phases.

**Introduction** Participants were informed that they would observe pairs of students from different classes whose academic performance had been ranked by the university, and that their goal was to learn the ranking system employed. They were given the definition of "average" – the sum of all scores divided by the number of students in a class – and "average deviation" – the sum of all the (absolute) differences between student scores and the average, divided by the number of students.

**Ranked examples for study** During study, five examples of ranked student pairs were presented. A ranked example (see Figure 1a and 1b) stated which student was ranked higher by the university, and reported each student's: (1) name (e.g., Sarah); (2) class (e.g., Sociology); (3) class's mean score (e.g., 79%); (4) class's average deviation (e.g., 8%); (5) class's minimum score (e.g., 67%); and (6) class's maximum score (e.g., 90%).

**Principles for ranking students** Participants could interpret or predict the rank of each student pair using at least four principles. The three *non-normative* principles were incorrect but designed to correspond to intuitive statistical misconceptions.

---

<sup>1</sup> We included a question that assessed whether participants were actually reading instructions (Oppenheimer et al, 2009). The pattern of results was the same if participants who did not pass this test were excluded.

(a) Sarah got 85% in a Sociology class, where the average score was 79%, the average deviation was 8%, the minimum score was 67%, and the maximum score was 90%.

Tom got 69% in a Art History class, where the average score was 65%, the average deviation was 3%, the minimum score was 42%, and the maximum score was 87%.

Sarah was ranked more highly by the university than Tom.

(b) Sarah got 85% in a Sociology class, where the average score was 79%, the average deviation was 3%, the minimum score was 67%, and the maximum score was 90%.

Tom got 69% in a Art History class, where the average score was 65%, the average deviation was 8%, the minimum score was 42%, and the maximum score was 87%.

Tom was ranked more highly by the university than Sarah.

Figure 1: (a) A *consistent* ranked example for which all four principles predicted the same ranking. (b) An *anomalous* ranked example constructed by switching the class average deviations of the consistent example from Figure 1a. The switch means that the correct relative-to-deviation ranking is now the opposite of what is predicted by the raw-score, relative-to-average, and relative-to-range principles. Emphasis is added for illustration and was not provided to participants.

We term the principles (1) *raw-score*: the higher ranking went to the student with the higher score, irrespective of mean, average deviation, and minimum or maximum score; (2) *relative-to-average*: the higher ranking went to the student whose score was the farthest above (or least below) the class's mean score; (3) *relative-to-range*: the higher ranking went to the student whose score was farther from the average *relative to the range* in class scores, where this was calculated as the difference from the mean divided by the range.

The *relative-to-range* principle privileges the score that is farther from the mean as measured in "range-units," capturing some notion of variability (when range is correlated with variability), and could be approximated by looking at a score's distance from the maximum score.

The fourth and more accurate *relative-to-deviation* principle favored whichever score was a greater number of average deviations above the mean. This was calculated as the difference from the mean divided by the average deviation, and is closely related to normative measures such as the standard deviation and z-score, indicating the person's score relative to the distribution of scores in the class.

**Consistent vs. anomalous examples** All five ranked examples conformed to the relative-to-deviation principle. However, a ranked study example could be *consistent* with or *anomalous* with respect to the ranking given by the raw-score, relative-to-average, and relative-to-range principles, all of which always generated identical rankings on study examples (see Figure 1a and 1b). Five consistent examples were constructed so that each could be converted to an anomalous example by switching the average deviation of the two students' classes. This permitted a close match between consistent and anomalous examples on all other dimensions (compare Figure 1b to Figure 1a).

In the *single anomaly* condition there were four consistent examples and one anomalous example. The *multiple anomalies* condition had the opposite ratio: one consistent example and four anomalous examples.

**Pre-test** To provide a baseline measure of belief before study, participants were presented with ten unranked student pairs. They judged which student the university would rank higher, and rated confidence in their judgment on a scale from 1 ("not at all") to 7 ("extremely").

The ten student pairs were designed to identify the principle(s) that participants used to rank students, and thus pitted candidate principles against each other. Specifically, there were two instances of each of the following types of pairs, pitting (1) the relative-to-deviation principle against the three non-normative principles (like anomalous study examples); (2) the raw-score principle against the other three principles; (3) the relative-to-average principle against the other three principles; (4) the relative-to-range principle against the other three principles; and (5) the two principles that were most sensitive to variability, relative-to-range and relative-to-deviation, against the raw-score and relative-to-average principles.

**Study** Each of the five ranked examples was presented onscreen for exactly 90 seconds in a format similar to Figure 1a and 1b. Participants in the *explain* condition were prompted to explain why the higher-ranked student was ranked more highly, typing their explanation into a text box onscreen. Participants in the *free study* control condition were told to type their thoughts during study into an equivalent text box.

**Post-test** To assess belief after study, participants' ranking judgments and confidence ratings were solicited for ten unranked student pairs. All names and grades were changed from the pairs used in pre-test, but five points were added to each grade to generate novel numbers while preserving the way in which the items pitted the principles against each other.<sup>2</sup>

<sup>2</sup> Additional questions were asked at the end of the experiment (e.g. demographics, sufficient time for task, strategy) but are not further discussed here in the interest of space.



## Results

**Overall pre- and post-test accuracy** Learning was assessed by comparing accuracy on the pre-test and post-test items. Correct responses were considered to be those that were consistent with the relative-to-deviation principle. Figure 2 reports an overall measure of accuracy across all pre-test and post-test items as a function of learning task and number of anomalies. Accuracy improved from pre- to post-test: A 2 (task: explain vs. free study)  $\times$  2 (number of anomalies: single vs. multiple)  $\times$  2 (test: pre-test vs. post-test) repeated measures ANOVA found a main effect of test (pre- vs. post-) on overall accuracy,  $F(1, 269) = 4.33, p < 0.05$ .

The ANOVA additionally revealed interactions between test and learning task,  $F(1,269) = 4.95, p < 0.05$ , and between test and number of anomalies,  $F(1,269) = 3.88, p < 0.05$ . These effects were driven by a greater boost in pre- to post-test accuracy for participants in the *explain – multiple anomalies* conditions, which in turn was driven primarily by rankings on anomalous items, as discussed further below.

**Anomalous items: Change in pre- to post-test accuracy** Figure 3 reports accuracy on *anomalous* pre-test and post-test items. These items were analogous to the anomalies at study in pitting the relative-to-deviation principle against all three non-normative principles. Accuracy on these items was critical to testing our hypotheses about the effects of explanation and anomalies on the revision of beliefs to favor the relative-to-deviation principle over the non-normative alternatives. Accuracy on anomalous items improved from pre- to post-test: A 2 (task: explain vs. free study)  $\times$  2 (number of anomalies: single vs. multiple)  $\times$  2 (test: pre-test vs. post-test) ANOVA on accuracy on anomalous items revealed a main effect of pre- vs. post- test,  $F(1, 269) = 63.85, p < 0.05$ .<sup>3</sup>

Subsequent analyses directly examined the pre-test to post-test *change* in accuracy on the anomalous items. Figure 4 reports performance on this measure, calculated as post-test accuracy minus pre-test accuracy.

A 2 (task: explain vs. free study)  $\times$  2 (anomalies: single vs. multiple) ANOVA on the pre- to post-test change in accuracy on anomalous items revealed a significant effect of number of anomalies,  $F(1, 266) = 12.8, p < 0.05$ . This main effect was qualified by an interaction between learning task and number of anomalies,  $F(1, 266) = 7.77, p < 0.05$ . No other effects were significant.

<sup>3</sup> The ANOVA on accuracy for anomalous items revealed a number of additional effects, the relevance of which is more readily communicated in our subsequent analyses on the pre- to post- test *change* in accuracy. These include a two-way interaction between test and number of anomalies,  $F(1,266) = 12.80, p < 0.001$ , a three-way interaction between test, learning task and number of anomalies,  $F(1,266) = 7.77, p < 0.01$ , and main effects of task,  $F(1,266) = 4.80, p < 0.05$ , and number of anomalies,  $F(1,266) = 8.45, p < 0.005$ .

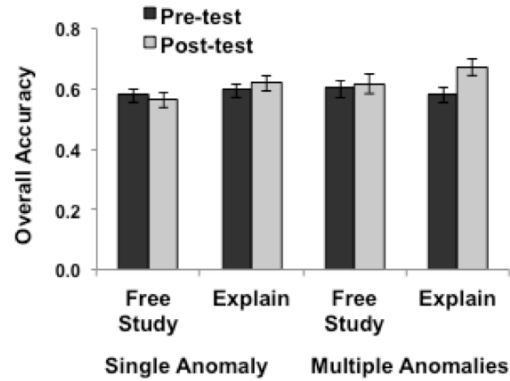


Figure 2: Accuracy on *all* pre-test and post-test items, by learning task and number of anomalies.

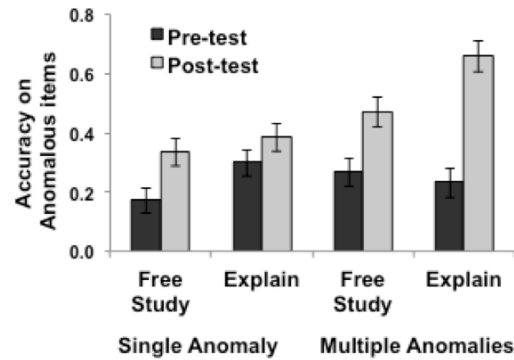


Figure 3: Accuracy on *anomalous* pre-test and post-test items, by learning task and number of anomalies.

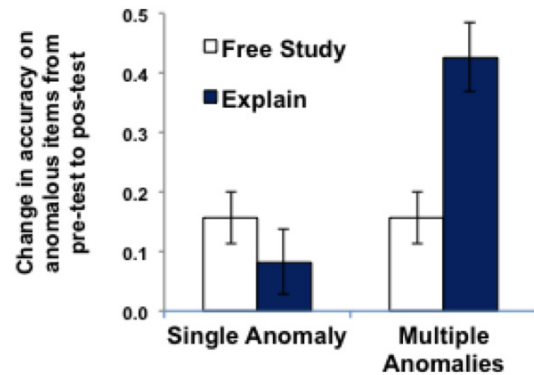


Figure 4: Change from pre-test to post-test in accuracy on *anomalous* items, by learning task and number of anomalies.

When there were multiple anomalies, participants prompted to explain showed greater learning (relative to free study) about the relative-to-deviation principle,  $t(126) = 2.44, p < 0.05$ . But when only a single anomaly was present, there was no significant effect of explanation,  $t(143) = 1.13, p = 0.26$ .

In sum, although explaining helped participants learn the challenging relative-to-deviation principle over more intuitive alternatives, this benefit only exceeded that observed in the control condition when many anomalies were explained. It could be that the effects of explaining one anomaly were too small to yield a statistically significant effect. But a more intriguing possibility – suggested by the (non-significant) trend for control participants to show greater learning gains than explain participants in the single anomaly condition (see Figure 4) – is that explaining actually hindered belief revision by encouraging participants to discount the single anomalous observation.

It is worth emphasizing that the learning benefits observed in the *explain – many anomalies* condition cannot be attributed simply to the effects of receiving more anomalies. Receiving multiple anomalies (relative to observing a single anomaly) promoted greater learning when participants were prompted to explain,  $t(135) = 2.24$ ,  $p < 0.05$ . However, in the free study control condition, observing a greater number of anomalies did not produce a significant learning benefit,  $t(134) = 0.55$ ,  $p = 0.58$ . Without explaining, the potential learning benefits of anomalous information may not be realized.

## Discussion

In an experiment involving statistical judgments, we found that participants who were prompted to explain observations engaged in greater belief revision than participants who engaged in a control task matched for time, attention, and use of language. Specifically, participants who explained showed a greater increase from pre- to post-test in their use of a normative principle for ranking students' performance across courses (the relative-to-deviation principle). However, this benefit was only observed when the observations that participants explained involved multiple anomalies – observations inconsistent with non-normative principles that were arguably more intuitive and more consistent with prior beliefs. When a single anomaly was presented, participants who explained showed comparable or (nonsignificantly) less learning than those in the control condition.

In the introduction we presented several plausible hypotheses about the effects of explaining anomalies. One possibility was that explaining anomalous observations would lead learners to consult current beliefs in making sense of the unanticipated observation, and therefore be more likely to preserve their current beliefs by somehow explaining away or discounting the anomaly (Chinn & Brewer, 1993). In the present experiment, for example, participants could have invoked a clerical error to explain an unanticipated observation, or generated reasons for a ranking that went beyond the information provided (e.g., perhaps a given course was especially difficult and therefore taken by students who were already high achievers).

If explaining encouraged participants to engage in such belief-preserving strategies independently of the number of anomalous observations, then pre- to post-test performance

should have increased more for participants in the control condition than for those in the explain condition. While there was a trend in this direction when a single anomaly was presented, the findings do not provide clear support for this hypothesis.

Explanation-induced failures to revise belief in light of anomalies could be more likely in contexts where participants hold stronger prior beliefs. In fact, Williams, Lombrozo, and Rehder (2010) report an “explanation impairment effect” along these lines. It should be noted, however, that under some conditions maintaining current beliefs in the face of anomalous observations could be the correct or rational strategy. For example, when observations are erroneous or generated by probabilistic processes it could be preferable for learners to discount anomalous observations on the basis of (more accurate) prior beliefs.

A second hypothesis that we considered at the outset was that explaining anomalous observations would increase belief revision, perhaps by drawing attention to anomalies or forcing participants to identify patterns that would account for the anomalous observations and past observations in a unified way. While we found support for this hypothesis when many anomalies were presented, explaining did not have a measurable advantage when participants only observed a single anomaly.

Our findings are therefore consistent with a third possibility: that the effects of explanation interact with the number of anomalous observations. The *number* of anomalies per se may not be crucial, but rather serve (in the current experiments) as an indication of the strength of the evidence that current beliefs require revision. It could be that explaining *few* anomalous observations has no effect (relative to control), encourages belief-preserving strategies, or has variable effects across participants, while explaining multiple (or more problematic) anomalies more uniformly increases the extent to which participants revise beliefs to achieve consistency with observations.

It's also noteworthy that in the absence of explanation, encountering additional anomalies was insufficient to increase belief revision: In the free study condition, observing multiple anomalies (80% of observations) did not yield any significant learning benefit beyond observing just one. Chinn and Brewer (1993) point out that anomalies do not always lead to changes in belief given the number of belief-preserving strategies available to learners, and our findings are consistent with this observation. Explaining could therefore be especially valuable as a strategy for ensuring that learners benefit from anomalous observations, especially in pedagogical contexts in which anomalies are likely to highlight misconceptions and point to normative alternatives.

In previous work we have proposed a *subsumptive constraints* account of the effects of explanation on learning (Williams & Lombrozo, 2010), and we interpret the current results as broadly consistent with this proposal. According to this account, explaining does not provide a general boost to processing, but rather exerts a selective constraint to

interpret what is being explained as an instance of a broader pattern or generalization. One substantiated prediction of this account is that explaining guides people towards patterns that apply to more observations – that is, those that render fewer observations anomalous (Williams & Lombrozo, 2010). A second is that explaining increases learners' consultation of prior beliefs to privilege patterns that prior beliefs suggest will generalize to other contexts (Walker et al., under review; Williams & Lombrozo, 2010b; Williams & Lombrozo, under review).

In the current experiment, the correct relative-to-deviation principle involved fewer (zero) anomalies, but the alternative principles were more consistent with most people's prior beliefs concerning ranking. Explaining could therefore have favored the relative-to-deviation principle in the multiple anomaly condition because the evidence indicated that current beliefs were problematic. In contrast, participants in the single anomaly condition – depending on the strength of their prior beliefs – could have been inclined to favor current beliefs or to consider revision in the face of weak evidence for an alternative. These observations raise a number of important questions for future research concerning the precise conditions under which explaining anomalies will revise or entrench current beliefs.

### Acknowledgments

JJW was supported by a Natural Sciences and Engineering Research Council of Canada fellowship. This research was also supported by an NSF CAREER grant awarded to TL. (DRL-1056712). We thank Vanessa Ing and Sean Trott for helping with analyzing data and piloting the experiment. For helpful discussions about the design of the experiment we thank Cathy Chase, Daniel Benenky, Liz Ritchie, Timothy Nokes-Malach and other members of his lab.

### References

- Ahn, W., Brewer, W. F., & Mooney, R. J. (1992). Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 18(2), 391-412.
- Belenky, D. M., & Nokes-Malach, T. J. (in press). Motivation and transfer: The role of mastery-approach goals in preparation for future learning. *The Journal of the Learning Sciences*.
- Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science education. *Review of Educational Research*, 63, 1-49.
- Chi, M.T.H., de Leeuw, N., Chiu, M.H., LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439-477.
- Fonseca, B. & Chi, M.T.H. (2011). The self-explanation effect: A constructive learning activity. In Mayer, R. & Alexander, P. (Eds.), *The Handbook of Research on Learning and Instruction* (pp. 270-321). New York, USA: Routledge Press.
- Legare, C.H., Gelman, S.A., & Wellman, H.M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development*, 81, 92-944.
- Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. The MIT Press.
- Kuhn, T. (1962). *The structure of scientific revolutions* (3rd ed). Chicago, IL: University of Chicago Press.
- Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10, 464-470.
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11), 2098-2109.
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867-872.
- Renkl, A. (1997). Learning from Worked-Out Examples: A Study on Individual Differences. *Cognitive Science*, 21(1), 1-29.
- Schwartz, D.L., & Martin, T. (2004). Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction. *Cognition and Instruction*, 22(2), 129-184.
- Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31-58). New York: Cambridge University.
- Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition*, 52(1), 1-21.
- Wellman, H. M., & Liu, D. (2007). Causal reasoning as informed by the early development of explanations. *Causal learning: psychology, philosophy, and computation*, 261-279.
- Walker, C.M., Williams, J.J., Lombrozo, T., & Gopnik, A. The role of explanation in children's causal learning. Manuscript under review.
- Williams, J.J. & Lombrozo, T. (2010). The role of explanation in discovery and generalization: evidence from category learning. *Cognitive Science*, 34, 776-806.
- Williams, J.J., & Lombrozo, T. (2010). Explanation constrains learning, and prior knowledge constrains explanation. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2912-2917). Austin, TX: Cognitive Science Society.
- Williams, J.J., Lombrozo, T., & Rehder, B. (2010). Why does explaining help learning? Insight from an explanation impairment effect. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2906-2911). Austin, TX: Cognitive Science Society.

# Olfaction in a hunter-gatherer society: Insights from language and culture

**Ewelina Wnuk (ewelina.wnuk@mpi.nl)**

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands  
International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands

**Asifa Majid (asifa.majid@mpi.nl)**

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands  
Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

## Abstract

According to a widely-held view among various scholars, olfaction is inferior to other human senses. It is also believed by many that languages do not have words for describing smells. Data collected among the Maniq, a small population of nomadic foragers in southern Thailand, challenge the above claims and point to a great linguistic and cultural elaboration of odor. This article presents evidence of the importance of olfaction in indigenous rituals and beliefs, as well as in the lexicon. The results demonstrate the richness and complexity of the domain of smell in Maniq society and thereby challenge the universal paucity of olfactory terms and insignificance of olfaction for humans.

**Keywords:** olfaction; language of perception; smell terms; Maniq; Aslian.

## Introduction

For centuries, great thinkers and scientists have underestimated the sense of smell in humans. Olfaction is often singled out as the least useful perceptual sense, whose role in life is negligible. “Of all the senses it is the one which appears to contribute least to the cognitions of the human mind” (Condillac, 1754/1930, p. xxxi). Darwin (1874) deemed it to be “of extremely slight service” (p. 17), while to Kant (1798/2006) it appeared as “the most dispensable” (p. 50) of the senses. It has also been claimed that olfaction is of “little special value across cultures” (Gardner, 1993, p. 61) and that man “has left the world of smells” (Burton, 1976, p. 109). Neuroscientists have expressed the belief that smell is insignificant for humans and that, in fact, it is “extremely rudimentary” (Grinker, 1934, p. 313), vestigial (Pinker, 1997), or as Stanley-Jones (1957) phrased it, the human rhinencephalon is “untenanted” (p. 594).

Hand in hand with these ideas came the popularization of the belief that olfactory language is impoverished. Dan Sperber (1974/1975), a co-author of the cognitive approach to communication known as Relevance Theory, wrote:

Even though the human sense of smell can distinguish hundreds of thousands of smells and in this regard is comparable to sight or hearing, in none of the world’s languages does there seem to be a classification of smells comparable, for example, to colour classification... There is no semantic field of smells. (pp. 115–116)

According to Henning (1916), “olfactory abstraction is impossible” (p. 66), while Kant (1798/2006) remarks on a

margin of his manuscript: “Smell does not allow itself to be described, but only compared through similarity with another sense” (p. 51).

In spite of the fact that smell is either devalued or ignored in the accounts of many fields of science, there is a growing body of literature which attempts to bring to the fore the importance of smell across cultures (e.g. Classen, Howes, & Synnott, 1994). However, to date there are relatively few studies providing detailed descriptions of olfactory vocabularies in various languages. The current article is intended as a contribution to filling that gap by providing a description of the olfactory lexicon in the language of the Maniq, a group of nomadic hunter-gatherers living in southern Thailand. At the same time, it adds to the knowledge on olfaction of the larger linguistic group of Aslian (belonging to the Austroasiatic family), which is a locus of considerable olfactory elaboration in the cultural and linguistic realm (Burenhult & Majid, 2011). The Maniq data challenges the view that olfaction is of little value to humans as well as the idea that olfactory lexica are necessarily impoverished and lacking in abstract terms. This is important evidence, since the generalizations cited earlier are made primarily on the basis of WEIRD (Western, Educated, Industrialized, Rich, Democratic) communities (Henrich, Heine, & Norenzayan, 2010) and we know that even apparently basic processes such as visual perception may vary across populations (Segall, Campbell, & Herskovits, 1966).

In order to give as comprehensive account as possible of the complex domain of smell in the Maniq language and culture, the topic was explored with the use of multiple methods: ethnographic observation and interview, linguistic elicitation and experimentation. We begin by providing a cultural background to the role of olfaction in the beliefs and practices of the group. We then go on to discuss Maniq smell terminology and, finally, turn to the analysis of speakers’ similarity judgments of Maniq smell terms with the use of multidimensional scaling.

## The Maniq and their Language

Maniq [ma’niʔ] is spoken by 240–300 people living in scattered groups in the Banthad mountain range of southern Thailand (more specifically, in Trang, Satun and Phatthalung provinces). Maniq people belong to the larger ethnographic cluster of Semang with a traditionally mobile lifestyle and hunter-gatherer mode of subsistence. Despite

on-going deforestation and pressure towards sedentism, many Maniq are still nomadic and continue to hunt animals and forage wild plants. Their economy is further supported by income from tourists and small-scale exchange of forest produce. Maniq language belongs to the Northern Aslian branch of Aslian, which forms part of the Austroasiatic family.

### Indigenous Beliefs and Practices

Smell offers a heuristic method of making judgments about odor-emitting sources. Whether to approach something or stay clear of it might depend on the olfactory input one gets from the environment. This function of olfaction is said to be basic for all humans and is believed to be tightly related to the fact that we perceive and categorize smells according to their pleasantness (Yeshurun & Sobel, 2010).

There are numerous examples of odors in the Maniq world which have very clear associations in terms of whether their source is desirable or not. The Maniq make constant use of this information in a variety of contexts – ranging from everyday foraging activities, through indigenous medicine to the ritualized uses of scents. This section explores a number of instances which reflect the relevance of odor in the beliefs and practices of the tribe.

A large number of medicinal herbs collected by the Maniq have intense aromas, the majority of which can be described by the term *lspas* ‘to be fragrant’. Exemplars include: *kasay* ‘*Dianella ensifolia*’, *kupit* ‘turmeric (*Curcuma domestica*)’, *biha* ‘*Triomma malaccensis*’ and *p<sup>h</sup>ley* ‘Cassumunar ginger (*Zingiber montanum*)’ (from Thai *phlai*). The fact that pleasing odors and healing or disease-preventive powers come together in a sizeable group of plants seems to be perceived not as a coincidence, but instead speakers perceive a causal relationship between them. A Maniq woman asked whether a rhizome of the cassumunar ginger she wore in a necklace protected her against illness answered affirmatively adding *ʔeʔ lspas* ‘it is fragrant’. This links to an idea found among the closely-related Aslian groups (e.g. Jahai, Batek), namely that on some occasions odor is believed to be the curing agent of medicine (Burenhult & Majid, 2011; Endicott, 1979).

The beneficial properties of aromatic herbs extend beyond health-protective talismans such as necklaces, headbands and wristbands. A good example of this is the plant called *kasay* ‘*Dianella ensifolia*’, whose roots are ‘burnt in fire during windstorms’ (*tət buwaʔ ʔeʔ hayhəy*) in order to appease the wind (cf. the use of *kasay* and other fragrant plants during thunderstorms among some Aslian groups; Dallos, 2011; Endicott, 1979). At the same time, *kasay* is a multi-purpose medicine which apart from being boiled in water and used to treat stomachache is also burnt in fire in order to produce smoke to be inhaled by the sick.<sup>1</sup>

The Maniq do not offer detailed explanations of how

smoke counters disease or wind, but considering how frequently the term *lspas* is mentioned in such contexts one can be confident that smell plays an important role. A valuable insight into understanding these practices can be gained from the description of a similar act (blowing incense smoke on the body of a sick person) performed by the Batek:

The smoke is supposed to enter the body and cause the disease to flee. This is because the odour of the smoke is good (*bed’ət*) and that of the disease bad (*jebéc*), and they cannot mix. If the smoke goes in, the disease must leave. Alternatively, some say the good-smelling smoke draws the disease out of the body by attracting it, causing it to follow the smoke as it wafts upward from the patient's body. (Endicott, 1979, pp. 107-108)

Another situation in which a good-smelling smoke is used for fighting against a bad force is perhaps one of the most salient and common Maniq rituals of ‘burning animal hair’<sup>2</sup> (*tət sək ʔay*) and ‘bones’ (*ʔiyen*). It is performed on hot days when the sun has a yellow color and when it releases the characteristic smell *hamis*<sup>3</sup>. *Hamis* descends into the forest and spreads around causing illness among the people. Maniq, like other Semang (Benjamin, 1985; Endicott, 1979; Lye, 2004), believe that a cool and shaded environment is healthy and provides protection against disease. Exposed locations without too many trees, on the other hand, are dangerous since the sun heats people’s bodies and turns their eyes red. At such moments, the shelters provide safe refuge from both the heat as well as the malicious *hamis*. Burnt animal hair and bones give off a pleasant smell, *cajes*, which together with the smoke floats up to the sun and neutralizes *hamis*. This belief is a vivid illustration of how much power is attributed to odor in that it can have direct physical effects on the human body and the sun. Yet, people can actively defend themselves by releasing good odors thus forcing out harmful ones and bringing a balance to their immediate environment.

Smells are held as projections of their sources which can directly affect the human body or the environment. As Classen (1993) puts it noting the same phenomenon in a number of cultures across the world, “Involved here is the notion of odor as ‘essence’, containing the intrinsic identity of its source of origin” (p. 99). By this token, according to the Maniq viewpoint, invasive and dangerous odors constitute danger while benevolent ones can be employed as cures and defense mechanisms.

### Language of Olfaction

The cultural importance of olfaction is accompanied by a remarkably complex set of odor distinctions in the language.

<sup>1</sup> The latter practice was used by a man whose condition (immobilizing pain in the legs) was attributed to the ‘soil spirit’ (*tames tieʔ*).

<sup>2</sup> Burning hair (though in this case, it is human hair) is reported to be another thunderstorm-appeasing practice among the Batek, Lanoh and Temiar (Dallos, 2011; Endicott, 1979).

<sup>3</sup> The Batek, too, believe that the sun has an unpleasant odor (*pel’eng*) (Endicott, 1979).

These provide additional support for the claim that the domain of smell is of special value for Maniq society.

## Smell Terminology

In this section, we discuss the main Maniq smell terms. Twelve of these were elicited in a free naming task using “Sniffin’ Sticks” (Hummel, Sekinger, Wolf, Pauli, & Kobal, 1997), where Maniq speakers described different odor stimuli. Due to space limitations, the results of that task will not be reported here. Three additional terms (*paleŋ*, *caŋə*, *caŋes*) were attested during other language elicitation sessions.

In order to move towards understanding the meaning of the smell terms, an exemplar listing task was conducted with the speakers. In this task, consultants were presented with smell terms, one by one, and asked the question “What smells x?”. Participants were free to list as many exemplars as they wished. The task was carried out in Maniq. Table 1 lists the terms together with their exemplar sources elicited from 8 speakers. Numbers in brackets next to each exemplar indicate the number of consultants who gave that response. Six participants contributed responses to the entire (or almost entire) set of smell terms whereas 2 speakers commented on a limited number of terms while another participant was being interviewed. All data is taken into account, though in situations where one of the speakers repeated the response heard from another speaker, it is counted only once. Most plants were identified with the help of Maneenoon (2001). It was not possible to identify some of the plant and animal species – these are given in square brackets.

Maniq smell terms share a number of semantic properties. First, they are dedicated to describing olfactory sensations rather than being general descriptors applicable across sensory domains.<sup>4</sup> Second, they are abstract, meaning that they do not make direct reference to the source of the smell (like e.g. fruity), but rather denote an odor quality. This quality is often a common feature of a range of diverse objects, though examples of terms associated with essentially one referent seem to occur, too (e.g. *hamis*).

Note that some smell terms seem to have clearly identifiable prototypical sources, e.g. *kameh*, *paleŋ*, while others do not have such salient core exemplars, e.g. *mi?* *paŋtu?*, *mi?* *bayɔ̌ɔ̌*. On the whole, unique listings of exemplars are common, which may, to some extent, be an artifact of the listing task, or the small number of participants. It may also be indicative of a certain amount of subjectivity in the understanding of smell terms, but this is not clear at this point.

Another important aspect of odor terminology in Maniq is its presence in everyday conversation. The smell lexicon does not consist of specialist terms known to a limited group of people, nor is it restricted to particular contexts or

Table 1: Maniq smell terms with their corresponding exemplars. Numbers in brackets indicate the number of consultants who listed that exemplar. Unidentified animal and plant species are given in square brackets.

Term	Exemplars
caŋə	tubers ( <i>Dioscorea</i> spp.) (4), food (2), cooked meat (2), rice (1), pork (1)
caŋes	animal hair (2), burnt animal hair (2), burnt animal fat (1), sun (1)
caŋus	soap (3), washing oneself (2), fruit ( <i>Goniothalamus</i> sp.) (1), leaves (1), [kind of fruit] (1), clothes (1), talcum powder (1), sun (1)
hamis	sun (6), air/smoke coming from the sun (2)
haʔit	rotting animal (4), animal (1), plantain squirrel ( <i>Callosciurus notatus</i> ) (1), Prevost’s squirrel ( <i>Callosciurus prevostii</i> ) (1), [kind of squirrel] (1), bats (1)
kameh	[kind of millipede A] (5), [kind of millipede B] (1), [kind of millipede C] (1), ipoh poison (1), [kind of bat] (1), forest (1)
kamloh	smoke from fire (3), old shelter (1), bathing (1) tubers ( <i>Dioscorea</i> spp.) (3), bearcat ( <i>Arctictis binturong</i> ) (2), new shelter (1), clean and dry clothes (1), fruit ( <i>Ficus chartacea</i> ) (1), forest (1), tree (1), animal (1), food (1), medicine to drink (1)
lspəs	
paleŋ	blood (5), raw meat (1), [kind of plant] (1), searching for tubers (1), sun (1)
paʔɔ̌ɔ̌	pouring/getting water (2), tuber ( <i>Dioscorea daunea</i> ) (2), mud (1), digging tubers in mud (1), cooking muddy tubers (1), wet or dirty clothes (1), rotting bamboo tube (1), soil (1), mushroom (1), petai ( <i>Parkia speciosa</i> ) (1), <i>Parkia timoriana</i> (1), sweat (1), urine (1), old shelter (1)
mi?	old shelter (2), soil (2), shelter (1), mushrooms (1), skin of a dead animal (1), rotten wood (1), bamboo for water (1), rotting leaf (1), head of macaque/leaf monkey (1)
bayɔ̌ɔ̌	
mi?	mushrooms (2), rotten wood (2), rotten mushrooms (1), old shelter (1), animal bones (1), durian seed (1), snakes (1), forest (1), searching for tubers (1), soil (1)
danəw	
mi?	snakes (2), soil (2), searching for tubers (1), digging tubers (1), mushrooms (1), sweat (1), rotten wood (1), walking in the forest (1), making fire (1), smoke (1)
huhũɔ̌ɔ̌	
mi?	soil (2), burning fire (1), [type of fire wood A] (1), [type of fire wood B] (1), [kind of flower] (1), [kind of fruit] (1), mushrooms (1), tree (1), walking in the forest (1)
latiŋ	
mi?	tree sap (1), leaves (1), garlic (1), soil (1), forest (1)
paŋtu?	

<sup>4</sup> The only exception here is the term *bayɔ̌ɔ̌* which, apart from describing smell, refers to color – a specific kind of white, e.g. of fog or old individual’s hair.

registers of speech. Talking about smell is a mundane activity which all members of the community engage in on a daily basis. What is more, smell is an important reference point in a number of areas of life, such as medicinal practices and rituals.

Formally, Maniq smell terms can be subdivided into stative verbs and noun phrases. The verbs can take verbal affixes, though most frequently they do not bear any morphology (excluding the apparent frozen morphology in *ls-pəs*, where the initial half-syllable has the shape of the iterative affix). A few of the verbs, namely *caŋes*, *caŋus* and *caŋə*, apart from being close semantically, are phonologically similar. They do not, however, show evidence of a productive derivational relationship.

As for the noun phrases, all of them are headed by the noun *mi?* ‘smell’. They appear to be lexical noun phrases since none of the modifiers, with the exception of *bayɔ̌p*<sup>5</sup>, occurs outside of the “mi?+...” phrase. For that reason, it is rather difficult to establish the word class affiliation of these modifiers, though elicited aspect-inflected forms for three of them (*danɔw*, *latiŋ* and *ŋətu?*) suggest that they might be verbs.

All of the above terms serve as phenomenon-oriented descriptions. The controlled activity of smelling as well as the uncontrolled experience of perceiving smell are both expressed by the verb *ɔ̌n* ‘to smell’. It is worth noting that the Maniq *ɔ̌n* forms a clear and distinct category uniquely relating to olfaction with no extensions into other sense modalities.

## Organization of the Smell Lexicon

Taking into account the large number of smell terms as well as the considerable range of exemplars associated with most of them, it is unclear what principles might underlie the organization of the smell lexicon. One way of gaining insight into that organization is to investigate the relationships between the smell terms by collecting similarity judgments from speakers, following a similar procedure used to study color lexica (Shepard & Cooper, 1992). Since the Maniq are a non-literate community, this task was carried out with the use of the triadic comparison method, which does not require reading. Following data collection, the results from 11 Maniq speakers were pooled together to create a similarity matrix analyzed with multidimensional scaling analysis.

## Stimuli and Method

Stimuli for the experiment were the 15 smell terms given in Table 1. The experiment was based on the triadic comparison method as discussed by Weller and Romney (1988). A complete triad test with 15 items would involve 455 triads, which would be time-consuming and tiring for

participants. For that reason, we have used a balanced incomplete block design ( $\lambda=1$ ) of 35 triads. The letter  $\lambda$  represents “the number of triads in which each pair of items occurs” (Burton & Nerlove, 1976, p. 249). To increase the reliability of the design without adding more triads, we followed the recommendation of Burton and Nerlove (1976) to administer two different triad compositions, with each composition randomly presented to half of the participants. Smell terms were randomly assigned a number and the triads were created following the directions of Burton and Nerlove. The only modification of Burton and Nerlove’s design was the randomization of the order in which triads were presented – items were randomized within and across triads to avoid frequent repetition of terms in close proximity.

## Participants

The participants were 11 Maniq speakers (6 male, 5 female) aged approximately 20-45 years. All were native speakers of Maniq, who also had a good command of Southern Thai, the unrelated majority language of the region.

## Procedure

Each participant was tested individually. The task was run exclusively in Maniq to preclude the influence of Southern Thai. Speakers were orally presented with 3 smell terms at a time and asked the following question: “Which one is not the same/similar?” (the meaning of the Maniq term *min*, from Thai *měuan* ‘same, similar’, has scope over both sameness and similarity). The response was coded on a response sheet and the next triad was presented until all triads were complete.

In order to ensure that the task was proceeding as intended, a series of precautions were taken. Before starting the task, the researcher informed the participants that they would be presented with words relating to smell. The critical question “Which one is not the same/similar?” was used each time with the initial triads to make sure the participants remembered what they were asked to do. As they became accustomed to the task, the question was repeated every few triads. Three objects (three similar leaves from the same plant) were placed in a row in front of the consultants to act as anchors to the words in the triad. In order to avoid a situation in which a consultant fell into a response set, words were assigned to objects sometimes from right to left and sometimes from left to right. When presenting a triad, target words were pronounced slowly several times with neutral intonation, unless a consultant gave a response immediately after hearing the triad once. Many consultants responded with the following phrases: “These are together” and “This one is alone” or “These are similar” and “This one is not similar”. In case of any uncertainties on the side of the speaker or the researcher, the triad was read out again and the question was repeated. On the rare occasions when the consultant could not make a choice after being asked the question several times, the researcher proceeded to the next triad and came back to the

<sup>5</sup> *Bayɔ̌p* is a stative verb occurring in a variety of syntactic contexts with a number of verbal affixes. Note, however, that whenever the word is used to denote smell, it occurs in the nonderived form.



problematic one at the end. All participants were able to complete the study.

## Results and Discussion

A 15x15 similarity matrix was constructed by summing over all participants the number of times each pair of smell terms was judged similar. The matrix served as input into the scaling procedure carried out with the use of the PROXSCAL algorithm in SPSS. The resulting two-dimensional solution yielded a stress value (Stress-I) of .098, a dispersion-accounted-for value of .99 and a Tucker's coefficient of congruence value of .995. Figure 1 shows the overall results from all 11 participants.

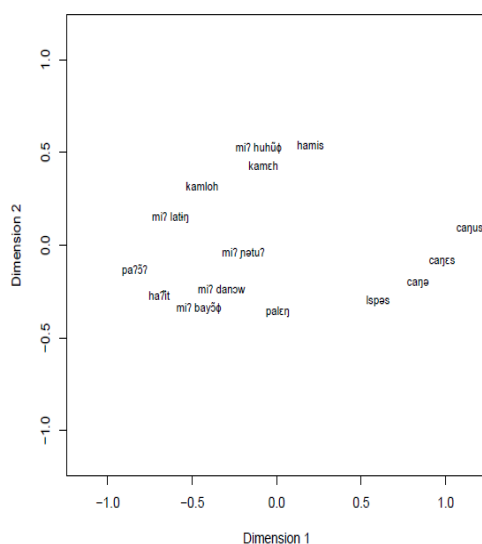


Figure 1: Two-dimensional MDS of 15 Maniq smell terms based on triadic comparison (N=11).

The distribution of smell terms is considerably more stretched on the first dimension. Items are more densely concentrated on the left-hand side, while the right-hand side is more sparsely populated, especially if we look at the almost empty area in the upper right quarter.

The first dimension is readily interpretable as distinguishing between pleasant and unpleasant smell terms, the former located on the right- and the latter on the left-hand side. Though the stimuli were words rather than actual smells, this aspect of the results is comparable with studies using odorants, which report the primary role of the hedonic dimension in smell perception and categorization (e.g. Dubois 2000; Schiffman, Robinson, & Erickson 1977). Pleasantness is also of great importance when considering neurophysiological responses to smells (cf. Yeshurun & Sobel 2010 for an overview) and there is some evidence suggesting this reflects the molecular structure of odorants (Khan et al., 2007).

The interpretation of the second dimension is less straightforward, yet a likely solution is the contrast of edibility vs. inedibility. Items at the bottom are associated

with food objects whereas those at the top are associated with nonfood objects. Again, this aspect appears to reflect odor perception since edibility was found to distinguish odorant samples (Chrea et al., 2004; Zarzo, 2008). An important caveat to this interpretation is that the focus is placed on the smell object rather than the smell quality itself. For instance, the terms *pale* or *ha?it* refer to raw or rotting animal meat, which are not edible within this community. Nevertheless, they refer to meat, which is an edible object.

Since many smell terms relate to multiple exemplars, some of which may be edible while others not, we focused on the smell term prototypes, which for the current purposes are defined as items listed by more than 1 speaker. Inspecting the plot, we see that most terms conform to the edibility distinction. All items in the upper part of the plot relate to exemplars which are considered to be inedible by the Maniq, e.g. *mi? huhu* (snakes and soil), *hamis* (sun and air/smoke coming from the sun), *kameh* (millipedes) and *kamloh* (smoke from fire). As for the opposite side, most items link to edible exemplars, e.g. *pale* (blood), *lspas* (tubers, bearcat), yet there are a few terms among whose prototypical exemplars we find inedible objects – *mi? bay* (old shelter and soil) and *mi? danow* (rotten wood). So, there is evidence consistent with viewing Dimension 2 in terms of edibility but a further examination is required to fully establish the facts.

A follow-up study could collect speakers' judgments on various possible semantic parameters, to see which best predicts the attested dimensions.

## Conclusions

This paper illustrates the richness and complexity of the domain of smell in Maniq society. The different methodologies employed provide insights into the smell lexicon, its underlying structure, and the deep cultural significance of different smell categories. Despite the fact that many cultures, especially those which are part of the developed world, are undergoing gradual deodorization, there appear to be a number of societies with a long tradition of vibrant interest in odor (cf. Classen, Howes, & Synnott 1994). As we hope to have demonstrated with this paper, Maniq adds to the literature regarding the special cultural value of smell across the world, and at the same time reinforces our observation that the Aslian-speaking communities of the Malay Peninsula provide a rewarding setting for studying such smell cultures and their linguistic elaboration of the domain (Burenhult & Majid, 2011). Moreover, it highlights the importance of looking beyond WEIRD people in our theories of cognition (Henrich et al., 2010).

Smell is an integral part of the intimate knowledge of the rainforest's fauna and flora shared by the tribe. What is more, for the Maniq, odor has a metaphysical dimension whereby it is treated as the projection of its source able to "act on its behalf". This is illustrated by the wind-appealing ritual involving the burning of *kasay* as well as the

medicinal practices of the group including both curing and prevention (the best example being the emission of smell to ward off the disease spreading with the odor of sun).

Contrary to the view that the language of odor is non-abstract and steeped in metaphors (Henning, 1916; Kant, 1798/2006; Sperber, 1974/1975), Maniq, and its Aslian brethren, possess rich smell vocabularies of over a dozen abstract terms. These terms are known to the whole speech community and are employed in everyday conversation.

Finally, the internal structure of the Maniq smell lexicon is remarkably similar to the dimensions of variance typically found in studies of odor categorization in speakers without an abstract olfactory lexicon. This suggests that odor lexica may reflect a pan-human olfactory space. Further investigation is needed to explore the extent to which olfactory language follows olfactory perception and cognition, and the extent to which perception and cognition might mirror language in the domain of olfaction.

## References

- Benjamin, G. (1985). In the long term: Three themes in Malayan cultural ecology. In K. L. Hutterer, A. T. Rambo, & G. W. Lovelace (Eds.), *Cultural values and human ecology in Southeast Asia*. Ann Arbor, MI: University of Michigan, Center for South and Southeast Asian Studies.
- Burenhult, N., & Majid, A. (2011). Olfaction in Aslian ideology and language. *The Senses and Society*, 6(1), 19–29.
- Burton, M. L., & Nerlove, S. B. (1976). Balanced designs for triads tests: Two examples from English. *Social Science Research*, 5(3), 247–267.
- Burton, R. (1976). *The language of smell*. London: Routledge & Kegan Paul.
- Chrea, C., Valentin, D., Sulmont-Rossé, C., Ly Mai, H., Hoang Nguyen, D., & Abdi, H. (2004). Culture and odor categorization: Agreement between cultures depends upon the odors. *Food Quality and Preference*, 15(7-8), 669–679.
- Classen, C. (1993). *Worlds of sense: Exploring the senses in history and across cultures*. London: Routledge.
- Classen, C., Howes, D., & Synnott, A. (1994). *Aroma: The cultural history of smell*. London: Routledge.
- Condillac, E. B. de. (1930). *Condillac's treatise on the sensations*. (M. G. S. Carr, Trans.). London: The Favil press. (Original work published 1754).
- Dallos, C. (2011). *From equality to inequality: Social change among newly sedentary Lanoh hunter-gatherer traders of Peninsular Malaysia*. Toronto: University of Toronto Press.
- Darwin, C. (1874). *The descent of man, and selection in relation to sex* (2nd ed.). London: John Murray.
- Dubois, D. (2000). Categories as acts of meaning: The case of categories in olfaction and audition. *Cognitive Science Quarterly*, 1(1), 35–68.
- Endicott, K. M. (1979). *Batek Negrito religion: The world-view and rituals of a hunting and gathering people of Peninsular Malaysia*. Oxford: Clarendon Press.
- Gardner, H. (1993). *Frames of mind: The theory of multiple intelligences*. New York, NY: Basic Books.
- Grinker, R. R. (1934). *Neurology*. Springfield, IL: C.C. Thomas.
- Henning, H. (1916). *Der Geruch*. Leipzig: J. A. Barth.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences*, 33(2-3), 61–83.
- Hummel, T., Sekinger, B., Wolf, S., Pauli, E., & Kobal, G. (1997). 'Sniffin' Sticks': Olfactory performance assessed by the combined testing of odor identification, odor discrimination and olfactory threshold. *Chemical Senses*, 22(1), 39–52.
- Kant, I. (2006). *Anthropology from a pragmatic point of view*. (R. B. Louden, Trans.). Cambridge: Cambridge University Press. (Original work published 1798).
- Khan, R. M., Luk, C.-H., Flinker, A., Aggarwal, A., Lapid, H., Haddad, R., & Sobel, N. (2007). Predicting odor pleasantness from odorant structure: Pleasantness as a reflection of the physical world. *Journal of Neuroscience*, 27(37), 10015–10023. doi:10.1523/JNEUROSCI.1158-07.2007
- Lye, T.-P. (2004). *Changing pathways: Forest degradation and the Batek of Pahang, Malaysia*. Lanham: Lexington Books.
- Maneenoon, K. (2001). *Ethnobotany of Sakai tribe in Trang, Phatthalung and Yala provinces* (Master's thesis). Prince of Songkla University, Hat Yai.
- Pinker, S. (1997). *How the mind works*. New York: Norton.
- Schiffman, S., Robinson, D. E., & Erickson, R. P. (1977). Multidimensional scaling of odorants: Examination of psychological and physicochemical dimensions. *Chemical Senses*, 2(3), 375–390.
- Segall, M. H., Campbell, D. T., & Herskovits, M. J. (1966). *The influence of culture on visual perception*. Indianapolis, IN: Bobbs-Merrill Co.
- Shepard, R. N., & Cooper, L. A. (1992). Representation of colors in the blind, color-blind, and normally sighted. *Psychological Science*, 3(2), 97–104.
- Sperber, D. (1975). *Rethinking symbolism*. (A. L. Morton, Trans.). Cambridge: Cambridge University Press. (Original work published 1974).
- Stanley-Jones, D. (1957). Posture and the rhinencephalon: A new interpretation. *The Journal of Nervous and Mental Disease*, 125(4), 591–598.
- Weller, S. C., & Romney, A. K. (1988). *Systematic data collection*. Newbury Park, CA: Sage Publications.
- Yeshurun, Y., & Sobel, N. (2010). An odor is not worth a thousand words: From multidimensional odors to unidimensional odor objects. *Annual Review of Psychology*, 61, 219–241.
- Zarzo, M. (2008). Psychologic dimensions in the perception of everyday odors: Pleasantness and edibility. *Journal of Sensory Studies*, 23(3), 354–376.

# Learning (to Learn) from Spatial Attention Cues During Infancy

## Winner of a Robert J. Glushko Dissertation/Thesis Prize in Cognitive Science

Rachel Wu (r.wu@bbk.ac.uk) and Natasha Z. Kirkham (n.kirkham@bbk.ac.uk)

Centre for Brain and Cognitive Development, Department of Psychological Sciences, Birkbeck, University of London  
Malet Street, London, WC1E 7HX, UK

### Abstract

Human infants develop a variety of attentional mechanisms that allow them to extract relevant information from a cluttered world. We know that both social and non-social cues shift infants' attention, but not how infants use these cues to learn basic events. With over 450 infants, four extensive eye-tracking studies in this thesis established a controlled paradigm for investigating how attention cues shape early learning. The results showed that infants' ability to learn about structures in their environment (i.e., predicting the appearance of audio-visual events and forming expectations about co-occurring features) is dependent on the presence and nature of attention cues. By 8 months, infants learned these events significantly better with social cues (e.g., eye gaze, infant-directed speech, expression of interest) than with non-social cues (e.g., flashing squares) or without any attentional cueing. Importantly, when presented with multiple events to learn and cued by a face to one specific event, infants learned the cued event and ignored the non-cued event. The last study found that familiar communicative social signals (i.e., an engaging face that spoke to the infant) boosted 9-month-olds' learning about cued events. In particular, the engaging face supported learning from non-social cues, providing evidence for a mechanism explaining how infants learn to learn from unfamiliar attention cues such as pointing or arrows. Our results showed that though social cues may temporarily detract attention away from certain learning events in the world, they appear to stimulate infants to display better learning about the cued event than when infants learn with other attention cues or on their own without attention cues.

**Keywords:** attention cues; pattern learning; infant eye-tracking; cognitive development; social cues.

### Introduction

The relationship between attention and learning is reciprocal: learning is enabled and facilitated by attentional selection, and attentional selection builds on previously learned knowledge. For example, studies with adults have shown that appropriate attentional selection can help filter distracters to learn about targets (e.g., Hillyard, Hink, Schwent, & Pinckton, 1973; Van Voorhis & Hillyard, 1977) and that learned knowledge about targets and distracters allows more efficient target selection among distracters (e.g., Chun & Jiang, 1998; Kruschke, 2011; Mazza, Turatto, Umiltà, & Eimer, 2007). This bi-directional relationship between attention and learning is highlighted in the developmental perspective, when considering such abilities in inexperienced learners (i.e., infants).

Human infants have to learn a great deal of information in a complex world that is filled with ambiguity. Not only are

many different features and dimensions present in the environment, but also they are often unrelated to any reinforcement or feedback. Selective attention, focusing on some information but ignoring other aspects, is a crucial component for learning, especially in situations involving uncertainty (e.g., Dayan, Kakade, & Montague, 2000). There are two broad solutions to this complexity and ambiguity inherent in the learning environment: (a) innate constraints on the cues selected for processing, or (b) rapid learning-to-learn mechanisms that assess cue-reliability and thus guide learning. In the bottom-up solution, there are environmental and innate biases that constrain what infants attend to (e.g., Brazelton, Scholl, & Robey, 1966; Johnson, Dziurawiec, Ellisc & Morton, 1991), and in the top-down solution, infants can sample their environment and quickly learn what sources of information make the most sense (e.g., Saffran, Aslin, & Newport, 1996; Fiser & Aslin, 2002). Mechanisms of learned information selection are generally considered top-down and thus may be tuned to specific task demands. Many experiments with older children and adults show that top-down selection produces efficient learning and allows for pre-activation of relevant neural circuits, which may lead to better processing of the stimulus (e.g., Driver & Frith, 2000). Would infants learn differently with different types of attention cues?

Both bottom-up and top-down (learned) attention cues shift infants' attention: social stimuli (e.g., eye gaze, infant-directed speech, faces), which infants learn to follow by four months of age, and non-social stimuli (e.g., bright lights/colors, motion), which capture infants' attention from birth. In following these stimuli, infants can develop attentional mechanisms for extracting relevant information from a cluttered multimodal world, because these stimuli can cue infants to the location of relevant events. In turn, relying on these attention cues can shape learning because what controls attention can gate processing (Moran & Desimone, 1985). Though it is well known that both social and non-social cues shift infants' attention, it is unclear how these attention cues differentially affect learning of basic events. We know that infants are capable of learning about the structure of their environment in the simplified laboratory setting. For example, infants recognize that certain sights and sounds belong together (e.g., toys dropping, cars driving, people talking, Bahrick, 2004; Lewkowicz, 2000) and track the co-occurrence of visual features (e.g., Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002). In the typical cluttered environment, however, infants are often presented with multiple events to

learn. How do infants know which events are important to learn? Would different attention cues similarly help infants select information to learn?

## Summary of Thesis Studies

The tight coupling of attention and learning, especially for the young learner, determines how information is selected, retained, and eventually applied. This thesis capitalized on recent advances in methodologies (i.e., eye-tracking) that allow for paradigms that include more ecologically valid environments with noise and distraction testing young age groups.

Learning from cues (specifically attention cues) is an ideal context for studying this interaction between selective attention and learning efficacy. The thesis investigated whether infants learn differently from different attention cues (or no cues), and how infants learn to learn from attention cues. In a distraction-filled environment, visual spatial attention cues (e.g., other's eye gaze or flashing lights) can highlight events in a particular location and facilitate processing of those events (see Posner, 1980). This is a critical aspect for the young learner, who may not know what to learn at a given moment. The components of a spatial cueing experimental paradigm include a cue, target, and distracter(s). Commonly, the cues shift attention to a particular location to prepare the viewer for the target event in that location. Cues can either attract attention (bright flashing lights, big moving objects) or shift attention from themselves to another location (learned attention cues: eye gaze, arrows).

My 3-year PhD investigated how different attention cues (or no cues) affect learning during infancy. My studies showed for the first time in one cohesive paradigm that the presence and nature of these attention cues mediate what infants learn about the structures in the environment (i.e., predicting the appearance of audio-visual events or forming expectations about co-occurring features). A sample of over 450 infants across four extensive eye-tracking studies (Wu & Kirkham, 2010; Wu, Gopnik, et al., 2011; Wu, Kirkham, et al., under revision) showed that by 8 months, infants learned the structures in their environment significantly better with subtle social cues (e.g., eye gaze, infant-directed speech) than with salient non-social cues (e.g., flashing squares) or without any attentional cueing. Importantly, when presented with multiple events to learn and cued by a face (rather than a flashing square) to one specific event, infants learned the cued event rather than the non-cued event. These results show that when naïve learners do not know what to learn, social objects (faces) shape the likelihood of learning cued targets, and that attending to social cues provides infants with an optimal strategy for learning appropriate events despite the presence of distractions.

**Study 1** In the first study, Experiment 1 used social cues to direct 8- and 4-month-old infants' attention to multimodal

events (i.e., animations of toys accompanied by specific sounds), while identical distracter events were presented in another location. Experiment 2 directed 8-month-olds' attention with colorful flashes to the same events. Experiment 3 measured 8-month-olds' baseline learning without attention cues. The test trials in all experiments played only the sounds previously associated with a particular animation, and looking time was measured to each now blank location that previously contained an object. Results showed that the 8-month-olds exposed to social cues learned about the cued audiovisual event (i.e., they predicted its appearance in the correct rather than incorrect cued location) (Figure 1). The 4-month-olds, however, displayed only general spatial learning from social cues (i.e., they looked to both correct and incorrect cued locations to predict its appearance), suggesting that infants *learn to learn* from a face stimulus between 4 to 8 months. Eight-month-old infants cued with the colorful flashes looked indiscriminately to both correct and incorrect cued locations during test (similar to the 4-month-olds learning from social cues) despite attending for equal duration to the training events as the 8-month-olds with social cues. Results from Experiment 3 (no learning) indicated that the learning effects from Experiments 1 and 2 resulted from exposure to the different cues and multimodal events. In summary, this first series of experiments shows that infants' attention to target events is captured equally well by both social and non-social cues, but learning is deeper and more precise with social cues. This study is published in the *Journal of Experimental Child Psychology* (Wu & Kirkham, 2010).

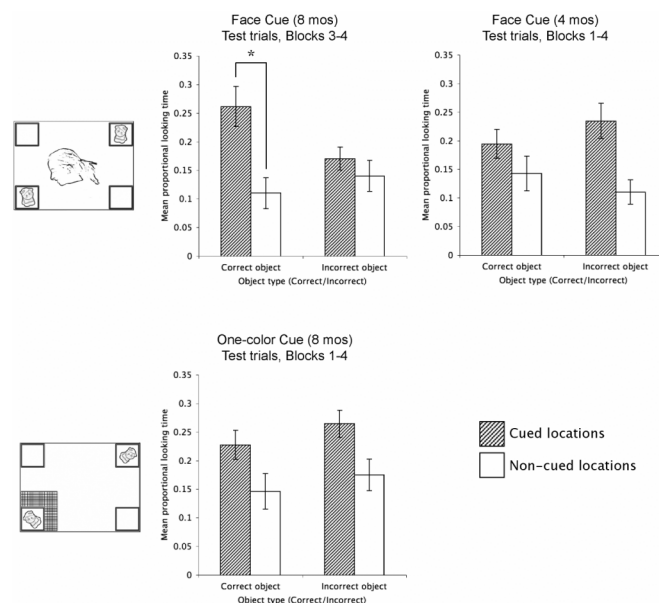


Figure 1: Stimuli and results from Study 1. Mean proportional looking times to locations during test trials from Blocks 3 and 4 in the Face Cue condition (8 months) and Blocks 1 to 4 in the Face Cue condition (4 months) and One-color Cue condition (8 months). The 8-month-olds in the Face Cue condition looked more to the cued correct object location, whereas the 4-month-olds in the same

condition looked longer to cued locations than to non-cued locations regardless of object–sound mappings. The 8-month-olds in the One-color Cue condition (flashing squares) also looked longer only to cued locations than to non-cued locations regardless of multimodal information.

\* $p=.01$

**Study 2** Study 2 used visual statistical learning as the dependent measure, and compared how infants learn from social cues to that with no cues. In laboratory experiments, infants can learn patterns of co-occurring visual features (e.g., Fiser & Aslin, 2002). Once infants learn the statistical regularities, however, how do they use the knowledge? In addition, which patterns do infants learn when presented with many options (as in the cluttered world outside of the laboratory)? In this study, infants were shown clusters of 3 shapes, where two always co-occurred and a third changed on every trial. Infants were either shown these shape patterns on their own (Experiments 1 and 2) or shown the pattern cued by a face (Experiment 3) or also with a distracter pattern in a non-cued location (Experiment 4). Test trials displayed shapes moving apart: either the co-occurring shapes (looking longer related to a preference for the inconsistent/violation of expectation) or the non-co-occurring shapes (looking longer related to a preference for the consistent). Tracking co-occurring units and inferring that they are larger fused units help identify integral object parts for object individuation, recognition, and categorization. Experiment 1 showed that 9-month-old infants interpret co-occurring features as larger fused units (i.e., infants looked longer when co-occurring features split apart). The other three experiments showed that social cues (compared to no cues) help 9-month-olds choose patterns among distracters during learning and test (Figure 2). These findings suggest that by 9 months, infants can use feature co-occurrence to learn about objects and that social cues shape such foundational learning in distraction-filled environments. In particular, though social cues may temporarily detract attention away from certain learning events in the world, they appear to stimulate infants to display the learning better in complex situations than when infants learn on their own without social cues. Task difficulty also mediates how inferences (made from visual statistical processing) are exhibited during test trials. This study is published in *Developmental Psychology* (Wu, Gopnik, et al., 2011).

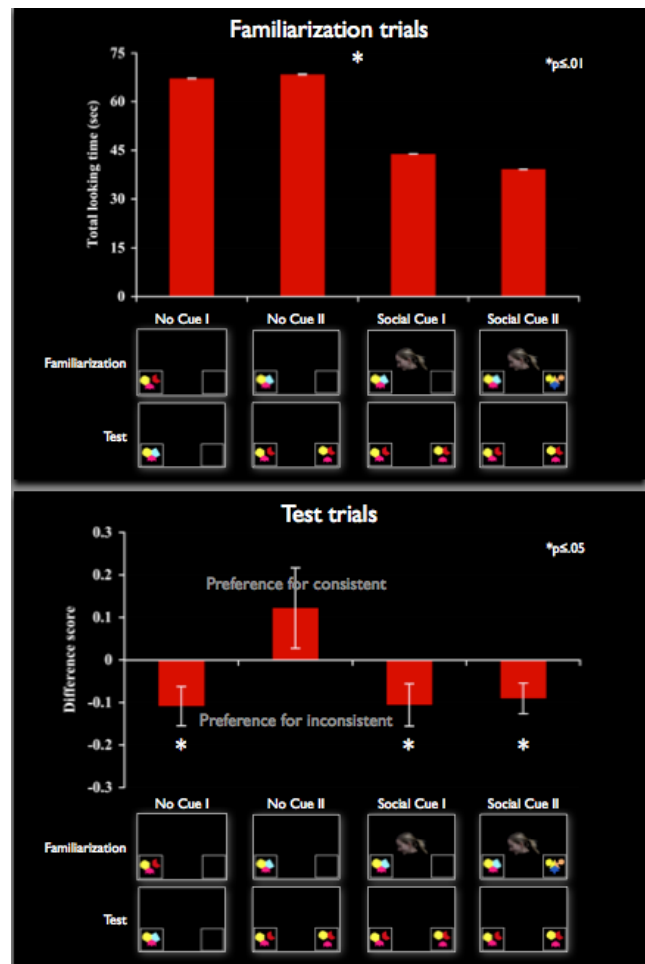


Figure 2: Stimuli and results from Study 2. The top half of the figure shows the total looking times to the target pattern for the familiarization trials across the four conditions (No Cue I, No Cue II, Social Cue I, and Social Cue II). Infants looked longer to the target pattern in the two No Cue conditions, and less in the two Social Cue conditions because they split their attention between the face and target pattern. \* $p\leq.01$  The bottom half of the figure displays the difference scores across four conditions (mean difference between proportional looking times for consistent minus inconsistent events during test). A negative value reflected a preference for the inconsistent splits (i.e., looking longer at events showing the separation of features that co-occurred rather than the separation of features that did not co-occur). Infants showed this preference in the easier test condition without a cue (No Cue 1), and with a social cue regardless of task difficulty. \* $p\leq.05$

**Study 3** Following up from these findings, that infants learn better from social than non-social cues, Study 3 investigated whether infants can *learn to learn* from typical non-social objects that are interactive like social cues (e.g., objects that move when infants look at it, e.g., Johnson, Slaughter, & Carey, 1998). The main procedure consisted of three phases: training, familiarization, and test. During the training phase, 9-month-olds interacted with a centrally presented teapot

(similar to Deligianni, Senju, Gergeley, & Csibra, 2011, who found that infants follow such interactive objects). Infants' fixations on the teapot caused it to jump or lift its lid. During the familiarization and test phases, infants were presented with the pattern events from Study 2, with the original face now replaced by the teapot. We found that if infants followed the teapot during familiarization, they seemed to have trouble learning about the cued event, only perhaps learning the cued location (general spatial learning similar to 4-month-olds cued by a face in Study 1). Infants who did not follow the teapot (and perhaps ignored this cue) seemed to learn about the non-cued event. While only preliminary observations can be made at this stage, this pilot study begins to address *how* infants learn to use cues, rather than describing cues they do or do not learn from.

**Study 4** In Study 1, the flashing square cue was unsuccessful at producing specific learning in infants (Wu & Kirkham, 2010). Study 4 investigated whether the pairing of familiar social cues with unfamiliar flashing cues could help infants learn from these novel flashing cues. Nine-month-old infants were eye-tracked during a Training phase, followed by a Generalization phase. In the Social Scaffolding condition, the Training phase consisted of an expressive face that spoke to the infant and then froze with a smile, while identical audio-visual animations appeared in diagonally opposite corners. At the same time, a red flashing square cued the infant to a specific target frame containing an object. In the Extended Practice condition, infants only saw the flashing square and multimodal events. The Generalization phase for both conditions displayed new audio-visual events with only the red flashing square cues. The test trials in each phase played one of the sounds previously associated with a particular animation, and looking time to each location was measured. In the Social Scaffolding condition, infants anticipated that the events would appear in the correct cued locations for both phases. In the Extended Practice condition, infants first displayed general spatial learning (replicating Experiment 2 in Study 1) during the Training phase, and then showed no learning during the Generalization phase (Figure 3). These findings suggest that initial exposure to familiar social cues can elicit and maintain specific learning from novel attention-orienting cues. Moreover, this could provide evidence for a mechanism explaining *how* infants and children can learn to learn from distal attention cues such as pointing fingers and arrows. This is an important first step towards elucidating an emerging ability to use familiar attention cues to support, enhance, and mediate learning about unfamiliar cues, going beyond documenting *which* cues guide attention and learning during infancy to proposing a mechanism for *how* this cascading learning effect occurs. Portions of this study are a CogSci proceeding and EuroCogSci proceeding (awarded Best Student Paper Prize), and is currently under revision for a journal submission.

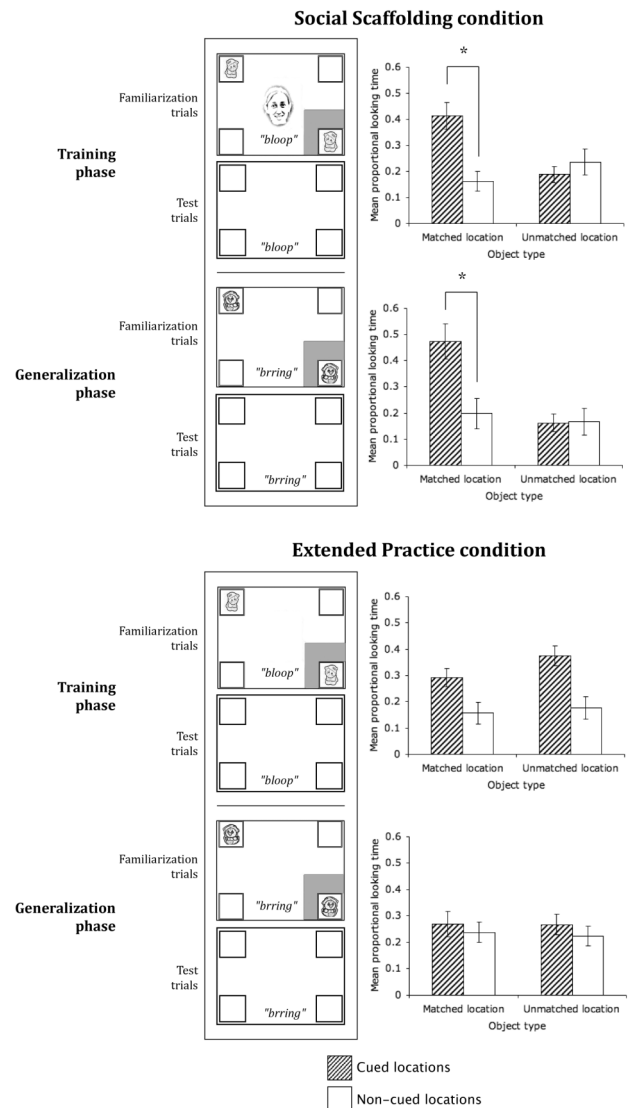


Figure 3: Stimuli and results from familiarization and test trials for both conditions in Study 4. All stimuli were in full color on a black background. When exposed to direct social signals paired with flashing squares, infants predicted objects would appear in matched cued locations (Social Scaffolding condition, Training phase). Infants continued learning specifically from flashing squares even after social signals were no longer presented (Social Scaffolding condition, Generalization phase). With only exposure to flashing squares, infants exhibited general spatial learning and no transfer of learning, suggesting that initial exposure to direct social signals is necessary to elicit and maintain specific learning from novel cues at this age.  $*p < .03$

## Conclusions

By studying how infants learn cued events (and do not learn distracter events) with attention cues, this thesis investigated the usefulness of attentional biases for infants' cognitive development. This research is important because it goes beyond describing singular events that infants can learn by exploring how they figure out *what* to learn in the

noisy environment. Investigating how infants learn to learn with cues provides a more accurate picture of infants' learning mechanisms in the rich natural environment. My thesis work contributed to the study of the optimal dynamics of selective attention and successful learning in typical development, which in turn would inform populations with learning difficulties. Though the studies in this thesis only investigated learning from visual spatial attention cues, research on the overall question of how one learns to learn among distraction is important for many areas of cognitive development. *Learning to learn* is essentially the interplay of gaining control over one's perceptual input based on previous experience (Aslin, 2008; Frank et al., 2009) and current interactions with the environment (Sarter, Givens, & Bruno, 2001). This is one way of describing how one gains expertise (Gopnik, 2009; Nodine, Kundel, Lauver, & Toto, 1996; Solso, 2001). We are trained to detect and learn from certain stimuli, whether it is learning from people (Bigelow, 1998; Csibra, 2010; Ghazanfar & Santos, 2004) or learning from other cues (e.g., arrows). With some rudimentary initial biases, infants *learn to learn* and become more adapted to function appropriately according to environmental norms.

Testament to the interdisciplinary contribution of this thesis, at least seven current projects on computational modelling, robotics, genetics, and atypical development are based directly on this work and dataset (please see CV). The large dataset (450+ infants) from this thesis is ideal for computational modelling and genetics projects, and the benchmark of typical behavior can be used to compare with data from atypical development.

This work also has led directly to the organization of two attention-learning workshops in London (Jan 2012 – organized by R. Wu, received £1818 to organize workshop) and Tokyo (March 2012 – organized by R. Wu, S. Shimojo, and T. Omori, funded by the Tamagawa-CalTech grant), encouraging discussions among computational modellers, developmental psychologists, neuroscientists, and roboticists to promote the emergence of this sub-field on the interaction of attention and learning.

## Acknowledgments

I owe a great deal of gratitude to my supervisors, Natasha Kirkham and Denis Mareschal for their guidance, dedication, and encouragement, as well as to the other collaborators on this project, Teodora Gliga, Alison Gopnik, Daniel Richardson, and Kristen Swan. Thank you to the Glushko-Samuelson Foundation and the Prize committee for the generous award, as well as providing travel grants to present this work at various Cognitive Science meetings. Thank you to Leslie Tucker and Marian Greensmith for help with data collection. This research was supported by a grant to NK from the British Academy, Grant No. SG-47879 and two grants to NK and RW from the University of London Central Research Fund.

## References

- Aslin, R. N. (2008). Headed in the right direction: A commentary on Yoshida and Smith. *Infancy*, 13(3), 275-278.
- Bahrick, L. E. (2004). Development of intermodal perception. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science* (Vol. 2, pp. 614–617). London: Nature Publishing Group.
- Bigelow, A. E. (1998). Infants' sensitivity to familiar imperfect contingencies in social interaction. *Infant Behavior and Development*, 21(1), 149-162.
- Brazelton, T. B., Scholl, M. L., & Robey, J. S. (1966). Visual responses in the newborn. *Pediatrics*, 37, 284-290.
- Chun, M. M., & Jiang, Y. (1998). Contextual cueing: implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, 36(1), 28-71.
- Csibra, G. (2010). Recognizing communicative intentions in infancy. *Mind and Language*, 25(2), 141-168.
- Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, 3 Supplement, 1218-1223.
- Deligianni, F., Senju, A., Gergely, G., & Csibra, G. (2011). Automated gaze-contingent objects elicit orientation following in 8-months-old infants. *Developmental Psychology*, 47, 1499-1503.
- Driver, J., & Frith, C. (2000). Shifting baselines in attention research. *Nature Reviews Neuroscience*, 1, 147-148.
- Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, USA*, 99, 15822-15826.
- Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, 110(2), 160-170.
- Ghazanfar, A. A., & Santos, L. R. (2004). Primate brains in the wild: The sensory bases for social interactions. *Nature Reviews Neuroscience*, 5(8), 603-616.
- Gopnik, A. (2009). *The Philosophical Baby: What Children's Minds Tell Us About Truth, Love, and the Meaning of Life*. New York: Farrar, Straus and Giroux.
- Hillyard, S. A., Hink, R. F., Schwent, V. L., and Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science*, 182, 177-180.
- Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40 (1-2), 1-19.
- Johnson, S. C., Slaughter, V., & Carey, S. (1998). Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. *Developmental Science*, 1(2), 233-238.
- Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, 83(2), B35-B42.
- Kruschke, J. K. (2011). Models of attentional learning. In: E. M. Pothos and A. J. Wills (eds.), *Formal Approaches*



- in *Categorization*, pp. 120-152. Cambridge University Press.
- Lepsien, J., & Nobre, A. C. (2007). Attentional modulation of object representations in working memory. *Cerebral Cortex*, 17(9), 2072-2083.
- Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological Bulletin*, 126(2), 281-308.
- Mazza, V., Turatto, M., Umiltà, C., & Eimer, M. (2007). Attentional selection and identification of visual objects are reflected by distinct electrophysiological responses. *Experimental Brain Research*, 181, 531-536.
- Mitchell, J. F., Sundberg, K. A., & Reynolds, J. H. (2009). Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. *Neuron*, 63, 879-888.
- Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715), 782-784.
- Nodine, C. F., Kundel, H. L., Lauver, S. C., & Toto, L. C. (1996). Nature of expertise in searching mammograms for breast masses. *Academic Radiology*, 3 (12), 1000-1006.
- Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32(1), 3-25.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926-1928.
- Sarter, M., Givens, B., & Bruno, J. P. (2001). The cognitive neuroscience of sustained attention: Where top-down meets bottom-up. *Brain Research Reviews*, 35(2), 146-160.
- Scerif, G., & Wu, R. (in press). Developmental Disorders. In A.C. Nobre & S. Kastner (Eds.) *The Oxford Handbook of Attention*. Oxford: OUP
- Schlesinger, M., Amso, D. & Johnson, S. P. (under review). Simulating the role of visual selective attention during development of perceptual completion.
- Solso, R. L. (2001). Brain activities in a skilled versus a novice artist: An fMRI study. *Leonardo*, 34(1), 31-34.
- Van Voorhis S., & Hillyard S. (1977). Visual evoked potentials and selective attention to points in space. *Perception & Psychophysics*, 22, 54-62.
- Wu, R., & Kirkham, N. Z. (2010). No two cues are alike: Depth of learning during infancy is dependent on what orients attention. *Journal of Experimental Child Psychology*, 107, 118-136.
- Wu, R., Gopnik, A., Richardson, D. C., & Kirkham, N. Z. (2011). Infants learn about objects from statistics and people. *Developmental Psychology*, 47 (5), 1220-1229.
- Wu, R., Kirkham, N. Z., Swan, K. A., & Gliga, T. (under revision). Direct social signals scaffold learning from novel attention cues during infancy.

# Adaptive Information Search and Decision Making over Single and Repeated Plays

**Dirk U. Wulff (Dirk.Wulff@unibas.ch)**

Department of Psychology, Missionstrasse 64A  
4055 Basel, Switzerland

**Thomas T. Hills (T.T.Hills@warwick.co.uk)**

Department of Psychology, University of Warwick  
Coventry CV4 7AL, UK

**Ralph Hertwig (Ralph.Hertwig@unibas.ch)**

Department of Psychology, Missionstrasse 64A  
4055 Basel, Switzerland

## Abstract

For over 50 years expected value and expected utility theory has been challenged by behavioral findings in repeated and single plays of risky gambles. The inherent long-term nature of these models has been found to be at odds with preferences indicating short-term maximization in single play situations. With the present study we provide further evidence on the distinction between long-term and short-term oriented behavior. Evaluating experienced-based decisions over repeated and single play situations we demonstrate that both choice preferences and search behavior change in response to long and short-term framing. This suggests different cognitive approaches for single and repeated play situations, with single decisions often favoring risk-aversion and therefore the underweighting of rare events. These findings are in line with alternative models of risky choice as for example proposed by Lopes (1996) and also the literature on state-dependent foraging.

**Keywords:** Decisions from experience, information sampling; risky choice, single-play and repeated-play.

## Introduction

Over lunch nobel-laureate Paul Samuelson offered his colleague the following bet: Win 200\$ with a 50% chance or lose 100\$. To his surprise the colleague rejected the offer but mentioned that he would agree to a series of a hundred such bets. Samuelson (1963) considered the choice of his colleague irrational. He later showed how this pattern of choice is incompatible with expected-utility maximization. A series of bets, which individually are all unacceptable, should not be accepted. But is the behavior of Samuelson's colleague irrational?

The prevailing answer after 50 years of scientific debate seems to be unanimously no. However, very different theoretical explanations have been invoked. On one hand, it has been demonstrated that the behavior of

Samuelson's colleague can indeed be captured by models of expected utility theory when the decision is made over the aggregated outcome of repeated plays (Aloysius, 2007). On the other hand, "distinct process differences between the situations" (Wedell, 2011) have been emphasized. Particularly Lola Lopes argued in favor of a two-stage decision process in which the qualitative feature of "coming-out-ahead" plays an important role (Lopes, 1996; Lopes & Oden, 1999). Formally Lopes assumes next to decumulative weighting a second process that maximizes the probability of achieving an aspiration level.

Along the lines of Lola Lopes' explanation, Hills and Hertwig (2010) recently provided evidence for a decision strategy that focuses on winning most of the time. Hills and Hertwig used an alternative paradigm of risky choice – decisions from experience – in which the information about options is acquired in a prior inconsequential sampling phase. They found that the choice behavior of people who sampled less and evaluated options in shorter intervals – were best described by a "round-wise" decision strategy. Instead of aiming at the overall higher mean of both options, this strategy compares the options over all sampling transitions and tallies the wins. Consequently, this strategy favors the option with a higher probability of coming out ahead—i.e., the option that wins most of the time. The results of Hills and Hertwig (2010) indicate that distinct processes are indeed at work when people aim for either long-run expected value maximization or for short-run maximization that focuses on coming-out ahead.

The distinction between such strategies is prevalent also outside the field of human decision-making. For

example, risk sensitivity and state-dependent foraging explain why a bird that needs food regularly to avoid starvation should concern itself with short-term outcomes and not seek for the better long-term average, which may not come in time (e.g., Caraco, 1980; Houston & McNamara, 1999; Stephens, 1981). In the present study we build on these ideas by examining information search and decision making in single and repeated play situations.

### Behavioral Findings Under Repeated Plays

Behavioral studies have shown that in repeated-play situations—where people get to play the same gamble multiple times—people are more likely to act according to the principles of expected utility maximization (Wedell, 2011). Generally, under repeated plays, the preference for the option with the higher expected value is substantially increased (e.g. Montgomery & Adelbratt, 1982). In addition, repeated plays reduce a number of well-known decision anomalies including possibility and certainty effects (Keren & Wagenaar, 1987; Keren, 1991), violations of procedural invariance (Wedell & Böckenholt, 1990) and ambiguity aversion (Liu & Coleman, 2009).

The effect of repeated plays on choice preference has also been established in applied settings. Redelmeier and Tversky (1990, 1992) found that physicians make differential decisions for individual versus aggregated instances. Bernartzi and Thaler (1990) have also demonstrated the relevance of the single repeated play distinction for investment decisions.

### Present Study

The goal of the present study is to evaluate the effect of single and repeated plays in the context of decisions from experience. By incorporating an active sampling phase decisions from experience account for the lack of full information about the available options in real life; a quality that is neglected in the commonly used paradigm of decisions from description. Based on the findings of Hills and Hertwig (2010) we expect the contrast between repeated and single play to impact pre-choice information search in addition to preferences. Specifically we expect people in single play as compared to repeated play situations to draw fewer samples and have shorter evaluation intervals for the available options. We also expect a higher preference in single play situations for

options that have a higher probability of winning most of the time irrespective of expected. For repeated plays we expect the opposite pattern. Such findings would indicate an overall differential cognitive approach to single and repeated plays of the same gamble and speak in favor of choice models such as Lola Lopes' two-stage account.

Additionally, we are interested in the degree to which the pattern of results is dependent on the certainty of the available options. In much of the risky choice literature a risky high outcome option is paired with an option that offers a smaller but certain outcome. Thereby, the quality of “coming-out-ahead” is often confounded with certainty. However, given the prior assertions we would expect the pattern of results to be largely independent over cases where the secure option is certain (or only relatively secure).

Finally, we want to evaluate how the standard application of decisions from experience relates to situations where the single and repeated-play character is made salient. Decisions from experience have recently received much attention as they depart from decisions that are made based on full description of the options (see Hertwig & Erev, 2009). In this context it will be very telling to observe if the overall pattern of results in the paradigm used most often in the literature (following Hertwig et al., 2004) resembles more a single or repeated play instantiation.

## Method

**Participants** We collected data from 124 participants. The mean age of the sample was about 24.2 years, 85% were students of the University of Basel. Participants were rewarded either by course credit or a fixed payment of about 13\$. Every participant also received a performance-based bonus as a result of his or her choices.

**Materials** A set of 12 target problems was created (see table 1). Each problem required a choice between two gambles. Every gamble was comprised of two outcomes – one positive outcome and zero. One gamble posed a relatively secure ( $p > .7$ ) or certain chance ( $p = 1$ ) to win a positive outcome. The other gamble was substantially more risky ( $p \leq .15$ ). This riskier option was always superior in expected value (1.5 to 1.8 times as high). We refer to these two options as the low (L) and high (H) expected value options, respectively. For control purposes, we included two problems where the high outcome options had the lower expected value (C1,

Table1: Study problems

Problem	H	L
1	92 with $p=.05$	3 with certainty
2	34 with $p=.05$	1 with certainty
3	120 with $p=.05$	5 with $p=.70$
4	44 with $p=.05$	2 with $p=.70$
5	70 with $p=.10$	4 with certainty
6	16 with $p=.10$	1 with certainty
7	54 with $p=.10$	4 with $p=.75$
8	23 with $p=.10$	2 with $p=.75$
9	35 with $p=.15$	3 with certainty
10	21 with $p=.15$	2 with certainty
11	48 with $p=.15$	5 with $p=.80$
12	9 with $p=.15$	1 with $p=.80$
C1	0 with certainty	1 with $p=.75$
C2	0 with certainty	1 with certainty
C3	9 with $p=.10$	3 with $p=.75$
C4	7 with $p=.10$	2 with certainty

C2) and two problems where the supposedly high outcome option was actually a sure event of zero (C3, C4). Those problems were created to exclude participants showing unsystematic or extreme risk seeking behavior.

**Procedure** The experiment was entirely computer based. Participants were given verbal and visual instructions that explicitly explained the payoff modality according to their assigned experimental condition. Participants then made choices for three practice trials. On every problem the participants were able to sample from the two options as much and in whatever fashion they liked. They were instructed to proceed to the decision whenever they felt confident enough. Finally the participants made their decisions on all 16 problems (12 target and 4 control problems) in random order.

**Payoff Manipulation** The single and repeated play character was induced between-subjects through different payoff modalities. The payoff in the repeated play condition was determined by one hundred draws from one of the participant's chosen options. In the single play condition the payoff, one random sample was taken from

one the participant's chosen options and then this value was multiplied by one hundred. A third condition (neutral) corresponded to the payoff modality usually applied in the literature (Hertwig et al., 2004), with the payoff equivalent to one random draw from each option chosen by the participant. As the payoff in the neutral condition is based on only 16 draws all presented outcomes were multiplied by a fixed factor to equal the expected values in the other two conditions.

## Results

We applied two criteria to exclude participants showing extreme risk seeking or unsystematic behavior. First, we excluded participants that neither preferred the higher mean nor the higher median in the control problems C1 and C2 (see table 1). Second, we also excluded participants that did not sample at least once from every option. As a result the following analyses are based on 82 out of 124 participants. Inspection of control problems C3 and C4 reveals that this restriction very efficiently reduces zero EV choices, which can be regarded as an indicator for either extreme risk seeking or not very systematic behavior. The excluded participants chose the zero EV option in about 27% of the cases, whereas the included participants chose did not choose this option at all. Thus, the remaining participants showed neither extreme risk-seeking nor unsystematic behavior.

The main focus of this study is on the contrast between the single and repeated play conditions. Statistical tests are therefore mainly reported for this contrast. Additionally, comparisons for the neutral condition are reported to see where usual experimental instantiations of this paradigm fall in the context of salient single and repeated play situations.

Figure 2A illustrates choice patterns over the experimental conditions. In support of our predictions, H preferences increased for the repeated-play as compared to single-play situation ( $t_{42.4}=3.58$ ,  $p<.001$ ). H preference in the neutral condition resembled the pattern in the single-play condition ( $t_{53.1}=0.54$ ,  $p=.594$ ), but there was a difference for H choices between the neutral and repeated play conditions ( $t_{43.8}=-3.05$ ,  $p=.004$ ). This effect is not affected by the inclusion of the certainty of the L option in the prediction of H choices ( $\chi^2_{2}=.51$ ,  $p=.776$ ). Because none of the following analyses were influenced by a comparison of certain and uncertain gambles, the following results are collapsed over both problem types.

Figure 2B shows the information search behavior of the participants. Descriptively, it matches our predictions. The repeated-play condition elicited higher total sample sizes as well as longer evaluation intervals for the individual options (samples per transition: average number of samples taken per transition in the sampling sequence). However, only the contrast for total number of samples reaches statistical significance (total:  $t_{32.9}=2.28$ ,  $p=.029$ ; per transition:  $t_{31.8}=1.58$ ,  $p=.124$ ). Again the neutral condition closely resembled the single-play condition in both sampling variables (total:  $t_{49.1}=.52$ ,  $p=.609$ ; per transition:  $t_{52.1}=.85$ ,  $p=.40$ ), but was at least in terms of total sample size marginally different from the repeated-play situation (total:  $t_{37.9}=-1.85$ ,  $p=.072$ ; per transition:  $t_{34}=-1.07$ ,  $p=.29$ ).

So far we have shown that the payoff structure independently affected choice and information search in the predicted direction. However, we generally predicted that information search and choice behavior were likewise affected by the induction of a repeated play situation in that it was expected to elicit both a higher preference for the long run winner as well as higher total and per transition sample sizes. To demonstrate this, two additional analyses have to be carried out. First, it has to be shown that the contrast for H preference is not entirely caused by a sampling bias. Hertwig et al. (2004) showed how small sample sizes can systematically obscure rare outcomes in decisions from experience. When rare outcomes are obscured, the option that secures winning most of the time can also appear as the option with the

higher long run expectation. Second, an association between the choice preferences and the sampling behavior needs to be established. Therefore we tested if this pattern of results holds when the qualities of winning most of the time – represented by the higher median – and higher long run expectation – represented by the higher mean – do not coincide in respect to actually observed outcomes. Then we evaluate if the preference for the higher mean option is associated with the sampling behavior.

Figure 3 shows the proportions of choices that were consistent with the higher experienced mean. The separate lines distinguish cases where the experienced median and mean predict the same (dotted lines) or different (solid lines) choices. Focusing on the cases where the mean and median do not fall together, it is apparent that participants in the repeated play situation opted more often for the option with the higher experienced mean as compared to the single play situation ( $t_{49.1}=2.72$ ,  $p=.001$ ). Thus, the differences in choices do appear to be associated with different decision strategies and not simply differences in observed outcomes. The neutral condition again matches the single ( $t_{48.3}=.12$ ,  $p=.90$ ) but not the repeated play condition ( $t_{48.8}=-2.26$ ,  $p=.028$ ). Overall, the observed differences are all the more convincing given the pattern of results for cases where the experienced mean and median fall together. Independent of the payoff condition we observed strong preferences for the option with the higher experienced mean pointing towards very systematic

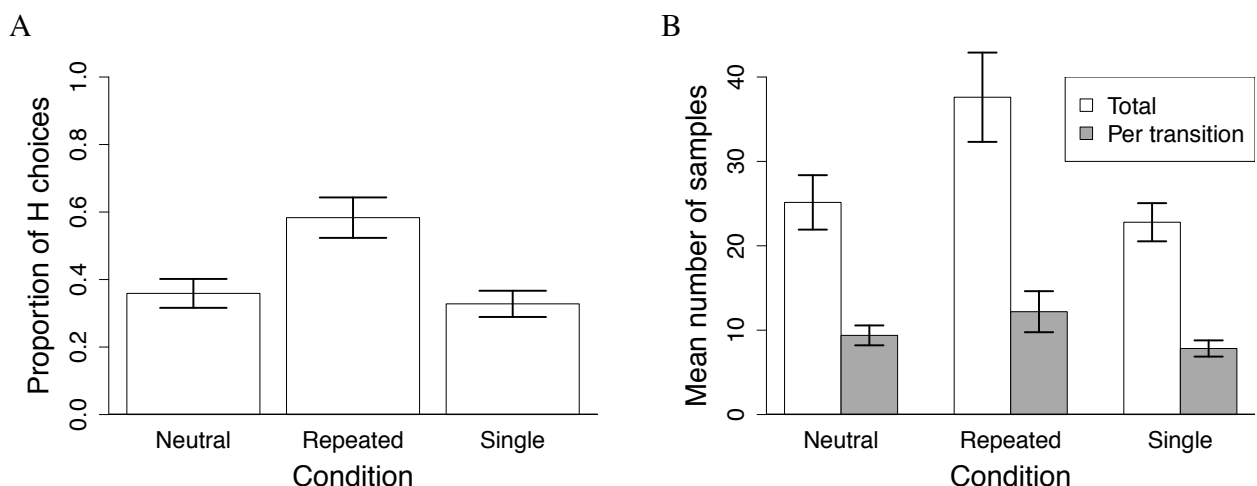


Figure 2: (A) Proportion of H choices as a function of the experimental condition. (B) Mean number samples in total and per transition in the sampling sequence as a function of the condition. Error bars represent standard error of the mean.

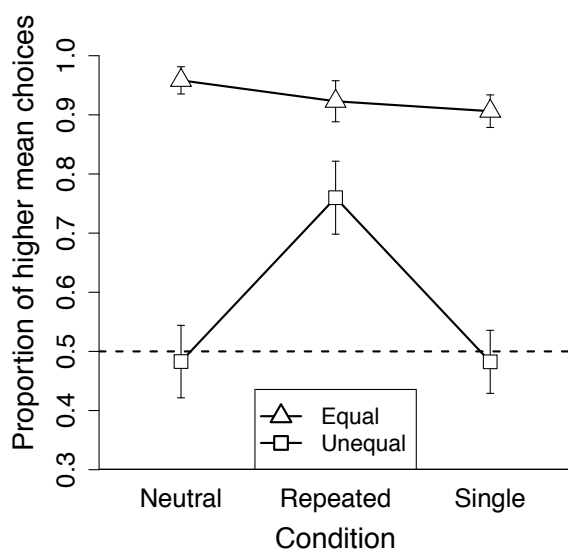


Figure 3: Higher mean choices as a function of the experimental condition.

behavior.

Now, is the shift in strategy use for the repeated play condition associated with changes in the sampling behavior? In line with our predictions we found this to be the case. Regressing higher mean choices on sample size and the group contrast of single and repeated play situations yields a significant interaction ( $t_{52} = -2.73$ ,  $p = .009$ ), while the main effect of the contrast vanishes ( $t_{52} = .54$ ,  $p = .59$ ). The same pattern was found for samples per transition (interaction:  $t_{52} = -2.36$ ,  $p = .022$ ; contrast:  $t_{52} = -0.33$ ,  $p = .746$ ). It was not possible to determine which of these sampling variables were more strongly associated with the strategy use because they were highly correlated ( $r = .83$ ). In sum, we thus were able to show, that the repeated play condition influenced both choice preference and sampling behavior.

## Discussion

The primary aim of our study was to use experience-based decisions to investigate information search and decision making over repeated and single plays. Based on prior findings of Hills & Hertwig (2010) we predicted that the situations of repeated and single plays would elicit information search behavior that has been found to be associated with short and long-term maximization, respectively. Our results confirm this prediction. Participants in the situation of repeated plays – particularly those that sampled more avidly and switched

less frequently between the options – exhibited higher preferences for the superior option in the long run. These results did not appear to be mediated by a simple sampling bias (Hertwig et al., 2004; Fox & Hadar, 2006). Rather, they were associated with different decision strategies.

We believe that our findings provide new evidence for the theoretical debate on single and repeated-play decisions. So far behavioral studies that separated these situations have focused entirely on decision outcomes. We extend this literature by showing that differences can also be demonstrated on more process-oriented measures such as information search. Our results imply that possibly more is changing over the situations of single and repeated play than the weighting of probabilities and outcomes (Aloysius, 2007), as suggested by unitary theoretical accounts, e.g. prospect theory (Kahneman & Tversky, 1979). We suggest that an overall different cognitive approach towards repeated plays is applied. These results appear to be consistent with theoretical accounts such as Lola Lopes' (1996, 1999) two-stage model.

Our findings also hold important insights for the literature on experienced-based decision making. On all of our measures, the standard payoff procedure resulted in behavior that resembled the condition where the single-play character was strongly emphasized. This might not appear surprising as in both cases only single draws were taken out of the chosen option. However, in the standard realization all decisions contribute to the overall payoff. Thus, the whole set of choices could be mentally combined (see Read, Loewenstein, & Rabin, 1999) and thereby the set of problems could more resemble a repeated play situation. In this context our findings suggest that under the standard paradigm of decisions from experience, all choices are evaluated independently. This finding might contribute in the explanation of the description-experience gap.

In conclusion, we find that single and repeated play situations impact more than just choice preferences. The way people search for information and specific decision strategies change as well. Our findings emphasize the distinction between long-term and short-term maximizing behavior in human decision-making - a distinction that is long established in the literature on animal foraging.

## References

Aloysius, J. A. (2007). Decision making in the short and

- long run: Repeated gambles and rationality. *British Journal of Mathematical and Statistical Psychology*, 60, 61–69.
- Benartzi S., & Thaler, R. H. (1999). Risk aversion or myopia? Choices in repeated gambles and retirement investments. *Management Science*, 45, 364–381.
- Caraco, T. (1980). On foraging time allocation in a stochastic environment. *Ecology*, 61, 119–128.
- Fox, C. R., & Hadar, L. (2006). “Decisions from experience” = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgment and Decision Making*, 1, 159–161.
- Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15, 534–539.
- Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, 13, 517–523.
- Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience: Do our patterns of sampling foreshadow our decisions? *Psychological Science*, 21, 1787–1792.
- Houston, A. I., & McNamara, J. M. (1999). *Models of adaptive behavior: An approach based on state*. Cambridge, England: Cambridge University Press.
- Liu, H. H., & Colman, A. M. (2009) Ambiguity aversion in the long run: Repeated decisions under risk and uncertainty. *Journal of Economic Psychology*, 20, 277–284.
- Lopes, L. L. (1996). When time is of the essence: Averaging, aspiration, and the short run. *Organizational Behavior and Human Decision Processes*, 65, 179–189.
- Lopes, L. L., & Oden, G. C. (1999). The role of aspiration level in risky choice: A comparison of cumulative prospect theory and SP/A theory. *Journal of Mathematical Psychology*, 43, 286–313.
- Montgomery, H., & Adelbratt, T. (1982). Gambling decisions and information about expected value. *Organizational Behavior and Human Performance*, 29, 39–57.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263–292.
- Keren G. (1991). Additional tests of utility theory under unique and repeated condition. *Journal of Behavioral Decision Making*, 4, 297–304.
- Keren, G., & Wagenaar, W. A. (1987). Violation of utility theory in unique and repeated gambles. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 13, 387–391.
- Read, D., Loewenstein, G., & Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, 19, 171–197.
- Redelmeier, D. A., & Tversky, A. (1992). On the framing of multiple prospects. *Psychological Science*, 3, 191–193.
- Redelmeier D. A., & Tversky, A. (1990). Discrepancy between medical decisions for individual patients and for groups. *New England Journal of Medicine*, 322, 1162–1164.
- Samuelson, P. A. (1963). Risk and uncertainty: A fallacy of large numbers. *Scientia*, 98, 108–113.
- Stephens, D. W. (1981). The logic of risk-sensitive foraging preferences. *Animal Behavior*, 29, 628–629.
- Wedell, D. H. (2011). Evaluations of single and repeated-play gambles. *Wiley Encyclopedia of Operations research and Management Science*.
- Wedell, D. H., & Böckenholt, U. (1990). Moderation of preference reversals in the long run. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 429–438.



# Stable Self-to-Object Spatial Relations Acquired from Sequential Spatial Learning

**Chengli Xiao (xiaocl@nju.edu.cn)**

Department of Psychology, Nanjing University, 22 Hankou Road  
Nanjing 210093, P.R. China

**Fudan Chen (cfdchashao@163.com)**

Department of Psychology, Nanjing University, 22 Hankou Road  
Nanjing 210093, P.R. China

## Abstract

Self-to-object spatial relations are generally considered to be transient and supported primarily by perceptual processes. The present study investigates whether people can acquire stable self-to-object spatial relations that are not disrupted by disorientation. Participants either simultaneously or sequentially viewed the object locations from a learning position amidst a geometrically irregular array. Next they were blindfolded and pointed to the objects under three conditions: before turning (baseline), after rotating 240° (updating), and after disorientation (disorientation). Finally, all participants were taken to another room to perform judgments of relative direction (JRDs) among remembered object locations. The internal consistency of pointing among objects was disrupted by disorientation following simultaneous viewing but not sequential viewing. However, participants' memories of object-to-object relations were equivalent in the two viewing conditions. Together, these results suggest that people construct stable self-to-object spatial relations when they sequentially view each object of the irregular layout.

**Keywords:** self-to-object spatial relations; sequential learning; disorientation

## Introduction

In everyday life, people use self-to-object (egocentric) and object-to-object (allocentric) spatial relations to encode the location of objects or landmarks in the environment, navigate effectively to significant places, and reorient themselves when getting lost. In a self-to-object reference system, locations are represented with respect to the particular perspective of a perceiver, whereas in an object-to-object reference system, locations are represented within a framework external to the holder of the representation and independent of his or her position e.g., Easton & Sholl, 1995; Klatzky, 1998).

Generally, it is believed that self-to-object spatial relations are transient and supported primarily by the perceptual processes, and that object-to-object spatial relations are stable and can be preserved in the memory system (e.g., Burgess, 2006; Holmes & Sholl, 2005; Mou, McNamara, Valiquette, & Rump, 2004). When moving around the environment, if people navigate by means of the self-to-object spatial relations, one has to add a common vector to each individual target vector to compute the new egocentric coordinates of the target locations. If this updating process is disrupted through procedures that induce

a state of disorientation, the coherence of relative locations among different targets is reduced. Therefore, a disoriented participant's pointing response will show an inconsistency among different targets (disorientation effect). To the contrary, if people navigate by means of the object-to-object spatial relations, which remains the same regardless of people's movements, the state of disorientation cannot reduce the coherence of relative locations among different targets. Therefore, the consistency of the pointing response among different targets will be equal between oriented and disoriented participants (absence of disorientation effect). By measuring the standard deviation across target objects of the mean signed pointing errors (configuration error), the accuracy of the localization of each target in relation to the others can be assessed. Following this logic, recent research has indicated that people can navigate by means of either the transient self-to-object or stable object-to-object spatial representations (Holmes & Sholl, 2005; Mou, McNamara, Rump, & Xiao, 2006; Sargent, Dopkins, Philbeck, & Modarres, 2008; Waller & Hodgson, 2006; Wang & Spelke, 2000). In most situations, people acquire both self-to-object and object-to-object spatial representations, and can navigate by means of one of them according to the layout geometry or instruction (Xiao, Mou, & McNamara, 2009).

However, in all the previous research, which indicated that people were able to navigate by means of transient self-to-object spatial representations, the pointing responses of disoriented participants still had a relatively high consistency among different targets. The configuration errors in the disorientation condition were no more than 30°, which were much less than the expected configuration error of randomly pointing (approaching 104°). It is possible that disorientation does not totally disrupt the self-to-object spatial representations, and that the disoriented participants can still locate objects based on the impaired self-to-object spatial representations that persist in their memory. One of our recent experiments provided circumstantial evidence for this hypothesis (Xiao, et al., 2009). After visually learning object locations amidst, or at the periphery of an irregular array (see Figure 1), blindfolded participants pointed to the objects before turning (baseline), after rotating 240° (updating), and after disorientation (disorientation). In both learning conditions, the configuration error significantly increased after disorientation, indicating that the participants located objects by means of the self-to-object spatial relations. When explicitly instructed to use the object-to-

object spatial relations (e.g., “Please keep track of all of the locations of the objects relative to other objects while you are turning to face the ball.”), after the baseline pointing test and before rotation, the participants who learned at the periphery of the irregular array could follow the instruction to prevent the disorientation effect, while the participants who learned amidst the irregular array could not follow the instruction to prevent the disorientation effect. These results suggest that after visually learning object locations at the periphery of the irregular array, the participants established both self-to-object and object-to-object spatial relations, but updated self-to-object spatial relations during rotation by default. They could also update object-to-object spatial relations when required. However, after visually learning the object locations amidst an irregular array, the participants can only establish the self-to-object spatial relations. They may only represent minimally, if at all, object-to-object spatial relations, which cannot be used during rotation. Therefore, there is little possibility that the participants used the object-to-object spatial relations after disorientation. The disoriented participants can only locate objects by means of the self-to-object rather than the object-to-object spatial relations, suggesting that the self-to-object spatial relations are preserved, to some extent, over disorientation in memory.

In the object-to-object spatial relations, object locations are represented with respect to another object or set of objects, while in the self-to-object spatial relations object locations are represented with respect to the perceiver (e.g., Easton & Sholl, 1995; Klatzky, 1998; Mou, Xiao, & McNamara, 2008). If the perceiver takes him or herself as a stable object, and refers every other object location relative to him or herself, a special kind of object-to-object spatial representation is built, and can be well preserved in memory as another kind of object-to-object spatial representation. In other words, the perceiver establishes a stable self-to-object spatial representation. After disorientation, the perceiver can recover object locations by retrieving the remembered self-to-object spatial information. In Xiao et al. (2009), the participants visually learned object locations amidst the irregular array, where they could not perceive the whole layout from a single viewpoint. However, participants could directly view inter-object spatial relations between objects separated by small angular distances, and thus fragmentary object-to-object spatial representations might be acquired (Sargent, Dopkins, Philbeck, & Chichka, 2010). Attending to and memorizing neighboring object-to-object spatial relations might interfere with the acquisition of self-to-object spatial relations. Therefore, participants might develop unstable self-to-object spatial representations after they visually learned amidst the irregular array. There is a high possibility that participants will construct more stable self-to-object spatial representations through the new learning methods, by which they can only directly perceive the self-to-object spatial relations but not the inter-object ones, such as through sequential learning. Previous research has demonstrated that participants can learn spatial locations

by viewing one object at a time (e.g., Yamamoto & Shelton, 2007, 2009). Compared with visually learning object locations amidst the array, sequentially viewing each object prevents participants from directly perceiving any inter-object relations, but compels them to focus on each object’s location relative to themselves. Therefore, the acquisition of object-to-object spatial representations is maximally reduced and the salience of self-to-object spatial relations is enhanced, and the participants may develop stable self-to-object spatial representations and minimal object-to-object spatial relations.

In the present study, participants either sequentially or simultaneously viewed object locations from a learning position amidst the same irregular layout as in Xiao et al. (2009). After learning, all participants were blindfolded and pointed to object locations in baseline, updating, and disorientation conditions. Before rotating to a new heading in the updating condition, half of the participants were explicitly instructed to use object-to-object spatial relations during locomotion as in Xiao et al. (2009). At last, all participants were taken to another room to perform judgments of relative direction (JRDs), which have been commonly used to assess the memory of the object-to-object relations in an environment (e.g., Mou, et al., 2004; Shelton & McNamara, 2001; Waller & Hodgson, 2006). Because we hypothesized that participants would use self-to-object spatial relations to locate objects before and after disorientation when learning amidst the irregular layout, and that the participants would establish more stable self-to-object spatial relations following sequential viewing than by following simultaneous viewing, we expected that the configuration error in sequential viewing condition would be smaller than that in the simultaneous viewing condition, and that the disorientation effect would be absent in the sequential viewing condition but present in the simultaneous viewing condition. Meanwhile, since we hypothesized that the participants would establish minimal object-to-object spatial relations in both the simultaneous and sequential viewing condition, we expected that the participants could not use the object-to-object spatial relations during locomotion and after disorientation. As in Xiao et al. (2009), the participants could not follow the object-to-object instruction to prevent the disorientation effect after simultaneous learning of the layout. Because we predicted that the participants would use the stable self-to-object spatial relations after sequentially learning the layout, we expected that there will be no disorientation effect in both the object-to-object instruction and the non-instruction group. Because the participants established minimal object-to-object spatial relations, a floor effect might be present, and the performance on JRDs in the sequential learning group might be equivalent or inferior to that in the simultaneous learning group.

## Method

### Participants

Thirty-two university students (16 men and 16 women) participated in this experiment in return for monetary compensation.

### Materials

The irregular layout of Xiao et al. (2009) was used in this experiment. As illustrated in Figure 1, nine objects were presented on the floor in a cylinder that was located in an experiment room. The cylinder was 3.0 m in diameter, made by black fabric and reinforced cloth. Objects were chosen with the restrictions that they were visually distinct, fit with approximately 0.3 m on each side, were familiar to people, and shared no obvious semantic association. The scissors, the hat, and the brush were placed in a line. Participants were standing 1 meter away from the brush, facing the scissors. The floor was covered with gray carpet. Four lights were placed on the ceiling near the side of the cylinder to illuminate the area. They were placed at equal intervals and at equal distance from the center of the cylinder to minimize directional illumination cues.



Figure 1: Layout of objects used in the study.

Test trials were presented by a computer via wireless earphone. A joystick was used as the pointing apparatus.

### Procedures and Design

Before entering the study room, each participant was instructed to learn the location of objects for a spatial memory test and trained in how to use a joystick. After that, the objects that would be encountered in the experiment layout were presented individually to all participants and the name of each object was given. Then, the blindfolded participants were escorted to the study room and led to the learning position by the experimenter. The participant was then asked to remove the blindfold.

**Simultaneous Viewing** All nine objects were presented on the floor. Half of the participants viewed the layout for 30 s and then closed their eyes and named and pointed to each

object with one of their fingers. Throughout the learning phase, the participants were stationary at the learning position. They were allowed to turn their heads to review the layout but were required to maintain their body orientation.

**Sequential Viewing** The other half of the participants viewed each object presented alone for 3 s in a spatially random sequence. To control the viewing time, the participant was asked to close his/her eyes while the experimenter replaced the just-viewed object on the floor with a new object in a new location. After viewing the last object, the participant was asked to close his/her eyes and to name and point to each object with one of their fingers. The learning sequence of the objects was randomized.

In both learning conditions, the learning-pointing session was repeated until participants could fluently name and point to the correct object locations twice in a row. (Fluency and accuracy were determined by the experimenter's visual inspection of the pointing performance.) The number of repetitions needed to achieve the learning criterion was recorded. During training and learning, the participants were not aware of what particular tasks they would perform in the testing stage.

**Egocentric Pointing Tasks** After learning the layout, all of the participants put on the wireless earphone, and held the joystick against their front waist. All of the participants were blindfolded and tested in the order of the baseline, updating, and disorientation conditions. In the baseline condition, participants maintained their heading to the scissors. In the updating condition, participants rotated 240° by themselves (e.g., "Please turn right until you are facing the candle"). Half of the participants turned right to face the candle, and half turned left to face the ball. Within each group, immediately before rotation, half of the participants were explicitly instructed to use object-to-object spatial relations during rotation (e.g., "Please keep track of all of the locations of the objects relative to other objects while you are turning to face the candle"). The other half were not given this instruction. In the disorientation condition, the participants rotated in place for 1 minute. Then they pointed to the location of an object named by the experimenter (e.g., "Please point to the ball"). This rotation and pointing procedure was repeated until the absolute pointing error was larger than 90°. Then the participants were instructed to turn to face the ball (or candle) if they faced the candle (or ball) in the updating condition ("Please turn right until you believe you are facing the ball"). They were allowed to adjust their position by themselves if they thought they had drifted off the testing location while rotating. A recovery period was given before the final pointing test.

In each rotation condition, four blocks of trials were included, and each block involved pointing to all nine objects once in a random order. After hearing the warning indication ("Start"), the participants pulled the joystick trigger, and the target object was immediately announced (e.g., "Please point to the candle"). Then the participants used the joystick to point to the direction of the target object.

The *configuration error* was measured as in Xiao et al. (see Table 1), which defined as the standard deviation of the means per target object of the signed pointing errors, which indicated the internal consistency of pointing response among different targets. As pointing data is inherently circular data, circular statistics (e.g., Jammalamadaka & SenGupta, 2001) were used.

Table 1: Definitional Formulas for Dependent Variables

Variable	Formula
Signed pointing error for object $i$ on trial $j$	$e_{ij} = \text{judged direction} - \text{actual direction}$
Mean signed pointing error for object $i$	$\bar{e}_i = \frac{\sum_j e_{ij}}{T}$
Configuration error	$\sqrt{\frac{\sum_i \left( \bar{e}_i - \frac{\sum_i \bar{e}_i}{N} \right)^2}{N-1}}$

Note:  $T$  = number of pointing trials per object;  
 $N$  = number of target objects.

**JRDs Task** After finishing the egocentric pointing tasks, the blindfolded participants were escorted to another room to perform JRDs. They were allowed to remove the blindfold and take a short break before proceeding to the JRDs. The participants first initiated each trial by pressing a button of the joystick. Trials began with the imagined standing location and facing object given aurally (e.g., “Imagine you are at the mug facing the ball”). After having a clear mental image of where he or she was standing and what he or she was facing, participants pulled the joystick trigger, and the target object was immediately presented (e.g., “Please point to the scissors.”). Then the participants used the joystick to point to where the target would be if he or she occupied the standing location and facing the direction as presented. The participants were instructed to hold the joystick exactly in the front of his or her waist and to keep the joystick forward when he or she pointed.

The JRDs test included 8 imagined headings. To facilitate exposition, we arbitrarily labeled headings counterclockwise from  $0^\circ$  to  $315^\circ$  in  $45^\circ$  steps. The learning direction was defined as  $0^\circ$ . Because the geometry of the layout was irregular, there were little pairs of objects established the imagined heading parallel to above 8 directions. Therefore, the imagined heading established by any pair of objects varied within  $\pm 15^\circ$  (that is,  $0^\circ \pm 15^\circ$ ,  $45^\circ \pm 15^\circ$ ,  $90^\circ \pm 15^\circ$ ,  $135^\circ \pm 15^\circ$ ,  $180^\circ \pm 15^\circ$ ,  $225^\circ \pm 15^\circ$ ,  $270^\circ \pm 15^\circ$ , and  $315^\circ \pm 15^\circ$ ) was included in the 8 imagined headings (e.g., at the candle facing the mug established the imagined heading  $1.56^\circ$ , and were taken as the imagined heading  $0^\circ$ ). The participants were given a total of 48 trials, six trials at each

of eight imagined headings. The dependent measures were the angular error of the pointing response, measured as the absolute angular difference between the judged pointing direction and the actual direction of the target.

In both egocentric pointing and JRDs tasks, pointing accuracy but not speedy response was emphasized.

## Results

### Egocentric pointing

Configuration error on egocentric pointing were subjected to mixed-model analyses of variance (ANOVAs), with the rotation condition (baseline, updating, and disorientation) as the within subject variable, the viewing type (sequential, simultaneous) and object-to-object instruction (yes, no) as the between subjects variables. The results revealed no main effect or interactions of the object-to-object instruction. Data were therefore collapsed across this factor for subsequent analyses.

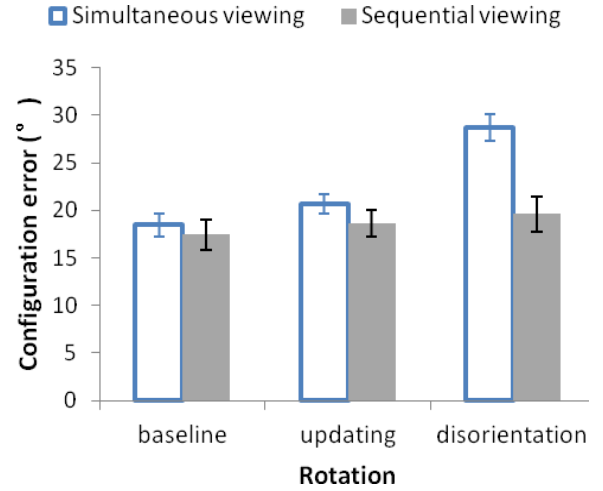


Figure 2: Configuration errors in egocentric pointing as a function of rotation condition and viewing type. Error bars are confidence intervals corresponding to  $\pm 1$  SEM.

The main effect of the rotation condition was significant,  $F(2, 60) = 13.13$ ,  $p < .001$ ,  $MSE = 24.92$ , the main effect of the viewing type was significant,  $F(1, 30) = 7.83$ ,  $p < .01$ ,  $MSE = 16.88$ , and the interaction of the rotation and the viewing type was significant,  $F(2, 60) = 6.09$ ,  $p < .005$ ,  $MSE = 24.92$ . As shown in Figure 2, three major findings were revealed. First, as in Xiao et al. (2009), the configuration error increased after disorientation when participants simultaneously viewed the layout, which indicated that participants had used the transient self-to-object spatial relations during rotation. This observation was supported statistically by a planned contrast comparing the participants' configuration errors in the updating condition with that in the disorientation condition following simultaneously viewing,  $F(1, 30) = 17.18$ ,  $p < .001$ ,  $MSE =$

59.05. Second, the configuration errors were equivalent before and after disorientation when participants sequentially viewed the layout, which indicated that participants used a kind of stable spatial relations before and after disorientation. This observation was supported statistically by a planned contrast comparing participants' configuration error for the updating condition with that for the disorientation condition following sequential viewing,  $F(1, 30) < 1$ . Third, the configuration errors were indistinguishable for simultaneous and sequential viewing in the baseline and updating conditions, but significantly higher for simultaneous viewing than for sequential viewing in the disorientation condition. These observations were supported statistically by planned contrast comparing participants' configuration errors for simultaneous viewing with those for sequential viewing at each rotation condition. There were no differences in the baseline and updating conditions,  $F_s(1, 30) \leq 1.39$ ,  $p_s \geq .24$ , but there was a significant difference in the disorientation condition,  $F(1, 30) = 15.01$ ,  $p < .001$ ,  $MSE = 43.69$ .

### JRDs

Performance data on JRDs were subjected to mixed-model analyses of variance (ANOVAs), with the imagined heading ( $0^\circ$  to  $315^\circ$  in  $45^\circ$  steps) as the within subject variable, viewing type (sequential, simultaneous) and object-to-object instruction (yes, no) as the between subjects variables.

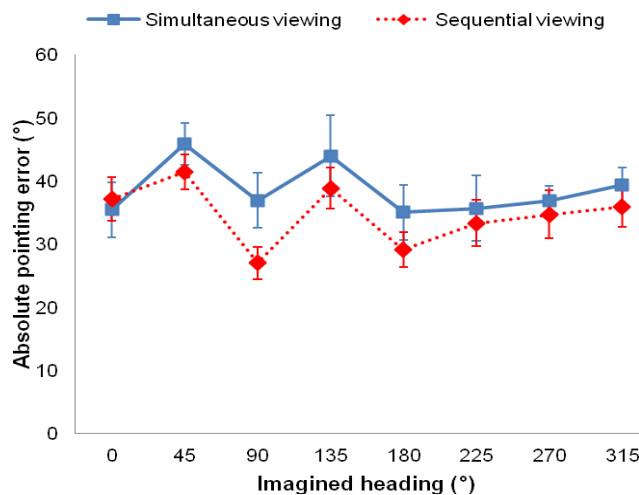


Figure 3: Mean absolute angular errors in JRDs as a function of imagined heading and viewing type. Error bars are confidence intervals corresponding to  $\pm 1$  SEM.

Three major findings were revealed. First, as shown in Figure 3, the participants' performance on JRDs was indistinguishable for simultaneous viewing and sequential viewing, which indicated that participants had constructed object-to-object spatial representations of equivalent fidelity through simultaneous viewing and sequential viewing. This observation was supported statistically by no main effect or

interactions of the viewing type,  $F_s \leq 1.36$ ,  $p_s \geq .25$ . Second, the participants' performance on JRDs was sensitive to the imagined headings, as supported statistically by the significant main effect of the imagined heading,  $F(7, 210) = 3.75$ ,  $p < .001$ ,  $MSE = 147.10$ . Third, the instruction before the updating condition affected the participants' performance on JRDs. The performances on JRDs were more accurate for the participants following object-to-object updating instruction than those following no instruction. This observation was supported statistically by the main effect of instruction,  $F(1, 28) = 5.30$ ,  $p < .05$ ,  $MSE = 727.92$ . The interaction between the imagined heading and instruction was marginally significant,  $F(7, 196) = 1.93$ ,  $p = .067$ ,  $MSE = 147.10$ . The simple effect of instruction was significant within the headings of  $135^\circ$ ,  $180^\circ$ , and  $225^\circ$ ,  $F_s(1, 28) > 4.93$ ,  $p_s < .05$ , but not significant within the headings of  $0^\circ$ ,  $45^\circ$ ,  $90^\circ$ ,  $270^\circ$ ,  $315^\circ$ ,  $F_s(1, 28) < 2.21$ ,  $p_s > .14$ .

### Discussion

The absence of the disorientation effect has been interpreted as evidence that object-to-object spatial relations have been used during rotation (Holmes & Sholl, 2005; Mou, et al., 2006; Sargent, et al., 2008; Xiao, et al., 2009). However, in the present study, the absence of the disorientation effect following sequential viewing can unlikely be explained by using the object-to-object spatial relations. The participants' performance on JRDs was equivalent across sequential and simultaneous viewing, suggesting that the participants established equivalent object-to-object spatial relations among each viewing group. If the object-to-object spatial relations could be used in one viewing group, there is little possible that the object-to-object spatial relations could not be used in another group. If the participants in the sequential viewing condition used the object-to-object spatial relations to prevent the disorientation effect, the participants in the simultaneous viewing condition should also be able to use them to prevent the disorientation effect. However, in the simultaneous viewing condition, the disorientation effect consistently appeared, even when the participants were explicitly required to use the object-to-object relations during rotation. This result is consistent with Xiao et al. (2009), indicating that the participants could not have used the object-to-object but rather used self-to-object spatial relations to perform egocentric pointing during rotation and after disorientation. Therefore, it is unlikely that the participants could have used the object-to-object spatial relations to perform egocentric pointing in the disorientation condition following sequential viewing. The absence of the disorientation effect following sequential viewing can only be explained by using the stable self-to-object spatial relations.

As the transient self-to-object spatial representation, the stable self-to-object spatial representation may make use of a special polar coordinate system in which the origin is at the participant and the reference direction is participant's front. Object locations are specified by egocentric distance

and egocentric bearing. During locomotion, the sensory-perceptual input enables the participant to update multiple self-to-object spatial relations. Although the procedures that induce a state of disorientation will disrupt this updating process, and the coherence of relative locations among different targets will plummet if the participant still relies on the sensory-perceptual system; the participant could recover object locations by retrieving the remembered angular and distance information from the learning view. If the test heading misaligned with the learning view, a mental rotation process would be involved to align the remembered self-to-object spatial representation with respect to the test heading. This retrieving process is similar to retrieving object-to-object spatial representations (Mou, Fan, McNamara, & Owen, 2008; Mou, Xiao, et al., 2008). At this point, the differences between object-to-object and stable self-to-object spatial representations become less dramatic, because both representations could be preserved in memory and be retrieved from a novel heading after disorientation. However, unlike the object-to-object spatial relations, it is difficult to use the stable self-to-object spatial relations to judge inter-object spatial relations. Because every single object is represented with respect to the self, the participants have to compute an inter-object spatial relation, and this computation process will introduce error (Klatzky, 1998). Therefore, in present study, the participants in the sequential learning condition could use the stable self-to-object spatial relations to avoid disorientation effect, but could not use them to improve their JRDs performance. Their performance on JRDs was not superior to the simultaneous viewing condition.

In summary, the present research indicates that by sequentially viewing every object location from a learning position, the participants constructed stable self-to-object spatial relations which could be preserved over disorientation. These results suggest that self-to-object spatial relations are not only transient and supported primarily by perceptual systems, but can also be stable and preserved in memory.

### Acknowledgements

Preparation of this article and the research reported in it were supported by a grant from the National Natural Science Foundation of China Grant 31000457 to Chengli Xiao. Correspondence concerning this article should be addressed to Chengli Xiao, Department of Psychology, Nanjing University, 22 Hankou Road, Nanjing 210093, China. E-mail: xiaocli@nju.edu.cn

### References

Burgess, N. (2006). Spatial memory: How egocentric and allocentric combine. *Trends in Cognitive Sciences*, 10, 551-557.

Easton, R. D., & Sholl, M. J. (1995). Object-array structure, frames of reference, and retrieval of spatial knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 483-500.

Holmes, M. C., & Sholl, M. J. (2005). Allocentric coding of object-to-object relations in overlearned and novel environments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 1069-1087.

Jammalamadaka, S. R., & SenGupta, A. (2001). *Topics in circular statistics*. Singapore: World Scientific Publishing Co. Pte. Ltd.

Klatzky, R. L. (1998). Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In C. Freksa, C. Habel & K. F. Wender (Eds.), *Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge LNAI 1404* (pp. 1-17). Berlin: Springer-Verlag.

Mou, W., McNamara, T. P., Rump, B., & Xiao, C. (2006). Roles of egocentric and allocentric spatial representations in locomotion and reorientation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 1274-1290.

Mou, W., McNamara, T. P., Valiquette, C. M., & Rump, B. (2004). Allocentric and egocentric updating of spatial memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 142-157.

Mou, W., Xiao, C., & McNamara, T. P. (2008). Reference directions and reference objects in spatial memory of a briefly viewed layout. *Cognition*, 108, 136-154.

Sargent, J., Dopkins, S., Philbeck, J., & Chichka, D. (2010). Chunking in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 576-589.

Sargent, J., Dopkins, S., Philbeck, J., & Modarres, R. (2008). Spatial memory during progressive disorientation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 602-615.

Shelton, A. L., & McNamara, T. P. (2001). Systems of spatial reference in human memory. *Cognitive Psychology*, 43, 274-310.

Waller, D., & Hodgson, E. (2006). Transient and enduring spatial representations under disorientation and self-rotation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 867-882.

Wang, R. F., & Spelke, E. S. (2000). Updating egocentric representations in human navigation. *Cognition*, 77, 215-250.

Xiao, C., Mou, W., & McNamara, T. P. (2009). Use of self-to-object and object-to-object spatial relations in locomotion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 1137-1147.

Yamamoto, N., & Shelton, A. L. (2007). Path information effects in visual and proprioceptive spatial learning. *Acta Psychologica*, 125, 346-360.

Yamamoto, N., & Shelton, A. L. (2009). Sequential versus simultaneous viewing of an environment: Effects of focal attention to individual object locations on visual spatial learning. *Visual Cognition*, 17, 457-483.



# Mutual Affects in Computer-mediated Collaborative Learning: Positive Feelings Shared by Collaborators Enhance System Evaluations

Takashi Yamauchi,  
Texas A&M University  
College Station, TX

takashi-yamauchi@tamu.edu

Takehiko Ohno, Momoko Nakatani, Yoichi Kato,  
NTT Cyber Solutions Laboratories  
Yokosuka-shi, Japan

ohno.takehiko@lab.ntt.co.jp

Arthur B. Markman  
University of Texas  
Austin, TX

markman@psy.utexas.edu

## Abstract

The authors employ behavioral theories of human motivation and affect and present an explanation for why some computer-mediated collaborative learning is satisfying for a user. In a longitudinal experiment, participants were divided into four groups and solved two open-ended problems together using a video-conference system. Traditional metrics of usability and product acceptance were examined with respect to psychological variables such as personality, background knowledge, and feelings toward group members (mutual affect). The results show that group-level mutual affect is a strong predictor of system acceptability judgments, even after controlling for other pragmatic variables such as opinion convergence. It is proposed that evaluating one's experience with a computer-mediated collaborative system is a sensemaking process and that the variables that modulate this process also influence subjective judgments of usability and acceptability of a system.

**Keywords:** User satisfaction, user experience, mutual affect

Cultivating positive emotions among collaborators is essential for the success of groupware applications because shared positive affects promote group coordination, common ground, and group awareness—key ingredients for successful online collaboration (Carroll et al., 2006). But what design features are critical to generate positive mutual affects? Do mutual affects influence user experience primarily by elevating pragmatic qualities of group interaction, such as group communication and coordination?

To help improve computer-mediated collaborative learning (e.g., learning collaboratively via video conferencing), researchers have identified important variables, such as group awareness, common ground, shared visual information and teamwork coordination (Carroll et al., 2006). However, to make a “good” collaborative learning system, these pragmatic variables should be supplanted; the product should be not only useful but also *engaging* and *satisfying* for the users (Hassenzahl & Tractinsky, 2006; Norman, 2004). But to make an engaging and satisfying product, it is crucial to know how users come to evaluate their experiences with a collaborative system.

Conceptual frameworks such as information processing, affordance, and cognitive architecture have generated testable hypotheses and guidelines instrumental for single-user product design. However, these pragmatic variables are not entirely feasible in collaborative settings because of

added complexity inherent in group interaction (Grudin, 1994).

Here, we propose a conjecture that psychological models of sensemaking can provide a useful framework for user experience in a groupware setting much in the same way that affordance and cognitive architecture helped the evolution of single-user interfaces. We argue that the evaluation of one's experience is basically a sensemaking process, and the variables that intervene this process influence “user experience.”

To test our framework, we developed an experimental study, where 29 college students were divided into four groups and solved two problems together in a 2-month period using a video-conference system. We examined subjective metrics of usability and product acceptance with respect to other psychological variables such as personality, background knowledge and group-coherence. The results showed that a positive mutual affect among group members led to increased product acceptance, even after controlling for other pragmatic variables such as opinion convergence and communication effectiveness.

## User Experience as a Sensemaking Process

Klein et al. (2006) and Pirolli and Card (2005) provide models of sensemaking. The two models differ in specifics but share some basic properties. Sensemaking consists of dynamic processes of data selection and frame/schema revision. Relevant data are selected according to one's frame (prior knowledge/beliefs/mindsets), the data are interpreted and the frame is revised according to the interpretation. Sensemaking goes through cycles of this data selection/interpretation and frame/schema revision loop. Our central hypothesis is that user experience is a sensemaking process. “Experience” does not come to people unambiguously. Experience is selected, sensed, represented, and interpreted by people (Pirolli & Card, 2005). In this process, affects play critical roles as affect seeps into the evaluation of the data.

**Group-level Mutual Affect** The importance of affect in product design is well known, but affect in human-computer interaction has pertained to a specific product. We think that group-level mutual affect (e.g., feeling of closeness of group members) can also be an important factor because affects are contagious and affects coming from unrelated sources can be easily fused into the evaluation of a product.



Much research has shown that relatively simple manipulations of inducing a positive affect, such as viewing a comedy film for a few minutes or writing about happy events, influence subsequent decision making of unrelated objects (Clore & Huntsinger, 2007). Schwarz and Clore (1983) present one of the most stunning demonstrations of affect contamination. In their experiment, the researchers interviewed subjects about their general happiness with their lives. Subjects were selected randomly and telephone-interviews were conducted on either a sunny day or a rainy day. Those who had an interview on a sunny day gave higher happiness ratings than those who had an interview on a rainy day. When the link between mood and weather was made clear to subjects, the ratings made on the rainy day went up, indicating that subjects' ratings about happiness were partly due to their erroneous generalization of their unhappy mood on the rainy day.

A similar misattribution is likely to happen in the judgment of usability and acceptability. Usability and acceptability of a product will be judged by pragmatic, hedonic, and aesthetic features of the product (Hassenzahl & Tractinsky, 2006). However, in making an actual judgment, a user will *interpret* his/her memories of experience. In this process, affective experience with group members can contaminate their evaluation (Clore & Huntsinger, 2007).

In the experiment described below, we examined the extent to which mutual affects formed among collaborators influence their usability and acceptability judgments of a video-conference system.

## Experiment

In our 2-month-long experimental study, four groups of participants (seven to eight participants per group) met eight times using MeetingPlaza, a multi-party Web conferencing and collaboration system (<http://www.meetingplaza.com>); each group worked together to solve two different open-ended problems (i.e., how to improve the university and recommendations for freshman job search) [15], and wrote two one-page white paper proposals together as a group using MeetingPlaza.

MeetingPlaza has web-, file-, whiteboard- and application-sharing functions that facilitate collaborative communication. For example, the web-share function allows participants in different locations to view the same web site on their own computers at the same time. The file-sharing and application-sharing functions help people in remote locations to view and manipulate the same file together (e.g., an MS Word file). The participants were encouraged to write papers together using these sharing functions.

**Hypothesis.** On the basis of the theoretical background discussed previously, we formed the following hypothesis. Group-level mutual affect (e.g., feelings of closeness toward group members) influences subjective judgments of system usability and acceptability. In particular, those who have high positive group-level affect should give high acceptability and usability ratings even when other group-

level variables such as opinion convergence and communication effectiveness are controlled for.



Figure 1. MeetingPlaza.

conference system, as a user makes a system evaluation based on his/her memory of the experience with a product. Thus, positive mutual affect will be translated into positive product evaluation.

**Participants.** Thirty-two participants were recruited from the Texas A&M University community. They were assigned randomly to four groups. Three participants chose not to take part in the experiment after the first meeting. Thus, a total of 29 participants completed the two problem solving sessions (Table 1). Participants received a payment of \$144 (\$12 per hour for a total of 12 hours for their involvement). Bonus payments of \$24–\$48 were made to group members who produced the best and second-best white papers. In a separate experiment, 47 undergraduate students were recruited from the university psychology subject pool for the evaluation of the white papers submitted by the four groups.

Table 1. Participant information

N=	29	Major: psychology = 16; public health = 2; political science, history, general studies, telecommunication, management, accounting, industrial engineering, electrical engineering, chemical engineering, nursing, nutrition = 1
(Male, Female)	(14, 15)	
Freshman	1	
Junior	5	
Sophomore	6	
Senior	12	
Graduate student	4	
Staff	1	
Average age	21.1	
Note. Three participants dropped after the first meeting and the data from 29 participants were analyzed.		

*Note.* The participants received a payment of \$144. Bonus payments of \$24–\$48 were made to group members who produced the best and second-best white papers.

**Materials.** We employed five questionnaires to assess participants' mindsets (implicit beliefs on intelligence, morality and world), personality (neuroticism, extraversion and psychoticism), technological literacy (computer-literacy and Internet-literacy), and expectations (expected ease of use and expected usefulness of the product). These questionnaires, which were given at the orientation meeting, were adopted to isolate the effect of group-level mutual affect as much as possible.

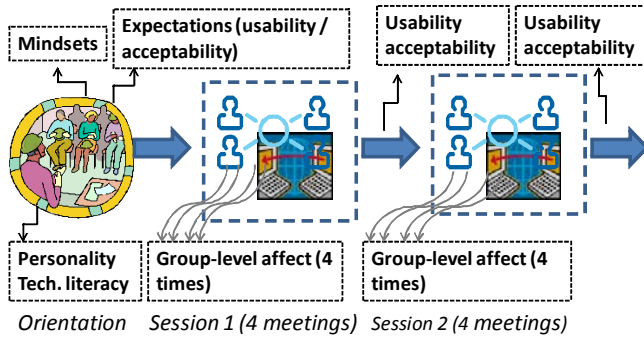


Figure 2. The logistics of the experiment.

Group-level coherence (affect, communication and opinion) was measured by electronic questionnaires given at the end of each meeting. Participants' subjective judgments of system usability and acceptability were collected three times in a two-month period, before using the system (at the orientation meeting), in the middle of using the system (after the fourth meeting), and at the end of the experiment (after the eighth meeting) (Figure 2). Participants' subjective judgments of system usability and acceptability were collected three times in a two-month period, before using the system (at the orientation meeting), in the middle of using the system (after the fourth meeting), and at the end of the experiment (after the eighth meeting) (Figure 2).

Below we explain the questionnaires used in the experiment.

**Implicit belief.** The implicit belief questionnaire assesses the extent to which people conceptualize intelligence, morality, or the world as a dynamic or fixed construct (Dweck, 1999). This questionnaire was included because people's implicit beliefs are known to influence their goal setting and learning experience.

**Personality.** Francis et al. (1992) developed an abbreviated version of the Eysenck personality questionnaire (EPQR), which has four dimensions (extraversion, neuroticism, psychoticism, and lie scale) with six questions each. The questionnaire assesses personality of a person with three dimensions, extraversion (high-low tendency to seek external stimulation), neuroticism (high-low level of negative affect), and psychoticism (high-low level of impulsiveness). Following the suggestion by Francis et al. (1992) we did not analyze lie-scale scores in the present experiment.

**Technology literacy.** The technology literacy questionnaire was developed for this experiment based on the digital literacy questionnaire by Hargittai (2009). Our questionnaire consists of two categories, computer literacy and Internet literacy, and a total of 10 questions. The computer literacy measure has four items related to knowledge about software (e.g., PowerPoint) and programming language (e.g., Java). Internet literacy consists of six items related to common Internet-based activities (e.g., tweeting or on-line shopping).

**Expected ease of use and expected usefulness of the product.** We modified Davis's Technology Acceptance

Model (TAM) (Davis, 1989) and developed questionnaires assessing expected ease of use and expected usefulness of the product (six questions for each). We included the term "expected" because our questionnaires were given shortly after MeetingPlaza was introduced to the participants but before they actually used the system.

**Usability.** We employed Lewis's Computer System Usability Questionnaire (CSUQ; 19 questions) (Lewis, 1995). The CSUQ consists of three factors, system usefulness, information quality, and interface quality. The pre-usability questionnaire was given at the orientation meeting shortly after participants were introduced to the system but before using the system. The post-usability questionnaire was given twice at the end of session 1 and at the very end of the experiment.

**Acceptability.** To measure participants' behavioral intention of adopting the video-conference system, we created 10 acceptability questions based on Davis et al., (1989) and Venkatesh and Morris (2003). These questions assessed participants' intention to continue to use MeetingPlaza if the system were made available to them.

**Group-level coherence.** We measured group-level coherence of individual members with three dimensions, mutual affect (e.g., how close do you feel with each member of your group?), opinion convergence (e.g., how close was your opinion with that of each member of your group?), and communication effectiveness (e.g., how effectively did you communicate with each member of your group?). Every participant rated how he/she felt about each group member at the end of every group meeting (a total of eight meetings), the ratings he/she gave to all group members were averaged over affect, communication effectiveness and opinion convergence dimensions, and these average values were treated as his/her group-level affect, communication effectiveness and opinion convergence (Strauss, 1997).

**Procedure.** The experiment consisted of four segments: orientation, session 1, session 2, and paper evaluation. Below, we describe the segments in chronological order (Figure 3).

**Orientation.** The orientation meeting was held in a large classroom. First, participants indicated their implicit beliefs, personality and technology literacy, and then the experimenter introduced MeetingPlaza. At this stage, participants were allowed to view MeetingPlaza, but they were not allowed to use the system. After this brief instruction, participants indicated their expected ease of use and expected usefulness of MeetingPlaza, along with their expected usability of the product (pre-usability) and their intention of using the product in the future (acceptability).

**Sessions 1 & 2.** Approximately 1 week after the orientation meeting, participants were assigned to four groups, and each group had its first on-line meeting using MeetingPlaza. In this segment, participants received extended instruction and demonstrations of MeetingPlaza functions and tested MeetingPlaza by themselves. Each group met twice a week, and discussed solutions for the assigned problems using MeetingPlaza. In one session,

participants as a group were required to write a one-page white paper describing ways to improve the university; in the other session participants as a group were required to write another white paper describing recommendations for job search for college freshmen. Each group was required to submit a paper at the end of the fourth meeting of each session. Each meeting lasted about 1 hour.

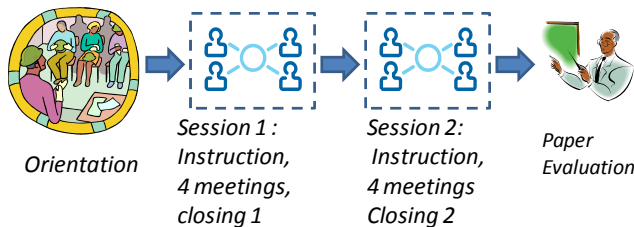


Figure 3. Four segments of the experiment

Session 2 was given 1 week after the end of session 1. The procedure of session 2 was identical to that described in session 1.

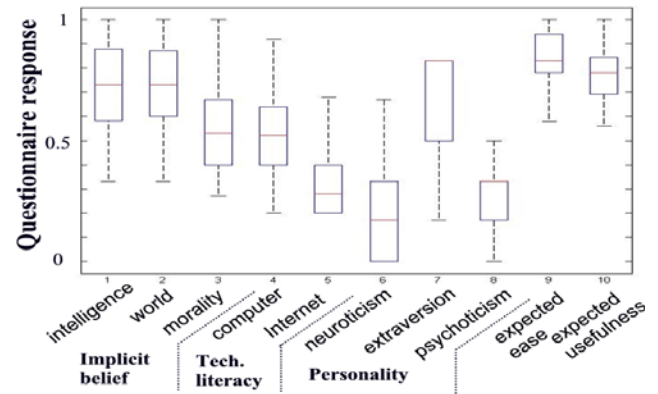
**Closings 1 & 2.** Two closing meetings, closing 1 and closing 2, were held at the end of sessions 1 and 2, respectively. In closings 1 and 2, participants filled out the usability and product acceptance questionnaires. Closing 2 was the final meeting.

**Paper evaluation.** In a separate experiment, 47 undergraduate students participated in the paper evaluation experiment (male=23, female=23, unknown=1) and rated the eight papers written by the four groups in six categories (creativity, implementation, coherence, effectiveness, cost, and communication) on a 0–100 scale. They were encouraged to rate the papers in the same way a professor grades their papers in a classroom.

## Results

All questionnaire responses were converted to a 0-1 scale such that the direction of the observed scores corresponded to the direction of the psychological dimensions in question. Thus, a high score corresponded to a high degree of the given dimension. This section begins with a summary of questionnaire responses, followed by a description of the longitudinal shifts of usability, product acceptability and group-coherence, and concludes with statistical analyses that examine the relationship between group-coherence and system evaluation.

**User profiles.** The responses on the 10 dimensions of the questionnaires (Figure 4) show that there were no ceiling or floor effects, except for the responses regarding expected ease of use and expected usefulness. This problem will be discussed in the next paragraph and later in the Results section. ANOVAs (analysis of variance) comparing the four groups in each of the ten user profile dimensions showed that the mean profile scores of the four groups were not statistically different:  $F(3, 25) < 2.2$ ,  $p > 0.11$ .



Note. The central mark and the edges of a box are the median and the 25<sup>th</sup> and 75<sup>th</sup> percentiles, respectively. The whiskers are the most extreme data points.

Figure 4. Boxplots for questionnaire responses.

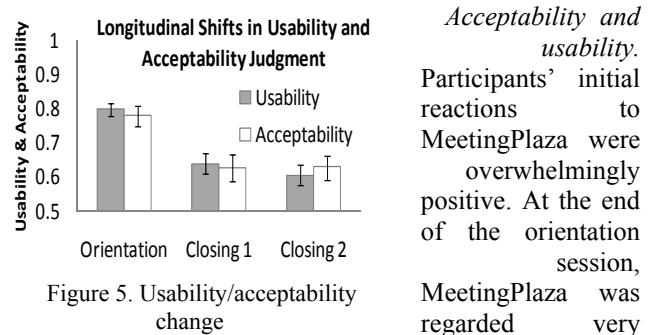


Figure 5. Usability/acceptability change

Participants' initial reactions to MeetingPlaza were overwhelmingly positive. At the end of the orientation session, MeetingPlaza was regarded very favorably (usability,  $M=0.80$ ,  $SD=0.11$ ; acceptability,  $M=0.78$ ,  $SD=0.17$ ; Figure 5). However, the ratings of MeetingPlaza dropped significantly at closing 1 (usability,  $M=0.64$ ,  $SD=0.64$ ; acceptability,  $M=0.63$ ,  $SD=0.21$ ) and closing 2 (usability,  $M=0.61$ ,  $SD=0.18$ ; acceptability,  $M=0.63$ ,  $SD=0.22$ ). Two 3 (sequence: orientation, closing 1, closing 2)  $\times$  4 (group: 1, 2, 3, 4) ANOVAs revealed that this drop occurred uniformly in all groups: usability,  $F(2, 50)=18.61$ ,  $MSE=0.02$ ,  $p<0.01$ ; acceptability,  $F(2, 50)=9.70$ ,  $MSE=0.02$ ,  $p<0.01$ . Neither the main effect of group nor the interaction between sequence and group was observed in both usability and acceptability measures:  $F$ 's  $< 1.5$ ,  $p > 0.24$ . These results suggest that MeetingPlaza created a positive impression on the college-age participants but the excitement dropped considerably once the participants used the product for problem solving, indicating that using the collaborative video-conferencing system was much more challenging than anticipated.

**Longitudinal shifts of group coherence.** The group-coherence scores (mutual affect, opinion convergence, communication effectiveness) all increased as the collaborative sessions progressed (Figure 6). Three sets of linear contrast analyses (shift; beginning, middle, end of the experimental sessions)  $\times$  (group; 1-4) applied to the three group-coherence measures revealed significant linear upward trends: communication effectiveness,  $F(1, 25)=7.72$ ,  $MSE=.01$ ,  $p<0.05$ ; opinion convergence,  $F(1, 25)=26.2$ ,  $MSE=.01$ ,  $p<0.001$ ; mutual affect,  $F(1, 25)=35.8$ ,

MSE=.01,  $p<.001$ , suggesting that our online meetings were indeed effective in developing a sense of common ground, better communication, and positive feelings. There was no interaction effect of group and shift:  $F's<1.0$ .

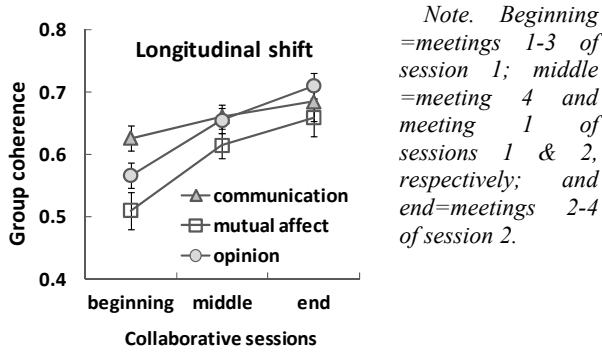


Figure 6. Longitudinal shifts of group coherence

**Evaluation of the hypothesis.** Regression analyses were employed to investigate the link between group-level affect and system evaluation. Both step-wise regression and regular multiple regression were adopted to ameliorate the problem of multicollinearity. In the step-wise regression procedure, the forward selection method was applied with the entry criterion of 0.1 to ensure that all relevant predictors were included in the regression models.

The step-wise regression analyses using group coherence variables suggest that mutual affects influenced system acceptability scores significantly;  $\beta=.52$ ,  $p<0.01$ ,  $R^2=.27$ ; but other group coherence variables—opinion convergence and communication effectiveness—did not influence acceptability scores,  $p>.1$ . The impact of the mutual affect variable remained strong on the system acceptability measure even after controlling for the effects of all other personal variables [mindsets (intelligence, morality, world); personality (neuroticism, extraversion, psychoticism), technology literacy (computer-literacy, Internet-literacy)];  $\beta=.48$ ,  $p=.05$ ; but not the usability scores;  $\beta=.35$ ,  $p=.15$ .

The results from multiple regression analyses were analogous to those found in the step-wise regression analyses. Even after the communicative variables—communication effectiveness and opinion convergence—were forced into the models, the strongest predictors were still mutual affect; the correlation between mutual affect and acceptability was significant after the effects of communication effectiveness and opinion convergence were partialled out ( $r=0.45$ ,  $p<0.05$ ).

**Cohort effects.** The predictor, group-level mutual affect, was evaluated with the data obtained from individual participants. Because MeetingPlaza is a collaboration tool, the impact of this variable should be scrutinized with respect to the properties obtained from each group. For this reason, we employed hierarchical linear regression models and estimated the beta coefficients of the mutual affect variable for each group and investigated if the impact of mutual affects remain robust after controlling for other group-specific properties—the ratio of female and male

participants in each group and the white paper evaluation score that each group received (Table 2).

Table 2. Hierarchical Linear Regression Model

Individual layer:

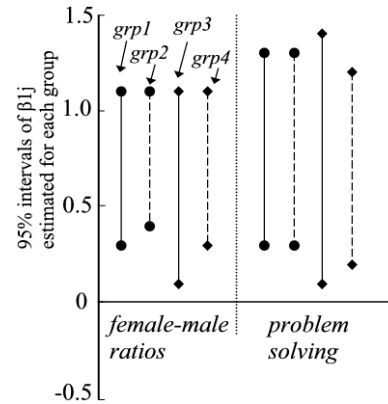
$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + u_{ij}$$

Group layer:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}$$

Note. Subscript  $j$  denotes group ID and  $Y_{ij}$  represents the acceptability score obtained from participant  $i$  of group  $j$ .  $\beta_{0j}$  and  $\beta_{1j}$  are intercepts and slopes of the regression line of group  $j$ , respectively.  $u$ 's are error terms.  $\gamma_{00}$  is the overall mean of the acceptability scores and  $\gamma_{10}$  is the mean of the slopes of the four groups.  $W_j$  represents group-specific values (e.g., either the female-male ratio of group  $j$  or the problem solving score of group  $j$ ) and  $\gamma_{11}$  is the coefficient for predictor  $W_j$ .



Note. The problem-solving performance was measured by the average paper evaluation score that each group received in a separate experiment.

Figure 7. 95% high density intervals of  $\beta_{1j}$  estimated for each group

Our hierarchical models had two layers, individual and group (Table 2). The model assumes that the coefficients  $\beta_{1j}$  of mutual affects ( $X_{ij}$ ) vary for each group and are modulated by a group-specific properties (problem-solving scores or female-male ratios). Note that the “group-level mutual affect” variable  $X_{ij}$  was included in the individual layer

because the values of this variable were calculated for individual participants. The values of group-specific variable  $W_j$  (e.g., problem-solving scores or female-male ratios) were calculated for each group, not for each participant. Because we had only four groups, the intervals of  $\beta_{1j}$  were estimated by a Bayesian method where the coefficients ( $\gamma$ ) in the group layer were treated as non-informative hyper-parameters and posterior distributions were obtained by the Markov Chain Monte Carlo algorithm with 1000 iterations (Gelman & Hill, 2007).

The results, which are summarized in Figure 7, suggest that even after the two group-specific properties (female-male ratios and problem-solving scores) were taken into account, the impact of mutual affect remained robust, as the 95% high density intervals of the coefficients  $\beta_{1j}$  were above 0 in all cases, suggesting that the effect of mutual affect occurred on top of the group-specific properties.



## Discussion

In computer-mediated collaborative learning, the focus has been to enhance pragmatic functionality of the system. The present study shows that fostering positive emotions among collaborators is no less important. The results suggest that mutual affects shared among collaborators influence the evaluation of product acceptability even after personal variables, such as personality and background knowledge, were taken into account, implying that the influence of mutual affect on a video-conferencing system is far reaching than previously thought. Because the effect of mutual affect was stronger than that of pragmatic variable such as group-level communication and opinion convergence, it is likely that mutual affects were fused into users' experience with the system.

Note that the fact that affective experience can be misattributed does not mean that affect is irrelevant in enhancing the functionality of a collaborative learning system. Positive emotions can unleash creative and flexible thinking (Isen, 2008) and shared feelings have a multiplicative effect on collaborators because emotions are highly contagious.

A considerable progress has been made in the area of affective computing of intelligent tutoring systems, primarily thanks to the pioneering studies by D'Mello, Graesser, and Conati (Conati & Maclaren, 2009; D'Mello et al., 2007). We suggest that similar affect detection technologies help advance groupware applications. In a large computer-based collaborative situation, it is difficult to assess participants' affective states in real time. A collaborative groupware system that can trace users' affective state can facilitate group participation and learning.

## Acknowledgments

This study was supported by a grant from NTT Cyber Solutions Laboratories.

## References

- Carroll, J. M., Rosson, M. B., Convertino, G., & Ganoë, C. H. (2006). Awareness and teamwork in computer-supported collaborations. *Interacting with Computers*, 26, 21-46.
- Clore, G. L., & Huntsinger, J. R. (2007). How emotions inform judgment and regulate thought. *Trends in Cognitive Sciences*, 11, 393-399.
- Conati, C., & Maclaren, H. (2009). Modeling User Affect from Causes and Effects. In G.-J. H. et al. (Ed.), *UMAP 2009, LNCS 5535* (pp. 4-15). Berlin Heidelberg: Springer-Verlag.
- D'Mello, S., Graesser, A., & Picard, R. (2007). Towards an Affect-Sensitive AutoTutor. *IEEE Intelligent Systems*, 22, 53-61.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13, 319-340.
- Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), 982-1003.
- Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality and development*. Ann Arbor, MI: Psychology Press.
- Francis, L. J., Brown, L. B., & Philipchalk, R. (1992). The development of an abbreviated form of the revised Eysenck Personality Questionnaire (EPQR-A): Its use among students in England, Canada, the U.S.A. and Australia. *Personality and Individual Differences*, 13, 443-449.
- Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.
- Grudin, J. (1994). Groupware and social dynamics: Eight challenges for developers. *Communication of the ACM*, 37(1), 92-105.
- Hargittai, E. (2009). An update on survey measures of web-oriented digital literacy. *Social Science Computer Review*, 27(1), 130-137.
- Hassenzahl, M., & Tractinsky, N. (2006). User experience – research agenda. *Behaviour & Information Technology*, 25, 91-97.
- Isen, A. M. (2008). Some ways in which positive affect influences decision making and problem solving. *Handbook of emotions* (pp. 548-573). New York: Guilford Press.
- Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, 21(5), 88-92.
- Lewis, J. R. (1995). IBM computer usability satisfaction questionnaire: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7, 57-78.
- Norman, D. A. (2004). Introduction to this special section on beauty, goodness, and usability. *Human-Computer Interaction*, 19, 311-318.
- Pirolli, P., & Card, S. K. (2005, May). *The sensemaking and leverage points for analyst technology as identified through cognitive task analysis*. Paper presented at the International Conference on Intelligence Analysis, McLean, VA.
- Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45, 513-523.
- Straus, S. G. (1997). Technology, group process, and group outcomes: Testing the connections in computer-mediated and face-to-face groups. *Human-Computer Interaction*, 12, 227-266.
- Venkatesh, V., & Morris, M. G. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27, 425-478.

# Implicit Learning: A Demonstration and a Novel SRT Paradigm

Fayme Yeates<sup>1</sup> (fy212@exeter.ac.uk)

F.W. Jones<sup>2</sup> A.J. Wills<sup>1</sup> M.R.F. Aitken<sup>3</sup> I.P.L. McLaren<sup>1</sup>

<sup>1</sup>School of Psychology, College of Life and Environmental Sciences,  
University of Exeter, UK.

<sup>2</sup>School of Psychology, Canterbury Christ Church University

<sup>3</sup>School of Psychology, University of Cambridge.



## Abstract

Evidence for human learning without awareness of what is learned has been sought in serial reaction time (SRT) tasks in which, unknown to participants, the locations of stimuli follow a particular rule or sequence (Willingham, Nissen & Bullemer, 1989). A number of criticisms have been levelled at such tasks, including a lack of adequate control for sequential effects and a discrepancy in sensitivity between measures of implicit and explicit knowledge about the task (Jones & McLaren, 2009; Shanks & St. John, 1994). In this study we provide a novel, two-choice SRT paradigm whereby the locations of the response stimuli are sometimes predicted by a separate set of stimuli on screen. A color-filled square appears before each stimulus requiring a response, with participants informed this is simply a fixation point to prepare for the next trial. Two out of eight colors are predictive on 80% of trials, and performance on these consistent trials was faster than on the other six colors that were equally likely to result in either of the two possible responses. All these trial types were faster and more accurate than the remaining inconsistent 20% of trials for the predictive colors, which also produce more errors than control colors. A prediction task and interview followed the task, on which participants performed at near (slightly below) chance levels. We suggest that this task is a useful tool for studying associative learning in humans, as it provides reliable effects that appear to demonstrate implicit learning with relatively brief training.

**Keywords:** Associative learning; implicit learning; SRT task

## Introduction

In 1994 Shanks and St. John published their seminal review of implicit learning in *Behavioral and Brain Sciences*. Their conclusion was that implicit learning in humans had not yet been conclusively demonstrated. One paradigm that seemed to hold some promise in this regard was the SRT task pioneered by Willingham, Nissen and Bullemer (1989). In this task participants had to simply respond to one of four possible stimulus locations by pressing a spatially compatible key. Unknown to them, the sequence in which the stimulus locations were presented was not random but instead repeated a particular ordering. Willingham et al found that participants trained on such a sequence learned to respond faster than controls, yet were unable to say much about the sequence or, importantly, to predict the next stimulus location given a number of preceding trials. Shanks and St. John analysed this result, and others like it, and concluded that whilst the prediction task was potentially a sufficiently sensitive method of assaying participants' knowledge of the sequence, the evidence suggested that the non-significant findings on the

prediction task had more to do with a lack of power than a lack of knowledge. Subsequent reviews of more recent evidence (e.g. Lovibond and Shanks, 2002; Shanks and Lovibond, 2002; Mitchell, De Houwer and Lovibond, 2009) have seen no reason to change this conclusion.

In this paper we attempt to provide the evidence needed to prove that learning can take place without awareness of what is being learned. We do this using a variant of the SRT task employed by Jones and McLaren (2009) developed by Aitken (1996) which, instead of using the preceding trials to determine the later ones, uses a separate stimulus to predict which of the location stimuli will be presented next. A colored square presented between two stimulus locations is the predictor stimulus, and which of the two stimulus locations changes from an open to a filled circle is what determines the response required. If the left circle fills, then a left key press is needed and if the right-hand circle fills then a right key press is appropriate. The colored square is presented just before a response is required (while both circles are unfilled) so that it can predict which response will shortly be needed. Rather than run a separate control group, we follow Cleeremans (1993) in using a within-subject control. Thus some colors are never predictive (they have a random relationship with the stimulus location) whereas others are correlated with the response needed on the next trial.

The detailed design of this paradigm was dictated by our assumption of a dual-process mechanism for human learning. Following McLaren, Green and Mackintosh (1994), we conceptualized learning as being driven by two quite different sets of processes, one Cognitive in nature, employing conscious, controlled, rule-based symbolic processes, the other Associative in nature, automatic in operation and based on simple algorithms that capture the correlations between events. The challenge, then, was to arrive at a paradigm that would readily allow learning on an associative basis but would be much less amenable to rule-based cognition. The Aitken SRT paradigm allows for a parameterization that meets these requirements. We used 8 colors in total, but made only 2 of them predictive (one left, one right). Thus, 6 out of the 8 bore no relationship to the next stimulus location to occur. Following the technique used by Posner and Snyder (1975), the 2 predictive colors were themselves only 80% reliable. Hence, if one of them typically predicted a left response would be the next needed, 20% of the time it was followed by a stimulus location that

required a right response. Our rationale for doing this was that it would make conscious detection of the contingencies in play much harder. Due to 6 of the 8 colors being non-predictive, the overall prediction rate possible in this experiment is a mere 57.5% if complete knowledge of the contingencies is assumed, and no stimulus is an entirely reliable predictor of anything. In these circumstances, we expected it to be very difficult to notice which colors had some predictive value. Any attempt to "work out" what was going on in this fast paced task, with many different colors involved, would overtax working memory and severely impair performance on the main, SRT task, which should lead to this type of strategy being quickly rejected.

We also used the fact that there were 6 non-contingent colors to configure a design in which half the blocks were control blocks, and the other half experimental. The latter included the 2 contingent colors and 2 of the non-contingent, whereas the control blocks had the other 4 non-contingent colors in them. Control and Experimental blocks were alternated. This meant that for much of the time our participants were not experiencing any contingency between the colored square and the stimulus location / response required. It also allowed us to define two, somewhat different, within subject controls for our experimental manipulation. The first are the control colors in the experimental blocks: these have the advantage of sampling our participants' behavior under similar circumstances to those in force for the contingent, experimental colors. The second are the colors in the control blocks: these have the advantage of sampling behavior over an entire block without any distortions caused by the contingent colors. The control blocks were given exactly the same sequence of right and left responses as the corresponding experimental blocks. Thus, any sequential effects, caused by a particular run of right and left responses, will be controlled for by these blocks. As we shall see, this proved to be a useful aspect of our design.

Despite our efforts to prevent conscious rule-based learning, the parameters chosen for this design should nevertheless support strong associative learning. The colors occur just before the stimulus location is presented and the response required, and feedback is available shortly afterward. Thus, given that there is a reliable contingency between a clearly perceived stimulus and stimulus location / response, it should be easily learned. We ensured that the colors were presented boldly at fixation and gave our participants good reason to be looking there at the start of each trial. If this type of learning is automatic, and stimulus specific, then there should be little difficulty in learning the association between the two predictive colors and their correlated stimulus locations / responses.

## **Method**

### **Participants**

The study was conducted with 32 participants, randomly divided into two groups (n=16 in each) who performed the task with slightly different parameters (different distances

between the stimuli on screen). The participants were all first year psychology students at the University of Exeter, aged from 18 to 22 years old, who were rewarded with a course credit in return for their participation.

### **Materials**

An Apple iMac computer was used to run the experiment, with participants seated roughly 50cm from the screen. The display during blocks contained two white, outline circles and a white, outline square on a black background. The circles were 1.9cm in diameter, and the square was 1.9cm in width. The square was positioned directly in the centre of the screen for both groups. One group saw the circles positioned 2.2cm to the right and left of the centre of the screen, consistent with the distance separating the two-choice SRT task stimuli in Jones & McLaren (2009). The other group saw the same size white circle outlines placed 7.5cm from the right and left of the centre of the screen, following the stimulus locations specified in Aitken (1996).

The predictor stimulus was a colored filled square 1.9cm in width that filled in the white square outline. A choice of eight possible colors: red, green, blue, yellow, pink, orange, brown and teal were used. The response signal was a white filled circle 1.9cm in diameter that replaced either the right or left outline circle during the trials. The participants were instructed to press the spatially compatible "x" key on a QWERTY keyboard if the target stimulus appeared on the left, and the ">" key if the stimulus appeared on the right.

### **Design**

The experiment consisted of a two-choice SRT task conducted over one session that lasted approximately an hour. The task consisted of 20 alternating blocks, half of which were control blocks and half experimental, each with 120 trials. The ordering of the blocks was counterbalanced across participants. On both experimental and control blocks participants received a random ordering of equal amounts of right and left response stimuli. In the experimental blocks participants saw four colors (red, green, blue, yellow) as the filled square predictor stimulus. One of these colors preceded a right circle fill with an 80% contingency, and another color appeared before the left circle filling 80% of the time. The other two colors in the experimental block occurred with equal probability before a right or left circle fill. The colors were randomized for each participant and the experimenter was blind to which colors were predictive, and which response they predicted. During control blocks, all four colors (pink, orange, brown, teal) were equally likely to occur before a right or left response signal. Each color that could appear in a block appeared on an equal number of trials in that block, and all trials were constructed so that a color would not repeat itself on the next trial.

### **Procedure**

Participants were instructed to fixate on the colored square and then to simply respond as quickly and accurately as possible to the white circle fills. They were told to use the



colored square to bring their focus back to the centre of the screen and therefore improve their overall performance on the task by avoiding bias to one side or the other. They were informed that the colors changed to attract their attention back to the centre of the screen: no mention was made of any contingencies or the role of the colored square as a predictive or instructive stimulus.

On each trial the square would fill with the stimulus color and remain on screen until the computer detected a response. A variable interval in the range of 250-500ms would then occur before the white circle response signal appeared on screen. Reaction time was measured from this stimulus' appearance on screen until a key press was detected. A 250ms response-stimulus interval (RSI) then followed before the appearance of the next colored square stimulus during which the screen was blank except for the square and circle outlines. On each trial the stimuli remained on the screen until the participant had responded or was timed-out for not having pressed a key within 4.25s.

If participants pressed an incorrect key or were timed-out then the trial terminated and the computer issued a short 'beep' sound. Following each block, participants rested during a 30 second break in which they were shown their average reaction time (in milliseconds) and their accuracy (as a percentage) for the block of stimuli just completed. They were also informed whether these scores were better or worse than those from the previous block. At the end of the 20<sup>th</sup> block, participants were instructed to fetch the experimenter.

A short verbal interview asking general questions about the task was given, and participants were then assessed to determine the extent of their knowledge about the contingencies during the experiment. They were first asked to describe any relationship that they had noticed between the colors and the circle fills. Then a prediction task was conducted that closely followed the procedures employed in the main task they had just experienced. Thus, participants were instructed to attend to a colored square at fixation in the middle of two outline circles, but now had to make a prediction about which circle they thought would fill by pressing the appropriate key. They were told that the circles would not fill, and that no response would be considered an error, apart from pressing keys other than the two response keys, and that there was no time limit on making their choice. The prediction task was presented as two short blocks of 16 trials per block, with each color from both control and experimental blocks occurring twice in each block in a randomized order.

## Results

Results were computed for both errors and RTs, with four Stimulus Types compared across Blocks and at the different Distances between the two response signals. Those colors that had an 80% contingency with a right or left circle fill (the Experimental colors) are split into both their Consistent and Inconsistent presentations. For example, if a red square preceded a right circle fill on 80% of the trials, those

responses would be Consistent, whilst when a red square was followed by a left circle fill on the other 20% of trials this would be Inconsistent with the trained contingency. The two other colors in the experimental block form one control set (denoted Control – Experimental Block) for within-subject comparison, as do the four control colors from the control blocks (Control – Control Block). Trials following an error were excluded from both RT and error analyses. The RT and error data are shown in Figures 1 and 2, respectively.

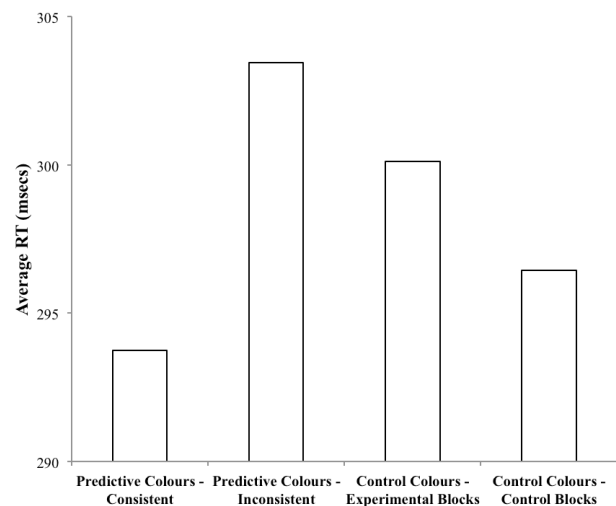


Figure 1. Average RT in msecs for different Stimulus Types.

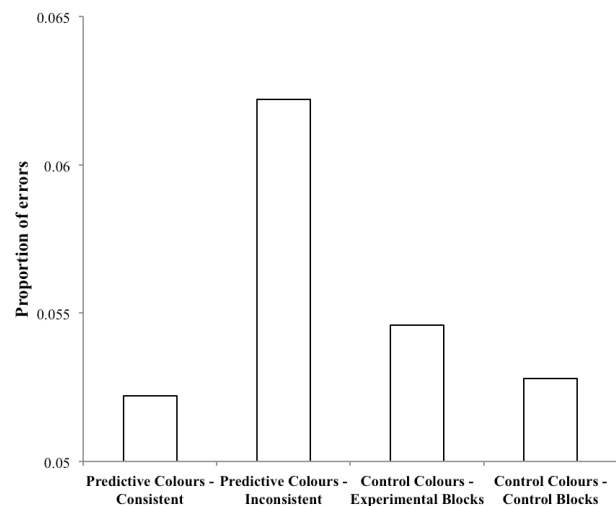


Figure 2. Proportion of errors for different Stimulus Types.

ANOVAs were conducted on both RT and proportion of errors, with Block, Distance and Stimulus Type as the independent variables. Blocks were collapsed into fours for the analyses, so each Block comprised two Experimental blocks and two Control blocks. The Distance of the stimulus did not have a main effect on the RTs,  $F(1,30) = .27$ ,  $p = .61$ , nor on the errors,  $F(1,30) = .56$ ,  $p = .46$ . Distance did not interact with any other variable, and therefore the location of the response stimuli was not a significant factor in determining learning of the color contingencies.

Figure 1 shows the average RT and proportion of errors for the four Stimulus Types, the main effect of Stimulus Type was significant for both the RTs,  $F(3,90) = 17.45, p < .001$ , and errors,  $F(3,90) = 3.72, p = .014$ . Both RTs and errors follow the same ordinal pattern, which was examined in more detail with a set of planned contrasts. Faster responses and fewer errors were made on Experimental Consistent trials. The difference between Experimental Consistent and Inconsistent trials is significant for both RTs,  $F(1,30) = 33.15, p < .001$ , and errors,  $F(1,30) = 5.463, p = .026$ , with Inconsistent stimuli responded to more slowly and with lower accuracy. Experimental Consistent stimuli are also responded to faster than both Control stimuli from Experimental,  $F(1,30) = 34.46, p < .001$  and Control blocks,  $F(1,30) = 6.173, p = .019$ , this trend is also apparent in the errors, but it is significant only in the RTs.

Experimental Inconsistent trials are slower than Control stimuli from Experimental blocks in RTs,  $F(1,30) = 4.85, p = .035$ , and show a trend toward more errors,  $F(1,30) = 3.34, p = .078$ . Experimental Inconsistent trials are also slower and less accurate than Control stimuli from the Control blocks, significantly so in RTs,  $F(1,30) = 14.51, p = .001$ , and again demonstrating a trend in the errors,  $F(1,30) = 4.05, p = .053$ . It seems therefore that learning of the contingencies occurred, with performance averaged across the experiment showing faster and more accurate responses to those trials that were predicted 80% of the time by the preceding color. This is further highlighted by the decrement in performance for those trials inconsistent with the learnt relationship, causing participants to make more errors and respond slower to these stimuli than both the Consistent stimuli and the Control stimuli.

The Control stimuli are interesting, in that they significantly differ from one another in the RTs,  $F(1,30) = 9.16, p = .005$ , although not in the errors,  $F(1,30) = .62, p = .44$ . Control stimuli from the Experimental blocks are responded to more slowly than Control stimuli from the Control block. We suspect that this is due to the presence of the Consistent and Inconsistent stimuli in the Experimental Block. It is not that more errors are being made (overall) in the Experimental blocks, instead, it may be that the conflict that occurs on inconsistent trials engages mechanisms that produce more cautious responding, though we note that this cannot simply be a speed-accuracy trade-off as the error rate is, if anything, lower for the Control block stimuli than for Control stimuli from the Experimental Blocks.

Block has a significant main effect in RTs,  $F(4,120) = 9.26, p < .001$ , and errors,  $F(4,120) = 19.02, p < .001$ . With practice, participants got faster at the task but exhibited something of a speed-accuracy trade-off, making more errors progressively throughout the experiment. Block did not interact with Stimulus Type in the RT analysis, thus Experimental Consistent stimuli in Experimental Blocks were responded to faster than Control stimuli in Control Blocks, which were in turn responded to faster than Control stimuli in Experimental Blocks, and all are faster than Experimental Inconsistent stimuli. The relative positions of

Inconsistent and Consistent stimuli are reversed in Block 1 in the errors, however, and the interaction between Block and Stimulus Type for errors is significant,  $F(12,360) = 2.68, p = .002$ . Thus, in this case, the Consistent / Inconsistent difference emerges over blocks.

The structured questionnaire and prediction task were completed by half of the participants, eight at each Distance value. The other participants were simply asked to pick which colors were predictive: they were invited to select two colors from the eight possible candidates. 14 of the 16 felt able to make the attempt, and out of the 28 responses that were generated 6 were correct (7 would be expected by chance).

The results of the structured questionnaire for the other 16 participants who completed the prediction task show that of the 32 responses, 8 were correct. If we take it that there are two predictive colors, and do not require the participant to remember which predicted left and which predicted right, then if we ask the participants to pick two colors out of the eight, the expected number of colors selected correctly is 0.5. With 16 participants taking part the total number of colors correctly selected is expected to be 8 by chance, and it is exactly 8, approximately evenly distributed across the different responses (right or left) and Distances.

Turning now to the prediction task itself, each participant experienced 32 trials in total. 8 of which involved the two predictive colors. Taking into account the requirement to give the appropriate response for the color, we can expect 4 correct responses by chance to these 8 presentations. Inspection of Figure 3 shows that the distribution of correct responses across participants is approximately binomial at each Distance, and is centered at a mean of 4 or less.

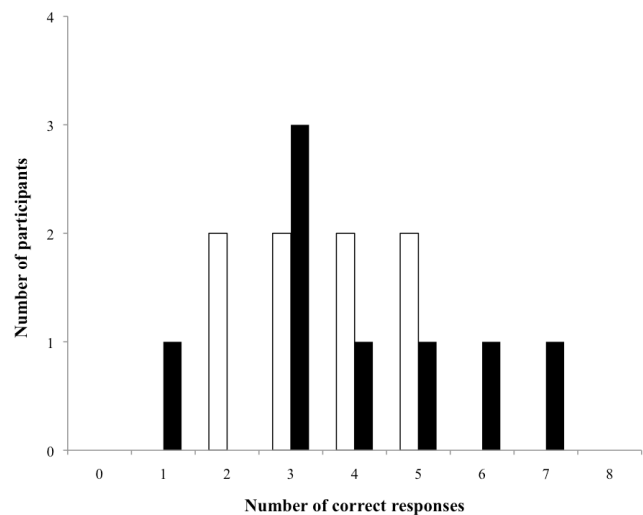


Figure 3. The number of correct responses participants gave for Consistent – Predictive colors. Data presented from 8 participants who performed the task with 2.2cm Distance (open bars) and 8 participants who performed the task with 7.5cm (solid bars) between response stimuli.

The actual means are 3.5 and 4 for Distances of 2.2 cm and 7.5 cm respectively, neither of which differ significantly from chance, and neither of which exceed mean chance

expectation. Our conclusion, based on these data, is that our participants do not know which colors are predictive, and do not know what they predict. This is in accord with their self-reports at the end of the experiment: they claimed to be unaware of any contingency between the color of the central square and the response that followed.

## General Discussion

Our basic claim is that our participants' behavior was affected by the contingencies between color and stimulus location / response in the absence of any conscious knowledge on their part about these contingencies. That their behavior was affected is beyond doubt; the effect is statistically highly significant and, by the standards of these things, rather easy to obtain after relatively little training. We believe that these characteristics make this paradigm a good candidate for studying associative learning in humans. In particular, the highly reliable difference between Consistent and Inconsistent responses to the contingent colors is noteworthy, and our results suggest that this difference is made up of a disadvantage for Inconsistent trials and an advantage for Consistent trials.

Our use of two different types of within-subject control enables us to be confident that the advantage for Consistent trials and the disadvantage for Inconsistent trials are real. They also allow us to speculate about the effect of having these types of trial in a block. In some ways the most appropriate controls are those colors used as controls in the Experimental blocks. They are subject to the same local conditions, in terms of motivation, mood etc., and so provide an appropriate baseline for comparison with Consistent and Inconsistent responses to the contingent stimuli. We included the separate control blocks because we suspected that simply having some predictable trials in a block might alter our participants behavior, and also to provide a control for sequential effects. This last was accomplished by using exactly the same sequence of left and right responses in paired Control and Experimental blocks, the only difference between the two was in what colors were used in the block, and how they were paired with responses. Thus, in a Control block our participants would have exactly the same sequence of responses to make, but no useful information from the colored squares that were presented to help (or hinder) them. The fact that our contingent stimuli also differ from these controls on Consistent and Inconsistent trials gives us considerable confidence that there is no hidden artifact responsible for our results.

The difference between the Experimental control stimuli and the Control block control stimuli is also revealing. We hypothesize that this difference, with Experimental block controls performing worse than Control block controls, is due to the conflict experienced on Inconsistent trials leading to a greater cautiousness on some trials (perhaps best described as some kind of post-stimulus presentation checking process) that does not shift performance in terms of any speed – accuracy trade-off. By this we mean that the

slower RTs do not lead to higher accuracy, but instead reflect an additional “cost” that is incurred in task performance. It is difficult to speculate any further on what the exact nature of this cost might be, but one possibility is that perceptual information specifying which stimulus location had occurred might continue to be gathered post-decision on some proportion of trials as part of the checking process. If this mostly occurred on trials where the decision was the correct one (which given the low absolute error rates of our participants is quite likely), then it would have relatively little impact on the error rate but a significant impact on RTs. This analysis can only apply if the chance of engaging the checking process is unrelated to the probability of making an error, and low enough to result in a low frequency of occurrence on trials that result in an error. In these circumstances our RT measure would be much more sensitive to such a process than the error estimate (because the latter would be based on so few data points), so a result in the RTs alone would be quite possible. All speculation aside, this result is, as far as we are aware, a novel one, and opens up a new line of enquiry that may well be related to recent work by Verbruggen et al (in press).

Now we turn to some consideration of the arguments that may be raised to suggest that our prediction test results do not prove that our participants had no conscious knowledge of the contingencies. The most obvious is that the prediction test results apply to only half the participants. This objection is of no concern, as the effects we report, i.e. the advantage for Consistent vs. Inconsistent, the advantage for Consistent trials relative to both types of Control trial and the disadvantage for Inconsistent trials are all significant if only the 16 participants who were given the prediction test are included. We've included all our data because this gives the best estimate of the pattern of performance on the main task, and because it gives the best estimate of the ability of participants to pick the contingent colors during the structured interview. Another objection that might be raised is that the prediction task itself is flawed, in that it a) takes place after the main task by which time participants have forgotten which are the contingent colors and b) is not sufficiently like the main task to help them recall this information. Our response is that our participants appear to have learned to respond faster to Consistent trials relatively early in the experiment, certainly by the time they reached the 6<sup>th</sup> block, and appear to have had no trouble remembering this information (if that was how they were doing it) from block to block from then on. And our prediction task used the same room, same computer, same display, same stimuli presented in a very similar fashion, it was simply the response requirements and consequent presentation rate (and feedback) that were different to the main task. If our participants consciously knew which were the contingent stimuli, this should have acted as a salient reminder of that information.

Perhaps more troublesome for our analysis is the objection that, whilst our prediction test does attempt to assess the correct information (Shanks & St. John, 1994), it

does not do so sensitively enough. We only gave two blocks of 16 trials each, compared to the 20 blocks of 120 trials each for the main task. Is this really a fair test of our participants' knowledge about the color-response contingencies? To which question our answer has to be "yes". The point here is that, in order to be effective in determining RT to respond to the stimulus location, participants' knowledge must be of the categorical variety if the idea of "conscious knowledge" is to mean anything different to "there's been some change in behavior". They must know that this color predicts this location / response, even allowing for a caveat to the effect that this prediction will not always be valid. In other words, they must know, when faced with the appropriate stimulus input, that some colors are predictive and that others do not seem to be. Only then can conscious knowledge produce the desired pattern of results. If this were the case, then our 8 trials for the contingent colors should be more than adequate to detect this type of categorical knowledge. If they get all 8 correct, then this is significantly better than chance on an individual basis. Even allowing for only some of the participants being aware of the contingencies, the means over all 16 participants who took the test should produce overwhelming evidence of such knowledge if only as many as 4 were aware of the contingencies. But we do not even have means that are greater than chance to report for our prediction test, let alone means approaching significance.

Another possibility is that the effect observed in the RTs might be due to our participants being aware of the contingencies on very few trials during the experiment (and then they forget), and that we have simply not given enough trials in the prediction task to capture any of these trials. This is implausible. Firstly, once the contingencies are explicitly coded – then this should in itself increase the probability of noticing them again until it becomes a frequent occurrence. Secondly, even a rate as low as say 1 trial in every 8 of the predictive colors (i.e. 2.5 on average a block in training) would produce a mean on our prediction test of 4.5 that is significantly different to that observed ( $Z=2.27$ ,  $p<.05$ ). And the effect on these individual trial RTs would have to be large (approx 40msec). There is no evidence for such a bimodal distribution in our data.

It would seem, then, that we have good evidence for implicit performance on our task. It is still possible that we may have not demonstrated implicit learning, if we adopt a position that results from a stringent application of the Shanks and St. John criteria (our thanks to Tony Dickinson for pointing this out). Imagine that during training on some trials our participants became aware of the contingencies, but then forgot this prior to being asked. They would have learned the contingencies, but not implicitly. This learning might then give rise to performance that reflected the learning, but still without conscious knowledge of the contingencies so that they passed our prediction test. This type of implicit performance, based on a form of implicit memory, would explain our results. In the limit, it is almost impossible to discount this type of critique, but we think

that it lacks plausibility when applied to our paradigm. There is not much to learn, just that two colors predict responses, and if participants were to become repeatedly aware of this (so that learning could occur), it's hard to see why they would forget this so easily.

Our conclusion, then, is that we have demonstrated learning in human participants without conscious awareness of what has been learned. This robust effect has been accomplished in a single, one-hour session with relatively few participants. These considerations suggest that this paradigm is ideally suited to the study of implicit processes, and we hope to report the results of further investigations with this paradigm in the near future.

### Acknowledgements

This research was supported by an ESRC grant to IPL McLaren and FW Jones.

### REFERENCES

- Aitken, M.R.F. (1996). *Peak shift in pigeon and human categorisation*. Unpublished PhD Thesis, University of Cambridge.
- Cleeremans, A. (1993). *Mechanisms of Implicit Learning: Connectionist Models of Sequence Processing*. Cambridge, MA: MIT Press.
- Jones, F.W., & McLaren, I.P.L. (2009). Human sequence learning under incidental and intentional conditions. *Journal of Experimental Psychology: Animal Behavior Processes*, 35(4), 538-553.
- Lovibond P.F., & Shanks, D.R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, 28(1), 3-26.
- McLaren, I.P.L., Green, R.E.A., & Mackintosh, N.J. (1994). Animal learning and the implicit/explicit distinction. In N.C. Ellis (Ed.) *Implicit and explicit learning of languages* (pp. 313-332). New York, NY: Academic Press.
- Mitchell, C.J., De Houwer, J., & Lovibond, P.F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, 32, 183-246.
- Posner, M.I. & Snyder, C.R.R. (1975). Attention and Cognitive Control. In Robert L. Solso (ed.), *Information Processing and Cognition: The Loyola Symposium*. Lawrence Erlbaum.
- Shanks, D.R. & St. John, M.F. (1994) Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, 17, 367-447.
- Shanks, D.R. & Lovibond, P.F. (2002) Autonomic and eyeblink conditioning are closely related to contingency awareness: Reply to Wiens and Öhman (2002) and Manns et al. (2002). *Journal of Experimental Psychology: Animal Behavior Processes*, 28, 38-42.
- Verbruggen, F., Adams, R., & Chambers, C.D. (in press). Proactive motor control reduces monetary gambling. *Psychological Science*.
- Willingham D.B., Nissen M.J. and Bullemer P. (2009). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15, 1047-1060.

# The Effect of Semantic Similarity is a Function of Contextual Constraint

**Hongoak Yun (hyun3@buffalo.edu)**

**Gail Mauner (mauner@buffalo.edu)**

Department of Psychology, 204 Park Hall  
Buffalo, NY 14260 USA

**Douglas Roland (droland@buffalo.edu)**

**Jean-Pierre Koenig (jpkoenig@buffalo.edu)**

Department of Linguistics, 609 Baldy Hall  
Buffalo, NY 14260 USA

## Abstract

We investigate how the degree to which a context constrains the words that could occur in a sentence affects the processing of the word that does occur. Roland et al. (2012) found that processing was facilitated when target words were more semantically similar to word alternatives that could have appeared. Because this effect is independent of word predictability, it suggests that comprehenders may have separate expectations for words and more general semantic features. We show that the semantic similarity effect is modulated by the degree of contextual constraint. We found that facilitation due to semantic similarity was greater when contexts were less constraining, and lower when contexts were more constraining, independent of word predictability. We interpret these results as suggesting that in highly constraining contexts, comprehenders may expect specific words, and face difficulties when these expectations are violated, while in less constraining contexts, they may have more general expectations for semantic properties shared between the words that could occur.

**Keywords:** sentence processing; semantic similarity; predictability; entropy; contextual constraint; expectation-based language comprehension

## Introduction

In expectation-based models of sentence comprehension, contextual information has an enormous effect on how words are integrated into sentences. These models predict that the degree of difficulty a reader encounters in integrating a new word into a sentence is either entirely or in large measure a function of how predictable that word is given prior context (e.g., Levy, 2008). Presumably this is because predicted words are activated by context in advance of when they are encountered, making them easier to retrieve from memory or because predictable words are easier to integrate into the representations being constructed during comprehension. The effect of predictability on processing time has been observed in many studies (e.g., Bicknell, Elman, Hare, McRae, & Kutas, 2010; Ehrlich & Rayner, 1981; Frisson, Rayner, & Pickering, 2005; Staub, 2011; DeLong, Urbach, & Kutas, 2005; Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Otten & Van Berkum, 2008; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005).

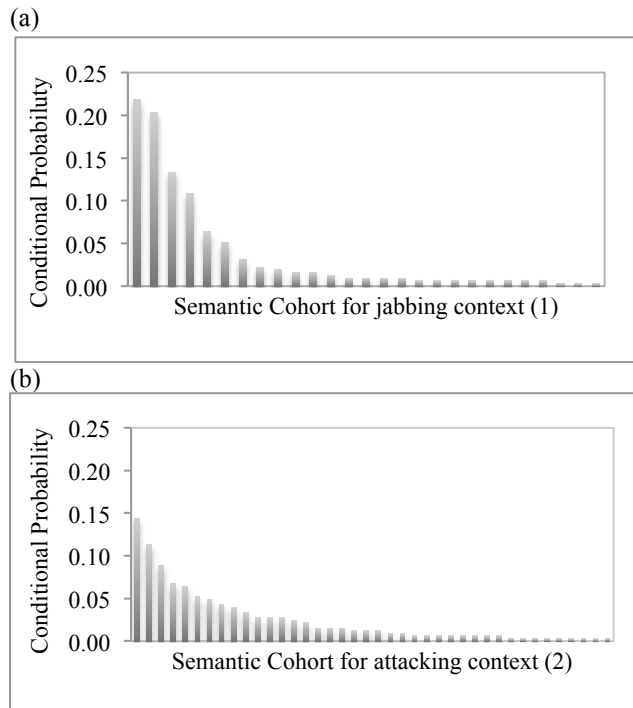
The relationship between a word's predictability and the amount of effort required to process it has been formalized in a number of computational models of language processing known as surprisal models (e.g., Boston, Hale, Patil, Kliegl, & Vasisht, 2008; Hale, 2001; Levy, 2008; Padó, Crocker, & Keller, 2009). In these models, the amount of cognitive effort required to integrate a word into a sentence depends on the negative log probability of that word given its preceding context. Surprisal models have had considerable success in predicting differences in reading times based on a word's predictability given its preceding context.

As it turns out, the amount of processing effort associated with integrating a word into a sentence cannot be entirely reduced to its predictability given its preceding context. Roland, Yun, Koenig, and Mauner (2012) examined the effects of the semantic cohort of a target word (i.e., the other words that could appear in the same position/context as the target word) on the processing of a target word. They found that words that are more semantically similar to their semantic cohort are easier to process when word predictability and other factors are controlled for. This result is important because it points to a limitation in expectation-based computational accounts of sentence processing that claim that a word's probability given its context is the sole predictor of processing effort (e.g., Levy, 2008). The results we present further constrain expectation-based accounts of sentence processing by showing that the effect of semantic similarity between a target word and its semantic cohort interacts with the degree of constraint provided by context.

The findings of Schwanenflugel and LaCount (1988) motivate the possibility that the benefits of semantic similarity on word integration might be modulated by contextual constraint. Schwanenflugel and LaCount found that unpredictable words that were semantically related to the most predictable word that could occur in the same position were processed faster than other equally unpredictable words that were also semantically unrelated to that most predictable word. What is crucial for this discussion is that the benefit of shared semantic information was not consistently observed for all unpredictable words. Shared semantic information only facilitated the processing of unpredictable words when contexts were weakly constraining.

To illustrate why the benefit of shared semantic similarity to other words activated by the preceding context would be greatest when a target word is unpredictable and its context is only weakly constraining, consider the sentence contexts in examples (1) and (2), for which we have obtained word completions (this study will be described in greater detail later).

- (1) The gladiator jabbed the African tiger with  
(2) The aborigine attacked the angry lion with



Figures 1a-b: Probability distributions for semantic cohorts for Examples (1) and (2).

In both contexts, an instrument noun is most likely to be the next word. However, the types of instruments in the semantic cohort differed across contexts. For context (1), instruments like *sword*, *spear*, *stick*, *knife*, and *spike* were mentioned. These instruments share a typical property, i.e., all can be used as “pokers”. In contrast, instruments like *sword*, *spear*, *knife*, *stick*, *fire*, *net*, *whip*, and *rock*, which were mentioned for context (2), have few salient characteristics that are common to instruments of attacking. This difference in the degree of shared characteristics suggests that context (1) places greater restrictions on the range of possible instruments than context (2). Using responses obtained from a completion study, we illustrate the distribution pattern of the probabilities of possible instruments for each context. In comparing Figure 1a to Figure 1b, two things become apparent. First, the most probable instruments for jabbing are more likely than the most probable instruments for attacking. Second, the probabilities of the jabbing semantic cohort drop more sharply than do the probabilities of the attacking semantic

cohort. One way of quantifying the greater degree of constraint provided by the jabbing context is to note that the top three items have a combined probability of .55. In the attacking context, even the first 6 items do not match that combined probability.

## Hypotheses and Prediction

Based on the findings of Schwanenflugel and LaCount (1988), we predict that the semantic similarity effect found by Roland et al. (2012) will be stronger in more weakly constraining contexts and weaker in highly constraining contexts. While Schwanenflugel and LaCount only examined the processing of unpredictable words, we do not expect interactions with word predictability, since Roland et al. found no interaction between similarity and predictability.

## Entropy as a measure of contextual constraint

In order to measure the effects of contextual constraint, we need a measure to quantify the degree of contextual constraint. Recall that in a more constraining context, there are larger differences in the probabilities of cohort members, because a small subset of the possible words is more likely, while the others are unexpected. Alternatively, in a less constraining context, there are a larger number of words that are more or less equally likely. We will use Entropy (H), a standard measure from information theory shown in Equation 1, to reflect these differences in the distributions of the probabilities the cohort members. Entropy is higher when the choices are more similar in probability, as in low constraint contexts, and is lower when choices are less similar in probability, as in more highly constraining contexts.

$$H(X) = - \sum_{i=1}^n p(x_i) \log p(x_i) \quad \text{Equation 1}$$

## Experiment to Generate Reading Times

**Participants** One hundred thirty native English-speaking undergraduates from the University at Buffalo received partial course credit for participation.

**Materials** We constructed 3 sets of 60 active declarative sentences with optional prepositional phrases similar to those in Example (3). Sets were differentiated by having an instrument noun that was highly likely (e.g., *sword*), moderately likely (e.g., *spear*), or unlikely (e.g., *spike*). To avoid wrap-up effects on instrument reading times (Just & Carpenter, 1980), all sentences included sentence-final phrases like *in the Colosseum*. Presentation regions are indicated in example (3) by vertical lines (|).

- (3) The gladiator |jabbed |the African tiger |with |a sword/spear/spike |in |the Colosseum.

Selection of target instruments was based on responses from a listing study in which 42 participants produced five instruments for sentence fragments like (1) and (2). Cloze probabilities for highly likely, moderately likely and unlikely instruments were  $M = .23$ ,  $S.D. = .06$ ,  $M = .10$ ,  $S.D. = .04$  and  $M = .02$ ,  $S.D. = .01$ , respectively. Results of plausibility rating revealed that all instruments were plausible. Co-occurrence frequencies between target verbs and instrument prepositional phrases, counted using the British National Corpus (BNC) (Burnard, 1995), were very low ( $M = .03$ ,  $S.D. = .03$ ), and separate modeling showed that the frequency with which each verb occurred with an instrument phrase played no role in our results.

Experimental sentences were counterbalanced across six presentation lists, each consisting of 10 experimental sentences for each level of predictability. To obscure systematicities, these sentences were intermixed with 90 distractor sentences with varied syntactic structures (e.g., subordinate clause, adverbial phrase, or relative-clause sentences) and prepositional phrases with different prepositions (e.g., *on*, *in*, or *from*). Finally, because participants judged whether each sentence made sense, 33% of the total number of trials were designed not to make sense.

**Procedure** Participant-paced, region-by-region reading was accompanied by a secondary make-sense judgment task. This task was used to increase sensitivity to subtle semantic effects that might not be observed in a straight reading paradigm. Trials were divided into two blocks with a two-minute break between blocks to lessen fatigue.

**Dependent Variables** While the primary dependent variable was the reading times for sentences that participants continued to judge acceptable, we examined “No” judgments to ensure that they did not differ as a function of instrument likelihood. Across conditions, percentages of “No” responses adjusted for remaining chances to say “No” (see Boland, Tanenhaus & Garnsey, 1990) were low (under 5% in all conditions) and their variances were small. “Yes” reading times for instrument noun phrases were filtered for outliers such that reading times greater than 4,000 ms or less than 200 ms were omitted. Filtering resulted in the removal of 27 of 3723 (0.7%) reading times.

### Measuring Effects of Predictability, Semantic Similarity, and Contextual Constraint

The goal of the modeling was to investigate how contextual constraints modulated the effect of semantic similarity. Reading times were submitted to a linear mixed-effects model for analysis using the R statistics program (version 2.14.0, R Development Core Team, 2011) using lme4 (version 0.999375-42, Bates & Maechle, 2011). Fixed factors consisting of Predictability, Similarity, Constraint, Length, and Frequency, described in more detail below, were used to predict reading time variances. Participants and

items were random factors. Fixed effects terms that did not contribute significantly to the fit of the model, including all 4-way and 5-way interactions, were removed. We simplified the initial fully crossed and fully specified random effects structure to yield the maximally justified random effect structure, as discussed by Jaeger (2009) and Baayen, Davidson, and Bates (2008). Outliers with a standardized residual at a distance greater than 2.5 standard deviations from zero were removed (Baayen, 2008).

### Model Predictors

**Predictability** We used log-transformed cloze probabilities from the above-mentioned listing study to estimate predictability. Each of the five responses was weighted by its order of mention. If an instrument was a participant’s first choice, it was weighted 5, if it was the second choice, it was weighted 4, and so on.

**Similarity** We measured the degree of semantic similarity between each target instrument and each member of its semantic cohort (i.e., the other words produced in the above-mentioned listing study) using Latent Semantic Analysis (LSA) cosines (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990) with a semantic space created from the BNC. LSA cosines were weighted by their cohort-frequencies to determine the average semantic similarity of a target instrument with its semantic cohort. Average LSA cosines between targets and their semantic cohorts ranged from .08 for the lowest similarity to .54 for the highest similarity. Our measure of semantic similarity differs from that used by Schwanenflugel and LaCount (1988), in that they compared the target word with only the most likely word, rather than with all of the words in the semantic cohort. In addition, they used human similarity judgments, while we used LSA cosines as a measure of similarity.

**Constraint** We used the entropy of the probability distribution of all possible instruments for a context to measure the degree of constraint provided by the preceding context. Entropy values ranged between 2.55 for the most constraining contexts and 5.02 for the least constraining contexts, with a mean of 3.88.

**Length** The lengths of instrument noun phrases were included as an additional factor to control for any potential reading time differences which might be due to this perceptual factor. Length was measured in number of characters, including spaces. Lengths ranged from 5 to 16 characters, with a mean of 8.36 characters.

**Frequency** We log-transformed the raw frequencies of the head nouns of the instrument noun phrases, which were obtained from the BNC. Base 10 log-transformed frequencies ranged from 0 to 4.56 (i.e., occurring between 1 and ~36K times in the BNC), with a mean of 2.85. Because frequency was correlated with Length ( $r = .65$ ) and



Predictability ( $r = .31$ ), we residualized Frequency for Length and Predictability, so that the predictors would only reflect the component of frequency that did not overlap with length and predictability. All other predictors had correlations of less than 0.30.

## Model Results and Discussion

We provide a summary of the linear mixed-effect regression model in Table 1 and a graphical representation of the interaction between Similarity and Constraint in Figure 2.

**Length** Longer words took longer to read. This is consistent with previous findings showing the effects of length (e.g., Juhasz & Rayner, 2003). Besides the interactions discussed below, there was a 3-way interaction between Length, Frequency, and Constraint. This was due to length effects being larger for low constraint, low frequency items and high constraint, high frequency items, and smaller for low constraint, high frequency items and high constraint, low frequency items. This possibly due to idiosyncrasies within our items, since we did not attempt to make sure that the same range of target word lengths were found in all conditions.

**Frequency** Unsurprisingly, more frequent words were read faster than less frequent words. This too is consistent with previous studies (e.g., Ashby et al., 2005; Juhasz & Rayner, 2003; Kliegl, Grabner, Rolfs, & Engbert, 2004; Staub, 2011). Frequency interacted with a number of other predictors as discussed below.

**Predictability** Consistent with many previous studies (Ashby, Rayner, & Clifton, 2005; Bicknell et al., 2010; DeLong et al., 2005; Ehrlich & Rayner, 1981; Federmeier et al., 2007; Frisson et al., 2005; Otten & Van Berkum, 2008; Rayner & Well, 1996; Staub, 2011; Van Berkum et al., 2005), more predictable instruments were processed more quickly.

There was a 3-way interaction between Predictability, Frequency, and Constraint, as well as a 2-way interaction between Predictability and Frequency, and a marginally significant 2-way interaction between Predictability and Constraint. These are due to low frequency unpredictable words taking longer to read in low constraint contexts than would be expected from the simple effects of frequency and predictability (i.e., when all factors combine to give the comprehender the least amount of help in predicting the word). This may have resulted in the model underestimating the reading times for low frequency unpredictable words in low constraint contexts, giving the appearance of a lack of a frequency effect for highly predictable words in low constraint contexts.

Table 1: Summary of fixed factors from the linear mixed-effect regression model, when the effects of random variables were maximized, for predicting reading times of that target noun.

	Estimated Coefficient	S.E.	<i>t</i> -value
Intercept	713.72 (713.64)	17.51	40.75
Predictability	-59.63 (-33.22)	5.49	-6.05
Similarity	-271.95 (-28.94)	7.67	-3.77
Constraint	17.25 (8.99)	10.31	0.87
Length	15.35 (33.28)	7.43	4.48
Frequency	-31.05 (-31.22)	7.10	-4.40
Predictability x Similarity	-57.68 (-3.59)	5.44	-0.66
Predictability x Constraint	-36.97 (-10.89)	5.61	-1.94
Predictability x Length	-3.92 (-4.60)	5.84	-0.79
Predictability x Frequency	23.89 (13.56)	5.31	2.55
Similarity x Constraint	-453.57 (-24.82)	6.55	-3.79
Similarity x Length	-52.25 (-12.09)	5.92	-2.04
Similarity x Frequency	35.14 (3.85)	6.57	0.59
Constraint x Length	-8.38 (-9.45)	8.10	-1.17
Constraint x Frequency	11.42 (6.18)	6.42	0.96
Length x Frequency	1.51 (3.44)	5.02	0.69
Predictability x Frequency x Constraint	44.19 (12.57)	10.66	2.29
Similarity x Constraint x Length	-138.28 (-16.18)	6.13	-2.37
Similarity x Frequency x Length	-55.33 (-12.90)	2.54	-2.35
Constraint x Frequency x Length	-26.41 (-29.67)	2.84	-4.82

Note: All predictors are centered. Parenthetical values below the coefficients are standardized coefficients from an alternate version of the model with standardized predictors. *t*-values with an absolute value greater than 2 are significant at an alpha level of .05 (Gelman & Hill, 2007).

**Similarity** Instruments were read faster when they were more similar to the members of their semantic cohort than when they were less similar. This result replicates Roland et al.'s (2012) results. There was also a 3-way interaction between Similarity, Frequency, and Length and a 2-way interaction between Similarity and Length. These interactions were due to the effect of similarity being larger for longer words and smallest for short, high frequency words. These are both consistent with the notion that similarity effects are due to spreading activation during processing, as the slower reading times for longer words provide more chance for activation to spread between pre-activated words, and short, fast words, being read quickly, provide the least time.

**Constraint** There was no main effect of contextual constraint. Importantly however, Constraint interacted with Semantic Similarity, just as hypothesized. We analyzed this interaction by performing separate analyses on the data where one standard deviation was either added or subtracted from the values for each of the predictors in the interaction to create models reflecting low and high conditions for each predictor, respectively (Aiken and West 1991). There was an effect of Semantic Similarity when Entropy was high (i.e., low constraint contexts) (Estimated coefficient = -500.28, S.E. = 103.43, t-value = -4.84), but no effect of Semantic Similarity when entropy was low (i.e., high constraint contexts) (Estimated coefficient = -34.14, S.E. = 85.70, t-value = -0.40). The estimated high and low reading times are shown in Figure 2. The fact that Semantic Similarity did not facilitate the integration of instruments in strongly-constraining contexts is consistent with Schwanenflugel and LaCount's (1988) results. In addition, there was a 3-way interaction between Similarity, Constraint, and Length, with the similarity effects in the low constraint conditions being larger for longer words than for shorter words. Again, this is consistent with the notion that the added reading times of longer words allows more time for activation to spread between pre-activated words.

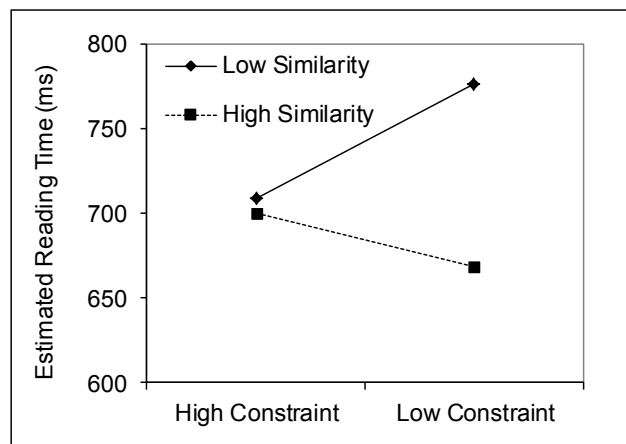


Figure 2: Interaction of Contextual Constraint and Similarity using standardized coefficients.

## General Discussion

We found that semantic similarity between a target word and its semantic cohort has a stronger effect on processing when the context provides fewer constraints on what may appear in the target position. Alternatively, the effects of semantic similarity become weaker as the context becomes more constraining. The effect of contextual constraint on the degree to which semantic similarity affects processing has important implications for models of processing. Roland et al. (2012) suggested two possible causes for the semantic similarity effect: spreading activation between the representations for the words that comprehenders were anticipating, and the possibility that expectations for words and expectations for semantic features could have independent effects on comprehension difficulty. Our results suggest that the nature of comprehenders' expectations may vary with the degree of contextual constraint. In a highly constraining context (i.e., low entropy), there is no effect of semantic similarity, and comprehension difficulty appears to be primarily determined by the predictability of the target word. If the target word is expected, it is easy to process. If the target word is unexpected, it is difficult to process.

On the other hand, in a less constraining context, semantic similarity and predictability both influence processing. Not only are more predictable words easier to process, but so are words that are more similar to the other members of the semantic cohort. Words are most difficult to process when they are both unexpected and semantically distant from their semantic cohort.

One possible explanation for why contextual constraint modulates the influence of semantic similarity for unpredictable words is that in a highly constraining context, comprehenders may be expecting specific words, and face difficulty when the expectations turn out to be wrong. In a less constraining context, comprehenders may have less specific expectations – anticipating semantic features in common between a set of possible words (in addition to, or as an alternative to anticipating specific words). Thus, they face less difficulty when the target word turns out to be something other than the most likely word – as long as the target word shares some level of semantic similarity with the other likely possible words. Overall, our data suggests that word predictability, semantic similarity, and contextual constraint all have an impact on language comprehension.

## References

- Aiken, L.S., & West, S.G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA: Sage.
- Asby, J., Rayner, K., & Clifton, C., Jr. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 58A(6), 1065-1086.
- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University

- Press.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.
- Bates, D., Maechler, M. & Bolker, B. (2011). lme4: Linear mixed-effects models using Eigen and Eigen++. R package version 0.999375-42. <http://CRAN.R-project.org/package=lme4>
- Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, 63, 489–505.
- Boland, J. E., Tanenhaus, M. K., & Garnsey, S. M. (1990). Evidence for the immediate use of verb control information in sentence processing. *Journal of Memory and Language*, 29(4), 413–432.
- Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1). 1, 1–12.
- Burnard, L. (1995). Users reference guide for the British National Corpus. Oxford: Oxford University Computing Services.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society For Information Science*, 41, 391–407.
- DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8, 1117–1121.
- Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20, 641–655.
- Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, 1146, 75–84.
- Frisson, S., Rayner, K., & Pickering, M. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31 (5), 862–877.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.
- Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 1–8). Pittsburgh, PA: Carnegie Mellon University.
- Jaeger, T. F. (2009, May 14). Random effect: Should I stay or should I go? [Web log post]. <http://hplab.wordpress.com/2009/05/14/random-effect-structure/> Retrieved 24.07.11.
- Juhász, B. J., & Rayner, K. (2003). Investigating the Effects of a Set of Intercorrelated Variables on Eye Fixation Durations in Reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1312–1318.
- Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), 329–354.
- Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1-2), 262–284.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106, 1126–1177.
- Otten, M., & Van Berkum, J. J. A. (2008). Discourse-based lexical anticipation: prediction or priming? *Discourse Processes*, 45(6), 464–496.
- Padó, U., Crocker, M., & Keller, F. (2009). A Probabilistic Model of Semantic Plausibility in Sentence Processing. *Cognitive Science*, 33(5):794–838.
- R Development Core Team. (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, 3(4), 504–509.
- Roland, D., Yun, H., Koenig, J.-P., & Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. *Cognition*, 122, 267–279.
- Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 14, 344–354.
- Staub, A. (2011). The effect of lexical predictability on distributions of eye fixation durations. *Psychonomic Bulletin & Review*, 18, 371–376.
- Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 443–467.

# Mutual Exclusivity and Vocabulary Structure

**Daniel Yurovsky**

[dyurovsk@indiana.edu](mailto:dyurovsk@indiana.edu)

Department of Psychological and Brain Sciences  
Indiana University

**Linda B. Smith**

[smith4@indiana.edu](mailto:smith4@indiana.edu)

Department of Psychological and Brain Sciences  
Indiana University

**Ricardo A. H. Bion**

[ricardoh@stanford.edu](mailto:ricardoh@stanford.edu)

Department of Psychology  
Stanford University

**Anne Fernald**

[afernald@stanford.edu](mailto:afernald@stanford.edu)

Department of Psychology  
Stanford University

## Abstract

The words that children learn can be characterized as a semantic network, with links connecting related words. Recent analyses have shown these networks to have small-world structure, with a few highly-connected hub words facilitating short paths between otherwise distant words. This structure contributes to network robustness, and differences in structure can predict differences in language learning outcomes. While previous studies have shown that semantic network structure reflects linguistic input structure, we provide the first evidence that it is related also to children's own language learning biases. Two-year old children who show a mutual-exclusivity bias have significantly more hub-like networks than children who do not, even when they know the same number of words. This finding contributes to our understanding of both semantic networks and the origins of mutual exclusivity.

**Keywords:** word learning; mutual exclusivity; semantic networks; language acquisition

## Introduction

Although the earliest analyses of human memory and learning concerned the learning of lists of unrelated words (Ebbinghaus, 1885/1962), researchers quickly discovered that the words people learn in more natural contexts are intricately connected. Vocabularies were conceptualized as richly structured networks, with links connecting semantically related words (Collins & Loftus, 1975). These connections play an important role in both learning and memory, and can be observed empirically in semantic priming experiments. Because activation spreads from words to their semantic neighbors, presenting a word, even subliminally, leads to faster processing of related words (Anderson, 1983). Even two-year old infants show semantic priming, suggesting that vocabularies have network structure early in language learning (Arias-Trejo & Plunkett, 2009).

Recently, the application of graph-theoretic methods to the study of these networks has begun to provide insight into their structural properties. For instance, Hills, Maouene, Maouene, Sheya, & Smith (2009a) analyzed the semantic network structure of 130 nouns typically learned before 30 months. Compared to randomly-connected control networks, these semantic networks showed significant small-world structure, in which most words are sparsely connected, but a few are highly-connected hubs. This kind

of structure results in networks robust to malfunction (e.g. forgetting a word; Albert, Jeong, & Barabási, 2000), and can help to explain some of the remarkable efficiency of human semantic memory (Raaijmakers & Shiffrin, 1981). Further, semantic networks lacking this structure are associated with slower language-learning, characterizing the vocabulary structure of late talkers (Beckage, Smith, & Hills, 2011). But why do children learn these words? Why do semantic networks have this structure?

Undoubtedly, one answer to this question is that structure comes from the environment. Because children learn words from the language they hear, language input is a strong predictor of the words that children will learn. For instance, the frequency with which a child hears a word in isolation can predict how likely a child is to learn that word (Brent & Siskind, 2001). Similarly, the semantic networks constructed from corpora of both adult-directed and child-directed language have many of the same structural properties as networks constructed from the words 30-month-old children are likely to know (Hills, et al., 2009a; Steyvers & Tenenbaum, 2005).

But perhaps a more complete explanation of the origin of semantic network structure is that it emerges from an interaction between structure in the linguistic environment and the child's own learning system. Because children are not unbiased samplers of linguistic input, their attentional and learning biases mediate the link between language input and language learned (Hudson Kam & Newport, 2005; Smith, 2000). For instance, children who learn to attend to shape are likely to learn shape-based categories, and those who learn to attend to other properties (e.g. material) learn other kinds of words (Colunga & Sims, 2011; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Can word-learning biases predict and explain semantic network structure? In this paper, we consider the case of disambiguation through mutual exclusivity.

In the disambiguation task, a child is presented with a novel object among one or more familiar object competitors. The child then hears a novel label (e.g. 'can you find the dax?') and is asked to select an object. Both toddlers and adults reliably select the novel object as the target of the novel label (Markman & Wachtel, 1988; Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992), and studies with infants suggest that this disambiguation may arise as early as 18 to 22.5 months (Halberda, 2003; Mather & Plunkett, 2009). Preferential

mapping of novel labels to novel objects over known objects, which we will refer to as mutual exclusivity (ME), could arise for a number of reasons, and its mechanism of action is the topic of significant debate (e.g. Diesendruck & Markson, 2001; Golinkoff et al., 1992; Markman & Wachtel, 1988). We explore this question in the general discussion, but will sidestep it here and instead consider the potential consequences of mutual exclusivity for semantic network structure.

Mutual exclusivity is a mechanism by which children can leverage prior knowledge to learn new words *in the context of known objects*. Consequently, children who show mutual exclusivity should have vocabularies that echo this kind of contextual structure. For these children, learning *fork* should ease the acquisition of *spoon*, *bowl*, and *plate*. In contrast, learning *fork* should have little effect on the acquisition of *dog* and *coat*. Thus, we propose that mutual exclusivity can help explain small-world structure of semantic networks, and those children who show mutual exclusivity will have more hub-like networks than those who do not. We begin by reporting empirical data from a disambiguation task with 24-month-old children, continue by describing a semantic network analysis of these children's vocabularies, and conclude with a discussion of how these results inform our understanding of the relationship between mutual exclusivity and vocabulary development, as well as the origins of mutual exclusivity itself.

## Experiment

### Method

**Participants.** Forty two-year-olds ( $M = 24.75$  months; range = 24-26; 20 female) participated. All were typically developing children from households in which parents reported English to be the dominant language. A subset of 34 infants ( $M = 24.9$  months; range = 22.4-27.5; 16 female) participated in the followup analysis (explained below).

**Stimuli.** Nine familiar objects (e.g. boat, glasses) were used in the warm-up trials. Twenty-five familiar (e.g. brush, cup) and 8 novel objects (e.g. massager, platypus) were used in the referent-selection task.

**Procedure.** Parents first completed the MCDI (Fenson, Dale, Reznick, Bates, Thal, & Pethick, 1994) and an SES measure (Hollingshead, 1975). After this, each child participated in three warm-up trials. On warm-up trials, the experimenter set a tray containing three familiar objects on the table, initially covered by an occluder. The experimenter asked for the target object (e.g. "which one is the dog?") three times: once while the items were occluded, again after lifting the occluder, and again three seconds later while pushing the tray towards the infant. The first reach, point, or grab, was scored as a response. On these trials, infants were praised for correct responses and corrected when necessary.

Subsequently, each child participated in sixteen referent-selection trials. On each trial, the experimenter presented a tray containing two familiar objects and one novel object.

The procedure was identical except that children received neutral feedback on all trials. On half the trials, the experimenter asked for a familiar object, while on the other half she asked for a novel object (e.g. *modi*, *taju*).

## Results and Discussion

Each participant made a total of 16 choices, picking 8 targets on familiar trials, and 8 targets on novel trials. Any trial on which the child did not know the label for the familiar target, or the label for one of the familiar distractors, was excluded from analysis. The proportion of targets correctly chosen on these remaining trials was then analyzed to determine the child's success in the task. Overall, children performed quite well, selecting the correct target on both familiar ( $M_f = .83$ ,  $t(39) = 15.31$ ,  $p < .001$ ) and novel trials ( $M_n = .545$ ,  $t(39) = 5.87$ ,  $p < .001$ ) at greater than chance levels. Thus, as a group, 24-month-old children used mutual exclusivity for disambiguation. Familiar trial performance, however, was significantly higher than novel trial performance ( $t(39) = 6.88$ ,  $p < .001$ ).

Because the central question in this study is about the relationship between learning mechanisms and vocabulary development, we measured both vocabulary size (MCDI - Fenson, et al., 1994) and mother's education (Hollingshead, 1975), a potential correlate of rich language input. Mother's education was reliably correlated with performance on familiar trials ( $r = .33$ ,  $p < .05$ ), but not novel trials ( $r = .01$ , *n.s.*), and vocabulary size was not significantly correlated with performance on either kind of trial ( $r_f = .19$ , *n.s.*;  $r_n = .15$ , *n.s.*). In the semantic network analysis to follow, we show that *vocabulary structure* is reliably related to novel trial performance. Because neither mother's education nor *vocabulary size* predict ME in this data set, the relationship between ME and structure is likely to be quite robust.

But perhaps this analysis is unfair. While most of the children had high levels of success on familiar trials, a few children did not perform as well. Since these children knew the words for all three objects on these familiar trials, their low levels of performance indicate that they may not have understood the task. Thus, for the same reason that response time analysis typically uses only correct response trials, excluding these children from individual-level analyses may give clearer correlations. In order to determine whether a child's performance was significantly better than expected by chance, we modeled chance behavior on each trial as random selection of one of the three objects.

The probability of success expected by chance is given by a binomial distribution with probability  $\frac{1}{3}$ . Consequently, a child should be counted as performing differently from chance if he or she made enough correct selections to be outside the 95% confidence interval for a binomial distribution. A child who made 8 choices, for instance, needed to make at least 5 correct choices to be counted as performing better than expected by chance. Each child's number of correct selections on familiar trials was thus submitted to a binomial test. Six of the 40 children were found to have performance levels on the familiar trials

indistinguishable from chance, and were thus excluded from further analysis. This left a subset of 34 children who could confidently be assumed to have understood the task. Figure 1 shows novel and familiar trial performance for children both from the full set, and from this reduced subset.

We also performed a similar analysis on novel trials, dividing children into two categories: those who reliably showed evidence of using mutual exclusivity (ME), and those who did not. Seventeen children were classified as ME users, and seventeen were classified as Nonusers. We are not arguing that ME is a binary phenomenon, but rather perform this binary split for technical reasons. Binarization loses some information separating children within the ME users category, but it also cleans up noise that may not meaningfully separate nonusers. Quantitative differences at or below chance levels are more likely to be generated by noise than they are to be generated by meaningful process differences, and thus are likely to dilute linear correlations. In subsequent analyses, because mutual exclusivity is analyzed as a binary phenomenon, we use Spearman's  $\rho$ , a non-parametric measure of correlation. In all cases, correlations were stronger for this binary measure.

In this subset, mother's education was still correlated with performance on familiar trials ( $r = .36, p < .05$ ), as was vocabulary size ( $r = .39, p < .05$ ). Neither mother's education nor vocabulary size predicted performance on novel trials ( $\rho = -.11, n.s.$ ;  $\rho = .08, n.s.$ ). In the analyses that follow, we compare the semantic network connectivity of ME users and nonusers. Because use of mutual exclusivity was uncorrelated with vocabulary size, differences in network connectivity are unlikely to be a simple reflection of network size. Further, because mother's education predicted performance on familiar, but not novel, trials, a relationship between ME and vocabulary structure arising from language input must come from more specific properties not indexed by mother's education in this sample.

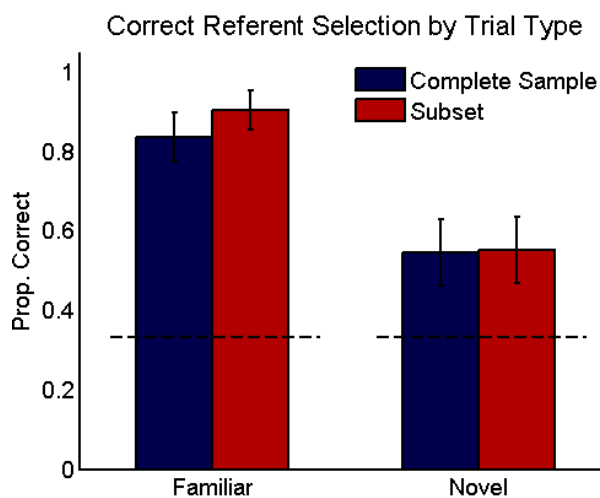


Figure 1: Proportion of correct choices by participants in both the familiar and novel conditions. Dark blue bars show the complete sample, light red bars the subset. Error bars indicate  $\pm 1$  standard error.

## Semantic Network Analysis

To understand how use of mutual exclusivity contributes to the structure of children's vocabularies, we formalize these vocabularies as semantic networks. In semantic networks, vertices represent the words that children know, and edges represent semantic relationships among these words. In any such analysis, the first step is to formalize 'semantic relatedness' – the relationship used to link two words.

Previous analyses have used a number of successful metrics of connectivity: co-occurrence in CHILDES (e.g. Beckage, Hills, & Smith, 2011), frequency of free-association by adults (e.g. Steyvers & Tenenbaum, 2005), and common perceptual and conceptual features (e.g. Hills, et al., 2009b). In our analysis, we adopt and extend the last approach, connecting two words if they share a number of common semantic features. Features were drawn from the set of McRae feature norms (McRae, Cree, Seidenberg, & McNorgan, 2005). McRae and colleagues asked 725 adults to freely list up to 14 features of 541 English nouns. The number of features shared by two words gives a measure of their semantic relatedness.

Although participants could generate any features they liked, McRae et al. (2005) subsequently divided the generated features into 4 categories: perceptual features accessible to the 5 senses (e.g. "has fur," "tastes sweet"), functional features (e.g. "used for writing," "is edible"), encyclopedic features (e.g. "is expensive"), and taxonomic features (e.g. "a crustacean"). Following Hills et al. (2009b), we analyze only features of the first and second kind, as these are the features likely to be available to two-year-old children. We create two different networks for each child: one in which connectivity is defined by *perceptual* feature overlap, and one in which connectivity is defined by *functional* feature overlap. This is because overlapping perceptual features indicate a very different kind of relatedness than overlapping conceptual features.

Hills et al. (2009b) analyzed the clusters produced by each of these kinds of networks to quantify these different kinds of relatedness. Defining connectivity by *perceptual* feature overlap produced networks that were dense, highly connected, put words into more than one category, and produced categories that were overly inclusive relative to human judgments (e.g. MCDI categories, Fenson, et al., 1994). In contrast, *functional* feature overlap produced networks that were sparser, had smaller, better defined categories, and were better at discriminating among near-category members. In general, words connected in the *functional* network are more likely to be encountered in a relational context, facilitating learning by mutual exclusivity (e.g. *cake-carrots*, *boots-coat*). In contrast, words connected in the *perceptual* networks are less likely to be encountered in such situations, and learning one is thus less likely to facilitate learning the other through mutual exclusivity (e.g. *sheep-sofa*, *pencil-stick*). Thus, we can test a specific prediction about how mutual exclusivity builds vocabulary structure: it facilitates the acquisition of *functionally* related words.

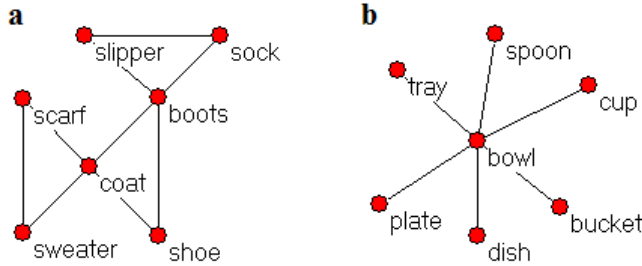


Figure 2: Two 7-vertex networks with different connectivity structures. The left network has a high clustering coefficient and a low degree centrality. The right network has a low clustering coefficient, but a high degree centrality.

In addition to using these two kinds of features to define connectivity, we measure their resulting structure in two different ways. These different connectivity measurements represent different ways in which mutual exclusivity could build structure. Consider the networks in Figure 2.

The first network (2a) has many local clusters, triangles in which any vertices with a common neighbor are likely to be neighbors themselves. One might predict mutual exclusivity to facilitate this kind of structure because using one word (e.g. *scarf*) to learn a semantic neighbor (e.g. *sweater*) should make a common neighbor even easier to learn (e.g. *coat*). This structure is measured by *clustering coefficient* (Equation 1), which has previously been used to distinguish the vocabulary structures of early and late talkers (Beckage, Hills, & Smith, 2011).

$$C = \frac{1}{|V|} \sum_{i=1}^{|V|} \frac{2|\{e_{jk}\}|}{d(v_i)(d(v_i) - 1)} \quad v_j, v_k \in N_i, e_{jk} \in E \quad (1)$$

In contrast, the second network (2b) does not have any local clusters, but rather has a single highly salient *hub*: a single vertex with many neighbors. This kind of structure might be even more likely to arise through mutual exclusivity, as learning the *hub* word (*bowl*) makes each of its neighbors easier to learn (*spoon*, *tray*, *cup*). This kind of structure is measured by *degree centrality* (Equation 2). This measure is new to semantic network analyses, but is a mainstay of social network science (Freeman, 1979), and measures the structural property intuitively most likely to be related to learning words through exclusion.

$$D = \frac{\sum_{i=1}^{|V|} [d(v^*) - d(v_i)]}{(|V| - 1)(|V| - 2)} \quad (2)$$

Thus, we test two hypotheses in the following analysis: mutual exclusivity should predict connectivity structure in *functional* but not the *perceptual* networks, and it should manifest more strongly in high *degree centrality* should than in *clustering coefficient*.

## Method

To construct semantic networks for each child, we used all words which are both measured by the MCDI, and for which McRae and colleagues collected feature norms. This resulted in a list of 130 nouns, encompassing animals, food, clothing, vehicles, etc. For a full list, see Hills et al. (2009b). Each child's semantic network was constructed by adding one vertex for each word on that child's productive MCDI. Vertices were connected if they shared a minimum number ( $w$ ) of semantic features. To be consistent with Hills et al. (2009b), we set this features threshold to all possible values 1-4. At  $w = 3$ , for instance, two words were connected only if they shared three or more semantic features. However, networks become increasingly sparse as  $w$  increases, and we thus urge caution in interpreting results at high thresholds.

Two networks were created for each child, one network in which only *perceptual* features defined connectivity, and one network in which only *functional* features were used to define connectivity (see above). Networks were defined by their set of vertices  $V$  and the set of edges  $E$  that connected them. A vertex's degree ( $d(v)$ ) is defined as the number of other vertices to which it is connected by an edge. These connected vertices are called neighbors, and together define a node's neighborhood ( $N$ ).

Once each network was constructed, two properties of its connectivity structure were measured. The first, clustering coefficient ( $C$ ), measures the proportion of vertices with a common neighbor that are also neighbors of each other. (Equation 1). The second, degree centrality ( $D$ ), measures the proportion of edges connected to a single dominant *hub* vertex (Equation 2). These measures of structure trade off, with high degree centrality necessitating a low clustering coefficient. Both measures always range between 0 and 1, and thus are independent of the size of a child's vocabulary. They are measures of structure independent of size.

## Results and Discussion

As in the analysis above, children were divided into two groups: Mutual Exclusivity Users who performed better than chance on the novel trials of the disambiguation task, and Nonusers who did not. Again, we reiterate that this is not a theoretical commitment, but rather a tool for noise reduction. The structure of each child's individual semantic networks – both perceptual and functional – was used to predict that child's category of mutual exclusivity usage.

Before presenting the results of network analyses, we recapitulate that vocabulary *sizes* were quite comparable between these groups. The 17 ME Users produced an average of 408.3 words on the MCDI while the 17 Nonusers produced an average of 388.1 ( $t(32) = .37, n.s.$ ). They also did not differ in the number of words they knew from the set of 130 used in the network analysis ( $M_u = 92.6, M_n = 88.1, t(32) = .51, n.s.$ ). However, the particular words they knew, and the semantic relationships among them, proved to be importantly different.



Figure 3 shows correlations between measures of network structure and the mutual exclusivity category to which each individual child belonged. For perceptual networks, constructed by connecting words by shared perceptual features (e.g. “has fur,” “tastes sweet”), neither clustering coefficient nor degree centrality were related to use of mutual exclusivity at any feature overlap threshold (Figure 3, left column). As predicted, perceptual networks, in which connections are not a good proxy for words likely to occur in contrastive contexts, have structures not well predicted by use of mutual exclusivity.

In contrast, for functional networks, those constructed by connecting words by shared relational, functional features, mutual exclusivity was a significant predictor of degree centrality when 2 or more overlapping features defined a connecting edge ( $w = 2$ ). At this threshold, children who used mutual exclusivity at above-chance levels had semantic networks with higher degree centrality ( $\rho = .34, p < .05$ ; Figure 3, bottom right). This same threshold is shown by Hills et al. (2009a) to best separate semantic categories in this set of words. Use of mutual exclusivity did not reliably predict clustering coefficient, but did show a positive trend, particularly at overlap threshold 3 ( $\rho = .28, p = .1$ ; Figure 3, top right). This trend should be interpreted cautiously, however, as conceptual networks were quite sparse at  $w = 3$ , having at most 12 edges.

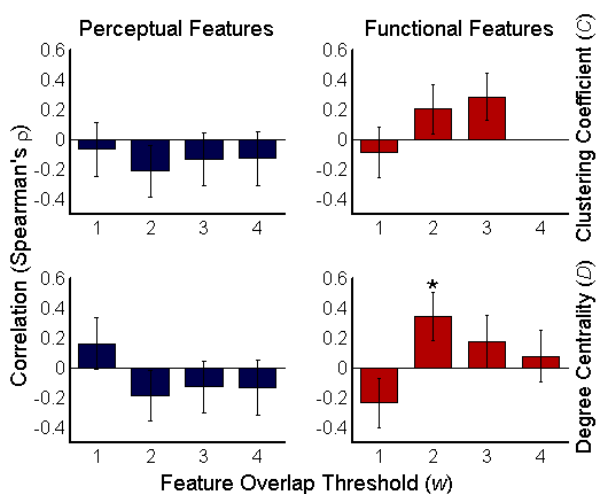


Figure 3: Correlation between network structure and mutual exclusivity performance. The top row shows correlations clustering coefficient, the bottom row for degree centrality. The left column shows perceptual features, the right shows functional. Individual bars shows correlations when particular thresholds ( $w$ ) define the minimum number of overlapping semantic features required to connect two words. Error bars show  $\pm 1$  standard error as measured by 1000 samples of bootstrap resampling (Lunnenborg, 1985). Mutual exclusivity was significantly correlated with degree centrality in the functional networks and showed a non-significant trend towards correlation with correlation coefficient in these same networks. As predicted, ME was uncorrelated with perceptual network structure.

Thus, the semantic network structures of children who reliably exhibit mutual exclusivity are predictably different from those of children who do not. Even though these children know the same number of words, the words they know are different. Semantic networks of ME users are characterized by more hub-like structure, a consequence of the kind of word learning facilitated by exclusion. Importantly, these differences are likely to matter (Beckage, Smith, & Hills, 2011). Differences in connectivity structure lead to differences in network robustness, with the networks of mutual exclusivity structure perhaps protecting them against forgetting and aiding future learning (Albert, Jeong, Barabási). These results represent a first step in understanding how children’s own learning mechanisms build the structure of their semantic networks.

## General Discussion

While the words that children learn are, of course, a function of the linguistic input to which they are exposed (Brent & Siskind, 2001; Hills et al., 2009a), this link is likely to be moderated by children’s own attentional and learning mechanisms (Hudson Kam & Newport, 2005; Smith, 2000). For instance, children learn to extend newly-learned object words to categories on the basis of particular feature dimensions. Normatively, children learn a bias to attend to shape, and this bias leads them to learn more categories organized by shape (Smith et al., 2002). However, children may learn a different bias, and consequently learn different words in the future (Colunga & Sims, 2011). We show that use of mutual exclusivity may play a similar role. Children who robustly use mutual exclusivity are likely to learn new words functionally related to words they already know. As atypical semantic network structure is related to slower language learning (Beckage, Smith, & Hills, 2011), these results point to a potential intervention for late-talking children. Learning to disambiguate the meanings of new words through exclusion could help late-talkers to catch up.

These results also lead to two further insights about mutual exclusivity and its role in vocabulary development. While mutual exclusivity is often thought to be critical to early word learning, its relationship to vocabulary size is unclear. For every study that finds a significant correlation between mutual exclusivity and vocabulary size (e.g. de Marchena, Eigsti, Worek, Ono, & Snedeker, 2011; Mervis and Bertrand, 1994), another finds no correlation between the two (e.g. Halberda, 2003; Mather & Plunkett, 2009). These results help to shed light on this inconsistency by pointing out that the relationship between vocabulary development and mutual exclusivity may be found not in size but in structure. While we do not mean to argue that mutual exclusivity is required for rapid word learning, we do suggest that their relationship can be better understood by considering semantic network structure.

Finally, these results may shed light on the origins of mutual exclusivity itself. Thus far, we have argued that mutual exclusivity builds vocabulary structure. But

vocabulary structure may also build mutual exclusivity. One can think of mutual-exclusivity as an overhypothesis, a probabilistic rule about the general structure of word-object mappings derived from the structure of individual word-object mappings (Kemp, Perfors, & Tenenbaum, 2007; Mervis & Bertrand, 1994). For instance, mutual exclusivity may have its roots in an understanding that labels are often contrastive, pointing to differences between otherwise similar objects. If this is true, vocabularies that make this overhypothesis more probable should lead to stronger mutual exclusivity biases. Thus, one can think of the hub-like structures characteristic of ME users in our sample as not only arising from mutual exclusivity, but helping to construct it as well. Hub words, which are connected to many semantically-related neighbors, may play an important role in discovery of this higher-order regularity. Thus, mutual exclusivity may operate much like the shape bias: being both built from regularities in the structure of linguistic input, and helping children to discover further regularities (Smith, et al., 2002). A deep understanding of the connection between mutual exclusivity and vocabulary structure, then, will come from understanding a three part relationship: how ME contributes to structure, how structure contributes to ME, and language input contribute to both.

### Acknowledgments

We are grateful to Cody Stitzel for her help coding the MCDI data, and to all of the members of the Smith, Yu, and Shiffrin labs whose feedback helped to improve this work. This work was supported by a NSF GRF to DY.

### References

- Albert, R., Jeong, R., & Barabási, A. (2000). Error and attack tolerance of complex networks. *Nature*, 406, 378-382.
- Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, 22, 261-295.
- Arias-Trejo, N., & Plunkett, K. (2009). Lexical-semantic priming effects during infancy. *Philosophical Transactions of the Royal Society B*, 364, 3633-3647.
- Beckage, N., Smith, L., & Hills, T. (2011). Small Worlds and Semantic Network Growth in Typical and Late Talkers. *PloS ONE*, 6, e19348.
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407-428.
- Colunga, E., & Sims, C. E. (2011). Early talkers and late talkers know nouns that license different word learning biases. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2550-2555). Austin, TX: Cognitive Science Society.
- de Marchena, A., Eigsti, I. M., Worek, A., Ono, K., & Snedeker, J. (2011). Mutual exclusivity in autism spectrum disorders: Testing the pragmatic hypothesis. *Cognition*, 119, 96-113.
- Diesendruck, G., & Markson, L. (2001). Avoidance of lexical overlap: A pragmatic account. *Developmental Psychology*, 37, 630-641.
- Ebbinghaus, H. (1885/1962). *Memory: A contribution to experimental psychology*. New York: Dover.
- Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S.J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59. Chicago: University of Chicago Press.
- Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1, 215-239.
- Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L., M. & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology*, 28, 99-108.
- Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, 87, B23-B34.
- Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. B. (2009a). Categorical structure among shared features in networks of early-learned nouns. *Cognition*, 112, 381-396.
- Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. B. (2009b). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition?. *Psychological Science*, 20, 729-739.
- Hollingshead, A. B. (1975). *Four factor index of social status*. Unpublished manuscript, Department of Sociology, Yale University, New Haven, CT.
- Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1, 151-195.
- Lunnenborg, C. E. (1985). Estimating the correlation coefficient: The bootstrap approach. *Psychological Bulletin*, 98, 209-215.
- Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10, 307-321.
- Markman, E., & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20, 121-157.
- Mather, E., & Plunkett, K. (2009). Learning words over time: The role of stimulus repetition in mutual exclusivity. *Infancy*, 14, 60-76.
- McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37, 547-559.
- Mervis, C. B., & Bertrand, J. (1994). Acquisition of the novel name-nameless category (N3C) principle. *Child Development*, 65, 1646-1662.
- Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, 88, 92-134.
- Smith, L. B. (2000). Learning to learn words: An associative crane. In R. M. Golinkoff & K. Hirsh-Pasek (Eds.), *Becoming a word learner: A debate on lexical acquisition* (pp. 51-80). New York: Oxford University Press.
- Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object naming provides on-the-job training for attention. *Psychological Science*, 13, 13-19.
- Steyvers, M., & Tenenbaum, J. B. (2005). The large scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29, 41-78.

# Quantitative Linking Hypotheses for Infant Eye Movements

**Daniel Yurovsky**

dyurovsk@indiana.edu  
Department of Psychological  
and Brain Sciences  
Indiana University

**Shohei Hidaka**

shhidaka@jaist.ac.jp  
School of Knowledge Science  
Japan Advanced Institute of  
Science and Technology

**Rachel Wu**

r.wu@bbk.ac.uk  
Centre for Brain and  
Cognitive Development  
Birkbeck, University of London

## Abstract

The study of cognitive development hinges, largely, on the analysis of infant looking. But analyses of eye gaze data require the adoption of *linking hypotheses*: assumptions about the relationship between observed eye movements and underlying cognitive processes. We develop a general framework for constructing, testing, and comparing these hypotheses, and thus for producing new insights into early cognitive development. We first introduce the general framework – applicable to any infant gaze experiment – and then demonstrate its utility by analyzing data from three studies investigating the role of attentional cues in infant learning. Finally, we discuss general implications for construction and testing of quantitative linking hypotheses.

Keywords: eye movement data; infancy methods; Bayesian data analysis; learning; attention

## Introduction

The study of infant cognitive development hinges, largely, on the analysis of infant looking data (Aslin, 2007). Since Fantz’s (1964) landmark demonstration of visual memory in 2-month-old infants, researchers have used his habituation technique, and other eye-movement methods, to ask deep theoretical questions about the ontogeny and development of human cognition. But analysis of eye-movements, like analysis of other high-dimensional cognitive measures (e.g. fMRI, EEG) carries particular challenges (Yu, Yurovsky, & Xu, 2012). In order to connect observed eye-movements to underlying cognitive processes, one must define a linking hypothesis that relates them (Aslin, 2007; Teller, 1984).

Every eye gaze paradigm used to study infant cognition commits to a particular *linking hypothesis*. In habituation studies, decreased looking is hypothesized to indicate encoding, and recovery of looking indicates discrimination of a novel stimulus (Gilmore & Thomas, 2002). In violation of expectation studies, increased looking is hypothesized to indicate noticing a surprising event. Intermodal preferential looking studies hypothesize that a difference in looking time to one sound-object mapping over another indicates a difference in their associations. Critically, these linking hypotheses are *qualitative*; they assert that a relationship exists, but do not specify its *quantitative*, metric properties.

Why should we prefer quantitative linking hypotheses? They help us, in several ways, to move from asking *if* a phenomenon occurs, to asking *how* and *why*. First, quantitative linking hypotheses allow researchers to clearly and unambiguously specify the assumptions and

mechanisms in their theories. As theories grow in complexity, correctly deriving their (sometimes counterintuitive) predictions can become difficult. Formalizing theories makes such prediction tractable (Shiffrin, 2010). Second, without quantitative predictions it can be impossible to distinguish competing theoretical accounts of the same data, fueling debates about “rich” (conceptual) vs. “lean” (perceptual) theoretical explanations (e.g., Spelke, 1998). Third, quantitative linking hypotheses allow researchers to test the same theoretical model across experiments, integrating multiple datasets within one self-consistent framework (Aslin, 2007; Schöner, & Thelen, 2006; Shiffrin, 2010).

Developmentalists who measure eye-movements, however, face several challenges to the construction of quantitative linking hypotheses. First, control of eye-movements is complex, and saccades are moderated by multiple systems (Aslin, 2007). Thus, quantitative linking hypotheses may need to integrate interacting mechanisms. Second, although fixation duration is likely related to learning, their relationship may not be a simple linear one. Instead, learning and looking may be linked non-monotonically, with a preference for familiarity appearing first, and a preference for novelty developing with further experience (Hunter & Ames, 1988). Linking hypotheses must be flexible enough to accommodate this kind of complexity. Third, early development is a time of rapid change, and the variability among infants of the same age may be surprisingly high. Thus, using the same linking hypothesis for each infant may distort true relationships in the data (Siegler, 1987). Because one cannot know apriori whether one’s data is best analyzed as one group, or two, or three or more, construction of linking hypotheses must adaptively accommodate this kind of variability.

Building on a growing body of statistical tools in Bayesian non-parametrics, this paper presents a rigorous, principled, empirically successful framework for the construction of quantitative linking hypotheses that meets the three challenges reviewed above. To demonstrate the utility of this framework, we analyze data from a set of experiments investigating the role of social and non-social cues in infant multi-modal learning (Wu & Kirkham, 2010). This analysis shows how quantitative linking hypotheses can provide leverage in understanding the development and operation of infant learning mechanisms. We begin by presenting the general framework, demonstrating its robustness in simulation studies, and then present the empirical data.

## General Model Framework

In any eye-tracking experiment, infants are exposed to stimuli that encode some structure of theoretical interest, and the researcher measures the influence of this structure on their behavior. For instance, in word-learning experiments, infants are exposed to consistent pairings between words and objects, and their discrimination for consistent vs. inconsistent mappings is measured (e.g. Yu & Smith, 2011). However, we are typically interested not in the change in observed behavior, but rather in the cognitive processes it implicates (Aslin, 2007). Quantitative linking hypotheses let us describe these processes directly.

For each infant, on each trial, the researcher observes some eye-gaze data ( $D$ ), and the researcher's goal is to determine the model ( $M$ ) that best explains these observed eye-movements ( $P(M|D)$ ). This can be formalized as a problem of Bayesian inference. The researcher can specify several possible models, each making different predictions about the gaze data likely to be observed ( $P(D|M)$ ). The researcher may also prefer simpler models apriori, in accord with Ockham's razor ( $P(M)$ ). These properties can then be combined via Bayes' rule to infer the model that best describes the infants' cognitive processes (Equation 1).

$$P(M|D) \propto P(D|M)P(M) \quad (1)$$

We present a graphical model (Figure 1) for connecting hypothesized cognitive models to observed eye gaze data. On each trial of an experiment, an infant ( $i$ ) is exposed to some experimental stimuli ( $e$ ) and produces observed eye movements ( $d$ ). This observed gaze data is encoded as proportion of dwell time over a set of hypothesized areas of interest (AOIs). The inference framework discovers the set of underlying cognitive processes ( $s$ ) that operate on the stimuli to generate the observed data. Intuitively, this is essentially a regression problem: inference finds the relationship between predictor variables ( $s, e$ ) and observed outcomes ( $d$ ). Because gaze data are a distribution over AOIs rather than a single continuous variable, we connect predictors to outcomes via the Dirichlet distribution ( $\theta$ ).

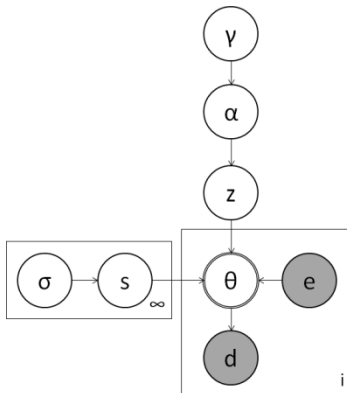


Figure 1: A graphical model for inferring the cognitive processes ( $s$ ) responsible for generated eye movements ( $d$ ) under particular experimental conditions ( $e$ ).

The introduction identified three challenges for quantitative linking hypotheses: multiple processes may drive eye-movements, linking functions may be complex, and a group of infants may be heterogeneous. This framework meets all three challenges. Because  $s$  can encode any hypothesized cognitive model, the contributions of multiple processes can be estimated together without forcing a dichotomy (Anderson, 2011). Nonetheless, if a process has little effect, this is found via the prior on parameter values  $\sigma$  (Figuerido, 2002). Second, cognitive processes and observed eye movements need not be linked in a simple, linear way. In this framework, the cognitive model  $s$  can encode any functional link. For simplicity, and to minimize assumptions, we do so through arbitrary degree polynomials (see Jackson & Sirois, 2009). Again, the model parameter prior ( $\sigma$ ) facilitates discovery of the most parsimonious linking function, penalizing complex polynomials.

Formally, each cognitive model parameter  $s$  is modeled as a draw from a 0-mean normal distribution whose standard deviation has a non-informative prior, making high values unlikely (Jeffreys, 1961). Each infant's data are modeled as a draw from a Dirichlet distribution over the AOIs whose parameters are defined as the exponentiated product of the cognitive model parameters and experimental settings  $e$  (Equations 2). This allows model parameters to be negative. A specific formulation is presented in the next section.

$$\begin{aligned} \sigma &\sim \frac{1}{|\sigma|} \\ s &\sim N(0, \sigma^2) \\ \theta_z &= e^{(s_z \times e)} \\ d &\sim \text{Dir}(\theta_z) \end{aligned} \quad (2)$$

Finally, infants in a sample may come from two or more different kinds of groups, (e.g. slow and fast learners: Yu & Smith, 2011). This framework automatically and adaptively determines the number of groups of infants, and the infants who belong to each group. Each distinct group is best represented by a different cognitive model. The estimation of unique groups is performed using the Chinese restaurant process (Aldous, 1985), which has been used successfully to determine unique groups in adult experiments (Navarro, Griffiths, Steyvers, & Lee, 2006). Clusters are discovered in this process by treating participants by analogy to customers in a Chinese restaurant. As each customer enters, he sits at each occupied table ( $z$ ) with probability proportional to the number of occupants, but also chooses a new table with some small probability ( $\alpha$ ). This implements a rich-get-richer scheme in which groups that account for the behavior of many infants become favored, and the most parsimonious number of groups is discovered. A hyper parameter ( $\gamma$ ) prevents us from having to make a direct decision about the probability of choosing a new table (Equations 3). Each cluster has different cognitive parameter values ( $s$ ).

$$\begin{aligned} \alpha &\sim \text{Exp}(\gamma) \\ z &\sim \text{CRP}(\alpha) \end{aligned} \quad (3)$$

## Case Study: Attentional Cues and Infant Learning

To demonstrate how this framework’s utility in a concrete case, we applied it to data from a set of studies investigating the role of attentional cues in infant multi-modal learning. In each experiment, 8-month-old infants watched videos in which sounds and objects’ on-screen locations were reliably related. When objects appeared in the top-left and bottom-right boxes, one sound was heard. When they appeared instead in the top-right and bottom-left boxes, a different sound was heard (Figure 2). In some conditions, infants were cued to one of the two objects. Subsequently, infants were exposed to test trials on which they heard a sound from training, but all four boxes were blank. If infants had learned sound-location regularities, they were expected to attend preferentially to locations consistent with each sound.

Submitting these test preferences to ANOVAs, Wu and Kirkham found reliable multi-modal learning only in the presence of the Face cue (2a), but not when infants were cued with a flashing square (2b) or received no cue (2c). We reanalyze this data to reveal significantly more structure, and to provide new insights into infant learning.

To this end, we define quantitative linking hypotheses for these experiments, formally specifying the connection between the observed eye-movement data ( $d$ ), observable experimental conditions ( $e$ ), and the unobservable, hypothesized cognitive processes ( $s$ ). By analogy to regression, the data are the dependent variable, experimental conditions are the independent variables, and the cognitive processes parameterize these independent variables. On each trial of the experiment – whether training or testing – infants saw a black screen containing four boxes, one in each corner of the screen (Figure 2). Thus, we define five areas of interest (AOIs): one for each of the four boxes, and a fifth to capture all other looks (including off-screen looks).

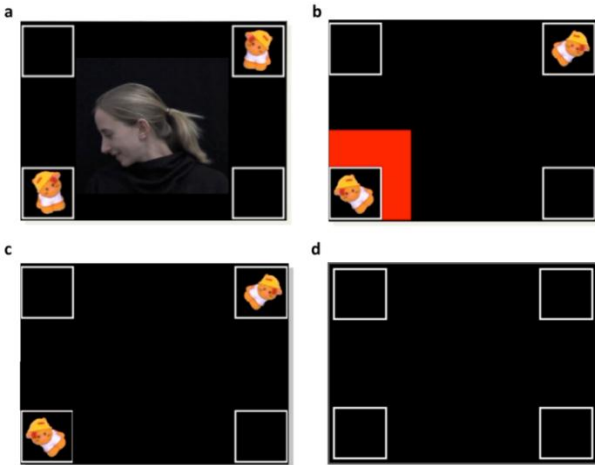


Figure 2: Training and testing trials from Wu & Kirkham (2010). In the Face condition (2a), a centrally-located face directed infants’ attention to one of the boxes. In the Square condition (2b), a red flashing square highlighted one of the boxes. In the No Cue condition (2c) only the multi-modal regularity was present. On test trials (2d), all boxes remained empty while infants heard a sound from training.

The total data ( $d$ ) for an individual infant is thus the entire set of gaze proportions observed on each trial. Formally, this is a matrix in which rows correspond to trials, columns to AOIs, and each cell to the proportion of looking to a particular AOI on a particular trial. This whole matrix is the outcome to be predicted from the experimental conditions ( $e$ ) and hypothesized cognitive processes ( $s$ ).

Next we specify the experimental conditions on each trial. While all four boxes were empty on test trials, on training trials two of the four boxes contained pictures of animals (Figure 2a-c). These are coded with binary indicator variables *salient* specifying whether a box ( $b$ ) contains a picture. Further, in the Face and Square conditions (Figure 2a-b), one of the boxes was highlighted by an attentional cue. We similarly define an indicator variable *cued*.

$$\begin{aligned} \text{salient}(b) &= \begin{cases} 1 & \text{box } b \text{ contains stim} \\ 0 & \text{otherwise} \end{cases} \\ \text{cued}(b) &= \begin{cases} 1 & \text{box } b \text{ cued} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (4)$$

In addition to the visual stimuli, each trial also included a sound that could alter looking patterns if infants learned sound-location contingencies. To formalize this learning process (below), we encode infants’ experience with these contingencies in the experimental conditions ( $e$ ). Thus, we let *contingent* be the cumulative time an infant has fixated a given box ( $b$ ) in the presence of a particular sound ( $n$ ). So, on trial  $t$  that plays sound  $n_t$  and on which we observe data  $d_t$ , where  $\delta(i, j)$  is Kronecker’s delta function,

$$\text{contingent}_t(b, n_t) = \sum_{r=1}^{t-1} \delta(n_r, n_t) \cdot d_r(b) \quad (5)$$

Last, we define cognitive processes that act on experimental condition variables to produce observed gaze data. First, infants may have a baseline preference for some locations over others. Thus, we include a preference constant  $p_b$  for each AOI, allowing the contributions of the other variables to be estimated relative to proper baselines. Second, an infant’s preference for a box may be altered by the presence of an object (*salient*), or the presence of an attentional cue (*cued*). We let the strength of these factors be linearly scaled by parameters  $l$  and  $c$  respectively, which function like slope terms in linear regression.

Finally, in these experiments, the question of interest is whether infants learn to associate sounds with objects/locations. We define the effect of association between a sound and location as a change in preference for that location through exposure to the contingent sound. Specifically, we define association between a sound and location as a function of time spent fixating that location in the presence of that sound (*contingent*). To avoid making assumptions about the association function (e.g. that it is linear, or monotonic), we let association between box  $b$  and sound  $n$  on trial  $t$  be an arbitrary degree polynomial function of cumulative looking time to  $b$  while hearing  $n$ .



Since polynomials can approximate any functional form, this is a general solution (Jackson & Sirois, 2009). As in testing for higher-order terms in standard regression, polynomial coefficients are pushed down to zero by model priors if they do not contribute to predictive power.

$$assoc_t(b, n) = \sum_{o=1}^o a_o \cdot [contingent_t(b, n)]^o \quad (6)$$

After formally specifying the experimental conditions and hypothesized cognitive process that act on this input, we can infer the effect of each hypothesized factor on infant the gaze data. As in regression, differences in parameters across conditions help us understand whether and how different cues affect infant multi-modal learning. To infer parameter values, we perform Bayesian inference in the model specified in Figure 1. Because this model has non-conjugate priors, we use an MCMC sampling algorithm that alternates Metropolis-Hastings updates with Split/Merge steps for cluster assignment (Jain & Neal, 2007). Sampling estimates the true distribution for each of these parameters, producing a set of credible intervals (similar to confidence intervals) that can be used to determine the likelihood that parameters are non-zero, as well as their likely range (Kruschke, 2011). Clustering was relatively insensitive to  $\gamma$ , so we let  $\gamma = 1$ .

## Simulations

In order to ensure that it behaves as expected, we validate the analysis in a set of simulation studies by generating gaze data from a known cognitive model and trying to recover its parameters. In these simulations, we show that this framework can deal with all three challenges for quantitative linking hypotheses: non-homogenous samples of infants, interactions among multiple cognitive processes, and non-linear functions linking learning to looking.

Infants in Wu & Kirkham’s (2010) study were simulated in training and testing trials like those in their experiments. Each simulated infant was exposed to four consecutive blocks, each consisting of six training trials and a test trial. On each training trial, objects appeared in two of the boxes (top-left and bottom-right, or top-right and bottom-left), and the lower box was cued. Each configuration of objects also co-occurred with a unique sound. Each of the two configurations occurred three times in each block of training trials, and order was randomized within a block. After all six training trials, infants saw one test trial where the screen was empty, but one of the two sounds was heard. Simulated infants then saw three more blocks, and each sound was tested twice in random order across the four test trials.

### Simulation 1

In Simulation 1, we generated gaze data from known models in which the infants in a sample were drawn from a mixture of one, two, three, or four distinct groups. Formally, the 30 infants on each run were drawn from a multinomial distribution with equal probability for each group.

Parameters for each group were drawn randomly without replacement from *cued* –  $c$ : {0, 1, 2, 3}, *salient* –  $l$ : {0, 1, 2, 3}, and *contingent* –  $a_1$ : {0, .2, .4, .6}. Baseline AOI preferences for each box were drawn uniformly from  $[-2, -1]$  and off-screen preference was drawn from  $[.5, 1.5]$ . These values were representative of those found in the empirical analysis (next section).

Across all 120 simulations (30 runs at each group size), the correct number of groups was identified in all but 1. On one run at group size 4, the analysis identified only 3 clusters. Further, individual infants were almost always assigned to the right group. Group assignment was perfect when the number of true groups was 1 or 2, and less than a quarter of one percent ( $<.0025$ ) of infants were misclassified at the higher group numbers. Thus, this framework deals well with heterogeneous groups of infants.

### Simulation 2

Simulation 2 tested the framework’s ability to recover correct quantitative parameter values when multiple processes interacted to produce eye movements. This time, all infants were drawn from one group, but group parameters were parametrically manipulated to sample the space of parameters recovered in the analysis of Wu & Kirkham’s empirical data. Six unique parameter values were chosen for each hypothesized cognitive processes, and one simulation was run at each combination. Baseline preferences on each run were drawn as in Simulation 1. Figure 3 (next page) shows parameter estimates and true values for each combination of parameter values. Inference was successful:  $r^2$  values were exceedingly high.

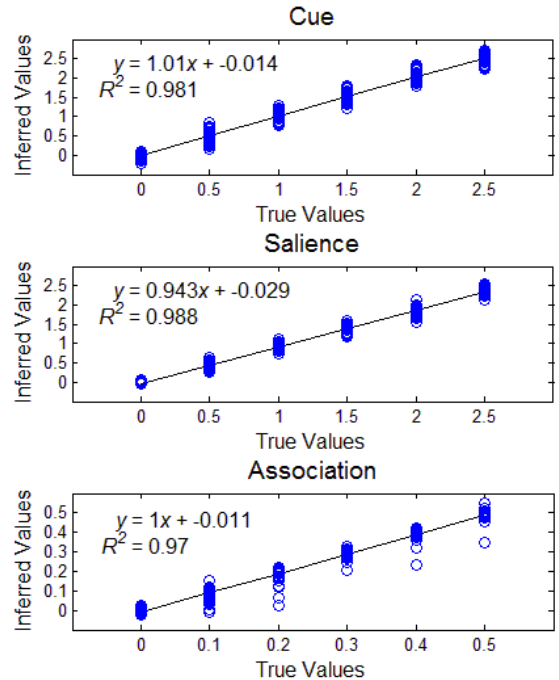


Figure 3: Best fit lines for true and inferred parameter values for each of the three factors hypothesized to affect infant gaze patterns in the experimental data.

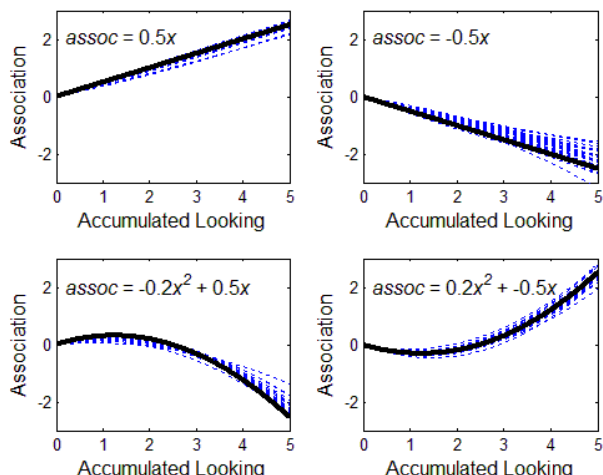


Figure 4: True functions (solid) and 30 inferred functions (dashed) for each learning function tested in Simulation 3.

### Simulation 3

Simulation 3 tested the framework’s ability to recover non-monotonic learning functions, for instance, a preference for familiarity followed by novelty. On each run of the simulation, 30 infants were generated with one of four possible learning functions: linear increasing, linear decreasing, u-shaped up, and u-shaped down (Figure 4). All other model parameters were drawn as in Simulation 2. Bayesian 95% credible intervals for estimated parameters were analyzed to determine how often a 0-valued parameter was estimated to be non-zero (0%) and how often a non-zero parameter was estimated to be 0 (2.5%). Thus, both Type I and Type II error rates were low (Kruschke, 2011).

### Empirical Analysis

We now apply the same inference procedure to gaze data from real infants. Instead of comparing the effects of different attentional cues on raw looking patterns, as in standard analyses (e.g. ANOVAs), inferring cognitive model parameters for each condition lets us analyze the effects of cues directly on attention and learning.

### Method

Inference was performed using the same model as in the Simulations above. Infants each saw a series of training and testing trials, and gaze data on each trial were coded as a proportion of looking to each on-screen box as well as a fifth AOI for all other looks (Figure 2). As before, inference recovered the joint distribution for all parameters ( $s$ ) explaining looking as a function of experimental conditions. Full parameter descriptions are in the Simulation section. Gaze data included 26 infants in the No Cue condition, 29 infants in the Face condition, and 30 infants in the Square condition (see Wu & Kirkham, 2010 for full details).

Before presenting the results, we review Wu & Kirkham’s ANOVAs for test trial looking. These analyses showed associative learning only in the Face condition. In contrast, infants in the No Cue condition showed no learning, and

infants in the Square condition preferred the cued locations, but did not learn to sound-location associations.

### Results and Discussion

Inference yields full posterior distributions for all cognitive model parameters, estimating the contribution of each factor in the context of all other factors. We focus on two key factors: attention to the cue ( $c$ ) and the association function ( $assoc$ ). Figure 5 shows estimated parameter values for both factors for infants in each experimental condition.

First, in no condition were infants best described as a single homogeneous group. Two distinct groups were identified in the Face and No Cue conditions, and four groups were found in the Square Condition. Thus, even within one condition, infants learned and used cues differently. Second, all learning functions were linear; credible intervals for all association coefficients  $\geq 2$  overlapped 0 in all conditions. Thus, Figure 5 shows the first-order association coefficient ( $a_1$ ) for each group.

Finally, we turn to the parameter values and their implications. First, all infants in the No Cue condition appeared to be learning ( $a_1 > 0$ ), although approximately  $\frac{2}{3}$  had low association values, indicating that they learned slowly. The Face condition had a comparable number of equally fast learners, and these fast learners did not show evidence of using the cue ( $c \approx 0$ ). However, the larger, slow group of learners did use the cue, and learned faster than the slow learners in the No Cue condition. Learners had two routes into learning the regularity: quickly and directly, or slowly and indirectly. This detailed level of structure underlies and explains Wu & Kirkham’s coarser analysis.

The Square condition also had a small group of fast learners who used the cue. However, in contrast to the other conditions, approximately  $\frac{1}{2}$  of the infants did not learn, and these infants all used the cue ( $c > 0$ ). These results directly confirm Wu and Kirkham’s hypothesis that the flashing square may interfere with learning by competing for attention, and that only the fastest learners may be able to learn from these kinds of competing cues. Together, these results both confirm the major findings from the standard analysis and provide deeper insight into how attentional cues guide (or interfere with) infant multi-modal learning.

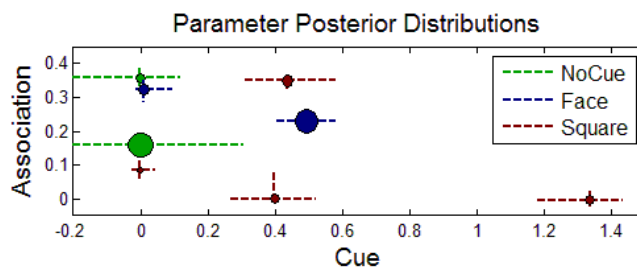


Figure 5: Posterior distributions for *cue* ( $c$ ) and *association* ( $a_1$ ) parameters for infants from Wu & Kirkham (2010). Each circle indicates a cluster, and its size indicates the proportion of infants in that condition in that cluster. Circles are centered at median parameter values, and dashed lines indicate 68% credible intervals, akin to  $\pm 1$  SE.



**Non-monotonic learning functions.** Simulation 3 showed that this framework can recover non-monotonic functions linking learning to looking when appropriate for the data (Hunter & Ames, 1988). However, no such functions appear in the Empirical Analysis above. Why? One possibility is that non-monotonic linking functions arise in a different kind of experiment or at a different age. An alternative possibility is that non-monotonic linking functions are seen when infants' baseline preferences are not controlled. In our analyses, we included a set of parameters  $p_b$  to encode baseline preferences for each location. When these parameters were not included, we *did* find non-monotonic linking functions in all conditions. Thus, we propose that, at least in some cases, observation of non-monotonic linking functions may be an artifact of different baseline preference rather than a core property of the learning system itself.

**Competing Hypotheses.** One strength of quantitative linking hypotheses is that they facilitate direct comparison of competing theories for the same data. In the previous sections, we argued that changes in looking preferences over the course of these experiments arise from associations between heard sounds and fixated locations, and modeled this learning with the *assoc* function. Alternatively, preferences could change over time through habituation; infants' preferences could change as a function of looking to a location independent of the concurrent sound. We tested this directly, by modeling habituation as an arbitrary-degree polynomial function of cumulative looking time to a location (Equation 7). However, 95% credible intervals for *habit* parameters overlapped 0 in all conditions, out this explanation for the data. Thus, quantitative looking hypotheses allowed us to directly compare two hypothetical explanations of this data and to choose the best alternative.

$$habit_t(b) = \sum_{o=1}^o h_o \cdot \left( \sum_{r=1}^{t-1} d_r(b) \right) \quad (7)$$

## General Discussion

Infant researchers have made tremendous progress by using eye gaze data to ask questions about early cognition and development. The majority of this work has used *qualitative* linking hypotheses, but we propose that even faster progress can be made through model-based analyses using *quantitative* linking hypotheses (Aslin, 2007; Teller, 1984). While quantitative linking hypotheses have been proposed for specific experiments (e.g. Gilmore & Thomas, 2002; Yu & Smith, 2011), this paper presents a general framework applicable to all eye movement experiments. We hope this work will facilitate asking and answering future questions about early cognitive processes and their development.

## Acknowledgments

This work was supported by a NSF GRF and NSF EAPSI to DY, a BPS Postgraduate Study Visits Award to RW, and Grant-in-Aid for Scientific Research to SH. We thank Natasha Kirkham, and the Smith, Yu, and Shiffrin labs.

## References

- Aldous, D. (1985). *Exchangeability and related topics*. In *École d'été de probabilités de Saint-Flour, XIII—1983* (pp. 1–198). Berlin: Springer.
- Anderson, B. (2011). There is no such thing as attention. *Frontiers in Psychology*, 2, 1-8.
- Aslin, R. N. (2007). What's in a look? *Developmental Science*, 10, 48-53.
- Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, 146, 668–670.
- Figuerido, M. A. T. (2002). Adaptive sparseness using Jeffreys prior. *Advances in Neural Information Processing Systems*, 14, 722-729.
- Gilmore, R. O. & Thomas, H. O. (2002). Examining individual differences in infants' habituation patterns using objective quantitative techniques. *Infant Behavior and Development*, 25, 399-412.
- Hunter, M.A., & Ames, E.W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. In L.P. Lipsitt (Ed.), *Advances in child development and behavior* (pp. 69-95). New York: Academic Press.
- Jackson, I., & Sirois, S. (2009). Infant cognition: going full factorial with pupil dilation. *Developmental Science*, 12, 670-679.
- Jain, S., & Neal, R. M. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis*, 2, 445-472.
- Jeffreys, H. (1961). *Theory of probability*. Oxford, England: Oxford University Press.
- Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6, 299-312.
- Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet Processes. *Journal of Mathematical Psychology*, 50, 101-122.
- Schöner, G., & Thelen, E. (2006). Using dynamic field theory to rethink infant habituation. *Psychological Review*, 113, 273-299.
- Shiffrin, R. M. (2010). Perspectives on modeling in cognitive science. *Topics in Cognitive Science*, 2, 736-750.
- Siegler, R. S. (1987). The perils of averaging over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, 116, 250-264.
- Spelke, E. S. (1998). Nativism, empiricism, and the origins of knowledge. *Infant Behavior and Development*, 21, 181-200.
- Teller, D. Y. (1984). Linking propositions. *Vision Research*, 24, 1233-1246.
- Wu, R., & Kirkham, N. Z. (2010). No two cues are alike: Depth of learning during infancy is dependent on what orients attention. *Journal of Experimental Child Psychology*, 107, 118-136.
- Yu, C., & Smith, L. B. (2011). What you learn is what you see: Using eye movements to study infant cross-situational word learning. *Developmental Science*, 14, 165-180.
- Yu, C., Yurovsky, D., & Xu, T. (2012). Visual data mining: An exploratory approach to analyzing temporal patterns of eye movements. *Infancy*, 17, 33-60.

# Mutual Exclusivity and Vocabulary Structure

**Daniel Yurovsky**

[dyurovsk@indiana.edu](mailto:dyurovsk@indiana.edu)

Department of Psychological and Brain Sciences  
Indiana University

**Linda B. Smith**

[smith4@indiana.edu](mailto:smith4@indiana.edu)

Department of Psychological and Brain Sciences  
Indiana University

**Ricardo A. H. Bion**

[ricardoh@stanford.edu](mailto:ricardoh@stanford.edu)

Department of Psychology  
Stanford University

**Anne Fernald**

[afernald@stanford.edu](mailto:afernald@stanford.edu)

Department of Psychology  
Stanford University

## Abstract

The words that children learn can be characterized as a semantic network, with links connecting related words. Recent analyses have shown these networks to have small-world structure, with a few highly-connected hub words facilitating short paths between otherwise distant words. This structure contributes to network robustness, and differences in structure can predict differences in language learning outcomes. While previous studies have shown that semantic network structure reflects linguistic input structure, we provide the first evidence that it is related also to children's own language learning biases. Two-year old children who show a mutual-exclusivity bias have significantly more hub-like networks than children who do not, even when they know the same number of words. This finding contributes to our understanding of both semantic networks and the origins of mutual exclusivity.

**Keywords:** word learning; mutual exclusivity; semantic networks; language acquisition

## Introduction

Although the earliest analyses of human memory and learning concerned the learning of lists of unrelated words (Ebbinghaus, 1885/1962), researchers quickly discovered that the words people learn in more natural contexts are intricately connected. Vocabularies were conceptualized as richly structured networks, with links connecting semantically related words (Collins & Loftus, 1975). These connections play an important role in both learning and memory, and can be observed empirically in semantic priming experiments. Because activation spreads from words to their semantic neighbors, presenting a word, even subliminally, leads to faster processing of related words (Anderson, 1983). Even two-year old infants show semantic priming, suggesting that vocabularies have network structure early in language learning (Arias-Trejo & Plunkett, 2009).

Recently, the application of graph-theoretic methods to the study of these networks has begun to provide insight into their structural properties. For instance, Hills, Maouene, Maouene, Sheya, & Smith (2009a) analyzed the semantic network structure of 130 nouns typically learned before 30 months. Compared to randomly-connected control networks, these semantic networks showed significant small-world structure, in which most words are sparsely connected, but a few are highly-connected hubs. This kind

of structure results in networks robust to malfunction (e.g. forgetting a word; Albert, Jeong, & Barabási, 2000), and can help to explain some of the remarkable efficiency of human semantic memory (Raaijmakers & Shiffrin, 1981). Further, semantic networks lacking this structure are associated with slower language-learning, characterizing the vocabulary structure of late talkers (Beckage, Smith, & Hills, 2011). But why do children learn these words? Why do semantic networks have this structure?

Undoubtedly, one answer to this question is that structure comes from the environment. Because children learn words from the language they hear, language input is a strong predictor of the words that children will learn. For instance, the frequency with which a child hears a word in isolation can predict how likely a child is to learn that word (Brent & Siskind, 2001). Similarly, the semantic networks constructed from corpora of both adult-directed and child-directed language have many of the same structural properties as networks constructed from the words 30-month-old children are likely to know (Hills, et al., 2009a; Steyvers & Tenenbaum, 2005).

But perhaps a more complete explanation of the origin of semantic network structure is that it emerges from an interaction between structure in the linguistic environment and the child's own learning system. Because children are not unbiased samplers of linguistic input, their attentional and learning biases mediate the link between language input and language learned (Hudson Kam & Newport, 2005; Smith, 2000). For instance, children who learn to attend to shape are likely to learn shape-based categories, and those who learn to attend to other properties (e.g. material) learn other kinds of words (Colunga & Sims, 2011; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Can word-learning biases predict and explain semantic network structure? In this paper, we consider the case of disambiguation through mutual exclusivity.

In the disambiguation task, a child is presented with a novel object among one or more familiar object competitors. The child then hears a novel label (e.g. 'can you find the dax?') and is asked to select an object. Both toddlers and adults reliably select the novel object as the target of the novel label (Markman & Wachtel, 1988; Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992), and studies with infants suggest that this disambiguation may arise as early as 18 to 22.5 months (Halberda, 2003; Mather & Plunkett, 2009). Preferential

mapping of novel labels to novel objects over known objects, which we will refer to as mutual exclusivity (ME), could arise for a number of reasons, and its mechanism of action is the topic of significant debate (e.g. Diesendruck & Markson, 2001; Golinkoff et al., 1992; Markman & Wachtel, 1988). We explore this question in the general discussion, but will sidestep it here and instead consider the potential consequences of mutual exclusivity for semantic network structure.

Mutual exclusivity is a mechanism by which children can leverage prior knowledge to learn new words *in the context of known objects*. Consequently, children who show mutual exclusivity should have vocabularies that echo this kind of contextual structure. For these children, learning *fork* should ease the acquisition of *spoon*, *bowl*, and *plate*. In contrast, learning *fork* should have little effect on the acquisition of *dog* and *coat*. Thus, we propose that mutual exclusivity can help explain small-world structure of semantic networks, and those children who show mutual exclusivity will have more hub-like networks than those who do not. We begin by reporting empirical data from a disambiguation task with 24-month-old children, continue by describing a semantic network analysis of these children's vocabularies, and conclude with a discussion of how these results inform our understanding of the relationship between mutual exclusivity and vocabulary development, as well as the origins of mutual exclusivity itself.

## Experiment

### Method

**Participants.** Forty two-year-olds ( $M = 24.75$  months; range = 24-26; 20 female) participated. All were typically developing children from households in which parents reported English to be the dominant language. A subset of 34 infants ( $M = 24.9$  months; range = 22.4-27.5; 16 female) participated in the followup analysis (explained below).

**Stimuli.** Nine familiar objects (e.g. boat, glasses) were used in the warm-up trials. Twenty-five familiar (e.g. brush, cup) and 8 novel objects (e.g. massager, platypus) were used in the referent-selection task.

**Procedure.** Parents first completed the MCDI (Fenson, Dale, Reznick, Bates, Thal, & Pethick, 1994) and an SES measure (Hollingshead, 1975). After this, each child participated in three warm-up trials. On warm-up trials, the experimenter set a tray containing three familiar objects on the table, initially covered by an occluder. The experimenter asked for the target object (e.g. "which one is the dog?") three times: once while the items were occluded, again after lifting the occluder, and again three seconds later while pushing the tray towards the infant. The first reach, point, or grab, was scored as a response. On these trials, infants were praised for correct responses and corrected when necessary.

Subsequently, each child participated in sixteen referent-selection trials. On each trial, the experimenter presented a tray containing two familiar objects and one novel object.

The procedure was identical except that children received neutral feedback on all trials. On half the trials, the experimenter asked for a familiar object, while on the other half she asked for a novel object (e.g. *modi*, *taju*).

## Results and Discussion

Each participant made a total of 16 choices, picking 8 targets on familiar trials, and 8 targets on novel trials. Any trial on which the child did not know the label for the familiar target, or the label for one of the familiar distractors, was excluded from analysis. The proportion of targets correctly chosen on these remaining trials was then analyzed to determine the child's success in the task. Overall, children performed quite well, selecting the correct target on both familiar ( $M_f = .83$ ,  $t(39) = 15.31$ ,  $p < .001$ ) and novel trials ( $M_n = .545$ ,  $t(39) = 5.87$ ,  $p < .001$ ) at greater than chance levels. Thus, as a group, 24-month-old children used mutual exclusivity for disambiguation. Familiar trial performance, however, was significantly higher than novel trial performance ( $t(39) = 6.88$ ,  $p < .001$ ).

Because the central question in this study is about the relationship between learning mechanisms and vocabulary development, we measured both vocabulary size (MCDI - Fenson, et al., 1994) and mother's education (Hollingshead, 1975), a potential correlate of rich language input. Mother's education was reliably correlated with performance on familiar trials ( $r = .33$ ,  $p < .05$ ), but not novel trials ( $r = .01$ , *n.s.*), and vocabulary size was not significantly correlated with performance on either kind of trial ( $r_f = .19$ , *n.s.*;  $r_n = .15$ , *n.s.*). In the semantic network analysis to follow, we show that *vocabulary structure* is reliably related to novel trial performance. Because neither mother's education nor *vocabulary size* predict ME in this data set, the relationship between ME and structure is likely to be quite robust.

But perhaps this analysis is unfair. While most of the children had high levels of success on familiar trials, a few children did not perform as well. Since these children knew the words for all three objects on these familiar trials, their low levels of performance indicate that they may not have understood the task. Thus, for the same reason that response time analysis typically uses only correct response trials, excluding these children from individual-level analyses may give clearer correlations. In order to determine whether a child's performance was significantly better than expected by chance, we modeled chance behavior on each trial as random selection of one of the three objects.

The probability of success expected by chance is given by a binomial distribution with probability  $\frac{1}{3}$ . Consequently, a child should be counted as performing differently from chance if he or she made enough correct selections to be outside the 95% confidence interval for a binomial distribution. A child who made 8 choices, for instance, needed to make at least 5 correct choices to be counted as performing better than expected by chance. Each child's number of correct selections on familiar trials was thus submitted to a binomial test. Six of the 40 children were found to have performance levels on the familiar trials

indistinguishable from chance, and were thus excluded from further analysis. This left a subset of 34 children who could confidently be assumed to have understood the task. Figure 1 shows novel and familiar trial performance for children both from the full set, and from this reduced subset.

We also performed a similar analysis on novel trials, dividing children into two categories: those who reliably showed evidence of using mutual exclusivity (ME), and those who did not. Seventeen children were classified as ME users, and seventeen were classified as Nonusers. We are not arguing that ME is a binary phenomenon, but rather perform this binary split for technical reasons. Binarization loses some information separating children within the ME users category, but it also cleans up noise that may not meaningfully separate nonusers. Quantitative differences at or below chance levels are more likely to be generated by noise than they are to be generated by meaningful process differences, and thus are likely to dilute linear correlations. In subsequent analyses, because mutual exclusivity is analyzed as a binary phenomenon, we use Spearman's  $\rho$ , a non-parametric measure of correlation. In all cases, correlations were stronger for this binary measure.

In this subset, mother's education was still correlated with performance on familiar trials ( $r = .36, p < .05$ ), as was vocabulary size ( $r = .39, p < .05$ ). Neither mother's education nor vocabulary size predicted performance on novel trials ( $\rho = -.11, n.s.$ ;  $\rho = .08, n.s.$ ). In the analyses that follow, we compare the semantic network connectivity of ME users and nonusers. Because use of mutual exclusivity was uncorrelated with vocabulary size, differences in network connectivity are unlikely to be a simple reflection of network size. Further, because mother's education predicted performance on familiar, but not novel, trials, a relationship between ME and vocabulary structure arising from language input must come from more specific properties not indexed by mother's education in this sample.

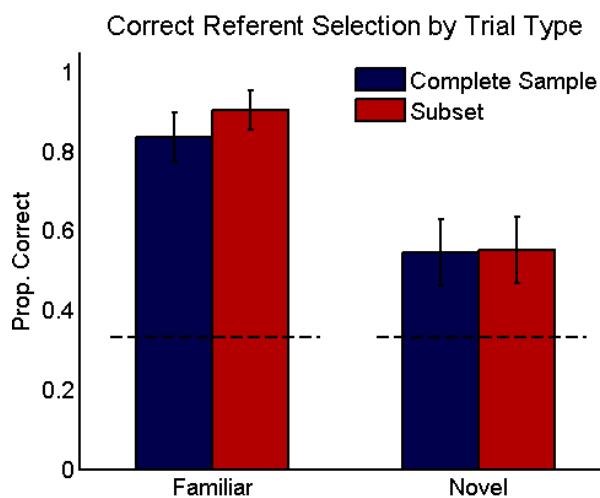


Figure 1: Proportion of correct choices by participants in both the familiar and novel conditions. Dark blue bars show the complete sample, light red bars the subset. Error bars indicate  $\pm 1$  standard error.

## Semantic Network Analysis

To understand how use of mutual exclusivity contributes to the structure of children's vocabularies, we formalize these vocabularies as semantic networks. In semantic networks, vertices represent the words that children know, and edges represent semantic relationships among these words. In any such analysis, the first step is to formalize 'semantic relatedness' – the relationship used to link two words.

Previous analyses have used a number of successful metrics of connectivity: co-occurrence in CHILDES (e.g. Beckage, Hills, & Smith, 2011), frequency of free-association by adults (e.g. Steyvers & Tenenbaum, 2005), and common perceptual and conceptual features (e.g. Hills, et al., 2009b). In our analysis, we adopt and extend the last approach, connecting two words if they share a number of common semantic features. Features were drawn from the set of McRae feature norms (McRae, Cree, Seidenberg, & McNorgan, 2005). McRae and colleagues asked 725 adults to freely list up to 14 features of 541 English nouns. The number of features shared by two words gives a measure of their semantic relatedness.

Although participants could generate any features they liked, McRae et al. (2005) subsequently divided the generated features into 4 categories: perceptual features accessible to the 5 senses (e.g. "has fur," "tastes sweet"), functional features (e.g. "used for writing," "is edible"), encyclopedic features (e.g. "is expensive"), and taxonomic features (e.g. "a crustacean"). Following Hills et al. (2009b), we analyze only features of the first and second kind, as these are the features likely to be available to two-year-old children. We create two different networks for each child: one in which connectivity is defined by *perceptual* feature overlap, and one in which connectivity is defined by *functional* feature overlap. This is because overlapping perceptual features indicate a very different kind of relatedness than overlapping conceptual features.

Hills et al. (2009b) analyzed the clusters produced by each of these kinds of networks to quantify these different kinds of relatedness. Defining connectivity by *perceptual* feature overlap produced networks that were dense, highly connected, put words into more than one category, and produced categories that were overly inclusive relative to human judgments (e.g. MCDI categories, Fenson, et al., 1994). In contrast, *functional* feature overlap produced networks that were sparser, had smaller, better defined categories, and were better at discriminating among near-category members. In general, words connected in the *functional* network are more likely to be encountered in a relational context, facilitating learning by mutual exclusivity (e.g. *cake-carrots*, *boots-coat*). In contrast, words connected in the *perceptual* networks are less likely to be encountered in such situations, and learning one is thus less likely to facilitate learning the other through mutual exclusivity (e.g. *sheep-sofa*, *pencil-stick*). Thus, we can test a specific prediction about how mutual exclusivity builds vocabulary structure: it facilitates the acquisition of *functionally* related words.

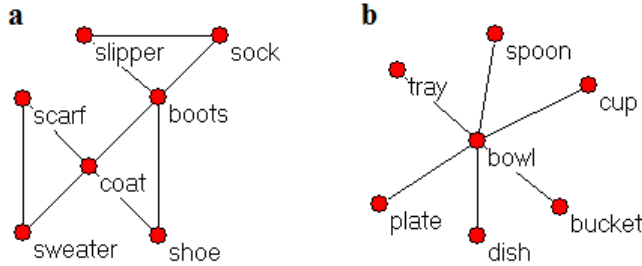


Figure 2: Two 7-vertex networks with different connectivity structures. The left network has a high clustering coefficient and a low degree centrality. The right network has a low clustering coefficient, but a high degree centrality.

In addition to using these two kinds of features to define connectivity, we measure their resulting structure in two different ways. These different connectivity measurements represent different ways in which mutual exclusivity could build structure. Consider the networks in Figure 2.

The first network (2a) has many local clusters, triangles in which any vertices with a common neighbor are likely to be neighbors themselves. One might predict mutual exclusivity to facilitate this kind of structure because using one word (e.g. *scarf*) to learn a semantic neighbor (e.g. *sweater*) should make a common neighbor even easier to learn (e.g. *coat*). This structure is measured by *clustering coefficient* (Equation 1), which has previously been used to distinguish the vocabulary structures of early and late talkers (Beckage, Hills, & Smith, 2011).

$$C = \frac{1}{|V|} \sum_{i=1}^{|V|} \frac{2|\{e_{jk}\}|}{d(v_i)(d(v_i) - 1)} \quad v_j, v_k \in N_i, e_{jk} \in E \quad (1)$$

In contrast, the second network (2b) does not have any local clusters, but rather has a single highly salient *hub*: a single vertex with many neighbors. This kind of structure might be even more likely to arise through mutual exclusivity, as learning the *hub* word (*bowl*) makes each of its neighbors easier to learn (*spoon*, *tray*, *cup*). This kind of structure is measured by *degree centrality* (Equation 2). This measure is new to semantic network analyses, but is a mainstay of social network science (Freeman, 1979), and measures the structural property intuitively most likely to be related to learning words through exclusion.

$$D = \frac{\sum_{i=1}^{|V|} [d(v^*) - d(v_i)]}{(|V| - 1)(|V| - 2)} \quad (2)$$

Thus, we test two hypotheses in the following analysis: mutual exclusivity should predict connectivity structure in *functional* but not the *perceptual* networks, and it should manifest more strongly in high *degree centrality* than in *clustering coefficient*.

## Method

To construct semantic networks for each child, we used all words which are both measured by the MCDI, and for which McRae and colleagues collected feature norms. This resulted in a list of 130 nouns, encompassing animals, food, clothing, vehicles, etc. For a full list, see Hills et al. (2009b). Each child's semantic network was constructed by adding one vertex for each word on that child's productive MCDI. Vertices were connected if they shared a minimum number ( $w$ ) of semantic features. To be consistent with Hills et al. (2009b), we set this features threshold to all possible values 1-4. At  $w = 3$ , for instance, two words were connected only if they shared three or more semantic features. However, networks become increasingly sparse as  $w$  increases, and we thus urge caution in interpreting results at high thresholds.

Two networks were created for each child, one network in which only *perceptual* features defined connectivity, and one network in which only *functional* features were used to define connectivity (see above). Networks were defined by their set of vertices  $V$  and the set of edges  $E$  that connected them. A vertex's degree ( $d(v)$ ) is defined as the number of other vertices to which it is connected by an edge. These connected vertices are called neighbors, and together define a node's neighborhood ( $N$ ).

Once each network was constructed, two properties of its connectivity structure were measured. The first, clustering coefficient ( $C$ ), measures the proportion of vertices with a common neighbor that are also neighbors of each other. (Equation 1). The second, degree centrality ( $D$ ), measures the proportion of edges connected to a single dominant *hub* vertex (Equation 2). These measures of structure trade off, with high degree centrality necessitating a low clustering coefficient. Both measures always range between 0 and 1, and thus are independent of the size of a child's vocabulary. They are measures of structure independent of size.

## Results and Discussion

As in the analysis above, children were divided into two groups: Mutual Exclusivity Users who performed better than chance on the novel trials of the disambiguation task, and Nonusers who did not. Again, we reiterate that this is not a theoretical commitment, but rather a tool for noise reduction. The structure of each child's individual semantic networks – both perceptual and functional – was used to predict that child's category of mutual exclusivity usage.

Before presenting the results of network analyses, we recapitulate that vocabulary *sizes* were quite comparable between these groups. The 17 ME Users produced an average of 408.3 words on the MCDI while the 17 Nonusers produced an average of 388.1 ( $t(32) = .37$ , *n.s.*). They also did not differ in the number of words they knew from the set of 130 used in the network analysis ( $M_u = 92.6$ ,  $M_n = 88.1$ ,  $t(32) = .51$ , *n.s.*). However, the particular words they knew, and the semantic relationships among them, proved to be importantly different.



Figure 3 shows correlations between measures of network structure and the mutual exclusivity category to which each individual child belonged. For perceptual networks, constructed by connecting words by shared perceptual features (e.g. “has fur,” “tastes sweet”), neither clustering coefficient nor degree centrality were related to use of mutual exclusivity at any feature overlap threshold (Figure 3, left column). As predicted, perceptual networks, in which connections are not a good proxy for words likely to occur in contrastive contexts, have structures not well predicted by use of mutual exclusivity.

In contrast, for functional networks, those constructed by connecting words by shared relational, functional features, mutual exclusivity was a significant predictor of degree centrality when 2 or more overlapping features defined a connecting edge ( $w = 2$ ). At this threshold, children who used mutual exclusivity at above-chance levels had semantic networks with higher degree centrality ( $\rho = .34, p < .05$ ; Figure 3, bottom right). This same threshold is shown by Hills et al. (2009a) to best separate semantic categories in this set of words. Use of mutual exclusivity did not reliably predict clustering coefficient, but did show a positive trend, particularly at overlap threshold 3 ( $\rho = .28, p = .1$ ; Figure 3, top right). This trend should be interpreted cautiously, however, as conceptual networks were quite sparse at  $w = 3$ , having at most 12 edges.

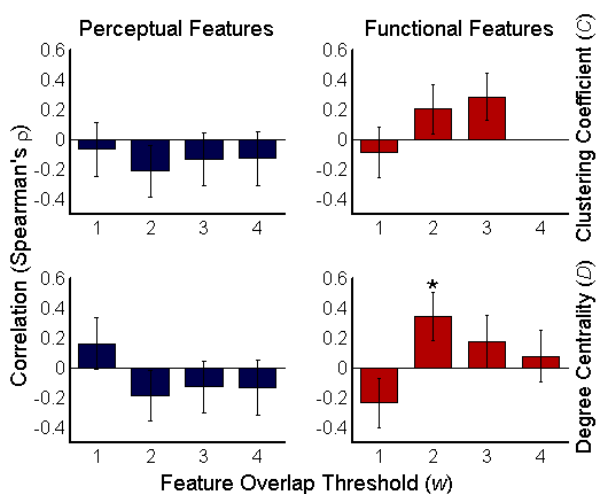


Figure 3: Correlation between network structure and mutual exclusivity performance. The top row shows correlations clustering coefficient, the bottom row for degree centrality. The left column shows perceptual features, the right shows functional. Individual bars show correlations when particular thresholds ( $w$ ) define the minimum number of overlapping semantic features required to connect two words. Error bars show  $\pm 1$  standard error as measured by 1000 samples of bootstrap resampling (Lunnenborg, 1985). Mutual exclusivity was significantly correlated with degree centrality in the functional networks and showed a non-significant trend towards correlation with correlation coefficient in these same networks. As predicted, ME was uncorrelated with perceptual network structure.

Thus, the semantic network structures of children who reliably exhibit mutual exclusivity are predictably different from those of children who do not. Even though these children know the same number of words, the words they know are different. Semantic networks of ME users are characterized by more hub-like structure, a consequence of the kind of word learning facilitated by exclusion. Importantly, these differences are likely to matter (Beckage, Smith, & Hills, 2011). Differences in connectivity structure lead to differences in network robustness, with the networks of mutual exclusivity structure perhaps protecting them against forgetting and aiding future learning (Albert, Jeong, Barabási). These results represent a first step in understanding how children’s own learning mechanisms build the structure of their semantic networks.

## General Discussion

While the words that children learn are, of course, a function of the linguistic input to which they are exposed (Brent & Siskind, 2001; Hills et al., 2009a), this link is likely to be moderated by children’s own attentional and learning mechanisms (Hudson Kam & Newport, 2005; Smith, 2000). For instance, children learn to extend newly-learned object words to categories on the basis of particular feature dimensions. Normatively, children learn a bias to attend to shape, and this bias leads them to learn more categories organized by shape (Smith et al., 2002). However, children may learn a different bias, and consequently learn different words in the future (Colunga & Sims, 2011). We show that use of mutual exclusivity may play a similar role. Children who robustly use mutual exclusivity are likely to learn new words functionally related to words they already know. As atypical semantic network structure is related to slower language learning (Beckage, Smith, & Hills, 2011), these results point to a potential intervention for late-talking children. Learning to disambiguate the meanings of new words through exclusion could help late-talkers to catch up.

These results also lead to two further insights about mutual exclusivity and its role in vocabulary development. While mutual exclusivity is often thought to be critical to early word learning, its relationship to vocabulary size is unclear. For every study that finds a significant correlation between mutual exclusivity and vocabulary size (e.g. de Marchena, Eigsti, Worek, Ono, & Snedeker, 2011; Mervis and Bertrand, 1994), another finds no correlation between the two (e.g. Halberda, 2003; Mather & Plunkett, 2009). These results help to shed light on this inconsistency by pointing out that the relationship between vocabulary development and mutual exclusivity may be found not in size but in structure. While we do not mean to argue that mutual exclusivity is required for rapid word learning, we do suggest that their relationship can be better understood by considering semantic network structure.

Finally, these results may shed light on the origins of mutual exclusivity itself. Thus far, we have argued that mutual exclusivity builds vocabulary structure. But

vocabulary structure may also build mutual exclusivity. One can think of mutual-exclusivity as an overhypothesis, a probabilistic rule about the general structure of word-object mappings derived from the structure of individual word-object mappings (Kemp, Perfors, & Tenenbaum, 2007; Mervis & Bertrand, 1994). For instance, mutual exclusivity may have its roots in an understanding that labels are often contrastive, pointing to differences between otherwise similar objects. If this is true, vocabularies that make this overhypothesis more probable should lead to stronger mutual exclusivity biases. Thus, one can think of the hub-like structures characteristic of ME users in our sample as not only arising from mutual exclusivity, but helping to construct it as well. Hub words, which are connected to many semantically-related neighbors, may play an important role in discovery of this higher-order regularity. Thus, mutual exclusivity may operate much like the shape bias: being both built from regularities in the structure of linguistic input, and helping children to discover further regularities (Smith, et al., 2002). A deep understanding of the connection between mutual exclusivity and vocabulary structure, then, will come from understanding a three part relationship: how ME contributes to structure, how structure contributes to ME, and language input contribute to both.

### Acknowledgments

We are grateful to Cody Stitzel for her help coding the MCDI data, and to all of the members of the Smith, Yu, and Shiffrin labs whose feedback helped to improve this work. This work was supported by a NSF GRF to DY.

### References

- Albert, R., Jeong, R., & Barabási, A. (2000). Error and attack tolerance of complex networks. *Nature*, 406, 378-382.
- Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, 22, 261-295.
- Arias-Trejo, N., & Plunkett, K. (2009). Lexical-semantic priming effects during infancy. *Philosophical Transactions of the Royal Society B*, 364, 3633-3647.
- Beckage, N., Smith, L., & Hills, T. (2011). Small Worlds and Semantic Network Growth in Typical and Late Talkers. *PloS ONE*, 6, e19348.
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407-428.
- Colunga, E., & Sims, C. E. (2011). Early talkers and late talkers know nouns that license different word learning biases. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2550-2555). Austin, TX: Cognitive Science Society.
- de Marchena, A., Eigsti, I. M., Worek, A., Ono, K., & Snedeker, J. (2011). Mutual exclusivity in autism spectrum disorders: Testing the pragmatic hypothesis. *Cognition*, 119, 96-113.
- Diesendruck, G., & Markson, L. (2001). Avoidance of lexical overlap: A pragmatic account. *Developmental Psychology*, 37, 630-641.
- Ebbinghaus, H. (1885/1962). *Memory: A contribution to experimental psychology*. New York: Dover.
- Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S.J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59. Chicago: University of Chicago Press.
- Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1, 215-239.
- Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L., M. & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology*, 28, 99-108.
- Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, 87, B23-B34.
- Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. B. (2009a). Categorical structure among shared features in networks of early-learned nouns. *Cognition*, 112, 381-396.
- Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. B. (2009b). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition?. *Psychological Science*, 20, 729-739.
- Hollingshead, A. B. (1975). *Four factor index of social status*. Unpublished manuscript, Department of Sociology, Yale University, New Haven, CT.
- Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1, 151-195.
- Lunnenborg, C. E. (1985). Estimating the correlation coefficient: The bootstrap approach. *Psychological Bulletin*, 98, 209-215.
- Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10, 307-321.
- Markman, E., & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20, 121-157.
- Mather, E., & Plunkett, K. (2009). Learning words over time: The role of stimulus repetition in mutual exclusivity. *Infancy*, 14, 60-76.
- McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37, 547-559.
- Mervis, C. B., & Bertrand, J. (1994). Acquisition of the novel name-nameless category (N3C) principle. *Child Development*, 65, 1646-1662.
- Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, 88, 92-134.
- Smith, L. B. (2000). Learning to learn words: An associative crane. In R. M. Golinkoff & K. Hirsh-Pasek (Eds.), *Becoming a word learner: A debate on lexical acquisition* (pp. 51-80). New York: Oxford University Press.
- Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object naming provides on-the-job training for attention. *Psychological Science*, 13, 13-19.
- Steyvers, M., & Tenenbaum, J. B. (2005). The large scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29, 41-78.



# Inferring Covert Events in Logical Metonymies: a Probe Recognition Experiment

Alessandra Zarcone (alessandra.zarcone@ims.uni-stuttgart.de)

Institut für Maschinelle Sprachverarbeitung, Azenbergstr. 12  
70174 Stuttgart, Germany

Sebastian Padó (pado@cl.uni-heidelberg.de)

Institut für Computerlinguistik, Universität Heidelberg, Im Neuenheimer Feld 325  
69120 Heidelberg, Germany

Alessandro Lenci (alessandro.lenci@ling.unipi.it)

Dipartimento di Linguistica “T. Bolelli”, Via Santa Maria 36  
56126 Pisa, Italy

## Abstract

It has been widely acknowledged that the interpretation of logical metonymies involves the interpretation of covert events (*begin the book* → *reading / writing*). Whether this implicit content is part of our lexicon or rather derives from general pragmatic inference, it is currently subject of debate. We present results from a probe recognition experiment, providing novel evidence in support of early metonymy processing, consistent with the hypothesis that covert events are retrieved from knowledge of typical events activated by lexical items.

**Keywords:** Logical metonymy; generalized event knowledge; qualia structure; covert events; probe recognition.

## Logical metonymy:

### lexicon, world knowledge, typical events

In logical metonymies, an event-subcategorizing verb is combined with an entity-denoting patient. Contrast the German non-metonymic “long variant” (Ex. 1) with the metonymic “short variant” (Ex. 2), which is understood to have the same meaning:

1. Peter begann das Bier **zu trinken**. (“long variant”)  
Peter began the beer **to drink**.  
Peter began **drinking** the beer.
2. Peter begann das Bier. (“short variant”)  
Peter began the beer. → drinking the beer

In (2), the event-subcategorizing verb *beginnen* (begin) combines with an entity, *das Bier* (the beer), but the clash is resolved and the interpretation constructed by the recovery of a covert event which mediates between matrix verb and object (e.g. *begin* → *drinking the beer*).<sup>1</sup> The reconstruction of the covert event has well-known behavioral correlates – processing metonymic sentences is more costly than processing non-metonymic ones (Pykkänen & McElree, 2006; Baggio, Chroma, Lambalgen, & Hagoort, 2010).

On the level of theory, logical metonymies pose a challenge to compositionality (Partee, ter Meulen, & Wall, 1993; Baggio, Lambalgen, & Hagoort, 2012) and therefore touch on a foundational principle to language research. One of the points

<sup>1</sup>This presumed process, namely the coercion of the object into an event associated with it, explains the use of the term “metonymy”.

of debate is where covert events are retrieved from. The two main accounts of logical metonymy have suggested that covert events are part of our lexical knowledge (“lexical hypothesis”) or that they are retrieved through post-lexical inferences triggered by our general world knowledge and communication principles (“pragmatic hypothesis”).

We present experimental evidence in support of a third hypothesis, namely that covert events are retrieved from knowledge of typical events stored in our long-term memory. A previous self-paced reading study (Zarcone & Padó, 2011) had presented evidence for generalized event knowledge integration in logical metonymy, but its results were relying on a strong methodological assumption, namely that the same cognitive processes are used to interpret “long variants” (Ex. 1) and “short variants” (Ex. 2). The present study uses a probe recognition paradigm in order to avoid this assumption. Additionally, the new paradigm allows us compare reaction times for low and high inter-stimulus intervals to assess the time course of metonymy interpretation in more detail. We find evidence for early, expectation-driven processing for metonymy as opposed to later, strategy expectancy generation, which points towards a central role of generalized event knowledge in logical metonymy interpretation.

## Covert events from lexical knowledge

Pustejovsky (1995) and Jackendoff (1997) provide an account of logical metonymy which we call the “lexical hypothesis”: Logical metonymy is a type mismatch between the (semantic) subcategorization of a metonymic verb for an event and the entity denoted by the object on the other side, which requires the integration of an event to be resolved. The event is retrieved from complex lexical entries (qualia structures) associated with the object in the mental lexicon. In particular, the “agentive quale” (the event that brings about the object) and the “telic quale” (the main purpose of the object) are the relevant components of the qualia structure which can be retrieved as covert events in metonymic contexts.

Being part of the mental lexicon, qualia model linguistic knowledge – in opposition to world knowledge and pragmatic inferences. Psycholinguistic work has identified experimental correlates for the lexical hypothesis (see Pykkänen and McEl-

ree (2006) for a review), however processing studies have focused predominantly on the type mismatch and have largely ignored the question of what covert events are accessible to metonymic interpretation and what their nature is (lexical vs. pragmatic).

The lexical hypothesis provides an economical solution to the problem of covert event retrieval, and it seems very plausible that we associate typical events with lexical items referring to entities in our mental lexicon. However, this solution seems to undergenerate the range of potential interpretations for logical metonymy. Consider the following examples:

3. John is a famous wrestler. He really enjoys a good fight.  
(→ fighting)
4. John is a wrestling fan. He really enjoys a good fight.  
(→ watching)

These examples show a few shortcomings which seem to be associated with this account: first, *fight* is an event-denoting item and, since metonymy is typically claimed to arise from a type clash, we wouldn't expect an event reconstruction here, nevertheless *watching* seems to be the normal scenario involving wrestling fans and wrestling fights; secondly, qualia structures are only defined for artifacts, but not for event-denoting items such as *fight*; lastly, a mechanism to select one or another event depending on the agent involved (e.g. *wrestler* vs. *wrestling fan*) is not specified. Facts like this are well known at least since Lascarides and Copestake (1998), who claim that qualia structure alone (as defined in Pustejovsky (1995)) is not enough to explain the range of reconstructed events in metonymic sentences, which instead derive from the contribution of wider contextual information.

### Covert events from pragmatic inferences

An alternative approach (the "pragmatic hypothesis") has argued that metonymy resolution is driven by dynamic inferences based on context and world knowledge rather than static lexicon entries. This model assumes that lexico-conceptual representations are atomistic (Fodor & Lepore, 1998) and that metonymic event reconstruction derives from post-lexical pragmatic inferences (Cartson, 2002; R. D. de Almeida & Dwivedi, 2008; R. G. de Almeida et al., 2009). Without resorting to lexical atomism, Asher (2010) similarly models logical metonymy in terms of general discourse principles for presupposition accommodation. This approach has the advantage of placing logical metonymy within a broader picture of inference-driven processes, and it accounts for the most problematic cases for strict lexicalist models. However, it currently lacks a concrete characterization of the type and organization of knowledge involved in metonymy interpretation.

### Covert events from generalized event knowledge

Recent work on *generalized event knowledge* (McRae & Matsuki, 2009) has shown that inferential world knowledge about typical scenarios plays an early and crucial role in sentence

comprehension processes. There is solid experimental evidence that language understanding makes extensive use of global plausibility information and event knowledge (e.g. Altmann and Kamide (1999)). McRae and Matsuki (2009) established that speakers make use of prototypical knowledge about events (generalized event knowledge) when rapidly building expectations about upcoming input. Generalized event knowledge is assumed to be built from first and second-hand experience: for instance, we learn that a scenario of *washing hair* typically includes a shampoo, a sink, a bathroom, and happens indoor; a scenario of *washing car* would include different elements (an outdoor environment, a hose). Such scenarios are available in our memory and can be cued by linguistic input, e.g. "action verbs as well as nouns referring to agents, patients, instruments, locations, and events" (McRae & Matsuki, 2009). Generalized event knowledge can be also thought as default information associated with lexical items, according to the proposal in Lascarides and Copestake (1998).

Generalized event knowledge can provide the basis for a third account of covert event recovery in logical metonymy, suggesting that covert events are retrieved from knowledge of typical events stored in our long-term memory. Similarly to the lexical hypothesis, this third hypothesis links objects to associated events. The difference between the accounts is that the lexical hypothesis associates each noun with a fixed set of events, whereas the picture for generalized event knowledge is more flexible. In many cases, our experience will comprise events that do not correspond to classical qualia – in the case of cars, we know that cars need to be filled up, that they need to be washed, and that many people lease their cars. Word meaning is tied up to this sort of scenario knowledge, which typically contains the qualia as a proper subset. This is not always true, however – people will only have rich representations for events and objects that they are familiar with. For example, Matsuki et al. (2011) found that U.S. college undergraduates were not familiar with the event of "dusting off" and failed to build expectations about plausible scenarios. Also, some events which are part of our generalized event knowledge for a given entity (a few examples from our experimental materials: *pizza - deliver, fix - car, peel - apple*) do not fit well with the traditional qualia roles.

It follows that the picture painted by generalized event knowledge is also considerably different regarding the status of this knowledge: There is no distinction between linguistic and world knowledge. Resorting to world knowledge is an element of similarity with the pragmatic hypothesis, from which our proposal however differs with regards to an important aspect. The "pragmatic hypothesis" assumes that metonymic event reconstruction is carried out by general communication or discourse based devices, like many other types of pragmatic inferences; contrastively, generalized event knowledge is activated immediately during sentence processing. Consequently, the metonymic event reconstructions can be based on exactly the same type of knowledge responsible for generating predictions during on-line language comprehension and for the

Table 1: Example materials for the self-paced reading study (Zarcone & Padó, 2011) and the present probe recognition study.

	Self-paced reading	Probe recognition	
		Sentence	Probe
high-typicality agent	Der <i>Konditor</i> begann, die <i>Glasuren</i> <b>aufzutragen</b> . The baker started the icing <b>to spread</b> .	Der <i>Konditor</i> begann mit der <i>Glasuren</i> . The baker started with the icing.	<b>AUFTRAGEN</b>
low-typicality agent	Das <i>Kind</i> begann, die <i>Glasuren</i> <b>aufzutragen</b> . The child started the icing <b>to spread</b> .	Das <i>Kind</i> begann mit der <i>Glasuren</i> . The child started with the icing.	<b>AUFTRAGEN</b>
high-typicality agent	Das <i>Kind</i> begann, die <i>Glasuren</i> <b>zu essen</b> . The child started the icing <b>to eat</b> .	Das <i>Kind</i> begann mit der <i>Glasuren</i> . The child started with the icing.	<b>ESSEN</b>
low-typicality agent	Der <i>Konditor</i> begann, die <i>Glasuren</i> <b>zu essen</b> . The baker started the icing <b>to eat</b> .	Der <i>Konditor</i> begann mit der <i>Glasuren</i> . The baker started with the icing.	<b>ESSEN</b>

predicate-argument thematic fit (McRae & Matsuki, 2009). We regard the account proposed here as a generalization of purely lexicalist approaches, which is able to provide a more dynamic model of covert event interpretation triggered by event knowledge associated with lexical items, while keeping it distinct from other, genuinely pragmatic processes.

### Evidence from self-paced reading

While type clashes and type shifting have received great attention in the experimental literature on logical metonymy, there is comparatively little work on the source of covert events (lexical vs. inferential). Offline work (Lapata, Keller, & Scheepers, 2003; Zarcone & Padó, 2010) has established that the range of interpretations for metonymy is larger than predicted from qualia structure. In an important online study, Frisson and McElree (2008) “assume that coerced senses are computed from a broader range of properties than the Qualia structure of the complement noun”. In order to exclude the possibility that increased processing costs in logical metonymies might be determined by competition between different possible interpretations, they carried out an eye-tracking experiment, contrasting (a) sentences like *The teenager began the novel*, where one interpretation is strongly preferred (*reading*), (b) sentences like *The waitress started the coffee*, where multiple interpretations are plausible (*drinking*, *preparing*) and possibly competing, and (c) base forms of (a) and (b) like *The teenager read the novel*. However, the authors do not commit to a specific hypothesis regarding the range of covert events.

Zarcone and Padó (2011) have suggested *generalized event knowledge* as an alternative source of interpretation, providing results from a self-paced reading experiment. The study capitalized on the verb-final word order in German subordinate phrases, by likening the recovery of a covert event in a “short variant” metonymy (Ex. 2) to the process of building expectations about the sentence-final event in its “long variant” (Ex. 1) and by analyzing reaction times for long variants, where the event was explicit.

The self-paced experiment contrasted a high-typicality agent condition with a low-typicality agent condition (*Der Konditor / das Kind hörte auf, die Glasuren aufzutragen*. - *The baker / the child finished spreading the icing*). In the high-typicality condition the target event was cued by the preceding agent-patient pair (*baker-icing*), creating rich expectation on the upcoming event. The experiment found that these expecta-

tions in fact yield a facilitation effect and shorter reading times for the target verb compared to the low-typicality condition.

Also, a crucial difference with Frisson and McElree (2008) is that in this study the same patient noun is used in different context conditions, in order to restrict variability due to item idiosyncrasies. Thus, the experiment in Zarcone and Padó (2011) provided evidence towards a generalized event knowledge account of logical metonymy. However, it was heavily based on the assumption that the same cognitive resources are involved when recovering covert events in logical metonymies and when predicting sentence-final sentences (Lapata et al., 2003). This assumption is of course debatable.

### Experiment

The goal of the present study is to strengthen the case for generalized event knowledge in logical metonymy by avoiding any assumptions about the relative processing of “short variant” and “long variant” sentences. We employ a different experimental paradigm, namely probe recognition, and concentrate on unequivocally metonymical experimental materials. More specifically, we build “short variant” sentences from the items used in the self-paced reading study, contrasting two typicality conditions as before. The main difference is in the cued event, which is now not part of the sentence but is presented as a probe after the metonymical sentence. The old and new materials are contrasted in Table 1.

Crucially, in the probe recognition study the probe is not part of the test sentence: to enhance this and avoid sentence completion effects (i.e. that participants would be influenced by verb-final word order in German subordinate phrases and perceive the probe as part of the sentence), the *zu* particle was omitted. Since the cued event is never part of the sentence, participants are required to answer “no” for all probes. To balance the responses, the material sentences are complemented by fillers for which the answer is “yes” (see below for details).

Our expectation is that in the high-typicality condition (*baker-icing-spread*), the covert event *spread* is cued by the agent-patient combination, causing participants to require longer decision latencies to recognize that the verb was not part of the sentence than in the low-typicality condition (*child-icing-spread*), where the cued event is a different one (*eat*).

A second element of novelty is the addition of a second factor: inter-stimulus interval (ISI). Contrasting a short and a long ISI can provide insights about the nature of covert event

Table 2: Triplets for *Glasur (icing)*.

	Agent	Patient	Event
high-typicality triplet	Konditor	Glasur	auftragen
	baker	icing	spread
	Kind	Glasur	essen
low-typicality triplet	child	icing	eat
	Kind	Glasur	auftragen
	Konditor	icing	spread
	baker	Glasur	essen
		icing	eat

retrieval: at longer ISI, an expectancy strategy is employed (strategic expectancy generation) and participants tend to use the input received to generate a potential set of upcoming targets whose processing is facilitated (Becker, 1980; Ferretti, McRae, & Hatherell, 2001; Van Der Meer, Krüger, & Nuthmann, 2005); an effect at a short ISI would suggest that covert events are available online and early on during processing.

### Creation of Materials

As mentioned above, we adapted the German materials from the self-paced reading experiment in Zarcone and Padó (2011) for the probe recognition task. In the earlier study, elicitation tasks were used to tap into generalized event knowledge scenarios when preparing materials for the self-paced reading experiment, according to an established method in research on generalized event knowledge (see also McRae, Hare, Elman, and Ferretti (2005)).

**Event elicitation** We elicited typical events for a set of 50 patients in German, by asking 20 participants in a web experiment to generate verbs in response to typical patients (“list the things that these objects have done to them”). Space for 10 responses per item was provided and no time limit was imposed. For each item, we chose four events from those named early by many participants (i.e., those with highest mean reciprocal rank measure), ensuring that the four events referred to different scenarios. Then we paired each patient to the infinite form of its four selected verbs (200 patient-event pairs): e.g., the four events selected for *Auto (car)* were *fahren* (drive), *reparieren* (fix), *verkaufen* (sell), *waschen* (wash).

**Agent elicitation** We elicited typical agents for the resulting 200 patient-event pairs from 10 participants in a web experiment (“list who typically performs these actions”). For each item, space was provided for 10 responses; no time limit was imposed. From the initial list of 200 patient-event pairs, we extracted 24 patients paired with 2 events each, and per each patient-event combination we selected one of the best agents (those named early by many participants, i.e. with highest mean reciprocal rank measure), obtaining 48 agent-event-patient high-typicality triplets.<sup>2</sup> 48 low-typicality triplets were obtained by crossing agents between the two events in the

high-typicality triplets, for a total of 96 agent-event-patient triplets (48 high-typicality, 48 low-typicality). Table 2 shows examples for high- and low-typicality triplets.

**Test sentences** The aim of the probe recognition task was to replicate the results from the self-paced reading tasks with strictly metonymical sentences. For this reason, we only used “short version” main clause sentences (Ex. 2) constructed with metonymical verbs, as shown on the right-hand side in Table 1.

### Method

**Participants** Thirty-six students of Universität Stuttgart volunteered to participate in the experiment and were paid for their participation. All participants were native speakers of German and had normal or corrected-to-normal vision.

**Procedure** On each trial, a sentence appeared in the middle of the screen and the participants were to press a key after reading it. Pressing the key elicited the presentation of the probe word with a short (100 ms) or long (900 ms) inter-stimulus interval (ISI). The participants were instructed to decide as quickly and accurately as possible whether or not the probe appeared in the sentence by pressing the green or the red key respectively. The green key was the left key for left-handed participants and the right key for right-handed participants, so that the “no” answers were always given with the non-dominant hand.

Responses and decision latencies for each probe were recorded. The experimental session lasted approximately 30 minutes; participants were allowed to take two breaks during the experiment, one after the first third of sentences and one after the second third.

Each participant saw all 48 sentence-probe combinations, half of them in the high-typicality condition and the other half in the low-typicality condition. The experimental items were intermixed with 72 filler items, which were the same for both lists. Since the 48 test probes in each list were never in the sentence (i.e. the answer was “no”), 60 of the fillers did include the probe in the sentence and 12 did not, for a total of 60 “yes” answers and 60 “no” answers in each list.

**Design** The study employed a 2x2 mixed factorial design. One factor, inter-stimulus interval (ISI: long / short), was varied between subjects, the other factor, typicality (TYP: high / low), was varied within subjects.

### Results

All participants scored better than 95% correct on the probe recognition task. Data points corresponding to the wrong answers and outliers ( $> 2.5$  SD from the mean) were excluded from the analysis (2% of the data points). The mean of reading times on the sentences preceding the probes was 2629 ms (SD 1280). At both short and long ISI, mean decision latencies were longer for the high typicality condition (Table 4).

We examined the effect of ISI and typicality on decision latencies through 2 x 2 by-subject ( $F_1$ ) and 2 x 2 by-item ( $F_2$ ) analyses of variance, which yielded main effects of

<sup>2</sup>An additional sensibility verification test was run, in order to check that low-typicality triplets were, although not typical, still sensible (i.e., did not violate any selectional restriction). See Zarcone and Padó (2011) for more details on the pre-tests).

Table 3: Fixed effects for the mixed-effect model:  $\log(dl) \sim ISI * TYPICALITY + dlPrecProbe + rtSent + order + (1|subject) + (1|item)$

	Estimate	MCMCmean	HPD95lower	HPD95upper	pMCMC	$Pr(>  t )$
(Intercept)	7.4848	7.4766	7.3939	7.5584	0.0001	0.0000
ISIshort	-0.5490	-0.5452	-0.6342	-0.4624	0.0001	0.0000
TYPICALITYlow	-0.0108	-0.0106	-0.0351	0.0155	0.4192	0.4062
dlPrecProbe	0.0000	0.0000	0.0000	0.0001	0.0282	0.0506
rtSent	0.0000	0.0000	0.0000	0.0000	0.0952	0.0879
order	-0.0039	-0.0039	-0.0046	-0.0031	0.0001	0.0000
ISIshort:TYPICALITYlow	-0.0360	-0.0366	-0.0735	-0.0019	0.0472	<b>0.0505</b>

Table 4: Mean decision latencies (dl, measured in ms).

	low typicality		high typicality		
ISI	Mean dl	SD	Mean dl	SD	Mean diff.
short	969	296	1026	363	+57
long	1735	351	1746	362	+12

ISI ( $F_1(1, 35) = 111.03, p < .001; F_2(1, 47) = 2553, p < .001$ ) and of typicality ( $F_1(1, 35) = 7.7616, p = .009; F_2(1, 47) = 6.02, p = .015$ ). The difference in decision latencies between low and high typicality is larger for the short ISI condition, but the interaction fails to reach significance.

Mixed-effect models have been shown to be more powerful for reading studies, because they allow on the one hand for separating random effects of item and participant, and on the other hand for taking into account trial-to-trial longitudinal dependencies between individual observations, by including covariates such as response latencies at preceding trials. Following the procedure suggested by Baayen, Davidson, and Bates (2008); Baayen and Milin (2010), we performed a mixed-effect analysis using as covariates the order of presentation (rank-order of a trial in its experimental sequence), the reading times at the sentence preceding the probe and the decision latencies at preceding probe. The mixed-effect analysis (Table 3) shows a number of significant effects. Most important for our current purpose is the marginally significant interaction of ISI and typicality (shown in boldface), which indicates that the effect of typicality indeed diminishes for longer inter-stimulus intervals.

## Discussion

In our experiment, high typicality agents in logical metonymies cue covert events that are coherent with the generalized event knowledge scenario associated with those agents and with the given patient. Cued events were integrated in the sentence meaning and were therefore more difficult to reject as probes, leading to significantly longer decision latencies for the high typicality condition than for the low typicality condition. This supports the hypothesis that covert events can be predicted by generalized event knowledge scenarios, which constitute a broader range of interpretation than traditional qualia-based accounts, which are typically a subset of them. Elicitation tasks were used to retrieve two generalized event knowledge scenarios per item, and in some cases these did

not correspond to traditional qualia roles: the events used for *pizza* were *baking* and *delivering*, the events for *apple* were *picking* and *peeling*, the events for *car* were *driving* and *fixing*.

The emergence of the effect at long ISI could indeed be explained as a result of an expectancy strategy (that is, pragmatic inferences driven by the perceived need to perform the task), but the effect at short ISI provides evidence for early integration of generalized event knowledge, suggesting that covert event interpretation emerges early on in processing. Firm conclusions on this point require further investigation, but there is clear parallelism to short SOA effects of generalized event knowledge typicality in Ferretti et al. (2001).

The early emergence of the typicality effect in Zarcone and Padó (2011) and in the current experiment is in contrast with the "inference hypothesis" (if by inference we mean a post-lexical late-onset process), in that covert events emerge early on in processing. Promising work on ERPs (Baggio et al., 2010; Schumacher & Weiland, 2011) can therefore be of crucial importance in the near future to shed new light into more fine-grained processes involved in covert event resolution.

In the current experiment, we were able to replicate the activation effect for events which can be plausibly assumed to be present in generalized event knowledge observed in Zarcone and Padó (2011). By using a different experimental paradigm, we avoided the need to compare "long" and "short" variants. The parallel outcome of the two experiments provides evidence for a picture of language processing in which metonymic event reconstruction is carried out by the same resource – i.e., general event knowledge – which is normally employed for predicate-argument integration during on-line comprehension of non-coercion sentences.

To test this hypothesis, we are currently running a further experiment with non-metonymic verbs and the same probe recognition design used in this study. There exists indeed a long standing tradition of experiments contrasting coercion conditions and control conditions which do not involve coercion (Traxler, Pickering, & McElree, 2002; R. D. de Almeida & Dwivedi, 2008; Frisson & McElree, 2008), and in such studies the question of coercion plays a central role. We have sidestepped this issue in our experiments so far, but our next step will directly address this question by contrasting the same sentence-probe combinations illustrated above with sentences containing a non-coercion predicate (e.g. *The man rolled a cigarette*), such that the probe (e.g., *smoke*) expresses an event

part of the general knowledge activated by the sentence (e.g., smoking is the purpose of rolling a cigarette). Our model predicts no significant differences in decision latencies between the non-coercion sentences of the new experiment and the coercion sentences of the present study, given that the same general event knowledge is responsible for on-line expectation activation during sentence comprehension. In other terms, we expect that both with *The man rolled a cigarette* and with *The man enjoyed a cigarette*, general event knowledge activates the cued event of smoking, irrespective of the presence or not of a coercive predicate.

## Conclusion

We have provided evidence in support of the hypothesis that generalized event knowledge can predict covert event interpretation. Also, the current study constitutes a further step towards a characterization of the phenomenon of logical metonymy within the broader frame of early online integration of typical event knowledge in comprehension.

**Acknowledgments.** We acknowledge the support of Deutsche Forschungsgemeinschaft (DFG) for the project D6 (SFB 732) at the University of Stuttgart. We thank Valentina Bambini, Berry Claus, Nick Gaylord and Ken McRae for helpful discussions and our anonymous reviewers for useful contributions to improve the analysis.

## References

- Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73, 247–264.
- Asher, N. (2010). *Lexical meaning in context*. Cambridge University Press.
- Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.
- Baayen, R., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3, 12–28.
- Baggio, G., Chroma, T., Lambalgen, M. van, & Hagoort, P. (2010). Coercion and compositionality. *Journal of Cognitive Neuroscience*, 22(9), 2131–2140.
- Baggio, G., Lambalgen, M. van, & Hagoort, P. (2012). The processing consequences of compositionality. In *The Oxford Handbook of Compositionality*. Oxford University Press.
- Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory and Cognition*, 8, 483–512.
- Cartson, R. (2002). *Thoughts and utterances*. Blackwell.
- de Almeida, R. D., & Dwivedi, V. D. (2008). Coercion without lexical decomposition: Type-shifting effects revisited. *Canadian Journal of Linguistics*, 53(2/3), 301–326.
- de Almeida, R. G., Riven, L., Manouilidou, C., Lungu, O., Dwivedi, V., Jarema, G., et al. (2009). *Coercion effects are pragmatic: fMRI and behavioral evidence*. Poster presented at the 15th AMLaP. Barcelona, Spain.
- Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas and thematic role concept. *Journal of Memory and Language*, 44, 516–547.
- Fodor, J. A., & Lepore, E. (1998). The emptiness of the lexicon: Reflections on James Pustejovsky's *The Generative Lexicon*. *Linguistic Inquiry*, 29(2), 269–288.
- Frisson, S., & McElree, B. (2008). Complement coercion is not modulated by competition: Evidence from eye movement. *Journal of Experimental Psychology: Language, Memory, and Cognition*, 34, 1–11.
- Jackendoff, R. (1997). *The architecture of the language faculty*. MIT Press.
- Lapata, M., Keller, F., & Scheepers, C. (2003). Intra-sentential context effects on the interpretation of logical metonymy. *Cognitive Science*, 27(4), 649–668.
- Lascares, A., & Copestake, A. (1998). Pragmatics and word meaning. *Journal of Linguistics*, 34, 387–414.
- Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae, K. (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Language, Memory, and Cognition*, 37(4), 913–934.
- McRae, K., Hare, M., Elman, J. L., & Ferretti, T. R. (2005). A basis for generating expectancies for verbs from nouns. *Memory and Cognition*, 33, 1174–1184.
- McRae, K., & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass*, 3/6, 1417–1429.
- Partee, B. H., ter Meulen, A., & Wall, R. E. (1993). *Mathematical Methods in Linguistics*. Kluwer.
- Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.
- Pylkkänen, L., & McElree, B. (2006). The syntax-semantic interface: On-line composition of sentence meaning. In *Handbook of Psycholinguistics* (p. 537–577). Elsevier.
- Schumacher, P., & Weiland, H. (2011). Reading Brecht and talking to the espresso: Electrophysiological investigations of conventional and novel metonymy. In *Proceedings of the Metonymy 2011 workshop*. Stuttgart, Germany.
- Traxler, M. J., Pickering, M. J., & McElree, B. (2002). Coercion in sentence processing: evidence from eye-movements and self-paced reading. *Journal of Memory and Language*, 47, 530–547.
- Van Der Meer, E., Krüger, F., & Nuthmann, A. (2005). The influence of temporal order information in general event knowledge on language comprehension. *Zeitschrift für Psychologie*, 213(3), 142–151.
- Zarcone, A., & Padó, S. (2010). "I like work: I can sit and look at it for hours" – Type clash vs. plausibility in covert event recovery. In *Proceedings of VERB 2010*. Pisa, Italy.
- Zarcone, A., & Padó, S. (2011). Generalized event knowledge in logical metonymy resolution. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (p. 944–949). Austin, TX: Cognitive Science Society.

# Sparse Population Code Models of Word Learning in Concept Drift

Byoung-Tak Zhang<sup>1,2</sup> (btzhang@bi.snu.ac.kr)

Jung-Woo Ha<sup>1</sup> (jwha@bi.snu.ac.kr)

Myunggu Kang<sup>1</sup> (mgkang@bi.snu.ac.kr)

<sup>1</sup>School of Computer Science and Engineering

<sup>2</sup>Cognitive Science and Brain Science Programs

Seoul National University, Seoul 151-744, Korea

## Abstract

Computational modeling has served a powerful tool for studying cross-situational word learning. Previous research has focused on convergence behaviors in a static environment, ignoring dynamic cognitive aspects of concept change. Here we investigate concept drift in word learning in story-telling situations. Informed by findings in cognitive neuroscience, we hypothesize that a large ensemble of sparse codes flexibly represents and robustly traces drifting concepts. We experimentally test the population coding hypothesis on children's cartoon videos. Our results show that learning the meanings of words over time is hard, especially when the concept evolves slowly, but the sparse population coding can handle the concept drift problem effectively while hypothesis elimination and simplistic parametric models have difficulty.

**Keywords:** Cross-situational word learning; statistical language learning; concept drift; meaning change; population coding.

## Introduction

Children learn the meaning of words rapidly and robustly across multiple situations (Smith & Yu, 2008). Computational modeling has served a powerful tool for precise investigation of the hypothesized mechanisms of word learning. Many computational models of word learning have been used to simulate and account for the observed patterns such as reference disambiguation, blocking, and long-term memory (Frank *et al.*, 2009, Kachergis *et al.*, 2010; Vlach & Sandhofer, 2010).

Existing computational models for word learning can be broadly divided into hypothesis elimination and associative learning (Fazly *et al.*, 2010). In the hypothesis elimination approach the learning process consists of eliminating incorrect hypotheses about word meaning, on the basis of a combination of a priori knowledge and observations of how words are used to refer to aspects of experience, until the learner converges on a single consistent hypothesis. For instance, Siskind (1996) presented an efficient algorithm for keeping track of just the necessary and possible components of word-meaning hypotheses consistent with a set of examples. A weakness of this approach is that some logically possible hypotheses may be ruled out a priori or the concepts cannot be recovered once they are eliminated.

Another approach to computational modeling of word learning is associative learning. Yu (2005), for example, studied a word-object association model in a unified framework of lexical and category learning. This model demonstrated the emergence of patterns observed in early word learning. Xu and Tenenbaum (2007) proposed a

probabilistic model of word learning. The Bayesian account aims to explain inductive learning at the level of computational theory rather than to describe psychological processes involved. Fazly *et al.* (2010) uses a probabilistic framework to propose an incremental associative model that deals with referential uncertainty. The proposed model is demonstrated to converge over time on the most likely meaning of the word in CHILDES data sets. However, this model does not incorporate alignment ambiguity and it is not clear how the model behaves if the concept drifts in the course of learning.

Concept drift is a fundamentally important phenomenon in language acquisition. It means that the statistical properties of the target concept, which the learner is trying to learn, change over time (Widmer & Kubat, 1996). For example, a child might think that all birds can fly until he/she observes an ostrich, at which time the child revises the concept of bird. This causes problems because the learning process needs some mechanisms to unlearn or revise the learned concepts. Simple hypothesis elimination cannot account for this since it lacks a mechanism for recovering the eliminated concepts. Both the associative learning and its probabilistic versions have difficulties since they strive to model global patterns, not modeling local patterns that might be necessary at a later stage.

Here we propose a computational model of word learning that deals with concept drift under alignment ambiguity and referential uncertainty. The model borrows ideas from neuroscience and uses a population coding (Pouget *et al.*, 2000; Ma *et al.*, 2006). We propose a sparse population-code network in which meanings of the words are represented as a large collection of sparse microcodes. Since each microcode is sparse, it describes a general concept. There are many of the microcodes and, thus statistically, only a few parts of them are updated on a single observation, maintaining stability by the remaining microcodes in the population. We test this population coding hypothesis on naturalistic children's cartoon video data. To make the experiments more realistic, we use state-of-the-art image processing techniques to represent the scene as a bag of image patches. This is contrasted with the previous studies of cross-situational word learning in which the scene representation adopts hand-coded semantic features. Our experimental results show that learning the meaning of words over time is hard, especially when the concept is drifting slowly. We demonstrate that the sparse population coding can handle the concept drift problem effectively



while simplistic parametric models have difficulty in dealing with the problem.

## Materials and Experimental Setup

### Video Data Sets

We used a series of children’s cartoon videos, *Maisy*, consisting of 6 episodes. Each episode plays for 48 to 105 minutes and the total play time is 475 minutes. From this video set, we prepared a total of 972 utterance-scene pairs as described in the following subsections. Cartoon videos provide naturalistic story-telling situations that children face in language acquisition (Zhang & Kang, 2011). An additional advantage of cartoons is that its image processing is relatively easy, allowing for automated generation of a large data set to study the long-term learning behavior in situated word learning.

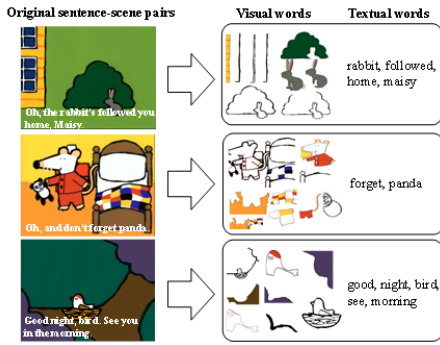


Figure 1: Examples of utterance-scene representation

### Utterance-Scene Representation

The material for cross-situational learning consists of utterance-scene pairs, where each pair is represented as a vector of the form

$$\mathbf{x}(t) = (\mathbf{w}(t), \mathbf{v}(t)) = (w_1, \dots, w_{|\mathbf{w}(t)|}, v_1, \dots, v_{|\mathbf{v}(t)|})$$

Here,  $|\mathbf{w}(t)|$  and  $|\mathbf{v}(t)|$  are the number (vocabulary size) of textual words and visual words in the  $t$ -th example, respectively. Figure 1 shows the examples of utterance-scene pairs extracted from the original videos. The following subsection describes how the textual words were processed.

### Language Processing

We collected all utterances in the text captions of the video set, which amounts to approximately 2,800. Removing simple utterances such as ‘Hi’ gives a total of 972 sentences. We determined the vocabulary for textual words by computing the standard TF-IDF (term-frequency and inverse-document-frequency) values. TF-IDF gives higher weights to the terms that frequently occur and are uncommon between episodes. This results in 1,049 words. We chose the top 448 textual words which defines the utterance vocabulary. The sound modality was not used in the experiments.

### Image Processing

We extracted image frames from the video, one frame for each of the 972 sentences extracted by language processing. Out of a stream of image frames played for the duration of speech of an utterance, we chose the image frame corresponding to the start of the utterance. This results in an image corpus of 972 scenes. Each scene was described by a subset of 7,520 image patches (i.e., visual words), each composed of the SIFT (scale-invariant feature transform) features and the color histogram extracted as follows. To define the visual words, we first used the MSER (maximally stable extremal region) feature extractor to segment and extract salient and informative regions from the images. SIFT was then used to find salient features in the extracted regions. The resulting features are grouped by K-means clustering to remove redundancy.

### Experimental Paradigm

Given the set of learning examples  $D_N = \{(\mathbf{w}(t), \mathbf{v}(t)) \mid t = 1, \dots, N\}$ , the goal of the learner is to form the concepts in the training set by finding the relationships between the words and the visual words (i.e. image patches). Learning proceeds incrementally, i.e. the examples are presented in sequence. Each time an example is presented the learner updates its model before the next example comes in.

Figure 2(b) shows the paradigm we adopt in this study. As indicated by the connections between textual words and those between visual words, we consider the fully interconnected relationship between different words and visual words. Note that this paradigm is contrasted with the standard paradigm shown in Figure 2(a), where the learner is to learn the relationship between the words and the referents or meanings, but do not attempt to learn the relationship among the words or among the referents.

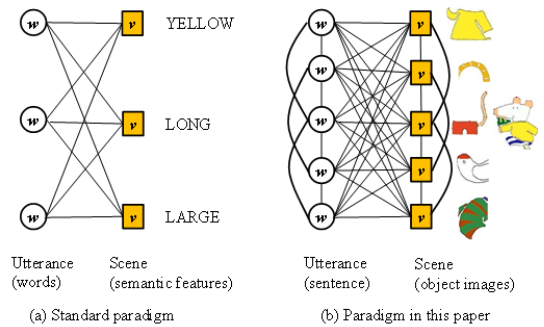


Figure 2: Experimental paradigms in comparison

## The Model

### Concept Representation

Meaning of words can be defined as a set of contexts in which word occurs in running text (Burgess & Lund, 1998) or represented in network connectivity revealed by statistical analysis of a text corpus (Steyvers & Tenenbaum,

2005). The textual domain can be extended to include the visual domain by taking into account the full contexts in which the word and images (visual words) co-occur in scenes (Zhang, 2008). Figure 3 illustrates this type of concept representation we adopt in this work. Here, the concept of MOUSE, for example, is defined as a collection of words ( $w$ -nodes), i.e. {yellow, run, dark, tall}, and a collection of visual words or visual patches ( $v$ -nodes) linked to the ‘mouse’-node in the figure. Thus, we consider the learner to acquire the visually-grounded linguistic concepts or the joint vision-language concepts, similar to the perceptual symbol systems *a la* Barsalou (1999).

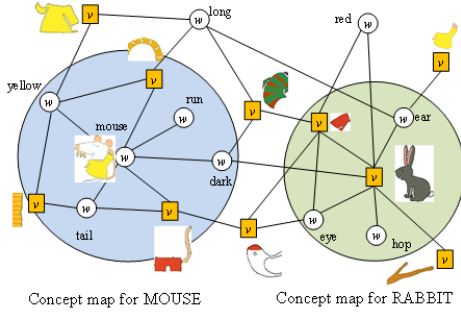


Figure 3: The concept of a word is defined by other words  $w$  as well as visual words  $v$ . In this representation the concepts are defined as a relationship among the primitives (words and visual words).

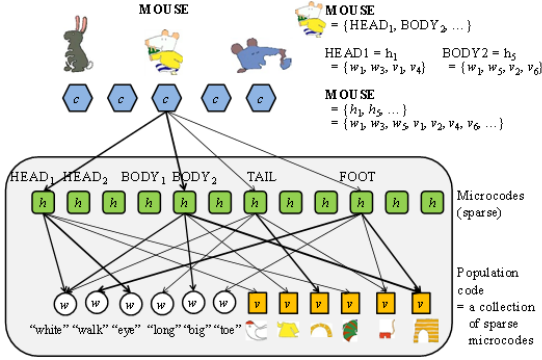


Figure 4: Sparse population coding scheme

### Sparse Population Coding

We represent the joint vision-language concept using sparse population codes. Figure 4 shows the basic units of the coding scheme, i.e. microcodes. Each microcode represents a prototype, exemplar, or common pattern for a set of similar examples. For instance, a microcode  $h = \{\text{'white'}, \text{'eye'}, v_1, v_4\}$  represents a class of objects (or concept HEAD1 as indicated in the figure) that have white eyes and image features of  $v_1$  and  $v_4$ , where  $v_1$  and  $v_4$  are image patches. The textual words, ‘white’ and ‘eye’, and the visual words,  $v_1$  and  $v_4$ , are instances of the textual and visual word vocabulary, respectively. Since the number of words or visual words chosen to define the specific microcode is small compared to their vocabulary size, this is a sparse

coding scheme. Typically we use a large number of microcodes to describe complex concepts.

The population of sparse microcodes can be considered as a three-layer network as shown in Figure 4. The first (bottom) layer consists of the  $w$ -nodes for words (e.g. “white”) and the  $v$ -nodes for visual words (image patches). The second (middle) layer represents the  $h$ -nodes for microcodes or micro-concepts such as HEAD1. A formal concept is represented as an ensemble of micro-concepts (or microcodes), as indicated by  $c$ -nodes at the third (top) layer of the network. This network can be learned from the data. Before describing the learning procedure we see the statistical background underlying this representation.

### Finite Mixture Model Formulation

Formally, a large collection of microcodes represents the empirical distribution of the concepts in the form of a finite mixture model (McLachlan & Peel, 2000). To see this, we suppose that the density of data  $\mathbf{x} = (\mathbf{w}, \mathbf{v})$  can be written in the form:

$$P(\mathbf{x} | \theta) = \sum_{j=1}^M \alpha_j f_j(\mathbf{x} | h_j) \quad (1)$$

where  $f_j(\mathbf{x} | h_j)$  are densities and  $\alpha_j$  are nonnegative quantities that sum to one:

$$0 \leq \alpha_j \leq 1 \quad (j=1, \dots, M) \quad \text{and} \quad \sum_{j=1}^M \alpha_j = 1.$$

Equation (1) is called  $M$ -component finite mixture density. Roughly, the configuration of the microcode defines the shape of the mixture component  $f_j(\mathbf{x} | h_j)$  and the weight associated with the microcode defines the mixing weight  $\alpha_j$ . We denote the complete collection of all distinct parameters occurring in the mixture model by  $\theta = (\boldsymbol{\alpha}, \mathbf{h})$ , where  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_M)$  and  $\mathbf{h} = (h_1, \dots, h_M)$ . We note that by designing the microcodes  $h_j$  appropriately to be the parameters of the component density  $f_j(\mathbf{x} | h_j)$ , the mixture density can be represented by the sparse population code.

In other words, if the microcode has an associated component density  $f_j(\mathbf{x} | h_j)$ , the distribution of the data set  $D_N = \{\mathbf{x}(1), \dots, \mathbf{x}(N)\}$  can be represented by the population code:

$$\begin{aligned} &P(\mathbf{w}(1), \mathbf{v}(1), \dots, \mathbf{w}(N), \mathbf{v}(N) | \theta) \\ &= P(\mathbf{x}(1), \dots, \mathbf{x}(N) | \theta) = \prod_{t=1}^N \sum_{j=1}^M \alpha_j f_j(\mathbf{x}(t) | h_j) \end{aligned} \quad (2)$$

which is a sum of  $M^N$  products of component densities. Each term in the summation is interpreted as the probability of obtaining a given one of the MN possible divisions of the observations among the groups.

### Learning Algorithm

Learning proceeds incrementally by observing each utterance-scene pair in sequence. On each observation of an

example  $(\mathbf{w}, \mathbf{v})$  the learner predicts and updates the meanings or concepts  $\theta$ . This is an inductive process and can be formally described as Bayesian inference:

$$P_t(\theta | \mathbf{w}, \mathbf{v}) = \frac{P(\mathbf{w}, \mathbf{v} | \theta) P_{t-1}(\theta)}{P(\mathbf{w}, \mathbf{v})} \quad (3)$$

At each time step  $t$ , the prior distribution  $P_{t-1}(\theta)$  of the hypothesis  $\theta$  is updated to the posterior distribution  $P_t(\theta | \mathbf{w}, \mathbf{v})$  of the hypotheses by computing the likelihood function  $P(\mathbf{w}, \mathbf{v} | \theta)$  and normalizing by

$$P(\mathbf{w}, \mathbf{v}) = \sum_{\theta'} P(\mathbf{w}, \mathbf{v} | \theta') P_{t-1}(\theta') \quad (4)$$

to make  $P_t(\theta | \mathbf{w}, \mathbf{v})$  back to a probability distribution. The posterior is then used as the prior for the next time step. Making the data set explicit, we can rewrite (3) in a recursive form:

$$\begin{aligned} P_t(\theta | \mathbf{w}(t), \mathbf{v}(t), \mathbf{w}(1:t-1), \mathbf{v}(1:t-1)) \\ = \frac{P(\mathbf{w}(t), \mathbf{v}(t) | \theta) P_{t-1}(\theta | \mathbf{w}(1:t-1), \mathbf{v}(1:t-1))}{P(\mathbf{w}(t), \mathbf{v}(t) | \mathbf{w}(1:t-1), \mathbf{v}(1:t-1))}, \end{aligned} \quad (5)$$

where  $\mathbf{w}(t)$  and  $\mathbf{w}(1:t-1)$  denote the word vector at time step  $t$  and the sequence of word vectors from time step 1 to  $t-1$ , respectively. Expectation-maximization (EM) style algorithms are usually used to solve the estimation problem (McLachlan & Peel, 2000). In the following we describe the method we implemented as a sparse population coding network. Recall that  $\theta = (\boldsymbol{\alpha}, \mathbf{h})$ , i.e. the concepts are represented as a collection of microcodes  $\mathbf{h}$  with weights  $\boldsymbol{\alpha}$  in the network. The population code is a mechanistic representation for psychological processes since it describes the memory encoding and decoding mechanisms more explicitly than simplistic parameter tuning models.

We first describe the learning algorithm in pseudocode and then explain it.

```

1  $H(0) \leftarrow \{\}, V_T \leftarrow \{\}, V_I \leftarrow \{\}$ 
2  $t \leftarrow 1$  ; prior  $P_{t-1}(\theta)$ 
3 Perceive  $\mathbf{x}(t) = (\mathbf{w}(t), \mathbf{v}(t))$ 
4  $E = \{h_1, \dots, h_m\} \leftarrow \text{Sample}(\mathbf{x}(t))$  ; microcodes
5  $V_T \leftarrow V_T + \{\text{new } w's\}, V_I \leftarrow V_I + \{\text{new } v's\}$ 
6  $H \leftarrow H + E$  ; accommodation
7 Repeat
8    $H' \leftarrow \text{Predict}(H)$  ; sampling prior  $P_{t-1}(\theta)$ 
9    $\mathbf{x}' \leftarrow \text{Generate}(H')$  ; likelihood  $P(\mathbf{x} | \theta)$ 
10   $H'' \leftarrow \text{Correct}(H', \mathbf{x}', \mathbf{x}(t))$  ; assimilation
    (resampling)
11   $H \leftarrow H''$ 
12 Until reconstruction_satisfactory( $\mathbf{x}(t)$ )
13  $H(t) \leftarrow H$  ; posterior  $P_t(\theta | \mathbf{x}(t))$ 
14  $t \leftarrow t+1$ 
15 Go to 3
```

Given an utterance-scene instance (line 3), a subset of words and a subset of visual words are selected to build a

microcode (line 4). For each utterance-scene pair, a number  $m$  of microcodes are generated randomly and repeatedly (line 4). Duplications are permitted and, in fact, the number of duplications represents the strength of the code (we will use this later on in decoding the referents or meanings of the words). The set  $E$  of new microcodes is then added to the existing set  $H$  of microcodes (line 6). This step is equivalent to accommodating new memory elements. Then the model is trained to tune or assimilate the incoming concepts into the existing concepts (lines 7-12). First, a collection  $H'$  of the microcodes is sampled to be used to generate an example  $\mathbf{x}'$ . The generated example is then compared to the training example. The difference is used to correct the model  $H$  or the population code. This results in the update of the posterior distribution (line 13).

The algorithm consists of basically three steps: i) sampling new microcodes (line 4), ii) merging them with the old (existing) population of microcodes (line 6), and iii) resampling of the whole microcode population (line 10) to correct the conflicts and interferences. To correct predictive errors in an unsupervised way, the algorithm test-generates the samples from the current model (lines 8 and 9) and compare the resulting data with the perceived data (line 10).

## Connection to Probabilistic Models of Cognition and Monte Carlo

Recall that the population of sparse codes approximates the probability distribution of the examples if the population size is big. Recall also that the learning algorithm is implemented by repeatedly sampling the sparse codes (microconcepts or hypotheses) like a Monte Carlo simulation does. In terms of Bayesian inference the learning algorithm updates the distribution of the concepts from prior to posterior distribution by Monte Carlo simulation. Shi *et al.* (2010) suggested that exemplar models are a successful class of psychological process models that can be used to perform a sophisticated form of Monte Carlo approximation. The similarity of the sparse coding representation with the exemplar model suggests that our sparse population code model offers a concrete process model of Bayesian cognition.

## Simulations and Results

### Parameter Setting for Experiments

Experiments were performed using the following parameter settings. Given a new observation, 10 new sparse codes were sampled and added to the population. Each microcode consists of three textual words and one visual word. 5 iterations of error correcting steps were executed to tune the whole population code to the new observation. To see the effects of memory capacity we experimented with two sizes of populations:  $|H| = 100, 500$ . When the population size exceeds the memory capacity, we replace 10 microcodes with the lowest weight values by 10 new microcodes. We define two scores for measuring the similarity between visual and textual concepts as follows:

$$S(w, v) = \frac{1}{\sum_{h \in H_w} \alpha(h)} \cdot (|H_v|)^{\frac{1}{2}} \cdot \sum_{h \in H_v} \alpha(h) \text{ and } S(w) = \sum_{h \in H_w} \alpha(h)$$

where  $\alpha(h)$  is the weight of microcode  $h$ , and  $H_w$  and  $H_v$  are the subsets of  $H$  consisting of the microcodes with textual word  $w$  and visual patch  $v$ , respectively.

### Vocabulary Growth

Figure 5 shows the growth of visual and textual vocabularies as learning proceeds. When the memory size is unlimited (left), the size of both visual words and textual words increases continuously (linearly). When the maximum memory capacity is set to be limited to 500 (right), the size of visual words increases first and then decreases while the number of textual words grow continuously but in two stages of fast growth and then slow growth. The difference in vocabulary growth pattern seems in part due to the difference in vocabulary size of visual and textual words, i.e. in this experimental setting, 7520 visual words and 448 textual words were used for candidates.

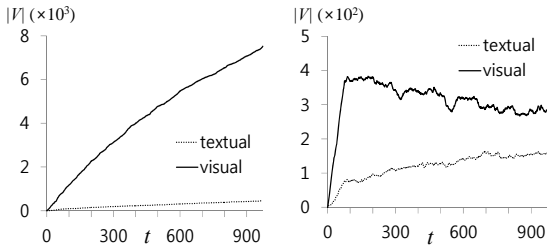


Figure 5: Growth of vocabulary. (left) unlimited memory size. (right) limited memory size.

### Word Learning in Concept Drift

Figure 6 shows the trace of concept memory for the 4 separate focus objects which appear in all episodes.

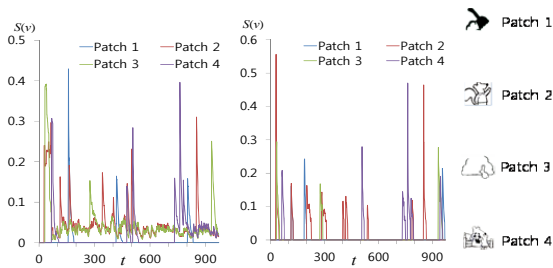


Figure 6: Emergence, extinction, and re-emergence of concepts in drift. (left) larger memory capacity ( $|H| = 500$ ). (right) smaller memory capacity ( $|H| = 100$ ).

The results show the emergence, extinction, and re-emergence of different visual concepts as the video runs. If the memory size is relatively big (500 in this case), the concepts do not extinct totally and remain in the backend to re-emerge when new similar observations are made. In contrast, when the memory size is small (100 in this case), the concepts disappear entirely from the memory,

suggesting the difficulty of the problem, especially if the memory capacity is small. However, this problem can be solved by dynamically varying the population size to balance exploration and exploitation. In contrast to this sparse population coding approach, a localist, eliminative method would have a fundamental difficulty in recovering once-eliminated concepts due to its lack of associative connections between concepts.

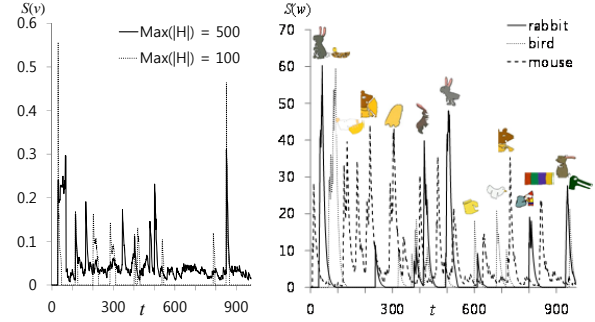


Figure 7: (left) Emergence patterns of concepts for different memory size. Plotted are the weight values for the specific visual concept shown. (right) Emergence of different visual concepts for given three textual concept (rabbit, bird, and mouse).

Figure 7 shows the change of concepts in the course of learning. (left) shows the change of weight distribution for the specific visual concept (patch 2 in Figure 6) shown. (right) is the reverse, i.e. the query is given by a textual word (rabbit) and the graph shows how the corresponding image concept changes as learning proceeds. It can be observed that, since the concept of rabbit drifts, different types of rabbit images and, sometimes very different (and wrong) images, are retrieved by the same word.

### Concept Generalization and Specialization

Figure 8 shows the joint vision-language concept maps around the ‘rabbit’ as they evolve over the 6 episodes. The maps (a)-(d) are the snapshots after watching 1, 2, 4, 6 videos, respectively. Note that the map contains visual words as well as textual words. We observe that the connectivity of the visual-linguistic map grows as more episodes are learned. Careful examination of the map shows the role of visual words or concepts for the specialization and generalization of the textual concepts and vice versa. For example, in Figure 8(a) we observe that a visual word connects the three textual words of ‘enjoy’, ‘lunch’, and ‘rabbit’ together. This adds an additional, visually-grounded, connection (association) between the words ‘enjoy’ and ‘lunch’. We also observe that the word ‘rabbit’ is connected to multiple images, again grounding and refining the meaning of the textual word. Formally, the former is the one-image to many-words relationship and the latter is the many-images to one-word relationship. This again shows the effect of visual generalization of the textual words and that of visual specialization, respectively, which cannot be observed in language-only concept maps (Zhang & Kang, 2011).





# The role of recent versus future events in children's comprehension of referentially ambiguous sentences: Evidence from eye tracking

Lu Zhang (lzhang@cit-ec.uni-bielefeld.de)

Lily Kornbluth (lilykornbluth@fulbrightmail.org)

Pia Knoeferle (knoeferl@cit-ec.uni-bielefeld.de)

Cognitive Interaction Technology Excellence Center, University of Bielefeld,  
Bielefeld, Germany

## Abstract

Findings from recent eye-tracking studies suggest that adults prefer to rely more on recently seen events than possible future events during sentence comprehension: When the verb in an NP1-VERB-ADV-NP2 sentence was referentially ambiguous between a recent action and an equally possible future action, adults fixated the target of the recent action more often than the not-yet-acted upon object (Knoeferle & Crocker, 2007; Knoeferle, Carminati, Abashidze, & Essig, 2011). We examined whether this preference for the recent event generalizes to five-year-old children. In an eye-tracking study, five-year-olds were presented a display with an animal and two other objects. On the next picture frame, the animal was depicted as performing an action (e.g., a horse galloped to a blue barn). Next, a spoken sentence referred either to an event involving the acted upon target object (the blue barn) or to an equally plausible future action event (e.g. galloping to the red barn). At the adverb in NP1-VERB-ADV-PP sentences, children fixated more often the recent (vs. future) event target. This result replicates the findings from the adult studies and suggests that, just like adults, children rely more on the recent event than expectations of an event that could happen next. At the same time, visual context effects of the recent events were subtly delayed for children (vs. adults). For adults, the recent-event preference emerged during the verb; for children, by contrast, it emerged post-verbally during the adverb. Thus, similar attentional mechanisms underlie visual context effects in both 5-year old children and adults but their time course differs.

**Keywords:** eye tracking; child language comprehension; visual context; depicted events.

## Introduction

Adults have been shown to rapidly and efficiently integrate all sorts of contextual cues during real-time spoken language comprehension. Referential contrast between objects can incrementally influence syntactic disambiguation (e.g., Tanenhaus, Spivey, Eberhard, & Sedivy, 1995), as well as semantic interpretation (Sedivy et al., 1999). Object affordances (Chambers et al., 2004) and depicted events (Knoeferle et al., 2005; Knoeferle et al., 2008) can also rapidly affect structural disambiguation.

Sometimes contextual information can even permit adult comprehenders to actively derive expectations about upcoming information. One source of evidence for predictive processes has been “anticipatory” eye movements to target objects (i.e., eye movements to objects just before they are mentioned). Verb selectional restrictions (Altmann & Kamide, 1999), compositional noun and verb meaning

and associated world knowledge (Kamide et al., 2003a,b), prosody (Weber et al., 2006), and information structure (Kaiser & Trueswell, 2005) can each restrict the range of target objects that can be mentioned next, as evidenced by participants inspecting a target object before its mention relative to a control condition.

Like adults, children have also been shown to derive expectations about upcoming information. When children (aged 10-11) listened to a sentence in which the verb *eat* restricted the domain of reference to edible objects, they fixated the only edible object in context shortly after hearing *eat* and before that object was actually mentioned (Nation, Marshall, & Altmann, 2003). This was true for both (verbally and visually) less-skilled children as well as for normally developing children although the former (vs. the latter) made more and shorter fixations to the edible object.

At first glance this might give the impression that child and adult language comprehension and expectation formation are governed by similar mechanisms. Results from another study with younger children (mean age of 4.7), by contrast, suggest marked differences between child and adult language comprehension. In a study by Trueswell et al. (1999), children either heard a locally structurally ambiguous sentence such as *Put the frog on the napkin in the box* or an unambiguous sentence such as *Put the frog that's on the napkin in the box*. For the ambiguous sentence, the prepositional phrase *on the napkin* can either modify the noun indicating the location of the frog, or attach into the verb phrase and specify the destination of the action. Children saw either only one possible referent for *frog* in the 1-referent condition (e.g. a frog on the napkin, an empty napkin, a distractor object and a box) or two referents in the 2-referent condition (e.g. a frog on the napkin, another frog, an empty napkin and a box). A 2-referent context should bias comprehenders to look at the frog on the napkin upon hearing *on the napkin* rather than to the empty napkin.

However, five-year old children frequently looked at the *incorrect* destination (the empty napkin) in both one-referent and two-referent contexts. Adults, by contrast, looked first at the correct target (the frog on the napkin) when hearing *Put the frog on the napkin* and then to the correct destination (the box) rather than the empty napkin in a context with two frogs. These findings were taken as support for the claim that children – unlike adults – incorrectly interpreted the prepositional phrase *on the napkin* as the destination for *put*, and that they were unable

to use the referential contrast (between two frogs) for structural disambiguation. Moreover, children's actions indicated that they never revised this initial misanalysis: On 60% of the trials they performed an action that involved the incorrect destination (e.g., moving a frog to the empty napkin before putting it in the box).

Accordingly, at least some aspects of 5-year-old children's and adults' real-time language comprehension appear to differ. What is not clear, however, is to what extent children's (vs. adults') use of (visual) context for spoken language comprehension is indeed limited, and, more broadly, what demarcates child-adult comprehension differences. Perhaps 5-year-old children and adults are not that dissimilar in their comprehension mechanisms and only employ different (attention) mechanisms in a few isolated instances. Alternatively, children at that age still differ fundamentally from adults in how they use (visual) contextual cues for language comprehension. This is an interesting research question since processing accounts of situated language comprehension (e.g., Knoeferle & Crocker, 2006, 2007) will ultimately want to accommodate language processing from infancy to young-adulthood to older age.

Existing findings suggest similarities in how children versus adults process language in (visual) context, but there are also some differences. Just like adults, infants as young as six months of age can track moving objects with their gaze (Richardson & Kirkham, 2004). 36-month-olds also exhibit adult attention behavior in that they shift their visual attention more quickly to a target picture when they hear *blue car* in a context with a blue and a red car than when the context shows a blue car and a blue house (Fernald, Thorpe, & Marchman, 2010). This suggests that they can rapidly use linguistic input to fixate relevant referents. In younger children, by contrast, this behavior is not yet apparent. Furthermore, when 19-months-old infants listened to nouns as they saw matching (vs. mismatching) objects, their event-related brain potentials to the noun exhibited an N400 (a negativity approximately 400 ms after stimulus onset, see Kutas & Hillyard, 1984) that was larger for mismatches than matches. That negativity was also found in adults, but the scalp distribution and latency of that effect differed in children relative to adults (Friedrich & Friederici, 2004). In summary, it is unclear to what extent children throughout language development and adults share the same mechanisms in language comprehension, language-mediated visual attention, and visual context effects on comprehension.

The present research contributes to this emerging evidence about real-time situated language processing in children by examining how recently-depicted action events guide children's visual attention and spoken language comprehension. We know that adults can rapidly draw on recent action events in informing language comprehension and in interrogating visual context (Knoeferle & Crocker, 2007). Participants saw a character (a waiter) move toward an object, interact with it (e.g., polish candelabra), and move

away from it. They then listened to an utterance that referred either to the recent action (polishing the candelabra: simple past tense: *Der Kellner polierte kürzlich die Kerzenleuchter*, "The waiter recently polished the candelabra") or to an equally plausible action that hadn't yet been performed (e.g., polishing crystal glasses; present tense with future meaning: *Der Kellner poliert sogleich die Kristallgläser*, "The waiter will soon polish the crystal glasses"). At the verb *poliert* . . . ('polish...') the comprehension system and visual attention had a choice between anticipating the recent action target versus anticipating (and thus inspecting) the target of the as-yet-unseen future action. Adult participants preferentially anticipated the target of the recent (vs. the other, future) action, a gaze pattern that continued even as future tense information became available through the adverb (e.g., *sogleich*, 'soon'). Verb meaning and future tense information did not elicit expectations of future events, and adults relied on the recently inspected events. Recent research has replicated these results with real-world stimuli. In addition, the recent-event preference replicated even when both 'recent' and 'future' events were equally frequent but tense effects were then more pronounced (i.e., participants always saw one action before and another action after sentence comprehension, Knoeferle, Carminati, Abashidze, & Essig, 2011).

The present experiment used eye tracking to see to what extent 5-year-olds can also rely on recent events in directing their visual attention and language comprehension. To this end, 5-year-olds saw clipart depictions such as a horse and two stables, one red and one blue (see Fig. 1). The horse moved to the blue stable (Fig. 1b). Subsequently the child would hear *Das Pferd galoppierte gestern zu der blauen Scheune*, (literal translation: "The horse galloped yesterday to the blue barn", "Yesterday, the horse galloped to the blue barn") or *Das Pferd galoppiert morgen zu der roten Scheune* (literal translation: "The horse gallops tomorrow to the red barn", "Tomorrow, the horse will gallop to the red barn"). If 5-year-olds rely on recent events with the same time course as adults, then we should see them inspect the target of the recent event (the blue barn) more often than the target of the future event (the red barn) during the verb and post-verbal adverb. While tense information is available post-verbally, there was only a (non-reliable) tendency for tense effects post-verbally in the adults (Knoeferle & Crocker, 2007, Experiment 3; Knoeferle et al., 2011, Experiment 1). Inspections to the target of the future event (the red barn) in children should thus only increase as that target is mentioned.

## Experiment

### Participants

24 kindergarten children (10 4-year-olds and 14 5-year-olds, range: 4-5;9) took part in the Experiment and received a small toy for their participation. All participants had German as their only mother tongue and normal or



corrected-to-normal vision. All were unaware of the experiment purpose. Children and one of their parents gave informed consent.

## Materials and Design

There were sixteen items, and two sentence conditions (Figure 1 and Table 1). Each item consisted of a series of three clipart scenes and four sentences. We created the pictures by using commercially available clipart and graphics programs. The first frame of the scene displayed a central animal agent (a horse) and two objects (e.g., a blue barn and a red barn, Figure 1a). The objects on either side of the animal were identical mirror images that only differed in their color or size (e.g., red barn, blue barn). The verb of the sentence (e.g., *galoppieren* ‘gallop’, see example in Table 1) was always a motion verb. Both of the two objects (e.g. the blue barn and the red barn) were equally plausible targets of the event (e.g. horse-galloping). However, the agent approached only one of the two objects (e.g. galloping to the blue barn, Figure 1b) and then moved back to another center position (Figure 1c). Each frame was presented for 1500 ms. The sentence could either refer to a past event (Table 1a, *Das Pferd galoppierte gestern zu der blauen Scheune*. ‘The horse galloped yesterday to the blue barn.’) or a future event (Table 1b, *Das Pferd galoppiert morgen zu der roten Scheune*. ‘The horse gallops tomorrow to the red barn.’) Figure 1a’-c’ and Table 1a’-b’ were the counterbalanced version in which the red barn was the target of the recent action. Therefore, each object was the target of a past and a future action once. This ensured that visual characteristics of the post-verbally referenced target object contributed equally to each critical condition.

We also counterbalanced the presentation side of each object. As shown in Figure 1, the blue barn was on the left side and the red barn was on the right side. In the counterbalancing version (not shown), the red barn was on the left side and the blue barn was on the right side.

In addition to the 16 experimental items, we created 8 filler items to ensure that children were exposed to a range of other sentence structures and actions. The two conditions of the sentence (past vs. future tense), the counterbalancing of the target object, and the counterbalancing of the target

object presentation side led to eight basic lists. Lists were pseudo-randomized and each participant saw an individually randomized version of one of the eight experimental lists.

## Procedure

An EyeLink1000 remote eye-tracker with a sampling rate of 500 Hz monitored participants’ eye movements. Images were presented on a 22" LCD color monitor at a resolution of 1680×1050 pixels concurrently with a spoken sentence. We only tracked the right eye, but viewing was binocular. At the beginning of the experiment, each child was instructed to play a game. In this game, children were asked to inspect the images and to listen to the sentences. After each trial, they heard a question about the previous sentence and were asked to try to answer it correctly.

Each trial started with the display of a series of three frames which depicted an action (e.g., Fig. 1a-c). Each of the three frames in Fig. 1 was presented for 1500 ms (totaling 4500 ms). After that, the third image remained on the screen and the sentence was played via speakers. Five hundred milliseconds after the offset of the sentence, a spoken question asked for the target object of the verb in the previous sentence (for example, *Wohin galoppierte das Pferd?/Wohin galoppiert das Pferd?* ‘Where did the horse gallop?/Where does the horse gallop?’). Participants’ task was to answer the question by naming the correct destination.

At the start of the experiment, each participant was shown two example image sequences and sentences. Next, participants were set up and calibrated manually using a five-point fixation stimulus. The black dot that is used to calibrate adults was replaced by a smiley face to attract children’s attention. The EyeLink software validated calibration; if validation was poor, the calibration procedure was repeated until validation was good. Between the individual trials, participants saw a centrally-located smiley on the screen which they were asked to fixate. This allowed the eye-tracking software to perform a drift correction if necessary. The entire experiment lasted approximately 25 minutes.

Table 1: Example item sentences

Picture	Condition	Sentence
Figure 1a-c	Past tense	(a) <i>Das Pferd galoppierte gestern zu der blauen Scheune..</i> The horse galloped yesterday to the blue barn.
	Future tense	(b) <i>Das Pferd galoppiert morgen zu der roten Scheune.</i> The horse gallops tomorrow to the red barn.
Figure 1a’-c’	Past tense	(a’) <i>Das Pferd galoppierte gestern zu der Roten Scheune..</i> The horse galloped yesterday to the red barn.
	Future tense	(b’) <i>Das Pferd galoppiert morgen zu der blasuen Scheune.</i> The horse gallops tomorrow to the blue barn.

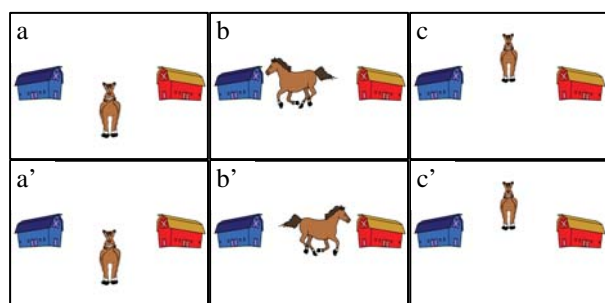


Figure 1: Example item pictures. In Figure 1 a-c, the horse gallops to the blue barn. To counterbalance visual characteristics of the target (e.g., its color), the horse gallops to the red barn in Figure 1 a'-c'.

## Analysis

For the purpose of inferential analyses, we defined three time windows: an exact verb region (from verb onset until its offset); the extended adverb region (from verb offset until adverb offset) and the PP region (from preposition onset until sentence end). We coded participants' fixations to four areas of interest in the scene: the agent (e.g., the horse in Figure 1); the recently acted upon object (e.g., the red barn in Figure 1a-c); the future object (e.g. the blue barn in Figure 1a-c); and the background. Of those, the recent and future objects were our target areas of interest.

The proportions of fixation on the target areas of interest (the recent and the future targets) were entered into log-ratio analyses (c.f., Arai, van Gompel & Scheepers, 2007; Carminati, van Gompel, Scheepers, & Arai, 2008; Knoeferle et al., 2011). We computed mean log gaze probability ratios for the recent object relative to the future object  $\ln(P(\text{future target})/P(\text{recent target}))$  for each condition and for each time window. Then we entered the log probability ratios into a one-factor (*tense*) ANOVA. Separate models were fitted for log-ratios averaged over participants and items respectively. We report the *p*-values for these analyses. To test whether the log probability ratios of each condition differs significantly from zero, we conducted simple *t*-tests. We adjusted the significance level of the *p*-values using the Bonferroni correction.

For the descriptive overview of the time course of the eye-movement data, we divided the utterance from sentence onset into time slots of 250 ms each. For each time slot and target object, we computed the number of fixations that fell within a given time slot. Then we plotted the mean proportion of fixation counts per time slot separately for each sentence condition and each target object.

## Results

Figures 2a) and 2b) plot the mean proportion of fixations to the two objects in the future and past tense conditions using time slots of 250ms. Figure 3 zooms in on one region of

interest and presents the mean log gaze probability ratios ( $\ln(P(\text{future target})/P(\text{recent target}))$ ) per condition for the adverb region.

Figures 2a) and 2b) illustrate an overall preference for fixating the acted-upon object rather than the not-acted-upon object from the offset of the verb until well into the NP2, irrespective of tense condition. As illustrated in Figure 2, the preference for looking at the acted-upon object is much reduced and reverses as children hear the second noun phrase in the future compared to the past tense condition.

In agreement with the descriptive pattern, the inferential analysis revealed no significant main effect of tense at both the verb and the adverb region. To test whether children had a preference to inspect one of the two targets, we examined whether the intercept was significantly different from zero. At the verb region this was not the case. By contrast, simple *t*-tests confirmed that log probability ratios of both the future tense condition and the past tense condition by subjects ( $ps < 0.05$ ) and of the past tense condition by items ( $p < 0.005$ ) were significantly different from zero for the adverb region (see Figure 3). This corroborates the findings from the descriptive analysis and indicates that children looked more often at the recently-acted-upon object than at the not-yet-acted-upon object in both the past and future tense conditions. For the PP region, by contrast, analyses confirmed that the children inspected the target objects as they were named (both  $ps < 0.002$ ).

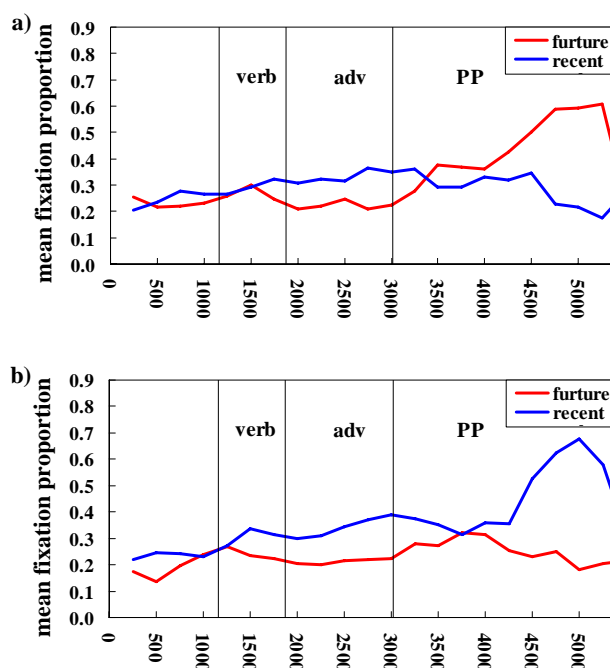


Figure 2: Eye movements to the target of the recent event (recent target: blue lines) and the target of the future event (future target: red lines) from sentence onset to sentence end for a) future tense condition and b) past tense condition

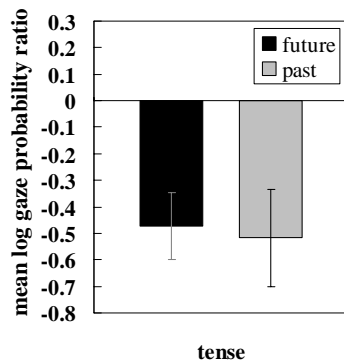


Figure 3: Children's mean log gaze probability ratios ( $\ln(P(\text{future target})/P(\text{recent target}))$ ) per condition for the adverb region (Error bars represent the standard error of the mean log gaze probability ratios)

## Discussion

The present research assessed whether 5-year-old children resemble adults in how and when they make use of recently-inspected clipart events during spoken language comprehension. We conducted an eye-tracking experiment in which we monitored 5-year-olds' eye movements to target objects in a clipart picture as they listened to related sentences. The children saw an animal move towards one of two equally plausible objects (e.g., the horse would gallop to the blue barn when a blue and a red barn were depicted). When the motion verb in an ensuing spoken sentence referred either to that recent action or to another action that hadn't yet happened, children preferred to inspect the target of the recent event (e.g., the blue barn) over the target of the as-yet-unseen but plausible other event (e.g., the red barn).

This finding confirmed clear similarities in how children and adults (Experiment 3 in Knoeferle & Crocker, 2007) direct their visual attention during spoken language comprehension: both of these participant groups preferred to inspect the recent-event (vs. future-event) target. The time course of visual attention, however, was delayed for children relative to adults. While adults in Experiment 3 by Knoeferle and Crocker (2007) began to inspect the recent-event target more often during the verb, the recent-event inspection preference only emerged post-verbally for children.

One open question is what underlies the recent-event preference in both children and adults. The experiment procedure introduced a frequency bias for the recent events. While participants in Experiment 3 by Knoeferle and Crocker (2007) and in the present study saw an event before each experimental trial, they never saw a post-sentence future event acted out. The procedure of never depicting the future event may have created a within-experiment

frequency bias toward relying more on recently depicted than on equally plausible future events for comprehension. It is possible that this bias led participants to preferentially inspect the recent (vs. future) event target.

Indeed, statistical regularities play an influential role in a range of cognitive processes for both children and adults. At 8 months of age, children can already use statistical regularities in linguistic input to segment words in fluent speech (Saffran, Aslin, & Newport, 1996). Statistical factors also play a role in children's visual attention to novel (vs. known) object patterns. When circles appeared in a pattern, infants at 11 months inspected novel (vs. known) circle sequences longer; by contrast, that behavior was not yet present at 8 months of age (Kirkham, Slemmer, Richardson, & Johnson, 2007, Experiment 1). For adults, statistical regularities play a role in language processing and other cognitive and motor processes. Adults' short-term linguistic experience can modulate their language production (Kaschak, Loney, & Borreggine, 2006; Haskell, Thornton, & MacDonald, 2010) and sentence reading (Wells, Christiansen, Race, Acheson, & MacDonald, 2009). It also affects adults' action execution (e.g., Chapman, Gallivan, Wood, & Milne, 2010) and visual perception (e.g., Chun & Jiang, 1999).

For the recent-event inspection bias in adults, however, frequency biases appear to play no causal role. When recent and future events were performed equally frequently within the experiment, effects of tense appeared somewhat earlier than in Knoeferle and Crocker (2007, Experiment 3), during the post-verbal adverb. By contrast, adults' recent-event inspection bias during the verb remained largely unchanged (Knoeferle et al., 2011, Experiment 2).

To what extent this preference generalizes to 5-year-olds when both recent and future events are equally frequent is unclear. What is clear, however, is that children, like adults, rapidly used recently inspected clipart events during comprehension, but that the time course of these event effects was delayed in children. For accounts of situated language comprehension (e.g., Coordinated Interplay Account, Knoeferle & Crocker, 2006, 2007), the present findings together with the other results that we discussed suggest that the closely temporally coordinated interplay of language comprehension, visual attention, and visual context effects on comprehension has a developmental basis.

## Acknowledgments

This research was funded by the Cognitive Interaction Technology Excellence Center (German Research Foundation, DFG). We thank Linda Krull and Eva Mende for their assistance with preparing the stimuli and collecting data. We thank Maria Nella Carminati for advice regarding the analyses and Helene Kreysa for help with the Experiment

Builder software. We also thank all the participating families and students for their support.

## References

- Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73, 247–264.
- Arai, M., van Gompel, R., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, 54, 218–250.
- Carminati, M. N., Gompel, R. P. G. van, Scheepers, C., & Arai, M. (2008). Syntactic priming in comprehension: the role of argument order and animacy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 1098–1110.
- Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *JEP: LMC*, 30, 687–696.
- Friedrich, M. & Friederici, A.D. (2004). N400-like semantic incongruity effect in 19-month olds: processing known words in picture contexts. *Journal of Cognitive Neuroscience*, 16, 1465–1477.
- Chapman, C. S., Gallivan, J. P., Wood, D. K., & Milne, J. L. (2010). Reaching for the unknown: Multiple target encoding and real-time decision-making in a rapid reach task. *Cognition*, 116, 168–176.
- Chun, M. M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, 10, 360–365.
- Haskell, T. R., Thornton, R., & MacDonald, M. C. (2010). Experience and grammatical agreement: statistical learning shapes number agreement production. *Cognition*, 114, 151–164.
- Kaiser, E., and Trueswell, J. C. (2005). The role of discourse context in the processing of a flexible word-order language. *Cognition*, 94, 113–147.
- Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003b). Integration of syntactic and semantic information in predictive processing: cross-linguistic evidence from German and English. *JPR*, 32, 37–55.
- Kamide, Y., Altmann, G. T. M., & Haywood, S. (2003a). The time course of prediction in incremental sentence processing. *Journal of Memory and Language*, 49, 133–156.
- Kaschak, M. P., Loney, R. A., & Borreggine, K. L. (2006). Recent experience affects the strength of structural priming. *Cognition*, 99, B73–B82.
- Kirkham, N., Slemmer, P., Richardson, D., & Johnson, S. P. (2007) Location location location: Development of spatiotemporal sequence learning in infants. *Child Development*, 78, 1559–1571.
- Knoeferle, P. 2007. “Comparing the time-course of processing initially ambiguous and unambiguous German SVO/OVS sentences in depicted events”, in: R. Gompel van, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movement research: insights into mind and brain*. Oxford: Elsevier, 517 – 533.
- Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment. *Cognition*, 95, 95–127.
- Knoeferle, P., & Crocker, M.W. (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science*, 30(3), 481 – 529.
- Knoeferle, P., & Crocker, M. W. (2007). The influence of recent events on spoken language comprehension: evidence from eye movements. *JML*, 57 (2), 519–543.
- Knoeferle, P., Habets, B., Crocker, M. W., & Münte, T. F. (2008). Visual Scenes Trigger Immediate Syntactic Reanalysis: Evidence from ERPs during Situated Spoken Comprehension. *Cerebral Cortex*, 18, 789–795.
- Knoeferle, P., Carminati, M., Abashidze, D., & Essig, K. (2011). Preferential inspection of recent real-world events over future events: evidence from eye tracking during spoken sentence comprehension. *Front. Psychology* 2:376. doi: 10.3389/fpsyg.2011.00376
- Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307, 161– 163.
- Nation, K., Marshall, C. M., & Altmann, G. (2003). Investigating individual differences in children’s real-time sentence comprehension using language-mediated eye movements. *Journal of Experimental Child Psychology*, 86, 314–329.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632–1634.
- Trueswell, J., Sekerina, I., Hill, N., & Logrip, M. (1999). The kindergartenpath effect: Studying on-line sentence processing in young children. *Cognition*, 73, 89 –134.
- Richardson, D. C. & Kirkham, N.Z. (2004). Multi-modal events and moving locations: Eye movements of adults and 6-month-olds reveal dynamic spatial indexing. *JEP: General*, 133, 46–62.
- Saffran, J., Aslin, R. N., & Newport, E. (1996). Statistical learning by 8-month old infants. *Science*, 274, 1926–1928.
- Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., and Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71, 109–148.
- Weber, A., Grice, M., & Crocker, M. W. (2006). The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements. *Cognition*, 99, B63–B72.
- Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. C., & MacDonald, M. C. (2009). Experience and sentence processing: statistical learning and relative clause comprehension. *Cognition*, 58, 250–271.

# Updating: Learning versus supposing

**Jiaying Zhao (jiayingz@princeton.edu)**

Department of Psychology, Green Hall,  
Princeton University, NJ 08540 USA

**Vincenzo Crupi (vincenzo.crupi@unito.it)**

Department of Philosophy, University of Turin, Turin, Italy

**Katya Tentori (katya.tentori@unitn.it)**

DiSCoF, CIMeC, University of Trento, Rovereto, Italy

**Branden Fitelson (branden@fitelson.org)**

Department of Philosophy, Rutgers University,  
New Brunswick, NJ 08901 USA

**Daniel Osherson (osherson@princeton.edu)**

Department of Psychology, Green Hall,  
Princeton University, NJ 08540 USA

## Abstract

Bayesian orthodoxy posits a tight relationship between conditional probability and updating. Namely, the probability of an event  $A$  after learning an event  $B$  should equal the conditional probability of  $A$  given  $B$  prior to learning  $B$ . We examine whether ordinary judgment conforms to the orthodox view. In three experiments we found substantial differences between the conditional probability of an event  $A$  supposing an event  $B$  compared to the probability of  $A$  after having learned  $B$ . Specifically, supposing  $B$  appears to have less impact on the credibility of  $A$  than learning that  $B$  is true. Thus, Bayesian updating seems not to describe the relation between the probability distribution that arises from learning an event  $B$  compared to merely supposing it.

**Keywords:** belief updating, reasoning, probability

## Introduction

Let  $Pr_1$  represent the beliefs of an idealized agent who is considering at time 1 the credibilities of events over an outcome space  $\Omega$  (finite, for simplicity). Suppose that for some event  $B \subseteq \Omega$  with  $Pr_1(B) > 0$  experience intervenes at time 2 to convince the agent that  $B$  is (definitely) true. What new distribution  $Pr_2$  should embody the agent's revised beliefs? The Bayesian response (Hacking, 2001, Ch. 15) is that  $Pr_2$  should be the result of conditioning<sup>1</sup>  $Pr_1$  on  $B$ , that is:

(1) BAYESIAN UPDATING:

If  $B \subseteq \Omega$  is learned between times 1 and 2 (and nothing else relevant is learned) then for all events  $A \subseteq \Omega$ ,  $Pr_2(A) = Pr_1(A | B)$  (provided that  $Pr_1(B) > 0$ ).

It is easy to check that  $Pr_2$  as defined by (1) is a genuine probability distribution and that  $Pr_2(B) = 1$  (as expected). Also, (1) is a consequence of compelling axioms on belief change (Gärdenfors, 1988, §5.2), and its violation exposes the agent to sure-loss betting contracts (Harman, 1999, §4.12).

<sup>1</sup>By conditioning we refer to simple or strict Bayesian updating, rather than associative learning.

Such normative virtues suggest a psychological question. One way of formulating (1) is that *supposing* an event  $B$  should have the same impact on the credibility of an event  $A$  as *learning*  $B$ . Is this true for typical assessments of chance? For example, is the judged probability of a Democratic victory in 2012 *supposing* that Hilary Clinton is the vice presidential candidate the same as the judged probability of a Democratic victory in 2012 after *learning* that Clinton, as a matter of fact, is the vice presidential candidate?

The foregoing question is orthogonal to the provenance of conditional probability in the mind, that is, to the way such probabilities are mentally computed. Thus, even if people fail to respect the standard definition:

$$Pr(A | B) \stackrel{\text{def}}{=} \frac{Pr(A \cap B)}{Pr(B)}$$

it is still possible for (1) to hold.<sup>2</sup> All that matters is whether the same degree of confidence in event  $A$  is reached when supposing event  $B$  compared to learning it. On the other hand, recent literature on conditional reasoning has suggested a difference between supposing vs. learning the antecedent of a conditional (Oaksford & Chater, 2007). We investigated the difference between learning and supposing in three experiments.

## Experiment 1

### Participants

Forty undergraduates (27 female, mean age 19.4 yrs, SD = 1.3) from Princeton University participated in exchange for course credit.

<sup>2</sup>An earlier study focussed on the fidelity of the standard definition (above) to the numbers people report as conditional probabilities (Zhao, Shah, & Osherson, 2009).

## Materials

Five decks of cards served as stimuli, each with 20 cards. Each card presented an animal and a colored square on one side and was blank on the other side. The animal was marked on the bottom half of the card and could be either a dog or a duck. The colored square was marked on the top half and could be either green or yellow. Thus, each deck contained four types of cards: green dog, green duck, yellow dog, and yellow duck. Table 1 summarizes the respective frequencies of the types of cards for each deck.

Table 1: Number of cards in each deck used in Experiment 1.

Deck	Green Dog	Green Duck	Yellow Dog	Yellow Duck
1	5	4	6	5
2	9	2	6	3
3	4	8	2	6
4	7	8	2	3
5	3	3	6	8

## Procedure

There were two conditions in the experiment: *learn* and *suppose*, each with 20 participants. In both conditions, the five decks were presented to the participant in random order. For each deck, the experimenter first showed the cards to the participant, with the animals and colors in plain view. Cards were presented briefly (around 0.5 second apiece) to prevent counting. After all cards in the deck were presented, the participant shuffled the deck, drew one card at random, and put it on the table blank side up. Thus, neither the participant nor the experimenter knew what the card was.

The procedure then differed between the two conditions. In the *learn* condition, the experimenter covered the card drawn from the deck, turned the card over while still covered, and then revealed one half of the card to the participant. Whether the animal or the color was thereby revealed was randomly determined. If the revealed half was an animal then the participant estimated the probability that the unrevealed half was a certain color; whether they were asked for the probability of “green” versus “yellow” was determined randomly. If the revealed half was a color then the participant estimated the probability that the unrevealed half was a certain animal; whether they were asked for the probability of “dog” versus “duck” was determined randomly. The covered half was never revealed to the participant. This procedure was repeated for all five decks.

The *suppose* procedure was identical to the foregoing up to placing one card from the shuffled deck face down on the table. In the *suppose* procedure, neither side of the card was revealed, and the experimenter proceeded instead to ask a question of the form: “What is the probability that so-and-so appears on the card supposing that such-and-such appears?” The content of the question (so-and-so and such-and-such) was determined by yoking each *suppose* participant to the immediately preceding participant, who was in the *learn* con-

dition. Specifically, if for decks 1 through 5 the *learn* participant was asked for the probabilities of  $A_1 \dots A_5$  upon learning  $B_1 \dots B_5$  then the *suppose* participant was asked for the conditional probabilities  $Pr(A_1 | B_1) \dots Pr(A_5 | B_5)$  in the order corresponding to the presentation of the five decks to the *learn* participant.

Thus, the first participant was assigned to the *learn* condition, the second to the *suppose* condition (and yoked to the first participant), and likewise for succeeding pairs of participants. The crucial difference was that participants in the *learn* condition estimated  $A$  after learning  $B$ , whereas participants in the *suppose* condition estimated  $A$  while supposing  $B$ .

## Results and Discussion

We computed three statistics over the five probabilities that a given participant produced, namely, (a) the average of the five raw responses, (b) the average absolute deviation from 0.5, and (c) the average absolute deviation of a response from the objective probability of the event under consideration (where the objective probability was derived from the composition of the deck employed in that trial). Statistic (b) quantifies confidence inasmuch as extreme probabilities signify presumed knowledge about an event whereas 0.5 represents ignorance. The statistics produced by the two groups were then compared via paired  $t$ -tests. There were thus 20 pairs, defined by yoking each *suppose* participant to his/her *learn* participant.

As seen in row (a) of Table 2, the average responses across the 20 *learn*-*suppose* pairs were virtually identical [ $t(19) = 0.60$ ,  $p = 0.56$ ,  $d = 0.13$ ]. Row (b) shows, however, that the absolute deviation from 0.5 was reliably greater for the *suppose* group compared to *learn* group [ $t(19) = 2.61$ ,  $p = 0.02$ ,  $d = 0.58$ ]. The absolute deviation from objective probability also differed reliably between the *learn* and *suppose* conditions [ $t(19) = 4.15$ ,  $p < 0.001$ ,  $d = 0.93$ ] with more accurate responses from the *learn* participants; see row (c). Moreover, in 16 of the 20 pairs, *learn* participants were more accurate than *suppose* participants ( $p = 0.01$  by binomial test).

Table 2: Comparison of *learn* and *suppose* groups in Experiment 1

Statistic	Learn	Suppose	$p$
(a) Raw estimate of $Pr(A B)$	0.50(0.10)	0.49(0.12)	0.56
(b) Abs. dev. from 0.5	0.14(0.05)	0.19(0.06)	0.02
(c) Abs. dev. from $Pr(A B)$	0.09(0.05)	0.16(0.08)	0.00

Means for the two groups, relative to various statistics. Standard deviations are given in parentheses;  $p$ -values reflect paired  $t$ -tests ( $N = 20$ ). (Abs. dev. = Absolute deviation)

The results of Experiment 1 thus suggest limitations to the Bayesian model of updating, at the descriptive level. To check the robustness of our findings, we repeated Experiment 1 with new decks, involving different frequencies of the four events.

## Experiment 2

### Participants

A new group of forty undergraduates (26 female, mean age 19.5 yrs, SD = 1.8) from Princeton University participated in exchange for course credit.

### Materials and procedure

The procedure was identical to Experiment 1 except for the use of five different decks shown in Table 3.

Table 3: Number of cards in each deck used in Experiment 2.

Deck	Green Dog	Green Duck	Yellow Dog	Yellow Duck
1	9	2	1	8
2	2	8	9	1
3	7	1	3	9
4	3	8	7	2
5	8	3	2	7

### Results and Discussion

We computed the same statistics as in Experiment 1; see Table 4. As before, there was virtually no difference in the average responses of the suppose versus learn groups. And once again, suppose participants were less accurate than learn participants in terms of absolute deviation from the objective value; 16 of the twenty learn/suppose pairs showed this pattern ( $p = 0.01$  by binomial test). This time, however, learn rather than suppose participants issued more extreme probabilities; see row (b) of Table 4.

Table 4: Comparison of learn and suppose groups in Experiment 2

Statistic	Learn	Suppose	$p$
(a) Raw estimate of $Pr(A B)$	0.50(0.14)	0.48(0.10)	0.53
(b) Abs. dev. from 0.5	0.23(0.05)	0.18(0.06)	0.00
(c) Abs. dev. from $Pr(A B)$	0.13(0.04)	0.21(0.08)	0.00

Means for the two groups, relative to various statistics. Standard deviations are given in parentheses;  $p$ -values reflect paired  $t$ -tests ( $N = 20$ ). (Abs. dev. = Absolute deviation)

Why did suppose participants issue more extreme probabilities than learn participants in Experiment 1 while the reverse is true in Experiment 2? Tables 1 and 3 report the objective distributions of the cards in the two experiments, and reveal greater extremeness for the second experiment compared to the first. So, the switch in extremeness might be a corollary to the greater accuracy of the learn compared to suppose groups.

In sum, the results of Experiment 2 reveal once again a gap between learning and supposing that is not foreseen by Bayesian updating.

## Experiment 3

In Experiments 1 and 2, probabilities were grounded in frequencies and therefore *extensional*. The third experiment was

designed to evaluate the impact of learning versus supposing in an *intensional* setting involving probabilities of non-repeatable events. In particular, participants in the third experiment specified their confidence (as a probability) that Bill Clinton won/lost a specified state given that he won/lost another state in the 1992 presidential election.

### Participants

A new group of sixty undergraduates (41 female, mean age 20.4 yrs, SD = 1.9) from Princeton University participated in exchange for course credit.

### Materials

A deck of 50 cards served as stimuli. One side of a given card was marked with a U.S. state, the other side left blank.

### Procedure

As in the previous experiments, there was a learn and a suppose condition, each with 20 participants. In both conditions, the participant examined the deck then shuffled it and placed two cards face down on the table (without looking at them). Despite the appearance of randomness, the experimenter examined but then ignored the contents of the cards, and instead asked about two states from a pre-selected list. The list consisted of 20 swing states (electoral outcome not easily predictable); the two swing states figuring in a given trial were drawn randomly from the list.<sup>3</sup>

In the learn condition the experimenter picked up one of the two drawn cards and looked at its underside (preventing the participant from seeing the content). The state was announced (actually, the announced state was preselected from the list of 20 swing states), and then the electoral outcome for that state was determined by consulting a website. Specifically, with the participant watching, the experimenter discovered the outcome for that state via <http://uselectionatlas.org/RESULTS/>, and showed the result to the participant. Note that the participant was only shown whether Clinton won or lost the specified state; information about other states was masked. The experimenter then examined the underside of the remaining card, announced this second state (actually, preselected from the list of swing states), and asked the participant to estimate the probability of Clinton winning or losing that state. The framing of the question in terms of winning or losing was consistent with the outcome for the first state. For example, if Clinton won the first state then the participant estimated the probability of Clinton winning the second state, and likewise for losing. The two cards were then put aside, never revealed to the participant. This procedure was performed five times per participant.

In the suppose condition, each participant was yoked to the immediately preceding learn participant. To start the trial, the experimenter announced that it was a winning (or losing) round, meaning that the participants were to estimate the

<sup>3</sup>The swing states were taken to be AL, AZ, GA, ID, IN, KS, KY, LA, MI, MN, MO, MS, MT, NC, ND, NM, OH, TN, VA, WV.



probability that Clinton won (or lost) the second state, supposing that he won (or lost) the first state. The choice of framing (win/lose) appeared to be random, but in fact matched the questions in the learn condition. To finish the trial, the participant shuffled the deck and placed two cards face down on the table (without looking). The experimenter pretended to look at the undersides of the two cards and asked the participant to estimate the probability of Clinton winning (or losing) the second state supposing that he won (or lost) the first state. The two states were yoked to those in the learn condition. The two cards were then set aside, never revealed to the participant. This procedure was performed five times (yoked to the preceding learn participant).

A third group ( $N = 20$ ) served as a control condition in which just  $Pr(A)$  was estimated (no conditioning event  $B$  was evoked). In this condition, each participant was yoked to the preceding suppose and learn participants, and gave probabilities to the five states that were target events  $A$ . For each trial, the experimenter announced that it was a winning (or losing) round, meaning that the probability to be estimated was that Clinton won (or lost) the state in question. The framing was yoked to the questions in the suppose and learn conditions. The participant then shuffled the deck and placed one card face down on the table. The experimenter pretended to look at the card and asked the participant to estimate the probability of Clinton winning (or losing) the state. The procedure was performed for each of the five states.

## Results and Discussion

As seen in row (a) of Table 5, the average responses of the learn participants were reliably higher than those of the suppose participants [ $t(19) = 4.41$ ,  $p < 0.001$ ,  $d = 0.99$ ]. Row (b) shows that the learn group offered more extreme probabilities than the suppose group [ $t(19) = 3.11$ ,  $p < 0.01$ ,  $d = 0.69$ ].

Table 5: Comparison of learn and suppose groups in Experiment 3

Statistic	Learn	Suppose	$p$
(a) Raw estimate of $Pr(A B)$	0.64(0.09)	0.53(0.10)	0.00
(b) Abs. dev. from 0.5	0.21(0.06)	0.15(0.06)	0.00
(c) Quadratic penalty	0.18(0.08)	0.25(0.08)	0.02

Means for the two groups are presented, relative to various statistics. Standard deviations are given in parentheses;  $p$ -values reflect paired  $t$ -tests ( $N = 20$ ). (Abs. dev. = Absolute deviation)

To quantify the accuracy of the probability assigned to event  $A$  upon learning or supposing  $B$ , we computed the *quadratic penalty* for  $Pr(A)$ . To illustrate, the quadratic penalty for  $Pr(\text{wins Virginia})$  is  $(1 - Pr(\text{wins Virginia}))^2$  in the event that Clinton won Virginia, and it is  $(0 - Pr(\text{wins Virginia}))^2$  in case he lost. (Note that the conditioning event  $B$  was true in every case.) Thus, low penalty signifies accuracy of a stochastic forecast whereas high penalty signifies inaccuracy; assigning the noncommittal probabil-

ity 0.5 guarantees a penalty of 0.25, indicating ignorance. The quadratic penalty was introduced by Brier (1950); see Predd et al. (2009) for a justification of its use in measuring accuracy. For every participant, we computed her average quadratic penalty over the five trials. Row (c) of Table 5 shows that learners were closer to the truth than supposers were [ $t(19) = 2.57$ ,  $p = 0.02$ ,  $d = 0.58$ ]. This holds for 17 of the 20 pairs of participants ( $p = 0.01$  by binomial test). It is striking that the mean quadratic penalty for the suppose group is almost exactly 0.25, the accuracy level guaranteed by issuing 0.5 probabilities.

We note that a majority (60%) of the pairs figuring in the experiment had consistent outcomes in the election (Clinton winning both or losing both). For the learn condition, the average probabilities assigned to consistent and inconsistent pairs were 0.72 and 0.50, respectively, whereas they were 0.55 and 0.48 for the suppose condition. A two-way ANOVA reveals a reliable interaction, the difference between the learn probabilities exceeding that for the suppose probabilities [ $F(1, 19) = 17.7$ ,  $p < .001$ ]. Since a majority of pairs were consistent (as noted above), these facts explain the lower quadratic penalty for learn participants, and highlight their greater sensitivity to the conditioning event  $B$ .

Finally, in the control condition, the average raw estimate of  $Pr(A)$  was 0.51 ( $SD = 0.13$ ). This is close to the 0.53 estimate of  $Pr(A | B)$  in the suppose group [ $t(19) = 0.51$ ,  $p = 0.61$ ,  $d = 0.12$ ] but reliably different from the 0.64 estimate of the learn group [ $t(19) = 4.09$ ,  $p < 0.001$ ,  $d = 0.91$ ]. The quadratic penalty for the control condition was 0.29 ( $SD = 0.09$ ), reliably different from learn [ $t(19) = 4.18$ ,  $p < .001$ ,  $d = 0.94$ ] but not suppose [ $t(19) = 1.50$ ,  $p = 0.15$ ,  $d = 0.34$ ]. These results indicate once again that the conditioning event  $B$  had greater impact on the judgements of the learn participants compared to suppose.

## General Discussion

Bayesian updating (1) seems not to describe the relation between the probability distribution that arises from learning an event  $B$  compared to merely supposing it. For, in our three experiments, the probabilities that issue from learning  $B$  are more accurate than those resulting from conditioning, and they also differ in their deviation from 0.5. In Experiment 3, moreover, the average probabilities in the two groups differed significantly.

In the latter experiment, learn participants seem to have made greater use of the conditioning event  $B$  than did suppose participants. This is revealed by the greater difference in updated compared to prior probabilities for  $A$  in the learn compared to the suppose conditions. Specifically, learn estimates were reliably higher than the prior, suggesting that learn participants interpreted a win [loss] of one swing state to increase the chance of a win [loss] of another. In contrast, suppose participants' estimates of  $Pr(A | B)$  were almost identical to the control group's  $Pr(A)$ .

Insensitivity to  $B$  may reflect a deficit of imagination, the

suppose participants being unable to simulate the effect of genuinely believing *B*. In fact, Bayesian updating imposes a heavy burden on the reasoner's ability to foresee the impact of experience. Suppose that lions are discovered roaming your neighborhood; can you anticipate the probabilities you would attach to other events if such startling circumstances actually came to pass? Analogous difficulties arise when attempting to predict future affective states (Wilson & Gilbert, 2003).

At the normative level, the Bayesian doctrine (1) is supported by the considerations mentioned in the introduction, yet it remains contentious (see, e.g., Bacchus, Kyburg, and Thalos (1990)). Recent work has begun to provide a necessary critique of Bayesian inference (Jones & Love, 2011). The Bayesian doctrine may also prove to be unsuited to situations in which the agent, albeit rational, loses track of her position in time or space (Arntzenius, 2003). But the debate about (1) might be of limited relevance to the typical transition from one probability distribution to another. Such transitions need not depend on adding an event *B* to one's beliefs without probabilistic qualification. Rather, experience might lead us to revise our confidence in *B* without driving it to zero or one. The rule proposed by Jeffrey (1983) is suited to this kind of case. Recent work has begun to examine Jeffrey's rule from the psychological point of view (Over & Hadjichristidis, 2009; Zhao & Osherson, 2010).

### Acknowledgments

Osherson acknowledges support from the Henry Luce Foundation.

### References

- Arntzenius, F. (2003). Some problems for conditionalization and reflection. *The Journal of Philosophy*, 356-370.
- Bacchus, F., Kyburg, H., & Thalos, M. (1990). Against conditionalization. *Synthese*, 85, 475-506.
- Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1-3.
- Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge MA: MIT Press.
- Hacking, I. (2001). *An Introduction to Probability and Inductive Logic*. Cambridge UK: Cambridge University Press.
- Harman, G. (1999). *Reasoning, meaning and mind*. Oxford UK: Oxford University Press.
- Jeffrey, R. C. (1983). *The Logic of Decision (2nd Edition)*. Chicago IL: The University of Chicago Press.
- Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, 34, 169-231.
- Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.
- Over, D., & Hadjichristidis, C. (2009). Uncertain premises and jeffrey's rule. *Behavioral and Brain Sciences*, 32, 97 - 98.
- Predd, J., Seiringer, R., Lieb, E. H., Osherson, D., Poor, V., & Kulkarni, S. (2009). Probabilistic coherence and proper scoring rules. *IEEE Transactions on Information Theory*, 55(10), 4786 - 4792.
- Wilson, T., & Gilbert, D. (2003). Advances in experimental social psychology. In M. P. Zanna (Ed.), (Vol. 35, p. 345-411). Academic Press.
- Zhao, J., & Osherson, D. (2010). Updating beliefs in light of uncertain evidence: Descriptive assessment of Jeffrey's rule. *Thinking & Reasoning*, 16, 288-307.
- Zhao, J., Shah, A., & Osherson, D. (2009). On the provenance of judgments of conditional probability. *Cognition*, 113(1), 26 - 36.

# Humor, Emotions and Communication: Human-like Issues of Human-Computer Interactions

**Pawel Dybala (paweldybala@res.otaru-uc.ac.jp)**

JSPS Research Fellow / Otaru University of Commerce, Midori 3-5-21, 047-8501 Otaru, Japan

**Kohichi Sayama (sayama@res.otaru-uc.ac.jp)**

Otaru University of Commerce, Department of Information and Management Science,  
Midori 3-5-21, 047-8501 Otaru, Japan

## Introduction

The research described below is unique within both cognitive science and computer science (especially in AI and HCI-related variants) as, to the best of our knowledge, it is the first to combine the issues of humor generation and emotion recognition in an interaction between humans and computers. We present the first conversational system which recognizes user emotions and on this basis decides whether or not to use humor in response.

## Humor and Emotions

This research is based on an assumption that what can work well in interactions between humans can also be beneficial for HCI. It was demonstrated that humor can induce positive moods and reduce negative moods in humans; see Dybala (2011); Dybala et al. (2010c) for a summary. This leads to the conclusion that it should perform the same role in HCI. This was shown in existing research, however, only to some extent. While non-computer-science related research usually considers the humor-emotion relationship in what we defined as three-stage chains, i.e. emotive state I → humorous stimulus → emotive state II, in HCI, all research so far focuses only on doublets, i.e. humorous stimulus → emotive reaction (for more details regarding this approach see Dybala et al., 2010a). Our research bridges this gap and is the first in HCI in which not only emotions following but also those preceding humorous acts were taken into consideration (Dybala, 2011).

## Components

The system developed in our research is named MAS-Punda. Punda is the name of our pun generator (Dybala et al., 2008), and MAS stands for “multiagent system”, as it is composed of multiple intelligent agents: 1) a conversational agent, 2) an emotion recognition agent and 3) a humor generator.

### Conversational Agent

The first agent in our research is Maru-chan, a conversational agent for Japanese (Takahashi, 2009). The agent uses the Internet to extract word associations for users' utterances, and then uses them to generate a relevant response (Dybala et al., 2010d). For more details, see Dybala, Ptaszynski & Sayama (2012a), Dybala (2011).

### Emotion Recognition Agent

The second agent used in our research is Ptaszynski's et al. Emotion Recognition Agent (2010). Besides performing the decisive role in the multi-humoroid, it is also used in automatic evaluation of the system. The agent uses databases with emotive expressions and emotive elements for Japanese. It performs two types of text analysis: 1) it detects if a sentence is emotive, and 2) it determines the type and valence of the conveyed emotion(s). For more details, see Dybala (2011) and Dybala et al. (2010c).

### Humor Generation Agent

The second agent used in this research is Pundalin, a pun generator for Japanese (Dybala et al., 2008), which also uses the Internet. This agent represents the class of “humoroids”, defined by Dybala et al. (2009d). Against a user utterance, the agent generates phonetic candidates for puns, selects the most appropriate, and then integrates it into an adequate humorous response. In the generation step, the agent uses phonetic patterns based on an innovative classification of Japanese puns, proposed by Dybala (2011). For more details about this agent, see (Dybala, 2011; Dybala et al., 2010c).

### Multiagent Emotion Aware Joking System

The agents described above were merged to create a multiagent humor-equipped joking conversational system. It represents a high level class of humoroids, defined and named by Dybala et al. (2010d) as “multi-humoroids”. Basing on existing literature (see above), we assumed that in HCI, the computer agent can use humor to make the interlocutor feel better. In order to do that, first the system detects human emotions (performed by ML-Ask), and on this basis makes a decision whether a joke should be told.

To summarize the decision rules used in this experiment, the assumptions were that: 1) if a human's emotive state is negative or neutral, the agent can use humor in order to make him / her feel better, and 2) otherwise, the response is generated by the non-humorous agent. If ML-Ask decides that a joke should be told, the response is generated by the joke generator. If ML-Ask decides otherwise, the response is generated by the Maru-chan (Dybala et al., 2010c).

### Evaluation Experiments

The multiagent system was evaluated using a novel chatterbot evaluation methodology proposed by Dybala et al.

(2010b). In this particular case we conducted two experiments: user-oriented and automatic.

### User-oriented evaluation

In the first experiment we asked 13 users to perform conversations with Maru-chan and MAS-Punda in order to compare their performance in a questionnaire completed afterwards. Some results are summarized in Table 1.

Table 1. User-oriented evaluation experiment results

A) Did the agent try to make the conversation more interesting?, B) Did you find the conversation interesting?, C) Did the agent try to make you feel better?, D) Did the agent use humor in appropriate moments?, E) Describe your feelings towards the agent after the interaction (sum), F) Which agent was better?, cont? – do you want to continue the dialogue? (option to choose)

Question	A	B	C	D	E	F	cont?
Maru	1.69	2.08	1.69	1.00	-9	38%	2
MAS	2.85	2.69	2.69	2.45	+8	62%	5

### Automatic evaluation

In the second experiment, the chat logs acquired in the first experiment were analyzed by the ML-Ask agent to investigate user emotions and their changes during the conversations. Some results are summarized in Table 2.

Table 2. Automatic evaluation experiment results

	Maru-chan			MAS-Punda		
Emotiveness	91 (average: 7.0 per utt.)			125 (average 9.6 per utt.)		
Valence	to positive		to negative	to positive		to negative
	68%		32%	94%		6%
Final emotion	pos.	neg.	neutr.	pos.	neg.	neutr.
	69%	31%	0%	85%	0%	15%

### Summary and Future Directions

The experiments showed that the humor- and emotion-equipped multiagent system was evaluated as better, more interesting and, perhaps most importantly, making users feel better than the baseline agent. This leads to the conclusion that our goal to construct a conversational system able to properly react with humor to users' emotions was achieved.

That said, there is still much to be done in this area. Currently we are working on a user-adaptive humor sense model, which should lead to personalization of the system (Dybala et al., 2009b). We also plan to construct a metaphor processing agent and implementing it into our system.

### Acknowledgements

This work was supported by KAKENHI (Project Number: 23-01348).

### References

Dybala, P., Ptaszynski, M., & Sayama, K. (2012a). Reducing Excessive Amounts of Data: Multiple Web

- Queries for Generation of Pun Candidates. *Advances in Artificial Intelligence*, Hindawi Publishing (to appear).
- Dybala, P., Ptaszynski, M., & Sayama, K. (2012b). A Step Towards Joking AI: Multiagent Humor-Equipped Conversational System. Chapter in *Agent-Based Approaches to Ambient Intelligence*, IOS Press (to appear).
- Dybala, P. (2011). *Humor to Facilitate HCI: Implementing a Japanese Pun Generator into a Non-task Oriented Conversational System*, Lambert Academic Publishing.
- Dybala, P., Ptaszynski, M., Rzepka, R., & Araki, K. (2010a). Extending the Chain: Humor and Emotions in Human Computer Interaction. *International Journal of Computational Linguistics Research*, Vol. 1, No. 3, 116-128, Digital Information Research Foundation.
- Dybala, P., Ptaszynski, M., Rzepka, R., & Araki, K. (2010b). Evaluating subjective aspects of HCI on an example of a non-task oriented conversational system, *Int. J. Artif. Intell. T., Special Issue on AI Tools for HCI Modeling 19*(6), 819-856.
- Dybala, P., Ptaszynski, M., Maciejewski, J., Takahashi, M., Rzepka, R., & Araki, K. (2010c). Multiagent system for joke generation: Humor and emotions combined in human-agent conversation. *Journal of Ambient Intelligence and Smart Environments* 2, 31-48.
- Dybala, P., Ptaszynski, M., Rzepka, R., & Araki, K. (2010d). Multi-humoroid: Joking System That Reacts With Humor To Humans' Bad Moods. *Proc. of AAMAS 2010* (pp. 1433-1434), Toronto, Canada.
- Dybala, P., Ptaszynski, M., Rzepka, R., & Araki, K. (2009a). Activating Humans with Humor - A Dialogue System that Users Want to Interact With. *IEICE T. Inf. Syst., Special Issue on Natural Language Processing and its Applications*, Vol.E92-D, No.12, 2394-2401.
- Dybala, P., Ptaszynski, M., Rzepka, R., & Araki, K. (2009b). Humorized Computational Intelligence - towards User-Adapted Systems with a Sense of Humor. *LNCS*, Vol. 5484, 452-461, Springer Berlin & Heidelberg.
- Dybala, P., Ptaszynski, M., Rzepka, R., & Araki, K. (2009c). Subjective, but not worthless - Non-linguistic features of chatterbot evaluations. *Proc. of KRPD-09, IJCAI-09* (pp. 87-92), Pasadena, California, USA.
- Dybala, P., Ptaszynski, M., Rzepka, R., & Araki, K. (2009d). Humoroids - Talking Agents That Induce Positive Emotions with Humor. *Proc. of AAMAS'09* (pp. 1171-1172), Budapest, Hungary.
- Dybala, P., Ptaszynski, M., Higuchi, S., Rzepka, R., & Araki, K. (2008). Humor Prevails! - Implementing a Joke Generator into a Conversational System. *LNAI*, Vol. 5360, 214-225, Springer Berlin & Heidelberg.
- Ptaszynski, M., Dybala, P., Rzepka, R., & Araki, K. (2010). An Automatic Evaluation Method for Conversational Agents Based on Affect-as-Information Theory. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Special Issue on Emotions*, Vol. 22, No. 1, 73-89.
- Takahashi, M. (2009). *Web ni yoru kyoukihindo to n-gram moderu wo mochiita hatsuwabunnnseishuhou (Utterance Generation Method Using Web Search and N-grams)*. Bachelor diss., Hokkaido Univ., Sapporo, Japan, 2009.

# Meta-Representational Competence as an Aspect of Spatial Intelligence

Mary Hegarty (hegarty@psych.ucsb.edu)

Department of Psychology, University of California, Santa Barbara  
Santa Barbara, CA 93106 USA

## Abstract

Meta-representational competence (diSessa, 2004) encompasses ability to choose the optimal external representation for a task and to use novel external representations productively. Research on this aspect of spatial intelligence reveals large individual differences in ability to adaptively choose and use external visual-spatial representations for a task. This research suggests that we should not just think of interactive external visualizations as ways of augmenting spatial intelligence, but also consider the types of intelligence that are required for their use.

**Keywords:** Spatial ability; intelligence; external representations.

## Introduction

The concept of spatial intelligence brings to mind measures of individual differences in spatial ability, which measure performance in tasks (e.g., mental rotation, paper folding) involving the storage and manipulation of internal spatial representations. While this is certainly a central aspect of spatial intelligence, with developments in information technologies, spatial thinking increasingly involves interacting with external visualizations. For example, medical students now learn three dimensional (3-D) anatomy by interacting with computer visualizations that they can rotate at will, scientists gain insight into their data by visualizing and interacting with multidimensional plots, and chemists use a variety of external representations, including 3-D physical and virtual models and 2-D diagrams, to reason about the structures and reactive properties of molecules. While new technologies are typically seen as means of augmenting human intelligence, this paper considers the ways in which they also make new demands on our intelligence.

diSessa (2004) coined the term *meta-representational competence* to encompass ability to choose the optimal external representation for a task, use novel external representations productively, and invent new representations as necessary. This competence goes beyond the capacity to understand the conventions of a particular type of representation (such as a graph, map, or diagram). It is a form of metacognition about visual-spatial displays. diSessa was mostly concerned with children's native representational competence in using and inventing scientific representations, and how to foster this in instruction. In this paper, I consider meta-representational competence as a component of adult spatial intelligence.

## Using Representations

In one line of research inspired by the use of new technologies in medicine, I and my colleagues have found large individual differences in use of interactive visualizations (Cohen & Hegarty, 2007; Keehner, Hegarty, Cohen, Khooshabeh & Montello 2008; Stull, Hegarty & Mayer, 2009). In a series of studies, participants were given the opportunity to manipulate a virtual 3-D object while accomplishing a spatial task, namely inferring the appearance of a cross section of the 3-D object. On each cross section trial, participants were shown a picture of the 3-D object with a line drawn through it, and an arrow pointing to the line and their task was to draw the cross section that would result if the object were sliced at the line and one was viewing the cross section from the direction of the arrow.

The most common use of the virtual object was to rotate it to the view of the object that one would see if one was viewing it from the perspective of the arrow. Rotating the external visualization in this way relieves the participant of the need to mentally rotate the object or mentally change his or her perspective with respect to the object. Participants who used the model in this way had better performance, but many students did not use the model. In general, there were large individual differences in ability to discover how to best use the virtual object, ability to actually manipulate it, and ability to benefit from the most task-relevant view of the object.

We are finding similar results in current studies on use of 3-D models in chemistry (Stull, Hegarty, Dixon & Stieff, submitted). In these studies the task (for chemistry students) is to translate between different diagrams of molecules that use different conventions to depict the 3-D structure of molecules in the two dimensions of the printed page, and depict the molecule from different spatial perspectives. On each trial participants are given one diagram of a molecule and their task is to draw a different diagram of the same molecule. The most common use of the model is to first match it to the perspective of the given diagram, then rotate it to match the perspective of the diagram to be drawn, and then draw this diagram. However, when they are provided with a concrete 3-D model of the molecule, many students do not use the model and perform poorly on the representation translation task. These studies indicate that ability to use an external representation is not always a given, and provide evidence for individual differences in adult meta-representational competence.

## Choosing Representations

Another aspect of meta-representational competence is choosing the most effective representation for a given task. We have examined this aspect in the domain of meteorology. The design of a weather map, such as its complexity or the relative salience of different depicted variables, can have significant effects on performance of map comprehension tasks (Canham & Hegarty, 2010; Fabrikant, Rebich-Hespanha & Hegarty, 2010; Hegarty, Canham & Fabrikant, 2010). However, when given a choice of displays, people do not always choose the optimal display for their task. In a series of studies we asked naïve undergraduate students and experienced weather forecasters to perform read-off and inference tasks with maps that varied in complexity (the number of displayed variables) and realism (the extent to which they looked like their real-world referents). We also asked our participants to choose (from a set of maps varying in complexity and realism) the maps that they would prefer to use when accomplishing these tasks.

Both naïve students and experienced forecasters performed the tasks more efficiently with simple maps that displayed only the task-relevant information and naïve students were more accurate with these displays. When asked to choose which map they would prefer to use, or the map with which they would be most efficient, the majority also chose these relatively simple maps. However about one third of naïve students and expert weather forecasters chose more realistic and complex weather maps than they needed, even though complexity and realism impaired performance (Hegarty, Stull & Smallman, in press; Hegarty, Smallman, Stull & Canham, 2009). More generally, surveys of undergraduate students indicate that they prefer realistic, animated, and detailed displays over more abstract, static, and sparse displays for a range of tasks (Hegarty, 2010; Hegarty et al., 2009).

## Conclusions

It is important to consider the broader context of these research findings. For example, efficiency may not be an individual's paramount goal when he or she chooses a visual display. There is also a tradeoff between the time required to find or design the most efficient display and the time saved by using that display. Finally, complex displays that are less efficient in the short term may lead to a more elaborated understanding of the situation represented by the display

Nevertheless, these studies highlight the fact that although new visualization technologies have the potential to augment human spatial intelligence, intelligence is also required for their use. With the current interest on how to foster spatial intelligence, (National Research Council, 2006) and increased availability of complex visualizations, they suggest that more attention should be paid to teaching people to use, design, and critique external spatial representations, in addition to training their internal visualization abilities.

## Acknowledgments

This research was funded by National Science Foundation Grants 0313237 and 1008650 and grant N000140610163 from the Office of Naval Research.

## References

- Canham, M. & Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction*, 20, 155-166.
- Cohen, C. A. & Hegarty, M. (2007). Individual differences in use of an external visualization while performing an internal visualization task. *Applied Cognitive Psychology*, 21, 701-711.
- diSessa, A. A. (2004). Metarepresentation: Native Competence and Targets for Instruction. *Cognition and Instruction*, 22, 293-331.
- Fabrikant, S. I., Rebich-Hespanha, S., & Hegarty, M. (2010). Cognitively inspired and perceptually salient graphic displays for efficient inference making. *Annals of the Association of American Geographers*, 100, 13-29
- Hegarty, M. (2010). Components of spatial intelligence. In B. H. Ross (Ed.) *The Psychology of Learning and Motivation*. San Diego: Academic Press (pp. 265-297).
- Hegarty, M., Canham, M. & Fabrikant, S. I. (2010). Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 36, 37-53.
- Hegarty, M. Smallman, H. S., & Stull, A. T. (2012) Choosing and using geospatial displays: Effects of design on performance and metacognition. *Journal of Experimental Psychology: Applied*, 18, 1-17.
- Hegarty, M. Smallman, H. S., Stull, A. T. & Canham, M. (2009). Naïve Cartography: How intuitions about display configuration can hurt performance. *Cartographica*, 44, 171-187.
- Keehner, M. Hegarty, M., Cohen, C. A., Khooshabeh, P. & Montello, D. R. (2008). Spatial reasoning with external visualizations: What matters is what you see, not whether you interact. *Cognitive Science*, 32, 1099-1132.
- National Research Council (2006). *Learning to think spatially: GIS as a support system in the K-12 curriculum*. Washington, DC: National Research Council Press.
- Stull, A. T. Hegarty, M., Dixon, B., Stieff, M. (submitted) Representational translation with concrete models.
- Stull, A. T., Hegarty, M. & Mayer, R. E. (2009). Orientation references: Getting a handle on spatial learning. *Journal of Educational Psychology*, 101, 803-816.

# Rationality-Guided AGI as Cognitive Systems

Ahmed Abdel-Fattah, Tarek R. Besold, Helmar Gust,  
Ulf Krumnack, Martin Schmidt, Kai-Uwe Kühnberger

({ahabelfatta | tbesold | hgust | krumnack | martisch | kkuehnbe}@uni-osnabrueck.de)

Institute of Cognitive Science, University of Osnabrück,  
Albrechtstr. 28, 49076 Osnabrück, Germany

Pei Wang

(pei.wang@temple.edu)

Department of Computer and Information Sciences, College of Science & Technology, Temple University,  
1805 N. Broad Street, Philadelphia, PA 19122 USA

## Abstract

The integration of artificial intelligence (AI) within cognitive science (CogSci) necessitates further elaborations on, and modelings of, several indispensable cognitive criteria. We approach this issue by emphasizing the close relation between artificial general intelligence (AGI) and CogSci, and discussing, particularly, “rationality” as one of such indispensable criteria. We give arguments evincing that normative models of human-like rationality are vital in AGI systems, where the treatment of deviations from traditional rationality models is also necessary. After conceptually addressing our rationality-guided approach, two case-study systems, NARS and HDTP, are discussed, explaining how the allegedly “irrational” behaviors can be treated within the respective frameworks.

**Keywords:** Rationality; intelligence; AGI; HDTP; NARS

## Motivations and Background

For more than five decades, artificial intelligence (AI) has always been a promising field of research on modeling human intelligence. The success of projects like IBM’s Watson (Ferrucci et al., 2010), for instance, increases the hopes in achieving not only language intelligence but also inference mechanisms at a human-level and paves the way for solving more baffling tasks. However, AI has turned into a vague, un-specific term, in particular because of the tremendous number of applications that belong, in fact, to seemingly orthogonal directions. Philosophers, psychologists, anthropologists, computer scientists, linguists or even science fiction writers have disparate ideas as to what AI is (or should be). The challenge becomes more obvious when AI is looked at from a CogSci perspective, where the focus is mainly on explaining processes of general cognitive mechanisms (not only on how one or another intelligence task can be solved by a computer). We think that from a CogSci perspective the kind of intelligence characterizing classical AI problems is not yet exhaustive enough. Solutions to most of the problems are not cognitively inspired: neither do they consider essential cognitive mechanisms (or general intelligence results) nor do they show the biological plausibility of the solutions.

*Artificial General Intelligence* (AGI) refers to a research direction that takes AI back to its original goals of confronting the more difficult issues of human-level intelligence as a whole. Current AGI research explores all available paths, including theoretical and experimental computer science, cognitive science, neuroscience, and innovative interdisciplinary

methodologies (Baum, Hutter, & Kitzelmann, 2010). Here, we approach cognition in AGI systems by particularly promoting “rationality” as one of such indispensable criteria, and analyze some divergent, sometimes seemingly irrational, behaviors of humans.

In this article, our goal is twofold. We first concern ourselves with explicitly allocating ideas from AGI within CogSci. Second, we give a conceptual account on some principles in normative rationality-guided approaches. After explaining our approach at a general level, we explain how two cognitively inspired systems, namely NARS and HDTP, have the potential to handle (ir)rationality. We conclude by giving some remarks and future speculations.

## Why AGI?

In current AGI research, there are approaches following different paths, including those (1) inspired by the structure of human brain or the behavior of human mind, (2) driven by practical demands in problem solving, or (3) guided by *rational principles* in information processing. We are concerned with the latter approach, which has at least three essential advantages. One advantage of the rationality-guided approach, from an AGI perspective, is that it is less bound to exactly reproducing human faculties on a functional level. Another advantage is that it gives AI the possibility of being established in a way similar to other disciplines, where it can give a theoretical explanation to intelligence as a process that can be realized both in biological systems and computational devices. The third advantage of the rationality-guided approach is that it is not limited to a specific domain or problem.

## Rationality

The term *rationality* is used in a variety of ways in various disciplines. In CogSci, rationality usually refers to a way a cognitive agent deliberatively (and attentively) behaves in, according to a specific normative theory. The prototypical instance of cognitive agents that can show rational behavior is humans, who so far are also the ultimate exemplar of generally intelligent agents. When modeling intelligence, it is reasonable to initially take the remarkable abilities of humans into account with respect to rational behavior, but also their apparent deficiencies that show up in certain tasks.



Surprisingly little attention has been paid so far in AI towards a theory of rationality. A reason might be that the concept of rationality was too broad in order to be of interest to AI, where for a long time usually relatively specific cognitive abilities were modeled and heuristics were suggested. Moreover, an artificial cognitive agent is usually intended to reproduce rational behavior, not to act in seemingly irrational ways. Consequently, AI researchers are not interested in results of some *classical rationality puzzles*. Still, we think that a move towards integrating AGI in CogSci cannot ignore rationality issues, neither the remarkable abilities nor the originalities human subjects show in rationality tasks.

### Traditional Models of Rationality

Different models of rationality use significantly different methodologies. Clustering such models according to the underlying formalism usually results in at least the following four classes: (1) logic-based models (Evans, 2002), (2) probability-based models (Griffiths, Kemp, & Tenenbaum, 2008), (3) heuristic-based models (Gigerenzer, 2008), and (4) game-theoretically based models (Osborne & Rubinstein, 1994). Several of these models have been proposed for establishing a *normative theory of rationality*, normally by judging a belief as rational if it has been obtained by a formally correct application of the respective reasoning mechanism, given some background beliefs or knowledge (cf. e.g. also (Gust et al., 2011; Wang, 2011)). Therefore, such theories of rationality are not only intended to model “rational behavior” of humans, but to postdictively decide whether a particular belief, action, or behavior is rational or not. Nonetheless, although a conceptual clarification of rational belief and rational behavior is without any doubts desirable, it is strongly questionable whether the large number of different (quite often orthogonal) frameworks makes this task easier, or if the creation of a more unified approach wouldn’t be recommendable. From our perspective, basic cognitive mechanisms seem to offer a basis for such an endeavor.

### Some Rationality Challenges and Puzzles

Although the models mentioned above have been proven to be quite successful in modeling certain aspects of intelligence, all four types of models have been challenged. For example, in the famous Wason selection task (Wason & Shapiro, 1971) human subjects fail at a seemingly simple logical task (cf. Table 1.a). Similarly, Tversky and Kahneman’s Linda problem (Tversky & Kahneman, 1983) illustrates a striking violation of the rules of probability theory in a seemingly simple reasoning problem (cf. Table 1.b). Heuristic approaches to judgment and reasoning try to stay closer to the observed behavior and its deviation from rational standards (Gigerenzer, 2008), but they fail in having the formal transparency and clarity of logic-based or probability-based frameworks with regard to giving a rational explanation of behavior. Game-based frameworks can be questioned due to the various forms of optimality concepts in game-theory that can support different “rational behaviors” for one and the same situation.

In order to make such challenges of rationality theories more precise, we discuss some aspects of the famous Wason selection task and the Linda problem in more detail.

**Wason Selection Task** This task shows that a large majority of subjects are seemingly unable to evaluate the truth of a simple rule of the form “*if p then q*” (Wason & Shapiro, 1971). In the version depicted in Table 1.a, this rule is represented by: “*If on one side of the card there is a D, then on the other there is the number 3*”. According to classical logic, in order to assign a truth-value to this rule, subjects need to turn D and 7. What is interesting is the fact that a slight modification of the content of the rule to a setting more familiar from daily life, while keeping the structure of the problem isomorphic, makes subjects perform significantly better, as e.g. shown in (Cosmides & Tooby, 1993).

Table 1: a. A description of the Wason selection task. b. An abbreviated version of the Linda problem setting.

<b>a. Wason Selection Task (Wason &amp; Shapiro, 1971):</b>
Every card which has a D on one side has a 3 on the other side (and knowledge that each card has a letter on one side and a number on the other side), together with four cards showing respectively D, K, 3, 7, hardly any individuals make the correct choice of cards to turn over (D and 7) in order to determine the truth of the sentence. This problem is called “selection task” and the conditional sentence is called “the rule”.
<b>b. Linda Problem (Tversky &amp; Kahneman, 1983):</b>
Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
(F): Linda is active in the feminist movement.
(T): Linda is a bank teller.
(T&F): Linda is a bank teller and is active in the feminist movement.

**Linda Problem** With respect to the Linda problem (Tversky & Kahneman, 1983) it seems to be the case that subjects have problems to prevent the so-called *conjunction fallacy*: subjects are told a story specifying a particular profile about someone called Linda. Then, some statements about Linda are shown and subjects are asked to order them according to their probability (cf. Table 1.b). 85% of subjects decide to rank the statements “*Linda is a bank teller and is active in the feminist movement*” (T & F) as more probable than the statement “*Linda is a bank teller*” (T). This ranking conflicts with the laws of probability theory, because the probability of two events (T & F) is less than or at most equal to the probability of one of the events (e.g. (T)).

### Classical Resolution Strategies of Irrationality

Many strategies have been proposed to address the mentioned challenges, ranging from the use of non-classical logics to model subjects’ behavior in the Wason selection task (Stenning & van Lambalgen, 2008), to considerations involv-

ing reasoning in semantic models instead of (syntactic) deductions (Johnson-Laird, 1988) in the case of the Wason selection task. With respect to the Linda problem it has been argued that pure probability theory is not appropriate for addressing the problem properly, but a foundation of the analysis of this problem in coherence theories would be necessary (Pfeifer, 2008). Another resolution strategy applicable to both puzzles is to question whether tasks were appropriately phrased in the respective experiments. In the Wason selection task the “if-then” rule presented in natural language is usually not equivalent to its interpretation in classical logic, and in the Linda puzzle the term “probable” can be interpreted differently by the subjects (Gigerenzer, 2005). In any case, although there are many proposals to address the challenges, there is no generally accepted rationality concept available yet. Moreover, specific frameworks can address specific challenges, but do not generalize to the breadth of the mentioned problems.

For a generally intelligent cognitive system a question that can be raised is: *which principles of rationality can be transferred to and modeled in AGI systems, in order to achieve intelligence on a human scale?* We will argue for models that link rationality to the ability of humans to establish analogical relations (continuing a line of reasoning started in (Besold et al., 2011)), and to the ability to adapt to the environment by making good use of previously obtained experiences.

### Non-Standard, CogSci-Based Approaches

The two examples discussed above definitely show that humans have sometimes problems to apply rules of classical logic correctly (at least in rather abstract and artificial situations), and to reason according to the Kolmogorov axioms of probability theory. Nonetheless, the most that can be concluded from the experiments is that human agents are neither classical deduction machines nor probability estimators, but perform their indisputable reasoning capabilities by other means, necessarily linked to their cognitive capacities.

#### Resolving the Selection Task by Cognitive Mechanisms

As mentioned above, subjects perform better (in the sense of more according to the laws of classical logic) in the Wason selection task, if content-change makes the task easier to access for subjects. We think that the performance of subjects has a lot to do with the ability of subjects to establish appropriate analogies. Subjects perform badly in the classical version of the Wason selection task, probably because they fail to establish a correct analogy. Therefore, subjects fall back to other (less reliable) strategies to solve the problem. In a content-change version of the task the situation is different, because subjects can do what they would do in an everyday analogous situation. In short, the success or failure of managing the task is crucially dependent on the possibility to establish a meaningful analogy.

Another related resolution is to study the mode of the inference that should underly a normative theory of rationality. When a system has sufficient knowledge and resources

(with respect to the problems to be solved), an axiomatic logic (such as classical logic) can be used, which treats the available knowledge as axioms, and derives theorems from them to solve a given problem. When the system has insufficient knowledge, it has no absolute truth to be used as axioms, so has to follow some “non-axiomatic” logic, whose premises and conclusions are all revisable by new evidence. In Wason’s task, the expected results are the ones assuming an axiomatic system, while the actual results may be consistent with a non-axiomatic one. Therefore, the “mistake” here is mainly the misunderstanding between the psychologists who run the tests and the subjects who take the tests. In this artificially structured experiment, it is valid for the psychologists to assume sufficient knowledge and resources, therefore to expect the application of an axiomatic type of inference mechanism. Their mistake, however, is the failure to see the result as coming from another type of inference. On the side of subjects, since non-axiomatic reasoning is used more often in everyday life, most of them fail to understand the experiment setting as a testing of their capacity of using an axiomatic inference mechanism. This explains why many subjects admit their mistake afterwards, and do better in the content-change task (as soon as they realized that the expected way of reasoning is not their default one, they have less problem to adapt to follow it).

#### Resolving the Linda Problem by Cognitive Mechanisms

Here, a natural explanation of subjects’ behavior is that there is a lower degree of coherence of Linda’s profile plus the statement “*Linda is a bank teller*” in comparison to the degree of coherence of Linda’s profile plus the statement “*Linda is a bank teller and is active in the feminist movement*”, as in the conjunctive statement, at least one conjunct of the statement fits quite well to Linda’s profile. *Coherence* (Thagard, 2002) is a complicated concept that needs to be discussed in more detail (as does its connection to notions like the idea of representativeness proposed as an explanation for the Linda problem by Tversky and Kahneman themselves), but it can be mentioned that coherence is important for the successful establishment of an analogical relation, as well as for guiding adaptation of obtained knowledge and experiences. In order to make sense out of the task, subjects tend to rate statements with a higher probability where facts are arranged in a theory with a higher degree of coherence. Also, this can be thought of as a form of coherently adapting beliefs, which also depends heavily on subjects’ experiences rather than on their knowledge of Kolmogorov axioms of probability theory.

### Modeling Rationality: Case Studies

Formal and computational models in CogSci can be roughly divided into two major types: *descriptive* and *normative*. A descriptive model explains how a system actually works, and its establishment is based on empirical data. A descriptive model’s quality is evaluated according to its behavior’s *similarity* to that of humans. A normative model, on the other hand, specifies how a system should work, and its estab-

lishment is based on certain general principles or postulates. Such a normative model's quality is evaluated according to its behavior's *coherence* with these basic assumptions. Though the two types of models are closely related, they are still built and evaluated differently (Wang, 2011). When building a model of rationality, a central issue is the selection of the assumptions on which the model is based, since all conclusions about the model are derived from, and justified against, these assumptions.

In the following, we give two examples for cognitively inspired systems: NARS and HDTP. Both stand in a certain tradition to classical cognitive architectures like the well-known models ACT-R (Anderson & Lebiere, 1998) and SOAR (Laird, Newell, & Rosenbloom, 1987), because they attempt to model cognition in breadth and not relative to highly specialized abilities. Nevertheless, because NARS and HDTP stand in a tradition of modeling the competence aspect of general intelligence, they attempt to integrate a bunch of different human-inspired reasoning abilities, and they try to integrate these abilities in uniform models, they also differ significantly from the mentioned classical cognitive architectures. We briefly introduce NARS and HDTP and discuss how they can account for "irrational" behaviors in tasks, such as the Selection Task and the Linda problem.

**AGI with Relative Rationality (NARS)** NARS (Non-Axiomatic Reasoning System) is an AGI system designed under the assumption that the system usually has insufficient knowledge and resources with respect to the problems to be solved, and must adapt to its environment. Therefore, the system realizes a "relative rationality", that is, the solutions are the best the system can get *under the current knowledge-resource restriction* (Wang, 2011). Since this system has been described in a book (Wang, 2006) and many papers (most of which are available at the the last author's website<sup>1</sup>), here we only briefly explain the treatment of the "Selection Task" and "Conjunction Fallacy" in NARS.

Since NARS has insufficient knowledge and resources, its beliefs are not "absolute truth" but summary of the system's experience. Especially, the *truth-value* of a statement measures its *evidential support*, and the evidence can be either *positive* or *negative*, depending on whether the evidence agrees with the statement. Concretely, for statement "*If on one side of the card there is a D, then on the other there is the number 3*", the D card always provides evidence (positive if the other side is 3, otherwise negative); the 3 card may provide positive evidence (if the other side is D); the 7 card may provide negative evidence (if the other side is D); the K card provides no evidence. To determine the truth-value of the statement, all cards except K should be checked, but due to insufficient resources, the system may fail to recognize all evidence. In this case, D is the easiest, while 7 the hardest. This result is consistent with the common responses of human beings. It is labeled as "irrational", because in classical logic

the truth-value of a statement only depends on the existence of *negative* evidence, and whether there is *positive* evidence does not matter. Furthermore, classical logic does not consider resource restriction at all. For a detailed discussion on evidence and truth-value in NARS, see (Wang, 2009).

In NARS, the meaning of a concept, such as "Linda" or "feminist bank-teller", is determined by the available information about it, in terms of how it relates to other concepts, as far as the system knows. For a given concept, such information may be either *extensional* (indicating its instances or *special cases*) or *intensional* (indicating its properties or *general cases*). To decide the extent to which a concept, "Linda", is a special case of another one, "bank-teller" or "feminist bank-teller", the system will consider all available evidence. In this example, the most accessible evidence about all three concepts are *intensional* (i.e., about their properties), so the system reaches its conclusion by checking if Linda has the properties usually associated with "bank-teller" and "feminist bank-teller", respectively. Since according to the given information Linda has more common properties with "feminist bank-teller" than with "bank-teller", her "degree of membership" is higher to the former than to the latter. This is judged as a "fallacy" when probability theory is applied *extensionally* to this situation, so only the *base rates* matters, while the properties do not. For a detailed discussion on the categorization model in NARS, see (Wang & Hofstadter, 2006).

In summary, as soon as a normative model of rationality or intelligence makes more realistic assumptions, many "heuristics", "bias", and even "fallacies" follow from them. In the above examples, there are strong reasons for assuming that the truth-value of a statement should depend on both positive and negative evidence (rather than negative only), and the meaning of a concept should depend on both extensional and intensional relations (rather than extensional only). We believe these examples mainly show the limitations of traditional models (classical logic, probability theory), rather than human errors. The practice of NARS and similar systems shows that it is possible for a new normative model to explain and reproduce similar results in a unified way.

**Rationality Through Analogy (HDTP)** As a second case study, we want to sketch how *Heuristic-Driven Theory Projection* (HDTP), an analogy-engine, can be used to implement some crucial parts of our cognitively-based theory of rationality (for an expanded elaboration cf. e.g. (Besold et al., 2012)). HDTP is a framework for computing analogical relations between two domains that are axiomatized in many-sorted first-order logic (Schwering, Krumnack, Kühnberger, & Gust, 2009). It provides an explicit generalization of the two domains as a by-product of establishing an analogy. Such a generalization can be a base for concept creation by abstraction. HDTP proceeds in two phases: in the *mapping phase*, the source and target domains are compared to find structural commonalities, and a generalized description is created, which subsumes the matching parts of both domains. In the *transfer phase*, unmatched knowledge in the source domain

<sup>1</sup>At <http://www.cis.temple.edu/~pwang/papers.html>.

is mapped to the target domain to establish new hypotheses. HDTP is therefore similar in spirit to the well-known Structure-Mapping Engine (SME) (Falkenhainer, Forbus, & Gentner, 1989), e.g. with respect to the mentioned mapping and transfer phases and the symbolic representation of domains. Nevertheless, HDTP also differs significantly from SME, e.g. with respect to the strong expressive power of the underlying domain theories (many-sorted first-order logic in HDTP vs. propositional logic in SME), the establishment of the analogy relation as a by-product of an abstraction, and the massive usage of heuristics differ from the ones used in SME.

HDTP implements a principle (by using heuristics) that maximizes the coverage of the involved domains (Schwering et al., 2009). Intuitively, this means that the sub-theory of the source (or the target) that can be generated by re-instantiating the generalization is maximized. The higher the coverage the better, because more support for the analogy is provided by the generalization. A further heuristics in HDTP, for which the motivation is to prevent arbitrary associations, is the minimization of substitution lengths in the analogical relation, i.e. the simpler the analogy the better (Gust, Kühnberger, & Schmid, 2006). There is a trade-off between high coverage and simplicity of substitutions: An appropriate analogy should intuitively be as simple as possible, but also as general and broad as necessary, in order to be non-trivial. This kind of trade-off is similar to the trade-off that is usually the topic of model selection in machine learning and statistics.

The modeling of the Wason selection task with HDTP is quite simple as long as appropriate background knowledge is available, in case an analogy should be established, or the lack of appropriate background knowledge prevents analogy making, in case no analogy should be established. In other words, the availability of appropriate resources in form of background knowledge is crucial. If appropriate background knowledge for an analogous case is missing, then there is no chance to establish an analogical relation or a potential analogy (with low coverage and complex substitutions) is misleading the subject. Hence, subjects have to apply other strategies. This is the situation when subjects are confronted with the original Wason selection task based on properties of cards. Most subjects have problems to establish a meaningful analogy with a well-known domain due to the high degree of abstractness of the task itself. In the other case, if there is a source theory with sufficient structural commonalities, then the establishment of an analogical relation is straightforward. This happens if the task is changed in the following way: the rule that needs to be checked is now: *“If someone is drinking beer in a bar, he / she must be older than 21”*. In the experiment, subjects can choose between “drinking beer”, “drinking coke”, “25 years old”, and “16 years old” (Cosmides & Tooby, 1993). In the corresponding experiments, subjects behave significantly better than in the original selection task. With analogy making the improvement of the subjects in mastering the task can be explained. They can establish an analogy between the sketched set-up of the

experiment and a standard situation in daily life, in which they would simply do the necessary actions to check whether there is someone who is drinking beer in the bar without being older than 21: check people who are drinking beer, and check what people are drinking who are 16. As both situations are very similar to each other, the generalization is straightforward, substitutions length are minimal, and coverage is high.

The Linda problem is structurally different in comparison to the Wason selection task. In an analogy making context, an explanation of subjects’ behavior in terms of coherence maximization is promising. Coherence aspects of input theories are crucial for establishing analogies in several ways. Roughly speaking, the statement *“Linda is a bank teller”* has less coherence with Linda’s profile than the statement *“Linda is a bank teller and is active in the feminist movement”*. Therefore, it is easier to establish an analogy between Linda as given in Linda’s profile and Linda as described in *“Linda is a bank teller and is active in the feminist movement”* than in the pure “bank teller” case. Notice that from an abstract point of view the coherence-based resolution of the task is rather similar with the intensional interpretation of the task in NARS, where “feminist bank teller” has a higher degree of membership with Linda’s profile than “bank teller”.

## Conclusion and Future Work

There are multiple models of rationality, each with its own assumptions and applicable situations. The traditional models are based on certain idealized assumptions, and thus are limited to the domains where the latter are satisfied. Since human cognition has evolved in and is usually used in realistic situations where those idealized assumptions do not hold, those models of rationality are not universally applicable, and violations should not be deemed “irrational” per se. The seemingly irrational behaviors are there not because the intelligent systems (e.g. humans) are irrational, but because the traditional normative theories do not cover rationality very well.

Instead, we believe what is needed are new models of rationality that are based on more realistic assumptions and developed in a more holistic framework. Such models should be able to provide an adequate and feasible positive account of actual human rationality, also accommodating particularities of human-style reasoning. Such a framework could form a cornerstone of a closer connection between AGI and CogSci, embedding important parts of the AGI program within a CogSci context, whilst making the more general methods and theories of AGI accessible to the CogSci side.

The overall appeal for a “more cognitive” view on rationality models and systems is infrequent, but not unusual. Amongst others, already Kokinov (2003) reaches the conclusion that the concept of rationality as a theory in its own right ought to be replaced by a multilevel theory based on cognitive processes involved in decision-making. On the more technical side, there is a growing body of evidence that analogy engines (like HDTP) and general-purpose reasoning engines (like NARS) can be used for implementing these cognitive

mechanisms and, thus, also as foundations of a rationality-guided approach to general intelligence.

This paper should merely be considered as a point of departure, leaving questions for future research galore. For example with respect to the present proposal concerning HDTP, it seems recommendable to figure out to which extent different types of coherence concepts can be integrated into the framework. In particular, the challenges mentioned above need to be addressed, and a formal treatment of coherence needs to be fleshed out. Furthermore, an implementation of coherence principles for retrieval, mapping, and re-representation purposed in the analogy making process needs to be formulated. Concerning NARS, amongst others the following issues would merit work and effort: real-time temporal inference, procedural inference, and self-control. Regarding competing theories for rationality, clarifying to what extent cognitive capacities and limitations have already been taken into account (implicitly as well as explicitly) when designing the theories, and to what extent the classical frameworks can be re-instantiated by a cognitively-based approach, has to be considered one of the principal questions for future research. Finally, on a fundamental conceptual level, a broader definition of rational beliefs is still needed.

## References

- Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.
- Baum, E., Hutter, M., & Kitzelmann, E. (Eds.). (2010). *Artificial General Intelligence*. Lugano, Switzerland: Atlantis Press.
- Besold, T. R., Gust, H., Krumnack, U., Abdel-Fattah, A., Schmidt, M., & Kühnberger, K. (2011, July). An Argument for an Analogical Perspective on Rationality & Decision-Making. In J. van Eijck & R. Verbrugge (Eds.), *Proc. of the Workshop Reasoning About Other Minds (RAOM-2011)*. CEUR-WS.org, Vol. 751.
- Besold, T. R., Gust, H., Krumnack, U., Schmidt, M., Abdel-Fattah, A., & Kühnberger, K.-U. (2012). Rationality Through Analogy - Towards a Positive Theory and Implementation of Human-Style Rationality. In I. Troch & F. Breitenacker (Eds.), *Proc. of MATHMOD 12 Vienna*.
- Cosmides, L., & Tooby, J. (1993). Cognitive adaptations for social exchange. In J. H. Barkow and L. Cosmides and J. Tooby (Ed.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 163–228). Oxford.
- Evans, J. (2002). Logic and Human Reasoning: An Assessment of the Deduction Paradigm. *Psychological Bulletin*, 128, 978–996.
- Falkenhainer, B., Forbus, K., & Gentner, D. (1989). The Structure-Mapping Engine: Algorithm and Example. *Artificial Intelligence*, 41, 1–63.
- Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., et al. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3), 59–79.
- Gigerenzer, G. (2005). I think, therefore I err. *Social Research*, 72(1), 195–218.
- Gigerenzer, G. (2008). *Rationality for Mortals: How People Cope with Uncertainty*. Oxford University Press.
- Griffiths, T., Kemp, C., & Tenenbaum, J. (2008). Bayesian Models of Cognition. In R. Sun (Ed.), *The Cambridge Handbook of Computational Cognitive Modeling*. Cambridge University Press.
- Gust, H., Krumnack, U., Martínez, M., Abdel-Fattah, A., Schmidt, M., & Kühnberger, K.-U. (2011). Rationality and General Intelligence. In J. Schmidhuber, K. Thorisson, & M. Looks (Eds.), *Artificial General Intelligence* (pp. 174–183).
- Gust, H., Kühnberger, K.-U., & Schmid, U. (2006). Metaphors and Heuristic-Driven Theory Projection (HDTP). *Theor. Comput. Sci.*, 354, 98–117.
- Johnson-Laird, P. (1988). *Cognitive science*. Cambridge University Press.
- Kokinov, B. (2003). Analogy in Decision-Making, Social Interaction, and Emergent Rationality. *Behavioral and Brain Sciences*, 26(2), 167–168.
- Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). SOAR: An Architecture for General Intelligence. *Artificial Intelligence*, 1–64.
- Osborne, M., & Rubinstein, A. (1994). *A Course in Game Theory*. MIT Press.
- Pfeifer, N. (2008). A Probability Logical Interpretation of Fallacies. In G. Kreuzbauer, N. Gratzl, & E. Hiebl (Eds.), *Rhetorische Wissenschaft: Rede und Argumentation in Theorie und Praxis* (pp. 225–244). LIT-Verlag.
- Schwering, A., Krumnack, U., Kühnberger, K.-U., & Gust, H. (2009). Syntactic Principles of Heuristic-Driven Theory Projection. *Journal of Cognitive Systems Research*, 10(3), 251–269.
- Stenning, K., & van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. MIT Press.
- Thagard, P. (2002). *Coherence in Thought and Action*. MIT Press.
- Tversky, A., & Kahneman, D. (1983). Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, 90(4), 293–315.
- Wang, P. (2006). *Rigid Flexibility: The Logic of Intelligence*. Dordrecht: Springer.
- Wang, P. (2009). Formalization of Evidence: A Comparative Study. *Journal of Artificial General Intelligence*, 1, 25–53.
- Wang, P. (2011). The Assumption on Knowledge and Resources in Models of Rationality. *International Journal of Machine Consciousness (IJMC)*, 3, 193–218.
- Wang, P., & Hofstadter, D. (2006). A Logic of Categorization. *Journal of Experimental & Theoretical Artificial Intelligence*, 18(2), 193–213.
- Wason, P. C., & Shapiro, D. (1971). Natural and Contrived Experience in a Reasoning Problem. *Quarterly Journal of Experimental Psychology*, 23, 63–71.

# Musical Relevance: a Computational Approach

Edoardo Acotto (acotto@di.unito.it)

Università di Torino - Dipartimento di Informatica, Corso Svizzera 185  
10149, Torino ITALY

Daniele P. Radicioni (radicion@di.unito.it)

Università di Torino - Dipartimento di Informatica, Corso Svizzera 185  
10149, Torino ITALY

## Abstract

This study is a first attempt at formalizing the concept of *Musical Relevance* from a cognitive and computational perspective. We elaborate on Sperber and Wilson's Relevance Theory, and extend it to account for musical cognition, involving both listening and understanding. Our claim is that the application of the concept of *Cognitive Relevance* to music would permit us to partially explain hearers' behavior and composers' choices. A computational model of Musical Relevance could also contribute to the formulation of a general computational theory of musical cognition. In turn, formulating an algorithm to compute Musical Relevance can shed light on the computational nature of the broader cognitive principle of relevance. We propose to unify Relevance Theory with the Generative Theory of Tonal Music, in order to compute Musical Relevance. We started implementing a system to test the proposed approach over simple examples and report about the results in a preliminary experimentation.

**Keywords:** Relevance Theory; Computational Approach; A Generative Theory of Tonal Music; Tonal Pitch Space.

## Introduction

The *Relevance Theory* was initially formulated as cognitive-pragmatic theory of communication (Sperber & Wilson, 1986); lately it has been viewed and developed as a general theory of human cognition (Wilson & Sperber, 2004; Caruthers, 2006). The relevance of an input to an individual (or a cognitive system) is defined as the ratio between the cognitive effect and the processing effort. In the authors words:

“a.) Other things being equal, the greater the positive cognitive effects achieved by processing an input, the greater the relevance of the input to the individual at that time. b.) Other things being equal, the greater the processing effort expended, the lower the relevance of the input to the individual at that time.”

An input is relevant for a cognitive agent in a given context when it can be related with the information registered in memory and accessible, and when this relation yields a “positive cognitive effect”.<sup>1</sup> Relevance of an input is a continuous (non-categorical) variable. The concept is comparative and non quantitative: e.g., “x is more relevant than y, for P in the context C”.<sup>2</sup> The greater the cognitive effects are, the greater

<sup>1</sup>“A positive cognitive effect is a worthwhile difference to the individual's representation of the world: a true conclusion, for example. False conclusions are not worth having. They are cognitive effects, but not positive ones” (Wilson & Sperber, 2004).

<sup>2</sup>On the comparative/quantitative notion of relevance, and on Carnap's distinction of comparative and quantitative concepts, see (Sperber & Wilson, 1986).

the relevance of a given input is (*ceteris paribus*); on the other side, the smaller is the processing effort, the greater is the relevance of a given input (*ceteris paribus*).

It is a matter of fact that Relevance Theory has a lot of opponents. For example, in a footnote Jerry Fodor let us know that according to his opinion a Relevance Theory doesn't even exist: “As for a theory of relevance, saying that if we had one it would solve the frame problem is as pointless as saying that if we solved the frame problem, that would give us a theory of relevance: Both are true, of course, because ‘assessing relevance’ and ‘framing’ are two terms for the same thing. [...] If cognition is to attain true beliefs with any efficiency, it's got to be the case both that what's importantly relevant is generally in the frame, and that what's not importantly relevant generally isn't. Maybe meeting these conditions is tractable within the assumptions of Classical theories, but I don't know of any current proposal for a cognitive architecture, Classical or otherwise, that seems likely to tract it” (Fodor, 2000, p. 114). Fodor takes correctly Sperber and Wilson's theory as a semantic-pragmatic theory of linguistic comprehension, and he poses the question of how a cognitive system can attain true beliefs in an efficient way. However, if we conceive that Relevance Theory can be a general theory of (human) cognition, we have to remark that not all mental representations have a truth-based semantics, and mental representations of music seem to be a good candidate for representation without truth value (Acotto, 2011 (in press)). So, the efficiency of the cognitive system faced with non-semantic representations has to be analyzed with different criteria than those that Fodor has in mind.

One chief problem with Relevance Theory is the difficulty to formalize it: however, in a restricted and formal domain like music, this seems to be possible and psychologically plausible. Modeling musical relevance we have to shift from a “subjective” concept of relevance to an “objective” one: instead of modeling the musical relevance for a given individual in a given context, we'll model the relevance for an idealized listener familiar with the Western tonal music idiom. That is, we are presently concerned with a restricted subset of all possible music.

In their original formulation of Relevance Theory Sperber and Wilson (1986) do not propose any method for calculating relevance, so we had to provide relevance with a quantitative counterpart to design a computational model. This is a key contribution of the present work.

The formulation of an algorithm to compute Musical Relevance would lead to improve the general computational nature of the cognitive principle of relevance: “Human cognition tends to be geared to the maximization of relevance” (Wilson & Sperber, 2004). If Relevance Theory is empirically plausible, and if the musical mind yields a kind of thought comparable with other forms of mental life, Relevance Theory can apply to the musical thinking as well. In order to explore such hypothesis, we propose to put together Relevance Theory and Lerdahl and Jackendoff’s Generative Theory of Tonal Music, GTTM (Lerdahl & Jackendoff, 1983).<sup>3</sup>

The GTTM describes the musical comprehension of a hearer familiar with the Western tonal idiom. It postulates the existence of mental representations of music, structured on five levels: first the musical surface, then two horizontal structures (meter and grouping), and finally two hierarchical structures, the time-span reduction and the prolongational reduction, which can be formalized as binary branching trees (Hamanaka, Hirata, & Tojo, 2006). Generative Theory of Tonal Music finds in Lerdahl’s Tonal Pitch Space theory a partial readjustment (Lerdahl, 2001), especially concerning the formalization and the quantification of the musical dimensions.

Although other musicological theories exist that are related to musical salience (e.g., by Deliège (1996)), we chose the notion of relevance by Sperber and Wilson because it seemed to be more naturally suited to a computational implementation. A major assumption is that it allows formalizing and quantifying musical relevance via the computation of the musical cognitive effect and of the processing effort. We presently do not explore the connections of our work with related investigations of a notion similar to Musical Relevance and grounded on information theory (Conklin & Witten, 1995; Pearce, Conklin, & Wiggins, 2004): we defer to future work the exploration on such links.

The paper is structured as follows: we first qualify Musical Relevance as the ratio between the cognitive effect and the processing effort, and explore both cognitive (musical) effect and (musical) processing efforts. We then provide an example to show how such concepts fit to the musical context. Then a preliminary experimentation is illustrated, and the results are reported and discussed. Finally we conclude by pointing out present limitations and future works.

## Computing Musical Effect

According to Relevance Theory, in order to be more relevant than another, a music excerpt has to offer a greater cognitive/emotional effect than another one requiring the same pro-

cessing effort; alternatively, a musical excerpt has to require a minor processing effort than another one that yields the same effect. The Musical Relevance (*MR*) is defined as the ratio between the Musical Effect (*ME*) and Processing Effort (*PE*): that is,  $MR = ME/PE$ .

GTTM individuates three types of tonal tension: *surface*, *sequential* and *hierarchical* tension. Some experimental tests have been carried out, confirming that sequential tension is not sufficient to represent the effective musical understanding, and that hearers perceive hierarchical tension as well (Lerdahl & Krumhansl, 2007). The Musical Effect yielded by the tonal tension is complemented by the tonal attraction:<sup>4</sup> in other words, the “forces” that constitute musical effect are both *tensional* and *attractive*. In order to calculate the musical effect some rules can be applied, that were devised as Tonal Pitch Space Rules (Lerdahl, 2001). The following rules can be used to compute the musical effect.

### Surface tension rule

$$T_{diss}(y) = scale\_degree + inversion + non\_harmonic\_tones \quad (1)$$

where the tension score of the target chord  $y$  is computed as the sum of three elements *scale\_degree*, *inversion* and *non\_harmonic\_tones*. *scale\_degree* is 1 if  $3^\wedge$  or  $5^\wedge$  is present in the melodic voice, 0 otherwise; *inversion* is 2 if  $3^\wedge$  or  $5^\wedge$  in the bass, 0 otherwise; *non-harmonic tone* is 3 if a pitch class is a diatonic non-chord tone, 4 if it is a chromatic non-chord tone, 0 otherwise.

### Sequential tension rule

$$T_{seq}(y) = \delta(x_{prec} \rightarrow y) + T_{diss}(y) \quad (2)$$

where  $y$  is the target chord,  $x_{prec}$  is the chord that immediately precedes  $y$  in the sequence,  $T_{seq}(y)$  is the tension associated with  $y$ , and  $\delta(x_{prec} \rightarrow y)$  is the distance from  $x_{prec}$  to  $y$ .

### Hierarchical tension rule

$$\begin{aligned} T_{loc}(y) &= \delta(x_{dom} \rightarrow y) + T_{diss}(y); \\ T_{glob}(y) &= T_{loc}(y) + T_{inh}(x_{dom}) \end{aligned} \quad (3)$$

where  $y$  is the target chord,  $x_{dom}$  is the chord that directly dominates the prolongational tree;  $T_{loc}(y)$  is the local tension associated to  $y$ ;  $\delta(x_{dom} \rightarrow y)$  is the distance from  $x_{dom}$  to  $y$ ;  $T_{glob}(y)$  is the global tension associated to  $y$ ;  $T_{inh}(x_{dom})$  is the sum of the values of the distances inherited by the chords that dominate  $x_{dom}$ .

### Melodic attraction rule

$$\alpha(p_1 \rightarrow p_2) = \frac{as_2}{as_1} \cdot \frac{1}{n^2} \quad (4)$$

where  $p_1$  and  $p_2$  are pitches, with  $p_1 \neq p_2$ ;  $\alpha(p_1 \rightarrow p_2)$  is the attraction of  $p_1$  to  $p_2$ ;  $as_1$  is the anchoring strength of  $p_1$  and

<sup>3</sup>It is noteworthy that Musical Relevance model is not directly committed to the GTTM for computing the effect and the effort. E.g., we could employ different theories descending from (Meyer, 1956), such as (Narmour, 1990, 1992) and (Huron, 2006). However, Narmour’s Implication/Realization theory does not account for the *hierarchical* structure (i.e., binary branching tree) of music perception (Margulis, 2005, p. 688), and the same holds for Huron’s Expectation theory.

<sup>4</sup>The model by Lerdahl and Krumhansl is a quantitative theory of tonal tension made out of four components: “1. A representation of hierarchical (prolongational) event structure. 2. A model of tonal pitch space and all distances within it. 3. A treatment of surface (largely psychoacoustic) dissonance. 4. A model of voice-leading (melodic) attractions” (Lerdahl & Krumhansl, 2007).



$as_2$  is the anchoring strength of  $p_2$  in the current configuration of the basic space;  $n$  is the number of semitone intervals between  $p_1$  and  $p_2$ .<sup>5</sup>

*Harmonic attraction rule*

$$\alpha_{rh}(C_1 \rightarrow C_2) = c \cdot \frac{\alpha_{rvl}(C_1 \rightarrow C_2)}{\delta(C_1 \rightarrow C_2)} \quad (5)$$

where  $\alpha_{rh}(C_1 \rightarrow C_2)$  is the harmonic attraction of  $C_1$  toward  $C_2$ ; the constant  $c = 10$ ;  $\alpha_{rh}(C_1 \rightarrow C_2)$  is the sum of the attraction of the leading voices for all the voices in  $C_1$ ;  $\delta(C_1 \rightarrow C_2)$  is the distance of  $C_1$  a  $C_2$ , with  $C_1 \neq C_2$ .

Such rules have found an experimental corroboration in (Lerdahl & Krumhansl, 2007). We assume that these rules represent a good approximation of the musical effect in the overall computation of musical relevance.

For sake of simplicity and because of the greater complexity of music, in this paper we are concerned with melodic music only. Even though this is a clear simplification, and it is the first step of a more complete and complex model, our present work allows us to make experimental tests and to compute a musical relevance score for simple melodies.

## Computing Processing Effort

Concerning the *PE*, no methods to calculate it are given nor suggested in (Lerdahl & Jackendoff, 1983) nor in (Lerdahl, 2001). Nevertheless, following GTTM we can surely identify a vertical, hierarchical, dimension of the *PE* represented by the binary branching trees of the musical structure. Against the “concatenationism” hypothesis (Davies, 2011), musical surface is not enough to understand music, and the structural properties of a melody are a key element for its understanding. We can then assume that at least a great portion of the *PE* is involved in detecting the structural properties of the heard melody.

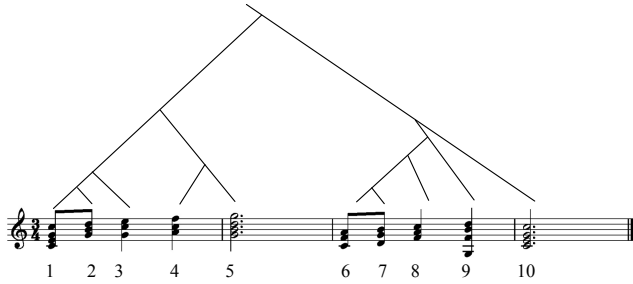


Figure 1: Structure of the toy melody represented in the GTTM notation, by Katz and Pesetsky (2009).

How to compute the binary branching trees that represent the hierarchical musical structure? In their reinterpretation

<sup>5</sup>In our implementation we adopted a correction to handle the case of repeated notes, with  $p_1 = p_2$ , that otherwise would produce a division by zero. Since the underlying rationale is that in case of repeated notes the ratio  $\frac{1}{n^2}$  should approach 1, we presently set the value of  $n$  to 0.9.

of the GTTM, Katz and Pesetsky observe that in the GTTM time-span trees the more relevant information is the hierarchical distance from the root of the tree (see Figure 1). This distance is measured through a Root Distance (*RD*) number: “The *RD* number of an event  $e$  in a structure  $K$ ,  $RD(e)$ , is the number of nodes that nonreflexively dominate the maximal projection of  $E$  (i.e.  $eP$ ) in  $K$ ” (Katz & Pesetsky, 2009).

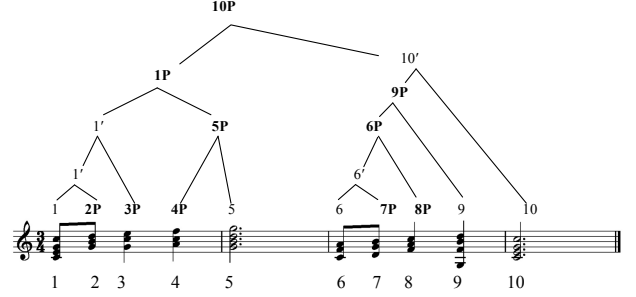


Figure 2: Structure of the toy melody in the standard linguistic notation.

The reinterpretation of Lerdahl and Jackendoff’s theory is made within the framework of generative linguistics, where the concept of “projection” plays a central role: “A constituent whose head is  $H$  is called a projection of  $H$ , and is conventionally labeled  $H'$  (‘H-bar’) if it is dominated by another projection of  $H$ ; and  $HP$  otherwise.  $HP$  is called the maximal projection and  $H'$  is called an intermediate projection of  $H$ .  $H$  itself is sometimes called the zero-level projection of  $H$ ”. As it is showed in Figure 2, the linguistics notation allows a graphic translation of the structure of the GTTM.<sup>6</sup>

We will take into consideration the *RD* number of each sound event as a part of the *PE* (we increase by 1 the *RD* numbers). So, we shall measure the *PE* by using the rules of time-span reduction formulated in GTTM.

## An example with melody

Searching for a first implementation of our model, we focus initially on the case of the melodies, in particular the leading voice of the toy melody, and illustrated in Figure 3.

Since we are considering a melody, we only make use of the *melodic attraction rule* (please refer to Equation 4), since the other rules are concerned with events where multiple notes are present at a time.

**Music effect** Since melodic attraction is between each two musical events, an attraction number is not referred to a single

<sup>6</sup>“Variations in the notation with which one expresses a theory can influence one’s thinking about the actual topics under investigation. Even when different sorts of diagrams represent exactly the same information (as is the case here), the differences among them may reflect and reinforce differing working hypotheses or hunches about the kinds of phenomena one expects to model. Differences of this sort between GTTM and common practice in linguistics arise in two important domains: the relevance of projection level and the amount of information that project from terminal nodes to the constituents that they head.” (Katz & Pesetsky, 2009)



Figure 3: The leading voice of the toy melody.

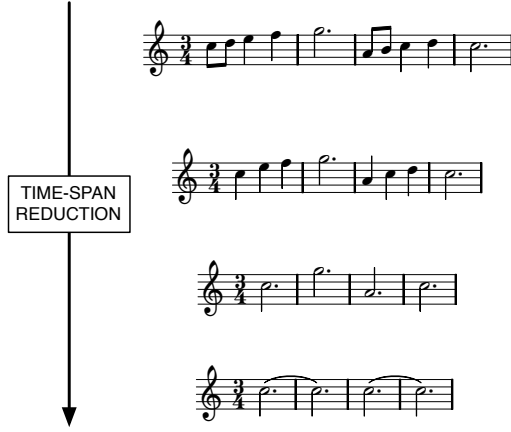


Figure 4: The three levels of the toy melody time-span reductions.

pitch, but represents a transition between two adjacent pitches  $x$  and  $y$ . The attraction values (i.e. the musical effect) between each two notes in the toy melody are as follows:

$$\begin{aligned}
 \alpha(C \rightarrow D) &= 0.125 \\
 \alpha(D \rightarrow E) &= 0.375 \\
 \alpha(E \rightarrow F) &= 0.667 \\
 \alpha(F \rightarrow G) &= 0.375 \\
 \alpha(G \rightarrow A) &= 0.007 \\
 \alpha(A \rightarrow B) &= 0.25 \\
 \alpha(B \rightarrow C) &= 2.0 \\
 \alpha(C \rightarrow D) &= 0.125 \\
 \alpha(D \rightarrow C) &= 0.5
 \end{aligned}$$

**Processing Effort** In order to calculate the  $PE$ , we compute the  $RD$  numbers by following the binary branching tree: for each pair of events the  $PE$  will be the average of the two  $RD$  numbers (augmented by a unit). To compute processing effort implies the possibility to automatically “reduce” a given melody to its more fundamental schema, as it is shown in Figure 4.

This kind of reduction relies on a set of “preference rules”. However, these are not easily implemented because if there are cases where multiple rules are triggered at the same time, unfortunately in the GTTM no criteria are proposed to resolve such conflicts. Deepening the computation of the  $PE$  component of  $MR$  will be one major focus of future work.

**Musical Relevance** We can now calculate the musical relevance of the transition from one event to another in a melody (without considering, for the moment, the relevance of similarity). As each note has a value of  $PE$ , but the attraction value



Figure 5: Excerpt from the Piano Sonata in C minor (KV. 457) by W.A. Mozart.



Figure 6: Excerpt from the “Petit pièce” from the Opus 68 n.5 by R. Schumann.

is between two events, we’ll calculate the average  $PE$  of each couple of notes, and then we’ll calculate the relevance (the ratio between  $ME$  and  $PE$ ) of the passage from one musical event to another.

## Experimental Assessment

In order to provide the proposed approach with an experimental assessment, we devised the following experimental setting.

We implemented a system that computes the *Musical effect* ( $ME$ ) component in the equation  $MR = ME/PE$ . We studied how the computed musical effect differs by varying two simple melodies.

A core hypothesis of the experimentation is that melodies by historical composers maximize  $ME$ , and  $MR$  as well.<sup>7</sup> Let us consider the *cognitive constraints* by Lerdahl (1988), that postulates that ‘good’ music is composed according to the cognitive nature of human mind/brain. By following this theoretical framework we stipulate that the original melodies from historical composers have higher  $ME$  than ‘experimental’ variations composed by ourselves.

In this setting, we expect the system to compute lower  $ME$  for our variations; also, we expect it to be able to distinguish between grammatical and ‘ungrammatical’ variations, by assigning lower scores to ungrammatical ones.

**Experimental setting** We selected the music excerpts illustrated in Figures 5 and 6. Such pieces were chosen in order to capture (and test the system in) two widely different experimental conditions. In particular, they can be thought of as two paradigmatic examples of themes opposite in spirit. The first one is rather percussive and jumps over the main degrees of the C minor key. On the other side, the second one is a typical *cantabile* theme: it is more regular under a rhythmic viewpoint, and the melody mostly moves stepwise.

<sup>7</sup>In accord with the given definition, music relevance ( $MR$ ) grows *ceteris paribus*— as the music effect ( $ME$ ) grows, and vice versa it decreases *ceteris paribus*— as the processing effort ( $PE$ ) grows. Since we added new events by interleaving existing events with new ones, this makes the input more complex. Then we know in advance that new nodes will produce further levels to the reduction tree, thus increasing the  $PE$  component. That is, we know *a priori* that by increasing  $PE$  and by decreasing  $ME$ , the final  $MR$  will result reduced.

Table 1: The Musical Effect scores for the six considered pieces.

<i>Excerpt</i>	<i>Score</i>
Mozart excerpt	0.4259
Mozart Var 1	0.3888
Mozart Var 2	0.3267
Schumann excerpt	0.3182
Schumann Var 1	0.2460
Schumann Var 2	0.1877

Therefore, different musical aspects are accounted for by the considered excerpts.

We then elaborated two variations for each excerpt (Figure 7). In both cases the first variation (indicated as *Var 1* in Figure 7-a) and 7-b)) is only slightly different from the original excerpt. As regards as the second variation, the Mozart excerpt has been modified in a ungrammatical fashion (see *Var 2* in Figure 7-a)), whilst Schumann excerpt has been modified through the insertion of musically plausible notes that transform it into a rhythmically regular arpeggio (see *Var 2* in Figure 7-b)).

The implemented system takes in input the excerpts encoded as MIDI files, and computes the associated musical effect –through the formula in Equation (4)– as the sum of the melodic attraction between each two music events:

$$ME_{excerpt} = \left[ \frac{\sum_{i=1}^{excerpt.length-1} \left( \frac{as_{i+1}}{as_i} \cdot \frac{1}{n^2} \right)}{excerpt.length} \right]$$

By adding new notes we expected a reduction in the musical effect. Furthermore, since the second variation of each input excerpt was more different from each ‘original’ source, we expected to observe a decrease in the musical effect and, relatedly, in musical relevance.

**Results** In accord with our intuition, the implemented system computed the maximum ME scores for the original excerpts, reduced scores for each first variation, and the lowest scores for both second variations. The final figures are reported in Table 1.<sup>8</sup> Provided that this experimentation represents only the very first step towards a psychological validation (that would require considering in how far the results approach human responses), the results seem to corroborate Lerdahl’s hypothesis. Tonal music is governed by an attraction-based syntax. This sort of attraction, which is maximally exploited by composers and which is maximal in the original music excerpts, is at least partly grasped by the proposed model. Further work is needed to investigate whether and how classical western tradition as a whole ‘incorporates’ a criterion to maximize the Musical Effect (independently of the associated processing effort).

<sup>8</sup>The material employed in the experimentation (MIDI files, printable scores and Lilypond sources) along with the results file is available at the URL <http://www.di.unito.it/~radicioni/datasets/cogsci12/>.

Also, if we compare the original excerpts (Figure 5 and 6), the system accounts for the greater ‘dramatical’ salience of Mozart’s excerpt (which is a first theme of a C minor sonate). Schumann’s excerpt is a simple piece: its value results perhaps from the balance between its simple effect and its structural simplicity.

## Conclusions

The paper illustrates a modeling attempt, and an initial implementation of a complex phenomenon such as relevance-guided music understanding. The presented implementation only accounts for the musical effect; coping with the computation of the processing effort is left for future work.

Due to such limitation we considered for experimentation only melodies with differing surface, but with similar underlying structures. Notwithstanding this limitation, the preliminary experimentation provided some evidence that the musical effect captures meaningful aspects of Western tonal music.

Another relevant point for completing the Musical Relevance model involves dealing with musical similarity. Similarities and repetitions in music are a frequent and important structural phenomenon. They affect important musical features like style (Meyer, 1989), but they permit also to affirm that without similarity there would be no music, since similarity is a center of gravity for perception and comprehension (Cambouropoulos, 2009). Similarities and repetitions influence musical effect, and therefore they have impact on the cognitive relevance of a piece of music. The relevance of a musical event  $E_2$  should be a function both of the relevance of the similar event  $E_1$ , and of the relevance of  $E_2$  as taken in isolation. The similarity increases the relevance of a musical event; otherwise, similarities should be avoided for the risk to diminish the relevance effect. Starting from existent systems for computing musical similarity (Meredith, Lemström, & Wiggins, 2002; Radicioni & Botta, 2006), in future works we will focus on detecting similar patterns in music pieces.

## Acknowledgments

This work has been partly supported by the project *Speak2Home*. We are grateful to two anonymous reviewers for their valuable advices that helped to substantially improve the work, and to Jelle Gerbrandy for discussions on some subtle aspects of the paper.

## References

- Acotto, E. (2011 (in press)). Mental Representations of Music in Cognitive Science. In *Proceedings of the Italian Cognitive Science Society (AISC) Conference*.
- Cambouropoulos, E. (2009). How similar is similar? *Musicae Scientiae, Discussion Forum 4B*, 7–24.
- Carruthers, P. (2006). *The Architecture of the Mind*. Oxford: Oxford University Press.
- Conklin, D., & Witten, I. H. (1995). Multiple Viewpoint Systems for Music Prediction. *Journal of New Music Research*, 24(1), 51–73.



Figure 7: a) variations of the Mozart excerpt; b) variations of the Schumann excerpt.

- Davies, S. (2011). *Musical Understandings: And Other Essays on the Philosophy of Music*. Oxford, UK: Oxford University Press.
- Deliège, I. (1996, October). Cue Abstraction as a Component of Categorisation Processes in Music Listening. *Psychology of Music*, 24(2), 131–156.
- Fodor, J. (2000). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge: MIT Press.
- Hamanaka, M., Hirata, K., & Tojo, S. (2006). Implementing “A Generative Theory of Tonal Music”. *Journal of New Music Research*, 35(4), 249–277.
- Huron, D. (2006). *Sweet anticipation. music and the psychology of expectation*. Cambridge: MIT Press.
- Katz, J., & Pesetsky, D. (2009). *The Identity Thesis for Language and Music*. draft available at <http://ling.auf.net/lingBuzz/000959>.
- Lerdahl, F. (1988). Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition. In J. Sloboda (Ed.), (pp. 231–59). Oxford University Press.
- Lerdahl, F. (2001). *Tonal pitch space*. Oxford: Oxford University Press.
- Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge: MIT Press.
- Lerdahl, F., & Krumhansl, C. L. (2007). Modeling tonal tension. *Music Perception*, 24, 329–366.
- Margulis, E. (2005). A Model of Melodic Expectation. *Music Perception*, 22(4), 663–714.
- Meredith, D., Lemström, K., & Wiggins, G. (2002). Algorithm for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4), 321–345.
- Meyer, L. (1956). *Emotion and meaning in music*. Chicago, USA: Chicago University Press.
- Meyer, L. (1989). *Style and Music*. Chicago: Chicago University Press.
- Narmour, E. (1990). *The analysis and cognition of basic melodic structures*. Chicago, USA: University of Chicago Press.
- Narmour, E. (1992). *The analysis and cognition of melodic complexity*. Chicago, USA: Chicago University Press.
- Pearce, M., Conklin, D., & Wiggins, G. A. (2004). Methods for Combining Statistical Models of Music. In *Computer Music Modeling and Retrieval: Second International Symposium, CMMR 2004* (pp. 295–312).
- Radicioni, D. P., & Botta, M. (2006). A Methodological Contribution to Music Sequences Analysis. In F. Esposito & Z. W. Ras (Eds.), *Foundations of Intelligent Systems* (pp. 409–418). Berlin: Springer-Verlag.
- Sperber, D., & Wilson, D. (1986). *Relevance. Communication and Cognition*. Oxford: Blackwell.
- Wilson, D., & Sperber, D. (2004). Handbook of Pragmatics. In G. Ward & L. Horn (Eds.), (chap. Relevance Theory). Oxford: Blackwell.

# Examining the Connection Between Dynamic and Static Spatial Skills and Video Game Performance

Deanne M. Adams ([adams@psych.ucsb.edu](mailto:adams@psych.ucsb.edu))

Richard E. Mayer ([mayer@psych.ucsb.edu](mailto:mayer@psych.ucsb.edu))

Department of Psychological and Brain Sciences  
University of California, Santa Barbara, CA 93106-9660

## Abstract

Previous research has found a connection between spatial cognition and success in STEM areas and that spatial cognition skills can be trained using video games. The present study explores whether a relationship exists between non-trained spatial cognition skills and video game performance. Non-video game players first completed four spatial cognition tasks and then played two video games, Tetris and Unreal Tournament (UT). Results showed significant correlations between performance on UT and mental rotation, paper-folding, and the Race dynamic spatial task. In contrast, Tetris performance only correlated with paper-folding. These results indicate that performance on action video games such as UT may be related to more spatial skills than Tetris.

**Keyword:** spatial cognition, video games

## Objective and Rationale

The goal of the present study is to examine the relationship between video game performance and performance on tests of static and dynamic spatial skills. The rationale is that (a) spatial skills may be instrumental for success in science, technology, engineering, and mathematics (STEM), and (b) video game playing may be related to the development of spatial skills.

## Spatial Cognition and STEM Areas

In a longitudinal study Wai, Lubinski, and Benbow (2009) examined the connection between adolescents' spatial ability and later achievement in STEM fields (i.e., science, technology, engineering, and mathematics). Cognitive ability measures of mathematical, verbal, and spatial ability for 400,000 participants from the Project TALENT data bank of 9<sup>th</sup>-12<sup>th</sup> grades were compared to follow-up academic data from 11 years later. Three major conclusions were made from this study: (1) high spatial ability was found among almost all adolescents who went on to achieve educational and occupational credentials in STEM areas; (2) spatial ability was critical for students in the general population as well as those deemed

intellectually talented; and (3) restricting talent searches to verbal and mathematical ability may miss many spatially gifted individuals. If we wish to encourage students to go into STEM fields the educational system may need to adapt to include spatial measures in talent assessment as well as to include interventions and training that encourage the development of spatial skills.

Studies in different areas of STEM have shown different spatial abilities are used on these tasks. For physics, Kozhevnikov, Hegarty, and Mayer (2002) found that there was a significant correlation between students' spatial abilities and accuracy on kinematic problems. Further research by Kozhevnikov, Motes, and Hegarty (2007) found several differences between high- and low-spatial students and the answer they gave to physics problems. High-spatial students could integrate several motion parameters while low-spatial students only considered one. High-spatial students used kinematic graphs as abstract representations of motion while low-spatial students interpreted graphs as being picture-like representations. For representations of the problems, high-spatial students could reorganize representations of spatial problems into other corresponding representations while low-spatial students used multiple, uncoordinated representations of the same problems. Eye-tracking also showed that high-spatial students made eye movements that corresponded to elements of the problem while low-spatial students did not. The researchers suggested this is due to the high-spatial individuals visualizing the correct movement produced when the two movement components were integrated.

To help improve performance in engineering, Sorby and Baartmans (2000) developed a 10-week course to help teach spatial visualization skills to freshman engineering majors who were identified as having lower scores on the Purdue Spatial Visualization Test: Rotations (PSVT:R). Twenty-four students took the course while 72 acted as the control group. During the course students were taught topics such as rotations of objects, cross-sections of solids, and translation and scaling. Those who participated in the course showed significant gains on the PSVT:R beyond simple practice effects. Furthermore, after later analyzing the transcripts for the all of the 96

students, the researchers found that students who took the course scored higher on later graphics courses and had higher retention rates in the graphically oriented engineering program.

While higher spatial ability often facilitates STEM learning for novices, experts in the field often do not use spatial strategies when solving problems. For example, Stieff (2007) found that students used mental rotation strategies when determining if molecular diagrams were identical or enantiomers. In contrast, experts typically used an analytical strategy to complete the task. Uttal and Cohen (in press) propose that spatial ability actually acts as a gateway to getting into STEM fields. While experts develop contextualized spatial abilities and can also use prior semantic knowledge, novices must rely on de-contextualized spatial abilities. Therefore spatial training such as Sorby and Baartmans's (2000) can be used to help novices develop these skills and prevent dropout.

### **Video Games and Spatial Cognition**

Spence and Feng (2010) propose that if students possess poor spatial skills, they will avoid learning in academic areas that require spatial cognition, such as STEM subjects. Similar to verbal or mathematic ability, we must try to improve spatial ability through training either through early education or through play. Research has shown that video games can be used to improve spatial cognition skills. Terlecki et al. (2008) found long lasting, transferable effects on spatial skills after participants were trained using Tetris. For the study, both control and experimental subjects were given a mental rotation task (MRT) once a week over a 12 week period. Participants in the experimental condition also played Tetris for one hour a week while the control participants played Solitaire. At the end of the 12 weeks, large improvements were found for both the repeat MRT exposure control group and the video game training group. Participants in the videogame training condition, however, showed significantly greater transfer effects on the Guilford-Zimmerman Spatial Visualization Task (Guilford & Zimmerman, 1947) and the Surface Development Test (SDT) (Ekstrom, French, & Harman, 1976).

Feng, Spence, and Pratt (2007) also found improvement on a mental rotation task after participants played a first-person shooter action game called Medal of Honor: Pacific Assault. First-person shooter games involve simulated combat in which the player competes against other players or computer controlled enemies. Improvement in mental rotation correlated with improvement on a useful-field-of-view task (UFOV). The authors suggest that the improvement in mental rotation was due to improving lower level attention skills.

Furthermore, Subrahmanyam and Greenfield (1994), in a study on improving both dynamic and static skills in school aged children, found that those with the best initial

spatial skills were best at playing the video game. The researchers proposed that strong dynamic spatial skills helped the participants master a game while practice strengthened weak dynamic spatial skills. The fact that already existing strong dynamic spatial skills facilitate game mastery relates back to Uttal and Cohen's (in press) argument that spatial ability can act as a gateway to STEM areas. In these training studies students/participants are completing these training regimes either because they are being paid or as part of a class exercise. In the real world there is no external motivator to encourage individuals to continue playing video games if they find them too difficult, perhaps due to a lack of spatial ability, or because they find video games unappealing in general. Studies with expert video game players that display high spatial and visual attention skills are also dealing with individual that were self-selected game players (Feng et al., 2007). These individuals may have become regular game players because they possessed higher relative spatial skills to begin with. This could be causing a Matthews effect in which individuals with higher spatial skills enjoy playing action video games therefore they play more of them and further improve their spatial abilities. The question then becomes whether lacking preexisting spatial ability could affect video game performance and therefore affect the motivation to continue playing.

### **Present Study**

The present study examines the relationship between performance on static spatial tasks (i.e., mental rotation and paper-folding) and dynamic spatial tasks (i.e., race task and interception task) on the one hand and novice performance on commercial video games (i.e., Unreal Tournament and Tetris) on the other. Prior research by Terlecki et al. (2008), Feng et al. (2007) and Subrahmanyam and Greenfield (1994), have found that playing video games can increase performance on tasks involving different cognitive skills. However, no study has explored how an individual's performance on any of these cognitive tasks may relate to their ability to play video games.

**Participants.** The participants were 69 college students from the University of California, Santa Barbara (44 women, 25 men). Ages ranged from 18 to 27 years old with a mean age of 19.29 years. All participants were classified as non-video game players, meaning that that did not regularly play commercial video games during their free time.

**Materials and Procedure.** The experiment took place over two sessions that were scheduled within two days of one another. In the first session, participants first filled out a survey assessing their video game usage to remove

any regular video game players. Next, they completed battery of tests including two dynamic spatial tasks and two static spatial tasks. All tasks included in the battery were administered on computers using electronic versions. Static spatial ability was assessed using a version of the Shepard and Metzler (1970) mental rotation test (MRT) and the paper-folding test (Ekstrom et al., 1976). The first cognitive task that participants completed was paper-folding. During the task, on each item, a series of pictures is presented demonstrating how a piece of paper is being folded. The last picture on the left includes circles that signify where holes are punched all the way through the folded piece of paper. To the right of the vertical line are five figures with possible configurations for the holes in the paper once it has been completely unfolded. Participants indicated which of the five figures they believed had the correct configuration by pressing the corresponding key (i.e., 1-5). Participants had 3 minutes for each of the 2 sets of 10 trials.

The paper-folding task was followed by the mental rotation task, in which two 3D block figures were presented simultaneously. The participants were asked to indicate whether the two items were the same or different (i.e., mirror-reversed images). On some trials the block figure on the right hand side of the screen was a version of the left figure rotated in depth varying by 20° intervals. Participants completed a total of 120 trials for this task.

Dynamic spatial skills were assessed using variations of the Race2 and Interception tasks as described by Hunt et al. (1988). Specifics for the Race2 task were taken from D'Oliveira (2009). During the Race task, two oval objects (one black, one white) race horizontally toward their own finish lines, as exemplified in Figure 1. Three relative speed differences between the objects were used to vary trial difficulty. Participants completed one block with a total of 108 trials. For the Interception task, the goal is to hit a moving target as it passes across the top of the screen with a 'missile' fired from a fixed location along the bottom of the screen. The target appeared in four different locations and traveled at four different speeds. Participants completed a total of 64 trials. The intent of the task was to measure how accurately the participants can judge the amount of time it will take for the 'missile' to intercept the target.

During the second session, participants played an hour of a first-person shooter action video game called Unreal Tournament 2004 (UT) as well as 45-50 minutes of the puzzle game Tetris. Which game was played first was counterbalanced between sets of participants.

During the 1 hour action video game session, game play was separated into three 20 minutes intervals. During the first two 20 minute game intervals participants played at the lowest level of difficulty (novice). For the last 20 minute game interval, the difficulty level was raised to the second lowest difficult level (average). The difficulty level determine how effective the 16 enemy

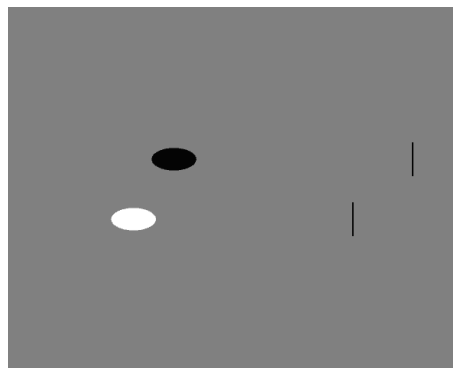


Figure 1: Sample trial from Race task adapted from D'Oliveira (2009)

“bots” controlled by the game were at firing and tracking down the player. In order to be successful the players must be able to avoid enemy fire while aiming and returning fire. The game keeps track of how many times the player kills an enemy as well as how many times the player is killed.

Tetris is a puzzle game in which players use 7 different block shapes in order to create lines. Every time a complete line is created the line disappears and the player is awarded points. The more lines that the player completes at a time the higher the point value awarded. If incomplete lines fill the given area the game ends. Participants are given one block shape at a time and have to place it somewhere at the bottom of the play area. As the player's score increases, the falling rate of the blocks increase making the game harder. The 7 block shapes can be rotated in increments of 90 degrees. Five of the block shapes are asymmetrical allowing for 4 different possible configurations that can be used to complete lines.

After finishing playing each game, participants were asked questions relating to their game satisfaction including how much they enjoyed the game, how likely they would play the game again, and how difficult they found playing the game. To measure performance on UT, total number of kills across all three games was used. For Tetris, performance was assessed by recording the highest score achieved while playing the game.

## Results

To first examine whether the order in which the participants played the game affected their performance independent sample t-tests were conducted. No significant differences were found between participants who played UT first or second for either UT performance,  $t(63) = .224, p = .823$ , or highest score on Tetris,  $t(64) = .906, p = .368$ . Looking at performance on just the first two games of Tetris, since all participants played at least 2 games, revealed that participants in the UT first group scored significantly higher than those that played Tetris



Table 1: Correlations between the game performance measures and performance on the four cognitive tasks.

Game performance	Mental rotation		Paper folding	Race RT Combined Easy	Intercept
	Total Errors	Mean RT	Score	Trials	Accuracy
Tetris (Top Score)	-.206	.002	.244*	-.196	.185
Unreal Tournament (Total Kills)	-0.267*	-0.334**	.271*	-.269*	.212

\*Designates a significance level of .05

\*\* Designates a significance of .01

first,  $t(64) = .2.087, p = .041$ , on their second game. This suggests that playing UT first may actually increase performance on Tetris playing. To examine whether performance on one video game related to performance on the other a Pearson correlation was conducted and revealed no significant correlation between UT kills and highest Tetris score,  $r(64) = -.007, p = .957$ .

**Static Spatial Tasks.** Table 1 shows the correlation between measures for mental rotation and paper-folding and UT and Tetris performance. For paper-folding, performance was assessed by combining the score for the two segments. Participants were penalized (-.2) for attempted items that they got wrong. As shown in the first column of Table 1, there was a modest, positive correlation between UT total kills and paper-folding score,  $r(63) = .271, p = .032$ , as well as a modest, positive correlation between high Tetris score and paper-folding score,  $r(64) = .244, p = .050$ . Next we examined whether response time (RT) performance on mental rotation correlated with performance on either UT or Tetris. As shown in the second column of Table 1, there was a moderate, negative correlation between UT total kills and mean response time across all mental rotation items,  $r(64) = -.334, p = .007$ . This indicates that participants who had higher scores on UT had faster response times for the mental rotation. There was no significant correlation between highest Tetris score and response time for mental rotation,  $r(66) = .002, p = .990$ . The same pattern was obtained for error rates on the mental rotation task: there was a significant negative correlation between total kills in UT and error rate for mental rotation,  $r(64) = -.267, p = .031$ , but no significant correlation between highest score in Tetris and error rate for mental rotation,  $r(66) = .002, p = .990$ .

**Dynamic Spatial Tasks.** For the race task we examined both accuracy (indicating the correct object would win the race) and response time (how quickly participants

determined which object would reach the goal first). Response times were only used for correct trials. In general there was a high accuracy across all participants, especially at the higher relative speed increases (easier trials). For Tetris, there was no significant correlation between Tetris score and race accuracy,  $r(67) = .144, p = .245$ , or race response times,  $r(67) = -.177, p = .152$ . For UT, there was no significant correlation between total kills and race accuracy,  $r(67) = .148, p = .239$ . When looking at response times, analysis revealed a weak, negative correlation between total kills and average response times for the higher speed difference ratios,  $r(65) = -.269, p = .030$ , but not the smallest speed increase,  $r(65) = -.178, p = .157$ , or overall average response times,  $r(65) = -.242, p = .052$ . This indicates that at the easier levels, when the relative speed between the two objects was greater, participants that had higher scores on UT made faster responses on the race task.

Finally, for the intercept/missile task, performance was measured by looking at the participant's accuracy at hitting the target over the 64 trials. No significant correlation was found between missile accuracy and game performance for either total kills in UT,  $r(64) = .212, p = .091$ , or highest score in Tetris,  $r(64) = .185, p = .135$ .

## Discussion

The results from the present study suggest that individuals who already possess certain spatial skills, such as indicated by faster response times for dynamic spatial skills, mental rotation, and paper-folding, are more likely to do better at playing an action video game such as Unreal Tournament, whereas the connection between spatial skills and early Tetris appears to be less pronounced.

Unlike prior research by Terlecki et al. (2008) this study found no connection between Tetris playing and mental rotation. This is similar to results found by Sims and Mayer (2002) in which there were no significant differences between high-skill and low-skill Tetris players

on any spatial tasks except for those that involved Tetris shapes or Tetris like shapes. In a separate study they also found that after 12 hours of training Tetris, participants did not significantly differ on mental rotation tests from those who had not practiced except for using different strategies when rotating Tetris shapes. Research by Kirsh and Maglio (1994) suggests that Tetris may not be the best environment to increase mental rotation performance. Tetris players often offload the mental rotation effort by using the game mechanics to rotate the figures instead of mentally doing the rotation. According to their model, the player should first compute the best place to put the Tetris piece before planning any rotation or movement to place it. The data shows that participants actually begin rotating the shape very early, before a possible placement could be decided upon. Rotating a shape in the external world is classified as an epistemic action, which are actions that are designed to change the input to the information-processing system and a way that an individual can alter the external environment to provide information needed to complete the task. In Tetris, if the game can complete the rotations faster than it takes the individual to do the rotations mentally, then it make sense from a limit cognitive resources standpoint for the individual to rely on the external rotation that only requires a simple motor action to complete (Kirsh & Maglio, 1994). Therefore, if participants are not actually practicing mental rotation during Tetris, it may not be the best game to use to increase spatial cognition.

Our findings did show that Tetris performance correlated significantly with paper-folding. One possible reason for this is that while Tetris does not require skills in mental rotation, it does require other visualization skills. Good Tetris players may be better able to visualize all possible configurations of the Tetris pieces therefore being able make quicker decisions about where to place pieces.

Both Tetris and action video games have been used to successfully train spatial cognition skills such as mental rotation (Terlecki et. al. 2008, Feng et al., 2007). The results from this study lead to two questions that should be addressed in future research. First, would playing one type of game lead to higher, more transferable, or longer lasting spatial skills compared to playing another? Four measures from our static and dynamic spatial skills were related to performance in game playing for Unreal Tournament while only one was related to Tetris. Feng et al. (2007) theorize that improvements in spatial tasks depend on the skills required during the game. The action video game improves lower-level spatial attention capacities, which in turn improves MRT performance. Their control condition did use a 2D puzzle game but because the game did not sufficiently exercise spatial attention capacities there was no benefit from playing. This suggests that the best way to improve mental rotation is to improve spatial skills in a way that is more

generalizable. This could involve improving lower level cognition skills such as attention. Spence and Feng (2010) also proposed that along with improving spatial selective attention, one other key difference between puzzle games like Tetris and many action video games is the perspective. The majority of puzzle games are played from a more allocentric perspective making the visuomotor coordination in Tetris is less natural compared to the more egocentric perspectives used in most first person shooter action games. Playing an action video game may therefore be more likely to increase spatial ability. A comparison of two comparable training regimes using either Tetris or an action game like Unreal Tournament should be done to determine whether there is any benefit to using one game over the other.

Another possible question is if both games can increase spatial ability, as prior research has shown, could the required level of preexisting spatial ability to play affect whether it is a viable choice for training? As proposed by Spence and Feng (2010), a student with poor spatial skills may avoid tasks that require them. As mentioned previously, this suggests the potential for a Matthew effect with spatial cognition and video games. Those that already possess high spatial skills do better at playing action video games and are more likely to continue to playing. By continuing to play these individuals increase their spatial ability. Therefore, a game such as Tetris which appeared to require less preexisting spatial skill to be successful may encourage students to play more than one like Unreal Tournament.

## References

- Ekstrom, R.B., French, J.W., & Harman, H.H. (1976). *Kit of factor referenced cognitive tests*. Educational Testing Service: Princeton, NJ.
- Epic Games (2004) Unreal Tournament 2004 [computer software]
- Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science*, 18(10), 850-855.
- D'Oliveira, T.C. (2009). Dynamic spatial ability: An exploratory analysis and a confirmatory study. *The International Journal of Aviation Psychology*, 14(1), 19-38.
- Guilford, J., & Zimmerman, W. (1947). *Guilford-Zimmerman Aptitude Survey*. Orange, CA: Sheridan Psychological Services.
- Hunt, E., Pellegrino, J.W., Frick, R.W., Farr, S.A., & Alderton, D. (1988). The ability to reason about movement in the visual field. *Intelligence*, 12, 77-100.
- Kirsh, D., & Maglio, P. (1994). On the distinguishing epistemic from pragmatic action. *Cognitive Science*, 18, 513-549.
- Kozhevnikov, M., Hegarty, M., & Mayer, R.E. (2002).

- Visual/spatial abilities in problem solving in physics. In M. Ander, B. Meyer, & P. Oliver (Eds.). *Diagrammatic Representations of Reasoning*. Springer-Verlag.
- Kozhevnikov, M., Motes, M.A., & Hegarty, M. (2007). Spatial visualization in physics problem solving. *Cognitive Science*, 31, 549-579.
- Shepard, R.N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171, 701-703.
- Sims, V.K. & Mayer, R.E. (2002). Domain specificity of spatial expertise: The case of video game players. *Applied Cognitive Psychology*, 16, 97-115.
- Sorby, S.A., & Baartmans, B.J. (2000). The development and assessment of a course for enhancing the 3-D spatial visualization skills of first year engineering students. *Journal of Engineering Education*, 301-307.
- Spence, I., & Feng, J. (2010). Video games and spatial cognition. *Review of General Psychology*, 14 (2), 92-104.
- Stieff, M. (2007). Mental rotation and diagrammatic reasoning in science. *Learning and Instruction*, 17 219-234.
- Subrahmanyam, K. & Greenfield, P.M. (1994). Effect of video game practice on spatial skills in girls and boys. *Journal of Applied Developmental Psychology*, 15, 13-32.
- Terlecki, M.S., Newcombe, N.S., & Little, M. (2008). Durable and generalized effects of spatial experience on mental rotation: Gender difference growth patterns. *Applied Cognitive Psychology*, 22, 996-1013.
- Tetris Holdings, LLC. (2009). Tetris Zone [computer software]
- Wai, J., Lubinski, D., & Benbow, C.P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, 101 (4), 817-835.
- Uttal, D.H. & Cohen, C.A., & (in press). Spatial thinking and STEM educationL When, why, and how? In B, Ross (Ed.), *Psychology of Learning and Motivation*, (Vol .57). New York: Academic Press.

# Erroneous Examples Versus Problem Solving: Can We Improve How Middle School Students Learn Decimals?

Deanne Adams<sup>1</sup> ([adams@psych.ucsb.edu](mailto:adams@psych.ucsb.edu))

Bruce M. McLaren<sup>2</sup> ([bmclaren@cs.cmu.edu](mailto:bmclaren@cs.cmu.edu))

Kelley Durkin<sup>3</sup> ([kelley.l.durkin@vanderbilt.edu](mailto:kelley.l.durkin@vanderbilt.edu))

Richard E. Mayer<sup>1</sup> ([mayer@psych.ucsb.edu](mailto:mayer@psych.ucsb.edu))

Bethany Rittle-Johnson<sup>3</sup> ([bethany.rittle-johnson@vanderbilt.edu](mailto:bethany.rittle-johnson@vanderbilt.edu))

Seiji Isotani<sup>4</sup> ([sisotani@icmc.usp.br](mailto:sisotani@icmc.usp.br))

Martin van Velsen<sup>2</sup> ([vvelsen@cs.cmu.edu](mailto:vvelsen@cs.cmu.edu))

<sup>1</sup>Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106-9660

<sup>2</sup>Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA 15213

<sup>3</sup>Department of Psychological Sciences, Vanderbilt University, Nashville, TN 37240-7817

<sup>4</sup>Department of Computer Systems, University of Sao Paulo (ICMC-USP)

## Abstract

Worked examples have been found to be effective tools in reducing cognitive load and supporting learning. Erroneous examples are worked examples that include incorrect steps and are intended to help students learn how to identify important principles and errors to avoid. The current study examines whether using erroneous examples in an online intelligent tutoring system can help middle-school children learn decimals beyond simple problem solving with feedback. Results showed that although students did not differ between the two conditions on an immediate posttest, students in the erroneous examples group performed better on a delayed posttest. This suggests that working with errors, and thus processing the decimal problems at a deeper level, helped students retain more about decimals and build upon that understanding over time.

**Keywords:** erroneous examples, math learning, computer-based tutors

## Worked Examples and Math Learning

One effective method that has been applied to mathematics education to increase learning is worked-out-examples (also called *worked examples*). Worked examples consist of a problem formation, the steps taken to reach the solution, and the final solution (Cooper and Sweller, 1987; McLaren, Lim, and Koedinger, 2008; Renkl, 2005, 2010; Renkl and Atkinson, 2010; Zhu and Simon, 1987). Worked examples may be effective because they facilitate learning by helping to manage intrinsic processing levels (i.e. cognitive processing required to learn the material presented in a lesson); decreasing extraneous processing (i.e., cognitive processing that does not support the instructional goal); and by encouraging generative processing (i.e., cognitive processing that enables deeper learning). According to the cognitive theory of multimedia learning (Mayer, 2009)

and cognitive load theory from which it is derived (Moreno and Park, 2010) learners have a limited processing capacity in working memory and every learning task has an intrinsic level of processing required to understand and learn the task. During problem solving such as mathematics, students use strategies such as means-ends analyses to solve problems, comparing the state of the problem to the goal state and trying to reduce the differences (Renkl and Atkinson, 2010). Over time they develop procedural and schematic knowledge that facilitates problem solving. Worked examples can decrease both intrinsic and extraneous cognitive processing during learning by showing the students the solution procedures to follow. The freed up cognitive resources can then be applied to understanding and eventually to automatizing the different steps in the problem's procedure.

A study by Cooper and Sweller (1987) compared learning by doing/traditional problem solving and learning from worked examples. The results showed that participants in the learning by examples group could answer transfer problems much faster than students who learned by doing although the later group actually had more practice in solving problems.

An important issue with worked examples is that although students may have freed up cognitive resources, this does not mean that the freed cognitive capacity will be used for generative processing (also called *germane processing*) which requires deeper processing of the material (Renkl and Atkinson, 2010). Students may need further assistance in fully absorbing and learning solution methods or principles. Self-explanation is one way to achieve this. Chi et al. (1989) found that good problem solvers are more likely to generate self-explanation statements while thinking out loud when reading a lesson on physics. In addition, other research has shown the importance of explicitly prompting for self-explanation

(Hausman and Chi, 2002). Explanations can therefore be used to encourage further processing of the material and increase learning.

### **Erroneous Examples and Learning**

One other proposed way to encourage deeper processing while using worked examples is to present students with incorrect (or erroneous) examples. Erroneous examples may encourage students to use more explanations since they must identify and explain to themselves why the solution is incorrect and how it can be corrected. Erroneous examples may also help students focus on each step of a solution method separately to identify where the error occurred. However erroneous examples could also place additional processing demands on learners, overloading working memory. The student may have to simultaneously represent both the correct and incorrect solution steps while searching for what is wrong in the worked example (Grosse and Renkl, 2007). Therefore, learners with low prior knowledge may be more likely to be adversely affected by incorrect examples because they would be unable to hold large chunks of new information in memory while also looking for an error. Grosse and Renkl (2007) suggest relieving this processing demand by highlighting the error. Reiss, Hellmich, and Thomas (2002) found that learners only had a .35 probability of identifying a math false argument as being false while correct arguments had a .67 probability of being identified as correct.

Yet research has shown that erroneous examples can facilitate learning of mathematics. In a study by Kawasaki (2010), 170 5<sup>th</sup> grade students were presented with either a correct or incorrect solution to a math problem by one of the participants. The teacher then explained the correct solution either contrasting the two procedures for the incorrect or displaying the correct. Students who had used similar incorrect solutions benefitted the most from the instruction with the incorrect example. Tsovaltzi et al. (2010) found mixed results for whether erroneous examples facilitated learning of fractions. For 6<sup>th</sup> graders they found that including erroneous example, especially with help, increased metacognitive skills such as answering conceptual questions. With 9<sup>th</sup> and 10 graders, on standard problem solving tests, students in the erroneous examples with help condition outperformed students in the erroneous examples without help and the no erroneous examples groups. They propose that this was due to the low prior knowledge level of the students.

Grosse and Renkl (2007) also found an effect of prior knowledge on the effectiveness of erroneous examples. College level students were taught a lesson on probability. In their first experiment half of the conditions were presented with correct solutions only while the other half were presented with both correct and incorrect solutions. For groups with both incorrect and correct solutions, half

of the participants had the error highlighted while the other half did not. The study found an interaction between the prior knowledge of the individual and the inclusion of incorrect solutions. High prior knowledge students benefitted from having both correct and incorrect solutions and scored higher on far transfer problems that did not have solution structures similar to the problems presented during the lesson. In contrast, low prior knowledge students did worse on a far transfer test when given both correct and incorrect solutions. For highlighting the error, high prior knowledge students did not benefit from having errors highlighted (presumably because they were already able to identify the error on their own). Low prior knowledge individuals did significantly better when the errors were highlighted than when they were not. Grosse and Renkl's (2007) second study replicated the prior knowledge incorrect solution interaction but also found that including errors changes the sort of self-explanation statements students made. Students made more elaborations that were error related such as identifying the error or the reasons for the error, however, students in this group also made less principle-based self-explanation. Principle-based explanations have been proven to foster learning outcomes (Renkl, 1997).

In a recent study by Isotani et al. (2011) an online tutoring system with erroneous examples was used to teach decimals to middle school students. Six commonly held misconceptions dealing with decimals were identified, such as decimals being treated as negative numbers or students treating the two sides of a decimal as separate numbers. Participants were separated into three conditions: problem solving, worked examples, and erroneous examples. During the problem solving condition students had to at least attempt to answer a problem once and were given feedback in the form of green or red lettering as to whether their answer was correct. If the student supplied an incorrect answer they could choose to have the correct answer displayed. In the worked example condition students were given a word problem in which the correct answer was given. The students were then asked to complete two sentences that described how the problem was solved and what knowledge about decimals was needed to answer the problem. Students would select responses for the two blanks in the sentence and then receive feedback from the tutor as to whether their created explanation was correct. The erroneous examples problems were similar to the worked examples except that an incorrect solution was presented. It was the job of the students to fill in the blank to generate two sentences: the first identifies the particular decimal misconception while the second sentence prompts the student to explain how the individual in the problem could correctly solve the problem.

The results uncovered no significant differences among the three groups for either immediate posttest or the delayed posttest and unlike Grosse and Renkl (2007) there

was no interaction between high and low prior knowledge and condition. One possible reason for no significant differences among the three groups is the amount of cognitive load that the sentence completion task required of the participants. Instead of focusing on the math, the students may have been devoting their cognitive processing to selecting the correct sentence portions and reading their completed sentence.

## Present Study

For the current study we have streamlined the materials from Isotani et al. (2011) to increase the focus on finding and fixing errors in erroneous examples. In particular, we simplified that design to compare problem solving to erroneous examples. This study focused on whether erroneous examples could encourage more generative processing than problem solving, even though both conditions encourage at least some problem solving (for erroneous examples: finding and fixing errors). The two groups were presented with isomorphic problems, but with different ways of interacting with those problems. The erroneous examples subjects were presented with an incorrect solution, were prompted to explain and correct the error and reflect on the correct answer, and received feedback on their responses. The problem solving subjects were asked to solve the same problems, reflect on the correct answers, and received feedback on their work. The additional steps in the erroneous examples condition of explaining and correcting the error/misconception made in each problem was intended to improve learning outcomes by encouraging learners to engage in generative processing concerning decimal principles. The problems were also simplified from Isotani et al. (2011) by

providing more complete explanations for the students to choose from. Previous research on self-explanation prompts by Johnson and Mayer (2010), demonstrated that providing the explanation statements, rather than having learners generate their own, facilitated learning from an educational game. By providing the students with possible complete explanations to choose from rather than parts of sentences, processing demands should decrease.

**Participants.** Participants consisted of 208 (Male = 101, Female = 107) middle-school students from Pittsburgh, PA. Of those students, 105 were in the 6<sup>th</sup> grade while 103 were in 7<sup>th</sup> grade. Ages ranged from 11 to 13 ( $M = 11.99$ ,  $SD = .722$ ).

**Materials.** The computer-based materials consisted of 6 components, three tests (pretest, posttest, and delayed posttest), two surveys (demographic/math experience and evaluation), and the intervention problems. For the pretest and two posttests, three separate but isomorphic tests were constructed. Question types including placing decimals on a number line, putting a group of three or four decimal numbers in order, providing the next two numbers in a sequence, and answering true/false statements. All three tests contained 46 problems with a total of 50 points possible. For the demographic survey, along with basic information about age and grade level, students were asked about their experience with decimals and computers. They were also asked a few self-efficacy questions such as, "I am good in math at school", with 5-point Likert answers, ranging from "Strongly Agree" to "Strongly Disagree." For the evaluation survey, students were

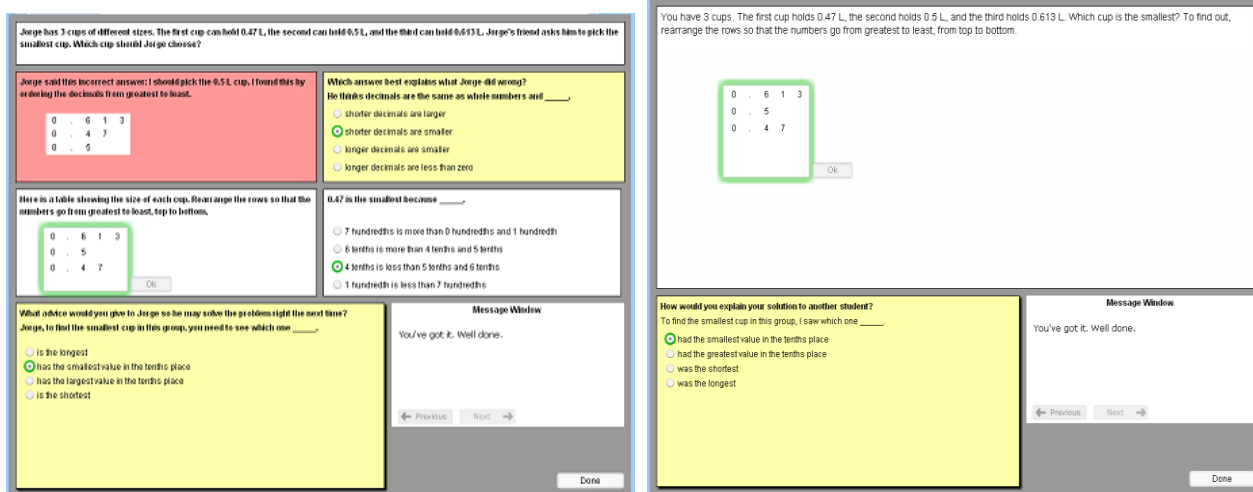


Figure 1: Side-by-side comparison of the isomorphic questions from the two intervention conditions. An erroneous example problem is on the left and the equivalent problem to solve is on the right.

asked how they felt about the intervention using a 5-point Likert scale ranging from 1 to 5. Questions included items such as, “I would like to do more lessons like this.”

During the intervention students completed a total of 36 problems, with interaction and feedback implemented by intelligent tutoring software (Aleven et al, 2009). The problems were arranged in four groups of three (with each group targeted at one of four misconception types) making a total of 12 groups. For the erroneous condition, students would first receive two problems dealing with a misconception such as “shorter decimals are smaller.” The third problem was then a problem to solve (with feedback) related to the misconception (i.e. putting decimals of different lengths in order from largest to smallest). The erroneous problems contained up to 5 components (not including the problem statement) for the students to interact with (see Figure 1 for a comparison between the two interventions). In the top left box students read the error made by the individual in the word problem. After pressing a “Next” button students were asked to identify what the subject had done wrong from a list of 3-4 options, one of which was the misconception exhibited by that student. In the left middle panel students were then asked to correct the mistake. This involved either placing the decimal correctly on a number line, changing a decimal addition, correctly ordering a list of decimals (largest to smallest or smallest to largest), or correctly completing a sequence of decimals. In the right middle panel participants explained why the new answer was correct. Finally, in the bottom left panel the students were asked to give advice to the fictional student that had gotten the answer incorrect. For every panel that required the student to make a selection feedback was provided (green = correct; red = incorrect). Students also received text feedback from a message window that was placed at the bottom right corner of the intervention. Messages include encouragement for students to try incorrect steps again or feedback for students to continue on to the next step or problem.

In the problem-solving version of the intervention, students were given the same problems as in the erroneous examples condition except they were asked to provide the solutions themselves. These problems were also arranged in groups of three with a simple correct / incorrect feedback for the third problem in each sequence. On the first two problems of the problem solving condition, after solving the problem students were asked how they would explain their solution to another student. These options included the correct procedure along with misconception distracters. Students in this group also received feedback from a message window in the bottom right panel as well as green / red feedback on their solution and multiple-choice selections.

**Procedure.** Students were randomly assigned to one of the two conditions (PS = 108; ErrEx = 100). Students in both conditions were given a total of five 43-minute sessions to complete the entire intervention. The students were randomly assigned to either the problem solving or the erroneous worked example condition. Students were also randomly assigned to receive one of the six possible pretest / posttest / delay-posttest orderings (ABC, ACB, BAC, BCA, CAB, CBA). On the second day the students answered the demographic and math/computer experience questionnaire before starting the intervention. The students were given two days to complete the problem solving/worked example problems. Upon completion they were given the intervention assessment questionnaire. The next day students were given the immediate posttest. Finally, during the following week, students were given the delayed posttest.

## Results

Due to an error in data recording for four of the problems, the data for those problems was removed from the pretest, posttest, and delayed posttest scores making the total possible score out of 46. To first examine whether the problem solving (PS) and erroneous examples (ErrEx) condition performed similarly on the pretest an independent sample t-test was conducted. It was found that the ErrEx group performed significantly better on the pretest than the PS group,  $t(206) = 3.045, p = .003$  (See Table 1 for means and standard deviations). An ANOVA revealed that there was no significant difference between the test orders,  $F(5,202) = 1.293, MSE = .057, p = .268$ .

In general students significantly improved their test performance after the intervention, regardless of condition,  $t(207) = -8.058, p < .001$ , with a mean increase of 9%. Students continued to significantly improve between the immediate and delayed posttest,  $t(207) = -8.230, p < .001$ , with a mean increase of 6%. Overall students increased their performance an average of 15% between the pretest and the delayed posttest, yielding a medium-to-large effect,  $d = .75$ .

To examine whether one condition increased learning more than the other gain scores were calculated between the pretest and posttest, pretest and delayed posttest, and posttest and delayed posttest. An ANCOVA with pretest as a covariate revealed that for the pretest-to-immediate-posttest gain did not differ significantly for the two groups,  $F(1,205) = .768, MSE = 34.97, p = .382$ . There were significant differences between the two conditions for pretest-to-delayed-posttest gains,  $F(1,205) = 9.896, MSE = 349.08, p = .002$ , and between immediate posttest and delayed posttest,  $F(1,205) = 7.027, MSE = 163.07, p = .009$ , with participants in the ErrEx condition having higher gain scores. That is, although participants



## Discussion

The results of this study show that although using erroneous examples did not facilitate learning gains for an immediate pretest, students in the erroneous group had significantly higher gains on the delayed posttest. These results suggest that students taught with erroneous examples may have had a deeper learning experience, one that helped them build upon their initial understanding of decimals to gain a deeper understanding by the time they took the delayed posttest.

Previous research by Grosse and Renkl's (2007) found that prior knowledge interacted with incorrect examples; higher prior knowledge students performed better when presented with incorrect solutions. For our study, however, no significant interaction was found between prior knowledge and condition. The data showed that both low and high prior knowledge individuals did better in the erroneous examples condition than the problem solving condition. This might have occurred because the erroneous example students, both low and high prior knowledge, were enticed to engage in more generative processing than the problem solving students, through the prompted explanation and correction of errors.

One limitation of our study is that we did not include a correct worked examples condition. The reasons for this were straightforward. First, in the present study we wanted to compare the most common ecological control condition – that of students solving problems – to the much less typical learning experience of working with erroneous examples. Second, as we revised the instructional materials from the Isotani et al (2011) study, we realized that erroneous examples and problem solving were more comparable from a cognitive load perspective. As designed, they both require active problem solving – in the case of erroneous examples, the correction step; in the case of problem solving, generating the solution from the given problem – something worked examples does not require. Renkl and Atkinson (2010) mention a reversal of the worked examples effect when students already have sufficient knowledge. Studying just the examples without any sort of active problem solving may become redundant for the students therefore decreasing the amount of mental effort they put into the lesson. Nevertheless, to compare other possible instructional approaches, in a future study we intend to include a worked examples condition.

Table 1: Test performance for the two conditions

Test score	Groups	
	Problem Solving	Erroneous Examples
	<i>M</i> <i>SD</i>	<i>M</i> <i>SD</i>
Pretest	24.68 (9.42)	28.69 (9.58)
Posttest	29.07 (9.48)	32.58 (8.95)
Delayed Posttest	31.06 (9.20)	36.23 (7.47)

in the ErrEx condition may not have scored higher on the immediate posttest, they showed superior gains when tested after the week delay.

To determine whether the intervention had a different effect for students with high prior knowledge versus those with low prior knowledge, similar to Grosse and Renkl's (2007), we conducted an additional analysis. Participants were first classified as high versus low by using a median split on the pretest scores (8-25 points for low and 26-45 points for high). This divided the groups so that there were 107 students classified as low prior knowledge and 101 as high prior knowledge. For high prior knowledge individuals an ANCOVA with pretest as a covariate revealed the participants did not differ for pretest to immediate posttest gains,  $F(1,98) = .122$ ,  $MSE = 3.76$ ,  $p = .728$ , or posttest test to delayed posttest gains,  $F(1,98) = 2.01$ ,  $MSE = 46.15$ ,  $p = .160$  (see Table 2 for means and standard deviations). There was a significant difference with ErrEx showing greater gains between the pretest and delayed posttest,  $F(1,98) = 4.75$ ,  $MSE = 76.27$ ,  $p = .032$ . For low prior knowledge individuals there was still not significant difference between pretest and posttest gains,  $F(1,104) = .489$ ,  $MSE = 28.49$ ,  $p = .486$ . However there the ErrEx condition did have significantly higher gains between the pretest and delayed posttest,  $F(1,104) = 5.21$ ,  $MSE = 265.73$ ,  $p = .025$ , and the posttest and delayed posttest,  $F(1,104) = 5.02$ ,  $MSE = 120.21$ ,  $p = .027$ . Thus, the pretest-to-delayed posttest gain was greater for the ErrEx condition for both low and high prior knowledge learners.

Table 2: Test performance for low/high prior knowledge individuals for the two conditions

Test Score	Low Prior Knowledge		High Prior Knowledge	
	PS <i>M</i> ( <i>SD</i> )	ErrEx <i>M</i> ( <i>SD</i> )	PS <i>M</i> ( <i>SD</i> )	ErrEx <i>M</i> ( <i>SD</i> )
Pretest	17.73 (3.89)	19.25 (4.01)	34.40 (5.36)	36.11 (5.03)
Posttest	24.24 (8.82)	26.80 (8.06)	35.84 (5.33)	37.13 (6.75)
Delayed Posttest	26.30 (8.55)	31.07 (7.47)	37.71 (5.96)	40.29 (4.34)

In summary, our study provides evidence that presenting students with errors that they are prompted to analyze, explain, and correct can facilitate learning decimals from a computer-based tutor.

## Acknowledgments

The U.S. Department of Education (IES), Award No: R305A090460, provided support for this research. We also thank the Pittsburgh Science of Learning Center, NSF Grant # 0354420, for technical support of our work.

## References

- Aleven, V., McLaren, B.M., Sewall, J., & Koedinger, K.R. (2009). A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education*, 19(2), 105-154.
- Chi, M.T.H., Bassok, M., Lewis, M.W., & Reimann, R. & Glaser, R. (1989). Self explanations: How students study and used examples in learning to solve problems. *Cognitive Science*, 13, 145-182.
- Cooper, G., & Sweller, J. (1987). The effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, 79, 347-362.
- Grosse, C.S. & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction*, 17(6), 612-634.
- Hausmann, R.G.H. & Chi, M.T.H. (2002). Can a computer interface support self-explaining? *Cognitive Technology*, 7(1),4-14.
- Isotani, S., Adams, D., Mayer, R.E., Durkin, K., Rittle-Johnson, B., & McLaren, B.M. (2011). Can erroneous examples help middle-school students learn decimals? In: Proceedings of the *Sixth European Conference on Technology Enhanced Learning: Towards Ubiquitous Learning (EC-TEL-2011)*.
- Johnson, C.I. & Mayer, R.E. (2010). Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior*, 26, 1246-1252.
- Kawasaki, M. (2010). Learning to solve mathematics problems: The impact of incorrect solutions in fifth grade peers' presentations. *Japanese Journal of Developmental Psychology*, 21 (1), 12-22.
- Mayer, R. E. (2009). *Multimedia learning* (2<sup>nd</sup> ed). New York: Cambridge University Press.
- McLaren, B.M., Lim, S., & Koedinger, K.R. (2008). When and how often should worked examples be given to students? New results and a summary of the current state of research. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*,. Austin, TX: Cognitive Science Society.
- Moreno, R., & Park, B. (2010). Cognitive load theory: Historical development and relation to other theories. In J.L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive Load Theory*, Cambridge: Cambridge University Press.
- Reiss, K., Hellmich, F., & Thomas, J. (2002). Individual and scholastic factors for argumentations and proofs in mathematics instruction]. In M. Prenzel, & J. Doll (Eds.), *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen. Zeitschrift für Pädagogik (45. Beiheft)*
- Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, 21, 1-29.
- Renkl, A. (2005). The worked-out examples principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning*. New York: Cambridge University Press.
- Renkl, A. (2010). Instruction based on examples. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction*. New York: Routledge.
- Renkl, A., & Atkinson, R.K. (2010). Learning from worked-out examples and problem solving. In J.L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive Load Theory*,. Cambridge: Cambridge University Press.
- Tsovaltzi, D., Melis, E., McLaren, B.M., Meyer, A-K., Dietrich, M. & Gogvadze, G. (2010). G. Learning from erroneous examples: When and how do students benefit from them? In Proceedings of the *European Conference on Technology Enhanced Learning*, LNCS vol. 6383,. Springer, Heidelberg.
- Zhu, X., and H.A. Simon. 1987. Learning mathematics from examples and by doing. *Cognition and Instruction* 4 (3),137-66.

# A Multi-Category Theory of Intention

Henny Admoni (henny@cs.yale.edu) and Brian Scassellati (scasz@cs.yale.edu)

Department of Computer Science, 51 Prospect Street  
New Haven, CT 06511 USA

## Abstract

People excel at attributing intentionality to other agents, whether in simple scenarios such as shapes moving in two dimensions or complex scenarios such as people interacting. We note that intentionality attributions seem to fall into two categories: low-level intentionality in which an observer has a theory of mind about an agent, and high-level intentionality in which an observer believes the agent has a theory of mind about something else. We introduce the terms *L-intentionality* and *H-intentionality* to refer to these attributions, respectively, and describe this division by using examples from previous research. Social robots provide a particularly good platform for evaluating the presence of different types of intentionality, and we discuss how robots can help distinguish the relationship between H- and L-intentionality, based on a number of possible models that we enumerate. We conclude by highlighting some interesting questions about intentionality in general and the interplay between H- and L-intentionality in particular.

**Keywords:** intention; animacy; computer model; human-robot interaction; robotics

## Introduction

Much research in psychology has focused on people's ability—and eagerness—to attribute intentions and animacy to simple shapes based on motion. From Michotte's (1963) and Heider and Simmel's (1944) experiments with animacy and intention to recent work decomposing intentional actions such as chasing (Gao, Newman, & Scholl, 2009), psychologists have found that intention attributions to moving shapes appear to be immediate and irresistible. Animacy is often observed in a display of simple shapes when the motion in the display cannot be explained as ordinary inanimate motion, for instance when speed and direction change without direct contact with other objects (Tremoulet & Feldman, 2000).

At the same time, evidence shows that people attribute intentions based on high-level behavioral evaluations, as well. For instance, 18-month-old toddlers can recognize and imitate intentional actions performed by adults, even if those actions are unsuccessful (Meltzoff, 1995). By pre-school age, children begin to represent others' beliefs, even when those beliefs are mistaken, in order to correctly predict a person's intentional action (Wellman, Cross, & Watson, 2001). As adults, neurological evidence indicates that a certain region of the brain is sensitive to whether peoples' motions are consistent or inconsistent with their purported intentions (Pelphrey, Morris, & McCarthy, 2004).

While abundant evidence demonstrates peoples' attributions of intentionality, the types of attributions they make seem to differ. Cues that prompt intention attributions come in two categories: low-level, perceptual cues, such as motion, and high-level cues that must be reasoned about, such as facial expression. To distinguish between intentions cued

in these different ways, we introduce two novel terms, referring to intention attributions made from low-level cues as *L-intentionality* and to attributions made from high-level cues as *H-intentionality*. To date, little work has explored such categorical differences of intentionality. In the Types of Intentionality section, we use examples from previously published research to define our hypothesis that L-intentionality and H-intentionality are separate kinds of intention attributions.

Robotics has provided a valuable experimental platform to test perceptions of intentionality. Because robots are extremely flexible (in terms of appearance, motions, sounds, and so on), researchers can manipulate specific variables of a human-robot interaction to test specific features of intentionality attributions. In the Social Robots as Experimental Platforms section, we describe past work with robots and other computational models of intentionality, and we discuss the benefits social robotics can offer intentionality research.

The next section, Models of Intentionality, enumerates possible models for the relationship between H-intentionality and L-intentionality based on the hypothesis that these are distinct observations. We describe what each model implies about real-world intentionality attributions, and we note how each model can be tested to confirm or deny our hypothesis.

We conclude this paper by discussing some likely starting points for research on the different categories of intention, and describing some interesting questions about intentionality that have yet to be addressed.

## Types of Intentionality

In this paper, we define an intentional action as a goal-directed action that is performed deliberately. Intentionality is the capacity to express or perform intentional actions. A theory of mind for other agents enables us to attribute intentionality to those agents (Leslie, 1987; Baron-Cohen, 1995), an ability that develops early in life (Meltzoff, 1995). Note that for our purposes, animacy and intentionality are strongly correlated, in that it is impossible to attribute animacy without the presence of intentional, goal-directed behavior (Tremoulet & Feldman, 2006).

In this section, we distinguish L-intentionality and H-intentionality as distinct but related categories. We can define each category by how an observer perceives and recognizes intentionality. To put the categorical difference simply, L-intentionality in an agent involves an observer having a theory of mind for that agent; H-intentionality involves an observer believing that the agent has a theory of mind for something else (Figure 1).

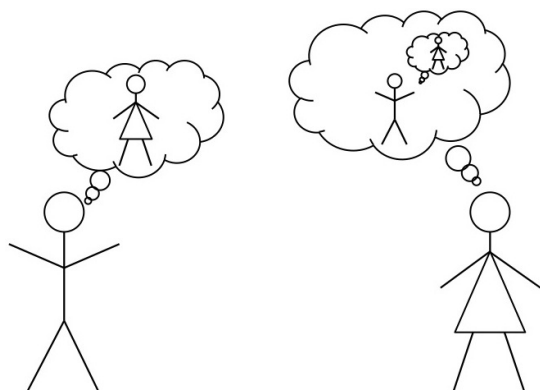


Figure 1: Low-level intentionality (left) is attributed when an agent has a theory of mind for another agent. High-level intentionality (right) is attributed when an agent believes another agent has a theory of mind for something else.

### L-Intentionality

To illustrate the different types of intentionality, picture a preschool boy named Billy. Billy is good at putting away his toys: he readily brings them to the toy chest whenever it is time to clean up. This kind of goal-directed behavior—carrying toys directly to the toy chest and depositing them—involves a series of coordinated actions and knowledge, but it can occur independently of theories of mind about others in the environment. Observing this, we might attribute L-intentionality to Billy. (Assume for the sake of the example that we cannot infer that Billy is H-intentional based solely on what we know about human beings.)

Actions that cue L-intentionality reveal goal-directed, deliberate behavior. L-intentionality is often elicited from low-level perceptions, such as those arising from visual displays of moving shapes. The perception of intentionality and animacy from simple moving shapes has been well-established in psychology literature (see (Scholl & Tremoulet, 2000) for a review); even when people do not know the type of intentional action they are looking for, they show high validity with the ground truth and high reliability with each other when evaluating the motion of animated shapes (Pantelis et al., 2011). Most often, these shapes exhibit basic actions such as chasing, fighting, or foraging.

For example, in Gao et al. (2009), an animated “wolf” chases after an animated “sheep” by moving toward the sheep within some degree of direct heat-seeking behavior. When the degree of chase is sufficiently small, participants identify chasing consistently. When the wolf deviates more strongly from direct heat seeking, however, the perception of intentionality disappears. In these L-intentionality experiments, goal-directed motion leads to an attribution of intentionality.

### H-Intentionality

Now imagine that we observe Billy hiding his shoes inside the toy chest. He watches his mother place his shoes in a cov-

ered cubby. While his mother looks away, Billy retrieves the shoes, walks to the chest, looks to make sure his mother is not watching and puts his shoes in the closed chest. The action of watching his mother put the shoes in one place, and then covertly moving them to another, suggests that Billy has a theory of mind for his mother, understanding that she has her own (mistaken) beliefs; Billy is displaying H-intentionality.

While L-intentionality is based on a theory of mind for the agent in question, H-intentionality is based on that agent’s theory of mind for others. Therefore, H-intentionality is seen in more complex visual scenes than the simple moving shapes of L-intentionality; it is often be cued by a combination of stimuli such as facial expression, vocal prosody, and physical motion. H-intentionality does not “pop out” in the same way as L-intentionality because attributions of H-intentionality require additional cognitive processing to account for the agent’s beliefs about its environment.

## Social Robots as Experimental Platforms

Experiments with human-robot interaction are a particularly rich source of intentionality attributions. In one experiment, a robot received greater attributions of animacy and intelligence when it cheated at a game of rock-paper-scissors than when it played the game correctly (Short, Hart, Vu, & Scassellati, 2010). In each round of the game, a human participant and the robot both selected an item (rock, paper or scissors) in secret, then simultaneously displayed their selection through hand signals. Each of the items loses to one other item and wins over one other item, so that one’s performance in the game depends on the opponent’s selection as well as one’s own. In conditions where the robot verbally cheated—declaring “I win!” when it had lost—participants tended to report that the robot was broken and less intelligent; in conditions where the robot physically cheated—changing its losing hand signal after viewing the participant’s selection—participants were significantly more likely to use active verbs when describing the robot. In other words, a small change in behavior (switching hand signals) led to a dramatic increase in intentionality attributions.

Social robots are a valuable platform for experiments involving intentionality. Robots are available in a huge variety of appearances (from anthropomorphic to simple shapes) and motion abilities (fully mobile to stationary; with a broad range of physical capabilities). Being programmable, robot behavior can be carefully manipulated to alter individual features (such as moving an arm at a particular speed) and to ignore subtle (and potentially subconscious) social cues from others, a feat that a human experimenter might find difficult. As machines, robots can perform exactly the same action again and again, but as social tools, robots appear socially neutral: while most participants in experiments from our laboratory have seen or heard about robots, most have no experience with actual robot capabilities, allowing robots to act a blank slate on which social characteristics (such as intentionality) can be drawn at will.

Animations or videos of people provide many of the same benefits as robots, but they lack an embodiment that may affect interactions. Research has shown that people follow commands more consistently from physically present robots than from virtual robots, even when the virtual robot looks and acts exactly like the embodied robot (Bainbridge, Hart, Kim, & Scassellati, 2011). This reflects the common wisdom of sales, which asserts that you should visit someone in person to close a deal. For intentionality research, in which subtle features may make a difference in intention attributions, having an embodied agent observed in real time allows for highly realistic experimental setup while maintaining strict control over experimental variables.

H-intentionality has been of particular interest to social robotics researchers, who are motivated to design human-robot interactions that are natural and communicative. One part of natural interaction is identifying and displaying a theory of mind for others through non-verbal intentional behavior, whether with gaze (Mutlu, Yamaoka, Kanda, Ishiguro, & Hagita, 2009), hand gestures (Nehaniv, Dautenhahn, Kubacki, Haegle, & Parlitz, 2005), or facial expression (Breazeal & Scassellati, 1999). Therefore, understanding intentionality is important for the design of robots that will interact with people, such as service or assistive robots.

### Models of Intentionality

In this section, we enumerate the possible relationships between H-intentionality and L-intentionality (Figure 2); in the following Discussion section, we explain which models we believe are most viable. To better describe the models, we will return to the previous example with Billy and his toys.

Researchers have attempted to computationally recognize intentions through Bayesian models (Baker, Tenenbaum, & Saxe, 2006; Schrempf, Albrecht, & Hanbeck, 2007), hidden Markov models (Aarno & Kragic, 2006), and algorithmic methods (Feldman & Tremoulet, 2008). As observed actions and possible intentions become more complex, specifying a reasonably-sized state space for intention-action mapping becomes increasingly challenging, which limits the power of current computational models of intention recognition. The models in this paper are intended to be abstract, high-level views of how intention attributions can be conceptually organized, not algorithmic specifications for functional programs.

The descriptions of the models in this section are based on *features of intentionality*—observations that can be empirically measured. For instance, goal-directed movement toward another agent in an approximately heat-seeking manner (as in Gao et al. (2009)) is one feature of chasing, an L-intentional action. Behavior based on anticipation of others' responses, as when Billy looked to see whether his mother was watching him hide his shoes, is a feature of H-intentional actions. Many features for each type of intention have yet to be identified, and are part of the novelty of this area of research.

Of all of the models for the relationship between H- and L-intentionality, the simplest option is that there is no cor-

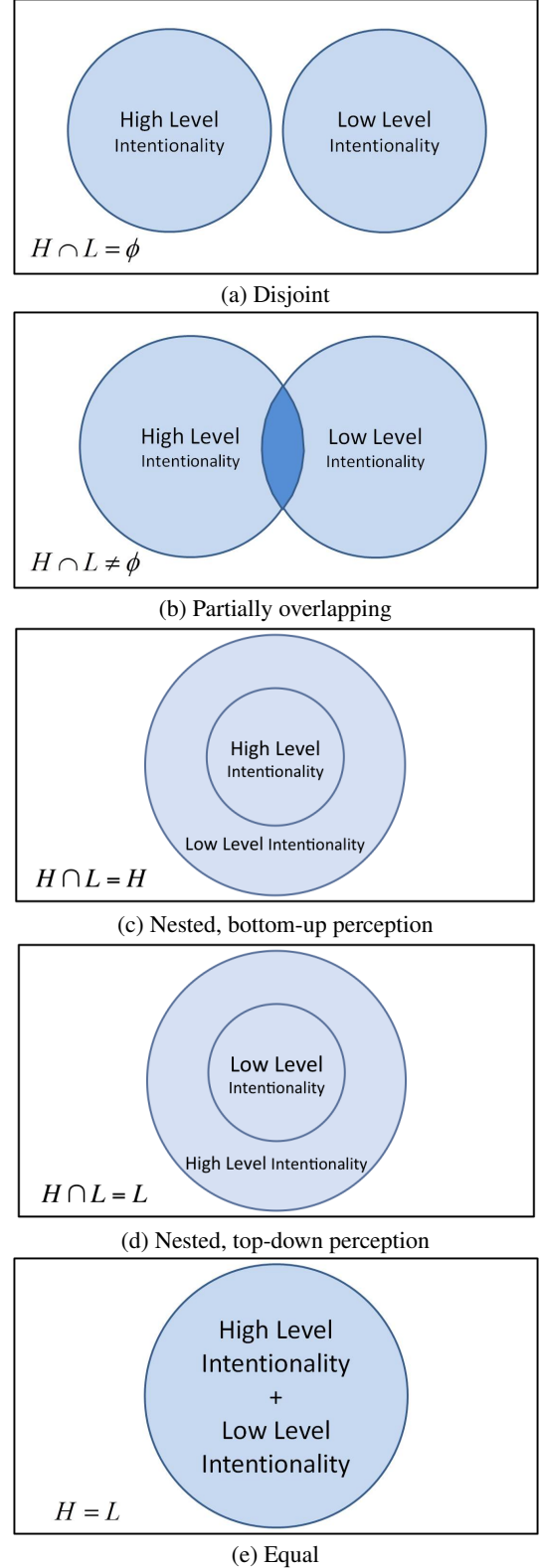


Figure 2: Potential models of the interaction between H- and L-intentionality. Each labeled circle represents the set of features that cue that type of intentionality. Set notation below each image mathematically describes the relationship between  $H$ , the set of features that cue H-intentionality, and  $L$ , the set of features that cue L-intentionality.

relation between the two (Figure 2a). In this model, the two types of perceptions share no features. Seeing Billy put his toys in the chest (and subsequently attributing L-intentionality to Billy) would not affect later judgements about H-intentionality, and vice versa. In the disjoint case, recognizing a feature of one type of intentionality would immediately identify the type of intentionality present, because that feature could not correspond to the other type of intentionality. To falsify this model, researchers must show that both H- and L-intentionality are cued by some feature.

A second possibility is that the two types of intentions share some features, but do not overlap completely (Figure 2b). In this case, seeing shared features would cue both types of intentionality attributions, though not every feature would be a shared feature. Identification of shared features would not confirm which type of intention is being perceived, but identification of disjoint features would allow experimenters to pinpoint the type of intention present. To prove that this model is correct, researchers would have to identify one feature that is unique to each type of intentionality, and one feature that is shared by both types.

Another possibility is that the types of intention are nested, so that one is wholly contained in the other. In the first form of nesting, which we'll call bottom-up perception, H-intentionality features are a strict subset of L-intentionality features (Figure 2c). In this case, L-intentionality is cued whenever H-intentionality is perceived, though the former can be present without the latter. In our example with Billy, merely perceiving H-intentionality—from watching Billy hide his shoes—would be enough to perceive L-intentionality, even without watching Billy put away his toys. This model can be falsified by identifying a scenario in which H-intentionality but not L-intentionality is perceived.

The complementary model posits that L-intentionality features are a strict subset of H-intentionality features (Figure 2d). We call this top-down perception, because the more complex H-intentionality can be cued without L-intentionality. In this model, H-intentionality is automatic whenever L-intentionality is perceived. This model can be falsified by identifying a feature of L-intentionality that does not cue H-intentionality.

A final possibility is that the feature sets of H-intentionality and L-intentionality are equal; that is, there is no difference between the features involved in cueing each type of intentionality. In this model, H-intentionality arises whenever L-intentionality is cued, and vice versa. To falsify this model, researchers would need to identify at least one feature from one type of intentionality that does not cue the other. Failing to falsify this model would challenge the hypothesis that H-intentionality and L-intentionality are separate processes.

### **The Importance of Intentional Duality**

Researchers have long established that people attribute intentionality to other people and to simple moving shapes under some conditions. The existence of H- and L-intentionality as

distinct types of intention attributions, if proven, might indicate a categorical distinction that runs more deeply in peoples' cognitive systems. H-intentionality and L-intentionality may not only be perceived differently; they may be understood and processed in different ways as well. After all, if the perceptual pathways for the two types of intentionality are different, perhaps the cognitive pathways that process them are also different. Perhaps there are distinct brain regions or neural pathways that process L-intentional and H-intentional stimuli. Perhaps, even, there is a different developmental time course for each type of intentionality, and infants can perceive one type of intentionality before the other. This difference might even extend to the perception of one agent as more complete or more animate than another, based on the type of cues that were used to establish its intentionality. All of these possibilities are consequences of our two-intention hypothesis, and will need to be further explored.

Understanding intentionality is also important for designing human-robot interactions. For robots that must interact naturally with people, being able to both recognize and perform intentional behavior is essential. To date, most computational approaches for intention recognition rely on statistical or probabilistic methods that do not scale well with increasing actions and intentions. Our model is a first step toward comprehensive understanding of intentionality that may lead to more complete and flexible computational models.

If cognitive differences underlie the different kinds of intentional attributions, these differences can be manipulated in interesting ways for human-robot interactions. Because robots are programmable, they can be made to display only cues from one type of intention, leading an observer to cognitively categorize them in a particular way. For instance, suppose researchers establish that unfamiliar L-intentional agents do not invoke as much shyness or fear as H-intentional agents, by virtue of their apparently less complicated mental structure. Robots that interact with children can then be manipulated to display only L-intentionality cues, reducing the likelihood that children will be afraid of them by controlling how they are perceived. The ability to craft human-robot interactions in a completely unprecedented way becomes possible if the categorical difference between intention types extends to cognitive processing levels.

### **Discussion**

Based on our distinction between H- and L-intentionality, some models are more likely than others. Clearly, a model in which the feature sets for both intentionality types are identical is impossible if we are to maintain the distinction between the two. We previously defined attributing H-intentionality to an agent as believing that the agent has a theory of mind for some other target. Inherent in this definition is the idea that we also believe the agent in question has its own goals and is capable of intentional actions to achieve those goals; in other words, that we have a theory of mind for that agent. Therefore, L-intentionality seems to be inherent in attribu-

tions of H-intentionality. For this reason, the disjoint model and the overlapping model seem unlikely candidates for our purposes. If L-intentionality is inherent in H-intentionality, then it would be impossible for the latter to be perceived without the former, which rules out the top-down model. In fact, we believe that the bottom-up model (Figure 2c), in which all features of H-intentionality are also features that cue L-intentionality, has the most promise as a model of intention attributions. In this model, L-intentionality can be present on its own—as supported by the many intention-from-motion studies described in the Introduction—but the presence of H-intentionality presupposes the presence of L-intentionality. For completeness, we have listed all possible models in this paper, but only the bottom-up model really serves the distinction we draw between H- and L-intentionality.

Though we distinguish between H- and L-intentionality, we have not made any claims about how these types of intentionality might be treated differently once they are perceived. Because our distinction is novel, we do not yet have evidence to ascertain whether or not H-intentionality affects peoples' reasoning differently from L-intentionality. The characteristics of intentionality described here might vary based on the type of intentionality that is perceived.

One characteristic that has yet to be explored is whether intentionality is revokable. Though many experiments identify the presence of intentionality, very few explore conditions under which intentionality disappears. It is possible that once intentionality is perceived, it remains for the duration of the exposure; on the other hand, perception of intentionality might also fade or be removed through some mechanism, such as time since the last intentional action. Anecdotal evidence from our lab suggests that attributions of H-intentionality are hard to remove, while attributions of L-intentionality may not be. In the rock-paper-scissors experiment described earlier (Short et al., 2010), once participants made attributions of H-intentionality to the robot, those attributions persisted despite the absence of additional H-intentional behavior. On the other hand, in the chasing studies (Gao et al., 2009), it seems easy to imagine that an agent that no longer chases would stop appearing L-intentional after some time.

Philosophers have identified yet another interesting characteristic of intentionality attributions: actions with harmful side-effects seem to be perceived as more intentional than actions with beneficial side-effects. For instance, if a company owner institutes a new manufacturing process that increases profits but damages the environment, the owner is seen as intentionally harming the environment. On the other hand, if the owner institutes a process that increases profits and helps the environment, the owner is not seen as intentionally helping the environment (Knobe, 2005). Does this type of disparity hold for both types of intentionality? Would a side-effect from an L-intentional action be seen as equally harmful to the same side effect from an H-intentional action? Or does the complexity of reasoning perceived in H-intentional agents endow them with more responsibility for side effects?

Along the same vein, observing intentional actions from an outsider's perspective may be different from identifying intentionality based on actions that personally affect one's self. Are attributions of intentionality different based on whether the effect of the action is personally relevant? Does it matter if the personally relevant effect is positive or negative? Are there differences between H-intentionality and L-intentionality in these cases?

Animals, like robots, present an interesting boundary condition for intentionality. Animals are clearly sentient and are generally attributed beliefs and goals. Whether animals can be said to be intentional, however, is open to debate (Heyes & Dickinson, 1990). It would be interesting to determine whether animals fall into the L-intentional or H-intentional category, and whether this separation of intention types might make it easier to attribute intentionality to animals.

Descriptions of models in the previous section refer to features that cue intentionality, but the precise nature of these features is still to be determined. These features need not be exclusively visual; they might include auditory features like prosody or kinesthetic features like heat. Some features, such as “approaches a target with a velocity between  $x$  and  $y$ ” may be quite specific; these are often used in rule-based intention detectors, but lack flexibility to account for novel or stochastic situations. Other features might be more general, such as “orients eyes toward the target.” Enumerating and sorting these features into H- and L-intentionality cues is a significant task, but it has the potential to dramatically increase understanding about intention perception.

Identifying and evaluating features of intentionality will require carefully designed experiments, and robots as intentional actors may be useful in teasing apart attributions of intentionality. Establishing features of intentionality will also help those who design interactions between people and objects, such as those working with social robots. By understanding features of intentionality, roboticists will be able to design robots that detect and exhibit relevant intentional behaviors, which would strengthen non-verbal communication in human-robot interactions.

## Conclusions

This paper presents a new representation and vocabulary for classifying different types of intentionality. Using examples from the extensive psychology literature on intention recognition, we hypothesize that intention attributions can be categorized into two types, L-intentionality and H-intentionality, based on the kinds of perceptions that cue those attributions. We describe the benefit of social robotics as a platform for experimenting with intentionality perceptions, and we mention some past research from robotics that explores intention attributions under various conditions. We then outline possible models for the relationship between H- and L-intentionality based on the set of features that elicit the perception of each: completely disjoint, partially overlapping, nested with H-intentionality as a proper subset, nested with



L-intentionality as a proper subset, and identical. Along with each model description, we specify how that model could be falsified. We posit some important consequences of proving our intention duality hypothesis, and we discuss the potential validity of these models, identifying that bottom-up processing is the most likely model given our separation of H- and L-intentionality. We discuss characteristics of intentionality that might vary between H- and L-intentionality, and we pose questions for future exploration of this area of research.

## Acknowledgments

Thanks to Greg Trafton for inspiring conversations, to Brad Hayes for help with diagrams, and to anonymous reviewers for helpful comments. This material is based upon work supported by grants from the National Science Foundation under contracts No. 1139078, No. 1117801, and No. 0835767. The first author is supported by an National Science Foundation Graduate Research Fellowship.

## References

- Aarno, D., & Kragic, D. (2006). Layered HMM for motion intention recognition. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS 2006)* (pp. 5130–5135). Beijing, China.
- Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, 3, 41–52.
- Baker, C. L., Tenenbaum, J. B., & Saxe, R. R. (2006). Bayesian models of human action understanding. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems 18*. MIT Press.
- Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.
- Breazeal, C., & Scassellati, B. (1999). How to build robots that make friends and influence people. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS '99)* (Vol. 2, pp. 858–863). Kyongju, South Korea.
- Feldman, J., & Tremoulet, P. D. (2008). *The attribution of mental architecture from motion: Towards a computational theory* (Tech. Rep. No. 87). Rutgers University Center for Cognitive Science (RuCCS).
- Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, 59, 154–179.
- Heider, F., & Simmel, M. (1944, April). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2), 243–259.
- Heyes, C., & Dickinson, A. (1990). The intentionality of animal action. *Mind and Language*, 5(1).
- Knobe, J. (2005, August). Theory of mind and moral cognition: exploring the connections. *Trends in Cognitive Sciences*, 9(8).
- Leslie, A. M. (1987). Pretense and representation: The origins of “theory of mind”. *Psychological Review*, 94(4), 412–426.
- Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31(5), 838–850.
- Michotte, A. (1963). *The perception of causality*. Oxford, England: Basic Books.
- Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., & Hagita, N. (2009, March). Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior. In *Human robot interactions (HRI'09)*. La Jolla, California: ACM.
- Nehaniv, C. L., Dautenhahn, K., Kubacki, J., Haeghele, M., & Paritz, C. (2005). A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction. In *Proceedings of the IEEE international workshop on robot and human interactive communication (ROMAN 2005)* (pp. 371–377).
- Pantelis, P. C., Cholewiak, S., Ringstad, P., Sanik, K., Weinstein, A., Wu, C.-C., et al. (2011). Perceptions of intentions and mental states in autonomous virtual agents. *Journal of Vision*, 11(11), 1990–1995.
- Pelphrey, K. A., Morris, J. P., & McCarthy, G. (2004). Grasping the intentions of others: The perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *Journal of Cognitive Neuroscience*, 16(10), 1706–1716.
- Phillips, W., Baron-Cohen, S., & Rutter, M. (1998). Understanding intention in normal development and in autism. *British Journal of Developmental Psychology*, 16, 337–348.
- Scholl, B. J., & Tremoulet, P. D. (2000, August). Perceptual causality and animacy. *Trends in Cognitive Sciences*, 4(8).
- Schrempf, O. C., Albrecht, D., & Hanbeck, U. D. (2007). Tractable probabilistic models for intention recognition based on expert knowledge. In *Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS 2007)* (pp. 1429–1434). San Diego, CA.
- Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair!! an interaction with a cheating robot. In *5th ACM/IEEE international conference on human-robot interaction* (pp. 219–226).
- Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, 29, 943–951.
- Tremoulet, P. D., & Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception and Psychophysics*, 68(6), 1047–1058.
- Wellman, H. M., Cross, D., & Watson, J. (2001, May/June). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3), 655–684.

# A Narratological Approach for Narrative Discourse: Implementation and Evaluation of the System based on Genette and Jauss

**Taisuke Akimoto (g236i001@s.iwate-pu.ac.jp)**

Graduate School of Software and Information Science, Iwate Prefectural University, 152-52 Sugo  
Takizawa, Iwate 020-0193 Japan

**Takashi Ogata (t-ogata@iwate-pu.ac.jp)**

Faculty of Software and Information Science, Iwate Prefectural University

## Abstract

This paper proposes a computational system of narrative discourse generation and its implementation. In the system, Genette's discourse theory is reconstructed as discourse techniques which transform the tree structure for a story into discourse structures. Also, we introduce Jauss's reception theory to construct the control mechanism, which continues discourse generation through generation cycles based on the interaction between both narrator and narratee mechanisms. Moreover, we attempt two kinds of performance checks and two types of evaluation experiments and confirmed that the system generates diverse discourse structures on the rough correspondence with generative parameters. And furthermore, this study shows that two different types of literary knowledge are organically integrated into a system's framework.

**Keywords:** Narrative generation system; narrative discourse; story; narratology; Genette; Jauss.

## Introduction

The research of narrative generation system is a challenging theme in artificial intelligence and cognitive science. It has a close relationship to various topics such as problem solving, planning, schema, story grammar, natural language generation, creativity, etc. Moreover, in recent years, interdisciplinary approaches with narratology and literary theories are also emerging. We have proceeded on a narrative generation system based on this kind of mixed approach since early 1990s. A common framework for the narrative generation system (Ogata, 1994; Ogata & Kanai, 2010; Akimoto & Ogata, 2011) consists of three stages: story, discourse, and surface representations (by language, animated movie, and music). Story is the content or a temporal sequence of events to be narrated, and discourse means how to organize a story or a narrated structure of events. They are generated as the conceptual representation forms or deep structures of narrative. Therefore, discourse phase does not equal natural language generation phase. The discourse in this paper especially means the internal structure of narrative representation. For example, many of the objectives treated by Callaway and Lester (2002) belong in natural language generation phase in the architecture of our narrative generation system. This paper deals with the part of discourse and proposes a computational model of structural narrative discourse processing and its implementation. As a fundamental standpoint, we use the discourse theory of Genette (1972). In addition, reception theory of Jauss (1970) is introduced into the system to control the generation and transformation of narrative discourse structure. First, this paper introduces the system architecture. And second, we present results of the system's evaluations, which focuses on the correctness of structure transformation and the control mechanism based on the interaction

between narrator and narratee inside the system. Last, the problems and future directions are discussed.

In the area of researches on narrative generation system, there is no attempt that utilizes Jauss's reception theory. Moreover, most of previous systems (e.g., TALE-SPIN by Meehan (1980), BRUTUS by Bringsjord and Ferrucci (2000), and so on) focused on the aspect of "story" generation mainly. However, recently, Montfort (2007) applied Genette's discourse theory to develop an interactive fiction system, and Lönneker-Rodman (2005) introduced the category of "voice" in Genette theory into the conceptual design of natural language generation system. As stated above, the computational application of Jauss provides an original design which can be not comparable in other narrative generation systems. And, the introduction of Genette has the character based on systematic and comprehensive design more than the other attempts. Such introduction of the knowledge in literary area contributes to narrative generation system and artificial intelligence regarding the providing of more precise and pragmatic domain specific knowledge and can guide the exploration for the developing computational techniques in creative areas. Especially, we show that two different and separate narratologies are organically integrated into one computational mechanism. This is a worthy contribution that the introduction of narratology into computational simulation has.

## Genette's Narrative Discourse Theory

Gérard Genette (1930-, France) is a representative literary theorist and narratologist mainly associated with structuralism. The discourse theory by Genette (1972) comparatively clearly categorizes various types of discourse techniques through the analysis of a novel. The theory consists of following three broad categories: "tense" relevant to the relationship between story's time and discourse's time, "mood" relevant to the modality for regulating narrative information, and "voice" relevant to the relationship among narrating, story and discourse. Each category is further divided into many subcategories. In the proposed system, discourse techniques are mainly relating to both categories for tense and mood.

## Jauss's Reception Theory

Reception theory is one standpoint in modern literary theories and narratology, which focuses on the reception or reading process of literary works. In this theory, readers contribute strongly to the production process of literary works as a whole. Hans Robert Jauss (1921-1997, German) is a representative theorist of this area by proposing an idea to characterize literary history based on the concept of "horizon of expectation", which means a kind of previous knowledge for positioning a new work on the context of readers' experiences of reading. Artistic character of a new work is grasped by the disparity between the

given horizon and the work, and the appearance of a new work may result in the change of an old horizon. We grasp this theory as a model which literary works are continuously changing through the interaction between authors and readers.

### Proposing a Narrative Discourse Mechanism

We propose a narrative discourse system using both ideas of Genette and Jauss. This system is intended to be positioned in the part of narrative discourse in the common framework for the narrative generation system (Ogata, 1994; Ogata & Kanai, 2010; Akimoto & Ogata, 2011). In the proposed discourse system, each category in Genette theory is elaborately formalized as a discourse technique for transforming a story structure or the part into a discourse structure, and Jauss theory is simply interpreted as a mechanism in which above discourse construction process is controlled through the interaction between narrator mechanism with generative parameters and narratee mechanism with expectation parameters.

These narrator and narratee do not mean real existences but virtual agents inside the system. In the current implementation, both narrator and narratee is individual model. However, reception theory covers both individual model and collective one, and we should consider other possibilities about the concepts in the future. For example, there are multiple models such as the narrator as an individual & the narratees as multiple individuals, and the narrators as multiple individuals & the narratee(s) as a collection. Our narrative generation research is an exploratory approach through the incremental revision of a variety of elements or modules and the integration and a flexible framework for the step-by-step expansion and conversion is prepared.

Following cycle continues according to the interaction of narrator and narratee. The narrator mechanism performs the processing of discourse generation and transformation using discourse techniques and a set of rules for controlling the application based on generative parameters. On the other hand, the narratee mechanism evaluates the result based on the comparison of expectation parameters and generative ones. In the next cycle, referencing the narratee's evaluation, the narrator tries to do the generation in an effort to come close to the narratee's expectation or higher degree of the satisfaction. However, the processing eventually comes at a point where the narratee's satisfaction turns to fall from rise or the narratee gets tired. In such timing, the narrator abandons a part of old generative parameters ("deviation") and moves to a new cycle of discourse generation according to a new strategy, and narratee's expectation is also altered.

As this process is a principled and elaborate computational application based on the idea and concept of reception theory, it is characterized as a comprehensive and general control mechanism for narrative generation system to be able to be expanded to other narrative generation stages such as story generation and natural language generation.

### Structural Representations for Story and Discourse

Both structures for a story and a discourse have a same tree form. In the story tree, each leaf node is corresponding to an event described with conceptual representation, which is really described by a case frame consisting of one verb concept and eight kinds of cases such as agent and object. Each internal node in the story tree is corresponding to a "relation" combining with the child nodes like "cause-effect" and "serial". On the

other hand, a discourse is described as a tree structure transformed from a story tree. And next seven kinds of relations are used for only the discourse tree: "recall", "present-backward", "prophecy", "present-feature", "episode", "description", and "repetition\_discourse".

### Discourse Techniques

In computational perspective, since each technique for discourse by Genette is respectively corresponding to a type of discourse structure outputted from an input story, the process is able to define with a kind of transformation procedure. For the procedural definition of techniques, we prepare next five kinds of procedural primitives for operating any intermediate or terminal node in the input structure: deletion, copy, conjunction, substitution, and creation. Current version of the system has 13 kinds of discourse techniques using the primitives as shown in Table 1. Although techniques for tense cover the main part, a few techniques for mood are also contained. Figure 1 shows the operation of "complementary analepsis\_ellipsis" as an example of transformation.

### Control Mechanism

By reference to the comparatively vague description about effects of discourse techniques by Genette (1972), we originally defined discourse parameters including  $p_1$ :supplement,  $p_2$ :complexity,  $p_3$ :suspense,  $p_4$ :length,  $p_5$ :hiding,  $p_6$ :descripttiveness,  $p_7$ :repetition,  $p_8$ :diffuseness,  $p_9$ :implication, and  $p_{10}$ :temporal-independency. These parameters are associated with the feature and the effect of constructed discourse structures, and are used for generative goals for narrator and expectations for narratee. Each parameter takes the value of 1 (small), 2 (medium), or 3 (large). Moreover, we defined quantitative criteria for measuring the degree of attainment of each parameter in a generated discourse. These criteria are not based on the cognitive effects for recipient, but structural features which can be calculated from the number and order of specific leaf/internal nodes in the tree structure of discourse. For example, "length" is measured by the total number of leaf nodes in a discourse structure. The quantitative criteria are used for the system's evaluation experiments in the following section. And also, as mentioned later, the narrator decides discourse techniques to be applied based on the rules for selecting techniques by values of the generative parameters. These rules are defined according to the correlation coefficient between each generative parameter and each measured value using the above criteria.

Figure 2 shows the overview of control mechanism. The list of Table 2 is the explanation of important terms used in the process. For the process, an input story is given by user or previous story phase. Other necessary data are the saturation point in the degree of satisfaction ( $n_p$ , 1 or more), the number of generation cycles (1 or more), and some kinds of initial values including generative parameters, expectation parameters, the degree of desire in narratee (0 or more), and the number of sufficiency in narratee (0 or more). According to the input data, system repeats following five steps.

**(1) Selection of Techniques** Narrator decides techniques to be applied according to generative parameters and rules for selecting techniques to be used. These rules define 0 or more techniques corresponding to each parameter's value, such as [If "supplement" is 1 then nothing, 2 then "external analepsis",

and 3 then “external analepsis” & “external prolepsis”]. When a same kind of technique is selected by more than one rule, the narrator takes the number of times of this technique to be used, from one rule which has the greatest number of the technique.

Table 1: 13 kinds of discourse techniques

<b>External analepsis:</b> Narrating past events which are positioned outside of story’s time range (i.e., not contained in the story).
<b>Complementary analepsis_ellipsis:</b> Narrating past events which are lacked of the original position.
<b>Complementary analepsis_paralipsis:</b> Narrating past events which are partially lacked of the original position.
<b>Repetitive analepsis:</b> Narrating past events once more.
<b>External prolepsis:</b> Narrating prospective events which are positioned outside of story’s time range (i.e., not contained in the story).
<b>Complementary prolepsis_ellipsis:</b> Narrating prospective events and these events are lacked of the original position.
<b>Complementary prolepsis_paralipsis:</b> Narrating prospective events and these events are partially lacked of the original position.
<b>Repetitive prolepsis:</b> Narrating prospective events and these events are narrated at the original position once more.
<b>Achronie:</b> Narrating events which have unidentified temporal relation with time of story.
<b>Pause:</b> Pausing temporal progress of the story by inserting descriptions.
<b>Implicit ellipsis:</b> Skipping one part of story.
<b>Repeating:</b> Narrating same events twice.
<b>Paralipsis:</b> Narrating less information than original sequence of the events.

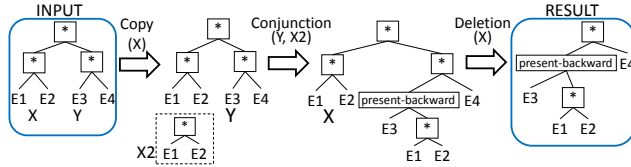


Figure 1: The transformation process of “complementary analepsis\_ellipsis”

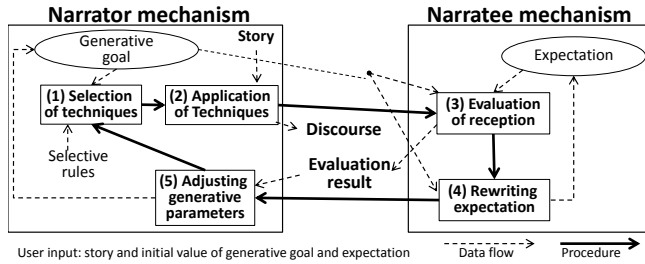


Figure 2: The overview of control mechanism

Table 2: Important terms in the control mechanism

Generative goal	Discourse parameters which represent the narrator’s direction of discourse generation.
Expectation	Discourse parameters which represent the expected discourse features by the narratee. Each parameter has two kinds of attributions which are “degree of desire” and “number of sufficiency”.
Degree of desire	This is numerical value and it represents the strength of expectation which is represented by the value of the parameter.
Number of sufficiency	This is number of time the parameter was sufficed. Suffice means the accordance of a value of generative parameter and a value of expectation parameter.
Degree of satisfaction	Degree of satisfaction in narratee’s each expectation parameter to narrator’s generative goal.
Indication	An annunciation to the narrator about the parameter which was the least “degree of satisfaction”.
Getting tired	The narratee is boring with the expected discourse, namely “degree of satisfaction” is decreased.
Deviation	The narrator intentionally sets a generative goal which counters the expectation of the narratee.

(2) **Application of Techniques** Narrator mechanism applies all selected techniques to the structure. The interference among techniques sometimes arises. For example, an inserted node by “repeating” is removed by “implicit ellipsis”. For avoiding such phenomena, we determined the order of priority that techniques are applied. Removal techniques like “implicit ellipsis” have higher priority than additional techniques like “pause”. The position for a technique to be applied is decided using some heuristic constraint rules relevant to the node’s size mainly. For instance, the large internal nodes containing more than 7 leafs can not be the position “implicit ellipsis” is applied.

(3) **Evaluation of Reception** First, the narratee calculates the degree of satisfaction in every expectation’s parameter. Here, higher satisfaction can be obtained when a more strongly desired parameter is satisfied. Next, narratee calculates and indicates one parameter with the lowest satisfaction. If two or more parameters have the lowest satisfaction, the parameter which has smaller subscript number will be selected. For example, when length ( $p_4$ ) and hiding ( $p_5$ ) were the lowest, the former is selected. The result is described as a pair of the name of parameter and the evaluation value.

(4) **Rewriting Expectation** The narratee rewrites the expectation parameters through following two processes. First process rewrites the number of sufficiency and the degree of desire. The former is increased when the narratee received the sufficed discourse in each time. And this change causes the rise and fall in the degree of satisfaction as shown in Figure 3. And,  $n_p$  in the figure sets a turning point of the degree of satisfaction. Smaller  $n_p$  means the narratee is get tired easily. Through the generation cycle, such change occurs in each of ten parameters independently and repeatedly. In another process, a parameter’s value caused by the reception of a deviated discourse is renewed. (5) **Adjusting Generative Parameters** The narrator rewrites one of generative parameter’s values according to the indication from the narratee. If there are no parameters to be changed, this process is skipped. “Deviation” is done in this step when narratee got tired in the expectation parameter. It randomly alters the value of parameter got tired with a value from 1 to 3 except for the current value. In the next cycle (step 4), narratee will rewrite the deviated parameter’s value.

## Implementation and Execution Example

We implemented the system with Common Lisp. It mainly consists of three main elements: discourse techniques, narrator mechanism, and narratee mechanism. The program contains about 60 kinds of defined functions. Story and discourse are described with the same form of list. Moreover, we preliminarily provide supplemental data for events and descriptive information to use in “external analepsis”, “external prolepsis”, “achronie”, and “pause”. The system finally outputs a list of generated discourse and a Japanese natural language text. The latter is generated by a simple natural language sentences generation program we have originally developed. Table 3 shows an execution example which contains an input story, generative parameters, and a generated discourse. Both input/output are also described in natural language. In the generated discourse, 13 kinds of discourse techniques are applying. In addition, the input story directly uses a generated result of a story generation system by (Ogata & Terano, 1991) which uses a story grammar based on Propp theory (Propp, 1969). This shows an actual case this proposed system can be combined with other components in the narrative generation system architecture.

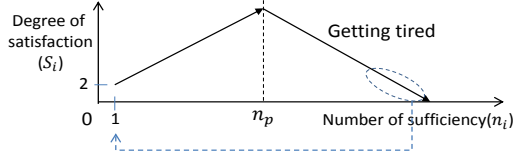


Figure 3: Changing pattern of the degree of satisfaction

Table 3: An execution example

Input story
A snake abducted a princess. An older woman sang a grief-stricken song. Ivan resolved at searching the princess. Ivan departed from the town. Ivan battled with the snake. Ivan got arm wounded by the snake. Ivan won against the snake. Ivan abducted the princess. Ivan departed from the snake's country. The snake flew in the air to the snake's country. The snake followed Ivan. Ivan hid himself to a rock. The princess found Ivan's wound. Dwarfs built a palace. Ivan lived in the palace. Ivan married with the princess.
Generative goal
supplement:3, complexity:2, suspense:3, length:1, hiding:1, descriptive-ness:2, repetition:2, diffuseness:2, implication:3, temporal-independency:2
Generated discourse
An older woman sang a grief-stricken song. The woman was born. (And) the woman grew in a town. (Now,) Ivan battled with a snake in the snake's country. Ivan got arm wounded by the snake. (But) Ivan won against the snake. (And) Ivan departed from the snake's country. The snake followed Ivan. (A few years later,) the older woman will get weak and die. (By the way,) the snake prophesied that dwarfs will build a palace. Ivan hid himself to a rock. A princess found Ivan's wound. The snake's country is slightly dark. Ivan's wound is bleeding. The princess is beautiful. (Soon after,) Ivan lived in a palace. (A short time ago,) dwarfs built this palace. Dwarfs built this palace. Ivan married with the princess. (Go back to the time of beginning,) the snake abducted the princess. The older woman sang a grief-stricken song. Ivan abducted the princess. Ivan departed the snake's country. (At that time,) Ivan prophesied that dwarfs will build a palace, (and) dwarfs will build a palace. (By the way,) somewhere a bear went to farming village. The bear attacked a farmer.

## Operation Checks and Evaluations

We think that the evaluation of narrative generation system should be extremely done with a goal of narrative qualitative progress such as the improvement of creativity and interestingness. For example, Callaway and Lester (2002) proposed some evaluation items although the aspect of narrative comparatively surface language generation and Akimoto and Ogata (2009) organized evaluation criteria comprehensively. However, as the previous step, we attempt fundamental checks of the performance and simple evaluations. First two attempts are for the performance confirmation. In the first check, we analyze the aspect of logical structure in generated discourse representations. Next, an important purpose of the current system is the realization of no arbitrary diversity in the generation. First is a simple attempt for confirming whether changing generative goals results in the diversity of generated texts. And, second is an experiment for investigating narrative diversity through a generation cycle based on the interaction between narrator and narratee. In the last experiment, we quantitatively verify the correspondence relationship between used parameters and generated discourses. All experiments use the input story in preceding section.

## A Structural Analysis of Generated Discourse

First, as a confirmation of the system's performance, we explain the overall structure of the result shown in Table 3. The outline of input story is that "A snake abducts a princess, and then Ivan rescues the princess from the snake, and then Ivan married the princess". Whereas, the output discourse has some features: (a) some events relating to the princess are hidden at the early part of the discourse, and are revealed after the marriage, and, (b) the discourse is longer than the input obviously. Moreover, we confirmed although generated discourse is struc-

turally different from the story, both have a same semantic content. This coincides with the definition of the relationship between story and discourse. Next, we analyzed the transformation process of above result to check the logical correctness in the processing of used techniques and confirmed that each technique was correctly functioning as the individual level. However, we found out some matters at the level of combinatorial application of techniques. For example, a node inserted by "complementary analepsis\_paralipsis" was additionally moved by "complementary analepsis\_ellipsis", that is, the result of the former was negated by the latter. This topic is generalized that a part of tree constructed by previously applied discourse techniques is additionally transformed by the later techniques. Such phenomenon, i.e., the interference among techniques may bring logical errors in a discourse structure. For instance, under the "present-backward" relation, right side's child nodes may contain posterior events to left side's child nodes. To solve such problem, we prepared some heuristic constraints and priority order rules to apply techniques in step 2 in the control mechanism.

## Diverse Generation by Changing Parameters

Although it is no wonder in a sense, one of the merits in the proposed mechanism is that the diversity of generation brought by changing generative parameters. This characteristic is relating to a basic concept in the narrative generation system project of the flexible generation from fragmentary narrative elements and techniques. We confirmed generation diversity using "measured values" which means numerical numbers calculated based on the measurement criteria for each parameter. These values are automatically calculated from generated discourse using an embedded program routine. The various kinds of discourses are generated by different generative parameters such as very short one, longer and relatively complex ordered one, and so on. The obvious changes of measured values were caused by the change of values of the parameters. On the other hand, we confirmed that the system generates discourse structures in a certain range from same generative parameters. For example, 100 results generated from a same story (Table 3) and a set of generative parameters which generates very short discourse, the range of measured value "length" was 6 to 11. In summary, we could confirm that a certain degree of generative diversity is actualized from a generative goal and the change of parameters causes the change of the range of generation.

## Diverse Generation through Continuous Cycle

The objective of this experiment is to investigate the different changing patterns of a circulative discourse generation process based on the different values of  $n_p$ . We executed the program respectively 10000 cycles for two kinds of  $n_p$ , 20 and 200. Figure 4 shows the change of four measured values in generated discourses. However, although the only first 500 cycles are shown, a similar pattern of change was continued after that. Two types of changing patterns exist. First is a micro level changing pattern based on same generative parameters occurring in each cycle. However, in this figure, "supplement" has not such kind of change. Another type is a macro level changing pattern caused by different generative parameters. The frequency in this kind of change is influenced by the value of  $n_p$ , and more frequent changes occur with smaller value of  $n_p$ . This fact indicates that intentional or strategic control of the system's behavior may become possible by adjusting the value of



$n_p$ . For example, if we want to generate in a wide range of discourses, we can set smaller values of  $n_p$ . In contrast, if we want to generate in a narrow range of discourses, we can set larger values of  $n_p$ . Next, we focus on the range of generated discourses. In each measured value, there was a certain range as shown in Figure 4. On the other hand, as the combinations of measured values, 18765 patterns of discourse texts were generated from above 20000 discourses as the pattern which has completely similar measured values. In summary, the grasp of generative characteristics by parameters is connected to the development of a variety of control strategies. In addition, expansions of the existing control mechanism are also conceivable. The circulative generation process in the mechanism is thought of as a kind of closed-loop because same a changing pattern is repeating through a cycle. Regarding this, we programmed a simple mechanism which manipulates a specific parameter intentionally or randomly to create a part exceeding circulative loop. For example, by increasing a parameter's value incrementally, the corresponding aspect in discourse generation deviates from the closed-loop. Such a kind of breakdown or mutation may connect to narrative creativity or interestingness. These are implications for system implementation the experiments have. On the other hand, this kind of trial exploits a possibility of theoretical approach to creative genres like literature in terms of providing an experimental method to narratology, reception theory in this case.

### Correspondence between Generative Parameters and Generated Discourse

We programmed a function which generates discourse texts by all (59049) combinations of generative parameters to quantitatively confirm the correspondence between used parameters and generated texts. Table 4 shows all of the correlation coefficient between each generative parameter and each measured value. In this table, vertical line and horizontal line respectively shows generative parameters and measured values. Each intersection means their correlation coefficient. In each parameter,

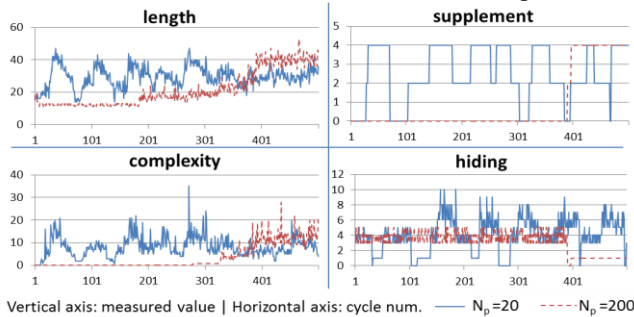


Figure 4: Changing parameters through a generation cycle

Table 4: Correlation coefficients between generative parameters and measured values

		Measured value									
		Supplement	Complexity	Suspense	Length	Hiding	Descriptiveness	Repetition	Diffuseness	Implication	Temporal-independency
Generative goal	Supplement	0.76	0.19	-0.01	0.22	-0.00	-0.01	0.00	0.00	0.22	0.00
	Complexity	-0.00	0.41	0.26	0.22	-0.00	0.00	0.21	0.17	0.11	0.14
	Suspense	-0.00	0.29	0.35	0.14	0.00	-0.00	0.16	0.13	0.01	0.00
	Length	0.19	0.05	0.01	0.39	-0.28	0.45	0.04	0.19	0.00	0.14
	Hiding	0.00	0.06	-0.07	-0.26	0.69	0.01	0.02	-0.37	-0.01	0.00
	Descriptiveness	-0.00	-0.00	-0.00	0.23	-0.00	0.61	-0.00	-0.00	0.00	0.00
	Repetition	-0.00	-0.06	-0.09	0.23	-0.00	0.00	0.40	0.33	0.00	0.00
	Diffuseness	0.00	0.00	-0.02	0.25	-0.28	-0.00	0.22	0.34	0.00	0.00
	Implication	0.38	0.40	-0.04	0.28	-0.00	-0.01	0.19	0.16	0.81	0.00
	Temporal-independency	0.00	0.21	-0.02	0.19	0.00	0.00	0.23	0.19	-0.00	0.88

higher value in the corresponding measured value's column (shaded one) and smaller values in other columns mean that the parameter is clearly reflected into the generated text. As a result, these are not complete correlations because the measured value by a parameter is sometimes influenced by the effect of other parameters. For example, the measured value by "length" is influenced by "descriptiveness" parameter. And, as correlations in "complexity", "suspense", "length", "repetition", and "diffuseness" were relatively weak, a reorganization of the correspondences between parameters and discourse techniques is required. As an idea, we are considering a method for hierarchically organizing parameters according to the abstraction level. For instance, parameters like "length" influenced by parameters such as "supplement" and "descriptiveness" may be positioned at the lower or more concrete layer in the hierarchy. In this hierarchical model, each technique is linked to one or more lower parameters which are influenced by the technique.

### Overall Discussion and Future Issues

This section shows more general discussions and future directions from several standpoints.

#### Toward the Expansion of Control Mechanism

Proposed system has the capacity to generate diverse discourses within a certain definite range based on the automatic change of both sets of parameters for generation (narrator) and evaluation (narrate). The changing pattern in discourse structures in a circulative process is based on variable  $n_p$ . In this mechanism, discourse generation is principally executed automatically. In contrast, as stated in previous section, we are considering a method with more intentional and conscious mechanism to control generation. We prepared an experimental program repeating discourse generation based on a direction which a real user directly gave. By increasing or decreasing the values of "length" and "complexity" intentionally, this program repeats the rises & falls or continues to extend the length. For instance, we confirmed that the program performed just as a given direction and very long outputs were generated. The measured value of "length" was 149 as maximum. Another idea is about "deviation" process in the narrator (Table 2). Although current system changes only one parameter's value randomly in deviation process, we experimentally modified the system to be able to adjust the number of changing parameters in deviation process. As a result, we confirmed more number of changing parameters causes more rapid change of outputs. These experiments show that proposed control mechanism is an expandable framework based on the strategic and flexible adjustment of parameters.

#### An Application for Narrative Creation Support

Although this research is directly aiming at automatic narrative generation, at the same time, we are involved in the planning of system for supporting user's narrative creation with the automatic generation function. For example, the system automatically generates diverse narrative texts and then the user selects one or more outputs to complete or expand by processing them as a kind of narrative material. In the case, narrative generation mechanism is corresponding to a function for stimulating user's thought and inspiration. To investigate the idea, next simple experiments is attempted. First, we selected two preferred texts from many outputs by the system (by the input information in Table 3). At this time, we can see graphs of measured values as

shown in Figure 4 to grasp parameters' characteristics in generated discourses. And then, we created a new story by the processing, expansion, and elaboration, moreover by combining two materials into one story. Human strong point is the creating of complex or complicated psychological narrative simulations and the processing of rhetorical techniques on surface text representation. On the other hand, machine's special skill is the generation of complex or complicated sequences in temporal progression and the logical processing of other discourse elements, and machine has the capacity to be able to generate superhuman texts, though, they may sometimes unnatural. Collaborative narrative creation or creation support by computer is one of the future directions in this research.

### Issues for Introducing Narratology

In this proposal of the application of Genette theory, we do not reach the comprehensive introduction. However, as we provided a common method for implementing discourse techniques, we can directly use it to extend the range of covering Genette based techniques and other categories. Previously, we have been developing several elemental systems for discourse techniques including order, distance, focalization, and other categories (Ogata et al. (2004) shows their overviews). However, these programs were not integrated as a whole system because of different data forms and processing methods. The first step for the integration at the level of discourse becomes the integration of existing functions into the proposed framework. On the other hand, the proposition of computational formation and implementation inspired by Jauss reception theory is also an original result in this paper. Although we stated the topic of individual narrator/narratee and collective ones, other various issues. For example, the narratee mechanism does not refer to generated discourse itself, and the narratee's reception pattern, which is the process of rise and fall in the degree of satisfaction, is preliminarily fixed. These issues will be solved through an exploratory approach in the future.

### The Integration with Narrative Generation System

Proposed system is positioned as a module of the narrative generation system. To integrate it into the system, first, it is necessary that the output data by the story generation phase becomes the input data of the discourse phase by unifying the form of knowledge representations. Regarding this, a tentative experimental version of integrated narrative generation system is already implemented by Akimoto and Ogata (2011). Next issue is the revision and expansion of discourse mechanism itself. Our immediate goal is to develop a systematic set of discourse techniques including all categories of Genette theory. At the same time, other types of techniques the theory does not describe are also required to be added into the comprehensive discourse mechanism. For example, Genette did not refer to the concrete way of "description" and "explanation" for character, object, and so on, at least systematically or formally. This sort of micro discourse techniques is a topic that has been discussing in the field of AI and natural language processing, and by the medium of this part, literary or narratological knowledge and AI-based knowledge will be blended in a narrative generation system. Moreover, we plan to expand the proposed narrative control method based on the interaction between narrator and narratee inside the system to an entire system including story phase and surface representation phases.

## Conclusions

This paper proposed a computational system of narrative discourse generation and its implementation. In the system, Genette's discourse theory is reconstructed as discourse techniques which transform the tree structure for a story into discourse structures. Also, we introduced Jauss's reception theory to construct the control mechanism, which continues discourse generation through generation cycles based on the interaction between both narrator and narratee mechanisms. Moreover, we did two kinds of performance checks and two types of evaluation experiments and confirmed that the system generates diverse discourse structures on the rough correspondence with generative parameters. And furthermore, this study showed that two different types of literary knowledge are organically integrated into a system's framework. At last, although this research does not directly treat the aspect of human cognition, as mentioned above, this indicates that advanced literary knowledge which is an important part at human cognition can be studied as a system, especially a computational system, more essentially beyond the traditional boundaries of fields of study. This is also a significance of cognitive science in a wide sense.

## References

- Akimoto, T. & Ogata, T. (2009). On the evaluation of a narrative discourse system: surveys of evaluation methods and an exploratory evaluation. *Proc. of the 19<sup>th</sup> Conference of LCCII, Japanese Cognitive Science Society*. 19G-01. (in Japanese)
- Akimoto, T. & Ogata, T. (2011). A consideration of the elements for narrative generation and a trial of integrated narrative generation system. *Proc. of the 7<sup>th</sup> NLPKE* (pp. 369-377).
- Bringsjord, S. & Ferrucci, D. A. (2000). *Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS, a Storytelling Machine*. New Jersey: Lawrence Erlbaum.
- Callaway, C. B. & Lester, J. C. (2002). Narrative prose generation. *Artificial Intelligence*, 139(2), 213-252.
- Genette, G. (1972). *Narrative discourse: an essay in method*. Transl. Lewin, J. E. (1980). New York: Cornell University Press.
- Jauss, H. R. (1970). *Toward an aesthetic of reception*. Transl. Bahti, T. (1982). Minneapolis: University of Minnesota Press.
- Lönneker-Rodman, B. (2005). Narratological knowledge for natural language generation. *Proc. of the 10<sup>th</sup> European Workshop on Natural Language Generation* (pp. 91-100).
- Meehan, J. R. (1980). *The Metanovel: Writing Stories by Computer*. New York: Garland Publishing.
- Montfort, N. (2007). *Generating narrative variation in interactive fiction*. A dissertation in computer and information science, University of Pennsylvania.
- Ogata, T., Hori, K., & Ohsuga, S. (1994). Towards narrative text generation based on narrative techniques and strategies. *Proc. of International Federation for Information and Documentation* (pp. 296-300).
- Ogata, T. & Kanai, A. (2010). *An introduction of informatics of narratology: on thought and technology of narrative generation*. Tokyo: Gakubunsha. (in Japanese)
- Ogata, T. & Terano, T. (1991). Explanation-based narrative generation using semiotic theory. *Proc. of Natural Language Processing Pacific Rim Symposium 91* (pp. 321-328).
- Ogata, T., Umehara, S., Yamakage, S., Ueda, K., & Hosaka, Y. (2004). Aspects of narrative discourse process and their integration by computer simulation. *Cognition, text and society of narrative: notes of intensive course at university of Yamaguchi 2002* (Edited by Ogata, T., Tech. Rep. JCSS-TR-52, pp. 136-143). Japanese Cognitive Science Society.
- Propp, V. (Пропп, В. Я.) (1969). *Морфология сказки, Из, 2е. Москва: Наука*. (Transl. Scott, L. (1968). *Morphology of the Folktale*. Austin: University of Texas Press.)



# A Bayesian Model of the Effect of Object Context on Visual Attention

**Ben Allison** (ballison@inf.ed.ac.uk)

**Frank Keller** (keller@inf.ed.ac.uk)

**Moreno I. Coco** (mcoco@inf.ed.ac.uk)

Institute for Language, Cognition and Computation

School of Informatics, University of Edinburgh

10 Crichton Street, Edinburgh EH8 9AB, UK

## Abstract

Research in visual cognition has demonstrated that scene understanding is influenced by the contextual properties of objects, and a number of computational models have been proposed that capture specific context effects. However, a general model that predicts the fit of an arbitrary object with the context established by the rest of the scene is until now lacking. In this paper, we explain the contextual fit of objects in visual scenes using Bayesian topic models, which we induce from a database of annotated images. We evaluate our models firstly on synthetic object intrusion data, and then on eye-tracking data from a spot-the-difference task and from an object naming experiment. For the synthetic data, we find that our models are able to detect object intrusions accurately. For the eye-tracking data, we show that context scores derived from our models are associated with fixation latencies on target objects.

**Keywords:** visual attention; object context; Bayesian modeling; eye-tracking data.

## Introduction

Real-world objects are often related to each other and typically form a coherent scene. For example, a toothbrush is likely to occur with a tube of toothpaste, a mirror, a sink; it is unlikely to occur with a sauce pan, a salt shaker, a cooker. For a given object, it is therefore possible to determine whether it is in context in a scene (toothbrush in bathroom), or out of context (toothbrush in kitchen). Experimental evidence shows that context information facilitates human object recognition (Bar, 2004). In visual search tasks, eye fixations are targeted towards contextually appropriate regions (Torralba et al., 2006), and out-of-context objects attract fixations earlier than in-context objects (Underwood et al., 2008).

In computer vision, being able to detect out of context objects is useful for object labeling. The local detectors standardly used for this task only consider the visual features of the pixels within the bounding box of the object of interest (Felzenszwalb et al., 2010). Local detectors are therefore prone to confusing objects that are visually similar (e.g., fork and toothbrush). This problem can be addressed by combining a local detectors with a model of object context, i.e., a model that determines which objects occur together. While this approach has been shown to increase object labeling performance (Choi et al., 2010; Galleguillos et al., 2008), the context models used are simple, typically relying on co-occurrence statistics over object labels. Furthermore, the context models used in computer vision are not designed to capture human performance (e.g., in visual search). Therefore, these models have not been evaluated on tasks such as detecting out-of-context objects.

In this paper, we present a new model of object context based on a more complex notion of object label co-occurrence that makes use of latent (i.e., unlabeled and unobserved) scene types: the Latent Scene Type model. This model allows us to exploit the common structure of scenes in order to estimate reliable parameters even for infrequently occurring objects. We investigate two model variants: the first is Latent Dirichlet Allocation (LDA, Blei et al. 2003), a standard model of word-topic co-occurrence, which we use to capture object-scene type co-occurrence. The second model variant is formulated as a Bayesian mixture of multinomials, which assumes one latent scene type per scene (rather than one per object, as in LDA).

We test both model variants on the task of producing context judgments for objects in scenes. We first use a synthetic data set for evaluation (in this data, context objects have been artificially inserted). In the second evaluation study, we use our model to mimic the data from an eye tracking experiment in which human participants had to spot out-of-context objects. Finally, we demonstrate that our model can predict fixation latencies in an object naming experiment which included out-of-context objects.

## Related Work

To our knowledge, ours is the first model to attempt to quantify the degree of fit between arbitrary objects in a scene, and to correlate the predictions of such a model with human behavior in scene viewing tasks. However, a number of models have been proposed to capture context effects on visual attention; a prominent example is the Contextual Guidance Model (CGM, Torralba et al. 2006), which combines bottom-up saliency with global scene information (scene gist, Oliva & Torralba 2006). The model is trained on a set of images in which the target objects are labeled; from this data a probability distribution of typical positions of objects is learned. This distribution is conditioned on the scene gist, essentially a coarse-grained representation of global image features. Gist is a latent variable in the model, comparable to scene type in our approach. The CGM has been evaluated on eye-tracking data from visual search experiments, and can successfully predict the scene-type-specific search behavior that participants exhibit. However, the model is not specifically designed to detect out-of-context objects, and has not been evaluated on tasks that require an estimate of the contextual fit of an object.

In the computer vision literature, the work closest to ours in spirit, if not in ultimate task, is that of Choi et al. (2010). The authors use a generative model of images features, scene gist, the set of objects in the image, and their locations, to re-rank the output of a local object detector to respect contextual interactions between objects, and show an improvement over the baseline detector. The co-occurrence model used by Choi et al. (2010) is a fairly simple binary tree of presence features whose principal purpose is to facilitate inference on other aspects of the image.

The model of object context proposed in this paper is formulated as a topic model. While topic models have been studied extensively in both the language and vision literature, they originate from applications to text, beginning with Latent Dirichlet Allocation. LDA assumes a document is sampled from a mixture of multinomials, where the multinomial from which words are drawn is sampled once per word, and the mixture co-efficients are sampled once per document. A corpus is then a distribution over mixture co-efficients. This approach can be adapted fairly straightforwardly for modeling objects instead of words, and scenes instead of documents. However, we note that instead of being used as descriptive tools to provide insight into collections, in this paper we are interested in the predictive aspects of a topic model and want to test how well they correlate with human scene viewing data. In this respect, while the models we used are standard, the purpose for which we use them is novel and we derive new metrics to correlate with human behavior.

Topic models have been applied to images in the computer vision literature (Wang et al., 2009; Li et al., 2009), but rather than describing the sampling of object labels, these models specify how discrete-valued image patches are sampled (by quantizing continuous image features), and the relation between these patches and the labels applied to the image.

## Models of Latent Scene Type

This paper presents a model of context, and by extension contextual fit, which rests solely on the set of object labels in the image. The method we employ has two components: a distribution over the set of labels in the scene, and the application of such a model to a continuous measure of how well any object fits with that scene. The observation of sets of objects is explained through latent scene types, which can be thought of as simple clusters of objects which are likely to co-occur. We then use the predictive distribution over new sets of objects, as derived from our latent scene type models, to determine the fit of target objects to the scenes.

### A Model Of the Probability of a Set of Object Labels

We experiment with two models of the probability of a set of object labels, both of which are topic models. Topic models comprise mixtures over multinomial distributions, where in this case the multinomial outcomes correspond to object labels. The first topic model, Latent Dirichlet Allocation, is fairly standard, while the second one, the mixture of multinomials model, is less commonly used. Each of the models

describes a distribution over a vector of counts, which we call  $o$ , such that  $o_i$  is the count of the  $i$ -th object in the current image (note that for most images, most of these elements will be zero).

**Latent Dirichlet Allocation** There has been much interest in the use of topic models as descriptive tools, able to infer structure in collections of documents, or other collections of discrete entities (Blei et al., 2003; Wang et al., 2009). For the LDA model, the predictive distribution over new sets of object labels is given by:

$$p_{LDA}(o|\alpha, \beta) = \int p(\theta|\alpha) \left( \prod_n \sum_{z_n} p(z_n|\theta) p(l_n|z_n, \beta) \right) d\theta \quad (1)$$

where  $l_n$  is the label for the  $n$ -th object in the image, and  $z_n$  is the (latent) topic assignment for this label—counting the  $l_n$  gives  $o$  as defined above. The  $z$ s are indicators explaining which latent scene type was used to generate the current label. As in the original paper,  $\alpha$  is the Dirichlet prior on  $\theta$ , and  $\beta$  is the topic–word probability matrix that gives the probability of each object label in each topic. The above is evaluated using a particle-filter-inspired Monte Carlo method described by Wallach et al. (2009).

**Mixture of Multinomials** The mixture of multinomials model is defined over the same count vector  $o$  as above, but for this model the scene type  $z$  is sampled only once per scene. The parameters to the model are  $\phi$  (the mixture coefficients) and  $\theta$  (the parameters for the component multinomials). The distribution over the observable variables,  $o$ , is:

$$p(o) = \sum_z \phi_z p(o|\theta_z) \quad (2)$$

where the distribution  $p(o|\theta_z)$  is a multinomial parametrized by the vector  $\theta_z$  (its components giving the probabilities of each possible label occurring within that component).

We explore two variants of this model—the first uses maximum a posteriori (MAP) estimation to fix the parameters to the (approximate) posterior mode—the single best estimate of model parameters. This can be done using EM, and we employ uniform priors on both sets of parameters. Conditioned on some observations (training data, which we label  $D$ ), the maximum likelihood method stipulates:

$$\hat{\phi}, \hat{\theta} = \arg \max_{\phi, \theta} p(D|\phi, \theta) \quad (3)$$

$$p_{ml}(o|D) = p(o|\hat{\phi}, \hat{\theta}) \quad (4)$$

This predictive distribution (4) is what we are interested in: exploring the probability of new scenes given our training data.

However, from a computational as well as cognitive perspective, given only limited samples from the process we should feel uneasy about saying with any certainty what the values of the parameters are. Instead, we suggest that given

our experience we have beliefs about what is likely to happen, but we retain uncertainty and factor this in to our predictions. In light of this, we also employ a Bayesian version of this model, which integrates over the full parameter space given our training data  $D$ :

$$p_{\text{bayes}}(o|D) = \int p(\phi, \theta|D) p(o|\phi, \theta) d\phi, d\theta \quad (5)$$

For the mixture of multinomials model we cannot evaluate this integral in closed form, so we sample mixture models from their posterior  $p(\phi, \theta|D)$ —i.e., we retain uncertainty about which model best explains our data, and average over this uncertainty in deriving our predictions. Assuming Dirichlet priors (in our case, uniform) on  $\phi$  and  $\theta$  leads to Dirichlet posteriors over these same parameters, conditioned on assignments of training observations to latent mixture components; it is these assignments that we sample. We then evaluate the  $p(o|\phi, \theta)$  at each of these sampled points; in practice we do not sample different mixture models for each new  $o$  we wish to evaluate, but run the sampler once in training and store all mixture models sampled. This allows us to simply average over the sampled components for the predictive distribution, leading to a deterministic evaluation of (5) as simply the mean of the probability  $p(o|\phi, \theta)$  under the sampled models.

As a final note for all these models, while inference techniques are approximate in all cases, and different between the MoM and LDA models, we are confident that the particular approximations do not overly sway the models' ultimate performance. While using heldout probability as the metric of concern show disparities between different approximations for the LDA model in Wallach et al. (2009), our uses of the models are different. Most of the problems we consider are decision problems where the exact probability is less of a concern than the relative probabilities of the scenes under two models, and in the final section we are interested in the correlation between the probabilities and some other continuous measure which is unlikely to be affected by (relatively) small changes due to approximation error.

### Detecting Out-of-context Objects with Scene Probability

The previous section presented models which have been used previously in other fields for describing the co-occurrence of entities. We turn in this section to the manipulation of these models to derive quantities which we will correlate with human performance.

There are two distinct tasks we explore in this paper: which of two objects is more probable given a scene, and whether a given object belongs to a scene or not. Here, we briefly describe the use of the models we defined in the previous section to achieve these tasks.

Firstly, the conditional probability of some object (label) in question  $o'$  given a set of object labels  $o$  (where  $o$  is the count vector as above) is:

$$p(o'|o) = \frac{p(o \cup o')}{\sum_{o_{\text{new}}} p(o_{\text{new}} \cup o)} \quad (6)$$

That is, the probability of the count vector which includes the new label  $o'$ , normalized by the probability of the context for all possible objects which could be added to the scene.

To determine which of  $o^1$  and  $o^2$  better fits some context  $o$  we can compare  $p(o^1|o)$  and  $p(o^2|o)$  computed as in (6)—we may simply interested in which of these is the larger, or perhaps in the ratio between these two quantities. Note that in either case, the normalizing constant can be dropped since it is common to both (this speeds up computation considerably).

Secondly, to determine whether  $o'$  is in context or not, we can compare  $p(o'|o)$  with the quantity obtained by marginalizing out the extra object, namely:

$$p(o^{\text{new}}|o) = \sum_{o^n} p(o^n) p(o^n|o) \quad (7)$$

where  $p(o^n)$  is the probability of  $o^n$  occurring in any scene, for which we use simply the fraction of all objects across all scenes which are  $o^n$ . For this paper, we explore both the decision problem (is  $p(o'|o) > p(o^{\text{new}}|o)$ , i.e., is the object in context or not) and the continuous scores derived as above.

### Evaluation on Synthetic Out-of-context Objects

We construct our first test set based on the Spatial Envelope data set (Oliva & Torralba, 2001). Here, the models will be used to determine whether an object is in context with respect to the rest of a scene, or not (Equation (7)). The images in the data set contain full object annotations, but also scene type labels. These allow us to construct test data for the scenario we are interested in. (Note, however, that this is the only use of overt scene type labels in this paper; the scene types in our model are latent.) The data set is annotated using LabelMe conventions, but does not overlap with the LabelMe data from which our models are estimated. In terms of objects per scene, there are on the order of ten objects in each image, and the number of images is reasonably balanced between scene types. We extract scenes which are either rural or urban (the two top level scene types). We produce frequency counts of objects within these two categories, and compute a  $\chi^2$  statistic for each to measure the distinctiveness of that object in that class. We then select the 25 most distinctive objects for each class which occur in at least ten scenes, and extract all scenes containing each of these objects. These distinctive objects are treated as the targets, and the other objects in the image form the contexts.

The original scenes form examples of in-context objects—to produce out-of-context ones, for each scene we replace the in-context target with a randomly selected member of the distinctive list for the other category. This produces a set of just over 26,000 scenes, equally balanced between in- and out-of-context objects, to use for further experimentation. We divide this into 6,000 scenes for development (model selection and parametrization), with the remainder being used for held-out testing. In all cases, the held-out data are unobserved until all model parameters are fixed. Table 1 shows examples of the data we produce.

Target	In/Out Context	Context
<b>stone</b>	<i>in</i>	stick:1 stone:1 tree_trunk_fallen:2 trees:1 ground:2 brushes:1
<b>buildings</b>	<i>in</i>	skyscraper:1 building_occluded:2 buildings:1 sky:1 skyscraper_occluded:1
<b>road</b>	<i>out</i>	tree:1 stone:3 river_water:1 trees:1 field:1 sky:1 stones:2 rocky_mountain:1
<b>sea_water</b>	<i>out</i>	window:11 car_occluded:2 pot_plant_occluded:1 sidewalk:1 person_occluded:1 arcade:1 palm_tree:1 car:1 window_occluded:1 person_walking:3 person_woman_walking:1 traffic_light:1 hall:1 building:1 road:1

Table 1: Some examples of the synthetic data—the context is depicted as a sparse vector over the outcomes in the form  $[label:count]$ , which is then reduced as appropriate for a trimmed vocabulary

	LDA		ML-MoM		B-MoM	
$ T $	500V	1000V	500V	1000V	500V	1000V
50	0.737	0.747	0.674	0.679	0.896	0.895
100	0.759	0.801	0.660	0.662	0.897	0.899

Table 2: Accuracy (proportion of decisions where the correct determination is made) on the synthetic data

**Results** Table 2 shows results on the synthetic dataset. The Bayesian Mixture of Multinomials is clearly superior to the other two models, and the larger vocabulary size and greater dimensionality improves this slightly. The LDA model shows greater sensitivity to parametrization than the other two, and the maximum likelihood model is considerably worse than the others across all parameter settings. Of particular note is the maximum likelihood model getting *worse* as the dimensionality increases; this is a classic result for non-Bayesian models, where as the parameter space expands it is less and less well summarized by a single point (the mode) and that mode becomes harder to find.

## Modeling Human Experimental Data

The evaluation study presented in the previous section used artificially generated data. It showed that the Latent Scene Type Model is highly accurate at detecting out-of-context objects which have been inserted into a scene. In the present study, we validate this result using a data set from an eye tracking experiment by Underwood et al. (2008). In this experiment, participants had to perform a search task (determine whether two scenes are the same or different); in the different-scene condition, the target object was either out of context or in context, with saliency being controlled. An example pair of scenes can be found in Figure 1. The results show that scenes with in-context objects are inspected for longer and received more fixations than scenes with out-of-context objects. Also the in-context objects themselves were detected later and required more fixations prior to detection than out-of-context objects.

We expect our Latent Scene Type Model to capture the behavioral effect of out-of-contextness demonstrated by Underwood et al.’s study: out-of-context objects should receive

lower probabilities than in-context objects in their data set.

Underwood et al.’s study contains 80 pairs of scenes. In one scene in the pair, the target object is in context (congruent, in the language of that paper) and in the other it is out of context. (Saliency was also manipulated in the study, but this is not of interest here.) We manually listed the objects in each scene (the contexts are identical between pairs, and two pairs are identical save for their targets). Checking the labels against our LabelMe training data revealed that 25% of target objects were observed in LabelMe, and just over 30% of all objects. LabelMe contains mainly outdoor scenes, while the experimental data set are all indoor scenes, predominantly kitchen, utility room or bathroom scenes, in which the objects have been carefully arranged.

We therefore iteratively relabeled the target objects to establish a closer match with the LabelMe database, choosing in some cases synonyms and in others (direct) hypernyms. This was a manual process which relied on linguistic resources such as WordNet. This produced a target coverage rate of just over 70%, making it possible to use 45 of the 80 scene pairs, with each scene having on average approximately 60% of its context object appear in the training data (note that this increase was incidental, as we optimized the coverage of the target objects and simply propagated corrections through to contexts as well so as to reduce the amount of manual engineering). The selected scenes contain an average of ten objects in total.

Given the small size of the test set, we were not able to split off a separate development set, and therefore retained the parameters as set in the previous section on the synthetic data set.

**Results** Table 3 shows results on the Underwood et al. data for the task of detecting which of two possible objects is out of context. As in previous sections, we note that the Bayesian version of the mixture of multinomials model performs better than the maximum likelihood version of the model, but given the small dataset it is not possible to compare the B-MoM and LDA models except to say that both are significantly different from a random (50%) baseline as established by a binomial test. Note that while seeming disappointing initially, the performance of the models here is limited because the Un-

Method	LDA	ML-MoM	B-MoM
Accuracy	31/45	24/45	29/45

Table 3: Proportion of scenes in the Underwood et al. (2008) data where the correct determination was made. The LDA and B-MoM models are significantly different from a random (50%) baseline, but not one another



Figure 1: A pair of example scenes from the eye tracking experiment of Underwood et al. (2008). The target object in the left hand image is the sock (in context), while in the right hand image it is the can of soup (out of context)

Underwood et al. scenes are staged indoor shots featuring many objects that occur infrequently, if at all, in our training data (only 60% of context objects appeared at all). The next section presents an evaluation where objects are more frequently observed.

## Modeling an Object Naming Dataset

The third evaluation of our models used eye-tracking data from an object naming experiment by Coco et al. (2012). In this study, 24 participants were presented with 28 photo-realistic scenes and asked to name the five most important objects in the scene. In each scene, an object of interest and two competitors were inserted using Photoshop. The Saliency (Salient, Non-Salient) and Contextual Fit (In-Context, Out-of-Context) of the object of interest was manipulated. In contrast to Underwood et al. (2008), this study shows that out-of-context objects are less likely to be named than in-context objects. Moreover, first fixation latency, i.e., the time to land on a target object for the first time from scene onset, is longer for out-of-context than for in-context objects. A naming task demands a joint evaluation of both linguistic and visual information, thus even if an out-of-context object might be visually more informative, it is linguistically less relevant.

We first evaluate our models on the task of determining which of two objects is in context, identical to that presented in the previous section. Then, we investigate whether the contextual scores calculated by the models are correlated with the visual responses observed on the associated objects. We employ linear mixed effects model (LME, Baayen et al. 2008)



(a) In context target



(b) Out of context

Figure 2: An example of a scene with an in-context target (cup) and the same scene with an out-of-context target (fish)

analysis to investigate how first fixation latency (the dependent measure) correlates with model score (our predictor). LME is more appropriate than simple correlation because there were many other factors considered in the experimental data which affect the dependent measure, including frequency and saliency of objects in the scene and size of the objects. Employing LME means we are able to control for these factors by including them as covariates in the analysis.

On the basis of the experimental data, we expect the model score to be negatively associated with first fixation latency, i.e., the more out-of-context an object is, the longer it takes to fixate it. Together with the Score, we include as predictors the Saliency of the object, and the type of Model. As a random effect, we include Scene. We residualize first fixation latency by the area of the object (in pixel square) to reduce the effect of area on the dependent measure. We select the final LME model by following a forward step-wise procedure, where nested models are compared on the basis of log-likelihood improvement. In the following, we report the coefficients of the predictors found significant after model selection.

Method	LDA	ML-MoM	B-MoM
Accuracy	46/56	33/56	50/56

Table 4: Proportion of scenes in the Underwood data where the correct determination was made. The LDA and B-MoM models are significantly different from a random (50%) baseline, but not one another

	LDA		ML-MoM		B-MoM	
	I	O	I	O	I	O
S	0.0106	0.0001	0.0010	0.0010	0.0071	0.0006
NS	0.0064	0.0002	0.0010	0.0010	0.0054	0.0012

Table 5: Average context scores across the conditions—I is in context, O out of context, S is the salient condition and NS is the non-salient condition. ML-MoM scores are in fact slightly different to one another, but both contexts are highly improbable under the model

**Results** We first present the results for the decision problem. There are fifty-six decisions to be made (28 pairs of scenes, each in the salient and non-salient condition), with the goal being to determine which of the pair is out of context. Table 4 shows the results on this task, where we once again see that the LDA and Bayesian models are significantly above chance, but given the limited sample size not significantly different from one another. Table 5 shows the mean context scores across the conditions for each of the three models, where we see that the effects of the models in the decision problem (Table 4) are equally visible on the continuous scale.

When using the LME to check whether model score is a predictor of first fixation latency, we find a significant effect with  $\beta_{Score} = -0.1309$ ;  $p < 0.0001$ : the more in-context an object is, the shorter the latency. We do not find an effect of Saliency and Model. This result also echoes the experimental finding obtained by Coco et al. (2012), and shows that the scores generated by our models can capture the patterns in the eye-movement responses.

## General Discussion

This paper introduced the Latent Scene Type models for describing the fit of objects to scenes. Our models quantify how well a target object fits an observed context (the other objects in the image). Sets of objects are generated by latent scene types, with scene types representing objects which tend to co-occur. We choose a Bayesian formulation for our models, as this is attractive from a cognitive point of view: a cognitive process operates with finite experience, which means that it has to estimate a model of the world based on a limited sample (in our case of context and objects). Committing to a single parameter setting based on a limited sample is difficult; it therefore seems more plausible to integrate over the full parameter space, which is the hallmark of Bayesian models. The Bayesian approach therefore captures the uncertainty

faced by a cognitive process with access to limited data.

We showed that the Latent Scene Type models perform well on the task of detecting out-of-context objects in a synthetic dataset. Furthermore, we successfully applied the models to two eye-tracking datasets, one involving a spot-the-difference task, the other involving object-naming. In both cases, the models were able to successfully detect out-of-context objects, and in the case of the naming data, we also showed that model scores are associated with first fixation latencies on a target object (either in or out of context).

## Acknowledgments

The work reported here was funded by the European Research Council under award number 203427 “Synchronous Linguistic and Visual Processing”.

## References

- Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.
- Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5, 617-629.
- Blei, D., Ng, A., & Jordan, M. (2003, March). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- Choi, M., Lim, J., Torralba, A., & Willsky, A. (2010). Exploiting hierarchical context on a large database of object categories. In *Proceedings of cvpr'10* (p. 129-136).
- Coco, M. I., Malcolm, G. L., & Keller, F. (2012). The interplay of bottom-up and top-down mechanisms in visual guidance during object naming. *Journal of Vision*. (under review)
- Felzenszwalb, P., Girshick, R., McAllester, D., & Ramanan, D. (2010, September). Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence*, 32(9), 1627-1645.
- Galleguillos, C., Rabinovich, A., & Belongie, S. (2008). Object categorization using co-occurrence, location and appearance. In *IEEE conference on computer vision and pattern recognition (cvpr)*. Anchorage, AK.
- Li, L.-J., Socher, R., & Fei-Fei, L. (2009). Towards total scene understanding: classification, annotation and segmentation in an automatic framework. In *Proceedings CVPR'09*.
- McCallum, A. (2002). *MALLET: A Machine Learning for Language Toolkit*.
- Oliva, A., & Torralba, A. (2001, May). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42, 145-175.
- Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. In *Progress in brain research* (p. 2006).
- Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. M. (2006). Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review*, 113(4), 766-786.
- Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition*, 17, 159-170.
- Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings ICML'09* (pp. 1105-1112). New York, NY, USA: ACM.
- Wang, C., Blei, D., & Fei-Fei, L. (2009). Simultaneous image classification and annotation. In *In proceedings CVPR'09*.

# The Theory of Visual Attention without the race: a new model of visual selection

**Tobias S. Andersen (ta@imm.dtu.dk)**

Informatics and Mathematical Modeling, Technical University of Denmark  
2800 Kgs. Lyngby, Denmark

**Søren Kyllingsbæk (sk@psy.ku.dk)**

Center for Visual Cognition, Department of Psychology, Øster Farimagsgade 2A, 1353 Kbh. K.,  
University of Copenhagen, Denmark

## Abstract

The Theory of Visual Attention (TVA; Bundesen, 1990) is a comprehensive quantitative account of visual attention, which accounts for many empirical findings and has been extensively applied to clinical studies of attention. According to TVA, perceptual processing of objects occurs in parallel constrained by a limited processing capacity or rate, which is distributed among target and distractor objects with distractor objects receiving a smaller proportion of resources due to attentional filtering. Encoding into a limited visual short-term memory is implemented as a race model. Given its major influence it is surprising that few studies have compared TVA directly to alternative models. Here we insert an algebraically simpler model of encoding into TVA as an alternative to the race model and show that this provides a better fit to Shibuya and Bundesen's (1988) whole and partial report data, which have been a keystone test bed for TVA.

**Keywords:** Attention; working memory; Theory of Visual Attention; Vision; Psychophysics; Modeling

## Introduction

The Theory of Visual Attention (TVA; Bundesen, 1990) incorporates visual perceptual processing, attentional filtering and encoding into visual short-term memory (VSTM) in a unified quantitative model. The model has been extended to account for results from a wide variety of experimental paradigms (Logan, 1996; Logan & Gordon, 2001), and the neural implementation of TVA (NTVA) has been applied to results from single cell studies (Bundesen, Habekost, & Kyllingsbæk, 2005). Despite the extensiveness of the TVA based theoretical framework, we are aware of only a few recent studies (Dyrholm, Kyllingsbæk, Espeseth, & Bundesen, 2011; Kyllingsbæk, Markussen, & Bundesen, 2011; Petersen & Andersen, 2012) challenging the specific details of the model using standard model assessment methods. Of these studies we will include Petersen and Andersen's (2012) findings

that the log-logistic psychometric function inserted into TVA leads to improved performance in the current study.

Computational models of cognition such as TVA offer both theoretical and practical advantages. The theoretical advantages include the strict quantitative formulation of cognitive modules, the definition of which can otherwise prove to be elusive. Computational models can also be applied to a range of experimental paradigms and help arrive at a unified interpretation. This can be of practical use as the assessment of the function of cognitive modules is of great importance in clinical psychology and neuro-pharmacology. In this vein, TVA has been extensively applied to studies of clinical populations (Habekost & Starrfelt, 2009) and to the effect of psychoactive drugs (Finke, et al., 2010; Vangkilde, Bundesen, & Coull, 2011). Many of these studies base their assessment on estimates of the parameters in TVA and therefore rely on TVA precisely reflecting the actual computational mechanisms underlying visual attention. This makes it the more pressing to assure that this is indeed the case by comparing the specifics of TVA to competing models.

Whole and partial report tasks have been a keystone test bed for TVA. In whole report tasks, a number of objects, typically letters or digits, are presented to the observer. The task of the observer is to identify and report the objects presented. The exposure duration is typically brief (<200 ms) in order to avoid eye movements so that the information available can be assumed to be near constant across the stimuli and throughout the stimulus duration. Partial report tasks are like whole report except that in addition to the target



objects, a number of distractor objects are also presented. Some characteristic, like color, location or object category (e.g. letters vs. digits) distinguishes targets from distractors. The task of the observer is to report only the target objects and ignore the distractors.

Performance in whole report tasks is limited by perception and memory. In order for the target objects to be correctly reported, they must be perceived. This depends on stimulus attributes such as contrast, exposure duration, size, complexity and the number of stimulus categories (Pelli, Burns, Farell, & Moore-Page, 2006). Since these limitations exist also when only a single object is present the effect of these stimulus attributes can be studied in single letter identification experiments (Petersen & Andersen, 2012).

When multiple objects are presented the single letter psychometric functions cannot explain performance. Instead, the psychometric function needs to be adjusted. In TVA the adjustment is based on the assumption that the sum of processing resources, defined as the sum of hazard rates, is constant (Shibuya & Bundesen, 1988).

In partial report tasks, performance depends also on the ability to filter out the irrelevant distractor objects through selective attention in order to avoid their interference with perceptual processing and their taking up working memory capacity. If filtering is perfect, performance in partial report tasks should match that of whole report tasks with the same number of target objects. Shibuya and Bundesen (1988) showed that this is not the case and that the filtering process is imperfect. TVA models filtering as a smaller amount of processing resources being allocated to distractor objects.

Even when contrast and exposure duration are more than sufficient for all letters to be correctly identified according to the adjusted psychometric functions, observers fail to base their report on more than about four objects (Sperling, 1960). This seems to be due to limitations on VSTM rather than on perception *per se*. In TVA the mechanism of encoding is a race, so that objects are encoded into VSTM when they are

perceptually processed but only if VSTM capacity is still available, i.e. if it has not already been occupied by other objects.

TVA is thus able to describe performance in whole and partial report tasks with a given number of targets and distractors based on performance in single object identification in the form of the psychometric function. It does this based on assumptions of how multiple targets affect perceptual processing, the process of filtering and encoding into a limited VSTM. We find it difficult to envision a model that would not partition visual perception, attention and short-term memory into these components as does TVA but we find that there is room to examine the specific implementation of these stages.

In the following we shall examine the encoding stage of TVA, the race model. We will insert a different model of the encoding stage into TVA and compare the two encoding models' abilities to describe Shibuya and Bundesen's (1988) whole and partial report data. We will do this using either the exponential psychometric function conventionally used in TVA or the log-logistic function that Petersen and Andersen (2012) found to improve performance.

## Methods

### Modeling

#### The psychometric function and distributing resources

In TVA, perceptual processing of a single object is typically described by the exponential psychometric function

$$F(t) = 1 - \exp(-v_t(t - t_0)), t > t_0$$

$$F(t) = 0, t > t_0$$

where  $F$  is the probability of correctly identifying the object,  $v_t$  is the rate of processing for the target object,  $t$  is the exposure duration and  $t_0$  is a short time interval between stimulus onset and the beginning of perceptual processing. In terms of probability theory, the rate,  $v_t$ , is the hazard rate and  $v_t(t - t_0)$  is the cumulative hazard rate, the hazard rate integrated over time. When only a

single target is presented the sum of processing resources, or hazard rates,  $C$ , is allocated to that target so that  $v_t = C$ . In whole report, when multiple targets are presented, the objects are typically arranged at equal distances from the fixation point so that it is reasonable to assume that they receive equal shares of the processing resources, i.e.  $v_t = C/T$ , where  $T$  is the number of targets. In partial report, distractor objects are assumed to receive a proportionally smaller share of processing resources due to attentional filtering so that  $v_d = \alpha v_t$ . From this, we can deduce that  $v_t = C/(T + \alpha D)$  where  $D$  is the number of distractors (Bundesen, 1990).

In a recent study Petersen and Andersen (2012) showed that other psychometric functions can be inserted into TVA and that this, in general, improves the performance of the model. The log-logistic function gave the best fit of those functions having two free parameters like the exponential function. Therefore we will use it here. The log-logistic can be expressed as

$$F(t) = \frac{1}{1 + \left(\frac{t}{t_0}\right)^{-v_t}}$$

Although the parameters  $t_0$  and  $v_t$  describe the shift and the slope of the psychometric function respectively just as for the exponential function, their exact meaning is different than for the exponential function. The shift,  $t_0$ , is here the 50% correct threshold. Unlike the exponential function, the hazard rate is not explicit in the expression for the log-logistic function but the cumulative hazard rate,  $\Lambda_t$ , can be derived to be

$$\Lambda_t = -\log(1 - F) = \log\left(1 + \left(\frac{t}{t_0}\right)^{v_t}\right)$$

Distributing processing resources according to TVA with the log-logistic function becomes simpler if we notice that the assumption of a constant sum of hazard rates is equivalent to a constant sum of cumulative hazard rates. When

only a single object is presented the cumulative hazard rate is thus  $\Lambda_t = C_{cum}$ . From this the response probabilities in whole and partial report can be calculated by setting the cumulative hazard rate to  $C_{cum}/(T + \alpha D)$ .

### Encoding into a limited VSTM

The previous section outlined TVA applied to the case of whole and partial report when the total number of objects does not exceed the capacity of VSTM. In that case we can calculate the probability of the score,  $j$ , which is the number of correctly reported target objects, as

$$P(j) = \binom{T}{j} [F(t)]^j [1 - F(t)]^{T-j}$$

This expression is derived from the binomial distribution giving the probability of encoding  $j$  targets. The number of encoded target objects is termed the *score*.

When the number of objects exceeds VSTM capacity selection of the objects to encode is needed. According to TVA the selection happens as a race for free slots in VSTM; a race that ends when all slots are occupied or when perceptual processing ends. Inserting the race model into TVA is somewhat algebraically complex but allows calculating the score probability, i.e. the probability of correctly reporting a certain number of target objects. Detailed expressions and derivations are given in Petersen and Andersen (2012).

Here we introduce a different model of selection of objects to be encoded by conditioning on the total number of objects encoded being no greater than VSTM capacity, i.e.  $j + m \leq K$  where  $m$  is the number of distractor objects encoded. This probability is calculated by calculating the score probabilities for  $j = 1, \dots, T$  and  $m \leq \min(D, K - j)$

$$P(j) = \binom{T}{j} [F(t)]^j [1 - F(t)]^{T-j} \times \sum_{m=0}^{\min(D, K-j)} \binom{D}{m} [G(t)]^m [1 - G(t)]^{D-m}$$

Conditioning on  $j+m \leq K$  is then implemented by normalization of the probability mass function  $P(j)$ . Here, the psychometric function for distractor objects is denoted  $G(t)$ . Note that the number of encoded distractor objects,  $m$ , is considered an unobservable nuisance parameter, which is summed out.

For both encoding models, VSTM capacity,  $K$ , is allowed to take non-integer values, which are implemented as a mixture model where the VSTM capacity is the ceiling value of  $K$ ,  $\lceil K \rceil$ , with a probability of  $\text{mod}(K, \lfloor K \rfloor)$  where  $\lfloor K \rfloor$  is the floor value of  $K$  and  $\lfloor K \rfloor$  with a probability of  $1 - \text{mod}(K, \lfloor K \rfloor)$ .

### Model evaluation

As testing ground for comparing the two models of encoding we choose Shibuya and Bundesen's (1988) whole and partial report data that have been influential in the development of TVA (Bundesen, 1990). The data set consists of score counts for two observers each performing 6,480 trials with varying number of target and distractor elements and exposure durations. The observers were instructed to report the identity of targets only when they were reasonably confident in order to minimize the effect of guessing.

Only very rarely did the observers achieve scores greater than 4. Following the example of Bundesen (1990) we have registered these responses as scores of 4. The encoding models can be extended to account for these higher scores by allowing the VSTM capacity to vary between three integer values rather than just two but this requires an additional free parameter, which is difficult to justify by the ability to model only few of thousands of trials.

### Results

Table 1 displays the goodness of fits in terms of the negative logarithm of the likelihood for the two models of encoding and the two psychometric functions fitted to both observers in Shibuya and Bundesen's (1988) data. Note that the encoding

models and psychometric functions have the same number of free parameters.

The goodness of fits in Table 1 confirms that the log-logistic psychometric function provides a better fit than the exponential psychometric function as found by Petersen and Andersen (2012) and also that the conditioning model offers an additional, although slight, improvement in the goodness of fit.

Table 1: Goodness-of-fits

Psychometric function	Selection model	
	Race	Conditioning
Exponential	1579	1552
Log-logistic	1331	1273

To further examine the fits of the encoding models Figure 1 displays the cumulative score proportions, i.e. the proportion of responses to a given stimulus type with at least  $j$  correctly reported targets along with model fits for both encoding models with the log-logistic psychometric function for subject HV. As is evident from Figure 1, the model fits are very similar. It takes careful inspection to see that there are, in fact, systematic differences. The clearest difference is that when six targets are presented both encoding models tend to overestimate the cumulative score proportion but the conditioning model less so than the race model. Also, when the number of distractors is no greater than two, both models tend to underestimate the cumulative score proportion for exposure durations between 30–70 ms but the conditioning model less so.

For the briefest exposure durations of 10 ms observers rarely reported any targets. In Bundesen's (1990) analysis the few trials in which they did were discarded so that the score was assumed to be zero. This might favor the exponential psychometric function as it constrains the score to be zero for exposure durations shorter than  $t_0$ . We therefore fitted the models to the data with this data adjustment. The conditional model still fitted the data better but more so with the exponential psychometric function than with the log-logistic.

Table 2 lists the parameter values for the fits. Note that the VSTM capacity,  $K$ , and filter-parameter,  $\alpha$ , are comparable across both encoding model and psychometric function. They seem, however, to vary very little with these model variations within observers. The temporal threshold,  $t_0$ , and processing capacity,  $C$ , are not comparable across psychometric functions, only across encoding models. The temporal threshold seems also to vary very little with encoding model within observer and psychometric function. For the log-logistic function the processing capacity, i.e. the sum of hazard rates, varies over time and is therefore given for  $t = t_0$ . The processing capacity is slightly, but consistently, greater for the conditioning model than for the race model. This difference may seem slightly more pronounced for the log-logistic psychometric function ( $4 \text{ s}^{-1}$  averaged over the two observers) than for the exponential function ( $2 \text{ s}^{-1}$  averaged over the two observers) but this might be due to the difference in magnitude of  $C$  as the relative

differences were similar (7% for the log-logistic model and 5% for the exponential model averaged over the two observers).

Table 2: Estimated parameters for psychometric functions (Psy. F.), observer (Obs), and model.

Psy. F.	Obs.	Model	K	$\alpha$	C	$t_0$
Log-log.	MP	Race	3.8	0.40	57	0.036
		Cond.	3.9	0.41	61	0.036
	HV	Race	3.3	0.56	50	0.033
		Cond.	3.2	0.52	53	0.034
Exp.	MP	Race	3.9	0.39	37	0.010
		Cond.	4.0	0.38	38	0.010
	HV	Race	3.3	0.55	35	0.010
		Cond.	3.2	0.52	37	0.010

## Discussion

The differences that we found between the fits of the encoding models are consistent. They are, however, also small. This warrants care in model selection and this is, in fact, our main point. The

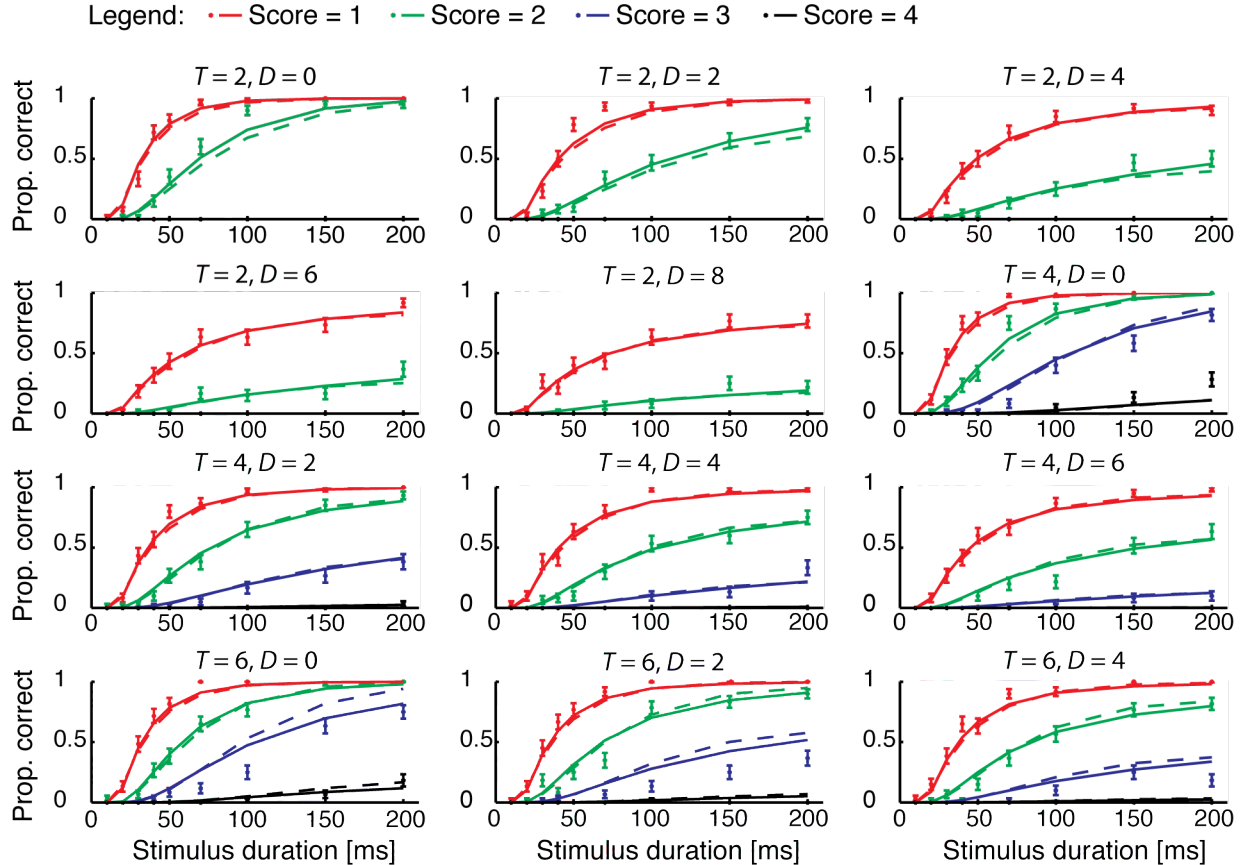


Figure 1: Cumulative score proportions for Shibuya and Bundesen's (1988) whole and partial report experiment with estimates from the conditional (solid line) and race (dashed line) encoding models using the log-logistic psychometric function.  $T$  and  $D$  at the top of each graph indicate the number of presented target and distractor objects respectively.

model fits do not provide strong evidence in favor of one model of selection over another. We find that this is a strong point as the race model has remained unchallenged as the model of selection for two decades of TVA based research.

Parameter estimates varied very little with the type of psychometric function and encoding model. The only consistent differences in parameter estimates between the two encoding models were in the processing capacity,  $C$ . This difference should however be compared to the variability within observers estimated by Finke et al. (2005) to be as high as 20% although variability between observers can also be as little as 10% (Vangkilde, et al., 2011), which is far less than the differences observed between clinic populations and normal controls (Starrfelt, Habekost, & Leff, 2009; Vangkilde, et al., 2011) yet greater than the differences between the models tested here. We therefore preliminarily conclude that parameter estimation is robust to variations in the type of psychometric function and encoding model with the caveat that studies of greater populations than the two observers studied may reveal greater variability.

Model comparison should be based on the models' ability to describe the data, here given by the goodness-of-fit; model flexibility, here given by the number of free parameters; but also on model interpretability. The interpretation of the race model is straightforward; it explicitly gives a mechanism for selection of objects to be encoded into VSTM. On this point the conditioning model is vague. We do not understand what mechanism, cognitive or neural, that could implement selection by conditioning but find that this is an interesting topic for future studies.

### Acknowledgments

The authors thank Claus Bundesen and Hitomi Shibuya for making the data from their whole and partial experiment (Shibuya & Bundesen, 1988) available.

### References

Bundesen, C. (1990). A Theory of Visual Attention. *Psychological Review*, 97(4), 523–547.

- Bundesen, C., Habekost, T., & Kyllingsbæk, S. (2005). A Neural Theory of Visual Attention: Bridging Cognition and Neurophysiology. *Psychological Review*, 112(2), 291–328.
- Dyrholm, M., Kyllingsbæk, S., Espeseth, T., & Bundesen, C. (2011). Generalizing parametric models by introducing trial-by-trial parameter variability: The case of TVA. [doi: 10.1016/j.jmp.2011.08.005]. *Journal of Mathematical Psychology*, 55(6), 416–429.
- Finke, K., Dodds, C. M., Bublak, P., Regenthal, R., Baumann, F., Manly, T., et al. (2010). Effects of modafinil and methylphenidate on visual attention capacity: a TVA-based study. *Psychopharmacology (Berl)*, 210(3), 317–329.
- Habekost, T., & Starrfelt, R. (2009). Visual attention capacity: a review of TVA-based patient studies. *Scand J Psychol*, 50(1), 23–32.
- Kyllingsbæk, S., Markussen, B., & Bundesen, C. (2011). Testing a poisson counter model for visual identification of briefly presented, mutually confusable single stimuli in pure accuracy tasks. *J Exp Psychol Hum Percept Perform*.
- Logan, G. D. (1996). The CODE theory of visual attention: an integration of space-based and object-based attention. *Psychol Rev*, 103(4), 603–649.
- Logan, G. D., & Gordon, R. D. (2001). Executive control of visual attention in dual-task situations. *Psychol Rev*, 108(2), 393–434.
- Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Res*, 46(28), 4646–4674.
- Petersen, A., & Andersen, T. S. (2012). The effect of exposure duration on visual character identification in single, whole and partial report. *Journal of Experimental Psychology: Human Performance and Perception*, In Press.
- Shibuya, H., & Bundesen, C. (1988). Visual Selection from Multielement Displays: Measuring and Modeling Effects of Exposure Duration. *Journal of Experimental Psychology Human Perception Performance*, 14(4), 591–600.
- Sperling, G. (1960). The Information Available in Brief Visual Presentations. *Psychological Monographs*, 74(11), 1–29.
- Starrfelt, R., Habekost, T., & Leff, A. P. (2009). Too little, too late: reduced visual span and speed characterize pure alexia. *Cereb Cortex*, 19(12), 2880–2890.
- Vangkilde, S., Bundesen, C., & Coull, J. T. (2011). Prompt but inefficient: nicotine differentially modulates discrete components of attention. *Psychopharmacology (Berl)*, 218(4), 667–680.

# The Development of Second-order Social Cognition and its Relation with Complex Language Understanding and Memory

**Burcu Arslan (barslan.cogs@gmail.com)**

**Annette Hohenberger (hohenberger@ii.metu.edu.tr)**

Department of Cognitive Science, Middle East Technical University, Üniversiteler Mahallesi,  
Dumlupınar Bulvarı, No: 1, 06800, Ankara, Turkey

**Rineke Verbrugge (l.c.verbrugge@ai.rug.nl)**

Institute of Artificial Intelligence, University of Groningen, P.O. Box 407,  
9700 AK Groningen, The Netherlands

## Abstract

In this study, the development of second-order social cognition and its possible relationship with language and memory were investigated. For this reason two second-order false belief tasks (FBT\_2), a short term memory task (WST), a complex working memory task (LST), a linguistic perspective-taking test (PTT) and a double-embedded relative clause task (REL\_2) were used with 21 Turkish kindergarten children (aged 4-5 years), 47 primary school children (aged 6-12 years) and 10 adults. A general developmental trend was found for all tasks. However, a multiple linear regression showed that once age was partialled out, none of the other tasks could predict FBT\_2 scores. Our findings are consistent with the modularity view that Theory of Mind (ToM) is a faculty of the human mind that does not share intrinsic content with other faculties such as language and memory (Leslie et al., 2004) and also with Apperly's (2011) 'two-systems' account of Theory of Mind. However, it develops together with those other faculties which may constrain the expression of children's false belief reasoning as a manifestation of their social cognitive abilities.

**Keywords:** Second-order Social Cognition; Cognitive Development; Theory of Mind; Language; Memory

## Introduction

In daily life, we are constantly in interaction with other agents, such as co-workers, friends and family members. As a result of this interaction, we form models pertaining to the different mental states of other agents. Social cognition of individuals is shaped based on these models. The ability to understand that different agents have different mental states, such as desires, beliefs, knowledge and intentions, which can be different from one's own, is called Theory of Mind (ToM) (Premack & Woodruff, 1978). Zero-order, first-order, second-order and higher-order reasoning are different levels of social cognition. The objects of zero-order reasoning are the rules of nature and our real-life environment. For instance, if David knows: "There is an apple on the table", he is applying zero-order reasoning. However, in daily life we are not just talking about world facts. Social interaction covers statements such as "David thinks Jessica knows that there is an apple on the table". In this situation David is applying first-order reasoning by attributing a mental state to Jessica. In addition to first-order reasoning, there are even more complex social situations

like "Jack thinks David knows that Jessica knows that there is an apple on the table". This time, Jack is applying second-order reasoning by attributing first-order reasoning to David who attributes a mental state to Jessica who reasons about an object in the real world; therefore we, in turn, are attributing third-order reasoning to Jack. In this study we follow Verbrugge (2009) in using the term 'second-order social cognition' in the same sense as 'second-order theory of mind'.

First-order ToM develops between the ages of three and five (Wimmer & Perner, 1983). Interestingly, second-order ToM develops much later, between the ages of six and nine (Perner, 1988). The reason for this gap is not entirely clear yet and attracts the curiosity of researchers who are working on ToM. In Verbrugge (2009), it is hypothesized that the developmental latency between first and second-order social reasoning is due to children's constraints on serial processing rather than limitations in simple working memory capacity. More explicitly, 6 year-old children may have the ability to represent another person's mental state about their own mental state. However, they cannot apply this ability because of the lack of efficiency in applying the related mental processes serially (cf. Hendriks, Van Rijn, & Valkenier, 2007).

One of the most widely applied verbal paradigms for studying ToM is the false-belief task (FBT), which has first been studied by Wimmer and Perner (1983). The main idea of the FBT is to examine whether children can attribute a false belief to another agent in a given story where they know the reality and the other agent has a false belief. Using language comprehension tasks is another verbal paradigm for studying the development of social cognition. These tasks generally test listeners' semantic and/or pragmatic inferential abilities. In these tasks, the listener has to take the speaker's linguistic alternatives and his/her choice into account to understand the correct meaning of the sentence. In the present study, a complex language comprehension task was used to test children's ability to meet a questioner's expectations of an appropriate answer to his / her questions in a given context, by taking the questioner's perspective.

The development of ToM has been intensely investigated and documented in the literature. However, one of the debatable issues is still if other factors contribute to ToM

understanding during development. There is one influential factor, namely language development (Astington & Baird, 2004; Flobbe, Verbrugge, Hendriks, & Krämer, 2008): Does language have an effect on acquiring ToM, or not? In order to elucidate the relationship between language and social reasoning during development, two language tasks were used in this study. The first one is a complex language comprehension task in which the morphological structure in particular zero- vs. accusative-marked nouns had to be mastered and the second one is a double-embedded relative clause task in which complex syntactic structure was required. Generally, complement clauses are studied in the literature in order to investigate the relationship between the syntactical component of language and ToM. Unlike complement clauses, relative clauses do not necessarily involve mental state predicates. Using relative clauses instead of complement clauses allows us to specifically focus on the structural format of 2-way embedding. This is a purely structural parallel between second-order embedding in the thought domain and second-order embedding in the language domain.

## Method

### Participants

A total of 68 (35 female, 33 male,  $M=7.53$  yrs,  $SD=2.53$ ) children and 10 (5 female, 5 male,  $M=33.48$  yrs,  $SD=10.00$ ) adults participated in the experiments. Children's grades varied from kindergarten to fifth-grade, and their age range varied from 4 to 12 years. There were 21 kindergarten children ( $M=4.43$  yrs,  $SE=.07$ ), 17 first graders ( $M=6.99$  yrs,  $SE=.09$ ), 15 third graders ( $M=9.01$  yrs,  $SE=.08$ ) and 15 fifth graders ( $M=11.00$  yrs,  $SE=.10$ ). A group of 10 adults served as a control group.

### Design

A cross-sectional study with the four above-mentioned age groups was conducted. All subjects participated in the following five tests: word span task, second-order false belief task, perspective-taking test, second-order relative clause task, and listening span task.

All of the tests were completed in one session, which varied from 25 minutes to 35 minutes. Children were tested in a quiet empty classroom at their school.

### Materials and Procedures

**Word Span Task (WST)** Children's short term memory was tested with Ünal's (2008) Turkish version of the WST. Mono-syllabic Turkish words such as "saç, tuz" and "yurt" (hair, salt and country) were selected considering their frequency in daily usage and easy pronunciation. There are a total of seven sets, which consist of 2 to 8 words. Each set comprised 3 sub-sets. An example of a set of 2 words is as follows: i) köşk muz (manor banana); ii) pil üst (battery upper); iii) buz dört (ice four).

The words from these sets were read to the participants starting from the 2-word set. After reading one set (i.e. köşk muz), the participant repeated the words in that order. If the participant made two errors, i.e., any error in two of the three sub-sets of that level, the experiment was terminated. If he/she made fewer than two errors, the subsequent, next higher, set was read (i.e. the 3-word set). The word span equals the correct number of words at the respective level at which the child makes fewer than two errors. Thus, in the analysis the word span range may vary between 1 and 8.

**Second-order False Belief Task (FBT\_2)** This task consists of two different second-order false belief stories, namely the 'Birthday Puppy' Story and the 'Chocolate Bar' Story. Both stories were adapted from English to Turkish from Flobbe et al. (2008) with the authors' permission. These stories were told to the subjects by presenting Flobbe et al.'s (2008) drawings. Second-order embedding structures such as "Mary thinks that John thinks the chocolate is in the drawer" were not used in the stories.

For both stories, the drawings were shown to the participants while the stories were being told. The order of stories in the false belief task was balanced. If a participant gave correct answers to the reality control (Where is the chocolate now?), first-order ignorance (Does John know that Mary has hidden the chocolate in the toy chest?), linguistic control (Does Mary know that John saw her hide the chocolate?), second-order false belief (Where does Mary think that John will look for the chocolate?) and justification (Why does she think that?) questions, the participant's score of the first story was 1. The total score for both of the false belief stories is therefore minimally 0 and maximally 2. Since the questions preceding the second-order false belief question are control questions, they need to be answered correctly in order for the false belief question to be possibly answered correctly.

**Perspective-taking Test (PTT)** The perspective-taking test is a complex language comprehension task including two close-ended questions with two options. In the story, Ali tells Ayşe that he is planning to go to the bookstore today. Ayşe wants Ali to buy a storybook. Ali goes to the bookstore and buys the book. While Ali is going back home, he sees his friend Mehmet on the road. Mehmet asks Ali what he did today. At this point, the experimenter asks the participant which of the following answers, (a) or (b), Ali gives to Mehmet:

- a) Kitab-ı al-dı-m. (I bought the book.)  
Book-ACC buy-PAST-1PSg
- b) Kitap al-dı-m. (I bought a book.)  
Book buy-PAST-1PSg

After that, the experimenter continues to tell the story. Ali goes back home. Ayşe opens the door and asks Ali what he did today. This time the experimenter asks the participant which of the following answers, (a) or (b), Ali gives to Ayşe: a) Kitabı aldım or b) Kitap aldım. The order of the answers to the close-ended questions provided to the



subjects was balanced across participants. Since Mehmet asks the question without having been introduced to the book before, the expected answer for the first question was the answer with zero-marking: “Kitap aldım” (referring to “a book”). The reasoning behind this answer is as follows: Ali knows that Mehmet does not know that Ali went to the bookstore to buy a storybook for Ayşe. However, the expected answer for the second question was “Kitabı aldım” (referring to “the book”) rather than “Kitap aldım”, since Ayşe had told Ali that she wanted him to buy a storybook. The reasoning behind this answer is as follows: Ali knows that Ayşe wants to know whether Ali bought the storybook or not.

If the participant gave the expected answer to the two questions, s/he received a score of 2 points.

**Double-embedded Relative Clause Task (REL\_2)** This task is related to the comprehension of relative clauses (RC) in Turkish. This task was adapted from Özge (2010) with the author’s permission. The questions and the drawings were modified to double-embedded RCs in order to analyze the participants’ second-order language embedding abilities, on a par with their second-order ToM abilities. One practice and 6 experimental items were used. Figure 1 demonstrates the drawings for one of the questions related to this task. The critical positions for finding the correct answers were equally distributed across the drawings (3 times in the first row and 3 times in the second row) and between right (2 times), left (2 times), and central position (2 times).

First, introductory pictures were shown to the participants in order to familiarize them with the animals in the action by telling the name of the animals and the actions (e.g., “this is a pushing sheep”). After that, the pictures representing the questions were shown one by one (see Figure 1).

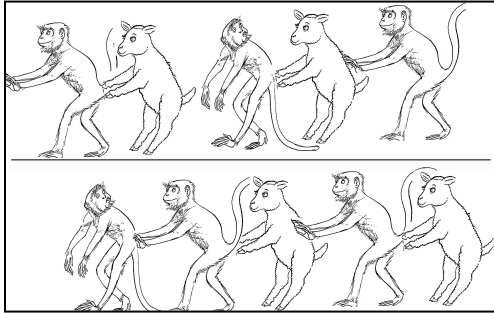


Figure 1: Hangi resimde kuzuyu iten maymunu iten bir kuzu var? (“In which picture there is a sheep pushing a monkey that is pushing a sheep?”)

The first and second rows of the picture were pointed out in order to make it clear that there are two separate lines of pictures by saying, “This is the first picture and this is the second picture”. In the trial session, it was explained that the participants were required to point out the row with the animals related to their answer. If they could not answer correctly in the trial, the correct animals were pointed out by the first author with necessary explanations. The sentences

were repeated up to 4 times. Participants’ scores could range from 0-6.

**Listening Span Task (LST)** In order to measure complex working memory, Ünal’s (2008) English-to-Turkish adaptation of the original LST was used with the author’s permission. The task consists of sets of sentences read out to the participants one by one. There are a total of five collections, each of which consists of six sets of sentences. The first collection contains six sets of two sentences each, the second collection contains six sets of three sentences each, and so forth, until the fifth collection, which contains six sets of six sentences each. An example of a 3-sentence set of LST is as follows: i) Muzlar bisiklete biner. (Bananas ride bicycles); ii) Elimiz beş parmaklıdır. (Our hands have five fingers); iii) Soğan acıdır. (Onions are hot).

The participants were expected to first judge the truthfulness of the sentences by saying “Yes” or “No”. Secondly, they had to recall the last word of all the sentences of a set told to them, in reverse order. After they gave an answer to the first sentence, the next sentence was told to them. For example, for the 2-sentence set, if the first sentence is “Muzlar bisiklete biner.” (Bananas ride bicycles), the participants were required to say “Hayır<sup>1</sup>;biner”. After that, if the second sentence is “Elimiz beş parmaklıdır.” (Our hands have five fingers), they were required to say “Evet<sup>2</sup>;parmaklıdır, biner.”. If the participant made less than two mistakes in a sentence collection, the subsequent sentence set, which comprised one more sentence, was told to the participant. The score of the participants equaled the number of sentence collections in which they did not make more than one mistake. Participants’ scores could range from 0-6.

## Hypotheses

We hypothesized main effects of grade for all tasks. Children’s performance should increase with increasing grade. Furthermore, we hypothesized that FBT\_2 could be predicted by the remaining tasks, in particular by the complex language tasks, PTT and REL\_2, to the extent that those share variance with it.

## Results

First, the statistical analyses of children’s responses to the five tasks are presented. Later, the results of the fifth graders will be compared with the results of adults. The p values are two-tailed, unless stated otherwise (in which case they are one-tailed). In order to analyze the developmental trend in the tasks used in the experiment, the data was divided into four groups according to the children’s grades (kindergarten, 1st, 3rd, 5th grade). Since the data violates normality assumptions, non-parametric Kruskal-Wallis and Mann-Whitney tests were used. Since six Mann-Whitney Tests were used to test the difference across the grades, the

<sup>1</sup> ‘Hayır’ means ‘No’.

<sup>2</sup> ‘Evet’ means ‘Yes’.

alpha level for the Bonferroni correction was set to .008. Although the data was not normally distributed, linear regression analysis was used to investigate the relationship between the second-order false belief task and the other tasks. Error bars in Figures 2-6 represent standard errors.

### FBT\_2

There is a significant difference in performance between the grades ( $\chi^2(2) = 40.22, p = .000$ ). According to the Mann-Whitney Tests, while there is a steady increase in performance, there is no significant difference between the first and third grades and between the third and fifth grades. However, there is a significant difference between kindergarten and grade one ( $Z = -3.73, p = .000$ ), kindergarten and grade three ( $Z = -4.73, p = .000$ ), kindergarten and grade five ( $Z = -5.36, p = .000$ ), and grade one and five ( $Z = -2.99, p = .003$ ). Figure 2 shows the mean values of the FBT\_2 score according to the grades. Since all of the adults and all of the fifth graders answered all of the questions correctly, there is no significant difference between the adults' and fifth graders' FBT\_2 performance ( $\chi^2(2) = 0.00, p = 1.00$ ).

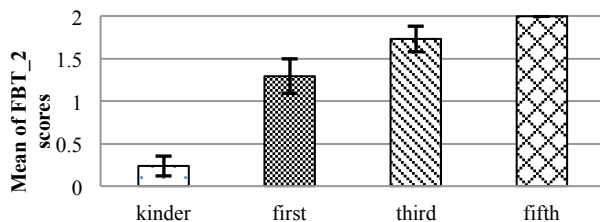


Figure 2: The development of FBT\_2

### WST

There is a significant difference between the grades ( $\chi^2(2) = 24.67, p = .000$ ). According to the Mann-Whitney Tests, there is no difference between the first, third and fifth grades, while there is a significant difference between kindergarten and grade one ( $Z = -3.06, p = .002$ ), kindergarten and grade three ( $Z = -4.14, p = .000$ ), and kindergarten and grade five ( $Z = -3.59, p = .000$ ). Figure 3 shows the mean values of the Word Span Task score according to the grades. The analysis showed that there is a significant difference between the adults' and fifth graders' performance ( $\chi^2(2) = 8.925, p = .003$ ).

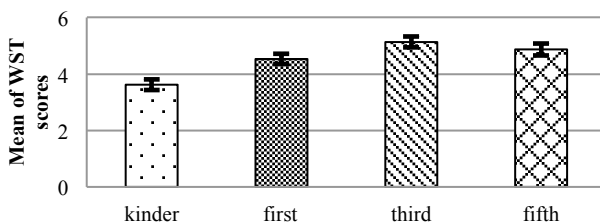


Figure 3: The development of WST

### PTT

There is a significant difference between the grades ( $\chi^2(2) = 8.53, p = .036$ ). According to the Mann-Whitney Tests, there is no difference between the kindergarten and grade one, grade one and three, grade one and five, grade three and five, while there is a significant difference between the kindergarten and grade five ( $Z = -2.473, p = .006$ , one-tailed). Figure 4 shows the mean values of the perspective-taking test score according to the grades. The analysis showed that there is no significant difference between the adults' and fifth graders' performance ( $\chi^2(2) = 1.778, p = .182$ ).

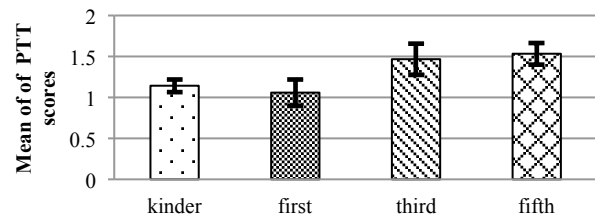


Figure 4: The development of PTT

### REL\_2

There is a significant difference between the grades ( $\chi^2(2) = 27.37, p = .000$ ). In order to see which grades differ significantly from each other, Mann Whitney Tests were used. Figure 5 shows the mean values of the double-embedded relative clause score according to the grades. The analysis showed that there is a significant difference between the adults' and fifth graders' performance ( $\chi^2(2) = 6.096, p = .014$ ).

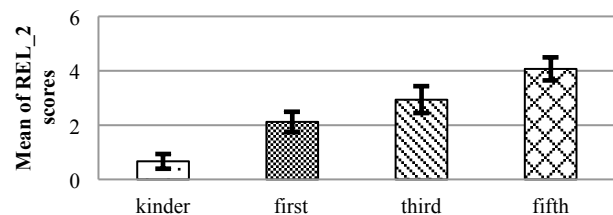


Figure 5: The development of REL\_2

### LST

There is a highly significant difference between the grades ( $\chi^2(2) = 30.87, p = .000$ ). According to the Mann-Whitney Tests, there is no difference between the kindergarten and first grade, nor between third and fifth grades, while there is a significant difference between the kindergarten and grade three ( $Z = -3.53, p = .000$ ), kindergarten and grade five ( $Z = -4.64, p = .000$ ), grades one and three ( $Z = -2.92, p = .003$ ), and grades one and five ( $Z = -4.08, p = .000$ ). Figure 6 shows the mean values of the Listening Span Task score according to the grades. The analysis showed that there is a significant difference between the adults' and fifth graders' performance ( $\chi^2(2) = 4.729, p = .030$ ).

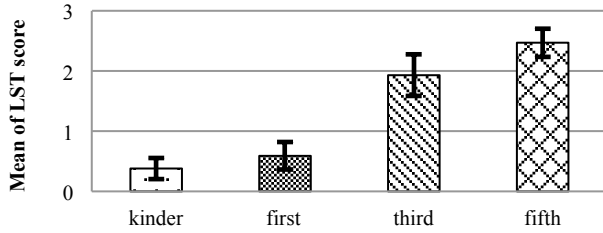


Figure 6: The development of LST

### PTT Predicting FBT\_2

Since the data violates normality, the non-parametric Spearman's Rank Order Correlation was used to test the relationship between total FBT\_2 and PTT scores. This analysis showed that there is no significant relationship between total FBT\_2 and PTT ( $r_s = .19$ ,  $p = .126$ ). Partial correlation was also used in order to control the other variables. When age is controlled for, the previous correlation of  $r_s = .19$  between PTT and FBT\_2 drops to  $-.095$  ( $p = .444$ ); when WST is controlled for, the correlation drops to  $.12$  ( $p = .922$ ), when REL\_2 is controlled for, the correlation drops to  $.036$  ( $p = .772$ ) and when LST is controlled for, the correlation drops to  $-.22$  ( $p = .860$ ).

### REL\_2 Predicting FBT\_2

Since the data violates normality, the non-parametric Spearman's Rank Order Correlation was used to test the relationship between total FBT\_2 and REL\_2. This analysis showed that there is a significant relationship between total FBT\_2 and REL\_2 scores ( $r_s = .54$ ,  $p = .000$ ). Bivariate regression was also used in order to predict the model of REL\_2 score predicting the FBT\_2 score. Using the enter method, the FBT\_2 score could be predicted from the REL\_2 score by the following formula:  $0.24 \times \text{REL}_2 + 0.673$  ( $F_{66,1} = 26.196$ ,  $p = .000$ ,  $r = .533$ ,  $R^2 = .284$ ). Partial correlation was also used in order to control the other variables. When age is controlled for, the previous correlation of  $r = .533$  between REL\_2 and FBT\_2 drops to  $.10$  ( $p = .42$ ); when WST is controlled for, the correlation drops to  $.39$  ( $p = .001$ ); when PTT is controlled for, the correlation drops to  $.52$  ( $p = .000$ ) and when LST is controlled for, the correlation drops to  $.25$  ( $p = .041$ ).

In the light of the partial analyses, multiple linear regression was used with age and REL\_2 scores as independent variables and FBT\_2 as dependent variable. Using the enter method, the FBT\_2 score could be predicted from age and REL\_2 score by the following formula:  $\text{FBT}_2 = 0.039 \times \text{REL}_2 + 0.25 \times \text{age} - 0.751$  ( $F_{65,2} = 42.091$ ,  $p = .000$ ,  $r = .751$ ,  $R^2 = .564$ ). However, only the contribution of age is significant ( $\beta = .692$ ,  $t = 6.47$ ,  $p = .000$ ).

### Multiple Linear Regression for FBT\_2

Two models were constructed by using multiple linear regression to predict the FBT\_2 score: first just with age as predictor and second with age and all tasks. Table 2 shows the correlations of all tasks for FBT\_2. Using the enter

method, the FBT\_2 score could be predicted by age by the following formula:  $0.27 \times \text{age} - 0.814$  ( $F_{66,1} = 83.965$ ,  $p = .000$ ) and by age and all tasks by the following formula:  $0.236 \times \text{age} + 0.145 \times \text{WST} + 0.045 \times \text{REL}_2 - 0.034 \times \text{LST} - 0.130 \times \text{PTT} - 1.098$  ( $F_{62,5} = 17.519$ ,  $p = .000$ ,  $r = .765$ ,  $R^2 = .586$ ). However, only the contribution of age is significant ( $\beta = .655$ ,  $t = 5.45$ ,  $p = .000$ ). Collinearity diagnostics showed that age (94%) and WST (90%) each load highly on a different single dimension. This means that age and the WST do not share variance with each other. On the other hand, PTT, REL\_2 and LST share some variance with one another. Still they mainly load on their own distinctive dimension. This is because they are also related to different abilities. Moreover, the LST (60%) and the REL\_2 (75%) load highest on the same dimension which shows that both tasks tap into the same cognitive ability.

Table 2: Correlations of all tasks and age for FBT\_2

Variable	Correlation (r)	p
Age	.748	.000**
WST	.518	.000**
PTT	.160	.096 n.s.
REL_2	.533	.000**
LST	.503	.000**

### General Discussion and Conclusions

As can be seen clearly from Figure 2, a linear developmental trend was found for the second-order false belief task score from grade one (6- 7 years) to grade five (10- 12 years), preceded by a big step between kindergarten and grade 1. We can say that second-order false belief reasoning starts to develop around the age of 6, and reaches adult-like understanding at around the age of 9;5 (grade 5). These findings are compatible with Perner and Wimmer's (1985) study, which states that second-order false belief understanding occurs after the age of 6. Although kindergarten children failed in the second-order false belief task on average, there were three of them who succeeded in the 'Birthday Puppy' Story and one of them who succeeded in both the 'Birthday Puppy' Story and 'Chocolate Bar' Story. Since their Listening Span Task and double-embedded relative clause task scores were better than the others, this is compatible with the view that children before the age of 6 may indeed have second-order social cognition which may show itself if the respective cognitive resources have also developed.

The results of the Word Span Task showed a significant and clear developmental trend from kindergarten to third grade. Fifth graders' score was somewhat lower than that of the 3rd graders, but only insignificantly.

The results of the Perspective Taking Test showed that kindergarten children and first graders had scores around 1 which is the score expected by chance. The salient development occurs between 1st and 3rd grade. Making pragmatic inferences by picking up morpho-syntactic clues

like case-marking is a very advanced meta-linguistic skill. Giving correct answers to the questions requires a comparison between the two case forms and a decision which of them is better suited for the given context. Even adults' performance was not perfect and did not significantly differ from that of 5th graders. However, unlike children, some of the adults changed their first wrong answer and gave a correct answer after hearing the second question. This shows that some of the adults took the hearer's perspective and/or the experimenter's intention of asking those questions into account.

To the best of our knowledge this is the first time that a double-embedded relative clause task has been devised in a Turkish developmental study. Our result revealed a very strong developmental trend in the task. Also, adults outperformed fifth graders in this task.

In Listening Span Task, participants were expected to judge the semantic truth of the sentences, to report it, to remember the last word of that sentence, and then repeat the same steps again for the next sentence by also reporting the last word of the previous sentence, and so on. Since in Turkish the present form of the verb takes the suffixes –er, -ar, -ir, -ür, -ur for positive sentences and takes the suffixes –maz, -mez for the negative ones, the most challenging part of the task for children and even for some adults was to repeat the last word of the sentence when its semantic truth was false. So, they must inhibit the regular way of reporting, and have to report it in the instructed form. This inhibition in the Listening Span Task is thought to be similar to the inhibition necessary in false belief reasoning. The results showed a strong developmental trend, again particularly between first and third graders.

Even though second-order false belief scores could be significantly predicted by all other tasks (except for Perspective Taking Test), the regression analyses showed that only the contribution of age was significant. Once age was taken out, none of the other tasks could predict second-order false belief task anymore. In view of theoretical accounts of ToM, our findings are compatible with Leslie et al.'s (2004) modular account of ToM. He and his colleagues argue that ToM is a separate cognitive faculty as compared to language or memory. It is innate, i.e., in principle in place from early on, however, in order to manifest itself it may need to await the cognitive maturation of the child in those other domains. The “serial bottleneck” (Verbrugge 2009) might be one of those constraints that is overcome during development. Our findings are also compatible with Apperly's (2011) 'two-systems' account of ToM. Apperly posits that low-level efficient processing modules take care of simple ascriptions of perception, knowledge and belief, while high-level ToM makes use of less efficient and slower to develop general knowledge and inferential reasoning processes.

Since in our study we found concurrent development in all the cognitive abilities that we tested, that is, no delay between any of them, ToM may at any time have been supported just sufficiently enough to manifest itself at that

level. Indeed, it might be impossible to prove the relation between ToM and the other cognitive domains in a cross-sectional study like ours, but only in a longitudinal study where such delays may be observed within rather than across individuals.

## Acknowledgments

Rineke Verbrugge is grateful to the Netherlands Organization for Scientific Research for Vici grant NWO-277-80-01.

## References

- Apperly, I (2011). *Mindreaders: The Cognitive Basis of "Theory of Mind"*. Psychology Press; Hove.
- Astington, JW., and Baird, J. (2005). *Why Language Matters for Theory of Mind*. Oxford University Press; Oxford.
- Flobbe, L., Verbrugge, R., Hendriks, P. and Krämer I. (2008). Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information*, 17 (4), 2008, 417-442. Special issue on formal models for real people, edited by M. Counihan.
- Hendriks, P., van Rijn, H., and Valkenier, B. (2007). Learning to reason about speakers' alternatives in sentence comprehension: A computational account. *Lingua*, 117(11), 1879–1896.
- Leslie, Alan M., Friedman, Ori, and German, Tim P. (2004). Core mechanisms in 'theory or mind'. *Trends in Cognitive Sciences*, 12, 528-533.
- Özge, D. (2010). *Mechanisms and Strategies in the Processing and Acquisition of Relative Clauses in Turkish Monolingual and Turkish-English Bilingual Children*. PhD thesis, Middle East Technical University, Ankara-Turkey.
- Perner, J. (1988). Higher-order beliefs and intentions in children's understanding of social interaction. In J.W. Astington, P.L. Harris, and D.R. Olson, (Eds.). *Developing Theories of Mind*. Cambridge University Press, Cambridge. 271–294.
- Perner, J. and Wimmer, H. (1985). "John thinks that Mary thinks that...": Attribution of second-order beliefs by 5- to 10-year old children. *Journal of Experimental Child Psychology*, 5, 125-137
- Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 4, 515-526.
- Ünal, G. (2008). *Release from Proactive Interference and its relations to Executive Functions: A developmental study on Turkish children*. Master thesis, Middle East Technical University, Ankara-Turkey.
- Verbrugge, R. (2009). Logic and social cognition: The facts matter, and so do computational models. *Journal of Philosophical Logic*, 38 (6), 649-680.
- Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 103–128.

# Learning of motor maps from perception: a dimensionality reduction approach

Ankit Awasthi, Sadbodh Sharma, Amitabha Mukerjee (aawasthi, sadbodh, amit@cse.iitk.ac.in)

Department of Computer Science and Engineering, IIT Kanpur

## Abstract

The role of perception in sighted infant motor development is well-established, but what are the processes by which an infant manages to handle the complex high-dimensional visual input? Clearly, the input has to be modeled in terms of low-dimensional codes so that plans may be made in a more abstract space. While a number of computational studies have investigated the question of motor control, the question of how the input dimensionality is reduced for motor control purposes remains unexplored. In this work we propose a mapping where starting from eye-centered input, we organize the perceptual images in a lower-dimensional space so that perceptually similar arm poses remain closer. In low-noise situations, we find that the dimensionality of this discovered lower-dimensional embedding matches the degrees-of-freedom of the motion. We further show how complex reaching and obstacle avoidance motions may be learned on this lower-dimensional motor space. The computational study suggests a possible mechanism for models in psychology that argue for high orders of dimensionality reduction in moving from task space to specific action.

**Keywords:** motor development, perception-action models, dimensionality reduction, reach learning

## Introduction

That the apparently random arm movements by neonates may have a role in terms of learning visuo-motor control is a position that has gained strength in recent years (A. Van der Meer, Weel, & Lee, 1995; Adolph & Berger, 2006). Visual control of arm movement is evident almost immediately after birth, when neonates appear to struggle with artificial handicaps to keep their arm in the field of view, such as when small weights are tied to their arms, or when they appear to be purposefully keeping their arms within a narrow beam of light (A. van der Meer, 1997).

The emphasis on motor development over the last few decades, arising from the linkage between perception and action that has driven new theories of dynamic cognition (Thelen, Smith, Lewkowicz, & Lickliter, 1994), has further highlighted this connection. Indeed, there is evidence to suggest that the very representation of action has some relation to mental imagery which may provides the prospective structure for organizing motions (Caeyenberghs, Wilson, Van Roon, Swinnen, & Smits-Engelsman, 2009). Studies on monkeys report that a large majority of neurons implicated in the reach planning area in the posterior parietal cortex, encode location in eye-centred coordinates (Batista, Buneo, Snyder, & Andersen, 1999). In addition, nearly half the neurons in the ventral premotor area - thought to be implicated in visual grasp planning - are responsive to eye-centered image data (Mushiake, Tanatsugu, & Tanji, 1997). It is thought that the eye-centric neural computations work together with other body-centric and arm-centric systems (Graziano, 2011) but much of the planning is thought to invoke eye-centered or visual images of the arm.

However, the exact mechanisms by which the complex high-dimensional visual stimulus is used for motor control of the arm, remain unclear. While there have been many computational studies of different aspects of motor development, several involving robots with embodied cameras (Beltrán-González, 2005; Hoffmann, Schenck, & Moller, 2005), such works usually address the question of managing the complexity and dimensionality reduction only peripherally (Jordan & Wolpert, 1999). A particularly interesting approach was the modeling of reaching and grasping via neural networks (Kuperstein, 1991), and a related attempt to map the workspace using self-organizing maps (Saxon & Mukerjee, 1990). However, these works do not deal directly with image models.

On the other hand, another body of work in robot motion planning attempts to construct paths through a workspace that are optimal in various senses of reducing path length, kinematic smoothness, dynamic smoothness, energy costs, etc. A class of efficient methods are based on sampling the motion space, which may also be rather high-dimensional, though of the order of tens, not thousands as in the image space. Such stochastic approaches include probabilistic roadmaps or branching trees of possible paths through the free space. However, these algorithms, some of which also have cognitive ramifications, work primarily with inputs defined directly in the workspace, and do not work on the image space as the input (Choset, 2005).

The objective of this work is then to try to develop a model that would reflect motor properties in an embedding derived purely from the visual space. Thus, we consider the mechanisms by which the system may observe certain similarities between perceived arm configurations to construct a local neighbourhood of visual configurations. These local neighbourhoods are presumed to lie on a low-dimensional surface in the very high-dimensional image space. In the ISOMAP algorithm being used here (Tenenbaum, Silva, & Langford, 2000), these local neighbourhoods are assumed to be linear in the dimensionality of the manifold, and they are then “composed” to construct the global relations between distant arm configurations. For the purposes of the demonstration we have used a simulated two degree of freedom arm.

## Reach planning and obstacle shifts

The input consists of a large number of images showing the arm in random poses in its workspace. These images are then mapped to a lower-dimensional manifold. We shall use the term *visuo-motor map* and *low-dimensional embedding* for this image to low-dimensional map. Note that the variables in the low-dimensional map are just mathematical abstractions that emerge from the computation and are not based on

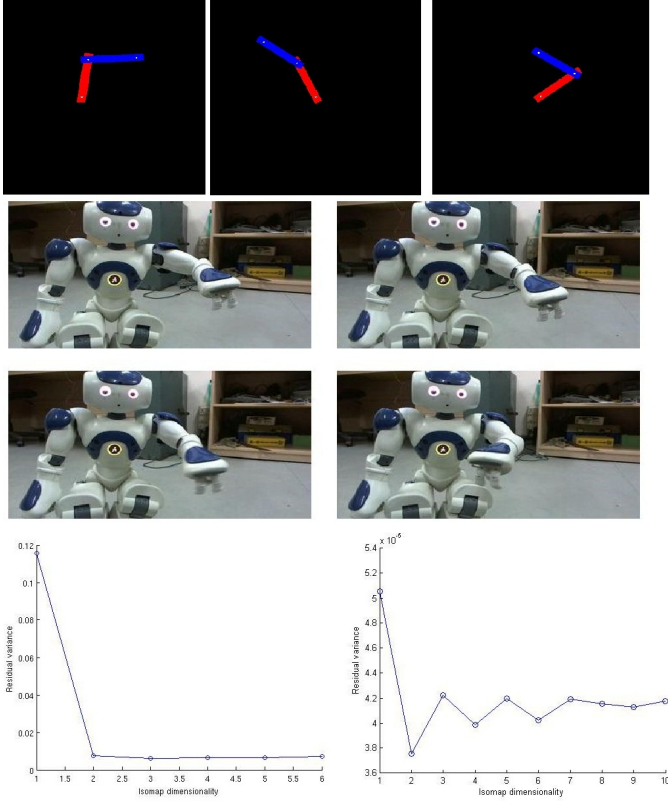


Figure 1: Image data from (a) two-degree of freedom planar arm, (b) humanoid robot nao moving one arm along a linear trajectory. The residual variance error in the lower-dimensional space after ISOMAP, versus the dimensionality of the embedding is shown below. Left: planar arm (2-DOF). Right: humanoid robot nao moving arm in straight line (1-DOF); errors are noisy,  $\sim 10^{-5}$ .

prior knowledge about any physical variable. Thus, we use no knowledge of the kinematics or geometry of the arm, except the initial image data. Despite this paucity of supervisory knowledge, we are able to identify the number of degrees of freedom (two) of the motion system, and also a set of (two) parameters that can be controlled to move the arm to different configurations within its workspace. We also show that these parameters can be easily mapped to physical variables such as joint angles.

The visuo-motor map is obtained as a graph embedded in the lower-dimensional space, each node is mapped to an image in the much higher-dimensional image space. The structure of the graph is shared between both the original image space as well as the lower-dimensional embedding space. In the following, we shall consider a) constructing such a map (with no prior knowledge about the physics of the arm), b) using it to map a path from a source to a target in the absence of obstacles, and c) map obstacles that may impinge on the path, and plan trajectories that may avoid it. A further aspect is that with increasingly dense samples in the image space,

the models become increasingly accurate.

We demonstrate the dimensionality reduction process - part (a) - on three domains - images of a simulated two-link planar arm moving to various positions in its workspace, a humanoid robot moving its arm along a straight line, and two coordinated planar arms moving a box. A two-dimensional interpolation is shown to work well for the first situation (two degree of freedom in the arm). For the other two situations, we find that one dimension is enough. For motion planning (part b), and obstacle avoidance (c), we restrict ourselves to only the first simulated arm - the two degree of freedom simulation of the upper and lower limbs.

For motion planning, both initial and desired configurations, as well as obstacles, if any, are known only in the visual space. The initial and target images are mapped to the manifold embedding using the out-of-sample mapping - i.e. they are interpolated based on the “nearest” images and the same interpolation is applied to the embedded points. Now, a mapping to the goal can be discovered in terms of edges in the graph, which actually constitute a roadmap, but in the visuomotor space and not in the usual configuration space of robot motion planning. The problem of having the visuomotor map represent obstacles is quite challenging, since some obstacles (such as the own body) may be permanent, but in most situations other constraints would arise in the workspace. As the obstacle positions shift with respect to the arm, must it relearn the visuo-motor embedding every time? This is indeed what happens to most robot-motion planning algorithms, and handling dynamic obstacles remains a considerable challenge in robotics. In our case, we assume that our map is a base map of the maximal free space, constructed in the absence of all movable obstacles. This results in an embedding from which nodes can be deleted if an obstacle is introduced to a region of the image space which was otherwise occupied by an arm in one of the data points. Such situations are handled by marking as *blocked* the non-free nodes of the graph in the lower-dimensional space. To start with therefore, we consider full rotations for each of the two joints of our simple 2-DOF arm, and images are randomly sampled across this entire  $360 \times 360$  degree rotation range.

As obstacles are introduced into the workspace, the system marks those nodes of the graph that visually overlap with the obstacle as *blocked*, and motions are restricted to the remaining “free” parts. Of course, this method assumes that if two nodes are in free space, the path between them is also free, which is not true for the non-linear mapping space. However, considering how densely sampled the image space of a limb in daily use is for an organism, we may argue that in real situations, the visuomotor map will have a dense mapping so that this assumption is far more likely to hold.

An additional characteristic of the learning process reflects increasing accuracy in the visuo-motor map. As the system encounters more and more motions in the visual space, the graph becomes more dense, and the accuracy of plans improves. This may also contribute to the observations of in-



creasing fluency in infant motions, though much of it is also due to development in the musculatory system.

## Dimensionality Reduction

We consider a computational system exposed to a large number of images reflecting different motor configurations of its arms. At the same time, it has muscle sensations of how the arm motion is executed. In this work, we focus primarily on the visual input space. We first observe that the data there, captured here in terms of a set of large images, constitute an extremely high dimensional image space. Every pixel may be considered to be independent, so that a  $800 \times 800$  image (figure. 1(a)), is a point in a  $64000$  dimensional space. That is each image is one point out of  $N^{64000}$ , where  $N$  is the number of values a pixel can take. Clearly, the system needs to reduce the complexity of this data drastically. In the biological system, this is achieved by a combination of neural data compaction processes starting in the retina, as well as a number of learned responses that eventually result in responses in the pre-motor areas, primarily in the posterior-parietal cortex (PPC) (Batista et al., 1999). Here we consider a computational model for constructing a lower-dimensional manifold from the image data.

Many approaches have been tried for dimensionality reduction. One class of approaches assume that the underlying manifold from which the observations are drawn is linear. Linear dimensionality reduction methods include linear methods such as Principal Component Analysis or Multi-dimensional Scaling (Bishop, 2006). On the other hand, non-linear dimensionality reduction assumes that the data is drawn from a manifold - a surface that is mappable everywhere to a disk in a euclidean space of much lower-dimension. Such NLDR algorithms include Isomap, Locally-Linear Embedding (LLE), Local Tangent Space Alignment, etc. (Lee & Verleysen, 2007). In this work, we have used the ISOMAP algorithm, which attempts to preserve geodesic distances in the lower-dimensional mapping.

### Isomap on the image space

To see how the ISOMAP algorithm is applied to our image space, we observe that a continuous motion will result in a sequence of images that lie along a one-dimensional curve in the image space. This is because images from successive instants in time are likely to be very similar, and extending this similarity locally would result in a 1-manifold. This applies to complex real images (figure. 1b) Now, let us consider a two-degree of freedom arm. For the two-degree of freedom, it turns out that all such one-dimensional trajectories in the image space are found to lie on a surface, which is everywhere mappable to a disk in the euclidean plane ( $R^2$ ). Thus, all these images lie on a curve 2-dimensional manifold in the image space (figure. 1). Distances between configurations can be thought of as distances between the shortest paths (geodesics) that lie on this manifold. The Isomap essentially solves an optimization problem where the lower-dimensional points are to

be chosen so as to minimize the error in these geodesic distances. This is the *residual error* which is reported in the graphs in figure 1c. If there is a sharp drop in the error at some value of the dimension used, then one assumes that this dimension is a good estimate for the dimensionality of the underlying manifold. In some cases (e.g. for the Nao robot images), all the errors are less than  $10^{-5}$  - this implies that the error is already very low at dimension one, so  $d=1$  is a good estimate for its dimensionality.

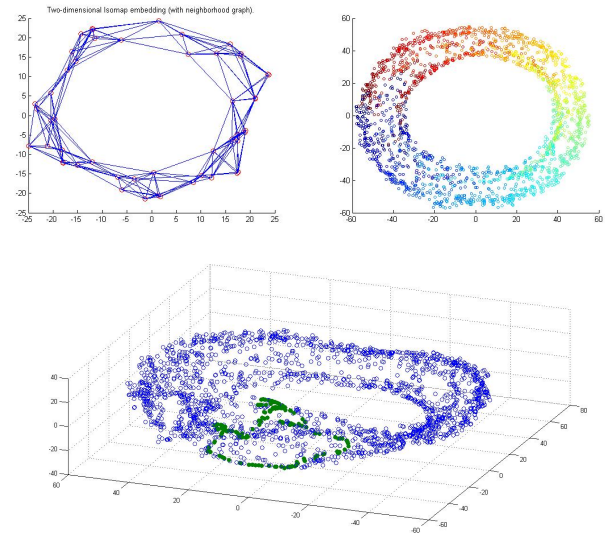


Figure 2: Two-dimensional embeddings generated by ISOMAP. Top Left: map from 50 robot images. Top Right: map from 2000 images showing variation of  $\theta_1$  along the manifold. Bottom: map in 3d showing a cross-section with  $\theta_1$  kept constant and  $\theta_2$  varying. Note that the topology of the embedding actually reflects the topology of the  $\theta_1, \theta_2$  space - in that the  $\theta_1$  variation occurs along the ring (top right figure), while the  $\theta_2$  variation is approximately radial (bottom). Note that the 50-case also captures the cyclic nature of  $\theta_1$ , though the resolution is poor. This suggests the torus topology of the  $\theta_1, \theta_2$  space.

### Out of sample points

A significant difficulty with non-linear dimensionality approaches is that these models are not very good for mapping novel (unseen) data points. In order to overcome this very significant difficulty, some approaches such as the Locality Preserving Projections (LPP) (Lee & Verleysen, 2007). Here a linear mapping is constructed based on a quadratic function that attempts to preserve some aspects of the non-linear data. However, the inverse mapping (from lower dimension to the original space) in such systems is poor, and we have restricted ourselves to true non-linear models.

In order to solve the out of sample interpolation problem, we have borrowed an idea from Locally-Linear Embedding, where a local approximation can be constructed for novel



points based on a weighted linear approximation in the image space. Now, the same weights are used in the lower-dimensional space to obtain a mapping for the point in that space. Indeed, in the LLE, the fact that these weights remain the same in the lower-dimensional space is the constraint that is the basis for its optimization. The underlying assumption is that locally, the nearest neighbours around the novel point may be approximated by a linear surface in both the image space and the target space. While the Isomap algorithm is premised on a somewhat different (global geodesic) constraint, the locally linear approximation is not too far off for the isomap, especially if the sampling is dense. Thus, this method is here adopted to our isomap data; though it is only approximate, the method appears to work well in practice.

Thus, the objective of this work is to use visual input that is as general as possible, but to show that for certain functional situations, it can lead to much reduced dimensionalities. Thus, here, the very fact that it reflects different configurations of the same physical arm restricts its variation in the image space. Although we have experimented with several other dimensionality reduction techniques, we report here only the results based on ISOMAP. The approach is to first reduce the dimensions of the visual input and then use this compact representation to do path planning in the presence of obstacles. Also, we show that a simple mapping can be learnt between the low dimensional embedding generated by ISOMAP and the motor signals and thus effectively eliminating the problem mentioned above.

## Visuo-Motor embedding

We conducted the following experiments on 100 x 100 images of a robot. For our experiments, we assumed that the robot is free to move all around - that is there are no restrictions on  $\theta_1$  and  $\theta_2$ , i.e.,  $0 \leq \theta_1 \leq 360$ ,  $0 \leq \theta_2 \leq 360$ . The lower dimensionality embedding generated by the Isomap, as shown in figure 3, is the visuo-motor embedding since it captures motor signals using visual data and both can be arrived at using the embedding.

### Mapping Visuo-Motor Embedding to Motor Signals

In this experiment, we illustrate that the ISOMAP not only reduces the number of dimensionality optimally, but this reduction is not arbitrary and geometry of the embedding captures important insights into parameterization of the data -  $\theta_1, \theta_2$  (figure 3). In order to further substantiate this point we show that there exists simple mapping from the embedding to motor signals -  $(\theta_1, \theta_2)$ . We used a 1-hidden layer (10 hidden units) feed forward neural network with linear output to learn the mapping with small errors. Moreover the parameters governing the data -  $\theta_1, \theta_2$ , can be seen as motor signals given to manipulate the robot. Note that, once a path is found in the embedding space, such a mapping is required to realize it in real i.e., generating the actual motor signals to traverse the path. While there exists a simple mapping from the embedding space to the motor signals, it is very hard to

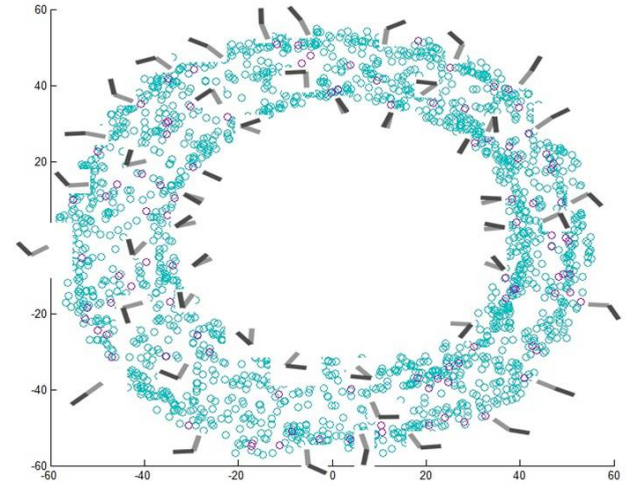


Figure 3: ISOMAP embedding (torus) with points with a sampling of images. The base angle (what we call  $\theta_1$ ) changes along the circumference of the torus and  $\theta_2$  along the radius of the torus.

learn a mapping directly from the sensory input (image space) to motor signals ( $\theta$  space) (see figure 4)

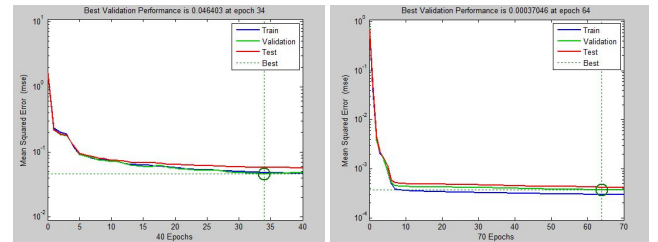


Figure 4: Comparison of the performance of neural network when learnt from high dimensional images (left) and low dimensional embedding (right) to  $\theta_1, \theta_2$  for the two-link manipulator.

### Mapping Obstacles to Visuo-Motor Embedding

Given an obstacle, we mark all the points which are blocked by the obstacle as colliding and others as collision free configurations (this can be checked very easily in the image space) resulting a set of collision free points in which the path can be planned. An important advantage of this method is that it avoids recomputation of the embedding and obstacle mapping for each obstacle separately. The set of collision free points can be easily updated in the event of a moving obstacle (figure 5), new obstacles appearing in the workspace, and old obstacles going away. This also preserves the topology of the embedding.

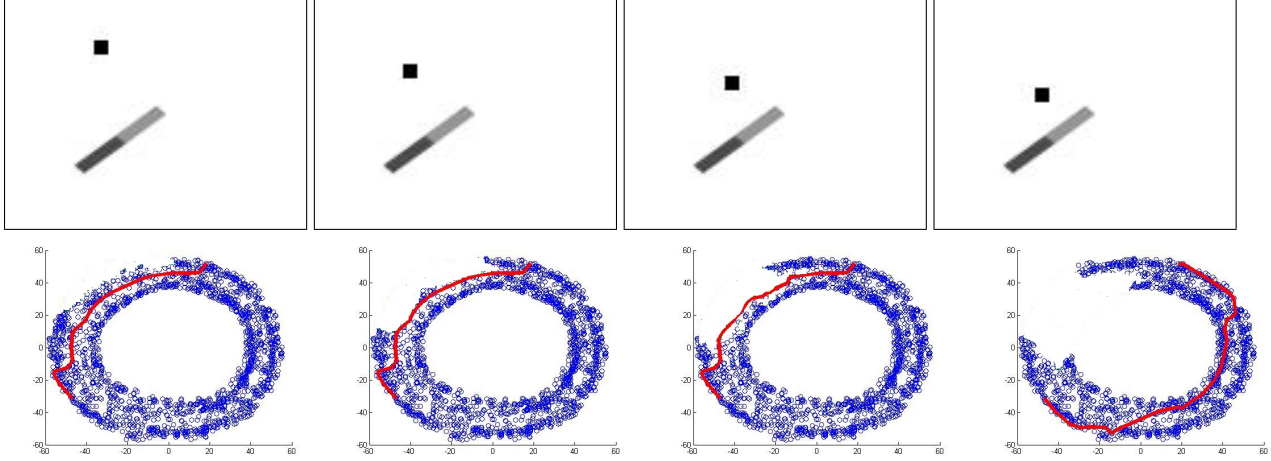


Figure 5: Robot and obstacle configurations and corresponding mapping to embedding space. The set of collision free points can be easily updated for a moving obstacle as shown. Moreover, if a previously free path becomes obstructed due to movement of an obstacle or appearance of a new obstacle, a new path can be easily be planned with the updated set of collision free points. Blue points here are the set of collision free points and green points the set of colliding points. Red color shows the points along the path from the initial configuration to final configuration

### Path Planning in Visuo-Motor Embedding

Once we have the embedding points and a set of collision free points we can plan a path between any pair of start and end points, using any of the graph search algorithms. In case, the start or end point is not in the embedding, a nearby point is found in the embedding and then path planning is done between those points. One of the important observations is that a shortest path query in the graph, results in a path which avoids any redundant movements which is consistent with how a human would do if presented with same planning task. In other words, the shortest path in the embedding space corresponds to least changes to parameters -  $\theta_1, \theta_2$ . (see figure 6) Moreover, as an obstacle changes its position in the workspace, the obstructed points can be removed as visually perceived and a new path can be planned avoiding the obstacle.

### Learning a representation?

We have shown that this process, starting from a set of images and no other information, is able to discover a number of facts. We present these facts, which we call a *naive representation*, presenting in parentheses what a robotics text may call an analog for these.

- a two dimensional manifold obtains then best reduction in residual error. (Robotics: it has two degrees of freedom)
- the structure of the low-dimensional 2-D manifold is that of a torus, so it is cyclic in both dimensions, i.e. it is a  $S^1 \times S^1$  topology. (R: it is an  $S^1 \times S^1$  topology)
- if we associate two variables with these 2 DOFs, we may have one go around the torus in the main direction, and the

other along the thickness, then these capture some aspects of the variation (R:  $\theta_1, \theta_2$ )

- any node in the graph associated with two variables is mapped to the image space, and any image used in the original input is mapped to a node. (R: forward and inverse kinematics)
- the space of these variables is connected with edges that denote nearest neighbours. (R: there is a roadmap)
- given any images for the start and end positions, it is possible to find a path connecting them using these edges. (R: roadmap-based path planning)
- given an obstacle, the system can determine which configurations hit the obstacle in the image space. The corresponding nodes are removed in the manifold graph. (R: the C-space map of an obstacle).
- given an obstacle and a start and goal image, a path can be found via those edges not incident on any of the obstacle nodes. (R: C-space obstacle avoidance).

The above analogies are of course very coarse. Only the topology is really preserved; none of the metric properties are guaranteed to hold; thus the path obtained may not be very short. Nonetheless, considering that no prior knowledge of any kind was used in constructing this motor map, the above naive representation is surely an impressive analogy to present models of robotics.

Note however, that the system, in using these routines, need not be “conscious” that it is using such a representation - it just has to be able to use it effectively.

A final note on the precision of the process. We note that the space need not be sampled uniformly, as was done here. In many situations, certain configurations would be occurring more frequently, and there would higher accuracies would hold in these part of the space (as in the rough diagram of the 50-torus earlier). This is also true of animals and humans, where more frequently executed actions are much more precise.

## Conclusion

In this work, we have demonstrated that although visual input is extremely high-dimensional, the data lies on a nonlinear manifold that is much lower dimensional. The dimensionality of this manifold, and its structure, reveals much about the problem domain and may be called a *naïve representation*. We also propose an approximate mechanism for handling out-of-sample data in the NLDR algorithm - first, by constructing the manifold for the entire workspace and then deleting points, as opposed to trying to construct a new manifold for every workspace. This (figure ) also explains the fact that during the early stages of motor learning, the resolution of the map is poor, and this may also contribute to some aspects of the jerky movements demonstrated in early infancy. As the system populates the visual space with more data points, the resulting surface becomes smoother and permits fluent motions.

Is there cognitive evidence that such a method is actually implemented in the cortex? It is perhaps too early to comment on this. We can only point to a clear role for eye-centered coordinates in the reach planning process, and suggest that such a representation, involving “nearby” images from the image space, may constitute at least a plausible model for part of the computation involved in reach planning.

## References

- Adolph, K., & Berger, S. (2006). Motor development. In *Handbook of child psychology*.
- Batista, A., Buneo, C., Snyder, L., & Andersen, R. (1999). Reach plans in eye-centered coordinates. *Science*, 285(5425), 257.
- Beltrán-González, C. (2005). *Toward predictive robotics: the role of vision and prediction on the development of active systems*. Unpublished doctoral dissertation.
- Bishop, C. (2006). *Pattern recognition and machine learning*. Springer.
- Caeyenberghs, K., Wilson, P., Van Roon, D., Swinnen, S., & Smits-Engelsman, B. (2009). Increasing convergence between imagined and executed movement across development: evidence for the emergence of movement representations. *Developmental science*, 12(3), 474–483.
- Choset, H. (2005). *Principles of robot motion: theory, algorithms, and implementation*. The MIT Press.
- Graziano, M. (2011). Is reaching eye-centered, body-centered, hand-centered, or a combination? *Reviews in the Neurosciences*, 12(2), 175–186.
- Hoffmann, H., Schenck, W., & Moller, R. (2005). Learning visuomotor transformations for gaze-control and grasping. *Biological Cybernetics*, 93(2), 119–130.
- Jordan, M., & Wolpert, D. (1999). Computational motor control. *The cognitive neurosciences*, 601.
- Kuperstein, M. (1991). Infant neural controller for adaptive sensory-motor coordination. *Neural Networks*, 4(2), 131–145.
- Lee, J., & Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer.
- Meer, A. van der. (1997). Keeping the arm in the limelight: advanced visual control of arm movements in neonates. *European Journal of Paediatric Neurology*, 1(4), 103–108.
- Meer, A. Van der, Weel, F. Van der, & Lee, D. (1995). The functional significance of arm movements in neonates. *Science*, 267(5198), 693.
- Meer, v. d. W. F. R. . L. D. N. van der, A. L. H. (1996). Lifting weights in neonates: Developing visual control of reaching. *Scandinavian journal of psychology*, 37(4), 424–436.
- Mushiake, H., Tanatsugu, Y., & Tanji, J. (1997). Neuronal activity in the ventral part of premotor cortex during target-reach movement is modulated by direction of gaze. *Journal of neurophysiology*, 78(1), 567–571.
- Oztop, E., Bradley, N., & Arbib, M. (2004). Infant grasp learning: a computational model. *Experimental Brain Research*, 158(4), 480–503.
- Saxon, J., & Mukerjee, A. (1990). Learning the motion map of a robot arm with neural networks. In *Neural networks, 1990., 1990 ijcnn international joint conference on* (pp. 777–782).
- Tenenbaum, J. B., Silva, V. de, & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2319.
- Thelen, E. (2000). Motor development as foundation and future of developmental psychology. *International Journal of Behavioral Development*, 24(4), 385–397.
- Thelen, E., Smith, L., Lewkowicz, D., & Lickliter, R. (1994). *A dynamic systems approach to the development of cognition and action*. MIT Press.
- Von Hofsten, C. (2004). An action perspective on motor development. *Trends in cognitive sciences*, 8(6), 266–272.

# Establishing a Database for Studying Human Face Photograph Memory

**Wilma Alice Bainbridge\* (wilma@mit.edu)**

Department of Brain and Cognitive Sciences, MIT, 77 Massachusetts Avenue  
Cambridge, MA 02139 USA

**Phillip Isola\* (phillipi@mit.edu)**

Department of Brain and Cognitive Sciences, MIT, 77 Massachusetts Avenue  
Cambridge, MA 02139 USA

**Idan Blank (iblack@mit.edu)**

Department of Brain and Cognitive Sciences, MIT, 77 Massachusetts Avenue  
Cambridge, MA 02139 USA

**Aude Oliva (oliva@csail.mit.edu)**

Computer Science and Artificial Intelligence Lab (CSAIL), MIT, 77 Massachusetts Avenue  
Cambridge, MA 02139 USA

## Abstract

Contemporary visual environments bombard us with hundreds of face images every day, and this places a non-trivial demand on long-term memory. However, little is known about what makes certain faces remain in our memories, while others are quickly forgotten. To establish a basis for face memorability exploration, we assembled a database of 8,690 face photographs from online sources, spanning diverse face and image characteristics. Workers on Amazon's Mechanical Turk were asked to identify repetitions within a stream of these stimuli. Variations in image memorability (hit rates, false alarm rates, and their interactions) were reliable across participants, suggesting that face images may have different intrinsic levels of memorability. We discuss future directions in using this database to quantify face photograph memorability, as well as potential scientific and commercial applications.

**Keywords:** face recognition; image memorability; face photograph memory database

## Introduction

Every day, we encounter an overwhelming number of photographs and images of people's faces. Many interpersonal interactions are mediated by such images: we view people's Facebook profile pictures; memorize photographs of our students; browse personals on dating websites; skim through pictures attached to job applications; and encounter countless face images published on advertisements on billboards, in magazines, and online. As social creatures, we remember many of these faces.

Large-scale visual memory experiments have shown that people have a remarkable ability to remember which specific image they saw even after seeing thousands of pictures depicting objects, scenes or events (Konkle, Brady, Alvarez, & Oliva, 2010a; Standing, 1973). Importantly, these studies have shown that we do not just remember the gist of a picture, but we are able to recognize which precise image we saw and some of its visual details (Brady, Konkle, Alvarez, & Oliva, 2008; Konkle, Brady, Alvarez, & Oliva, 2010b). In addition to remembering particular images as icons, we also have the intuition that not all images are remembered equally. While the reasons why some images are remembered are varied, recent works have found that

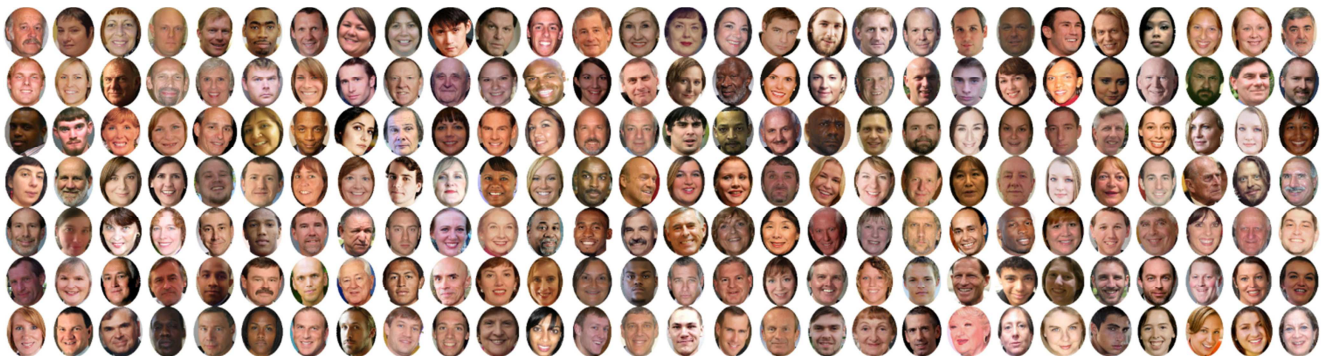


Figure 1: An example set of 196 random images from the face photo database used for this study.





Figure 2: An illustration of the behavioral procedure. Participants were required to identify repeats amongst a stream of face photos.

images containing people with visible faces are highly memorable (Isola, Parikh, Torralba, & Oliva, 2011a; Isola, Xiao, Torralba, & Oliva, 2011b).

Despite the fact that the memorability of face photos is of both psychological and commercial significance, it is not clear how findings illuminating scene and object memorability will generalize to face images. First, memorability has been shown to be heavily influenced by the distinctiveness of stimuli (Konkle et al., 2010a, 2010b). Compared to scenes and objects, faces are a relatively homogeneous category and have low variation in visual features. However, faces could be coded with rich sub-categorical structure (e.g., gender, race, age, emotional expression, dominance level, attractiveness) that may render their representations more distinguishable in memory. Second, evidence suggest that faces are processed by specialized cognitive (Duchaine, Yovel, Butterworth, & Nakayama, 2006; Robbins & McKone, 2007) and neural (Kanwisher & Yovel, 2006) mechanisms (c.f., McKone, Crookes, & Kanwisher, 2009). For these reasons, face memorability deserves special attention.

In this study, we establish a large-scale face photograph database on which we have quantified performance on a repetition detection task. We examined inter-image variability, and its reliability, on this task. Specifically, we analyzed two memory-related behavioral measures – hit rate and false alarm rate – which we term “memorability scores”.

## Methodology

We conducted a large-scale experiment that used photos from a database of diverse faces, run on 337 participants on Amazon's Mechanical Turk. The following section describes the assembly of the database and the experiment run on Mechanical Turk.

### Face Photo Database Generation

We assembled a diverse database of 10,000 photos of faces. First, we generated a list of approximately 25,000 first and last name pairs from a database of names from the United States census (Kleimo, 2011), using parameters for a balance of both genders and names of high commonality. Use of the US census allowed us to collect names from a diverse range of ethnic backgrounds, representing the general gender, racial, and age distribution of the United States adult population. However, because the first and last

names were generated randomly, they did not necessarily represent specific people from the US population. Example names included “Wilma Reno,” “Phillip Robichaux,” “Lori Blank,” and “Arlene Olivarez”.

Each of the 25,000 names was used as a search query, and, for each query, approximately 10 photos were automatically downloaded from Google Images. Our Google Image Search parameters included that all photos be at least 400×300 pixels, full-color, and of faces. The experimenters went through the set of photos and deleted those that were low-quality, depicted children, were obscured by other objects, included accessories such as hats and glasses, or had unusual makeup. The database was filtered down to over 10,000 photos of faces that were diverse over a wide range of ages, genders, races, and attractiveness levels. Faces had both eyes visible and open and, in general, expressions tended to range from neutral to smiling. Five experimenters then went through the set and deleted recognizable celebrities for the purposes of this study, bringing the set used for this experiment to a final size of 8,690 photos. We expect that only a small percent of our database should be celebrity photos that were not identified through our initial screening. The stimuli for the experiment were then generated by placing ovals around the faces to frame them and to diminish the influence of irrelevant background features in the photo. All photos were resized to a standard of 256 pixels in height with variable width to preserve aspect ratio. Figure 1 shows a collection of example photos from the database.

### The Behavioral Experiment

Face memory performance was measured through a behavioral study called the “Face Memory Game” run on Amazon's Mechanical Turk. Mechanical Turk is a tool belonging to Amazon.com's Web Services that allows researchers to crowdsource tasks and experiments for monetary compensation to a large Internet population. Mechanical Turk served as an ideal environment for this study, allowing us to obtain memory scores for thousands of images.

The methodology for this game is based off the methodology from a previous image memorability study conducted with scenes (Isola et al., 2011b; see Figure 2). The task was structured into a series of 30 levels, each taking about 4.8 minutes and consisting of 120 photos. Although labeled “levels” to give a sense of progress to the participant, the levels did not differ from each other in

difficulty or stimulus type. For each level, the participant saw a constant stream of stimuli, each displayed for 1 second and then followed by a 1.4 second fixation point before the next stimulus was presented. Stimulus presentation order was different for each participant. Participants had to press the key ‘r’ (for “repeat”) whenever the current stimulus was the same photo as one they had seen before (sometimes across levels). When they responded correctly to a repeat, a green cross appeared as feedback. When participants missed a repeated photo or pressed ‘r’ for a novel photo, a gray X appeared to indicate an error. The game was first preceded by a short qualification and training round of 30 photos. Between levels, participants were given a brief break of up to five minutes and were presented with their correct response score for that level. After 30 levels of the game were over, the game ended. However, participants could choose to end the game at any time, and their data was used up to that point.

From the face stimulus database, 2222 photos were randomly selected as target photos, while the remaining 6468 photos were used as filler photos. Repetitions of photos in the task happened with both target and filler photos. The memory performance measures are based off the results from the target photos, where repetitions were spaced 91-109 photos apart. The repetition with the filler photos acted as a “vigilance task” to test the reliability of participants, with repetitions spaced 1-7 photos apart. The filler photos were also used as spacing between the target photos, and some had no repetitions. Neither target photos nor filler photos had more than one repetition across the entire study.

A total of 337 Mechanical Turk workers participated in the game, and 90% of the data came from 168 workers. The average worker played over 8 levels. We limited the game to only Mechanical Turk workers in the US, so that the workers’ demographics would approximately match the demographics of the faces used as stimuli. Workers were paid \$0.40 per level, or approximately \$5 an hour. Workers were screened in several ways throughout the study to ensure they were attentive to the task. First, only workers with at least a 95% Mechanical Turk approval rate were allowed to participate in the study. During the study, if a participant’s error rate for false alarms exceeded 50% for the last 30 photos, or if their hit rate for vigilance task repeats fell below 50% for the last 10 photos, then the data from that level were discarded and the participant received a flag. Rejection criteria were reset for each level. If the participant received three flags, they were blocked from continuing in the experiment. Otherwise, participants could restart the game as many times as they liked, until they had completed 30 levels. When restarting the game, unseen photos were always selected as stimuli.

## Results

We collected an average of 30.4 hit rate (HR) scores per photo and 35.4 false alarm rate (FAR) scores per photo. The

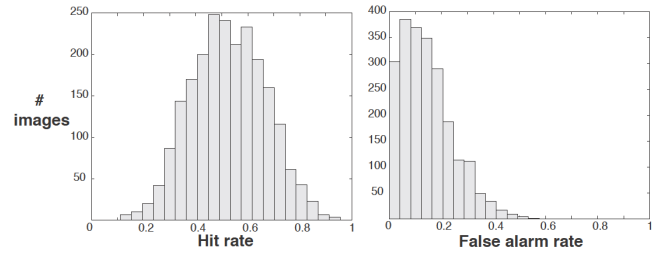


Figure 3: Hit rate and false alarm rate histograms over the target photos in our experiment.

average HR was 53.6% ( $SD=14.3\%$ ), and the average FAR was 14.5% ( $SD=9.9\%$ ). The distributions of these memorability scores followed simple unimodal forms (Figure 3).

## Is Memory Performance on Some Images Reliably Different than on Other Images?

To evaluate the reliability of our measurements, we split our participant pool into two independent halves, and quantified how well memorability scores measured on the first half of the participants matched memorability scores measured on the second half of the participants. Averaging over 25 random split-half trials, we calculated a Spearman's rank correlation  $\rho$  of 0.44 between HRs on the two halves and a  $\rho$  of 0.48 on FARs. The strength of these correlations demonstrates that we have characterized real differences between photos.

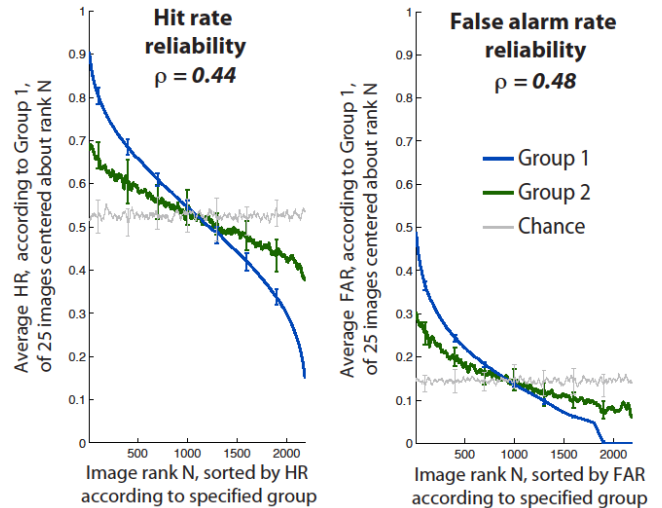


Figure 4: Data split-half reliability. Photos are ordered on the x-axis by the HR (left) and FAR (right) of a random half of the participants, and are plotted against these measures on the same half (blue line) or the remaining half (green line) of participants. Chance reliability is shown by randomly ordering the photos on the x-axis (gray line). Plots are averaged across 25 such random splits of the participant pool.

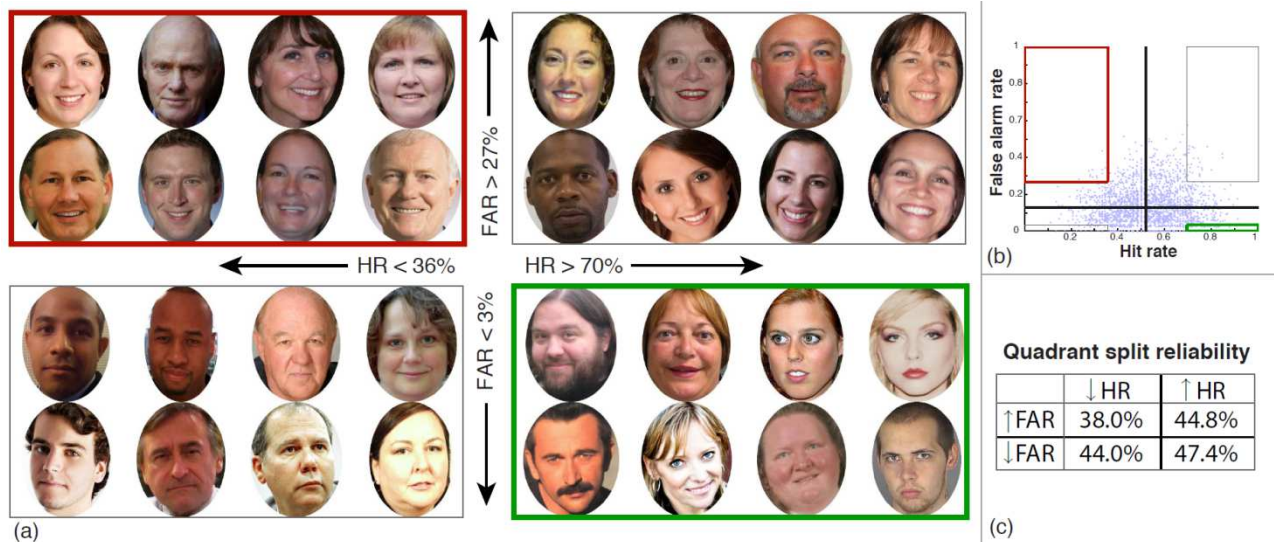


Figure 5: (a) Sample images of four performance profiles. The image set was broken into a  $5 \times 5$  grid of HR quintiles crossed with FAR quintiles. Each quadrant shows a random sample of the photos at each of the four corners of this distribution (highest/lowest HR/FAR). The set outlined in green can be characterized as more memorable than the set outlined in red since the green set has both a higher HR and a lower FAR than the red set. (b) A scatterplot showing HR versus FAR. Rectangles indicate the same corners of the quintile grid as in (a). The black lines split the distribution along the median HR and FAR creating four performance profiles. (c) Reliability computed as percent overlap of HR/FAR profile assignments of photos between two halves of the participants (averaged across 25 random splits of the participant pool). Profiles correspond to the quadrants defined by the black lines in (b).

Are these differences large enough to be interesting? We examined the reliability of the size of the memorability differences as follows. We sorted photos by their scores given by the first half of the participants and plotted this against memorability scores according to the second half of the participants (Figure 4). For clarity, we convolve the resulting function with a length-25 box filter. This shows that, for example, if a repeat is correctly detected 80% of the time by one half of the participants, we can expect the other half of the participants to correctly detect this repeat around 66% of the time, corroborating that this photo is truly memorable. At the other end of the spectrum, if a repeat is only detected 30% of the time by one half of the participants, the other half will tend to detect it only 42% of the time – this photo is consistently forgotten. It thus appears that there really is sizable variation in face photo memorability.

Thus, our data show enough variation and enough reliability that it should be possible to use these data to model detailed aspects of photo memorability in later work (c.f., Isola et al., 2011a, 2011b). Individual differences and random variability in the sequence of photos each participant viewed add noise to these estimations; nonetheless, this level of reliability suggests that information intrinsic to the photos plays a key role in determining which photos are remembered.

## False Memories versus True Memories

Our data allow us to look at both false memories and true memories. False memories may arise in response to highly typical faces, because they resemble many other faces (Vokey & Read, 1992). True memories should relate to specific encodings of the photos seen in our experiment. Can we separate these two signals in our data? If a photo receives both a high hit rate and a high false alarm rate, it may be highly memorable, but it also may just be a face that always feels familiar, regardless of whether or not it has been previously seen. A stronger case for high memorability can be made when we find photos that have high hit rates and low false alarm rates – what is termed a "mirror effect" (Glanzer & Adams, 1985, 1990). If one photo consistently has both a higher hit rate and a lower false alarm rate than another photo, then we can confidently say that the first photo evoked a stronger true memory than the second.

To isolate truly memorable photos, we split our photo set about the median HR and then again about the median FAR, producing four performance profiles (high/low HR/FAR) (see Figure 5). Are some photos consistently assigned to the high-HR/low-FAR profile, whereas others are consistently assigned to the low-HR/high-FAR profile? If so, we can say the former photos are more memorable than the latter. We tested this level of consistency by splitting our photos into profiles according to one random half of the participants and comparing these assignments to those given by the other half of the participants. Averaging over 25 such trials, the



two halves of the subjects agree 47% of the time on assignments to the high-HR/low-FAR profile (chance level would be 25%). Interestingly, we see similar levels of agreement in each of the remaining quadrants, as reported in Figure 5c.

These quadrants may reflect different types of photos with respect to memory: some photos may be distinctive and strongly remembered; some may be prototypical and produce both strong memories and many false alarms; others may evoke many false memories while, interestingly, generating relatively few true memories (low-HR/high-FAR); and still others may simply be ignored all together (low-HR/low-FAR).

## Discussion

This study has established a database for the exploration of face photograph memory, and shows that memorability of face photographs can be reliably measured. We found an average hit rate of 53.6% across the target face photos, compared to a false alarm rate of only 14.5%. In contrast, Isola et al. (2011b) used the same experimental protocol and found an average hit rate of 67.5% and false alarm rate of 10.7% for scene photo memory capacity. Do these numbers for face photos indicate that we are worse at remembering faces than scenes? Or, is the face photo performance high, considering that faces vary at the exemplar level (i.e., all belong to the same basic-level category), while the scenes used by Isola, et al. (2011b) vary at the categorical level (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976)? It is difficult to compare across separate studies and participant pools – for example, Isola et al. (2011b) recruited international participants, while the current study limited participants to the United States. It will also be essential to find a way to quantify the differences between face and scene photos in order to meaningfully compare memorability between the two different groups of stimuli.

A second interesting question to explore is what attributes lead to the separation of photos into the four performance profiles we identified based on hit rate and false alarm rate (Figure 5). Previous research has suggested that more distinct faces have high hit rates and low false-alarm rates in an old/new task (Deffenbacher, Johanson, Vetter, & O'Toole, 2000; Light, Kayra-Stuart, & Hollander, 1979). In contrast, both hit and false-alarm rates are high for typical faces, due to the effect of "context-free familiarity", a sense of familiarity not related to a specific previous encounter with a face (Vokey & Read, 1992). The other two profiles we explored may also have interesting qualifying characteristics to examine that were not explicitly addressed in the past literature.

Beyond distinctiveness and typicality, we advocate the exploration of several other attributes and their correlations with memorability. Previous research has noted that memorability of a face, both perceived and actual, may differ based on viewer characteristics, such as race (Chiroro & Valentine, 1995; Meissner, Brigham, & Butz, 2005) or recent experience with other face images (Lewis &

Johnston, 1997); however, the current study shows surprising reliability across subjects of diverse backgrounds, viewing a widespread distribution of photos. This suggests there are similarities across participants in how they represent different photos in memory. One next important step will be to examine how the demographic characteristics of the participant (e.g., race, gender, and age) may or may not predict the memorability of face photos with matching or non-matching characteristics. Other properties to examine in the context of memorability include perceived memorability (do people actually remember what they think they will remember?), attractiveness, and eye contact. While the current work focuses on memory for photos of faces, future work will also explore memory for face identity across different photos of the same person.

The future possibility of quantifying "memorability" of a face lends itself to many useful applications in both the field of psychology and mainstream society. For instance, Todorov (2011) identified features in faces linked to different subjective judgments of those faces, such as attractiveness and trustworthiness. These were used to build computer models that generated faces varying along these featural dimensions. A score of memorability could similarly be added to the feature set of a face, and thus be used to rate, manipulate, and generate face images. For animated films, animators could create cartoon characters with different levels of memorability (c.f., Gooch, Reinhard, & Gooch, 2004), such as a highly memorable protagonist surrounded by forgettable extras. Makeup artists could use software that would identify where to apply makeup to make celebrities memorable for a photoshoot. Algorithms could automatically identify the most memorable face photographs out of an album to use in textbooks, magazines, or even social network profiles.

## Conclusion

This study serves as an initial, empirical look at a new large, diverse database of face photos and the average rate and reliability of memorability measurements across this database. When viewing a stream of hundreds, sometimes thousands, of novel face photos, participants in our experiment were able to accurately identify repeats about half the time they appeared, while making relatively few false alarms. This suggests that participants were holding in memory detailed representations of hundreds of face photos even though each photo was presented with just a single one-second view. In addition, we found that photos of faces vary substantially in memorability; these reliable differences indicate the importance of memorability for understanding how we process face images. This research opens the door to future investigation in various fields, from cognitive psychology to cognitive neuroscience to computer vision, as to what makes some face images or facial features more memorable than others.

## Acknowledgments

We would like to thank Marc Howard for helpful discussions and advice. This work is partly funded by a NSF grant (1016862) and a Google research award to A.O. W.A.B. is funded by the Leventhal Graduate Fellowship, P.I is funded by an NSF graduate research fellowship, and I.A.B. is funded by the Henry E. Singleton Fund.

The face photograph database will be publicly available on the author's website.

## References

- Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38), 14325-14329.
- Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology*, 48(4), 879-894.
- Deffenbacher, K. A., Johanson, J., Vetter, T., & O'Toole, A. J. (2000). The face typicality-recognizability relationship: encoding or retrieval locus? *Memory & Cognition*, 28(7), 1173-1182.
- Duchaine, B. C., Yovel, G., Butterworth, E. J., & Nakayama, K. (2006). Prosopagnosia as an impairment to face-specific mechanisms: elimination of the alternative hypotheses in a developmental case. *Cognitive Neuropsychology*, 23(5), 714-747.
- Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, 13(1), 8-20.
- Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(1), 5-16.
- Gooch, B., Reinhard, E., & Gooch, A. (2004). Human facial illustrations: creation and psychophysical evaluation. *ACM Transactions on Graphics (TOG)*, 23(1), 27-44.
- Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011a). *Understanding the intrinsic memorability of images*. Paper presented at the 25th Conference on Neural Information Processing Systems (NIPS), Granada, Spain.
- Isola, P., Xiao, J. X., Torralba, A., & Oliva, A. (2011b). What makes an image memorable? *24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 145-152.
- Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1476), 2109-2128.
- Kleimo, A. (2011). The Random Name Generator, from <http://www.kleimo.com/random/name.cfm>
- Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010a). Scene Memory Is More Detailed Than You Think: The Role of Categories in Visual Long-Term Memory. *Psychological Science*, 21(11), 1551-1556.
- Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010b). Conceptual Distinctiveness Supports Detailed Visual Long-Term Memory for Real-World Objects. *Journal of Experimental Psychology-General*, 139(3), 558-578.
- Lewis, M. B., & Johnston, R. A. (1997). Familiarity, target set and false positives in face recognition. *European Journal of Cognitive Psychology*, 9(4), 437-459.
- Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, 5(3), 212-228.
- McKone, E., Crookes, K., & Kanwisher, N. (2009). The cognitive and neural development of face recognition in humans. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences IV*. Cambridge, MA: MIT Press.
- Meissner, C. A., Brigham, J. C., & Butz, D. A. (2005). Memory for own- and other-race faces: a dual-process approach. *Applied Cognitive Psychology*, 19(5), 545-567.
- Robbins, R., & McKone, E. (2007). No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition*, 103(1), 34-79.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3), 382-439.
- Standing, L. (1973). Learning 10,000 Pictures. *Quarterly Journal of Experimental Psychology*, 25, 207-222.
- Todorov, A. (2011). Evaluating faces on social dimensions. In A. Todorov, S. T. Fiske & D. A. Prentice (Eds.), *Social Neuroscience: Toward Understanding the Underpinnings of the Social Mind*. New York, NY: Oxford University Press.
- Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition*, 20(3), 291-302.

# The Representation and Processing of Tense, Aspect & Voice across Verbal Elements in English

Jerry T. Ball (Jerry.Ball@wpafb.af.mil)

711<sup>th</sup> Human Performance Wing, Air Force Research Laboratory,  
Wright-Patterson Air Force Base, OH 45431

## Abstract

We consider the representation and processing of the English verb features tense, aspect and voice, within a computational cognitive model of human language processing. We assume that a collection of features is associated with each verbal element and that these features may project to the clauses in which they occur. When multiple verbal elements occur, it is possible for the features to conflict, necessitating mechanisms of feature blocking and overriding to determine feature projection. The alternative of having multiple entries in the mental lexicon for each verbal element with different feature sets is avoided due to the ambiguity that would be introduced, and the weak grammatical motivation for doing so. However, we do assume an ambiguity in the case of most v-ed and v-base verb forms, with the past tense v-ed form being distinct from the past participle v-ed form and the present tense v-base form being distinct from the non-finite v-base form. We assume that every finite clause expresses a tense and voice feature and many finite clauses express an aspect feature as well. We consider the case of transitive and intransitive verbs in combination with the auxiliary verbs “be” and “have” in finite clauses. For intransitive verbs, we introduce an active/inactive voice feature distinction which aligns with the transitive distinction between active and passive voice.

**Keywords:** grammatical feature, tense, aspect, voice

## Introduction

We consider the representation and processing of the English verb features tense, aspect and voice, within the context of a *pseudo-deterministic* model of human language processing (Ball, 2011a) implemented in the ACT-R cognitive architecture (Anderson, 2007). The pseudo-deterministic model reflects the integration of a highly parallel, probabilistic, and context dependent, activation and selection mechanism and non-monotonic *context accommodation* mechanism (with limited parallelism) with what is otherwise an incremental processor which pursues the best analysis. The overall effect is a human language processor (HLP) which presents the appearance and efficiency of deterministic processing, despite the rampant ambiguity which makes truly deterministic processing impossible. Our non-monotonic context accommodation mechanism replaces the monotonic *look-ahead* mechanism of Marcus’s deterministic parser (Marcus, 1980) and is argued to be more cognitively plausible (Ball, 2011a).

We assume that a collection of verb features is associated with each verbal element (cf. Gazdar et al., 1985) and that

these features may project to the clauses in which they occur. We consider the composition of verb features across verbal elements within a clause. When multiple verbal elements occur, it is possible for the verb features to conflict. The context accommodation mechanism, which has been independently motivated (Ball, 2010a), is crucial for handling conflicts. In particular, we propose specialized mechanisms of *feature blocking* (i.e. a feature of a preceding verbal element precludes projection of a conflicting feature of a subsequent verbal element) and *feature overriding* (i.e. a feature of a subsequent verbal element overrides a conflicting feature of a preceding verbal element) to handle conflicts. Feature overriding is non-monotonic in that it changes the incrementally evolving representation.

Our non-monotonic approach can be contrasted with approaches which rely on monotonic unification of non-conflicting features (Gazdar et al., 1985; Sag et al., 1986; Sag, Wasow & Bender, 2003). To avoid feature conflicts, such approaches tend to posit alternative entries in the mental lexicon which are structurally ambiguous, often linguistically unmotivated and sometimes grammatically inadequate. For example, “a few books” is grammatical in English despite the fact that “a” is singular and “few” and “books” are plural. In a monotonic unification-based approach, the number feature of “a” must somehow unify with the number feature of “few” and “books”. To handle this, one could posit a plural or number lacking version of “a”. But this introduces ambiguity and lacks linguistic motivation. In our non-monotonic approach, the plural feature of “few” and “books” is allowed to override the singular feature of “a” (Ball, 2010b). Feature blocking and overriding are concerned with the composition of features across lexical items within constructions and differ from non-monotonic *default constraint inheritance* (cf. Sag, Wasow & Bender, 2003, 229ff.) which is concerned with defeasible inheritance of features within individual lexical items—which we also use (Ball, 2011b).

English has a highly restricted number of distinct verb forms which include the following:

- V-base (or V-plain) form (e.g. “give”, “go”)
- V-s form (e.g. “gives”, “goes”)
- V-ed form (e.g. “gave”, “went”, “kicked”)
- V-en form (e.g. “given”, “gone”)
- V-ing form (e.g. “giving”, “going”)

“Goes” is a slightly irregular v-s form, “gave” and “went” are irregular v-ed forms, and “gone” is an irregular v-en form. We also treat the combination of the infinitive marker “to” and the base verb form as a distinct verb form, abbreviated as to+v-base (e.g., “to give”). Having a distinct infinitive form allows the model to unambiguously recognize infinitives as multi-word units and reduces overall ambiguity. In total, we claim the existence of six distinct verb forms. By comparison, Quirk et al. (1985, p. 96) claim only five regular verb forms

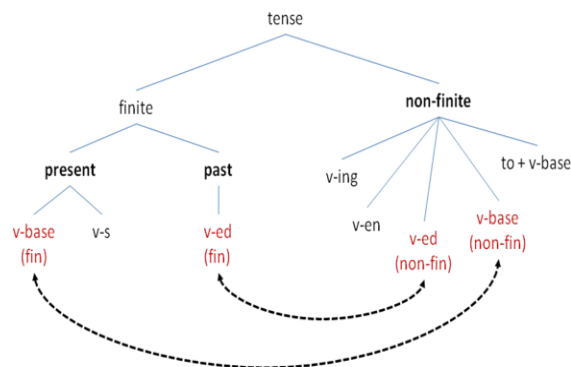
- Base form (v-base)
- -s form (v-s)
- -ing participle (v-ing)
- Past form (v-ed)
- -ed participle (v-ed or v-en)

not recognizing to+v-base as a distinct form, treating v-en as an irregular -ed participle, and calling v-ed the past form (distinct from the -ed participle). Huddleston & Pullum (2002, p. 74) recognize six verb forms, three primary forms and three secondary forms:

- Primary
  - preterite (v-ed)
  - 3<sup>rd</sup> singular present tense (v-s)
  - plain present tense (v-base)
- Secondary
  - plain form (v-base)
  - gerund-participle (v-ing)
  - past participle (v-ed or v-en)

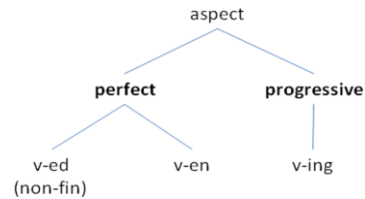
We follow Quirk et al. and Huddleston & Pullum in allowing the v-ed form to map to two distinct tenses: *past* tense and *non-finite* or untensed. We follow Huddleston & Pullum in allowing the v-base form to map to two different tenses: *present* tense and *non-finite*. Quirk et al. and Huddleston & Pullum treat the v-ed (*non-finite*) and v-en forms as alternative forms of the past participle. We keep them distinct since the v-en form is unambiguous. Huddleston & Pullum, like Quirk et al., do not recognize to+v-base as a distinct form.

In terms of the mapping from different verb forms to the tense feature, we propose the following ontology:



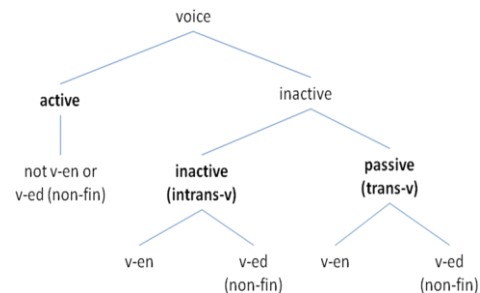
There are eight tense differentiated verb categories corresponding to the six different verb forms. The v-base (*present* tense and *non-finite*) and v-ed (*past* tense and *non-finite*) forms are ambiguous with respect to tense.

For aspect, we propose the following ontology:



We categorize *perfect* as a type of aspect in agreement with Quirk et al. (1985), but contrary to Huddleston & Pullum (2002) who treat *perfect* as a type of tense. Grammatically, there is a clear contrast in form between *progressive* and *perfect* aspect in English with the v-ing verb form corresponding to the *progressive* and the v-en or v-ed (*non-finite*) verb form corresponding to the *perfect*. *Perfect* aspect encodes the completion of an action in contrast to *progressive* aspect which encodes its continuation. However, *perfect* aspect is also closely associated with *past* tense since completed actions typically occur in the past, although the completion may be co-intensive with the present.

For voice, we propose the following ontology:



We assume that voice is a grammatical feature of intransitive as well as transitive verbs. *Active* voice indicates that the subject is actively involved in the action of the verb. *Passive* voice indicates that the subject of the transitive verb corresponds to one of the affected objects (object, indirect object) of the active equivalent. *Inactive* voice indicates that the subject is an inactive participant of an intransitive verb.

Combining features across the six forms and distinguishing transitive and intransitive verbs, the following feature combinations exist:

- V-base (fin): *present* tense, *active* voice
- V-s (fin): *present* tense, *active* voice
- V-ed (fin): *past* tense, *active* voice
- V-base (non-fin): *non-finite*, *active* voice
- To+v-base (non-fin): *non-finite*, *active* voice
- V-ed (non-fin, trans-verb): *non-finite*, *perfect* aspect, *passive* voice
- V-ed (non-fin, intrans-verb): *non-finite*, *perfect* aspect, *inactive* voice

- V-en (non-fin, trans-verb): *non-finite, perfect* aspect, *passive* voice
- V-en (non-fin, intrans-verb): *non-finite, perfect* aspect, *inactive* voice
- V-ing (non-fin): *non-finite, progressive* aspect, *active* voice

Any verbal entry in the mental lexicon will contain the features associated with one of these combinations. All forms of the auxiliary verb “be” encode *inactive* voice. All forms of the auxiliary verb “have” encode *active* voice. When used as a transitive verb, “have” follows the transitive verb pattern. Modal auxiliaries (e.g. “He *can* go”) encode a modal feature in addition to *present* tense and *active* voice. V-base (*present* tense, *active* voice) is the combination associated with imperative (e.g. “*give* me it”) and subjunctive uses (e.g. “I desire that he *give* me it”).

We consider the combining of tense, aspect and voice across the verbal elements in a clause, restricting the discussion to main verbs and the auxiliary verbs “be” and “have” in finite clauses. A key assumption is that the features of verbal elements may conflict, necessitating mechanisms for feature *blocking* and *overriding*, and prohibiting unification as the primary integration mechanism (i.e., conflicting features cannot unify).

With respect to feature blocking, we assume that the grammatical features of the first of two immediately adjacent verbal elements normally suppress expression of conflicting grammatical features of the second. A preceding verbal element expressing *active* voice is incompatible with an immediately following verbal element expressing *passive* or *inactive* voice. For example, in “he has kicked the ball” expression of *active* voice by “has” suppresses expression of *passive* voice by “kicked” (i.e. “he” is actively involved in kicking), and in “he has gone” expression of *active* voice by “has” suppresses expression of *inactive* voice by “gone” (i.e. “he” is actively involved in going). However, the combination of “have” with “been” is special in that the *inactive* voice of “been” overrides the *active* voice of “have”. In addition, the *inactive* voice of “been” is compatible with either the *inactive* voice of a main intransitive verb or the *passive* voice of a main transitive verb. For example, in “the ball has been kicked”, the *passive* voice of “kicked” can project to the clause since the *inactive* voice of “been” overrides the *active* voice of “has”, and the *passive* voice of “kicked” is compatible with the *inactive* voice of “been”. Likewise in “he has been gone”, the *inactive* voice of “gone” can project to the clause (i.e. “he” is not actively involved in going). Feature blocking and overriding are the most distinctive elements of the approach presented in this paper. Both are incompatible with monotonic unification of features.

## Feature Projection for Transitive Verbs

We start by considering the encoding and projection of features in clauses containing the transitive verb “give” as

the main verb. First, we consider clauses with a single main verb, starting with *present* and *past* tense “give”.

1. He gives (**pres+act**) me the ball
2. He gave (**past+act**) me the ball

In 1, “gives” encodes and projects the *present* tense and the *active* voice features. In 2, “gave” encodes and projects the *past* tense and *active* voice features.

If we add the auxiliary verb “be” to “give”, things start to get more interesting:

3. He is (**pres+inaet**) giving (**prog+act**) me the ball
4. He was (**past+inaet**) giving (**prog+act**) me the ball

In 3, “is” encodes and projects *present* tense and *inactive* voice and “giving” encodes and projects *active* voice—overriding the *inactive* voice of “is”—and *progressive* aspect. The overriding of the *inactive* voice of “is” by the *active* voice of “giving” is an exception to the rule that the competing features of the preceding verbal element block projection of the features of the following verbal element (specific to “be”+verb). Example 4 only differs in that “was” encodes and projects *past* tense.

5. He is (**pres+inaet**) given (**perf+pass**) the ball
6. He was (**past+inaet**) given (**perf+pass**) the ball

In 5, “is” encodes and projects *present* tense and *inactive* voice, allowing “given” to project *passive* voice to the clause. Allowing “be” to encode *inactive* voice which can be overridden by an immediately following verbal element, allows “giving” to project *active* voice and “given” to project *passive* voice. In addition to projecting *passive* voice, “given” also encodes and projects *perfect* aspect. There is a clear sense in which “He is given the ball” implies completion of the act of giving, which comes from the *perfect* aspect of “given”. Example 6 only differs in that “was” encodes and projects *past* tense. One might think that completion of the act of giving comes from “was” and not “given” in this example. However, note that “he was giving me the ball” does not imply completion even though “was” is *past* tense.

7. He is (**pres+inaet**) to give (**nonfin+act**) me the ball
8. He was (**past+inaet**) to give (**nonfin+act**) me the ball

In 7, “is” encodes and projects *present* tense and *inactive* voice and “to give” encodes *non-finite* tense (i.e. the absence of tense) and *active* voice, but only projects *active* voice since *present* tense is projected by “is” and blocks the *non-finite* feature of “to give”. The overall effect is that the clause is *present* tense and *active* voice similar to “he gives me the ball”. However, “He is to give me the ball” also implies a modal obligation which is not captured by the current analysis. Example 8 only differs in that “was” is *past* tense.

Adding the auxiliary verb “have” to “give” also has interesting effects.

9. He has (**pres+act**) given (**perf+pass**) me the ball  
 10. He had (**past+act**) given (**perf+pass**) me the ball

In 9, “has” encodes and projects *present* tense and *active* voice. Projection of *active* voice by “has” blocks the possibility of “given” projecting *passive* voice. This leaves only *perfect* aspect to project from “given”. In contrast with the more usual treatment in which “have” combines with a v-en or v-ed (non-finite) form verb to project *perfect* aspect, we propose that “have” instead has the effect of suppressing projection of *passive* voice from the immediately following v-en or v-ed (non-finite) verb form, by projecting *active* voice. Example 10 only differs in that “had” projects *past* tense. If there were separate entries for the passive and perfect variants of “given”, then “have” could bias selection of the perfect entry, whereas “be” could bias selection of the passive entry and this ambiguity is manageable. However, with separate entries, it would not be possible to project both *perfect* aspect and *passive* voice from a single verbal element. Under our current approach, “have” suppresses *passive* voice, but allows *perfect* aspect to project, whereas “be” allows both *passive* voice and *perfect* aspect to project.

11. He has (**pres+act**) to give (**nonfin+act**) me the ball

In 11, “has” encodes and projects *present* tense and *active* voice. “To give” also encodes and can express *active* voice, but this is redundant (but not incompatible) with “has”. Like “be”, “have” combines with an infinitive to express a modal obligation to complete the act. In both cases, this effect appears to derive from the construction (e.g. “is” + “to give”, “have” + “to give”) rather than the individual lexical items. Constructional effects can become encoded in complex lexical items and it is likely that “have to” is encoded in the mental lexicon as a multi-word unit (in spoken language as the reduced form “hafta”) and expresses an obligation as part of its idiomatic meaning as shown in example 12.

12. He has to (**pres+act+must**) give (**nonfin+act**) me it  
 13. He had to (**past+act+must**) give (**nonfin+act**) me it

Example 13 with *past* tense “had” expresses a past obligation rather than a present obligation.

The combination of *perfect* aspect and *passive* voice may also be realized across verbal elements. Consider

14. He has (**pres+act**) been (**perf+inact**) given (**perf+pass**) the ball

As an exception, the *inactive* voice of “been” overrides the *active* voice of “has” allowing the *passive* voice of “given” to project. Note that both “been” and “given” encode and may express *perfect* aspect. At the clausal level, we have *perfect* aspect whether it comes from one or more verbal elements.

As the preceding example shows, it is possible to combine verb features across verbal elements in ways that are not allowed within a single verb (e.g. present perfect), although one would like to assume that conflicting features cannot be simultaneously expressed, even across verbal elements.

However, besides the combining of *present* tense and *perfect* aspect—which represent different dimensions of meaning that do not conflict—surprisingly, *perfect* aspect and *progressive* aspect can also be combined across verbal elements.

15. He has (**pres+act**) been (**perf+inaet**) giving (**prog+act**) me the ball

In this example, “has” expresses *present* tense and *active* voice, “been” expresses *perfect* aspect, with *inactive* voice overriding the *active* voice of “has”, and “giving” expresses *progressive* aspect and *active* voice which overrides the *inactive* voice of “been”. It may be that the combination results in an iterative interpretation that is at once *progressive* in iterating and *perfect* in the completion of each iteration (e.g. “He has been giving me the ball over and over”). It is an open research question how to represent the projection of two aspectual features (i.e. *perfect* and *progressive*) in a single clause. The computational model currently supports projection of a single aspectual feature.

*Progressive* aspect can be combined with *passive* voice across verbal elements.

16. He is (**pres+inact**) being (**prog+inact**) given (**perf+pass**) the ball

In 16, “is” projects *present* tense, “being” projects *progressive* aspect, and “given” projects *passive* voice. It is unclear if “given” projects *perfect* aspect in this example—it appears not to (the gray font for **perf** indicates this).

*Perfect* aspect can combine with *progressive* aspect and *passive* voice across verbal elements.

17. He has (**pres+act**) been (**perf+inact**) being (**prog+inact**) given (**perf+pass**) the ball

In 17, “has” projects *present* tense and *active* voice, but *active* voice is subsequently overridden by the *inactive* voice of “been”. “Been” projects *perfect* aspect, “being” projects *progressive* aspect (perhaps overriding the *perfect* aspect of “been”), and “given” projects *passive* voice, with *perfect* aspect questionable. This clause expresses a complex collection of tense, aspect and voice features across four verbal elements.

## Feature Projection for Intransitive Verbs

When we consider intransitive verbs like “go”, the introduction of the *inactive* voice feature becomes especially important. The intransitive v-en form is particularly revealing. Consider the verb “gone”.

18. He has (**pres+act**) gone (**perf+inaet**)

Like typical v-en forms of transitive verbs, “gone” expresses *perfect* aspect when preceded by “has”. But why do we need *inactive* voice for intransitive verbs? Because intransitive verbs can occur with “be” just like transitive verbs:

19. He is (**pres+inact**) gone (**perf+inact**)

There is clearly an expression of completion in this example, reflected in the projection of *perfect* aspect from “gone”, but the active involvement of the referent of “he” is de-emphasized. This de-emphasis is the intransitive verb equivalent of passivization in transitive verbs. In the intransitive verb case, there is no object available to be promoted to the subject function. Instead, the subject of the intransitive verb is demoted from active participant to inactive participant, but remains the subject.

Now consider a set of even more revealing examples:

- 20. He has (**pres+act**) tired (**perf+inact**)
- 21. He is (**pres+inact**) tired (**perf+inact**)
- 22. He is (**pres+inact**) very tired (**perf+inact**)

In “he has tired”, “tired” is the v-ed (non-finite) verb form. Since “has” projects *active* voice, the *inactive* voice of “tired” is blocked, but *perfect* aspect projects. In “he is tired (all of a sudden)”, it is unclear if *perfect* aspect projects. If it doesn’t, then the clause is *present* tense and *inactive* voice. Since “tired” is an intransitive verb, *inactive* voice demotes the subject making it an inactive participant. We are left with an expression that has essentially the same force as an adjectival expression—a single subject argument that is an inactive participant, and an auxiliary + verb combination that lacks any aspectual feature. If we view *stative* force as the lack of any aspect (either perfect or progressive), then the expression is effectively *stative*. Many researchers, including Huddleston & Pullum (2002) and Quirk et al. (1985) treat “tired” in 21 and 22 as an adjective. Huddleston & Pullum (2002, p. 1436) claim that the ability of a word like “tired” to combine with the adverb “very” is a definitive test for an adjective. Quirk et al. (1985, p. 167) make a similar claim. However, it is hard to see how this test is definitive given that “tired” has the form of a v-ed verb. The assumption that “tired” is an adjective when combined with “be” and a verb when combined with “have” necessitates two entries in the mental lexicon to represent “tired”. The approach advocated here requires a single verb entry, but allows the context to control the projection of grammatical features such that an intransitive verb can function very much like an adjective. As a challenge to the claim that “very” definitively identifies an adjective, consider

- 23. He is (**pres+inact**) very worn out (**perf+inact**)

It is atypical of adjectives, and typical of verbs to combine with prepositions to form verb-particle constructions. “Worn out” appears to be a typical verb-particle construction, except that it can occur with “very”. There is also a sense in which “worn out” implies completion of the process of wearing out as encoded by *perfect* aspect.

In general, we argue against the dual treatment of inflected verbs, including stative verbs, as adjectives since this introduces an ambiguity that does not facilitate processing. However, this does not mean that there is never an ambiguity between verbs and adjectives. Consider

- 24. The door is (**pres+inact**) open

“Open” appears to be a genuine adjective in that it does not have any verb inflection and it occurs after “is” where v-base verb forms do not occur. (Note that “\*He is tire” is not grammatical.) If “open” is genuinely ambiguous, how does the incremental, pseudo-deterministic processor deal with it? If we restrict “is” to setting a bias for non-finite inflected verb forms (e.g. v-ing, v-en or v-ed (non-finite)), adjectives and prepositions, then “open” will be biased to the adjective, rather than the v-base verb form, in the context of “is”. Note that this bias will not be sufficient if “gone” is both a v-en verb form and adjective, or “tired” is both a v-ed (non-finite) verb form and adjective.

Huddleston & Pullum (2002, p. 1436) note that expressions like “they were married” are ambiguous between an adjectival and a verbal interpretation. In “they were married last week” the verbal interpretation dominates, and in “they were married for ten years” the adjectival interpretation dominates. Is it possible to handle this ambiguity without positing distinct entries in the mental lexicon?

- 25. They were (**past+inact**) married (**perf+pass**)
- 26. They were (**past+inact**) married (**perf+pass**)

If the verbal interpretation corresponds to the projection of *perfect* aspect and *passive* voice, and the adjectival interpretation corresponds to suppression of *perfect* aspect and *passive* voice, then we can represent the distinction without positing separate entries in the mental lexicon. One immediate advantage of this approach is an ability to handle post verbal modification via feature overriding:

- 27. They were (**past+inact**) married (**perf+pass**) last week
- 28. They were (**past+inact**) married (**perf+pass**) for ten years

In the first example, the relatively punctual nature of “last week” encourages the expression of *perfect* aspect, whereas in the second example, the durative nature of “for ten years” discourages and perhaps overrides the expression of *perfect* aspect—although there still appears to be an implication that they are no longer married. The “adjectival” use also lacks *passive* voice. In the case of transitive verbs like “marry”, *passive* voice applies to the event reading in which the agent of the event (e.g. the priest) is demoted from subject to optional oblique argument. In the case of “they were married for ten years”, we have a durative event that is stative-like and lacking an agent. Note that at the processing of the word “married” we do not know what affect post verbal modifiers will have or even if there will be any. In an approach which has separate verb and adjective entries for “married”, an incremental, pseudo-deterministic processor will run into problems. It is not possible to decide at “married” which entry is needed. Either both entries will need to be carried forward in parallel, or the processor must have some mechanism for backing up and trying the alternative. From an incremental processing perspective, neither of these is



attractive. The human language processor does not have sufficient resources to carry forward multiple options in parallel—at least not across multiple choice points where additional parallelism might be required. Backtracking is equally problematic. Resources are needed to store the alternatives to be considered on backtracking, and knowing when to backtrack is indeterminate. Our pseudo-deterministic processor eschews backtracking and constrains parallel propagation of alternatives, relying instead primarily on non-monotonic adjustment of the evolving representation via feature overriding and feature blocking to deal with many forms of ambiguity without positing multiple entries in the mental lexicon.

As a final example with “married”, consider

29. They are (**pres+inact**) being (**prog+inact**) married (**perf+pass**) by a priest

In this example, “are” expresses *present* tense and “being” expresses *progressive* aspect. Since “are” and “being” are forms of “be”, they express *inactive* voice. This allows “married” to express *passive* voice, but the *perfect* aspect of “married” is blocked by the *progressive* aspect of “being”. The result is a clause that is *present progressive* and *passive*. There is an ambiguity here: are they in the act of being married by a priest or is the event just planned for the future? Since the *present* tense ranges over future events in English, this ambiguity may not be resolvable in terms of feature projection or suppression.

There is a related ambiguity in the meaning of expressions with *progressive* verb forms. According to Huddleston & Pullum (2002, p. 80)

30. Her parents are entertaining

is ambiguous between “entertaining” as a progressive verb form and “entertaining” as a stative adjective. If we allow the *active* voice feature of “entertaining” to be suppressed then these two uses can result from a single verb entry:

31. Her parents are (**pres+inaet**) entertaining (**prog+ act**)  
 32. Her parents are (**pres+inact**) entertaining (**prog+ aet**)

In 31, the parents are actively involved in entertaining, whereas, in 32, the parents are not actively involved. It does not seem necessary to suppress *progressive* aspect in this example since *progressive* aspect is already stative-like compared to *perfect* aspect. Note that this allows us to handle “her parents are entertaining tomorrow” and “her parents are entertaining to be around” via feature projection or suppression without multiple entries in the lexicon.

## Summary

We described the representation and processing of the inflectional verb features tense, aspect and voice within the context of an incremental, pseudo-deterministic human language processor (Ball, 2011a). Diagrammatic trees generated during the execution of the processor which show

verb feature projection on a broad range of different inputs are available at <http://www.doublertheory.com/compiler/compiler-grammar.htm>.

Verbs, including auxiliary and modal verbs, are encoded with tense, aspect and voice features in the mental lexicon and these features can project to, or be expressed by, the clauses in which they occur. When the verb group contains multiple elements, the grammatical features of the verbal elements must be reconciled. Monotonic unification of grammatical features is not possible when the grammatical features conflict. Mechanisms of feature blocking and overriding are needed to handle the reconciling of incompatible features and to minimize the amount of ambiguity in the mental lexicon—at least when localist representations (cf. Sag, Boas & Kay 2012) in which all verbal features compete for expression at the clausal level are assumed. A non-localist alternative of using hierarchically organized features (as suggested by a reviewer) such that in “He has been kicked”, “has” takes “been” as a complement with “has” expressing *present* tense and “been” expressing *perfect* aspect on a second level, and “been” takes “kicked” as a complement which expresses *perfect* aspect on a third level, may handle the case of feature blocking, but doesn’t explain how the overall expression is *passive*, which requires overriding the higher level *active* feature of “have” at the clausal level—if verbs express a voice feature as is assumed.

## References

- Anderson, J. (2007). *How Can the Human Mind Occur in the Physical Universe?* NY: Oxford University Press.
- Ball, J. (2010a). Context Accommodation in Human Language Processing. *Proceedings of the NLPCS Workshop*. INSTICC.
- Ball, J. (2010b). Projecting Grammatical Features in Nominals. *Proceedings of the 19<sup>th</sup> BRIMS Conference*.
- Ball, J. (2011a). A Pseudo-Deterministic Model of HLP. *Proceedings of the Cognitive Science Conference*.
- Ball, J. (2011b). Explorations in ACT-R Based Cognitive Modeling: Chunks, Inheritance, Production Matching and Memory in Language Analysis. *Proceedings of the AAAI Fall Symposium: Advances in Cognitive Systems*.
- Gazdar, G., Klein, E., Pullum, G. & Sag, I. (1985). *Generalized Phrase Structure Grammar*. Cambridge: Harvard.
- Huddleston, R. & Pullum, G. (2002). *The Cambridge Grammar of the English Language*. Cambridge, UK: Cambridge Univ Press.
- Marcus, M. 1980. *A Theory of Syntactic Recognition for Natural Language*. Cambridge, MA: The MIT Press.
- Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Essex, UK: Pearson Education Limited.
- Sag, I., Boas, H & Kay, P. (2012). Introduction to Sign-Based Construction Grammar. In *Signed-Based Construction Grammar*, edited by H. Boas & I. Sag. Stanford: CSLI.
- Sag, I., Kaplan, R., Karttunen, L. Kay, M., Pollard, C., Shieber, S. & Zaenen, A. (1986). Unification in Grammatical Theory. *Proceedings of the 5<sup>th</sup> WECOL*. 228-254. Stanford: CSLI.
- Sag, I., Wasow, T. & Bender, E. (2003). *Syntactic Theory, a Formal Introduction*. Stanford: CSLI.

# Cognitive reserve and intelligence: Modulating the effects of damage in ageing dynamical systems

Frank D. Baughman<sup>1</sup> (frank.baughman@curtin.edu.au)

Natalie Baughman<sup>2</sup> (natalie.baughman@uwa.edu.au)

Simon A. Mills<sup>1</sup> (simon.mills@curtin.edu.au)

<sup>1</sup>Curtin University, School of Psychology, Bentley, Perth 6102

<sup>2</sup>University of Western Australia, School of Psychology, Crawley, Perth 6009

## Abstract

The term cognitive reserve (CR) is used to describe the lack of direct relationship between the severity of brain damage, or pathology and subsequent levels of observed impairments. It has been suggested by Stern (2009), that CR may reflect differences in (a) pre-existing levels of some reserve capacity of the brain (the *passive* form); or, (b) differences in the underlying functional architectures supporting cognitive processes (the *active* form). In this paper, we explore two implementations of cognitive reserve that seek to target both these forms, extending recent work using dynamical systems framework (Baughman & Thomas, 2008; van der Maas et al., 2006). We examine how variability in cognitive reserve may modulate the effects of damage, at different levels of intelligence. The resulting simulations showed that level of intelligence does not differentially modulate the pattern of cognitive change following complete destruction of a single cognitive process, but that the effects of damage are proportionate across each level of intelligence. Following the two implementations of cognitive reserve that we tested, we found: (1) higher levels of connectivity within a given architecture resulted in greater spread of damage and lower endstate performance; and, (2) functional architectures that are characterized by greater specialization of function, rather than distributed function, differentially protected against the effects of damage, with these models also exhibiting better recovery.

**Keywords:** Cognitive reserve; intelligence; ageing; damage; recovery; dynamical systems; functional architecture.

## Introduction

The term *cognitive reserve* (CR) is often used in relation to the pattern of general cognitive decline found in normally ageing adults, and to the more extreme forms of cognitive breakdown seen following brain damage (e.g., stroke), or disease (e.g., dementia and Alzheimer's). In healthy ageing adults, the term is used to refer to the variability observed between individuals of the same age. In clinical samples, the term refers to the observation that levels of brain damage, or pathology have no clear relationship to the severity of subsequent impairments. This is to say, two individuals with similar levels of brain damage may exhibit different cognitive profiles (e.g., the impairments for one individual may be subtle, while for the other they may be much more pronounced). The lack of direct relationship between the degree of pathology, or brain damage and clinical manifestation, has led to the suggestion that individuals differ with respect to their pre-existing levels of cognitive reserve (Stern, 2002; Stern, 2009).

A number of studies have reported mixed findings concerning the extent to which factors such as ones levels of intelligence, educational attainment, occupation and activity are associated with reduced risk of dementia, stroke, and lower levels of general decline (Kaplan et al., 2009; Koenen et al., 2009; Nithianantharajah & Hannan, 2009; Tucker-Drob, Johnson, & Jones, 2009; Whalley, Deary, Appleton, & Starr, 2004; Zahodne et al., 2011). However, these studies have not yielded causal accounts detailing how variability in CR may directly influence cognitive performance. Theoretical accounts of CR have however distinguished two broad forms (Stern, 2009). The *passive model* posits that CR may be delivered through differences in pre-existing reserve levels of some capacity of the brain (e.g., this might be number of neurons, or number of connections). Under this view, damage to a cognitive system with lower pre-existing levels of capacity, will lead to poorer outcomes, compared to cognitive systems where these levels are higher. The *active model* describes that differences in CR may be explained by differences in functional architectures underlying cognition. Under this view, it is hypothesized that some functional architectures are more efficient, and thus more resilient to the effects of damage, than others (Stern, 2009). Computational approaches provide an ideal platform from which to examine these issues because they provide an explicit framework for testing how various neurocomputational properties may directly lead to changes in a cognitive system. Here, we describe one approach using dynamical systems theory which aims to capture a broad pattern of development across a range of cognitive profiles and which allows for the consequences of damage to be assessed at the level of the whole cognitive system and across time.

## Computational approaches to the study of ageing and damage

Computational studies to ageing, and to damage in ageing systems, have mostly focused on the effects of variation to three main parameter manipulations: (1) reducing the slope of gradients in activation functions (Li, Von Oertzen, & Lindenberger, 2006); (2) reducing the connectivity between processes (Alstott, Breakspear, Hagmann, Cammoun, & Sporns, 2009); and, (3) removal, or deletion of processing units to simulate neuronal death (Rubinov, McIntosh, Valenzuela, & Breakspear, 2009). The effects of these parameter manipulations can be subtle and varied. However,

their effects are generally that they show reduced levels of performance, require that networks need more time to learn (akin to older adults needing more time to learn, compared to younger adults) and result in a more protracted process of recovery, following damage. Individual differences in ageing and damage within a cognitive system with more CR might thus be explained by: (a) steeper gradients in the activation functions; (b) a greater number of pre-existing levels of connections (or, weights) between processing units; or, (c) lower rates of cell death.

Thomas (2008) recently examined issues concerning ageing and cognitive reserve within a connectionist model of English past tense learning. In these simulations, aging was implemented separately via: (1) a reduction of gradient in processing units; and, (2) a reduction (loss) of connections. CR was implemented via manipulating the number of hidden units within the model. Specifically, low cognitive reserve models were assigned 50 hidden units (a level just sufficient to allow the model to learn) and high cognitive reserve models were assigned 100 hidden units. Damage, applied at various different timepoints, was implemented by probabilistically removing 50% of the connection weights in the network. This work is notable in that it provides an explicit test of one role of CR in modulating the effects of damage within a cognitive domain. There are relatively few studies that have sought to develop on this approach. Furthermore, most computational approaches to date appear to have targeted the capacity reserve (passive) form of CR proposed by Stern. We argue that a better understanding is needed for how the use of different functional architectures may modulate the effects of damage.

In this paper we examine the effects of damage within ageing dynamical systems models. Our central goal is to test two implementations of CR. We implement passive and active forms of CR proposed by Stern (2009). In the first instance, we assess the effects of varying the degree of connectivity between processes in a given architecture. In the second instance, we examine how the use of different functional architectures may modulate the effects of damage. Our target architectures are the Fully distributed, Hemispheric, Central processor, Bi-directional and Uni-directional architectures, represented in Figure 1. We further aim to examine how intelligence levels may modulate the patterns of damage, given the different implementations of CR.

### Dynamical systems theory

Dynamical systems theory (DST) provides one way of addressing these questions as it offers a framework for exploring the interaction between multiple component processes in a cognitive system. This then allows the possibility of tracing the consequence of changes to a given system over time. By specifying the relationship between component processes, we may stipulate exactly what the functional architecture is, and then test how the effects of ageing and damage unfold in a particular architecture. We base our approach, on the ‘mutualism model’ of intelligence

first proposed by van der Maas and colleagues (2006) and which was subsequently extended by Baughman and Thomas (2008) to explore the effects of early focal impairments to a process within a range of different functional architectures.

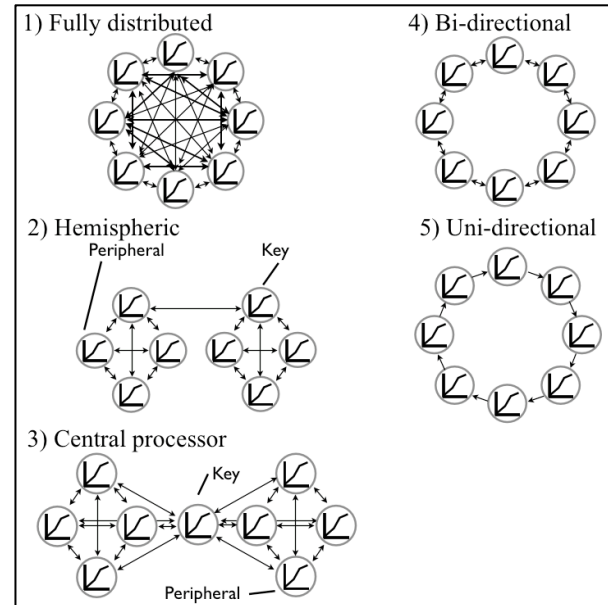


Figure 1: Five alternative model architectures. Note, this figure is illustrative of the architectural design. The actual models contained 16 processes each, and 17 in the case of the Central processor model.

### The mutualism model

van der Maas et al. (2006) offered a fully connected dynamical systems model of the development of intelligence that simulates cognitive development for a number of components (depicted by Model 1, of Figure 1). The model provides a number of parameters that influence development for each individual processes, but where development of the model, as a whole, is influenced dynamically by all processes within the model. A key feature of their model, is that the processes which are connected to each other within a system interact with one another and influence each other, in a mutually beneficial way throughout development. Hence, the model is called the ‘mutualism’ model. Equation 1 gives the dynamics of the mutualism model.

$$\frac{dx_i}{dt} = a_i x_i \left( \frac{1 - x_i}{K_i} \right) + a_i \sum_{\substack{j=1 \\ j \neq i}}^w M_{ij} x_j x_i / K_i$$

Equation 1. The mutualism model (van der Maas (2006))

The mutualism equation is derived from population dynamics and the Lotka-Volterra equation. Briefly, the equation states that at each point in time ( $t$ ) the change in the performance level  $x$  of a given process  $i$  ( $dx_i$ ) is a

product of the sum of the interaction weights of each process  $j$  to which it is functionally connected ( $M_{ij}x_j$ ), multiplied by the rate of growth of process  $i$  ( $a_i$ ), multiplied by the current level of performance of process  $x_i$ , divided by the asymptote level for that process ( $K_i$ ). Changes in  $x_i$  at each time step are thereby constrained by the performance (and thus the individual properties) of all other processes to which it is connected.

Extending the mutualism model, Baughman and Thomas (2008) showed that following impairments to a single process, early on in development, architectures characterized by greater connectivity between processes offered greater *compensation* and showed reduced levels of *spread* of damage. Additionally, they showed that compensation and spread were further modulated by where in the cognitive system impairments were applied. Baughman and Thomas distinguished *peripheral* processes from those that occupied *key* positions within a given architecture. For example, while in the Fully distributed model all processes are equal (and so impairment to one process is equivalent to damage to any other process), this is not the same for the Hemispheric and Central processor architectures. Both these models contain peripheral processes (e.g., processes in one hemisphere which do not directly influence processes in the other hemisphere) and key processes (e.g., processes within one hemisphere share a direct connection to processes in the other hemisphere). Figure 1 illustrates the distinction between peripheral and key processes. The effects of damage to peripheral versus key processes within different functional architectures remains largely untested. As such, it is not obvious whether the same architectures that offer advantages following damage to processes early in development, will also offer advantages to damage later in development.

## Simulations

In both the Normal and Damaged models, Ageing and IQ were implemented by manipulating values of the capacity for each process ( $K$ ).

**Ageing:** General cognitive decline was simulated by applying a fixed level of decay (0.075%) to the capacity ( $K$ ) of each process from 400 timesteps onwards. For the present simulations, we did not examine the consequences of variability in the rate, or the onset of decay.

**IQ:** To create Low IQ, Average IQ and High IQ models, models were calibrated to begin with different starting values of  $K$  (Low IQ=2, Average IQ=3, and High IQ=4).

**Cognitive Reserve:** For the passive form of CR, we tested three levels of Connection strength between processes ( $M$ ). However, because the boundaries of values that this parameter accepts without exhibiting catastrophic effects are limited, the range we implemented was small. We used  $M=0.049$ ,  $M=0.050$ , and  $M=0.051$ , to simulate Low, Average and High Connectivity, respectively. For the active form of CR, we compared the effects of damage in Fully distributed, Hemispheric, Central processor, Bi-directional, and Uni-directional architectures (see Figure 1).

**Damage:** In the damaged models, a single process was removed from the cognitive system to simulate total destruction of that process. Damage was applied separately to a peripheral process in each architecture, then to the key processes in the Hemispheric and Central Processor architectures. We held constant the level of damage (one process was damaged in under all architectures) and the onset of damage. Damage was applied to either a peripheral or a key process at timepoint 550, just over half-way through the models 'lifetime'. The damaged process was thus removed from the network and the relevant connections to and from it, also removed. All other parameters specified in the mutualism model, namely those relating to the growth rates of processes ( $a$ ), and the initial starting states of each process ( $x$ ) were also held constant and did not vary in these simulations ( $x=0.05$ ,  $a=6.0$ ). Finally, because one of our primary concerns was examining specific levels of IQ, we were not concerned with population variability. Thus, we did not require the models to be run for many pseudosubjects and only a single model was run for each architecture in Figure 1 for 1000 time steps. The full set of models that we tested totaled 108. These were comprised of: (i) Normal ageing models at 3 levels of IQ (Low, Average and High) within 3 levels of Connectivity (Low, Average and High) and 5 Architectures (Fully distributed, Hemispheric, Central processor, Bi-directional, and Uni-directional); (ii) Peripherally-damaged ageing models (as Normal, but with one process damaged); and, (iii) Key-damage ageing models (IQ: Low, Average and High x Connectivity (Low, Average and High) x 2 Architectures (Hemispheric, Central processor). Figure 2 shows the trajectories for Normal and Damaged models for the Fully distributed, Central processor and Uni-directional architectures, at Average IQ, Average Connectivity levels.

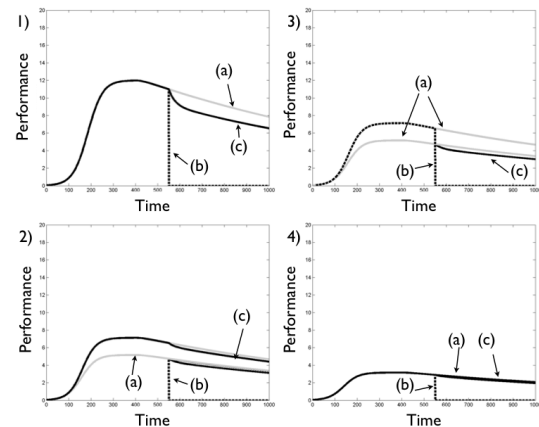


Figure 2: Trajectories of Normal and Damaged models for the Fully distributed (Tile 1), Central Processor (Tile 2: damage to peripheral process; Tile 3: damage to key process), and Uni-directional (Tile 4) architectures at Average Connectivity ( $M=0.05$ ) and Average IQ ( $K=3$ ) levels. Tiles depict processes in the Normal models (a) with a grey line, and the damaged (b) and affected processes (c) in the Damaged models, with dashed and solid black lines, respectively.

## Measures

Asymptote levels in the architectures differ as a consequence of the number of processes that are connected within it. As such, comparisons between the absolute levels reached by two architectures would be misleading. Instead, we use each Normal model as the benchmark for which to compare the performance of its damaged counterpart. This allows for relative comparisons across the different architectures. The two key metrics we use to assess the effects of damage are: (1) Area - the extent to which the trajectories of processes in the Damaged model resembles those in the Normal model (we compute the area under the curve, for each Damaged processes, and this is turned into a proportion of the area of the processes in the Normal model); and, (2) Endstate level - the extent to which the endstate levels of the Damage model reaches the functional endstate of the Normal model. Thus, area gives a measure of models attempt to compensate for damage, and endstate provides a measure of the models ability to recover.

## Results

Table 1 provides the Area data for Normal and Damaged models, at each level of intelligence and each level of Connectivity. The table shows effects of manipulations to IQ and Connectivity, across each of the architectures tested. The uppermost part of the table provides the data for comparisons for Normal versus Peripherally-Damaged models, the lowermost part of the table shows these comparisons for Normal versus Key-Damaged processes, in the Hemispheric and Central Processor models.

### Intelligence

As expected, varying the level of intelligence (IQ) in a model had direct effects on the overall level of performances reached. Table 1 shows that for each architecture higher IQ models performed better compared to lower IQ models (e.g., the level of performances reached in the Uni-dimensional architecture at each level of IQ, under Low Connectivity, are 11769.8, 17654.6 and 23539.5, respectively). However, the results of the simulations showed that IQ level did not modulate the effects of damage within architecture, at the various levels of CR. That is, within a given level of Connectivity, the effect of damage was proportionate at each level of IQ. For example, in the Low Connectivity Fully distributed model, the proportion of area reached by the Damaged models in Low IQ, Average IQ and High IQ models were all 80.2% of Normal levels.

### Cognitive reserve as differences in connectivity

Varying CR, when implemented as level of Connectivity, showed small, but consistent effects on level of performance reached (e.g., the levels reached in the Normal Hemispheric model at each level of Connectivity, under High IQ, are 35425.8, 35852.7 and 362900.0. However, greater levels of CR did not protect models from the effects of damage. In

fact, the reverse was found to be the case. Increased connectivity between processes resulted in higher proportion of spread of damage and poorer endstate recovery. This outcome was true for all architectures, but most apparent in the Fully distributed model. Figure 3 shows the proportion of area and endstate levels reached for each architecture, at each level of Connectivity, following peripheral damage.

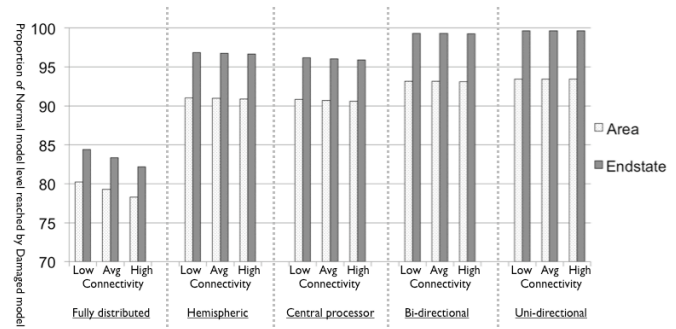


Figure 3: Proportion of Area and Endstate obtained in peripherally-damaged models by level of Connectivity.

Figure 4 shows that these effects are further exaggerated by damage to the key processes, in the Hemispheric and Central processor models. This figure shows that following damage to the key process, the Hemispheric model reached levels of recovery that were similar to the peripherally-damaged model (the greatest difference between key and peripherally-damaged process endstate was 1%). In the Central processor model, endstates differed by approximately 9%. The figure also shows that in the Central Processor model, key damage resulted in both lower recovery (Endstate) and more protracted form of recovery (Area).

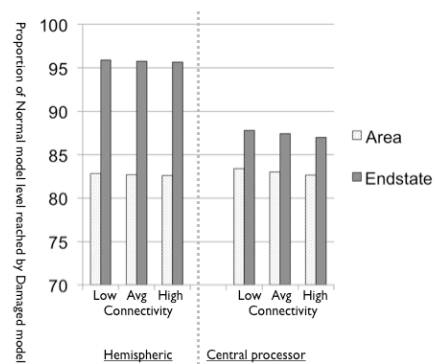


Figure 4: Proportion of Area and Endstate obtained in key-damaged models by level of Connectivity.

Table 1. Calculations of area under the curve for trajectories from the Normal and Damaged models

Normal models	Fully Distributed Connectivity			Hemispheric Connectivity			Central Processor Connectivity			Bi-Directional Connectivity			UniDirectional Connectivity		
	Low	Average	High	Low	Average	High	Low	Average	High	Low	Average	High	Low	Average	High
Low	42237.8	44772.1	47629.9	17712.9	17926.3	18145.0	20610.9	20929.6	21258.5	12409.1	12436.7	12464.4	11769.8	11782.1	11794.6
Average	63356.7	67158.1	71444.8	26569.3	26889.5	27217.5	30916.3	31394.4	31887.8	18613.7	18655.1	18696.6	17654.6	17673.2	17691.8
High	84475.6	89544.1	95259.7	35425.8	35852.7	36290.0	41221.8	41859.2	42517.0	24818.2	24873.4	24928.8	23539.5	23564.3	23589.1
mean	63356.7	67158.1	71444.8	26569.3	26889.5	27217.5	30916.3	31394.4	31887.8	18613.7	18655.1	18696.6	17654.6	17673.2	17691.8
Damaged Peripheral	Low			Average			High			Low			Average		
	Low	Average	High	Low	Average	High	Low	Average	High	Low	Average	High	Low	Average	High
Low	33881.7	35499.0	37278.8	16127.0	16306.4	16489.9	18718.7	18984.4	19258.0	11559.8	11583.9	11608.1	10999.4	11010.2	11021.1
Average	50822.6	53248.4	55918.1	24190.5	24459.6	24734.8	28078.1	28476.6	28887.0	17339.7	17375.9	17412.2	16499.1	16515.4	16531.6
High	67763.4	70997.9	74557.5	32253.9	32612.8	32979.8	37437.4	37968.8	38516.0	23119.7	23167.8	23216.2	21998.8	22020.5	22042.2
mean	50822.6	53248.4	55918.1	24190.5	24459.6	24734.8	28078.1	28476.6	28887.0	17339.7	17375.9	17412.2	16499.1	16515.4	16531.6
Damaged Key		Connectivity			Connectivity										
		Low	Average	High	Low	Average	High								
IQ		Low	14669.3	14827.4	14988.8	17187.1	17378.1	17573.5							
		Average	22070.7	22308.5	22551.6	25780.6	26067.1	26360.2							
		High	29490.6	29808.5	30133.5	34374.1	34756.2	35146.9							
		mean	22076.9	22314.8	22558.0	25780.6	26067.1	26360.2							

## Cognitive reserve as differences in functional architecture

Implementing CR, as different functional architectures, did modulate the effects of damage. However, it was not those architectures characterized by more connectivity between processes that proved most resilient to damage. Indeed, it was those architectures comprised of more limited connectivity where the effects of damage were minimized and the endstate levels of recovery most complete. In the architectures tested here, this was the Uni-directional architecture. Damage to any process in this architecture had effects on processes downstream of the damaged process. But these effects became increasingly small, over the remainder of the models lifetime. Figure 5 shows Area (left) and Endstate levels (right), respectively, for each of the architectures tested.

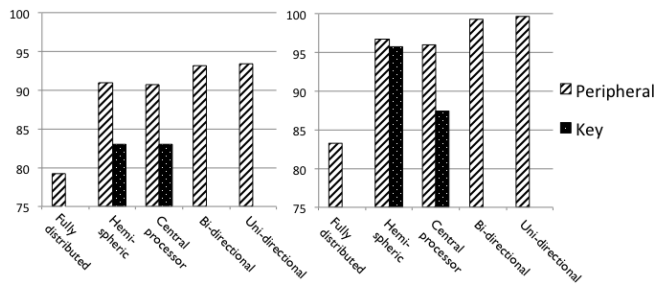


Figure 5: Comparisons of functional architectures by proportion of Area (left) and Endstate (right) obtained in peripherally-damaged and key-damaged models.

## Conclusions

Previous simulation studies have showed that following early forms of focal impairment, architectures characterized by greater levels of connectivity offer superior levels of protection compared to those with more limited connectivity (see Baughman & Thomas, 2008). However, in the simulations reported here, where permanent damage occurred to a system late on in its development, it was those models characterized by less connectivity (i.e., more specialized in function) that offered greatest resilience to damage. Examples of those offering the greatest protection are the Bi-directional and Uni-directional models, with the Fully distributed architecture offering the least protection following damage. These results indicate that throughout the process of development, similar events that impair just a limited number of processes to a system may have very different consequences for its outcome. These results are consistent with the notion that different functional architectures may underlie different stages of development (Fransson, Aden, Blennow, & Lagercrantz, 2011), possibly through a process of emergent specialization (Karmiloff-Smith, 2009). Future work is needed to investigate how the parameters we held constant (such as rate of decline, cognitive growth, and the severity and onset of damage) might provide a more complete account of the factors that contribute to real-world variability in ageing.

## References

- Alstott, J., Breakspear, M., Hagmann, P., Cammoun, L., & Sporns, O. (2009). Modeling the impact of lesions in the human brain. *Plos Computational Biology*, 5(6).
- Baughman, F. D., & Thomas, M. S. C. (2008). Specific impairments in cognitive development: A dynamical systems approach. In *Cognitive science*. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1819-1824). Austin, TX: Cognitive Science Society.
- Fransson, P., Aden, U., Blennow, M., & Lagercrantz, H. (2011). The functional architecture of the infant brain as revealed by resting-state fmri. *Cerebral Cortex (New York, N.Y. : 1991)*, 21(1), 145-154.
- Kaplan, R. F., Cohen, R. A., Moscufo, N., Guttmann, C., Chasman, J., Buttaro, M., . . . Wolfson, L. (2009). Demographic and biological influences on cognitive reserve. *Journal of Clinical and Experimental Neuropsychology*, 31(7), 868-876.
- Karmilof-Smith, A. (2009). Nativism versus neuroconstructivism: Rethinking the study of developmental disorders. *Developmental Psychology*, 45(1), 56-63.
- Koenen, K. C., Moffitt, T. E., Roberts, A. L., Martin, L. T., Kubzansky, L., Harrington, H. L., . . . Caspi, A. (2009). Childhood IQ and adult mental disorders: A test of the cognitive reserve hypothesis. *The American Journal of Psychiatry*, 166(1), 50.
- Li, S. C., Von Oertzen, T., & Lindenberger, U. (2006). A neurocomputational model of stochastic resonance and aging. *Neurocomputing*, 69(13-15), 1553-1560.
- Nithianantharajah, J., & Hannan, A. J. (2009). The neurobiology of brain and cognitive reserve: Mental and physical activity as modulators of brain disorders. *Progress in Neurobiology*, 89(4), 369-382.
- Rubinov, M., McIntosh, A. R., Valenzuela, M. J., & Breakspear, M. (2009). Simulation of neuronal death and network recovery in a computational model of distributed cortical activity. *American Journal of Geriatric Psych*, 17(3), 210.
- Stern, Y. (2002). What is cognitive reserve? Theory and research application of the reserve concept. *Journal of the International Neuropsychological Society*, 8(3), 448-460.
- Stern, Y. (2009). Cognitive reserve. *Neuropsychologia*, 47(10), 2015-28.
- Thomas, M. S. C. (2008). Ageing, plasticity, and cognitive reserve in connectionist networks. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2089-2094). Austin, TX: Cognitive Science Society.
- Tucker-Drob, E. M., Johnson, K. E., & Jones, R. N. (2009). The cognitive reserve hypothesis: A longitudinal examination of age-associated declines in reasoning and processing speed. *Developmental Psychology*, 45(2), 431-46.
- van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, 113(4), 842-61.
- Whalley, L. J., Deary, I. J., Appleton, C. L., & Starr, J. M. (2004). Cognitive reserve and the neurobiology of cognitive aging. *Ageing Research Reviews*, 3(4), 369-82.
- Zahodne, L. B., Glymour, M. M., Sparks, C., Bontempo, D., Dixon, R. A., MacDonald, S. W., & Manly, J. J. (2011). Education does not slow cognitive decline with aging: 12-Year evidence from the victoria longitudinal study. *Journal of the International Neuropsychological Society*, 17(6), 1039-46.



# Verb omission errors: Evidence of rational processing of noisy language inputs

Leon Bergen (bergen@mit.edu)<sup>1</sup>, Roger Levy (rlevy@ucsd.edu)<sup>2</sup>, Edward Gibson (egibson@mit.edu)<sup>1</sup>

<sup>1</sup>Department of Brain and Cognitive Sciences, MIT, Cambridge MA 02139,

<sup>2</sup>Department of Linguistics, UC San Diego, La Jolla CA 92093

## Abstract

We investigate the mechanisms that allow people to successfully understand language given noise in the world and in their own perceptual inputs. We address two parts of this question. First, what knowledge do people use to make sense of language inputs that may have been corrupted? Second, how much of this knowledge is used while people are processing sentences? We conduct a sentence production experiment and an on-line reading experiment in order to answer these questions. Both experiments provide evidence that syntactic knowledge can drive top-down reinterpretations of word identities, as well as syntactic reanalyses that are incompatible with people's language input. In addition, Experiment 2 provides evidence that this knowledge is deployed on-line, as people process sentences.

**Keywords:** Language Understanding; Sentence Production; Rational Analysis

## Introduction

People typically understand language effortlessly and successfully despite the fact that the language input can be corrupted in many ways between the speaker's planning of an utterance and the listener's comprehension of it. This suggests that the language comprehension mechanism has developed methods for correcting noise added to the input, in order to recover the speaker's true intent. We will be viewing this problem from the perspective of rational Bayesian inference. According to this position, people identify the intended language input by rationally integrating their prior linguistic expectations with the possibility of noise. The claim that people are at least somewhat rational when correcting for noise seems obviously true – if someone hears “The teacher write”, they are unlikely to perceive this as a totally different phrase, but are instead likely to interpret it as the very similar “The teachers write” or “The teacher writes.” A more interesting question therefore is: how thoroughly do rational noisy-channel effects permeate real-time comprehension at different levels and detail of linguistic representation?

This question can be broken into two parts:

1. What kinds of knowledge are deployed during processing in order to determine if an error has occurred in the input?
2. What kinds of reanalyses due to noise does the sentence processing mechanism pursue? Is it restricted to the reanalysis of single words, or will it pursue more radical reanalyses of the syntactic structure?

In order to address these questions, we will first discuss what is entailed by an ideal-observer perspective on sentence processing in the presence of noise. We will then survey some open questions about the extent to which the sentence processing mechanism is adapted to noise. Finally, we will present the results of a sentence production and a sentence comprehension experiment that each bear on these questions.

## Noisy channel models

According to a noisy channel model of sentence comprehension, a rational comprehender uses their perceptual input, which may have been corrupted, to identify the language input intended by the speaker. There are two sides to the comprehender's inference problem. First, prior linguistic knowledge should constrain people's inferences about what the speaker actually meant. If a particular phrase is very unlikely *a priori*, then it is even more unlikely that it was intended by the speaker but later corrupted by noise. For example, people's syntactic knowledge will inform how they interpret an ungrammatical sentence like “The voter hope there will be a recount.” There is a conflict between the two parts of this sentence: the phrase “The voter hope” is a noun-noun compound, but the phrase “there will be a recount” is the argument of a verb. This conflict can be resolved by a simple syntactic reanalysis of the first phrase. If the first phrase is reanalyzed as a noun-verb construction, then the second phrase will have its required verb, and the sentence will be grammatical. As a result, it seems very compelling to reanalyze the first phrase as “The voter hopes” or “The voters hope.”

While people's linguistic knowledge constrains their inferences about the intended input on one side, their knowledge about the process that generated the noise constrains them on the other. People do not usually intend to say one thing and then accidentally say something completely unrelated. Consequently, the noise process makes certain hypotheses about how the input was corrupted very unlikely. Similarly, our perception does not usually introduce massive errors into the input we receive. Hence in the example above, people only consider similar sentences when they are trying to figure out the intended meaning of “The voter hope.”

The problem of inferring the true intended input can be posed in terms of optimal Bayesian inference. The comprehender's prior linguistic information can be represented by a probability distribution  $P_L$ , which is defined over phrases or sentences. The model of the noise process is given by the distribution  $P_N$ . This distribution specifies how likely a particular noise event is to occur, e.g. the deletion of a letter in the perception of the input, or the accidental insertion of a morpheme by the speaker. By Bayes' Theorem, the probability that a sentence or phrase  $s$  was intended given the perceived sentence or phrase  $s'$  is equal to:

$$P(s|s') = \frac{P_L(s)P_N(s \rightarrow s')}{P(s')} \quad (1)$$

where  $P_N(s \rightarrow s')$  is the probability that  $s'$  will be perceived when  $s$  is intended by the speaker.

In general, having perceived the input  $s'$ , we can measure the comprehender's evidence for phrase  $s_1$  relative to  $s_2$  by

looking at the ratio

$$\frac{P(s_1|s')}{P(s_2|s')} = \frac{P_L(s_1)P_N(s_1 \rightarrow s')}{P_L(s_2)P_N(s_2 \rightarrow s')}. \quad (2)$$

The higher this value, the more evidence that  $s_1$  and not  $s_2$  was intended. We can apply this formula to the case that  $s_2 = s'$ . The resulting ratio

$$\frac{P(s|s')}{P(s'|s')} = \frac{P_L(s)P_N(s \rightarrow s')}{P_L(s')P_N(s' \rightarrow s')}. \quad (3)$$

can be interpreted as the probability that an error occurred and  $s$  was intended relative to the probability that no error occurred and  $s'$  was in fact intended.

These formulas capture some important intuitions about how people should infer the intended input, and also have some interesting consequences. Only sentences  $s$  that have a relatively high probability of causing the perceptual input  $s'$  will be plausible candidates for the intended meaning of the speaker. If this probability  $P_N(s \rightarrow s')$  is low, then by equation 1, the probability  $P(s|s')$  that  $s$  was intended must also be low. For example, if sentence  $s$  is very different than  $s'$ , then it would require a large number of specific errors to transform  $s$  into  $s'$ . As a result,  $P_N(s \rightarrow s')$  would be low, and there would be low probability that  $s$  was actually intended.

The model also captures an interesting tradeoff between the comprehender's linguistic knowledge and the noise model. It will always be the case that the easiest way to generate a perceptual input is for that perceptual input to have been intended: it is not necessary to posit any errors in this case in order to explain the perceptual input. However, the comprehender may infer the presence of noise when there is an alternative sentence  $s$  that is sufficiently similar to  $s'$  and has higher probability according to the language model  $P_L$ . Formula 3 states that as the ratio  $\frac{P_L(s)}{P_L(s')}$  of the probability of  $s$  relative to  $s'$  increases, the comprehender will be more likely to infer that  $s$  was intended. One simple consequence of this is that spelling mistakes will be corrected by the comprehender: misspelled words will receive low probability under the model  $P_L$  relative to nearby, correctly spelled words.

A more interesting consequence is that under certain conditions, the comprehender should infer that noise was added to the input even when the actual sentence has a legitimate analysis. This can happen when the perceived sentence is well-formed but highly unlikely because of semantic implausibility or because it contains low-frequency syntactic constructions. If there is a sufficiently similar sentence  $s$  that has higher probability under the language model, then the ratio in formula 3 may be high enough for the comprehender to infer that  $s$  was actually intended.

Finally, the noisy channel model should lead us to predict that comprehenders will treat the presence of a language-unit (e.g. a letter, morpheme, or word) differently from its absence. In particular, comprehenders should be more likely to infer that a perceived language unit was intended by the speaker than that the absence of a language unit was intended.

The Bayesian size principle of (Xu & Tenenbaum, 2007) explains this asymmetry. This principle states that when a hypothesis allows for many possibilities to occur, it must place a small amount of probability on each of these possibilities. More formally, the principle states that

$$P(h|H) = \frac{1}{|H|} \quad (4)$$

where  $H$  is a hypothesis containing  $h$ , and  $|H|$  is the number of possibilities contained in  $H$ . In the present setting, this implies that there is a lower chance of any specific language unit being accidentally inserted into a perceived sentence than there is of a specific language unit being accidentally deleted. To see this, imagine that because of perceptual noise, a random letter is either deleted from or inserted into part of the sentence. While there are 26 English letters that could be inserted at a given location, there is only one way to delete a specific letter. It follows that there is a higher probability that a particular sentence  $s_1$  was intended, but a letter was deleted from it, than that another sentence  $s_2$  was intended, but a letter was inserted into it. The same reasoning applies to noise generated from speech errors, or insertions and deletions that occur on larger language units such as morphemes and words.

## Previous work

While these types of comprehension behavior are predicted by an ideal-observer model, it is unclear if they are borne out in human comprehension. Previous work has addressed parts of this question. A number of studies have shown the influence of non-syntactic factors on how people determine word identity. In the speech perception literature, researchers have demonstrated context effects in word recognition (Marslen-Wilson, 1975, 1987; Dahan & Tanenhaus, 2004). However, our understanding of two aspects of word recognition remains more limited: (i) the role of syntactic ambiguity as a contextual factor for word recognition; (ii) the consequences of misrecognizing a word, or remaining uncertain as to its identity, for downstream comprehension of the sentence. Event plausibility has also been shown to induce word misrecognition and subsequent processing difficulty, but such processing difficulty has not yet been demonstrated when different syntactic analyses are involved (Slattery, 2009).

Other modeling and experimental studies have provided evidence for the effect of syntax on the correction of noise. Levy (2008) proposed a Bayesian model of noisy-channel sentence comprehension similar to the present one, and suggested that such models might shed light on a number of syntactic-comprehension phenomena difficult to reconcile with traditional sentence-processing models. Gibson and Bergen (2012) provided evidence suggesting that comprehenders can come to global misinterpretations of complete sentences that are incompatible with the true string on the basis of both semantic and syntactic information, but this work does not show how these misinterpretations unfold in real time, and only considers cases in which a word is inserted or

swapped, and not those in which a word is substituted. There is also evidence that comprehenders can be induced to disregard orthographic material (commas) in reading on the basis of strongly biased grammatical expectations, and to adopt a garden-path syntactic analysis that should be incompatible with the actual orthographic material (Levy, 2011), but this has not been shown to happen when actual word identities are at stake. Finally, there is evidence that comprehenders can entertain the possibility that a previous word was misrecognized on the basis of alternative syntactic analyses of an input prefix (Levy, Bicknell, Slattery, & Rayner, 2009). However, this does not show that comprehenders will actually pursue these alternative syntactic analyses further in the sentence.

The present studies aim to fill several gaps in this literature. We will provide evidence that the misrecognition of words can be driven by the comprehender considering grammatical analyses that are incompatible with the true input and taking into account fine-grained (word-word-collocation-dependent) preferences for syntactic analysis. Our evidence will suggest that comprehenders pursue these alternative analyses downstream in the sentence. Finally, we will provide evidence that comprehenders' inferences about noise are rationally sensitive to the asymmetry between insertions and deletions.

### Experiment 1: Verb omission errors

We discussed two distinctive claims of the noisy channel model in the previous section:

1. Comprehenders should infer that a perceived phrase contains an error when there exists a nearby syntactic construction with a much higher base-rate of occurrence. This is true even when the phrase appears well-formed.
2. Comprehenders should be more likely to infer that part of the intended phrase was deleted than that part was inserted. As a result, they should be more likely to believe that a phrase was intended when the only nearby alternatives could have only produced the phrase via insertion.

We test these claims using a timed sentence completion task, with participants asked to complete short sentence preambles. Participants were briefly shown a sentence preamble. This preamble then disappeared, and they were asked to type a complete sentence beginning with this preamble.

Participants were shown preambles that were unambiguously noun-verb ("NV"), unambiguously noun-noun ("NN"), or NN/NV ambiguous constructions (Frazier & Rayner, 1987; MacDonald, 1993). Crucially, each NV preamble differed by only a single morpheme ("-ed") from an NN preamble, and similarly for all of the NN preambles. For example, "The immigrant feared" was used as an unambiguous NV construction, because "feared" must be interpreted as the main verb of the resulting sentence, while "The immigrant fear" was used as an unambiguous NN construction, because "fear" must be interpreted as a noun in the resulting sentence. In addition, the preambles varied in their degree of bias towards the NV or NN constructions. A summary of these con-

ditions is provided in Table 1. The purpose of the ambiguous-preamble continuations was to validate the categorization of each item as NN-biased or NV-biased (see Results below).<sup>1</sup>

Table 1: Experiment 1 conditions

Condition	Example
NV-biased/NN preamble	The immigrant fear
NV-biased/NV preamble	The immigrant feared
NV-biased/ambiguous preamble	The immigrant fears
NN-biased/NN preamble	The almond roll
NN-biased/NV preamble	The almond rolled
NN-biased/ambiguous preamble	The almond rolls

Our noisy channel model predicts a specific pattern of *rational misidentification* of the unambiguous sentence preambles. Participants should primarily misidentify less likely constructions in favor of more likely constructions, and should primarily infer that a morpheme was dropped from a construction, not that it was added. Under the model, therefore, there are two criteria both of which must hold for misidentifications to be most frequent. The first is that the preamble be NN, since NV can be converted to NN by deletion of a single letter or past-tense morpheme, whereas NN-to-NV conversion requires an insertion. The second is that the preamble be NV-biased, since the NV construction is *a priori* more likely in this case. We thus predict that NV-biased NN preambles should be the most likely to be misidentified; in each of the other conditions at least one of the two above criteria fails to hold, and we therefore expect fewer misidentifications of the preambles in these cases.

There are several ways that such a pattern of rational misidentification could manifest itself in our sentence completion paradigm. First, if participants misidentify the preamble, then we expect *repetition errors* in which they retype the preamble incorrectly. For example, if they misidentify an NN preamble as an NV preamble, then we expect them to retype a preamble containing a verb with a past-tense morpheme.

We may also find a more interesting effect which results from participants' rational *uncertainty* about the identity of the sentence preamble. The evidence that participants receive from an NV-biased NN preamble may lead them to be rationally uncertain about whether an NN or NV preamble was intended, as there is still a tradeoff between positing noise and moving to a more probable construction. An optimal sentence processing mechanism would maintain representations of both of these possibilities after encountering the preamble.

We may find evidence of these multiple representations in the form of *verb omission errors*. In such sentences, participants would correctly reproduce the NN preamble, but complete the sentence as though they had already introduced a

<sup>1</sup> Some of our unambiguous preambles do have alternate syntactically permissible readings, e.g. "The almond rolled ice cream was good," but these are low-frequency constructions, and participants never used them in their completions.

main verb. The following sentence provides an example:

- (1) The immigrant fear being deported.

The word “fear” is a noun here, though it would need to be a verb for the sentence to be grammatical. If participants produce such sentences, then this would be evidence that they are maintaining uncertainty about the identity of the preamble, and are repeating the preamble according to one representation, but completing it according to the other.

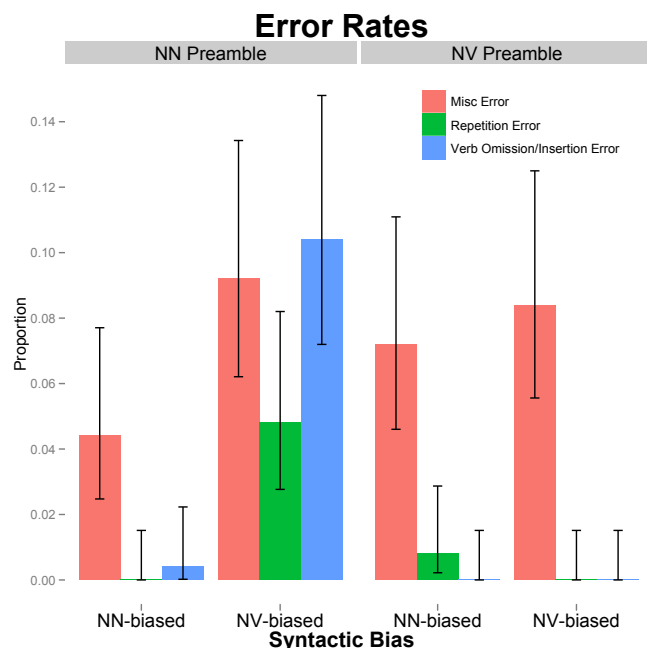


Figure 1: Experiment 1 results. The left panel shows the error rates in the NN condition, depending on the syntactic bias of the construction, while the right panel shows the error rates in the NV condition. Error bars are 95% confidence intervals. Note that, according to the coding criteria, verb omission errors occurred only in the NN-preamble condition, while verb insertion errors occurred only in the NV-preamble condition.

## Methods:

**Participants:** Sixty native English speakers from the United States were recruited from Amazon Mechanical Turk. They were paid a small amount of money for participation.

**Materials and Procedure:** Participants were shown each sentence preamble for 1.5 seconds. After this, the preamble disappeared from the screen, and participants were given 13 seconds to retype the preamble exactly and complete the sentence. During the instructions, participants were shown several instances of incorrect completions; for example, they were told that if they were asked to complete “The dog in the park”, then they could not add anything to the preamble as in “The dog in the parks.”

Items consisted of 12 NN-biased and 12 NV-biased preambles that were selected for their bias. These preambles were presented in one of three conditions: NN, NV, or an ambiguous condition. These items were presented in a within-subjects design. The ambiguous condition consisted of plural or present-tense items of the form “The immigrant fears”, which is consistent with either an NN or NV reading. This condition was used for norming the items: higher rates of NV completions indicated NV-bias, and similarly for NN completions. Finally, 12 unambiguous NN and 12 unambiguous NV items were used as fillers; these were distinct from the test items in not being in the morphological neighborhood of an alternative construction.

**Coding:** Completed sentences were coded as correct if the sentence preamble was repeated correctly and the sentence was grammatically well-formed. A sentence was coded as a repetition error if its preamble was repeated with a morphological error that switched its grammatical role (e.g. if an NN preamble was repeated with past-tense morphology) and the rest of the sentence was grammatically consistent with the repeated preamble. In the NN conditions, a sentence was coded as a verb omission error if the preamble was repeated correctly but the rest of the sentence grammatically required a main verb to appear in the preamble. In the NV conditions, a sentence was coded as a verb insertion error if the preamble was repeated correctly but the rest of the sentence grammatically required the preamble to be a noun phrase.

Of the total responses, 7% contained miscellaneous errors, which did not fall under the other criteria. For 74% of these errors, the sentence did not contain a complete independent clause; most of the remainder contained word substitutions in the preamble, or number or tense agreement errors.

## Results and discussion

We first analyzed whether the NN and NV-biased items were biased in the correct direction. This was determined using the completions for the ambiguous condition. All 12 of the NN-biased items received NN completions most frequently; 11 of the 12 NV-biased items received NV completions most frequently, and 1 was equi-biased. Of the 229 NN-biased items in the ambiguous condition completed correctly, 197 (86%) were given NN completions. Of the 228 NV-biased items completed correctly, 184 (81%) were given NV completions.

We next looked at whether repetition errors and verb omission errors occurred more frequently for the NN preambles. We tested this using repeated measures ANOVAs; Figure 1 shows the frequencies of each type of error for each experimental condition.<sup>2</sup> There were significantly more repetition errors in the NN preambles than in the NV preambles (12 vs. 2;  $p < 0.05$ ). In addition, there were significantly more verb omission errors in the NN preambles than noun omission errors in the NV preambles (27 vs. 0;  $p < 0.001$ ). We note that for the NN preambles, the repetition errors were approx-

<sup>2</sup>We did not use mixed logit models because of convergence issues with random slopes.

imately evenly distributed among errors to the plural marking of the first noun, past-tense errors on the second noun, and present-tense errors on the second noun.

Our final question was whether the error rate for the NN items was higher for the NV-biased items (we did not analyze this for the NV items, because these contained so few errors). Because the bias was manipulated between-items, we tested this using ANOVAs with subjects as random factors. All of the repetition errors for the NN items occurred on NV-biased items (12 vs. 0;  $p < 0.01$ ). All but one of the verb omission errors occurred on NV-biased items (26 vs. 1;  $p < 0.001$ ).

These results provide evidence for both predictions of the noisy channel model. First, we observed the predicted asymmetry between insertions and deletions: nearly all of the observed errors were consistent with the participants inferring that a morpheme had been dropped from the preamble they observed. Second, we observed that nearly all errors were made in the direction of the more probable construction.

We also found evidence for a more tentative prediction of the model, which is that participants would maintain uncertainty about the identity of the preamble. Verb omission errors patterned in the same manner as the repetition errors; moreover, they occurred at twice the rate of the repetition errors. This suggests that participants were maintaining multiple representations of the sentence preamble; verb omission errors occurred when more than one representation was deployed in the completion of the sentence.

## Experiment 2

In Experiment 2, we used a different method to evaluate the predictions of the noisy channel model. The model predicts that people will adopt incorrect syntactic analyses if there exist similar phrases that could have easily generated them. In such cases, we should be able to observe the effects of misidentification downstream in sentence comprehension: specifically, comprehenders should have difficulty if the later parts of a sentence are inconsistent with their interpretation.

We used NN and NV preambles like those in Experiment 1 to probe such garden-path effects. If people misidentify an NN preamble as an NV preamble, then they should be surprised when a main verb is used later in the sentence; conversely, if they misidentify an NV preamble as an NN preamble, then they should be surprised when the clause ends without a main verb. For example, in the Dense-neighborhood/NN condition in Table 2 – which is so named because there are other grammatical phrases in the morphological neighborhood of its preamble – people should sometimes infer that the past tense “chauffeured” was intended instead of “chauffeur”, and therefore they should be surprised when they arrive at the main verb “hoped.” On the other hand, if they infer that “chauffeur” was intended instead of “chauffeured” in the Dense-neighborhood/NV condition, then we would expect difficulty at “but”, which indicates the end of the first clause. Such effects have been shown for truly ambiguous NN/NV preambles such as “The voter hopes” (Frazier &

Rayner, 1987), and have also been shown to be affected by the collocation’s lexical bias (MacDonald, 1993), but have never been demonstrated when the preamble is unambiguous.

We used self-paced reading to evaluate whether people had difficulty at these “disambiguating” regions (note, however, that these items are only ambiguous given the possibility of noise). People’s performance at these regions was evaluated against two control conditions, Sparse-neighborhood/NN and Sparse-neighborhood/NV, which were given this name because their preambles were not in the morphological neighborhood of alternative syntactic constructions. These conditions were identical to the Dense-neighborhood conditions, except for this difference in the density of the morphological neighborhood. For the Sparse-neighborhood/NN condition, this was done by using an adjective before the head noun, and for the NV condition, this was done by using the quantifier “some” and a plural morpheme before the main verb.

The noisy channel model predicts an interaction between the density of the morphological neighborhood and grammatical structure at the disambiguating region. In particular, because of the asymmetry between morpheme insertion and deletion, people misidentify the preamble most frequently in the Dense-neighborhood/NN condition. We should expect less difficulty at the disambiguating region of the Dense-neighborhood/NV condition because the syntactic alternative of the preamble would require an insertion to produce the perceived wordform. We should similarly expect less difficulty in both Sparse-neighborhood conditions, because these preambles are far from alternative syntactic constructions.

Table 2: Experiment 2 conditions

Condition	Example
Dense-neighborhood/NN	The intern chauffeur for the governor hoped for more interesting work.
Dense-neighborhood/NV	The intern chauffeured for the governor but hoped for more interesting work.
Sparse-neighborhood/NN	The inexperienced chauffeur for the governor hoped for more interesting work.
Sparse-neighborhood/NV	Some interns chauffeured for the governor but hoped for more interesting work.

## Methods:

**Participants:** We recruited 120 native English speakers from the United States from Amazon Mechanical Turk. They were paid a small amount of money for participation.

**Materials and Procedure:** We tested participants using a self-paced reading program that ran in participants’ web browsers. Words were presented one at a time. After each sentence, participants were asked a comprehension question.

We constructed 16 items in the conditions shown in Table 2. These conditions varied in a within-subjects design. We also included 32 filler sentences during testing.

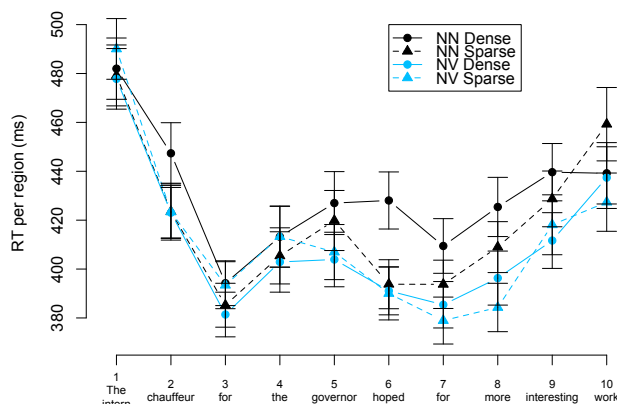


Figure 2: Reading-times from Experiment 2. Region 6 is the critical disambiguating region.

## Results and discussion

Before analyzing the data, we excluded participants that answered fewer than 80% of the comprehension questions correctly, and excluded trials on which reading times were farther than 2 standard deviations from the mean. Accuracies on comprehension questions were above 90% on all conditions. The results are shown in Figure 2. For the analysis, words were aligned relative to the disambiguating word, which is labelled as region 6 and is the first place we can expect to find any critical effects of disambiguation.

Our first question was whether there was an effect of neighborhood density for the NN items, for which we expected difficulty at the disambiguating region. We performed our analysis with a linear mixed-effects model with random slopes and interactions for participants and items. We found a significant increase in RTs at the disambiguating region for the Dense-neighborhood items ( $\beta=33.08$  ms,  $t=2.49$ ,  $p < 0.05$ ). We next looked at whether this effect was strongest for the NN items. There was a significant interaction between neighborhood density and syntactic structure ( $\beta=35.06$  ms,  $t=2.21$ ,  $p < 0.05$ ), indicating a superadditive effect of density and structure on RTs. No interactions were significant ( $p>0.05$ ) at any region prior to the critical region.

These results provide evidence that participants misidentified the sentence preambles, and that this misidentification was rational. Specifically, we found evidence that participants were most likely to misidentify preambles that could have been produced by an alternative phrase via a deletion. This is consistent with the asymmetry between insertions and deletions implied by the Bayesian size principle.

## Discussion

We have investigated the knowledge that people use to correct noise in their language input, as well as the on-line process-

ing mechanisms that support this error correction. Regarding the first topic, our experiments provide evidence that comprehenders entertain and even adopt syntactic reanalyses of their language input to account for the possibility of noise, even when these reanalyses are inconsistent with the true input. Moreover, we found evidence that these reanalyses are driven by the rational integration of grammatical expectations and expectations about the noise process. The results of Experiment 1 suggest that people prefer alternative explanations for their input that involve higher probability syntactic constructions, and noise consisting deletions rather than insertions.

We have also found evidence that people employ these alternative syntactic analyses during on-line sentence processing. In Experiment 2, we found downstream effects of noise-driven reinterpretations at the point when they contradicted the input sentence. This suggests that the sentence processing mechanism is not delayed in positing or making use of syntactic reanalyses during the course of comprehension. This is the first result demonstrating that comprehenders will pursue garden-path syntactic analyses differing from the true sentence preamble by a word substitution, and that garden-path disambiguation in these cases incurs measurable costs; this result is directly predicted by our noisy-channel model. Together with other recent work, these results raise new questions regarding the full breadth of sentence-processing phenomena that may be best understood as the consequence of rational, noisy-channel probabilistic inference.

## References

- Dahan, D., & Tanenhaus, M. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 498.
- Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of memory and language*, 26(5), 505–526.
- Gibson, E., & Bergen, L. (2012). The rational integration of noise and prior semantic expectation: Evidence for a noisy-channel model of sentence interpretation. *Submitted*.
- Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 13th conference on empirical methods in natural language processing* (pp. 234–243).
- Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *Proceedings of the 49th annual meeting of the association for computational linguistics*.
- Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50), 21086.
- MacDonald, M. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*, 32, 692–692.
- Marslen-Wilson, W. (1975). Sentence perception as an interactive parallel process. *Science*, 189(4198), 226.
- Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition*, 25(1), 71–102.
- Slattery, T. (2009). Word misperception, the neighbor frequency effect, and the role of sentence context: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 35(6), 1969.



# Gestural Alignment in Natural Dialogue

Kirsten Bergmann (kirsten.bergmann@uni-bielefeld.de)

Stefan Kopp (skopp@techfak.uni-bielefeld.de)

Faculty of Technology, Center of Excellence "Cognitive Interaction Technology" (CITEC)

Collaborative Research Center "Alignment in Communication" (SFB 673)

Bielefeld University, P.O. Box 100 131, D-33501 Bielefeld, Germany

## Abstract

A well-known phenomenon in natural interaction is that speakers adapt their linguistic and nonverbal behaviors. Research on *gestural* alignment is, however, still in its early stages based on evidence from experimental settings. This paper provides a first systematic study of gesture form convergence based on a large sample of naturalistic dialogue data. We found evidence for gestural alignment, but not all form features of co-speech gestures are subject to this effect. In a detailed analysis of those sensitive features we further address questions of how gestural alignment depends on the temporal distance between gestures, and whether *intra*-speaker or *inter*-speaker influences on gesture form are stronger.

**Keywords:** Alignment, co-speech gestures, natural interaction

## Introduction

Co-speech gesturing is an integral part of human communication, but it is not well understood why and how gestures take on their particular physical form. This holds especially for iconic gestures that apparently communicate by virtue of iconicity, i.e., through a correspondence between their form and geometrical or spatial properties of what they refer to. Empirical studies, however, revealed that similarity with the referent cannot fully account for all occurrences of iconic gesture use (Kopp, Tepper, Ferriman, Striegnitz, & Cassell, 2007). Findings also indicate that a gesture's form is influenced by other contextual constraints such as discourse (Holler & Stevens, 2007) or the linguistic context (Kita & Özyürek, 2003). In addition there are considerable differences in how speakers gesture, partly assumed to be due to different cognitive abilities (Hostetter & Alibali, 2007).

In addition to intra-speaker sensitivities, co-speech gesturing in dialogue may also be influenced by the gestures of the interlocutor. A large body of work has demonstrated inter-personal sensitivities in verbal and nonverbal behavior in natural social interaction, leading often to coordination and alignment between interlocutors (cf. (Kopp, 2010)). For example, linguistic coordination has been reported with respect to words, phrase structures, speech rate, tones of voice, speech rhythms, etc. (cf. Chartrand, Maddux, and Lakin (2005); Branigan, Pickering, Pearson, and McLean (2010)). Pickering and Garrod (2004) ascribed this linguistic alignment largely to an automatic priming of interlocutors lexical, syntactic, or semantic representations and a percolation of activation between adjacent representational levels. Others, e.g. (Brennan & Clark, 1996), suggested that speakers strategically design utterances for an addressee and thereby prefer previously used (grounded) constructions. Regarding nonverbal behavior, interactants can likewise be found to mimic each

other, e.g., in posture, body movements like foot shaking, mannerism, or facial expressions (cf. Chartrand and Bargh (1999); Lakin and Chartrand (2003)). This kind of mimicry is assumed to be largely non-conscious and automatic, being mediated by perception-action links that involve the own motor system in the perception of others actions (Dijksterhuis & Bargh, 2001). Only recently, researchers have started to look at whether speakers also align in their *co-speech* gestures, i.e., the spontaneous and meaningful hand movements that accompany speech. Such gestures stand out as they are very closely linked to the speech they accompany, in both content and timing (McNeill, 1992). Investigating whether speakers align in and via their gestures can thus help, first, to understand what shapes these gestures and, second, to shed light on the role of interpersonal coordination in dialogical communication.



Figure 1: Example of alignment of two successive gestures (left: router's assertion; right: follower's acknowledgement).

In this paper we present results from the first large-scale investigation of gestural alignment in natural dialogue. Fig. 1 shows an example, in which some properties of the first gesture (left) are being mimicked (e.g., handshape, trajectory) while others are not (e.g. relative movement direction). This suggests a feature-based, multi-level analysis of gestural alignment. We will thereby focus on the *form-based aspects* of gestural alignment here. In the next section we review the few existing studies that have looked at occurrences of form convergence in co-speech gestures, so far. Then we present the corpus data and the approach taken to investigate the phenomenon of gestural alignment in it, and present results of three analyses meant to answer the following questions: (1) Is there gestural alignment, compared against a baseline, and are there differences between different form features of a gesture? (2) What is stronger, the influence of the interlocutor's gestures or of one's own previous gesturing? (3) How does gestural alignment depend on (temporal) distance between the gestures? Finally, we discuss our findings in light of these questions and draw conclusions.



## Related Work

Based on initial evidence by Kimbara (2006), who reported a couple of examples of gesture form convergence among interlocutors, some recent studies addressed the phenomena of gestural alignment (in this context often termed ‘mimicry’) more deeply. Parrill and Kimbara (2006) investigated the question to what extent observing mimicry affects people’s behavior. They found that participants who observed mimicry in a video-recorded interaction were subsequently more likely to reproduce the mimicked behavior in their own descriptions, whereby a gesture was assessed as a reproduction if it corresponded with the stimulus gesture in handshape, motion and location.

In a similar setting, Mol, Krahmer, Maes, and Swerts (2012) provided evidence for the alignment of handshapes in co-speech gestures: Participants who saw a speaker in a video stimulus using gestures with a particular handshape were more likely to produce gestures with these handshapes later on, while retelling the story. This evidence is, however, limited to a particular kind of gestures (‘path gestures’ in directions), distinguishing between two different handshape classes (index finger extended vs. more than one finger extended). Mol et al. further addressed the role of meaning in this context. They found that gesture forms were only repeated across speakers if they had occurred in a meaningful context as expressed in concurrent speech. It is concluded that gesture form adaptation resembles adaptation in speech, rather than it being an instance of automated motor mimicry.

Kimbara (2008) studied triadic interaction with two co-narrators providing a joint narration to a third person, while manipulating the mutual visibility between co-narrators. Greater convergence in one gesture form feature (handshape) was found when participants could see each other. However, in this setting the two narrators were required to provide a coherent description for the recipient which might enhance the likability for gesture form convergence. Holler and Wilkin (2011) showed that gesture mimicry also occurs in face-to-face dialogue. In repeated references to the same figure-like stimuli, participants were found to be more likely to use similar gestures when they could see each other (vs. a non-visible condition). Holler & Wilkin concluded that gestures seem to play an active role in the process of grounding, because the vast majority of mimicked gestures occurred in phrases devoted to the presentation or acceptance of information.

In sum, existing studies lend considerable evidence that a speaker’s gesture use is influenced by others’ gestures. However, this quantitative empirical evidence is limited to experimental settings with video-based stimuli or elicited repeated references—an caveat often put forward against studies in linguistic alignment (cf. Howes, Healey, and Purver (2010)). To date, there is no analysis of gestural alignment based on a large sample of naturalistic dialogue data. The present study aims to close this gap.

## Present Study

We have conducted statistical analyses on a large data corpus of spontaneous speech and gesture in dialogue (SaGA corpus (Lücking, Bergmann, Hahn, Kopp, & Rieser, 2010)). With these analyses we aimed for a systematic investigation of gesture form convergence going beyond previous studies in several respects. First of all, our corpus provides a detailed coding of the gestures’ physical form including handshapes, palm and finger orientation, wrist movement, and position. This allows for addressing the degree to which *single* gesture form features are sensitive to influences of an interlocutor’s gestures, instead of considering the “same overall form” (Holler & Wilkin, 2011), one particular form feature only (Kimbara, 2008), or the sum of several form features (Parrill & Kimbara, 2006). Second, some of the above mentioned experimental studies manipulated the visibility between interactants to create a baseline for gestural alignment occurring by chance. Investigating natural dialogues allows for an alternative baseline by creating artificial dialogues, as previously done in corpus analyses of linguistic alignment (Howes et al., 2010); for details see sect. ‘Control Data’. Third, a characterizing feature for alignment in speech corpora is that the repetition probability is increased immediately after the prime and decreases toward the global mean with greater distances between prime and target (Reitter, 2008). A corpus analysis on extended dialogue allows to address this issue for gestural alignment, too. Fourth, we are able to investigate the *contingencies* involved in gestural alignment. Pickering and Garrod (2004) suggested to treat alignment not only as an *inter*-subjective phenomenon (‘other-alignment’), but also *intra*-subjectively (‘self-alignment’). Given the fact that the use of co-speech gestures is subject to major inter-individual differences (Hostetter & Alibali, 2007), the relationship between self- and other-alignment is important to assess the strength of inter-speaker gesture form convergence. Finally, the above mentioned studies have in common that they are limited to gestural alignment in repeated references to the same referent. Analyzing a large data sample allows to study the degree of gesture adaptation on a level of gesture form beyond the connection to specific referent objects, allowing to delineate grounding and mere motor resonances. In the following, we will briefly describe the corpus and explain how we framed the problem of detecting and measuring alignment between gestures occurring in dialogue.

**Data Corpus** The SaGA corpus consists of 25 dyads (21 female, 29 male participants) engaged in a spatial communication task combining direction-giving and sight description. This task required participants to convey the shape of objects and the spatial relations between them. The stimulus was an artificial town presented in a Virtual Reality environment, affording experimental control for the content of speaker messages. After taking a “bus ride” through the town, a router explained the route to an unknown and naïve follower. In total, the SaGA corpus consists of 280 minutes of

video material containing 4449 iconic/deictic gestures. All dialogues are completely and systematically annotated based on an annotation grid, tested and refined using multi-coder agreement tests; for details see Lücking et al. (2010). Each gesture is demarcated by the beginning and end of the expressive, so-called ‘stroke’ phase. The gesture’s form during the stroke phase, as far as relevant here, is annotated in terms of the following distinct form features: (1) **HANDEDNESS**: one-handed (either left- or right-handed), or two-handed gestures; (2) **HANDSHAPE**: ASL-based coding of hand configurations like ASL-B, ASL-C, etc. + modifiers (e.g. bent, loose); (3) **PALM- AND FINGER ORIENTATION**: up, down, sideways, towards body, away + combinations and sequences; and (4) **WRIST MOVEMENT TYPE**: static, linear, or curved + sequences. A further important and characterizing feature of an iconic gesture is the more general **REPRESENTATION TECHNIQUE** (e.g. Kendon (2004)). For the spatial domain of the SaGA dialogues, the following set of techniques proved to be adequate: indexing, placing (as if putting a virtual object somewhere), shaping (as if sculpting a 3D shape), drawing (as if sketching a 2D outline), and posturing (using the hand/arm as a model for something).

**Prime-target Pairs** From 4449 iconic/deictic gestures in the SaGA corpus, a total of 17130 prime-target pairs<sup>1</sup> were extracted. Figure 2 exemplifies the possible alignment-relevant influences between different prime-target pairs that can occur in dialogue. Each of these pairs is characterized by a distance **DIST** between prime and target gesture, taken to be the number of other gesture occurrences in-between plus 1. Each prime-target pair is further characterized by a **CONTINGENCY TYPE**: whether prime and target gestures are produced by the *same* speaker (‘self-pair’) or by *different* speakers (‘other-pair’), respectively. In the SaGA corpus there are 17362 self-pairs and 3993 other-pairs.

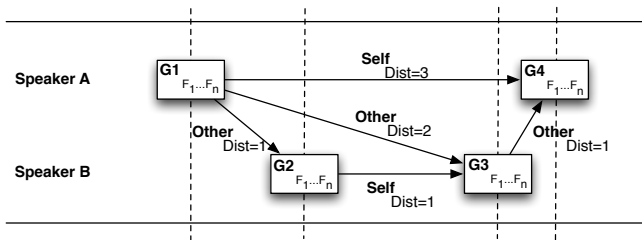


Figure 2: Possible alignment influences between gestures: Speaker A produces two gestures (G1, G4), while speaker B makes two gestures (G2, G3) in-between. Gestures are characterized by features ( $F_1 \dots F_n$ ) and can influence each other both within a speaker (‘self’) and across speakers (‘other’), where the relation’s distance (DIST) is determined by the occurrences of gestures in-between.

<sup>1</sup>We employ the term ‘prime-target pair’ in lack of a better one. This is not to imply that alignment is due to *priming*.

**Control Data** A common problem in studies on behavioral coordination is to lay down a baseline of how much coordination can occur simply by chance, regardless of any contingencies between primes and targets. Adopting the approach of Howes et al. (2010) in their corpus analyses of lexical and syntactic alignment in speech, we created ‘fake’ dialogues by re-combining the gestures of two speakers from originally different dialogues. This is done in an interleaved fashion, i.e., the whole sequence of gestures produced by one particular direction-giver is kept, but merged with the complete gesture sequence produced by a different direction-follower. This way we created 25 control dialogues with randomly chosen participants while respectively maintaining the participants’ role (direction giver vs. direction follower). As a matter of course, although the total number of gestures remains the same, this results in a different number of prime-target pairs in the control data set with regard to **CONTINGENCY TYPE**: 16523 self-pairs and 2407 other-pairs.

**Metric** Considering gestural alignment necessitates to define a metric estimating the similarity between prime and target gesture. Since we want to be able to assess alignment even at the level of single features of a gesture, we define a metric for each particular gesture feature. To make results comparable with each other, we employ a binary metric for all variables: it scores 1 if prime and target gesture are identical in a particular gesture feature, and 0 otherwise. For some features this definition can be applied straightforwardly (e.g. **HANDEDNESS**: one-handed vs. two-handed), for others it is reasonable to allow some minor variation between prime and target gesture. Palm and finger orientation, for instance, are coded as combinations of five basic values (up, down, sideways, towards, away). That is, a palm orientation of ‘down’ and an orientation of ‘down/away’ would count as a mismatch although the actual difference in palm orientation is 45° which can be regarded a slight deviation given the natural fuzziness of human gesture use. Accordingly, the binary metric is applied to the gesture features as follows, whereby for features which allow sequential coding the final segment of the prime’s value and the first segment of the target’s value are considered:

- **REPRESENTATION TECHNIQUE** and **HANDEDNESS**: A score of 1 is given only if the values for prime and target gesture are identical, 0 otherwise.
- **HANDSHAPE**: Any modifiers of ASL handshapes like ‘spread’ or ‘loose’ are omitted, i.e. ‘ASL-B-spread’ and ‘ASL-B-loose-spread’ both fall into the basic category ‘ASL-B’. A score of 1 is given only if prime and target are identical in this basic category for both hands, 0 otherwise.
- **PALM AND FINGER ORIENTATION**: A score of 1 is given if prime and target match in at least one part of the annotation value for both hands, 0 otherwise.

- **WRIST MOVEMENT TYPE:** A score of 1 is given if prime and target are identical or – in case of a two-handed gesture with different movement types – if the value for one hand is identical with the other gesture’s value, 0 otherwise.

## Results

**Is there gestural alignment in dialogue?** This analysis aims to show whether gesture use in real dialogues shows reliably more other-alignment than would occur by chance. To this end, we compare similarity scores in real vs. control dialogues for each gesture feature with a one-way analysis of variance for each of the gesture features. We only consider prime-target pairs with DIST=1 here, since it is more likely that alignment occurs in consecutive gestures than in more distant pairs. With regard to Figure 2 this means that we take prime-target pairs like (G1,G2) or (G3,G4) into account ( $N=950$  pairs; 579 from the original data, and 371 from control dialogues). Exact means and standard deviations are given in Table 1.

For REPRESENTATION TECHNIQUE ( $F_{(1,948)} = 24.61, p < .001$ ), HANDSHAPE ( $F_{(1,948)} = 17.92, p < .001$ ), and PALM ORIENTATION ( $F_{(1,948)} = 6.65, p = .01$ ), there is a reliable difference between the two groups such that the mean similarity in control dialogues is significantly lower than in real dialogues. For HANDEDNESS ( $F_{(1,948)} = 3.47, p = .063$ ) the analysis marginally fails to reach significance, but by trend the mean similarity in control dialogues is significantly lower than in real dialogues. For FINGER ORIENTATION ( $F_{(1,948)} = .16, p = .69$ ) and WRIST MOVEMENT TYPE ( $F_{(1,948)} = .06, p = .94$ ) the analysis shows no significant main effect between real and control data.

This means that there exists other-alignment in gesture use, but not all gesture features are subject to this effect. Only for the features REPRESENTATION TECHNIQUE, HANDSHAPE, PALM ORIENTATION, and HANDEDNESS the mean similarity of prime and target is higher as to be expected by chance. We continue with these features to a finer analysis.

Table 1: Mean similarity of gesture features for real and control dialogues (standard deviations in parentheses).

	Real	Control
Representation Technique	.31 (.37)	.16 (.46)
Handedness	.68 (.47)	.62 (.49)
Handshape	.37 (.48)	.24 (.43)
Palm Orientation	.49 (.50)	.41 (.50)
Finger Orientation	.61 (.49)	.60 (.49)
Wrist Movement Type	.40 (.49)	.40 (.49)

**Self- vs. Other-Alignment?** To compare the effects of self- and other-alignment, we now investigate the difference between prime-target pairs in the same speaker (CONTINGENCY TYPE = ‘self’) vs. prime-target pairs with different speakers (CONTINGENCY TYPE = ‘other’). We only consider adjacent prime-target pairs with DIST=1 (for instance, in Figure 2 self-pairs like (G2,G3) with other-pairs

like (G1,G2). The total number of pairs amounts to  $N=4317$  (3738 self-pairs, and 579 other-pairs). Again we employ a one-way analysis of variance, exact means and standard deviations are given in Table 2.

For all variables under consideration the analysis reveals significant main effects: REPRESENTATION TECHNIQUE ( $F_{(1,4315)} = 25.05, p < .001$ ), HANDSHAPE ( $F_{(1,4315)} = 51.86, p < .001$ ), HANDEDNESS ( $F_{(1,4315)} = 39.38, p < .001$ ), PALM ORIENTATION ( $F_{(1,4315)} = 67.95, p < .001$ ). These effects are due to the fact that mean similarity of prime and target are higher for self-pairs than in other-pairs. That is, the alignment between gestures is reliably stronger *within* speakers than it is *across* speakers.

Table 2: Mean similarity of gesture features for self- and other-speaker pairs (standard deviations in parentheses).

	Self	Other
Representation Technique	.41 (.49)	.31 (.46)
Handedness	.79 (.40)	.68 (.46)
Handshape	.53 (.50)	.37 (.48)
Palm Orientation	.67 (.47)	.49 (.50)

**Effect of temporal distance on other-alignment?** To elucidate how gestural alignment is affected by temporal distance, we analyze how the similarity score depends on the distance between prime and target gestures. For this analysis we consider other-pairs of distance 1-4. In Figure 2 examples of other-pairs with DIST=1 would be (G1,G2) or (G3,G4), examples of an other-pair with DIST=2 are (1,3) or (2,4). We employ a one-way analysis of variance for the dependent variable SIMILARITY and the independent variable DIST. A total of 3081 primed-target pairs is analyzed ( $N(\text{DIST}=1)=579$ ,  $N(\text{DIST}=2)=758$ ,  $N(\text{DIST}=3)=843$ ,  $N(\text{DIST}=4)=901$ ).

For REPRESENTATION TECHNIQUE there is a main effect of DIST and similarity score  $F_{(3,3077)} = 6.22, p < .001$ : the similarity score is smaller the greater the distance between prime and target. This is due to significant differences between prime-target pairs with DIST=1 and others (DIST=2:  $t_{(1335)} = 1.96, p = .05$ ; DIST=3:  $t_{(1420)} = 2.51, p = .012$ ; DIST=4:  $t_{(1478)} = 4.30, p < .001$ ), as well as between distances 2 and 4 ( $t_{(1657)} = 2.40, p = .017$ ). Likewise, for HANDSHAPE the similarity scores decrease significantly with increasing distance between prime and target DIST ( $F_{(3,3077)} = 7.10, p < .001$ ). This is due to the fact that the similarity of prime-target pairs with DIST=1 is higher than for prime-target pairs with higher distances (DIST=2:  $t_{(1335)} = 2.63, p = .09$ ; DIST=3:  $t_{(1420)} = 3.37, p = .001$ ; DIST=4:  $t_{(1478)} = 4.51, p < .001$ ). In contrast, for HANDEDNESS ( $F_{(3,3077)} = .045, p = .99$ ), and PALM ORIENTATION ( $F_{(3,3077)} = .41, p = .75$ ) there is no main effect of distance between prime and target gesture.

That is, the more gestures occur between prime and target, the smaller is their similarity with respect to REPRESENTATION TECHNIQUE and HANDSHAPE, which is corroborated when checking the actual temporal distances in milliseconds.

By contrast, the similarity score remains more or less constant for the features HANDEDNESS and PALM ORIENTATION.

Table 3: Mean similarity for varying distances between prime and target gesture (standard deviations in parentheses).

	DIST=1	DIST=2	DIST=3	DIST=4
Representation Technique	.31 (.46)	.26 (.44)	.25 (.43)	.21 (.41)
Handedness	.68 (.47)	.67 (.47)	.68 (.47)	.67 (.47)
Handshape	.37 (.48)	.30 (.46)	.29 (.45)	.26 (.44)
Palm Orientation	.49 (.50)	.49 (.50)	.48 (.50)	.47 (.50)

Together with results from the first analysis that revealed highest  $F$ -scores for the former two features in comparison with control data, this provides the following picture: For REPRESENTATION TECHNIQUE and HANDSHAPE there is a strong other-alignment effect which decreases with greater distances from the prime gesture. For HANDEDNESS and PALM ORIENTATION there is a rather weak difference when comparing original and control data and the effect is more or less constant. In other words, there seems to be a general (weak) tendency to produce gestures with a certain amount of similarity in these two features, but this is not biased by the other's directly preceding gesture use.

## Discussion

In this paper we reported results from the first fine-grained and systematic analysis of alignment in co-speech (iconic) gesturing in natural direction-giving dyads. What did we find? First, there is significant gestural alignment in dialogue. That is, a speaker's use of co-speech gestures is affected by the other's gestures in the dialogue. Remarkably, not all gesture features seem to be equally sensitive, with WRIST MOVEMENT and FINGER ORIENTATION being most resistant. Second, alignment effects are significantly stronger *within* speakers than *across* speakers. That is, a speaker's gestures influence each other more than the gestures the interlocutor performs, albeit the effectiveness of other-alignment. Third, regarding the relation between the strength of other-alignment and the prime-target distance, a multi-faceted picture emerges: alignment in handshape or representation technique becomes weaker with greater distances, while alignment in handedness and palm orientation remain constant.

These findings can shed new light on iconic co-speech gestures, as well as the cognitive processes underlying their production in dialogue. To start with, how can we make sense of the heterogeneity of feature-based gestural alignment? A closer look at the role of gestural representation techniques might be informative. Each of these techniques is characterized by a specific pattern of how meaning is depicted. For example, in drawing gestures as in Figure 1 it is the *wrist trajectory* that conveys most of the intended meaning, while in indexing or placing gestures the *position* of the hands is of major importance to convey meaning. That is, some features can be considered more communicatively significant than others. Indeed, variation within the less significant features has also been reported to reflect individual gesturing style (Bergmann,

2012). Our results here suggest that those features, like handshape, are also more amenable to inter-personal coordination, while the communicatively more significant features tend to be more resistant. This suggests a notion of gestural alignment as adaptation within the degrees of freedom available under given communicative constraints.

Existing cognitive models of speech and gesture production lack an account of other-alignment. However, our findings suggest a row of implications for such models. At first it is important to note that our analysis of gestural alignment at the level of different features supports a view that a gesture is not produced as a whole, but in different steps that can exert influences over the constitution of different features of a gesture. We thus hypothesize that different gesture form features are determined at different points in time with other-alignment arising potentially from high-level mechanisms in terms of full grounding, i.e. signaling established links between form and meaning, as well as low-level mechanisms of priming or motor resonance (Montgomery, Isenberg, & Haxby, 2007). That is, such a model provides (at least) two routes than could mediate alignment. However, our finding that communicatively significant features are less affected by alignment may be a consequence of a hypothesized more "ego-centric" nature of early, high-level stages of the production process, which may be more concerned with the necessary, communicatively intended functions and less with corrective or audience-design functions (Keysar & Henly, 2002). Lower-level processes, on the other hand, involve connected sensory and motor processes which have been shown to be effective also in gesture (Montgomery et al., 2007) and have often been assumed to mediate interpersonal coordination. Our empirical findings suggest that the sensorimotor route is particularly effective. To further distinguish between the two routes, we are concerned with measuring the degree of form similarity in relation to referent similarity in ongoing work.

Our observation of strong self-alignment effects may be explained by a strong role of internal priming or caching in the cognitive speech-gesture production process. This conforms hypotheses of self-routinization or *expert performance* effects, which state that over repeated encounters with a particular problem, memory traces build up that directly map a problem stimulus to a solution (e.g., Logan (1988)). It is important to note, however, that gesture use also seems to be subject to adaptations taking place in extended dialogue with repeated references. Such processes have been found, e.g., to lead to considerable *reduction* of the complexity of speech and gestures (Hoetjes, Koolen, Goudbeek, Krahmer, & Swerts, 2011). That is, gestures can become simpler or less precise over repeated uses when referring to the same entity. Further research is needed to elucidate how different mechanisms and driving forces (e.g., alignment through repetition vs. simplification through reduction) compete and interact with each other.

With all these raised questions, we think that computational simulation can provide a valuable tool to test whether

different kinds of cognitive mechanisms result in the effects we observe empirically. We have developed a computational model for speech and gesture production in previous work (Kopp, Bergmann, & Wachsmuth, 2008). Opening this to the effects of dialogical interaction, e.g. endowing it with perceptive abilities, will enable us to complement this empirical work with computational studies.

### Acknowledgments

This research is partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 “Alignment in Communication” and the Center of Excellence in “Cognitive Interaction Technology” (CITEC).

### References

- Bergmann, K. (2012). *The production of co-speech iconic gestures: Empirical study and computational simulation with virtual agents*. PhD Thesis: Bielefeld University.
- Branigan, H., Pickering, M., Pearson, J., & McLean, J. (2010). Linguistic alignment between humans and computers. *Journal of Pragmatics*, 42, 2355–2368.
- Brennan, S., & Clark, H. (1996). Lexical choice and conceptual pacts in conversation. *Journal of Experimental Psychology*, 22(6), 1482–1493.
- Chartrand, T., & Bargh, J. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76, 893–910.
- Chartrand, T., Maddux, W., & Lakin, J. (2005). Beyond the perception-behavior link: The ubiquitous utility and motivational moderators of unconscious mimicry. In R. Hassin, J. Uleman, & J. Bargh (Eds.), *The new unconscious* (pp. 334–361). New York: Oxford University Press.
- Dijksterhuis, A., & Bargh, J. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology*, 33, 1–40.
- Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (2011). Greebles greeble greeb: On reduction in speech and gesture in repeated references. In *Proc. of the 33rd annual conference of the cognitive society*.
- Holler, J., & Stevens, R. (2007). An experimental investigation into the effect of common ground on how speakers use gesture and speech to represent size information in referential communication. *Journal of Language and Social Psychology*, 26, 4–27.
- Holler, J., & Wilkin, K. (2011). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, 35, 133–153.
- Hostetter, A., & Alibali, M. (2007). Raise your hand if you're spatial—relations between verbal and spatial skills and gesture production. *Gesture*, 7, 73–95.
- Howes, C., Healey, P., & Purver, M. (2010). Tracking lexical and syntactic alignment in conversation. In *Proc. of the 32nd annual conference of the cognitive science society*.
- Kendon, A. (2004). *Gesture—visible action as utterance*. Cambridge University Press.
- Keysar, B., & Henly, A. (2002). Speakers' overestimation of their effectiveness. *Psychological Science*, 13, 207–212.
- Kimbara, I. (2006). On gestural mimicry. *Gesture*, 6, 39–61.
- Kimbara, I. (2008). Gesture form convergence in joint description. *Journal of Nonverbal Behavior*, 32, 123–131.
- Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48, 16–32.
- Kopp, S. (2010). Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. *Speech Communication*, 52, 587–597.
- Kopp, S., Bergmann, K., & Wachsmuth, I. (2008). Multimodal communication from multimodal thinking—towards an integrated model of speech and gesture production. *Semantic Computing*, 2(1), 115–136.
- Kopp, S., Tepper, P., Ferriman, K., Striegnitz, K., & Cassell, J. (2007). Trading spaces: How humans and humanoids use speech and gesture to give directions. In *Conversational informatics* (pp. 133–160). New York: John Wiley.
- Lakin, J., & Chartrand, T. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, 14, 334–339.
- Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, 95, 492–527.
- Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2010). The Bielefeld Speech and Gesture Alignment corpus (SaGA). In *Proceedings of the LREC 2010 workshop on multimodal corpora*.
- McNeill, D. (1992). *Hand and mind—What gestures reveal about thought*. Chicago: University of Chicago Press.
- Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language*, 66, 249–264.
- Montgomery, K., Isenberg, N., & Haxby, J. (2007). Communicative hand gestures and object-directed hand movements activated the mirror neuron system. *Social Cognitive and Effective Neuroscience*, 2(2), 114–122.
- Parrill, F., & Kimbara, I. (2006). Seeing and hearing double: The influence of mimicry in speech and gesture on observers. *Journal of Nonverbal Behavior*, 30, 157–166.
- Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27, 169–226.
- Reitter, D. (2008). *Context effects in language production: Models of syntactic priming in dialogue corpora*. PhD Thesis: University of Edinburgh.

# E Pluribus Multa In Unum: The Rationality Multiverse

Tarek R. Besold and Kai-Uwe Kühnberger  
(`{tbesold | kkuehnbe}@uni-osnabrueck.de`)

Institute of Cognitive Science  
University of Osnabrück  
49069 Osnabrück, Germany

## Abstract

The paper argues for a new view on and an approach to rationality as a concept of study and modeling paradigm of human behavior. After critically reviewing classical (normative) approaches to rationality, decision-making, and rational behavior, we present cornerstones of a positive, integrative, and holistic conception of these cognitive capacities. A discussion of key elements of this new view is given, and possible consequences and implications are considered.

**Keywords:** Rationality; Human-level intelligence; Subject-centered model; Cognitive capabilities; Integrative model.

## Introduction

With Aristotle's famous characterization of man as a rational animal ("zoon logikon") in his *Metaphysics* (Tredennick, 1933-35), and the ascription of a rational principle to the human being in his *Nicomachean Ethics* (Broadie & Rowe, 2002), the idea of seeing rationality and rational behavior as indispensable parts of our humanity has been introduced into human self-conception, starting centuries of inquiry into the nature and properties of this alleged *conditio sine qua non*. Still millennia after Aristotle, Descartes is only one amongst many famous scholars explicitly mentioning the question in his writings: "*But what is a man? Shall I say 'a rational animal'? No; for then I should have to inquire what an animal is, what rationality is, and in this one question would lead me down the slope to other harder ones (...)*" (Cottingham, Stoothoff, & Murdoch, 1984).

Nowadays, with the background of modern sciences, and especially with the advent of cognitive science, new questions and perspectives have been added to the question for the nature of rationality. Instead of merely studying properties and features of its manifestation in humans and their behavior, and measuring those against standards for rationality derived from normative theories, prediction and modeling aspects of theories of rationality are gaining more and more importance. These efforts can take various forms, ranging from formal studies of customer preferences and behavior, over cognitive modeling of decision-making and choice processes, to rational agency projects within AI.

Nonetheless, although this shift of emphasis within the study of rationality is clearly happening, it is a change of priorities within the existing overall setting, but not a revolutionary process overthrowing existing paradigms, or creating new approaches and ideas. In the following, we want to argue that the latter is what would be needed, on the one hand for real progress with respect to the aforementioned modeling and prediction tasks, but on the other hand also for the sake

of a deeper understanding and progress within rationality research itself.

The paper is structured as follows: The next section gives an overview of classical accounts and paradigms within rationality research, together with well-known objections and caveats to these standard frameworks. Then, a review of recent empirical and theoretical findings is presented, indicating a way to a new understanding and modeling of rationality and rational behavior. This new conception, for the time being called "subject-centered rationality", is sketched and elaborated on in a dedicated section. The penultimate section discusses basic features of this new stance in a juxtaposition of (expected) standard objections and reservations with solutions and answers to the former. Concludingly, the paper is put into a bigger context within the respective fields.

## Standards in Rationality

In the following, when talking about rationality, we mostly want to refer to the manner in which people derive conclusions when considering things deliberately in the domain of individual problem-solving, also including the conformity of one's beliefs with one's reasons for those beliefs or of one's actions with one's reasons for those actions. In consequence, at least for our considerations, beliefs and knowledge (which are also seen as presuppositions for respective actions), that is the epistemic aspects of rationality, are our main concerns. Within this context, rationality is intrinsically connected to an optimality principle, making a decision rational if it is not just reasoned, but if it is also in certain ways optimal for achieving a goal or solving a problem.

Several centuries of thought and investigation in relation to rationality and its manifestations in humans and their behavior resulted in the formation of mainly four abstract general models (most of which also bring along corresponding normative theories, and even definitions, of rationality):

- Logic-based systems (cf. e.g. (Evans, 2002)): A belief is rational, if there is a logically valid reasoning process to reach this belief relative to available/given background knowledge.
- Probability-based frameworks (cf. e.g. (Griffiths, Kemp, & Tenenbaum, 2008)): A belief is rational, if the expectation value of this belief is maximized relative to given probability distributions of background beliefs.
- Game theory-based models (cf. e.g. (Osborne & Rubinstein, 1994)): A belief is rational, if the expected payoff of

maintaining the belief is maximized relative to other possible beliefs.

- Accounts based on the use of heuristics (cf. e.g. (Gigerenzer, Hertwig, & Pachur, 2011)).

Unfortunately, when comparing the different conceptions, it shows that the frameworks (and also the resulting definitions of rationality) are in many cases almost orthogonal to each other, making them in the best case incommensurable, if not inconsistent or even partly contradictory in their modeling assumptions. Also, in many cases the predictive power of these classical theories of rationality turns out to be rather limited (at least when applied to real-world examples instead of artificially simplified and constructed scenarios), as they have more normative or postdictive-explanatory character.

Even more, although each of the listed accounts has gained merit in modeling certain aspects of human rationality, the generality of each such class of frameworks has at the same time been challenged by psychological experiments or theoretical objections. On the one hand, studies by Byrne on human reasoning with conditionals (Byrne, 1989) indicated severe deviations from classical logic (see below), a finding also supported by human subjects failing at a seemingly simple logical task in the famous Wason-selection task (Wason, 1966). Similarly, Tversky and Kahneman's Linda problem (Tversky & Kahneman, 1983) illustrates a striking violation of the rules of probability theory. On the other hand, game-based frameworks are questionable due to the lack of a unique concept of optimality in game-theory which could possibly support different "rational behaviors" for one and the same situation (just think of the plethora of different equilibrium concepts which have been derived from the original Nash equilibrium, cf. e.g. (Halpern, 2008)). Finally, heuristic approaches to judgment and reasoning (Gigerenzer, 2008) are often seen as approximations to a rational ideal and in some cases could work in practice, but mostly lack formal transparency and explanatory power. Also, from a methodological or philosophical point of view, severe reservations can be put forward: Due to the open nature of the collection of heuristics (the "heuristic toolbox") propagated in most current accounts, the possibility of falsification and refutation of modeling assumptions and theories is not guaranteed (as always another heuristics could be introduced, covering cases previously not accounted for), and a (reasonable) completion of the model can neither be checked for, nor guaranteed at any point.

### A More Cognitive Perspective

Already starting out with Simon's seminal work on "*A Behavioral Model of Rational Choice*" (Simon, 1955) more than half a century ago, conceptually different takes on rationality and rational behavior have been introduced into the discussion. Where logic-based, probability-based, and game-based accounts normally do not take into account limitations and cognitive properties of the reasoner (i.e., in our case, the human agent), the last decades have seen a growth of awareness for the importance and indispensability of these factors

in models of (human) rationality. Nonetheless, from our point of view, understanding the full meaning and implications of the commitment to a more "cognitively adequate" theory of rationality still is in an early stage, with researchers in different disciplines, both on the more theoretical and the empirical side, only starting to fully integrate these concepts into their accounts of rationality and rational behavior. In the following, we want to sketch some important developments and insights, preparing the ground for the subsequent presentation of our account of "subject-centered rationality".

### Theoretical Considerations

In (Simon, 1955), Simon articulates a simple but groundbreaking insight: (Human) agents are bounded in their resources and computational power. Once accepted, this proposition has far reaching consequences for the entire conceptual endeavor of formalizing and modeling rationality. All of a sudden, internal limitations of the reasoner, like limitations on working memory and computational power, but also external constraints, like limited time for decision-making, or incomplete and possibly also false information, have to be accounted for when creating a framework for rationality. Where this might not necessarily be a problem at first sight, it casts more than just a slight shadow of doubt on some of the fundamental assumptions underlying many "classical" accounts of rationality: It might be the case that reaching a conclusion via *modus ponens* in a logic-based model, once the preconditions are fulfilled, can computationally be realized rather simple - but what if the reasoner has to deal with incomplete information? If expectations have to be maximized in a probability-based model, where actually do the priors come from, once their existence cannot just be imported as given by the modeling assumptions? For many game-theoretic settings, even modern computers have a hard time finding equilibria or optimal solutions to not overly complex problems - how then can a mere human take the corresponding decision within a split second?

Although these issues clearly were identified as urgent questions, subsequent attempts at solutions mostly tried to solve the appearing problems within the original models, instead of putting the foundations to the proof and (possibly) having to construct entirely new models of rationality. Nonetheless, this has changed over the course of the last years. In a reply to Colman's article "*Cooperation, psychological game theory, and limitations of rationality in social interaction*" (Colman, 2003), Kokinov challenges traditional views on rationality (Kokinov, 2003). Taking an initial stance similar to Colman's, that is agreeing that rationality fails as both, descriptive theory of human decision-making and normative theory for good decision-making, Kokinov reaches a different, more radical conclusion than Colman did before. Instead of trying to fix the concept of rationality by redefining it, adding formerly unconsidered criteria for optimization of some kind, he proposes to replace the concept of rationality as a theory in its own right by a multilevel theory based on cognitive processes involved in decision-making. Where



Colman proposes a collection of ad-hoc strategies for explaining the deviations from rationality which people exhibit in their behavior, Kokinov proposes analogy as means of unifying the different, formerly unconnected parts of Colman's attempt at describing the mechanisms of decision-making. In Kokinov's view, the classical concept of utility making has to be rendered as an emergent property, which will emerge in most, but not all, cases, converting rationality itself into an emergent phenomenon, assigning rational rules the status of approximate explanations of human behavior.

Of course, this also defines a rather extreme position. But also scholars in "more standard" disciplines of rationality research, as for example decision theory, have become aware of fundamental problems with the traditional conception of rationality. In (Gilboa, 2010), Gilboa (in accordance with an earlier definition in (Gilboa & Schmeidler, 2001)) defines rationality as a subjective concept: "(...) *a mode of behavior is rational for a given decision maker if, when confronted with the analysis of her behavior, the decision maker does not wish to change it.*". This of course has far-reaching consequences. First of all, rationality becomes subject-centered, in that what is rational might vary with the population in question. But probably the most important implication is the dependence on the individual subject's abilities and limitations. If the decision maker does not understand the analysis, or why her behavior is not judged as rational, she cannot be judged irrational for not complying with the alleged norm of rationality. If limited cognitive capacities do not allow the reasoner to understand the rules he should follow in his reasoning (e.g. the Neumann-Morgenstern theory (von Neumann & Morgenstern, 1944)), but would always take the same decision again, he has to be called rational.

### Experimental Evidence

But also on the more empirical side, there is evidence galore that cognitive capacities and limitations have a clear influence on behavior and decision-making. Evidence for a crucial role of analogy as cognitive ability in decision-making can be found in psychological studies on decision-making and choice processes. An overview by Markman and Moreau (2001), based on experiments and observations from psychological studies (amongst others on consumer behavior and political decision-making), reaches the conclusion that there are at least two central ways how analogy-making influences choice processes. Analogies to other domains can provide means of representation for a choice situation, as generally speaking the making of a decision relies on a certain degree of familiarity with the choice setting. In many cases of this kind, analogy plays a crucial role in structuring the representation of the choice situation, and thus may strongly influence the outcome of a decision. Also, structural alignment (a key process of analogy-making) plays a role when comparing the different possible options offered by a decision situation, with new options being learned by comparison to already known ones. An experimental study by Kokinov (2005) demonstrated that people actually do use analogies in the process

of decision-making, with significant benefit already if only one case is found to be analogous to the choice situation under consideration. Furthermore, evidence has been found that there is no significant difference between close and remote analogies in this process, and that people are not limited to rely only on analogous cases from their own experience, but that also cases which were only witnessed passively (e.g., by being a bystander, or learning about a situation from reports in the media) may have beneficial influence.

Another example can be given in form of well-known studies on human decision-making under time pressure, which show a change in the applied inference procedure. In (Rieskamp & Hoffrage, 2008), the authors report that, whilst the best predicting model of human inference for decision making in an unstressed condition was a weighted linear model integrating all available information, when time pressure was induced, best predictions were obtained by using a simple lexicographic heuristic (Fishburn, 1974). When speculating about the precise way in which the induced pressure influences the reasoning process, the presumed change from a more complex strategy using complex relational structures to a simple single-attribute-based procedure, one possible explanation can again be found in research on analogy-making: In (Tohill & Holyoak, 2000), it is reported that anxiety made participants of an analogical-reasoning experiment switch from a preference for complex relational mappings to simple attribute-based mappings. So presupposing that analogy to already familiar situations serves as a basis for the decision-making, this reduction of the complexity of the mapping would be in line with the observed change in strategies.

### Cornerstones of "Subject-Centered Rationality"

Continuing and expanding thoughts already started for example in (Besold et al., 2011) and (Besold et al., 2012), in this section, we want to present some key features and cornerstones of a new account and understanding of rationality which we call "subject-centered rationality". What shall be presented is not yet another model of rationality and rational behavior, but an overall view and meta-conception of rationality, defining a supporting and limiting background and context for the construction of new models:

1. Rationality in a human context clearly has to be considered as a subject-centered notion, demanding for the integration of subject-related properties and constraints. This goes in line with Simon's and Gilboa's already discussed positions in that there is no use establishing models and norms which can never be implemented or fulfilled by human reasoners due to limitations of the agent or of its environment. Going back to the very beginnings, taking Aristotle's characterization of humans as rational animals as defining statement, a framework of rationality which cannot be applied at any moment by a human in everyday life has to be considered as limited in its usefulness and adequacy. This means that especially limiting constraints of human cognition, as for

example the computational boundedness of human agents and thus also the computational complexity of the rationality theory at hand, have to be accounted for.

2. The main aspect of theories and models of rationality has to be their use and applicability as positive theories, and not as mere normative or postdictive explanatory accounts. A valuable and adequate account of rationality has to provide a feasible prediction of human rational behavior and decision-making when being provided with the information the human reasoner can access at the moment of reasoning, that is, when possibly only having access to incomplete knowledge, when having to deal with ambiguities and possible false assessments of situations, etc.
3. A feasible theory of rationality does not have to be committed to one single formal modeling paradigm, but instead of being monolithic should pursue a holistic approach. Different formalisms and approaches to modeling should be unified and integrated into one account, providing the amalgamation of the different approaches with the union of their respective advantages and particular strengths, whilst mitigating each others deficiencies and weaknesses. Logic-based formalisms can be used alongside probability-based techniques alongside heuristic elements alongside models of cognitive core capacities alongside game-theoretic means for utility maximization. This integration of means and paradigms is governed by two guiding principles, the appropriateness with respect to the boundedness and the limited resources of the human reasoner, and the respect for and reproduction of human particularities in rational behavior and decision-making.

As should have become obvious from the above listing, we clearly consider rationality as a concept which is not connected to a particular formalism or theoretical modeling paradigm, but instead see it as possibly a plethora of different mechanisms competing, interacting and contributing to what we externally observe as a single capacity. Of course, this brings along challenges and questions, but from our point of view at the same time offers even more chances and opportunities. In the following, we want to give two examples where incorporating different (formal) paradigms into (distinct) traditional contexts has shown to be highly profitable:

With “algorithmic rationality”, Halpern and Pass (2011) proposed a framework including computational costs in otherwise game-theoretical notions of rationality, allowing to directly take into account the complexity dimension of agents’ boundedness when modeling rationality with game-theoretical means.

“Probabilistic dynamic epistemic logic” by Kooi (2003) presents an attempt at amalgamation between probabilistic and logical views of rationality, which can for example be used when treating with aspects of rationality at the intersection between epistemic logic and game theory.

Also, we strongly advocate the integration of factors and mechanisms describing the influence of cognitive capacities

and abilities into models of rationality, possibly offering entirely new perspectives and explanations for classical paradoxes of human rationality, as illustrated by the following examples:

Well-known empirical studies by Byrne (1989) question whether human reasoning can be covered by a classical logic-based framework. Presented with the information given in Table 1 and asked for what can be concluded from this, from 1. 46% of subjects conclude that Marian will not study late in the library, erring with respect to classical logic (as denial of the antecedent does not validate a negation of the consequent). Also, from 2. 96% of subjects conclude that Marian will study late in the library, whilst only 38% of subjects reach the same conclusion from 3.. Thus an introduction of another antecedent (without any indication that the antecedent should not hold) dramatically reduced the number of subjects applying a simple modus ponens in their process of forming a conclusion. Giving the task and the find-

Table 1: Inferences and Conditionals (Byrne, 1989)

- |   |
|---|
| <ol style="list-style-type: none"> <li>1. If Marian has an essay to write, she will study late in the library. She does not have an essay to write.</li> <li>2. If Marian has an essay to write, she will study late in the library. She has an essay to write.</li> <li>3. If Marian has an essay to write, she will study late in the library. She has an essay to write. If the library stays open, she will study late in the library.</li> </ol> |
|---|

ings a more cognitive capacities-oriented view than supported by the logic-based framework, the results concerning conclusions drawn by the subjects can for example be explained through analogy. People faced with the information given in 1. will recall similar conversations they had before, using these known situations as basis for their decision on what to conclude. According to Grice (Grice, 1975), in conversations, speakers are supposed to provide the hearer with as much information as is needed for exchanging the necessary information, a rule which also goes in accordance with our everyday observation. Thus, when being given the additional information that “Marian does not have to write an essay.”, the set of candidate situations for establishing an analogy to the present one will be biased towards situations in which this information had an impact on the outcome, resulting in the conclusion that Marian would not study late in the library either. Regarding 2. and 3., a similar conjecture seems likely to hold: By additionally mentioning the library, similar situations in which the library might actually have played a crucial role (e.g., by being closed) will be taken into account as possible base domains of the analogy, causing the change in conclusions made.

Already more than a decade ago, in the field of decision theory and economics Gilboa and Schmeidler (1995) devel-

oped an (at least partly) case-based theory and model for decision-making under uncertainty. In their model, cases are primitive and provide a simple axiomatization of a decision rule that selects an act to be performed based on the act's past performance in similar cases. Each act is evaluated by the sum of the utility levels that resulted from using this act in past cases, where the degree of (dis)similarity between the past cases and the problem at hand is accounted for by weighting the respective utility by the value of a similarity measure between both situations. Remarkably, this formal approach in a natural way gives rise to (amongst others) the notions of satisficing decisions and aspiration levels (cf. also (Gilboa & Schmeidler, 2001) for a detailed account).

### **Solutions for Problems of a Subject-Centered Notion of Rationality**

Of course, a positive subject-centered approach to rationality brings along quite some ground for reservations and skepticism. In the following, we want to address some of the most probable objections.

Doesn't your account collapse into pure subjectivity, making it not usable as basis for a general theory and framework anymore? No, it does not. The idea to overcome the problem of pure subjectivity is to identify central cognitive mechanisms, limitations and properties, common to all humans, and use those as basis for building up the theory. This is not a commitment to any particular modeling paradigm, but rather a strong statement assigning the positive aspects of the model (i.e. the adequate reflection of actual human properties) a higher priority than detail decisions for how to model a particular capacity, or overall methodological modeling consistency.

But then, how do you want to (theoretically) explain and (practically) preserve the individual aspects of your subject-centered modeling notion? In real life, each human comes equipped with different gifts and talents. Distinct cognitive capacities, although present in (almost) every human, are developed to a different extent. This also gives a cognitively-based model of rationality the possibility to account for different individual behaviors and decisions. Although there is an overall unified framework on a general human level, once reliable predictions shall be made on an individual scale, the weights and levels of developments between these different capacities will have to be assessed and adapted.

Your account does not provide any normative power, a key feature one would expect from a theory of rationality? To the contrary, normativity can be introduced on two levels, allowing for a distinction between a more general form of rationality, and a subjective one. Even more, contrary to some of the classical approaches, it is even possible to provide a quantitative account of performance on a normative scale. On a higher level, normativity can be introduced via the question "Given the general cognitive mechanisms and model, how well does the individual perform compared to the general optimal case?", that is, by assessing whether, tak-

ing into account goal-oriented behavior, the chosen way of acting resulted in the best outcome possible for an ideal representative of the species. If this is the case, the respective behavior or decision has to be considered as generally rational. If this is not the case, comparing the quality of the outcome to the performance of other individuals of the species can provide a quantification of the quality of the decision or behavior on a normative scale (e.g. "the subject performed at least as rational as 80% of his species"). On the individual level, normativity can be introduced via the question "Given the individual distribution of properties and limitations of the individual, how well does the subject perform compared to the individually optimal case?", that is, by assessing whether, taking into account goal-oriented behavior, the chosen way of acting resulted in the best outcome possible for the individual (a notion reminiscent of Gilboa's already aforementioned idea of rationality (Gilboa, 2010)). Also here, similar to the more general case, a quantitative aspect can be introduced to judging a behavior or decision rational: Provided that the model has accurately been fitted to the individual, reflecting its cognitive capacities, properties and limitations to a sufficient degree, it can be assessed to what extent the subject made use of its theoretical capacities.

### **Conclusion**

In the present paper, we give an account of basic principles and cornerstones of our positive conception of a theory and framework for rationality and rational behavior, envisioning a modeling paradigm which integrates different perspectives and approaches into a holistic system, giving rise to an integrated multiverse of rationality, replacing the multiple (mostly) mutually exclusive competing universes which there currently are.

From our point of view, our perspective on rationality offers several advantages not only within the field of cognitive science, but also for neighboring disciplines. A positive, predictively usable theory and framework for rationality (moreover if equipped with a quantitatively accessible notion of normativity) would allow for manifold applications, for example within decision theory and psychology (serving as an initial test bed for conjectures and research hypotheses), but also in more technical fields such as human-computer interaction (allowing for more natural and better adapting interfaces between man and machine) or artificial intelligence (greatly contributing to an overall model of human intelligence; cf. e.g. (Besold, 2011) for an AI-centric perspective).

Of course, there still are numerous open questions left for future investigation: How can the given cornerstones, characterizations and properties of "subject-centered rationality" be developed into a completely worked out meta-theory? Which are the key particularities, properties and limitations of human cognition that have to be integrated and accounted for by a subject-centered theory of rationality? How compatible are the already existing "classical" frameworks for rationality with our proposed view? How do our meta-level considera-

tions relate to conceptual work done in other relevant fields modeling (and possibly predicting) human behavior, and thus most likely dealing with similar questions? What would be a promising paradigm for an implementation: a highly modular bottom-up approach starting out by modeling one facet of rationality, consecutively adding more modules later, an entirely hybrid top-down approach, applying an amalgamated broad mixture of formalisms and a very general modeling paradigm from the very beginning, addressing different forms and facets of rationality by specialization within the overall framework, or something in between?

Although these are challenging and demanding questions, and a complete answer to any of those still is far from being visible, we are convinced that each single one of them is worth scientific effort and attention already by itself, in their totality moreover promising key insights into a core concept of human intelligence and cognition.

## References

- Besold, T. R. (2011). Rationality in/through/for AI. In J. Romportl, P. Ircing, E. Zackova, R. Schuster, & M. Polak (Eds.), *Proc. of Extended Abstracts of Beyond AI 2011*.
- Besold, T. R., Gust, H., Krumnack, U., Abdel-Fattah, A., Schmidt, M., & Kühnberger, K. (2011, July). An Argument for an Analogical Perspective on Rationality & Decision-Making. In J. van Eijck & R. Verbrugge (Eds.), *Proc. of the Workshop on Reasoning About Other Minds (RAOM-2011)*. CEUR-WS.org, Vol. 751.
- Besold, T. R., Gust, H., Krumnack, U., Schmidt, M., Abdel-Fattah, A., & Kühnberger, K.-U. (2012). Rationality Through Analogy - Towards a Positive Theory and Implementation of Human-Style Rationality. In *Proc. of MATHMOD 12 Vienna*. (to appear)
- Broadie, S., & Rowe, C. (Eds.). (2002). *Aristotle Nicomachean Ethics: Translation, Introduction, and Commentary*. Oxford University Press.
- Byrne, R. (1989). Suppressing valid inferences with conditionals. *Cognition*, 31(1), 61–83.
- Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences*, 26(2), 139–198.
- Cottingham, J., Stoothoff, R., & Murdoch, D. (Eds.). (1984). *The Philosophical Writings of Descartes* (Vol. II). Cambridge University Press.
- Evans, J. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, 128, 978–996.
- Fishburn, P. (1974). Lexicographical orders, utilities and decision rules: A survey. *Management Science*, 20, 1442–1471.
- Gigerenzer, G. (2008). *Rationality for Mortals: How People Cope with Uncertainty*. Oxford University Press.
- Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The Foundation of Adaptive Behavior*. Oxford University Press.
- Gilboa, I. (2010). Questions in Decision Theory. *Annual Reviews in Economics*, 2, 1–19.
- Gilboa, I., & Schmeidler, D. (1995). Case-Based Decision Theory. *The Quarterly Journal of Economics*, 110, 605–639.
- Gilboa, I., & Schmeidler, D. (2001). *A Theory of Case-Based Decisions*. Cambridge University Press.
- Grice, H. P. (1975). Logic and Conversations. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics, Vol. 3: Speech Acts* (pp. 41–58). Academic Press.
- Griffiths, T., Kemp, C., & Tenenbaum, J. (2008). Bayesian Models of Cognition. In R. Sun (Ed.), *The Cambridge Handbook of Computational Cognitive Modeling*. Cambridge University Press.
- Halpern, J. Y. (2008). Beyond Nash Equilibrium: Solution Concepts for the 21st Century. In *Proc. of the 27th Annual ACM Symposium on Principles of Distributed Computing*.
- Halpern, J. Y., & Pass, R. (2011, June). Algorithmic Rationality: Adding Cost of Computation to Game Theory. *ACM SIGecom Exchanges*, 10(2), 9–15.
- Kokinov, B. (2003). Analogy in decision-making, social interaction, and emergent rationality. *Behavioral and Brain Sciences*, 26(2), 167–169.
- Kokinov, B. (2005). Can a Single Episode or a Single Story Change our Willingness to Risk? The Role of Analogies in Decision-Making. In B. Kokinov (Ed.), *Advances in Cognitive Economics*. NBU Press.
- Kooi, B. P. (2003). Probabilistic Dynamic Epistemic Logic. *Journal of Logic, Language and Information*, 12, 381–408.
- Markman, A., & Moreau, C. (2001). Analogy and analogical comparison in choice. In D. Gentner, K. Holyoak, & B. Kokinov (Eds.), *The Analogical Mind: Perspectives from Cognitive Science* (pp. 363–399). MIT Press.
- Osborne, M., & Rubinstein, A. (1994). *A Course in Game Theory*. MIT Press.
- Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*, 127, 258–276.
- Simon, H. A. (1955, February). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1), 99–118.
- Tohill, J., & Holyoak, K. (2000). The impact of anxiety on analogical reasoning. *Thinking & Reasoning*, 6(1), 27–40.
- Tredennick, H. (Ed.). (1933–35). *Metaphysics*. Harvard University Press.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review*, 90(4), 293–315.
- von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.
- Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New Horizons in Psychology*. Penguin.

# The Role of Semantic Transparency in the Processing of Verb-particle Constructions by French-English Bilinguals

Mary-Jane Blais (mary-jane.blais@mail.mcgill.ca)

School of Communication Science and Disorders, McGill University, 1266 Pine Ave W  
Montreal, QC, H3G 1A8 CAN

Laura M. Gonnerman (laura.gonnerman@mcgill.ca)

School of Communication Science and Disorders, McGill University, 1266 Pine Ave W  
Montreal, QC, H3G 1A8 CAN

## Abstract

Verb-particle constructions (phrasal verbs) are a notoriously difficult aspect of English to acquire for second-language (L2) learners. This study was conducted to assess whether L2 English speakers would show sensitivity to the subtle semantic properties of these constructions, namely the gradations in semantic transparency of different verb-particle constructions (e.g., *finish up* vs. *chew out*). L1 French, L2 English bilingual participants completed an off-line (explicit) survey of similarity ratings, as well as an on-line (implicit) masked priming task. Bilinguals showed less agreement in their off-line ratings of semantic similarity, but their ratings were generally similar to those of monolinguals. On the masked priming task, the more proficient bilinguals showed a pattern of effects parallel to monolinguals, indicating similar sensitivity to semantic similarity at an implicit level. These findings suggest that the properties of verb-particle constructions can be both implicitly and explicitly grasped by L2 speakers whose L1 lacks phrasal verbs.

**Keywords:** Verb-particle constructions; bilingualism; semantic ratings; second language; masked priming.

## Introduction

Verb-particle constructions, also known as phrasal verbs<sup>1</sup>, are semantic units composed of a verb and a particle, which may be superficially similar to either a preposition (e.g., *turn out of the house*) or an adverb (e.g., *break the question down*). Common examples in English include *THROW OUT*, *LOOK UP*, *CHEW OUT*, *FINISH UP*, *PULL OVER*, and hundreds of others. These expressions are extremely common in some languages (e.g., English, German), though notably absent in others (e.g., French, Spanish, Italian). The language-specific properties of this phenomenon make it of interest to research in both monolingual and bilingual psycholinguistics. Current bilingualism research has demonstrated that non-native speakers have particular difficulty using these constructions, but has not yet identified the source of this difficulty. The present study was thus designed to investigate one aspect of verb-particle constructions that has been shown to affect monolinguals' processing: semantic transparency of the construction, which ranges from transparent (e.g., *finish up*) to opaque (e.g., *chew out*).

<sup>1</sup>While some authors prefer one or the other for various reasons, in this text the terms "phrasal verb" and "verb-particle construction" will be used interchangeably.

Semantic transparency was investigated using both an explicit and an implicit measure, to determine the level of processing where monolinguals and bilinguals differ.

## The Nature and Processing of Phrasal Verbs

Semantically, phrasal verbs are generally assumed to be stored as units in the lexicon, similarly to words or idioms (e.g., Jackendoff, 1995). That is, the meanings of such expressions are memorized holistically, separately from the meanings of the component words. There is much less consensus, however, as to whether these units are processed lexically in the same way as any other word, or whether syntactic processing is also necessary. Arguments based on traditional linguistic analysis have shed some light on this issue, but have been ultimately inconclusive. For example, phrasal verbs are amenable to processes of derivational morphology, changing from verbs into nouns in expressions such as "a show-off," "a fixer-upper" or "a passer-by" (e.g., Farrell, 2005). On the other hand, the verb and particle are clearly distinct units in the sentence that can be separated both by a noun phrase (e.g., *throw it out*) and by an adverb (e.g., *fixed it right up*). This type of insertion should not be possible within a single word, according to the so-called Lexical Integrity Principle (Chomsky, 1970); thus, in this sense verb-particle constructions behave similarly to syntactic phrases.

More recently, researchers have approached this question of whether verb-particle constructions are more phrase-like or word-like, using psycholinguistic and neuroimaging techniques. For example, Konopka and Bock (2009) showed that word order preferences for verb particles can be structurally primed; participants were more likely to remember a sentence as having an adjacent (or non-adjacent) verb and particle if they had just seen a different sentence with the same structure. This finding, which held regardless of the idiomaticity of the construction, was taken as evidence for more structurally-based accounts of phrasal verb processing. A different conclusion was drawn by Cappelle, Shtyrov and Pulvermuller (2010), who used magnetoencephalography (MEG) to record neural responses to verb-particle pairs that were congruent (e.g., *heat up*) or incongruent (e.g., *heat down*). The mismatch negativity responses to these pairs were comparable to responses patterns typically elicited by words, rather than sentences.

The authors concluded that at a neural level, phrasal verbs are processed lexically rather than syntactically. Thus, both linguistic and neuro-cognitive methods have yielded mixed results with regard to the nature of phrasal verb processing.

An alternative perspective holds that this strictly modular view of the lexicon versus the syntax creates a false dichotomy that fails to account for the behavior of verb-particles. For example, in an effort to conform them to these designations, many researchers have categorized phrasal verbs as either “transparent,” that is, interpretable based on knowledge of the component words, or “idiomatic,” having an opaque meaning that can only be memorized (e.g., Dagut & Laufer, 1985). However, it has recently been recognized that an entire spectrum exists between these two extremes. Gonnerman and Hayes (2005) asked native English speakers to rate, on a scale of 1-9, the degree of similarity between a verb-particle construction and its component verb alone (e.g., “How similar is *carry off* to *carry*?”). Their participants gave highly consistent ratings that were distributed through the entire range of the scale. For example, the pair *finish up/finish* was considered to be very similar while *chew out/chew* was rated as highly dissimilar. Other items, such as *look up/look* were generally rated around the middle of the scale.

In the same study, the authors tested participants’ implicit sensitivity to dependency using masked priming, an on-line task. Participants were asked to make a lexical decision to target words presented visually on a computer screen. Before each target, a prime consisting of another word or word combination was presented for 35ms, long enough to be processed subliminally but too short to be recognized consciously. Lexical decisions were facilitated when a target verb (e.g., *finish*) was primed by a low-dependency verb particle construction (e.g., *finish up*), but not when the target (e.g., *chew*) was primed by a high-dependency construction (e.g., *chew out*). Thus, these participants were shown to recognize dependency variations in both offline and on-line semantic processing.

### Processing in Second-language (L2) Learners

Phrasal verbs have long been recognized as among the most difficult aspects of English to acquire for second-language (L2) learners, and are also therefore of interest to those in the English as a Second Language (ESL) teaching profession (Neagu, 2007). Several researchers have investigated this phenomenon in bilinguals, though most of this work has focused on the avoidance of verb-particles in production. For example, Dagut and Laufer (1985) found that in written English tasks, native Hebrew speakers tended to avoid phrasal verbs (e.g., *let down*) in favor of single-verb synonyms (e.g., *disappoint*). While the authors attributed this effect to the lack of verb-particle constructions in Hebrew, subsequent studies have shown that similar difficulties are experienced by learners whose native languages include phrasal verbs, such as Dutch (Hulstijn & Marchena, 1989) and Swedish (Laufer & Eliasson, 1992). For these speakers, however, phrasal verbs seem to be more

easily acquired as a function of proficiency; advanced Dutch and Swedish learners display more native-like behaviour than either intermediates with the same L1s or advanced learners with L1 Hebrew. Thus, the difficulty of L2 English phrasal verbs appears to result from a compounding of factors that are both syntactic (inter-language differences) and semantic (inherent difficulty of acquiring idiomatic vocabulary). Later research (e.g., Liao & Fukuya, 2005; Gonzalez, 2010) has strengthened the hypothesis that avoidance of phrasal verbs decreases as English proficiency increases for all speakers, but that it does so more quickly for speakers with verb-particle constructions in their L1s.

Thus far, most investigations of verb-particles in L2 speakers have focused on production, particularly on the phenomenon of avoidance. However, it is equally important to investigate these structures at the level of receptive language processing. Comprehension of various linguistic structures precedes their production, both in first language (e.g., Benedict, 1977) and second language (e.g., Ringbom, 1992) acquisition, making this an important aspect of determining bilinguals’ competence with phrasal verbs.

In one of the few studies of phrasal verbs in on-line L2 comprehension, Matlock and Heredia (2002) measured the time it took for non-native speakers with various L1s to read English sentences involving the same phrase in either a verb+preposition context (e.g., *John ate up the street*) or a verb+particle context (e.g., *John ate up the pizza*). While native English speakers and early bilinguals (i.e., having acquired English before age 12) reacted more quickly to verb-particle constructions, late bilinguals seemed to process phrases involving a literal preposition most easily. This was taken as evidence that in processing figurative language, native speakers and early bilinguals can activate a figurative meaning instantly while late bilinguals must first retrieve the literal meaning before seeking alternate interpretations. While promising, however, this study had several limitations. First, the authors’ “on-line” measure was response time to an entire sentence, a relatively crude method which was unable to isolate the processing of the verb-particle construction itself. Moreover, first language and current proficiency level were not carefully controlled in this experiment.

A different, though related line of research is the study of idioms in second language comprehension. Like phrasal verbs, idioms consist of words that appear in other contexts but which take on a new meaning in a particular combination and a particular context. Given this similarity, it is not surprising that both types of constructions are difficult for second language learners. Models of monolingual idiom comprehension differ in the role they attribute to compositional versus non-compositional processes (see Titone & Connine, 1999, for a review); however, most current theories agree that native speakers may access either the literal or non-literal meaning of an idiom first depending on the construction itself as well as contextual and discourse factors (Giora, 2002; Titone & Connine, 1999). There is somewhat less consensus about

whether non-native speakers take full advantage of this complex processing strategy. One proposal (Cieslicka, 2006; Cieslicka & Heredia, 2011) is that the literal meanings of idioms enjoy universal salience for non-native speakers; that is, these speakers will always activate a literal interpretation before seeking an alternative reading. This “Literal Salience Hypothesis” is proposed to hold regardless of the context, familiarity, or decomposability of an idiom. However, not all researchers agree with this account (e.g., Bulut and Çelik-Yazici, 2004).

Thus, psycholinguistic studies suggest that both phrasal verbs and other types of non-literal language are processed in fundamentally different ways by native versus non-native speakers. However, there remains a significant need for more work describing the comprehension of L2 phrasal verbs. First, while work on idiom processing has made valuable contributions to this line of research, it must be recognized that full idioms, such as *kick the bucket* and *let the cat out of the bag*, differ from phrasal verbs in several important respects. While idioms constitute a large class of expressions with a great deal of variation in their syntax and flexibility, verb-particles pattern fairly regularly and behave much like literal verb-preposition combinations syntactically (Dixon, 1982). Some particles also behave more like morphemes in the sense that they can be applied productively; for example, the perfective UP can be applied to any verb that can be thought of as completive, yielding FINISH UP, WASH UP, GROW UP, ROLL UP, WRITE UP and many more. Thus, it might be expected that in interpreting verb-particles, as opposed to idioms, second-language learners would have additional sources of information (from regularities in the language) and may not rely so heavily on an initial literal interpretation.

Second, research on second language learning in general must distinguish between explicit and implicit language processes. The importance of dissociating these aspects of comprehension has been recognized at least as far back as Bialystok (1979), who found that while learners acquired both explicit and implicit knowledge of a new language, it was largely the explicit component that improved with increased instruction. This study also found that learners employed either their implicit or explicit knowledge depending on the processing demands of the task. More recently, Ellis (2005) emphasized the difference between these types of knowledge, which he defined using a variety of criteria including awareness, time available, attention, systematicity, certainty, metalinguistic knowledge, and learnability. This study found that explicit language ability was more strongly related to years of instruction, while implicit competence was correlated with age of acquisition.

Taken together, these results support the need to measure acquisition of a particular structure both implicitly and explicitly, an approach we have taken in the present study. The following experiments were conducted to test whether native speakers of French, a language that lacks verb-particle constructions, are sensitive to the same semantic variations recognized by native speakers.

## Verb-particle Similarity Ratings

To measure bilinguals’ sensitivity to the semantic transparency of verb-particle constructions, we used an explicit, off-line, similarity rating task. Past research (Gonnerman & Hayes, 2005) has shown that when asked to rate the similarity between verbs and their corresponding verb-particle constructions, native English speakers provide consistent ratings across a spectrum ranging from low (*chew out/chew*) to mid (*look up/look*) to high (*chew out/chew*) similarity. To determine whether L2 speakers are sensitive to this variability, we administered a similar survey to French dominant English bilinguals. This metalinguistic task was designed to measure participants’ explicit knowledge of verb-particle semantics, which we predicted would be similar to, but less accurate than that of monolinguals.

### Participants

34 adult (age 18-40) native speakers of Canadian French were recruited through web-based advertisements on a university research mailing list, and participated voluntarily. English proficiency was self-rated as either Beginner (n=1), Intermediate (n=9), Advanced (n=20) or Near-native (n=4). Participants also reported their age of first exposure to English, which ranged from 1 to 20 years, with a mean of 8.21 years.

### Materials

78 verb-particle pairs were presented in an internet-based survey. Stimuli were selected from a larger set of 212 verb-particle constructions that were rated by monolinguals in Gonnerman & Hayes’ (2005) study, and contained an even distribution of low (mean rating < 4), medium (4-6) and high (>6) similarity items as rated by the monolinguals. Particles (e.g., *up*, *on*, *off*) were evenly distributed among high, medium and low similarity items. In addition, items in each group were matched for the frequency (Kucera & Francis, 1967) of the verbs (e.g., *throw*), as well as for the frequency of verb-particle constructions in their entirety (e.g., *throw up*).

### Procedure

Each participant rated all 78 items. Participants were asked to rate the similarity in meaning of verb particle/verb pairs on a scale from 1 (very dissimilar) to 9 (very similar). Instructions for this task included examples of highly similar as well as dissimilar pairs with corresponding ratings. Ratings were compiled electronically and analyzed for comparison with the ratings obtained from monolinguals by Gonnerman & Hayes (2005).

### Results & Discussion

Similarity ratings of the 78 items from monolinguals and bilinguals are shown in Figure 1. Results are arranged in ascending order of the monolinguals’ ratings. Monolingual and bilingual ratings are positively correlated with



correlation coefficient 0.707 ( $p < .01$ ), indicating that bilinguals can make similar judgments of semantic similarity to native English speakers. Ratings from the bilinguals were fairly evenly distributed across the range of the scale; on average, participants chose each point on the scale between 6 and 10 times. Interestingly, ratings of the two groups agreed more consistently in the middle of the scale than at either end, with those of the lowest-similarity items being most discrepant.

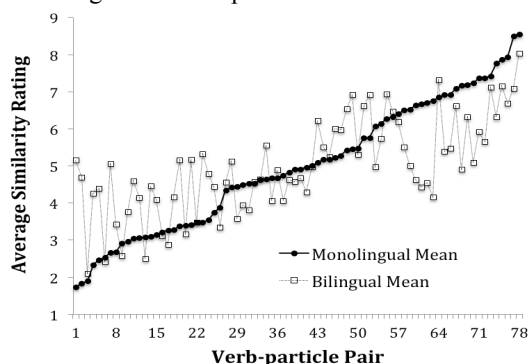


Figure 1: Mean semantic similarity ratings obtained from French-English bilinguals and English monolinguals from Gonnerman & Hayes' (2005). Verb/verb-particle pairs are arranged in ascending order of the monolinguals' mean ratings.

In other respects, bilinguals' ratings differed from those of the monolinguals. Bilingual speakers' ratings were significantly less consistent than monolinguals, with an average standard deviation of 2.33, as opposed to 1.96 for the monolinguals ( $F(1,77)=56.16$ ,  $p < .001$ ). Bilinguals' ratings also showed a reduced range (5.94), as compared to the monolinguals' range of 6.82. These results indicate that while second language speakers do recognize a range of semantic transparency across verb-particle constructions, they are generally more variable in their responses.

### Masked Priming

As an on-line measure of semantic processing, participants completed a masked priming task in which a target verb (e.g., LOOK) was primed by its corresponding verb-particle construction (e.g., *look up*). In past research (Gonnerman & Hayes, 2005), priming has been found to be strongest for verb-particle constructions rated as highly similar in meaning to their isolated verbs. This task was designed to determine whether bilinguals' implicit processing of verb-particle constructions would be predicted by the degree of semantic transparency, as has been shown for monolinguals. In addition, the task serves as an implicit comparison to the explicit data obtained from the ratings task. If applicable to verb-particle constructions, the Literal Salience Hypothesis (Cieslicka, 2006; Cieslicka & Heredia, 2011), would predict poor performance on the priming task; Literal Salience holds that non-literal language is always first interpreted literally and only then re-analyzed, a

process which would not have time to occur in a masked priming paradigm.

### Participants

30 native speakers of French, aged 18 to 35, participated for monetary compensation. Inclusion criteria were identical to those of the similarity rating experiment: participants were required to consider themselves non-native speakers of English but to have functional proficiency in English. Age of first exposure to English ranged from 1 to 15 years, with a mean of 7.79 years. English proficiency was self-reported as Intermediate ( $n=7$ ), Advanced ( $n=16$ ) or Near-native ( $n=5$ ).

### Materials

The same 78 verb-particle constructions were used as related primes for their corresponding verbs (e.g., *cover up/cover*). For each construction, an unrelated control prime was created to match in frequency and number of letters (e.g., *show off/cover*). Control primes did not overlap with test primes in meaning or orthography. Finally, identity primes (e.g., *cover/cover*) were included for each item. Stimuli were divided into three lists, with one of these conditions in each list so that no participant responded to any verb more than once. In order to reduce the proportion of related prime-target pairs, 78 real word prime-target filler items were added to each list. In addition, 156 non-word filler items were included, matching the real words in frequency and orthography as closely as possible. Of these, half employed verb particle primes with non-words that were either "related" (e.g., *keep out/keem*) or "unrelated" (e.g., *live down/bool*), while the other half used single words as primes. Thus, each participant responded to 312 items, of which 39 were related prime-target pairs containing verb particle constructions.

### Procedure

Participants were tested individually in a quiet room with dim, natural lighting. Stimuli were presented using PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993) software on CRT monitors running at 85 HZ. Each trial consisted of a fixation point (\*) displayed for 1000ms, after which a mask (%#@!&^\$) was displayed for 500ms; subsequently, the prime appeared briefly for 35ms followed immediately by the target, which remained on the screen for 200ms. Participants made a lexical decision to the target by pressing the yes/no buttons on a button box, from which reaction times were recorded. After the participant's response, a 500ms delay occurred before presentation of the next trial. Stimuli appeared in white on a black background, with primes in lower case letters and targets in upper case letters.

### Results & Discussion

Four participants, who made errors on more than 40% of the items, were excluded from the analyses. For all other participants, only correct responses were included in the

analyses. Data were trimmed to exclude outliers; that is, response times slower than 300ms or faster than 1000ms. A 3 (Prime Type: Related vs. Unrelated vs. Identity) by 3 (Prime-target similarity: Low vs. Mid vs. High) repeated measures ANOVA was conducted to determine whether priming effects were modulated by semantic similarity. Because we were interested in priming effects specifically, we also planned comparisons between the unrelated and related response times. Priming effect for the monolinguals (from Gonnerman & Hayes, 2005) and bilinguals are shown in Tables 1 and 2 below. An identity condition was also included for the bilinguals to rule out the possibility that bilinguals are only reading the first word in the verb-particle primes, that is, reading only the first element (*e.g.*, *throw*) and ignoring the particle separated by a space (*e.g.* *up*). Including the identity condition therefore allowed us to determine whether priming for related targets reflects the whole prime, since otherwise *throw off/throw* would simply elicit the same identity priming as *throw/throw*).

Results showed a significant main effect of Prime Type ( $F(2, 50)=7.96, p<.01$ ) and a significant main effect of Similarity ( $F(2, 50)=7.19, p<.01$ ). The interaction of these factors was non-significant. Planned comparisons revealed significant ( $p<.05$ ) differences between the unrelated and identity primes across all conditions, and significant differences between the unrelated and related primes in the mid and high-similarity conditions (see Table 2 below).

Table 1: Monolinguals' response latencies for target words by prime type and similarity (from Gonnerman & Hayes, 2005).

Monolinguals	Prime-target Similarity		
	Low	Mid	High
Prime Type			
Unrelated ( <i>cast off/throw</i> )	550	553	557
Related ( <i>throw up/throw</i> )	543	532	537
<b>Unrelated-Related</b>	<b>7</b>	<b>21*</b>	<b>20*</b>

Table 2: Bilinguals' response latencies for target words by prime type and similarity.

Bilinguals	Prime-target Similarity		
	Low	Mid	High
Prime Type			
Unrelated ( <i>cast off/throw</i> )	605	619	605
Related ( <i>throw up/throw</i> )	592	599	576
Identity ( <i>throw/throw</i> )	583	595	569
Unrelated-Identity	22*	24*	36*
<b>Unrelated-Related</b>	<b>13</b>	<b>20*</b>	<b>24*</b>

Tables 1 and 2 show the response latencies from monolingual and bilingual participants to unrelated, related and (for the bilinguals) identity primes. Responses from the bilinguals were slower overall, consistent with the increased processing cost of responding in one's second language. In

all other respects, however, results from the two groups are strikingly similar. In addition, for the bilinguals identity priming across all three conditions was higher than Unrelated-Related priming, indicating that the bilinguals did in fact respond differently to the verb-particle constructions than to the verbs alone. As did the monolinguals, the bilingual speakers showed no priming effect for low similarity items, but significant facilitation from verb-particles rated as having mid or high similarity to the target verbs. These results suggest that, contrary to our expectations, at an implicit level L2 speakers are sensitive to the same gradations in semantic transparency that are reflected in monolingual priming effects.

## General Discussion

The present study was designed to investigate the performance of non-native English speakers on implicit and explicit measures of phrasal verb comprehension. Based on past research, we hypothesized that the bilinguals would have difficulty with both tasks, showing decreased sensitivity to the variations in verb/verb-particle similarity that are easily recognized by monolinguals.

Somewhat surprisingly, responses of the L2 speakers approximated those of monolinguals on both the explicit and implicit semantic tasks. This native-like behaviour supports the findings of past research (*e.g.*, Laufer & Eliason, 1992; Liao & Fukuya, 2005) demonstrating that non-native speakers can improve their competence with verb-particle constructions regardless of L1. Importantly, it also extends this literature from production to comprehension, suggesting that use of these constructions reflects their mastery even at a subconscious level. Nevertheless, it should be noted that bilingual responses were not identical to those of monolinguals, especially in the variability between participants on the similarity rating task. More research is needed to determine whether this reflects a fundamental difference between monolinguals and bilinguals, or whether even this effect might disappear in high-proficiency L2 speakers.

The results from the masked priming experiment do not support an extension of the Literal Salience Hypothesis (Cieslicka, 2006; Cieslicka & Heredia, 2011) to verb-particle construction processing in L2. Being below the consciousness threshold, the presentation length of the primes in this experiment was considered brief enough to measure initial, automatic interpretation. Thus, if bilingual speakers universally activated the literal meaning of a verb without considering it in conjunction with a particle, then identical priming would be expected for verb-particle constructions and identity primes across conditions. In contrast, our participants showed consistently higher priming for identity primes than for related verb-particle primes. Additionally, the difference between high and low/mid similarity items can only be explained if participants were responding to the construction as a unit and not simply to the literal combination of words. These data suggest that the literal salience account of idiom

processing in bilinguals does not apply to processing of verb-particle constructions.

When comparing the present study to past research, it should be noted that the bilingual participants in this study had a somewhat different language experience than those in most previous studies of phrasal verb acquisition (e.g., Dagut & Laufer, 1985; Hulstijn & Marchena, 1989; Laufer & Eliason, 1992). While past research has largely focused on speakers learning English in a formal or foreign-language setting, our participants were inhabitants of Montreal, where both French and English are regularly used in formal/educational as well as informal contexts. Thus, although context of exposure was not explicitly controlled in our study, it is reasonable to expect that most of our participants had (either currently or at some point in the past) some degree of contact with and use of English in everyday speaking situations. The present study therefore offers an important extension of work on L2 phrasal verbs to a bilingual population more apt to use English in informal as well as formal contexts.

Several possible directions for future research are suggested by the present study. Gonnerman & Hayes (2005), have noted that variations in verb-particle similarity can influence speakers' word-order preferences, for instance, deciding whether a verb and particle should appear in an adjacent (e.g., *throw out the garbage*) or shifted (e.g., *throw the garbage out*) construction. A logical extension of this experiment would be to investigate whether bilinguals use semantic similarity to influence their word-order preferences. In addition, past work on avoidance of verb-particle constructions in bilinguals suggests a need for more careful comparison of bilinguals with different proficiency levels and ages of acquisition. Finally, evidence from both bilingual and monolingual processing must ultimately be integrated with theoretical models of cognitive/linguistic function, addressing such issues as the interface between the lexical and semantic systems.

### Acknowledgments

This research was supported by two graduate research awards to the first author, from the Natural Sciences and Engineering Research Council of Canada (NSERC), and from the Fonds québécois de recherche sur la nature et les technologies (FQRNT).

### References

Bialystok, E. (1979). Explicit and implicit judgements of L2 grammaticality. *Language Learning*, 29, 81-103.  
 Bulut, T. & Çelik-Yazici, I. (2004). Idiom processing in L2: Through rose-colored glasses. *The Reading Matrix*, 4, 105-116.  
 Cappelle, B., Shtyrov, Y., & Pulvermüller, F. (2009). Heating up or cooling up the brain? MEG evidence that phrasal verbs are lexical units. *Brain & Language*, 115, 189-201.

Cieslicka, A.B. (2006) Literal salience in on-line processing of idiomatic expressions by second-language learners. *Second Language Research*, 22, 115-144.  
 Cieslicka, A.B., Heredia, R.R. (2011). Hemispheric asymmetries in processing L1 and L2 idioms: Effects of salience and context. *Brain & Language*, 116, 136-150.  
 Chomsky, N. (1970). Remarks on nominalization. In R. A. Jacobs & P. S. Rosenbaum (Eds.), *Readings in English transformational grammar* (pp. 184-221). Waltham, MA.: Ginn.  
 Cohen, J. D., MacWhinney, B., Flatt, M. & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, and Computers*, 25, 257-271.  
 Dagaut, M. & Laufer, B. (1989). Avoidance of phrasal verbs: A case for contrastive analysis. *Studies in Second Language Acquisition*, 7, 73-79.  
 Dixon, R.M.W. (1982). The grammar of English phrasal verbs. *Australian Journal of Linguistics*, 2, 1-42.  
 Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. *Studies in Second Language Acquisition*, 27, 141-172.  
 Farrell, P. (2005). English verb-preposition constructions: Constituency and order. *Language*, 80(1), 96-137.  
 Giora, R. (2002) Literal vs. figurative language: different or equal? *Journal of Pragmatics*, 34, 487-506.  
 Gonnerman, L.M., & Hayes, C.R. (2005). The professor chewed the students... out: Effects of dependency, length, and adjacency on word order preferences in sentences with verb particle constructions. In *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*. (pp. 785-790). Mahwah, NJ: Erlbaum.  
 Jackendoff, R. (1995). The boundaries of the lexicon. In M. Everaert, E. van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 133-165). Hillsdale, NJ: Erlbaum.  
 Konopka, A.E. & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology*, 58, 68-101.  
 Laufer, B. & Eliasson, S. (1992). What causes avoidance in L2 learning? *Studies in Second Language Acquisition*, 15, 35-48.  
 Liao, Y. & Fukuya, Y.J. (2004). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning*, 52, 193-226.  
 Neagu, M. (2007). English verb particles and their acquisition: a cognitive approach. *Revista Espanola de Lingüística Aplicada*, 20, 121-138.  
 Titone, D.A. & Connine, C.M.. (1999) On the compositional and non-compositional nature of idiomatic expressions. *Journal of pragmatics*, 31, 1655-1674.

# Perception of Ambiguous Drawings and the Construction and Inhibition of its Alternative Interpretation – Reflections on Consciousness

Svetoslav Bliznashki (valsotevs@gmail.com)

Mariya Popova (popova.mariya@gmail.com)

Boicho Kokinov (bkokinov@nbu.com)

Central and East European Center for Cognitive Science, Department of Cognitive Science and Psychology, New Bulgarian University 21, Montevideo Str., Sofia 1618Bulgaria

## Abstract

Based on some of our previous experimental findings concerning the negative priming which occasionally occurs during perception of ambiguous drawings we propose a simple mathematical model which we believe may account for important aspects of our data. We tested some (but not all) of the crucial predictions of our model using a newly developed set of asymmetric ambiguous drawings. The results further supported the phenomenon of negative priming taking place during perception of ambiguous drawings and were consistent with the predictions of the model.

**Keywords:** ambiguous drawings; negative priming; vector reflection.

## Introduction

Early gestalt psychologists pointed out that human beings are incapable of perceiving two interpretations of an ambiguous drawing simultaneously. Only one of the interpretations seems to be consciously available at any given moment. Numerous studies show that the conscious experience of a particular interpretation of an ambiguous drawing is subject to context effects (e.g. Balceits & Dunning, 2006; Fisher, 1968; Long & Olszweski, 1999, 2004; Rock & Mitchener, 1992; Kokinov et al., 2007, 2009), i.e. depending on the context we can experience either of the two interpretations. These context effects raise the question whether both interpretations are built in parallel but only the stronger is experienced, or depending on the context only the relevant interpretation is built. This question is analogous to the question asked by Swinney (1979) whether both meanings of a homophone are accessed when the word is encountered or only the meaning relevant to the context. The answer provided by Swinney is that both meanings are accessed. His explanation is in terms of an autonomous lexical access device that does not interact with the context. However, in the case of perception of a picture it is impossible to think in terms of access to pre-stored meanings of the picture, it is clear that the interpretations have to be constructed on the fly. Thus if it turns out that the alternative interpretation is also primed, that would mean that both interpretations are constructed in parallel. Such results would potentially cause a re-interpretation of Swinney's results since that would mean that this is not a specific linguistic phenomenon that could be explained by the specific organization of the language system and more general explanations should be sought for.

Kokinov, Biznashki, Kosev, and Hristova (2007) were interested in how analogy could cause re-representation and re-interpretation of an ambiguous picture. In a further extension of Experiment 1, Bliznashki and Kokinov decided to use additionally Swinney's methodology (a post-test Lexical Decision Task, called LDT for short) in order to evaluate whether the unseen interpretation will be primed. If this was the case, that would be evidence in favor of the idea that parallel processes are constructing both interpretations. The results of this experiment confirmed the hypothesis of parallel representation building of both interpretations, however, they were rather surprising since we obtained negative priming, i.e. we obtained evidence that the unseen interpretation was inhibited (Kokinov, Vankov, Bliznashki, 2009). This was in accordance with some simulation data with the AMBR model which also exhibited inhibition of the corresponding concepts.

We were happy with these results, but there was a puzzle that remained unresolved and which we found very challenging. In this experiment we used an asymmetric drawing with one interpretation being easy to perceive and the other one being hard to notice. We also varied the context in which the drawing was provided by pushing the participants in the study to perceive either the easy or the hard interpretation in order to solve a complex analogical task. We were happy to prove that the relational structure of the task exerted strong influence upon subjects' perception (which was the main goal of our studies) but we were puzzled to find out that the negative priming was not always present. More specifically the negative priming occurred *if and only if* subjects were contextually prompted to perceive the weak (hard-to-see) interpretation of the drawing. In that case the strong (easy-to-see) interpretation showed pronounced negative priming during a Lexical Decision Task which followed immediately after the presentation of the ambiguous drawing, while if the participants saw the easy to see interpretation, there was no negative priming of the alternative interpretation. In our current study we try to explain that finding in formal terms as well as to replicate the results in an experiment specifically designed to detect priming of an unperceived interpretation of ambiguous drawings (which was not the case with our previous studies which focused primarily on exploring the parallel nature of different processes which took place during analogical reasoning).

## Mathematical Model and Predictions

Modeling our results was a challenging enterprise. To see why consider the simple competitive learning network (e.g. Knight 1990) shown in figure 1.

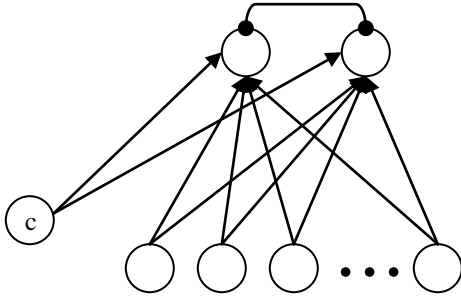


Figure 1. A standard competitive learning network; the two output units are inhibiting each other; all other connections are positive; the node “c” represents external context.

Let’s suppose such a network is trained to classify two sets of drawings – one set consisting of drawings of rabbits and another consisting of drawings of dogs. Now suppose that after a successful learning phase the network is presented with a linear combination of two exemplars with the weights of the combination reflecting the degree to which the elements of the input vector retain their similarity to the original exemplars. In most cases the network will “see” the easier interpretation and (given a relevant activation function) will inhibit the harder interpretation. How can we make the network “see” the harder interpretation? One way we might achieve that is by allowing a context node to pour activation towards the relevant output node. This would be analogous to our previous experiments during which we pressed subjects to see the harder interpretation of an ambiguous drawing by placing the drawing in a particular context. As long as the inhibitory connection/s between the output nodes remain constant and uninfluenced by the inputs to the network, however, all our manipulations will result in the more relevant interpretation winning the competition and the inhibition of the alternative. This contradicts with our data since as already explained negative priming seemed to be present only for the strong interpretation when the weak one was supported by context. Because of that we tried to explain our data by a model which will not rely on constraint satisfaction.

### Simple Vector Reflection Achieving Conscious Appraisal (SVRACA): conceptual description.

In SVRACA we regard concepts as orthogonal vectors residing in a high dimensional space. Each newly perceived entity is represented as a linear combination of the already existing concepts in memory. The vectors representing concepts are considered analogous to Long Term Memory (LTM) while vectors representing newly perceived entities are supposed to be analogous to the content of Working Memory (WM). SVRACA makes a clear distinction

between conscious and subliminal perception. Subliminal perception is modeled by simply constructing a vector in WM as a linear combination of concepts/percepts in LTM. Conscious perception is modeled by the categorization of the constructed vector as an example of one of the existing concepts in LTM. The vector from LTM which shows the highest positive correlation with the vector in WM is supposed to represent its designated category. In cases of ambiguity the newly constructed vector in WM exhibits non negligible positive correlation with more than one vectors residing in LTM. In such cases SVRACA *reflects* the WM vector along all axes (LTM vectors) which the WM vector correlates positively with except for one. The one exception is the single axis which the WM vector still correlates positively with after the reflection operation. All other axes correlate negatively or negligibly with the WM vector. In that state the system is said to have resolved the ambiguity. In other words *conscious perception of an entity is said to occur when a WM vector is fully constructed and when it is roughly orthogonal or negatively correlated with all vectors in LTM except for one with which it exhibits a relatively high correlation*. Thus in SVRACA vector reflection is considered a metaphor for consciously perceiving only one interpretation of an ambiguous stimulus at any given time. In SVRACA the main operation – reflection is executed in a probabilistic fashion (i.e. a stochastic process decides around which axes the WM vector is reflected and which axis remains constant) and in determining the probabilities of reflection both the internal and external context play crucial role. Internal context (or perceptual context) is represented by the weights of the linear combination which constitutes the WM vector. The larger a weight the more probable the WM vector will *not* be reflected along the axis the weight is associated with. External context is represented by a vector of probabilities with each probability reflecting the relative strength with which environment pushes towards a particular interpretation. The interaction between internal and external context probabilistically determines which interpretation of an ambiguous stimulus is perceived. Figure 2 illustrates how SVRACA works and some of its predictions:

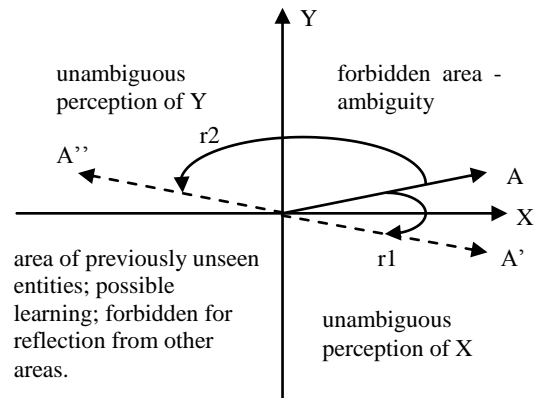


Figure 2. LTM vectors are X and Y, WM vector is A. A is ambiguous since it correlates positively (makes sharp angles) with both X and Y.

SVRACA will reflect A either around the X axis (this reflection is indicated as r1 in the figure) or around the Y axis (r2). In the first case the result is A' – a vector unambiguously perceived as X while in the second the result is A'' – a vector unambiguously perceived as Y. Which reflection takes place – r1 or r2 is determined by taking into account both perceptual and external context. Ignoring external context for the moment we see that A is much closer to X and consequently it will be much more probable that r1 occurs. If that is the case A' will result and X but not Y will be considered to have been consciously perceived. A' now is practically orthogonal to Y (a very small and possibly undetectable negative correlation exists between A' and Y). Thus the system has perceived the strong interpretation (X) of the ambiguous stimulus (A) and the alternative interpretation (Y) is neither primed nor suppressed. In the other scenario r2 occurs (because of external context or by pure chance) and A'' is perceived unambiguously as Y. Here however the unperceived interpretation (X) correlates very strongly and negatively with A'', i.e. SVRACA predicts that in this case (perceiving the weak interpretation for whatever reason) the unperceived interpretation will be negatively primed. That is exactly what we have observed in our previous studies. It may seem quite arbitrary that we chose vector reflection as the main operation which deals with ambiguity so let's state our reasons at the very beginning. First of all reflection is a computationally cheap operation. Second it makes a direct use of a meaningful representation of ambiguous stimuli. If an ambiguous stimulus is represented as a linear combination of unambiguous stimuli then reflection is probably the cheapest way to attain an unambiguous representation while maintaining the base representational scheme. Third, reflection allows us to model our stimuli as vectors in a high dimensional space which in turn opens the door for exploiting many of the advantages of using highly distributed representations. Fourth, vector reflection may serve as a useful and mathematically sound metaphor for conscious experience. Another such metaphor is reaching a stable state in a recurrent neural network but as we already saw at least some neural architectures don't provide a straightforward way of explaining the selective negative priming which is the topic of our study. Moreover it is not at all inconceivable representing SVRACA as a complex recurrent neural net in the future. Last but not least a model based on vector reflection not only makes sense of currently available data but also makes useful and testable predictions which will be discussed later.

### SVRACA: semi formal description.

Currently SVRACA supports 10 LTM vectors residing in a 10000 dimensional space. LTM vectors are orthogonal to each other and can be said to form a basis of a 10 dimensional space. Each vector is standardized to have a mean of 0 and a standard deviation of 1. Thus all vectors can be conceived as being of unit length. Here we will discuss the simplest case of ambiguity where a WM vector

is a linear combination of only two LTM vectors (e.g. an ambiguous drawing which can be interpreted as a rabbit or as a dog). A simulation begins with the system being supplied with a vector of coefficients  $\mathbf{w}$  which serve as weights for the linear combination. In the simplest case the system is supplied with two coefficients. The sums of squares of the coefficients must be equal to 1 ( $\|\mathbf{w}\|=1$ ) and thus the length of the linear combination is also equal to 1. The system proceeds by forming a linear combination of as many vectors from LTM as there are coefficients in  $\mathbf{w}$ . This construction stage reflects encoding a stimulus by the perceptual system and it requires a certain amount of time. Our simulations involve applying  $\mathbf{w}$  to randomly chosen vectors from LTM since the next stage involves SVRACA determining the relationship of the WM vector to each of the LTM vectors. SVRACA performs a linear regression on WM and collects all coefficients larger than some threshold value (e.g. 0.01). The vectors in LTM corresponding to these coefficients are those which make a significant contribution to the linear combination in WM. The larger a coefficient the closer the WM vector to the LTM vector associated with that coefficient. If all coefficients in  $\mathbf{w}$  are close to 0 (i.e. below a threshold of 0.1) or negative, SVRACA identifies the WM vector as a previously inexperienced entity and the simulation stops (see the lower left quadrant in figure 2). If all entries in  $\mathbf{w}$  are close to 0 or negative except for one than SVRACA identifies the WM vector as belonging to the category in LTM associated with the only positive coefficient (see the lower right and the upper left quadrants in figure 2). If more than one entry in  $\mathbf{w}$  is positive SVRACA interprets the situation as ambiguous and tries to resolve the ambiguity by reflection (see the upper right quadrant in figure 2). In the absence of any external context SVRACA reflects the WM vector around all axes associated with positive coefficients except for one. The probability of any axis remaining unaffected by the reflection operation is proportional to its squared coefficient. For example the probability of WM being reflected around all axes but the third is equal to  $w(3)^2$ . External context is represented by a set of weights each ranging from 0 to infinity. The elements of the vector of context weights, denoted  $\mathbf{c}$ , are subject to only one constraint namely that each of them should be 0 or positive. When external context is present the probability  $\mathbf{p}$  of each axis being the only one unaffected by reflection is equal to:

$$\mathbf{p} = ((\mathbf{w} * \mathbf{c} + \mathbf{w})^2) / \text{sum}((\mathbf{w} * \mathbf{c} + \mathbf{w})^2) \quad (1)$$

where “.” designates element-wise operations,  $\mathbf{p}$  is the vector of probabilities that each axis is the only one unaffected by reflection,  $\mathbf{w}$  is the vector containing the weights determining the WM vector and  $\mathbf{c}$  is the vector containing the weights representing the strength of external context towards each LTM axis. After a single axis is chosen to be unaffected by reflection SVRACA reflects the WM vector around all other axis with positive weights. This is achieved by setting to 0 all elements in  $\mathbf{w}$  which were initially negative as well as the element corresponding to the unaffected axis. This newly formed vector is called

reflection vector and is denoted by  $\mathbf{wr}$ . Now all positive elements in  $\mathbf{wr}$  refer to axes WM is to be reflected around and all other elements are 0. If we denote the collection of vectors in LTM which are involved in the linear combination WM as  $\mathbf{V}$  ( $\mathbf{V}$  is now a matrix with dimensions  $10000 \times \text{length}(\mathbf{w})$ ) the reflection WMr is achieved by:

$$\mathbf{WMr} = \mathbf{WM} + \mathbf{V} * (-2 * \mathbf{wr}) \quad (2)$$

where WMr is the reflected version of WM,  $\mathbf{V}$  is the collection of all vectors in LTM originally participating in the formation of WM and  $\mathbf{wr}$  is the reflection vector. The final step in a SVRACA simulation is to examine the relationship of WMr to the vectors in LTM in order to decide which LTM concept is the only positively related one to WMr (i.e. which concept WMr is perceived as). The relationship between any two concepts in SVRACA (LTM, WM, WMr) is defined as the cosine of the angle between the vectors in question. Since all vectors are defined as being of unit length the cosine of the angle between two vectors is also equal to the product moment correlation between them. Thus after reflection WMr will in most cases be positively correlated with only a single vector in LTM, negatively correlated with all other vectors in LTM which participated in the formation of WM and orthogonal to all vectors in LTM which were not involved in the construction of WM in the first place. The signs and magnitudes of these correlation coefficients are supposed to predict priming (positive or negative) in real world situations.

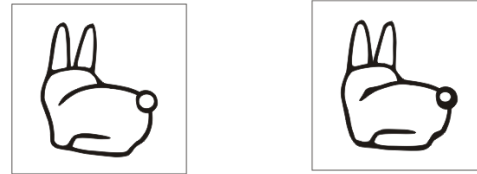
**A Typical Simulation** is described next. We supply SVRACA with a highly asymmetrical concept which is a linear combination of two LTM vectors with coefficients  $\mathbf{w} = [0.95 \ 0.3122]$ . We see that the WM vector is much closer to the LTM1 vector than to LTM2 vector. The context vector, however reflects the opposite picture:  $\mathbf{c} = [0 \ 5]$ , i.e., the LTM1 vector is absolutely irrelevant to the situation while the LTM2 would be very useful. Since both numbers in  $\mathbf{w}$  are positive SVRACA recognizes the ambiguity (i.e. identifies the LTM vectors correlating positively with WM) and calculates the probabilities  $\mathbf{p}$  of consciously perceiving each of the two possible interpretations according to (1):  $\mathbf{p} = [0.2045 \ 0.7955]$ . We see that the contextual influence radically changes the odds towards a conscious perception of LTM2. Next SVRACA stochastically determines around which axis to reflect WM based on the just obtained probabilities. Among 100 simulations with the same input parameters SVRACA interpreted the ambiguous WM as LTM1 16 times, and as LTM2 84 times. The reflection operation (2) resulted in WMr correlating -0.3122 with LTM2 and 0.95 with LTM1 in the 16 cases (conscious perception of LTM1) and 0.3122 with LTM2 and -0.95 with LTM1 in the 84 cases (conscious perception of LTM2). If we take the square of these correlation coefficients while keeping in mind their signs we see that when LTM1 is consciously perceived LTM2 remains practically unprimed – its overlap with WMr is equal to -0.0975. In the second case however when LTM2 is consciously perceived LTM1 exhibits pronounced negative priming – the magnitude of

the negative overlap (i.e. the signed coefficient of determination) between WMr and LTM1 is -0.9025. In other words SVRACA predicts practically no priming of the unperceived interpretation when the system perceives the strong interpretation and strong negative priming of the unperceived interpretation when the weak interpretation is perceived due to external context. We proceed with an empirical study testing these predictions.

## Experiment

### Design and Stimuli

We attempted to test our current predictions and further verify our previous findings by developing a new set of ambiguous drawings. We developed 3 pairs of asymmetrical drawings. Each pair contained an easy-to-see and a hard-to-see version of a particular picture. The first pair of ambiguous drawings can be seen as either a rabbit or a dog (figure 3). The two variants of this drawing included an easy-to-see rabbit (hence a hard-to-see dog) and an easy-to-see dog (hence a hard-to-see rabbit). The second ambiguous drawing can be seen as either a duck or a goat. The third drawing embodied a mouse and a frog and its two variants depicted an easy-to-see mouse (a hard-to-see frog) and an easy-to-see frog (a hard-to-see mouse).



Easy rabbit/hard dog

Easy dog/hard rabbit

Figure 3. The two variants of the first ambiguous drawing.

Table 1: Percentages of people who saw each interpretation

Version:	% who saw the strong interpretation	% who saw the weak interpretation
Strong rabbit (weak dog):	95	5
Strong dog (weak rabbit):	78	22
Strong duck (weak goat):	71	29
Strong goat (weak duck):	95	5
Strong mouse (weak frog):	84	16
Strong frog (weak mouse):	88	12

The variants of each drawing were validated in a simple picture naming task in which each variant of each drawing was presented on a computer screen (among many fillers



depicting unambiguous drawings) and subjects were simply required to name what they saw. Each subject saw only one version of an ambiguous drawing (to avoid priming effects) and each version was seen by 100 subjects. Thus 200 subjects participated in the standardization procedure. Table 1 shows the percentages of people who interpreted each version in each way.

For each variant of a picture we compared statistically (chi squared tests) the percentages of people seeing the strong and the weak interpretation of a drawing and all comparisons yielded highly significant results. Thus our stimuli indeed represented asymmetrical ambiguity. In the study we used the drawings as follows: each subject was presented with a single version of each ambiguous drawing only. The drawing was placed in a context which favored either the easy or the hard interpretation. There were three conditions in a repeated measures design: in the first condition subjects were contextually prompted to see the easy version of a drawing; in the second condition subjects were prompted to perceive the hard version; in the control condition subjects saw an unambiguous version of a drawing.

## Procedure

Each subject saw three drawings in three conditions in three target trials. Each target drawing appeared as a possible solution to a simple analogical task of the type A:B::C:? with two possible answers. The ambiguous drawing was one of the possible answers and the only way to solve the task correctly was to perceive the ambiguous drawing in its intended interpretation. The analogical task served as a context which prompted participants to perceive consciously only one of the possible interpretations in each drawing. In the first condition (called Easy Condition or EC from now on) subjects were required to perceive the easy interpretation of a drawing in order to solve the task correctly. In the second condition (called Hard Condition or HC) subjects were required to perceive consciously only the hard interpretation of a drawing. In the third condition (Control Condition or CC) subjects were presented with an unambiguous version of the drawing which was obviously the correct response to the analogical task. Each subjects saw only one version of an ambiguous drawing in order to avoid priming effects. Subjects were required to respond with the solution to each analogical task verbally in front of a microphone by naming the option they felt best fitted the analogy. After the response the microphone triggered the start of a LDT. A word appeared at the center of the computer screen and subjects were required to indicate whether it was a meaningful word in Bulgarian or a random sequence of letters by pressing a button. After each target trial a meaningful word appeared which referred to the supposedly unperceived interpretation of the just presented ambiguous drawing. In the CC where an unambiguous version of one of the ambiguous drawings was presented the word in the LDT referred to the other interpretation of the corresponding ambiguous drawing. The timing of the study

was as follows: an analogical task remained on the screen until the subject responded to it in front of the microphone. Immediately after the response a white screen with a fixation cross at the center appeared and remained for 350ms. After that a string of letters appeared at the center of the screen and remained there until subjects indicated by pressing one of two buttons whether the letters comprised a meaningful word or not. After the response a blank screen lasted 7sec. before the beginning of the next trial. The experiment contained three target trials and seventeen filler trials (the filler trials were the same as the target ones with the exception that the analogical tasks didn't contain ambiguous drawings and the LDT following them could contain both words or non-words after each analogical task; in contrast only words followed the target trials). The order of presentation of the three conditions as well as the order of presentation of the three drawings and the variants of each drawing were counterbalanced between participants resulting in 24 different lists. Each subject participated in only one list. We give examples of the first two lists in order to clarify the procedure: A subject in the first list was firstly supposed to see an easy duck (i.e. the easy-to-see duck variant of the duck/goat drawing was presented in a task that required the subject to see a duck in order to solve the task correctly) – EC, then she was supposed to see a hard frog (i.e. a hard-to-see frog variant of the frog/mouse drawing was presented in a task that required the subject to see a frog in order to solve the task correctly) – HC and finally she saw an unambiguous dog in a task that required her to see a dog in order to solve the task correctly – CC. In the EC condition she responded to the word “goat” during the LDT, in the HC she responded to the word “mouse”, and in the CC she responded to the word “rabbit”. A participant in the second list was firstly supposed to see a hard dog (respond to the word “rabbit” during LDT) – HC, then she was supposed to perceive an unambiguous goat (respond to “duck”) – CC and finally she was supposed to see an easy frog (respond to “mouse”) –EC. The remaining 22 lists exhausted all other possible combinations of variants of pictures to be perceived (and hence words to react to during LDT) and orders of presentation of conditions. In total we collected Reaction Times (RT) to six words all appearing in each of the three conditions. Those RTs constituted our dependent measure.

## Participants

Sixty eight undergraduate students (45 females and 23 males) from New Bulgarian University participated in the study for obtaining credits.

## Results and Discussion

Fifteen subjects failed to perceive the hard-to-see interpretation in the HC condition and were replaced. Five subjects were replaced because some of their RTs to target words during the LDT exceeded our threshold of 1500ms. After replacement of those subjects we were left with 48 participants – exactly two participants per list. The decision

to have two valid participants per list was made prior to the beginning of the study. The average RTs for each condition are presented in table 2.

Table 2: Average RTs for the LDT in each condition. Standard deviations are presented in parentheses.

Condition:	RT
Control Condition (CC)	774ms. (134)
Easy Condition (EC)	777ms. (130)
Hard Condition (HC)	824ms. (146)

We analyzed our data with a linear mixed model (e.g. Hoffman, 2007; Brysbaert, 2007) in which we entered subjects and items (i.e. words in the LDT) as random factors simultaneously. Our design allowed only for the inclusion of random intercepts for each random factor. The fixed factor was “Condition” with three levels for each subject (EC, HC, CC). The analysis showed a highly significant overall result for our independent variable –  $F(2, 89.301)=5.718, p=0.005$ . Multiple comparisons (performed via the SIDAK method) revealed a significant difference between CC and HC ( $p=0.01$ ), EC and HC ( $p=0.017$ ) and virtually no difference between CC and EC ( $p=0.997$ ). We see that the results corroborate our previous findings and are perfectly consistent with the predictions of SVRACA: when subjects perceived consciously the hard-to-see interpretation words referring to the easy-to-see interpretation become negatively primed compared to a control condition involving no ambiguity but not vice versa (i.e. no priming is detected when subjects perceive consciously only the easy-to-see interpretations). It is important however not to overgeneralize these findings. Fifteen subjects in our study (22%) failed to perceive the hard-to-see interpretation in the HC (i.e. perceived only the easy interpretation or reported perceiving both interpretations) even in the presence of strong context. This wasn’t unexpected since our study involved a condition in which subjects were required to perceive consciously only a very hard interpretation of an ambiguous figure. No contextual influence exists which can assure that this happens 100% of the time. Nevertheless we feel compelled to point out that given these circumstances our findings extend only to subjects which were “context sensitive” enough to comply with our experimental manipulation. The future directions of our work include testing other explicit predictions of SVRACA. Such predictions include: priming positively only the strong interpretation of an ambiguous drawing when the figure is presented subliminally; smaller effects compared to the just presented when symmetric ambiguous drawings are involved (i.e. when dealing with drawings with two equally easy to perceive interpretations SVRACA predicts that negative priming will again occur due to reflection but the amount of priming will be considerably smaller since reflection of a bisecting angle vector will result in a smaller negative correlation with the unperceived concept in LTM);

generally smaller effects of positive priming during subliminal presentation of ambiguous stimuli.

The obtained results are coherent with analogous results obtained by Nievas and Mari-Beffa (2002) who discovered negative priming of the non-selected meaning of a homograph, but only if the non-selected meaning is the dominating one. The combination of their and our data shows that the phenomenon of negative priming of the alternative representation is broader and holds both for linguistic and perceptual task and thus should be interpreted as unconsciously building a representation of both alternatives and inhibiting the stronger one if for contextual reasons we are pressed to use the weaker interpretation.

## References

- Balcetis, E., Dunning, D. (2006). See What You Want to See: Motivational Influences on Visual Perception. *Journal of Personality and Social Psychology* 2006, Vol. 91, No. 4, 612–625
- Brysbaert, M. (2007). The language-as-fixed-effect fallacy: Some simple SPSS solutions to a complex problem. Report written for RTN-LAB Version 2.0.
- Fisher, G. H. (1968). Ambiguity of form: Old and new. *Perception & Psychophysics*, 4, 189–192.
- Hoffman, L., Bovaird, J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Beh. Res. Methods*, 39, 101–117
- Locker, L., Hoffman, L., Bovaird, J. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods* 2007, 39 (4), 723–730
- Long, G. M., & Olszewski, A. D. (1999). To reverse or not to reverse: When is an ambiguous figure not ambiguous? *Am. Journal of Psychology*, 112, 41–71.
- Long, G. M., & Toppino, T. C. (2004). Enduring interest in perceptual ambiguity: Alternating views of reversible figures. *Psychological Bulletin*, 130, 748–768.
- Nievas, F., Mari-Beffa, P. (2002). Negative priming from the non-selected meaning of the homograph. *British Journal of Psychology* (2002), 93, 47–66
- Knight, K. (1990). Connectionist ideas and algorithms. *Communications of the ACM*, vol. 33, no. 11, 59–74
- Kokinov, B., Bliznashki, S., Kosev, S., Hristova, P. (2007). Analogical Mapping and Perception: Can Mapping Cause a Re-Representation of the Target Stimulus? In: *Proc. of the 29th Annual Conference of the Cognitive Science Society*. Erlbaum, Hillsdale, NJ.
- Kokinov, B., Vankov, I., Bliznashki, S. (2009). How Analogy Could Force Re-representation of the Target and Inhibition of the Alternative Interpretation. In: *New Frontiers in Analogy Research*. Sofia: NBU Press.
- Rock, I., & Mitchener, K. (1992). Further evidence of failure of reversal of ambiguous figures by uninformed subjects. *Perception*, 21, 39–45.
- Swinney, D. (1979). Lexical Access during Sentence Comprehension: (Re)consideration of context effects. *J. of Verbal Learning and Verbal Behavior*, 18, 645–659

# Change in Foreign Language Skills Over Time

Amber N. Bloomfield ([abloomfi@umd.edu](mailto:abloomfi@umd.edu))

Megan C. Masters

Steven J. Ross

Stephen P. O'Connell

Kassandra Gynther

University of Maryland, College Park  
7005 52<sup>nd</sup> Avenue, College Park, MD 20901

## Abstract

Foreign language professionals invest considerable time and effort in acquiring foreign language skills. Of key interest is how these skills change over time, and which sustenance, or language training activities, are efficacious in maintaining or improving proficiency. This paper discusses the results of mining more than 800 test/re-test records of foreign language professionals. Analyses investigated the extent to which lag time between test occasions and formal language training impacted changes in listening and reading proficiency ratings. Results indicate that certain factors, such as initial proficiency level, affect both patterns of change and the rate at which foreign language skills manifest evidence of loss.

**Keywords:** foreign language attrition, foreign language assessment, language training

## Introduction

The study of how foreign language proficiency changes over time is a relatively new area of research, often noted as receiving its first major impetus from a conference at the University of Pennsylvania in May 1980 (e.g., Clark & Jorden, 1984; De Bot & Weltens, 1995; Lambert & Freed, 1982; Weltens, 1987). In the few decades during which research on this topic has been active, a variety of papers have been published, though many of these studies have focused on which aspects of the foreign language are lost, and in what order (e.g., syntax vs. lexical knowledge first; Jordens, De Bot, & Trapman, 1989). Although this approach is valuable for describing the language aspects most vulnerable to loss, it does not target the question of what factors (e.g., conversing informally with friends in the L2) increase or decrease the rate of loss or how quickly general language abilities, such as reading comprehension, begin to show loss. Studies which have investigated factors influencing the loss of foreign language skills have examined the duration of the period of reduced input (i.e., time elapsed since peak language ability was attained), achieved proficiency level prior to the period of reduced input, and target language use during the period of reduced input, as well as other factors. The current study expands on previous research by examining the language skills of adults with foreign language proficiency in a variety of languages at multiple points of time to determine the rate at which loss

occurs, and how change in proficiency over time is affected both by formal language training and starting proficiency level.

## Period of reduced input and change in language skills

Duration of the period of reduced input has been defined as time since the end of formal training (e.g., Bahrck, 1984) or time since the end of intensive language exposure, such as an immersion experience (e.g., Snow, Padilla, & Campbell, 1988). More generally, this factor has been conceptualized as the amount of time since learners achieved their peak proficiency (Bardovi-Harlig & Stringer, 2010), though this characterization can be problematic when learners' abilities actually improve during the period (Gardner, Lalonde, Moorcroft, & Evers, 1987; Murtagh & van der Slik, 2004). The amount of time the learner has had to lose language skills is a particularly intuitive factor in explaining degree of foreign language loss, with the common sense prediction being that loss increases with elapsed time. Several studies have found evidence for such a relationship (e.g., Murtagh & van der Slik, 2004; Nagasawa, 1999; Reetz-Kurashige, 1999), though there are some findings suggesting the rate of loss over time may not be linear (Bahrck, 1984). However, it is important to take into consideration other factors, such as the target language-specific activities the learner has engaged in during the period of reduced input, rather than just the time elapsed since some benchmark of language learning was attained.

One issue with research on the duration of the period of reduced input is that these studies explore duration as a discrete factor, investigating language skills at a few specific points in time (often at only one or two time points after the period of reduced input has begun). In general, there is a tradeoff in the literature between sample size and the number of time points at which the language skills of the participants are measured: those studies with sizable *n* (e.g., Clark & Jorden, 1984; Gardner et al., 1987; Murtagh & van der Slik, 2004) tend to measure language skills only twice, once at the beginning of the period of reduced input (i.e., the baseline) and again a set amount of time later. Cross-sectional studies, like Bahrck (1984) and Snow et al. (1988), take only one measure of language skills for each participant and compare across groups that have

experienced different durations of reduced input (note that this method does not take into account potential differences in initial proficiency levels). By contrast, some studies compare language skills for the same individual at a number of time points, but tend to involve only very small groups of participants, generally children (but see Russell, 1999, for an exception), and to focus on the specific aspects of the language that are lost rather than loss of general language ability (e.g., Hansen-Strain, 1990; Reetz-Kurashige, 1999; Yoshitomi, 1999). The reason for the small number of studies exploring multiple time points with adult foreign language learners is likely attributable to practical constraints: it is difficult to longitudinally track and maintain contact with the same group of language learners over a long period of time. Yet, to explore how foreign language skills change over time, it is desirable to examine the skills of the same set of individuals repeatedly during the period of reduced input.

### Language use during the period of reduced input

The extent to which foreign language is used during the period of reduced input is likely to be an important determiner of the amount of knowledge lost at the end of this period because it determines just how “reduced” the input during this period is for the learner. Clark and Jorden (1984) found that the learners of Japanese who did not show loss months after formal language training ended reported using the language more regularly than those who did show loss. In a similar study, Murtagh and van der Slik (2004) demonstrated that use of the target language after leaving formal language training predicted strength of language skills for learners of Irish eighteen months after leaving school. In a study with employees of the Canadian government, French-dominant bilinguals reported more opportunities to use their less dominant language (English) and also showed less loss of skills in their weaker language than did English-dominant bilinguals (Edwards, 1977, discussed in Oxford, 1982). Snow (1982) found that, at the group level, the lowest amount of loss was exhibited by Spanish immersion student groups that had the highest proportion of learners who continued to study the target language after the immersion ended (reported in Snow et al., 1988).

### Achieved proficiency and change in language skills

Achieved proficiency in the L2 is important for assessing change in foreign language skills because it provides the baseline against which to compare current ability. This factor has also been considered in its own right as a potential predictor for foreign language loss. Having higher target language proficiency may lead to decreased loss over time (for reviews, see Bardovi-Harlig & Stringer, 2010; Weltens, 1987). One potential reason for this is that having greater foreign language proficiency may provide a learner with more available strategies to compensate for loss of specific foreign language knowledge. For example, a learner could use target language morphological knowledge to

uncover the meaning of a forgotten lexical item in the same way children use this type of information to comprehend unfamiliar words (Carr & Johnston, 2001). In addition, some theories of foreign language acquisition suggest that language knowledge, once it reaches a critical threshold, simply becomes resistant to loss (Bardovi-Harlig & Stringer, 2010). However, there are somewhat mixed results for the relationship between achieved proficiency and change in language skills over time, with several studies suggesting that higher proficiency learners do indeed experience less loss over time than do lower proficiency learners (Clark & Jorden, 1984; Gardner et al., 1987; Kaufman, 1995; Nagasawa, 1999) and others suggesting there is no difference in rate of language loss between higher and lower proficiency learners (Bahrick, 1984; Weltens & Grendel, 1993).

## The Current Study

The current study explored change in general foreign language skills (reading and listening comprehension) for a large number of foreign language professionals who were tested multiple times (2-7 test occasions, with the majority having 3 or more) over a period as long as 6 years. The dataset included test histories for nearly 50 different languages; because of the small number of people testing in any one language, all analyses were completed in aggregate across the tested languages. Most people in the dataset were tested annually, but there was considerable variation in the frequency of testing. In addition to test records, information on participation in official language training was available for individuals in the dataset.

### Method

The dataset included 1084 test histories for listening and 1085 for reading.<sup>1</sup> Each test event was associated with a rating of proficiency based on the raw score from the test; the raw score was not included in the dataset. Ratings are based on the Interagency Language Roundtable (ILR) scale for that particular skill: possible scores can range from 0 to 5, with 0 = *No Proficiency* and 5 = *Functionally Native Proficiency*; between each pair of adjacent levels is a “plus level” (e.g., 0+, 2+), *assigned when proficiency substantially exceeds one skill level and does not fully meet the criteria for the next level* (Interagency Language Roundtable, 2011).<sup>2</sup> For the purpose of analyses, all ILR ratings were recoded into numeric values.<sup>3</sup> All scores were associated with a test date, so it was possible to calculate the amount of time, in days, from the first test administration to each subsequent test. Whether or not a person received language training, dummy coded as 0 = *no training* and 1 = *training*, was included in all analyses. Only a small subset

<sup>1</sup> A minority of individuals tested in more than one language; analyses treat each test history as a separate case.

<sup>2</sup> For more information on the ILR proficiency scales, see <http://www.govtirl.org/Skills/ILRscale1.htm>.

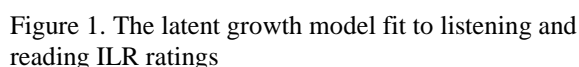
<sup>3</sup> For example, 1 = 10, 1+ = 16, 2 = 20, 2+ = 26, etc.

Initial target language proficiency was defined as the first ILR rating on record for any given language in the relevant skill (i.e., reading or listening). Due to the sparseness of individuals at some levels of proficiency, each person's first rating was coded as either *high*/2 = a rating of 2+ or greater or *low*/1 = a rating of 2 or lower. This recoded variable was used in analyses of skill loss. In addition, information about whether or not the test version changed during the testing history was available for the dataset. This factor is likely to be important when examining change in ratings, as the introduction of a new, unfamiliar test may decrease ratings even in the absence of loss of language skills. This factor was coded as 1 = *version change* and 0 = *no version change*.

**Overall patterns of change** To explore the pattern of change across test administrations, a latent growth analysis (Bollen & Curran, 2006) was used to examine the direction and trajectory of change over time. In these latent growth analyses, the first score or intercept (*ICEPT*) and the growth over time (*SLOPE*) were dependent variables, while participation in formal language training (*Training*) and the occurrence of a test version change (*verdif*) were included as independent variables (see Figure 1). Because few test histories contained more than four tests, only the first four tests in the dataset were included in the latent growth analyses.

the slope was also significant for listening scores ( $r = -0.37$ ,  $p < .001$ ), indicating that people who started with lower initial ILR ratings had a faster rate of improvement over time than did those who started with higher ILR ratings. The *training* variable had a significant negative relation ( $-0.15$ ) to the intercept for listening scores ( $t = -4.27$ ,  $p < .001$ ), indicating the people who participated in formal training had lower initial ILR ratings, perhaps indicating that these individuals self-selected for training. In addition, training had a significant positive relationship ( $0.20$ ) with slope ( $t = 3.09$ ,  $p < .01$ ), indicating that the trajectories of improvement for professionals who had some kind of language training were higher than for those who had not participated in a training activity. The version change variable also had a significant relationship with slope, but this relationship was negative ( $-0.19$ ;  $t = -2.88$ ,  $p < .01$ ). The direction of this relationship is intuitive: the introduction of a new test version decreases the rate of improvement. What is less intuitive, however, is the significant positive relationship ( $0.09$ ) between the version change variable and the intercept ( $t = 2.76$ ,  $p < .01$ ). This relationship indicates that people who experienced a version change at some time in their testing history also tended to have higher initial ILR ratings. A new test version was introduced for only a subset of languages during the time covered by the test history data, so whether a version change occurred was partially dependent on the language tested. It may be that initial proficiency in this dataset was higher in those languages that experienced a version change, leading to the relationship between version change and the intercept. In fact, chi-square tests revealed that significantly more people with starting proficiency of 2+ or above experienced a version change for both listening and reading ( $\chi^2(1) = 8.08$  and  $\chi^2(1) = 8.32$ , respectively; both  $ps < .01$ ).

As for listening, the *training* variable has a significant negative relationship ( $-0.20$ ) with the intercept for reading ( $t = -5.71, p < .001$ ), indicating that people who participated in training tended to have lower initial reading ILR ratings, and has a significant positive relationship ( $0.16$ ) to the slope ( $t = 2.67, p < .01$ ), suggesting that improvement was faster for those individuals who have received some type of language training. The version difference variable has a significant negative relationship ( $-0.19$ ) to slope ( $t = -2.88, p < 0.01$ ); as for listening, the introduction of a new test version leads to a slower rate of growth in reading ratings



over time. The same non-intuitive positive relationship between version change and the intercept was also present for reading scores (0.12;  $t = 3.42$ ,  $p < .001$ ). Again, the most likely explanation is that those languages for which a new test version has been introduced also tend to be languages where the individuals have a higher initial rating.

Ratings improved over time for both reading and listening, as indicated by the significant positive slopes in the latent growth analyses. Version change negatively impacted this trend, leading to a slower rate of improvement over time for listening scores but did not affect change over time for reading scores. Training, on the other hand, positively affected this trend, indicating that the formal language training resulted in faster improvement in ILR ratings for both listening and reading. While people with lower initial ratings showed a faster rate of improvement in listening, this pattern was reversed for reading. It may be that attaining higher proficiency levels in reading becomes easier as proficiency increases, while attaining higher levels for listening becomes more difficult as proficiency increases. This is a topic for future research.

**Loss in ILR ratings** To examine how the amount of time between test occasions affected the incidence of loss in ratings, event history analyses were conducted separately for reading and listening. A test history was coded as showing a loss if any subsequent test occasion produced a lower ILR rating than the first test occasion. In an event history analysis, the time between the first test and the test that yielded a lower score (i.e., lag time) is modeled to capture the amount of lag time for cases coded as losses compared to the lag time for cases where scores are sustained or increased from the first to the most recent test (i.e., non-loss cases).<sup>4</sup> The goal of the event history analysis is an estimation of the average time lag associated with increasing incidences (events) of proficiency loss and the effect of any covariates on the rate of loss.

The version change factor was revised somewhat for the event history analyses. The factor captured whether a person who experienced a loss experienced a version change *prior to that loss*, rather than at another point in their test history. For individuals who did not show loss, this factor continued to indicate whether they had experienced a version change at any time in their test history.

Across all sets of listening test records, 17.7% of the cases were coded as showing a loss (i.e., at least one rating was lower than the first ILR rating on record). The event history analysis estimates the time lag associated with survival (i.e., not showing a loss) as a function of time beginning with 100% of the cohort surviving. The event history model assesses whether loss is a possible function of time between test occasions. For an event phenomenon

impervious to time, the rate of loss would be expected to remain close to zero across all observed time lags. However, the prediction based on previous studies (e.g., Nagasawa, 1999; Reetz-Kurashige, 1999) would be for the rate of loss to increase as time lag increases, which was in fact found for listening ratings. Event history analysis revealed that the projected survival rate for listening was fairly long, with over 80% of cases found to maintain or improve listening ratings with three years' lag between test occasions (see Figure 2).

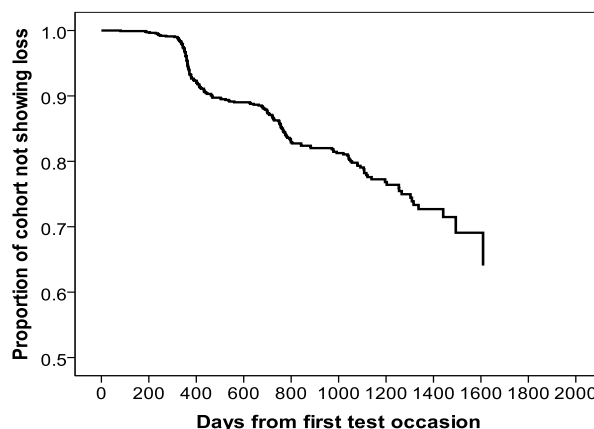


Figure 2. Loss in listening ratings over time

Version change was a marginally significant covariate in the event history analysis ( $B = 0.32$ ,  $p < 0.07$ ), with the predictable impact of increasing the rate of loss for listening: the odds of showing loss were projected to be roughly 1.37 times higher for people experiencing a version change than those who did not experience a version change. In addition, whether or not the person engaged in formal language training was entered as a covariate into the analysis. Participation in training did not significantly affect rate of loss for listening ILR ratings ( $B = -0.15$ ,  $p = 0.51$ ). However, the initial proficiency of the individual, coded as *low* for those with an initial ILR rating of 2 or lower and as *high* for those with first ILR rating of 2+ or higher, was a significant covariate ( $B = 0.71$ ,  $p < 0.01$ , see Figure 3).

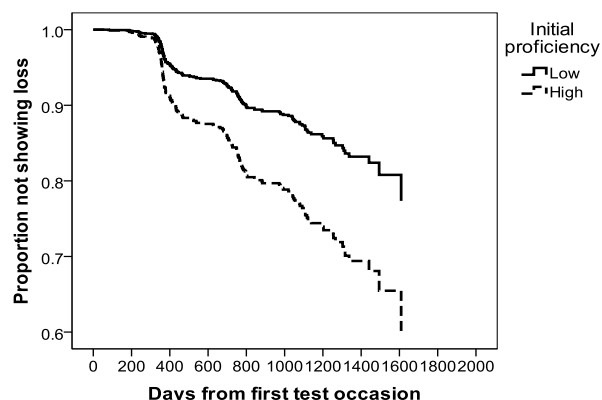


Figure 3. Initial proficiency and loss in listening ratings

<sup>4</sup> Cases where the first rating on record was 0 (providing no opportunity to see a pattern of loss over time) and those where two test ratings were listed for the same date (suggesting the person was tested twice in one day) were excluded from event history analyses.

Surprisingly, however, those with higher scores exhibited a faster rate of loss, with the odds of this group showing loss projected as roughly 1.99 times higher than the odds for those with lower initial ratings. This finding is contrary to what has been found in previous studies (Clark & Jorden, 1984; Gardner et al., 1987; Kaufman, 1995; Nagasawa, 1999), and will be discussed in the Summary below.

Across all sets of reading test records, 17.6% of the cases were coded as showing a loss compared with the first score. As for listening, the incidence of loss increased with the amount of time between test occasions for reading tests scores (see Figure 4). Event history analysis projected the survival rate for reading to be comparable to that seen for listening: about 80% of cases were found to maintain or improve listening scores with three years' lag between test occasions.

Consistent with the analyses completed for listening scores, test version change was entered as a covariate into the event history analysis for reading scores. This factor failed to approach significance for reading scores, however ( $B = 0.13, p = 0.48$ ). Whether or not the individual engaged in formal language training also failed to approach significance as a covariate for loss of reading ratings ( $B = 0.12, p = 0.56$ ). However, as for listening ILR ratings, the first rating on record, coded as *high* = 2+ or higher and *low* = 2 or lower, was again a significant covariate ( $B = 0.55, p < .05$ ; see Figure 5). The difference between the high and low proficiency groups was similar to that seen for listening scores, with the higher proficiency group showing a faster rate of loss than the lower proficiency group.

The results of the event history analyses reveal that the proportion of test cases in the dataset showing loss increased as the amount of time since the first test increased. However, training did not have a significant effect on the rate of loss for listening or reading. Because only a small portion (~14%) of the individuals included in the current analyses had formal language training on record, and only a subset of this group experienced a loss in ratings, it is possible that there was not enough power in these analyses to detect an impact of training on the rate of skill loss. Further, experiencing a test version change was a marginally significant covariate for the rate of loss of listening scores, but not for reading scores.

## Summary and Conclusion

The findings from the current study are in line with previous results from studies examining change in foreign language skills in several ways. Consistent with previous research investigating the duration of the period of reduced input (e.g., Murtagh & van der Slik, 2004; Nagasawa, 1999; Reetz-Kurashige, 1999), the probability of loss increased as the amount of time since the first test increased for both reading and listening. These results indicate that, all other factors being equal (amount of foreign language use, motivation, etc.), ILR ratings for these individuals will tend to decrease as the amount of time between test administrations increases.

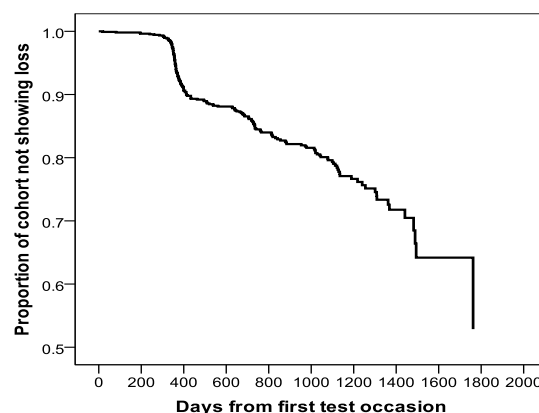


Figure 4. Loss in reading ratings over time

In contrast, the investigation of achieved proficiency (the degree of proficiency achieved in the language) and its effects on the rate of loss were not consistent with previous research findings. In the event history analyses, initial proficiency was a significant covariate for the rate of loss for all three skills, with higher proficiency individuals projected to have *faster* loss rates than lower proficiency individuals. This result is in the opposite direction from what is typically found when there is a difference in change for higher and lower proficiency learners (e.g., Clark & Jorden, 1984; Gardner et al., 1987). The current dataset contained many more individuals with *high* proficiency (a first score of 2+ or higher) than with *low* proficiency (a first score of 2 or lower), so it is possible that there was something unique about the individuals with lower initial ILR ratings that led to the relationship between initial proficiency and rate of loss. It is also possible that maintaining a higher rating is simply more difficult, with a smaller margin for error, leading to a faster rate of loss for this group. Further, there may be differences in motivation between the two groups, with the lower proficiency individuals more motivated to work to improve their language skills, and so slower to show loss over time. These possibilities will be investigated in future studies.

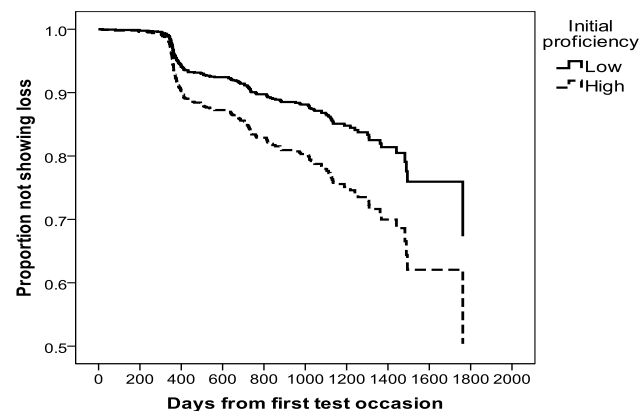


Figure 5. Initial proficiency and reading attrition



The current study also found no effect of whether an individual engaged in formal language training on the rate of loss for either listening or reading, though engaging in training did lead to a faster rate of improvement over time. This result runs counter to previous findings showing that the extent of target language use and exposure during the period of reduced input predicts retention (e.g., Clark & Jorden, 1984; Edwards, 1977; Murtagh & van der Slik, 2004). However, it is important to note that the information on formal language training was available for only a small subset of all people (14%), and that it is unclear why this subset of the sample participated in language training. It may be that these individuals were selected for training specifically because they struggled to maintain their foreign language skills. Further, the records of formal language training were the only information available about what the people included in the dataset might have been doing with their foreign language skills since the first language test in the dataset. It is very possible that large individual differences exist in the sample in terms of informal language training and other types of use and exposure (e.g., watching TV in the target language) during the period of time since the first test.

In conclusion, the current study introduced a method for investigating factors affecting change in adult foreign language skills in a longitudinal design, and described the results of an analysis of change in proficiency ratings for one group of foreign language professionals. Although this dataset was limited in several ways, including providing little information about use of the foreign language between proficiency tests, it did offer two or more data points for most individuals that were spread across a number of years. Further, the first ILR rating in the dataset offered a proxy for achieved language proficiency, so a given individual's current performance could be compared against his or her own previous abilities. A survey is currently being distributed to collect additional information about language learning history and current language use for professionals in this population to provide a more complete picture of the factors that affect change in foreign language skills.

### Acknowledgements

This research was supported by the University of Maryland Center for Advanced Study of Language with funding from the Department of Defense.

### References

- Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General*, 113(1), 1-29.
- Bardovi-Harlig, K., & Stringer, D. (2010). Variables in second language attrition. *Studies in Second Language Acquisition*, 32(1), 1-45.
- Bollen, K. A., & Curran, P. J. (2006). *Latent curve models*: Wiley-Interscience.
- Carr, L., & Johnston, J. (2001). Morphological cues to verb meaning. *Applied Psycholinguistics*, 22(4), 601-618.
- Clark, J. L. D., & Jorden, E. H. (1984). A Study of Language Attrition in Former US Students of Japanese and Implications for Design of Curriculum and Teaching Materials. Retrieved August 4, 2011 from ERIC Document Reproduction Service, <http://www.eric.ed.gov>, No. ED243317.
- De Bot, K., & Weltens, B. (1995). Foreign language attrition. *Annual Review of Applied Linguistics*, 15(1), 151-164.
- Edwards, G. (1977). Second language retention in the public service of Canada. *Ottawa: Research Section, Official Languages Directorate*.
- Gardner, R. C., Lalonde, R., Moorcroft, R., & Evers, F. (1987). Second language attrition: The role of motivation and use. *Journal of Language and Social Psychology*, 6(1), 29-47.
- Hansen-Strain, L. (1990). The attrition of Japanese by English-speaking children: An interim report. *Language Sciences*, 12(4), 367-377.
- Interagency Language Roundtable. (2011). Descriptions of proficiency levels. Retrieved January 1, 2012, from <http://www.govtllr.org/Skills/ILRscale1.htm>
- Jordens, P., De Bot, K. D., & Trapman, H. (1989). Linguistic aspects of regression in German case marking. *Studies in Second Language Acquisition*, 11(2), 179-204.
- Kaufman, D. (1995). Where Have All the Verbs Gone? Autonomy and Interaction in Attrition. *Southwest Journal of Linguistics*, 14, 43-66.
- Lambert, R. D., & Freed, B. F. (1982). *The Loss of Language Skills*. Rowley, MA: Newbury House Publishers, Inc.
- Murtagh, L., & van der Slik, F. (2004). Retention of Irish skills: A longitudinal study of a school-acquired second language. *International Journal of Bilingualism*, 8(3), 279-302.
- Nagasawa, S. (1999). Learning and losing Japanese as a second language: A multiple case study of American university students. In L. Hansen (Ed.), *Second Language Attrition in Japanese Contexts* (pp. 169-212). Oxford: Oxford University Press.
- Oxford, R. (1982). Technical issues in designing and conducting research on language skill attrition. In R. D. Lambert & B. F. Freed (Eds.), *The Loss of Language Skills* (pp. 119-137). Rowley, MA: Newbury House.
- Reetz-Kurashige, A. (1999). Japanese returnees' retention of English-speaking skills: Changes in verb usage over time. In L. Hansen (Ed.), *Second language attrition in Japanese contexts* (pp. 21-58). Oxford: Oxford University Press.
- Russell, R. A. (1999). Lexical maintenance and attrition in Japanese as a second language. In L. Hansen (Ed.), *Second language attrition in Japanese contexts* (pp. 114-141). Oxford: Oxford University Press.
- Snow, M. A. (1982). *Graduates of the Culver City Spanish Immersion Program: A follow-up report*. Paper presented at the Sixteenth Annual TESOL Convention, Honolulu, HI.
- Snow, M. A., Padilla, A., & Campbell, R. (1988). Factors influencing language retention of graduates of a Spanish immersion program. *Applied Linguistics*, 9, 182-197.
- Weltens, B. (1987). The Attrition of Foreign-Language Skills: A Literature Review. *Applied Linguistics*, 8(1), 22-38.
- Weltens, B., & Grendel, M. (1993). Attrition of vocabulary knowledge. In R. Schreuder & B. Weltens (Eds.), *The bilingual lexicon* (pp. 135-156). Amsterdam: Benjamins.
- Yoshitomi, A. (1999). On the loss of English as a second language by Japanese returnee children. In L. Hansen (Ed.), *Second language attrition in Japanese contexts* (pp. 80-111). Oxford: Oxford University Press.

# Look-Ahead Monte Carlo with People

**Charles Blundell (c.blundell@gatsby.ucl.ac.uk)**

Gatsby Computational Neuroscience Unit, University College London, London, UK

**Adam Sanborn (a.n.sanborn@warwick.ac.uk)**

Department of Psychology, University of Warwick, Coventry, UK

**Thomas L. Griffiths (tom\_griffiths@berkeley.edu)**

Department of Psychology, University of California, Berkeley, Berkeley, CA USA

## Abstract

Investigating people's representations of categories of complicated objects is a difficult challenge, not least because of the large number of ways in which such objects can vary. To make progress we need to take advantage of the structure of object categories – one compelling regularity is that object categories can be described by a small number of dimensions. We present Look-Ahead Monte Carlo with People, a method for exploring people's representations of a category where there are many irrelevant dimensions. This method combines ideas from Markov chain Monte Carlo with People, an experimental paradigm derived from an algorithm for sampling complicated distributions, with hybrid Monte Carlo, a technique that uses directional information to construct efficient statistical sampling algorithms. We show that even in a simple example, our approach takes advantage of the structure of object categories to make experiments shorter and increase our ability to accurately estimate category representations.

**Keywords:** category representation; Markov chain Monte Carlo; directional judgements

## Introduction

Categories are an essential component of how people reason about the world, allowing us to act intelligently when we encounter new objects, new people, or new situations. Natural stimuli, such as images or text, tend to be very complicated. In contrast, much of our understanding of categorisation behaviour has been built on experiments using well-controlled stimuli that vary along only one or two dimensions (e.g., Nosofsky, 1986). A major driver of the disconnect between experimental investigations and real-world behaviour is that standard experimental methods do not allow the experimenter to adaptively focus on informative stimuli. Attempting to map out a complicated category with standard methods would require participants to complete an unendurable number of trials.

Identifying connections between cognitive science and statistics can help us develop new methods to extend our experimental reach. Many of the computational models of categorisation developed from carefully-controlled laboratory studies can be interpreted as representing categories as probability distributions over their constituent stimuli (Ashby & Alfonso-Reese, 1995). Statisticians have developed sophisticated methods to sample from high-dimensional distributions and Sanborn and Griffiths (2008) identified how to use one of these methods, Markov Chain Monte Carlo (MCMC), to draw samples from categories that were represented in the mind. This method, termed Markov Chain Monte Carlo with

People (MCMCP), focused trials on informative regions of stimuli – greatly reducing the number of trials a participant needs to perform.

While MCMCP exploited the local nature of categories, it could not exploit the way that many categories lie on manifolds. Many seemingly complex stimulus representations possess a simpler embedded representation. For example, images, when represented as a grid of pixel values, inhabit a high-dimensional space, typically hundreds to hundreds of thousands of dimensions. Yet modifications to these images, such as changing a facial expression or moving a limb, require simultaneously modifying only some of the dimensions. Thus often stimuli possess many irrelevant dimensions and many “linked” dimensions, where changing one implies proportionally changing many other connected dimensions. Since such relationships are often smooth among dimensions, suggesting a manifold structure.

In this paper, we present an extension to MCMCP that is able to take advantage of the manifold structure of continuous-valued stimuli and apply it to a low-dimensional stimulus. We take inspiration from the large literature in statistical computing that has explored how MCMC algorithms can be modified to increase their efficiency in solving specific problems. For example, Hybrid Monte Carlo (Neal, 1996) is a successful method for sampling in high-dimensional spaces that incorporates directional information into MCMC methods. We adapt MCMCP to include directional information, allowing participants to “look ahead” at the consequences of their decisions so they can guide stimulus generation, increasing the efficiency with which we can investigate categories.

The remainder of this paper is organised as follows. We first introduce the MCMC and explain how it was adapted to people. Next we introduce Look-Ahead Monte Carlo with People (LAMCP), first motivating it intuitively and then presenting the technical justification of our method. We compare LAMCP to MCMCP empirically in an experiment exploring the golden ratio in rectangles—a simple example of a low-dimensional manifold in a higher-dimensional space. We then conclude with a brief discussion how our method relates to hybrid Monte Carlo and other sampling techniques.

## Background: MCMC with People

Using the connection between computational models of categorisation and probability distributions (Ashby & Alfonso-

Reese, 1995), let  $p(x|c)$  denote the probability of stimulus  $x$  under the distribution associated with category  $c$ . We could attempt to elicit people's category representation by asking them to rate the typicality of item  $x$  in category  $c$ , but that would require us to present participants with every stimulus of interest. Instead we would like to draw samples from  $p(x|c)$ .

The Metropolis sampling scheme (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) is a commonly used algorithm for generating a sequence of samples from a designer-supplied probability distribution. It constructs a Markov chain whose stationary distribution is the provided probability distribution and then uses this Markov chain to, eventually, generate a sequence of samples from the stationary distribution. The algorithm has two parts that are provided by the designer: a proposal distribution  $q(x^*|x)$ , and an acceptance function  $A(x^*, x)$ . The proposal distribution is typically a simple distribution (such as a Gaussian or uniform distribution) and is used to propose possible samples,  $x^*$ , given the current sample,  $x$ . The acceptance function tests how similar samples from this proposal distribution are to the desired stationary distribution, guiding the algorithm towards this distribution. Both the proposal distribution and acceptance function must be carefully selected to ensure the Markov chain converges correctly (see Neal, 1993). At each step  $t$  of the algorithm, a new state of the chain is proposed,  $x_t^*$ , by sampling it from the proposal distribution,  $q(x_t^*|x_{t-1})$ . With probability  $A(x_t^*, x_{t-1})$ , the state of the Markov chain at step  $t$  is the proposal  $x_t^*$ , otherwise it is the previous state, i.e.,  $x_t = x_{t-1}$ . The initial state  $x_0$  of the Markov chain is picked at random. It can be shown that if this procedure is repeated for long enough, the Markov chain it defines will eventually converge on the desired stationary distribution.

MCMCP (Sanborn & Griffiths, 2008; Sanborn, Griffiths, & Shiffrin, 2010) transformed the Metropolis sampling scheme into an experimental method for cognitive science. MCMCP is a sequential paradigm in which participants construct the stimuli themselves, in small, manageable steps. The equilibrium distribution of interest is the distribution over stimuli belonging to a single category  $p(x|c)$ , which we shall write as  $\pi(x)$  for short-hand, the particular category being implicit.

On each trial of the MCMCP procedure, a new stimulus sample  $x_t^*$  is generated by a computer from an experimenter-provided proposal distribution, such as a Gaussian or uniform distribution. Participants are presented with a choice of two possible samples; the current state,  $x_{t-1}$ , or the proposed new state,  $x_t^*$ . Their selection becomes the new state of the Markov chain,  $x_{t+1}$ . The next trial has the same form, and the procedure repeats until the Markov chain is deemed to have equilibrated.

If participants choose  $x_t^*$  according to a valid acceptance function, then one can show that, just like Metropolis sampling, this scheme forms a Markov chain whose stationary distribution is the distribution over stimuli in a particular category  $\pi(x)$ . Fortunately there is an exact correspondence be-

tween the ratio rule of human decision making (Luce, 1963) and a valid acceptance function for use with Metropolis sampling, namely the Barker acceptance function (Barker, 1965):

$$A(x_t^*, x_{t-1}) = \frac{\pi(x_t^*)}{\pi(x_t^*) + \pi(x_{t-1})} \quad (1)$$

Assuming participants' choices are Markov, the paradigm forms a Metropolis sampling scheme whose states are samples from people's distribution of objects in a particular category at the equilibration of the Markov chain.

The proposal distribution used by MCMCP is fixed and provided by the experimenter; typically it is an isotropic Gaussian distribution. This induces a random-walk Markov kernel, which, as Neal (1993) notes for general MCMC, and as Martin, Griffiths, and Sanborn (2012) note for MCMCP, can be inefficient for exploring large, correlated spaces. Gains in efficacy are to be had by removing this random walk, whilst maintaining the properties that allow it to converge to the category distribution.

## An Intuitive View of MCMCP and LAMCP

Before presenting our novel Look-Ahead Monte Carlo with People (LAMCP) method, we will provide some intuitions as to how it differs from MCMCP. Intuitively, we can think about exploring a probability distribution by following a Markov chain in terms of a hiker attempting to travel along a ridge path. The ridge represents an interesting low-dimensional manifold embedded in a largely irrelevant higher-dimensional space, and we wish participants to explore this manifold efficiently.

Suppose our hiker is standing upon a bumpy ridge in an otherwise large flat landscape. He wishes to explore the ridge, whilst only descending into the lower parts of the landscape fleetingly. Suppose also that he cannot see anything, and must be told about the terrain by MCMCP or LAMCP.

MCMCP allows our hiker to know about the terrain where he currently stands and also at another location, picked at random by MCMCP. He must then choose whether he wishes to step to this new location or stay where he is. MCMCP does not know of our hiker's intention to follow the ridge, and so when MCMCP proposes locations far away from him, the new location is likely to be in the flat and so he will often elect not to move. Proposed locations very close to our hiker are likely to be on the ridge so he will be willing to make a step. Walking along the ridge in this random fashion will take a great many small steps.

LAMCP gives our hiker a guide. This guide walks away from the hiker along a randomly oriented straight line. The guide then returns to the hiker and tells him on average, how much of the ridge she saw on her travels. If the hiker feels that the guide saw a lot of the ridge, then LAMCP randomly picks a location along this straight line. The hiker can then either stay where he is, or walk to this new location offered by LAMCP. With LAMCP, the hiker can take advantage of a longer view. In addition, our guide will propose the same

direction in which our hiker just travelled, which will be advantageous in following a straight ridge path. Both aspects allow our hiker to travel more quickly.

### Look-Ahead Monte Carlo with People

Look-Ahead Monte Carlo with People (LAMCP) is a sequential paradigm that, like MCMCP, samples from a participant's distribution over stimuli for a particular category. Unlike MCMCP, LAMCP has two kinds of trials. The first kind are just like the trials of MCMCP; participants are asked to choose between a generated stimulus and the previous stimulus as the next state of a Markov chain. In the second kind of trial, however, participants are asked to choose a direction in the stimulus space to explore. This directional information is then used to generate a stimulus to be presented to the participant in the first kind of trial. Thus LAMCP produces two kinds of samples; stimuli, which we shall denote  $x$ , and also directions, which we shall denote with  $d$ .

LAMCP alternates between proposing stimuli using the current direction and previous stimulus and proposing directions using the current stimulus and previous direction. In this way, LAMCP is able to capture and to some extent remember local manifold structure when generating stimuli.

Recalling the analogy of the hiker trying to follow a ridge; new stimuli are generated from directions by starting at the current stimulus and advancing some distance according to the current direction. The distance advanced is sampled at random. More precisely, the two kinds of trials of LAMCP operate as follows:

**Stimulus trial:** Suppose that a direction  $d_t$  has been sampled already. The direction has an additive effect upon the current stimulus and so the proposal for the new stimulus,  $x_t^*$ , is:

$$x_t^* = x_{t-1} + \epsilon_t d_t$$

where  $\epsilon_t$  is a random value, sampled once for each stimulus trial, with distribution which we shall denote  $q_\epsilon(\epsilon_t)$ , determining for how far the direction  $d_t$  should be followed. The variable  $\epsilon_t$  is the distance travelled, in direction  $d_t$  to obtain the new stimulus.

Participants are then presented with the previous stimulus  $x_{t-1}$  and the proposal  $x_t^*$  and asked which of these stimuli looked like they belonged more to the category of interest  $\pi(x)$ . As in MCMCP, we assume this choice is made by participants according to the ratio rule and thus corresponds to the Barker acceptance function (Equation 1).

**Direction trial:** In this step, participants pick a suitable direction for advancing the current stimulus  $x$ . A direction proposal is sampled from the direction proposal distribution  $q_d(d_t^* | d_{t-1}, x_{t-1})$ . Direction values are presented to participants as animations, showing how the proposed stimuli will be derived from the direction. This is the look-ahead part of our paradigm: this animation provides participants with insight into how selecting a direction would

affect the proposal of stimuli during future stimulus trials. Participants can decide to continue along a single direction for multiple stimulus trials by continuing to select the current direction during a direction trial.

Suppose  $N$  frames are to be generated for a proposed direction  $d_t^*$ . First  $N$  values of  $\epsilon_t$  are selected, typically uniformly spaced within some task-specific interval:  $\epsilon_t^1, \epsilon_t^2, \dots, \epsilon_t^N$ . Then frame  $n \in \{1, \dots, N\}$  is the stimulus produced by  $x_t + \epsilon_t^n d_t^*$ . The animation loops, first increasing the frame number  $n$  from 1 to  $N$ , then decreasing the frame number from  $N$  to 1.

Participants are presented with two animations: one corresponding to the proposed direction  $d_t^*$  and another for the previous direction  $d_{t-1}$ . Participants are asked to select the animation in which the stimuli look most like they belongs to the category of interest. By doing so, participants are picking the direction that is most likely to offer a stimulus belonging to the category during the next stimulus trial.

Again, as in MCMCP, we assume this choice is made by participants according to the ratio rule and thus corresponds to the Barker acceptance function (Equation 1).

In summary, the LAMCP paradigm is as follows:

1. Generate new direction proposal from direction proposal distribution  $q_d$ :

$$d_t^* \sim q_d(d_t^* | d_{t-1}) \quad (2)$$

The experimenter provides  $q_d$ : the distribution is required by the Barker acceptance function to be symmetric,  $q_d(d^* | d) = q_d(d | d^*)$ .

2. Generate an  $N$ -frame animation of the resulting samples:  $x_t^* = x_{t-1} + \epsilon_n d_t^*$  for each  $n \in \{1, 2, \dots, N\}$ .
3. Participants asked to choose between  $d_t^*$  and  $d_{t-1}$  for the new direction  $d_t$ , based upon the animated stimuli. In particular they select the new direction with probability:

$$\frac{\pi(d_t^* | x_{t-1})}{\pi(d_t^* | x_{t-1}) + \pi(d_{t-1} | x_{t-1})}$$

4. Generate new stimulus proposal from direction  $d_t$  and previous stimulus  $x_{t-1}$ :

$$\epsilon_t \sim q_\epsilon(\epsilon_t) \quad (3)$$

$$x_t^* = x_{t-1} + \epsilon_t d_t \quad (4)$$

The experimenter provides  $q_\epsilon$ . Typically it will be a Gaussian distribution or a scale mixture of Gaussian distributions.

5. Participants are asked to choose between  $x_t^*$  and  $x_{t-1}$  as the new stimulus  $x_t$ , choosing  $x_t^*$  with probability:

$$\frac{\pi(x_t^*)}{\pi(x_t^*) + \pi(x_{t-1})}$$

In terms of Markov chains, LAMCP is a Metropolis sampler, where each kind of trial can be understood to be updates to either the stimulus or the direction. The overall stationary distribution is  $\pi(x, d) = \pi(x)\pi(d|x)$ . Participants are asked to pick directions whose animation show a high probability stimulus for the longest. Stated more formally, the stationary distribution for directions is:

$$\pi(d|x) \propto \int \pi(x^*) \delta(x^* = x + \epsilon d) dx^* d\epsilon \quad (5)$$

where  $x^*$  are the look-ahead points and  $\pi(x^*)$  is the stimulus distribution evaluated at these look-ahead points, whilst  $\epsilon$  is their distance along the direction  $d$  from the original stimulus  $x$ .

### Extending LAMCP to Many Choices

As originally developed, MCMCP and LAMCP consisted of two alternative forced choice (2AFC) trials. This follows naturally from the Barker acceptance function (Equation 1). However, a larger number of alternatives could provide even more informative judgements.

At the start of each trial, instead of generating one proposal  $x_t^*$ , a set of  $n$  proposals is generated  $\{x_t^p : p \in \{1, \dots, n\}\}$ . Let  $c$  denote the index of the selected choice; instead of the Barker acceptance function in Equation 1, we use:

$$A(x_t^c, x_{t-1}) = \frac{\pi(x_t^c)}{\pi(x_{t-1}) + \sum_{p=1}^n \pi(x_t^p)} \quad (6)$$

Just as the Barker acceptance function (Equation 1) naturally arises in people's decision making from the Luce's choice axiom (Luce, 1963), so too does Equation 6.

It is not obvious, however, whether such an acceptance function leads to a valid Markov chain Monte Carlo sampler. By considering the multiple choice Markov transition kernel, one can show that the condition of *detailed balance* is satisfied by this acceptance function, and hence the above acceptance function maintains the equilibrium distribution  $\pi(x)$ . In particular, the Markov transition kernel is:

$$T(x \rightarrow x') = \sum_{p=1}^n \delta(x', x^p) A(x^p, x) S(x, x^p) + \delta(x', x) \left[ 1 - \sum_{p=1}^n \int A(x^p, x) S(x, x^p) dx^p \right] \quad (7)$$

where  $S$  is the symmetric proposal distribution for generating new proposals. Detailed balance requires one to show that  $\pi(x)T(x \rightarrow x') = \pi(x')T(x' \rightarrow x)$  holds for all  $x$  and  $x'$ , which is easily achieved by substitution of definitions and algebra.

Though we have shown that the connection between the ratio rule and a valid acceptance function holds for more than two choices, there is evidence that people do not follow the ratio rule in this case (Wills, Reimers, Stewart, Suret, & McLaren, 2000). Rouder (2004) suggests that behaviour is better explained by raising the choice probabilities to a power

that varies from trial to trial. In our experiment, we shall only use this new regime for direction trials, where we do not need to sample from a stable distribution. In future, however, it would be interesting to explore how robust MCMCP is to the deviations from the ratio rule that people display.

### Experiment: Testing LAMCP

We evaluated whether LAMCP is able to produce proposals that are more commonly accepted than MCMCP, generated higher quality samples than MCMCP, and use fewer trials to achieve the same quality of estimates as MCMCP. Each trial shall correspond to one decision made by a participant. For LAMCP, this means it will take two trials to produce one stimulus.

We applied LAMCP and MCMCP to a simple task where stimuli are parameterized by two dimensions—rectangles with width and height. Our aim is to elicit from participants samples from the category of golden rectangles, where the height is equal to the golden ratio (1.618) times the width. This scenario is a manifestation the ridge example that we motivated our approach with earlier. The golden ratio lies along a ridge in the two dimensional space form by all widths and heights.

Participants were asked to select the rectangle that looked most like a golden ratio rectangle on stimulus trials, or in the case of direction trials, to pick the animation that looked most like a golden ratio rectangle for the longest amount of time. We evaluated how their output deviated from ideal golden ratio rectangles.

**Participants.** A total of 43 participants were recruited from Amazon Mechanical Turk: each having at least a 95% task approval rate, and had at least 100 approved tasks. Each participant was required to contribute at least 100 decisions to the task to be paid \$0.05. To check for consistency among participants, approximately 10% of participants' decisions were repeats of their own or other participants' decisions. These repeats were not used in the analysis. Participants' decisions were incorporated in real-time, and so some participants discontinued the experiment before 100 decisions but their results were still included (five participants for LAMCP, six for MCMCP24, and two for MCMCP).

**Stimuli.** Stimuli were black rectangles rendered in the participant's web browser. Each stimulus was drawn in a 232 pixel by 232 pixel light grey box, with an internal border of 5 pixels. The light grey box had a border of 5 pixels surrounding it, and was on a white background. For animated stimuli, 25 frames were generated, and the frame was advanced, in a loop, changed approximately every 100 ms. For 2AFC trials, stimuli were shown side by side. For 4AFC trials, stimuli were shown in a grid of two rows and two columns. The stimuli parameters (width and height) were generated by either MCMCP or LAMCP, using truncated Gaussians for the

initial stimulus and proposal distributions to ensure the parameters remained within the range 0 to 1. For MCMCP, the variance of the truncated Gaussians was randomly chosen at each trial to either be 0.01 or 0.25. For LAMCP, the variance of the truncated Gaussians was randomly chosen at each trial to either be 0.1 or 0.5. Variance parameters for LAMCP were higher than those for MCMCP.

**Procedure.** Participants were asked to study 24 examples of rectangles, six of which were golden ratio rectangles and 18 non-golden ratio rectangles. Six of the counter-examples were the six examples rotated by 90 degrees, with explicit instructions highlighting that golden ratio rectangles are tall and thin, not short and wide.

We ran three regimes—MCMCP with just 2AFC trials, MCMCP with alternating 2AFC and 4AFC trials (to account for the effects of including alternating 2AFC/4AFC trials, which we shall call MCMCP24), and LAMCP with 2AFC for stimulus trials and 4AFC for direction trials. For each regime, we ran 10 chains, each chain being 100 trials long. Each trial consisted of choosing between either the stimulus selected by the previous trial and one or three new stimuli.

**Results.** The median acceptance rates of MCMCP and MCMCP24 were 38% ( $\pm 2\%$ ; semi-interquartile range) and 37% ( $\pm 4\%$ ), respectively, whilst the median acceptance rate of stimulus for LAMCP was 46% ( $\pm 2\%$ ), suggesting that the proposals generated by LAMCP were typically more representative than those generated by MCMCP.

The median absolute difference between the estimated golden ratio and the true golden ratio was 0.72 ( $\pm 0.44$  semi-interquartile range) and 0.92 ( $\pm 2.73$ ) for MCMCP and MCMCP24, respectively. The median absolute difference of LAMCP, 0.52 ( $\pm 0.38$ ) is closer to the golden ratio than both MCMCP methods. The absolute differences of MCMCP and MCMCP24 are significantly ( $p < 0.001$ ) different to the absolute differences of LAMCP under the Wilcoxon rank-sum test.

Effective sample size is a heuristic for determining the number of independent samples yielded by an MCMC procedure. We compared effective sample size estimates of LAMCP, MCMCP and MCMCP24 using R-CODA (Plummer, Best, Cowles, & Vines, 2006). We found that the median effective sample size for LAMCP was 5 ( $\pm 1$ ) whilst it was 11 ( $\pm 4$ ) and 19 ( $\pm 15$ ) for MCMCP and MCMCP24, respectively. This suggests that whilst LAMCP produces more favourable samples, the samples are more correlated with one another than those produced by MCMCP. This could be a consequence a linear correlation introduced by the generative process of the stimuli used by LAMCP. Interestingly the effective sample size for directional samples was  $15 \pm 3$  with an acceptance rate of  $75\% \pm 2\%$ .

Figure 1 shows the estimated golden ratio and distance of samples to the golden ratio produced by MCMCP and LAMCP. The top row of Figure 1 shows that participants us-

ing LAMCP quickly find the golden ratio and are easily able to explore and follow this correlation in the stimulus parameters, compared to MCMCP participants. The bottom row of Figure 1 shows that throughout the evolution of both Markov chains, LAMCP participants generate samples that are closer to the golden ratio than MCMCP participants.

We recorded the time between trials for each participant, and for LAMCP, compared the difference between these times for stimulus and direction trials. We could find no significant difference in these times under a variety of two-sample tests (Wilcoxon rank-sum, t-, and Kolmogorov-Smirnov tests;  $p > 0.05$ ,  $n = 500$ ). A likely explanation for this is that network transmission time dominates participants decision time: the median time between trials was 2.8 seconds ( $\pm 1$  second semi-interquartile range). Thus for Mechanical Turk experiments, using directional trials do not appear to take more of a participants' time than stimulus trials.

## Discussion

Look-Ahead Monte Carlo with People is an extension to MCMCP that exploits local manifold structure found in continuous valued stimuli by soliciting direction judgements from participants. This method will allow us to more efficiently explore and understand complicated categories in higher dimensions than previously attempted, by extracting more useful information per trial from participants. Whilst our simple experiment demonstrated the efficacy of our techniques in even the simplest case, for tasks such as images of faces, similar local structure likely exists in stimuli and so we can hope for similar gains.

Compared to other procedures for eliciting distributions from people, LAMCP's two kind of trial paradigm is also similar to iterated learning (Kirby, 1998; Griffiths & Kalish, 2007) where people either sample an internal representation or a physical manifestation of that representation. Directions are akin to an internal representation of what lies ahead, whilst the stimuli are the manifestations of the directions. The analogy is particularly apt if different people participate in each pair of LAMCP trials.

LAMCP not only produces samples from people's stimulus distribution, but also samples from their distribution over directions in stimulus space. Whilst our motivation for sampling from this distribution is purely incidental—we wish to obtain directions in some principled fashion so as to inform the stimulus generation process—the directions give a direct hint as to what people estimate to be the shape of the manifold in which samples lie. This extra piece of statistical information may be useful in gaining a better estimate of structure of stimuli, as well as aiding the estimation process.

**Acknowledgements:** We wish to thank Peter Dayan and Charles Sutton for several useful discussions.

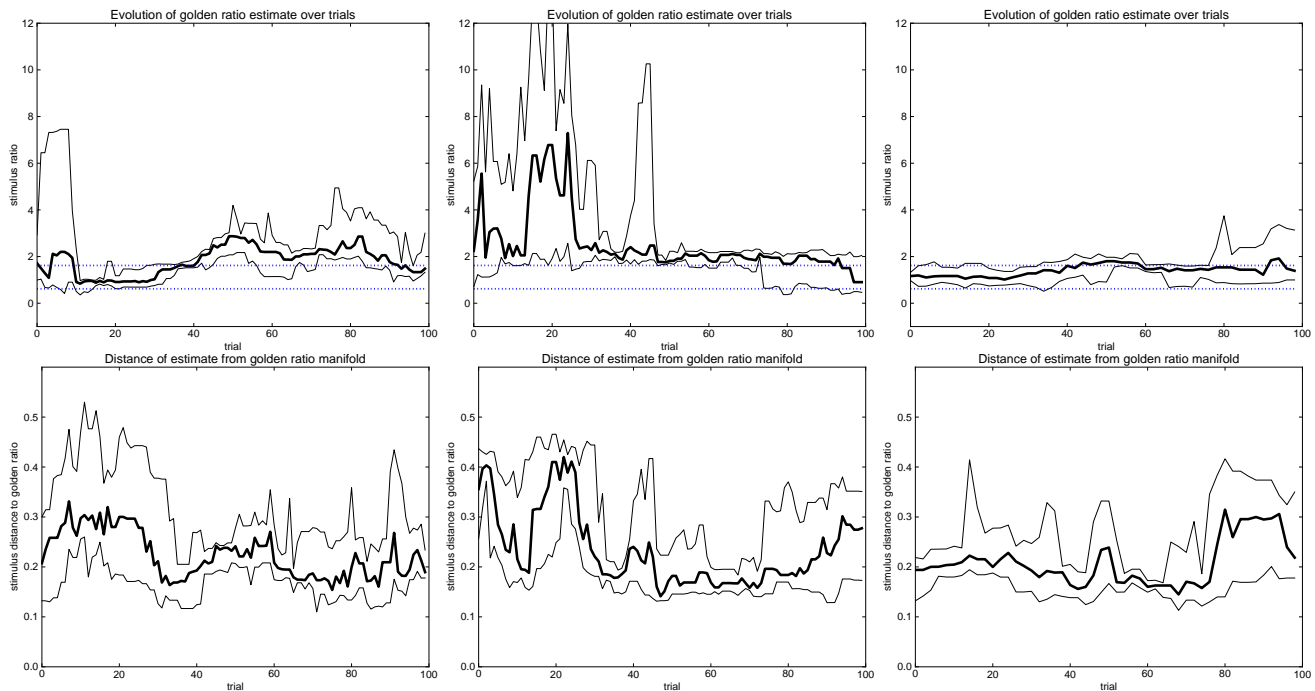


Figure 1: Estimates of the golden ratio (top) and distance of samples to the golden ratio manifold (bottom) under three regimes: MCMCP (left), MCMCP24 (middle), and LAMCP (right). Heavy solid lines correspond to the median over 10 chains, whilst lighter solid lines are the interquartile range. Dashed straight lines in the top plots correspond to the golden ratio (top) and its reciprocal (bottom).

## References

- Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39, 216-233.
- Barker, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18, 119-133.
- Belisle, C. E., Romeijn, H. E., & Smith, R. L. (1993). Hit-and-run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, 18(2).
- Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31, 441-480.
- Kirby, S. (1998). *Language evolution without natural selection: From vocabulary to syntax in a population of learners* (Tech. Rep.). Language Evolution and Computation Research Unit, University of Edinburgh.
- Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*, 1 (p. 103-189). Wiley.
- Martin, J. B., Griffiths, T. L., & Sanborn, A. N. (2012). Testing the efficiency of Markov Chain Monte Carlo with People using facial affect categories. *Cognitive Science*, 36, 150-162.
- Metropolis, A. W., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087-1092.
- Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Tech. Rep. No. CRG-TR-93-1). Department of Computer Science, University of Toronto.
- Neal, R. M. (1996). Bayesian learning for neural networks. *Lecture Notes in Statistics*, 118.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, 115(1), 39-57.
- Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnostics and output analysis for MCMC. *R News*, 6(1), 7-11.
- Rouder, J. N. (2004). Modeling the effects of choice-set size on the processing of letters and words. *Psychological Review*, 111, 80-93.
- Sanborn, A. N., & Griffiths, T. L. (2008). Markov chain Monte Carlo with people. In J. C. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems 20* (pp. 1265-1272). MIT Press.
- Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with Markov Chain Monte Carlo. *Cognitive Psychology*, 60, 63-106.
- Wills, A. J., Reimers, S., Stewart, N., Suret, M., & McLaren, I. P. L. (2000). Tests of the ratio rule in categorization. *The Quarterly Journal of Experimental Psychology*, 53(4), 983-1011.



# A Study of Loan Color Terms Collocation in Modern Japanese

Anna V. Bordilovskaya (firstwave.anna@gmail.com)

Graduate School of Humanities, Kobe University, 1-1 Rokkodai-machi,  
Nada-ku, Kobe 657-8501 JAPAN

## Abstract

The Japanese lexicon consists of Japanese-origin words (WAGO), Chinese-origin words (KANGO) and words borrowed from English and other European languages (GAIRAIGO). The acquisition of words from three sources results in the abundance of near synonyms without any clear rules when a particular synonym should be used. Loveday has hypothesized that WAGO/KANGO and GAIRAIGO concrete nouns are used to address similar phenomena of Japanese and Western origins, respectively. This is referred as Hypothesis of Foreign vs. Native Dichotomy (HFND). However, the matter of abstract nouns, adjectivals and their collocations remains unstudied. In contrast to the previous studies, based on questionnaires, our approach stems from statistical analysis of corpus data. Our results illuminate a distinguishable bias in the structure of collocations – nouns and adjectivals of the same origin tend to appear together more often than the ones of the different origins. Our finding implies extension of HFND to the level of collocations.

**Keywords:** cognitive linguistics, Modern Japanese, loanwords, corpus study

## Introduction

The structure of the Japanese lexicon is a very complex and rapidly developing system. It consists of Japanese-origin words (WAGO), Sino-(Chinese)origin words (KANGO) and words borrowed from English and other European languages (GAIRAIGO). The adoption of words from Chinese has a long history, while borrowing of GAIRAIGO loanwords intensified starting from Meiji period (1868 - 1912).

However, GAIRAIGO increased in number notably and became widely used only in the second half of the 20th century: “English words have become especially important since WWII, and these loanwords have become genuine parts of the Japanese lexicon, found in daily conversation and the world of letters.” (Daulton, 2008).

The acquisition of words from three sources (WAGO, KANGO and GAIRAIGO) results in the abundance of near synonyms. The differentiation between such synonyms in some cases is clear-cut. For example, there is a historically developed stylistic constraint for Sino-origin synonyms to be used mainly in written speech as opposing to Japanese-origin words, which are widely used in oral speech.

However, the differentiation between WAGO/KANGO and GAIRAIGO near synonyms in Modern Japanese is a more complicated issue.

English loanwords in Japanese have been a topic of various studies by both native and foreign linguists for about 100 years.

Some researchers are more interested in the assimilation processes of loanwords (Kay, 1995; Irwin, 2011), other linguists focus on semantic changes (Daulton, 2008), third mainly study sociolinguistic background and functions (Loveday, 1986, 1996).

At present, the number of GAIRAIGO is increasing rapidly and loanwords penetrate into different spheres of life. Dictionaries (Katakanago Jiten Consaizu (The Concise Dictionary of Katakana Words), etc.) in most cases do not state any clear differences in the meaning and usage for the abovementioned near synonyms.

On the other hand, the experience of studying and communicating in Japanese shows that it is not possible to substitute WAGO/KANGO and GAIRAIGO near synonyms freely.

Many of GAIRAIGO words are concrete nouns borrowed from Western culture that came to be used abundantly with the modernization of life in Japan. Along with concrete nouns, numerous abstract nouns and adjectives were also borrowed into the basic vocabulary. Loanwords integrated into the Japanese language had passed through difficult assimilation processes: orthographical, phonetic, grammatical changes etc.

The issue of loanwords increase and native words substitution has always been a topic not only of semantic but also of sociolinguistic research. The comprehensive results on this topic are presented by Loveday (1996). In this work, Loveday gives a historical overview starting with the contact of Japanese with the Chinese language and finishing by describing the role, types and functioning of English loans in Japanese mass media, etc.

According to Loveday (1996), GAIRAIGO words comprise about 10 per cent of the Modern Japanese lexicon. Although, the ratio of GAIRAIGO depends on the sphere of usage. GAIRAIGO are rarely used in official documents, political, religious, law lexicon. On the other hand, the percentage of English loanwords is higher in daily routine, fashion, mass media language: GAIRAIGO words “are particularly high in the areas of fashion, cosmetics, food, audio technology, sport, housing, music, art, business management, and engineering.” (Loveday, 1996).

Loveday (1996) considers that the occurrence of the synonymic WAGO-GAIRAIGO pairs is the result of the “Westernization of Japanese culture”. Such pairs exist “in semantic opposition where a word referring to a Western phenomenon is English-based and ‘complementary’ with a

word deriving from (Sino-) Japanese and referring to a related version of the phenomenon belonging to native culture".

Therefore, Loveday (1996) has hypothesized Foreign vs. Native Dichotomy: GAIRAIGO is "a word referring to a Western phenomenon", while WAGO/KANGO is "referring to a related version of the phenomenon belonging to native culture."

Thus, Modern Japanese has a clear-cut opposition between WAGO/KANGO and GAIRAIGO near synonyms for concrete nouns, i.e., GAIRAIGO concrete nouns are used to name foreign-originated phenomena, while (Sino-) Japanese words are used to name native-originated phenomena. Loveday provides the following examples: *tō/shōji* (= sliding door) – *doa* ('door'); *futon* (= quilted bedding) – *beddo* ('bed'), etc.

On the other hand, Loveday's study does not consider any abstract nouns and adjective-derived (adjectivals) loanwords. However, GAIRAIGO abstract nouns and adjectivals are also of extreme importance.

One of the central places among the adjectivals is occupied by the color terms. Japanese color nomenclature is not a new object for studying. There was some research on loan color terms in Modern Japanese as well.

Stanlaw (1997) stated that "the Japanese colour lexicon actually consists of two sets of mutually exclusive terms, one of native origin, the other borrowed from English". His research demonstrates that some of Japanese native color terms (such as MOMO-IRO (pink), DAIDAI-IRO (orange)) are replaced by English loanwords (PINKU and ORENJI respectively). He also suggested that the replacement took place "in reverse order to the Berlin and Kay evolutionary sequence"<sup>1</sup> (Stanlaw, 1997).

The major finding by Stanlaw is that in Modern Japanese some of the loan color terms are more salient in the minds of the Japanese than native color terms, where the substitution of native color terms for loanwords develops in a certain order.

The shortcoming of Stanlaw's study is that neither the semantic and stylistic differences between native and loan words nor constraints on color terms usage are discussed.

The layer of loan color terms in modern Japanese has been poorly investigated. At present the correct application of loan color terms for non-native speakers, who do not possess any implicit rules or so-called native speakers' linguistic sense of intuition, is a difficult task.

Considering the importance of color terms and absence of any explicit roles of application, we focus our study on illuminating the regularities which can provide a hint for such explicit rules.

### Extended Hypothesis

Loveday's hypothesis is easily applicable to concrete nouns. However, Loveday's research does not cover the

differences between GAIRAIGO derived from adjectives describing existing cultural phenomena and their WAGO/KANGO counterparts. For example, a phenomenon of 'white' color has already existed before borrowing GAIRAIGO adjectival HOWAITO, which is used to address the same color. The same is true for phenomena other than basic color terms. For instance, YANGU and WAKAI, which both refer to 'young' attribute.

In present study we focus on the descriptive role of adjectivals, therefore we consider the collocation of nouns and adjectivals in attributive positions. Consequently, in our study we focus not on the single words (nouns), but on collocations: pairs of nouns and adjectivals.

Our hypothesis is the extension of Loveday's Foreign vs. Native Dichotomy for collocations of nouns and adjectives/adjectivals. We assume that GAIRAIGO adjectivals are used for the description of foreign-originated phenomena and, thus, are more likely to collocate with GAIRAIGO nouns in attributive position. Similarly, WAGO/KANGO adjectives are used to describe the native-originated phenomena, that is, we expect WAGO/KANGO adjectives to be more inclined to collocate with WAGO/KANGO nouns rather than with GAIRAIGO nouns.

We call this hypothesis to be Extended Hypothesis of Foreign vs. Native Dichotomy (EHFND).

The important difference between Loveday's original hypothesis and our *extended hypothesis* is that the former one refers to the difference in usage of single words (concrete nouns), while the later one considers collocations. Therefore, our extended hypothesis deals with a more general linguistic level of collocation.

Thus, the problem is to differentiate the usage of collocations between GAIRAIGO and WAGO/KANGO near synonyms in Modern Japanese. To our knowledge, this matter has not been previously addressed.

## Methodology

### Objectives

Our primary objective is to test Extended hypothesis for the color terms in Modern Japanese. The secondary objective is to examine the matter of substitution of native color terms for the GAIRAIGO ones.

### Materials: Corpus-based analysis

The objective of the given research is to define the tendency of collocation of English adjectives borrowed and assimilated to Modern Japanese.

The number of WAGO/KANGO vs. GAIRAIGO near synonymic pairs is considerably numerous. Since we are interested in the presence of correlation between attributes and words they modify the object of research is GAIRAIGO adjectivals derived from English adjectives and Japanese words corresponding to them, we focus on basic color terms. For the purpose of the present study, we have selected eight loan basic color terms, which are frequently used in attributive position: HOWAITO, BURAKKU,

<sup>1</sup> Berlin and Kay (1991) speculate that as languages evolve, they develop new basic color terms in a strict chronological sequence.

REDDO, IERO:, BURAUN, PA:PURU, PINKU, GURE:/GUREI. They correspond to (Sino-) Japanese color terms SHIROI (white), KUROI (black), AKAI (red), KIIRO (yellow), CHAIRO (brown), MURASAKIIRO (purple), MOMOIRO (pink), HAIIRO (grey), respectively.

The reasons for choosing these color terms are

- all of the abovementioned GAIRAIGO color terms have long-term assimilation to the Japanese language (since Meiji period, more than 100 years);
- these GAIRAIGO color terms have a clear-cut correspondence to WAGO/KANGO near synonyms, that is why GURI:N (green) and BURU: (blue) were excluded from the study as their correspondence to native color terms (AOI and MIDORIIRO) is ambiguous;
- selected loanwords belong to one semantic category ‘basic color terms’, that is why they give a wider outlook of the Modern Japanese vocabulary.

## Method

The corpus-based analysis of GAIRAIGO collocation patterns is the main method of the analysis in the present study. The corpus used for data collection is KOTONOHA corpus (Balanced Corpus of Contemporary Written Japanese) available on-line<sup>2</sup>. This corpus was selected because it is considered to be a well-balanced corpus of Modern Japanese, including not only printed sources (such as newspapers, magazines and literature), but also web resources (such as Yahoo blogs and Yahoo answers).

KOTONOHA corpus was searched to compare the frequencies for four following collocation patterns: 1) attributive GAIRAIGO + NO + GAIRAIGO noun; 2) attributive GAIRAIGO + NO + WAGO/KANGO noun; 3) attributive WAGO/KANGO + (NO)<sup>3</sup> + GAIRAIGO noun; and 4) attributive WAGO/KANGO + (NO) + WAGO/KANGO noun.

The cases when selected GAIRAIGO were referring to a person’s name (like HOWAITO NO IKEN - ‘the opinion of Mr. White’), were eliminated from the search results. The idiomatic expressions like ‘blue blood’, ‘White House’, ‘black funds’, etc. have been excluded from the study as well.

## Results and Data Analysis

**Collocation Frequency Data.** The collocation frequency data for each color term are presented in Tables 1 through 8. The main means for statistic analysis are Chi-square test of independence of categorical data and Binomial test. However, if the conditions of Chi-square application were violated, Fisher’s exact test was used instead.

**HOWAITO vs. SHIROI pair.** The results on HOWAITO and SHIROI collocations are presented in Table 1.

Table 1: HOWAITO vs. SHIROI collocation.

	Gairaigo nouns (%)	Wago/kango nouns (%)	Total
Howaito	23 (77%)	7 (23%)	30 (100%)
Shiroi	47 (30%)	109 (70%)	156 (100%)
Total	70	116	186

The Chi-square test of independence of categorical data has been applied to the data in Table 1. The Chi-square test revealed statistically significant dependence between origin of adjective/adjectival and origin of noun ( $\chi^2(1, 186) = 23.218, p < 0.001$ ). On the other hand, the Binomial test for pairs (HOWAITO + GAIRAIGO nouns) vs. (HOWAITO + WAGO/KANGO nouns) reveals significant difference in frequencies ( $p < 0.05$ ). The Binomial test for pairs (SHIROI + GAIRAIGO nouns) vs. (SHIROI + WAGO/KANGO nouns) also reveals significant difference in frequencies ( $p < 0.001$ ). Therefore, for HOWAITO and SHIROI, there is obvious preference of WAGO/KANGO nouns vs. GAIRAIGO.

The Binomial test reveals significant difference ( $p < 0.001$ ) in frequencies of HOWAITO vs. SHIROI appearance – 30 vs. 156 hits, respectively. Therefore, SHIROI is used more often.

**BURAKKU vs. KUROI pair.** The results for BURAKKU and KUROI collocations are presented in Table 2.

Table 2: BURAKKU vs. KUROI collocations.

	Gairaigo nouns (%)	Wago/kango nouns (%)	Total
Burakku	14 (70%)	6 (30%)	20 (100%)
Kuroi	87 (23%)	295 (77%)	382 (100%)
Total	101	301	402

The Chi-square test of independence of categorical data has been applied to the data in Table 2. The Chi-square test revealed statistically significant dependence between origin of adjective/adjectival and origin of noun ( $\chi^2(1, 402) = 22.531, p < 0.001$ ). On the other hand, the Binomial test for pair (BURAKKU + GAIRAIGO nouns) vs. (BURAKKU + WAGO/KANGO nouns) reveals no significant difference in frequencies ( $p = 0.115 > 0.05$ ). However, the Binomial test for pair (KUROI + GAIRAIGO nouns) vs. (KUROI + WAGO/KANGO nouns) reveals significant difference in frequencies ( $p < 0.001$ ). Therefore, for KUROI there is obvious preference of WAGO/KANGO nouns vs. GAIRAIGO. Although the same tendency is developing for the BURAKKU, there is no statistical evidence to support our hypothesis.

<sup>2</sup> KOTONOHA Balanced Corpus of Contemporary Written Japanese <http://www.kotonoha.gr.jp/shonagon/>

<sup>3</sup> Some of the WAGO/KANGO in attributive position do not require NO-case marker

The Binomial test reveals significant difference ( $p < 0.001$ ) in frequencies of BURAKKU vs. KUROI appearance – 20 vs. 382 hits, respectively. Therefore, KUROI is used more often.

**REDDO vs. AKAI pair.** The results for REDDO and AKAI collocations are presented in Table 3.

Table 3: REDDO vs. AKAI collocations.

	Gairaigo nouns (%)	Wago/kango nouns (%)	Total
Reddo	2 (50%)	2 (50%)	4 (100%)
Akai	87 (26%)	247 (74%)	334 (100%)
Total	89	249	338

The Chi-square test of independence of categorical data is not applicable for the data from Table 3, since for two cells the expected count (2) less than 5. Therefore, Fisher's exact test has been applied. This test has reveals no significant dependencies ( $p = 0.284 > 0.05$ ).

In this case, the sample size for pairs including REDDO is only 4 occurrences. Therefore, due to small sample size the proper analysis of pairs (REDDO+GAIRAIGO nouns) vs. (REDDO+WAGO nouns) cannot be conducted. On the other hand, we have tested the frequencies of pairs (AKAI+GAIRAIGO nouns) vs. (AKAI+WAGO nouns). The Binomial test reveals statistically significantly difference ( $p < 0.001$ ). This result illustrates a bias towards preference of WAGO nouns to be used with AKAI.

The Binomial test reveals significant difference ( $p < 0.001$ ) in frequencies of REDDO vs. AKAI appearance – 4 vs. 334 hits, respectively. Therefore, AKAI is used more often.

**IERO: vs. KIIRO pair.** Next, we consider IERO: and KIIRO collocation data. The results are presented in Table 4.

Table 4: IERO: vs. KIIRO collocations.

	Gairaigo nouns (%)	Wago/kango nouns (%)	Total
Iero:	16 (67%)	8 (33%)	24 (100%)
Kiirō	136 (32%)	294 (68%)	430 (100%)
Total	152	302	454

The Chi-square test of independence of categorical data has been applied to the data in Table 4. The Chi-square test revealed statistically significant dependence between origin of adjective/adjectival and origin of noun ( $\chi^2(1, 454) = 12.531, p < 0.001$ ). On the other hand, the Binomial test for pair (IERO: + GAIRAIGO nouns) vs. (IERO: + WAGO/KANGO nouns) reveals no significant difference in frequencies ( $p = 0.152 > 0.05$ ). However, the Binomial test for pair (KIIRO + GAIRAIGO nouns) vs. (KIIRO + WAGO/KANGO nouns) reveals significant difference in frequencies ( $p < 0.001$ ). Therefore, for KIIRO there is obvious preference of WAGO/KANGO nouns vs. GAIRAIGO. Although the same tendency is developing for

IERO:, there is no statistical evidence to support our hypothesis. The Binomial test reveals significant difference ( $p < 0.001$ ) in frequencies of IERO: vs. KIIRO appearance – 24 vs. 430 hits, respectively. Therefore, KIIRO is used more often.

**BURAUN vs. CHAIRO pair.** The results for BURAUN and CHAIRO collocations are presented in Table 5.

Table 5: BURAUN vs. CHAIRO collocations.

	Gairaigo nouns (%)	Wago/kango nouns (%)	Total
Buraun	26 (65%)	14 (35%)	40 (100%)
Chairo	110 (28%)	289 (72%)	399 (100%)
Total	136	303	439

The Chi-square test of independence of categorical data has been applied to the data in Table 5. The Chi-square test revealed statistically significant dependence between origin of adjective/adjectival and origin of noun ( $\chi^2(1, 439) = 23.822, p < 0.001$ ). On the other hand, the Binomial test for pair (BURAUN + GAIRAIGO nouns) vs. (BURAUN + WAGO/KANGO nouns) reveals no significant difference in frequencies ( $p = 0.081 > 0.05$ ). However, the Binomial test for pair (CHAIRO + GAIRAIGO nouns) vs. (CHAIRO + WAGO/KANGO nouns) reveals significant difference in frequencies ( $p < 0.001$ ). Therefore, for CHAIRO there is obvious preference of WAGO/KANGO nouns vs. GAIRAIGO. Although the same tendency is developing for BURAUN, there is no statistical evidence to support our hypothesis.

The Binomial test reveals significant difference ( $p < 0.001$ ) in frequencies of BURAUN vs. CHAIRO appearance – 40 vs. 399 hits, respectively. Therefore, CHAIRO is used more often.

**PA:PURU vs. MURASAKIIRO pair.** Table 6 presents the results for PA:PURU and MURASAKIIRO collocations.

Table 6: PA:PURU vs. MURASAKIIRO collocations.

	Gairaigo nouns (%)	Wago/kango nouns (%)	Total
Pa:puru	12 (71%)	5 (29%)	17 (100%)
Murasakiirō	50 (24%)	161 (74%)	211 (100%)
Total	62	166	228

The Chi-square test of independence of categorical data is not applicable for the data from Table 6, since for one cell the expected count equals 5. Therefore, Fisher's exact test has been applied. This test has revealed significant dependencies ( $p < 0.001$ ).

On the other hand, the Binomial test for pair (PA:PURU + GAIRAIGO nouns) vs. (PA:PURU + WAGO/KANGO nouns) reveals no significant difference in frequencies ( $p = 0.143 > 0.05$ ). However, the Binomial test for pair (MURASAKIIRO + GAIRAIGO nouns) vs.

(MURASAKIIRO + WAGO/KANGO nouns) reveals significant difference in frequencies ( $p < 0.001$ ). Therefore, for MURASAKIIRO there is obvious preference of WAGO/KANGO nouns vs. GAIRAIGO. Although the same tendency is developing for PA:PURU, there is no statistical evidence to support our hypothesis.

The Binomial test reveals significant difference ( $p < 0.001$ ) in frequencies of PA:PURU vs. MURASAKIIRO appearance – 17 vs. 211 hits, respectively. Therefore, MURASAKIIRO is used more often.

**GUERE:/GUEREI vs. HAIIRO pair.** The search results for GUERE:/GUEREI and HAIIRO collocations are presented in Table 7.

Table 7: GUERE:/GUEREI vs. HAIIRO collocations.

	Gairaigo nouns (%)	Wago/kango nouns (%)	Total
Gure:/Gurei	207 (60%)	136 (40%)	343 (100%)
Haiiro	66 (23%)	225 (74%)	291 (100%)
Total	273	361	634

The Chi-square test of independence of categorical data has been applied to the data in Table 7. The Chi-square test revealed statistically significant dependence between origin of adjective/adjectival and origin of noun ( $\chi^2(1, 634) = 91.114, p < 0.001$ ). On the other hand, the Binomial test for pairs (GUERE:/GUEREI + GAIRAIGO nouns) vs. (GUERE:/GUEREI + WAGO/KANGO nouns) reveals significant difference in frequencies ( $p < 0.001$ ). The Binomial test for pairs (HAIIRO + GAIRAIGO nouns) vs. (HAIIRO + WAGO/KANGO nouns) also reveals significant difference in frequencies ( $p < 0.001$ ). Therefore, for both GUERE:/GUEREI and HAIIRO, there is obvious preference of WAGO/KANGO nouns vs. GAIRAIGO.

The Binomial test reveals significant difference ( $p = 0.043 < 0.05$ ) in frequencies of GUERE:/GUEREI vs. HAIIRO appearance – 343 vs. 291 hits, respectively. Therefore, GUERE:/GUEREI is used more often.

**PINKU vs. MOMOIRO pair.** Finally, the results for PINKU and MOMOIRO collocations are presented in Table 8.

Table 8: PINKU vs. MOMOIRO collocations.

	Gairaigo nouns (%)	Wago/kango nouns (%)	Total
Pinku	232 (58%)	169 (42%)	401 (100%)
Momoiro	15 (35%)	28 (65%)	43 (100%)
Total	247	197	444

The Chi-square test of independence of categorical data has been applied to the data in Table 8. The Chi-square test revealed statistically significant dependence between origin of adjective/adjectival and origin of noun ( $\chi^2(1, 444) = 8.303, p < 0.001$ ). On the other hand, the Binomial test for pairs (PINKU + GAIRAIGO nouns) vs. (PINKU +

WAGO/KANGO nouns) reveals significant difference in frequencies ( $p < 0.001$ ). The Binomial test for pairs (MOMOIRO + GAIRAIGO nouns) vs. (MOMOIRO + WAGO/KANGO nouns) also reveals significant difference in frequencies ( $p < 0.001$ ). Therefore, for PINKU and MOMOIRO, there is obvious preference of WAGO/KANGO nouns vs. GAIRAIGO.

The Binomial test reveals significant difference ( $p < 0.001$ ) in frequencies of PINKU vs. MOMOIRO appearance – 401 vs. 43 hits, respectively. Therefore, PINKU is used more often.

## Discussion and Conclusion

In present study we have introduced a new approach to differentiation between (Sino-)Japanese and English-origin near synonyms. In contrast to the previous studies, which were based on questionnaires, our approach stems from statistical analysis of corpus data.

We have shown that there is an obvious bias in the structure of collocations: nouns and adjectives/adjectivals of the same origin (WAGO/KANGO or GAIRAIGO) tend to appear together more often than nouns and adjectives/adjectivals of different origins. In all eight cases considered, we have found statistical evidence for such bias to exist for WAGO/KANGO adjectives. For GAIRAIGO adjectivals, in the case of HOWAITO, GUERE:/GUEREI and PINKU it was also possible to support this assumption with statistical evidence. In the case of REDDO, the sample size was too small. In five remaining cases, there was no statistical evidence, but the same tendency of preferring GAIRAIGO nouns against WAGO/KANGO nouns by GAIRAIGO adjectivals can be observed.

In general, for seven pairs out of eight, except for REDDO vs. AKAI, we have found statistical evidence for dependencies in categorical data.

Summarizing, we consider that this volume of evidence is enough to support our Extended hypothesis, derived from original Loveday's Foreign vs. Native Dichotomy (Loveday, 1996). On the other hand, our hypothesis refers to the structure consisting of two words, i.e. adjectives/adjectivals plus nouns, while Loveday (Loveday, 1996) has investigated only concrete GAIRAIGO and WAGO/KANGO nouns referring either to foreign or native objects, respectively.

Therefore, we illustrated an existence of Foreign vs. Native Dichotomy at the level of collocations.

We have also compared frequencies of the appearance of color terms. Our results show that among the selected pairs of color terms, there are both cases when native color terms prevail the borrowed ones and vice versa (GUERE:/GUEREI and PINKU).

The results for GUERE:/GUEREI and PINKU are in coherence with the Stanlaw's assumption of the replacement of native words with borrowed ones (Stanlaw, 1997). According to Berlin and Kay (1991), the color terms appear in language in the hierarchical order starting from more general terms covering white, black, red color categories

and ending with more specific color categories like pink, orange, grey and purple. Stanlaw (1997) suggests that the replacement of native color terms starts from the end of the Berlin and Kay's hierarchy (Berlin and Kay, 1991). That is colors like GURE:/GUREI and PINKU should be substituted earlier than other color terms. The results of our statistical analysis indicate that GURE:/GUREI and PINKU are used more often than corresponding native color terms. This finding does not provide exhaustive evidence for replacement dynamics, but we consider that the more frequent usage of one color term (loanword) against another one (native color term) can be regarded as the indication of ongoing process of substitution in accordance with the results of Stanlaw's study (Stanlaw, 1997).

The corpus-based method employed for the data collection is different from Loveday's approach (Loveday, 1996). We consider that this method offers a more profound quantitative analysis of the gathered data. To our knowledge, such approach to the analysis of loanword collocations has not yet been implemented elsewhere.

On the other hand, the results of the present research are in coherence with both Loveday's (Loveday, 1996) and Stanlaw's approaches (Stanlaw, 1997).

Therefore our method exhibits successful application for two different tasks, thus our method symbiotically amalgamates abilities and power of inference of the previously considered research approaches.

Many native speakers cannot explain a particular rule to use either of near synonyms. According to Dienes and Berry (1997), implicit learning means that information is acquired without intension and the resulting knowledge is difficult to express. Therefore, native-like linguistic sense of usage of near synonyms can be considered as an implicit knowledge possessed by native speakers. On the other hand, our statistical data uncovers essential regularities in usage of near synonyms. We assume these regularities to be a part of implicit language knowledge. Thus, the approach for analysis of near synonyms has a potential to connect the pure linguistic study of collocation with field of implicit learning and knowledge. Therefore, we consider that the results of our study can be used as a starting point for constructing the explicit rules to articulate the native-like linguistic sense (implicit knowledge).

For the purpose of our study, only the group of basic color terms has been employed. We have provided evidence that Extended Hypothesis of Foreign vs. Native Dichotomy is likely to be true for such data. However, to prove that this tendency is general among GAIRAIGO more data on various loanwords needs to be analyzed. On the other hand, corpus data represents only a part of the whole scope of Modern Japanese.

To improve the quality of analysis it is important to integrate native speakers' introspection-based analysis. Thus, two main directions for further research are increasing data set and extending methods of analysis.

## References

- Berlin, B. and Kay, P. (1991). *Basic Color Terms: Their universality and evolution*. Berkley & Los Angeles: University of California Press.
- Daulton, F.E. (2008). *Japan's built-in lexicon of English-based loanwords*. Clevedon: Multilingual Matters Ltd.
- Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin & Review*, 4, 3-23.
- Hardin, C.L., & Maffi, L. (1997). *Color categories in thought and language*. Cambridge: Cambridge University Press.
- Irwin, M. (2011). *Loanwords in Japanese*. Amsterdam: John Benjamins Publishing Company.
- Katakanago Jiten Consaizu daisanhan. 2005. The Concise Dictionary of Katakana Words, Third edition. Sanshou-douhenshuujo.
- Kobayashi, S. (1975). *Nihonjin no kokoro to iro*. (The Soul of the Japanese and Colors). Kodansha. (in Japanese)
- Loveday, L.J. (1986). *Explorations in Japanese Sociolinguistics*. Amsterdam: John Benjamins Publishing Company.
- Loveday, L.J. (1996). *Language contact in Japan: a sociolinguistic history*. Oxford: Clarendon Press.
- Nagata, Ya. (2002). *Iro no techo. Iromihon to bunkenrei de tsudzuru shikimeji gaido*. (Color notebook. The Guide of Color Terms Illustrated with Color samples and Literature Examples). Shogakkan. (in Japanese)
- Rossotti, H. (1983). *Colour. Why the world isn't grey*. Princeton, NJ: Princeton University Press.
- Sato, Ch. (2008). *Shikisai gaidobukku*. (Color Guidebook). Nagaokashoten. (in Japanese)
- Stanlaw, J. (1997). Two observations on culture contact and the Japanese color nomenclature system. In C.L. Hardin & L. Maffi (Eds.), *Color categories in thought and language*. Cambridge: Cambridge University Press.
- Stanlaw, J. (2004). *Japanese English: Language and culture contact*. Hong Kong: Hong Kong University Press.

# Intelligibility is Necessary for Scientific Explanation, but Accuracy May Not Be

**Mike Braverman (braverm2@illinois.edu)**

Department of Psychology, 603 E. Daniel Street  
Champaign, IL 61820 USA

**John Clevenger (jcleveng2@illinois.edu)**

Department of Philosophy, 810 S. Wright Street  
Urbana, IL 61801 USA

**Ian Harmon (iharmon2@illinois.edu)**

Department of Philosophy, 810 S. Wright Street  
Urbana, IL 61801 USA

**Andrew Higgins (higgins9@illinois.edu)**

Department of Philosophy, 810 S. Wright Street  
Urbana, IL 61801 USA

**Zachary S. Horne (horne2@illinois.edu)**

Department of Philosophy, 810 S. Wright Street  
Urbana, IL 61801 USA

**Joseph Spino (spino2@illinois.edu)**

Department of Philosophy, 810 S. Wright Street  
Urbana, IL 61801 USA

**Jonathan Waskan (waskan@illinois.edu)**

Department of Philosophy, 810 S. Wright Street  
Urbana, IL 61801 USA

## Abstract

Many philosophers of science believe that empirical psychology can contribute little to the philosophical investigation of explanations. They take this to be shown by the fact that certain explanations fail to elicit any relevant psychological events (e.g., familiarity, insight, intelligibility, etc.). We report results from a study suggesting that, at least among those with extensive science training, a capacity to render an event intelligible is considered a requirement for explanation. We also investigate for whom explanations must be capable of rendering events intelligible and whether or not accuracy is also viewed as a requirement.

**Keywords:** science; explanation; psychologism; intelligibility.

## Introduction

The nature of explanation has been a major topic of investigation in the philosophy of science for at least sixty years. While much is still disputed, a consensus has emerged that scientific explanations are not constituted by psychological events. Philosophers of science first made arguments to this effect in the middle of the 20th century in response to the charge that explanations play no useful role in science because they are constituted by subjectively variable psychological states (viz., familiarity and empathy).

Hospers (1946), Miller (1947), and Hempel (1965), for instance, argued that many legitimate explanations appeal to principles (e.g., Newton's law of gravitation) that were utterly unfamiliar when they were first introduced. Hempel (1942) also noted that historical explanations sometimes refer to individuals (e.g., paranoiacs) with whom most are incapable of empathizing. On such grounds Hempel famously rejects psychologistic theories of explanation in favor of a more "objective" account (1965, 426).

More recently, philosophers of science have viewed psychologistic theories as equating explanations with other feelings, such as insight, satisfaction, or "aha" feelings (Craver 2007; Salmon 1984; Trout 2007). Like Miller and Hempel, they rebut such psychologistic proposals by pointing to cases of explanation where the relevant feelings are absent. One common strategy is to point to a putative explanation, whether it be a passage of text describing a process of speciation (Trout 2007) or a computer simulation of the human nervous system (Craver 2007), whose complexity so outstrips the limits of human memory and attention that humans find it incomprehensible. Humans thus fail to derive from these putative explanations any feelings of insight, satisfaction, etc. It is frequently concluded on the basis of such examples that explanations are non-psychological - we term this the *objectivity*



hypothesis - and hence that psychological research will contribute little to our understanding of explanations.<sup>1</sup>

We believe that this anti-psychologistic attitude is wrongheaded for a few reasons. One is that the very process of justifying theories of explanation on the basis of how well they track philosophers' classifications (i.e., as explanations or non-explanations) of various representations are laden with psychological presuppositions. Philosophers appear to assume that they have, through their exposure to science and scientists, come to possess tacit knowledge of the norms regarding the proper use of 'explanation.' Whether or not philosophical judgments mirror scientific ones is, however, an empirical matter that is best resolved through psychological investigation. We suspect, moreover, that scientists do regard a certain kind of conscious psychological event as necessary for explanation. The type of event we have in mind is not an affect-laden *feeling*, but rather the more intellectual process of understanding how or why, at least possibly, an event came about. This state is sometimes known as finding a happening intelligible or making sense of it (cf. Machamer & Woody 1994). Familiar objections have been raised against this proposal as well - namely, that there are (e.g., hyper-complex) explanations that never render anything intelligible to anyone. Our specific hypothesis, however, is that scientists will not consider a representation an explanation unless it has the *capacity* to render intelligible, which we term *intelligibility*. It would thus not undercut our position if there were explanations that never actually render the event intelligible to anyone. Consider, by comparison, that a liquid may be a solvent of salt even if at a given time it happens not to be dissolving any salt at all or even if it never does so (viz., because the opportunity never arises). It must merely have the capacity to do so, even if it is not exercised. Likewise, we suspect that scientists require not that a representation actually renders a happening intelligible to someone, but merely that it has the capacity to do so.

Notice that even if a liquid that has the capacity to dissolve salt never exercises that capacity, it would still tell us a lot about what makes that liquid a solvent of salt if we had information about the process by which it would dissolve salt were the opportunity to arise. This is best accomplished by studying cases in which solvents of salt actually do exercise their capacity to dissolve salt. Matters are somewhat more complicated with regard to explanations, but we believe that even if a certain representation that has the capacity to make things intelligible to people never actually exercises that capacity, it may well tell us a lot about what makes that representation an explanation if we had information about the process by which it would make sense of things were the opportunity to

arise. This would be best accomplished by studying cases in which explanations actually do exercise their intelligibility.

Notice also that if the reason why a liquid has failed to dissolve any salt is that there is something about the nature of the liquid that precludes it from ever dissolving any salt, the liquid then lacks even the capacity to dissolve salt and is thus no solvent of it. Likewise, if there is something about a representation that precludes it from ever rendering an event intelligible - for instance, it is too complex to ever be comprehended by anyone at all or it refers to things like extra dimensions that are utterly incomprehensible to anyone - then it lacks even the capacity to render intelligible and scientists will thus not regard it as an explanation. We suspect that laypeople also view intelligibility as a requirement and will classify such cases in a similar way.

If a representation must have the capacity to render intelligible in order for it to be considered an explanation, a further question naturally arises: Intelligible to whom? Scientists generally interact with at least some colleagues who exceed their own intelligence, so they are almost certainly cognizant of the fact that some representations are too complex to make events intelligible to everyone. Thus, they likely do not require that a representation must have the capacity to render things intelligible to just anyone in order to be considered an explanation. However, since scientists presumably do not interact with beings that possess utterly different perceptual and cognitive abilities, they might think that a representation must have the capacity to make things intelligible to beings basically like themselves in order to be considered an explanation. Then again, they may turn out to be even more liberal, merely requiring that a representation makes sense of things for sentient beings of some sort, even ones with completely alien thought processes. There are numerous possibilities here, with regard to both scientists and laypeople, that are worth investigating.

Another widely accepted view among philosophers of science is that a high degree of accuracy is essential in order for a representation to be an explanation (Craver 2007; C.G. Hempel 1965; Humphreys 1989; Salmon 1998; Trout, 2007). We call this the *accuracy* hypothesis. We think a representation must merely specify a possible way, even if it is not the actual way, in which an event occurred for the representation to be considered an explanation. We call this the *plausibility* hypothesis. An implication of the accuracy hypothesis is that scientists will be less likely to judge that a representation is an explanation if they are told that it is merely possibly accurate (which is to say that it might be inaccurate) or that it once seemed possibly accurate but was later falsified. Cases of the latter sort mirror many cases from this history of science where a theory seemed to make sense of things (e.g., the Ptolemaic theory of planetary retrograde motion) but was ultimately proven false. If scientists and laypeople do regard such cases as explanations, telling them that a representation is merely possibly accurate or conceivably accurate but factually

---

<sup>1</sup> Trout, admittedly, seems in places to be opposed to the idea that explanations are constituted by *conscious* psychological states in that he allows that explanations sometimes involve implicit learning.

inaccurate should not undermine their tendency to regard a representation as an explanation.

We tested whether people treated either intelligibility or accuracy as a necessary condition for something to be an explanation. We tested this with regard to two groups of participants: those with and without extensive science training.

## Methods

### Participants

The participants in this study were 297 workers recruited using Amazon Mechanical Turk with varying levels of science training.

### Materials

We used stories about the origins of either life, color experiences, or gamma ray bursts, and manipulated the characteristics of these stories. Vignettes were grouped into following three sets:

Set 1: A potential explanation (viz., a passage of text) is described that has various theoretical virtues and that also supplies some level of understanding. Specifically, the representation either supplies understanding of how the target happening actually occurred (A), of a possible way in which it occurred (PA), or of a possible way in which it occurred that is eventually shown to be false (PAF).

Set 2: A potential explanation is described which is said to be incapable of rendering a happening intelligible to humans because of our cognitive limitations - that is, either the representation defies the limits of normal human working memory and attention or it refers to highly exotic, hyper-dimensional properties. The possibility is then introduced that the representation would render the happening intelligible to beings with cognitive capacities that are, roughly speaking, quantitatively better (i.e., involving augmented memory and attention span) or qualitatively different (i.e., involving the ability to think in extra dimensions). Cases of the former sort involve hyper-complex passages of text (AM). Cases of the latter sort (HD) refer to passages of text that describe hyper-dimensional properties.

Set 3: This set closely mirrors Set 2 except that the complexity and exotic nature of the putative explanation is such that it precludes the representation from rendering the target intelligible to anyone at all. This set includes passages of text (IQuant) or computer simulations (Sim) that are hyper-complex, and passages of text that refer to exotic properties (IQual).

There were three storylines and nine vignette types, yielding twenty-seven possible vignettes. Each participant read and responded to three vignettes. One vignette came from each set, and each followed a different storyline.

### Collection

Participants were recruited through the Amazon Mechanical Turk (MTurk) work-distribution website. To be

eligible, workers had to be in the U.S. and have at least a 75% approval rate. Eligible workers were redirected to SurveyMonkey, where they completed the study. Afterwards, workers were directed back to MTurk, where they were compensated with \$.50.

### Procedure

Participants were presented with a vignette (story) on a computer screen. Each vignette referred either to a typewritten description of a physical process or to a computer simulation thereof. For example, some participants saw the following vignette<sup>2</sup>:

Dr. Nikro is a little-known, very-gifted scientist investigating the manner in which color experiences arise in the brain. He spends years tinkering with the complex equations of neurochemistry and subatomic physics and considering the different possible locations in which color experiences might originate. He eventually chances upon a remarkable series of calculations, which he posts to his rarely visited public webpage. They indicate that conditions like those found in the pyramidal cells of the cortex would, over the course of hundreds of milliseconds, reliably undergo a series of changes resulting in the creation of color experiences. The calculations fit quite nicely with the most widely accepted theories and observations from a variety of related fields. Yet they merely refer to ordinary things like membranes and neural firings, of which anyone could easily conceive. As a result, Dr. Nikro comes to understand the actual manner in which color experiences are generated in the brain, as would any specialist from his field who were to study the details of the material he posted.

After reading the vignette, participants were asked to rate the extent to which they agree (-3 strongly disagree, 0 neutral, +3 strongly agree) that the representation described in the vignette constitutes an explanation. There was some concern that participants might confuse this question with the question of whether or not the representation constitutes a *good* or a *satisfying* explanation. Participants were thus informed at the outset that, insofar as they do agree that the representation constitutes an explanation, they should also specify the extent to which they agree that the representation constitutes a *good* and a *satisfying* explanation (cf. Lombrozo and Carey 2004).

Mishra & Brewer (2003) found that participants paid closer attention to the contents of their study materials (i.e., real-world explanations) if they were forewarned that they would be asked a series of simple comprehension questions. We likewise informed participants in advance that they would be asked a series of simple questions. Some of these

---

<sup>2</sup> A demonstration version of the study can be found at [www.surveymonkey.com/s/PPG6DW6](http://www.surveymonkey.com/s/PPG6DW6).

concerned aspects of the vignettes (e.g., accuracy or intelligibility) that were particularly salient to the experiment.

This sequence of instructions, vignettes, and questions repeated two more times, each time with vignettes from a different set and on a different topic (i.e., life, color experiences, or gamma-ray bursts). Participants completed a distractor task in between each vignette. At the end of the study, participants were asked some follow-up questions, including questions about their level of science training.

Six separate studies were constructed to balance the order in which participants saw the various storylines, with one study for each permutation (i.e., Life-Color Experiences-Gamma Rays, Life-Gamma Rays-Color Experiences, etc.).

## Results

Of the 297 completed studies, data from 38 of those studies were excluded because the participants involved had already completed the study at least once. After this exclusion there were 259 completed studies or 777 sets of responses to particular vignettes. If a participant failed to answer at least two of three comprehension questions correctly for a particular vignette, responses to that vignette were excluded. Using this criterion, 61 individuals had responses to one vignette eliminated, 9 individuals had responses to two vignettes eliminated, and 5 individuals had responses to all three vignettes eliminated. After eliminating these problematic responses, responses to 683 vignettes were recorded and analyzed.

We divided our sample into high-science (5 or more college-level science courses) and low-science (less than 5 college-level science courses) groups to investigate the ways in which laypeople and scientists conceive of explanations. After the division, the high-science group contained the following number of responses to the ‘constitutes an explanation’ question for each of the following conditions: 26 accurate (A), 32 possibly accurate (PA), 22 possibly accurate but false (PAF), 28 augmented memory (AM), 28 hyper-dimensional thinking (HD), 22 unintelligible-quantitative (IQuant), 30 unintelligible-qualitative (IQual), and 27 unintelligible-quantitative simulation (SIM).

The low-science group contained the following number of responses to the ‘constitutes an explanation’ question for each of the following conditions: 37 accurate (A), 58 possibly accurate (PA), 51 possibly accurate but false (PAF), 57 augmented memory (AM), 62 hyper-dimensional thinking (HD), 21 unintelligible-quantitative (IQuant), 71 unintelligible-qualitative (IQual), and 59 unintelligible-quantitative simulation (SIM).

The mean responses and standard deviations to the explanation rating task are located in Tables 1 and 2.

Table 1: Mean Responses and Standard Deviations for High-Science Participants

	A	PA	PAF	AM	HD	IQuant	IQual	SIM
Avg.	.769	.968	.090	.285	-.357	-.727	-1.1	-.370
SD	1.77	1.80	2.02	1.95	1.66	1.72	1.78	1.94

Table 2: Mean Responses and Standard Deviations for Low-Science Participants

	A	PA	PAF	AM	HD	IQuant	IQual	SIM
Avg.	1.29	1.31	-.137	.350	-.241	-.714	-.859	-.610
SD	1.66	1.42	1.78	1.67	1.64	1.87	1.87	1.65

After the division of the sample, mean ratings were measured for differences using 18 two-tailed independent samples *t*-tests. To guard against increased chance of Type I errors, we performed a Bonferroni correction to adjust  $\alpha$  (significant results were  $p < .0027$ ).

In the low-science group, descriptions – be they written text or simulations – that were unintelligible to anyone, resulted in lower ratings than descriptions that were intelligible to the scientists in the vignettes (A and IQuant ( $t(56) = 4.22, p < .0027$ ), A and IQual ( $t(106) = 5.87, p < .0027$ ) and A and SIM ( $t(94) = 5.48, p < .0027$ )). Low-science participants also gave lower ratings when *only* cognitively advanced beings – whether they be humans with augmented memory or beings capable of perceiving extra dimensions – found the descriptions intelligible (PA and AM ( $t(113) = 3.30, p < .0027$ ), PA and HD ( $t(118) = 5.49, p < .0027$ )). Significance obtained when comparing low-science participants’ ratings of descriptions described as false to descriptions described as accurate or possibly accurate (A and PAF ( $t(86) = 3.82, p < .0027$ ), PA and PAF ( $t(107) = 4.68, p < .0027$ )). Whether a description was accurate or possibly accurate, however, made no difference to low-science participants’ judgments ( $t(93) = 41, p = .96$ ).

For the high-science group, descriptions incapable of rendering the event intelligible to anyone resulted in lower ratings than those that made the event intelligible to the scientists in the vignettes (A and IQual ( $t(54) = 3.911, p < .0027$ )). However, these differences did not hold across all Set 3 vignettes (A and IQuant ( $t(46) = 2.95, p = .004$ ), A and Sim ( $t(51) = 2.25, p = .028$ )). For high-science participants, ratings were not significantly different for descriptions described as false, or only possibly accurate, as compared to those described as accurate (A and PAF ( $t(46) = 1.24, p = .222$ ), A and PA ( $t(56) = .422, p = .675$ ), PA and PAF ( $t(52) = 1.67, p = .10$ )). Likewise, these participants’ ratings were not significantly affected when a description was described as intelligible only to beings with quantitatively or qualitatively different cognitive capacities (PA and AM ( $t(58) = 1.40, p = .16$ ), PA and HD ( $t(58) = 2.94, p = .004$ )).

## Discussion of Results

This study sought to answer the following questions:

1. Does intellig-ability matter? If the objectivity hypothesis captures how laypeople and practicing scientists conceive of explanations, then altering whether a representation is described as capable of rendering a happening intelligible should not affect the judgments of either low-science or high-science participants. If the

intellig-ability hypothesis is correct, participants should be less likely to regard a representation as an explanation when told that it is, whether due to sheer complexity or to exotic constructs, incapable of rendering a happening intelligible to anyone. Low-science participants were less likely to judge a representation to be an explanation when told that the representation lacks intellig-ability. The judgments of low-science participants thus indicate that laypeople conceive of explanations in a way consistent with the intellig-ability hypothesis and inconsistent with the objectivity hypothesis.

High-science participants were less likely to regard a passage of text as an explanation when told that the representation, because of qualitative barriers (IQual), lacks intellig-ability. They were, however, not significantly less likely to regard a passage of text (IQual) or a simulation (SIM) as an explanation when told that the representation, because of quantitative barriers, lacks intellig-ability. One possibility is that quantitative barriers to intelligibility seem far less daunting than qualitative ones. Indeed, those with extensive science training should be well aware that myriad techniques have been developed for analyzing and visualizing information regarding the behaviors of complex systems with the precise point of rendering those systems intelligible to humans. This may have had enough of an impact on judgments to weaken the intellig-ability effect. Regardless, the differences in high-science ratings between (PA), on the one hand, and (IQuant) and (SIM) on the other were large enough to warrant further investigation.

The difference between (A) and (IQual) is, by itself, clearly inconsistent with the objectivity hypothesis. Admittedly, we cannot be sure that the judgments of our high-science participants mirror those of professional scientists, so further study of practicing scientists is also needed to better support our proposal that intellig-ability is a requirement for explanation. However, taken as a whole, this set of results strongly suggests that manipulating intellig-ability alters explanation judgments, and we take this to be inconsistent with the objectivity hypothesis.

2. Intellig-ability to Whom? The results above indicate that people treat intellig-ability as in some way necessary for explanation, but what is not clear is for whom a representation must be intellig-able (i.e., able to make sense of things). To address this question, participants were asked to consider cases in which potential explanations are unintellig-able to humans because of cognitive limitations. In the (AM) case, the representation exceeds the limits of normal human working memory and attention. In another case (HD), the representation refers to exotic properties such as hyper-dimensionality (see Set 2 in Materials). The possibility is then introduced that the representation may be intellig-able to beings with, roughly speaking, quantitatively augmented memory and attention or qualitatively different cognitive capacities. We take such beings to lie along a similarity continuum such that creatures who merely have augmented memory and attention are more similar to present-day humans than creatures with qualitatively different cognitive capacities.

As it turns out, relative to judgments regarding (PA) vignettes, low-science participants are less inclined to regard a description as an explanation in the (AM) and (HD) conditions. Relative to that same baseline (PA), there was no indication that high-science participants are less inclined to regard a description as an explanation in the (AM) condition. Nor were high-science participants significantly less likely to regard a description as an explanation in the (HD) condition than in the (PA) condition. Although the difference in judgments is not significant ( $p = .004$ ), it is suggestive enough to warrant further investigation.

There are a number of ways of interpreting these findings. Starting with low-science individuals, one possibility is that they believe representations must be intellig-able to present-day humans. This condition was satisfied in the (PA) vignettes but not in the Set 2 (i.e., (AM) or (HD)) vignettes. Another possibility is that these individuals' expectations are somewhat more flexible, requiring merely that representations be intellig-able to beings fairly similar to present-day humans. It may just be that they consider the differences between present-day humans and the beings described in the Set 2 vignettes to be too stark to meet this condition. It is also possible that the reason why they were less inclined to regard the representations in the Set 2 vignettes as explanations had nothing to do with intellig-ability. It may have been, rather, that the representations did not *actually* render the target happening intelligible to anyone. On this view, it is still an open question as to what sorts of beings a representation must actually make sense of things.

Within the high-science group, there were not significant differences between ratings of (PA) and any of the Set 2 vignettes (though, as mentioned, the comparison to (HD) ( $p = .004$ ) is suggestive). Thus, it appears that high-science individuals may be quite flexible about to whom a representation must be intellig-able in order for it to count as an explanation. They do not expect that representations must be intellig-able to present-day humans, but it is unclear whether there is some upper bound on their expectations. They may think a representation that is intellig-able to any kind of sentient being at all counts as an explanation, or they might require that it be intellig-able to beings relevantly similar to present-day humans. In the latter case, high-science individuals may have such a liberal understanding about what counts as relevantly similar that the beings described in the Set 2 vignettes still meet this condition.

Clearly, further studies are also warranted here in order to determine for whom, precisely, a representation must be capable of making sense of things in order to be considered an explanation and whether or not low-science individuals require that this capacity be exercised.

3. Does accuracy matter? The accuracy hypothesis suggests that high-science participants should be more likely to judge something an explanation when it is accurate than when its accuracy is in question or when it is false. This pattern of results did not obtain. There was no difference in judgments about descriptions depicted as accurate (A) and

those depicted as possibly accurate (PA). As mentioned above, the possibly accurate but false (PAF) vignettes mirror many historical examples in which a representation of a way that things could have occurred is eventually discredited. We hypothesized that there would be no difference between judgments regarding the (A) and (PAF) vignettes. Results for high-science participants were consistent with this prediction in that there was no significant difference in their judgments about the two types of vignettes. However, we did find that low-science participants were significantly less likely to judge that the descriptions in the (PAF) vignettes are explanations compared with the descriptions in the (A) vignettes. The fact that the (A) versus (PAF) manipulation produced an effect in the low-science group alleviates some of our concerns about the relevance of the null result found in the high-science group. We thus once again take our findings to supply tentative evidence that the accuracy hypothesis does not reflect the views of practicing scientists, though we acknowledge that additional research on the matter is desirable.

### Anti-Psychologism Revisited

We were motivated to undertake this project in large part because we reject the anti-psychologistic stance regarding the study of explanation that still pervades much of the philosophy of science. Philosophers often take this stance to be justified by appeal to cases which they regard as explanations but which seem not to elicit any relevant psychological events (e.g., familiarity, insight, intelligibility, etc.). Our data suggest, however, that scientists and laypeople have different intuitions from philosophers regarding some of these cases. Insofar as there is something about a representation, whether sheer complexity or exotic constructs, that positively precludes it from rendering a certain happening intelligible, high-science and low-science participants are significantly less inclined to regard it as an explanation. Thus, the anti-psychologistic position appears to rest upon intuitions that are, without discernible justification, idiosyncratic; psychological investigations are the proper methodology for determining how scientists conceive of explanations. In addition, given that the capacity to produce a certain kind of mental state (i.e., finding intelligible) seems to be the crucial factor, empirical investigation into the actual exercise of this capacity will surely be a part of any complete portrait of why certain representations are regarded as explanations and others are not. Indeed, insofar as scientists and laypeople have, with their concepts of explanation, correctly demarcated the boundaries of a real natural or sociocultural kind, then psychological research into what it means to find a happening intelligible will probably contribute much to our understanding of what explanations are, in and of themselves. Thus, for philosophers of science wishing to know what explanations are and what role they play in our lay and scholarly lives, it would seem inadvisable to turn their backs on empirical psychology and retreat to

evaluating theories based upon how well they track their own classifications of supposedly clear-cut cases of explanation and non-explanation.

### References

- Craver, C. F. (2007). *Explaining the brain*. New York: Oxford University Press.
- Hempel, C. G. (1942). The function of general laws in history. *The Journal of Philosophy*, 39(2), 35-48.
- Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: The Free Press.
- Hospers, J. (1946). On explanation. *The Journal of Philosophy*, 43(13), 337-356.
- Humphreys, P. (1989). Scientific explanation: The causes, some of the causes, and nothing but the causes'. *Minnesota studies in the philosophy of science*, 13, 283-306.
- Lombrozo, T., & Carey, S. 2004. Functional explanation and the function of explanation. *Cognition*, 99, 167-204.
- Machamer, P., & Woody, A. (1994). A model of intelligibility in science: Using Galileo's balance as a model for understanding the motion of bodies. *Science & Education*, 3(3), 215-244.
- Mishra, P., & Brewer, W. F. (2003). Theories as a form of mental representation and their role in the recall of text information. *Contemporary Educational Psychology*, 28, 277-303.
- Miller, D. L. (1947). Explanation versus description. *The Philosophical Review*, 56(3), 306-312.
- Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.
- Salmon, W. (1998). *Causality and explanation*. New York: Oxford University Press.
- Trout, J. (2007). The psychology of scientific explanation. *Philosophy Compass*, 2(3), 564-591.

# Real-time expectations based on context speech rate can cause words to appear or disappear

**Meredith Brown (mbrown@bcs.rochester.edu)**

Department of Brain & Cognitive Sciences, University of Rochester  
Meliora Hall, Box 270268, Rochester, NY 14627-0268

**Laura C. Dilley (ldilley@msu.edu)**

Department of Communicative Sciences & Disorders, Michigan State University  
116 Oyer, East Lansing, MI 48824

**Michael K. Tanenhaus (mtan@bcs.rochester.edu)**

Department of Brain & Cognitive Sciences, University of Rochester  
Meliora Hall, Box 270268, Rochester, NY 14627-0268

## Abstract

To test predictions of a forward modeling framework for spoken language processing, we characterized effects of context speech rate on the real-time interpretation of indefinite noun phrases using the visual world paradigm. The speech rate of sentence material distal to the onset of the noun phrase was manipulated such that the segments surrounding the determiner *a* in singular noun phrases had a faster speech rate than the surrounding context and the segments surrounding the onset of plural noun phrases had a relatively slow rate. These manipulations caused listeners to fail to perceive acoustically present determiners and to falsely perceive determiners not present in the signal. Crucially, fixations to singular and plural target pictures revealed effects of distal speech rate during the real-time processing of target expressions, strongly suggesting a locus in perceptual expectations. These results set the stage for quantitative tests of forward models of spoken language processing.

**Keywords:** Speech rate; prosody; expectations; speech processing; visual world paradigm

## Introduction

Expectation-based approaches in which perceptual input is evaluated with respect to internally generated forward models provide compelling and increasingly influential explanations of phenomena in the perception and motor control literatures (e.g. Jordan & Rumelhart, 1992; Kawato, 1999). For example, DIVA, an influential model of speech production, incorporates a forward model component that predicts the auditory signal likely to result from a particular configuration of articulators within the vocal tract (Guenther & Micci Barreca, 1997). The forward model component accounts for the speed and efficiency with which the system can control speech movements, given the relatively slow mechanisms by which acoustic feedback influences speech production.

Similar forward modeling approaches may also provide a promising explanatory framework for spoken language comprehension. As in the domain of motor control, forward modeling of the perceptual attributes of upcoming speech provides a compelling explanation for the remarkable speed and efficiency of speech perception and spoken word recognition in the face of considerable variability. This is a particularly attractive feature of expectation-based approaches to higher-level language comprehension as well, which propose a central role for expectations in processes such as syntactic com-

prehension and lexical processing (e.g. Levy, 2008; Altmann & Kamide, 1999). We propose that comprehension also involves developing expectations about the acoustic realization of upcoming speech, conditioned on various contextual factors such as prosodic phrasing, speech rate, discourse history, and speaker-specific characteristics. Previous work suggests that these expectations are best characterized as probability distributions, with listeners representing not only the expected form of a spoken word given the set of contextual conditioning factors, but also a measure of the variance or uncertainty of their estimate (Clayards et al., 2008; Levy et al., 2009). The degree of congruence between these perceptual expectations and the incoming acoustic signal then contributes to the differential activation of competing lexical alternatives. Finally, when perceptual expectations are incongruent with the actual realization of a word, listeners should update their beliefs about the cues that condition their perceptual expectations, resulting in adaptation that more closely aligns listeners' expectations with the characteristics of the signal in context.

Speech prosody generates acoustic regularities that are likely to foster expectations about the acoustic realization of upcoming speech sounds, including pitch and temporal characteristics that listeners perceive as patterning. This perceived patterning has been shown to influence real-time spoken language processing. For example, manipulations of pitch and duration early in an utterance influence the interpretation of cues to prosodic structure several syllables downstream (Dilley & McAuley, 2008; Dilley et al., 2010; Brown et al., 2011). The distal locus of these effects suggests that they are rooted in listeners' expectations about the acoustic-phonetic realization of upcoming segments.

Speech rate is particularly likely to systematically influence listeners' expectations about upcoming material. Speech sounds are interpreted relative to the global speech rate of an utterance, affecting perceived phoneme distinctions such as voicing contrasts (e.g. Miller, 1987; Reinisch et al., 2011). Therefore, listeners must evaluate the incoming signal with respect to a speaker's estimated rate. Dilley and Pitt (2010) demonstrated that effects of context speech rate scale up to the perceived rate of articulation of larger constituents, in-

cluding syllables and words. They found that when the speech rate of a function word (e.g. *or* in the phrase *leisure or time*) is increased relative to the rate of the surrounding context by either speeding up the function word or slowing down the surrounding context, participants are less likely to report hearing a function word in a sentence transcription task. Conversely, when the speech rate of segments surrounding a location in which a function word would be licensed grammatically (e.g. *leisure time*) is slowed down relative to the surrounding context, participants are more likely to report hearing a function word within the relatively slow portion of speech, effectively hallucinating having heard this item.

The findings of Dilley and Pitt (2010) suggest that listeners rapidly entrain to the rate of an utterance and develop speech rate expectations that influence the perceived number of morphophonological constituents within a spectrally ambiguous stretch of speech of a certain duration. However, the task used in this study does not directly address the prediction of the forward modeling account that the observed speech rate effects have an expectation-based locus. Indeed, it is possible that the sentence elicitation task itself contributed to the observed effects by engaging the production system and inviting explicit comparison of the perceived utterance with different alternative parses. Without time course information, it is unclear whether the observed effects of speech rate on the appearance and disappearance of words are based on perceptual expectations or post-perceptual reinterpretation of the input.

## Experiment overview

As a first step toward testing the predictions of the forward model, we used the visual world paradigm (Tanenhaus et al., 1995) to assess the effect of context speech rate on the interpretation of indefinite singular and plural noun phrases. In these noun phrases, the presence or absence of the plural morpheme *-s* was phonemically ambiguous, due to the presence of a sibilant-initial word following the target expression. In addition, the speech rate of sentence material distal (i.e. non-local) to the onset of the target expression was manipulated such that the segments surrounding the determiner in singular noun phrases had a faster speech rate than the distal context and the segments surrounding the onset of plural noun phrases had a slower speech rate than the distal context. These distal context manipulations were predicted to bias listeners to fail to perceive acoustically present determiners within singular noun phrases and to falsely perceive determiners prior to plural noun phrases, thereby shifting whether the noun phrases were judged to be singular or plural. We hypothesized that listeners' expectations about the acoustic-phonetic realization of upcoming segments within an utterance would be the source of these perceptual effects, and that manipulating the speech rate of material distal to the onset of the target expression would therefore influence fixations to pictures of singular vs. plural referents during the real-time processing of the noun phrase.

## Methods

### Participants

Thirty-two native English speakers from the University of Rochester participated in the visual world experiment. All participants had normal hearing and normal or corrected-to-normal visual acuity.

### Materials

The speech stimuli used in the experiment were 24 grammatical declarative sentences containing a target noun phrase consisting of an adjective and plural noun (e.g. *(a) brown hen(s)*). Each target expression was preceded by at least six syllables of utterance context ending in a vowel or rhotic consonant (e.g. *The Petersons are looking to buy*), a phonetic context in which a high degree of coarticulation with the determiner *a* would be expected. The target expression was followed by at least two additional syllables, beginning with a sibilant-initial word (e.g. *soon*) to increase participants' reliance on the determiner, rather than the presence or absence of plural *-s*, as a cue to number.

	preceding context	determiner region	target expression
singular			
no manipulation	1480	381	598
distal manipulation	2812	381	1137
plural			
no manipulation	1480	313	598
distal manipulation	888	313	359

Table 1: Mean durations of each sentence region by target expression type and condition. For each recording, duration of the determiner region was identical between distal- and no-manipulation versions of each recording.

Spoken sentences containing singular and plural versions of the target expression were elicited from 12 speakers and recorded using a Marantz PMD 660 digital recorder sampling at 44.1 kHz in a sound-attenuated booth. A singular and plural version of two critical items were selected from each speaker's recordings. The pitch synchronous overlap-and-add algorithm was then used to create two resynthesized versions of each recording (Moulines & Charpentier, 1990), in addition to two versions not discussed here. In the *distal-manipulation condition*, the speech rate of sentence material distal to the potential location of a determiner was altered; in the *no-manipulation condition*, the speech rate of the utterance was unaltered. For singular expressions, the distal speech rate manipulation involved temporally expanding the utterance context preceding and following the determiner region (the region beginning with the word preceding the determiner and ending with the following phoneme, e.g. *buy a b-*) by a factor of 1.9. This manipulation resulted in the



determiner region having a faster speech rate than the surrounding context. Likewise, the distal speech rate manipulation for plural expressions involved temporally compressing the determiner region by a factor of .6, to slow the rate of the determiner region relative to the surrounding context and thereby encourage the perception of an acoustically absent determiner. The mean duration across each region of the stimuli in each condition is provided in Table 1.

To reduce the salience of the singular-plural ambiguity in the critical items, 18 singular and 18 plural filler items were included in which the number of the target expression was more clearly signaled (e.g. *that rusty knife*). Of these filler items, one third were temporally compressed by a factor of .6 and one third were expanded by a factor of 1.9.

## Procedure

Each trial began with the presentation of a visual display containing four clip-art pictures. Two pictures depicted singular and plural versions of the target expression. The other two were singular and plural versions of a different object whose labels were phonologically unrelated to the target word. To ensure visual similarity between singular and plural pictures, the two versions of each picture were created by manipulating a single picture, either by duplicating and superimposing copies of a single target object or by isolating a single object within a picture of multiple target objects.

After 500 ms of display preview, participants heard a spoken sentence over Sennheiser HD 570 headphones. Their task was to click on the picture referred to in the sentence. Participants were not given feedback on their performance, and incorrectly selected a distractor picture on fewer than .5% of all critical trials. Throughout the study, eye movements were tracked and recorded using a head-mounted SR Research EyeLink II system sampling at 250 Hz, with drift correction procedures performed after every fifth trial.

Two sets of three lists were constructed by randomizing picture positions and trial order, dividing the experiment into three blocks, and rotating through permutations of the blocks. Within each list, an equal number of items were assigned to each of four singular and four plural conditions. The pairing of items with conditions was counterbalanced across participants, and each participant encountered each sentence only once. All lists started with four filler items to familiarize participants with the referent identification task.

## Analyses

Response choices were analyzed with separate multilevel logistic regression models for singular and plural items. Condition, the duration of the sibilant -s in each recording, and their interactions were included as fixed effects, and random intercepts and slopes were included for participants and items. For the measure of sibilant duration, we used the duration of the sibilant in each recording prior to manipulation, to minimize collinearity between sibilant duration and condition and to account for known effects of rate normalization on the relation between phoneme duration and segmentation (e.g. Miller,

1987). The duration of the sibilant was standardized by subtracting the mean value and dividing by the standard deviation. The final model was selected by removing fixed and random effects stepwise and comparing each smaller model to the more complex model using the likelihood ratio test (Baayen, Davidson, & Bates, 2008).

Growth curve analysis was used to evaluate the proportions of fixations to singular and plural target pictures over time by condition (Mirman et al., 2008). This analysis method is a variant of multi-level regression modeling that has emerged as an alternative to other statistical methods used to evaluate effects of experimental manipulations in the visual world paradigm. In growth curve analysis, the proportions of fixations to a picture are first aggregated by participants or by items for each condition at each time point sampled within the region of interest. Orthogonal power polynomial terms are then fit to the resulting fixation proportion curves to model variations in curve parameters (e.g. the intercept and slope) that can be attributed to the independent variable(s) and to participant-wise or item-wise variation. Because it explicitly models the trajectory of change in proportions of fixations over time, growth curve analysis is a more appropriate and temporally sensitive analysis technique than traditional analysis of variance approaches, which frequently entail the violation of assumptions about the independence and distribution of data points and the loss of fine-grained temporal detail.

Separate growth curve analyses were conducted for fixations to singular and plural target pictures in response to singular and plural tokens, with the no-manipulation condition used as the reference condition in all analyses. For analysis of fixations during the processing of the target noun phrase, data were aggregated and analyzed by participants and by items (indicated throughout as  $B_1$  and  $B_2$ , respectively). Each model used a third order polynomial to capture the generally sigmoidal shape of the curves.

For analysis of fixations during the processing of the target noun phrase, we characterized the effects of condition on the shape of the fixation curve by adding to the base model the effects of condition on the intercept and linear term, which represent the overall mean curve height and the slope of the curve. The onset of the adjective was used as the reference point for consistency across analyses of singular and plural conditions. Because items varied in the extent to which the adjective in the target expression biased listeners toward the target picture relative to the distractor picture, the onset of signal-driven fixations relative to the start of the adjective was determined by visually evaluating the point of divergence between the mean proportion of fixations to either version of the target picture and the mean proportion of fixations to either version of the distractor picture, averaged across conditions. Averaging across singular and plural target pictures and across conditions permitted unbiased evaluation of the mean point of divergence (Barr, 2008), estimated to be 300 ms after the onset of the adjective. Fixation curves were therefore analyzed between 300 and 1500 ms after adjective onset.

## Results

### Responses in the picture selection task

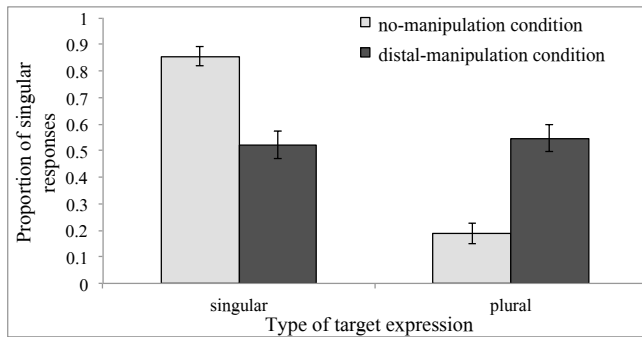


Figure 1: Proportions of correct picture selections.

Results from the picture selection task are shown in Figure 1. The regression model of responses for singular items revealed significant effects of condition. Fewer singular target pictures were selected in the distal-manipulation condition than in the no-manipulation condition ( $B=2.12$ ,  $z=5.24$ ,  $p<.0001$ ). Analysis of responses to plural items showed the opposite effect, with more singular target pictures selected in the distal-manipulation condition than in the no-manipulation condition ( $B=-2.09$ ,  $z=-5.45$ ,  $p<.0001$ ). For both models, the duration of the sibilant and by-participants and by-items random slopes did not contribute significantly to variance in response data and were not included in the final model.

These data show a pattern similar to those in off-line response choice data obtained by Dilley and Pitt (2010) in a sentence transcription task using a wider variety of function words. Increasing the relative speech rate of the determiner by slowing the rate of the distal context decreased the likelihood of listeners perceiving the determiner, whereas speeding up the distal context biased listeners to perceive a determiner that was not present in the signal. In addition, these results confirmed that the sibilant following the target noun was acoustically ambiguous, causing listeners to base their judgments primarily on the perception of the determiner.

### Fixations during processing of target expression

Figure 2 shows the ratio of mean fixation proportions for singular target pictures to the sum of mean fixation proportions for both singular and plural target pictures, averaged across the window starting at 300 ms after the onset of the adjective and ending 300 ms after the offset of the target expression. The pattern of results was similar to the effects observed in responses in the picture selection task. During the processing of singular expressions, the proportion of fixations to the singular picture was higher overall in the no-manipulation condition than in the distal-manipulation condition. Fixations during the processing of plural expressions showed the opposite effect. The proportion of fixations to the singular picture was higher in the distal-manipulation condition than in the no-manipulation condition. This pattern of results supported

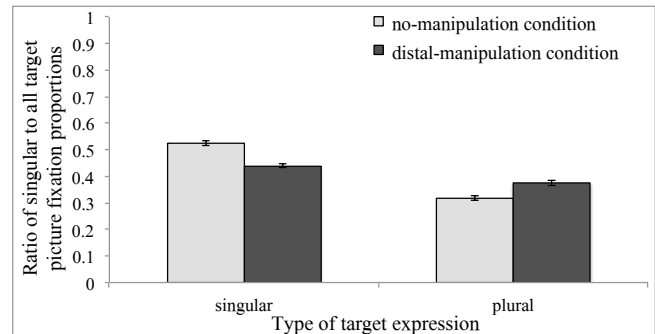


Figure 2: Ratio of mean proportions to singular target pictures to the sum of mean proportions to singular and plural target pictures between 300 ms following adjective onset and 300 ms following the offset of the target expression.

the prediction that effects of distal speech rate manipulation on listeners' judgments would manifest during real-time comprehension.

Figure 3 (left) provides more detailed information about the time course of fixations to singular and plural pictures during the processing of singular items. Growth curve analysis revealed that distal manipulation had a significant negative effect on the intercept term for fixations to the singular picture ( $B_1=-.16$ ,  $t_1=-5.33$ ,  $p<.0001$ ;  $B_2=-.15$ ,  $t_2=-4.36$ ,  $p<.0005$ ) and a significant positive effect on the intercept term for fixations to the plural picture ( $B_1=.07$ ,  $t_1=2.19$ ,  $p<.05$ ;  $B_2=.07$ ,  $t_2=1.79$ ,  $p<.1$ ), although this effect was marginal in the items analysis. Relative to the no-manipulation condition, the distal-manipulation condition elicited a lower overall proportion of fixations to the singular picture and a higher overall proportion of fixations to the plural picture. Further, the linear term showed significant effects of distal manipulation on fixations to both the singular picture ( $B_1=-1.23$ ,  $t_1=-4.07$ ,  $p<.0001$ ;  $B_2=-1.18$ ,  $t_2=-4.28$ ,  $p<.0001$ ) and the plural picture ( $B_1=.87$ ,  $t_1=2.84$ ,  $p<.0001$ ;  $B_2=.79$ ,  $t_2=19.83$ ,  $p<.0001$ ), suggesting that the rate of change was less steep overall for fixations to singular pictures and steeper for fixations to plural pictures in the distal-manipulation condition.

Figure 3 (right) shows the time course of fixations to singular and plural target pictures during the processing of plural items. For the most part, analyses of fixations during the processing of plural items yielded a pattern of results similar to that obtained for singular items. Relative to the no-manipulation condition, the intercept term was significantly higher in the distal-manipulation condition than in the no-manipulation condition for fixations to singular pictures ( $B_1=.08$ ,  $t_1=3.22$ ,  $p<.005$ ;  $B_2=.08$ ,  $t_2=3.21$ ,  $p<.005$ ), and lower for fixations to plural pictures ( $B_1=-.14$ ,  $t_1=-4.43$ ,  $p<.0001$ ;  $B_2=-.14$ ,  $t_2=-4.26$ ,  $p<.0005$ ). Further, the linear term for fixations to the singular picture differed significantly across conditions ( $B_1=.81$ ,  $t_1=3.65$ ,  $p<.0005$ ;  $B_2=-.79$ ,  $t_2=-19.83$ ,  $p<.0001$ ), but the linear term for fixations to the plural picture did not ( $B_1=.15$ ,  $t_1=.45$ ,  $p>.1$ ;  $B_2=.23$ ,  $t_2=.70$ ,  $p>.1$ ).

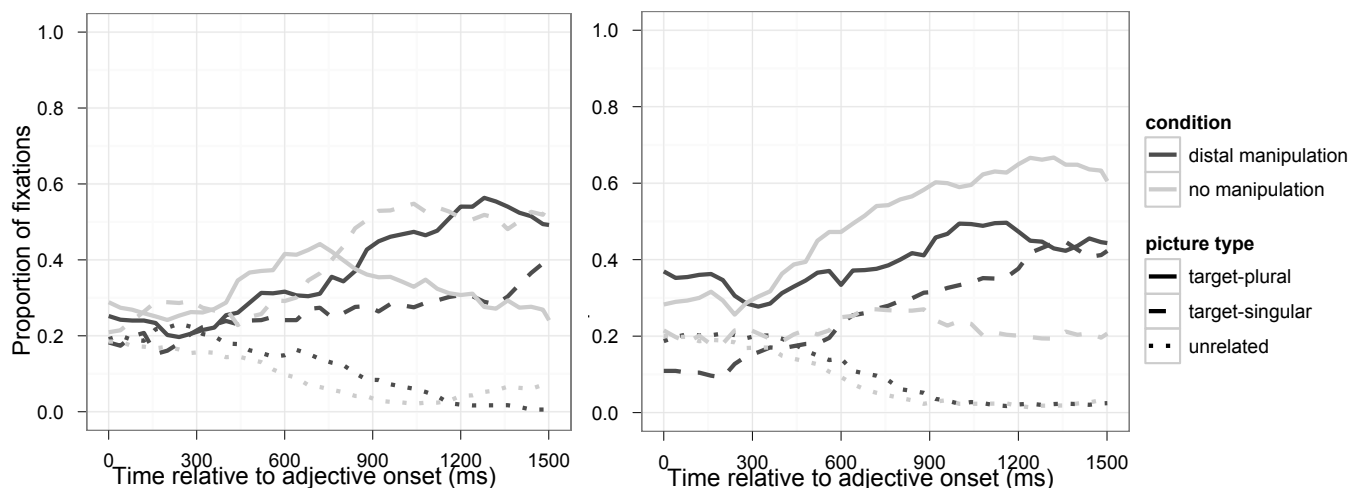


Figure 3: Proportions of fixations to pictures in response to singular (left) and plural (right) target noun phrases. Black and grey lines represent distal- and no-manipulation conditions; solid and dashed lines represent fixations to plural and singular pictures.

In summary, participants were less likely overall to fixate plural pictures and more likely to fixate singular pictures in the distal-manipulation condition than in the no-manipulation condition. This result demonstrates that manipulating the speech rate of an utterance distal to the potential location of a determiner influences whether a determiner is perceived during the real-time processing of indefinite noun phrases, consistent with the predictions of the forward modeling account.

## Discussion

When speech distal to the onset of an indefinite determiner at the onset of a singular noun phrase was temporally expanded, listeners were less likely to perceive the determiner and were more likely to select a plural picture as the referent of the noun phrase. Conversely, when speech distal to the onset of a plural noun phrase was temporally compressed, listeners were more likely to select a singular picture, suggesting that the distal context manipulation induced the perception of a determiner that was not acoustically present in the signal. The absolute speech rate of the determiner region of each item was identical across conditions, demonstrating that it was the speech rate of the determiner relative to its surrounding context, rather than its absolute rate, that drove these effects.

Crucially, fixations to singular and plural pictures revealed that effects of distal speech rate occurred during the real-time processing of the target expression, strongly suggesting a locus of the effect in perceptual expectations. The observed time course of speech rate effects demonstrates that listeners entrain to the overall rate at which speech sounds are articulated within an utterance and expect this speech rate to persist in upcoming material, biasing them to expect relatively long stretches of speech to contain more morphophonological constituents and relatively short stretches of speech to contain fewer. Indeed, expectations based on context speech rate may be a powerful cue in the online interpretation of natural speech, which is generally readily interpretable despite

frequently degraded or ambiguous spectral cues to segmental content.

According to a forward modeling account of spoken language processing, the language processing system includes a component that integrates relevant contextual information from a variety of sources, such as speech rate and speaker identity, to predict the acoustic-phonetic attributes of different lexical alternatives. These expectations crucially influence listeners' perception of the incoming acoustic signal and the relative activation of competing lexical alternatives. Further, mismatch between expectations and the signal is predicted to result in feedback that continuously updates the probabilistic links between contextual factors and expected outcomes.

Although most of the growing body of work demonstrating fine-grained sensitivity to acoustic detail has not been framed in terms of perceptual expectations (e.g. Gow, 2001; Hawkins, 2003; Salverda et al., 2003; though see McMurray et al., 2011), such findings are congruent with this framework. Particularly suggestive are effects of acoustic detail being conditioned by context manipulations (e.g. Salverda et al., 2007; Sumner & Gaftner, 2011) and by speaker-specific characteristics (e.g. Kraljic et al., 2008; Creel et al., 2008).

Our findings set the stage for further tests of predictions of the forward modeling account. For example, if perceptual expectations about upcoming speech are represented probabilistically, the variance of the distribution of expected acoustic forms should influence the magnitude of expectation-based effects on perception. Moreover, incongruence between expectations and the actual realization of a word in context should result in perceptual adaptation bringing expectations more in line with relevant characteristics of the signal (e.g. speaker-specific accent or idiosyncratic prosody). To test these predictions, we are manipulating the relative speech rate not only of the determiner region, but also of segments surrounding the sibilant following the target expression.

## Conclusions

This study demonstrates that perceptual expectations have sufficiently strong effects on perception to make words appear within and disappear from the signal during real-time language processing. Forward modeling of the acoustic realization of upcoming speech may be a crucial mechanism enabling listeners to cope with and exploit variability in the input, particularly when spectral cues are degraded or ambiguous. These findings establish distal speech rate manipulation as a suitable experimental paradigm for testing further predictions of the forward modeling account.

## Acknowledgments

This research was supported by an NSF predoctoral fellowship (MB), NSF grant BCS-0847653 (LCD), and NIH grants HD27206 and DC0005071 (MKT). We gratefully acknowledge Dana Subik for participant recruitment and testing, Anne Pier Salverda for valuable discussions, and Chelsea Marsh for assistance with stimulus creation.

## References

- Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73, 247–264.
- Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.
- Barr, D. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, 59, 457–474.
- Brown, M., Salverda, A., Dilley, L., & Tanenhaus, M. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin and Review*, 18, 1189–1196.
- Clayards, M., Tanenhaus, M., Aslin, R., & Jacobs, R. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108, 804–809.
- Creel, S., Aslin, R., & Tanenhaus, M. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, 106, 633–664.
- Dilley, L., Mattys, S., & Vinke, L. (2010). Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. *Journal of Memory and Language*, 63, 274–294.
- Dilley, L., & McAuley, J. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, 59, 291–311.
- Dilley, L., & Pitt, M. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, 21, 1664–167.
- Gow, D. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language*, 45, 133–159.
- Guenther, F., & Micci Barreca, D. (1997). Neural models for flexible control of redundant systems. In P. Morasso and V. Sanguineti (eds.), *Self-organization, Computational Maps and Motor Control* (pp. 383–421). Amsterdam: Elsevier-North Holland.
- Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31, 373–405.
- Jordan, M., & Rumelhart, D. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16, 307–354.
- Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9, 718–727.
- Kraljic, T., Samuel, A., & Brennan, S. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science*, 19, 332–338.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106, 1126–1177.
- Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106, 21086–2109.
- McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118, 219–246.
- Miller, J. (1987). Rate-dependent processing in speech perception. In A. Ellis (ed.), *Progress in the psychology of language* (pp. 119–157). London: Erlbaum Associates.
- Mirman, D., Dixon, J., & Magnuson, J. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, 59, 475–494.
- Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9, 666–688.
- Reinisch, E., Jesse, A., & McQueen, J. (2011). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, 37, 978–996.
- Salverda, A., Dahan, D., & McQueen, J. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, 90, 51–89.
- Salverda, A., Dahan, D., Tanenhaus, M., Crosswhite, K., Masharov, M., & McDonough, J. (2007). Effects of prosodically modulated sub-phonetic variation on lexical competition. *Cognition*, 105, 466–476.
- Sumner, M., & Gafter, R. (2011). Integrating frequency, formality and phonology in the perception of spoken words. Talk presented at the 85th annual meeting of the Linguistic Society of America, Pittsburgh, PA.
- Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632–1634.

# Metrical expectations from preceding prosody influence spoken word recognition

**Meredith Brown (mbrown@bcs.rochester.edu)**

Department of Brain & Cognitive Sciences, University of Rochester  
Meliora Hall, Box 270268, Rochester, NY 14627-0268

**Anne Pier Salverda (asalverda@bcs.rochester.edu)**

Department of Brain & Cognitive Sciences, University of Rochester  
Meliora Hall, Box 270268, Rochester, NY 14627-0268

**Laura C. Dilley (ldilley@msu.edu)**

Department of Communicative Sciences & Disorders, Michigan State University  
116 Oyer, East Lansing, MI 48824

**Michael K. Tanenhaus (mtan@bcs.rochester.edu)**

Department of Brain & Cognitive Sciences, University of Rochester  
Meliora Hall, Box 270268, Rochester, NY 14627-0268

## Abstract

Two visual world experiments tested the hypothesis that expectations based on preceding prosody influence the perception of suprasegmental cues to lexical stress. Experiment 1 showed that phonemically overlapping words with different initial stress patterns compete for recognition. Experiment 2 further demonstrated that fundamental frequency and syllable timing patterns across material preceding the target word can influence the relative activation of competing alternatives with different initial stress patterns. The activation of alternatives with initial stress was higher when preceding stressed syllables had suprasegmental acoustic characteristics similar to the initial syllable of the target word. These findings suggest that expectations about the acoustic realization of an utterance include information about metrical organization and lexical stress, and that these expectations constrain the initial interpretation of suprasegmental stress cues. These results are interpreted as support for expectation-based forward models in which acoustic information in the speech stream is interpreted based on expectations created by prosody.

**Keywords:** Prosody; spoken word recognition; lexical stress; visual world paradigm; expectations; lexical competition

## Introduction

A growing body of work indicates that expectations about the acoustic realization of the phonemes and prosody of a spoken sentence influence how listeners initially interpret incoming acoustic-phonetic cues during spoken language processing. For example, manipulations of pitch and duration early in an utterance influence the interpretation of cues to prosodic and morphophonemic constituency several syllables downstream (Dilley & McAuley, 2008; Dilley et al., 2010; Brown et al., 2011; Dilley & Pitt, 2010). However, little is known about the types of representations that contribute to these perceptual expectations. The present study investigates whether perceived prosodic and metrical patterning across preceding portions of an utterance can influence listeners' expectations about the metrical organization of upcoming material, modulating their interpretation of proximal cues to lexical stress and therefore influencing the activation of potential lexical candidates.

Lexical stress is a key contributor to sentence-level prominence patterns and rhythmicity. Listeners are sensitive to

segmental and suprasegmental cues to stress during spoken word recognition (Cutler, Dahan & van Donselaar, 1997). Although vowel quality is the most potent stress cue to influence lexical processing in English, other suprasegmental cues such as duration also distinguish stressed from unstressed syllables, and judgments about these suprasegmental stress cues are modulated by surrounding prosody in off-line tasks (Niebuhr, 2009). Cues to stress may influence not only the recognition of particular words, but also the process of segmenting the speech stream more generally. For example, listeners are more likely to misperceive phrases like "she's a must to avoid" as "she's a muscular boy" than they are to mishear "in closing" as "enclosing", suggesting that listeners preferentially posit word boundaries prior to prominent syllables (Cutler & Butterfield, 1992). This metrical segmentation strategy is substantiated by the distribution of stressed syllables within the English lexicon: approximately 90% of content words in conversational English have initial stress (Cutler & Carter, 1987).

Perceived metrical patterning is a potentially powerful source of expectations in speech perception. Speech prosody often exhibits characteristics that listeners perceive as patterning (Couper-Kuhlen, 1993; Pierrehumbert, 2000). For example, listeners tend to hear stressed syllables in English as perceptually isochronous, i.e., as occurring at regular intervals (e.g., Lehiste, 1977). In addition, previous work using non-linguistic auditory stimuli (e.g. sequences of alternating tones) has demonstrated that pitch, temporal, and/or amplitude patterning in distal (i.e. non-local) auditory context can influence the processing of proximal material (e.g. the perceived relative prominence of high vs. low tones; Woodrow, 1911; Thomassen, 1982). The tendency for speakers to use recurring sequences of pitch accents within an intonational phrase (Couper-Kuhlen, 1993; Pierrehumbert, 2000) may likewise contribute to the perceived metrical structure across syllables in an utterance.

We conducted two visual world experiments to test the hy-

pothesis that expectations based on preceding prosody influence the perception of suprasegmental cues to lexical stress. Experiment 1 verified that phonemically overlapping words with different initial stress patterns compete for recognition. Experiment 2 further demonstrated that fundamental frequency and syllable timing patterns across material preceding the target word can influence the relative activation of competing alternatives with different initial stress patterns. The activation of alternatives with initial stress was higher when preceding stressed syllables had suprasegmental acoustic characteristics similar to the initial syllable of the target word, and vice versa. These findings suggest that expectations about the acoustic realization of upcoming speech include information about metrical organization and lexical stress, and that these expectations constrain the initial interpretation of suprasegmental stress cues during spoken word recognition.

## Experiment 1

The main goal of Experiment 1 was to establish that phonemically overlapping words with different initial stress patterns compete for recognition. We demonstrate that materials containing a target word with relatively neutral segmental and suprasegmental stress cues on the initial syllable elicit initial activation of both initially-stressed and initially-unstressed lexical alternatives.

### Methods

**Participants** The participants were 16 students from the University of Rochester. All participants were native English speakers with normal hearing and corrected-to-normal visual acuity, and received \$7.50 for their participation in the study.

**Materials** The 24 speech stimuli used in the experiment were grammatical declarative sentences containing either a word with initial strong-weak stress (e.g. *jury*) or a phonemically overlapping word with initial weak-strong stress (e.g. *giraffe*). The initial syllables for each related pair of words were produced with as close to the same vowel quality and pronunciation as possible. Each stimulus began with at least two disyllabic words with initial primary stress, followed by one or two monosyllabic words (e.g. *Heidi sometimes saw*). This distal context material was followed by another monosyllabic word (e.g., *that*) followed by the target word. Whereas the preceding context for each item was the same for both SW and WS target words, the sentence material following the target word differed to maximize semantic coherence (e.g. *the jury leaving the courthouse* vs. *the giraffe in the city zoo*).

To discourage participants from noticing the stress pattern manipulation, we included 48 filler items for which the visual display contained two pictures whose labels had a different phonological relation (e.g. words with onset-embedded competitor words, like *antlers* and *ant*). For half of the filler items, one of the phonologically related words was mentioned in the utterance. In addition, an equal number of SW, WS, and

monosyllabic target words were used in the filler items. Eight filler items were identity-spliced between the first and second syllables of the target word, whereas the rest were spliced at some point prior to the target word.

The first author recorded multiple tokens of each sentence as WAV files at 44.1 kHz, producing each sentence with minimal F0 excursions and slight F0 declination. Each recording was split into two halves at the end of the first syllable of the target word, at a point in the waveform with an amplitude of zero. Identity- and cross-spliced versions of the item containing the SW target word were created by splicing together the last half of a SW recording either with the first half of another SW recording or with the first half of a WS recording. Likewise, identity- and cross-spliced versions of the item containing the WS target word were created by splicing the last half of a WS recording together with the initial SW and WS sentence fragments used to create the SW items.

**Procedure** The study was divided into three phases. In the first phase, each of the 304 clip-art pictures used in the visual world experiment was presented to the participant in the center of the display, with its label printed underneath. Each picture-label pair was presented for a minimum of 3 seconds. Participants proceeded at their own pace through the set of picture-label pairs by pressing the space bar.

The visual world experiment began immediately following the picture-label exposure phase. Each trial started with the presentation of a visual display containing four pictures, two of which corresponded to the phonologically related SW and WS words on critical trials. The remaining two distractor images were selected such that they were distinct from the two potential target pictures with respect to visual and semantic properties and the phonological properties of their labels. After 500 ms of display preview, participants heard a spoken sentence, and their task was to click on the picture that was referred to in the sentence. They were not given feedback on their performance during the experiment. Throughout the study, eye movements were tracked and recorded using a head-mounted SR Research EyeLink II system sampling at 250 Hz, with drift correction procedures performed every five trials.

Immediately after the completion of the visual world experiment, we assessed participants' ability to generate the appropriate labels for both members of each stress-alternating word pair. Participants named each of the 48 associated pictures, presented in a randomized sequence, and their responses were recorded. Responses were scored as correct if they preserved the phonemes and stress pattern across the initial two syllables (e.g. *jury*, *jury box*, and *jury members* were all considered correct, but *jurors* was not).

For the visual world experiment, two lists were constructed by randomizing the positions of the images on the screen within each trial and pseudorandomizing the order of trials within the list. Within each list, half of the experimental trials had SW target words and half had WS target words, and of these, half were identity-spliced and half were cross-spliced.

The assignment of items to each of the four conditions was counterbalanced across participants, for a total of eight lists. Six practice trials were included at the start of the experiment to familiarize participants with the picture selection task.

**Analyses** For statistical analysis, proportions of fixations to the target, competitor, and distractor pictures on experimental trials were averaged across the window starting at 200 ms after the onset of the target word and ending 750 ms later, i.e., 200 ms after the mean offset of the target word. The mean proportions were then transformed using the empirical logit function (Cox, 1970). Effects of target word type and splicing condition on logit-transformed fixation proportions were analyzed in a multilevel linear regression model. Fixed effects included picture type, target word type, splicing condition, trial number, and interactions between these factors. Random effects included intercepts and slopes for participants and items. Trial number was standardized by subtracting the mean value and dividing by the standard deviation. To select the final model, effects were removed stepwise and each reduced model was compared to the more complex model using the likelihood ratio test (Baayen et al., 2008).

## Results and discussion

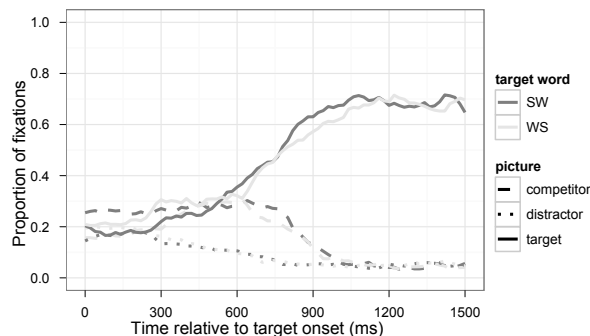


Figure 1: Proportions of fixations to target, competitor, and distractor pictures in Experiment 1. Line color denotes target word type (SW vs. WS); line texture denotes picture type.

Overall, participants found the referent identification task easy, and clicked on the incorrect picture on less than 3% of all experimental trials. These trials were excluded from further analysis. In addition, we excluded trials for which the target or competitor picture was not correctly named by the participant in the third phase of the experiment (18.9%), since their performance on this task revealed whether they associated the intended names with the pictures.

Figure 1 shows the proportion of fixations to the target, competitor, and distractor pictures as a function of condition starting at the onset of the target word. As expected, hearing the initial sounds of a WS word elicited transient competition from a phonemically overlapping SW competitor, starting at approximately 250 ms after the onset of the target word. The proportions of target and competitor fixations were roughly equivalent until approximately 600 ms af-

ter target word onset, when fixations began to converge on the target picture. Crucially, hearing the initial sounds of a SW word also elicited competition effects from WS competitors with approximately the same magnitude and time course.

Multilevel linear regression analyses confirmed the prediction that words with initial WS stress initially compete for recognition with phonemically overlapping words with initial SW stress. The logit-transformed proportion of fixations to the distractor pictures was significantly lower than the transformed proportions of fixations to the target ( $B=.59$ ,  $SE=.11$ ,  $t=5.16$ ,  $p<.0001$ ) and competitor ( $B=.29$ ,  $SE=.08$ ,  $t=3.70$ ,  $p<.0005$ ) pictures, after taking into account by-participant and by-item random intercepts, by-participant random slopes for picture type, and by-item random slopes for the interaction between picture type and splicing condition. Neither the target word stress pattern nor trial number contributed significantly to model fit, suggesting that competition effects were similar for SW and WS words and were stable across the experiment. Neither factor was included in the final model. Further, fixed effects of splicing condition did not contribute significantly to variance in fixation proportions. Taken together, these findings indicate that, for our materials, the competition between phonemically overlapping words with different initial stress patterns was similar for SW and WS target words.

These results verified that phonemically overlapping words with different initial stress patterns compete for recognition, and suggested that statistically-based heuristics or biases to interpret lexically stressed syllables as the onsets of content words do not dominate processing, at least when corresponding weak and strong syllables have similar vowel quality and segmental pronunciation. The overall similarity of lexical competition effects in identity- vs. cross-spliced items further suggested that we succeeded in creating items with relatively neutral segmental and suprasegmental cues to the lexical stress of the initial syllable.

## Experiment 2

The goal of Experiment 2 was to characterize the effects of preceding prosody on listeners' initial interpretation of different-stress cohort pairs. In this experiment, the acoustic characteristics of preceding portions of the utterance distal to the target word were manipulated, leaving the acoustic realization of the target word and its immediately surrounding context unchanged. Syllables in the distal context with lexical or sentence-level stress were manipulated to have the same relative F0 level (either low or high with respect to surrounding syllables) and to be roughly isochronous. We hypothesized that this prosodic manipulation would bias listeners to expect upcoming syllables with similar pitch and timing characteristics as preceding prominent syllables to be lexically stressed, and conversely to expect upcoming syllables with different pitch and timing characteristics to be unstressed.

## Methods

**Participants** The participants were 32 students from the University of Rochester who met the same inclusion criteria



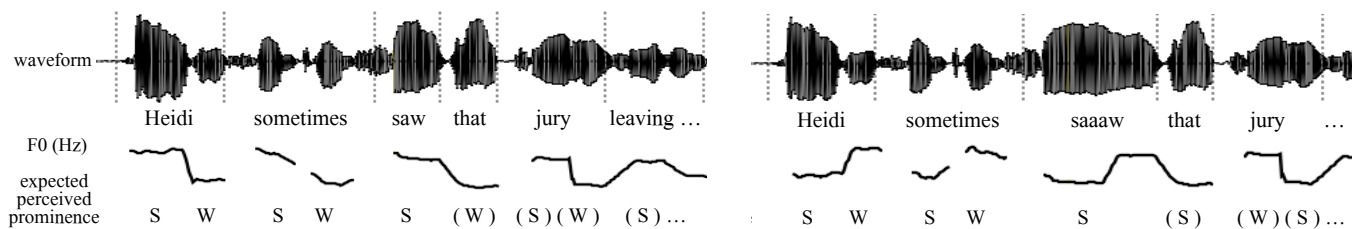


Figure 2: Example item in SW-biasing (left) and WS-biasing (right) conditions. The fundamental frequency (F0) and timing of stressed syllables within preceding distal context were manipulated to bias listeners to perceive the initial syllable of the target word as stressed or unstressed. Acoustic properties of the target word and its proximal context were the same across conditions.

as for Experiment 1.

**Materials** The stimuli used in Experiment 2 were created by manipulating the prosody of the sentence material preceding the target word in each of the identity- and cross-spliced items used in Experiment 1, using the pitch synchronous overlap-and-add algorithm (Moulines & Charpentier, 1990). All distal F0 manipulations involved removing non-vocalic pitch points and shifting the remaining pitch points within each context syllable up by 35 Hz (for high syllables) or down by 25 Hz (for low syllables). This manipulation preserved the natural microvariation and F0 declination of the original recording while imposing salient periodic alternations onto the F0 contour. F0 manipulations were performed for all syllables through at least the second syllable following the offset of the target word. Whether the first syllable of the target word had low or high F0 varied between items. Temporal manipulations involved compressing or expanding the rime of the first monosyllabic word within the preceding context such that the duration of the fifth intervocalic interval (i.e. the interval between vowel onsets in the fifth and sixth syllables) either matched the mean duration of the following two intervocalic intervals or was twice the mean duration of the preceding four intervocalic intervals, following Dilley and McAuley (2008). Similar prosodic manipulations were performed on filler items.

Two versions of each item were created with different acoustic characteristics across the distal context preceding the target word but the same acoustic characteristics across its proximal context (i.e. the preceding adjacent syllable) and all subsequent material (Figure 2). In the *SW-biasing condition*, syllables in the distal context with lexical and/or sentence-level stress (e.g. *Heidi sometimes saw...*) were manipulated to have the same relative F0 level as the initial syllable of the target word (e.g. high, cf. Figure 2), and the duration of the fifth syllable was manipulated such that the timing of the sequence of prominent syllables was approximately isochronous with the timing of the first syllable of the target word. Whether the fifth syllable was shortened or lengthened to accomplish this regularity depended on the structure of the preceding context: It was shortened when the distal context ended in one monosyllabic word and lengthened when it ended in two. The SW-biasing context was predicted to bias listeners to perceive the

initial syllable of the target word as lexically stressed (consistent with e.g. *jury* rather than *giraffe*). This bias was predicted to increase the initial proportion of fixations to the SW competitor when the target word began with an unstressed syllable, and to decrease the initial proportion of fixations to the WS competitor for SW target words.

Conversely, in the *WS-biasing condition*, the F0 of stressed syllables in the distal context was manipulated to have the opposite relative F0 level as the initial syllable of the target word (e.g. low, cf. Figure 2), and the duration of the fifth syllable was manipulated such that the timing of the sequence of prominent syllables was approximately isochronous with the timing of the syllables immediately preceding and following the first syllable of the target word. The WS-biasing context was predicted to bias listeners to perceive the initial syllable of the target word as unstressed, and therefore to increase the initial proportion of fixations to the WS competitor for target words beginning with a stressed syllable and to decrease the proportion of fixations to the SW competitor for target words beginning with an unstressed syllable.

**Procedure** The procedure was the same as Experiment 1.

**Analyses** Proportions of fixations to target, competitor, and distractor pictures on experimental trials were computed and averaged across the same time window as in Experiment 1. Effects of word stress, preceding prosody, and splicing condition on logit-transformed fixation proportions were analyzed using multilevel linear regression. Fixed effects and random slopes included picture type, target word type, distal prosody condition, splicing condition, trial number (i.e. the position of the item in the sequence encountered by the participant), the initial F0 of the target word (low vs. high), and interactions between factors. Since we were primarily interested in the effects of distal prosody on the relative proportions of fixations to the target and competitor pictures, fixations to the distractor pictures were not included in the analysis. Trial number was included in the analyses due to recent work suggesting that listeners rapidly adapt to the reliability of prosodic cues, particularly in counterbalanced experimental designs (e.g. Kurumada et al., to appear; Brown et al., under review). A full explication of prosodic adaptation effects, however, is beyond the scope of the present paper.

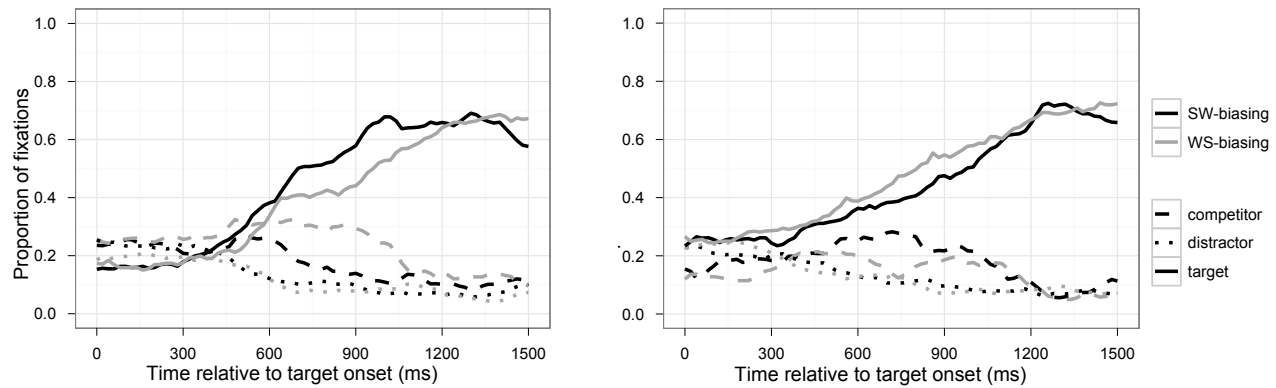


Figure 3: Proportion of fixations to target, competitor, and distractor pictures upon hearing SW (left) and WS (right) target words in SW- vs. WS-biasing contexts in Experiment 2. Line color denotes target word type; line texture denotes picture type.

## Results and discussion

Trials in which participants clicked on the incorrect picture (<2%) or for which participants generated an incorrect label for the target or competitor picture in the post-test (29.1%) were excluded from analysis. Figure 3 shows the proportion of fixations to the target, competitor, and distractor images as a function of condition starting at the onset of the target word. Starting at around 350 ms after the onset of SW target words (Figure 3, left), SW-biasing prosody resulted in higher proportions of target fixations and lower proportions of competitor fixations, with effects persisting for approximately 800 ms. For WS words (Figure 3, right), SW-biasing prosody had opposite effects, with relatively low proportions of target fixations and relatively high proportions of competitor fixations.

	<i>B</i>	<i>SE</i>	<i>t</i>	<i>p</i>
intercept	-.65	.09	-7.20	<.0001
target picture	.37	.12	3.11	<.005
WS-biasing prosody	.17	.09	1.91	<.1
trial number	-.06	.03	-2.06	<.05
target picture x WS-biasing prosody	-.29	.12	-2.39	<.05

Table 1: Parameters of the final multilevel regression model of logit-transformed fixation proportions in response to SW target words in Experiment 2. The model included by-item intercepts and slopes for picture type and prosody condition.

Multilevel linear regression analyses revealed not only the predicted three-way interaction between picture type, distal prosody condition, and target word stress pattern ( $B=.50$ ,  $SE=.25$ ,  $t=2.02$ ,  $p<.05$ ), but also a significant interaction between these three factors and trial number ( $B=-.46$ ,  $SE=.17$ ,  $t=-2.73$ ,  $p<.01$ ). To investigate the source of this significant four-way interaction, separate analyses were conducted for SW and WS words. Analysis of fixations during the processing of SW words revealed a significant interaction between target picture and prosody condition (Table 1). Participants were more likely to fixate the competitor picture in the WS-

biasing condition than in the SW-biasing condition ( $B=.17$ ,  $SE=.08$ ,  $t=2.13$ ,  $p<.05$ ); target fixations did not differ significantly by condition. Although trial number had a significant main effect, it did not enter into any significant interactions.

Analysis of fixations during the initial processing of WS words revealed not only a significant interaction between target picture and prosody condition, but also a significant three-way interaction between target picture, prosody condition, and trial number (Table 2). We explored this interaction by fitting two additional models to the data, with trial number centered one standard deviation above and below the mean. These models revealed that the interaction between target picture and prosody condition was significant early in the experiment ( $B=.56$ ,  $SE=.18$ ,  $t=3.18$ ,  $p<.005$ ). In the WS-biasing condition, target fixations were significantly higher ( $B=.32$ ,  $SE=.13$ ,  $t=2.49$ ,  $p<.05$ ), whereas competitor fixations were significantly lower ( $B=-.24$ ,  $SE=.10$ ,  $t=-2.46$ ,  $p<.05$ ). However, the interaction between target picture and prosody condition was not significant late in the experiment.

## General discussion

Results from two visual world experiments supported the hypothesis that expectations based on preceding prosody influence the perception of suprasegmental cues to lexical stress. Experiment 1 showed that initially unstressed words compete for recognition with phonemically overlapping words with initial stress, and vice versa. Experiment 2 further demonstrated that F0 and syllable timing patterns across material preceding the target word influence the relative activation of competing SW and WS alternatives. The activation of SW alternatives was higher when preceding stressed syllables had suprasegmental acoustic characteristics similar to the initial syllable of the target word, while the activation of WS alternatives was higher when preceding stressed syllables had suprasegmental characteristics dissimilar to the initial syllable of the target word.

These findings show that expectations about the acoustic realization of upcoming material within an utterance include

	<i>B</i>	<i>SE</i>	<i>t</i>	<i>p</i>
intercept	-.61	.06	-10.28	<.0001
target picture	.30	.10	3.05	<.005
WS-biasing prosody	-.10	.08	-1.20	>.1
trial number	-.10	.06	-1.60	>.1
target picture x WS-biasing prosody	.23	.12	1.96	<.05
target picture x trial number	.16	.08	1.93	<.1
WS-biasing prosody x trial number	.14	.08	1.70	<.1
target picture x WS-biasing prosody x trial number	-.33	.12	-2.82	<.005

Table 2: Parameters of the final multilevel regression model of logit-transformed fixation proportions in response to WS target words in Experiment 2. The model included random intercepts and picture type slopes for participants and items.

information about metrical organization and lexical stress, and that these expectations constrain the initial interpretation of suprasegmental stress cues during spoken word recognition. This indicates, in turn, that cues to lexical stress in sEnglish are not restricted to word-internal cues such as a syllable's F0, duration and amplitude, but can also include sentence-level patterning. This observation suggests that expectations from a variety of sources influence listeners' interpretation of suprasegmental stress cues.

The observation that prosodic expectations influence the interpretation of suprasegmental stress cues raises questions about the mechanisms by which various sources of contextual information are integrated with the incoming signal. Our findings are congruent with forward-modeling approaches in which perceptual input is evaluated with respect to internally generated hypotheses about the acoustic-phonetic realization of upcoming material. Forward models have been fruitfully explored within influential theories of motor control (e.g. Wolpert et al., 1995; Guenther & Micci Barreca, 1997), and may likewise provide a promising explanatory framework for aspects of spoken language understanding.

### Acknowledgments

This research was supported by a NSF predoctoral fellowship (MB), NSF grant BCS-0847653 (LCD), and NIH grants HD27206 and DC0005071 (MKT). We gratefully acknowledge Dana Subik for participant recruitment and testing.

### References

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.

Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin and Review*, 18, 1189–1196.

Brown, M., Salverda, A. P., Gunlogson, C., & Tanenhaus, M. K. (under review). Rapid integration of prosodic and discourse cues during spoken-word recognition.

Couper-Kuhlen, E. (1993). *English speech rhythm: Form and function in everyday verbal interaction*. Amsterdam: John Benjamins.

Cox, D. R. (1970). *The analysis of binary data*. London: Chapman and Hall.

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, 31, 218–236.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133–142.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40, 141–201.

Dilley, L., Mattys, S. L., & Vinke, L. (2010). Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. *Journal of Memory and Language*, 63, 274–294.

Dilley, L., & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, 59, 291–311.

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, 21, 1664–1670.

Guenther, F., & Micci Barreca, D. (1997). Neural models for flexible control of redundant systems. In P. Morasso and V. Sanguineti (eds.), *Self-organization, Computational Maps and Motor Control* (pp. 383–421). Amsterdam: Elsevier-North Holland.

Kurumada, C., Brown, M., & Tanenhaus, M. K. (to appear). Pragmatic interpretation of contrastive prosody: It looks like speech adaptation. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, 5, 253–263.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9, 666–688.

Niebuhr, O. (2009). F0-based rhythm effects on the perception of local syllable prominence. *Phonetica*, 66, 95–112.

Pierrehumbert, J. (2000). Tonal elements and their alignment. In M. Horne (ed.), *Prosody: Theory and experiment* (pp. 11–36). Dordrecht: Kluwer Academic Publishers.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632–1634.

Thomassen, J. M. (1982). Melodic accent: Experiments and a tentative model. *Journal of the Acoustical Society of America*, 71, 1596–1605.

Wolpert, D. M., Gharamani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, 269, 1880–1882.

Woodrow, H. (1911). The role of pitch in rhythm. *Psychological Review*, 18, 54–77.

# Eliciting a Sensemaking Process from Verbal Protocols of Reverse Engineers

Adam R. Bryant<sup>1,2</sup> (adam.bryant@wpafb.af.mil), Robert F. Mills<sup>2</sup> (robert.mills@afit.edu),  
Gilbert L. Peterson<sup>2</sup> (gilbert.peterson@afit.edu), and Michael R. Grimaila<sup>2</sup> (michael.grimaila@afit.edu)

<sup>1</sup>Cognitive Models and Agents Branch, 711th Human Performance Wing

<sup>2</sup>Department of Electrical and Computer Engineering, Air Force Institute of Technology  
Wright-Patterson AFB, OH 45433 USA

## Abstract

A process of sensemaking in reverse engineering was elicited from verbal protocols of reverse engineers as they investigated the assembly code of executable programs. Four participants were observed during task performance and verbal protocols were collected and analyzed from two of the participants to determine their problem-solving states and characterize likely transitions between those states. From this analysis, a high-level process of sensemaking is described which represents hypothesis generation and information-seeking behaviors in reverse engineering within a framework of goal-directed planning. Future work in validation and application of the process is discussed.

**Keywords:** Sensemaking; Information seeking; Human computer interaction; verbal protocol studies.

## Introduction

Sensemaking is a term used to describe a broad family of cognitive activities in which a person comes to develop a mental model to represent the elements of some situation of interest (Klein, Phillips, Rall, & Peluso, 2007; M. R. Endsley, 2000). Whereas situation awareness refers to the ability to attend to relevant information elements in a task as a situation unfolds (M. Endsley & Rodgers, 1994), sensemaking refers to the process that enables one to come to an understanding of the meaning and relevance of the elements that make up the situation (Klein, Moon, & Hoffman, 2006; M. R. Endsley, 2000). The sensemaking process has been described as an ongoing integration of knowledge from a mental model of a situation, available information about the context of a situation, and perceptual data from the environment (D. M. Russell, Stefik, Pirolli, & Card, 1993; M. R. Endsley, 2000).

Sensemaking is described as connecting inferences and observations, integrating knowledge and conjecture, finding explanations for ambiguous data, diagnosing ambiguous symptoms, and identifying problems (Klein et al., 2007). Sensemaking also refers to comprehension of the significance of ambiguous events and data in the environment (Weick, 1995). The many functions of sensemaking describe a class of distinct cognitive processes involving interactions between knowledge, information, and actions in an environment which may encompass a number of different inference and learning behaviors (Menzies, 1996; Josephson, Chandrasekaran, Smith, & Tanner, 1987). Sensemaking has been studied in naturalistic settings (Klein et al., 2007), in computer-human interaction (Pirolli & Card, 2005), in intelligence analysis (Zhang, Soergel, Klavans, & Oard, 2009), and in organizations (Weick, 1995), but no process-level description exists to characterize how people make sense of executable programs.

As part of a larger study to understand and model how people make sense of programs from executable representations, we wanted to understand the general interaction process that is involved as a person works with a debugger to make sense of a program. While sensemaking might be a general process of interacting with an environment to develop a mental model, we investigated sensemaking in a particular applied domain (reverse engineering assembly language) because it provides a restricted set of semantics to ground the investigation without being a “toy” problem. This was also essential because understanding sensemaking in that domain was the major focus of the larger study. So while the term ‘sensemaking’ has been used to mean many different things, we were interested in studying how people connect information from the task environment with background knowledge in order to develop and modify a mental model, all with a focus on applying this process to improve the information processing automation in software reverse engineering tools.

To understand this element of understanding executable programs, we undertook a study to collect verbal protocols from reverse engineers while they analyzed programs from assembly language representations. Four participants were observed performing a challenging reverse engineering task, and verbalizations from two of the participants were transcribed and coded. From the coded data, two participants’ state transitions were extracted to determine a process involving the state transitions.

First, the methods used in the study are briefly described. Next, the methods used in the analysis of verbal protocols is presented. Following that, the process of sensemaking used in the reverse engineering tasks is described in the context of complex problem-solving.

## Method

The participants were instructed to complete a problem called *Angler* that was downloaded from the *crackme.de* website (Schneider, T., 2011) in February 2011<sup>1</sup>. *Angler* is a type of crackme called a *keygen* (which stands for key generation), a type of program that typically presents text fields for the user to enter a name and a serial number and asks the participant to reverse engineer the algorithm which processes the user’s data. In keygen tasks, people disassemble and reverse engineer the program to discover a valid key for a given user

<sup>1</sup>Shortly after obtaining the crackme, the crackme.de website was taken down from the Internet. The *Angler* program can also be obtained from the website of the program’s author at <http://cyclops.ueuo.com>.

name or to write an algorithm that, when given a user name, produces a valid license key that unlocks the functionality of the program.

The Angler crackme program presents the participant with a semi-transparent visual window with two entries for text. It has three buttons, labeled “Check”, “About”, and “Exit.” When a user submits a name and serial number combination, the application tests whether it is a valid combination and informs the user.

### Selection of Participants

The researcher solicited reverse engineers to participate in a verbal protocol study through an e-mail invitation to the Air Force Institute of Technology and to a cross-organizational reverse engineering working group at Wright-Patterson Air Force Base. The solicitation requested that participants have knowledge of reverse engineering and experience using tools such as OllyDbg, WinDbg, Immunity Debugger, and the IDA Interactive Disassembler. The solicitation produced four reverse engineers from Wright-Patterson Air Force Base who participated without remuneration.

### Data Collection

Video data of the participants’ computer screen was captured and combined with audio recordings of the participants’ verbalizations during the task. The participant computer ran a Windows XP operating system that was hosted within a VirtualBox virtual machine, and which was preloaded with the software tools and documentation identified from a previous subject matter expert study of the reverse engineering task. The experimenter viewed the participant’s task environment remotely over a virtual network computing (VNC) connection using the TightVNC server and client software (TightVNC Software, 2011). The experimenter’s computer was outfitted with a microphone which was wired into the participant’s cubicle to enable the combined collection of audio and video data.

Each participant was seated at a small cubicle in an unoccupied, quiet room in front of the participant computer which contained a mouse, keyboard, monitor, and the microphone. The participants were instructed as to the different reverse engineering tools and documentation available, and were permitted as much time as was needed to become familiar with the task environment. Paper was also available so participants could make notes, as recommended by Wood (1997).

Each participant was given instructions on how to verbalize thoughts during task performance and was instructed not to try to explain the task or thought processes. The experimenter stressed this point and demonstrated examples of poor, acceptable, and high-quality concurrent verbalizations with a simulated coffee-making task to help the participants understand how they should verbalize during task performance. During the performance of each task, the experimenter was seated out of the participant’s view in the opposite cubicle and reminded participants to verbalize when they fell silent for more than a few seconds using simple prompts such

as “please remember to verbalize during the task” and “remember to talk aloud” as discussed in Trickett and Trafton (2007). Other than those reminders, the experimenter was silent throughout the task, and took notes about each participants’ goals, strategies, concepts, and problems in the task.

Audio data of each participant’s concurrent verbalizations and video data of the experimenter’s computer monitor (showing all actions on the participant’s computer monitor) were recorded using the CamStudio software. After the task, participants were asked to recall what they thought their strategies were, what parts of the task they thought were the most difficult, what would have made the task easier, and what they felt they needed to pay attention to.

### Overview of the Reverse Engineering Task

Participants were to investigate, modify, and re-implement an algorithm that an executable Windows program called “Angler.exe” (Cyclops, 2011) uses to process a user-provided serial number. Participants worked from assembly language representations of the program provided through the IDA interactive disassembler and the OllyDbg debugger without access to source code or debugging symbols.

The code in Angler consists of 15 major subroutines, including the subroutine called `WinMain` which starts the windowing process and other subroutines to present dialog boxes. The program runs within a single thread of execution in memory and does not have hidden sections, encrypted code, or code obfuscations. Angler’s file header contains pointers to four program sections that are mapped into memory at run time: the `.text`, `.rdata`, `.data`, and `.rsrc` sections, which are typical for portable executable programs compiled to run on Win32-based operating systems (Eilam, 2005).

Though there are many strategies to approaching the challenge, successful completion required that participants do the following:

1. Read and understand the goal of the task
2. Determine that a system function handles input
3. Isolate the input handling function
4. Determine the format for the serial number input
5. “Catch” the input as the program executes
6. Craft data so the program executes the success message
7. Translate the function into pseudocode
8. Write pseudocode for a key generator

An algorithm in the program composed of 27 basic blocks of assembly instructions processes the user’s serial number. The algorithm takes the first four characters of the person’s name and performs a cyclic redundancy check (CRC) to produce an even-numbered value from that character, finds four pairs of prime factors that sum to the even-numbered values, and assembles an eight-value number separated by dashes (Cyclops, 2011). Through a series of programmatic checks, the algorithm determines if the user’s serial number matches the computed serial value. In these checks, the last instructions in each basic block of code check for a data value from

Verbalization Describes	Coded As
Desired future state	Make goal representation
Activities to accomplish goal	Plan approach
Status of an ongoing activity	Carry out plan
Noticing something	Sense information
Recognizing relevance	Interpret information
More abstract statement	Update knowledge
An assumption	Create hypothesis
Question about something	Create hypothesis

Table 1: Rules Used to Code Segments

the input and then transfer the program's execution based on the result of that check. If the reverse engineer does not understand the meanings of these behaviors, the program appears to be making a large number of arbitrary numerical checks in a long sequence of assembly language instructions.

### Verbal Protocol Analysis

Participant B's video and audio data recordings were accidentally destroyed during a problem with saving the video file to disk and were not able to be recovered or transcribed. Additionally, Participant A lacked familiarity with the tools and the task and was not able to make sense of the program. Although observations of all of the participants provided valuable context and examples, only verbal data from Participants C and D were coded and analyzed to elicit the sensemaking process.

Concurrent verbal protocols were collected from Participants C and D following the method outlined in (Ericsson & Simon, 1980). We reviewed the video and verbal data from Participants C and D and transcribed it into a spreadsheet, broken into one verbal segment per row as discussed in Trickett and Trafton (2007). The participants' verbalizations were segmented during transcription in order to take advantage of other contextual clues from the audio and video. Verbalizations were segmented based on whether they represented a single idea, and when a segment contained a shift from one idea to another, the second idea was recorded in its own row as its own new segment. Where significant verbal breaks occurred, the subsequent verbalizations were recorded on a new row as a separate segment.

After all of the available data was transcribed, it was coded according to the following taxonomy of sensemaking in reverse engineering, established from a previous literature review of sensemaking in reverse engineering in Bryant et al. Bryant, Mills, Peterson, and Grimaila (2012): *Make the goal representation, Plan an approach, Carry out a plan, Sense information, Interpret information, Update knowledge, and Create a hypothesis*. This taxonomy of sensemaking steps is similar to the information-processing loop used in programming artificial agents to interface with an environment (S. Russell & Norvig, 2003), with the addition of a state to generate goals and state to generate hypotheses.

The data from the two participants were coded according to

the coding rules in Table 1. To ensure that the coding scheme was appropriate for the data, interrater reliability was computed between two independent coders. One researcher coded all of the data (592 segments) and afterwards a second coder independently coded 29.2 percent of the segments (173 sequential segments) without having seen the original coder's data, and from a starting point randomly selected by the second coder. Cohen's Kappa statistic (Cohen et al., 1960) was computed to measure interrater reliability for the 29.2 percent of segments coded by both coders. Cohen's Kappa measures the agreement between coders on positive and negative instances while taking into account the likelihood of agreement based on chance. Cohen's Kappa is computed as:

$$\kappa = (P_o - P_c) / (1 - P_c)$$

$P_o$  is the proportion of agreements between the coders and  $P_c$  is the proportion of agreement which would be predicted by chance. Generally, a Cohen's Kappa value of 0.0 to 0.4 indicates zero to very little agreement, 0.6 to 0.8 indicates significant agreement, and 0.8 and above represents near perfect agreement, though there is disagreement in the literature about specific ranges of values (Bakeman & Gottman, 1997; Trickett & Trafton, 2007). After both coders independently coded the data, the interrater reliability was calculated and the coders met to discuss disagreements. If codes had weak interrater reliability (0.4 or below), the categories were removed or changed and the data was recoded. The final interrater reliability for the dual-coded verbalizations was 0.82, which demonstrates significant to "near perfect" agreement in all categories (Table 2). Following standard practice, the remainder of the verbalizations were coded by the researcher (Trickett & Trafton, 2007).

### Computing State Transitions

The state transitions from the two participants' data were computed to determine how the reverse engineers made sense of the programs. Transitions between the states indicated movement through the problem-solving process

State	Category	Cohen's Kappa
<b>a</b>	Make goal representation	0.93
<b>b</b>	Plan approach	0.82
<b>c</b>	Carry out plan	0.78
<b>d</b>	Sense information	0.72
<b>e</b>	Interpret information	0.75
<b>f</b>	Update knowledge	0.81
<b>g</b>	Create hypothesis	0.92
	Average agreement	0.82

Table 2: Interrater Reliability of Coding Scheme (173 segments)

As described in Bakeman and Gottman (1997), matrices of state transition probabilities were computed to determine the process used in the task. For  $m$  states  $S_j$  and  $S_k$ , and  $n$  seg-

ments  $i$ , the total transitions between each state  $S_j$  and state  $S_k$  are computed as:

$$Tr(S_j, S_k) = \sum_{i=1}^n (S_{i,j} \times S_{i+1,k}) : S \in (0, 1) \quad (1)$$

The total transitions departing a state  $S_j$  are computed as:

$$Tr(S_j, out) = \sum_{k=1}^m Tr(S_j, S_k) \quad (2)$$

The total transitions entering a state  $S_k$  are computed as:

$$Tr(in, S_k) = \sum_{j=1}^m Tr(S_j, S_k) \quad (3)$$

The overall transition probabilities for state  $S_j$  to  $S_k$  are computed as:

$$P(Tr(S_j, S_k)) = \frac{1}{2} \left( \frac{Tr(S_j, S_k)}{Tr(S_j, out)} + \frac{Tr(S_j, S_k)}{Tr(in, S_k)} \right) \quad (4)$$

Using these equations to compute the state transitions, the state transition probabilities for the two participants are shown in Table 3 and Table 4.

	a	b	c	d	e	f	g
a	0.17	0.24	0.10	0.10	0.15	0.00	0.04
b	0.14	0.20	0.21	0.21	0.03	0.00	0.12
c	0.10	0.13	0.05	0.33	0.09	0.10	0.00
d	0.10	0.08	0.12	0.37	0.37	0.10	0.14
e	0.00	0.11	0.25	0.22	0.25	0.28	0.18
f	0.05	0.00	0.00	0.10	0.10	0.21	0.29
g	0.23	0.15	0.07	0.10	0.17	0.07	0.38

Table 3: Transition Probability Matrix (Participant C)

	a	b	c	d	e	f	g
a	0.21	0.26	0.22	0.15	0.14	0.00	0.04
b	0.18	0.19	0.26	0.18	0.09	0.00	0.11
c	0.08	0.12	0.07	0.22	0.11	0.10	0.13
d	0.08	0.21	0.09	0.45	0.28	0.18	0.18
e	0.18	0.06	0.08	0.22	0.24	0.20	0.03
f	0.13	0.03	0.07	0.12	0.10	0.31	0.16
g	0.25	0.07	0.04	0.15	0.03	0.08	0.10

Table 4: Transition Probability Matrix (Participant D)

The mean transition probabilities were computed as  $1/N \sum_i P(Tr(S_j, S_k))$  for  $N$  participants. The mean transition probability was  $\mu = 0.14$  and the standard deviation was  $\sigma = 0.09$ . The threshold for significance was set at  $\mu + \sigma = 0.23$ . The significant transitions at the threshold  $P(Tr) \geq 0.23$  are shown in Figure 1.

Because only two participants' verbalizations were coded, inferences cannot be made about the how the processes from these two samples apply to the broader population of reverse

engineers or the broader sensemaking process. Nevertheless, previous studies have used observations and verbal data from small samples during exploratory research as way to generate hypotheses which are to be verified with further investigation and more participants (Newell & Simon, 1972; Trickett & Trafton, 2007). The verbal protocols from the two participants and the observed problem-solving processes from all four participants are useful to provide a framework for understanding how reverse engineers make sense of executable programs.

## Discussion

Figure 1 shows a process of how the sensemaking behaviors were used by reverse engineers attempting to solve the Angler task. When the participants were working on problems in the task, they continually moved through a loop of activities, which included the establishment of a goal representation, a plan to achieve the goal, carrying out the actions of the plan, sensing information from the task environment, interpreting the information, potentially updating knowledge if the information was relevant, and developing hypotheses based on the new knowledge. The sensemaking loop in Figure 1 shows this cycle as a Markov model populated with the probabilities that each state transition would occur.

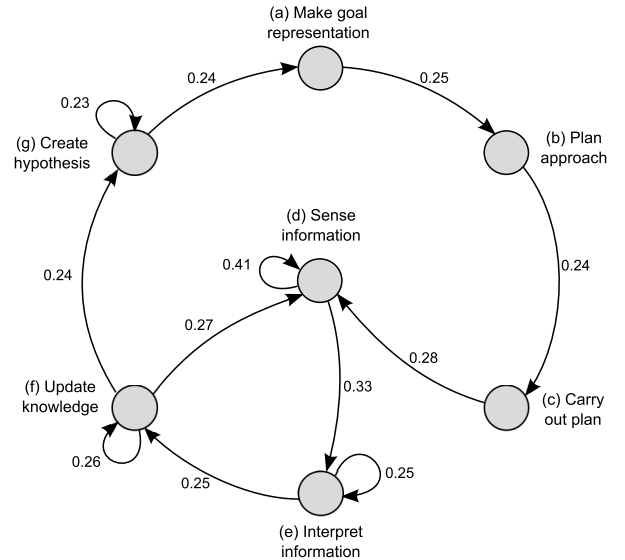


Figure 1: Sensemaking Processes from the Reverse Engineering Task

While progressing through this process, the participants gathered information to help them construct and refine their mental models of the Angler program. As the participants worked on the problem, they gained information about components such as functions, the program's execution paths, data in the program, and sequences of instructions. They were then able to relate these elements to items in the task environment.



## Plans, Goals, and Actions

Once the participants had verbalized a goal, this often was immediately followed with the development of a partially-constructed plan of actions which would enable the attainment of the goal. If the goal was to gather information from the task environment, the plan involved actions which the person could use to help gather the required information. Likewise if the goal was to configure the program or the reverse engineering tools in a particular way, then the plan involved sequences of actions which led the person to be able to change elements in the situation.

In some cases, a participant's goal did not directly lend itself to a plan, so the person deliberated over different ideas to construct and evaluate an approach that would generate a usable plan. Sometimes the deliberation was verbalized, and others it was inferred by the presence of long pauses.

Participants appeared to determine the best actions for their situation by thinking through hypothesized behaviors and inferred future states of the program. When participants verbalized plans, their plans involved sets of actions in which some of the actions were sequenced. Sometimes the participants did not order their actions into a sequence until they were already taking some initial actions and detecting conflicts in how those actions would affect the state of the situation.

Other plans the participants expressed included searching information from the reverse engineering tool about text strings, debugging the program to see how the program's memory stack changes, labeling a function so it can be identified later, inserting a breakpoint after the initialization routine to "catch" the program, and tracing data as it flows through a function during simulated program execution.

## Information Processing and Interpretation

Participants processed information by actively seeking it out and by passively noticing information during the performance of some other action. When participants passively sensed the information, they executed the Angler program to gather information about its behaviors or looked through the disassembled code to gather clues that might be useful.

When participants actively sought out information, they set a goal for the information they were interested in obtaining, made a plan to acquire that information, and followed the plan by carrying out actions in the task environment. In one example, a participant wrote down the addresses of system calls and used the search features of the debugger to find each of the calls.

Participants actively sought out information about the program's behavior by isolating phenomena. To do this, participants used the debugger to move the program's execution past a system call (which performs some action for the program), and then looked through the current register values and disassembled code in the debugger to determine what had changed in the state of the program.

In either active or passive information processing, the participants seemed to perform some initial processing to deter-

mine whether the information was relevant to one of their goals, and consequently ignored or attended to the information. If participants deemed the information as relevant, they interpreted it to connect it to concepts they already understood. If the information element and its connected concept provided the person with a new perspective or more insight that was relevant to one of their goals, the person verbalized a phrase indicating they had updated their knowledge. These phrases were typically summarized or distilled statements that captured the essence of how the different concepts were related.

During coding discussions, the second coder characterized this process as when the person "compiled" their information in an analogous manner to how a compiler converts a program's source code into executable code.

## Mental Models and Hypotheses

Once participants had added new summarized knowledge to their mental model of a program, they often came up with a hypothesis directly afterward. Participants appeared to generate hypotheses in the task after deducing the logical conclusions of new knowledge they had acquired.

Hypotheses and assumptions were generated mainly after participants sensed and interpreted information and updated their knowledge with the implications of this information. The participants' hypotheses took the form of verifiable statements such as: "it looks like `GetDlgItem` creates a handle to some part of the dialog that's open." In this case, the participant started a subsequent loop through the sense-making process with the goal of verifying whether or not the `GetDlgItem` function creates such a Window handle.

The hypotheses that resulted from this process were typically used to generate a new goal, such as to seek out information from the environment to confirm or refute a fact or to investigate another line of investigation about how the program works. However, when participants did not progress through the process to the development of a hypothesis, they were not able to generate information-seeking goals and got "stuck" in the task. In these cases, participants reverted to exploring instructions or behaviors of the code, since they did not have specific hypotheses about the Angler program to investigate.

## Conclusions

A process of sensemaking in reverse engineering was described and characterized through observations and analysis of verbal protocols from participants reverse engineering programs from assembly language representations. Verbal protocols were analyzed to extract state transition patterns from two reverse engineers' performance, and from that data a process of sensemaking was elicited and the steps of that process were described in a theory of how people make sense of executable programs. Participants were observed forming goals, creating plans to achieve their goals, and carrying out plans. Participants also sensed information from the task environment, interpreted the information, updated their mental

model with the information, and generated hypotheses from that integration which led to new goals.

This research represents a necessary step toward increasing the autonomy of reverse engineering tools, and can be used to determine a general theory of sensemaking that can improve the ways in which people interact with other complex systems. The work is limited in that it only involved observations of four participants and collection of verbal protocols from two participants, and its generality is potentially limited by the nature of the task; nevertheless the results provide insight which can be used to develop a more general computational theory of sensemaking through future empirical study.

Future work is needed to determine the generality of how people work through this process in similar information-processing tasks. Future research is also needed in determining how these processes can be realized as models within established computational cognitive architectures. Finally, work is needed in employing these findings to improve humans' interactions with complex systems such as reverse engineering tools.

### Acknowledgments

The Sensors Directorate at Wright-Patterson Air Force Base supported this research. The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

### References

- Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge: Cambridge University Press.
- Bryant, A., Mills, R., Peterson, G., & Grimaila, M. (2012). (in press) software reverse engineering as a sensemaking task. *Journal of Information Assurance and Security*.
- CamStudio Developers. (2011). *Camstudio*. Available from <http://camstudio.org>
- Cohen, J., et al. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Cyclops. (2011). *Crackmes by cyclops*. Available from <http://cyclops.ueuo.com/crackme.html>
- Eilam, E. (2005). *Reversing: Secrets of Reverse Engineering*. Wiley.
- Endsley, M., & Rodgers, M. (1994). Situation awareness information requirements analysis for en route air traffic control [Conference proceedings (article)]. In *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting* (Vol. 38, pp. 71–75).
- Endsley, M. R. (2000). Situation models: An avenue to the modeling of mental models. In *Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting*.
- Ericsson, K., & Simon, H. (1980). Verbal reports as data. *Psychological Review*, 87(3), 215.
- Hex-Rays. (2011). *The IDA Pro Disassembler and Debugger*. Available from <http://www.hex-rays.com/idaipro/>
- Immunity, Inc. (2011). *Immunity Debugger*. Available from <http://www.immunityinc.com/products-immdbg.shtml>
- Josephson, J. R., Chandrasekaran, B., Smith, J. W., & Tanner, M. C. (1987). A mechanism for forming composite explanatory hypotheses. *IEEE Transactions on Systems, Man, and Cybernetics*, 17(3), 445–454.
- Klein, G., Moon, B., & Hoffman, R. (2006). Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems*, 21(4), 70–73.
- Klein, G., Phillips, J., Rall, E., & Peluso, D. (2007). A data-frame theory of sensemaking. In *Expertise Out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making* (pp. 113–155).
- Menzies, T. (1996, September). Applications of abduction: Knowledge-level modelling. *International Journal of Human-Computer Studies*, 45(3), 305–335.
- Microsoft, Inc. (2011). *WinDbg*. Available from [http://msdn.microsoft.com/en-us/library/ff561300\(v=vs.85\).aspx](http://msdn.microsoft.com/en-us/library/ff561300(v=vs.85).aspx)
- Newell, A., & Simon, H. (1972). *Human Problem Solving*. Prentice-Hall Englewood Cliffs, NJ.
- OllyDbg. (2011). *OllyDbg*. Available from <http://www.ollydbg.de/>
- Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of the International Conference on Intelligence Analysis* (pp. 2–4).
- Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI93)* (pp. 269–276). New York, New York, USA: ACM Press.
- Russell, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education.
- Schneider, T. (2011, jan). *crackme.de*. Available from <http://crackme.de>
- TightVNC Software. (2011). *TightVNC*. Available from <http://www.tightvnc.com>
- Trickett, S., & Trafton, J. (2007). A primer on verbal protocol analysis. In D. Schmorow, J. Cohn, & D. Nicholson (Eds.), *The PSI Handbook of Virtual Environments for Training and Education: Developments for the Military and Beyond*. Westport, CT Praeger Security International.
- Weick, K. (1995). *Sensemaking in Organizations*. Sage Publications, Inc.
- Wood, L. E. (1997, March). Semi-structured interviewing for user-centered design. *Interactions*, 4(2), 48–61.
- Zhang, P., Soergel, D., Klavans, J. L., & Oard, D. W. (2009, June). Extending sense-making models with ideas from cognition and learning theories. *Proceedings of the American Society for Information Science and Technology*, 45(1), 23–23.

# Cooing, Crying, and Babbling: A Link between Music and Prelinguistic Communication

Michael Byrd, Casady Bowman, and Takashi Yamauchi

(mybrd@neo.tamu.edu, casadyb@neo.tamu.edu, takashi-yamauchi@tamu.edu)

Department of Psychology, Mail Stop 4235

Texas A&M University, College Station, TX 77843 USA

## Abstract

Like language, the human capacity to create music is one of the most salient and unique markers that differentiates humans from other species (Cross, 2005). In the following study, the authors show that people's ability to perceive emotions in infants' vocalizations (e.g., cooing and babbling) is linked to the ability to perceive timbres of musical instruments. In one experiment, 180 "synthetic baby sounds" were created by rearranging spectral frequencies of cooing, babbling, crying, and laughing made by 6 to 9-month-old infants. Undergraduate participants (N=145) listened to each sound one at a time and rated the emotional quality of the "synthetic baby sounds." The results of the experiment showed that five acoustic components of musical timbre (e.g., *roll off*, *mel-frequency cepstral coefficient*, *attack time* and *attack slope*) could account for nearly 50% of the variation of the emotion ratings made by undergraduate students. The results suggest that the same mental processes are probably applied for the perception of musical timbres and that of infants' prelinguistic vocalization.

**Keywords:** Emotion; Language; Music

## Introduction

Infants use a variety of vocal sounds, such as cooing, babbling, crying, and laughing, to express their emotions. Infants' prelinguistic vocal communications are highly affective in the sense that they evoke specific emotions—happiness, frustration, anger, hunger, and/or joy—without conveying concrete ideas. In this sense, infants' vocal communication parallels music. Music is highly affective; yet it is conceptually limited (Cross, 2005; Ross, 2009).

The interaction between music and language has attracted much attention recently (Chen-Haffteck, 2011; Cross, 2001; Masataka, 2007). However, despite their similarities, little attention has been paid to the relationship between music and prelinguistic vocalizations (Chen-Haffteck, 2011; Cross, 2001; He, Hotson, & Trainor, 2007; Masataka, 2007). If music and language are highly related, what is the relationship between infants' vocal communications such as babbling, and music?

In the study described below, we analyze acoustic cues of infants' vocalization and demonstrate that emotions created by prelinguistic vocalization can be explained to a large extent by the acoustic cues of sound that differentiate timbres of musical instruments, potentially implicating that the same mental processes are applied for the perception of musical timbres and that of infants' vocalizations.

The paper is organized as follows: we review related work examining the link between prelinguistic vocalization

and music followed by an overview of the experiment. After discussing our timbre extraction and sound creation method, we introduce one experiment that investigates the connection between music and prelinguistic communication.

## Related Work

Infants begin life with the ability to make different sounds—first cooing and crying, then babbling. Next they form one word, and then two, followed by full sentences and speech. In the first ten months, infants progress from simple sounds that are not expressed in the phonetic alphabet, to babbling, which is an important step in infants learning how to speak (Gros-Louis, West, Goldstein, & King, 2006; Oller, 2000).

Musical instruments and infants' vocalizations both elicit emotional responses, while conveying little information on what the sender is trying to express. Music can have a very powerful effect on its listeners, as we all have a piece of music that will bring back emotions. Music can convey at least three universal emotions, happiness, sadness and fear (Fritz et al., 2009). These emotions are similar to the emotions expressed by infants with their limited sounds (Dessureau, Kurowski, & Thompson, 1998; Zeifman, 2001; Zeskind & Marshall, 1998). Both infants and music convey meaning without the use of words. Infants rely on their voices and non-verbal/non-word sounds to communicate and it is these sounds that inform the listener of how important and of what type of danger the infant is facing, such as being too cold, hungry or of being left alone (Dessureau et al., 1998; Zeifman 2001; Zeskind & Marshall, 1998).

Across cultures, songs sung while playing with babies are fast, high in pitch, and contain exaggerated rhythmic accents, whereas lullabies are lower, slower and softer. Infants will use cues in both music and language to learn the rules of a culture. Motherese, a form of speech used by adults in interacting with infants, often consists of singing to infants using a musical, sing-song voice, that mimics babies' cooing by using a higher pitch. An infant's caregiver will use higher pitch when speaking to an infant, as it helps the infant learn and also draws their attention (Fernald 1989).

In summary, research shows that there is a close link between infants' vocal communication and music. This link is demonstrated through the babbling and cooing sounds used by infants' to communicate, and also by mothers' use of motherese to assist infant's learning of language in a sing-song manner. Infants are able to use the same cues

from both music and language to facilitate learning in both domains. Given these close connections, it is likely that the same mental processes are involved for the perception of instrumental sounds and the perception of infants' vocalizations. The beginning stages of this idea are investigated in one experiment by examining the emotion perception of synthetic baby sounds.

## Overview of the Study

In the Emotion Rating Experiment described below, we tested the general hypothesis that the same mental process is involved for the perception of infants' vocalization and that of timbres of musical instruments. More specifically, we hypothesize that the acoustic components of timbre will be significant predictors of emotion. If this is true, then there should be a plausible link between musical timbre and prelinguistic vocal timbre, also indicating a link for mental processing in the two domains. We employed an audio synthesizer program and created 180 different "synthetic baby sounds" by combining spectral frequencies of real baby sounds. In the experiment, our undergraduate participants ( $N=145$ ) listened to the "synthetic baby sounds" one at a time and rated affective qualities of these sounds. Later, we extracted "musical timbres" from the synthetic baby sounds, and examined the extent to which the emotional ratings made by our undergraduate students were accounted for by the timbres of the synthetic baby sounds.

Timbre is an important perceptual feature of both music and speech. Timbre is defined as the "acoustic property that distinguishes two sounds"—for example, those of the flute and the piano—"of identical pitch, duration, and intensity" (Hailstone et al., 2009; McAdams & Cunible, 1992). The classic definition of timbre states that two different timbres result from the sound of different amplitudes (of harmonic components) of a complex tone in a steady state" (Helmholtz, 1885). Timbre is a sound quality that encompasses the aspect of a sound that is used to distinguish it from other sounds of the same pitch, duration, and loudness.

The timbre properties of *attack time*, *attack slope*, *zero-cross*, *roll off*, *brightness*, *mel-frequency cepstral coefficients*, *roughness*, and *irregularity* are well known in music perception research as the main acoustic cues that correlate with the perception of timbre of musical instruments (Hailstone et al., 2009). Our assumption is that if infants' vocal sounds are perceived in the same manner as the timbres of musical instruments are perceived, these same acoustic properties can account for the perception of emotions in infants' vocalization.

Using principal components analysis (PCA), we summarized emotional ratings made by our undergraduate participants into two principal dimensions, to reduce the data, and applied a stepwise regression to evaluate the extent to which our predictors—the acoustic timbre components—accounted for emotion ratings for synthesized baby sounds.

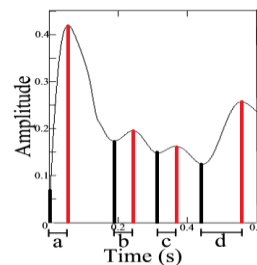
Below, we briefly describe our timbre extraction method and the method of creating "synthetic baby sounds."

## Timbre Extraction

This section describes acoustic cues relating to timbre in detail, as well as the computational procedure of extracting these cues. The purpose of using these acoustic cues is to act as predictors in regression analyses that can explain perceived emotions of our "synthetic baby sounds." The acoustic cues were chosen based on their use in musical timbre (see Lartillot & Toivainen, 2007).

Eight acoustic properties of timbre: attack time, attack slope, zero-cross, roll off, brightness, mel-frequency cepstral coefficients, roughness, and irregularity were extracted from all sound stimuli using MIRToolbox in Matlab (Lartillot, Toivainen, & Eerola, 2008). These acoustic properties are known to contribute to the perception of timbre in music independently of melody and other musical cues (Hailstone et al., 2009). The acoustic features were extracted from synthesized sounds rated in the Emotion Rating Experiment.

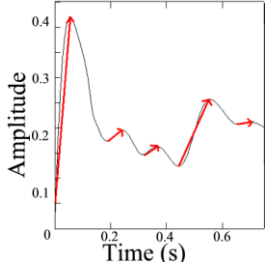
*Attack time* is the time in seconds it takes for a sound to travel from amplitude of zero, to the maximum amplitude of a given sound signal, or more simply the temporal duration. Some features of timbre such as attack time contribute to the perception of emotion in music (Gabrielsson & Juslin, 1996; Juslin, 2000; Loughran, Walker, O'Neill & O'Farrell, 2001); which suggests that features of timbre can at least in part determine the emotional content of music (Hailstone et al., 2009).



**Figure 1.** Attack times of an audio file. *A* through *d* are separate attack times; indicated by the distance from the black line, to the red line.

Attack time is computed using the equation of a line,  $y = mx + b$ , it is part of a sounds amplitude envelope where  $m$  is the slope of the line and  $b$  is the point where the line crosses the vertical axis ( $t=0$ ). Figure 1 gives an illustration of attack time. The horizontal segments below the x-axis indicate the time it takes in seconds to achieve the maximum peak of each frame for which the attack time was calculated.

*Attack slope* is the attack phase of the amplitude envelope of a sound, also interpreted as the average slope leading to the attack time. This can also be calculated using the equation of a line  $y = mx + b$ , where  $m$  is the slope of the line and  $b$  is the point where the line crosses the vertical axis ( $t=0$ ), see Figure 2. The red line in Figure 2 indicates the slope of the attack.



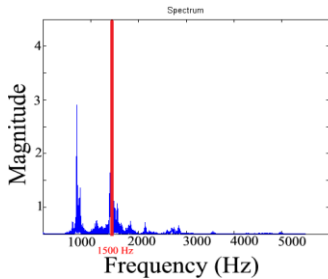
**Figure 2.** Attack slope of a audio file. The red arrow indicates the duration (attack time) for which the attack slope is calculated.

$$Z_t = \frac{1}{2} \sum_{n=1}^N |sign(x[n]) - sign(s[n-1])|$$

*Roll off* is the amount of high frequencies in a signal, which is specified by a cut-off point. The roll-off frequency is defined as the frequency where response is reduced by -3 dB. This is calculated using the following equation where  $M_t$  is the magnitude of the Fourier transform at frame  $t$  and frequency bin  $n$ .  $R_t$  is the cutoff frequency, see Figure 3.

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{R_t} M_t[n]$$

*Brightness* is the amount of energy above a specified frequency, typically set at 1500 Hz – this is related to spectral centroid. The term brightness is also used in discussions of sound timbres, in a rough analogy with visual brightness. Timbre researchers consider brightness to be one of the strongest perceptual distinctions between sounds. Acoustically it is an indication of the amount of high-frequency content in a sound, and uses a measure such as the spectral centroid, see Figure 3.



**Figure 3.** Brightness of an audio file. To the right of the red dashed line is the amount of energy above 1500 Hz, or the brightness of the sound.

peaks, dissonant sounds have irregularly placed spectral peaks as compared to consonant sounds with evenly spaced spectral peaks.

Formally, roughness is calculated using the following equation where  $a_j$  and  $a_k$  are the amplitudes of the

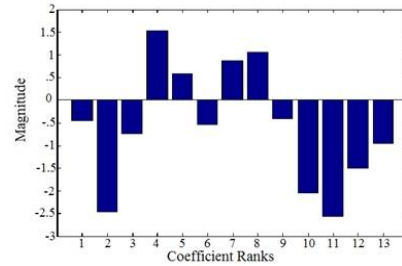
*Zero-cross* is the number of times a sound signal crosses the x-axis, this accounts for noisiness in a signal and is calculated using the following equation where sign is 1 for positive arguments and 0 for negative arguments.  $X[n]$  is the time domain signal for frame  $t$ .

components, and  $g(f_{cb})$  is a ‘standard curve.’ This was first proposed by (Plomp & Levelt, 1965).

$$\rho = \frac{\sum_{j,k}^n a_j \cdot a_k \cdot g(f_{cb})}{\sum_j^n a_j^2}$$

Following extraction of the value for roughness from the sound stimuli, principal components analysis was used to reduce the dimensions of the roughness data.

*Mel-frequency Cepstral Coefficients* (mfcc) represent the power spectrum of a sound. This power spectrum is based on a linear transformation from actual frequency to the Mel-scale of frequency. The Mel scale is based on a mapping between actual frequency and perceived pitch as the human auditory system does not perceive pitch in a linear manner. Mel-frequency cepstral coefficients are the dominant features used in speech recognition as well as some music modeling (Logan, 2001). Frequencies in the Mel scale are equally spaced, and approximate the human auditory system more closely than a linearly spaced frequency bands used in a normal cepstrum. Due to large data output, prior to analyses mfcc data were reduced using principal components analyses to create a workable set of data. A cutoff criterion of 80% was used to represent the variability in the original mfcc data. Figure 4 shows the numerical Mel-frequency cepstral coefficient rank values for the 13 mfcc components returned. Thirteen components are returned due to the concentration of the signal information in only a few low-frequency components.



**Figure 4.** Mel-frequency cepstral coefficients (mfcc) of an audio file. This figure shows the acoustic component mfcc. Each bar represents the numerical (rank coefficient) value computed for the thirteen components returned.

*Irregularity* of a spectrum is the degree of variation between peaks of a spectrum (Lartillot et al., 2008). This is calculated using the following equation where irregularity is the sum of the square of the difference in amplitude between adjoining partials in a sound.

$$\frac{\sum_{k=1}^N (a_k - a_{k+1})^2}{\sum_{k=1}^N a_k^2}$$

## Creating Synthetic Baby Sounds

We created 180 short, 2 second, synthetic baby sounds from ten real infant sounds: five males and five females ranging from ages 6 to 9 months making screaming, laughing, crying, cooing and babbling sounds. These sounds were chosen to create novel stimuli emulating human prelinguistic sounds. Among these sounds, four (one screaming boy, one crying boy, one screaming girl and one crying girl) were audio-recorded directly from two volunteer infants in Nacogdoches, Texas using an Olympic Digital Voice WS-400S recorder. The sounds of babbling and cooing boys and girls were taken from audio-files downloaded from a sound effects website (<http://www.freesounds.org>), and the sounds of laughing boy and girl were taken from files downloaded from YouTube (<http://www.youtube.com>).

These infant sounds were decomposed by four laboratory assistants into amplitude and spectral frequency components by applying fast Fourier transform using a sound editing software program (SPEAR, Klingbeil, 2005). Arbitrarily chosen spectral frequencies of one sound (e.g., a babbling sound of a boy) were mixed with arbitrarily chosen spectral frequencies of another sound (e.g., cooing girl) and then modified by means of amplitude, or shifting frequencies, to convey one of the basic emotions, happy, sad, anger, or fear (Ekman, 2002).

For each sound pair, four sounds were created to sound happy, sad, angry, and fearful. In this manner, each sound pair (45 pairs in total, all possible pairs of the 10 real sounds), was used to create four affective sounds, which was decided subjectively by the laboratory assistants. The total 180 sound stimuli were normalized and white noise was taken out prior to and after creation of each sound stimulus.

## Emotion Rating Experiment

The goal of the experiment was to obtain empirical ratings of college students examining the emotional quality of the synthetic baby sounds that we created. To analyze the link between emotion ratings and acoustic cues, a stepwise regression analysis was employed.

**Participants.** A total of 145 undergraduate students (73 males, 73 females) participated in this experiment for course credit. Participants were randomly assigned to one of two groups that listened to 90 or 89 sounds of 179 total sounds. Stimuli were randomly assigned to one of two groups; no participants were in both groups.

**Materials.** Stimuli were taken from the 180 synthetic baby sounds that were created from a group of a total of ten recorded real infants' sounds (see the "Creating Synthetic Baby Sounds" section for the details of the sound creation).

**Procedure.** Participants were presented with 90/89 sounds using customized Visual Basic software through JVC Flats stereo headphones. Each stimulus's maximum volume was

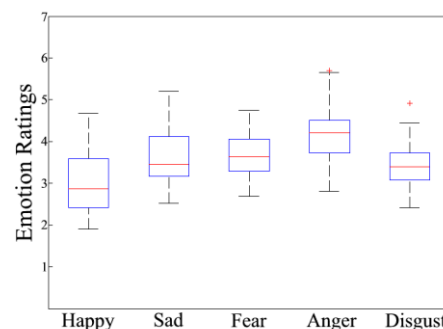
adjusted and normalized. Participants were instructed to listen to sound stimuli, and rate each sound on five emotion categories, happy, sad, angry, fearful, and disgusting (Ekman, 1992; Johnson-Laird & Oatley, 1989). Each scale ranged from 1 to 7—1 being *strongly disagree* (the degree to which the stimuli, sounded like one of the five emotions), and 7 being *strongly agree*. Stimuli were presented in a random order.

## Results

This section starts with descriptive statistics of emotion ratings followed by the results from stepwise regression analysis, which examined the extent to which emotion ratings given to the synthetic baby sounds were explained by their timbre properties. For the regression analysis, average emotion scores were calculated for individual synthetic sounds by collapsing over individual participants, yielding a 179 sounds x 5 emotion dimension matrix. By applying principal component analysis (PCA), this matrix was summarized in a 179 x 2 matrix with the two columns corresponding to two principal components identified by the PCA procedure. The first two orthogonal components explained 88.1% and 7.1% of the variance of the emotion rating data, respectively.

*Descriptive Statistics.* Behavioral data, Figure 5, shows overall observations for each emotion from the emotion rating data. From the whiskers of the box plot for the emotion data, it is apparent that there is variation within the data. The highest rating for the emotion data did not exceed a value of 6, on the scale of 1-7. The median of the ratings for emotion varied between approximately 2.5 and 4.75 within the emotion rating data.

For all 179 sounds rated, most were rated as angry, indicated by the median of the data for anger. The sounds were rated least like the emotion happy, as the median for this emotion was the lowest for all sounds rated on the five emotions.



**Figure 5.** Box plot of observations for emotion ratings. Each box in the figure indicates one emotion rated by participants. The median is indicated by the red line in the center of each box, and the edges indicate the 25<sup>th</sup> and 75<sup>th</sup> percentiles, the whiskers of each plot indicate the extreme data points, and outliers are plotted outside of the whiskers.



**Regression analysis.** A step-wise regression analysis was used to analyze the collected rating data and timbre components, to determine which component could best explain the emotion rating data. Seventeen total predictors were used in the stepwise regression to analyze the emotion ratings made by participants. These were *attack time*, *attack slope*, *zero-cross*, *roll off*, *brightness*, *mel-frequency cepstral coefficients 1-6*, *roughness 1-4*, and *irregularity*. Due to large data output, mfcc data were reduced using principal components analyses to create a workable set of data. There were originally 13 numerical Mel-frequency cepstral coefficient rank values returned. These 13 rank values were reduced to 6, accounting for 78% of the total mfcc data. Roughness was also reduced in the same way using PCA, from 79 components to four components that described 80% of the original roughness data. These predictors were used to analyze the emotion ratings made by participants.

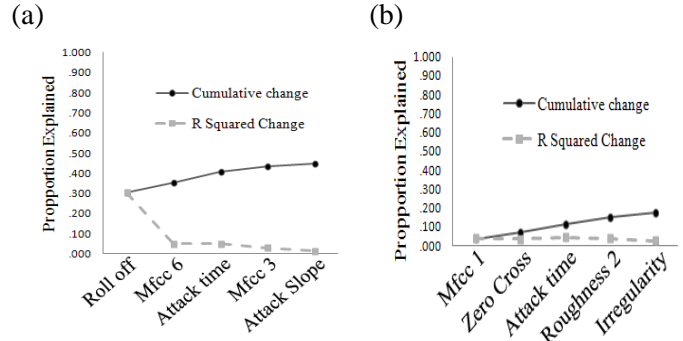
The results of the regression for the first principal component (PCA1) indicated four acoustic features significantly predicted emotion ratings; roll off ( $\beta = -.386$ ,  $p < .001$ ), mfcc 6 ( $\beta = .218$ ,  $p < .001$ ), attack time ( $\beta = .248$ ,  $p < .001$ ), and mfcc 3 ( $\beta = -.202$ ,  $p < .002$ ), and attack slope ( $\beta = .034$ ,  $p < .034$ ), see Table 1 for percent explained by principal component 1.

**Table 1:** Significant acoustic components for emotion PCA 1 and PCA 2

Predictors	PCA 1	PCA 2
% explained	88%	7.1%
Attack time	.23***	.31***
Attack slope	.12*	
Irregularity		-.16*
Mfcc 1		-.24**
Mfcc 3	-.19**	
Mfcc 6	.22***	
Roughness		.21**
Zero-cross		.25**
Roll off	-.41***	

\*  $p < .05$ , \*\*  $p < .01$ , and \*\*\*  $p < .001$ .

The second principal component (PCA 2) showed that five acoustic features significantly predicted emotion ratings; mfcc 1 ( $\beta = -.244$ ,  $p < .001$ ), zero cross ( $\beta = .250$ ,  $p < .002$ ), attack time ( $\beta = .305$ ,  $p < .000$ ), roughness 2 ( $\beta = .208$ ,  $p < .006$ ), and irregularity ( $\beta = -.159$ ,  $p < .024$ ), (Table 1).



**Figure 6.** R-squared. Emotion judgment principal components 1 (PCA 1) and 2 (PCA 2). This figure shows the proportion of R-squared contributed for each addition of a predictor to the model for principal component I and II from the emotion judgments.

Figure 6 shows the proportion of R-squared contributed for each addition of a predictor to the model for PCA 1 – (a) and PCA 2 – (b). Looking at the values of R-squared, it is apparent that roll-off was best able to describe the emotion ratings, accounting for 30% of the emotion ratings for PCA 1. The second principal component does show several significant acoustic cues that predict emotion; however, none are as strong as in the first principal component.

## General Discussion

Music and language are perhaps two of the most cognitively complex and emotionally expressive sounds invented by humans. Recently, the evolutionary origins of music and language have attracted much attention in researchers of a broad spectrum (Cross, 2001, 2005; Hauser et al., 2002; Kirby, 2007). The present study, examining the relationship between infants' vocalizations—cooing, babbling, crying and screaming—and the perception of musical timbres, suggests that the link between music and language can go even further back to the prelinguistic level of development.

Our Emotion Rating Experiment indicates that nearly 50% of emotions created by synthetically produced infant sounds can be explained by a small number of acoustic cues pertaining to musical timbres. Among those, *roll off*, which quantifies the amount of high frequencies in a signal, turned out to be the most important cue. The second most important property, *mfcc* (*mel-frequency cepstral coefficients*), corresponds to perceived pitch in the human auditory system, and are the dominant features used in speech recognition and music modeling (Logan, 2001). Given these findings, we conjecture that high-frequency sounds are probably taken as the robust cue of emotion attribution, and more fine-grained distinctions of emotion are made by extracting speech-related cues.

The ability to discriminate sounds is said to be present even in primitive animals such as carp (Chase, 2001), implying that this ability evolved early in history. Some animals have sounds and or calls that can convey the emotions of finding something of interest or of fear (Hauser,



Chomsky, & Fitch, 2002). Such abilities were probably present even before music was fully developed in the current form.

## Acknowledgments

We would like to thank Na Yung Yu and Ricardo Gutierrez-Osuna for their valuable comments. The first two authors, MB and CB, contributed to this study an equal amount and the order of their authorship was determined by a coin toss.

## References

- Chase, A. R. (2001). Music Discriminations by carp (Cyprinus carpio). *Animal Learning and Behaviour*, 29, 336-353.
- Cross, I. (2001). Music, mind and evolution. *Psychology of Music*, 29, 95-102.
- Cross, I. (2005). Music and meaning, ambiguity and evolution. In D. Miell, R. MacDonald, D. Hargreaves (Eds.), *Musical Communication* (pp. 27-43). New York: Oxford University Press.
- Dessureau, B. K., Kurowski, C. O., & Thompson, N. S. (1998). A reassessment of the role of pitch and duration in adults' responses to infant crying. *Infant Behavior and Development*, 21, 367-371.
- Ekman, P. (1992). Are there basic emotions? *Psychological Review*, 99, 550-553.
- Fernald, A. (1989). Intonation and Communicative Intent in Mothers' Speech to Infants: Is the Melody the Message? *Child Development*, 60, 1497-1510.
- Fritz, T., Jenschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Koelsch, S. (2009). Universal Recognition of Three Basic Emotions in Music. *Current Biology*, 19, 573-576.
- Gabrielsson, A., & Juslin, P. N. (1996). Emotional expression in music performance between the performer's intention and the listener's experience. *Psychology of Music*, 24, 68-91.
- Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, 30, 112-119.
- Hailstone, J. C., Omar, R., Henley, S., Frost, C., Kenward, M., & Warren, J. D. (2009). It's not what you play, it's how you play it: Timbre affects perception of emotion in music. *Quarterly Journal of Experimental Psychology*, 62, 2141-2155.
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, 22, 1569-1580.
- He, C., Hotson, L., & Trainor, L. J. (2007). Mismatch Responses to Pitch Changes in Early Infancy. *Journal of Cognitive Neuroscience*, 19, 878-892.
- Helmholtz, H. v. (2005). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. London: Longmans, Green, and Co.
- Johnson-laird, P. N., & Oatley, K. (1989). The language of emotions: An analysis of a semantic field. *Cognition & Emotion*, 3, 81-123.
- Juslin, P. N. (2000). Cue utilization in communication of emoting in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 1797-1813.
- Kirby, S. (2007). The evolution of language. In R. Dunbar & L. Barrett (Eds.), *Oxford handbook of evolutionary psychology* (pp. 669-681). Oxford: Oxford University Press.
- Klingbeil, M. (2005). Software for spectral analysis, editing, and synthesis *Proceeding of the ICMC* (pp. 107-110). Barcelona Spain.
- Koelsch, S. (2005). Neural substrates of processing syntax and semantic in music. *Current Opinion in Neurobiology*, 15, 207-212.
- Lartillot, O., Toivainen, P., Eerola, T. (2008) A Matlab Toolbox for Music Information Retrieval. In, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker (eds.), *Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Organization*, (pp 261-268). New York: Springer.
- Logan, B., & Robinson, T. (2001). Adaptive model-based speech enhancement. *Speech Communication*, 34, 351-368.
- Loughran, R., Walker, J., O'Neill, M., O'Farrell, M. (2001). The Use of Mel-frequency Cepstral Coefficients in Musical Instrument Identification. in Proc. of the 6th International Conference on Music Information Retrieval (ISMIR), 2005, Finland, (pp. 1825-1828).
- Masataka, N. (2007). Music, evolution and language. *Developmental Science*, 10, 35-39.
- McAdams, S., & Cunible, J. C. (1992). Perception of timbral analogies. *Philosophical Transactions: Biological Sciences*, 9, 336-383.
- Olivier Lartillot, Petri Toivainen, "A Matlab Toolbox for Musical Feature Extraction From Audio", *International Conference on Digital Audio Effects*, Bordeaux, 2007.
- Oller, D. K. (2000). *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum
- Plomp, R., & Levelt, W. J. M. (1965). *Tonal consonance and critical bandwidth*. Soesterberg: Institute for Perception RVO-TNO, National Defense Research Council T.N.O.
- Ross, B. (2009). Challenges facing theories of music and language co-evolution. *Journal of the Musical Arts in Africa*, 6, 61-76.
- Zeskind, P., S., & Marshall, T., R. (1988). The Relation between Variations in Pitch and Maternal Perceptions of Infant Crying. *Child Development*, 59, 193-196.

# Mothers Do Not Drive Structure in Adult Homesign Systems: Evidence from Comprehension

**Emily M. Carrigan (emily.carrigan@uconn.edu)**

Department of Psychology, 406 Babbidge Road  
Storrs, CT 06269-1020 USA

**Marie Coppola (marie.coppola@uconn.edu)**

Department of Psychology, 406 Babbidge Road  
Storrs, CT 06269-1020 USA  
Department of Linguistics, 337 Mansfield Road  
Storrs, CT 06269-1145 USA

## Abstract

Some profoundly deaf individuals without conventional linguistic input develop gestures, called “homesign,” to communicate. We examined homesign systems (HSs) used by four deaf Nicaraguan adults (ages 15-27), and evaluated whether homesigners’ mothers are potential sources for these systems. Study One measured mothers’ comprehension of descriptions of events (e.g., “A man taps a woman”) produced in homesign and spoken Spanish. Mothers comprehended spoken Spanish descriptions (produced by a hearing child) better than homesign descriptions, suggesting a greater degree of sharedness for spoken Spanish. Study Two compared the homesign comprehension of each homesigner’s mother to that of a native user of American Sign Language (ASL). ASL Signers performed better than mothers, confirming that homesign productions contain comprehensible information, to which mothers are not fully sensitive. Taken together, these results suggest that mothers are not the source of their deaf child’s HS, and add to evidence that HSs are more like language than like gesture.

Keywords: Language acquisition; homesign; deafness; language creation; gesture; sign language; Nicaragua

## Introduction

The language sciences have long grappled with the question of what drives language acquisition. At the heart of this debate is the question: What are the contributions of language input versus the contributions of the learner? It can be difficult to disentangle these two factors in typical language acquisition situations, as such situations do not offer the opportunity to experimentally manipulate the presence of linguistic input.

Studying spontaneously occurring cases of degraded linguistic input can help discern human predispositions for language learning. Previous research has shown that children can surpass their linguistic input (e.g., Singleton & Newport, 2004, Senghas & Coppola, 2001). In addition, some deaf children born into hearing families have no access to signed or spoken linguistic input. While their parents primarily speak, and do not use gesture with them, these children nevertheless use a system of manual gestures, called “homesign,” to communicate. Homesign has many, but not all, of the features of fully developed languages (e.g., a stable lexicon, basic syntax and morphology, Goldin-Meadow,

2003). The gestures and gesture combinations produced by the mothers of child homesigners lack the morphological and syntactic structure observed in the children’s productions (e.g., Goldin-Meadow & Mylander 1984, 1990; Goldin-Meadow et al. 1994, Goldin-Meadow 2003). While mothers’ gestures may serve as an initial foundation for their deaf child’s homesign system, children surpass whatever “input” they might receive from their mother. These mothers tended not to engage in gestural communication with the child homesigners, and such child homesigns are used for a relatively short time, until the children reach school age. It is possible that, given increased gestural communication and a lengthier period of use, mothers might play a greater role in the development of homesign systems.

This research examines homesign systems developed *with communication partners who engage homesigners using gesture*, unlike the young deaf children studied by Goldin-Meadow and her colleagues. In Nicaragua, we can locate rare cases in which deaf individuals develop and use homesign systems as their primary means of communication for their entire lives. These individuals are not part of the Deaf community that uses the recently emerged Nicaraguan Sign Language (NSL) (Senghas, 1995; Senghas & Coppola, 2001). The homesigners do not learn conventional sign language, and they have not acquired the Spanish spoken around them.

These mature homesign systems display many linguistic features, for instance, the grammatical relation of Subject (Coppola & Newport, 2005), proto-pronouns (Coppola & Senghas, 2010), use of space, and devices for establishing reference (Coppola & So, 2005). Given the accumulated interactions over time between a homesigner and his or her family members, and the more abstract, language-like devices that homesigners produce, it is possible that these family members may have contributed more to the development of the homesign system than the mothers of the homesigners Goldin-Meadow & colleagues observed. One step in determining the source of the linguistic features we observe in mature homesign is evaluating family members’ potential contributions.

## Study One

We begin by evaluating homesigners’ mothers as potential sources, because they each have significant gestural

communication experience with their deaf child, and because this type of transmission (mother to child) parallels that of typical language acquisition situations. One approach would be to compare Mothers' and homesigners' gesture productions. However, comparing their productions now does not allow us to assess Mothers' role in the *development* of their child's homesign system, especially in cases where Mothers' and homesigners' productions are similar. Looking instead at how well Mothers comprehend homesign productions can address this. We reason that, if mothers served as models for homesigners' abstract linguistic devices, they should comprehend their child's homesign productions.

### Study One: Predictions

If mothers invent and pass down homesign systems to their deaf children in the same way that they serve as models for the spoken Spanish acquired by their hearing children, we would expect mothers to comprehend descriptions of events produced by their deaf child at least as well as they comprehend spoken Spanish descriptions of the same events produced by one of their hearing children.

### Participants

**Producers:** Four deaf adult Nicaraguan homesigners (1 female), ages 16-26 *at the time of production*, produced the descriptions used as stimuli for this task. All four homesigners were deaf, with *very* minimal knowledge of spoken or written Spanish. Some could produce and/or comprehend a limited number of common spoken Spanish words, such as "mamá," "papá," and "agua" (water). All find writing their names effortful. They had had little to no formal education, had not acquired Nicaraguan Sign Language (NSL), and *do not interact with each other*.

Four hearing siblings of homesigners (1 female), ages 17-43. The siblings were native monolingual Spanish speakers, had an average of 8.5 years of education (range 0-14), and had not acquired NSL.

**Receivers:** Four hearing mothers (henceforth "Mothers") of homesigners, ages 45-60. The mothers were native monolingual Spanish speakers, had an average of 2.25 years of education, and had not acquired NSL.

### Materials

The stimuli were **descriptions** of 83 simple videotaped events involving live actors and real, everyday objects. The events had one or two participants; the two-participant events included all combinations of animate and inanimate. The two animate participants in the events were the same man and woman throughout, and the inanimate participants were objects such as "cup," "banana," and "flower." Example events include "A man kisses a woman" and "A sheet of paper falls".

The comprehension array used in this task included four pictures. One picture always depicted the target event. For one-participant events, the non-target foil pictures could

depict: a) the same participant/object involved in a different action *or* state ("Other Action"); b) a different participant/object involved in the same action/state ("Other Entity"); or c) a different participant/object involved in a different action/state ("Unrelated"). For two-participant events, the non-target foil pictures could depict: a) the same participants involved in reversed thematic roles ("Reverse"); b) one participant involved in the same action with a different entity ("Other Entity"); c) the same two participants involved in a different action ("Other Action"); or d) one participant involved in an unrelated action (either with or without a second entity; "Unrelated"). Because these materials were originally designed as an elicited production task (Coppola & Newport, 2005, which also lists the stimulus items), the comprehension arrays are not standardized across all items, and contain different combinations of foil types.

**Homesign descriptions** were produced by the homesigners described above. We videotaped these descriptions, then clipped and compiled them into QuickTime video files.

**Spoken Spanish descriptions** of these same events were produced by a hearing sibling of each homesigner, in the presence of their Mothers.

### Procedure

Each Mother watched the videotaped homesign descriptions (83 total) produced by her own deaf child. The task is divided into two subtests, each beginning with 3 practice items, to ensure that mothers understood how to do the task<sup>1</sup>. Mothers watched each description as many times as they wanted, then selected, from an array of four pictures, the picture that best matched that description. One picture was a still from the target (correct) event, and the others were distracters.

Mothers also completed the task using spoken Spanish descriptions of events produced by one of their hearing children in real time. The order of the homesign and spoken Spanish tasks was counterbalanced.

### Results & Discussion

The Mothers comprehended homesign descriptions at rates significantly better than chance (25%; exact binomial test,  $p < 0.001$ ).

However, despite performing above chance, Mothers comprehended spoken Spanish descriptions better than they comprehended homesign descriptions of the same events (For 3 mothers,  $p < 0.05$ , McNemar's Test for Correlated Proportions; fourth mother,  $p = 0.057$ ). This result acts as a built-in control, showing that Mothers are not having trouble with the task itself, but rather with the content of the homesign descriptions.

---

<sup>1</sup> We are confident that mothers understand how to do the task, as all of them have completed it in the past, though typically with live descriptions from their deaf child.

Comparing Receivers' comprehension to different reference levels of performance (e.g., 25%, 33%, 50%) can give us clues as to how much of the descriptions (homesign or spoken Spanish) they understand. For example, in an event like "A man kisses a woman," the picture choices show: a) A man pushing a chair; b) A man sitting; c) A man kissing a woman (the correct choice); and d) A woman kissing a man. One homesigner's description of this event was glossed as MAN WOMAN KISS. If the Receiver understands the gestures for the participants, or even just the action gesture the homesigner produced for this event description, picture choices (a) and (b) could be eliminated. It is also possible for Receivers to narrow their choices based solely on a general, non-linguistic strategy. They might have noticed, for instance, that two of the picture choices contained the same two actors, engaged in the same action (although in different thematic roles), and reasoned that the correct choice must be one of those two pictures.

The performance of three of the four Mothers does *not* differ significantly from 50%. Regardless of the strategy they might be using to complete the task, Mothers' performance indicates that they do not understand enough of the homesign description to reliably select the right picture; however, we may not be able to make claims from these data about exactly what Mothers do understand. In future work, will more carefully control participants' cognitive ability (including use of general strategies to complete tasks such as these), and be designed to isolate which aspects of the homesign descriptions mothers do understand (see the Results & Discussion section of Study Two for a brief attempt at the latter with these data).

Mothers' poorer comprehension of homesign versus spoken Spanish descriptions suggests that Mothers play a different role in the communicative development of their hearing and deaf children: they share, and are likely a main source of spoken Spanish input for their hearing children, but do not share or transmit homesign to their deaf children. This finding accords with previous studies of the structure in child homesign, which is also not attributable to the deaf children's mothers (Goldin-Meadow & Mylander, 1984).

## Study Two

Each mother has had between 20 and 30 years of experience communicating with her deaf child; why should their comprehension levels be so low? One could argue that the homesign descriptions themselves are the cause; perhaps Mothers fully understood the descriptions, but the descriptions themselves did not contain sufficient information for mothers to succeed at the task. It also might be the case that factors such as age of exposure to a visual communication system, and quantity and/or length of time of experience with that system, play a role in comprehension.

### Study Two: Predictions

If ASL Signers comprehend homesign productions at levels equal to or worse than homesigners' Mothers, it might be

the case that those productions do not contain sufficient information to allow *any* Receiver to succeed at the task. If, however, ASL Signers comprehend homesign descriptions better than Mothers, the descriptions do contain enough information to succeed at the task.

## Participants & Methods

Four fluent Deaf users of ASL (3 females), ages 21-66, who did not know the homesigners or their homesign systems, participated in this study. The ASL Signers had all been exposed to ASL before the age of five, used ASL every day, and had an average of 15.25 years of education. We paired each ASL Signer with one homesigner's Mother; that ASL Signer watched the same homesigner's productions as did the Mother, and chose, from the same picture array, the photo that matched each event description.

Unlike the Mothers, at the start of the task the ASL Signers saw the 6 practice items, which all involved the man and/or the woman, to ensure that they had learned the homesigner's lexical items for "Man" and "Woman." As previously mentioned, the man and the woman in the events were always played by the same male and female actor; thus, neither the producers nor receivers in the task ever had to distinguish one man or woman from another. ASL Signers were, as Mothers were, allowed to watch each description as many times as they wanted (they watched most descriptions no more than once). ASL Signers and Mothers thus had roughly equal exposure to these stimuli (although each Mother still had vastly more experience with her deaf child's homesign system than the ASL Signer with whom she was matched).

## Results & Discussion

Like homesigners' Mothers, ASL Signers comprehended homesign descriptions at rates significantly better than chance (25%; exact binomial test,  $p < 0.0001$ ). Furthermore, ASL Signers comprehended the homesign descriptions they viewed better than that homesigner's Mother (For 3 pairs,  $p < 0.01$ , McNemar's Test for Correlated Proportions; fourth pair,  $p = 0.851$ ). Thus, the homesign descriptions *did* contain sufficient information to allow a receiver to successfully complete the task. Mothers did not succeed for some other reason; we explore this further in subsequent analyses.

**Item Type and Error Analyses** In order to better understand which aspects of homesign production drive comprehension (or non-comprehension) by receivers, we examined how features of the items themselves might influence comprehension of homesign descriptions. The events had one or two animate or inanimate participants; all events with two animate participants, that is, "Reversible" events (e.g. "A man kisses a woman"), included a distracter picture that depicted reversed roles for the participants (e.g. "A man kisses a woman").

To show correct comprehension of "Reversible" event descriptions Receivers need to understand how the arguments represented by the lexical items relate to the

verb. That is, they need to understand the structure of these descriptions. Comprehension of “Non-Reversible” event descriptions, in contrast, only necessitates that the Receiver recognize and remember the lexical items produced by the homesigner.

On the Non-Reversible events (collapsing across number of participants,  $n=52$  items), all four ASL Signers performed significantly better than chance (25%; exact binomial test,  $p<0.0001$ ). Mother also performed significantly above chance on this subset of items ( $p<0.01$ ).

Comparing Mothers to ASL Signers, we see the same pattern as for the overall analyses: the same three out of four ASL Signers do significantly better than the Mothers with whom they are paired at comprehending the non-reversible events (For three pairs,  $p<0.001$ , McNemar’s Test for Correlated Proportions; fourth pair,  $p=0.359$ ). This indicates that even when comprehension relies only on recognizing and remembering the lexical items, Mothers did not succeed. This is particularly surprising given that: a) Mothers were allowed to view the descriptions as many times as they wanted; and b) Mothers have had much more practice with homesigners’ lexical items (indeed, with each homesign system in general) than did the ASL Signers (20-30 years interacting with the homesigner and using the homesign system, as opposed to one hour viewing homesign descriptions for the ASL Signers). The Mothers’ apparent lack of comprehension of lexical items does not necessarily indicate that they did not recognize them—they might have difficulty processing the lexical items in real time, even with repeated viewings.

Table 1 presents an analysis of the incorrect foil types chosen for Non-Reversible items (again, collapsing across 1- and 2-Participant items).

Table 1: Errors on Non-Reversible Items  
(Proportion of each foil type chosen)

Participant	Unrelated	Other Entity	Other Action	Total Number of Errors
ASL 1	0.22	0.22	0.56	9
Mother 2	0.31	0.45	0.24	29
ASL 2	0.38	0.38	0.25	8
Mother 2	0.23	0.77	0.00	13
ASL 3	0.33	0.00	0.67	6
Mother 3	0.43	0.30	0.26	23
ASL 4	0.13	0.38	0.50	8
Mother 3	0.32	0.48	0.20	25

Each Mother chose the “Other Entity” foil more often than the ASL signer with whom she was paired. This indicates that Mothers are poorer than ASL Signers at understanding or remembering the homesign gestures produced for the participants in these items.

Comparing Mother-ASL Signer pairs on the Reversible 2-participant events ( $n=16$ ), only 1 ASL signer performed significantly better than the Mother with whom he was paired ( $p<0.05$ , McNemar’s Test for Correlated

Proportions). However, this lack of a difference between Mothers and ASL Signers may be due to the small number of items on which the comparison is based.

If, as discussed in Study One, Receivers understand something about the lexical items the homesigner produces in a description (either for the participants *or* the action in an event), they should be able to narrow their picture choices to two for the Reversible items. Previous work with three of the four homesigners who produced these event descriptions has demonstrated that they reliably place the noun phrase expressing the subject in clause-initial position (Coppola & Newport, 2005). If Receivers are further sensitive to the systematic way homesigners represent argument structure, we would expect to see comprehension of these reversible event descriptions at rates significantly above 50%.

Table 2: Proportion Correct on Reversible Items

Participant	Number Correct	Number of Reversible Items	Difference from 50%, Exact Test, p-value (2-tailed)
ASL1	8	16	0.598
Mother 1	9	16	0.402
ASL 2	11	15	0.059†
Mother 2	11	15	0.059†
ASL 3	15	16	<0.001*
Mother 3	8	16	0.598
ASL 4	11	15	0.059†
Mother 4	5	15	0.941

\*: significant †: marginally significant

Three out of the four ASL signers performed significantly or marginally better than 50% correct on the reversible items, compared with only one of the Mothers (Table 2).

An analysis of the incorrect foil types chosen for Reversible items can be found in Table 3.

Table 3: Errors on Reversible Items  
(Proportion of each foil type chosen)

Participant	Unrelated	Other Entity	Other Action	Reverse	Total Errors
ASL 1	0.13	0	0	0.88	8
Mother 1	0.14	0.14	0	0.71	7
ASL 2	0	0	0	1.00	4
Mother 2	0	0	0	1.00	4
ASL 3	0	0	0	1.00	1
Mother 3	0	0.13	0	0.88	8
ASL 4	0	0	0	1.00	4
Mother 4	0.30	0.10	0	0.60	10

Both Mothers and ASL Signers chose the “Reverse” foil most often when they erred, although Mothers showed slightly more variability than ASL Signers. The fact that Mothers mostly selected the “Reverse” foil means that they were likely narrowing the picture choices for this type of item to just the “Correct” and “Reverse” pictures. As

outlined in the Results & Discussion section of Study One, for most of the Reversible items, it is possible to narrow choices using three strategies: 1) By understanding the gesture produced for the two participants in the event; 2) By understanding the gesture produced for the action in the event; or 3) By using a general test-taking strategy that assesses the similarity of the different foil pictures. The data from the Error Analysis of Non-Reversible items indicate that Mothers do *not* always recognize or remember the gestures produced for the participants in an event, which allows us to eliminate the first candidate strategy. It is likely, therefore, that Mothers are able to narrow their options for the correct picture either via an understanding of the action or event gesture produced by the homesign (which tends to be highly iconic), or via a direct comparison of the picture foils to one another.

The results of the comparison to fifty percent and the Error Analysis for Reversible items indicate that, although ASL Signers' comprehension of homesign descriptions is not error-free, they comprehended enough of both the lexical items and the structure to outperform the homesigners' Mothers. Indeed, the ASL Signers' errors are understandable, since their experience with the homesign systems is so limited.

The success of ASL signers indicates that the homesign descriptions do contain systematic, comprehensible information. Homesigners' Mothers, despite their much greater experience with the individual homesign systems, are apparently not sensitive to the information that ASL Signers are presumably using to succeed in the task.

## General Discussion

Taken together, the results of Studies 1 and 2 suggest that Mothers do not directly transmit homesign systems to their deaf children. Mothers' comprehension of their adult child's homesign system is relatively poor; this lack of understanding apparently persists despite the fact that Mothers report regularly using gesture to communicate since their now-adult offspring were children. These findings do not preclude the possibility that Mothers' gestures served as a foundation for their child's homesign system. However, these data do indicate that Mothers are *not masters* of their adult child's homesign system, which tells us that, at present, homesign is qualitatively different for Mothers than it is for homesigners. This, in turn, suggests that it is homesigners rather than Mothers who drive the development of their homesign systems.

Our failure to find a statistically significant difference between Mothers' and ASL Signers' comprehension of reversible events is likely due to the small number of items of that type. We are currently designing new stimuli that will include greater numbers of these informative events.

Our dataset is obviously limited by the small number of homesigning participants, and the logistical difficulty of working with them. Though we have a small number of Mother-ASL Signer pairs, each participant contributes a relatively large number of data points. A hierarchical linear

model will allow us to better account for sources of variability at the levels of Item (Reversible vs. Non-reversible events); Receiver Type (ASL Signer, Homesigner Sibling, etc.) and Homesigner Family Unit.

Mothers' poor comprehension of their child's homesign systems compared with that of ASL Signers raises several questions: Which factors distinguish Mothers from ASL Signers? Which of these factors drives comprehension (or non-comprehension) of homesign descriptions?

One difference between Mothers and ASL Signers is their length of experience with the homesign system itself; Mothers have significantly greater experience with the homesign system than do ASL Signers. However, given this, we would expect mothers to comprehend homesign production *better* than ASL Signers, which they do not.

Mothers and ASL Signers also differ in their age of exposure to a visual communication system. This factor could explain Mothers' poorer comprehension of homesign. Brentari, Coppola, Mazzoni & Goldin-Meadow (2012) show that the handshapes produced by homesigners pattern more like those of established sign languages than like the gestures produced by hearing individuals in terms of phonological features (specifically, finger complexity in Object and Handling handshapes used in classifier predicates). These comprehension data could provide further evidence that homesign is closer to a linguistic system than to gesture; that is, homesign must be acquired beginning at an early age in order to reach proficiency (Johnson & Newport, 1989; Newport, 1990).

The mother of Homesigner 2, who shows the best comprehension compared to the ASL Signer with whom she is paired, began using homesign with her son at an earlier age than the other mothers in our study (she was 16 years old when he was born). Although well beyond the critical period for language acquisition established by Newport and colleagues, her relative youth might have conferred a crucial advantage in acquiring her deaf child's homesign system.

It might also be the case that ASL Signers' early and significant experience with an established visual-manual language helped them perform better than Mothers in comprehending homesign productions. Perhaps the ASL Signers are drawing on their (implicit) knowledge of how visual languages are structured to understand homesign.

Comparing our two current groups with two additional groups could further distinguish the effects of the age of exposure to a visual communication system, and the type of system (homesign vs. an established language), on homesign comprehension (summarized in Table 4).

First, we can measure the comprehension of homesign by siblings of those homesigners. The siblings of homesigners who are close to them in age likely started using homesign at a young age to communicate with their deaf sibling. Comparing homesigners' siblings to their mothers will tell us whether early exposure to homesign can also drive better comprehension. In addition, comparing the comprehension of homesigners' siblings to that of native ASL Signers will reveal whether the nature of the visual communication

matters (e.g., homesign vs. an established visual language like ASL).

Second, we can look at comprehension of homesign by signers who acquired ASL later in life (e.g. in adolescence or beyond). In our current groups of native ASL Signers, age of exposure is confounded with knowledge of a complex, established language. Measuring the comprehension of late-learning ASL signers can tell us whether (and if so, how much) experience with an established visual language supports homesign comprehension.

Table 4: Comparison groups and their characteristics

Group	Type of visual system	Age of Exposure	Length of exposure
Mothers of Homesigners	HS	late	significant
Native ASL Signers	ASL	early	significant
Siblings of Homesigners	HS	early	significant
Non-native ASL signers	ASL	late	Range from minimal to significant

These comparisons will help elucidate the nature of homesign systems. If more experience with an complex, established visual communication system (ASL) supports better comprehension of homesign, we can, in conjunction with data regarding the systematicity of homesign production, provide further support that homesign systems are themselves linguistic. Moreover, if better comprehension of homesign is predicted by factors that are associated with acquiring linguistic systems (namely, age of exposure), such results would accord with Brentari et al.'s findings, providing evidence that homesign systems are more like language than like gesture.

More work must be done to create a full and accurate characterization of the structure and development of homesign systems. If such research supports our claim that homesigners and not their mothers drive the development of homesign systems, this will have interesting implications regarding the capacity of the human brain for language. To the degree homesigners *are* innovating their systems and to the degree that their systems resemble existing visual languages, we can say there is something in the learner that is capable of producing language. As others have suggested (e.g., Senghas & Coppola 2001), it may be that the capacities that evolved to support language acquisition are also capable of creating language, to some degree. Future work involving converging methods—spontaneous and elicited production and comprehension—and different populations—such as homesigners and different cohorts of users of Nicaraguan Sign Language—will help further clarify the specific contributions of the brain, and the environmental conditions necessary for different features of language to emerge.

## Acknowledgments

We gratefully acknowledge the Nicaraguan and American participants. We thank Deanna Gagne, Julia Fanghella, Diane Lillo-Martin, Letitia Naigles, Cris Ortega, and Russell Richie for assistance with data collection and processing, as well as helpful discussions. This research was supported by NIH grant P30 DC010751 to the second author and Diane Lillo-Martin.

## References

- Brentari, D., Coppola, M., Mazzoni, L., & Goldin-Meadow, S. (2012) When does a system become phonological? Handshape production in gesturers, signers, and homesigners. *Natural Language and Linguistic Theory*, 30, 1-31.
- Coppola, M. & Newport, E. L. (2005). Grammatical Subjects in home sign: Abstract linguistic structure in adult primary gesture systems without linguistic input. *Proceedings of the National Academy of Sciences*, 102, 19249-19253.
- Coppola, M. & Senghas, A. (2010). Deixis in an emerging language. In D. Brentari (Ed.). *Sign Languages: A Cambridge Language Survey*. Cambridge, UK: Cambridge University Press.
- Coppola, M., and W. C. So. (2005). Abstract and Object-Anchored Deixis: Pointing and spatial layout in adult homesign systems in Nicaragua. In A. Brugos, M. R. Clark-Cotton, and S. Ha (Eds.) *Proceedings of the Boston University Conference on Language Development*, 29. Boston: Cascadia Press.
- Goldin-Meadow, S. (2003). *The resilience of language*. New York: Psychology Press.
- Goldin-Meadow, S. & Mylander, C. (1984). Gestural communication in deaf children: the effects and noneffects of parental input on early language development. *Monographs of the Society for Research in Child Development*, 49, 1-151.
- Goldin-Meadow, S. & Mylander, C. (1990). Beyond the input given: The child's role in the acquisition of language. *Language*, 66, 323-355.
- Goldin-Meadow, S., Butcher, C., Mylander, C., Dodge, M. (1994). Nons and Verbs in a Self-Styled Gesture System: What's in a Name? *Cognitive Psychology*, 27, 259-319.
- Senghas, A. (1995). *Children's contribution to the birth of Nicaraguan Sign Language*. Doctoral dissertation, Brain and Cognitive Sciences, Massachusetts Institute of Technology.
- Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. *Psychological Science*, 12, 323-328.
- Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, 49, 370-407.



# Information Foraging in the Unknown Patches across the Life Span

Jessie Chin<sup>1</sup> (chin5@illinois.edu), Brennan Payne<sup>1</sup> (payne12@illinois.edu),  
Andrew Battles<sup>2</sup> (battles2@illinois.edu), Wai-Tat Fu<sup>3</sup> (wfu@illinois.edu),  
Daniel Morrow<sup>1</sup> (dgm@illinois.edu), Elizabeth A. L. Stine-Morrow<sup>1</sup> (eals@illinois.edu)

<sup>1</sup>Department of Educational Psychology, <sup>2</sup>Department of Electrical & Computer Engineering,

<sup>3</sup>Department of Computer Science

University of Illinois at Urbana Champaign

405 N Mathews Ave

Urbana, IL 61801 USA

## Abstract

This study used a word search puzzle paradigm to examine the effects of task environment and individual differences in ability on information foraging. Younger and older adults attempted to maximize the number of items found in a set of 4 puzzles in which they were at liberty to search within a puzzle or switch between them. Younger adults demonstrated faster uptake (i.e., number of words found as a function of time) from individual puzzles than older adults but experienced more deceleration of rates during the search. Additionally, older adults switched less often and their switching was less dependent on the uptake rate compared to younger adults. Both younger and older adults stayed longer than was optimal in a patch, older adults were especially likely to persevere suboptimally. Collectively, these results suggest that individuals may differentially optimize information gain through self-regulation of exploration and exploitation.

**Keywords:** Information foraging; information uptake; cognitive aging; adaptive behavior.

## Introduction

Self-regulation of cognition in natural environments almost always involves alternating phases of *exploration*, which entails search in the service of deciding how effort will be allocated, and *exploitation*, or task engagement in which effort is allocated to meet task-specific goals. Information Foraging (IF) models are predicated on an analogy between these regulatory processes and the way in which animals forage for food in the wild. Information foraging has been used to account for how people search for information in external environments, such as the WWW (e.g., Fu & Pirolli, 2007; Payne et al., 2007; Pirolli & Card, 1999) and in memory (Hills et al., 2010, 2012). However, even though IF presents a compelling metaphor, there is actually very little empirical research investigating the alignment between IF principles and how people interact with the environment to search and make use of information sources (Metcalf & Jacobs, 2010). There is also little work that has examined how individual differences afford or constrain search in and uptake from information sources. In this study, we used a simple word search puzzle to explore these issues.

According to the IF theory (Fu & Pirolli, 2007; Pirolli & Card, 1999), certain basic properties of animal foraging can be applied to the way human seek and consume information. First, food is distributed in the wild in clusters, or “patches,” that vary in their profitability (i.e., potential yield) and in

their tractability (i.e., how much of an investment of resources is needed for exploitation; e.g., apples on low branches or high branches). Resources in the patch are often finite and unknown to the foragers in advance, though “scent cues” may provide hints about profitability of the patch. Second, as patches become depleted, the rate of uptake decelerates. Third, the forager faces a tradeoff between gaining nutrients from exploiting a patch and consuming energy from exploring for food (e.g., to move among patches). The optimal foraging theory predicts that animals will stay in a patch until the expected rate of gain falls below the overall rate of gain, which takes into account the cost of moving to a new patch (Charnov, 1976; Stephens & Krebs, 1986). Finally, because food is crucial to survival, foragers work to maximize their food uptake and rarely revisit a depleted patch (Stephen, Brown & Ydenberg, 2007; Stephens & Krebs, 1986).

There are similarities and differences between animal foraging and human information foraging. For example, information is often clustered into patches (e.g., particular forms of print resources, webpages), though units of information are often hard to quantify in everyday life. Although information seekers may sometimes find it difficult to estimate profitability and tractability before visiting a patch, they may judge the richness or relevance of information based on their knowledge or expertise. Learners often selectively allocate their attention to materials as long as they perceive themselves to be learning, and disengage if they perceive their rate of learning to decrease below a threshold (e.g., Metcalfe, 2002; Metcalfe & Kornell, 2005). While information seekers have been found to adjust their search behavior to the statistical structures of the task environments (e.g., Fu & Pirolli, 2007), given the limited computational capacity and imperfect knowledge of human beings, the decision to explore a new task or exploit the current one is often suboptimal due to the biased representation of the local environment (e.g., Simon, 1956). For example, Payne, Duggan and Neth (2007) found, in a series of cognitive foraging experiments, that switch decisions could not be entirely predicted by the rate of gain from a patch. Rather, people tended to switch more than optimal without monitoring the real-time change of expected gain. Finally, empirical studies show that information seekers often revisit information patches (e.g. Payne et al., 2007). In fact, unlike food, information will not be exhausted after consumption. Therefore, the benefit of

“revisiting a patch” is particularly ecologically important in information foraging.

Little research has examined adult age differences in foraging behavior. Aging brings changes in both processing capacity and knowledge that would likely impact both uptake rates and exploratory behavior (Beier & Ackerman, 2005). In fact, older information seekers have been found to adopt different strategies to adapt to the environment. Mata, Wilke and Czienskowski (2009) showed that older adults were adaptive to the task characteristics in a fish foraging task, such as staying longer in one pond while between-ponds travel time was high. Interestingly, older adults have been found to search/explore less information but use simple heuristics or knowledge-driven strategies to achieve good performance in decision-making or ill-defined information search tasks (e.g., Chin, Fu & Kannampallil, 2009; Mata & Nunes, 2010). However, older adults’ information uptake behavior in a foraging task has generally received little attention. To investigate information foraging behavior in unknown environments, the goals of the current research were to examine: 1) the effects of task environments and individual differences on information uptake (measured as the rate of information gain), and 2) the effects of task environments and individual differences on the decisions to switch between sources.

Methods

The word search puzzle paradigm was modified from previous research (e.g., Chin, Fu & Stine-Morrow, 2011; Experiment 4 in Payne, Duggen & Neth, 2007). Participants were asked to maximize the number of items found in a set of 4 word search puzzles on an iPad. One puzzle was visible at a time and participants switched between puzzles at liberty, with a 10-minute limit (See Figure 1).

Participants

Sixty-one participants were recruited from the community. Four participants (3 young, 1 old) were

excluded due to technical problem or failure to comply with the instructions. Among remaining 57 participants, 28 young adults (Mean Age = 19.79, SD = 1.23; 19 female) and 29 old adults (Mean Age = 70.57, SD = 6.33, Range = 62-85; 20 female) were analyzed. All participants had graduated from high school. There was no age difference in the frequency of iPad use ( $t(56)=0.55$ ,  $p=0.59$ ). Young adults used computers more often than old adults ( $t(56)=2.83$ ,  $p<.01$ ), and old adults did word puzzles more often than young adults ( $t(56)=-2.63$ ,  $p<.05$ ). Older adults had better vocabulary than younger adults as measured by the Advanced Vocabulary Test (Ekstrom et al., 1976) ( $t(56)=-4.77$ ,  $p<.001$ ). On the other hand, younger adults had better working memory than older adults, as measured by Reading Span task (Stine & Hindman, 1994) ( $t(56)=2.87$ ,  $p<.01$ ).

Materials

The 4 puzzles, each containing 16 words from a different semantic category, were presented in three conditions: all easy, containing mostly high-prototypical category exemplars in canonical orientations in the puzzle (forward, down, left-right diagonal); all difficult, containing mostly low-prototypical exemplars in any orientations; and mixed (2 easy, 2 difficult). Measurement of exemplar prototypicality was based on category norms from Van Overschelde, Rawson, and Dunlosky (2004), in which prototypicality was indexed as the proportion of participants generating the word when given the category; there was significant difference in the mean prototypicality of words in the easy and hard puzzles ( $F(1,10)=20.82$ ,  $p<.001$ ). There were no differences in the mean log word frequency (Balota et al., 2007,  $F(1,10)=0.69$ ,  $p=.42$ ) or mean word length ( $F(1,10)=0.20$ ,  $p=.66$ ) between items in the easy and hard puzzles. Thus, given that the words in the easy puzzles were easier to generate from semantic memory and in a canonical orientation, they were more likely to “pop-out” than those in the hard puzzles. While controlling the density of the easy

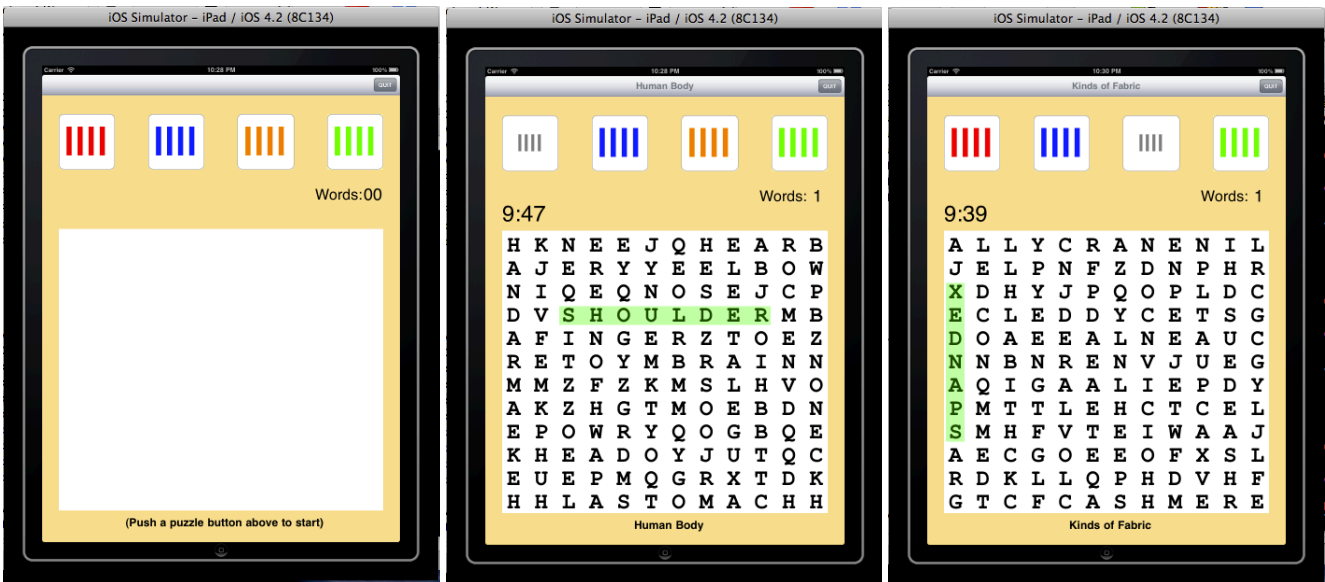


Figure 1. Layout of the word search puzzle experiments

and hard puzzles, we manipulated the profitability of the puzzles to see if participants were effective in monitoring their uptake rates.

The interface for word search puzzle was programmed in iPad (see Figure 1). Participants first saw the interface with four colored buttons. Each button referred to a puzzle of different semantic category. Participants could press any of the four buttons to start the experiment. When the participant pressed any of the four buttons, a countdown timer of 10 minutes started. A word search puzzle appeared with its category name shown on the top and bottom of the interface. Participants saw one puzzle at a time, and used their fingers to swipe the words they found. The found words were highlighted in different colors and remained highlighted during the whole session. Participants could check the number of words they found in each puzzle on the right corner, but would not know the number of words remaining in each puzzle. During the experiments, participants could press the button to switch to the other puzzles. In the mixed condition, the order of buttons of easy and hard puzzles was in counterbalanced order. Every meaningful touch (such as button touch, letter touch) on the iPad was recorded with time stamps.

### Experimental Design

The experiment followed a 2 x 3 mixed factor design with between-subject variable, age (young vs. old) and within-subject variable, task condition (all easy vs. mixed vs. all hard). The order of the three conditions was counterbalanced across participants.

### Procedures

At the beginning of the experiment, participants completed cognitive measures after the consent process. Participants then practiced locating words in the puzzles and switching among puzzles for 20 minutes. After the practice, participants performed the experimental task. Each condition took 10 minutes. Participants had been told explicitly that some puzzles might be easier than others, and they could go back and forth among four puzzles and decide how long they want to spend in each puzzle on their own. After all three conditions, the experimenter briefly interviewed the participants about their self-observed search and switch strategies. Participants were debriefed at the end.

## Results

A 2 x 3 Repeated Measures Analysis of Variance showed significant main effects of age and condition on the number of words found in each condition (Age:  $F(1,55)=35.37$ ,  $p<.001$ ; Condition:  $F(2,55)=191.78$ ,  $p<.001$ ). Both younger and older adults found the most words in the Easy condition, then the Mixed condition, followed by the Hard condition. Younger adults found more words than older adults across all the conditions. The Age by Condition interaction was not significant (see Table 1). However, younger and older adults varied in the extent to which they found words on their first encounter (Bout 1) versus successive encounters (Bout>1) with the puzzles (Figure 2). Older adults tended to find most words in their first bout at

the puzzle, while younger adults tended to find relatively more words in later bouts (i.e., more revisiting), especially in the hard puzzle (Age:  $F(1,54)=11.00$ ,  $p<.005$ ).

Table 1. Descriptive statistics of word search performance

Mean (SD)	Easy	Mixed	Hard
Young	38.93(6.35)	31.86(5.63)	23.39(7.40)
Old	29.24(6.95)	23.34(5.47)	15.72(6.15)

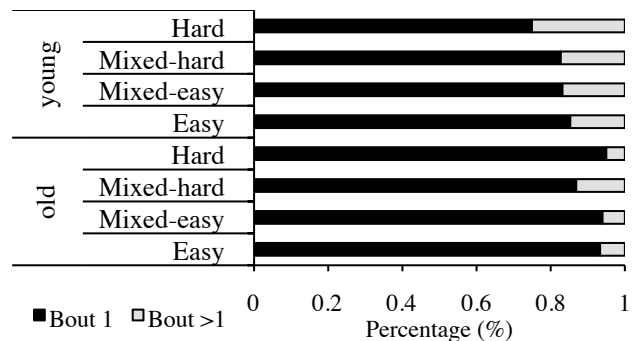


Figure 2. Age difference in the percent of words found in the first attempt

### Age Differences in Uptake Rates

Mixed-effects modeling was conducted to estimate uptake rates in the different conditions. Uptake rate was defined as the cumulative number of words found as a function of time with data modeled based on 2-sec intervals. As showed in Figure 2, participants found most words in their first bouts across different conditions; thus, we modeled the uptake rates for the first bout only. There were 37,763 observations in total. Following the growth curve analysis method (Mirman, Dixon & Magnuson, 2008), we started with the

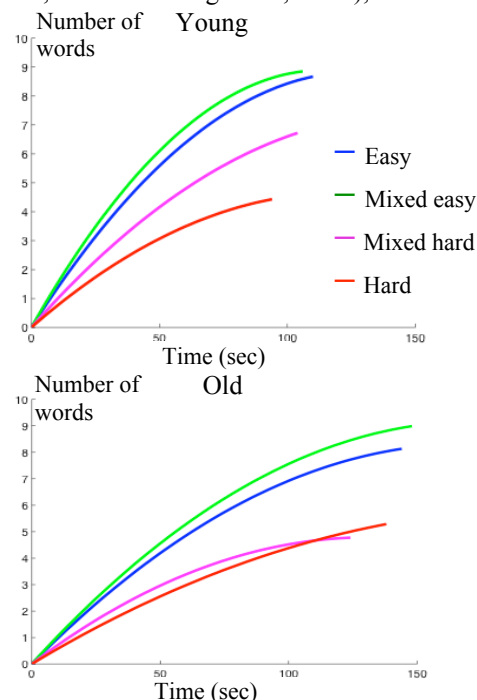


Figure 3. The uptake rates for younger and older adults in different puzzles

“average uptake rate model.” The uptake rate function (cumulative number of words per unit time) was calculated as :

$$Y_{ij} = \gamma_1(\text{time}) + \gamma_2(\text{time}^2) + U_{1j}(\text{time}) + U_{2j}(\text{time}^2) + e_{ij} \quad (1)$$

In (1),  $Y_{ij}$ ,  $\gamma$ ,  $U$ , and  $e_{ij}$  represented the cumulative number of words, the fixed effects, the random effects of subjects, and the error term respectively. Because we are interested in capturing both the linear and non-linear components of the random effects of subjects, it was divided into:  $U_{1j}$  – the linear “rate” and  $U_{2j}$  – the non-linear “rate of change”. Then we added fixed effects of age and condition and its interaction terms to the model “conditional uptake rate model”, as follow:

$$Y_{ij} = \gamma_1(\text{Time}) + \gamma_2(\text{Time}^2) + \gamma_3(\text{Age} \times \text{Time}) + \gamma_4(\text{Age} \times \text{Time}^2) + \gamma_5(\text{Condition} \times \text{Time}) + \gamma_6(\text{Condition} \times \text{Time}^2) + \gamma_7(\text{Age} \times \text{Condition} \times \text{Time}) + \gamma_8(\text{Age} \times \text{Condition} \times \text{Time}^2) + U_{1j}(\text{time}) + U_{2j}(\text{time}^2) + e_{ij} \quad (2)$$

The condition update rate model in (2) was developed to test how uptake rates changed (both linearly and non-linearly) with conditions and age. The model shows that the uptake rate (which measured how quickly subjects found a word in a puzzle) for the easy puzzles was higher than for the hard ones ( $F=2377.28$ ,  $p<.001$ ). Interestingly, the uptake rate for the hard puzzles was higher when they were embedded in the mixed condition with easier puzzles relative to those in the pure condition. This was true for both younger and older adults, suggesting a facilitation effect in the mixed condition, in which there were 2 easy and 2 hard puzzles. Figure 3 showed best fitting curves of uptake rates of younger and older adults in four puzzles to the empirical data. The length of curves represents the mean duration of uptakes (exploitation). As shown in these plots, older adults stayed longer in the puzzle than younger adults.

Younger adults had higher uptake rates than older adults, especially in the easy puzzles (Age x condition x time:  $F=108.32$ ,  $p<.001$ ). Younger adults also showed a larger rate of change, such as quicker deceleration of uptake rate across time, than older adults ( $F=16.30$ ,  $p<.001$ ). The difference in rates of change was larger in the easy, mixed puzzles than the hard ones ( $F=93$ ,  $p<.001$ ). Thus, the uptake rates grew more quickly for younger adults but reached the asymptote quicker (with larger reduction of rates across time) than older adults.

### Age Differences in Switch

Given the individual difference in uptake rates across different conditions, we examined whether age differences in uptake rates were related to frequency of switching. A 2 x 3 Repeated Measures Analysis of Variance (Age x Condition) was conducted on the number of switches in the easy, mixed, and hard condition. Younger adults switched more often than older adults in all conditions (Figure 4)

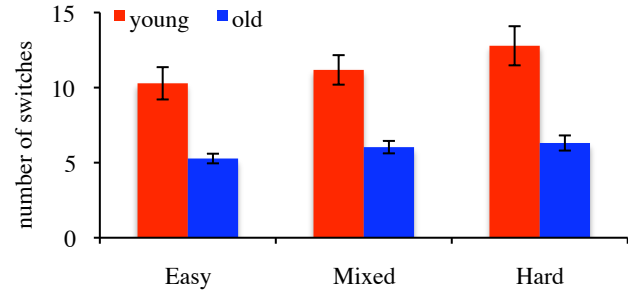


Figure 4. Age differences in number of switches of different conditions

( $F(1,55)=30.39$ ,  $p<.001$ ). There was also a main effect of condition showing that people switched more in the hard condition, then the mixed condition, followed by the easy condition ( $F(2, 55)=5.21$ ,  $p<.01$ ). The Age x Condition interaction was not significant.

Given that younger and older adults experienced different degrees of rates of change, we examined if the age differences in rates of change over time were associated with their switch behavior, and the extent to which they were moderated by individual differences in working memory and verbal ability. We first extracted the best linear unbiased predictors of rates of change from the average uptake rate model ( $U_{2j}$ ). Then we did a median split on the estimates of rates of change to create two groups – those with uptake rates dropping more and those with uptake rates dropping less. We did a 2 (Age) x 2 (dropping more or less) ANCOVA to examine the relationship between the number of switches and the deceleration of uptake rates across time by treating individual differences in working memory and verbal ability as covariates.

Results showed a significant Age x Rate of change interaction ( $F(1,51)=7.21$ ,  $p<.01$ ) in addition to the effects of age, rate of change, and working memory (Age:  $F(1,51)=19.29$ ,  $p<.001$ ; Rate of change:  $F(1,51)=9.87$ ,  $p<.01$ ; Working memory:  $F(1,51)=6.28$ ,  $p<.05$ ). The interaction of Age x Rate of change on the number of switches was shown in Figure 5. People with more reduction of uptake rates (the dropping more group) across time tended to switch more, and the difference was bigger in younger adults than older adults. In the other words, older adults were less sensitive to their rates of change in uptake—they were less likely than younger adults to switch puzzles as the rate of uptake diminished. On the other hand, younger adults were more sensitive to changes in uptake rates, which led to more switches. Also, the covariates of working memory had association with switch behavior – people with higher working memory capacity tended to switch more often. However, the age-differences in associations between rates of change and switch behavior were shown regardless of the individual differences in working memory.

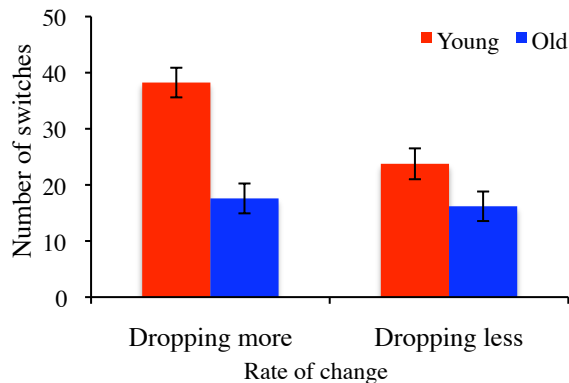


Figure 5. Interaction of age and rate of change on switch

### Suboptimal Leaving and Longer Perseverance for Older Adults

Given that younger adults switched more than older adults, and also showed higher uptake rates and rates of change, the next question we addressed was whether younger and older participants switched optimally. According to the optimal foraging theory, the marginal value theorem predicts the optimal patch departure time – the time at which the marginal uptake rate is equal to the mean uptake rate of the entire habitat (Charnov, 1976). We calculated the ratio of marginal uptake rate at each word and the mean uptake rate of the corresponding patch for each participant. The optimal time to switch to a different puzzle is when the ratio equals 1. When the ratio is larger than 1, it is advantageous to stay because the current marginal rate is higher than the average expected return (estimated based on previous experiences). As the marginal value decreases with decelerated uptake rates, the value becomes increasingly smaller than 1, and it is advantageous to switch because the expected uptake from the habitat as a whole exceeds the current marginal value.

Both younger and older adults were suboptimal based on the criterion derived from the marginal value theorem, as they left the puzzle late (Figure 6a). Mean ratio of the last word in the puzzle was smaller than 1, suggesting that the marginal uptake rate of the last word was slower than the mean uptake rate in the corresponding puzzle. Though people tended to leave the puzzle when the uptake rate was low, Figure 6a shows that participants would have been more optimal if they left the puzzle about 2 words earlier (the ratio of the third word back was close to 1). Additionally, among puzzles of different profitability, people tended to switch more optimally in the hard puzzles than in the easier puzzles. This finding suggests that both younger and older adults were more sensitive to the change of uptake rates and switched earlier in the hard puzzle condition (Figure 6b).

While both younger and older adults switched later than was optimal, they persevered differently in the puzzles. Perseverance was measured by the give up time, which was defined as the duration from finding the last word to leaving

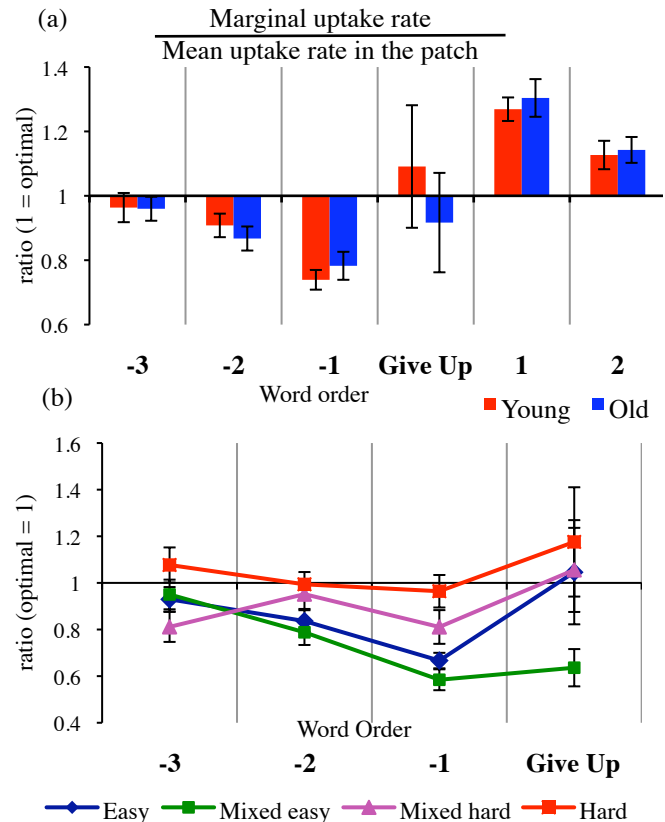


Figure 6. The ratio of marginal uptake rate of the word and mean uptake rate in the corresponding patch for younger and older adults (a) and different puzzles (b)

the puzzle (e.g., Payne et al, 2007) – i.e., the amount of time participants persevere in a patch without finding a word.

A 2 x 4 Repeated Measures Analysis of Variance (Age x Puzzle type: easy, mixed easy, mixed hard, hard) was conducted to explore the effects of age and puzzle type on give up time. Give up time was longer for older adults than for younger adults. In other words, older adults persevered longer in the current patch before moving to a new patch compared to younger adults. Figure 6a also showed that while younger adults tended to persevere for a shorter time than their mean uptake time for the puzzle, older adults tended to persevere longer than their mean uptake time. Furthermore, people persevered longer in the hard puzzles than the easy ones (puzzle type:  $F(1,53)=2.55$ ,  $p<.05$ ; age:  $F(1,53)=13.15$ ,  $p<.001$ ). Interestingly, the give up time in the mixed easy puzzles was relatively longer than the time in the all easy condition, suggesting that participants were influenced by the mixed context.

### Conclusion

The study used the word search puzzle paradigm to study the information search behavior of younger and older adults in the patches of different profitability. Although the gain functions of puzzles were unknown to the participants, individuals were able to allocate their effort to uptake and switch when uptake decreased. Older adults showed slower



uptake rates and smaller change of rates than younger adults across different puzzles. Thus, older adults relied less on the deceleration of uptake rates to decide when to switch to a different puzzle. Older adults switched less often and persevered longer in the puzzles, especially in the difficult condition. To maximize the search performance, older adults allocated more time to exploitation (i.e., task engagement in the puzzles) and younger adults did more exploration to the new puzzles than the older adults. Overall, older and younger adults showed adaptive self-regulation patterns through differential attention to exploitation and exploration.

Older adults were found to be less explorative in information search in decision making (Mata & Nunes, 2010) and web information search (Chin, Fu & Kannampallil, 2009), and they explored (i.e., switched to another puzzle) less often in the current study as well. Less exploration might be adaptive given the heavy demands on processing capacities of switching behavior in information search (e.g., Chin et al., 2009, 2011). In addition to the higher switch cost of older adults, results suggested that older adults seemed to use different policies (i.e., less relying on the rates of change) to make switch decision than younger adults. As the optimal foraging model suggests, foragers will leave while the marginal uptake rates is lower than the mean uptake function of a patch. However, given the uptake function of the puzzles were unknown to the participants, people needed to track their uptake behavior after entering a puzzle across time to estimate the expected gain of the puzzle. This process was so information intensive and resource demanding that older adults might experience more difficulty executing which was partly shown in our results. Thus, age differences in learning from experiences in a given information patch and the corresponding patch-leaving policy should be further examined in future studies.

Despite the age differences in switch, both younger and older adults were suboptimal in terms of the later departure time in the patches. Interestingly, past studies also found that foragers were suboptimal in external search task (e.g., Mata et al., 2009), but closer to optimal in memory search (Hills et al., 2012). Hills and his colleagues used cross-modal priming to show that external search patterns can be transferred to internal search patterns, suggesting that there is a central executive control process monitoring both internal and external search behavior. Therefore, the difference of patch-departure behavior in internal and external search task might be due to the fact that foragers have more knowledge about the gain function of a patch in the internal search task than the external search task. Similarly, in the condition of mixed uptake functions, results showed that people were farther away from the optimal (i.e., late departure) in the mixed easy, mixed hard puzzles than the easy and hard puzzles respectively suggesting that the knowledge of a patch might be important to determine the optimal departure in the task.

## References

- Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445-459.
- Beier, M.E. & Ackerman, P.L. (2005). Age, ability and the role of prior knowledge on the acquisition of new domain knowledge. *Psychology & Aging*, 20, 341-355.
- Charnov, E. L., (1976). Optimal foraging, the Marginal Value Theorem. *Theoretical Population Biology*, 9, 129-136.
- Chin, J., Fu, W-T. & Kannampallil, T. (2009). Adaptive information search: Age-dependent interactions between cognitive profiles and strategies. In *Proceedings of the 27th ACM Conference on Human Factors in Computing Systems CHI'09* (pp. 1683-1692). ACM Press.
- Chin, J., Fu, W-T. & Stine-Morrow, E.A.L. (2011). To go or to stay: Age differences in cognitive foraging. *Poster of the 33<sup>rd</sup> CogSci Conference*, Boston, MA.
- Ekstrom, R. B., French, J. W., & Harmon, H. H. (1976). *Manual for the Kit of Factor-Referenced Cognitive Tests*. Princeton, NJ: Educational Testing Service.
- Fu, W.-T., & Pirolli, P. (2007). SNIF-ACT: A cognitive model of user navigation on the World Wide Web. *Human-Computer Interaction*, 22, 355-412.
- Hills, T., Todd, P. M., & Goldstone, R. L. (2010). The central executive as a search process: Priming exploration and exploitation across domains. *Journal of Experimental Psychology: General*, 139, 590-609.
- Hills, T., Jones, M., & Todd, P.M. (2012). Optimal foraging in semantic memory. *Psychological review*, 119, 431-440.
- Mata, R., & Nunes, L. (2010). When less is enough: Cognitive aging, information search and decision quality in consumer choice. *Psychology and Aging*, 25, 289-298.
- Mata, R., Wilke, A., & Czienskowski, U. (2009). Cognitive aging and adaptive foraging behavior. *Journal of Gerontology: Psychological Sciences*, 64B, 474-481.
- Metcalf, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General*, 131, 349-363.
- Metcalf, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language*, 52, 463-477.
- Mirman, D., Dixon, J.A., & Magnuson, J.S. (2008). Statistical and computational models of visual word paradigm: Growth curves and individual differences. *Journal of Memory and Language*, 59, 475-494.
- Payne, S. J., Duggan, G. B., & Neth, H. (2007). Discretionary task interleaving: Heuristics for time allocation in cognitive foraging. *Journal of Experimental Psychology: General*, 136, 370-388.
- Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, 106, 643-675.
- Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychology review*, 63, 129-138.
- Stine, E. A. L., & Hindman, J. (1994). Age differences in reading time allocation for propositionally dense sentences. *Aging and Cognition*, 1, 2-16.
- Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.
- Stephens, D.W., Brown, J.S., and Ydenberg, R.C. (2007). *Foraging: Behavior and Ecology*. Chicago: University of Chicago Press.
- Van Overschelde, J. P., Rawson, K. A. & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Batting and Montage (1969) norms. *Journal of Memory and Language*, 50, 289-335.

# Connectivity Asymmetry Can Explain Visual Hemispheric Asymmetries in Local/Global, Face, and Spatial Frequency Processing

**Ben Cipollini (bcipolli@cogsci.ucsd.edu)**

Department of Cognitive Science, University of California San Diego  
9500 Gilman Dr 0515, La Jolla, CA 92093 USA

**Janet Hsiao (jhsiao@hku.hk)**

Department of Psychology, University of Hong Kong  
604 Knowles Building, Pokfulam Road, Hong Kong SAR

**Garrison Cottrell (gary@eng.ucsd.edu)**

Department of Computer Science and Engineering, University of California San Diego  
9500 Gilman Dr 0404, La Jolla, CA 92093 USA

## Abstract

Left-right asymmetries have been noted in tasks requiring the classification of many visual stimuli, including Navon figures, spatial frequency gratings, and faces. The Double Filtering by Frequency (DFF) model (Ivry & Robertson, 1998), which postulates asymmetric frequency filtering on task-relevant frequency bands, has been implemented to account for asymmetric processing of each stimulus type above, but does not provide a fully mechanistic explanation, nor does it have direct neural correlates. The Differential Encoding (DE) model (Hsiao, Shahbazi, & Cottrell, 2008), which postulates that a known asymmetry in patch connectivity drives visual processing asymmetries, has previously been used to account for only one stimulus type. Here, we refine the DE model to match the published patch asymmetry more precisely and show that the DE model generalizes to three of the four datasets mentioned above. Examination of the failure to match all datasets suggest a possible reinterpretation of the original dataset itself.

**Keywords:** local/global processing, left-side bias, hemispheric asymmetry, visual perception, Differential Encoding, Double Filtering by Frequency, computational model

## Introduction

A large literature of experimental psychology and cognitive imaging studies has established the existence of a wide range of left-right asymmetries in the classification of many visual stimuli. A typical paradigm consists of briefly presenting a stimulus to the left or right of fixation, then requiring subjects to perform a classification task, such as whether a target stimulus was present or not. Because information from the left visual field (LVF) is initially directed only to the right cerebral hemisphere (RH), and right visual field (RVF) to the left cerebral hemisphere (LH), comparisons of task performance between LVF/RH and RVF/LH can indicate asymmetries in hemispheric processing. Visual stimuli which have shown such asymmetries in these types of tasks include Navon figures (Sergeant, 1982) consisting of a large, “global”-level figure composed of smaller, “local”-level figures (see Figure 3b), spatial frequency gratings (Christman, Kitterle, & Hellige, 1991; Kitterle, Hellige, & Christman, 1992), and faces (Young & Bion, 1981; Brady, Campbell, & Flaherty, 2005).

Ivry and Robertson (1998) developed the Double Filtering by Frequency (DFF) theory to account for these asymmetries.

The computational model they implemented aimed to account for three particular experiments from the literature, thought to express core features of the data:

- Sergeant (1982): showed the basic hemisphere  $\times$  level interaction of the local/global literature, with responses to targets presented at the smaller, “local” level showing faster reaction times when presented to the RVF/LH, and responses to targets presented at the larger, “global” level showing faster reaction times when presented to the LVF/RH.
- Kitterle et al. (1992): showed that reaction times in two different classification tasks, using the same stimuli but requiring use of information at different spatial frequencies, interacted with the visual field/hemisphere of presentation. Responses to the task requiring high spatial frequency (HSF) information showed faster reaction times when presented to the RVF/LH, and responses to the task requiring low spatial frequency (LSF) information showed faster reaction times when presented to the LVF/RH.
- Christman et al. (1991): showed that discrimination between two stimuli which differed by a single spatial frequency interacted with the visual field/hemisphere of presentation, based on the *relative* frequency of the discriminative spatial frequency compared to the rest of the spatial frequencies contained in the stimuli. When the discriminative frequency was higher than the frequency content of the rest of the stimulus, responses were faster for presentation to the RVF/LH; when the discriminative frequency was lower than the frequency content of the rest of the stimulus, responses were faster for presentation to the LVF/RH.

The DFF model replicated the core features of each of the above studies. Later, Hsiao, Shieh, and Cottrell (2008) showed that the DFF theory could also account for the so-called ‘left-side bias’—the tendency for people to associate face identity with the right-side of a person’s face (appearing in the LVF of the viewer) (Brady et al., 2005).

The DFF model, while it accounts for all these data, requires a homunculus to input the frequency range of interest for the particular task being modeled, rather than discovering



the task-relevant frequency range through training. In addition, no neurophysiological evidence has been found for spatial frequency filtering in cortex.

Hsiao, Shahbazi, and Cottrell (2008) took a very different approach to the problem of explaining visual processing asymmetries. Rather than starting with a theory of the algorithms behind the asymmetries (Marr’s “algorithmic” level), they created a model of an anatomical asymmetry (Marr’s “implementation” level) and asked whether it could account for the asymmetries in classification of visual stimuli observed in behavioral studies. They used an asymmetry in inter-patch connectivity (Galuske, Schlote, Bratzke, & Singer, 2000), one of the few known network-level asymmetries, and the only one suggested to be related to local/global processing asymmetry (Galuske et al., 2000; Hutsler & Galuske, 2003). This asymmetry was found in BA22, an auditory association area, and not in primary auditory cortex. This matches current theory in the local/global literature, where it has been suggested that local/global processing differences occur beyond early primary sensory areas (Sergent, 1982; Ivry & Robertson, 1998), motivating the use of this asymmetry to model tasks involving local/global processing.

Hsiao et al. used this connectivity pattern as inspiration to implement a simple “autoencoder” neural network with differential connectivity (see Figure 2) to encode the stimuli, and a simple perceptron to perform the task using the autoencoders’ learned representations (see Methods section for details on this model). They tested it to see whether this anatomical asymmetry could account for a subset of the data modeled by Ivry and Robertson. Using precisely the same small, 1-dimensional inputs that Ivry and Robertson created for modeling a reduced version of Sergent’s study, Hsiao et al.’s “Differential Encoding” (DE) model showed a hemisphere  $\times$  level interaction that matched Sergent’s human data more closely than that of Ivry and Robertson’s DFF model. Hsiao et al. then constructed realistic 2D bitmaps of Sergent’s Navon stimuli, trained 2D versions of the DE models on these realistic stimuli, and again showed a hemisphere  $\times$  level interaction that was a better quantitative match to the original human data than the results published by Ivry and Robertson using their DFF model.

The DE model was able to address issues with the DFF model by implementing an anatomical asymmetry. The use of a secondary network for classification allows the learning algorithm to select task-relevant information, rather than manually selecting frequency bands as does the DFF model.

However, the DE model did not address most of the data accounted for by the DFF, including two local/global studies (Kitterle et al., 1992; Christman et al., 1991), one face-processing study (Young & Bion, 1981), and the relationship between local/global processing and spatial frequency processing. In addition, the DE model used parameters for number of connections and distribution shape that were very different from the parameters reported in the literature.

Here we improve the model’s fidelity to neural data. We

show that this model accounts for three of the four studies described above. We also show that this model filters frequencies in a manner consistent with the human literature.

## Methods

The “Differential Encoding” (DE) model is based on an asymmetry in “patch” connectivity found in BA 22 (Galuske et al., 2000). “Patches” are found in many cortical areas across species (monkeys, humans, cats, rodents) and across sensory and association areas. Patches are thought to be a level of organization akin to a macro-column, consisting of thousands of selectively interconnected neurons within a cortical area. These patches selectively interconnect to a small subset of other local patches through horizontal connections through the grey matter. These patches are named because an injection of dye into the cortical surface will label cortex at the injection site, as well as rather discrete “patches” of surrounding cortex (see Figure 1) (Amir, Harel, & Malach, 1993; Levitt & Lund, 2002).

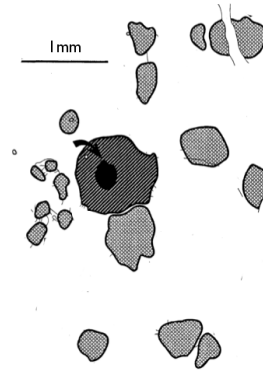


Figure 1: Drawing of “patches” in V4. Dark arrow indicates site of dye injection. Reproduced (without permission) from Amir et al. (1993).

The function of these inter-patch connections is not known. Briefly, we propose here that horizontal connections lead to interconnected patches, biasing each other to process features shared across the interconnected group. We therefore implement a feed-forward model, where the hidden units discover the correlated features shared across interconnected input “patches” across a set of input stimuli.

The “Differential Encoding” (DE) model includes two autoencoder neural networks with differences in connectivity, one for each hemisphere. Unlike most autoencoders, the hidden units of these models connect to a small subset of the input and output banks (see Figure 2). Each hidden unit has a position in the input (and output) arrays, and a fixed number of connections to the input (and output) arrays are sampled from a Gaussian distribution centered at that hidden unit’s position in the input (and output). The LH and RH autoencoders have the same number of hidden units and sample the same number of connections to the input (and output) for each hidden unit. The only difference between the networks, then,

is the width ( $\sigma$ ) of the Gaussian distribution. In accordance with the findings of Galuske et al. (2000), the left hemisphere network had a wider distribution than the right hemisphere network (see Figure 2).

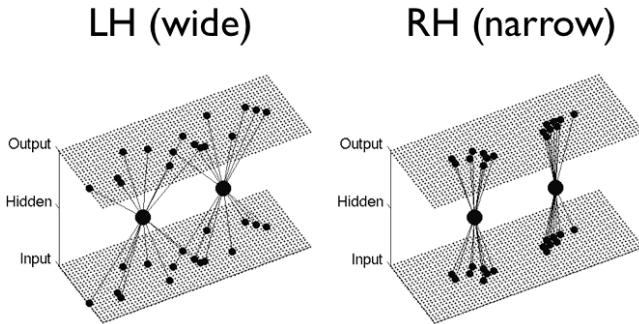


Figure 2: Representation of two hidden units for LH (left) and RH (right) autoencoder networks, along with their connections. The connections are randomly sampled from a Gaussian distribution centered on each hidden unit's position in the input array. The Gaussian distribution used for the LH is wider than that used for the RH. Not pictured are the classification networks which operate on the hidden unit encodings extracted from the autoencoder networks after training.

The number of hidden units was varied from extremely small (13) to extremely large (800) values. Results did not differ qualitatively when the number of connections per hidden unit varied to allow for the same number of overall weights to be used to learn the images. If too few weights were used, the networks could not learn the training set well enough for a meaningful analysis. After these initial explorations, the final simulations were run with the number of connections per hidden unit fixed to be within the range reported by Amir et al. (1993) in visual cortex (between 8 and 15), and the number of hidden units were chosen to allow equal spacing across the input image with enough total parameters to learn the images.

Each LH and RH network is constructed by randomly sampling connections for each hidden node. Gaussian distributions were used such that inter-patch distance values were similar to those calculated from data reported in (Galuske et al., 2000). On average, patches had 1.75 times the width of a single patch between them in the RH, and 2.05 in the LH. Here, we considered the size of a patch to be a pixel, and chose sigmas such that when inter-patch distance was measured after randomly sampling connections, on average there were 1.75 pixels between connections in the RH, and 2.05 in the LH.

Greyscale images are constructed for each task stimulus. The autoencoders are trained via of backpropagation of error (Rumelhart, Hinton, & Williams, 1986) to reproduce these greyscale images from the input to the output. Once the autoencoders reach a pre-determined performance level, training stops. Each stimulus image is then presented to the

trained autoencoder network, and the activation of the hidden units is recorded. These encodings, which differ only due to the differences in connectivity structure between LH and RH networks, are then used as inputs to a separate feed-forward neural network, which is trained to classify these encodings according to the behavioral task for the experiment.

For a single experiment, multiple “instances” of each LH and RH network are constructed and trained; each “instance” differs only in the random samples of its connections. The number of instances is determined by matching the total number of trials used both in the modeling experiment and in the corresponding behavioral experiment, such that the statistical power of each experiments are equated. Performance is evaluated on each model individually, then performance over all “instances” of a hemisphere are averaged. Average model error for each model hemisphere is compared to average reaction time of the corresponding visual field in the human experiment, with both measures conceived as measures of difficulty or uncertainty in processing.

In order to examine how the different connectivity structures affect spatial frequency encoding, each stimulus image is presented to a trained autoencoder. Each output image produced is examined for spatial frequency content, and a 2D power spectrum across all images in the stimulus set is constructed. Each 2D power spectrum is translated to a 1D power spectrum by measuring the linear distance from the pixel carrying the  $f_0$  (intensity) power to each pixel in the 2D power spectrum, then sorting all linear distances and averaging over any pixels with the same linear distance. Each 1D power spectrum is then compared to the power spectrum of the original image. The difference in 1D power spectrums is then compared across hemispheres, showing for each frequency which hemisphere has encoded information closer to the original image than the other.

## Experiments and Results

### Sergent (1982) simulations

16 binary images (31x13 pixels) of Navon stimuli (letters [H, T, F, L] each appeared at local and global levels in all possible combinations) (see Figure 3b for example stimuli) were presented to 68 LH ( $\sigma = 6.0$ ) and RH ( $\sigma = 3.0$ ) autoencoder models to match the total number of trials in Sergent's human data. Each autoencoder network had 360 hidden units, with each hidden unit connecting to 8 input and output units. Each autoencoder network was trained to 0.005 average error per output unit, then hidden unit encodings were extracted. A perceptron classifier with 360 input units and one output unit was trained to classify each of the 16 Navon stimuli as containing a target or not.

As in Hsiao, Shahbazi, and Cottrell (2008), the network showed a significant hemisphere  $\times$  level interaction (Two-factor, within-subject repeated measures ANOVA;  $F(1,67)=8.62$ ;  $p < 0.01$ ). We ran this same modeling experiment and analysis for all 6 combinations of target and distracter letters, to see if results would generalize. This only

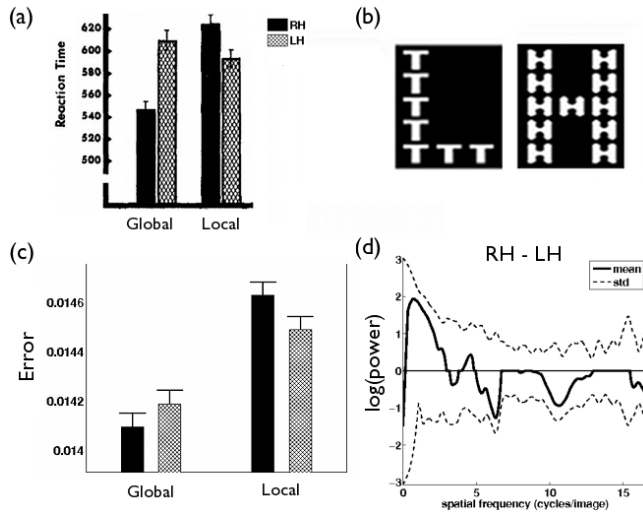


Figure 3: Original and model results for Sergent task

- (a) Original hemisphere  $\times$  level interaction; reproduced (without permission) from Sergent (1982)
- (b) Sample Navon stimulus used in our modeling experiment
- (c) DE model hemisphere  $\times$  level interaction
- (d) DE model spatial frequency analysis of output images, showing a RH advantage (above zero on Y-axis) for LSF (towards left side of X-axis) and a LH advantage (below zero on Y-axis) for HSF (towards right side of X-axis)

required training new classification neural networks, as the stimuli in each experiment remained the same. Each of these experiments showed a statistically significant hemisphere  $\times$  level interaction.

Comparing the 1D power spectrums created from the output images of the autoencoder neural networks, we saw a clear tendency for the RH network to be closer to the power spectrum of the original image for LSFs, and the LH network to be closer to the power spectrum of the original image for HSF (see Figure 3d). These trends matched the large literature reporting better performance on LSF for LVF/RH and better performance on HSF for RVF/LH.

### Kitterle et al (1992) simulations

40 greyscale images (31x13 pixels), each consisting of a low or high frequency grating, the grating either being a sine or square wave, and shown at one of 10 phases (see Figure 4b for example stimuli), were presented to 40 LH ( $\sigma = 6.0$ ) and RH ( $\sigma = 3.0$ ) autoencoder models. Each autoencoder network had 360 hidden units, with each hidden unit connecting to 8 input and output units. Each autoencoder network was trained to 0.005 average error per output unit, then hidden unit encodings were extracted. Two identical neural networks with 360 input units, ten hidden units, and one output unit were trained to classify each of the 40 stimuli. One classification network was trained to discriminate between wave type (sine vs square) and to ignore the frequency of the waves; this task required use of HSF information. The other classification net-

work was trained to discriminate between the two frequencies of the waves and to ignore the wave type; this task required use of LSF information. Both networks were trained with the same training parameters.

The first classification network showed a significant effect of hemisphere ( $F(1,39)=4.06$ ;  $p < 0.05$ ), with the LH network showing better performance. The second classification network showed a significant effect of hemisphere ( $F(1,39)=4.53$ ;  $p < 0.04$ ), with the RH network showing better performance. Across the two tasks, the networks showed a significant hemisphere  $\times$  task interaction (Two-factor, within-subject repeated measures ANOVA;  $F(1,39)=9.89$ ;  $p < 0.002$ ). Note that the main effect of task-type was not preserved; our models found discriminating the wave type to be an easier task than discriminating between two frequencies; Kitterle et al. (1992) reported the opposite result for their human participants.

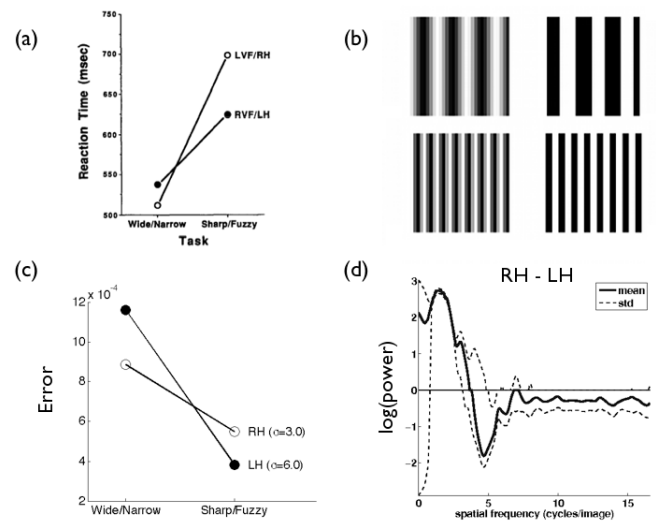


Figure 4: Original and model results for Kitterle task

- (a) Original hemisphere  $\times$  task interaction; reproduced (without permission) from Kitterle et al. (1992)
- (b) Sample stimuli used in the modeling study
- (c) DE model hemisphere  $\times$  task interaction
- (d) DE model spatial frequency analysis of output images, showing a RH advantage (above zero on Y-axis) for LSF (towards left side of X-axis) and a LH advantage (below zero on Y-axis) for HSF (towards right side of X-axis)

### Face Processing simulations

**Young and Bion (1981): Face Recognition** Young and Bion (1981) (and other) studies have shown a RH advantage for face recognition. We set out to replicate this general finding.

The same face stimuli used in Hsiao, Shieh, and Cottrell (2008), which used the DFF model to show a left-side bias, were used to construct greyscale images in this study. The dataset contained 30 individuals with 8 expressions each; 4

expressions used in training, and 4 different expressions were used in the data collection/testing phase. These face stimuli were more complex, and so required more parameters to train to a lesser error criterion. The face stimuli were presented to 40 LH ( $\sigma = 8.0$ ) and RH ( $\sigma = 3.0$ ) autoencoder models. Each autoencoder network had 360 hidden units, with each hidden unit connecting to 12 input and output units; ; the wider LH gaussian was selected to space out the greater number of connections per hidden unit. Each autoencoder network was trained to 0.01 average error per output unit, then hidden unit encodings were extracted. A neural network with 360 input units, 25 hidden units, and 30 output unit was used to classify each of the 120 test images as one of the 30 individuals.

A significant effect of hemisphere on face identification accuracy was found (ANOVA;  $F(1,39) = 10.33$ ,  $p < 0.002$ ). These effects were consistent across the training and test sets.

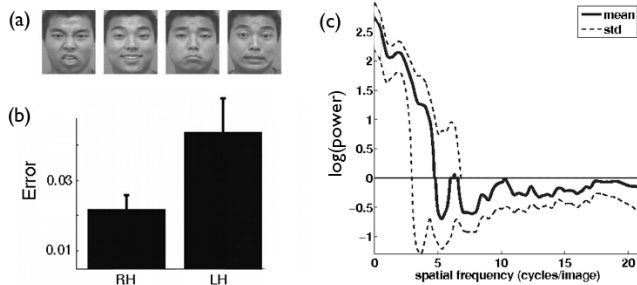


Figure 5: Stimuli and power spectrum for left-side bias task  
(a) Sample stimuli for one individual across four expressions (from the CAFE dataset)  
(b) DE model classification of individual face recognition identity (Young & Bion task); error-bars represent standard error of the mean  
(c) DE model spatial frequency analysis of output images, showing a RH advantage (above zero on Y-axis) for LSF (towards left side of X-axis) and a LH advantage (below zero on Y-axis) for HSF (towards right side of X-axis)

**Brady et al. (2005): Left-Side Bias** The same face stimuli used in Hsiao, Shieh, and Cottrell (2008), which used the DFF model to show a left-side bias, were used to construct greyscale face images. 240 greyscale face images (34x25 pixels; 30 individuals; 4 expressions used in training, 4 different expressions used in testing) were used to create left and right chimeric faces: faces with one side duplicated across the mid-line to the other. The same network parameters were used as the face recognition simulation above for training.

For each set of chimeric faces, a significant effect for hemisphere was found, with a RH advantage for face recognition in each case (left chimeric:  $F(1,39) = 7.58$ ,  $p < 0.01$ ; right chimeric:  $F(1,39) = 8.83$ ;  $p < 0.01$ ), again replicating a RH advantage for face identification. Comparing across left and right chimeric faces, both hemispheres showed a significant preference for left chimeric images, replicating the left-side bias effect.

## Christman et al. (1991) simulations

Two sets of 16 greyscale images (31x13 pixels) were constructed, each consisting of a two types of stimuli. The first type stimulus consisted of two frequency gratings at different relative phases to each other. The second stimulus type consisted of the first set of stimuli, with a third frequency grating superimposed upon them. In one stimulus set, the third frequency grating was at a higher spatial frequency than the other two frequency gratings; in the second stimulus set, it was at a lower spatial frequency than the other two. Importantly, the third frequency grating was of exactly the same spatial frequency in both stimulus sets (see Figure 6b for example stimuli). There were 4 phase variations for each stimulus type in each stimulus set.

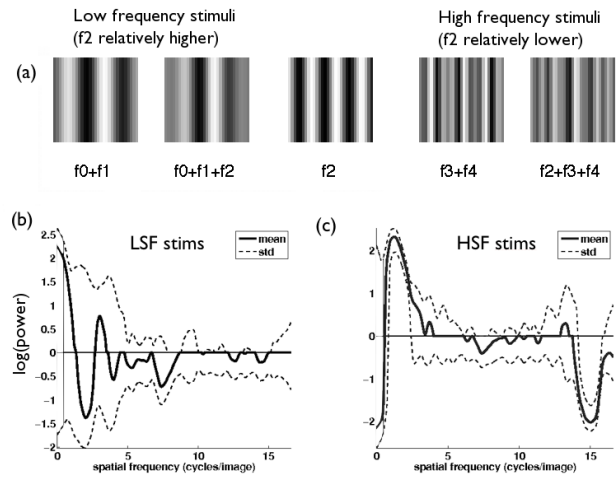


Figure 6: Original results, sample model stimuli, and model power spectrum  
(a) Sample stimuli created to train DE models  
(b) DE model spatial frequency analysis of output images, showing much flatter power spectrum differences than all other stimuli, with weak RH advantage (above zero on Y-axis) for LSF (towards left side of X-axis) and a weak LH advantage (below zero on Y-axis) for HSF (towards right side of X-axis)

Each set of greyscale images was presented to 64 LH ( $\sigma = 6.0$ ) and RH ( $\sigma = 3.0$ ) autoencoder models. Each autoencoder network had 360 hidden units, with each hidden unit connecting to 8 input and output units. Each autoencoder network was trained to 0.005 average error per output unit, then hidden unit encodings were extracted. For each set of greyscale images, a neural network with 360 input units, ten hidden units, and one output unit were trained to classify each of the 16 stimuli.

For both stimulus sets (LSF and HSF), there was a significant effect of stimulus class, with the 3-component stimulus being harder to classify than the 2-component stimulus. However, there was no hemisphere  $\times$  stimulus class interaction. Looking at the power spectrum differences between

the two stimulus classes (Figure 6b and 6c), and comparing them to the previous power spectrum differences, it is clear that there is much less encoding asymmetry between the two model hemispheres for these stimuli as compared to all other stimuli used in model experiments within this paper.

We tested a few possible explanations for this. We tried many different combinations of spatial frequency gratings; this varied which model hemisphere showed better performance, but no reliable hemisphere  $\times$  stimulus class interaction. We tried larger images, to expand the range of spatial frequencies that could be encoded, but again no consistency was found. Lastly, we tried training the autoencoder on separate dataset, then extracting hidden unit encodings on the task-relevant stimuli. Again, this did not show any consistent interaction.

Further work is warranted to better characterize whether the DE model can account for this critical dataset. We have created (elsewhere) a developmental model of this asymmetry which suggests that this dataset may be modeled by engagement of more than one cortical area showing asymmetry. This pattern is seen in neuroimaging results reported by (Hopf et al., 2006), for example, where a leftward asymmetry of an earlier visual processing area is engaged by a task at the “local” level, and a rightward asymmetry of a later visual processing area is engaged by a task at the “global” level. Variations in average inter-patch distance based on cortical area Amir et al. (1993) suggest that different areas may have different frequency preferences. This would suggest that “relative frequency” processing may in fact be simply selecting different absolute frequency filters, implemented in different cortical areas, based on task demands. We are currently investigating whether this might provide an alternate explanation to these data, and the idea of relative frequency encoding in general.

## Conclusions

Here, we showed that an asymmetry in local connectivity can account for local/global behavioral data, face processing data, and matches spatial frequency asymmetries reported in the literature. This model provides a biologically grounded implementation for these phenomena, and the analyses here showing consistent frequency filtering differences in the model hemispheres are consistent with the current algorithmic explanation for visual processing asymmetries via frequency filtering. Unlike the DFF model, however, these frequency filtering differences are found at a post-sensory encoding stage. Further work must be done to investigate whether our failure to model the results of Christman et al. (1991) is due to practical modeling concerns, or suggests a fundamentally different approach to modeling local/global processing asymmetry.

## Acknowledgments

This work was partly funded by a Center for Academic Research and Training in Anthropogeny (CARTA) fellowship,

as well as by NSF grant SMA 1041755 to the Temporal Dynamics of Learning Center, an NSF Science of Learning Center

## References

- Amir, Y., Harel, M., & Malach, R. (1993). Cortical hierarchy reflected in the organization of intrinsic connections in macaque monkey visual cortex. *The Journal of Comparative Neurology*, 334(1), 19–46.
- Brady, N., Campbell, M., & Flaherty, M. (2005). Perceptual asymmetries are preserved in memory for highly familiar faces of self and friend. *Brain and Cognition*, 58(3), 334–342.
- Christman, S., Kitterle, F. L., & Hellige, J. (1991). Hemispheric asymmetry in the processing of absolute versus relative spatial frequency. *Brain and Cognition*, 16(1), 62–73.
- Galuske, R. A., Schlote, W., Bratzke, H., & Singer, W. (2000). Interhemispheric asymmetries of the modular structure in human temporal cortex. *Science (New York, N.Y.)*, 289(5486), 1946–1949.
- Hopf, J., Luck, S. J., Boelmans, K., Schoenfeld, M. A., Boehler, C. N., Rieger, J., et al. (2006). The neural site of attention matches the spatial scale of perception. *The Journal of Neuroscience*, 26(13), 3532–3540.
- Hsiao, J., Shahbazi, R., & Cottrell, G. (2008). Hemispheric asymmetry in visual perception arises from differential encoding beyond the sensory level. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*.
- Hsiao, J., Shieh, D., & Cottrell, G. (2008). Convergence of the visual field split: Hemispheric modeling of face and object recognition. *Journal of Cognitive Neuroscience*, 20(12), 2298–2307.
- Hutsler, J., & Galuske, R. A. W. (2003). Hemispheric asymmetries in cerebral cortical networks. *Trends in Neurosciences*, 26(8), 429–35.
- Ivry, R. B., & Robertson, L. C. (1998). *The two sides of perception*. The MIT Press.
- Kitterle, F. L., Hellige, J. B., & Christman, S. (1992). Visual hemispheric asymmetries depend on which spatial frequencies are task relevant. *Brain and Cognition*, 20(2), 308–314.
- Levitt, J., & Lund, J. (2002). Intrinsic connections in mammalian cerebral cortex. In A. Schuez & R. Miller (Eds.), *Cortical areas: unity and diversity*. CRC Press.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536.
- Sergent, J. (1982). The cerebral balance of power: confrontation or cooperation? *Journal of Experimental Psychology. Human Perception and Performance*, 8(2), 253–72.
- Young, A. W., & Bion, P. J. (1981). Accuracy of naming laterally presented known faces by children and adults. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, 17(1), 97–106.



# **The face inversion effect and evoked brain potentials: Complete loss of configural information affects the N170**

**Ciro Civile (cc413@exeter.ac.uk)**

School of Psychology, College of Life and Environmental Sciences,  
University of Exeter, UK.

**Heike Elchlepp (H.Elchlepp@exeter.ac.uk)**

School of Psychology, College of Life and Environmental Sciences,  
University of Exeter, UK.

**R. McLaren (R.P.McLaren@exeter.ac.uk)**

School of Psychology, College of Life and Environmental Sciences,  
University of Exeter, UK.

**Aureliu Lavric (A.Lavric@exeter.ac.uk)**

School of Psychology, College of Life and Environmental Sciences,  
University of Exeter, UK.

**I.P.L. McLaren (I.P.L.McLaren@exeter.ac.uk)**

School of Psychology, College of Life and Environmental Sciences,  
University of Exeter, UK.

## **Abstract**

The face inversion effect (FIE) is a reduction in recognition performance for inverted faces compared to upright faces that is greater than that typically observed with other stimulus types (e.g. houses; Yin, 1969). Nevertheless, the demonstration that the inversion effect in recognition memory can be as strong with images of dogs as with faces when the subjects are experts in specific dog breeds (Diamond & Carey, 1986), suggests that there may be other factors, such as expertise, which give rise to the FIE. Event-related potentials (ERPs) were recorded while subjects performed an Old/New recognition study on normal and scrambled faces presented in upright and inverted orientations. We obtained the standard result for normal faces: The electrophysiological activity corresponding to the N170 was larger and delayed for normal inverted faces as compared to normal upright ones. On the other hand, the ERPs for scrambled inverted faces were not significantly larger or delayed as compared to scrambled upright stimuli. These results, in combination, show how the effect of inversion on the N170 is reliably greater when the faces are normal compared to scrambled, which suggests the disruption of configural information affects the FIE.

**Keywords:** Face recognition; Face inversion effect; N170; First and second order relational information; Old/new recognition task

## **Introduction**

Recognition of objects that are usually seen in one orientation is sometimes strongly impaired when the same objects are turned upside down, showing how intrinsically difficult it is to identify them. This has been found to be particularly the case for faces, leading to a phenomenon known as the face inversion effect (FIE). Thus, the fact that recognition of human faces is more impaired by inversion than is recognition for other stimuli has underlined how

faces are, in some sense, special. Some of the first evidence for the FIE reported by Yin (1969) presented participants with upright or inverted pictures of faces, airplanes, houses, and other stimuli. Following the study phase, participants were then tested with stimuli in the same orientation in a recognition task paradigm. The results showed that when the stimuli were studied and tested in an upright orientation, faces were better recognized than other sets of stimuli. However, when the same stimuli were presented and tested in an inverted orientation, recognition for faces decreased more than was the case for the other classes of stimuli. Yin (1969, experiment 3) replicated this result in an experiment using line drawings of facial stimuli and period costumes, thus controlling for the effect of subtle shadow information in an inverted face as a potential explanation for the large effect of inversion. In this experiment faces were not the easiest stimuli to be recognized when presented in an upright orientation. Therefore, the large FIE could not be attributed to the overall difficulty in discriminating within a stimulus category. Yin interpreted his results in terms of a face-specific process.

Over the past two decades more behavioral evidence has emerged that challenges the assumption that facial stimuli are special, not the least of which is the demonstration presented by Diamond and Carey (1986) that the inversion effect on recognition memory can be as strong with pictures of dogs as with faces when the subjects are experts in the identification and assessment of specific dog breeds. Given that the only stimuli that result in a substantial inversion effect are the ones for which the subjects have the necessary expertise, this suggests that the FIE may not be due to the fact that facial stimuli are subject to special processing because they are facial in nature, but instead that there are

other factors, such as expertise, which give rise to this effect. They distinguished between three types of information that can be used in recognition: isolated features, first-order relational features and second-order relational features. Isolated or local features are the independent constituent elements of an object. First-order information consists of spatial relations between constituent elements of an object, for example, the arrangement of the nose above the mouth. It is the first-order information that organizes a set of facial features into a face. Second-order information defines the relative size of these spatial relationships with regard to a base prototype. All faces tend to possess the same set of first-order relations, the essential manner in which faces differ from each other is captured by second-order relational information. These two kinds of relational information are both examples of configural information. Diamond and Carey suggested that large inversion effects will be obtained only if three conditions are met. Firstly, the members of the class of stimuli must share a configuration. Secondly, it must be possible to individuate the members of the class through second-order information. Finally, observers must have the expertise to exploit such second-order information. They proposed that the elements that distinguish faces lie on a continuum from isolated/local to second-order relational. Thus, recognition of members within the class differs from other types of recognition in its reliance on second-order relational features and in requiring expertise to use these features.

Searcy and Bartlett (1996) and Leder and Bruce (1998) have provided very clear evidence on the effect of disrupting configural information by inversion. In one of their experiments, Searcy and Bartlett (1996) made faces grotesque by either changing local elements, such as blackening teeth, blurring the pupils, or by changing the facial configuration. When shown in an inverted orientation, faces that were distorted through configural changes seemed to be more similar to the normal version, while the “locally distorted face” still looked grotesque. Thus, configural changes did not survive the inversion process as well as local ones. In another experiment, Leder and Bruce (1998) distorted faces so as to be more distinctive, either changing local features by giving them darker lips, bushier eye brows, etc. or by changing configural information to give a shorter mouth to nose spatial relation, etc. Distinctiveness impressions caused by distorted configural information disappeared when faces were presented in an inverted orientation compared to both upright faces and faces distorted in their local aspects. These results all provide evidence for the powerful effect that relational information has in the processing of upright faces relative to inverted faces. But there is still a question as to what precisely is the difficulty caused by any disruption of configural information consequent on inversion. The suggestion from some theories of perceptual learning (e.g. McLaren, 1997) is that expertise for faces can act directly on configural information, and confers the ability to make better use of it by effectively reducing the salience of first order relational

information (which is also configural but shared by most faces), leaving second order relational information relatively salient which aids discrimination. Thus, once configural information in upright faces has been disrupted (or at least our ability to make use of it), the benefits conferred by our expertise with those faces would tend to decrease, making them less easy to discriminate from one another. This explanation for the effect of expertise in face processing has some empirical support. The key finding is that it has been shown that experience with exemplars of a category that can be represented by a prototype (and have second order relational structure as a result of their variation about that prototype) leads to an increased ability to discriminate between members of that category. This improvement is lost when the stimuli are presented in an inverted orientation (McLaren, 1997). Recently, Civile *et al.*, (2011) provided some evidence that disrupting second order relational information by inverting (rotating by 180°) the eyes and the mouth, producing Thatcherised faces, whilst leaving other types of information (first order and local) intact reduces, even if it does not entirely eliminate, the FIE. However, in that same study they proposed that the FIE was still present for Thatcherised baseline stimuli (but significantly smaller than for normal faces) because Thatcherised faces still had some second-order information which had not been disrupted by the manipulation. Thus, in a second experiment they created a set of faces with all the second-order information disrupted. To do this they scrambled the faces by shuffling at random each of the features within a face. The effect of this was, in part, the expected one in that any inversion effect for the scrambled faces disappeared. The new finding was that performance for scrambled faces, whether in an upright or inverted orientation, was now below not only that for upright normal faces but also below that obtained for inverted normal faces. A possible explanation for this finding was that using scrambled faces may have affected both first and second order relational structure. In particular, when the ears were moved inside the face and replaced with other features, the typical shape of every face was changed to the point where it was no longer easily recognizable as a face.

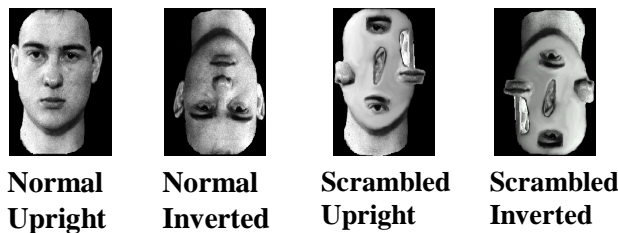
In this present study, we aimed to replicate behaviorally those results obtained by Civile *et al.*, (2011), but this time we used a slightly different design. Participants were presented with two old/new recognition tasks, one including male faces and another female faces. All together the sample of faces seen in this experiment was more than double that used in the Civile *et al.*, (2011) study. This was done so that we could measure event-related potentials (ERPs) recording subjects’ neural activity while performing the tasks. There have been previous ERPs studies that have compared the presentation of normal upright faces and normal inverted faces. Rossion *et al.*, (1999) recorded neural activity in a delayed-matching task. A larger amplitude and delayed activity on the N170 was found for inverted faces compared to upright faces suggesting that inversion may slow down and increase difficulty in face processing.



Following the ERP literature on the N170 (de Haan *et al.*, 2002; Eimer, 2000; Tanaka & Curran, 2001; Rossion *et al.*, 2002) we predicted a larger inversion effect on the N170 for normal faces compared to scrambled faces which suffer from disrupted configural information. We expected to obtain significantly larger and delayed N170 amplitudes for normal inverted faces compared to normal upright faces. This follows from the assumption that inversion effects the expertise needed to exploit configural information, and that the N170 depends on the ability to make use of this information. Thus, this difference is expected to be significantly larger than the one between the amplitudes for scrambled upright and inverted faces as the influence of expertise here will be minimal. We also looked for neural activity correlates to the disadvantage that scrambled faces have compared to normal inverted faces.

## Materials

The study used 320 images in total, half female and half male. These were photographs of faces of former students at the University of Cambridge. The faces were standardized in grey scale format using Adobe Photoshop. A program called Gimp 2.6 was used to manipulate the 320 stimuli. Any given face stimulus was prepared in four different versions i.e. normal upright, normal inverted, scrambled upright and scrambled inverted which were used in a counterbalanced fashion across participants so that each face was equally often used in each condition of the experiment. Six facial features were used for scrambling i.e. the mouth, nose, two ears and the two eyes (including eyebrows). Scrambling was done by selecting at random one feature of the face and moving it to the forehead (chosen because this is the widest space inside the face and so can accommodate any feature). Following this, a second feature was selected and moved to the space left empty by the first feature, and so on until all the six facial features had been moved. Examples of the stimuli used are given in Figure 1. The experiment was run using E-prime software Version 1.1 installed on a PC computer.



**Figure 1;** Examples of stimuli used in the experiment showing the four different facial conditions. The dimensions of the stimuli were 5.63cm x 7.84cm. The stimuli were presented at a resolution of 1280 x 960. Participants sat 1m away from the screen on which the images were presented.

## Participants

24 undergraduates and postgraduates at the University of Exeter took part in the experiment.

## Procedure

The experiment consisted of an initial ‘study phase’ followed by an ‘old/new recognition phase’ using only male faces, and then another ‘study phase’ and ‘old/new recognition phase’, but this time using only female facial stimuli. After the instructions, the first part of the experiment involved participants looking at 80 male faces (presented one at a time in random order). The participants saw a fixation cross in the centre of the screen that was presented for 500 ms, followed by a black screen for 500 ms and then by a facial stimulus that was presented for 3000ms. Then the fixation cross and the black screen were repeated, and another face presented, until all stimuli had been seen. These faces will be termed the “familiar” (designated as type 1) faces for that participant because they were presented again later on in the old/new recognition task. The face types were: Normal Inverted faces (1NI); Normal Upright faces (1NU); Scrambled Inverted faces (1SI) and Scrambled Upright faces (1SU). Following the study phase, after further instructions, there was an old/new recognition task in which participants were shown (in random order) the 80 male faces they had already seen (i.e. the familiar faces) intermixed with a further 80 unseen male faces which were designated as type 2 and split into the same four face sub-types as in the study phase. During this old/new recognition task participants indicated whether or not they had seen the male face during the study phase by pressing the ‘.’ key if they recognized the face or to press ‘x’ if they did not. Each face never appeared as more than one face sub-type at a time during the experiment. The facial stimuli available were divided into sets of 20 giving 8 sets of stimuli, and each participant group was shown a different combination of the 160 facial stimuli split over the 8 sets as shown in Table 1. Because there were 160 male faces to consider (80 in the study phase and 80 in the recognition task), four participant breaks were incorporated. These allowed participants to rest their eyes after they had viewed 40 male faces. The second part of the experiment followed the same procedure as that used in the first part of the experiment. The only difference this time was that participants saw female faces.

Face Type	Part. Group1	Part. Group2	Part. Group3	Part. Group4	Part. Group5	Part. Group6	Part. Group7	Part. Group8
1(1NI)	Set1	Set4	Set3	Set2	Set5	Set8	Set7	Set6
2(1NU)	Set2	Set1	Set4	Set3	Set6	Set5	Set8	Set7
3(1SI)	Set3	Set2	Set1	Set4	Set7	Set6	Set5	Set8
4(1SU)	Set4	Set3	Set2	Set1	Set8	Set7	Set6	Set5
5(2NI)	Set5	Set8	Set7	Set6	Set1	Set4	Set3	Set2
6(2NU)	Set6	Set5	Set8	Set7	Set2	Set1	Set4	Set3
7(2SI)	Set7	Set6	Set5	Set8	Set3	Set2	Set1	Set4
8(2SU)	Set8	Set7	Set6	Set5	Set4	Set3	Set2	Set1

**Table.1.**Combinations of facial stimuli presented to each

participant group. The same face set combinations were used in the first and second half of the experiment for the male and female faces.

## EEG Apparatus

The EEG was sampled continuously during study and recognition phases at 500 Hz with a bandpass of 0.016-100 Hz, the reference at Cz and the ground at AFz using 64 Ag/AgCl active electrodes and BrainAmp amplifiers. There were 61 electrodes on the scalp in an extended 10-20 configuration and one on each earlobe. Their impedances were kept below 10 k $\Omega$ . The EEG was filtered offline with a 20 Hz low-pass filter (24 dB/oct) and re-referenced to the linked ears.

## EEG Analysis

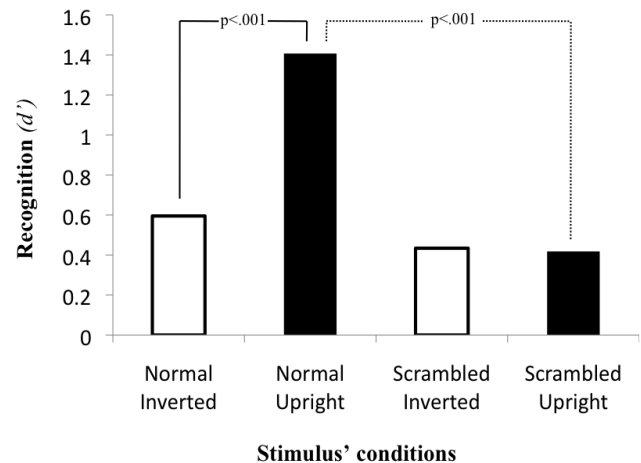
Peak amplitudes of the N170 in study and recognition phases were examined for differences between the experimental conditions. To improve the estimates of N170 amplitude and latency given the relatively small number of ERP segments in each condition (leading to a low signal-to-noise ratio), N170 extraction was aided by linear decomposition of the EEG by means of Independent Component Analysis (ICA, Bell & Sejnowski, 1995). ICA was run separately for each subject using all scalp channels and the entire dataset. For analyses of the recognition phase, segments associated with incorrect responses were discarded (there were no responses in the study phase). The remaining EEG segments were averaged for every participant and experimental condition. In each subject, we identified ICA components that: (1) showed a deflection (peak) in the N170 time-range (at 150-200 ms following stimulus onset), and (2) had a scalp distribution containing an occipital-temporal negativity characteristic of N170 (the scalp distributions of components are the columns of the inverted unmixing matrix). This resulted in 1-4 ICA components corresponding to the N170 identified in most subjects (mean 2.6; SD 1) - these were back-transformed into the EEG electrode space (by multiplying the components with the inverted unmixing matrix that had the columns corresponding to other components set to zero) and submitted to statistical analysis of N170 peak amplitude and latency.

## Results

### Behavioral Results

The data from all 24 participants contributed to the signal detection  $d'$  analysis. Responses for male and female faces were collapsed and transformed into  $d'$  measures. A significant interaction was found between face type and orientation,  $F(1,23) = 20.77$ ,  $p < .01$ . Figure 2 shows the results for the mean  $d'$  obtained for each face type. A planned comparison was used to examine whether or not there was a significant inversion effect for normal facial stimuli. This gave a highly significant advantage  $F(1,23) = 34.37$ ,  $p < .001$ , for normal upright faces vs. normal

inverted faces, and another planned comparison revealed no significant effect of inversion for scrambled upright vs. scrambled inverted faces,  $F(1,23) = 0.026$ ,  $p = \text{ns}$ . The effect of face type on the recognition of upright faces was also analyzed. Normal upright faces were recognized significantly better than scrambled upright faces  $F(1,23) = 56.75$ ,  $p < .001$ , but there was no significant difference in the recognition of normal inverted faces and scrambled inverted faces. Finally, scrambled upright were recognized significantly above chance,  $F(1,23) = 19.63$ ,  $p < .01$ , as were scrambled inverted faces,  $F(1,23) = 28.04$ ,  $p < .01$ .



**Figure 2;** Behavioral results from old/new recognition task. The X-axis shows the four different stimulus' conditions, whereas the Y-axis shows the  $d'$  means for each of the four facial conditions.

### N170 analysis

N170 latency and amplitude analyses were run in electrode PO8 which was the electrode showing most of the activity during our experiment. We attempted to run the same analyses on the N170 data as on the  $d'$  behavioral data considered earlier to facilitate comparison.

### Study phase (see Figure 3)

Latency analysis: The Face Type by Orientation interaction was significant, i.e. the effect of face inversion on N170 latencies was reliably larger when faces were Normal compared to Scrambled,  $F(1,23) = 7.79$ ,  $p < .05$ . In particular, the face inversion effect was highly reliable for Normal faces,  $F(1,23) = 24.54$ ,  $p < .001$ , with N170 latencies peaking 10 ms earlier for upright faces (at 175 ms) compared to inverted faces (186 ms). For scrambled faces, peaks for inverted faces were delayed compared to upright faces by less than 1 ms failing to reach significance,  $F(1,23) = 0.18$ ,  $p = \text{ns}$ . Latencies of upright faces peaked earlier (by 5 ms) when faces were Normal compared to Scrambled. This difference was reliable,  $F(1,23) = 5.36$ ,  $p < .05$ .

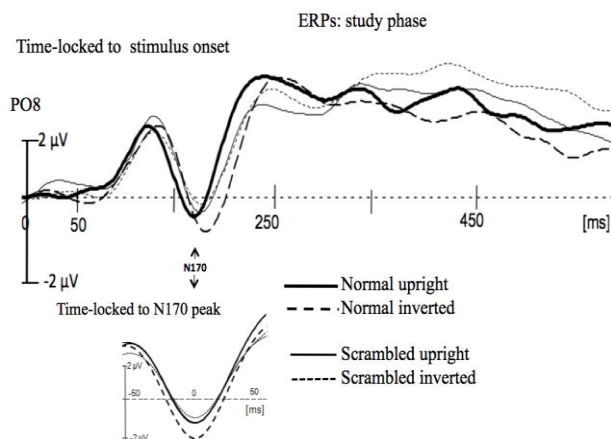
Peak amplitude analysis: The difference in peak amplitudes between upright and inverted faces was larger when faces were Normal ( $-0.61\mu\text{V}$ ) than when they were Scrambled ( $0.18\mu\text{V}$ ), but this was only marginally reliable,

$F(1,23) = 3.13$ ,  $p < .1$ . The effect of inversion neared significance for Normal faces,  $F(1,23) = 3.28$ ,  $p < .1$ , with more negative amplitudes for inverted ( $-1.56\mu V$ ) compared to upright ( $-0.94\mu V$ ) faces. For scrambled faces, the inversion effect did not approach significance  $F(1,23) = .075$ ,  $p = ns$ . The effect of Face Type was not reliable for upright faces,  $F(1,23) = 0.30$ ,  $p = ns$ . Amplitudes for inverted faces were significantly larger when the faces were Normal compared to Scrambled ( $-0.711\mu V$ )  $F(1,23) = 4.23$ ,  $p < .05$ .

#### **Old/new recognition task (see Figure 4)**

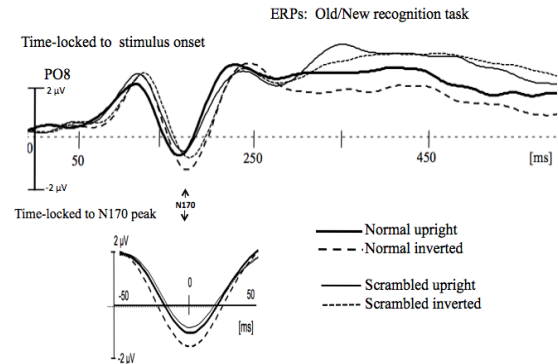
**Latency analysis:** A significant Orientation by Face Type interaction was found  $F(1,23) = 6.45$ ,  $p < .025$ . A significant inversion effect was observed for normal faces  $F(1,23) = 37.34$ ,  $p < .001$ , with N170 latencies peaking 9 ms earlier for upright faces (at 167 ms) compared to inverted faces (178 ms). A trend towards significance was found for the inversion effect related to scrambled faces  $F(1,23) = 2.51$ ,  $p = .13$  with N170 latencies peaking at nearly 4 ms earlier for upright Scrambled faces (at 176.3 ms) compared to inverted (179.90 ms). A final comparison revealed a significant effect for upright normal stimuli compared to scrambled upright stimuli  $F(1,23) = 9.06$ ,  $p < .01$ .

**Peak amplitude analysis:** No reliable Orientation by Face Type interaction was found. Means show a near significant inversion effect for Normal faces, with more negative amplitudes for inverted ( $-1.815\mu V$ ) vs. upright ( $-1.200\mu V$ ),  $F(1,23) = 3.67$ ,  $p = .06$ . No reliable difference was found for scrambled faces amplitudes  $F(1,23) = 0.79$ ,  $p = ns$ . No significant effect was found for upright normal stimuli compared to upright scrambled ones,  $F(1,23) = 1.03$ ,  $p = ns$ . However a significant effect was found for normal inverted faces compared to scrambled inverted ( $-1.216\mu V$ ),  $F(1,23) = 5.91$ ,  $p < .05$ .



**Figure.3.** The X-axis shows the elapsed time after a stimulus was presented, whereas the Y-axis shows the amplitudes ( $\mu V$ ) of the electrophysiological reactions in the study phase of the experiment. The insert in this figure is the ERP time-locked to the N170 peak, as identified in individual subjects. The time-scale of the inserts is stretched

relative to the main stimulus-locked ERP, the amplitude scale is the same in the insert as in the main figure.



**Figure.4.** The X-axis shows the elapsed time after a stimulus was presented. The Y-axis shows the amplitudes ( $\mu V$ ) of the electrophysiological reactions in the old/new recognition phase of the experiment. The insert in this figure is the ERP time-locked to the N170 peak, as identified in individual subjects during the old/new recognition task.

#### **Discussion**

On the behavioral side, and in agreement with the literature, we have obtained a strong inversion effect for normal faces. This has been eliminated entirely with scrambled faces. We have clear evidence here that configural information does indeed play an important role in driving the inversion effect for faces. Analyses on both the amplitude and latency of the N170 indicate a numerically larger inversion effect for normal faces than for scrambled faces. Running the same planned comparisons on the ERP data as for the behavioral data produces a very similar pattern of results, i.e. a strong inversion effect for the normal faces, and no inversion effect for scrambled faces, and a difference in N170 latencies between the upright normal and scrambled faces. The new finding here is that the scrambled stimuli (both upright and inverted) elicit a very similar N170 to one elicited by normal upright stimuli.

#### **General Discussion**

From the behavioral results of this study we have confirmed we can obtain a significant inversion effect with normal faces that can be eliminated entirely by disrupting both sources of configural information in the scrambled faces. This is consistent with our hypothesis that participants when presented with scrambled faces in an upright orientation would have no applicable expertise for those upright faces. Thus, when the same scrambled faces are presented in an inverted orientation, participants would not suffer any loss of expertise, as there was none to start with. Hence, we do not observe any inversion effect with scrambled faces. This supports the idea that the inversion effect observed with normal faces can at least in part be

explained by our ability to exploit configural information for categories of stimuli that possess both the necessary structure and are sufficiently familiar. If this structure is disrupted, then so is the inversion effect.

The ERP results provide neural correlates of our behavioral findings. In particular, in the study phase where participants were only asked to look at the faces and try to memorize them, analyses on both the amplitude and latency of the N170 gave a larger inversion effect for normal faces than for scrambled faces, and this result was highly significant for the latencies. Running the same planned comparisons on the ERP data as used for the behavioral data produces a very similar pattern of results, i.e. a strong inversion effect for the normal faces, none for the scrambled faces. However, If we study the waveforms that are time-locked to stimulus onset, then the new finding here is that both upright and inverted scrambled stimuli elicited a similar N170 to that for normal upright stimuli. This presents us with something of a mismatch with the behavioral patterns of results. According to the literature on face inversion and the N170, the ability to use configural information facilitates face processing, and this is supported by our behavioral results. The loss of configural information on inversion could have resulted in a selective amplification of the neural activity linked to faces because of an increase in difficulty due to a decrease in expertise for those faces presented in an inverted orientation (de Haan *et al.*, 2002; Eimer, 2000; Tanaka & Curran, 2001; Rossion *et al.*, 2002). In favor of this hypothesis is the correspondence between the behavioral data for the effects of inversion on normal faces and N170 for normal faces. This can also explain the lack of an inversion effect for the scrambled faces. Participants do not have expertise for these latter stimuli, thus the level of difficulty in processing them whether upright or inverted is the same leading to a similar N170 for both. However, according to this hypothesis we should have expected to obtain a larger and delayed N170 for scrambled faces compared to the N170 for normal upright faces and we did not. Instead, they are more similar to the N170 for normal upright faces.

Our results do agree with ERP studies using normal upright faces and familiar objects such as shoes or houses or chairs, in which it was found that the N170 elicited for these objects was more similar to that for normal upright faces. We can contrast this with the result obtained by Rossion *et al.* (2000), which compared the N170 elicited by a novel class of stimuli called Greebles which shared a common configuration to that obtained with faces and found it to be more like the N170 to inverted faces. This may suggest that our study may suffer from a lack of a correct baseline. Our scrambled stimuli were constructed by shuffling at random each of the features presented within a face. If our normal faces could be represented by a prototype, this was not the case for our scrambled faces, which instead varied a great deal in configuration because of the many different ways we shuffled the features within the face. It may be that our results show that participants perceived those scrambled

stimuli as many different types of object rather than as a set of new stimuli that could be represented by a prototype and shared a new configuration.

## Conclusion

In conclusion further research will be needed to evaluate the full implications of our results, but our data clearly suggest that there is a role for both first and second order structure in face recognition, that we argue can be understood in terms of experience-based expertise. And we have also shown that the elimination of the FIE can be correlated with a reduction of differences in neural activity in the N170.

## Acknowledgments

The research reported in this paper was supported by a Postgraduate studentship and an Exeter Graduate Fellowship awarded to Ciro Civile.

## References

- Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7, 1129-59.
- de Haan, M., Pascalis, O., & Johnson, M.H. (2002). Specialization of neural mechanisms underlying face recognition in human infants. *Journal of Cognitive Neuroscience*, 14, 199-209.
- Civile, C., R.P. McLaren., and I.P.L. McLaren (2011). Perceptual learning and face recognition: Disruption of second order relational information reduces the face inversion effect. In L. Carlson, C. Hoelscher, & T. Shipley (Eds.), *Proceedings of the 33<sup>rd</sup> Annual Conference of the Cognitive Science Society* (2083-2088). Austin, TX: Cognitive Science Society.
- Diamond, R. & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, 115, 107-117.
- Eimer, M. (2000). The face-specific N170 component reflects late stages in the structural encoding of faces. *NeuroReport*, 11, 2319-2324.
- McLaren, I.P.L. (1997). Categorization and perceptual learning: An analogue of the face inversion effect. *The Quarterly Journal of Experimental Psychology* 50A (2), 257-273.
- Rossion, B., Delvenne, J., Debatisse, D., Goffaux, V., Bruyer, R., Crommelinck, M., Guerit, M. (1999). Spatio-Temporal localization of the face inversion effect: an event-related potentials study. *Biological Psychology* 50, 173-189.
- Rossion, B., Gauthier, I., Goffaux, V., Tarr, M.-J., Crommelinck, M. (2002). Expertise training with novel objects leads to face-like electrophysiological responses. *Psychological Science*, 13, 250-257.
- Tanaka JW, Curran T. A neural basis for expert object recognition *Psychol Sci*. 2001 Jan 12(1):43-7.
- Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, 81, 141-145.





## **Face recognition and brain potentials: Disruption of configural information reduces the face inversion effect.**

**Ciro Civile (cc413@exeter.ac.uk)**

School of Psychology, College of Life and Environmental Sciences,  
University of Exeter, UK.

**Heike Elchlepp (H.Elchlepp@exeter.ac.uk)**

School of Psychology, College of Life and Environmental Sciences,  
University of Exeter, UK.

**R. McLaren (R.P.McLaren@exeter.ac.uk)**

School of Psychology, College of Life and Environmental Sciences,  
University of Exeter, UK.

**Aureliu Lavric (A.Lavric@exeter.ac.uk)**

School of Psychology, College of Life and Environmental Sciences,  
University of Exeter, UK.

**I.P.L. McLaren (I.P.L.McLaren@exeter.ac.uk)**

School of Psychology, College of Life and Environmental Sciences,  
University of Exeter, UK.

### **Abstract**

The face inversion effect (FIE) refers to the decline in performance in recognizing faces that are inverted compared to the recognition of faces in their normal upright orientation (Yin, 1969). Event-related potentials (ERPs) were recorded while subjects performed an Old/New recognition study on normal and Thatcherised faces presented in upright and inverted orientation. A large difference in processing between normal upright faces and normal inverted faces was observed at occipital-temporal sites about 165 ms following stimulus onset, mainly in the right hemisphere. Thus electrophysiological activity, which corresponds to the previously described N170, had larger amplitude and was delayed for normal inverted faces as compared to normal upright ones. By contrast, the activity for Thatcherised inverted faces was not significantly changed or delayed as compared to Thatcherised upright stimuli. These results combine to show how the effect of face inversion on the N170 is reliably greater when the faces are normal rather than Thatcherised. Finally, these findings complement, at a neural level, our behavioral studies which suggest that the loss of some configural information affects the FIE.

**Keywords:** Face inversion effect; N170; configural information.

### **Introduction**

The face inversion effect (FIE) is a reduction in recognition performance for inverted faces compared to upright faces that is greater than that typically observed with other stimulus types (e.g. pictures of houses; Yin, 1969). Nevertheless, the demonstration that the inversion effect in recognition memory can be as strong with images of dogs as with faces when the subjects are experts in specific dog breeds (Diamond & Carey, 1986), suggests that there may be other factors, such as expertise, which give rise to the FIE. Diamond and Carey (1986) proposed that there is a special type of information, “second order relational information” that we depend on with increasing expertise.

Their analysis was that human faces all have the same group of features (eyebrows, eyes, nose, mouth, etc.). All these faces tend to have in common the same basic disposition of components, such that the eyes are always above the nose and so on. Thus, “first order relational information” corresponds to the spatial relationship between the features of a face, and “second order relational information” corresponds to the small variations in the spatial relationships between these features that individuate the faces. This information can also be considered to be a type of configural information. Diamond and Carey (1986) suggested that a large inversion effect will be obtained only if three conditions are met. Firstly, the members of the class of stimuli must share a basic configuration. Secondly, it must be possible to individuate the members of the class through second-order information. Finally, individuals must have the expertise to exploit such second-order information. Thus, recognition of exemplars of such a class differs from other types of recognition in its reliance on second-order relational features and requires a certain expertise to use these features. This interpretation of the effect of expertise is supported by the role of a prototype in face recognition. In one of their papers Valentine and Bruce (1986a) suggested that a face prototype was a result of overlaying many examples of faces in a distributed memory network (e.g. as in McClelland & Rumelhart, 1985). Therefore, the emergence of a face prototype is not something special for faces, but occurs simply because facial stimuli constitute a homogeneous category of which many exemplars are experienced. Thus, prototype extraction would be expected to arise for any set of stimuli that satisfies the three conditions previously described for a large inversion effect. Conversely then, evidence of prototype extraction can be used to determine whether or not an observer possesses expertise in discriminating within a stimulus category. The

suggestion from some theories of perceptual learning (e.g. McLaren, 1997) is that expertise for faces acts directly on the representation of the information in a face, and confers the ability to make better use of it by effectively reducing the salience of first order relational information, leaving second order relational information relatively salient which aids discrimination. Thus, if the configural information in upright faces is disrupted, or our ability to extract it is (e.g. by inversion), the benefits conferred by our expertise with those faces would tend to decrease, making them less easy to discriminate from one another. This explanation for the effect of expertise in face processing has some empirical support. The key finding is that it has been shown that experience with exemplars of a category that can be represented by a prototype (and so have second order relational structure as a result of their variation about that prototype) leads to an increased ability to discriminate between members of that category (McLaren, Leewers and Mackintosh, 1994). This improvement is lost when the stimuli are presented in an inverted orientation (McLaren, 1997). Thus, the results from these studies taken together support the view that experience with stimuli may have a role in driving the specialization of processes subserving learning and memory.

This view receives support from event-related potentials (ERPs) studies such as Rossion, Gauthier, Goffaux, Tarr and Crommelinck (2002) who have shown that it is possible to obtain an electrophysiological inversion effect for an experimental non-face stimulus class called 'Greebles' once participants are trained in recognizing them. Rossion *et al.* (2002) trained participants with a three-phase experiment in which there was first, a baseline phase, where ERPs were recorded from responses to face and Greeble presentations in both upright and inverted orientations. Following this, there was a training phase using only upright Greebles. Finally, during the last phase of the experiment ERPs were measured using new faces and new Greebles presented in both upright and inverted orientations. ERPs prior to the training phase revealed the inversion effect to be larger for faces than for Greebles. Following training with upright Greebles, the N170 (negative deflection occurring between 150-200 ms) latencies for the upright faces and Greebles were similar. The ERPs for inverted faces remained roughly constant before and after the training phase with Greebles, but ERPs to Greebles showed a significant training effect, in that there was an increased delay and increased amplitude for inverted Greebles as compared with Greebles presented in an upright orientation. In conclusion, although the inversion effect for faces was larger in both experimental sessions, the inversion effect for Greebles increased with increasing expertise with that category of stimuli. Furthermore, Tanaka and Curran (2001) investigated the neural basis of object expertise while recording the brain activity of experts when categorizing images of common dogs and birds. Results showed that the magnitude of the N170 was larger when the participants categorized objects in the domain in which they were expert than when they

categorized objects in the domain in which they were novices. Finally, de Haan, Pascalis & Johnson (2002) investigated the inversion effect and the link to expertise using human and monkey faces, as the latter have a similar configuration of features to human faces. These two categories of stimuli were presented to participants in both upright and inverted orientations. Results revealed that the N170 amplitude evoked by upright faces was smaller than for other stimuli, and the amplitudes for monkey faces both upright and inverted, and inverted human faces did not differ significantly from one another. Thus, inversion increased the amplitude and latency for human faces but not for monkey faces. The same experiment conducted on 6-month-old infants produced a component with similar morphology to the N170. However, this infant component differed from the N170, both because it peaked 100 ms later and it was not affected by inversion. Thus, for adults the orientation of faces played a role in determining the N170 (Eimer, 2000), but for infants the influence of orientation appeared only at later processing stages. This absence of an inversion effect in the infant ERPs is consistent with the idea that adults develop expertise for face processing, including both species and orientation, as a consequence of experience with that stimulus category (de Haan *et al.*, 2002). These results also suggest that ERP inversion effects are tied to expertise with a suitable category, rather than to the category of faces *per se*.

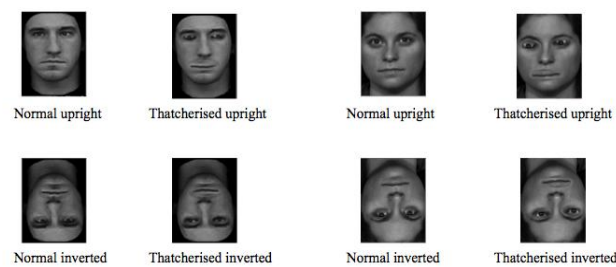
## EXPERIMENT

In this study we investigated the link between second-order relational structure and the face inversion effect suggested by Diamond and Carey (1986). The argument is that the improvement brought about by our expertise with faces is lost on inversion because this disrupts the ability to exploit second order relational information, leading to a strong inversion effect. In the behavioral part of this study, we aimed to demonstrate the typical strong inversion effect for normal face stimuli (for which we have expertise), and for comparison purposes ran a condition using what are known as Thatcherised face stimuli (see Fig. 1 for examples). These latter stimuli serve as our experimental manipulation in the sense that they suffer from somewhat disrupted second order-relational information (even when upright) caused by the 180° rotation of the eyes and the mouth, which should reduce at least some of the effect of expertise in the upright orientation. Another useful characteristic of these stimuli is that they are still faces, and are well matched for complexity with the normal faces. We also investigated the electrophysiological responses to normal faces in comparison with the responses obtained to Thatcherised faces and predicted that the N170 would correlate with our behavioral results. That is, the N170 for upright normal faces was expected to be different from that obtained in our other conditions. We expected to observe larger and delayed N170 amplitudes for inverted normal faces, as well as for upright and inverted Thatcherised faces, by analogy with the results of de Haan *et al.* (2002). This

follows from the assumption that the disrupted second order relational information in Thatcherised faces in part reduces the effect of expertise in the same way that inversion also reduces its impact, and that the N170 depends, at least in part, on the effect of expertise. Hence we expect the effect of expertise to only be evident for normal upright faces, and to manifest as a smaller amplitude and latency, leading to a large inversion effect (in the ERPs) for normal faces but not for Thatcherised faces.

### Materials

The study used 320 images in total, half female and half male. These were photographs of faces of former students at the University of Cambridge. The faces were standardized in grey scale format using Adobe Photoshop. A program called Gimp 2.6 was used to manipulate the 320 stimuli. Any given face stimulus was prepared in four different versions i.e. normal upright, normal inverted, Thatcherised upright and Thatcherised inverted, which were used in a counterbalanced fashion across participants so that each face was equally often used in each condition of the experiment. For the Thatcherised faces, each of the eyes and the mouth were flipped about the horizontal axis. Examples of the stimuli used are given in Figure 1. The experiment was run using E-prime software Version 1.1 installed on a PC computer.



**Figure 1;** Examples of stimuli used in the experiment showing the four different conditions for male and female faces. The dimensions of the stimuli were 5.63cm x 7.84cm. The stimuli were presented at a resolution of 1280 x 960 . Participants sat 1m away from the screen on which the images were presented.

### Participants

32 undergraduates and postgraduates at the University of Exeter took part in the experiment.

### Procedure

The experiment consisted of a ‘study phase’ and an ‘old/new recognition phase’ using only male faces, followed by another ‘study phase’ and ‘old/new recognition phase’, but this time using only female facial stimuli. After the instructions, the first part of the experiment involved participants looking at 80 male faces (presented one at a time in random order).The participants saw a fixation cross

in the centre of the screen that was presented for 500 ms. This was followed by a black screen for 500 ms and then by a facial stimulus that was presented for 3000ms. Then the fixation cross and the black screen were repeated, and another face presented, until all stimuli had been seen. These faces will be termed “familiar”(designated as type 1) faces for that participant because they will be presented again later on in the old/new recognition task. The face types during the study phase were: Normal Inverted faces (1NI); Normal Upright faces (1NU); Thatcherised Inverted faces (1TI) and Thatcherised Upright faces (1TU). Following the study phase, after further instructions, there was an old/new recognition task in which participants were shown (in random order) the 80 male faces they had already seen (i.e. the familiar faces) intermixed with a further 80 unseen male faces which were designated as type 2 (novel) and split into the same four face sub-types as the familiar set. During this old/new recognition task participants indicated whether or not they had seen the male face onscreen during the study phase by pressing the ‘.’ key If they recognized the face or by pressing ‘x’ if they did not. Each facial stimulus had a unique identifying number, to make sure that individual faces never appeared in more than one face type at a time during the experiment. To simplify their use in the experiment, the facial stimuli available were divided into sets of 20 giving 8 sets of stimuli, and each participant group was shown a different combination of the 160 facial stimuli rotated over the 8 sets as shown in Table 1. Because there were 160 male faces to consider (80 in the study phase and 80 in the recognition task), four participant breaks were incorporated. These allowed participants to rest their eyes after they had viewed 40 faces. The second part of the experiment followed the same procedure as that used in the first part of the experiment. The only difference this time was that participants saw female faces.

Face Type	Participant Group 1	Participant Group 2	Participant Group 3	Participant Group 4	Participant Group 5	Participant Group 6	Participant Group 7	Participant Group 8
1 (1NI)	Set 1	Set 4	Set 3	Set 2	Set 5	Set 8	Set 7	Set 6
2(1NU)	Set 2	Set 1	Set 4	Set 3	Set 6	Set 5	Set 8	Set 7
3(1TI)	Set 3	Set 2	Set 1	Set 4	Set 7	Set 6	Set 5	Set 8
4(1TU)	Set 4	Set 3	Set 2	Set 1	Set 8	Set 7	Set 6	Set 5
5(2NI)	Set 5	Set 8	Set 7	Set 6	Set 1	Set 4	Set 3	Set 2
6(2NU)	Set 6	Set 5	Set 8	Set 7	Set 2	Set 1	Set 4	Set 3
7(2TI)	Set 7	Set 6	Set 5	Set 8	Set 3	Set 2	Set 1	Set 4
8(2TU)	Set 8	Set 7	Set 6	Set 5	Set 4	Set 3	Set 2	Set 1

**Table.1.**Combinations of facial stimuli presented to each participant group. The same face set combinations were used in the first and second half of the experiment for the male and female faces.

### EEG Apparatus

The EEG was sampled continuously during both the study and test phases at 500 Hz with a bandpass of 0.016-100 Hz, the reference at Cz and the ground at AFz using 64 Ag/AgCl active electrodes and BrainAmp amplifiers. There were 61 electrodes on the scalp in an extended 10-20 configuration and one on each earlobe. Their impedances



were kept below 10 k $\Omega$ . The EEG was filtered offline with a 20 Hz low-pass filter (24 dB/oct) and re-referenced to the linked ears.

## EEG Analysis

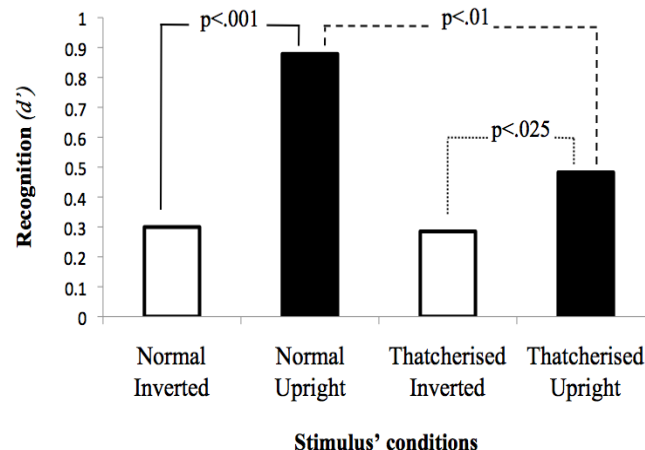
Peak amplitudes of the N170 in study and recognition phases were examined for differences between the experimental conditions. To improve the estimates of N170 amplitude and latency given the relatively small number of ERP segments in each condition (leading to a low signal-to-noise ratio), N170 extraction was aided by linear decomposition of the EEG by means of Independent Component Analysis (ICA, Bell & Sejnowski, 1995). ICA was run separately for each subject using all scalp channels and the entire dataset. For analyses of the recognition phase, segments associated with incorrect responses were discarded (there were no responses in the study phase). The remaining EEG segments were averaged for every participant and experimental condition. In each subject, we identified ICA components that: (1) showed a deflection (peak) in the N170 time-range (at 150-200 ms following stimulus onset), and (2) had a scalp distribution containing an occipital-temporal negativity characteristic of N170 (the scalp distributions of components are the columns of the inverted unmixing matrix). This resulted in 1-4 ICA components corresponding to the N170 identified in most subjects (mean 2.6; SD 1) - these were back-transformed into the EEG electrode space (by multiplying the components with the inverted unmixing matrix that had the columns corresponding to other components set to zero) and submitted to statistical analysis of N170 peak amplitude and latency.

## Results

### Behavioral Results

The data from all 32 participants contributed to the signal detection  $d'$  analysis. Responses for male and female faces were collapsed and transformed into  $d'$  measures. There was a significant interaction between face type and orientation,  $F(1,31) = 8.30$ ,  $p < .01$ . This reflected the fact that the inversion effect in the normal faces was significantly greater than that in the Thatcherised faces. Figure 2 shows the results for the mean  $d'$  obtained for each face type. A planned comparison gave a highly significant advantage  $F(1,31) = 29.99$ ,  $p < .001$ , for normal upright faces vs. normal inverted faces, and another planned comparison showed a similar (although smaller) inversion effect for Thatcherised upright vs. Thatcherised inverted faces,  $F(1,31) = 6.24$ ,  $p < .025$ . To further investigate this result, the effect of face type on the recognition of upright faces was also analyzed. Normal upright faces were recognized significantly better than Thatcherised upright faces  $F(1,31) = 13.71$ ,  $p < .01$ , but there was no significant difference in the recognition of normal inverted faces and Thatcherised inverted faces. Thus, it would seem that the reduction in the inversion effect for Thatcherised faces is more due to the impact that

Thatcherisation has on the upright faces rather than on the inverted ones.



**Figure 2;** Results for the old/new recognition task. The X axis shows the four different stimulus' conditions, the Y axis shows the mean  $d'$  for each condition.

### N170 analysis

Three participants had to be excluded because ICA did not find any components containing the N170 (nor was there an N170 visible in the original ERP). N170 latency and amplitude analyses were run in electrode PO8 which was the one showing most of the activity during our experiment. We attempted to run the same analyses on the N170 data as on the  $d'$  behavioral data considered earlier to facilitate comparison.

### Study phase (see Figure 3)

**Latency analysis:** The Orientation x Face Type interaction, i.e. the effect of inversion on N170 latencies, was reliably larger when faces were Normal compared to Thatcherised,  $F(1,28) = 4.73$ ,  $p < .05$ . In particular, the effect was highly reliable for Normal faces,  $F(1,28) = 21.19$ ,  $p < .01$ , with N170 latencies peaking 9 ms earlier for upright faces (at 165 ms) compared to inverted faces (174 ms). For Thatcherised faces, peaks for inverted faces were delayed compared to upright faces by 3 ms. This delay did not reach significance,  $F(1,28) = 1.54$ ,  $p = \text{ns}$ . Latencies of upright faces peaked earlier (by 4 ms) when faces were Normal compared to Thatcherised. This difference was only marginally reliable,  $F(1,28) = 3.24$ ,  $p = .082$ .

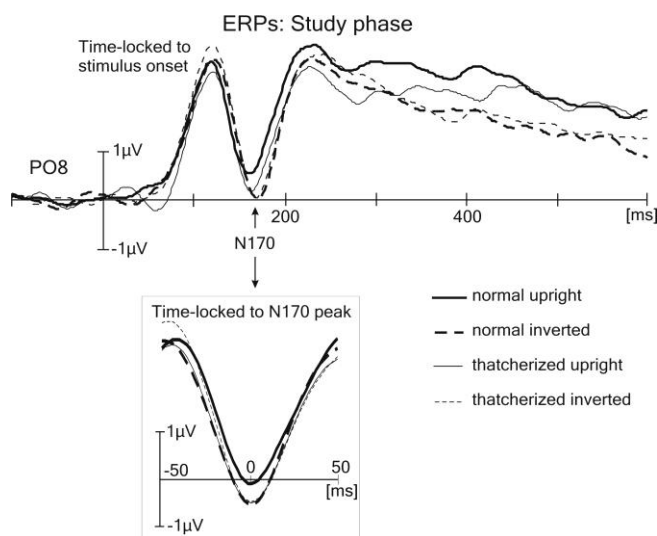
**Peak amplitude analysis:** The difference in peak amplitudes between upright and inverted faces was significantly larger when faces were Normal ( $-0.46\mu\text{V}$ ) than when they were Thatcherised ( $0.002\mu\text{V}$ ),  $F(1,28) = 4.18$ ,

$p=.05$ . The effect of inversion was reliable for Normal faces,  $F(1,28) = 7.06$ ,  $p < .025$ , with more negative amplitudes for inverted ( $-0.513\mu V$ ) compared to upright ( $-0.046\mu V$ ) faces. For Thatcherised faces the inversion effect did not approach significance  $F(1,28) = .0001$   $p=ns$ . The effect of Face Type was marginally reliable for upright faces,  $F(1,28) = 3.82$ ,  $p=.06$ , with more negative amplitudes for Thatcherised ( $-0.451\mu V$ ) compared to Normal ( $-0.046\mu V$ ) faces.

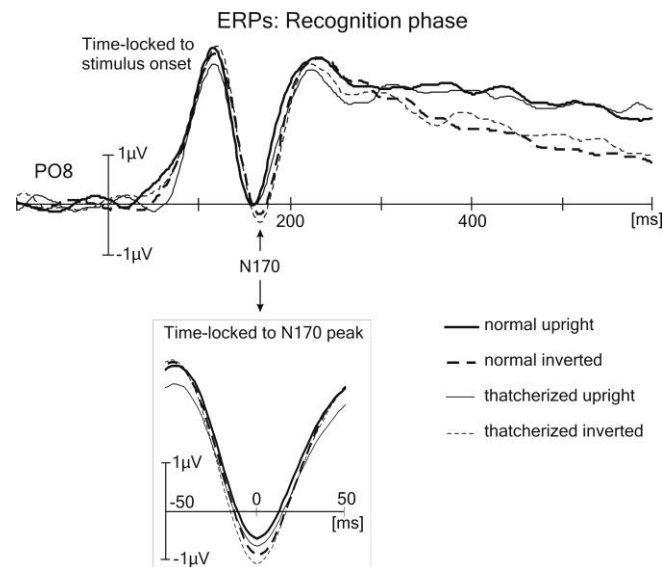
#### Old/new recognition task (see Figure 4)

**Latency analysis:** No significant Orientation by Face Type interaction was found. A significant inversion effect was obtained for normal faces  $F(1,28) = 16.36$ ,  $p < .01$  with N170 latencies peaking 5 ms earlier for upright faces (at 163 ms) compared to inverted faces (168 ms). A reduced but still significant inversion effect was found for Thatcherised faces  $F(1,28) = 6.62$ ,  $p < .025$  with N170 latencies peaking at nearly 5 ms earlier for upright Thatcherised faces (at 165.31 ms) compared to inverted (169.72 ms). A planned comparison revealed a trend towards significance for upright normal stimuli compared to Thatcherised upright ones  $F(1,28) = 2.27$ ,  $p=.15$ .

**Peak amplitude analysis:** As for latencies, no reliable Orientation by Face Type interaction was found. Means show a trend towards significance for Normal faces, with more negative amplitudes for inverted ( $-0.73\mu V$ ) vs. upright ( $-0.39\mu V$ ),  $F(1,28) = 2.50$ ,  $p=.13$ . For Thatcherised faces amplitudes are reliably more negative when they are inverted ( $-0.91\mu V$ ) vs. upright ( $-0.54\mu V$ ),  $F(1,28) = 4.59$ ,  $p < .05$ .



**Figure.3.** The X axis shows the elapsed time after a stimulus was presented, whereas the Y axis shows the amplitudes ( $\mu V$ ) of the electrophysiological reactions in the study phase of the experiment. The insert in this figure is the ERP time-locked to the N170 peak, as identified in individual subjects. The time-scale of the inserts is stretched relative to the main stimulus-locked ERPs, the amplitude scale is the same in the insert as in the main figure.



**Figure.4.** The X axis shows the elapsed time after a stimulus was presented. The Y axis shows the amplitudes ( $\mu V$ ) of the electrophysiological reactions in the old/new recognition phase of the experiment.

#### Discussion

This study has, in essence, confirmed our predictions. On the behavioral side we have obtained a strong inversion effect for normal faces and a reduced one for Thatcherised faces. The ERP results provide the sought after correlates of our behavioral findings in the study phase where participants were only asked to look at the faces and try to memorize them. Analyses on both the amplitude and latency show a larger inversion effect on the N170 for normal faces than for Thatcherised faces. Running the same planned comparisons on the ERP data as for the behavioral data produces a very similar pattern of results, i.e. a strong inversion effect for the normal faces, a greatly reduced effect for the Thatcherised faces, and a difference in N170 amplitude between the upright normal and Thatcherised faces but not between the two face types when inverted.

#### General Discussion

The behavioral results of this study show that we have obtained a significant inversion effect with normal faces, and have demonstrated that it is significantly larger than the inversion effect obtained with Thatcherised faces. To some extent, then, we have confirmed the basic face inversion finding. We have some evidence here that second order relational information plays a role in driving the inversion effect for faces. The most straightforward explanation of the difference in performance to the two face types when upright is that the Thatcherised faces have lost some (but not all) of the benefit of our expertise in dealing with second order structure. Because the Thatcherised faces are still essentially faces, then the application of our expertise with normal faces may lead to positively unhelpful results for upright Thatcherised faces, in that the changed features

stand out and command processing. Because these features are not those best suited to individuate faces, i.e. our processing is being dominated to a greater extent by what is common to Thatcherised faces (because they are surprising) rather than what would aid us in discriminating between them, performance for upright Thatcherised faces would be expected to be worse than for normal upright faces. The lack of any difference in recognition performance between normal and Thatcherised faces when inverted can be explained by arguing that in these circumstances second order relational information is not in play, and the two types of face are otherwise equated in terms of features and other factors (e.g. overall shape of the face).

The results from the ERPs bolster our interpretation of the effects we obtained in the behavioral results. As we predicted, the N170 to upright normal faces was different to that of our other stimuli, an effect that we can now argue reflects in part the high degree of expertise participants had for them. One of our findings is that this difference was a great deal clearer in the study phase of our experiment than in the test phase. This is not an entirely unexpected result. Firstly, if the modulation of the N170 reflects an effect of expertise, then this should occur when simply perceiving the stimulus – the effect is not tied to having to do anything in particular, except perhaps attend to the stimulus. Secondly, as a result of the study phase, the Thatcherised stimuli will start to become familiar, in particular the Thatcherised upright faces will tend to become progressively more equivalent to normal upright faces. Thus, any effect in the study phase will be a relatively pure comparison of the two stimulus types, one highly familiar, the other novel (at least in part); but in the test phase this distinction, and the effects that flow from it, will be attenuated by participants' increasing familiarity with the Thatcherised stimuli. If we study the waveforms that are time-locked to stimulus onset then the pattern at the N170 exactly corresponds to that observed in the behavioral data. As we predicted, upright normal faces occur earlier and with smaller amplitude in the N170, upright Thatcherised faces are somewhat later and have greater amplitude, and both the inverted face types are slightly later still and have slightly greater amplitude than upright Thatcherised faces. We suggest that the N170 is indexing, at least in part, the effect of expertise with the stimulus category. Inversion of the faces increases the amplitude of the N170 and delays its onset in agreement with a number of other studies which have found a greater delay and larger amplitude for the inverted stimulus (Rossion *et al*, 2002; Tanaka and Curran, 2001; de Haan *et al*, 2002). We note that the FIE for our Thatcherised stimuli is still significant, suggesting that simply disrupting second order information does not completely eliminate the FIE. A possible explanation for this is that by rotating the eyes and the mouth we have not disrupted all the second order information in a face. Thus, our baseline stimuli still have some second order information which participants may have expertise for. Another explanation could be that not only second order information is involved in the FIE but

there may be an important role for other types of information. Perhaps by disrupting both first and second order configural information we would be able to eliminate the FIE entirely. Our claims about the magnitude of the inversion effect are secure, but we cannot tell if performance in all our conditions is still benefiting from the effects of expertise (all the stimuli are, after all, recognizable as faces). One obvious way in which this might happen is by virtue of all the face types containing standard facial features that have not been themselves changed apart from a rotation or reflection. Another would be to appeal to the basic envelope of the stimuli remaining unchanged under Thatcherisation and inversion. Clearly it would be unwise to assume that all effects of expertise disappear under inversion, Thatcherisation or a combination of the two manipulations. What we can conclude, however, is that Thatcherisation interacts with stimulus inversion in a way that strongly suggests that experience with these stimuli helps us to better exploit that information.

## References

- Bell, A. J. & Sejnowski, T. J. (1995). An information–maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7, 1129-59.
- de Haan, M., Pascalis, O., & Johnson, M.H. (2002). Specialization of neural mechanisms underlying face recognition in human infants. *Journal of Cognitive Neuroscience*, 14, 199-209
- Diamond, R. & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, 115, 107-117.
- Eimer, M. (2000). The face-specific N170 component reflects late stages in the structural encoding of faces. *NeuroReport*, 11, 2319-2324.
- McClelland, J. L. & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, 114, 159-197.
- McLaren, I.P.L. (1997). Categorization and perceptual learning: An analogue of the face inversion effect. *The Quarterly Journal of Experimental Psychology* 50A (2), 257-273.
- McLaren, I.P.L., Leivers, H.L. & Mackintosh, N.J. (1994). Recognition, categorisation and perceptual learning. *Attention & Performance XV*. Cambridge, MA: MIT Press.
- Rossion, B., Gauthier, I., Goffaux, V., Tarr, M.-J., Crommelinck, M. (2002). Expertise training with novel objects leads to face-like electrophysiological responses. *Psychological Science*, 13, 250-257
- Tanaka JW, Curran T. (2001) A neural basis for expert object recognition *Psychol Sci*, 2001 Jan 12(1):43-7.
- Valentine, T., & Bruce, V. (1986a). Recognizing familiar faces : The role of distinctiveness and familiarity. *Canadian Journal of Psychology*, 40, 300-305.
- Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, 81, 141-145.

# Strength of Perceptual Experience Predicts Word Processing Performance Better than Concreteness or Imageability

Louise Connell ([louise.connell@manchester.ac.uk](mailto:louise.connell@manchester.ac.uk))

School of Psychological Sciences, University of Manchester  
Oxford Road, Manchester M13 9PL, UK

Dermot Lynott ([dermot.lynott@manchester.ac.uk](mailto:dermot.lynott@manchester.ac.uk))

Decision and Cognitive Sciences Research Centre, Manchester Business School, University of Manchester  
Booth Street West, Manchester M15 6PB, UK

## Abstract

Abstract concepts are traditionally thought to differ from concrete concepts by their lack of perceptual information, which causes them to be processed more slowly and inaccurately than perceptually-based concrete concepts. We examined this assumption by comparing concreteness and imageability norms to a set of perceptual strength norms in five separate modalities: sound, taste, touch, smell and vision. Results showed that that concreteness and imageability do not actually reflect the perceptual basis of concepts: concreteness ratings appear to be based on two different intersecting decision criteria, and imageability ratings are visually biased. Analysis of lexical decision and word naming performance showed that maximum perceptual strength (i.e., strength in the dominant perceptual modality) consistently outperformed both concreteness and imageability in accounting for variance in response latency and accuracy. We conclude that so-called concreteness effects in word processing emerge from the perceptual strength of a concept's representation and discuss the implications for theories of conceptual representation.

**Keywords:** abstract concepts; imageability; concreteness effects; perceptual strength; lexical decision; word naming

## Introduction

What exactly constitutes an abstract concept? Traditionally, abstract words such as *truth* or *impossible* are assumed to refer to things that are not perceptually experienced, while concrete words such as *chair* or *turquoise* are assumed to refer to perceptible, material entities. A long history of research has examined processing differences between such abstract and concrete concepts. In particular, concreteness effects refer to a behavioral advantage for words that refer to concrete concepts, which are processed more quickly and accurately than abstract concepts in tasks such as lexical decision and word naming (e.g., Binder et al., 2005; James, 1975; Kroll & Merves, 1986; Schwanenflugel, Harnishfeger & Stowe, 1988; Schwanenflugel & Stowe, 1989).

A number of different theories have been proposed to account for concreteness effects in word processing performance. Dual coding theory (Paivio, 1986, 2007) holds that both concrete and abstract concepts have a verbal code representation, but that concrete concepts alone also have a nonverbal, perceptual code that “gives rise to conscious (reportable) imagery when activated” (Paivio, 2007, p. 39). Abstract words are slower to process because they can only be imaged indirectly, via related concrete words. Context availability theory (Schwanenflugel &

Shoben, 1983; Schwanenflugel et al., 1988) instead argues that the type of information is less important than the quantity, and that concrete concepts are strongly linked to a narrow range of supporting contexts in memory whereas abstract concepts are weakly linked to a wide range. People are slower to process abstract words because they find it more difficult to retrieve associated contextual information. More recently, situated simulation views of conceptual representation (Barsalou & Wiemer-Hastings, 2005; Barsalou, Santos, Simmons & Wilson, 2008; see also Kousta, Vigliocco, Campo, Vinson & Andrews, 2011) have drawn together several aspects of both dual coding and context availability theories. Concrete concepts are represented in a narrow range of situations that focus on perceptual and motor information, while abstract concepts have a wide range of situations that focus on social, introspective and affective information. Abstract words are slower to process because people find it more difficult to access their situations.

However, despite their reputation as a textbook effect, concreteness effects do not always reliably emerge in semantic processing. Null effects are rarely publishable, but lack of concreteness effects in response times and error rates are not uncommon in cognitive neuroscience studies where significant findings on other measures are reported alongside null behavioral results (e.g., Fiebach & Friederici, 2003; Papagno, Fogliata, Catricalà & Miniussi, 2009; Sabsevitz, Medler, Seidenberg & Binder, 2005; Tyler, Russell, Fadili, Moss, 2001). Furthermore, reverse concreteness effects – a processing advantage for abstract concepts rather than concrete – have also been found in studies of healthy adult participants (e.g., Adelman, Brown & Quesada, 2006; Kousta et al., 2011). Such null and reversed concreteness effects are problematic for theories that claim fundamental representational differences between concrete and abstract concepts.

One reason for inconsistencies in empirical tests of concreteness effects may be that the intuitive and theoretical assumption is valid (i.e., that concepts with perceptual information are faster to process), but that the typical basis for selecting experimental items (i.e., concreteness or imageability ratings) does not offer an accurate measure of the perceptual basis of concepts. Most researchers select items from published norms such as the MRC psycholinguistic database (available online at [http://www.psy.uwa.edu.au/MRCDataBase/uwa\\_mrc.htm](http://www.psy.uwa.edu.au/MRCDataBase/uwa_mrc.htm)).

However, when we examined the original norming instructions used to collect these norms, we found it questionable that participants would have simultaneously considered their sensory experience across all modalities and then managed to aggregate this experience into a single, composite rating per word. Instructions for concreteness ratings, for example, define concrete words as referring to “objects, materials, or persons” and abstract words as referring to something that “cannot be experienced by the senses” (Paivio, Yuille & Madigan, 1968, p. 5). The resulting ratings, therefore, may reflect different decision criteria at the concrete and abstract ends of the scale, which is consistent with previous observations that the concreteness ratings scale has a bimodal distribution (e.g., Kousta et al., 2011). Imageability ratings are frequently used interchangeably with concreteness ratings (e.g., Binder et al., 2005; Sabsevitz et al., 2005) because of their high correlation and theoretical relationship in dual coding theory. Instructions for imageability ratings repeatedly refer to arousing a “mental image” (Paivio et al., 1968, p. 4), which is likely to lead naïve participants to focus on vision at the expense of other modalities. Both concreteness and imageability ratings could therefore add considerable noise to any dataset that assumed the ratings reflected a smooth continuum of perceptual experience across all modalities.

Our goals in the present paper were twofold. First, we aimed to establish whether concreteness and imageability norms actually reflect the degree with which concepts are perceptually experienced, as is commonly assumed. Second, we examined whether so-called concreteness effects in word processing are better predicted by concreteness/imageability ratings or by strength of perceptual experience. If the former, then forty years of empirical methodology have been validated but the reasons for null and reverse concreteness effects remain unclear. If the latter, then concreteness and imageability ratings are unsuitable for the tasks in which they are employed, and null and reverse concreteness effects are due to the unreliability of perceptual information in these ratings.

## Experiment 1

Rather than ask participants to condense their estimations of sensory experience into a single concreteness or imageability rating, modality-specific norming asks people to rate how strongly they experience a variety of concepts using each perceptual modality in turn (i.e., auditory, gustatory, haptic, olfactory or visual: Lynott & Connell, 2009, in prep.; see also Connell & Lynott, 2010; Louwerse

& Connell, 2011).

If concreteness and imageability are a fair reflection of the degree of perceptual information in a concept, then ratings of perceptual strength in all five modalities should be positively related to concreteness and imageability ratings, and these relationships should remain consistent across the rating scale. On the other hand, if we were correct in our hypothesis to the contrary, then we would expect some perceptual modalities to be neglected (i.e., no relationship) or even misinterpreted (i.e., negative relationship) in concreteness and imageability ratings. Specifically, concreteness norming instructions may have led to different decision criteria and therefore distinctly different modality profiles at each end of scale, whereas imageability instructions may have led to a predominantly visual bias.

## Method

**Materials** A total of 592 words were collated that represented the overlap of the relevant sets of norms, so each word had ratings of perceptual strength on five modalities as well as concreteness and imageability (see Table 1 for sample items). Perceptual strength norms came from Lynott and Connell (2009, in prep.), in which participants were asked to rate “to what extent do you experience WORD” (for nouns) or “to what extent do you experience something being WORD” (for adjectives) through each of the five senses (i.e., “by hearing”, “by tasting”, “by feeling through touch”, “by smelling” and “by seeing”), using separate rating scales for each modality. Perceptual strength ratings therefore took the form of a 5-value vector per word, ranging from 0 (low strength) to 5 (high strength). Concreteness ratings were taken from the MRC psycholinguistic database for 522 words, with ratings for the remaining 70 words coming from Nelson, McEvoy and Schreiber (2004). Imageability ratings for 524 words also came from the MRC database, and were supplemented with ratings for a further 68 words from Clark and Paivio (2004). All concreteness and imageability ratings emerged from the same instructions as Paivio et al.’s (1968) original norms, and ranged from 100 (abstract or low-imageability) to 700 (concrete or high-imageability).

**Design & Analysis** We ran stepwise regression analyses with either concreteness or imageability rating as the dependent variable, and ratings of auditory, gustatory, haptic, olfactory and visual strength as competing predictors. For analysis of consistency across the scales, each dependent variable was split at its midpoint before

Table 1: Sample words, used in Experiments 1 and 2, for which perceptual strength ratings [0-5] match or mismatch ratings of concreteness and imageability [100-700].

Word	Perceptual strength					Concreteness	Imageability
	Auditory	Gustatory	Haptic	Olfactory	Visual		
soap	0.35	1.29	4.12	4.00	4.06	589	600
noisy	4.95	0.05	0.29	0.05	1.67	293	138
atom	1.00	0.63	0.94	0.50	1.38	481	499
republic	0.53	0.67	0.27	0.07	1.79	376	356

regression: concreteness ratings formed abstract (rating [100-400],  $N = 294$ ) and concrete ([401-700],  $N = 298$ ) groups, whereas imageability ratings formed low ([100-400],  $N = 167$ ) and high ([401-700],  $N = 425$ ) groups. A priori sensitivity analysis confirmed that the sample size of the smallest group (low-imageability words) was still large enough to capture even a low degree of fit (minimum  $R^2 = .074$ ) in a five-predictor regression model at power of 0.8.

## Results & Discussion

**Concreteness** Analysis showed clear dissociations between concreteness and modality-specific perceptual experience, with little consistency across abstract and concrete groups (see Figure 1). Abstract words' ratings were predicted by three of the five modalities,  $F(3, 290) = 8.64$ ,  $p < .0001$ ,  $R^2 = .082$ , but with a low degree of fit and inconsistency in the direction of the relationship: positively related to vision, and negatively to auditory and olfactory strength. In contrast, concrete words' ratings were predicted positively by olfactory and visual strength,  $F(2, 295) = 33.52$ ,  $p < .0001$ ,  $R^2 = .185$ , but these two perceptual modalities offered a higher degree of fit than the model for abstract words.

Most perceptual modalities therefore failed to retain a consistent relationship with concreteness across the scale (auditory, olfactory) or had no predictive value at all (gustatory, haptic). However, the most serious conflict concerned the inversion of the olfactory effect: more olfactory meant more abstract, but more olfactory also meant more concrete. Such inconsistency in behavior poses serious problems for the assumption that abstractness and concreteness represent two ends of the same continuum, and rather indicates that participants applied different decision criteria at the concrete and abstract ends of the norming scale. While perceptual strength can explain more than twice the concreteness variance in concrete words (19%) than it did in abstract words (8%), participants are clearly basing their concreteness rating on non-perceptual information. It therefore appears that participants in concreteness norming studies treated the scale as two intersecting continua, neither of which reliably reflects the extent of sensory experience.

**Imageability** Analysis of imageability showed a clear visual bias at the expense of other perceptual modalities (see Figure 2). Ratings of low-imageability words were predicted by two perceptual modalities: visual strength (positively) and olfactory strength (negatively),  $F(2, 164) = 16.42$ ,  $p < .0001$ ,  $R^2 = .167$ . High-imageability ratings, on the other hand, were related to three perceptual modalities with a similar degree of fit,  $F(3, 421) = 36.32$ ,  $p < .0001$ ,  $R^2 = .206$ : positively for both visual and olfactory information, and negatively for gustatory.

Participants therefore tend to rely on visual experience when generating imageability ratings: visible things are highly imageable and invisible things are not. However, this focus on vision led other modalities to be neglected or misinterpreted. Neither auditory nor haptic experience was reflected at either end of the scale, and people tended to

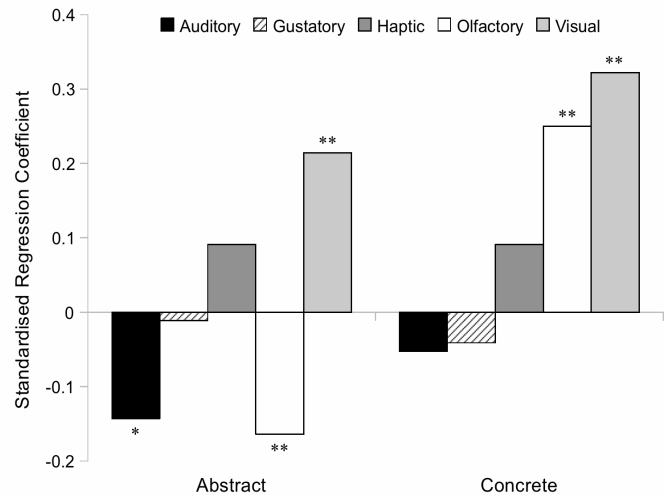


Figure 1: Modality predictors of concreteness ratings in Experiment 1 (\*  $p < .05$ , \*\*  $p < .01$ ).

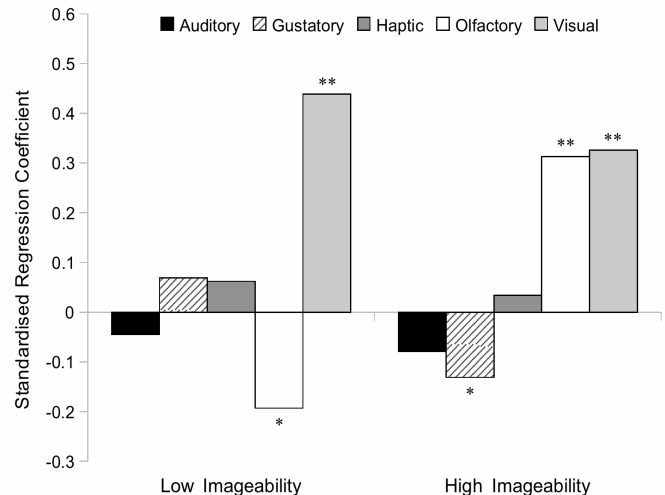


Figure 2: Modality predictors of imageability ratings in Experiment 1 (\*  $p < .05$ , \*\*  $p < .01$ ).

misconstrue olfactory information and ignore gustatory information for low-imageability concepts, yet follow olfactory strength while misinterpreting gustatory strength for high-imageability concepts. Results indicate that people do not find it equally easy to generate imagery across the range of modalities that constitute perceptual experience. Participants in imageability norming studies seem to have had difficulty in extending the meaning of “image” beyond its conventional interpretation as a visual impression.

## Experiment 2

Since neither concreteness nor imageability ratings reflect the full range of sensory experience, it raises the question of whether textbook concreteness effects in word processing are actually due to (a) the degree of perceptual information in each referent concept's representation, or (b) some other conceptually meaningful information that makes up most of



the variance in concreteness and imageability ratings. The present experiment aimed to resolve this question by comparing the unique predictive abilities of concreteness ratings, imageability ratings and perceptual strength in lexical decision and word naming performance.

If concreteness effects are due to the degree of perceptual information in each referent concept's representation, then perceptual strength should outperform concreteness and imageability in predicting latency and accuracy in word processing. In this case, we would expect perceptual strength to exhibit an independent effect in the presence of concreteness / imageability predictors, but not vice versa, because most of the variance in concreteness and imageability ratings reflects decision criteria that are unrelated to processing performance. On the other hand, if concreteness effects are actually due to some other non-perceptual representation that is captured by concreteness and imageability ratings, then they would maintain an independent effect even when perceptual strength has already been partialled out. In this sense, concreteness effects would subsume perceptibility effects, because the variance in concreteness and imageability ratings would reflect conceptually meaningful information in addition to perceptual differences.

## Method

**Materials** The same set of 592 words from Experiment 1 was used in this study, along with lexical decision and naming data from the Elexicon database (Balota et al., 2007; available online at <http://elexicon.wustl.edu>), which also provided lexical characteristics for each word to act as independent regression variables (see below).

**Design & Analysis** Hierarchical regression analyses determined the proportion of variance each candidate rating could explain. The dependent variables were mean lexical decision and naming times for each word ( $M = 633$  ms,  $SD = 64$  ms;  $M = 622$  ms,  $SD = 53$  ms; respectively), and their accompanying mean accuracy rates ( $M = 96.6\%$ ,  $SD = 5.4\%$ ;  $M = 98.9\%$ ,  $SD = 3.3\%$ ). As well as raw RT in ms, we also analysed standardized RT based on the mean z-scores of the original participants in the Elexicon data, which offers a more reliable measure of latency by partialing out individual differences in overall speed and variability (Balota et al., 2007). As independent variables in all regressions, we used a basic model found by Brysbaert and New (2009) to provide the best fit for RT and accuracy: log contextual diversity,  $\log^2$  contextual diversity, number of letters in the word, and number of syllables in the word.

The independent (unique) effects of concreteness ( $M = 427$ ,  $SD = 107$ ) and imageability ( $M = 461$ ,  $SD = 92$ ) were ascertained by adding the relevant predictor to a model containing maximum perceptual strength<sup>1</sup> and examining

<sup>1</sup> Maximum perceptual strength represents the highest rating in the concept's dominant modality, which analysis showed was the best method of compressing the five-value vector of perceptual strength into a single variable (necessary for equitable comparison with concreteness and imageability).

whether it led to an increase in fit. The independent effect of perceptual strength ( $M = 3.78$ ,  $SD = 0.75$ ) was calculated twice: once by entering it in a model that already contained concreteness, and once by adding it to a model that contained imageability. The correlation between imageability and concreteness was high,  $r(590) = .828$ ,  $p < .0001$ , and comparable to previous studies (e.g.,  $r = .83$  in Paivio et al., 1968). Maximum perceptual strength had a much weaker relationship with both concreteness,  $r(590) = .429$ ,  $p < .0001$ , and imageability,  $r(590) = .502$ ,  $p < .0001$ .

## Results & Discussion

Only perceptual strength emerged as a unique predictor of variance in word processing (see Figure 3 and Table 2). When either concreteness or imageability had already been included in the model, perceptual strength still accounted for an extra proportion of variance in all measures except naming accuracy. Critically, the inverse was not true. When maximum perceptual strength was included as a predictor, there was no model where the addition of concreteness or imageability produced an increase in  $\text{fit}^2$ . It is important to note that perceptual strength was in all cases acting in the expected direction (see Table 2): RT decreased and accuracy increased with higher perceptual strength. In other words, the independent predictive ability of perceptual strength never counteracted any facilitation by concreteness or imageability. Maximum perceptual strength thus captures meaningful information about conceptual structure that other ratings do not, and this information impacts directly on word processing performance.

One other striking difference emerged. Previous research has found that contextual diversity is inversely correlated with concreteness and imageability (i.e., abstract words appear in more diverse contexts than do concrete words: Schwanenflugel & Shoben, 1983; Schwanenflugel et al., 1988). Our data was consistent with this established pattern: zero-order correlations showed that concreteness was negatively related to log contextual diversity,  $r(590) = -.108$ ,  $p = .009$ , though the weaker trend for imageability was not significant,  $r(590) = -.024$ ,  $p = .560$ . Yet, in sharp contrast, perceptual strength was *positively* correlated with contextual diversity,  $r(590) = .117$ ,  $p < .0001$ . That is, although concrete-rated words have a narrower variety of contexts than abstract-rated words, perceptually strong words have a *wider* variety than perceptually weak words. We return to this issue in the general discussion.

## General Discussion

In the present paper, we show that concreteness and

<sup>2</sup> Our approach and findings thus differ considerably from those of Juhasz et al., (2011), who found that a sensory experience rating predicted lexical decision times after imageability was partialled out. However, they did not examine concreteness ratings or word naming times, their rating asked people to aggregate all sensory experience on a single scale (which our Experiment 1 indicates that people find very difficult) rather than collecting ratings on separate modalities, and they did not show which of imageability and their own rating had better predictive power.



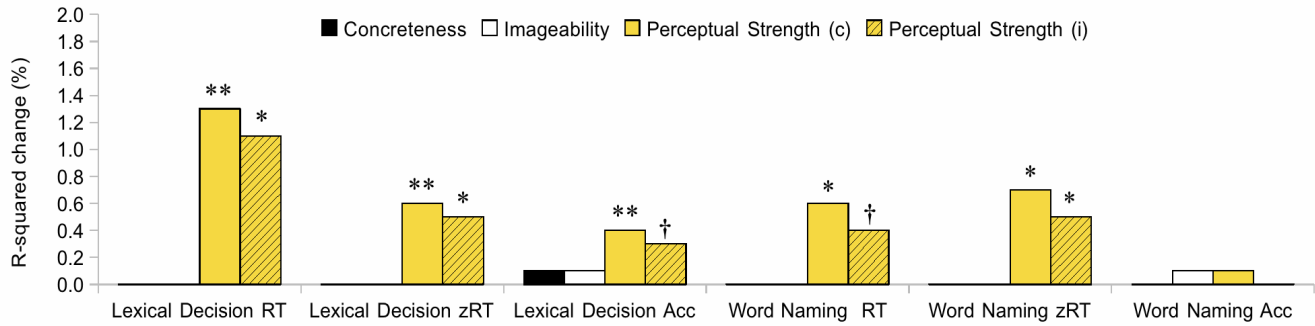


Figure 3: Independent (unique) effects of each predictor in Experiment 2, showing proportion of explained variance ( $R^2$  change in %) of Elexicon reaction time and accuracy data, over and above that of perceptual strength, concreteness (c) or imageability (i) (†  $p < .1$ , \*  $p < .05$ , \*\*  $p < .01$ ).

Table 2: Standardized regression coefficients for independent effects of each predictor over and above that of perceptual strength, concreteness (c) or imageability (i), in models of Elexicon reaction time and accuracy data in Experiment 2.

Predictor	Lexical Decision RT	Lexical Decision zRT	Lexical Decision Acc	Word Naming RT	Word Naming zRT	Word Naming Acc
Concreteness	−0.016	−0.004	−0.031	+0.025	+0.002	−0.023
Imageability	−0.023	−0.018	+0.036	−0.013	−0.024	+0.045
Perceptual strength (c)	−0.132**	−0.091**	+0.097**	−0.089*	−0.095*	+0.038
Perceptual strength (i)	−0.127**	−0.080*	+0.067†	−0.072†	−0.082*	+0.006

†  $p < .1$ , \*  $p < .05$ , \*\*  $p < .01$

imageability ratings do not accurately reflect the perceptual basis of concepts, and that concreteness effects in lexical decision and naming are better predicted by perceptual strength ratings than by concreteness or imageability ratings. These findings support the intuition that perceptual concepts are faster to process, and show that textbook concreteness effects in word processing are actually a function of the degree of perceptual information in each referent concept's representation. However, our results also suggest that concreteness and imageability ratings are unsuitable for the tasks in which they are employed, because most of their variance comes from non-perceptual decision criteria that is unrelated to word processing performance. Concreteness effects could therefore be better characterized as perceptibility effects, which can be sometimes nullified or inverted (e.g., Kousta et al., 2011; Papagno et al., 2009) when elicited from relatively noisy concreteness or imageability ratings.

While the connection between concreteness effects and perceptual information might at first glance seem like old news (e.g., Barsalou & Wiemer-Hastings, 2005; Paivio, 1986, 2007), the present findings have some important ramifications for how such effects should be interpreted. Concreteness effects, by their very name, are assumed to result from an ontological difference between concrete and abstract concepts carrying through to a representational difference that affects speed and accuracy of processing. Labeling a word as “concrete” or “abstract” has an intuitive appeal, but we would argue that these terms lacked proper operationalization during norming and hence it is unclear exactly what information is captured by concreteness

ratings. Of course, any set of ratings can only ever be an approximation of an underlying representation, and we are not suggesting that one should expect a perfect fit between concreteness ratings and behavioral effects. That said, the poor performance of concreteness ratings in the current data lies in sharp contrast to the robust performance of perceptual strength ratings. We suggest that the concrete / abstract ontological distinction must be disentangled from concreteness / imageability norms because empirical concreteness effects are not in themselves well predicted by concreteness / imageability.

Theoretically, the present results poses some problems for dual coding, context availability, and situated simulation explanations of concreteness effects. It is a central tenet of dual coding theory that highly perceptual concepts are those with the most direct connections between the verbal and nonverbal imagery codes, and people therefore find it difficult to generate perceptual imagery for words that lack these direct connections (Paivio, 1986, 2007). However, imageability (i.e., the ease of consciously generating imagery) is not well related to perceptual experience (Experiment 1), and its effects were entirely subsumed by larger effects of perceptual strength (Experiment 2). In other words, it is the extent of perceptual information in a concept's representation that matters to word processing, not the ease of generating imagery, which casts some doubt on the idea that processing delays for abstract concepts emerge from their lack of direct inter-system connections. Both context availability (Schwanenflugel & Shoben, 1983; Schwanenflugel et al., 1988) and situated simulation (Barsalou & Wiemer-Hastings, 2005; Barsalou et al., 2008)

views share the idea that abstract concepts are slowed in processing because they have a wider variety of potential situational contexts. This idea, however, is not borne out by our data. Strongly perceptual concepts (i.e., those that are generally assumed to be concrete, regardless of what concreteness ratings say) actually have greater contextual diversity than weakly perceptual concepts.

In sum, we believe that the operationalisation of abstract and concrete concepts deserves much closer scrutiny than it has received to date. Whether researchers want to investigate the ontological distinction between abstract and concrete concepts, or the variables that affect latency and accuracy in word processing, then they should reconsider the automatic tendency to reach for concreteness and imageability ratings that have little to do with the perceptual basis of concepts. Strength of perceptual experience has a powerful bearing on how people represent concepts during word processing, and these perceptibility effects are stronger than those elicited by concreteness or imageability.

## References

- Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17, 814-823.
- Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445-459.
- Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. A. Zwaan, *Grounding cognition: The role of perception and action in memory, language, and thinking* (pp. 129-163). Cambridge, UK: Cambridge University Press.
- Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, & A. C. Graesser, A. (Eds.). *Symbols, embodiment, and meaning* (pp. 245-283). Oxford, UK: Oxford University Press.
- Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract words. *Journal of Cognitive Neuroscience*, 17, 905-917.
- Brysbaert, M., & New, B. (2009). Moving beyond Ku era and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41, 977-990.
- Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods*, 36, 371-383.
- Connell, L., & Lynott, D. (2010). Look but don't touch: tactile disadvantage in processing modality-specific words. *Cognition*, 115, 1-9.
- Fiebach, C. J., & Friederici, A. D. (2004). Processing concrete words: fMRI evidence against a specific right-hemisphere involvement. *Neuropsychologia*, 42, 62-70.
- James, C. T. (1975). The role of semantic information in lexical decisions. *Journal of Experimental Psychology: Human Perception and Performance*, 1, 130-136.
- Juhász, B. J., Yap, M. J., Dicke, J., Taylor, S. C., & Gullick, M. M. (2011). Tangible words are recognized faster: The grounding of meaning in sensory and perceptual systems. *Quarterly Journal of Experimental Psychology*, 64, 1683-1691.
- Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, 140, 14-34.
- Kroll, J., & Merves, J. S. (1986). Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12, 92-107.
- Louwerse, M. M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, 35, 381-398.
- Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, 41, 558-564.
- Lynott, D., & Connell, L. (in prep.). Why noun concepts are more multimodal than adjective concepts: Modality exclusivity norms for objects. *Manuscript in preparation*.
- Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida word association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36, 402-407.
- Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford, UK: Oxford University Press.
- Paivio, A. (2007). *Mind and its evolution: A dual coding theoretical approach*. Mahwah, NJ: Erlbaum.
- Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76(1, Pt. 2), 1-25.
- Papagno, C., Capasso, R., Zerboni, H., & Miceli, G. (2007). A reverse concreteness effect in a subject with semantic dementia. *Brain and Language*, 103, 90-91.
- Sabsevitz, D. S., Medler, D. A., Seidenberg, M., & Binder, J. R. (2005). Modulation of the semantic system by word imageability. *Neuroimage*, 27, 188-200.
- Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, 27, 499-520.
- Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9, 82-102.
- Schwanenflugel, P. J., & Stowe, R. W. (1989). Context availability and the processing of abstract and concrete words in sentences. *Reading Research Quarterly*, 24, 114-126.
- Tyler, L. K., Russell, R., Fadili, J., & Moss, H. E. (2001). The neural representation of nouns and verbs: PET studies. *Brain*, 124, 1619-1634.

# Learning of Relational Categories as a Function of Higher-order Structure

Daniel Corral (daniel.corral@colorado.edu) & Matt Jones (matt.jones@colorado.edu)

Department of Cognitive Psychology, University Colorado Boulder,  
Boulder, CO 80309 USA

## Abstract

Higher-order relations are important for various cognitive tasks, such as analogical transfer. The current study tested people's ability to learn new relational categories, using a learning test of pure higher-order relations. Each stimulus consisted of 4 objects varying on 3 dimensions. Each category was defined by three binary relations between pairs of objects, producing six logically different conditions. Every category was composed of the same number of relations, but differed in the manner that the relations were linked (i.e., by operating on shared objects). Various learning models were compared and the significance of their performance on the experimental task is discussed. The current findings may advance understanding of the cognitive mechanisms involved in relational learning and the manner in which people naturally represent higher-order relational structures.

**Keywords:** higher-order relations; schema refinement; schema elaboration; structure acquisition.

## Introduction

The ability to generalize and transfer knowledge from a given problem to an analogous task has been of great interest to cognitive scientists and has led to an extensive amount of research. The large body of work on analogical transfer has converged on the idea that transfer is driven by discovering the common relational structure between two analogous scenarios (Gentner, 1983; Gick & Holyoak, 1983). Penn, Holyoak, and Povinelli (2008) posit higher-order relations are critical for most other higher cognitive processes as well, including inference, causal reasoning, and theory of mind. Nevertheless, there is little understanding of the cognitive mechanisms that subserve learning and recognition of higher-order relations.

The purpose of the current study is to explore how people learn different higher-order relations. We define a *higher-order relation* to be a system of first-order (i.e., primitive) relations operating on a common set of objects. Different higher-order relations differ in how the first-order relations are linked together by shared role-fillers. We report an experiment using a relational category-learning task, in which each subject learned a category defined by a higher-order relation, by learning to distinguish category members from non-members. The category in each experimental condition was defined by three binary relations among four objects. In the spirit of Shepard, Hovland, & Jenkins' (1961) classic study on learning feature-based categories, we conduct an exhaustive comparison of the six logically different categories of this type.

The dominant view of how people acquire relational concepts is schema refinement (e.g., Doumas, Hummel, & Sandhofer, 2008), but the present results highlight a number

of conceptual problems with this approach. As an alternative, we introduce schema elaboration as a mechanism that is more psychologically plausible and better able to match human performance. We consider four variants of schema elaboration, motivated by different theoretical perspectives, and compare their ability to predict the relative learnability of different higher-order relations.

## Structure-Mapping Theory

Since its initial proposal (Gentner, 1983), structure-mapping theory has provided a great deal of insight into the process of analogical learning and transfer. Structure-mapping theory posits that analogy involves aligning the relational structures of two scenarios. A relational structure is composed of multiple first-order relations that are linked together in a specific manner (i.e., the manner in which they operate on shared objects). Consider the classic solar system-atom analogy (Figure 1): Planets revolve around the sun, and planets are smaller than the sun; electrons revolve around the nucleus, and electrons are smaller than the nucleus. Although the same first-order relations are present in both scenarios (i.e., *smaller than* and *revolves around*), the analogy only works because the first-order relations share objects in the same way, such that their first roles are filled by the same object (i.e., planet and electron). In structure-mapping theory, this property is formally known as *parallel connectivity* (Gentner, 1983).

Thus, analogy can be viewed as the recognition that two scenarios are instances of the same higher-order relation, that is, first-order relations connected in the same manner. When the same first-order relations are present in two or more scenarios, but are shared differently between objects, different higher-order relations are formed. Hence, to successfully transfer between two analogous scenarios, people must learn the exact manner in which the first-order relations are connected to form a specific higher-order relation.

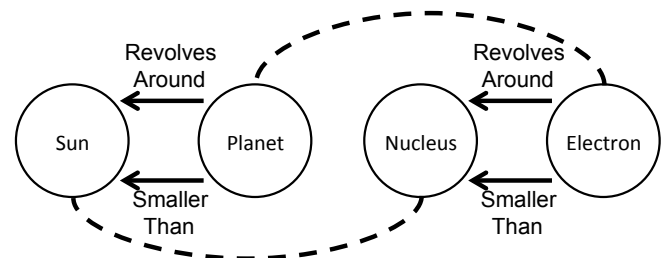


Figure 1. Diagram of solar-system-atom analogy.

## Schema Refinement

Formation of an analogy has been proposed to lead to induction of a schema, an abstract representation that captures the relational structure common to both analogues (Gick & Holyoak, 1983; Hummel & Holyoak, 2003; Kuehne, Forbus, Gentner, & Quinn, 2000). Subsequent analogy between a schema and a new episode can result in a new schema (replacing or supplementing the original schema) that contains only the structure that is common to the original schema and the new episode. This process is referred to as *schema refinement*, and it has been proposed to operate by a mechanism of *intersection discovery* (Doumas et al., 2008). An analogy between two episodes may lead to a “dirty” schema that includes idiosyncratic properties common to both episodes but not universal to other instances of the abstract concept being acquired (Doumas et al., 2008; Hummel & Holyoak, 2003). Through comparison to successive instances of the abstract concept (as they are encountered), the schema can be refined to contain only information that belongs to the concept.

As a model of relational learning, schema refinement has several shortcomings. First, because refinement models only allow for a schema to decrease in size, the model cannot add new information. Consequently, upon its first encounter with a member of a relational category, the model must retain all information contained in the exemplar, as it may be necessary in defining the category. Contrary to the results presented below, this assumption leads to a prediction of no false alarms during learning of a relational category. A false alarm can only occur when a subject’s current schema is missing relational constraints for what constitutes category membership. Under an idealized model of pure schema refinement, there is no way for the model to delete relations that are present in all category members.

A related assumption of schema refinement is that the model can start off with and maintain highly complex schemas. Given the processing constraints of working memory (Baddeley, 2003), such an assumption seems psychologically implausible. Instead, subjects should be expected to quickly forget a large amount of the information that was initially processed. When the number of objects and predicates contained within a higher-order relation exceeds the processing capacity of working memory, schema refinement may not accurately reflect how the concept is acquired. Thus, we propose that a more complete model of relational learning must incorporate forgetting, and, consequently, the ability to add information to the schema rather than only simplifying it.

## Schema Elaboration

When a learning model contains processing constraints similar to those of working memory, the ability to elaborate upon a schema (i.e., to add new information) may be better suited than schema refinement alone for the acquisition of higher-order relations. We propose that in cases where a schema is insufficiently complex (i.e., is missing appropriate relational constraints), people are capable of

updating their schema by adding new relations. We refer to this process as *schema elaboration*.

Because the pure schema refinement model is unable to add new information, the model has no room for error if true relational constraints are mistakenly discarded. This makes schema refinement an unrealistically rigid learning model. The schema elaboration model described in detail below is more flexible, as it is able to reincorporate information that it has mistakenly discarded or forgotten.

## Experiment

The current study investigated people’s ability to learn arbitrary new higher-order relations. The study used a standard category-learning paradigm, with an A/not-A design in which subjects were asked to decide whether each stimulus did or did not belong to the category. The category to be learned was manipulated between subjects. The category structures all contained the same number and types of first-order relations but differed in how those relations were connected (i.e., in the higher-order relation they formed). We aimed to test models of relational concept acquisition by assessing how this manipulation of higher-order structure affects learning.

Figure 2 shows an example stimulus. Each stimulus comprised four objects, known in the literature as *Shepard circles*, arranged in a square configuration. The objects varied along three separable dimensions: brightness, size, and radius tilt. Each dimension had four levels, assigned without replacement to the four objects on each trial.

Each dimension defines a comparative binary relation among the objects (i.e., brighter, larger, steeper). The category to be learned by each subject was defined by three such relations, one on each dimension. Thus, a stimulus was a member of the category if it satisfied all three of these relations (e.g., upper-right object must be larger than upper-left object, lower-left object must be brighter than lower-right object, and lower-right object must have its radius more tilted than upper-left object). The category structures varied in how the relations were connected to each other, in terms of the objects they were defined on. For example, any two relations could operate on the same pair of objects, on disjoint objects, or on one shared object with one unique object for each relation. This design leads to six topologically unique category structures, shown schematically in Figure 3. The manner in which these topological structures were instantiated (i.e., the roles of the four spatial locations) was counterbalanced across subjects within each condition.

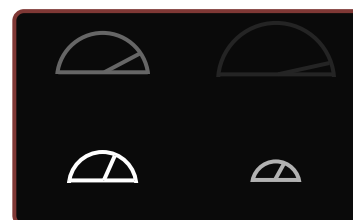


Figure 2. Example stimulus from main task.

## Method

137 undergraduates were randomly assigned to six (between subjects) conditions, differing in the category structure to be learned. Subjects were given a cover story in which the stimuli were optical key cards for a building; their task was to learn which key cards would open a door.

To familiarize subjects with each of the first-order relations, they were given three training tasks prior to the main task, one for each first-order relation (i.e., brighter, larger, and steeper). The training tasks were the same as the main task, except that each stimulus contained only two objects instead of four, and each category was defined by only one relation (e.g., right object must be darker than left object). Each training task ended once the subject answered eight consecutive correct responses. The order of training tasks was counterbalanced.

All four tasks followed the same procedure. On each trial, a stimulus was sampled randomly, subject to equal probability of choosing a stimulus in or out of the category. The subject responded by pressing Y or N (indicating the key card does or does not open the door), and then the correct answer was displayed. The instructions for each task indicated the categories were different (i.e., each task was about a different door of the building) and included a random, positive example (i.e., a key card that opens the current door). The full experiment (i.e., training and main tasks combined) was programed to end after 55 minutes.



Figure 3. Diagram of each category condition. Lines connecting objects indicate first-order relations.

## Models

Before presenting the results, we describe a series of models that were compared to the data. These models were designed to test the need for augmenting theories of schema refinement with mechanisms of forgetting and schema elaboration. Only the main task was modeled.

### Control Models

Three control models—pure refinement, refinement with forgetting, and refinement with forgetting and elaboration—were formulated to provide a baseline for the more sophisticated elaboration models discussed below.

All of the models operate by maintaining a schema from trial to trial that contains some set of relations among the objects within the stimuli. Each stimulus is classified as in the category if it satisfies all relations currently in the

schema. The schema is initialized as a complete representation (i.e., all 18 binary relations) of the example stimulus provided in the instructions for the main task.

The pure refinement (PR) model learns only following a miss, meaning a trial on which the stimulus belongs in the category but is mistakenly classified as a nonmember. This can occur when the schema includes relational constraints that are not part of the true category rule. Feedback after a miss causes the schema to be updated (refined) by intersection discovery, discarding all relational constraints the stimulus violates. All other relations in the schema are retained. This learning process will continue until all incorrect relational constraints are removed, at which point the schema will necessarily coincide with the true category rule.

The refinement-with-forgetting (RF) model incorporates processing constraints meant to mimic those of working memory. Soft capacity limitations result in the model losing (i.e., forgetting) relational constraints prior to each trial. Each relation has an independent probability  $p$  of being forgotten, which depends on the total number of relations currently in the schema ( $r$ ):

$$p = 1 - \frac{L}{r} (1 - e^{-r/L}), \quad (1)$$

where  $L$  is a processing-capacity parameter. This formulation has the property that the expected number of retained relations equals  $L * (1 - \exp(-r/L))$ , that is, exponential approach to some limiting capacity  $L$ .

The random elaboration model (RE) includes refinement, forgetting, and elaboration. The interplay between refinement and elaboration leads the model to add and remove constraints one at a time until the schema converges on the true category structure. The forgetting mechanism in the model allows for false alarms, as true relational constraints can be lost, making the schema under-constrained. Indeed, a subject may commit a false alarm if his or her working schema lacks a relational constraint that is part of the category rule.

According to the elaboration assumption, false alarms lead to the appendage of a new relation, which the stimulus satisfies but was not part of the initial hypothesis. This mechanism allows the schema to increase in size and complexity. Unlike with misses, after receiving feedback of a false alarm the subject does not know which relational constraints must be added (i.e., which relation in the stimulus constitutes a violation of the rule). Therefore, the model identifies all relations in the stimulus that are absent from the schema, and treats each as a candidate to be added. For simplicity, we assume exactly one relation is added to the schema following any false alarm. In the RE model, this choice is made at random among the candidates.

The PR model is an ideal observer for the present task, and hence performance was expected to be high for all conditions. In the RF model, once a true relational constraint is forgotten, it has no way of being reincorporated into the schema; hence all relations will eventually be lost and the

model should asymptote chance performance. In the RE model, elaboration and forgetting can combine to produce intermediate levels of performance (depending on the capacity parameter  $L$ ). However, it was expected that none of the three control models would predict any learning differences among the different category conditions. Indeed, all relations are treated independently (except for a global effect of schema size on the forgetting probability), and hence the manner in which relations are linked through operating on shared objects should have no effect on model performance.

### The Search for Relational Constraints

The RE model assumes schema elaboration involves random selection of a candidate relation that is in the stimulus but not in the schema. Alternatively, the selection could be preferential, sensitive to higher-order structure. Different assumptions about preferences guiding the relations that are added lead to different models of preferential elaboration, each making unique predictions about the relative learnability of the category structures in the current study. Here we consider four possibilities, motivated by different theoretical perspectives in the literature. Importantly, each model is inspired by the corresponding theoretical perspective but is not meant to be a formal implementation of that theory.

Each of the four models presented below works in the following manner. Following a false alarm, each candidate relation for addition to the schema is assigned a score that determines its probability of being selected. The probability each candidate is selected is given by

$$\frac{e^{\phi s}}{\sum_{s'} e^{\phi s'}} \quad (2)$$

where  $s$  is the score for the candidate,  $s'$  ranges over all candidates, and  $\phi$  is a parameter determining the degree of stochasticity in the decision process. The models differ in how the scores are determined by higher-order structure.

### Conceptual Coherence

Murphy and Medin (1985) proposed that people's lay theories about the world make categories conceptually coherent. One reason for this may be that a theory provides a conceptual filter through which relational information can be processed and organized around. Furthermore, research has shown that features that are central to a concept more strongly influence the concept's conceptual coherence (Sloman, Love, & Ahn, 1998). Taken together, these ideas suggest that a category composed of a central object that participates in all three relations will be more conceptually coherent than other category structures, as the category representation can be organized around the central object, providing a critical conceptual reference point. Thus, performance should be higher for conditions 2, 4, and 6 than for other category structures. To formalize this principle, the score ( $s$ ) for each candidate relation was defined as the sum,

over the two objects that relation operates on, of how many relations already in the schema each object participates in. This assumption leads the conceptual coherence (CC) model to prefer relations built on already more central objects, thus favoring categories with a centralized structure.

### Economy of Objects

Due to the processing constraints of working memory (Baddeley, 2003), it may be easier to discover an analogy that requires mapping fewer objects between scenarios. Therefore, when elaborating a schema, subjects may be inclined to select relations that minimize the total number of objects involved. This hypothesis predicts that learning will be superior for category structures involving a smaller number of objects, such as condition 6 and to a lesser extent 3 and 4. This principle was formalized in the economy of objects (EO) model, by defining the score for each candidate relation as the number of its objects (0, 1, or 2) that participate in other relations already in the schema.

### Plurality of Objects

In contrast to the EO model, cognitive load theory (van Merriënboer & Sweller, 2005) suggests that objects that do not participate in any of the category's relations will act as extraneous distractors. Therefore, subjects' attention may be drawn to such objects, making them more likely to add relations on new objects when elaborating the schema. Because people often struggle to recognize surface features as irrelevant information (Cooper & Sweller, 1987), irrelevant objects may place an unnecessary amount of strain on working memory, while concurrently obscuring the category's higher-order structure. Consequently, category structures that contain the greatest number of irrelevant objects (condition 6, followed by 3 & 4) would be most difficult to learn. This principle was formalized in the plurality of objects (PO) model, by defining the score for each candidate relation as the number of its objects (0, 1, or 2) that do not participate in any relations currently in the schema. This scoring rule is opposite that used in the EO model, and the two models are equivalent under a substitution  $\phi \rightarrow -\phi$ .

### Relational Chaining

Lastly, learning may be better for category structures composed of relations that are chained together (e.g., condition 1), as people may be intuitively inclined to link known relational structure to new objects. Such a preference might arise from a causal learning perspective, in which subjects seek to discover causal chains among the objects. For example, upon learning that object A must be bigger than object B, people may be inclined to test whether object B must be brighter than object C. Thus, structures composed of relations that can be more readily chained together may be easier to acquire. This principle was formalized in the relational chaining (RC) model, by defining the score for each candidate as



$$s = \frac{1}{c_{\min} + 2} + \frac{\frac{1}{c_{\min} + 1} - \frac{1}{c_{\min} + 2}}{c_{\max} - c_{\min} + 1}, \quad (3)$$

where  $c_{\min}$  and  $c_{\max}$  are the counts of relations currently in the schema in which the candidate's two objects participate. This rule was designed to implement a lexicographic preference for small values of  $c_{\min}$  followed by small values of  $c_{\max} - c_{\min}$ . As a special case, the score for relations with  $c_{\min} = c_{\max} = 0$  was set to zero, to implement a preference not to add isolated relations. Thus, the ideal candidate is one that extends an existing chain:  $c_{\min} = 0$ ,  $c_{\max} = 1$ .

### Summary of Models

The preceding subsections fully specify the models tested. The categorization response is determined by whether the stimulus satisfies all relations currently in the schema. Refinement follows misses, by intersecting the schema with the stimulus. Forgetting precedes each trial, following Equation 1. Elaboration follows a false alarm, adding a single relation from the stimulus, chosen by Equation 2 (or randomly, in the RE model). The models differ in whether they include forgetting and elaboration, and in the preferences guiding elaboration.

### Experiment and Model Results

Subjects varied in how much time they took to learn the training tasks, and hence in how much time they had for the main task. To reduce statistical noise and ensure all subjects had enough time to learn their condition's category structure, a selection criterion was used to exclude subjects who spent over 35 minutes on the training tasks (leaving less than 20 minutes for the main task). Because this criterion is based on events prior to the experimental manipulation, it introduces no bias in estimating differences among conditions. The selection left 104 subjects in the analysis. Of these subjects, the fewest number of trials completed on the main task was 245.

An ANOVA comparing average proportion correct on the first 245 trials across conditions revealed a non-significant trend,  $F(5, 98) = 1.83$ ,  $p = .114$ ,  $MSE = .004$ . Because of the complexity of the main task, 245 trials may be insufficient for learning. Therefore, we repeated the analysis while excluding the 6 additional subjects who had completed the fewest trials. The remaining 98 subjects all completed over 350 trials. An ANOVA on these subjects' performance on the first 350 trials indicates a significant main effect of category structure,  $F(5, 92) = 2.76$ ,  $p = .023$ ,  $MSE = .012$ . Figure 4 shows the mean performances by condition, compared to model predictions. Importantly, the ordering among conditions remained unchanged from the initial analysis, and means were nearly unchanged. Finally, the 6 excluded subjects were re-included, with their proportions correct defined based on the number of trials actually completed. The analysis again revealed a significant effect of condition,  $F(5, 98) = 2.75$ ,  $p = .023$ ,  $MSE = .012$ . Again, the ordering of performance between

conditions was unchanged, and condition means were nearly identical to the previous analyses.

Figure 4 shows the behavioral results and the simulated predictions of all models. Model parameters (as applicable) were the same for all models and were chosen by hand, with  $L = 9$  and  $\phi = 10$ .

Evaluation of the models in sequence provides support for each of our theoretical proposals (see Table 1). As predicted, the PR model far outperformed the subjects, suggesting the need for some sort of forgetting mechanism in addition to schema refinement. However, the RF model performed nearly at chance, because it eventually forgot all relations, suggesting a further need for some sort of schema elaboration mechanism. The RE model can match subjects' intermediate performance level, but it fails to predict differences among conditions. The last four models predict condition differences because they are sensitive to higher-order structure. However, the differences are all weaker than in the empirical data. The predicted condition differences are greater when  $L$  is increased to produce levels of performance (Figure 5, with  $L = 40$ ), but still none of the models reproduces the correct ordering among conditions. Therefore, further work is required to understand exactly how higher-order structure affects relational learning.

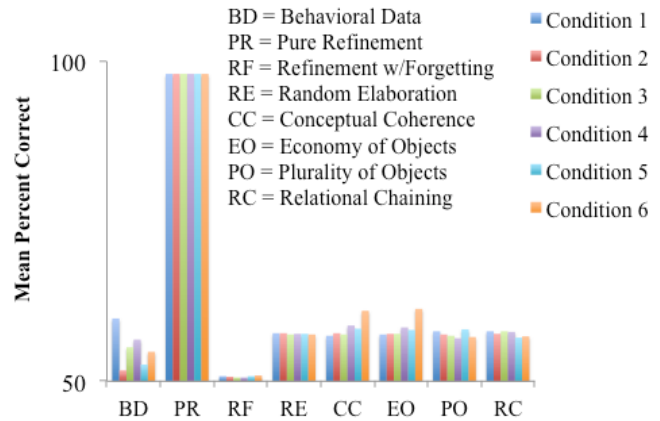


Figure 4. Mean performance for subjects and models.

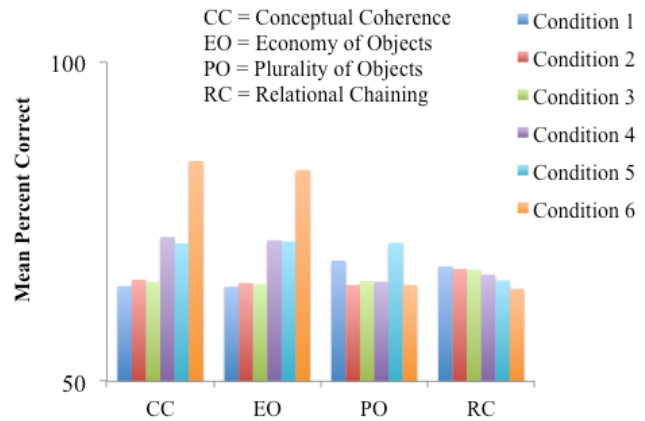


Figure 5. Predictions of structure-sensitive models at higher levels of performance, to accentuate condition differences.



Table 1. Strengths (+) and weaknesses (–) of the simulated models.

	Model						
	PR	RF	RE	CC	EO	PO	RC
Performs within the range of subject data	–	–	+	+	+	+	+
Predicts differences across conditions	–	–	–	+	+	+	+
Predicted differences match subjects	–	–	–	–	–	–	–

## Discussion

Although previous research has not directly addressed how readily people learn different types of higher-order relations, the acquisition of such concepts is integral to the development of expert representations (Chi, Feltovich, & Glaser, 1981). The current behavioral data suggest that acquisition of higher-order relations is indeed affected by the manner in which the elementary relations within a relational structure are connected. Subjects' performance was best for conditions in which relations could be chained together and where single objects participated in multiple relations (i.e., Conditions 1, 4, and 3).

Although schema refinement has been the dominant model of relational learning (e.g., Doumas et al., 2008), the PR model incorrectly predicts no learning differences across the different category conditions. Further, the model's performance differed dramatically from that of subjects. As expected, when processing constraints were introduced, the RF model failed to retain any of the relational constraints in the category, performing at chance in all conditions. Taken together, these results suggest that schema refinement alone is an insufficient explanation of human relational learning.

The predictions from the elaboration models allow us to address several important issues. That the elaboration models make predictions within the range of the subjects' performance supports the proposal that people employ elaboration mechanisms (in addition to schema refinement) when acquiring higher-order concepts. Additionally, the differences that were found across conditions in the behavioral data provide support for the idea that people indeed have preferences for seeking out certain types of higher-order relations, as formalized in the four structure-sensitive elaboration models.

However, the condition differences predicted by these models were weaker than those exhibited by subjects, and none of the models reproduces the correct ordering across conditions. Thus, it remains an open question as to the specific mechanisms that drive people's search for higher-order relations.

Understanding what drives the differences among the present experimental conditions may provide important theoretical insight into the mechanisms of relational learning, as well as the manner in which people acquire

more abstract, higher-order concepts. Such results may also have practical applicability for areas where the recognition of higher-order structures is important for deep learning, such as education, problem solving, and decision-making.

## Acknowledgements

This research was supported by AFOSR grant FA9550-10-1-0177.

## References

- Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews: Neuroscience*, 4, 829-839.
- Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152.
- Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, 79, 347-362.
- Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, 115, 1-43.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15, 1-38.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110, 220-264.
- Kuehne, S., Forbus, K., Gentner, D., & Quinn, B. (2000). SQL: Category learning as progressive abstraction using structure mapping. *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, 770-775.
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289-316.
- Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31, 109-178.
- Shepard, R. N., Hovland, C. I., & Jenkins, H.M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, 75, Whole No. 517.
- Sloman, S. A., Love, B. C., & Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, 22, 189-228.
- van Merriënboer, J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, 17, 147-177.

# Gaussian Process Regression for Trajectory Analysis

Gregory E. Cox (grcox@indiana.edu)

George Kachergis (gkacherg@indiana.edu)

Richard M. Shiffrin (shiffrin@indiana.edu)

Department of Psychological and Brain Sciences, Indiana University  
1101 E. Tenth St., Bloomington, IN 47405 USA

## Abstract

Cognitive scientists have begun collecting the trajectories of hand movements as participants make decisions in experiments. These response trajectories offer a fine-grained glimpse into ongoing cognitive processes. For example, difficult decisions show more hesitation and deflection from the optimal path than easy decisions. However, many summary statistics used for trajectories throw away much information, or are correlated and thus partially redundant. To alleviate these issues, we introduce Gaussian process regression for the purpose of modeling trajectory data collected in psychology experiments. Gaussian processes are a well-developed statistical model that can find parametric differences in trajectories and their derivatives (e.g., velocity and acceleration) rather than a summary statistic. We show how Gaussian process regression can be implemented hierarchically across conditions and subjects, and used to model the actual shape and covariance of the trajectories. Finally, we demonstrate how to construct a generative hierarchical Bayesian model of trajectories using Gaussian processes.

**Keywords:** Trajectory analysis; Gaussian processes; Bayesian statistics.

## Introduction

Cognitive scientists are gradually turning toward more fine-grained measures to gain more insight into the continuous nature of the cognitive processes that underly behavior. Perhaps the most widespread of these measures is eye tracking, in which we assume that where people gaze is the current focus of attention and processing. For example, when reading a syntactically ambiguous sentence, people tend to make eye movements back toward the function word or pronoun that best helps resolve the ambiguity (Frazier & Rayner, 1987). Or, when hearing continuous speech, people will tend to look more at objects whose names are consistent with a partially-heard word (e.g., people will look at either a “ball” or a “bear” if they have just heard the syllable “b”), indicating that people make continuous predictions about the content of speech based on partial information (Spivey, Grosjean, & Knoblich, 2005). Thus, a continuous measure of behavior, like eye tracking, appears to provide insight into ongoing cognitive processes.

More recently, researchers have begun to collect explicit continuous behavioral measures in the form of mouse or stylus movements (e.g., Freeman & Ambady, 2010). These may easily be used in place of any task that requires an explicit choice on the part of the participant, which includes most experimental paradigms in cognitive psychology. Rather than simply pressing a key to make their response, a participant can instead move their hand (as well as an attached mouse or stylus) toward the option of their choice before selecting

(clicking) it. Similar to eye tracking, the trajectories of these continuous motor movements provide a way of measuring the ongoing cognitive processes that lead to the participant’s final choice.

A major hurdle with any new measure is the need for appropriate analytical tools and statistical tests that allow researchers to draw inferences from trajectory data. Due to the richness of this data, many measures are possible and can lead to principled inferences (for an overview, see Freeman, Dale, & Farmer, 2011). When moving their hand while making a decision, people may deviate more from a straight trajectory if there is a tempting alternative, making viable such measures as maximum deviation, curvature area, and switches in direction.

In this paper, we introduce a new method for analyzing trajectory data. Our method is based on treating trajectories as a Gaussian process, for which there is much well-developed statistical theory. We begin by providing a brief overview of Gaussian process regression and show how it may be applied to motor response trajectories and—more fruitfully, we argue—their derivatives. Finally, we show how Gaussian process regression can be incorporated into a generative hierarchical Bayesian model of trajectories.

## Gaussian Process Regression

Gaussian process regression (GPR) is a statistical technique with a long history in spatial statistics, and more recently in function estimation and prediction (Griffiths, Lucas, Williams, & Kalish, 2009). The interested reader is directed to the excellent text on Gaussian processes by Rasmussen and Williams (2006).

## Gaussian Processes

A Gaussian process (GP) is simply a collection of random variables, all of which are jointly Gaussian distributed. What differentiates a Gaussian process from the more familiar multivariate Gaussian distribution is the fact that a Gaussian process may have an infinite index set, that is, it may specify an infinite number of jointly Gaussian variables. Thus, it is possible to define a Gaussian process over a continuous variable, like time. Just as a multivariate Gaussian distribution is defined entirely by its mean vector and covariance matrix, a Gaussian process is defined by its mean *function*  $m(x)$  and covariance kernel,  $k(x, x')$ , where  $x$  and  $x'$  are two (possibly multidimensional) values of some predictor variable  $\mathcal{X}$  (e.g., time). We denote the fact that a function  $f(x)$  is a Gaussian

process by

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')).$$

A Gaussian process can be considered a distribution over functions, with  $m(x)$  expressing the mean value of all of these functions at  $x$  and  $k(x, x')$  represented the expected covariance between the function value at  $x$  and that at  $x'$ , i.e., the amount of “information” that the function  $f(x')$  carries about the value at  $f(x)$  (and vice versa). Thus, if we encounter data (like trajectory data) for which we do not know or cannot guess the form of the function that generated it, we can *infer* the form of this function if we assume that it is a Gaussian process. This kind of inference is termed “Gaussian process regression”.

**Bayesian Inference with GPs** Gaussian process regression (GPR) seeks to model an unknown function  $f(x)$ , which is assumed *a priori* to be a Gaussian process. To do this, we need two things: a set of function observations  $\mathbf{f}(\mathbf{x})$  at some known values of the predictor  $\mathbf{x}$ ; and an expression for the covariance kernel  $k(x, x')$ . The data come from some experiment (e.g., a set of cursor coordinates  $f(x)$ ). We must, however, assume a particular covariance kernel. Although many kernels are possible, for the purposes of this paper, we will confine ourselves to the squared exponential (SE) or “radial basis function” kernel:

$$k(x, x') = f \exp \left[ -\frac{1}{2} \left( \frac{|x - x'|}{l} \right)^2 \right]. \quad (1)$$

The SE kernel is symmetric, is strictly positive, and most important for our purposes later, is infinitely differentiable. Notice that this kernel has two “hyperparameters”:  $f$ , which scales the maximum possible covariance; and  $l$ , which functions as a length scale. Later, we will consider how the values of these hyperparameters may themselves be estimated from data, but for the moment we shall assume they are known and fixed.

Armed with a set of observations and knowledge of the covariance kernel, we now wish to perform inference on the function that is presumed to have generated the observations. In other words, we are following the logic of Bayes’ rule:

$$p(\theta | \mathbf{x}, \mathbf{f}(\mathbf{x})) = \frac{p(\mathbf{x}, \mathbf{f}(\mathbf{x}) | \theta) p(\theta)}{\int p(\mathbf{x}, \mathbf{f}(\mathbf{x}) | \theta) p(\theta) d\theta},$$

where  $\theta$  are the parameters of the Gaussian process. Unlike in other regression settings (e.g., linear regression), where the parameters are a finite number of regression coefficients, the parameters of a Gaussian process may be infinite in number, since a GP prior allows nonzero probability to any functional form. We can, however, express our knowledge of the parameters of the GP *implicitly* via the posterior predictive distribution over *novel* function observations  $\mathbf{f}(\mathbf{x}^*)$ . This distribution is obtained by marginalizing over the parameters of the GP:

$$\mathbf{f}(\mathbf{x}^*) | \mathbf{x}^*, \mathbf{x}, \mathbf{f}(\mathbf{x}) = \int p(\mathbf{f}(\mathbf{x}^*) | \theta) p(\theta | \mathbf{x}, \mathbf{f}(\mathbf{x})) d\theta. \quad (2)$$

This distribution captures both the residual uncertainty about the underlying function  $f(x)$  and the knowledge gained about it from the observed data.

**Posterior Predictive Distribution** Computing the posterior predictive distribution begins with a prior on the mean and covariance functions of the GP, i.e.,  $p(\theta)$ . For the moment, we shall assume that the underlying function has a constant mean of zero, with a SE covariance function (equation 1). Expressing the likelihood of the observed function values,  $p(\mathbf{x}, \mathbf{f}(\mathbf{x}) | \theta)$ , is straightforward because they are assumed to come from a Gaussian process, and hence are jointly normally distributed. The parameters of this distribution come from our prior, i.e., the prior mean of each observation is taken to be zero, and the covariance between function values is dictated by our prior covariance kernel (the SE kernel given in eq. 1). Denoting the matrix of pairwise covariances between each observed datum as  $K(X, X)$ , we have

$$\mathbf{f}(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, K(X, X)).$$

Now, say we wish to express a posterior predictive distribution over function values at set of novel predictor values, denoted  $X^*$ . We can similarly compute a matrix of covariances between these points,  $K(X^*, X^*)$ , and between these novel points and the observed points,  $K(X, X^*)$ . Because both these novel points and the previously observed data values are presumed to have been generated by the same GP, they are all jointly normally distributed with mean zero and block covariance matrix:

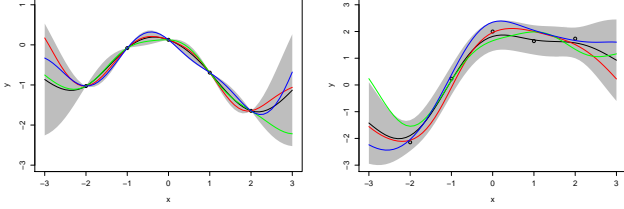
$$\begin{bmatrix} \mathbf{f}(\mathbf{x}) \\ \mathbf{f}(\mathbf{x}^*) \end{bmatrix} \sim \mathcal{N} \left( \mathbf{0}, \begin{bmatrix} K(X, X) & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix} \right).$$

We can then express the conditional posterior over  $\mathbf{f}(\mathbf{x}^*) | \mathbf{f}(\mathbf{x})$  as another multivariate Gaussian distribution, using known identities regarding the Gaussian distribution:

$$\mathbf{f}(\mathbf{x}^*) | \mathbf{f}(\mathbf{x}) \sim \mathcal{N}(K(X^*, X)K(X, X)^{-1}\mathbf{f}(\mathbf{x}), K(X^*, X^*) - K(X^*, X)K(X, X)^{-1}K(X, X^*)). \quad (3)$$

The posterior predictive distribution given just a few data points is shown in figure 1a. This figure also depicts three functions randomly drawn from this posterior. Note that they all pass through the observed function values (and the posterior variance at those points goes to zero). This is because we have assumed thus far that our function observations are noiseless; thus, we have absolute certainty that, whatever the true generating function is, it *must* pass through the values we have thus far observed. In addition, our (assumed) knowledge of the covariance kernel allows us to estimate the function’s behavior between and, to a certain extent, beyond the observed values.

In reality, we will rarely have noiseless observations of our function of interest. Luckily, observation noise is easily incorporated into the GPR framework by adding a noise term,  $\sigma^2$ , to the diagonal elements of the observed covari-



(a) No observation noise. (b) With uniform, uncorrelated Gaussian noise ( $\sigma^2 = .1$ ).

Figure 1: Examples of GPR given a set of function observations. The open circles are the observed values. The black line is the mean of the posterior predictive distribution, while the gray region is the 95% confidence region around that mean. The three colored lines are functions randomly drawn from the posterior. Covariance was assumed to be SE with  $f = 1$  and  $l = 1$ .

ance matrix, i.e.,  $K(X, X) + \sigma^2 I$ . The resulting joint observed/predicted distribution becomes

$$\begin{bmatrix} \mathbf{f}(\mathbf{x}) \\ \mathbf{f}(\mathbf{x}^*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X, X) + \sigma^2 I & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix}\right)$$

and the posterior predictive distribution changes accordingly:

$$\begin{aligned} \mathbf{f}(\mathbf{x}^*)|\mathbf{f}(\mathbf{x}) &\sim \mathcal{N}\left(K(X^*, X) [K(X, X) + \sigma^2 I]^{-1} \mathbf{f}(\mathbf{x}), \right. \\ &\quad \left. K(X^*, X^*) - K(X^*, X) [K(X, X) + \sigma^2 I]^{-1} K(X, X^*)\right). \end{aligned} \quad (4)$$

This assumes that noise is uniformly distributed and independent between observations, but if there is correlated noise between observations, this may be incorporated directly into the covariance kernel. An example of a posterior predictive distribution with observation noise is given in Figure 1b.

**GP Likelihood** In order to fit a GPR model to data, we require an expression for the likelihood of a set of function observations that are assumed to come from a GP. Luckily, as is clear from above, these observations can be treated as coming from a multivariate Gaussian with mean zero and covariance matrix  $K(X, X)$ . Thus, the likelihood is merely the multivariate Gaussian likelihood:

$$p(\mathbf{f}(\mathbf{x})) = (2\pi)^{-\frac{n}{2}} |K(X, X)|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} \mathbf{f}(\mathbf{x})^T K(X, X)^{-1} \mathbf{f}(\mathbf{x})\right] \quad (5)$$

where  $n$  is the number of observed data points and  $K(X, X)$  may be replaced by  $K(X, X) + \sigma^2 I$  if observation noise is assumed.

**Multiple Observed Functions** If multiple functions are observed simultaneously, e.g., the  $x$  and  $y$  coordinates of a cursor on a screen, they can each be treated as *a priori* independent Gaussian processes and the above reasoning applied to each individually.

## Derivatives of Gaussian Processes

Because differentiation is a linear operation, the derivative of a GP is itself a GP. Function derivatives are useful in the event that we actually have observations of the derivative (as in Solak, Murray-Smith, Leithead, Leith, & Rasmussen, 2003). However, we also argue that the derivatives of a continuous response like a mouse movement are more informative about the underlying cognitive process that generates them. For example, the acceleration is critical for finding inflection points, which could indicate that the participant is considering changing his or her mind, or that they have just incorporated new information into their decision process.

In the cases we consider below, we have direct observations only of position information, not of its derivatives (e.g., velocity and acceleration). To compute a posterior predictive distribution over function derivatives, we need only compute the covariances between each function observation and its derivatives at the points at which we are seeking predictions. This, in turn, requires expressions for the covariances between function values and derivatives, which are given for the SE covariance kernel below:

$$\frac{\partial}{\partial x} k(x, x') = -\frac{k(x, x')}{l^2} (x - x') \quad (6)$$

$$\frac{\partial^2}{\partial x^2} k(x, x') = \frac{k(x, x')}{l^2} \left[ \left( \frac{x - x'}{l} \right)^2 - 1 \right] \quad (7)$$

$$\frac{\partial^2}{\partial x \partial x'} k(x, x') = \frac{k(x, x')}{l^2} \left[ \left( \frac{x - x'}{l} \right)^2 + 1 \right] \quad (8)$$

$$\begin{aligned} \frac{\partial^4}{\partial x^2 \partial x'^2} k(x, x') &= -\frac{k(x, x')}{l^4} \left[ \left( \frac{x - x'}{l} \right)^2 \left( 3 - \left( \frac{x - x'}{l} \right)^2 \right) \right. \\ &\quad \left. + 3 \left( \frac{x - x'}{l} \right)^2 - 3 \right]. \end{aligned} \quad (9)$$

We can then compute a posterior predictive distribution over any desired derivative, given only raw function observations, by constructing the covariance matrices  $K(X^*, X)$  and  $K(X^*, X^*)$  from equation using the appropriate partial derivative above, rather than the original SE kernel  $k(x, x')$ . For example, to compute the posterior predictive distribution for the velocity,  $\dot{\mathbf{f}}^*(\mathbf{x}^*)|\mathbf{f}(\mathbf{x})$ , compute  $K(X^*, X)$  using equation 6 for each pair of predicted and observed  $x$  values and  $K(X^*, X^*)$  using equation 7 for each pair of predicted  $x$  values.

## Applications of GPR to Trajectory Analysis

In this section, we provide several examples of applications of GPR to trajectory analysis. In so doing, we introduce several extensions to the GPR modeling framework that place it in the realm of hierarchical generative models which can enable principled Bayesian inferences regarding the cognitive processes that underly observed motion trajectories.

### Estimating Hyperparameters

Although the posterior distribution in GPR is easily expressed analytically given knowledge of the covariance kernel and its

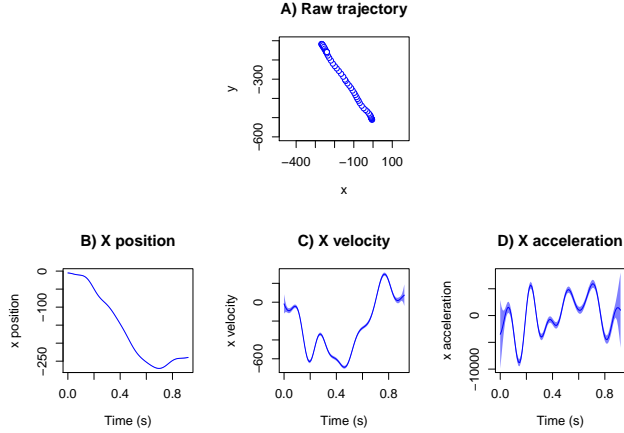


Figure 2: Example of GPR on a single two-dimensional trajectory. In B, C, and D, the light blue region depicts a 95% credible region about the mean posterior predictions.

hyperparameters, we are left with having to estimate  $f$ ,  $l$ , and  $\sigma^2$  (if we assume there is noise in the observations). In particular, we would like to be able to express our beliefs over these hyperparameters in the form of a posterior distribution. Unfortunately, this posterior will not in general be expressible analytically. Thus, we must turn to Monte Carlo methods to estimate the posterior over these parameters.

**A Single Trial** Let us assume we have a single mouse trajectory in two dimensions, as shown in Figure 2A. This trajectory is a single trial from the experiment reported by Spivey et al. (2005). On each trial of this experiment, participants saw two objects, the names of which either had similar phonological onsets (i.e., were members of the same “cohort”) or had phonologically unrelated names (the control condition). An audio recording instructed the participant to move their mouse cursor from a box in the lower center of the screen and click on one of the two objects (thus ending the trial).

The single trajectory consists of a series of  $(t, x, y)$  triples, with  $x$  and  $y$  coordinates and the times  $t$  at which they were observed. We treat the times  $t$  as a univariate predictor (i.e., in the role of  $x$  in the previous section) and  $x$  and  $y$  as conditionally independent Gaussian processes operating on  $t$ , with zero mean and SE covariance kernel (i.e., in the role of  $f(x)$  in the previous section). There are three hyperparameters that must be estimated: the parameters of the covariance kernel,  $f$  and  $l$  (see equation 1), and a noise term,  $\sigma^2$ , which is assumed to apply to measurements in both the  $x$  and  $y$  directions (isotropic noise is assumed here merely for simplicity). We choose very vague priors on each of these hyperparameters, such that they are informed almost entirely by the data, rather than our priors (although these priors could be informed by knowledge, e.g., of the accuracy of mouse position measurements). We assign a Gamma prior to  $f$  and  $l$  with shape and scale parameters set to 0.001 and an inverse-Gamma prior to  $\sigma^2$  (also with shape and scale parameters of 0.001). The

likelihood of the observed trajectory, conditional on particular values of the hyperparameters, is then given by equation 5. This model was implemented in JAGS (Plummer, 2011), drawing 1000 samples from the joint posterior over hyperparameters after 1000 steps of “burn-in”.

The estimated posterior mean of each hyperparameter is  $\bar{f} = 14760$ ,  $\bar{l} = 0.1146$ , and  $\bar{\sigma}^2 = 0.9471$ . The bottom three graphs of Figure 2 (B, C, and D) show the mean and 95% credible region of the posterior predictive distribution for the  $x$  coordinate (as well as its velocity and acceleration), marginalized over the samples of the hyperparameters.

**Multiple Trials** While this simple example illustrates how GPR can be applied to a single trajectory, we usually have several trials per participant per condition. In this case, we have multiple sets of triples,  $\{(\mathbf{t}_1, \mathbf{x}_1, \mathbf{y}_1), (\mathbf{t}_2, \mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{t}_n, \mathbf{x}_n, \mathbf{y}_n)\}$ , and we can treat them all as having been generated by the same underlying GP. In other words, even if two observations  $(x_1, y_1)$  and  $(x_2, y_2)$  were from different trials, we can still compute their covariance  $k(t_1, t_2)$  as a function of the times  $t_1$  and  $t_2$  at which they were observed, as if they were part of the same trial (and thus they also share hyperparameters). Collecting the observed function values  $\mathbf{x}_i$  and  $\mathbf{y}_i$  (where  $i$  indexes the trial), we can write

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} \sim \mathcal{N} \left( \mathbf{0}, \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) & \cdots & K(\mathbf{x}_1, \mathbf{x}_n) \\ K(\mathbf{x}_2, \mathbf{x}_1) & K(\mathbf{x}_2, \mathbf{x}_2) & \cdots & K(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ K(\mathbf{x}_n, \mathbf{x}_1) & K(\mathbf{x}_n, \mathbf{x}_2) & \cdots & K(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \right)$$

where  $K(\mathbf{x}_i, \mathbf{x}_j)$  denotes the covariance matrix between each sample in trial  $i$  and trial  $j$  (and a similar multivariate Gaussian likelihood is defined for  $\mathbf{y}$ ).

The assumption leading to the above likelihood is only valid if we assume that trajectories generated by the same participant in the same condition in fact represent samples from the same underlying process. If we assume that different trials from the same participant may be come from different processes that nonetheless share some underlying characteristics, the hierarchical extension of GPR that we introduce in the next section may be employed instead.

## Hierarchical GPR

Having shown how GPR can be applied to single trajectories and to multiple trajectories that may be assumed to share the same hyperparameters (i.e., to have been generated by the same underlying GP), we now turn to the case of multiple conditions and multiple participants per condition.

**Multiple Conditions** When there are multiple conditions in an experiment, we assume that a trajectory produced in one condition is conditionally independent of a trajectory produced in another condition, that is, that the trajectories are generated by different GP’s that nonetheless share hyperparameters. The rationale for sharing hyperparameters across conditions is simple: measurement noise (the  $\sigma^2$  parameters,

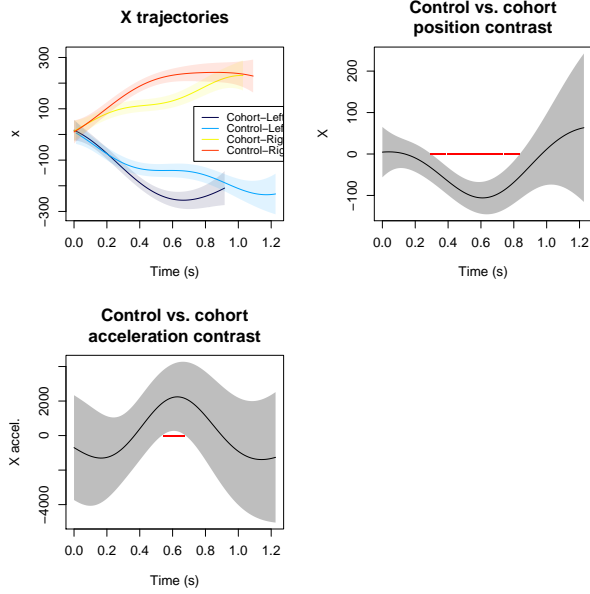


Figure 3: Posterior predictive distributions for trajectories inferred from a single subject from Spivey et al. (2005). The upper right plot shows the contrast computed between the  $x$ -positions in two cohort and two control conditions, while the lower left plot depicts the same contrast with the  $x$ -accelerations. Solid lines show the posterior predictive mean while the colored regions depict 95% credible regions around the corresponding mean.

one for each observed function) should depend only on the apparatus (e.g., the mouse or stylus). The hyperparameters of the covariance kernel, meanwhile, may be interpreted to reflect properties of the motor system of the participant, which are, of course, shared across conditions:  $f$  reflects the degree of “hysteresis”, or the tendency for the participant to produce trajectories with points that lie near one another, while  $l$  is indicative of the typical size of deviations from a straight line<sup>1</sup>.

We can again express the conditional likelihood of a set of function observations as a zero-mean multivariate Gaussian. Similar to the multiple-trial situation above, we can denote the observed trajectory points in condition  $j$  by  $\{\mathbf{x}_i\}_j$  and the covariance between each observation in condition  $j$  as  $K(\{\mathbf{x}_i\}_j, \{\mathbf{x}_i\}_j)$ . Then, we construct a block covariance matrix for the likelihood that reflects our assumptions about conditional independence between conditions:

$$\begin{bmatrix} \{\mathbf{x}_i\}_1 \\ \{\mathbf{x}_i\}_2 \\ \vdots \\ \{\mathbf{x}_i\}_n \end{bmatrix} \sim \mathcal{N} \left( \mathbf{0}, \begin{bmatrix} K(\{\mathbf{x}_i\}_1, \{\mathbf{x}_i\}_1) & 0 & \cdots & 0 \\ 0 & K(\{\mathbf{x}_i\}_2, \{\mathbf{x}_i\}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K(\{\mathbf{x}_i\}_n, \{\mathbf{x}_i\}_n) \end{bmatrix} \right)$$

And, again, we can follow the same logic to construct a similar likelihood for  $y$  or any other observed component of the trajectory.

To assess differences in trajectories between each condition, we can compute functional “contrasts” by taking the difference of the posterior predictive distributions for two conditions. This is done for one subject’s data from Spivey et

al. (2005) and shown in Figure 3. In this case, the contrast is between two pairs of conditions: the two cohort conditions (left and right) and the two control conditions. Zero lies outside the 95% credible region of the position contrast function between roughly .30 and .83 seconds, indicating that the trajectories produced by this subject to cohort and control stimuli are credibly (“significantly”) divergent over this region. This divergence results from the additional complexity of the cohort trajectories, which is shown by the acceleration contrast: The cohort trajectories include an additional “nudge” between .55 and .67 seconds after stimulus onset, as the subject reconsiders what he or she has heard.

This example illustrates two useful features of GP’s as trajectory models: First, when analyzing contrasts, they do not risk inflating the probability of false alarms due to comparisons at multiple time-points. Because a GP represents a distribution over *functions*, there is only one comparison actually taking place. Second, because GP’s allow one to compute the higher derivatives of a trajectory, they afford greater insight into the functional behavior that gives rise to observed differences between trajectories, leading to potentially useful insights into the cognitive processes that generate them.

**Multiple Participants** We further expand the scope of the analysis by allowing for multiple participants, each of whom contributes data in multiple conditions, perhaps in many trials. Researchers typically obtain trajectory measurements from multiple participants in the same condition in order to better estimate a general property that is presumed to hold across the population. In a memory experiment, this general property might be the probability of correctly recognizing a previously studied item. There may be great variability between participants in their ability to recognize the item, but each observation is presumed to be a sample from a general group tendency.

In the case of trajectory analysis, we similarly assume that each participant produces a trajectory (or trajectories) that are samples from a distribution of possible trajectories. A GP expresses just such a distribution. Hence, we assume that there is a group-level GP for each condition, the covariance kernel of which has its own hyperparameters  $f_G$  and  $l_G$ . This group level GP captures the covariance between different trials generated by different participants in the same condition. Meanwhile, the covariance between different trials generated by the *same* subject have their own covariance structure that is added to the group-level covariance. For example, if  $x_1$  and  $x_2$  are two data points observed in different conditions, their covariance  $k(x_1, x_2) = 0$ , as before. If, however,  $x_1$  and  $x_3$  come from the same condition, but different subjects, their covariance will be a function of the group-level covariance, parameterized by  $f_G$  and  $l_G$ , denoted  $k_G(x_1, x_3)$ . Finally, if  $x_1$  and  $x_4$  are two data points generated by the same subject (subject  $s$ ) in the same condition, their covariance will be the group covariance plus the covariance resulting from individual variation around the group trajectory, i.e.,  $k_G(x_1, x_4) + k_s(x_1, x_4)$ , where  $k_s(\cdot, \cdot)$  is a covariance kernel parameterized by subject-

<sup>1</sup>Of course, other choices of covariance kernel would have their own parameters which would have their own characteristic interpretations.



specific parameters  $f_s$  and  $l_s$ . As before, we can construct from these terms a covariance matrix for the entire dataset.

To perform inference in this case requires placing priors on both the group-level hyperparameters and the subject-level hyperparameters. When using vague priors, as we have thus far, it is often advisable to make use of hyperpriors. Thus, we let  $\mu_f \sim \text{Gamma}(0.001, 0.001)$  and  $\sigma_f^2 \sim \text{Inverse Gamma}(0.001, 0.001)$  be top-level priors on the mean and variance of the distribution of  $f_s$  values per subject. Then, by moment matching, we draw each  $f_s \sim \text{Gamma}\left(\frac{\mu_f^2}{\sigma_f^2}, \frac{\mu_f}{\sigma_f^2}\right)$ .

We do the same for each  $l_s$  (i.e., place a hyperprior on the mean and variance). By using hyperpriors in this way, we obtain “shrinkage” of the estimates of the subject-specific parameters, such that they can mutually inform one another.

## Generative GPR

Thus far, we have employed GPR solely in the way it was originally intended: as a nonparametric approach to function approximation—as a purely *descriptive* model. However, we can use GPR as a *generative* model in the following way: Say that we expect all trajectories in a particular condition to possess characteristic *landmarks*. These landmarks may be actual positions, or they may be particular values of one of the derivatives of the trajectory. For example, an inflection point—a point where the acceleration of the trajectory in a particular direction reverses—may have a special cognitive interpretation. In our ongoing example from Spivey et al. (2005), such a point may reflect the instant at which the word in the cohort condition has been completely processed, and the participant moves his or her cursor away from the distractor and toward the named object.

To implement this idea in a Bayesian fashion, we place a prior on the number of inflection points in the group-level trajectory. In principle, this number could be infinite, but in practice we assume this is a multinomial draw between 1 and 8 (the maximum allowed number of inflection points may, of course, vary depending on application). This multinomial is, in turn, parameterized by a draw from a Dirichlet distribution, which itself reflects a prior on the overall probability that the trajectory has a certain number of inflection points (from 1 to 8). Finally, for a given number of inflection points, the points themselves are presumed to be *a priori* uniformly distributed in time across the range of data points.

We can make use of the same formalism we have previously used to obtain the posterior predictive distribution for GPR to compute the likelihood, conditional on a certain set of sampled inflection points  $i_1, i_2, \dots, i_n$ . This involves computing the covariance matrix  $K(i, i)$  between the inflection points using the kernel in equation 9 and between the inflection points and observed values  $K(X, i)$  via equation 8. The conditional covariance of the data is then  $K(X, X) - K(X, i)K(i, i)^{-1}K(i, X)$ . The posterior predictive distribution, along with a sample of inferred inflection points, is shown in Figure 4. Notice that only the cohort conditions have inflections in the central region, reflecting the greater complexity of

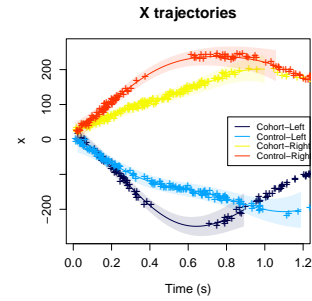


Figure 4: Posterior predictive distributions with a sample of the inflection points (+) inferred by the generative model.

both the trajectories and their underlying cognitive processes.

## Discussion

We have presented a general statistical method for modeling trajectories, and shown how it can be used to capture effects at multiple levels. A main advantage of Gaussian process regression over the various summary statistics used previously (e.g., maximum deviation) is that less information is thrown away: looking at the posterior density shows a normatively correct summary of the data, given the general assumptions made by the model. GPR balances functional complexity with capturing the underlying data, and is thus both more general and more principled than other forms of regression. Hierarchical GPR may be used to distinguish individual, group, and condition differences.

By accurately tracing and modeling the movement of the body, we can find evidence of ongoing cognitive processes, and literally see the shape of their influence. Many have wondered when psychology will reach paradigmatic maturity—like physics. Trajectories, tracing movement through space over time, are a fundamental property that all organisms and matter create.

## References

- Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, 26(5), 505–526.
- Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, 42(1), 226–241.
- Freeman, J. B., Dale, R., & Farmer, T. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, 2(0).
- Griffiths, T. L., Lucas, C. G., Williams, J. J., & Kalish, M. L. (2009). Modeling human function learning with Gaussian processes. *Advances in Neural Information Processing Systems*, 21.
- Plummer, M. (2011). *JAGS: Just another gibbs sampler*. Available from <http://mcmc-jags.sourceforge.net/>
- Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, MA: The MIT Press.
- Solak, E., Murray-Smith, R., Leithead, W. E., Leith, D. J., & Rasmussen, C. E. (2003). Derivative observations in Gaussian process models of dynamic systems. In S. T. Becker & K. Obermeyer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 1033–1040). MIT Press.
- Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences*, 102(29), 10393–10398.



# Mathematical Modeling of a Biological Odometry

Somayeh Danafar (somayeh@idsia.ch)

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA),  
Galleria 2, Manno-Lugano, 6928, Switzerland

## Abstract

Flexible and robust biological navigation are role models for robots. Biological odometry data from experiments with human subjects are explained by our novel mathematical model of biological path integration. We show the equivalence of neural representations of Polar and Cartesian egocentric path integration.

**Keywords:** Odometry; Path integration; Egocentric; Mathematical modeling.

## 1 Introduction

Navigation can be defined as a process that answers the following questions (a) “where am I?” (b) “where are other places with respect to me?” (c) “how do I get to other places with respect to me?” (Levitt and Lawton, 1990). Navigation is different from other forms of spatial behavior such as exploration, or foraging, in that there is an explicit reference to a goal location (Franz & Mallot, 2000). While many animals normally use landmarks or familiar positions to navigate, arthropods, many mammals and humans can reach their goal relying solely on their own locomotion signals. This type of navigation is known as *path integration* in biology or *odometry* in robotics.

Path integration has been studied extensively in desert arthropods and mammals (Weber et al., 1997; Séguinot et al., 1998; Etienne & Jeffery, 2004; Merkle, 2007). For humans, path integration is normally studied through triangle completion experiments (e.g. Riecke et al.). Wiener and Mallot (2006) studied visual path integration on human subjects using more complex paths with a greater number of segments and turning angles.

In robotics, sensory inputs are used to build and update a global representation of the environment. Thereafter, motor actions are derived by an inference procedure from this representation (McKerrow, 1991). The flexibility and navigation performance of biological organisms (e.g. migrating birds, arthropods) has motivated robotics researchers to adopt biologically-inspired approaches in order to achieve more accurate and robust navigation. Viewed in the opposite direction, such robots can help us to understand the behavior and biomechanics of biological systems. For instance, Möller et al. (1998) used an autonomous agent to study path integration in a type of desert ant. Lambrinos et al. (1997) studied the encoded signals of robot's wheels to estimate the moved distance. Polarized light was used as an allothetic signal. The navigation ability of a mobile robot using only visual

sensory input was investigated by Chahl and Srinivasan (1996). Weber et al. (1997) studied image motion information to estimate the travelled distance by the robot. In this paper, we address the problem of odometry through the mathematical modeling of a path integration system which matches the results from experiments conducted on human subjects (Riecke et al., 2002). In this way, we investigate what is happening at the neuronal level during the execution of the task which can later be used in biomimetic robots.

Generally speaking mathematical models of path integration can be divided in two types: geocentric and egocentric. In the present work, we focus on the egocentric model, described in section 2.1. Section 2.2 defines different sources of noise that arise in path integration. In section 2.3. the experimental data obtained from path integration with human subjects is described. In section 2.4 the mathematical model of this system is explained. To find the noise parameters which define the best mathematical model according to the experimental data, we need to solve an optimization problem which we elaborate on in section 3. The results are provided in section 4.

## 2 Path Integration (Odometry)

Mittelstaedt and Mittelstaedt (1980) established the term “path integration” and were the first to study it from a computational standpoint. They hypothesized that the signals derived from locomotion are used continuously to estimate the so-called global vector (travelled distance). This vector connects the reference point (e.g. the nest position) to the current position of the agent (e.g. the goal or target point) in a fixed coordinate system. These models of path integration are known as *Geocentric models*.

In contrast, *Egocentric models* center the coordinate system on the body of the moving agent. The agent computes and updates the sensory signals pertaining to its position and orientation in each time step (Gallistel, 1990; Benhamou and Séguinot, 1995). This approach is computationally efficient and particularly important in e.g. ants, given their limited computational resources.

Both models can be defined in terms of Polar and Cartesian coordinates. The models investigated here are based on an egocentric computation to formulate the path integration task conducted in an experiment on human subjects.

## 2.1 Egocentric Models

For path integration in egocentric models, two velocities are measured, the forward (translational) velocity,  $v$ , and the angular velocity,  $\omega$  (Figure 1). Egocentric related differential equations formulated by Banhamou and Séguinot (1995) are obtained considering small time steps. In the polar coordinates they are,

$$\frac{dr}{dt} = -v \cos \delta \quad (1)$$

$$\frac{d\delta}{dt} = v \frac{\sin \delta}{r} - \omega \quad (2)$$

The differential equations in Cartesian coordinates (by Banhamou and Séguinot, 1995) are,

$$\frac{dx}{dt} = -v + \omega y, \quad (3)$$

$$\frac{dy}{dt} = -\omega x. \quad (4)$$

This egocentric model linearly applies the parameters  $v$ , and  $\omega$ , as additive or multiplicative terms.

## 2.2 Noise Type

Homing in mammals and arthropods is imperfect. The lack of familiar positions or salient objects in identifying the starting position produces errors during path integration (Riecke et al., 2002). There are two types of errors in path integration that should be distinguished: random and systematic errors. Merkle (2007) mentioned that “there is evidence that random errors can originate from the inaccurate measurement of angles or distances, whereas systematic errors probably arise at the neural level of the organism”.

We examine the effect of both types of noise in path integration which is modelled by Monte-Carlo simulation in each unit of path movement. The first type of noise affects the sensors which measure  $v$  and  $\omega$ . This is considered due to the imperfectness of sensors. As the agent moves, it uses path integration to update its position across movement steps in relation to the reference point (nest position). The second type of noise is added to these calculated values to obtain the agent's position (Figure 2).

## 2.3 Experimental Data

To examine whether only vestibular cues are required for navigation, Riecke et al. (2002) conducted experiments on spatial orientation tasks. The experiments were conducted in the 180° Virtual Reality (VR) environment lab, with a half-cylindrical screen, where the participant is seated behind a

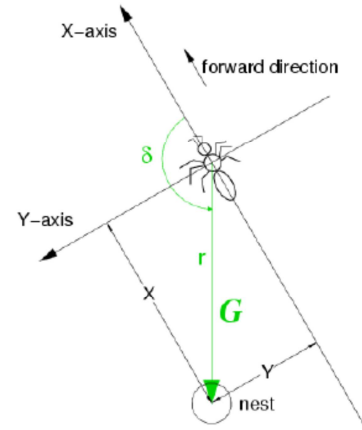


Figure 1: Egocentric path integration schema. The global vector,  $G$ , represented in Cartesian coordinates as  $G = (X, Y)$ , and the polar coordinates  $G = (r, \delta)$  (Merkle, 2007).

Sensors measure the actual translational speed  $v$  and rotational velocity  $\omega$

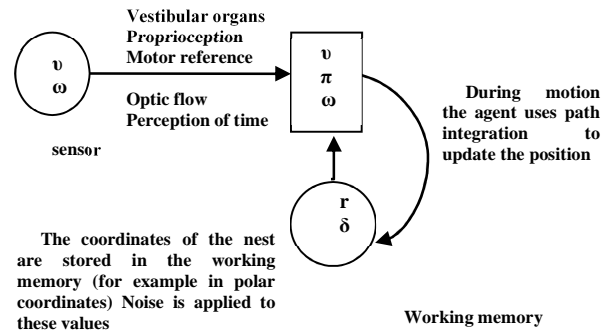


Figure 2: The types of noise in the navigation system.

table in the centre of the screen with a three button mouse, and is presented with visual cues. Pressing the middle button was used for forward translation and releasing for ending the motion. The left and right buttons were used for left or right rotations, respectively. Since there is minimum proprioceptive feedback in the button-based motion model, it is normally used as a model in VR related tasks. The experimental landscapes were streets, trees and houses. In each trial participants were presented with yellow and blue light beams, respectively, as the first and the second goal. The goals disappeared on contact. After the second goal disappeared there was a 2 second period of darkness. The task was then to return to the starting position accurately. The experiment was also done without reliable landmarks in a 3D field of blobs and with the naturalistic town environment and temporal landmarks. The reader is referred to Reicke et al. (2002) for a more detailed description.

Reicke et al. (2002) chose triangle completion since this task is “the simplest nontrivial combination of translations and rotations”. Each participant was presented with sixty isosceles triangles in random order; five different turning

angles (30°, 60°, 90°, 120°, and 150°) and two turning directions (left or right) which were repeated six times. Experimental results showed that participants could use their proprioceptive signals to estimate their travelled distance and turn back to the starting point with some bias (Figure 3). If we look at the homing trajectory end points for each participant over all his/her trials in Figure 3, we end up with a distribution over these sets of end points. We mainly work with this distribution in the next sections.

### 3 Mathematical Modeling

As in the experiment above, we model movement along isosceles triangles with 20 units and five different rotation angles between equal sides of a triangle. After passing one side of a triangle and reaching the first goal, the agent rotates and crosses the second side to reach the second goal and now it has to compute the third side of triangle. We used Monte Carlo simulations to simulate the path integration equations of section 2.1 and the noise of section 2.2. Sensor noise was added by Monte Carlo simulation as follows:

$$v' = v + N(\alpha_1 | v | + \alpha_2 | \omega |), \quad (5)$$

$$\omega' = \omega + N(\alpha_3 | v | + \alpha_4 | \omega |), \quad (6)$$

where  $v'$  and  $\omega'$  are the noisy sensor values, and  $\alpha_1$ ,  $\alpha_2$ ,  $\alpha_3$  and  $\alpha_4$  are user-defined free parameters.

The second type of noise was added to the calculated parameters which define the position of the agent in Polar or Cartesian coordinate systems. The noise can be added by Monte Carlo simulation in two ways: in a partial form, eqs. 7, 8 (Cartesian coordinates), 9 and 10 (Polar coordinates),

$$X' = X + N(\sigma).X \quad (7)$$

$$Y' = Y + N(\sigma).Y \quad (8)$$

$$r' = r + N(\sigma).r \quad (9)$$

$$\delta' = \delta + N(\sigma).\delta \quad (10)$$

or in an absolute form, eqs. 11, 12 (Cartesian coordinates), 13 and 14 (Polar coordinates).

$$X' = X + N(\sigma) \quad (11)$$

$$Y' = Y + N(\sigma) \quad (12)$$

$$r' = r + N(\sigma) \quad (13)$$

$$\delta' = \delta + N(\sigma) \quad (14)$$

Tuning the noise parameters of the Monte Carlo simulation yields different ending distributions around the reference point of the modelled triangular path. To evaluate the simulated results predicted by the mathematical model and the real experimental data, we

compared the home-ending distributions by means of a Homogeneity test (section 3.1). Determining the best noise parameter that provides the distribution closest to the real home-ending distribution of experimental data required solving an optimization problem (section 3.2). An example simulated path from our mathematical model, and the home-ending distributions are depicted in Figure 4.

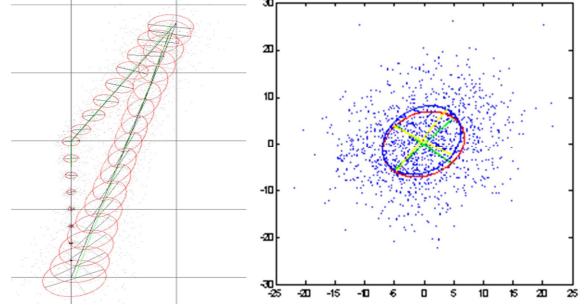


Figure 4: Left: Monte Carlo simulation generated noise. The agent's rotation angle of the agent is 45 degrees. Right: the red ellipse indicates the home-ending distribution of experimental data, the blue ellipse is the one obtained by Monte Carlo simulation.

#### 3.1 Homogeneity Testing to Compare the Distributions

Suppose  $\{x_1, \dots, x_m\}$  and  $\{y_1, \dots, y_n\}$  are two-samples drawn i.i.d. from distributions  $P$  and  $Q$ , respectively, a two-sample test tests whether  $P \neq Q$  (hypotheses are,  $H_0: P=Q$ , against the alternative  $H_1: P \neq Q$ ). We used Maximum Mean Discrepancy (MMD; Gretton et al., 2007) as our homogeneity test.

**Theorem1.** *Let  $(X, B)$  be a metric space, and let  $P$ , and  $Q$  be two Borel probability measures defined on  $X$ . The kernel function  $k: X \times X \rightarrow \mathbb{R}$  embeds the points  $x \in X$  into the corresponding reproducing kernel Hilbert space  $H$ . Then  $P = Q$  if and only if  $\text{MMD}[P, Q] = 0$ , where*

$$\text{MMD}[P, Q] := \left\| \mathbb{E}_p[k(x, \cdot)] - \mathbb{E}_q[k(y, \cdot)] \right\|_H,$$

where  $\| \cdot \|_H$  represents the RKHS norm.

For a predefined significance level (e.g. 5%), MMD values closer to zero indicate higher similarity between the distributions.

#### 3.2 Optimization

To find the noise parameters providing the modeled distribution closest to real home-ending distributions, we need to solve an optimization problem, i.e. find the maxima or minima of a so-called objective function. If the objective function is differentiable, we can use derivative-based methods to solve the optimization problem. Direct search methods are used in cases in which we do not have explicit information about the objective function, or are unable to compute the derivatives. The Nelder-Mead simplex method (Lagarias et al., 1998) is a direct search method which is widely used to optimize multidimensional objective functions with no constraints. We use the Nelder-Mead

simplex method since our objective function is a routine that does not have an analytical form. The input arguments of this routine are the noise parameters and the output is an MMD value. It has 6 free noise parameters ( $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \sigma_x$ , and  $\sigma_y$ ), e.g., of the noise which is generated for instance in a Cartesian coordination system. Without loss of generality we simplify equations 5 and 6 by setting  $\alpha_1, \alpha_2, \alpha_3, \alpha_4$  equal to 1:

$$v' = v + N(\sigma) \quad (15)$$

$$\omega' = \omega + N(\sigma) \quad (16)$$

Then our goal is minimization of the simplified objective function  $f(\sigma_v, \sigma_\omega, \sigma_x, \sigma_y)$  under the constraints of non-negative parameters (standard deviations). The solution is the minimal MMD values. As mentioned before, the Nelder-Mead simplex method is used to solve unconstrained problems; then, we need to convert our constrained problem to an unconstrained one. This is done with the algorithm introduced by J. D'Errico which uses the transformation values related to each bound, such as a quadratic function for inge bounds and a sinusoidal function for dual bounds.

## 4 Results

To make sure our objective function is not affected by sudden unexpected changes due to changing noise parameters, we approximately cover the variable space by changing steps of 0.01 to plot the function values (Table 1). To get a smoothly changing objective function we tuned the triangular side lengths to 20 meters. This value is 40 m in real experiments. There are 1000 Monte Carlo-generated home-ending data points. For MMD we use the Gaussian kernel with automated standard deviation tuning by the median sample data distance in distributions. We report results for the 5% significance level. Results for relative noise (eq. 7 and 8) and absolute noise (eq. 11 and 12) are obtained in both Cartesian and Polar coordinate systems. Results of simulated distributions in polar coordinates with absolute noise are depicted in Figure 5. Table 2 and 3 compare simulated distributions around the reference point of the navigation path with the distribution of home-ending points in experimental data for absolute and relative noise types.

An interesting question is whether polar or Cartesian coordinates are used on the neural level. Our results show they provide similar results. We also generated final distributions compatible with experimental data.

## 5 Conclusion

We introduced a novel mathematical model of egocentric path integration that uses Monte Carlo simulation of both the path integration equation and the noise. The home-ending distributions of data collected from experiments with human subjects were compared to those predicted by Monte Carlo. The closest matching distribution simulated by the model was found using the Nelder-Mead simplex method to

minimize the Maximum Mean discrepancy between the model and human data. We showed that at the neuronal level, the perceived advantage, in terms of both computational overhead and representational power, between Polar and Cartesian representations, is non-existent.

## 6 Acknowledgments

I want to thank Prof. H.A. Mallot, Prof. J. Schmidhuber, Dr. K. Basten, Dr. F. Gomez, and M. Aschoff for the profound discussions and help to prepare this paper.

## 7 References

- Benhamou, S., & Séguiot, V. (1995). How to Find One's way in the Labyrinth of Path Integration Models. *In Biol.*, 174, 463-466. (Ed.) Theor, J.
- Chahl, J.S., & Srinivasan, M.V. (1996). Visual Computation of Egomotion Using an Image Interpolation Technique. *Biological Cybernetics*, 74(5), 405-411.
- D'Errico, J. The Fminsearch bound, available at: <http://www.mathworks.com/matlabcentral/fileexchange/authors/679>.
- Etienne, A.S., & Jeffery, K.J. (2004). Path Integration in Mammals. *In Hippocampus*, 14(2), 180-192.
- Franz, M.O., & Mallot, H.A. (2000). Biomimetic Robot Navigation. *In Elsevier, Robotics and Autonomous Systems*, 30, 133-153.
- Gallistel, C.R. (1990). The organization of learning. *In Cambridge, MA: Bradford books, MIT press*.
- Gretton, A. Borgwardt, K.M., Rasch, M., Smola, A., & Schölkopf, B. (2007). A Kernel Method for the two-sample problem. *In Advances in Neural Information Processing Systems*, 19, 513-520. (Eds.) Schölkopf, B., Platt, J., Hoffman, T., MIT Press, Cambridge, MA, USA.
- Lagarias, J.C., Reeds, J.A., Wright, M.H., & Wright, P.E. (1998). Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. *In SIAM Journal of Optimization*, 9, 112-147.
- Lambrinos, D., Kobayashi, H., Pfeifer, R., Maris, M., Labhart, T., & Wehner, R. (1997). An Autonomous Agent Navigation with a Polarized Light Compass. *In Adaptive Behavior*, 6, 131-161.
- Levitt, T.S. & Lawton, D.T. (1990). Qualitative Navigation for Mobile Robots. *In Artificial Intelligence*, 44, 305-360.
- McKerrow, P.J. (1991). Introduction to Robotics. *Addison Wesley, New-York*.
- Merkle, T. (2007). Orientation and Search Strategies of Desert Arthropods: Path Integration Models and Experiments with Desert Ants, *Cataglyphis Fortis* (Forel 1902). *Dissertation for PhD, University of Bonn*.
- Mittelstaedt, M.L., & Mittelstaedt, H. (1980) Homing by Path Integration. *In Avian Navigation*, 290-297 (Eds.) Papi, F. Wallraff, H.G., Springer, Berlin.
- Möller, R., Lambrinos, D., Pfeifer, R., Labhart, T., Wehner, R. (1998) Modeling Ant Navigation with an Autonomous Agent. *In Proc. 5<sup>th</sup> Conference of Simulation of adaptive behavior*.

- Riecke, B.E., Van Veen, H.A.C., & Bülthoff, H.H. (2002). Visual Homing is Possible Without Landmarks: a Pass Integration Study in Virtual Reality. *In Presence MIT*, 11 (5), 443-473.
- Séguinot, V., Cattet, J., & Benhamou, S. (1998). Path Integration in Dogs. *In Animal Behavior*, 55, 787-797.
- Weber, K. Venkatesh, S., & Srinivasan, M.V. (1997). Insect Inspired Behaviors for the Autonomous Control of Mobile Robots. *In From Living Eyes to Seeing Machines*. (Eds.) Srinivasan, M.V., Venkatesh, S. Oxford University Press, Oxford, 226-248.
- Wiener, J.M., and Mallot, H.A. (2006). Path Complexity Does not Impair Visual Path Integration. *Spatial cognition and computation*, 6(4), 333-346.

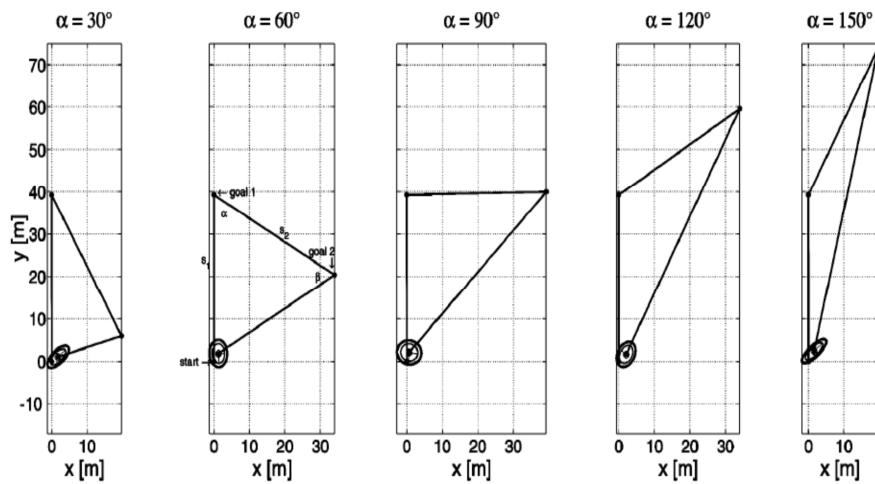


Figure 3. Homing performance in the Landmarks experiment. The data is pooled over the turning direction (left/right) as it had no significance influence on homing performance. Plotted are the mean (centroid), the 95% confidence ellipse (outer ellipse with thick line), and the standard ellipse (inner ellipse with thin line) for the homing endpoints (Reicke et al, 2002).

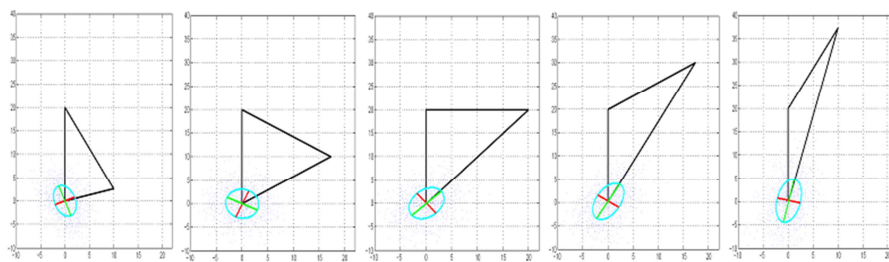


Figure 5: The results of sensor noise with absolute standard deviation production in the polar coordinate system.

Table 1: An example of how the home-ending distributions change with respect to the parameters of the Monte Carlo simulation. Entries are shown for increments of 0.01 for both  $\sigma_x$ ,  $\sigma_y$ , in Cartesian coordinates. Note how the generated Monte Carlo distributions change smoothly with  $\sigma_1, \sigma_2$ . The confidence interval over 30 runs is reported. The rotation angle of the path is  $120^\circ$ .

$\sigma_x \backslash \sigma_y$	0.01	0.02	0.03	0.04	0.05
	$\sigma_1, \sigma_2$	$\sigma_1, \sigma_2$	$\sigma_1, \sigma_2$	$\sigma_1, \sigma_2$	$\sigma_1, \sigma_2$
0.01	1.63 $\pm$ 0.01, 0.77 $\pm$ 0.01	3.24 $\pm$ 0.02, 1.55 $\pm$ 0.03	4.77 $\pm$ 0.02, 2.37 $\pm$ 0.03	6.26 $\pm$ 0.05, 3.28 $\pm$ 0.02	7.66 $\pm$ 0.1, 4.21 $\pm$ 0.08
0.02	1.63 $\pm$ 0.03, 0.77 $\pm$ 0.01	3.27 $\pm$ 0.03, 1.56 $\pm$ 0.01	4.77 $\pm$ 0.01, 2.37 $\pm$ 0.02	6.32 $\pm$ 0.06, 3.30 $\pm$ 0.02	7.65 $\pm$ 0.13, 4.23 $\pm$ 0.03
0.03	1.64 $\pm$ 0.01, 0.79 $\pm$ 0.01	3.27 $\pm$ 0.03, 1.59 $\pm$ 0.03	4.75 $\pm$ 0.10, 2.37 $\pm$ 0.01	6.28 $\pm$ 0.02, 3.29 $\pm$ 0.08	7.61 $\pm$ 0.15, 4.25 $\pm$ 0.08
0.04	1.65 $\pm$ 0.02, 0.81 $\pm$ 0.02	3.24 $\pm$ 0.04, 1.59 $\pm$ 0.04	4.78 $\pm$ 0.06, 2.38 $\pm$ 0.04	6.25 $\pm$ 0.09, 3.21 $\pm$ 0.08	7.61 $\pm$ 0.11, 4.22 $\pm$ 0.05
0.05	1.64 $\pm$ 0.03, 0.81 $\pm$ 0.01	3.25 $\pm$ 0.06, 1.60 $\pm$ 0.02	4.80 $\pm$ 0.11, 2.40 $\pm$ 0.06	6.27 $\pm$ 0.06, 3.25 $\pm$ 0.05	7.64 $\pm$ 0.10, 4.32 $\pm$ 0.05

Table 2: Comparison between simulated distributions around the initial point of the navigation path (Cartesian coordinates, various noise types). The  $\sigma_x$ , and  $\sigma_y$  to generate the Monte Carlo distributions are 0.1 and 1.6 respectively. The null hypothesis  $H_0$ , means the two distributions are similar.

Rotation Angel	Simulated dist. Cartesian ( $\mu_x, \sigma_x, \mu_y, \sigma_y$ )	Simulated dist. Polar ( $\mu_x, \sigma_x, \mu_y, \sigma_y$ )	MMD
30°	0.12, 2.41, 0.09, 3.29	-0.04, 2.36, 0.02, 3.28	$\sigma = 3.3$ Accept $H_0$
60°	0.09, 3.29, -0.01, 3.09	0.06, 3.22, -0.07, 3.08	$\sigma = 3.72$ Accept $H_0$
90°	-0.01, 3.42, -0.09, 3.37	0.18, 3.42, -0.03, 3.31	$\sigma = 3.89$ Accept $H_0$
120°	0.12, 3.11, 0.15, 3.79	0.05, 3.11, 0.24, 3.89	$\sigma = 3.95$ Accept $H_0$
150°	-0.06, 2.60, -0.01, 4.51	0.15, 2.54, 0.19, 4.66	$\sigma = 4.10$ Accept $H_0$

Table 3: Comparison of simulated distributions around the initial point of the navigation path (polar coordinates, various noise types). The  $\sigma_x$ , and  $\sigma_y$  to generated the Monte Carlo distributions are both 0.01. The null hypothesis  $H_0$ , means the two distributions are similar.

Rotation Angel	Simulated dist. Cartesian ( $\mu_x, \sigma_x, \mu_y, \sigma_y$ )	Simulated dist. Polar ( $\mu_x, \sigma_x, \mu_y, \sigma_y$ )	MMD
30°	-0.15, 6.3, - 0.21, 6.50	-0.05, 6.55, 0.19, 6.31	$\sigma = 7.59$ Accept $H_0$
60°	1.02, 6.82, 2.02, 5.14	0.82, 6.82, 1.85, 5.33	$\sigma = 6.75$ Accept $H_0$
90°	0.87, 6.04, 0.86, 4.19	0.29, 6.99, 1.4, 2.5	$\sigma = 5.58$ Accept $H_0$
120°	0.82, 6.75, 0.75, 4.88	0.29, 6.99, 0.96, 4.7	$\sigma = 6.67$ Accept $H_0$
150°	-0.03, 8.64, 0.31, 8.61	-0.03, 8.76, 0.27, 8.7	$\sigma = 10.23$ Accept $H_0$

# Solving nonogram puzzles by reinforcement learning

**Frédéric Dandurand (frederic.dandurand@gmail.com)**

Department of Psychology, Université de Montréal, 90 ave. Vincent-d'Indy  
Montréal, QC H2V 2S9 Canada

**Denis Cousineau (denis.cousineau@uottawa.ca)**

École de psychologie, Pavillon Vanier, Université d'Ottawa  
136 Jean Jacques Lussier, Ottawa, Ontario, K1N 6N5, Canada

**Thomas R. Shultz (thomas.shultz@mcgill.ca)**

Department of Psychology and School of Computer Science, McGill University, 1205 Penfield Avenue  
Montreal, QC H3A 1B1 Canada

## Abstract

We study solvers of nonogram puzzles, which are good examples of constraint-satisfaction problems. Given an optimal solving module for solving a given line, we compare performance of three algorithmic solvers used to select the order in which to solve lines with a reinforcement-learning-based solver. The reinforcement-learning (RL) solver uses a measure of reduction of distance to goal as a reward. We compare two methods for storing qualities ( $Q$  values) of state-action pairs, a lookup table and a connectionist function approximator. We find that RL solvers learn near-optimal solutions that also outperform a heuristic solver based on the explicit, general rules often given to nonogram players. Only RL solvers that use a connectionist function approximator generalize their knowledge to generate good solutions on about half of unseen problems; RL solvers based on lookup tables generalize to none of these untrained problems.

**Keywords:** Nonograms; problem solving; reinforcement learning; distance-based reward; SDCC.

## Nonogram puzzles

Invented in Japan in the 1980s, nonograms (also called Hanjie, Paint by Numbers, or Griddlers) are logic puzzles in which problem solvers need to determine whether each cell of a rectangular array is empty or filled, given some constraints. Nonograms are interesting problems to study because they are good examples of constraint satisfaction problems (Russell & Norvig, 2003), which are ubiquitous in real life (Shultz, 2001). Furthermore, despite their popularity among puzzle players, little work on nonograms exists in cognitive science, either in the form of empirical studies or modeling work. But nonograms have attracted attention in other areas. For instance, as we will see in the literature review section, solving nonograms has been studied mathematically, and a number of machine solvers exist. Finally, many rules and strategies for human players are described in web sites.

In nonograms, constraints take the form of series of numbers at the head of each line (row or column) indicating the size of blocks of contiguous filled cells found on that line. For example, in Figure 1, the first row contains 2 blocks of 2 filled cells, whereas row 5 contains no block of filled cells. Blocks have to be separated by at least one

empty cell. At the beginning, the state of all cells is unknown (often portrayed visually by a grey color), and the goal is to determine if each cell is empty (white) or filled (black), while satisfying all of the numerical constraints.

		1	2			
		1	1	0	4	3
2	2					
1	2					
	2					
2	1					
	0					

Figure 1 - Example of a 5x5 nonogram puzzle. In the initial state presented here, all cells are grey to indicate that the problem solver does not know yet if they should be filled (black) or empty (white).

## Strategies for solving nonograms

To solve nonograms, two important activities are necessary. First, the problem solver needs to decide which line (row or column) to solve next, and then to actually solve that line. Problem solvers typically need to iterate through the lines, progressively gathering more and more information about whether cells are empty or filled, until the actual state of every cell is known. Just as in crossword puzzles where the found words provide letters as clues or constraints for the orthogonally intersecting words, partially solving the cells on a nonogram line provides additional constraints for the intersecting lines.

A survey of popular web sites giving advice and tips on how to solve nonogram puzzles was performed, focusing on categorizing advice on *selection* of a line to solve, or on how to *solve* a given line. The majority of the advice relates to solving lines. For instance, an exhaustive set of rules can be found on Wikipedia (January 10, 2012 version). In contrast, there is comparatively little advice on how to appropriately select the next line to solve, and much of this



advice is given implicitly in commented solutions of specific problems. When available, explicit and general-purpose advice for line selection can be summed up as follows: begin with lines for which the constraint is either 0 (in which case all cells on that line are empty) or equal to the length of the line (in which case, all cells are filled). Occasionally, advice is also given to the effect that, by adding the different constraints (including an empty cell between each block) one can look for those that fill up a complete line (e.g., the first row in Figure 1 with constraints 2 2 is completely known as XX\_XX, with X as filled cells and \_ as empty cells). Skilled players also realize that lines that contain blocks of large sizes are often a good place to start. Finally, another general piece of advice is to pay attention to lines that have changed due to updates of the cells of intersecting lines. Except for the last one, these general advice rules often consider block constraints only, and do not take into account additional constraints imposed by cells already known to be filled or empty.

To sum up, strategies for solving lines are well-described as explicit, symbolic rules. In contrast, strategies for selecting lines appear more difficult to capture in explicit, symbolic terms, except for simple cases.

		1	2			
		1	1	0	4	3
2	2					
1	2					
	2					
2	1					
	0					

Figure 2 - Partial solution of a 5x5 nonogram puzzle.

		1	2			
		1	1	0	4	3
2	2					
1	2					
	2					
2	1					
	0					

Figure 3 - Solution of the example 5x5 nonogram, where all cells are known to be filled (black) or empty (white) and where the solution satisfies all the block sizes constraints.

Figure 2 presents an example of a partial solution of a nonogram puzzle after three steps. Column 3 and row 5 contain no block of filled cells, and thus all cells on them are empty. As described above, the first row is completely known. At this point, the position of the block of size 4 in column 4 is known, and so is the position of block 3 in column 5. Even though position of the blocks of size 1 in rows 2 and 4 cannot be determined yet, by propagating constraints from solving columns, their positions can

eventually be determined. The final solution of the puzzle is given in Figure 3.

In solving nonograms, the order in which lines are solved influences how many steps are necessary to complete the solution. A step is defined here as a single iteration on a specific line to extract the maximum information possible.

## Research on solving nonograms

Nonograms have been studied mathematically, and are known to be NP-complete (Benton, Snow, & Wallach, 2006), making search-based solution techniques practical only for small problems. More sophisticated solving approaches include rule-based techniques (e.g., Yu, Lee, & Chen, 2011), use of some intersection mechanism to prune inconsistent configurations (e.g., Yen, Su, Chiu, & Chen, 2010), linear programming (Mingote & Azevedo, 2009), genetic algorithms (e.g., Batenburg & Kusters, 2004) and a combination of relaxations and 2-satisfiability approaches (Batenburg & Kusters, 2009).

Nonograms also have been used as a tutorial for teaching university students about optimization using evolutionary or genetic algorithms (Tsai, Chou, & Fang, 2011).

To our knowledge, there have been no attempts to use reinforcement learning to solve nonograms.

## Research objectives

Our objective is to compare a solver for selecting the order in which to solve lines in nonograms based on reinforcement learning with three algorithmic methods: randomly, heuristically, and optimally (in the shortest number of steps).

We thus ask if an RL-based solver can learn good solutions, that is, solutions that are close to the optimum (i.e., shortest solution); and how they generalize to unseen problems.

## Methods

### Generated nonogram puzzles

Puzzles used for training and testing of the system have a size of 5 rows by 5 columns. The state of each cell (filled or empty) is randomly selected. The puzzle presented in Figure 1 was generated in this way, and used in the simulation. Only puzzles that have a unique solution are kept that is, puzzles for which block values correspond to one and only one board configuration. An example of non-unique problem is presented in Figure 4.

		1	1			
		1	1	0	4	3
2	2					
1	2					
	2					
2	1					
	0					

Figure 4 - Example of a non-unique puzzle. The right and the left configurations both satisfy the block size constraints.

## Training and testing regimes

RL solvers are incrementally trained on 3 different problems (starting with a single problem, and gradually increasing to three problems interleaved in the training set), and tested on a novel problem. A justification for these choices is presented in the Discussion section. Training proceeds on a single problem for 40 episodes, then training proceeds with problems 1 and 2 for 40 more episodes and finally the solver is trained on all three problems for 40 final episodes. Thus training involves 120 episodes in total. The reason for using this interleaved, incremental training (which rehearses problems already learned) is to avoid catastrophic interference (McCloskey & Cohen, 1989).

## Optimal solver for lines

Before discussing approaches to the selection problem, it is important to emphasize that simulations use an optimal line solver for solving a given line once it is selected. This optimal line solver can find all the new cells that can be declared as filled or empty.

To find these cells, the optimal line solver module first generates every possible position of all blocks (constraints) in the line consistent with the cells already determined as filled or empty. Second, it computes the intersection of all these possible positions. Finally, it identifies cells that are always filled or always empty in these intersections as such cells are now known for sure to be filled or empty.

This approach covers rules and strategies typically given to players for solving lines of nonograms, and implements rules described in other solvers (e.g., Yu et al., 2011). For instance, with a line that is currently blank (all unknowns), this method implements the rules described in Wikipedia (January 10, 2012 version) under "Simple Boxes" and "Simple spaces".

## Modules for selecting lines to solve

Given this optimal line solver for solving lines, we turn to the issue of selecting the next line (either a row or a column) to solve.

### Random solver

A random solver randomly selects the next line to solve, with replacement, from lines that are not completely solved already. Selection with replacement allows taking the same line twice in a row.

### Heuristic solver

A heuristic solver, inspired by the advice given to humans described above, selects order as follows. First, it chooses lines that have no block (easy case to solve because all cells are blank) and lines that are filled completely (e.g., 2 2). Then other lines are sorted and chosen in decreasing order of the largest block value. Lines with the same score (i.e., ties) are selected in the order in which they appear in the puzzle (rows then columns). Selection of the next line to solve is done without replacement until all lines have been

visited, after which a new round of visits is performed starting from the best remaining candidate lines.

Similarly to the general advice and rules that can be found online, the heuristic solver does not take into account the current state of the problem.

## Optimal nonogram solver

Here, a breadth-first search finds the minimal number steps. While this is feasible for the 5x5 puzzles considered here, experiments with larger puzzle sizes suggest that search time is prohibitively long -- a well-established result for NP-complete problems, as mentioned<sup>1</sup>.

## Reinforcement learning (RL) solver

The reinforcement learning (RL) solver learns the expected value of lines using a reinforcement algorithm called SARSA (Sutton & Barto, 1998). SARSA learns estimates of the expected value or quality (Q) of future rewards for every state-action pair. Its learning rule is

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha [r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$
where  $Q$  is the predicted reward (sometimes called Quality),  $s$  is a state,  $a$  is an action,  $r$  is a reward, and indices  $t$  and  $t+1$  are used for current and next states and actions respectively;  $\alpha$  is the learning rate, and  $\gamma$  is the discount factor. The name SARSA is based on the quintuple that the algorithm uses ( $s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1}$ ).

In the present simulations, we use a learning rate of 0.9, and a discount factor of 0.9 to reward the shorter solutions more highly. States, actions and rewards are described below.

### States

For a given action (a choice of a line), the state given to the RL solver consists of the states of the cells on the corresponding line. In other words, the RL solver gets information that is directly relevant to the line considered, as the action will only affect the cells in the line considered. In human problem solvers, this corresponds to an attention mechanism that focuses on the relevant elements, rather than seeing the complete problem configuration.

The state of a line consists of two elements: (1) the value of the block sizes coded using the integer corresponding to the number of cells in the block; and (2) states of cell on the line, which can take three values: unknown (U), filled (F) or empty (E). For example, in Figure 2, the state of column 1 is [1, 1, 0, F, U, U, U, E]. Note that learning begins in the initial state in which all cells of the board are Unknown, and that, for lines or columns composed of 5 cells, there cannot be more than three blocks

<sup>1</sup> As an approximation, if each line had to be solved once and only once, there would be  $(n+n)!$  combinations of sequences, which quickly increases with  $n$ . For  $n = 5$ , we get 3.6 million combinations, for  $n = 6$ , we get 479 millions; and  $n = 7$ , we already have  $8.7 \times 10^{10}$  (In practice, lines often need to be visited multiple times; or in rare cases not need to be visited at all).

## Actions

An action is a choice of the next line to solve. As there are 5 rows and 5 columns, the maximal number of actions available is 10. As soon as a line is completely solved, it is not further included in the list of actions considered for selection.

Given a list of actions and associated  $Q$  values, the RL solver needs to choose an action to execute. Action selection is performed with replacement (i.e., the same action can be taken again before all actions are visited). In testing mode, the solver uses a greedy technique called Hardmax which always selects the action with the largest  $Q$  value. In contrast, the solver uses a Softmax approach when in learning mode to allow exploration of the problem space. Under Softmax, every possible action can potentially be selected with a probability of selection increasing with the  $Q$  value.

We limit the number of steps that are allowed for finding a solution to 25, a value that is large enough to find a solution by randomly selecting actions (see Random in Figure 5). When failing to solve within 25 episodes in testing mode, the RL solver is typically stuck repeatedly selecting an action which does not yield any progress.

## Rewards

To provide rewards for line selection, we use a variant of a distance-reduction heuristic called distance-based rewards (Dandurand, Shultz, & Rey, 2012). Here, we return a reward equal to the proportion of the currently unknown cells that are determined as filled or empty as a result of selecting this action (i.e., solving this line). For instance, if cell states of the line are [U F U U E] before performing the action and [U F F U E] after, then 1 of the 3 unknowns were discovered, and this action would be rewarded with 0.33.

## Storing $Q$ values

We compare two systems for storing the rewards associated with state-action pairs. The first one is a classical lookup table which, for each unique state-action pair encountered, stores the value of the reward. Values need to be initialized to a non-zero value (here, we used 0.05) so that probabilities of selection are non-null under Softmax.

The second system consists of a connectionist function approximator. Instead of explicitly storing all  $Q$  values, a neural network is used to generate an approximation of the  $Q$  value as a single, real-valued output, taking the state-action pair as input. The three possible cell states are coded using 2 bits: Filled (1 0), Empty (0 1) or Unknown (0 0). Thus, a line state is a 13x1 vector (3 block values + 5 cells \* 2 bits per cell). Actions are coded as a 10x1 vector (5 rows then 5 columns) of binary values with the bit at the corresponding location set to 1, all others set to 0. Thus, for column 1 in Figure 2, the neural network function approximator receives as input the vector [1 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0], corresponding to the concatenation of state and action.

The neural network used is called sibling-descendent cascade-correlation (SDCC: Baluja & Fahlman, 1994), a variant of cascade correlation (CC: Fahlman & Lebiere, 1990) with reduced network depth. CC has been successfully used to model learning and cognitive development in numerous tasks (Shultz, 2003). Whereas default SDCC parameters are optimized for pattern classification, more appropriate parameter values were selected here for function approximation, namely to allow for longer input and output phases (see details in: Dandurand et al., 2012). Input and output phases were allowed to last for 200 epochs with a patience of 50 epochs. Change threshold was set to 0.01 for input phases, and 0.002 for output phases. Finally, the score threshold parameter was set to 0.025, a value that is small enough to approximate the targets well while limiting overfit.

A cache system is used to interface SARSA and SDCC because they have different processing requirements. More specifically, SARSA updates its approximation function  $Q$  after every action (called online learning). In contrast, learning in SDCC involves multiple patterns (input-output pairs) at once (called batch learning). SARSA updates the cache buffers until there are enough patterns to make a batch to train CC; details can be found in Rivest and Precup (2003).

## Results

First, we investigate the characteristics of the puzzles used for testing and training using the three algorithmic solvers. A sample of 20 simulations was run, with each simulation learning 3 different puzzles and different random initializations of neural networks. The numbers of steps necessary to solve these puzzles, plotted in Figure 5, suggest that puzzles are well-matched across training sessions and testing.

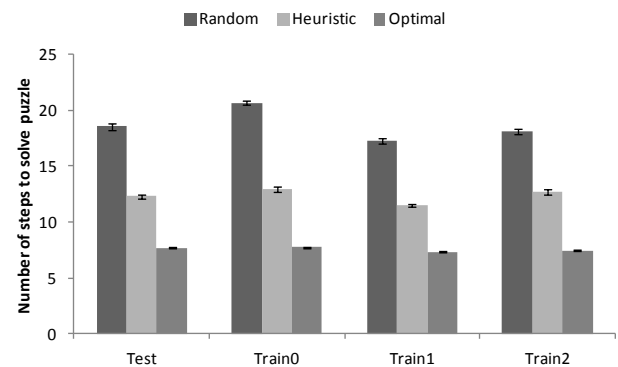


Figure 5 - Number of steps necessary to solve test and training nonogram puzzles using the three algorithmic solvers. Error bars represent standard errors (SE).

Figure 6 shows performance results for the lookup table. As we can see, it learns near-optimal solutions for the training material ( $M = 8.5$  and  $M = 7.6$  for RL and optimal nonogram solvers respectively); outperforming the heuristic

solver. However, RL solvers based on a lookup table do not generalize at all; performance on the test set reaching the maximal 25 steps on all 20 replications.

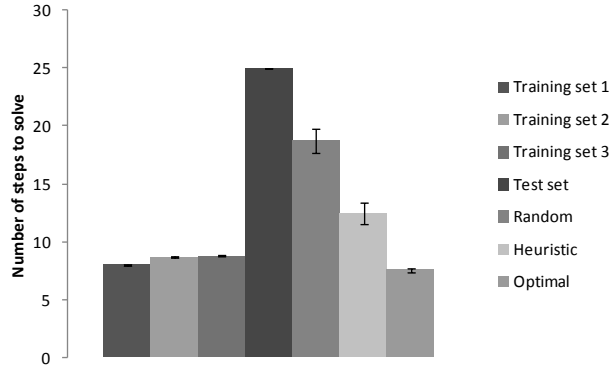


Figure 6 - Number of steps to solve nonogram puzzles with the RL solver using a lookup table, compared with the three algorithmic solvers, with standard error (SE) bars

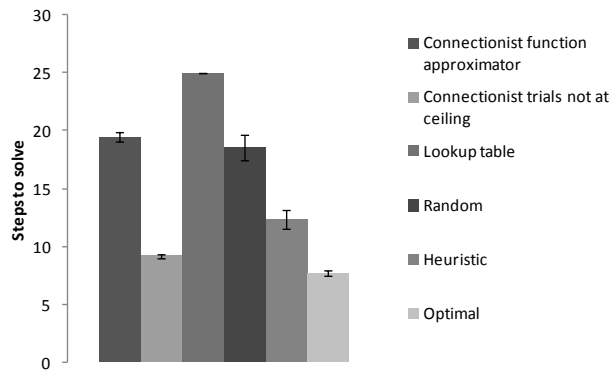


Figure 7 - Generalization performance (i.e. test set) of RL solvers using connectionist function approximators and lookup tables, with a ceiling at 25 steps, compared with the three algorithmic solvers, with standard error (SE) bars

Next, we investigate generalization performance comparing the lookup table with the connectionist function approximator (see Figure 7). As we can see, the connectionist function approximator does generalize to good solutions of 9.1 steps (in the second column) for 7 of the 20 simulations. The other 13 are at ceiling, which, when included makes the average shown in the first column. Networks that reach the ceiling value typically get stuck selecting repeatedly an action that does not yield progress, due to the use of Hardmax and a selection strategy with replacement. Because Hardmax is only sensitive to the action with the largest Q value, the choice of action does not reflect learning all other actions<sup>2</sup>. In particular, Hardmax may select a poor action that obtains the best Q value, even

<sup>2</sup> We have considered presenting results with Softmax, which does reflect learning of all actions. However, we found too much variability, and that performance was inflated due to capitalization on chance (randomly selecting good actions within the ten available).

when good action is just below. We get back to this issue in the discussion.

Finally, we compare the rewards learned by the SDCC-based RL solver with the scores given by the heuristic solver for the first move (i.e., step 1). We find a small but significant correlation of  $r = 0.18$ ,  $p = 0.01$ . This suggests that RL solvers learned good solutions without needing to fully implement the advice often given to human players for solving the initial board condition. This is unsurprising, as the rules given to humans may not be optimal, and, with many steps necessary to solve these puzzles (at least about 9 steps), there are many possibilities to the optimization process beyond step 1.

## Discussion

In general, performance differences between the heuristic and the RL solvers can be explained by the better choices that RL solvers make beyond the first solution step when the board is in its initial state. As mentioned, it is difficult to describe explicit and general purpose strategies for choosing lines to solve when the cells already solved place additional constraints. In fact, none of the online advice gives such explicit rules, but some advice implicitly provides guidance when describing solutions for specific problems. Similarly, RL solvers implicitly learn good choices for the next line to solve at any point in the solution, leading to near optimal solutions.

## Constraints on the choice of actions

As mentioned, by choosing the most flexible selection strategy, the one with replacement, networks need to learn their way out of repeatedly selecting the same action, leading to no further progress. In future research, different ways to handle this issue could be explored by placing additional constraints on action choice. For instance, RL solvers could keep track of the last  $N$  action choices made, and avoid selecting from these. With  $N = 0$ , we get the present "with replacement" scenario, whereas  $N = 10$  (5 rows and 5 columns) would correspond to the "without replacement" scenario. Imposing this constraint would sidestep the need to learn their way out of repeated action choices.

## Cognitive modeling

To our knowledge, there are no experimental studies of humans solving nonogram puzzles. The present simulations make testable quantitative predictions about how humans choose which lines to solve, thus providing some guidance for such research

As a cognitive model, the present system is hybrid, with a symbolic system for solving a given line, and a reinforcement learning with neural network support for choosing an appropriate ordering of lines. This modeling approach is grounded in evidence for both implicit and explicit cognitive processes (e.g., Reber, 1989). With Clarion (Sun, 2006) as a notable exception, modeling work on problem solving has mostly focused on explicit symbol

manipulation. The present work addresses the implicit aspects of learning how to solve problems. The fact that advice and strategies found online have little to say about the order in which to select lines suggests that it may well be an implicit task, difficult to verbalize as explicit rules.

### Towards a universal solver

Our reinforcement learning (RL) solvers learn near-optimal ordering of lines to solve nonogram puzzles, outperforming a heuristic solver based on general purpose rules for line selection. Our solvers show that multiple problems (here, 3) can be learned by a single system, and that about half of them generalize to a novel problem when coupled with a function approximator to compute, rather than merely stores, expected rewards. These results are so far limited to puzzles of relatively small size.

For future work, we could study generalization to larger puzzle sizes. Long term goals could include the design a universal solver that could solve any nonogram puzzle of various sizes nearly optimally. We tried training a RL solver with different, randomly generated nonograms on every learning episode, exploring some of the many simulation parameters (e.g., learning rates). These early attempts suggest that the task is difficult. In addition to the search space being very large, the function approximator appears to have difficulty learning stable representations when there is very high variability. Future research could also explore how to learn a larger number of problems (e.g., 10 or 100) in a reasonable time. Preliminary results suggest that it may be beneficial to gradually increase puzzle size.

### Acknowledgments

We thank Arnaud Rey for helpful comments on an earlier draft. This work is supported by a post-doctoral fellowship to FD and a grant to TRS, both from the Natural Sciences and Engineering Research Council of Canada.

### References

- Baluja, S., & Fahlman, S. E. (1994). *Reducing network depth in the cascade-correlation* ( No. CMU-CS-94-209). Pittsburgh: Carnegie Mellon University.
- Batenburg, K., & Kusters, W. (2004). A discrete tomography approach to Japanese puzzles. *Proceedings of the 16th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC)* (pp. 243–250).
- Batenburg, K., & Kusters, W. (2009). Solving Nonograms by combining relaxations. *Pattern Recognition*, 42(8), 1672–1683.
- Benton, J., Snow, R., & Wallach, N. (2006). A combinatorial problem associated with nonograms. *Linear algebra and its applications*, 412(1), 30–38.
- Dandurand, F., Shultz, T. R., & Rey, A. (2012). Including cognitive biases and distance-based rewards in a connectionist model of complex problem solving. *Neural Networks*, 25, 41–56.
- Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in neural information processing systems 2* (pp. 524–532). Los Altos, CA: Morgan Kaufmann.
- McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (pp. 109–165). San Diego: Academic Press.
- Mingote, L., & Azevedo, F. (2009). Colored nonograms: an integer linear programming approach. *Progress in Artificial Intelligence*, 213–224.
- Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of experimental psychology: general*, 118(3), 219–235.
- Rivest, F., & Precup, D. (2003). Combining TD-learning with Cascade-correlation networks. *the Proceedings of the twentieth International Conference on Machine Learning (ICML)* (pp. 632–639).
- Russell, S., & Norvig, P. (2003). *Artificial intelligence, a modern approach. Second edition*. Upper Saddle River, NJ: Prentice Hall.
- Shultz, T. R. (2001). Constraint satisfaction models. In J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences* (pp. 2648–2651). Oxford: Pergamon.
- Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge, MA: MIT Press.
- Sun, R. (2006). The CLARION cognitive architecture: Extending cognitive modeling to social simulation. In R. Sun (Ed.), *Cognition and multi-agent interaction*. New-York, NY: Cambridge University Press.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.
- Tsai, J. T., Chou, P. Y., & Fang, J. C. (2011). Learning Intelligent Genetic Algorithms Using Japanese Nonograms. *Education, IEEE Transactions on*, (99), 1–1.
- Yen, S. J., Su, T. C., Chiu, S. Y., & Chen, J. C. (2010). Optimization of Nonogram's Solver by Using an Efficient Algorithm. *Technologies and Applications of Artificial Intelligence (TAAI), 2010 International Conference on* (pp. 444–449).
- Yu, C. H., Lee, H. L., & Chen, L. H. (2011). An efficient algorithm for solving nonograms. *Applied Intelligence*, 35(1), 18–31.

# Rational Search of Associative Memory

**Eddy J. Davelaar (eddy.davelaar@gmail.com)**

Department of Psychological Sciences, Birkbeck University of London  
Malet Street, WC1E 7HX, London, UK

**J. Isaiah Harbison (isaiah.harbison@gmail.com)**

**Erica C. Yu (ericayu@umd.edu)**

**Erika K. Hussey (erikahussey@gmail.com)**

**Michael R. Dougherty (mdougher@umd.edu)**

Department of Psychology, University of Maryland at College Park  
College Park, MD 20742 USA

## Abstract

An important component of many, if not all, real-world retrieval tasks is the decision to terminate memory search. Despite its importance, systematic evaluations of the potential rules for terminating search are scarce. Recent work has focused on two variables: the total time spent in memory search before search is terminated and the exit latency (the time between the last retrieved item and the time of search termination). These variables have been shown to limit the number of plausible rules for terminating memory search. Here, we introduce an alternative stopping rule based on a rational moment-to-moment cost-benefit analysis. We show its ability to capture critical latency data and make testable predictions about the influence of changing the relative costs and benefits of memory search. Results from an experiment are presented that support the model's predictions. We conclude that the decision to terminate memory search is based on moment-to-moment changes in subjective value.

**Keywords:** Stopping rule; memory retrieval; free recall.

## Introduction

One of the most influential developments in cognitive psychology and cognitive science is that of a detailed theoretical framework of memory processes. In the late 1960s, Murdock (1967) summarized a view held by many theorists into the “modal model”, a model in which information (memoranda) transfers from sensory memory to short-term memory and then to long-term memory, with each subsequent system having greater memory persistence. The modal model was mainly a framework of memory encoding and the details of memory retrieval were left less-specified. Later theories explicated the retrieval processes in more detail (Anderson, 1972; Metcalfe & Murdock, 1981; Raaijmakers & Shiffrin, 1980, 1981). A common aspect in these theories is the assumption that retrieval from memory can be seen as a search process (Yntema & Trask, 1963) which takes time to complete. Importantly, in order to characterize this search process, models of memory were endowed with stopping rules that prevent the models from continuing search indefinitely. Despite the fact that theoreticians have been quick to incorporate stopping rules into models of memory, research evaluating the class of

stopping rules that might characterize people's decision to terminate memory search is limited.

The evaluation of stopping rules in models of recall is of both theoretical and practical interest. From a theoretical perspective, the goal of developing a comprehensive model of memory retrieval necessitates that we specify the control systems that operate on the memory representations (Newell, 1973). Any particular memory model might yield qualitatively different predictions depending on the specification of the control structures. This is particularly true for stopping rules, since the particular stopping rule employed will affect how long the model will persist in search, which can potentially affect the output of the model (number of items retrieved) and retrieval latencies.

From a practical perspective, understanding stopping rules in the domain of memory retrieval can be informative for the development of artificial intelligence and decision support systems, as well as for cognitive models of diagnostic hypothesis generation and judgment (Thomas, Dougherty, Sprenger, & Harbison, 2008). Within these systems, different stopping rules may yield qualitatively different solutions to diagnostic problems, with optimal solutions requiring different stopping rules depending on the task requirements.

In this paper, we extend the analytical work by Harbison et al., (2009) and implemented a stopping rule that is motivated by a rational analysis of memory (Anderson & Milson, 1989). The resulting rational model is tested against new data.

## Stopping rules

Atkinson and Shiffrin (1968, page 121) suggested a number of stopping rules, which have been implemented in models by a number of authors. These different stopping rules are: an internal time limit (Davelaar, et al., 2005; Davelaar, 2007; Diller, Nobel & Shiffrin, 2001; Farrell & Lewandowsky, 2002; Metcalfe & Murdock, 1981), a strength threshold (Anderson, et al. 1998; Diller, Nobel & Shiffrin, 2001), and an event-counter that would terminate search after a prespecified number of events (Raaijmakers & Shiffrin, 1980; Shiffrin, 1970).

Given the various stopping rules employed in the literature, it is clear that little heed has been paid to how a chosen stopping rule might affect the model's retrieval dynamics. Furthermore, the empirical research on which to test candidate stopping rules has been missing. The presence of self-terminating stopping rules in models of memory is in recognition of the fact that human observers are often required to self-terminate retrieval. Yet, most empirical studies of free recall have masked the contribution of stopping rules by providing participants with a pre-set retrieval interval. The use of pre-set retrieval intervals eliminates the need for the participant to utilize a stopping rule and even if participants were to use such a rule there would be no method of measuring it.

In order to address stopping rules in recall, one needs to allow participants to terminate their own retrieval episode. Consequently, the procedure of interest here is one in which the participant is given all the time they need for retrieval, but allowed to terminate retrieval whenever they wish (Dougherty & Harbison, 2007; Harbison, et al., 2009). This paradigm yields two temporal variables anticipated by models of memory that are important for understanding search termination, but which have received relatively little attention in the literature. The first of these reaction time measures is total time. Total time indexes the elapsed time between the onset of a retrieval cue (i.e., the initiation of the retrieval episode) and the decision to terminate retrieval (i.e., termination of the retrieval episode). The fact that models of memory incorporate stopping rules suggests that these models yield total time predictions. Obviously, different stopping rules will yield different total time predictions, but on an intuitive level one would expect total time to be monotonically related to total number of items retrieved: Total time should increase with the number of items retrieved.

The second reaction time measure is what Dougherty and Harbison (2007) called the exit latency. Exit latencies index the amount of time between the final successful retrieval and the decision to terminate of search. In contrast to total time, there is no obvious, intuitive prediction regarding how long participants will persist in retrieval as a function of number of successful retrieval attempts. Thus, exit latencies provide a potentially diagnostic source of data for evaluating stopping rules, particularly when considered in conjunction with the total time measure.

Few published studies report data on the two temporal variables relevant for measuring termination decisions (Dougherty & Harbison, 2007; Harbison, et. al., 2009; Unsworth, Brewer & Spiller, 2011). In the study by Dougherty and Harbison (2007), participants were visually presented with a cue word and 10 target words (A-X<sub>1</sub>, A-X<sub>2</sub>, ..., A-X<sub>10</sub>). They were told to remember the target words that were presented with each cue word. Each list of 10 target words had a unique cue word. Twelve such lists were presented in blocks of three. After each block of lists were presented, participants were given a cue word and had to report verbally as many words studied with that cue word

(A-?) as they could. Responses were recorded and participants pressed the space-bar to indicate that they could not generate additional words. The total time participants spent in search was measured as the time between presentation onset of the cue for retrieval and the time of pressing the space-bar. The exit latency was measured as the time interval between the last retrieved item and the time of pressing the space-bar.

The pattern of results regarding the stopping and exit latencies as a function of the number of words retrieved in that trial has been shown to be consistent across experimental manipulations (Harbison, et al., 2009). Typically, total time is an increasing function of the number of words retrieved in that trial, whereas exit latency is a negatively decelerating function of the number of words retrieved in that trial.

### Evaluating Stopping Rules

Harbison et al. (2009) conducted a simulation study to compare several of stopping rules suggested by Atkinson and Shiffrin (1968). They used the Search of Associative Memory (SAM; Raaijmakers & Shiffrin, 1981) and implemented the different stopping rules. The models were evaluated on their fit to data. Of the rules tested, only the total number of failures rule fitted the data both qualitatively and quantitatively. This is the rule that was used in the original SAM paper. The total number of failures rule is a special case of an iterative rule that is only concerned with the current sample from memory and the total accumulated number of failures. This lends itself to a rational analysis of the same rule which can make novel predictions.

We see memory retrieval as a form of information sampling for which a cost is incurred with every sampling attempt and a benefit is obtained for successful retrievals. We define the memory value function in which the total net value during the retrieval phase is a function of the total number of items retrieved at the elapsed retrieval time.

Elsewhere (Davelaar, Yu, Harbison, Hussey & Dougherty, submitted) we have derived a closed-form expression for the exit latency, where the decision to terminate search depends only on the information of the last time-step. We converged on the following rule (cf. Anderson & Milson, 1989):

*Terminate search when the additional cost of retrieving the next item starts to outweigh the relative or marginal benefit of having retrieved that item.*

We assume that a cost,  $a$ , is incurred with every sampling attempt,  $t$ , and a benefit,  $b$ , is obtained with every successful retrieval. We define the memory value function as:

$$V_t = Q + bN(t) - at \quad (1)$$

where  $b$  and  $a$  are the benefit and cost parameters.  $N(t)$  is the total number of items retrieved at time  $t$ . The net\_value,  $V_t$ ,



has a constant,  $Q$ , which is interpreted as the starting value that is related to factors such as motivation or time-pressure.

This stopping rule is based on the additional cost of retrieving the next item compared to the relative benefit of having retrieved that item. In other words:

$$\text{cost}(t + 1) - \text{cost}(t) > b/V_t \quad (2)$$

This equation states that when the difference in cost at time  $t$  and time  $t + 1$  is greater than the relative benefit, the memory search will be terminated.

We implemented this rule in SAM, replacing the retrieval failures rule. Figure 1 (top panel) shows the latency functions for the original SAM model. The retrieval failures rule captures both the increase in total time with total number of words recalled and the convex exit latency function. The bottom panel of Figure 1 shows the latency functions of SAM with relative benefit stopping rule. This model also captures the typical data patterns. In addition, when the relative cost is increased, the model predicts that both latency functions are lowered. That is, increased cost decreases the total time spent in memory search and decreases the time spent after the last item before deciding that further retrieval is futile. Importantly, these changes are independent of the total number of items retrieved.

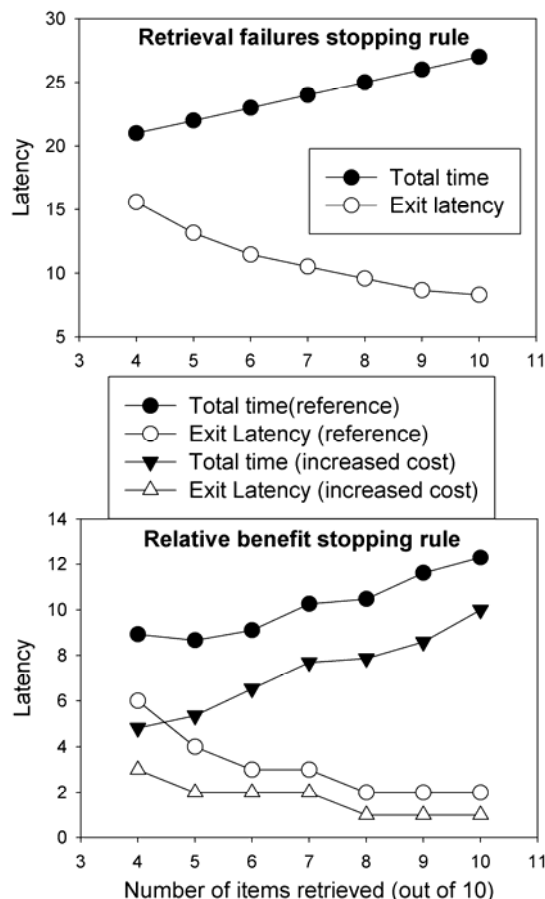


Figure 1. Simulation results of SAM with a retrieval-failures (top) and a relative-benefit stopping rule (bottom).

To summarize, a SAM implementation in which the decision to stop memory search is based on a moment-to-moment cost/benefit analysis predicts that when the cost increases (or benefit decreases) the search will terminate sooner. The next experiment tested these predictions.

## Experiment

### Methods

#### Participants

Forty-five college-aged participants were recruited from the University of Maryland subject pool and received performance-based compensation (\$15 or \$20) for participation in the study. Two participants were removed from analysis due to data collection errors.

#### Design and materials

The design used a delayed free recall paradigm whereby participants studied word lists, completed distractor math problems, and verbally recalled words from the most recent list using a PC-based microphone. The session was presented in two blocks. The first was a baseline block of 16 trials with the same payoff structure across participants (+100 for a correct recall, -100 for each second spent and incorrect recall). In the second block, cost and reward were varied between participants: one group was given an increase in reward (+150) for a correct recall and a simultaneous decrease (-50) in each second spent and each incorrect recall; the other group was given the inverse (+50 rewards, -150 cost). Retrieval protocol followed the self-terminated search paradigm used by Dougherty and Harbison (2007): participants were instructed that they had unlimited time to recall words and could end the recall period at any time by pressing the spacebar. The experimenter monitored the participant's recall and updated the participant's score in real-time, providing feedback to the participant on screen. Thirty-two lists of monosyllabic words were randomly created for each participant. List length was varied between 5, 7, 9, and 11 words and presentation order was randomized to prevent strategy use.

#### Procedure

Participants were informed they would complete a verbal recall task. The study words were sequentially presented in the center of the computer monitor for 2 s each. Following each study list, a distractor task was presented, which consisted of two simple, timed math problems. Problems contained three digits and two operands (e.g.,  $3 * 2 + 1 = ?$ ) and always resulted in a single-digit answer (digits 0-9). A question mark prompted the participant to enter an answer. Components of the math problem were presented sequentially for 1 s each. After two math problems, participants were prompted to begin verbally recalling words from the most recent study list and press the spacebar when they were finished retrieving. After the spacebar press, participants were prompted to press the spacebar again to begin the next study list when they were ready.

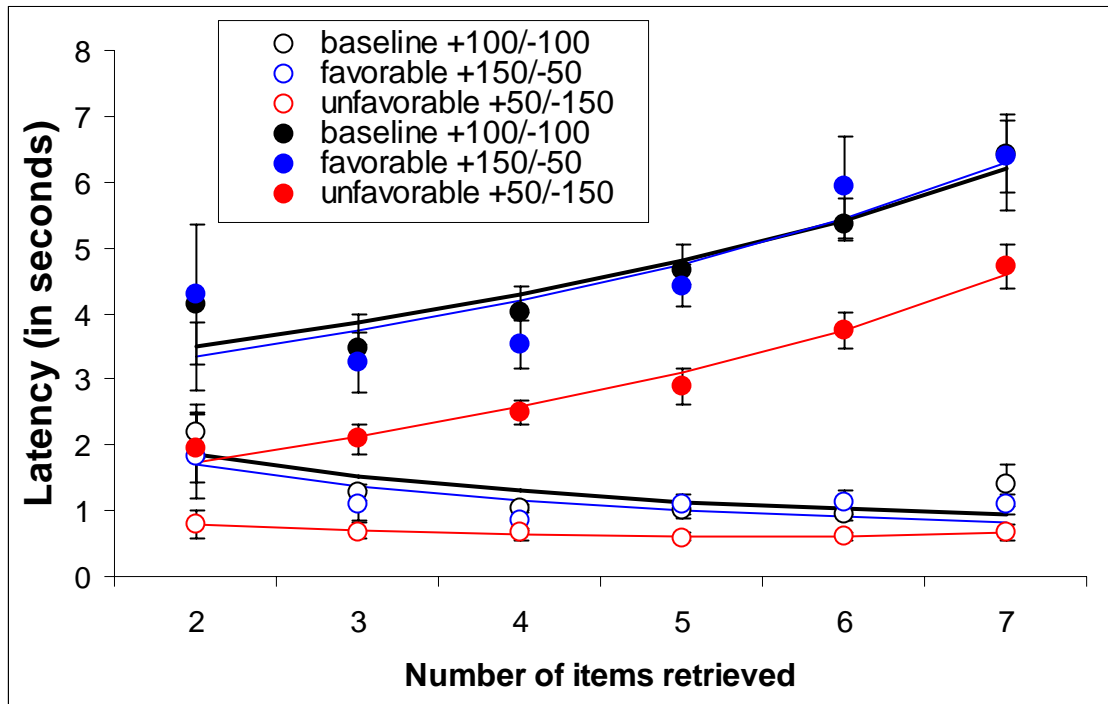


Figure 2. Total time and exit latency functions for the baseline block (both groups combined) and the second block (favorable and unfavorable condition). Error bars are standard errors of the mean. Only the last 8 trials of each block were used. Lines represent the best-fitting regression equation (Davelaar, et al., 2012).

### Coding

Using PennTotalRecall audio-analysis software, verbal retrieval data were retrospectively analyzed with millisecond accuracy. Two coders independently coded: 1) all words that were produced by each participant on each trial, 2) the time stamps of the verbal onset of all generated words, and 3) the time stamps of retrieval termination (i.e., times associated with spacebar presses). From these data, number of items retrieved, number of intrusions including repetitions and intra- and extra-list false alarms, inter-retrieval times, and exit latencies (i.e., time between end of final word retrieved and retrieval termination) were calculated. Each subject's trials were averaged before summarizing across subjects.

### Results

A 2x2 mixed design included an initial baseline control environment (+100 correct recall, -100 second spent or incorrect recall) and a second payoff environment varied between subjects (favorable: +150, -50; unfavorable: +50, -150). Due to steep learning curves in each new environment, only the last 8 of the 16 trials in each block were included in the following repeated measures ANOVA analyses.

The net points (rewards for correct recalls less the penalties for incorrect recalls and time spent) were updated in real-time for participants to use as feedback to monitor their own retrieval performance. As predicted, net points earned in each block increased over time [ $F(1,41) = 6.77$ ,  $p < .013$ ,  $\eta_p^2 = .14$ ] and the participants for whom the rewards

increased and costs decreased earned more points overall [ $F(1,41) = 15.23$ ,  $p < .001$ ,  $\eta_p^2 = .27$ ]; while net points in the baseline block were equivalent across conditions (favorable:  $M = -23.21$ ,  $SE = 41.04$ ; unfavorable:  $M = -35.80$ ,  $SE = 40.10$ ), performance splits drastically in the second block (favorable:  $M = 281.85$ ,  $SE = 54.97$ ; unfavorable:  $M = -161.08$ ,  $SE = 53.71$ ; condition x time:  $38.80$ ,  $p < .001$ ,  $\eta_p^2 = .49$ ), showing that the manipulation worked.

Total number recalled, including intrusions and repetitions, did not vary due to time, payoff environment, or an interaction of the two [conditions:  $F(1,41) = 1.61$ , ns,  $\eta_p^2 = .04$ ; time:  $F(1,41) = 3.36$ , ns,  $\eta_p^2 = .08$ ; condition x time:  $F(1,41) = 3.84$ , ns,  $\eta_p^2 = .09$ ]. Overall, the rate of intrusions was low (0.3 intrusions per list).

Temporal measures were sensitive to learning across the experiment: total time and exit latency both improved significantly for all participants [total time:  $F(1, 41) = 22.19$ ,  $p < .001$ ,  $\eta_p^2 = .35$ ; exit latency:  $F(1,41) = 12.95$ ,  $p < .001$ ,  $\eta_p^2 = .24$ ]. This performance improvement came primarily from the participants for whom the rewards decreased and the costs increased: the interaction between time and payoff structure was significant for both measures [total time:  $F(1,41) = 29.01$ ,  $p < .001$ ,  $\eta_p^2 = .41$ ; exit latency:  $F(1,41) = 9.98$ ,  $p < .003$ ,  $\eta_p^2 = .20$ ], but the main effects of condition were not significant [total time:  $F(1,41) = 1.14$ , ns,  $\eta_p^2 = .03$ ; exit latency:  $F(1,41) = 2.54$ , ns,  $\eta_p^2 = .06$ ].

Figure 2 shows the data on the retrieval latencies broken down by block and condition. Only those trials for which there were sufficient datapoints were included for the model

fit. The solid lines are the best-fitting regression equation derived in Davelaar et al. (2012). This regression model is based on the rational analysis and is a closed-form expression of the simulation rule in equation 2. The use of a closed-form expression facilitates identification of misfits that are due to theoretical misfits instead of sampling noise. The regression model also speeds up the simplex fitting procedure, which requires extremely large samples to fit the latencies at very low and very high total recall. The prediction was that increase in cost or decrease in benefit would lower the latencies. Compared to the baseline condition, making the test hard by increasing the cost and decreasing the benefit did indeed lower all retrieval latencies. Nevertheless, the opposite manipulation, decreasing the cost while simultaneously increasing the benefit, did not change the latencies compared to baseline. We address this asymmetry in the general discussion.

## General Discussion

The purpose of this paper was to extend our earlier work on stopping rules by proposing a stopping mechanism that is motivated by a rational analysis of decisions made on a moment-to-moment basis. The resulting rational SAM model produces the typical latency functions that several commonly used stopping rules failed to capture. The model makes testable predictions about the influence of monetary payoff structure on retrieval latencies and the decision to stop memory search.

The prediction was that making it harder to gain points would lower the retrieval latencies due to higher probability of stopping, whereas the reversed would be the case when it was easier to gain points. Interestingly, only the former prediction was borne out by the data and model fits. The results might be seen as an instance of loss aversion by suggesting what could be called an “it-ain’t-broke” hypothesis. Loosely put, when it is harder to obtain points, the cognitive system readjusts itself to avoid losing too much. However, when the environment changes to such an extent that it becomes easier to gain points, the system will not calibrate itself to then minimize the gains. Hence, if the cognitive system is not losing by what it does (i.e., it-ain’t-broke) then there is no reason for adjusting the cognitive parameter (i.e., don’t-fix-it).

Anderson and colleagues provided a rational analysis of the free recall task (Anderson & Milson, 1989; Anderson & Schooler, 1991), in which each item has a need probability associated with it. Only those items are retrieved whose need probability is larger than a certain criterion, which increases with the time spent inspecting an item before accepting or rejecting it. Anderson and Milson (1989) were able to capture a number of basic memory phenomena using their adaptive perspective. However, their analysis only provided the time of the last retrieved item and not of the exact time of terminating memory search. A possibility would be to use the criterion to estimate the termination time, but this would require knowing the functional form of how the criterion changes during item inspection.

Nevertheless, the success of Anderson’s rational analysis and our current results warrants investigating how these can be combined and would allow analyzing the consequences of different retrieval processes on stopping rules. This also applies to research based on the animal foraging literature, such as problem solving (Payne & Duggan, 2011) and information foraging (Pirolli & Card, 199). We leave such an endeavor for the future.

Our analysis suggests that stopping rules should play a more central role in the development and testing of models of memory. The choice of stopping rule has major impact on the overall model behavior. Obviously, one of the ultimate goals of memory theory is to characterize memory retrieval in general, both in and out of the lab. By focusing more on how people terminate memory search, we can bring our models more in line with the type of retrieval tasks that characterize retrieval tasks outside of the free-recall paradigm.

Investigating stopping rules has important implications for understanding tasks other than free-recall. For example, within the medical decision making literature, it is clear that physicians entertain costs when determining when to terminate their retrieval of diagnostic hypotheses from memory (Weber et al., 1993). More recently, Dougherty and Hunter (2003a; 2003b) showed that the perceived probability of any particular event (a hypothesis) is partially dependent on the number of alternatives retrieved from memory, which was affected by time pressure. This suggests that the decision to terminate memory search will affect his or her perceived probability of a particular hypothesis. Within the frequency judgment literature, Brown and colleagues (Brown, 1995; 1997; Brown & Sinclair, 1999; Conrad, Brown, & Cashman, 1998) have shown that participants’ responses to survey questions often are a monotonically increasing function of total time spent searching memory. Thus, the magnitude of participants’ frequency judgments on behavioral survey questionnaires should be affected by when they terminate search of long-term memory. Although the above tasks are all quite distinct, they serve to underscore the ubiquity of stopping rules in real-world retrieval tasks. Therefore, understanding how people terminate memory search, and the psychological variables that affect search termination, is paramount to the development of comprehensive models of memory retrieval and to understanding the dynamics of memory retrieval outside the lab.

In summary, in this paper we obtained further evidence for the view that participants are making adaptive choices to search termination that are based on a cost-benefit analysis.

## Acknowledgments

Part of this material is based on work supported by the National Science Foundation under Grants SES-0134678 and BCS-1030831. The authors thank Dave Huber and Rick Thomas for their helpful comments.

## References

- Anderson, J. R. (1972). FRAN: a simulation model of free recall. In G. H. Bower, & J. T. Spence (Eds.), *The psychology of learning and motivation: Vol. 5. Advances in research and theory* (pp. 315-378). New York: Academic Press.
- Anderson, J. R., & Milson, R. (1989). Human memory: an adaptive perspective. *Psychological Review*, 96, 703-719.
- Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2, 396-408.
- Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, 38, 341-380.
- Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: a proposed system and its control processes. In K. W. Spence, & J. T. Spence (Eds.), *The psychology of learning and motivation Vol. 2. Advances in research and theory* (pp. 89-195). New York: Academic Press.
- Brown, N. R. & Sinclair, R. C. (1999). Estimating number of lifetime sexual partners: Men and women do it differently. *Journal of Sex Research*, 36, 292-297.
- Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1539-1553.
- Brown, N. R. (1997). Context memory and the selection of frequency estimation strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 898-914.
- Conrad, F., Brown, N. R., & Cashman, E. (1998). Strategies for estimating behavioral frequency in survey interviews. *Memory*, 6, 339-366.
- Davelaar, E. J. (2007). Sequential retrieval and inhibition of parallel (re)activated representations: a neurocomputational comparison of competitive queuing and resampling models. *Adaptive Behavior*, 15, 51-71.
- Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: empirical and computational investigations of recency effects. *Psychological Review*, 112, 3-42.
- Davelaar, E. J., Yu, E. C., Harbison, J. I., Hussey, E. K., & Dougherty, M. R. (2012). A rational analysis of memory search termination.
- Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC-REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 414-435.
- Dougherty, M. R. P., & Hunter, J. E. (2003a). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, 113, 263 – 282.
- Dougherty, M. R. P., & Hunter, J. E. (2003b). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition*, 31, 968 – 982.
- Dougherty, M. R., & Harbison, J. I. (2007). Motivated to retrieve: how often are you willing to go back to the well when the well is dry? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 1108-1117.
- Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review*, 9, 59-79.
- Harbison, J. I., Dougherty, M. R., Davelaar, E. J., & Fayyad, B. (2009). On the lawfulness of the decision to terminate memory search. *Cognition*, 111, 146-421.
- Metcalf, J., & Murdock, B. B. (1981). An encoding and retrieval model of single-trial free recall. *Journal of Verbal Learning and Verbal Behavior*, 20, 161-189.
- Murdock, B. B. (1967). Recent developments in short-term memory. *British Journal of Psychology*, 58, 421-433.
- Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (ed.), *Visual Information Processing*. New York: Academic Press.
- Payne, S. J., & Duggan, G. B. (2011). Giving up problem solving. *Memory & Cognition*, 39, 902-913.
- Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review*, 106, 643-675.
- Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. Bower (Ed.), *The Psychology of Learning and Motivation, Vol 14*. New York: Academic Press.
- Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, 88, 93-134.
- Shiffrin, R. M. (1970). Memory search. In D. A. Norman (Ed.), *Models of human memory*. New York: Academic Press.
- Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, 115, 155-185.
- Unsworth, N., Brewer, G. A., & Spillers, G. J. (2011). Factors that influence search termination decision in free recall: an examination of response type and confidence. *Acta Psychologica*, 138, 19-29.
- Yntema, D. B., & Trask, F. P. (1963). Recall as a search process. *Journal of Verbal Learning and Verbal Behavior*, 2, 65-74.

# Strong structure in weak semantic similarity: A graph based account

Simon De Deyne (simon.dedeyne@psy.kuleuven.be)<sup>a, b</sup>

Daniel J. Navarro (daniel.navarro@adelaide.edu.au)<sup>b</sup>

Amy Perfors (amy.perfors@adelaide.edu.au)<sup>b</sup>

Gert Storms (gert.storms@psy.kuleuven.be)<sup>a</sup>

<sup>a</sup> University of Leuven, Department of Psychology, Tiensestraat 102, 3000 Leuven, Belgium

<sup>b</sup> University of Adelaide, School of Psychology, 5005 Adelaide, Australia

## Abstract

Research into word meaning and similarity structure typically focus on highly related entities like CATS and MICE. However, most items in the world are only weakly related. Does our representation of the world encode any information about these weak relationships? Using a three-alternative forced-choice similarity task, we investigate to what extent people agree on the relationships underlying words that are only weakly related. These experiments show systematic preferences about which items are perceived as most similar. A similarity measure based on semantic network graphs gives a good account for human ratings of weak similarity.

**Keywords:** similarity; semantic networks; word associations.

Although similarity is a fundamental concept in cognitive science, it is still not yet well understood. Any two entities have a potentially infinite number of features or predicates in common, making it always possible to construct *post hoc* explanations for why any items are similar to each other (Goodman, 1972; Medin, Goldstone, & Gentner, 1993). Even if similarity is logically vacuous, of course, it is not necessarily psychologically vacuous: there may indeed be a small or at least finite number of shared *represented* predicates (Medin & Ortony, 1989). However, while shared representations may well explain why people share clear intuitions about the similarity of strongly related items like CATS and MICE, the notion of shared representations may not apply when the items are only weakly related. After all, the only predicates that apply to such disparate items as RAINBOW and TUNAFISH are so vague and generic that appealing to them to explain similarity begins to make it nearly as underconstrained psychologically as Goodman first showed it was in a logical sense.

Despite the questions that weak similarity raises about the nature of our underlying mental representations, it is almost entirely unstudied. Almost all investigations into stimulus similarity have focused on items that tend to be quite similar to one another – we ask people to compare the similarity of CATS to MICE, or of MICE and MEN. Rarely if ever do we ask people questions about weak similarities. We can get a sense of how extreme this bias is by examining the empirical data for a set of 372 concepts belonging to 15 natural categories (e.g., fruit, tools, sports), as in Ruts et al. (2004). We used numerical methods to calculate theoretical values for the similarities between all pairs of words in a database of 12,000 word associations. Comparing the two, we found that the *weakest* similarities for which we have empirical data were *stronger* than 97% of the similarities that were predicted according to the word association data base. This suggests

that research into similarity has focused almost exclusively on similarities between only the most related items.

From a methodological point of view, this is not surprising: if asked to rate how similar HAIL and TEACHER are to each other, most people would struggle to know how to answer. Yet this struggle does not necessarily imply that no underlying representation of similarity exists. As Goodman (1972) and others have pointed out, it is always possible to find some basis for saying that HAIL and TEACHER are similar. The real question is which of these bases form part of human mental representations, and whether there exist any systematic regularities in how people spontaneously assess these weak relationships. The goal in this paper is to investigate (a) whether these regularities exist, and (b) whether they can be accommodated by existing theories of semantic representation.

Viewed as a problem of rating the stimuli between two entities that are only weakly related, the challenge seems intractable. Intuitively it feels like the similarity between HAIL and TEACHER is zero, and there is little underlying structure to be found. However, suppose the task were framed as a three-alternative forced-choice problem (e.g., Navarro & Lee, 2002). Which of the following three concepts is the odd one out: CUP, TEACHER and HAIL? Framed in this fashion, the problem seems less intractable, and many people have very strong intuitions about what the answer should be. Sometimes the intuition can be so strong that it may be difficult to see why the answer to the question is not obvious.

As an illustration, in our discussions of this specific triple, one author strongly felt that TEACHER was obviously the odd one out because teachers are people and the other two are not (an “animate vs inanimate” distinction). Another strongly felt that HAIL is the odd one out because it is a mass noun and the other two are count nouns (a “things vs stuff” distinction). In both cases the choice also invokes quite abstract ontological categories, and relies on very broad general knowledge about the world. Obviously the decision to rely on a particular category to guide the decision making is the result of “on the fly” reasoning about the items. Although nobody felt that CUP was the odd one out, it is interesting that for both authors the intuitions were quite strong, so much so that they were somewhat surprised to discover that the supposedly “obvious” choice was not, in fact, so obvious.

This leaves us with an open question: how deep does the structure in our mental representations go? One possibility is that there is significant agreement and constraint in our mental representations only when considering the relationship between entities that are strongly related to each other.

In other words, the Medin and Ortony (1989) argument about shared predicates may only apply between items that are already highly related. If this is the case, then one might expect Goodman's problem to arise when we try to measure weak similarities, causing each person's judgment to be essentially arbitrary and there to be few stable preferences across people. The other possibility is that there is enough shared structure in our mental representations that there is a strong agreement even for such strange pairings as RAINBOW and TUNAFISH, HAIL and TEACHER and so on.

In the first half of this paper we present two experiments exploring weak similarity structure in humans. We show that similarity ratings of weakly related items are nevertheless surprisingly regular across people, and moreover that similarity judgments can be manipulated in sensible ways. In the second half, we investigate the nature of the underlying representations that might give rise to these similarity judgments. Computational modelling demonstrates that weak similarities like those found in our experiments can be at least partially captured by semantic network models constructed from word association data.

## Experiment 1

Our main goal in this experiment was to investigate whether people reliably agree in their similarity judgments even between weakly related entities. In order to avoid the difficulties inherent in asking for similarity ratings between two very different items, we had participants choose which pair out of three possible pairs in a triple was the most similar one.

### Method

**Participants** Sixty-nine native Dutch speaking psychology students participated in exchange for course credit.

**Stimuli and Materials** The stimuli were 300 nouns taken from a set of 12,000 Dutch words used as cues in the word association task described in De Deyne and Storms (2008b) and De Deyne, Voorspoels, Verheyen, Navarro, and Storms (2011).<sup>1</sup> These items were used to produce triples, which were sampled at random given two constraints. Each item in a triple was required to have approximately the same frequency and imageability rating, in order to ensure that participant responses reflected underlying semantic relatedness rather than superficial similarities in concreteness or familiarity. Word frequency was calculated based on the log-transformed lemma frequencies taken from the CELEX database (Baayen, Piepenbrock, & van Rijn, 1993), while imageability was derived from judgments on a seven-point scale found in De Deyne and Storms (2008a). Within any triple, the maximum standard deviation was  $SD_{max} = 0.52$  for lemma frequency and  $SD_{max} = 0.84$  for imageability.

**Procedure** On each trial three words were presented at the corners of a triangle, as shown in Figure 1. Participants were instructed to click on the circle corresponding to the side of the triangle that connected the most related pairs. We stressed in these instructions that we were interested in the meaning of words rather than their orthographic similarity or phonological relatedness. To illustrate what we meant, we gave par-

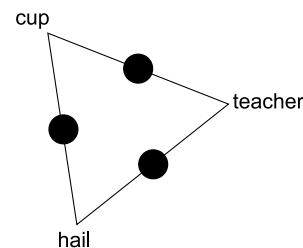


Figure 1: Example triple stimulus used in Experiment 1. The black circles indicate the controls used to select a pair with the mouse.

ticipants two example triples: in the first one (COLD - HOT - SQUARE) the first two words are related, and in the second one (MOIST - COLD - COOL) the last two are. Participants were asked to do their best even if the task seemed difficult, and not to dwell too long on a single trial but to complete the task in a spontaneous manner. The task was presented as a web questionnaire during a collective testing session.

### Results

Our key question was to what extent people tended to select the same pairs. If weak similarities do not exist or are not reliably shared by different people, we would expect all three possible pairs from every triple to be selected equally frequently. We test this in two different ways.

The first test of inter-rater reliability is to measure how often the most frequent pair from every triple is chosen. Since there are three possible pairs in any given triple, chance responding is 0.33. However, the median value was 0.67 – well above what one might expect by chance. Moreover, as Figure 2 illustrates, for 97 of the 100 triples the most commonly chosen response was selected significantly more frequently than would be expected by chance.<sup>2</sup>

Instead of just looking at the most frequent pair of any triple, we can also measure how much people's weak similarity judgments agree with one another in a more conventional way. We therefore ran  $\chi^2$  goodness-of-fit tests comparing the observed frequencies across the three responses to a null hypothesis that all three responses are equally likely for each triple separately. Taking this approach, the frequencies of 89 out of the 100 triples were significantly different from the null hypothesis,  $\chi^2(2)$ ,  $p < 0.05$ .

The results so far suggest that people encode weak regularities from the environment and do this in a systematic and measurable way. How robust is this finding? We consider this question in the next experiment.

## Experiment 2

The goal of this experiment is to investigate how robust the results from the first experiment are. If weak similarities are not "hard coded" in some way, then they must be derived or constructed somehow. Perhaps people are deriving them by searching a semantic network for the proximity of the two

<sup>1</sup>The complete list of stimuli including English translations is available at <http://ppw.kuleuven.be/concat/simon/>

<sup>2</sup>Note that the hypothesis tests here were conducted using a numerically simulated null distribution, since the sampling distribution of the maximum frequency is an extreme-value statistic and is not correctly described by a binomial distribution; it is, however, trivial to simulate numerically. Using this sampling distribution, the critical value was 0.39, corresponding to the cutoff shown in Figure 2.

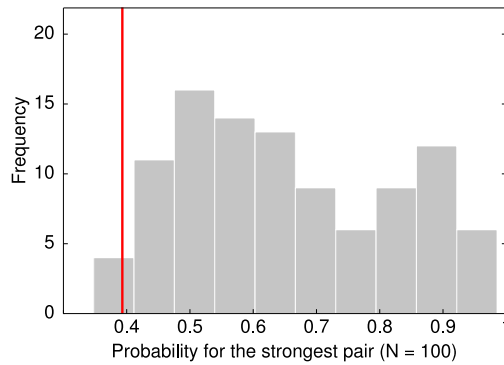


Figure 2: Distribution of the most frequently chosen pairs in Experiment 1. The vertical line indicates the 95% confidence boundary for the frequency one would expect if pairs were chosen randomly. Participants agreed with each other in selecting the pairs for almost all of the stimuli in the experiment.

items to each other, or constructing them on the fly based on some other underlying representation. In either case, we would expect that time pressure would cause less accurate estimations and more disagreement between individuals, resulting in more uniform choice probabilities than were found in Experiment 1. We therefore repeated the experiment, with the variation that this time we put participants under time pressure by asking them to decide which pair is more related as quickly as possible.

## Method

**Participants** Thirty native Dutch speaking students participated in exchange for course credits.

**Stimuli and Materials** The stimuli and materials were identical to those presented in Experiment 1.

**Procedure** The procedure was based on Experiment 1, but a few changes were made to allow for the accurate measurement of reaction times. Instead of using the mouse, participants were asked to use the keyboard, and to decide as quickly as possible which pair of words was related. At the beginning of each trial, the triple triangle was presented without the words until the participant pressed the space bar, which displayed the words at each corner. Unlike in Experiment 1, the black circles in Figure 1 were now labeled with either *J*, *K*, or *L*, and participants indicated which pair was most related by pressing the corresponding *J*, *K* or *L* key.

In order to make sure participants understood the task and were answering as quickly as possible, the main test was preceded by 20 practice items that had the words *word1*, *word2* and *word3* as labels at randomized locations. The participants were asked to click on the letter connecting word1 and word2 as quickly as possible. During this time a warning was shown when reaction times were slower than 3600ms, and participants were asked to try to make a faster response.

## Results

Before evaluating what effect the time pressure manipulation had, we first needed to clean up the reaction time data. For each individual we therefore removed any responses with reaction times higher than three standard deviations above their

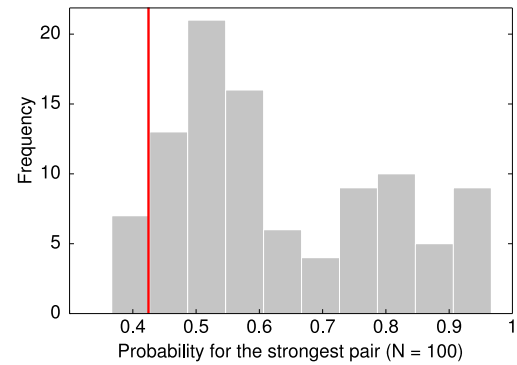


Figure 3: Distribution of the most frequently chosen pairs in Experiment 2. The vertical line indicates the 95% confidence boundary for the frequency one would expect if pairs were chosen randomly. As before, participants agreed with each other in selecting the pairs for most of the stimuli in the experiment. However, agreement was somewhat lower, suggesting that the time pressure made them unable to fully access their semantic representations, adding noise to their responses.

average, as well as reaction times faster than 300ms. The average reaction time was 3771ms ( $SD = 2131$ ). A log-transformation was used to reduce the skew in the reaction times. Next, for each participants the reaction times were transformed to *z*-scores, resulting in a Spearman-Brown split reliability ( $r_{RT}$ ) of 0.83. Since we did not record reaction times in the first experiment, it is not certain that the participants actually payed attention to the instructions and responding faster, as asked. We investigated if this was the case by running 18 new participants in Experiment 1, this time measuring their reaction times registered by keyboard response. The resulting reaction times had a mean of 4705ms ( $SD = 2864$ ), about one full second slower than the speeded judgments in Experiment 2.

We can now explore the answer to our central question: what effect did time pressure have on the reliability of weak similarity judgments? As before, we can measure how often the most frequent pair from every triple was chosen. Remembering that chance responding would be reflected in a value of 0.33, we find a median value of 0.57. As predicted, this is lower than the 0.67 of Experiment 1, but higher than what one would find if responses were random. We also found that for the vast majority of triples (93 out of 100), the most common response was selected significantly more often than would be expected by chance. Figure 3 shows the distribution of responses. It is evident that, while putting people under time pressure increases the uniformity of the distribution of responses, there is still substantial agreement. This intuition is supported by the  $\chi^2$  goodness-of-fit tests done for each triple, which finds that the frequencies within 89 out of the 100 triplets were significantly different from a null hypothesis under which all three pairs would be equally likely ( $\chi^2(2)$ ,  $p < 0.05$ ).

The results of these two experiments demonstrate that there is consistent and reliable agreement on similarity judgments, even when the entities involved are only weakly related, like CUP and TEACHER. On some level, this agreement is surprising, because such items only share features if so many



features are represented that we begin to run afoul of Goodman's problem. Even then, it is unclear that the items on which there is agreement are the items with more shared features. A more likely explanation of this finding may be that people show strong agreement because they share the kind of semantic representation that is at least partially captured by a semantic network. In the next section, we explore this possibility by modelling the similarities from the two previous experiments based on semantic graphs.

## Graph based models for weak similarity

In this section we investigate the hypothesis that at least some of the agreement between people about weak similarities arises due to shared semantic network representations. Network-based models for similarity have been proposed in related domains (e.g., the NETSCAL model by Hutchinson (1989) or the feature centrality model of Sloman and Rips (1998)), but the most similar models in psychology are the spreading activation models which accounted for a number of interesting semantic effects (e.g. Quillian, 1968).

Why might we expect semantic networks to capture some of the representation with which weak similarities are generated? Such networks probably reflect something about the way words are combined and used in the real world. For instance, the average American is exposed to about 100,500 words every day (Bohn & Short, 2009). The numerous ways that this vast amount of information can be combined may lead to an immense amount of mostly weak contingencies between items. Indeed, in recent years, the increasing availability of co-occurrence information to researchers has led to the development of models that derive representations of meaning from that co-occurrence. These models, which include latent semantic analysis (LSA, Landauer & Dumais, 1997) and topic models (Griffiths, Steyvers, & Tenenbaum, 2007), use statistical methods to extract the regularities underpinning the co-occurrence data. They thereby produce structured, meaningful representations that can be used to capture and explain human behavior and performance.

The goal behind network models is similar, though the approach is different. The network itself is derived from word associations which presumably reflect the patterns of co-occurrence in the world. We can then use the network as the core representation from which similarity measurements are derived. We theorize that although associations between individual entities may be too sparse to account for people's judgments about triples like CUP-TEACHER-HAIL, the network may capture broader relationships that *can* account for such judgments. If broad ontological distinctions like animacy or the count/mass noun distinction are reflected in the structure of the semantic network, we might expect a suitably chosen measure of network-based similarity to be able to capture, at least in part, the manner in which humans resolve the weak similarity questions that we asked in our experiments. How, though, can we measure similarity within a network? We address this problem in the next section.

## Similarity in semantic networks

Large-scale semantic models are typically extremely sparse. In the case of networks derived based on word associations, this means that the number of edges connecting any two

nodes (words) is very low. This is less of a problem when dealing with very similar concepts, because they are likely to share some edges despite the overall sparsity of the network. However, sparsity is a serious problem for other concepts. The same problem arises for non-network-based representations like feature overlap, because the number of features shared by weakly related items is very low, if not zero.

Given the problems imposed by sparsity, how can we measure similarity in a semantic network in a sensible way? We consider two different approaches here. The first is the widely used cosine measure of similarity (e.g., Landauer & Dumais, 1997; Steyvers, Shiffrin, & Nelson, 2004), which measures the extent to which two nodes in the graph share the same immediate neighbors. Two nodes that share no neighbors have a similarity of 0, and nodes that are linked to the exact same set of neighbors have similarity 1. Formally, it is defined as follows. Let  $\mathbf{A}$  denote a weighted adjacency matrix, whose  $ij$ -th element  $a_{ij}$  contains a count of the number of times word  $j$  is given as an associate of word  $i$  in a word association task. Each row in  $\mathbf{A}$  is therefore a vector containing the associate frequencies for word  $i$ . The cosine measure of similarity is the angle between these vectors, calculated as follows: because some words can have more associates than others, we normalize each row so that all of these vectors are of length 1. This gives us a new matrix  $\mathbf{G}$ , where  $g_{ij} = a_{ij} / (\sum_j a_{ij}^2)^{1/2}$ , and the matrix of all pairwise similarities is now

$$\mathbf{S} = \mathbf{G}\mathbf{G}^T \quad (1)$$

The key thing to recognize about the cosine measure is that it depends solely on the *local* structure of the graph: the similarities between two words is assessed by looking only at the words to which they are immediately linked.

Our second approach to similarity aims to take into account the overall structure of the entire network graph, and thus to reflect a broader view of the relationship between two nodes. This measure, similar to Leicht, Holme, and Newman (2006), is an example of a "random walk" approach to assessing similarity (see Kemeny & Snell, 1976; Van Dongen, 2000; Griffiths, Steyvers, & Firl, 2007, for related measures). In general terms, the idea is quite similar to the classical notion of spreading activation (e.g. Quillian, 1968). Similarity is thought to be related to the the number and length of the paths through the network that connect two nodes. If there are a lot of short paths that connect two nodes, then it is easy for a random walk through the graph to start at one node and end at the other; these nodes are therefore more similar. Formally, the measure is specified by beginning with the weighted adjacency matrix  $\mathbf{A}$ . This time, however, we normalize the rows so that each one expresses a probability distribution over words. That is, we use the matrix  $\mathbf{P}$  where  $p_{ij} = a_{ij} / \sum_j a_{ij}$ , and then calculate

$$\mathbf{S} = (\mathbf{I} - \alpha\mathbf{P}^{-1}) \quad (2)$$

where  $\mathbf{I}$  is a diagonal identity matrix and the  $\alpha$  parameter governs the extent to which similarity scores are dominated by short paths or by longer paths. A path of length  $r$  is assigned a weight of  $\alpha^r$ , so when  $\alpha < 1$ , longer paths get less weight than shorter ones.<sup>3</sup> Note that under this approach the

<sup>3</sup>As noted by Minkov (2008), this kind of mechanism can help avoid one of the major criticisms of the spreading activation mech-

similarities can be asymmetric (i.e.,  $s_{ij} = s_{ji}$ ). Since our experimental design forces the empirical similarities to be symmetric we use the average of  $\mathbf{S}$  and  $\mathbf{S}^T$  in our evaluations. Interestingly, our approach is very similar to the PageRank measure:  $\mathbf{X} = (\mathbf{I} - \alpha \mathbf{P}^{-1})\mathbf{1}$ . For PageRank it is standard practice to set  $\alpha$  to a fixed value of 0.85 (Page, Brin, Motwani, & Winograd, 1998), where  $\alpha$  is bounded between 0 and 1. Our choice of  $\alpha$  was 0.6 and represents a reasonable trade-off between some degree of decay and a non-trivial contribution of longer paths.<sup>4</sup>

For both measures the similarity indices for each triplet are normalized to sum to 1. This allows the model predictions to be directly comparable to the empirical choice probabilities, which also sum to 1.

## Evaluating the similarity measures

In order to assess whether the semantic network based measures of similarity are capable of capturing the pattern of weak similarities we observed in our experiments, it is first necessary to construct a semantic network. In other words, we must determine the weighted adjacency matrix  $\mathbf{A}$  from which our measures are derived. We constructed this network from a large dataset of word associations consisting of 12,571 cues and over 3 million responses. The data come from a task in which participants were given a short list of cue words and asked to generate three different responses to each single cue (see De Deyne & Storms, 2008b; De Deyne et al., 2011). From this data set we constructed two different weighted directed adjacency matrices. The graph  $\mathbf{A}_1$  only counts the *first* response given by the participant, whereas  $\mathbf{A}_3$  counts all three responses. The graph based on  $\mathbf{A}_1$  is the more conventional approach, and its sparsity is comparable with previous word association studies (Nelson, McEvoy, & Schreiber, 2004). Because it is based on more responses  $\mathbf{A}_3$  is somewhat denser, but in both cases the graphs were quite sparse. The graph  $\mathbf{A}_1$  included 11,969 nodes and only 0.416% of the possible links, whereas  $\mathbf{A}_3$  included 12,420 nodes and 1.176% of possible links.

To evaluate how well the weak similarities from our experiments can be captured using the semantic network models, we calculated the Spearman rank order correlations between the network-derived similarities and the empirical data. The results, summarized in Table 1, demonstrate that both measures of similarity are significantly correlated with the empirical data. As one might expect, the more global measure of similarity (the random walk measure) performs considerably better than the local cosine measure; and the richer network ( $\mathbf{A}_3$ ) tended to produce higher correlations than the network based on less data. Taken together, the general finding is that the more data one uses to define the network, and the more that the similarity measure takes account of the structure in that network, the better one is able to capture human intuitions about weak semantic similarity.<sup>5</sup>

anism, namely the fact that the entire network is quickly activated (e.g. Ratcliff & McKoon, 1994).

<sup>4</sup>Other values of  $\alpha$  were tried as well, but did not substantially change the pattern of results of our experiments

<sup>5</sup>Within the human data from Experiment 1, there are 28 triplets that did not share a single first association in our semantic network, and 72 that did. Because we were concerned that these results might simply be capturing this difference, we re-calculated the correlations

Table 1: Spearman rank order correlations ( $\rho$ ) between the graph-derived similarities and the empirical similarities from both experiments. All correlations are significant at  $p < .001$ , indicated by the double stars. The more global measure of similarity (random walk) consistently outperforms the more local measure (cosine), and that the correlations are stronger for the denser network (i.e.,  $\mathbf{A}_3$ ).

Graph	Cosine		Random walk	
	Exp 1	Exp 2	Exp 1	Exp 2
$\mathbf{A}_1$	.19**	.22**	.48**	.49**
$\mathbf{A}_3$	.38**	.37**	.55**	.57**

For Experiment 2, we can extend the analysis to see if the network measures can account for decision latencies as well. In general, one would expect that more difficult pairs should result in longer decision latencies. For each pair, we calculated the absolute similarity of the strongest pair and compared it with the decision latency of that pair. Restricting our results to the random walk measure of similarity, we found a significant correlation between network-based similarities and decision latencies ( $\rho = -.22$  for the  $\mathbf{A}_1$  network, and  $\rho = -.24$  for the  $\mathbf{A}_3$  network,  $p < .05$  in both cases). This is again consistent with the hypothesis that the semantic network encodes at least some information used to derive weak similarity.

## Discussion

The work in this paper demonstrates that there is substantial agreement between people about the similarity structure of even weakly related items, like HAIL and TEACHER or RAINBOW and TUNAFISH. Moreover, at least some of this agreement can be accounted for by semantic networks constructed from word association data.

The most striking thing about this finding is that there is any agreement about weak similarity at all. In the abstract, there appears to be very little in common between any three items that are randomly thrown together, and it is not an obvious conclusion that people would agree on how they are related. In practice, many people have strong intuitions about any given triplet, just as two of the authors of this paper had strong intuitions about CUP, TEACHER, and HAIL. Two aspects of this are most intriguing. First, there isn't always agreement about these intuitions (just as one author thought TEACHER was the obvious odd one out, and one thought it should be HAIL). Second, as the data from our two experiments show, there is nevertheless substantial agreement (nobody thought CUP should be the odd one out).

The main question we are left with is *why* people should agree on something like this. There is almost certainly no external pressure in the environment to do so; it is difficult to think of any situations in which random unrelated things are thrown together or used, and people must agree with each other about them without communicating explicitly. Rather,

separately for these two subsets of the data. The results did not differ in any substantive way from those reported in Table 1. Interestingly, 27 out of the 28 strongest pairs from these zero-overlap triplets were agreed upon by the human observers more than one would expect by chance. This amount of agreement was similar in Experiment 2, in which 25 of the strongest pairs from 28 triplets were agreed upon more than chance would predict.

such agreement probably stems from commonalities in the shared representations underlying the concepts. But what are those shared representations, and why should they exist at all? It is clear why it would be useful to represent similarities between entities that commonly co-occur, or that are often thought about together – but what benefit is there to building a representation that will probably never be used, and why do people seem to build similar ones?

Part of the answer to these questions may come from our analyses showing that semantic networks built from word associations can account for at least some of the agreement between people. This suggests that perhaps the shared representations measured in our weak similarity task don't occur because they offer some benefit, but rather occur as a by-product of the fact that the mind represents other things. In this case, it is interesting that networks formed from word associations capture some of those other things. We can be somewhat assured that the agreement accounted for by the networks is not the result of trivial or superficial similarities, since denser networks did better and things like frequency and imageability of the words was controlled for. Rather, it may be that these networks capture, at least to some extent, the kind of deep ontological similarities and abstract relationships that drove our intuitions about triples like CUP, TEACHER, and HAIL.

In light of this possibility, there are a number of areas that would be interesting to explore in future work. While our networks did account for a significant amount of the variance in people's weak similarity ratings, a substantial amount remains without explanation. One possibility for this is that our networks, despite being constructed from 12,000 associations, are still almost certainly much sparser and under-specified than people's actual semantic networks. Indeed, we found that the denser network constructed from more associations accounted for the data better. How much improvement is possible with increasingly dense networks and more items and associations? That is, to what extent is a large part of the variance in weak similarity ratings due to the same thing underlying the associations people make in word association tasks? How would this compare to networks constructed in other ways, like co-occurrence in language? How would this change if the networks were constructed in a more robust way, for instance, addressing the sparsity problem by inferring missing links, as in Miller, Griffiths, and Jordan (2009)? Is performance better or worse for different kinds of words, like abstract vs concrete? Work on all of these questions will help us to address the fundamental issue of what kind of semantic representation humans have – and how that representation underlies people's ability to estimate weak similarity.

## Acknowledgments

This work was supported by a research grant funded by the Research Foundation - Flanders (FWO) to the first author and by the interdisciplinary research project IDO/07/002 awarded to Dirk Speelman, Dirk Geeraerts, and Gert Storms. Special thanks to Dinis Gökaydin and Steven Verheyen for helpful comments.

## References

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database [CD-ROM]*. Philadelphia: University of Pennsylvania, Linguistic Data Consortium. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

- Bohn, R. E., & Short, J. E. (2009). *How much information? 2009. Report on American Consumers* (Tech. Rep.). Global Information Industry Center. University of California, San Diego.
- De Deyne, S., & Storms, G. (2008a). Word Associations: Network and Semantic properties. *Behavior Research Methods*, 40, 213-231.
- De Deyne, S., & Storms, G. (2008b). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, 40, 198-205.
- De Deyne, S., Voorspoels, W., Verheyen, S., Navarro, D., & Storms, G. (2011). Graded structure in adjective categories. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (p. 1834-1839). Austin, TX: Cognitive Science Society.
- Goodman, N. (1972). Problems and projects. In N. Goodman (Ed.), (p. 437-447). New York: Bobbs-Merrill.
- Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the Mind. *Psychological Science*, 18, 1069-1076.
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114, 211-244.
- Hutchinson, J. (1989). NETSCAL: A network scaling algorithm for nonsymmetric proximity data. *Psychometrika*, 54, 25-51.
- Kemeny, J., & Snell, J. (1976). *Finite markov chains*. Springer-verlag.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211-240.
- Leicht, E., Holme, P., & Newman, M. (2006). Vertex similarity in networks. *Psychical Review E*, 73, 026120.
- Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100, 254-278.
- Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (p. 179-195). New York: Cambridge University Press.
- Miller, K. T., Griffiths, T. L., & Jordan, M. I. (2009). Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems* (Vol. 22, p. 1276-1284).
- Minkov, E. (2008). *Adaptive graph walk based similarity measures in entity-relation graphs*. Unpublished doctoral dissertation, School of Computer Science Carnegie Mellon University, Pittsburgh, PA 15213.
- Navarro, D. J., & Lee, M. D. (2002). Commonalities and distinctions in featural stimulus representations. In W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (Vol. 24, p. 685-690). Mahwah, NJ: Lawrence Erlbaum.
- Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, and Computers*, 36, 402-407.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The pagerank citation ranking: Bringing order to the web*. (Tech. Rep.). Computer Science Department, Stanford University.
- Quillian, M. (1968). Semantic information processing. In M. Minsky (Ed.), (p. 227-270). Cambridge, MA: MIT Press.
- Ratcliff, R., & McKoon, G. (1994). Retrieving information from memory: Spreading-activation theories versus compound-cue theories. *Psychological Review*, 101, 177-184.
- Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behaviour Research Methods, Instruments, and Computers*, 36, 506-515.
- Sloman, S. A., & Rips, L. J. (1998). Similarity as an explanatory construct. *Cognition*, 65, 87-101.
- Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Experimental Cognitive Psychology and its Applications. In A. Healy (Ed.), (chap. Word association Spaces for Predicting Semantic Similarity Effects in Episodic Memory.). Washington, DC: American Psychological Association.
- Van Dongen, S. (2000). *Graph clustering by flow simulation*. Unpublished doctoral dissertation, University of Utrecht.

# Conceptual Event Units of Putting and Taking in Two Unrelated Languages

**Rebecca Defina (rebecca.defina@mpi.nl)**

Max Planck Institute for Psycholinguistics &  
International Max Planck Research School for Language Sciences, 6500AH Nijmegen, The Netherlands

**Asifa Majid (asifa.majid@mpi.nl)**

Max Planck Institute for Psycholinguistics, 6500AH Nijmegen, The Netherlands  
Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

## Abstract

People automatically chunk ongoing dynamic events into discrete units. This paper investigates whether linguistic structure is a factor in this process. We test the claim that describing an event with a serial verb construction will influence a speaker's conceptual event structure. The grammar of Avatime (a Kwa language spoken in Ghana) requires its speakers to describe some, but not all, placement events using a serial verb construction which also encodes the preceding taking event. We tested Avatime and English speakers' recognition memory for putting and taking events. Avatime speakers were more likely to falsely recognize putting and taking events from episodes associated with take-put serial verb constructions than from episodes associated with other constructions. English speakers showed no difference in false recognitions between episode types. This demonstrates that memory for episodes is related to the type of language used; and, moreover, across languages different conceptual representations are formed for the same physical episode, paralleling habitual linguistic practices.

**Keywords:** Conceptual event units; event segmentation; serial verb constructions; linguistic relativity.

## Introduction

Events occur in a continuous stream with no clear boundaries between them. Despite this continuity, we think and talk about events in terms of discrete and divisible units. Previous research has largely focused on the factors influencing the segmentation of events. This paper examines the question from a complementary perspective: what factors might lead event elements to be grouped together into a single conceptual event unit.

When we perceive ongoing activity, we segment it automatically and unconsciously (Kurby & Zacks, 2008; Zacks et al., 2001a). The conceptual event units thus created are structured hierarchically. Each event unit is made up of smaller units, which in turn combine to form larger units (Zacks, Tversky, & Iyer, 2001b). So, what counts as a single conceptual event unit depends to some extent on which level of granularity we are talking about. The choice of granularity level appears to be made at the point of reporting. Prior to that, people segment events at multiple levels of granularity simultaneously (Zacks, Speer, Swallow, Braver, & Reynolds, 2007).

Previous research shows that event units are determined by at least three main factors. First, the inherent properties

of events, such as points of greater motion, have a large effect on where event boundaries are placed (Newton, Engquist, & Bois, 1977; Zacks, 2004). Second, repeated co-occurrence, particularly in different contexts, encourages event elements to be grouped together, regardless of their inherent properties (Avrahami & Karev, 1994). Finally, the particular event schema that the person engages for an event affects the way they segment it (Zacks et al., 2007); for instance, whether or not a person understood the actor's goal influences the way a participant segments the actor's behavior (Zacks, 2004). The fact that event schemas influence conceptual event structure suggests that language may also play a role here. This paper explores this possibility.

Previous cross-linguistic research on the role of language in event cognition has largely focused on differences in the encoding of manner and path in motion events. The results have been mixed: Some studies have found language effects (e.g., Filipović, 2011; Finkbeiner, Nicol, Greth, & Nakamura, 2002; Kersten et al., 2010), but others have not (e.g., Gennari, Sloman, Malt, & Fitch, 2002; Loucks & Pederson, 2011; Papafragou, Massey, & Gleitman, 2002). More recently, scholars have begun to explore other aspects of language and how they might influence event cognition, particularly with respect to causal actions (e.g., Fausey & Boroditsky, 2011; Wolff, Jeon, & Li, 2009). For example, Wolff et al. (2009) found that the semantic property of whether or not a language allowed an intermediary actor to function as an agent affected both the syntactic and non-linguistic partitioning of events, consistent with the proposal that language may play a role in event segmentation.

In Wolff et al.'s (2009) study both the semantic and the syntactic differences are potential instigators of the non-linguistic event segmentation patterns. The current study narrows in on the potential link between syntactic encoding in particular and the concomitant non-linguistic partitioning of events.

One type of syntactic structure that is particularly interesting for event cognition is serial verb constructions (SVCs). These constructions allow multiple verbs to be placed within a single clause without coordination or subordination (Aikhenvald, 2006; Durie, 1997). The particular syntactic features vary across languages, though there is a shared set of core, prototypical features

(Aikhenvald, 2006; Foley, 2010). Though generally absent in European languages, SVCs are common cross-linguistically. Some languages have a particularly high rate of SVC use and these are called serializing languages. Languages with no SVCs, such as English, are called non-serializing languages.

It has been claimed that SVCs always refer to conceptualizations of a single event (Aikhenvald, 2006; Comrie, 1995). Take the examples below, from Avatime, a Ghana-Togo Mountain language from the Kwa branch of the Niger-Congo language family. The SVC in example (1a) describes what appears to be a single event: a man cutting firewood with the axe he picked up for that purpose. In contrast, the two Avatime simple, single verb clauses in example (1b) describe a less integrated scene of a man picking up an axe (maybe not with the immediate or sole purpose of cutting firewood) and then cutting firewood (not even necessarily with the axe just mentioned).

1. (a) *A-kò kàwɛ-à tsǎ ɪnyɪ-nɛ.*  
3S<sup>1</sup>-take axe-DEF cut firewood-DEF  
'He cut the firewood with the axe.'
- (b) *A-kò kàwɛ-à. A-tsǎ ɪnyɪ-nɛ.*  
3S-take axe-DEF 3S- cut firewood-DEF  
'He took the axe. He cut the firewood.'

While there is a strong feeling among linguists that SVCs should – and do – refer to single conceptual event units (Aikhenvald, 2006; Comrie, 1995; Durie, 1997), the relationship has not been directly tested.

The best evidence for this relationship, to date, comes from a study conducted by Givón (1990, 1991). He tested conceptual event units by investigating the production process. Speakers pause when they are encoding the next unit of speech (Goldman-Eisler, 1968). Givón thus took pauses in speech as an indication of conceptual cohesion: speech that was encoded together, and so between pauses, was taken to refer to single event units. The frequencies of speech internal pauses in different clause types were compared across three languages of Papua New Guinea (Kalam, Tairora and Tok Pisin), which use verb serialization to different degrees. Givón found that pauses were no more frequent within SVCs than they were within simple clauses with a single verb. From this, he concluded that SVCs and simple clauses both refer to single conceptual event units (contra Pawley, 1987). Note that this study only tests chunking at the linguistic level. It does not provide evidence about cognitive event segmentation. To do that, an independent test of conceptual event structure is required.

The present study aims to conduct just such an independent test. It focuses on placement events in the

serializing language Avatime, to test the following two hypotheses: 1) that SVCs correspond to single conceptual events and 2) that differences in linguistic descriptions of events correlate with differences in conceptual event units.

In Avatime, most placement actions, like putting a cup on a table, or a banana in a basket, must be described using both a take verb and a put verb in an SVC<sup>2</sup>, as in example (2). The grammar of the language requires speakers to encode the taking part as well as the placing part of the event, even if the person only saw the placing. Note that it is logically necessary for an object that is being placed to have been taken at some earlier point in time. So, it is not as strange as it may at first appear for a language to require the preceding taking event to also be encoded. The same construction is also used to describe cases when both the taking and placing events are seen. So an alternative interpretation of (2) is: *S/he took the banana and put it in the basket.*

2. *A-kò kòrantì-ɛ kpe ní kàsɔ-yà mè*  
3S-take banana-DEF put LOC basket-DEF inside  
'S/he put the banana into the basket.'

There is a small set of placement events that are described without a take-put SVC. These exceptional events include putting an article of clothing or jewelry on a body part (in its canonical location), and pouring liquids. These are described using either a put verb in a simple clause (3) or a put verb combined with a pouring manner verb in an SVC (4). It is strongly dispreferred to describe such actions using an SVC with a take verb.

3. *A-kpe likùto-lè*  
3S-put hat-DEF  
'S/he put the hat on.'
4. *E-nyi kùni-ò kpe ní kèzi-à mè*  
3S-pour water-DEF put LOC bowl-DEF inside  
'S/he poured the water into the bowl.'

The patterns of placement event descriptions in Avatime, and the claim that SVCs describe single conceptual events, lend themselves to experimental testing. Previous research has shown that people mentally fill in parts of event units that they have not actually seen (Strickland & Keil, 2011). We can build on this finding to test whether Avatime speakers treat take-put episodes as single event units. Specifically, if Avatime speakers see a videoclip showing a

<sup>1</sup> Abbreviations used: 3 '3<sup>rd</sup> person', DEF 'definite', LOC 'locative', S 'singular', ` 'low tone', ^ 'high tone', mid tone is unmarked.

<sup>2</sup> As with many languages with this type of construction, the take verb acts like an object marker and allows the two objects (thing placed, and location where placed) to be expressed (Lord, 1993). However, unlike some languages, in Avatime the take verb still maintains much of its original lexical semantics in these cases. Different take verbs will even be used to mark differences in the type of taking done.

general placement action, which they would describe using a take-put SVC, they should be more likely to falsely recognize a corresponding taking action. In contrast, if Avatime speakers see a videoclip showing a placement action, which they would not describe with a take-put SVC, such as putting on clothing or pouring a liquid, they should not falsely recognize a corresponding taking action.

To control for the possibility that putting events and their corresponding taking events are generally more cohesive than the donning of clothing or pouring of liquids and their corresponding taking events, we tested a control group of English speakers. English speakers describe general placement events with a single put verb which takes the thing moved as the object and the location as a prepositional phrase. For instance, *She put the book on the table*. The pouring of liquids is described using the same structure as general placements, but the verb is specific to pouring. For instance, *He poured water into the glass*. The putting on of clothing and jewelry is described using essentially the same structure but the location is often not expressed. For instance, *She put the necklace on*. There are no cases where the grammar of English requires the corresponding taking event to also be encoded. Hence, English speakers are not predicted to have differences in false recognition rates to these take events.

## Methods

### Participants

Thirty-four native speakers of Avatime, aged 11-16 (mean 14.1 years), were recruited at Vane Junior High School, Ghana. Four Avatime speakers were tested but excluded due to technical difficulties or for consistently answering either yes or no for all items. Thirty-three native speakers of English, aged 11-17 (mean 14.2 years), were recruited in the Blue Mountains and Sydney, NSW, Australia.

All Avatime speakers were fluent in Ewe and English and 11 additionally spoke Twi. One English speaker was also fluent in German, two spoke Spanish, one fluently and the other moderately. Of the remaining English speakers, 9 were completely monolingual and 21 had very limited knowledge of another language (French, German, Italian, Japanese, Korean or Latin).

### Materials

80 paired putting and taking events were filmed in a single location inside the Max Planck Institute, Nijmegen. They were acted out by two Dutch university students, one male and one female. Each videoclip lasted 3-4 seconds.

A paired putting and taking episode showed the same actor removing an object from one location and placing it in another. For instance, in Figure 1(a) a man takes a banana from the shelf and places it on a plate, in Figure 1(b) a woman takes a necklace from a bag and places it on her

neck. Across episodes, the camera angle and position of the actor in the room were kept constant.



Figure 1: Sample frames from the two videoclips (a) ‘man takes banana from shelf’ and ‘man puts banana on plate’; (b) ‘woman takes necklace from bag’, and ‘woman puts necklace on.’

Objects and locations were selected so as to be familiar to both Avatime and English speakers. The source location of the taking event was always different from the goal location of the putting event. Across episodes, the object, locations, position of the actors, and camera angle varied.

Of the 40 episodes, half had general placement events of the type described using take-put SVCs in Avatime, while the other half did not (the donning of clothing and pouring). Descriptions of the items by Avatime participants at the end of the experiment confirmed this distinction: The placement events in the SVC category were described using take-put serial verb constructions 96.2% of the time ( $SD = 1.8$ ). The placement events in the Non-SVC category were described using take-put serial verb constructions 6.5% of the time ( $SD = 1.7$ ). For ease of reference, both the putting and taking events in an episode will be referred to as either SVC or Non-SVC according to the type of putting event.

### Design

The 40 put and take episodes, resulted in 80 individual items, each consisting of a sole put event or take event. The 80 items were divided into two sets: only one part of a put-take episode featured in each set. Pilot testing with Avatime speakers showed that remembering all 40 learning items in one go was too difficult, so testing was divided into two blocks. In each block of a given set, there were 5 SVC put events, 5 Non-SVC put events, 5 SVC take events, and 5 Non-SVC take events. Blocks were counterbalanced across participants. Within each block, items appeared in one of four random orders.

### Procedure

Participants were asked to watch a series of videoclips and to remember them as best they could. They were told that



they would later be shown more videos, some exactly the same as the ones they had seen and some different, and that their task was to tell the experimenter which videoclips were the same and which were not.

Participants watched videoclips one at a time. The videoclips were separated by a black screen lasting 1 second. After the learning phase, there was a 5 minute distraction task unrelated to the experiment. Participants were then tested for their memory of the 20 videoclips they had just seen, plus their 20 unseen counterparts. So, if a participant saw a girl put on a necklace in the learning phase, they now, in the testing phase, also saw the girl taking the necklace out of the bag. Participants indicated whether each event was the same or different to the events they had seen previously. After finishing testing for the first block, participants saw the second block of 20 items and were tested for memory of those as described above.

After completing the memory experiment, participants viewed all the videoclips again and were asked to describe "what the person did".

Avatime instructions were translated by a native Avatime speaker fluent in English in consultation with the experimenter. Instructions and responses were given verbally in the participant's native language. Responses were recorded using an Olympus LS-10 flash recorder with a headset microphone.

Participants were tested individually and the same procedure was used for English and Avatime participants. The whole experiment lasted approximately 45 minutes.

## Results

Responses to seen and new items were analyzed separately using 2 construction-type (SVC or Non-SVC) x 2 event-type (put or take) x 2 language (Avatime or English) x 2 block order (AB or BA) mixed ANOVAs, with construction and event type being within-participant factors, and language and block order between-participant factors. The dependent variable was the number of reported recognitions. Block order was not significant for seen ( $F(1,59) = 0.62, p = 0.43, \eta_p^2 = 0.01$ ) or new items ( $F(1,59) < 0.01, p = 0.94, \eta_p^2 = 0.01$ ), so we collapsed over this factor.

We first tested whether participants were able to correctly recognize the items they had seen. The overall accuracy was 80.7% for Avatime speakers and 83.6% for English speakers. The difference between language groups was not significant,  $F(1, 61) = 0.92, p = 0.34, \eta_p^2 = 0.02$ . There was a main effect of event-type,  $F(1,61) = 9.20, p < 0.01, \eta_p^2 = 0.13$ . Putting events were remembered more accurately ( $M = 8.50$ ) than taking events ( $M = 7.92$ ). There was also a main effect of construction,  $F(1,61) = 9.81, p < 0.01, \eta_p^2 = 0.14$ , and a just significant interaction between construction-type and language  $F(1,61) = 3.94, p = 0.05, \eta_p^2 = 0.06$ . English speakers remembered Non-SVC events more accurately ( $M = 8.73$ ) than SVC events ( $M = 7.99$ ). Avatime speakers showed no difference in recognition between

previously seen SVC events ( $M = 7.98$ ) and Non-SVC events ( $M = 8.15$ ). There were no other interactions.

Our hypothesis concerned false recognitions to previously unseen or new items. It was predicted that there would be a three-way interaction between construction-type, event-type and language. Avatime speakers would have more false recognitions for taking events if the corresponding put event was one that they would describe using a take-put serial verb construction. English speakers should have the same rates of false recognition for SVC and Non-SVC type events. There was no statistically significant 3-way interaction,  $F(1,61) = 0.01, p = 0.92, \eta_p^2 < 0.01$ . However, there was a main effect of language,  $F(1,61) = 14.34, p < 0.01, \eta_p^2 = 0.19$ . Avatime speakers, in general, had more false recognitions ( $M = 2.83$ ) than English speakers ( $M = 1.58$ ). There was also a main effect of construction-type  $F(1,61) = 4.36, p = 0.04, \eta_p^2 = 0.07$ . SVC events had more false recognitions ( $M = 2.37$ ) than Non-SVC events ( $M = 2.04$ ). More interestingly, there was a significant interaction between language and construction-type,  $F(1,61) = 4.36, p = 0.04, \eta_p^2 = 0.07$ , see Figure 2. Avatime speakers had more false recognitions for SVC type events in general ( $M = 3.17$ ) than for Non-SVC type events ( $M = 2.50$ ). English speakers, on the other hand, had the same false recognition rates for SVC ( $M = 1.58$ ) and Non-SVC events ( $M = 1.58$ ). This suggests that Avatime speakers remember events described with SVCs differently to those which are not; and that this effect is not due to properties of the events themselves, since English speakers fail to show a difference across these event types.

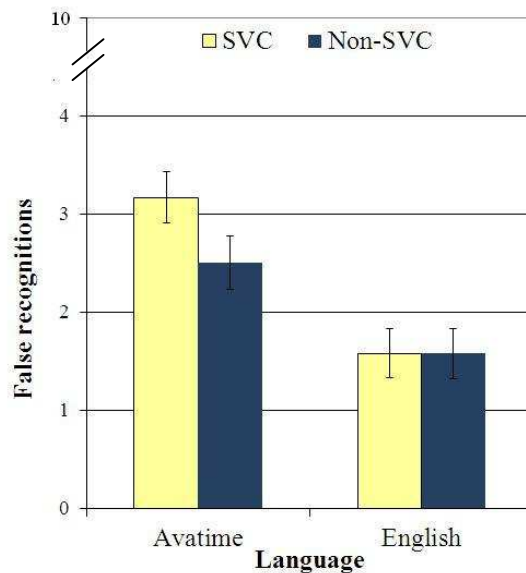


Figure 2: Average false recognitions as a function of language and construction type.

Finally, there was an unpredicted interaction between language and event type,  $F(1,61) = 4.51, p = 0.04, \eta_p^2 =$



0.07. Avatime speakers showed more false recognitions for put events ( $M = 3.08$ ) than take events ( $M = 2.58$ ). In contrast, English speakers had slightly more false recognitions for take events ( $M = 1.71$ ) than for put events ( $M = 1.44$ ). There were no other significant interactions.

## Discussion

Avatime speakers, but not English speakers, displayed more false recognitions for put and take events from SVC episodes than from the equivalent events in non-SVC episodes. This is consistent with the suggestion that language may play a role in conceptual event structure. Avatime speakers appear to construct a single conceptual event unit that includes the taking and putting event segments precisely when the putting event is one that they would describe with a take-put SVC.

Our initial prediction was that false recognitions would occur only with the take part of the episode. This was because it is the put event that determines whether or not an SVC is used, and it is this use of the SVC which is predicted to determine whether or not the take action is included in the event unit. For example, picking up a necklace should only be combined with its corresponding putting event if the putting event is something like putting the necklace in a bag, not putting it around your neck. Our results show, however, that Avatime speakers falsely recognize both the take and put parts of the SVC episodes regardless of which part they saw first. This indicates that as soon as both parts have been seen and understood to form an SVC type episode, Avatime speakers join the taking and putting actions together into a single conceptual event unit. Familiarity with either part then spreads to the unit as a whole, resulting in false recognition of the unseen part, be it a putting or a taking action.

These results show a correlation between conceptual event units and linguistic structure, but from these results alone we cannot say whether language influences conceptual structure, conceptual structure influences language, or whether some third factor is involved. There is some ancillary evidence that language may play a causal role here. For example, Trueswell and Papafragou (2010) found that people under high cognitive load directed attention to event elements considered important in their language; while Zacks et al. (2001b) found that there was greater alignment between fine- and coarse-level segmentation when speakers described events while they segmented them, rather than describing them later. To determine whether linguistic encoding is critically involved in this experiment further study would be needed.

The link between SVCs and single conceptual event units has often been suggested as a definitional criterion for SVCs (Aikhenvald, 2006; Comrie, 1995; Durie, 1997). This paper provides the first language external evidence in favor of this often cited relationship. However, SVCs were only compared with sets of separate clauses. To determine

whether or not the link between SVCs and single conceptual event units is useful as a definitional criterion, SVCs should be compared to other types of complex clauses as well. This paper has shown that testing recognition memory is a viable method for investigating the relationship between conceptual event units and syntactic structures.

In addition to the main result discussed above, we found three other effects. 1) Putting events were remembered more accurately than taking events by speakers of both languages. This is in line with predictions based on the asymmetry of sources and goals in motion events (Regier & Zheng, 2007; Papafragou, 2010) and research concerning put and take lexicons (Narasimhan, Kopecka, Bowerman, Gullberg, & Majid, in press; Regier, 1995). 2) Avatime speakers displayed more false recognitions for put events than for take events while English speakers showed the reverse pattern. This is not immediately interpretable and will require further investigation. 3) English speakers remembered both putting and taking events from Non-SVC episodes more accurately than those from SVC episodes. This shows that there may be differences between the episode types which are noticeable by English speakers, even though they do not use SVCs. It seems likely that actions involving clothing as well as pouring actions could be more salient than general taking and putting actions. Although English speakers were sensitive to the distinction between episode types, they nevertheless performed equivalently with respect to memory for new events. So although there may be differences between SVC and Non-SVC episodes these differences alone cannot predict our final results.

## Conclusion

This study provides the first evidence for the often claimed connection between serial verb constructions and single conceptual event units. It demonstrates that event elements grouped together in language are grouped together as conceptual event units: Avatime speakers conceptualize a take-put episode as a single event unit exactly when the placement event is one they would describe with a take-put SVC but not if it is from a Non-SVC. English speakers, on the other hand, do not distinguish the two types of events in their syntax, nor do they demonstrate greater event cohesion for the events described by take-put SVCs in Avatime. Thus, speakers' event conceptualisations parallel the linguistic structures used to describe those events.

## References

- Aikhenvald, A. (2006). Serial verb constructions in typological perspective. In A. Aikhenvald & R. Dixon (Eds.), *Serial Verb Constructions: A Crosslinguistic Typology*. Oxford: Oxford University Press.
- Avrahami, J., & Kareev, Y. (1994). The emergence of events. *Cognition*, 53, 239-261.

- Comrie, B. (1995). Serial verbs in Haruai (Papua New Guinea) and their theoretical implications. In J. Bouscaren, J. Franckel, & S. Robert (Eds.), *Langues et langage: Problèmes et raisonnement en linguistique, mélanges offerts à Antoine Culioli*. Paris: University Presses of France.
- Durie, M. (1997). Grammatical Structures in Verb Serialization. In A. Alsina, J. Bresnan, & P. Sells (Eds.), *Complex Predicates*. Stanford, CA: CSLI Publications.
- Fausey, C. M., & Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin & Review*, 18(1), 150–157.
- Filipović, L. (2011). Speaking and remembering in one or two languages: bilingual vs. monolingual lexicalisation and memory for motion events. *International Journal of Bilingualism*, 15(4), 466–485.
- Finkbeiner, M., Nicol, J., Greth, D., & Nakamura, K. (2002). The role of language in memory for actions. *Journal of Psycholinguistic Research*, 31(5), 447–457.
- Foley, W. A. (2010). Events and serial verb constructions. In B. Baker & M. Harvey (Eds.), *Complex Predicates: Cross-Linguistic Perspectives on Event Structure*. Cambridge: Cambridge University Press.
- Gennari, S. P., Sloman, S., Malt, B., & Fitch, T. (2002). Motion events in language and cognition, *Cognition*, 83, 49–79.
- Givón, T. (1990). ‘Verb serialization in Tok Pisin and Kalam: a comparative study of temporal packaging.’ In J. Verhaar (Ed.) *Melanesian Pidgin and Tok Pisin*. Amsterdam: Benjamins.
- Givón, T. (1991). Serial verbs and event cognition in Kalam: an empirical study of cultural relativity. In C. Lefebvre (Ed.), *Serial verbs: grammatical, comparative and universal grammar*. Amsterdam: John Benjamins.
- Goldman-Eisler, Frieda. 1968. *Psycholinguistics: experiments in spontaneous speech*. New York: Academic Press.
- Kersten, A. W., Meissner, C. A., Lechuga, J., Schwartz, B. L., Albrechtsen, J. S., & Iglesias, A. (2010). English Speakers Attend More Strongly Than Spanish Speakers to Manner of Motion When Classifying Novel Objects and Events. *Journal of Experimental Psychology: General*, 139(4), 638–653.
- Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, 12(2), 72–79.
- Lord, C. (1993). *Historical change in serial verb constructions*. Typological studies in language. Amsterdam: Benjamins.
- Loucks, J., & Pederson, E. (2008). Linguistic and non-linguistic categorization of complex motion events. In J. Bohnemeyer & E. Pederson (Eds.), *Event representation in language and cognition*, Language Culture and Cognition. Cambridge: Cambridge University Press.
- Narasimhan, B., Kopecka, A., Bowerman, M., Gullberg, M. & Majid, A. (in press). Putting and taking events: a cross-linguistic perspective. In A. Kopecka & B. Narasimhan (Eds.), *Events of ‘putting’ and ‘taking’: a cross-linguistic perspective*. Amsterdam: John Benjamins.
- Newton, D., Engquist, G., & Bois, J. (1977). The objective basis of behaviour units. *Journal of Personality and Social Psychology*, 35(12), 847–862.
- Papafragou, A. (2010). Source-goal asymmetries in motion representation: Implications for language production and comprehension. *Cognitive Science*, 34, 1064–1092.
- Papafragou, A., Massey, C., & Gleitman, L. (2002). Shake, rattle, “n” roll: the representation of motion in language and cognition. *Cognition*, 84, 189–219.
- Pawley, A. (1987). Encoding events in Kalam and English: different logics for reporting experience. In R. Tomlin (Ed.), *Coherence and grounding in discourse*, Typological Studies in Language. Amsterdam/Philadelphia: John Benjamins.
- Regier, T. (1995). A model of the Human Capacity for Categorizing Spatial Relations. *Cognitive Linguistics*, 6, 63–88.
- Regier, T. & Zheng, M. (2007). Attention to endpoints: A cross-linguistic constraint on spatial meaning. *Cognitive Science* 31,705–719.
- Strickland, B., & Keil, F. (2011). Event completion: Event based inferences distort memory in a matter of seconds. *Cognition*, 121, 409–415.
- Trueswell, J., & Papafragou, A. (2010). Perceiving and remembering events cross-linguistically: Evidence from dual-task paradigms. *Journal of Memory and Language*, 63, 64–82.
- Wolff, P., Jeon, G.-H., & Li, Y. (2009). Causers in English, Korean, and Chinese and the individuation of events. *Language and Cognition*, 1(2), 167–196.
- Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science*, 28, 979–1008.
- Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L. & Raichle, M. E. (2001a). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, 4, 651–655.
- Zacks, J. M., Tversky, B., & Iyer, G. (2001b). Perceiving, remembering and communicating structure in events. *Journal of Experimental Psychology: General*, 130(1), 29–58.
- Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological Bulletin*, 133(2), 273–293.

# Interpersonal Effects of Emotions in Morally-charged Negotiations

Morteza Dehghani (morteza@ict.usc.edu), Jonathan Gratch (gratch@ict.usc.edu)

Institute for Creative Technologies, University of Southern California,  
1205 Waterfront Dr., Playa Vista, CA 90094-2536, USA

Peter J. Carnevale (peter.carnevale@marshall.usc.edu)

Marshall School of Business, University of Southern California,  
3670 Trousdale Parkway, Los Angeles, CA 90089-0808

## Abstract

The majority of research on emotion and moral decision-making has focused on the intrapersonal effects of emotion. However, witnessing and displaying emotional expressions is also known to play a significant role in the facilitation and coordination of our social interactions. In this work, we hypothesize that interpersonal emotions effect moral appraisals by prioritizing different moral concerns. We investigate the impact of facial displays of discrete emotions, specifically anger and sadness, in a morally charged multi-item negotiation task. The results of our experiment support our hypothesis that moral appraisals can be strongly affected by interpersonal emotional expressions. We show that displays of anger may backfire if one of the parties associates moral significance to the objects of the negotiation, whereas displays of sadness promote higher concession-making. Overall, we argue that emotional expressions can shift moral concerns within a negotiation in ways that can promote cooperation.

**Keywords:** Emotion, Sacred Values, Moral Decision-Making, Negotiations

## Introduction

Recent research into emotion and moral decision-making reveals a consistent pattern. When confronted with possible threats to moral or sacred concerns, people tend to become emotional (e.g. Ginges et al., 2007), uncompromising (e.g. Tetlock, 2003), and act in ways contrary to traditional formalizations of rational self-interest (e.g. Atran, 2010). This article adds a bit of hope to this otherwise gloomy picture. Building on findings from both moral decision-making and the interpersonal effects of emotion, we show how emotion can sometimes foster cooperation rather than conflict. Our findings have potential important implications for negotiation and conflict resolution in sacred domains.

In this article, we build on the social-functional framework of emotions (Keltner & Haidt, 1999; Frijda & Masquita, 1994) which claims that different sociomoral concerns are prioritized based on the distinct emotions that are experienced (Horbeg, Oveis & Keltner, 2011). This theory argues that our perception of the permissibility of actions in moral situations is affected by the emotions experienced, as different emotions heighten the salience of different moral domains. For example, disgust has been linked to violations of purity-sancity (Rozin, et al., 1999), and research shows that experimentally predisposing individuals to disgust increases their tendency to focus on

purity related issues (e.g. sexuality) as opposed to other moral concerns such as justice (e.g. Tapias, Glaser, Keltner, Vasquez & Wickens, 2007). However, the research on emotion-related moral appraisals has been mostly limited to the intrapersonal effects of emotion in decision-making.

In contrast to research on moral decision-making, research on negotiation and conflict resolution has largely emphasized both the intrapersonal as well as interpersonal effects of emotion (Carnevale, 2008; Forgas, 1998). Work on interpersonal aspects of emotion argues that emotional expressions by one party can change how the other party construes and reacts to a situation. The evidence from this line of research suggests that cognitive emotional appraisals are not only antecedents of emotions experienced, but they may also follow from the perception of emotional expressions in others (Lerner, Han, & Keltner, 2007). For example, it has been widely documented that perceiving expressed anger during a course of a negotiation can elicit more concessions compared to other emotions (such as happiness) or no emotions at all (e.g. Van Kleef, De Dreu & Manstead, 2004a, 2004b, 2010; Sinaceur & Tiedens, 2006). These authors argue that anger communicates that a party has high aspirations and that concessions would be required from the other party to reach an agreement. The majority of these findings, however, rely on negotiations only involving issues that might be of interest to people but have no sentimental or moral significance to them (e.g. negotiating over the price of a cellphone and its duration of service). In this article, we build on these findings and show that interpersonal effects of emotion unfold somewhat differently in moral contexts.

Here, we investigate the impact of facial displays of discrete emotions, specifically anger and sadness, in a morally-charged multi-item negotiation task. We hypothesize that perceiving different emotional expressions in others can influence moral cognition by prioritizing different moral domains and shifting interpretive-frames. In other words, our interpretation of a moral issue can be subjected to the emotional expressions conveyed by other individuals involved in the negotiation. Anger is connotated with the prioritization of ethics of autonomy concerned with rights and justice (Rozin, et al., 1999), and sadness is linked to eliciting sympathy and heightening the salience of need, weakness and harm/care (Horbeg, Oveis & Keltner, 2011). We predict that when an object is perceived as a sacred (or protected) value (with intrinsic moral significance) (Tetlock,

2003; Baron & Spranca, 1997), angry or sad facial displays expressed by an opponent will have opposing effects on the behavior of individuals, as these emotions heighten different moral concerns. Recent work in social and cognitive psychology suggests that people with sacred values (SVs) tend to reject tradeoffs with other values (especially with secular ones) and will express anger when considering such tradeoffs (e.g. Tetlock et al. 2000).

We hypothesize that when facing angry opponents, SV participants (those associating moral significance to the item) will show the typical rejection of tradeoffs and concede very little, as their concerns about their sacred values will be amplified by the anger expressed in the other party. In other words, we predict perceiving expressed anger will back-fire for SV participants, that is, will lessen the likelihood of concession. However, when interacting with an opponent who displays sadness, concerns about need and care will become salient, and participants will concede more than SV participants in the anger condition. For non-SV participants we expect to see the known pattern of concession due to perceived anger consistent with findings of Van Kleef and others in non-moral domains (e.g. Van Kleef et al., 2004a, 2004b; Sinaceur & Tiedens, 2006).

We begin by discussing the negotiation task used in this study. Then we explain our hypotheses, and describe our experiments. We close with a discussion of our findings and its implications.

### Sacred-Objects Negotiation Task

It has been argued that research on morally motivated decision-making relies heavily on a “narrow empirical base”, in regards to subject populations as well as the

stimuli used in experiments (Medin & Atran, 2004). The research populations used in these studies mostly consist of undergraduates at major research universities. And the scenarios and the stimuli materials used mainly focus on single-shot trade-offs scenarios where participants are asked about the permissibility of set hypothetical actions (e.g. killing one person instead of five). However, many real-life moral situations unfold over repeated interactions (which can sometimes span years), such as socio-political conflicts involving sacred values (e.g. Israel-Palestine conflict: Ginges et al., 2007; Iran nuclear conflict: Dehghani, et al., 2009, 2010).

In order to overcome some of the above shortcomings, we have recently developed a new web-based multi-round negotiation task involving a participant and an opponent (computer agent), where different objects are placed on a board and the participant and the agent take turns in taking ownership of some of the items and giving away the others (Carnevale et al., 2011) (Figure 1). Participants can move items around the board by grabbing them with a mouse and putting the items either on their own side or on the opponent's side. After each new proposal is extended by the participants, the agent evaluates the offer, expresses an emotional reaction to the offer and decides whether or not to accept the offer or propose a new offer. Participants can express emotional reactions at any point by choosing one the emotional facial displays at the bottom right corner of the screen (Figure 1).

Aspects of this task are easily configurable in order to consider a variety of experiment questions. In the context of this article, all items are initially placed in the middle section of the board and are up for grabs. The negotiation

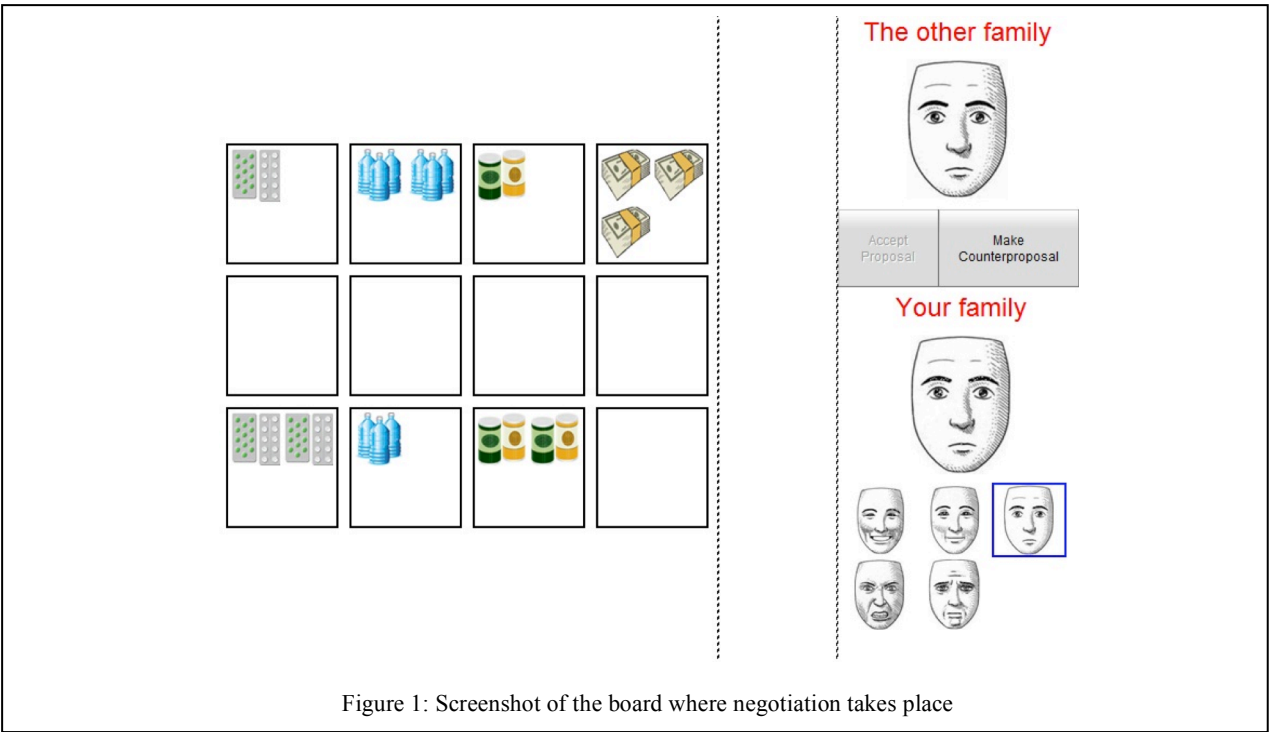
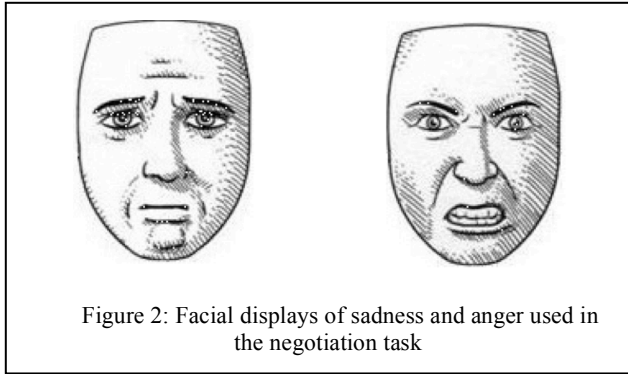


Figure 1: Screenshot of the board where negotiation takes place



consists of 12 rounds with each player taking turns making or receiving offers six times. When a participant makes an offer, the computer opponent decides to accept or reject the offer based on a pre-programmed strategy unknown to the participant. If the opponent decides to reject the offer, it will make a new proposal that the participant can in turn accept or reject.

### Agent Offers

All agents in this study follow the same strategy -- a fixed, non-contingent series of offers that has been designed to simulate resistance to tradeoffs involving sacred values. In pilot testing, most participants perceived this policy to be tough but plausible. There are four different groups of items involved in the negotiations (medicine, water bottles, food cans, money), with three items per group. The negotiation strategy of the agent is as follows ([medicine, water, food, money]): Round 2: [0, 0, 0, 0]; Round 4: [0, 0, 0, 1]; Round 6: [0, 0, 0, 1]; Round 8: [0, 1, 0, 1]; Round 10: [0, 1, 0, 1]; Round 12: [0, 2, 0, 2], where the numbers in the brackets represents how many items in each group the agent chooses to give to the participant. In the decision-making algorithm of the agent, the items are given the following qualitative payoff values: [50, 10, 5, 1]. These payoff values are only used for internal calculations and are not shown to the participants. The agent will accept a participants' offer if it has a higher or equal overall utility than the offer that the agent was about to make. Otherwise, it will reject it and make its next offer.

### Agent Expressions

Agents follow one of two possible facial display policies depending on the condition. The angry agent follows a fixed, non-contingent policy, displaying anger on rounds 2, 6 and 10, and returning to a neutral face after five seconds (i.e., the policy is the same no matter what the participant offered). The sad agent follows the same policy but displays sadness rather than anger. Figure 2 illustrates the expressions. In all other rounds, both agents display a neutral face (Figure 1).

## Experiments

In the following experiment we investigate the interpersonal effects of angry and sad facial displays in the sacred-objects

Imagine the following hypothetical scenario:

There has been an earthquake in the town you live in and many have been injured. All roads to your town have been blocked and as a result aid is coming in very slowly. Because of this every family has to split packages of aid sent using helicopters with another family.

You and the family that have to split the aids with each other, both have babies who have [A: been injured and have developed infections] [B: have caught minor colds]. [A: The only way to control the spread of infection, which if not stopped will become lethal, is to use penicillin] [B: In order to relieve the cold you can give your child acetaminophen]. You are also running low on food, but have enough clean water that would last you for several days. All the shops in the town are closed, so it is uncertain whether you can use the money to purchase goods.

Given the circumstances, you know that no other aid package will be received for another week. The aid packages include medicine including [A: penicillin] [B: acetaminophen], canned food, some money and water bottles.

In the task that follows, you have to negotiate how these items have to be split between your family and the other household. You do not know how much food and water the other family has.

The negotiation is done in a sequence of alternating offers. You will make the first offer. The other negotiator may or may not accept your offer. If it does not accept it, that is, if it rejects your offer, it will send you a new offer. You can either accept or reject its offer. If you accept it, you will get to keep the items that you did not give them. If you reject their offer, you can make another offer and submit it to them. If after 12 rounds there is no agreement, the negotiation will end in no agreement. In this case, you both will only receive one of each item and the rest will be given away.

Try to get as many items as you can.

Figure 3: Participants were presented either with scenario A (deadly-infection) or scenario B (minor-cold)

task discussed in the previous section. Similar to other negative emotions, both anger and sadness serve as calls for adjustment of behavior (Van Kleef et al., 2004). Anger, which is the most common emotion in conflict situations, signals potential confrontation (Allred, 1999) and is related to the ethics of autonomy concerned with rights and justice (Rozin et al., 1999). On the other hand, the hypothesized communicative function of sadness, especially when combined with tears, is to elicit sympathy, signal appeasement, indicate a social need for help and to prioritize the salience of need, harm and care (Shariff & Tracy, 2011; Hasson, 2009; Tiedens, 2001; Horbeg, Oveis & Keltner, 2011). As discussed previously, we hypothesize that perceiving different emotional facial displays in others would prioritize different moral domains and affect moral decision-making. Therefore, we expect that the participants' moral appraisals would be influenced by the angry or sad emotional expressions displayed by the agent during the negotiation. We specifically predict that participants who view an object of the negotiation as a sacred value would show the typical rejection of tradeoffs seen in SV

participants (e.g. Tetlock et al. 2000) but only when they interact with the agent displaying anger. On the other hand, SV participants interacting with the sad agent would concede significantly higher than SV participants interacting with the angry agent. Moreover, the backfiring effect of anger should only be seen for participants who perceive the negotiation object as sacred.

### Participants

Two hundred and fourteen American Amazon-Turk workers (age: 33.71, gender: 56% female) were paid \$1 each to participate in our study. On average it took each worker 6 minutes and 16 seconds to complete our task.

### Design

The study employed a between subject 2 X 2 X 2 full factorial design, where the first factor was agent's expressed emotion (anger/sadness), the second factor was the experimental scenario (deadly-infection scenario/minor-cold scenario, Figure 2), and the last factor was whether or not participants held a SV for the medicine package. After reading one of the two scenarios described in Figure 3, we assessed participants' values regarding the medicine package using Baron and Spranca's (1997) measure. In accordance with this measure we asked our participants: "How do you feel about giving up the medicine package?", and they were provided the following four choices to choose from:

a. I think this definitely needs to happen.

b. I do not object to this.

c. This is acceptable only if the benefits of trading the medicine are great enough.

d. This shouldn't be done no matter how great the benefits are.

Participants who answered “d” were categorized as holding a SV for the medicine package. Participants then played the Sacred-Objects task as described in the last section.

### Results

Participants who dropped out of the negotiation before Round 4 (made only one or two offers and were exposed to the emotional displays of the agent only once) were excluded from the analysis ( $N = 21$ ). From the participants who read the deadly-infection scenario, 62.77% ( $N = 59$ ) perceived the medicine package as a sacred value, compared to 44.44% ( $N = 44$ ) in the minor-cold scenario ( $\chi^2(1, N = 193) = 5.7884, p = 0.0161$ ). However, the scenario manipulation did not have an effect on the course of the negotiation and there was no difference in participants' responses and offers between the two scenarios. Therefore, for the rest of the analysis we combine the data from the two scenarios into one condition.

We examined demands for the items through the course of the negotiation (Figure 4). As predicted, SV participants conceded less on medicine than Non-SV participants throughout the negotiation (Demand 1:  $t(191) = 2.8485, p = 0.0292$ ; Demand 2:  $t(191) = 5.1783, p < 0.0001$ ; Demand 3:

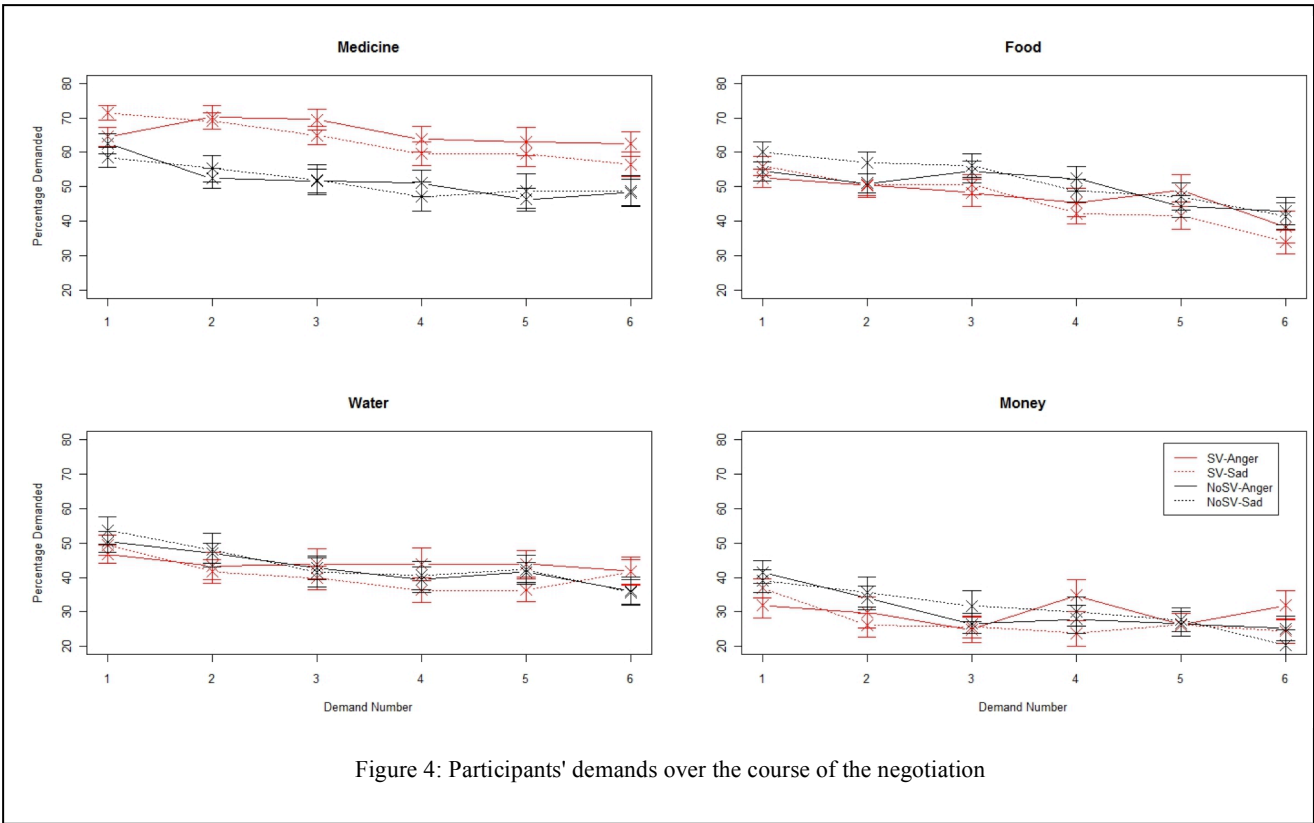


Figure 4: Participants' demands over the course of the negotiation



$t(191) = 4.5055, p < 0.0001$ ; Demand 4:  $t(191) = 3.2353, p = 0.0086$ ; Demand 5:  $t(186) = 3.7789, p = 0.0013$ ; Demand 6:  $t(185) = 2.8966, p = 0.0254$ ;  $p$ -values have been adjusted for multiple comparisons using Bonferroni correction). Corresponding effects were not obtained for the non-sacred objects (water bottles, food cans, money). Repeated measures ANOVA with Greenhouse-Geisser correction, with proposal round number as the within-subjects factor, and SV/NoSV and displayed-emotion as between-subjects factors, determined an overall main effect of time ( $p < 0.001$ ) for all the four objects, and an interaction between time, SV/NoSV and agent's emotion for medicine ( $F(4.585, 839.99) = 2.082, p = 0.053$ ). Again, corresponding effects were not obtained for the non-sacred objects.

To further analyze the differences in concession rates between the groups for different items, we used demand difference (number of packages in the first offer deducted from last offer) for medicine as a dependent variable in a 2 X 2 ANOVA where the first factor was the displayed emotional reaction (sadness/anger) and the second factor was the presence or absence of sacred values. For medicine, there was a significant interaction between SV and the agent's displayed emotion ( $F(1,189) = 4.7615, p = 0.0303$ ) (Figure 5). As predicted, SV participants who interacted with the sad agent conceded significantly higher than SV participants who interacted with the angry agent ( $t(101) = 2.3809, p = 0.0191$ ). Interestingly, SV participants who interacted with the angry agent conceded much less on medicine than Non-SV participants interacting with the same agent ( $t(94) = 2.1191, p = 0.0367$ ).

Given that the trends of negotiation for the rest the objects were similar, we combined them into a single group called non-sacred objects. A 2 X 2 ANOVA (SV/NoSV X AngryAgent/SadAgent) with average demand difference for non-sacred objects as the dependent variable, revealed a significant main effect of SV/NoSV, where SV participants conceded less ( $F(1,189) = 3.8550, p = 0.0511$ ), and a main effect of agent's emotion, where displayed sadness induced more concession ( $F(1,189) = 3.8915, p = 0.0499$ ). There was no interaction between SV and Agent's emotion for the non-sacred objects.

We also analyzed the participants' expressed emotion throughout the length of the negotiation. A planned comparison revealed that the difference in expressed anger between SV participants who interacted with the angry agent and non-SV participants interacting with the same agent was marginal ( $t(86) = 1.2613, p = 0.1053$ , one-tailed). Also, another planned comparison showed that SV participants who interacted with the agent that displayed sad facial expressions, expressed more sadness than SV participants who interacted with the angry agent ( $t(93) = 1.6826, p = 0.0479$ , one-tailed).

## Discussion

Our experiment shows that there was an overall concession over time for all items. However, for the medicine package, the amount of concession made by participants depended on

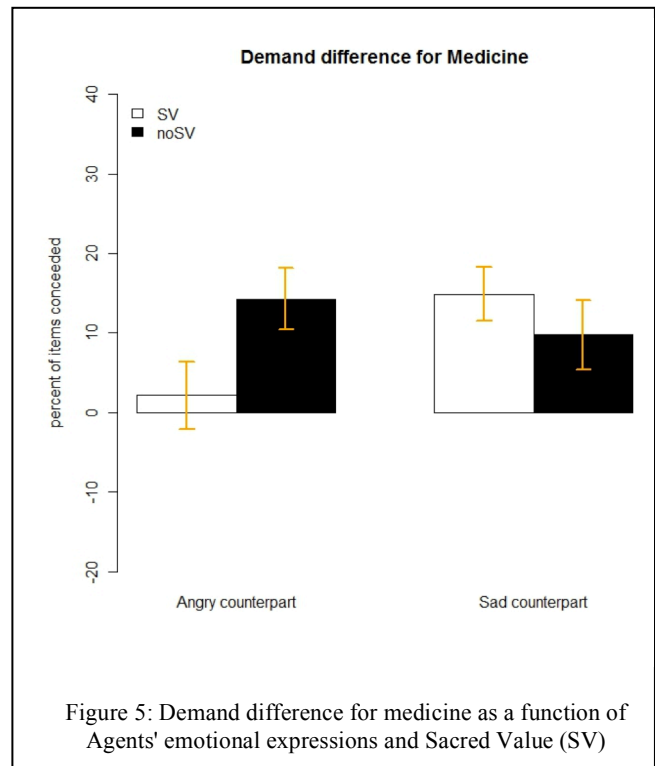


Figure 5: Demand difference for medicine as a function of Agents' emotional expressions and Sacred Value (SV)

whether they viewed the item as sacred and the agent's displayed emotion. As predicted, SV participants who interacted with the agent that displayed anger made significantly smaller concessions compared to SV participants who interacted with the sad agent. This finding supports our hypothesis that not only are moral concerns prioritized based on experienced intrapersonal emotions (Horbeg, Oveis & Keltner, 2011), but interpersonal emotions also affect moral appraisals by heightening different moral concerns. Specifically, we had predicted that witnessing sad facial displays would affect the decision-making of SV participants, by heightening the salience of need, weakness and harm/care (Horbeg, Oveis & Keltner, 2011). On the other hand, SV participants interacting with the angry agent showed the typical rejection of tradeoffs, as witnessing anger expressions amplified their concerns about their sacred values. Interestingly, consistent with findings of Van Kleef and others in non-moral domains (e.g. Van Kleef et al., 2004a, 2004b; Sinaceur & Tiedens, 2006), anger expressions for non-SV participants resulted in higher concession rates.

Overall, the contribution of our work is two-fold. First, our result emphasizes that moral appraisals can be strongly affected by interpersonal emotional expressions of other parties. Second, we showed that expressing anger may not be the best strategy to achieve higher concession rates in negotiation. Our result demonstrates that displays of anger may backfire if one of the parties associates moral significance to the objects of the negotiation. Displaying sadness was found to be a far more effective strategy to elicit concessions from the other party.



Sacred values play important roles in many cultural and political conflicts (e.g. Ginges et al., 2007; Dehghani et al., 2009, 2010). In this work we argued that moral concerns can be shifted within a negotiation in ways that promote cooperation and concession-making. One real-world implication of our research is that in negotiation involving sacred values, displaying anger and aggression might backfire, in the sense that it will result in conflicts to escalate and anger to reciprocate. However, non-aggressive and non-confrontational signals may result in better outcomes and larger concessions in these circumstances.

## Acknowledgments

This research was supported by AFOSR FA9550-09-1-050.

## References

- Allred, K. G. (1999). Anger driven retaliation: Toward an understanding of impassioned conflict in organizations. In R. J. Bies, R. J. Lewicki, & B. H. Sheppard (Eds.), *Research on negotiations in organizations* (Vol. 7). Greenwich, CT: JAI Press
- Atran, S. (2010). Talking to the Enemy: Violent Extremism, Sacred Values, and What it Means to be Human. London: Penguin
- Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes*, 70, 1
- Carnevale, P.J. (2008). Positive affect and decision frame in negotiation. *Group Decision and Negotiation*, 17, 51-63
- Carnevale, P., Kim, Y., de Melo, C., Dehghani, M. & Gratch, J. (2011). These Are Ours: The Effects of Ownership and Groups on Property Negotiation. Appeared in The 24th Annual Conference of the IACM, Istanbul, Turkey.
- Dehghani, M., Atran, S., Iliev, R., Sachdeva, S., Medin, D. & Ginges, J. (2010). Sacred values and conflict over Iran's nuclear program. *Judgment and Decision Making*, 5, 7, pp. 540-546.
- Dehghani, M., Iliev, R., Sachdeva, S., Atran, S., Ginges, J. & Medin, D. (2009). Emerging sacred values: Iran's nuclear program. *Judgment and Decision Making*, 4, 7, pp. 930-933.
- Forgas, J. P. (1998). On feeling good and getting your way: Mood effects on negotiating strategies and outcomes. *Journal of Personality and Social Psychology*. 74, 565-577
- Frijda, N. H., & Mesquita, B. (1994). The social roles and functions of emotions. In S. Kitayama & H. R. Markus (Eds.), *Emotion and culture: Empirical studies of mutual influence*. Washington, DC: American Psychological Association
- Ginges, J., Atran, S., Medin, D., Shikaki, K. (2007). Sacred bounds on rational resolution of violent political conflict. *Proceedings of the National Academy of Science*, 104, 7357
- Hasson, O. (2009). Emotional tears as biological signals. *Evolutionary Psychology*, 7, 363-37
- Horberg, E. J., Oveis, C., & Keltner, D. (2011). Emotions as moral amplifiers: An appraisal tendency approach to the influences of distinct emotions upon moral judgment. *Emotion Review*, 3, 237-244.
- Keltner, D., & Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition and Emotion*, 13, 505-521
- Lerner, J. S., Han, S., & Keltner, D. (2007). Feelings and consumer decision making: Extending the appraisal-tendency framework. *Journal of Consumer Psychology*, 17, 184-187.
- Medin, D. L., & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review*, 111, 960-983
- Rozin, P., Lowery, L., Imada, S., and Haidt, J. (1999). The cad triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, 76, 574-586.
- Shariff, A. F., & Tracy, J. L. (2011). What are emotional expressions for? *Current Directions in Psychological Science*, 20, 395-399.
- Sinaceur, M., & Tiedens, L. Z. (2006). Get mad and get more than even: When and why anger expression is effective in negotiations. *Journal of Experimental Social Psychology*, 42(3), 314-322.
- Tapias, P. M., Glaser, J., Keltner, D., Vasquez, K., & Wickens, T. (2007). Emotions and prejudice: Specific emotions toward outgroups. *Group Processes & Intergroup Relations*, 10, 27-39.
- Tetlock, P. E. (2003). Thinking the unthinkable: sacred values and taboo cognitions. *Topics in Cognitive Science*, 7, 32
- Tetlock, P. E., Kristel, O., Elson, B., Green, M., & Lerner, J. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, 78:853-870.
- Tiedens, L.Z. (2001). Anger and advancement versus sadness and subjugation: The effect of negative emotion expressions on social status conferral. *Journal of Personality and Social Psychology*, 80, 86-94
- Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2004a). The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology*, 86, 57-76.
- Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2004b). The interpersonal effects of emotions in negotiations: A motivated information processing approach. *Journal of Personality and Social Psychology*, 87, 510-528.
- Van Kleef, G. A., De Dreu, C. W., & Manstead, A. S. R. (2010). An interpersonal approach to emotion in social decision making: The Emotions as Social Information Model. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 42, pp. 45-96). Burlington, VT: Academic Press.

# Using Accent to Induce Cultural Frame-Switching

Morteza Dehghani, Peter Khooshabeh, Lixing Huang, Angela Nazarian, Jonathan Gratch

morteza,khooshabeh,lhuang,nazarian,gratch@ict.usc.edu

Institute for Creative Technologies

University of Southern California

12015 Waterfront Drive

Playa Vista, CA 90094

## Abstract

Language plays a crucial role in the formation and categorization of one's ethnic identity. Recent work on linguistic accent emphasizes the role of accent in person perception and social categorization, demonstrating that accent serves as a meaningful ethnic category indicator. In this paper, we examine whether accent can be used to implement socio-cultural markers. We investigate whether the accent of a virtual character, as a marker for culture, can cause cultural frame-shifts in individuals. We report an experiment, performed among bicultural and monocultural individuals, in which we test the above hypothesis. Our results show that accent can have a socio-cultural effect on people's cognition.

**Keywords:** accent, culture, frame-switching, biculturalism

## Introduction

There is a substantial amount of research focusing on the influence of language on human cognition. Language, at its core, is a cognitive tool that helps us construct shared meanings, categorize our knowledge, and thus piece together the various associative networks of collective experiences to form our perception of the world around us. According to the Whorfian hypothesis (Whorf, 1956), the grammar of a language and the manner in which the language is processed can provide a glance into the implicit organization of knowledge in that culture. This inherent influence is reflected in the belief that the grammatical structure of a language shapes our interactions due to the speakers inherent exposure to certain observations and mental representations, otherwise known as the principle of linguistic relativity (Chiu, Leung, & Kwan, 2007).

The results from previous studies with bilingual participants support the use of language as a means of activating associated cultural constructs. For instance, Chinese bilinguals randomly assigned to respond in Chinese reported more collective self-statements in open-ended self-descriptions, lower self-esteem, and more agreement with Chinese cultural views (Ross, Xun, & Wilson, 2002). These qualities are in line with traits commonly associated with Eastern collectivist cultures. The findings also suggest that cultural identities may be stored in separate knowledge structures and activated by the associated language. Marian and Neisser (2000) examined the retrieval of autobiographical memories in bicultural individuals and found that memories become more accessible when the language used at retrieval matched the language used at encoding. In other words, Russian-English bilinguals were able to remember more events from the Russian-speaking period of their lives when they responded in Rus-

sian and more experiences from the English-speaking period of their lives when they responded in English.

In their meta-analysis of cross-cultural studies of self-enhancement, Heine and Hamamura (2007) point out the limitations and confounds in existing research regarding whether language can independently prime a cultural mindframe. Lee, Oyserman, and Bond (2010) attended to these shortcomings in their study by exploring the effect of experimentally manipulated language use on value endorsements and self-descriptions. By unobtrusively manipulating language as a prime, they found that bilingual participants randomly assigned to use English reported significantly more self-enhancing social comparative judgements than those using Chinese. In addition, English users demonstrated more social distancing after under-performance and standardized their failure in the context of the behavior of their peers in order to protect their positive self-regard. This is in line with the expected behavior of someone from an individualist culture and touches on the work done by Tesser (2000) and Kwan et al. (2004) regarding the mechanisms employed to protect, augment, or maintain self-esteem. The above findings also support the claim that societies are socialized for both individual and collective mindsets, but responses are dependent on the cultural mindset salient at the moment of self-reflection (Lee et al., 2010).

This literature demonstrates the cognitive implications of experimentally manipulating language as a prime to prompt a shift in an individuals cultural mindset. The findings from these studies support the use of language as a marker for culture, however, in this paper we aim tease apart the specific aspects of language that act as cultural markers, e.g., by manipulating accents. Ethnolinguistic identity theory indicates that language plays a crucial role in the formation and categorization of one's ethnic identity (Giles, Bourhis, & Taylor, 1977). Rakič, Steffens, and Mummendey (2011a) compared the strength of vocal accents compared to facial visual appearance as indicators for ethnic categorization. They found that not only did accent serve as a meaningful ethnic category indicator, but participants also overlooked visual stimuli, such as facial features typically associated with a culture, in the presence of this more meaningful auditory information (Rakič et al., 2011a). In other words, sociopsychological processes were a more salient tool in determining memorability than objective stimulus features. In a later study, they examined the effect of accent workplace context. Job applicants with a standard German accent were compared to those with

nonstandard regional accents to test for biases influenced by the auditory information. Rakič, Steffens, and Mummendey (2011b) found that speakers with a standard German accent were perceived as more competent and more hireable than regional accent speakers. In addition, speakers with a Bavarian accent and the standard German accent were perceived as having higher socio-intellectual status than the remaining speakers. These results bolster the role of linguistic accent in person perception and social categorization.

In this paper, we examine whether accent can be used as a marker for culture by evoking cultural *frame-switching* (Hong, Morris, Chiu, & Benet-Martinez, 2000) in bicultural individuals. Cultural frame-switching refers to the idea that interpretive frames, in individuals who have internalized two cultural identities, can shift due to situational cues (Benet-Martinez, Leu, Lee, & Morris, 2002). For example, Hong et al. (Hong et al., 2000) prime Chinese-Americans with American, Chinese or neutral iconic images and demonstrate that participants in the Chinese prime condition interpreted the next task with more of a Chinese interpretive lens (focused more on external attributions) than those in the American or control conditions. On the other hand, participants in the American prime condition projected more American cultural values by focusing on individual attributions for the same task. A plausible explanation for frame-switching is that multi-cultural individuals have different norms and culturally specific systems of meaning (D'Andrade, 1984), which are activated depending on the context and the social environment. Hence, activation of different cultural interpretive frames can result in varied constraints on the individual's psychophysical behaviors.

The shift in the interpretive frames can be especially notable if there are significant differences in normative behavior between an individuals two cultures. There is growing evidence in the social sciences that culturally normative behaviors vary across cultures (see Henrich, Heine, & Norenzayan, 2010 for a review). In other words, behaviors that are considered typical in one culture may be considered abnormal in another. The variability of culturally normative behaviors and cognitions have been noted in different aspects of human behavior. Related to this work, researchers have shown cultural differences in morally motivated decision-making by identifying moral domains that are present (or salient) in some cultures but not in others (Shweder, Much, Mahapatra, & Park, 1997; Haidt, Koller, & Dias, 1993). Domains such as respect for authority and the saliency of the distinction between purity and impurity are some that have been identified in helping people characterize certain situations as morally tinged within one cultural group but not another. In this work, we carefully control for non-verbal behavior and manipulate only accent of spoken English, in order to experimentally evoke frame-switching and measure its effect on the participants' perception and cognition. Based on the previous literature, we predict individuals interacting with a counterpart having a culturally congruent accent will use moral frames specific to that

culture. Our method differs from studies that use language as a prime in the sense that in our experiment participants only listen and do not generate language explicitly.

## Embodied Conversational Agents

Virtual agents, particularly when rendered as embodied conversational characters, are capable of providing a compelling multimedia platform that serves as an effective interface for research purposes, educational applications or entertainment. Embodied conversational agents (ECAs) make it possible to manipulate external features such as visual appearance, speech type, and contextual graphical environments. This ability makes ECAs a convenient platform to isolate unique cultural characteristics and realize them through simulation. Along with enhanced experimental control, ease of manipulations, consistency and controlled measurements (Loomis, Blascovich, & Beall, 1999), these features make ECAs useful and reliable tools for studying cultural cognitions. For example, we can objectively pinpoint certain social and behavioral characteristics that are relevant to specific cultures and implement them using virtual agent technology. There have been a small number of studies investigating how culturally congruent virtual agent characteristics can affect users' cognition. In an effort to examine the interaction between culture and ECA design in the domain of education, Rader, Echelbarger, and Cassell (2011) developed virtual peers that matched the dialect of children speaking African-American English and asked the children to complete a bridge building exercise. The children alternated playing the role of student and teacher as they explained the building process. Rader et al. found that students who tend to speak more dialected English did so less when they played the role of a teacher. This work suggests that the virtual peer and culturally congruent context, coupled with the role switch, influenced students to speak mainstream English, which is shown to be related to higher student achievement. In another line of work, Yin, Bickmore, and Cortes (2010) report that individuals who process information using peripheral cues perceived an agent tailored to their own culture as more persuasive and trustworthy.

## Experiment

In this paper, we experimentally model frame-switching among bicultural individuals using the accent of an ECA and measure if their preference for certain moral acts are affected by this manipulation. Our hypothesis is that the accent of a virtual agent should affect people's perception of the culture of the agent. If this is true, then a virtual human that has an accent that is congruent with a participant's culture will elicit use of the congruent cultural frame. In order to test our hypothesis, we designed an experiment in which we control for non-verbal behavior of an ECA while manipulating only its accent. We recruited Iranian-American and American (US majority culture) participants and had them read a story which included Iranian and US cultural values and customs. The participants were then asked to summarize the story and answer a few questions about the material.

## Participants

Fifty-two Americans (mean age = 40) and Fourteen Iranian-Americans (Iranians living in Southern California for more than 5 years) (mean age = 34.61) participated in this study. The participants were recruited using craigslist.com and snowball sampling, which consisted of asking subjects to refer other subjects for the study. Each subject received \$25 at the end of the experiment for participating. The participants were not aware that they were participating in a culture study. Each participant completed the task in individual experimental sessions.

## Design

The study employed a between subject 2 X 2 full factorial design. The first factor is the culture of the participants (American or Iranian-American). The second factor is the agent's accent, which was either a standard American English or Iranian English accent, spoken by the virtual agent.

## Stimuli

The participants were first asked to read a short story (Figure 2) about a student named Anthony who was asked to go to dinner at this classmate's (Shawn) house. After arriving in Shawn's house, Anthony comes to the conclusion that Shawn's parents were not expecting a guest for dinner. Anthony complements Shawn's dad about an art piece in the house, and Shawn's dad insists that Anthony should take the picture. Finally, Anthony's friend picks him up from Shawn's house before dinner was served. The story included a balanced number of American and Iranian cultural products (e.g. proverbs), values, and events that could be tied to the celebration of Iranian New Year and to Saint Patrick's day<sup>1</sup>. None of these idea units were explicitly labeled with their cultural referent (there was no explicit reference to St. Patrick's day as such) and the idea units were interleaved so as to minimize memory distortions due to recency or primacy. After reading the story, participants interacted with an ECA.

## Rapport Agent

The agent used in this experiment, Utah (Hartholt, Gratch, Weiss, & Team, 2009) (Figure 1), is designed to establish rapport with human participants by providing contingent feedback while a user is speaking. To produce feedback, the agent first detects and analyzes in real-time the human speakers' audiovisual features, which are silence, head nod, eye-gaze (looking at the agent or not) and smile. The audio feature detector extracts intensity from the raw signal every 100ms using Praat<sup>2</sup>. With the intensity information, it outputs a binary feature, speech or silence, every 100ms. The visual feature detector<sup>3</sup> tracks the position of the face, the facial feature points, the direction of eye gaze and the smile level. With this information, it outputs visual features indicating whether the



Figure 1: The ECA used in our Experiments

human is nodding or not, looking away or not, and smiling or not. Based on the perceived audiovisual features, the response model (Huang, Morency, & Gratch, 2011) decides, in real-time, the most appropriate responses, such as head nod and smile. These different styles of animations are first converted into Behavior Markup Language (BML) (Kopp et al., 2006) and then sent to an action scheduler, which keeps track of the duration of each animation. If the current animation has not been completed, the new animation will be ignored. The BMLs are passed to Smartbody (Thiebaux & Marsella, 2007), a virtual human animation system designed to seamlessly blend animations and procedural behaviors. Finally, the byproducts of Smartbody are rendered by a commercial game engine, Gamebryo<sup>4</sup>, and displayed to users. For the experiment, the voice of the ECA was prerecorded using the voice of the second author, whom is familiar with both Iranian and American cultures. The second author did not participate in recruiting nor in running the experiments.

## Procedure

After participants finished reading the story, the virtual agent greeted them, explained an overview of the research center and asked them to verbally summarize the story they had just read. Next, they filled out a questionnaire about the appropriateness of certain actions and intentions of the characters within the story. They were specifically asked the following two questions: 1. Was it appropriate for Anthony to leave before dinner? 2. Do you think Shawn's parents really wanted to give the picture to him? Our hypothesis predicts that participants should use culturally congruent frames to interpret and answer these moral questions. The Iranian cultural frame suggests that it is not appropriate to refuse someone's generosity and hospitality. If cultural frame-shifting does indeed take place for Iranians-Americans when interacting with the culturally congruent agent, then they should say it is inappropriate for Anthony to leave dinner early. For the second

<sup>1</sup> Saint Patrick's day is widely celebrated in North America

<sup>2</sup>Praat, <http://www.fon.hum.uva.nl/praat/>

<sup>3</sup>OKAO Technology, [http://www.omron.com/r\\_d/coretech/vision/okao.html](http://www.omron.com/r_d/coretech/vision/okao.html)

<sup>4</sup><http://www.gamebryo.com>

... My classmate Shawn, who I'm not really good friends with and don't know too well, invited me to a bonfire at the beach. I told him that I was hungry and needed to get something to eat before going to the beach. He was going to his parent's house for dinner and invited me over. I don't know him well, so I initially refused his offer. But he kept on insisting that it's the beginning of Spring and I should go and have dinner with his family. On the way to his house, Shawn asked why I was wearing almost all green. I thought it was a strange question as a lot of students were wearing green that day...

I went over to his house and met his dad at the living room. I thought to myself that the apple doesn't fall far from the tree. Upon seeing his father, Shawn introduced me to him, saying this is my good friend Anthony. He seemed surprised by my presence. Now I wasn't sure if they were expecting a guest for dinner. Then Shawn's mother came to the living room. I had met her before. She used to work in the registrar of the school. I said hi to her and she greeted me back saying that it looked like water had gone under my skin...

... I didn't understand why they were asking me to stay for dinner, because I think they didn't have any dinner prepared.

... I saw a very small art piece on the wall of their hallway. It was a picture of some Chinese looking guys playing with a ball on horses. I told his dad that this is a lovely picture. He thanked me and told me that I could take it. I first thought he was kidding, but he seemed serious and told me that he wants me to take it. Shawn's mother also said that it will look better in my house and I should take it. Given that they were insisting so much, and the piece didn't look expensive I took it and thanked them for it.

... Shawn's parents told Shawn that they had to wait for another hour and a half to serve dinner because some family friends had just called and were coming to visit them for the first day of spring. Given that I had plans to go watch the game and couldn't wait that long, I got up thanked them for the picture and salad, and had my friend pick me up. Shawn and his parents insisted that I should wait and have dinner with them. But my friend was already there to pick me up...

Figure 2: Excerpt from the story read by participants

question, the Iranian frame could interpret the event as an instance of Iranian hospitality, especially when it comes to sharing their cultural artifacts (in this case Persian miniature). Next, participants were asked several questions about different emotions of the characters in the story. Lastly, to check the effectiveness of our manipulation, participants answered the following two questions in a random order: 3. Did the agent have more of an American accent or Middle-Eastern accent? 4. Did the agent appear more Western or more Middle-Eastern? Each question was answered on a 6-point scale (1 = No he did not, 6 = Yes he did; 1 = Not at all appropriate, 6 = Completely appropriate; 1 = Very much American, 6 = Very much Middle-Eastern; 1 = Very much Western, 6 = Very much Middle-Eastern).

## Results

For both manipulation check questions, we used the responses to questions 3 and 4 as dependent variables in a 2 X 2 ANOVA, where the first factor was the culture of the participants (American or Iranian-American) and the second factor was the accent of the agent (American or Iranian). There was a main effect of agents' accent for both questions (appearance:  $F(1, 62) = 11.038, p = 0.0015$ ; accent:  $F(1, 62) = 68.1434, p < 0.001$ ). The agent with an Iranian accent was viewed to not only have a more Middle-Eastern accent but also appeared more Middle-Eastern. Also, there was a main effect of culture for the appearance question ( $F(1, 62) = 4.276, p = 0.0428$ ) where Americans ranked the agent as more Middle-Eastern looking than did the Iranian-Americans. The responses to the first question were used as the dependent variable in a 2 X 2 ANOVA, with similar factors as above. There was a significant interaction between the two factors ( $F(1,62) = 4.3649, p = 0.0408$ ). A planned comparison revealed that Iranian-American participants who interacted with the agent with an American accent viewed Anthony leaving before dinner as more appropriate than the Iranian-Americans who interacted with the agent with an Ira-

nian accent (Welch Two Sample one-tailed t-test:  $t(10.596) = 10.596, p = 0.0512$ )<sup>5</sup> (Figure 3). However, this difference did not reach significance for Americans ( $t(49.48) = 1.4179, p = 0.1625$ ).

Next, using the responses to the second question as the dependent variable, we ran the same 2 X 2 ANOVA. There was a main effect of culture ( $F(1, 62) = 9.5759, p = 0.0029$ ), where Iranian-Americans ranked this question lower than Americans. There was an interaction between culture and accent of the agent ( $F(1, 62) = 9.5759, p = 0.0029$ ), where Iranians who interacted with the American accent agent indicated that Shawn's parents didn't want to give the picture to Anthony compared to those in the Iranian accent condition (Welch Two Sample t-test:  $t(9.008) = 3.1186, p = 0.0123$ ). However this difference did not reach significance for Americans ( $t(49.983) = -1.2908, p = 0.2027$ ).

There were also interactions between culture and agent's accent for Anthony's feeling of happiness ( $F(1, 62) = 5.0474, p = 0.0282$ ) and satisfaction ( $F(1, 62) = 12.7468, p < 0.001$ ). For happiness, Iranian-Americans interacting with the agent with an Iranian accent rated Anthony's happiness higher than those Iranian-Americans who interacted with an American accented agent ( $t(11.725) = 2.4057, p = 0.0336$ ), and vice versa for American participants ( $t(49.408) = -2.2325, p = 0.0301$ ). The same significant trend held for Anthony's satisfaction (Iranian-Americans:  $t(11.855) = 3.2666, p = 0.0068$ ; Americans:  $t(48.815) = -3.543, p < 0.001$ ). Similarly, there was a two-way interaction between culture and accent for Shawn's parents' satisfaction ( $F(1, 62) = 7.0429, p = 0.0101$ ) and the effect also approached significance for Shawn's satisfaction ( $F(1, 62) = 3.0044, p = 0.0880$ ). Similar to the emotions reported above, participants ranked the satisfaction of Shawn's parents and Shawn higher when the accent

<sup>5</sup> A power test revealed that if we had the same number of Iranian-American participants as American participants, with probability of 99.15% a two-tailed test with the means of the above sample would have reached significance.



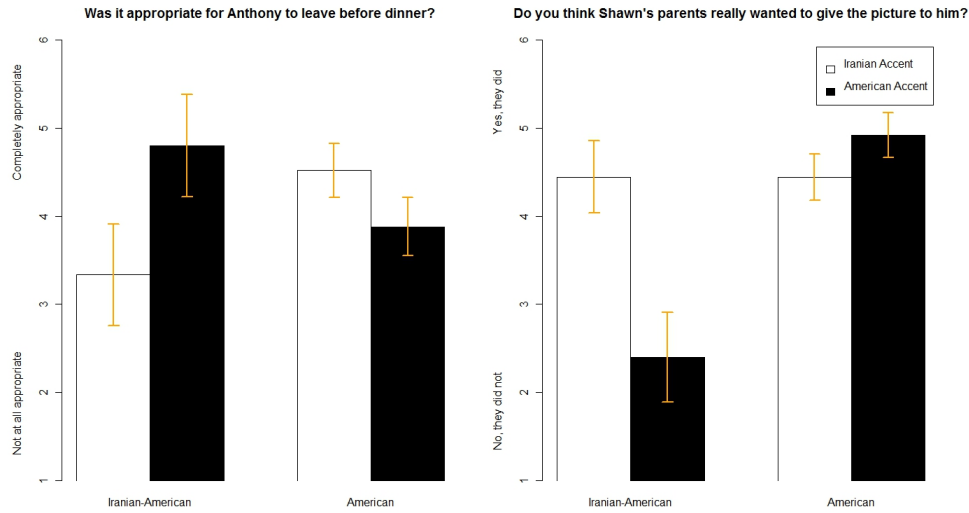


Figure 3: Interaction between culture and accent of the agent

of the agent matched their own accent (Shawn's parents': Iranian-Americans:  $t(10.121) = 3.6028$ ,  $p = 0.0047$ ; Americans:  $t(49.362) = -1.4053$ ,  $p = 0.1662$ , Shawn's: Iranian-Americans:  $t(5.953) = 0.7476$ ,  $p = n.s.$ ; Americans:  $t(49.054) = -2.2229$ ,  $p = 0.0309$ ).

## Discussion

This study provides evidence of how spoken accent can make a socio-cultural effect on people's cognition. In a fully factorial design, Iranian-Americans who interacted with a virtual agent that spoke Middle Eastern accented English were more likely to use a congruent cultural frame to interpret a morally charged scenario. The compelling aspect of this effect is that the accented virtual agent's visual appearance was identical across experimental conditions and the only manipulation was the agent's accent. Although simple effects were not significant for the Americans, the trends were in the correct direction. The trends could be due to the fact that most of the American participants recruited in this study have a multicultural background (33% African-American, 14% Latino).

To our surprise, our manipulation also affected people's evaluations of the emotions of the characters in the story. We speculate that this effect might be due to the fact that participants who interacted with an agent whose accent matched their own attributed intentions and goals that are congruent to their own cultural mindset to the characters and, as a result, appraised the situation for the characters more positively. Therefore, when Iranian-Americans interacted with the ECA that had an Iranian accent, they appraised the intentions and goals of the characters to be more in line with the Iranian culture, hence they evaluated the situation more positively, compared to Iranian-Americans who interacted with the American accented ECA.

We acknowledge the low number of Iranian-Americans

participants in our study. However, we would like to note that the probability of replication of a result is dependent on p-levels but not affected by sample size (e.g. Killeen, 2005).

In summary, contributions of this work are two-fold. First, the study adds to the literature in cross-cultural psychology by showing that just spoken language accent can induce cultural frame shifts. Second, this work makes a methodological contribution to the field of human-computer interaction and experimental psychology. Our results also have implications for teaching cross-cultural fluency and competency.

**Acknowledgments.** This research was supported by NSF IIS-0916858, a postdoctoral fellowship to PK from the Army Research Lab, and U.S. Army Research, Development, and Engineering Command (RDECOM). The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. MD, PK and LH were involved in designing and analyzing the experiment and also in writing the paper. AN was involved in running the experiment and writing the paper. JG was primarily involved in the writing and analysis stages of this work.

## References

- Benet-Martinez, V., Leu, J., Lee, F., & Morris, M. W. (2002). Negotiating biculturalism: Cultural frame switching in biculturals with oppositional versus compatible cultural identities. *Journal of Cross-Cultural Psychology*, 33 (5), 492-516.
- Chiu, C., Leung, A. K., & Kwan, L. (2007). The handbook of cultural psychology. In S. Kitayama & D. Cohen (Eds.), (p. 259-276). Guilford Press.
- D'Andrade, R. (1984). Cultural meaning systems. In

- R. Schweder & R. Levine (Eds.), *Culture theory. essays on mind, self and emotion* (p. 88-119). Cambridge: Cambridge University Press.
- Giles, H., Bourhis, R., & Taylor, D. M. (1977). Language, ethnicity and intergroup relations. In H. Giles (Ed.), (p. 307-348). Academic Press.
- Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, 65, 613-628.
- Hartholt, A., Gratch, J., Weiss, L., & Team, T. G. (2009). At the virtual frontier: Introducing gunslinger, a multi-character, mixed-reality, story-driven experience. In Z. Ruttkay, M. Kipp, A. Nijholt, & H. Vilhjálmsson (Eds.), *Intelligent virtual agents, lecture notes in computer science* (p. 500-501). Springer.
- Heine, S. J., & Hamamura, T. (2007). In search of east asian self-enhancement. *Personality and Social Psychology Review*, 11 (1), 4-27.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61-83.
- Hong, Y., Morris, M. W., Chiu, C., & Benet-Martinez, V. (2000). Multicultural minds. a dynamic constructivist approach to culture and cognition. *American Psychologist*, 55 (7), 709-20.
- Huang, L., Morency, L. P., & Gratch, J. (2011). Virtual rapport 2.0. In *International conference on intelligent virtual agents*.
- Killeen, P. R. (2005). An alternative to null-hypothesis significance testing. *Psychological Science*, 16, 345-353.
- Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., ... Vilhjálmsson, H. (2006). Towards a common framework for multimodal generation: The behavior markup language. In *International conference on intelligent virtual agents* (pp. 21-23).
- Kwan, V. S. Y., John, O. P., Kenny, D. A., Bond, M. H., & Robins, R. W. (2004). Reconceptualizing individual differences in self-enhancement bias: An interpersonal approach. *Psychological Review*, 111 (1), 94-110.
- Lee, S., Oyserman, D., & Bond, M. H. (2010). Am i doing better than you? that depends on whether you ask me in english or chinese: Self-enhancement effects of language as a cultural mindset prime. *Journal of Experimental Social Psychology*, 46, 785-791.
- Loomis, J. M., Blascovich, J. J., & Beall, A. C. (1999). Immersive virtual environment technology as a basic research tool in psychology. *Behavior research methods instruments computers a journal of the Psychonomic Society Inc*, 31(4), 557-64.
- Marian, V., & Neisser, U. (2000). Language-dependent recall of autobiographical memories. *Journal of Experimental Psychology: General*, 129 (3), 361-368.
- Rader, E., Echelbarger, M., & Cassell, J. (2011). Brick by brick. In *Proceedings of the 2011 annual conference on human factors in computing systems - CHI '11* (p. 2971). Vancouver, BC, Canada. doi: 10.1145/1978942.1979382
- Rakič, T., Steffens, M. C., & Mummendey, A. (2011a). Blinded by the accent! the minor role of looks in ethnic categorization. *Journal of Personality and Social Psychology*, 100 (1), 16-29.
- Rakič, T., Steffens, M. C., & Mummendey, A. (2011b). When it matters how you pronounce it: The influence of regional accents on job interview outcome. *British Journal of Psychology*, 102 (4), 868-883.
- Ross, M., Xun, W. Q. E., & Wilson, A. (2002). Language and the bicultural self. *Personality and Social Psychology Bulletin*, 28 (8), 1040-1050.
- Schweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). Morality and health. In A. Brandt & P. Rozin (Eds.), (p. 119-169). New York: Routledge.
- Tesser, A. (2000). On the confluence of self-esteem maintenance mechanisms. *Personality and Social Psychology Review*, 4 (4), 290-299.
- Thiebaux, M., & Marsella, S. (2007). Smartbody: Behavior realization for embodied conversational agents. In *In 7th international conference on intelligent virtual agents (iva)*.
- Whorf, B. (1956). *Language, thought, and reality*. Cambridge, MA: Technology Press of MIT.
- Yin, L., Bickmore, T. W., & Cortes, D. E. (2010). The impact of linguistic and cultural congruity on persuasion by conversational agents. In *Proceedings of IVA, lecture notes in computer science* (pp. 343-349). Philadelphia, PA: Springer.



# How Function Assignment and Word Order are Determined: Evidence from Structural Priming Effects in Japanese Sentence Production

**Ying Deng (dengying0105@gmail.com)**

Elisabeth University of Music

4-15 Nobori-cho, Naka-ku, Hiroshima-Shi, 730-0016 Japan

**Hajime Ono (onohajime@kindai.ac.jp)**

Faculty of Science and Engineering, Kinki University

3-4-1 Kowakae, Higashiosaka-Shi, Osaka, 577-8502 Japan

**Hiromu Sakai (Hsakai@hiroshima-u.ac.jp)**

Graduate school of Education, Hiroshima University

1-1-1 Kagamiyama, Higashihiroshima-Shi, Hiroshima, 739-8524 Japan

## Abstract

Using a structural priming paradigm, the details of sentence production model have been investigated substantially, specifically the processes in grammatical encoding level. Many studies provide evidence that the function assignment stage and the constituent assembly stage are processed separately in grammatical encoding. However, it is less known whether these two stages interact with each other during the processes, and if so, how the processes are executed. In this study, we report three structural priming experiments in Japanese, in which function assignment and word order were manipulated independently and simultaneously in order to examine the processes at two stages directly. Our results revealed that priming effects patterns were different depending on whether the effects occur at function assignment stage or at constituent assembly stage. Based on the current findings, implications for the recent models of sentence production are discussed, from the perspective of the grammatical function assignment and word order determination.

**Keywords:** Structural Priming; Sentence Production; Active/Passive sentence

## Introduction

People can express thoughts by conveying them through language. Speakers can often express their meaning in several ways by using different linguistic expressions. For example, a transitive event in which a dog is chasing a cat can be expressed by an active sentence (the dog is chasing the cat) as well as by a passive sentence (the cat is being chased by the dog). How do speakers put words together, and choose a syntactic structure? Many decisions must be made during the production process. How these decisions are made and how these decisions generate a linguistic expression, especially how the grammatical function roles and word order are determined during the processing are the central issues of sentence production.

Current models of sentence production widely assume that language production has three levels: message encoding level, grammatical encoding level, and a phonological and phonetic encoding level (e.g., Bock and Levelt, 1994, Ferreira and Slevc, 2007). Grammatical function assignment

and word order are assumed to be determined during the grammatical encoding level, in which speakers encode the preverbal message into a grammatically well-formed linguistic message. The grammatical encoding level is comprised of two separate encoding stages: the function assignment stage in which thematic roles such as agent and patient are assigned to grammatical functions and the constituent assembly stage in which word order is processed. For example, the choice of active/passive alternations is assumed to take place in the function assignment stage, in which the agent role is mapped to the subject and the patient role to the non-subject in the active voice, and the opposite mapping takes place in the passive voice. However word order is assumed to be left unspecified at this point. In the subsequent stage, the constituent assembly stage, word order is processed, such as in the case of the canonical/scrambled word order alternations in Japanese.

In contrast several studies have provided empirical evidence that suggests function assignment and word order are computed simultaneously (Bernolet, Hartsuiker, & Pickering, 2007; Branigan, Pickering, & Tanaka, 2008). Most production models assume that these two stages in the grammatical encoding level are computed separately, (e.g., Hartsuiker & Westenberg, 2000; Vigliocco & Nicol, 1998). Therefore, there is still controversy about whether grammatical functions and word order are encoded separately in different stages. Furthermore, less is known about how these two stages are computed during the production processes, that is whether or not these two stages interact with each other and if so, how the processes are executed.

A structural priming paradigm has often been used to investigate the details of the production model. Structural priming is known as the speaker's tendency to reuse syntactic structures that they have recently produced or comprehended (Bock, 1986). Using a picture description task, Bock (1986) found that after repeating a prime sentence participants were more likely to describe a subsequent target picture with the structure that they had just repeated in the prime sentence. For example, more

passive sentences were produced after passive primes than after active primes. Since previous studies have shown that structural priming is sensitive to syntactic/structure factors between the prime and target pairs (e.g., Bock, 1989, Bock & Loebell, 1990), the structural priming paradigm has been used to investigate the details of the processes in the grammatical encoding level specifically. However empirical evidence supporting the two separate stages model and the one stage model both come from research using the structural priming paradigm (e.g., the separate stages model: Hartsuiker & Westenberg, 2000; Shin & Christianson, 2009; the one stage model: Bernolet, Hartsuiker & Pickering, 2007; Tanka, 2007).

In this study, we investigated whether the function assignment stage and the constituent assembly stage are independent of each other in the grammatical encoding level. Moreover we aimed to investigate the details of how the function assignment stage and the constituent assembly stage are computed during the production processes, that is whether the processes of grammatical function assignment and word order determination interact with each other and if so, how the processes are executed.

In order to answer these questions we used Japanese and conducted three structural priming experiments. The relatively free word order of Japanese allowed us to manipulate function assignment and word order independently. In Japanese, besides mapping a thematic role to the subject (active/passive sentence), one can choose to place the subject at the first position before the non-subject in the sentence, namely, canonical word order (SOV word order), or after the non-subject, in scrambled word order (OSV word order). This word-order variation in Japanese allows us to manipulate the grammatical functions independent of overlap in the word order between prime and target pairs and enable us to investigate the function assignment stage and the constituent assembly stage separately.

If the two stages are computed simultaneously, we would expect to see structural priming effects only when both function assignment and word order match between the prime and the target pairs. If the function assignment stage and the constituent assembly stage are independent of each other, we would expect to see structural priming effects when only one function assignment or word order is shared between the prime and the target pairs (i.e., OSV-passive and SOV-passive pairs; OSV-passive and OSV-active pairs). Moreover, if the processes of two stages interact with each other, we would expect to find structural priming effects between the prime and target pairs interacted with grammatical function assignment and word order determination process.

In Experiment 1, we examined whether the pure structural priming effects of active and passive sentences in Japanese can be observed when conceptual factors were controlled. In Experiment 2, we looked at whether priming effects can be found when function assignment is matched between prime-target pairs, even without sharing the word order, like OSV-

passives and SOV-passives pairs. In Experiment 3, we manipulated function assignment and word order independently and simultaneously in order to look at whether the priming effects can be observed across the two stages.

## Experiment 1

In Experiment 1, we examined structural priming effects of active and passive sentences in Japanese. If the locus of the voice priming effects occurs at the function assignment stage, i.e. the assignment of the subject to agent role in active sentences or patient role in passive sentences, we expect to observe priming effects.

## Method

**Participants** Twenty students, all native speakers of Japanese at Hiroshima University participated in the experiment (13 females and 7 males, the mean age was 22.1 years). Participants were paid 500 yen for their participation. Throughout the experiment, a female native speaker of Japanese acted as the confederate.

**Materials** The prime was either an active or a passive sentence: (1) active prime and (2) passive prime. A target picture (Figure 1) was presented immediately after the prime. Two sets of 80 items were created: a confederate set and a participant set. The confederate set consisted of 80 sentences (20 prime and 60 filler sentences), and the participant set consisted of 80 simple black and white line drawings (20 target and 60 filler pictures). In addition, we prepared two sets of 36 simple black and white line drawings for a picture recognition task.

- (1) sapootaa-ga sakkaa sensyu-o ooensi-teiru.  
fans-NOM soccer player-ACC cheer  
“The fans are cheering the soccer player.”
- (2) sakkaa sensyu-ga sapootaa-ni ooen-sare-teiru.  
soccer player-NOM fans-OBL cheer-PASSIVE  
“The soccer player is being cheered by the fans.”

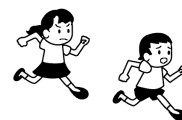


Figure 1: Example of a target picture. A girl chasing a boy.

To ensure that the observed effects are purely structural in nature, conceptual factors such as event type, animacy of NPs, and viewpoint shifts were carefully controlled in the prime/target pairs. That is, our experimental items only involved human entities for both agent and patient, minimizing the bias of conceptual accessibility. The event type between primes and targets was paired among positive, negative or neutral types, in order to eliminate the possibility that the event similarity facilitates the use of same viewpoint between the prime/target pairs.

**Procedure** We adopted a confederate-participant dialogue-style setting (Branigan, Pickering & Cleland, 2000). Participants were told that the study was investigating how people communicate to each other. Participants were asked to describe pictures in turn with the confederate. The experiment always began with the confederate's description. The confederate pretended to describe the picture even though the prime sentence was visually presented on screen. Next, participants were asked to repeat the prime sentence loudly, while memorizing and creating a mental image of the picture. Then, participants described the target picture on the screen and the confederate repeated the description and pretended to memorize it. The experiment consisted of three blocks, and after each block, participants and the confederate were asked to complete a picture recognition task in which a set of pictures was presented. Participants and the confederate were asked to decide whether these pictures matched the description given by the partner during the block.

## Results and Discussion

Priming effects were observed even after conceptual factors being strictly controlled (Figure 2). The results showed that, more passive sentences were produced after passive primes (11.5%) than active primes (1.5%), indicating a clear priming effect ( $F_1(1,19)=10.94, p<.001$ ;  $F_2(1,19)=17.36, p<.001$ ).

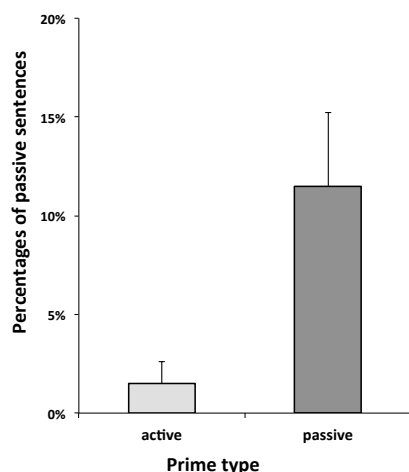


Figure 2: Percentages of passives in each condition.

After participants heard passive primes (i.e., assign the sentence subject to patient role) from the confederate, they tended to use passive sentences more often than after they heard active primes (i.e., assign the sentence subject to agent role). Because the conceptual factors between prime and target pairs have been carefully controlled in Experiment 1, the priming effects are most likely due to the processes in the function assignment stage.

However, there is an alternative explanation to the priming effects in Experiment 1. The passive primes and the produced passive sentences shared the same patient-agent thematic role order, but the active primes and the produced passive sentences do not. Since previous studies have demonstrated that the thematic role order did affect the priming effects (e.g., Chang, Bock, & Goldberg, 2003), it is possible that the locus of the priming effect occurs outside the function assignment stage. In experiment 2 we addressed these concerns.

## Experiment 2

In Experiment 2, we looked at whether priming effects can be found when function assignment is matched between prime-target pairs, even without sharing the word order, like OSV-passives and SOV-passives pairs. If function assignment and word order are computed simultaneously, the effects observed in Experiment 1 are expected to disappear in Experiment 2. In contrast, if function assignment is responsible for the priming effects, and it can be processed separately from word order, then we expected to observe the priming effects again.

## Method

**Participants** Twenty students, all native speakers of Japanese, at Hiroshima University participated in the experiment (11 females and 9 males, the mean age was 22.8 years). Participants were paid 500 yen for their participation. Throughout the experiment, a female native speaker of Japanese acted as the confederate.

**Materials** The materials were similar to those in Experiment 1, except we changed the word order of the subject and non-subject (oblique object) in the passive prime as in (3). This time the agent-patient order in active primes and passive primes was controlled.

- (3) sapootaa-ni sakkaa sensyu-ga ooen-sare-teiru.  
fans-OBL soccer player-NOM cheer-PASSIVE  
“The soccer player is being cheered by the fans.”

**Procedure** The procedure was identical to Experiment 1.

## Results and Discussion

The voice priming effects observed in Experiment 1 was replicated (Figure 3). Participants produced more passive sentences after passive primes (6%) than active primes (2%). The main effect of prime type was marginally significant in the participant analysis and significant in the analysis on items ( $F_1(1,19)=3.43, p=.08$ ;  $F_2(1,19)=4.63, p<.05$ ). Participants showed a tendency to reuse a passive sentence after passive primes, even though function assignment only was shared between the passive primes and the produced passive targets, and the thematic order was also unmatched, priming effects were found again. Thus the results from

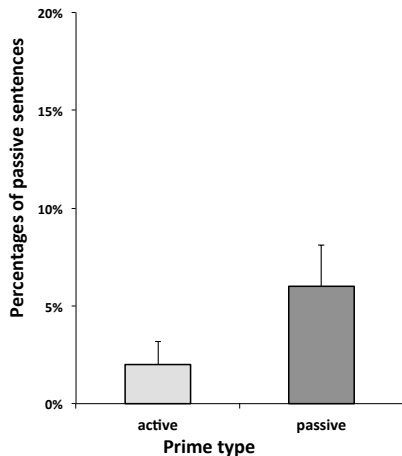


Figure 3: Percentages of passives in each condition.

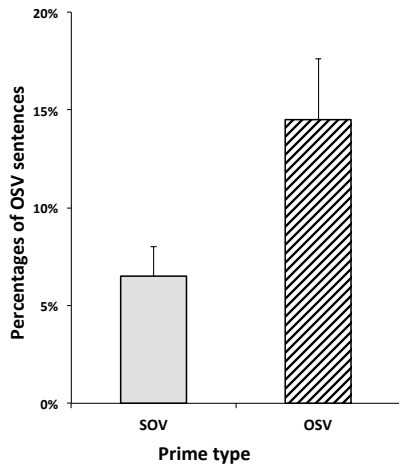


Figure 4: Percentages of OSVs in each condition.

Experiment 2 suggest that the voice priming effects in Experiment 2 are due to the processes in the function assignment stage.

In addition to the voice priming effects, we discovered another interesting finding. Participants produced more scrambled sentences (scrambled active sentences: OSV-active) after passive primes (OSV word order, 14.5%) than after active primes (SOV word order, 6.5%) (Figure 4). Analyses revealed a main effect of word order priming effect was marginally significant in the participant analysis and significant in the analysis on items (SOV-active) ( $F_1(1,19)=3.94, p=.06$ ;  $F_2(1,19)=11.50, p<.005$ ), meaning that more OSV-active were produced after the scrambled passive primes (OSV-passive) than canonical active primes.

The word order priming effects that were found between OSV-passive and OSV-active suggest that the word order determination processes occurred in the constituent assembly stage (place the subject before or after the non-subject) and could be primed separately from the function

assignment stage (assign an agent role to subject or non-subject NPs).

Results from the Experiment 1 and Experiment 2 strongly support that the function assignment stage and the constituent assembly stage are independent of each other in the grammatical encoding level, and the grammatical function roles and word order are determined separately during the processes.

However, since we did not manipulate the conditions in active voice systematically, whether the word order priming effects across voice between OSV-passive and OSV-active pairs in Experiment 2 reflected purely word order priming effects is unknown. Moreover whether function assignment and word order determination processes interacted with each other and if so how the processes are computed is left unknown. We examine these questions in Experiment 3.

### Experiment 3

In Experiment 3 we manipulated function assignment and word order independently and simultaneously in order to examine the interaction between two stages directly.

#### Method

**Participants** Thirty-three students, all native speakers of Japanese at Hiroshima University participated in the experiment (25 females and 8 males, the mean age was 21.5 years). Participants were paid 500 yen for their participation. Throughout the experiment, a female and a male native speakers of Japanese acted as the confederate.

**Materials** The prime was either an active or a passive voice sentence with canonical (SOV) or scramble (OSV) word order. (1) SOV-active prime, (2) SOV-passive prime OSV-active prime, and (3) OSV-passive prime were similar to those in Experiment 1 and 2. In addition, we created a (4) OSV-active prime condition. Two sets of 120 items were created: a confederate set and a participant set. The confederate set consisted of 120 sentences (32 prime and 88 filler sentences), and the participant set consisted of 120 simple black and white line drawings (32 target and 88 filler pictures). In addition, we prepared two sets of 60 simple black and white line drawings for a picture recognition task.

- (4) sakkaa sensyu-o      sapootaa-ga      ooensi-teiru.  
 soccer player-ACC      fans-NOM      cheer  
 “The fans are cheering the soccer player.”

**Procedure** The procedure was identical to Experiment 1 and 2.

#### Results and Discussion

Again, the voice priming effects and the word order priming effects were observed.

Figure 5 shows the proportion of SOV-passive sentences in each conditions. The interaction of voice and word order

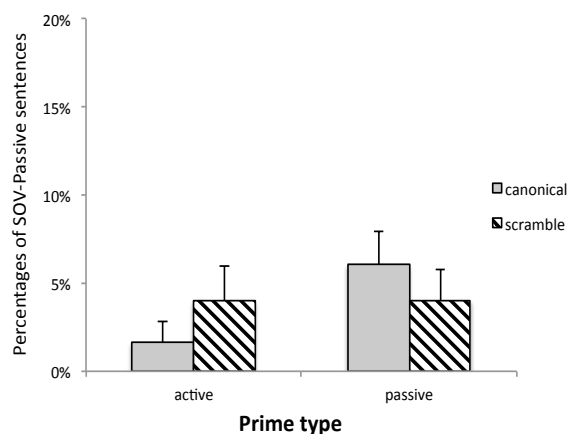


Figure 5: Percentages of SOV-passives in each condition.

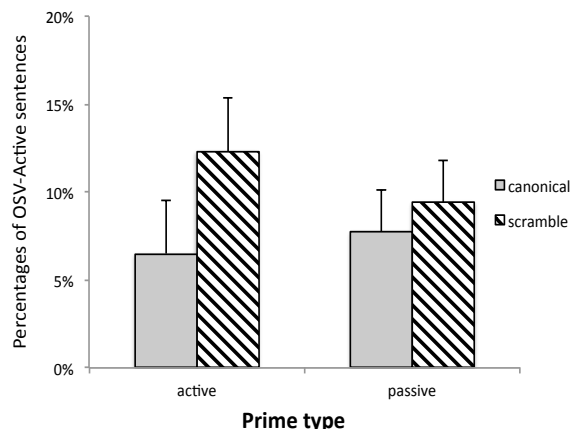


Figure 6: Percentages of OSV-actives in each condition.

was significant in the participant analysis and marginally significant in the analysis on items ( $F_1(1,32)=7.43$ ,  $p<.05$ ;  $F_2(1,31)=3.31$ ,  $p=.08$ ).

Contrasts across SOV word order conditions revealed that more SOV-passive sentences were produced after SOV-passive primes (6%) than after SOV-active primes (1.6%) ( $F_1(1,32)=9.06$ ,  $p<.01$ ;  $F_2(1,31)=7.30$ ,  $p<.05$ ). This pattern of result is identical to those in Experiment 1, which reflected a voice priming effect.

The tendency of difference was also found in active prime conditions – more SOV-passive sentences were produced after OSV-active (4%) than after SOV-active (1.6%) ( $F_1(1,32)=4.73$ ,  $p<.05$ ;  $F_2(1,31)=2.17$ ,  $p=.15$ ). The voice priming effects between OSV-active prime and SOV-passive target suggests that the similarity between these sentence types, although both function assignment and word order were different between the prime/target pairs. Although the conceptual factors were strictly controlled in Experiment 3, this finding suggests the locus of this priming effects might occur outside the grammatical encoding level.

In contrast, the results showed that in OSV-active responses only the main effect of word order was significant ( $F_1(1,32)=7.01$ ,  $p<.05$ ;  $F_2(1,31)=5.28$ ,  $p<.05$ ). More OSV-actives were produced after OSV primes than after SOV primes (Figure 6). This result suggests that after an active voice is selected at function assignment stage, the word order is determined at the constituent assembly stage, and affect by both OSV-active and OSV-passive primes. Again the results showed that word order determination occurs indecently from the function assignment stage.

Taken together, the findings from the voice priming suggest that various factors affect and complicate the processes at function assignment stage (assign an agent role to subject or non-subject NPs). However, the foundlings from the word order priming suggest that the processes at constituent assembly stage are rather straightforward only be influenced by word order determination.

## General Discussion

Three experiments investigated the production processes of the two stages in the grammatical encoding level. In particular, we examined whether the function assignment stage and the constituent assembly stage are processed separately or simultaneously and how the processes are computed.

Using a structural priming paradigm in Experiment 1, we demonstrated that the priming effects of passive voice occur in the function assignment stage, which has a major influence on the choice of active or passive voice, even when the active voice is highly favored. The results from Experiment 2 further confirmed that the priming effects are raised due to the processes in the function assignment stage. Moreover the word order priming effects were found across active and passive voice, which suggested that word order was determined independently of the function assignment stage, in the constituent assembly stage. The results from Experiment 3 further confirmed that the function assignment stage and the constituent assembly stage are computed separately during the production processes. Moreover the pattern of the voice priming effects and the word order priming effects are different according to the stages.

The results from Experiment 3 show the similarity between the OSV-actives and SVO-passives, in this case both function assignment and word order were different between the prime/target pairs. This is consistent with previous studies (e.g., Prat-Sala & Branigan, 2000). Prat-Sala and Branigan (2000) demonstrated that Spanish speakers tend to produce dislocated active sentences (OVS word order) instead of using passive sentences when the patient is more salient than the agent (a language setting which facilitates the production of passives). The authors interpreted that the results suggest that conceptual factors may be associated with variations in word order directly, and the way in which the factors influence is constrained by the syntactic options available within that language.

Prat-Sala and Branigan (2000) suggested the similarity between the dislocated actives and passives when producing an event in which the patient is more salient than the agent (e.g., a sentence emphasis on the patient). In Japanese, both passive sentence and OSV-active are assumed to put emphasis on the patient (Shibatani, 1985, 1990). An explanation of the priming effects found in Experiment 3 between OSV-actives and SVO-passives, might occur outside the grammatical encoding level, that it is due to the similarity between the OSV-active and SOV-passive, which both put emphasis on the patient. For example, after participants heard SOV-passive primes from the confederate, both passive voice and OSV word order were more available than usual at the grammatical encoding level. Although the event type between prime and target pairs were controlled, processing a passive prime might explain the emphasis on the patient roles. If the emphasis of the patient role is persistent between the prime and target pairs, because the active voice was highly favored as conceptual factors were controlled in our study, participants might have chosen an OSV-active over an OSV-passive or SOV-passive, in order to make the patient more salient than the agent. This explanation is consistent with the findings in Bernolet, Hartsuiker, and Pickering (2009), and can be explained by the current model of sentence production which assumes that perspective meaning, a sub-component in the preverbal message level, can affect constituent assembly stage directly. Further studies are needed to examine whether the priming effects between SOV-passive and OSV-active in our study were due to word priming effects or the persistence of emphasis in the patient role. Further studies are needed to examine whether the persistence of emphasis in the patient role can influence the priming effects in the grammatical encoding level.

In conclusion, our findings provide support for a two-stage production model with two separate stages within the grammatical encoding level, namely the function assignment stage and the constituent assembly stage. Moreover, our findings suggest that various factors affect and complicate the processes at the function assignment stage; however, the processes at the constituent assembly stage are rather straightforward by the syntactic options available in that language.

### Acknowledgments

This research was supported by Grant-in-Aid for Scientific Research #20320060 (PI: Hiromu Sakai, Hiroshima University) from the Japan Society for the Promotion of Science. I would like to thank Mikihiro Tanaka, Jeff Pannell, Kyoko Sakamoto, Hiroe Maeda, Kanako Ono and Takuya Kubo for helpful discussions and helping with data collection.

### References

Bernolet, S., Hartsuiker, R.J., and Pickering, M.J. (2007) Shared syntactic representations in bilinguals: evidence

for the role of word-order repetition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5), 931-949.

Bernolet, S., Hartsuiker, R.J., and Pickering, M.J. (2009) Persistence of emphasis in language production: A cross-linguistic approach. *Cognition*, 112, 300-317.

Bock, J. K. (1986) Syntactic persistence in language production. *Cognitive Psychology*, 18, 355-387.

Bock, K. (1989) Closed-class immanence in sentence production. *Cognition*, 31, 163-186.

Bock, J.K., & Levelt, W.J.M. (1994). Language production: Grammatical encoding. In M. Gernsbacher (Ed.), *Handbook of psycholinguistics*. San Diego, CA: Academic Press.

Bock, J.K., and Loebell, H. (1990) Framing sentences, *Cognition*, 35, 1-39.

Branigan, H.P., Pickering, M.J., and Cleland, A.A. (2000) Syntactic co-ordination in dialogue. *Cognition*, 75, B13-25.

Branigan, H.P., Pickering, M.J., & Tanaka, M. (2008). Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, 118, 172-189.

Chang, F., Bock, K., & Goldberg, A.E. (2003). Can thematic roles leave traces of their places? *Cognition*, 90, 29-49.

Ferreira, V.S., & Slevc, L.R. (2007). Grammatical encoding. In G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics*. Oxford: Oxford University Press.

Hartsuiker R.J., & Westenberg, C. (2000). Persistence of word order in written and spoken sentence production. *Cognition*, 75, B27-B39.

Shin, J.-A., and Christianson, K. (2009) Syntactic processing in Korean-English bilingual production: evidence from cross-linguistic structural priming. *Cognition*, 112, 175-180.

Shibatani, M. (1985). Passives and related constructions: A prototype analysis. *Language*, 64, 821-848.

Shibatani, M. (1990). *The languages of Japan*. Cambridge: Cambridge University Press.

Tanaka, M. (2007) *The representation of conceptual and syntactic information during sentence production*. Unpublished Ph.D. dissertation, The University of Edinburgh.

Vigliocco, G., and Nicol, J. (1998) Separating hierarchical relations and word order in language production: is proximity concord syntactic or linear? *Cognition*, 68, B13-B29.

# A Window of Perception When Diverting Attention? Enhancing Recognition For Explicitly Presented, Unattended, and Irrelevant Visual Stimuli by Target Alignment

**Andrew D. Dewald** (adewald@hawaii.edu)

Department of Psychology, University of Hawaii at Manoa  
2530 Dole Street, Honolulu, HI 96822 USA

**Scott Sinnett** (ssinnett@hawaii.edu)

Department of Psychology, University of Hawaii at Manoa  
2530 Dole Street, Honolulu, HI 96822

## Abstract

Irrelevant, but overtly presented, stimuli that are temporally aligned with an attended target in a separate task are later inhibited in a recognition task (Dewald, Dumas, & Sinnett, 2011). This is contrary to findings in the perceptual learning literature where facilitation has been observed for later recognition of irrelevant motion stimuli, albeit after extensive exposure rates. Here, we adapted previous work to include higher exposure rates, and subsequently observed a reversal in inhibition in favor of enhanced recognition performance. Participants responded to immediate picture repetitions in a stream of line drawings while ignoring simultaneously presented superimposed words. A surprise test measured recognition for the unattended words. Words that had previously appeared simultaneously with a repeated picture were recognized significantly *more often* than words that had appeared with non-repeating pictures. The findings suggest that the exposure rate and the quantity of irrelevant stimuli can have a significant impact on whether perception is inhibited or facilitated.

## Introduction

Throughout the past decade, researchers have explored how information is processed when it is explicitly or implicitly presented, and the fate of this information when it receives or does not receive direct and focused attention (Dewald, Sinnett, & Dumas, 2011; Mack & Rock, 1998; Rees, Russell, Frith, & Driver, 1999; Seitz & Watanabe, 2003, 2005; Sinnett, Costa, & Soto-Faraco, 2006; Tsushima, Sasaki, & Watanabe, 2006; Tsushima, Seitz, & Watanabe, 2008; Swallow & Jiang, 2010). For instance, several investigations have shown significant perceptual learning enhancements in the absence of focused attention for stimuli that are, in fact, presented below the threshold for visual awareness (i.e., implicitly presented) (Seitz & Watanabe, 2003; 2005; Watanabe, Náñez, & Sasaki, 2001). More recently and contrary to these findings, Tsushima and colleagues (Tsushima et al., 2006; Tsushima et al., 2008) presented evidence suggesting that when the implicit stimulus is made explicit (i.e., observable), a later inhibition is observed. Thus, it would appear that facilitation or inhibition is dependent on whether or not stimulus presentation is sub- or suprathreshold. Furthermore, all of these investigations purport that a synchronous temporal

relationship between the irrelevant stimulus (motion in these investigations) and a separate but attended target in the exposure stage is critical to observing these facilitatory or inhibitory effects in a later recognition task (i.e., the nonsynchronous condition is baseline).

Demonstrating learning enhancements for irrelevant stimuli, Seitz and Watanabe (2003) had participants take part in a series of experiments in which improved motion perception for an irrelevant, subthreshold motion, was postulated to be due to the establishment of a temporal relationship between task-relevant and task-irrelevant stimuli (Seitz & Watanabe, 2003; 2005, see also Watanabe et al., 2001 for a further example using a similar paradigm). Briefly, participants were required to identify a differently colored letter in a rapid serial visual presentation of letters. This primary task was superimposed over an irrelevant background motion stimulus that involved an array of moving dots, of which a small subset moved in coherence. Note that the coherently moving dots (5%) were implicit in nature, demonstrated by chance motion discrimination during pre-testing. While every letter was accompanied with an array of moving dots, the direction of the subthreshold coherent motion was always the same for the target letters of the primary task, while remaining random for non-target letters. The implicit motion synchronized with the presence of the task- target (the different colored letter) was later identified significantly more often than the other motions (i.e. those accompanying non-target letters) in a motion detection task (see, Seitz and Watanabe, 2003). It was hypothesized that when the irrelevant motion and task-target were presented simultaneously during exposure, the learning associated with attention being directed to the detection of the task-relevant features of the task-target would also be applied to the task-irrelevant stimulus of background motion, despite the motion being subthreshold and attention being explicitly directed towards the primary attention-demanding task. Further bolstering this account is the fact that significant improvements in the motion discrimination task were observed only for the motion direction that was paired with the presence of the task-target (i.e., simultaneously presented). However, it is important to recall that the irrelevant stimuli temporally aligned with the presence of the task-target were subthreshold in nature.



Given the differences in perception for subthreshold and suprathreshold information, a logical ensuing question would be to explore what happens when above threshold irrelevant motion stimuli are presented during the exposure stage.

Addressing this very question, Tsushima et al (2008) conducted a similar experiment using explicit rather than implicit motion. Specifically, their experiment included a condition with suprathreshold motions (i.e., 50% coherence) during the primary task (i.e., exposure stage), in addition to a subthreshold motion condition (i.e., 5% coherence). Although one might perhaps expect that higher motion coherence would lead to stronger learning effects (when compared to lower motion coherence signals) due to an arguably strengthened perceptual signal (Britten, Shadlen, Newsome, & Movshon, 1992), the opposite occurred. Facilitation was found only for the subthreshold stimulus levels, while an inhibition was observed for suprathreshold exposure (i.e., explicit presentation). Combined, these findings suggest that strong (overt) target-aligned irrelevant features are subject to attentional inhibition, as the initial task requires attention to be directed to the letter stream. This possibly prevents the strong, but irrelevant, feature from being learned. However, subthreshold motion is not subject to this same inhibitory control, and might therefore be learned and facilitate later perceptual performance.

Other experimental paradigms have utilized different approaches and stimuli to further investigate the way information is processed during dual-task performance. Interestingly, despite the explicit presentation of their stimuli, opposite findings from Tsushima et al (2006) have been observed, with a facilitation for overt stimuli that was presented simultaneously with an attended target from a separate task. For instance, Swallow and Jiang (2010; see also Lin et al., 2010 for a similar example of a paradigm utilizing temporally aligned targets) completed a series of experiments suggesting an “attentional boost” (i.e., facilitation) for simultaneously presented information in a dual-task paradigm, rather than an inhibition as witnessed by Tsushima et al (2006; 2008). In their experiment, participants monitored a stream of pictures of various scenes. A series of distractor items (small black superimposed squares) were simultaneously paired with the presentation of each picture. Participants were required to remember as many of the presented scenes as possible, in addition to monitor the distractor stream for the presence of an “odd-ball” color change (i.e., the presence of a white square rather than a black square). In a subsequent forced choice recognition test for the picture scenes, an enhanced recognition for pictures that had been presented simultaneously with the presence of the target (i.e., the ‘odd-ball’ color change) in the distractor stream was observed (i.e., an attentional boost).

Of particular note to Swallow and Jiang’s (2010) findings is that participants were required to attend to both streams of information simultaneously (encode the pictures as well as detect an “odd ball” target). Recall that in the paradigm utilized by Seitz and Watanabe (2003) as well as

Tsushima et al (2008), participants were instructed to detect a target in one stream (i.e., identify a differently colored letter) while being exposed to the background coherent motion, but not actively attend to the background motion (i.e., the motion was in fact irrelevant at this stage in the experiment). Regardless of these procedural differences, it is important to note that Swallow and Jiang’s (2010) findings are based on the presentation of above threshold stimuli, much like the 50% coherent motion in the Tsushima et al (2008) paradigm. A second important difference is the nature of the stimuli. Swallow and Jiang used pictures (see also Lin et al., 2010) while irrelevant motion was used in the other examples. It is likely that explicitly presented pictures are processed much differently than irrelevant and implicitly presented motion. Combined, the procedural differences between these two paradigms may be a contributing reason as to why a contradictory pattern of results has been observed (i.e., facilitation for Swallow & Jiang, 2010 and inhibition for Tsushima et al., 2008).

Recent work from our laboratory (Dewald et al., 2011) examined the temporal pairing of highly salient and overtly presented task-relevant and task-irrelevant stimuli with an adapted inattention blindness (IB) task (see Rees et al., 1999; Sinnett et al., 2006 for similar examples), requiring participants to monitor a stream of pictures with synchronized superimposed words, and respond to immediate repetitions in the picture stream while ignoring the word stream. Following the picture repetition detection task, participants were administered a surprise recognition test for the (ignored) words that had been superimposed over the pictures during the repetition detection task. Words that had been temporally aligned with the presence of a task-target (i.e., an immediately repeating picture) were subsequently recognized significantly below chance levels (i.e., inhibited), while words that had been temporally aligned with non-targets (i.e., a non repeating picture) were recognized at chance levels. These findings dovetail with the conclusions of Tsushima et al (2006, 2008) and suggest that suprathreshold presented stimuli will be inhibited rather than facilitated if presented simultaneously with an attended target from a separate task.

Critical the experiment presented here, a key component of the paradigm used by Tsushima et al (2006, 2008) and Seitz and Watanabe (2003, 2005) is that the exposure rates of implicit and explicit motion were extremely high, often including multiple days of exposure including thousands of trials. On the other hand, our previous research included a mere 200 total trials lasting approximately 10 minutes. Furthermore, while our research (see also Swallow & Jiang, 2010) paired many different irrelevant stimuli (i.e., words) with the relevant task target (picture repetition), the sub- and superthreshold motion paradigms paired only a single motion direction with all targets in the primary task. This important point is explained further below. The present investigation therefore aims to further extend this research by exploring whether recognition for a highly salient stimulus will in fact be facilitated if presented more

frequently, and without other competing stimuli. That is, it will address the question of whether using a higher frequency of presentation for a salient irrelevant stimulus could in fact modulate the previously observed inhibition and lead to facilitation effects (i.e., akin to the attentional boost effect; see Swallow & Jiang, 2010).

There are a number of possible outcomes in the present investigation. First, based on the conclusions of Seitz and Watanabe (2003), in regards to the temporal pairing of task-relevant stimuli (e.g., immediate picture repetitions) and task-irrelevant stimuli (e.g., superimposed words), it could be predicted that this synchronization will at the least establish a relationship that will affect perception for task-irrelevant stimuli aligned with task-relevant targets. It should be noted however, that studies by Tsushima et al (2006; 2008) as well as Dewald et al. (2011) suggest that if an irrelevant stimuli is explicitly presented (rather than implicitly, as used by Seitz & Watanabe, 2003), the temporal relationship between task-relevant and task-irrelevant stimuli may lead to an inhibited performance in a later recognition task for the irrelevant items. Critical to the question at hand, our previous demonstration used 50 different words rather than a single word (i.e., akin to a single motion; see Dewald et al., 2011). This increased quantity of words could have augmented the salient nature of the stimuli, thereby necessitating that they be ignored in order to complete the primary task of detecting picture repetitions, and consequently lead to inhibitory effects. To better approximate the conditions used by Seitz and Watanabe (2003), in the present experiment we utilize only one high-level irrelevant stimulus (a specific written word) that may lead to an enhanced performance rather than inhibited. This would therefore be analogous to the same unchanging suprathreshold motion always paired with the presence of the task-target in Tsushima et al (2008). We argue that although explicit and suprathreshold in nature, coherent motion detection fails to be processed to the same level as semantic words (see Borst & Egelhaaf, 1989 for a review of visual motion detection). Therefore, it can be predicted that the repeated exposure of a single word that is temporally aligned with an attended target will lead to a later facilitation in recognition for that word when compared to words that were not temporally aligned with an attended target (i.e., similar to the attentional boost, see Swallow & Jiang, 2010). We predict that these results will surface due to a synergy of higher salience for words along with increased exposure levels to the word that is temporally aligned with the presence of a task-target (note, all words were presented in equal frequency).

## Method

**Participants.** Sixteen participants (n=16) were recruited from the University of Hawai'i at Manoa in exchange for course credit. Participants were naïve to the experiment and had normal or corrected to normal vision.

**Materials.** A total of 50 pictures were selected from the Snodgrass and Vanderwart (1980) picture database. The pictures (on average 5 to 10 cm's) were randomly rotated  $\pm 30$  degrees from upright so as to ensure the difficulty of the task in each version of the experiment (see also Rees et al., 1999). Each of these pictures was combined with eight one to two syllable, high frequency English words (average length of 5 letters; range 4-6) selected from the MRC psycholinguistic database (Wilson, 1988). The overall average frequency of the eight selected words was 361 per million, ranging between 135 and 782. The words were displayed in bold, capitalized letters in Arial font at a size of 24 points. Each word was superimposed over a picture and the picture-word stimuli did not exceed 10 cm horizontally or vertically. Care was taken to ensure that picture-word combinations did not have any semantic relationship.

A stream of 960 picture-word concatenated items was created. Repeated pictures acted as the task relevant-targets. The presentation stream was broken into eight blocks of 120 trials in which an immediate picture repetition occurred on average of one out of every eight trials, equating to an average of 15 task-relevant target repetitions per block, for a total of 120 trials of exposure to a task-relevant target (and specific word, see below).

Eight words were selected to be superimposed over the 960 trial picture stream. This was done to parallel the quantity of items and exposure to irrelevant stimuli as well as mimic the dependent measure employed by Watanabe et al. (2001; see also Seitz & Watanabe 2003, 2005). The eight selected words can be thought of as the eight coherent motions. That is, the same single word was always temporally aligned with the presentation of an immediately repeated picture target. All eight words were presented equally. The presentation was pseudorandomized so that on average one out of every eight trials was an immediate picture repetition (and therefore the presentation of the same superimposed task-irrelevant target word). Only one superimposed word was aligned with all of the immediately repeated pictures for each participant. This single word was randomized between the eight words between participants (2 participants per word) so as to control for any possible differences that may have existed regarding particular word saliency.

A surprise recognition test was administered after the completion of the repetition detection task. The test consisted of a total of sixteen words from which half came from the previously viewed visual stream, while the other half consisted of foil words that had never been seen before. The foils were words that had never been used in the exposure stage of the experiment, but were taken from the same database and had an average frequency of 236 per million with a range of 165-399. The eight non-foil (previously seen in the picture repetition task) words were words that were either temporally aligned with the task-relevant target, (i.e., superimposed over the immediate repetition of a picture), or were not temporally aligned with the task-relevant target (i.e., superimposed over non-

immediately repeating pictures). For ease of explaining and reference, words synchronized with task-relevant targets will be referred to as *target-aligned* words and those not aligned with task-relevant targets will be referred to as *non-aligned* words (see also Dewald et al., 2011).

Both the repetition detection and word recognition tasks were randomized and presented by DMDX software (<http://www.u.arizona.edu/jforster/dmdx.htm>) one at a time, written in bold, capitalized letters in Arial font at a size of 24 points, in an identical fashion as they were displayed in the previous stream. The words in the recognition test remained on the screen until a response was made.

## Procedure

Participants were required to attend to the picture stream (i.e., ignore the simultaneously presented superimposed words) and respond to immediate picture repetitions by pressing the ‘G’ key on the keyboard of the computer. Each item in the picture-word presentation was presented for 350 ms with a 150-ms inter-stimulus interval (ISI; blank screen) between each item for a stimulus onset asynchrony (SOA) of 500 ms (see Figure 1). Before the first experimental block, a training block of eight trials was given and repeated until participants were familiar and comfortable with the task.

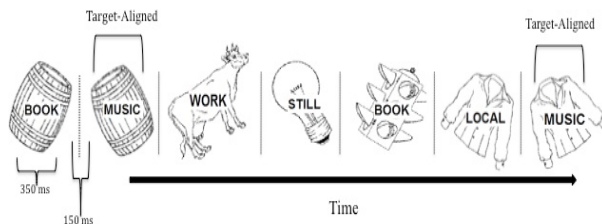


Figure 1. Each picture–word stimulus was presented for 350 ms and then replaced by a blank screen for 150 ms before the next stimulus. The task is to respond to picture repetitions and ignore the words, for which a surprise recognition test is later given. Note that the *target-aligned* word (in the present case ‘MUSIC’) remains the same across all repetition trials.

Immediately after the repetition detection task, the surprise word recognition test was administered to all participants. Words were displayed individually on the center of the screen in the same size and font as previously presented in the repetition detection task, and remained on the screen until the participant made a response. Participants were instructed to press the “B” key if they had seen the word during the repetition detection task or, instead, the “V” key if they had not seen the word before.

## Results

Overall task performance for word recognition was 78%, which was significantly above chance levels ( $t(15) = 6.58$ ,  $p < .01$ ). In order to address the question at hand, that is, if recognition performance is enhanced for words that had

appeared with a picture repetition, the average correct recognition score for *target-aligned* words (words superimposed over immediately repeated pictures) and *non-aligned* words was compared against chance, and also against each other. In this case, recognition for *target-aligned* (87.5%,  $SE=.85$ ) and *non-aligned* words (68.4%,  $SE=.37$ ) was significantly better than chance ( $t(15) = 10.91$ ,  $p < .001$  and  $t(15) = 4.89$ ,  $p < .001$  respectively). Most importantly, recognition for *target-aligned* words was significantly better than performance for *non-aligned* words ( $t(15) = 2.31$ ,  $p = .03$ ; see Figure 2).

Additionally, the correct rejection of foil words was compared with overall performance for *target-aligned* and *non-aligned* words. No significant differences between recognition for *target-aligned* words and correct rejections surfaced (*target-aligned*: 87.5%,  $SE=.85$  vs.  $CR: 88.6\%$ ,  $SE=.03$ ,  $t(15) = -.07$ ,  $p = .994$ ). There was a significant difference between correctly recognizing *non-aligned* words and correctly rejecting foil words (*non-aligned*: 68.4%,  $SE=.37$  vs.  $CR: 88.6\%$ ,  $SE=.03$ ,  $t(15) = 3.69$ ,  $p < .002$ ). Further demonstrating the overall accuracy of word recognition, there were significantly fewer false alarms (FA) (i.e., incorrectly identifying a foil word as having been present during the picture repetition task), when compared with correct foil rejections ( $FA = 10.4\%$  vs.  $CR = 88.6\%$ ,  $t(15) = 17.08$ ,  $p < .001$ ), *target-aligned* words (87.5%,  $t(15) = 17.65$ ,  $p < .001$ ), and *non-aligned* (68.4%,  $t(15) = 13.79$ ,  $p < .001$ ) words as well as significantly fewer false alarms than to be expected by chance ( $t(15) = 19.11$ ,  $p < .0001$ ).

Lastly, confirming that participants were able to successfully perform the initial repetition task, an analysis was also conducted on the accuracy of the primary task of detecting immediate target repetitions. Overall, participants accurately detected target repetitions (**Hits**: 75%,  $SE=0.20$  vs. **Misses**: 25%,  $SE=0.79$ ,  $t(15) = 11.83$ ,  $p < .001$ ).

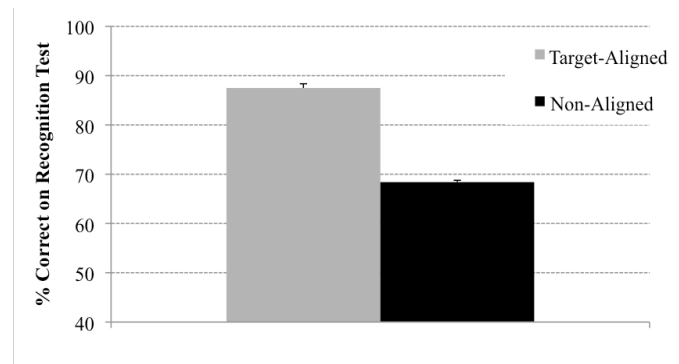


Figure 2. Recognition percentages and standard error bars for Target-Aligned (grey bar) and Non-Aligned (black bar) words in the surprise word recognition test after attending to the picture stream.

## Discussion

The present findings extend investigations exploring how above threshold, but unattended information is processed

when it appears simultaneously with an attended target (Dewald et al, 2011; Lin et al, 2010; Seitz & Watanabe, 2003, 2005; Swallow & Jiang, 2010; Tsushima et al., 2006, 2008). Critically, despite all task-irrelevant words being recognized better than chance, words that were temporally aligned with a picture target (i.e., repetition) were recognized at significantly higher rates when compared with words that did not appear with a picture target. Specifically, performance for both the *target-aligned* (88%) and *non-aligned* (68%) words was significantly better than chance, but *target-aligned* words were recognized significantly better than *non-aligned* words. Accordingly, this suggests that, at least in the present case, temporally pairing explicitly presented irrelevant stimuli with relevant target stimuli facilitates subsequent recognition of the irrelevant stimuli.

Performance on the ability to correctly reject foil words further bolsters the notion that target-alignment is critical for enhanced perception. Participants were significantly more accurate at correctly rejecting foil words in the recognition test than they were at correctly identifying *non-aligned* words (88% vs. 68%, respectively), while no significant differences were observed when comparing performance for rejecting foils with accuracy for *target-aligned* words (88% vs. 87%). These data suggest an “attentional boost” (see Swallow & Jiang, 2010) for irrelevant stimuli as long as the stimuli were presented simultaneously with a target in the picture repetition task, despite not receiving direct attention.

The superior recognition of *target-aligned* words is analogous to the enhanced motion detection for coherent motion displays aligned with relevant task-targets observed by Seitz and Watanabe (2003). Interestingly, this enhancement was found only after exposure to implicitly presented motion directions. When explicitly presented (i.e., suprathreshold) an inhibition was observed for motion recognition, rather than a facilitation (see Tsushima et al., 2006; 2008). These findings would seemingly indicate that the enhancement or inhibition of performance is contingent on whether the initial irrelevant but target-aligned stimulus was implicitly or explicitly presented. The present findings fail to support either notion. That is, *facilitation* for target-aligned stimuli was observed, despite all irrelevant stimuli being overtly presented.

Aligning with our result, other recent investigations have observed an “attentional boost” (i.e., facilitation) for simultaneously and overtly presented information in a dual-task paradigm (Lin et al., 2010; Swallow & Jiang, 2010). However, significant procedural differences warrant discussion. Specifically, Swallow and Jiang’s (2010) task (see also, Lin et al, 2010) required participants to attend to both the irrelevant and relevant streams of information, rather than only one stream as done here (see also Seitz & Watanabe, 2003; but see Swallow & Jiang, 2011, discussed below). Thus, the key distinguishing feature between the respective paradigms is whether or not attention was simultaneously directed to both streams of information or if

only a single stream receives attention while the other is ignored.

Addressing the differing approaches, Swallow and Jiang (2011) required to only attend to one stream of information while ignoring the other. In this case, participants were required to direct their attention to the detection of the “odd ball” target only and not required to pay attention to the concurrently presented picture scenes (i.e., the images were now *irrelevant* to the task). A surprise recognition test administered for the pictures (task-irrelevant) revealed that the attentional boost effect did not occur when the background scenes were made task-irrelevant, thereby suggesting that temporal alignment was not sufficient to foster the attentional boost of target-aligned irrelevant stimuli.

This elimination of the attentional boost seemingly contradicts the present findings demonstrating an enhanced performance for the recognition of target-aligned stimuli. A potential explanation for this could be found in the differing amounts of irrelevant stimuli utilized in each respective paradigm. For instance, in Swallow and Jiang’s (2011) experiment over 100 different stimuli (pictures) served as the irrelevant items, while presently there was only one target-aligned stimulus and seven non-aligned items (i.e., analogous to Tsushima et al., 2008). This considerable difference in stimulus set size could be why we observed an attentional boost (despite attention not being directed to both streams). Note, the elimination of the attentional boost effect is actually a null effect when comparing aligned with non-aligned items, therefore it is difficult to make a strong claim regarding these findings.

Further support for the speculation that the quantity of irrelevant items modulates whether the boost is observed or not, comes from previous research conducted by our laboratory (Dewald et al., 2011). In this work, the paradigm was identical to the present (i.e., detect picture repetitions followed by a word recognition test), but an inhibition for target-aligned stimuli was observed. Importantly, the number of irrelevant items (50) was more in line with Swallow and Jiang’s (2011) recent work. Furthermore, as stated before, using semantic words rather than pictures could also be a contributing reason why we continue to see an attentional boost here, despite attending to only one stream of information (pictures).

Setting aside the null effect found by Swallow and Jiang (Experiment 4, 2011), studies involving a specific analysis of recognition performance for target-aligned vs. non-aligned stimuli show recognition for target-aligned stimuli to be either *inhibited* (i.e., recognized significantly below chance levels) or *facilitated*, with a key difference being whether the target stimulus was implicitly or explicitly presented. It is evident that in the present experiment, words were recognized at high levels (78% of the time), and when accounting for target alignment, those synchronized with repetitions were indeed better recognized (a 20% improvement). In fact, when attention was most utilized in the repetition detection task, subsequent performance for the

target-aligned word was best. Recall however, that Tsushima et al. (2008) observed an inhibition for target-aligned, suprathreshold irrelevant stimuli. Although both Tsushima et al. (2008) and the present investigation utilized an above threshold, irrelevant stimuli (written words/ high motion coherence, respectively), it may be that the increased saliency and frequency of presentation of the written words lead to a facilitation rather than an inhibition.

Combined, the present findings and previous research offer insight into how irrelevant information is processed when it is presented simultaneously with an attended target. Under certain circumstances, unattended stimuli can be perceived and affect behavior (see also Dewald et al., 2011; Seitz & Watanabe, 2003; Tsushima et al., 2008). When using a low salience, irrelevant stimulus, there appears to be a relationship between explicit or implicit presentations, fostering either an inhibition (Tsushima et al., 2008), or facilitation (Seitz & Watanabe, 2003) respectively. Regardless of the explicit or implicit nature of the stimuli presentation, synchronization of task-relevant and task-irrelevant signals is crucial to establishing a relationship that will affect perception for task-irrelevant stimuli aligned with task-relevant targets, however, this relationship can be modulated when dividing attention across streams of information (Swallow Jiang, 2010, i.e., a facilitation is observed for explicitly presented stimuli). Most importantly to the present investigation however, when using a high exposure rate to salient, explicitly presented irrelevant stimuli in a limited stimulus set a perceptual window is created in which an “attentional boost” (Swallows & Jiang, 2010) surfaces for stimuli that do not receive direct and focused attention as long as they are presented simultaneously with the relevant target of a separate task.

## References

- Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Science U.S.A.*, 90, 5718–5722.
- Borst, A., & Egelhaaf, M. (1989). Principles of visual motion detection. *Trends in Neurosciences*, 12(8), 297-306.
- Dewald, A.D., Sinnett, S., & Dumas, L.A.A. (2011). Conditions of directed attention inhibit recognition performance for explicitly presented target-aligned irrelevant stimuli. *Acta Psychologica*.138, 60-67.
- Driver, J., & Spence, C. (2004). Crossmodal spatial attention: Evidence from human performance. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention*. Oxford, UK: Oxford University Press.
- Lin, J.Y., Pye, A.D., Murray, & Boynton, G.M. (2010). Enhanced memory for scenes presented at relevant points in time. *PLoS Biol*, 8(3), E1000337.
- Rees, G., Russell, C., Frith, C. D., & Driver, J. (1999). Inattention blindness versus inattentional amnesia for fixated but ignored words. *Science*, 286, 2504-2507.
- Roelfsema, P. R., van Ooyen, A., & Watanabe, T. (2009). Perceptual learning rules based on reinforces and attention. *Trends in Cognitive Sciences*, 14(2), 64-71.
- Seitz, A. R., Kim, R., & Shams, L. (2006). Sound facilitates visual learning. *Current Biology*, 16, 1422-1427.
- Seitz, A. R. & Watanabe, T. (2003). Psychophysics: Is subliminal learning really passive? *Nature*, 422, 36.
- Seitz, A. R. & Watanabe, T. (2005). A unified model for perceptual learning. *Trends in Cognitive Science*, 9 (7), 329-334.
- Sinnett, S., Costa, A., & Soto-Faraco, S. (2006). Manipulating inattention blindness within and across sensory modalities. *Quarterly Journal of experimental Psychology*, 59(8), 1425-1442
- Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 1 74–215.
- Swallow K. M., & Jiang, Y. V. (2010). The attentional boost effect: Transient increases in attention to one task enhance performance in a second task. *Cognition*, 115, 118-132.
- Swallow K.M., & Jiang, Y. V. (2011). The role of timing in the attentional boost effect. *Attention, Perception, and Psychophysics*, 73, 389-404.
- Tootell, B., Silverman, M. S., & De Valois, R. L. (1995). Spatial frequency columns in primary visual cortex. *Science*, 214(4522), 813-815.
- Tsushima, Y., Sasaki, Y., & Watanabe, T. (2006). Greater disruption due to failure of inhibitory control on an ambiguous distractor. *Science*, 314, 1786-1788.
- Tsushima, Y., Seitz, A. R., & Watanabe, T. (2008). Task-irrelevant learning occurs only when the irrelevant feature is weak. *Current Biology*, 18(12), 516-517.
- Watanabe, T., Náñez, Y., & Sasaki, S. (2001). Perceptual learning without perception. *Nature*, 413, 844–848.
- Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, version 2. *Behavioural Research Methods, Instruments and Computers*, 20, 6-11.

# A Computational Logic Approach to the Suppression Task

Emmanuelle-Anna Dietz and Steffen Hölldobler (`{dietz,sh}@iccl.tu-dresden.de`)

International Center for Computation Logic, TU Dresden  
D-01062 Dresden, Germany

Marco Ragni (`ragni@cognition.uni-freiburg.de`)

Center for Cognitive Science, Friedrichstraße 50  
D-79098 Freiburg, Germany

## Abstract

A novel approach to human conditional reasoning based on the three-valued Łukasiewicz logic is presented. We will demonstrate that the Łukasiewicz logic overcomes problems the so far proposed Fitting logic has in reasoning with the suppression task. While adequately solving the suppression task, the approach gives rise to a number of open questions concerning the use of Łukasiewicz logic, unique fixed points, completion versus weak completion, explanations, negation, and sceptical versus credulous approaches in human reasoning.

**Keywords:** Łukasiewicz logic; computational logic; suppression task; human reasoning.

## Introduction

An interesting study is the *suppression task*, in which Byrne (1989) has shown that graduate students with no previous exposure to formal logic did suppress previously drawn conclusions when additional information became available. Interestingly, in some instances the previously drawn conclusions were valid whereas in other instances the conclusions were invalid with respect to classical two-valued logic. Consider the following example: *If she has an essay to finish then she will study late in the library* and *She has an essay to finish*. Then most subjects (96%) conclude: *She will study late in the library*. If subjects, however, receive an additional conditional: *If the library stays open she will study late in the library* then only 38% of the subjects conclude: *She will study late in the library*. This shows, that, although the conclusion is still correct, the conclusion is suppressed by an additional conditional. This is an excellent example for human capability to draw *non-monotonic* inferences.

Table 1 shows the abbreviations that will be used throughout the paper, whereas Table 2 gives an account of the findings of Byrne (1989). As we are using a formal language, propositions like “She will go to the library” (abbreviated  $L$ ) will be represented by propositional variables like  $l$ , with the intended interpretation that if  $l$  is *true* ( $\top$ ), then “She will go to the library”. Taking a naive propositional approach, we can represent  $A$  by the implication  $e \leftarrow l$ , where the propositional variables  $e$  and  $l$  represent the facts  $E$  and  $L$ , respectively, and so on.

It is straightforward to see that classical two-valued logic cannot model the suppression task adequately: Applying the classical logical consequence operator to some instances of the suppression task (like  $A, C, E$ ) yields qualitatively wrong answers, due to the monotonic nature of the classical logic.

Table 1: The suppression task (Byrne, 1989) and used abbreviations. Subjects received conditionals  $A, B$  or  $C$  and facts  $E, \bar{E}, L$  or  $\bar{L}$  and had to draw inferences.

$A$	<i>If she has an essay to finish then she will study late in the library.</i>
$B$	<i>If she has a textbook to read then she will study late in the library.</i>
$C$	<i>If the library stays open she will study late in the library.</i>
$E$	<i>She has an essay to finish.</i>
$\bar{E}$	<i>She does not have an essay to finish.</i>
$L$	<i>She will study late in the library.</i>
$\bar{L}$	<i>She will not study late in the library.</i>

Table 2: The drawn conclusions in the experiment of Byrne.

Conditional(s)	Fact	Experimental Findings
$A$	$E$	96% of subjects conclude $L$ .
$A, B$	$E$	96% of subjects conclude $L$ .
$A, C$	$E$	38% of subjects conclude $L$ .
$A$	$\bar{E}$	46% of subjects conclude $\bar{L}$ .
$A, B$	$\bar{E}$	4% of subjects conclude $\bar{L}$ .
$A, C$	$\bar{E}$	63% of subjects conclude $\bar{L}$ .
$A$	$L$	53% of subjects conclude $E$ .
$A, B$	$L$	16% of subjects conclude $E$ .
$A, C$	$L$	55% of subjects conclude $E$ .
$A$	$\bar{L}$	69% of subjects conclude $\bar{E}$ .
$A, B$	$\bar{L}$	69% of subjects conclude $\bar{E}$ .
$A, C$	$\bar{L}$	44% of subjects conclude $\bar{E}$ .

Consequently, at least a non-monotonic operator is needed. As argued by Stenning and van Lambalgen (2008)<sup>1</sup> human reasoning should be modeled by, first, reasoning towards an appropriate representation and, second, by reasoning with respect to this representation. As appropriate representation Stenning and van Lambalgen propose logic programs under completion semantics based on the three-valued logic used by Fitting (1985), which itself is based on the three-valued Kleene (1952) logic.

Unfortunately, some technical claims made by Stenning and van Lambalgen (2008) are wrong concerning their second step. Hölldobler and Kencana Ramli (2009b) have shown that the three-valued logic proposed by Fitting is inadequate for the suppression task. Somewhat surprisingly, the suppression task can be adequately modeled if the three-valued

<sup>1</sup>There is an earlier publication (Stenning & van Lambalgen, 2005), but Michiel van Lambalgen advised us to refer to their textbook.

Łukasiewicz (1920) logic is used. The paper gives an account of this finding and discusses a variety of consequences of this new logic and some open questions.

### Adequacy

Computational approaches to explain human reasoning should be *cognitively adequate*. Usually, the concept of adequacy is measured by distinguishing between conceptual and inferential adequacy (Strube, 1996). In our context, a system is *conceptually adequate* if it appropriately represents human knowledge. *Inferential adequacy* measures whether the computations behave similarly to human reasoning. It is common in Cognitive Science to evaluate theories by performing reasoning experiments on subjects. For instance, Knauff (1999) investigates which kind of information humans use when representing and remembering spatial arrangements in Allen’s interval calculus. In Computer Science, one commonly used hypothesis is, that if computational models are biologically plausible then they should also behave similar to the biological brain (Herrmann & Ohl, 2009). However, until now there are no implemented models which easily process computations given a large amounts of data or efficiently deal with incomplete information. These aspects are fundamental for elementary reasoning processes. In this paper, we evaluate the inferential adequacy of our computational logic approach by examining that our approach gives the same answers as subjects in the suppression task experiments.

### A Computational Logic Approach

Stenning and van Lambalgen (2008) have proposed to use logic programs under completion semantics and based on a three-valued logic to model the suppression task. In particular, they suggest that human reasoning is modeled by, first, reasoning towards an appropriate representation or logical form (conceptual adequacy) and, second, reasoning with respect to this representation (inferential adequacy).

In the following we introduce three-valued logics and, in particular, the Łukasiewicz logic. As the chosen representation are logic programs, such programs are introduced next together with their (weak) completion. We adopt the reasoning step towards an appropriate logical form from Stenning and van Lambalgen (2008). Thereafter, we discuss three-valued models for logic programs under the Łukasiewicz semantics and, in particular, the model intersection property which entails the existence of least models. We show that the conclusions drawn with respect to these least models correspond to the findings in (Byrne, 1989) and conclude that the derived logic programs under Łukasiewicz semantics are inferentially adequate for the suppression task.

In order to investigate inferential adequacy we consider the semantic operator associated with logic programs as defined by Stenning and van Lambalgen (2008). For each program  $\mathcal{P}$ , this operator admits a least fixed point, which is equal to the least Łukasiewicz model of  $\mathcal{P}$ . At this point we are able to discuss the technical problems in (Stenning & van Lambalgen, 2008), while showing that they do not occur if we use

Łukasiewicz semantics. Finally, we add abduction to the approach and show that sceptical reasoning is needed in order to model the suppression task adequately.

### Three-Valued Logics

Three-valued logics were introduced by Łukasiewicz (1920). Table 3 gives the truth tables for different three-valued logics. The symbols  $\top$ ,  $\perp$ , and  $\text{U}$  denote *true*, *false*, and *unknown*, respectively. For instance, if  $F$  is mapped to  $\perp$  and  $G$  is mapped to  $\top$  then their conjunction ( $F \wedge G$ ) is mapped to  $\perp$  and their disjunction ( $F \vee G$ ) is mapped to  $\top$ . By introducing

Table 3: The three-valued logics

$F \parallel \bar{F}$	$F \parallel \bar{F}$	$F \parallel \bar{F}$	$\wedge$	$\vee$	$\leftarrow_L$	$\leftrightarrow_L$	$\leftarrow_K$	$\leftrightarrow_S$
$\top \parallel \perp$	$\top \parallel \perp$	$\top \parallel \perp$	$\top$	$\top$	$\top$	$\top$	$\top$	$\top$
$\perp \parallel \top$	$\perp \parallel \top$	$\perp \parallel \top$	$\perp$	$\top$	$\top$	$\perp$	$\top$	$\perp$
$\text{U} \parallel \text{U}$	$\text{U} \parallel \text{U}$	$\text{U} \parallel \text{U}$	$\text{U}$	$\top$	$\top$	$\text{U}$	$\top$	$\perp$
$\perp \parallel \top$	$\perp \parallel \top$	$\perp \parallel \top$	$\perp$	$\top$	$\top$	$\perp$	$\top$	$\perp$
$\perp \parallel \text{U}$	$\perp \parallel \text{U}$	$\perp \parallel \text{U}$	$\perp$	$\top$	$\top$	$\text{U}$	$\top$	$\perp$
$\top \parallel \text{U}$	$\top \parallel \text{U}$	$\top \parallel \text{U}$	$\top$	$\top$	$\top$	$\text{U}$	$\top$	$\perp$
$\text{U} \parallel \perp$	$\text{U} \parallel \perp$	$\text{U} \parallel \perp$	$\perp$	$\top$	$\top$	$\text{U}$	$\top$	$\perp$
$\text{U} \parallel \top$	$\text{U} \parallel \top$	$\text{U} \parallel \top$	$\perp$	$\top$	$\top$	$\text{U}$	$\top$	$\perp$
$\text{U} \parallel \text{U}$	$\text{U} \parallel \text{U}$	$\text{U} \parallel \text{U}$	$\text{U}$	$\top$	$\top$	$\text{U}$	$\top$	$\perp$

a third truth value, there are various options to define the truth tables for the connectives. For example, Kleene (1952) introduced an implication ( $\leftarrow_K$ ), whose truth table is identical to the Łukasiewicz implication ( $\leftarrow_L$ ) except in the cases where precondition and conclusion are both mapped to  $\text{U}$ : In this case, the implication itself is mapped to  $\text{U}$  by Kleene, but to  $\top$  by Łukasiewicz. The set of connectives under Łukasiewicz semantics is  $\{\neg, \wedge, \vee, \leftarrow_L, \leftrightarrow_L\}$ . Kleene also introduced a so-called *strong equivalence*, where the truth value  $\top$  is assigned to  $F \leftrightarrow_S G$  if  $F$  and  $G$  are assigned to identical truth values, and  $\perp$  is assigned otherwise. Fitting (1985) combined the truth tables for  $\neg$ ,  $\vee$ ,  $\wedge$  from Łukasiewicz with the Kleene implication and strong equivalence for investigations within Logic Programming. We will call this combination the *Fitting semantics* where the set of connectives is  $\{\neg, \wedge, \vee, \leftarrow_K, \leftrightarrow_S\}$ <sup>2</sup>. Stenning and van Lambalgen (2008) use Fitting semantics without giving a reason for this particular choice.

### Logic Programs

A *logic program* is a finite set of expressions of the form

$$A \leftarrow B_1 \wedge \dots \wedge B_n, \quad (1)$$

where  $n \geq 1$ ,  $A$  is an atom, and each  $B_i$ ,  $1 \leq i \leq n$ , is either a literal,  $\top$ , or  $\perp$ .  $A$  is called *head* and  $B_1 \wedge \dots \wedge B_n$  is called *body* of the *clause* (1). A clause of the form  $A \leftarrow \top$  is called *positive fact*, whereas a clause of the form  $A \leftarrow \perp$  is called *negative fact*. In the sequel,  $\mathcal{P}$  shall denote a logic program.

Consider the following transformation for a given  $\mathcal{P}$ :

<sup>2</sup>We believe that Fitting had termination analysis of logic programs in his mind when he selected this particular logic.



1. All clauses with the same head  $A \leftarrow Body_1, A \leftarrow Body_2, \dots$  are replaced by  $A \leftarrow Body_1 \vee Body_2 \vee \dots$ .
2. If an atom  $A$  is not the head of any clause in  $\mathcal{P}$  (and, thus, is *undefined* in  $\mathcal{P}$ ) then add  $A \leftarrow \perp$ .
3. All occurrences of  $\leftarrow$  are replaced by  $\leftrightarrow$ .

The resulting set is called *completion* of  $\mathcal{P}$  ( $c\mathcal{P}$ ). If step 2 is omitted, then the resulting set is called *weak completion* of  $\mathcal{P}$  ( $wc\mathcal{P}$ ). Consider  $\mathcal{P} = \{p \leftarrow q\}$ , then  $c\mathcal{P} = \{p \leftrightarrow q, q \leftrightarrow \perp\}$ .  $c\mathcal{P}$  entails that  $p$  and  $q$  are mapped to  $\perp$ . Reasoning with respect to the completion of a logic program is non-monotonic. For instance, if  $\mathcal{P}' = \mathcal{P} \cup \{q \leftarrow \top\}$ , then  $c\mathcal{P}'$  entails that  $p$  and  $q$  are mapped to  $\top$ . The process of weak completion can be associated with the human interpretation of conditionals as biconditionals (Evans, Newstead & Byrne, 1993).

### Reasoning Towards an Appropriate Logical Form

Stenning and van Lambalgen (2008) have argued that the first step in modeling human reasoning is reasoning towards an appropriate logical form. In particular, they argue that conditionals shall not be encoded by implications straight away but rather by licenses for implications. For example, the conditional  $A$  should be encoded by the clause  $l \leftarrow e \wedge \overline{ab}_1$ , where  $ab_1$  is an *abnormality* predicate which expresses that something abnormal is known. In other words,  $l$  holds if  $e$  holds and nothing abnormal is known.

We think that Stenning and van Lambalgen (2008) adequately model the representational part of the suppression task and adopt this reasoning step. Our focus is on the inferential aspect of their approach. In the first two columns of Table 4 the programs obtained for the first six examples of the suppression task are depicted. For instance, in  $\mathcal{P}_{ACE}$  we have that *She will study late in the library* if either *She has an essay to finish* and *Nothing abnormal ( $ab_1$ ) is known* or *She has a textbook to read* and *Nothing abnormal ( $ab_3$ ) is known*. The predicates  $ab_1$ ,  $ab_2$  and  $ab_3$  represent different kinds of abnormality. For instance,  $ab_1$  is true when *The library does not stay open* and  $ab_3$  is true when *She does not have an essay to finish*.

### Three-Valued Models for Logic Programs

A (*three-valued*) *interpretation* is a mapping from a propositional language to the set  $\{\top, \perp, \text{U}\}$  of truth values. It is quite common to represent interpretations by tuples of the form  $\langle I^\top, I^\perp \rangle$ , where  $I^\top$  contains all atoms which are mapped to  $\top$ ,  $I^\perp$  contains all atoms which are mapped to  $\perp$ ,  $I^\top \cap I^\perp = \emptyset$ , and all atoms which occur neither in  $I^\top$  nor in  $I^\perp$  are mapped to  $\text{U}$ . Let  $\mathcal{P}$  be a program and  $I$  an interpretation.  $I$  is a (*three-valued*) *model under Łukasiewicz semantics* for  $\mathcal{P}$  ( $I \models_L \mathcal{P}$ ) if and only if each clause occurring in  $\mathcal{P}$  is mapped to  $\top$  using the truth table depicted in Table 3. Likewise,  $\models_F$  can be defined with respect to the Fitting semantics. For instance, consider  $\mathcal{P} = \{p \leftarrow q\}$ . Then under Łukasiewicz semantics we have three different models  $\langle \{p, q\}, \emptyset \rangle$ ,  $\langle \emptyset, \{p, q\} \rangle$  and  $\langle \emptyset, \emptyset \rangle$ . Under Fitting semantics only the first two interpretations are models, because if  $p$  and  $q$  are mapped to  $\text{U}$  then  $p \leftarrow q \in \mathcal{P}$  is mapped to  $\text{U}$  as well.

Table 4: A summary of the computational logic approach to the suppression task (Part 1).

$\mathcal{P}$	clauses	$wc\mathcal{P}$	$\text{lm}_L wc\mathcal{P}$	Byrne
$\mathcal{P}_{AE}$	$l \leftarrow e \wedge \overline{ab}_1$ $ab_1 \leftarrow \perp$ $e \leftarrow \top$	$l \leftrightarrow e \wedge \overline{ab}_1$ $ab_1 \leftrightarrow \perp$ $e \leftrightarrow \top$	$\langle \{e, l\}, \{ab_1\} \rangle$	96% $L$
$\mathcal{P}_{ABE}$	$l \leftarrow e \wedge \overline{ab}_1$ $l \leftarrow t \wedge \overline{ab}_2$ $ab_1 \leftarrow \perp$ $ab_2 \leftarrow \perp$ $e \leftarrow \top$	$l \leftrightarrow (e \wedge \overline{ab}_1) \vee (t \wedge \overline{ab}_2)$ $ab_1 \leftrightarrow \perp$ $ab_2 \leftrightarrow \perp$ $e \leftrightarrow \top$	$\langle \{e, l\}, \{ab_1, ab_2\} \rangle$	96% $L$
$\mathcal{P}_{ACE}$	$l \leftarrow e \wedge \overline{ab}_1$ $l \leftarrow o \wedge \overline{ab}_3$ $ab_1 \leftarrow \overline{o}$ $ab_3 \leftarrow \overline{e}$ $e \leftarrow \top$	$l \leftrightarrow (e \wedge \overline{ab}_1) \vee (o \wedge \overline{ab}_3)$ $ab_1 \leftrightarrow \overline{o}$ $ab_3 \leftrightarrow \overline{e}$ $e \leftrightarrow \top$	$\langle \{e\}, \{ab_3\} \rangle$	38% $L$
$\mathcal{P}_{A\overline{E}}$	$l \leftarrow e \wedge \overline{ab}_1$ $ab_1 \leftarrow \perp$ $e \leftarrow \perp$	$l \leftrightarrow e \wedge \overline{ab}_1$ $ab_1 \leftrightarrow \perp$ $e \leftrightarrow \perp$	$\langle \emptyset, \{e, l, ab_1\} \rangle$	46% $\overline{L}$
$\mathcal{P}_{AB\overline{E}}$	$l \leftarrow e \wedge \overline{ab}_1$ $l \leftarrow t \wedge \overline{ab}_2$ $ab_1 \leftarrow \perp$ $ab_2 \leftarrow \perp$ $e \leftarrow \perp$	$l \leftrightarrow (e \wedge \overline{ab}_1) \vee (t \wedge \overline{ab}_2)$ $ab_1 \leftrightarrow \perp$ $ab_2 \leftrightarrow \perp$ $e \leftrightarrow \perp$	$\langle \emptyset, \{e, ab_1, ab_2\} \rangle$	4% $\overline{L}$
$\mathcal{P}_{AC\overline{E}}$	$l \leftarrow e \wedge \overline{ab}_1$ $l \leftarrow o \wedge \overline{ab}_3$ $ab_1 \leftarrow \overline{o}$ $ab_3 \leftarrow \overline{e}$ $e \leftarrow \perp$	$l \leftrightarrow (e \wedge \overline{ab}_1) \vee (o \wedge \overline{ab}_3)$ $ab_1 \leftrightarrow \overline{o}$ $ab_3 \leftrightarrow \overline{e}$ $e \leftrightarrow \perp$	$\langle \{ab_3\}, \{e, l\} \rangle$	63% $\overline{L}$

### Reasoning with Respect to Least Models

In order to identify the desired model of a certain program, we reason with respect to their least models. Least models are guaranteed to exist if the model intersection property holds:

$$\begin{aligned} \cap \{I \mid I \models_L \mathcal{P}\} &\models_L \mathcal{P}, \\ \cap \{I \mid I \models_L wc\mathcal{P}\} &\models_L wc\mathcal{P}. \end{aligned}$$

In Hölldobler and Kencana Ramli (2009b) it was shown that the model intersection property holds for (weakly completed) programs under Łukasiewicz semantics. The model intersection property for programs does not hold under Fitting semantics. Consider again  $\mathcal{P} = \{p \leftarrow q\}$ , then both,  $\langle \{p, q\}, \emptyset \rangle$  and  $\langle \emptyset, \{p, q\} \rangle$ , are models for  $\mathcal{P}$ , whereas their intersection  $\langle \emptyset, \emptyset \rangle$  is not a model for  $\mathcal{P}$  under Fitting semantics.

The third column of Table 4 shows the weak completions of the programs encoding the first six examples of the suppression task. Column 4 in Table 4 depicts the corresponding least models where  $\text{lm}_L$  denotes the least model of its argument under Łukasiewicz semantics. The last column shows the results of the suppression task. Specifically we find that

$$\begin{aligned} \text{lm}_L wc\mathcal{P}_{AE} &= \langle \{e, l\}, \{ab_1\} \rangle && \models_L l \\ \text{lm}_L wc\mathcal{P}_{ABE} &= \langle \{e, l\}, \{ab_1, ab_2\} \rangle && \models_L l \\ \text{lm}_L wc\mathcal{P}_{ACE} &= \langle \{e\}, \{ab_3\} \rangle && \not\models_L l \vee \overline{l} \\ \text{lm}_L wc\mathcal{P}_{A\overline{E}} &= \langle \emptyset, \{e, l, ab_1\} \rangle && \models_L \overline{l} \\ \text{lm}_L wc\mathcal{P}_{AB\overline{E}} &= \langle \emptyset, \{e, ab_1, ab_2\} \rangle && \not\models_L l \vee \overline{l} \\ \text{lm}_L wc\mathcal{P}_{AC\overline{E}} &= \langle \{ab_3\}, \{e, l\} \rangle && \models_L \overline{l} \end{aligned}$$

where  $\not\models_L$  means that a given formula cannot be concluded. Our approach coincides with the seemingly favored results of the suppression task and thus appears to be adequate.

### Computing Least Models

In Computational Logic, least models are usually computed as least fixed points of appropriate semantic operators (see, e.g., Apt & Emden, 1982). Stenning and van Lambalgen (2008) devised such an operator for programs discussed herein: Let  $I$  be an interpretation in  $\Phi_P(I) = \langle J^\top, J^\perp \rangle$ , where

$$\begin{aligned} J^\top &= \{A \mid \text{there exists } A \leftarrow \text{body} \in \mathcal{P} \text{ with } I(\text{body}) = \text{true}\}, \\ J^\perp &= \{A \mid \text{there exists } A \leftarrow \text{body} \in \mathcal{P} \text{ and} \\ &\quad \text{for all } A \leftarrow \text{body} \in \mathcal{P} \text{ we find } I(\text{body}) = \text{false}\}. \end{aligned}$$

As shown in Hölldobler and Kencana Ramli (2009b) for any  $\mathcal{P}$ , the least fixed point of  $\Phi_P$  is identical to  $\text{lm}_{3LWC} \mathcal{P}$  and can be computed by iterating  $\Phi_P$  starting with the empty interpretation. The following example shows how the least model of  $\mathcal{P}_{ACE}$  is computed starting with interpretation  $I_0 = \langle \emptyset, \emptyset \rangle$ :

$$\begin{aligned} I_1 &= \Phi_{\mathcal{P}_{ACE}}(I_0) = \langle \{e\}, \emptyset \rangle \\ I_2 &= \Phi_{\mathcal{P}_{ACE}}(I_1) = \langle \{e\}, \{ab_3\} \rangle = \Phi_{\mathcal{P}_{ACE}}(I_2) \end{aligned}$$

where  $I_2$  is the least fixed point of  $\Phi_{\mathcal{P}_{ACE}}$ . This is not a model under Fitting semantics because the clause  $l \leftarrow o \wedge ab_3 \in \mathcal{P}_{ACE}$  is mapped to  $\perp$  and not to  $\top$  such as under Łukasiewicz semantics. This is a counter example for Lemma 4(1.) in Stenning and van Lambalgen (2008) which states that the least fixed point of the  $\Phi_P$  operator under Fitting semantics is the minimal model of  $\mathcal{P}$ . Another statement made by Stenning and van Lambalgen (2008), Lemma 4(3.) states, that all models of  $c\mathcal{P}$  are fixed points of  $\Phi_P$  and every fixed point is a model. Consider the completion of  $\Phi_{\mathcal{P}_{ABE}}$ , then  $t$  is mapped to  $\perp$  and therefore  $l$  is mapped to  $\perp$  as well. However, its least fixed point is  $\langle \emptyset, \{e, ab_1, ab_2\} \rangle$  where  $t$  and  $l$  are undefined. This example also shows that reasoning under Fitting semantics and with respect to the completion of a program is not adequate as only 4% conclude  $\bar{L}$  in this case.

### Unique Fixed Point

As mentioned in the previous subsection, the least fixed point of the operator  $\Phi_P$  can be computed by iterating  $\Phi_P$  starting with the empty interpretation. However, if the operator is a contraction then by the Banach Contraction Theorem (Banach, 1922) the operator has a unique fixed point which can be computed by iterating the operator starting with an arbitrary interpretation. As shown in Hölldobler and Kencana Ramli (2009a),  $\Phi_P$  is a contraction if  $\mathcal{P}$  is acyclic<sup>3</sup>. All programs shown in Table 4 are acyclic.

### Abduction

The second part of the suppression task deals with the affirmation of the consequent and modus tollens. These reasoning

processes can best be described as abductive, that is, a plausible explanation is computed given some observation. Following Kakas, Kowalski, and Toni (1993) we consider an *abductive framework* consisting of a program  $\mathcal{P}$  as knowledge base, a set  $\mathcal{A}$  of abducibles consisting of the (positive and negative) facts for each undefined predicate symbol in  $\mathcal{P}$ ,<sup>4</sup> and the logical consequence relation  $\models_L^{lmwc}$ , where  $\mathcal{P} \models_L^{lmwc} F$  if and only if  $\text{lm}_{3LWC} \mathcal{P}(F) = \top$  for the formula  $F$ . As *observations* we consider literals.

Let  $\langle \mathcal{P}, \mathcal{A}, \models_L^{lmwc} \rangle$  be an abductive framework and  $O$  an observation.  $O$  is *explained* by  $\mathcal{E}$  if and only if  $\mathcal{E} \subseteq \mathcal{A}$ ,  $\mathcal{P} \cup \mathcal{E}$  is satisfiable, and  $\mathcal{P} \cup \mathcal{E} \models_L^{lmwc} O$ . Usually, minimal explanations are preferred. In case there exist several minimal explanations, then two forms of reasoning can be distinguished.  $F$  follows *sceptically* from program  $\mathcal{P}$  and observation  $O$  ( $\mathcal{P}, O \models_s F$ ) if and only if  $O$  can be explained and for all minimal explanations  $\mathcal{E}$  we find  $\mathcal{P} \cup \mathcal{E} \models_L^{lmwc} O$ , whereas  $F$  follows *credulously* from  $\mathcal{P}$  and  $O$  ( $\mathcal{P}, O \models_c F$ ) if and only if there exists a minimal explanation  $\mathcal{E}$  such that  $\mathcal{P} \cup \mathcal{E} \models_L^{lmwc} O$ .<sup>5</sup> For instance, consider the following two programs under sceptical reasoning:

1.  $\mathcal{P}_{AB}$  where  $O = l$ :  $\mathcal{A} = \{e \leftarrow \top, e \leftarrow \perp, t \leftarrow \top, t \leftarrow \perp\}$  and  $\text{lm}_{3LWC} \mathcal{P}_{AB} = \langle \emptyset, \{ab_1, ab_2\} \rangle$ . There are two minimal explanations with either  $\{e \leftarrow \top\}$  and  $\{t \leftarrow \top\}$ . Thus, we cannot conclude whether *She has an essay to finish* or not.
2.  $\mathcal{P}_{AC}$  where  $O = l$ :  $\mathcal{A} = \{e \leftarrow \top, e \leftarrow \perp, o \leftarrow \top, o \leftarrow \perp\}$  and  $\text{lm}_{3LWC} \mathcal{P}_{AC} = \langle \emptyset, \emptyset \rangle$ . There is only one minimal explanation  $\{e \leftarrow \top, o \leftarrow \top\}$  and thus *She has an essay to finish*.

Table 5 depicts the programs, the observations and the minimal explanations for the second part of the suppression task in the second, third, and fourth column, respectively. The second last column shows the least model of the weak completion of the union of the program and the minimal explanation under the Łukasiewicz semantics and the final one shows the results of the suppression task. If we reason sceptically with respect to these least models, then we obtain

$$\begin{array}{ll} \mathcal{P}_A, l \models_s e, & \mathcal{P}_A, \bar{l} \models_s \bar{e}, \\ \mathcal{P}_{AB}, l \not\models_s e, & \mathcal{P}_{AB}, \bar{l} \models_s \bar{e}, \\ \mathcal{P}_{AC}, l \models_s e, & \mathcal{P}_{AC}, \bar{l} \not\models_s \bar{e}, \end{array}$$

which are adequate answers if compared to the seemingly favored results of the suppression task. One should observe that a credulous agent concludes  $e$  from  $\mathcal{P} = \mathcal{P}_{AB}$  and  $O = l$ , which according to Byrne (1989) only 16% of the subjects did.

## Open Questions

### Łukasiewicz Logic

This logic was selected because the technical bugs in Stenning and van Lambalgen (2008) can be solved by switching from Fitting to Łukasiewicz semantics. In particular, the

<sup>3</sup>A program  $\mathcal{P}$  is acyclic if there exists a numbering for all propositional variables such that for all clauses in  $\mathcal{P}$  the value of the head is strictly larger than the value of the literals in the body.

<sup>4</sup>Recall that  $A$  is *undefined* in  $\mathcal{P}$  if and only if  $\mathcal{P}$  does not contain a clause of the form  $A \leftarrow \text{Body}$ .

<sup>5</sup>See (Hölldobler, Philipp, & Wernhard, 2011) for more details.

Table 5: A summary of the computational logic approach to the suppression task (Part 2). The cases  $\mathcal{P} = \mathcal{P}_{AB}, O = l$  and  $\mathcal{P} = \mathcal{P}_{AC}, O = \bar{l}$  have two minimal extensions.

$\mathcal{P}$	clauses	$O$	$\mathcal{E}$	$\text{Im}_{\text{wc}}(\mathcal{P} \cup \mathcal{E})$	Byrne
$\mathcal{P}_A$	$l \leftarrow e \wedge \overline{ab_1}$ $ab_1 \leftarrow \perp$	$l$	$e \leftarrow \top$	$\langle \{e, l\}, \{ab_1\} \rangle$	53% $E$
$\mathcal{P}_{AB}$	$l \leftarrow e \wedge \overline{ab_1}$ $l \leftarrow t \wedge \overline{ab_2}$ $ab_1 \leftarrow \perp$ $ab_2 \leftarrow \perp$	$l$	$e \leftarrow \top$	$\langle \{e, l\}, \{ab_1, ab_2\} \rangle$	16% $E$
			$t \leftarrow \top$	$\langle \{l, t\}, \{ab_1, ab_2\} \rangle$	
$\mathcal{P}_{AC}$	$l \leftarrow e \wedge \overline{ab_1}$ $l \leftarrow o \wedge \overline{ab_3}$ $ab_1 \leftarrow \bar{o}$ $ab_3 \leftarrow \bar{e}$	$l$	$e \leftarrow \top$ $o \leftarrow \top$	$\langle \{e, l, o\}, \{ab_1, ab_3\} \rangle$	55% $E$
$\mathcal{P}_A$	$l \leftarrow e \wedge \overline{ab_1}$ $ab_1 \leftarrow \perp$	$\bar{l}$	$e \leftarrow \perp$	$\langle \emptyset, \{e, l, ab_1\} \rangle$	69% $\bar{E}$
$\mathcal{P}_{AB}$	$l \leftarrow e \wedge \overline{ab_1}$ $l \leftarrow t \wedge \overline{ab_2}$ $ab_1 \leftarrow \perp$ $ab_2 \leftarrow \perp$	$\bar{l}$	$e \leftarrow \perp$ $t \leftarrow \perp$	$\langle \emptyset, \{e, l, t, ab_1, ab_2\} \rangle$	69% $\bar{E}$
$\mathcal{P}_{AC}$	$l \leftarrow e \wedge \overline{ab_1}$ $l \leftarrow o \wedge \overline{ab_3}$ $ab_1 \leftarrow \bar{o}$ $ab_3 \leftarrow \bar{e}$	$\bar{l}$	$e \leftarrow \perp$	$\langle \{ab_3\}, \{e, l\} \rangle$	44% $\bar{E}$
			$o \leftarrow \perp$	$\langle \{ab_1\}, \{l, o\} \rangle$	

model intersection property holds under Łukasiewicz semantics. Hence, for each program  $\mathcal{P}$  a least model does exist which can be computed as least fixed point of the associated semantic operator  $\Phi_{\mathcal{P}}$ . Moreover, a rigorous study has revealed that the suppression task can be adequately modeled under Łukasiewicz semantics, whereas this does not hold for Fitting semantics. Nevertheless, the main question of whether Łukasiewicz logic is adequate for human reasoning is still open. For example, in the Łukasiewicz logic the Deduction Theorem does not hold<sup>6</sup>. Hence, it would be interesting to see how humans deal with the deduction theorem. Can other typical human reasoning problems like the Wason (1968) selection task be adequately modeled under Łukasiewicz semantics?

### Unique Fixed Point

For each program  $\mathcal{P}$  shown in Table 4 the operator  $\Phi_{\mathcal{P}}$  is a contraction. Thus, there is a unique fixed point, which can be computed by iterating  $\Phi_{\mathcal{P}}$  on some initial interpretation. Consequently, if in the suppression task subjects are influenced towards some initial non-empty interpretation, their performance should not differ provided that they have enough time to compute the least fixed point; it should differ, however, if they are interrupted before the least fixed point is computed and asked to reason with respect to the interpretation com-

<sup>6</sup>A logic satisfies the *Deduction Theorem* if for any finite set of formulae  $\Phi = \{\phi_1, \phi_2, \dots, \phi_n\}$  and any formula  $\psi$  the following holds:  $\Phi \models \psi$  if and only if  $\models (\phi_1 \wedge \phi_2 \wedge \dots \wedge \phi_n) \rightarrow \psi$ .

puted so far.

### Completion versus Weak Completion

The program  $\mathcal{P}_{AB\bar{E}}$  served as an example to illustrate that completion is inadequate for the suppression task whereas weak completion is adequate. Likewise, Hölldobler et al. (2011) have shown in a detailed study that the programs mentioned in Table 5 together with their minimal explanations must be weakly completed in order to adequately model the suppression task, whereas completion does not. Are there other human reasoning episodes which support the claim that weak completion is adequate? Even if so, the problem remains to explicitly add negative facts (in the reasoning step towards an appropriate logical form) for those predicates, which should be mapped to  $\perp$  like  $ab_1$  in the program  $\mathcal{P}_{AE}$ .

### Sceptical versus Credulous Reasoning

The case of program  $\mathcal{P} = \mathcal{P}_{AB}$  and observation  $O = l$  in Table 5 shows that agents must reason sceptically in order to adequately model this case. Whereas this is a striking case for sceptical reasoning, the case  $\mathcal{P} = \mathcal{P}_{AC}$  and  $O = \bar{l}$  is less convincing. A sceptical agent will not conclude  $\bar{e}$ , whereas a credulous agent will conclude  $\bar{e}$ . Compared to the corresponding case  $(A, C, \bar{L})$  shown in Table 2, 44% of the subjects conclude  $\bar{E}$ . Unfortunately, Byrne (1989) (and related publications that we are aware of) gives no account of the distribution of the answers given by those subjects who did not conclude  $\bar{E}$ . Hence, at the moment we can argue in favor of a sceptical agent (*the majority of the subjects did not conclude  $\bar{E}$* ), but – given the complete distribution – it may be the case that one can argue in favor of a credulous agent (*there are more subjects concluding  $\bar{E}$  than subjects concluding  $E$  and subjects answering “I don’t know”*).

### Explanations

The approach presented in this paper is based on minimal explanations. Although, there are findings corroborating the human preference of minimal explanations (over non-minimal ones) (Ormerod, Manktelow, & Jones, 1993) – this holds only partially (Johnson-Laird, Girotto, & Legrenzi, 2004). Computational models of abduction typically generate explanations iteratively such that minimal explanations are generated first. How are minimal explanations computed by humans? What happens if there are more than one minimal explanation?

### Negation

In the presented approach positive information is preferred over negative one. Consider, for example, the program  $\mathcal{P} = \{q \leftarrow \top, q \leftarrow \perp\}$ . The least model of  $\text{wc}\mathcal{P}$  is  $\langle \{q\}, \emptyset \rangle$  and, hence, an agent reasoning with respect to this model will conclude  $q$ . Is this consistent with human reasoning? The presented approach could be extended to include integrity constraints like  $\perp \leftarrow q$ . Any model for a program containing such an integrity constraint must map  $q$  to  $\perp$ . Is this adequate for human reasoning? If so, under which conditions shall such

integrity constraints be added within the reasoning step towards an appropriate logical form?

### Connectionist Realization

As shown in (Hölldobler & Kencana Ramli, 2009c), the computation of the least fixed point of the semantic operator  $\Phi_{\mathcal{P}}$  associated with a program  $\mathcal{P}$  can be realized within the core-method (Bader, Hitzler, Hölldobler, & Witzel, 2007). In this connectionist realization,  $\Phi_{\mathcal{P}}$  is computed by a feed-forward network, whose output units are recurrently connected to the input units. Whereas this network is trainable by backpropagation and, thus,  $\Phi_{\mathcal{P}}$  can be learned by experience, there is no evidence whatsoever that backpropagation is biological plausible. The approach can be extended to handle abduction following (Garcez, Gabbay, Ray, & Woods, 2007). However, in this setting, explanations are generated in a fixed, hard-wired sequence, which does not seem to be plausible either.

### Summary

We have presented an adequate computational logic approach for the suppression task. It is based on weakly completed logic programs under Łukasiewicz semantics. Such programs admit least models which can be computed by iterating an appropriate semantic operator. Reasoning is performed with respect to the least models. The approach is extended by sceptical reasoning within an abductive framework. Moreover, it can be realized in a connectionist setting. The approach has been carefully tested against alternatives like completed logic programs, Fitting semantics, and credulous reasoning, but none of these variations was found to be adequate.

### Acknowledgments

Many thanks to Bertram Fronhöfer, Caroline Dewi Puspa Kencana Ramli, Tobias Philipp, and Christoph Wernhard.

### References

- Apt, K., & Emden, M. van. (1982). Contributions to the theory of logic programming. *J. of the ACM*, 29, 841-862.
- Bader, S., Hitzler, P., Hölldobler, S., & Witzel, A. (2007). The core method: Connectionist model generation for first-order logic programs. In B. Hammer & P. Hitzler (Eds.), *Perspectives of neural-symbolic integration* (Vol. 77, p. 205-232). Berlin, Heidelberg: Springer.
- Banach, S. (1922). Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fund. Math.*, 3, 133-181.
- Byrne, R. (1989). Suppressing valid inferences with conditionals. *Cognition*, 31, 61-83.
- Evans, J.S.B.T., Newstead, S.E., & Byrne, R.M.J. (1993). *Human reasoning: The psychology of deduction*. Hillsdale, NJ England: Lawrence Erlbaum Associates.
- Fitting, M. (1985). A Kripke-Kleene semantics for logic programs. *Journal of Logic Programming*, 2(4), 295-312.
- Garcez, A. d'Avila, Gabbay, D., Ray, O., & Woods, J. (2007). Abductive reasoning in neural-symbolic learning systems. *TOPOI*, 26, 37-49.
- Herrmann, C. S., & Ohl, F. W. (2009). Cognitive adequacy in brain-like intelligence. In B. Sendhoff, E. Körner, O. Sporns, H. Ritter, & K. Doya (Eds.), *Creating Brain-Like Intelligence*. Springer.
- Hölldobler, S., & Kencana Ramli, C. D. P. (2009a). Contraction properties of a semantic operator for human reasoning. In L. Li & K. K. Yen (Eds.), *Proceedings of the fifth international conference on information* (p. 228-231). International Information Institute.
- Hölldobler, S., & Kencana Ramli, C. D. P. (2009b). Logic programs under three-valued Łukasiewicz's semantics. In P. M. Hill & D. S. Warren (Eds.), *Logic programming* (Vol. 5649, p. 464-478). Springer Berlin Heidelberg.
- Hölldobler, S., & Kencana Ramli, C. D. P. (2009c). Logics and networks for human reasoning. In C. Alippi, M. M. Polycarpou, C. G. Panayiotou, & G. Ellinasetal (Eds.), *Artificial neural networks – ICANN* (Vol. 5769, p. 85-94). Springer Berlin Heidelberg.
- Hölldobler, S., Philipp, T., & Wernhard, C. (2011). An abductive model for human reasoning. In *Proceedingth tenth international symposium on logical formalizations of commonsense reasoning*. ([commonsensereasoning.org/2011/proceedings.html](http://commonsensereasoning.org/2011/proceedings.html))
- Johnson-Laird, P., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, 111(3), 640-661.
- Kakas, A. C., Kowalski, R. A., & Toni, F. (1993). Abductive Logic Programming. *Journal of Logic and Computation*, 2(6), 719-770.
- Kleene, S. (1952). *Introduction to metamathematics*. North-Holland.
- Knauff, M. (1999). The cognitive adequacy of Allen's interval calculus for qualitative spatial representation and reasoning. *Spatial Cognition and Computation*, 1, 261-290.
- Łukasiewicz, J. (1920). O logice trójwartościowej. *Ruch Filozoficzny*, 5, 169-171. (English translation: On Three-Valued Logic. In: *Jan Łukasiewicz Selected Works*. (L. Borkowski, ed.), North Holland, 87-88, 1990.)
- Ormerod, T., Manktelow, K., & Jones, G. (1993). Reasoning with three types of conditional: Biases and mental models. *Quarterly Journal of Experimental Psychology*.
- Stenning, K., & Lambalgen, M. van. (2005). Semantic interpretation as computation in nonmonotonic logic: The real meaning of the suppression task. *Cognitive Science*, 29(6), 919-960.
- Stenning, K., & Lambalgen, M. van. (2008). *Human reasoning and cognitive science*. MIT Press.
- Strube, G. (1996). *Wörterbuch der Kognitionswissenschaft*. Klett-Cotta.
- Wason, P. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, 20, 273-281.

# Confidence in Causal Inferences: The Case of Devaluation

Uwe Drewitz (uwe.drewitz@tu-berlin.de)

Stefan Brandenburg (stefan.brandenburg@tu-berlin.de)

Berlin Institute of Technology, Department of Cognitive Psychology and Cognitive Ergonomics

Franklinstraße 28/29, 10587 Berlin, Germany

## Abstract

When people have to make predictions and diagnosis they make use of their causal knowledge. This knowledge refers to two constituting aspects of causality: sufficiency and necessity. In standard theories both aspects are considered as being independent from each other. The present research tests this assumption. In an experiment we examined how peoples confidence in one of both aspects is affected, if they receive negative evidence for the complementary aspect. The presented data show that peoples confidence related to the aspect that has not been challenged by negative evidence decreases under such conditions. This devaluation effect is not predicted by standard theories.

**Keywords:** causal models; causal learning; reasoning under uncertainty; induction

## Introduction

When people make a causal statement like: *A causes B*, they attribute a causal relation. This attribution can be based on various cues to causality (Einhorn and Hogarth, 1986) like spatial and temporal contiguity. However, in many situations people need more information than these. Being repeatedly confronted with a phenomenon, people (can) look for regularities as well. Psychological theories claim, that under such circumstances causal attributions rely on contingency information. Contingency information describe how the occurrence or absence of one event (i.e. event C) goes together with the occurrence or absence of another event (i.e. event E). Based on this information people can determine how likely an effect of interest will occur, given the presence or absence of a putative cause. According to standard psychological theories (e.g. Waldmann & Holyoak, 1992; Waldmann & Hagmayer, 2001; Griffith & Tenenbaum, 2005) people integrate the information about the (co-)occurrence and (co-)absence to either infer a causal relation or estimate it's strength, respectively. Therefore standard psychological theories claim that people base their judgments on all available data for contingency information. This is a reasonable assumption for situations where people do causal judgments. In contrast, in many real-world tasks people do not have to do such integrative judgments. They apply their knowledge to forecast events (i.e.  $E+$  /  $E-$ ) based on given data (i.e.  $C+$  /  $C-$ ). In probability calculus this is captured by conditional probabilities. The prediction of  $E+$  for example can be made based on  $P(E+/C+)$  or  $P(E+/C-)$  (see Fig. 1) depending on whether  $C+$  or  $C-$  is present. These conditional probabilities are independent of each other. Given these facts, standard theories do not predict effects of information integration over all contingency data.

Hence, for a prediction of  $E$  given  $C$ , persons would not integrate over all the four possible pairings of the two events (i.e.  $C$  &  $E$ ). However, we present experimental data that contradict this position.

## Sufficiency and Necessity

Various so called rule-based models (see Allan, 1980) have been proposed in research literature (e.g. Jenkins & Ward, 1965; Cheng & Novick, 1992; Cheng, 1997; White, 2003). They assume that persons rely on frequencies of (co-) occurrence and (co-) absence of two events (i.e.  $C$  &  $E$ ). The four cells in the contingency table in Figure 1 represent their four possible pairings. With respect to these two events, every observation can be assigned to one pairing and as such, to one cell of the contingency table. Every observation gives either positive or negative evidence to one of both aspects of causality: *sufficiency* and *necessity*. Positive evidence can be understood as strengthening an aspect (either sufficiency or necessity). Comparably, negative evidence weakens an aspect. Sufficiency and necessity are complementary building blocks of causality (e.g. Mill, 1869). An event  $C$  is recognized as *sufficient* to produce another event  $E$ , if the latter always follows the occurrence of the former. The same event  $C$  is considered as *necessary* to bring forth the event  $E$ , if its absence of  $C$  is always accompanied by the absence of  $E$ .

		$E$		
		+	-	
$C$	+	$a$	$b$	<i>sufficiency</i> $P(E+/C+) = a / (a+b)$
	-	$c$	$d$	<i>necessity</i> $P(E+/C-) = c / (c+d)$

Figure 1. 2x2 contingency table (+ indicates presence, - indicates absence).

Moreover, sufficiency and necessity are statistically independent of each other. Whereas the *sufficiency* of a putative cause for an effect depends on the frequencies in the cells  $a$  and  $b$ , the *necessity* is determined by the frequencies in the cells  $c$  and  $d$  (see Fig.1). Two different, statistically independent conditional probabilities capture these facts (see Fig.1): the probability of the presence of  $E$  given the presence of  $C$ ,  $P(E+/C+)$ , and the probability of

the presence of E given the absence of C,  $P(E+/C-)$ . These probabilities are complementary in the sense of causality. People are willing to attribute a causal relation between two events if both aspects are met. This idea goes back to John Stuart Mill (1869) who claimed that causal knowledge does not arise from the repeated observation of the sequence of two events only. Instead people also acknowledge what happens if a putative cause *fails* to appear. From this perspective, causes can be characterized in terms of *sufficiency* and *necessity* and both of these aspects have to be satisfied. Every observation that belongs to the pairing of cell a gives positive evidence to the *sufficiency* of the putative cause C for E, the effect of interest. Just as all observations that belong to the pairing of cell d give evidence to the *necessity* of C for E.

How do the described facts fall into the scope of standard theories of causal learning (see introduction)? These theories describe how people come up with a judgment, when they are requested to rate the strength of a causal relation in a causal attribution task. In such tasks people base their judgments on both aspects (sufficiency and necessity), which means that they consider all four frequencies that can be presented in contingency information (see Figure 1). Of course, people do integrative judgments in real-world tasks. But very often they have to make predictions based on given data. In turn, as soon as people can rely for example on the presence or absence of C, i.e. on C+ or C-, their prediction is related to only one of both aspects. For example, given the presence or absence of C (C+ or C-), people act differently as if they were asked to rate the strength of the relation between C and E. Let us assume people have seen numerous pairings where one event C precedes another event E (frequency in cell a). Based on these observations people will predict E+ (presence of E) given C+ (presence of C). In such a case, there is no need to integrate the information about C- (absence of C), which is captured by the frequencies in cells c and d. On the other hand, if C- (absence of C) precedes E- (absence of E), which is represented by the frequency in cell d, people might use this information to predict that E will not occur given the absence of C. In that case information with respect to C+ (cells a and b) can be ignored. Consequently, given the independence of both aspects, neither positive nor negative evidence related to one of the aspects should affect inferences related to the complementary aspect. In contrast, we claim that such an effect exists. We tested this hypothesis based on the representation of causal knowledge, which is introduced in the next section.

### Mental Causal Models

Several ways have been proposed to represent causal knowledge. For example Thüring and Jungermann (1992) suggest that people acquire mental models of causation in terms of conditional rules (e.g. If C+ then E+.). The conditional rules of a model reflect the characteristics of sufficiency and necessity of a causal relationship. This is in

line with the conception of causes as sufficient and necessary conditions for their effects. As shown by Thüring, Drewitz and Urbas (2006) these conditional rules can be obtained by mere induction. In the case of the model of unique causation (see Table 1), which states that "C causes E", the event C is framed as a sufficient as well as a necessary condition for the event E. This is captured by the rules R1 and R2 in Table 1. When a situation calls for a causal inference, the available data (for instance C+ or C-) are matched with the rules (R1 and R2) and the required information is deduced.

Table 1: Model of unique causation.

Model statement: "C causes E"

R1:  $C+ \rightarrow E+$

R2:  $C- \rightarrow E-$

The importance of rules like R1 and R2 lies in the savings they provide. Rules save costs such as time, attention or memory capacity. However, to get all the benefit rules entail, they have to be linked into higher-order knowledge, like models. Let us have a look on both our rules R1 and R2. Neither R1 nor R2 tell us whether there is a causal relation between C and E, or not. Only when they are linked together one possesses this knowledge. We call the linking of rules the construction of a mental model. In this sense the statement "C causes E" is knowledge acquired by building the model, not by having the two rules R1 and R2 only. That also means that as soon as the rules are linked into a model, there is more than there was before. Or, in other words: The whole is more than the sum of its parts. Assuming that our considerations are right, we can ask the following question: If the whole - the mental model - is questioned because one of its parts fails, is there an effect on the other parts as well? In terms of the model of unique causation (see. Tab. 1): When people observe that one rule fails to predict the outcome, is there an effect on how they use the complementary other rule?

### Causal Inferences under Uncertainty

Before we have a closer look on this question we want to make clear what known effects the failure of rule has. Depending on how successful the application of a rule was in the past, people will place more or less confidence into their predictions deduced from that rule. Let's think, for example, of a person that has rather limited causal knowledge as expressed by the model of unique causation. Whenever this person faces a situation where C+ is present, she will apply R1 and predict E+. Vice versa she will predict E- if C- is present, based on R2. As long as E+ goes always together with C+ all her predictions deduced from R1 are confirmed. Hence, her faith in R1 and therefore the confidence she places in her predictions should be high. The same holds for R2 and the related predictions as long as E- goes always together with C-. However, that will change as soon as it turns out that a rule is wrong. If people have build

up incomplete or incorrect rules they will come up with wrong inferences in the course of events. Let's assume that in truth  $C+$  in conjunction with another event  $X+$  may be sufficient for  $E+$  instead of  $C+$  alone. In such a case people will observe  $E-$  subsequent to  $C+$  whenever  $X$  is absent ( $X-$ ). An observation like this will impair the *sufficiency* of  $C$  for  $E$ . Moreover, such an observation will *discredit* the rule  $R1$ . In general, every prediction that is not confirmed but contradicted by a subsequent observation *discredits* the rule it is derived from. Consequently, as long as a person cannot expand her model, she will loose confidence in the respective rule. All a person can do in such an uncertain situation is to reduce the confidence she places in her predictions based on that rule. Hence, the question we raised at the end of the last section remains unanswered. But now, we can reformulate and ask more specifically: Is the effect of the reduction of confidence always limited to the rule that was discredited?

### Discrediting and Devaluating Causal Rules

To start with let us return to contingency information. Figure 1 shows a contingency table for the model of unique causation. A person's observations that fall into cell a provide evidence for the reliability of rule  $R1$ , while observations that fall into cell 'd' provide evidence for the reliability of  $R2$  (see Table 1). On the other hand, all observations made in cell b discredit  $R1$ , while all observations made in cell c discredit  $R2$ . Therefore, the first row of the table provides information about the sufficiency of the cause and the second row about its necessity. As depicted in Figure 1 a cause is only *completely* sufficient, if observations are made in cell a, but not in cell b, and it is only *completely* necessary, if observations are made in cell d, but not in cell c. Only in these cases, the conditional probabilities are at their optimum with respect to a causal relation between  $C$  and  $E$ . From this point of view, the optimum of  $P(E+/C+)$  equals one and equals zero for  $P(E+/C-)$ . Additionally, an increase or decrease of  $P(E+/C+)$  does not affect  $P(E+/C-)$  and vice versa. What does this mean from a psychological perspective? The first implication is consistent with the mechanism of *discrediting* a rule. When the sufficiency or necessity is weakened, the certainty of inferences based on the respective rule should decrease. For instance, if an observation of  $C+$  together with  $E-$  (see cell b in Fig.1) is made, the aspect of sufficiency of  $C$  for  $E$  is weakened. Subsequently, inferences based on  $R1$  go along with a reduced certainty. The second implication touches the central issue of this paper. It illustrates our assumption of devaluation. We assume that negative evidence for one aspect of causality will be reflected by increased uncertainty about the complementary aspect of causality. For instance, if  $R1$  is discredited by negative evidence (observations that fall into cell b), confidence in  $R2$  decreases as well. We call this the effect of devaluation. Table 2 shows which observation discredits and devaluates the rules of the model of unique causation.

So far, we have described the consequences of positive evidence that strengthens a rule and the consequences of negative evidence that weakens a rule in terms of discrediting and devaluation. This leads to three hypotheses:

1. **Strengthening:** Observations that fall into cell a and d of the contingency tables provide positive evidence for the respective rules of the models and should increase the confidence in inferences drawn from these rules.
2. **Discrediting:** Observations that fall into cell b and cell c provide negative evidence and *discredit* the rules as shown in table 2. In all these cases, the confidence in inferences drawn from the rules should get reduced.
3. **Devaluation:** Observations that fall into cell b and cell c should *devalue* the complementary rules as shown in table 2. Again, the confidence in inferences from the affected rules should decrease.

The following experiment serves to test these hypotheses.

Table 2: Discrediting and devaluating causal rules.

Rule	Observation	Discrediting of	Devaluation of
$R1: C+ \rightarrow E+$	$C+, E-$	$R1$	$R2$
$R2: C- \rightarrow E-$	$C-, E+$	$R2$	$R1$

### Experiment

In our study, participants had to acquire causal knowledge about a simulated technical system based on inductive learning. Over the course of the experiment, positive as well as negative evidence was presented to investigate the consequences of discrediting and devaluation.

#### Method

**Participants.** Sixty graduate and undergraduate students at the Berlin Institute of Technology were recruited for the experiment. All of them were paid for their participation.

**Material.** Figure 2 shows the schematic screen layout of the simulated system that was presented to the participants. It was introduced as an electrical system of a power plant. The system was built up from four subsystems that were responsible for two output systems. Information about the state of these subsystems was displayed on four dials (for top boxes in Fig.2). Each dial represented the state of one subsystem, which was either DOWN ( $C+$ ) or UP ( $C-$ ), or unknown because its dial was switched off. Only one subsystem was causally relevant and served as cause  $C$  for the outcome of the relevant output system (either  $E+$  or  $E-$ ). The other three subsystems were irrelevant for the task. One of them was unused (the dial was switched off) while the other two were used as distractors to give the system a more diversified appearance. In the lower half of the screen, the displays for the output systems were shown. In some of the trials participants had to predict the outcome of only one of them and in the remaining trials they had to predict the outcome of both. If only the outcome of one system had to be predicted, the display of the other output system was not shown. Whereas one output system ( $E$ ) was relevant for the



experiment the other was used to make the task more realistic.

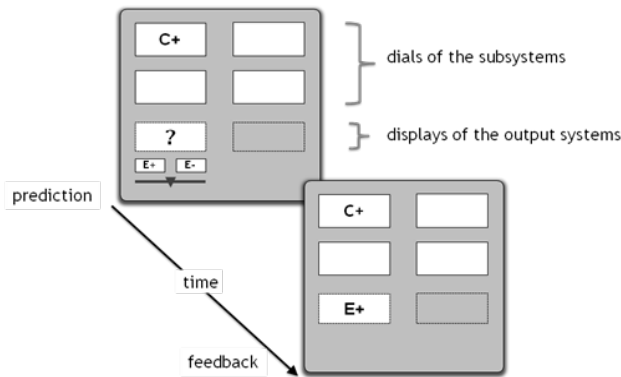


Figure 2. Screen layout (schematic) and sequence of one trial of the experiment.

Below the display of each output system two buttons were shown for the prediction of the outcome. One button served the prediction of MALFUNCTION (E+) and the other one the prediction of working operation (OK) (E-). Clicking on one of them was necessary to make the prediction. Finally, below these buttons a slider was presented that could be adjusted to rate the confidence of the judgments. The lowest confidence (0%) was set in the middle of the slider. Subjects were instructed to place the slider on the very right to indicate full confidence (100%) that E- will occur, and on the very left to mark full confidence (100%) for E+.

**Procedure.** The participants' task was to predict the outcomes (E+ or E-) of the output system(s). To solve this task, they had to understand the underlying causal relation between the subsystems and the output systems. In each trial, they were shown the layout of the device as presented in Figure 2. First, subjects had to check the operation of the subsystems. Then, based on the information, which was shown on the dials, they were requested to predict the state of the output system(s) by clicking on the respective buttons (OK or MALFUNCTION). Finally, they rated their confidence for each prediction by adjusting the respective slider(s). After participants finished their prediction and confidence rating, they had to click on a 'send' button and subsequently received feedback that showed the actual outcome(s). The experiment consisted of thirty-three trials. These trials were split up in a *reinforcement phase* and a *test phase*. Figure 3 depicts the experimental procedure schematically. Note that the frequencies in the cells of the contingency tables in Figure 3 (b) are summed up for both phases. In the reinforcement phase, which consisted of twenty-six trials, participants received information that enabled them to acquire a model of unique causation with two rules (R1 & R2, see Table 2). This was accomplished by providing positive evidence for R1 (eight trials, see Fig.3) and R2 (eight trials, see Fig.3). Additionally, there were two distractor trials in which information about an irrelevant subsystem was shown only. In the remaining eight trials participants had to predict only the outcome of the second output system that was irrelevant for the test of

the hypothesis. After the twenty-six trials of the reinforcement phase the test phase started that consisted of seven trials. In four of these seven trials, negative evidence for one of the two rules (R1 or R2) was presented. The negative evidence always opposed the rule reinforced in the last trial of the reinforcement phase. In these trials people had to predict the outcome of the relevant output system (E) only. Another two trials were used as distractor trials presenting information about one of the irrelevant subsystems. In the seventh and last trial of the test phase the post-measure for the relevant test was recorded. Therefore data were presented that matched the same rule as in the last trial of the reinforcement phase (see Fig.3).

**Independent and dependent variables.** Since the model of unique causation consisted of two rules, both were used to investigate the issues of discrediting and devaluation. For this purpose, the sample of sixty participants was split into two groups of thirty participants each. One group received negative evidence about R1, the other half about R2.

(a) Reinforcement phase.

	E+	E-	
C+	8	0	measure of confidence for prediction based on:  C-
C-	0	7	

positive evidence      pre-measure      time

(b) Test of Devaluation Effect phase.

	E+	E-	
C+	8	4	measure of confidence for prediction based on:  C-
C-	0	8	

negative evidence      post-measure      time

Figure 3. Experimental procedure (schematic). Reinforcement phase and test phase were presented in sequence.

Values are exemplary for one of the two experimental groups. Contingency table in (a) displays frequencies for reinforcement phase. The trial of the pre-measure also was the eighth presentation of (C-, E-), which is updated in (b). The contingency table in (b) displays summed frequencies for reinforcement phase and test phase.

To investigate the strengthening of rules, the amount of *positive evidence* ranged from one to eight trials (see Fig. 3a, positive evidence) for each rule (R1 & R2). To test the impact of discrediting, the amount of *negative evidence* ranged from one trial to four trials (see Fig.3b, negative evidence) for each rule (R1 & R2). The factor measurement with the factor levels *pre* and *post* served the investigation of devaluation as described in the procedure (see Fig.3). Throughout the experiment, confidence ratings of inferences

predicting the states of the relevant output system were used as dependent variable.

## Results

For statistical analysis, we computed three ANOVAs with repeated measures, one for each effect. Additional to the significance of effects we report effect sizes after Cohen (1988). Cohen (1988) defines small effects from  $0.10 < f < 0.25$ , medium effects from  $0.25 < f < 0.40$  and large effects from  $f > 0.40$ . The effect of *strengthening* was analyzed with a one-factorial ANOVA with repeated measures for each rule. We used the number of occurrences of positive evidence (1-8) as independent variable. Strengthening greatly affected subjects confidence ratings for R1 ( $F(7,413)=57.12$ ,  $p<0.01$ ,  $f=0.98$ ) as well as R2,  $F(7,413)=46.83$ ,  $p<0.01$ ,  $f=0.89$ . Figure 4 shows the effects of strengthening on subjective confidence for both rules. As depicted, subjects' confidence in their prediction of the state of the output system strongly increases over time.

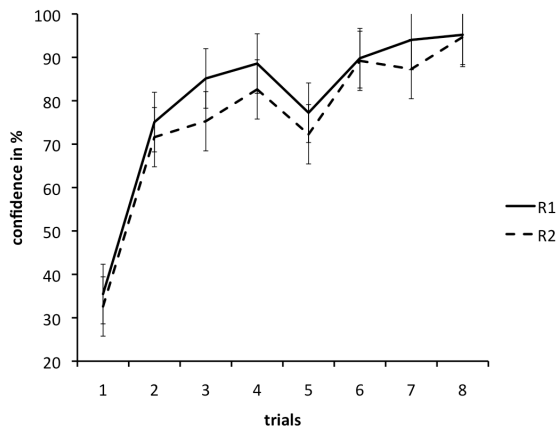


Figure 4. Effect of positive evidence on confidence ratings, depending on the number of trials. Error bars represent standard error.

For *discrediting* rules 1 and 2 (R1 & R2), the four trials with *negative evidence* were run to weaken subjects' confidence in their predictions. Since rule one was discredited for half the subjects and rule two was discredited for the other half, rule became a factor in the analysis. Therefore a 2x2 ANOVA with repeated measurement was calculated in which the *rules* of the model (R1 and R2) served as between subjects factor and *negative evidence* (trials 1-4) was a within subjects factor. We found a significant large main effect of *negative evidence* ( $F(3,174)=18.19$ ,  $p<0.01$ ,  $f=0.56$ ), but no effect of *rules* ( $F(1,58)=0.03$ ,  $p=0.95$ ,  $f=0.00$ ) nor an interaction effect ( $F(3,174)=1.96$ ,  $p=0.12$ ,  $f=0.18$ ). Figure 5 visualizes the results. To investigate the effect of *devaluating* a rule (Fig. 6), it seems necessary to highlight how we achieved the data for this computation. For all subjects rule 1 and rule 2 were strengthened. The last trial of the strengthening phase for each rule (trial 8) served as pre-measure. However, only for half of the subjects rule 1 was discredited. If these subjects' confidence for the

prediction of rule two (post-measure) was lower after discrediting rule one, devaluation took place. Reversely, for the other half of the sample rule 2 was discredited. Hence, if subjects' confidence for rule 1 (post-measurement) also decreases, devaluation worked as well.

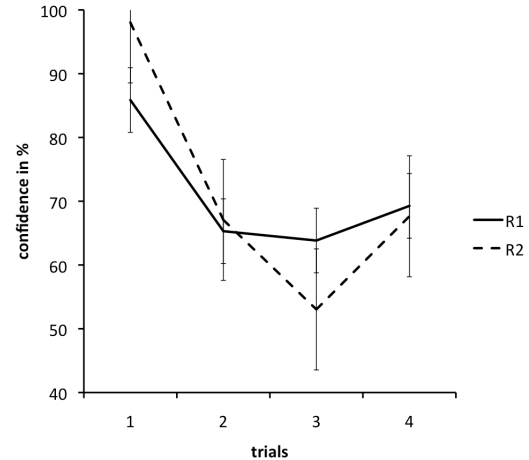


Figure 5: Effect of negative evidence on confidence ratings. Error bars represent standard error.

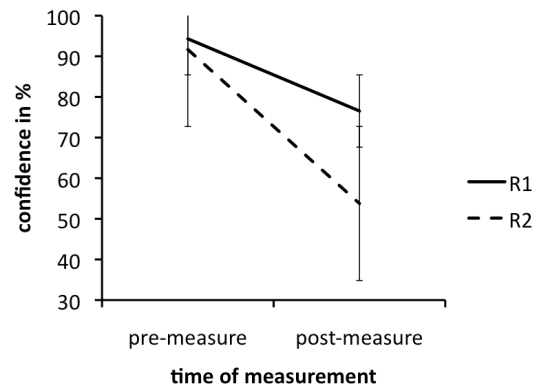


Figure 6: Effect of devaluation on confidence ratings for both rules. Error bars represent standard error.

A 2x2 ANOVA was calculated over the between subjects factor *rule* (either R1 or R2 was discredited) and the within subjects factor *measurement* (pre- and post-measure). This analysis revealed a medium main effect of *rule* ( $F(1,58)=4.47$ ,  $p=0.03$ ,  $f=0.28$ ) and a large main effect of *measurement* ( $F(1,58)=42.66$ ,  $p<0.01$ ,  $f=0.85$ ). Additionally, we observed a medium significant interaction,  $F(1,58)=5.58$ ,  $p=0.02$ ,  $f=0.31$ . Figure 6 visualizes these effects.

## Discussion

In the present paper we tested three. First, we assumed that positive evidence strengthens subjects' confidence for predictions they derived from a set of rules that was acquired in the course of an experiment. Empirical evidence supported that hypothesis. At the end of the strengthening

phase peoples' confidence was close to 100%. Second, we expected a decrease in participants' confidence in their causal inferences if negative evidence discredited the respective rules. This hypothesis was empirically confirmed as well. Finally, hypothesis three claimed, that negative evidence for one aspect of causality results in decreased confidence in the complementary aspect as well. Empirical findings clearly supported this assumption. This result opposes a normative view that would require people to base their predictions solely on the given facts. For example, given C- subjects should predict E- with high confidence. In contrast, despite their correct prediction of E- (in case of C-) participants confidence decreased with respect to the critical test in the post-measure. This effect emphasizes the idea that humans do not consider sufficiency and necessity as independent of each other. Instead, once people have acquired causal knowledge, they take evidence for both aspects into account. They do so, even if the predictions they make are solely based on one of them. Hence, we conclude that people mentally construct causal models that relate sufficiency and necessity. These models can be seen as a whole. If one part or aspect of such a model proves to be wrong, subjects lose their confidence for the complementary part as well. Existing models of causal learning and reasoning aim to explain integrative judgments. Hence people are required to integrate information over all four cells of the contingency table. Thus, they always have to consider both aspects of causality. Therefore these models do not fit to the conditions of the experimental task. Nevertheless, assuming that subjects frame the task in our experiment as to judge the strength of the relation of C and E, the Power PC model (Cheng, 1997) would predict a confidence level of 66% for the post-measure. This is within the range of our results. Hence, if subjects are asked to make predictions in a causal learning paradigm, they reframe the task to judge the strength of a causal relation of two events. According to Griffith and Tenenbaum (2005) parts of the experimental task can be described as causal structure learning. From this point of view presenting negative evidence for one aspect would favor a different causal structure (compound causation or alternative causation respectively). Hence, it might be that peoples' post-measure judgments reflect their preference for the new structure compared to the previous one. Alternatively the post-measure judgments might reflect participants' uncertainty regarding the new structure. Again, these alternative explanations require people to integrate over all contingency information. In contrast to these alternative explanations there are models of inductive causal learning that are based on cognitive architectures and that emphasize the role of declarative memory (Drewitz & Thüring, 2009; Drewitz & Brandenburg, 2012). These models account for peoples' judgments and their confidence ratings given positive as well as negative evidence. They provide a possible explanation for peoples' performance in inductive learning based on memory processes. Additionally they do not assume that people reframe the experimental task from

prediction to integrative judgments. To discriminate between these explanations, future research should focus on the replication of the devaluation effect for more complex causal models and different dependent variables like reaction times and pupil dilation. If we can replicate the effect we also might be able to differentiate between possible alternative explanations.

## References

- Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, 15, 147-149.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104, 367-405.
- Cheng, P.W. & Novick, L.R. (1992). Covariation in natural causal induction. *Psychological Review*, 99, 365-382.
- Cohen J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, New Jersey.
- Drewitz, U. & Brandenburg, S. (2012). Memory and Contextual Change in Causal Learning. In N. Rußwinkel, U. Drewitz & H. van Rijn (eds.), Proceedings of the 11th International Conference on Cognitive Modeling, Berlin: Universitätsverlag der TU Berlin.
- Drewitz, U. & Thüring, M. (2009). Modeling the Confidence of Predictions: A Time Based Approach. In A. Howes, D. Peebles, R. Cooper (Eds.), 9th International Conference on Cognitive Modeling, Manchester, UK.
- Einhorn, H. J. & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, 99, 3-19.
- Griffith, T. L., & Tenenbaum, J. B. (2005). Structure and Strength in causal induction. *Cognitive Psychology* 51, 334-384.
- Jenkins, H.M. & Ward, W.C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, 79, 1, 1-17.
- Mill, J. S. (1869). *A system of logic, ratiocinative and inductive*. London: Longman, Roberts & Green.
- Thüring, M. & Jungermann, H. (1992). Who will catch the Nagami Fever? Causal inferences and probability judgments in mental models of diseases. In D.A. Evans & V.L. Patel (Eds.), *Advanced models of cognition for medical training and practice* (307-325). Berlin: Springer.
- Thüring, M., Drewitz, U., & Urbas, L. (2006). Inductive Learning, Uncertainty and the Acquisition of Causal Models. In R. Sun & N. Miyake (Eds.), Proceedings of the 28th Annual Cognitive Science Society. NJ: LEA.
- Waldmann, M. R. & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition*, 82, 27-58.
- Waldmann, M. R. & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, 121, 222-236.
- White, P. A. (2003). Making causal judgments from contingency information: the pCI rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 710-727.

# Spatial Co-ordination in Music Tuition

**Sam Duffy & Patrick G. T. Healey**

Queen Mary University of London

Media and Arts Technology Programme

Interaction Media and Communication Group

School of Electronic Engineering and Computer Science

London, E1 4NS, United Kingdom

s.duffy@eecs.qmul.ac.uk, ph@eecs.qmul.ac.uk

## Abstract

Playing an instrument is a physical skill learned through observation, repetition and rehearsal. Students of orchestral instruments seek one-to-one tuition from expert musicians. However as they become more accomplished, the number of suitable tutors becomes more concentrated, especially for less common instruments. Often a tutor-student relationship develops over several years and temporary separation due to overseas performing, auditioning and teaching commitments is problematic. Some music education organisations use video conferencing as a solution to these problems, however it has long been recognised that interaction mediated by video conferencing is not analogous to a co-present experience. In this paper, ethnographic video analysis is used to study the interactions in co-present and separated instrumental music lessons. We find that the musical score represents more than a physical embodiment of the music - it plays an important role in coordinating activity and interaction. In video mediated lessons a single physical score can no longer be shared and interaction is changed as a result.

**Keywords:** video conference; interaction; ethnography.

## Introduction

A recognised method for learning to play an orchestral musical instrument is through regular one-to-one lessons with an experienced tutor. Playing an instrument is a physical practice requiring dextrous manipulation of a complex tool. Marchand (2010) proposes that the interpretation, understanding, and realisation of practice is based in motor cognition. As the student watches the tutor, visually processed signals are paired with observed actions, gestures, and postures. These may be co-ordinated with verbal instruction and commands from the tutor. However learning a musical instrument cannot be achieved purely by verbal description or observation, students learn a practice by 'doing'. Through observation followed by repetition and rehearsal, with iterative feedback from the tutor, the student develops motor and kinaesthetic cognition of how to play their instrument. This is a collaborative process, involving the co-ordination of understanding (Clark & Brennan, 1991).

At an undergraduate level of study, music students seek professionally recognised performers as tutors and their choice of where to study could be influenced by resident tutors and professors at an institute. The number of qualified professional tutors in any particular field is finite, but becomes more limited the more accomplished a student becomes, especially for less common instruments. Once a teaching relationship has been established, musicians tour

and travel frequently, so temporary separation of tutor and student can occur at critical times, such as prior to an important audition or performance. One solution to these problems is video conferencing. This is popular in geographically remote areas such as Australia (Lancaster, 2007) but is also part of urban mainstream conservatoires such as The Manhattan School of Music in New York. However interaction when video is the medium of communication is not analogous to the co-present experience. The belief at the inception of video conferencing that technology which replicated face-to-face interaction, simply at a distance, would enhance communication, contained a fundamental misunderstanding about how people interact when working collaboratively to achieve a task (Whittaker, 2003; Edigo, 1988; Hollan & Stornetta, 1992). Heath et al (1997) found that the visual focus of collaborative work is likely to be aligned to the focal point of the activity, such as a document or object, rather than face-to-face.

Existing research concerning separated musicians has focused on collaborative performance, the impact of latency and delays and tools to enable distributed ensembles to perform, improvise and compose (Chew et al., 2004; Sarkar & Vercoe, 2007; Hamilton, Iyer, Chafe, & Wang, 2008; Barbosa, 2003; Bryan-Kinns & Healey, 2006). However the activity taking place during music tuition is not the same as performance due to the educational frame of reference.

## The Use of Shared Space

Individuals in shared space coordinate their actions through spatial awareness, peripheral monitoring of non-verbal signals and the ability to joint reference; where gaze and gesture around a shared point coordinates the attention of participants (Whittaker, 2003). They use their position relative to each other to create mutually recognised shared space. Kendon (1990) describes how two or more people can organise themselves to create and sustain a shared space, called 'o-space', to maintain a common focus of attention. He goes on to describe sustained clusters and patterns as formations, an 'F formation' being where participants have equal, direct and exclusive access to their o-space. When two people are performing a collaborative task through the medium of video, they no longer have a concept of negotiated mutual distance (Sellen, 1992) and they cannot easily manage their position relative to each other or objects in their environment, there-

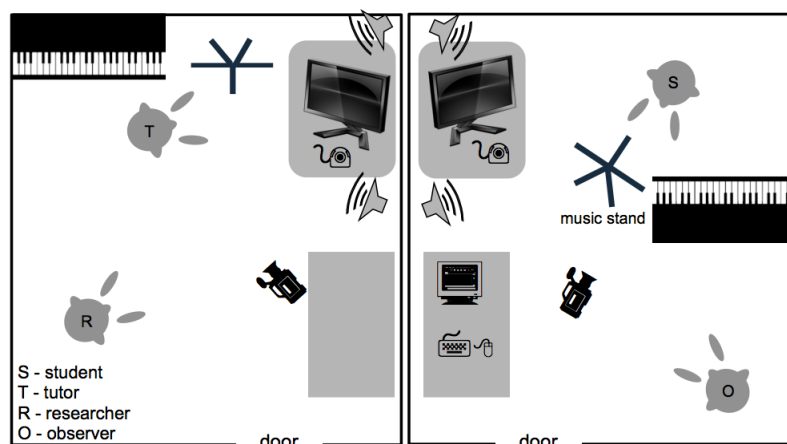


Figure 1: The layout for video lessons.

fore participants cannot use formations to create an o-space. Gestures and gaze are also shown to be less efficient in video mediated communication (Heath & Luff, 1991b). There is a body of work investigating the importance of gestures and non-verbal communication to teaching (Roth, 2001), to performing musicians (Vines, Wanderley, Krumhansl, Nuzzo, & Levitin, 2004; Wanderley, Vines, Middleton, McKay, & Hatch, 2005; Broughton & Stevens, 2009) and even specifically to instrumental music tuition (Kurkul, 2007; West & L. Rostvall, 2003). In this paper we analyse the interactions of co-present musicians in a learning environment and compare them to the interactions seen when student and tutor interact through the medium of video.

## Methodology

Ethnography requires a researcher to participate in people's daily lives, watching what happens, listening to what is said and gathering data to understand the issues emerging (Hammersley & Atkinson, 2007). It can offer fine-grained, detailed qualitative insight into how users interact with technology (Whittaker, 2003) and is often used as a tool to assess HCI and CSCW systems, for example the studies by Heath et al (1997; 1991a; 2005). It is the only way to study embodied social practice as it naturally occurs, rather than in conditions created by the researcher.

**Co-present Studies** We observed co-present music lessons taking place at the educational establishments where the students would normally have their weekly lessons. Three thirty-minute lessons were observed and filmed, two clarinet lessons and a trumpet lesson. The students observed were preparing for Grade 7 or 8 exams<sup>1</sup>. A researcher was present since it was not possible to know in advance how much the participants would move around the room, necessitating repositioning of the camera (see researcher position R1 and R2 in Fig-

ure 2), however the researcher took no part in the lessons. The footage was analysed using ELAN and a detailed transcript produced for each class.

**Video Mediated Studies** The video conference data was obtained from a study run by British Telecom Research and Development<sup>2</sup> to evaluate a video conference prototype, designed specifically to support instrumental music tuition. We were invited to observe tests which involved students and visiting tutors at Aldeburgh Music in Suffolk. Six one-hour lessons using the prototype were observed and filmed over three days, including harp, cello, violin, oboe and french horn. The tutors had a photocopy of the student's music, or their own editions of the score to be worked on. Some of the tutors had previous experience of teaching via video conference and some of the student-tutor pairings had worked together previously. A researcher observed from the tutor room, there already being an observer from the prototype team in the student room. Video footage was obtained from both rooms (see camera positions in Figure 1) and analysed synchronously.

## Results and Discussion

Professional musicians interviewed as part of this work believed latency to be the biggest barrier to teaching via video conference, as the delay makes it very difficult to play together (Chew et al., 2004). However analysis of the three co-present lessons showed that synchronous activity (singing or playing together, accompanied playing, the tutor conducting) made up only 11 percent of the lesson on average. This activity was used largely to resolve specific rhythmic problems. Where video conference is used to manage temporary separation of a student-tutor pair, many normal lesson activities can still take place, synchronous tools being saved for the next co-present lesson. The impact of the medium on interaction seemed to be a more significant problem as this affected

<sup>1</sup>Grade 8 from a recognised exam board such as the Associated Board of the Royal Schools of Music is often an entry requirement for music performance undergraduate degrees

<sup>2</sup>As part of the EU FP7 project Together Anywhere Together Anytime ("TA2") <http://www.ta2-project.eu/Pages/overview.html>

all lesson activities.

### Co-present Lesson Interactions

The rooms where lessons took place were small, the space constrained by the piano and the music stand. Nonetheless, in each case, tutors used their position relative to the student and the music stand to communicate their intention to act. This led to the establishment of specific zones within the space, which both participants could be seen to observe. To illustrate this we present vignettes illustrating examples of the use of these zones.

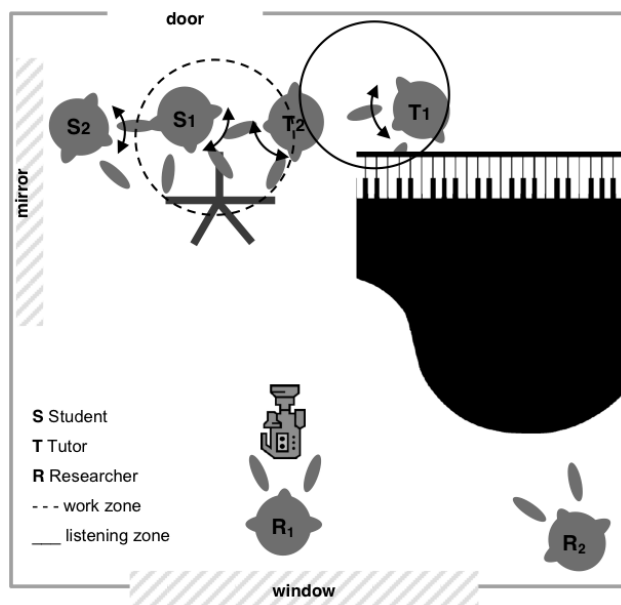


Figure 2: The work zone and listening zone.

**The Work Zone** The music stand became the focal point of a side-by-side F-formation (Kendon, 1990) as participants shared the student's score, the o-space created between them being designated as a 'work zone' (Figure 2). In one example, towards the end of the lesson the student and tutor briefly relax into social conversation. The student moves back slightly and moves her upper body to face the tutor, rather than the music (position S2 Figure 2). The tutor moves her upper body to face the student, the stand no longer the focus of their o-space. The student rubs her shoulder and moves about, relaxing her muscles (Figure 3). The tutor then puts her left hand on the music stand, between two short utterances, whilst still maintaining eye contact with the student. In this way the tutor holds both the stand and the student, triangulating her position (Healey & Battersby, 2009) as she begins the transition back to the work zone. Finally, the tutor turns her head back towards the music stand, pulling her body round, facing back into the work zone (Figure 4). The student also turns back to the stand (position S1 Figure 2), and swings her clarinet up towards her face, having understood the signal that they are going to go back to work.



Figure 3: Moving out of the work zone.



Figure 4: The tutor signals 'back to work'.

**The Listening Zone** In each case, the tutor defined a listening zone (for example T1 in Figure 2). When the tutor occupied this zone, the student understood that they could play without immediate interruption as the tutor wanted to hear a longer section. The tutor also defined a listening position within the listening zone, for example one tutor stood with feet slightly apart, hands loosely folded in front of her body, shoulders relaxed; attending to, but not bidding for, the floor.

**Transitions Between the Zones** When a tutor wanted to give detailed feedback on a passage, they moved forward into the work zone (position T2 Figure 2). When they stepped backwards into the listening zone again, the student understood that the specific topic of detailed work was finished and they should play a longer passage again, for the tutor to listen to and assess.

The transition from listening position to interruption could be sudden or more gradual. In some cases, the first indication that a tutor had diagnosed a problem from their listening position was when they lifted their gaze from the score to look at the student's face or instrument. Sometimes the student had already looked up, aware of their error, to see if it had been detected. In other cases the student demonstrated peripheral awareness; looking up in response to the tutor's movement, returning their gaze, and sometimes even stopping playing. The intent to make a more significant interruption could be indicated in advance by the tutor leaning into the piano to pick up a pencil whilst still in the listening zone, signalling intention to write on the score; or picking up their own instrument, indicating that they wished to demonstrate something the student had played incorrectly.

The duration of the tutor's planned interruption was indicated through the extent of her movement into the work zone.

When the tutor intended only a brief interruption she leaned forward into the work zone, without moving her feet, and pointed to the score whilst giving feedback. Then leaned back into her 'home position' having used body torsion to indicate a temporary movement into the work zone, the lower body remaining in the listening zone or 'base position' (Schegloff, 1998).

### **Control of Musical Turns**

When a student is playing, they are emotionally engaged with their performance and concentrating on the score. However the tutor frequently interrupts to provide immediate feedback on an identified problem. Frequent, unexpected interruptions could become frustrating, however student-tutor pairs managed interruptions in order to reduce the impact. In a music lesson, where a verbal utterance is often followed by a musical response, it is reasonable to assume that a musical phrase is analogous to a conversational turn and we should therefore be able to see the characteristics of turn management (Sacks, Schegloff, & Jefferson, 1974) such as transition-relevant places for a new turn, back-channelling (Moran, 2011), turn breakdown and repair.

We could see evidence of musical turn management whilst the student was playing. Musicians have been shown to have awareness of anticipation of other musician's intentions with respect to musical structure (Moran, 2011) and the position between two adjacent musical phrases or ideas was observed to be preferred by the tutor as a transition-relevant place to interrupt. Even if a change in the tutor's posture indicated that they had diagnosed a problem earlier in a phrase, they would often wait for the long note at the end of the phrase before initiating interruption of the student's performance. This provided an ideal opportunity for the tutor to speak, whilst the student had a natural point to finish, before moving on to the next musical idea.

The transition from verbal instruction, to visual observation, to motor cognition (Marchand, 2010) was observed. After verbally explaining a point, in some cases the clarinet tutor then demonstrated it for the student on her own instrument. The student was asked to imitate. If the tutor was not satisfied with the performance she played the phrase again, exaggerating the aspect not yet corrected. The alternating musical turns increased in intensity, the student's technique converging over time with that of the tutor's. In one example a student impatiently copied the tutor's demonstration, starting to play before she had finished demonstrating. The tutor did not detect any noticeable improvement and musically admonished him, interrupting his performance and playing it again herself, taking back the turn.

### **The Shared Score**

The score played a much greater role in the lesson than a physical embodiment of the music, also providing a shared reference to coordinate joint attention (Whittaker, 2003). In co-present lessons the participants shared a score, usually that which belonged to the student, and both pointed to parts of the

score as they spoke as a convenient way to reference without having to mention bar numbers specifically. Tutors gestured over the score as they talked, linking their comments to specific notes and putting phrases into context of the whole piece. For example, in one lesson the tutor pointed to the music 32 times, half of the instances for navigation purposes such as "from here" and half to reference feedback against musical notation.

Direct eye contact was shared for less than 5 percent of the time, and was made up of brief glances (for example one lesson contained 112 instances of shared gaze, the average duration being less than one second). More often both looked at the score, even when in conversation together, whilst exhibiting a high level of peripheral awareness. For example, when one student performed for the tutor and stumbled over a note, the tutor immediately moved in towards the score with her pencil, starting to speak. However the student interrupted before her pencil reached the score saying "change my right?", his gaze not having moved from the music. The tutor stopped moving, looked up at him nodding and said "you read my mind, yeah" then moved back to her listening position.

Whilst the students had all made pencil annotations on their music outside of the lesson, such as marking fingerings, phrase marks, accents and breathing; during the lesson it was the tutor who annotated the score. For example the clarinet tutor used the character 'O', writing it on specific notes to indicate the 'open throat' required to control tone in some registers or 'X' to signify a particular spacing of fingers on the keys. Through annotation, the student's score built into a permanent and cumulative record of the learning imparted by the tutor; a record of how they had developed.

### **Video Mediated Lesson Interactions**

In the video mediated lessons, students demonstrated awareness of their need to monitor both their music and the tutor at the same time, by initially positioning themselves directly in front of their screen so that they could see both the tutor, and their music, without significantly moving their head. However the score and music stand obstructed the main camera view (Figure 5) and in all cases tutors asked the students to turn around so that they could see their hands on the instrument. The students were then turned between 45 and 90 degrees away from their screen (position S in Figure 1) and no longer had peripheral awareness of the tutor when performing.

### **The Divided Score**

Now that the student's gaze was divided between their score and the screen, there was a dramatic impact on turn control. With a separate music stand and score in each room, a shared work zone could no longer be established, and the tutor could not create a listening zone, removing communication through spatiality. Whilst some tutors still adopted a listening position (for example one tutor formally placed her folded hands in her lap - see Figure 6) they now used exaggerated gestures such as a raised arm or a wave to indicate when they





Figure 5: The score obstructs the camera view.

wanted the student to stop playing, and when these were also unseen, resorted to a vocal request. This was sometimes not heard by the student who was absorbed in their playing and not facing the video conference system speakers, and the tutor had to raise their voice. Even when they were able to rapidly switch their attention, students missed cues through looking in the wrong place at the wrong time. From their perspective, they were continually being stopped unexpectedly by a raised voice, requiring a significant twist of their upper body to see the screen (Figure 7) and this quickly led to frustration.



Figure 6: Tutor's gaze divided between screen and score.

Tutors also struggled with dividing their gaze between the screen and the score (Figure 6). Often they would discuss a phrase looking down at their score, as the student looked at their own separate score. Neither party could monitor their video screen at the same time. Previously the tutor could use peripheral awareness of the student's gaze as evidence of continued attention and the student could use indicative gestures to confirm their understanding of the feedback and its relation to the score (Clark & Brennan, 1991). Navigation became problematic as a result, requiring detailed reference to page and bar numbers to establish precisely where in the score feedback related to, or for the tutor to establish where they would like the student to play from. For example, as shown in the following extract of dialogue.

*Tutor:* that wasn't quite right, let's try it again...  
 [the student starts to play]  
*Tutor:* ...from the beginning of the bar.  
 [student stops, looks up]

*Student:* from the, sorry? From the?  
*Tutor:* from the beginning of the bar.  
 [the tutor starts playing the phrase that she wants to hear, the student is looking hesitant]  
*Tutor:* can you play it from the beginning of the bar and stop on the B and the E?  
*Student:* OK [with instrument raised to playing position]  
*Tutor:* Do you see where I mean?  
 [the student looks at the music intensely, wiggling her fingers on the fret board]  
*Student:* "uh hum" [hesitantly]

The tutor could no longer directly annotate the student's score and it was noted that, in comparison to the co-present lessons, notes and annotations were not frequently made by either participant. One tutor made reference to a student's annotations where they were available on their photocopy of the student's music, confirming their value.



Figure 7: The student must switch gaze from score to screen.

## Conclusions and Further Work

Ethnographic analysis of co-present lessons provided a useful framework to assess the effectiveness of video mediated communication to teach a practice based skill. The importance of the shared score to lesson interaction was evidenced by problems managing interaction such as turn control when participants were separated and could no longer share the same physical representation of the music.

A further study is planned where the instrument class will be confined to woodwind (for example clarinet or oboe) and an additional camera will be placed behind the participants to capture gesturing over the score in more detail. The score will be photographed and the annotations discussed with participants during post-observation interviews. Technological solutions to the problem of interaction lost through the divided score will be suggested. These are likely to involve an interactive visual layer over a digitised representation of the physical score, which shows the separated participants where each person is gesturing on the music. Ideally both participants should be able to mark their layer in a way which allows the student to take an annotated copy away, and return with it for the next lesson. There should be a way for the tutor to communicate intent to interrupt the student's performance through visualisation of gestures on the music. In this way

some of the functions of the shared score can be introduced to the separated lesson.

## Acknowledgments

The author would like to thank Doug Williams at British Telecom Research and Technology Future Content Group; Jonathan Reekie, Bill Lloyd and Marie Bennell at Aldeburgh Music; Ruth Aldred at Manchester Grammar School and Nicola Baigent at Trinity School of Music. Thank you to all of the students and tutors who kindly allowed their lessons to be filmed and the reviewers for their helpful feedback. The Media and Arts Technology programme is supported by an EPSRC Doctoral Training Centre EP/G03723X/1. Arts Council England supported Aldeburgh Music's purchase of equipment.

## References

- Barbosa, A. (2003, December). Displaced Soundscapes: A Survey of Network Systems for Music and Sonic Art Creation. *Leonardo Music Journal*, 13, 53–59.
- Broughton, M., & Stevens, C. (2009, April). Music, movement and marimba: an investigation of the role of movement and gesture in communicating musical expression to an audience. *Psychology of Music*, 37(2), 137–153.
- Bryan-Kinns, N., & Healey, P. (2006). Decay in collaborative music making. In *Proceedings of the 2006 conference on new interfaces for musical expression* (pp. 114–117). Paris: NIME.
- Chew, E., Zimmermann, R., Sawchuk, A. A., Kyriakakis, C., Papadopoulos, C., François, A. R. J., et al. (2004). Musical interaction at a distance: Distributed immersive performance. In *Music network 2004* (pp. 1–10). Barcelona: DIP Publications.
- Clark, H., & Brennan, S. (1991). Grounding in communication. *Perspectives on socially shared cognition*, 13(1991).
- Edigo, C. (1988). Videoconferencing as a technology to support group work: A review of its failure. In *Proceedings of the 1988 conference on computer-supported cooperative work* (pp. 13–24). Portland, Oregon: ACM.
- Hamilton, R., Iyer, D., Chafe, C., & Wang, G. (2008). To the Edge with China : Explorations in Network Performance. *Computer*, 7–8.
- Hammersley, M., & Atkinson, P. (2007). *Ethnography: Principles in practice* (Third ed.). London: Routledge.
- Healey, P., & Battersby, S. (2009). The interactional geometry of a three-way conversation. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 785–790). Amsterdam: COGSCI.
- Heath, C., & Lehn, D. (2005). Interaction and interactivities: collaboration and participation with computer-based exhibits. *Public Understanding of Science*, 1–23.
- Heath, C., & Luff, P. (1991a). Collaborative activity and technological design: Task coordination in London Underground control rooms. In *Proceedings of the second conference on european conference on computer-supported cooperative work* (pp. 65–80). Amsterdam: Kluwer Academic Publishers.
- Heath, C., & Luff, P. (1991b). Disembodied conduct: communication through video in a multi-media office environment. *Proceedings of ACM CHI 91 Human Factors in Computing*, 99–103.
- Heath, C., Luff, P., & Sellen, A. (1997). Reconfiguring Media Space. *Video-mediated communication*, 323–347.
- Hollan, J., & Stornetta, S. (1992). Beyond being there. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92*, 119–125.
- Kendon, A. (1990). *Conducting Interaction: Patterns of behavior in focused encounters*. Cambridge: CUP.
- Kurkul, W. W. (2007, February). Nonverbal communication in one-to-one music performance instruction. *Psychology of Music*, 35(2), 327–362.
- Lancaster, H. (2007). Are we (virtually) there yet? Face-to-face v. virtual learning landscapes in musical instrumental teaching. *CAUCE 2007*, 1–16.
- Marchand, T. H. (2010, May). Embodied cognition and communication: studies with British fine woodworkers. *Journal of the Royal Anthropological Institute*, 16, S100–S120.
- Moran, N. (2011, May). Music, bodies and relationships: An ethnographic contribution to embodied cognition studies. *Psychology of Music*.
- Roth, W.-M. (2001, January). Gestures: Their Role in Teaching and Learning. *Review of Educational Research*, 71(3), 365–392.
- Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 696–735.
- Sarkar, M., & Vercoe, B. (2007). Recognition and prediction in a network music performance system for Indian percussion. In *Proceedings of the 7th international conference on new interfaces for musical expression* (Vol. 2, pp. 317–320). New York, New York, USA: ACM.
- Schegloff, E. (1998). Body torque. *Social Research*, 65(3).
- Sellen, A. (1992). Speech patterns in video-mediated conversations. In (pp. 49–59). Monterey: ACM.
- Vines, B., Wanderley, M., Krumhansl, C., Nuzzo, R., & Levitin, D. (2004). Performance gestures of musicians: What structural and emotional information do they convey? *Gesture-based communication in human-computer interaction*.
- Wanderley, M., Vines, B., Middleton, N., McKay, C., & Hatch, W. (2005, June). The Musical Significance of Clarinetists' Ancillary Gestures: An Exploration of the Field. *Journal of New Music Research*, 34(1), 97–113.
- West, T., & L. Rostvall a. (2003, May). A Study of Interaction and Learning in Instrumental Teaching. *International Journal of Music Education*, 40(1), 16–27.
- Whittaker, S. (2003). Theories and Methods in Mediated Communication. *The handbook of discourse processes*(973), 243–286.

# Increased Vigilance in Monitoring Others' Mental States During Deception

Nicholas D. Duran (nduran2@ucmerced.edu)

Rick Dale (rdale@ucmerced.edu)

Cognitive and Information Sciences, The University of California, Merced, Merced, CA, 95343

## Abstract

Communicating false information during social interactions places unique cognitive demands on the speaker. One such demand is the increased need to track what another knows to avoid introducing contradictory information. The present study examines the minimal conditions required for vigilant mental state monitoring by using a game-like task that required participants to respond to virtual partners' questions with true or false information. In this task, there were no explicit demands to engage in mental state monitoring. We found increased response times when the answers potentially violated a partner's belief states - but only if participants believed their partner to be an actual cognitive agent. These effects were also shown to be additive to the simple demands in suppressing a truth bias while responding falsely. We argue that participants exert greater cognitive effort as an automatic response elicited by being situated in an interactive, deceptive context.

**Keywords:** deception; belief monitoring; self; other

## Introduction

In this paper, we examine the underlying cognitive processes that are involved in deceptive responding during a real-time social interaction. We propose two primary processes: 1) the inhibition of a prepotent truth response while responding falsely, and 2) the need to track the knowledge of the person to whom one is falsely responding. Both of these processes are commonly associated with executive control and working memory, and thus are hypothesized to involve considerable cognitive costs. Previous studies on deceptive behavior have examined these processes in isolation. In this paper, we show the integrated effects of each process in a single task. Moreover, we provide evidence that when participants merely know that they may have to convey false information, even when the risk of detection is negligible, it is sufficient to induce vigilant tracking of the knowledge of the false information's recipient. However, as an important caveat, this effect only occurs when a person takes an *intentional stance* toward another (Dennett, 1987). As we argue below, such tracking is likely an adaptive response that ensures false or true information conveyed throughout an interaction remains consistent, thereby minimizing the risk of detection.

## Background

Human interaction often involves the exchange of deceptive information. Although these fabrications are mostly innocuous, as in lying to boost one's own credentials or to protect another's feelings, they are routinely introduced into conversations (DePaulo & Kashy, 1998). Nevertheless, the risk of detection when communicating false information is always present. To minimize this risk, plausibility must be maintained, and as such, sophisticated cognitive control processes are likely recruited. A primary need is the suppression

of a truth bias while responding falsely (Duran, Dale, & McNamara, 2010), as well as a need to remember what information is false and maintain this falsity as the ostensible truth in the mind of another (Sip, Roepstorff, McGregor, & Frith, 2008). A number of neuroimaging and reaction time studies have explored truth suppression, showing increased executive function processes involving the inhibition of predominant responses and the resistance to interference (see Spence et al., 2004, for a review).

Much less is known, however, about the processes underlying real-time deception in social interactions. Some work has focused on the recipients of deception and their beliefs about the hidden motivations of their partners (e.g. Schul, Mayo, & Burnstein, 2004). Conversely, others have looked at those doing the deception and have found evidence for active monitoring of their partners' suspicion, purportedly for the purpose of manipulating others' beliefs about their own goals and intentions (Bhatt, Lohrenz, Camerer, & Montague, 2010; Carrion, Keenan, & Sebanz, 2010). In these studies, the modeling of another's mental state consists of the impressions that the other is likely to form. Participants are given opportunities to mislead a partner about the true nature of some privileged knowledge, as a poker player does when bluffing about the cards in their hand. Successful participants are those that strategically fend off another's suspicion with well-timed true responses. These true responses, construed as "deceptive truth," are believed to result from mental state monitoring, and invoke neural regions of cognitive effort similar to that associated with various theory of mind tasks. Beyond considerations about the general impressions or motivations another might possess, deception likely involves the specific content one believes another to believe about events and objects in the world following the deceptive act.

We explore this aspect of mental state monitoring in deceptive interaction; specifically, the need to encode and maintain the factual content of what another might know. This is particularly difficult during an extended interaction where new information is introduced and must be integrated into deceiver's knowledge of another's supposed knowledge. If one is to give another false information about their whereabouts, such as saying they were not at the local Sears on Saturday, the deceiver needs to maintain this model of events for the other, and apply it consistently downstream in the conversation to avoid saying "yes" to a question like, "Were you shopping on Saturday?" From this example, implications that might be drawn from the earlier presented false information also need to be maintained by the deceiver, such as knowing that being at Sears implies shopping. Such a model of others' belief states is likely an adaptive evolutionary response that minimizes the risk of detection (Premack, 2007), but one that

also enacts considerable cognitive costs that combines with those associated with truth suppression (a process that can occur regardless of an audience). To evaluate both these processes in an integrated manner, we turn to a novel *guessing game* task.

The basic structure of this task involves a partner attempting to guess the identity of an object of which another partner is solely aware. This task is similar to the game of “Twenty Questions,” where a “questioner” tries to guess a person, place, or thing by asking yes or no questions. After the allotted number of questions have been asked, the questioner must wager a guess. This task presents a situation where the mental state of another (i.e., the questioner) incrementally converges on what another has in mind (i.e., “answerer”). Although the uninformed questioner is making explicit attempts to converge, the same does not necessarily hold for the partner who is answering yes or no. That is, to succeed in this task, the answerer does not need to track the evolving image emerging in the questioner’s mind. However, in our version of the Twenty Questions game, the answerer is told that there is the possibility of having to give the questioner false information. This simple instructional manipulation is hypothesized to elicit increased vigilance in tracking the mental state of the questioner, largely because the answerer now has to maintain the veracity of what the other believes. Important to the task set-up, there are no explicit instructions for monitoring another’s mental state, nor does it affect the success of completing the task.

Mechanisms involved in such a situation are likely automatic and triggered in response to particular contexts (German, Niehaus, Roarty, Giesbrecht, & Miller, 2004) - one of which, as we argue, is the communication of false information. However, merely being situated in a context that requires transmission of false information is not sufficient. The participant must also take an intentional stance toward their partner, a partner who is thought to have their own set of beliefs, desires, and knowledge states. In other words, participants must consider their partners as having minds worth tracking (Gallagher, Jack, Roepstorff, & Frith, 2002). One test of this claim is to compare participants (i.e., answerers) who believe their partner to be real versus a computer simulation, while holding all other features of the interaction equivalent. Accordingly, on critical trials all participants should find the false responses more challenging than the truth (due to suppression of a truth bias); however, for those who take an intentional stance, they will experience added difficulty because they will also be violating knowledge their partner possesses.

In what follows, we describe in greater detail the method used to assess other-directed mental state monitoring during deception. We then report the findings from two experiments that incorporate crowdsourcing techniques. We end with a brief discussion of the findings and limitations.

## Mental State Monitoring in a Guessing Game

### Initial Set-up

The current task was implemented as an online, Flash-based game that makes use of key features in Amazon Mechanical Turk (AMT). AMT is a crowdsourcing platform that allows participants (i.e., “Workers”) to sign up to complete tasks posted by other users (i.e., “Requesters”). There are many advantages of using crowdsourcing techniques (see Munro et al., 2010, for a review), with one notable advantage being the ability to create an illusion of connectivity, whereby simulated, recorded partners can act as convincing interactive partners (Duran, Dale, & Kreuz, 2011). To achieve this illusion in this study, participants were recruited under the pretext of examining “how people solve problems while receiving misleading information.” They were then told that the task involved beta software that allows us, the Requesters, to connect two Mechanical Turk Workers, but with software that only allows a one-way transmission of audio. The participant’s partner (a recorded simulation of a male’s voice) was always designated as the role of questioner and was the one who would transmit audio. A series of validity checks were then presented that highlighted the connection, such as a “connection screen” where participants ostensibly waited for the software to locate and connect them to their partner, and an “introduction screen” where the recorded partner introduced himself and provided a secret codeword for the participant to enter, ostensibly verifying to the recorded partner that they were “connected” to a real person. As described further below, checks were used to ensure participants believed this sham connection.

### Basic Game Structure

In the initial instructions, participants were provided a demonstration of how the task was structured (see Figure 1 for a flow diagram). First, participants were told that they would be presented with one of two objects (an alarm clock or a red apple) that only they could see. Their partner would then ask a yes or no question to attempt to guess the identity of the object, and once the question was asked, they would trigger a “GO” button that would appear at the bottom-center position of the participant’s screen. When the participant clicked this button, a response screen would appear with “YES” and “NO” response buttons positioned in the top-right and top-left corners. On this screen, a prompt to respond with a “TRUTH” or “LIE” also appeared in the middle of the screen.<sup>1</sup> Thus, if the partner had asked, “Is it a person?,” the participant had to navigate their computer mouse from the bottom of the screen to the “YES” response button. The response was then transmitted to the partner and a short pause was introduced to allow the partner to “formulate” another question before the next trial began.

<sup>1</sup>The use of a “LIE” response prompt has been used in previous research to approximate deceptive behavior and has been shown to invoke similar physiological and neurological reactions as unsolicited deception. However, we acknowledge that this is still an approximation of deceptive behavior, which we address in Study 2.

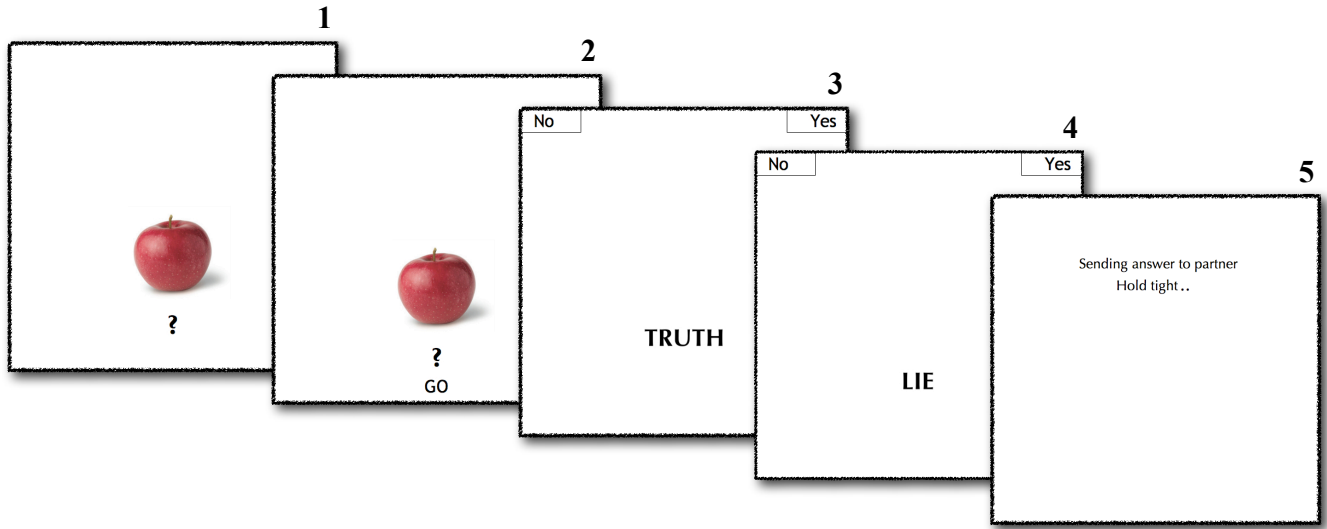


Figure 1: Flow of interaction in experimental task. In (1), after initial instructions, a typical trial begins by presentation of the to-be-guessed object. The partner then asks a question about the identity of the object; (2) After the question is asked, “GO” appears and participants click to respond; (3) and (4) Participants are given a prompt to respond truthfully or falsely, and do so by clicking “YES” or “NO” boxes; (5) Response is “sent” to partner.

### Monitoring and Contradicting Given Information

Two rounds of the guessing game were played, counterbalanced across the object to be guessed (clock or apple). In each round, the partner asked 5 questions before they made a guess.<sup>2</sup> There were 16 variations of each question set for each object, with each set randomly selected for each participant. In both rounds, participants were prompted to answer truthfully for the first three questions, but for the 4th or 5th questions, they were prompted to respond falsely. These falsification questions presented a situation where participants might violate their partner’s possible belief states. What this means is that by the 4th or 5th questions, the partner has eliminated numerous options of what the object might be, allowing a greater possibility for some objects, like an alarm clock, to be mentally represented. If participants take into consideration this reduced mental set of their partner, a response that denies the possibility of an alarm clock acts to contradict what the partner was increasingly likely to believe. This type of *general* violation constituted one round of the guessing game.

In another round, a *specific* violation was presented. In the corresponding scenario, the question being falsified contradicted information explicitly given earlier in the informational exchange. For example, if an early question like, “Can it be used on a daily basis?,” is initially confirmed by the participant as being true, the participant can infer that their partner believes that the unknown object is “common.” When a later-occurring question such as, “Is it fairly common?,” must then be falsified, the participant is now violating the more

specific belief their partner was inferred to have. An example sequence of questions is given in Table 1. These two types of violation, general and specific, were presented to the participant across the two rounds, in counterbalanced order.

<i>Question Sequence</i>	<i>Expected Response</i>	<i>Prompt</i>
1. Is it a person?	No	True
<b>2. Did someone invent it?</b>	<b>No</b>	<b>True</b>
3. Do people eat it?	Yes	True
4. Does it grow underground?	No	True
<b>5. It is a man-made product?</b>	<b>Yes</b>	<b>False</b>

Table 1: Sample questions asked by simulated partner who is attempting to “guess” an apple. The later-occurring Question 5, in which participants are prompted to falsify, contradicts specific information given earlier in Question 2. Questions were also counterbalanced for whether critical trials required a “yes” or “no” response.

### Establishing Intentional Stance

To examine the incurred processing costs of violating the mental state of another, we compared participants who believed that they were interacting with a real partner versus those who did not. Participants who did not believe they were interacting with a real person were hypothesized to be less likely to attribute mental states to the simulated partner, and

<sup>2</sup>Limited to 5 questions to avoid memory interference or decay that would have likely resulted from 20 questions.

thus should not experience the associated processing costs on the critical false trials. The only costs these participants should experience is that attributable to the suppression of a truth bias when responding falsely. To identify participants who did or did not believe they were interacting with a real partner, we asked two critical follow-up questions at the end of the task. The first probed whether the participant would give a small monetary bonus (paid by us, the Requesters) to their partner for the quality of questions asked. The rationale for including this question is that a participant who suspects that their partner is a mere recording is unlikely to give a reward. The second question was more direct, and asked participants to rate on a scale from 1 to 7 the degree to which they thought they were connected to a real person. Participants who would give a reward, and were on the upper end of the scale for their belief that they were interacting with a real person, were considered those who would take an intentional stance.

## Experiment 1

We collected data from 104 participants. One subject was excluded for answering two of the 10 questions incorrectly. We also removed excessively long trials that were over 8000 ms (0.73% of data), and from this truncated set, we removed trials that were more than three SDs (855 ms) above the mean response time (2308 ms). This resulted in a loss of 2.48% of the data. Forty-eight participants also self-selected into a group who believed that their partner was real, with the remaining participants believing that their partner was not real.

## Results

A mixed effects ANOVA was used to compare the difference in response times for the two groups of self-selected participants, with a within-subjects factor of whether a trial required a true or false prompt<sup>3</sup>.<sup>4</sup> Subject was entered as a random effect. The analysis was conducted using the lmer package in the R statistical software. In this package,  $p$ -values are computed with 10,000 Monte Carlo Markov Chain simulations, using lmer's pvals.fnc function (Baayen, Davidson, & Bates, 2008). We report these  $p$  values and the unstandardized effect estimates for the main effects and interactions.

The results indicate a main effect for participant type (those who believe vs. not believe),  $B = 342$  ms,  $p = .003$ ; as well as for prompt type (false vs. true),  $B = 481$  ms,  $p < .001$ . There was also an interaction between participant and prompt type,  $B = 277$  ms,  $p = .05$ . In follow-up tests to examine this interaction, it appears that for both believers and non-believers,

<sup>3</sup>Because of the smaller number of false critical trials compared to true trials, and because the false trials always occurred near the end of the interaction, we only analyzed the true trials that occurred immediately before the presentation of the false prompt trials. The inclusion of all true trials does not radically alter the reported findings, as the response pattern in the data remains consistent. However, by doing so, the significant interaction is now only marginally significant ( $p = .10$ ).

<sup>4</sup>The comparison between the false prompts that draw on general and specific belief violations showed no statistically significant differences, and thus are combined into one condition.

the false response critical trials showed greater response latencies than true response latencies: believers,  $B = 622$  ms,  $p < .001$ ; and non-believers,  $B = 345$  ms,  $p = .003$ . This effect corresponds to the greater cognitive difficulty associated with suppressing a false response. Moreover, the magnitude of the false response time latencies for believers was much greater than those who did not believe,  $B = 500$  ms,  $p < .001$  (see Figure 2). This finding suggests that believers noticed a contradiction forced by the false prompt that the non-believers did not. The likely reason is that believers had an active model of the other's mental state and experienced greater cognitive effort in consulting and ultimately violating this knowledge.

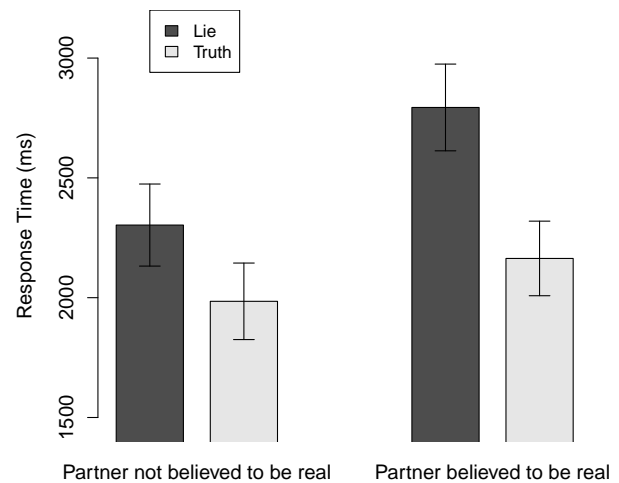


Figure 2: Overall, the response time latencies are much higher for false (lie) responses compared to true responses, with the greatest response latencies for participants who believed their partner to be real. There are no significant differences across true responses.

## Experiment 2

In Experiment 1, participants appear to use true knowledge provided earlier *upstream* in an interaction to influence later *downstream* responses. As a further test, we examined whether participants encode and consistently maintain false information that is provided upstream. The set-up of this task is such that participants should respond falsely to a later-trial critical question to verify the accuracy of earlier presented false information. To do so, participants are allowed to freely respond without prompts. We hypothesize that participants who believe they are interacting with a real partner will be more likely to choose to respond falsely. In other words, they are more likely to respond falsely *again* to preserve the consistency of their partner's beliefs (because of false information they presented upstream). This response will enact similar cognitive demands as evidenced in Experiment 1.

## Modified Method

Participants first provided their partner with a particular belief about the to-be-guessed object early in the interaction.

This is done by prompting the participant to respond falsely on the second question asked by their partner. Later, downstream in the questioning (on the critical fourth question in the questioning sequence), participants are allowed to choose whether to respond falsely or tell the truth. In one round of the guessing game, the downstream question is related to the earlier false information, and thus if a participant is monitoring what the other knows, and wants to stay consistent with the false depiction of the object, they will voluntarily respond falsely (see Table 2 for a sample sequence of questions). By consulting and acting on the knowledge state of the other, we expect this to incur a processing cost. As a type of control, in a second round of the guessing game, the downstream question is unrelated to the earlier false information, and thus there is no explicit need to consult what the other knows on the critical question. Thus, the differences in the *information related* round should no longer be present in this *information unrelated* round.

Question Sequence	Expected Response	Prompt
1. Is it a thing	Yes	True
2. Can you easily pick it up?	No	False
3. Is it found in people's homes?	Yes	True
4. Can it be moved?	?	--
5. Is it something people might use on a daily basis	?	--

Table 2: Sample question set asked by simulated partner who is attempting to “guess” an alarm clock. The downstream Question 4, which requires a free response, relates to information falsified in Question 2. There are 8 variations of question sets for each object (total of 16).

## Results

One hundred and seventy-six participants supplied data via Mechanical Turk. Five participants were removed for answering two or more of the 10 question incorrectly, and an additional 16 were removed for failing to provide at least one false response in the final unprompted questions. Outlier trials, those trials that exceeded 3 SDs above the mean were also removed. Furthermore, based on follow-up questions, 80 participants were self-selected as those who believed they were interacting with a real partner, 71 believed they were interacting with a simulation, and four participants’ beliefs were undetermined. Thus, 151 participants provided data where their false and true responses could be evaluated.

The main analysis specifically targets the round in the guessing game where there was an opportunity to maintain the false information introduced earlier in the question sequence (i.e., “information related” round; see Figure 3a). By freely answering false on the fourth (critical) question in the

sequence, the false belief state of the questioner is maintained. We hypothesized that this process requires greater processing time because the deceiver must consult what the other knows, recognizing that to respond true would elicit a contradiction. Importantly, such behavior is likely to occur only when a participant believes they are interacting with a real partner. We found evidence for this hypothesis in a simple t-test evaluating the critical false response trials between participants who did or did not believe they were interacting with a real partner,  $t(76) = 2.20$ ,  $p = .03$  (Figure 3a).

It should also be noted that the false response trials in the above analysis represented responses from 41 of the 80 participants who could be classified as believers. This number is fewer than expected given the hypothesis of increased vigilance in maintaining the partner’s belief states. However, when participants who believed they were interacting with a real partner did answer truthfully (thereby contradicting a partner’s mental state), these response times were the second highest of all trial groups (Figure 3a). These elevated scores suggest that participants are aware, at some level, that they are violating their partner’s mental state. This is supported by a significant main effect in a mixed effects ANOVA for participant type (belief vs. not belief), showing that participants who believed they were interacting with another, despite answering truthfully or falsely, had increased response times compared to those who did not believe,  $B = 621$  ms,  $p = .02$ .

Finally, as hypothesized for the round in the guessing game where the upstream information was unrelated to the downstream information, no differences were found between participants who did or did not believe they were interacting with a real partner (see Figure 3b)

## General Discussion

Across two experiments we examined response behavior in a context where participants had to transmit false information to another. We found evidence that participants experience greater cognitive effort in suppressing a truth bias; and furthermore, show evidence of increased effort when they are confronted with a response that violates or potentially violates the mental state of another (as measured by response latencies). For the latter, we argued this increased effort results from an active, or vigilant, monitoring of what another believes. Participants appear to do so as long as they think they are interacting with a “mindful” agent, and also do so despite instructions that have no explicit requirement to consider others’ mental states.

A limitation in this study, found in Experiment 2, is that participants who believed their partner to be real did not overwhelmingly choose to maintain the false information provided upstream in the interaction. One reason is that being detected as providing contradictory information carried little consequence, thus there was little motivation to choose a false response. Other research also suggests strong individual differences in whether participants consider their partners’ sus-



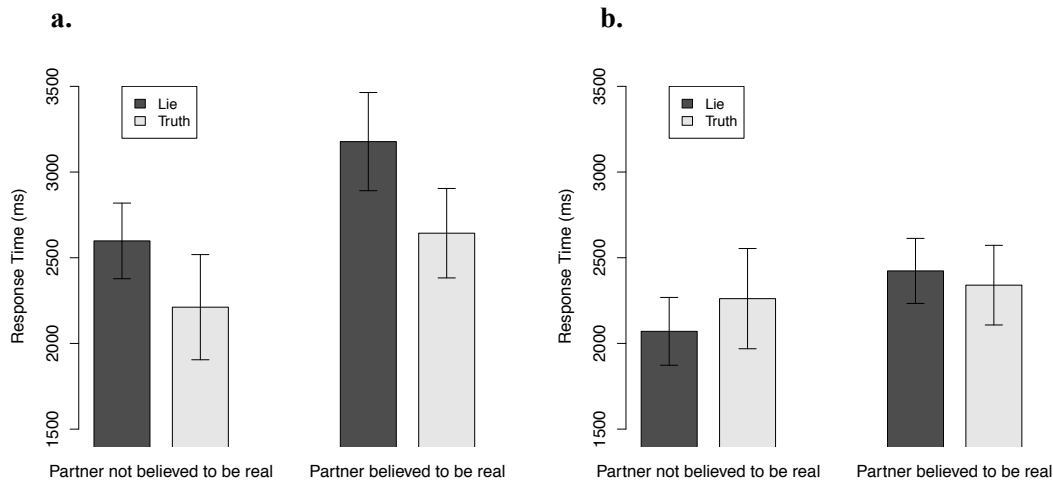


Figure 3: In (a), response times for participants who answered falsely on the critical free response questions is highest for those who believed they were interacting with a real partner versus those who did not believe. There are no differences for true responses. In (b), the control with unrelated information downstream showed no significant effects.

picion in low risk interactions (Bhatt et al., 2010). Despite this limitation, those participants who did think they were interacting with another, and who responded falsely, were the only group to be influenced by the contradictory information in the information stream.

In sum, we explored how truth biases and social factors are jointly involved in simulated “deceptive acts.” While we grant that these are basic cognitive experiments that can only loosely approximate naturalistic contexts, we would argue that the approach opens new avenues of investigation. The underlying *cognitive mechanisms* of deception are still being sought. By employing basic cognitive experimentation in simple but controllable tasks, we could gain a more systematic understanding of the mechanisms underlying deceptive acts. These experiments are a step in that direction.

## References

- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.
- Bhatt, M. A., Lohrenz, T., Camerer, C. F., & Montague, P. R. (2010). Neural signatures of strategic types in a two-person bargaining game. *Proceedings of the National Academy of Sciences*, 107, 19720.
- Carrion, R. E., Keenan, J. P., & Sebanz, N. (2010). A truth that's told with bad intent: An ERP study of deception. *Cognition*, 114, 105-110.
- Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: The MIT Press.
- DePaulo, B. M., & Kashy, D. A. (1998). Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology*, 74, 63-79.
- Duran, N. D., Dale, R., & Kreuz, R. J. (2011). Listeners invest in an assumed other's perspective despite cognitive cost. *Cognition*, 121, 22-40.
- Duran, N. D., Dale, R., & McNamara, D. S. (2010). The action dynamics of overcoming the truth. *Psychonomic Bulletin & Review*, 17, 486-491.
- Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *Neuroimage*, 16, 814-821.
- German, T. P., Niehaus, J. L., Roarty, M. P., Giesbrecht, B., & Miller, M. B. (2004). Neural correlates of detecting pretense: Automatic engagement of the intentional stance under covert conditions. *Journal of Cognitive Neuroscience*, 16, 1805-1817.
- Munro, R., Bethard, S., Kuperman, V., Lai, V., Melnick, R., Potts, C., et al. (2010). Crowdsourcing and language studies: the new generation of linguistic data. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 122-130.
- Premack, D. (2007). Human and animal cognition: Continuity and discontinuity. *Proceedings of the National Academy of Sciences*, 104, 13861.
- Schul, Y., Mayo, R., & Burnstein, E. (2004). Encoding under trust and distrust: The spontaneous activation of incongruent cognitions. *Journal of Personality and Social Psychology*, 86, 668-679.
- Sip, K. E., Roepstorff, A., McGregor, W., & Frith, C. D. (2008). Detecting deception: The scope and limits. *Trends in Cognitive Sciences*, 12, 48-53.
- Spence, S. A., Hunter, M. D., Farrow, T. F. D., Green, R. D., Leung, D. H., Hughes, C. J., et al. (2004). A cognitive neurobiological account of deception: Evidence from functional neuroimaging. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359, 1755-1762.

# Mining Relatedness Graphs for Data Integration

Jeremy T. Engle  
([jtengle@indiana.edu](mailto:jtengle@indiana.edu))

Ying Feng  
([yingfeng@indiana.edu](mailto:yingfeng@indiana.edu))

Robert L. Goldstone  
([rgoldsto@indiana.edu](mailto:rgoldsto@indiana.edu))

Indiana University  
Bloomington, IN. 47405 USA

## Abstract

In this paper, we present the AbsMatcher system for schema matching which uses a graph based approach. The primary contribution of this paper is the development of new types of relationships for generating graph edges and the effectiveness of integrating schemas using those graphs. AbsMatcher creates a graph of related attributes within a schema, mines similarity between attributes in different schemas, and then combines all information using the ABSURDIST graph matching algorithm. The attribute-to-attribute relationships this paper focuses on are semantic in nature and have few requirements for format or structure. These relationships sources provide a baseline which can be improved upon with relationships specific to formats, such as XML or a relational database. Simulations demonstrate how the use of automatically mined graphs of within-schema relationships, when combined with cross-schema pair-wise similarity, can result in matching accuracy not attainable by either source of information on its own.

**Keywords:** Data integration; graph matching; ABSURDIST; semantic relatedness.

## Introduction

Data integration has application to a wide variety of fields from e-commerce to bioinformatics. One of data integration's subtopics is the attribute matching problem which finds mappings between attributes in source and target data sets. This paper presents the AbsMatcher framework which concentrates on one-to-one attribute matches as an initial effort, leaving complex n-to-one matches to future work. AbsMatcher finds matching results based on graphs of within-schema attribute relationships and cross-schema comparisons of attribute similarity. The focus of in this paper is the process and relationships used to create a within-schema graph for each data set.

The AbsMatcher framework has two distinct phases. The first is a mining phase which produces a graph for each data set where edges relate within-schema attributes and an aggregated matrix of cross-schema attribute similarities measures. We refer to these graphs as internal information, because a graph only contains information relating attributes within the same schema. Information which is aggregated into the cross-schema matrix is referred to as external information because it involves a comparison between

attributes in different schemas to determine how semantically similar they are.

Secondly, AbsMatcher's matching phase uses the ABSURDIST (Feng, Goldstone, & Menkov, 2004; Goldstone & Rogosky, 2002) algorithm to combine mined information and determine match correspondences using an iteratively converging global optimization algorithm. ABSURDIST was originally developed to translate between conceptual systems in a psychologically plausible manner. Additionally, ABSURDIST has a weighting ratio to determine the balance of influence on the outcome of internal and external information. Though we focus on specific sources of internal and external information in this paper both AbsMatcher and ABSURDIST were designed in a way so that additional sources could easily be added.

In a graph based approach to schema matching the matching process uses graph matching to determine mappings between attributes/nodes based on the similarity of their sets of relationships/edges. Edges of graphs are labeled with different relationship types, which represent different forms of information. Relationship types can broadly be divided into structural relationships which are based on how attributes are organized in a data set and semantic relationships which are based on meaning of the information associated with an attribute. An example of a structural relationship is a parent/child relationship between nested attributes from an XML data set. An example of a semantic relationship is a general/specific relationship for the concepts represented by two attributes. These examples highlight that one of the challenges in how to create graphs is that what relationships can be used is tied to the format of a data set and the available (meta)data.

Previous systems (Aumüller, Do, Massmann, & Rham, 2005; Giunchiglia & Shvaiko, 2003; Melnik, Garcia-Molina, & Rahm, 2002) address the problem of how to create graphs by only using metadata which can be intuitively translated to graph form. As a result, the graphs created by these translations predominately represent the structural design of a data set. Though structural relationships can be useful, their disadvantage is that factors such as missing metadata, differing metadata formats, or different database designers can remove these structural relationships' usefulness.

This paper presents a set of relationships which can be used as general practice but more importantly are still applicable when metadata is limited or datasets with differing formats are being integrated. Of particular contribution are Yahoo semantic relatedness relationships which leverage Yahoo query results to measure the semantic relatedness of attributes' names. Yahoo relationships are an improvement over the use of tools like Wordnet because of their ability to handle attribute names that have abbreviations, words unique to a domain, and/or multi-term phrases. All three of these factors commonly occur in data sets. Together, Yahoo relationships and the other relationships we present offer tools to use when structural metadata is absent or of no benefit.

We use the terms AbsMatcher and ABSURDIST throughout this work. AbsMatcher is the overall system which formulates graphs after mining internal information and aggregates mined sources of external similarity. ABSURDIST refers specifically to the matching phase which iteratively combines internal and external information to determine a set of correspondences.

### ABSURDIST Background

ABSURDIST was developed to solve the general problem of translating between two conceptual systems. We adapt this approach to data integration by treating attributes as concepts to be matched. A complete discussion of ABSURDIST and how information factors into the iterative process can be found in Goldstone and Rogosky (2002). Information in ABSURDIST is classified as internal (within-schema) or external (cross-schema). External information provides the ability to input cross-schema similarity into the ABSURDIST algorithm. Different external sources are aggregated into an NxM matrix of values between 0 and 1, where N and M are the sizes of the schemas to be matched. The dividing line between internal and external is that internal information is relationships between attributes in the same schema, whereas external similarity is a comparison between attributes in two separate schemas.

ABSURDIST iteratively updates correspondences using internal and external information until reaching a stable point, terminates, and selects the final matches. ABSURDIST as an error minimization algorithm selects the set of matches that result in the least total link error. This section discusses the conceptual motivations of ABSURDIST and leaves specific examples of internal and external information for later sections.

#### Internal Information as Graphs

Internal information in ABSURDIST represents intra-system information about how nodes in each conceptual system relate to other nodes in the same system. Internal information for a system is independent of the system with which it is being aligned. For each schema, ABSURDIST takes internal information as input in the form of: information on relationship types, node types, node

information, and a graph of relationships. Internal information factors into the  $R$  and  $I$  terms of Equation 1. A node in a conceptual system must have a unique identifier and a categorical type. If only one type exists then the effects of node types become irrelevant. Relationships in ABSURDIST represent a conceptual association between intra-system nodes creating a generalized interpretation of structure. A relationship type has a categorical label and is defined as being either directed or undirected. Relationships are instantiated as edges, which collectively form a graph of continuously valued weighted edges. If the same weight is used for every edge, these weights become irrelevant.

#### Iterative Algorithm

ABSURDIST is an iterative algorithm which updates an NxM matrix of correspondences where N and M refer to the number of attributes in the source schema,  $A$ , and target schema,  $B$ , respectively. Each cell in the correspondence matrix,  $C_t(A_q, B_r)$ , represents how strong a match is at iteration step  $t$  for attribute  $q$  in schema  $A$  and attribute  $r$  in schema  $B$ . The algorithm terminates when the matrix has converged or a maximum number of iterations is reached. For each iteration, ABSURDIST updates each  $C_t(A_q, B_r)$  by a net input defined by

$$N(A_q, B_r) = \alpha E(A_q, B_r) + \beta R(A_q, B_r) - \chi I(A_q, B_r)$$

#### Equation 1. Correspondence Update Equation

Equation 1 shows how internal ( $R$  and  $I$ ) and external ( $E$ ) information combine to update the correspondence from attribute  $q$  in schema  $A$  to attribute  $r$  in schema  $B$ . The  $E$  term represents similarity based on external information, the  $R$  term represents similarity based on internal information, and the  $I$  term uses internal information to inhibit incorrect correspondences. As a global optimization algorithm, both  $R$  and  $I$  take into account the state of the system at each iteration  $t$ .  $\alpha$ ,  $\beta$ , and  $\chi$  are weights that control the influence of forms of information, where  $\alpha$  and  $\beta$  are set as a ratio to each other and  $\chi$  is set independently of the others. For example, when  $\alpha$  is one and  $\beta$  is zero only external information is used to find correspondences.

#### Related Research

A number of surveys have been done which cover the different aspects of the schema matching problem (Shvaiko & Euzenat, 2005). One of the established approaches to schema matching is to use candidate matchers to generate candidate matches which are aggregated into a final set. Graph-based systems, including AbsMatcher, have multiple modules to generate edges in the graph, multiple modules to generate the equivalent of external information, and then use a graph matching algorithm to generate correspondences based on graphs. It is possible that correspondences generated using a graph matching algorithm could be used as a candidate matcher in a system. Cupid (Madhavan, Bernstein, & Rahm, 2001), and Similarity Flooding (Melnik, Garcia-Molina, & Rahm, 2002) systems all use

graph matching to accomplish schema matching. COMA++ (Aumuellner et al., 2005) is a generalized framework for schema matching which was used in the Similarity Flooding system to combine the results from graph matching with non-graph-oriented candidate matchers. The difference between AbsMatcher and these previous systems is the generality of AbsMatcher and generating graphs based on semantics instead of data model metadata.

Previous graph-based schema matchers construct graphs based on the metadata for the data model. These systems have modules specifically built for translating different data models -- such as relational databases, XML, ontologies, or conceptual hierarchies -- into a graph form. This approach makes the graphs generated dependent on the thoroughness of the data set creator, and completely different graphs will be generated even when the same data set is stored in different data models. The advantage of these systems is that they leverage the effort of data set creators. For example considerable effort is generally put into the design phase of a relational database. Examples of using metadata would be creating a relationship between parent and child XML attributes or the fact that an attribute is a primary key in a relational database. The disadvantage of basing graphs on metadata is that derived relationships often have more to do with how data is stored and less about semantic relationships. The goal of the information sources we present in this paper is that they can be used regardless the data model and still generate semantic relationships.

The Semantic Matching (Giunchiglia & Shvaiko, 2003) system provides the closest comparison to AbsMatcher. It creates a graph based on metadata and a limited number of semantic relationships. Semantic Matching uses electronic thesauri in order to create overlap, mismatch, and general/specific relationships. The one issue with electronic thesauri is that they only work with words in their index and are unable to handle abbreviations or phrases which are often used to name attributes. AbsMatcher shares the same motivation as Semantic Matching, but uses the web to create semantic relatedness relationships and mines the data sets for statistical relatedness relationships. Additionally, ABSURDIST was designed with a general idea of relationships, which makes adding new forms of internal relationships a simple process.

We mine semantic relatedness using Yahoo query results (Bollegala et al., 2007) and Information Dependencies (Dalkilic & Robertson, 2000), however, neither has been used for schema matching.

## Mining ABSURDIST Graphs

The focus of this paper is on the process and relationships types used to create within-schema graphs. The unifying characteristic for all of the relationships we present is that they are not specific to a data model nor represent structural information. We present two categories of relationships; ones which use the entropy of the data and the second which uses Yahoo query results based on attribute names to measure semantic relatedness.

Mining an ABSURDIST graph is a two-stage process. The first is mining edges of the desired relationship type and the second is filtering out noisy edges. Filtering is done by using thresholds to eliminate mined edges whose values are not statistically significant enough to represent something beyond noise. For brevity's sake we limit the discussion of filtering to describing what the threshold checks for each relationship type.

## Entropy Relations

Entropy-based relationships use an information theoretic approach to look at the information content of attributes based on their data. The goal is to look for patterns which defy statistical trends and therefore are more likely to represent user intended relationships. We use the Information Dependency (InD) measure (Dalkilic & Robertson, 2000), which is based on Shannon's Entropy, to look at the information content of attributes. Entropy relationships require at least a sample of the data. The discussion of Entropy relationships includes approximate attribute entropy relationships, data set key relationships, and approximate functional dependencies.

Attribute entropy relationships measure the degree to which attributes resemble keys, which have a different value in each record in the data set for the attribute, or constants, which have the same value in each record in the data set for the attribute. An attribute being close to a key or constant is a unique statistical property which is a result of how data is created, e.g. an ISBN is purposefully defined as a key. Attributes in other data sets that are semantically similar are likely to also have similar statistical properties, so when keys or constants occur they are strong indicators of a likely match. In Table 1, *PersonName* is an example of a key and *Gender* is an example of an attribute that is almost a constant. Attribute Entropy relationships are filtered based on their entropy values and only kept when those values are either above (approximate key) or below (approximate constant) defined thresholds.

Data set keys are sets of attributes that together have a unique set of values for the data set and therefore form a key. Data set key relationships are created between pairs of attributes that together are close to forming, or do form, a data set key, but neither attribute is a key on its own. An example from Table 1 is that by combining *Address* and *Gender* a unique set of values exists for every row. The above example would result in an edge *PairKey*(*Address*, *Gender*) to be created in the graph. A data set key relationship creates undirected edges between attributes and uses the entropy value as the weight. Data set approximate key relationships are filtered using a threshold which defines how close to a primary key the attribute set must be.

The last Entropy relationship type uses Approximate Functional Dependencies (AFDs). AFDs are probabilistic rules,  $X \rightarrow Y$ , which measure the ability of values for a left hand side (LHS) attribute set to determine values of the right hand side (RHS) attribute set. The closer an AFD's measured value is to 1 the better the LHS is at predicting the

RHS. AbsMatcher’s use of AFDs as an information source for schema matching presents a novel application for AFDs. We use AFDs which have a single attribute LHS and a single attribute RHS in creating dependency relationships. By only using single attributes on each side the search space is reduced from  $2^{N+M}$  to  $N \times M$ . Though Functional Dependencies (FDs), which AFDs extend, have been used in schema matching, this is to our knowledge the first use of AFDs. Filtering dependency relationships uses a threshold which parameterizes the number of standard deviations that an AFD’s value must be away from the average value of all AFDs with the same LHS or RHS.

Table 1. A sample data set of people

PersonName	Address	Gender
Santa Claus	100 North Pole	Male
Mrs. Claus	100 North Pole	Female
Jeremy Engle	215 Lindley Hall	Male
Rob Goldstone	338 Psychology	Male

### Semantic Relationships

The premise behind using semantic relatedness is to create a relationship between attributes that are thematically related. A trivial example of this would be attributes for the first and last name of a person. If the respective attribute labels are “first” and “last” then a graph edge is created between these attributes based on the thematic association of these labels.

$$WebJaccard(P,Q) = \begin{cases} 0, & H(P \cap Q) < c \\ \frac{H(P \cap Q)}{H(P) + H(Q) - H(P \cap Q)}, & H(P \cap Q) \geq c \end{cases}$$

### Equation 2. WebJaccard Using Yahoo! Query Hits

One of the common tools for mining semantic relatedness is using WordNet (Fellbaum, 1998). Semantic relationships are found for two words according to their common membership in sets of synonyms, or synsets. Though WordNet has a large dictionary, the tools that rely on it fail when one of the two words is not in the dictionary. There are two common scenarios which increase the likelihood of WordNet failing. The first is that data sets commonly have domain specific terms that are less likely to be in a general dictionary like WordNet. The second problem is that data sets commonly have attribute names that are multiple words and/or use abbreviations. The tools making use of WordNet are not capable of handling either of these cases. In order to overcome these issues, we use tools that query the World Wide Web instead of WordNet.

We use the WWW as a source of information and adapt existing information retrieval measures to use the number of results from queries to compute similarity. Our semantic relatedness relationships are based on work by Bollegala et al. (2007) which queried Google and used the number of query results in computing existing similarity measures, however they only tested its use on single words.

The first step in mining semantic relatedness relationships is to tokenize attribute names. Attribute names are tokenized on occurrences of underscores and capital letters to create a multi-term query. Though not sophisticated these simple rules provide a best effort for creating multi-term queries. The relatedness of two attributes is then found using the WebJaccard measure as expressed in Equation 2, where  $P$  and  $Q$  are the multi-term queries for each attribute name. When available we also include the data set name as a query term to provide sense disambiguation. We use Yahoo as a source for querying because of the open availability of their search API. Yahoo semantic relatedness relationships are filtered to include edges only when the WebJaccard value is above a threshold.

### Mining the External Similarity Matrix

We use existing sources of external information, and therefore only discuss them briefly. External information directly compares attributes in the source and target schemas to look for similar attributes. While mining external similarity both attribute names and values from the data are used. We tested basic sources of external information to investigate the effects of combining internal and external information. Two sources of external similarity were prototyped and tested.

The first source of external similarity is string edit distance, which is a lexical comparison of attribute names. String edit distance represents a method for finding matches that are “low hanging fruit.” We use the jSimlib (<https://jsimlib.dev.java.net/>) library that normalizes string edit distance by the sum of the length of the two strings.

The second source of external similarity is cosine similarity, which is commonly used to compare the similarity of two free text documents. The similarity of the two documents is computed as the cosine value between the term frequency vectors for each document. For attribute-to-attribute schema matching, when the attributes contain text we treat them as documents and create term frequency vectors. The Lucene (<http://lucene.apache.org/java/docs/index.html>) framework was used to calculate the cosine similarity.

We tested three groups of data sets that vary in domain and size which come from the Illinois Semantic Integration Archive (ISIA) at <http://pages.cs.wisc.edu/~anhai/wisc-si-archive/>. The Courses data sets have listings of classes from four different universities, data sets sizes range from twelve to sixteen attributes. The second group of data sets is the Real Estate I (REI) data sets, which includes the homeseekers, nky, windermere, and yahoo data sets. Three of the data sets have sizes in the mid-thirties and the final one is in the sixties. The third group of data sets is the Real Estate Core (REC) data sets. REC data sets are the same as the REI data sets, but only include attributes that have a match in one of the other data sets. This reduced the number of attributes in the data sets to the low twenties, except one having twenty-eight attributes. The REC group is used to test the effects on matching performance when attributes with no matches are removed.



## Validation Experiments

The goals in evaluating AbsMatcher are to look at the performance of internal information by itself and whether the combination of internal and external information provides better cumulative performance. WebJaccard and Entropy internal relationships are meant to provide a baseline ability for schema matching so performance is judged first by whether consistent evidence of an ability to find matches, and second by looking for evidence that combining internal and external information is better than only external information. Finding evidence of these two points would indicate matches being found which internal information can uniquely contribute to finding. Performance is measured using recall. Many schema matching systems provide statistical matches, as opposed to absolute matching, so we present recall for correct matches made and for the correct match being one of the top 3 best matches. This more liberal scoring criterion provides information on whether AbsMatcher has partial information that could be leveraged by future improvements to the algorithm or information sources. Precision is not included because currently AbsMatcher returns a match for each attribute in the smaller of the two schemas. This means that the number of matches returned for a pair of schemas will remain constant no matter what other parameters change. This point is discussed further in future work.

For the initial tests, we first explored schema matching using only the previously described internal relationships, in three combinations. The Entropy combination includes attribute entropy, data set key, and dependency relationships. The WebJaccard results consist of semantic relatedness relationships based on Yahoo results. Finally, the “All” combination includes both Entropy and WebJaccard relationships.

We first look at the extent to which schemas can be matched using only the mined graphs for the two data sets. When using only this limited source of information a high level of performance cannot be expected. However, this limitation is useful in making an initial judgment of whether mined graphs contain useful information. For each group of data sets we select the best performing parameters and present the results in Figure 1 for all three combinations of internal relationships and all three groups of data sets.

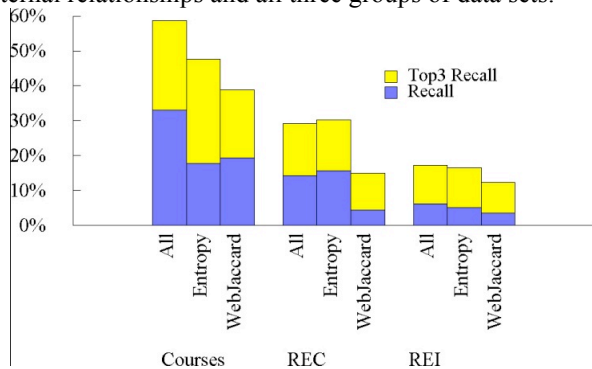


Figure 1. Data sets by types of internal relationships

The first result to examine is AbsMatcher’s ability to find correct matches. Though the results in Figure 1 are relatively low in the context of overall performance of schema matching systems, the more appropriate context is as a source of matches which would be used in a broader system. In this context WebJaccard and Entropy relationships do show consistent ability to find at least some matches. The performance of the top 3 correspondences improves over just correct matches indicating that AbsMatcher can provide supporting evidence which would affirm or discredit correspondences from other candidate matchers. As seen in Figure 1 the top 3 correspondences can provide useful results on a third to half of all matches. The top 3 matches can be useful when considering that the weights of correspondences in the top 3 can often be very close.

The second result to examine is what sources or combinations of sources of internal relationships are the most effective. Neither the Entropy nor WebJaccard relationships were consistently the best between the different data set groups. Though neither was consistently the best, the positive result is that when combined in *All*, performance improved or matched the performance of the best performing source of internal relationships. The fact that adding sources of internal relationships does not degrade performance strengthens the potential that when other existing forms of internal relationships are added, performance could be improved.

For the second set of tests, we combined both internal and external sources of information. For some matches the information which best indicates the correct match is derived by comparing an attribute from each data set. In ABSURDIST this means the use of external information that is combined with internal information using Equation 1. In Equation 1 there are two weighting coefficients,  $\alpha$  and  $\beta$ , which determine the balance between external and internal information. The  $\alpha:\beta$  ratio represents the comparative weights of external:internal information. We tested AbsMatcher with different ratios, where each represented a different balance between external and internal information. Figure 2 presents results for a representative three of those ratios. The 0:1 data point represents using only internal information, which corresponds with the results in Figure 1. The 1:0 data point represents only using external information. The 3:1 data point tested the effort to combine the use of internal and external information. The goal in this evaluation is to determine whether combining internal and external information has a benefit over just using external similarity.

Figure 2 provides evidence that combining internal and external information can for some data sets provide better results than either one in isolation. Though the improvement for Courses and REC data sets is small the fact that it occurs for both supports the claim that internal structure can improve matching performance. It must be remembered that the results for Courses and REC represent the average performance across twelve different pairs of

data sets matched. The ability of internal structure to find correct matches and the additional beneficial effect that it can have when combined with external similarity indicates that internal structure is to some extent finding both unique and useful information for schema matching.

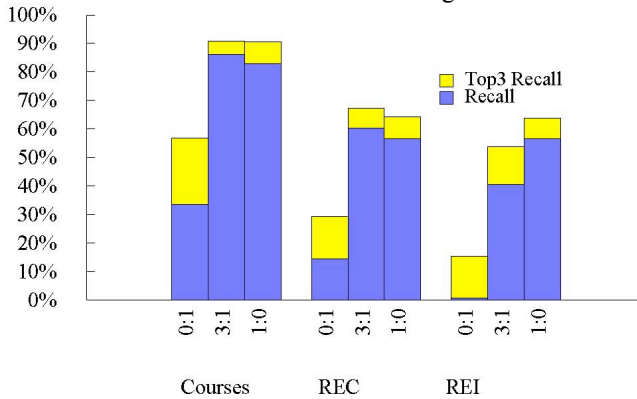


Figure 2. Data Sets by Ext:Int Ratio

The REI data sets do not benefit from internal information. This could in part be due to the fact that REI alignments leave more attributes unmatched. The REC data sets are versions of the REI data sets where attributes with no matches removed. They average 15.5 correct matches between a pair of data sets, meaning that on average half of the attributes in a data set for REI are not being matched, yet information is still mined for them. Courses and REC have a different scale yet both show similar trends in the ability to find correct matches. The only difference between REC and REI data sets is the existence of unmatched attributes, so the difference in performance can be unambiguously attributed to this. This indicates that information which indicates invalid matches could be an important feature to add to AbsMatcher.

## Conclusions

The goal in developing AbsMatcher was to create a schema matching system that used a graph based approach, but was not reliant on a specific data model as a source of information. To this end, we propose Entropy and WebJaccard relationships which can be used even when more descriptive metadata, such as XML or metadata from a relational database, is unavailable. Additionally, these relationships emphasize non-structural relationships in an effort to create graphs which are more conceptual in nature. We then tested these graphs using the ABSURDIST graph matching system. ABSURDIST is ideally suited because of its ability to accept graphs with a wide variety of forms (weighted, unweighted, directed, undirected, labeled, and unlabeled) and ABSURDIST was designed specifically with the idea of combining internal and external information together.

The goals in testing AbsMatcher were to look at whether Entropy and WebJaccard relationships are useful for schema matching on their own and whether they have benefits when

combined with external similarity. Experiments demonstrated that to varying extents the tested relationships are able to accomplish both of the goals. The results presented in this paper were aggregated over multiple individual experiments. The additive benefit of our sources of internal structure is important because it argues that internal structure holds unique information for finding correspondences.

These results were based on aggregating results from a number of matching pairs. It is important to note that there were outliers on both the positive and negative side. This is a common problem in schema matching, where sources of information perform well in certain scenarios and poorly in others. It is this point which motivated the approach of aggregating many disparate measures of similarity. This leads to the idea that by adding new information sources into AbsMatcher we can improve even beyond the baselines presented in this work.

## Acknowledgments

This research was supported by National Science Foundation REESE grant 0910218, Lockheed Martin, and DARPA.

## References

- Aumuellner, D., Do, H.-H., Massmann, S. and Rahm, E. (2005). Schema and ontology matching with COMA++ *Proceedings of the ACM SIGMOD international conference on Management of data*, ACM, Baltimore, Maryland.
- Bollegala, D., Matsuo, Y. and Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines *Proceedings of the 16th international conference on World Wide Web*, ACM, Banff, Alberta, Canada.
- Dalkilic, M.M. and Roberston, E.L. (2000). Information dependencies *Proceedings of the 19th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ACM, Dallas, Texas, United States.
- Fellbaum, C. (1998). *Wordnet: An Electronic Lexical Database*. Bradford Books.
- Feng, Y., Goldstone, R.L. and Menkov, V. (2004). ABSURDIST II: A Graph Matching Algorithm and its Application to Conceptual System Translation *Proceedings of the 17th International Florida Artificial Intelligence Research Symposium Conference (FLAIRS)*, AAAI Press, Miami Beach, Fla., USA, 640-645.
- Giunchiglia, F. and Shvaiko, P. Semantic Matching (2003). *Knowledge Engineering Review*, 18 (3). 265-280.
- Goldstone, R.L. and Rogosky, B.J. (2002). Using Relations within Conceptual Systems to Translate across Conceptual Systems, *Cognition*, 295-320.
- Madhavan, J., Bernstein, P.A. and Rahm, E. (2001). Generic Schema Matching with Cupid *VLDB*.
- Melnik, S., Garcia-Molina, H. and Rahm, E. (2002). Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching *ICDE*.
- Shvaiko, P. and Euzenat, J. (2005). A Survey of Schema-Based Matching Approaches *Journal on Data Semantics IV*, 3730. 146-171.



# Lexical Redundancy, Naming Game and Self-constrained Synonymy

**Kerem Eryilmaz (kerem@ii.metu.edu.tr)**

Cognitive Science, Middle East Technical University (METU), Ankara 06800, Turkey

**Cem Bozsahin (bozsahin@metu.edu.tr)**

Cognitive Science, METU, Ankara 06800, Turkey

## Abstract

Language games are tools to model some aspects of the social aspects of language and communication. Our approach aims to cover the ground between the elementary naming game and the complex models for social use, for the growth of possibly redundant community and personal lexicons. It uses weighted lists of words for the personal lexicon, probabilistic choice as a selection mechanism and lateral inhibition as the weight update scheme. The results demonstrate that the model is a generalization of the elementary naming game, and it provides a good picture of how big a lexicon agents use for the task, how this size can be controlled using the model parameters and a possible way of explaining how synonymy is kept under control.

**Keywords:** emergence; stochastic naming game; synonymy; language game; semiotic dynamics

## Introduction

Language is a social phenomenon. Although there are uses to language that are not social, and there are modes of social interaction other than language use, a very clear and salient function of language is collaboration and communication among individuals in a population. However, computational studies on language until mid-90s had been more focused on modelling the individual capacities in language acquisition and performance, and did not provide much insight into what language might look like from the viewpoint of a population of language users.

A new generation of research addresses this problem by investigating language as it is produced, used and propagated in a community of language users. A fruitful approach has been to see language as a complex adaptive system. What this basically implies is that only at the population level can we adequately characterize language use, and only by taking into account that language is constantly reshaped by its users. How an individual uses language and how the population generally uses language can and do affect one another. In short, the claim is that some features of language do not stem directly from linguistic or cognitive capacity but from a society of interacting agents. One such feature is the lexicon and its emergence as a common vocabulary of a community.

This new branch of research proved to be quite fertile and produced a large body of work. One path of this research is exemplified by Kirby, who approaches language as “the result of an interaction between three complex adaptive systems that operate on different timescales: the timescale of biological evolution (phylogeny), the timescale of individual learning (ontogeny) and the timescale of language change (glotogeny)” (Kirby, 2002).

Another avenue is that of Steels, who also sees language as a complex adaptive system (Steels, 2000). His work started as an investigation of semiotic dynamics in simple language games played among artificial autonomous agents, and eventually led to an elaborate construction grammar formalism called Fluid Construction Grammar (FCG) for simulating a population of language users who bootstrap their own language (Steels & De Beule, 2006).

Steels’s original models were kept elementary to make them easier to analyze. FCG, however, is quite comprehensive. We feel that there is much to be discovered at various levels of complexity, from the trivial, original “naming games” to the full-blown grammatical complexity of FCG. For example, one feature, the negotiation for names of objects, remains simple: if two agents agree on a name, they agree to adopt that one and discard the alternatives, therefore their lexicons are nonredundant if and when convergence arises. Since complex adaptive systems are very unpredictable in terms of how things might change if we change the structure of interactions, we think it is important that we explore much of the degrees of freedom as we can, trying to characterize the emerging structures and phenomena at each level of complexity added.

In that spirit this paper presents an investigation of a naming game among artificial agents with a slightly more complex, weight-based lexicon scheme rather than the original, and a consequent probabilistic word selection scheme. Our focus here is on the tension between lexicon growth and synonymy in the population. The paper first describes the original naming game. Our extension is described next. Our experiments and results are then presented.

## The Original Naming Game

The naming game focuses on vocabulary formation and agreement in a population. The agents try to bootstrap a vocabulary of (proper) nouns that they associate with the objects they try to name (De Vylder & Tuyls, 2006). It is assumed that the agents already know how to send and receive signals, and possess the motivation to do so. It is further assumed that the objects are uniquely identifiable by all agents, so feature sets as in the discrimination game are not employed.<sup>1</sup> The final assumption is that the agents have an independent channel of communication (such as pointing) through which they can reveal their word-object pairings to each other.

<sup>1</sup>Note that it is perfectly possible to combine the two games, as done by Steels (2003).

Formally, the game involves a set of objects  $O = \{o_1, \dots, o_m\}$ , and a set of agents  $A = \{\alpha_1, \dots, \alpha_n\}$ . Each agent  $\alpha_j$  possesses a lexicon  $L_{\alpha_j} = \{E_{\alpha_j,1}, \dots, E_{\alpha_j,k}\}$  where  $k \leq m$ , and each entry in the lexicon consists of a list of words associated with the object  $o_i$ ,  $E_{\alpha_j,i} = \{w_{i,1}, \dots, w_{i,q}\}$ . All agents also possess a function that maps from the global set of objects to the agent's repertoire of words ( $\phi_a : O \mapsto E_{\alpha_j}$ ).<sup>2</sup> Therefore, an agent is characterized only by its lexicon. A game consists of these steps:

1. Two agents are chosen, one as the hearer ( $\alpha_h$ ) and one as the speaker ( $\alpha_s$ ).
2. The speaker chooses an object ( $o_i$ ) to refer to, and points to it (i.e. makes his choice explicit without using the system we try to bootstrap)
3. The speaker chooses a word in the lexicon for the object ( $\phi(o_i)$ ), or creates one if necessary.
4. The hearer tries to decode the word into the object being referred to ( $w_{k,j} \in E_{h,k}$ ), and uses the independent channel to signal this to the other (e.g. points to it).
5. The speaker agent assesses if the response complies with its lexicon, and makes the hearer aware of this assessment. Agents update their lexicons accordingly.<sup>3</sup>
6. If all agents have identical lexicons, the game stops.

Each game results in success ( $\exists w_{i,n} \mid w_{i,n} = \phi_s(o_i); w_{i,n} \in E_{h,k}$ ) or failure ( $\nexists w_{i,n} \mid w_{i,n} = \phi_s(o_i); w_{i,n} \in E_{h,k}$ ). Upon success, both agents purge their entries for that object of all but the successful word. Otherwise, the hearer adds the new word to its entry of the object, or creates one if necessary. There is no intermediary between a word exchange being successful and a word dominating an agent's lexical inventory for an object; it operates on an all-or-nothing basis.

## The Stochastic Naming Game

The current proposal has four key differences from the original model:

1. **Lexical entries:** The lexical entries in the original model are simple lists of words. The proposed model implements lexical entries as *weighted* lists of words, updated upon interactions. This allows a graded behaviour in which words are preferred, and constitutes a more realistic situation in which convergence should be achieved, compared to plain lists.

More formally, for each agent  $\alpha_i$ , an additional value function  $\theta_{\alpha_i}$  is added to retrieve the weight:

$$\theta_{\alpha_i} : w_{k,q} \mapsto \mathbb{R}$$

<sup>2</sup>The function  $\phi_a$  returns  $E_a$  which is a list of *all* the words used for an object from the perspective on agent  $a$ , and not the most successful word.

<sup>3</sup>Note that this assessment does not require global knowledge; it is a local decision determined solely by agent's lexicon and its assessment of the interaction with another agent.

Basically, this is just a lookup for the weight associated with the word in an agent's lexicon. By adding this, the characterization of an agent is a tuple of the lexicon and  $\alpha_i$ , instead of only the lexicon as in the original game:

$$\alpha_i : \langle L_{\alpha_i}, \theta_{\alpha_i} \rangle$$

2. **Word selection:** The word selection scheme in the original model simply picks a word from the set of words present in the lexicon. It does not specify how to pick the word. Although there are some suggestions for schemes that optimize convergence (Baronchelli, Dall'Asta, Barrat, & Loreto, 2005), there is no set practice. The proposed model has a specific scheme that makes a weighted, probabilistic choice of the word to use at each round. This introduces a number of advantages. First, it introduces some noise by not guaranteeing the leading word to be chosen at every round. Some level of noise is often beneficial to convergence in dynamical systems. Second, it is a more realistic scheme, especially when top words have similar scores. Third, it makes the system more fault tolerant by minimizing the impact of successful rounds caused by words that are ultimately going to fail and of unsuccessful rounds caused by words that are ultimately going to succeed.

Formally, the function  $\phi_{\alpha_i}$  is changed to return a word by a weighted random choice. To this end, we first define a probability distribution  $P$  where:

$$P(w_{k,q}) = \frac{\theta_{\alpha_i}(w_{k,q})}{\sum_y \theta_{\alpha_i}(w_{k,y})} \quad (1)$$

Subsequently, the word  $\phi_{\alpha_i}$ 's returns can be characterized as a random variable  $X$  with distribution  $P$ .

$$\phi_{\alpha_i}(o_k) = X \quad (2)$$

for which:

$$X \sim P; X \in E_{\alpha_i,k} \quad (3)$$

3. **Parameters:** As a consequence of the update scheme, there are more parameters in this version of the game than the original one. In particular, three  $\delta$ -values ( $\delta_{\text{success}}$ ,  $\delta_{\text{failure}}$  and  $\delta_{\text{inhibition}}$ ) are added for use in updating the lexicon, whose precise roles are elaborated in the following paragraphs. Additionally, two  $\theta$ -values ( $\theta_{\text{max}}$  and  $\theta_{\text{min}}$ ) are added as the maximum and minimum values for any score in the lexicon.
4. **Update scheme:** The agents in the proposed model no longer discard the competing synonyms (i.e. the other words in the lexical entry) upon a successful interaction. Instead, the agents update the weights of their lexical entries for the object upon every interaction.

A function  $\omega$  is added to each agent which returns a new, updated weight function after an interaction:

$$\omega : \theta_{\alpha_i} \mapsto \theta'_{\alpha_i}$$

This function  $\omega$  adds or subtracts from scores some predefined  $\delta_{\text{success}}$ ,  $\delta_{\text{failure}}$  and  $\delta_{\text{inhibition}}$ , based on lateral inhibition (Lenaerts, Jansen, Tuyls, & De Vylder, 2005). Upon a successful interaction with word  $w_{k,p}$ , this function returns a new function  $\theta'_{\alpha_i}$  and optionally modifies the lexicon of the agent. The modification is that if the resulting score for a word is less than a predefined value  $\theta_{\min}$ , that word is removed from the lexical item for that object. Also, there is a set limit  $\theta_{\max}$  on how large the weight may grow, at which point no weight is added. More formally,  $\omega$  returns the following upon success:

$$\theta'_{\alpha_i}(w_{k,q}) = \begin{cases} \min(\theta_{\alpha_i}(w_{k,q}) + \delta_{\text{success}}, \theta_{\max}) & \text{if } q = p \\ \theta_{\alpha_i}(w_{k,q}) - \delta_{\text{inhibition}} & \text{if } q \neq p \end{cases} \quad (4)$$

where

$$\omega(\theta_{\alpha_i})(w_{k,q}) = \theta'_{\alpha_i}(w_{k,q}) \quad (5)$$

and the following upon failure:

$$\theta'_{\alpha_i}(w_{k,q}) = \begin{cases} \theta_{\alpha_i}(w_{k,q}) - \delta_{\text{failure}} & \text{if } q = p \end{cases} \quad (6)$$

where

$$\omega(\theta_{\alpha_i})(w_{k,q}) = \theta'_{\alpha_i}(w_{k,q}) \quad (7)$$

It then modifies the lexicon as follows:

$$L'_a = (L_a / E_{\alpha_i,k}) \cup E'_{\alpha_i,k} \quad (8)$$

where

$$E'_{\alpha_i,k} = \{w | \theta'_{\alpha_i}(w) \geq \theta_{\min}; \forall w \in E_{\alpha_i,k}\} \quad (9)$$

With this scheme, it is possible to mimick the original model of Steels by using  $\delta_{\text{success}}$ ,  $\delta_{\text{failure}}$  and  $\delta_{\text{inhibition}}$ , values of 10.0, 0.0, 10.0 respectively. Informally, this makes sure that success always maximizes the weight of a word and always eliminates other synonyms, and that failure does not have an impact on the weights. This, in effect, is the behaviour of the original model.

## Methodology

Each parameter set, that is, a tuple of  $(\delta_{\text{success}}, \delta_{\text{failure}}, \delta_{\text{inhibition}})$  was considered a unique case, and the simulation was run 50 times for each case, using 50 agents and 2 objects.<sup>4</sup> The model is considered to have reached convergence when there is 100% success over a success window of 100 rounds or when it reaches the limit for number of rounds,

<sup>4</sup>Originally, up to 5 objects were going to be tested but as far as we could tell from the pilots, this did not provide any interesting insights that 2 objects did not provide for our intentions. Since more objects dramatically increased the simulation time and analysis, the simulations were run using just 2 objects.

which is chosen as 500,000 for this study. This is the product of the number of agents and the number of objects, making it very likely that all agents will have taken part in at least one interaction regarding each object, making the success window more meaningful. Also, note that our concept of “convergence” is different from what is used in the literature. It does not mean that all lexicons are identical, it simply means that all lexicons are “similar enough” for the intended purpose, “similar enough” defined as above. This way, it is possible to see if synonyms can exist at the point where the success of any given communication is almost certain.

The model was run with various  $\delta$  parameters. The method of choosing them was fixing a set of ratios in the form  $\delta_{\text{failure}}:\delta_{\text{success}}$  and  $\delta_{\text{inhibition}}:\delta_{\text{failure}}$ , and then producing the actual  $\delta$  values by choosing a value for  $\delta_{\text{success}}$  and calculating the rest using that chosen value.

There were five values for  $\delta_{\text{success}}$  denoted by the set  $\{1.0, 3.0, 5.0, 8.0, 10.0\}$ . For the ratio  $\delta_{\text{failure}}:\delta_{\text{success}}$ , the ratios picked were  $0.0:1.0$ ,  $0.5:1.0$ ,  $1.0:1.0$ ,  $1.5:1.0$  and  $2.0:1.0$ . The ratios used for  $\delta_{\text{inhibition}}:\delta_{\text{failure}}$  were  $0.0:1.0$ ,  $0.5:1.0$ ,  $1.0:1.0$  and  $1.5:1.0$ . If both  $\delta_{\text{failure}}$  and  $\delta_{\text{inhibition}}$  are 0.0 for a case, it is not possible to calculate  $\delta_{\text{inhibition}}$  from the ratio, so for those cases  $\delta_{\text{inhibition}}$  was set to  $\delta_{\min}$  to provide some negative feedback to the model so that it can converge.

The cases in which  $\delta_{\text{inhibition}} > 1.25 \times \delta_{\text{failure}}$  are excluded since this corresponds to the vicinity of the original model and our aim is at exploring different areas of the parameter space. Only the case represented by the tuple (10.0, 0.0, 10.0) is included for comparison since it exactly corresponds to the behaviour of the original model.

The values  $\delta_{\max}$  and  $\delta_{\min}$  were fixed at 10.0 and 0.1, respectively.

## Results

The results make it clear that choosing the original model parameters is not the only viable option for our model. In fact, of the total of 70 parameter sets, only 16 performed worse than the original model parameters in terms of time of convergence.

In the following, we will present the results on how model parameters interact. We are not going to present all the results because of space considerations. The analysis will be made both in terms of time of convergence, relative convergence rate and lexicon size. Relative convergence rate is defined as the time it takes for the system to converge once the average size of the lexicons are maximized (this time point of maximum lexicon size is represented as  $t_{\max}$  in the simulation). Time of convergence is the total number of rounds from the start of the simulation to converge. Lexicon size is the total number of words in an agent’s lexicon.

The results confirm that  $\delta_{\text{inhibition}}$  functions as a way to shrink the lexicon to have a greater relative convergence rate, that is, to reduce the time it takes for the convergence to be reached once  $t_{\max}$  is reached. The functions of  $\delta_{\text{success}}$  and  $\delta_{\text{failure}}$  are straightforward as positive and negative feedback,

respectively.

### Interaction of $\delta_{\text{success}}$ and $\delta_{\text{failure}}$

The interaction of  $\delta_{\text{success}}$  and  $\delta_{\text{failure}}$  is easier to explain. These are directly counteracting forces, and therefore if  $\delta_{\text{failure}}$  is greater than  $\delta_{\text{success}}$ , the system fails to converge save a few exceptional cases. This is not surprising since the system, especially before  $t_{\text{max}}$ , mostly learns by failing the exchanges therefore learning new words. If the impact of failure is higher than that of success, then it becomes really difficult to disseminate some words to all agents so that later they can converge to that word. Basically, they all get discarded before their commonality causes them to become successful across many interactions.

This effect is further compounded by non-zero  $\delta_{\text{inhibition}}$  values, which, upon occasional initial success, acts effectively as a failure penalty for all non-successful words, further decreasing the number of common words. However, there are a few cases where models with  $\delta_{\text{success}} \leq \delta_{\text{failure}}$  actually converge to a common vocabulary. In a closer look, there are two conditions under which convergence occurs. One is when all  $\delta$  values are quite small, so that success can accumulate to save at least a couple of alternatives for a word from being discarded. In contrast, larger values mean that one or two failures guarantee discarding of a word, and the number of successes have little impact on this since the scores stop increasing once they hit the upper limit of  $\theta_{\text{max}}$ . In other words, small  $\delta$  values give the agents some room to keep a greater amount of interaction history, and therefore they become better at evolving their lexicon in tandem with the population trends.

The other case is where  $\delta_{\text{inhibition}} = 0$ . This leaves only  $\delta_{\text{success}}$  and  $\delta_{\text{failure}}$  to battle each other, and occasionally leaves room for convergence unless  $\delta_{\text{failure}} < 1.25 \times \delta_{\text{success}}$  or  $\delta_{\text{failure}} > 0.75 \times \theta_{\text{max}}$ . This is equivalent to saying that  $\delta_{\text{failure}}$  should leave room for at least two failures until a previously successful word is discarded, i.e. the word is not discarded right away upon failure. Since there is no other mechanism to decrease the weights, this allows some dissemination with a bit of luck.

### Interplay of $\delta_{\text{failure}}$ and $\delta_{\text{inhibition}}$

The key to the relationship between  $\delta_{\text{failure}}$  and  $\delta_{\text{inhibition}}$  is that they can replace one another with slightly different effects.  $\delta_{\text{inhibition}}$  is a stronger form of  $\delta_{\text{failure}}$  which affects not only one but almost all (save the successful one for the round) words associated with an object each round. In fact, this is why the original model, with the parameter set (10.0,0.0,10.0), can function without  $\delta_{\text{failure}}$ .

This great impact of  $\delta_{\text{inhibition}}$  effectively shrinks the lexicon, and this makes the  $t_{\text{max}}$  smaller but also reduces the relative convergence rate. The reason is that at  $t_{\text{max}}$  there are less alternatives that the system may converge to for an object. This means that any perturbation in the system, such as an interaction where the speaker agent prefers a not-to-be-successful word, needs to be counteracted so that the sys-

tem returns to moving towards the word it originally had been converging to.

In contrast, a big lexicon at  $t_{\text{max}}$  means there are many alternatives, and a disturbance need not be fully counteracted; the system might just converge to another word-object pairing that is salient in the population. Accordingly, our results show that a relatively large  $\delta_{\text{failure}}$  combined with a small  $\delta_{\text{inhibition}}$  produces the fastest convergences, with a large lexicon at  $t_{\text{max}}$  since  $\delta_{\text{inhibition}}$  does not get a chance to shrink the lexicons as much. After that, applications of  $\delta_{\text{inhibition}}$  mostly help convergence to the pairings that will ultimately dominate.

### Lexicon Size

Lexicon size can be used as an indicator of game dynamics. Since it is not a parameter but a quantity that manifests itself during the game, it is difficult to test. Nonetheless, there are some clear tendencies that are important.

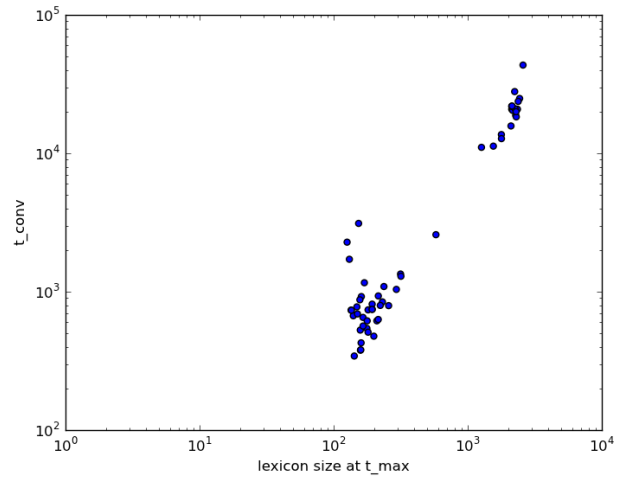


Figure 1: A log-log plot of time of convergence versus average lexicon size for all convergent parameter sets.

The first such tendency is between the time of convergence and average lexicon size at  $t_{\text{max}}$ . It turns out that there is a direct proportion between the logarithm of average lexicon size and the logarithm of the time of convergence. Roughly, this means the bigger the lexicon size at  $t_{\text{max}}$ , the longer the convergence takes. This indicates that the time of convergence and relative convergence rate are not necessarily correlated. Although parameter sets that produce bigger lexicons converge fast after  $t_{\text{max}}$ , they also take longer to put together (i.e. to reach  $t_{\text{max}}$ ) and therefore do not necessarily represent the sets that allow convergence in the minimum number of rounds.

The second tendency is best represented by a plot of rounds (i.e. the time series) versus average lexicon size (see Figure 2). The plots fork into two fairly distinct groups, and all of the plots with bigger lexicon size belong to parameter sets in

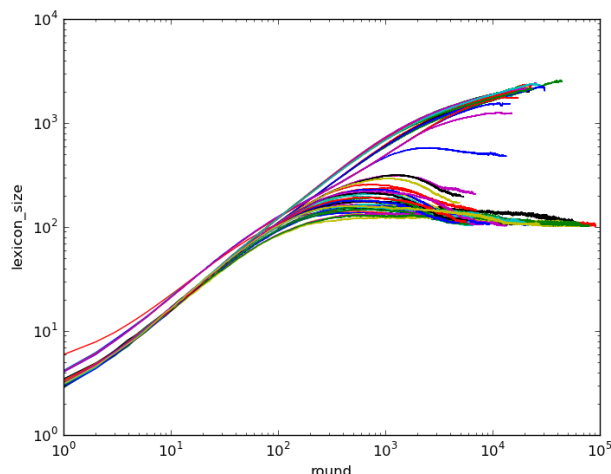


Figure 2: A log-log plot of time of convergence versus average lexicon size for all convergent parameter sets.

which  $\delta_{\text{inhibition}} \leq 0.1$ . This demonstrates that  $\delta_{\text{inhibition}}$  has the function of shrinking the lexicon. In the figure, it is apparent that for all cases, there is a peak (whose position in time we call  $t_{\text{max}}$ ) after which convergence is achieved (which is also a finding encountered in the original naming game literature). For very small values of  $\delta_{\text{inhibition}}$ , this peak is more or less identical to the time point at which convergence is achieved. In other words,  $\delta_{\text{inhibition}}$  shrunk the lexicon so little that the system converged with a larger average lexicon size. In other words, on average, at least some of the objects have more than one alternative word associated with them although the success rate climbed to 100% over a window of 100 rounds. This may hold an insight into how synonymy may be controlled in human languages.<sup>5</sup> In our model, synonymy does not necessarily hinder mutual communication. In more complex models, if they have analogous tendencies at all, it might be beneficial for large populations to be able to communicate very successfully without a full agreement on which object is referred to by which word.

This also suggests that memory requirement for this task is not necessarily determined by the amount of memory available to the agents but by the type of learning a population adopts (or is born with), and the fact that the task is undertaken by a population. Even if they had very big memory capacities (i.e. more than they need to use), part of this capacity would be redundant because of the decay in weights of the unsuccessful or non-successful words would get them discarded from the lexicon.<sup>6</sup> However, parameter sets with small  $\delta_{\text{inhibition}}$  (similar to those in the top partition of Figure

<sup>5</sup>Of course, these are results from a single study with very strict and narrow boundaries in terms of what it deems possible “language”; hence the verb *inspiration* instead of *insight*.

<sup>6</sup>Non-successful words are those whose weights are decremented not because they failed, as unsuccessful words’ are, but because some other word succeeded and  $\delta_{\text{inhibition}}$  was nonzero.

2) can conceivably make use of more memory than available here. This indicates that  $\delta_{\text{inhibition}}$  is a way of controlling the memory use of the population for the task.

## Conclusion

We believe that this line of research is quite relevant to cognitive science and to study of linguistics in general. Although the communicative capacities of our agents are very limited in that they do not contain any capabilities for syntax, discourse etc., our results are somewhat reminiscent of Elman’s work on importance of starting with a small working memory in language acquisition (1993). Our model shows that using a reasonable  $\delta_{\text{inhibition}}$  actually facilitates learning although it does reduce the amount of memory used for the task. In our model, it is not even about the availability of more memory; the memory simply does not need to be larger.

This would not make sense for learning static knowledge where having as many samples as possible at any point in time is the goal. But such a finding is arguably not as surprising for learning population-generated knowledge which may change from one point in time to the other. The memory constraint comes not from the learners but from the nature of what is learned, how it changes and the relationship between the learners and what is learned. The constraint is not on any given specific lexicon but on the “lexicon of the population”. In other words, this effect does not really stem from the agents themselves but from the fact that what they are trying to agree on is malleable by this very process of negotiation.

It also confirms the previous findings in the semiotic dynamics literature that the sudden popularity of a word upon coinage is an illusion. The conclusion from this literature, including this study, is that there are often many competing synonyms for the same concept, and, after a word hits a popularity threshold, the system falls into a spiral of making the most popular word even more popular. This is even true in a model such as ours where there is a net negative feedback,  $\delta_{\text{failure}}$ , and varying ratios of the strength of other mechanisms of pressure, unlike other models in the literature. There are grey areas where words are neither in disuse nor popular, and they need not become either unused or popular for convergence to occur.

Finally, these simulations can give us an idea about controlling synonymy. Our model does not necessarily have much pressure towards eliminating synonymy; it just works by looking at how successful communications are. The only pressure on synonymy is controlled by  $\delta_{\text{inhibition}}$ . Such inhibition effectively means that success of one word for an object is the failure of others for that same object. However, unlike simpler models in the literature, synonymy can be preserved while still achieving very high communicative success, though only when using low  $\delta_{\text{inhibition}}$  values and with the side effect of having a larger vocabulary to hold the synonyms. The fact that the model also converges with multiple synonyms for an object both when  $\delta_{\text{inhibition}} \neq 0$  and  $\delta_{\text{inhibition}} = 0$  demonstrates that this ability to maintain syn-

onymy is not about the existence but the magnitude of the pressure towards no synonymy, as exerted by  $\delta_{\text{inhibition}}$ .

The main lesson to learn from this is that synonyms are eliminated only if they seriously hinder communicative success, and they are eliminated more or less globally when they do. It would be more feasible to use synonyms if the agents had contextual cues as humans do to constrain the search space, but synonymy can be preserved even without this additional information. If the ambiguity they bring is manageable within the task we are trying to achieve with our language (and these models demonstrate that there really exist such tasks), they need not be eliminated. This is analogous to the difference between colloquial text and legislation in terms of ambiguity, which arguably has something to do with what the population of language users intend to use the language for. What we are trying to achieve in such tasks is not necessarily identical lexicons but communicative success, and what this means has to be defined on a per-case basis.

Figure 2 shows that we pay the price for synonymy if we let it loose: it causes either late convergence or no convergence. The natural equivalent of these results are open to discussion. Our suggestion is that they seem to indicate keeping it low among a community while not completely eliminating it seems to be a way of balancing expressivity and communicative success. If either aspect begins to dominate, it will be equivalent to widespread synonymy and no synonymy, respectively. The conclusion that these may be counteracting forces are shown by synonym's relation to number of iterations, rather than assumed cognitively.

## References

- Baronchelli, A., Dall'Asta, L., Barrat, A., & Loreto, V. (2005). Strategies for fast convergence in semiotic dynamics. *Word Journal Of The International Linguistic Association*, 6.
- De Vylder, B., & Tuyls, K. (2006). How to reach linguistic consensus: A proof of convergence for the naming game. *Journal of theoretical biology*, 242(4), 818–831.
- Elman, J. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*.
- Kirby, S. (2002, April). Natural Language From Artificial Life. *Artificial Life*, 8(2), 185–215.
- Lenaerts, T., Jansen, B., Tuyls, K., & De Vylder, B. (2005). The evolutionary language game: An orthogonal approach. *Journal of Theoretical Biology*, 235(4), 566–582.
- Steels, L. (2000, December). language as a complex adaptive system. In M. Schoenauer et al. (Eds.), *Parallel problem solving from nature-ppsn vi* (Vol. 6).
- Steels, L. (2003, July). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, 7(7), 308–312.
- Steels, L., & De Beule, J. (2006). A (very) brief introduction to fluid construction grammar. In *Proceedings of the third workshop on scalable natural language understanding scanlu 06* (p. 73). Association for Computational Linguistics.

# Learning Unattested Languages

Sara Finley (finleys@elmhurst.edu)

Department of Psychology, Elmhurst College  
Elmhurst, IL, 60126, USA

## Abstract

This paper demonstrates the role of morphological alternations in learning novel phonotactic patterns. In an artificial grammar learning task, adult learners were exposed to a phonotactic pattern in which the first and last consonant agreed in voicing. Long-distance phonotactics encoded as strictly piecewise languages suggest that first-last phonotactic patterns should be unattested in natural language. However, recent theories of morphologically induced phonological patterns predict that long-distance agreement between the first and last consonant of a word can occur when the agreement is induced as a morphological alternation. The results of two experiments support the prediction that first-last harmony patterns are more easily learned when morphological cues to the pattern are present. Participants only learned the first-last pattern when presented as a morphological alternation.

**Keywords:** statistical learning, phonotactics, morphology.

## Introduction

One of the major goals of generative linguistics is to explain the nature of language in terms of computational constraints on the cognitive capacity for human languages. Computational models of phonotactic patterns work to understand the restrictions that underlie the set of patterns that are possible in natural language and the set of patterns that are not possible. Recent work has argued that phonotactic patterns, defined as constraints on the co-occurrence of different sounds within a word, are subject to a very specific set of computational constraints (Heinz, 2007, 2011a, 2011b). In particular, it has been demonstrated that long-distance phonotactic patterns derived from consonant harmony can be simulated using strictly piecewise languages (Heinz, 2010; Heinz & Rogers, 2010; Rogers et al., 2010), a subset of regular languages.

Computational models of phonotactic patterns raise three important questions for a theory of the cognitive science of language. First, is there a correlation between the patterns that are learnable and the patterns that can be generated by computational models? There is a prediction that any long-distance consonant agreement pattern that does not fall within the set of strictly piecewise languages should not be learnable. Second, do computational models of phonotactic patterns capture the intricacies of generative models of phonological representations? If morphological and syntactic constructions require more complex computational machinery to generate, then there is a question of whether patterns at the interface between phonology and morphology, and phonology and syntax are subject to the same computational constraints as purely phonotactic patterns (Heinz & Idsardi, 2011). Third, is there a way to

reconcile apparent exceptions to general tendencies linguistic typology? Linguistic tendencies are typically proposed as opposed to universals because almost any ‘universal’ has exceptions (Evans & Levinson, 2009).

These questions are particularly salient for a specific, hypothetical phonotactic pattern: first-last agreement (Lai, 2012). For the purposes of this paper, a first-last agreement pattern is any phonotactic pattern in which the first and last segment in a word must agree in terms of some phonological feature. For example, a first-last consonant voicing pattern requires that the first and the last consonant in a word share the same value for the feature [Voice]. In such a pattern, the word [boteg] would be a possible word because the first consonant ([b]) and the last consonant ([g]) are voiced, even though the medial consonant ([t]) is voiceless. However, \*[boget] would not be a possible word because the first consonant ([b]) is voiced and the final consonant ([t]) is voiceless.

First-last patterns are said to be unattested in natural language (Lai, 2012). One possible explanation for the failure to find a true case of first-last agreement is that such patterns can not be generated with a strictly piecewise grammar (Heinz & Rogers, 2010). If long-distance phonotactic patterns must be generated with a strictly piecewise grammar, patterns that fall outside of the cognitive constraints on phonological patterns may not be learnable.

While purely phonotactic first-last patterns have not been described in natural languages, there are some possible cases of first-last agreement patterns when morphology is considered. A morphologically controlled phonological alternation is any sound pattern that occurs only in the presence of a specific morphological environment. For example, the alternation between /o/ and /e/ in ‘goose’ vs. ‘geese’ is induced by the alternation between singular and plural. Such morphologically controlled patterns may manifest as a first-last agreement pattern. For example, in Lokaa, a Benue-Congo language spoken in Nigeria, the future tense is marked with a low tone on the final syllable and a prefix containing a low tone (e.g., [nà-à-fúkà] ‘you will gather’). In this case, the first and last vowels of a word must agree in tone, but only in the future tense. (Iwara, Akinlabi, & Truckenbrodt, 2003). Finley (2009) accounts for this morphological alternation using morpheme-specific constraints that target specific edges of the word. Finley’s analysis suggests that morphologically controlled patterns, also referred to as ‘featural affixation’, are subject to different constraints than purely phonotactic patterns. The possibility that featural affixation can target the first and last element of a word leads to the prediction that long-distance



patterns that cannot be generated with a strictly piecewise model of phonotactics may be generated at the interface between phonology and morphology.

There are three reasons to believe that phonotactic patterns and morphologically controlled phonological patterns are subject to different representational and learning constraints. First, as discussed above, the typological restrictions on morphologically controlled patterns tends to be more open than the restrictions placed on phonotactic patterns (Finley, 2009). Second, infants appear to learn phonotactic patterns earlier than morphologically controlled phonological patterns (Jusczyk, Friederici, Wessels, Svenkerund, & Jusczyk, 1993). Third, Lai (2012) demonstrated that adult learners are worse at learning a first-last consonant agreement pattern than a typical consonant harmony pattern that targeted all relevant segments of the word.

The problem with understanding the difference between morphological and phonotactic patterns in terms of representation and typology is that there are reasons why the typology of phonotactic constraints may be different from the typology of morphologically controlled phonological patterns. For example, the lack of existence of a first-last phonotactic agreement pattern may reflect constraints on phonotactic representations, or it could simply reflect an accidental gap. In addition, phonotactic patterns may be learned faster than morphologically controlled phonological patterns because phonotactic patterns apply to a large range of words, while morphologically controlled patterns only apply to specific morphological environments. In this case, the infant must learn both the phonological pattern, but also the morphological environment.

One possible way to understand the relationship between typological and computational constraints on long-distance phonotactic patterns is to explore the existence of learning biases for long-distance patterns. Previous research suggests that first-last phonotactic agreement patterns may not be learnable (Lai, 2012). While Onnis, Monahan, Richmond, and Chater (2005) showed learning of first-last phonotactics, this pattern was based on syllables, rather than features, and therefore may be subject to different constraints. However, there is a question of whether adults may be able to learn first-last agreement patterns if they are presented as a morphologically controlled phonological alternation. In an artificial grammar learning task, it is possible to compare learners with the same language backgrounds (American English) with two languages that are minimally different (phonotactic first-last agreement vs. morphologically controlled first-last agreement). If morphologically controlled patterns are subject to different constraints on learning and representation, one should expect that in the case of first-last agreement patterns, morphologically controlled patterns should be easier to learn than a phonotactic agreement pattern. This prediction is particularly interesting because it goes against the general findings that phonotactic patterns are learned before morphological patterns. In an artificial grammar learning

paradigm, adult participants were exposed to a first-last agreement pattern that was induced either as a morphological alternation or as a phonotactic pattern. Participants who were exposed to the pattern as a phonotactic pattern did not differ significantly from chance or control participants. This is similar to Lai's (2012) results, which showed that an unattested first-last agreement pattern is less easily learnable than a version of an attested consonant harmony pattern. However in the study reported here, participants were exposed only to the first-last agreement pattern, either presented as a phonotactic constraints or as part of a morphological alternation.

## Experiment 1

### Participants

All participants were adult native speakers of English with no previous exposure to a language involving first-last agreement or consonant harmony. Forty-six University of Rochester undergraduate students and affiliates and were paid \$10 for their participation. Two additional participants were from the Elmhurst College Psychology Department Human Subject Pool, and were given extra course credit for their participation.

### Design

Participants were trained on a first-last voicing agreement pattern via auditory exposure. In this pattern, the first and the last consonants of every word agreed in voicing. All words were of the form CVCVC, where C refers to stop consonants drawn from the set /p, t, k, b, d, g/, and V refers to vowels drawn from the set /i, e, o, u, a/. The first and last consonants were either both voiced /b, d, g/ or both voiceless /p, t, k/, with no restriction on the voicing of the medial consonant.

Participants in the Morphological Training condition were exposed to 24 pairs of CVCVC items (repeated five times each) in which the first CVCVC item contained voiceless stops in the first and last positions and the second item contained voiced consonants in the first and last positions (e.g., /kidat gidad/ and /topak dopag/). Participants in the Morphological Training condition were told that they were listening to a novel language, and that they would hear pairs of words, the first of which was a 'singular' form and the second of which was a 'plural' form. The use of 'singular' and 'plural' labels was designed to create the effect of a morphologically controlled alternation. Because adult English speakers are familiar with the distinction between singular and plural, it was assumed that participants recognized that the pairs of items were morphologically related. There was no other semantic information accompanying the training items.

Participants in the Phonotactic Training condition were exposed to the same 48 words that were presented to the participants in the Morphological Training condition, and were told that they would be listening to words from a novel language. There were two main differences between the

Morphological Training condition and the Phonotactic Training condition, reflecting the two main differences between phonotactic and morphologically controlled phonological patterns. First, participants in the Phonotactic training condition were given no semantic information about the items. Second, items in the Phonotactic Training condition were not presented as pairs of items, but as single words presented in a random order.

The medial consonant varied between voiced and voiceless such that half of the items showed voicing agreement for all consonants, and the other half of exposure items showed voicing agreement only between the first and last consonant. In addition, the distribution of consonants was even, such that each consonant appeared in an equal number of items in both final and initial positions. One third of training items contained identical consonants in both first and last positions. Examples of the training stimuli can be found in Table 1.

Table 1: Example Training Items.

Voiceless	Voiced
<b>kidat</b>	<b>gidad</b>
<b>topak</b>	<b>dopag</b>
<b>pibot</b>	<b>bibod</b>

In the Morphological Training condition, all ‘singular’ words had voiceless consonants in first and last positions, and all ‘plural’ words had voiced consonants in the first and last positions. There are two reasons why this design was chosen (as opposed to adding a suffix that alternated depending on the quality of the first consonant, as in /kida-kidat, dopa-dopad/). First, the present design allows the Morphological Training and the Phonotactic Training conditions to use the exact same set of training items, as opposed to two different sets for each condition. Second, the voiceless-voiced alternation mirrors the morphological harmony patterns described in Finley (2009). For example, Kanembu shows an alternation in which the incomplete form are all [–ATR], while the complete forms are all [+ATR] (Akinlabi, 1996).

Following exposure, all participants were given a two-alternative forced choice test. This test was designed to probe whether participants had learned the agreement pattern. All participants received the same set of 40 test items that contained ten Old Items (items heard in the training set), ten New Items (items not heard in the training set), and 20 filler items that contained the voiceless alternation.

All test items were of the form CVCVC in which the ‘correct’ (harmonic) item contained a voiced consonant in the first and the last position of the word, and the ‘incorrect’ (disharmonic) item contained a voiced segment in the first position and a voiceless consonant in the final position. Examples of test items can be found in Table 2, below.

Participants in both training conditions were given identical instructions for how to complete the test phase.

Participants were told that they would hear two words. One word was from the language they had just heard, and the other word was not from the language they had just heard; if they believed the first word was from the language, they were instructed to press the ‘a’ key; if they believed the second word was from they language, they were instructed to press the ‘l’ key. Participants did not hear pairs of words in the test phase.

Table 2: Example Test Items.

Old Items	
Harmonic	Disharmonic
<b>gidad</b>	<b>gidat</b>
<b>dopag</b>	<b>dopak</b>
New Items	
Harmonic	Disharmonic
<b>bikad</b>	<b>bikat</b>
<b>depod</b>	<b>depot</b>
<b>gutub</b>	<b>gutup</b>

A female native speaker of English produced the spoken materials that were used in the experiment, and had no knowledge of the design or purpose of the experiment. The speaker produced all sounds in a sound-attenuated booth. All bi-syllabic stimuli were produced with stress on the first syllable, but instructions were given to the speaker to pronounce all vowels (as English vowels in unstressed position tend to be reduced). All stimuli items were normalized for intensity (set at 70dB) using Praat (Boersma & Weenink, 2005).

All phases of the experiment were run in Psyscope X (Cohen, MacWhinney, Flatt, & Provost, 1993). Participants were given both written and verbal instructions. The entire experiment took approximately 20 minutes.

### Results

Proportion of correct responses (i.e., choosing the item that contained a voiced stop in the first and last position) for all conditions are given in Figure 1. Responses were compared via a 2x2 mixed design ANOVA. There was a significant effect of Training,  $F(1, 34) = 4.24, p < 0.05$ , in that participants in the Morphological Training condition (mean = 0.63,  $CI \pm 0.084$ ) selected the harmonic option more often than participants in the Phonotactic Training condition (mean = 0.51,  $CI \pm 0.085$ ). There was no effect of Test Item,  $F < 1$ , and no significant interaction,  $F(1, 34) = 2.30, p = 0.14$ .

Responses to Old and New items were compared to 50% chance via Bonferroni corrected one-sample t-tests. There was a significant effect in the Morphological Training condition for both Old Items, with a mean of 0.64,  $CI \pm 0.11, t(17) = 2.75, p < 0.05$ , and New Items, with a mean of 0.62,  $CI \pm 0.083, t(17) = 2.99, p < 0.01$ . This suggests that participants in the Morphological Training condition learned the harmony pattern at a level greater than chance. There

was no significant differences in the Phonotactic Training condition for either Old Items, with a mean of 0.47,  $CI \pm 0.12$ ,  $t(17) = -0.49$ ,  $p = 0.63$ , or New Items, with a mean of 0.54,  $CI \pm 0.082$ ,  $t(17) = 1.14$ ,  $p = 0.27$ . This suggests that participants in the Phonotactic Training condition failed to learn the harmony pattern at a level greater than chance.

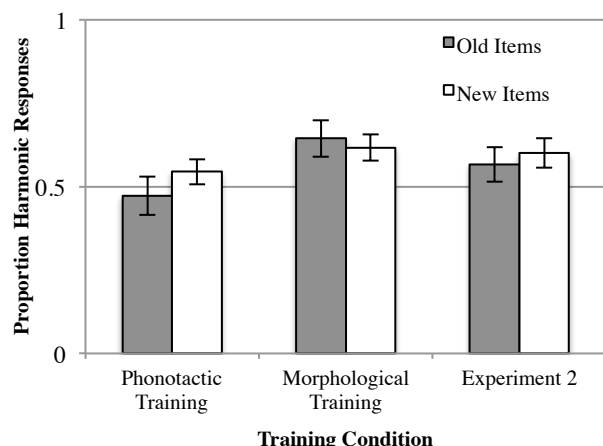


Figure 1: Results.

It is important to note that the success of the participants in the Morphological Training condition was not limited to correct items that were fully harmonic. Participants in the Morphological Training condition selected the correct item 64%, of the time when the medial item was voiceless and the first and last item was voiced (e.g., /beteg/),  $CI \pm 0.11$ ,  $t(17) = 2.86$ ,  $p = 0.011$  (because of the small number of items, Old and New items were combined in this analysis, with a mean of 67% for Old items and 59% for New Items). This analysis rules out the possibility that participants only learned a harmony pattern that required all consonants to share the same feature for voicing.

There was a high amount of individual variation in the present experiment. Three of 18 participants in the Morphological Training condition showed a mean below 50%, while nine of the 18 in participants in the Phonotactic training condition showed a mean lower than 50%. The fact that so many participants in the Phonotactic Training condition scored below chance suggests that these participants were not attending to the relevant aspects of the stimuli. These participants may have simply been ‘guessing’ incorrectly more often than correctly, or they may have inferred a pattern that was not actually present in the data.

The results of the present experiment support the hypothesis that a first-last agreement pattern is more easily learnable as a morphologically controlled phonological alternation than as a phonotactic pattern. Participants in the Morphological Training condition responded correctly to harmonic items at a level greater than chance, and significantly outperformed participants in the Phonotactic Training condition, who failed to learn the first-last agreement pattern.

There were two major differences between the Morphological Training condition and the Phonotactic

Training condition. First, participants in the Morphological Training condition received information about the morphological status of the items in the training (singular-plural pairs). Phonotactic patterns are not morphologically restricted, and morphological information is therefore irrelevant to the phonotactic pattern. Second, the training items in the Morphological Training condition were presented as pairs of words, voiceless followed by voiced. This reflects the fact that morphologically controlled phonological alternations are typically described in terms of an alternation. While both pieces of information are necessary to differentiate between phonotactic patterns and morphologically controlled alternations, it is unclear whether the unattested phonotactic first-last agreement pattern might be learnable if words were simply presented as pairs of words that differed in voicing. Presentation of items as pairs of words may highlight the regularities present in the word, regardless of the morphological status of the pairs of words. Experiment 2 tests whether adult English speakers are able to learn the unattested phonotactic first-last alternation if the items are presented in pairs of voiceless-voiced ‘alternations’ without morphological information.

## Experiment 2

### Participants

All participants were adult native speakers of English with no previous exposure to consonant harmony. All 18 participants were Elmhurst College undergraduate students, recruited from the Elmhurst College Psychology Department Human Subject Pool, and were given extra course credit for their participation.

### Design

Participants in Experiment 2 were given the same exposure items as participants in the Morphological Training condition in Experiment 1. Participants were exposed to 24 pairs of items that reflected an alternation between CVCVC words in which the first and last consonants agreed in voicing; the first word of each pair contained voiceless stops, and the second word of each pair contained the corresponding voiced stops. Unlike Experiment 1, participants in Experiment 2 were not given any information about the morphological status of the pairs of items. Participants were simply told that they would be listening to words from a novel language. They were not told that the items were presented in pairs. Participants in Experiment 2 received the same test items as participants in Experiment 1.

### Results

Proportion of correct responses (i.e., choosing the item that contained the voiced stops in first and last position) were recorded, and are present in Figure 1, above.

Responses to Old and New items were compared to 50% chance via Bonferroni corrected one-sample t-tests. There was no significant effect for Old Items, with a mean of 0.57,

$CI \pm 0.11$ ,  $t(17) = 1.27$ ,  $p = 0.22$ . There was, however, a marginal difference for New Items, with a mean of 0.60,  $CI \pm 0.093$ ,  $t(17) = 2.26$ ,  $p = 0.074$ . These results suggest that participants in Experiment 2 did not reach full criterion for learning, but did show some evidence of learning beyond the chance level.

In addition, a 2x3 ANOVA was performed comparing results for Experiment 1 with results for Experiment 2. There was a marginal effect of Training,  $F(2, 51) = 2.21$ ,  $p = 0.080$ , no effect of Test Item,  $F < 1$ , and no significant interaction  $F(1, 51) = 1.10$ ,  $p = .34$ . Pairwise comparisons revealed no significant differences between Experiment 2, with a mean of 0.58,  $CI \pm 0.084$ , and the Phonotactic Training condition of Experiment 1,  $p = 0.21$ , or the Morphological Training condition of Experiment 1,  $p = 0.42$ . These lack of significant differences suggest that participants in Experiment 2 performed at a level intermediate between that of the Phonotactic Training condition and that of the Morphological Training condition in Experiment 1.

There are potentially many reasons why participants in Experiment 2 did not perform significantly different than either the Morphological Training condition or the Phonotactic Training condition. First, it is possible that some of the learners imposed morphological structure on the pairs of words. In informal debriefing, several participants noted that they had analyzed the pairs of words as being related ‘like singular and plural’. If some learners naturally impose morphological structure on alternating pairs of words, it would suggest that learners use morphological cues when they have the potential to be helpful. Second, it is possible that the presence of cues to a morphologically controlled phonological pattern (alternations and morphological information) provide the best learning environment for the most people. If only one of the cues is present, learning will be intermediate between having both cues and no cues at all. Third, it is possible that the high degree of individual differences across both experiments made finding a significant effect difficult. Of the 18 participants in Experiment 2, five showed means lower than 50%. These individual differences may have been compounded the additional factors that lead to an intermediate result for Experiment 2. Fourth, the lack of a difference may simply reflect a floor effect. It may be difficult to show substantial differences between training conditions, due to the fact that learning in the Morphological Training condition of Experiment 1 was significant, but not highly robust.

## Discussion

The present study explored the role of alternations and morphological information in learning phonotactic patterns. First-last agreement patterns, which fall outside of the strictly piecewise grammars, are predicted to be unlearnable (Heinz 2010; Lai, 2012). However, linguistic analyses that demonstrate the possibility of a morphologically controlled alternation that targets the first and last segments of a word

(Finley, 2009), along with the existence of morphologically controlled first-last agreement patterns, leads to the prediction that morphologically controlled phonological patterns may not be subject to the same constraints on learning and representation as purely phonotactic patterns. This prediction was tested using an artificial grammar learning paradigm in which adult native English speakers were exposed to an artificial first-last agreement pattern that was either presented as a morphological alternation or as a phonotactic alternation. Participants failed to learn the phonotactic pattern, but successfully learned the morphologically controlled phonological alternation. This result suggests that unattested phonotactic patterns may be possible given the right morphological cues.

The results of the present experiment have important consequences for theories of typological linguistic universals. One of the major issues with proposing a linguistic universal is that it is very difficult to interpret potential counter-examples, or a lack of counterexamples (Evans & Levinson, 2009). For example, if there are no cases of first-last agreement patterns in natural language, is it because of a cognitive restriction or because of an accidental gap? In the case of first-last agreement patterns, potential counter-examples can often be ‘explained-away’ in terms of morphological restrictions. Using an artificial grammar learning paradigm, it is possible to tease apart issues of the source of a typological restriction (Nevins, 2009). First, if two patterns that are minimally different except for a predicted restriction on language, there is a clear prediction that one pattern will be learned more easily than the other. For example, Finley (in press) compared learning between minimally different vowel harmony languages. One language had typologically (and phonetically) salient mid vowels as the source for harmony, while the other language had typologically (and phonetically) less salient high vowels. Participants who were exposed to the typologically salient cues were able to learn the harmony pattern, while participants who were exposed to the less salient cues failed to learn the pattern. Second, it may be possible to find explanations for potential counterexamples to proposed linguistic universals. Additional social and cognitive cues may support learning a pattern that falls outside a predicted learning space. Artificial grammar learning experiments provide a mechanism to control for these factors. In the present study, it was demonstrated that the proposed restriction that strictly piecewise languages patterns form part of the cognitive constraints on phonological grammars may not hold in the case of morphologically controlled alternations.

Jusczyk and colleagues (1993) suggest that infants may learn phonotactic patterns faster than morphological ones. This appears to be at odds with the results of the present study, in which the morphologically controlled pattern was learned with greater ease than the phonotactic pattern. There are two possible explanations for this difference. First, the present study addressed phonotactic patterns that are outside of the range of naturally occurring attested phonological

patterns. Thus, it may be possible that phonotactic patterns are easier to learn than morphologically controlled alternations, but only when the phonotactic pattern falls within the set of strictly piecewise languages. Second, the artificial nature of the present study may have provided a shortcut to learning. Participants were told that they were hearing morphological alternations. In a natural learning situation, the learner has to discover both the morphological component to the pattern as well as the phonological component. Infants may be better at learning phonotactic constraints simply because there is less information to attend to. Adults in a language learning task can use morphology as a cue to learning in a way not possible in infant language learning.

The present study leaves open the question of why morphologically controlled phonological alternations might allow for a larger range of possible languages than phonotactic patterns. One possibility is that morphology provides additional cues to learning that may not be possible when learning a purely phonotactic pattern. This falls in line with theories that predict that metalinguistic cues such as social factors and communicative intent play an important role in the typology of language and language learning. Another possibility is that the computational power of morphological and syntactic processes exceeds that of purely phonotactic patterns. Thus, patterns at the interface of phonology and morphology/syntax may thus fall outside of the computational power of purely phonological patterns (Heinz & Idsardi, 2011). If this is the case, there is a question of how to integrate phonotactic patterns at the interface between morphology and syntax.

This issue has important implications for computational models. If morphologically controlled phonological patterns are governed by a different set of constraints than phonotactic patterns, there is a question of how to incorporate both into a computational model of language and language learning. The ultimate goal of linguistics is to provide a model of language that explains the mechanisms that underlie the processes that are found (and are not found) in natural language that is both cognitively plausible and computationally elegant. Understanding the factors that learnability of various types of phonological patterns will ultimately lead to an understanding of the cognitive capacity for language.

### Acknowledgments

The author would like to thank Jeffrey Heinz, Regine Lai, Patricia Reeder, Elissa Newport, Gary Dell, members of the Aslin-Newport Lab, as well as participants at the 2011 MidPhon Conference for helpful comments and discussion. In addition, we would like to thank Kelly Johnston, Anna States and Emily Kasman. I assume all responsibility for any errors. Funding was provided by NIH Grants DC00167 and T32DC000035.

### References

- Akinlabi, A. (1996). Featural affixation. *Journal of Linguistics*, 32, 239-289.
- Boersma, P., & Weenink. (2005). Praat: Doing phonetics by computer.
- Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments and Computers*, 25, 257-271.
- Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5), 429-492.
- Finley, S. (2009). Morphemic vowel harmony as featural correspondence. *Lingua*, 119(3), 478-501.
- Finley, S. (in press). Typological asymmetries in round vowel harmony: Support from artificial grammar learning. *Language and Cognitive Processes*.
- Heinz, J. (2007). *Inductive learning of phonotactic patterns*. Ph.D. dissertation, UCLA.
- Heinz, J. (2010). Learning long-distance phonotactics. *Linguistic Inquiry*, 41(4), 623-661.
- Heinz, J. (2011a). Computational phonology part I: Foundations. *Language and Linguistic Compass*, 5(4), 140-152.
- Heinz, J. (2011b). Computational phonology part II: Grammars, learning, and the future. *Language and Linguistics Compass*, 5(4), 153-168.
- Heinz, J., & Idsardi, W. (2011). Sentence and word complexity. *Science*, 333(6040), 295-297.
- Heinz, J., & Rogers, J. (2010). Estimating strictly piecewise distributions *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics.
- Iwara, A., Akinlabi, A., & Truckenbrodt, H. (2003). The tonal phonology and phonetics of the future negative in Lokaa. *Proceedings of the 33rd Annual Conference on African Linguistics*, 33, 103-115.
- Jusczyk, P., Friederici, A., Wessels, J., Svenkerund, V., & Jusczyk, A. M. (1993). Infants' sensitivity of the sound patterns of native language words. *Journal of Child Language*, 32, 402-420.
- Nevins, A. (2009). On formal universals in phonology. *Behavioral and Brain Sciences*, 32, 461-462.
- Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech perception. *Journal of Memory and Language*, 53, 225-237.
- Rogers, J., Heinz, J., Bailey, G., Edlefsen, M., Vissher, M., Wellcome, D., et al. (2010). On languages piecewise testable in the strict sense. In C. Ebert, G. Jaeger & J. Michaelis (Eds.), *The mathematics of language* (Vol. 6149, Lecture Notes in Artificial Intelligence, pp. 255-265). Berlin: Springer.

# Systemic Expertise: Instructing Non-Artists on Depicting Human Figures in 3-D

Nick V. Flor

Virtual World Research Lab  
Anderson School of Management  
University of New Mexico  
nickflor@unm.edu

## Abstract

I use the theoretical framework of distributed cognition to develop a repeatable procedure that non-artists can follow to model 3-D human figures—manikins—on a computer, which are sufficient for prototyping. The approach was not to train artists to be expert human figure modelers, but rather to derive a distribution of abilities across person and computer such that the system of person-in-interaction-with-technology exhibited expertise. These abilities were discovered through an analysis of two equivalent functional systems: figure modeling on a computer and figure drawing on paper. I report a test of the procedure on a group of non-artists, which yielded a high success rate. This research contributes to our understanding of applying distributed cognition to the design of instructional procedures, and to our understanding of the sciences of the artificial.

**Keywords:** distributed cognition; science of the artificial; 3-D human figure modeling; virtual worlds.

## Introduction

Drawing the human figure on paper is a difficult task, as anyone who has picked up a pencil and tried to draw a figure can attest. Modeling the human figure on a computer is also a difficult task, since it either starts with a drawing of a human figure (Oliverio, 2007; Patnode, 2008; Russo, 2006), or it requires detailed knowledge of surface anatomy (de la Flor & Mongeon, 2010; Ratner, 2009; Spencer, 2010). Moreover, one must learn how to operate a complex 3-D software package.

The general problem is that of representing an imagined human figure in some medium, which requires a complex skillset typically acquired through years of practice. For this reason when there is a task that requires either human or human-like figures, such as an animated movie or a video game, one hires artists who are experts in figure modeling.

The specific research question explored in this paper is as follows: is it possible to design a replicable procedure that *non-artists* can learn in a short amount of time for modeling human figures on a computer? This research contributes to our understanding of distributed cognition (Hutchins, 1995), and the sciences of the artificial (Simon, 1996).

The answer to this question has practical implications as well. We live in a highly networked, digital ecosystem where almost every mobile device, notebook, or desktop computer has a graphics processor capable of displaying 3-d models. There is significant opportunity for designing scientific and educational applications that make use of this capability such as virtual worlds (Bainbridge, 2007; see Figure 1). There are also many individuals that have business and information systems backgrounds that would

like to develop applications that take advantage of these 3-D capabilities, but lack the skills in figure modeling. Certainly one could always hire an artist, but knowing how to model figures in 3-D and to model objects generally, allows one to more quickly explore the spaces of virtual world prototypes and possible 3-D representations. Artists can still be utilized later in the process to add finishing aesthetics.

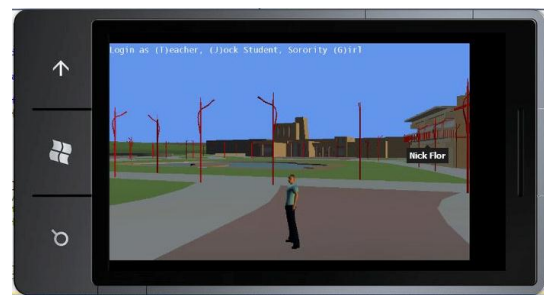


Figure 1. Virtual UNM Campus Developed by the Author on a Mobile Device—Human Figure in the Center

Hutchins (1995, p. 155), provided a guiding framework for answering this research question when he wrote:

*“...a good deal of the expertise in the system is in the artifacts, (both the external implements and the internal strategies)—not in the sense that the artifacts are themselves intelligent or external agents, or because the act of getting into coordination with the artifacts constitutes an expert performance by the person; rather, the system of person-in-interaction-with-technology exhibits expertise. These tools permit the people using them to do the tasks that need to be done while doing the kinds of things the people are good at: recognizing patterns; modeling simple dynamics of the world, and manipulating objects in the environment” (p. 155).*

In the context of this framework the research question becomes not can we make a non-artist into an expert figure modeler, but rather can we design a system of external implements and internal strategies, such that the system of person-in-interaction-with-technology exhibits expert figure modeling skills?

## The Processes of Human Figure Modeling and Human Figure Drawing

The task of modeling a human figure on a computer consists of a person (the modeler) and the computer running the 3-D modeling software. The challenge is to discover a distribution of abilities across person and computer, which allow a non-artist to model a human figure in a short amount of time.



A comprehensive cognitive ethnography of the process of modeling and drawing the human figure is beyond the scope of this paper. Instead, I describe the general processes and key representations that occur during modeling, based on my experience building 3-D virtual worlds, as well as working with and observing professional Hollywood modelers, animators, and visual FX personnel, as a faculty member and director of the University of New Mexico’s interdisciplinary film & digital media program.

### The Process of Human Figure Modeling

The process most commonly used in industry to model human figures is commonly referred to as *character modeling with reference pictures*, although I use the term “3-D tracing” because it is analogous to the technique of placing see-through paper over a photograph and creating a drawing by tracing. The method can be seen in a variety of books including, Guindon (2007, pp. 15-83), Ingrassia (2009, pp. 269-288), Oliverio (2007, pp. 47-323), Patnode (2008, pp. 133-207), and Russo, (2006, pp. 83-152). The general process is as follows (see Figure 2 for depictions of the steps):

- 1.The modeler first obtains reference pictures. In the movie and videogaming industries, these reference pictures are typically drawn by a concept artist in the art department. The concept artist either draws the pictures on paper then scans them into an electronic format, or uses a paint program to draws them directly into an electronic format. The reference pictures depict front and side views of the character to be modeled, with features lined up in both photos including the top of the head, the waistline, and the bottom of the feet.
- 2.The reference pictures are then imported into a modeling program. The pictures are placed at right angles to one another, and centered.
- 3.The modeler adds a simple box to the workspace so that it overlaps the torso in both the front and the side reference-picture views.
- 4.Finally, the modeler molds the box to the reference drawings—extruding limbs, adding edge loops, and moving vertices so that the edges of the model line up with the edges in the reference pictures. Figure 3 details the molding process.

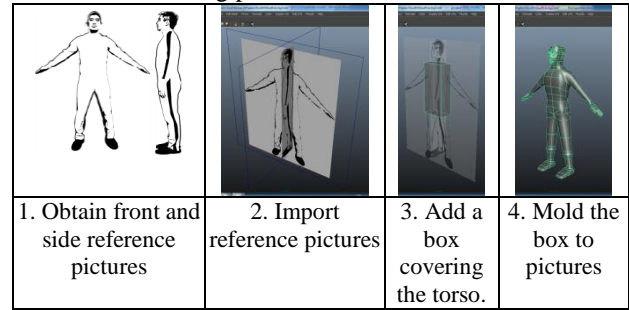


Figure 2. Figure Modeling with Reference Pictures, viz., 3-D Tracing. A modeler (1) obtains front and side pictures of a human figure; (2) imports them into a 3-D modeling program; (3) adds a basic box that covers the torso; (4) molds the box to the pictures.

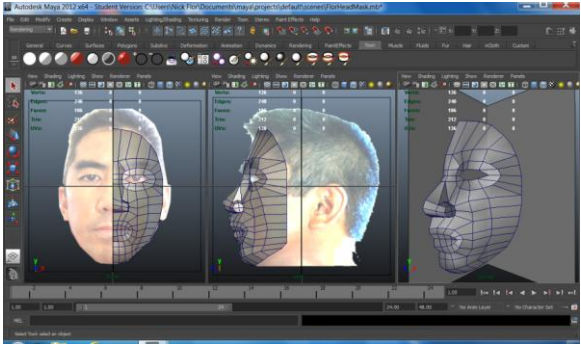


Figure 3. Molding to Reference Images—a time-intensive activity where an artist molds a box or a plane to images, by aligning vertices to landmarks on front and side pictures. In this example, a plane is molded to a face, but the same actions apply to molding a box to a body. The pictures are scaffolds that are removed after modeling is completed.

**Abilities and Issues for Non-Artists.** For a non-artist to successfully perform the figure modeling process, the main abilities he or she must acquire are to draw the human figure in front and side views (step 1), and to align vertices to landmarks in front and side views (step 4). As mentioned, figure drawing is a difficult skill to learn. Moreover, the alignment process is time consuming. For this reason, figure modeling with reference pictures is not a suitable method for the non-artist.

There is no research that analyzes these abilities and sheds light on developing a new figure modeling process for non-artists. However, an analysis of the figure drawing process reveals abilities that are helpful for developing a new figure modeling process via a redistribution of those abilities across person and computer.

### The Process of Human Figure Drawing

The output of figure drawing is both an input to the figure modeling process (step 1 in Figure 2), and is also an equivalent functional system to figure modeling. It is equivalent because both figure drawing and figure modeling result in a representation of the human figure. The difference is the medium—paper in the case of figure drawing, and a computer in the case of figure modeling.

Observations of artists reveal that the drawing of a human figure on paper results from at least three distinct kinds of representations layered on top of one another (refer to Figure 4). These observations are corroborated by information in drawing instructional books (Hampton, 2009, pp. 54-55; Lee & Buscema, 1978, pp. 70-71; Loomis, 1943, pp. 43-45; Reed, 1984, p. 38).

- 1.The first representation the artist creates is a rough sketch of the figure that depicts broadly the action of the body and limbs. Examples of rough figure sketches include *stick figures*, which use lines to represent the body and limbs; and *gestures*, which are drawings that are completed quickly, in several minutes or less, that are sufficient for a viewer to understand the meaning of the action depicted.



2. The artist then layers a representations consisting of geometric primitives based on spheres, cubes, and cylinders over the rough sketch in order to give the figure both mass and depth (Hampton, 2009, p. 54; Hogarth, 1996, p. 8; Lee & Buscema, 1978, p. 21; Ozawa, 1999; Reed, 1984, pp. 22,24,26). These primitives serve an important function which is known in the psychological literature as scaffolding (Wood, Bruner, & Ross, 1976). As one artist put it: “Square boxes ... are necessary for proper depiction of the human form as well. Picture the 3-dimensional human body and think of it in terms of boxes... Once you become accustomed to drawing, there will be no need to draw the boxes, but you should always keep them in mind to make certain you think in three dimensions.” (Ozawa, 2005, p. 15)
3. Finally the artist layers representations of muscles, clothing, and shadows over the geometric primitives in order to add realistic detail to the figure. It should be noted that this last representation varied among artists, with some listing it as one representation and others breaking it into as many as three separate representations. Moreover, these are not the only representations that an artist can layer next. For example, if the publisher allows colors, the artist layers color over the detailed figure drawing. However, I list this as one representation—detail—because the general consensus was to layer various kinds of detail over the second representation of the figure-as-geometric primitives.

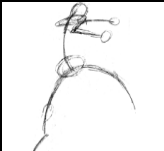


		
Representation 1: Rough action sketch	Repr. 2: Mass via shapes	Repr. 3: Detail, which can include muscles, clothing and shading

Figure 4. Figure Drawing as a Sequence of Layering Three Representations (figures adapted from Lee & Buscema, 1978)

The artist’s use of these intermediate representations is a good example of the principle of cognitive effort conservation in a distributed cognitive system. Without using these intermediate representations an artist would have to imagine a fully-detailed figure mentally and then draw it. This kind of imagining is a computationally-intensive process for skills like perspective. Instead, when using intermediate representations, an artist can work on different aspects of the figure in smaller more manageable steps, which then serves as a kind of guiding framework for adding more complexity in subsequent representations. In particular, the rough sketch allows the artist to focus on depicting the meaning of the pose. Once complete, this rough sketch serves as a framework for the piecemeal addition of body masses with depth using simple boxes, cylinders, and other geometric shapes, which are easier for

the artist to draw in perspective than actual body parts. Once the figure with mass and depth is completed, the basic geometric shapes serve as a framework for the piecemeal placement of muscles, clothing, and lighting details. In this manner the use of intermediate representations changes the task from a detailed thinking then drawing task, into an easier and incremental looking and drawing task.

**Abilities for Non-Artists.** The observable representations are produced by specific abilities, which in the case of figure drawing on paper are located mainly in the head of the artist and that take many years of practice to develop. According to Loomis (1943), the fundamental abilities required for human figure drawing include “proportion, anatomy, perspective, value, color, and knowledge of mediums and materials.” Briefly, an artist uses *proportion* ability to draw body parts in the proper, aesthetically pleasing, sizes relative to one another (Repr. 1). Ability in *anatomy* is used to detail the surface of the body as influenced by the underlying skeletal and muscular structures (Repr. 3). *Perspective* ability allows the artist to draw the body parts from different observation points and with *foreshortening* (Repr. 2), where body parts that are closer to an observer appear bigger than those parts that are further away. It is important to note that foreshortening is the single most difficult ability for drawing figures (Hogarth, 1996, p. back cover). *Value* skill is used to apply the appearance of light and shading to the figure (see last image in Repr. 3). *Impact* is the ability to draw figures in poses that are visually interesting (Repr. 1; Lee & Buscema, 1978, p. 60). Finally, knowledge of mediums and materials is an operational skill—the ability to use paper and pencil or other media and instruments for drawing. Figure 5 depicts the abilities needed for figure drawing on paper.



Figure 5. A Visual Depiction of the Abilities for Figure Drawing.

## Redistributing Abilities, Manikins & Satisficing

All the abilities for drawing a human figure on paper, with the exception of proportion, are difficult or time-intensive for an individual to learn.

However, when modeling a human figure on a computer one can have the computer perform most of the difficult abilities, including rendering the figure with the proper perspective, with the proper shading (value), and with color.

The problem of learning anatomy remains, but one can satisfice and remove anatomical detail as a goal, yet still

produce human-like figures that are useful for virtual world prototyping. This is not without precedence in figure drawing or animation. Both artists (Loomis, 1943) and animators (Williams, 2001) emphasize the use of manikins (similar to Figure 6), devoid of anatomical detail, for prototyping poses or movements. Detail is added only after the poses and movements have been worked out.

When anatomical ability is removed as a requirement, a non-artist need only learn proportion and impact to model human-like figures on a computer. Like in drawing and animation, detail can always be added later to a manikin. Figure 6 depicts a distribution of abilities suitable for a novice to model 3-D figures on a computer, and where the computer does the more difficult representational abilities.



Figure 6. A visual depiction of the distribution of abilities needed to model figures for *prototyping* purposes. Abilities in parentheses are added abilities that are not part of the figure drawing skill set (not covered in this paper)

### A Procedure for 3-D Figure Modeling

The redistribution of abilities as depicted in Figure 6, changes the task of modeling a human figure, from drawing and molding boxes to reference pictures (see Figure 2), to the easier task of creating body parts in proportion. The values for the following procedure were derived from the body proportions specified by (Hamm, 1963).

*3-D modeling* is an activity where an individual creates representations of real or imagined things—that appear to have width, height, and depth—using a special kind of computer program known as modeling software. Popular examples of modeling software include Maya and 3ds Max, both by Autodesk Corporation. The representations are known as models, and the person that creates them is referred to as a modeler.

There are two basic ways to model, using polygons or using curves. This instruction covers polygon modeling—the most popular kind of modeling for games and virtual worlds. Using polygon modeling, the modeler starts with a basic object — e.g., a cube, sphere, or pyramid with a certain *width x height x depth* and with a certain number of faces — much like a sculptor starts with a lump of clay. The modeler then molds this source object into a target object using the operations of moving, cutting, extruding, wedging, merging, deleting, and mirroring, which are applied to the object’s vertices, edges, and faces.

### Part 1. Creating the Body

To create a body (Figure 7): (1) Create a cube that is 1x2.5x1 units with 2x5x2 faces; (2) Wedge the top back face 45 degrees, yielding an arm face; (3) Move the outer-bottommost edges .1 units in the x direction yielding a leg face; (4) Extrude the arm face 2.5 units, tapered by .1 units; (5) Extrude the leg face 2 units yielding a thigh; (6) Move the thigh face’s inner-bottommost edge +.1 units in the x-direction; (7) Extrude the thigh face 2 units, tapered by .1 units, yielding a calf; (8) Move the middle back vertices to a z-value of zero and merge the overlapping vertices, yielding a back wedge; (9) Mirror to see what the completed body will look like; (10) Save file.

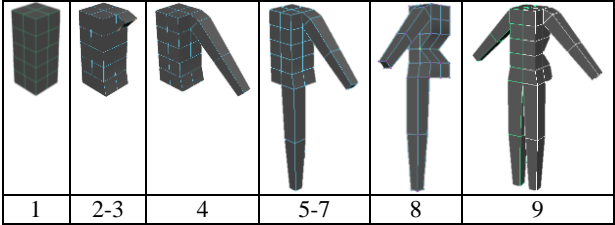


Figure 7. Modeling the Body (see text for explanation)

### Part 2. Creating the Hand

To create a hand (refer to Figure 8): (1) Create a cube .5x.125x.5 with 2x1x4 faces; (2) Wedge the thumb face 30 degrees; (3) Inset the stubs .01; (4) Extrude the fingers .5 with 3 divisions; (5) Extrude the thumb .25, then .125; (6) Move back the pointer and ring finger .0625 and move the pinky back 0.125; (7) Move the pinky joint edges back -.0625 and the middle finger joint edges up .0625; (8) Remove the back edges and vertices; (9) Move in the wrist edges inwards by .1; (10) Save file.

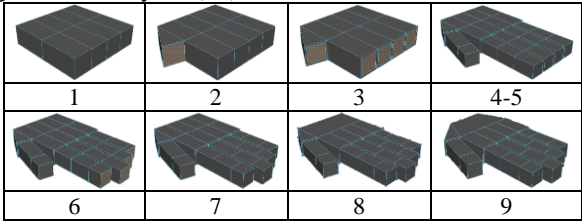


Figure 8. Modeling the Hand (see text for explanation)

### Part 3. Creating the Feet

To create a foot (Figure 9): (1) Create a cube that is .5x.25x1, with 1x2x2 faces; (2) Merge the toe vertices down; (3) Move the ankle sides in by .1, and the front ankle by .2; (4) Extend the toes by 1.25; (5) Save file.

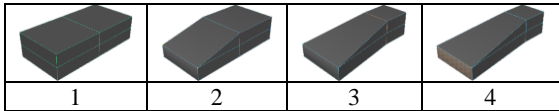


Figure 9. Modeling the Foot (see text for explanation)

## Part 4. Creating the Head

To create a head (refer to Figure 10), simply create a cube that is .7x1x1 with 2x3x3 faces, then save the file. The head can be shaped after it is attached to the body in part 5.



Figure 10. Modeling the Head (see text for explanation)

## Part 5. Combining the Pieces

The final step is to merge all the pieces into one (refer to Figure 11). After saving all the separate body parts: (1) Import the half-body; (2) Import the foot; (3) Snap foot to leg; (4) Merge objects & vertices; (5) Import hand; (6) Snap hand to upper limb; (7) Merge objects and vertices; (8) Mirror to get the right side of the body; (9) Attach the head; (10) Smooth figure if desired; (11) Save.

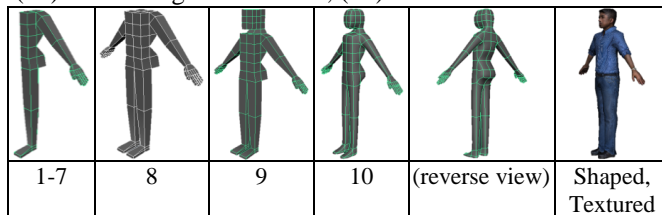


Figure 11. Attaching the Body Parts (see text for explanation)

The resulting manikin is suitable for inserting into a game for prototyping purposes, or one can take the time to further shape and texture the manikin to look more realistic (see Figure 11, last image).

## Method for Testing the Procedure

### Participants

Forty-three students participated in this study. The students came from MGMT 330—Fundamentals of Virtual Business Programming, and none of them had any experience with figure drawing or 3D modeling prior to taking the course, which was a required course for junior and senior level students majoring in management of information systems.

### Apparatus

For hardware, students used their personal laptops or desktop computers to run the 3-D modeling software. The software used was Maya 2010, which the students downloaded for free from the Autodesk.com website as part of Autodesk's free software program for students and faculty. All 43 students reported successfully loading Maya onto their personal computers. The University's online instructional system, WebCT, held a link to YouTube tutorial videos made by the instructor on the topic of modeling 3D characters and this video was made available to all students.

## Procedure

Students were given a lecture on how to model 3D figures based on the procedure described previously. They were then given an assignment where they had a week both to create a personalized robot avatar and to animate a walk cycle for the robot avatar (see Figure 12 for the written instructions). Note: I used the term "robot" instead of manikin since students were more familiar with the former.

- Based on the Professor's Robot as taught in class:
1. Model a robot body in Maya (filename: body.mb).
  2. Model a robot head in Maya (filename: head.mb).
  3. Model a robot hand in Maya (filename: hand.mb).
  4. Model a robot foot in Maya (filename: foot.mb).
  5. Personalize your robot body parts.
  6. Merge all the body parts to yield a robot avatar.
  7. Animate a walk cycle for your robot avatar (filename: robot.mb).
  8. Export your walking robot avatar as an fbx file (filename: robot.fbx).

Figure 12. Written Instructions Given to Students

Students received the lecture through YouTube videos that were embedded in the University's online instructional delivery system, WebCT. The total time for the video lectures was approximately one hour. MGMT 330 was a purely online course with no face-to-face component.

## Results

There were 43 students in MGMT330. Of these 43 students, 33 turned in a robot, while 10 students did not turn in a robot. Of the 10 students that did not turn in a robot, 7 did not turn in any assignments at all during the semester. Of the 33 students that turned in a robot, 3 did not save their files properly and only turned in an animated skeleton. However, when their work was checked in person, they indeed modeled a robot correctly. If you removed, the 7 students that did not show up to class, then 91.6% of the students successfully followed the procedure for modeling a robot. Figure 13 depicts the robots created by the students.

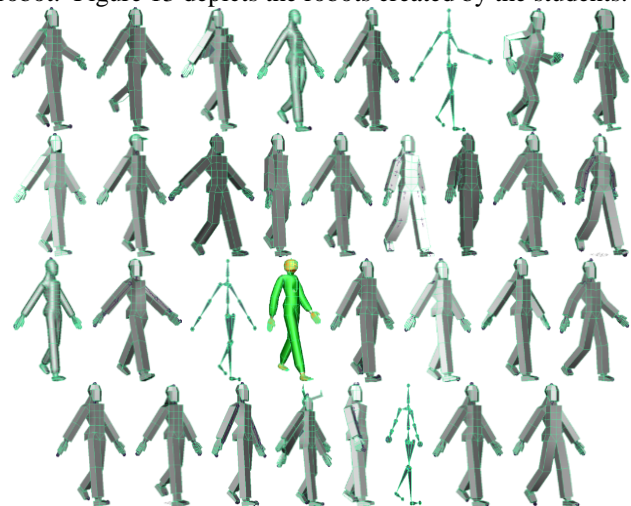


Figure 13. Figures Created by the Management Students

## Discussion and Conclusion

The research in this paper presents a repeatable procedure I developed for modeling human figures in 3-D based on a distributed cognition analysis, which can be used to create characters for prototype virtual worlds in academia and industry. Unlike teaching someone how to draw human figures on paper, teaching a person how to model human figures on a computer can be done reliably and in a relatively short amount of time—at least for sophomore-level college students and beyond.

While both drawing and modeling are instances of the general activity of representing human figures in some medium, the specific tools & media change the distribution of abilities needed by the artist. When the tool is a computer running the modeling software, many of the abilities that are difficult and time-consuming for a person to master such as drawing objects in the proper perspective on paper, get distributed to the computer.

While a procedure for quickly modeling character for virtual worlds and games is valuable, equally important is the approach used to design this instructional procedure. Having to design instruction outside of one's expertise is not limited to instructors in academia, but is an activity that both managers and employees may have to engage in especially in a project with new employees using new tools. While instructional procedures are not always written down, they are important to create nonetheless. In situations, where one must design instruction outside of one's expertise, the distributed cognition theoretical framework suggests that it is possible to:

1. Find a reference system of activity that is equivalent in terms of its output to the target system. In the current study, figure drawing on paper was equivalent to figure modeling on a computer. Both activities are known as functional systems in the distributed cognition framework.

2. Analyze observable representations in the reference system, to determine the fundamental abilities needed to produce the result in the reference system. In this study, the fundamental abilities included proportion, impact, perspective, value, and anatomy.

3. Decide on a new distribution of abilities based on: (a) the goal of the larger system that the target system is part of—or the intended use of the output of the target system; (b) the capabilities of the technology in the target system; and guided by (c) the principle of effort conservation. For figure modeling, abilities like proportion and impact were kept with the person, whereas abilities like perspective and value were distributed to the computer.

What are the limitations of this procedure? Consider that there were two fortunate consequences of my analysis: (1) the abilities that remained after I applied the procedure, i.e., proportion and impact, were easy for a non-expert to master; and (2) the abilities distributed to the computer, such as perspective, foreshortening, and shading (value)—abilities that are difficult for an artist to master when drawing figures on paper—were already implemented by the software.

If after ability distribution, the difficult abilities are not already implemented in the software, someone will have to implement their functionality. And if the remaining abilities are difficult for people to master, then designing an instructional procedure may not be an option.

However, instructors at least have another option to consider in designing their pedagogy. Furthermore, this procedure gives instructors the potential to create novel and possibly innovative forms of instructional material, which provide students the ability to create artifacts that were previously only possible by experts.

## References

- Bainbridge, W. (2007). Scientific Research Potential of Virtual Worlds. *Science*, 317, 472-476.
- de la Flor, M., & Mongeon, B. (2010). *Digital Sculpting with Mudbox*. Burlington, MA: Focal Press.
- Guindon, M. A. (2007). *Learning Autodesk Maya 2008*. San Rafael, CA: Autodesk.
- Hamm, J. (1963). *Drawing the Head and Figure*. New York: Berkley Publishing Group.
- Hampton, M. (2009). *Figure Drawing: Design and Invention*. M. Hampton Publishing.
- Hogarth, B. (1996). *Dynamic Figure Drawing*. New York: Watson-Guption Publications.
- Hutchins, E. (1995). *Cognition in the Wild*. Cambridge: MIT Press.
- Ingrassia, M. (2009). *Maya for Games*. Burlington: Elsevier.
- Lee, S., & Buscema, J. (1978). *How to Draw Comics The Marvel Way*. New York: Simon & Schuster.
- Loomis, A. (1943). *Figure Drawing for All It's Worth*. New York: Viking Press.
- Oliverio, G. (2007). *Maya 8: Character Modeling*. Plano: Wordware Publishing.
- Ozawa, T. (2005). *Let's Draw Manga: Bodies and Emotions*. Gardena: Digital Manga Publishing.
- Patnode, J. (2008). *Character Modeling with Maya and ZBrush*. Burlington, MA: Focal Press.
- Ratner, P. (2009). *3-D Human Modeling and Animation*. Hoboken, New Jersey: Wiley & Sons.
- Reed, W. (1984). *The Figure: The Classic Approach to Drawing & Construction*. Cincinnati: North Light Books.
- Russo, M. (2006). *Polygon Modeling: Basic and Advanced Techniques*. Plano: Wordware Publishing.
- Simon, H. (1996). *The Sciences of the Artificial*. Cambridge, MA: MIT Press.
- Spencer, S. (2010). *ZBrush Digital Sculpting: Human Anatomy*. Indianapolis: Wiley Publishing.
- Williams, R. (2001). *The Animator's Survival Kit*. New York: Faber and Faber.
- Wood, D., Bruner, J. S., & Ross, G. (1976). The Role of Tutoring in Problem Solving. *Journal of Child Psychiatry and Psychology*, 17, 89-100.



## Task switching without knowledge of the tasks.

C.L. Forrest (clf206@exeter.ac.uk)

H. Elchlepp (H.Elchlepp@exeter.ac.uk)

S. Monsell (S.Monsell@exeter.ac.uk)

I.P.L. McLaren (I.P.L.McLaren@exeter.ac.uk)

School of Psychology, College of Life and Environmental Sciences,  
University of Exeter, UK.

### Abstract

Task-cuing paradigms are typically taken to explore control of task-set. However, they can be construed as requiring not selection of a task-set, just retrieval of a cue+stimulus-->response (CSR) mapping. In this paper we considered performance in a task-cuing paradigm in which participants saw a color cue that indicated whether they should classify a digit as odd/even or high/low using one of two responses. Half the participants were instructed in terms of tasks (Task group) whilst the others were required to learn the CSR mappings without mention of tasks (CSR group). Predicted performance under CSR conditions was modeled using an APECS connectionist network. Both the model and CSR group produced small switch costs, mostly due to incongruent stimuli, and large congruency effects that reduced with practice. In contrast, the Task group produced a larger switch-cost and a smaller, stable congruency effect.

**Keywords:** task-switching, connectionist modeling, conditional discriminations, associative learning

### Introduction

We often think of our behaviour as being governed by both higher-level cognitive control processes and lower-level associative processes (McLaren, Green & Mackintosh, 1994). Typically these processes are thought to operate simultaneously but with a degree of independence. This paper takes a task-cuing paradigm, typically taken as measuring the higher level cognitive control processes involved in changing between tasks (Monsell, 2003), and asks if the performance typically seen could instead be accounted for by lower level associative processes. This paradigm has been used widely to measure control processes in areas as diverse as aging (Mayr, 2001) and schizophrenia (Meiran, 2000) It is also commonly included in brain training packages as a way to improve your ability to multitask and pay attention. Given such widespread use, it is important to assess if the paradigm actually measures control processes at all; it has been argued that it does not (Logan and Bundesen, 2003; Schneider and Logan, 2005).

The response contingencies in many task-cuing experiments can be construed without any reference to tasks. This paper examines what happens when participants approach such an experiment without knowledge of the task-sets. Data and simulation suggest that they can learn the statistical structure of the experiment through the use of

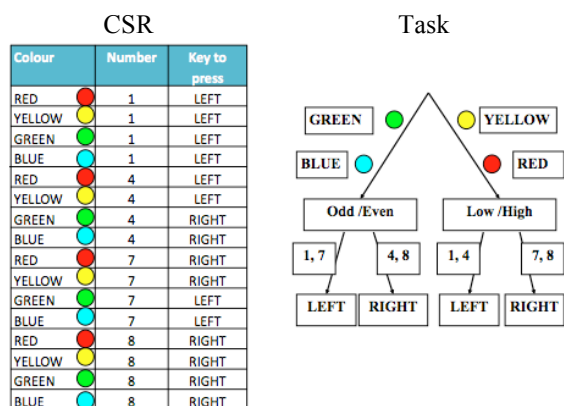


Figure 1 on the left shows the way in which the experiment was presented in the CSR condition and on the right in the Task condition.

associative learning mechanisms, but their performance differs from that of participants construing the situation as one requiring task-selection and switching.

To clarify this, let us consider the task-cuing paradigm that we used. Under standard instructions in this paradigm participants are told, for example, that if the background circle is blue or green then they should classify the digit they then see as odd/even, where odd requires a response with the left key and even with the right. However, if the background circle is red or yellow they should classify the digit as higher/lower than 5, with a right response for high, and a left response for low. This is the "task-set" construal of what is required, as illustrated on the right of Figure 1. Yet participants do not need knowledge of these tasks to know how to respond, as the color and the number combination is completely predictive of the required response, e.g. a 4 on yellow will always require a left response. Hence, especially with small stimulus set, it is entirely possible for participants in a task cuing experiment not to use the task-sets at all, in which case the experiment is not measuring task-based control processes.

In the experiment reported here we compare a group who are explicitly instructed to use the task-sets with one that has no knowledge of the underlying task-set structure. In order to examine whether, and in what ways, performance differs

between the two conditions we will consider three of the common effects found within the task-switching literature: the switch cost, the reduction in the switch cost with time to prepare (RISC effect) and the congruency effect (Monsell, 2003, Kiesel et al, 2011).

It is typically found that when participants change from performing one task to performing another task there is a switch cost; participants are generally slower and less accurate on a task-switch trial than a task-repeat trial. Participants are also able to reduce this switch cost when they are given more time to prepare the task set, i.e. when there is a longer time between the cue (colored circle), which indicates the task-set, appearing and the stimulus (number) appearing the switch cost declines. Explanations of these effects have appealed to task set inertia (Allport, Styles & Hsieh 1994) —conflict due to residual activation of the previous task set - and/or the need to perform a task set reconfiguration process (Rogers & Monsell, 1995) which reduces conflict if performed before the stimulus appears. But, according to the compound-cuing model of Logan and colleagues (Logan & Bundesen, 2003; Schneider & Logan, 2005) participants simply retrieve the response associated with the combination of cue and stimulus, so these effects cannot be taken as hallmarks of control.

In the task-cuing experiment already described, the responses for the two tasks are mapped onto the same keys, i.e. the left key represents odd and high, whilst the right key represents even and low. Hence for some numbers the response is always the same regardless of the task cued, e.g. 1 always requires a left response; these are called *congruent* stimuli. For other numbers the response changes with the task cued, e.g. 4 requires a left response if the task is high/low but a right response if the task is odd/even; these are *incongruent* stimuli. Typically, it is found that participants are faster and more accurate for congruent than incongruent stimuli. As with the switch cost and reduction in switch cost (RISC) effect there have been both task-set based and non-task-set based explanations of this congruency effect. Some researchers have argued that the congruency effect is due to response conflict from the currently irrelevant task-set (Kiesel et al, 2011). Other researchers have argued that it is caused by associative interference, as the incongruent stimuli are linked to both responses whilst the congruent stimuli are only linked to one (Kiesel, Wendt & Peters, 2007).

In this experiment we asked whether the switch cost, the RISC effect and the congruency effect depend on how the participants construe the experiment, i.e. whether in terms of tasks or cue + stimulus to response (CSR) mappings.

In addition to considering these standard task switching effects we also considered the effect of introducing novel stimuli (cf. Rogers & Monsell, 1995). This is particularly relevant for assaying the difference between switching among stimulus-classification task rules versus applying a single set of learned CSR rules. For participants using tasks there should be little impact of introducing new stimuli. There might be a slight novelty effect, but they should be

able to treat the new stimuli in the same way as the old, continuing to apply the same classification rules. However, participants with no knowledge of the task-sets have no way of knowing how to respond to the novel numbers; they should be reduced to learning how to respond by trial and error, and one would expect performance on the new numbers to be dramatically worse than performance on the old numbers.

### Modeling

As summarized above there is plenty of evidence to suggest how participants typically perform in a task-cuing paradigm with knowledge of the tasks (Monsell, 2003, Kiesel et al, 2010). In order to attempt to predict how participants would perform in the task-cuing experiment described above without knowledge of the task-sets we simulated performance using an associative model. The mappings for the congruent stimuli are shown in outline in Table 1. It is immediately evident that they should be easily captured by an associative model, as the stimuli in isolation predict the correct response.

		Cues (Color)			
		W (blue)	X (green)	Y (red)	Z (yellow)
Stimuli (Digit)	<b>A (1)</b>	<b>L</b>	<b>L</b>	<b>L</b>	<b>L</b>
	B (3)	L	L	L	L
	C (6)	R	R	R	R
	<b>D (8)</b>	<b>R</b>	<b>R</b>	<b>R</b>	<b>R</b>

*Table 1 The associative structure of the congruent trials. L indicates a left R a right response. Boldface rows indicate example initially trained stimuli; the others introduced later*

The incongruent stimuli, shown in Table 2, are more of a challenge for an associative model. There is evidence from rabbits (Saavedra, 1975) and humans (Livesey et al, 2011) that, although these stimuli are harder to learn than the congruent stimuli, they can be learned. However, a single layer error-correcting model, e.g. Rescorla-Wagner (1972) would be unable to learn this structure.

		Cues (Color)			
		W (blue)	X (green)	Y (red)	Z (yellow)
Stimuli (Digit)	E (2)	R	R	L	L
	<b>F (4)</b>	<b>R</b>	<b>R</b>	<b>L</b>	<b>L</b>
	<b>G (7)</b>	<b>L</b>	<b>L</b>	<b>R</b>	<b>R</b>
	H (9)	L	L	R	R

*Table 2 shows the associative structure of the incongruent trials using the conventions employed in Table 1.*

In addition to the difference in performance on incongruent and congruent trials, one might also expect effects of cue equivalence (Honey & Ward-Robinson, 2002; Hodder, George, Kilcross & Honey, 2003). These studies

trained rats or humans (respectively) with the same contingencies as the incongruent trials. They found that cues that indicated the same outcome from stimuli became equivalent e.g. here W and X would become equivalent as would Y and Z, in that there would be a greater degree of generalization between W and X than W and Y. Honey and Ward-Robinson (2002) found that a modified connectionist model was able to account for their data by allowing the same hidden unit to carry the mappings for equivalent cues. We used a model from the same class as their chosen model. The model is known as APECS (McLaren, 1994, 2011; LePelley & McLaren, 2001) and has a good record in modeling human learning and memory. APECS has the basic characteristics of a back-propagation network (Rumelhart, Hinton and Williams, 1986), i.e. it is a standard feedforward error correcting system with input, hidden and output layers that has been modified in two key ways:

**Learning algorithm and rates** The APECS learning algorithm allows the learning rates to change in an adaptive manner. On each trial, the hidden unit with the largest error receives a higher learning rate than the other hidden units. This effectively means that one (or a few) hidden unit(s) is (are) selected to carry each mapping from input to output.

**Bias** The APECS group of models also includes an adaptive bias whose learning rate is varied to prevent catastrophic interference to old learning occurring when new information is learnt (McCloskey and Cohen, 1989). The adaptive bias lowers the chances of the same hidden unit being used by a different mapping and hence prevents the previous learning being over-written.

### Modeling Method

**Sequencing** As in the experiment below, one third of trials were "switch" trials (defined with respect to the task-set representation). The cue changed color on every trial, and either of two colors signaled each task. The number of times a given stimulus appeared in a given task on a repeat or switch trial was constrained. There were 14 blocks of 49 trials in total. For the first 10 blocks only 4 stimuli were possible, whilst for the last 4 blocks 8 stimuli were possible.

**Representation and Architecture** The 4 cues and 8 digit stimuli were represented discretely with one input unit coding for each. The responses were also represented discretely, and the model was trained to 0.9 for the correct response and 0.5 for the wrong one. It was trained to auto-associate the input with the output, with certain output units active only if a specific input unit was active. The network had three layers: 16 input units, 14 hidden units and 18 output units.

**Learning parameters** The fast learning rate was set to 0.8 whilst the slow learning rate for the unselected units was 0.0005. For the bias the learning rate for selected hidden units was 0.5 and for others was 0.005.

**Output** The output of the model was assessed by subtracting the difference between the actual activations of the two response output units (desired response – undesired response) from the target difference (0.4). On this measure larger scores mean worse performance.

### Modeling Results

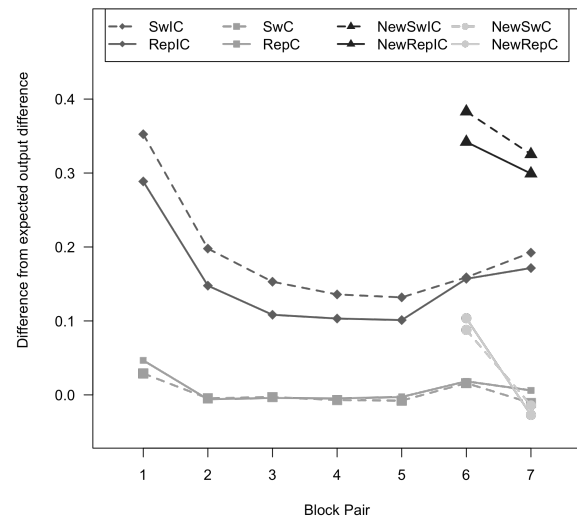


Figure 2 shows the performance of the model plotted as the difference between the desired output difference (0.9–0.5) and the actual output difference – hence 0.4 represents no learning whilst 0 represents perfect learning. The points are plotted by block pair, congruency, switch and new stimuli. Darker lines represent incongruent stimuli (IC) with diamonds representing the originally trained stimuli and triangles the transfer stimuli (New). Lighter lines represent congruent stimuli, with squares representing the originally trained stimuli and circles the transfer stimuli (New). Dotted lines represent switch trials (Sw) and solid lines repeat trials (Rep).

The results were analyzed across block pairs 2–5 (as block pair 1 was a practice block) using an ANOVA with the factors, block pair, congruency, and switch.

**Task switches** There was a small but significant effect of "task switch"; the model's performance was worse on switch than repeat trials (switch: 0.075, repeat: 0.055),  $F(1,31)=49.5$ ,  $p<0.001$  — see Figure 2.

**Congruency** There was a large and significant effect of congruency; the model's performance was worse on incongruent than congruent trials (congruent: 0.095, incongruent: 0.035),  $F(1,31)=168.5$ ,  $p<0.001$ .

**Switch by congruency.** The switch cost was significantly larger for incongruent trials (0.04) than the congruent trials (–0.002),  $F(1,30)=10.4$ ,  $p<0.01$ .

**Acquisition effects.** Overall performance reliably improved from block pair 2 to 5 (Figure 2),  $F(3, 93)=44.3$ ,  $p<0.001$ . The two-way interaction between block pair and congruency was significant  $F(3, 93)=43.3$ ,  $p<0.001$ . This interaction can be seen in Figure 2 which shows the congruent stimuli being learnt quickly whilst the incongruent stimuli take longer to learn.

**Transfer to new stimuli** The effect of transfer was analyzed by comparing the performance on the newly introduced stimuli with that on the old stimuli in block pairs 6 and 7. As expected, the model found the novel stimuli



(0.189) much harder than the previous stimuli (0.089),  $F(1,31)=509$ ,  $p<0.001$ .

### Modeling Discussion

The model predicts a large congruency effect which varies over blocks, a small switch cost which is only present in the incongruent trials and a significant disadvantage for newly introduced stimuli. This gives a clear indication what we might expect from participants if they were performing on an associative basis. It is also different from the typical task-cuing results where the switch cost is usually larger than the congruency effect. We now consider the empirical data obtained from participants trained on this task under Task or CSR instructions.

### Behavioral Method

**Participants** The participants were 35 psychology undergraduates (mean age = 20.3 years, 7 males) at the University of Exeter. Participants took part for course credit and a bonus payment, which was contingent on their performance (average payment £2.04, range £1.50-£2.50).

**Stimuli** The task cues were circles ( $6.7^\circ$  of visual angle), filled with: blue (RGB: 0, 0, 255), red (RGB: 255, 0, 0), green (RGB: 0, 255, 0) or yellow (RGB: 255, 255, 0); in the center of the cue, the digit stimulus was then displayed in 60-pt Courier bold font ( $1.3^\circ$  of visual angle). The two sets of digits used were 1,4,7,8 and 2,3,6,9 – these sets were used as on average the values are the same distance from 5 (the criterion value for 'high'/'low'). An iMac was used to display the stimuli using Matlab 2008a with Psychtoolbox.

**Design and procedure** The sequencing was constrained in the same way as for the model, with the addition of a variable CSI that was alternated by blocks to give a long CSI of 1200 ms and a short CSI of 100 ms. For the first block pair, participants were given a piece of paper with correct responses in the format of the relevant Task or CSR diagram (as in Figure 1); in addition participants in the Task condition were given standard task-set instructions verbally and on-screen, whereas participants in the CSR condition were directed to learn cue+stimulus  $\rightarrow$  response mappings on the basis of trial by trial feedback.

After 5 block pairs the second set of four stimuli was introduced in addition to the set already in use. No mention of the new numbers was made prior to their appearance. Participants were debriefed using a structured questionnaire, and replaced if their reported strategy differed from that instructed, i.e. if they induced the tasks in the CSR group, or failed to use the tasks as instructed in the Tasks group. Two participants in the Task group (who did not mention using tasks) and one participant in the CSR group (who induced one of the tasks) were replaced in this way.

### Behavioral Results

The results were analyzed using an ANOVA as for the model, with the additional between-subjects variable of instructions and within subjects variable of CSI.

**Task switches and instruction.** There was a much larger switch cost in the Task group (160 ms) than in the CSR group (18.6ms),  $F(1,30)=16.0$ ,  $p<0.001$  — see Figure 3. The switch costs were reliable for the Task group,  $F(1,15) =$

22.4,  $p<0.001$ , and nearly reliable for the CSR group,  $F(1,15) = 3.24$ ,  $p=0.092$ . For errors, there was a near reliable interaction between instruction group and task switch/repeat,  $F(1,30)=3.13$ ,  $p=0.087$ : the switch cost for the Task group was a reliable 2.9%,  $F(1,15)= 11.9$ ,  $p<0.01$ , and for the CSR group 1.2%, also reliable,  $F(1,15)=5.46$ ,  $p<0.05$ .

**Preparation and instruction.** As Figure 3 shows, preparation reduced the RT switch cost in the Task group from 213 ms (4.5%) in the short-CSI blocks to 107 ms (1.4%) in the long CSI blocks, this was significant in the RTs,  $F(1,15)=6.23$ ,  $p<0.05$  and nearly so in the errors,  $F(1,15)=3.96$ ,  $p=0.065$ . There was no such effect in the CSR group, for whom the switch cost was 16 ms (0.9%) in the short-CSI blocks and 21 ms (1.5%) in the long-CSI blocks  $F<1$ . The interaction was reliable in the RTs,  $F(1,30)= 5.67$ ,  $p<0.05$  and nearly significant in the errors,  $F(1,30)=3.91$ ,  $p=0.057$ . Participants in the Task condition also showed a general preparation effect, whereby if only the task-repeat trials are considered they were faster with a long-CSI (611ms) than with a short-CSI (853ms),  $F(1,15) = 63.8$ ,  $p<0.001$ . For the same contrast the CSR group was slightly, but not reliably, slower in the long-CSI (776ms) than at the short-CSI (745ms) condition,  $F(1,15) = 2.55$ .

**Congruency and instruction.** RT and error rate showed (Figure 3) a much larger effect of congruency in the CSR group (346 ms, 7.4%) than in the Task group (91 ms, 6.1%); the interaction was highly reliable for RTs,  $F(1,30)=23.9$ ,  $p<0.001$ , but not for error rate,  $F<1$ . In separate analyses, the congruency effect was reliable for both the Task group,  $F(1, 15) = 6.26$ ,  $p<0.05$ , for RTs, and  $F(1,15)=33.8$ ,  $p<0.001$ , for errors, and the CSR group,  $F(1,15) = 84.5$ ,  $p<0.001$ , for RT, and  $F(1,15)=11.3$ ,  $p<0.01$ , for errors.

**Switch by congruency.** In agreement with the model the switch cost was larger for incongruent trials for the CSR group (30ms, 2%) than congruent trials (7ms, 0.4%). Similarly for the Task group the switch cost was larger for incongruent trials (161ms, 4.8%) than congruent trials (69ms, 1.1%). There was an overall significant interaction between task switch and congruency in the errors,  $F(1,30)= 10.4$ ,  $p<0.01$ , but not in the RTs. This effect did not differ between the two experimental conditions in the error data or RTs.

**Acquisition.** Overall performance improved from block pair 2 to 5 (Figure 4), and this was reliable in RTs and errors,  $F(2.7,79.6)=43.3$ ,  $p<0.001$ ,  $F(2.7, 79.6)=4.60$ ,  $p<0.05$ . The three-way interaction between block pair, congruency and instructions was significant in the RTs only,  $F(2.7, 79.6)= 7.35$ ,  $p<0.01$  and marginally so in the errors,  $F(2.7, 79.6), 2.23$ ,  $p=0.095$ . Separate analyses revealed a highly significant block pair by congruency interaction in the CSR condition, RT:  $F(2.3,34.6)= 9.40$ ,  $p<0.001$ , errors:  $F(2.3,34.6)=6.94$ ,  $p<0.05$ , but not in the Task group,  $F<1$ . This interaction can be seen in Figure 4 which shows the congruent trials being learnt quickly by the CSR group whilst the incongruent trials took longer to learn, a pattern similar to the predictions made by the model.

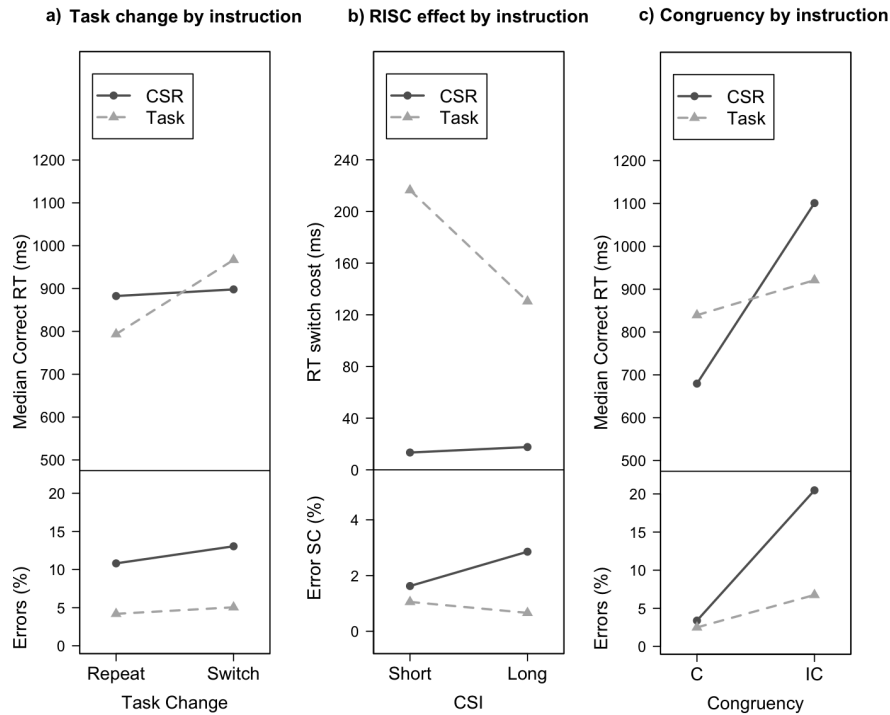


Figure 3 shows switch cost reductions in switch cost and the congruency effect, for the Task and CSR groups.

**Introduction of new stimuli.** Figure 4 also illustrates that the two groups were differentially affected by the introduction of novel stimuli at block pair 6. The new stimuli seem to be accommodated with ease by the Task group, but for the CSR group they clearly cause problems, especially incongruent stimuli. As for the model, the effect of new stimuli was analyzed by comparing the performance on the novel stimuli with that on the old. As expected, the CSR group was more affected by the introduction of new stimuli; their RT (error rate) was much larger for the new stimuli 1201 ms (21.7%) than the old stimuli 946 ms (4.1%), whereas in the Task group performance on the new stimuli, at 780ms (7.5%) was more equivalent to that of the old stimuli, 731 ms (4.8%). This difference was supported by a significant interaction in both the RTs,  $F(1,25)=23.8$ ,  $p<0.001$  and errors,  $F(1,30)=37.7$ .

### Discussion

There was a clear difference in the performance of the two groups. The Task group exhibited a large switch cost, which was substantially reduced by the opportunity to prepare. In contrast, the CSR group had a smaller switch cost, which derived largely from the incongruent stimuli and was unaffected by CSI. The CSR group had a much larger congruency effect, which was modulated with practice because congruent stimuli were learnt much faster than incongruent stimuli. In contrast the Task group exhibited a smaller congruency effect which was much more stable over practice.

These differences in the performance of participants with

and without knowledge of the task-sets suggest that there is merit in theories of performance in task-cuing paradigms that appeal to task-set. However, given that participants who had no knowledge of the tasks showed significant "switch costs" and congruency effects also indicates that these phenomena are not *per se* indices of top-down control of task-set (as Logan & Bundesen, 2003, have also argued, for different reasons). Hence, part of the switch cost seen in the Tasks group might have the same source as for the CSR group, and the congruency effect in the Task group might be an ameliorated version of that seen in the CSR group, with top down task-set control helping to shield against associative interference (Dreisbach & Haider, 2009). However, the marked differences in performance between the groups — the much larger switch cost and its reduction with preparation in the Tasks group, and the much larger congruency effects in the CSR group — clearly suggested a qualitative difference in processing strategy between them. The effects of practice and transfer, with the CSR group's rapid learning of the congruent stimuli and difficulty with the transfer test contrasting with the relatively stable switch costs over practice, and good transfer for the Tasks group, also pointed to a substantial difference in processing strategy between groups, and highlights one of the advantages of a task-set strategy — the ability to generalize to novel cases.

Moreover, the data of the CSR group seem in agreement with the behavior of an associative learning network. All of the effects predicted by the model were present in the CSR

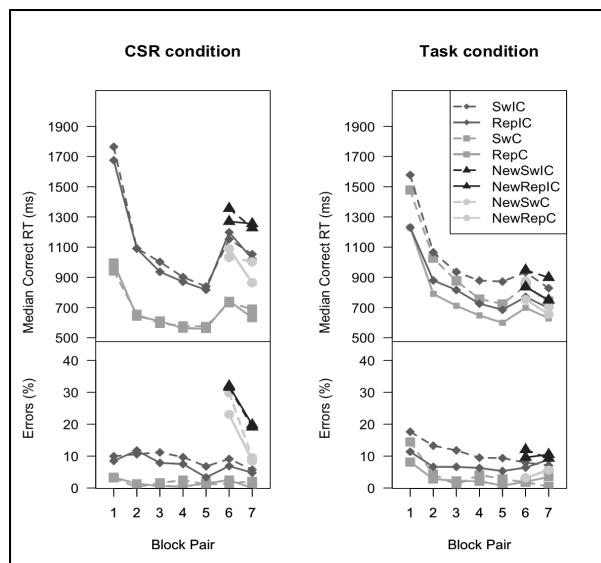


Figure 4 shows the performance over block pairs in the same way as Figure 2.

group: a large congruency effect and its modulation by practice, a modest "switch cost" due mostly to the incongruent stimuli, and a marked disadvantage in coping with new stimuli. This is certainly consistent with the suggestion that this group's performance was dependent on associative learning. We conclude that there is evidence to suggest that when participants perform in a task-cuing paradigm without knowledge of the tasks, they produce a distinctive pattern of results which is in line with the predictions of an associative model. If one is interested in using task-cuing to measure control processes, it may be wise to check for use of a CSR strategy, and to use conditions (e.g. larger stimulus sets) that discourage it.

## REFERENCES

- Allport, D. A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umiltà, & M. Moscovitch (Eds.), *Conscious and Nonconscious Information Processing: Attention and Performance XV* (pp. 421-52). Cambridge, MA: MIT Press.
- Dreisbach, G., & Haider, H. (2009) How task representations guide attention: Further evidence for the shielding function of task sets. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 477-86.
- K. I. Hodder, D. N. George, A. S. Killcross & R. C. Honey (2003): Representational blending in human conditional learning: Implications for associative theory, *The Quarterly Journal of Experimental Psychology Section B*, 56:2, 223-28.
- Honey, R. C., & Ward-Robinson, J. (2002). Acquired equivalence and distinctiveness of cues: I. Exploring a neural network approach. *Journal of Experimental Psychology: Animal Behavior Processes*, 28, 378-87.
- Kiesel, A., Wendt, M., & Peters, A. (2007). Task switching: on the origin of response congruency effects. *Psychological Research* 71, 117-25.
- Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A.M. & Koch, I. (2011) Control and Interference in Task Switching—A Review. *Psychological Bulletin*.
- Le Pelley, M.E. and McLaren, I.P.L. (2003) Retrospective revaluation in humans: Learning or memory? *The Quarterly Journal of Experimental Psychology*, 54B, 311-52.
- Livesey, E.J., Thorwart, A., De Fina, N.L. & Harris, J.A. (2011) Comparing Learned Predictiveness Effects Within and Across Compound Discriminations. *Journal of Experimental Psychology Animal Behavior Processes*, 37, 446-65.
- Logan, G. D., & Bundesen, C. (2003). Clever homunculus: Is there an endogenous act of control in the explicit task-cuing procedure? *Journal of Experimental Psychology: Human Perception and Performance*, 29, 575-99.
- Mayr, U. (2001). Age differences in the selection of mental sets: The role of inhibition, stimulus ambiguity, and response-set overlap. *Psychology and Aging*, 16, 96-109.
- Meiran, N., Levine, J., Meiran, N. & Henik, A. (2000). Task-set switching in schizophrenia. *Neuropsychology*, 14, 471-82.
- McLaren, I.P.L. (1994). Representation development in associative systems. In J.A. Hogan & J.J. Bolhuis (Eds.), *Causal Mechanisms of Behavioural Development* (pp. 377-402). Cambridge, UK: Cambridge University Press.
- McLaren, I.P.L., Green, R.E.A. & Mackintosh, N.J. (1994) Animal Learning and the Explicit/Implicit Distinction. In N.C. Ellis (Ed.), *Implicit and Explicit Learning of Languages*. London: Academic Press.
- McLaren, I.P.L. (2011). APECS: An adaptively parameterised model of associative learning and memory. In Alonso, E. & Mondragón, E. (Eds.), *Computational Neuroscience for Advancing Artificial Intelligence: Models, Methods and Applications*. Hershey, PA: IGI Global.
- Monsell, S. (2003) Task switching. *Trends in Cognitive Science* 7, 134-40.
- Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, 124, 207-231.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York, NY: Appleton-Century-Crofts.
- Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart and J.L. McClelland (Eds.) *Parallel Distributed Processing*. Vol.1. Cambridge, Mass. Bradford Books.
- Saavedra, M. A. (1975). Pavlovian compound conditioning in the rabbit. *Learning and Motivation*, 6, 314-26.
- Schneider, D. W., & Logan, G. D. (2005) Modeling task switching without switching tasks: A short-term priming account of explicitly cued performance, *Journal of Experimental Psychology: General*, 134, 343-367.

# Early effects of word surprisal on pupil size during reading

**Stefan L. Frank (s.frank@ucl.ac.uk)**

Department of Cognitive, Perceptual and Brain Sciences  
University College London  
26 Bedford Way, London WC1H 0AP, United Kingdom

**Robin L. Thompson (robin.thompson@ucl.ac.uk)**

Deafness, Cognition and Language Research Centre  
Department of Cognitive, Perceptual and Brain Sciences  
University College London  
49 Gordon Square, London WC1H 0PD, United Kingdom

## Abstract

This study investigated the relation between word surprisal and pupil dilation during reading. Participants' eye movements and pupil size were recorded while they read single sentences. Surprisal values for each word in the sentence stimuli were estimated by both a recurrent neural network and a phrase-structure grammar. Higher surprisal corresponded to longer word-reading time, and this effect was stronger when surprisal values were estimated by the neural network. In addition, there was an early, positive effect of surprisal on pupil size, from about 250 ms before word fixation until 100 ms after fixation. This early effect, which was only significant for the network-based surprisal estimates, is suggestive of a preparation-based account of surprisal.

**Keywords:** Reading; Eye tracking; Pupillometry; Sentence comprehension; Surprisal; Recurrent neural network; Phrase-structure grammar

## Introduction

Language comprehension is mostly incremental: When listening to or reading a sentence, each word is immediately integrated with information from the sentence so far (e.g., Just, Carpenter, & Woolley, 1982). It has been argued that the amount of cognitive effort required to process a given word can be quantified by its *surprisal* (Hale, 2001; Levy, 2008), an information-theoretic measure of the extent to which the word's occurrence was unexpected. Formally, if  $w_{1..t}$  denotes the sentence's first  $t$  words, the surprisal of the following word is:  $\text{surprisal}(w_{t+1}) = -\log P(w_{t+1}|w_{1..t})$ . These values can be estimated by any language model that assigns probabilities to word sequences.

The relationship between surprisal and cognitive load (i.e., relative difficulty in processing) has indeed been observed in reading studies: Words with higher surprisal values take longer to read, which accounts for several phenomena in sentence comprehension, such as garden-path effects (Brouwer, Fitz, & Hoeks, 2010) and anti-locality effects (Levy, 2008). More generally, reading times at each word in sentences or texts have been shown to correlate with surprisal (e.g., Boston, Hale, Patil, Kliegl, & Vasishth, 2008; Demberg & Keller, 2008; Fernandez Monsalve, Frank, & Vigliocco, 2012; Frank & Bod, 2011; Smith & Levy, 2008).

Here, we investigate an alternative empirical index of cognitive load; one that can be measured continuously and with precise time-resolution: pupil size. By analyzing how and

when effects of word surprisal appear in pupillometry data, we are able to use a physiological measure to investigate the fine-grained time course of sentence-comprehension processes.

A large number of studies, using a variety of tasks, have looked at the relationship between cognitive load and pupil dilation (for a recent overview, see Laeng, Sirois, & Gredebäck, 2012). Although these studies differ in how cognitive load is operationalized, increased cognitive load is invariably found to result in larger pupil size. In a non-linguistic context, Preuschoff, 't Hart, and Einhäuser (2011) showed that pupil size (and therefore, presumably, cognitive load) increases when a stimulus is less expected. They had participants perform a simple gambling task and found that experiencing surprise causes pupil dilation: Pupil size correlated not with the gambling outcome itself but with its unexpectedness.

Whether unexpectedness of words in sentences also results in pupil dilation is still an open question. In fact, there has been only very little pupillometry research in psycholinguistics. Engelhardt, Ferreira, and Patsenko (2010) found that a mismatch between syntactic and prosodic structure of auditorily presented sentences results in larger pupil size compared to a condition in which the two structures matched. In another sentence-listening study, Piquado, Isaacowitz, and Wingfield (2010) found a pupil response to both syntactic complexity and sentence length. To the best of our knowledge, there exists only two published studies in which pupillometry is applied during sentence reading: Raisig, Hagendorf, and Van der Meer (2012) presented participants with written descriptions of simple events in everyday activities and found increased pupil dilation when the order of presentation was incongruent with the actual temporal order of the described activities. Just and Carpenter (1993) compared object- and subject-relative clauses and found increased reading times and pupil dilation on the object-relatives, which have long been known to be more difficult to process (Hakes, Evans, & Brannon, 1976). Moreover, the occurrence of a semantically implausible word resulted in increased pupil size compared to a plausible-word condition.

Here, we did not compare particular sentence pairs but, instead, investigate the general relation between word surprisal

and pupil size, looking for effects on each word within a large set of visually-presented sentences. The goals of this study were to explore pupillometry as a methodology for investigating sentence-comprehension processes during reading; to uncover the time-course of surprisal effects; and to assess the suitability of two very different model types for surprisal estimation: recurrent neural networks (RNNs) and phrase-structure grammars (PSGs). We found a very early, positive effect of surprisal on pupil size, which was only significant for the surprisal values generated by the RNN. These findings suggest that surprisal effects are caused by a process of word prediction rather than word integration.

## Method

### Eye tracking and pupillometry

**Materials** The self-paced reading study by Fernandez Monsalve et al. (2012) and Frank (2012) used 361 sentence stimuli, semi-randomly selected from three novels published on [www.free-online-novels.com](http://www.free-online-novels.com). Two hundred and five of these sentences (comprising 1931 word tokens) could fit on a single line of the display and were therefore used in the current eye-tracking experiment. Of those 205 sentences, 110 had a corresponding yes/no comprehension question to ensure that subjects were reading for meaning.

**Participants** Seventeen monolingual, native English speakers were recruited from the University College London subject pool. One participant was excluded due to technical issues, leaving 16 participants (11 women, mean age 27.6) with analyzable data.

**Procedure** Subjects were seated 50 cm from the monitor with their chin on a chin rest. Both eyes were tracked using a head-mounted eye-tracker (SR Research, EyeLink II). Individual sentences were presented in 18-point Courier font, left-aligned on the display. Each sentence was preceded by a left-aligned fixation cross that was presented for 800 ms. Gaze direction and pupil area were sampled at a rate of 500 Hz.

After initial calibration (nine fixation points) and five practice trials, subjects were invited to ask clarification questions and the experiment began. Another calibration check was performed after the practice items and then again after every 35 trials (the final set had only 30 trials), at which time subjects took a self-paced break (total 205 trials, six sets). Additionally, drift correction on a single centrally located fixation point was performed at the start of each trial. Responses were recorded using a mouse (center button to continue after finishing a sentence; right and left buttons to respond 'yes' or 'no', respectively, to comprehension questions). The entire experiment (with instructions and calibration) took approximately 50 minutes to complete. The order of trial presentation was randomized throughout.

### Surprisal estimation

For each word in the experimental sentences, surprisal values were generated by the same set of probabilistic language

models as used by Fernandez Monsalve et al. (2012). All models were trained on 702,412 sentences (comprising 7.6 million word tokens; 7,754 word types) from the written-text part of the British National Corpus.

**Recurrent neural network** Although RNNs are often used for learning the statistics of language, they are nearly always applied to artificial toy languages. Training such models on a large, English-language corpus, as we do here, requires something more advanced than the standard Simple Recurrent Network (SRN; Elman, 1990). The solution was to first encode each word as a distributed vector and train the network on sequences of those word representations. More precisely, network training was divided into three distinct stages (see also Fernandez Monsalve et al., 2012; Frank, 2012):

1. A co-occurrence matrix  $\mathbf{P} = (p_{ij})$  was constructed, where each  $p_{ij}$  is the (smoothed) probability that word types  $i$  and  $j$  occur adjacently in the training data. These values were then transformed into  $q_{ij} = \log p_{ij} - \log(\sum_k p_{ik} \sum_k p_{kj})$ . The 400 columns of  $\mathbf{Q}$  with highest variance were selected, and formed the 400-dimensional vectors for each of the 7,754 word types. This representational space captures the paradigmatic relations between words (e.g., words of the same syntactic category tend to receive similar representations), which boosts generalization to untrained input.
2. The 702,412 training sentences, in the form of word-vector sequences, were given as input to an SRN that learned to predict the vector representation of the upcoming word  $w_{t+1}$  after each sentence-so-far  $w_{1...t}$ . The SRN used standard backpropagation and received the complete training corpus five times.
3. A two-layer feedforward network with 200 hidden units learned to 'decode' the SRN's output vectors into localist representations, that is, into 7,754-dimensional vectors where each element corresponds to a word type. It received the training data two times and, like the SRN, used standard backpropagation for connection-weight update. Its output units have softmax activation functions, so each output vector forms a probability distribution over word types.

The complete model, combining these three stages, generates estimates of the probabilities  $P(w_{t+1}|w_{1...t})$  for all word types, from which the surprisal of the actual next word follows directly. These surprisal values were obtained at ten intervals during training of the decoder network, resulting in ten sets of surprisal estimates (by an increasingly well-trained model) each of which was analyzed independently.

**Phrase-structure grammar** Grammars are usually not induced from 'flat' word sequences but require complete syntactic tree structures as training material. It was therefore necessary to first obtain such structures by parsing the training sentences. This was done by the Stanford Parser (version 1.6.7; Klein & Manning, 2003). The resulting collection of tree structures served as the PSG training corpus.



In a standard probabilistic context-free grammar, the probability of a production rule is conditioned on the rule’s left-hand side. For example, the rule ‘NP  $\rightarrow$  Det N’ would be associated with the probability that a phrase consists of a determiner (Det) followed by a noun (N), given that it is a noun phrase (NP). A grammar’s structural sensitivity can be increased by also conditioning on other parts of the tree structure, for example, by estimating the probability of ‘Det N’ given that the current phrase is an NP that belongs to a verb phrase. In this manner, many different grammars, with different structural sensitivities, can be induced from the same set of training data. Here, we applied Roark’s (2001) grammar-induction algorithm to obtain eight different grammars (see also Fernandez Monsalve et al., 2012; Frank & Bod, 2011). Next, an incremental parser (Roark, 2001) processed the experimental sentences. At each word, it computed the probabilities of possible syntactic structures<sup>1</sup> (under each of the eight grammars) given the sentence-so-far  $w_{1..t}$ . The sum of those probabilities equals  $P(w_{1..t})$ , and surprisal values follow because  $-\log P(w_{t+1}|w_{1..t}) = \log P(w_{1..t}) - \log P(w_{1..t+1})$ . That is, for each word we obtain eight grammar-based surprisal values, in addition to the ten RNN-based surprisals discussed above.

## Results

All participants displayed adequate comprehension by answering at least 80% of the comprehension questions correctly. We excluded from consideration the first and last word of each sentence, clitics, words attached to a comma, the first fixated word, and non-fixated words. Further, data corresponding to fixations outside the sentence presentation region, as well as regressions (i.e., fixations to words earlier in a sentence after a fixation on a later word) were discarded.

### Word-reading time

**Analysis** As a measure of word-reading time, we took total fixation time on a word before fixation on any other word (i.e., the first-pass reading time, or gaze duration; av. 231 ms, s.d. 116 ms). A mixed-effects regression model was fitted to this dependent variable (14,304 data points), using as predictor variables: sentence presentation order (both linear and quadratic factors), word position in sentence (linear and quadratic), word length, log of word frequency, and log of forward transitional probability (i.e., the word’s probability given the previous word). Also, all significant two-way interactions were included,<sup>2</sup> as were all significant random slopes of main effects.<sup>3</sup>

The effect size of surprisal is defined as the decrease in regression model deviance when surprisal is included as an

<sup>1</sup>The least probable structures were ignored to make this computation feasible.

<sup>2</sup>These were determined by first including all two-way interactions and then removing the least significant ones until all  $|t| > 2$ .

<sup>3</sup>These were a by-item slope of sentence order and by-subject slopes of all factors except forward probability and quadratic sentence order.

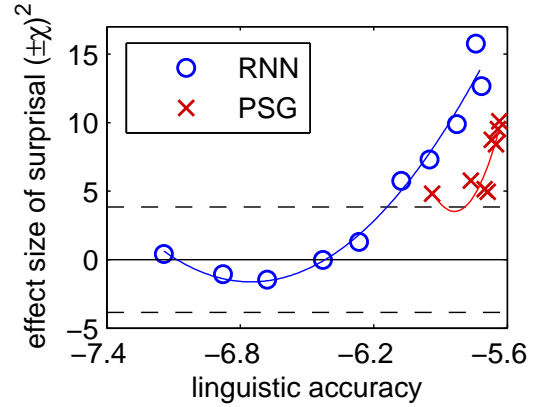


Figure 1: Effect of surprisal (as estimated by either RNN or PSG) on gaze durations as a function of linguistic accuracy (average  $\log P(w_{t+1}|w_{1..t})$ ). Plotted are the estimated  $\chi^2$ -statistics (where negative values denote effects in the negative direction) and best fitting second-degree polynomials. The dashed lines at  $\chi^2 = \pm 3.84$  denote the level beyond which  $p < .05$ .

additional predictor. This quantity is the  $\chi^2$ -statistic of a log-likelihood test for the significance of surprisal. Effect size can be contrasted with ‘linguistic accuracy’: the extent to which the model has learned the statistical patterns of the language. Linguistic accuracy is quantified as the average of  $\log P(w_{t+1}|w_{1..t})$  estimated over the experimental sentences, weighted by the number of times  $w_{t+1}$  occurs in the analysis.

**Surprisal effect** Figure 1 plots the size and direction of the surprisal effect as a function of linguistic accuracy. As expected, all the statistically significant effects are in the positive direction: More surprising words take longer to read. Moreover, models that capture the statistics of the language more accurately also account for more variance in reading time.

Surprisal as estimated by the RNN model (after sufficient training) shows stronger effects than does PSG-based surprisal. We compared the RNN and PSG that showed the strongest effects by testing whether one set of surprisal estimates had an effect over and above the other. The RNN’s surprisals did have an additional effect over the PSG’s ( $\chi^2 = 7.6$ ;  $p < .01$ ) but the reverse was not the case ( $\chi^2 = 1.93$ ;  $p > .15$ ). That is, the grammar does not yield surprisal values that explain any unique variance in reading times.

### Pupil size

**Analysis** As the eyes move across the screen, the angle between the eye gaze and camera changes, affecting the observed pupil size. This was corrected for by fitting a second-degree polynomial to the measured pupil sizes during saccades as a function of the horizontal gaze direction. Correction was performed for each presentation block (i.e., between recalibrations), participant, and individual eye (left or right).

The fitted values then served as a baseline of pupil size at each horizontal location on the display. If both eyes were successfully tracked, pupil size was averaged over the two. For each subject and sentence separately, pupil sizes were then rescaled to a percentage of the average over the sentence.

The effect of word surprisal on pupil size was analyzed at every 2 ms sample (i.e., the sampling rate of the eye-tracker), from 500 ms before the first fixation on a word, up to 1000 ms after that fixation. If any pupil size during that 1500 ms time window was below 70% or above 130%, the data for those 1500 ms were discarded.

When we analyzed reading times, a baseline regression model was fitted to the gaze durations. In the case of pupil sizes, however, it is not possible to fit just one baseline model because the values of the dependent variable differ across samples. Alternatively, a different model could be fitted to each sample but that would make it impossible to track the surprisal effect over time. Therefore, the same, simplified baseline model is used for all samples. It contained the main effects from the reading-time analysis, except that the factor ‘word position’ was replaced by the fixated letter’s position in the sentence (both linear and quadratic factors). Letter position allows us to take into account differences in luminosity across the display, which can affect pupil dilation. In addition, because samples of pupil dilation are taken up to 500 ms before fixation on the current word, the length, log frequency and log forward probability of the *previous* word are also included. As before, the effect size of surprisal was defined as the decrease in regression model deviance due to surprisal. Surprisal estimates were taken from the RNN and PSG model that explain the most variance in gaze duration.

**Surprisal effect** Figure 2 shows how strongly a word’s surprisal affects pupil size, time-locked to the moment of first fixation on that word. There is a positive relation between surprisal and pupil size, which arises very early, even before fixation (i.e., parafoveally).

Considering that the effect of a word’s surprisal arises before fixation on that word, it makes sense to discard cases in which the previous word was not fixated. Specifically, it is unlikely that enough information about a word can be obtained if it is still more than one word ahead. Indeed, as shown in Figure 3, the effect of surprisal remains as strong even when we only take into account cases in which there is a fixation on the previous word (in spite of a 30.3% reduction in the amount of data).

**Entropy effect** Alternatively, the early effect of surprisal could be due to readers’ uncertainty about the upcoming word.<sup>4</sup> If uncertainty about  $w_{t+1}$  correlates positively with its surprisal, and being in a state of increased uncertainty causes the pupils to dilate, then the apparent effect of surprisal may actually be an effect of uncertainty. Such an effect can appear during processing of  $w_t$ , without any information about the

upcoming word  $w_{t+1}$ .

We investigated this possibility by estimating how much uncertainty about  $w_{t+1}$  a reader may experience after processing  $w_{1...t}$ . In information theory, uncertainty about the value of a random variable is quantified by its *entropy*. In the context of incremental sentence comprehension, the uncertainty about  $w_{t+1}$  is defined as:

$$H(w_{t+1}) = - \sum_{w_{t+1}} P(w_{t+1}|w_{1...t}) \log P(w_{t+1}|w_{1...t}).$$

The entropy  $H(w_{t+1})$  is based on the probability distribution  $P(w_{t+1}|w_{1...t})$ , which is exactly the output of the RNN model. Note that, unlike the word’s surprisal, the entropy over  $w_{t+1}$  does not require knowledge of the actual upcoming word  $w_{t+1}$ . Crucially,  $H(w_{t+1})$  equals the expected value of surprisal( $w_{t+1}$ ) so the two values correlate positively ( $r = .38$  in our data set). A positive effect on pupil dilation of uncertainty about  $w_{t+1}$  could therefore be misinterpreted as an effect of the surprisal of  $w_{t+1}$ . However, as Figure 4 shows, the relation between entropy (as estimated by the RNN) and pupil size is (if anything) negative. Consequently, the effect of surprisal in Figure 3 is not an entropy effect in disguise.

## Discussion

Our reading-time results corroborate earlier findings: More surprising words take longer to read; this effect grows stronger as surprisal values are estimated by linguistically more accurate models; and RNN-based surprisals account for more variance than do grammar-based estimates. Like Frank and Bod (2011), we found no additional effect of the grammar-based surprisals. However, applying the same surprisal estimates to data from a self-paced reading study, Fernandez Monsalve et al. (2012) did find an additional effect of the PSG’s surprisals, possibly because their data set was almost ten times larger than our current set.

Importantly, predictions by computational models of language have never before been applied to the analysis of pupil-lometric data. Hence, the effect of word surprisal on pupil size had not yet been demonstrated. This effect confirms that surprisal is indeed a cognitively relevant measure of processing load, and not merely of processing *time*.

Two explanations have been proposed for the relation between word surprisal and cognitive load: According to Levy (2008), integrating a new word into the interpretation of the sentence so far comes down to updating a probability distribution over all possible sentence interpretations. He proves that the extent of this update, expressed as the Kullback-Leibler divergence from the old distribution to the new, equals the word’s surprisal. Alternatively, Smith and Levy (2008) argue that the surprisal effect is due to the reader’s processing system being more prepared for more expected words. Under that account, we may expect surprisal effects to occur sooner than if they result from integration of the new word with the current sentence interpretation. Therefore, the very early, pre-fixation effect we found here seems most compatible with Smith and Levy’s preparation account.

<sup>4</sup>We would like to thank an anonymous reviewer for this suggestion.



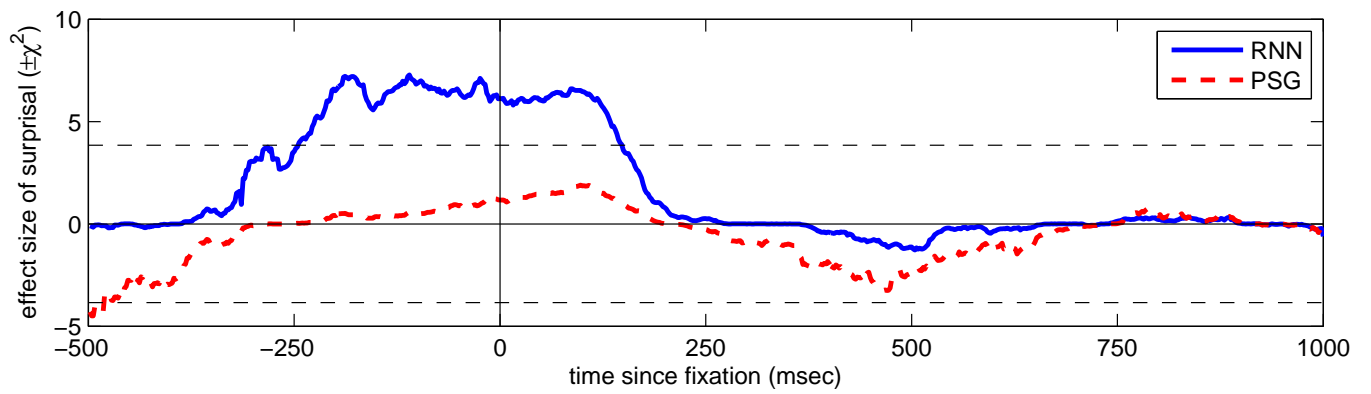


Figure 2: Effect of surprisal on pupil size over time.

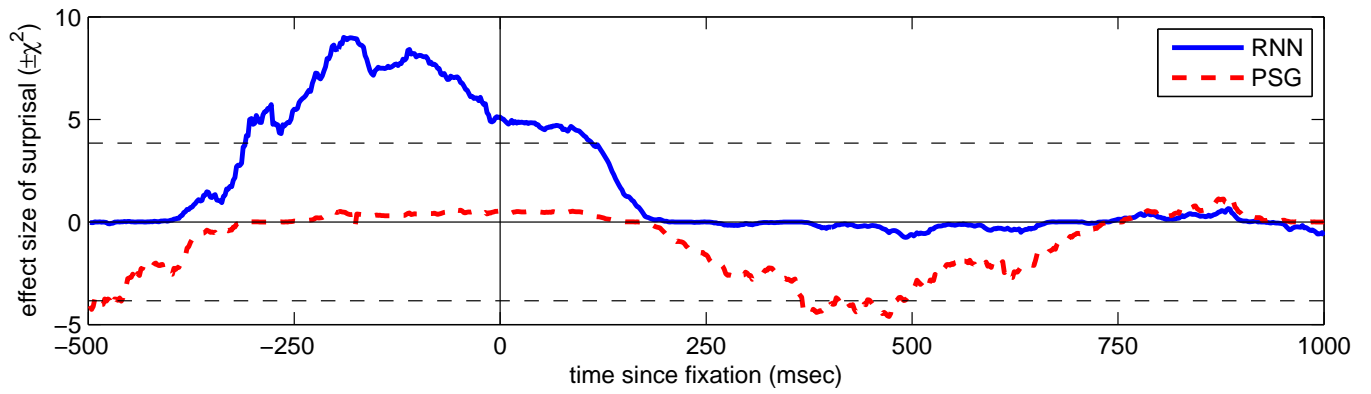


Figure 3: Effect of surprisal on pupil size over time, taking only cases where the previous word was fixated.

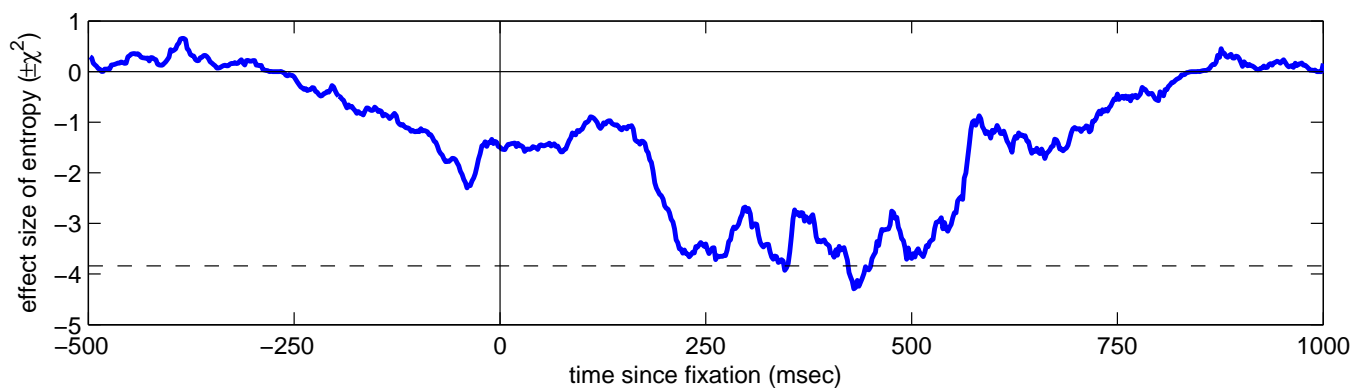


Figure 4: Effect of entropy (uncertainty about the upcoming word) on pupil size over time.

The early occurrence of a surprisal effect may also explain why only the RNN predicted pupil size. Presumably, RNNs simulate early, predictive processing whereas applying a PSG (i.e., parsing) generates syntactic structure and therefore models later ‘integrative’ processing. Hence, an early effect on pupil size that does not depend on integrative processing would only be predicted by RNNs and not by PSGs.

## Conclusion

A word’s surprisal has a very early effect on pupil size during reading: At about 250 ms *before* the word is fixated, its surprisal is a significant predictor of the reader’s pupil size. This suggests that surprisal effects are due to preparation (Smith & Levy, 2008) rather than integration (Levy, 2008). Moreover, it may explain why surprisal estimates by RNNs have a stronger effect than those from PSGs. Perhaps more importantly, however, we have established that pupillometry is a viable paradigm for studying the fine-grained time course of reading processes.

## Acknowledgments

The research presented here was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant number 253803, and by a grant from the Economic and Social Research Council of Great Britain (RES-620-28-6001) awarded to the Deafness Cognition and Language Research Centre. We are grateful to Naima Ansari for her assistance with data collection.

## References

- Boston, M. F., Hale, J., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2, 1–12.
- Brouwer, H., Fitz, H., & Hoeks, J. (2010). Modeling the noun phrase versus sentence coordination ambiguity in Dutch: Evidence from surprisal theory. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (pp. 72–80). Uppsala, Sweden: Association for Computational Linguistics.
- Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109, 193–210.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14, 179–211.
- Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology*, 63, 639–645.
- Fernandez Monsalve, I., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398–408). Avignon, France: Association for Computational Linguistics.
- Frank, S. L. (2012). Uncertainty reduction as a measure of cognitive processing load in sentence comprehension. *Manuscript submitted for publication*.
- Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22, 829–834.
- Hakes, D. T., Evans, J. S., & Brannon, L. L. (1976). Understanding sentences with relative clauses. *Memory & Cognition*, 4, 283–290.
- Hale, J. T. (2001). A probabilistic Early parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.
- Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology*, 47, 310–339.
- Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111, 228–238.
- Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics* (pp. 423–430).
- Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, 7, 18–27.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106, 1126–1177.
- Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, 47, 560–569.
- Preusschoff, K., ‘t Hart, B. M., & Einhäuser, W. (2011). Pupil dilation signals surprise: evidence for noradrenaline’s role in decision making. *Frontiers in Neuroscience*, 5.
- Raisig, S., Hagendorf, H., & Van der Meer, E. (2012). The role of temporal properties on the detection of temporal violations: insights from pupillometry. *Cognitive Processing*, 13, 83–91.
- Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27, 249–276.
- Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 595–600). Austin, TX: Cognitive Science Society.

# The Plausibility of Semantic Properties Generated by a Distributional Model: Evidence from a Visual World Experiment

Diego Frassinelli (d.frassinelli@sms.ed.ac.uk)

Frank Keller (keller@inf.ed.ac.uk)

Institute for Language, Cognition and Computation

School of Informatics, University of Edinburgh

10 Crichton Street, Edinburgh EH8 9AB, UK

## Abstract

Distributional models of semantics are a popular way of capturing the similarity between words or concepts. More recently, such models have also been used to generate properties associated with a concept; model-generated properties are typically compared against collections of semantic feature norms. In the present paper, we propose a novel way of testing the plausibility of the properties generated by a distributional model using data from a visual world experiment. We show that model-generated properties, when embedded in a sentential context, bias participants' expectations towards a semantically associated target word in real time. This effect is absent in a neutral context that contains no relevant properties.

**Keywords:** Distributional models of semantics; concepts and properties; context effects; eye movements; visual world.

## Introduction

The representation of semantic concepts has been the subject of an intense debate over the last few decades (Murphy, 2002). An emerging consensus is that the internal structure of a concept can be represented as a set of semantic properties (Garrard, Lambon Ralph, Hodges, & Patterson, 2001; Baroni & Lenci, 2008). These properties can be accessed in the form of semantic feature norms elicited from experimental participants (McRae, Cree, Seidenberg, & McNorgan, 2005). In the computational modeling literature, this idea has been taken up by distributional models of semantics. Such models have traditionally been used to compare the similarity of words or concepts. However, recently, a distributional model has been proposed that is able to generate properties associated with a concept (Baroni, Murphy, Barbu, & Poesio, 2010). These properties are computed based on corpus data, and have been shown to overlap with those generated in feature elicitation experiments. Distributional models can therefore be claimed to provide a cognitively plausible representation of concepts in terms of semantic properties.

In the present paper, we propose a novel way of testing this claim using the visual world paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), which allows the study of conceptual processing in real time. We embed the properties generated by Baroni et al.'s model for a given target word into a sentential context. If the model-generated properties are cognitively plausible, then they should bias participants' expectations towards a target word, compared to a competitor word not associated with the properties. As a baseline, we also embed the target and competitor in a neutral context; the contextual expectation effect should be absent in this case.

## Background

The idea of testing the predictions of distributional models using the visual world paradigm goes back to Huettig, Quinlan, McDonald, and Altmann (2006). They were interested in validating the semantic similarity measures generated by two distributional models: Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) and Contextual Similarity (McDonald, 2000). Huettig et al. (2006) demonstrated that the similarity scores generated by both models are significantly correlated with fixation probabilities in a visual world experiment.

Huettig et al. used a list of 26 target/competitor pairs of semantically related but not strongly associated words. In every pair, one of the words corresponded to a target object depicted in a visual scene (the target word); the other one (the competitor word) was semantically related to the depicted object. For every pair of words, a spoken sentence was recorded that contained either the target or the competitor. Huettig et al. focused on the effect of hearing the target vs. the competitor as critical word. For this reason, the context sentences they used were neutral, providing background information that did not bias the participants towards either the target or the competitor. One of their contexts is given in (1) as an example.

- (1) At first, the man laughed loudly, but then he saw the elephant (target)/alligator (competitor) and understood that it was dangerous.

The crucial manipulation in our experiment, however, concerns the context sentence. We run Huettig et al.'s neutral context as a baseline condition, but we add two context conditions: a context containing properties associated with the target, and a context containing properties associated with the competitor. These context sentences were constructed using three properties produced by the distributional model Strudel (Structured Dimension Extraction and Labeling; Baroni et al., 2010). Strudel is a model trained on the lemmatized and part-of-speech tagged version of Ukwac, an English corpus of two billion tokens extracted from the Web (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009).

Strudel shares with other distributional models the assumption that it is possible to represent the meaning of a word in terms of other words that frequently appear in its linguistic context. Unlike traditional distributional approaches, Strudel describes concepts not only in terms of their most frequent

context words, but can also represent a word’s internal structure in terms of semantic properties (e.g., visual features, the functions of an artefact). The output of Strudel is a list of properties linked to the corresponding concept through a pattern describing the relation between the concept and the property. An example is the relation *elephant\_in\_jungle*, in which the concept *elephant* is related to the property *jungle* via the pattern *in*. The set of properties for each concept is computed based on the number of co-occurrences in the corpus, taking into account the number of relevant patterns. The properties that Strudel generates this way are cognitively plausible in the sense that they overlap with human-generated feature norms such as the McRae et al. (2005) norms, as Baroni et al. (2010) demonstrate.

## Experiment

This experiment had two main goals. Firstly, we wanted to test Strudel’s ability to produce semantic properties for concepts. We evaluated this by using the properties to create sentential contexts, which we predict should bias participants towards the target concept. Secondly, we wanted to establish the effect that such contexts have on the processing of the target concept.

Huetting et al. used a neutral context and found that participants are more likely to fixate a target object when they hear its name, but they also show an increased fixation probability for the name of a semantically associated object. We expect this effect to be modulated by context. More specifically, the processing of properties associated with the target should build up an expectation for the target, and as a consequence, there should be more fixations on the target object when the target word is spoken, compared to the neutral context condition. This effect should be attenuated for the competitor, which is distinct from the target, but semantically related (as in Huetting et al.’s design).

## Method

**Materials** The visual world paradigm requires both visual and linguistic stimuli. We used the same visual scenes as Huetting et al. Each scene contained black and white line drawings of the target object and three distractors; the pictures were extracted from the Snodgrass and Vanderwart (1980) collection. Huetting et al. removed phonological competitors and matched the pictures according to naming and image agreement, familiarity, visual complexity, and word frequency of the correspondent noun. Moreover, they tested the visual similarity between pictures. In our experiment, we used the same scenes used in the original experiment: this allowed us to skip the norming process.

We used the same linguistic materials as Huetting et al. for the neutral context condition. We added to this two context conditions: one for the target concept, and one for the competitor. For each of the 52 concepts (26 competitor/target pairs) in the Huetting et al. materials, we extracted from the output of Strudel the first 20 semantic properties

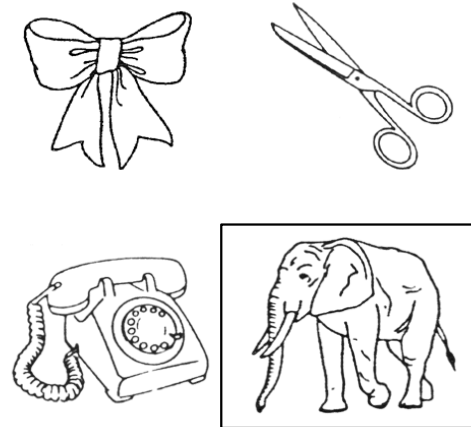


Figure 1: Example scene for the pair elephant (target)/alligator (competitor) in the experiment. The box highlights the target object (not shown to participants).

(nouns, verbs, and adjectives) ordered according to their log-likelihood ratio. We constructed a context sentence for each concept using three of these properties (excluding those associated with words that are part of the same target and competitor pair).

The context sentences had a standard pattern: a temporal subordinate clause introducing the situation followed by the main clause. The target concept is embedded at the end of the main clause and followed by an adverb (which serves as a spill-over region for the analysis). As an example, Figure 1 depicts the scene associated with the pair elephant (target)/alligator (competitor). The sentences associated with this scene are:

- (2) Neutral Context: At first, the man laughed loudly, but then he saw the **elephant** and understood that it was dangerous.
- (3) Target Context: While the man was crossing the *jungle*, he saw a *poacher capturing* an **elephant** ferociously.
- (4) Competitor Context: While the man was crossing the *swamp*, he saw a hippo *attacking* a *gigantic* **elephant** ferociously.

The critical word is given in **bold**; the properties are in *italics*. For every sentence there was also a counterpart that included the competitor word (in this case **alligator**), resulting in six conditions in total.

The quality of the materials was evaluated in two norming studies performed using Amazon Mechanical Turk. In a sentence plausibility judgment task, 33 native English speakers rated the sentences on a scale from 1 (completely implausible) to 7 (completely plausible). The mean rating for the con-

cept in the sentence with the corresponding properties was 5.67 ( $SD = 0.63$ ) and in the opposite sentence, it was 4.70 ( $SD = 1.07$ ); the opposite sentences were created by swapping the critical words across conditions (target for competitor and vice versa). An Anova showed no main effects, but a significant interaction of concept (target or competitor) and sentence (target or competitor) ( $F_1(1, 35) = 27.86, p < .001$ ;  $F_2(1, 32) = 53.81, p < .001$ ).

In a sentence completion task, we removed the critical words from the sentences and asked 21 participants to complete each of the 52 sentences (two groups of 36 sentences) by typing the most plausible noun. After a process of synonym reduction, we counted the number of occurrences for each word. Good sentences had to elicit primarily the nouns they were associated with and only a small percentage of competitor or unrelated words.

The combination of these two norming studies was used to ensure that a given context was sufficiently associated with the target word, and not with the competitor word. Based on the norming data, we excluded eight pairs of concepts: these were cases in which Strudel had produced properties for a different sense of the word than the one in the Huettig et al. materials, as well as cases in which the target sentences were too different from the competitor ones so that the properties could not be plausibly swapped.

The sentence materials were recorded by a native English speaker at a normal speech rate for presentation in the experiment.

**Procedure** The entire experiment included 108 sentences: 18 word pairs (36 words in total) embedded in a neutral context and two biasing contexts. We rotated the position of the four objects on the screen to control for order or position effects. In total we therefore obtained 432 distinct items that we split in 24 lists of 18 items. The distribution of items across lists was based on a Latin square design, ensuring that each list included exactly one word from each target/competitor pair. Twenty-five filler items were added and a random presentation order generated for each list.

Twenty-four native English speakers from the University of Edinburgh were paid five pounds for taking part in the experiment. Each participant saw the items of one of the 24 lists, randomly interspersed with nine yes/no questions about the sentence or the scene. The questions were there to ensure that participants paid attention throughout the experiment.

Participants were seated in front of a 21" multi-scan monitor with a resolution of 1024 x 768 pixels and their eye movements were recorded using an EyeLink II head-mounted eye-tracker with a sampling rate of 500 Hz. Only the dominant eye was tracked. At the beginning of the experiment and after every ten trials, the eye-tracker was recalibrated using a nine-point randomized calibration. Before each trial, drift correction was performed. At the beginning of each trial the scene appeared on the screen, and the sentence began to play at the same time; the scene disappeared after 1500 ms after the end

of the sentence. The experiment was explained using written instructions and preceded by practice trials. The instructions asked participants to listen carefully to the sentences and look wherever they wanted on the screen. The experiment lasted approximately 30 minutes.

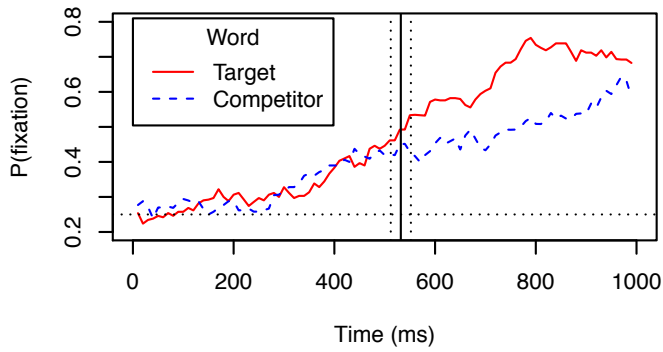
## Results and Discussion

**Fixation Probabilities** Our analysis is based on the fixations on the target object compared to the fixations on the three distractor objects on the display. We excluded out-of-screen fixations and blinks from the analysis. Figure 2 plots the probability of fixating the target object across the three context conditions. The neutral context condition used the sentences of Huettig et al.; the target and competitor conditions used the contextually biased sentences produced based on the Strudel properties. In each plot, 0 ms corresponds to the acoustic onset of the critical word; our analysis takes into account the first 1000 ms after this onset. The vertical line shows the average offset of the critical words, with confidence intervals. The horizontal line at .25 indicates the probability of randomly fixating one of the four objects.

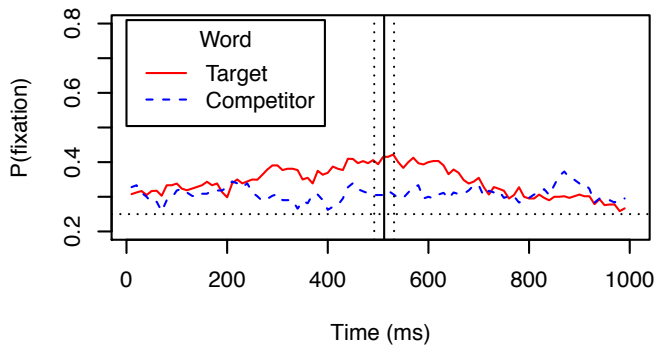
An inspection of the plots reveals a broadly similar trend across the three context conditions. The critical words require some time before they are recognized, which means that the fixation probabilities for the target and the competitor words take between 200 and 500 ms before they diverge. After that, we observe an increase in fixations to the target word compared to the competitor. The point of divergence is about 200 ms later in the neutral context; a semantically related context seems to aid the recognition of the critical word and triggers early fixations to the corresponding object. (Bear in mind that the competitor context is also semantically related to the target, as our norming studies showed.)

In the neutral context condition (Figure 2(a)), we observe a steady increase in fixation probability for both the target and the competitor word, which start to diverge at the offset of the critical word (this is presumably the point at which the critical word has been recognized by the participants). From that point on, we see more fixations on the target than on the competitor. This is in line with what Huettig et al. (2006) found: a competitor word triggers fixations to a semantically related target object, but less fixations than the target word corresponding to the target object. Our neutral context condition therefore provides a replication of Huettig et al.'s results. (The original paper also showed that the difference in fixation probability between target and competitor correlates with their semantic similarity, but we will not test this claim.)

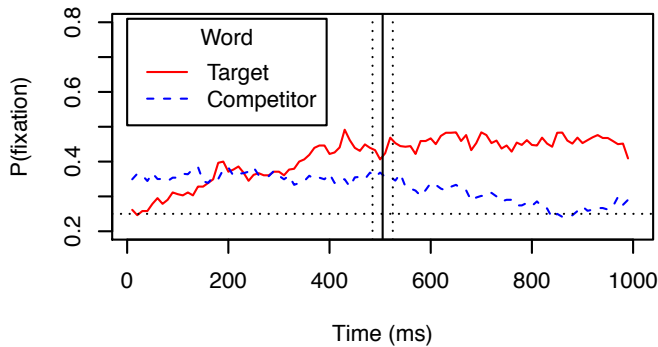
In the target context condition (Figure 2(b)), participants had heard a sentence containing properties of the depicted objects. Presumably this enables them to predict the target word with some accuracy (and our sentence completion study confirmed this). As the target is expected (and hence less interesting) at this point, we only observe a small increase of fixation probability for the target compared to the competitor, which starts early, at around 200 ms. This early start is consistent with the fact that participants are able to predict the critical



(a) Fixation probability in a *neutral context* sentence.



(b) Fixation prob. in a sentence associated with the *target* object.



(c) Fixation prob. in a sentence associated with the *competitor* object.

Figure 2: Fixation probabilities on the target object over time for the target (continuous red line) and competitor (dotted blue line) words. The onset of the critical word is at 0 ms. The vertical lines indicate the mean of the offset of the critical word with confidence interval. The horizontal line shows a probability of .25 (random baseline for four objects).

word in this condition based on the context sentence.

In the competitor context condition (Figure 2(c)), participants had heard a context sentence that is not directly associated with the depicted target object, but is instead associated with the semantically related competitor. In this case, hearing the target word (rather than the contextually appropriate competitor word) is unexpected, i.e., it generates interest and

a larger increase in the number of fixations compared to the competitor word. This means that the two conditions diverge more than in the target context condition, and the divergence remains high for the whole period of analysis.

**Inferential Statistics** To statistically analyze the effect of the experimental manipulation on participants' fixations, we adopted the framework of linear mixed effect models (LME, Baayen, Davidson, & Bates, 2008). As suggested by Barr (2008), the dependent variable was the empirical logit of the fixation probability, calculated for each bin as:

$$\text{emplog} = \log \left( \frac{Y + .5}{N - Y + .5} \right)$$

where  $Y$  is the number of fixations on the target object and  $N$  is the total number of fixations in the bin.

Our model included the factor *Word* representing the nature of the critical word, coded as *Competitor* =  $-.5$  and *Target* =  $.5$ . To determine context effects, we included two factors in contrast coding: the factor *Context* coded the difference between the neutral context =  $-.5$  and the biasing context =  $.25$  conditions; the factor *TargetSentence* differentiated the biasing context sentences further by distinguishing *Competitor* =  $-.5$  and *Target* =  $.5$ . We have also included *Region* as a factor that indicates if the bin is in the critical region (coded as  $-.5$ ) or in the region after the offset of the critical word (coded  $.5$ ). Finally, the continuous predictor *Time* was discretized into 10 ms bins (range 1–100).

The random effects we included were *Participant* and *Item*, which were intercepts in the model. We also included random slopes for all the main effects (*Word*, *Context*, *TargetSentence*, *Region*, and *Time*). We used the model selection procedure of Coco and Keller (2012) to find the minimal model that best fits our data. Table 1 gives the coefficients and significance levels for the minimal model; main effects or interactions not listed in this table were not included in the minimal model by the selection procedure.

**Effect of Context** The factor *Context* compares fixation probabilities in the neutral context and in the biasing context, collapsing the competitor and the target context in the biasing context condition. We find a significant, positive main effect of this factor, suggesting that participants make more fixations on the target object in the biasing context condition. This is modulated by a negative interaction *Time:Context*, which indicates that fixation probability increases over time in the neutral context condition. This explains the upwards trend in Figure 2(a), but not in the biasing context conditions (Figures 2(b) and 2(c)).

While there is no general effect of whether the context is the competitor or the target sentence (no main effect of *TargetSentence*), we do find a significant positive interaction *Time:TargetSentence*. This confirms that there is a larger increase in fixations to the target object in the target context compared to the competitor context.

Table 1: Coefficients for the mixed effects model for the data in Figure 2.

Predictor	Coefficient
(Intercept)	-1.15***
Time	0.17*
Context	0.80*
Time:Context	-0.64***
TargetSentence	-0.47
Time:TargetSentence	0.11**
Word	-0.06
Time:Word	0.18***
Region	0.09
Region:Context	-0.41**
Region:TargetSentence	-0.61***
Word:TargetSentence	0.84***
Time:Word:TargetSentence	-0.43***
Region:Context:Word	-0.60***

\*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$

**Effect of Critical Word** While there is no main effect of Word, we find a significant positive interaction Time:Word that indicates that fixations on the target word increase more quickly than fixations on the competitor word. This is not surprising: when participants hear a word that matches the target object on the screen, they fixate this object more frequently (recall that the target object is depicted in all conditions, the competitor object is never on the screen).

**Effect of Region** There is no significant main effect of Region: whether the fixations are in the critical region (between the onset and the offset of the critical word) or in the post-critical region. However, we find a significant negative interaction Region:Context, suggesting that the neutral context sentences receive more fixations in the post-critical region compared to the biasing context sentences. This is compatible with the observation that context facilitates the processing of the critical word, which thus receives fixations earlier in the context condition.

The interaction Region:TargetSentence confirms that in the post-critical region participants fixate the target object more in the competitor context, presumably because it conflicts with their contextual expectations in this case. In the target context, however, contextual expectations and target object match, which means there is no reason to fixate the target object more frequently (compare Figures 2(b) and 2(c)).

**Interaction of Context and Critical Word** The most important interactions with respect to our experimental hypothesis are those involving Context and Word or TargetSentence and Word. These interactions demonstrate

that context has an effect that is specific to the critical word.

We find a significant positive interaction Word:TargetSentence, which demonstrates that the target object receives more fixations when the target word occurs in the target context (rather than in the competitor context). This effect changes over time (significant negative interaction Time:Word:TargetSentence): the increase in fixations in the target word condition is larger in the competitor context than in the target context. For the competitor word, the opposite tendency emerges. This confirms the prediction that an expected critical word (i.e., one matching the context) is less interesting, and thus less likely to be fixated.

Finally, we can report a significant negative interaction Region:Context:Word, suggesting that the effect of Word in the neutral context condition is limited to the post-critical region, while in the biasing condition, it is stronger in the critical region. This corresponds to the observation that the fixation curves for the target and the competitor word diverge earlier for the biasing context conditions (see Figure 2).

## General Discussion

First of all, our results replicate the findings of Huettig et al. (2006). In the neutral context condition, we find that participants fixate the target object both when they hear the critical word, and when they hear the semantically related competitor. While we observe less fixations on the target for the competitor word, Figure 2(a) clearly indicates that it is fixated more than chance (corresponding to a probability of .25).

However, the main purpose of our experiment was to test the ability of a distributional model of semantics to generate properties of concepts that are cognitively plausible. We therefore included two context conditions in our experiment, one in which the context sentence contained properties related to the target word, and one in which it contained properties related to the competitor word. In both cases, the properties were created by Strudel, a model of semantic representation.

When we compared these two biasing context conditions to the neutral context condition, we found two main effects. Firstly, a biasing context facilitates the processing of the critical word. Over time, the context builds up an expectation of the critical word, resulting in less fixations to the target object when it is contextually expected. This effect occurs for both types of biasing contexts, which is in line with the fact that the target and the competitor words were semantically related, which presumably implies that their properties are also semantically related. In the neutral context, in contrast, no expectations can be computed, as participants cannot guess the identity of the target word before its onset. The target object is unexpected and hence more interesting and receives more fixations, but these fixations appear later, once the recognition of the target word is complete.

Our second finding is that a biasing context makes it possible to anticipate the critical word: in a target context, we get more fixations to the target during the target word, com-



pared to the competitor word (Figure 2(b)). In the competitor context, we also initially find more fixations during the competitor word than during the target word. However, the pattern reverses after about 200 ms, presumably because of the match between the target word and the target object on the screen, which overrides the contextual expectation of the competitor word. Fixations for the target word remain high, however, compatible with a violation of contextual expectations (Figure 2(c)).

Both effects provide confirmation for the claim that we started out to prove: distributional models of semantics can generate properties that are cognitively plausible. They are plausible in the sense that they can be used to construct contexts that successfully bias participants towards a word that is compatible with the context. This contrasts with a neutral context, in which differences in fixation probabilities are purely driven by the semantic similarity with the target word. We therefore conclude that models like Strudel are a first step towards modeling linguistic context in a distributional way, which contrasts with the single-word approach that most of the distributional semantics literature has taken so far.

### Acknowledgments

The work reported here was funded by the European Research Council under award number 203427 “Synchronous Linguistic and Visual Processing”. We are grateful to Moreno I. Coco and Christoph Scheepers for their essential support and suggestions and to Desmond Elliott and David Matthews for the help in the production of the linguistic stimuli used during the experiment.

### References

- Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.
- Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, 43(3), 209–231.
- Baroni, M., & Lenci, A. (2008). Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1), 53–86.
- Baroni, M., Murphy, B., Barbu, E., & Poesio, M. (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2), 222–254.
- Barr, D. J. (2008). Analyzing visual world eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, 59(4), 457–474.
- Coco, M. I., & Keller, F. (2012). Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science*, in press.
- Garrard, P., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, 18(2), 125–174.
- Huetting, F., Quinlan, P. T., McDonald, S. A., & Altmann, G. T. M. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychologica*, 121(1), 65–80.
- Landauer, T., & Dumais, S. (1997). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211–240.
- McDonald, S. A. (2000). *Environmental determinants of lexical processing effort*. Unpublished doctoral dissertation, University of Edinburgh.
- McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4), 547–59.
- Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
- Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology. Human Learning and Memory*, 6(2), 174–215.
- Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632–34.

# Concepts in context: Evidence from a feature-norming study

Diego Frassinelli (d.frassinelli@sms.ed.ac.uk)

Institute for Language, Cognition and Computation  
School of Informatics, University of Edinburgh  
10 Crichton Street, Edinburgh EH8 9AB, UK

Alessandro Lenci (alessandro.lenci@ling.unipi.it)

Dipartimento di Linguistica "T. Bolelli", Via Santa Maria 36  
56126 Pisa, Italy

## Abstract

Concepts are typically conceived as context-free knowledge structures. Recently, a different view has emerged according to which subjects produce situation-specific conceptualizations, thereby raising important questions about the level of contextual dependency in conceptual representations. In this paper, we present a feature-norming study in which subjects are asked to generate properties of concepts presented in context. Collected data are analysed to investigate the actual amount of conceptual variation induced by contexts and the effect of context modality.

**Keywords:** Semantic feature norms; property generation; context.

## Concepts and contexts

Both in classical and in post-classical models, concepts have been conceived as substantially context-free knowledge structures. Regardless of the particular theory (e.g. exemplar, prototype, and connectionist), it is generally assumed that concepts result from abstracting critical information about an entity *per se* (such as shape, colour, etc.), leaving behind background situations (i.e. the contexts) in which these entities are experienced. Concepts thus become invariant to different contexts of use. Accordingly, the same representation of an apple is used both when categorizing an entity on a tree, and when categorizing the same entity in a supermarket.

Recently, this view has been overtly criticized. For instance, Yeh and Barsalou (2006) argue that concepts not only contain a large array of situational information about the physical settings, events, and subjective perspectives of agents, but they also produce different conceptualizations in different contexts. For instance, the supermarket situation would activate context-specific information concerning an apple, different from that activated by a different context, such a tree in a garden. These two claims directly follow from the perceptual simulation model adopted by the authors, but more in general they raise important questions about the level of contextual dependency in our conceptual representations. Wu and Barsalou (2009) used a property generation task to investigate the situated nature of concepts, and reported that approximately 26% of the features produced by subjects were indeed situation-related. Subjects generated properties (semantic feature norms) provide interesting evidence about conceptual representations, but one intrinsic limit of the study in Wu and Barsalou (2009) is that stimuli

were presented *out of context*, as it is customary in semantic norming. This way, it becomes impossible to address and test the more specific and crucial issue concerning the relation between concepts and context, that is the actual effect of the context in modulating and biasing conceptual representations.

In this paper, we present a feature-norming study in which subjects are asked to generate properties of concepts presented *in context*. To the best of our knowledge this is the first property generation task with this design. While we do not commit ourselves to any specific model of conceptual representation, collected data allow us to address directly three key issues concerning the effects of different contexts on concepts: i.) the actual amount of conceptual variation induced by contexts, and ii.) the property types that are more subject to contextual variation, and iii.) the effect of the context modality. In particular, we will investigate the effect of both linguistic contexts (i.e. a sentence in which the context noun appears) and extralinguistic contexts (i.e. an image of a situation in which an entity can be experienced).

## Semantic Feature Norms

Nowadays there is a strong consensus on the fact that it is possible to describe the internal structure of a concept in terms of a set of semantic properties (Garrard, Lambon Ralph, Hodges, & Patterson, 2001; Baroni & Lenci, 2008). A traditional way to access and study the structure of conceptual knowledge is the use of semantic features norms. These are lists of properties that participants produce describing and defining a specific concept; moreover they include several measures and statistics calculated according to feature production frequencies.

As suggested by McRae and colleagues (McRae, Cree, Seidenberg, & McNorgan, 2005) these lists do not provide a static and definitive representation of concepts, however, they are the most direct way to study the dynamics associated with the online process that takes place when subjects have to process a specific concept.

Different researchers used these lists to investigate various aspects of human cognition. They have been used to test the psychological validity of cognitive theories (Wu & Barsalou, 2009), and as stimuli for different experiments such as semantic similarity (McRae, Sa, & Seidenberg, 1997) and property verification tasks (Cree, McNorgan, & McRae, 2006).

One of the most widely used norms is the collection of McRae et al. (2005). This is the largest set of semantic properties freely available: it includes properties for 541 living and non-living concepts. Another smaller example is represented by the collection of Vinson and Vigliocco (2008). In this case, the authors extended their analysis to the domain of actions and events. They collected norms for 167 living and non-living objects and for 287 events and actions.

The pros of these collections are relatively straightforward; however they exhibit also different limitations (McRae et al., 2005). The process of collection, normalization and classification is extremely long and expensive. Moreover, the linguistic nature of the task favours the information which is easily verbalized, penalising spatial and temporal relations between entities. During the classification phase, the annotators have to reinterpret the intents of the subjects and cannot always preserve the original information. Finally, as we said above, all existing norms were collected by presenting words in isolation and not associated with a specific context. In this work, we will focus our attention on this last feature.

### Collecting context-sensitive feature norms

The main goal of this work is to describe the collection of semantic feature norms for 8 concrete concepts and to analyse the effects that contextual variability exerts on the number and types of properties produced.<sup>1</sup>

### Design

The collection was performed on-line using a website interface.

**Stimuli** The 8 normed concepts correspond to the following English nouns: *apple*, *banana*, *bear*, *horse*, *bike*, *car*, *hammer*, and *knife*. The nouns were sampled in order to have an equal number of animate and inanimate concepts belonging to the semantic classes traditionally used in these studies, that are fruits, animals, vehicles, and tools.

For each concept, we identified two alternative situations frequently associated with the correspondent object. We downloaded from the Web 16 colour pictures depicting the two contexts for each concept and we downsized them (288\*320 pixels). The pictures do not include the target object unless it is strictly necessary for the correct interpretation of the context (e.g. a showroom without some cars inside would not be identifiable). This way, participants are not biased in their descriptions by a specific instance of the concept appearing in the picture. A native English speaker produced 16 sentences describing the context depicted in the correspondent picture. Unlike the visual contexts, the sentences include the target concept noun (written in capital letters).

For every trial, the target concept noun (in capitals) appears on the top of the screen and is followed by 10 blank lines that participants have to fill in with concept properties. In the case

of the linguistic context, the sentence containing the target word appears instead of the target word. For the visual context, the picture appears on the left of the blank lines. Figure 1 shows the visual and linguistic contexts for *apple*.

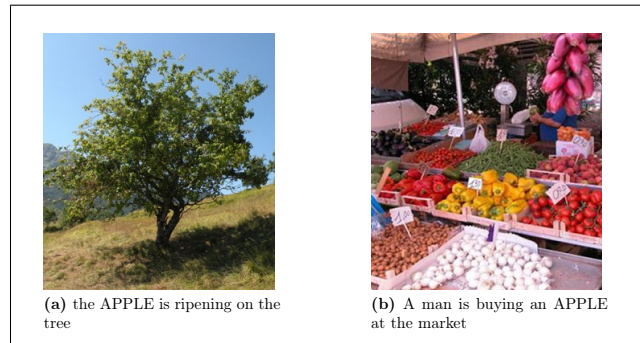


Figure 1: *apple*: visual and linguistic contexts.

**Procedure** The entire experiment included 40 different combinations of the 8 concepts and the 5 different context types (2 visual, 2 linguistics, and 1 no context). 125 lists of 8 items each were created: the distribution of the items across lists was based on a Latin Square design, ensuring that every list comprised only one occurrence of each concept and one of each specific context. Every list included the 8 concepts as follow: one or two out of context, three or four in a visual and linguistic context respectively. The order of the trials was semi-random: all the out of context trials appeared before the linguistic ones and those before the visual ones. In this way, the complexity of the stimulus increased during the experiment.

125 native English speakers recruited online performed the experiment. Each participant saw the 8 items of one of the 125 lists; in this way, every concept and context were seen only once. The experiment started after written instructions of the task and 3 examples. The task required to read the word (or the sentence) at the top of the screen and to produce a maximum of 10 properties per concept describing different aspects of it. The instructions clearly stated that the aim of the experiment was to study how people process the meaning of words; subjects were not instructed to take contextual variability into account. Moreover, we provided a list of possible qualities of the concepts to take into account during the actual experiment: colours (e.g. CHERRY *red*)<sup>2</sup>, tastes (e.g. ICE CREAM *good*), shapes (e.g. BALL *round*), functions (TRAIN *transportation*), typical locations (SHARK *ocean*), emotions (CHRISTMAS *excitement*), evaluations (SOUP *hate*), etc. We did not set a maximum amount of time for a single trial however, on average, the entire experiment lasted about 15 minutes.

<sup>1</sup>The collection is freely available at <http://sesia.humnet.unipi.it/norms>

<sup>2</sup>In this work, we use caps to indicate CONCEPTS and italics to indicate the *properties produced*.

## Post Processing

**Data Codification** After collecting the features, a process of data filtering and normalization was carried out. We identified all the synonym properties and we normalized them to the same feature (e.g., “bike” and “bicycle” were coded as *a bike*). Coordinate or disjunctive features providing more than one piece of information were split into separate tokens. For example, “is red, green or yellow” became *is red, is green, and is yellow*. We removed all the quantifiers (e.g. “can be”, “generally”, “usually”) and other materials not relevant to the analysis (e.g. miscellanea, incoherences, and free associations). Finally, one of the authors coded the resulting properties according to a specific set of patterns (e.g. BEAR *beh\_eats\_honey* codes “to eat honey” as a prototypical behaviour (beh) of bears) and classified them according to the scheme described below.

**Coding Scheme** The properties were classified according to a partially simplified version of the coding scheme proposed by Wu and Barsalou (2009). The scheme includes 24 property types grouped into 4 main categories:

- *Taxonomic properties* (TAX): properties describing taxonomic relations (hypernyms, hyponyms, synonyms, and coordinates).
- *Entity properties* (ENT): properties describing the entity per se (e.g. internal and external properties and elements, prototypical behaviours).
- *Situation properties* (SIT): properties associated with the contextual background (e.g. locations, time, participants, functions).
- *Introspective properties* (INT): properties describing feelings and mental states (e.g. evaluations, contingencies).

## Results

Participants produced 6922 properties in total: 3619 entity properties, 2025 situation properties, 644 introspective properties, and 634 taxonomic properties. Table 1 reports the average number of features (and Standard Error) produced by every subject for each item and grouped by property class and modality. There are no noticeable differences among modalities (visual, linguistic, and no context): this suggests that different contexts are not exerting a significant effect on the number of properties generated. The differences arise analysing the property classes. The properties describing different aspects of the target concept (ENT) are the most frequent (52% of the total). Properties providing contextual information related to the target concept (SIT) are produced the 29% of the time (interestingly, this figure is close to the one reported by Wu and Barsalou (2009)). Properties describing mental states (INT) and taxonomic relations (TAX) are less frequent (around 9%).

Table 1: Average (AVG) and Standard Error (SE) of the number of features produced by each subject for each concept grouped according to broad property class and modality.

Class	No Context		Visual		Linguistic	
	AVG	SE	AVG	SE	AVG	SE
TAX	1.20	0.09	1.34	0.15	1.24	0.11
ENT	3.82	0.40	3.64	0.36	3.86	0.41
SIT	2.34	0.26	2.66	0.31	2.15	0.25
INT	1.61	0.19	1.59	0.19	1.64	0.19

## Analysis

**Model** We analysed the data adopting the framework of the linear-mixed effects models with a Poisson linking function (Baayen, Davidson, & Bates, 2008). The dependent variable was the property frequency. Table 2 presents the coefficients and p-values of the mixed model. To investigate the effects exerted by contextual variability we included two factors in contrast coding: the factor *Modality* for the effects of visual (+.5) and linguistic context (-.5), and the factor *Context* for the effects produced in the out-of-context (-.5) and in the in-context (+.25) conditions. We also analysed the effects associated with the type of feature produced: the factor *Property* compared the object related properties such as entity and taxonomic properties (-.25) and the context related properties such as situation and introspective properties (+.25); the factor *ObjectProp* coded the effects of taxonomic (-.5) and entity (+.5) properties; and the factor *SituationProp* the effects of situation (-.5) and introspective (+.5) properties. The random effects were *Subject* and *Concept*, which were intercepts in the model. We also included random slopes for all the main effects (*Modality*, *Context*, *Property*, *ObjectProp*, and *SituationProp*).

Table 2: The coefficients for the linear-mixed effects model.

Predictor	Coefficient	Signif.
(Intercept)	0.66	***
Property	-0.28	
ObjectProp	1.10	***
SituationProp	-0.43	***
Context	0.01	
Modality	0.04	
Context:Property	-0.09	
Context:Object	-0.17	
Context:Situation	-0.06	
Modality:Property	0.14	
Modality:Object	-0.13	
Modality:Situation	-0.29	**

\*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$

**Property Types** The factor `Property` compares the overall mean of entity and taxonomic properties (properties more associated with the object) with the overall mean of situation and introspective properties (properties that are more associated with the context). We find a slightly significant effect ( $p < .1$ ) in favour of the first group of properties: object related properties are produced more frequently than context related properties. The factor `ObjectProp` shows a significant positive main effect for the entity properties compared to the taxonomic ones. The factor `SituationProp` reveals a positive effect for situation properties compared to the introspective ones.

**Context and Modality** As expected, there is no significant main effect for `Context` and `Modality`. We only find a not significant effect associated with the visual context. Participants produced almost the same number of properties independently of the presence and type of contextual information they are exposed to.

**Interactions** The interactions reported in table 2 do not reveal significant effects. The only significant effect is described by the interaction `Modality:SituationProp` where situation properties are positively biased by visual contexts, while introspective properties are more biased by linguistic properties.

### Qualitative Analysis of Feature Types

In this section we present a qualitative analysis of the data to determine the main effects exerted by contextual information on specific property types.

For each concept, we divided the features produced in both contexts from those associated only with a specific context. We did the same procedure for visual and linguistic contexts independently. The aim of this analysis is to determine which property types are more dependent on a specific context (context dependent) and which are produced in both contexts (context independent). We are interested in a general evaluation of this effect without taking into account inter-conceptual variability: for this reason we combined the results obtained for each concept. After a preliminary analysis, we discovered that property types show an almost coherent trend both in a visual and linguistic context. We analysed the effects exerted by the two modalities using a linear model. We did not find a significant difference between visual and linguistic modality as main effect ( $\beta_{Visual} = 3.90, p = 0.94$ ) and also in interaction with the context\_dependent vs context\_independent variable ( $\beta_{Visual:ContextInd} = 4.54, p = 0.88$ ). For that reason, table 3 reports the results from a general point of view, without modality distinction. If there is a difference between the two modalities we will discuss it separately. We report the percentage of context dependent and context independent properties out of the total number of properties of the same type (e.g. the 92% of hypernyms are context independent). We also present in bold the percentage for the entire

class at the end of each group of properties (e.g. the 80% of taxonomic properties are context independent). In brackets there is the number of properties of each type out of the total number of properties in the same class (e.g. the 78% of taxonomic properties are superordinates).

Table 3: Percentage of the features that are context dependent (associated with only one context) and context independent (associated with both contexts).

Property	Dependent	Independent
C-super (.78)	0.08	0.92
C-subord (.19)	0.57	0.43
C-coord (.02)	1	0
C-syn (.02)	0.5	0.5
<b>Taxonomic</b>	<b>0.2</b>	<b>0.8</b>
E-exsurf (.27)	0.15	0.85
E-excomp (.24)	0.12	0.88
E-sys (.21)	0.28	0.72
E-beh (.07)	0.39	0.61
E-incomp (.06)	0.24	0.76
E-insurf (.06)	0.17	0.83
E-mat (.06)	0.07	0.93
E-quant (.02)	0.53	0.47
E-whole (.01)	0.21	0.79
<b>Entity</b>	<b>0.2</b>	<b>0.8</b>
I-cont (.68)	0.62	0.38
I-eval (.30)	0.38	0.62
I-emot (.02)	0.5	0.5
<b>Introspective</b>	<b>0.54</b>	<b>0.46</b>
S-func (.47)	0.25	0.75
S-assoc (.15)	0.58	0.42
S-loc (.15)	0.44	0.56
S-action (.08)	0.33	0.67
S-particip (.08)	0.48	0.52
S-origin (.06)	0.1	0.9
S-time (.01)	0.59	0.41
S-socart (<.01)	1	0
<b>Situation</b>	<b>0.35</b>	<b>0.65</b>

**Taxonomic Properties** Taxonomic properties describe highly stable relations among concepts. As expected, the 80% of these properties are equally produced in different contexts. The hypernyms (C-super, e.g. CAR *a vehicle*) are the 78% of the entire taxonomic class. These properties are represented by a small number of highly frequent feature types (in total only 22 for the visual and linguistic modalities) describing associations strictly language related. Hyponyms (C-subord, e.g. APPLE *Granny Smith*) include a high number of infrequent property types (in total 48 subordinates) and are more context dependent. This can be expected, given that each concept is associated with many hyponyms, which in turn might become differently prominent depending on the context. Co-

ordinates (C-coord, e.g. APPLE *a pear*) and synonyms (C-syn, e.g. CAR *an automobile*) are only a small group of properties.

**Entity properties** The trend of this class is consistent: all the properties describing objects' qualities are not sensitive to contextual variability (the 80% of the total). The only exception is represented by those properties describing frequency or intensity (E-quant, e.g. APPLE *different varieties*); however this group includes a very small number of features and it is valid only for the linguistic modality (65% of context dependent properties).

**Introspective properties** The most substantial group among introspective properties is represented by contingency properties (I-cont, e.g. APPLE *is good with cinnamon*). These properties describe the "common sense knowledge" associated with a specific object in specific conditions. For this reason, it is not surprising to see a strong contextual effect. On the other hand, evaluations about the object (I-eval, e.g. APPLE *is delicious*) are less context dependent: participants have a personal opinion about every object that is unlikely to change in different situations. Emotions (I-emot, e.g. BEAR *is scary*) are very few cases.

**Situation Properties** The behaviour of this group of data is more various, given also the high heterogeneity of the properties in this class. Some properties are intrinsically related to an entity, and, therefore, less variable across situations: for instance, typical functions (S-func, e.g. CAR *used for transportation*), actions (S-action, e.g. APPLE *used by cooking*), origins (S-origin, e.g. APPLE *grows on trees*) and locations (S-loc, e.g. BANANA *grows in tropical climates*). Instead, other property types are more context-related, and, therefore, subject to stronger cross-situation variation, such as associations (S-assoc, e.g. CAR *associated with speed*). Participants (S-particip, e.g. BANANA *eaten by monkeys*) are almost equally present in both sets.

### Analysis of Feature Density

The last analysis we carry on compares the distribution of specific features in terms of feature density: how many subjects produce the same feature for the same stimulus. We analysed the features divided into the context dependent and context independent sets. In the context dependent set the 85% of specific properties are produced by only one person, the 11% by two different subjects. The remaining 4% is shared by properties produced from 3 to 10 people. On the other hand, in the context independent set we have the 21% of features produced only by 2 subjects (the minimum value for having an overlap). The maximum number of subjects that produced the same feature is 47 (KNIFE *is sharp*).

## General Discussion

The experiment described in this work was aimed to test the effects exerted by contextual variability on the production of semantic properties by human beings. We gave both quantitative and qualitative evidence of these effects. Neither modality nor context variability has significant effects on the number of features produced. People list almost the same number of properties in different contexts. However, these properties are not equally distributed and the differences among them are statistically significant. As already emerged in the literature, subjects produce more entity properties than taxonomic ones and more situated properties than introspective ones. It is interesting to note that merging together the properties more object related (entity and taxonomic) and the properties more context related (introspective and situation) the difference decreases considerably with only a slightly significant effect in favour of the first group. This suggests that people are including in their dynamic representation of concepts both information describing the object *per se* but also almost the same amount of background information. To gather more evidence, we performed also a qualitative analysis. In this case, we extracted from the collection only the properties produced in context and we identified those occurring with both contexts and those associated with only one. The results are straightforward for the taxonomic and entity properties: almost the 80% of all the properties classified in this way are produced in both contexts. We assist to an opposite effect when we move to the introspective properties. More complex is the dynamic of situation properties: some of them are more related to the object, some others to the context.

These results suggest that the context sensitivity of concepts is strongly limited to certain property types. A possible explanation can be found in Barsalou (1982). In this work, the author suggests the existence of two different kinds of properties: *context independent properties* strictly associated with the object *per se*, and *context dependent properties* associated with the specific context in which the word appears. Our data point in the same direction. It is possible to identify a large group of "core" properties that are not biased by contextual variability (in particular entity and taxonomic properties, as expected) and a smaller group of more dynamic properties produced less frequently and only associated with specific contexts (introspective properties, and partially situation properties).

## Acknowledgements

The experiment was funded by the Computational Linguistics Lab (University of Pisa) and the European Research Council under award number 203427 "Synchronous Linguistic and Visual Processing". We thank Larry Barsalou, Marco Baroni, Frank Keller and Moreno I. Coco for their help and suggestions.

## References

- Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.
- Baroni, M., & Lenci, A. (2008). Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1), 53–86.
- Barsalou, L. (1982). Context-independent and context-dependent information in concepts. *Memory and Cognition*, 10(1), 82–93.
- Cree, G. S., McNorgan, C., & McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Learning, Memory, and Cognition*, 32(4), 643–658.
- Garrard, P., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, 18(2), 125–174.
- McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4), 547–59.
- McRae, K., Sa, V. R. de, & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology*, 126(2), 99–130.
- Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, 40(1), 183–190.
- Wu, L., & Barsalou, L. (2009). Perceptual simulation in conceptual combination: evidence from property generation. *Acta Psychologica*, 132(2), 173–89.
- Yeh, W., & Barsalou, L. (2006). The situated nature of concepts. *American Journal of Psychology*, 119(3), 349–84.



# Learning transfer in small group coordination

Seth Frey (sethfrey@indiana.edu)

Cognitive Science, 1101 E. 10th St., rm. 290B  
Bloomington, IN 47401 USA

## Abstract

This work tests the adaptation of groups from two generalizations of the multiple-player stag hunt to a difficult third version, the notorious weakest-link game. The two training conditions either encouraged or discouraged the development of stable subgroups. Theories of modularization predict that stable subgroups will facilitate coordination in larger groups by helping them “scale up.” However, internal structure may also cause “overfitting,” or adaptation to only spurious features of training. In this experiment, experience with internal structure prevented coordination at larger scales, while experience in environments that discourage internal structure led to performance at least as high as in the control environments. I offer the analogy from individual learning transfer, that distracting details from superficially-similar domains may transfer and interfere with coordination. This work has implications for the development and adaptability of small coordinating groups. In particular, it demonstrates that coordination is not a monolith, and experience with one sense may impair performance under others.

**Keywords:** stag hunt; learning transfer; n-player games; coordination games; group structure.

## Introduction

Adaptability and transfer have been a focus of group research since its inception (Bavelas, 1950; Guetzkow & Simon, 1955). The research in “endogenous” groups applies modern experimental methods and game theoretic approaches to understanding adaptability and learning during group formation.

Traditional group experiments impose groups by design, and they use careful matching procedures and information conditions to minimize interaction and reputation effects, and to maintain the independence of individual reasoners. Investigations of *endogenous* group formation encourage interactions and attend to the processes of group formation within an experiment. Endogenous group experiments permit the study of internal (often “network”) structure. In groups, *internal structure* describes the heterogeneous but systematic pattern of coordinated behavior between subgroups. Internal structure can be measured by observing the behavior of group members over time. As the demands on a group change, these internal patterns should also change. For large networks, group structures can be compared along innumerable dimensions, but for the small groups featured in this study it is possible to meaningfully quantify structure with fewer variables.

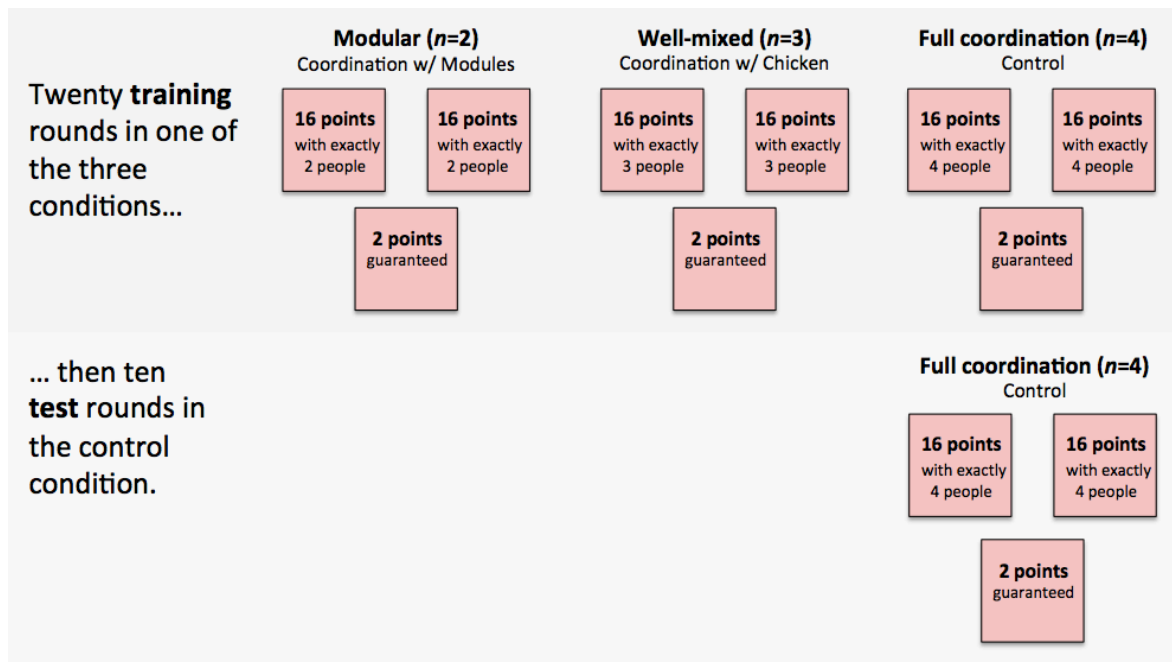
Ahn, Isaac, & Salmon (2008) documented segregation processes as a result of reputation formation and individual choice. Camerer and Weber elicited unprecedented

coordination in the weakest-link game by starting with successful groups of two and “growing” them slowly up to the standard 7-person case (Camerer & Weber, 2008). Their work shows that coordination can be attained if the learning process occurs in parallel with the group formation process. Across a diversity of paradigms, many other experiments have found conditions in which the internal structure created by local information improves group-level outcomes (Ahn, Esarey, & Scholz, 2009; Goldstone, Roberts, Mason, & Gureckis, 2010; Kearns, Suri, & Montfort, 2006; Mullen, Johnson, & Salas, 1991; Mason, Jones, & Goldstone, 2008).

There is also conflicting evidence that internal structure interferes with coordination among endogenous groups. Weber & Camerer (2003), complementing their “slow growth” result above, showed that combining two successful medium-sized groups can induce coordination failure in the large merged group. Rick, Weber, & Camerer (2006) hypothesized that decentralization would promote adaptability. After their result failed to materialize, they reframed the experiment in terms of learning transfer (Rick, Weber, & Camerer, 2007). Frey & Goldstone (2010) show the simultaneous emergence of internal structure and coordination failure in a multiple-player stag hunt. Participants in groups of many sizes could coordinate on equilibria that required larger quora for larger payoffs. Despite many trials and large groups, coordination failure was common, and the successes were most typical for the smallest subgroups of pairs and triplets, despite the much larger rewards available with more concentrated groups. There is no theory to reconcile the conflicting effects of group formation on individual outcomes.

This experiment explores whether internal structure will help or hinder groups in coordinating at larger scales. The complex experimental designs and large networked groups of most endogenous group experiments make it difficult to carefully distinguish hypotheses. I introduce a 4-player coordination game that elicits three types of coordination, depending on the value of an experimentally manipulated parameter. An experiment with only four players can be represented as a very small network. The design focuses on learning transfer from each of three coordination games.

Learning transfer is a growing area of behavioral game theory. Studies suggest that experiences of cooperation or coordination can transfer across very different types of games (Stahl, 2000; Devetag, 2005). In the most ambitious study, Woolley, Chabris, Pentland, & Hashmi (2010) give evidence for a general Collective



**Figure: Structure of the game, and design of the experiment**

Groups of four picked from three choices over thirty rounds of a Stag Hunt game, moving from twenty training rounds to ten identical test rounds. Two of the three choices (the top two) required coordination with exactly  $n$  other group members for 16 points (Stag), and the remaining choice guaranteed 2 *secure* points (Hare).

Intelligence factor by which certain groups seem to transfer easily across many unrelated collective tasks. But the processes of individual learning and transfer are complex, and we cannot cherry-pick concepts from the individual-level without acknowledging this complexity. For example, one individual-level phenomenon that complicates the idea of general intelligence is *negative transfer*, a type of learning transfer. In cases of negative transfer, a superficial similarity between two tasks will induce transfer of spurious concepts that actually hinder performance (Gagne, Baker, & Foster, 1950; Day & Goldstone, 2012). For example, tennis and badminton are superficially similar sports, but skills developed in one should not necessarily transfer to the other; Some tennis techniques, like maintaining firm wrists, will not improve performance in badminton. Negative transfer complicates general intelligence factors by introducing order effects, and generally by raising questions about what it means for two tasks to be similar or distant.

With the right design, a failure of internal structure to improve large-scale coordination implies negative transfer. Groups in this environment did in fact show negative transfer from their experience of a seemingly congruent coordination game. The consequence is that transfer is better from a less similar game—one that discouraged the slow, modular development of large-scale coordination.

## Method

Participants played a novel multiple player generalization of the stag hunt, a classic two-player game. This paradigm was designed to explore the relationship between a group's internal patterns of behavior and their ability to coordinate at larger scales.

In the prototypical stag hunt, two players choose blindly to either hunt *Hare* or *Stag*. Hare offer little meat but can be caught without another's assistance. Stag reward both hunters with more food per person, but they cannot be successfully hunted by an individual alone. A decision to hunt hare instead of stag reflects an aversion to risk, because a lone stag hunter gets nothing while hare hunters are guaranteed a small but secure reward. The stag hunt captures many of the general incentives and pressures behind coordination, and the properties of agents in the stag hunt are well documented, both theoretically and empirically (Harsanyi & Selten, 1988; Camerer, 2003; Skyrms, 2004). However, upon expanding to multiple players, the number of generalizations explodes, and most extant research is in simulation.

In the multiple player stag hunt reported here—the *structured stag hunt*—players have three choices: they may hunt hare alone for a small payoff (2 points) or they may hunt either of two stag for a larger payoff (16 points). Group size in this game is fixed at four players.

The payoff structure of the two stag strategies depends on parameter  $n$ , which may equal 2, 3, or 4. At

$n=4$ , the *Full coordination* condition, players who select one of the Stag strategies receive the large payoff only if all 4 players select it (See Figure). At this equilibrium, four hunters have successfully coordinated to hunt one of two large stag. At  $n=4$ , there are three pure strategy Nash equilibria, corresponding to the decision of all four players to all choose one of the three strategies. Previous work suggests that, as the game is iterated, players will settle into one of these three pure strategy equilibria (Camerer & Weber, 2007). The game also has many mixed-strategy equilibria—at all parameterizations—but I will attend only to the more salient pure equilibria. The Full coordination condition functioned as a control baseline condition for comparison with the other two conditions.

For  $n=3$ , the *Well-mixed* condition, only 3 players may hunt stag in order to receive a payoff. If all four players chose the same Stag strategy, then they each receive 0 points. In this condition, the proverbial stag has plenty of meat for three people, but is, perhaps, very likely to be startled by the stirrings of a larger crowd. At  $n=3$ , the structured stag hunt resembles the game of Chicken, in which the most profitable strategy for individual players gives the lowest payoff if all players select it.

Both theory and experiment suggest that coordination will be unstable in this condition of the game (Bornstein, Budescu, & Zamir, 1997). There are nine pure-strategy Nash equilibria in the Well-mixed condition but, unlike the other two conditions, there is no equilibrium in which all four players receive the largest payoff. Because the only symmetric equilibria involve selecting randomly from a distribution over the three strategies (mixed strategies), experience in the Well-mixed condition should select against the development of stable internal structure.

Finally, in the *Modular* condition, when  $n=2$ , an internal group structure becomes possible. When  $n=2$  the same penalties apply if more than 2 players select the same stag strategy—players earn nothing. However, the  $n=2$  condition is special because it is possible to split four players evenly between the two Stag such that all four players receive the higher payoff. The condition is called Modular because it requires division into two subgroups. With social learning, experience in the Modular condition should promote the development of an internal group structure. Based on results from previous work (Frey & Goldstone, 2010), my prediction is that groups in the Modular condition will naturally stratify between the two stag strategies over time. Whether this facilitated transfer to the Full coordination condition was an open question.

In all three conditions, Pareto dominance fails to select between the two pure equilibria involving Stag strategies, and there is an equilibrium selection problem on top of the traditional problem of selecting between Hare and a single Stag. The complexity of introducing two Stag was necessary to create the three nearly identical conditions.

## Experiment

These three versions of the structured stag hunt provide the basic ingredients for distinguishing between two competing theories of how a group's internal structure influences its ability to coordinate.

Participants played the structured stag hunt in three conditions: a Full control condition and the Modular and Well-mixed conditions.

Participants played thirty rounds of the game. In all three conditions, the last ten rounds functioned to *test* performance in the Full coordination version of the game. In the control condition, the first 20 *training* rounds were also played with  $n=4$ . In the Well-mixed condition, groups played these first twenty rounds at  $n=3$ . In the Modular condition, groups played the first twenty rounds at  $n=2$ .

## Subjects and procedure

52 psychology undergraduates played the game in 13 groups of four. Experiments were conducted over networked computers concealed in separate cubicles. Complete instructions were read aloud before participants were given the opportunity to review them individually. After the first round, participants saw their group's previous round's choices before the next round began. However, participant icons were made identical to make it more difficult to use reputation to form internal structure. Experimental sessions lasted just under five minutes on average, with about 10 seconds per round. Participants were paid a small bonus of 1¢ per point, and mean earnings were \$1.25, or \$15.00 an hour. The experiment was always run in the free time after other collective behavior experiments.

## Measures

The main dependent variable in this experiment is the number of test rounds, out of ten, for which groups settled on one of the pure equilibria containing the Stag strategy. If internal group structure can improve higher-level coordination, then Modular groups will coordinate more successfully in test trials than the Well-mixed groups.

Other dependent variables were influenced by condition and number of rounds, including *group clustering* and *group internal structure*, *payoff*, *opting out*, and *fixation*.

*Group clustering* and *internal structure* were measured as the mean and standard deviation of the distances between participants. In each round, two participants shared distance 0 if they had made the same choice, and 1 otherwise. Summed over all rounds, two participants have distance 30 if they never selected the same choice, and distance 0 if they always selected the same choice. Using pairwise distances between group members, the structure of the group can be represented as a fully connected distance network. If the mean of these distances is low, participants are tending toward clustering on the same choice. For small groups, the standard deviation of distance is equivalent to closeness centrality (Frey & Goldstone, 2010), a measure of relational structure in networks. If the standard deviation is high, participants tend

Dependent measures	Grand Mean	Full ( <i>n</i> =4; control)			Well-mixed ( <i>n</i> =3)			Modular ( <i>n</i> =2)		
		Train	change	Test	Train	change	Test	Train	change	Test
Performance	1.92	0.25		3	<b>3.5 *</b>	↑	4.25	1.25		0
Group distance	3.56	3.8	↓	1.0	3.81	↓	2.25	<b>4.6 *</b>		<b>4.25 *</b>
Group structure	2.82	3	↓	0.92	2.9	↓	2.25	<b>3.48 *</b>		<b>3.41 *</b>
Payoff	4.19	1.6		5.9	<b>4.8 *</b>		7.2	<b>5.56 *</b>		0.712
Opting out [0,1]	0.09	0.15		0.14	<b>0.074 *</b>		0.045	<b>0.045 *</b>		0.089
Fixation [0.33,1]	0.7	0.66	↑	0.93	0.71	↑	0.79	0.57	↑	0.72

**Table: Summary measures of group performance**

Distance is reported instead of clustering. Distance and clustering are inversely related.

**bold\*** reports a significant effect ( $p < 0.01$ ). In a “Train” column, comparison is with respect to training in the control condition. In a “Test” column, comparison is with respect to tests in the control condition.

Arrows report significant effect directions of experience, within a condition, from Train to Test ( $p < 0.01$ ). All regressions report statistics distributed on  $F(5,30)$ .

*Italics* in upper right Test columns represent one significant post-hoc test between Modular and Well-mixed test performance.

to be close to some group members and distant from others. By this definition, a modular arrangement of two groups of two will register as having the highest internal structure.

*Payoff* was the mean number of points per participant. *Opting out* was measured as the percentage of secure Hare choices, out of thirty. *Fixation* measured pure strategy play, the phenomenon of choosing the same strategy repeatedly. Fixation was defined as the maximum value in a participant’s observed distribution of choices. By this, pure-strategy play will register a maximum fixation of 1.0, and random mixed-strategy play will register the minimum value of 0.33.

## Results

Two separate ANOVAs were used to compare the three conditions in the training trials separately from in the test trials. These analyses encoded three dummy variables, one for each condition. Additionally, a multiple regression tested the effects of experience within each condition. The adjusted  $R^2$  of the regression over performance was 71%, and the model was significant ( $F(6,30)=11.2$ ,  $p < 0.01$ ). To maintain the independence of observations, these analyses were conducted at the group-level. Means and significance values are summarized in the Table.

Looking first at the 20 training trials of the control condition, it is clear that Full coordination is difficult. Participants successfully coordinated on Stag in an average of 0.5 of the 20 training rounds. Groups performed much better during training in the Well-mixed condition, coordinating successfully on a Stag in 7 of the 20 training rounds (and significantly more often than groups in the Full coordination condition;  $F(1,21)=24$ ,  $p < 0.01$ ).

During training in the Modular condition, participants settled upon the stable configuration (two subgroups of two) in only 2.5 of the 20 training rounds. This was not significantly higher than control. However, partial solutions were more common, with 21 instances (out of a possible 40, or two per round), of exactly two subjects successfully coordinating on a Stag.

Performance in the control condition increased (non-significantly) from 0.5/20 to 2.75/10 during the final ten test rounds. The first ten rounds of the Full condition may be understood as an alternative baseline, corresponding to an entirely untrained experience of the game. Across all sessions of the Full coordination condition, no groups successfully coordinated on Stag in the first ten rounds of training. Variance in performance was much higher in this condition than in the other two; Groups in the control condition tended either to converge stably on coordination success (10/10 test trials) or coordination failure (0/10 test trials), with little behavior in between these extremes.

Groups trained in Well-mixed trials showed improvement upon advancing to the test trials ( $t(32)=3.16$ ,  $p < 0.01$ ). Mean performance on test was 4.25 out of 10 rounds. This was not significantly different than the test performance of the control group. However, a t-test showed that after training in the Well-mixed condition, groups exhibited better test performance than Modular groups ( $t(3)=-5.67$ ,  $p=0.010$ ), even though the Well-mixed game is less similar to the Full coordination game.

Across all test rounds of all groups in the Modular condition, no group successfully coordinated on Stag during test (0/20). Groups in the Modular condition did not show significantly different performance between training and test trials.

Group clustering, group internal structure, payoff, opting out, and fixation were modeled in the same manner as performance, with a linear model to establish the effects of increasing experience in the game, and separate ANOVAs for train and test trials to establish differences between the three conditions. These dependents were also modeled at the group level. Though many of these results may not become motivated until the discussion, means and significance values are summarized in the Table.

Groups exhibited significant internal structure. In the Well-mixed and Full coordination conditions, experience in the game (number of rounds played) predicted increased clustering ( $t(32)=-5.86$ ;  $t(32)=-7.6$ , respectively—both

$p < 0.01$ ) and decreased internal structure ( $t(32) = -3.95$ ;  $t(32) = -5.95$ , both  $p < 0.01$ ). Clustering and structure did not change for Modular groups, and during test trials, only Modular groups showed less clustering and more internal structure than control groups ( $F(1,33) = 6.22$ ,  $p < 0.01$ ;  $F(1,33) = 6.76$ ,  $p < 0.01$ ).

Payoffs in the training rounds were significantly higher in both the Modular and Well-mixed conditions than in the Full coordination condition ( $F(1,21) = 14.9$ ,  $p < 0.01$ ;  $F(1,21) = 20.5$ ,  $p < 0.01$ ). These differences did not persist between the test trials, and the change between train and test trials was not significant in the regression.

Opting out (choosing the secure payoff) showed the same pattern as payoffs. Opting out was significantly below control during the training of both Modular and Well-mixed groups ( $F(1,21) = 13.2$ ;  $F(1,21) = 12.7$ , both  $p < 0.01$ ), but these changes did not persist in test and were not reflected in the model of experience. Over all trials and conditions, only 9% of choices were to the assured payoff.

Over all subjects, mean fixation was 70%; the most common strategy selected by a given subject was selected for 70% of the rounds. Fixation increased with experience in all three conditions, but did not differ significantly across conditions (Control:  $t(32) = 6.06$ ,  $p < 0.01$ ; Well-mixed:  $t(32) = 4.83$ ,  $p < 0.01$ ; Modular:  $t(32) = 2.99$ ,  $p < 0.01$ ).

## Discussion

In the Full coordination control condition, groups face a more complex version of the already-difficult weakest-link game (Van Huyck, Battalio, & Beil, 1990). They performed moderately well in the test trials.

In the Well-mixed condition, groups showed moderate performance solving the  $n=3$  coordination problem, and their performance on test was at least as good as that of groups trained in the control condition. Data support the prediction that these groups would not form modules or other manifestations of an internal structure, or at least not to the degree exhibited by groups in the Modular condition.

For groups trained in the Modular condition, there was a stable efficient outcome for two subgroups of two whereby all four players receive the largest payoff. Evidence supports the prediction that this environment promotes the emergence of stable subgroups; groups in the Modular condition did in fact exhibit high internal structure consistent with a modular group structure. However, while groups in this condition earned more during training than other groups, they were unable to settle consistently upon two groups of two simultaneously. Modular groups were also unable to use their experience within a structured group to coordinate upon stag during the test rounds.

Why did groups with internal structure test worse than groups without internal structure? I will reject a few possibilities. The first is that subgroups of two are small, while subgroups of three are almost as large as subgroups

of four. Previous work has shown that incremental growth can aid coordination at higher scales (Camerer & Weber, 2008) while larger growth spurts can hinder it (Weber & Camerer, 2003). However, this perspective ignores that fact that coordination in all three conditions is coordination among four people. The structure of the Well-mixed condition, which creates conditions like those in the game of Chicken, makes successful coordination more than a matter of finding three willing risk-takers; it also involves coordinating with a fourth who will forgo the greatest payoff. Perhaps subjects in the structure condition learned to randomize. This would be consistent with the moderately efficient mixed strategies that exist in this condition. But mixed strategies are even more efficient in the well-mixed condition, which should have elicited randomized behavior most effectively. Subjects in the structure conditions could not have fixated too intently on specific strategies because fixation was not higher than in the other conditions. Poor performance can also not have been due to Modular groups opting out and selecting Hare; opting out was also not significantly higher than in any other condition during test, and it did not change with experience.

One more possible explanation is that internal structure itself caused the coordination failure at higher levels. The measures of group structure indicate that participants in the Modular condition overcame the difficulties of positively identifying each other and managed to form groups with internal structure. The structures that the Modular condition selected for were congruent with the demands of Stag in the Full coordination condition; subgroups need only merge into one larger group. However, the experience of stable subgroups seems to have transferred negatively to the Full coordination trials.

## Conclusion

This work looks at the interaction of individual reasoning processes as group members interact and learn to coordinate. Some previous work suggests that building up small coordinating subgroups will aid growth to full-scale coordination. This work supports competing claims, that the spontaneous emergence of local stable patterns of coordination may interfere with large-scale coordination. While groups without stable internal structure performed as well as control groups, groups that adapted to match the modular structure of their problem found that this structure interfered with full-scale coordination in the test environment. While groups are certainly adaptive in an important sense, adaptability is not a universal, or even well defined property of group behavior. Similarly, not all experiences of coordination and cooperation are the same, and experience with one type of coordination can impair performance in others.

## Acknowledgements

The author wishes to thank Tatsuya Kameda, Keigo Inukai, Tom Wisdom, and Robert L. Goldstone for their support and Benjamin Marchus for his assistance in conducting this research. This project was funded by NSF EAPSI 1108165 and National Science Foundation IGERT training grant 0903495 in the Dynamics of Brain-Body-Environment Systems at Indiana University.

## References

- Ahn, T., Esarey, J., & Scholz, J. (2009). Reputation and Cooperation in Voluntary Exchanges: Comparing Local and Central Institutions. *The Journal of Politics*, 71(02), 398–413.
- Ahn, T., Isaac, R., & Salmon, T. (2008). Endogenous group formation. *Journal of Public Economic Theory*, 10(2), 171.
- Bavelas, A. (1950). Communication patterns in task-oriented groups. *Journal of the Acoustical Society of America*, 22(6), 725–730.
- Bornstein, G., & Yaniv, I. (1998). Individual and group behavior in the ultimatum game: Are groups more “rational” players? *Experimental Economics*, 1(1), 101–108.
- Bornstein, G., Budescu, D., & Zamir, S. (1997). Cooperation in intergroup, N-person, and two-person games of chicken. *Journal of Conflict Resolution*, 41(3), 384.
- Camerer, C. F. (2003). Behavioral game theory: Experiments in strategic interaction. Princeton University Press.
- Camerer, C. F., & Weber, R. (2007). Experimental Organizational Economics. in *The Handbook of Organizational Economics*, eds. R. Gibbons and J. Roberts. New Jersey: Princeton University Press.
- Camerer, C., & Weber, R. (2008). Growing organizational culture in the laboratory. *Handbook of Experimental Economics*
- Day, S., & Goldstone, R. L. (2012). The Import of Knowledge Export: Connecting Findings and Theories of Transfer of Learning. *Educational Psychologist*.
- Devetag, G. (2005) Precedent transfer in coordination games, *Economic Letters* 89, 227–232.
- Feri, F., Irlenbusch, B., & Sutter, M. (2009). Efficiency gains from team-based coordination—Large-scale experimental evidence. *Preprints of the Max Planck Institute for Research on Collective Goods*.
- Frey, S., & Goldstone, R. (2010). Group Stratification and Coordination Failure in a Continuous N-Player Stag Hunt. In 2010 Proceedings of the Cognitive Science Society. Presented at the 2010 *Proceedings of the Cognitive Science Society*.
- Gagne, R. M., Baker, K. E., & Foster, H. (1950). On the relation between similarity and transfer of training in the learning of discriminative motor tasks. *Psychological Review*.
- Goldstone, R. L., Roberts, M., & Gureckis, T. (2008). Emergent processes in group behavior. *Current Directions in Psychological Science*, 17(1), 10–15. doi:10.1111/j.1467-8721.2008.00539.x
- Goldstone, R. L., Roberts, M., Mason, W., & Gureckis, T. (2010). Collective Search in Concrete and Abstract Spaces. in *Decision Modeling and Behavior in Complex and Uncertain Environments* (Eds. T. Kugler, J. C. Smith, T. Connolly, Y. Son ). New York: Springer Verlag (pp. 277–308).
- Guetzkow, H., & Simon, H. (1955). The impact of certain communication nets upon organization and performance in task-oriented groups. *Management Science*, 1(3/4), 233–250.
- Harsanyi, J. C., Selten, R. (1988) A General Theory of Equilibrium Selection in Games, MIT Press Books.
- Kearns, M., Suri, S., & Montfort, N. (2006). An experimental study of the coloring problem on human subject networks. *Science*, 313(5788), 824.
- Mason, W., Jones, A., & Goldstone, R. (2008). Propagation of innovations in networked groups. *Journal of Experimental Psychology*, 137(3), 422–433.
- Mullen, B., Johnson, C., & Salas, E. (1991). Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and Applied Social Psychology*, 12(1), 3–23.
- Rick, S., Weber, R. A., & Camerer, C. (2006). The Effects of Organizational Structure and Codes on the Performance of Laboratory “Firms.” *Department of Social and Decision Sciences Working Paper*.
- Rick, S., Weber, R., & Camerer, C. (2007). Knowledge Transfer in Simple Laboratory Firms: The Role of Tacit vs. Explicit Knowledge. *Department of Social and Decision Sciences Working Paper*.
- Schelling, T. C. (1960). The strategy of conflict. Massachusetts: Harvard University Press.
- Shupp, R., Williams, A., & Hall, W. (2008). Risk preference differentials of small groups and individuals. *Economic Journal*, 118 (525), 258–283.
- Skyrms, B. (2004). The stag hunt and the evolution of social structure. Cambridge University Press.
- Stahl, D. O. (2000) Action reinforcement learning versus rule learning, Department of Economics, University of Texas.
- Van Huyck, J., Battalio, R., & Beil, R. (1990). Tacit coordination games, strategic uncertainty, and coordination failure. *The American Economic Review*, 80(1), 234–248.
- Weber, R., & Camerer, C. (2003). Cultural conflict and merger failure: An experimental approach. *Management Science*, 49(4), 400–415.
- Woolley, A., Chabris, C., Pentland, A., & Hashmi, N. (2010). Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science*.

# Society Functions Best with an Intermediate Level of Creativity

**Liane Gabora (liane.gabora@ubc.ca)**

University of British Columbia (Okanagan campus)  
Department of Psychology, Arts Building, 3333 University Way  
Kelowna BC, V1V 1V7, CANADA

**Hadi Firouzi (hadi.firouzi@ubc.ca)**

University of British Columbia (Okanagan campus)  
Department of Engineering, EME Building, 3333 University Way  
Kelowna BC, V1V 1V7, CANADA

## Abstract

In a society, a proportion of the individuals can benefit from creativity without being creative themselves by copying the creators. This paper uses an agent-based model of cultural evolution to investigate how society is affected by different levels of individual creativity. We performed a time series analysis of the mean fitness of ideas across the artificial society varying both the percentage of creators,  $C$ , and how creative they are,  $p$  using two discounting methods. Both analyses revealed a valley in the adaptive landscape, indicating a tradeoff between  $C$  and  $p$ . The results suggest that excess creativity at the individual level can be detrimental at the level of the society because creators invest in unproven ideas at the expense of propagating proven ideas.

**Keywords:** adaptive landscape; agent-based model; creativity; cultural evolution; discounting; EVOC; imitation; individual differences; time series analysis

## Introduction

Our capacity for self-expression, problem solving, and making aesthetically pleasing artifacts, all stem from our creative abilities. Psychologists have almost universally converged on the definition of creativity proposed by Guilford over sixty years ago at his annual address to the American Psychological Association (Moran, 2011). Guilford defined creativity in terms of two criteria: originality or novelty, and appropriateness or adaptiveness, *i.e.*, relevance to the task at hand. Individuals vary from not particularly creative to highly creative. Not only are humans individually creative, but we build on each other's ideas such that over centuries, art, science, and technology, as well as customs and folk knowledge, can be said to evolve (Cavalli-Sforza & Feldman, 1981; Gabora, 1996; A. Mesoudi & Laland, 2006; A. Whiten & Stringer, 2011). Creativity has long been associated with personal fulfillment (May, 1975; Rogers, 1959), self-actualization (Maslow, 1959), and maintaining a competitive edge in the marketplace, and it is often assumed that more creativity is necessarily better.

However, there are significant drawbacks to creativity (D. H. Crompton & Runco, 2010; Ludwig, 1995). Generating creative ideas is time consuming, and a creative solution to one problem often generates other problems, or has unexpected negative side effects that may only become apparent after much effort has been invested. Creative people often reinvent the wheel, and may be more likely to bend rules,

break laws, and provoke social unrest (Sternberg & Lubart, 1995; Sulloway, 1996). They are more prone to affective disorders such as depression and bipolar disorder, and have a higher incidence of schizophrenic tendencies, than other segments of the population (Andreasson, 1987; Flaherty, 2005; Goodwin & Jamieson, 1990).

Given these negative aspects of creativity, it is perhaps just as well that, in a group of interacting individuals, not all of them need be particularly creative for the benefits of creativity to be felt throughout a social group. The rest can reap the rewards of the creator's ideas by simply copying, using, or admiring them. Few of us know how to build a computer, or write a symphony, but they are nonetheless ours to use and enjoy. Clearly if everyone relied on the strategy of imitating others, the generation of cultural novelty would grind to a halt. This raises the following questions: what is the ideal ratio of creators to imitators, and how creative should the "creative types" be?

## The Model

We investigated this using an agent-based model of cultural evolution referred to as "EVolution of Culture", abbreviated EVOC. To our knowledge, EVOC is the only computational model that enables one to create an artificial world with agents of varying levels of creativity and observe the effect of varying creativity on mean fitness and diversity of ideas in the artificial society. It uses neural network based agents that (1) invent new ideas, (2) imitate actions implemented by neighbors, (3) evaluate ideas, and (4) implement successful ideas as actions. EVOC is an elaboration of Meme and Variations, or MAV (Gabora, 1995), the earliest computer program to model culture as an evolutionary process in its own right, as opposed to modelling the interplay of cultural and biological evolution<sup>1</sup>. The approach was inspired by genetic algorithm (Holland, 1975), or GA. The GA is a search technique that finds solutions to complex problems by generating a population of candidate solutions through processes akin to mutation and recombination, selecting the best, and repeating until a satisfactory solution is found. The goal behind MAV,

<sup>1</sup>The approach can thus be contrasted with computer models of how individual learning affects biological evolution (Best, 1999, 2006; Higgs, 1992; Hinton & Nowlan, 1992; Hutchins & Hazlehurst, 1991).



and also behind EVOC, was to distil the underlying logic of not biological evolution but cultural evolution, *i.e.*, the process by which ideas adapt and build on one another in the minds of interacting individuals. Agents do not evolve in a biological sense—they neither die nor have offspring—but do in a cultural sense, by generating and sharing ideas for actions. In cultural evolution, the generation of novelty takes place through invention instead of through mutation and recombination as in biological evolution, and the differential replication of novelty takes place through imitation, instead of through reproduction with inheritance as in biological evolution. EVOC has been used to address such questions as how does the presence of leaders or barriers to the diffusion of ideas affect cultural evolution.

We now summarize briefly the architecture of EVOC in sufficient detail to explain our results; for further details on the model, we refer the reader to previous publications (Gabora, 1995, 2008a, 2008b; Gabora & Leijnen, 2009; Leijnen & Gabora, 2009; Gabora & Saberi, 2011).

### Agents

Agents consist of (1) a neural network, which encodes ideas for actions and detects trends in what constitutes a fit action, (2) a ‘perceptual system’, which carries out the evaluation and imitation of neighbours’ actions, and (3) a body, consisting of six body parts which implement actions. The neural network is composed of six input nodes and six corresponding output nodes that represent concepts of body parts (LEFT ARM, RIGHT ARM, LEFT LEG, RIGHT LEG, HEAD, and HIPS), and seven hidden nodes that represent more abstract concepts (LEFT, RIGHT, ARM, LEG, SYMMETRY, OPPOSITE, and MOVEMENT). Input nodes and output nodes are connected to hidden nodes of which they are instances (*e.g.*, RIGHT ARM is connected to RIGHT.) Each body part can occupy one of three possible positions: a neutral or default positions, and two other positions, which are referred to as active positions. Activation of any input node activates the MOVEMENT hidden node. Same-direction activation of symmetrical input nodes (*e.g.*, positive activation—which represents upward motion—of both arms) activates the SYMMETRY node. In the experiments reported here the OPPOSITE hidden node was not used.

### Invention

An idea for a new action is a pattern consisting of six elements that dictate the placement of the six body parts. Agents generate new actions by modifying their initial action or an action that has been invented previously or acquired through imitation. During invention, the pattern of activation on the output nodes is fed back to the input nodes, and invention is biased according to the activations of the SYMMETRY and MOVEMENT hidden nodes. (Were this not the case there would be no benefit to using a neural network.) To invent a new idea, for each node of the idea currently represented on the input layer of the neural network, the agent makes a probabilistic decision as to whether the position of that body part

will change, and if it does, the direction of change is stochastically biased according to the learning rate. If the new idea has a higher fitness than the currently implemented idea, the agent learns and implements the action specified by that idea.

### Imitation

The process of finding a neighbour to imitate works through a form of lazy (non-greedy) search. The imitating agent randomly scans its neighbours, and adopts the first action that is fitter than the action it is currently implementing. If it does not find a neighbour that is executing a fitter action than its own current action, it continues to execute the current action.

### Evaluation

Following Holland (1975), we refer to the success of an action in the artificial world as its *fitness*, with the caveat that unlike its usage in biology, here the term is unrelated to number of offspring (or ideas derived from a given idea). Fitness of an action is determined using a predefined equation, Eq. 1, that ascribes a range of fitness values from 0 to 10 to the 729 possible actions. (Six body parts that can be in three possible positions gives a total of 729.) The fitness function used in these experiments rewards activity of all body parts except for the head, and symmetrical limb movement. Total body movement,  $m$ , is calculated by adding the number of active body parts, *i.e.*, body parts not in the neutral position. The fitness  $F$  of an action is calculated as follows:

$$F_{nc} = m + 1.5(s_a + s_l) + 2(1 - m_h) \quad (1)$$

$s_a = 1$  if arms are moving symmetrically; 0 otherwise

$s_l = 1$  if legs are moving symmetrically; 0 otherwise

$m_h = 1$  if head is stationary; 0 otherwise

Note that actions have a cultural version of what in biology is referred to as epistasis, wherein what is optimal with respect to one component depends on what is done with respect to another. Epistasis occurs because what is optimal for the left arm depends on what the right arm is doing, and *vice versa*, and same for the legs.

### Learning

Invention makes use of the ability to detect, learn, and respond adaptively to trends. Since no action acquired through imitation or invention is implemented unless it is fitter than the current action, new actions provide valuable information about what constitutes an effective idea. Knowledge acquired through the evaluation of actions is translated into educated guesses about what constitutes a successful action by updating the learning rate. For example, an agent may learn that more overall movement tends to be either beneficial (as with the fitness function used here) or detrimental, or that symmetrical movement tends to be either beneficial (as with the fitness function used here) or detrimental, and bias the generation of new actions accordingly.

## A Typical Run

Fitness of actions starts out low because agents are initially immobile. They are all implementing the same action, with all body parts in the neutral position; thus action diversity is at a minimum. Soon some agent invents an action that has a higher fitness than immobility, and this action gets imitated, so fitness increases. Fitness increases further as other ideas get invented, assessed, implemented as actions, and spread through imitation. The diversity of actions increases due to the proliferation of new ideas, and then decreases as agents hone in on the fittest actions. In the version of the model used here, fitness values hit a ceiling and converge<sup>2</sup>. Thus, over successive rounds of invention and imitation, the agents' actions improve. EVOC thereby models how "descent with modification" occurs in a purely cultural context.

## Experiments

To carry out our investigation of how varying the level of creativity of individuals affects the fitness of ideas in society as a whole, these experiments used a default artificial world: a toroidal lattice with 1024 nodes, each occupied by a single, stationary agent, and a von Neumann neighborhood structure (agents only interacted with their four adjacent neighbors). Creators and imitators were randomly dispersed.<sup>3</sup> Runs lasted 100 iterations.

In an earlier version of EVOC, in which the ratio of inventing and imitating was always the same for all agents, we found that the society as a whole did best when the ratio of creating to imitating was approximately 2:1 (Gabora, 1995). To incorporate individual differences in degree of creativity, we constructed a version of EVOC that enables us to distinguish two types of agents: *imitators*, that only obtain new actions by imitating neighbors, and *creators*, that obtain new actions by either inventing one or by imitating a neighbor. Imitators never invent at all; they simply copy the creators' successful inventions. Thus all new actions are generated by creators. We also made it possible to vary the probability that creators create versus imitate; each agent can be a pure imitator, a pure creator, or something in between. Whereas any given agent is either a creator or an imitator throughout the entire run, the proportion of creators innovating or imitating in a given iteration fluctuates stochastically. The proportion of creators relative to imitators in the society is referred to as  $C$ . The creativity of the creators – that is, the probability that a creator invents a new action instead of imitating a neighbor – is referred to as  $p$ . If a creator decides to invent on a particular iteration, the probability of changing the position of any body part involved in an action is  $1/6$ .<sup>4</sup>

The society consists of three subgroups:

- $C \times p \times N$  creators attempting to innovate
- $C \times (1 - p) \times N$  creators attempting to imitate
- $(1 - C) \times N$  imitators attempting to imitate

where the number of agents,  $N$  is 1024.

In previous investigations we measured, for different values of  $C$  and  $p$ , the diversity of ideas over the course of a run (Gabora, Leijnen, & vonGhyczy, 2012). We found that the cultural diversity, *i.e.*, the number of different ideas implemented by one or more agent(s), was positively correlated with both the proportion of creators to imitators, and with how creative the creators were. We also obtained suggestive evidence that when creators are relatively uncreative, the mean fitness of ideas increases as a function of the percentage of creators in the society, but when creators are highly creative, the society appears to be better off with fewer creators (Leijnen & Gabora, 2009). However, those simulations were performed with small societies (100 agents), and since action fitness was obtained at only one time slice (after 50 iterations) for all ratios of creators to inventors, these results did not reflect the dynamics of the time series. Given a set of series of accumulating value over time, it is unclear which series is most representative. The series cannot be unambiguously ordered unless for each pair of series one strictly dominates the other, and that is not the case here; the curves representing mean fitness at different values of  $\{C, p\}$  increase monotonically but they often cross and re-cross as time progresses. Thus here we present a more extensive investigation of the relationship between creativity and society as a whole that employs a sophisticated solution to the time series problem.

## Analysis

We used time series discounting which associates a "present value" with any future benefit such that the present value of any given benefit diminishes as a function of elapsed time until the benefit is realized (McDonald & Siegel, 1986). The standard approach in financial settings is exponential discounting. Given a series of benefits  $b_t$ , the Net Present Value (NPV) is defined as:

$$NPV(b) = \sum_{t=1}^N r^{t-1} b_t \quad \text{with} \quad 0 < r \leq 1 \quad (2)$$

The discount rate  $r$  is normally set as  $r = (\frac{100+i}{100})^{-1}$  where  $i$  is the interest rate (in percentage) for the unit period that an investor can obtain from a safe investment.

This basic idea was adapted to analyze the benefit accrued by attaining fit actions for different values of  $C$  and  $p$  in EVOC. The first discounting method used was Time-to-Threshold (TTT) discounting. Since all fitness trajectories were monotonically increasing, those that reached a reasonably high threshold  $\tau$  sooner should be valued higher. We measured how many iterations (time to threshold) it took for fitness to reach  $\tau$ . For these runs,  $\tau = 9$  was used as a measure of optimal fitness to allow for a realistic averaging over time.

<sup>2</sup>This is not the case for another version of the model (Gabora & Saberi, 2011).

<sup>3</sup>In other experiments (Leijnen & Gabora, 2009) we investigated the results of clustering creators.

<sup>4</sup>This gave on average a probability of one change per newly invented action, which previous experiments (Gabora, 1995) showed to be optimal.

Whereas imitators need creators, creators should ignore others if they could do better on their own ( $p = 1$ ). In other words, the fitness prospects of creators working alone can be viewed in a manner analogous to the interest yield of treasury bonds in investment decisions. This logic suggests another kind of modification of the standard discounting method. The second adaptation to the basic notion of discounting we refer to as Present Innovation Value (PIV) discounting. Let  $F_t^{C,p}$  be the mean action fitness at period  $t$  for parameter setting  $\{C, p\}$ . Note that  $F_t^{1,1}$  is the fitness expectation with no interaction amongst agents. We define the PIV for any fitness curve as:

$$PIV(F^{C,p}) = -N + \sum_{t=1}^N \frac{F_t^{C,p}}{F_t^{1,1}} \quad (3)$$

Therefore,  $PIV(F^{1,1}) = 0$ ; creators are indifferent to working alone or in a community with imitation.

## Results

All results are averages across 100 runs. The 3D graph and contour plot for the  $\log_{10}$  TTT discounting analysis of the time series for different  $C, p$  settings are shown in Figures 1 and 2 respectively. Note that by definition a low TTT value corresponds to high mean fitness of actions across the society. The TTT method clearly demonstrates a valley in the adaptive landscape. The line running along the bottom of the valley in Figure 2 indicates, for any given value of  $p$  the optimal value for  $C$ , and *vice versa*. When  $p = 1$  the optimal values of  $C = 0.38$ . When  $C = 1$  the optimal values of  $p$  is 0.19. The global optimum is at approximately  $\{C, p\} = \{0.4, 1.0\}$ .

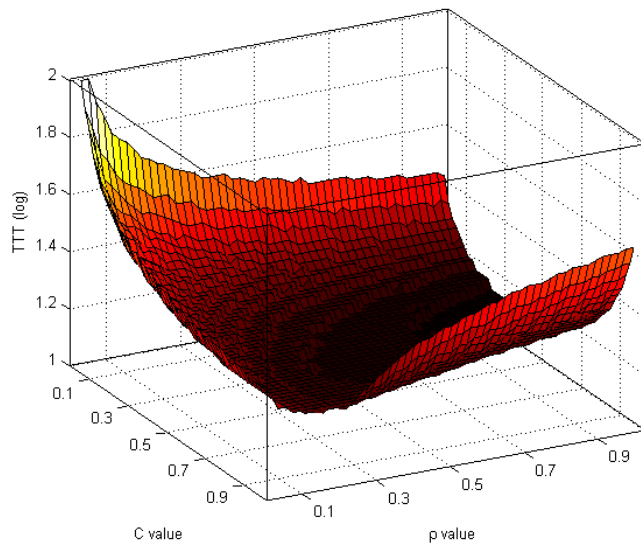


Figure 1: 3D graph of the  $\log_{10}$  Time-to-Threshold (TTT) landscape of the average mean fitness for different values of  $C$  and  $p$ , with  $\tau = 9$ . The valley in the fitness landscape indicates that the optimal values of  $C$  and  $p$  for the society as a whole are less than their maximum values for most  $C, p$  settings.

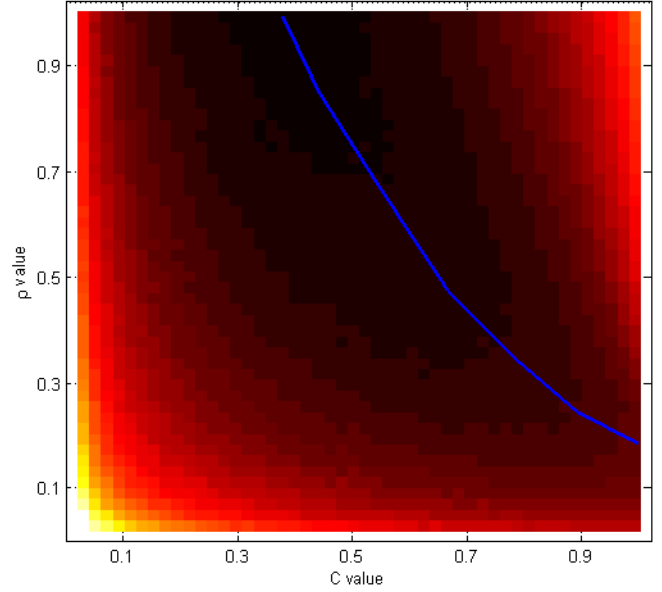


Figure 2: Top-view contour plot of the  $\log_{10}$  Time-to-Threshold (TTT) landscape of the average mean fitness for different values of  $C$  and  $p$ , with  $\tau = 9$ . The line, obtained by visually extrapolating over minimum values  $C$  and  $p$ , indicates the set of optima.

The 3D graph and contour plot for the PIV discounting analysis of the time series for different  $C, p$  settings are shown in Figures 3 and 4 respectively. The pattern is very similar to that obtained with the  $\log_{10}$  TTT discounting analysis.

Thus both  $\log_{10}$  TTT and PIV analyses of the time series showed that, although some creativity is essential to get the fitness of cultural novelty increasing over time, more creativity is not necessarily better. For optimal mean fitness of agents actions across the society there is a tradeoff between  $C$ , the proportion of creators in the artificial society, and  $p$ , how creative these creators are.

## Discussion and Future Directions

This investigation yielded results that contradict the widespread assumption that creativity is necessarily desirable. The model is highly idealized, and caution must be taken in extrapolating to human societies. The PIV results assume that creators avoid input from neighbors if doing so would maximize the fitness of their actions. In reality, creative individuals may not behave so rationally. However, the PIV results were corroborated by the TTT results, indicating that the basic pattern does not depend on the assumption of economic rationality.

EVOC agents are too rudimentary to suffer the affective penalties of creativity but the model incorporates another important drawback to creativity mentioned in the introduction: an iteration spent inventing is an iteration not spent imitating. Creative agents, absorbed in their creative process, effectively rupture the fabric of the artificial society; they act as

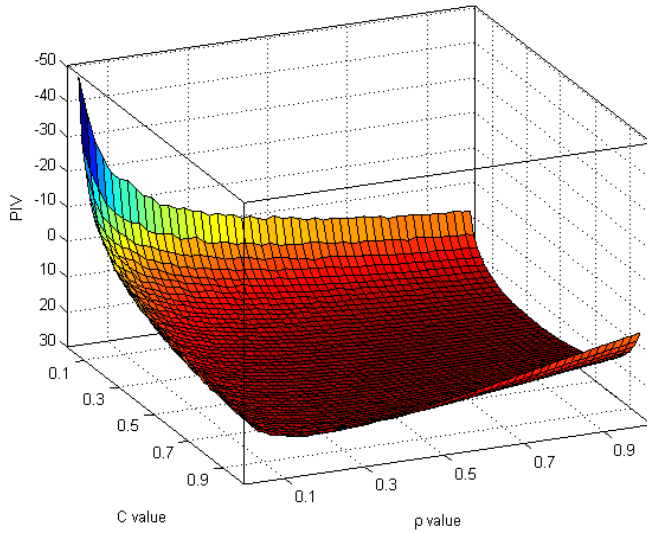


Figure 3: 3D graph of the Present Innovation Value (PIV) landscape of the average mean fitness for different values of  $C$  and  $p$ . Since the  $x$  axis has been inverted to aid visibility of the adaptive landscape, the valley again indicates that the optimal values of  $C$  and  $p$  for the society as a whole are less than their maximum values for most  $C, p$  settings.

insulators that impede the diffusion of proven solutions. Imitators, in contrast, serve as a cultural memory that ensures the preservation of successful ideas. This suggests that the reason people are not more creative than they are is not just because it is difficult to be creative; there is a cost to society as well.

Our results suggest that families, organizations, or societies may self-organize to achieve levels of both imitation and creativity that are intermediate in order to achieve a balance between continuity and change. The results suggest that imitation is neither just the greatest compliment, nor a form of free-riding, but a valuable social mechanism that serves innovators and imitators alike. Without invention there is nothing to imitate, but invention is considerably more effective in conjunction with imitation.

Limitation of this work include that the fitness function was static throughout a run, and agents had only one action to optimize. In real life, there are many tasks, and a division of labor such that each agent specializes in a few tasks, and imitates other agents to carry out other tasks. Another limitation is that EVOC currently does not allow an agent to imitate only certain features of an idea while retaining features the idea it is currently implementing. Creative change can break up co-adapted partial solutions. Recall that actions have a cultural version of what in biology is referred to as epistasis, wherein what is optimal with respect to one component depends on what is done with respect to another. Once both components have been optimized in a mutually beneficial way (for example, the arms are moving symmetrically), excess creativity can cause co-adapted partial solutions to break down. In future studies we will investigate the effects of using a dynamic

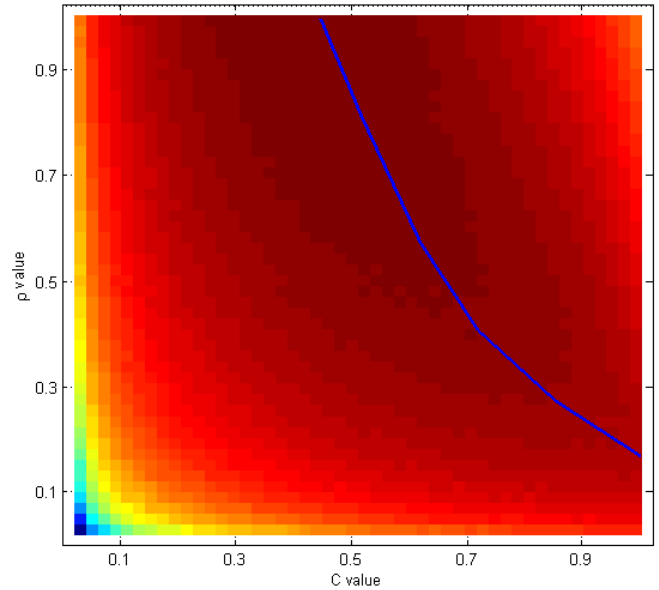


Figure 4: Top-view contour plot of the Present Innovation Value (PIV) landscape of average mean fitness for different values of  $C$  and  $p$ . The line, obtained by visually extrapolating over maximum values  $C$  and  $p$ , indicates the set of optima.

fitness function, and enabling partial imitation. We will also compare our findings to real world data.

If it is the case that social groups can be too creative for their own good, then expensive and widely used programs to enhance creativity through methods such as brainstorming may be counterproductive. The results of these experiments help make sense of findings that creativity is often suppressed in the classroom and in society at large, and that creative individuals often experience discrimination, or worse (Craft, 2005; Cropley & Cropley, 2005; Scott, 1999; Torrance, 1963b, 1963a). (It is well-known that Einstein's dissertation was rejected by the Technische Hochschule in Vienna; he wrote his papers on relativity while working at a patent office.) On the other hand, once the merits of ones' creative efforts become known, this individual's creativity is generally supported or even idolized. In future work we plan to investigate the hypothesis that the social practice of discouraging creativity until the creative individual has proven him- or herself serves as a form of social self-regulation ensuring that creative efforts are not squandered. Specifically, we will use EVOC to test the hypothesis that if individuals who generate creative outputs of low fitness are exposed to social pressures that discourage creativity, and individuals who generate creative outputs of high fitness are encouraged to be creative, the society may self-organize such that it achieves a balance of creative and uncreative individuals (such as the  $C, p$  values indicated by the red line in our experiments).

## Acknowledgments

This work was supported by grants to the first author from the Social Sciences and Humanities Research Council of Canada, the Natural Sciences and Engineering Research Council of Canada, and the Flemish Fund for Scientific Research, Belgium. We thank Tiha von Ghyczy for help with the analysis.

## References

- A. Mesoudi, A. W., & Laland, K. (2006). Toward a unified science of cultural evolution. *Behavioral & Brain Sciences*, 29, 329–347.
- Andreason, N. C. (1987). Creativity and mental illness. prevalence rates in writers and their first degree relatives. *American Journal of Psychiatry*, 144, 1288–1292.
- A. Whiten, K. N. L., R. A. Hinde, & Stringer, C. B. (2011). Culture evolves. *Philosophical Transactions of the Royal Society B*, 366, 938–948.
- Best, M. (1999). How culture can guide evolution: An inquiry into gene/meme enhancement and opposition. *Adaptive Behavior*, 7, 289–293.
- Best, M. (2006). Adaptive value within natural language discourse. *Interaction Studies*, 7, 1–15.
- Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton, NJ: Princeton University Press.
- Craft, A. R. (2005). *Creativity in schools: Tensions and dilemmas*. London: Routledge.
- Cropley, D. H., & Cropley, A. J. (2005). Engineering creativity: A systems concept of functional creativity. In J. C. Kaufman & J. Baer (Eds.), *Faces of the muse: How people think, work and act creatively in diverse domains*. Hillsdale, NJ: Lawrence Erlbaum.
- D. H. Cropley, J. C. K., A. J. Cropley, & Runco, M. (2010). *Creativity in schools: Tensions and dilemmas*. Cambridge: Cambridge University Press.
- Flaherty, A. W. (2005). Frontotemporal and dopaminergic control of idea generation and creative drive. *Journal of Computational Neurology*, 493, 147–153.
- Gabora, L. (1995). Meme and variations: A computational model of cultural evolution. In L. Nadel & D. Stein (Eds.), *1993 lectures in complex systems*. Reading MA: Addison-Wesley.
- Gabora, L. (1996). A day in the life of a meme. *Philosophica*, 57, 901–938.
- Gabora, L. (2008a). Evoc: A computational model of cultural evolution. In *Proceedings of the 30th annual meeting of the cognitive science society* (pp. 18–25). New York: Sheridan Publishing.
- Gabora, L. (2008b). Modelling cultural dynamics. In *Proceedings of the association for the advancement of artificial intelligence (aaai)* (pp. 18–25). Menlo Park, CA: AAAI Press.
- Gabora, L., & Leijnen, S. (2009). How creative should creators be to optimize the evolution of ideas? a computational model. *Electronic Proceedings of Theoretical Computer Science*, 9, 108–119.
- Gabora, L., Leijnen, S., & vonGhyczy, T. (2012). The relationship between creativity, imitation, and cultural diversity. *International Journal of Software and Informatics*, 9, in press.
- Gabora, L., & Saberi, M. (2011). How did human creativity arise? an agent-based model of the origin of cumulative open-ended cultural evolution. In *Proceedings of the acm conference on cognition and creativity* (pp. 299–306). Atlanta, GA: ACM Press.
- Goodwin, F., & Jamieson, K. (1990). New York: Oxford University Press.
- Higgs, P. G. (1992). The mimetic transition: a simulation study of the evolution of learning by imitation. *Proceedings of the Royal Society B - Biological Sciences*, 267, 1355–1361.
- Hinton, G. E., & Nowlan, S. J. (1992). How learning can guide evolution. *Complex Systems*, 267, 495–502.
- Holland, J. (1975). *Adaptation in natural and artificial systems*. Ann Arbor: University of Michigan Press.
- Hutchins, E., & Hazelhurst, B. (1991). Learning in the cultural process. In D. F. C. Langton J. Taylor & S. Rasmussen (Eds.), *Artificial life ii*. Redwood City: Addison-Wesley.
- Leijnen, S., & Gabora, L. (2009). The artist loft effect in the clustering of creative types: A computer simulation. In *Proceedings of the seventh creativity and cognition conference* (pp. 389–390). New York: ACM Press.
- Ludwig, A. M. (1995). *The price of greatness*. New York: Guilford Press.
- Maslow, A. H. (1959). Creativity in self-actualizing people. In H. . Brothers (Ed.), *Creativity and its cultivation*. New York: McGraw-Hill.
- May, R. (1975). *The courage to create*. New York: Bantam.
- McDonald, R., & Siegel, D. R. (1986). The value of waiting to invest. *Quarterly Journal of Economics*, 101, 707–728.
- Moran, S. (2011). The roles of creativity in society. In J. Kaufman & R. Sternberg (Eds.), *Cambridge handbook of creativity*. Cambridge UK: Cambridge University Press.
- Rogers, C. (1959). Toward a theory of creativity. In H. Anderson (Ed.), *Creativity and its cultivation*. New York: Harper & Row.
- Scott, C. L. (1999). Teachers biases toward creative children. *Creativity Research Journal*, 12, 321–337.
- Sternberg, R. J., & Lubart, T. I. (1995). *Defying the crowd: Cultivating creativity in a culture of conformity*. New York: Free Press.
- Sulloway, F. (1996). *Born to rebel*. New York: Pantheon.
- Torrance, E. P. (1963a). *Education and the creative potential*. Minneapolis: University of Minnesota Press.
- Torrance, E. P. (1963b). *Guiding creative talent*. Englewood Cliffs, NJ: Prentice-Hall.

# Does domain size impact speech onset time during reference production?

Albert Gatt (albert.gatt@um.edu.mt)

Institute of Linguistics, University of Malta  
Tilburg center for Cognition and Communication (TiCC), Tilburg University

Roger P.G. van Gompel (r.p.g.vangompel@dundee.ac.uk)

School of Psychology, University of Dundee

Emiel Krahmer (e.j.krahmer@uvt.nl)

Tilburg Center for Cognition and Communication (TiCC), Tilburg University

Kees van Deemter (k.vdeemter@abdn.ac.uk)

Department of Computing Science, University of Aberdeen

## Abstract

In referring to a target referent, speakers need to choose a set of properties that jointly distinguish it from its distractors. Current computational models view this as a search process in which the decision to include a property requires checking how many distractors it excludes. Thus, these models predict that identifying descriptions should take longer to produce the larger the distractor set is, independent of how many properties are required to identify a target. Since every property that is selected is checked, they also predict that distinguishing a target should take longer the more properties are required to distinguish it. This paper tests this prediction empirically, contrasting it with two alternative predictions based on models of visual search. Our results provide support for the predictions of computational models, suggesting a crucial difference between the mechanisms underlying reference production and object identification.

**Keywords:** Referring expressions, language production, visual search, computational modeling

## Introduction

When a speaker refers to a target referent in a visual domain, she identifies it for an addressee by using properties which distinguish it from its distractors. For example, in order to identify the object surrounded by a red border in Figure 1, a speaker needs to refer to it using both its colour and its size (*the large blue aeroplane*); leaving out either of these properties would result in an underspecified description.

Most psycholinguistic accounts of reference in such domains assume that the discriminatory value of properties plays an important role, since the objective is to identify an object for the addressee (Olson, 1970). On the other hand, it is also well-established that certain properties are ‘preferred’ in that speakers often include them when they are not required to distinguish the target, thus producing overspecified descriptions (Pechmann, 1989 ; Belke & Meyer, 2002 ; Arts, 2004).

The present paper is concerned with the mechanisms underlying the selection of properties. Specifically, we ask whether this process is best viewed as a *search*, along

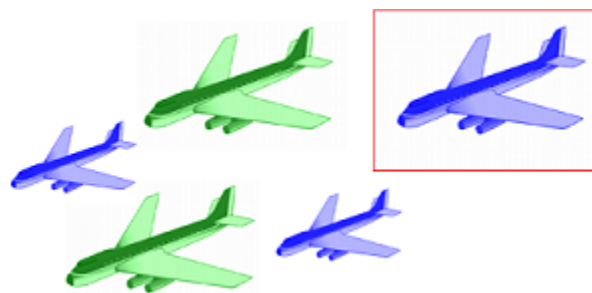


Figure 1: An example domain

the lines suggested by current computational models of Referring Expression Generation (REG; see Krahmer & van Deemter, 2012, for a survey). In these models (described more fully in the next section), the decision to include a property in a description requires checking it against the distractor set to determine whether it excludes at least some of them. If speakers do perform such a procedure, then larger domain sizes should result in more effort (and this should be indicated, for example, by increased speech onset times). This is because more objects have to be checked every time a property is considered for inclusion.

This prediction is compatible with a classic finding in the visual search and attention literature, where search time has been shown to increase linearly with domain size (Treisman & Gelade, 1980). However, whereas REG models predict an impact of domain size irrespective of the number of properties required to distinguish a target referent, the task used by Treisman and Gelade only evinces a linear increase with targets distinguished by a *conjunction* of properties (e.g. *blue and large*). When targets are distinguished by a single property, a ‘pop-out’ effect is observed and domain size has no impact.

Yet a third possibility is suggested by more recent visual search models (e.g. Itti & Koch, 2001), which give a more central role to parallel processing. In these models,



an initial, first-pass overview of a visual domain results in the parallel activation of salient features, forming a saliency map. From a production perspective, such models would predict that, irrespective of how many properties are needed for identification, domain size should have no impact on search time because the salient, contrastive features of a referent could be ‘read off’ the saliency map without exhaustive checking against the distractors.

In short, there are at least three alternative models that could account for how a speaker selects properties, each with different predictions concerning the impact on search time of (i) the size of the visual domain in which the referent must be distinguished, (ii) the number of properties that are required to achieve this. In the remainder of this paper we first discuss these models in more detail and then describe an experiment that sought to investigate these relationships. In our experiment, we focus on speech onset time as an indicator of the amount of search effort required to produce a distinguishing description.

### Three alternative models

Computational models of Referring Expression Generation (REG) are core components of systems which automatically generate text or speech from non-linguistic data (Reiter & Dale, 2000). Their aim is to determine the content of a distinguishing description of a referent in a given domain. REG algorithms usually view this process of content selection as a search (cf. Bohnet & Dale, 2005) and this is often modelled as an incremental procedure (e.g. Dale, 1989 ; Dale & Reiter, 1995, as well as several models based on these). The models that this paper is concerned with focus on ‘first-pass’ references, that is, the generation of initial descriptions in domains which are assumed to be mutually known between speaker and hearer.

The input to a REG algorithm is a domain of objects (such as Figure 1), one of which is the target referent. These algorithms iterate through the available properties (for example, the size and colour) of the objects. Each property is considered as a possible candidate for inclusion in the description.<sup>1</sup>

For each property, the algorithm checks whether there is at least one distractor that it excludes. If this is the case, then the property is included in a description and the distractor set is updated. For example, suppose the first property to be considered in Figure 1

is  $\langle \text{COLOUR:blue} \rangle$ . A check of each of the four distractors shows that there are two non-blue objects. Since it has some discriminatory value, this property is added to the description and the distractor set updated to leave only the two remaining distractors. Since, at this stage, the target referent is not yet distinguished (it is not the only blue object, as shown by the presence of the two remaining distractors), the process does not terminate, but considers the next available property,  $\langle \text{SIZE:large} \rangle$ .<sup>2</sup> Upon checking, the algorithm discovers that the two remaining distractors are both excluded by this property. Since, at this stage, there are no remaining distractors, the procedure terminates with a description that contains both size and colour. From the perspective of the present paper, this content selection procedure makes the following important predictions:

1. Search time should increase linearly with the number of distractors, since each candidate property has to be checked against the distractor set to determine its discriminatory value;
2. Since search is incremental and every candidate property is checked, the effect of domain size should be observed irrespective of whether a distinguishing description contains a single property or a conjunction;
3. Independently of (2) above, the more properties are required to distinguish the target referent, the longer the search time should be, because each property represents a cycle in an iterative search procedure.

The first of these predictions is compatible with classic findings in the visual search literature. Treisman et al. (1980) reported a steep linear increase in search time with increasing domain size in tasks in which participants have to determine whether some object is present in a visual domain in response to a question presented beforehand (e.g. *Is there a red vertical?*). Various replications of this effect have shown that reaction time increases by as much as 31ms with every new object (e.g. Spivey, Tyler, Eberhard, & Tanenhaus, 2001). However, the effect only holds when participants search for a target defined by a conjunction of properties. Single property search (e.g. *Is there a vertical?*) evinces a ‘pop-up’ phenomenon, attributed to parallel activation of salient features, which obviates the need for serial search and integration of multiple features. Interestingly, in the case of conjunction search, the linear increase in search time with domain size is altered if participants hear the description of a target *concurrently* with the presentation of a visual scene (Spivey et al., 2001). In this case, the slope is significantly shallower, perhaps because concurrent presentation of description and domain allows listeners to incrementally circumscribe the search domain

<sup>1</sup>An important factor that distinguishes algorithms from each other is how they prioritise properties during search. For example, Dale et al. (1995) propose to prioritise properties based on the preferences evinced by humans in psycholinguistic experiments (e.g. Pechmann, 1989). By contrast, Dale (1989) proposes a model which prioritises properties based on their discriminatory value. Here, we abstract away from these differences, focusing on the basic search mechanisms common to all of them.

<sup>2</sup>We are assuming that the *large* value of the SIZE attribute is determined as part of this procedure, perhaps by an algorithm along the lines described by Deemter (2006).



as each property is processed (cf. Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995 ; Sedivy, Tanenhaus, Chambers, & Carlson, 1999, for related observations).

The third model we alluded to in the Introduction makes neither of the predictions of REG models. It is possible that speakers default to a ‘fast and frugal’ heuristic whereby, instead of searching for a distinguishing property or combination thereof, they rely on a first-pass overview of the coarse visual features of the domain to select all the properties that seem to have some contrastive value. Such a procedure would be compatible with more recent models of visual search and attention, where salient contrasts are activated through an initial, parallel process that results in a saliency map (Itti & Koch, 2001). Under this model, we would expect the number of properties required to distinguish a referent to have no impact on search time. We would also not expect the function modeling the impact of domain size on search time to increase linearly.

As the foregoing discussion suggests, REG models view the mechanism underlying reference production as fundamentally similar to that of object identification, namely, as a search process. However, there is a crucial difference which accounts for the different predictions made by the two classes of models.

In reference production, search is *object-driven*. A speaker knows which target is intended. Her task, as it is modelled in REG, is to identify a set of properties which are individually discriminatory (each contributes to the overall goal of identification) and jointly distinguishing. Thus, even determining whether a single property is discriminatory requires a check against the distractor set. By contrast, in the standard identification task, search is *property-driven*: a description of the object serves as an instruction to pick out a particular entity. If the description contains only a single feature, a pop-out search is sufficient, for the listener need not verify that the feature in question has discriminatory value – that assumption should follow from the fact that the speaker is being co-operative and is not including redundant information. Indeed, recent work has suggested that the inclusion of properties with no discriminatory value – a strong tendency among speakers, as we discussed above – results in increased processing effort for listeners, suggesting that they do in fact make this assumption (Engelhardt, Baris Demiral, & Ferreira, 2011).

## Experiment

In the experiment, participants were shown visual domains of the kind displayed in Figure 1 and asked to produce a distinguishing description of the target referent. We measured the time it took participants to initiate a description, as a function of the size of the domain and the number of properties (one or two) required to

distinguish the target.

## Participants

The experiment was conducted at the Tilburg center for Cognition and Communication. Forty native speakers of Dutch participated in return for course credit.

## Materials and design

The experimental stimuli consisted of 64 items selected from a version of the Snodgrass and Vanderwart set of line drawings with colour and texture (Rossion & Pourtois, 2004). The items were selected on the basis of a pretest in which seven native speakers of Dutch were asked to name greyscale versions of the pictures. For the items, we selected only those pictures for which at least 5 out of the 7 speakers agreed on the name of the object. These were subsequently manipulated to create versions in different sizes and colours. For each item, 8 versions of a visual domain were constructed, each consisting of a target referent indicated by a red border, and a number of distractors. The 8 versions represented combinations of the following two factors:

- *Properties* (2 levels): Either size only (s) or both colour and size (CS) were required to distinguish the target. Figure 1 is an example of the CS condition.
- *Distractors* (4 levels): There were 2, 4, 8 or 16 distractors in addition to the target, representing increasing domain size.

In each domain, all objects (target and distractors) were of the same type (e.g. all were aeroplanes). In s trials, distractors were identical to the target except for size. Distractors were also identical to each other (e.g. the target was a small blue aeroplane and all distractors were large blue aeroplanes). In the CS trials, half the distractors were identical to the target except for their size and the other half were identical to the target except for their colour (e.g. the target was a large blue aeroplane, half the distractors were small green aeroplanes and the other half were large blue aeroplanes).

In addition to the experimental items there were 108 fillers. In 64 of these, the target could be distinguished using size only or both size and colour, as in the trials. However, there was variation in the types of distractors (not all distractors were of the same type as the target). In the remaining 64 fillers, the target could be distinguished by using its type only. There were equal numbers of fillers containing 2, 4, 8 or 16 distractors.

In each trial, objects were presented in a sparse grid. For each item, the position of the target was fixed in advance and was the same in all conditions. The position of the distractors was also fixed in the 2-, 4-, 8- and 16-distractor conditions. Both items and participants were randomly divided into 8 groups. Each participant saw exactly 8 items in each condition; item and participant groups were rotated through a latin square so that each

item was seen in a condition by an equal number of participants.

## Procedure

Participants did the experiment individually in a sound-proof booth, wearing a headset through which their descriptions were recorded. The experiment was run using the DMDX package for stimulus presentation (Forster & Forster, 2003). They were asked to imagine that they were describing objects for a listener who could see the exact same objects but did not know which one was the target referent. In order to avoid the use of descriptions containing locative expressions, participants were also told that their putative listener would see the objects in different positions (none of the participants used locatives).

A trial was initiated with a warning bell and a fixation cross appearing for 500ms in the middle of the screen. Subsequently, the visual domain appeared. After they had described the target, participants pressed the Enter key on their keyboard to move to the next trial.

Trials were presented in two blocks to allow participants to take a break. Speech onset time was measured using the DMDX voice trigger from the point when the visual domain was presented to the point when a participant began to speak.

## Data pre-processing

Descriptions were transcribed and annotated for whether they contained size, colour or both. Descriptions in the S condition which contained both size and colour were classified as overspecified. Descriptions in the S condition which contained only colour, or those in the CS condition which contained only one of the two properties, were classified as underspecified. All other descriptions were classified as well-specified. Data from two participants was excluded because they produced utterances which compromised the calculation of speech onset time (for example, starting all of their descriptions with *I see a...*). The remaining 38 participants produced well-specified descriptions 71% of the time, with 27% overspecified descriptions and 2% underspecifications. The relative frequency of over- compared to underspecifications is to be expected, given previous work (see the Introduction).

Speech onset times were manually tuned using Check-Vocal (Protopapas, 2007), a program for the detection and correction of voice key mistriggers (due to lip smacks, coughs, background noise etc) in DMDX result files. For each sound file, we ensured that the speech onset time was taken at the precise point where the participant's description began. In case the description included a determiner, this meant the onset of the determiner. In case a description began with a hesitation (e.g. *uhhhh het kleine rode bed*), the onset time was still the onset of the description, that is, following the initial hesitation.

Following tuning, an onset time was defined as an outlier if it exceeded the mean  $\pm 2SD$  in its condition. 106 data points (4.4%) were considered outliers by this criterion and were treated as missing.

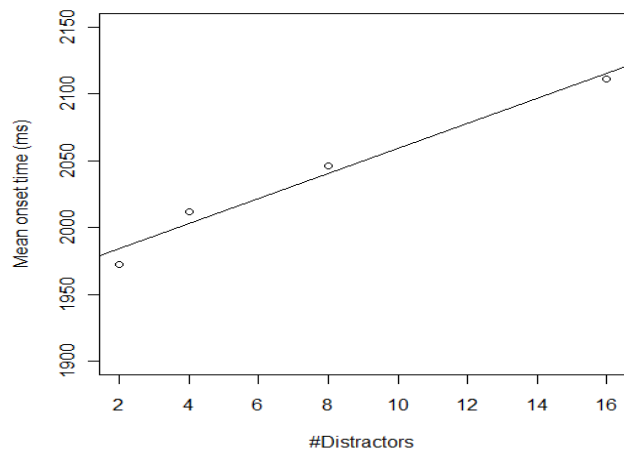


Figure 2: Mean onset times by number of distractors.

## Results

In what follows, we report results based on descriptions which were well-specified, excluding over- and underspecified cases.<sup>3</sup> This is because in the case of underspecified descriptions, participants presumably did not check against the distractor set to see whether a selected property combination was distinguishing; where participants overspecified, the inclusion of a redundant property may not have involved such a check because it was extra information.

Table 1 displays mean speech onset times in each Distractor condition for different levels of Property as well as overall, while Figure 2 displays the relationship between onset times and Distractors. The means show an increase in speech onset time in the CS compared to the S condition. As the Figure shows, the relationship between speech onset time and domain size appears linear.

We report a linear mixed effects analysis with Properties and Distractors as fixed effects, and random intercepts for participants and items.<sup>4</sup> The Properties factor was scaled and centred; Distractors was treated as a continuous variable, since our aim is to model the change in

<sup>3</sup>The statistical tests reported here were also conducted on the full dataset; the general trends are identical to the ones reported here.

<sup>4</sup>A comparison of the model we report here to one with a random intercept and slope for each participant showed that the latter did not provide a better fit to the data (model  $\chi^2 = 8.05, p > .5$ ); neither did adding a random intercept and slope for each item ( $\chi^2 = 3.34, p > .9$ ).

	2	4	8	16	overall
CS	2022.91 (492)	2022.44 (507)	2105.93 (525)	2139.67 (809)	2073 (538)
S	1872.30 (458)	1990.95 (566)	1921.73 (209)	2046.73 (473)	1955 (506)
overall	1972.35 (486)	2011.97 (527)	2046.45 (527)	2111.16 (572)	—

Table 1: Mean speech onset times and standard deviations in each condition

speech onset times as a function of continuous increases in domain size.

There were strong main effects of both Properties ( $t = -3.40, p < .001$ ) and Distractors ( $t = 3.72, p < .001$ ), but no interaction ( $t = .22, p > .8$ ). Thus, both the increase in speech onset time with two properties compared to one, and the increase with domain size are reliable.<sup>5</sup> To further investigate the nature of the effect of Distractors, we carried out planned comparisons by re-running a linear mixed effects analysis with Distractors as the only fixed effect and random intercepts for participants and items. For this model, Distractors was recoded as a factor using forward difference coding to perform contrasts between adjacent levels. None of these contrasts proved significant, although the difference between domains with 8 and 16 distractors approached significance ( $p = .06$ ). Post-hoc pairwise comparisons using t-tests with Bonferroni adjustment showed a significant difference between domains with 2 distractors and those with 8 ( $p = .03$ ) and those with 16 ( $p = .004$ ) distractors, but no other differences. This suggests that the primary contrast is between relatively small domains and those with many more objects in the visual display.

## Discussion

Our experimental results show that the time speakers take to produce references is influenced both by the number of distractors from which they have to distinguish the referent and by the number of properties that they need to include in their description in order for it to be identifying.

The effect of domain size was very strong and crucially did not interact with the number of properties required to distinguish the target referent. This is compatible with the predictions of computational REG models, in which every property needs to be checked against the distractor set in order to determine whether or not it contributes to the goal of identifying a target referent. By contrast, the literature on visual search and object identification only reports robust effects of domain size with conjunctions. In our initial discussion, we suggested that this is in part due to the difference between the task of reference production and that of object identification or online reference resolution. A speaker needs to ensure that every selected property contributes to the

overall referential goal.<sup>6</sup> By contrast, there is no need to check contrastive value in a task involving search and identification for an object based on an instruction (e.g. Treisman & Gelade, 1980) or a definite description (as in the reference resolution experiments of Tanenhaus et al., 1995). Here, reliance on pop-up search is a good strategy for targets identified by a single feature (Treisman & Gelade, 1980) whereas, when a description is presented concurrently with a visual domain, each property in a description can be used to circumscribe the set of relevant objects (Tanenhaus et al., 1995 ; Spivey et al., 2001). Thus, the search mechanisms of speakers and listeners are qualitatively different.

A second prediction in relation to domain size was that its effect on speech onset time would be linear. Figure 2 does suggest a linear effect, although the increase becomes less steep as domain size grows larger. We also do not find differences between adjacent levels of the Distractors factor, with post-hoc analyses showing that the primary differences are between the smallest and the largest domains. Further research is required to confirm these findings. Nevertheless, the main effect, which was obtained by modelling Distractors as a continuous variable, does support the predictions of REG algorithms. It also runs counter to the predictions of models of visual search based on parallel activation of salient, contrastive properties (the third class of models discussed at the beginning of this paper), which would predict no main effect.

Finally, the finding that speakers take longer to initiate descriptions containing two properties, compared to only one, also supports the incremental search procedure in REG models, where each candidate property is checked against the distractor set. Although our experimental design does not allow us to determine whether the effect of properties is linear or not, REG models do make an interesting prediction in this regard.

Consider the effect of an incremental procedure of the

<sup>5</sup>The p-values for the LME models were estimated using Baayen's (2008) `pvals.fnc` function, included in the `languageR` package.

<sup>6</sup>The fact that speakers overspecify (that is, add properties that do not contribute to identification) may be due to the incremental nature of reference production, whereby speakers include properties one by one, starting from properties (such as colour) which are highly 'preferred' (Pechmann, 1989). Such properties may have discriminatory value individually, but may turn out to be redundant once the description has been fully formulated. Note, however, that even under this account – which is essentially the account incorporated by the incremental REG models we reviewed above (Dale & Reiter, 1995) – speakers would still check a property for its individual contrastive value.

sort we have described (Dale, 1989 ; Dale & Reiter, 1995) when it selects two or more properties. Since discriminatory value is a prerequisite for selection, every property that is selected results in a decrease in the size of the distractor set, an effect akin to the incremental domain circumscription that Spivey et al. (2001) suggest as an interpretation of their results. Thus every property that is selected leaves fewer objects against which to check the next candidate property, predicting that the effect of properties should be non-linear. This is a possibility that we intend to explore in future work, by including conditions where more than two properties are required to identify the referent.

## Conclusions and future work

This paper focused on the predictions of computational models of reference production, comparing them to some well-known findings in the visual search and attention literature. We reported an experiment that showed that speakers take longer to refer to an object the more properties they require to distinguish it, and the more objects there are in the domain. Our findings lend some support to current computational models, and also highlight some important differences between the search mechanisms involved in reference production and those involved in object identification or reference resolution.

We have identified a number of avenues for future work. In the medium term, we plan to investigate the effect of domain size further in order to determine more precisely the nature of the relationship between domain size and search time. We also plan to investigate the nature of the effect of properties on search time, testing the predictions made by incremental computational models on the effect of adding properties to a description.

## Acknowledgments

This work forms part of the project *Bridging the gap between psycholinguistics and computational linguistics: The case of Referring Expressions*. Albert Gatt and Emiel Krahmer are supported by a grant from the Netherlands Organization for Scientific Research (NWO). Kees van Deemter is supported by the EPSRC Platform Grant *Affecting people with Natural Language*.

## Références

- Arts, A. (2004). *Overspecification in instructive texts*. Thèse de doctorat non publiée, Tilburg University.
- Baayen, R. (2008). *Analyzing linguistic data*. Cambridge : Cambridge University Press.
- Belke, E., & Meyer, A. (2002). Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing times during same-different decisions. *European Journal of Cognitive Psychology*, 14(2), 237–266.
- Bohnet, B., & Dale, R. (2005). Viewing referring expression generation as search. In *Proc. IJCAI'05*.
- Dale, R. (1989). Cooking up referring expressions. In *Proc. ACL'89*.
- Dale, R., & Reiter, E. (1995). Computational interpretation of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(8), 233–263.
- Deemter, K. van. (2006). Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2), 195–222.
- Engelhardt, P. E., Baris Demiral, S., & Ferreira. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, 77(2), 304–314.
- Forster, K. I., & Forster, J. C. (2003). Dmdx: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35, 116–124.
- Itti, L., & Koch, C. (2001, March). Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3), 194–203.
- Krahmer, E., & van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1), 173–218.
- Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, 77, 257–273.
- Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27, 89–110.
- Protopapas, A. (2007). Check vocal: A program to facilitate checking the accuracy and response time of vocal responses from dmdx. *Behavior Research Methods*, 39(4), 859–862.
- Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge : Cambridge University Press.
- Rossion, B., & Pourtois, G. (2004). Revisiting snodgrass and vanderwarts object databank : the role of surface detail in basic level object recognition. *Perception*, 33, 217–236.
- Sedivy, J. G., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71, 109–147.
- Spivey, M., Tyler, M., Eberhard, K., & Tanenhaus, M. (2001). Linguistically mediated visual search. *Psychological Science*, 12, 282–286.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. G. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632–1634.
- Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136.

# Ping Pong in Church: Productive use of concepts in human probabilistic inference

Tobias Gerstenberg<sup>1</sup> (t.gerstenberg@ucl.ac.uk) & Noah D. Goodman<sup>2</sup> (ngoodman@stanford.edu)

<sup>1</sup>Cognitive, Perceptual and Brain Sciences, University College London, London WC1H 0AP

<sup>2</sup>Department of Psychology, Stanford University, Stanford, CA 94305

## Abstract

How do people make inferences from complex patterns of evidence across diverse situations? What does a computational model need in order to capture the abstract knowledge people use for everyday reasoning? In this paper, we explore a novel modeling framework based on the probabilistic language of thought (PLOT) hypothesis, which conceptualizes thinking in terms of probabilistic inference over compositionally structured representations. The core assumptions of the PLOT hypothesis are realized in the probabilistic programming language Church (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008). Using “ping pong tournaments” as a case study, we show how a single Church program concisely represents the concepts required to specify inferences from diverse patterns of evidence. In two experiments, we demonstrate a very close fit between our model’s predictions and participants’ judgments. Our model accurately predicts how people reason with confounded and indirect evidence and how different sources of information are integrated.

**Keywords:** inference; reasoning; causality; language of thought; probabilistic programming

## Introduction

People often make surprisingly accurate inferences about a person’s latent traits from very sparse evidence. If the second author (NG) loses to the first author (TG) in a ping pong match and afterwards wins against two other lab members, we are fairly confident that TG is a strong player despite only having observed him winning a single game. However, if we consequently find out that NG felt a bit lazy in his match against TG and did not try as hard as he normally does, our belief about TG’s strength might change. This reasoning is not limited to a particular set of potential players, it can be generalized to related situations (such as team matches), and it supports inferences from complex combinations of evidence (e.g. learning that NG was lazy whenever he played a match against a team that included TG) – human reasoning is remarkably *productive*.

How can we best model the flexible inferences people draw from diverse patterns of evidence such as the outcomes of matches in a ping pong tournament? What assumptions about the cognitive system do we need to make to be able to explain the productivity and gradedness of inference? What is the minimum level of abstraction that mental representations need to exhibit in order to support the inferential flexibility that our cognitive machinery displays?

There are two traditional, but fundamentally different ways of modeling higher-level cognition, each with its own strengths and drawbacks: Statistical approaches (e.g. Rumelhart & McClelland, 1988) support graded probabilistic inference based on uncertain evidence but lack some of the representational powers of more richly structured symbolic approaches. Symbolic approaches (e.g. Newell, Shaw, & Simon, 1958), on the other hand, are confined to operating

in the realm of certainty and are ill-suited to modeling people’s inferences in a fundamentally uncertain world. More recently, researchers have started to break the dichotomy between statistical and symbolic models (Anderson, 1996) and have shown that much of cognition can be understood as probabilistic inference over richly structured representations (Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

For instance, causal Bayesian networks (CBN; Pearl, 2000) have been proposed as a modeling framework that combines the strengths of both statistical and symbolic approaches. Given a particular representation of a task that the cognitive system faces, a CBN supports inferences about the probability of competing hypotheses for many different patterns of evidence. However, a CBN is limited to the specific situation it was designed to model, allowing inferences from different observations of existing variables, but not from fundamentally different combinations of objects or events. While some attempts have been made to model more abstract knowledge by constructing CBNs with richer, hierarchical structures (Kemp & Tenenbaum, 2009) or by combining CBNs with propositional logic (Goodman, Ullman, & Tenenbaum, 2011; Griffiths, 2005), CBNs have only coarse-grained compositionality insufficient to support productive extensions over different objects and situations.

Human thought, in contrast, is characterized by an enormous flexibility and productivity (Fodor, 1975). We can flexibly combine existing concepts to form new concepts and we can make use of these concepts to reason productively about an infinity of situations. The *probabilistic language of thought* (PLOT) hypothesis (Goodman & Tenenbaum, in prep) posits that mental representations have a language-like compositionality, and that the meaning of these representations is probabilistic, allowing them to be used for thinking and learning by probabilistic inference. This view of the representation of concepts provides a deeper marriage of the statistical and symbolic view. Because they are probabilistic, they support graded reasoning under uncertainty. Because they are language-like, they may be flexibly recombined to productively describe new situations. For instance, we have a set of concepts, such as “strength” and “game”, in the ping pong domain that we may compose together and apply to symbols such as TG. These combinations then describe distributions on possible world states, which we may reason about via the rules of probability. The PLOT hypothesis has been realized in existing computational systems, including the probabilistic programming language Church (Goodman et al., 2008). Church has several features that enable it to model productive inference from a small set of concepts – in particular, it allows reasoning about placeholder symbols and the forming of complex evidence by composing the concepts.

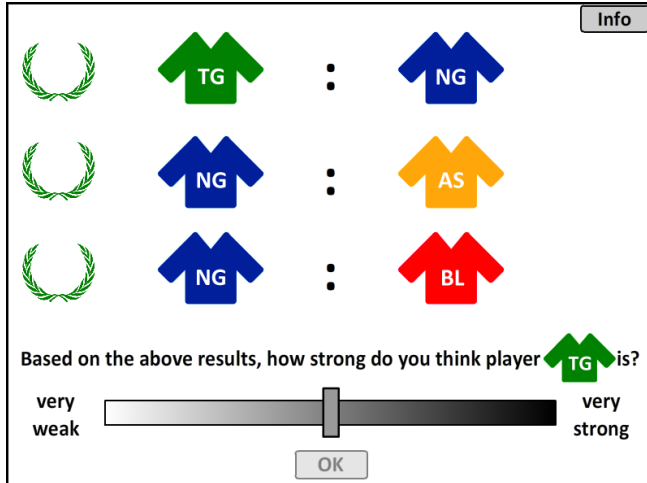


Figure 1: Screenshot of a single player tournament. The winner of each match is indicated by a laurel wreath.

In this paper, we use Church (Goodman et al., 2008), as an instantiation of the PLoT, to explain aspects of people’s flexible concept use, and use the ping pong scenario as a simple case study to illustrate our key points while admitting quantitative empirical evaluation. In two separate experiments, we test the predictions of our modeling approach by examining people’s inferences based on complex patterns of causal evidence. We conclude by pointing out areas of research that are likely to benefit from this modeling framework.

## Modeling probabilistic inferences in Church

Figure 1 shows an example of the inference task that participants faced in the experiments which we will describe below. What representation would be needed to (a) be sensitive to the statistical nature of the evidence and (b) capture the abstract, symbolic structure that remains invariant between this particular situation and other similar situations that could involve different players and different outcomes? Figure 2 shows the Church code that we used to model people’s inferences about a player’s strength based on the results of ping pong tournaments. We chose the ping pong environment because it can be summarized by a relatively simple but rich set of concepts that support productive inferences from a variety of evidence in a variety of situations. We will first introduce the Church language and then explain how this representation captures our intuitive concepts of ping pong.

Church is based on the  $\lambda$ -calculus, with a syntax inherited from the LISP family of languages (McCarthy, 1960). Thus operators precede their arguments, and are written inside grouping parentheses: `(+ 1 2)`. We use `define` to assign values to symbols in our program and `lambda` for creating functions. We could, for example, create a function `double` that takes one number as an input and returns its double. The code would look like this: `(define double (lambda (x) (+ x x)))`. What differentiates Church from an ordinary programming language is the inclusion of random primitives. For example, the function `(flip 0.5)` can be interpreted as a simple coin flip with a weight outputting either

```
(mh-query 1000 100 ;Monte Carlo Inference
;CONCEPTS
(define personstrength (mem (lambda (person) (gaussian 10 3))))
(define lazy (mem (lambda (person game) (flip 0.1))))
(define (teamstrength team game)
  (sum (map (lambda (person)
    (if (lazy person game)
      (/ (personstrength person) 2)
      (personstrength person)))
    team)))
(define (winner team1 team2 game)
  (if (< (teamstrength team1 game)
    (teamstrength team2 game))
    'team2 'team1))
;QUERY
(personstrength 'A)
;EVIDENCE
(and
 (= 'team1 (winner '(TG) '(NG) 1))
 (= 'team1 (winner '(NG) '(AS) 2))
 (= 'team1 (winner '(NG) '(BL) 3))
 (lazy '(NG) 1) ;additional evidence, used in Experiment 2
)
```

Figure 2: Church model of the ping pong scenario.

true or false. Every time the function is called, the coin is flipped afresh. A Church program specifies not a single computation, but a distribution over computations, or sampling process. This *sampling semantics* (see Goodman et al., 2008, for more details) means that composition of probabilities is achieved by ordinary composition of functions, and it means that we may specify probabilistic models using all the tools of representational abstraction in a modern programming language.

We now turn to describing the concepts (see `CONCEPTS` in Figure 2) that are required to represent the ping pong domain (Figure 1). This simple sports domain is built around people, teams and games. In Church, we can use symbols as placeholders for unspecified individuals of these types. This means that we do not need to define in advance how many people participate, what the size of the teams will be, or how many games a tournament will have. We define an individual player’s strength, `personstrength`, via a function that draws from a Gaussian distribution with  $M = 10$  and  $SD = 3$ . The memoization operator `mem` ensures that the strength value assigned to a person is persistent and does not change between games. We next make the assumption that players are sometimes `lazy`. The chance of a person being lazy in a particular game is 10%, specified by using the function `flip` with a weight of 0.1. As mentioned above, we also want to allow for the possibility that individual players form teams – we thus need the overall strength of a team,

Table 1: Modeling assumptions.

concept	description	assumption
<code>personstrength</code>	strength of a player	normally distributed, persistent property
<code>lazy</code>	chance that a player is lazy	$p(\text{lazy}) = 10\%$ , not persistent
<code>teamstrength</code>	strength of a team	individual strengths combine additively
<code>winner</code>	winner of a match	team with greater strength wins



Table 2: Patterns of observation for the single player tournaments. *Note:* An additional set of 4 patterns was included for which the outcomes of the games were reversed. The bottom row shows the omniscient commentator’s information in Experiment 2.

confounded evidence (1,2)	strong indirect evidence (3,4)	weak indirect evidence (5,6)	diverse evidence (7,8)
A > B	A > B	A > B	A > B
A > B	B > C	B < C	A > C
A > B	B > D	B < D	A > D
lazy,game: B,2	B,1	B,1	C,2

*Note:* A > B means that A won against B.

**teamstrength.** Here, we define the team’s strength as the sum of the strength of each person in the team. If a person in the team is lazy, however, he only plays with half of his actual strength. The way in which we can define new concepts (e.g. **teamstrength**) based on previously defined concepts (**personstrength** and **lazy**) illustrates the compositionality of Church. Finally, we specify how the **winner** of a game is determined. We simply say the the team wins who has the greater overall strength. This set of function definitions specifies a simple lexicon of concepts for reasoning about the ping pong domain. The functions are built up compositionally, and may be further composed for specific situations (see below). What’s more, the set of concept definitions refers to people (teams, etc.) without having to declare a set of possible people in advance: instead we apply generic functions to placeholder symbols that will stand for these people. Table 1 concisely summarizes our modeling assumptions.

Now we have a lexicon of concepts (**CONCEPTS**) that we may use to model people’s inferences about a player’s strength (**QUERY**) not only in the situation depicted in Figure 1 but in a multitude of possible situations with varying teams composed of several people, playing against each other with all thinkable combinations of game results in different tournament formats (**EVIDENCE**). This productive extension over different possible situations including different persons, different teams and different winners of each game, renders the Church implementation a powerful model for human reasoning.

A program in Church can be seen as a formal description of the process that generates observed or hypothesized evidence. The **mh-query** operator specifies a conditional inference. Both the evidence provided and the question we are asking are composed out of the concepts that specify the domain. Church completely separates the actual process of inference from the underlying representations and the inferences they license. This allows the modeler to focus on defining the conceptual representation of the domain of interest without having to worry about the exact details of how inference is carried out; it also provides a framework for psychological investigation of representations and the inferences that may be drawn, without committing to *how* these inferences are made – a well-formed level of analysis between Marr’s computational and algorithmic levels (Marr, 1982).

Table 3: Patterns of observation for the two-player tournaments. *Note:* An additional set of 6 patterns was included in which the outcomes of the games were reversed.

confounded with partner (9,10)			confounded with opponent (11,12)			strong indirect evidence (13,14)		
AB	>	CD	AB	>	EF	AB	>	EF
AB	>	EF	AC	>	EG	BC	<	EF
AB	>	GH	AD	>	EH	BD	<	EF
weak indirect evidence (15,16)			diverse evidence (17,18)			round robin (19,20)		
AB	>	EF	AB	>	EF	AB	>	CD
BC	>	EF	AC	>	GH	AC	>	BD
BD	>	EF	AD	>	IJ	AD	>	BC

Hence, in contrast to other frameworks for building psychological models of cognition, such as ACT-R (Anderson, 1996), Church does not incorporate any assumptions about how exactly the cognitive system carries out its computations but merely postulates that inference accords with the rules of probability.

## Experiment 1: Bayesian Ping Pong

In Experiment 1, we wanted to explore how well our simple Church model predicts the inferences people make, based on complex patterns of evidence in different situations. Participants’ task was to estimate an individual player’s strength based on the outcomes of different games in a ping pong tournament. Participants were told that they will make judgments after having seen single player and two-player tournaments. The different players in a tournament could be identified by the color of their jersey as well as their initials. In each tournament, there was a new set of players. Participants were given some basic information about the strength of the players which described some of the modeling assumptions we made (cf. Table 1). That is, participants were told that individual players have a fixed strength which does not vary between games and that all of the players have a 10% chance of not playing as strongly as they can in each game. This means that even if a player is strong, he can sometimes lose against a weaker player.

**Participants** 30 (22 female) recruited through Amazon Mechanical Turk participated in the experiment. The mean age was 31.3 (*SD* = 10.8).

**Materials and Procedure** The experiment was programmed in Adobe Flash CS5.<sup>1</sup> Participants viewed 20 tournaments in total. First, one block of 8 single player tournaments and then another block of 12 two-player tournaments. The order of the tournaments within each block was randomized. Participants could remind themselves about the most important aspects of the experiment by moving the mouse over the Info field on the top right of the screen (see Figure 1). Based on the results of

<sup>1</sup>Demos of both Experiments can be accessed here: [http://www.ucl.ac.uk/lagnado-lab/experiments/demos/BPP\\_demos.html](http://www.ucl.ac.uk/lagnado-lab/experiments/demos/BPP_demos.html)



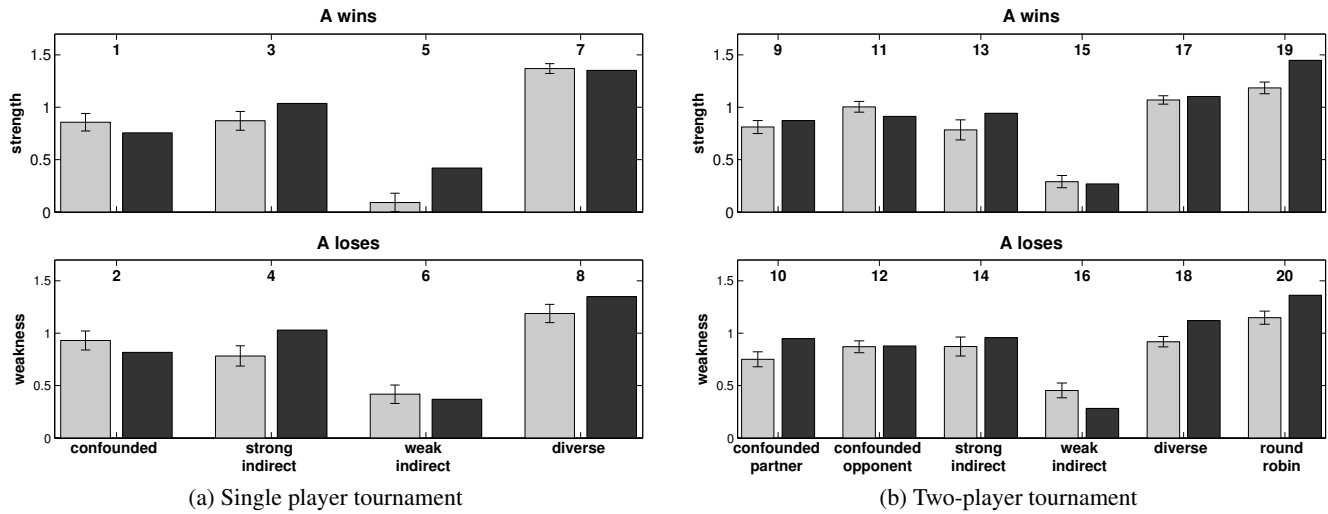


Figure 3: Mean strength estimates (grey bars) and model predictions (black bars) for the single player (left) and two-player tournaments (right). Numbers above the bars correspond to the patterns described in Tables 2 and 3. Error bars are  $\pm 1$  SEM.

the three matches in the tournament, participants estimated the strength of the indicated player on a slider that ranged from -50 to 50. The endpoints were labelled “very weak” and “very strong”. It took participants 7.4 ( $SD = 3.3$ ) minutes to complete the experiment.

**Design** Table 2 shows the patterns of evidence that were used for the single player tournaments. Table 3 shows the patterns for the two-player tournaments. In all tournaments, participants were asked to judge the strength of player A.

For the single player tournaments, we used four different patterns of evidence: *confounded evidence* in which A wins repeatedly against B, *strong* and *weak indirect evidence* where A only wins one match herself but B either continues to win or lose two games against other players and *diverse evidence* in which A wins against three different players. For each of those patterns, we also included a pattern in which the outcomes of the games were exactly reversed.

For the two player tournaments, we used six different patterns of evidence: In some situations A was always in the same team as B (*confounded with partner*) while in other situations A repeatedly played against the same player E (*confounded with opponent*). As in the single player tournaments, we also had patterns with mostly indirect evidence about the strength of A by having his partner in the first game, B, either win or lose against the same opponents with different teammates (*weak/strong indirect evidence*). Finally, we had one pattern of *diverse evidence* in which A wins with different teammates against a new set of opponents in each game and one *round robin* tournament in which A wins all his games in all possible combinations of a 4-player tournament.

## Results and Discussion

In order to directly compare the model predictions with participants’ judgments we z-scored the model predictions and each individual participant’s judgments. Furthermore, we reverse coded participants’ judgments and the model predic-

tions for the situations in which the outcomes of the games were reversed so that both strength and “weakness” judgments go in the same direction.

Figure 3 shows the mean strength estimates (gray bars) together with the model predictions (black bars) for the single and two-player tournaments. The top panels display the situations in which A won his game(s). The bottom panels show the situations in which A lost. Our model predicts participants’ judgments in the single and two-player tournaments very well with  $r = .98$  and  $RMSE = .19$ . A very high median correlation with individual participants’ judgments of  $r = .92$  shows that the close fit is not merely due to an aggregation effect.

In describing the data qualitatively, we will focus on the strength judgments in the top panels (strength and weakness judgments were highly correlated,  $r = .96$ ). In the single player tournaments, A is judged equally strong when he repeatedly wins against the same player (situation 1) or when strong indirect evidence was provided (3). A is judged weakest when only weak indirect evidence is provided (5). A is judged to be strongest when she won against three different players (7). In the two-player tournaments, A is judged equally strong when the evidence is confounded with the partner or opponent and when strong indirect evidence is provided (9, 11 and 13). A is judged to be relatively weak when only weak indirect evidence is provided (15). A is judged to be strong for the situations in which participant’s received diverse evidence about A’s strength (17) and even stronger for the round robin tournament (19).

There appears to be only one prediction that the model makes which is not supported by the data. In the single player tournaments, the model predicts that participants should be slightly more confident about the strength of A when provided with strong indirect evidence (situations 3, 4) compared to when confounded evidence is given (situations 1, 2). However, there is no significant difference between participants’

judgments for strong indirect evidence ( $M = 26.2$ ,  $SD = 15.4$ ) compared to confounded evidence ( $M = 27.8$ ,  $SD = 13.8$ ),  $t(29) = 0.44$ ,  $p > .05$ .

The results of Experiment 1 show that our model predicts participants' inferences very accurately. We have demonstrated that a single and concise representation of the task is sufficient to predict people's inferences for a great diversity of patterns of evidence.

The close fit between our model and participants' inference also shows that our modeling assumptions (e.g. that the team's strength is a linear combination of the individual team members' strengths) generally matched participants' implicit assumptions (cf. Table 1). However, the fact that the model's prediction of a difference between strength judgments based on strong indirect evidence versus confounded evidence was not supported by the data, suggests that participants might have differed in the extent to which they took the chance of laziness into consideration. In fact, only 16 out of 30 participants showed the pattern in the predicted direction. If we increase the probability of a person being lazy in a particular game in the model, it matches participants' average judgments for these situations. Intuitively, if the chances of a person having been lazy in a particular game are increased, there is a higher chance that player A won his game against player B in situation 3 because B was lazy in this round. However, when A wins repeatedly against B, there is hardly any effect of changing the probability of laziness. For example, it is very unlikely when A won three times against B, that B (and not A) was lazy three times in a row.

## Experiment 2: Omniscient Commentator

In Experiment 1 we have shown that our model accurately predicts participants' inferences for a great variety of patterns of evidence from different combinations of teams and outcomes. A still greater variety of evidence is available by composing the basic concepts together in different ways: there is no reason for evidence not to directly refer to a player's strength, laziness, etc. While in Experiment 1, the match results were the only source of information participants could use as a basis for their strength judgments, Experiment 2 introduced an omniscient commentator who gave direct information about specific players. After participants saw a tournament's match results, an omniscient commentator, who always told the truth, revealed that one player was lazy in a particular game. We were interested in how participants updated their beliefs about the strength of player A given this additional piece of evidence. Importantly, we do not need to change anything in the Church code to derive predictions for these situations since all the necessary concepts are already defined.

**Participants** 20 (11 female) recruited through Amazon Mechanical Turk participated in the experiment. The mean age was 34 ( $SD = 9.8$ ).

**Materials, Procedure and Design** Participants viewed 10 single player tournaments which comprised the 8 situations

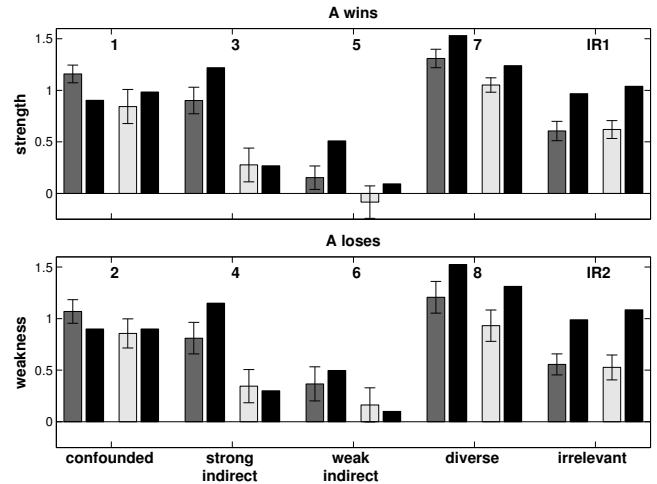


Figure 4: Mean strength estimates and model predictions. Dark grey bars = estimates after tournament information only, light grey bars = estimates after omniscient commentator info, black bars = model predictions. Error bars are  $\pm 1$  SEM.

used in Experiment 1 plus two additional patterns (IR 1, 2). Participants first judged player A's strength based merely on the match results in the tournament. Afterwards, participants received information from the omniscient commentator about one player who was lazy in a particular match. Participants then rated A's strength for a second time, whereby the slider was initialized at the first judgment's position. It took participants 9.4 ( $SD = 4$ ) minutes to complete the experiment.

The bottom row of Table 2 shows what information the omniscient commentator revealed in each situation. For example, in situation 3 in which participants first saw strong indirect evidence, the commentator then said: "In game 1, Player B was lazy." In the additional pattern (IR 2), A wins against B, B wins against C and D wins against E. The commentator then reveals that E was lazy in game 3. For the patterns in which A lost his game, the results of each match as shown in Table 2 were reversed and the corresponding losing player was indicated as having been lazy. For example, in situation 2, A lost all three games against B and the commentator revealed that A was lazy in game 2.

## Results and Discussion

Figure 4 shows the mean strength judgments (gray bars) together with the model predictions (black bars). The dark gray bars indicate participants' first judgments based on the tournament information only. The light gray bars indicate participant's second judgments after they received the commentator's information. The model predicts participants' ratings very accurately again with  $r = .97$  and  $RMSE = 0.29$ . The model's median correlation with individual participants' judgments is  $r = .86$ . Again, strength and weakness judgments for the corresponding patterns were highly correlated,  $r = .98$ .

Generally, participants lowered their estimate of A's strength (top panel) and weakness (bottom panel) after having received the commentator's information. The fact that

participants do not lower their estimates of A's strength for the two cases in which they received *irrelevant evidence* by the commentator about a player's laziness who was in no relationship with A (IR 1, 2), shows that participants did not just have a tendency to regress towards the mean of the scale in their second judgments.

As predicted by the model, the degree to which participants lowered their strength estimates as a result of the laziness information differed between situations. While participants only marginally lowered their estimates for the *confounded evidence* patterns, estimates went down considerably for the *strong indirect evidence* patterns. As mentioned in the discussion of Experiment 1, finding out in the *strong indirect evidence* situation that A's win against B might have only been due to the fact that B was lazy in this match undermines the relevance of the additional evidence about B's performance in match 2 and 3 for A's strength.

The results of Experiment 2 show that participants, as well as our model, have no difficulty in integrating different sources of evidence to form an overall judgment of a player's likely underlying strength. The model predicts participants' judgments very accurately by being sensitive to the degree to which the initial strength estimate should be updated in the light of new evidence provided by the commentator.

## General Discussion

In this paper, we have demonstrated a novel modeling framework that conceptualizes people's reasoning as probabilistic inference over compositionally structured representations. With a handful of concepts that can combine compositionally and support productive extensions over novel situations and objects, we predict participants' judgments in two experiments with thirty different patterns of evidence in total.

The fact that people can reason flexibly based on different patterns and sources of evidence illustrates the importance of modeling our representational capacities on a sufficiently abstract level. People's use of concepts are not tied to particular situations but extend productively over different contexts. The concept of a *winner*, for example, applies to a whole range of possible games or even to domains outside of games entirely such as winning an election. We have provided a concrete working-example of how such a representation could look like, using the probabilistic programming language Church (Goodman et al., 2008). The fact that our model's predictions corresponded very closely to people's judgments can be taken as evidence that the assumptions we had to make when writing the program, generally matched the intuitive assumptions that people brought to the task. A Church program makes the modeling assumptions explicit and thus allows them to be scrutinized. Furthermore, particular modeling assumptions can also be treated as parameters in the model. For example, as outlined above, different participants seemed to have given unequal weight to the probability that a player might be lazy in a game. Without changing the general structure of our representation, we could account

for these individual differences by allowing for flexibility in our modeling assumptions through, for example, treating the chance of laziness as a free parameter.

In our experiments, we have focused on a single query and only used a small number of the possible patterns of evidence. However, our representation supports many more combinations of queries and evidence. For example, we could ask about the probability that a particular player was lazy in a certain game. Or we could ask which of two teams is likely to win given that we have observed the players perform in some previous games or based on some direct information about their strength. Furthermore, it would require only minimal additions to the concept lexicon to handle evidence such as, "all players in the red jerseys were lazy" or "at least one of the players in the green jerseys is very strong."

To conclude, we have provided only a small glimpse into what we see as a broad research program that investigates people's flexible use of everyday concepts using the tools of probabilistic programming – the probabilistic language of thought hypothesis. We are convinced that this research program has the potential to greatly benefit our understanding of how higher-level capacities of human cognition (such as concept learning, naive physics, and theory of mind) are possible.

## Acknowledgments

We thank Andreas Stuhlmüller for insightful comments and for helping with the Church implementation. This work was supported by a doctoral grant from the AXA research fund (TG), a John S. McDonnell Foundation Scholar Award and ONR grant N00014-09-1-0124 (NG).

## References

- Anderson, J. R. (1996). Act: A simple theory of complex cognition. *American Psychologist*, 51(4), 355–365.
- Fodor, J. A. (1975). *The language of thought*. Harvard University Press.
- Goodman, N. D., Mansinghka, V. K., Roy, D., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: A language for generative models. In *Uncertainty in artificial intelligence*.
- Goodman, N. D., & Tenenbaum, J. B. (in prep). *The probabilistic language of thought*.
- Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, 118(1), 110.
- Griffiths, T. L. (2005). Causes, coincidences, and theories. (Unpublished doctoral dissertation)
- Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, 116(1), 20.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.
- McCarthy, J. (1960). Recursive functions of symbolic expressions and their computation by machine, Part I. *Communications of the ACM*, 3(4), 184–195.
- Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, 65(3), 151.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.
- Rumelhart, D. E., & McClelland, J. L. (1988). *Parallel distributed processing*. MIT Press.
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279.

# Speech Act Recognition in Conversation: Experimental Evidence

**Rosa S. Gisladdottir** (rosa.gisladdottir@mpi.nl)

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands  
International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands

**Dorothee J. Chwilla** (d.chwilla@donders.ru.nl)

Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

**Herbert Schriefers** (h.schriefers@donders.ru.nl)

Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

**Stephen C. Levinson** (stephen.levinson@mpi.nl)

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

## Abstract

Recognizing the speech acts in our interlocutors' utterances is a crucial prerequisite for conversation. However, it is not a trivial task given that the form and content of utterances is frequently underspecified for this level of meaning. In the present study we investigate participants' competence in categorizing speech acts in such action-underspecific sentences and explore the time-course of speech act inferencing using a self-paced reading paradigm. The results demonstrate that participants are able to categorize the speech acts with very high accuracy, based on limited context and without any prosodic information. Furthermore, the results show that the exact same sentence is processed differently depending on the speech act it performs, with reading times starting to differ already at the first word. These results indicate that participants are very good at "getting" the speech acts, opening up a new arena for experimental research on action recognition in conversation.

**Keywords:** Action; Speech acts; Implicature; Pre-Offers; Conversation Analysis; Self-paced reading.

## Introduction

Knowing a language doesn't just require syntax or semantics, but the ability to extract speech acts from our interlocutors' utterances. This is crucial in conversation, since all actions – be they non-verbal or verbal – have implications for how we should respond (Schegloff, 2007): A greeting calls for another greeting, an offer is followed by an acceptance or declination. Scholars in Conversation Analysis were the first to reveal the systematicity of courses of action in turn-taking (e.g. Sacks, Schegloff, & Jefferson, 1974; Schegloff, 2007), moving away from the single act as the fundamental unit of analysis – the perspective of Speech Act theory (Austin, 1976; Searle, 1969) – to the role of sequential context. One of the main observations from this literature is that turns tend to come in pairs of actions and in such *adjacency pairs* the first (pair) part sets up powerful constraints on what type of action can follow (Schegloff, 2007). Moreover, not all actions have equal status. Reflecting an orientation towards *preference structure*, dispreferred actions, such as rejections to invitations, tend to

be delivered with inter-turn gaps, turn-initial delay, hedges or other discourse markers (*uhm*, well) (Schegloff, 2007), while preferred actions (e.g. acceptances) do not. This suggests that conversationalists not only monitor speech for actions but also orient to sequential constraints.

Given that action is the *sine qua non* of conversation, how do we map speech acts onto our interlocutors' utterances, bridging the gap between sentence meaning and action? In some cases this is a simple matter. In the utterance *Please close the door*, for instance, the imperative mood and the adverb *please* function as "special markers" or "illocutionary force indicating devices" (Levinson, 1983; Clark, 1979; Schegloff, 2007) that clearly indicate this is a request. In most utterances, however, the absence of such dedicated vocabulary leaves the propositional content underspecified for the speech act level of meaning. As an example, an assertion like *I have a credit card* can deliver different speech acts, depending on context. When responding to a question from our interlocutor (e.g. *How are you going to pay for the ticket?*), *I have a credit card* functions as an information-giving answer. If it follows an offer of payment (e.g. *I can lend you money for the ticket*), it is used to indirectly decline it. In this case it could be characterized as an indirect speech act, in which "one illocutionary act is performed indirectly by way of performing another" (Searle, 1975). In yet another interchange, if our interlocutor has expressed the need or desire for some means of payment (e.g. *I don't have any money to pay for the ticket*), the same statement of ownership can function as a prelude to an offer, called a Pre-Offer in Conversation Analysis (Schegloff, 1988; Schegloff, 2007). In all three cases, the form and semantic content is underspecified for the action such that the full import of the utterance *I have a credit card* can only be ascertained relative to the context, in this case the prior speech act in the conversation.

There is some psychological evidence that people do extract speech act information online. Using a recognition probe task and lexical decision task, Holtgraves (2008a) addressed whether the comprehension of a sentence like

*Don't forget to go to your dentist* (an “implicit speech act”) entails automatic activation of the speech act performed (reminding). He found that the recognition of such speech acts is automatic, both in written and spoken utterances. A further study (Holtgraves, 2008b) suggests that people recognize and retain in long-term memory the actions that people perform with their utterances. In line with Speech Act Theory and Conversation Analysis, Holtgraves (2008a) argues that “in conversation there is an action dimension, a dimension that does not exist for isolated sentences or texts. Speakers are usually constructing utterances with the intention to perform certain actions, and with the intention of having the recipient recognize those actions” (p. 640).

Clearly, action recognition crosscuts research on topics such as communicative intention and implicature in Pragmatics (Grice, 1975; Levinson, 1983; Sperber & Wilson, 2004), the study of indirect speech acts (e.g. Gibbs, 1979; Clark, 1979; Clark, 1996; Coulson & Lovett, 2010) and discourse processing (e.g. Graesser, Singer, & Trabasso, 1994) in Psycholinguistics, as well as a more general discussion of action and theory of mind in the cognitive sciences. There is limited experimental research, however, on speech act recognition in spoken dialogue. The experimental approach used in Holtgraves (2008a, 2008b) involves artificial tasks (lexical decision and recognition probe) and does not unravel the time-course of action recognition. The puzzle remains: how is it that we can extract speech acts from utterances so efficiently, as evidenced by extraordinarily fast turn transitions (Sacks et al., 1974; Stivers et al., 2009; Levinson, in press)? To address this we are currently planning an Event-Related Potential (ERP) study on action recognition in auditorily presented dialogues to track the time course of action comprehension in real time. The experiment presented in this paper was designed to assess the feasibility of such a study.

## The Experiment

The aim of the present experiment was to investigate participants' competence in identifying speech acts in action-underspecific sentences and explore the time-course of speech act inferencing. To do this we presented target sentences using the self-paced reading paradigm and asked participants to categorize the speech acts and rate how sure they were in the categorization. In the domain of Pragmatics, self-paced reading has been used to investigate the processing of phenomena such as scalar implicatures (Breheny, Katsos, & Williams, 2006), with longer reading times (in comparison to a control) interpreted as indicating the generation of an inference. The self-paced reading paradigm allows us to obtain information on the word-by-word processing of action-underspecific utterances, thereby exploring the time-course of speech act inferencing.

The stimuli in our study consist of a context sentence which is presented auditorily, followed by a target sentence designed to be interpreted as an Answer, Pre-Offer or Declination depending on the context (see Table 1). These

actions are commonly found in conversation and their form and function has been described in the conversation analytic literature.

Table 1: Examples of stimuli in Dutch and translations.

Condition	Context	Target Sentence
Answer	Hoe ga je voor het ticket betalen? <i>How are you going to pay for the ticket?</i>	Ik heb een creditcard. <i>I have a credit card.</i>
Declination	Ik kan je wat geld lenen voor het ticket. <i>I can lend you money for the ticket.</i>	Ik heb een creditcard. <i>I have a credit card.</i>
Pre-Offer	Ik heb geen geld om het ticket te betalen. <i>I don't have any money to pay for the ticket.</i>	Ik heb een creditcard. <i>I have a credit card.</i>

The Answers in our study complete an adjacency pair by responding to a *wh*-question in the first turn. This condition serves as a benchmark for inferencing in the reading time analysis since the gap between literal (sentence) meaning and the action intended is the smallest. Moreover, since the other actions in the study can superficially be viewed as Answers, because they respond to the prior turn, this condition provides a check on whether participants go beyond a simple characterization of the sentences and identify the correct speech act.

The Declinations satisfy an adjacency pair by responding to a proposal (an offer or invitation) in the first turn, but require a somewhat complex inference to infer the action. Conversation analysts have noted that, at least in English, such indirect responses “need not be polite, nor unclear or obfuscatory. For certain activities, in specific sequential locations, responding indirectly may be the most efficient form of communication,” (Walker, Drew, & Local, 2011, p. 17).

The third action is the Pre-Offer, which is a type of pre-sequence. Pre-sequences are preliminary to, or project, the main course of action - in this case an offer (Schegloff, 2007; Schegloff, 1988), demonstrated in the following example:

Bookstore, 2.1: 107 (modified from Schegloff, 2007, p. 35)

- 1 A: I'm gonna buy a thermometer though because I  
2 B: but  
3 A: think she's got a temperature.  
4 C: *Pre-Offer* we have a thermometer.  
5 A: *Go-ahead* yih do?  
6 C: *Offer* wanta use it?  
7 A: *Acceptance* yeah.

Only if the response to the Pre-Offer is positive (line 5) is the offer put forward (line 6). This strategy allows conversationalists to check whether an offer would be welcome or not, preventing them from embarrassment that would arise if an offer were to be rejected.

Crucially, the Pre-Offers differ from the Answers and Declinations in that they do not complete an adjacency pair but rather open up or project a continuation of the sequence (with the response and possibly a subsequent offer). By including Pre-Offers in our study we can not only investigate the impact of sequential context on processing, but also explore whether the distinction between projection (Pre-Offers) and a backward directed inference (Declinations) is borne out in reading times.

Given that the same utterance can be used as an Answer, Declination or Pre-offer depending on the sequential context, in this study we investigate: 1) Can participants reliably categorize action-underspecific speech acts? 2) Does the time-course of speech act inferring differ for these actions as reflected in self-paced reading times? Due to the exploratory nature of the study and lack of research in this area – in particular on Pre-Offers – we do not make specific predictions regarding reading times. However, we speculate that the reading time pattern of Pre-Offers and Declinations may differ relative to Answers, based on the structural properties described above.

## Methods

**Participants** 39 native speakers of Dutch were recruited from the student population in Nijmegen, The Netherlands. Participants were paid 8 euros for participating.

**Materials and Design** The stimulus materials were 126 target sentences, presented visually one word at a time (self-paced reading), which were preceded by 378 auditory context-setting utterances that biased the interpretation of the target as an Answer, Pre-offer or Declination. To maintain a balance of variety and control in the stimulus materials, half of the target sentences started with the pronoun “I” (Dutch *ik*) and the verb “have” (*heb*), e.g. “I have a credit card”. The other half was more varied (including simple utterances like “I am going to the market” and “My brother is a mechanic”). We varied word-length to make the stimuli as natural as possible, but constructed the target sentences such that the final word is critical for understanding the propositional content of the utterance (irrespective of speech act level meaning). In line with reported characteristics of indirect replies (Walker et al., 2011), the target sentences do not involve ellipsis or pronominalization.

To maintain consistency in the way the Declinations and Pre-offers are connected to their contexts, we ensured that there is at least one clear *implicated premise* and an *implicated conclusion* for each sentence-pair: when presented with an utterance that is indirect, the hearer needs to access an implicated premise and combine it with the

proposition expressed to derive the implicated conclusion (Blakemore, 1992).<sup>1</sup>

In order to get a measure of the semantic relatedness between context and target sentence in each condition, Latent Semantic Analysis (LSA) values (Landauer, Foltz, & Laham, 1998) were computed for the English translation of each sentence pair using document-to-document mode with “General reading up to 1st year of college” as the semantic space. The average LSA values for each condition were: Answers 0.13, Declinations 0.32, and Pre-Offers 0.42 (the higher the value, the more semantic similarity).

The stimuli were translated from English into Dutch and checked by two native speakers of Dutch. The sentences were recorded by four native speakers. The recordings of the target sentences were not used in this experiment, since they are presented visually in self-paced reading.

The stimuli were pseudo-randomized and balanced across three lists, such that participants saw each target sentence only once, in one context. After each trial (sentence pair), participants were given a comprehension and rating task. They were first asked to indicate what the second speaker was doing with his response and were given the options of Answering, Offering and Declining (D. *antwoorden, aanbieden, weigeren*). Since Pre-Offer is not a colloquial term, the broader term of offering was chosen. Participants were then asked to rate how sure they were in their categorization decision on a rating scale from 1 (very uncertain) to 7 (very certain). The purpose of the rating task was to assess the feasibility of using the items in future studies.

**Procedure** Participants were given instructions that included one example of each action. They were instructed to imagine that they were listening to a conversation between friends or colleagues, and to read the sentences as quickly as possible, but not too quickly as they would have to “judge the underlying meaning” of the sentences. They were then seated in a chair in front of a monitor in a soundproof experimental booth. On each trial the context sentence was played while a small picture of a loudspeaker was presented at the middle of the screen. 500 msec after the end of the spoken sentence participants were presented with the target sentence in a moving window self-paced reading format (Just, Carpenter, & Woolley, 1982). A series of lines appeared on the screen representing each word in the target sentence. When participants clicked on the mouse the first word appeared and upon subsequent button presses a new word was shown, while the previous word was again replaced by a line. When participants clicked the mouse after the last word had been shown, they were presented with the action categorization question, immediately followed by the certainty rating. There were 126 experimental trials, preceded by a brief practice session.

<sup>1</sup> In the dialogue (A): *I can lend you money for the ticket.* – (B): *I have a credit card*, the implicated premise is that a credit card can pay for things, including tickets. The implicated conclusion is that speaker B does not need A’s help with paying for the ticket.

## Results

**Accuracy** Overall accuracy (number of correct responses in the action categorization question divided by the total number of responses) was very high, 95.8 percent. Accuracy percentages (summarized in Table 2) were very similar across conditions. A repeated measures ANOVA revealed that accuracy was not affected by action [ $F(2, 76) = .07, p = .93$ ].

Table 2: Accuracy and mean certainty ratings.

	Accuracy	Certainty Ratings	
		Mean	SD
Answer	96.0%	6.60	0.36
Declination	96.0%	6.50	0.39
Pre-Offer	95.6%	6.35	0.51

**Ratings** Participants rated how certain they were in answering the action categorization question on a scale from 1 (very uncertain) to 7 (very certain). The overall mean certainty rating was 6.48. Mean certainty ratings for each condition are summarized in Table 2. We conducted a repeated-measures ANOVA on the mean ratings and found an effect of action [ $F(1.42, 53.99) = 11.20, p < .01$  (Greenhouse-Geisser)]. Pairwise comparisons using Sidak adjustment for multiple comparisons revealed that Pre-Offers ( $M=6.35, SD=.51$ ) had lower ratings than Answers ( $M=6.60, SD=.36$ ), [ $p < .01$ ], and Pre-Offers were also rated lower than Declinations ( $M=6.50, SD=.39$ ), [ $p < .01$ ]. The comparison between Answers and Declinations was not significant, [ $p = .15$ ].

**Reading Times** The time between button presses was recorded as the reading time for each word. Extreme values below 100 msec were excluded, as well as values above 1200 msec for non-final words and above 7000 msec for final words (in total 12 outliers). Since online speech comprehension and the subsequent off-line categorization task tap different types of information, error trials were not excluded from the reading time analysis.

Mean reading times for the first word, the verb and the final word of the target sentences were used for the analysis, in addition to the mean reading time of the entire sentence and mean reading time per word (sentence reading time divided by number of words). A repeated-measures ANOVA was carried out to examine the effect of action (Answer, Pre-Offer, Declination) on reading times and post-hoc comparisons were performed using the Sidak adjustment. The conditions started to differ already at the first word, with mean reading time being affected by the action manipulation [ $F(1.69, 64.26) = 5.24, p = .01$  (Greenhouse-Geisser)]. Post-hoc comparisons revealed that first word reading times in Pre-Offers ( $M=259, SD=65$ ) were longer than in Answers ( $M=251, SD=56$ ), [ $p = .03$ ], and Pre-Offers tended to be longer than Declinations ( $M=252, SD=57$ ), [ $p = .06$ ]. The comparison between Answers and Declinations was not significant [ $p = .98$ ].

Table 3: Mean reading times in msec.

		Answer	Declination	Pre-Offer
First Word	Mean	251	252	259
	RT			
Verb	Mean	260	267	265
	RT			
Final Word	Mean	564	622	593
	RT			
Word <sup>2</sup>	Mean	339	354	352
	RT			
Entire Sentence	Mean	1459	1528	1501
	RT			
	SD	584	652	603

At the verb, there was also an effect of action on mean reading time [ $F(2, 76) = 3.41, p = .04$ ]. Pairwise comparisons revealed that verb RTs in Declinations ( $M=267, SD=69$ ) were longer than in Answers ( $M=260, SD=60$ ), [ $p = .048$ ]. The comparison between Pre-Offers and Answers [ $p = .25$ ] was not significant, nor between Pre-Offers and Declinations [ $p = .80$ ].

At the final word, mean reading times also differed between actions, [ $F(1.63, 61.83) = 4.28, p = .03$  (Greenhouse-Geisser)]. Pairwise comparisons on the final RTs revealed that Declinations ( $M=622, SD=447$ ) were marginally significantly longer than Answers ( $M=564, SD=384$ ) [ $p = .052$ ], while there were no differences between Answers and Pre-Offers [ $p = .17$ ], nor Pre-Offers and Declinations [ $p = .41$ ].

An ANOVA on mean RTs per word revealed an effect of action [ $F(2, 76) = 5.73, p = .01$ ]. Declinations ( $M=354, SD=146$ ) took longer than Answers ( $M=339, SD=135$ ) [ $p = .02$ ], and Pre-Offers ( $M=352, SD=147$ ) took longer than Answers ( $M=339, SD=135$ ), [ $p = .04$ ]. There was no difference between Pre-Offers and Declinations [ $p = .95$ ].

Finally, although the RTs for the entire sentence differed descriptively, these differences were not reliable [ $F(2, 76) = 2.73, p = .09$ ].

## Discussion

The present experiment demonstrates that participants categorize the speech acts of sentences whose form and semantic content is underspecified for action with very high accuracy (95.8%). They are able to do so based on limited context (the prior speech act) and without any prosodic information in the target sentence. Importantly, the accuracy was the same for all three actions (Answer, Declination, Pre-Offer). If participants had processed the target sentences superficially, ignoring the speech act content, they could have categorized Declinations and Pre-Offers as Answers.

<sup>2</sup> RT of the entire sentence divided by number of words.



This is the case since the Dutch term for answering (*antwoorden*) also means *to respond* and all three speech acts can superficially be seen as responses. This should have resulted in lower accuracy for Pre-Offers and Declinations vis-à-vis Answers. The high accuracy rate across actions shows that participants go beyond a simple characterization of the target sentences as responses and “get” the correct action. Participants were also very confident in categorizing all actions and rated the certainty of their categorizations on average 6.48 (out of 7). These results provide further support that participants orient to the action content of sentences.<sup>3</sup>

The reading time results demonstrate that the exact same sentence is processed differently depending on the speech act it performs. In all conditions the reading time increased throughout the sentence, but Declinations and Pre-Offers had different trajectories relative to Answers, which had shortest RTs on all measures. Reading times differed already at the first word, with first word RTs in Pre-Offers being longer than in both Answers and Declinations. It should be pointed out that descriptively the difference between the means is small and standard deviation large. Event-related brain potentials might be a more sensitive measure to reveal processing differences at early positions in the sentence. At the verb and the final word however, RTs were longest in the Declination condition.

What could explain the different reading times across conditions? One possible source of difference is the *amount* of inferencing required. In text processing, reading times are predicted to be the longest for words in the text that generate “many online inferences” (Graesser, Swamer, & Baggett, 1996). The fact that Declinations have the longest RTs at the final word may be because *more* inferences (e.g. Gricean implicatures) are needed to relate the sentence to the prior context.

Another source for differences may be the *kind* of inferencing required. Given the exploratory nature of this study we did not make predictions regarding reading time results. However, based on differences in sequence organization, we speculated that Declinations and Pre-Offers would exhibit different reading time patterns relative to Answers. While recognizing the speech act in a Declination requires computation of how the utterance can be understood as the second pair part to the prior proposal, identifying a Pre-Offer involves knowing that an offer will follow in the sequence. Because Declinations close an adjacency pair, they do not heavily constrain the relevant next action. Pre-Offers, on the other hand, call for either go-

ahead (e.g. *you do?*) and a subsequent offer, or a blocking response (*that’s ok*). Pre-Offers, therefore, invite stronger predictive inferences about the next speech act. The distinction between an inference based on a backward bridge to the prior turn in an adjacency pair and an inference based on forward projection of a sequence is akin to the difference between causal antecedent and causal consequence inferences in text processing (e.g. Magliano, Baggett, Johnson, & Graesser, 1993). The reading time differences between Pre-Offers and Declinations provide some indication that the distinction between forward projection and a backward directed inference plays a role in the online processing of speech acts. Why the projective nature of Pre-Offers would call for more processing at the first word, however, is less clear. More research is needed to investigate whether this finding holds for spoken language processing as well.

Considering that action comprehension is crucial for everyday conversation, it seems likely that people can predict upcoming speech acts, making the fast transitions between turns (Sacks et al., 1974; Stivers et al., 2009; Levinson, in press) possible. Neuroimaging studies suggest that people use their knowledge of the wider discourse context to predict specific upcoming words and that prediction is not the result of relatively low-level, word-based priming mechanisms, “but involves a more sophisticated message-level mechanism that can take into account the actual nuances of the preceding discourse,” (Otten & Van Berkum, 2008). Whether and how sequential context and the implicit knowledge of the organization of actions guides the interpretation of utterances is a topic for further investigation. We will explore this in future research using event-related brain potentials.

An alternative explanation for the reading time results is that the experimental manipulation does not address speech act recognition per se, but some other confounding variable such as semantic priming from the context. Latent Semantic Analysis can be used to determine semantic relatedness of two texts and LSA similarity relations have been found to correspond well with the pattern of results in priming studies (Landauer et al., 1998). If semantic priming from the context is the main factor governing the reading times one would expect the condition with the lowest LSA values (least amount of priming) to have the longest mean reading times. The opposite is true: Answers had the lowest average LSA value but the shortest reading times on all measures. This suggests that the differences in reading times across conditions in our study were not due to lexico-semantic relationships between the content words of the context and the target sentences.

## Conclusion

In this study on speech act comprehension we investigated the processing of sentences that perform different speech acts depending on prior context. In each case an assertion is used as a vehicle for some other action, and it is “part of competent membership in the society/culture and being a

<sup>3</sup> Note that the certainty ratings were lower for Pre-offers (6.35) than Answers (6.60) and Declinations (6.50). Since “pre-offer” is not a colloquial term, participants were instructed to categorize Pre-offers as “offering”. The ratings may reflect participants’ awareness that the speech acts they were instructed to label as “offering” are preparatory to, or project, offers and in most cases do not constitute offers in themselves. However, for the purpose of this study it was sufficient that participants could identify speech acts as belonging to one of the defined categories.

competent interactant to analyze assertions of this sort for what (else) they may be doing at this moment, at this juncture of the interaction, in this specific sequential context” (Schegloff, 2007, p. 35). Our study tapped into this competence by addressing two primary questions: how reliably participants can categorize action-underspecific speech acts, and whether the time-course of speech act inferencing differs for the actions as reflected in self-paced reading times.

Participants in our study categorized the speech acts of action-underspecified utterances with very high accuracy, based on limited context and without any prosodic information. Furthermore, the exact same sentence was processed differently depending on the speech act it performed, with reading times starting to differ already at the first word. These findings open up a new arena for experimental research on speech act recognition in conversation.

As a crucial component of social behavior, communication involves actions. Being a competent member of society must require a cognitive architecture that is oriented to speech acts. However, given that the form and content of utterances is frequently underspecified for this level of meaning, assigning speech acts to our interlocutors is not a trivial matter. Having demonstrated that participants orient to the action content of sentences and can categorize speech acts with high accuracy, the next experimental step is to shed light on this ability in spoken dialogues – the foundation of doing things with words.

## References

- Austin, J. L. (1976). *How to do things with words*. Oxford: Oxford University Press.
- Blakemore, D. (1992). *Understanding utterances: An introduction to pragmatics*. Oxford: Blackwell.
- Breheny, R., Katsos, N., & Williams, J. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100(3), 434–463.
- Clark, H. H. (1979). Responding to indirect speech acts. *Cognitive Psychology*, 11(4), 430–477.
- Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.
- Coulson, S., & Lovett, C. (2010). Comprehension of nonconventional indirect requests: An event-related brain potential study. *Italian Journal of Linguistics*, 22(1), 107–124.
- Gibbs, R. W. (1979). Contextual effects in understanding indirect requests. *Discourse Processes*, 2(1), 1–10.
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3), 371–395.
- Graesser, A. C., Swamer, S. S., & Baggett, W. B. (1996). New models of deep comprehension. In B. Britton & A. C. Graesser (Eds.), *Models of understanding text*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics, 3: Speech Acts*. New York: Academic Press.
- Holtgraves, T. (2008a). Automatic intention recognition in conversation processing. *Journal of Memory and Language*, 58(3), 627–645.
- Holtgraves, T. (2008b). Conversation, speech acts, and memory. *Memory & Cognition*, 36(2), 361–374.
- Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111(2), 228–238.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3), 259–284.
- Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
- Levinson, S. C. (in press). Action formation and ascription. In T. Stivers & J. Sidnell (Eds.), *Handbook of conversation analysis*. Malden, MA: Wiley-Blackwell.
- Magliano, J. P., Baggett, W. B., Johnson, B. K., & Graesser, A. C. (1993). The time course of generating causal antecedent and causal consequence inferences. *Discourse Processes*, 16(1-2), 35–53.
- Otten, M., & Van Berkum, J. J. A. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes*, 45(6), 464–496.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735.
- Schegloff, E. A. (1988). Presequences and indirection: Applying speech act theory to ordinary conversation. *Journal of Pragmatics*, 12(1), 55–62.
- Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis*. Cambridge: Cambridge University Press.
- Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.
- Searle, J. R. (1975). Indirect speech acts. In P. Cole & J. Morgan (Eds.), *Syntax and semantics, 3: Speech Acts*. New York: Academic Press.
- Sperber, D., & Wilson, D. (2004). Relevance theory. In G. Ward & L. Horn (Eds.), *Handbook of pragmatics*. Oxford: Blackwell.
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26), 10587–10592.
- Walker, T., Drew, P., & Local, J. (2011). Responding indirectly. *Journal of Pragmatics*, 43(9), 2434–2451.

## Towards historical cognitive science: the case of Ancient Greece

V.V. Glebkin ([gleb1514@gmail.com](mailto:gleb1514@gmail.com))  
Gymnasium 1514, 12 Krupskoi Street  
Moscow, 119311 Russia

### Abstract

This study rests on the two basic ideas: that there has been a visible development of cognitive skills from the Antiquity to nowadays, and that the text analysis is the only way to bring it out. The author addresses the three eminent works: Euclid's *Elements* and the historical treatises by Herodotus and Thucydides to reveal the notable peculiarities of the Ancient Greeks' cognitive style in comparison with the current cognitive models.

### Introduction

Cognitive scientists traditionally look at language as a cognitive system. However, it is not the only acceptable view on language within the scope of cognitive science. The distinguished German historian Reinhart Koselleck emphasized «a methodologically irresolvable dilemma: that every history, while in process and as occurrence, is something other than what its linguistic articulation can establish; but that this "other" in turn can only be made visible through the medium of language» (Koselleck 2004, 223; cf. Pocock 2009, 106–119). Admittedly, this position may just as well be applied to cognition: the way people think is different from verbal representation, but we cannot comprehend how the people of the 19<sup>th</sup> century and earlier used to think and perceive the world without a scrupulous analysis of the texts they created. Although such a view on text as a medium for revealing cognitive models is not widespread amongst cognitive scientists, there are a number of branches within cognitive science that have emerged recently (cognitive stylistics, cognitive poetics etc.), in which scholars apply a cognitive analysis to particular texts, mostly fiction (Attardo 2002; Culpeper 2002; Semino 2002; Tsur 2002; Emmott et al. 2007; Semino 2007). However, we can hardly encounter any works that provide us with a precise analysis of different types of texts, created in the same historical epoch in order to explicate general for these texts cognitive models. To develop an elaborate methodology for such analysis is not a matter of cognitive stylistics or cognitive poetics; it is merely a matter of a special branch of cognitive science which can be called historical cognitive science.

This paper focuses on two case studies, but its bottom line is to provide some conceptual pillars for investigations in this field, in other words, to discover some correlations between the narrative models of the text construction and the cognitive models used by its author. The concept of *cognitive style* will be the main methodological tool for that. This concept is widely applied in different types of researches (Rubin 1970; Berzonsky & Ondrako 1974; Witkin et al. 1977; Logan 1983; Roberge & Flexer 1983; Fuchs 1991; McIntyre & Meloche 1995; Judice 1997; Riding &

Rayner 1998; Riding & Rayner 2000; Tomes 2004; cf. concept of *mind style* in Semino 2002; Semino 2007); although it cannot be called fairly clear-cut (see some criticism in Tomes 2004, 47–48), its gist is quite transparent: a) *cognitive styles* characterize the form rather than the content of cognitive activity; b) they are pervasive dimensions which cut across the disciplinary boundaries; c) they are stable over time; d) they are bipolar, that is, they can be sorted out into opposite pairs (field-dependency – independency; holist – serialist thinking; adaptors – innovators etc.) (Witkin et al. 1977, 15–16; Riding & Rayner 1998, p. 20). The two peculiarities of my applying this term here should be featured as follows: firstly, in this context I mean a cognitive style of a particular culture, but not a particular person; in other words, I address the mode of thinking common to a notable number of culture bearers, involved in different types of intellectual activity; secondly, I seek for some criteria to compare these modes and establish the foundations for cultural-historical typology of cognitive styles.

The only parameter we will focus on is *field-dependency – independency*. Following H. Witkin, the field-independent cognitive style characterizes the tendency to differentiate objects from their surroundings whereas the field-dependent one stresses the strict connection between surroundings and objects. It concerns the subjects themselves as well; people of the field-dependent style are more likely to follow external instructions while field-independent style people prefer to rest on the internal basis for their actions (Witkin et al. 1977, 2–14).

The *field-dependency – independency* opposition, with some necessary corrections, seems to be a fruitful pathway to fit the process of cognitive evolution in phylogenesis. Furthermore, it correlates quite well with some classical researches in this field. I mean the investigations of Vygotsky's school of cultural-historical psychology or, more precisely, the distinction between complex and conceptual thinking (Vygotsky 1986(1934), 96–145) and the idea of field binding (Samukhin et al. 1934; Vygotsky 1984 (1933/34)). Briefly, the essence of this approach can be formulated as follows: unlike concepts which are characterized by a rigid structure and a set of objective features, complexes have flexible and contextually dependent frames. In the case of complex thinking, subjects' cognitive decisions are influenced by their unique experience, and they cannot be described by any general abstract model such as Aristotelian logic. Vygotsky created this model mainly to explain pre-school and primary school child cognitive development, and only outlined its application to phylogenesis. However, his followers applied this approach to different types of cultures and got the im-

portant results. Thus, Alexander Luria (1976), researching cognitive scenarios basic for Central Asia dekchans, pointed out that those scenarios were triggered by psychological fields of their everyday activity. Particularly, the subjects could not see the abstract principles used to classify a given set of objects and failed to identify the odd one out; they found all the objects useful for everyday life. As well, they could not solve syllogisms, conceiving their elements as independent propositions. Furthermore, their answers were based on their everyday experience, and they insisted that they could speak only about the things they had seen before. Further investigations (e.g., Mikheev 1985; Tulviste 1991) confirmed that complex thinking and rigid links with the psychological field of everyday experience can be called the bottom line of the traditional cultures' cognitive style.

At first sight our mind operates in a radically different way. Nevertheless, the investigations of R. Frumkina and her colleagues (Frumkina & Mirkin 1986, Frumkina et al. 1991, Frumkina & Mikheev 1996, Frumkina 2007) found out that complex thinking characterizes cognitive decisions of educated persons in modern culture in a great number of everyday situations. The only difference from the traditional culture is that they can explain their decisions and accommodate them to the experimenter's requests. As a generalization of these results, we can suppose that our cognitive structure has several levels, where complex thinking occupies the lowest, strongly field-dependent level, while different types of theoretical thinking are on the upper ones. In a concrete situation, we, guided by circumstances, resort to the relevant «floor» of our cognitive construction.

Given these standpoints as the background for the further discussion, we have a reason to ask in which cultures these floors emerge. It might seem that they emerge alongside the emergence of a written language and complex forms of social-economic activity in such large-scale civilizations as Ancient Babylon or Ancient Egypt. However, this point is the subject of serious objections. For instance, in Glebkin 2011 I address the Code of Hammurabi, i.e. the Babylonian law code, dated back to around the 18<sup>th</sup> century B.C., and argue that, in dissonance with our expectations, the code structure and its layout are accounted for by a complex thinking model. Consequently, the point that the theoretical mode of thinking is a feature of the Ancient Babylon culture cannot be taken for granted and needs convincing evidence based on a scrupulous analysis of concrete texts.

What cannot be cast in doubt is the fact that theoretical thinking is an important element of the Ancient Greek culture. However, it is the beginning but not the end of the investigation. The question is whether the cognitive style dominant in this culture is similar to the modern one, or it has some notable peculiarities. And, if the latter hypothesis is correct, can we track the trajectory of cognitive evolution within the theoretical mode of thinking? In order to answer this question, I would like to consider the three eminent works: Euclid's *Elements* and the historical treatises by Herodotus and Thucydides.

## The cognitive style of Euclid's *Elements*

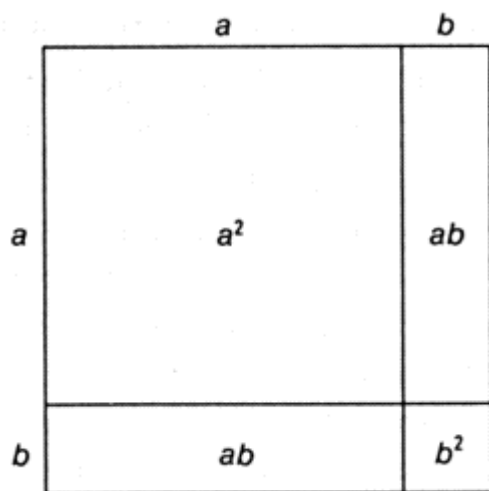
It is not a novel insight that modern mathematics (at least, mathematics at school) rests on Euclid's *Elements*. If so, we might expect to see in this work the familiar to us conceptual ideas and basic attitudes. According to common sense, mathematics is not grounded on any socio-cultural environment. Nevertheless, there are persuasive arguments for the opposite view. Thus, the eminent German historian and philosopher Oswald Spengler in his book *The Decline of the West (Der Untergang des Abendlandes)* claimed that every culture has its own mathematics, and the difference between the Ancient Greek mathematical style and that of Modernity is crucial (Spengler 1991 (1918), 41-69). So, let us have a more precise look at the text of the *Elements*.

We start up with the definitions of the first book. Here we encounter some surprises. For example, the definition of a triangle goes like this: *Of trilateral figures, an equilateral triangle is that which has its three sides equal, an isosceles triangle that which has two of its sides alone equal, and a scalene triangle that which has its three sides unequal* (Heath 1956, 1, 2). In modern understanding an equilateral triangle is a particular case of a triangle, and it sounds strange to mark out a scalene triangle as a special type of triangles. However, for the Ancient Greeks the more perfect cannot be a particular case of the less perfect, and it is an equilateral triangle which is a triangle par excellence. The next illustration of this principle is the difference between the concepts of number and magnitude. The definitions related to magnitude are placed in the fifth book of the *Elements*, whereas those related to number are located in the seventh one (Heath 1956, 2, 113-114, 277-278). Some of them are identical, and we can render number as a particular case of magnitude. However, for the Ancient Greeks it is not the case. Number is more perfect than magnitude; unlike the latter, the former has its own visual image, its own εἶδος. The idea of the perfect form, based on a visual perception, is extremely important for the Ancient Greek philosophy and culture. The more perfect the entity is the more perfect form it has. Thus, Parmenides' Being (τὸ ὄν) and Plato's Universe (κόσμος) have the most perfect form, that is the form of a sphere (Parmen. Fr. 7; Pl. Tym. 33b-34a).

The visual ground for cognition can be also illustrated by the "geometric algebra" of the second book. Here, elementary algebraic formulas, such as  $(a+b)^2 = a^2 + 2ab + b^2$ , are proved by employing the language of geometry (see fig. 1). This proof looks rather cumbersome (it occupies two pages, whereas an algebraic proof fits into one line), but here we encounter the fundamental limitations of the Ancient Greek mathematics. Such a "geometrization" of mathematics, its dependence on the visual field determined its frontiers<sup>1</sup>; solution of some third and forth equations was

<sup>1</sup> The "visual vector" of the Ancient Greek mathematics can also be revealed in terminology. Thus, according to Liddell-Scott's Greek-English Lexicon the basic meaning of

the maximum to reach in that scope. The only way to take a further step in this field was to develop the abstract notation system of algebra which meant breaking the links between numbers and their visual ground. Such breaking demanded radical cultural transformations provided by medieval culture.



**Fig. 1. A geometrical proof of the formula**  
 $(a+b)^2 = a^2 + 2ab + b^2$

It may therefore be interesting to sketch the bottom line of this process. From the Ancient Greek perspective, both the Universe and particular natural things were self-sufficient entities, and such self-sufficiency was perceived as perfection (see, e.g., Arist. Phys II 192 b8-30). It means that they contained within themselves a principle of their motion and transformation. In the medieval Christian culture, however, such a principle turns out to be situated outside the Universe. The Universe and particular things become there signs of the transcendental reality, the means to understand the scheme of God. The functional paradigm stands for the eidetic one. What it means for mathematics can be clearly seen if we compare views on number by Plotinus, whose “Enneads” is considered to be the outcome of the Ancient Greek philosophical attitudes, and Augustine, the key person of the early medieval philosophy. For Plotinus numbers are placed between *ἓν* (the One) and *νοῦς* (the Divine Mind), having the higher rank than the other *εἰδωτά* (ideas) (Enn. 6, 6, 8-14), whereas for Augustine numbers are transformed into tools in God’s hands, loosing in this their unique forms. Thus, he introduces numbers, perceived by sense (*numeri sensibilis*), numbers, moving over time (*temporales numeri*) etc. (Epistola III, 2; De musica, VI, 57). The diversity of types and forms of number entails the release from visual-field-dependence, which in turn

gives new opportunities for mathematics, particularly, for the theory of functions.

So, summing up this part of the paper, I would like to highlight the visual-field-dependence as the important feature of the mathematics cognitive style. Now, we move on to the Ancient Greek historical treatises.

## The cognitive style of Herodotus and Thucydides

Let me start with the *History* by Herodotus. This historian is called the “Father of History” because his treatise is the first example of the elaborate systematic analysis of a huge amount of historical data. To some extent his status in science is similar to Euclid’s one. Here, we focus on the first book of the *History* and start up with the methodology. In order to give a more precise analysis of the narrative structure of this book, I first marked out three levels of narration: *external* (the basic level where people are considered social role-holders and their behavior is influenced by their surroundings), *internal* (the level corresponding to feelings, thoughts and decisions of humans as free will persons), *transcendent* (the level characterizing gods’ actions, fate, predestination and other factors which are believed transcendent to the human world). Then I divided the text into some structural elements, namely: *events*; *causal remarks*, situated both within an event and between the events (they explain why the event develops or the subsequent events are connected in this particular way), *expositions*, introducing unknown for the readers, but important for the further narration information about the venue of an event, people engaged in it, etc.; *philosophical, existential et al. reflections and explanations*. The events in turn were sorted out into three groups: “time markers”, i.e. some bare mentions to fill the time gap (e.g., “Not long after the overthrow of the tyrants by the Lacedaemonians, the battle of Marathon was fought between the Athenians and the Persians” (Thuc. Hist., 1, 18; tr. by B. Jowett)) (E0); the single events described briefly (E1); the events described in detail (E2) (although it was not the absolute criterion, the detailed description commonly had more than 2000 characters). Additionally, I recorded whether the event is single or iterative.

Since a comprehensive analysis of all the aspects of the narrative structure would doubtlessly lead to another long article I will restrict myself to the analysis of the 1<sup>st</sup> chapter, just focusing on some observations that should be definitely included in such a debate.

Firstly, 19% of the 1<sup>st</sup> chapter is taken by the events described in detail (E2)<sup>2</sup>. Given that 21% of the chapter is devoted to ethnographic descriptions within expositions, we can stress a slow speed of narration; the historian’s view here is a sensitive to details view of a spectator, but not a bird’s-eye view of a long-term researcher.

qewrśw is “to look at, view, behold”, e.g., “to view the public games”; qewr...a basically means “sending the state-ambassadors to the oracles or games or being a spectator at the theatre or games”, qeərhma – “sight, spectacle, object of contemplation”. Thanks to Aristotle these concepts were shifted from the material world to the ideal one to characterize the process, product and object of intellectual contemplation.

<sup>2</sup> I counted the figure of characters in the Greek text.

The additional evidence for such visual-field-dependence is provided by the fact that 67% (55 from 82) of the events, described briefly (E1), turn out to be spectacular descriptions resting on a visual perception, or, put another way, a kind of performance the audience visualize at the theatre. Let me illustrate it with an episode of the tyrannus Pisistratus returning to Athens: “*Presently his enemies who together had driven him out began to feud once more. Then Megacles, harassed by factional strife, sent a message to Pisistratus offering him his daughter to marry and the sovereign power besides. When this offer was accepted by Pisistratus, who agreed on these terms with Megacles, they devised a plan to bring Pisistratus back which, to my mind, was so exceptionally foolish that it is strange (since from old times the Hellenic stock has always been distinguished from foreign by its greater cleverness and its freedom from silly foolishness) that these men should devise such a plan to deceive Athenians, said to be the subtlest of the Greeks. There was in the Paeanian deme a woman called Phya, three fingers short of six feet, four inches in height, and otherwise, too, well-formed. This woman they equipped in full armor and put in a chariot, giving her all the paraphernalia to make the most impressive spectacle, and so drove into the city; heralds ran before them, and when they came into town proclaimed as they were instructed: “Athenians, give a hearty welcome to Pisistratus, whom Athena herself honors above all men and is bringing back to her own acropolis.” So the heralds went about proclaiming this; and immediately the report spread in the demes that Athena was bringing Pisistratus back, and the townsfolk, believing that the woman was the goddess herself, worshipped this human creature and welcomed Pisistratus*” (Her. Hist., 1, 60; tr. by A. Godley). We can see that the pivot component of this episode is the visual image of Phya-Athena, and its structure in general addresses us to Aristophanes’ or Menander’s comedy.

Secondly, the philosophical reflections are expressed here not through the author’s words, but for the most part by the extended remarks of the characters in the dialogues. For instance, the idea of happiness, extremely important for Herodotus and the Ancient Greek culture in general, is put into the mouth of the eminent Athenian legislator Solon in his talk with Croesus, king of Lydia (Her. Hist., 1, 30-33). The behavior of the characters and the context of the talk are fairly close to Homer’s epos or the ancient tragedy, where the spectator is expected to watch it.

Thirdly, in order to reveal the reasons for historical events, Herodotus refers to both, transcendent powers (fate, gods’ envy) and human intentions dependent on their character, social rank, view on the situation, etc. Most frequently his interpretation is guided by the cumulative principle, in other words, he gives a number of versions without reconciling them. Importantly, however, transcendent factors proved to be involved in human life as initial reference points, and from the matter of fact human quick-wittedness or stupidity appear the main reason for the historical development. A good illustration for

that is Herodotus’ view on oracles and signs. Given the truth of the oracles as a point beyond doubt, he points at the capacity to render oracles and signs as a deciding factor to the successful action and puts the reason for human failures in people themselves rather than in fate or destiny (Hist. 1, 65; 1, 67-68; 1, 71; 1, 91 etc.).

So, in sum, we can conclude, that for Herodotus the cloth of history is woven by particular people who implement their intentions and projects and take into account various circumstances, from weather to oracles and signs, to do their best in that.

Let us look now at the Thucydides’ treatise. At first sight, his narrative manner has nothing in common with the Herodotus’ one. The notable part of events in the 1<sup>st</sup> chapter of his *History* is described with time markers, and the descriptions, resting on a visual perception, occupy just 13% (8 from 62) of the briefly described events. However, in comparison with Herodotus, the events described in detail occupy here much more space (41.5%). Part of them (12.3%) look quite “cinematic” stories (e.g., sea battle between Corinth and Kerkyra (1, 48-53), or constructing the walls around Athens (1, 89-93)), but the key place here (29.2%) is occupied by talks and dialogues, invented by Thucydides. In these dialogues the characters state their views on the situation trying to convince the audience to follow their suggestions. Taking into account their length and position within the text, we can call them the core elements of Thucydides’ treatise. The analysis of these talks leads us to the two main conclusions. Firstly, their composition resembles Euripides’ tragedies. Similar to Herodotus, these talks address a listener, but not a reader. Secondly, even much more intensively than Herodotus, Thucydides insists that human intentions and reasons are the main factor of the historical development. Transcendent level happens to be omitted in his text.

The situation changes radically if we resort to the medieval historiography. Let me illustrate these transformations with *The History of the Franks* by Gregory of Tours. Indeed, we can find here a number of descriptions resting on a visual perception. However, all of them appear signs of transcendent reality, the testimony of its presence in the material world. Here is the illustration: “*At that time Quirinus, bishop of the church of Sissek, endured glorious martyrdom in Christ's name. The cruel pagans cast him into a river with a millstone tied to his neck, and when he had fallen into the waters he was long supported on the surface by a divine miracle, and the waters did not suck him down since the weight of crime did not press upon him. And a multitude of people standing around wondered at the thing, and despising the rage of the heathen they hastened to free the bishop. He saw this and did not permit himself to be deprived of martyrdom, and raising his eyes to heaven he said: "Jesus lord, who sittest in glory at the right hand of the Father, suffer me not to be taken from this course, but receive my soul and deign to unite me with thy martyrs in eternal peace."* With these words he gave up

*the ghost, and his body was taken up by the Christians and reverently buried*" (1, 35; tr. by E. Brehaut).

Another important feature of this text is the lack of direct causal links between events. Similar to Augustine's numbers historical events turn out for Gregory of Tours the tools in God's hands, which leads us to breaking of the visual-field-dependence and gives new opportunities for historiography.

### Conclusion

Now, it is time to return to the general issue raised in the introduction. It is not in doubt that the Ancient Greek culture is theoretical, where we can find most cognitive operations that we perform. However, we can also encounter some special features like visual grounding of cognitive operations. All in all, the question is whether it is correct to speak here about the cognitive development from antiquity to nowadays, or to compare different cognitive styles for the sake of revealing cognitive evolution means to put the shoe on the wrong foot. There is some evidence to support the former hypothesis. Thus, M. de Vega (2008) argues for the existence of two levels of embodiment: a first-order embodiment is "strongly grounded on current perception and action", whereas a second-order embodiment "is much more detached from current perception and action" (ibid., 300). Similarly, we can single out at least two levels of embodiment for mathematics: the one for Euclid's geometry and the other for, say, a functional analysis. So, the general point is that, following the more complex challenges of modern life, the cognitive structure of modern people has got more "floors", and their cognitive styles have much more variations than they used to have in Ancient Greece. The opposition "field-dependency – independency" seems quite productive to describe this development.

### Acknowledgements

I appreciate Mrs. Tatiana Malitskaya for help in preparing this paper.

### References

Attardo, S. (2002). Cognitive stylistics in humorous texts. In E. Semino, J. Culpeper (eds.) *Cognitive Stylistics. Language and cognition in text analysis*. Amsterdam; Philadelphia: J. Benjamins Pub. Co., 231-250.

Berzonsky, M. & Ondrako, M. (1974). Cognitive Style and Logical Deductive Reasoning. *The Journal of Experimental Education*, 43, 1, 18-24.

Culpeper, J. (2002). A cognitive stylistic approach to characterization. In E. Semino, J. Culpeper (eds.) *Cognitive Stylistics. Language and cognition in text analysis*. Amsterdam; Philadelphia: J. Benjamins Pub. Co., 251-277.

de Vega, M. (2008). Levels of embodied meaning: From pointing to counterfactuals. In de Vega M., Glenberg A., Graesser A. (eds.) *Symbols and embodiment: debates on meaning and cognition*. Oxford; New York: Oxford University Press, 285-308.

Emmott, C., Sanford, A. & Dawydiak, E. (2007). Stylistics meets Cognitive Science: Studying Style in Fiction and Readers' Attention from an Interdisciplinary Perspective. *Style*, 41, 2, 204-224.

Frumkina, R. (2007). Social'noe poznanie v kontekste lingvistiki i psihologii. *Obschestvennye nauki i sovremennost'* 1, 145-156.

Frumkina R. & Mirkin B. (1986). Semantika "konkretnoi" leksiki: psiholingvisticheskii podhod. *Izvestiya Akademii Nauk SSSR. Seriya Literatury i yazyka*. 45, 1, 12-22.

Frumkina, R., Mikheev, A., Mostovaya, A. & Ryumina, N. (1991). *Semantika i kategorizatsiya*. M.: Nauka.

Frumkina, R. & Mikheev A. (1996). *Meaning and categorization*. New York: Nova Science.

Fuchs, S. (1991) Metatheory as Cognitive Style. *Sociological Perspectives*, 34, 3, 287-301.

Glebkin, V. (2009). Chislo u Plotina i Avgustina. In Krichevets A. (ed.) *Chislo*. M.: MAKSS Press, 264-272.

Glebkin, V. (2011). Hermeneutics and cognitive science: a preliminary approach. In B. Kokinov, A. Karmiloff-Smith, N. J. Nersessian (eds.). *European Perspectives on Cognitive Science. Proceedings of the European Conference on Cognitive Science EuroCosSci2011*. Sophia: New Bulgarian University Press, 1-4.

Heath, Th. (1956). *The thirteen books of Euclid's Elements*. New York: Dover Publications.

Judice, N. (1997). *Cognitive style: A three-dimensional model*. Ph.D. dissertation. Arizona State University.

Koselleck, R. (2004). *Futures past: on the semantics of historical time*. New York: Columbia University Press.

Logan, J. (1983). Cognitive Style and Reading. *The Reading Teacher*, 36, 7, 704-707.

Luria, A. (1976). *Cognitive Development: Its Cultural and Social Foundations*. Cambridge, Mass.: Harvard University Press.

McIntyre, R. & Meloche, M. (1995). Cognitive Style and Customer Orientation. *Journal of Business and Psychology*, 10, 1, 75-86.

Mikheev, A. (1985). Svobodnaya klassifikatsiya nabora predmetov (eksperiment v Nagornom Karabahe). In Frumkina, R. (ed.). *Lingvisticheskie i psiholingvisticheskie struktury rechi*. M.: Institut yazykoznaniya, 78-93.

Pocock, J. (2009). *Political thought and history: essays on theory and method*. Cambridge, UK; New York: Cambridge University Press.

Riding, R. & Rayner, S. (1998). *Cognitive styles and learning strategies: understanding style differences in learning and behaviour*. L.: D. Fulton Publishers.

Riding, R. & Rayner, S. (2000). *Cognitive styles*. Stamford, Conn.: Ablex Publishing Corp.

Roberge, J. & Flexer, B. (1983). Cognitive Style, Operativity, and Mathematics Achievement. *Journal for Research in Mathematics Education*, 14, 5, 344-353.

Rubin, R. (1970). Cognitive Style in Pregnancy. *The American Journal of Nursing*, 70, 3, 502-508.



- Samuhin, N., Birenbaum, G. & Vygotskii, L. (1934). K voprosu o strukture demencii pri bolezni Pi-ka. *Sovetskaya nevrologiya, psikiatriya i psikhigiya*, 3,6, 97-136.
- Semino, E. (2002). A cognitive stylistic approach to mind style in narrative fiction. In E. Semino, J. Culpeper (eds.) *Cognitive Stylistics. Language and cognition in text analysis*. Amsterdam; Philadelphia: J. Benjamins Pub. Co., 95-122.
- Semino, E. (2007). Mind Style Twenty-five Years On. *Style*, 41, 2, 153-173.
- Spengler, O. (1991 (1918)). *The decline of the West*. New York: Oxford University Press, 1991.
- Tomes, Y. (2004). *Cognitive style, achievement, and ethnicity: A study in higher education*. Ph.D. dissertation. Richmond, Virginia: Virginia Commonwealth University.
- Tsur, R. (2002). Aspects of Cognitive Poetics. In E. Semino, J. Culpeper (eds.) *Cognitive Stylistics. Language and cognition in text analysis*. Amsterdam; Philadelphia: J. Benjamins Pub. Co., 279-318.
- Tulviste, P. (1991). *The cultural-historical development of verbal thinking*. Commack, N.Y.: Nova Science Publishers.
- Vygotsky, L. 1984 (1933/34). Rannee detstvo. In Vygotsky, L. *Sobranie sochinenii*, 4. M.: Pedagogika, 340-367.
- Vygotsky, L. (1986(1934)). *Thought and language*. Cambridge, Mass.: MIT Press.
- Witkin H., Moore C., Goodenough D., Cox P. (1977). Field-Dependent and Field-Independent Cognitive Styles and Their Educational Implications. *Review of Educational Research*, 47, 1, 1-64.

# Development of Category-Based Reasoning in Preschool-Age Children: Preliminary Results of a Longitudinal Study

**Karrie E. Godwin (kegodwin@andrew.cmu.edu)**

Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue  
Pittsburgh, PA 15213 USA

**Bryan J. Matlen (bmatlen@andrew.cmu.edu)**

Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue  
Pittsburgh, PA 15213 USA

**Anna V. Fisher (fisher49@andrew.cmu.edu)**

Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue  
Pittsburgh, PA 15213 USA

## Abstract

Category-based reasoning is central to mature cognition; yet, the developmental course of this fundamental ability remains unclear. We designed a longitudinal study to investigate the development of category-based reasoning. We also took an individual differences approach to identify possible cognitive factors that may facilitate category-based reasoning. In this paper we report preliminary results of our longitudinal investigation into the development of category-based reasoning.

**Keywords:** Induction. Reasoning. Categories. Cognitive Development

## Introduction

A great deal of prior research has investigated the development of category-based reasoning. This work suggests that the fundamental ability to make inferences on the basis of category labels (i.e., category-based reasoning) is early developing (Gelman & Coley, 1990; Gelman & Markman, 1986; Jaswal, 2004; Jaswal & Markman, 2007; Welder & Graham, 2001). In a simple test of this skill, Jaswal and Markman (2007, Experiment 1) presented 24-month-old children with pairs of familiar animals (e.g., *dog* and *cat*). The children watched as the animals engaged in specific activities (e.g., the cat drinks milk and the dog chews on a bone). A third hybrid-animal was then presented (e.g., a cross between a cat and dog) and children were asked to use the props to demonstrate which action the hybrid animal would make (e.g., drink milk or chew on a bone). Importantly, the hybrid animal was designed to look more similar to one of the targets (e.g., The hybrid animal was designed to look more similar to the cat). In the no-label condition the hybrid was referred to generically (e.g., the experimenter labeled the cat-like animal as “*this one*”). In the label condition the hybrid animal was labeled counter-intuitively (e.g., the experimenter labeled the cat-like animal a “*dog*”). In the no-label condition Jaswal and Markman found that 24-month-olds generalized based on perceptual similarity 69% of the time. However, in the label condition, when perceptual similarity was pitted against category information, perceptually-based generalizations

dropped to 37%. These results suggest children as young as 24-months of age can utilize labels to infer category membership.

In a seminal study, Gelman and Markman (1986) examined children’s ability to make inductive inferences using category information that was conveyed by synonymous labels. In this experiment, preschool-aged children were presented with triads of objects and provided with respective labels. The children were told that two of the objects possessed particular properties, and the children were asked to infer which property the third object possessed. For example, children were presented with a bunny and a squirrel and told that the bunny eats grass, and the squirrel eats bugs. Subsequently, the children were asked to determine whether the rabbit ate grass like the bunny or bugs like the squirrel. Gelman and Markman found that children made category-based inductions 63% of the time, which is slightly above chance, and posited that preschool children are sensitive to the cues synonymous labels provide about category membership.

Despite these intriguing findings, there is mounting evidence demonstrating that the course of category-based reasoning follows a more protracted developmental course (Fisher, 2010; Fisher, Matlen, & Godwin, 2011). For example, Fisher et al. (2011) found that children’s ability to make inductive inferences using synonyms is limited to a small set of semantically-similar words that co-occur in child-directed speech according to the CHILDES database (MacWhinney, 2000). In particular, Fisher et al. found that most 4-year-old children were able to perform category-based inferences with synonyms that are likely to co-occur in child-directed speech (e.g., *bunny-rabbit*, *puppy-dog*); however, they were unlikely to make category-based inferences with non co-occurring synonyms (e.g., *alligator-crocodile*, *rock-stone*). This pattern of results was found with both natural kinds and artifacts. Additionally, children’s reliance on category information was found to improve gradually with age. Although 5-year-olds evidenced improvement in their reliance on category information compared to 4-year-olds, the majority of children did not reliably utilize category information

conveyed by non-co-occurring semantically-similar labels until six years of age.

If category-based reasoning has a protracted developmental course, an important question to be addressed is identifying what actually develops that enables children to utilize labels as windows into categories and reliably use this information in the course of induction.

One possibility is that advances in category-based reasoning are facilitated by changes in how children organize knowledge. There is evidence that children begin to organize concepts into networks by 21 months (Arias-Trejo & Plunkett, 2009). There is also evidence that conceptual organization changes over the course of development, with associative networks emerging prior to semantic networks (McCauley, Weil, & Sperber, 1976; Plaut & Booth, 2000). It is also possible that development of executive functioning may facilitate category-based reasoning by allowing children to disengage their attention from – often misleading or irrelevant – surface similarities and consider deeper relational similarities (Sloutsky & Fisher, 2005; Sloutsky, 2010).

The goal of the present research is to examine possible cognitive factors contributing to the development of category-based reasoning. Towards this goal we designed a longitudinal investigation taking an individual differences approach. Specifically, we collected measures of children’s category-based reasoning at Time 1, verbal working memory, IQ, and semantic knowledge organization. Collection of additional measurements (i.e., inhibitory control, non-verbal working memory, semantic priming, and category-based reasoning at Time 2) is currently in progress. In what follows we report the preliminary results of this study.

## Method

### Participants

Participants were 43 four-year-old children from a local preschool (*Age*=4.32 years, *SD*=0.28 years, 20 females, 23 males).

### Materials & Procedure

Children were tested individually in a quiet room adjacent to their classroom by a trained research assistant. The tasks were administered across 6 sessions over the course of approximately 2 weeks. A detailed description of each task is provided below.

#### Category-Based Reasoning Task

The category-based reasoning task consisted of a triad induction task. Visual Stimuli were sets of three identical doors which were presented on a computer; see Figure 1. Verbal stimuli included 9 label triads: 3 triads referring to animate natural kinds, 3 triads referring to inanimate natural kinds, and 3 triads referring to artifacts (see Table 1). The properties participants were asked to generalize consisted of two-syllable blank predicates. Each trial was comprised of a target item, a category-choice, and an unrelated lure (e.g., *rock-stone-grass*). The children were told that the objects

were hiding behind doors. This design was employed to encourage children’s reliance on category information conveyed via labels. This procedure has been successfully used in prior research (Fisher et al., 2011). On every trial children were told what was hiding behind each door.

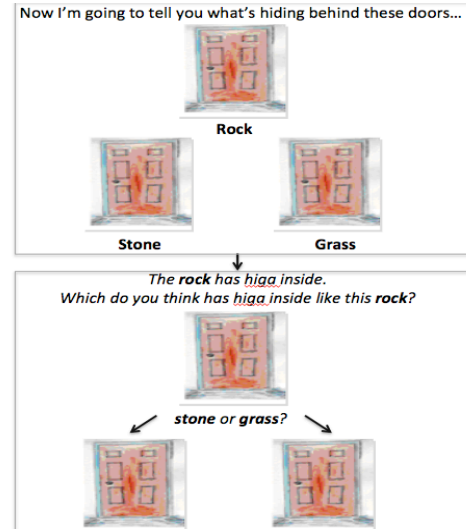


Figure 1: Schematic depiction of the category-based reasoning task. All instructions were given verbally by the experimenter.

Then, children were told that the target had a novel-property and they were asked to generalize the property to one of the test items (the category-choice or lure).

Table 1: Linguistic Stimuli for the Category-Based Reasoning Task

Target	Category Choice	Lure	Property
Rock	Stone	Grass	Higa
Alligator	Crocodile	Butterfly	Omat
Rug	Carpet	Window	Koski
Rat	Mouse	Fish	Lignin
Hill	Mountain	Flower	Erwin
Sea	Ocean	Apple	Manchin
Sofa	Couch	Cup	Creighan
Shoe	Boot	Car	Troxel
Lamb	Sheep	Frog	Matlen

The trials were presented in one of two orders: all trials were randomized for order 1, and for order 2 the presentation was reversed. Presentation order was counterbalanced across participants. The reasoning task was administered during session 2 and again in session 4 to assess stability in children’s generalization performance.

#### Picture Identification Task

The picture identification task served to assess children’s familiarity with the labels used in the reasoning task. Verbal

stimuli included 27 labels (the target, category-choice, and lure from the reasoning task). Visual stimuli consisted of a set of 108 pictures presented on a computer. All instructions and labels were given by hypothesis-blind experimenters. The picture identification task is similar to the Peabody Picture Vocabulary Test (Dunn & Dunn, 1997). On each trial children were asked to point to the object labeled by the experimenter from 4 pictorial response options (the target and 3 lures). The trials were presented in one of two orders. Presentation order was counterbalanced across participants.

#### Intelligence Test

IQ materials consisted of a commercially purchased intelligence test, the *Wechsler Preschool and Primary Scale of Intelligence* (WPPSI). The IQ test was administered in order to assess if children's reasoning performance was related to their general intelligence and/or a particular intelligence component. Eight of the WPPSI subscales were administered over 3 testing sessions in order to obtain an index of children's Verbal IQ, Performance IQ, Processing Speed Quotient, and Full Scale IQ. The WPPSI was administered by the first author of this paper and two trained research assistants.

#### Verbal Working Memory Tasks

Children's verbal working memory capacity was assessed using a simple and complex word-span task. Verbal stimuli entailed 60 words that were arranged into 6 sets. Set length ranged in size from a list length of 2 words to 6 words. Each set was comprised of 3 lists of the same length.

In the simple word-span task, children listened to the experimenter read a series of familiar count nouns, as judged by the MacArthur Communicative Development Inventory (Dale & Fenson, 1996). Then, children were asked to recite the words in the same order in which they were presented. The number of words in each set increased monotonically after children correctly completed two out of three trials within a given set size. For example, if children correctly completed 2 trials with set size 2, they then moved on to set size 3 (for a minimum of 2 trials or a maximum of 3 trials), and then set size 4 (for a minimum of 2 trials or a maximum of 3 trials), and so on until children made two errors within a set at which point testing stopped. The child's score is the longest list length he or she recited successfully. The complex word-span task was identical to the simple word-span task, except that children were asked to repeat items in the reverse order in which they were presented (e.g., If children were given the string, "*duck, house, chair*", the correct response would be "*chair, house, duck*"). The word lists were presented in one of two orders. Presentation orders were counterbalanced across participants.

The simple word-span task was included in the assessment battery in order to assess if children's performance on the category-based reasoning task was related to their general working memory capacity. The complex word-span task was included as it was

hypothesized that the complex word-span task more closely resembled the demands of the reasoning task itself (e.g., both the category-based reasoning task and the complex word-span task contain a memory component as well as a transformation/processing component).

#### Semantic Space Task

Visual stimuli included a game board consisting of a 9x9 grid. Two 1" wooden cubes served as the game pieces. Verbal stimuli consisted of 24 animal pairs. The list of linguistic stimuli is provided in Table 2. In the semantic space task, children are asked to help *Zibbo the Zookeeper* organize his zoo. Children are told that Zibbo wants to put animals of the same kind close together. Children are presented with 24 animal pairs (i.e., a target animal and a test item). Of the 24 animal pairs, 6 dyads were semantically-similar (e.g., *lamb-sheep*), 6 dyads shared a common habitat or setting (e.g., *lamb-horse*), 6 dyads were physically similar – according to size and/or color (e.g., *lamb-swan*), and 6 dyads served as filler trials. Note that the target animal was paired with 3 different animals throughout the game - the category-choice, a physically-similar item, and the habitat match. On each trial, the experimenter shows the child where Zibbo put the target animal (e.g., the experimenter places the game piece on a designated space on the board and tells the child, "*The zookeeper put the crocodile here*"). Then, the experimenter hands the child the second game piece and asks the child where the test item should go (e.g., "*Where do you think the grasshopper should go?*"). The board is then cleared and the experimenter presents the next dyad. The child's response on each trial is recorded so the distance between the target animal and test item can be calculated.

Placement of the 18 critical trials (i.e., semantically-similar dyads, physically similar dyads, or similar habitat dyads) was pseudo randomized to eight potential squares; see Figure 2. Each square was utilized at least twice and no more than three times. The 6 filler trials were randomly assigned to one of the remaining 24 squares in order to encourage participants to use the entire game board. Trials were presented in one of two orders and presentation orders were counterbalanced across participants.

Table 2: List of Stimuli for the Semantic Space Task

Critical Trials			
Target	Category - Choice	Physical Similarity	Habitat
Crocodile	Alligator	Grasshopper	Fish
Chick	Hen	Goldfish	Goat
Lamb	Sheep	Swan	Horse
Whale	Dolphin	Elephant	Octopus
Monkey	Gorilla	Chipmunk	Parrot
Mouse	Rat	Hippo	Pig
Filler Pairs			
1. Zebra/Turkey; 2. Bear/Snake; 3. Panther/ Turtle; 4. Tiger/Butterfly; 5. Frog/Lion; 6. Giraffe/Seal			

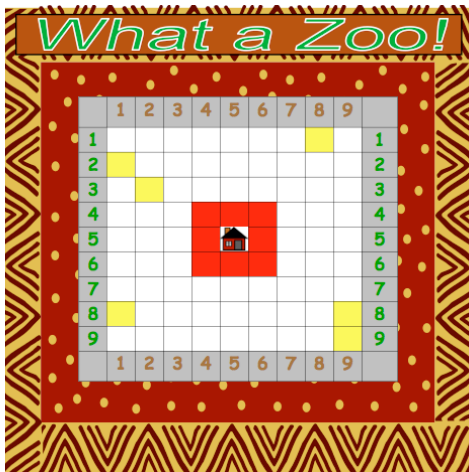


Figure 2: Depiction of the game board for the semantic space task. Red squares indicate the location of the critical trials and the yellow squares mark the location of the filler trials.

The semantic space task was included in the assessment battery to assess whether the organization of children's semantic space was related to their performance on the category-based reasoning task. Specifically, we were interested in identifying how children weight different dimensions (e.g., semantic-similarity, physical similarity, and habitat) and whether the distribution of weights to various dimensions enhances or hinders children's ability to successfully make category-based inductions.

## Results

### Picture Identification

The results of the picture identification task suggest that children possessed the prerequisite knowledge to perform category-based induction as children were highly familiar with the labels used in the reasoning task ( $M=0.92$ ,  $SD=0.14$ ). Additionally, the correlation between children's performance on the picture identification task and children's average reasoning score was only marginally significant ( $r=.28$ ,  $p=0.07$ ).

### Category-Based Reasoning Task

As stated previously, performance on the category-based reasoning task was measured twice over the course of 1 week in order to examine the stability of this measure. Mean category-based reasoning at Time 1a and 1b were very similar ( $M=0.62$ ,  $SD=0.22$ ;  $M=0.63$ ,  $SD=0.26$  respectively) and these measures were significantly correlated ( $r=.483$ ,  $p=0.001$ ). Proportions of category-based responses were compared to chance level (0.5) using single-sample t-tests. All mean scores (scores at Time 1a & 1b and average reasoning score) were significantly above chance; all  $t$ 's  $> 3.30$ , all  $p$ 's  $< 0.0022$ .

To investigate individual patterns of responses, participants were classified as either category-based or non-category-based responders. A category-based responder was defined as a participant who gave a category-based response on at least 7 out of 9 (78%) trials (binomial probability = 0.09). At Time 1a and 1b only a small percentage of children were classified as category-based responders (33% and 37% respectively).

To further investigate stability in children's category-based reasoning performance we also examined whether children's classification remained stable across Time 1a and 1b. We found that 67% (29 out of 43) of children were categorized as stable across Time 1a and 1b. Of these children only 19% (8 out of 43) were classified as consistently category-based responders, 49% (21 out of 43) were consistently non category-based, and 33% (14 out of 43) were considered unstable responders; See Figure 3. For the purposes of the remaining analyses the average reasoning score was utilized ( $M=0.63$ ,  $SD=0.21$ ).

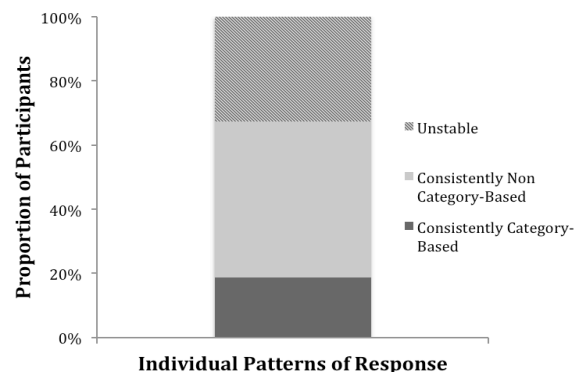


Figure 3: Proportion of children classified as consistently category-based, consistently non-category-based, or unstable responders.

### Intelligence Test

Children's mean composite IQ scores and their Full Scale IQ were in the average range ( $MVIQ=107.36$ ,  $SD=24.30$ ;  $MPIQ=107.26$ ,  $SD=14.78$ ;  $MPSQ=92.78$ ,  $SD=23.56$ ;  $MFSIQ=107.05$ ,  $SD=15.79$ ). Children's Verbal IQ and Performance IQ composite scores were significantly correlated with their average performance on the category-based reasoning task ( $r=.33$ ,  $p=0.03$ ;  $r=.31$ ,  $p=0.05$  respectively). However, Processing Quotient was not significantly correlated with children's average reasoning score ( $r=.15$ ,  $p=0.35$ ). Children's Full Scale IQ was also significantly correlated with their average reasoning score ( $r=.50$ ,  $p=0.001$ ).

### Verbal Working Memory Tasks

Children's performance on the simple word-span task was better than their performance on the complex word-span task ( $M=3.07$ ,  $SD=1.32$ ;  $M=1.28$ ,  $SD=1.14$  respectively). A

mean score of 3.07 on the simple word-span task indicates that on average children were able to successfully recall a list length of 3 words. Children's score on the simple word-span task was found to be correlated with their average performance on the category-based reasoning task ( $r=.35$ ,  $p=0.02$ ). A mean score of 1.28 on the complex word-span task suggests that many children obtained a score of 0 on the task as the smallest list length was 2 words. Performance on the complex word-span task was not significantly correlated with children's average induction performance ( $r=.08$ ,  $p=0.60$ ), possibly due to floor effects on the complex word-span task.

### Semantic Space Task

Children's semantic space score was calculated in the following way: First, for each child an average score for each category was calculated (i.e., an average score for semantically-similar dyads, an average score for similar habitat dyads, and an average score for physically similar dyads). Children's scores for similar habitat dyads and physically similar dyads were averaged together to create an average score for non-semantically-similar dyads. This score was subtracted from the average score for semantically-similar dyads to obtain a difference score. Larger difference scores indicate that children placed semantically-similar dyads closer together and non-semantically-similar dyads farther apart. Smaller difference scores indicate that children did not reliably discriminate between semantically-similar dyads and non-semantically-similar dyads.

Table 3: Correlation Matrix

(Note: CB Reasoning = Category-based reasoning, Pic ID = picture identification, SWS = simple word-span, CWS = complex word-span, SS = semantic space, VIQ = verbal IQ, PIQ = performance IQ, PSQ = processing speed, FSIQ = full scale IQ)

	1	2	3	4	5	6	7	8	9
1. CB Reasoning	-	.283	.345*	.082	.460**	.332*	.305*	.149	.496**
2. Pic ID	-	-	.706**	.275	.228	.723**	.100	-.157	.584**
3. SWS	-	-	-	.415**	.313*	.671**	.182	-.038	.578**
4. CWS	-	-	-	-	.126	.170	.188	.009	.202
5. SS	-	-	-	-	-	.384*	.432**	.145	.526**
6. VIQ	-	-	-	-	-	-	.049	-.314*	.800**
7. PIQ	-	-	-	-	-	-	-	.598**	.540**
8. PSQ	-	-	-	-	-	-	-	-	.123
9. FSIQ	-	-	-	-	-	-	-	-	-

Note. \*. Correlation is significant at the 0.05 level. \*\*. Correlation is significant at the 0.01 level

Children's mean score for semantically-similar dyads was 4.37 ( $SD=1.35$ ). Children's score for physically similar dyads and similar habitats was 5.64 ( $SD=1.76$ ) and 5.83 ( $SD=1.60$ ) respectively. Children's mean score for non semantically-similar dyads was 5.74 ( $SD=1.51$ ). Difference scores ranged from -2.58 to 5.67 suggesting considerable variability in children's performance on this task. The average difference score was 1.37 ( $SD=1.88$ ). Children's performance on the semantic space task was found to be significantly correlated with their average performance on the category-based reasoning task ( $r=.46$ ;  $p=0.002$ ).

### Predicting Category-Based Reasoning Performance

There were a total of 8 possible predictors of children's reasoning performance. As can be seen in Table 3, several of the predictors were significantly correlated with each other. Concerns regarding collinearity are allayed as tolerance values for predictors entered into the regression model were within the acceptable range. Children's scores on the simple word-span task, semantic space task, and FSIQ were entered into the model as predictors. Children's average score on the category-based reasoning task was the dependent variable.

The regression model significantly predicted children's average reasoning score,  $F(3, 38)=5.47$ ,  $p=0.003$ . The  $R$  squared value indicates that 30% of the variance in children's performance on the category-based reasoning task was explained by the model. Simple word-span was not found to be a significant predictor ( $\beta=0.08$ ,  $p=0.62$ ); however, semantic space ( $\beta=0.26$ ,  $p=0.11$ ) and FSIQ ( $\beta=0.31$ ,  $p=0.10$ ) were marginally significant predictors of children's reasoning performance.

## Discussion

Overall, the results from this study, although preliminary, point to several findings. First, the analysis of individual patterns of response on the reasoning task replicate previous work (Fisher et al., 2011; Godwin, Matlen, & Fisher, 2011). Specifically, we found that when young children are presented with non-co-occurring semantically-similar labels only a small percentage of children spontaneously engage in category-based reasoning.

Second, the present findings suggest that several factors are related to children's induction performance. We found that children's Full Scale IQ, Verbal IQ, Performance IQ, simple working memory, and semantic space performance were all significantly correlated with children's average reasoning score. Additionally, children's Full scale IQ and performance on the semantic space task were identified as unique predictors of children's performance on the category-based reasoning task according to the regression model. The correlation between semantic organization and category-based reasoning is also consistent with related research on the development of analogical reasoning, which has suggested that children's shift from focusing primarily on perceptual similarity to relational similarity is mediated by increases in domain knowledge (Rattermann & Gentner,



1998). On-going research will examine whether other cognitive factors (e.g., inhibitory control, non-verbal working memory, etc.) are also related to the development of category-based reasoning.

Third, the present study provides novel information on the stability in children's induction performance. This study is the first to our knowledge to look at the stability in children's performance on a conceptual development measure. The findings from this study suggest that children's category-based reasoning performance between Time 1a and 1b is correlated; however, there is still a great deal of variability in children's performance as indicated by the small percentage of children who were classified as consistently category-based across the two time points. Additionally, the longitudinal component of this study will enrich our understanding of the stability in children's inductive reasoning. Once data collection is complete, we will be able to examine whether the percentage of children who are classified as consistently category-based increases with age.

In conclusion, the present study contributes to our understanding of children's emerging ability to engage in category-based reasoning. The contributions of this work include identifying potential factors that may be predictive of children's induction performance as well as the opportunity to investigate the stability of children's category-based reasoning. Future research is needed to extend these findings and disentangle the different hypotheses put forth to explain this fundamental aspect of conceptual development.

### Acknowledgements

We thank Malika Sinha, Laura Pacilio, Alyssa Montanaro, Anna Loiterstein, Like Li, and Kayoung Joung for their help collecting data. We also thank the children, parents, and teachers who made this project possible. This work was supported in part by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education (R305B090023 and R305B040063).

### References

- Arias-Trejo, N. & Plunkett, K. (2009). Lexical-semantic priming effects during infancy. *Philosophical Transactions of the Royal Society Biological Sciences*, 364, 3633-3647.
- Dale, P. S., and Fenson, L. (1996). Lexical development norms for young children. *Behavioral Research Methods, Instruments, & Computers*, 28, 125-127.
- Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test-Third Edition*. Circle Pines, MN: AGS Publishing.
- Fisher, A.V. (2010). What's in the name? Or how rocks and stones are different from dogs and puppies. *Journal of Experimental Child Psychology*, 105, 198-212.
- Fisher, A.V., Matlen, B.J., & Godwin, K.E. (2011). Semantic similarity of labels and inductive generalization: Taking a second look. *Cognition*, 118, 432-438.
- Gelman, S.A. & Coley, J.D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology*, 26, 796-804.
- Gelman, S.A., & Markman, E. (1986). Categories and induction in young children. *Cognition*, 23, 183-209.
- Godwin, K., Matlen, B., & Fisher, A. (2011). The influence of co-occurrence and inheritance information on children's inductive generalization. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33<sup>rd</sup> Annual Conference of the Cognitive Science Society* (2009-2011). Austin, TX: Cognitive Science Society.
- Jaswal, V.K. (2004). Don't believe everything you hear: Preschoolers' sensitivity to speaker intent in category induction. *Child Development*, 3, 279-300.
- Jaswal, V. K. & Markman, E. M. (2007). Looks aren't everything: 24-month-olds' willingness to accept unexpected labels. *Journal of Cognition and Development*, 8(1) 93-111.
- MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.
- McCauley, C., Weil, C. M., & Sperber, R. D. (1976). The development of memory structures as reflected by semantic-priming effects. *Journal of Experimental Child Psychology*, 22, 511-518.
- Plaut, D.C. & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, 107(4), 786-823.
- Rattermann, M. J. & Gentner, D. (1998). More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task. *Cognitive Development*, 13, 453 - 478.
- Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive Science*, 34, 1244-1286.
- Sloutsky, V. M., & Fisher, A. V. (2005). Similarity, Induction, Naming, and Categorization (SINC): Generalization or verbal inductive reasoning? Response to Heit and Hayes. *Journal of Experimental Psychology: General*, 134, 606-611.
- Welder, A.N., & Graham, S.A. (2001). The influences of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. *Child Development*, 72, 1653-1673.



# Is that your final answer? The effects of neutral queries on children's choices

Aaron Gonzalez<sup>1</sup>, Patrick Shafto<sup>2</sup>, Elizabeth Bonawitz<sup>1</sup>, & Alison Gopnik<sup>1</sup>

<sup>1</sup>Department of Psychology, University of California, Berkeley

<sup>2</sup> Department of Psychological and Brain Sciences, University of Louisville

## Abstract

Preschoolers often switch a response on repeated questioning, even though no new evidence has been provided (Krahenbuhl, Blades, & Eiser, 2009). Though apparently irrational, this behavior may be understood as children making an inductive inference based on their beliefs about whether initial responses were correct and the knowledgeability of the questioner. We present a probabilistic model of how the questioners' knowledge and biases to be positive should affect inferences. The model generates the qualitative prediction that an ideal learner should switch responses more often following a "neutral query" from a knowledgeable questioner than following queries from an ignorant questioner. We test predictions of the model in an experiment. The results show that four-year-old children are sensitive to questioners' knowledge when responding to a neutral query, demonstrating more switching behavior when the query is provided by a knowledgeable questioner. We conclude by discussing the practical and theoretical implications for cognitive development.

When should a learner abandon their current hypothesis in favor of a new one? It is becoming clear that even preschool-aged children rationally update beliefs and generate new explanations following informative evidence (Gopnik, Glymour, Sobel, Schulz, & Danks, 2004; Schulz, Bonawitz, & Griffiths, 2007; Denison, Bonawitz, Gopnik, & Griffiths, 2010; Bonawitz, Schijndel, Friel, & Schulz, 2012; Bonawitz, Fisher, & Schulz, in press). These tasks often involve assessing children's starting belief state, either presenting the child with new evidence or allowing the child to generate their own evidence, and then eliciting an explanation or judgment. What constitutes "evidence" in these tasks is fairly intuitive; for example, children may be presented with a storybook containing information about two variables that tend to co-occur (Schulz et al., 2007) or they may observe a toy that reacts when certain objects are placed on top of it (Bonawitz, Denison, Gopnik, & Griffiths, 2011).

In addition to revising beliefs following correlational evidence and interventions, children also learn from others. Interestingly, even preschool-aged children do not do so indiscriminately; they use information about informants' knowledge and intent to guide who they trust. For example, children do not trust informants who label familiar objects incorrectly (Koenig & Harris, 2005), who dissent from groups (Corriveau, Fusaro, & Harris, 2009), and even with familiar informants, children update social inferences (Corriveau & Harris, 2009). Similarly, recent research suggests that children use information about informants' intent to guide inferences (Mascaro & Sperber, 2009; Shafto, Eaves, Navarro, & Perfors, in press). Other research has suggested that even four-year-old children make different causal inferences depending on the social context when evidence is presented: direct instruction by a knowledgeable and helpful informant

provides strong constraints on likely hypotheses as compared to accidental information by a not-knowledgeable informant. Even when contrasted with intentional (but not instructional) actions and interrupted demonstrations, preschoolers make stronger inferences about causal structure from direct instruction by leveraging the informant's knowledge and intent (Bonawitz, Shafto, et al., 2011; Buchsbaum, Gopnik, Griffiths, & Shafto, 2011; Butler & Markman, 2010). These pedagogical assumptions have been captured by probabilistic models (e.g. Shafto & Goodman, 2008; Bonawitz, Shafto, et al., 2011), which suggest a rational account of how learners update their beliefs following evidence generated in the context of teaching.

These literatures suggest that preschool children are sophisticated and powerful social learners; they revise their beliefs when evidence is sufficient and use social or instructional contexts to help interpret the strength of the evidence. However, other research suggests that children may abandon hypotheses too capriciously. For example, the effects of repeated questioning on eyewitness testimony in young children have been studied extensively in the context of the judicial system, and work done by Poole and White (1991) found that, in contrast to adults, four-year-olds were more likely to change their responses to repeated yes or no questions. More recently Krahenbuhl et al. (2009) found that children as young as four changed over a quarter of their responses to repeated questions, resulting in a decline in accuracy. In Howie, Sheehan, Mojarrad, and Wrzesinska (2004) four-year-old children were more likely than older children to change responses toward an incorrect answer on repeated questioning. That is, although no additional evidence was provided, by simply asking children the same question a second time, children were very likely to switch their predictions. How might we reconcile these findings with those suggesting that children should only rationally update beliefs following informative evidence?

One explanation for this apparently irrational switching behavior is that seemingly neutral questions such as "Is that your final answer?" may provide strong information in certain social contexts. If preschoolers are sensitive to the potential communicative intent behind such queries, the question itself may be a source of evidence about whether an initial guess was correct. Consider a game in which a sticker is hidden under one of two cups. Suppose an informant asks the child which cup they believe the sticker is under and after the child's initial guess, the teacher asks, "Is that your final answer? Would you like to change your guess?" In which contexts does such a question provide information about the true location of the sticker? Intuitively, it seems obvious that if the

questioner does not know the actual location of the sticker, then the repetition of the question provides little additional evidence; however, a *knowledgeable* questioner might have a good reason for giving the learner with a second chance at answering the question. In these cases, this apparently neutral query is not neutral at all; it is a strong cue to the learner that they should change their answer.

In what follows we will explore the idea that even a “neutral” query carries information about the state of the world when a questioner is knowledgeable. We present a probabilistic model that demonstrates how an ideal learner might evaluate such “neutral” queries in scenarios in which the questioner is knowledgeable and scenarios in which she is ignorant. With the model, we evaluate the conditions under which switching guesses is the rational choice for the learner. We then test the basic prediction with preschoolers in an experiment in which the informant’s knowledge or ignorance is made explicit. We suggest that repeated questioning does indeed lead a learner to switch responses and that even preschoolers are sensitive to the knowledge state of others when making such inferences.

### Modeling learners’ responses to neutral queries

Here we consider a model of “neutral queries”. Bayesian inference provides a natural framework in which to consider how an ideal learner should update her beliefs following this kind of information. In Bayesian inference, the learner’s goal is to update their beliefs about hypotheses,  $h$ , given data,  $d$ , where the degree of belief in a hypothesis given data is denoted  $P(h|d)$ . These updated posterior beliefs are determined by two factors: the learner’s prior beliefs in hypotheses,  $P(h)$ , and the probability of sampling the observed data, assuming the hypothesis is true,  $P(d|h)$ . Specifically, the updated posterior belief in a particular hypothesis is proportional to the product of the prior belief in that hypothesis and the probability of sampling the data given that hypothesis,  $P(h|d) \propto P(d|h)P(h)$ .

Because we are considering only two hypotheses, we can use Bayes Odds to simplify the problem:

$$\frac{P(h_1|d)}{P(h_2|d)} = \frac{P(d|h_1)P(h_1)}{P(d|h_2)P(h_2)}, \quad (1)$$

where  $P(h_1|d)$  is the probability that the sticker is in the first location, given the statement from the informant (“correct”, “incorrect”, “is that your final answer”) and  $P(h_2|d)$  is the probability that the sticker is in the second location given the statement. It is reasonable to assume that the learner has no a priori assumptions about either location, which allows us to cancel out the prior beliefs (i.e.  $P(h_1) = P(h_2)$ ), thus  $\frac{P(h_1)}{P(h_2)} = 1$ .

The main issue is the probability of the statement given the location of the sticker,  $P(d|h)$ . We can model this likelihood with a simple causal graphical model (see Figure 1). Causal graphical models consist of a structure indicating the causal relationships among a set of variables, where nodes are

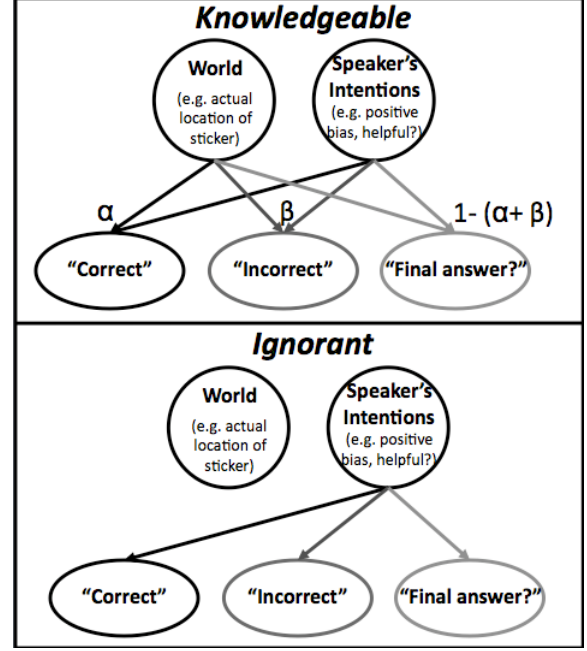


Figure 1: Graphical model depicting dependencies in cases when the informant is *knowledgeable* and *ignorant*.

variables and dependence relationships are indicated by arrows from causes to effects. To complete a graphical model, conditional probability distributions give the probability that each variable takes on a particular value given the value of its causes (Pearl, 2000; Spirtes, Glymour, & Schienens, 1993, see Table 1).

In our model of the problem, the variables include the actual state of the world (“World”, i.e. location of the sticker), the intention of the speaker (to provide helpful feedback, to avoid negative feedback, etc.), and the possible statements the informant can make (“Correct”, “Incorrect”, “Is that your final answer?”). In the first model (*Knowledgeable*) the informant is aware of the actual state of the world. As a result, both the true location of the sticker and the intention of the speaker influence the statement given to the learner. In the second model (*Ignorant*), the informant is not aware of the actual state of the world. As a result, the true state of the world does not influence the statement given to the learner. That is, the state of the world and the information provided are causally independent of each other.

The dependence assumptions captured by the graphical model generate predictions about the behavior of learners. In the case of a knowledgeable informant, given information about the actual state of the world and the informant’s statement, a learner can infer something about the informant’s goals (e.g. to provide positive or negative feedback). Given information about the state of the world and the informant’s goals, a learner could also predict (with some probability) the likelihood that the informant would produce different statements. In our problem, given the informant’s goals and the statement provided, the learner can make an inference about

Table 1: Conditional probability table for knowledgeable graphical model.

Guess	“Correct”	“Incorrect”	“Final answer?”
Correct	$\alpha_c$	$\beta_c \approx 0$	$1 - (\alpha_c + \beta_c) \approx 1 - \alpha_c$
Incorrect	$\alpha_i \approx 0$	$\beta_i$	$1 - (\alpha_i + \beta_i) \approx 1 - \beta_i$

the state of the world. A learner could also make more abstract inferences: given information about the goals of the informant, the statement provided, and the actual location of the sticker, a learner could infer which model (*Knowledgeable* vs *Ignorant*) best captures the knowledge state of the informant.

The specification of the conditional probability distribution provides additional qualitative predictions. In the *knowledgeable* graphical model, we might reasonably argue that if the child chooses the correct location initially, the speaker is very unlikely to say “incorrect”; in this case  $P(\text{“incorrect”}|\text{correct choice}) = \beta_c \approx 0$ . Similarly, if the child chooses the incorrect location initially the speaker is very unlikely to say “correct”,  $P(\text{“correct”}|\text{incorrect choice}) = \alpha_i \approx 0$ . Given these intuitive assumptions, we can compare three possible ways biases about the goals of the teacher might play out in the model’s predictions.

The first possibility is that the informant is *unbiased*. Let us consider the case when the informant is knowledgeable. In this case, the *unbiased informant* is just as likely to say “correct” when the initial guess is correct as “incorrect” when the initial guess is incorrect,  $\alpha_c \approx \beta_i$ . If this is the case, then the learner can not infer whether their initial guess is correct or not if they hear the statement “is that your final answer”. This is because  $P(\text{“final answer?”}|\text{correct}) = P(\text{“final answer?”}|\text{incorrect})$ . That is, the statement “is that your final answer” provides no additional information about the location of the sticker (Equation 1 is approximately equal to 1). Now consider the case where the informant is ignorant. In this case, because the informant has no information about the actual state of the world, the true location is conditionally independent of the statements made by the informant, and the learner cannot make any inferences about the state of the world. Thus, assuming unbiased informants, learners should make the same inferences if asked “is that your final answer” in a *knowledgeable* condition as if asked “is that your final answer” in an *ignorant* condition. This model does not predict the degree to which the learners should switch responses, but it does predict no difference between conditions.

A second possibility is that the informant is *positively biased*. In this model, the knowledgeable informant may be inclined to want to say “correct” following correct initial guesses, but would be reluctant to say “incorrect” following an incorrect initial guess,  $\alpha_c > \beta_i$ . If this is the case, then the statement “is that your final answer” provides support for the hypothesis that the learner’s initial guess was incorrect because she is more likely to hear “is that your final answer” given an incorrect guess than “is that your final answer” given a correct guess (Equation 1 > 1). Thus, the *positively biased* model predicts that a learner should show increased switch-

ing in a *Knowledgeable* condition as compared to an *Ignorant* condition (in which the state of the world is still conditionally independent of the statements and thus does not provide additional information).

The third possibility is that the informant is *negatively biased*. In this model, the informant may be inclined to say “incorrect” following an incorrect initial guess, but would be comparatively reluctant to say “correct” following a correct initial guess,  $\alpha_c < \beta_i$ . If this is the case, then the statement “is that your final answer” provides support for the hypothesis that the learner’s initial guess was correct (Equation < 1) and the learner should show a decrease in switching responses in the *knowledgeable* condition as compared to an *Ignorant* condition.

Note that the precise values of  $\alpha$  and  $\beta$  are not important for the predictions of this model, but the relationship between these variables drives the predictive differences. We investigate three implications of this model: first, do we replicate the finding that preschoolers tend to switch responses following what might be considered a “neutral query”; second, do preschoolers take the knowledge state of the informant into account when inferring whether or not to switch hypotheses; third, do preschoolers assume that the informant is neutral, positively, or negatively biased when they provide a query?

## Experiment: Preschoolers’ switching behavior in response to a neutral query

To investigate the predictions of our model we invited preschoolers to participate in a game where the goal was to guess the location of a sticker under one of two cups. After their initial guesses, children were given some feedback and the opportunity to change their guess. After two training trials in which the experimenter told the child that their first guess was either correct or incorrect, children were given three test trials. In the test trials the experimenter asked the child “Is that your final guess?” after children’s initial guesses. Some children participated in a condition in which the experimenter looked under the cups before generating the query and others participated in a condition in which the experimenter did not look before the query. The critical measure is simply on what percentage of trials children switched their prediction to the other cup by condition.

## Method

**Participants** Thirty-two preschoolers (mean age: 58.6 months; range: 48-79 months) were recruited from local preschools and museums for participation.

**Design** Preschoolers were randomly assigned to either the *Knowledgeable* condition or *Ignorant* condition. Four children were dropped and replaced for demonstrating a side bias (see results) in the *Knowledgeable* condition and five children were dropped and replaced for side biases in the *Ignorant* condition. There were no differences in ages between condition,  $t(30) = 0.35, p = ns$ .

**Materials** 5 pairs of colored cups (pink, blue, yellow, green, orange) were used in the conditions. An animal sticker was placed on the inside of one of the cups in each set.

**Procedure** The experimenter began by pulling aside a pair of cups and showing the child that there was a sticker inside one and no sticker inside the other. The experimenter then said, “In this game, it is going to be your job to guess which cup has the sticker inside. For each set of cups I’m going to ask you twice which cup you think has the sticker inside. After you make your first guess I will ask you once more and you can either keep your guess the same or guess the another cup. If your second guess is right then you get a point and for every point you get we will play a game at the end. So remember you want to try and get your second guess right so you can get a point!” The experimenter then proceeded to take a different pair of cups and said, “Let’s take a look at these two cups. One of them has the sticker inside and it’s going to be your job to guess which one. I’m going to look inside so I know which cup has the sticker.” The experimenter then looked inside and then asked the child which cup they believed had the sticker. Regardless of the accuracy of the children’s guess, the experimenter randomly responded either “Yes that’s right!” or “Hmmm that’s not right” and then asked the child again which cup they believed had the sticker inside. Children did not see the contents of the cup immediately after the trials, so they did not receive feedback as to whether their guesses were correct. The second training trial was the same as the first with the exception that the experimenter reversed the response provided after the child’s initial guess.

The experimenter then began the test trials. In the *Knowledgeable* condition the experimenter said, “I’m going to look so I know which cup has the sticker inside” making their knowledge state explicit. She then proceeded to ask the child which cup they believed had the sticker inside; when the child responded, the experimenter provided no explicit feedback as in the training trials, but instead said, “Okay, you said this cup had the sticker inside. Is that your final guess; which cup do you think has the sticker inside?” Children did not see the location of the sticker. In the *Ignorant* condition the experimenter said, “I’m not going to look inside so I don’t know which cup has the sticker either” making their ignorance explicit; the rest of the condition proceeded as with the *Knowledgeable* condition. At the end of the experiment, the experimenter brought back all the pairs of cups and let the children discover which cups contained the stickers.

## Results

**Coding** Children’s responses were video taped and recoded by an assistant blind to condition; seven children were coded live because either no video consent was provided by the parents or because the view of the children’s pointing response was obstructed. For the remaining 25 children, there was 92% agreement; the errors were caused by obvious Left/Right coding errors and were resolved by a third coder.

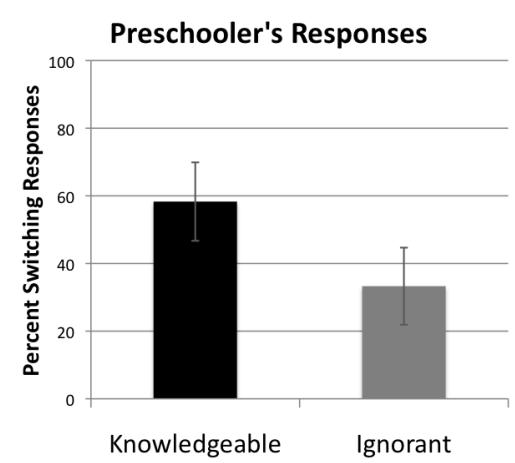


Figure 2: The percent of total test trials on which preschoolers switched their guesses in the *Knowledgeable* and *Ignorant* conditions following the neutral query.

**Inclusion Criteria** Because the order of “correct” and “incorrect” feedback was randomly assigned and counterbalanced, and because the experimenter could not control which cup the children would initially guess, some children were unintentionally “trained” to believe that the sticker always was located in the cup to one side. We classify two scenarios in which this side-bias occurs as follows: (1) During the first training trial the participant selects side A and is told that he is incorrect so he switches his guess to side B. On the second training trial, the participant chooses side B again and is told he is correct and so he continues to select side B for all subsequent guesses, both before and after the query. (2) During the first training trial the participant selects side A and is told he is correct and stays with side A. On the second training trial, the participant then chooses side B and is told he is incorrect, which leads him to switch his guess to side A and continue to select side A without switching on all subsequent guesses. In these two cases we have (by virtue of our random design) given the participant evidence that the sticker will always be on the same side. As discussed above nine children were dropped and replaced for this reason.

**Test Results.** Children’s performance on the initial feedback trials was nearly perfect, with children appropriately switching responses following the feedback that they were incorrect on training trials and appropriately staying with their response following the feedback that they were correct (94%). Our measure of interest was whether children switched their predictions in the test trials following the neutral query of “Is that your final answer; which cup has the sticker inside?” On aggregate, there were no differences between the three test trials within condition (*knowledgeable* Fisher Exact  $p = 1$ ; *Ignorant* Fisher Exact  $p = 1$ ), so the switching response of each child was scored as a total of switching responses by condition; children were more likely to switch their responses in the *Knowledgeable* condition (58% of the trials) than were

children in the *Ignorant* condition (33% of the trials; Pearson  $\chi^2 = 6.04, p = .01$ , see Figure 2). Children in the *Ignorant* condition were less likely to switch responses than would be predicted by chance (Binomial,  $p = .01$ ). No age effects were observed within or across conditions.<sup>1</sup>

Taken together, these results show that children take the knowledge state of the questioner into account; they are more likely to switch responses when the questioner is knowledgeable than when she is ignorant. These results are consistent with the *positive bias* model, suggesting that in addition to taking the knowledge state of the informant into account, preschoolers also have the assumption that the informant prefers to provide positive (over neutral) feedback following correct guess outcomes and a preference to provide a neutral query (over negative) following incorrect guess outcomes.

## General Discussion

Recent research investigating children's learning of causal relationships and inferences about knowledgability suggests that while children's behavior is captured by rational models, in certain cases children's behavior appears less sensible, almost capricious. We have focused on children's seemingly irrational tendency to switch their responses when subjected to repeated questioning, proposing that this behavior may be the rational consequence of reasoning about the questioner's knowledge.

In our example, we presented a relatively simple model that accounts for switching behavior by considering the relationship between the kinds of responses informants may give and the relationship of those responses to the true state of the world. The main prediction of the model is that children's tendency to switch in response to questioning should depend on the epistemic state of the questioner. We presented a simple experiment which showed that preschool children switch their guesses more in response to neutral queries by a knowledgeable informant than by neutral queries by an ignorant informant.

Given that we told participants that they would be asked to guess twice, one might wonder whether the neutral queries should provide any information. By virtue of our design, we made it clear to children that the experimenter could also provide positive ("that's right") or negative ("that's wrong") feedback to children. Thus, by instead asking "is that your final answer" on test trials, children could reason about the implications of *not* hearing one of the alternatives. By providing neutral feedback when other feedback was available, the experimenter provided implicit information to the learner.

Our results leave open the question of why children should switch their hypotheses at all in the *Ignorant* condition. The switching pattern observed in the *Ignorant* condition suggests that social inferences about informant's intentions are not the

only explanation for children's switching behavior. One possibility is that regardless of the epistemic state of the questioner, the opportunity to re-evaluate an answer leads children who have a degree of uncertainty about their answers to reconsider their response. If guessing the outcome of a coin toss, there is no necessary reason to guess heads repeatedly. Similarly, in the ignorant condition, children's switching behavior may simply reflect their uncertainty about the correct answer. Recent research suggests that questioning may simply present the opportunity to reconsider hypotheses (Abbot & Griffiths, 2011), which can lead to a new response; deciding to stay or switch are both rationally viable (Denison et al., 2010; Bonawitz, Denison, et al., 2011).

The model highlights the kinds of assumptions a learner must bring to the table. Our empirical results suggest that preschool children must (1) take knowledge state of informant into account, and (2) assume a positive bias in informants. A large and growing body of evidence suggests that children keep track of others' knowledge and use it for various reasoning problems (e.g. Corriveau et al., 2009; Corriveau & Harris, 2009; Koenig & Harris, 2005). To our knowledge, the argument that children assume this positive bias in informants is novel and has interesting implications for learning and cognitive development. For instance, one might reasonably ask whether this bias is a product of experience and whether such experiences would affect the interpretation of neutral queries.

Our model makes additional predictions that may be explored in future work. For example, training children that the questioner has negative biases should lead to a decrease in switching responses. Switching behavior should also be dependent on the degree to which the informant has information about the actual location; neutral queries from an informant with partial knowledge may lead children to switch responses less often than an informant with complete knowledge. Another possibility is to explore the model's predictions over time to see how children learn about the intentions and knowledge of the informant over a series of trials. Finally, older children and adults may bring different assumptions about informants, so finding age-dependent changes may help characterize how social-causal inferences change with development.

Our results raise important practical considerations in science and education. Often times researchers, teachers, and parents ask children questions as a means of evaluating what the child believes. These questions are assumed to be inert, providing no additional information to the child, and the responses are treated as windows into the child's thinking. Our results call both of these assumptions into question. Even seemingly neutral queries provide information to children when posed by someone who can reasonably be assumed to be knowledgeable. And, because they are savvy social operators, children may change their minds in response to these neutral queries. In our experiments, we intentionally chose situations in which children would have weak beliefs about the location of the sticker, because previous research has sug-

<sup>1</sup>Comparing the average number of switches per child across young and old (by median split), we find no differences (Knowledgeable:  $t(14) = -.35, p = .73$ ; Ignorant:  $t(14) = 0, p = 1$ , Aggregate:  $t(30) = -.25, p = .80$

gested that children's prior beliefs play an important role in how they interpret ambiguous evidence (Schulz et al., 2007) and how they account for evidence that conflicts with strongly held beliefs (Bonawitz et al., 2012). We do not know the degree to which children's prior beliefs will interact with inferences in these settings; However, regardless of prior beliefs, neutral queries may act as prod for the child to reconsider what she previously believed and thus may not provide a true window into the child's thinking. Thus, children's responses following repeated questioning in research and education settings should be interpreted with caution.

## Conclusions

Children use various assumptions about a teacher's knowledge and intent to guide their reasoning. Our work takes this idea and expands on it, suggesting that questions originally assumed to be inert and provide no feedback may in fact serve as cues for children to draw inferences from. Learning about the world is an immensely daunting task for young minds, yet children are able to learn rapidly and accurately even in light of limited information. In school settings, at home, and even in eyewitness testimony, we employ repeated questioning as a means to assess a child's beliefs. When asked by a knowledgeable informant, such as a teacher, parent, or attorney, these questions may not simply elicit information about what the child believes, but instead may give the child reason to change their beliefs all together.

**Acknowledgements:** Thanks to Sophie Bridgers and Jaclyn Harris for data collection, to Swe Tun, Alvin Chan, and Sonia Spindt for coding, and to participating daycares, parents and children. This research was additionally supported by grant IIS-0845410 from the National Science Foundation and the James S. McDonnell Foundation.

## References

- Abbot, J., & Griffiths, T. L. (2011). Exploring the influence of particle filter parameters on order effects in causal learning. In *Proceedings of the 33rd annual conference of the cognitive science society*.
- Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. (2011). A simple sequential algorithm for approximating bayesian inference. In *Proceedings of the 33rd annual conference of the cognitive science society*.
- Bonawitz, E., Fisher, A., & Schulz, L. (in press). Teaching three-and-a-half year olds to reason from ambiguous evidence. *Journal of Cognition and Development*.
- Bonawitz, E., Schijndel, T. van, Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, 64(4), 215-234.
- Bonawitz, E., Shafto, P., Gweon, H., Goodman, N., Spelke, E., & Schulz, L. (2011). The double-edge sword of pedagogy: Teaching limits children's spontaneous exploration and discovery. *Cognition*.
- Buchsbaum, D., Gopnik, A., Griffiths, T., & Shafto, P. (2011). Childrens imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, 120(3), 331 - 340.
- Butler, L., & Markman, E. (2010). Pedagogical cues influence children's inductive inference and exploratory play. In *Proceedings of the 32nd annual conference of the cognitive science society*.
- Corriveau, K. H., Fusaro, M., & Harris, P. L. (2009). Going with the flow: Preschoolers prefer non-dissenters as informants. *Psychological Science*, 20, 372-377.
- Corriveau, K. H., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and past accuracy. *Developmental Science*, 12, 426-437.
- Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. (2010). Preschoolers rationally sample hypotheses. In *Proceedings of the 32nd annual conference of the cognitive science society*.
- Gopnik, A., Glymour, C., Sobel, D., Schulz, L., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111, 1-31.
- Howie, P., Sheehan, M., Mojarrad, T., & Wrzesinska, M. (2004). Undesirable and desirable shifts in children's responses to repeated questions: age differences in the effect of providing a rationale for repetition. *Applied Cognitive Psychology*, 18(9), 1161 - 1180.
- Koenig, M., & Harris, P. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, 76, 1261-1277.
- Krahenbuhl, S., Blades, M., & Eiser, C. (2009). The effect of repeated questioning on children's accuracy and consistency in eyewitness testimony. *Legal and Criminological Psychology*, 14(2), 263-278.
- Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, 112, 367-380.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.
- Poole, D., & White, L. (1991). Effects of question repetition on the eyewitness testimony of children and adults. *Developmental Psychology*, 27(6), 975 - 986.
- Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared make your tummy ache? naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*, 43, 1124-1139.
- Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (in press). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*.
- Shafto, P., & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for pedagogical situations. In *Proceedings of the 30th annual conference of the cognitive science society*.
- Spirtes, P., Glymour, C., & Schienens, R. (1993). *Causation prediction and search*. New York: Springer-Verlag.

# Abstract language comprehension is incrementally modulated by non-referential spatial information: evidence from eye-tracking

Ernesto Guerra (ernesto.guerra@uni-bielefeld.de)

Pia Knoeferle (knoeferl@cit-ec.uni-bielefeld.de)

Cognitive Interaction Technology Excellence Cluster, Bielefeld University,  
Morgenbreede 39, 33615, Bielefeld, Germany.

## Abstract

Research on situated language processing has examined how visually depicted objects or concrete action events inform the comprehension of concrete sentences. By contrast, much less is known about how *abstract* sentence comprehension interacts with non-linguistic visual information. Moreover, while non-linguistic information can rapidly inform language comprehension when it is related to sentence content through reference or lexical-semantic associations, it is unclear to which extent this is the case when the visual context is ‘non-referential’ (i.e., not related to the sentence through reference or lexical semantic associations). We conducted two eye-tracking reading experiments to address these two open issues. In both experiments, reading times were shorter when sentences about conceptually similar abstract ideas were preceded by objects (words-on-cards in Experiment 1 and blank playing cards in Experiment 2) that were depicted close together (vs. far apart); and when sentences about conceptually dissimilar abstract ideas were preceded by objects that were depicted far apart (vs. close together). This happened rapidly (first-pass reading times) and incrementally (as the sentence unfolded). Thus, (a) comprehension of abstract language can be modulated by non-linguistic visual information (spatial distance between depicted objects) at the sentence level, and (b) online language comprehension can be informed by visual context even in the absence of an overt referential or lexical-semantic link.

**Keywords:** semantic interpretation; spatial information; non-referential visual context; eye tracking.

## Introduction

Studies in the ‘visual world paradigm’ have contributed extensively to our understanding of how non-linguistic visual information affects sentence comprehension (e.g., syntactic disambiguation: Tanenhaus et al., 1995; semantic interpretation: Sedivy et al., 1999). In ‘visual world studies’, listener’s eye movements are tracked during comprehension of a spoken sentence that describes a given visual environment. Findings from such studies have shown that visual presentation of objects or concrete action events can facilitate incremental structural disambiguation (e.g., Tanenhaus et al., 1995; Knoeferle, Crocker, Scheepers, & Pickering, 2005); that language can rapidly guide visual attention to semantically relevant objects as evidenced by anticipatory eye-movements (e.g., Altmann & Kamide 1999; Kamide, Scheepers, & Altmann, 2003; Kamide, Altmann, & Haywood, 2003); and that distractor objects are inspected more often when they are semantically related (vs. unrelated) to a target word (e.g., Huettig & Altmann, 2005,

2011; Huettig & McQueen, 2007). Visual context not only affects spoken language comprehension rapidly, but also sentence comprehension during reading. Evidence from picture-sentence verification has revealed rapid visual context effects for concrete visual stimuli (e.g., red dots) and sentence content (e.g., *The dots are red*, see Clark & Chase, 1972; also Gough, 1965; Knoeferle, Urbach, & Kutas, 2011; Underwood, Jebbet, & Roberts, 2004).

However, most of these studies have concentrated on sentences about concrete objects and events. While evidence suggests that visual context can rapidly and incrementally inform comprehension of concrete spoken and written sentences, it is unclear to which extent non-linguistic visual context information can influence the processing of abstract language rapidly and incrementally. In examining situated language comprehension, most visual world studies have further relied on a *referential* linking hypothesis (e.g., a noun referencing an object or a verb an action). By contrast, it’s unclear whether visually presented information can influence sentence comprehension when there is no overt referential or lexical-semantic link with sentence content.

## Spatial Distance and Semantic Similarity

Conceptual metaphor theory proposes that abstract meaning is grounded in physical experience through metaphorical mapping (Lakoff & Johnson, 1999). Similarity, for instance, would be grounded in the physical experience of spatial distance. Recent behavioral studies have provided first evidence for a link between spatial distance and similarity. In one study, two visually presented abstract words (e.g., *loyalty* and *boredom*) were judged to be more similar when they were presented close together (vs. far apart), but more dissimilar when they were presented far apart (vs. close together, Casasanto, 2008). In another, similarity-judgment task (on whether two squares on a screen had similar colors or not) speeded decision times were shorter when similarly-colored squares were presented close to each other (vs. far apart), and when differently-colored squares were presented far apart (vs. close to each other, Boot & Pecher, 2010). These rating and response time effects support the view that there is a relationship of some sort between spatial information (the distance between two stimuli) and semantic and visual similarity.



## Accounting for situated language comprehension

The nature and time course of spatial distance effects on cognitive, and in particular, language comprehension processes, however, remains unclear. In summary, we have identified several open issues in research on the interaction between non-linguistic visual information and language comprehension, which we conceptualize as two research questions: (1) Can non-linguistic information rapidly and incrementally modulate the semantic interpretation of abstract sentences? (2) Can non-linguistic visual information modulate language comprehension even in the absence of referential or lexical-semantic links?

Addressing these and other questions is important to advance accounts of situated language comprehension (e.g., the Coordinated Interplay Account, CIA, Knoeferle & Crocker, 2006; Knoeferle & Crocker, 2007). The CIA accommodates visual context effects during spoken language comprehension. It consists of three informationally and temporally dependent stages. A first stage accommodates the processes of incremental sentence comprehension that are the focus of traditional sentence processing accounts. A second stage describes utterance-mediated shifts in (visual) attention. ‘Scene integration’, finally, integrates the linguistic and scene input and informs interpretation based on visual representations. The CIA makes no assumptions regarding the modular status of either the linguistic or visual processes involved. Rather, it outlines the interaction of utterance interpretation, (visual) attention and scene information.

The CIA has been derived from eye-tracking findings on the comprehension of concrete sentences in non-linguistic visual contexts. The rapid and incremental time course with which information in visual context interacts with spoken language comprehension appears to generalize to reading when there is a referential link between visual context and sentence meaning (see Knoeferle et al., 2011 for evidence). Knowing whether rapid and incremental visual context effects are also observed when there is no referential link and when sentences are abstract, would be important for extending the account and refining its language-context linking mechanism. Based on the close time locking of visual context effects and language comprehension in the CIA, we would expect to see spatial distance effects emerge time-locked to when information about semantic similarity becomes available in the sentence. The present research addressed the two research questions (1) and (2) in two eye-tracking reading experiments. The studies examined whether, and if so with which time course, spatial distance effects on semantic similarity processing occur during comprehension of abstract sentences.

### Experiment 1

In Experiment 1, participants inspected a visual context that depicted words on cards either close together or far apart (Fig. 1). Then they read a sentence that was either about

similarity or dissimilarity between abstract nouns (Table 1). After reading the sentence and judging its veracity, they saw a picture and verified whether it was the same as the one that they had inspected before the reading task.

If spatial distance between the cards can modulate the interpretation of semantic similarity, we should see this reflected in reading times. To the extent that the existing findings on spatial distance effects (Boot & Pecher, 2010; Casasanto, 2008) generalize to language comprehension we should see faster reading times for similarity-conveying sentences when the preceding words-on-cards are close together (vs. far apart), and for dissimilarity-conveying sentences when the words-on-cards are far apart (vs. close together). Moreover, if effects of spatial distance are incremental, we should see them at the adjective region of the sentence (see Table 1 for examples) since this is when similarity relations are made explicit and could thus be related to spatial distance from the recent visual context. In principle, effects could appear even earlier, namely at the second noun phrase, since semantic similarity could become available as soon as the two abstract nouns are integrated. Finally, to the extent that these effects are immediate, we should observe them in first-pass reading times.

### Method

**Participants** Thirty-two native speakers of German (mean age: 23.6; range 19-33) with normal or corrected-to-normal vision participated in the experiment for a compensation of 6 Euro. None of them had been exposed to a second language before age 6, and all gave informed consent.

**Materials and Design** We created 48 sentences<sup>1</sup>, each of which had two versions. In one version the sentence expressed *similarity* between two abstract nouns, and in the other version it expressed *dissimilarity* between two abstract nouns (see Table 1 for examples). In addition, we created visual contexts using commercial graphics programs. The visual context showed two playing cards each of which presented an abstract word (see Fig. 1). Card depictions did not change between items but the words on the cards did. The words on the cards always appeared as the first and second noun phrase in the sentence. The two visual contexts and the two sentences made up an item.

A 2x2 within-subjects Latin square experimental design was implemented with two factors (spatial distance and semantic similarity), each with two levels (close vs. far, similar vs. different, respectively). Combinations of the two factors and levels resulted in four experimental conditions: cards far apart vs. close together with a similarity-conveying sentence; and cards far apart vs. close together with a dissimilarity-conveying sentence (see Table 1).

We constructed 96 filler sentences. All of them were grammatical and semantically legal German sentences. However, 72 of them described unrealistic situations (e.g.,

---

<sup>1</sup>One item was removed due to an error in the order of presentation of words.

‘a presentation without good rhetoric should be given more often’), and the other 24 described plausible situations (e.g., ‘on the tram, passengers show their ticket to the inspector’). All filler sentences were preceded by cards in different positions on the screen (e.g., in the upper left and lower right corner) and most of them were blank (N=72). Twenty-four filler sentences, however, had cards with words on them. There were in addition 14 practice trials. A list consisted of 144 trials (48 experimental and 96 fillers trials), which were all pseudo-randomized in four lists. Each list contained only one version of every item. There was at least one filler trial in between two items.

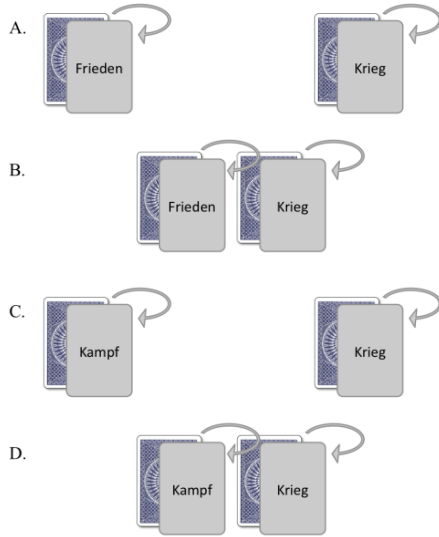


Figure 1: Example visual contexts for the sentences in Table 1 (Experiment 1). The cards moved from the center of the screen either far apart (A, C) or close to each other (B, D). After two seconds, cards turned around (as represented by the semi-circular arrow) and presented two abstract words. The words were either semantically dissimilar (e.g., *Frieden* [‘Peace’] and *Krieg* [‘War’], (A) and (B)) or similar (e.g., *Kampf* [‘Battle’] and *Krieg* [‘War’], (C) and (D)).

**Procedure** Upon arrival in the laboratory, participants received information about the study. After that they were calibrated using a 9-point calibration procedure. Then they completed 14 practice trials. After practice, the experiment began. Each trial was presented as a three-step task. First, participants saw a visual context for six seconds, with two playing cards in different positions. For half of the trials in the experiment (i.e., 24 filler and all experimental trials), cards turned around after two seconds, each showing a word for four seconds. For the other half of the trials, cards turned and showed a blank front for 500 ms. After inspecting the visual context participants read a sentence (see Table 1). They were instructed to try to understand the sentence and to judge its veracity (by pressing either a “yes” or a “no” button). For the ensuing picture verification, participants saw a picture of two cards, and verified (by pressing “yes”/“no” buttons) whether they were identical with the two cards they had seen before the sentence.

Table 1: Example of visual contexts, corresponding sentences, and the resulting condition.

Picture	Sentence type	Condition
Fig.1A	(1) <i>Frieden</i> <sub>NP1</sub> und <sub>coord</sub> <i>Krieg</i> <sub>NP2</sub> <i>sind</i> <sub>VP1</sub> <i>bestimmt</i> <sub>ADV</sub> <i>verschieden</i> <sub>ADJ</sub> .	Far-Dissimilar
Fig.1B	<i>das verriet</i> <sub>VP2</sub> <i>der Anthropologe</i> <sub>NP3</sub> .	Close-Dissimilar
Fig.1C	(2) <i>Kampf</i> <sub>NP1</sub> und <sub>coord</sub> <i>Krieg</i> <sub>NP2</sub> <i>sind</i> <sub>VP1</sub> <i>freilich</i> <sub>ADV</sub> <i>entsprechend</i> <sub>ADJ</sub> .	Far-Similar
Fig.1D	<i>das verriet</i> <sub>VP2</sub> <i>der Anthropologe</i> <sub>NP3</sub> .	Close-Similar

Translation: (1) ‘Peace<sub>NP1</sub> and<sub>coord</sub> war<sub>NP2</sub> are<sub>VP1</sub> certainly<sub>ADV</sub> different<sub>ADJ</sub>, suggested<sub>VP2</sub> the anthropologist<sub>NP3</sub>.’ (2) ‘Battle<sub>NP1</sub> and<sub>coord</sub> war<sub>NP2</sub> are<sub>VP1</sub> surely<sub>ADV</sub> similar<sub>ADJ</sub>, suggested<sub>VP2</sub> the anthropologist<sub>NP3</sub>.’

**Data Analysis** Log-transformed reading times were analyzed using a linear mixed effect regression (LMER), including in a single step main and interaction effects of the factors. We implemented full models with random intercepts for participants and items, and fixed effect random slopes and their interactions for both random intercepts. We analyzed the second noun phrase (NP2) and the adjective region (ADJ), where we should see spatial distance effects on semantic interpretation if those occur time-locked to when semantic similarity between the first two noun phrases becomes available during reading (see Table 1). We further analyzed the VP2 and NP3 regions to see if any spatial distance effects also occur at subsequent regions of the sentence. In these regions we examined three eye-tracking reading measures; first-pass reading (the duration of all fixations from first entering an interest area and prior to moving to another interest area), regression path duration (the time from first entering a region until moving past that region to the right; unlike first-pass reading time, this measure includes reading time following regressions out of the region), and total reading times (the duration of all fixations in a given region, see, e.g., Rayner, 1998).

## Results

For the ADJ region, a similarity main effect was observed across all measures; reading times were shorter for sentences expressing similarity compared to those expressing dissimilarity (all  $t$ -values  $> 2$ ). We also observed a reliable interaction between spatial distance and semantic similarity in first-pass times for the ADJ region ( $t$ -value = -2.04, see Fig. 2): first-pass times were shorter when similarity-conveying sentences were preceded by cards-with-words presented close to each other (vs. far apart). By contrast, first-pass times for sentences that expressed dissimilarity were shorter when they were preceded by cards-with-words presented far apart (vs. close to each other). No interaction effects were observed in other measures. Both VP2 and NP3 regions showed reliable

interaction effect in first-pass reading times and NP3 also in total reading times. These interaction effects were similar as for the ADJ region. For the NP2 region, neither main effects nor interaction effects involving the manipulated factors were observed in any gaze measure.

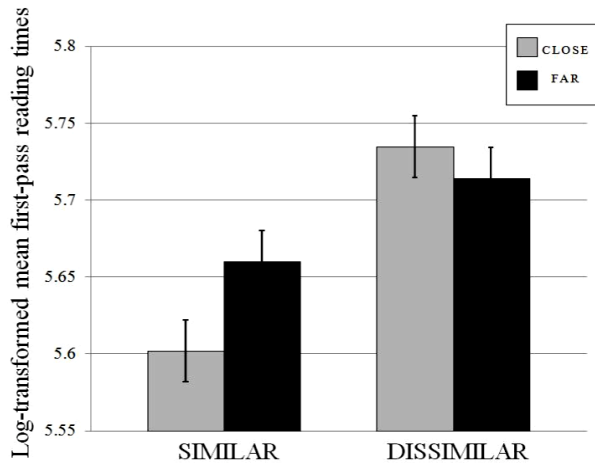


Figure 2: Log-transformed mean first-pass reading time (with error bars plotting the standard error of the mean) for the ADJ region as a function of sentence type and spatial distance between cards-with-words in Experiment 1.

## Discussion

In Experiment 1, we presented participants with a visual context for which the distance between cards-with-words, was manipulated. Cards were either presented close together or far apart, and they were followed by a sentence that either expressed similarity or dissimilarity of abstract nouns. We observed spatial distance effects on reading times as a function of the semantic content of the sentence. These results suggest that non-linguistic information from the visual context (spatial distance) can modulate interpretation of abstract language (semantic similarity) during online sentence comprehension. Crucially, spatial distance effects on semantic interpretation appeared both rapidly (first-pass) and incrementally (at the ADJ region).

Results from Experiment 1, inform our first research question (whether abstract semantic interpretation can be rapidly and incrementally modulated by non-linguistic information). However, since the visual context for critical items was related to the following sentence through words on the cards, it is possible that effects of spatial distance on semantic similarity interpretation were mediated by, or even depended upon, that link. To address this concern and to answer our second research question (can non-linguistic visual information modulate language comprehension even in the absence of referential or lexical-semantic links?), Experiment 2 relies on the same design and presentation but the cards did not show any words and remained blank.

## Experiment 2

Experiment 2 was identical to Experiment 1 but instead of presenting words on cards, the cards that participants saw

before sentence reading were blank. Seeing the noun phrases on the cards in Experiment 1 could permit participants to integrate similarity of the nouns with spatial information even before sentence reading. If so, this could facilitate and speed up any effects of spatial distance during reading. Using blank cards in Experiment 2 permitted us to see to which extent the effects of spatial distance on sentence reading in Experiment 1 depended upon the repetition of sentential noun phrases on the cards.

## Method

**Participants** Thirty-two further native speaker of German with normal or corrected-to-normal vision (mean age: 24.4; range 20-31) participated in the experiment for a compensation of 6 Euro. All gave informed consent.

**Materials, Design, Procedure and Data Analysis** The experimental design, procedure and data analysis were the same as in Experiment 1. The visual context, however, was modified. While participants in Experiment 1 saw cards-with-words for four seconds, participants in Experiment 2 saw blank cards for three seconds. In both experiments, however, visual context presentation duration was the same (six seconds). We delayed card turning by one second in Experiment 2, since participants did not have to read any words.

## Results

At the NP2 region we observed a similarity main effect, such that sentences that expressed similarity had shorter first-pass times, compared to sentences that expressed dissimilarity ( $t$ -values = 2.04). Moreover, analyses of first-pass times confirmed a reliable interaction between spatial distance and semantic similarity ( $t$ -value = -2.07). Figure 3 shows the interaction pattern in Experiment 2.

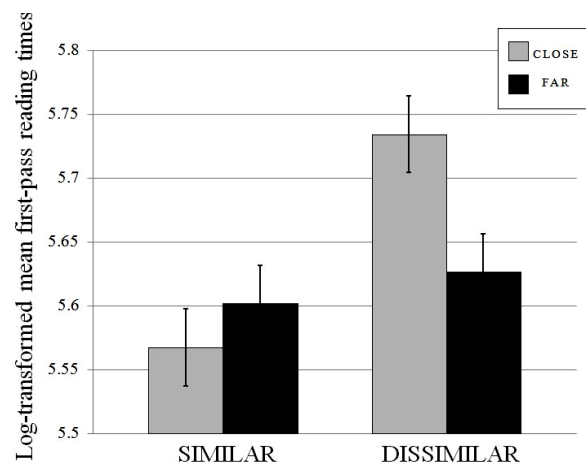


Figure 3: Log-transformed mean first-pass reading times (with error bars plotting the standard error of the mean) for the NP2 region as a function of sentence type and spatial distance between cards in Experiment 2.

First-pass times for sentences that expressed similarity were shorter when preceded by blank cards close to each

other (vs. far apart), while reading times for sentences that expressed dissimilarity were shorter when preceded by blank cards far apart (vs. close to each other). No other interaction effects were observed in other measures in this region.

Unlike in Experiment 1, a main effect of similarity appeared in regression path duration at the ADJ region, but no interaction effects were observed for that region. For the subsequent VP2 region, a marginal interaction effect emerged ( $t$ -value = -1.84), with a similar pattern to that at the NP2 region in Experiment 2 and the ADJ region in Experiment 1. Main effects of spatial distance and similarity were observed in regression path duration for the NP3 region, but no interaction effect was found.

## Discussion

In Experiment 2, we manipulated the distance between objects that did not have any overt relation to the semantic meaning of the ensuing sentence. The visual context can thus be described as non-referential in its relationship to the sentence. Moreover, spatial distance between cards was irrelevant for performing the comprehension task (sentence veracity judgments). Yet, we still found rapid and incremental interaction effects between spatial distance and the interpretation of semantic similarity, as reflected in reading times. These findings provide strong support for the role of non-referential spatial information in abstract language comprehension.

## General Discussion

In research on situated language processing, abstract language comprehension has received substantially less attention than concrete language. While results from visual world paradigms have shown that non-linguistic information can rapidly and incrementally modulate comprehension of sentences that relate to objects and events (through referential or associative links), it was unclear whether this effect would generalize to abstract sentences.

We assessed this open issue in two eye-tracking-reading experiments. Experiment 1 provided evidence that spatial information (distance between words depicted on cards) can rapidly and incrementally influence semantic interpretation of abstract sentences. First-pass times at the adjective and subsequent regions of sentences such as ‘battle and war are surely similar, suggested the anthropologist’ were shorter when a preceding display showed cards-with-words close together (vs. far apart). Since objects in the visual context were still related to the sentence in Experiment 1 (through words on cards), Experiment 2 examined whether the rapid spatial distance effects from Experiment 1 extend to a situation in which the visual context (playing cards) was entirely unrelated to the sentence. When cards remained blank rather than showing words (Experiment 2), we observed rapid and incremental effects of spatial distance on reading times as a function of the meaning of the sentences. First-pass times at the second noun phrase and at the second verb in sentences such as ‘peace and war are certainly

different, suggested the anthropologist’ were shorter when a preceding display showed cards-with-words far apart (vs. close together).

Overall, thus, non-linguistic information can rapidly and incrementally influence semantic interpretation of abstract language. These results extend existing similarity-judgment and response-times results (Boot & Pecher 2010; Casasanto 2008), and clarify that non-linguistic information (i.e., spatial distance) can modulate language comprehension and not just similarity judgments and ratings. In this regard, our results are compatible with theories of embodied cognition that attribute an important role to perceptual information in language processing (see Lakoff & Johnson 1999). They also extend findings that suggest abstract language can be related to visual context through lexical-semantic associations. Duñabeitia et al. (2009) showed that listeners rapidly inspected objects (e.g., a nose) that were associated with abstract words (e.g., Spanish *olor*, ‘smell’). What the present findings add is the insight that visual context can in turn influence abstract language comprehension, and that its effects are mediated via subtle mappings between the semantic similarity of nouns and object distance.

The present findings also have important implications for processing accounts of situated language (e.g., the CIA, Knoeferle & Crocker, 2006, 2007). Consider how concrete language is related to visual context in the CIA. The emerging interpretation guides (visual) attention to relevant information in visual context or working memory. As people process a sentential verb, they engage in a search for a matching action either in the immediate environment or in their working memory. When they have found a matching action, verb and action are co-indexed and action information can inform sentence interpretation. Comprehenders can further develop expectations about referents based on lexical-semantic associations between the verb and objects in context.

The present findings corroborate and extend the referential and associative linking mechanism of the Coordinated Interplay Account. First, the spatial distance effects occurred not anywhere during reading but they first emerged at sentence regions that contained information about semantic similarity in both experiments. This was as expected based on the CIA’s close time lock between when the utterance identifies relevant visual context information, and when that context information impacts sentence interpretation. What the present results add is the insight that this closely temporally coordinated interplay extends to non-referential visual context effects (Experiment 2) and abstract language comprehension (Experiments 1 and 2). Future research will examine the subtle differences in the time course of spatial-distance effects in Experiment 1 (ADJ, NP2) compared with Experiment 2 (NP2, VP2).

The findings moreover emphasize the necessity of assuming a fine-grained linking of linguistic and visual information during language comprehension. As semantic similarity is computed during comprehension, it is reconciled with representations of spatial distance between

objects that were neither mentioned in the sentence nor relevant for the sentence judgment task. This ties in with other recent results. Kreysa and Knoeferle (2011) observed rapid visual context effects (of a speaker's gaze and head movements) on spoken language comprehension, and this despite the fact that the speaker was never explicitly referenced to and that comprehenders hardly inspected the speaker during comprehension. Clearly, explicit reference is not necessary for incremental visual context effects.

Together the present and other recent findings argue for highly active visual context effects on language comprehension and highlight the need for multiple (referential and non-referential) mechanisms in informing language comprehension through non-linguistic visual information.

### Acknowledgments

This research was funded by the Cognitive Interaction Technology Excellence Cluster (German research foundation, DFG) and by a PhD scholarship awarded to EG by the Ministry of Education, Government of Chile. The authors want to thank Maria Nella Carminati for advice with the analyses, and the members of the Language and Cognition Lab (CITEC, University of Bielefeld) for their valuable comments on the reported research.

### References

- Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73, 247-264.
- Boot, I. and Pecher, D. (2010). Similarity is closeness: metaphorical mapping in a conceptual task. *Quarterly Journal of Experimental Psychology*, 63, 942-954.
- Casasanto, D. (2008) Similarity and proximity: When does close in space mean close in mind. *Memory and Cognition*, 36, 1047-1056.
- Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, 3, 472-517.
- Duñabeitia, J. A., Avilés, A., Afonso, O., Scheepers, C., & Carreiras, M. (2009). Qualitative differences in the representation of abstract versus concrete words: evidence from the visual-world paradigm. *Cognition*, 110, 284-292.
- Gough, P. B. (1965). Grammatical transformations and speed of understanding. *Journal of Verbal Learning & Verbal Behavior*, 4, 107-111.
- Huetting, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition*, 96, B23-32.
- Huetting, F., & Altmann, G. T. M. (2011). Looking at anything that is green when hearing "frog": how object surface colour and stored object colour knowledge influence language-mediated overt attention. *Quarterly Journal of Experimental Psychology*, 64, 122-145.
- Huetting, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic, and shape information in language-mediated visual search. *Journal of Memory and Language*, 54, 460-482.
- Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49, 133-156.
- Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, 32, 37-55.
- Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance, and world knowledge: evidence from eye tracking. *Cognitive Science*, 30, 481-529.
- Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language*, 57, 519-543.
- Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events. *Cognition*, 95, 95-127.
- Knoeferle, P., Urbach, T. P., & Kutas, M. (2011). Comprehending how visual context influences incremental sentence processing: Insights from ERPs and picture-sentence verification. *Psychophysiology*, 48, 495-506.
- Kreysa, H., & Knoeferle, P. (2011). Effects of speaker gaze on spoken language comprehension: Task matters. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1557-62). Austin, TX: Cognitive Science Society.
- Lakoff, G., & Johnson, M. (1999) *Philosophy in the flesh: The embodied mind and its challenge to western thought*. University of Chicago Press.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372-422.
- Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71, 109-147.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632-1634.
- Underwood, G., Jebbett, L., & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search. *The Quarterly Journal of Experimental Psychology*, 56, 165-182.

# Trading in a multiplayer board game: Towards an analysis of non-cooperative dialogue

Markus Guhe (m.guhe@ed.ac.uk), Alex Lascarides (alex@inf.ed.ac.uk)

School of Informatics, Informatics Forum, 10 Crichton St, Edinburgh EH1 2HX, Scotland

## Abstract

We collected and analysed a small dialogue corpus of people playing *The Settlers of Catan*. Dialogues are trading negotiations where Gricean maxims of cooperativity often break down as players adopt conflicting intentions in their attempt to win the game. This has consequences for what information players are sharing and for the sincerity of their contributions.

In this paper, we motivate and describe a two-level scheme for analysing non-cooperative dialogues, where both levels are interdependent. Each dialogue move is a move in the game (e.g., an offer to trade), and a coherent contribution to the dialogue so far, connected to a prior segment with a coherence relation, such as indirect answerhood or rejection. Parsing and generating coherence relations is computationally feasible (e.g., Baldridge & Lascarides, 2005), and here we'll argue that their semantics help to identify the game move, even when it is implicated rather than linguistically explicit.

**Keywords:** non-cooperative dialogue; dialogue move; negotiation; *The Settlers of Catan*; JSettlers; multiplayer game.

## Non-cooperative dialogue

Standard accounts of dialogue usually assume that the agents cooperate with one another on at least two levels (Grice, 1975): they coordinate the conventions that govern linguistic meaning (basic cooperativity); and they share attitudes towards what is said, including shared intentions (content cooperativity). Tutoring is a widely investigated example where content cooperativity is a reasonable assumption.

But there are many dialogue scenarios where conflicting preferences compel agents to perform dialogue moves that aren't content cooperative. Such situations range from differences in preferences with relatively low stakes (e.g., whether to go to a French or an Italian restaurant) to very high, life changing, stakes (e.g., a cross examination of a defendant by a prosecutor in court). Lying in court has potential high rewards (if the lie is not uncovered) as well as high penalties (if it is). But telling the truth or lying are just the two extremes in a spectrum of non-cooperative behaviour. To illustrate this, consider the following cross-examination of a defendant by a prosecutor (from Solan & Tiersma, 2005):

- (a) Prosecutor: Do you have any bank accounts in Swiss banks, Mr. Bronston?
- (b) Bronston: No, sir.
- (c) Prosecutor: Have you ever?
- (d) Bronston: The company had an account there for about six months, in Zurich.

The locutionary content of both (b) and (d) are true, but Bronston succeeds in *deflecting* the prosecutor's enquiry with a misleading implicature with (d): (d) implicates that Bronston never had any Swiss bank account, and this is false.

In this paper, we describe dialogue moves in non-cooperative situations, drawing on data from a dialogue cor-

pus of people playing *The Settlers of Catan*. In the corpus, each dialogue move is aligned with a machine readable version of the game state. We present a scheme of the types of dialogue moves that occur in our corpus. Our eventual goal is to build a dialogue system that plays *The Settlers of Catan*, engaging in non-cooperative dialogues like humans do. For such a system, the agent playing the game will use the dialogue moves from that scheme to interact with the other players so as to forward its chosen game moves. This will require us to parse and generate the semantic content of these dialogue moves; we will exploit coherence relations from SDRT for this purpose (Asher & Lascarides, 2003).

## The Settlers of Catan

*The Settlers of Catan* is a multiplayer board game set on the fictional island *Catan*. *Catan* is represented by a map consisting of hexes (see Fig. 1; see [catan.com](http://catan.com) for the full set of rules). Two to four players settle on the island by building settlements and cities connected by roads. It is a zero-sum game: the first player to win 10 victory points wins and all others lose. One obtains victory points by, for example, building a settlement (1 point) or a city (2 points). To build, one needs resources: clay, wood, rock, ore and sheep.

Players take turns and attempt to obtain resources and to build. A turn starts with the roll of dice. Each player potentially obtains or loses resources through dice rolls. This depends on a combination of: the number rolled, the location of game pieces on the board, and the resources currently held. (So, future game states are non-deterministic, necessitating players to calculate the risks of moves). The player whose turn it is can trade resources with the bank or other players and can use resources to build roads, settlements or cities.

Dialogues where the players trade resources surface during game play as the players try to exchange resources with one another. Their decisions about what resources they want and what they are prepared to give up are influenced by what they need to build, e.g. a road requires 1 clay and 1 wood. A player can only build on a location if it touches one of his already existing pieces (road, settlement or city); all settlements and cities must be separated by at least 2 roads (of any player). Thus, players' decisions about trades are not only determined by the decisions about what they want to build and where they want to build it, but also by their estimates of what will most advance, or undermine, the building strategies of the other players (see Thomas, 2003). Players can agree any trade; they can even lie or bluff, as found in classic conceptions of bargaining games (Osborne & Rubinstein, 1990).



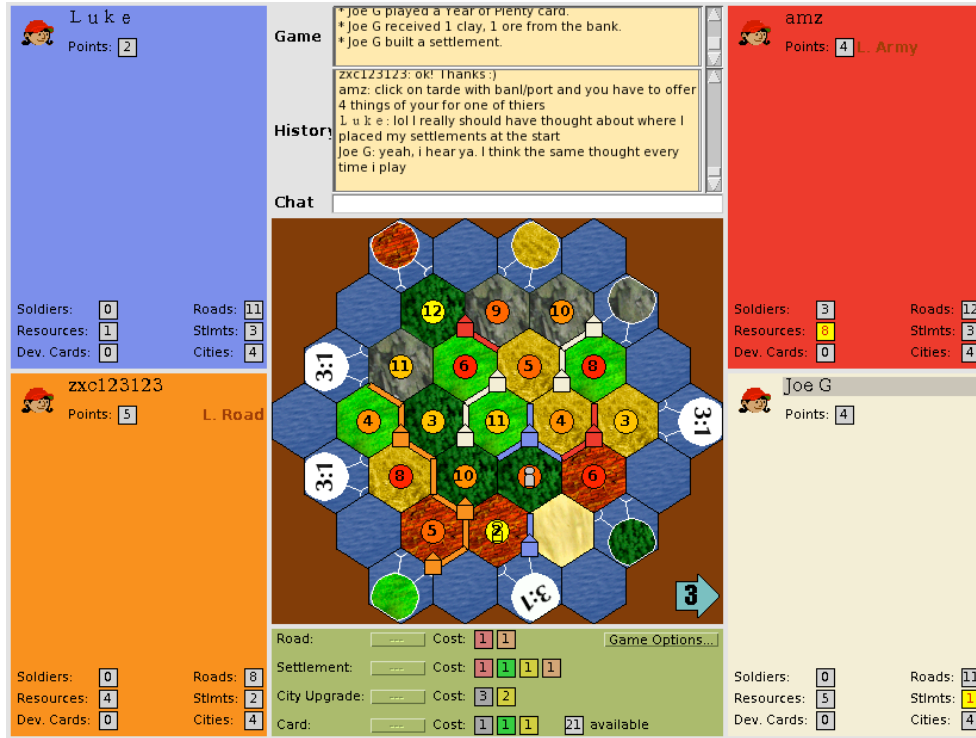


Figure 1: A game of *The Settlers of Catan* in the JSettlers interface.

## Modifying JSettlers for Corpus Collection

To analyse trading dialogues and to perform inferences about an agent's underlying game strategy, one needs a corpus where each utterance is aligned with a machine readable version of its synchronous game state. To this end, we adapted an existing open source implementation of Settlers, cf. Fig. 1 (JSettlers2, version 1.1.12). JSettlers mimics the physical version of the game in that each player can see the whole board but only his own hand (in a colour-coded panel). In JSettlers, trades are done graphically via the interface, see Fig. 2a. The player whose turn it is has two panels containing 'I Give', 'I Get' and colour coded buttons for the other players – he can offer a trade by adding appropriate values for these three items. This prompts the players in question, who can then accept or reject the trade through button clicking.

We adapted this interface to enforce trades to be negotiated linguistically prior to an actual exchange of resources. Specifically, the buttons required to exchange resources are hidden by a 'Register a Trade' button (Fig. 2b), and only one other player can be selected via the colour coded buttons. Players were instructed to first agree a trade via the chat interface and only then 'register' the trade. Clicking 'Register a Trade' revealed the trading interface (Fig. 2a), and enacting the trade is achieved in the same way as the original version of JSettlers.

## Negotiation dialogues

In contrast to other task-oriented dialogues, e.g., the map task (Anderson et al., 1991; Guhe & Bard, 2008), *The Settlers*



Figure 2: (a) A player's hand panel in JSettlers. (b) The modified hand panel during normal game play. Trade is agreed via the chat interface prior to using the panel to register the trade.

of *Catan* is not a cooperative game but a zero sum game: only one player can win, all others lose. Even so, it is sometimes necessary for players to form alliances to block another player from winning, e.g., *don't give tomm wheat, he can win the game if he gets wheat* (from dialogue 1). Because of this, players avoid dialogue moves that might undermine their trustworthiness; they like to *appear* to be cooperative even if they aren't actually cooperating at the content level (i.e., adopting a shared intention), e.g., they may reject a trade even when they are able to perform it by saying *no sorry* (when asked *can I buy wood for 1 ore?* and the player having wood; from dialogue 2). Joint action takes place on multiple levels



simultaneously (Clark, 1996), and we find cooperativity on many levels in our corpus, e.g. all players share a common language, they follow the rules of the game, and they do actually trade with each other.

To analyse trade interactions, it is not necessary to treat the dialogue exchanges within a single game as a single dialogue. Instead, they appear as chunks, especially since the game state changes between trades (thanks to dice rolls, for instance), leading the players to adopt different preferences about what resources they need at the start of a new *dialogue game*, as compared to their preferences at the end of the last one. In addition, the game is too complex for players to reason about the complete game tree anyway; so, they structure their game space, aiming to achieve shorter term goals by using heuristics about how to advance their goal of winning (Thomas, 2003). Because of this structuring of the game space, it is natural to analyse the dialogues in discrete chunks, each addressing a short term goal, e.g. to obtain wood. We adopted this subdivision into dialogue games from the analysis of map task dialogues (Kowtko, Isard, & Doherty, 1993).

We collected an exploratory pilot corpus of 21 games, each containing multiple dialogue games, which we will make available publicly in due course. All participants were native speakers of English and naive to the purpose of the study. Most were students at the University of Edinburgh.

The corpus is not currently annotated because the purpose of collecting it was to obtain a sufficient sample so as to inform decisions about an annotation scheme for the corpus collection proper, which we are engaged in at the moment. For this reason, we have not analysed this pilot corpus in quantitative detail; e.g. with respect to differences between individual players, linguistic features, strategies, or level of cooperativity in the dialogue moves. Instead, we report here on the range of moves that surfaced in this corpus. There are several types of dialogue games (sub-dialogues), e.g., clarifying the rules of the game, discussions of the game state, small-talk, and, most importantly, trade negotiation dialogue games, which also are the most frequent in the corpus. Each trade negotiation constitutes a dialogue game.

We analyse dialogues on two levels, a game-strategic level and a dialogue level. Put differently, we distinguish the level of *what* game move is being made and *how* that move is realised through lexical and compositional semantics plus an assumption that the contribution is coherent in context. This distinction is particularly useful for building an agent that will play *The Settlers of Catan* and be able to communicate with other players using natural language. For building this agent, the game-strategic computations should be separate from the computations needed for understanding and generating dialogue moves for achieving those game strategies.

Please be aware of the potential confusion between the terms *dialogue move* (or: *move in a dialogue game*), which we adopt from the map task, and *game move*. The latter refers to a move in the game *The Settlers of Catan*, e.g., building a road, or trading 1 ore for 3 wheat, while the former refers to

a move in the dialogue or equivalently a type of speech act; this might be a *property* of the utterance (e.g., to question) or a *relation* (e.g., to give an indirect answer), where the successful performance of the speech act depends on the content of both the current utterance and a prior segment of the dialogue. In trading dialogues, game moves (such as to get 1 ore for 3 wheat) are achieved by performing dialogue moves (e.g., by *accepting* a prior offer through uttering *It's a deal*). In trading dialogues, game moves are performed by performing a dialogue move: so accepting a trade may be performed by performing the speech act *Accept*. If not further specified, we are talking of dialogue games and dialogue moves.

## Dialogue Games

### Dialogue moves in the map task

We will take the twelve types of dialogue moves from the map task (Kowtko et al., 1993) as our starting point. Six dialogue moves initiate a new game: **Instruct**: Give an instruction to the fellow player of what to do. **Check**: Request confirmation of whether a previous utterance has been understood correctly. **Query-YN, Query-W**: A yes/no or wh-question about some aspect of the task, except clarification of an instruction (which is a Check move). **Explain**: Provide new information about the goal of the task with respect to the current state.

Then, there are six further moves: **Clarify**: Clarify or repeat a prior utterance or known information. **Reply-Y, Reply-N**: Elicited response to Query-YN, Check or Align. **Reply-W**: Elicited response to Query-W or Check. **Acknowledge**: Acknowledge understanding (can close a dialogue game.) **Ready**: Indicate intention to start a new game.

Not all dialogue moves can be uniquely assigned to just one of these classes; some fulfil multiple functions.

Dialogue games can contain sub-games, creating a nested structure. Again, we follow Kowtko et al. (1993): 'Nesting is considered to occur when a sub-game is initiated whose purpose can clearly be seen as contributing to the goal of the current game ... A break, such as the announcement of a misunderstanding, puts the status of the current game in doubt; the game might continue if the misunderstanding is cleared up, or it might be abandoned.' (p. 3)

We also follow Kowtko et al. in defining dialogue moves as 'an utterance, a partial utterance or a group of utterances, which convey the same specific intent.' (p. 3) In other words, a dialogue move can consist of one constituent or several, and if it is several constituents they may likewise be individual constituents in a further game, thereby creating a hierarchical structure of moves. Dialogue moves are thus defined by their function rather than linguistic form. We provide the analysis of this more detailed level, i.e. how utterances are related, in terms of SDRT (Asher & Lascarides, 2003) below.

### Dialogue moves in *The Settlers of Catan*

The dialogues in our corpus bear many similarities with map task dialogues, but there are also significant differences.

The fundamental structure of a trade negotiation consists

of one player making an offer and another player accepting or rejecting it. Thus, we replace the map task's Instruct move with the Offer and Counteroffer moves, and the moves Acknowledge and Ready are supplemented by Accept, Reject and Retract.<sup>1</sup> The full scheme, including the dialogue moves retained from the map task scheme, is given in Fig. 3.

---

#### Moves initiating a dialogue game

- Check
- Query-YN, Query-W
- Explain
- Align
- Offer
  - Make Complete Offer
  - Make Incomplete Offer
    - Give Specified
    - Get Specified
  - Undo Trade
- Counteroffer
  - Cannot Accept Proposal
  - Could Accept Proposal
- Game-Strategic Comment

#### Other moves

- Clarify
  - Reply-Y, Reply-N
  - Reply-W
  - Acknowledge
  - Ready
  - Accept
  - Complete Offer
  - Reject
    - Cannot Trade
    - Won't Trade
  - Retract
  - Null-move
- 

Figure 3: Dialogue moves in *The Settlers of Catan*.

The Offer, Counteroffer and Reject moves have subtypes. For **Offers**, apart from the rare **Undo Trade** move, where a player requests to reverse a trade, the main distinction is whether an offer is complete. With **Make Complete Offer** a player specifies precisely both what he is prepared to give and what he wants in return, and from whom. In **Make Incomplete Offer** a player only specifies some values of the trade (e.g., the resource she wants, but not how much, nor from whom, nor for what). **Give Specified** dialogue moves only specify what the player wants to trade (*Does anybody want clay?*) **Get Specified** (*I need clay.*) is similar.

As we are investigating the non-cooperative aspect of dialogue, we subdivide **Reject** moves into those where the player **Cannot Trade** because he lacks the required resources, and **Won't Trade**, where he could make the trade but chooses not to. Because we record the game states, we can distinguish Cannot Trade from Won't Trade by looking at the resources of the speaker. Analogously, **Counteroffers** can be made be-

---

<sup>1</sup>*Instruct* is a very map task specific dialogue move – in the map task one person instructs another person to draw a route on a map. While in *The Settlers of Catan* a player may give an instruction to another player, no player is obliged to follow it. Particularly in trade negotiations, no player can instruct another player to do something, which is why we are dropping it here.

cause the player **Cannot Accept the Proposal** or **Could Accept the Proposal** but does not want to do so.

As in the map task scheme, dialogue moves aren't mutually exclusive, e.g., a Counteroffer (of either type) can also be a Make Incomplete Offer (from dialogue 19): *Hardie: i'll trade 2 wheat for 1 wood – Gytis: I need clay or ore*. Taken by itself, Gytis's dialogue move is simply a Make Incomplete Offer, but in this context, it is also a Counteroffer (he wants clay or ore instead of wheat). Note that the Counteroffer is drawn from both what Gytis says and what Hardie says, and to infer this particular Counteroffer one needs to draw on the analysis at the second (the discourse) level, see the next section. It is through recognising the coherence relation between Gytis's contribution and Hardie's, and reasoning about that relation's *semantics* that one infers that Gytis has implicated acceptance of a part of Hardie's offer, i.e. give Hardie 1 wood.

We add some more dialogue moves to this scheme. Players can start a new sub-game by making **Game-Strategic Comments** in the context of trade dialogues, e.g. *don't give tomm wheat, he can win the game if he gets wheat* (dialogue 1).

A **Complete Offer** move completes the trade offer of a Make Incomplete Offer move by specifying the missing information: i.e., it differs from Counteroffer in that no part of the original offer is rejected. Further, as long as a trade has not been concluded, a player who has offered a trade may **Retract** the offer. Finally, because *The Settlers of Catan* is a multiplayer game, a player can decide to stay silent and not to respond to a trade offer. This is a **Null-move**.

Because dialogues are entirely unrestricted, trade dialogue games can be interrupted by other types of dialogue games, e.g. a dialogue game on how to use the hand panel, cf. Fig. 2. We do not consider these cases here.

## Sincerity and Informativeness

Players can be non-cooperative in at least two ways. First, a player  $P_2$ 's response to  $P_1$ 's prior move may reject the goal that underlies  $P_1$ 's prior utterance, even though he was in a position to fulfil that goal: a Won't Trade move, e.g.,  $P_1$  asks for wheat and  $P_2$  responds *no, sorry* despite having wheat. Second, a speaker's utterance may semantically entail and/or implicate content they know to be false. While a player can in principle perform an insincere dialogue move other than Reject, e.g. Offer, in practice such moves are rare (they have little foreseeable benefit but foreseeable penalties should the lie be exposed). So insincere moves tend to be rejections; indeed, they tend to be Won't Trade (and not Can't Trade) rejections.

Our corpus exhibits the whole range from sincere statements to outright lies. We mainly distinguish sincere rejections from deflections and lies. In the following examples from our corpus, we specify the resources a player has when making an utterance by using the abbreviations c (clay), o (ore), s (sheep), wh (wheat) and wo (wood).

**Sincere rejection.** Sincere rejections of both subtypes are possible. A move is sincere only if declarations of the player's

resources are truthful. For example, amz's utterance below is a sincere Can't Trade move; Joe's and Luke's utterances are sincere Won't Trade moves (all taken from dialogue 8):

amz[c=0|o=0|s=1|wh=1|wo=1] i dont have any clay  
 Luke[c=1|o=0|s=0|wh=1|wo=0] I have one clay but I want it  
 for myself?  
 Joe G[c=1|o=0|s=1|wh=0|wo=0] sorry, i'm holding on to  
 mine

**Deflection.** Deflections are misleading implicatures as described earlier; the following extract from dialogue 2 yields a misleading scalar implicature (since Tomm has wheat):

Cat[c=0|o=2|s=0|wh=2|wo=1] anyone for ore or wheat?  
 Thomas[c=2|o=2|s=0|wh=2|wo=3] i have ore

**Lie.** Lies are moves where a player declares false information, e.g., this Reject from dialogue 7:

Lorelei1292[c=0|o=1|s=3|wh=9|wo=1] can anyone give me  
 some clay for some wheat?  
 AM123[c=2|o=0|s=0|wh=2|wo=1] sorry have none of that!

Even assuming AM123 misunderstood the direction of the trade offer, the only available antecedents for resolving the anaphoric expression *none of that* are *wheat* and *clay*; thus, AM123's utterance declares content she knows to be false.

**Null-move.** Trade offers can also be rejected by not answering at all (from dialogue 11):

Britt[c=1|o=1|s=5|wh=0|wo=0] anyone in need of sheep?  
 Simon[c=0|o=0|s=2|wh=4|wo=0] no thanks  
 David A[c=0|o=1|s=1|wh=2|wo=0] ⟨null⟩  
 Din[clay=0|ore=1|s=1|wh=1|wo=1] ⟨null⟩

**Limited informativeness.** It is strategically important to decide how much information about one's own hand a player is willing to reveal. Deflections are a prime example for this.

**Volunteering information.** In contrast to the limited cooperativity exhibited in, for instance, Null Moves, there are also many instances of players behaving much more cooperatively than strictly speaking they need to in order to advance the domain-level game. For instance (from dialogue 3):

Euan[c=0|o=0|s=2|wh=2|wo=1] Anyone got any ore?  
 Cardlinger[c=0|o=1|s=0|wh=7|wo=3] only the 1  
 Joel[c=0|o=0|s=2|wh=2|wo=1] havent had it in awhile

One possible explanation for why players might choose to provide more specific content than simply a direct answer to the question is that by doing so they are more likely to be perceived as cooperative and providing positive face to their interlocutors (Brown & Levinson, 1978), which may be useful later in the game.

## The Dialogue Level: SDRT

The previous section detailed dialogue moves on the agent level. Now we describe how those moves interact with con-

tent at the dialogue level. This link between dialogue moves on the one hand and coherence relations and semantics on the other, is needed for identifying which dialogue moves (or communicative intentions) an agent has revealed during language interpretation, and it is needed for deciding which utterances (or dialogue moves) to make during language production.

Consider the utterance *I need clay*. Without any context, this is a Make Incomplete Offer move. But in the exchange *A: I want wood. B: I need clay*, the same utterance has to be understood in relation to the content of the previous utterance. Reasoning about its coherent use in context leads one to infer that it is a Complete Offer dialogue move.

We are using the framework of *Segmented Discourse Representation Theory* (SDRT, Asher & Lascarides, 2003; Lascarides & Asher, 2009) for specifying such coherence relations. In SDRT's model of dialogue, each dialogue agent *publicly commits* to coherence relations between his own utterance and prior ones, even if they were said by another agent. Shared public commitments then reveal agreed content; inconsistent public commitments reveal disputes. For instance, above, *B* commits to the coherence relation *Plan-Elab(a,b)* (where *a* is *A*'s utterance and *b* is *B*'s). This means that *b* provides information that elaborates a plan for achieving the communicative intention underlying *a*. *Plan-Elab* is a veridical relation: a public commitment to *Plan-Elab(a,b)* entails a commitment to the contents of *a* and *b*. Thus *B* is committed to *a*: he implicates endorsement of it by the illocutionary effects of the speech act *Plan-Elab* that he performed.

Such moves reveal commitments to preferences as well; in our domain, these preferences are (partial) information about offers that an agent would prefer to perform. Cadilhac, Asher, Benamara, and Lascarides (2011) describe a detailed algorithm for computing such preferences from the dialogue's content. It exploits the recursion over the discourse structure that is engendered by coherence relations to build a partial description of each player's preferences, as it evolves through the dialogue exchange, e.g.:

A: I am prepared to give you rock for wood.  
 B: I need clay.

The relation between these utterances is a *Plan-correction*, because *B* corrects *A*'s offer. The Cadilhac algorithm for constructing *B*'s declared preference in context yields that the implied trade is clay for wood. Thus, it computes *which part* of the prior trade was rejected (i.e., getting rock) and captures implicatures about what is accepted; here, the implicature that *B* accepts giving wood. While the *Plan-correction* relation is between whole utterances, the bits that are accepted and rejected are computed on the basis of the semantics of the relation and the semantics (and information structure) of the arguments (Lascarides & Asher, 2009). In contrast, in

A: [I want wood.]<sub>a1</sub> [What do you want?]<sub>a2</sub>  
 B: [I need clay.]<sub>b</sub>

two discourse relations are at work *Q-Elab(a1,a2)* and *IQAP(a2,b)*. *Q-Elab(a1,a2)* means that *a2* is a question

whose possible answers all elaborate a plan for achieving the underlying communicative intention of  $a_1$ . *IQAP* (Indirect Question Answer Pair) relates a question ( $a_2$ ) to a segment ( $b$ ) whose content, together with shared knowledge, defeasibly implies a direct answer to the question.

Default principles for identifying the scope of implicit endorsements and denials from Lascarides and Asher (2009) also mean that B accepts  $Q\text{-Elab}(a_1, a_2)$ . In other words, by committing to the question (via the veridical *IQAP*), B implicates a commitment also to the question's *illocutionary effects*, here  $Q\text{-Elab}(a_1, a_2)$ . This relation, not the content of  $a_2$  itself, yields an interpretation of the question equivalent to *What do you want in exchange for wood?* Thus B's trade offer  $b$  is a combination of A's and B's contributions.

A typical case from our pilot corpus (dialogue 19) demonstrates that such interactions can be even more complex:

Hardie: [anyone got ore?] $_{a_1}$  [1-1 for either clay or wheat?] $_{a_2}$   
Gytis: [I could give 1 ore for 2 clay] $_b$

Gytis's utterance is related in multiple ways to Hardie's  $Plan\text{-Elab}(a_1, a_2)$ . First, through  $b$ , Gytis commits to a *Plan-correction* on Hardie's  $Plan\text{-Elab}(a_1, a_2)$ : he chooses clay over wheat and wants 2 ore instead of 1. Second, he commits to  $IQAP(a_1, b)$ : he indirectly answers he has ore. Finally, he commits to  $Plan\text{-Elab}(a_1, b)$ :  $a_1$  reveals Hardie's incomplete offer for getting ore, which Gytis's response completes.

Thus, without a principled account of discourse relations like SDRT and an algorithm like the one by Cadilhac et al. (2011) for extracting declared preferences from dialogue content, it will be far from straightforward to map actual dialogue into a representation of the underlying trade, or generate dialogues that express one's intended trade.

## Conclusion

We have collected a pilot corpus of trading dialogues within the zero-sum game *The Settlers of Catan*. These dialogues reveal non-cooperative dialogue to be a nuanced affair: while Gricean maxims of cooperativity break down, basic cooperativity is adhered to and speakers often share intentions.

We argued for two interleaved levels of analysis: a game level on which dialogue moves are particular types of game moves; and a dialogue level where the agents' public commitments to coherent interpretations of their contributions are recorded. Both levels are needed for a computationally feasible mapping between utterances and the agents' preferences during language interpretation and production. Thus, for an agent playing *The Settlers of Catan*, dialogue moves are one type of game move the agent can make and must extract from other players' utterances. The semantic representation that is computed by means of coherence relations is the interface to the linguistic subsystems.

Due to the non-cooperativeness of such dialogues, the dialogue moves that players make vary in their sincerity and informativeness. We highlighted moves from our corpus that demonstrate the spectrum, ranging from fully sincere moves through deflections to lies.

In the STAC project (*STRAtegic Conversation*, see acknowledgments), we are currently in the process of collecting the corpus proper, which our research partners in Toulouse will fully annotate with SDRT relations. An implementation of the Cadilhac algorithm will make it possible to extract preferences from discourse structure to semi-automatically annotate the corpus with the agents' preferences (expressed as offers). The relations among preferences, plus the game state (in particular, a player's resources) then serve to determine the game moves. This will be a rich resource for building a dialogue agent that plays *The Settlers of Catan*, empirically grounding the agent's decisions to match what people do.

## Acknowledgments

This work is supported by ERC grant 269427 (STAC). We are particularly grateful for discussions with our STAC research partners in Toulouse: Stergos Afantenos, Nicholas Asher, Farah Benamara, Anaïs Cadilhac, Cedric Dégremont, Philippe Muller, Soumya Paul and Laure Vieu.

## References

- Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., et al. (1991). The HCRC Map Task Corpus. *Language and Speech*, 34(4), 351–366.
- Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge: Cambridge University Press.
- Baldridge, J., & Lascarides, A. (2005). Probabilistic head-driven parsing for discourse structure. In *Proc. of the 9th Conf. on Comp. Natural Language Learning* (pp. 96–103).
- Brown, P., & Levinson, S. (1978). *Politeness: Some universals and language usage*. Cambridge University Press.
- Cadilhac, A., Asher, N., Benamara, F., & Lascarides, A. (2011). Commitments to preferences in dialogue. In *Proc. of SIGDIAL 2011* (pp. 204–215).
- Clark, H. H. (1996). *Using language*. Cambridge, MA: Cambridge University Press.
- Grice, P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts* (Vol. 3, pp. 41–58). New York: Academic Press.
- Guhe, M., & Bard, E. G. (2008). Adapting referring expressions to the task environment. In *Proc. of the 30th CogSci* (pp. 2404–2409).
- Kowtko, J. C., Isard, S. D., & Doherty, G. M. (1993). *Conversational games within dialogue*. [http://groups.inf.ed.ac.uk/hcrc\\_publications/rp-31.ps.gz](http://groups.inf.ed.ac.uk/hcrc_publications/rp-31.ps.gz). Retrieved on 15 December 2011.
- Lascarides, A., & Asher, N. (2009). Agreement, disputes and commitments in dialogue. *J. of Semantics*, 26(2), 109–158.
- Osborne, M., & Rubinstein, A. (1990). *Bargaining and markets*. Academic Press.
- Solan, L., & Tiersma, P. (2005). *Speaking of crime: The language of criminal justice*. Chicago: U. of Chicago Press.
- Thomas, R. S. (2003). *Real-time decision making for adversarial environments using a plan-based heuristic*. Unpublished doctoral dissertation, Northwestern University.

# The Characteristics of Usability and Users' Eye movements in Searching for Information in a Hierarchically Organized Information Structure

**Yoshiko Habuchi (habuchi@tokuyama-u.ac.jp)**

Department of Welfare Information, Tokuyama University,  
843-4-2 Gakuendai Shunan Yamaguchi, 745-8566 JAPAN

**Haruhiko Takeuchi (takeuchi.h@aist.go.jp)**

National Institute of Advanced Industrial Science and Technology (AIST),  
1-1-1 Higashi Tsukuba Ibaraki, 305-8566 JAPAN

## Abstract

Finding information by successively following hyperlinks on web pages is a typical task performed on the web. A number of web users search for specific information and several studies have concluded that following the "information scent" is the fundamental process involved in the behavior. The aim of this study was to investigate the relationship between the strength of the "information scent" and user behavior by applying a usability inspection method developed for web usability studies. Three typical usability problems of navigation, (a) a page with a weak scent correct link problem, (b) a page with an unfamiliar correct link problem, and (c) a page with a competing links nested under competing headings problem, were compared with a problem-free page. In this study, we applied the Cognitive Walkthrough for the Web method to simulate a website, and analyze user behavior along with usability problems. Participants were asked to find an article on a simulated encyclopedia website. The success rate, total clicks, total time, fixation count and gaze time were analyzed. The results showed that the critical issues caused by usability problems appear on the target-link page. The results of this study confirm the effect of "information scent" and provide a scientific insight into web navigation.

**Keywords:** information scent; LSA: latent semantic analysis; hierarchical information structure; web usability; eye tracking

## Introduction

Over the last decade, eye-tracking studies have provided detailed insights into the way users interact with websites. In earlier usability studies in web navigation, the level of difficulty of a task for a user was measured by finding the number of clicks required or the time taken to complete the task, where a high difficulty level pointed to usability problems in the web page. Eye-tracking methods have made it possible to analyze user behavior in detail.

Several eye-tracking studies have provided an overall understanding of the behavior of website users based on their eye movements. Nielsen (2006) demonstrated that the eye-movement patterns of website visitors are F-shaped by aggregating a large amount of eye-movement data from web pages. Cutrell & Guan (2007) and Guan & Cutrell (2007) presented a series of detailed studies examining the relationship between the fixation duration, and the ranking and presentation of search results i.e. whether the search results were accompanied by short, medium, or long descriptions. They discovered that providing more

information next to the search results significantly improved performance in information tasks, but degraded performance in navigation tasks, where the performance was assessed in terms of selecting the correct search result and the time taken to complete the task. They argued that the difference in performance was because when users were given longer descriptions, they paid more attention to the description and less to the URL of the search result, which would be of help in assessing the usefulness of the search result. These eye-tracking studies have provided a broad understanding of the behavior of web users.

Habuchi, Kitajima, & Takeuchi (2008) focused on users' cognitive activities, especially their goals and knowledge when searching for information on the web. They explored the relationship between usability problems and web users' eye-movements by independently assessing usability problems using Cognitive Walkthrough for the Web (CWW)<sup>1</sup> (Blackmon, Polson, Kitajima, & Lewis, 2002; Blackmon, Kitajima, & Polson, 2003, 2005; Blackmon, Mandalia, Polson, & Kitajima, 2007). They considered three typical usability problems: the weak scent correct links nested under competing headings problem, the unfamiliar correct link problem, and the competing links nested under competing headings problem. They found that if a webpage had any usability problems, user performance deteriorated noticeably in terms of the total number of fixations in the subsequent link selection stage. The performance was worst on pages with a weak scent problem or an unfamiliarity problem, which resulted in longer fixation durations because users examined the link carefully before selecting it. However, no difference in performance was observed in the initial heading selection stage between pages with usability problems and those without. The study showed that it is

---

<sup>1</sup> Cognitive Walkthrough for the Web (CWW) is a web usability inspection method that can detect several usability problems that a website visitor may encounter while navigating through the website in search of specific information by successively selecting hyperlinks on intermediate navigation pages. It uses a Latent Semantic Analysis (LSA) semantic space (Landauer & Dumais, 1997) to measure the "information scent" i.e. local proximal cues. Web users use the "information scent" to make navigation choices while navigating the web by following hyperlinks (Pirolli & Card, 1999; Blackmon et al., 2002, 2003, 2005; Chi & Suppattanasiri, 2003; Pirolli, 2005).

possible to distinguish between such pages based on the task completion time and total number of fixations. Furthermore, it showed that users' gaze patterns varied depending on the whether the problem was in the link itself, or elsewhere. However, the test material they used consisted of just two samples for each usability problem. To verify their results and draw a general conclusion, results obtained using a larger number of samples must be examined.

The aim of this study is to investigate the relationship between usability problems and information search behavior by applying the CWW method to handle various types of usability problems. In this study, we conducted more experiments and controlled other factors such as familiarity of the desired information. We built a controlled website to study the relationship between usability problems and user behavior.

## Experiment

The task used in this study is an information search task performed within a hierarchically organized information structure. The task difficulty depends primarily on the strength of information scent on each navigation page for a given search target. The three types of usability problems compared are, (1) weak scent correct link problem, (2) unfamiliar correct link problem, and (3) competing links nested under competing headings problem; where problem-free items existed. These usability problems can be detected through CWW.

## Method

**Materials** The Japanese LSA semantic space is necessary for finding usability problems in CWW. It was constructed using corpora from the Japanese language Wikipedia abstract containing 116,038 words and 129,937 contexts. The semantic space of the Japanese LSA consists of 116,038 words and 300 dimensions. Please refer to Takeuchi, & Habuchi (2008) for details.

A simulated encyclopedia website was constructed for this study. It has the same link structure as the Microsoft Encarta website. The menu was two levels deep, with 9 top-level links (called headings) and 93 second level links (called links). Each entry word and its description from the Wikipedia abstract were categorized into one of the 93 links. Usability problems were identified by applying CWW on these articles (Takeuchi, & Habuchi, 2008).

We selected 97 candidate articles, such that each article either had no usability problems, or had one of the following problems: weak scent correct link problem; unfamiliar correct link problem; and competing links nested under competing headings problem (three competing links). The predicted link-click count in this experiment was 2.29 for no usability problems, 3.81 for weak scent, 4.05 for unfamiliar links, and 4.26 for competing links nested under competing headings, according to the CWW model (Blackmon et al. 2007).

The entry words from the selected articles varied in their degree of familiarity for users. As it is desirable for each

entry word from the candidate material to have the same level of familiarity, we conducted the word concept familiarity rating task for the 97 entry words with 48 additional subjects (aged between 20 and 33 years with a mean age of 24.3 years). The familiarity rating scale was from 1 (least familiar) to 5 (most familiar). Finally, we selected 20 articles such that each problem type had 5 articles from which unfamiliar entry words would be chosen (see Table 1). These articles were used as the search targets in the experiment.

**Participants** Thirty-five university students participated in this study. Nine of the participants were excluded: six due to not having sufficient eye-tracking data and three others who reported after the experiment that they misunderstood the task. Twenty-six university students (11 males and 15 females; aged between 18 and 24 years with a mean age of 20.6 years) participated in the experiment. All participants were native Japanese speakers and received compensation for their participation. They had normal or corrected to normal vision. They were all regular users of the internet and were used to browsing with Internet Explorer. They were ignorant of the hypothesis being investigated in the study.

**Apparatus and Procedure** The eye-tracking equipment used was a Tobii X60 eye-tracker with Tobii studio 1.24 software. The minimum fixation duration was set to 40 ms and the fixation radius to 20 pixels. Eye movements were recorded at a sampling rate of 60Hz. The experimental website was opened in Internet Explorer 6 under Windows XP. URL visibility events, mouse movement, and click events were recorded.

The task was to find the relevant article within 130 seconds. The search topic, along with a short description was provided to the participants. The expectation was that participants would build a mental representation of the item, which they would use to evaluate the "information scent" of the navigation pages in the website.

Participants were tested individually. They were first seated in front of a screen-mounted eye-tracking system. Next, they were required to search for the desired information on the simulated encyclopedia website. In each test, the participant was given a target word. On clicking, the target word was replaced on the screen by the heading page containing a description of the search target and 9 headings. Participants made their choice by clicking on one of the sections. If a participant reached the target-link, the target page was shown, and then the next trial began when the participant clicked the "next" button (see Figures 1 and 2). To eliminate order effects, the 20 items were allocated among four blocks and the order of the blocks was counterbalanced. Each participant was given four blocks of five trials each. There was a short break of a few minutes between each block. Participants were tested each session lasted approximately 25 minutes.

The dependent variables were the success rate, time taken to complete the task, total number of clicks, total fixation count and gaze time.

Table 1: Properties of experimental materials.

Property	No. of items	Concept familiarity <sup>1</sup>	Target and Target-link Cosine value	Target-link Vector length
Problem-free	5	1.14 (.41)	.52 (.31)	.16 (.30)
Weak Scent	5	1.20 (.48)	.09 (.00)	.22 (.22)
Unfamiliar	5	1.12 (.42)	.58 (.23)	.08 (.13)
Competing links (3)	5	1.18 (.47)	.45 (.27)	.42 (.00)

<sup>1</sup> Concept familiarity was rated on scale from 1 (least familiar) to 5 (most familiar).

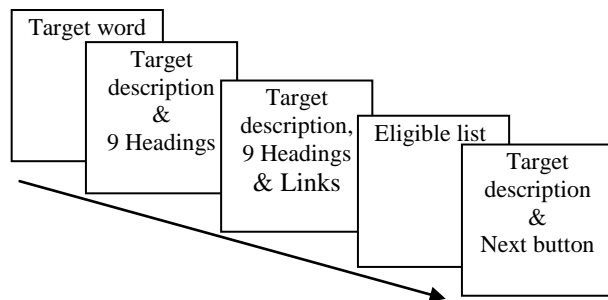


Figure 1: Sequence of events

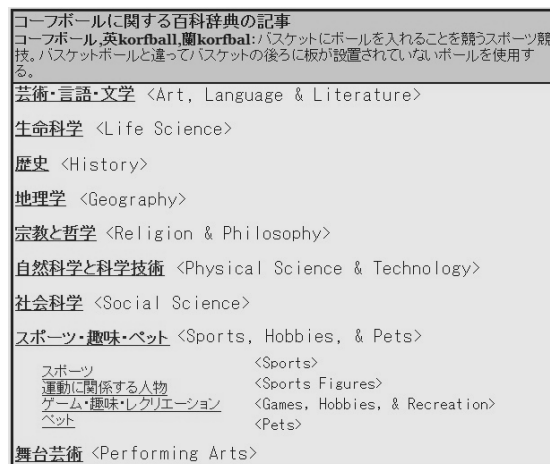


Figure 2: Menu on the target-link page. The top part of the screen shows the target description and the area under it shows the menu. The links that appear under each heading became visible on clicking on the appropriate heading, and the links were only shown for one heading at a time.

## Results

We analyzed the data from when a heading page was opened to when the target link was clicked. An analysis of variance (ANOVA) and several comparisons were performed separately for the success rate, total click count, and the task completion time. These results are shown in Tables 2 and 3, grouped by usability problem type. An ANOVA and several comparisons were also performed

separately for the success rate, total clicks, and total time. The significance level for ANOVA results was  $p < .05$ . A summary of the success rate, total clicks, and total time results for the experiment is presented in Table 2.

**Success rate** Before an angular transformation was performed, the success rate data were skewed: Problem-free = -2.56, Weak scent problem -1.19, Unfamiliarity problem -.53, Competing links problem 1.36. Post-transformation, the skewness was reduced to -2.56, -1.11, -.03, and 1.36, respectively. An analysis of the variance indicated a significant difference between the angular-transformed success rates [ $F(3, 75)=30.20$   $p < .001$ ]. Post hoc testing indicated that the transformed success rate for the “problem-free” case were significantly higher than for cases with a usability problem.

**Click count** Prior to the log transformation, the total clicks data were skewed: Problem-free 1.03, Weak scent problem 0.97, Unfamiliarity problem 0.24, Competing links problems 0.45. Post-transformation, the skewness was reduced to 0.50, 0.36, -0.31, and -0.38, respectively. An analysis of the variance indicated a significant difference among the log-transformed total clicks for usability problems [ $F(3, 75)=51.20$ ,  $p < .001$ ]. Post hoc testing indicated that the transformed click count for the “problem-free” case were significantly less than for cases with a usability problem.

**Solution time** Prior to the log transformation, the solution time data were skewed: Problem-free 1.18, Weak scent problem 0.25, Unfamiliarity problem 0.12, Competing links problems 0.18. Post-transformation, the skewness was reduced to 0.51, -0.07, -0.64, and -0.38, respectively. An analysis of the variance indicated a significant difference in the log-transformed success time for usability problems [ $F(2.58, 64.38)=59.63$ ,  $p < .001$ ]. Post hoc testing indicated that transformed solution times for the “problem-free” case were significantly less than for cases with a usability problem.

## Eye tracking analysis

We conducted eye-tracking analysis for two regions of the web page, the target description area, and the navigation area during each target selection stage. Tables 4 and 5 show the results of the eye tracking analysis.

**Fixation counts** Before analysis, all data were logarithmically transformed.



Table 2: Success rate, total click count, and solution time.

	Success rate	Total Clicks	Solution time
Problem-free	97.7%	3.92 (1.7)	17.1 ( 9.7)
Weak Scent correct link	72.3%	9.45 (2.6)	53.0 ( 9.7)
Unfamiliar correct link	87.7%	6.93 (2.2)	36.6 (15.8)
Competing link nested under the competing heading (3)	84.6%	7.15 (1.6)	35.4 ( 7.2)

Note: Time measurements in seconds. Number shown in parentheses is SD.

**Target-heading selection stage** A two-way analysis of variance was conducted for the fixation counts until the initial heading was clicked. An information area main effect was not obtained [ $F(1, 25) = 2.23$ , n.s.]. Also, the usability problem main effect was significant [ $F(2.77, 69.34) = 25.02$ ,  $p < .001$ ], confirming that the Weak scent problem link pages had significantly more fixations than other pages. Finally, the information area and usability problems interaction was significant [ $F(2.57, 64.35) = 20.19$ ,  $p < .001$ ]. A Bonferroni post hoc analysis of this interaction revealed that the Weak scent and Unfamiliar link pages had more fixations than that the other two cases, regardless of page area. In addition, although the Competing links problem pages had fewer fixations in the goal description area, they had more fixations in the navigation area.

**Transit-link stage** Not all participants saw the extra transit-link page; analysis reduced the participant number from 26 to 19. A two-way analysis of variance was conducted for the total fixation count in the transit-link selection page. An information area main effect [ $F(1, 18) = 111.64$ ,  $p < .001$ ] was significant, confirming that the navigation area was responsible for far more fixations than the target description area. The usability problem main effect was significant [ $F(2.48, 44.63) = 13.02$ ,  $p < .001$ ], which confirmed that the Weak scent or Competing problem link pages had significantly more fixations than the Problem-free pages. The interaction of information area and usability problems was significant [ $F(2.65, 47.67) = 4.81$ ,  $p < .01$ ]. A Bonferroni post hoc analysis was performed and it showed that in total, the Weak scent, and Competing links problem pages had more fixations than the Problem-free pages.

**Target-link selection stage** A two-way analysis of variance was conducted for the fixation counts in the target-link selection page. An information area main effect [ $F(1, 25) = 156.28$ ,  $p < .001$ ] was significant, confirming that the navigation area was subject to significantly more fixations than the goal description area. The usability problem main effect was significant [ $F(2.98, 74.57) = 15.87$ ,  $p < .001$ ], confirming that the Weak scent and Unfamiliarity problem link pages had significantly more fixations than the Problem-free pages. However, Competing links problem pages had fewer fixations than Problem-free pages. The interaction of information area and usability problems was also significant [ $F(3, 75) = 5.16$ ,  $p < .01$ ]. A Bonferroni post hoc analysis confirmed the difference in fixation distribution. Weak scent link problem pages in total had more fixations than Problem-free pages. In the target-description area,

Unfamiliar link pages had more fixations than Problem-free pages. However, Weak scent link pages gathered more fixations in the navigation area.

**Gaze time** Before analysis, all data were logarithmically transformed.

**Target heading selection stage** A two-way analysis of variance was conducted for the gaze time until the target-link was clicked. An information area main effect was obtained [ $F(1, 25) = 4.83$ ,  $p < .05$ ], confirming that the target-description area had significantly longer gaze time than the navigation area. The usability problem main effect was significant [ $F(2.94, 73.59) = 24.18$ ,  $p < .001$ ], confirming that the Weak scent problem link pages had significantly longer gaze time than the other cases. The information area and usability problems interaction was significant [ $F(2.55, 63.72) = 18.05$ ,  $p < .001$ ]. A Bonferroni post hoc analysis of this interaction revealed that the gaze time at the target-description area for Weak scent correct link problem items was longer than that in other cases. However, in the navigation area, Weak scent, and Unfamiliar link problem pages had longer gaze time than the Problem-free pages.

**Transit-link stage** We summed up surveyed areas individually and divided total gaze time by total pages opened. As a result, the number of participants was reduced from 26 to 19. A two-way analysis of variance was conducted for the gaze time for transit-link pages. An information area main effect was obtained [ $F(1, 18) = 83.12$ ,  $p < .001$ ]. The usability problem main effect was significant [ $F(1.62, 29.24) = 10.96$ ,  $p < .001$ ], confirming that Weak scent and Competing links problem pages had significantly longer gaze time than the Problem-free pages. The information area and usability problems interaction was obtained [ $F(1.44, 25.94) = 7.86$ ,  $p < .01$ ]. A Bonferroni post hoc analysis of this interaction revealed that in the target-description area, all usability problem pages had a longer gaze time than Problem-free pages. However, in the navigation area, Weak scent, and Competing links problem pages had longer gaze times than the Problem-free case.

**Target-link selection stage** A two-way analysis of variance was conducted for the gaze time of the target-link selection page. An information area main effect [ $F(1, 25) = 7.55$ ,  $p < .001$ ] was significant, confirming that the navigation area was subject to significantly longer gaze time than the target description area. The usability problem main effect was significant [ $F(2.1, 52.53) = 7.55$ ,  $p < .001$ ], confirming that Competing links problem pages had a lower gaze time than other cases. Unfamiliar link problem pages had a longer

gaze time than the Problem-free case. The interaction between the information area and usability problems was not significant [ $F(1.6, 39.88)=2.32$ , n.s.].

## Discussion

The results indicated clear differences between problem-free and usability problem pages, in the success rate, total clicks and total time. The eye-tracking data showed that users' eye movements for selecting a target link varied depending on the nature of the usability problem. Two main conclusions can be drawn from this study. The first is that usability problems occur from the first heading selection stage in the navigation area. Pages with a weak scent problem had more issues than other pages. Web pages with the latter problems forced users to examine links and the target description more times than for problem-free pages.

The other point was certain characteristics of usability problems were clear at the target-link selection stage. In pages with a weak scent problem or an unfamiliarity problem, it took users longer to select a link. This is presumably because participants had to discern the meaning of the correct link by carefully reading the link label. A possible explanation for this phenomenon is that for the weak scent correct link problem, participants looked at the navigation area several times. In contrast, for the unfamiliar correct link problem, participants remained at one link page for longer. These situations differed from cases of competing links nested under competing headings, where users could immediately select a link because it was a good match for the target and less effort was needed to understand the meaning of the link. We can also say users' eye movement before clicking a link at the target-link selection stage varied according to the usability problem of the targeted link.

Table 3: Mean percentage of shortest path achievement.

Processing Stage	From the start until Target-heading	From Target-heading until Target-link	From the start until Target-link
Problem-free	83.8% (13.9)	72.3% (18.8)	63.8% (16.0)
Weak Scent correct link	54.6% (18.4)	34.6% (16.5)	23.8% (15.0)
Unfamiliar correct link	55.4% (19.8)	51.5% (24.1)	23.8% (15.0)
Competing links nested under the competing heading (3)	53.8% (17.7)	57.7% (16.3)	36.2% (18.8)

Note: Number shown in parentheses is SD.

Table 4: Mean Fixation counts.

Processing Stage	Target-Heading selection stage (n=26)		Transit Link page (n=19)		Target-Link selection stage (n=26)	
Area of web page	Target description	Navigation	Target description	Navigation	Target description	Navigation
Problem-free	15.2 (8.1)	18.9 (12.2)	7.6 (10.7)	53.5 (30.0)	8.1 (6.9)	31.6 (15.8)
Weak Scent correct link	31.7 (14.3)	26.8 (14.5)	23.3 (17.7)	100.5 (48.7)	11.1 (7.3)	55.4 (24.0)
Unfamiliar correct link	16.0 (8.8)	21.6 (10.6)	11.4 (10.7)	65.5 (37.0)	15.9 (13.1)	42.6 (25.2)
Competing links nested under competing heading (3)	12.2 (7.6)	23.2 (12.8)	16.0 (12.9)	73.9 (50.2)	5.4 (5.5)	25.7 (13.4)

Note: Number shown in parentheses is SD.

Table 5: Mean gaze time.

Processing Stage	Target-Heading selection stage (n=26)		Transit Link page (n=19)		Target-Link selection stage (n=26)	
Area of web page	Target description	Navigation	Target description	Navigation	Target description	Navigation
Problem-free	2612 (1650)	3433 (1667)	1306 (1828)	10196 (7101)	1422 (1203)	6112 (2673)
Weak Scent correct link	5556 (2823)	4738 (2205)	4306 (3613)	16806 (9807)	1853 (1269)	9346 (3307)
Unfamiliar correct link	2956 (1606)	4089 (1798)	1830 (1882)	11953 (5767)	2604 (2153)	8731 (3754)
Competing links nested under competing heading (3)	2173 (1446)	3960 (1741)	2668 (2458)	12057 (6333)	1017 (1178)	4521 (1878)

Note: All measurements are on the millisecond time scale. Number shown in parentheses is SD.

## Conclusion

This study aimed to investigate how website visitors search for desired information in web pages with usability problems. Three types of usability problems, the weak scent correct link problem, the unfamiliar correct link problem, and the competing links nested under competing headings problem, were compared against problem-free pages. The results showed that participants' behavior when faced with usability problems was distinguishable from behavior on problem-free pages in terms of the success rate, total number of clicks and total time taken. In addition, it was possible to distinguish the usability problem facing the user from the fixation counts and gaze times at the target-link selection stage.

The eye-tracking analysis expanded on the information gained in previous studies. This study showed that users' eye movement before clicking a link in the target heading and target-link selection stages varies depending on the usability problem. Pages with competing links nested under competing headings have users spending less time before selecting the correct link, but they involve more clicks due to confusion from incorrect links. Pages with a weak scent problem or an unfamiliarity problem have users taking more time to select the correct link.

This study has attempted to explain the relationship between "information scent" and user behavior. In future work, we aim to extract the discriminative behavior when two or more problems exist simultaneously.

## Acknowledgments

This study was conducted while the first author was in the National Institute of Advanced Industrial Science and Technology (AIST). We are grateful to Muneo Kitajima, Professor, Nagaoka University of Technology, for his support and guidance.

## References

- Blackmon, M. H., Polson, P. G., Kitajima, M., & Lewis, C. (2002). Cognitive walkthrough for the web. *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves* (pp. 463-470). New York, NY: ACM.
- Blackmon, M. H., Kitajima, M., & Polson, P. G. (2003). Repairing usability problems identified by the cognitive walkthrough for the web. *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 497-504). New York, NY: ACM.
- Blackmon, M. H., Kitajima, M., & Polson, P. G. (2005). Tool for accurately predicting website navigation problems, nonproblems, problem severity, and effectiveness of repairs. *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 31-40). New York, NY: ACM.
- Blackmon, M. H., Mandalia, D. R., Polson, P. G., & Kitajima, M. (2007). Automating usability evaluation: Cognitive walkthrough for the web puts LSA to work on real-world HCI design problems., In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Chi, E., Rosien, A., & Suppattanasiri, G. (2003). The bloodhound project: Automating the discovery of web usability issues using the infoscent simulator. *Proceedings of the ACM Conference on Human Factors in Computing Systems* (pp. 505-512). Fort Lauderdale, NY: ACM.
- Cutrell, E., & Guan, Z. (2007). What are you looking for?: an eye-tracking study of information usage in web search. *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 407-416). New York, NY: ACM.
- Guan, Z., & Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 417-420). New York, NY: ACM.
- Habuchi, Y., Kitajima, M., & Takeuchi, H. (2008). Comparison of eye movements in searching for easy-to-find and hard-to-find information in a hierarchically organized information structure. *Proceedings of the 2008 symposium on Eye tracking research & applications* (pp. 131-134). New York, NY: ACM.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
- Nielsen, J. (2006, April 17). F-shaped pattern for reading web content. Jakob Nielsen's Alertbox, Retrieved from [http://www.useit.com/alertbox/reading\\_pattern.html](http://www.useit.com/alertbox/reading_pattern.html).
- Pirolli, P. (2005). Rational analysis of information foraging on the web. *Cognitive Science*, 29, 343-373.
- Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, 106, 643-675.
- Takeuchi, H., & Habuchi, Y. (2008). An automatic evaluation of category names based on Latent Semantic Analysis. *Proceedings of Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on advanced Intelligent Systems* (pp. 423-428). Nagoya, Japan: SCIS & ISIS.
- Wikipedia: The Free Encyclopedia. (n.d.). Retrieved from <http://ja.wikipedia.org/wiki/>

# A time-invariant connectionist model of spoken word recognition

**Thomas Hannagan (thom.hannagan@gmail.com)**

CNRS & Aix-Marseille University  
3, place Victor Hugo, 13331 Marseille, France

**James S. Magnuson (james.magnuson@uconn.edu)**

Department of Psychology, University of Connecticut, 406 Babbidge Road, Unit 1020, Storrs, CT 06269-1020 USA  
and Haskins Laboratories, 300 George St., New Haven, CT 06511 USA

**Jonathan Grainger (i.jonathan.grainger@gmail.com)**

CNRS & Aix-Marseille University  
3, place Victor Hugo, 13331 Marseille, France

## Abstract

One of the largest remaining unsolved mysteries in cognitive science is how the rapid input of spoken language is mapped onto phonological and lexical representations over time. Attempts at psychologically-tractable computational models of spoken word recognition tend either to ignore time or to transform the temporal input into a spatial representation. This is the approach taken in TRACE (McClelland & Elman, 1986), the model of spoken word recognition that has the broadest and deepest coverage of phenomena in speech perception, spoken word recognition, and lexical parsing of multi-word sequences. TRACE reduplicates featural, phonemic, and lexical inputs at every time step in a potentially very large memory trace, and has rich interconnections (excitatory forward and backward connections between levels and inhibitory links within levels). This leads to a rather extreme proliferation of units and connections that grows dramatically as the lexicon or the memory trace grows. Our starting point is the observation that models of visual object recognition – including visual word recognition – have long grappled with the fundamental problem of how to model spatial invariance in human object recognition. We introduce a model that combines one aspect of TRACE – time-specific phoneme representations – and higher-level representations that have been used in visual word recognition – spatially- (here, temporally-) independent diphone and lexical units. This reduces the number of units and connections required by several orders of magnitude relative to TRACE. In this first report, we demonstrate that the model (dubbed TISK, for Time-Invariant String Kernel) achieves reasonable accuracy for the basic TRACE lexicon and successfully models the time course of phonological activation and competition. We close with a discussion of phenomena that the model does not yet successfully simulate (and why), and with novel predictions that follow from this architecture.

**Keywords:** Keywords: Spoken Word Recognition; Time invariance ; Computational models; TRACE.

## Background

Could it be that despite very salient differences, the auditory and visual systems actually rely on the same mechanisms in order to recognize words? One signal has a temporal dimension and is carried by transient sound waves, the other is spatially extended and travels at the speed of light. One signal travels sequentially (over time) through the

cochlear nerve, the other in parallel through the optic nerve. In their own dedicated primary cortical regions, however, both arrive at spatial representations – tonotopic for the auditory system, retinotopic for the visual system. What happens next, according to computational models of visual and spoken word recognition, further hints at some possible unification.

## Modeling spoken and visual word recognition: TRACE and IA

From a psycholinguistic point of view, two early models of word recognition based on the same computational framework have been enormously successful. In the visual domain, the Interactive Activation (IA) model and its extensions (McClelland & Rumelhart, 1981; Grainger & Jacobs, 1996) can account for a large number of robust and sometimes counterintuitive behavioral findings, in a simple and elegant hierarchical structure where units at any level compete to represent the stimulus, and engage in "lobbying" up and down in the hierarchy. In the auditory domain, TRACE (an extension of the IA framework for speech; McClelland & Elman, 1986) continues to produce new insights into human behavior, including close fits to fine-grained estimates of the time course of spoken word recognition from the visual world paradigm (Allopenna et al., 1998; Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Dahan, Magnuson, & Tanenhaus, 2001).<sup>1</sup>

One probably superficial difference between the two models is that between-level connections in IA models of reading typically include both inhibitory and excitatory connections, whereas between-level connections in TRACE

---

<sup>1</sup> It is important to note that current, psychologically tractable models of spoken word recognition do not take real speech as their input. While Grossberg & Myers (2000) have modeled aspects of speech and word processing using real speech inputs, these efforts have not yet yielded a model that can handle speech input and a broad range of phenomena in spoken word recognition. In order to be able to address complex issues in word recognition without first solving all fundamental problems in speech perception, TRACE's inputs (for example) are "pseudo-spectral" acoustic-phonetic features that ramp on and off over time, with temporal overlap between adjacent phonemes providing a coarse analog of coarticulation.

are only excitatory. The evidence that this is superficial comes from demonstrations in visual letter identification that performance is at least as good without inhibitory connections between levels (Rey et al., 2009). A much more serious difference, however, is that the IA model can only recognize words at one location on the retina, whereas TRACE has to recognize words at any point in time.

But this impressive ability of TRACE is only achieved at the price of duplicating each unit for as many time slices as needed in the simulation. That is, the processing units in TRACE form a large memory, with units aligned with time 'slices'. Essentially, there is a copy of every feature, phoneme, and word unit at every time slice (the complete details are more complex – for example, words are only duplicated every 3 time slices; see McClelland & Elman, 1986, for details). When input begins, the first instant of the input aligns with and activates units in the first time slice in memory. As the input continues, it activates nodes aligned with specific time slices. Those units can become and remain active for a considerable time after the input has continued on. Conceptually, this is like marks on a page left by a seismograph – the memory banks contain a trace of the input that has come along. But these are not passive traces, since unit activations flux as a function of excitatory and inhibitory input from other units, and a decay parameter.

Having reduplicated units allows TRACE to solve the temporal alignment problem by brute force; given the input /dad/, it can tell that the phoneme /d/ should be activated twice and how far apart in time the two occurrences are – because the two instances of /d/ are encoded by completely independent /d/ detectors aligned with different points in time. The same applies at the word level; TRACE can tell that /dag/ (the TRACE representation of DOG) occurs twice in /dagtsdag/ (DOG EATS DOG) because the two instances are encoded by independent /dag/ detectors aligned with different points in time.

But this comes at a cost. Consider the number of units per slice: 63 x 3 features, 14 phonemes, and, in the basic TRACE lexicon, 212 words, for 415 units. If we ballpark the number of connections by assuming an average of 8 featural connections per phoneme, and 3 phonemes per word, and allowing for connections between units at adjacent time slices, we would have approximately 47,000 connections per time slice with a 200-word lexicon. If we make the trace approximately 2 seconds long (the duration of echoic memory), we need approximately 83 thousand units and 9.4 million connections. If we increase the lexicon to a more realistic size of 20,000 words and the phoneme inventory to 40, these figures reach approximately 4 million units and 80 billion connections.

One might argue that this may not be an unreasonable scale, given the number of neurons and connections in the brain. However, principles of parsimony (might there be a simpler solution?) and evolutionary pressures to minimize energy consumption would be reasonable motivations to seek a less costly solution to time-invariance. Exploring such an alternative is the purpose of this paper, and we

report first results on a model that achieves decent performance using many fewer nodes and connections than TRACE. With a 2 second layer of time-invariant input nodes and TRACE's 14 phonemes and 212 words, TISK requires 9.7 thousand units and 62 thousand connections. This represents a 9-fold improvement over TRACE for units, and 2 orders of magnitude for connections. Critically, the orders of magnitude in improvement turn out to be proportional to lexicon size: with 20,000 words and 40 phonemes, TISK would require 48 thousand units (TRACE requires 84 times more) and 3.3 million connections (TRACE requires 24 *thousand* times more).

## String kernels

In the machine learning literature, one computational technique that has been very successful at representing sequences of symbols independently of their position goes under the name of *string kernels* (Hofmann et al., 2007). Symbols can be amino-acids, nucleotides, or letters in a webpage: in every case the gist of string kernels is to represent strings (such as "TIME") as points in a high-dimensional space of symbol combinations (for instance as a vector where each component stands for a combination of two symbols, and only the components for "TI", "TM", "TE", "IM", "IE", "ME" would be non-zero). It is known that this space is propitious to linear pattern separations and yet can also capture the (domain-dependent) similarities between them. Although it has been argued in the visual modality that string kernels can account for masked priming effects and are thus likely involved in the early stages of processing, there has been very little investigation of String kernels in the auditory domain (Gales, 2009, being a yet unpublished exception).

Given the demonstrated versatility of the technique, there is every reason to suspect that string kernels could also work in spoken word recognition, where symbols would then be discrete and time-specific phonemes, which would be turned into vectors in the space of time-invariant phoneme combinations. This would entail that the same type of representations are in fact at work in spoken and visual word recognition. However, while one can find some appeal in this unification (this would for instance pave the way to establishing connections between sublexical orthography and sublexical phonology), there remains the nagging problem of how to turn sequences of time-specific phonemes into time-invariant phoneme combinations – that is, how to compute the string kernel for spoken words. Thinking in the unified framework of string kernels suggests that similar problems across modalities can receive similar solutions, and we now introduce our time-invariant alternative to the TRACE model, which handles the transition between time-specific and time-invariant units in much the same way as location-specific and location-invariant units are activated in the visual modality, through the use of symmetry networks (Shawe-Taylor, 1989).

## Model

### Architecture and dynamics

The model is illustrated in Figure 1. It uses the same lexicon and basic activation dynamics as the TRACE model, but a radically different architecture. It is comprised of four levels: inputs, phonemes, nphones (currently, nphones are single phones or diphones) and words. Inputs consist of a bank of time-specific feature units as in TRACE, through which a wave of transient activation pattern travels. However, this input layer is deliberately very simplified compared to its TRACE analog, given that at any time there is always at most one input unit active – inputs do not overlap in time, and do not code for phonetic similarity (that is, the /d/ unit is equally similar to /a/ and /t/, as each unit can either be on or off; we will address phonetic grain in future work). This input level sends activation forward to the phoneme level. The time-specific phoneme level consists of 10 banks of 14 phonemes that serve as input to the network (the limitation to 10 is completely arbitrary, but sufficient for single-word recognition; there are only 14 phonemes because we are using the 14 phonemes implemented in TRACE). Input phonemes are introduced one at a time and activate the time-invariant nphone level via feedforward connections. Phoneme-to-nphone connection weights are set according to a gradient weighting scheme that we will shortly describe. The nphone level consists of  $196 + 14$  units, one for each phoneme and for each of the 142 possible diphones that can be composed with the set of phonemes. Units at this level compete with one another via lateral inhibition, and send activation forward to the time invariant word level through excitatory connections, whose weights were normalized by the number of nphones of the destination word. The word level consists of 212 units, one for each of the original words in the TRACE lexicon, with lateral inhibition between words, and feedback excitatory connections from words to nphones.

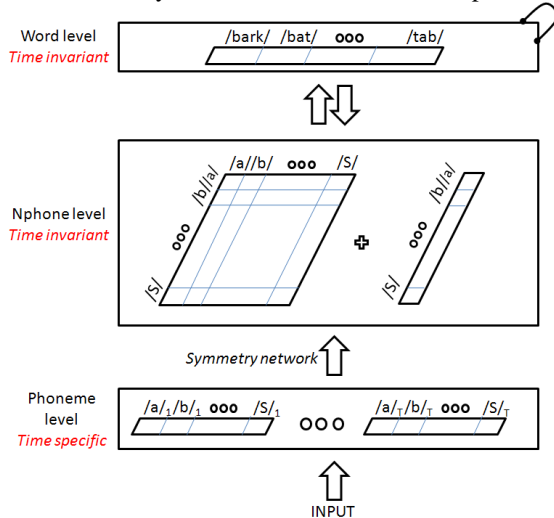


Figure 1: The TISK model - a time-invariant architecture for spoken word recognition.

Note that feedback serves several functions, as does lexical-phonemic feedback in TRACE: it provides a basis for lexical effects on phoneme decisions; it makes the model more efficient and robust against noise (Magnuson et al., 2005); and it provides an implicit sensitivity to phonotactics – the more often a phoneme or nphone occurs in lexical items, the more feedback it potentially receives. Feedback in models of spoken word recognition is a controversial topic (see McClelland et al., 2006; McQueen et al., 2006; Mirman et al., 2006), which we do not address here; our aim is to see whether a model with a radically simpler computational architecture compared to TRACE can (begin to) account for a similar range of phenomena in spoken word recognition.

Units in the model are leaky integrators: at each cycle, all units are activated according to the net input they receive and to their previous activation, minus a decay term, as described in equation 1:

$$A_i(t) = \begin{cases} A_i(t-1) * (1 - Decay) + Net_i(t) * (1 - A_i(t-1)), & \text{if } Net_i > 0 \\ A_i(t-1) * (1 - Decay) + Net_i(t) * A_i(t-1), & \text{if } Net_i < 0 \end{cases}$$

and where the net input of unit  $i$  at time  $t$  is given by:

$$Net_i(t) = \sum_{j=1}^k w_{ij} a_j(t)$$

The python code for the model as well as the list of parameters are available upon request to the first author. We now describe how the connections between phonemes and nphones are set in the model.

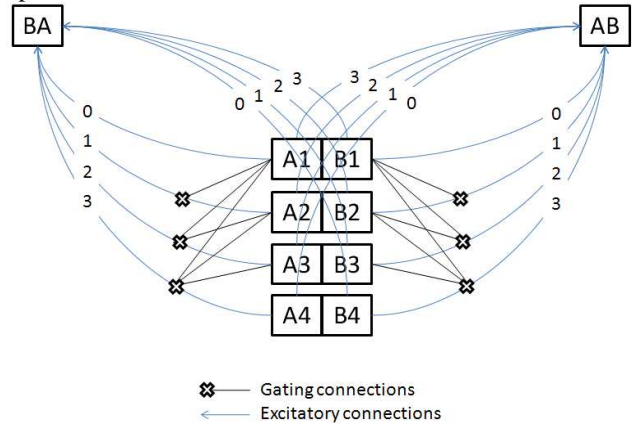


Figure 2: A symmetry network for time-invariant nphone recognition that can distinguish between anaphones.

### A symmetry network for phonological string kernels

The problem we are confronting here is to achieve time-invariant recognition while still distinguishing between transposed phoneme combinations. Since we must recognize a succession of phonemes like  $[/a_t, /b_{t+1}]$  whatever time “ $t$ ” is, we need to be able to recognize each phoneme /a/ and /b/ at any “ $t$ ”. But since each unit must activate at any time, how then can we activate unit /ab/ rather than /ba/ at the nphone level?

This issue of selectivity (here, between “anaphones”: words with the same phonemes in different order) versus invariance (here, to position-in-time) has long been identified in the fields of visual recognition and computer vision, and has recently received attention in a series of articles investigating invariant visual word recognition (Dandurand, Grainger, & Dufau, 2010; Dandurand, Hannagan, & Grainger, 2010; Hannagan, Dandurand & Grainger, 2011). The solution adopted in the present model is illustrated in Figure 2, and was inspired by what has been learned through this recent work on the way various backpropagation networks deal with the selectivity versus invariance dilemma (to our knowledge this solution has not yet been proposed in spoken word recognition models). Briefly stated, this consists of correlating phoneme-to-nphone connection strengths with phoneme position-in-time, as illustrated in Figure 2 (blue excitatory connections). If the weights from phoneme units  $/a_1/$ ,  $/a_2/$ , ...,  $/a_T/$  to diphone unit  $/ab/$  decrease linearly from  $T-1$  to zero, and if on the contrary the weights from phoneme units  $/b_1/$ ,  $/b_2/$ , ...,  $/b_T/$  to diphone unit  $/ab/$  increase at the same pace from zero to  $T-1$ , then presenting the input sequence  $[/a_t/ \ /b_{t+1}]$  will always result in a constant net input for  $/ab/$  whatever the time “ $t$ ” is, and it will result in a smaller constant net input to  $/ba/$ . By setting the weights from these phoneme units to the transposed diphone  $/ba/$  in exactly the opposite pattern, and by setting once and for all a common activation threshold for every diphone unit anywhere between these two net inputs, one can ensure that the network can always neatly distinguish between  $/ab/$  and  $/ba/$ . To prevent sequences with repeated phonemes like  $[/b_1/ \ /a_2/ \ /b_3/]$  from activating large sets of unwanted nphones like  $/bi/$ ,  $/b^{\wedge}/$ , it is however necessary to introduce gating connections (black connections in Figure 2), whereby for instance  $/b_1/$  disables the connection between all future  $/b_{t>1}/$  and diphones  $/*b/$  (where “ $*$ ” stands for any phoneme but  $b$ ).

Other architectures exist that can operate the transition between time-specific phonemes and time-invariant nphones, but the symmetry network we introduce within this model builds on a solution found by the backpropagation algorithm, and has thus arguably a headstart in learnability. It also seems to provide a faithful implementation of the “string kernel” code recently described by Hannagan & Grainger (2011).

## Results

### Recognition rate

We presented the model with every word in its 212-word lexicon. A word was counted as correctly recognized if it had been the most active lexical unit for ten cycles in a row before the deadline, which was set to 100 cycles. Recognition performance was similar across different operational measures of recognition. With these settings, the model correctly recognizes 98% of the 216 words. We consider this satisfactory for a first test of a new computational approach, although the TRACE model

reaches 100% recognition. A consideration of the few unrecognized words, like  $/triti/$  and  $/st^{\wedge}di/$ , is instructive in that they were often confused with their cohort candidates (e.g.  $/trit/$  and  $/st^{\wedge}di/$ ), which activate exactly the same nphones but one (resp.  $/ti/$  and  $/id/$ ). This confusion can only happen in the current model when two phonemes are closely repeated at the end of a relatively long word, since the importance of any one nphone for recognition of a word is currently inversely proportional to how many nphones it activates. We note that a model whose nphone-to-word weights would be set following other criteria (for instance, the conditional probability of the word given the nphone) would give more importance to diagnostic nphones and reach perfect accuracy.

### Competitor effects: Cohort, rhyme and embedding

Figure 3 shows the average cohort and rhyme effects in the model (left panel) and in TRACE (right panel). The curves were calculated by averaging across trials the activation levels of all targets (“target” curve in black), of all words that started with the same phonemes as the target (“Cohorts” curve in red), of all words that ended with the same phonemes as the target (“Rhymes” curve in blue), of all words contained in the target (“Embeddings” curve in purple) and of all other words (“Mean of all words” curve in grey).

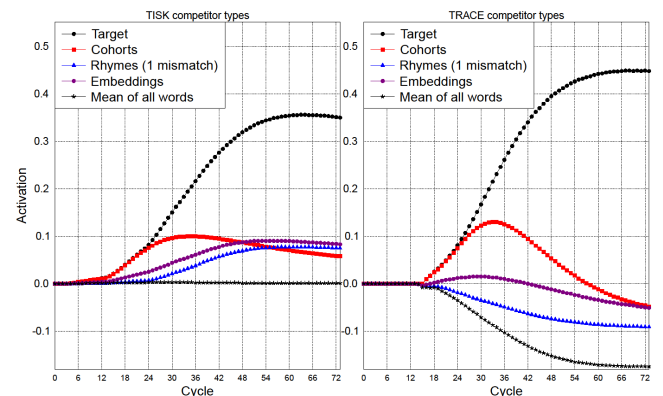


Figure 3: Average activations in the lexicon, when partitioned for each trial as Target, Cohort words, Rhyming words, embedded words and All words.

Left panel: TISK model.

Right panel: TRACE model.

Apart from superficial differences in zero-valued versus negative resting levels, Figure 3 shows that the agreement between models is good on competitor effects. Indeed the magnitude and ordering of the Cohorts, Rhymes and Embeddings effects is similar in the two models, relative to the baseline Mean of all words.

The behavior exhibited by both models also mirrors the cohort and rhyme effects that have been reported in humans performing for instance the so-called “visual world” task. In a nutshell, overall candidate words that begin like the target are more active early on during processing while those that



end like the target are more active later one during processing, without ever rising to the activation level of the target, or going below the activation level of unrelated words.

## Discussion

The previous results tentatively suggest that a time-specific model of spoken word recognition like TRACE could in principle be replaced by a time-invariant alternative (TISK). This raises the questions of whether there is indeed any kind of evidence for time-invariant phonological representations in the brain, above and beyond considerations of parsimony, and whether one could find predictions that would allow us to uniquely distinguish between the time-invariant and time-specific candidate models. We now address these two questions.

### Neural evidence for time invariant spoken word recognition?

Researchers interested in the neural representations for visual words are blessed with the Visual Word Form Area, a well-defined region in the brain that sits at the top of what is still known as the ventral visual stream, and is demonstratively the locus of our ability to read words (Gaillard et al., 2006), but critically not to hear them. Until recently, the common wisdom was that by the mere virtue of its situation in the brain – if not by its purported hierarchical architecture with increasingly large receptive fields – the VWFA was bound to achieve complete location invariance for word stimuli. However recent fMRI studies show that, and computational modeling explains why, a significant degree of sensitivity to location is present in the VWFA Rauschecker et al. (2011). A trained, functional model of location invariance for visual words explains why this can be so: in this model the conflicting requirements for location invariant and selectivity conspire with limited resources, and force the model to develop in a symmetry network with broken location symmetry on its weights. This in turn produces “semi-location invariant” distributed activity patterns, which are more sensitive to location for more confusable words (Hannagan, Dandurand & Grainger, 2011). Thus brain studies have already been highly informative and have helped constrain our thinking on how the brain recognizes visual words.

But attempts to proceed in the same way for the auditory modality quickly run into at least two brick walls. The first is that there is no clear homologue of the VWFA for spoken words. This might be because the speech signal varies in more dimensions than the visual signal corresponding to a visual object; a VWFA homologue for speech might need to provide invariance not just in temporal alignment, but also across variation in rate, speaker characteristics, etc. While there have been reports of hints of such invariance and/or multidimensional sensitivity in the superior (Salvata et al., in press) and medial (Chandrasekaran et al., 2011) temporal

gyri, a VWFA homologue for speech has not yet been detected.

The second is that paradigms for testing time-invariance are less easily designed than those which test location-invariance in the visual case. Varying on Rauschecker et al. (2011) however, we can propose a task that tests for the presence of time-specific word representations, in which subjects would be presented with a sequence of meaningless sounds where one spoken word would be embedded. By manipulating the position of this word in the sequence, one could then test whether a “blind” classifier could be trained on the evoked fMRI activation patterns to discriminate which activation patterns correspond to which positions-in-time. A clear demonstration that a classifier is unable to “blindly” map phonological patterns to position-in-time would be good evidence for the model we have introduced. In the alternative scenario, a successful blind classifier would be a smoking gun for this model. Following on our work in the visual modality, we would then need to consider a revised version with limited and shared units that could possibly achieve semi-time invariant representations.

### Specific predictions

A specific prediction of this model concerns the treatment of repeated phonemes in a word. As we have seen, the TRACE model deals with both cases by assigning activation to different time-specific units, whereas the model we have introduced must activate for instance the same “na” unit in “banana” at two different times. Finding evidence against this central feature would plainly falsify the model. However it is still unclear at this point how this would really manifest in the model (for instance would words with repeated diphones such as “banana” get more activation from the diphone level than in the TRACE model?). In fact one critical test for the current model will reside in its ability to handle such inputs in a way that is consistent with humans. If the expected differences with TRACE are indeed obtained, experimental evidence could then be gathered with the “visual word paradigm” by presenting targets and distractors with or without repeated diphones. Similarly, one would expect the same phenomena to be within reach of empirical investigations for repeated words in a sentence.

## Conclusions

We have presented a computational model of spoken word recognition (TISK) that achieves a close-to-perfect word recognition rate (98%), while also exhibiting the ability to account for basic aspects of phonological competition – the time course of cohort and rhyme effects. This time-invariant alternative uses vastly (orders of magnitude) less computational resources than its time-specific counterpart, TRACE, the economy in number of units being inversely proportional to the number of time steps allowed as input and (in TRACE) memory at all levels or (in our model) at the phoneme level. A notable property of the model is that the computational mechanisms involved – string kernels and symmetry networks – are exactly the same as have been

proposed in the visual word recognition literature, paving the way to a possible unified account of word recognition across modalities. Finally we have pointed to where we think specific predictions of the model should arise, and we have put forward a new task that makes the model more easily falsifiable.

### Acknowledgments

This work was conducted under the European Research Council grant #230313 awarded to Jonathan Grainger.

### References

- Alloppenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition: evidence for continuous mapping models. *Journal of Memory and Language*, 38, 419 – 439.
- Chandrasekaran, B., Chan, A.H.D., & Wong, P.C.M. (2011). Neural processing of what and who information during spoken language processing. *Journal of Cognitive Neuroscience*, 23(10), 2690-2700.
- Cohen, L., Dehaene, S., Naccache, L., Lehericy, S., Dehaene-Lambert, G., Henaff, M., et al. (2000). The visual word-form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Cognition and Brain Theory*, 123, 291 – 307.
- Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, 42, 317 – 367.
- Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, 16, 507-534.
- Dandurand, F., Grainger, J., & Dufau, S. (2010). Learning location invariant orthographic representations for printed words. *Connection Science*, 22(1), 25 – 42.
- Dandurand, F., Hannagan, T., & Grainger, J. (2010). Neural networks for word recognition: Is a hidden layer necessary? In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 688-693.
- Gaillard, R., Naccache, L., Pinel, P., Clémenceau, S., Volle, E., Hasboun, D., et al. (2006). Direct intracranial, fMRI, and lesion evidence for the causal role of left inferotemporal cortex in reading. *Cognition and Brain Theory*, 50(2), 191 – 204.
- Gales, M. J. F. (2009). Sequence kernels for speaker and speech recognition. In Proc. Technology Workshop at Johns Hopkins University, Baltimore.
- Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple readout model. *Psychological Review*, 103, 518 – 565.
- Grossberg, S., & Myers, C. W. (2000). The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review*, 107, 735 – 767.
- Hannagan, T., Dandurand, F., & Grainger, J. (2011). Broken symmetries in a location invariant word recognition network. *Neural Computation*, 23 (1), 251–283.
- Hofmann, T., Schölkopf, B., & Smola, A. J. (2007). Kernel methods in machine learning. *Annals of Statistics*, 36, 1171 – 1220.
- Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114, 1 – 37.
- Magnuson, J. S., Strauss, T. J., & Harris, H. D. (2005). Interaction in spoken word recognition models: Feedback helps. In *Proc. Ann. Cognitive Science Society*.
- McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, 18, 1 – 86.
- McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Science*, 10, 363 – 369.
- McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. an account of basic findings. *Psychological Review*, 88, 375 – 407.
- McQueen, J., Norris, D. & Cutler A. (2006). Are there really interactive processes in speech perception? *Trends in Cognitive Sciences*, 10(12), 533.
- Mirman, D., McClelland, J. L., & Holt L.L. (2006). Theoretical and empirical arguments support interactive processing. *Trends in Cognitive Sciences*, 10(12), 534.
- Rauschecker, A. M., Bowen, R. M., Parvizi, J., & Wandell, B. A. (2011). Position-sensitivity in the VWFA measured using fMRI pattern-classification and intracranial recordings in humans. In *Society for Neuroscience Proc.*
- Rey, A., Dufau, S., Massol, S., & Grainger, J. (2009). Testing computational models of letter perception with item-level ERPs. *Cognitive Neuropsychology*, 26, 7 – 22.
- Salvata, C., Blumstein, S. E., & Myers, E. B. (in press). Speaker invariance for phonetic information: An fMRI investigation. *Language & Cognitive Processes*.
- Shawe-Taylor, J. (1989). Building symmetries into feedforward networks. In *Proceedings of the First IEE Conference on Artificial Neural Networks*, London.

# N-back Performance: Comparing Assessment and Training Performance

**J. Isaiah Harbison (jiharb@umd.edu)**

Center for Advanced Study of Language and Department of Psychology, University of Maryland  
7005 52<sup>nd</sup> Avenue, College Park, MD 27642 USA

**Sharon M. Atkins (smatkins@umd.edu)**

Neuroscience & Cognitive Science Program and Department of Psychology, University of Maryland  
Biology/Psychology Building, College Park, MD 27642 USA

**Michael R. Dougherty (mdougherty@psyc.umd.edu)**

Department of Psychology and Center for Advanced Study of Language, University of Maryland  
Biology/Psychology Building, College Park, MD 27642 USA

## Abstract

Despite its frequent use, much is unknown about how the n-back task is performed and how it relates to working memory. We conducted a detailed analysis of the accuracy and reaction time data from a 4-back version of the task and compared the results with previous results from an adaptive training version of the task. The experiment was also designed to test the novel predictions of a computational model of n-back performance. The assessment results were largely consistent with both the training data and the model predictions.

**Keywords:** working memory; executive functioning; n-back; working memory training; computational model.

## N-back and Cognition

The n-back task is used both to measure (Owen et al., 2005) and improve (Jaeggi et al., 2008) working memory (WM). It is considered a memory updating task, and updating is thought to be a core component of working memory (Miyake et al., 2000). However, the task is not consistently or strongly correlated with performance on complex working memory span tasks, such as operation span or reading span (Kane et al., 2007). Furthermore, despite transfer to measures of fluid intelligence, n-back training has not been found to transfer to other measures of WM (Jaeggi et al., 2008; Li et al., 2008).

To better understand n-back performance and its relation to WM, the present study provides a detailed analysis of 4-back data. This study builds on a previous analysis of an n-back training task (Harbison, Atkins, & Dougherty 2011) by testing if the results from an adaptive, training version of the n-back task are replicated in a non-adaptive, assessment version of the task. The present study also tests new predictions made by the computational model of n-back performance based on that training data (Harbison et al., 2011).

## The N-back Task

In the n-back task participants are presented with a sequence of stimuli (e.g., letters). As each stimulus is presented, participants are asked to compare the current stimulus with the stimulus that occurred  $n$  items prior in the sequence. For example, in the 4-back version of the task, participants

might be presented with the letter sequence “H-G-S-M-L-T-...”. If the next letter in the sequence is “S” then participants should respond “target” as the current letter matches the letter occurring four letters prior. If the next letter is anything else, then the correct response is “non-target”. Not all non-matching letters are the same in terms of difficulty. Lures, stimuli that match an item near to but not at the target location, are more difficult than fillers (stimuli that are neither lures nor targets). Participants are less accurate and take longer to respond to lures relative to fillers (Gray, Chabris, & Braver, 2003; Harbison et al., 2011; Kane et al., 2007; McCabe & Hartmen, 2008; Oberauer, 2005). From the example, the letters “H”, “G”, “M”, and “L” are lures. They match the 6<sup>th</sup>, 5<sup>th</sup>, 3<sup>rd</sup>, and 2<sup>nd</sup> letter back, respectively, but not the 4<sup>th</sup> letter back. Letters such as “F”, “P”, and “R” are fillers.

In the training version of the n-back task the level of  $n$  varies as a function of participant performance. The  $n$  level is increased when participants perform well and decreased when participants perform poorly at their current  $n$  level. In contrast, assessment versions of the task are non-adaptive; participants are given a set number of trials at predetermined levels of  $n$ .

## Previous Results

Performance on the n-back task is not often the focus of the experiments in which the task is used. Instead, the n-back task is either used to measure or to improve WM. Therefore, despite its frequent use, there remains a lack of detailed data on n-back task performance (for exceptions see Gray et al., 2003; Kane et al., 2007; McCabe & Hartmen, 2008; Oberauer, 2005).

Previously, we (Harbison et al., 2011) identified four results that characterize n-back training task performance. First, accuracy for target trials varies as a function of serial position. Figure 1a shows the results for sequences of 4-back from the training data; participants demonstrated primacy for target trials whereas this effect was weak to non-existent for lure and filler trials. Here the lures were one position away from the target, so they matched either the 3<sup>rd</sup> or 5<sup>th</sup> back stimuli. Second, in the reaction time (RT) data, we found that participants were faster making correct than

incorrect responses on lure and filler trials. This was not found for target trials. Figure 2a shows the mean RT data from 4-back sequences of the training data. Third, correct responses to targets and lures were made at approximately the same rate. Fourth, and perhaps least surprising, we found that participants made correct responses more quickly to filler stimuli than to either targets or lures. While only the results from 4-back are shown, the results are generally consistent across n levels of 3- to 7-back in the training data, with minor discrepancies at 1-, 2-, and 8-back.

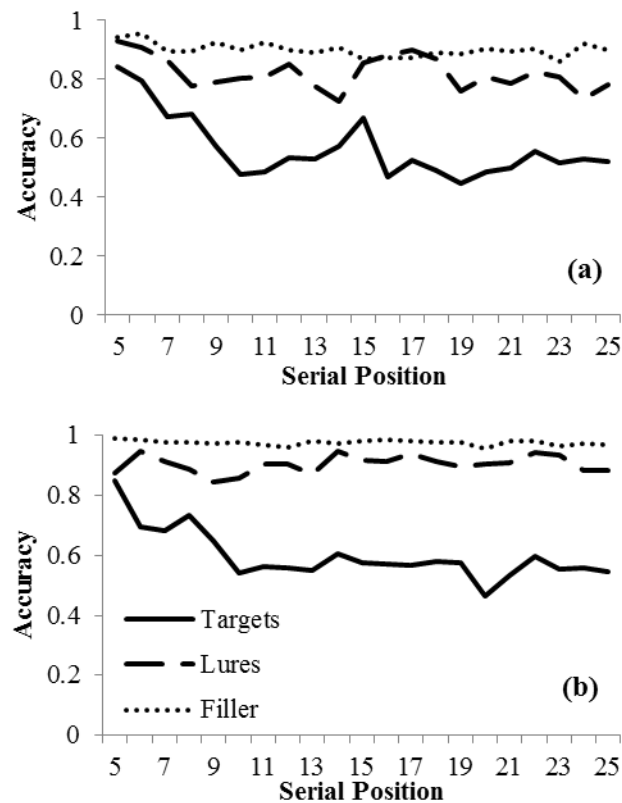


Figure 1. Participant (a) and Model (b) accuracy across serial position from the 4-back performance in a training experiment (Harbison et al., 2011).

We developed a two-process model of recognition to account for these accuracy and RT results (Harbison et al., 2011). The model assumes that when each stimulus in a sequence is presented, participants first generate an estimate of familiarity. If the stimulus is not familiar, the response is “non-target”. If the stimulus is familiar, then an attempt is made to determine if it does indeed match the stimulus occurring n items back through the process of recollection. If the recollected item matches the current stimulus, then a “target” response is made. If the recollected item does not match, then the “non-target” response is made. Finally, if recollection fails, the model guesses. RT predictions are based on the number of processes necessary to respond (familiarity = 1, familiarity and recollection = 2, familiarity, recollection, and guessing = 3). The model’s performance

on the 4-back training stimuli is shown in Figures 1b and 2b. The model captures the main qualitative patterns observed in the participant data. For example, according to the model the observed primacy for targets is due to the interference of previous items in the sequence on the maintenance of subsequent items (i.e., proactive interference). While both targets and lures are reliant on the same processes, familiarity and recollection, primacy is predicted more for targets as participants are expected to be much more likely to guess “non-target” than “target” when recollection fails as targets are much less frequent (only 20% of the stimuli are targets) Therefore, guesses are most likely to lead to correct responses for lure trials and incorrect responses for target trials.

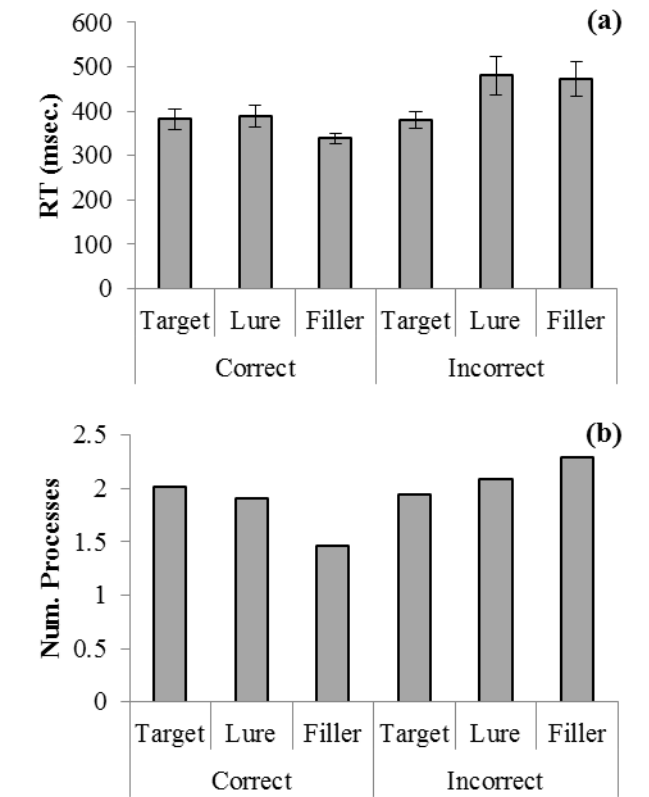


Figure 2. Participant (a) and Model (b) reaction time data from 4-back performance in a training experiment (Harbison et al., 2011).

### N-back Model Details

The n-back model is implemented within the HyGene framework (Thomas et al., 2008) and consists of three components: a representation of the current stimulus, the active subset in memory, and semantic memory. Stimuli are represented by a concatenation of an item vector and the current context vector. The elements in both the item and context vectors can take on the value of 1, -1, or 0. Here 0 represents lack of information about a feature, 1 indicates a feature’s presence and -1 its absence. Each item’s representation and the initial context vector are generated

randomly. However, the current context changes with each new stimulus. Specifically, when a new stimulus is presented each element of the current context has some probability of randomly changing to a new value. This probability is a parameter in the model (pDrift).

**Familiarity** The first step in processing a new stimulus is judging its familiarity to items in the active subset by

$$S_i = \frac{\sum_{j=1}^M P_j T_{ij}}{N_i}, \quad \text{Eq. 1}$$

where  $S_i$  is the similarity of the probe (P) and the  $i$ -th trace in memory ( $T_i$ ).  $j$  is the index of the element in the item representation for both the probe and the trace.  $N_i$  is the number of elements that are non-zero in the trace, the probe, or both.  $M$  is the number of traces in the active subset. The similarity is cubed to calculate the activation ( $A_i$ ) of each trace. Finally, the activations of all the traces in the active subset are summed to get the echo intensity for the probe. If the echo intensity is less than or equal to 0, then the item is unfamiliar and the “non-target” response is made. Otherwise, the model moves to the recollection process.

**Recollection** The model attempts to recollect the stimulus that occurred in the  $n$ -th back location when the current stimulus is familiar. This is performed by first trying to reinstate the  $n$ -th back context. Each element in the current context is changed to the  $n$ -th back context with some probability, pReinstate. This is the second parameter in the model.

Next, the (partially) reinstated context is used to probe the active subset. Equation 1 is again used but now the context portion of the representation serves as the probe instead of the item portion. Also, instead of summing the activations to get the echo intensity, the activations are used to create an echo content, a noisy representation of the item that occurred with the  $n$ -th back position by

$$C = \sum_{i=1}^M A_i T_{ij}. \quad \text{Eq. 2}$$

To identify the item from the noisy representation, the model uses the item representation from the echo content as the probe for activating the item representations stored in semantic memory. Again the results of equation 1 are used to generate the similarity and the activation, but this time semantic memory is probed and instead of using the activations to generate echo intensity or echo content, the activations are used to determine the probability of sampling and recovery from semantic memory. Specifically, the probability of sampling an item in semantic memory is calculated by

$$p(\text{Sample}_i) = \frac{A_i}{\sum_{j=1}^W A_j}, \quad \text{Eq. 3}$$

where  $W$  is the number of items in semantic memory. Therefore, the probability of sampling an item in semantic memory is equal to its relative activation. After sampling an item, an attempt is made to recover that item. Recovery is successful if the activation of the sampled item is greater than the threshold tRetrieval, the third parameter in the model. If the recovered item matches the current stimulus, then the response is “target”. If it does not match, the response is “non-target”. If retrieval fails then the model guesses.

**Guessing** The probability of guessing target is equal to the base rate probability of targets in the sequence. This probability was .2 in both the training study and in the present experiment.

**Encoding** After a response is made the current stimulus is encoded by the model. The representation of the item and the current context are stored in the active subset of memory. Each item in the active subset competes with every other item. Specifically, each feature in an item’s representation can only be non-zero for one item in the active subset. This assumption is based on the process of overwriting (Oberauer & Lewandowsky, 2008). To reduce competition, the model attempts to remove irrelevant items. In the case of 4-back, any item that occurred more than 4 items prior, from the active subset is irrelevant. Each time a new stimulus is encoded an attempt is made to remove all the irrelevant items currently in the active subset of memory. The probability of removing irrelevant items is the final parameter of the model, pRemove.

## Limitations of Previous Results

The results from the previous training study provided a starting point but there are a number of reasons why a replication and extension is needed. The present study is motivated by a desire to get cleaner data than is acquired from training studies. In training versions of the  $n$ -back task, the level of  $n$  fluctuates as a function of participant performance. Therefore, the amount of data that each participant provides for each level of  $n$  can vary substantially. For example, in the previous training study some participants never reached 4-back (i.e., were never successful enough at 2- and 3-back to reach 4-back). Some participants quickly advanced past 4-back to get to higher levels of  $n$ . Finally, some participants were stuck at 4-back for a while, as their accuracy was not high enough for  $n$  to increase or low enough for  $n$  to decrease. More generally, at lower levels of  $n$ , the majority of data is from participants that have the most difficulty performing the task. At higher levels of  $n$ , there is only data from participants that either excelled at the task from the beginning or participants that improved and are near the end of their training.

Another limitation of the reported training data was that it was drawn from a larger WM training study in which participants performed a number of different WM and WM-related training tasks and assessments. Extensive practice on

these tasks might have changed how they approached the n-back task.

In addition, the n-back model makes a number of predictions that are not tested by the previous data. First, it predicts gradual improvement in accuracy as lures move further from the target position. Lures one away from the target position (3- and 5-back when n is 4) should be more difficult than lures two positions away (2- and 6-back). Furthermore, lures the same distance from the target position are predicted to have the same approximate difficulty ( $n+2$  lures =  $n-2$  lures,  $n+1$  lures =  $n-1$  lures). The predictions are shown in Figure 3b. These predictions, like all other predictions presented, are made using the same parameter values as used in Harbison et al. (2011) for matching the training data ( $p\text{Drift} = .33$ ,  $p\text{Remove} = .15$ ,  $p\text{Reinstate} = .75$ ,  $t\text{Retrieval} = .10$ )

Second, unlike accuracy predictions, RT predictions are not symmetric around the target position. RTs for lures closer to the current stimulus should take longer to respond to correctly than lures further away from the current stimulus. That is, lures that match the 2-back position should take longer to reject than lures in the 6-back position. In contrast, the time it takes to make incorrect responses to 2-back and 6-back lures should not differ. These predictions are shown in Figure 5b.

We conducted a new experiment in which all participants had extensive experience at a moderately high level of n, 4-back. 4-back was chosen because in the training study most participants were able to reach that level, 4-back allowed lures two positions away that were not the immediately prior stimulus (2-back), and because the previous 4-back data showed the same reaction time profile as was shown at higher levels of n. This pattern was not as consistent at lower levels of n, specifically 1- and 2-back.

## Experiment

One hundred and forty-seven participants were randomly assigned into one of two counterbalanced conditions which determined if the participants performed sequences with lures first or second. Seventy-four participants were in the lure-first condition, seventy-three lure-second. Both conditions performed 16 sequences with lures and 16 sequences without lures. Each sequence was 25 letters long and contained five targets and either eight or zero lures. When the lures were present, there were two of each type in the sequence (2-, 3-, 5-, and 6-back lures). After completing the 4-back task, participants performed the block span and letter-number sequencing (LNS) tasks as measures of WM (Atkins et al., 2009).

## Results

Note that all differences reported have a p value of .05 or less. Also, unless otherwise noted, within-participant analyses were used. As such, the figures showing results averaging over participants can be misleading. Finally, there were no significant differences due to condition assignment

(lures first or lures second). Therefore, order is ignored in the reported analyses.

**Accuracy** The mean accuracy data by trial type is shown in Figure 3a. With or without lures, participants were most accurate with filler items and least accurate with target items. Performance on lures two away from the target (2- and 6-back) was worse than filler and better than performance on lures one away (3- and 5-back) from the target position. There was not a significant difference between lures the same distance away from the target. Comparing performance on sequences with lures against sequences without lures, there was not a significant difference in target performance, but participants were significantly better on filler items when there were lures.

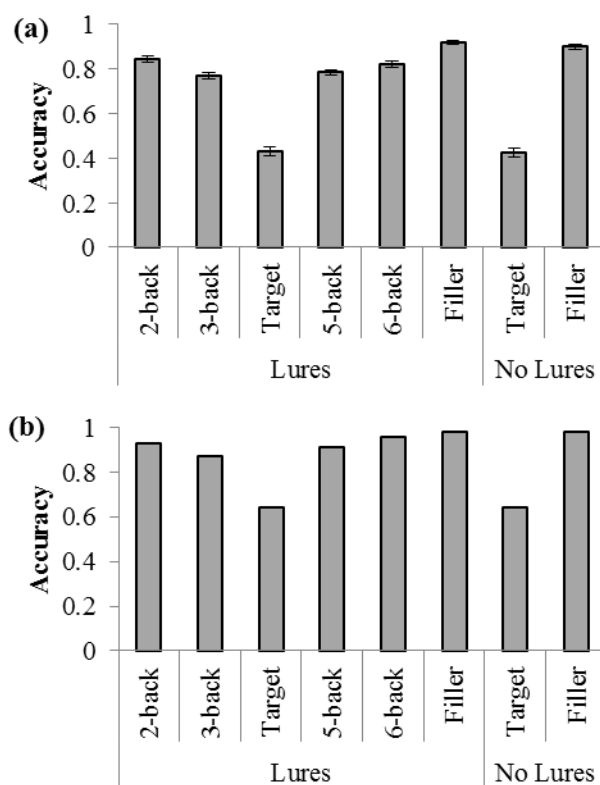


Figure 3. Participant (a) and Model (b) mean accuracy by trial type.

**Serial Accuracy** As shown in Figure 4a, participants showed primacy for target trials. Also, performance on lures two away from the target position were consistently better than performance on lures one away from the target position. Note that while target accuracy dropped below 50% in middle and later serial positions, this is not really chance performance, as participants would be expected to respond “target” only 20% of the time if they guessed “target” with the same probability as targets in the sequence.

**Reaction Times** As in the training data, participants were significantly faster to respond correctly to lure and filler items than they were to respond incorrectly, as shown in Figure 5a. In contrast, target RT was not significantly different for correct and incorrect responses. Also as in the training data, participants were quickest to respond to filler items correctly.

There was not a significant difference between 6-back and 2-back lures for incorrect responses, but there was for correct responses. This pattern of results was predicted by the model. However, there were also some inconsistencies with the previous data. Inconsistent with both the training data and the model's predictions, the present experiment found incorrect filler responses were faster, not slower, than the incorrect responses to lures, on average. Also, participants were quicker to respond to target items than predicted by the model. Both correct and incorrect target responses were significantly faster than the average lure responses.

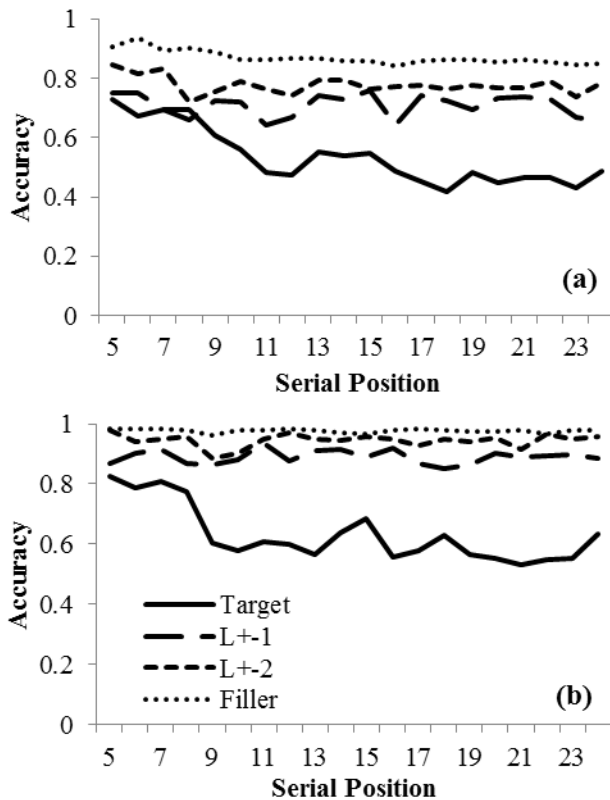


Figure 4. Participant (a) and Model (b) serial accuracy results by trial type.

**Working Memory** There was a weak but significant correlation of both LNS and block span with target performance ( $r$ 's from 0.188 to 0.283). Lure and filler accuracy were not correlated with these WM measures ( $r$ 's < 0.135). This result is consistent with previous assessment versions of the n-back Oberauer (2005) but not previous

training data (Harbison et al., 2011) which found the relationship with lure but not target performance.

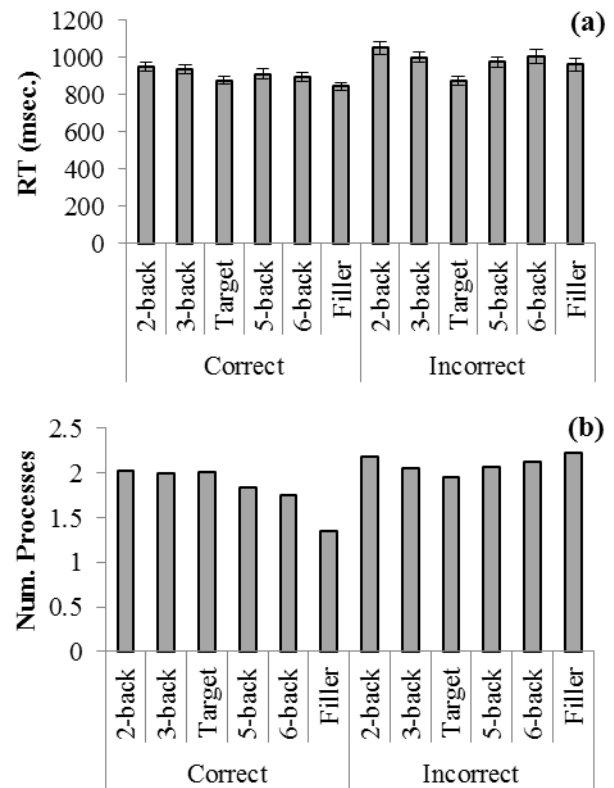


Figure 5. Participant (a) and Model (b) reaction time results by trial type and accuracy.

## Discussion

The results from the 4-back task are largely consistent with the results of the adaptive, training version of the n-back task where difficulty is adjusted based on participant performance. In the present experiment three of the four results were replicated: target accuracy showed primacy, incorrect responses took longer than correct responses for lure and filler stimuli, and correct responses to filler items were faster than responses to any other trial type. In these ways the results are consistent with both the training data and the n-back model that was based on the training data.

In addition, the new data supported two novel predictions made by the n-back model. First, lure accuracy fit the predicted pattern, with lures one away from target position being more difficult than lures two away from the target position, while lures the same distance away were performed with approximately the same accuracy. Second, reaction times were predicted by the model to be longer for correct responses to 2-back than 6-back lures despite equivalent accuracy and equivalent RT for incorrect responses. It should be noted that the model was constructed using training results with lures only in positions one away from the target position, and yet was able to accurately



predict performance on lures two away from the target position both in terms of accuracy and reaction time.

The results were not without discrepancies. Participants in the present experiment were faster at responding both correctly and incorrectly to targets than either found in the training data or predicted by the model. Also, incorrect filler responses were not the slowest overall responses in the present data. Instead, incorrect lure responses were the slowest. Both of these results, plus the fact that participants showed a different RT profile in the training version of the n-back task at n levels of 1 and 2 from the general trend found at n levels 3 and above indicate the model is at best incomplete. One natural extension of the model which could account for at least some of these results is to include the area of direct access in addition to the activated subset of long-term memory currently implemented (Cowan, 1988). Items in the area of direct access would be able to forgo the recollection process, as they would be immediately available.

The present study provides additional support for the account of n-back performance as driven by recognition processes. Both target (Oberauer, 2005; the present study) and lure (Harbison et al, 2011) performance have been found to correlate with other WM assessments. Both of these trial types rely on recollection according to the present model. In contrast, filler trial performance can be accounted for by familiarity alone and has not been found to be related to other measures of WM. This could account for the inconsistent and/or weak relationship between overall n-back performance and other measures of WM as a large portion, often more than half, of the stimuli in a given n-back sequence are filler trials.

The purpose of this study is to improve the understanding of the cognitive mechanisms behind performance on the n-back task. As with other working memory tasks which correlate with many higher level cognitive processes, it is important to determine what is being measured by the n-back assessment and what might be improved by training versions of this task (Shipstead, Redick, & Engle, 2012). The results suggest that the relationship between n-back performance and other measures of working memory are dependent on a specific process, recollection, or the ability to comply with the demands of the task and inhibit responses based on familiarity alone in order to use recollection as the basis for response.

### Acknowledgments

This research was supported in part by the University of Maryland Center for Advanced Study of Language with funding from the Department of Defense.

### References

Atkins, S. M., Harbison, J. I., Bunting, M. F., Tuebner-Rhodes, S., & Dougherty, M. R. (2009, November). *Measuring working memory with automated block span and automated letter-number sequencing*. Poster

presented at the 50<sup>th</sup> Annual Meeting of the Psychonomic Society.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin*, 104, 163-191.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2002). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, 6, 316-322.

Harbison, J. I., Atkins, S. M., & Dougherty, M. R. (2011). N-back training task performance: Analysis and model. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33<sup>rd</sup> Annual Conference of the Cognitive Science Society* (pp.120-125). Austin, TX: Cognitive Science Society.

Jaeggi, S. M., Buschkuhl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 6829-6833.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 615-622.

McCabe, J., & Hartman, M. (2008). Working memory for item and temporal information in younger and older adults. *Aging, Neuropsychology, and Cognition*, 15, 754-600.

Miyake, A., Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A., Wager, T. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, 41, 49-100.

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology: General*, 134, 368-387.

Oberauer, K. & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review*, 115, 544-576.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, 25, 46-59.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012, March 12). Is working memory training effective? *Psychological Bulletin*. Advanced online publication. doi:10.1037/a0027473.

Thomas, R. P., Dougherty, M. R., Sprenger, A., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, 115, 155-185.

# Pedagogical agents that support learning by explaining: Effects of affective feedback

Yugo Hayashi ([yhayashi@fc.ritsumei.ac.jp](mailto:yhayashi@fc.ritsumei.ac.jp))  
Mariko Matsumoto ([is039081@ed.ritsumei.ac.jp](mailto:is039081@ed.ritsumei.ac.jp))  
Hitoshi Ogawa ([ogawa@airlab.ics.ritsumei.ac.jp](mailto:ogawa@airlab.ics.ritsumei.ac.jp))

College of Information Science and Engineering, Ritsumeikan University,  
1-1-1 Nojihigashi, Kusatsu, Shiga, 525--8577, Japan

## Abstract

The present study investigates how a conversational agent can facilitate explanation activity. An experiment was conducted where pairs of participants, who were enrolled in a psychology course, engaged in a task of explaining to their partners the meanings of concepts of technical terms taught in the course. During the task, they interacted with a conversational agent, which was programmed to provide back-channel feedbacks and metacognitive suggestions to encourage and facilitate conversational interaction between the participants. Results of an experiment suggested that affective positive feedbacks from conversational agent facilitate explanation and learning performance. It is discussed that a conversational agent can play a role for pedagogical tutoring and triggers a deeper understanding of a concept during an explanation.

**Keywords:** pedagogical agents; explanation activities; affective learning.

## Introduction

The ever-evolving information and communication technology has made it possible to support human cognition by using systems which aids human interaction. Many researchers in computer science are tackling on the theme of developing embodied conversational agents to support education. Recent studies on cognitive science and learning science show that collaborative learning facilitates understanding or acquisition of new concepts depends greatly on how explanations are provided. In this study a collaborative activity of making explanation is experimentally investigated by using a conversational agent that serves as a teaching assistant. The goal of the experiment is to find out what kind of feedback from the agents is most conducive to successful learning performance.

## Related work

### Explanations during collaborative activities

Number of studies on collaborative problem solving in cognitive science revealed how concepts are understood or learned. Researchers have shown that asking reflective questions for clarification to conversational partners is an effective interactional strategy to gain a deeper understanding of a problem or a concept (e.g. Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Miyake, 1986; Salomon,

2001; Okada & Simon, 1997). It has also been demonstrated that the use of strategic utterances such as asking for explanation or providing suggestions can stimulate reflective thinking and meta cognition involved in understanding a concept. Playing different roles during explanation is also said to help problem solvers reconstruct external representation and concepts (Shirouzu, Miyake, & Masukawa; 2002).

Studies that are discussed above suggest that how well one can explain is the key to understanding and learning of a concept. However, explanation becomes successful if people have difficulties in retrieving and associating relevant knowledge required for explanation activity. Researches on collaborative learning have reported that these difficulties rise among novice problem solvers (Coleman, 1998; King, 1994). Also, it may not help learn a concept if people cannot communicate with each other as in when, for example, they use technical terms or phrases unknown to others (Hayashi & Miwa, 2009).

It is assumed that one of the ways to help collaborative problem solvers is to introduce a third-person or a mentor who can facilitate the task by using prompts such as suggestions and back-channels. However, it is often difficult for one teacher to monitor several groups of collaborators and to supervise their interaction during explanation in actual pedagogical situations. Recently there are studies which demonstrate that the use of conversational agents that act as educational companions or tutors can facilitate learning process (Holmes, 2007; Baylor & Kim, 2005). Unfortunately, it has not been fully understood if and what kinds of support by such agents would be more helpful for collaborative learners. In this paper, the author will further investigate this question through the use affective expressions.

### Pedagogical conversational agents as learning advisers

Researchers in the field of human computer interaction have conducted a number of experimental studies which involve the use of pedagogical agents (e.g. Kim, Baylor & Shen, 2007; Reeves & Nass, 1996; Graesser & McNamara, 2010). One point to be taken into consideration in studies of human performance is the affective factor. This factor influences people's performance in either negative or positive ways and

several studies reported that such factors are especially important in learning activities (Baylor & Kim, 2005). For example, Bower & Forgas (2001) revealed that positive moods can increase memory performance. Mayer & Turner (2002) also demonstrated that positive state of mind can improve text comprehension.

Moods may affect the performance of human activities both verbally and non-verbally. In a study by Kim, Baylor, & Shen (2007), which examined how positive and negative comments from conversational agents affect learning performance, a pictorial image of an agent was programmed to project a textual message to the participant; in the positive condition, a visual avatar produced a short comment like "this task looks fun", while in the negative condition, it produced a short comment like "I don't feel like doing this, but we have to do it anyway". The results showed that the conversational agents that provided the participants with comments in a positive mood furnished them with a higher motivation of learning.

The studies discussed above suggest that the performance of explanation would also be enhanced if suggestions are given in positive mood either verbally or through visual feedbacks.

## Research Goal

This study investigates how conversational agents can facilitate understanding and learning of concepts. This paper will focus on an agent which has a role that assists paired participants to explain concepts to their partners during the collaborative peer-explanation activity. A natural language processing agent monitored the interaction between the participants and provided prompts to them which were generated by pre-defined rules. The research goal of this study is to understand if the use of positive expressions provided by a conversational agent facilitates collaborative learners' understanding of concepts.

## Method

### Experimental task and procedure

The experiment was conducted in a room where the computers were all connected by a local area network. Participants were given four technical terms presented on the screen. They were: 'schema', 'short-term / long-term memory', 'figure-ground reversal', and 'principle of linguistic relativity', which had been introduced in a psychology class. Along with the keyterms, a brief explanation of the concept was described by a few sentences. They were asked to describe the concepts of these words. After this pre-test, they logged in the computer and used the program installed in a USB flash drive (see the next section for detail). The pairs of participants were communicated through the chat program and one of the paired participants was instructed to explain to their partner the meanings of the words presented on their computer screen one by one. When two of the four concepts were explained to their partner, they switched the roles and the other partner explained the

rest of the two words to his/her partner. This was repeated but the words they explained the second time were different from those in the first time. All participants received the same prompts of suggestions from the agent on how explanations should be given and how questions should be asked about the concepts. After this pre-test, they took the same test in the post-test. The descriptions of the concepts they provided in the post-test were compared with those of the pre-test to analyze if the participants gained a deeper understanding of the concepts after the collaborative activity. The whole process of the experiment took approximately 80 minutes (see Figure 1).

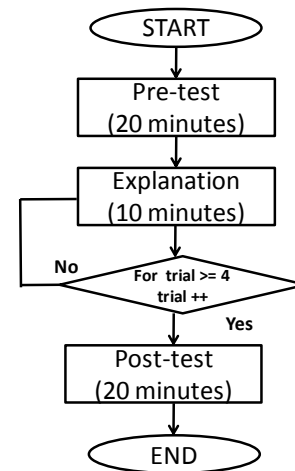


Figure 1: Experiment flow.

## Experimental system

In the experiments, a computer-mediated chat system was set up through computer terminals connected via a local network and the interactions of the participants during the activity were monitored. The system used in the experiments was programmed in Java (see Figure2).

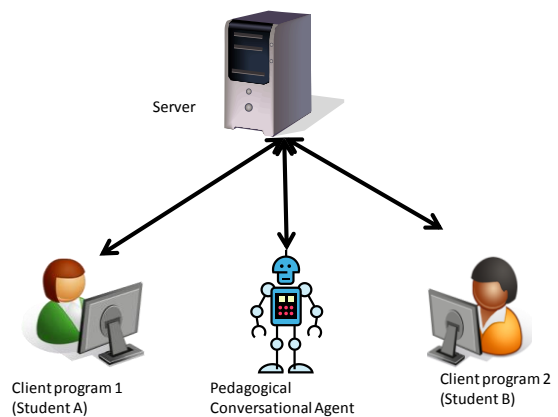


Figure 2: Experimental Setting.

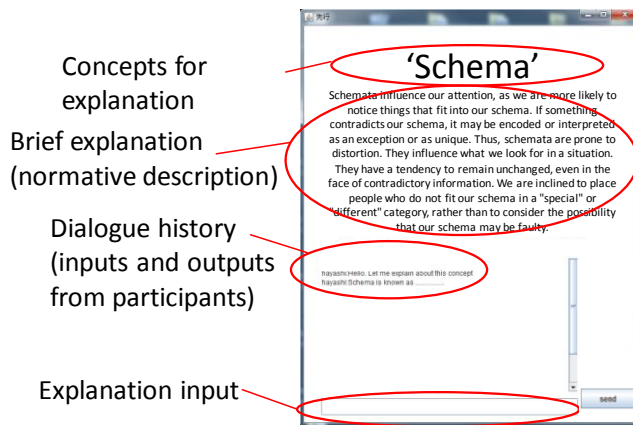


Figure 3: Screenshot of the chat system.

The system consists of three program modules of Server, Chat Clients, and Agent, all of which are simultaneously activated. Multi-threads are used so that the server program can send all messages to the clients' chat system and the agent simultaneously.

The pedagogical agent used in this study is a simple rule-based production system typical of artificial intelligence. It is capable of meaningfully responding to input sentences from users and consists of three main modules: Semantic Analyzer, Generator, and Motion Handler (see Figure 4).

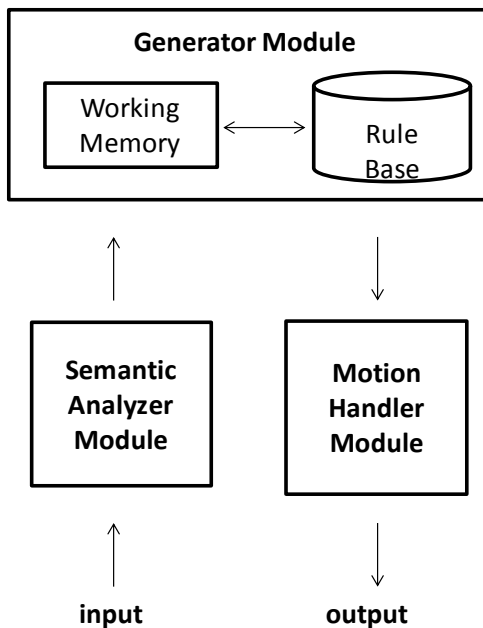


Figure 4: Architecture of message production.

Textual input of all conversational exchanges produced by paired participants is sent to the semantic analyzer of the conversation agent. The semantic analyzer then scans the text and detects keywords relevant to the concepts if they

are being used in the explanation task (e.g. "I think that a *schema* is some kind of *knowledge* that is used based on one's own *experience*." (detected key words are shown in bold italic). Next, the extracted keywords are sent to the working memory in the generator and processed by the rule base, where various types of rule-based statements such as 'if X then Y' are stored to generate prompt messages (if there are several candidates of matching statements for the input keywords, a simple conflict-resolution strategy is utilized). When the matching process is completed, prompt messages are selected and sent back to the working memory in the generator. The messages generated by the rule base are also sent to the motion handler module to activate an embodied conversation agent, a computer-generated virtual character which can produce human-like behaviors such as blinking and head-shaking. Each output message is textually presented in a text filed on the computer display (See next sections for details).

Several types of output messages are presented by the agent depending on the content of input text from the participants (see Table 1 below for examples). Only short back-channels are sent when there are several related key words in a text (Type 1 output); Messages of encouragement are given when the agent detects some keywords related to the target concept (Type 2 output, Type 3 output, Type 4 output).

Table 1: Types of output messages from the agent.

Type of messages	Examples
<u>Input messages</u> (Detected key words are in Bold)	"I think that a <b>schema</b> is some kind of <b>knowledge</b> that is used based on one's own <b>experience</b> ."
<u>Output: type 0</u> Back-channels	"That's the way", "Keep going!", "Um-hum"
<u>Output: type 1</u> Positive Suggestion (Used in Positive condition)	"Wow! You used a few very good keywords. That's great! It is better if you explain it from a different perspective!"
<u>Output: type 2</u> Negative Suggestion(Used in Negative condition)	"Well, you used few keywords. That is not enough. It is not satisfactory unless you explain it from a different perspective."
<u>Output: type 3</u> Normal Suggestion (Used in Neutral condition)	"You used few important keywords. Try to explain from a different perspective."

## Participants and conditions

In this study, 90 participants participated in the experiment. The participants were all undergraduate students who were taking a psychology course and participated in them as part of the course work. They were randomly assigned to three conditions, which varied with respect to how prompts of

suggestions were presented and how conversational agents were used (see the sections below for details).

To find out how affective factors influence the task of explanation, three types of avatars were created: one is the positive agent with friendly facial expression which was used for the "positive condition", and the negative agent with unfriendly facial expression which was used for the "negative condition", and finally the neutral agent with no facial expression which was used for "neutral condition". In the positive condition ( $n = 31$ ), the participants were given positive suggestions, which were synchronized with the facial expressions of the positive agent. In the negative condition ( $n = 28$ ), the participants were given negative suggestions, which were synchronized with the facial expressions of the negative agent. In the neutral condition ( $n = 31$ ), the participants were given suggestions without emotional expressions.

The messages were given through chat dialogue and the virtual character moved its head gestures while the participants chat on the computer (For examples of suggestion for the conversational agent see Table 1). Since there was odd number of participants in positive and neutral condition, one group was composed by three.

### Dependant variables

To evaluate the outcome of (1) quality of the performance of learning, and (2) interaction process, two types of measures were used.

First, for the learning performance, the results of the pre- and post- tests were compared to find out how the explanation task with different conditions facilitated their understanding or learning of the concepts. For the comparison, their descriptions were scored in the following way: 1 point for a wrong description or no description, 2 points for a nearly-correct description, 3 points for a fairly-correct description, 4 points for an excellent description, and 5 points for an excellent description with concrete examples. It was judged that the greater the difference in scores between the two tests the higher the degree of the effect of explanation.

Second, for the analysis of explanation process, all the dialogs during the task were analyzed. The main focus of the analysis was to investigate what kind of explanations were used during their interaction. Each dialog sentences that included explanations were coded by the following two categories: (1) explanations that were made by using terms and phrases presented by the system (see Figure 3 for an example of the description), and (2) explanations that were generated based on subjective inference. The former is called "normative explanations". On the other hand, the utterance in the latter is called "subjective explanations".

## Results

### Quality of performance

The results showed that the participants' understanding of the concepts improved after the explanation task in all

conditions (see Figure 5). The vertical axis in Figure 5 represents the average scores of the tests for the three groups at the times of pre- and post- tests. A statistical analysis was performed using a 2 x 3 mix factor ANOVA with the two evaluation period (the pre-test vs. the post-test) and the three conditions with different feedback (Positive vs. Negative vs. Neutral) as independent factors.

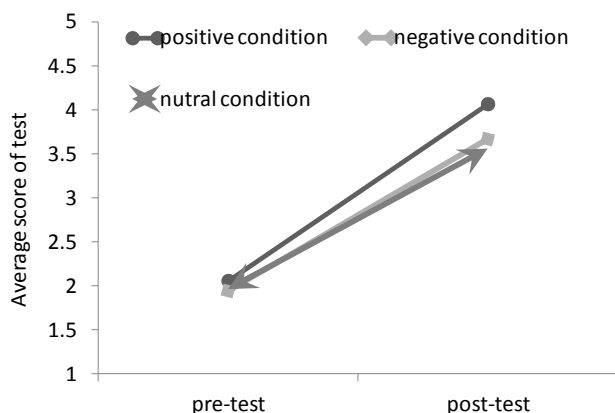


Figure 5: Results of the quality of the performance of learning.

There was significant interaction between the two factors ( $F(2, 87) = 3.388, p < .05$ ). First, an analysis of the simple main effect was done on each level of the feedback factor. In the Positive, Negative, and Neutral condition, the average scores in post-test was higher than pre-test respectively ( $F(1,87) = 254.397, p < .01$ ;  $F(1,87) = 172.796, p < .01$ ;  $F(1,87) = 155.812, p < .01$ ). Next, an analysis of the simple main effect was done on each level of the period factor. In the pre-test, there was no differences between conditions ( $F(2,174) = 0.202, p = .82$ ). Although in the post-test there were differences between conditions ( $F(2,174) = 9.094, p < .01$ ). Further analysis on the post-test was conducted using the Ryan's method. Results indicate that the average score of Positive condition was higher than Negative condition and the average score of Positive condition was higher than Neutral condition respectively ( $p < .01$ ;  $p < .01$ ). There were no differences between Negative condition and Neutral condition ( $p = .51$ ).

The overall result suggests that the collaborative activities facilitated the participants' understanding or learning of the concepts more when the positive suggestions were presented to the participants.

### Interaction process

Figure 6 indicates the relationships between the usage of normative explanations and subjective explanations. The vertical axis represents the average ratio of each participant's explanation type. The horizontal axis shows each of the three conditions.

The analysis of ANOVA with the factor of explanation type (normative explanations vs. subjective explanations)

was conducted on each condition. The results show that participants in the Neutral condition and Negative conditions used more subjective explanations than normative explanations, respectively ( $F(1, 27) = 7.326, p < .05$ ;  $F(1, 30) = 25.116, p < .01$ ). On the other hand, there were no statistical differences between the two conditions in the Positive condition ( $F(1, 30) = 0.46, p = .50$ ). These results indicate that participants in the Negative and Neutral conditions made explanations mostly based on subjective explanations.

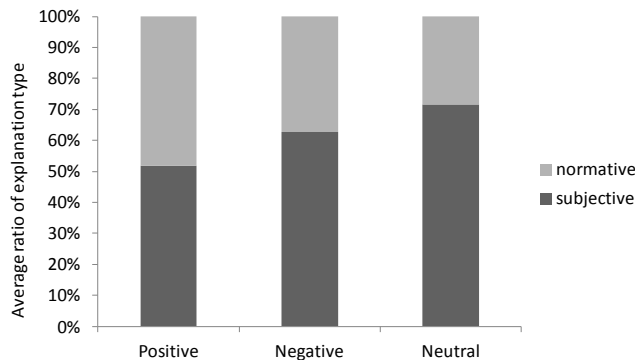


Figure 6: Results of the type of explanation activities.

## Discussion

### Affective expressions of the conversational agent

The results of the experiment suggested that the greater the positive affective expressions from the conversational agent the more it can facilitate explanation activities which leads to a deeper understanding of concepts (i.e., Positive condition > Negative condition, Positive condition > Neutral condition). The results of the dialogue analysis somewhat support these result. That is, participants in the Negative and Neutral condition used more subjective inferences and interpretations about the key concept instead of using normative phrases, which might led to construction of misunderstandings on the concepts. On the other hand, participants in the Positive condition used normative expressions that were on track. It is assumed that the affective expressions generated by the agents facilitated the participants' motivation to keep their attention to the computer system which provided important information.

These results provide more reliable findings than those compared with experiments conducted by the authors' previous work (Hayashi, 2012). In those experiments, the influences of affective feedbacks were examined during collaborative activities. Unfortunately, there was no neutral condition and dialogue analyses were not further conducted. The present study makes it clear that positive emotions expressed by a pedagogical agent facilitated explanation activities at the interaction level. This suggests that the participants might have paid more attention during the interaction process and worked harder when they received positive comments than they received neutral and negative

comments. One of the interesting finding is that, the learning performance of the participants in the Negative condition and Neutral condition were the same in this experiment. It is assumed that negative affective feedbacks were not able to trigger such motivation and enhance performance as much as the Positive condition. This may be affected by the lack of attention to the agent. This point will be further investigated elsewhere.

### Awareness towards the conversational agent

Studies in social psychology have suggested that work efficiency is improved when a person is being watched by someone, or, that the presence of an audience facilitates the performance of a task. This impact that an audience has on a task-performing participant is called the "audience effect". Another relevant concept on task efficiency, but from a slightly different perspective, is what is called "social facilitation theory". The theory claims that people tend to do better on a task when they are doing it in the presence of other people in a social situation; it implies that person factors can make people more aware of social evaluation. Zajonc (1965), who reviewed social facilitation studies concluded that the presence of others have positive motivational affects.

Holmes (2007) is one of the experimental studies which investigated the effects of a tutoring agent. In this experiment, an agent, which played the role of an assistant, was brought in to help a participant who explained a concept. In the experiment, three different environments were set up for the 'explaining activity'. They were: (1) two participants working with a text-based prompt, (2) two participants working with a visual image of pedagogical agent which produced a text-based prompt (3) one participant working with a visual image of pedagogical agent which produced a text-based prompt (in this setup, participants did not have a human co-learner and directly interacted with the agent). The result showed that the participants in the last two conditions did better than the first where only textual prompts were presented. It also showed that the participants in the second condition did not engage in the explanation activity as much as those in the third. The first finding of Holmes (2007) is that the participants in the last two conditions, who worked with the agent, performed better may be attributed to the fact that their task of explanation was being watched or monitored by the agent. Also, the second finding that the effect of the agent for the participants in a pair was not as high as for those directly interacting with it alone may be because the level of attention of the participants in the second condition was not as great as that in the third condition; it may be that the participants in the second were less conscious of the presence of the monitoring agent than those in the third group.

These results of the present study suggest that participants would do better in the task of explanation if they are more conscious of the presence of the agents or if they are given an explicit direction to pay attention to the agent. The

results of the present experiment provide new evidence that the positive feed backs made by the agents can facilitate such "audience effect".

### Conclusion and future work

The present study investigated the effectiveness of the use of a conversational agent in a collaborative activity, where paired participants explained each other the meaning of technical terms taught in a psychology class for a better understanding. Conversational agents were used to encourage and facilitate the students' interaction through both verbal and visual input. The experimental results suggested that the presence of a conversational agent with positive expressions can trigger a deeper understanding of a concept during an explanation.

Pedagogical agent can play several different roles for collaborative learning activities and several studies have looked into the effectiveness of the use of a pedagogical agent with different roles. For example, Baylor & Kim (2007) investigated the effectiveness of the use of a pedagogical agent which plays the roles of an expert teacher, a motivator, and a mentor (both an expert and motivator). However, not much is known yet about what roles it can play effectively. Another issue to be further investigated is the effect of the personality of the agent upon these roles. These and other related topics need to be further studied in future.

### Acknowledgments

I appreciate all students who participated in this experiment. I also want to thank my student advisee Shin Takii (Ritsumeikan University) and Rina Nakae (Ritsumeikan University), Yuichi Mizuno (Ritsumeikan University) for helping me to conduct the experiment.

### References

1. Baylor, A. L., & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education*, 15, 95-115.
2. Bower, H. G., Forgas, P. J. (2001). Mood and social memory. In J. P. Forgas (Ed). *Handbook of affect and social cognition*, NJ, LEA, 95-120
3. Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145-182.
4. Coleman, E. B.(1998). Using explanatory knowledge during collaborative problem solving in science. *The Journal of Learning Sciences*, 7, 387-427.
5. Graesser, A., McNamara, D. (2010). Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist*, 45, 234-244.
6. Gulz, A., Haake, M. (2006). Design of animated pedagogical agents – A look at their look, *International journal of Human-Computer Studies*, 64, 322-339.
7. Hayashi, Y., Miwa, K. (2009). Prior experience and communication media in establishing common ground during collaboration. In *Proceedings of the 31th annual conference of the cognitive science society*, 528-531.
8. Hayashi, Y. (2012). On pedagogical effects of learner-support agents in collaborative interaction. In S.A. Cerri and B. Clancey (Eds.): *Proceeding of the 11th International Conference on Intelligent Tutoring Systems (ITS2011)*, Lecture Notes in Computer Science, Springer-Verlag, Vol 7315, pp. 22-32.
9. Holmes, J. (2007). Designing agents to support learning by explaining. *Computers & Education*, 48, 523-547.
10. Kim, Y., Baylor, A. L., & Shen, E. (2007). Pedagogical agents as learning companions: The impact of agent emotion and gender. *Journal of Computer Assisted Learning*, 23, 220-234.
11. King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal*. 30, 338-368.
12. Mayer, D. K., Turner, J.C. (2002). Discovering emotion in classroom motivation research, *Educational Psychologist*. 37, 107-114.
13. Miyake, N. (1986). Constructive interaction and the interactive process of understanding. *Cognitive Science*. 10, 151-177.
14. Okada, T., & Simon, H. (1997). Collaborative discovery in a scientific domain. *Cognitive Science*. 21, 109-146.
15. Reeves, B., Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York: Cambridge University Press
16. Salomon, G. (2001). *Distributed cognition: Psychological and educational considerations*. New York: Cambridge University Press
17. Shirouzu, H., Miyake, N., & Masukawa, H. (2002). Cognitively active externalization for situated reflection. *Cognitive Science*. 26, 469-501.
18. Zajonc, R. B. (1965). Social facilitation. *Science*. 149, 271-274.



# Knowledge and Political Categorization

Evan Heit (eheit@ucmerced.edu)

Stephen P. Nicholson (snicholson@ucmerced.edu)

School of Social Sciences, Humanities and Arts  
University of California, Merced  
Merced CA 95343 USA

## Abstract

A nationally representative sample of US adults completed two political categorization tasks. The first was to identify the political parties for hypothetical candidates with information given about demographics and stands on issues. The second task was to decide whether to vote for each candidate. On the identification task, judgments about whether a person is a Democrat were almost a perfect mirror image of judgments of whether a person is a Republican. In general, respondents were very successful in the identification task; there was a strong correlation with objective probabilities. Likewise, respondents were successful at the voting task, in terms of their own party interests. Success at these two tasks was positively correlated with a measure of political knowledge. The pattern of responses was also influenced by the political party of the respondent; suggesting that feature weights depended on party membership. Implications for models of categorization and reasoning are discussed.

**Keywords:** Categorization; Expertise; Probability Judgment; Political Cognition.

## Introduction

We propose that political parties should be conceived of as categories. Following Rosch & Mervis's (1975) seminal work on categorization, political parties have a horizontal dimension corresponding to typicality structure, e.g., Mitt Romney is a more typical Republican than is Ron Paul. It is then appropriate to ask what is the function of political categories (cf., Anderson, 1991; Billman & Heit, 1988; Markman & Ross, 2003), beyond labeling individuals as party members. One key function is to support voting, which can be seen as a category-based inference, e.g., knowing that Mitt Romney is a typical Republican would lead many people to vote for him in a Presidential election.

In previous research (Heit & Nicholson, 2010) we have collected typicality judgments for a set of real political candidates. College students rated the individuals either on typicality as a Democrat or typicality as a Republican. The relation between the two sets of ratings was strong, negative, and linear, with a remarkable correlation of -0.9957. Essentially, whatever made an individual more typical of one party was seen to make that individual less typical of the other party (cf., Rosch & Mervis, 1975; Verbeemen, Vanoverberghe, Storms, and Ruts, 2001). It was not possible to be typical of both parties, or atypical of both parties. The results contrasted with other opposing

pairs of categories, male versus female jobs and healthy foods versus junk foods. We concluded that for political categories, there is a highly systematic and polarized representation of knowledge.

Although the results were extremely strong, the study itself had limitations. For example, students may not be representative of voters at large. We did not systematically study the effects of demographic variables such as level of political knowledge (which might be low for college students) and party of the respondent. Because the stimuli were simply names of public figures, we could not tell which information about these figures was being used. Also, the dependent variable, typicality, has disadvantages, because it is not objective and it may not map directly onto real political behavior such as voting.

Hence, the present experiments substantially improved upon Heit and Nicholson (2010). Each experiment involved several hundred adults from a nationally representative sample of US adults, with information collected about political knowledge and party membership. The stimuli were descriptions of hypothetical candidates in terms of demographic information (race, gender, number of children) and stands on issues (government spending and abortion). Information about each candidate's political party was omitted from the stimuli; however the objective probability of being a Democrat or Republican based on demographics and stands on issues could be determined from national survey data. In Experiment 1, the task was to identify each candidate's party. In effect, we were examining whether respondents could correctly categorize candidates as Democrats or Republicans when this information is withheld. In Experiment 2, the task was voting; respondents were asked how likely they would be to vote for each candidate. A key measure of interest was whether respondents voted the party line, i.e., Democrats voting for Democrats and Republicans voting for Republicans. In general, we were interested in whether performance on these two tasks depended on political knowledge and party membership of the respondent. We also examined the influence of various cues, to see if different cues were used for the two tasks and by different sub-groups of respondents.

In the cognitive science literature on categorization, perhaps the most closely related work addresses the effects of expertise on biological categorization. For example, Johnson and Mervis (1997) studied categorization of

songbirds, reporting shifts due to expertise. In effect, what was the subordinate level for non-experts became the basic level for experts. Medin and Atran (2004) reviewed an extensive set of studies showing effects of knowledge and group membership on biological categorization, cautioning against conclusions drawn just from Western college students. For example, Medin, Lynch, Coley, and Atran (1997) conducted a study of tree experts including taxonomists, landscapers, and maintenance workers. They reported differences in categorization and reasoning due to the goals of each type of expert. These differences appeared to be mediated by differences in feature weighting for particular areas of expertise. (See Hayes, Heit, & Swendsen, 2010, for a further review of knowledge effects on category-based reasoning.)

With regard to the issue of political knowledge, political science research has generally been pessimistic. Political scientists have emphasized that the US public is largely ignorant of politics (e.g., Delli Carpini & Keeter, 1996). The mass public is reported to have minimal levels of political attention and information as well as incoherent and unstable attitudes (e.g., Converse, 1964). Despite these low levels of information and political understanding, most citizens appear to make do with simple political heuristics such as relying on single cues, e.g., party labels (e.g., Lau & Redlawsk, 2006). Notably, in the present study, we require participants to make judgments from multiple cues, while information about party labels is withheld. Hence, we are addressing whether the pessimistic view from political science is supported.

## Experiment 1

### Method

**Participants.** A total of 598 US adults participated, in September-October 2010, as part of the Cooperative Congressional Election Study (CCES). Only self-identified Democrats or Republicans were included, and because the political knowledge measure referred to party control of

both houses of the state legislature, adults from Nebraska and Washington, DC were excluded.

**Materials.** The key stimuli were the nine hypothetical candidate descriptions; examples are shown in Table 1. Candidates were described in terms of gender, race, number of children, position on government spending, and position on abortion. The objective probability of each person being a Democrat or a Republican was determined with survey data from the 2008 American National Election Study (ANES). Across the nine descriptions, the objective probability ranged from 10% to 91% in increments of about 10%. On the survey itself, respondents were informed that each candidate was either a Democrat or a Republican. Approximately half of the respondents were asked to judge the probability that each was a Democrat; the remainder judged the probability that each was a Republican.

The knowledge measure was based on eight questions. Four questions required the respondent to correctly identify the party controlling the US Senate, the US House of Representatives, and the two legislative chambers in the respondent's home state. Four more questions asked respondents if they recognized the names of public officials (governor, two US senators, and US representative). Based on a rough median split, respondents with seven or eight correct responses were considered high knowledge, and the remainder were considered low knowledge.

**Procedure.** Respondents participated at their own pace, as part of a larger Internet-based survey.

### Results and Discussion

In a conceptual replication of Heit and Nicholson (2010), the relation between these two kinds of judgments was extremely strong, negative, and linear,  $r = -0.9977$ . (See Figure 1.) Consistent with the findings of Tversky and Koehler (1994), the results showed binary complementarity, that is, there was neither evidence for subadditivity or superadditivity, and complementary pairs summed to an average of 99.2%.

Table 1: Sample Stimuli for Experiments 1 and 2.

Description	Obj. Prob. of Democrat
Joanna is a white female with no children who enjoys watching television, exercising, and discussing politics with friends. In a recent political discussion he voiced the opinion that government provides about the right amount of services and that, by law, abortion should never be permitted.	20.9%
Emily is a white female with no children who enjoys watching television, exercising, and discussing politics with friends. In a recent political discussion she voiced the opinion that government should provide many more services and that, by law, abortion should be allowed under some circumstances.	49.5%
George is an African-American male with no children who enjoys watching television, exercising, and discussing politics with friends. In a recent political discussion he voiced the opinion that government provides about the right amount of services and that, by law, a woman should always be able to obtain an abortion.	72.7%

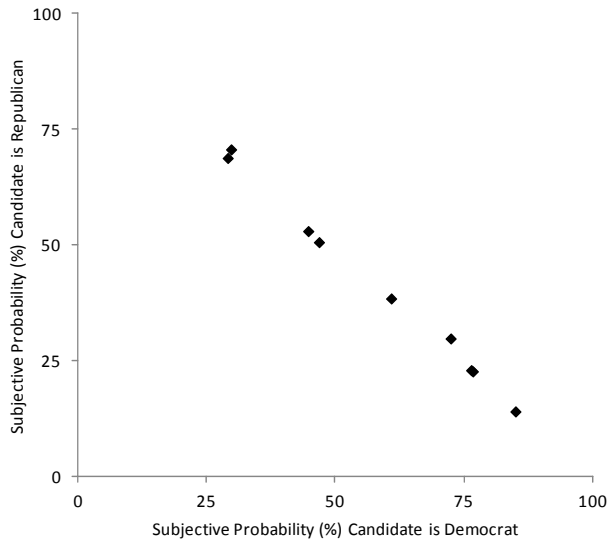


Figure 1. Subjective probability candidate is a Democrat versus subjective probability candidate is a Republican.

Figure 2 shows average subjective probability judgments plotted against objective probabilities. For this composite measure, all judgments were pooled; judgments about whether a candidate was a Republican were subtracted from 100% to put them on the same scale as judgments about whether a candidate was a Democrat. The correlation between subjective and objective probabilities is remarkable,  $r=0.9557$ , indicating that collectively, respondents were able to identify candidates' party affiliation based on very limited information. Most of the data points fall above the main diagonal, indicating that the proportion of Democrats in the stimulus set was somewhat overestimated overall. The subjective probability judgments have a somewhat smaller range than the objective probabilities.

Of course, the remarkable success of respondents at judging party membership in the aggregate need not be reflected at the individual level. Still, individual respondents were successful. The median correlation between objective and subjective probability, at the individual level, was 0.7523, and the mean correlation was 0.6203. The fact that the mean is lower than the median reflects that a small number of respondents did very poorly at this task.

The mean correlations varied as a function of knowledge and partisanship of the respondents. The mean correlations for high knowledge Democrats, low knowledge Democrats, high knowledge Republicans, and low knowledge Republicans were 0.7214, 0.5597, 0.6265, and 0.5575, respectively. A two-way ANOVA revealed a main effect of knowledge, with high knowledge respondents showing higher correlations,  $F(1, 591)=15.32$ ,  $p<.001$ . Neither the effect of party membership or the interaction between knowledge and partisanship reached the level of statistical significance.

We next conducted analyses of the cues used by respondents in each sub-group. The question addressed was what information was used in making these political categorization judgments, and whether use of information varied across groups. Essentially, we conducted four regression analyses, predicting probability judgments based on the cues of gender, race, number of children, position on government spending, and position on abortion for each sub-group. Because each respondent contributed judgments for nine items to the analysis, we used a version of the generalized linear model that accommodates clustered data (as implemented in the generalized estimating equation module in SPSS Version 18). Gender was coded 0 for male and 1 for female; race was coded 0 for white and 1 for African-American; position on government spending was arbitrarily coded as a 1, 2, or 3 with higher values indicating a more favorable position; and position on abortion was arbitrarily coded as a 1, 2, or 3 with higher values indicating a more permissive position.

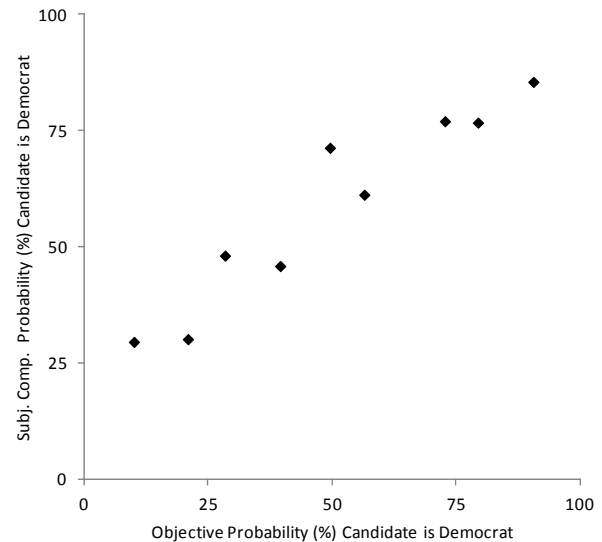


Figure 2. Objective probability candidate is a Democrat versus subjective probability candidate is a Democrat.

Before describing the findings, it is worth noting that as in the real world, within these stimuli there was multicollinearity among the cues. We had created stimuli with the aim of covering a wide range of probabilities in intervals of 10%, rather than breaking up the usual correlations. In some cases, the demographic and issue variables were strongly correlated with each other. Hence, regression coefficients should be interpreted with caution. With this point made, Figure 3 shows the standardized regression coefficients across the five cues. In general, the regression coefficients are rather similar across sub-groups. Perhaps the most interpretable difference is that stand on abortion, a highly predictive cue, has more weight for high expertise Democrats than for low expertise Democrats, and for high expertise Republicans than for low expertise Republicans. Paying more attention to this cue would lead

to greater success for the high expertise respondents at the identification task. Unexpectedly, the African-American cue shows negative weights. In fact, this cue had a strong positive correlation with identification as a Democrat. For example, in a simple regression for all respondents, predicting judgments from just the African-American cue, the standardized regression coefficient was 23.16. However, stand on abortion was correlated with African-American, and acted as a suppressor variable. In a regression with just these two predictor variables, the standardized regression coefficient for abortion is 30.60 and the coefficient for African-American drops to -13.69. Therefore, we would emphasize the similarity of regression coefficients across sub-groups, and avoid overinterpretation of specific values.

As an interim summary, we note that so far there is evidence for main effects of expertise, with higher knowledge respondents being more successful at the categorization task. There is little evidence for group (party) differences or differences in feature weighting. Overall, respondents' success at using multiple cues to identify party membership suggests a much more optimistic view than the standard view from political science, that people can, at best, make basic judgments if party label information is supplied.

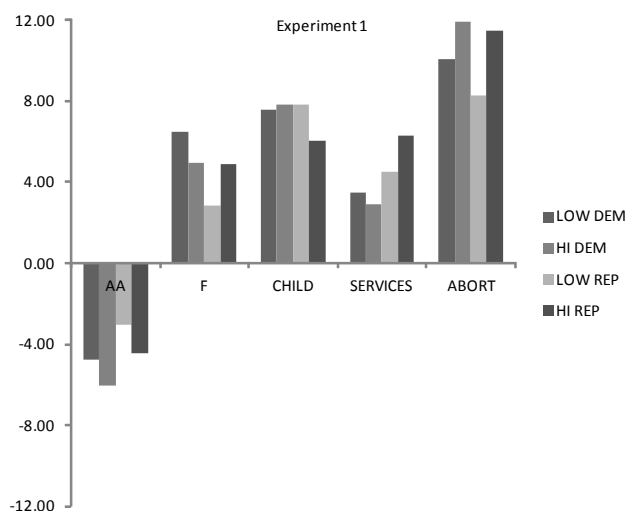


Figure 3. Estimated regression coefficients for low-knowledge and high-knowledge Democrats, and low-knowledge and high-knowledge Republicans.

## Experiment 2

Having shown in Experiment 1 that respondents can successfully identify party membership of hypothetical candidates, in Experiment 2 we investigated voting judgments on these same candidate and compared these responses to identification judgments.

### Method

From the same survey as in Experiment 1, a different set of 573 US adults participated, screened according to the same

criteria. Political knowledge was measured as in Experiment 1.

Again, the key stimuli were the nine candidate descriptions shown in Table 1. However, respondents were asked how likely they would be to vote for each candidate, on a scale from 0% to 100%.

## Results and Discussion

Figures 4 and 5 show the average voting probability judgments across the nine descriptions as a function of objective probability of being a Democrat, for respondents who identified themselves as Democrats and Republicans, respectively. For Democratic respondents, there was a strong, positive relation between a candidate's objective probability of being a Democrat and the average probability of voting to support. The correlation was 0.9000. The figure is suggestive of a threshold function, with the three candidates least likely to be Democrats attracting a low level of votes, and the five candidates most likely to be Democrats attracting a level of votes above 50%. For Republican respondents, there was a negative relation, although not quite as strong as for Democrats,  $r = -0.6606$ . Hence, the results suggest that both Democrats and Republicans tended to vote the party line (Democrats more so), even when explicit party information was not given. It is interesting to compare these correlations to the overall correlation for Experiment 1, in which respondents' probability judgments for party identification had a 0.9557 correlation with objective probability. Clearly, voting judgments are not the same as party identification judgments. Any lack of voting the party line in Experiment 2 is not due to respondents' inability to identify candidates' political parties.

We next examined these correlations at the level of individual respondents. For Democrats, the median correlation was 0.7823, and the mean correlation was 0.5004. For Republicans, the median correlation was -0.4057, and the mean correlation was -0.2971. As in Experiment 1, the median and mean correlations at the individual level are lower than the aggregate correlations, but they still suggest more party-line voting by Democrats. For a finer-grained analysis, we looked at mean correlations as a function of knowledge and partisanship of the respondents, with high or low knowledge operationalized as in Experiment 1. The mean correlations for high knowledge Democrats, low knowledge Democrats, high knowledge Republicans, and low knowledge Republicans were 0.6009, 0.4013, -0.3351, and -0.2396, respectively. For the purpose of an ANOVA examining tendency to vote the party line, correlations for Republican participants were multiplied by -1, for this analysis only. A two-way ANOVA revealed a main effect of knowledge, with high knowledge respondents showing stronger correlations,  $F(1, 569) = 11.09$ ,  $p < .001$ , and a main effect of party membership, with Democrats showing stronger correlations,  $F(1, 569) = 23.24$ ,  $p < .001$ . The interaction between knowledge and partisanship did not reach the level of statistical significance.

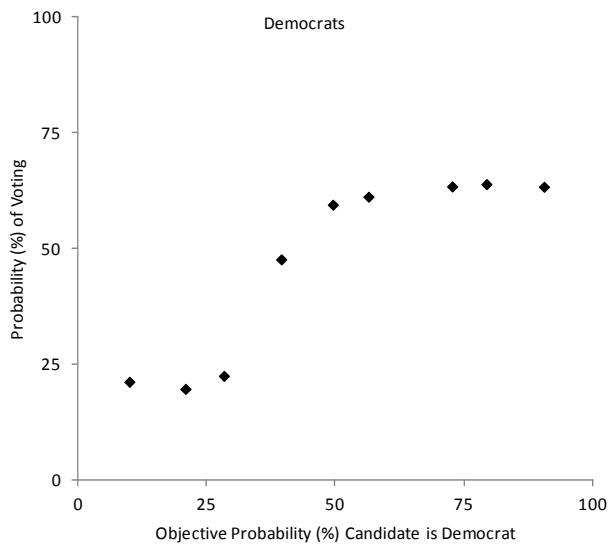


Figure 4. Objective probability candidate is a Democrat versus voting probability, for Democrats.

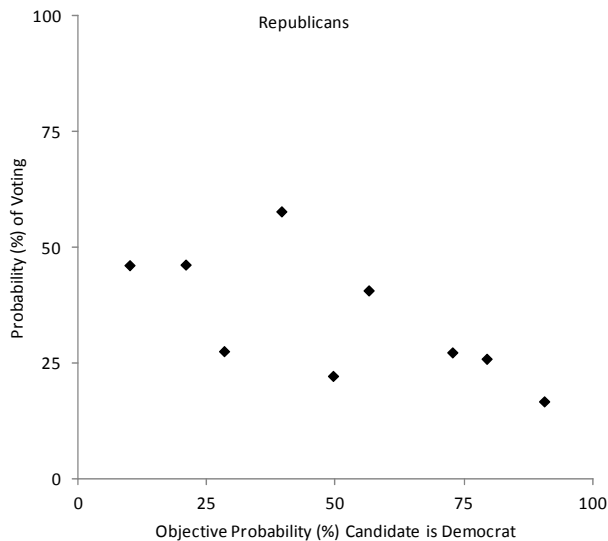


Figure 5. Objective probability candidate is a Democrat versus voting probability, for Republicans.

As in Experiment 1, we next conducted analyses of the cues used by respondents in each sub-group. Because of multicollinearity, the results should be taken as suggestive rather than definitive. Figure 6 shows the standardized regression coefficients across the five cues. Unlike Experiment 1, the regression coefficients varied considerably across sub-groups. It appears that Democrats were more influenced by demographic cues such as gender and number of children than are Republicans, although high knowledge Democrats were less influenced by demographic cues than low knowledge Democrats. It appears that Democrats were more influenced by stand on abortion and Republicans were more influenced by stand on government

spending. Use of these issue cues appears to be greater for high knowledge participants than for low knowledge participants. Again, the African-American cue shows negative weights for Democratic respondents. In fact, Democrats were much more likely to vote for African-Americans than for whites. For example, in a simple regression for all Democrat respondents, predicting judgments from just the African-American cue, the standardized regression coefficient was 15.91. In a simple regression for all Republican respondents, predicting judgments from just the African-American cue, the standardized regression coefficient was -0.86. Hence, Republicans were barely influenced by this demographic cue. Therefore, we would emphasize that cue utilization for voting appeared to differ considerably as a function of partisanship and political knowledge, and indeed the cues for voting are not the same as the cues for party categorization.

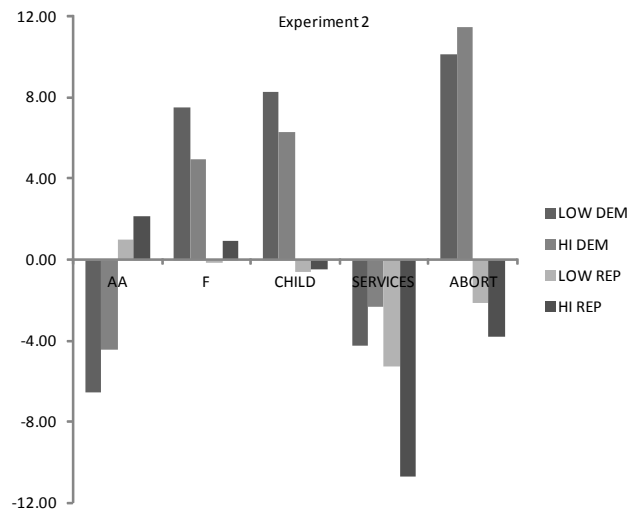


Figure 6. Estimated regression coefficients for low-knowledge and high-knowledge Democrats, and low-knowledge and high-knowledge Republicans.

In sum, like Experiment 1, Experiment 2 suggests an optimistic view of US voters, using multiple cues to vote for their party interests, even when party labels are omitted from descriptions. Interestingly, the pattern of responses for voting was different than for identification, so it did not seem that respondents treated the two tasks as the same. In Experiment 2, there were also robust differences in responses due to expertise and party membership.

## General Discussion

The results of the present experiments replicate Heit and Nicholson (2010) in terms of showing a highly polarized electorate. Just as our previous study found an almost perfect negative correlation between typicality in Democrat and typicality in Republican, here in Experiment 1 we found an almost perfect negative correlation between estimated probability that a candidate is a Democrat and estimated

probability that a candidate is a Republican. In Experiment 2, we found that Democrats and Republicans not only showed different patterns of voting for the same candidates, but also used different cues or feature weights. Democrats paid more attention to candidate's personal information and stand on the abortion issue, whereas Republicans focused on government spending. Although we previously concluded that "The opposite of Republican is Democrat," here we found that Democrats and Republicans did not simply disagree with each other, but actually cared about different issues and characteristics of candidates.

The respondents were remarkably successful at the identification and voting tasks. In the aggregate, the correlation between subjective judgments and objective probabilities was nearly .96, and the correlation for the median respondent very respectable, about .75. (We would refer to the "wisdom of crowds" phenomenon documented by Surowiecki, 2004, to explain the stronger performance at the aggregate level.) On the voting task, respondents were able to vote correctly—vote in their own party interests—even when labels were omitted.

In terms of connections to categorization research, we see commonalities with research on expertise in biological categorization (e.g., Medin et al., 1997, Medin & Atran, 2004). Democrats and Republicans can be seen as experts who see the same candidate but have different goals, just as taxonomists, landscapers, and maintenance workers would see the same tree with different goals. These differences are mediated by the level of expertise of each voter, suggesting there are different feature weights for identification and voting tasks, and for Democrats and Republicans. At this point, we can only pose the question of whether these feature weights are optimized for the tasks in the sense of Nosofsky (1986) and Kruschke (1992).

Indeed, what appears on the surface to be a feature weighting effect might have a different underlying explanation. For example, Heit (1998) showed that a Bayesian model of inductive reasoning can explain what appears to be a selective weighting effect in reasoning about either anatomy or behavior of animals (Heit & Rubinstein, 1994) not in terms of selective weighting but in terms of a hypothesis space that reflects feature co-occurrences. This account can be generalized to address a variety of selective effects in both induction and categorization (Kemp, Shafto, & Tenenbaum, 2012).

To conclude, we believe that studies of political cognition provide an interesting opportunity for the development and testing of computational models of categorization and reasoning.

## References

- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98, 409-429.
- Billman, D., & Heit, E. (1988). Observational learning without feedback: A simulation of an adaptive method. *Cognitive Science*, 12, 587-625.
- Converse, P. E. (1964). The nature of belief systems in mass publics. In D. E. Apter (Ed.), *Ideology and Discontent*. New York: Free Press.
- Delli Carpini, M. X., & Keeter, S. (1996). *What Americans Know About Politics and Why It Matters*. New Haven: Yale University Press.
- Hayes, B. K., Heit, E., & Swendsen, H. (2010). Inductive reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1, 278-292.
- Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational Models of Cognition*. Oxford: Oxford University Press.
- Heit, E., & Nicholson, S. (2010). The opposite of Republican: Polarization and political categorization. *Cognitive Science*, 34, 1503-1516.
- Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 411-422.
- Johnson, K. E., & Mervis, C. B. (1997). Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology: General*, 126, 248-277.
- Kemp, C., Shafto, P., & Tenenbaum, J. B. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, 64, 35-73.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.
- Lau, R. R., & Redlawsk, D. P. (2006). *How Voters Decide: Information Processing During Election Campaigns*. New York: Cambridge University Press.
- Markman, A.B., & Ross, B.H. (2003). Category use and category learning. *Psychological Bulletin*, 129, 592-615.
- Medin, D. L. & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review*, 111, 960-983.
- Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology*, 32, 49-96.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39-57.
- Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573-605.
- Surowiecki, J. (2004). *The Wisdom of Crowds*. New York: Doubleday.
- Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101, 547-567.
- Verbeemen, T., Vanoverberghe, V., Storms, G., & Ruts, W. (2001). Contrast categories in natural language concepts. *Journal of Memory and Language*, 44, 1-26.

# Going to Extremes: The influence of unsupervised categories on the mental caricaturization of faces and asymmetries in perceptual discrimination

Andrew T. Hendrickson; Paulo F. Carvalho; Robert L. Goldstone

([athendri](#), [pcarvalh](#), [rgoldsto](#)) @indiana.edu

Department of Psychological and Brain Sciences, Indiana University

1101 East Tenth Street, Bloomington, IN 47405 USA

## Abstract

Recent re-analysis of traditional Categorical Perception (CP) effects show that the advantage for between category judgments may be due to asymmetries of within-category judgments (Hanley & Roberson, 2011). This has led to the hypothesis that labels cause CP effects via these asymmetries due to category label uncertainty near the category boundary. In Experiment 1 we demonstrate that these “within-category” asymmetries exist before category training begins. Category learning does increase the within-category asymmetry on a category relevant dimension but equally on an irrelevant dimension. Experiment 2 replicates the asymmetry found in Experiment 1 without training and shows that it does not increase with additional exposure in the absence of category training. We conclude that the within-category asymmetry may be a result of unsupervised learning of stimulus clusters that emphasize extreme instances and that category training increases this caricaturization of stimulus representations.

**Keywords:** Categorical Perception, Category Labels, Perceptual Learning, Category Learning, and Language

## Introduction

**Categorical perception.** Our perceptual systems fail overwhelmingly to be precise replicators of reality in the way a camera or a microphone is, because these systems have not evolved to create a veridical representation of reality. Though constrained by overall neural architecture and the inertia of representations in primary sensory areas (Petrov et al., 2005), our perceptual systems consistently learn to create useful, but potentially distorted, representations of reality (Landy & Goldstone, 2005).

Often, this perceptual learning produces experiences that do not reflect the continuous variation of reality. Instead they warp that variability into discrete groupings such that entities that fall within a group are less discriminable than physically equally spaced entities that fall in different groups, a process known as categorical perception (CP; Harnad, 1987).

While some of the focus in CP research has been on assessing if particular categories are innate through cross-cultural studies (Kay & Reiger, 2003; Roberson & Davidoff, 2000; Sauter et al., 2011), early studies of CP focused on phonemes (Liberman et al., 1957) which show systematically different category boundaries based on an individual’s native language (Logan et al., 1991).

Learned CP has been shown in the visual modality across a variety of dimensions including hue and saturation (Goldstone, 1994), line drawings (Livingston et al., 1998), and morphs between arbitrarily paired faces (Kikutani et al., 2008; 2010).

**Category labels and CP.** An alternative framework suggests that the presence of category labels, and not perceptual changes, are responsible for CP effects (Pisoni & Tash, 1974). In this view the category label can be seen as an additional feature: entities in different categories have different labels thus having an additional feature unique for each category. This causes similarity to decrease and discrimination accuracy to rise. Items in the same category have the same label and thus either their similarity increases or remains constant leading to discrimination accuracy that does not increase.

Hanley and Roberson (2011) point out that the accuracy in assigning category labels is not constant across distance to the category boundary. Items farther away from the boundary are more likely to be categorized correctly than items closer to the category boundary. This viewpoint is consistent with many models of category learning that do not incorporate perceptual learning, including decision boundaries (Ashby & Maddox, 1990) and many exemplar-based (Nosofsky, 1986) models of category learning.

**Within-category discrimination asymmetries.** In perceptual discrimination testing in which a target object (X) must be held in memory and compared to itself and a foil object (A and B, respectively), if A is more likely to be assigned the same category label as X than B, then the probability of selecting A as the answer should increase relative to if A and B are equally likely to be assigned to categories. Therefore, when the target object is farther away from the category boundary than the foil and thus more consistently labeled in the category, accuracy will increase because the target object is more likely to be selected. Similarly, when the foil object is farther away, accuracy will decrease because the foil object will be selected more frequently (compared in both cases to cases in which no labeling asymmetry exists).

Hanley and Roberson (2011; see also Roberson et al., 2007) find this asymmetric within-category advantage for more perceptually extreme targets across a wide array of stimuli for which CP effects have been shown, including color across cultures (Roberson & Davidoff, 2000; Roberson et al., 2000; Roberson et al., 2005), facial emotions (Roberson et al., 2007), morphed celebrity faces and morphed unfamiliar but trained faces (Kikutani et al., 2008; 2010). They failed to find an advantage for more extreme faces among morphed unfamiliar and either untrained (Kikutani et al., 2008) or covertly exposed (Kikutani et al.,



2010) conditions. Recently, Sauter et al. (2011) failed to replicate the within-category asymmetry across cultures for morphed facial emotions despite showing CP effects.

**CP within categories.** Recent evidence has demonstrated (Gureckis & Goldstone, 2008; Hendrickson et al., 2010) that CP effects emerge not only between categories but also within categories. For example, two objects that belong to the same learned category (receiving the same label) may nonetheless have increased discriminability if they belong to different clusters within the category when compared to the case in which they belong to the same cluster. Within-category CP effects occur when the distribution of members of a category is structured into clusters (sub-groups within each category) rather than distributed uniformly (e.g. Goldstone, 1994) or normally (e.g. Ashby & Maddox, 1990).

These within-category CP effects are consistent with models of categorization in which the discriminability of items is not only affected by their category label but also by the learned clustering of items regardless of their labels (Love et al., 2004; McDonnell & Gureckis, 2011). These learning processes account for both within and between category CP effects through representational change: learning new clusters or prototypes that warp the similarity between entities either within or between categories (Goldstone & Hendrickson, 2009).

**Within-cluster discrimination asymmetry.** Interestingly, Gureckis and Goldstone (2008; see also Hendrickson et al., 2010) also found that the magnitude of the CP effects on both the category relevant and the category irrelevant dimensions increased as categorization accuracy improved. Importantly, neither of these CP effects were found before training.

Using this kind of stimuli space, the label ambiguity account of CP hypothesizes that the within-category asymmetry should emerge with the CP effect along the category-relevant dimension and is in fact causing the categorical perception effect. This would be for discriminations perpendicular to the category boundary in a two-dimensional space (see Fig. 1).

What remains unclear is if, within each category, a similar asymmetry should emerge parallel to the category boundary. A strict category label account suggests this should not occur because all stimuli would be equidistant to the category boundary and thus categorized equally accurately. This strict viewpoint would need to postulate a second mechanism to account for within-category CP effects.

Conversely, a category label ambiguity theory of CP that allows each cluster within a category to have a unique label would predict that the asymmetry will occur along the category-irrelevant dimension and that the emergence of the asymmetry will cause the within-category CP effect. The main purpose of this work was to investigate the emergence of within-category asymmetries along both the category relevant and irrelevant dimensions. A pre-post design was used to assess the relative timing during training of the emergence of CP effects and within-category asymmetries.

## Experiment 1

In the first experiment we tested these predictions by measuring perceptual discrimination accuracy along both relevant and irrelevant dimensions before and after category training. The stimuli and category structures were identical to previous studies (Gureckis & Goldstone, 2008; Hendrickson et al., 2010) that showed CP effects both dimensions. The perceptual discrimination task was a two alternative forced choice (2AFC) XAB task similar to those reported by Hanley and Roberson (2011). The within-cluster asymmetry and the standard CP effect were measured along the category relevant and category irrelevant dimensions before and after category training.

## Method

**Participants.** 80 Indiana University undergraduates participated in this experiment for course credit. 1 participant was excluded from analyses for failing to conclude the experiment within the allotted time (60 min).

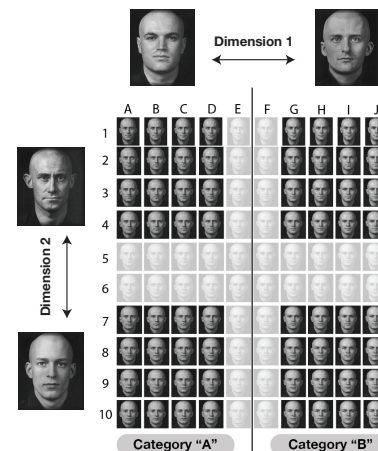


Fig. 1: Stimuli varied along two arbitrary dimensions (1 and 2) forming a 10-by-10 grid of blended faces. The light grey stimuli were not used in discrimination or categorization, introducing a source of within-category structure (two clusters of faces within each category). The vertical line between columns E and F shows an example category boundary used during category learning (the other boundary was a horizontal line between rows 5 and 6).

**Materials.** A 2-dimensional 10 by 10 matrix of bald male faces was created using a standard morphing technique (Steyvers, 1999). Each dimension was created by morphing between two faces selected from Kayser (1997). The two selected faces were roughly equally spaced in the multi-dimensional space based on a pilot similarity judgment task. The 100 stimuli that constitute the full matrix were created by equally morphing between all unique pairs of 10 faces in each of the two dimensions of faces (see Fig. 1).

**Procedure.** The task consisted of a block of 192 discrimination trials (pre-categorization phase), followed by 8 blocks of 16 categorization trials (categorization phase),

and a second block of 192 discrimination trials (post-categorization phase).

Each discrimination trial followed the XAB pattern: a target stimulus (X) was presented for 500 ms in the center of the screen followed by a response screen containing a target and a foil (A and B) stimulus presented horizontally until a response was made. A 500 ms blank screen was presented between the two screens and between trials there was a pause of 1000 ms. Participants were instructed to determine whether A or B was identical to X. The “target” is the option identical to X, and the “foil” is the other choice.

Target and foil face stimuli were selected such that they were identical along one dimension and were separated by 2 face stimuli in the 10 X 10 stimulus space along the other dimension. This spacing was determined by pilot studies to avoid ceiling or floor performance. The two central rows and columns were not used as either targets or foils.

Participants completed 384 discrimination trials broken up into the two blocks). Each block of 192 trials consisted of 12 unique trials in each row (and each column): the first and fourth stimuli in the row were compared four times, the fourth and seventh were compared four times, and the seventh and the tenth stimuli were compared four times (see Fig.1). Within each pair each stimulus was the target twice and with the target occurring equally often on the left and right position. These comparisons were made for 8 rows (excluding the middle two) and 8 columns (or rows), both parallel and perpendicular to the category boundary.

Each categorization trial consisted of a face stimulus appearing in the center of the screen. The two category labels appeared below the stimulus indicating which key (“q” or “p”) should be pressed to indicate that category label. The assignment of labels to keys was randomized on each trial. After participants respond, feedback indicating the correct category label was presented for 1000 ms followed by a pause of 1000 ms between trials. Each non-grey stimulus from Fig. 1 was presented twice in random order during category training.

## Results

**Categorization Performance.** A repeated measures ANOVA with block as a factor revealed a significant effect on categorization accuracy  $F(7,546) = 21.75, p < .0001$ , categorization accuracy improved throughout training.

A linear regression between distance to the center of the category space and categorization accuracy was performed separately for each dimension (category relevant and irrelevant). There was a significant improvement in accuracy for stimuli more distant on the category relevant dimension,  $F(1,236) = 73.7, p < .0001$  but no significant change in categorization accuracy as a function of distance along the irrelevant dimension  $F(1,236) = 1.47, p = .23$ .

**Discrimination Performance.** All discrimination trials were coded in three ways. Half the trials varied along the category relevant dimension (perpendicular to the category boundary) and half along the irrelevant dimension. Discrimination trials

were also coded on the relative extremeness of the target and foil objects: an equal number of trials were coded as “foil more extreme”, “target more extreme” and “equal.” Finally, for traditional CP analyses, the “foil more extreme” and “target more extreme” trials were grouped as Within trials, “equally” extreme trials were coded as Between trials. Between and Within trials could be relative to the category relevant or irrelevant dimension.

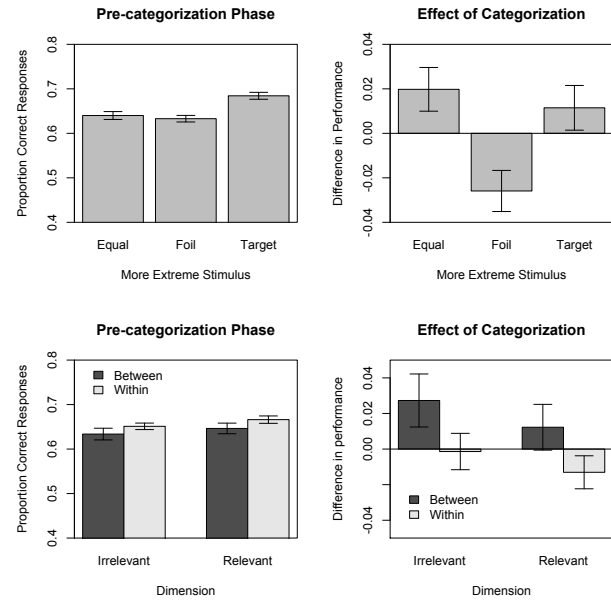


Fig. 2: Experiment 1 Results. Top-left: pre-categorization target-foil extremeness. Top-right: change in target-foil extremeness after categorization. Bottom-left: pre-categorization CP effects, split by dimension. Bottom-right: change in CP effects after categorization, also split. Error bars represent standard errors.

**Pre-categorization phase.** The graph in the top-left panel of Fig. 2 depicts the pre-categorization results, divided by extremeness condition. A 3 x 2 repeated measures ANOVA with relative extremeness (Equal vs. Foil vs. Target) and dimension (Relevant vs. Irrelevant) revealed a main effect of stimulus extremeness,  $F(2,156) = 16.15, p < .0001$ , but no main effect of dimension,  $F(1,78) = 2.06, p = .16$ , or interaction,  $F(2,156) < 1$ . Pairwise comparisons revealed that discrimination accuracy is higher when the target is the more extreme stimulus when compared to when the foil is more extreme,  $p = .0001$ , and when they are equally extreme,  $p = .001$ . The last two types of discrimination trials did not differ,  $p = 1$ . All  $p$  values were adjusted for multiple comparisons using a Bonferroni correction.

The results from the pre-categorization task considering the traditional CP analyses are depicted in the bottom-left panel of Fig. 2. A 2 x 2 repeated measures ANOVA with CP type (Within vs. Between) and dimension (Relevant vs. Irrelevant) as factors revealed a main effect of CP type,  $F(1,78) = 4.34, p = .04$ , with Within more accurate than Between, but no effect of dimension,  $F(1,78) = 1.68, p = .20$ , or interaction between the two variables,  $F(1,78) < 1$ .

**Change in discrimination performance after learning.** The pre-categorization analyses of extremeness and CP type were performed on the change in discrimination performance due to categorization. The change was computed by subtracting the pre-categorization discrimination accuracy from the post-categorization. The top-right panel in Fig. 2 depicts the results for the extremeness effects and the bottom right depicts the same results in terms of CP type.

A 3 x 2 repeated-measures ANOVA with relative extremeness and dimension as factors revealed a main effect of extremeness,  $F(2,156) = 6.89, p = .001$  but no main effect of dimension,  $F(1,78) = 1.33, p = .25$  or interaction,  $F(2,156) < 1$ . Pairwise comparisons revealed that performance changed equally for target more extreme and equally extreme,  $p = 1$ . The change in accuracy for the foil more extreme condition was significantly less than the other types: target more extreme ( $p = .03$ ) and equal ( $p = .005$ ).

To further investigate if accuracy performance improved with categorization, we performed a series of one-sample *t*-tests for each one of the extremeness conditions. The change in accuracy did not significantly differ from 0 for the target more extreme condition,  $t(78) = 1.11, p = .27$  but was significantly lower for foil more extreme,  $t(78) = -2.75, p = .007$ . The change in accuracy of the equal condition was marginally greater than 0,  $t(78) = 1.87, p = .06$ .

A 2 x 2 repeated-measures ANOVA revealed a main effect of CP type,  $F(1,78) = 5.75, p = .02$  but no main effect of dimension,  $F(1,78) = 1.38, p = .24$ , or interaction between the two variables,  $F(1,78) < 0$  (Fig. 2, bottom-right).

## Discussion

The results of Experiment 1 are not consistent with the hypothesis that category label ambiguity causes CP patterns. The pre-categorization phase in Experiment 1 indicates that the asymmetries seen in 2AFC tasks do not depend on the category or verbal codes assigned. More specifically, the results show that discrimination accuracy is higher when the target is more extreme than the foil alternative in the absence of any previous categorization learning. CP patterns were not observed before categorization despite the presence of the within-category asymmetry; in fact the reverse of the CP effect was marginally significant before categorization.

That the asymmetry exists before categorization suggests that it is a result of unsupervised learning processes rather than explicit category labels (Gureckis & Goldstone, 2008; Love et al., 2004). It remains unclear if the unsupervised mechanism is cluster labeling or perceptual change. We revisit this point in the general discussion.

Extremeness along the category relevant dimension predicted categorization accuracy but extremeness along the irrelevant dimension did not. This suggests that the asymmetry along the irrelevant dimension, both before and after categorization training, was not produced by differences in category labeling accuracy.

Categorization training did produce the expected CP effect: Between improved more than Within. This effect was modulated by an asymmetry among the Within trials, the foil

more extreme trials showed decreased performance and the target more extreme showed significantly higher change. This asymmetry is consistent with the category label ambiguity hypothesis and occurred after category training.

The changes in discrimination performance differ between the relevant and irrelevant dimensions. This may have been due to the extensive opportunity for unsupervised learning of cluster structure during pre-categorization discrimination.

## Experiment 2

One hypothesis that must be tested is if the asymmetric change in discrimination performance found after categorization training in Experiment 1 can be accounted for by the increased exposure to the stimuli instead of learning categories. This hypothesis is tested in Experiment 2, which is similar to Experiment 1 in that it consists of two critical blocks of discrimination judgments. However, another block of discrimination trials was substituted for the categorization task. Thus, by comparing performance in the first and last blocks of discrimination trials, which had roughly the same number of exposures to the stimulus as in Experiment 1, we can test the effect of experience with the stimulus space in the absence of categorization experience.

## Method

**Participants.** 76 Indiana University undergraduate students participated in this experiment for course credit. Two participants were excluded from analyses because they did not conclude the experiment in the allotted time (60 min).

**Procedure.** This experiment followed the same general procedure of Experiment 1 except for the exclusion of the categorization phase. Participants completed 3 blocks of discrimination trials. Each block was identical to those in Experiment 1.

## Results

All discrimination trials were coded similar to Experiment 1 but collapsed across dimension because no category boundary was learned. Discrimination trials were coded on the relative extremeness of the target and foil stimuli: an equal number of trials were coded as “foil more extreme”, “target more extreme” and “equal.” Finally, to compare to traditional CP analyses, the “foil more extreme” and “target more extreme” trials were grouped as Within trials, “equally” extreme trials were coded as Between trials.

**1<sup>st</sup> discrimination block.** The accuracy results for the first block of discrimination by extremeness condition are shown in the upper left corner of Fig. 3.

A repeated measures ANOVA revealed a main effect of stimulus extremeness,  $F(2, 146) = 9.03, p < .0001$ . Pairwise comparisons further revealed that trials in which the target was more extreme resulted in better discrimination than trials in which the foil was more extreme,  $p = .006$ , and also trials in which the two stimuli were equally extreme,  $p = .01$ . Finally, there is no difference in accuracy between trials in

which the foil was more extreme and trials in which both stimuli were equally extreme,  $p = 1$  (Fig. 3, top-left panel).

In the traditional CP analysis (despite no category training occurring), discrimination accuracy is higher for Within than Between discrimination trials,  $t(73) = -2.01$ ,  $p = .05$  (see Fig. 3, bottom-left panel).

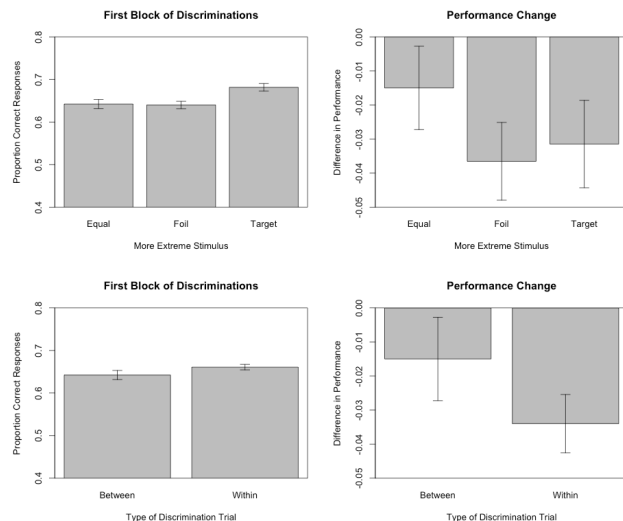


Fig. 3: Experiment 2 results. Top-left: Block 1 target-foil extremeness. Top-right: change in target-foil extremeness after prolonged exposure. Bottom-left: Block 1 CP effects. Bottom-right: change in CP effects after exposure. Error bars represent standard errors.

**Change in discrimination performance.** We computed the difference in accuracy between the last and first blocks to analyze the possible learning effect through successive exposure to discrimination trials. The top-right panel in Fig. 3 depicts the results considering the extremeness analysis while the bottom-right panel shows the results organized in terms of CP analyses.

A within-subjects ANOVA performed on these data revealed no main effect of stimulus extremeness,  $F(2,146) = 1.22$ ,  $p = .30$ . Similarly, when analyzing the change in performance between the last and first blocks of discrimination trials for Within and Between discriminations (see bottom right panel of Fig. 3) there are no significant differences in performance between the two types of discrimination trials,  $t(73) = -1.65$ ,  $p = .1$ .

**Categorization (Exp. 1) vs. Exposure (Exp. 2)** Categorization training ( $M = 0.011$ ) led to significantly higher change in discrimination performance relative to exposure ( $M = -0.031$ ) for target more extreme,  $t(151) = 2.62$ ,  $p = .01$ , as well as for equal trials, (cat.  $M = 0.020$ , exp.  $M = -0.015$ ,  $t(151) = 2.16$ ,  $p = .03$ ). On the contrary, there was not a significant difference between the change in discrimination accuracy for categorization ( $M = -0.026$ ) and exposure ( $M = -0.036$ ) for foil more extreme trials,  $t(151) = 0.72$ ,  $p = .47$ .

## Discussion

The results from the first block of discrimination trials replicate those found in the pre-categorization phase of Experiment 1. The asymmetry between the target and foil more extreme trials existed without category training and when CP patterns were not found.

Without category training, however, comparing the first and last blocks of discrimination in Experiment 2 did not show a change in performance consistent with the CP effect or a change in the difference between equal, target or foil more extreme trials. Performance for all trial types decreased in a consistent way across all trial types. This is likely due to fatigue considering the great number of trials participants complete without any feedback.

Finally, the categorization in Exp. 1 resulted in significantly different performance change for equal and target more extreme trials than what was seen with exposure alone (Exp. 2). However, this was not the case for foil more extreme trials. This suggests that category training improves discrimination for between-category judgments as well as for within-category judgments in which the target is more extreme than the foil.

## General Discussion

The presence of the within-category asymmetry before categorization and for each of the clusters refutes the hypothesis that the asymmetry alone can account for CP patterns or that the asymmetry is a direct result of explicit category labels. Instead these results are consistent with an unsupervised learning mechanism that is sensitive to the distribution of items within categories (Love et al., 2004; McDonnell & Gureckis, 2011) and a decision process for discriminations that distorts extreme exemplars to produce category caricatures in the distribution of items (Goldstone, 1996; Goldstone et al., 2003; Roberson et al., 2007).

The change in discrimination performance after categorization shows an increase in the asymmetry in Experiment 1 but not in Experiment 2. The fact that the asymmetry increases on the category relevant dimension as well as the irrelevant dimension is a challenge to the category label ambiguity hypothesis (Hanley & Roberson, 2011). To account for this behavior, the labeling hypothesis must be expanded to allow individual clusters within categories, learned via unsupervised mechanisms, to be assigned unique labels as in SUSTAIN (Love et al., 2004) or other semi-supervised learning models (McDonnell & Gureckis, 2011).

However, the fact that the effect of extremeness in a 2AFC task is observed before any category learning has taken place points to a biasing effect of extremeness within a stimulus set rather than category learning *per se*. Consequently, the relative change in performance seen after category learning might result from category learning processes that produce warped caricatures by shifting perceptual representations toward the stimulus extremes (Goldstone, 1996; Goldstone et al., 2003). This account is consistent with the relative improvement in between-category (and cluster) judgments as well as target more extreme judgments after categorization

relative to exposure alone for both the relevant and irrelevant dimensions.

We believe the strongest message from this work is the critical importance of measuring the change in perceptual discrimination performance to understand the learning mechanisms that underlie CP. Going forward, we plan to expand this analysis to look at changes in within-category asymmetries under conditions of verbal interference that may impair label usage (Hendrickson et al., 2010; Roberson et al., 2007) and formalize the unsupervised learning predictions in an extension of the SUSTAIN computational modeling framework (Love et al. 2004; Gureckis & Goldstone, 2008).

## Acknowledgments

This research was supported in part by National Science Foundation REESE grant 0910218 and Department of Education IES grant R305A1100060. PFC was supported by Graduate Fellowship SFRH/BD/68554/2010 from the Portuguese Foundation for Science and Technology (FCT). All data was collected by Jeremy Falkmann.

## References

- Ashby FG, Maddox WT. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance*.
- Goldstone, R. (1994). Influence of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*.
- Goldstone, R. L. (1996). Isolated and Interrelated Concepts. *Memory & Cognition*.
- Goldstone, R. L., & Hendrickson, A. T. (2009). Categorical Perception. *Interdisciplinary Reviews: Cognitive Science*.
- Goldstone, R. L., Steyvers, M., & Rogosky, B. J. (2003). Conceptual interrelatedness and caricatures. *Memory & Cognition*.
- Gureckis, T. M., & Goldstone, R. L. (2008). The effect of internal structure of categories on perception., *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.
- Hanley, J.R. & Roberson, D. (2011) Categorical perception effects reflect differences in typicality on within-category trials. *Psychonomic Bulletin & Review*.
- Harnad, S. (Ed.). (1987). *Categorical perception: The groundwork of cognition*. New York: Cambridge University Press.
- Hendrickson, A. T., Kachergis, G., Gureckis, T. M., & Goldstone, R. L. (2010). The effect of verbal interference and the internal structure of categories on perceptual discrimination., *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.
- Kikutani, M., Roberson, D., & Hanley, J. R. (2008). What's in the name? Categorical perception of unfamiliar faces can occur through labeling. *Psychonomic Bulletin & Review*.
- Kikutani, M., Roberson, D., & Hanley, J. R. (2010). Categorical perception for unfamiliar faces: Effect of covert and overt face learning. *Psychological Science*.
- Kay, P. and Regier, T. (2003) Resolving the question of color naming universals. *Proceedings of the National Academy of Science*.
- Kayser, A. (1997). *Heads*. New York: Abbeville Press.
- Landy, D., & Goldstone, R. L. (2005). How we learn about things we don't already understand. *Journal of Experimental and Theoretical Artificial Intelligence*.
- Lieberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*.
- Livingston, K., Andrews, J., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*.
- Logan, J., Lively, S., & Pisoni, D. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*.
- McDonnell, J. and Gureckis, T.M. (2011). Adaptive Clustering Models of Categorization. in *Computational Models of Categorization*. edited by Pothos and Willis, Cambridge University Press, Oxford, UK.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*.
- Petrov, A., Doshier, B., & Lu, Z. L. (2005). The dynamics of perceptual learning: an incremental reweighting model. *Psychological Review*.
- Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*.
- Roberson, D., Damjanovic, L., & Pilling, M. (2007). Categorical perception of facial expressions: Evidence for a "category adjustment" model. *Memory & Cognition*.
- Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition*.
- Roberson, D., Davidoff, J., Davies, I. R. L., & Shapiro, L. (2005). Colour categories in Himba: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*.
- Roberson, D., Davies, I. R. L., & Davidoff, J. (2000). Colour categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*.
- Sauter, D. A., LeGuen, O., & Haun, D. B. M. (2011). Categorical Perception of Emotional Facial Expressions Does Not Require Lexical Categories. *Emotion*.
- Steyvers, M. (1999). Morphing techniques for generating and manipulating face images. *Behavior Research Methods, Instruments, & Computers*.
- Webster, M. A., & Kay, P. (2011) Color categories and color appearance. *Cognition*.

# Re-learning labeled categories reveals structured representations

Andrew T. Hendrickson; George Kachergis; Caitlin M. Fausey; Robert L. Goldstone  
(`{athendri, gkacherg, cfausey, rgoldsto} @indiana.edu`)

Department of Psychological and Brain Sciences, Indiana University  
1101 East Tenth Street, Bloomington, IN 47405 USA

## Abstract

How do people learn to group and re-group objects into labeled categories? In this paper, we examine mechanisms that guide how people re-represent categories. In two experiments, we examine what is easy and what is hard to relearn as people update their knowledge about labeled groups of objects. In Study 1, we test how people learn and re-learn to group objects that share no perceptual features. Data suggest that people easily learn to re-label objects when the category structure remains the same. In Study 2, we test whether more general types of labeling conventions -- words that do or do not correspond with object similarities -- influence learning and re-learning. Data suggest that people are able to learn either kind of convention and may have trouble switching between them when re-structuring their knowledge. Implications for category learning, second language acquisition and updating representations are discussed.

**Keywords:** categories, labels, learning and transfer, knowledge change

## Introduction

An eighth-grade science student will happily tell you that she has just learned the electrons and nucleus of an atom are very similar to the planets and sun in our solar system. She now has a “multi-body orbiting systems” category with both of these systems as members. This approach is a classic example of transferring knowledge: information she previously learned about the solar system can be applied to the atom. But the story doesn’t end there: a physicist will tell you the truth is that an atom doesn’t really work that way, electrons are quantum wave functions that do not orbit a central point. New information causes the analogy to break down and if the eighth-grader continues to study physics she will learn new categorizations because the solar system belongs with systems explained by Newtonian mechanics while the atom belongs with quantum systems.

The process of relearning categories and reshaping knowledge is important not only for shifting from novice to expert but also for learning new ways to use words. Many languages group objects, relations and events in different ways (e.g., Majid, Boster, & Bowerman, 2008; Malt, Sloman, & Gennari, 2003; Wolff, Jeon, & Yu, 2009) and second language learners must categorize in new ways in order to speak their second language conventionally.

In laboratory category learning tasks, relearning has been studied using highly structured, binary dimensions. Using 3-dimensional binary stimuli, Kruschke (1996) found a clear hierarchy in the speed of relearning. Preserving the same

classification rule but reversing the response options is relearned the fastest, and switching to a new classification rule that involves a previously relevant dimension is more quickly learned than a rule using a previously irrelevant dimension. This pattern of results is consistent with the reversal learning literature (e.g., Kendler & D’Amato, 1954) and successfully modeled with straightforward extensions of many attention-shifting exemplar-based categorization models (e.g., Kruschke, 1996).

The main goal of this work is to examine category re-learning for categories that are not clearly defined by rules. Many categories in the world and in language do not obey a simple rule-based classification scheme, and it is interesting to consider the challenges that learners may face as they structure and re-structure their knowledge about these categories. In this work, we focus on how people learn to group and re-group objects into labeled categories.

Examining how people re-learn to group objects into categories has the potential to reveal insights into three important issues. First, does existing evidence about category learning and re-learning (e.g., with rule-based categories) extend to other kinds of categories? This is important to understand for more general theories of knowledge development. Second, how do people represent labeled groups of objects? What role does a label play in structuring knowledge (e.g., label-as-feature vs. label-as-category-marker, Deng & Sloutsky, 2012; Gelman & Markman, 1986)? This hotly-debated issue may benefit from data about patterns of re-learning because transfer paradigms are a useful way to assess what has been represented, on the logic that people are better able to re-learn structures that closely match what they had originally represented (e.g., Kruschke, 1996). Finally, second language learners are a large part of the world’s population and when people learn a new language they often must learn to categorize objects in different ways. What are the mechanisms that guide this re-representation?

As a first step toward these aims, here we consider a variety of potential changes between initial learning and relearning. In the experiments described below, people first learn a category structure and then are asked to relearn across a variety of potential changes. These changes reflect potential real-world relearning situations and the question becomes: what is easy and what is hard to re-learn as people re-structure their knowledge about labeled groups of objects?



## Study 1: Re-learning with unrelated objects

In Study 1, we examined several learning and re-learning relationships. In all scenarios, participants first learned to label nine objects. These nine objects were grouped into three categories; three labels were paired with three objects. After learning the initial categorization, participants re-learn. We manipulated what information remained constant and what changed between learning phases. Participants faced re-learning scenarios in which either the objects, the labels, the grouping of objects, and/or the mapping between groups and labels changed (see Table 1). Scenarios in which people quickly adapt to new object-label mappings are likely to be scenarios that conserve whatever representations they learned from labeling the original nine objects. Scenarios that people find more difficult likely overlap less with the learned representations. Thus, how easy it is to adapt to new object-label mappings has the potential to reveal some aspects of how people represent labeled objects.

Several patterns of re-learning data would be informative. In particular, the data may distinguish whether learners represent the information in the following ways: (a) Learners associate each distinct object with its appropriate label, or (b) Learners associate objects with labels, and also represent that the three similarly labeled objects are related to each other. That is, despite the fact that objects never appear together participants may learn the object groupings by virtue of sharing a label. Relearning patterns can also disambiguate between the potential interference or benefit of relearning new pairings or groupings of old objects compared to novel objects. We examine these questions in the following experiment.

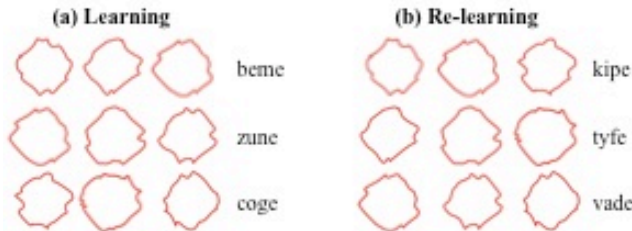


Figure 1: (a) An initial learning structure with 9 objects randomly assigned to 3 labels (by row). (b) A re-learning structure in the Re-Group & Re-name condition. Old objects grouped in a different way with novel labels.

## Method

**Participants.** 180 Indiana University undergraduates participated for course credit. 38 participants failed to complete the study in the allotted time and were excluded from analyses. Participants were randomly assigned to one of seven conditions ( $n = 19$  to  $23$  for each condition).

**Materials.** Figure 1 shows the objects and labels that one participant might see during (a) learning and (b) re-learning. For object stimuli, 72 unique segments were formed by fitting a spline through 8 randomly perturbed points along a 90-degree arc. Objects were created by combining 4 segments; a unique set of 18 objects was created for each

participant by sampling without replacement from the set of segments and arranging them to form a continuous outline. Labels were novel words from the set of {*beme*, *vade*, *kipe*, *coge*, *zune*, *tyfe*}. Other than the final “e” for all words, no letter was repeated across labels.

**Design.** All participants completed an original learning phase and then a re-learning phase. For each participant, nine of 18 objects and three of six labels were randomly selected for the initial learning phase. The remaining objects and labels were used in re-learning if needed. Conditions were defined by the changes between learning and re-learning (see Table 1).

Table 1: Re-learning Conditions in Study 1

Re-learning Condition	Items		Grouping Structure Conserved		Object-Label Associations Conserved	
	Objects	Labels	Yes	No	Yes	No
Learned	Old	Old	✓		✓	
Re-map	Old	Old	✓			✓
Re-name	Old	New	✓			✓
Re-group	Old	Old		✓		✓
Re-group & Re-name	Old	New		✓		✓
Recycled Name	New	Old		✓		✓
Novel	New	New		✓		✓

In some cases, new mappings between objects and labels were prompted only by substituting either new objects or new labels for the old objects and labels already learned. For example, a participant in the **Recycled name** condition might have used “*beme*” to label three objects during learning and then learned to re-use “*beme*” to label a novel three objects during re-learning. Similarly, a participant in the **Re-name** condition might have first learned that three objects were all called “*beme*” and then re-learned that they are all called “*zune*,” a name that had not been presented before. In these conditions, participants saw one set of objects [or labels] during learning and a different set of objects [or labels] during re-learning.

In other cases, new mappings between objects and labels reorganized the structure of previously learned categories using old objects. For example, consider a participant who first learned that three objects are each called “*beme*,” another three are each called “*zune*,” and a final three are each called “*coge*.” In **Re-map**, they might later re-learn that the first three are now called “*zune*,” the second three “*coge*,” and the third three “*beme*.” The grouping of objects is intact, the mapping between objects and labels is not. In **Re-group**, they might later re-learn that “*beme*” is now used for one object previously called “*beme*,” one object previously called “*coge*,” and one object previously called “*zune*.” This breaks the grouping of objects. In **Re-group & Re-name**, they might later learn to use “*kipe*,” a label not previously presented, to label three objects that had previously been called “*beme*,” “*coge*,” and “*zune*.” This breaks the grouping of objects and uses new labels.

Two other conditions gave participants no conflict with what was learned before – in one they continued with the



same objects, labels, and mapping between them as they had originally learned (**Learned**) and in another they learned about totally new objects and labels (**Novel**).

**Procedure.** On every trial, participants saw one object and three possible labels (Figure 2). The starting position of the cursor was equidistant from all response options and the location of labels was randomly determined on every trial. Participants were asked “Which category does this belong in?” The object and labels remained on the screen until the participant made a response. Afterward, feedback appeared above the object for 1200 ms: “Correct [Incorrect]! This is a \_\_\_\_.” Feedback included the correct label for all responses. There was a 400 ms pause between trials.

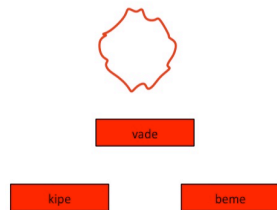


Figure 2: A sample trial.

Participants completed the learning phase in blocks of 9 trials. In each block, every object was presented once. Participants continued in this learning phase until they were correct on at least 8 out of 9 trials in a block for 4 consecutive blocks. Thus, everyone learned the original categories to criterion.

Participants then started the re-learning phase, reading these instructions: “You are doing great. In the next section the categories may change.” The re-learning phase consisted of 5 blocks in which all of the 9 stimuli were presented 3 times.

## Results

**Initial Learning.** The minimum number of blocks to reach criterion during the initial learning was 7 and the maximum number was 94. Participants reached criterion with a mean of 34.1 and a median of 31.0 blocks.

**Re-learning.** People’s performance in the relearning phase depended on block and condition. Re-learning data were analyzed using an ANCOVA, with categorization accuracy as the dependent variable, condition (7 levels) as a factor, and block (5 blocks) as a covariate. There were main effects of Condition ( $F(6, 135) = 23.1, p < 0.0001$ ) and Block ( $F(1,561) = 491.5, p < 0.0001$ ), and a significant interaction between Condition and Block ( $F(6,561) = 12.9, p < 0.0001$ ).

In order to interpret the interaction between Condition and Block, the trajectory of accuracy across block for each subject was clustered into groups and the distribution of each category within clusters was compared. The trajectories were clustered with methods to estimate the appropriate number of clusters (Kaufman & Rousseeuw, 1990). This technique identified two clusters, where one cluster was characteristic of three conditions: Learned (all 19 subjects), Re-name (19 of 20), and Re-map (16 of 21). These were conditions in which the original grouping of

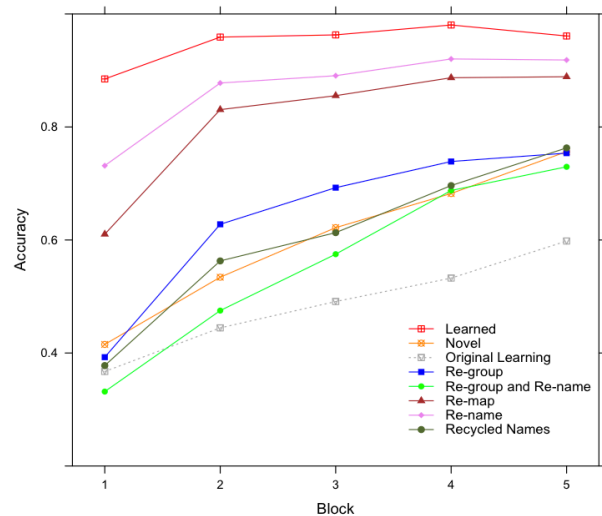


Figure 3: Participants’ mean accuracy by condition across re-learning in Study 1. Re-learning was faster than Original Learning. Of the re-learning conditions, Learned, Re-name, and Re-map show higher performance than the others.

objects were conserved from learning to re-learning (i.e., the first three rows of Table 1). The other cluster was characteristic of three conditions: Novel (14 of 19), Recycled Name (14 of 20), and Re-group & Re-name (17 of 23). The Re-group condition was equally split between both clusters (10 of 20 in each).

## Discussion

Performance during relearning depended on whether or not the groupings of objects were maintained from learning to re-learning. The three conditions in the higher accuracy cluster – Learned, Re-map, and Re-name – all conserved the grouping of objects from learning to relearning. The conditions strongly consistent with the lower accuracy cluster – Re-Group & Rename, Recycled Name, and Novel – did not preserve any groupings, either because they consisted of novel objects or because the three objects mapped to one label in relearning had not been mapped to a common label during learning.

The participants in the Re-group condition were equally split between the two clusters. In this condition, the nine objects that people saw during learning and re-learning were the same. Three of these were associated with the same name (one per label) during learning and re-learning, while the remaining six became associated with a different name during relearning (due to re-grouping the objects). The split between clusters suggests participants in this condition may have used different strategies, with some individuals less disrupted by the re-grouping because they were able to use object-label mappings that were identical between learning and re-learning but others learned representations that strongly relied on groupings, which were not conserved.

Study 1 suggests that labels — novel or re-purposed — are relatively easy to re-map to a group of objects. Learning new object groupings or re-groupings is more difficult.

These data suggest that learners may have successfully represented more than object-to-label mappings for each individual object and instead capitalized on structure among objects. We discuss this idea further after considering another interesting case of learning and re-learning.

## Study 2: Re-learning with related objects

The objects in Study 1 could be grouped into categories in only by using the labels because the objects shared no segments. How do people learn and re-learn *related* objects? In what way(s) does similarity among objects influence object grouping and impact re-learning?

To examine these issues, we again used a category learning and re-learning paradigm in Study 2. There are three main differences in Study 2: (a) objects were created such that they had structured similarity because some objects contained the same segments in the same locations; (b) the nature of mapping between objects and labels was manipulated to be consistent or inconsistent with object similarity. The mapping from training could either persist or switch from learning to re-learning; (c) all re-learning scenarios used novel labels and old objects. The non-control conditions in Study 2 all involved breaking the grouping of objects with the goal of understanding how relearning is influenced by the mapping between object similarity and category labels.

## Method

**Participants.** 107 Indiana University undergraduates participated in this study for course credit. 30 participants (evenly distributed across conditions) did not complete the study in the allotted time and were excluded from analyses. Participants were randomly assigned to one of six conditions ( $n = 11$  to 14 per condition).

**Materials.** Objects were created by combining four segments from a set of 72 segments. Nine objects were created for every participant. Instead of randomly selecting unique segments without replacement for each location of every object, some segments repeated across objects.

Specifically, two segments of every object also appeared in exactly two other objects (see columns 2 and 4 of the stimuli dimensions in Table 2). No two objects shared more than one segment and the location of the repeated segments was constrained so that they were not adjacent to each other. The other two locations were unique segments sampled without replacement (columns 1 and 3).

Further, in order to avoid possible preferences for a particular spatial location (e.g., whatever appears in the “top left quadrant” is easier to learn because of looking tendencies), a random number between 1 and 360 was selected for every participant and all objects for that person were rotated that many degrees.

In all conditions, different labels were used during learning and re-learning. For learning, three labels were randomly sampled without replacement from the same set of labels used in Study 1. The remaining three labels were used during re-learning.

Table 2: Stimuli and condition structure in Study 2

Stimuli	Learning		Re-Learning	
	Many	One	Many	One
1 1 1 1	A	A	D	D
2 1 2 2	C	A	E	E
3 1 3 3	B	A	F	F
4 2 4 1	B	B	E	D
5 2 5 2	A	B	F	E
6 2 6 3	C	B	D	F
7 3 7 1	C	C	F	D
8 3 8 2	B	C	E	E
9 3 9 3	A	C	D	F

*Objects:* Four segments per object, columns indicates a location. Within a column, the same number indicates an identical segment. Across columns, numbers are unrelated. *Learning & Relearning:* Letters indicate labels. In “One” columns, a label matches one segment (Learning-One, Segment 2; Relearning-One, Segment 4). In “Many” columns, no single feature matches a label.

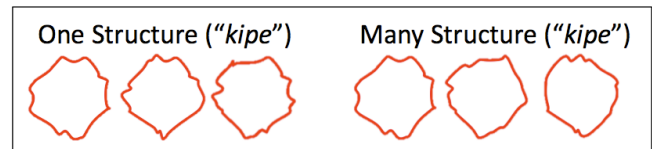


Figure 4: Example categories (Study 2). “One”: shared segment in upper right. “Many”: All unique segments.

**Design.** Categories were defined by the nature of segment-to-label mapping. The One category structures had a one-to-one mapping between segments in a location and labels (see Table 2). The Many structures did not have a one-to-one mapping, and never shared a segment within a category.

**Conditions were defined by the type of structures in learning and re-learning.** Learning and re-learning structures were combined so that some participants continued in the same style of mapping (One-One; Many-Many conditions) while others switched (One-Many; Many-One conditions). All participants in these conditions learned new ways to group objects during re-learning regardless of the type of mapping in learning and re-learning, the group of objects that shared the same label changed.

Two additional control conditions did not change the category structure between learning and re-learning though novel labels were introduced. In “Same One” participants started and remained in a “One” structure and in “Same Many” participants stayed in a “Many” structure.

**Procedure.** Learning and re-learning was exactly like Study 1. After re-learning, people did a re-test in which they tried to recall the label that they had *originally* learned for every object. They were instructed “In this final section we will ask you about the first labels you learned. You will not be told if you were correct or not, please do the best you can.” The re-test trials only displayed the original labels and no feedback was provided. People saw each object 8 times.

## Results

**Initial Learning.** People reached criterion in about the same number of blocks during each kind of initial training (One:  $M = 35.2$ , Many:  $M = 41.7$ ,  $t(75) = 1.5$ ,  $p = 0.14$ ).

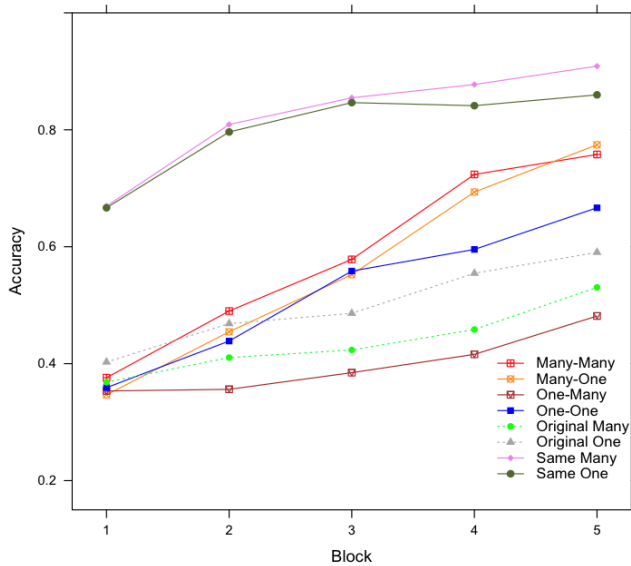


Figure 5: Participants' mean accuracy in Re-learning (Study 2). The Same conditions are easy, followed by Many-Many, Many-One, and to some extent One-One. One-Many is more difficult, like the Original Learning.

**Relearning.** People's performance in the relearning phase depended on block and condition (Figure 5). Re-learning data were analyzed using an ANCOVA, with accuracy as the dependent variable, Condition (6 levels) as a factor and Block (5 blocks) as a covariate. There were main effects of Condition ( $F(5,71) = 12.7$ ,  $p < 0.0001$ ) and Block ( $F(1,302) = 275.4$ ,  $p < 0.0001$ ), and a significant interaction between Condition and Block ( $F(5, 302) = 9.5$ ,  $p < 0.0001$ ).

Like Study 1, a clustering analysis was performed using the accuracy data across block for every participant and was best fit with 3 clusters. The highest accuracy cluster was most consistent with the two conditions in which object groupings did not change from learning to re-learning: Same Many (10 of 12) and Same One (12 of 14). A second cluster was most consistent with the two conditions that did change groupings but started with the Many mapping: Many-One (7 of 11) and Many-Many (9 of 12). The third cluster was most consistent with the One-Many condition (10 of 13). The One-One condition was equally consistent with the second and third clusters (6 of 13 in each). Thus, switching from a One structure, where a segment and label are paired, to a Many structure, where many segments map to a label, appeared to be particularly challenging.

**Re-test.** After re-learning, people's memory for the originally learned labels did not appear to be strongly influenced by Condition or Block. Memory accuracy was analyzed using an ANCOVA with accuracy in the re-test as the dependent measure, Condition (6 levels) as a factor and Block (4 blocks) as a covariate. A marginal effect of Condition was found ( $F(5,70) = 2.2$ ,  $p = 0.062$ ). No

significant effect of Block was found ( $F(1,222) < 1$ ,  $p = 0.5$ ). There was no significant interaction between Condition and Block ( $F(5, 222) = 1.4$ ,  $p = 0.21$ ).

To better understand the marginal effect of Condition, pair-wise post-hoc tests showed that the accuracy in the Many-Many condition ( $M = 0.73$ ) was significantly lower than Same-Many ( $M = 0.91$ ,  $p = 0.006$ ). All other conditions were not significantly different ( $p > 0.01$ ).

## Discussion

When people learned novel labels for objects, they had the easiest time if the original groupings of objects were preserved. Using new words to talk about old groups was equally easy whether or not the original groupings were mapped to labels based on repeated, shared segments or based on individual items. If the grouping of objects was disrupted, however, people who had originally learned a Many mapping (i.e., item-specific association with labels) re-learned faster than those who had learned a One mapping (i.e., repeated-segment association with labels).

Why is it easier to learn new labels for new groups after having first learned an item-specific labeling convention than after having learned a labeling convention that capitalizes on perceptual similarity? Learning that words are used for similar objects (i.e., One-One and One-Many) may have directed attention to a common, predictive feature for each word. Re-learning to label these objects may be hard when this feature is no longer predictive. But shifting attention to different features might be easier if you have learned that words are used on an object-by-object basis – this is a useful approach during “re-learning” even when the particular mappings change. Thus, generalizations about “labeling conventions” may influence later re-learning.

Interestingly, people may learn labeling conventions that capitalize on perceptual similarity in different ways. In this study, participants in the One-One condition were equally split between the two clusters. The group of people who were clustered with the Many-Many and Many-One conditions may have been able to re-learn easily because they discovered the new shared feature during re-learning. In addition to the original associations between objects and labels, these people may have learned that there *is* a feature that predicts a label. They then successfully transferred this generalization during re-learning. The other group of people in this One-One condition may have learned the original association between a shared feature and its label but not the higher order pattern and could not later shift their attention.

Re-learning is even more challenging if people must overcome not only a learned association between a specific feature and a label, but also the higher-order generalization that there *will be* one feature that predicts each label. This is the situation of learners in the One-Many condition, and indeed, they performed the worst during re-learning.

Is there more than one way to successfully re-learn after having learned a labeling convention that is item-specific? This is an open question. In the Many-Many and Many-One conditions, people initially learned item-specific mappings

between objects and labels. In the re-learning, even though people in the “Many-One” condition could have learned to associate labels with shared features, they also could have learned these associations in an item-specific way. The fact that people in the Many-Many and Many-One conditions showed similar re-learning trajectories suggests that this may have been a common strategy. Thus, it is possible people in all re-learning scenarios were likely to continue using whatever labeling convention they had originally used (either item-specific, or shared feature). It just happens to be that an item-specific strategy leads to success throughout learning and re-learning, while a labeling convention that relies on perceptual similarity (shared features) does not.

## General Discussion

Data from two experiments suggest that people easily learn to re-label objects when the category structure remains the same from original to subsequent learning. Further, people can learn multiple labeling conventions – words do or do not correspond with object similarities – but may have trouble switching between them when restructuring their knowledge. Like advantages for using relevant dimensions when re-learning rule-based categories (Kruschke, 1996), there are advantages to using existing category structures when re-learning arbitrary or similarity-based categories.

The advantage for shared structure, as well as shared labeling conventions, may help to explain some difficulties that adults face when learning a second language. It should be especially hard to learn new labels for objects that are organized differently in the two languages compared to objects that are categorized similarly. Moreover, an intriguing possibility is that once learners have made higher-order generalizations about the kinds of non-linguistic structure that predict labels, they may find it hard to “start from scratch” and build new associations as they construct different similarity spaces. Interesting test cases of this idea would be to see if L2 learners make systematic labeling errors based on L1 structure (and change over the course of L2 learning), and also whether object similarity spaces are predictably different in bilinguals than monolinguals.

Labels may play a critical role in forming categories and shaping representations (e.g., Goldstone, 1994; Goldstone & Hendrickson, 2009) and these representational changes may persist during subsequent use. The re-learning tasks used here raise interesting questions about how labels may wax and wane as drivers of representation and re-representation. We suggest that a focus on *change over time*, together with approaches that test knowledge structures at a single moment will enrich our understanding of these issues.

Learning any particular structure can be a double-edged sword. Subsequently learning a very similar structure may be easy, but it may be much harder to learn very different structures. In the case of language, different labeling conventions may promote relatively richer or shallower encoding of individual object representations. For example, if it is possible to successfully label objects by paying attention to a single shared feature, then using labels to learn

these categories may promote a representation that prioritizes this feature. But, if it is necessary to pay attention to multiple aspects of objects in order to talk about them, then people will (e.g., Slobin, 2003). Thus, the complexity of the mapping between labels and structures might shape subsequent representations via perceptual learning (Goldstone & Hendrickson, 2009). It may be easy to use them in some situations, but harder in others.

In general, transferring knowledge to new situations is easier after deeply encoding the relevant structure and certain kinds of initial exposure promote this kind of encoding (e.g., Day & Goldstone, 2011; Gentner & Markman, 1997). It is an interesting open question to consider what aspects of language learning – what labeling conventions, used at what points of learning about the relevant structures and categories in one’s world – support different trajectories of developing object representations.

## Acknowledgements

This research was supported in part by National Science Foundation REESE grant 0910218, Department of Education IES grant R305A1100060, and National Institute of Health and National Research Service Award HD007475-17.

## References

- Day, S., & Goldstone, R. L. (2011). Analogical transfer from a simulated physical system. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Deng, W., & Sloutsky, V.M., (2012). Carrot eaters or moving heads: Inductive inference is better supported by salient features than by category labels. *Psychological Science*.
- Gelman, S.A., & Markman, E. (1986). Categories and induction in young children. *Cognition*.
- Gentner, D., & Markman, A. B. (1997). Structural alignment in analogy and similarity. *American Psychologist*.
- Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*.
- Goldstone, R. L., & Hendrickson, A. T. (2009). Categorical Perception. *Interdisciplinary Reviews: Cognitive Science*.
- Kaufman, L., & Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
- Kendler, & D’Amato, (1954). A comparison of reversal shifts and non-reversal shifts in human concept formation behavior. *Journal of Experimental Psychology*.
- Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*.
- Majid, A., Boster, J.S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition*.
- Malt, B.C., Sloman, S.A., & Gennari, S.P. (2003). Universality and language specificity in object naming. *Journal of Memory and Language*.
- Slobin, D.I. (2003). Language and thought online: Cognitive consequences of linguistic relativity. In D. Gentner & S. Goldin-Meadon (Eds.), *Language in mind: Advances in the investigation of language and thought*. MIT Press.
- Wolff, P., Jeon, G., & Yu, L. (2009). Causal agents in English, Korean and Chinese: The role of internal and external causation. *Language and Cognition*.

# A Matter of Process Accuracy: Observing or Inferring the Criterion of Few or Many Exemplars

Maria P. Henriksson (maria.henriksson@psyk.uu.se)

Department of Psychology, Uppsala University  
SE-751 42, Uppsala, Sweden

## Abstract

Can we tailor fit the training to enhance judgment accuracy by changing to the learning format that invites the most effective cognitive process for the task environment at hand? The results from a study on multiple-cue judgments revealed that observing the cues and the criterion of exemplars simultaneously with no feedback involved in the training, a learning format predicted to invite exemplar memory processes, was the better learning option when there were few unique exemplars in training. Inferring the criteria of different exemplars and receiving outcome feedback during training, a learning format predicted to invite cue-abstraction, was the better learning option when there were many unique exemplars in training. Implications for the notion of an initial “rule bias” suggested by several previous studies are discussed.

**Keywords:** rule-bias; observation; feedback; cue-abstraction; exemplar memory

## Introduction

Virtually all research on multiple-cue judgment has involved the learning format *feedback learning* and multiple-cue learning has often, more or less explicitly, been regarded as an analytic or rule-based process, where outcome feedback is used to adjust cue weights and to test hypotheses about the cue-criterion relations (Klayman, 1988). The *Cue-Abstraction Model* (CAM; e.g. Juslin, Karlsson & Olsson, 2008) is a cognitive model capturing many of these properties of the judgment process, assuming explicit knowledge about cue-weights and a controlled integration of information by an additive rule (see the Method section for more information about this model). If the analytic, abstract knowledge assumed with the CAM accurately reflects (or well approximates) the task environment, the judgments become independent of the concrete exemplars encountered during training. Judgment accuracy for old exemplars experienced in training and new exemplars should thus be similar, with ability to extrapolate the judgments beyond the observed range of training exemplars (DeLosh, Bussemeyer, McDaniel, 1997).

However, feedback learning is not the only learning format. In the related domain of category learning there is a growing interest in investigating the effects of *observation learning* where no feedback is involved and people learn from observing the cues and the criterion (see, e.g., Ashby, Maddox, & Bohil, 2002). There is also some evidence that exemplar memory processes can better describe the performance with observation learning than with the

standard feedback learning format (Estes, 1994). A recent study on multiple-cue judgments revealed evidence that observation learning invites exemplar processes and is able to exploit more complex task environments with resulting superior performance when the cues are multiplicatively related to the criterion in the task environment (Henriksson, Enkvist & Juslin, 2012). This suggests that exemplar processes might be used by observation learners who only have to store the information about exemplars in memory and use the similarity to these stored exemplars when assessing the criterion value of new exemplars in subsequent judgments. In contrast to the predictions for CAM, exemplar processes predicts that judgment accuracy for old exemplars experienced in training is superior to the judgment accuracy for new exemplars, and that judgments cannot extrapolate beyond the training range of exemplars (DeLosh et al., 1997; Medin & Schaffer, 1978; Nosofsky, 1986). See Method section for more information about this model.

In categorization, Rouder and Ratcliff (2004) have found evidence that exemplar processes provides a better account of the data when there are few and distinct exemplars and rule-based processes provides a better account of the data when the exemplars are confusable and not distinct from each other, for example when exemplars are probabilistically assigned to a category. It is conceivable that exemplar processes might be more vulnerable to the number of unique exemplars in the task environment. As the number of different exemplars increases with experience, the memorization might become difficult with interference between exemplars. Thus, the accuracy of exemplar processes might be constrained to task environments where there are a limited number of training exemplars. On the other hand, a cue-abstraction process might require experience of many different exemplars varying on the cue-dimensions for testing and fine-tuning hypotheses about the relative cue-weights and the relationship between cues and the criterion. The prediction for the multiple-cue judgment task is therefore that the better relative fit of EBM (i.e., clearer advantage for EBM over CAM), the better the judgment accuracy when there are few exemplars in training. When there are many training exemplars the prediction is that the better relative fit of CAM the better judgment accuracy.

The recurring “rule bias” (Ashby, Alfonso-Reese & Turken., 1998, p. 467), an initial inclination for analytical processes, is perhaps not surprising considering that the feedback format is often applied in studies on categorization and multiple-cue judgment. It is possible that feedback



learning per se invites relatively more cue-abstraction (CAM) or at least reinforces that kind of process. However, it is reasonable to appreciate that exemplar processes can act as a back-up process whenever rule-based processes fails to exploit the task environment (Juslin et al., 2008; Karlsson, Juslin & Olsson, 2008). It is possible that the previous reported shifts to exemplar processes (Ashby et al., 1998; Erickson & Kruschke, 1998; Kalish, Lewandowsky & Davies, 2005) may in part be mediated by a spontaneous shift to observation learning. For example, if the task is difficult to learn by testing explicit hypotheses against feedback, the participant could start to randomly guess the missing value and wait for the correct outcome feedback to appear. Then, the participant will have the same information as an observation learner who only has to store the information in memory for subsequent use.

In sum, the predictions are that with few exemplars, observation is predicted to produce higher accuracy than feedback by inviting the EBM. With many exemplars, feedback learning is predicted to produce higher judgment accuracy than observation by inviting the CAM (see Figure 1 for the predictions).

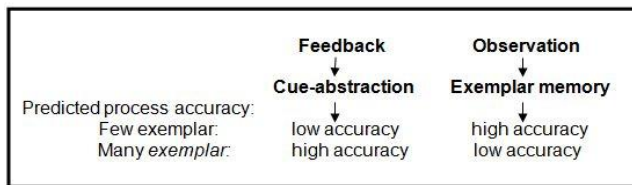


Figure 1. Predicted judgment accuracy and invited processes for feedback and observation learners after training with few or many exemplars.

## Method

### Participants

Sixty-four students from Uppsala University volunteered. Seven were excluded since their test performance indicated no learning. Of the remaining 57 participants, 40 were women and 17 men with the average age of 24.86 ( $SD=7.20$ ).

### Materials, Procedure, and Design

A computerized multiple-cue learning task was presented to the participants instructing them to learn the suitability for an unspecified job based on values ranging from 0-10 on four cues describing different applicants. The cues were *independent*, *thoughtful*, *detailed-oriented*, and *practical* and they were stated on the computer screen along with the cue-value describing each individual exemplar.

The criterion  $c$ , the degree of suitability of the exemplar is a linear, additive function of the cues  $C$ , with the most important cue with a relative weight of .4 and the second-most important cue with a relative weight of .3 and so forth

(see Equation 1)<sup>1</sup>. The assignment of the labels of the cues to the relative cue-weights was counterbalanced across the participants.

$$c = 500 + .4 \cdot C_1 + .3 \cdot C_2 + .2 \cdot C_3 + .1 \cdot C_4 \quad (1)$$

Among the  $11^4$  possible exemplars that can be generated, two sets of training exemplars were sampled, each with a criterion ranging from 510 to 590. The 16 training exemplars in the condition *few exemplars* were presented 10 times in a randomized order. In the condition *many exemplars* the 16 exemplars were presented only once along with 144 other exemplars in a randomized order (in total 160 training trials in each condition). At test, the 16 exemplars reoccurred together with 14 new exemplars, all with criterion values ranging from 500 to 600. The 30 test exemplars were presented twice in a randomized order. At test, 12 of the old exemplars experienced in training were matched to 12 new exemplars with the same criterion in order to examine new-old differences.

A 2 x 2 factorial design was used and participants were randomly assigned to one of four experimental conditions. The independent variables were the learning format (observation and feedback) and the numbers of exemplars in training (few or many). All participants were told that they were going to learn the degree of suitability of different presented applicants. Half of the participants were told that they should learn by observing the cues and the criterion of different exemplars, similar to screening lists of previous employees' characteristics and degrees' of suitability (observation learning). The other half was told that they should observe the cues describing each individual exemplar and predict the missing criterion value. After each judgment, outcome feedback about the criterion was provided to the participant (feedback learning). Half of the participants in each learning condition experienced few unique training exemplars and the other half experienced many unique exemplars. After the training phase a test phase followed that was identical for all participants. All participants were informed that no feedback should be received during or after this test phase.

### The Models and Dependent variables

The *Cue-Abstraction Model* (CAM; e.g. Juslin et al., 2008) assumes that the participants abstract cue-weights in training, analogue to linear regression weights. When they later judge the criterion of a probe, they use the knowledge of the cue-weights to integrate the linear additive impact of the cues. For each cue  $C_i$ , the weight  $w_i$  ( $i=1 \dots 4$ ) is used when adjusting the criterion  $\hat{c}$  of a probe  $p$ ,

<sup>1</sup> Two cues were positively related to the criterion and two cues were negatively related to the criterion so as not to make identification of cue-directions trivial (i.e., with high cue-values always predicting high suitability and low cue-values always predicting low suitability).

$$\hat{c}_p = a + \sum_{i=1}^4 \omega_i \cdot C_i \quad (2)$$

where the intercept  $a$  and the weights  $\omega$  are parameters in the model.

The *Exemplar-Based Model* (EBM) refers to a version of the generalized context model (Nosofsky, 1986) that is applicable to multiple-cue learning (e.g., Juslin et al., 2008). As many exemplar-based models assume (e.g., Medin & Schaffer, 1978; Nosofsky, 1986), people store memory traces of concrete exemplars together with the outcome. At the time of judgment, people retrieve similar exemplars from long-term memory. According to the Generalized Context Model (GCM: Nosofsky, 1986), the similarity to stored exemplars depends on the attention to the cue dimensions and the sensitivity for the distance between the exemplars in the psychological space. The distance between the probe  $p$  and an exemplar  $j$  is given by,

$$d_{pj} = h \left[ \sum_{m=1}^4 w_m |x_{pm} - x_{jm}|^r \right]^{1/r} \quad (3)$$

where  $x_{pm}$  and  $x_{jm}$  are values of the probe and the exemplar on cue dimension  $m$  ( $m=1..4$ ),  $w_m$  are attention weights on cue dimension  $m$ , and  $h$  is a parameter that captures the sensitivity for the distance between the exemplars in the psychological space. The sensitivity varies from 0 to  $\infty$ . The attention weights on cues vary between 0 and 1 and are constrained to sum up to 1. Euclidian metric is used and  $r$  is set to 2. The overall similarity between a probe  $p$  and exemplar  $j$  is assumed to be a nonlinear decreasing function of their distance  $d_{pj}$  in the psychological space,

$$S(p, x_j) = e^{-d_{pj}} \quad (4)$$

EBM implies that the criterion  $\hat{c}$  of a probe  $p$  is assessed by,

$$\hat{c}_p = \frac{\sum_{j=1}^4 S_j \cdot c_j}{\sum_{j=1}^4 S_j} \quad (5)$$

where  $S_j$  is the similarity to exemplar  $j$ , and  $c_j$  is the criterion of exemplar  $j$ . The estimated criterion of a probe is the weighted average of the criteria of similar exemplars retrieved from long-term memory, where the similarity is the weight (see Juslin et al., 2008).

This exemplar model and the cue-abstraction model (Juslin et al., 2008) were fitted individually to the responses by each participant in the test phase. A cross-validation procedure was used in the modeling and the model fit is measured by *Root Mean Squared Deviation* (*RMSD*) between the model prediction and the judgment<sup>2</sup>. Judgment accuracy in the test phase is measured by *Root Mean Squared Error* (*RMSE*) between the judgment and the

criterion. Hence, the lower value of *RMSE*, the better judgment accuracy. Deltafit ( $\Delta$ fit), a measure of the relative differences in fit of EBM and CAM was computed by subtracting the *RMSD* for CAM from the *RMSD* for EBM so that negative values corresponds to a relatively better fit for EBM and positive values corresponds to a relatively better fit for CAM. Separate analyses of the correlations between the deltafit and judgment accuracy (*RMSE*) can therefore be calculated in order to explore how useful the two cognitive processes are for achieving accuracy when experiencing few or many training exemplars.

## Results

A split-plot ANOVA revealed only a significant within-effect of *RMSD* for CAM and EBM,  $F(1, 53)=20.82$ ,  $p<.001$ . The model with the average best fit was the CAM ( $RMSD_{CAM}=11.10$ ,  $SD=3.82$  vs.  $RMSD_{EBM}=14.2$ ,  $SD=5.59$ ). The variance explained by the CAM was significantly higher for feedback learners regardless of the number of training exemplars ( $r^2_{CAM}=.73$  and  $.77$ ;  $r^2_{EBM}=.61$  and  $.61$ ). Though CAM was found to be the better fitting model for observation by the *RMSD*, the variance explained by the CAM was not significantly higher than for EBM ( $r^2_{CAM}=.86$  and  $.80$ ;  $r^2_{EBM}=.77$  and  $.71$ ).

In line with the predictions, a split-plot ANOVA with the judgment accuracy of the matched old and new exemplars at test revealed that there was a significant interaction between the learning formats and the matched exemplars,  $F(1,53)=4.76$ ,  $p=.034$ . Observation learners had significantly better accuracy judging the old exemplars that had been experienced in training compared to judging the matched new exemplars. This result suggests that exemplar based processes are used by observation learners. No systematic difference in judgment accuracy between matched old and new exemplars was found for feedback learners, suggesting that cue-abstraction is used by feedback learners (see Figure 2 for illustration of the results).

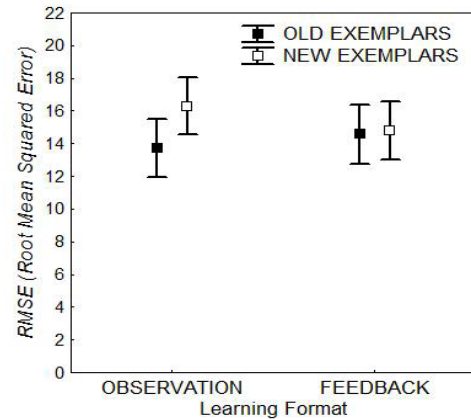


Figure 2. Judgment accuracy (*RMSE*) for the matched old and new exemplars at test for observation and feedback learners after training with few or many exemplars. Lower value of *RMSE* signifies better judgment accuracy. Vertical bars denote 95% Confidence intervals.

<sup>2</sup> The 2 x 30 judgments of the test exemplars were randomly split into two sets for each participant so that each exemplar occurred once in each set.



In line with the predictions, the deltafit (i.e., the relative fit of the models) had a positive correlation with *RMSE* when there were few training exemplars,  $r_s = .33$ ,  $t(26)=1.78$ ,  $n=28$ ,  $p=.04$  one-sided, a result that suggest that the better fit for EBM, the lower the *RMSE* (thus better judgment accuracy). With many experienced exemplars in training, the deltafit had a negative correlation with *RMSE*,  $r_s = -.34$ ,  $t(27)= -1.88$ ,  $n=29$ ,  $p=.04$  one-sided, a result that suggest that the better fit for CAM, the lower the *RMSE* (thus better judgment accuracy). The two correlation coefficients differed significantly ( $p < .01$ ).

A two-way ANOVA with judgment accuracy (*RMSE*) as dependent variable revealed no main effects of the number of experienced exemplars or learning formats. However, there was a significant interaction effect,  $F(1, 53) = 4.17$ ,  $p = .046$ . In line with the prediction, observation learners had marginally significantly better judgment accuracy than feedback learners after training with few exemplars ( $M = 13.41$  vs.  $15.92$ ,  $SD = 3.69$  vs.  $5.08$ ,  $p = .06$  by planned comparison). On the other hand, feedback learners had marginally significantly better judgment accuracy than observation learners after training with many unique exemplars ( $M = 11.58$  vs.  $13.69$ ,  $SD = 4.76$  vs.  $3.39$ ,  $p = .09$  by planned comparisons). As illustrated in Figure 3, the number of exemplars affects more the overall performance for feedback learners than for observation learners. This is consistent with the claim by Juslin et al. (2008) that whenever a cue-abstraction fails to exploit the task environment, it is better to shift to exemplar memory that can act as a back-up process.

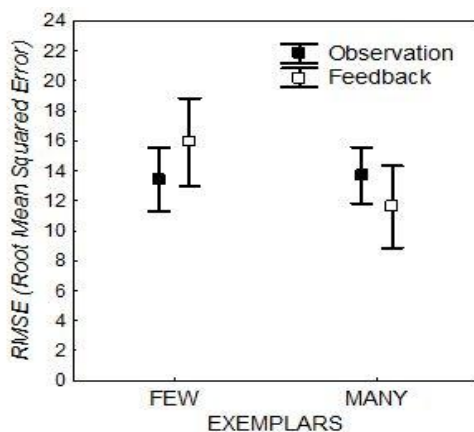


Figure 3. Overall judgment accuracy (*RMSE*) for observation and feedback learners after training with few or many exemplars. Lower value of *RMSE* signifies better judgment accuracy. Vertical bars denote 95% confidence intervals.

## Discussion

This study revealed support for the hypotheses that observation learning is more efficient when few unique exemplars had been experienced, whereas feedback learning is more efficient when many unique exemplars had been experienced in training. The modelfit revealed a dominance

of CAM for both observation and intervention. It is possible that the hypothesized processes invited by the learning formats are more easily detected early in the training as in Henriksson et al. (2012). However, the relative fit of the models and the accuracy of judging old and the matched new exemplars was in line with the predictions and suggested that exemplar memory is invited by observation learning and cue-abstraction is invited by feedback learning. The result is in line with the results reported by Rouder and Ratcliff (2004) suggesting that a rule-based process is better able to exploit a task environment when there are many exemplars and that exemplar memory is better able to exploit a task environment when there are few exemplars.

The results in this paper opens for the possibility that the “rule bias” in many studies on categorization and multiple-cue judgment (e.g., Ashby et al., 1998; Erickson & Kruschke, 1998; Juslin et al., 2008; Kalish et al., 2005) may in part be reinforced by the frequent use of the feedback learning format in experiments. However, with a different learning format such as observation there might have been a “bias for exemplar memory” instead. As Juslin et al. (2008) suggest exemplar processes might act as a back-up process that are used whenever rule-based processes fails.

Ashby et al. (2002) has suggested that observation learning might be a learning format that captures many learning situations for children, as when parents teach their children by pointing to objects or persons in the environment and the child is assumed to learn by observing the characteristics of the object. The results from this experiment are in line with previous research that observation is associated with more exemplar processes (Estes, 1994; Henriksson et al., 2012). There is some evidence that 9 to 11 years olds compared to adults have difficulties using cue-abstraction and instead rely on exemplar processes even when a task environment facilitates cue-abstraction. Not fully matured frontal lobe structures, important for working memory, is one explanation for the observed difficulties in using cue-abstraction among the preteen children (Von Helversen, Mata, & Olsson, 2010). Aging might also affect working memory capacity, and as has been shown in categorization, younger adults and elderly perform at similar levels when learning is based on observation learning. But when learning is based on feedback, younger adults outperform older, suggesting that working memory and set-shifting abilities are important in feedback learning (Schmitt-Eliassen et al., 2007). One successful application of the idea of different learning formats is that observation learning seems to offer patients with Parkinson’s disease a way to learn that circumvents their deficits for rule-based processing in categorization (Shohamy et al., 2004). The result from my study suggest that you also can tailor fit the training with few or many exemplars to enhance judgment accuracy by changing the learning format.

In this study, two generic or archetypical cognitive processes in their pure form have been compared, but it is of course possible that people rely on a mix of processes. It is

possible that the invited processes in different learning formats can in combination with demands from the task environment transform into a hybrid process and in the future it is reasonable to incorporate models such as SUSTAIN (Love, Medin & Gureckis, 2004) or the Varying Abstraction Model (Vanpaemel & Storms, 2008) to name a few. In terms of such mixed or hybrid models, the results reported here can be understood as a change in the relative dominance of the two processes, where observation learning invites relatively more EBM and feedback learning invites relatively more CAM.

## References

- Ashby, G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, 105, 442-481.
- Ashby, G., Maddox, T., & Bohil, C. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, 30, 666-677.
- DeLosh, E. L., Busmeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 968-986.
- Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, 127, 107-140.
- Estes, W. K. (1994). *Classification and Cognition*. New York: Oxford University Press.
- Henriksson, M. P., Enkvist, T., & Juslin, P. (2012). Bias for rules: A question of learning format and task environment. Manuscript submitted for publication.
- Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple-cue judgment: A division-of-labor hypothesis. *Cognition*, 106, 259-298.
- Kalish, M. L., Lewandowsky, S., & Davies, M. (2005). Error driven knowledge restructuring in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 846- 861.
- Karlsson, L., Juslin, P., & Olsson, H. (2008). Exemplar-based inference in multi-attribute decision making: Contingent, not automatic, strategy shifts? *Judgment and Decision Making* 3, 244-260.
- Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 317-330.
- Love, B., Medin, D., & Gureckis, T. (2004). SUSTAIN: A network model category learning. *Psychological Review*, 11, 309-332.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning, *Psychological Review*, 85, 207-238.
- Nosofsky, R. M. (1986). Attention, Similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology: General*, 115, 39-57.
- Rouder, J. N., & Ratcliff, R. (2004). Comparing categorization models. *Journal of Experimental Psychology: General*, 133, 63-62.
- Shohamy, D., Myers, C. E., Grossman, S., Sage, J., Gluck, M.A., & Poldrack, R.A. (2004). Cortico-striatal contributions to feedback-based learning: Converging data from neuroimaging and neuropsychology. *Brain*, 127, 851- 859.
- Schmitt-Eliassen, J., Ferstl, R., Wiesner, C., Deuschl, G., & Witt, K. (2007). Feedback-based versus observational classification learning in healthy aging and Parkinson's disease. *Brain Research*, 1142, 178-188.
- Vanpaemel, W., & Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, 15, 732-749.
- Von Helversen, B., Mata, R., & Olsson, H. (2010). Do children profit from looking beyond looks? From similarity-based to cue-abstraction in multiple cue judgment. *Developmental Psychology*, 46, 220- 229.

# Identifying Kinematic Cues for Action Style Recognition

Shohei Hidaka (shhidaka@jaist.ac.jp)

Japan Advanced Institute of Science and Technology,  
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

## Abstract

Recognition of emotional states from other's actions is one of key capability for smooth social interaction. The present study provides a computational-theory-level analysis on which feature may take a crucial role in recognition of emotional attributes in human actions represented as point-light display. Lead by the previous theoretical works and empirical findings, the velocity and acceleration profile was investigated as a major feature of emotional attributes classification. The results showed that emotional attributes in actions as well as action types could be identified by covariance of velocity profiles among multiple body parts. Since, despite different velocity profiles in different actions, these features for emotional attributes were found commonly in multiple different actions, it suggests that the action styles may be mediated by an information channel parallel to action types per se.

**Keywords:** Action style recognition; biological motion; emotion; social cognition.

## Introduction

Our bodily motion is coherent, smooth and effortless. From bodily motion, we perceive other's state such as mood, emotional expression, and intention (Blake & Shiffrar, 2007). Perception of other's state takes a crucial role in social context. Although most of us can easily "read" what others intend to do through their actions, there is a significant gap from the physical motion – a set of trajectories of multiple body parts with a large degree of freedom (Bernstein, 1967). Recognition of motion is vitally important to any animal kinds. Detection of another animal, possibly a pray, a predator, or a conspecific, and the following detailed identification what it is and how it may behave is essential to take an emergent actions to it (Johnson, Bolhuis, & Horn, 1985). Humans are social animals. Not surprisingly, our visual system is highly specialized to recognize others' state. The present study aims to provide a computational-level description on how people recognize emotional status in others' actions.

### Perception of biological motion

How do we recognize implicit patterns in different styles of actions? The past experimental literature has explored capacity of motion perception using point-light displays (Johansson, 1973) in which the point-lights attached in major joints are only visible in the dark background (Figure 1a). Thus the available information is point-wise kinematic motion in multiple body parts. Despite of the limited information, people can recognize identity (Troje, Westhoff, & Lavrov, 2005), gender (Kozlowski & Cutting, 1977; Troje, 2002), emotions (Pollick et al., 2001; Atkinson, 2009; Hobson & Lee, 1999), dynamics such as the weight of a

lifted object (Bingham, 1987) of actions from point-light displays.

Not only demonstrating human capacity, the studies using point-light display have suggested features extracted in action perception. Accumulating empirical studies on action perception have suggested that velocity and its higher order derivatives in a single or multiple body parts as one of major correlates to emotional attributes in actions: duration of action (Pollick et al., 2001), velocity (DeMeijer, 1989), acceleration (force or the second order time derivatives) (Chang & Troje, 2008; 2009) and jerk or the third order time derivatives (Cook, Saygin, Swain, & Blakemore, 2009), pairwise counter-phase oscillation (Chang & Troje, 2008; 2009). In particular, we highlight the contribution of the higher order derivatives of velocity and importance of its covariational structure. Of relevance, Chang & Troje (2009) found that, not one of either but a pair of feed motion was a major cue for discrimination of walking direction.

### Past computational models on action recognition

Consistent to these empirical findings, most of the theoretical approach works on some kind of statistical regularities among motion profiles. According to a recent review (Troje, 2008), perception of biological motion has the multi-level processing on local and global motion properties. The feature processing consists of four layers from early (low-level) to late (high-level) processing: life detection, structure-from-motion, action recognition, and style recognition. The system detect autonomous agent, and construct body structure from its detailed analysis, then is followed by more detailed action analysis.

A couple of computational models are available for structure-from-motion and action recognition (Giese & Poggio, 2003; Lange & Lappe, 2006), and a few for post-action-recognition-level style perception (Troje, 2002; Pollick, Lestou, Ryu, Cho, 2002; Davis & Gao, 2004) in vision science. In the model of structure-from-motion and action recognition, the model identifies body structure and subsequently actions from the pixel-based visualization of point-light displays. In Giese & Poggio (2003), the model was built based on neuro-physiological findings on visual cortex, and was applied to recognition of action types and action direction in normal, masked, or scrambled point-light displays.

While, the post-action-recognition-level models for style perception typically assume the either/both 2D or 3D point-light on the major joints and also which action is to be executed is readily available prior to the recognition of action style (Troje, 2002; Davis & Gao, 2004; Pollick et al., 2001). For example, Troje (2002) have proposed a computational model of gender identification in gait

presented as point-light display. The model was built upon the three stages: First a set of postures is encoded based on Fourier decomposition, the low-dimensional projection of extracted features is obtained by principal component analysis (PCA), and then it is fed to classifier (as a similar model, see also Davis & Gao (2004)).

### Simple, transparent, yet general model

Although the previous theoretical works have offered successful pattern recognizer for biological motion, there are three shortcomings. First, most of the studies have been closed in one special type (or its slight variant) of action which often has a unique constraint such as periodicity (e.g., walking, running; e.g., Troje, 2002). Second, related to the first point, a limited number of body parts specialized for each action tends to be (e.g., arm movement for tennis swing (Pollick et al., 2002) ). Although not all the model is specialized, in turn, such a generalized model typically loses transparency of mechanism as a cost of generality (For example, multi-layer physiologically-plausible model, Giese & Poggio, 2003). One of drawback of complex models (using nonlinear filters or feature decomposition technique such as Fourier decomposition and PCA) is that the estimated parameters do not necessarily offer interpretation on which natural features are informative such as body parts and time course (Pollick & Paterson, 2008). Moreover, such model often outperformed human recognition (Troje, 2002; Davis. & Geo, 2004; Pollick & Paterson, 2008) rather than explaining use of features in human recognition.

### The theoretical assumptions in the model

In the present study, we employed the simplest possible framework – a variant of linear regression – in order to characterize the motion cues in whole body interaction for multiple types of actions and emotional attributes. The model has the three major assumptions as follows. (1) The major joint (point-light) is specified and readily available prior to action and style recognition as well as the previous post-structure-from-motion models. Specifically, the point-light coordinate was directly fed to the model. (2) The velocity profile (and its higher-order derivatives) is supposed a primary source of information for style recognition. (3) The model integrates local (single-joint motion) and global (multi-joint motion) in form of linear combination. This is simply implemented as linear regression in which the best linear combination of them was estimated by optimization of recognition/classification performance.

On the other hand, we *do not* assume that action is specified prior to recognition of action style, instead we rather expect to find generalizable features of action style common in multiple types of actions. Since people can recognize different styles in unconventional actions (Moore, Hobson, & Lee, 1997; Hobson & Lee, 1999), a model for human biological motion is required to be general for multiple actions.

Specifically, in the present study, we analyzed the human bodily actions while the actors were given different emotional contexts (Ma et al., 2006). These actions are

experimentally manipulated which emotional context was intended to be under each action performance. Given such a set of human actions, our first goal is to recover the latent emotional attributes which the actor intended to hold (or so experimentally manipulated) from the physical motions. By doing so, we describe how the emotional attributes are expressed in the different bodily actions. More specifically, we focus on the following questions: Is it possible to find a general features regardless of different types of actions? If so, which types of features take crucial roles?

## Biological motion library

Ma et al. (2006) have created an open-access biological motion library, consisting of data recorded from a motion tracking system (point-light actors: Figure 1a). The dataset contains 30 naïve actors each performing 5 actions (walk, knock, lifting, throw, and combinations of the four actions) in 4 emotional contexts (angry, happy, sad, and neutral) (see Ma et al., 2006 for more detail). Each action was performed after the subject was given a background story manipulating the emotional context how the subject is supposed to perform the action. In the present study, we used a subset of actions mainly using right hand, i.e., knock, lift, and throw. As an example, the joint angle of right arm and its angle velocity while 5 repeating the same actions is drawn in 1b.

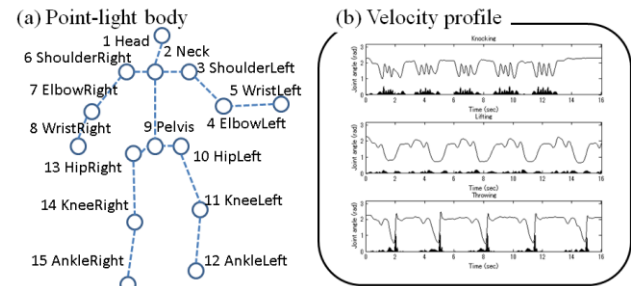


Figure 1: (a) Point-light actor (no link in the model and behavioral test), (b) A temporal profile of right-hand-elbow-shoulder joint angle (solid line) and its velocity profile (black) in 5 repeating knock, lift, and throw actions.

## Model of Emotion Recognition from Actions

We analyzed a subset of the data from biological motion library (Ma et al., 2006). This subset included 15 male and 15 female actors performing 3 different actions (knock, lift, throw) in each of 4 different emotional contexts (neutral, angry, happy, sad). Although the model was trained with actions with neutral emotion as well as actions with emotional attributes, only the three non-neutral emotions were used in the test to facilitate comparison to behavioral data. The additional neutral actions in the training data give the model a chance to learn actions in emotionally neutral context which human subjects have experienced out of laboratory experiment. Each action in each emotional context was repeated 5 times on each of 2 trials, producing 3600 actions in total.

### Features: covariance of velocity, acceleration, and jerk.

In order to implement previous theoretical findings regarding velocity profiles as a cue for biological motion perception, we used the velocity, acceleration, and jerk covariance profiles to identify actions and emotions. Each of these was used to define features for the regression model. Features came in two kinds: local and global. For instance, variance of acceleration is a local motion property (single-point motion) that captures smoothness of motion over an interval. The covariance of acceleration between multiple points is a global motion property that captures the degree of temporal coordination between two body parts. Variance/covariance was evaluated for each action defined by joint-angle of right arm (see also Ma et al., 2006 for details). We used a nested model structure to identify the contribution of each kind of information to action and emotion identification. The simplest model was a velocity-only model that included only the single-point variance and two-point covariances for each joint. This model was subsumed by an acceleration model that also included acceleration variance and covariance, and both were nested in a model that included jerk variance/covariance information. Since, at each moment, velocity and acceleration of 15 body parts were obtained, 15 variances and 105 covariances were obtained for each action. Thus, a total 120, 240, or 360 dimensions across pairwise body parts were used for classification in the Velocity, Acceleration and Jerk models, respectively. Because it produced the most parsimonious fits to behavioral data, the Acceleration model was analyzed most extensively, and is the model discussed if a different model is not mentioned specifically.

### Classification with automatic dimension reduction

These normalized variance and covariance features were used to classify emotions and actions using a multi-class sparse logistic regression model (Yamashita et al., 2008). Model parameters were estimated in a hierarchical Bayesian framework, which penalizes parameters that do not contribute significantly to improving prediction. This is done with a sparsity that reduces the likelihood of the model in proportion to the number of non-zero parameters ( $w$ ) multiplied by a scaling parameter  $\lambda$  (Figure 1c for its graphical model). Specifically, the probability of each action  $i$  belonging to class  $k$   $P(y_{ik})$  follows the multinomial distribution with probability represented as logistic function of linear combination of the given features  $x_{ij}$  for data  $i$  of dimension  $j$  with the weights  $w_{jk}$  as follows.

$$p(y_{ik}) \propto \left( 1 + \exp \left( \sum_j x_{ij} w_{jk} \right) \right)^{-1} \quad (1)$$

The loglikelihood of class of data given by Equation 1 combined with the prior probability on the weights  $w$  is maximized. The sparseness prior is given as follows.

$$p(w_{jk}) \propto \sqrt{\lambda_{jk}} \exp(-2^{-1} \lambda_{jk} w_{jk}^2)$$

$$p(\lambda_{jk}) \propto \lambda_{jk}^{-1}$$

where weights follows the gamma distribution with the hyper parameter  $\lambda_{jk}$  which follows a fixed-parameter gamma distribution (Jeffrey's prior). This prior prefers zero-value

weights, and thus penalize non-zero weights without sufficient information to classification of the given data.

Thus, without any free parameter to adjust, most of weights on non-relevant dimensions were supposed to be excluded from the model on course of optimization.

The each action in the dataset is randomly assigned either training or test samples. The 3300 training samples were used to estimate the parameters in the classification model, and its performance with the 300 test samples was evaluated. The reported results were averaged across 10 randomly generated sets of test/training samples.

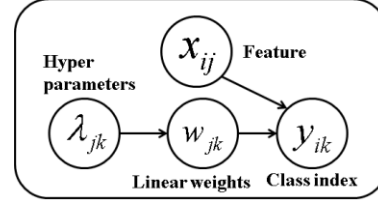


Figure 2: The sparse logistic regression linking the velocity features to given emotion/action class.

### Classification performance in the model

The average correct classification of the Acceleration model with velocity and acceleration profile as features was 97.5% for action types and 69.73% for emotion classes at 33% as chance level of both action and emotion classification. The response patterns of the model in emotion classification were shown at the bottom panel in Figure 2. In order to evaluate the model's prediction on the action/emotion classification, we conducted the behavioral study on action/emotion classification. The detail of the model's prediction would be discussed with the behavioral data.

### Behavioral study: action/emotion classification

In order to test the prediction of the proposed model, we run behavioral experiment in which adult participants were asked to classify the type of actions and emotions given human actions presented as a point-light display. A subset data from the biological motion library was used as stimuli.

#### Participants

10 graduate students at Indiana University were recruited.

#### Material

Action-emotion stimuli were sampled from the biological motion library (Ma et al., 2006). Nine pairwise combinations of 3 actions (knock, lift, and throw) and 3 emotions (angry, happy, sad) were sampled from each of 3 selected actors. This yielded 27 video clips in total. The viewing angle was fixed to look down the actor from actors' left side, so that the view capture the both front and side aspects of actions. From this fixed angle, point-lights from different joints rarely overlapped.

#### Procedure

The experimental procedure consisted of two separated phases. In familiarization phase, 9 video clips (3 actions by 3 emotions) which were not used in the following phase were presented on a computer monitor simultaneously. Each was accompanied by a label identifying both the action and emotion in the clip. Participants were told that they would

be asked to categorize similar clips by action and emotion, and that they should watch the clips until they instructed to watch the clips until they were satisfied that they felt they could do so.

A test phase followed immediately after the familiarization phase. Each participant watched a series of 15 second in which a point-light actor performed one of the 3 actions in the style of one of the three emotions. Participants were asked to determine the action and emotion in each video. Presentation of the stimulus on each trial was ended either when the clip ended or when the participant pressed a button to advance to the next trial. The test phase consists of 27 trials, with presentation order randomized by participant. Together, the familiarization and test phases lasted approximately 10 minutes for each participant.

## Results and Discussion

The proposed model provides quantitative prediction on classification of emotional attributes based on statistical structure in velocity profiles. Here we compared classification performance of action and emotion in the human behavior and the models. The correct ratio in action classification was nearly perfect in both human (98.61%) and the Acceleration model (97.5%) to chance level 33%. The correct ratio in emotion recognition was comparably medium level in both human (68.98%) and model (69.73%) to chance level 33%. The result showed the model achieved comparable performance in biological motions in both action and emotion classification.

### Data fitting and comparison of models

Since the action classification was nearly at ceiling, we analyzed the classification error patterns emotional classification in detail (Figure 3). Figure 3 shows the proportion of responses for each type of emotion. In order to analyze which kind of feature human subjects utilized, we compared the three variant of the models with nest-structure feature sets: *Velocity* model with only velocity profile, *Acceleration* model with velocity and acceleration profile, *Jerk* model with velocity, acceleration, and jerk (up to the third order derivative). The goodness-of-fit for each model was evaluated to what extent human responses in the behavioral data followed a multinomial distribution with the average proportion of responses in each model as parameters. Note that, although the feature set in the models were different, none of the three models were optimized for fitting of the behavioral data (thus no free parameter). Instead they were optimized to classify the emotional attributes in actions. The log-likelihood of data for Velocity, Acceleration, Jerk model were -93.931 ( $R^2 = 0.810$ ), 90.051 ( $R^2 = 0.890$ ), and -89.116 ( $R^2 = 0.900$ ), respectively. The likelihood ratio test revealed significant difference in likelihood of Velocity model from the other two models ( $\chi^2(1) > 3.8807$ ,  $p < 0.05$ ), but did not find significant difference between Acceleration and Jerk model ( $\chi^2(1) = 0.7479$ ,  $p = 0.33$ ). This result of model comparison suggested that velocity profile alone was not sufficient to capture behavioral patterns, but velocity and acceleration profile

might be sufficient since the additional jerk profile made little additional contribution for data fitting. Therefore, hereafter we analyzed the Acceleration model as the representative model.

### Action-specificity of emotion attributes

Next, we tested the hypothesis that recognition of emotion attributes is specific to each action types. If so, the model trained to classify emotional attributes for each action (*Action-specific* model) would capture behavioral patterns better than the model trained to classify them for all the three actions together. The log-likelihood of Action-specific acceleration model was -90.6381 ( $R^2 = 0.890$ ) which is slightly worse but not significantly different from that of non-action-specific acceleration model ( $\chi^2(1) = 0.5871$ ,  $p = 0.444$ ). Therefore, the action-specific model did not necessarily offer a better account for human recognition.

### Classification with only average velocity

One of largest qualitative difference between human and model was found in the proportion of response “happy” to angry action: human recognizers confused angry with happy more than with sad, whereas the model recognizers confused it with sad more than happy. According to the post-experimental interview to the participants, many of them reported that they relied on average velocity of actions. Typically angry actions tended to be fast, sad actions were slow in the current stimuli, and happy ones were in the middle of them. This may be a potential reason why for human perceiver angry actions tended to be confused with happy ones rather than with sad ones, and also the model did not included as its features for classification.

Therefore, in order to evaluate the contribution of average velocity, we performed additional analysis. A past study has reported that the average velocity, or duration of from beginning to end of action, (and its correlated factors such as duration) was one of major correlates to subjects’ rating of emotion attribute (Pollick et al., 2001). Indeed, we found the angry actions were fast and sad actions tended to be slow on average in the data used in the present study. The four-way ANOVA on the factor (emotion types, action types, repetition, and trial), revealed the significant main effect of all the factors but trials ( $p < 0.01$ ).

However, the average velocity of the actions alone was not enough for classification of action style. The correct ratio using the average velocity, duration, peak velocity of each action was 35.9% (chance level 25%) which was not comparable to human performance (68.98%). Even after we classify the subset of data separate for each action, the classification performance did not improve significantly (Average of three actions: 36.4, Knock: 34.1, Lift: 38.0%, Throw: 37.1% for the chance level 25%). This result suggested the average velocity of actions alone could not fully explain the action style recognition.

In sum, the current model-based analyses suggested that the covariance profile of acceleration in multiple body parts carried significant amount of information on emotion attributes in actions. Despite of very different velocity profiles in three actions, knock, lift, and throw, the



classification of emotion regardless of the actions was as successful as action-specific classification. This result suggested that, to some extent, emotional attributes in actions were more general rather than specific for each action.

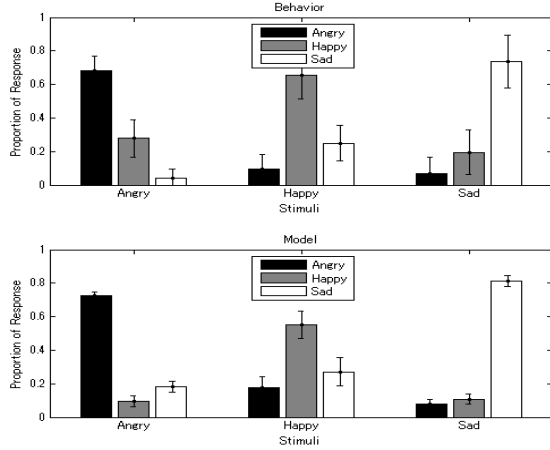


Figure 3: The response patterns for each emotion type in human subjects (upper panel) and the model (bottom panel).

### Distributed cues in emotion recognition

Next, we analyzed the effective feature dimensions for each emotion attribute in the Acceleration model. In the sparse logistic regression, the fewest possible feature dimensions were automatically selected among all the given dimensions. The selected dimensions, variance and covariance of acceleration profiles, were supposed to be a spatiotemporal “template” informative to action/emotion attributes. Thus we analyzed which features are specified for each of emotion attributes. Figure 4 depicted the inter-connection between body parts which were identified as significant features for emotion discrimination (see also criterion of the dimension selection in the model). In each panel of Figure 4, the thin and thick lines indicate covariance of velocity and acceleration for a pair of body parts respectively which is also coded by intensity in the adjacent matrix. The number of effective dimensions for velocity/acceleration and local (variance)/global (covariance) was shown at the bottom right panel.

We found the numbers of effective velocity dimensions (either local or global) were consistent with the average velocity: the largest number of effective dimensions in angry actions which tended to be performed fast, meanwhile the smallest number of them in sad actions which tended to be performed slow. Also the total number of effective dimensions for each action was consistent to the classification performance (Figure 3): the model found fewest effective dimensions for happy actions and had lowest accuracy in identifying them. Overall, we found more global features (pairwise covariance) than the local features (single-point variance). This result suggested that the emotional attributes were distributed rather than specific to a small number of body parts. Since these patterns found in the present model were directly interpreted as those on

body parts or their relationship, they would offer a specifically testable prediction on which body parts may be potentially informative.

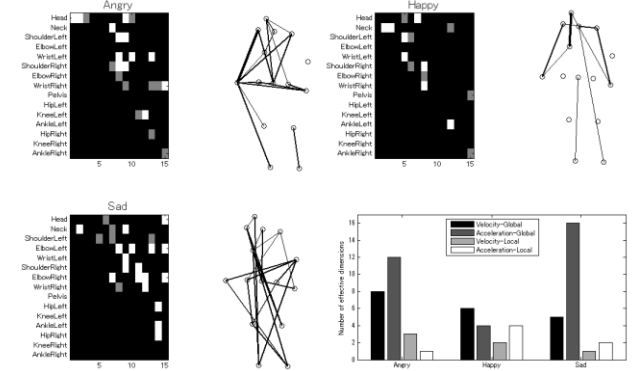


Figure 4: The variance/covariance in velocity profile significantly relevant to each emotion attribute mapped on a body scheme. The white and gray cell indicates effective variance/covariance of velocity and acceleration, respectively. No lower triangle cells were presented due to its symmetricity. The bottom right panel showed the number of effective dimensions for each emotion attribute.

### General Discussion

In the present study, we provided a computational model which specifies characteristic kinematic features for recognition of emotional attributes in actions. Following the lead of the past studies, we analyzed the velocity profile with special attention. Classification with covariance of velocity and acceleration profile among multiple joints showed comparable performance as good as human classification. Moreover, by comparing multiple models trained with different feature types, it suggested that (1) velocity alones was not sufficient but combined with acceleration or higher order derivatives might characterize the human emotion recognition, and (2) there may be common emotional attributes invariant to action-specific motion profiles.

#### Action style as parallel process rather than hierarchy

The present analysis showed that, based on covariance of velocity profile across whole body, emotion attributes may be characterized beyond specificity of each action. However, recent review on action recognition offers a contradictory view to the present study: recognition of action style needs action recognition in prior to it. According to a recent review (Troje, 2008), perception of biological motion is multi-level processing on local and global motion properties. The feature processing consists of four layers from early (low-level) to late (high-level) processing: life detection, structure-from-motion, action recognition, and style recognition. Once both agent and action are identified, pattern recognition at a “subordinate” level (Rosch, 1988) helps to retrieve further information about the details (i.e., action style) of both.

In the present study, we propose an alternative view on action style perception: The emotion attributes can be identified with or without pre-specification of action types.



In the present analysis, we found that the model without action-specific features can account for behavioral classification performance as well as that with action-specific features. Therefore, we speculate that action style is “parallel” process rather than hierarchical one to action recognition which may be coded independently from the action types.

### Future works

One of the future directions is to extend the behavioral study so that we can evaluate subjects’ attention to body parts and its time course using an additional measure (e.g., eye movements). The extended behavioral study would allow us to directly test the model’s detailed prediction about which body parts and their relationship may be informative to action/emotion classification (Figure 4).

### Acknowledgments

The author is grateful to Daniel Yurovsky for his discussion on the current manuscript. This study was supported by Artificial Intelligence Research Promotion Foundation and Grant-in-Aid for Scientific Research B No. 23300099.

### References

- Atkinson, A. P. (2009). Impaired recognition of emotions from body movements is associated with elevated motion coherence thresholds in autism spectrum disorders. *Neuropsychologia*, 47, 3023–3029.
- Bernstein, N. A. (1967). *The coordination and regulation of movements*. Oxford: Pergamon.
- Bingham, G. P. (1987). Kinematic form and scaling: Further investigations on the visual perception of lifted weight. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 2, 155–177.
- Blake, R. & Shiffrar, M., (2007). Perception of Human Motion. *Annual Review of Psychology*, 58, 47–73.
- Chang, D. H. F., & Troje, N. F. (2008). Perception of animacy and direction from local biological motion signals. *Journal of Vision*, 8, (5):3, 1–10.
- Chang, D. H. F., & Troje, N. F. (2009). Acceleration carries the local inversion effect in biological motion perception. *Journal of Vision*, 9, (1):19, 1–17.
- Chang, D. H. F., & Troje, N. F. (2009). Acceleration carries the local inversion effect in biological motion perception. *Journal of Vision*, 9(1):19, 1–17.
- Cook, J., Saygin, A. P., Swain, R., & Blakemore, S-H., (2009). Reduced sensitivity to minimum-jerk biological motion in autism spectrum conditions. *Neuropsychologia*, 47, 14, 3275–3278.
- DeMeijer, M. (1989). The contribution of general features of body movement to the attribution of emotions. *Journal of Nonverbal Behavior*, 13, 4, 247–268.
- Giese, M. A. & Poggio, T. (2003). Neural Mechanisms for the recognition of biological movements, *Nature Reviews Neuroscience*, 4, 179–192.
- Hobson, R. P. & Lee, A. (1999). Imitation and Identification in Autism, *Journal of Child Psychological Psychiatry*, 40, 4, 649–659.
- Hubert, B., Wicker, B., Moore, D. G., Monfardini, E., Duverger, H., Fonseca, D. Da, Deruelle, C. (2006). Recognition of Emotional and Non-emotional Biological Motion in Individuals with Autistic Spectrum Disorders. *Journal of Autism Developmental Disorders*, 37, 7, 1386–1392.
- Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14, 2, 201–211.
- Johnson, M. H., Bolhuis, J. J., & Horn, G. (1985). Interaction between acquired preferences and developing predispositions during imprinting. *Animal Behaviour*, 33, 1000–1006.
- Lange, J., & Lappe, M. (2006). A model of biological motion perception from configural form cues. *Journal of Neuroscience*, 26, 11, 2894–2906.
- Ma, Y., Paterson, H. M., Pollick, F. E. (2006). A motion capture library for the study of identity, gender, and emotion perception from biological motion, *Behavior Research Methods*, 38, 1, 134–141.
- Moore, Hobson, & Lee (1997). Components of person perception: An investigation with autistic, non-autistic retarded and typically developing children and adolescents., *British Journal of Developmental Psychology*, 15, 401–423.
- Pollick, F. E., Lestou, V., Ryu J. Cho, S-B. (2002). Estimating the efficiency of recognizing gender and affect from biological motion., *Vision Research*, 42, 2345–2355.
- Pollick, F. E., Paterson, H., Bruderlin, A. & Sanford, A. J. (2001) Perceiving affect from arm movement. *Cognition*, 82, B51–B61.
- Pollick F. E., Paterson, E. (2008). Movement style, Movement features, and the recognition of affect from human motion, In Shipley, T. F. & Zacks, J. M., *Understanding Events from Perception to Action*, New York: Oxford University Press, 286–307.
- Rosch, E. (1988). Principles of categorization. In A. Collins & E. E. Smith (Eds.), *Readings in cognitive science* (pp. 312–322). Sam Mateo: Morgan Kaufmann.
- Troje, N. F. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 2, 371–387.
- Troje, N. F. (2008). *Biological motion perception*. In Basbaum, A. et al. (Eds.), *The senses: A comprehensive reference* (pp. 231–238). Oxford: Elsevier.
- Troje, N. F., Westhoff, C., & Lavrov, M. (2005). Person identification from biological motion: effects of structural and kinematic cues. *Perception & Psychophysics*, 67 (4), 667–675.
- Pollick, F. E., Paterson, H. M., Bruderlin, A., Sanford, A. J., (2001). Perceiving affect from arm movement. *Cognition*, 82, B51–B61.
- Yamashita, O., Sato, MA., Yoshioka, T., Tong F., Kamitani Y. (2008). Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *Neuroimage*. 42, 4, 1414–29.

# The Atoms of Cognition: A Theory of Ground Epistemics

Seng-Beng Ho ([hosengbeng@nus.edu.sg](mailto:hosengbeng@nus.edu.sg))

Temasek Laboratories, National University of Singapore  
5A Engineering Drive 1, #09-02, Singapore 117411

## Abstract

We propose a set of “atomic cognitive operational representations” on which higher level cognitive representations and processes can be built, thus providing fundamental building blocks for cognitive mechanisms necessary for intelligent actions. The fundamental concepts involved are elemental temporal changes of some quantities and in this representational scheme the temporal dimension is explicitly represented to fully characterize the meanings of the concepts involved at the epistemic ground level. This provides full grounding for all subsequent concepts that are built upon them, allowing cognitive systems embodying these concepts to have full and complete understanding and characterization of the concepts involved that it can use for various cognitive ends. This provides a firm theoretical foundation for the study of cognition and intelligence.

**Keywords:** representation; operational representation; spatiotemporal representation; conceptual grounding; fundamental building blocks of cognition; experiential memory

## Introduction

Unlike the physical sciences, “standard” paradigms for scientific investigation of cognitive phenomena still do not exist for the sciences of cognition and intelligence (Arbib, 2002; Gazzaniga, 2008; Russell & Norvig 2010). Chief among the achievements of the physical sciences is the discovery of various fundamental particles and forces that provide the foundation for the understanding of physical reality. These fundamental particles and forces form the necessary and sufficient building blocks upon which the characterization of all other higher level physical phenomena can be constructed.

We show that a set of representations, which we refer to as “atomic cognitive operational representations” or *ACORs*, can perform an equivalent function of providing the fundamental building blocks for building cognition. Like elementary particles and forces, *ACORs* allow ground level semantics to be represented and all higher level semantics and cognitive operations to be based on them. As a dual of physical ontology, the atomic cognitive operational representations provide the *cognitive ontology* for building all cognitive phenomena.

In our theory, a set of correctly formulated ground level “atomic” building blocks of cognition will be necessary for allowing all cognitive processes to be constituted. Furthermore, the same fundamental building blocks apply at all epistemic scales to enable useful knowledge to be derived through cognition for intelligent functioning. Consider the following.

Before the advent of science, we dealt with the world as best we could, at a level of description provided by our natural sensory systems. Take for example the case of a tree as perceived by our human senses. From a distance, a tree can be sensed and conceptualized as consisting of a trunk, some branches and many leaves. However close up, our senses can tell the texture and the detailed shapes of the trunk, the branches and the leaves, and perceive their movements. The detailed perception of these subparts of a tree in turn provides us with the necessary knowledge to be able to use those subparts for various purposes, such as for decoration or other more functional ends. When the wind blows and the leaves on a tree move, the softness of the movement might allow us to “imagine” using the leaves as a broom for the purpose of sweeping, or as a fan to fan oneself; in due course, we might proceed to act on those ideas when the need for them arises. This would be the characterization of the tree at an epistemic ground level at which normal perception operates to provide useful information for a cognitive system to function intelligently.

As science improves our understanding of the natural world, our conceptualization of trees goes beyond just trunk, branches and leaves as they appear to our natural senses. We discovered chlorophyll and the photosynthetic processes, for example. These biological processes involve much smaller entities – various molecules – and attendant complicated interactions that our natural senses cannot detect directly. With this deeper understanding, we reach another epistemic ground level. It is the thesis of this paper that no matter which ground level we are looking at – i.e., whether it be the earlier one reached by our natural senses or the deeper ones reached through scientific means – the same fundamental building blocks of cognition are involved in subserving cognitive mechanisms from which intelligent actions emerge.

With the fundamental building blocks, we put the sciences of cognition and intelligence on a firm theoretical foundation, much as what the theory of fundamental particles and forces has done for the physical sciences.

## The Atomic Operational Representations of Appearance and Existence

The vast canvas of space-time is the arena in which physical and mental processes take place. One fundamentally important concept that any mental process must deal with is appearance and existence (of objects, events and processes). We begin with the formulation of the atomic concept of appearance.

Figure 1a shows a one dimensional space with 5 discrete positions  $x_0 - x_4$ . At time  $t_0$  it contains nothing. At time  $t_1$ , something appears at position  $x_1$ . It could be an elemental bit of substance or a point of light. Now, assuming an intelligent system has perceptual detectors that can sense and signal something appearing at a certain spot at a certain time and has an explicit *experiential memory* that can store its perceptual experiences and lay them out in a spatial extent for simultaneous processing, Figure 1b shows what the explicit experiential memory looks like in the time interval  $t_0$  to  $t_1$ . Figure 1b also shows a spatiotemporal “appear” template that can pick up this change in the physical situation in the one dimensional space. Basically this template encodes the meaning of “appear.” It captures the situation in which in a moment in time just prior to the appearance of something, there was nothing in that point in space and in the next moment something came into existence. In the spatiotemporal template, time is spatialized and the temporal dimension is explicitly represented.

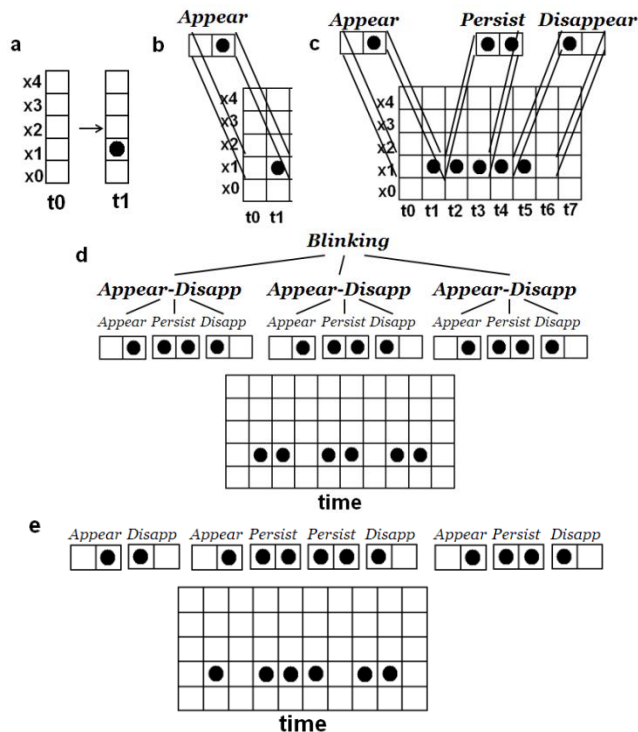


Figure 1: Atomic operational representations for “appearance” and related concepts. a. Appearance of an object at a specific time and location. b. The experiential memory and the “appear” operational representation. c. “Persist” and “disappear” operators. d. Concept of “blinking.” e. Re-composing operators – cognitive manipulatability.

It should be stressed that the elemental “blob” that is picked up and represented in the experiential memory of Figure 1b may or may not correspond exactly to an elemental “blob” in the physical world. It is an “elemental”

bit of occurrence as characterized by the sensory organ of the cognitive system relevant to the cognitive task at hand. Depending on the resolution of the sensory system and the cognitive task involved, the “blob” could correspond to an atom in the physical world, a vehicle or a person, or the recognizable points on a human body or on the leaves of a tree. It could also be something created by other internal mental processes in a “mental space.”

We submit that this spatiotemporal *appear* template captures the full meaning of *appear* at the ground level. For the purpose of identifying an *appear* event, the template is used as follows: it is matched to the spatiotemporal patterns in the experiential memory as shown in Fig 1b. If there is a match, an *appear* event is identified to have taken place.

The *appear* template, other than being useful for the purpose of identifying or recognizing the occurrence of an *appear* event, is also usable in the opposite direction – that of *generating* the action or idea of *appear*. Assuming that an intervening system is available to translate an intelligent system’s intention into physical realization, the intelligent system can act, under the direction of the *appear* template, to move a material substance into a specific point in space and thus making it “appear” at that location, to materialize a bit of substance from nowhere at a specific point in space, or to cause a point of light to appear at a specific point, etc. These are examples of “make-appear” – the *appear* template used in the generation direction. The *appear* template can also be used to generate an “idea” (i.e., a bit of “mental substance” – e.g., the mental analog of a corresponding physical quantity – or something more abstract) at a point in a mental space for the purpose of mental manipulation. An example of a mental space would be the spatiotemporal representational structures that hold the experiential memory for mental manipulation, such as the structure shown in Figure 1c that might be implemented using a computer memory. Generation of ideas in the mental space is also sometimes referred to as “imaging” or “imagining” (Kosslyn 1994). We will refer to this characteristic of these atomic operational representations as “operational bidirectionality” – i.e., they can be used for recognition as well as generation.

We thus term the *appear* spatiotemporal template an “operational representation” as it encodes the meaning of an operational concept – *appear* – directly in terms of the operations that it performs. Such an operational representation is *atomic* in that it captures knowledge at the ground level. An atomic operational representation is *elemental* in that it represents a smallest discrete change in the spatial and temporal dimensions relevant to the cognitive task at hand. A more succinct description of “atomic operational representations” is “atomic operators.”

Figure 1c depicts two other concepts related to *appear* – “persist” and “disappear.” Together with *appear*, they can be used to capture the process of something appearing at a point, persisting (existing) for some time, and then disappearing. Figure 1b and 1c also suggest that these

templates or operators can be picked up from the environment, in the same manner as the “cookie-cutter” approach described by Uhr and Vossler (1981) – i.e., when a subpattern is perceived in the environment, it is simply “cut-out”/“picked-up” and stored as an elemental pattern to be used later to match with further pattern information coming in through the perceptual system to recognize future occurrences of the subpatterns. This is usually known as an unsupervised learning process (Duda, Hart & Stork, 2001; Fukushima, 1988; Malsburg, 1973). Whereas in Uhr and Vossler (1981), the “cookie-cutters” work on static images containing certain spatial patterns, thus they extract sub-features of the *spatial* patterns for the purpose of characterizing these patterns, the corresponding application here would be extracting *spatiotemporal* sub-patterns that serve to characterize activities or changes in the environment, external or internal to the cognitive system.

Figure 1d shows how higher level concepts can be built upon the atomic operators of *appear*, *persist* and *disappear* – if something alternately appears and disappears over time, the concept of “blinking” can be used to encapsulate the process. The higher level concepts are firmly grounded through the atomic operational representations. This process is termed “cognitive hierarchy construction.”

Figure 1e shows how the atomic operators can be recombined into novel sequences that may not have been encountered in the environment earlier in a generation process through the conceptual hierarchy, directed by reasoning and problem solving requirements. This characteristic is referred to as “cognitive manipulability” – i.e., these representations can be directly manipulated in cognitive processes.

Base on the foregoing discussion, the critical characteristics of operational representations in general and atomic operational representations in particular are summarized as follows:

1. *Explicit Temporal Representation* - the temporal dimension is explicitly represented in operational representations – this requires the intelligent system utilizing the representation to have an explicit experiential memory – temporal changes are laid out in a spatial extent for simultaneous processing.
2. *Elemental Representation* – atomic operators are elemental in that they represent the smallest discrete changes in the respective dimensions.
3. *Cognitive Hierarchy Construction* – higher level concepts can be built directly upon the lower level as well as the atomic level operational representations.
4. *Operational Bidirectionality* - operational representations can be used for recognition as well as generation purposes. Recognition and generation can involve things, events and processes in both the physical environment and the mental space.
5. *Unsupervised Extraction* – atomic level as well as higher level operational representations can be extracted from the environment in an unsupervised learning process.

6. *Grounding of Representations* – atomic operational representations are grounded directly in the environment or the mental space, higher level representations are grounded on the atomic representations.

7. *Cognitive Manipulability* – The operational representations, atomic or higher level, can be recomposed and cognitively manipulated in cognitive processes.

The atomic operational representations capture and encode “meaning” in the operations themselves. For example, *appear* is not defined in terms of other atomic or non-atomic concepts but is instead defined directly in the recognition and generation operations that it stipulates.

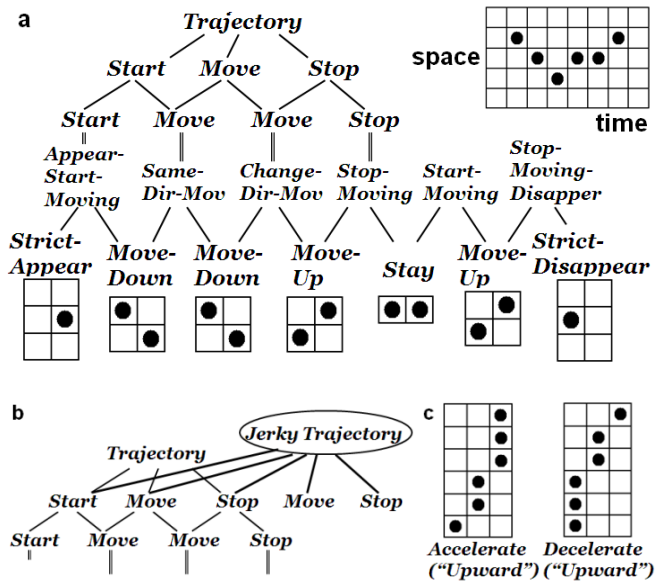


Figure 2: Operational representations of movements. a. Up and down movements in a one-dimensional space and the associated operational representations and conceptual hierarchy. b. A “jerky trajectory.” c. Accelerate and decelerate operators

## Atomic Operational Representations of Movement

Again, using events in a discrete one dimensional space, we illustrate the atomic operational representations associated with movements in Figure 2. In Figure 2a, we show the operational representational characterizations of a point object appearing at a certain location, moving “up” and “down” in the one dimensional space, and then disappearing. There are basically two kinds of spatiotemporal templates associated with atomic movements – “move-up” and “move-down.” Here we also introduce a more specific kind of *appear* template called “strict-appear” – i.e., the appearance of something in a specific location is not caused by things moving into the location. If the earlier, more general *appear* in Figure 1 is used, then at every location where the point object moves into, the *appear*



template will be triggered. Similarly for *disappear* – we introduce a “strict-disappear” template. The “stay” template is similar to the *persist* template in Figure 1. It is interesting to note that the concept of *stationarity* – “stay” – can only be represented adequately if there is an explicit temporal dimension that can capture and encode the meaning of “no change in position in time.” Similarly for the concept of “persist” – “no change in the state of existence” – depicted in Figure 1c.

Figure 2a shows a cognitive hierarchy built upon the atomic operational representations of appearance and movement, reaching a level where the concept of “trajectory” is formed. Other than *trajectory*, concepts such as “jerky trajectory” can emerge from the conceptual hierarchy, as shown in Figure 2b.

Figure 2c shows the concepts of *acceleration* (an “upward” acceleration) and *deceleration* (an “upward” deceleration) captured in their corresponding operational representations. Again, here we can see that the *meanings* of *move-up*, *move-down* or *acceleration* and *deceleration* are captured directly in the corresponding associated operations stipulated by the operational representations.

### Atomic Operational Representations of Scalar and General Parameters

At any given point in physical space, a physical quantity, such as the hardness of some substance or the brightness of a point of light, may change elementally. Similarly, in mental space, the strength of an idea or feeling may also change elementally. These non-spatial scalar parameters are represented in the same manner as the spatial parameters as atomic operational representations as shown in Figure 3a, using the examples of a physical parameter such as “brightness” or a mental parameter such as “pain” that increases and decreases over time.

Atomic movement through space (physical or mental), which involves an elemental change in spatial position, can be considered a special case of a change in something, so does a change in the state of existence in space – i.e., *appear* or *disappear*. Therefore, the most general characterization of all the atomic operators including those that capture non-spatial parameters is that they characterize an elemental change of these parameters, spatial or otherwise, across elemental time. Figure 3b shows a generalized atomic operational representation of parameter change – like a “God particle” of atomic operators – that subsumes all possible atomic operators.

The number and kind of atomic operators that can be found in a particular intelligent organism or system depends on the parameters that are necessary for its survival or intelligent operations. For example, if an organism needs salt for survival and has the necessary sensory apparatus to detect the presence of salt with varying intensities over time, then for the purpose of cognitive processing it would need “saltiness” atomic operators. In conscious perception such as that typically experienced and reported by human beings, the atomic level parameters typically appear as indivisible

cognitive entities endowed with certain “qualia” such as pain, sadness, anxiousness, brightness, darkness, redness, saltiness, sweetness, spatiuousness, etc. These atomic level cognitive parameters as appear to consciousness can vary in their intensities over time but otherwise have no further internal structures scrutinizable by consciousness.

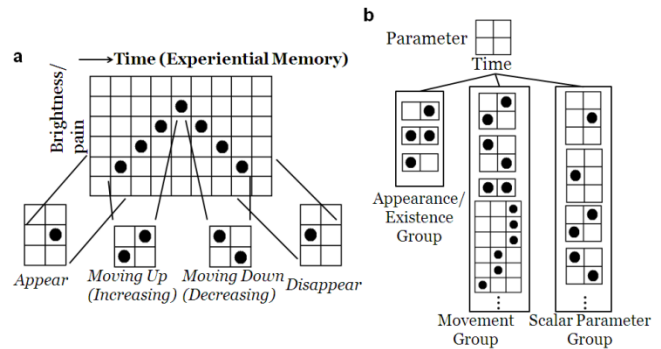


Figure 3: Other operational representations. a. Atomic operational representations of scalar parameters (brightness or pain) increasing and decreasing over time. b. Generalized atomic representation of parameter change.

### Representing and Reasoning about Time

If a cognitive system provides for the explicit representation of time through an experiential memory such as that described above, it can lead to another interesting mental operation, which is the conceptualization of time as a “thing” and the “movement” of it mapped onto spatiotemporal dimensions in the cognitive system’s *mental space-time* that allows the cognitive system to cogitate about time itself. Figure 4 shows the representations and operations involved for this.

In Figure 4 we re-label the horizontal axis as “cognitive time,” to distinguish from the “real” physical time out there in the real world. (All previous labels of “time” in Figures 1 - 3 should rightfully be “cognitive time” as these operational representations are *mental* entities.) Here, time is characterized as a “thing” that can “move.” Normally, we would conceptualize time as something that moves on inexorably with a “constant speed.” In the physical universe, it is hard to ascribe “speed” to time as one needs a time reference to talk about speed, and such “super-time” is non-existent. However, in our mental processes, we can freely cogitate about these possibilities. Firstly, we often imagine real-world events going faster or slowing down. Therefore time can “accelerate” or “decelerate,” and that would be applying the *acceleration* or *deceleration* operators as depicted in Figure 2c in the time representation here. We sometimes wonder if time has a beginning or end, and these would correspond to applying the *appear* and *disappear* operators as shown in Figure 4. If we cogitate whether time can come to a halt, that would correspond to applying the *persist/stay* operator. And in physics the concept of time reversal has been invoked to explain certain physical phenomena and that would correspond to applying

a *move-down* operator as shown in Figure 4. The concept of time travel to the past or future would correspond to a sudden jump or accelerated movement to a different “point” in “time” from the present point. Our reasoning processes with time thus use similar mental processes as we use for space, once time is conceptualized as a “thing” and its “movement” mapped onto spatiotemporal dimensions in our mental space-time.

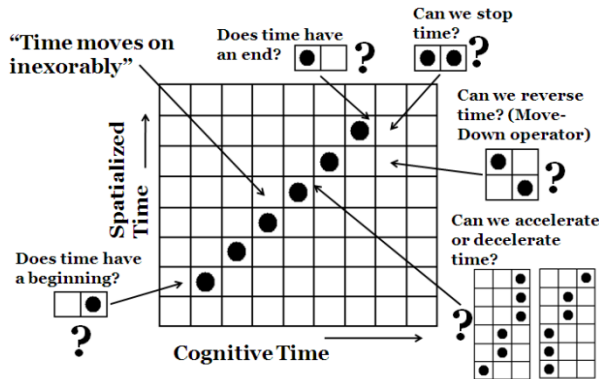


Figure 4: Cogitating about time. Representations capturing the movement, acceleration, deceleration, halting, beginning (appearing), ending (disappearing), and reversal of time for the purpose of cogitation about time.

## Representing Interactions

The interactions between objects can be characterized using operational representations. Figure 5a shows a situation in a one dimensional world in which 2 objects hit each other and are reflected. The concepts of “meet”, “part” and “reflect” can arise in this situation as shown in Figure 5a and be encoded in the form of operational representations. Figure 5b shows the situation with an alternative physics where a moving object penetrates a stationary object and no reflection takes place. The concept of “penetration” can likewise be characterized at the ground level through operational representations.

## Extended Spatial Objects and Higher Dimensional Representations

The object to be represented by an operational representation can also be an extended object consisting of more than one point. Each point on the object can be represented by an atomic operational representation. If all the points on an object move in unison, it is a rigid object (the concept of “rigidity” can be characterized by some higher level operational representations building on top of the atomic operators in the same vein as the *blinking* and *trajectory* concepts shown in Figures 1d and 2a respectively). Otherwise, the object is deformable and the deformations can be characterized by how the points on the object move relative to each other likewise through the use of some higher level characterizations of the deformations built-up from the atomic operators.

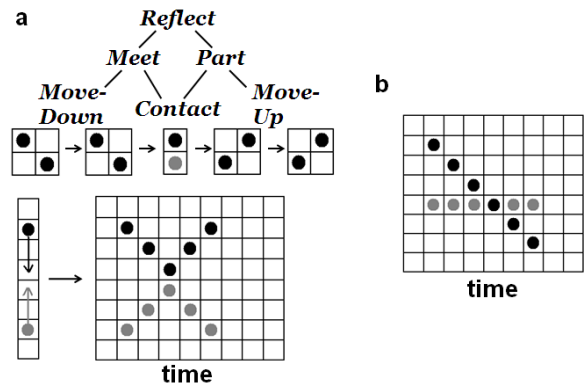


Figure 5: Interactions captured in operational representations. a. Two objects reflect off each other. b. Penetration of one object through another one.

An atomic operational representation that represents elemental movements in one of 8 directions in a 2-dimensional (2D), discrete, and Cartesian space would be a cube of 3X3X1 as shown in Figure 6a. If some liquid on a flat surface moves and changes shape more or less in a plane, its complex shape changes can be captured by the 3D (2D space + 1D time) atomic movement operator as shown in Figure 6a. Higher level characterizations of the liquid movement can be built from the 3D atomic operators.

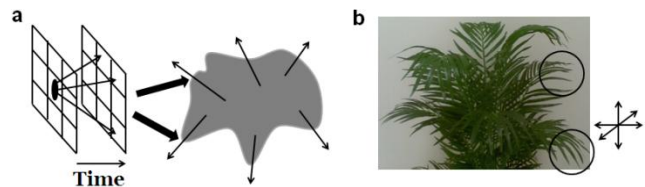


Figure 6: Complex object representations. a. Liquid moving and changing shape in 2D and the associated 3D atomic operator. b. Leaves and 3D micro trajectories of all their elemental movements.

Figure 6b shows a plant’s leaves whose movements can be characterized using the atomic operators that can lead to the intelligent construction of a “broom” or “fan” made from leaves – i.e., having captured and understood a certain soft character of the potential elemental movements of the leaves on the tree, the cognitive system is then able to *imagine* the use of the leaves for the purpose of sweeping dirt or fanning air to cool oneself. For characterizing the leaves, because of 3D movements, a 4D atomic movement operator (3D space + 1D time) is needed. Bare branches will not function correctly as a broom or a fan because their elemental parts lack certain surfaces that can move in a certain “soft” manner. This understanding requires the high resolution ground level characterization of the potential movements of the leaves as captured by the 4D atomic movement operators and higher level characterizations of the movements built on the 4D atomic operators.

The atomic operational representations are the fundamental building blocks of cognition. All events and processes – including the operations of natural phenomena ranging from the micro to the cosmic scales (e.g., from the operations of bacteria to trees, weather, events and processes in the heavens, etc.), operations of human-made machineries (e.g., the operations of the finest mechanical contraptions to the operations of a gargantuan rocket, etc.), and mental operations (e.g., thinking processes in general, abstract mathematical manipulations in the mind, etc.) – are fundamentally cognitively describable based on the atomic operational representations.

## Discussion

It is not a coincidence that the “mental” atomic operational representations described in this paper for the purpose of cognitive processing bear a strong resemblance to the usual “physical” spatiotemporal representations many a scientist and engineer employs to represent physical events and processes for the ease of understanding and manipulating them. This is because there is an intimate link between the physical and the mental world. Spatiotemporal representations are fundamental and lie at the very foundation of the descriptions of all events and processes that can take place in this reality. Cognitive processes are the dual of the physical processes in that they seek to represent the physical processes but they have additional and unique characteristics and operational requirements as described in this paper as they need to manipulate the representations in certain cognitively useful ways.

The grounding of concepts, and hence the learning of the atomic operators, would take place in the early days of a cognitive system’s interaction with the environment (e.g., the infant stage in humans). Future studies of cognitive developmental processes could direct their investigative efforts to uncover whether or how these atomic operators are learned and used.

Our atomic operational representational scheme dictates that the temporal dimension needs to be represented explicitly for the adequate capture of the ground meanings of concepts. Hence, it follows that for a system to qualify as “cognitive,” it must have the capability of representing and manipulating the temporal dimension explicitly. A critical mechanism that subserves explicit temporal representation is the experiential memory (Figure 1b). Neuroanatomically, the experiential memory can be identified with the hippocampus, which is supposedly the neural structure responsible for memory over time – episodic memory (Anderson et al, 2007). In this paper, for the sake of exploring the basic principles we describe the experiential memory as a structure that simply lays out the temporal information in an uncompressed manner. In reality, because of the amount of information in the temporal domain that a cognitive system needs to process, compression in the time dimension is probably necessary. We submit that the hippocampus performs both the experiential memory function as well as the compression function necessary for

its practical functioning. We also submit that for a system to have “cognitive abilities” and be considered a “cognitive system”, it must possess a structure that functions like the hippocampus or the experiential memory that processes and represents temporal information explicitly. This allows it to truly conceptualize and understand events, processes and causality in the world. Even the concept of *stationarity* and *changelessness* can be captured adequately only if there is an explicit temporal representational dimension as discussed above. Hence, artificial or natural neural networks systems that merely reproduce certain input-output mappings of some seemingly “intelligent” processes are “reflexive” and do not qualify to be characterized as “cognitive systems.”

A last and interesting fact to note is that the spatialization of time in operational representations parallels the spatialization of time in relativity which brought us a deeper understanding of the reality in which we inhabit (Lorentz et al, 1923).

## Acknowledgments

I thank Dr. Kenneth Kwok and the anonymous reviewers for their suggested changes to the manuscript.

## References

- Anderson, P., Morris, R., Amaral, D., Bliss, T., and O’Keefe, J. (2007). *The Hippocampus Book*. Oxford UK: Oxford University Press.
- Arbib, M. A. (Ed.) (2002). *The Handbook of Brain Theory and Neural Networks*. Cambridge, Massachusetts: MIT Press.
- Duda, R. O., Hart, P. E. & Stork, D. G. (2001). *Pattern Classification*. New York: John Wiley & Sons, Inc.
- Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1,119-130.
- Gazzaniga, M. S., Ivry, R. B. & Mangun, G. R. (2008). *Cognitive Neuroscience: The Biology of the Mind*. New York: W. W. Norton & Company.
- Griffiths, D. J. (2008). *Introduction to Elementary Particle*. Weinheim: Wiley-VCH.
- Kosslyn, S. M. (1994) *Image and Brain: The Resolution of the Imagery Debate*. Cambridge, Massachusetts: MIT Press.
- Lorentz, H. A., Einstein, A., Minkowski, H., and Weyl H. (1923) *The Principle of Relativity: A Collection of Original Memoirs on the Special and General Relativity*, with notes by A. Sommerfield. Methuen and Company, Ltd.
- Malsburg, C. V. D. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik* **14**, 85-100.
- Russell, S. & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. New Jersey: Pearson Education, Inc.
- Uhr, L. & Vossler, C. (1981) A pattern-recognition program that generates, evaluates, and adjusts its own operators. In Feigenbaum, E. A. & Feldman, J. (eds.) *Computers and Thought*. Malabar, Florida: Robert E. Krieger Publishing Company, Inc.



# Simple heuristic or knowledge-based inference?

## Model comparison of binary choice inference

Hidehito Honda (hito@muscat.L.chiba-u.ac.jp)

Toshihiko Matsuka (matsukat@muscat.L.chiba-u.ac.jp)

Department of Cognitive and Information Science, Chiba University  
1-33, Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522, Japan

### Abstract

We investigated the processes of making inference in binary choice tasks. We proposed statistical models for two simple heuristics and two knowledge-based inferences, and compared how well these models could explain the observed patterns of inferences. It was found that the best model for explaining choice patterns varied depending on the inference problem. In particular, results suggested that participants used a simple heuristic when they had difficulty in retrieving clues pertaining to the inference problem. In contrast, when participants could retrieve enough clues pertaining to inference problems, they made inferences based on these clues.

**Keywords:** simple heuristics; knowledge-based inferences; binary choice inferences; model comparison

### Introduction

In the last decade, a highly controversial topic in research on judgment and decision-making has been the *recognition heuristic* (Goldstein & Gigerenzer, 2002). When the recognition heuristic is applied to a binary choice problem, the inference rule is described as follows:

“If one of two objects is recognized and the other is not, then infer that the recognized object has the higher value with respect to the criterion.”  
(Goldstein and Gigerenzer, 2002, p.76)

For example, consider the following question, “Which city has the larger population, Tokyo or Chiba?” In this problem, the recognition heuristic predicts that when a person recognizes Tokyo but not Chiba, the person will infer that Tokyo has the larger population.

Many researchers have discussed the recognition heuristic from theoretical, empirical, and rational perspectives. For example, in the journal *Judgment and Decision Making*, a special issue on the recognition heuristic was published in 2010 and 2011, which included 24 papers in three volumes.

In particular, researchers have been interested in whether the recognition heuristic is actually used in making inferences. Some researchers have claimed that people often make inferences based on the recognition heuristic in binary choice situations, in which people recognize one of the objects in a pair<sup>1</sup> (Goldstein & Gigerenzer, 2002; Pachur, Bröder, & Marewski, 2008; Pachur & Hertwig, 2006; Reimer &

katsikopoulos, 2004; Snook & Cullen, 2006; Volz et al., 2006). In contrast, others have argued that although people make inferences using the recognition heuristic in some R-U pairs, people often make inferences using knowledge pertaining to the inference problem when such knowledge is available (Bröder & Eichler 2006; Hilbig, Pohl, & Bröder, 2009; Glöckner & Bröder, 2011; Newell & Fernandez, 2006; Newell & Shanks, 2004; Oeusoonthornwattana & Shanks, 2010; Oppenheimer, 2003; Pohl, 2006; Richter & Späth, 2006).

The recognition heuristic can be applied only to R-U pairs. Thus, although many studies have accumulated evidence for both the arguments, they have focused on the processing of inference in R-U pairs. Then, how do people make inferences when they recognize both objects of the pair? Previous studies have assumed that in the case of R-R pairs, inferences are made by using knowledge (e.g., Goldstein & Gigerenzer, 2002; Hilbig, 2010). As described above, many studies have suggested that when people have knowledge in addition to mere recognition, inferences are affected by the knowledge. However, specific models of knowledge-based inferences have not been proposed to date.

Furthermore, people may make inferences using a simple heuristic such as the recognition heuristic, even in the case of R-R pairs. For example, Hertwig, Herzog, Schooler, and Reimer (2008) proposed a *fluency heuristic*, which is an inference strategy using the retrieval fluency of objects, and Honda, Abe, Matsuka, and Yamagishi (2011) proposed a *familiarity heuristic*, which is the inference strategy using the familiarity of objects.

The goal of the present study was to examine binary choice inferences in not only R-U pairs, but also in R-R pairs. Brighton and Gigerenzer (2009) argued that when examining the validity of the recognition heuristic, alternative models should be proposed and compared. Accordingly, we approached the above issue by comparing models. In particular, we examined binary choice of population inferences, by comparing the validity of models representing a simple heuristic and knowledge-based inference.

### Models of binary choice inference

For each pair of cities, the two cities can be ordered by their actual populations, and we name them accordingly. For example, if cities X and Y are presented, and the city X has a larger population than city Y, we call city X the “larger city,” and city Y the “smaller city.”

<sup>1</sup> Hereafter, depending on the recognition of objects in a pair, we use following abbreviations: R-U (Recognized-Unrecognized), R-R (Recognized-Recognized), and U-U (Unrecognized-Unrecognized).

## Simple heuristic

We examined inferences in not only R-U pairs, but also in R-R pairs. Simple heuristics that can be applied to both R-U and R-R pairs are appropriate for this purpose. Hence, we used the fluency heuristic and the familiarity heuristic for modeling simple heuristics.

**Fluency heuristic.** This model is based on Hertwig et al. (2008). They have defined the model as follows: If two objects, a and b, are recognized, and one of the two is fluently retrieved, then infer that this object has the higher value with respect to the criterion (p. 1192). Based on this definition, the response pattern in pair  $i$  in the binary choice task is represented by the following variable,  $Flu_i$ :

$$Flu_i = \frac{Flu_{Si} - Flu_{Li}}{Flu_{Li} + Flu_{Si}} \quad (1)$$

where  $Flu_{Li}$  and  $Flu_{Si}$  represent the fluency for the larger and smaller cities in pair  $i$ . The range of this variable is  $-1 < Flu_i < 1$ <sup>2</sup>. Given a pair in which participants were more fluent with the smaller city,  $Flu_i$  took a value less than 0. In contrast, for a pair in which participants were more fluent with the larger city,  $Flu_i$  took a value larger than 0. If participants were equally fluent with the larger and smaller cities,  $Flu_i$  approached 0.

**Familiarity heuristic.** This model is based on Honda et al. (2011). It assumes that when there is a difference in familiarity between two cities, people will infer that the more familiar city has the larger population. Response patterns in pair  $i$  in the binary choice task are assumed to be represented by the following variable,  $Fam_i$ :

$$Fam_i = \frac{Fam_{Li} - Fam_{Si}}{Fam_{Li} + Fam_{Si}} \quad (2)$$

where  $Fam_{Li}$  and  $Fam_{Si}$  represent the familiarity for the larger and smaller cities in pair  $i$ . The range of this variable is  $-1 \leq Fam_i \leq 1$ <sup>3</sup>. Given a pair in which participants were more familiar with the smaller city,  $Fam_i$  took a value less than 0. In contrast, for a pair in which participants were more familiar with the larger city,  $Fam_i$  took a value larger than 0. If participants were equally familiar with the larger and smaller cities,  $Fam_i$  approached 0.

## Models: Knowledge-based inference

Some researchers have pointed out that people use available knowledge in binary choice inferences. Although it might be difficult to clarify the specific processes of knowledge-based inferences, previous findings have been straightforward: When people have knowledge pertaining to population size, that knowledge directly affects the inference processes. For example, when one knows that a city has a professional soccer team, implying a larger population, that city

is more likely to be picked in a binary choice. Contrarily, when one knows that a city is rapidly aging, implying a declining population, it is unlikely to be chosen.

We assumed that people are clued to retrieve information when making inferences about the population of cities, as to whether the cities are large, or small. Based on this assumption, we modeled two knowledge-based inferences, a Z-score model and a Decision by Sampling model.

**Z-score model (ZM).** This model assumes that people have correct knowledge of absolute population sizes of cities, and that the availability of knowledge depends on their familiarity with cities. In each list, we calculated the z-scores for 15 cities (See Table 1). Using the z-scores, the response pattern in pair  $i$  is represented by the following variable,  $ZM_i$ :

$$ZM_i = \frac{Fam_{Li}z_{Li} - Fam_{Si}z_{Si}}{\alpha(z_{\max} - z_{\min})} \quad (3)$$

where  $z_{Li}$  and  $z_{Si}$  denote z-scores for the larger and smaller cities in pair the  $i$ , and  $z_{\max}$  and  $z_{\min}$  denote maximum and minimum z-scores in the list. In this equation,  $\alpha$  is a positive constant determining the range of this variable. The most important feature of  $ZM_i$  is that when people are familiar with a small city (i.e., the city with a negative z-score), that city is unlikely to be chosen. This feature differs from the familiarity heuristic, which assumes that familiarity leads to choice.

**Decision by Sampling (DbS).** As noted above, ZM assumes that people have correct knowledge about population sizes. However, this assumption may be inappropriate. Hilbig, Pohl, and Bröder (2009) suggested that participants have knowledge about the relative ranks of a criterion value within a given set. Thus, we have proposed another model of knowledge-based inference, Decision by Sampling (DbS), which was originally proposed by Stewart, Chater, and Brown (2006).

In this model, the subjective value of an object is determined by relative rank. In the present study, we assumed that knowledge about population sizes is determined by the relative rank in the list. The rank of each city in the list is calculated using following equation:

$$r = \frac{R-1}{14} - 0.5 \quad (4)$$

where  $R$  is the rank of the population in a list in an ascending order (the specific value of  $r$  for each city is shown in Table 1). The value,  $r$ , ranges from -0.5 to 0.5 depending on population size. Using  $r$ , the response pattern in pair  $i$  is represented by following variable,  $DbS_i$ :

$$DbS_i = \frac{fam_{Li}r_{Li} - fam_{Si}r_{Si}}{\beta} \quad (5)$$

where  $r_{Li}$  and  $r_{Si}$  denote  $r$  for the larger and smaller cities in pair  $i$ , respectively. Value  $\beta$  is a positive constant determining the range of this variable. The feature of this variable is analogous to  $ZM_i$ . Thus, when people are familiar with a small city, it is unlikely to be chosen in binary choice inferences.

<sup>2</sup> In the present study, fluency is operationally defined using elapsed time for city recognition. See the Results and Discussion.

<sup>3</sup> We set the minimum values of  $Fam_{Li}$  and  $Fam_{Si}$  with zero. See the Results and Discussion section.

Table 1. Two lists used in the experiment

List A	Population	Z-score	DbS (r)	List B	Population	Z-score	DbS (r)
Kawaguchi-shi	479,486	2.364	0.500	Yokohama-shi	3,544,104	2.608	0.500
Machida-shi	405,142	1.588	0.429	Osaka-shi	2,506,456	1.351	0.429
Kohriyama-shi	334,756	0.853	0.357	Nagoya-shi	2,145,208	0.913	0.357
Takasaki-shi	317,686	0.675	0.286	Sapporo-shi	1,869,180	0.579	0.286
Tsu-shi	283,167	0.315	0.214	Kobe-shi	1,498,805	0.130	0.214
Sasebo-shi	260,348	0.077	0.143	Kyoto-shi	1,392,746	0.002	0.143
Hachinohe-shi	248,776	-0.044	0.071	Fukuoka-shi	1,352,221	-0.047	0.071
Matsumoto-shi	223,472	-0.308	0.000	Hiroshima-shi	1,141,304	-0.303	0.000
Hitachi-shi	201,607	-0.536	-0.071	Sendai-shi	998,402	-0.476	-0.071
Yamaguchi-shi	187,539	-0.683	-0.143	Chiba-shi	905,199	-0.589	-0.143
Takaoka-shi	182,408	-0.736	-0.214	Niigata-shi	804,873	-0.710	-0.214
Imabari-shi	176,966	-0.793	-0.286	Hamamatsu-shi	786,776	-0.732	-0.286
Miyakonojo-shi	174,473	-0.819	-0.357	Kumamoto-shi	662,599	-0.883	-0.357
Ogaki-shi	159,661	-0.974	-0.429	Okayama-shi	659,561	-0.886	-0.429
Ashikaga-shi	159,040	-0.980	-0.500	Kagoshima-shi	601,675	-0.957	-0.500

## Experiment

### Method

**Participants.** Japanese undergraduates from Wako University, Keio University, and Toho University ( $n = 79$ ; 26 men and 53 women) participated in this experiment. They were given course credits for participation.

**Tasks and Materials.** We conducted three tasks, a binary choice task of population inference, measurement of familiarity, and a recognition task.

In the binary choice task, participants were presented with two Japanese city names and were asked to choose the city that they thought had the larger population.

In this task, we used Lists A and B (Table 1) that were used in Honda et al. (2011). Honda et al. (2011) suggested that these two lists differ in availability of knowledge pertaining to populations, and participants made inferences using different strategies. Inference patterns for List A could be well explained by the familiarity heuristic. In contrast, patterns of population inference using List B could be well explained by differences in the actual populations, indicating that participants had a good knowledge about the population sizes of cities on List B.

For the recognition test, participants were asked whether they knew the 30 cities that were presented in the binary choice task. For the measurement of familiarity, participants were asked how well they knew the 30 cities presented in the binary choice task.

**Procedure.** The three tasks were individually conducted using a computer, and they were presented in the following order: binary choice task, recognition task, and measurement of familiarity.

In the binary choice task, when participants pressed the key named, “Next,” the focal point “\*” was presented for 2000 ms on a computer screen. Then, two city names were

presented on the scree. Participants responded to the question by pressing one of two keys assigned to make a choice. Participants were instructed to respond as quickly and correctly as possible. Half of the participants first received 105  $\left(\frac{15 \times 14}{2}\right)$  pairs from List A, then 105 pairs from List B.

The other participants were presented the lists in the opposite order. The presentation order of the listed cities was randomized.

In the recognition task, by pressing the key named, “Next,” the focus point “\*” was presented for 2000ms, then a single city name was presented. Participants responded to the question by pressing one of two keys named, “Recognized,” or “Unrecognized.” The time that elapsed between the presentation of city name and the participants’ key-press was recorded. Participants were instructed to respond as quickly and correctly as possible.

For the measurement of familiarity, pressing the key named, “Next,” presented the participants with a city name. They were asked to rate their familiarity with the city using a 100-point scale shown on the screen. The scale ranged between (*not know at all*) on the far left to (*know a lot*) on the far right.

The above three tasks took 45-60 minutes to complete.

## Results and Discussion

In the following analysis, we operationally define familiarity with cities based on participants’ responses during the measurement of the familiarity task. Unrecognized cities were assigned a zero. Retrieval fluency for a city was operationally defined, based on the elapsed time recorded in the recognition task, as in Hertwig et al. (2008). It was assumed that the more quick the recognition response was, the more fluent was a participant’s retrieval of an object.

The constant values,  $\alpha$  and  $\beta$ , were set at 100. The range of  $ZM_i$  and  $DbS_i$  are  $-1 < ZM_i \leq 1$  and  $-0.5 \leq DbS_i \leq 1$ .

Table 2. Correlation coefficients among four models.

	List A			List B		
	$Fam_i$	$ZM_i$	$DbS_i$	$Fam_i$	$ZM_i$	$DbS_i$
$Flu_i$	0.628	0.070	0.059	0.622	0.229	0.384
$Fam_i$	-	-0.113	-0.017	-	0.451	0.597
$ZM_i$	-	-	0.857	-	-	0.826

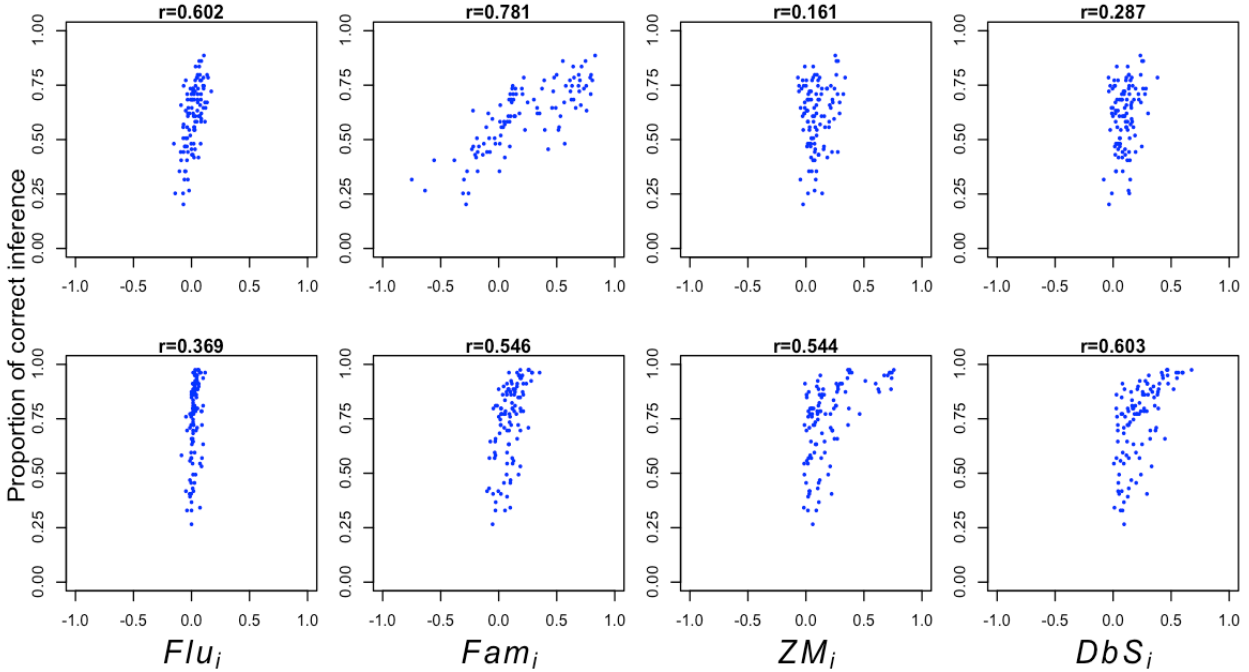


Figure 1. Relationship between choice pattern and variables of the four models. The upper four figures demonstrate results for List A, and the lower for List B.

### Analysis of aggregated data

**Similarity among models.** Although the four models were based on different assumptions about the processes of inference, predictions of binary choice might be analogous. Thus, we analyzed the similarity among the four models.

For 105 pairs on Lists A and B respectively, we calculated  $Flu_i$ ,  $Fam_i$ ,  $ZM_i$ , and  $DbS_i$  using mean familiarity or fluency for each city. Then we analyzed similarities among the four models in terms of correlation.

Table 2 shows the correlation coefficients among the variables. It was found that two heuristic models were analogous to each other. So were two knowledge-based inference models. Similarity between the heuristic and knowledge-based inference models varied depending on the list. With List A, heuristic models did not show a similarity with knowledge-based inferences. However, heuristic models showed a moderate similarity with the knowledge-based inference models in List B.

**Predictions of choice patterns.** Next, we examined the relationship between variables of the four models and choice patterns. For choice patterns, we used proportion of correct choice (i.e., rate of the larger city choice in a pair). For 105

pairs from Lists A and B respectively, we calculated the proportion of correct choice, and then examined correlations between the proportions and  $Flu_i$ ,  $Fam_i$ ,  $ZM_i$ , and  $DbS_i$ .

Figure 1 shows this relationship. As to the results of List A (the upper four figures), the two heuristic models predicted the choice patterns better than the knowledge-based inference models. In particular, a strong relationship between  $Fam_i$  and choice patterns was observed. Hence, these results suggest that the familiarity heuristic explains inference processes in List A.

The results for List B showed a different picture. Of the knowledge-based inference models,  $DbS_i$  showed a stronger relationship with the choice patterns than the others. Although  $Fam_i$  also showed a relationship with choice patterns comparable with  $DbS_i$ ,  $DbS_i$  will be a more appropriate model than the familiarity heuristic to explain choice patterns in List B. List B consists of well-known cities, and therefore, it is quite likely that participants could retrieve many clues pertaining to population size (Honda, et al., 2011). As a result, participants could have knowledge about population size such as relative rank, which is assumed in  $DbS_i$ .

Table 3. Median of correlation coefficients among four models.

	List A			List B		
	$Fam_i$	$ZM_i$	$DbS_i$	$Fam_i$	$ZM_i$	$DbS_i$
$Flu_i$	0.914	-0.027	0.062	0.142	0.071	0.098
$Fam_i$	-	-0.024	0.116	-	0.102	0.132
$ZM_i$	-	-	0.908	-	-	0.823

Table 3. Results of multilevel-logistic regression analysis (log likelihood).

List A			
Fluency	Familiarity	ZM	DbS
-4093	-4022	-4687	-4658
List B			
Fluency	Familiarity	ZM	DbS
-4638	-4498	-4504	-4484

Taken together, it was found that choice patterns were explained by different models depending on which list was used. In List A, heuristic models explained choice patterns better than knowledge-based inference models. In contrast, in List B, where participants were predicted to easily access many clues pertaining to the population size, choice patterns were explained by knowledge-based inference models better than heuristic models.

#### Analysis of individual data<sup>4</sup>

**Similarity among models.** As in the aggregated analysis, we first analyzed the similarity of correlations among the four models. We calculated correlation coefficients for each participant, and examined the distribution of the coefficients.

Table 3 provides medians of correlation coefficients for the six pairs. In List A, results were analogous to those in the aggregated data analysis. However, the observed similarities were much stronger than with the aggregated data analysis. For List B, similarity between two knowledge-based inference models was observed as in the aggregated data analysis. However, similarities were not observed between the fluency heuristic and familiarity heuristic. Moreover, similarity was not observed between the two heuristics and two knowledge-based inferences.

Thus, we found that there was no similarity between the two models at the individual level, even when it was observed with the aggregated data.

**Predictions of choice patterns.** Next, we examined choice patterns using individual data. Specifically, we adopted a model-based approach using a multilevel-logistic regression analysis (Gelman & Hill, 2007). We compared the four models that predicted the choice patterns in terms of model fitting.

The four models, fluency heuristic, familiarity heuristic, ZM, and DbS are represented as follows:

$$\log \frac{P_{CLi}}{1 - P_{CLi}} = aFlu_i + b \quad (6)$$

$$\log \frac{P_{CLi}}{1 - P_{CLi}} = aFam_i + b \quad (7)$$

$$\log \frac{P_{CLi}}{1 - P_{CLi}} = aZM_i + b \quad (8)$$

$$\log \frac{P_{CLi}}{1 - P_{CLi}} = aDbS_i + b \quad (9)$$

where  $P_{CLi}$  denotes the choice rate for the larger city in pair  $i$ . Values  $a$  and  $b$  denote free parameters for weight and intercept, respectively. For Lists A and B, these four models were regressed, based on individual data from 79 participants. Then we assessed the goodness of fit of the models using log-likelihood values. Table 3 shows the result of this analysis. For List A, the familiarity heuristic resulted in the best fit among the four models. For List B, DbS was a better fit than the other models.

Hence, findings based on individual data were analogous to those based on aggregated data. It was found that participants changed their inference strategies, depending on the list. Choice patterns with List A were well explained by the familiarity heuristic, and those with List B were well explained by DbS.

### General Discussion

In the present study, we examined the processes of inference regarding binary choices by model comparisons. We proposed statistical models for simple heuristic and knowledge-based inferences. It was found that the familiarity heuristic well explained the inference patterns for List A, and that DbS well explained those for List B. These findings suggest that people use different strategies depending on the situation.

The most important difference between Lists A and B is how easily people can retrieve clues relevant to population size. Given that List B consisted of well-known cities (Hon-

<sup>4</sup> In the following analyses, we excluded U-U pairs.

da et al., 2011), participants could easily retrieve clues pertaining to population size. For larger cities, participants could retrieve many clues suggesting that the population size would be large. Similarly, participants could retrieve many clues suggesting smaller population sizes for small cities. As a result, inference patterns could be well explained by knowledge-based inference models. Note that DbS explained the inference patterns better than ZM. This finding suggests that participants could have a good sense of population size in relative level. This is consistent with the findings of Hilbig, Pohl, and Bröder (2009).

On the other hand, participants might not have retrieved clues relevant to population size, and then used a simple inference strategy, such as the familiarity heuristic. In other words, a simple heuristic is the likely strategy when people have difficulty in retrieving clues relevant to inferences. This finding is consistent with the framework for understanding heuristics proposed by Shah and Oppenheimer (2008).

### Acknowledgments

This work was in part supported by Foundation of Fusion of Science and Technology (FOST).

### References

- Bröder, A., & Eichler, A. (2006). The use of recognition information and additional cues in inferences from memory. *Acta Psychologica, 121*, 275-284.
- Brighton, H., & Gigerenzer, G. (2011). Towards Competitive Instead of Biased Testing of Heuristics: A Reply to Hilbig and Richter (2011). *Topics in Cognitive Science, 3*, 197-205.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.
- Glöckner, A., & Bröder, A. (2011). Processing of recognition information and additional cues: A model-based analysis of choice, confidence, and response time. *Judgement and Decision Making, 6*, 23-42.
- Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*, 75-90.
- Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1191-1206.
- Hilbig, B. E. (2010). Precise models deserve precise measures: A methodological dissection. *Judgement and Decision Making, 5*, 272-284.
- Hilbig, B. E., Pohl, R. F., & Bröder, A. (2009). Criterion knowledge: A moderator of using the recognition heuristic? *Journal of Behavioral Decision Making, 22*, 510-522.
- Honda, H., Abe, K., Matsuka, T., & Yamagishi, K. (2011). The role of familiarity in binary choice inferences. *Memory and Cognition, 39*, 851-863.
- Newell, B. R., & Fernandez, D. (2006). On the Binary Quality of Recognition and the Inconsequentiality of Further Knowledge: Two Critical Tests of the Recognition Heuristic. *Journal of Behavioral Decision Making, 19*, 333-346.
- Newell, B. R., & Shanks, D. R. (2004). On the Role of Recognition in Decision Making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 923-935.
- Oeusoonthornwattana, O., & Shanks, D. R. (2010). I like what I know: Is recognition a non-compensatory determiner of consumer choice? *Judgement and Decision Making, 5*, 310-325.
- Oppenheimer, D. M. (2003). Not so fast! (and not so frugal!): Rethinking the recognition heuristic. *Cognition, 90*, B1-B9.
- Pachur, T., Bröder, A., & Marewski, J. N. (2008). The recognition heuristic in memory-based inference: Is recognition a non-compensatory cue? *Journal of Behavioral Decision Making, 21*, 183-210.
- Pachur, T., & Hertwig, R. (2006). On the psychology of the recognition heuristic: Retrieval primacy as a key determinant of its use. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 983-1002.
- Reimer, T., & Katsikopoulos, K. V. (2004). The use of recognition in group decision-making. *Cognitive Science, 28*, 1009-1029.
- Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin, 134*, 207-222.
- Snook, B., & Cullen, R. M. (2006). Recognizing National Hockey League greatness with an ignorance-based heuristic. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 60*, 33-43.
- Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53*, 1-26.
- Volz, K. G., Schooler, L. J., Schubotz, R. I., Raab, M., Gigerenzer, G., & von Cramon, D. Y. (2006). Why you think Milan is larger than Modena: Neural correlates of the recognition heuristic. *Journal of Cognitive Neuroscience, 18*, 1924-1936.

# Whose turn is it anyway?

## Same- and cross-person compound contributions in dialogue

Christine Howes, Patrick G. T. Healey, Matthew Purver

{chrizba, ph, mpurver}@eecs.qmul.ac.uk

Queen Mary University of London

Interaction, Media and Communication Research Group

School of Electronic Engineering and Computer Science

London E1 4NS, UK

### Abstract

In natural conversation people sometimes build larger grammatical, semantic and pragmatic units out of multiple turns or installments. The incremental and collaborative character of these ‘compound contributions’ presents challenges for theories of natural language processing. Compounds produced over successive turns by one person have often been analysed in essentially the same way as compounds produced by multiple people. In some recent accounts this putative equivalence has been taken as evidence for the claim that within- and cross-person language processing are fundamentally interchangeable. However, in this paper we present an analysis of compound contributions in a corpus of ordinary dialogues which shows that same- and cross-person compound contributions are constructed in different ways and have different semantic and pragmatic effects on the organisation of dialogue. In particular, we show that they differ in the pragmatic environments in which they occur and that they have different consequences for subsequent turn-taking and interpretation. This asymmetry highlights the need for models of dialogue that account for not just the inherent incrementality of dialogue, but the different status of each contributor towards a turn-in-progress.

**Keywords:** Dialogue; compound contributions.

### Introduction

*Compound contributions (CCs)* – dialogue contributions that continue or complete an earlier contribution, see e.g. (1) – are the paradigm case of coordination in dialogue and constitute a critical test case for theories of *natural* language processing.

- (1) **Daughter:** Oh here dad, a good way to get those corners out  
**Dad:** is to stick yer finger inside.  
**Daughter:** well, that’s one way.  
[from Lerner (1991)]

CCs are of interest to dialogue theorists because they provide evidence about how contributions can cohere with each other at multiple levels – syntactic, semantic and pragmatic (though of course they are not the only way). They also indicate the radical context-dependency of conversational contributions, which can, in general, be highly elliptical without disrupting the flow of the dialogue. CCs are a dramatic illustration of this: speakers must rely on the dynamics of the unfolding context (linguistic and extra-linguistic) in order to guarantee successful processing and production.

Much of the work on CCs has studied cross-person cases, in different disciplines and under a variety of different names, including *collaborative completions* (Clark, 1996; Poesio and Rieser, 2010), *co-constructions* (Sacks, 1992), *joint productions* (Helasvuo, 2004), and *split utterances* (Purver et al., 2006).

Linguistic studies show that grammatical constraints are respected across speaker and hearer (see e.g. Gregoromichelaki et al., 2009). In Finnish (which has rich inflectional morphology), and Japanese (a verb-final language), cross-person CCs within a single clause conform to the strict syntactic constraints of the language, despite the change in speaker (Helasvuo, 2004; Hayashi, 1999).

From a psycholinguistic point of view, the phenomenon of CCs seems compatible with mechanistic approaches as exemplified by the Interactive Alignment model of Pickering and Garrod (2004), which claims that, all things being equal, it should be as easy to complete another’s sentence as one’s own. According to this model, speaker and listener ought to be interchangeable at any point. A similar stance is taken by the grammatical framework of Dynamic Syntax (DS: Cann et al., 2005). In DS, parsing and production are taken to employ the same mechanisms, leading to a prediction that CCs ought to be strikingly natural (Purver et al., 2006).

From an organisational point of view, it has been claimed that turn-taking operates not on individual conversational participants, but on ‘parties’ (Schegloff, 1995). For example, a couple talking to a third person may organise their turns as if they are one ‘party’, rather than two separate individuals. Lerner (1991) speculates, following Sacks (1992), that cross-person compound contributions can clarify the formation of such parties, as they reveal a relationship between syntactic mechanisms and social organisation. He claims that this provides evidence of one way in which syntax can be used to organise participants into “groups”.

Because a sentence is obviously a prototypical instance of that thing which is done by a unit. Normally, some single person. That then permits it – for those who have the wit to do it – to be a way that some non-apparent unit may be demonstrated to exist.



We get, then, a kind of extraordinary tie between syntactic possibilities and phenomena like social organization. That is, an extremely strong way that these kids go about demonstrating that, for one, there is a group here, is their getting together to put this sentence together, collaboratively. (Sacks, 1992, p145)

These different approaches all treat cross-person compound contributions as being in some sense equivalent to turns produced by a single participant, in syntactic, semantic or pragmatic terms. However, there are few studies of same-person CCs, and those that there are (e.g. Goodwin, 1979; Walker, 2004) focus on the subset of *expansions*, which add material to an already potentially complete contribution (2), excluding *completions*, which involve the addition of syntactic material which is required to make the whole compound contribution (syntactically) complete (3).

- (2) **T:** It'll be an E sharp.  
**G:** Which will of course just be played as an F.  
[BNC G3V 262-263]
- (3) **D:** Well I do know last week thet=uh Al was  
 certainly very (pause 0.5s)  
**R:** pissed off [Lerner (1996), p260]

Like cross-person expansions, same-person ones are viewed as a highly productive way of utilising grammatical constraints for interactional purposes. Walker (2004) notes "it would seem that increments can be added to almost any possibly complete turn at talk, placing the practice alongside other generic conversational practices such as self- and other-initiated repair" (p167).

This type of treatment again suggests that there should be no differences to supplying a continuation to a prior turn, regardless of who produced the initial contribution. However, none of these studies have directly compared same- and cross-person CCs. We here present a corpus study to bridge this gap.

## Hypotheses

We examine two basic questions. First, whether the internal construction of CCs i.e., the syntactic and pragmatic ways in which the component parts are tied together, is the same in the same- and cross-person cases. Second, the external organisation of CCs; whether CCs as a whole are integrated into conversational organisation in the same way as as a conventional turn.

Following the existing literature we analyse the internal structure of the CC in terms of; a) the syntactic relationship between the components i.e. whether they are expansions or completions (as described above), b) repair i.e. whether the continuation of the CC is used to perform an edit or amendment on the antecedent and c) separation i.e. how closely together the antecedent

and continuation parts of a CC are normally placed. For the external organisation of CCs we examine the pragmatic organisation of the sequences in which CCs occur. Specifically, the patterns of turn-taking (who will speak next) and ratification (who acknowledges or responds to the CC). We also consider the placement of CCs with respect to backchannels. Backchannels are short acknowledgements like 'aha' or 'mmm', often produced in overlap with a speakers turn, which provide feedback to a speaker but don't typically lead to a change of speaker.

If there is no fundamental syntactic or cognitive difference between same- and cross-person CCs (e.g. Pickering and Garrod, 2004; Cann et al., 2005) then, all things being equal, we would expect that they should have the same distribution of expansions/completions, repairs and antecedent-continuation separations.

**Hypothesis 1** *Same-person and cross-person CCs have the same patterns of internal construction.*

In addition, if a CC functions as a single turn that just happens to have been produced in two (or more) parts (potentially by more than one person), rather than being a distinctive form of contribution, then they should be integrated into the conversation as a whole in the same way as non-compound turns. This would predict that the patterns of backchannels, turn-taking and ratification should, all things being equal, be the same for CC and non-CC turns. Moreover, this external organisation should be the same for both same-person and cross-person CCs. In a typical conversational sequence once a speaker has finished someone else will usually – although not always – speak next. Similarly, people do not normally acknowledge or ratify their own turns.

**Hypothesis 2** *The people who produce a CC should be less likely than other participants to speak next. They should also be less likely than other participants to ratify or acknowledge the CC.*

## Method

To investigate similarities and differences between same- and cross-person CCs, a corpus analysis of CCs in the spoken portion of the British National Corpus (BNC: Burnard, 2000) was carried out. This part of the corpus contains a large number of genuine spoken dialogues across a wide range of people and situations, allowing us to examine the prevalence of CCs in a variety of dialogues not restricted to the task-based dialogues which previous corpus studies tend to have analysed.

## Materials and procedure

For this exercise, the portion of the BNC annotated by Fernández and Ginzburg (2002), chosen to maintain a balance between what the BNC defines as *context-governed* (drawn from a particular domain e.g. business meetings, school classes, radio interviews) and *demographic* (recorded by volunteers during their daily lives)

dialogue, was used. This portion comprises 11,469 *s-units* – roughly equivalent to sentences<sup>1</sup> – taken from 200-turn sections of 53 separate dialogues.

**Annotation scheme** The dialogues were annotated according to the protocol outlined in Purver et al. (2009), and summarised in table 1, below.

Tag	Value	Explanation
end-complete	y/n	For all s-units: does this s-unit end in such a way as to yield a complete proposition or speech act?
continues	s-unit ID	For all s-units: does this s-unit continue the proposition or speech act of a previous s-unit? If so, which?
repairs	number of words	For continuations: does the start of this continuation explicitly repair words from the end of the antecedent? If so, how many?
start-complete	y/n	For continuations: does this continuation start in such a way as to be able to stand alone as a complete proposition or speech act?

Table 1: Annotation tags

## Results

person:	Same-		across-		Cross-	
	all				(all)	
	N	%	N	%	N	%
overlapping	0	0	0	0	18	5
adjacent	840	44	0	0	262	80
sep. by overlap	320	17	0	0	10	3
sep. by backchnl	460	24	456	63	17	5
sep. by 1 s-unit	239	13	229	32	16	5
sep. by 2 s-units	31	2	31	4	4	1
sep. by 3 s-units	5	0	3	0	1	0
sep. by 4 s-units	4	0	4	1	0	0
sep. by 5 s-units	1	0	1	0	0	0
sep. by 6 s-units	2	0	2	0	1	0
Total	1902		726		329	

Table 2: BNC antecedent/continuation separation

As discussed in Howes et al. (2011), the transcription conventions used when compiling the corpus can affect the raw results; in particular, the BNC convention of dividing contributions into “sentence-like units”, and in transcribing overlapping interruptions by interlocutors in linear time order, may result in an over-estimate of the number of same-speaker within-turn CCs. However, even excluding such within-turn and overlapping cases, and looking only at across-turn cases, there are over

<sup>1</sup>*S-units* are defined as “sentence-like divisions of a text”, and *utterances* are defined as “stretches of speech usually preceded and followed by silence or by a change of speaker”. Utterances may consist of many s-units; s-units may not extend across utterance boundaries. While s-units are therefore often equivalent to complete syntactic sentences, or complete functional units such as bare fragments or one-word utterances, they need not be: they may be divided by interrupting or overlapping material from another speaker.

twice as many same-person CCs (726) as cross-person CCs (329) – see table 2.

Many CCs have at least one s-unit intervening between the antecedent and continuation. In same-person cases, once we have excluded the within-turn CCs, this must always be the case; the intervening material is usually a backchannel (63%) or single other s-unit (32%, often e.g. a clarification question), but two intervening s-units are possible (4%) with up to six being seen. In cross-person cases, 88% are adjacent or separated only by overlapping material, but again up to six intervening s-units were seen, with a single s-unit most common.

## Completeness

As can be seen from table 3, the end- and start-completeness figures for same- and cross-person CCs are strikingly similar. The majority of both same- and cross-person continuations (71% to 72%) continue an already complete antecedent, with only 28-29% therefore being *completions*.

person:		Same-				Cross-	
		all	across-			(all)	
		N	%	N	%	N	%
Antecedent	Y	1367	72	513	71	236	72
end-complete	N	535	28	213	29	93	28
Continuation	Y	224	12	99	14	45	14
start-complete	N	1678	88	627	86	284	86
Repair	Y	77	4	34	5	32	10
	N	1825	96	692	95	297	90
Total		1902		726		329	

Table 3: BNC completeness and repair

These figures are even more striking when we consider the placement of arbitrary split points. In the experiment reported in Howes et al. (2011), artificial CCs were created in text chat dialogues, resulting in only 37% of fake CCs having a split point at a point where the antecedent could be considered end-complete (i.e. expansions) with 63% therefore appearing to be continuations.

This can be taken as evidence that participants in dialogue tend to wait for syntactic cues that suggest a possible opportunity for speaker change (referred to by conversation analysts as ‘transition relevance places’) before taking the floor – even where they construct their contribution as a continuation to a prior utterance.

## Repair

Although we are using only limited notion of **repair**, which only takes into account the amount of repetition or reformulation of words from the end of the antecedent at the start of the continuation, we believe that repair, as formulated, can be taken as an index of the difficulty of integrating the continuation to the syntactic material offered in the antecedent. This being the case, then under a model in which speakers and hearers are interchangeable (such as the interactive alignment model) the

proportion of repairs should be the same in same-person CCs as cross-person CCs, as there should be no increased difficulty in integrating one's continuation to another's antecedent as there would be to one's own. There should also not be any effects of where the split point occurs on the prevalence of repair.

However, cross-person continuations are significantly more likely to repair their antecedents than same-person cases (32/329, 10% vs. 34/726, 5%;  $\chi^2_{(1)} = 9.82, p = 0.002$ ), showing that there are differences between distributions of same- and cross-person CCs. In other words, although the distributions regarding completeness were equivalent and supported Hypothesis 1, it isn't this simple, and there appear to be additional constraints associated with continuing another's prior contribution that do not necessarily apply when continuing one's own.

## Backchannels

Similarly, there are different distributions of CCs in which the continuation follows a backchannel between same- and cross-person CCs. Same-person cases are more often produced as a response to a backchannel (63% of across-turn cases follow a backchannel, whilst even discounting adjacent cases only 40% of cross-person CCs do) suggesting that shaping one's next turn as a response to feedback is a common strategy in dialogue. Note also that 13% of all s-units in the corpus sample were backchannels<sup>2</sup> so there are actually a greater proportion of same-person cases following a backchannel than would be expected by chance, suggesting that backchannels may be used as a cue for participants to perform a *continue* grounding act (Traum, 1994).

## CA categories

In terms of specific types of CCs, the most common of the CA categories are Lerner's (1996) hesitation-related *opportunistic* cases, which make up 3-5% of same- and 10% of cross-person CCs, meaning cross-person opportunistic cases are more common than same-person ones (same, across-turn 36/726, 5% vs. other 32/329, 10%;  $\chi^2_{(1)} = 8.53, p = 0.003$ ). Interestingly, the breakdown of cases into those where the antecedent ends with an unfilled versus a filled pause also shows a difference between same- and cross-person cases: an other person is more likely to offer a continuation after an unfilled pause, than after a filled pause (antecedents ending in '*er(m)*' 35 continued by same, 13 by other; ending in '<pause>' 19 continued by same, 19 by other;  $\chi^2_{(1)} = 4.77, p = 0.03$ ). This finding backs up claims by Clark and Fox Tree (2002) that filled pauses can be used to indicate that the current speaker's turn is not yet finished and thus have the effect of holding the floor.

<sup>2</sup>This figure is based on the BNC part of speech tags, and as such may incorrectly include some answers to yes/no questions.

Lerner's compound TCU cases (*instead of*, *said/thought* etc, *if-then* and *when-then*) account for 2-3% of same-person and 1% of cross-person CCs, though note that these could be underestimates, as his non-syntactic cues (e.g. contrast stress and prefaced disagreements) could not be extracted. Rühlemann's (2007) *sentence relative* cases come next with 1-2%.

In contrast, the most common pattern (for same- and cross-person CCs) is the addition of an extending clause, either a conjunction introduced by '*and/but/or/nor*' (36-42%), or other clause types with '*so/whereas/nevertheless/because*', and the (other) category.

## Next speaker

To see if there are any effects on turn-taking or apparent party-membership, the 329 cross-person CCs were further annotated according to who spoke after the continuation and whether the conversation was dyadic or multiparty. Of the 53 dialogues, 34 were dyadic and 19 multiparty (though as observed in Eshghi (2009), many segments of multiparty dialogue are also dyadic in nature, we leave this to one side). This equates to 4919 turns in dyadic dialogues, in which there were 204 cross-person CCs (4.15%) and 2961 turns in multiparty dialogues with 125 cross-person CCs (4.22%). These proportions are not different (204/4919, 4% vs. 125/2961, 4%;  $\chi^2_{(1)} = 0.03, p = 0.87$ ), which is surprising – if cross-person CCs are used to indicate party-membership we might expect a greater proportion in the multiparty dialogues. This could be taken to suggest that parties are not common in these annotated dialogues.

There is no difference in the proportion of occasions in which the participant who contributed the continuation also provides the next contribution, thus holding the floor (50/204, 25% vs. 26/125, 21%;  $\chi^2_{(1)} = 0.600, p = 0.44$ , in line with the figure of 3/15, 20% reported in Szczeppek, 2000a). However, in all dialogues the proportion in which the supplier of the continuation retains the floor is lower than in general. For all annotated s-units in the dialogues there is no change of speaker in 41% of cases, compared to 23% of cases following a cross-person CC (4791/11469, 41% vs. 76/329, 23%;  $\chi^2_{(1)} = 44.424, p < 0.001$ ), suggesting that the continuation is treated as a separate turn and that interlocutors supplying a continuation do not assume they have a right to retain the floor.

## Ratification

Supporting the idea that ratifications ought to be more common in dyadic dialogues, if only appropriate when addressed to the original speaker, cross-person CCs are ratified or rejected by the initial speaker in (marginally) more cases in dyadic than multiparty cases (82/204, 40% vs. 37/125, 30%;  $\chi^2_{(1)} = 3.769, p = 0.052$ ). This does suggest that in dyadic dialogues cross-person CCs are

more often interpreted by the antecedent owner as addressed towards them, potentially as a form of repair which requires acknowledgement or ratification, and not interpreted as simply the mechanistic articulation of predictable material by another.

Cross-person CCs are also more likely to be ratified or rejected in completions than expansions (59/93, 63% vs. 79/236, 34%;  $\chi^2_{(1)} = 24.600, p < 0.001$ ). This is surprising if completions are merely the vocalisation of already predicted material (as in the interactive alignment model, for example), or if they are taken to be explicit acknowledgements (in a grounding model) as they should not then need either explicit evaluation or additional completion by the contributor of the antecedent.

In total, 138/329 (42%) of cross-person CCs are ratified; which is not rare, suggesting that cross-person continuations are often treated as not part of the same single unit as the antecedent. In a grounding model this suggests that these cases are those which are taken to be repairs, or new discourse units though note that we cannot distinguish between these possibilities. However, if they are treated as repairs then they are not treated as within-party repairs analogously to self-repairs, because these should also not require ratification.

## Discussion

Contrary to the predictions of Hypothesis 1, although there are similarities between same- and cross-person CCs, for example in the distributions of expansions versus completions, there are also significant differences. Cross-person continuations are more likely to start with explicit repair/reformulation of the antecedent. This suggests that people use CCs to do different things in the same and cross-person cases. Repairs are interesting in this context because they are not predictable continuations of the preceding contribution and therefore provide a counter-example to the idea that CCs are appropriately analysed as essentially the same turn but with a switch of producers. Perhaps more interestingly, they also suggest the operation of a more strategic process. It is generally observed that people avoid repairing other people's turns (known in Conversation Analysis as the preference for self repair; Schegloff et al., 1977). The observed pattern might be one that we would expect if cross-person CCs, in virtue of being constructed as a continuation of the speakers utterance, provide a device that enables a less exposed or 'face-threatening' form of other repair (as Lerner (1993) hypothesised). This works as a repair strategy to the extent that the completed CC is understood as 'belonging' to the producer of the antecedent of the CC. However, as we go on to discuss, our observations about patterns of turn-taking suggest that this is not straightforwardly the case.

Opportunistic CCs (after a '<pause>' or 'erm') are in general more likely to be cross-person cases; however

there are again pragmatic constraints – cross-person CCs are more likely where the antecedent ends in an unfilled rather than a filled pause. This suggests participants are aware of turn-taking or sequential expectations, and that speaker and hearer roles carry different responsibilities.

There also seem to be different places when same- and cross-person continuations are offered; the majority of cross-person continuations are adjacent to their antecedents, whilst even considering within-turn cases this is not so for same-person continuations. Same-person continuations are far more likely to follow a backchannel or single other s-unit than cross-person cases, suggesting that it may be the feedback from one's interlocutor(s) that leads to producing something syntactically tied to one's own prior contribution.

Contrary to Hypothesis 2, ratifications are offered following a cross-person CC; something which should be rare if the speaker of the antecedent were treating the continuation as if they themselves had just finished their turn. The assumption that cross-person CCs operate as a single turn that just happens to have been produced by more than one interlocutor, is perhaps an artefact of syntactic analyses that idealise away from the key organisational features of conversation. The evidence from CCs suggests that they might be better characterised as separate contributions that build parasitically on prior contributions, meaning that syntax is not an organising structure in the production of dialogue, but a resource that can be flexibly exploited by participants.

That ratifications were also more likely to be offered following a completion rather than an expansion, suggests that completions can also not be taken to be solely grounding devices, but must also be being treated by the antecedent owner as at least potentially repairing the incomplete antecedent (in which case an acknowledgement is appropriate).

This means that although one can unproblematically finish or continue another's utterance, this does not give it the same status as if they had completed or continued it themselves and has consequences for how we model CCs in particular and dialogue in general. There are several ways in which this might be approached. Either continuations by another are generally treated as repairs (and not exclusively as particularly strong forms of acknowledgement) or they are not taken to be acknowledgements at all. Given that continuations tend to be offered when common ground is presumed to be shared it could be the case that it is the presumption of shared common ground which requires acknowledging, or rejecting. Alternatively, it might be the fact that the incoming participant is aligning themselves with the initial speaker that requires acknowledgement, and not the content itself.

From a dialogue modelling perspective, we would want to be able to tell when a human agent's contribution con-

tinues some prior contribution – either their own or the system’s – in order to correctly analyse the semantics of the discourse, which is non-trivial given that antecedents do not have to be (and often are not) incomplete, or adjacent to the continuation.

The system should also be able to produce naturalistic continuations, and respond appropriately (e.g. by acknowledging a continuation from the user) including in terms of turn-taking. One example is the use of expansions – the system need not compute a complete sentence, but could use previously parsed input as a starting point. As dialogue models are very often in highly constrained contexts in which the system seeks information from the user, appropriate strategies involving CCs could be using incomplete antecedents to invite a user completion (for example, the travel agent system might ask “You want to go to...?”) and appendor questions (“...by bus?”) – see Hough (2011), for a preliminary outline of such a system.

### Acknowledgments

This was part of my PhD. This research was supported by the *Dynamics of Conversational Dialogue* project, funded by the UK ESRC (RES-062-23-0962).

### References

- Burnard, L. *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services <http://www.natcorp.ox.ac.uk/docs/userManual/>, 2000.
- Cann, R., Kempson, R., and Marten, L. *The Dynamics of Language*. Elsevier, Oxford, 2005.
- Clark, H. H. *Using Language*. Cambridge University Press, 1996.
- Clark, H. H. and Fox Tree, J. E. Using *uh* and *um* in spontaneous speaking. *Cognition*, 84(1):73–111, 2002.
- Eshghi, A. *Uncommon Ground: The Distribution of Dialogue Contexts*. Ph.D. thesis, Queen Mary University of London, 2009.
- Fernández, R. and Ginzburg, J. Non-sentential utterances: A corpus-based study. *Traitement Automatique des Langues*, 43(2), 2002.
- Goodwin, C. The interactive construction of a sentence in natural conversation. In Psathas, G., editor, *Everyday Language: Studies in Ethnomethodology*, pages 97–121. Irvington Publishers, New York, 1979.
- Gregoromichelaki, E., Sato, Y., Kempson, R., Gargett, A., and Howes, C. Dialogue modelling and the remit of core grammar. In *Proceedings of IWCS*. 2009.
- Hayashi, M. Where grammar and interaction meet: A study of co-participant completion in Japanese conversation. *Human Studies*, 22(2):475–499, 1999.
- Helasvuoto, M.-L. Shared syntax: The grammar of co-constructions. *Journal of Pragmatics*, 36(8):1315–1336, 2004.
- Hough, J. Incremental semantics driven natural language generation with self-repairing capability. In *Recent Advances in Natural Language Processing (RANLP)*, pages 79–84. Hissar, Bulgaria, 2011.
- Howes, C., Purver, M., Healey, P. G. T., Mills, G. J., and Gregoromichelaki, E. On incrementality in dialogue: Evidence from compound contributions. *Dialogue and Discourse*, 2(1):279–311, 2011.
- Lerner, G. Collectivities in action: Establishing the relevance of conjoined participation in conversation. *Text-Interdisciplinary Journal for the Study of Discourse*, 13(2):213–246, 1993.
- Lerner, G. H. On the syntax of sentences-in-progress. *Language in Society*, pages 441–458, 1991.
- Lerner, G. H. On the “semi-permeable” character of grammatical units in conversation: Conditional entry into the turn space of another speaker. In Ochs, E., Schegloff, E. A., and Thompson, S. A., editors, *Interaction and Grammar*, pages 238–276. Cambridge University Press, 1996.
- Pickering, M. and Garrod, S. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226, 2004.
- Poesio, M. and Rieser, H. Completions, coordination, and alignment in dialogue. *Dialogue and Discourse*, 1:1–89, 2010.
- Purver, M., Cann, R., and Kempson, R. Grammars as parsers: Meeting the dialogue challenge. *Research on Language and Computation*, 4(2-3):289–326, 2006.
- Purver, M., Howes, C., Gregoromichelaki, E., and Healey, P. G. T. Split utterances in dialogue: A corpus study. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue (SIGDIAL 2009 Conference)*, pages 262–271. Association for Computational Linguistics, London, UK, 2009.
- Rühlemann, C. *Conversation in Context: A Corpus-Driven Approach*. Continuum, 2007.
- Sacks, H. *Lectures on Conversation*. Blackwell, 1992.
- Schegloff, E. A. Parties and talking together: Two ways in which numbers are significant for talk-in-interaction. *Situated Order: Studies in the Social Organization of Talk and Embodied Activities*, pages 31–42, 1995.
- Schegloff, E. A., Jefferson, G., and Sacks, H. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382, 1977.
- Traum, D. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester, 1994.
- Walker, G. On some interactional and phonetic properties of increments to turns in talk-in-interaction. In Couper-Kuhlen, E. and Ford, C. E., editors, *Sound Patterns in Interaction: Cross-linguistic Studies from Conversation*. John Benjamins, 2004.

# Language acquisition in Down Syndrome from embodied perspective: How body constrains language acquisition?

P. Hristova (phristova@cogs.nbu.bg), H. Toushek (chrissidt@yahoo.com), G. Petkov (gpetkov@cogs.nbu.bg)

Department of Cognitive Science and Psychology, New Bulgarian University, 21 Montevideo Street  
Sofia 1618, Bulgaria

## Abstract

Parents of children with Down syndrome (DS) were asked to fill a questionnaire about how much their children understand and how well they use words. It was found that word acquisition is affected not only by word frequency, but also by whether a word is related to eventual misbalance of the body. The results are in favor of the hypothesis that the constraints of the human body may cause systematic variations of language acquisition as long as keeping the body balance is a typically difficult motor task for DS children. Additionally, we also found an asymmetry of the acquisition of the verbs and the nouns, depending on their frequency and relatedness to eventual misbalance of the body.

## Introduction

Usually learning of a native language takes several years. During their first years of life children learn thousands of linguistic symbols and constructions used by people around them. How does this extensive learning happen? Glenberg, A.M., Havas, D., Becker, R., & Rinck, M. (2005) conclusively argue backed up with the Searl's Chinese room and the Harnad's symbolic marry-go-around argument, that language must be grounded outside the linguistic system. Language gets meaning from what is already meaningful for children, i.e. perceptions, actions and emotions. Many cognitive psychologists, however, would agree that the initial states of language acquisition are grounded in perception and action, since at least, the first symbols can be considered as gaining meaning from these modalities. The problem mainly concerns the mature language processing where the role of bodily states for language use is still debatable (Glenberg et al, 2005). Recently, however, a study on body-specificity hypothesis has shown that right-handers activate the left-premotor cortex during lexical decision on manual-action verbs, while left-handers activate the right premotor cortex (Willems, Hagoort, Casasanto, 2010). Hence, it seems reasonable to assume that different bodies lead to different representation of environmental categories and therefore, to predict that body differences are connected to specific differences in language usage.

The goal of this study is to provide preliminary evidence that specific body constraints correspond to specific patterns of language acquisition. The possibility that the human mind and body were evolutionary shaped by language in a way that allows language production and comprehension is largely recognized nowadays (for a review, Pinker, S., 2000), but the opposite direction is still fairly underestimated. In other words, there has been no attempt, at least to our knowledge, to investigate systematic variations in *language acquisition* due to particular body constraints.

We start from the consistent evidence for specific difficulties in the Down Syndrome population (Lauteslager, 2004; Winders, 1997) for problems with maintaining balance and posture (i.e., insufficient stability, lack of trunk rotation and delayed reaction speed) that according to Lauteslager (2004) leads to a peculiar compensatory symmetrical manner of moving, mainly characterized by lack of variability. Every move that threatens the balance seems to be problematic for children with Down Syndrome (DS). Thus almost every stage of the typical motor development is a hardship for a DS child, who usually is able to maintain head control, sitting, standing, climbing, walking, jumping etc. considerably late in his/her development. But importantly, the way DS child acquires the motor skills important for independent locomotion is different from the typically development trajectory and is characterized by preference toward symmetrical movements with a lot of external support. The main reason considered to be responsible for the specific motor development of DS children is the low muscle tone (i.e. hypotonia) that affects individuals from this population in different degree but is present from birth and persist throughout life.

To summarize, the low muscle tone constraints motor development of DS children, causing problems with maintaining balance and posture, which in turn leads to development of specific symmetrical movements, movements with a lot of external support and static motor behaviour (Lauteslager, 2004). Thus we hypothesize that word for actions and objects that threaten body posture and balance will be represented differently from DS population compared to normally developing one.

Unfortunately, it is difficult to predict how these specific body constraints would change language acquisition. It is possible that children with DS will acquire the meaning of words that are problematic for their motor planing and behaviour much later than other words which do not tread body posture and balance. In other words, because it is difficult for them to walk, run, climb, swing etc. they will avoid such behaviors and the respective situations that require these actions. If this hypothetical scenario really takes place, then the direct sequence should be a substantial underrepresentation of such nouns and verbs that in some respect threaten their balance and body posture.

But we may expect just the opposite result as well, namely that children with DS will know better the words, associated with maintenance of posture and balance than the other words. If children with DS are raised in a stimulating environment, which encouraged them to walk, to bring the ball, to jump, to ride etc. they will finally master these actions. Moreover, during this process of pure motor

development child's attention necessarily should be focused on the accomplishment of movements that are difficult for them. Hence everything that happens at this moment of keen attention can be potentially encoded better, including the labels for the body movement, hopefully provided from people around them. Since, it was reported that they have substantial deficits in sustained attention (i.e. attention toward a given stimulus that allow its efficient processing) (Brown, Johnson, Paterson, Rick Gilmore, Longhi and Karmiloff-Smith, 2003), the attention toward a difficult action seems to provide a prolonged period of focused attention that may facilitate encoding at least of the label of the action and the associated with this action objects.

To sum up, we expect a substantial difference in both understanding and production of words, associated with body posture and balance in DS children. At this point, we don't have any strong theoretical reasons to narrow down our expectations. But, since we plan to study understanding and production of word in home-raised children with DS (i.e. hopefully, these children were stimulated to interact actively with the surrounding physical and social environment) we assume that the mere association of a given word with posture and balance maintaining will facilitate its knowledge.

### **Language acquisition in children with Down Syndrome**

Overall, children with DS acquire language at slower rates than typically developing children on the same chronological age. They produce shorter sentences and tend to omit function words (i.e., articles, prepositions, pronouns, etc.) (Chapman, 1995). Language comprehension in DS is usually superior to language production, but still significantly behind their level of cognitive development (Miller, 1992), mental age (Vicari, Caselli, Gagliardi, Tonucci, & Volterra, 2002) or vocabulary size (Singer et al, as cited in Tomasello, 2006). Overall the linguistic abilities of children with DS are surprisingly behind the ones expected on the bases of their cognitive abilities.

But actually the picture is much more complicated. A few studies point to the possibility that language acquisition in DS has its own profile, which turned to be different from the one of typically developing children. When asked to repeat words and sentences children with DS omit a significantly higher number of articles, verbs, and prepositions than typically developing children and children with specific language impairment, matched on mental age (Caselli, Trasciani, & Vicari, 2008). The authors explain these differences with the specific repetition task administered to children that certainly relies on different cognitive abilities, including verbal short-term memory, executive control etc., that are usually reported to be poorer in DS than in mental age-matched children with typical development. However, general cognitive impairments can hardly explain why exactly verbs, articles and prepositions are preferentially omitted by children with DS rather than nouns and modifiers.

On other hand, another line of research points to the fact that verb acquisition poses specific difficulties to DS children, since verb understanding strongly relies on syntactic development (Tomasello, 2006), which is reported to be dramatically delayed in this specific population (Fowler, 1990 as cited in Tomasello, M, 2006). Naigles, Fowler, & Helm (1995) point out that verbs incorporate both semantic and syntactic knowledge and thus mastering of verb meaning should depend on both of these components: "specific lexical and syntactic information concerning each individual verb must be accrued in order to establish stable verb representations".

Overall, based on the Caselli et al (2008) and Naigles et al. (1995) findings we may expect that children with DS will underrepresent verbs compared with nouns. With respect to our hypothesis, also seems much easier to imagine how verbs are learned through out specific actions than nouns, for example. Hence, the expected effect of embodiment on language acquisition, if any, should be predominantly expected in the domain of verb learning.

Then, our question is what will happen with word learning in general and verb learning in particular, if some actions are more difficult for an individual than others.

### **Investigation**

We asked parents of children with Down syndrome and parents of typically developing children, matched on age (control group) to judge on a 7-point scale how well their children understand and how well they use the respective words. We designed a list of 172 Bulgarian words, controlled for objective frequency, length, type (noun or verb), and balanced with respect to embodiment, i.e. the degree of association between a given word and maintenance of body posture and balance (see section Stimuli below).

Whereas, expectedly, the results for the control group reached a ceiling effect for all words, the data of DS group seems to follow an interesting trend.

### **Method**

#### **Design**

The design of the study was 2(diagnoses of trisomy 21: yes/not) x 2(objective frequency: high/low) x 2(length of the word: long/short) x 2 (type of the word: noun or verb) x 2(embodiment: high/low) factorial design. The dependent measures were the ratings, given from the parents of the children, to the understanding and to the production of the respective words.

An additional independent factor – concreteness/abstractness was measured for control considerations (see the next section - Stimuli).

#### **Stimuli**

We achieved the objective frequency of 355 Bulgarian words from a corpus of 70 stories, included in the training



program for public kindergartens in Bulgaria. The length of the words varied uniformly between 3 and 11 letters.

On the next step, we asked two native speakers to judge each of the word on a 7-point scale according to the instruction: “Please, rate on a 7-point scale how much each of the following concepts or actions disturbs body posture or balance”. For example, both experts gave high ratings to verbs like rush toward, jump, and scramble. Both experts gave high ratings to nouns like fight, stroke, and ball. On the other pole, low ratings were given for verbs like rill, mistake, love; and nouns like sky, gold, sign... The highest disagreements between the expert’s ratings were 3 (on a 7-point scale).

Finally, we chose 172 words from the whole list (100 verbs and 72 nouns) and formed 86 pairs of words with polar ratings according to their embodiment (relatedness to dis-balance of the body) and fixed objective frequency and length. More precisely, we ensured that for each verb (respectively noun) with a specific length, frequency, and low embodiment, there is another verb (respectively noun) in the list with same length and frequency but with high embodiment. The length of the words was distributed among 3 and 11 letters with a dominance of 4-8 letter words.

If the value of the words according to frequency and embodiment dimension are discretized onto “high frequent”, “low frequent”, and, respectively “high embodied” and “low embodied”, the overall distribution of the stimuli would be the one shown in Table 1.

Table 1. Distribution of the stimuli – verbs on the top panel; nouns on the bottom panel.

<i>VERBS</i>		frequency		
		low	high	total
embodiment	low	31	16	47
	high	35	18	53
	total	66	34	100
<i>NOUNS</i>		frequency		
		low	high	total
embodiment	low	21	19	40
	high	17	15	32
	total	38	34	72

For an additional control, we asked 14 adults (10 women and 4 men) on mean age of 22.5, to rate the words according to their concreteness/abstractness. This was an additional ad-hock variable, used to control for: First, whether our measurement of embodiment wasn’t actually a measurement of concreteness. Second, whether concreteness is the better predictor of understanding and usage of words then the other factors. Thus, for each of the 172 words we had also an average rating of its concreteness/abstractness on a 7-point scale (1- very concrete...7- very abstract).

## Procedure and participants

Eight parents of children with DS (seven mothers and one father) were asked to evaluate each of the words according to two criteria:

First, how much they think that their child understand the respective word.

Second, how appropriate their children use the respective word.

The ratings were given on a 7-point scale. The order of the words was random for every participant. The parents worked at home. At the beginning of the questionnaire, they should fill the age of the child as well as additional health problems, if any. All parents were duly confident that the data are confidential and will be used for statistical purposes only.

The same procedure was used for a control group of the parents of ten children without DS.

The age the children with DS were between 4 years and 4 years and 7 months. Half of the children (4) were girls and the other half (4) – boys. Two of the children had corrected to normal vision, 2 had heart problems (one of them was diagnosed also with West Syndrome). All children with DS were with full trisomy 21and not mosaic. All parents of DS children in our study were recruited from Bulgarian Down Syndrome Association and were with university education.

The group of the typically developing children was matched in age and sex to the DS children in our study. None of them did not have heart or vision problems. All parents again were with university degree of education.

## Results

First of all, we received a clear ceiling effect for the children from the control group. The mean rating for understanding was 6.662, st. dev. 0.623 for all words; the mean rating for usage was 0.659, st. dev. 0.708.

The respective mean ratings for the DS children, however, were:

For understanding: mean rating 3.887, st. dev. 1.883; for usage: mean rating 1.772, st. dev. 1.002

The rest of analyses were on the results from the DS children only and reflect the pattern of language understanding and production observed and reported from their parents. The data for all DS children were averaged by item (172 independent words) and we analyzed the impact of embodiment, frequency, length and type of the words two dependent variables – mean rating of understanding and mean rating of production.

For control (see below, at the end of the section Results), we aggregated the data by subject as well and repeated the analyses assuming within-subject design (each subject was an independent case).

**Results for understanding, depending on objective frequency, length of the words, type of the words**

(verbs/nouns) and embodiment (how much each concept or action disturbs body posture or balance)

The Univariate ANOVA analysis on *understanding* detected main effect of the frequency ( $F(1, 164) = 12.031, p = 0.001$ ) and of the embodiment ( $F(1, 164) = 8.730, p = 0.004$ ). There was not significant main effect of the type of the word (verb or noun): ( $F(1, 164) = 1.465, p = 0.228$ ). In other words, parents in our sample estimated words with high objective frequency as the words that their children understand better than the words with lower objective frequency. Interestingly, words with high embodiment were also rated as significantly more knowable for children with DS than words, associated with less posture disturbances and balance difficulties.

There were not significant interactions among pairs of the variables but it was a significant triple interaction among frequency, embodiment, and type of the word ( $F(1, 164) = 4.708, p = 0.031$ ). Figure 1 illustrates this interaction:

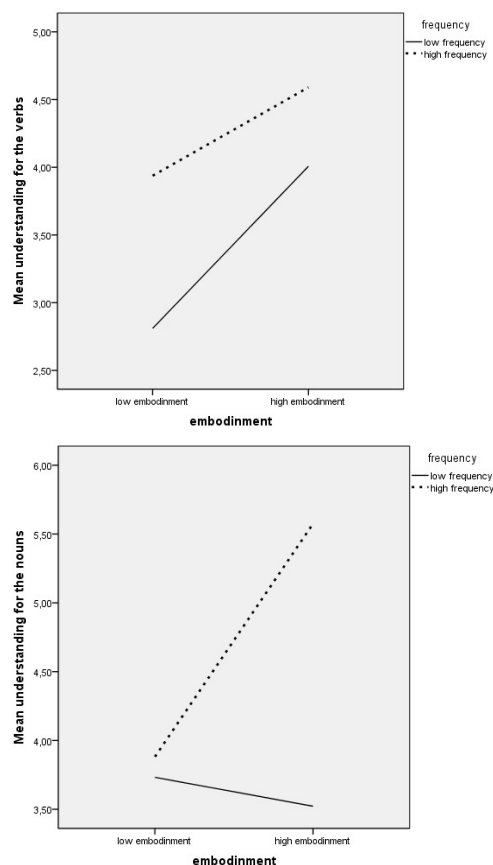


Figure1: The influence of the frequency and embodiment on the understanding of the words by the children with Down syndrome was different for the verbs (upper panel) and nouns (bottom panel)

Thus, the pattern of understanding of the words was found to be completely different. Whereas both high frequency and high embodiment support understanding (not much surprisingly), it happen that embodiment doesn't support understanding of low frequently nouns.

The length of the words influenced understanding of the verbs ( $F(1, 98) = 9.246, p = 0.003$ , obtained by a linear regression analysis): according to parents of DS children their children understand shorter verbs better than longer ones. However, the length doesn't influence significantly understanding of the nouns: ( $F(1, 70) = 0.125, p = 0.725$ ). One may argue that this is due to the fact that children learn first the nouns, thus the influence of the length became slower. However, we didn't found a significant difference between the overall level of understanding of the verbs and the nouns. It should be mentioned, however, that it was a significant correlation between length and frequency ( $r = -0.279, p < 0.001$ ).

The final control measurement – concreteness/abstractness of the words, correlated with embodiment ( $r = -0.244, p = 0.001$ ) and didn't correlated with the other independent variables.

Whereas it was a main effect of abstractness on understanding ( $F(1, 164) = 6.739, p = 0.010$ ), there were no any interactions (neither paired, neither triple) with the other factors (i.e., embodiment, frequency, length and type of the word). Thus, our interpretation is that embodiment is different measurement from concreteness and that the interesting interaction on Figure 1 is due to the relatedness of the words to a possible disturbance on the balance of the body, instead of concreteness/abstractness.

**Results for production, depending on objective frequency, length of the words, type of the words (verbs/nouns) and embodiment (how much each concept or action disturbs body posture or balance)**

Although the ratings for the other dependent measure – correct usage of the words – were much lower, the results followed similar pattern according to all analyses:

It was main effect of the frequency ( $F(1, 164) = 11.885, p = 0.001$ ) and of the embodiment ( $F(1, 164) = 4.582, p = 0.034$ ). In contrast with the results for the understanding, it was a significant main effect of type of the word too: ( $F(1, 164) = 11.216, p = 0.001$ ). The nouns received higher ratings.

There were not any pair interactions. The triple interaction between frequency, type of word, and embodiment was with a marginal significance only: ( $F(1, 164) = 3.046, p = 0.0083$ ).

Length of the word influenced usage of the verbs ( $F(1, 98) = 14.126, p = 0.000$ ) and doesn't influence the usage of the nouns ( $F(1, 70) = 0.055, p = 0.815$ ).

It was main effect of abstractness ( $F(1, 164) = 4.137, p = 0.044$ ) without any paired or triple interactions. For comparison with the marginal significance of the interaction between frequency, type, and embodiment, the results for the respective interaction between frequency, type, and abstractness was  $F(1, 164) = 0.194, p = 0.660$ .

**Repetition of analyses:** analyses on the data averaged by subject

The impact of embodiment, frequency, length and type of the word was measured with a Repeated Measures Analysis on data averaged by subject (i.e., eight DS children). The main effects of these factors were again estimated as significant. We obtained main effects of frequency ( $F(1, 7) = 66.714, p = 0.000$ ), of embodiment ( $F(1, 7) = 46.505, p = 0.000$ ), and also of type of the word ( $F(1, 7) = 9.239, p = 0.019$ ).

The triple interaction was also significant:  $F(1, 7) = 31.166, p = 0.001$ . The only difference was that Repeated Measures Analysis estimated as significant the interaction between frequency and embodiment:  $F(1, 7) = 12.071, p = 0.010$ .

## Discussion

According to their parents children with DS understand and use better words that involve difficult for them actions, namely the ones associated with greater posture disability and asymmetry. It seems that, while performing the difficult for them body movements, children learn better the words for these movements and the words for the objects that are typically connected to this movements. Having in mind that keeping the balance of their body is a typically difficult motor task for these children, the result is in favor of the hypothesis that the body constraints may cause systematic variations of language acquisition.

Interestingly, although both frequency and embodiment influence word understanding and production of DS children, the pattern for verbs and for nouns differs. The embodiment doesn't influence low frequent nouns. This was not due to a floor effect, as can be seen on Figure 1 and cannot be explained just by a complete misunderstanding of those words. This asymmetry raises new questions and requires further investigations.

The main effect of embodiment, however, points to an important trend in language acquisition in children with DS: the words that are related to eventual misbalance of the body were estimated as better understood and used from DS children.

The important question, however, is why this happens. We speculated at the beginning of this paper that children with DS may be recognize the situations that disturb their body posture or balance as more difficult and hence, requiring their attention. Then, their outperformance on words that are associated with such situations can be considered as a matter of attention, which is dedicated to the maintenance of body posture and balance. The extra attention dedicated to the movements that require maintaining of posture and balance may as a side effect improve the knowledge for the associated concepts. Attention, indeed, seems to be a problematic cognitive ability for DS population. Brown et al.(2003) conclusively argued that toddlers with DS has significant problems with maintaining attention to objects in the environment compared to a group of chronologically matched toddlers with Williams syndrome, a group of chronologically matched typically developing children and a

group of mentally matched typically developing children. Possibly, the extra attention and effort toward the difficult actions may overcome the reported sustained attention deficits of DS children. Instead of training the sustained attention of children with DS we may compensate it with educational techniques that back on their posture and balance difficulties.

It could be, however, that DS children recognize actions that disturb their body posture and balance as threatening ones, since at least at the beginning of their motor development these actions usually end up with incidents, associated with physical pain. Possibly, in order to avoid successfully this actions, children with DS learned better everything, associated with situations that treat their body posture and balance.

Of course, both explanations could be rephrased in a way that appreciates the role of the parents in raising their children. It is quite possible, indeed probable that parents, rather than children recognize, which are the threatening and the difficult situations for their DS toddlers and pay more attention in teaching them how to master such situations. Again, the prediction will be that children finally will learn better the words associated with such situations than with others.

Unfortunately, our study could not disentangle between those possible explanations, but we find the compensatory account the most interesting one. If extra intentional resources can be allocated to particular movements and as a side effect this extra attention can improve conceptual knowledge, we may design techniques for both native and foreign language learning appropriate for the DS peculiarities.

## Acknowledgments

This research was supported financially by the European Office for Aerospace Research and Development under grant FA8655-10-1-3061 (Adaptive Problem Solving by Analogy).

## Reference

- Brown, Johnson, Paterson, Rick Gilmore, Longhi and Annette Karmiloff-Smith (2003) Spatial representation and attention in toddlers with Williams syndrome and Down syndrome, *Neuropsychologia* 4, 1037–1046
- Caselli, Monaco and Manuela Trasciani, and Vicari, (2008) Language in Italian Children With Down Syndrome and With Specific Language Impairment, Vol. 22, No. 1, 27–35
- Chapman RS. Language development in children and adolescents with Down syndrome, In: Fletcher P, MacWinney B, editors. *The Handbook of Child Language*. Oxford: Blackwell, 1995.
- Glenberg, A.M., Havas, D., Becker, R., & Rinck, M. (2005). Grounding Language in Bodily States: The Case for Emotion. R. Zwaan and D. Pecher (Eds.) *The grounding*

- of cognition: The role of perception and action in memory, language, and thinking. Cambridge: Cambridge University Press.
- Lauteslager, P. (2004). Children with Down's syndrome; motor development and intervention. Amersfoort: 's Heeren Loo Zorggroep.
- Miller JF. Development of speech and language in children with Down syndrome. In: Lott IT, Mc Loy EE, editors. Down Syndrome, 1992: p. 39–50.
- Naigles, L. G. Fowler, A.E. , Helm, A. (1995). Syntactic bootstrapping from start to finish with special reference to Downsyndrome. In : Tomasello, M. & Merriman, W. E. (Eds.) Beyond names for things: Young children's acquisition of lverbs. pp. 299-330. Hillsdale, NJ, Lawrence Erlbaum Associates, Inc.
- Pinker, S. (2000) "Language Acquisition," in L. R. Gleitman, M. Liberman and D. N. Osherson eds., *An invitation to Cognitive Science* (<http://users.ecs.soton.ac.uk/~harnad/Papers/Py104/pinker.langacq.html>). MIT Press, Cambridge.
- Tomasello, M. (2006). Acquiring linguistic constructions. In D. Kuhn & R. Siegler (Eds.), *Handbook of Child Psychology*. New York: Wiley.
- Vicari, M.C. Caselli, C. Gagliardi, F. Tonucci, V. Volterra, (2002) Language acquisition in special populations: a comparison between Down and Williams syndromes, *Neuropsychologia*, vol 40 , p. 2461–2470
- Willems, R. M., Hagoort, P., & Casasanto, D. (2010). Body-specific representations of action verbs: Neural evidence from right- and left-handers. *Psychological Science*, 21, 67-74.
- Winders, P. (1997) Gross Motor Skills in Children with Down Syndrome: A Guide for Parents and Professionals, Woodbine House.

# The Upbeat of Language: Linguistic Context and Embodiment Predict Processing Valence Words

**Sterling Hutchinson (schthns@memphis.edu)**

Department of Psychology / Institute for Intelligent Systems, University of Memphis  
365 Innovation Drive, Memphis, TN 38152 USA

**Max M. Louwerse (mlouwerse@memphis.edu)**

Department of Psychology / Institute for Intelligent Systems, University of Memphis  
365 Innovation Drive, Memphis, TN 38152 USA

## Abstract

Previous studies have demonstrated that comprehension of conceptual metaphors elicits embodied representations. This finding is non-trivial, but begets the question whether alternative explanations are to be dismissed. The current paper shows how a statistical linguistic approach of word co-occurrences can also reliably predict metaphor comprehension. In two experiments participants saw word pairs with positive (e.g. *happy*) and negative (e.g. *sad*) connotations. The pairs were presented in either a vertical configuration (Experiment 1) or a horizontal configuration (Experiment 2). Results showed that response times could be explained by both the statistical linguistic approach and the embodied approach. However, embodied information was most salient in the vertical configuration and statistical linguistic information was most salient in the horizontal configuration. Individual differences modulated these findings, with female participants being most sensitive to the statistical linguistic approach, and male participants being most sensitive to the embodiment approach. These findings suggest that comprehension of conceptual metaphors can be explained by both linguistic and embodiment factors, but that their relative salience is modulated by cognitive task and individual differences.

**Keywords:** embodied cognition; symbolic cognition; linguistic context; valence words; gender differences; symbol interdependency

## Introduction

When we are happy, we are in high spirits; when we are sad, we are down in the dumps. Our mood is lifted when we are cheerful, but our enthusiasm drops when we are depressed. We have our high times and our low times. We reach for the sky, but sometimes our plans run into the ground. These are some examples suggesting that conceptual metaphors highlight associations between abstract concepts (e.g., happy, sad) and spatial properties (e.g., high, low). Lakoff and Johnson (1981; 1999) suggest that metaphors like these help automatically ground abstract concepts in bodily experiences. In this sense, metaphors inform the language user of the perceptual and biomechanical processes underlying the representation of those concepts. Such theories of embodied cognition have received an impetus in the last decade (Barsalou, 1999; Glenberg, 1997; Prinz, 2004; Zwaan, 2004) with considerable empirical support showing that cognitive processes are undoubtedly influenced by perceptual and spatial information (De Vega,

Glenberg, & Graesser, 2008; Pecher & Zwaan, 2005; Semin & Smith, 2008 for overviews). These findings allow for the conclusion that linguistic symbols are grounded in modality specific perceptual and motor systems.

If conceptual metaphors highlight associations between abstract concepts and physical or spatial properties, it is predicted that words with positive connotations are processed faster when presented higher in space, and words with negative connotations are processed faster when presented lower in space. This is indeed what a number of studies have shown, with participants being apt to process and remember matches between location and words (i.e., *joy* presented on the top of the screen) better than mismatches (i.e., *hate* presented on the top of the screen) (Meier & Robinson, 2004; Meier, Hauser, Robinson, Friesen, & Schjeldahl, 2007; Pecher, van Dantzig, Boot, Zanzolie, & Huber, 2010; Schnall & Clore, 2004; Schubert, 2005). More recently, Santana and de Vega (2011) found that matches facilitate comprehension of figurative language more so than that of literal language, suggesting that positive and negative metaphors are also processed through embodied mechanisms. Similarly, when pictures are presented in their expected spatial positions (i.e., an image of a positive concept presented on the top of screen) comprehension of such affectively salient pictures is facilitated (Crawford, Margolies, Drake, & Murphy, 2006; Meier et al., 2007). Furthermore, when participants are in a positive mood, they are more likely to exhibit upwards biases during line bisection tasks, with the opposite pattern holding true for negative moods (Wapner & Werner, 1957). Subjects even feel more successful (a positive feeling) when standing erect (a high vertical position) and less successful (a negative feeling) when slumped over (a low vertical position) (Stepper & Strack, 1993). In a review of the literature, Meier and Robinson (2008) summarized that affect is indeed understood through embodied relations, including vertical spatial representations.

However, such strong evidence supporting embodied representations diverts our attention away from other explanations for these findings. Paivio (1986) has extensively shown that both verbal and non-verbal representations play important roles in cognition. Barsalou, Santos, Simmons, and Wilson (2008) and Louwerse (2007; 2008; 2011) have similarly argued that both statistical

relationships between symbols and embodied representations work together to represent conceptual knowledge. Louwerse (2007; 2011) proposed the Symbol Interdependency Hypothesis, arguing that language encodes embodied representations. That is, embodied cues are encoded linguistically, so that language users can rely on the linguistic system as a shortcut to the perceptual system. Consequently, comprehension can be explained both by a statistical linguistic approach and an embodied approach, a conclusion supported by an increasing amount of empirical evidence (Louwerse, 2011).

Louwerse and Jeuniaux (2010) further argued that language processing relies on both linguistic and embodiment factors variably depending on the cognitive task. They showed that linguistic factors best explained response times (RTs) when participants made semantic judgments about word pairs, but perceptual factors best explained RTs when participants made iconicity judgments about pictures. Importantly, both linguistic and perceptual factors explained RTs in both semantic judgments and iconicity judgments, for both linguistic stimuli and pictorial stimuli, but their relative importance changed as a function of the task and stimulus.

In addition, Louwerse and Connell (2011) have shown that findings initially entirely attributed to embodied representations, such as the increased processing times for modality shifts, can in fact be attributed to statistical linguistic frequencies. Interestingly, faster processing can be best explained by the linguistic system, and slower processing can best be explained by the perceptual system.

In addition to cognitive tasks modifying statistical linguistic and embodiment effects on comprehension, individual differences might also play a role in how we represent information. Preexisting strategies or habits may impact an individual's propensity for processing information in different ways. For example, because embodied representations ground language through bodily experiences we could predict that those individuals with superior spatial ability are more likely to utilize such simulations. Similarly, those with enhanced language skill may show an inclination to process information in a linguistic fashion. This possibility is easily explored through gender differences, with males showing greater general spatial ability (Kimura, 2000; Linn & Peterson, 1985) and females showing greater general language ability (Kimura, 2000; Kramer, Delis, Kaplan, & O'Donnell, 1997). Based on such tendencies, we hypothesized that because males may have a greater affinity to encode information in an imagistic format, embodiment factors should better predict RTs for males. Likewise, females might be more likely to encode information in a symbolic manner, thus linguistic factors were hypothesized to better predict female RTs.

The current study had three goals. First, we aimed to investigate the extent to which a statistical linguistic approach and an embodied approach explained processing of valence words. As in Louwerse and Jeuniaux (2010) we identified embodied and linguistic factors and determined

how well each factor could explain RTs. Second, we aimed to investigate whether the effect of a statistical linguistic approach and an embodied approach is modified as a function of the cognitive task. As demonstrated by the aforementioned studies, presenting positive-negative word pairs in a vertical orientation seems to encourage subjects to perceptually simulate the words they are reading. However, we also know that positive words usually occur before negative words in texts (e.g., *plus and minus, good and bad, positive and negative*), thus if we present words horizontally (as we read them), we might expect subjects to instead rely upon statistical linguistic features. We therefore hypothesized that embodiment factors would be more salient when words were presented vertically whereas linguistic factors would be more salient when words were presented horizontally. Thirdly, we aimed to investigate whether the effect of a statistical linguistic approach and an embodied approach is modified as a function of individual differences, and more specifically as a function of participant gender. We answered these questions in two experiments whereby male and female participants responded to positive and negative valence word pairs. In Experiment 1 word pairs were presented in a vertical configuration (e.g., *happy* above *sad*) to encourage subjects to rely upon embodied features; in Experiment 2 word pairs were presented in a horizontal configuration (e.g., *happy* preceding *sad*) to encourage subjects to rely upon statistical linguistic information.

## Experiment 1

### Participants

Thirty-four undergraduate native English speakers at the University of Memphis (24 females) participated for extra credit in a Psychology course.

### Materials

The experiment consisted of 50 pairs of words that were opposites on a valence dimension (e.g., positive-negative) (see Table 1). One hundred filler items consisted of word pairs without a positive-negative relation, with half of the pairs having a high semantic association and half having a low semantic association as determined by latent semantic analysis (LSA), a computational linguistic technique that measures the similarity in meaning between word pairs, but ignores an order relation (Landauer, McNamara, Dennis, & Kintsch, 2007) (high semantic association:  $\cos = .44$ ; low semantic relation:  $\cos = .18$ ).

### Procedure

Participants were asked to judge the semantic relatedness of word pairs presented on an 800x600 computer screen running E-Prime software (Psychology Software Tools Inc., Pittsburgh, PA). Words were presented one above another, in a vertical orientation. Upon presentation of a word pair, participants indicated whether the pair was related in meaning by pressing designated yes or no keys. All word

pairs were randomly ordered for each participant to negate any order effects. Each participant saw all word pairs, but whether a participant saw a word pair in an iconic or a reverse iconic order was counterbalanced between two groups, such that all participants saw iconic and reverse-iconic word pairs, but no participant saw a word pair both in an iconic and a reverse-iconic order. To ensure participants understood the task, participants completed five practice trials before beginning the experimental task.

Table 1: Positive-Negative Critical Word Pairs

achievement – failure	add – subtract
angel – devil	angels – demons
appear – disappear	beautiful – ugly
bright – dim	birth – death
clean – dirty	confident – arrogant
day – night	dead – alive
dream – nightmare	excitement – boredom
fast – slow	freedom – slavery
friend – enemy	fun – boring
gain – loss	good – bad
grow – shrink	handsome – ugly
happy – sad	healthy – sick
heaven – hell	hero – villain
laugh – scream	life – death
love – hate	more – less
on – off	optimist – pessimist
pass – fail	plus – minus
positive – negative	pretty – ugly
progress – stagnation	protagonist – antagonist
regular – irregular	right – wrong
safe – danger	smile – frown
strong – weak	sunshine – rain
true – false	unite – split
victory – defeat	wealthy – poor
winner – loser	yes – no

## Results and Discussion

Subjects whose RTs fell more than 2.5 SD from the mean per condition, per subject were removed from the analysis, affecting 3.7% of the data.

### Linguistic and Embodiment Factors

We distinguished between a linguistic and an embodiment factor in order to determine whether a statistical linguistic approach or an embodiment approach would better explain the processing of valence words.

**Linguistic factor** The statistical linguistic factor was operationalized as the log frequency of a-b (e.g., *happy-sad*) or b-a (e.g., *sad-happy*) order of word pairs. The order frequency of all word pairs within 3-5 word grams was obtained using the large *Web 1T 5-gram* corpus (Brants & Franz, 2006).

**Embodiment factor** If conceptual metaphors are understood through embodied representations and processes (Lakoff & Johnson, 1981; 1999), then a concept must be perceptually simulated in order to determine if it is positive or negative. Therefore, we operationalized embodiment factor as a rating of how positive or negative a concept was, following previous studies (Louwerse, 2008; Louwerse & Jeuniaux, 2010). Thirty-eight participants at the University of Memphis were asked to what extent they agreed with the statement: *x is more positive than y* (e.g., *happy* is more positive than *sad* or *sad* is more positive than *happy*). Ratings were made for all word pairs (in both orders) on a scale of 1-6, with 1 being strongly disagree and 6 being strongly agree.

### Response Time Analyses

A mixed-effect regression model analysis was conducted on RTs with linguistic and embodiment factors as the fixed factors and participants and items as random factors (Baayen, Davidson, & Bates, 2008). The model was fitted using the restricted maximum likelihood estimation (REML) for the continuous variable (RT). F-test denominator degrees of freedom were estimated using the Kenward-Roger's degrees of freedom adjustment to reduce the chances of Type I error (Littell, Stroup, & Freund, 2002).

The embodiment factor was significantly related to the RTs,  $F(1, 86.93) = 10.40, p < .002$ , with higher ratings yielding lower RTs. This finding supports an embodied cognition account: the more a word pair marked a positive-negative relation, the stronger the effect on the RTs. However, the linguistic factor also explained the RTs,  $F(1, 85.511) = 70.96, p < .001$ , with higher frequencies yielding lower RTs. These findings show that the vertical configuration of conceptual metaphors can be explained by both the linguistic system and the embodied system, confirming previous findings and the claim that language processing is both linguistic and embodied.

### Gender Effects

Next, we investigated whether the linguistic and embodiment factors were modulated by individual differences. To test whether embodiment factors better predicted RTs for males, and linguistic factors better predicted RTs for females, we conducted a mixed effects analysis on RTs using the interaction between gender x linguistic factor and gender x embodiment factor as fixed factors and subject and item as random factors. The two interactions on RTs were significant, for both gender x linguistic factor,  $F(2, 113.86) = 38.68, p < .001$ , and gender and embodiment factor,  $F(2, 210.99) = 8.23, p < .001$ . Interestingly, and as predicted, the effect size was largest for females x linguistic factor, and males x embodiment factor, as shown in Figure 1.



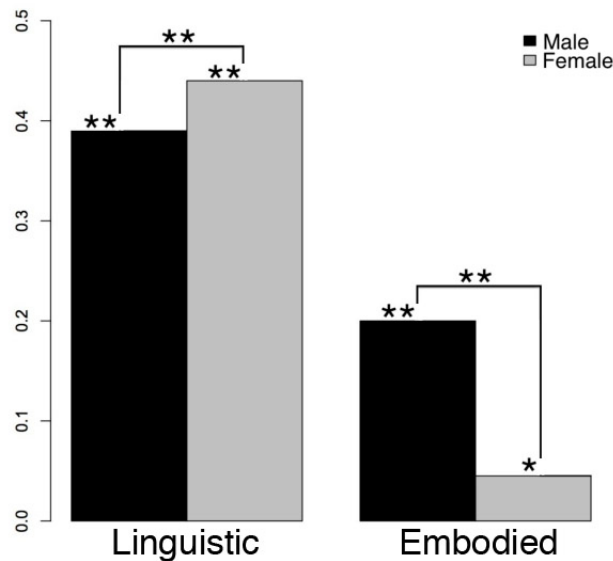


Figure 1. The effect sizes in  $R^2$  for the interaction of participant gender x linguistic factor and the interaction of participant gender x embodiment factor on RT in vertical configuration of valence word pairs.

The results of Experiment 1 show that both linguistic and embodiment factors explain processing of valence words. The effects of either factor are modulated by individual differences. Although both male and female participants depended more upon the linguistic factor, females relied more on the linguistic factor than males, and males relied more on the embodiment factor than females.

## Experiment 2

### Participants

Forty undergraduate native English speakers at the University of Memphis (29 females) participated for extra credit in a Psychology course.

### Materials

Materials were identical to those used in Experiment 1.

### Procedure

The procedure was identical to Experiment 1 except that items were now presented next to each other (horizontal configuration) rather than above each other (vertical configuration).

## Results and Discussion

### Linguistic and Embodiment Factors

As in Experiment 1 a mixed effects regression model was conducted with RT as the outcome variable, with the linguistic and the embodiment factors as fixed predictor factors, and with both participants and items as random factors. Again, the embodiment factor was significantly related to the RTs,  $F(1, 84.93) = 7.01, p = .01$ , with higher embodiment ratings yielding lower RTs. The linguistic

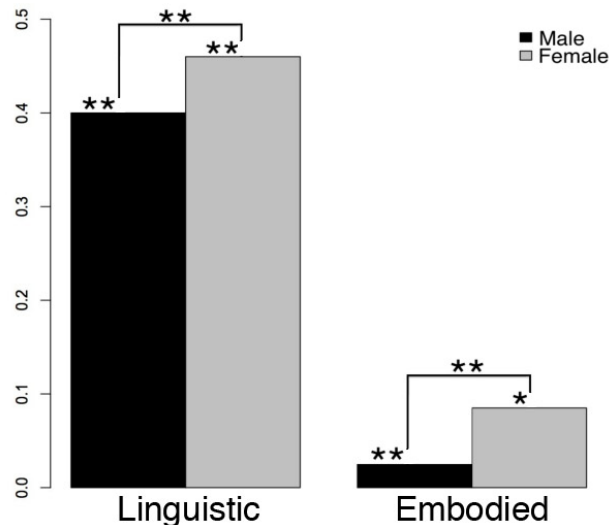


Figure 2. The effect sizes in  $R^2$  for the interaction of participant gender and the linguistic factor and the embodiment factor on RT in horizontal configuration of valence word pairs.

factor also explained RTs,  $F(1, 86.01) = 94.60, p < .001$ , with higher frequencies yielding lower RTs. These findings are similar to those obtained in Experiment 1, and support the conclusion that language processing is both linguistic and embodied.

### Gender Effects

As before, to determine interactions between gender x linguistic factor and gender x embodiment factor, we conducted a mixed effects model using the two interactions as fixed factors and subjects and items as random factors. Both interactions reached significance, for gender x linguistic factor,  $F(2, 142.89) = 48.12, p < .001$ , as well as gender x embodiment factor  $F(2, 207.74) = 4.19, p = .02$ . The effect size was again largest for female participants and the linguistic factor (Figure 2). However, effect size was now also larger for females and the embodiment factor. A possible explanation for this finding is that the embodiment factor may have played a lesser role in the horizontal configuration than in the vertical configuration, an explanation that fits the idea that cognitive task modulates the relative importance of linguistic and embodiment factors. This issue was investigated next.

Next, we were concerned with the relationship between the prominence of a linguistic factor and a embodiment factor during processing. To answer this question we conducted an analysis using both the linguistic and embodiment factor, comparing their relative importance in the vertical (Experiment 1) versus horizontal (Experiment 2) configurations. Again, both participants and items were used as random factors in the mixed effects regression, and the interaction of the linguistic factor x orientation (vertical versus horizontal), and the interaction of the embodiment factor x orientation were of interest. The interaction for both orientation x linguistic factor and orientation x embodiment

factor was significant,  $F(2, 133.13) = 53.75, p < .001$  and  $F(2, 163.86) = 6.32, p < .01$  respectively. Furthermore, an interaction between orientation  $\times$  linguistic factor  $\times$  embodiment factor reached significance,  $F(2, 165.28) = 4.90, p < .01$ . Although the linguistic factor played a more important role in both configurations, the linguistic factor was more salient in the horizontal configuration than in the vertical configuration, whereas the embodiment factor was more salient in the vertical configuration than in the horizontal condition (Figure 3). These results suggest that the relative importance of linguistic and embodiment factors is modulated by the cognitive task.

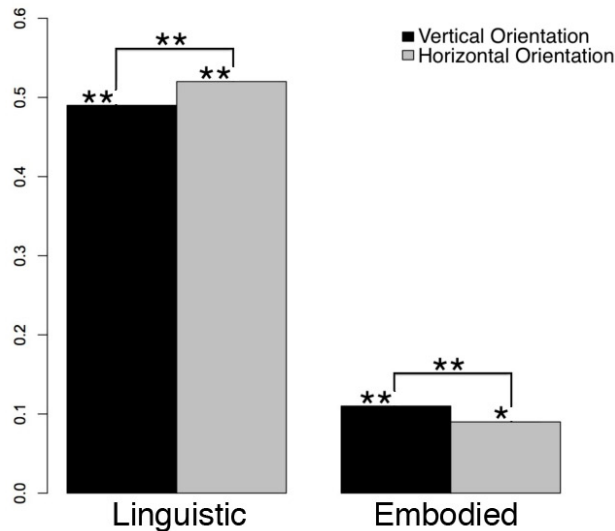


Figure 3. The effect sizes in  $R^2$  for both the linguistic factor and the embodiment factor in vertical orientation of valence word pairs (Experiment 1) and horizontal orientation of valence word pairs (Experiment 2).

## Discussion

Previous studies have shown that when comprehenders process conceptual metaphors, they activate perceptual simulations (Lakoff & Johnson, 1981; 1999). The notion that metaphor comprehension is embodied is non-trivial. However, it begets the question whether alternative explanations are to be dismissed. The two experiments reported here show that both linguistic and embodiment factors explain the processing of valence words. Furthermore, the dominance of one factor relative to the other was modified differentially based on the experimental task, such that linguistic factors better explained RTs for horizontal presentations, with embodied features better explaining RTs for vertical presentations. Moreover, we found evidence suggesting that male participants typically rely more on embodied representations, whereas female participants typically rely more on statistical linguistic patterns.

The findings reported in this paper are important for theories of cognition. Although there is much evidence supporting embodied accounts of mental representations (Barsalou, 1999; Glenberg, 1997; Pecher & Zwaan, 2005;

Semin & Smith, 2008; Zwaan, 2004), there is also an increasing amount of evidence suggesting that other factors can play an important role (Hutchinson, Johnson, & Louwerse, 2011; Louwerse, 2008; Louwerse & Jeuniaux, 2010; Barsalou, Santos, Simmons, & Wilson, 2008). It is important that researchers consider alternatives contributing to metaphoric conceptual processing rather than fixating on one explanation without leaving any room for alternative explanations. Our findings do not dispute the necessity of employing embodied representations but rather we call into question their dominance during cognition. Comprehension of conceptual metaphors can be explained by both linguistic and embodiment factors, but their salience as predictors of cognitive processing is modified by cognitive task and by individual differences.

## References

- Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.
- Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-660.
- Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. de Vega, A. M. Glenberg, & A. C. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 245-283). Oxford, UK: Oxford University Press.
- Brants, T., & Franz, A. (2006). Web 1T 5-gram version 1. Philadelphia: Linguistic Data Consortium.
- Crawford, E., Margolies, S., Drake, J., & Murphy, M. (2006). Affect biases memory of location: Evidence for the spatial representation of affect. *Cognition & Emotion*, 20, 1153-1169.
- de Vega, M., Glenberg, A. M., & Graesser, A. C. (Eds.). (2008). *Symbols and embodiment: Debates on meaning and cognition*. Oxford, UK: Oxford University Press.
- Glenberg, A. M. (1997). What memory is for: Creating meaning in the service of action. *Behavioral and Brain Sciences*, 20, 1-55.
- Hutchinson, S., Johnson, S., & Louwerse, M. M. (2011). A linguistic remark on SNARC: Language and perceptual processes in Spatial-Numerical Association. In L. Carlson, C. Hoelscher, & T. Shipley (Eds.), *Proceedings of the 33<sup>rd</sup> Annual Conference of the Cognitive Science Society* (pp. 1313-1318). Austin, TX: Cognitive Science Society.
- Kimura, D. (2000). *Sex and cognition*. Cambridge, MA: The MIT Press.
- Kramer, J. H., Delis, D. C., Kaplan, E., & O'Donnell, L. (1997). Developmental sex differences in verbal learning. *Neuropsychology*, 11, 577-584.
- Lakoff, G., & Johnson, M. (1981). *Metaphors we live by*. Chicago, IL: University of Chicago Press.
- Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*.

- New York, NY: Basic Books.
- Landauer, T., McNamara, D., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.
- Linn, M. C., & Peterson, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, 56, 1479-1498.
- Littell, R. C., Stroup, W., & Freund, R. J. (2002). *SAS system for linear models*. Cary, NC: SAS Publishing.
- Louwerse, M. M. (2007). Symbolic or embodied representations: A case for symbol interdependency. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 107-120). Mahwah, NJ: Erlbaum.
- Louwerse, M. M. (2008). Embodied relations are encoded in language. *Psychonomic Bulletin and Review*, 15, 838-844.
- Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *TopiCS in Cognitive Science*, 3, 273-302.
- Louwerse, M. M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, 35, 381-398.
- Louwerse, M. M., & Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. *Cognition*, 114, 96-104.
- Meier, B., Hauser, D., Robinson, M., Friesen, C., & Schjeldahl, K. (2007). What's "up" with god? Vertical space as a representation of the divine. *Journal of Personality and Social Psychology*, 93, 699-710.
- Meier, B. P., & Robinson, M. D. (2004). Why the sunny side is up. *Psychological Science*, 15, 243-247.
- Meier, B. P., & Robinson, M. D. (2008). The metaphorical representation of affect. *Metaphor and Symbol*, 20, 239-257.
- Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford, UK: Oxford University Press.
- Pecher, D., van Dantzig, S., Boot, I., Zanzolie, K., & Huber, D. E. (2010). Congruency between word position and meaning is caused by task Induced spatial attention. *Frontiers in Psychology*, 1, 1-8.
- Pecher, D., & Zwaan, R. A., (Eds.). (2005). *Grounding cognition: The role of perception and action in memory, language, and thinking*. Cambridge, UK: Cambridge University Press.
- Prinz, J. (2004). *Furnishing the mind: Concepts and their perceptual basis*. Cambridge, MA: MIT Press.
- Santana, E., & de Vega, M. (2011). Metaphors are embodied, and so are their literal counterparts. *Frontiers in Psychology*, 2, 1-12.
- Schnall, S., & Clore, G. (2004). Emergent meaning in affective space: Conceptual and spatial congruence produces positive evaluations. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual meeting of the cognitive science society* (pp. 1209-1214). Mahwah, NJ: Erlbaum.
- Schubert, T. W. (2005). Your highness: Vertical positions as perceptual symbols of power. *Journal of Personality and Social Psychology*, 89, 1-21.
- Semin, G., & Smith, E., (Eds.). (2008). *Embodied grounding: Social, cognitive, affective, and neuroscientific approaches*. New York, NY: Cambridge University Press.
- Stepper, S., & Strack, F. (1993). Proprioceptive determinants of emotional and nonemotional feelings. *Journal of Personality and Social Psychology*, 64, 211-220.
- Wapner, S., & Werner, H. (1957). The effect of success and failure on space localization. *Journal of Personality*, 25, 752-756.
- Zwaan, R. A. (2004). The immersed experiencer: Toward an embodied theory of language comprehension. *Psychology of Learning and Motivation*, 44, 35-62.

# Knowledge-based Modeling in Dynamic Decision Making

**Angel Iglesias (angel.iglesias@csic.es)**

Bioengineering Group, Spanish National Research Council (CSIC)  
Arganda del Rey, 28500-Madrid, Spain

**M. Dolores Del Castillo (lola@iai.csic.es)**

Bioengineering Group, Spanish National Research Council (CSIC)  
Arganda del Rey, 28500-Madrid, Spain

**J. Ignacio Serrano (jignacio.serrano@csic.es)**

Bioengineering Group, Spanish National Research Council (CSIC)  
Arganda del Rey, 28500-Madrid, Spain

**Jesus Oliva (jesus.oliva@csic.es)**

Bioengineering Group, Spanish National Research Council (CSIC)  
Arganda del Rey, 28500-Madrid, Spain

## Abstract

A knowledge-based model that emulates human behavior in a Dynamic Decision Making task is proposed. The model, MAIDEN-DSF, uses a connectionist representation of knowledge and a value function to compute the best alternative. In order to validate MAIDEN-DSF, two data sets have been used: a training set and a test set that contain the behavior of participants that performed the task with different conditions. The results suggest that MAIDEN-DSF is a considerable framework in order to model human behavior. The aim of this paper is to use MAIDEN-DSF to prove that participants do not perceive delay conditions when dealing with Dynamic Decision Making tasks.

**Keywords:** Decision Making; Computational Models; Connectionism.

## Introduction

Dynamic Decision Making lies in tasks that require a series of decisions where the state of the world changes, both autonomously and as a consequence of the decisions made (Brehmer, 1992). The Dynamic Stocks and Flows (DSF) task (Gonzalez & Dutt, 2011), emulates such situations with three elements: a single stock represented by a water tank; inflows, which increase the level of the stock; and outflows, which decrease the level of the stock. The goal of this task is to keep the stock at a certain level over 100 time periods. In every time period, a participant can control the stock adding or removing water via two inputs that represent the decision: the user inflow ( $UI$ ) and the user outflow ( $UO$ ) values. Besides, there are two external inputs that represent the environment inflow ( $EI$ ) and outflow ( $EO$ ) values, both of them are not directly controlled by the participant. The dynamics of the environment is unknown to the participant, so the  $EI$  and  $EO$  values have to be predicted using the experience acquired in previous time periods. When the participant makes a decision, the DSF determines the level of water in the tank by adding the  $UI$  and  $EI$  to the level in the tank and subtracting the  $UO$  and the  $EO$ . Then the DSF presents the resulting level, the goal level and the last  $EI$ ,  $EO$ ,  $UI$  and  $UO$  values within the following time period.

This seemingly simple stock problem is actually unintuitive and difficult. In fact, there is evidence that suggests that people do not perceive stocks and flows dynamics correctly. For example, in an experiment where graduate students were asked to sketch the evolution of the water level in a bathtub over time (given simple patterns for the environmental inflow and outflow), only 36% of the students answered correctly (Serman, 2002).

The DSF has been studied under different conditions (Cronin & Gonzalez, 2007) and there is evidence that the hardest condition for controlling the water tank is when the user inputs are delayed for three time periods (Lebiere, Gonzalez, & Warwick, 2010). Participants have some understanding about the dynamics of the tank, for example, if 1 gallon of water is added, then the level is increased 1 gallon. However, with this delay, participants seem to have incorrect beliefs about the relationship between stocks and delayed flows or participants do not perceive the delay and, therefore, they behave as if there was no delay.

The aim of this paper is to use a knowledge-based model that emulates human behavior in the DSF in order to find cues that prove that participants do not actually perceive the delay. The proposed model is based on MAIDEN, a Model of Assessment and Inference of DEcisions based on a Net of concepts (Iglesias, Del Castillo, Serrano, & Oliva, 2010).

## A knowledge-based model

MAIDEN-DSF, the implementation of MAIDEN for dealing with DSF, is divided in two main phases. The first phase lies in the estimation of  $EI$  and  $EO$  values using a connectionist representation of knowledge called decision net. In the second phase, MAIDEN-DSF uses a value function to choose the best alternative, which will be the one that sets the amount in the tank at the goal level corresponding with the estimated  $EI$  and  $EO$  values of the first phase.

## Decision net

MAIDEN-DSF is a knowledge-based model that uses a weighted net of concepts called decision net to predict the values of  $EI$  and  $EO$ . The concepts of the decision net are arranged in five layers: perception, short-term memory, working memory, deliberative and output layer. If the concepts of the decision net change, then the behavior of MAIDEN-DSF also changes. Therefore, the selection of suitable concepts is a key point in the implementation of the decision net. The decision net must be as generic as possible and the concepts of the decision net must represent general knowledge about the decision making task and the past experience.

1. The perception layer contains concepts that are directly shown by the DSF in every time period: the current level in the tank ( $Level_t$ ), the goal level ( $Goal$ ), the environment outflow and inflow in the previous time period ( $EO_{t-1}$  and  $EI_{t-1}$ ), and the last decision of the participant ( $UO_{t-1}$  and  $UI_{t-1}$ ).
2. The short-term memory layer takes into account the recent past experience and contains the last three values of the tank level ( $Level_{t-1}$ ,  $Level_{t-2}$  and  $Level_{t-3}$ ), the environment inflow ( $EI_{t-2}$ ,  $EI_{t-3}$  and  $EI_{t-4}$ ) and the environment outflow ( $EO_{t-2}$ ,  $EO_{t-3}$  and  $EO_{t-4}$ ).
3. The working memory layer provides elaborated information via concepts that represent the increase in the last environment inflow and outflow, the decrease in the last environment inflow and outflow and the positive or negative difference between the current level and the goal level.
4. The deliberative layer aggregate other concepts contained in the previous three layers. This layer is composed of two deliberative concepts.
5. The output layer contains the concepts used by MAIDEN-DSF in the value function: the estimated environment inflow and outflow ( $EI$  and  $EO$ ).

Every node of the decision net has got an associated positive activation value. The decision net propagates activation to one node by performing the weighted sum of the incoming activations from the nodes connected to it. The following expression shows how a node computes the total weighted input  $net_i$ :

$$net_i = \sum_j w_{ij} \cdot a_j \quad (1)$$

In expression 1,  $net_i$  represents the total weighted input of the  $i^{th}$  node,  $a_j$  is the activation of the  $j^{th}$  node and  $w_{ij}$  is the weight of the connection between the  $i^{th}$  and the  $j^{th}$  node. If  $net_i \geq 0$ , then the activation of a node is equal to its  $net$  value. Otherwise activation is zero.

The connections among the different nodes of the decision net have three constraints. First, nodes of the same layer are not interconnected. Second, nodes of the deliberative layer

can only propagate activation to the output layer. Third, a concept belonging to the perception, the short-term memory or the working memory layer can propagate activation to the deliberative layer or the output layer, but it cannot propagate activation to both layers according to neurophysiological evidence (Damasio, 1994; Romanski & LeDoux, 1992).

The concepts of the perception layer are easily identifiable in the DSF. The concepts of the short-term memory and the working memory layer have been extracted from interviews of participants who performed the DSF. Notice that the activation of these concepts must be updated every time period depending on the decisions and consequent outcomes up to the current time period. Figure 1 shows a graphical representation of the decision net.

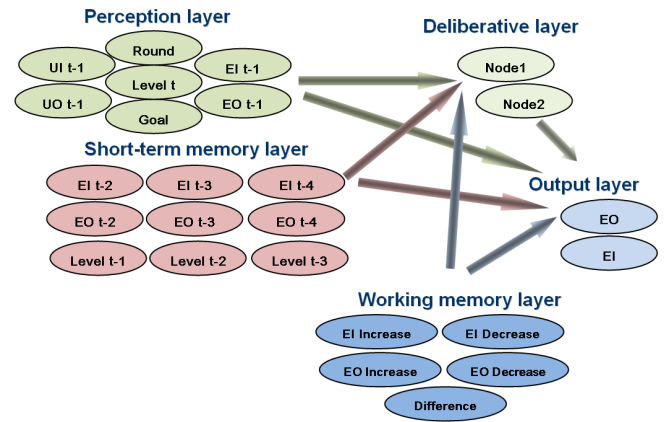


Figure 1: Representation of the decision net.

This design is coherent with Norman's definition of affordances (Norman, 1988): all action possibilities latent in the environment that are perceivable by a participant. These perceived affordances result from the mental interpretation of things based on past knowledge and experience. In MAIDEN-DSF, alternatives or actions are also evaluated based on past knowledge and experience.

## Value function

The goal of the task is to keep the water in the tank at a given level. The expression 2 represents the water level at time period  $t + 1$  ( $Level_{t+1}$ ), which depends on  $UI_t$ ,  $UO_t$ ,  $EI_t$ ,  $EO_t$  and  $Level_t$ .

$$Level_{t+1} = Level_t + UI_t - UO_t + EI_t - EO_t \quad (2)$$

The difference between the water level and the goal level at time period  $t + 1$  is calculated using the following expression:

$$Diff = Level_{t+1} - Goal \quad (3)$$

Since the best decision is the one that sets the stock at the goal level ( $Goal$ ), the value of  $Diff$  must be 0 and the best  $UI_t - UO_t$  value is given by the combination of equations 2 and 3:

$$UI_t - UO_t = Goal - Level_t - EI_t + EO_t \quad (4)$$

The expression 4 computes the best decision ( $UI_t - UO_t$ ) and requires the values of  $Goal$  and  $Level_t$ , which are known to the participant, and the values of  $EI_t$  and  $EO_t$ , which are unknown. Therefore, the value function used by MAIDEN-DSF replaces the unknown variables  $EI_t$  and  $EO_t$  with the activation of the concepts  $EI$  and  $EO$  of the decision net estimated with the information available in the current time period  $t$ , as shown in figure 2.

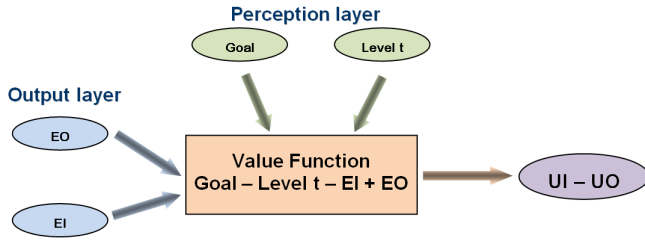


Figure 2: Scheme of MAIDEN's value function.

## Evaluation

The connection weights of the decision net of MAIDEN-DSF have been adjusted to model, as well as possible, the behavior of a set of participants who performed the DSF. The evaluation consists in the prediction of the behavior of another set of participants in different conditions of the DSF. The training and test data has been collected from the DSF challenge (Lebiere et al., 2010) whose homepage presents a wealth of information about the task (<http://www.hss.cmu.edu/departments/sds/ddmlab/modeldsf>). The measure used to evaluate the model fit was Pearson's linear correlation coefficient  $R$ , comparing the decisions of the model and the participants over all 100 time periods. In order to compute the decisions of MAIDEN-DSF, in each time period the activation of the concepts of the perception, short-term memory and working memory layers has been generated with the participant's actual sequence of decisions and consequent water levels up to the current time period. This method ensures that MAIDEN-DSF makes a decision with the same available knowledge as the participant whose behavior is wanted to be modeled.

The values of the connection weights that optimize the correlation of the model and the participants' decisions have been calculated by the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) (Hansen, 2006), which is a kind of Evolutionary Algorithm (Goldberg, 1989; De Jong, 2006) with considerable potential for searching in complex spaces. CMA-ES has been run ten times to find the best connection weights that optimize the correlation of the decisions of MAIDEN-DSF and the decisions of the participants who participated in the training conditions. Then, the correlation of the best connection weights is presented.

## Performance in the training conditions

The DSF with the training conditions used an environment inflow that was either an increasing function or a decreasing function over 100 time periods. The function could be linear or non linear and, therefore, there were four training conditions. The environment outflow function was constant and set to zero throughout the task. The following expression shows the linear increasing function:

$$EI_t = 2.0 + 0.08 \cdot t \quad (5)$$

The goal was to maintain the level of water in 4 gallons during all 100 time periods. The initial water level in the tank was fixed to 2 gallons. The training data corresponds with the behavior of 61 participants (linear increasing = 15 participants, linear decreasing = 11 participants, non linear increasing = 17 participants and non linear decreasing = 18).

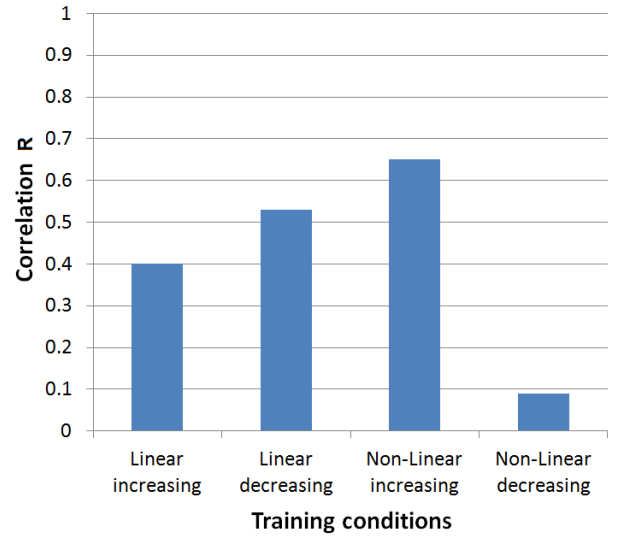


Figure 3: Correlation between MAIDEN-DSF and human decisions with the training conditions.

Figure 3 shows the correlation between the decisions of MAIDEN-DSF, whose connection weights have been optimized in order to fit the training data, and the decisions of the participants with the training conditions. All the  $R$  values are statistically significant using a Student's  $t$  distribution for a transformation of the correlation ( $p < 0.001$ ). These values correspond with the best set of connection weights found by CMA-ES after ten executions.

In the training set, the worst correlation is obtained in the non linear decreasing condition which appears to be the hardest task for the participants. The whole training set of participants has been modeled with the same connection weights. This low correlation value may be due to connection weights that model well the behavior of participants in different conditions but not in the non linear decreasing condition. Note



that CMA-ES finds the connection weights that maximize the overall correlation taking into account the four conditions.

### Performance in the test conditions

The values of the connection weights of MAIDEN-DSF were estimated to optimize the correlation of the model and the participants' decisions with the training conditions. This subsection shows the predictions made by this MAIDEN-DSF in the test conditions in order to validate the model.

The DSF with the test conditions used an environment inflow that was either a sequence-based function or a delayed function. There were three different sequence-based functions that generated a repeated sequence of length 2, a repeated sequence of length 4 and the same sequence of length 4 with noise. There were also two delayed functions where participant's decisions were delayed for either two or three time periods. Therefore, there were five test conditions. The environment outflow function was also constant and set to zero throughout the task. The goal was to maintain the level of water in 6 gallons during all 100 time periods and the initial water level in the tank was fixed to 4 gallons. The test data corresponds with 100 participants (delay 2 = 20 participants, delay 3 = 20 participants, sequence 2 = 20 participants, sequence 4 + noise = 20 participants and sequence 4 = 20 participants).

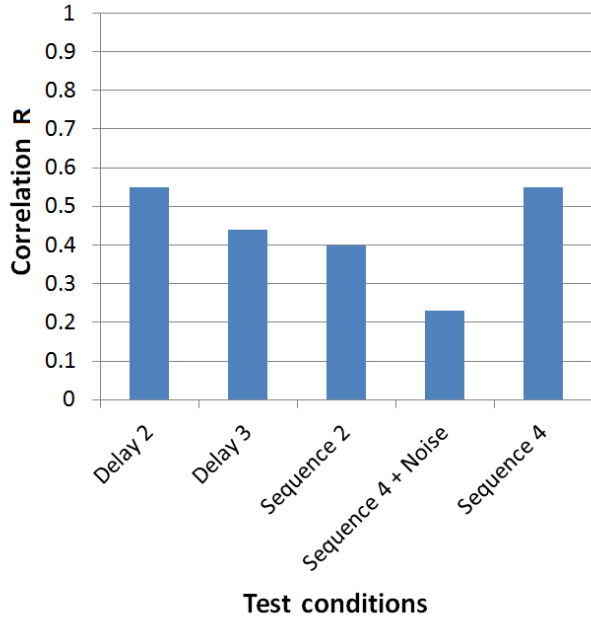


Figure 4: Correlation between MAIDEN-DSF and human decisions with the test conditions.

Figure 4 shows the correlation between the decisions of MAIDEN-DSF and the decisions of the participants with the test conditions. The  $R$  values obtained with the test conditions are all statistically significant ( $p < 0.001$ ).

It is noteworthy that the decision net of MAIDEN-DSF

does not contain any concept representing the delay; however, it models the human behavior with the delay conditions with a correlation higher than 0.4, even when the training conditions do not contain any delay. This result suggests that participants may not take into account the delay or may not understand it while performing the DSF.

In the delay 3 condition the correlation is reduced. This can be due to the high variability observed in the human decisions, for example, a participant decided to increase the level of water in  $1E + 09$  gallons.

With respect to the sequence-based conditions, the short-term memory layer of the decision net contains concepts from the previous four time periods up to the current one. A good performance in the sequence-based conditions implies the knowledge of the last two environmental inputs, if it is a sequence of length 2, or the last four environmental inputs, if it is a sequence of length 4. MAIDEN-DSF contains concepts representing the last four environmental inputs, so it may suitably model behaviors with these conditions.

### Delay conditions

The aim of this work is to find cues that prove whether participants do not perceive the delay during the DSF within the delay conditions. Although the results obtained in the previous experiment may already suggest that participants ignore delays, this section presents another cue to support the experiment.

A new version of MAIDEN-DSF with concepts that represent the delay has been also built. This new version will be denoted as MAIDEN-DSFd. The new decision net is composed of the concepts already explained in this paper with the addition of four new concepts:

- The short-term memory layer takes also into account the following user decisions:  $UI_{t-2}$ ,  $UI_{t-3}$ ,  $UO_{t-2}$  and  $UO_{t-3}$ .

Note that with the delay conditions the user inputs are delayed for two or three time periods, so the decisions relevant for the current tank level  $Level_t$  are taken in  $t - 2$  and  $t - 3$ .

The connection weights of MAIDEN-DSF and MAIDEN-DSFd have been adjusted to model the behavior of the 40 participants that participated in the delay conditions. CMA-ES has been applied ten times for each model version to estimate the connection weights that maximize the correlation between the model decisions and the decisions of the participants.

Table 1: Mean correlation (E) and standard deviation (SD) obtained by MAIDEN-DSF and MAIDEN-DSFd with the delay conditions.

Model	Delay 2		Delay 3	
	E	SD	E	SD
MAIDEN-DSF	0.54710	1.0E-06	0.43931	5.0E-09
MAIDEN-DSFd	0.54711	1.3E-06	0.43931	6.0E-09

Table 1 shows the mean correlation and the standard deviation of the ten solutions found by CMA-ES for each version



of MAIDEN. The  $R$  values obtained with both implementations are all statistically significant ( $p < 0.001$ ).

Table 2: Mean (E) and standard deviation (SD) of the root mean square error obtained by MAIDEN-DSF and MAIDEN-DSFd with the delay conditions.

Model	Delay 2		Delay 3	
	E	SD	E	SD
MAIDEN-DSF	24275.04	0.033	6227685.70	0.098
MAIDEN-DSFd	24274.98	0.036	6227685.59	0.059

Table 2 shows the mean and the standard deviation of the root mean square error obtained by the ten solutions found by CMA-ES for each version of MAIDEN. The error seems to be high due to the high variability of the human decisions within the delay conditions. In fact, in the DSF challenge (Lebiere et al., 2010) the best ranked model obtained a root mean square error of 25560.84 in the delay 2 condition and 5930065.84 in the delay 3 condition.

MAIDEN-DSF and MAIDEN-DSFd obtained similar correlations in both conditions. Within the delay 2 condition they only differ in 0.00001 and within the delay 3 condition the mean correlation is exactly the same. The standard deviation is also similar. These results point out that the behavior of the participants might be modeled without using the  $UI_{t-2}$ ,  $UI_{t-3}$ ,  $UO_{t-2}$  and  $UO_{t-3}$  decisions.

## Conclusion

A knowledge-based model that emulates human behavior in the Dynamic Stocks and Flows task has been proposed. MAIDEN-DSF operation is divided in two main phases. First, it estimates the environment inflow and outflow values using a decision net. Second, it uses a value function to choose the best alternative, which is the one that sets the water level of the tank at the goal level.

The model has been evaluated with two data sets from the DSF challenge (training set and test set). The evaluation lies in the prediction of human behavior in the test set given the training set. The results achieved support that MAIDEN-DSF is a considerable framework in order to model human behavior in the DSF. Besides, the proposed model participated in the DSF challenge and reached rank 2 regarding the ranking provided by the DSF challenge organizers based on the  $R^2$  correlation and the root mean square error.

MAIDEN-DSF has been used to find cues that point out that the behavior of the participants who performed the DSF task with the delay conditions can be modeled without taking into account any concept representing the delay. This evidence suggests that participants do not take into account or are not aware of the delay while making a decision during the delay conditions.

The results suggest that it is very difficult to configure a model in order to emulate the performance of several human beings with the same parameter values (connection weights in this case). This might lead to develop comparison procedures

focusing on individual modeling and adaptation. MAIDEN is mainly based on the knowledge acquired through past experience, the knowledge extracted from the environment and the relationships between the concepts that represent these two kinds of knowledge. A very interesting feature is that the connection weights of the decision net can show what concepts are the most important for a participant. In this experiment, the whole set of participants has been modeled with the same connection weights. A future work might consist in the individual modeling of each participant in order to study the connection weights that better fit each participant's behavior. An individual modeling may indicate which participants had realized that there was a delay in their decisions. In this experiment the connection weights of the concepts representing  $UI_{t-2}$ ,  $UI_{t-3}$ ,  $UO_{t-2}$  and  $UO_{t-3}$  in MAIDEN-DSFd affect the behavior of the model. An absence of connection between these concepts implies that the participants do not use this knowledge. However, CMA-ES seeks the best combination of connection weights that better fit the whole set of human behaviors, therefore, if there is one person that uses this knowledge, CMA-ES will find a connection weight for these concepts. This is the reason for the better suitability of individual modeling.

A global modeling is suitable for experiments where there are at least two groups: a control group and an experimental group. A future work may study the connection weights that better model the behavior of participants who performed well in the delay conditions (control group) and the participants who obtained poor results (experimental group). The connection weights can point out what concepts affect a decision for a certain participant. For instance, this information may be used to improve participants' performance by explicitly revealing relevant information that participants who obtained poor results tend to ignore but participants who performed well in the delay conditions take into account. This may lead to better understanding of dynamic decision making.

## Acknowledgments

This work has been supported by the JAE Program (Predoctoral research training leading to PhD Theses).

## References

- Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta Psychologica*, 81(3), 211-241.
- Cronin, M. A., & Gonzalez, C. (2007). Understanding the building blocks of dynamic systems. *System Dynamics Review*, 23(1), 1-17.
- Damasio, A. R. (1994). *Descartes' error : emotion, reason, and the human brain*. New York: Putnam.
- De Jong, K. A. (2006). *Evolutionary computation: A unified approach*. Cambridge, Massachusetts: The MIT Press.
- Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

- Gonzalez, C., & Dutt, V. (2011). A generic dynamic control task for behavioral research and education. *Computers in Human Behavior*, 27(5), 1904 - 1914.
- Hansen, N. (2006). The cma evolution strategy: A comparing review. In J. Lozano, P. Larranaga, I. Inza, & E. Bengoetxea (Eds.), *Towards a new evolutionary computation* (p. 75-102). Springer.
- Iglesias, A., Del Castillo, M. D., Serrano, J. I., & Oliva, J. (2010). A psychologically and neurophysiologically plausible model for emulating human behavior in decision making tasks. In *Proceedings of the brain inspired cognitive systems - bics 2010*. Madrid: Universidad Politecnica de Madrid.
- Lebiere, C., Gonzalez, C., & Warwick, W. (2010). Editorial: Cognitive architectures, model comparison and agi. *Journal of Artificial General Intelligence*, 2(2), 1-19.
- Norman, D. (1988). *The psychology of everyday things*. New York: Basic Books.
- Romanski, L. M., & LeDoux, J. E. (1992). Equipotentiality of thalamo-amygdala and thalamo-cortico-amygdala circuits in auditory fear conditioning. *Journal of Neuroscience*, 12, 4501-4509.
- Sterman, J. D. (2002). All models are wrong: Reflections on becoming a systems scientist. *System Dynamics Review*, 18(4), 501-531.

# Rapid entrainment to spontaneous speech: A comparison of oscillator models

**Benjamin Inden**

Faculty of Technology  
Bielefeld University  
binden@techfak.uni-bielefeld.de

**Zofia Malisz**

Faculty of Linguistics and Literary Studies  
Bielefeld University

**Petra Wagner**

Faculty of Linguistics and Literary Studies  
Bielefeld University

**Ipke Wachsmuth**

Faculty of Technology  
Bielefeld University

## Abstract

Oscillator models may be used for modeling synchrony between gestures and speech, or timing of backchanneling and turn-taking in dialogues. We find support for the hypothesis that oscillator networks can better predict rhythmic events on the syllable and foot level than single oscillators, but we do not find support for the hypothesis that phase resetting oscillators perform better than phase adapting oscillators. Overall, oscillators can be used to predict rhythmic events in speech, but higher level information needs to be integrated into such models to reach a satisfactory performance.

**Keywords:** speech rhythm, entrainment

## Introduction

Spontaneous speech, like music, exhibits temporal regularities, but these cannot be captured by simple descriptions. Rhythm, i.e., hierarchical structured temporal regularities, is widely believed to be an important principle for understanding both music and speech (Large, 2008; Cummins & Port, 1998). Regularity of timing greatly contributes to speech perception and understanding. Regular sequences of, e.g., inter-stress intervals in speech or tone sequences in music speed up perception by facilitating meaningful grouping and contrast within a very rich acoustic signal. The role of rhythmic expectancies both in speech and music perception has been the basis of Dynamic Attending Theory (Jones, 1990) and more specific phonological models such as PolySP (Hawkins, 2003). Humans can even perceive rhythms in music that do not directly correspond to any frequency found in a spectral analysis of the signals (Large, 2008). Similarly, the timing of speech production is coordinated via rhythmic principles. The same principles govern the neuro-physiological dynamics of all motor behavior. On the syllable level, the vocalic pulse represents the basic timing coordination of the articulatory system (Browman & Goldstein, 1992).

When measuring brain activity, one can easily find a number of prominent frequencies. Some of them are in the range of typical speech units: the theta band (3-12 Hz) corresponds to the typical duration of a syllable (100-300 ms),

whereas delta band oscillations (0.5-3 Hz) correspond to typical lengths of prosodic and metrical units. Many kinds of oscillators have the property of entraining to an externally provided periodic signal, i.e., they become phase-locked to the signal. Therefore, it seems plausible that neural oscillators might play a role in the production and perception of speech by synchronizing certain systems with the speech signal (Buzsáki & Draguhn, 2004; Ghitza & Greenberg, 2009). Previous research has also found that gestures may be synchronized with speech rhythms (Condon, 1986; Tuite, 1993; Wachsmuth, 1999; Loehr, 2007). Furthermore, listeners can become entrained to a speaker's rhythm, which helps them to provide backchanneling or take turns in a dialogue at a suitable moment in time (Wilson & Wilson, 2005).

A number of oscillator models have been proposed that can entrain to musical rhythms. Here we focus on models proposed by Large and McAuley (Large, 1994; McAuley, 1995). These have been shown to achieve entrainment to input signals not just in a period ratio of 1:1, but also in more complex ratios, which makes coupling between several levels of speech rhythm possible. Furthermore, oscillator banks have been shown to reproduce empirical findings about human perception of rhythm: they can resonate at frequencies that are not present in the input signals, but perceived by human subjects, too (Large, 2008; Large, Almonte, & Velasco, 2010).

We are interested in whether these oscillator models originally built for modeling music perception can also be used to model human entrainment to less regular speech signals. In general, we believe in the necessity to couple oscillators for different levels of the rhythmic hierarchy, so ultimately we will include this in the models discussed in this article. However, here we focus on the question what particular features might make an oscillator model more capable of correctly predicting syllable and foot onset times when considered separately from the other levels of the rhythmic hierarchy. After all, speech is less regular than music, so adapta-

tion to input signals should be very fast. Therefore, we will compare two previously proposed oscillator models and then make a number of changes to one of them to see whether prediction performance improves. In particular, we examine the following hypotheses: First, oscillators that reset their phase upon arrival of an input signal may be faster than those that adapt their phase gradually. Second, it may be better to have a bank of oscillators tuned to different frequencies than to have a single oscillator that adapts its period. This would be because period adaptation time is dependent on the amount of change necessary whereas in banks of oscillators, the time for a differently tuned oscillator to become activated is constant with regards to the amount of frequency change. To the degree these hypotheses turn out to be supported by the data, they can inform future modeling of human entrainment to speech rhythms. Besides addressing these two hypotheses, our experiments also show how much can be learned at all by oscillator models without considering the hierarchical organization of speech, i.e., from a pure low-level approach.

## The Data

The speech data comes from a corpus of spontaneous dialogue in German where one dialogue partner told a holiday story and the other was instructed to listen actively (Buschmeier, Malisz, Włodarczak, Kopp, & Wagner, 2011). The corpus was collected for the purposes of modeling entrainment in dialogue, multimodal behavior of the listener, i.e., feedback signals, head and manual gesture, as well as the prosody of the storyteller. The latter objective is addressed in the present paper.

Audiovisual recordings were made in a sound-treated studio. Participants were positioned approximately three meters apart to minimize crosstalk. Close talking high-quality headset microphones were used. The signal properties were annotated in Praat (Boersma & Weenink, 2012). Careful annotation of the acoustic signal enables to approximate emergent rhythmic phenomena (Gibbon & Fernandes, 2005). To represent the syllabic oscillator hypothesized for speech production, we first semi-automatically extracted vowel onsets from the data (Cummins & Port, 1998; Barbosa, 2006). Secondly, experts annotated rhythmic feet, representing the slower stress oscillator, where each prominent syllable is a pulse on that level. We also annotated interpausal units (IPUs) with a criterion that only minimally perceptible interruptions in the flow of speech were marked (not all acoustic pauses).

For the present simulations, two conversations (henceforth *dataset 1* and *dataset 2*) were used. Phrases (IPUs) consisting of at least two feet events were selected. Any phrase initial unstressed vowel events (anacrusis) were excluded as well as the phrase final vowel event. The trimmings were done to exclude any extra lengthening at the end of phrase and extra irregularity at the beginning of phrase that typically signal a boundary in German. The resulting phrases consist of fluent, spontaneous, uninterrupted speech with a minimal phrase length of one second. The mean duration of a syllable-sized

intervocalic interval was 125 msec in this material and 365 msec for the foot. 69 phrases each from dataset 1 and dataset 2 were provided as input to the different oscillator models, i.e., the resulting onset times for each vowel or foot event served as the input pulse.

For each conversation, a control set of regular phrases was created by generating completely regular pulses with frequencies equal to the mean frequencies of events in the corresponding individual phrases from the conversation data.

## Models of entrainment

### Phase adaptation oscillator (PAO)

This oscillator model is one of several similar models originally proposed by Large for entrainment to musical rhythms (Large, 1994). The phase of this oscillator is defined as  $\phi(t) = \frac{t-t_x}{p}$ , where  $t_x$  is the time of the last event (in the input or according to the oscillator's expectation) and  $p$  is the period of the oscillator. The phase is reset to 0.0 when it reaches 1.0. The output of the oscillator is modeled as a periodic function  $o(t) = 1 + \tanh(\gamma(\cos(2\pi\phi(t)) - 1))$ , where the output gain parameter  $\gamma$  controls the sharpness of the activity peaks. The oscillator has three adaptation rules that depend on the input signal  $s(t)$  as well as learning rates  $\eta_1$ ,  $\eta_2$ ,  $\eta_3$ . The first rule in effect adapts the phase:

$$\Delta t_x = \eta_1 s(t) \frac{p}{2\pi} \text{sech}^2(\gamma(\cos(2\pi\phi(t)) - 1)) \sin(2\pi\phi(t))$$

The second adapts the period:

$$\Delta p = \eta_2 s(t) \frac{p}{2\pi} \text{sech}^2(\gamma(\cos(2\pi\phi(t)) - 1)) \sin(2\pi\phi(t))$$

The third adapts an estimate  $\Omega$  of input variability:

$$\Delta \Omega = \eta_3 s(t) \text{sech}^2(\gamma(\cos(2\pi\phi(t)) - 1)) (\cos(2\pi\phi(t)) + 2\gamma(o(t) - 1) \sin^2(2\pi\phi(t)))$$

This estimate in turn determines the receptive field width  $\tau$  of the oscillator, i.e., the width of a window in time around its maximal activation where it is highly adaptive to input signals:  $\tau = \tau_{\max} + 0.5(\tau_{\min} - \tau_{\max})(1 + \tanh \Omega)$ . The output gain is inversely related to the receptive field width:  $\gamma = \frac{-0.416}{\cos(2\pi\tau) - 1}$ . So if there is less input variability, the receptive field shrinks, and the output peaks are sharper, whereas if there is more input variability, the receptive field grows, and the output peaks are softer. Finally, the output value  $o(t)$  is multiplied by a confidence value  $c = c_{\max} + 0.5(c_{\min} - c_{\max})(1 + \tanh \Omega)$ . Further explanations about the motivation behind these choices, and the behavior of the oscillator, can be found in the literature. The following parameter settings were also taken from the literature:  $\eta_1 = 1.0$ ,  $\eta_2 = 0.3$ ,  $\eta_3 = 0.3$ ,  $\tau_{\min} = 0.02$ ,  $\tau_{\max} = 0.5$ ,  $c_{\min} = 0.0$ ,  $c_{\max} = 1.0$ . Because we expect syllable periods to be in the range  $[0.1, 0.25]$ , and feet periods in the range  $[0.2, 0.5]$ , we set the initial periods of the period and feet oscillators to the middle of these ranges, i.e. 0.175 and 0.35.

## Phase reset oscillator (PRO)

This oscillator has been originally proposed by McAuley for the perception of music, and modified by Nerlich in the context of human-machine interaction (McAuley, 1995; Nerlich, 1998). Its output, like that of the PAO, is a periodic function modified to modulate the sharpness of the output peaks, together with a term for exponential decay of the output:

$$o(t) = \left( \frac{1 + \cos(2\pi\phi(t))}{2} \right)^{(1-\Omega(n))\gamma_{min} + \Omega(n)\gamma_{max}} \exp\left(-\frac{\beta t_x}{p_{ini}}\right)$$

The phase  $\phi(t)$  is always kept in the range  $[-0.5, 0.5]$ , and reset to 0.0 when an input event arrives. The synchrony  $\Omega(n) = (1 - \epsilon)\Omega(n-1) + \epsilon(1 - 2|\phi^r(n)|)$  is measured every time an input event arrives:  $\phi^r(n)$  is the phase of the oscillator at the reset, and  $\epsilon = 0.2$  is a parameter that weights the current impulse against the memory of earlier synchrony with input events.  $\gamma_{min} = 1$  and  $\gamma_{max} = 5$  constrain the range of output sharpening that is dependent on measured synchrony with the train of input events. The final term in the output equation dampens the output exponentially when no input arrives.  $\beta = 0.5$  is the decay rate,  $p_{ini}$  the initial period of the oscillator, and  $t_x$  the time since the last input event.

The period is adapted using  $\Delta p = \alpha \Delta t P M \frac{P}{2}$ , where  $\alpha = 1$  is the entrainment rate, the period coupling term  $P = \phi^r(n)(1 - \Omega(n))$  is dependent on the synchrony and on the phase at the last reset, and the impulse response function  $M = \frac{1}{1 + \exp(-\Gamma(\phi^r(n) \exp(-\Theta t) - 0.5))}$  (with impulse response gain  $\Gamma = 1000$  and impulse response bias  $\Theta = 2$ ) ensures that almost all adaptation is done shortly after an input event. Like in the PAO model, we set initial periods of the period and feet oscillators to 0.175 and 0.35, while all other parameters are taken from the literature.

## Phase reset oscillator network (PRN)

We use a network of 20 parallel oscillators that are similar to the PRO model. However, we let the output decay not when no input arrives as before, but when the individual oscillator is not synchronous with the train of input signals:

$$o_i(t) = \exp(c_d(1 - \sigma_i(t))) \left( \frac{1 + \cos(2\pi\phi(t))}{2} \right)^{c_s}$$

The constant  $c_s = 20$  determines the sharpness of the oscillator output signal (the more oscillators we have in the network for a given frequency range, the higher this constant should be to reduce blurring of the network output), while  $c_d = -20$  determines how much the oscillator output decays depending on its asynchrony. The synchrony is measured each time an input event arrives using  $\sigma_i(t_r) = (1 - c_p)\sigma_i(t_{r-1}) + c_p(1 - \exp(c_e\phi(t_r)^2))$ , where  $c_p = 0.2$  is a constant that weights the current impulse against the memory of earlier synchrony with input events just like in the PRO model, and  $c_e = -200$  determines how much prediction error is still considered synchronous. Using an exponential term here instead of a piecewise linear term as in the PRO model

ensures that only a few oscillators will consider themselves synchronous with the input signal, which again reduces blurring of the network output. Period adaptation is not used in the PRN model, but phase reset works just like in the PRO model.

The initial periods are logarithmically distributed in the range of  $[0.1, 0.25]$  for syllables, and  $[0.2, 0.5]$  for feet. There is an additional network output unit with a sigmoid output function  $n(t) = 1/(1 + \exp(-\sum_i o_i(t-1)))$ , where the  $o_i(t)$  are the outputs of the individual oscillators. This variant of the model is called PRN1. In the variant called PRN2, the output unit is also connected to the network input. After an input event, its output remains zero until the sum of its input has a positive slope. Because there may be high oscillator outputs immediately after an input event that could disrupt this behavior, an absolute refractory period of 5 simulation steps after an input event is enforced unconditionally.

## Results

In experiments presented elsewhere (Malisz, Inden, Wachsmuth, & Wagner, 2012), we fed event signals from the whole conversation into PAO and PRO models and measured their internal phases when an input signal arrived. In those experiments, we found a significant advantage of the PRO model over the PAO model. By contrast, here we feed data from individual phrases separately into the oscillators and measure their average output activation when an input signal arrives. We also measure average output activation when no input signal arrives and take the difference between the averages as a performance measure. As Tables 1 to 4 show, there is no significant advantage of the PRO model over the PAO model in this case. Furthermore, both are at or below random level on most of the real data sets.

As Tables 1 to 4 also show, using a bank of oscillators like PRN1 is a significant improvement over using a single oscillator (and is significantly above random level). When adding the refractory period rule to the oscillator network, performance further improves significantly for almost all datasets.

The output trajectories of the different oscillator models for an example phrase can be seen in Fig. 1.

## Discussion

Our experiments do provide some support to the hypothesis that oscillator networks may be better suited to speech data than single oscillators that adapt their period. Such insights can inform modeling of human rapid entrainment to spontaneous speech. However, the experiments provide no support for the hypothesis that phase resetting oscillators like the McAuley oscillator are better suited to the rather irregular speech data than phase adapting oscillators like the Large oscillator. This might be because performance of both models is so close to chance level when used in that way. More than anything else, these results show that the level of prediction performance that can be reached by considering just one level of speech rhythm is rather low regardless of the used oscillator models.

oscillator model	phrase data			regular control data		
	prediction at vowel onset	prediction at other times	difference	prediction at vowel onset	prediction at other times	difference
PAO	0.241±0.005	0.265±0.006	-0.024±0.007	0.593±0.030	0.186±0.010	0.408±0.041
PRO	0.296±0.011	0.327±0.004	-0.031±0.013	0.544±0.033	0.274±0.007	0.270±0.034
PRN1	0.311±0.013	0.273±0.006	0.039±0.010	0.854±0.012	0.257±0.005	0.597±0.014
PRN2	0.311±0.013	0.168±0.006	0.143±0.011	0.854±0.012	0.139±0.005	0.715±0.014

Table 1: Prediction of vowel onsets (mean oscillator output) for different oscillator models and dataset 1.

	phrase data			regular control data		
	prediction at foot event	prediction at other times	difference	prediction at foot event	prediction at other times	difference
PAO	0.260±0.010	0.288±0.004	-0.028±0.013	0.474±0.029	0.230±0.008	0.244±0.036
PRO	0.316±0.017	0.333±0.004	-0.018±0.018	0.561±0.028	0.322±0.005	0.239±0.030
PRN1	0.356±0.015	0.318±0.006	0.038±0.014	0.714±0.022	0.305±0.006	0.409±0.023
PRN2	0.356±0.015	0.207±0.008	0.149±0.015	0.714±0.022	0.187±0.007	0.527±0.022

Table 2: Prediction of foot events (mean oscillator output) for different oscillator models and dataset 1.

oscillator model	phrase data			regular control data		
	prediction at vowel onset	prediction at other times	difference	prediction at vowel onset	prediction at other times	difference
PAO	0.246±0.006	0.261±0.005	-0.015±0.007	0.689±0.023	0.146±0.008	0.543±0.030
PRO	0.295±0.011	0.323±0.003	-0.027±0.012	0.627±0.027	0.254±0.005	0.372±0.026
PRN1	0.329±0.013	0.273±0.005	0.056±0.010	0.879±0.006	0.252±0.003	0.628±0.008
PRN2	0.329±0.013	0.169±0.005	0.160±0.010	0.879±0.006	0.136±0.003	0.743±0.007

Table 3: Prediction of vowel onsets (mean oscillator output) for different oscillator models and dataset 2.

	phrase data			regular control data		
	prediction at foot event	prediction at other times	difference	prediction at foot event	prediction at other times	difference
PAO	0.307±0.011	0.293±0.004	0.015±0.013	0.344±0.019	0.281±0.006	0.063±0.024
PRO	0.376±0.020	0.358±0.005	0.018±0.020	0.470±0.026	0.372±0.006	0.098±0.029
PRN1	0.255±0.021	0.236±0.016	0.019±0.012	0.558±0.031	0.245±0.014	0.313±0.024
PRN2	0.255±0.021	0.172±0.013	0.083±0.014	0.558±0.031	0.172±0.012	0.386±0.026

Table 4: Prediction of foot events (mean oscillator output) for different oscillator models and dataset 2.

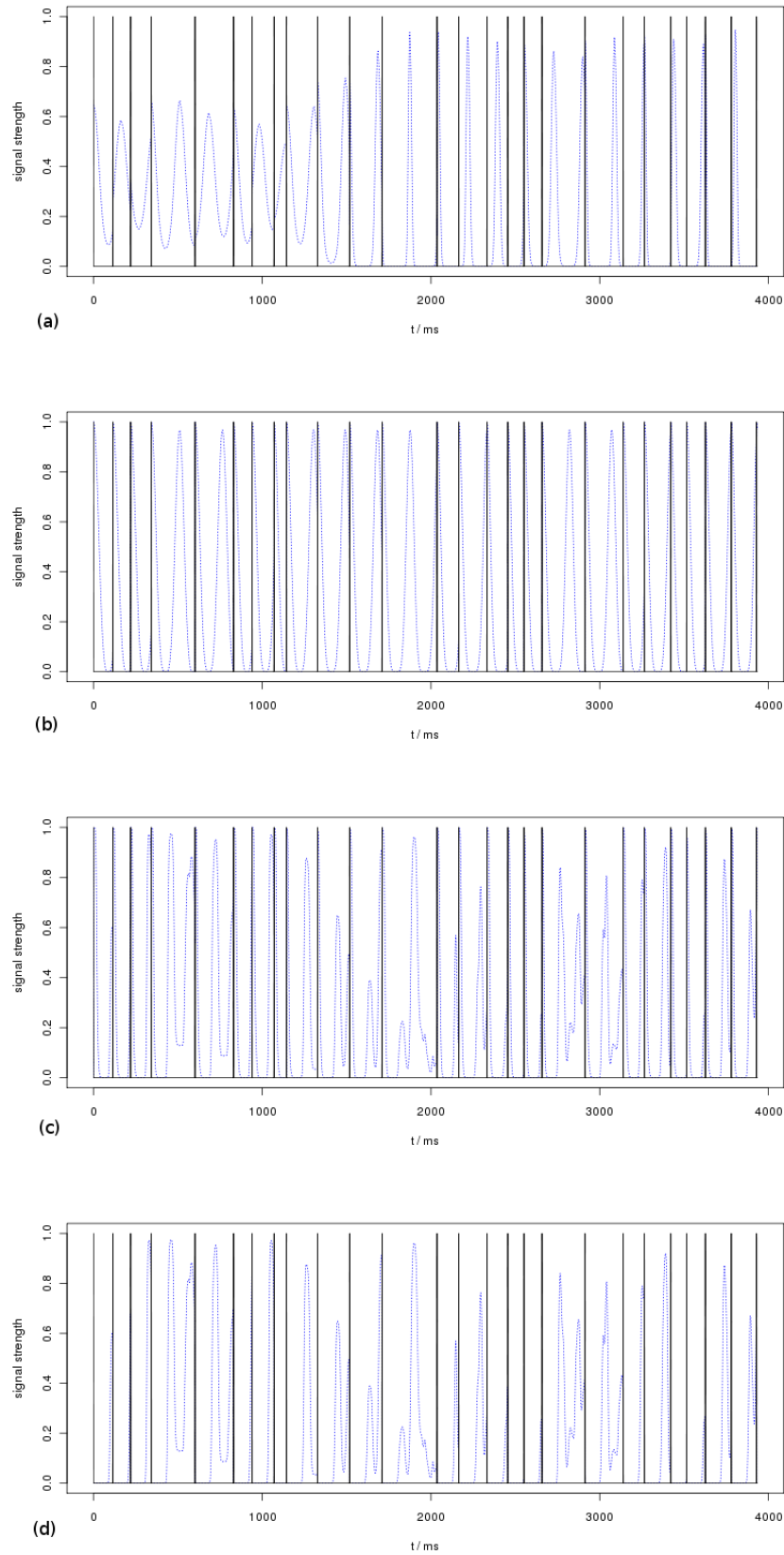


Figure 1: Example output trajectories (blue) and vowel onsets (black) for syllables in the German phrase “... eine Urlaubsreise mit meiner Familie, also ich war mit meiner Schwester und meiner Mutter dort.” (“... a vacation trip with my family, that is, I was there with my sister and my mother.”) (a) PAO model, (b) PRO model, (c) PRN1 model, (d) PRN2 model.



The parameters used for the oscillator models seem to be reasonable and have been found by looking at the literature (PAO and PRO models) or preliminary experiments (PRN model). However, it cannot be totally excluded that other parameter settings will lead to higher performance. Therefore, we searched the space of the most important parameters of the PAO and PRO models for better performance on a randomly selected subset of the data for one conversation using evolutionary algorithms (De Jong, 2006) (details and results not shown here). While the evolutionary algorithm found different parameter settings that performed better on the training set, the subsequent performance on the complete set of data was only marginally better in most cases, and did not change any of the previously mentioned conclusions.

Future work will include using coupled syllable and foot oscillators, and possibly using evidence for vocal activity rhythms, i.e., cycles in pauses and hesitations in dialogue, to model the structure of the interpausal units (McGarva & Warner, 2003; Merlo & Barbosa, 2010). Ultimately, we aim to use the output from entrained oscillators to control the timing of backchanneling and turn-taking in artificial embodied conversational agents (Kopp, Allwood, Grammer, Ahlsen, & Stockmeier, 2008; Poppe, Truong, Reidsma, & Heylen, 2010).

**Acknowledgments** This research is kindly supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673.

## References

- Barbosa, P. A. (2006). *Incursões em torno do ritmo da fala*. Campinas: Pontes.
- Boersma, P., & Weenink, D. (2012). *Praat: Doing phonetics by computer. version 5.3.04*. Retrieved 30 January 2012, from <http://www.praat.org/>
- Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49, 155-180.
- Buschmeier, H., Malisz, Z., Włodarczak, M., Kopp, S., & Wagner, P. (2011). 'Are you sure you're paying attention?' – 'Uh-huh'. Communicating understanding as a marker of attentiveness. In *Proceedings of Interspeech 2011* (pp. 2057–2060). Florence, Italy.
- Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, 304, 1926-1929.
- Condon, W. S. (1986). Rhythm in psychological, linguistic and musical processes. In J. Evans & M. Clynes (Eds.), (p. 55-77). Springfield, Ill.: Thomas.
- Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in english. *Journal of Phonetics*, 26, 145-171.
- De Jong, K. A. (2006). *Evolutionary computation — a unified approach*. MIT Press.
- Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, 66, 113-126.
- Gibbon, D., & Fernandes, F. R. (2005). Annotation-mining for rhythm model comparison in brazilian portuguese. In *Proceedings of interspeech*.
- Hawkins, S. (2003). Roles and representations of systematic phonetic fine detail in speech understanding. *Journal of Phonetics*, 31, 373-405.
- Jones, M. R. (1990). Learning and the development of expectancies: an interactionist approach. *Psychomusicology*, 9, 193-228.
- Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., & Stockmeier, T. (2008). Modeling embodied feedback with virtual humans. In I. Wachsmuth & G. Knoblich (Eds.), *Modeling communication with robots and virtual humans*. Springer-Verlag Berlin Heidelberg.
- Large, E. W. (1994). *Dynamic representation of musical structure*. Unpublished doctoral dissertation, The Ohio State University.
- Large, E. W. (2008). The psychology of time. In S. Grondin (Ed.), (chap. Resonating to musical rhythm: Theory and experiment). West Yorkshire: Emerald.
- Large, E. W., Almonte, F. V., & Velasco, M. J. (2010). A canonical model for gradient frequency neural networks. *Physica D*, 239, 905-911.
- Loehr, D. (2007). Aspects of rhythm in gesture and speech. *Gesture*, 7, 179-214.
- Malisz, Z., Inden, B., Wachsmuth, I., & Wagner, P. (2012). An oscillator based modeling of german spontaneous speech rhythm. In *Perspectives on rhythm and timing workshop*. Glasgow, UK.
- McAuley, J. D. (1995). *Perception of time as phase*. Unpublished doctoral dissertation, Indiana University, Bloomington.
- McGarva, A. R., & Warner, R. M. (2003). Attraction and social coordination: Mutual entrainment of vocal activity rhythms. *Journal of Psycholinguistic Research*, 32, 335-354.
- Merlo, S., & Barbosa, P. A. (2010). Hesitation phenomena: a dynamical perspective. *Cognitive Processing*, 11, 251-261.
- Nerlich, U. (1998). *Rhythmische Segmentierung sprachlicher Instruktionen in einem Mensch-Maschine-Kommunikations-Szenario*. Unpublished master's thesis, Faculty of Technology, Bielefeld University.
- Poppe, R., Truong, K. P., Reidsma, D., & Heylen, D. (2010). Backchannel strategies for artificial listeners. In *Proceedings of the intelligent virtual agents conference*.
- Tuite, K. (1993). The production of gesture. *Semiotica*, 93, 83-106.
- Wachsmuth, I. (1999). Communicative rhythms in gesture and speech. In *Proceedings of the international gesture workshop on gesture-based communication in human-computer interaction*.
- Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review*, 12, 957-968.

# An Empirical Study on the Mechanisms of Creativity in Visual Arts

Bipin Indurkha (bipin@agh.edu.pl)

Institute of Computer Science, AGH University of Science and Technology, Cracow, Poland  
Cognitive Science Lab, International Institute of Information Technology, Hyderabad, India

Shinji Ogawa (perfectworld@nyc.odn.ne.jp)

## Abstract

This collaborative research between a visual artist and a cognitive scientist is based on the assumption that the so-called *aha* moment actually emerges from a number of interacting micro-processes. The empirical study presented here focuses on the creative process involved in connecting two pictures by painting another picture in the middle. This technique was involved in four *Infinite Landscape* workshops conducted at Art Museums in Japan and Europe over the last five years. Based on the artist's verbal recollection of the ideas that occurred to him as he drew each of the connecting pictures, we identify the micro-processes and cognitive mechanisms underlying these ideas, and discuss their implications for modeling creativity.

**Keywords:** Creativity; emergence; perceptual features; similarity; surface features; visual art.

## Introduction

A central problem in creativity research is how new ideas are generated. In recent years, it is gradually being realized that creativity is an emergent property of many interacting micro-processes (Dunbar 1997; Sawyer 2006). These micro-processes can occur within a cognitive agent itself, or in different agents within a group or society. Our larger goal in this research is to study and model these micro-processes.

In particular, we are focusing on the creative processes in visual art. For this, one could consider the creative insights spanning over the entire career of an artist (for example, Dali 1993); or over a part of the career of an artist (for example, Okada *et al.* 2009); or across several artists (for example, Mace & Ward 2002). When a longer period is covered, it is difficult to get information about the micro-processes involved in the creation. Even when one focuses on the creation of a particular work, if the goal is too open-ended, the micro-processes are too unrestrained and divergent. For example, in the study of Mace & Ward (2002), twenty-five artists were interviewed to get data about their creative processes. But because the artists could create any work they wanted, the insights from their self-reflection are only useful for a macro-level model.

When a work is created under constraints, it often increases the level of creativity required (Stokes 2005); it also makes it easier to compare data across different works because they were created under the same constraint. With this in mind, we focused on the task of creating a picture to connect two given pictures seamlessly, as described below.

## Background: *Infinite Landscape* Workshops

This research is a collaborative effort between a visual artist [henceforth referred to as *the Artist*] and a cognitive scientist. Over the last five years, the Artist conducted four workshops at art museums in Japan and in Europe with the common theme *Connecting different spaces*. In each workshop, there were 15-19 participants, all children (8-14 years) except in one workshop there were six adults. Three workshops conducted in Japan followed the following *modus operandi*.

In the first step, the children were shown about 20 photographs of scenery from around the world, and then they were asked to draw imaginary landscapes using the building, people, animals etc. in these pictures as they liked. In the second step, the Artist brought the children's imaginary landscapes to his studio, and then he drew one picture to be inserted between every two pictures of children, so that all three pictures form a seamless scene. One such trio of pictures is shown in Fig. 1: scenes 9 and 10 were drawn by participants, and the Artist drew S9 to connect the two.

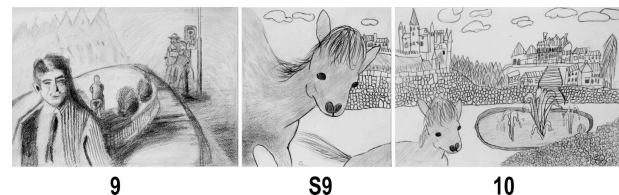


Figure 1

In the third and final step, all the pictures were connected in a ring without a beginning and an end, and the completed ring was suspended from the ceiling of the museum where the workshop was held. The ring was placed with the paintings on the inner side, so that the viewer is surrounded by the work while viewing it.

The fourth workshop conducted in Kraków was similar except for two differences. One is that the children were not shown any photographs in the first step, but half the group was asked to draw Kraków as they imagined it in the past; and the other half the future of Kraków, all based on their imagination. This was only suggested to them and the participants drew whatever they wished. The other difference was that in the final step, the completed ring was placed on a glass floor.

## Methodology

Our overall methodology for this research project is as follows. In the first step, the Artist recorded various ideas that occurred to him as he drew each of the connecting pictures. It should be emphasized that in this step, the Artist was not aware of any potential hypotheses as to what we might be looking for in this data. In the second step, we analyzed these self-reflections to identify various micro-processes and their interactions with each other that were instrumental in the creation of the macro-level connecting pictures. In the third step, we posit cognitive mechanisms underlying these micro-processes. Finally, we plan to model these mechanisms in a computational system.

The research presented in this paper focuses on Steps 2 and 3. From the self-reflection data collected about each of the four workshops, we identified instances where a new idea was generated that became a major theme in the finished picture. This identification itself is also based on the self-reflection data. In other words, we are relying on the Artist's own judgment of the novelty factor. We should emphasize here that because of the nature of the task, namely to connect the two pictures seamlessly, there were many cases where the Artist copied elements, extended texture, color or shape from one of the pictures to the middle picture, and so on. Though we have included such micro-processes in our complete analysis, they are not discussed here.

## Mechanisms of Creativity

We present here several examples of the Artist's thought processes as he sought to connect the given two pictures seamlessly. The Artist's original comments were in Japanese, and are translated here with minor editing by the other author of this paper. We have also labeled and categorized these examples based on the factor that played a key role in the overall theme and the composition of the connecting picture.

### Surface Similarity

In several instances, similarity with respect to color, shape or texture played a key role in the genesis of the connecting picture, and in such a way that a semantic construct was created. This is illustrated by the following examples.

**Similarity in shading or texture:** Consider the Artist's observations concerning Fig. 1: "These two had completely different atmosphere from each other. Sketch 9, drawn by an adult participant, is a scene set at dusk; a person looking at the artist is drawn wearing a sad expression. Sketch 10 has a bright atmosphere with flowers, fountains, buildings on a hill, and a horse. Moreover, each picture had an important character in the bottom left. The idea for connecting these sketches came to me while looking at the wonderful horse in 10. I thought of putting a parent horse running nearby. Because the background color of 9 and the body color of the horse in 10 was the same, I transformed the background of 9

into the parent horse in S9, which became a nested image structure. Then I extended the baby horse and the hill with the buildings."

Here the same shading for the horse's body in 10 and the background in 9 led to the idea that the background in 9 can be morphed into the mother horse in S9, which results in an Escher-like nesting of pictures. The same phenomenon is also seen in Fig. 2: "There was the ground and the sky in the left one-third of 11, but the sea covered the remaining part on the right. In 12, a vast meadow was drawn with rich pictorial details. Here my attention was drawn to the connection between the color of the giant bridge in 11 and the color of the sky in 12. In S11 I drew the enlarged bridge of 11 and connected it with the picture on 12, which resulted in a nested image structure."

These two examples show how texture or shading triggered an association that led to nested image structures.

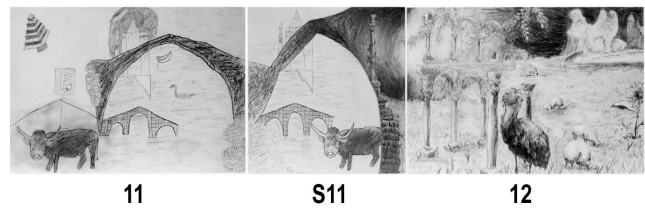


Figure 2

**Similarity in shape:** In Fig. 3, it is the shape of the curves that started a chain of thought: "I felt these two could not be connected with the techniques I had used so far. Then I noticed the wall on the top-right corner of 12 and the curved ledge surrounding the fountain in 13. Using these two curves, I drew a large Mobius strip in S12. As this Mobius strip divided S12 into four sections, in each section I extended the adjacent scenery. It felt like pouring in the scenery. Accordingly, I was able to connect them without blending, and this became the first work with this technique."

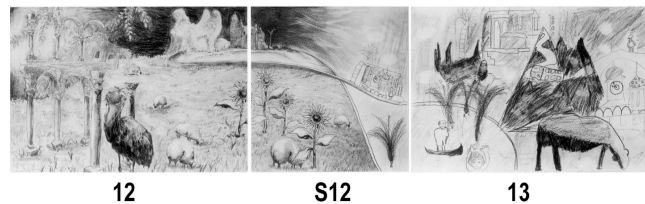


Figure 3

**Surface features trigger a new concept:** In Fig. 4, surface features of an object drawn by the participant reminded the Artist of a completely different object, and that became the theme of the connecting picture: "Suddenly my attention was caught by the strange-shaped, cage-like object drawn at the corner fence in the right pencil-sketch. I thought this shape was a piano. Once I could overlap these images, the

line of fence naturally transformed into a musical staff, and I could draw the dragon playing the piano. I made the particles of light in the sky of the left picture as if they are the sound emanating from the piano.”

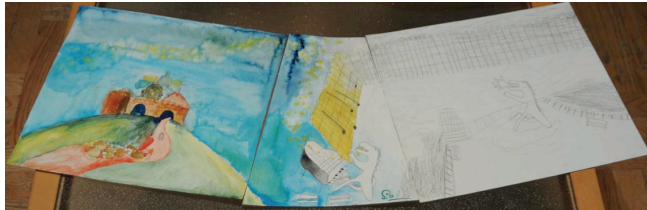


Figure 4

### Contrast

There were also several examples where the contrast or the opposition between the two pictures was instrumental in generating a new idea.

**Contrast in perspective:** In Fig. 5, the contrast between the viewpoints of the pictures was a major factor: “I thought it is not possible to connect 3 and 1. Picture 1 is clearly a bird’s eye-view, as if a bird is looking down towards the ground; in contrast, picture 3 has a distinct horizon with a clear separation between the earth and the sky. First I extended the broken train track and the tire tunnel. Then as I was drawing the dark blue river, I thought, ‘But where should I extend this river? Towards the top? Towards the bottom? If I extend it below, then I can connect it with the ground of 3. But...’ I felt lost. Finally, I resolved to bring the river up. It was a desperate effort. However, at that time I thought of a good way to solve this problem. In the remaining left edge of the picture, I extended the scenery from 3. Finally, to integrate the inconsistent parts of the picture, I floated a number of clouds from 3 on the river. Thus, by using clouds as intermediaries, I was able to connect a bird’s eye-view picture with a perspective picture.”



Figure 5

**Contrast in richness of details:** In Fig. 6, it was the contrast between the richness of details that lead to a very interesting result: “Because 8 was a richly detailed realistic presentation, to contrast it with the presentation in 7, I decided to stress dimensionality in the connection. The

realistic rocks and the bridge in 8 were rendered in 3-d and were connected with the bridge in 7 that was extended in 2-d. To make this connection smoother and give an accent to the picture, I drew 3 Russian onion domes from 7 into S7.”

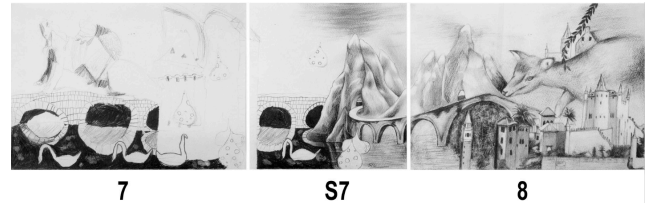


Figure 6

Fig. 7 provides another example: “I thought about how to connect picture 5 with picture 6 that had strong green with dark lines and was rich in details. I thought I could turn the contrast of a picture rich in details and a picture low in details into a pictorial effect. On the top right of 5, there is a game-character like man standing with a trident basking in the sunlight. First, to counter that, I drew a partner woman’s figure on top left of S5. But I laid down the trident by her side. I drew most of S5 as an extension of the dynamics of picture 6. On the bottom right of 5, there is an abstract painting-like area, and I placed this dark touch on the left edge of S5. This was a pleasure to work on. When I saw all of the paintings arranged in a ring at the Art Museum in Okazaki, it was obvious that it was picture 5 that was being the heretic and bringing out the effect of difference in richness of details in paintings. This realization was the most important lesson to me from this case.”



Figure 7

### Semantic similarity/association

There were several examples where similarity at the meaning level played a key role in generating ideas for the middle picture.

**Concept retrieval based on semantic association:** In Fig. 8, the wisp of smoke coming out of the chimney of a house, suggested the idea of a steam engine: “Perhaps my worst betrayal (in a good sense) of the participants is when I changed the brown house of the robot into a steam engine somewhat arbitrarily. The thread of smoke coming out of the chimney made me do this.”





Figure 8

**Similar objects:** Two objects, both trains, but with surface-level dissimilarities played a major role in Fig. 9: “I first noticed the train in 1 and 2. I admired that even though they both had drawn the same train, their drawing styles were very different, and I felt a strong urge to connect the two trains. First I connected the two train tracks that were cutting across 1 and 2, and then drew the gradual transformation of one train into the other. On the bottom left I drew a swan from 1, and on the top right I drew the water fountains and trees from 2, and then connected the backgrounds of the two pictures.”



Figure 9

### Deliberately Ignoring Meaning

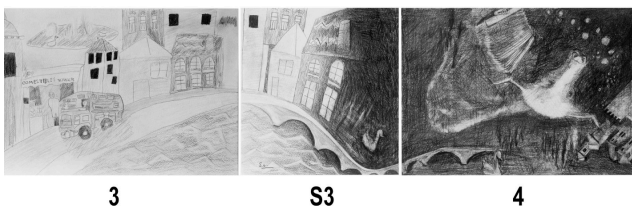


Figure 10

Fig. 10 provides an interesting example where the Artist deliberately chose to ignore the meaning and focused on the surface features only: “At first Sketch 4 was filled-in completely black, and then brightened by eraser. It had no earth and sky, but an ambiguous space from a dark fantasy. Normally, a picture like this cannot be connected with any picture. I decided to connect this dark picture with 3, which had a child-like pictorial space. However, it would be impossible to connect the two in an ordinary way. Here, I decided to ignore all the meanings in these pictures, but instead focus on the pattern of light and dark. I said to myself, ‘it is just a blotch’. The only connecting point in

both pictures was the street in 3 and the bridge on the bottom left of 4. I could connect this street and the bridge. Luckily, bottom left of 4 looks like the sea, and bottom right of 3 also looks like a body of water. In S3, I extended the road in 3 in S-shaped curve and connected it with the bridge in 4. Continuing, I also extended the sea. Until here it was traditional technique. The problem was what to do on top of this. On the left part of S3, the only possibility was to extend the street-side houses on 3, so I did that in the same touch. Then I gradually changed the color of houses from gray to black, while introducing spatial distortion, and changing them from solid to liquid. I floated a swan in the dark pond that the buildings were turned into.”

This example also illustrates the role of surface features (similarity between the shapes of the road and the bridge) and semantic association based on functionality (roads and bridge are both used for travelling.)

### Metaphor

In Fig. 11, an overarching metaphor was generated in trying



Figure 11

to connect a picture with two neighboring pictures. We include a long quote here to familiarize the reader with the context and the thought processes of the Artist: “The child who drew this seemed (and it is my personal impression) emotionally repressed, and who is not accepted as himself by the surroundings. I had decided to not consider the psychological problems of the children, but focus only on the expression of form and color. However, when I was confronted with this work that expressed intense psychological problem, I instinctively thought, ‘What can we do to help this heart?’ But I am well aware that the most I can do is to finish the work and by having the participants view it as an aesthetic experience send them some kind of message. I placed on both sides of 11 two of the brightest pictures, 10 and 12, and started to connect it with them. The sky in 10 is clear, but the top portion is dull and grey. Even more than that, the colors of gloomy 11 seem to entirely reject any possible connection. In S10, losing to this strong feeling of rejection, I connected the pictures rather abruptly and formally. As a result, I gave angel wings from the globe in the middle of 11 to the swan in the red area of S10. Continuing with connecting 11 with 12, on the spur of the moment, I thought of mixing the color of the water surface. In other words, in the middle of the blood red lake of 11, I poured in a stream of bright blue from 12. I was hoping that the same effect would also occur in the heart of the child who drew 11. I felt relieved instinctively when this work was finished. Furthermore, as a cheering party for the child

who drew 11, I added on the top right of S11 the street lamps and the acorn decorations of 12.”

### Influence of Other Ongoing Projects

Fig. 12 illustrates a case when an idea was borrowed from another ongoing project of the Artist: “In S6, I first drew the remaining portion of the cow in 6, and then the remaining portion of the cliffs of Cappadocia. Moreover, to connect the complex-shaped terrain in 7 with the savannah meadow of 6, I extended the grey building on the right edge of 6 and transformed the terrain. Incidentally, at the same time I was working on my ‘Moiré’ series, so I made the shape of the grey buildings like the silhouette of Mt. Saint Michel. This became an example of my incorporating a concept of my own in the scene through my pro-active involvement.”



Figure 12

## Discussion and Related Research

Having gleaned these bits of insights, we now identify four major themes underlying creativity that are highlighted by this study. We comment below on each theme and also discuss previous research related to it.

### Role of Surface Features

It has been widely recognized that similarities play a key role in the generation of new ideas (Kokinov *et al.* 2009; Ward 2011.) Although surface similarities are often found to influence memory access and recall (Barnden & Holyoak 1994), most of the research has focused on semantic aspects of the similarity, like structural alignment, for these are considered to be more helpful in problem solving and learning. In fact, surface similarities are often thought to be distracting (Faries & Sclossberg 1994). Our data, however, indicates that surface features can have a significant influence on creation of new ideas in at least two different ways.

**Surface similarities between two objects:** Here noticing surface similarities between two different objects triggers an exploration for a possible deeper meaningful relation between them, as we saw in many examples above. This is consistent with the results of our earlier studies (Indurkha *et al.* 2008; Ojha & Indurkha 2009), where we found that similarities with respect to color, shape, texture etc. facilitate generation of conceptual associations.

### Surface features of an object recall a different concept:

We saw above how the perceptual features of a cage-like object in Fig. 4 triggered the concept of piano, which became the motif of the connecting picture. This is consistent with a model of perceptual metaphors we had proposed in our earlier work (Indurkha 2006), where we argued that certain metaphors rely on a perceptual resonance between the images corresponding to the source and the target.

### Role of Contrast or Opposition

We found several instances where new ideas or perspective emerged in trying to connect contrasting elements. Many previous studies of creativity have also found that opposition can be a key to generating new insights. For instance, Schön (1963) emphasized that in order to get a new insight about a concept, it needs to be *displaced*, that is, put in the context of other unrelated concepts. Koestler (1964) emphasized that the pattern underlying a creative act is the perception of a situation or an idea in two self-consistent but habitually incompatible frames of reference. More recently, Shapira and Liberman (2009) suggest *psychological distance* as a mechanism for enhancing creativity. They and their colleagues (Jia, Hirt and Carpen 2009) have demonstrated that psychological distance can be induced by such simple devices as taking another person’s perspective or thinking of the problem as if it is unreal.

### Deliberate Deconstruction of Meaning

We presented one example above where the Artist deliberately chose to ignore the meaning, and focused on the perceptual features like shade and texture. This mechanism is also often acknowledged as a useful heuristic for creativity. For example, Gordon’s (1961) *making-the-familiar-strange* is essentially the process of deconstructing the familiar meaning associated with the problem. Similarly, the first step in one of the creativity mechanisms proposed by Rodari (1996) is *estrangement*, where you are asked to see the object as if for the first time, without associating familiar meanings with it.

### Interaction of Top-down and Bottom-up Influences

The metaphor and the moiré series examples (Fig. 11 and 12) illustrate the top-down influences in the creative processes. What we mean here is that the psychological state of the artist and her or his past experiences can also influence the particulars of a creative insight. There have been several accounts of creativity that emphasize the interaction of top-down and bottom-up processes (Fauconnier and Turner 2002; Hofstadter 1995; Indurkha 1997).

## Conclusions and Future Research

We analyzed data from the Artist’s verbal recollection of his thoughts as he drew the middle pictures to connect pairs of pictures seamlessly. From this analysis, we identified a

number of micro-processes that led to the *big picture* idea. In particular, we found that *surface features*, *contrast*, and *meaning deconstruction* play major roles in the generation of new ideas.

There are two lines of research that we are pursuing from here onwards. One is to develop a meme-based approach to formalize these micro-processes, and implement a computational model of them (Ogawa, Indurkha and Byrski 2012). Besides, we are also interested in studying the cognitive processes of the viewers as they look at the trio of pictures. The term creativity is generally restricted to the artist or the person who generates the work, design or the artifact; and one does not attribute it to the reader or the viewer. However, we have argued before that in some situations at least, some creativity is required from the reader or the viewer as well (Indurkha 2007). Moreover, our past research has shown that surface-level perceptual similarities influence how viewers connect pairs of images and relate them conceptually (Ojha and Indurkha 2009; 2012). It would be interesting to see how this process is affected when there is an intervening picture in the middle; and we would like to study the effect of contrast as well. We plan to conduct behavioral and eye-tracking experiments to measure the viewers' response and incorporate those observations in our model.

### Acknowledgments

We are grateful to the Meguro Museum of Art, Tokyo, the National Museum of Art, Osaka, Okazaki Mindscape Museum, Aichi, and Museum of Contemporary Art, Kraków for hosting these workshops. We are also thankful to all the participants for their efforts and contributions.

### References

- Barnden, J.A., & Holyoak, K.J. (Eds.) (1994). *Analogy, Metaphor, and Reminding*, Intellect Books.
- Dali, S. (1993). *The Secret Life of Salvador Dali*. New York: Dover.
- Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T.B. Ward, S.M. Smith and J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes*. Washington, DC: American Psychological Association.
- Faries, J.M., & Schlossberg, K.R. (1994). The effect of similarity on memory for prior problems, *Proceedings of the 16<sup>th</sup> annual conference of the Cognitive Science Society*, 278 – 282.
- Fauconnier, G., & Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.
- Gordon, W.J.J. (1961). *Synectics: The Development of Creative Capacity*. New York: Harper & Row.
- Hofstadter, D. (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. New York: Basic Books.
- Indurkha, B. (1997). On Modeling Creativity in Legal Reasoning. *Proceedings of the Sixth International Conference on AI and Law*, 180-189. Melbourne, Australia.
- Indurkha, B. (2006). Emergent representations, interaction theory, and the cognitive force of metaphor. *New Ideas in Psychology*, 24(2), 133–162.
- Indurkha, B. (2007). Creativity in Interpreting Poetic Metaphors. In T. Kusumi (Ed.) *New Directions in Metaphor Research*, Tokyo: Hitsuji Shobo, Tokyo.
- Indurkha, B. (2010) On the role of metaphor in creative cognition. *Proceedings of the International Conference on Computational Creativity: ICC3-X*, Lisbon, Portugal.
- Indurkha, B., Kattalay, K., Ojha, A., & Tandon, P. (2008). Experiments with a creativity-support system based on perceptual similarity. In H. Fujita and I. Zulkernan (Eds.) *New Trends in Software Methodologies, Tools and Techniques*, Amsterdam: IOS Press.
- Jia, L., Hirt, E.R., & Karpen, S.C. (2009). Lessons from a Faraway land: The effect of spatial distance on creative cognition. *Journal of Experimental Social Psychology* 45, 1127–1131.
- Koestler, A. (1964). *The Act of Creation*. London: Hutchinsons.
- Kokinov, B., Holyoak, K. & Gentner, D. (Eds.) (2009). *New Frontiers in Analogy Research*, Sofia: New Bulgarian University Press.
- Mace, M., & Ward, T. (2002). Modeling the creative process: A grounded theory analysis of creativity in the domain of art making. *Creativity Research Journal* 14, 179–192.
- Ogawa, S., Indurkha, B., & Byrski, A. (2012). A meme-based architecture for modeling creativity. *Proc. of the Int. Conf. on Computational Creativity*, Dublin, Ireland.
- Ojha, A., & Indurkha, B. (2009). Perceptual vs. conceptual similarities and creation of new features in visual metaphor. In B. Kokinov, K. Holyoak & D. Gentner (Eds.) *New Frontiers in Analogy Research*, Sofia: New Bulgarian University Press.
- Ojha, A., & Indurkha, B. (2012). On the role of perceptual features in metaphor. To appear in F. Ervas & E. Gola (Eds.) *Metaphor and Communication*.
- Okada, T., Yokochi, S., Ishibashi, K., Ueda, K. (2009). Analogical Modification in the Creation of Contemporary Art. *Cognitive Systems Research* 10 (3), 189–203.
- Rodari, G. (1996). *The Grammar of Fantasy* (J. Zipes, Trans.) New York: Teachers & Writers Collaborative.
- Sawyer, K. (2006). *Explaining Creativity*. Oxford (UK): Oxford University Press.
- Schön, D.A. (1963). *Displacement of Concepts*. New York: Humanities Press.
- Shapira, O., & Liberman N. (2009). An Easy Way to Increase Creativity. *Scientific American (Mind Matters)*, July 21, 2009.
- Stokes, P.D. (2005). *Creativity from Constraints: The Psychology of Breakthrough*. Berlin: Springer.
- Ward, T.B. (2011). Analogies. In M.A. Runco & S.R. Pritzker (Eds.) *Encyclopedia of Creativity*, (2<sup>nd</sup> ed.), New York: Academic Press.



# Emergence of control in artistic expressions and the process of expertise

Chiaki Ishiguro (qq116201@iii.u-tokyo.ac.jp)

Graduate School of Interdisciplinary Information Studies, University of Tokyo  
Tokyo, 113-0033, Japan

Takeshi Okada (okadatak@p.u-tokyo.ac.jp)

Graduate School of Education &  
Interfaculty Initiative in Information Studies, University of Tokyo  
Tokyo, 113-0033, Japan

## Abstract

The creation of a work of art has been indicated to result from 'expressive awareness', achieved as the artist matches images and methods. This study examines how novices, who tend to express reproductively, acquire such expressive awareness over several weeks of practice of photography. We conducted case studies with two conditions: 1) one participant reflected only her own creative activities, and 2) one participant imitated eminent works of creative expression in the domain. As a result, the participants acquired expressive awareness in both conditions, though the contents of the expressive awareness were different. The imitation participant started to practice creative expressions and tried to control her creation consciously, while the reflection participant started to focus on precision of methods of expression. The findings of this study are useful for developing educational practice in art schools.

**Keywords:** artistic creativity; expertise; imitation; reflection; artistic expression

## Introduction

Artistic creation has been one of the most significant activities of human beings. Recent psychological studies have focused on the process of artistic creation. Such studies provide useful insights for understanding creative cognition and have implications for creativity education. Previous studies on the cognitive processes of artistic creation have indicated that artistic creation consists of processes for generating ideas or concepts, and processes for externalizing them into artwork (e.g., Mace & Ward, 2002; Yokochi & Okada, 2005). These studies have also suggested that coordination of these two processes is important. In other words, when creating artworks, artists pay attention to whether or not their artwork matches with their art concepts, and whether or not the strategies that they choose are effective in actualizing their ideas as a form of art. The process of coordinating their intentions and actions to achieve artistic expressions is a type of monitoring process, i.e. a metacognitive process (Flavell, 1976). Though it is known that these processes feature in artistic creation by experts, they have rarely been seen in creation by novices (Fayena-Tawil, Kozbelt & Sitaras; 2011).

This sense of matching of images (hereafter *expression contents*) and methods (hereafter *expression methods*) to externalize them, referred to in this paper as *expressive awareness*, plays an important role in artists' creation (c.f. Gantner, 1979). However, no empirical research has been done to examine how novices acquire this expressive

awareness in the process of achieving artistic expertise. This study focuses on such an acquisition process of expressive awareness as an initial form of monitoring ones' own creative process. The findings from this study offer new insight for the development of education programs for art schools.

When we examine the process of artistic expertise, we first need to explain what artistic expression is. Before the modern era, expression meant giving a plausible impression of motifs relating to religion or history (Diderot, 1980). In contrast, after the modern era, the concept of artistic expression came to mean turning the creators' experiences, emotions and subconscious experiences into an entity with reality and impact (Croce, 1902; 1990). Nowadays both views of expression exist in society. This diversity of views of expression may affect the process of artistic expression itself. Therefore, in this study, we classify artistic expressions as *reproductive expression*, the contents of which are intended to represent real entities such as landscapes or still lifes, and *creative expression*, the contents of which are intended to represent the creators' ideas or emotions.

Novices who have no knowledge of artistic creation generally prefer realistic works (Cupchik & Gebotys, 1988; Kozbelt, 2006). Also, Ishibashi & Okada (2009; 2010) reported that novice subjects drew realistic drawings when they were asked to draw original works with natural materials as motifs. According to these findings, art novices tend to appreciate and create reproductive expressions.

What are the important factors affecting novices' acquisition of expressive awareness? One of the candidate factors is continuing participation in expressive activities. People use *self-explanation* (Chi et al., 1989) and *reflection* (Schön, 1983) on their own artwork and on their process of creation during such expressive activities. These processes can lead to the acquisition of expressive awareness. Therefore, it is assumed that novices are able to acquire expressive awareness if they have continuing opportunities to create artworks and reflect on their own creations.

However, in order to become expert in a domain, such an action-reflection cycle may not be enough. Csikszentmihalyi (1999) suggested that creativity is dependent not only on the creators' activities, but also relates to domain rules, representations and methods in the field of expertise. Hence, it might be assumed that profound encounters with existing artwork in an artistic domain play an extremely significant role in the acquisition of expressive awareness in addition to

reflection on the artist's own creations. In fact, it has for a long time been considered that copying masterpieces (i.e. a profound type of encounter with masterpieces) is a very efficient way to learn drawing techniques and the painters' intentions represented in masterpieces.

Ishibashi & Okada (2009; 2010) empirically examined the effects of imitation on artistic creation. They conducted psychological experiments that entailed novices copying unfamiliar abstract drawings. While copying, they speculated on the intentions and processes behind the drawings and acquired new perspectives for drawing. As a result, using new representations, participants who copied abstract drawings drew more creative pictures than those who did not. Though their study produced pioneering work that empirically examined the effect of a profound encounter with the artwork of others on creative drawing, they did not focus on the issue of expressive awareness described above.

Therefore, we decided to conduct empirical case studies focusing on the question of how profound experiences (such as imitation) of existing artworks affect novices' acquisition of expressive awareness. To elucidate this question rigorously, it would have been better for us to conduct experiments with a greater number of participants, as done by Ishibashi & Okada (2010). However, conducting experiments with many participants to observe such a long-term cognitive change would be extremely time-consuming. Finding at least twenty participants who were willing to spend seven weeks on this experiment was practically impossible for us. Since our main goal is to investigate how expressive awareness is acquired during a fairly long-term process of the mastery of art, we decided to conduct exploratory case studies analyzing the data in detail from various aspects. Such a case study method with a long-term span has been used to examine the process of acquisition of knowledge or strategies in cognitive psychology (e.g., Siegler & Jenkins, 1989).

We used artistic photography as the target domain for research to answer our question, because photography is one of the most familiar genres of artistic expression in our ordinary lives. Photography also has a very distinctive feature that people can take photographs in a very short time span; a photo is usually taken in less than a second. These features of photography enabled us to study the early process of artistic expertise easily, because novices have a low barrier to participation in this artistic activity.

It is important to note that photographs can express the creator's intentions or ideas even if the photographs realistically represent the motifs. For example, a famous photographer, Henri Cartier-Bresson introduced the concept of '*The Decisive Moment*' by creating artwork matching with a theme. Hence, it is possible for us to understand photographers' intentions deeply through viewing their photographs.

Although novices in photography may usually tend to appreciate or create reproductive expressions, they acquire expressive awareness through the process of repeated

practice of photography. Specifically, novices gradually acquire creative expressive awareness and try to control relationships between expression contents and methods consciously if they imitate photographs with creative expression. In contrast, they acquire reproductive expressive awareness if they only repeat photo taking and reflection. They pay more attention to expression methods to express precisely the reproductive contents, because they have less necessity to develop expression contents than in the case of the creative expression style. These considerations lead to the following two hypotheses. The first hypothesis: Novices in the domain of artistic photography acquire expressive awareness, through just repeating photo taking and reflection or by imitating eminent artistic photographs of creative expression. However, the former acquire reproductive expressive awareness and the latter acquire creative expressive awareness. The second hypothesis: Novices with creative expressive awareness try to control relationships between expression contents and methods consciously in the following creations. In contrast, those with reproductive expressive awareness emphasize the importance of expression methods.

## Method

In this study, we conducted two case studies that examine the changes in artistic creation over a period of several weeks. Each case study included four tests [a pre-test and post-tests 1-3] and three interventions [interventions 1-3]. The tests and interventions were repeated alternately.

**Participants and the period of observation** Two female students at the University of Tokyo volunteered for this study. One participant was 23 years old and an undergraduate student in science, and was assigned to the imitation condition. The other participant was 24 years old and a graduate student in pharmacology, and was assigned to the reflection condition. The length of the case study period of the former was 86 days from 28 October 2010 to 21 January 2011, and that of the latter was 46 days from 30 September to 14 November 2011. The reason why the participant in the imitation condition spent a longer time than the other one was because the Christmas holidays fell within the case study period and she stopped her photography for a while.

Both of the participants showed an interest in artistic photography before participation in this study, and the imitation participant visited an exhibition including photographic works just before participating in the study. However, they had never had any professional training in artistic photography. They took photographs mainly when they travelled or attended special events, and were not in the habit of visiting photo galleries or exhibitions to see collections of photographs.

Also they had never used a single-lens reflex camera. Therefore, the first author taught them the basic method of use of a digital single-lens reflex camera, such as the mechanism of the camera and how to use the diaphragm,

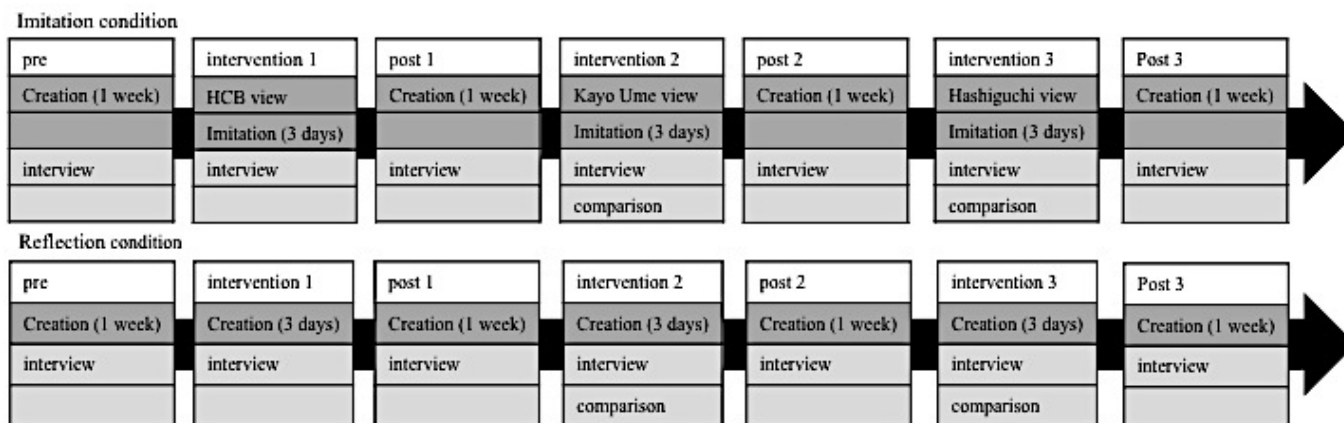


Figure 1 Procedure of case studies in each condition

shutter speed and exposure. They then practiced taking photographs with the first author.

**Procedure** The participants were required to take at least 40 artistic photographs per week during their own free time in each of the four test sessions: the pre-test and the post-tests 1-3. The three interventions were different in the two conditions. In each intervention session of the imitation condition, the participant first saw a collection of artworks by a proficient photographer, selected the one photograph that she liked the most, and then was given the following instructions, “Think about what the photographer pays attention to when taking this photograph and try to take the same style of pictures yourself.” The participant in the reflection condition continued to create artistic photographs of her own in the intervention sessions. Both of them took at least 20 photos in each intervention session (see Figure 1).

All of the photographs were taken with a digital single-lens reflex camera (Canon EOS KissX3 / EF-S18-55mm F3.5-5.6IS). The participants were directed to use only three modes of exposure that enable the user to control expression to a large extent: Manual Mode, Shutter Priority AE and Aperture Priority AE. They were not allowed to use Program Mode, since not much room is left for users to control the photographic expression if they use such an automated mode. All of the photographs taken were preserved for analyses.

After each session in the two conditions, the first author interviewed the participants. The interviews for the imitation condition took about 1 hour, and for the reflection condition took about half an hour. They were recorded by IC recorders and a video camera. In the interviews, the participants selected the ten best photographs in the session after each test, and the five best photographs in the session after each intervention. If there were any other photographs that the participants selected, they also reported on these. In addition, we interviewed them about their views on photography in each session. The participants were directed to reflect on their photographs in the session by comparing them with photographs taken in previous sessions.

**Imitation tasks** The photographs used for the imitation tasks were of styles different from those about which the

imitation participant had prior knowledge, as indicated in the interview after the pre-test. This was because it has been shown by Ishibashi & Okada (2009; 2010) that the copying of unfamiliar pictures stimulates the production of creative artwork. In order to identify styles of expression unfamiliar to the participant, the first author interviewed her about her favourite photographs, favourite photographers, and her own photographs taken in the pre-test. Her favourite photographer was Mika Ninagawa and her favourite photograph was one from ‘SHADOWS: Works from the National Museums of Art’. As their characteristics, she mentioned ‘colours’ and ‘shadows’. In addition, there were no photographs in which people were the main motif in pre-test sessions in either of the conditions. Therefore, photographs that depicted people as the main motifs and were not characterized by ‘colours’ or ‘shadows’ were selected for the imitation sessions.

**Data** The main data are protocols in interviews.

First, we checked in the interview data whether or not the imitation participant actually practiced imitation. Second, in order to examine the first and second hypotheses, we checked whether expressive awareness was observed and if so, how this awareness developed in the answers to the questions about the participants' views of photography and the interviews about each photograph taken by them.

**Expressive awareness and the process of expertise** To examine the first hypothesis, with the interview protocols on views of photography and of each photograph taken by the participants, we checked whether expressive awareness was acquired by the participants.

In the verification of the second hypothesis, we also checked whether the features of the process of expertise appear in the interview protocols on the views of photography and of each of the photographs taken by the participants. In addition, we also investigated the changes in expression methods, expression contents, and matching of them in chronological order in each condition.

**The changes in the participants' views of photography** The interviewer asked the participants, ‘What do you think is good photography?’ and ‘What should we do to take good photographs?’, in order to investigate their views on

photography in each session. In this protocol, we checked whether they mentioned ‘Expression contents’, ‘Expression methods’ and ‘Matching between the expression contents and methods’. ‘Expression contents’ are further divided into ‘Reproductive expression’ and ‘Creative expression’.

In this analysis, ‘Matching of expression contents and methods’ indicates expressive awareness. If the ‘Expression contents’ were ‘Reproductive expression contents’, we regarded the expressive awareness as ‘*Reproductive expressive awareness*’. Also, if the ‘Expression contents’ were ‘*Creative expression contents*’, we regarded the expressive awareness as ‘*Creative expressive awareness*’.

**The changes in expressive awareness in the interview protocols for each photograph** The data are interview protocols about the 10 best photographs of their choice (in total 40 photographs) taken by the participant in pre-test and post-tests 1, 2, 3, in both conditions. We identified the interview protocols about matching of expression contents and methods as expressive awareness. Then we examined how each protocol changed in chronological order in each condition.

## Results

We first checked whether imitation was actually practiced. The results showed that in the imitation participant’s reflection on her own photographs taken in the interventions, characteristics of imitation tasks appeared. We do not explain this result in detail in this paper due to limitations of space. (The detailed results are reported in Ishiguro, Okada & Ishibashi, 2011.)

For testing the first and second hypotheses, we checked interview protocols about views of photography and about

reflection on each photograph. As a result, statements about expressive awareness appeared after intervention 1 in both conditions. The reflection participant mentioned reproductive expressive awareness and the imitation participant mentioned creative expressive awareness (supporting the first hypothesis). However, the reflection participant occasionally stopped mentioning this in the interview on her view of photography, and statements about the matching of expression contents and methods were less frequent in her reflection on each photograph than that of the imitation participant.

The process of expertise of expression differed in the two conditions. The imitation participant always paid attention to expression contents, methods and matching of them after the acquisition of expressive awareness. In contrast, the reflection participant sometimes reported no matching of expression contents and methods after the acquisition of expressive awareness, and emphasized the importance of expression methods (supporting the second hypothesis). The specific results are the following.

**The views of photography** Both the participants mentioned expressive awareness after intervention 1 (see Table 1, 2). Additionally, the imitation participant, who experienced imitation of multiple varying artworks, showed a stronger interest in creative expression contents and mentioned the matching of expression contents and methods after intervention 2. In contrast, the reflection participant, who repeated only photo taking and reflection, emphasized reproductive expression contents from the pre-test, and acquired reproductive expressive awareness in intervention 1. However, she had a tendency to pay attention to precision of expression methods after post-test 2.

Table 1 The view of photography in the imitation condition

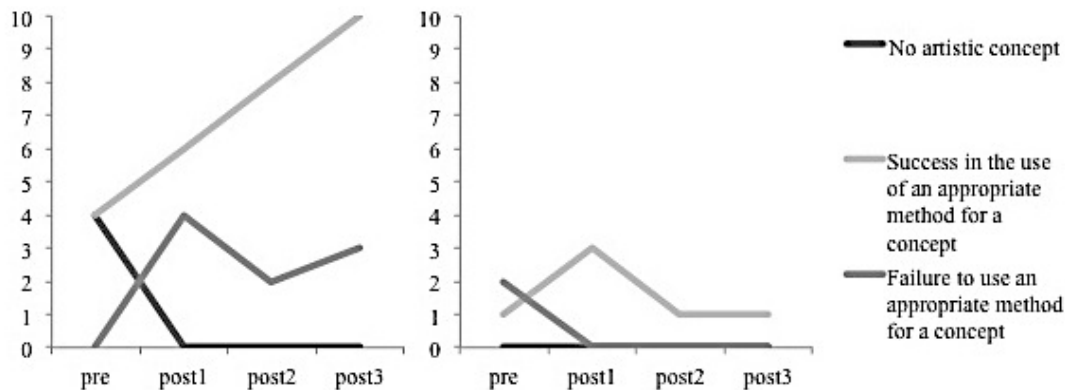
The imitation participant		pre	intervention1	post1	intervention2	post2	interventions3	post3
Expression contents	Reproductive contents	○						
	Creative contents		○	○	○	○	○	○
Expression methods			○	○	○	○	○	○
Matching of expression contents and methods (Expression awareness)	Reproductive expression awareness							
	Creative expression awareness					○	○	○

Note. ○ open circles mean that participants mentioned about each item

Table 2 The view of photography in the reflection condition

The reflection participant		pre	intervention1	post1	intervention2	post2	interventions3	post3
Expression contents	Reproductive contents	○	○	○	○			○
	Creative contents							
Expression methods			○	○	○	○	○	○
Matching of expression contents and methods (Expression awareness)	Reproductive expression awareness	○	○	○				○
	Creative expression awareness							

Note. ○ open circles mean that participants mentioned about each item



NB: The two graphs show the number of photographs mentioned in connection with each of the items on the vertical axis and sessions on the horizontal axis.

Figure 2 Expression awareness in the imitation condition

Figure 3 Expression awareness in the reflection condition

**Protocols for each photograph** We examined the interview protocols for each photograph from the aspect of ‘matching of expression contents and methods’ in order to check whether or not expressive awareness was actually utilized in each photograph (see Figure 2, 3). The results indicated that ‘matching of expression contents and methods’ was focused on more strongly in the imitation condition than in the reflection condition. The imitation participant seemed to be interested in creative expression because from the pre-test to post-test 1, she mentioned ‘No artistic concept’, which is defined as a statement indicating a lack of the concepts necessary for artistic photographs. However, she increasingly mentioned ‘Failure to use an appropriate method for a concept’, which means failure to take photos using appropriate methods despite having a certain expression content, and ‘Success in the use of an appropriate method for a concept’, which means success in taking photos using appropriate methods for a certain expression content. These results imply that the imitation participant had creative expressive awareness when taking each photograph, and such awareness became higher in the subsequent sessions. By contrast, in the reflection condition, statements about ‘Success in the use of an appropriate method for a concept’ increased in post-test 1, in which expressive awareness appeared for the first time in interview protocols about the views of photography. However, the number of the statements about the matching of expression contents and methods in this condition was less than in the imitation condition. Also, it was implied that she had less interest in creative expression contents, because there were no statements about ‘No artistic concept’.

## Discussion

Through two case studies, we have examined how novices acquire expressive awareness that controls expression contents and methods in a creation process when they continue to participate in artistic expression for a number of weeks. We have also investigated whether there are differences in expressive awareness and the process of

expertise between those who employ reproductive expression, repeating photo taking and reflection, and those who practice photography based on imitations of eminent works of creative expression.

Expressive awareness was acquired in both conditions after interventions. Creative expressive awareness was acquired in the condition of imitating eminent artworks, and reproductive expressive awareness was acquired in the condition of reflection (supporting the first hypothesis). Additionally, the result that in the reflection condition there were fewer statements about reproductive expressive awareness indicates that reproductive expression was not as consciously achieved as creative expression.

The two conditions differed in the process of expertise. The imitation participant paid attention to the matching between expression contents and methods after the acquisition of expressive awareness. By contrast, the reflection participant paid attention to expression methods very precisely (supporting the second hypothesis).

Given these results, it is suggested that novices are able to acquire reproductive expressive awareness by only repeating creation and reflection. However, through imitating eminent works of creative expression, their expression style changes to creative expressions and they become conscious of the matching of creative expression contents and methods.

In this study, we provide new findings in the study of expertise as follows. Previous studies of expertise have indicated that experts have structured domain specific knowledge and utilize their metacognition (Glaser & Chi, 1988), such as self-monitoring skills (Ericsson et al, 1993) or reflection in action (Schön, 1983) in various expert practices. Such knowledge and cognitive processes are acquired in the long-term processes of expertise.

This study defines expressive awareness as indicating the emergence of control of ones’ own creative process, and examines how this occurs. The results indicate that expressive awareness is acquired after continuous participation in creation. However, it is important not only

to repeat creation and reflection but also to have profound encounters (like imitation) with eminent creative expression works in the domain so that creators can utilize these actively. These findings support the claim that creators have to learn domain specific knowledge in order to create original works (Csikszentmihalyi, 1999).

**Limitations of this study and future work** We were not able to conduct experiments with a large number of participants. Therefore, we have to be careful in generalizing our findings. Also, for various practical reasons, this study failed to control differences in the academic backgrounds and photographic experience of the two participants. Therefore, the possibility that differences in critical thinking or motivation for photo taking between them affected the results of this study still remains. However, each of them was intelligent enough to secure a place at a highly prestigious university in Japan, and they both told us that they were highly motivated to take pictures. Future studies are required to exclude such confounding variables more carefully, by increasing the sample size of both the imitation and reflection condition and controlling variables that might affect differences in the conditions.

## References

- Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-Explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145-182.
- Glaser, R., & Chi, M. T.H.(1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr, (Eds.) *The nature of expertise*.(pp. xv-xxvii). Hillsdale, NJ: Erlbaum.
- Croce, B. (1909). *Estetica come scienza dell'espressione e linguistica generale*. [The aesthetic as science of expression and of general linguistic] (D, Ainslie. Trans.). Milan-Palermo-Naples: Sandron, (Original work published 1902)
- Csikszentmihalyi, M. (1999). Implications of a systems perspective for the study of creativity. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 313-335). New York: Cambridge University Press.
- Cupchik, G. C., & Gebotys, R. J. (1988). The search for meaning in art: Interpretive styles and judgments of quality. *Visual Arts Research*, 14, 38-50.
- Diderot, D. (2005). *Essais sur la peinture*. [The theory of drawings] (K, Sasaki. Trans.). In D. Diderot, & G. May (Eds.), *Œuvres complètes*, vol. 14, Hermann, (Original work published 1980)
- Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100, 363-406.
- Fayena-Tawil, F., Kozbelt, A., & Sitaras, L. (2011). Think global, Act local: A protocol analysis comparison of artists' and nonartists' cognitions, metacognitions, and evaluations while drawing. *Psychology of Aesthetic*
- Gantner, J. (1983). *Das Bild des Herzens über Vollendung und Un-Vollendung in der Kunst* [Images in minds: problems about completion and incompletion of artistic works] (N. Nakamura, Trans.). Gebr. Mann Verlag, Berlin. (Original work published 1979)
- Ishibashi, K., & Okada, T. (2006). Exploring the effect of copying incomprehensible exemplars on creative drawings. In R. Sun (Ed.) *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1545-1550), Mahwah, NJ: Lawrence Erlbaum Associates.
- Ishibashi, K., & Okada, T. (2009). Tasyasakuin ha souzou wo dou insupaia suruka: sakuin no shinkinsei oyobi kanyo houhou no kouka [Effects of contact with works of art by others on artistic creation] *Proceedings of the 26th Annual Meeting of the Japanese Cognitive Science Society*, 48-51.(in Japanese)
- Ishibashi, K., & Okada, T. (2010). Tasyasakuin no mosya niyoru byogasouzou no sokushin [Copying of other's artistic drawing promotes drawing creativity]. *Cognitive Studies*, 17(1), 196-223.
- Ishiguro, C., Okada, T., & Ishibashi, K. (2011). Influence on beginners' photography of imitation of artistic photographs by others. *Proceedings of the 28th Annual Meeting of the Japanese Cognitive Science Society* (pp. 213-222) (in Japanese).
- Kozbelt, A. (2006). Dynamic evaluation of Matisse's 1935 *Large Reclining Nude*. *Empirical Studies of the Arts*, 24, 119-137.
- Mace, M., & Ward, T. (2002). Modeling the creative process: A ground theory analysis of creativity in the domain of art making. *Creativity Research Journal*, 14, 179-192.
- Siegler, R. S., & Jenkins, E. A. (1989). *How children discover new strategies*. Hillsdale, NJ: Erlbaum.
- Schön, D. A. (1983). *The reflective practitioner*. New York: Basic Books.
- Yokochi, S., & Okada, T. (2005). Creative cognitive process of art making: A field study of a traditional Chinese ink painter. *Creativity Research Journal*, 17, 241-2

# Changes in Cognitive Processes upon Learning Mini-Shogi

**Takeshi Ito (ito@cs.uec.ac.jp)**

Department of Communication Engineering and Informatics  
The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo, JAPAN

**Daisuke Takeno (takano@minerva.cs.uec.ac.jp)**

Department of Communication Engineering and Informatics  
The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo, 182-8585, JAPAN

**Xiaohong Wan (xhwan@brain.riken.jp)**

Laboratory for Cognitive Brain Mapping  
RIKEN Brain Science Institute, Hirosawa 2-1, Wako\_shi, Saitama, 351-0198, JAPAN

**Keiji Tanaka (keiji@riken.jp)**

Laboratory for Cognitive Brain Mapping  
RIKEN Brain Science Institute, Hirosawa 2-1, Wako\_shi, Saitama, 351-0198, JAPAN

## Abstract

In this research, we investigated cognitive processes while playing Mini-Shogi through fMRI and cognitive experiments. Mini-Shogi is a Japanese chess-like game that uses a small board. In our cognitive experiment, the group of stronger Mini-Shogi players the stronger group's total thinking time shortened and their total number of eye movements decreased. However, our investigation of search depth revealed different results from those of past research. The results of our fMRI experiment revealed that after learning, activity in the caudate nucleus increased among stronger players. The results of our experiments suggested that intuitive ability and the capacity for careful consideration are not independent.

**Keywords:** Expertise; Mini-Shogi; Intuition; fMRI; Eye movement.

## Introduction

Since developments in measurement apparatus have recently made it easier for physiological data to be measured more correctly (and in greater detail), the amount of research comparing cognitive data with physiological data is increasing. One type of such research, research on learning, not only clarifies the mechanism of intelligence but will also be useful for education.

In the field of cognitive science, research examining the difference between beginners and experts using chess and puzzles has been executed. In chess, the most famous cognitive experiments have involved memorizing positions and were performed by de Groot (de Groot, 1965; de Groot & Gobet, 1996). The results of those experiments revealed that experts can memorize more chess pieces of a position quickly than beginners.

As a follow-up to de Groot's work, Chase and Simon introduced the theory of chunking to explain why expert game players perform so well in memory tasks (Chase & Simon, 1973). Chunking is the process of dividing a chess position into smaller parts that have meanings. Chase and

Simon showed that stronger players have bigger chunks of chess knowledge than do weaker players.

Ito and others have observed cognitive differences between experts and beginners in Shogi (Japanese chess) as a follow-up to research on chess (Ito, Reijer and Matsubara, 2004). In that research, they compared the thought processes of beginners, club players, and experts during the next move task. The results are shown in the following table (Table 1). The results suggested that there were not only spatial clusters (spatial chunks) but time clusters (time chunks) involved in the experiments on chess.

Table 1: The results of cognitive research on Shogi

	Beginners	Club players	Experts
Recognition of a position	slow	fast	very fast
Area of eye movements	wide	narrow	very narrow
Thought time	short	long	short
Generation of a candidate move	a few	many	a few
Depth of search	shallow	deep	more deep

This topic has also been researched in the field of brain science. With technological progress, measurement of brain activity came to be accomplished less via PET (positron emission tomography) and more via fMRI (functional magnetic resonance imaging). Paolo and others measured the active parts of the brain by using PET while subjects were solving chess-related problems (Paolo et al., 1994). The frontal lobes, the occipital lobes, and the left premotor area were activated when chess-related problems were considered.

Atherton, Chen, and others measured which parts of the brain were activated by solving problems in chess and Go using fMRI (Atherton et al., 2003; Chen et al., 2003). The



premotor area of the frontal lobes, the parietal lobes, and the occipital lobes are activated when solving both types of problems.

Wan and others have observed a physiological difference between amateur and expert Shogi players (Wan et al., 2011). They observed changes in brain activity by using fMRI. They found new brain activity.

When professional Shogi players saw the Shogi problem during the experiment, selective activity was observed in a part of the caudate nucleus; this activity was not seen in amateurs. On the basis of that peculiarity, we guess that this activity is related to expertise.

Much research using eye trackers has been executed in the field of cognitive science since the development of eye movement tracking devices.

Law and others analyzed the eye movements of beginners and experts in the use of training equipment for laparoscopic surgery (Law et al., 2004). Experts looked at the affected part, but beginners looked at the operating instruments. We can understand this by analyzing experts' eye movements and comparing the difference between experts and beginners. Much past research has separately measured experts and beginners and observed the differences between them.

Few studies have observed the learning process from beginner to expert. Therefore, in this research, we planned an experiment that makes the beginner learn and generates an expert. We also examined whether experts shared the features of expertise seen in past experiments.

In this research, we measured both cognitive and physiological changes. The physiological experiment was carried out at RIKEN. In this report, we explain the cognitive experiment and discuss the results of both the cognitive and physiological experiments.

## Mini-Shogi

### What is Mini-Shogi?

*Shogi* is a Japanese chess-like game. Mini-Shogi is also called 55-Shogi; it is similar to Shogi, but uses a  $5 \times 5$ -square board. Almost all rules are inherited from Shogi. Therefore, Mini-Shogi can be said to be "small-board Shogi."



Figure 1: Initial position of Mini-Shogi

The aim of the game is to checkmate the opponent's king. The initial position of the game is shown in Figure 1. There are six kinds of pieces: *fu* (Pawn), *kin* (Gold), *gin* (Silver), *kaku* (Bishop), *hisha* (Rook), and *ou* (King). Rook and Bishop can be promoted to Dragon and Horse, respectively,

as in Shogi rules, in case these pieces move to the end line of the board. Pawn and Silver can be promoted to Gold. The "drop rule" states that a taken piece can be reused, as in Shogi.

Since Shogi is a complicated game, a long period of time is taken to learn it. However, Mini-Shogi can be learned to a fixed level in a short period. In this research, since subjects needed to learn within a limited period, we found Mini-Shogi to be more suitable than Shogi for efficient acquisition of expertise.

## Experiment

### Subject

The subjects were 2<sup>nd</sup>- or 3<sup>rd</sup>-year students at the University of Electro-Communications, and 26 people became subject candidates. The contents of a questionnaire given to prospective subjects examined their experience with Shogi, Mini-Shogi, and Shogi terms. Six people whose questionnaires reflected too much expertise were excluded. After the experiment started, one person dropped out on his personal reasons. Finally, 19 people became subjects.

### Learning Environment

The learning term of this experiment lasted three months. The cognitive experiments were executed three times: the 1<sup>st</sup> period lasted from the start of the study to one week later, the 2<sup>nd</sup> period lasted from one week later to one and a half months later, and the 3<sup>rd</sup> period began after the end of the learning term. The experiments using fMRI were carried out in the early stage and the last stage of learning at RIKEN.

We gave the subjects the "K55" Mini-Shogi software package during the experimental period. "K55" was developed by Yoshikazu Kakinoki, who is a famous Shogi programmer. K55 was the champion program at the UEC Cup 55-Shogi Championship from 2007 to 2009 and the 2<sup>nd</sup> place program in 2010. Therefore, it is strong enough for beginners, and some of its interfaces are suitable for learning. It can be set to a difficulty level of 1–14, so it can be pitted against opponents of various ability levels. It is easy to operate the software, as positional evaluation by K55 can be displayed numerically, and it has a hint mode that can be used effectively for learning.

Moreover, we created an environment on which human players can play on the Internet. This system is called "55-floodgate." After a player completes 30 games, he/she receives a computed rating; we thought that these ratings would be useful for improving subjects' motivation. We asked subjects to learn about the game as much as possible (and for at least one hour each day). The duty to present a report was imposed upon the subjects once per week. The contents of this report were the following: time spent learning during the week, things noticed, winning rates against K55, etc.

In order to maintain a subject's motivation for learning during the experimental period, a lecture about Mini-Shogi was given to the subjects one month after the start of the

experiment. The contents of the lecture ranged from fundamental to strategic topics, such as the effective usage of each piece, the value of each piece, and advantageous openings. A collection of Mini-Shogi next move problems was distributed to two months after the start of learning, and subjects were made to answer these. Furthermore, a tournament was carried out at the end of the experimental period, and many prizes were awarded on the basis of the results. This tournament was thought to increase subjects' motivation.

## Cognitive Experiment

**Method** We showed the next move problems to subjects and asked them to speak freely regarding their thought processes. Eye movements were recorded simultaneously with verbal data using an eye tracker (QG-PLUS, Ditect, Inc.). We used a liquid crystal display monitor at a resolution of  $1280 \times 1024$  pixels (96 dots per inch). We recorded verbal protocol data according to the traditional cognitive science approach. We instructed the subjects that there were no time limits for thinking. We prepared 11 next move problems devised by an expert Mini-Shogi player. The next move problems shown to subjects used difficult positions from which multiple candidate moves can be considered. The order of presentation was randomized for each subject. The same problems were shown during each period, but the order of presentation was changed. (Incidentally, no subjects realized that the same problems were used.) A time interval was set between each problem, and the subject could decide to take breaks freely.

**Results** We selected six subjects for whom we obtained good verbal protocol and eye movement data in the cognitive experiment.

We executed three experiments: one each in the first period, the second period, and the third period of learning. In order to clarify the differences between subjects' mental processes before and after learning, we decided to compare the data from the first and third periods.

We classified verbal protocol data by categories of contents. We used the same classification system as did previous research, using five categories: "recognition of the position," "generation of the candidate move," "prediction," "evaluation," and "decision" (Ito, Reijer and Matsubara, 2004). On the basis of this analysis, we examined what kind of thinking was performed when the subject thought about the problems. The "depth of search" means the longest moves for which they searched, as reflected by verbal data.

The results are shown in Table 2, which displays the average changes between the 1<sup>st</sup> period and the 3<sup>rd</sup> period. Each entry in Table 2 consists of data averaged among six subjects. Statistical significance was assessed via t-tests of correlations between two samples.

Table 2: Changes between 1<sup>st</sup> period and 3<sup>rd</sup> period

	1 <sup>st</sup> period	3 <sup>rd</sup> period
Recognition of a position [s] *	37.8	52.1
Amount of eye movement [pixels] *	21089	27963
Thought time [s]	137.4	149.8
Generation of candidate moves [number] *	2.0	1.7
Depth of search [depth]	2.6	2.5

$p < 0.05$  (\*)

**Discussion** Subjects generated fewer candidate moves during the 3<sup>rd</sup> period than they did during the 1<sup>st</sup> period. This result corresponds with those of previous research.

However, other aspects of the results were different from those of previous research. We considered that the reason for these discrepancies may have been that the six subjects did not learn enough.

## fMRI Experiment

**Method** Brain activity was investigated using an fMRI scanner owned by RIKEN. The stimuli were 180 easy original next move problems (one-move checkmates, etc.).

One hundred sixty of the problem stimuli were between the early stage and the last stage of learning, and 20 of the problems were identical during the early stage and the last stage of learning. In the interrupt task used for initialization of brain activity, if a Gold piece appeared, subjects were instructed to press a button. Moreover, a task of reporting the King's position was set as the control task. The control task was used in order to check whether the subject understood the board position correctly.

We explained that a next move problem or a control problem were displayed for 2 seconds and required to answer it within 3 seconds.

**Results** In the experiment using fMRI, data from all 19 subjects were analyzed. Ten subjects were sorted into a low-rank group and nine subjects, into a high-rank group on the basis of the percentage of correct answers in the experiment conducted during the last stage of learning.

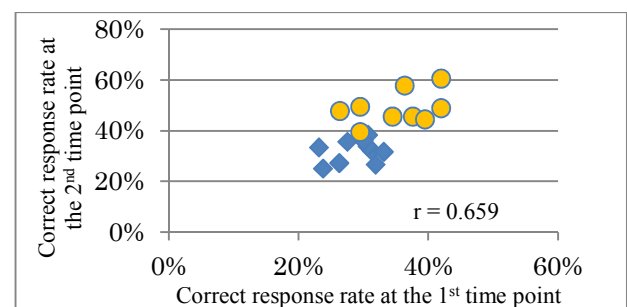


Figure 2: Correct response rates during the 1<sup>st</sup> and 2<sup>nd</sup> fMRI experiments

Figure 2 expresses the percentages of correct answers in the early stage and the last stage. The high-scoring group's data are represented by circles, and the low-scoring group's data are represented by diamonds.

Figure 3 expresses the percentage of correct answers in the last stage of learning and the strength of caudate nucleus activity. The high-scoring group's data are represented by circles, and the low-scoring group's data are represented by diamonds. Strength of activity expresses the ratio of activation in the last stage to that in the early stage. The activity was seen in the high score group. Figure 4 shows the region of enhanced activity.

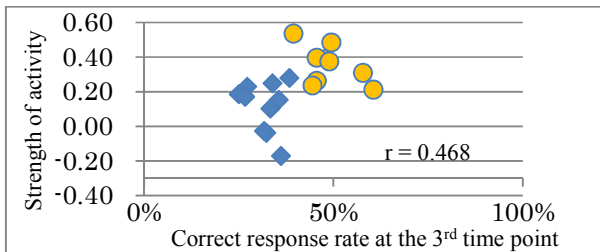


Figure 3: Strength of activity and correct response rate at the 3<sup>rd</sup> time point

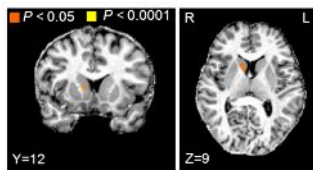


Figure 4: Region of enhanced activity

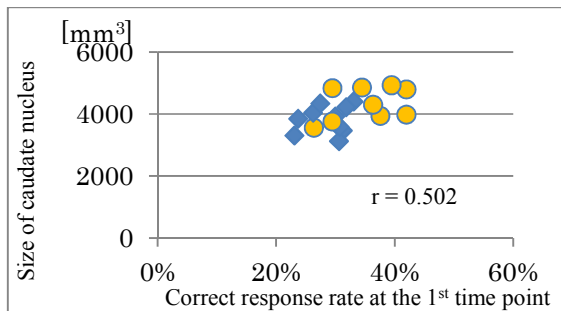


Figure 5: Size of caudate nucleus and correct response rate at the 1<sup>st</sup> time point

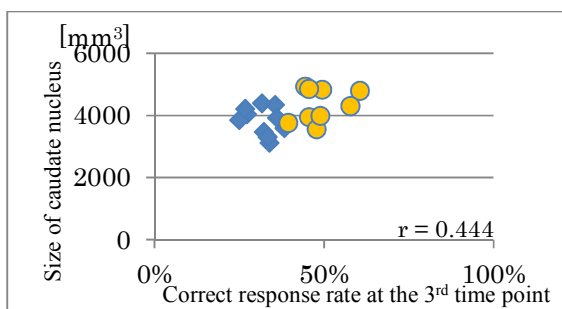


Figure 6: Size of caudate nucleus and correct response rate at the 3<sup>rd</sup> time point

Figure 5 expresses the rate of correct answers and the size of the caudate nucleus at the early stage. Figure 6 expresses the rate of correct answers and the size of caudate nucleus at the last stage. The high-scoring group's data are represented by circles, and the low-scoring group's data are represented by diamonds.

**Discussion** It is possible that good early learning rates could be caused by large caudate nuclei, as a correlation between the percentage of correct answers in the early stage and the size of the caudate nucleus was found.

The size has a weak correlation with the percentage of correct answers in the last stage.

The subjects whose percentage of correct answers was good have a tendency for caudate nucleus activity to be strong in the last stage. This result suggests that everyone can learn regardless of the size of the caudate nucleus.

## Additional Analysis

### Group analysis

The results of our cognitive experiment may have differed from the previous research because our subjects did not become strong enough. Therefore, we decided to take the subjects' strength into consideration.

As the scale that measures subjects' relative strength, we decided to use a combination of the rankings of the tournament held at the end of the experiment and the rate of correct answers to problems in the fMRI experiment performed at RIKEN. Figure 7 displays the subjects; the subjects who were used in the cognitive experiment are expressed as larger points.

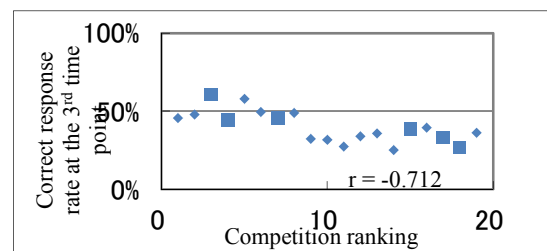


Figure 7: Correct response at the 3<sup>rd</sup> time point and competition ranking

The correlation coefficient between the competition rankings and the percentage of correct answers was -0.712. We divided the two groups on the basis of competition rankings. The high-scoring group was composed of the subjects placing 3<sup>rd</sup>, 4<sup>th</sup>, and 7<sup>th</sup> in the tournament, and the low-scoring group placed 15<sup>th</sup>, 17<sup>th</sup>, and 18<sup>th</sup> out of 19 participants.

### Results

Table 3 displays the performance characteristics of the high-scoring group when the subjects are placed and analyzed in these groups.

Compared with Table 2, recognition of a position, amount of eye movement, and thought time have decreased.

Table 3: Changes in the high-scoring group during the study

	1 <sup>st</sup> period	3 <sup>rd</sup> period
Recognition of a position [s] **	31.5	21.1
Amount of eye movement [pixels] *	15560	12411
Thought time [s] *	137.4	88.7
Generation of candidate moves [number]	2.2	2.0
Depth of search [depth]	3.2	2.9

$p < 0.05$  (\*)  $p < 0.01$  (\*\*)

**Discussion** It seems that these are the features of expertise, as our results corresponded with those of previous research on certain points, such as reduction of time necessary to recognize a position, reduction of the amount of eye movement, and reduction of thinking time obtained in the high-scoring group.

A different result from previous research was obtained regarding whether searching deeply implies that skill is high. Although the tendency for search to be deep was seen as subjects became experts in previous research, this tendency was not seen in our results.

This result indicates that deep searching ability was not acquired by learning.

The items that differ between the high- and low-scoring groups at the 1<sup>st</sup> time point are the following:

Table 4: Difference between low score and high score

	low	high
Recognition of a position [s] **	44.1	21.1
Amount of eye movement [pixels] **	26616	12411
Thought time [s]	137.5	88.7
Generation of a candidate moves [number] *	1.7	2.0
Depth of search [depth] **	2.0	2.9

$p < 0.05$  (\*)  $p < 0.01$  (\*\*)

We expected that these differences might express variation in ease of learning. Bransford suggested that there is a type of beginner called an intellectual beginner (Bransford et al., 1982). He compared an efficient learner with an inefficient learner by stating that an inefficient learner did not notice the difficulty of a study subject and did not change his/her strategy when study subjects changed. The difference seen in this experiment may support the existence of the intellectual beginner.

### Analysis of Cognitive and fMRI experiment

The cognitive data obtained in the cognitive experiment and the brain imaging data acquired in the fMRI experiment are set and examined. Percentage of correct answers, size of the

caudate head, playing strength, and cognitive activity data are mainly compared.

**Results** Figure 8 expresses the percentage of correct answers to the problem used in the cognitive experiment in the 1<sup>st</sup> period and the percentage of correct answers to the problem used in the fMRI experiment at the early stage. The high-scoring group's data are represented by circles, and the low-scoring group's data are represented by diamonds. Correlations were not seen between the two variables.

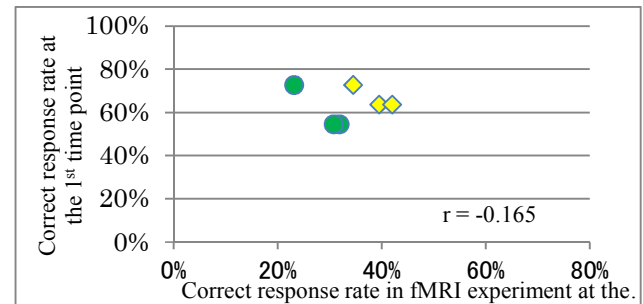


Figure 8: Relationship between fMRI and cognitive experiment at the 1<sup>st</sup> time point (at the early stage)

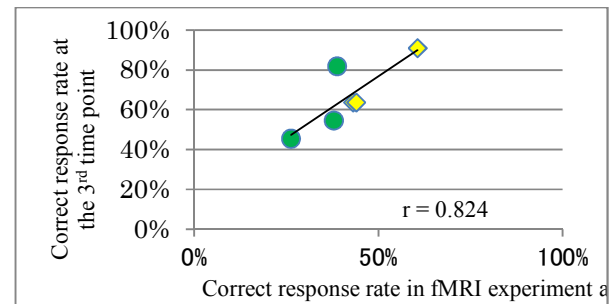


Figure 9: Relationship between fMRI and cognitive experiment at the 3<sup>rd</sup> time point (at the last stage)

Figure 9 plots the percentage of correct answers to the problems used in the cognitive experiment in the 3<sup>rd</sup> period and the percentage of correct answers to the problems used in the fMRI experiment at the last stage. The high-scoring group's data are represented by circles, and the low-scoring group's data are represented by diamonds. There was a correlation between the two variables.

Table 5: Correlation coefficients between results of the cognitive experiment and the fMRI experiment

	Strength of activity	The size of caudate nucleus
Recognition of time positions [s]	0.021	0.746
Amount of eye movement [pixels]	-0.292	0.436
Thought time [s]	0.103	0.807
Generation of candidate moves [number]	-0.534	-0.424
Depth of search [depth]	0.538	0.282

Table 5 shows correlations between the differences of results in the cognitive experiment and the fMRI experiment. Correlations were found a little at a time in order to recognize the position and size of the caudate nucleus. Correlation was also found a little at a time between thinking time and the size of caudate. However, correlations were seldom seen among other items. Moreover, no correlation was seen when activity at the 1<sup>st</sup> and 3<sup>rd</sup> time points, brain activity, and the size of the caudate head were compared.

**Discussion** Figure 8 shows that there was no difference between the two groups' rates of considering carefully or answering problems intuitively in the 1st period of study. However, Figure 9 shows the relationship between the percentage of correct answers on problems that were considered carefully and the percentage of correct answers that were answered by intuition in the last stage of the study. It is therefore possible that intuitive and deliberative thinking abilities are not totally separate, and that these two types of thinking have mutually influenced one another. Table 4 shows that there is a relationship between the time to understand positions and the volume of the caudate nucleus. Since there is a close relationship between time to understand positions and thinking time, it is thought that the relationship was caused by both types of thinking. Since there is no relationship between cognitive change and change of brain activity in other areas, it is suggested that the changes are independent.

In the fMRI experiment, we investigated what happens to brain activity when a subject solved a problem by intuition. Some subjects described intuition in the reports that they presented once per week. A certain subject reported the following description during the 10<sup>th</sup> week of the experiment:

***"A move may flash by intuition or not flash."***

Moreover, another subject gave an analogous description:

***"If I encounter an important point in the game, I generate a move by intuition and search deeply."***

Both of the subjects seem to have considered (in light of the fact that they gave such descriptions at the end of the experiment) that intuitive ability is supported by prolonged learning.

## Conclusions

In this research, changes of expertise in Mini-Shogi were investigated in a cognitive experiment and an fMRI experiment. In the cognitive experiment, the results were analogous to those of previous experiments in terms of the difference between beginners and experts, showing reduction in thought time and reduction of eye movement upon position recognition with increasing expertise while obtaining almost the same level of skill in the subject. In the fMRI experiment, the high-scoring group displayed higher caudate nucleus activity, as seen in previous work.

Comparing the results of the cognitive experiment and the fMRI experiment, a correlation was found between the time

to understand positions and the size of caudate nucleus. However, correlation was not seen among other variables.

On the basis of this report and the results of our experiments, we suggest that subjects' intuitive ability and deliberation ability are not completely independent and that they may mutually influence each other.

## References

- Atherton, O., Zhuang, J., Bart, W. M., Hu, X. P., & He, S. (2003). A functional MRI study of high-level cognition. I. The game of chess. *Cognitive Brain Research*, 16, 26-31.
- Bransford, J. D., Stein, B. S., Vye, N. J., Franks, J. J., Auble, P. M., Mezynski, K. J., & Perfetto, G. A. (1982). Differences in approaches to learning: An overview. *Journal of Experimental Psychology: General*, 111, 390-398.
- Bransford, J. D. (2000). *How people learn: Brain, mind, experience, and school.*, (expanded ed.). Committee on Developments in the Science of Learning. J. D. Bransford, A.L. Brown, and R.R. Cocking (Eds.), Washington, D.C: National Academy Press.
- Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4, 81.
- Chen, X., Zhang, D., Zhang, X., Li, Z., Meng, X., He, S., & Hu, X. P. (2003). A functional MRI study of high-level cognition. II. The game of Go. *Cognitive Brain Research*, 16, 32-37.
- de Groot, A. D. (1965). *Thought and choice in chess*. Mouton Publishers.
- de Groot, A. D., & Gobet, F. (1996). *Perception and memory in Chess –studies in the heuristics of the professional eye*. Van Gorcum.
- Ito, T., Matsubara, H., & Reijer, G. (2004). Cognitive science approach to Shogi playing processes (2) – Some results on next move test experiments. *Information Processing Society of Japan*, 45(5), 1481-1490.
- Ito Lab. 55Shogi portal [Internet monograph]. Retrieved January 23, 2012 from <http://minerva.cs.uec.ac.jp/~uec55/index.php>.
- Law, B., Atkins, M. S., Kirkpatrick, A. E., Lomax, A. J., & Mackenzie, C. L. (2004). Eye gaze patterns differentiate novice and expert in a virtual laparoscopic surgery training environment. *Proceedings of ACM Symposium of Eye Tracking Research & Applications*. Publisher: ACM Press.
- Osaka R., Nakamizo, Y., & Koga, K. (1993). *Experimental psychology of eye movements*. The University of Nagoya Press.
- Paolo, N., Jordan, G., Pietro, P., David, A., John, C. C. & Robert, M. (1994). Brain activity in chess playing. *Nature*, 369, 191.
- Wan, X., Nakatani, H., Ueno, K., Asamizuya, T., Cheng, K., & Tanaka, K. (2011). The neural basis of intuitive best next-move generation in board game experts. *Science*, 331, 341-346.

# One-shot lotteries in the park

Mordechai Z. Juni (mjuni@nyu.edu)<sup>1</sup>

Todd M. Gureckis (todd.gureckis@nyu.edu)<sup>1</sup>

Laurence T. Maloney (laurence.maloney@nyu.edu)<sup>1,2</sup>

<sup>1</sup>Department of Psychology, <sup>2</sup>Center for Neural Science, NYU

6 Washington Place, New York, NY 10003 USA

## Abstract

How do people manipulate their environment when balancing trade-offs between probability of success and payoff? Individuals in a city park played a simple lottery using a small set of marbles placed in an urn. Participants had the ability to actively improve their chances of winning but only by reducing the amount of money that they could possibly win. Hence, participants controlled the lottery's intuitive trade-off between probability of success and potential payout. Across four different lottery structures, participants, on average, behaved systematically *safer* than the optimal strategy that maximizes expected gain. We explore two different accounts of this sub-optimal choice behavior: probability distortion, and intrinsic utility of winning.

**Keywords:** decision making; one-shot lottery; probability distortion; intrinsic utility of winning.

## Introduction

Typical decision making studies offer participants choices between two fixed alternatives. Sometimes the prospects are explicitly described: "Would you prefer to draw from Deck-1 which awards \$2 with .8 probability or from Deck-2 which awards \$6 with .3 probability?" Other times the prospects must be learned from experience: "After sampling repeatedly from both decks and observing the outcomes, you will choose which deck to draw from for the trial that counts." Dissimilar results from these two kinds of experimental paradigms have led to an explosion of research concerning the description-experience gap in decision-making under risk (for a brief review, see Hertwig & Erev, 2009).

While this dichotomy has gathered considerable focus, some decisions are not clearly descriptive or clearly experience-based. We think of these situations as intuitive "everyday" decisions that implicitly select one out of many choices available. A key feature of such decisions is that people need not consider explicitly every possible alternative to make a decision.

In the present study, participants manipulated their environment to choose between a large number of different prospects that weren't explicitly described and weren't directly experienced. The experimental task was a one-shot lottery whose parameters (probability of winning and magnitude of monetary prize) were partially under participants' control. Critically, the lottery was carefully designed so that increasing the probability of winning automatically decreased the potential monetary prize, and increasing the potential monetary prize automatically decreased the probability of winning (an inverse relationship typical of many real-world lotteries).

Our goal in this study was three fold. First, we were interested in how people approach situations where they have

control over a potentially rewarding stochastic environment (see also Juni, Gureckis, & Maloney, 2011). In particular, do people manipulate their environment to maximize their expected gain? Second, we were interested in intuitive "everyday" decision-making where a large number of prospects must be discerned. Finally, we were interested in taking some of our recent decision making research out of the laboratory to consider a more diverse population of decision makers.

## Taking decision research out of the lab

The majority of psychological research in cognitive science on decision making is conducted using laboratory studies with college undergraduates. However, a number of recent arguments have been presented for why such populations may not be representative of the general human population (Henrich, Heine, & Norenzayan, 2010). In addition, a large percentage of decision making research is conducted in the laboratory on computers for either real or hypothetical amounts of money. One concern about computer-based studies is that participants may suspect that the odds or payouts are being manipulated as part of the study design.

To address these concerns, we ventured outside of the laboratory and into the streets of a large US city to elicit choice behavior from randomly chosen pedestrians who were posed with a single, non-hypothetical problem that was played out for real, in person. Those who agreed to participate were informed that the lottery would be played only once and that they would receive real money if they won. The lottery was performed using a physical urn and marbles that the participant could see and touch. This guaranteed that participants could be certain that there was no manipulation of the odds (e.g., by a computer program).

## The "marbles" game

The one-shot lottery we implemented is very intuitive. The urn initially contains several black marbles and no white marbles. After agreeing to participate, the subject is handed several white marbles. To win the lottery, the participant must, without looking, pull out a white marble from the urn. Thus, the participants must put at least some of the white marbles into the urn so that they have a chance of winning the lottery.

Of course, one strategy might be to place all the white marbles in the urn (to maximally increase the odds of successfully drawing a white marble). However, the monetary prize for pulling out a white marble from the urn is determined by the number of white marbles that the participant chooses to *not* put into the urn but rather set aside as potential prize money.



Each white marble that was not placed into the urn represents \$1 in prize money, which is only awarded if the marble that the participant pulls from the urn turns out to be white.

Four different conditions were tested which varied the number of black marbles that were first placed in the urn: 1, 2, 8, or 25. In all conditions the participant was handed 10 white marbles. If the participant were to put zero white marbles into the urn, they have no chance of winning. On the other hand, if they were to put all 10 white marbles into the urn, they win nothing even if they draw a white marble. Thus, the experimental design restricted participants to put anywhere between 1 and 9 white marbles into the urn. This number becomes our primary dependent measure (i.e., “how many white marbles do participants put into the urn as a function of the number of black marbles in the urn?”).

Figure 1 shows how the different probability structures of the four conditions affects the respective expected gain functions. The figure also shows the respective number of white marbles that should be put into the urn to maximize expected gain in each condition. As is visible, the lottery was deliberately designed so that the normative ideal rule regarding how many white marbles should be put into the urn is different for each condition. When there is one black marble in the urn, the participant can maximize expected gain by putting two white marbles into the urn. When there are two black marbles, expected gain is maximized by putting three white marbles into the urn. When there are eight black marbles, expected gain is maximized by putting four white marbles into the urn. And, finally, when there are 25 black marbles in the urn<sup>1</sup>, expected gain is maximized by putting five white marbles into the urn.

The basic question asked in our study is if people combine information about probability and reward to maximize their expected gain in this intuitive one-shot lottery.

## Methods

**Subjects** Data was collected from people walking through Washington Square Park in New York City. 120 people (65 males and 55 females) participated in the experiment (30 per condition). Ages ranged from 18 to 75 years, with an average age of 29.77 years ( $SD=14.65$ ) and a median age of 24 years. Participants were not compensated for their time, but they did receive the prize money (anywhere between \$1 and \$9) if they won their lottery.

**Materials** To conduct the lottery we used transparent cups, an opaque urn, 10 white marbles, and a varying number of black marbles (1, 2, 8, or 25). The black and white marbles were identical except for color and could not be identified through touch.

<sup>1</sup>In the limit, as the number of black marbles grows very large, the rule that maximizes expected gain is to put half of the white marbles into the urn and set the other half aside as prize money. Given that participants were given 10 white marbles to work with, it is impossible for the optimal rule to dictate placing more than five of them into the urn no matter how many black marbles there are in the urn. We thank Hang Zhang for pointing this out to us.

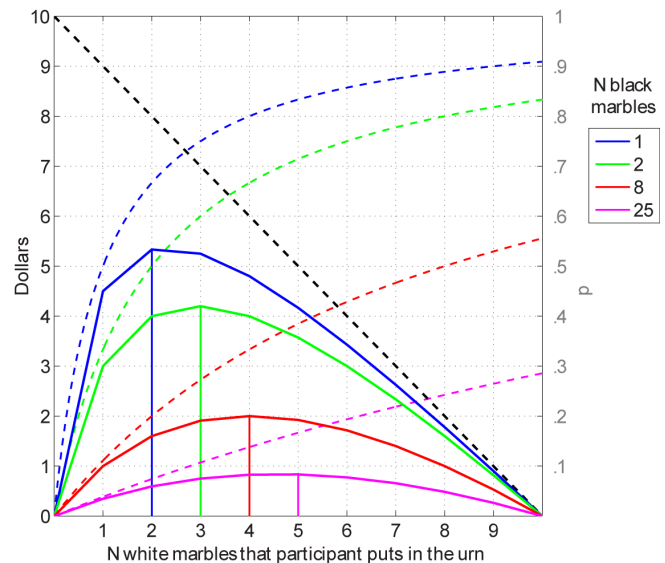


Figure 1: Experimental design. The participant decides how many white marbles to put into the urn and how many to set aside as the potential prize for winning the lottery. The black dashed diagonal shows the prize money that the participant receives if the marble that she pulls out from the urn turns out to be white. It starts at \$9 when only one white marble is put in the urn and declines to \$1 when nine white marbles are put into the urn. The colored dashed curves show the probability (p) of winning the lottery as a function of the number of black marbles in the urn and the number of white marbles that the participant chooses to put into the urn. Each probability function is multiplied by the gain function (i.e., the black dashed diagonal) to generate an expected gain function. The colored solid lines show the discrete expected gains of the lottery as a function of the number of black marbles in the urn and the number of white marbles that the participant chooses to put into the urn. The vertical colored lines show the maximum expected gain for each of the four experimental conditions.

**Procedure** The lottery was conducted next to a bench in Washington Square Park. On the bench there was an empty urn, a cup with white marbles, and a cup with black marbles. The experimenter handed the participant the cup with the white marbles and said the following:

“To conduct the lottery we will be using marbles. You are in control of the white marbles. There are 10 white marbles and each represents \$1. I am in control of the black marbles. There are (1, 2, 8, 25) of them.”

The experimenter then showed the participant an envelope with a large number \$1 bills and emphasized that if the participant won the lottery they would receive real money. Next, the experimenter poured the black marbles into the urn and said the following:

“I have poured the black marbles into the urn. To perform the lottery, you will be placing your hand into the urn and pulling out a single marble without looking. If the marble



that comes out is black you win nothing. If the marble that comes out is white you will win the number of white marbles that you chose not to put into the urn but rather set aside as your potential prize money. This lottery will be performed only once.”

Next the experimenter held the urn in one hand and an empty cup in the other hand and said the following:

“Take your white marbles and decide how many you want to put into the urn for the lottery and how many you want to set aside in this cup as your potential prize money if you win the lottery. Remember, each white marble is worth \$1.”

Once the participant divided up the white marbles between the urn and the prize cup, the experimenter asked the participant to confirm verbally what the monetary prize would be if they pulled out a white marble from the urn to ensure that there weren’t any misunderstandings.

Next the experimenter held up the urn above the participant’s eyes and asked the participant to pull out a single marble from the urn. If the participant pulled out a black marble they received no money; if they pulled out a white marble they were paid \$1 for each white marble that they set aside in the prize cup.

## Results

Each of the four experimental conditions had 30 different participants. Each participant provided one data point. We report how many white marbles they chose to put into the urn. The outcomes of the lotteries are irrelevant to our study and so we do not report them.

Figure 2 shows a box and whiskers plot for the number of white marbles put into the urn in each experimental condition. The colored asterisks mark the optimal number of white marbles that should be put into the urn to maximize expected gain in each condition. For each of the four conditions we used a single-sample t-test with the null hypothesis set to the optimal number of marbles for the given condition. The stars indicate the level of significance (see Figure 2 caption).

The number of white marbles that participants, on average, tended to put into the urn increased systematically with an increase in the number of black marbles in the urn. Furthermore the results indicate that, on average, participants systematically put one more white marble into the urn than dictated by the normative rule that maximizes expected gain (i.e., people, on average, are sub-optimal with respect to the normative rule, preferring a \$1 decrease to their potential prize in exchange for an increase to their probability of winning the lottery).

## Discussion

Our results seem to indicate that participants, on average, did not maximize expected gain. Curiously though, the number of white marbles that they tended to put into the urn was one more than the normative ideal rule irrespective of the experimental condition. This systematic tendency led us to explore possible explanations to account for their sub-optimal manipulation of the lottery’s parameters.

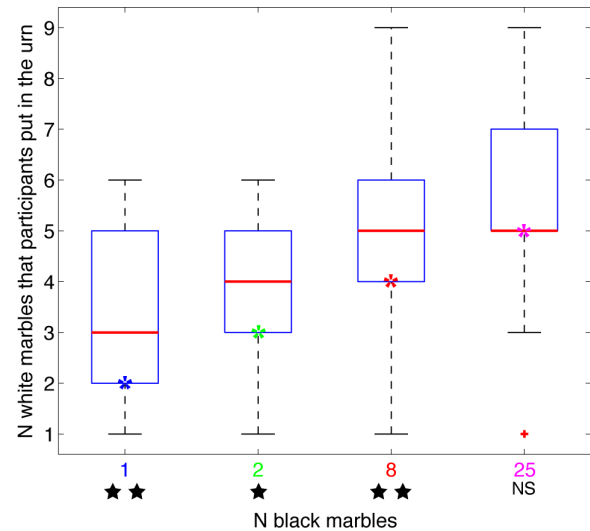


Figure 2: Box and whiskers plot showing the results in all four conditions. The colored asterisks mark the corresponding optimal number of white marbles that should be put into the urn to maximize expected gain. Participants, on average, tended to put “one too many” white marbles into the urn with respect to optimal. Two stars indicate that this was significant at the .01 level, while one star indicates that it was significant at the .05 level. The non-significant condition had a  $p = .08$ .

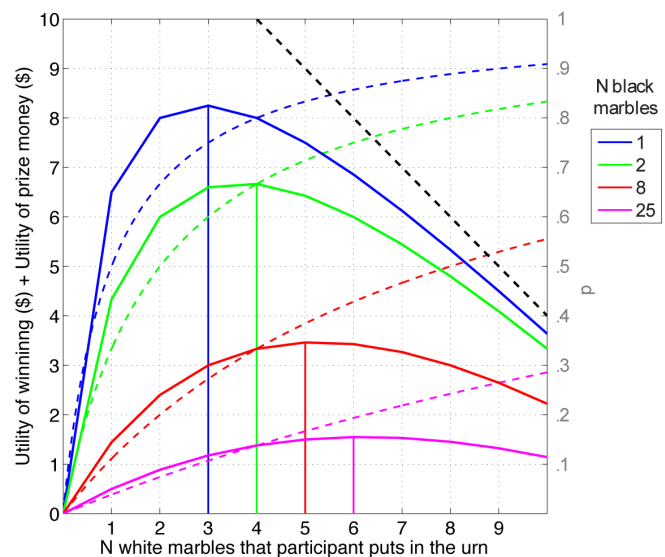


Figure 3: Expected utility functions if we take into account an additional utility of \$4 for winning the lottery. Compare this figure to Figure 1 that shows the expected gain functions without taking into account any additional utility of winning. Notice that the maximum expected utilities under this scheme are shifted one marble to the right in all four conditions relative to the maximum expected gains in Figure 1.

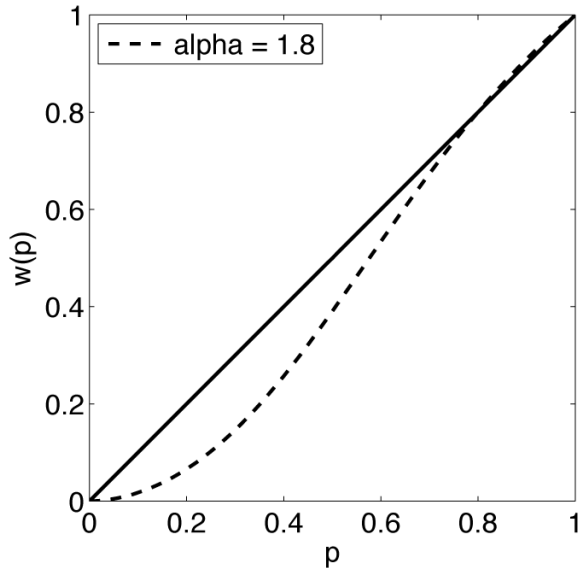


Figure 4: S-shaped probability distortion.

The first possibility we explored is that participants might have had an intrinsic utility for winning the lottery in addition to their utility for the actual money that they receive if they win (Parco, Rapoport, & Amaldoss, 2005). We explored this possibility by putting a fixed dollar value on the intrinsic utility of winning (a free parameter).

Figure 3 shows the expected utility functions when the intrinsic utility of winning is valued at \$4. Notice that the maximum expected utilities shift rightward one marble relative to the normative rule that maximizes expected gain. This rightward shift persists if the intrinsic utility of winning is anywhere between \$3.10 and \$4.60.

This analysis suggests that participants' sub-optimal behavior could be accounted for if they have an intrinsic utility for winning the lottery that is in addition to their utility for the actual money that they receive if they win.

The second possibility we explored is that participants might have had a distortion in subjective probability. A standard single-parameter model for distortion of probability in decision-making under risk is written as follows (Tversky & Kahneman, 1992):

$$w(p) = \frac{p^\alpha}{(p^\alpha + (1-p)^\alpha)^{1/\alpha}} \quad (1)$$

As our task resembles a decision from description more than a decision from experience, we hypothesized that an Inverse-S-shaped probability distortion might be more likely to account for the data than an S-shaped probability distortion (Ungemach, Chater, & Stewart, 2009; Wu, Delgado, & Maloney, 2009). In other words, we expected that participants' average behavior might be accounted for with an  $\alpha < 1$  as is commonly found in decisions from description, and not with an  $\alpha > 1$  as is commonly found in decisions from experience (but see Glaser, Trommershäuser, Mamassian, & Maloney,

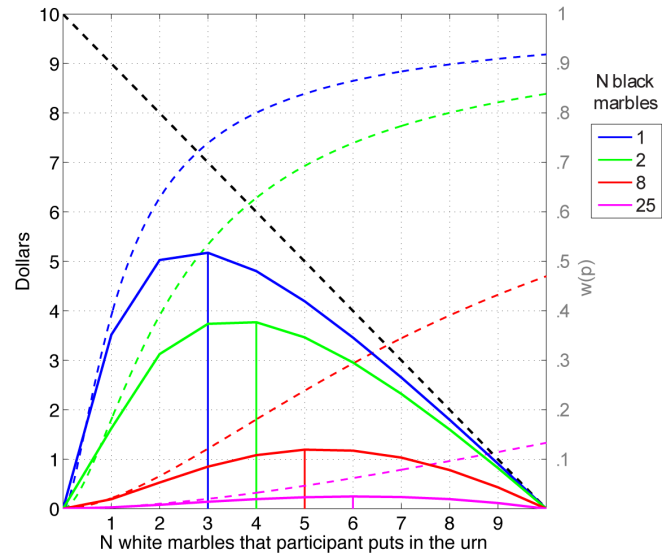


Figure 5: Expected gains based on the S-shaped probability distortion depicted in Figure 4. Compare this figure to Figure 1 that shows the true probabilities and expected gain curves. Notice that the maximum expected gains under this scheme are shifted one marble to the right in all four conditions relative to the maximum expected gains in Figure 1.

2012). However, the only way to shift the maximum expected gain one marble to the right in all four conditions is by having  $1.69 < \alpha < 1.92$ .

Figure 4 shows an S-shaped probability distortion with  $\alpha = 1.8$ . Figure 5 shows the consequent expected gain functions based on this distorted probability function with the maximum expected gains shifted one marble to the right.

This analysis suggests that participants' sub-optimal behavior could be accounted for if they have an S-shaped probability distortion, a form only rarely encountered in decisions from description tasks.

## Conclusion

Participants in typical decision making experiments select among a small number of lotteries each with fixed probabilities and rewards. In the experiment reported here we considered a task where participants could manipulate their environment to improve their chances of winning the lottery but only by reducing the amount of money that they could possibly win.

While a simple and preliminary study, our results may prove useful to researchers interested in how people balance risk and reward in simple, intuitive decision tasks. Our experiment is unique in a couple of ways that are worth pointing out.

First, participants could actively manipulate the odds of successfully winning the lottery (by placing more or less white marbles into the urn). While the literature on risky decision making is immense, we are unaware of previous studies

that considered this *active* manipulation of the decision environment (one exception is a related study we reported last year in Juni et al., 2011).

Second, the decision environment was set up to be intuitive, fast, and easy to conduct with everyday people walking through a park. As a result, we collected a more diverse sample than is typical for research on judgment and decision-making.

Our primary results are that participants, on average, did not maximize expected gain. In particular, participants, on average, tended to put one additional white marble into the urn than dictated by the normative ideal rule. By doing so they slightly increased their probability of winning relative to ideal, but also decreased their expected winnings slightly.

We considered two different accounts for participants' sub-optimal choice behavior. Participants in decisions from description typically over-weight small probabilities and under-weight large probabilities (Inverse-S-shaped probability distortion). However, we found that we could account for the results obtained only if we assumed that participants under-weighted small probabilities and over-weighted large probabilities (S-shaped probability distortion). While such a tendency has been found in decisions from experience, our task resembles a decision from description more than a decision from experience, making an S-shaped probability distortion surprising.

A second account for our data is that participants have an intrinsic utility for winning the lottery that is in addition to their utility for the actual money that they receive if they win. If this account of participants' behavior is correct, it would seem that their intrinsic utility of winning is approximately \$4.

Future studies could tease apart these two accounts of participants' behavior by keeping the objective probabilities of the lotteries the same and simply scaling up the value of each white marble. According to the probability distortion account we should see no change in participants' average behavior. But according to the intrinsic utility of winning account we should see participants' average behavior shift toward optimal as the intrinsic utility of winning is diluted relative to the increased potential winnings that the lotteries afford.

## Acknowledgments

MZJ was supported by NIH Grant T32 EY007136. LTM was supported NIH Grant EY019889. TMG was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

## References

- Glaser, C., Trommershäuser, J., Mamassian, P., & Maloney, L. T. (2012). Comparison of distortion of probability information in decision under risk and an equivalent visual task. *Psychological Science*, 23(4), 419-426.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61-135.
- Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, 13(12), 517-523.
- Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2011). Don't stop 'til you get enough: Adaptive information sampling in a visuomotor estimation task. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (p. 2854-2859). Austin, TX: Cognitive Science Society.
- Parco, J. E., Rapoport, A., & Amaldoss, W. (2005). Two-stage contests with budget constraints: An experimental study. *Journal of Mathematical Psychology*, 49(4), 320-338.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 297-323.
- Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighed or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, 20(4), 473-479.
- Wu, S. W., Delgado, M. R., & Maloney, L. T. (2009). Economic decision-making under risk compared with an equivalent motor task. *Proceedings of the National Academy of Sciences*, 106(15), 6088-6093.

# Children's acquisition of fraction knowledge from concrete versus generic instantiations

**Jennifer A. Kaminski (kaminski.16@osu.edu)**

Center for Cognitive Science, Ohio State University  
210A Ohio Stadium East, 1961 Tuttle Park Place, Columbus, OH 43210, USA

**Vladimir M. Sloutsky (sloutsky.1@osu.edu)**

Center for Cognitive Science, Ohio State University  
208C Ohio Stadium East, 1961 Tuttle Park Place, Columbus, OH 43210, USA

## Abstract

The goal of this experiment was to investigate elementary school children's ability to acquire basic fraction knowledge. The degree of concreteness of instantiations of proportions was varied between subjects. First-grade children learned to label proportions of objects with fraction. Proportions were presented either as concrete, colorful flowers or as generic black and white circles. Following instruction, participants were given a test of learning and an immediate or delayed test of transfer involving proportions of novel objects. Those who learned with the generic materials scored higher on learning and transfer than those who learned with the concrete materials. Differences between learning conditions were attenuated for the delayed transfer test. These findings suggest that concrete, perceptually rich instantiations of fractions may hinder children's acquisition of basic fraction knowledge in comparison to simple generic instantiations of fractions.

**Keywords:** Psychology; Education; Learning; Transfer; Relations, Mathematics Education.

## Introduction

Mathematical concepts are often difficult for children to acquire. One response to this challenge is to introduce concepts to students through concrete instantiations which include perceptually rich, familiar material. The use of concrete material is widespread in education (see McNeil & Uttal, 2009 for discussion). Concrete instantiations of mathematics may involve familiar contexts and can be visually appealing and engaging. For example, simple arithmetic concepts are often instantiated through sets of familiar objects, such as two apples plus three apples equals five apples. Such material may spark interest in the learning task and maintain attention on the learning material.

However, a primary goal of learning mathematics is the ability to apply mathematics to new situations. Therefore, successful acquisition of mathematical knowledge implies that the learner has not only acquired knowledge of the mathematical relations in the context of learning, but also has the ability to transfer the mathematical knowledge to novel isomorphic situations. There is evidence that concrete instantiations can hinder transfer of learning. Adults who learned a novel mathematical concept from a generic, perceptually sparse instantiation were better able to transfer

this knowledge to a novel isomorph than those who learned the same concept from a concrete instantiation (Kaminski, Sloutsky, & Heckler, 2008; Sloutsky, Kaminski, & Heckler, 2005; see also Goldstone & Sakamoto, 2003; Goldstone & son, 2005 for related findings).

In comparison to more abstract, generic instantiations, concrete instantiations of a given concept communicate more extraneous information. For example, a photograph of a person communicates more nonessential information than a simple stick figure drawing (see Kaminski & Sloutsky, 2011 for a discussion). Similarly, instantiating addition as the sum of apples communicates more information than instantiating it as the sum of tally marks. This additional information (e.g. the appearance, taste, etc. of apples) is extraneous to the mathematics and may present an obstacle for learning for the following reason. Mathematical concepts are defined by relational structure. Relations are less salient than objects (e.g. Gentner, 1988). Instantiating mathematical concepts through concrete material, in comparison to a more generic format, may increase the salience of superficial aspects of the learning material and consequently divert the learner's attention from the to-be-learned relational structure (see Goldstone, Medin, & Gentner, 1991 for a similar argument regarding similarity judgments), thus hindering learning.

Therefore, it appears that generic instantiations of mathematical concepts have an advantage over concrete instantiations with respect to transfer. However it could be argued that this advantage is limited to older learners. After all, in educational practice, older students are expected to learn and reason with abstract instantiations including symbols, equations, and other standard notation. It may be that younger learners (e.g. elementary school students) may need concrete instantiations to begin to acquire abstract knowledge.

In addition to the notion that concrete material may be more engaging for the learner, support for the use of concrete instantiations to teach abstract concepts to young children is often tied to theories of learning and development. Some developmental theories (Montessori, 1917; Piaget, 1970) posit that young children's thinking is inherently concrete and that they are not capable of reasoning about abstract concepts using symbols. According

to these theories, children proceed through developmental stages in which their reasoning becomes more abstract and less dependent on concrete material. Other theories (e.g., Bruner, 1966) tie the ability to reason about abstract material not to developmental stages but to levels of knowledge. From this perspective, all novices, regardless of age, would need concrete instantiations to begin to acquire knowledge of an abstract concept. Both accounts suggest that learning should begin by instantiating the mathematical concept through concrete, familiar material.

However, if the difficulty transferring knowledge from concrete instantiations is due to extraneous information diverting attention from the relevant relational structure, then concrete instantiations may be at least as detrimental for children's learning as they are for adults. Children have difficulty controlling their attentional focus and filtering irrelevant, potentially distracting information (Kemler, 1982; Shepp & Swartz, 1976; Smith & Kemler, 1978, see also Hanania & Smith, 2010). For example, Shepp and Swartz (1976) instructed 6- and 9-year-olds to sort items according to shape, with color being an irrelevant dimension. It was found that 6-year-olds (but not 9-year-olds) were slower when color varied independently of shape than when color co-varied with shape or did not vary at all. Therefore, the task-irrelevant dimension affected performance of younger, but not older participants.

There is also evidence that concrete, perceptually rich material can hinder preschool children's ability to perform simple relational tasks in comparison to performance with generic material. One line of evidence comes from studies of children's early symbol use. Successful symbol use requires the detection and transfer of common relations. For example, to effectively use a map as a symbol for a real location, one must recognize the common relations between entities on the map and their real-world analogs. Two- and three-year-old children are more successful transferring location information from a picture to the real world than from a 3-dimensional scale model to the real world (DeLoache, 1995a, 1995b). These findings suggest that preschool children have difficulty using concrete, perceptually rich objects as symbols than using less concrete objects as symbols.

In addition, preschool children are better able to detect relations of monotonic increase and monotonic decrease in size between displays that involve simple perceptually sparse objects than concrete, perceptually rich objects (Gentner, Ratterman, Markman, & Kotovsky, 1995; Kaminski, & Sloutsky, 2010). It has also been demonstrated that kindergarten children are better able to recognize common proportions across displays of different objects when first given instruction with generic, perceptually sparse objects than when given instruction with concrete, perceptually rich objects (Kaminski & Sloutsky, 2009). Taken together there is evidence that concrete, perceptually rich material may hinder the recognition of relational structure for kindergarteners and preschool children.

Less is known about how concreteness affects young school-aged children's acquisition of relational knowledge that is part of standard mathematics content. While concrete instantiations of mathematics may communicate distracting extraneous information, it is possible that school-aged children have developed sufficient inhibitory control to focus on the relevant relations and not be distracted by extraneous aspects of concrete material. If this is the case, then concrete instantiations will not hinder learning of mathematical concepts and may even facilitate learning by making the material more interesting for children. However, while executive function is maturing throughout childhood, the complexity of the relations we expect children to learn is increasing.

Higher-order mathematical concepts involve more complex relations than simpler mathematical concepts. For example, the concept of addition is relationally more complex than the concept of set cardinality (i.e. the use of a natural number to represent the number of elements in a set). Preschool children learn the concept of set cardinality, while school-age children learn the concept of addition which entails determining the cardinality of the union of two sets. Similarly, the concept of multiplication is relationally more complex than the concept of addition because multiplication is defined as repeated addition. It may be that when children reach a level of development at which they are capable of attending to some relations in the context of extraneous information, they may not be able to attend to more complex relations in the presence of the same extraneous information. As a result, the acquisition of more complex relations from concrete instantiations may be more susceptible to diverted attention than acquisition of simpler relations. Therefore, we propose that concreteness in the presence of more complex relations, such as arithmetic relations, can hinder knowledge acquisition in comparison to more generic instantiations of the same concepts.

## Overview

The purpose of the present research was to test the hypothesis that concreteness of the learning material will hinder young school-aged children's acquisition of mathematical knowledge. The present study examined initial learning and subsequent transfer of basic fraction knowledge when instruction involved either a concrete, perceptually rich instantiation versus a generic, perceptually sparse instantiation. First-grade students were taught to label proportions of discrete objects with fractions. Transfer was measured as students' ability to label proportion of novel objects with fractions. For half the participants, transfer was tested immediately after instruction. For the other half of participants, transfer was tested after a two-week delay.

## Experiment

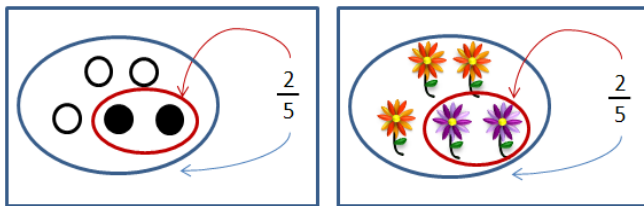
### Method

**Participants** Participants were 64 first-grade students recruited from middle-class, suburban schools in the Columbus, Ohio area (34 girls and 30 boys,  $M = 7.3$  years,  $SD = .40$  years).

**Materials and Design** The experiment had a 2 (Learning condition: Concrete vs. Generic) by 2 (Transfer Test: Immediate vs. Delayed) between-subjects design. Participants were randomly assigned to one of the two learning conditions and one of the two transfer test times. The timing of the transfer test was a between-subject factor to control for any potential testing effects on delayed transfer.

The task was to label proportions of discrete objects with fractions. The experiment had two phases. The first phase consisted of training and a test of learning. Training consisted of four examples of how to label a proportion of objects with a fraction, followed by six questions with corrective feedback. In the Generic condition, all training examples were proportions of black circles out of black and white circles. In the Concrete condition, all training examples were proportions as purple flowers out of purple and orange flowers. Figure 1 presents one of the examples used in training for both the Generic and Concrete conditions.

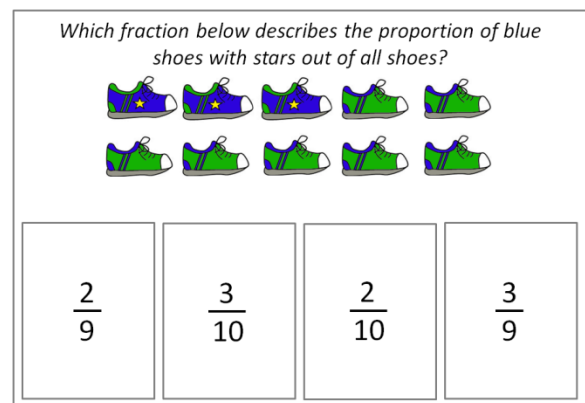
Following training, participants were given an eight-question test of learning which presented novel proportions in the same format as the training (i.e. circles for the Generic condition and flowers for the Concrete condition). Questions were multiple-choice. Four questions presented a proportion and participants were asked to select a fraction that described the proportion. The remaining four questions presented a fraction and participants were asked to select a collection of objects for which the proportion matched the fraction. Four response choices were given: (1) the correct response, (2) correct numerator, but incorrect denominator, (3) correct denominator, but incorrect numerator, and (4) incorrect numerator and incorrect denominator. The order of the answer choices was counterbalanced across question trials.



**Figure 1:** Example of labeling a proportion with a fraction from the training phase (Generic condition on left, Concrete condition on right).

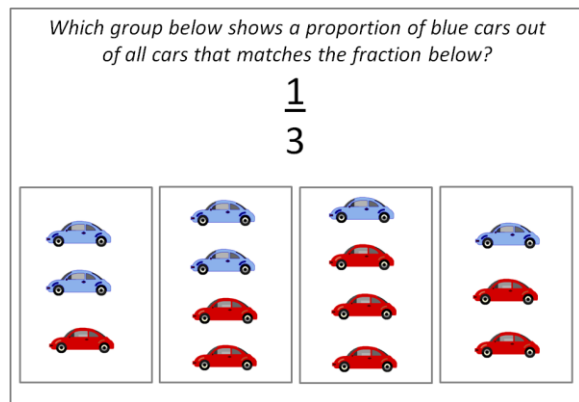
The second phase of the experiment was a transfer task in which participants were given 24 multiple-choice questions involving novel objects. For half of the participants the transfer test was given immediately after phase 1 (i.e. training and testing of learning) and for the other half of participants the transfer test was given two weeks after phase 1. Twelve questions presented a collection of objects and participants selected a fraction that described a specified proportion of the objects (see Figure 2); the other twelve questions presented a fraction and participants selected a corresponding collection of objects that showed a proportion matching the fraction (see Figure 3). For each question, there were four possible response choices, as described for the training questions. The questions involved proportions with denominators (i.e. total number of items in a display) ranging from 2 to 10. Many different items were used for response choices and included: red and blue cars, blue and green shoes, red and green fish, green and red bugs, bears with and without flags, cupcakes with and without sprinkles, slices of pizza (present or missing), light windows and dark windows of a house, partially full bus seats, partially full pencil box, partially full paint bucket and partially remaining chocolate bar.

**Procedure** All training and test questions were presented on the computer. During training, the experimenter gave a definition of proportion and explained that fractions can describe proportions. For example, in the Generic condition when showing the example of  $2/5$  (see Generic condition of Figure 1), the experimenter stated while gesturing to the circles, “The proportion of black circles in this group is  $2/5$  because there are five circles all together, 1, 2, 3, 4, 5, and two of them are black, 1, 2”. Explanations in the Concrete condition were completely isomorphic to those of the Generic condition. Participants proceeded through the test questions at their own pace. The experimenter recorded their responses through the computer.



**Figure 2:** Example of a transfer test question for which participants needed to choose a fraction that matched the proportion shown.





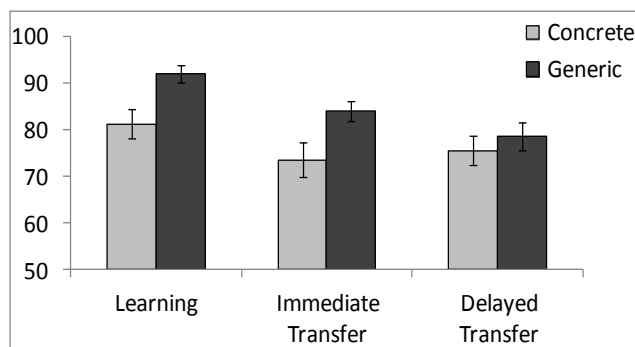
**Figure 3: Example of a transfer test question for which participants needed to choose a proportion that matched the fraction shown.**

## Results

Two participants, one from the Concrete Delayed condition and one from the Generic Immediate condition, were removed from the analysis because their learning scores were more than 2.5 standard deviations below the mean score in their conditions.

In both the Concrete and Generic conditions, children successfully learned. Learning scores in both conditions were well above a chance score of 25% (see Figure 4), one-sample *t*-tests,  $t_s > 18.0$ ,  $p_s < 0.001$ . However, participants in the Generic condition scored significantly higher than those in the Concrete condition ( $M = 92.1\%$ ,  $SD = 10.1\%$  for Generic and  $M = 81.2\%$ ,  $SD = 17.7\%$  for Concrete), independent samples *t*-test,  $t(60) > 2.93$ ,  $p < .006$ , Cohen's  $d = .752$ .

Transfer test scores were also above chance in both learning conditions (Concrete and Generic) and transfer test time conditions (Immediate and Delayed) (see Figure 4), one-sample *t*-tests  $t_s > 13.0$ ,  $p_s < 0.001$ . However, there was a significant difference in transfer scores between the learning conditions on the immediate transfer test, independent samples  $t_s(29) > 2.39$ ,  $p < .024$ , Cohen's  $d = .895$ . Participants in the Generic condition scored higher than those in the Concrete condition. The difference in transfer scores attenuated considerably on the delayed transfer test, independent samples  $t(29) = .704$ ,  $p = .487$ , Cohen's  $d = .200$ . An analysis of variance was performed with transfer test score as the dependent variable, learning condition and transfer test time as fixed factors and learning score as a covariate. The results reveal a significant effect of learning score,  $F(1, 57) = 24.2$ ,  $p < .001$ ,  $\eta_p^2 = .298$ , and no significant effect of learning condition,  $F(1, 57) = .366$ ,  $p = .548$ , and no effect of transfer time,  $F(1, 57) = .815$ ,  $p = .371$ . There was a moderate interaction between learning condition and transfer time  $F(1, 57) = 2.90$ ,  $p = .094$  (see Figure 4). For both immediate and delayed testing, transfer scores were



**Figure 4: Mean Test Scores (% Correct).**

Note: Error bars represent standard error of mean.  
Chance score is 25%.

positively correlated with learning scores, Pearson Correlations,  $r(29) = .600$ ,  $p < .001$  and  $r(29) = .554$ ,  $p < .002$  respectively. These findings suggest that transfer is a function of learning such that higher levels of learning result in higher levels of transfer.

Taken together the results of this experiment suggest that learning basic fraction knowledge from a concrete instantiation, in comparison to a more generic instantiation, can hinder initial learning which may in turn hinder subsequent transfer to novel material. With time, the negative effect of concreteness on transfer appears to be attenuated.

## General Discussion

This research considered first-grade children's ability to acquire basic knowledge of the concept of fraction. Participants were instructed on how to label proportions of objects with fractions. Instruction presented proportions either through generic black and white shapes or through colorful, familiar objects. Participants were tested on their ability to label proportions of novel objects with fractions. Participants who received instruction with either type of material successfully learned and applied their knowledge to novel objects. However, those who were instructed with the generic instantiation scored 10% higher on tests of learning and immediate transfer to novel objects than those who were instructed with the concrete instantiation. The difference in transfer scores due to instruction with the concrete versus generic learning material diminished when the transfer test was delayed for two weeks. Yet both immediate and delayed transfer test scores were strongly correlated with learning scores.

The results of this study support the hypothesis that concreteness of the learning material can hinder children's acquisition of mathematical knowledge. In particular, it appears that instruction involving concrete instantiations of proportions hinders initial learning and consequently hinders subsequent transfer in comparison to instruction involving generic instantiations of proportion. Instruction of basic fraction notation using generic material may help students gain a solid knowledge foundation which may in



turn benefit them when learning more advanced mathematical concepts involving fractions. Although concrete instantiations are often colorful and visually appealing, bland, generic instantiations are learnable by children and can offer an advantage for learning and subsequent transfer of mathematical knowledge.

These findings suggest that although aspects of executive function, including the ability to control attentional focus and inhibit irrelevant information, mature considerably in the preschool years, extraneous information included in concrete educational material may be difficult for elementary school children to ignore. Learning of mathematical concepts may be hindered because less attentional resources have been allocated to the relevant to-be-learned relations.

With respect to actual pedagogical practice, mathematics instruction is generally not limited to using only one instantiation of a concept and frequently involves multiple instantiations, including formal symbolization as well as familiar contextualization. Concrete and abstract instantiations of mathematical may both have advantages. However, it is not clear a priori when and how to include concrete instantiations and generic instantiations in instruction. For example, there appears to be a trade-off between grounded, concrete instantiations and abstract, symbolic instantiations when solving algebra problems where grounded, concrete formats facilitate solving simple problems and abstract, symbolic formats facilitate solving more complex problems (Koedinger, Alibali, & Nathan, 2008; Koedinger & Nathan, 2004). The results of the present study provide evidence of an advantage for generic material for acquiring knowledge of basic fraction notation. The challenge for researchers and educators is to develop a theoretical basis for the timing and use of both concrete and generic instantiations in instruction of mathematical concepts in general.

## Acknowledgments

This research is supported by a grant from the Institute of Educational Sciences of the U.S. Department of Education (#R305B070407.) to V. M. Sloutsky and J. A. Kaminski.

## References

- Bruner, J. S. (1966). *Toward a theory of instruction*. Cambridge Mass: Harvard University Press.
- DeLoache, J. S. (1995a). Symbolic functioning in very young children: Understanding of pictures and models. *Child Development*, 62, 736-752.
- DeLoache, J. S. (1995b). Early understanding and use of symbols: The model model. *Current Directions in Psychological Science*, 4, 109-113.
- Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, 59, 47-59.
- Gentner, D., Ratterman, M. J., Markman, A., & Kotovsky, L. (1995). Two forces in the development of relational similarity. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence* (pp. 263-314). Hillsdale, NJ: Erlbaum.
- Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, 23, 222-264.
- Goldstone, R. L. & Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology*, 46(4), 414-466.
- Goldstone, R. L. & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the Learning Sciences*, 14, 69-110.
- Hanania, R., & Smith, L. B. (2010). Selective attention and attention switching. *Developmental Science*, 13, 622-635.
- Kaminski, J. A., & Sloutsky (2009). The effect of concreteness on children's ability to detect common proportion. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the XXXI Annual Cognitive Science Society*. Amsterdam, the Netherlands: Cognitive Science Society.
- Kaminski, J. A., & Sloutsky, V. M. (2010). Concreteness and relational matching in preschoolers. In S. Ohlsson, & R. Catrambone (Eds.), *Proceedings of the XXXII Annual Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Kaminski, J. A., & Sloutsky, V. M. (2011). Representation and transfer of abstract mathematical concepts in adolescence and young adulthood. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, & J. Confrey (Eds.), *The adolescent brain: Learning, reasoning, and decision making* (pp. 67-93). Washington, DC: American Psychological Association.
- Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math. *Science*, 320, 454-455.
- Kemler, D. G. (1982). Cognitive development in the school years: Foundations and directions. In J. Worrell (Ed.), *Psychological Development in the Elementary School Years* (pp. 233-268). New York: Academic Press.
- Koedinger, K. R., Alibali, M. W., & Nathan, M. J. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. *Cognitive Science*, 32, 366-397.
- Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representation on quantitative reasoning. *Journal of the Learning Sciences*, 13, 129-164.
- McNeil, N. M., & Uttal, D. H. (2009). Rethinking the use of concrete material in learning: Perspectives from development and education. *Child Development Perspectives*.
- Montessori, M. (1917). *The advanced Montessori method*. New York: Frederick A. Stokes.

- Piaget, J. (1970). *Science of education and the psychology of the child*. New York: Orion Press.
- Shepp, B. E., & Swartz, K. B. (1976). Selective attention and the processing of integral and nonintegral dimensions: A developmental study. *Journal of Experimental Child Psychology*, 22, 73-85.
- Sloutsky, V. M., Kaminski, J. A., & Heckler, A. F. (2005). The advantage of simple symbols for learning and transfer. *Psychonomic Bulletin & Review*, 12, 508-513.
- Smith, L. B., & Kemler, D. G. (1978). Levels of experienced dimensionality in children and adults. *Cognitive Psychology*, 10, 502-532.

# The Role of Imagination in Augmenting Perceptual Representation

**Seokmin Kang (sk2587@columbia.edu)**

Teachers College, Columbia University, 525 W. 120<sup>th</sup> Street  
New York, NY 10027 USA

**Gregory L. Hallman Jr. (glh2103@columbia.edu)**

Teachers College, Columbia University, 525 W. 120<sup>th</sup> Street  
New York, NY 10027 USA

**John B. Black (jbb21@columbia.edu)**

Teachers College, Columbia University, 525 W. 120<sup>th</sup> Street  
New York, NY 10027 USA

## Abstract

Knowledge construction requires both perceptual information from external sources and our active interpretation of that information. Thirty one 5<sup>th</sup> grade elementary school students were asked to move plates in a Tower of Hanoi (TOH) task, which was displayed on a screen monitor. When the students were asked to respond to the weight of the plates, both imagination and non-imagination groups reported that they felt weight of the plates which actually had no weight. Also, students in imagination group reported that they felt the plates heavier than did those in non-imagination group. The result shows that haptic information can be created without providing any information through haptic channel. In addition, the result implies how knowledge is constructed based on previous knowledge and suggests that children create their own imaginary world even at a perceptual level. Also, it was discussed why imagination experience can maximize and help learning in embodiment activities.

**Keywords:** Imaginary world; Haptic illusion; Embodied cognition.

In Buddhism, it is said that “Everything depends on our mind.” This sentence can be interpreted as humans’ perceiving an object, or a given environment, depending on what we expect to perceive or interpret. According to Ungerleider and Pasternak (2004), when constructing knowledge we not only accept spatial and visually-guided action information (dorsal process stream), but process information of shape, form, and object identity (ventral process stream). This insight has been scientifically tried in studies about illusion in perception (Lee & van Donkelaar, 2002) and in memory (Roediger, 1996; Roediger & McDermot, 1995). Also, especially according to optical illusion studies, what we see does not always match what is actually visible. Our visual perception is influenced by our past experience and current expectations (Gregory, 1997; Gregory, 1998; Pendlebury, 1996).

According to Kant (1965), our knowledge is elicited from two fundamental sources of the mind: the capacity of receiving representations and the power of knowing an object through these representations. Intuition and concepts constitute the elements of all our knowledge, so that neither concepts without an intuition in some way corresponding to

them, nor intuition without concepts, can yield knowledge (Kant, p.92). Also, Kant mentioned that “...all perceptions are grounded a priori in pure intuition, association in pure synthesis of imagination, and empirical consciousness in pure apperception...” (Kant, p.141). This also can be interpreted as that perceptual information received via different modalities, becomes knowledge by our active cognitive process. This can be called knowledge construction. Therefore, by using this nature of human perception, it is possible to create perceptual illusions. In terms of knowledge construction from perceptual information, while previous studies have dealt with people’s creation of additional perceptual information or distortion based on what was given (Day, 1990; Ellis & Lederman, 1993; Fermüller & Malmb, 2004; Mack, Heurer, Villardi, & Chambers, 1985; Suzuki & Arashida, 1992; Otto-de Haart, Carey, & Milne, 1999), this study seeks to investigate knowledge construction from a certain activity which enriches memory representation. In other words, we argue that perceptual knowledge can still be created through another perceptual channel even if no perceptual information has been delivered.

While many studies have explored visual illusion, what has rarely been investigated is how perceptual information from other modalities is created. Furthermore, while the literature offers a handful of studies about haptic illusion (Gentaz, Camos, Hatwell, & Jacquet, 2004; Robles-De-La-Torre & Hayward, 2001), these studies have limitation that they were administered under situations where haptic information was already provided to participants.

## Knowledge construction by imagination

It is assumed that when we process information from text, we go through a perceptual process where spoken or written messages are originally encoded (Golden, 1986; Hubel & Wiesel, 1977; Kuffler, 1953; McClelland, 1976; Syrdal & Gopal, 1986). Then the words in the message are transformed into a mental representation of the combined meaning of the words (Graf & Torrey, 1966). At higher level of comprehending of a text, readers use the mental representation of the sentence’s meaning. According to

Kintsch (1998), this level of understanding adds inferences made from the reader's knowledge and these inferences come from semantic memory, knowledge about objects, events, people, goals, beliefs, etc. However, Black and Bower (1980) argued that there is another level of text comprehension, which they called the Story World because they were only discussing stories. This level of comprehension requires a representation of the world being referred to in a story, how it is laid out visually and spatially, and how it functions. We can generalize this beyond stories by calling this level of comprehension and representation an Imaginary World. For example, after reading one or more Harry Potter fantasy books, if one has reached this "imaginary world" level of comprehension, one would have some ideas about Harry Potter's world, such as how it looks, how it is laid out, and how it functions (information which actually may not be provided in the books). Then, when seeing a Harry Potter movie, one can judge to what extent it matches this Imaginary World. Black, Turner, and Bower (1979) already showed evidence of readers creating their own imaginary world when reading a story. In their study, when people read sentences like *John was working in the front yard, then he went inside*, they read them faster than when they read *John was working in the front yard, then he came inside*. The 'came' in the second sentence creates a switch in point of view and causes a longer reading time (and change in memory and change in comprehensibility ratings) because the reader has to switch the perspective from which they are watching the action in the imaginary world.

In imaginary world, people make plausible inferences or create a story that was not actually given. This is further activated when a goal is added, which results in a more coherent, interesting and memorable text. For example, Owens, Bower and Black (1979) had one group read an unsurprising script-based story just containing the expected actions of scripts, while another group had an introductory sentence that introduced a goal structure (the girl in the story wondering if she were pregnant). Adding this introduction, which allowed the readers to infer a goal structure with many linked story statements, leads to better recall of the story. Black (2008) took this a step further when he mentioned that learners can construct elaborate mental representations of Imaginary Worlds with minimal input such as text plus a floor plan. This implies that learners create imaginary worlds based on given information, place themselves into that situation, and retrieve a plausible story by filling out story gaps that are not given to them.

These findings can be linked to research on embodied cognition. Barsalou (1999) supports the idea that thoughts stem from dynamic interactions between the physical world and the body. These interactions, described as part of "grounded cognition," are not limited to bodily states, but are also found in situated action, social interactions, emotional states, the environment, and perceptual simulations. In this sense, active interaction can also be

realized in imaginary world, since one can situate himself in a given situation and enact what is given in that imaginary world. According to proponents of embodied cognition, acknowledging these interactions is the first step towards understanding human cognition.

Research on embodied cognition reveals that when people place their bodies and move around in physical space, they create references for understanding concepts. This is evident in a study by Glenberg and Kaschak (2002). The authors found that subjects, after hearing the command "open the drawer", responded faster when the response action was a pulling motion instead of a pushing one. Likewise, work by Spivey et al. revealed that during an experiment that measured eye movement, participants were more likely to look up upon hearing a description about the top of a building (Spivey, Tyler, Richardson, & Young, 2000).

These discoveries about human cognition and space are complimented by the seminal research of Shepard and Metzler on mental rotation and imagery in the 1970s. The researchers showed participants images of 3-D geometric figures. The participants were later shown similar yet rotated images and were asked to interpret the uniformity, or lack thereof, of the two images. Results revealed that participants rotated visual mental images in much the same way that the actual objects would be manipulated in an actual physical space. The research also revealed a proportional relationship between the degree of the rotation and the time it took participants to mentally rotate and respond (Shepard & Metzler, 1971).

Glenberg and colleagues have found that physical manipulation and imagined manipulation help children learn (Glenberg, Gutierrez, Levin, Japuntich, & Kaschak, 2004). Specifically these researchers have conducted experiments using a narrative of a farmer and his farm animals. The narrative is accompanied by a set of toys in the likeness of the story's characters. Participants (1<sup>st</sup> and 2<sup>nd</sup> graders) were asked to use the toys to re-enact the sentences they read. The children are then asked to imagine manipulating the toys. What resulted was an increase in reading comprehension and a more positive attitude about reading (Glenberg et al., 2004).

In the same vein, researchers on grounded cognition suggest that how people act influences how they think by grounding perception, affect, and even language comprehension in the sensorimotor systems used to interact with the surrounding world (Beilock, Lyons, Mattarella-Micke, Nusbaum, & Small, 2008; Glenberg & Robertson, 2000; Niedenthal, 2007; Zwaan, 1999). For example, learning to produce specific walking movements (without visual feedback) aids one's ability to later visually discriminate these movements, presumably because discrimination becomes tied to the sensorimotor systems used in moving (Casile & Giese, 2006). Therefore, perception and action lead to conceptual understanding. In addition, this process requires a representation of what is being referred to, how it is laid out visually and spatially,

which finally leads to the creation of “Imaginary Worlds” (Black, 2008).

Based on this “Imaginary World” concept, this study starts with a question about if learners can create their own imaginary world on a perceptual level. For example, readers develop their own mental images of a story while reading the story (Black et al., 1979) and learners can also build their own mental representation while learning from either a text, a text with floor plans plus static pictures, or a text with floor plans accompanied by a virtual tour picture (van Esselstyn & Black, 2001). In this study, we try to test that creating an imaginary world is possible on a perceptual level, which, in turn, enables us to see why imagination activity helps children more involved in embodiment activity.

So far, few empirical studies have been done to test the role of imagination in children’s perception and knowledge construction. Being that this study seeks to encourage children to create perceptual information which was not initially given to them, this study is different from previous perceptual illusion studies that showed how perceptual information is distorted or how illusion is created by sensory information which is transmitted via the same modality. Throughout this study, we try to observe how children, while in the process of constructing knowledge, represent knowledge which does not exist through modality and synthesize it into knowledge.

## Method

**Participants** Thirty-one (18 boys and 13 girls) 5<sup>th</sup> grade elementary school students from the New York City public schools district participated in this study and were divided into two groups: imagination group ( $n = 15$ ) and non-imagination (control) group ( $n = 16$ ). The study was IRB approved.

Children of this age were chosen because they were almost at the end of concrete operational stage. Therefore, they are expected to use the appropriate level of inductive logic, such as transitivity, decentering, ability to eliminate egocentrism, and etc. (Piaget, 1977).

**Apparatus** The Kinect sensor for the Microsoft Xbox 360 video game console (hereafter referred to as Kinect) was connected to an HP laptop through which an interactive Tower of Hanoi problem was displayed onto a Dell 19 inch LCD monitor with speakers attached. Kinect is a motion sensing input device by Microsoft for the Xbox 360 video game console. Based around a webcam-style add-on peripheral for the Xbox 360 console, it enables users to control and interact with the Xbox 360 without the need to touch a game controller, through an infrared natural user interface using gestures and spoken commands. However, in this study, we manipulated Kinect without Xbox 360. As a way of communicating with the Kinect, FFAST (Flexible Action and Articulated Skeleton Toolkit, version 0.07) software was used. FFAST is middleware which facilitates integration of full-body control with games and VR

applications. The toolkit relies on software from OpenNI and PrimeSense to track the user’s motion using the PrimeSensor or the Microsoft Kinect sensors. FFAST includes a custom VRPN server to stream the user’s skeleton over a network, allowing VR applications to read the skeletal joints as trackers using any VRPN client. Additionally, the toolkit can also emulate keyboard input triggered by body posture and specific gestures. This allows the user add custom body-based control mechanisms to existing off-the-shelf games that do not provide official support for depth sensors (Suma, Lange, Rizzo, Krum, & Bolas, 2011). Also, the interface of the TOH game was programmed using Scratch. Scratch is a programming language developed by the MIT Media Lab Lifelong Kindergarten Group.

**Experimental Design and Procedure** The experiment was administered in a quiet classroom with one student at a time. A student was guided and asked to stand in front of the Kinect and a screen monitor that was connected to a laptop computer. Firstly, the experimenter verbally explained how to move the plates in the TOH problem set on the screen. The children were told: “Your job is to move two plates from the first pole to the third pole. You can move only one plate at a time. Also, while moving the plate, you can put the small plate on top of the large one, but you cannot put a large plate on top of the small one.” The plates in TOH were created to look like real steel (Figure 1).

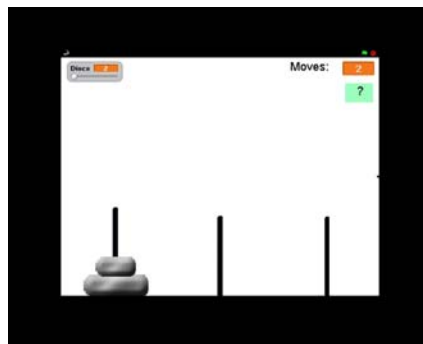


Figure 1. Snapshot of TOH problem used in the task

Afterwards, the experimenter demonstrated how to move plates on the screen. The student was then allowed to participate in a practice session of moving plates from one pole to another until he/she felt comfortable controlling the game using only hand gestures. Students were then given a challenge of two plates for the TOH problem. Unlike the practice session, in the experiment session, audio speakers attached to the LCD monitor were turned on, so as to provide students with the sound of plate being dropped (in this case, the sound of a heavy metal plate dropping to the ground).

Figure 2 shows a student moving a plate in the problem-solving task. All 31 students completed the given task. During the experimental session, children completed the TOH problem by carrying plates without touching any

objects such as a monitor screen, a keyboard, or a laptop computer. Immediately after they solved the problem, short survey questions were given.

All the procedures were identical for both imagination group and non-imagination group except that at this stage, students in imagination group were prompted to imagine how they moved the plates before answering how heavy the plates were. Non-imagination group received exactly the same survey except an imagination prompt, “Imagine how you moved the plates.” in question 1.

As dependent measure variables, students’ responses to the survey questions were compared. Students were asked to choose a number that best described what they felt while playing the game. The survey intended to measure the perceived weight the participants felt from moving the virtual plates, the level of interest in the activity, and the amount of difficulty encountered in controlling this Kinect-based embodiment tool. As they answered the questions, the experimenter read aloud each question and response options (i.e., labels of a Likert-scale). After the survey, students were sent back to their class. Totally participation time for one student amounted to about 15 minutes for the entire session.



Figure 2. A student solving Tower of Hanoi problem.

## Results and Discussion

Students’ responses in two groups for plate weight, level of motivation, and amount of difficulty in controlling the plates were compared and analyzed.

First, to find out whether students in each group felt the weight of the plate or not, one sample t-test was administered for responses asking how heavy the plate felt. In the analysis, it was found that students in both imagination group ( $M = 2.87$ ,  $SD = 1.36$ ,  $t(14) = 5.33$ ,  $p < .01$ ) and non-imagination group ( $M = 1.94$ ,  $SD = 1.12$ ,  $t(15) = 3.34$ ,  $p < .01$ ) reported that the virtual plate they moved indeed had a tangible weight.

Also, students in both groups reported that it was relatively easy to use their hands to move the plates. Average response was 2.27 ( $SD = 1.22$ ) in imagination group and 2.81 ( $SD = 1.47$ ) in non-imagination group. In

addition, students in both groups showed strong interest in the game task. All students in imagination group responded that they enjoyed the game very much ( $M = 5.00$ ,  $SD = 0.00$ ). Similarly, average response in non-imagination group was 4.94 ( $SD = 0.25$ ).

In group comparison analysis, independent sample t-test was administered. The students in imagination group felt the plates significantly heavier than non-imagination group,  $t(29) = 2.08$ ,  $p < .05$ ,  $d = 0.75$ . Figure 3 shows response averages of feeling of plates’ weight in two groups.

In terms of students’ interest, there was no difference between two groups ( $p = .34$ ). Also, in a question of asking how easy it was to use hands to carry the plates, there was no group difference in difficulty of carrying the plates ( $p = .27$ ).

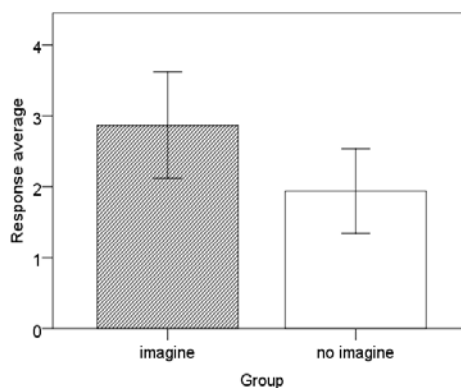


Figure 3. Response average of feeling of weight. Error bars represent standard errors of the means.

This study shows how children develop and organize perceptual information, and create their own imaginary world based on an experience. Surprisingly, children in both groups felt the weight of plates which actually had no weight. In this case, it would be a haptic illusion that falls at the intersection of perceiving and remembering.

It is even more surprising that children in imagination group weighted the plates heavier than non-imagination group. It is assumed that the memory of what plates would feel like was drawn into the procedure as expectancies and then affected performance, and the feeling of weight would be augmented by the imagination employed during the embodied activity. It may be like the case when someone thinks he is about to pick up a heavy object (e.g., a brick) but then he lifts it way too high too fast because it turns out to be light (e.g., styrofoam). According to Winograd, Peluso, and Glover (1998), self-reports of high degrees of vivid mental imagery correlate with enhanced false recall and false recognition. In our study, it is possible that, with a prompt, “imagine,” students in imagination group further developed haptic information than those in non-imagination group. As a result, children in both groups reported the weight of plates by activating what they expected to feel. This is realistic when considering that upon judging a situation, value, or an object, individuals rely more on

reasoning, which, in turn, is based on their prior knowledge or experience. Nonetheless, participants should not have responded that the plate had a weight in this situation.

What is interesting is that, even though the activities in this study were obvious and physically concrete, children in this study reported their haptic experience which was not provided to them. Based on our observations, it seems that when it comes to deciding the characteristics of a given object, children evolve into relying more on their own representation of the concept which was created by synthesizing subjective sensation and representation, rather than relying on pure sensation.

Another plausible explanation is that children might mistake a plate's supposed weight with the feeling of fatigue from arm movement caused by moving. Yet, considering that controlling plates was relatively easy and that activity was enjoyable, it is assumed that illusion of a plate's purported weight was created by the embodied activity and it was further augmented by the participants' imagination.

In our case, with imagination, children who acted out using their body added new perceptual information, creating their own representation of the TOH game and the plates in the task. This also corresponds to grounded cognition studies. It is maintained that new representation is based on individual actions and action plays an important role in the emergence of new representations (Beilock & Goldin-Meadow, 2010; Boncoddio, Dixon, & Kelley, 2010).

Children used information available to fill the knowledge gap with information that was not actually provided. In other words, as Black and Bower (1980) mentioned, children's representation of an object when derived from an imagined activity has information implicit in it beyond that which is available in the propositional listing of the scene. It is assumed that embodied cognition in a learning environment must first be physically enacted. Then, the learning activity is maintained through imagined embodiment, and is finally used within a task where transfer of learned content can occur.

Again, this study showed that by having children imagine, unsupplied information can be derived not only in text understanding, but also in perceiving an object's property which was not provided via any sensory channel. The study result implies that having children imagine activities that they involved or acted out can enrich their experience. The phenomenon we observed here has many implications of how teachers can use imagination in teaching concepts, such as ones that have information which is implicit or hard to be captured with a visually aided material.

Also, in our study, to observe how children construct knowledge from perceptual information, interactive media technology was used. Interactive media technology carries with it the benefit of enhancing students' comprehension across several frames of reference because of its ability to present information in varying ways. This ensures success for all types of learners: visual, active, auditory, and tactile. There are many technologies at use today that take

advantage of these varying styles of presentation and each caters to one or several simultaneous types of learners at once. With a help from these characters of technology, we could capture the moment at which children construct knowledge from perceptual information. The motivation observed in this study is also unprecedented and will no doubt enlighten pedagogical methods in future classrooms.

In future studies, by investigating how information in other modalities affects level of embodiment, we will learn more about the role of imagination and the interaction of imagination and perceptual channel in knowledge construction.

## References

- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral & Brain Sciences*, 22, 577-660.
- Beilock, S. L., & Goldin-Meadow, S. (2010). Gesture changes thought by grounding it in action. *Psychological Science*, 21, 1605-1611.
- Beilock, S. L., Lyons, I. M., Mattarella-Micke, A., Nusbaum, H. C., & Small, S. L. (2008). Sports experience changes the neural processing of action language. *Proceedings of the National Academy of Sciences, USA*, 105, 13269-13273.
- Black, J. B. (2008). Imaginary Worlds. In M. A. Gluck, J. R. Anderson, & S. M. Kosslyn (Eds.), *Memory and Mind: A Festschrift for Gordon H. Bower* (pp. 195-208). Mahwah, NJ: Lawrence Erlbaum Associates.
- Black, J. B., & Bower, G. H. (1980). Story understanding as problem-solving. *Poetics*, 9, 223-250.
- Black, J. B., Turner, T. J., & Bower, G. H. (1979). Point of view in narrative comprehension, memory, and production. *Journal of Verbal Learning and Verbal Behavior*, 18, 187-198.
- Boncoddio, R., Dixon, J. A., & Kelley, E. (2010). The emergence of a novel representation from action: Evidence from preschoolers. *Developmental Science*, 13, 370-377.
- Bower, G. H., Black, J. B., & Turner, T. J. (1979). Script in memory for text. *Cognitive Psychology*, 11, 177-220.
- Casile, A., & Giese, M. A. (2006). Non-visual motor learning influences the recognition of biological motion. *Current Biology*, 16, 69-74.
- Day, R. (1990). The bourdon illusion in haptic space. *Perception & Psychophysics*, 47, 400-404.
- Ellis, R. R., & Lederman, S. J. (1993). The role of haptic versus visual volume cues in the size-weight illusion. *Perception and Psychophysics*, 53, 315-324.
- Fermüller, C., & Malmb, H. (2004). Uncertainty in visual processes predicts geometrical optical illusions. *Vision Research*, 44, 727-749.
- Gentaz, E., Camos, V., Hatwell, Y., & Jacquet, A-Y. (2004). The visual and the haptic Müller-Lyer illusions: Correlation Study. *Current Psychology Letters* [Online], 13, URL:<http://cpl.revues.org/index431.html>
- Glenberg, A. M., Gutierrez, T., Levin, J. R., Japuntich, S., & Kaschak, M. P. (2004). Activity and imagined activity can



- enhance young children's reading comprehension. *Journal of Educational Psychology*, 96, 424-436.
- Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9, 558-565.
- Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory & Language*, 43, 379-401.
- Golden, R. (1986). A developmental neural model of visual word perception. *Cognitive Science*, 10, 241-276.
- Graf, P., & Torrey, J. W. (1966). Perception of phrase structure in written language. *American Psychological Association Convention Proceedings*, 83-88.
- Gregory, R. L. (1997). Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London*, 352, 1121-1128.
- Gregory, R. L. (1998). Brainy mind. *British Medical Journal*, 317, 1693-1695.
- Hubel, D. H., & Wiesel, T. N. (1977). Functional architecture of macaque monkey visual cortex. *Philosophical Transactions of the Royal Society of London*, 198, 1-59.
- Kant, I. (1965). *Critique of pure reason* (N. K. Smith, Trans.). New York: Macmillan. (Original work published 1781)
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.
- Kuffler, S. W. (1953). Discharge pattern and functional organization of mammalian retina. *Journal of Neurophysiology*, 16, 37-68.
- Lee, J. H., & van Donkelaar, P. (2002). Dorsal and ventral visual stream contributions to perception-action interactions during pointing. *Experimental Brain Research*, 143, 440-446.
- McClelland, J. L. (1976). Preliminary letter identification in the perception of words and nonwords. *Journal of Experimental Psychology: Human Perception and Performance*, 2, 80-91.
- Mack, A., Heuer, F., Vilardi, K., & Chambers, D. (1985). The dissociation of position and extent in Muller-Lyer figures. *Perception and Psychophysics*, 37, 335-344.
- Niedenthal, P. M. (2007). Embodying emotion. *Science*, 316, 1002-1005.
- Otto-de Haart, E. G., Carey, D. P., & Milne, A. B. (1999). More thoughts on perceiving and grasping the Müller-Lyer illusion. *Neuropsychologia*, 37, 1437-1444.
- Owens, J., Bower, G. H., & Black, J. B. (1979). The "soap opera" effect in story recall. *Memory and Cognition*, 7, 185-191.
- Pendlebury, M. (1996). The role of imagination in perception. *South African Journal of Philosophy*, 15, 133-138.
- Reder, L. M. (1982). Plausible judgment versus fact retrieval: Alternative strategies for sentence verification. *Psychological Review*, 89, 250-280.
- Robles-De-La-Torre, G., & Hayward, V. (2001). Force can overcome object geometry in the perception of shape through active touch. *Nature*, 412, 445-448.
- Roediger, H. L. (1996). Memory illusions. *Journal of Memory and Language*, 35, 76-100.
- Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 803-814.
- Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171, 701-703.
- Spivey, M. J., Tyler, M. J., Richardson, D. C., & Young, E. E. (2000). Eye movements during comprehension of spoken scene descriptions. The Proceedings of the 22<sup>th</sup> Annual Cognitive Science Society Meeting, 487-492.
- Suma, E., Lange, B., Rizzo, A., Krum, D., & Bolas, M. (2011). FFAST: The Flexible Action and Articulated Skeleton Toolkit. *Proceedings of IEEE Virtual Reality 2011*.
- Suzuki, K., & Arashida, R. (1992). Geometrical haptic illusions revisited: Haptic illusions compared with visual illusions. *Perception and Psychophysics*, 52, 329-335.
- Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, 79, 1086-1100.
- Ungerleider, L. G., & Pasternak, T. (2004). Ventral and dorsal cortical processing streams. In L.M. Chalupa, & J. S. Werner (Eds.), *The Visual Neurosciences* (pp. 541-562). MIT Press, Cambridge, MA.
- van Esselstyn, D., & Black, J. B. (2001). Learning through interactive panoramic imagery. Proceedings of ED-MEDIA. Norfolk, VA: AACE.
- Winograd, E., Peluso, J. P., & Glover, T. A. (1998). Individual differences in susceptibility to memory illusions. *Applied Cognitive Psychology*, 12, S5-S27.
- Zwaan, R.A. (1999). Situation models: the mental leap into imagined worlds. *Current Directions in Psychological Science*, 8, 15-18.

# The effects of amnesia on driving performance in elderly drivers

**Naoko Kawano (n-kawano@med.nagoya-u.ac.jp)**

Department of Psychiatry, Graduate School of Medicine,  
Nagoya University, Japan.

**Kunihiro Iwamoto (iwamoto@med.nagoya-u.ac.jp)**

Department of Psychiatry, Graduate School of Medicine,  
Nagoya University, Japan.

**Kazutoshi Ebe (whitejam@nifty.ne.jp)**

Toyota Central R&D Labs., Inc., Japan.

**Katsuyuki Ukai (ukai777@kamiida-hp.jp)**

Kamiida daiichi General hospital, Japan.

**Yusuke Suzuki (yus@med.nagoya-u.ac.jp)**

Department of Geriatrics, Graduate School of Medicine,  
Nagoya University, Japan.

**Hiroyuki Umegaki (umegaki@med.nagoya-u.ac.jp)**

Department of Geriatrics, Graduate School of Medicine,  
Nagoya University, Japan.

**Tetsuya Iidaka (iidaka@med.nagoya-u.ac.jp)**

Department of Psychiatry, Graduate School of Medicine,  
Nagoya University, Japan.

**Norio Ozaki (ozaki-n@med.nagoya-u.ac.jp)**

Department of Psychiatry, Graduate School of Medicine,  
Nagoya University, Japan.

## Abstract

Cognitive dysfunction caused by some neurodegenerative diseases is associated with an increased risk of traffic accidents. Previous studies have reported inconsistent results for prodromal and early stages of dementia. Few studies have directly compared the effects on driving performance of amnesic subtype of mild cognitive impairment (aMCI) with those by normal aging in elderly drivers. The present study examines the association between cognitive decline and driving ability in elderly drivers with aMCI. The participants were 19 healthy young adults (HYA), 26 healthy elderly adults (HEA), and 12 elderly patients with aMCI. All performed a road-tracking, a car-following, and a harsh-braking task on a driving simulator (DS). Elderly participants also completed cognitive assessment tasks including measures of memory performance. All MCI participants showed a well-defined memory decline, and demonstrated significantly decreased performance on the car-following and road-tracking tasks as compared with the HYA group. However, the aMCI group also demonstrated significantly decreased performance on the car-following task as compared with HEA. In elderly participants, the car-following performance was positive correlated with the score on the Trail Making Test-B. This evidence indicates a difference for driving ability between individuals with symptomatic memory impairment and the

aging-related memory of normal controls. This difference may be associated with flexibility of visual attention and executive function.

**Keywords:** mild cognitive impairment, driving simulator, Trail Making Test, elderly driver, normal aging.

## Introduction

Driving is a specialized and complex action that requires the use of extensive cognitive abilities. Age-related dysfunction of the central nervous systems in people with dementia, such as symptoms caused by Alzheimer's disease, may influence driving. Studies have found that drivers with dementia have 2.5 to 4.7 times the risk of an automobile accident compared to cognitively intact elderly drivers (Cooper, Tallman, Tuokko, & Beattie, 1993; Molnar, Patel, Marshall, Man-Son-Hing, & Wilson, 2006). Fatal motor vehicle crash rates (per miles driven) show a U-shape curve with the highest rates among the youngest and oldest drivers (Hakamies-Blomqvist, Sirén, & Davidse, 2004).

Cognitive limitations not only from some age-related neurodegenerative diseases but also from aging-related normal changes are associated with an increased risk of being involved in a traffic accident. Epidemiological studies demonstrated that elderly patients with mild dementia are

high-risk drivers, as a group, compared with cognitively intact drivers (Man-Son-Hing Marshall, Molnar, & Wilson, 2007). However, there is little evidence showing the difference in the actual driving behaviors of individuals in these two groups (Iverson et al., 2010). A substantial number of patients with a Clinical Dementia Rating Scale (CDR: Morris, 1993) scores of 0.5-1.0, which are in the preclinical or early stage of dementia, were still able to drive safely in an on-road driving test (Brown et al., 2005). Iverson et al. recommend that studies are needed to identify the appropriate predictive factors for risky driving in patient with prodromal and mild dementia, and then to develop a composite system of rating risk for drivers in the early stages of dementia.

“Mild cognitive impairment (MCI)” is conceptualized as a transitional state between normal cognitive aging and clinical dementia, with amnesic and non-amnesic subtypes of MCI have been defined (Petersen & Morris, 2005). The amnesic subtype of MCI (aMCI) is characterized by memory impairment, and is often operationalized as the Clinical Dementia Rating Scale score of 0.5. People with aMCI progress to Alzheimer’s disease (AD) at a rate of 8-15% per year, as compared to the normal aging with a dementia progression rate of 1-2% per year (Petersen et al., 2001). Therefore aMCI is considered a prodromal syndrome of AD, and often precedes the onset of dementia.

Individuals with MCI may be at risk for decline in everyday complex functions, including driving. However, we have limited information about the crash risk and driving behaviors of individuals with clinical MCI (Man-Son-Hing et al., 2007). Recently, Frittelli et al. (2009) reported that MCI had a limited effect on driving performance on a driving simulator, and that AD patients’ unsafe driving behavior was not predicted by their MMSE scores. They compared patients with mild AD and with MCI and age-matched neurologically normal controls. However, it is not clear which cognitive characteristics of individuals with MCI do endorse safer driving performance or which do not.

O’Connor, Edwards, Walley, and Crowe (2011) reported associations between driving behaviors, assessed by a self-report questioners, and classification of MCI. Their sample was a subset of the mobility data ( $n = 2381$ ) in the Advanced Cognitive Training for Independent and Vital Elderly (ACTIVE) study ( $N = 2802$ ), and the subset included 82 individuals with the aMCI, 140 individuals with the non-amnesic subtype of MCI, and 82 individuals with multi-domain subtype of MCI, and normal controls. They investigated psychometrically well-defined MCI at baseline as a predictor in a five years follow-up of changes in driving behaviors. Their results suggested that MCI status predicted declines in driving frequency and increases in driving difficulty. The classification of MCI subtypes predicted different trajectories of changes for driving frequency and driving difficulty. The amnesic and non-amnesic groups showed greater declines in driving frequency than the multi-domain group and the normal controls. The amnesic and non-amnesic groups showed increases for driving difficulty

in common situations, whereas the multi-domain group showed a greater increase for driving difficulty only in complex situations. The aMCI showed decline in driving frequency but did not self-report driving difficulty. The authors discussed these findings from the viewpoint of self-regulation and risk perception in their groups’ driving behaviors. They suggested that these findings may reflect impaired risk perception related to amnesia. Objective assessment data of driving performance is still need to determine actual risk assessment.

In the present study, we considered these remaining problems, and examined individuals with clinical aMCI and age-matched memory-intact individuals and normal young adults to evaluate driving performance using a driving simulator (DS). Although many consider road testing to be the gold standard to evaluate driving competence, road tests are costly and can be dangerous when the driver is incompetent. The DS appears to be a safe and cost-effective method for the objective evaluation of driving performance.

In order (1) to examine how cognitive state may impair driving performance in patients with a prodromal stage of dementia, and (2) to identify cognitive variables explaining the deterioration of driving performance in the aMCI group, we designed a case-control study to compare the driving performance decline between adults with clinical aMCI and elderly adults with an intact memory, based on the performance of normal young adults.

## Method

### Participants

We recruited 19 healthy young adults (HYA: 39.3 years old,  $SD = 6.5$ ), 26 healthy elderly adults (HEA: 70.0 years old,  $SD = 6.1$ ), and 12 elderly patients with aMCI (71.8 years old,  $SD = 7.6$ ). The participants were naïve with regard to this study, and were paid for their participation. All were active drivers with more than 10 years of driving experience. They all had normal or corrected-to-normal vision, and no history of cerebral vascular events.

The Nagoya University Graduate School of Medicine and Nagoya University Hospital ethics review committee approved this study. Written informed consent was obtained from all participant prior to their participation.

All participants were examined by an experienced psychologist who used the same task order. They had no history of psychiatric problem as assessed by the Structured Clinical Interview for DSM-IV (SCID: First, Spitzer, Gibbon, & Williams, 1997). The HYA and HEA individuals were recruited in non-clinical setting, and had no impairment in activities daily living (ADL) and no evidence of dementia on the Clinical Dementia Rating Scale (CDR = 0.0). They showed no evidence of cognitive decline on enrollment screening questionnaire. The patients with aMCI were recruited in clinical setting. All were diagnosed according to the criteria for MCI, provided by the Petersen group. All of the aMCI group had a CDR = 0.5.

**Amnesia confirmation** The confirmation of amnesia was identified using psychometric methods. All elderly participants were assessed using structured neuropsychological tests, including the Mini-Mental State Examination (MMSE: Folstein, Folstein, & McHugh, 1975), to their confirm general cognitive state, and the Logical Memory delayed recall subtests of the Wechsler Memory Scale-Revised (WMS-R: Wechsler, 1987) to confirm memory function.

**Cognitive functions** Additional neuropsychological measures, which have been used in previous studies related to complex driving tasks, were also used. The digit span subtests of the Wechsler Adult Intelligence Scale-Revised (WAIS-R: Wechsler, 1981) and the Trail Making Test (TMT) -A were used to assess simple attention and concentration function. The TMT-B and the Modified Stroop Test were used to assess attention flexibility and executive function. The Clock Drawing Test (CDT) was used to assess visuospatial function.

## Tasks

**Driving performance** Daily driving skills associated with traffic accidents were measured by a road-tracking, a car-following, and a harsh-braking task. These tasks were run on a DS manufactured by Toyota Central R&D Labs., Inc. (Nagakute, Japan).

The car-following task involved a straight two-lane road with no other traffic, except for a single preceding car. When the preceding car decelerated, its brake lights came on. As the preceding car accelerated (to 60 km/h) or decelerated (to 40 km/h), the participant was required to maintain a distance between the cars as close to 5 m as possible. The car-following distance (m) was recorded every 20 ms. Performance was measured as the coefficient of variation (CV) obtained by dividing the standard deviation of the distance between the cars by the mean value (see Uchiyama, Ebe, Kozato, Okada, & Sadato, 2003). Therefore, a smaller distance CV value (DCV) indicated better performance. The test duration was 5 min.

The road-tracking task required the participant to drive at a constant speed of 100 km/h, while stabilizing the vehicle at the center of a gently winding road. The standard deviation of the lateral position (SDLP; in cm), which indicates weaving, was used as performance measures. Recordings were made every 20 ms during the test, which lasted for a period of 5 minutes.

The harsh-braking task included a straight two-lane road with no traffic, but with humanoid models on either side of the left lane. The humanoid models randomly ran onto the road as the participant's car approached. The participant was instructed to maintain a constant speed of 50 km/h and to avoid hitting the humanoid models by harsh braking as quickly as possible. The brake reaction time (BRT; in ms) was used as a measure of the cognitive psychomotor performance, including attention efficiency (see Ridout & Hindmarch, 2001). Each test consisted of 7 BRT trials over

a 5-min period, and the mean BRT was calculated from these results.

The driving performance evaluation using the DS was performed after the neuropsychological testing was complete. Before starting the evaluation, each driver was familiarized with the DS in three practice driving trials. During the practice, the driving performances of the HEA participants had plateaued (Kawano et al., in press). Within one week of the practice session, the test session was performed for each DS task, in the following order: the road-tracking, car-following, and harsh-braking task.

## Statistical analyses

To compare the demographic data including age, education, and the neuropsychological test performances among the groups, Kruskal-Wallis tests were carried out. Post hoc multiple comparisons were computed by the Mann-Whitney U test with the ordinary Bonferroni adjustment. DS values were compared among groups by one-way ANOVAs, after applying log transformations to normalize the data. Post hoc multiple comparisons were computed by the Tukey HSD. To analyses the relationship between the performance of DS task and some neuropsychological values, correlational analyses based on the Spearman's  $\rho$  and multiple linear regression analyses were carried out.

All Statistical analysis was performed with the SPSS. A  $P$ -value of less than 0.05 was considered to indicate statistical significance.

## Results

The characteristics of participant group are shown in Table 1. No significant differences were shown in age and education level between NEA and aMCI groups. There were significant differences between the two elderly groups for the MMSE and the Logical Memory delayed recall subtests of the WMS-R. All NEA individuals had MMSE scores over 26, and Logical Memory delayed recall performances over 11. On the other hand, all individuals with aMCI had a score of 9 or below on the Logical Memory delayed recall task.

To examine whether aMCI affected DS performance, the performance on each of the three DS tasks were compared among the groups. Figure 1 displays the groups' mean performance for each driving task. For the car-following task, the ANOVA revealed a main effect of group ( $F(2, 54) = 16.61, p = 0.00, \eta^2 = 0.38$ ). Multiple comparisons by Tukey HSD tests showed significant differences among all groups. The performance of NYA was significant higher than aMCI and NEA ( $t = 5.76, p = 0.00, r = 0.73; t = 2.69, p = 0.03, r = 0.38$ ), and NEA was significant higher than aMCI ( $t = 3.76, p = 0.00, r = 0.53$ ). For the road-tracking task, the ANOVA revealed a main effect of group ( $F(2, 54) = 17.62, p = 0.00, \eta^2 = 0.40$ ). Multiple comparison by Tukey HSD tests showed significant differences between NYA and aMCI. The performance of NYA was higher than aMCI and NEA ( $t = 5.53, p = 0.00, r = 0.71; t = 4.51, p = 0.00, r =$

Table 1: Characteristics of each group: Mean (Standard Deviations)

	aMCI		NEA		NYA	
Sex (female/male)	2/10		11/15		1/18	
Age (years)	71.8	(7.6)	70.0	(6.1)	39.3	(6.5)
Education (years)	13.4	(3.5)	14.4	(2.8)	16.6	(1.7)
<i>Cognitive characteristics</i>						
Mini-Mental State Examination	25.5	(2.9)	28.5	(1.4)		
WMS-R Logical Memory: delayed recall	4.6	(3.7)	20.0	(5.3)		

*Note.* aMCI = amnesic type of mild cognitive impairment, NEA = normal elderly adults, NYA = normal young adults, WMS-R = Wechsler Memory Scale-Revised.

0.56), but there was no significant difference between NYA and aMCI ( $t = 1.95$ ,  $p = 0.14$ ,  $r = 0.31$ ). For the harsh-braking task, 12 persons had missing values caused by technical problems or the participants. A one-way ANOVA was performed for 10 aMCI participants, 19 NEA participants, and 16 NYA participants. The analysis revealed no main effect of group ( $F(2, 42) = 3.21$ ,  $p = 0.05$ ,  $\eta^2 = 0.13$ ). There were no significant differences between groups.

Correlational analyses were conducted to examine the relationship between the performance on the car-following task (DCV) and some neuropsychological values in elderly groups. In the elderly group mixed NEA and aMCI, there were significant positive correlations between the DCV on the car-following task and TMT-A, TMT-B, and the Modified Stroop Test ( $\rho = 0.47$ ,  $p = 0.00$ ;  $\rho = 0.54$ ,  $p = 0.00$ ;  $\rho = 0.38$ ,  $p = 0.02$ ). No significant correlations were found between DCV and CDT ( $\rho = -0.18$ ,  $p = 0.30$ ), or DCV and the WAIS-R digit span subtest ( $\rho = -0.31$ ,  $p = 0.06$ ). In the aMCI group only, there were significant positive correlations between DCV and TMT-A, TMT-B, and the Modified Stroop Test ( $\rho = 0.47$ ,  $p = 0.00$ ;  $\rho = 0.54$ ,  $p = 0.00$ ;  $\rho = 0.38$ ,  $p = 0.02$ ). No significant correlations were found between DCV and CDT ( $\rho = -0.18$ ,  $p = 0.30$ ), or DCV and the WAIS-R digit span subtest ( $\rho = -0.31$ ,  $p = 0.06$ ). In addition, to confirm whether these variables, which are significantly associated with DCV, predicted the car-following performance more than just memory impairment

did, multiple linear regression analyses were carried out. Adjusting for the delayed recall performance on the WMS-R Logical Memory, the predictive strength of TMT-A, TMT-B, and the Modified Stroop Test was confirmed using log transformation DCV as an independent variable. Results of multiple linear regression analyses are displayed in Table 2. The results showed that the TMT-B performance significantly predicts the car-following performance after adjusting for the severity level of amnesia ( $\beta = 0.40$ ,  $p = 0.04$ ,  $R = 0.63$ , adjusted  $R^2 = 0.40$ ).

## Discussion

This study has provided clear evidence that late-life amnesia harms driving performance. An elderly sample with these characteristics showed that significantly decreased performance on the car-following and road-tracking tasks as compared with the normal young adults. However, the results showed a difference between individuals with symptomatic memory impairment and normal aging-related characteristics for driving abilities. There was no significant difference among the elderly groups on the road-tracking task. In contrast, there was a significant difference between groups on the car-following task. These results indicate that aMCI affects driving performance in patients with a prodromal stage of dementia, but normal aging also affects performance.

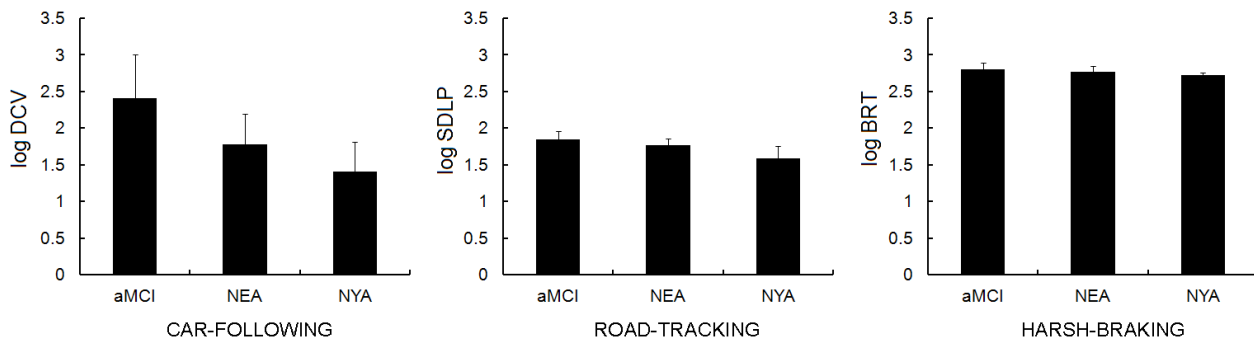


Figure 1: Means for DS task performances for each group.

*Note.* DCV = distance coefficient of variation, SDLP = standard deviation of the lateral position, BRT = brake reaction time, aMCI = amnesic subtype of mild cognitive impairment, NEA = normal elderly adults, NYA = normal young adults.

Table 2: Association between the performance for the car-following task and neuropsychological test scores in the elderly group.

	<i>R</i>	<i>R</i> <sup>2</sup>	<i>adjusted R</i> <sup>2</sup>	<i>B</i>	$\beta$	<i>P</i>	95% confidence interval of <i>B</i>		
TMT-A	0.577	0.333	0.292	0.011	0.167	0.318	-0.011	-	0.033
WMSR-delayed recall				-0.069	-0.475	0.007 *	-0.118	-	-0.021
TMT-B	0.629	0.396	0.359	0.007	0.395	0.040 *	0.000	-	0.013
WMSR-de				-0.042	-0.290	0.124	-0.097	-	0.012
Modified Stroop Test	0.574	0.329	0.290	0.023	0.260	0.140	-0.008	-	0.053
WMSR-delayed recall				-0.057	-0.382	0.033 *	-0.109	-	-0.005

Note. TMT = Trail Making Test, WMSR = Wechsler Memory Scale-Revised: Logical Memory.

Second, the different findings may be due to sample characteristics of MCI group and age matched control group. We checked that all the MCI participants showed a well-defined memory decline, and that all the normal elderly controls had intact memory and normal cognitive characteristics. In the present study, we examined these elderly people with amnesia and age-matched memory intact controls to evaluate driving performances using driving simulator, and compared their performance with normal young adults. Frittelli et al. (2009) compared among mild AD patients, people with MCI, and age-matched neurologically normal controls. We considered the problems involved with normal aging, and we found that not only aMCI but also normal-aging impaired driving performance. Our results indicated that MCI impaired car-following accuracy.

Simulated driving was found to be significant impaired in participants with aMCI. To identify cognitive variables explaining the decrement in car-following performance in late-life, correlational analyses were conducted. Although car-following performance was significantly related to performance on the TMT-A, TMT-B, and the Modified Stroop Test in our study sample, the correlations with the MMSE were not significant. These results are in agreement with previous studies showing that the MMSE, a widespread dementia screening tool, is limited as a predictive instrument. A few studies have reported a significant relation of MMSE scores to driving simulator and on-road driving performance (e.g. Fitten et al., 1995), but many studies have indicated that MMSE scores did not discriminate unsafe drivers from elderly drivers with unimpaired performance (Brown et al., 2005; Fox, Bowden, Bashford, & Smith, 1997; O'Neill et al., 1992). Also, general cognitive decline assessed by the MMSE is not found at the start of brain degeneration disease, and the aMCI operational criteria are conditional on a normal score on the MMSE. Thus, our elderly sample had less variability in MMSE scores. For patients with a prodromal stage of dementia, the MMSE is not useful as a predictor of safe driving, which has been shown in our results and numerous other studies with inconsistent results.

We found a linear relationship between performance on the TMT-B and the car-following task, and this relationship

remained after adjusting for the degree of symptomatic memory impairment. It is possible that the TMT-B is a useful tool for classifying elderly drivers as to whether they would pass or fail on the car-following task. The TMT is associated with attention flexibility and executive function (Lezak, 2004). The difference between unsafe drivers and persons with unimpaired performance may be associated with the flexibility of visual attention and executive function. Particularly, the TMT-B assesses more precisely the ability to alternate between two cognitive sets of stimuli, while the TMT-A of the test provides useful information concerning attention, visual scanning, speed of eye-hand coordination and information processing (Zalonis et al, 2008). These results suggest that persons with MCI who not only have memory impairment but also have difficulty gathering and processing information in parallel should drive under close supervision.

According to a cybernetic model of driving output, driver psychopathology is only one of a multitude of factors to consider (Moller, 2011). However, factors of driver psychopathology, normally limitations of psychological and neurophysiologic functioning, are associated with driving performance. Future studies need to evaluate the appropriate weighting of these risk factors in people with MCI, and develop criteria for re-evaluation of competency to drive.

**Limitation** The results of this study are constrained by certain limitations that are outlined below. Although we concluded this experiment based on the premise that participants in the three experimental groups were homogenous, there may have been differences in participants within each group. These differences may have been caused by misclassification of the amnesia status as a result of differences in age, educational level, or chronic physical disorder, although, the age and education level of the sample were not correlated with DS performance. We also did not consider the possibility that individuals with aMCI may have other impairments (a multiple-domain subtype) or not (amnestic single-domain subtype), because of the small sample size. Future studies need to consider these details of individuals in amnestic subtypes, non-amnestic subtypes, and multi-domain subtypes of MCI. Finally, whereas it is generally considered that road testing is the gold standard by which to evaluate driving

competence, we used the DS. There is a definite difference between on-road testing and DS tasks provide. The former can provide information on daily habits related to driving, whereas the latter provides objective assessment of competence. Future studies need to conduct longitudinal investigations of traffic incidents and compare them with changes in DS performance.

### Acknowledgments

Funding for this study was provided by research grants from the following: the Ministry of Health, Labor and Welfare of Japan; the Chiyoda Mutual Life Foundation; the Hori Sciences and Arts Foundation; the Conference for Expressway-related Social Contribution Activities; the Japan Health Foundation; the Suzuken Memorial Foundation, the ZENKYOREN, and the General Insurance Association of Japan.

### References

- Brown, L. B., Stern, R. A., Cahn-Weiner, D. A., Rogers, B., Messer, M. A., Lannon, M. C., Maxwell, C., Souza, T., White, T., & Ott, B. R. (2005). Driving Scenes Test of the Neuropsychological Assessment Battery (NAB) and on-road driving performance in aging and very mild dementia. *Archives of Clinical Neuropsychology*, 20, 209-216.
- Cooper, P. J., Tallman, K., Tuokko, H., & Beattie, B.L. (1993). Vehicle Crash Involvement and Cognitive Deficit in Older Drivers. *Journal of Safety Research*, 24, 9-17.
- First, M.B., Spitzer, R.L., Gibbon, M., Williams, J.B.W. & Benjamin, L. (1994). *The structured clinical interview for DSM-IV Axis II personality disorders (SCID-II) (Version 2.0)*. New York: Biometrics Research, New York State Psychiatric Institute.
- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12, 189-198.
- Fox, G. K., Bowden, S. C., Bashford, G. M., Smith, D. S. (1997). Alzheimer's disease and driving: prediction and assessment of driving performance. *Journal of the American Geriatrics Society*, 45, 949-953.
- Frittelli, C., Borghetti, D., Iudice, G., Bonanni, E., Maestri, M., Tognoni, G., Pasquali, L. & Iudice, A. (2009). Effects of Alzheimer's disease and mild cognitive impairment on driving ability: a controlled clinical study by simulated driving test. *International Journal of Geriatric Psychiatry*, 24, 232-238.
- Hakamies-Blomqvist, L., Sirén, A. & Davidse, R.J. (2004). Older drivers-a review. VTI report 497A. Swedish National Road and Transport Research Institute VTI, Linköping.
- Iverson, D. J., Gronseth, G. S., Reger, M. A., Classen, S., Dubinsky, R. M., & Rizzo, M. (2010). Practice parameter update: evaluation and management of driving risk in dementia. Report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*, 74, 1316-1324.
- Kawano, N., Iwamoto, K., Ebe, K., Aleksic, B. Noda, A., Umegaki, H., Kuzuya, M., Iidaka, T., & Ozaki, N. (in press). Slower adaptation to the driving simulator and simulator sickness in older adults. *Aging Clinical and Experimental Research*.
- Lezak, M. D. (2004). *Neuropsychological assessment (4th ed.)*. New York: Oxford University.
- Man-Son-Hing, M., Marshall, S. C., Molnar, F. J., & Wilson, K. G. (2007). Systematic review of driving risk and the efficacy of compensatory strategies in persons with dementia. *Journal of the American Geriatrics Society*, 55, 878-884.
- Moller, H. (2011). Psychiatric disorders and driving performance. In D. L. Fisher, M. Rizzo, & J. Caird (Eds.), *Handbook of driving simulation for engineering, medicine, and psychology*. London: Taylor & Francis.
- Molnar, F. M., Patel, A., Marshall, S. C., Man-Son-Hing, M., & Wilson, K. G. (2006). Clinical utility of office-based cognitive predictors of fitness to drive in persons with dementia: A systematic review. *Journal of the American Geriatric Society*, 54, 1809-1824.
- Morris, J.C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology*, 43, 2412-2414.
- O'Connor, M. L., Edwards, J. D., Wadley, V. G., & Crowe, M. (2010). Changes in mobility among older adults with psychometrically defined mild cognitive impairment. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 65B(3), 306-316.
- O'Neill, D., Neubauer, K., Boyle, M., Gerrard, J., Surmon, D., & Wilcock, G. K. (1992). Dementia and driving. *Journal of the Royal Society of Medicine*, 85(4), 199-202.
- Petersen, R. C. & Morris, J. C. (2005). Mild Cognitive Impairment as a Clinical Entity and Treatment Target. *Archives of Neurology*, 62, 1160-1163.
- Petersen, R. C., Doody, R., Kurz, A., Mohs, R. C., Morris, J. C., Rabins, P. V., Ritchie, K., Rossor, M., Thal, L., & Winblad, B. (2001). Current concepts in mild cognitive impairment. *Archives of Neurology*, 58(12), 1985-1992.
- Ridout, F., & Hindmarch, I. (2001). Effects of tianeptine and mianserin on car driving skills. *Psychopharmacology*, 154, 356-361.
- Uchiyama, Y., Ebe, K., Kozato, A., Okada, T., & Sadato, N. (2003). The neural substrates of driving at a safe distance: a functional MRI study. *Neuroscience Letters*, 352, 199-202.
- Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale-Revised*. New York: The Psychological Corporation.
- Wechsler, D. (1987). *Manual for the Wechsler Memory Scale-Revised*. San Antonio, TX: The Psychological Corporation.
- Zalonis, I., Kararizou, E., Triantafyllou, N. I., Kapaki, E., Papageorgiou, S., Sgouropoulos, P., & Vassilopoulos, D. (2008). A normative study of the trail making test A and B in Greek adults. *The Clinical Neuropsychologist*, 22(5), 842-850.



# From Vectors to Symbols to Cognition: The Symbolic and Sub-Symbolic Aspects of Vector-Symbolic Cognitive Models

Matthew A. Kelly (mkelly11@connect.carleton.ca)

Robert L. West (robert\_west@carleton.ca)

Institute of Cognitive Science, Carleton University  
1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6 Canada

## Abstract

To achieve a full, theoretical understanding of a cognitive process, explanations of the process need to be provided at both symbolic (i.e., representational) and sub-symbolic levels of description. We argue that cognitive models implemented in vector-symbolic architectures (VSAs) intrinsically operate at both of levels and thus provide a needed bridge. We characterize the sub-symbolic level of VSAs in terms of a small set of linear algebra operations. We characterize the symbolic level of VSAs in terms of cognitive processes, in particular how information is represented, stored, and retrieved, and classify vector-symbolic cognitive models in the literature according to their implementation of these processes. On the basis of our analysis, we speculate on avenues for future research, and suggest means for theoretical unification of existent models.

**Keywords:** Vector symbolic architectures; Holographic reduced representations; cognitive modelling; symbolic modelling; sub-symbolic modelling.

## Introduction

To achieve a full, theoretical understanding of a cognitive process and how it relates to the physical world, explanations of the process need to be provided at both symbolic (i.e., representational) and sub-symbolic levels of description. The classic symbolic approaches to modelling do not account for how the symbol manipulations described in the model could arise from neural tissue, or account for how the symbols themselves come into existence. Classic connectionist approaches are more concerned with neural plausibility, but are notoriously opaque, doing little to aid our understanding of the cognitive processes modelled. By contrast, the vector-symbolic approach to modelling explicitly provides an account at both levels of description.

Vector Symbolic Architectures (VSAs), a term coined by Gayler (2003; but see also Plate, 1995), are a set of techniques for instantiating and manipulating symbolic structures in distributed representations. VSAs have been used to successfully model a number of different cognitive processes (e.g., analogical mapping in Eliasmith & Thagard, 2001; letter position coding in Hannagan, Dupoux, & Christophe, 2011; semantic memory in Jones & Mewhort, 2007). It has been argued that VSAs provide a bridge between conventional symbolic modelling and both connectionist modelling (Rutledge-Taylor & West, 2008) and more realistic models of neural processing (Eliasmith, 2007). However, if we are to take the bridging metaphor seriously, it is important to clarify which parts of a VSA are symbolic in nature and which are sub-symbolic. We will

attempt to lay out a simple system for understanding VSAs in terms of basic operations and symbolic/sub-symbolic decisions, and thereby provide a comprehensive and comprehensible introduction to VSAs for newcomers, and a common frame of reference for those already using VSAs. By providing a high-level overview that integrates the techniques of existent VSA-based cognitive models into a coherent picture we hope to highlight as yet unexplored avenues of research and sketch what a VSA-based account of cognition as a whole would look like.

In this analysis, the vectors represent symbolic information. These vectors, or symbols, can be combined and manipulated using a small number of operations, which can be understood as sub-symbolic processes. However, the information processing models built from these operations are themselves, best characterized at a symbolic level of description. Importantly, the modelling decisions made at the sub-symbolic level are to some degree independent of the modelling decisions made at the symbolic level. This paper is divided into two parts to reflect these two levels of description, symbolic and sub-symbolic.

## The Sub-Symbolic Level

VSAs are closely related to the better-known tensor product representations (Smolensky, 1990), but unlike tensor product representations, VSAs can compactly represent symbolic expressions of arbitrary complexity. A number of VSA techniques exist in the literature, including Holographic Reduced Representations (HRRs; Plate, 1995), frequency-domain HRRs (Plate, 1994), some earlier forms of holographic associative memory (Eich, 1982; Murdock, 1982), as well as binary spatter codes (Kanerva, 1992), Multiply-Add-Permute coding (Gayler, 2003), and square matrix representations (Kelly, 2010).

Each VSA technique uses the same set of basic operations, but implements the operations differently. Thus the choice of a particular VSA dictates how symbols are instantiated and manipulated and defines the model at the sub-symbolic level. To ground the discussion, we mainly discuss Holographic Reduced Representations (HRRs) (Plate, 1995), as HRRs are the most widely used VSAs in the cognitive modelling literature. Also, HRRs are used as the basis for the Neural Engineering Framework (NEF; Eliasmith, 2007) and thus demonstrably have a clearly defined and plausible neural implementation. However, the other VSA techniques are similar and most anything that can be done with an HRR can be done with any VSA technique.

## ***n*-Space and Similarity**

In a VSA, a symbol, or representation, is an  $n$ -dimensional vector: a list of  $n$  numbers that defines the coordinates of a point in an  $n$ -dimensional space. VSAs work best for values of  $n$  in the hundreds or thousands (Plate, 1995).

A vector can be understood as a line drawn from the origin (the zero coordinates) to the coordinates specified by the vector. The length of the line is the vector's magnitude. The direction of the vector encodes the meaning of the representation. Similarity in meaning can thus be measured by the size of the angles between vectors. This is typically quantified as the cosine of the angle between vectors. The cosine of vectors **a** and **b** can be calculated as:

$$\text{cosine}(\mathbf{a}, \mathbf{b}) = (\mathbf{a} \cdot \mathbf{b}) / ((\mathbf{a} \cdot \mathbf{a})^{0.5} (\mathbf{b} \cdot \mathbf{b})^{0.5})$$

where  $\cdot$  is the dot product. A cosine of 1 means the vectors are identical, -1 means they are opposites, and 0 means they are completely dissimilar. If each vector has a magnitude of one, the cosine is just the dot-product of the vectors. Thus, some systems rescale all vectors to a magnitude of one after vector operations. In memory systems where new memories are superimposed on old memories, such re-scaling causes a recency effect and rapid forgetting because new memories will make-up a fixed fraction of memory, regardless of the quantity of previous experience.

While the cosine measures the *angle* between two vectors, the cosine is often described as a measure of *distance*. As it is more intuitive to describe similarity as a measure of distance than as a measure of the angle, for convenience, we can imagine the vectors as describing points on a *hypersphere*, such that the size of the angles are the distances between them.

## **Atomic versus Complex representations**

Representations in a VSA are either atomic or complex. An atomic representation is a unique representation, a symbol that cannot be broken down into sub-symbols. In an HRR, values for an atomic representation are typically generated by random sampling from a standard normal distribution. By assigning random values to the vectors, atomic representations will be uniformly distributed across the surface of the hypersphere, such that the atomic representations will have little to no similarity to each other.

Complex representations can be created by either combining atomic representations or recursively combining complex representations. Critically, in a VSA, a complex representation has the same dimensionality as an atomic representation, allowing representations both atomic and complex to be compared, or combined together to create representations of arbitrary complexity. VSAs have two operators for combining representations: superposition and binding. In HRRs, superposition is vector addition and binding is circular convolution. We denote vector addition by  $+$ , and circular convolution by  $*$ . Binding and superposition, along with random permutation, are the basic operations used to create complex representations in VSAs.

## **Superposition (+) versus Binding (\*)**

The key difference between superposition and binding is their effect on similarity. Superposition is similarity-preserving: the sum of two vectors is a vector that falls in the angle between them. Conversely, binding is similarity-destroying: the circular convolution of two vectors is roughly orthogonal to the two original vectors. The purpose of superposition is to combine representations to create a new representation that is similar to all of the combined representations. The purpose of binding, on the other hand, is to create "chunks": unique identifiers for combinations of representations.

Most VSA use a form of vector addition for superposition. Vector addition is computed by adding together the corresponding elements of the two vectors. So, for example,  $\{1,4,7\} + \{5,4,2\} = \{6,8,9\}$ .

To bind, HRRs use circular convolution,  $*$ , which can be computed rapidly using element-wise multiplication,  $\circ$ , and the fast Fourier transform,  $\text{fft}$ , and its inverse,  $\text{fft}^{-1}$ :

$$\mathbf{a} * \mathbf{b} = \text{fft}^{-1}(\text{fft}(\mathbf{a}) \circ \text{fft}(\mathbf{b}))$$

Essentially circular convolution is a lossy way of scrambling the information of the two vectors together to produce a new vector of the same dimensionality.

Consider the problem of learning the meaning of the phrase "kick the bucket", a colloquial euphemism for death. Suppose the cognitive model has a vector representation of the concept *kick* and a vector representation of the concept *bucket*. The sum (superposition) of those two vectors will produce a vector that is close to both *kick* and *bucket*, indicating that the phrase "kick the bucket" has a meaning similar to *kick* and to *bucket*. But in order for the cognitive model to be able to learn that the phrase "kick the bucket" has a distinct meaning that is not a function of its parts, the model needs to be able to assign to "kick the bucket" a distinct identifier. Binding is the operation that performs this function in VSA-based models. The vector  $\text{kick} * \text{bucket}$  is dissimilar to the vectors *kick*, *bucket*, and  $\text{kick} + \text{bucket}$ .

Binding and superposition can also be used jointly to address the binding problem (Gayler, 2003), that is, the question of how to couple sets of attributes together such that the attributes of one object are not confused with the attributes of another. For example, given a *small red square* and a *large blue circle*, the complex representation  $(\text{small} * \text{red} * \text{square}) + (\text{large} * \text{blue} * \text{circle})$  creates a single vector that distinctively represents the knowledge that the *square* is small and red and the *circle* is large and blue.

## **Unbinding**

Unbinding is an inverse of binding that allows vectors that have been bound together in a complex representation to be unpacked and recovered. Circular correlation,  $\#$ , is the unbinding operator for HRRs. Given a pair of vectors bound together, and one of the pair, referred to as the probe, unbinding produces an approximation of the other vector, referred to as the target, i.e.,

$$\mathbf{p} \# (\mathbf{p} * \mathbf{t}) = \mathbf{a} \approx \mathbf{t}$$

where  $\mathbf{p}$  is the probe,  $\mathbf{t}$  is the target, and  $\mathbf{a}$  is an approximation of the target.

Unbinding can be understood as binding with the inverse of the probe. The *inverse* of any vector  $\mathbf{x}$  is a re-ordering of the elements of  $\mathbf{x}$ , i.e. a permutation of  $\mathbf{x}$ , such that,

$$\mathbf{x} \# \mathbf{x} = \mathbf{x} * \text{inverse}(\mathbf{x}) \approx \delta$$

where  $\delta$  is the identity vector for binding, i.e. for any vector  $\mathbf{x}$ ,  $\mathbf{x} * \delta = \mathbf{x}$ . Thus binding with the inverse of a vector unbinds what that vector has been associated with:

$$\mathbf{a} \# (\mathbf{a} * \mathbf{b}) = \text{inverse}(\mathbf{a}) * (\mathbf{a} * \mathbf{b}) \approx \delta * \mathbf{b} = \mathbf{b}$$

In HRRs, the inverse of any vector  $\mathbf{x} = \{\mathbf{x}_1 \dots \mathbf{x}_n\}$  is:

$$\text{inverse}(\mathbf{x}) = \{\mathbf{x}_1, \mathbf{x}_n, \mathbf{x}_{n-1}, \dots, \mathbf{x}_3, \mathbf{x}_2\}$$

Circular convolution,  $*$ , is commutative, i.e., the order of binding does not matter when using circular convolution. Given vectors  $\mathbf{a}$  and  $\mathbf{b}$ , their association  $\mathbf{a} * \mathbf{b} = \mathbf{b} * \mathbf{a}$ , and likewise, when unbinding,  $\mathbf{b} \# (\mathbf{a} * \mathbf{b}) = \mathbf{b} \# (\mathbf{b} * \mathbf{a}) \approx \mathbf{a}$ .

## Permutation

Gayler (2003) describes random permutation as an operation used "to quote or protect the vectors from the other operations". Permutation of the numbers  $1 \dots n$  defines a unary function that can transform a vector. A randomly chosen permutation of a vector is unlikely to be similar to the original vector, but the permutation is also reversible. Given  $p$ , there is a permutation  $p^{-1}$  such that,  $p^{-1}(p(\mathbf{a})) = \mathbf{a}$ . When permuted, the information within a vector is essentially hidden and protected from being affected by other vector operations.

For example, as noted above, circular convolution is commutative, that is,  $\mathbf{a} * \mathbf{b} = \mathbf{b} * \mathbf{a}$ . This property of circular convolution can be useful, but it can be a hindrance in situations where the order of items matter, e.g. "dog feed" and "feed dog" are phrases which carry different meanings by virtue of differences in word order.

A non-commutative variant of circular convolution can be defined using a random permutation  $p$  and its inverse  $p^{-1}$ . By always randomly permuting one of the arguments before convolution, one defines a binding operation that is non-commutative, i.e. while  $\mathbf{a} * \mathbf{b} = \mathbf{b} * \mathbf{a}$ ,  $p(\mathbf{a}) * \mathbf{b} \neq p(\mathbf{b}) * \mathbf{a}$ . Unbinding then uses the inverse permutation  $p^{-1}$ , e.g.

$$\begin{aligned} \cosine(p^{-1}(\mathbf{a} \# (p(\mathbf{a}) * \mathbf{b})), \mathbf{b}) &\approx 0.71 \\ \cosine(p^{-1}(\mathbf{b} \# (p(\mathbf{a}) * \mathbf{b})), \mathbf{a}) &\approx 0 \end{aligned}$$

Non-commutative binding is used by the BEAGLE model (Jones & Mewhort, 2007) to bind vectors that stand for words in sentences in order to construct representations of the semantics of each of those words. For a variety of other uses of random permutation in VSAs, see Gayler (2003), Sahlgren, Holst, and Kanerva (2008), and Kelly (2010).

## The Symbolic Level

When making a vector-symbolic model, decisions need to be made at both the symbolic and sub-symbolic levels. At the sub-symbolic level, the modeller needs to decide how to instantiate symbols as vectors and symbol-manipulation as vector algebra. Conversely, at the symbolic level, the modeller needs to make decisions about how to structure, manage, store, and retrieve those symbols. Choosing to use HRRs rather than another kind of VSA can define the sub-symbolic level, but this choice is largely independent of the decisions to be made at the symbolic level.

In fact, we have already seen two examples of manipulations at the symbolic level. The first was combining binding and addition to create a vector that encodes information about bound entities (e.g., *small red square* and *large blue circle*). The second was combining permutation and binding to create a bound entity that maintained information about order. Essentially, all VSA systems work in the same way. Vectors encode the desired information according to some sort of scheme (i.e., by combining the operations discussed above), and then, when needed, the information is retrieved from the vectors.

## Encoding and Storage

BEAGLE (Jones & Mewhort, 2007) and DSHM (Rutledge-Taylor & West, 2008) use the terms *environmental vectors* and *memory vectors*. We extend the use of this terminology to other vector-symbolic models. An *environmental* vector is a vector that stands for atomic perceptions from the environment (e.g., a *red circle* needs two environmental vectors, one for *circle* and one for *red*). Environmental vectors are fixed and do not change. A *memory* vector is a complex representation stored by the model and used to produce behaviour. In some systems, memory vectors change with experience. Additionally, we use the term *experience vector* to refer to a representation that stands for the model's current experience of its environment created by combining environmental vectors (e.g., an experience vector could represent the perception of a *red circle* by convolving the environmental vectors of *circle* and *red*).

By examining the relationship between environmental, experience, and memory vectors across vector-symbolic models, we distinguish between three main approaches to storage. In a *many-to-one* vector model, all experience vectors are summed into a single memory vector for storage. In a *one-to-one* vector model, each experience vector is stored as a separate memory vector among an ever-growing number of memory vectors. In a *many-to-many* vector model, there are a fixed number of memory vectors, and incoming environmental vectors are used to update them. Each of these approaches has strengths and weaknesses.

**Many-to-one** In a *many-to-one* vector model, such as TODAM (Murdock, 1983) or CHARM (Eich, 1982), memory is modelled as a single, high-dimensional vector. All experience vectors are added to the memory vector. There is a limit to how much can be stored in the vector before mistakes start to be made. Mistakes are, of course, of interest to psychologists, and the pattern of mistakes made

by a *many-to-one* vector model allow it to mimic human forgetting in list-recall tasks. If the goal is to model how people store a small amount of recently learned or closely related information, a single memory vector suffices.

*Many-to-one* vector models also have the advantage of a clear neural implementation. In the Neural Engineering Framework (NEF; Eliasmith, 2007) binding, unbinding, and and superposition can all be implemented through neural connectivity. In the NEF interpretation, a *many-to-one* memory is a neural group with self-recurrent connections that acts as a working memory or buffer, and many such buffers could exist in the brain.

**One-to-one** In a *one-to-one* vector model, such as MINERVA (Hintzman, 1986), the Iterative Resonance Model (Mewhort & Johns, 2005), and the Holographic Exemplar Model (Jamieson & Mewhort, 2011), each experience vector is represented as a separate memory. While this approach to modelling memory is both simple and successful, the ever growing number of vectors that need to be stored and accessed by the memory system is both neurally implausible and computationally impractical for modelling tasks in which very large amounts of knowledge are relevant, e.g., semantic priming tasks (Jones & Mewhort, 2007). However, these models are able to reproduce a wide variety of memory effects, providing a unitary account of episodic, semantic, and implicit memory, indicating that, although their warehouse-style management of vectors is implausible, their processes of storage and retrieval provide a good analogue for biological memory.

**Many-to-many** *Many-to-many* vector models, such as BEAGLE (Jones & Mewhort, 2007) and DSHM (Rutledge-Taylor, 2008), can be understood as a hybrid of the earlier *many-to-one* and *one-to-one* approaches. In *many-to-many* memory, for each item of interest, there is a randomly generated environmental vector and a specially constructed memory vector. In BEAGLE the items of interest are words: the environmental vector stands for the word's orthography or phonology and the memory vector stands for the word's meaning. In DSHM, the items are objects relevant to the experimental task: the environmental vector stands for the percept of the object and the memory vector stands for the concept of the object.

Like the one-to-one models, the management of the vectors in many-to-many systems is computationally expensive and, at this point, neurally implausible. However, the ability to generate memory vectors that stand for particular concepts is very powerful (e.g., Rutledge-Taylor, Vellino, & West, 2008) and allows these systems to capture numerous different phenomena (e.g., Rutledge-Taylor & West, 2008) and represent vast quantities of data (Jones & Mewhort, 2007).

For example, to create an association between *keyboards* and *computers*, each time a computer and keyboard co-occur a copy of the environmental vector for keyboard can be added to the memory vector for computer and a copy of the environmental vector for computer can be added to the memory vector for keyboard. The effect to this would be to move the memory vector for computer closer to the

environmental vector for keyboard and move the memory vector for keyboard closer to the environmental vector for computer. Over time, the result of this is to organize the space so that memory vectors are clustered around environmental vectors that they co-occur with so that the distance between the vectors equals strength of association.

Another, more complicated example involves binding and the use of the *placeholder vector*. The placeholder vector is an atomic (i.e., random) vector, but it is used to encode all associations, and thus can be used as a universal retrieval cue. Consider the phrase or stimulus *blue triangle*. Without using the placeholder, we could update memory as follows:

$$\begin{aligned}\text{memory}_{\text{blue}} &+= \text{blue} * \text{triangle} \\ \text{memory}_{\text{triangle}} &+= \text{blue} * \text{triangle}\end{aligned}$$

By binding together the environmental vectors for **blue** and **triangle** and adding the result to the memory vectors for *blue* and *triangle* (an operation denoted by  $+=$ ), we move the two memory vectors towards the point in space described by the vector **blue \* triangle**, and thereby move **memory<sub>blue</sub>** and **memory<sub>triangle</sub>** closer together. But people almost *never* get the concepts *blue* and *triangle* confused with each other. This is because *blue* is a colour (or an adjective), and *triangle* is a shape (or a noun), i.e. they are different sorts of thing.

Conversely, consider updating using the placeholder:

$$\begin{aligned}\text{memory}_{\text{blue}} &+= \text{placeholder} * \text{triangle} \\ \text{memory}_{\text{triangle}} &+= \text{blue} * \text{placeholder}\end{aligned}$$

This moves **memory<sub>blue</sub>** towards **placeholder \* triangle**, i.e. towards all properties of triangles, and moves **memory<sub>triangle</sub>** towards **blue \* placeholder**, i.e. towards all things that are blue. Thus, by using a placeholder, the memory vectors for nouns will cluster together in one region of space, and the vectors for adjectives will cluster together in another region of space, and things that are *colours* will cluster separately from things that are *coloured*. This is a subtle distinction but Jones and Mewhort (2007) have shown it to be very important and very powerful.

## Retrieval

There are two categories of information retrieval processes used in vector-symbolic models: *unbinding*, which retrieves information from a particular vector, and *resonance*, which allows information to be retrieved from the entire library of vectors in memory. *Many-to-one* models, such as TODAM (Murdock, 1982) only use unbinding. *One-to-one* models that do not use binding to encode associations, such as MINERVA (Hintzman, 1986), only use resonance. In *many-to-many* systems, these two retrieval processes are complementary. For example, resonance can be used to retrieve a vector, which can then be unbound.

**Unbinding** Consider a simple example where the agent is given a set of coloured shapes to remember: *blue triangle*, *green square*, *red circle*. In a *many-to-one* vector model this could be encoded by binding (\*) the vectors for the shapes to the colours, then summing to create a memory vector:

$$\text{memory} = \text{blue} * \text{triangle} + \text{green} * \text{square} + \text{red} * \text{circle}$$

The colour of any one of these shapes could then be recalled by unbinding (#) using the shape to probe memory:

$$\text{triangle} \# \text{memory} \approx \text{blue}$$

In a *many-to-many* vector model, unbinding may use the *placeholder* as the probe. The placeholder is a special, randomly generated atomic vector that acts as a key to all of memory. The placeholder is initially used in binding:

$$\begin{aligned} \text{memory}_{\text{blue}} &= \text{placeholder} * \text{triangle} \\ \text{memory}_{\text{triangle}} &= \text{placeholder} * \text{blue} \end{aligned}$$

The placeholder can then be used in unbinding:

$$\text{placeholder} \# \text{memory}_{\text{triangle}} \approx \text{blue}$$

**Resonance (*one-to-one*)** The term resonance comes from MINERVA (Hintzman, 1986), but it is implemented differently across different models. In MINERVA, the process of resonance begins by measuring the similarity (cosine) of each vector in memory to the probe. Then resonance computes a weighted sum of all vectors in memory. This sum, termed the *echo*, is what the model retrieves from memory. Each vector in the sum is weighted by its similarity to the probe raised to an exponent.

For example, the three shapes might be represented as:

$$\begin{aligned} \text{memory}_1 &= \text{triangle} + \text{blue} \\ \text{memory}_2 &= \text{square} + \text{green} \\ \text{memory}_3 &= \text{circle} + \text{red} \end{aligned}$$

If the probe is **triangle**, then the echo would approximate:

$$\begin{aligned} \text{echo} &\approx 0.5^b \text{memory}_1 + 0.0^b \text{memory}_2 + 0.0^b \text{memory}_3 \\ \text{echo} &\approx 0.5^b (\text{triangle} + \text{blue}) \end{aligned}$$

such that the memory system would remember that the triangle is blue. The exponent  $b$  is a small, positive integer that is odd-numbered so as to preserve the sign of the similarity. Note that the similarity values of 0.5 and 0.0 are approximate. Random vectors in a high dimensional space have an expected cosine of 0, but the actual cosine between any two random vectors will be a little more or a little less.

The exponent  $b$  critically allows *one-to-one* vector models to function even when there is a very large amount of data in memory. If the exponent  $b$  is 1, the result of resonance roughly imitates decoding in a simple associative memory, such as a Hopfield network. With an exponent greater than one, resonance increases the signal to noise ratio in the echo by increasing the relative weighting of the memory vectors most similar to the probe. If  $b$  is too low, a large number of partial matches in memory could easily overwhelm an exact match to the probe, resulting in a poor echo. With a high  $b$ , the echo will essentially just be the most similar vector in memory to the probe. In MINERVA, a  $b$  of 3 is standardly used, but a  $b$  of 3 may be too low when modelling a larger

sum of knowledge than what is typically necessary to model a psychology experiment (e.g., in modelling word pronunciation, such as in Kwantes & Mewhort, 1999).

In the Iterative Resonance Model (IRM; Mewhort & Johns, 2005), resonance is iterated, and with each iteration  $b$  is increased until a decision to stop iterating is made, resulting in either successful retrieval or a failure to retrieve. This approach has two benefits: (1) the number of iterations can be used to predict response time in memory tasks, and (2) it eliminates  $b$  as a tweaking parameter by introducing a theory-driven approach to setting its value.

**Resonance (*many-to-many*)** Although the term *resonance* is used to describe retrieval in *many-to-many* vector models, the implementation is different and simpler: Essentially, the memory vector most similar to the probe is retrieved. This can be understood as a kind of spreading activation (Rutledge-Taylor & West, 2008). The probe and memory vectors can be understood as points on a hypersphere, such that the cosine measures the distance between them. One can imagine a ripple of activation spreading out from the probe across the surface of the hypersphere. The memory vectors closest to the probe become active in working memory, with the closer vectors becoming active sooner. This model of resonance allows BEAGLE (Jones & Mewhort, 2007) to make semantic priming reaction time predictions (e.g., that *doctor* is recognized faster when preceded by *nurse* than when preceded by an unrelated prime such as *stapler*) and to model the fan-effect in DSHM (Rutledge-Taylor & West, 2008).

## Conclusions

We hold that, in order to bridge the gap between human experience and neural connectivity, explanations at both the symbolic and sub-symbolic levels of description are necessary parts of theory in cognitive science. As we illustrate in this paper, cognitive models that use vector-symbolic architectures intrinsically operate at both of these levels of description and thereby provide a needed bridge between the two kinds of explanation.

At the sub-symbolic level is the vector-symbolic architecture itself, and the linear algebra operations on vectors that comprise the architecture: *similarity*, *superposition*, *binding*, *unbinding*, *permuting*, *unpermuting*. All of these operations are easily amenable to neural implementation, as in the NEF (Eliasmith, 2007).

At the symbolic level, we have the cognitive model itself, and the cognitive processes that define it. On the basis of their storage and retrieval mechanisms, we classify existing vector-symbolic cognitive models into *many-to-one*, *one-to-one*, and *many-to-many* vector models. This classification scheme highlights stark differences between these models.

*Many-to-many* vector models differ from the other two classes of model in two important ways. First, *many-to-many* models use a *placeholder* vector to stand for "this item I am thinking about". The placeholder acts as a symbol with an important functional role but no perceptual or conceptual meaning. It may be useful to incorporate other kinds of function vectors in future models, e.g., a *wildcard* vector to

stand for "that item that I'm not thinking about", vectors to stand for emotional states, for truth values, et cetera.

Secondly, in *many-to-many* models, memory vectors are labelled and stand for particular concepts, whereas in *one-to-one* models concepts are an emergent phenomenon produced by the echoes retrieved using resonance (Hintzman, 1986). While the conceptual representations in *many-to-many* models are powerful, having a predefined number of concepts is implausible and limiting.

*Many-to-one* vector models can be constructed in NEF as self-recurrent neural groups and are understood as working memory buffers. By contrast, *one-to-one* and *many-to-many* models are best understood as models of long-term memory, but as yet lack a neural explanation.

Finding a means of translating *one-to-one* and *many-to-many* vector models into neural models may provide a route to a unified, vector-symbolic account of memory storage and retrieval. As we noted earlier, a *one-to-one* model behaves somewhat like a Hopfield network when the resonance exponent  $b$  is set to 1. To implement a *one-to-one* model as a network, one needs to find a mechanism analogous to  $b$  that can act to increase the signal to noise ratio in the echo. We speculate that the vector-symbolic intersection circuit proposed by Levy and Gayler (2009) might provide a start for developing such a mechanism.

We suspect that the memory vectors that stand for concepts in *many-to-many* vector models are, in fact, the echoes in *one-to-one* vector models. That is to say, we agree with Hintzman (1986) that concepts are an emergent property of retrieval. Using a *one-to-one* model to do the kind of large scale modelling in *many-to-many* models is impossible because *one-to-one* models store all experiences without any form of compression. However, a neural implementation of a *one-to-one* model would naturally be lossy in its storage, and so could provide a plausible account of concept formation over a lifetime of experiences.

Unification in other areas, such as representation, is important too. Incorporating a vector-symbolic model of string encoding (Hannagan et al., 2011) into the BEAGLE model of semantics (Jones & Mewhort, 2007) could, for instance, allow BEAGLE to model how shared orthography can help and hinder in understanding the meaning of words.

Eventually, we hope to see developed a vector-symbolic cognitive architecture, which not only presents a unified and neurally plausible approach to representation, storage, and retrieval, but also extends the vector-symbolic account beyond its roots in memory theory, and integrating it into accounts of emotions, attention, perception, and consciousness. As cognitive scientists, it is important to keep in mind our ultimate, lofty, and collective goal of a theory that unifies not only all aspects of the cognition, but all relevant levels of description.

## References

- Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review*, 89, 627–661.
- Eliasmith, C., & Thagard, P. (2001). Integrating structure and meaning: a distributed model of analogical mapping. *Cognitive Science*, 25, 245–286.
- Eliasmith, C. (2007). How to build a brain: From function to implementation. *Synthese*, 159, 373–388.
- Gayler, R. (2003). Vector symbolic architectures answer Jackendoff's challenges for cognitive neuroscience. *ICCS/ASCS International Conference on Cognitive Science*.
- Hannagan, T., Dupoux, E., & Christophe, A. (2011). Holographic string encoding. *Cognitive Science*, 35, 79–118.
- Hintzman, D. L. (1986). "Schema abstraction" in multiple-trace memory models. *Psychological Review*, 93, 441–428.
- Jamieson, R. K., & Mewhort, D. J. K. (2011). Grammaticality is inferred from global similarity: A reply to kinder (2010). *The Quarterly Journal of Experimental Psychology*, 64, 209–216.
- Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114, 1–37.
- Kelly, M. A. (2010). Advancing the theory and utility of holographic reduced representations. (Master's thesis, School of Computing, Queen's University).
- Kanerva, P. (1996). Binary spatter-coding of ordered k-tuples. *Proceedings of the 1996 International Conference on Artificial Neural Networks*, 869–873.
- Kwantes, P. J., & Mewhort, D. J. K. (1999). Modeling lexical decision and word naming as a retrieval process. *Canadian Journal of Experimental Psychology*, 53, 306–315.
- Levy, S.D., & Gayler, R.W. (2009). A distributed basis for analogical mapping. *New frontiers in analogy research; Proceedings of the Second International Analogy Conference - Analogy 09*, 165–174.
- Mewhort, D. J. K., & Johns, E. E. (2005). Sharpening the echo: An iterative-resonance model for short-term recognition memory. *Memory*, 13, 300–307.
- Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89, 609–626.
- Plate, T. A. (1994). Distributed representations and nested compositional structure. (Doctoral dissertation, Department of Computer Science, University of Toronto).
- Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6, 623–641.
- Rutledge-Taylor, M. F., Vellino, A., & West, R. L. (2008). A holographic associative memory recommender system. *Proceedings of the Third International Conference on Digital Information Management*, 87–92.
- Rutledge-Taylor, M. F. & West R. L. (2008). Modeling the fan-effect using dynamically structured holographic memory. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 385–390.
- Sahlgren, M., Holst, A., & Kanerva, P. (2008). Permutations as a means to encode order in word space. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 1300–1305.
- Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46, 159–216.

# Sex Differences in the Discrimination of Non-Native Speech Sounds

**Vera Kempe (v.kempe@abertay.ac.uk)**

Division of Psychology, University of Abertay Dundee  
Dundee, DD1 1HG, United Kingdom

**John C. Thoresen (john.thoresen@epfl.ch)**

Brain Mind Institute, Ecole Polytechnique Fédérale de Lausanne,  
1015 Lausanne, Switzerland

**Patricia J. Brooks (patricia.brooks@csi.cuny.edu)**

Department of Psychology, College of Staten Island and the Graduate Center, City University of New York  
Staten Island, NY 10314, USA

## Abstract

This study examined sex differences in the discrimination of minimal pairs of foreign language (non-native) tonemes. Adult native speakers of English (237 women and 177 men), with no prior exposure to a tonal language, performed an AX-task, which required them to discriminate between rising and falling-rising Norwegian tonemes. When controlling for nonverbal intelligence, prior exposure to foreign languages, and age, sensitivity measures ( $d'$ ) showed a clear male advantage. Thus, the sex differences previously observed in non-linguistic temporal processing tasks appear to extend to the discrimination of unfamiliar non-native speech sounds. These sex differences in auditory processing may be due to anatomical differences between men and women in the ratio of white to grey matter in the left hemisphere, which, in turn, might affect speed of neural transmission. These findings contribute to the ongoing debate on cognitive effects of putative sex differences in intra- and inter-hemispheric connectivity.

**Keywords:** non-native speech perception; tonal contrast; sex differences; adult L2 learning; auditory processing.

## Introduction

Auditory processing of temporal sequences underlies the neural representation of speech and has been implicated in impairments in language development; e.g., dyslexia and Specific Language Impairment (Goswami et al., 2002; Talcott et al., 2000; Tallal, 1980). However, little is known about individual differences in auditory processing in the non-clinical adult population, and, specifically, individual differences in the ability of adults to discriminate the speech sounds of foreign (non-native) languages. Illuminating the basis of individual differences in non-native speech processing may help to explain some of the considerable variance in outcomes observed among adult foreign language (L2) learners (Johnson & Newport, 1989). So far, only a few studies have explored individual differences in the processing of non-native speech sounds (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Golestani & Zatorre, 2009). Thus, we know very little about which factors, besides age of first exposure (Flege, MacKay & Meador, 1999), make an adult more or less successful in processing non-native speech sounds.

Research on temporal processing as a predictor of psychometric intelligence (Rammsayer & Brandler, 2007) and working memory capacity (Troche & Rammsayer, 2009) has revealed a sex difference, with men outperforming women in temporal order judgments (Szelag et al., 2011; Wittman & Szelag, 2003) and temporal discrimination tasks (Rammsayer & Troche, 2010). Factor-analytical approaches have traced the male advantage to a latent variable – temporal resolution power, which has been linked to neural oscillation rate determining speed and accuracy of neural transmission (Jensen, 1982). The male advantage is not confined to the auditory modality, however, but has also been observed for tactile temporal processing (Rostad, Mayer, Fung & Brown, 2007), suggesting that it affects general temporal processing in the sub-second range.

In addition to sex differences in pure temporal information processing tasks, there is evidence for a male advantage in the discrimination of pitch contours of computer-generated waveforms, comprising a fundamental frequency and two formants, which were presented binaurally (McRoberts & Sanders, 1992). Pitch contour discrimination requires sensitivity to changes in pitch over time and therefore relies on temporal processing. Rapidly changing values of one or several acoustic parameters (e.g., formant transitions) play a crucial role in distinguishing different speech sounds—for example, notoriously difficult phonological contrasts like the dental-retroflex contrast (for English speakers) or the r/l contrast (for Japanese speakers) require sensitivity to rapid spectral changes. The present study therefore aims to examine whether a male advantage can also be found in the ability to discriminate natural non-native speech contrasts. Natural speech sounds differ from synthetic stimuli in their greater variability within speech sound categories and in the complexity of their acoustic characteristics.

We chose to examine sensitivity to lexical tones as one example of such a non-native speech contrast. We used Norwegian tonemes as many dialects of Norwegian have a simple tonal system with pitch accents that distinguish otherwise homophonous bisyllabic words. Detecting these tonal contrasts requires tracking temporal changes in pitch



contours of bi-syllabic words. We tested adult native English speakers' sensitivity to the tonal contrast between rising and falling-rising tonemes, which are illustrated in Figure 1.

If sex differences in non-linguistic temporal processing extend to linguistic stimuli we would expect to see a male advantage in the processing of an unfamiliar Norwegian tonal contrast by native English speakers.

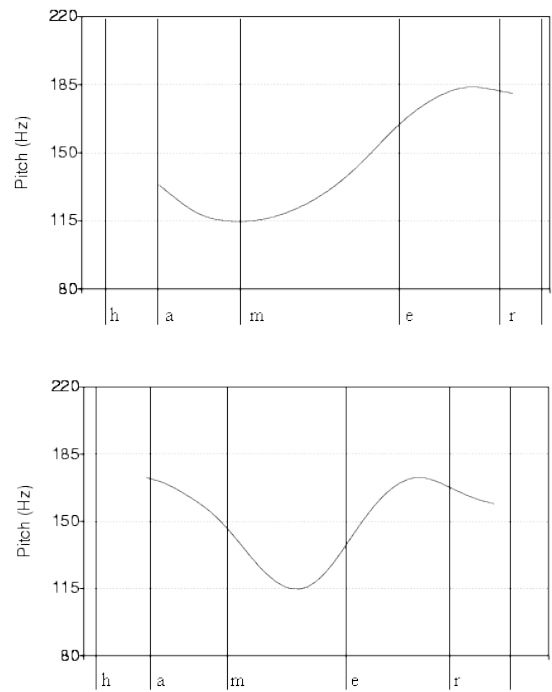


Figure 1: Illustrations of the different pitch contours of minimal pairs of Norwegian tonemes. The upper panel shows the rising tone for the word ‘Hammer’ [a proper noun]. The lower panel shows the falling-rising tone for the word ‘hammer’ [the tool]<sup>1</sup>.

### Method

We pooled data from six experiments on non-native discrimination of Norwegian tonal contrasts that were conducted over a period of five years (Kempe, Thoresen & Brooks, 2007, 2008; Kempe, Brooks, Marronaro & Thoresen, 2010; Kempe, Thoresen, Kirk, Brooks & Schaeffler, 2011). These experiments tested native speakers of English differing in dialectal background (American, English and Scottish), and varied with respect to other cognitive abilities tested (e.g., verbal working memory capacity) and other speech sound contrasts presented in addition to the tonal contrast. All six experiments controlled for nonverbal intelligence and prior exposure to other languages. It is necessary to control for nonverbal intelligence because of the well-established link between

processing speed and temporal processing on the one hand and psychometric intelligence on the other hand (e.g., van Raavenzwaaj, Brown & Wagenmakers, 2011; Rammsayer & Brandler, 2007; Sheppard & Vernon, 2008). It is also important to make sure that any observed differences cannot be accounted for by prior exposure to tonal contrasts.

**Participants:** A total of 458 participants (197 men) were tested in various locations in the United Kingdom and in the United States (New York City). Participants’ mean age was 22.8 years (range 17-61 years). Of the participants, 282 (117 men) were native speakers of American English and 176 (80 men) were native speakers of Scottish English.

An additional ten native speakers of Norwegian (four men), aged 20 to 22 years, were tested to confirm that the chosen tonal contrast can reliably be discriminated by native speakers of the language.

**Materials:** To capture the within-category variability characteristic of natural speech sounds, a male native speaker of Norwegian recorded two different instances of each of 16 bisyllabic Norwegian words comprising 8 minimal tonal contrast pairs. In half of the pairs, the first stressed syllables contained short vowels (mean length 80 ms); in the other half they contained long vowels (mean length 144 ms). These eight minimal pairs are listed in Table 1. Stimuli were recorded at a sampling rate of 44.1 kHz and presented to participants through Sennheiser headphones.

Table 1: Minimal pairs of Norwegian words used for tone discrimination. Note that the members of a pair are homophones despite differences in spelling.

rising tone	falling-rising tone
short vowel	
<i>bønder</i> [farmers]	<i>bønner</i> [beans]
<i>lammet</i> [lamb]	<i>lamme</i> [to paralyze]
<i>sulten</i> [hunger]	<i>sulten</i> [hungry]
<i>verket</i> [creation]	<i>verke</i> [to ache]
long vowel	
<i>bøter</i> [fines]	<i>bøter</i> [to repent]
<i>laget</i> [team]	<i>lage</i> [to make]
<i>suget</i> [suction]	<i>suge</i> [to suck]
<i>været</i> [weather]	<i>være</i> [to be]

To establish whether pitch contours were indeed sufficiently distinct between the two tonemes, we measured pitch of the steady-state part of the vowels in the first and second syllable. As Figure 1 indicates, there should be a larger difference between pitch on the first and the second syllable for a rising pitch contour than for a falling-rising pitch contour. Pitch measurements confirmed that the median pitch difference between syllables for the rising pitch contour (38 Hz) was significantly larger than the median pitch difference for the falling-rising pitch contour (4 Hz), Mann-Whitney  $U = 2.0$ ,  $p < .001$ ,  $r = .84$ , indicating

<sup>1</sup>Figure reprinted from the project Lingo resource at the Norwegian University of Science and Technology, Trondheim, [http://www.ling.hf.ntnu.no/ipa/no/tema\\_008.html](http://www.ling.hf.ntnu.no/ipa/no/tema_008.html)

a sufficiently large measurable difference in acoustic characteristics between the two tonemes.

**Procedure:** The Norwegian tonal contrasts were presented in an AX discrimination procedure requiring participants to make judgments about whether pairs of words sounded the ‘same’ or ‘different’. The 32 ‘same’ trials consisted of different instances of the same word spoken with the same pitch accent. The 32 ‘different’ trials consisted of minimal pairs of words spoken with different pitch accents. The two words in each pair were presented with an inter-stimulus interval of 200 ms.

Participants also completed the Cattell Culture-Fair Test of Nonverbal Intelligence, Scale 3, Form A (Cattell & Cattell, 1973), and a language background questionnaire, used to confirm participant status as a native English speaker and to inquire about prior exposure to languages other than English. Participants were asked to rate their reading, writing, speaking and comprehension abilities in each of their languages on a scale from 1 (rudimentary) to 6 (native-like).

## Results

**Norwegian native speakers:** Each participant’s performance was converted to an  $A'$  score, a measure of sensitivity that corrects for individual differences in bias.  $A'$  is a non-parametric analogue to  $d'$  and has values ranging from 0 and 1, with 0.5 corresponding to chance. The mean  $A'$  score for the native speakers of Norwegian was .93 ( $SD = 0.04$ ), which supports the validity of our stimuli and confirms that discrimination of the native tonal contrast did not pose any problems for native speakers.

**English native speakers:** Seventeen men and 18 women reported some familiarity with a tonal language (e.g., Chinese). A further three men and six women failed to provide proficiency ratings in the language background questionnaire. All these participants were excluded from the analyses leaving a total of 177 men and 237 women.

The mean  $A'$  score for the entire sample was 0.71 ( $SD = 0.14$ ). Table 2 presents means and standard deviations for men and women, along with results of Bonferroni-corrected t-tests comparing men and women on age, Culture Fair non-verbal intelligence test scores (CF IQ), number of learned foreign languages (L2s), and mean proficiency self-ratings for first and second L2s. If participants had not studied any L2, the corresponding rating scores were set to 0. These comparisons showed that while women had higher proficiency self-ratings in their first L2, women’s discrimination of Norwegian tonemes was significantly lower than men’s.

Table 2: Means and standard deviations (in parenthesis) of the various measurements for men and women. The last column shows the results of a t-test (\*\* indicates significance after Bonferroni correction at  $p < .01$ ).

	men	women	$t(412); p$
age	22.5 (7.1)	23.0 (7.3)	-0.56; .574
CF IQ score	25.0 (5.1)	24.0 (5.1)	1.96; .050
# L2s	1.4 (0.8)	1.6 (0.8)	-2.50; .013
self-rating 1 <sup>st</sup> L2	2.3 (1.3)	2.7 (1.5)	-2.90; .004**
self-rating 2 <sup>nd</sup> L2	0.7 (1.1)	1.0 (1.2)	-2.40; .018
$A'$	0.74 (0.13)	0.70 (0.14)	3.18; .002**

To account for potential uncontrolled effects of the different testing conditions and contexts in the six experiments, we computed standardized  $A'$  scores for each experiment separately. These standardized  $A'$  scores served as the dependent variable in a multiple regression analysis with age, Culture Fair nonverbal intelligence test scores, the various language background variables, and sex (coded as a dummy variable) as predictors (see Table 3).

Table 3: Results of a multiple regression analysis of all predictors and sex, coded as dummy variable, on standardized  $A'$  scores for toneme discrimination ability.

	$\beta$	$t$	$p$
age	-.073	-1.50	.134
CF IQ score	.161	3.24	.001
# L2s	.049	0.69	.492
self-rating 1 <sup>st</sup> L2	-.071	-1.28	.201
self-rating 2 <sup>nd</sup> L2	.076	1.04	.300
sex	-.152	-3.11	.002

The model accounted for a total of 5.8% of the variance,  $F(6,405) = 5.2$ ,  $p < .001$ , and showed men outperforming women, over and above a facilitative effect of non-verbal intelligence. A further stepwise regression analysis with all predictors entered at the first step, and sex entered at the second step, showed that sex accounted for a unique 2% of variance, cumulative  $F(1,405) = 9.7$ ,  $p < .01$ .

To check whether the effect of non-verbal intelligence was present in both sexes, we performed separate multiple regression analyses for men and women. In both analyses, the only significant effect was that of the Culture Fair test (men:  $\beta = .16$ ,  $p < .05$ ; women:  $\beta = .17$ ,  $p < .05$ ). The relationship between non-verbal intelligence and sensitivity to non-native tonal contrasts in men and women is depicted in Figure 2.

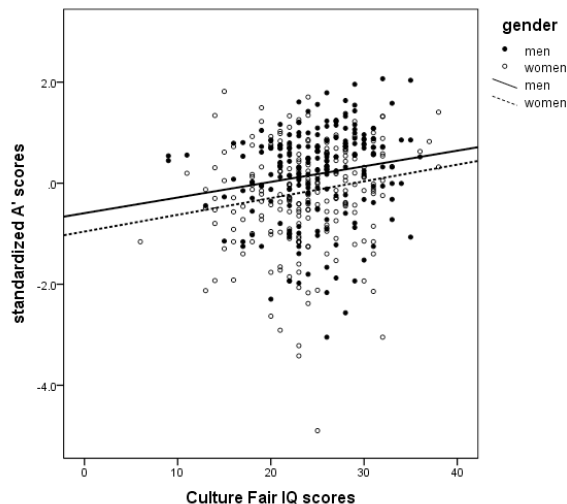


Figure 2: Correlation between Culture Fair scores and sensitivity to Norwegian tonal contrasts for men (black circles, solid line) and women (white circles, dashed line).

## Discussion

Our findings demonstrate a small but significant male advantage in non-native toneme discrimination, which cannot be attributed to sex differences in prior exposure to foreign languages or non-verbal intelligence. The lack of a link between prior language exposure and processing of non-native speech sounds is consistent with similar findings from other studies (Golestani & Zatorre, 2009). Note that the effect of non-verbal intelligence on toneme discrimination ability was independent of the effect of sex, confirming the general link between temporal information processing—which is one component of non-native speech sound discrimination—and psychometric intelligence.

The sex effect may seem unexpected because studies that employ non-native speech sound discrimination tasks or temporal auditory processing tasks typically do not compare the performance of men and women. Moreover, given that the sex effect is very small, the sample sizes in such studies are often not large enough for it to be detectable. However, recently, Bowles, Silbert, Jackson & Doughy (2011) reported a male advantage for discrimination of two Hindi consonant contrasts involving differences in Voice Onset Time in a large sample of 1,185 male and 395 female native speakers of American English. This suggests that the male advantage in temporal information processing clearly extends to the processing of difficult non-native speech sounds containing rapid spectral changes.

What mechanisms may be responsible for this sex effect? Golestani et al. (2007) have shown in a perceptual training study that faster learning of non-native speech sounds, involving rapid spectral changes, was associated with differences in brain anatomy. Specifically, faster learning was linked to larger overall white matter volumes in left Heschl's gyrus and increased degree of left > right asymmetry in white matter density in auditory cortex.

Increased white matter volume may indicate greater myelination, which would result in more rapid neural transmission crucial for perception of rapid spectral changes. It may also be due to a greater number of white matter fibers connecting language regions within and between cortical hemispheres, for example, connecting the auditory cortex with anterior and posterior language regions. Given that for these perceptual learning tasks performance at the outset was highly correlated with speed of learning (Golestani & Zatorre, 2009) it is reasonable to speculate that similar anatomical changes may also distinguish individuals who perform better in perceptual discrimination tasks when presented with a non-native speech sound for the first time. In addition, white matter volume has been shown to be negatively correlated with variability in isochronous tapping in the sub-second range (Ullén, Forsman, Blom, Karabanov & Madison, 2008) suggesting that white matter volume can be implicated in rapid temporal processing in other domains as well.

Comparisons of male and female brain anatomy and cytoarchitecture have revealed larger white matter to grey matter ratios in men than women, and less white matter asymmetry between hemispheres in women (Gur et al., 1999). Gur et al. (1999) suggested that maintaining grey matter volume consisting of somatodendritic tissue—responsible for computation—at the relative expense of myelinated connective tissue—responsible for information transmission—may be a reasonable evolutionary strategy for dealing with the smaller cranial volumes of females, where transmission occurs over relatively shorter distances than in males.

How can these conjectures about the anatomical substrate responsible for a male advantage in temporal processing be reconciled with findings of (a) somewhat higher verbal abilities in women and (b) the absence of mean sex differences in psychometric intelligence? Meta-analyses (Hyde & Linn, 1988; Lynn & Mikk, 2009) have shown a reliable albeit very small female advantage in verbal abilities, mainly related to speech production and reading and writing abilities. This seems to be at odds with the present finding of a male advantage in temporal processing which extends to the processing of non-native speech sounds. However, there is much more to verbal abilities than the processing of non-native speech sounds, making it unlikely that a sex difference in one capacity will dominate the complex interaction of skills required for the various aspects involved in language learning and processing.

Mean sex differences in general intelligence have generally proven to be elusive (Johnson, Carothers & Deary, 2009) despite sex differences in reaction times (Der & Deary, 2006) and temporal processing (Rammsayer & Troche, 2010), – parameters that have been shown to be predictive of general intelligence (Sheppard & Vernon, 2008). The present study is in agreement with these findings as the trend towards slightly higher Culture Fair non-verbal intelligence scores in men was not significant after Bonferroni correction, despite the fact that Culture Fair

scores correlated significantly with sensitivity to the Norwegian tonal contrast. A review of research on sex differences in various timed tests revealed that men are faster on reaction time and finger tapping tests while women are faster in naming and symbol copying and neither sex outperforms the other in general intelligence (Roivainen, 2011). Thus, while reaction time and temporal information processing appear to explain some of the variance in general intelligence, other performance components are also bound to play a role and these components do not necessarily favor men.

It is important to keep in mind that the observed sex effect in non-native speech sound processing was very small. Future research will have to explore to what extent the male advantage in non-native speech sound processing generalizes to other tasks; e.g., identification tasks or AXB-tasks which may be more taxing on working memory or on the ability to form long-term representations of novel speech sounds. It is even less clear whether the male advantage in non-native speech sound processing benefits other aspects of adult foreign language learning, such as morphosyntax or vocabulary acquisition. To clarify these issues, studies of individual differences in various aspects of language learning should include sex as a variable into their analyses.

Despite these limitations, the reported findings underscore the importance of studying sex differences in cognitive tasks as one of the domains that allow researchers to explore potential cognitive repercussions of neuro-anatomical differences.

### Acknowledgments

The authors would like to thank Felix Schaeffler for advice in constructing the stimuli and Neil W. Kirk and James Munro for their help in running the experiments. Parts of this study were funded by a Small Grant from the journal *Language Learning*.

### References

- Bowles, A. R., Silbert, N. H., Jackson, S. R., & Doughy, C. J. (2011). Individual differences in working memory predict second language learning success. Poster presented at the 52nd Annual Meeting of *The Psychonomic Society*, Seattle, WA.
- Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, 101, 2299–2310.
- Cattell, R. B., & Cattell, H. E. P. (1973). *Measuring Intelligence with the Culture-Fair Tests*. Champaign, IL: Institute for Personality and Ability Testing.
- Der, G., & Deary, I. J. (2006). Reaction time age changes and sex differences in adulthood. Results from a large, population based study: The UK Health and Lifestyle survey. *Psychology and Aging*, 21, 62–73.
- Flege, J. E., MacKay, I. R., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *Journal of the Acoustical Society of America*, 106, 2973–2987.
- Golestani, N., & Zatorre, R. J. (2009). Individual differences in the acquisition of second language phonology. *Brain & Language*, 109, 55–67.
- Golestani, N., Molko, N., Dehaene, S., LeBihan, D., & Pallier, C. (2007). Brain structure predicts the learning of foreign speech sounds. *Cerebral Cortex*, 17, 575–582.
- Goswami, U., Thompson, J., Richardson, U., Stainthorp, R., Hughes, D., Rosen, S., & Scott, S. K. (2002). Amplitude envelope onsets and developmental dyslexia: a new hypothesis. *Proceedings of the National Academy of Sciences USA*, 99, 10911–10916.
- Gur, R. C., Turetsky, B. I., Matsui, M., Yan, M., Bilker, W., Hughett, P., & Gur, R. E. (1999). Sex differences in brain gray and white matter in healthy young adults: Correlations with cognitive performance. *Journal of Neuroscience*, 19, 4065–4072.
- Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, 107, 139–155.
- Jensen, A. R. (1982). Reaction time and psychometric g. In H. J. Eysenck (Ed.), *A model for intelligence*. New York: Springer.
- Johnson, W., Carothers, A., & Deary, I. J. (2009). A role for the X chromosome in sex differences in general intelligence? *Perspectives in Psychological Science*, 4, 598–611.
- Johnson, J. S. & Newport, E. L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21, 60–99.
- Kempe, V., Brooks, P. J., Marronaro, R., & Thoresen, J. (2010). Individual differences in the perception of a non-native tonal contrast. Poster presented at the Workshop on *Psycholinguistic Approaches to Speech Recognition in Adverse Conditions*. Bristol, UK.
- Kempe, V., Thoresen, J. C., & Brooks, P. J. (2007). Differences in foreign language phoneme perception and production. Poster presented at the 48th Annual Meeting of *The Psychonomic Society*, Long Beach, USA.
- Kempe, V., Thoresen, J. C., & Brooks, P. J. (2008). Norwegian toneme perception by non-native speakers. Poster presented at the 49th Annual Meeting of *The Psychonomic Society*, Chicago, IL.
- Kempe, V., Thoresen, J. C., Kirk, N. W., Brooks, P. J., & Schaeffler, F. (2011). Perceptual and cognitive predictors of non-native phoneme discrimination. Poster presented at the 52nd Annual Meeting of *The Psychonomic Society*, Seattle, WA.
- Lynn, R., & Mikk, J. (2009). Sex differences in reading achievement. *Trames*, 13, 3–13.
- McRoberts, G. & Sanders, B. (1992). Sex differences in performance and hemispheric organization for a nonverbal auditory task. *Perception and Psychophysics*, 51, 118–122.

- Rammsayer, T., & Troche, S. (2010). Sex differences in the processing of temporal information in the sub-second range. *Personality and Individual Differences*, 49, 923-927.
- Rammsayer, T. H., & Brandler, S. (2007). Performance on temporal information processing as an index of general intelligence. *Intelligence*, 35, 123-139.
- Roivainen, E. (2011). Gender differences in processing speed: A review of recent research. *Learning and Individual differences*, 21, 145-149.
- Rostad, K., Mayer, A., Fung, T. S., & Brown, L. N. (2007). Sex-related differences in the correlations for tactile temporal thresholds, interhemispheric transfer times, and nonverbal intelligence. *Personality and Individual Differences*, 43, 1733-1743.
- Sheppard, L. D., & Vernon, P. A. (2008). Intelligence and speed of information processing: A review of 50 years of research. *Personality and Individual Differences*, 44, 535-551.
- Szelag, E., Szymaszek, A., Aksamit-Ramotowska, A., Fink, M., Ulbrich, P., Wittmann, M., & Pöppel, E. (2011). Temporal processing as a base of language universals: Cross-linguistic comparisons on sequencing ability with some implications for language therapy. *Restorative Neurology and Neuroscience*, 29, 35-45.
- Talcott, J. B., Witton, C., McClean, M., Hansen, P. C., Rees, A., Green, G. G. A., & Stein, J. F. (2000). Dynamic sensory sensitivity and children's word decoding skills. *Proceedings of the National Academy of Sciences USA*, 97, 2952-2962.
- Tallal, P. (1980). Auditory temporal perception, phonics, and reading disabilities in children. *Brain and Language*, 9, 182-198.
- Troche, S. J., & Rammsayer, T. H. (2009). The influence of temporal resolution power and working memory capacity on psychometric intelligence. *Intelligence*, 37, 479-486.
- Ullén, F., Forsman, L., Blom, Ö., Karabanov, A., & Madison, G. (2008). Intelligence and variability in a simple timing task share neural substrates in the prefrontal white matter. *Journal of Neuroscience*, 28, 4238-4243.
- van Raavenzwaaj, D., Brown, S., & Wagenmakers, E.-J. (2011). An integrated perspective on the relation between response speed and intelligence. *Cognition*, 119, 381-393.
- Wittmann, M., & Szelag, E. (2003). Sex differences in perception of temporal order. *Perceptual and Motor Skills*, 96, 105-112.

# Evaluative feedback can improve deductive reasoning

Sangeet Khemlani<sup>1</sup> and Adam Moore<sup>2</sup>  
{skhemlani, adam.moore457}@gmail.com

<sup>1</sup>Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC 20375 USA

<sup>2</sup>Center for Advanced Brain Imaging, Georgia Institute of Technology, Atlanta, GA 30318 USA

## Abstract

We examine whether reasoning is improved by evaluative feedback, i.e., the information of whether a reasoner's answer was correct or incorrect, and report two studies that show that evaluative feedback increases the chances that participants will produce normatively correct responses for deductive reasoning problems. In Experiment 1, participants who were given feedback about their performance did better on problems based on disjunctions that were designed to elicit illusory inferences. In Experiment 2, participants answered difficult syllogisms with more accuracy when they were provided with feedback. We conclude by contrasting the rule-, heuristics-, and model-based accounts of deduction on their ability to explain the effects of evaluative feedback.

**Keywords:** feedback, reasoning, illusory inferences, syllogisms

## Introduction

People often receive feedback after they have drawn an inference. Feedback can manifest in a contrarian's objection, a pat on the back, a heated argument, or a grunt of disapproval. In many cases, feedback can be *prescriptive*, i.e., it can be accompanied by further instructions and suggestions for improvement, such as what one might receive in a classroom environment. In other cases, feedback can be *evaluative* and devoid of any pedagogical value, such as a final grade in a course. Prescriptive feedback has been shown to improve participants' reasoning on a wide variety of tasks (Cheng, Holyoak, Nisbett, & Oliver, 1986; Khemlani & Johnson-Laird, 2009; Leevers & Harriss, 1999). The effect is robust but unsurprising: if prescriptive feedback could not make better reasoners out of humans, it would be difficult to explain the internalization of rules, heuristics, and insights. Our investigation focuses instead on evaluative feedback, i.e., feedback that neither explains nor characterizes performance as any more than a minimal description of whether performance was correct or incorrect on a particular trial (Neth, Khemlani, & Gray, 2008). We are interested in this impoverished form of feedback because it is unclear what effect, if any, it should have on a person's future performance on similar problems (Klayman & Ha, 1987; Wason, 1960).

Suppose reasoners were told that the deduction they drew from a set of premises was incorrect. Since they have no further information on why their answers were incorrect, it is not clear that the feedback could apply to different sets of premises. Reasoners may remember the structure of the premises so that if they encounter the same problem again, they can provide a correct answer, but there is little reason

to think that the feedback should produce any systematic improvement in reasoning beyond correcting an answer to a particular problem unless reasoners directly search for an explanation of why they went wrong (Walsh & Johnson-Laird, 2009). Moreover, given multiple-choice problems in which the elimination of one answer does not identify the correct answer, evaluative feedback might produce no effect whatsoever. Since memories are susceptible to interference and decay, it is uncertain whether evaluative feedback will have any impact on the ability to solve related but syntactically different problems. Few studies have examined how immediate evaluative feedback informs reasoning (but cf. Wason, 1964), and few psychological theories of reasoning explicitly permit evaluative feedback to modulate the way individuals reason (Braine & O'Brien, 1998; Oaksford & Chater, 2007; Rips, 1994; Stenning & Van Lambalgen, 2008) though there is evidence that the first thing individuals do upon learning that their conclusion is incorrect is to check their reasoning (Johnson-Laird, Girotto, & Legrenzi, 2004).

If feedback influences the way people make deductions, theories of reasoning ought to accommodate such effects by showing how individuals make use of the additional information with which they are supplied. In the following experiments, we show that immediate, evaluative feedback improves the way individuals reason. We conclude by explaining how three prominent theories of reasoning might account for improvements in performance due to evaluative feedback.

## Experiment 1: Sentential reasoning

Experiment 1 presented participants with a set of problems that were expected to yield "illusory" sentential inferences. Sentential inferences are those based on sentential connectives such as *and* (a conjunction) and *or* (a disjunction). Illusory inferences are systematic errors that are produced when people fail to consider all of the possibilities consistent with the premises (Khemlani & Johnson-Laird, 2009). Each problem was based on a set of premises in which one disjunction was embedded within another. The disjunctions were either exclusive or inclusive; for example, consider these premises based on two exclusive disjunctions:

Suppose one of the following assertions is true and one is false:

1. You have the blue candies and the red candies.
2. You have the red candies or else the orange candies, but not both.

Is it possible to have the blue candies and the orange candies only?

Table 1: The four types of problem in Experiment 1, their premises and corresponding questions, the predicted conclusions, and the correct conclusions to each question.

Problem			Conclusion	
Type	Premises	Question	Predicted	Correct
Exclusive-exclusive	One is true and one is false: 1. A and B. 2. B or else C	Is it possible to have A and B only?	Yes	No
Exclusive-inclusive	One is true and one is false: 1. A and B. 2. B or C or both.	Is it possible to have A and B?	Yes	No
Inclusive-exclusive	One or both are true: 1. A and B. 2. B or else C.	Is it possible to have A and C only?	No	Yes
Inclusive-inclusive	One or both are true: 1. A and B. 2. B or C or both.	Is it possible to have A and C only?	No	Yes

The rubric makes clear that there is an exclusive disjunction between assertions 1 and 2. In previous studies participants tended to respond “no”, that it was not possible to have only the blue and orange candies (Khemlani & Johnson-Laird, 2009, Experiments 2 and 3). The answer is illusory, however; the premises allow for the possibility of having only blue and orange candies. The difficulty of problems that yield illusory inferences is robust; even when participants received remedial instructions that explained how to overcome the illusions, they made errors more often than not.

In the present study, participants were provided feedback about their responses. They were randomly assigned to two different feedback conditions: *feedback*, in which participants were informed about whether their answers were correct or incorrect; and *no feedback*, in which they received no information about their performance but rather continued to the next problem after a brief delay.

## Method

*Participants and design.* 53 volunteers were recruited through a platform hosted by Amazon.com through which people participate in experiments over the Internet for monetary compensation (for a discussion on the validity of results from this platform, see Paolacci, Chandler, & Ipeirotis, 2010). None of the participants had received any training in logic. They received four sorts of problems based on disjunctive premises, and all of the problems were designed to elicit an illusory inference. Table 1 presents the four sorts of problems, each of which was presented twice using different materials. We tested two groups of participants; one group received feedback on their answers and the other did not.

*Procedure and materials.* On each trial, participants received a disjunctive set of premises and a question that was intended to elicit a fallacious response. Participants then selected buttons marked “Yes” or “No”. Once the participant responded, there was a delay for 2 seconds during which feedback, if appropriate, was displayed on the

screen. In the no feedback condition, participants received just a delay before moving on to the next problem. Whenever feedback was given to a participant, it replaced the text of the premises and conclusion so that participants did not have access to the problem itself, and could not re-evaluate the premises. The materials used in the study pertained to various combinations of colored candy, and participants received each set of materials only once.

## Results and discussion

Table 2 presents the percentages of correct responses for each group of participants. Participants found the problems quite difficult, and produced correct responses 30% of the time. They made more correct responses when presented feedback than when not (Mann-Whitney test: 38% vs 21%,  $z = 3.00$ ,  $p < .0001$ ).

Table 2: The percentage of correct responses to the four types of problem in Experiment 1 as a function of the type of feedback received.

Problem Type	Received feedback?	
	Yes	No
Exclusive-exclusive	28	12
Exclusive-inclusive	28	8
Inclusive-exclusive	55	27
Inclusive-inclusive	41	35

Participants in Experiment 1 performed better when presented evaluative feedback about the correctness or incorrectness of their answers on problems designed to yield illusory inferences. Likewise, performance did not differ as a function of the order in which the problems were presented; participants in the feedback condition did not do better on the last three trials compared to the first three trials in the experiment (33% vs. 37%, Wilcoxon test,  $z = .68$ ,  $p = .49$ ).



Table 3: The premises of the fourteen types of syllogistic problems used in Experiment 2 and a set of candidate conclusions, which include: a correct conclusion that necessarily follows from the first and second premises; a consistent conclusion that does not necessarily follow from the premises; and the most common erroneous conclusion that reasoners generate.

Problem		Candidate conclusion		
First premise	Second premise	Correct	Consistent	Common conclusion
Some A are B	No C are B	Some A are not C	Some C are not A	No A are C
All A are B	Some C are not B	Some C are not A	Some A are not C	Some C are A
No A are B	Some B are C	Some C are not A	Some A are not C	No A are C
All B are A	No B are C	Some A are not C	Some C are not A	No A are C
Some A are not B	All C are B	Some A are not C	Some C are not A	Some A are C
No B are A	Some B are C	Some C are not A	Some A are not C	No A are C
No B are A	All B are C	Some C are not A	Some A are not C	No A are C
All B are A	Some B are not C	Some A are not C	Some C are not A	Some A are C
Some B are A	No B are C	Some A are not C	Some C are not A	No A are C
All B are A	All B are C	Some A are C	All C are A	All A are C
No A are B	Some C are B	Some C are not A	Some A are not C	No A are C
No A are B	All B are C	Some C are not A	Some A are not C	No A are C
Some B are not A	All B are C	Some C are not A	Some A are not C	Some A are C
Some A are B	No B are C	Some A are not C	Some C are not A	Some A are C

One alternative explanation of the results in Experiment 1 is that instead of making participants better, feedback might have slowed them down so they could read the premises more carefully. A portion of the participants might have initially sped through the study, and if the effect of feedback was to get them to pay attention and stop responding erratically, then the results could be explained without recourse to theoretical claims about performance increases. We are skeptical of such an explanation for two reasons. First, most participants did not respond randomly; they performed reliably worse than chance. Second, every participant received a 2-second delay between trials, and so at the outset they were unable to rush through the study.

Another explanation of the results in Experiment 1 is that instead of making participants perform better, feedback made participants more erratic. The percentage of correct responses was not reliably greater than what would be expected if participants chose responses at random, which could have been driven by a reduction in participants' confidence in their initial answers due to the feedback they received. Likewise, one limitation of the present study is that erroneous disjunctive inferences, while representative of sentential reasoning, come about as a result of a tendency to overlook possibilities (Khemlani & Johnson-Laird, 2009, p. 622). To overcome these limitations, we used a different task and a more diverse set of materials in Experiment 2. Instead of having participants choose between just two alternatives, we provided participants several putative conclusions for syllogistic reasoning problems, only one of which validly followed from the premises.

## Experiment 2: Syllogistic reasoning

Experiment 2 examined whether feedback could help participants discover the correct response to a syllogism from a set of alternatives. Syllogistic reasoning is logically

simple but psychologically complex, and many theories have been proposed to deal with how humans process syllogisms. Modern theories of syllogistic reasoning are based on mental models (Johnson-Laird & Bara, 1984; Polk & Newell, 1995), formal rules of inference (Braine & O'Brien, 1998; Rips, 1994), or the mood of the most informative premise (Oaksford & Chater, 2007).

Not all syllogisms are created equal; some are easy and can be solved in a matter of seconds, and others are so vexing that reasoners may spend many minutes considering their premises. Consider one such problem:

All of the brewers are accountants  
All of the brewers are cashiers  
What must be true?

Reasoners often conclude that all accountants are cashiers, or else that no valid conclusion follows from the premises. The former conclusion is false because not all accountants are necessarily brewers. The latter is false as well, because a valid conclusion exists: it follows that some accountants are cashiers. The moral of the story is that syllogisms are not always easy to solve, and in the present study, we chose those syllogisms that pose the most trouble for reasoners (see Khemlani & Johnson-Laird, in press, for a review).

Participants were once again randomly assigned to the two feedback conditions that were used in Experiment 1, i.e., they either received feedback or did not.

## Method

*Participants and design.* 56 volunteers were recruited from the same participant pool that was used in Experiment 1. All of the participants were untrained in logic, and they completed the experiment using an interface written in Ajax. They received fourteen syllogistic reasoning problems; Table 3 presents the premises of the problems and their

corresponding alternative conclusions. As in the previous study, we tested a group of participants who received feedback against a control group that received no feedback.

*Procedure and materials.* Participants took the study over the Internet, and for each problem they received two quantified premises and four alternative conclusions. The problems were taken from syllogisms identified by previous research as being the most difficult for reasoners (Bucciarelli & Johnson-Laird, 1999; Chapman & Chapman, 1959; Oaksford & Chater, 1999). One of the four alternative conclusions was correct, and the other three were distractors. The distractors consisted of a) a conclusion that is consistent with, but does not follow necessarily from, the premises; b) the most common but incorrect response that participants had spontaneously generated in previous studies (see Bucciarelli & Johnson-Laird, 1999); and c) a “null” response, i.e., “no valid conclusion”. The order in which the alternative conclusions were displayed on the screen was randomized.

Participants were told that only one of the four responses was correct. They registered their response by selecting buttons assigned to one of the four conclusions. When the participant responded, there was a delay for 2 seconds during which feedback, if appropriate, was displayed on the screen. Whenever feedback was given to a participant, it replaced the text of the premises. The materials used in the study pertained to various combinations of occupations, e.g., “All of the brewers are accountants,” and participants received each set of materials only once.

### Results and discussion

Table 4 presents the proportion of agreement to the four different types of conclusions that were presented on each trial. The problems were difficult; across the study, participants agreed to correct conclusions 39% of the time. The feedback (44%) condition yielded reliably more correct responses than the no feedback condition (44% vs. 33%, Mann-Whitney test,  $z = 2.15$ ,  $p < .05$ ), and this pattern held for 10 of the 12 syllogisms (Binomial test,  $p < .05$ ). As in Experiment 1, performance in the feedback condition did not increase steadily; accuracy on the first five trials was not reliably lower than on the last five trials (41% vs. 45%, Wilcoxon test,  $z = .48$ ,  $p = .63$ ).

Table 4: The proportion of agreement to the four types of conclusions in Experiment 2 as a function of the type of feedback received.

Conclusion type	Received feedback?	
	Yes	No
Correct	44	33
Consistent	21	20
Common	26	26
No valid conclusion	9	16

As in Experiment 1, we consider the alternative explanation that instead of making participants perform better, feedback slowed participants down and forced them to read the premises more carefully. The present results are not consistent with this account, because regardless of presence or absence of feedback participants chose the consistent conclusion about 20% of the time. If the feedback motivated them to be more careful, then they would have made fewer errors of interpretation, and we would see a difference between the extent to which they agreed with consistent answers. The uniformity of their answers suggests that in fact, participants were reading and comprehending the problems at the same level of competence regardless of the feedback they were given.

### General Discussion

Across two different paradigms calling for deductive reasoning, evaluative feedback improved performance relative to no feedback. No psychological theory of deduction is constructed to explicitly make predictions about effects of feedback. However, we conclude by examining how the principles of various theories of reasoning might be used to account for the performance gains observed in our studies.

Psychological theories of deduction fall into three broad categories: those based on formal rules akin to those in the proof theory of logic (e.g., Rips, 1994; Stenning & Van Lambalgen, 2008), those based on the processing of subjective probabilities and probabilistic heuristics (Oaksford & Chater, 2007), and those based on models akin to those in the semantic theory of logic (e.g., Johnson-Laird, 1983; Polk & Newell, 1995). Each type of theory yields a different account of how feedback might be integrated into deductive processes to improve reasoning performance.

#### *Theories of deduction based on formal rules*

Theories based on the application of formal rules of inference propose that reasoning is a process of proof in which syntactic rules are used to derive conclusions from the premises. A precursor to reasoning is accordingly the recovery of the logical form of premises to allow the application of rules. Once the logical form has been recovered, rules are applied over the formal structure of the premises to yield conclusions. Theories based on formal rules posit only those rules that allow participants to draw valid deductions, but recognize that humans often make errors in reasoning. For instance, Rips (1994, p. 386) suggests that logical errors are made more often for problems that require more steps of proof or that require complex rules to be applied to the premises. To solve a particular problem, syntactic rules must be utilized to derive a proof of its conclusion, step by step. Thus improvement in reasoning on a given problem can be explained by a) an increased tendency to recognize that a particular rule is necessary, and b) the increased frequency with which the rule is applied. Rips (1994) reports studies of such improvements. If evaluative feedback improves the way

individuals reason, then it should affect the way particular rules are recognized and applied. Thus, rules theories predict that feedback makes it easier to recover the rules relevant to the problem at hand. But it is not clear how rules theories would account for generalized performance increases based on evaluative feedback, i.e., increases in reasoning that affect many rules at once. For complex reasoning problems that require several rules to be applied, a credit assignment problem exists: a reasoner does not know, based on evaluative feedback alone, which rule has been incorrectly applied. The reasoner's performance can increase only if credit is assigned to the rule that was incorrectly applied; otherwise, it is possible that the reasoner does worse on future trials. Rules theories offer no hint at how the credit assignment problem could be overcome, but one solution is to statistically abstract the conditional relationship between the use of each particular rule in all relevant contexts and the ultimate outcome. Such a solution relies on gathering and encoding massive amounts of data, and so it is incompatible with performance increases after only a few trials.

#### *Heuristic based theories of deduction*

The theory of deduction based on probabilistic heuristics (Oaksford & Chater, 2007) assumes that individuals reason by employing simple heuristics based on informativeness and probabilistic entailment. A claim is informative if it rules out possibilities; thus, the universal statement All of the swans are white is more informative than the existential statement Some of the swans are white, because the universal rules out the possibility that some swans are not white, whereas the existential statement has no such constraint. Oaksford and Chater (2007) argue that people use heuristics based on this knowledge of informativeness to select and test conclusions along with heuristics based on probabilistic entailment, i.e., knowledge about whether one premise probabilistically follows from another. For instance, the statement All swans are white probabilistically entails the statement Some swans are white. The authors detail several ways in which individuals might apply the heuristics based on informativity and probabilistic entailment to test and derive conclusions, and show that the predictions made by the heuristics are a good fit for the difficulty of certain syllogisms, i.e., arguments with two or more quantified premises (Oaksford & Chater, 2007, Ch. 7).

Oaksford and Chater argue that for a particular syllogism, individuals construct and test conclusions based on heuristics that require the following pieces of information: 1.) a complete ordering of premises on their informativeness; 2) the quantifier of the least informative premise; 3) a complete account of probabilistic entailments; 4) the most informative premise. Oaksford and Chater suggest that (1) and (3) are immutable whereas (2) and (4) are calculated from the premises of each new problem. According to their theory, a human's departure from a normative answer provided by logic need not be suboptimal, as logic is the wrong normative baseline by which to assess

rationality. If humans "err", they do so not because they do not provide the answer sanctioned by classical logic, but because they are equipped with inexpensive heuristics that are fallible. Chater and Oaksford's (1999) analysis of difficult syllogisms suggest that to provide probabilistically valid responses to syllogistic reasoning problems, reasoners are required to apply all of the probabilistic heuristics specified in the model. As Copeland and Radvansky (2004) observe, the need to apply more heuristics taxes working memory as it requires individuals to hold in mind both the heuristics themselves as well as the results of each heuristic. Feedback may trigger improved performance by inducing reasoners to apply all heuristics instead of just a subset.

#### *Theories of reasoning based mental models*

The mental model theory of deduction (Johnson-Laird, 1983) is based on the notion that individuals reason, both deductively and inductively, by constructing representations of possibilities. The theory proposes that the process of deduction goes through three stages: individuals first use the meaning of sentences and their knowledge to envisage what is possible given the propositions expressed in the premises, and they represent the possibilities as a single mental model. Second, the model is scanned for information not made explicit in the premises, and if any such information is found it is considered a putative conclusion. Third, individuals assess the conclusion by looking for counterexamples, i.e., alternative models of the premises where the conclusion is false. If a counterexample exists, then the conclusion is dismissed and individuals return to the second stage to construct an alternative explanation. Improved reasoning as a result of evaluative feedback can be attributed to the diligence with which individuals form models and search for counterexamples. Reasoners would then use feedback as a cue to fully flesh out multiple mental models and search for counterexamples. These processes require working memory resources to hold the relevant models in mind and operate on them. Thus, increases in executive control as a result of evaluative feedback may improve the ability to consider alternatives and search for counterexamples.

In summary, we showed how three accounts of reasoning can explain why deduction is enhanced by feedback. Mental rules theories predict that feedback must affect individual rules to improve general performance on non-identical problems; however, substantial experience would be required for the reasoner to learn the individual relations between feedback and the use of particular rules across many contexts. The probability heuristics theory holds that individuals apply a series of simple heuristics based on approximations of statistical calculations. Regardless of the normative baseline used, feedback for a particular problem should cue participants to apply all relevant heuristics instead of a subset. Mental models theory posits that individuals flesh out models and search for counterexamples in order to obtain correct answers, and can make mistakes

when they fail to do either. Thus, feedback may prompt the reasoner to search for counterexamples more assiduously, and could lead to general increases in performance across many types of problems.

The results we report demonstrate that feedback and reinforcement can improve the efficacy of conscious reasoning. Theories of reasoning can be extended to handle feedback effects explicitly in the ways we outlined above, and doing so may allow future studies to identify and test the ways in which feedback information implicitly changes reasoners' representations and inferential processes.

### Acknowledgments

This research was supported by a National Science Foundation Graduate Research Fellowship to both authors, and by National Science Foundation Grant No. DRMS 0844851. We thank Jeremy Boyd, Andy Conway, Sam Glucksberg, Adele Goldberg, Geoffrey Goodwin, Matt Johnson, Phil Johnson-Laird, Mike Oaksford, and Laura Suttle for their suggestions and critiques.

### References

- Chapman, L. J., & Chapman, A. P. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology*, 58, 220–226.
- Chater, N. & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, 38, 191–258.
- Cheng, P. W., Holyoak, K. J., Nisbett, R., & Oliver, L. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology*, 18, 293–328.
- Braine, M. D. S., & O'Brien, D. P. (Eds.). (1998). *Mental logic*. Lawrence Erlbaum Associates, Mahwah, NJ.
- Bucciarelli, M., & Johnson-Laird, P.N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, 23 (3), 247–303.
- Copeland, D. E., & Radvansky, G. A. (2004). Working memory and syllogistic reasoning. *Quarterly Journal of Experimental Psychology*, 57, 1437–1457.
- Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. N., & Bara, B. (1984). Syllogistic inference. *Cognition*, 16, 1–61.
- Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, 111 (3), 640–661.
- Khemlani, S., & Johnson-Laird, P.N. (2009). Disjunctive illusory inferences and how to eliminate them. *Memory & Cognition*, 35 (5), 615–623.
- Khemlani, S., & Johnson-Laird, P.N. (in press). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*.
- Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review*, 94, 211–228.
- Leevers, H.J., & Harris, P.L. (1999). Persisting effects of instruction on young children's syllogistic reasoning with incongruent and abstract premises. *Thinking and Reasoning*, 5 (3), 145–173.
- Neth, H., Khemlani, S., & Gray, W. (2008). Feedback design for the control of a dynamic multitasking system: Dissociating outcome feedback from control feedback. *Human Factors*, 50, 643–651.
- Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, 5, 349–357.
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5, 411–419.
- Polk, T.A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, 102, 533–566.
- Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.
- Stenning, K., & Van Lambalgen, M. (2008). *Human reasoning and cognitive science: Logical foundations for the psychology of reasoning*. Cambridge, MA: MIT Press.
- Walsh, C. R., & Johnson-Laird, P. N. (2009). Changing your mind. *Memory & Cognition*, 37 (5), 624–631.
- Wason, P.C. (1960). On the failure to eliminate hypothesis in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12, 129–140.
- Wason, P.C. (1964). The effect of self-contradiction on fallacious reasoning. *Quarterly Journal of Experimental Psychology*, 16, 30–34.

# When doing the wrong thing is right

David Kirsh

(kirsh@ucsd.edu)

Cognitive Science, 0515, UCSD  
La Jolla, CA 92093 USA

Richard Caballero

(riche117@gmail.com)

Cognitive Science, 0515, UCSD  
La Jolla, CA 92093 USA

Shannon Cuykendall

(scuykend@uci.edu)

Dept. of Dance, UCI  
Irvine, CA 9 92697 USA

## Abstract

We designed an experiment to explore the learning effectiveness of three different ways of practicing dance movements. To our surprise we found that partial modeling, called *marking* in the dance world, is a better method than practicing the complete phrase, called practicing *full-out*; and both marking and full-out are better methods than practicing by repeated *mental simulation*. We suggest that marking is a form of practicing a dance phrase aspect-by-aspect. Our results also suggest that prior work on learning by observation and learning by mental practice may not scale up to complex movements.

**Keywords:** Dance practice; Marking; Mental Simulation; Aspect-by-Aspect.

## Introduction

We report here on a surprising finding in an experiment that compared the relative effectiveness of three different ways of practicing dance phrases. We found that 1) partial modeling of a dance phrase by *marking* the phrase, as it is called in the dance world, is a better method than practicing the complete phrase, called practicing *full-out*; and 2) both marking and full-out are better learning methods than practicing by repeated *mental simulation*. This is surprising because when a dancer marks a phrase they are literally doing the wrong thing – like humming a piece of music instead of singing it. The result raises the interesting possibility that practicing a movement in a simplified manner, or aspect-by-aspect, rather than practicing all of its components at once, may be the best way to practice. In marking, subjects intentionally practice the phrase in an improper form, with distortions, exaggerations, simplifications, even with substitutions such as using hands for legs, or gestures for entire body movements, such as pirouettes. The official reason for marking is to save energy. But we believe that when cleverly mixed, this diversity may provide a powerful method for a dancer to explore the structure of a phrase more exhaustively than regular full-out practice.



Figure 1. The three conditions in the experiment.

This idea challenges common sense and previous work on complex motor learning. It is common sense that practicing something the way it should be performed ought to be more effective than practicing it with intentional distortions, or with essential components missing. If that were not so then repeatedly drawing a face in caricature rather than drawing it realistically ought to lead to drawing the face more realistically later. Similarly, practicing tennis strokes without a ball, or using the wrong approach and form ought to lead to better shots, at times, than always practicing in proper form. It is noteworthy that experiments have shown that both these marking-like methods are, at times, better forms of practice than always practicing in an undistorted, full way. In music performance, for example, using exaggeration in rehearsal is thought to be a helpful method of practicing, delivering results that surpass repeated full-out play [Hinz, 08]. Musicians practice passages both faster and slower than written. It is standard to manipulate phrasing, dynamics, articulation, intonation, and tempo, to name a few. [Chaffin et al 2002, Friberg et al, 06]. In sports viewing, [Hill & Pollack, 00; Pollack et al 01] found that subjects have learned to recognize complex actions better, such as certain types of tennis strokes, when some of the parts of the stroke have been exaggerated. Evidently, marking may have a place in training. But as a general method, practicing *only* distorted versions of the real thing, or versions that leave out essential components, is a counterintuitive method of rehearsal. Our unanticipated result is that this counterintuitive method is effective.

Our findings also challenge work on mental simulation. In sports psychology, imagery is often referred to as cognitive enactment or visualization, and is one of the most popular performance enhancement and rehabilitation techniques. It has been shown in numerous studies that mental simulation in sports contexts can significantly improve an athlete's performance on measures of style, speed and strategy. [Weinberg 08]. In music, Pascual-Leone [2001] reported a similar finding about learning to play a five-finger exercise on a piano keyboard. After five days, the group that mentally simulated playing, performed an exercise comparably to the third day level of those who practiced physically. All these experiments showed that mental practice leads to substantial improvement. We therefore came to the experiment believing our dancers would significantly benefit from their ten minutes of mental simulation.

Failure to find this improvement from mental simulation also bears on the findings of [Cross et al, 09], who, in several experiments, found that repeated observation of a target phrase – and hence ‘practice’ in the motor resonance system – leads to comparable performance to full-out physical practice. Simulation has been shown to facilitate in much the same way as observation – by activation of covert actions via the motor resonance system, [Jeannerod, 01]. The unexpected result by Cross et al [op cit] was found to hold for learning the rhythm and steps for pieces in a game like Dance Dance Revolution (DDR), where subjects must stamp their right or left foot onto footprints on a mat in time with music. Subjects watched the video repeatedly and may have played covertly. In our experiment, the phrases to be mastered were far more complex than DDR, involving movement of the entire body, with dynamics and feeling, and not driven in response to a stimulus. And they were simulated and not observed. But if observation works so well there is reason to suspect that mental simulation should not as well.

If our results about marking are true then marking during dance practice should not be seen as a sign of fatigue or laziness, as so often it is in dance studios. Rather, it may be a strategic method for selective training. This opens the door to developing more effective methods of selectively working on ‘aspects’ of a phrase. This likely applies to domains other than dance. We speculate that the success of marking also tells us something about how the body itself can be used to help manage attention, improve focus and even facilitate simulation in a selective way. The body may well draw attention to what is important in the way a hand in speed-reading drags the eyes along so that a speed reader can move through the page faster and more effectively. It is yet another way the body itself can be involved in cognition.

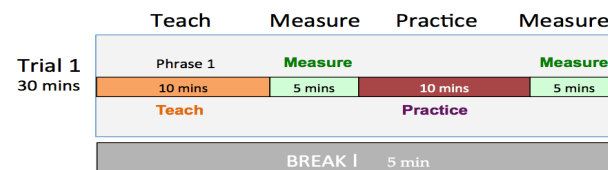
**Conjecture and Method.** In designing the experiment, our conjecture was that:

- practicing a dance phrase full-out would lead to better performance than mental simulation, and
- marking would lie somewhere in the middle: better than mental simulation but worse than full-out.
- Mental simulation would also lead to better performance.

Owing to the presumed power of the motor resonance system we wanted to see if anything *extra* would be gained by adding body activity to the mental simulation and projection we thought already occurred during marking. Our belief was that dancers would learn something from marking, just not as much as from practicing full-out. To test this idea we used the dancers from Random Dance, the contemporary company we have been studying. [Kirsh et al, 09; Kirsh, 12] All are super-experts, having been chosen from an audition pool of 800 professional dancers throughout Europe and the States.

**Procedure:** The design required dividing the ten dancers into three groups: A, B, C. All three groups were then brought into the studio and taught a dance phrase new to them, lasting about 55 seconds. The dancers were taught this phrase during a 10-minute teaching period, and at the end of it, the group left the studio and the dancers returned, one by one, to the studio and performed the dance in front of the teacher. As shown in figure 1 above, there were three conditions: practicing full-out, practicing by marking the phrase, and lying still mentally simulating the phrase. After the first round the dancers changed condition and were taught a second phrase of about the same duration and complexity as the first. The experimental design is a 3 by 3 Latin Square where each group is run in each condition. Thus, if group A started by Marking, they progressed to Full-Out, and then finished in the third trial in the Simulation condition.

Each dancer’s performance was graded according to established criteria (technicality, memory, timing, and dynamics – discussed below), first by the teacher in real-time and later by two independent expert observers who reviewed the video frame by frame. Once all dancers were graded, the group returned to the same large studio and practiced the dance for 10 minutes. While practicing they were asked to face in different directions and not look at each other. Once this 10-minute practice period was over they left the studio and, as before, returned one by one to be graded by the same criteria as before. See figure 2. Learning is understood as the change in grade acquired during the 10-minute practice phase.



**Figure 2.** The temporal structure of the experiment is displayed. There were three trials.

**Measures:** In mastering a dance phrase it is customary to be evaluated on technicality, memory, timing and dynamics.

**Technicality** means the level of precision in positions and transitions. Are the forms full and well-formed (e.g. juicy, fully rounded)?

**Memory**, or level of detail, refers to the completeness of each movement. Was something left out – a hand gesture, a turn, a foot angle?

**Timing** refers to the duration of individual steps and the duration of the transitions. Our timing coder used frame-by-frame measures of timing for great precision in comparing test conditions to the target standard.

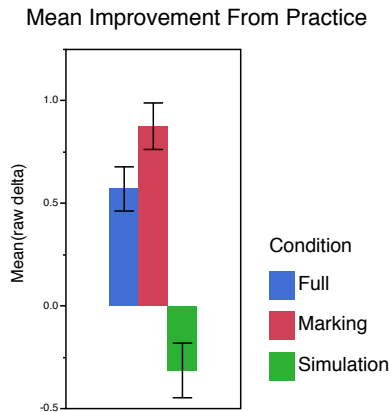
**Dynamics** refers to the force, speed and acceleration of movements. Also included are various qualities of motion – resistance, emotionality, and intentionality.

On analyzing the experimental results we found that:



- 1) Marking was the most effective method of practicing, being slightly more learning efficient than practicing full-out ( $p = .0189$ );
- 2) Both marking and full-out led to substantially more learning than mental simulation ( $p = .0001$ );
- 3) Mental simulation was not a strong form of practice; there was negligible improvement between pre and post tests in the simulation condition and in many cases it led to a decrease in performance.

Our finding both support and violate our hypotheses. We were correct that the learning achieved by marking is more effective than mental simulation (mean difference = 1.19, with  $p < .0001$ ) across the key dimensions of Memory, Technique and Timing. But we were surprised by its magnitude. We were greatly surprised that marking is more effective than Full-Out (mean difference = .31, with  $p = .0189$ ), though the difference is quite modest. We were also surprised that mental simulation did not facilitate at all. To compute these values we first performed one-way ANOVA's on all measures in all conditions and found highly significant differences throughout. We then ran pair-wise post-hoc comparisons (Tukey's HSD) and computed p values as shown in table 2. All p values were computed over z-scores to reduce noise caused by variability in dancers, measure-types, and graders.



**Table 1.** Mean improvement from practice (the learning delta), as measured on a 5-point scale. The absolute difference in delta between Marking and Full-out is 0.31, which is significant when measured by the z-score for Technicality, Memory and Timing ( $p = .0189$ ). Full is better for Dynamics but not significantly ( $p = .145$ ).

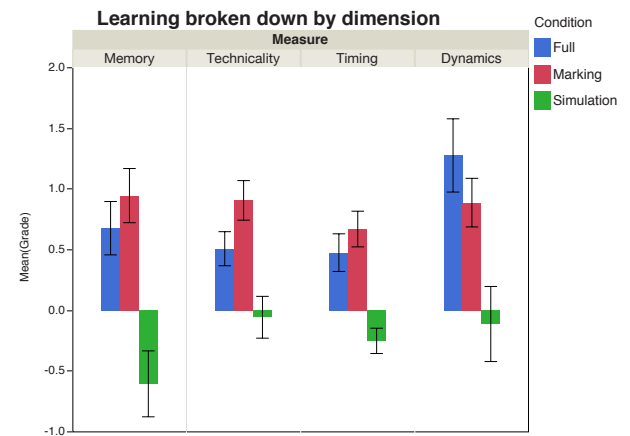
**Table 2 P values**

Measure	M>F	F>M	M>S	F>S
Memory	.7334		<.0001	<.0001
Technicality	.0029		<.0001	.0005
Timing	.0194		<.0001	<.0001
Dynamics		.145	.0003	<.0001
Mem, Tech, Timing	.0189		<.0001	<.0001

We assumed that marking would add something to mental simulation because somehow the process of marking would facilitate mental simulation rather than

interfering with it. Our main idea of a mechanism is that marking provides a physical anchor for mental simulation, thereby scaffolding imagination and leading to higher realism in simulation and increased priming of motor preparation. (See Kirsh, 10) We found qualitative support for this idea from interviews with the dancers. When asked what they think about when marking our subjects reported that they have in mind the full-out movement – though with fewer dynamics. They do not ‘see’ themselves as dancing in a distorted way, as they would if observing themselves in the mirror. They project off of their movement to the normative movement they want to be making. This is the movement they have in their mind’s eye. Marking seems to serve as a physical scaffold for projecting movement imagery. Thus, part of our conjecture was right: marking is better than simulation, though nothing we found proves our conjectured explanation of why it is better (i.e. projection). We were surprised, however, by just how much more effective marking is than mental simulation as a practice technique.

We also found that marking is better than full-out as a practice technique. This falsified our conjecture that full-out practice is the best form of practice.



**Table 3.** Marking was significantly better than Full-out for learning the aspects of a phrase related to technicality and memory and trending to significance in timing. Not surprisingly it was less effective at learning dynamics, which are rarely practiced in marking. Mental simulation was most effective (but still yielding zero or negligible improvement) for thinking about technical elements (precision in movement). It led to decreased performance – negative learning – for movement details.

**Marking vs. Full-Out Discussion:** There are a few possible explanations why marking is better than full-out. The simplest is that it is possible to mark more steps in a 10-minute period than it is to execute them full-out. To explore this idea we coded the video’ed activity of four dancers as they practiced their phrase in the experiment: two subjects in marking, were compared with two subjects in full-out, for each phrase. The results unambiguously show that the marking group performed significantly more steps and repetitions than the full-out groups. See table 4. The reason marking might be a

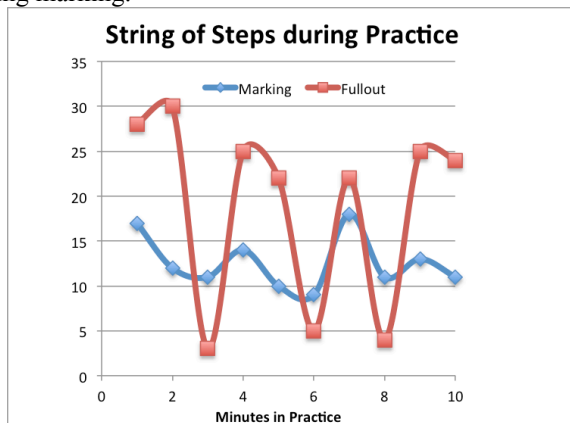


better way to practice, then, may be as simple as that dancers get in more trials in the same time by marking than by working full-out.

Phrase	Marking	Full-Out
I	351	275
II	317	300
III	317	188
<b>Mean</b>	<b>328</b>	<b>254</b>

**Table 4.** A simple enumeration of the number of steps executed in marking vs. full-out, matched by phrase.

Related to this number of steps argument is that marking might be a better form of practice because it is easier to fast forward or skip quickly through steps when marking, or to jump completely to new sequences. Full-out requires correct timing so there is no such thing as fast forward. Jumping to new sequences or sub-sequences is possible but it seems to be harder for dancers. From reviewing the steps that dancers practiced full-out we observed that the average sub-sequence was longer in full-out than in marking. See table 5. We observed this same phenomenon in actual dance sessions, where dancers jump to different parts of a phrase more often during marking.



**Table 5.** The string-of-steps, on average, is shorter for marking (12) than full-out (19). A string-of-steps is a sequence of steps performed in the right order. Dancers jump around within and between sequences more often when marking than when practicing full-out. In full-out, dancers alternated between very long and very short string-of-steps.

A second possible explanation of why marking is best is that in full-out practice there are more aspects to attend to at once. Not all aspects are equally in need of practice. Every step has many qualities. For instance, in Laban Movement Analysis [Newlove, 05], a distinction is drawn between ‘effort qualities’: flow (free/bound), weight (light/heavy), time (sustained/sudden), space (direct/indirect) and ‘shape qualities’: Rising/Sinking, Spreading / Enclosing, Advancing / Retreating, Growing / Shrinking. In practice, a dancer cannot attend equally to all these qualities simultaneously. Attention must be focused more narrowly. When practicing full-out,

however, dancers need to execute a movement as near to its full form as they can. This suggests that narrowing in on to a single aspect to practice will be harder because all aspects must be performed at once.

It seems, therefore, that marking offers dancers just what they want: a way of working on their movements aspect by aspect. Dancers do not think they are dancing incorrectly when they mark; they think they are dancing incompletely. They are focusing on some aspect of each step – its timing, extension, path or shape.

This ability to confine attention selectively may also explain why marking is better than full-out in remembering details. Intuitively, marking is akin to ephemeral sketching, instead of using a paper of pencil to sketch, dancers use their bodies, and the sketch is gone as soon as it made. But, ephemeral or not, dancers can still work on specific aspects of movement, the way their hands or feet specifically should move. They can cycle back to these parts while leaving everything else stationary. This is something dancers cannot do when dancing full-out. This reinforces the idea that by marking they can practice in a more incremental, piecemeal fashion than when practicing full-out. During one pass a particular aspect of a movement can be the center of attention, whereas another aspect can be the center of attention on a second pass. To be sure, the final conception of the target object requires the subject to integrate and assemble the aspects together in a unified whole. So there remains a puzzle about how a subject can come up with an effective whole movement from a set of disparate aspects that may interact in complex ways. This need for integration may impose limits on the effectiveness of marking as a learning method. But it also suggests that if aspects are relatively independent from each other, then marking can be an effective way of practicing because it facilitates a divide and conquer strategy: work on the problematic parts of a phrase and then assemble all parts into the final product. This is likely to be a more powerful method than practicing a target phrase holistically, whether through mental simulation or full-out.

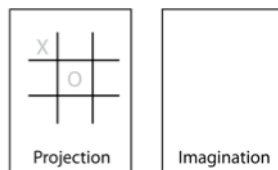
**Marking vs. Mental Simulation Discussion:** Prima facie, one reason mental practice – mental simulation – is less effective than marking is that when simulating, subjects do not receive sensory feedback from the body and the environment. In marking, by contrast, there is additional information available that a subject may use to reduce error. For instance, there is input about balance, gravity, weight and inertia. These physical features are not available through mental simulation, at least not in any realistic manner. This extra input from the physical world also means that dancers can re-evaluate their movement in a different way on the basis of how they interpret their perception of their own movement. The paradox of marking, however, is that the literal feedback from the body during marking is distorted feedback. Subjects are dancing incorrectly (in a very literal sense).

So the literal feedback they might use to determine an error measure, and so to sharpen their form, is not correct.

To resolve this paradox the place to start is with the dancers' own comment that when marking they have in mind their full-out movement, that marking is the physical scaffold for projecting to this normative imagery. To explain how an imperfect model of a movement – which is what marking literally is – can behave as a physical scaffold we need to introduce a few ideas. We begin with the concepts of projection, imagination and anchoring.

Projection is akin to attaching a mental image to a physical structure. When we project onto an object, we experience ourselves intentionally augmenting the object. The object anchors our mental image, and successful projection requires spatially locking the projected image onto the anchoring structure. To spatially lock, the mental image to be attached must be the right size and be connected to a specific location on the external structure.

When we *imagine* an object, we again are dealing with mental images but we do not attach it to anything in the seen world. It has no anchor and it need have no particular size. Mental simulation is a kind of imagination.



**Fig 4.** The differences between, projection and imagination can be understood as projecting an image of X and O onto a blank tic tac toe grid versus imagining an X and O on a grid while staring at a blank sheet of paper or better still, when blindfolded.

In Kirsh [09a], the results of running 20 subjects playing tic tac toe in the projection and imagination conditions was reported. To play the game all subjects first learned to name cells using 1 to 9 for a three by three board. They then would call out their move after hearing their opponent's. The numbers 1 to 3 were used for the top row, left to right, 4-6 for the middle row and 7 to 9 for the bottom row.

The results were not simple. Subjects did not play tic tac toe faster in the 3 by 3 condition in any condition, which we had predicted. Having a grid to anchor projection did nothing in the three by three game where subjects rarely needed to recall more than 5 or 6 moves. To challenge the subjects, we then taught them to play four by four games. Here the visual memory load is greater and we found that having a grid appears to facilitate subjects. Projection > Imagination ( $p = .002$ ). As predicted a grid now serves as an understructure or scaffold for projecting moves. However, given the unhelpfulness of a 3 by 3 scaffold it seems that the value of a scaffold increases with the complexity of task. In fact, scaffolding may be necessary for successful mental

simulation of harder problems. This may explain why Cross et al found observation to be facilitative in simple dance whereas we found that mental simulation of complex dance was not facilitative.

Anchoring projection, therefore, is one possible explanation why marking helps dancers. The major limitation of this view is that anchoring and projection are themselves inadequately understood. For instance, how does anchoring a projection differ from using an external structure or process as a mediator, an idea that regularly surfaces in discussion of the effects of culture and learning [Vygotsky, 78; Wertsch, 07]?

Here's a case in point. If a musician uses his foot to keep beat, does his tapping anchor his projection (and performance) or mediate it? In this instance, the reason to call it an anchor is that the target rhythm is not a regular beat per se – the rhythm he taps – it is the musical rhythm played 'on top of the beat' (e.g. da joom da joom, daah tika daah tika). He is thinking about the rhythm and using his beating as a stable pulse to help him. This is analogous with the tic tac toe grid, because, presumably, the beat is running on an automatic oscillator, [Eck et al, 00] liberating higher motor planning centers to work on different, but coordinated, sorts of covert actions. Beating is a way of scaffolding rather than mediating the correct rhythm.

Compare tapping a basic rhythm with the gestures an orchestra leader makes as he conducts a musical piece. Once again the underlying beat is embodied, though gesturally now rather than by tapping a foot. But a conductor also adds emphasis to help instrumentalists interpret the music. By gesturing a conductor directs musicians to attend to specific musical features. Are those gestures anchoring the musicians' projection? Or are they mediating their performance, without relying on a third thing called projection to help them perform? Projection seems a mental extra, pointless. The musicians can follow the conductor's directions immediately, without a further process of projecting what they need to do.

Contrast conducting with this last case. In [Frank & Barner, 12] elementary students in Gujarat, India, were taught to add and multiply using an abacus and then asked to perform calculations without the physical abacus. This practice, known as mental abacus, involves visual manipulations of an imagined abacus. Interestingly, when students work on their mental abacus they almost always flick their fingers, miming the movement of the beads. Performance suffers when abacus users are not permitted to use their hands (Frank & Barner, *ibid*; Hatano, 77]. Apparently, gesture plays a vital role in creating, or at least sustaining, mental abacus structures. Hand motions interact with the visual system, improving mental simulation. As before we cannot say whether this process involves projecting off of gestures or is better understood as some sort of meditational process. But projection seems the simpler account. Gestures scaffold mental imagery for the human calculator.

## Conclusion

In this study we set out to test whether marking is a more effective form of practice than mental simulation. It is. We also found that marking seems to be a better form of practice than the standard method of dancing full-out, and that mental simulation did not facilitate learning as it typically does.

When looking for the cause of marking's power we speculated that marking might function as the understructure for projection. When marking, a dancer creates a physical scaffold that facilitates projection. This would explain what 'extra' dancers get by physically marking a phrase rather than mentally rehearsing it. They get an external structure they can extrapolate from. This enables them to generate a conception of the final target that is more vivid, complete, and requiring less mental effort than the targets they imagine when they mentally rehearse without the support of overt movement. So it is not that a dancer is either marking or mentally simulating: marking is way to do mental simulation better.

We speculated further that mental simulation performed poorly because the target structure was a complex dance phrase about 1 minute long and this level of complexity exceeds most studies of the use of simulation.

Lastly, we conjectured that dancing is more effective than full-out because it allows dancers to focus on aspects of their movement rather than on all aspects at once, which is what is required during full-out. In music and most sports, it is customary to work on aspects of one's performance rather than working on everything all at once. Marking is tailor made for that purpose.

The success of marking warrants rethinking the best ways to practice motor activities.

**Acknowledgements.** We benefited from the and skillful help of Gina Bello, Leo Trottier, Ethan Soutar-Rao, Paul Zaino, the students in Cogs 160, and Wayne McGregor | Random Dance. Funding for this project under NSF: IIS-1002736 is gratefully acknowledged.

## References

- Chaffin, R., Imreh, G., & Crawford, M. (2002). Practicing perfection: Memory and piano performance. Mahwah, NJ: Erlbaum Associates.
- Cross, Emily, et al. (2009). Sensitivity of the Action Observation Network to Physical and Observational Learning. *Cerebral Cortex*. 19:315-326.
- Eck, D., Gasser, M., and Port, R. (2000). Dynamics and embodiment in beat induction. In P. Desain and L. Windsor (Eds.), *Rhythm perception and production*. Exton, PA: Swets and Zeitlinger
- Frank, M., Barner, D. (2012). Representing exact number visually using mental abacus. *Journal of Experimental Psychology: General*. Vol 141(1), pp. 134-149
- Friberg, A., Bresin, R., Sundberg, J., (2006) Overview of the KTH rule system for musical performance. Vol 2, No 2-3. pp 145-161
- Hatano, G., Miyake, Y., & Binks, MG. (1977). Performance of expert abacus operators. *Cognition*, 5, 47-55.
- Hill H. & Pollick F.E. (2000) Exaggerating temporal differences enhances recognition of individuals from point light displays *Psychological Science* Vol.11(3) pp 223-228
- Hinz, Bob. (2008). Practice Exaggeration for Large Intervals and Leaps. *Creative Keyboard*. <http://www.creativekeyboard.com/oct08/hinz.html>
- Jeannerod, Marc, (2001). Neural Simulation of Action: A Unifying Mechanism for Motor Cognition. *NeuroImage* 14, S103-S109
- Kirsh, D., et al. (2009) Kirsh, D., *Choreographic Methods for Creating Novel, High Quality Dance*. 5th International workshop: Design and Semantics of Form and Movement.
- Kirsh, D. (2010). Thinking with the Body, in (eds) S. Ohlsson R. Catrambone, *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society. Pp 2864-2869.
- Kirsh, (2012). How marking in dance constitutes thinking with the body. *Versus*.
- Krancioch, C. S., Mathews, P. Dean, A. Sterr. (2009). On the Equivalence of Executed and Imagined Movements: Evidence from Lateralized Motor & Nonmotor Potentials. *Human Brain Mapping* 30:3275-3286
- Newlove, J. & Dalby, J. (2005) *Laban for All*. Nick Hern Books, London.
- Pascual-Leone, A. (2001), *The Brain That Plays Music and Is Changed by It*. *Annals of the New York Academy of Sciences*, 930: 315-329
- Pollick F.E., Fidopiastis C. & Braden V. (2001) Recognising the style of spatially exaggerated tennis serves *Perception* Vol.30(3) pp 323-338
- Rhodes, Gillian; S. Brennan, S. Carey, Identification and ratings of caricatures: Implications for mental representations of faces, *Cognitive Psychology*, Volume 19, Issue 4, October 1987, Pages 473-497,
- Weinberg, R. (2008). Does Imagery Work? Effects on Performance and Mental Skills. *Journal of Imagery Research in Sport and Physical Activity*, 3, 1.
- Vygotsky, L., 1986. *Thought and Language*. The MIT Press, Cambridge, MA.
- Wertsch, J.V. (2007). Mediation. In H. Daniels, M. Cole, & J.V. Wertsch (Eds.), *The Cambridge companion to Vygotsky* (pp. 178-192). New York, NY: Cambridge University Press

# Tests and Models of Non-compositional Concepts

Kirsty Kitto and Peter Bruza

Information Systems School, Queensland University of Technology  
Brisbane, 4000, AUSTRALIA  
{kirsty.kitto,p.bruza}@qut.edu.au

## Abstract

The question of under what conditions conceptual representation is compositional remains debatable within cognitive science. This paper proposes a well developed mathematical apparatus for a probabilistic representation of concepts, drawing upon methods developed in quantum theory to propose a formal test that can determine whether a specific conceptual combination is compositional, or not. This test examines a joint probability distribution modeling the combination, asking whether or not it is factorizable. Empirical studies indicate that some combinations should be considered non-compositionally.

**Keywords:** conceptual representation; compositionality; context; probabilistic tests

## Conceptual Representation

Within cognitive science, the question of how to represent concepts is still being debated. Different positions have been put forward (e.g. the prototype view, the exemplar view, theory theory view), and Murphy (2002) contrasts some of these positions. He asks which is most supported by the various aspects of cognition related to conceptual processing, but concludes somewhat disappointingly, that “there is no clear, dominant winner”. Here, we take the position that it is possible to progress by asking a broader question about the nature of concepts; can they *always* be modeled compositionally? Or do they sometimes take a non-compositional form?

Some arguments for compositionality center around the systematicity and productivity of language; there are infinitely many expressions in natural language and yet our cognitive resources are finite. Compositionality ensures that this infinity of expressions can be processed, as it allows an arbitrary expression to be understood in terms of its constituent parts. Since compositionality is what explains systematicity and productivity, Fodor (1998) claimed that concepts *must be* compositional, however, this is at odds with prototypicality effects (Frixione & Lieto, [In Press]; Fodor, 1998). For example, consider the by now well known conceptual combination PET FISH. A “guppy” is not prototypical PET, nor a prototypical FISH, and yet a “guppy” is a very prototypical PET FISH (Hampton, 1997). Therefore, the prototype of PET FISH cannot result from the composition of the prototypes of PET and FISH, and so the characterization of concepts in prototypical terms is difficult to reconcile with compositionality (Hampton, 1997; Fodor, 1998). This supports a view put forward by Weiskopf (2007) when he observed that conceptual

combinations are “highly recalcitrant to compositional semantic analysis”.

Here, we take a novel approach to this debate, by providing a mathematical test which determines whether a conceptual combination can be considered compositionally, or not. We start with a consideration of what compositionality might mean probabilistically.

## Probabilistic Models of Compositionality

Figure 1 represents a basic probabilistic scenario involving a ‘black box’ composed of two proposed subsystems,  $A$  and  $B$ . What would it mean if this system was declared to be compositional? Acknowledging that it is the experiments which can be performed upon this system (and their likely outcomes) that will define this notion allows us to move beyond philosophy and into the realms of a mathematical definition.

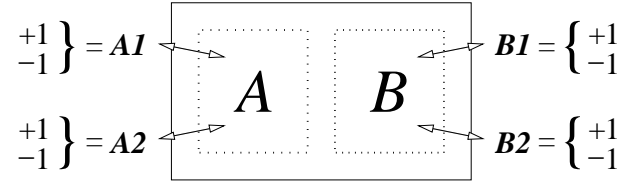


Figure 1: A potentially compositional system, consisting of two identifiable sub-components  $A$  and  $B$ . The system can perhaps be understood in terms of a mutually exclusive choice of experiments upon those sub-components, one represented by the random variables  $A1, A2$  (pertaining to an interaction between the experimenter and component  $A$ ), and the other by  $B1, B2$  (pertaining to an interaction between the experimenter and component  $B$ ). Each of these experiments can return a value of  $\pm 1$ , representing yes and no.

We define a compositional system as one which can be validly decomposed in such a manner that different experiments can be carried out upon each of its sub-systems, and that these will answer a set of ‘questions’ regardless of the experimental behavior of any other sub-systems. For the sake of simplicity, we shall assume that the answers to these questions are binary, they might be termed ‘yes’ and ‘no’, but are for generality labeled as  $+1$  and  $-1$ .<sup>1</sup> Standard probabilistic reasoning suggests that

<sup>1</sup>This assumption is more reasonable than it might at first appear: it is always possible to break a complex question into a set of simple binary questions, as the popular game

it is possible to describe this behavior in terms of four random variables representing the bivalent outcomes:  $\{\mathbf{A1}, \mathbf{A2}, \mathbf{B1}, \mathbf{B2}\}$ . What analysis can be brought to bear upon such a situation? As with many systems, the outcomes of our experiments will have a statistical distribution over all available outcomes, and it is possible to develop a set of probabilistic arguments about this scenario. For example, it is possible to consider the joint probability  $\Pr(\mathbf{A1}, \mathbf{A2}, \mathbf{B1}, \mathbf{B2})$  describing the likely behavior of our experimental black box, however, this very formulation forces us to consider what exactly a non-compositional probability distribution *would look like*. This paper is devoted to answering this question, but in order to approach the answer, we must first provide a model of non-compositional behavior, and this is not an easy task; almost all of our mathematical formalisms are based upon a notion of compositionality (Kitto, 2008). However, one mathematical model is widely accepted as non-compositional, Quantum Theory (QT), and so we take this formalism as the basis of our formulation of non-compositional behavior. Our reasons for this choice will become clearer from a psychological perspective as our argument progresses.

## Senses and Concepts

Our model takes Gärdenfors (2000) conceptual space as its starting point, extending this notion through the use of a vector space representation of concepts. For the purposes of this paper, we shall construct this representation through reference to the word association networks and vocabulary of the human mental lexicon, although this is not a necessary step for the formalism proposed; any sensible vector space construct would suffice if it has a similar structure to that discussed below. The Univer-

Associate	Probability	Associate	Probability
<b>ball</b>	<b>0.25</b>	<b>fighter</b>	<b>0.14</b>
cave	0.13	<b>gloves</b>	<b>0.14</b>
vampire	0.07	<b>fight</b>	<b>0.09</b>
fly	0.06	dog	0.08
night	0.06	<b>shorts</b>	<b>0.07</b>
<b>baseball</b>	<b>0.05</b>	<b>punch</b>	<b>0.05</b>
bird	0.04	<b>Tyson</b>	<b>0.05</b>
blind	0.04	...	...
animal	0.02		
...	...		

(a)

(b)

Figure 2: The free association data for two words, (a) bat, and (b) boxer. Both cases show a clear division of each concept into a sport sense (highlighted in bold), and an animal sense.

sity of South Florida (USF) word association data maps the strength of word associations displayed by a large sample of psychology students over a period of 30 years

of 20 questions illustrates. Quantum theory has provided a more sophisticated proof of this result using the Spectral Decomposition Theorem (Isham, 1995).

(Nelson et al., 2004). In Figure 2 we see a set of association strengths for two words, “boxer” and “bat”. Note the manner in which both words can be attributed a meaning that belongs to one of two *senses*; an animal sense and a sporting sense. Thus, we claim that the concepts BOXER and BAT are both ambiguous. Despite this ambiguity, humans are adept at recognizing the sense that is intended for an ambiguous word. They do this through reference to the context in which the word is being used, and this context might depend upon a wide range of factors (e.g. the co-occurrence of other words spoken before and after, the history of a conversation, the social context of the speaker). We note at this point that even our simple scenario has far more ambiguity than has appeared in the USF data (e.g. some people would interpret boxer as a pair of shorts, and someone could bat their eyes etc.), indeed, there are a wide range of very fine gradations in meaning that might be attributed to even these simple concepts. This added complexity can be dealt with in our model through an extension of the state space to higher dimensions, and through the use of a more sophisticated set of data<sup>2</sup> to construct the vector space model that we shall present.

In the next section we shall show that it is possible to construct a simple model of this ambiguity and its contextual dependency through use of the quantum formalism.

## A Quantum-Like Model of Word Associations

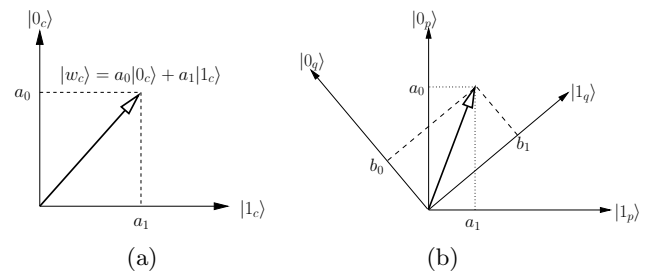


Figure 3: (a) A concept  $w$ , for example *bat*, is represented in some context  $c$  which takes the form of a basis  $\{|0\rangle, |1\rangle\}$ . (b) Changing the cue might change the chance of recall.

A simple model of the manner in which context might affect the interpretation that a subject ascribes to an ambiguous word can be constructed through the use of a *superposition state*, which is a novel concept of a state arising in Quantum Theory (QT). In Figure 3(a), an ambiguous word  $w$  is represented in some context  $c$ , as a superposition of recalled,  $|1\rangle$  and not recalled  $|0\rangle$  within the mind of a subject. When presented with a cue (rep-

<sup>2</sup>Such as the one being collected here: <http://www.smallworldofwords.com>

resented by the context  $c$ ) the subject might return word  $w$ , or not, with some probability. These probabilities can be estimated through reference to the online USF data,<sup>3</sup> which, in the context of a cue word “ball” suggests that a subject will recall the concept BAT with a probability  $Pr(BAT|ball) = .19$ , or they might recall something else ( $Pr(\overline{BAT}|ball) = .81$ ). We put this data into the quantum superposition state of Fig. 3(a) and so represent the cognitive state of our subject in the context of being presented the cue word “ball” as

$$|BAT\rangle_{ball} = \sqrt{0.81}|0\rangle_{ball} + \sqrt{0.19}|1\rangle_{ball}. \quad (1)$$

Figure 3(a) represents these probabilities *geometrically* using the measurement postulate of QT (Laloë, 2001; Isham, 1995), but this same state can be easily obtained through use of the Pythagorean theorem.

This simple model is made more interesting in Figure 3(b), where we have represented the fact that a different context (in this case a cue, but context could be a more complex semantic component) might result in a different set of recall probabilities. Thus, we could represent BAT as the superposition in the cognitive state of a subject when presented with the cue “cave”:  $\sqrt{0.94}|0\rangle + \sqrt{0.06}|1\rangle$  so giving a 6% probability that the word “bat” will be recalled by a subject who is presented with this different cue (or context). We see that the word “bat” is more likely to be retrieved from memory when a subject is presented with the cue “ball” than the cue word “cave”, and this change in probability can be obtained from the *same initial cognitive state* through a shift (i.e. a rotation) in the basis vectors representing the context in figure 3(b).

How should we consider the combination of two words in this model? While it is possible that a simple tensor multiplication of the two superposition vectors might suffice, this is not necessarily the correct mechanism (Bruza et al., 2009). Indeed, it seems possible that not all senses of a word remain accessible during conceptual combination. Thus, it might prove to be the case that a BOXER BAT is only ever interpreted by human subjects as “a small furry mammal with boxing gloves on”, or “a toy bat that a boxer dog chews on”, which would imply a case of perfect anti-correlation in the senses attributed by a subject to the combination. That is, considering the interpretation of the novel (i.e. non-lexicalized) conceptual combination BOXER BAT in the context of two priming conditions, one applied to each of the concepts in the combination (e.g. BOXER primed by “dog” and BAT by “ball”) we denote a concept which is recalled with the same sense as that for which it was primed as 1 and a failure to return in this sense by 0. For this scenario we might find that not all possible combinations of

the two senses in the combination can be realized. For example, we might find that a subject’s cognitive representation of BOXER BAT should be represented as

$$|BOXER\rangle \oplus |BAT\rangle = a|01\rangle + b|10\rangle, \text{ where } |a|^2 + |b|^2 = 1. \quad (2)$$

denoting a scenario where either BOXER has a sporting sense and BAT an animal sense ( $|01\rangle$ ) with probability  $|a|^2$ , or BOXER an animal sense and BAT a sporting sense ( $|10\rangle$ ) with probability  $|b|^2$ .

Such a cognitive state has profound consequences for the notion of compositionality. Indeed, QT has consistently shown that similar states *cannot* be interpreted compositionally (Isham, 1995; Laloë, 2001). Thus, if a similar set of experiments can be found that apply to human language processing, then this would give strong support for the claim that language cannot always be considered compositionally. The remainder of this paper will briefly sketch out recent work which attempts to test for such non-compositional conceptual behavior (Kitto et al., 2010, 2011; Bruza et al., 2012).

## Tests of (Non-)Compositionality

QT has a well developed suite of tests that can be applied to systems of the form shown in Figure 1, and these can be quickly adapted towards the the analysis of compositionality in language. For example, it is possible to construct a variation of the Clauser–Horne–Shimony–Holt (CHSH) inequality (Isham, 1995; Laloë, 2001), using an analysis derived from Cereceda (2000), which tells us that a system of this form can only be described as a combination of it’s subcomponents if:

$$\begin{aligned} 2 \geq \Delta = & |2(\Pr(\mathbf{A1} = +1, \mathbf{B1} = +1) \\ & + \Pr(\mathbf{A1} = -1, \mathbf{B1} = -1) + \Pr(\mathbf{A1} = +1, \mathbf{B2} = +1) \\ & + \Pr(\mathbf{A1} = -1, \mathbf{B2} = -1) + \Pr(\mathbf{A2} = +1, \mathbf{B1} = +1) \\ & + \Pr(\mathbf{A2} = -1, \mathbf{B1} = -1) + \Pr(\mathbf{A2} = +1, \mathbf{B2} = -1) \\ & + \Pr(\mathbf{A2} = -1, \mathbf{B2} = +1) - 2)| \quad (3) \end{aligned}$$

This formula (and a number of variations of it) has a substantial history in the physics and philosophy literature (Laloë, 2001; Shimony, 1984), and lack of space prevents a detailed explanation, however, we can briefly motivate its usage through a discussion of figure 1 and of the potentially compositional nature of the system it describes. Each subsystem  $A$  and  $B$  is represented by random variables:  $\{\mathbf{A1}, \mathbf{A2}\}$  and  $\{\mathbf{B1}, \mathbf{B2}\}$ , denoting whether a particular sense was observed (+1), or not (−1) under a given experimental arrangement. For this system, compositionality is expressed in terms of a *factorizable probability distribution*:  $\Pr(A, B) = \Pr(A) \Pr(B)$ . The syntax of this equation clearly shows how the model of the combined system is assumed to be expressed a product of the distributions corresponding to the individual subsystems  $A$  and  $B$ . When the inequality in (3) is violated, such a

<sup>3</sup>These numbers are obtained by finding the value for “bat” in the “cave” matrix that is depicted at <http://web.usf.edu/FreeAssociation/AppendixC/>.

compositionality assumption does not hold. Thus, the  $\Delta$  value in (3) gives us a clear criterion for deciding whether a given concept combination should be considered compositional or not. In a set of recent work (Kitto et al., 2010, 2011; Bruza et al., 2012), we have performed a number of experiments aimed at testing our formulation of the compositional hypothesis, and we shall now briefly discuss these results.

## Empirical Evaluation

We utilized four different priming regimes in order to generate the four different experimental scenarios required by Fig. 1. These experiments start by biasing subjects towards a particular interpretation of a non-lexicalized conceptual combination through exposure to words that have a particular *sense* representing the underlying concept. They then ask subjects to interpret the conceptual combination, and to designate the senses that they used in that interpretation. If conceptual combinations such as BOXER BAT are genuinely compositional, then it seems reasonable to assume that “vampire” primes BAT but has no priming effect on BOXER. A probabilistic analysis using (3) was performed upon the data obtained to test this assumption.

Table 1 lists the set of ambiguous conceptual combinations chosen, as well as the primes used to bias subjects towards each of two senses for the respective concepts. Primes were selected from the USF norms (Nelson et al., 2004) and the trials were composed of six phases.

**Phases 1-2:** Two consecutive double lexical decision tasks were carried out, where participants were asked to decide as quickly as possible whether two strings, a prime and the concept to be presented as a part of the compound given in Phase 3, were legitimate words, or if one of the strings was a non-word. Participants responded by pushing a button on the keyboard, labeled ‘word’ or a button labeled ‘non-word’ (left arrow and right arrow keys respectively). For instance, if given the strings “coil” and “spring”, then participants were expected to decide that both strings are words and so push the ‘word’ key, whereas if given “grod” and “church” then participants were expected to decide that they had been shown a non-word combination and to push the ‘non-word’ key. Each lexical decision consisted of the the two letter strings presented in the center of screen, one below the other. They were presented in this arrangement to discourage participants from interpreting the two words as a phrase. As soon as the participant responded, the screen was replaced by a blank screen for 800 ms, which was then immediately followed by the second lexical decision phase. The participant’s second lexical decision was followed by a 800 ms blank screen, and then immediately followed by phase 3. For example, one lexical decision task exhibited “coil” and “spring”, and was designed to prime the mechanical sense of the concept SPRING in

the conceptual combination SPRING PLANT. The order of the two double lexical decision tasks was counter-balanced, so that half were presented in the same order as the compound words (e.g., “coil” and “spring” are first presented, then “factory” and “plant”) and half were presented in the reverse order (e.g., first “factory” and “plant” are presented for lexical decision, followed by “coil” and “spring”). **Phase 3:** A bi-ambiguous conceptual combination was presented in the center of the screen (e.g. “spring plant”). Participants were asked to push the space bar as soon as they thought of an interpretation for the compound. Filler compounds were included for the filler (i.e. non-word) trials so as not to disrupt the participant’s rhythm in making two lexical decisions followed by an interpretation. **Phase 4:** Participants were asked to type in a description of their interpretation. **Phases 5-6:** Two disambiguation tasks were carried out, where participants choose what sense they gave to each word from a list (e.g., plant = A. ‘a living thing’; B. ‘a factory’; C. ‘other’). The order of test and filler trials were randomized. Participants completed 24 test trials and 24 filler trials, and the full procedure took 20-30 minutes. Experimental subcomponents utilizing non-words were discarded during the analysis presented here.

## Results

Table 1 lists a number of  $\Delta$  values, some of which violate equation (3). Confidence intervals around the CHSH value  $\Delta$  were computed using the bootstrap method that both removed and added data points that corresponded to interpretations that were either not present or added, and for each iteration, a pseudo  $\Delta$  was computed. Confidence intervals were computed using:  $\text{mean}(\text{pseudo}) \pm t_{0.975, n-1} \sqrt{\text{var}(\text{pseudo})/n}$ .

These results imply that there is good reason to believe that some conceptual combinations must be analyzed in a non-compositional framework. However, it is still possible to provide further details about *how* exactly the joint probability behaves during such a violation. In what follows, we shall analyze three specific examples from the three different categories of result: BOXER BAT (where  $\Delta < 2$ ); APPLE CHIP (where  $\Delta = 2$ ); and BANK LOG (where  $\Delta > 2$ ).

## Further Analysis

It is possible to write the joint probability in a form that starts to explain how violations of (3) occur. To do this we represent the four different random variables  $\{\mathbf{A1}, \mathbf{A2}, \mathbf{B1}, \mathbf{B2}\}$  in a matrix where each random variable contribution is split into a set of possible outcomes. This allows us to break down the results from Table 1 into a form that allows for a consideration of the underlying structure required for violations (or not) of (3). In this representation we can write the data gathered from the above experiments out as a set of joint distributions,



	Concept A		Concept B		Results	
Combination	Prime 1 ( <b>A1</b> )	Prime 2 ( <b>A2</b> )	Prime 3 ( <b>B1</b> )	Prime 4 ( <b>B2</b> )	$\Delta$	n
boxer bat	<i>dog</i>	<i>fighter</i>	<i>ball</i>	<i>vampire</i>	0.91 [0.74,1.09]	64
<b>bank log</b>	<i>money</i>	<i>river</i>	<i>journal</i>	<i>tree</i>	<b>2.13</b> [2.01,2.32]	65
apple chip	<i>banana</i>	<i>computer</i>	<i>potato</i>	<i>circuit</i>	2 [1.82,2.06]	65
stock tick	<i>shares</i>	<i>cow</i>	<i>mark</i>	<i>flea</i>	2.15 [1.98,2.41]	64
<b>seal pack</b>	<i>walrus</i>	<i>envelop</i>	<i>leader</i>	<i>suitcase</i>	<b>2.14</b> [2.01,2.32]	64
<b>spring plant</b>	<i>summer</i>	<i>coil</i>	<i>leaf</i>	<i>factory</i>	<b>2.29</b> [2.18,2.48]	64
<b>poker spade</b>	<i>card</i>	<i>fire</i>	<i>ace</i>	<i>shovel</i>	<b>2.15</b> [2.05,2.33]	65
slug duck	<i>snail</i>	<i>punch</i>	<i>quack</i>	<i>dodge</i>	1.41 [1.20,1.55]	63
<b>club bar</b>	<i>member</i>	<i>golf</i>	<i>pub</i>	<i>handle</i>	<b>2.28</b> [2.17,2.46]	64
web bug	<i>spider</i>	<i>internet</i>	<i>beetle</i>	<i>computer</i>	2 [1.82,2.06]	63
table file	<i>chair</i>	<i>chart</i>	<i>nail</i>	<i>folder</i>	0.38 [0.24,0.50]	63
<b>match bowl</b>	<i>flame</i>	<i>contest</i>	<i>disk</i>	<i>throw</i>	<b>2.21</b> [2.06,2.43]	64
<b>net cap</b>	<i>gain</i>	<i>volleyball</i>	<i>limit</i>	<i>hat</i>	<b>2.17</b> [2.04,2.39]	65
<b>stag yarn</b>	<i>party</i>	<i>deer</i>	<i>story</i>	<i>wool</i>	<b>2.24</b> [2.08,2.36]	61
mole pen	<i>dig</i>	<i>face</i>	<i>pig</i>	<i>ink</i>	1.44 [1.29,1.60]	63
battery charge	<i>car</i>	<i>assault</i>	<i>volt</i>	<i>prosecute</i>	2 [1.81,2.07]	63
count watch	<i>number</i>	<i>dracula</i>	<i>time</i>	<i>look</i>	1.54 [1.39,1.64]	65
bill scale	<i>phone</i>	<i>pelican</i>	<i>weight</i>	<i>fish</i>	1.77 [1.56,1.97]	64
rock strike	<i>stone</i>	<i>music</i>	<i>hit</i>	<i>union</i>	2.01 [1.84,2.18]	64
port vessel	<i>harbour</i>	<i>wine</i>	<i>ship</i>	<i>bottle</i>	1.53 [1.38,1.61]	65
crane hatch	<i>lift</i>	<i>bird</i>	<i>door</i>	<i>egg</i>	2.05 [1.89,2.24]	63
toast gag	<i>jam</i>	<i>speech</i>	<i>choke</i>	<i>joke</i>	1.23 [1.08,1.36]	63
star suit	<i>moon</i>	<i>movie</i>	<i>vest</i>	<i>law</i>	1.68 [1.50,1.84]	62
<b>fan post</b>	<i>football</i>	<i>cool</i>	<i>mail</i>	<i>light</i>	<b>2.13</b> [2.02,2.32]	63

Table 1: Results of the CHSH analysis:  $\Delta$  denotes the CHSH value with an associated confidence interval ( $\alpha = 0.05$ ),  $n$  the number of subjects. Conceptual combinations that significantly violate the CHSH inequality are bolded.

which allows for a further understanding of the resulting behavior.

For example, under this analysis, the joint probability of BOXER BAT can be written as (Bruza et al., 2012):

$$\begin{array}{c} \text{boxer} \end{array} \begin{array}{cc} & \text{bat} \\ & \begin{array}{cc} \mathbf{B1}(ball) & \mathbf{B2}(vampire) \\ +1 & -1 \\ -1 & +1 \end{array} \\ \begin{array}{cc} \mathbf{A1} & +1 \\ (dog) & -1 \end{array} & \begin{pmatrix} 0.43 & 0 & 0.1 & 0.2 \\ 0.28 & 0.28 & 0.1 & 0.6 \end{pmatrix} \\ \begin{array}{cc} \mathbf{A2} & +1 \\ (fighter) & -1 \end{array} & \begin{pmatrix} 0.13 & 0 & 0.31 & 0.06 \\ 0.33 & 0.53 & 0.31 & 0.31 \end{pmatrix} \end{pmatrix} \quad (4)$$

Here, we see no particular ordering or patterns when we compare equations (3) and (4). We can see that the data gathered does not center the distribution in such a way that it can violate the CHSH inequality.

In contrast, APPLE CHIP leads to a joint distribution that has a far more interesting structure:

$$\begin{array}{c} \text{apple} \end{array} \begin{array}{cc} & \text{chip} \\ & \begin{array}{cc} \mathbf{B1}(potato) & \mathbf{B2}(circuit) \\ +1 & -1 \\ -1 & +1 \end{array} \\ \begin{array}{cc} \mathbf{A1} & +1 \\ (banana) & -1 \end{array} & \begin{pmatrix} 1 & 0 & 0.73 & 0 \\ 0 & 0 & 0 & 0.27 \end{pmatrix} \\ \begin{array}{cc} \mathbf{A2} & +1 \\ (computer) & -1 \end{array} & \begin{pmatrix} 0.69 & 0 & 0.53 & 0 \\ 0 & 0.31 & 0 & 0.47 \end{pmatrix} \end{pmatrix} \quad (5)$$

In this case, we see a complete correlation between the subject responses. Thus, whenever a subject interprets APPLE as a fruit they decide that CHIP is a food, and when APPLE is interpreted as a computer then CHIP is

interpreted as an electronic device. This complete correlation of the senses attributed to the bi-ambiguous words leads to a value of  $\Delta = 2$ . This is still a compositional concept combination.

Finally, if we consider conceptual combination BANK LOG then we can see how a non-compositional value of  $\Delta > 2$  is obtained:

$$\begin{array}{c} \text{bank} \end{array} \begin{array}{cc} & \text{log} \\ & \begin{array}{cc} \mathbf{B1}(journal) & \mathbf{B2}(tree) \\ +1 & -1 \\ -1 & +1 \end{array} \\ \begin{array}{cc} \mathbf{A1} & +1 \\ (money) & -1 \end{array} & \begin{pmatrix} 0.94 & 0 & 0.53 & 0 \\ 0 & 0.06 & 0.07 & 0.4 \end{pmatrix} \\ \begin{array}{cc} \mathbf{A2} & +1 \\ (river) & -1 \end{array} & \begin{pmatrix} 0.21 & 0 & 0.8 & 0.13 \\ 0.03 & 0.76 & 0 & 0.07 \end{pmatrix} \end{pmatrix} \quad (6)$$

While this case is similar to the one illustrated in (5), it exhibits a key difference; a non-zero value has been returned by the ensemble of subjects for the off-diagonal case where  $\Pr(\mathbf{A2} = +1, \mathbf{B1} = -1) = 0.13$ , which corresponds to the case where the subjects interpret “bank log” as e.g. a financial institution made of wood. The off-diagonal term in (3) means that there is enough probability ‘mass’ for a violation. Comparing (4–6) with the set of equations typified by (3) we can understand that while it is necessary to for a system violating such inequalities to have some correlation in the random variables, it is just as important to have an anti-correlation.

## Conclusions

There is nothing in equation (3) that restricts its domain of application to quantum theory. Indeed, there are many systems that appear to be separated in a similar way, and so should adhere to the probabilistic behavior that it requires. Indeed, an early work by Aerts et al. (2000) proposed that the formalism of quantum theory could be widely applied to the description of a broad class of nonseparable systems, and this paper further contributes to this stream of work. More recently Bussemeyer et al. (2011) have applied the formalism of quantum theory to obtain a unified description of human decision making and the way in which it violates many of the axioms of standard probability theory. Together with the work presented here, these and many other results suggest that the formalism of QT is widely applicable to the analysis of psychological problems. We suggest that this is due to the ability of the formalism to incorporate a complex notion of *context* into its models, a significant advantage in cognitive modeling.

More specifically, the work presented here has considered only two possible senses for each word, and a very simple priming procedure, but we claim that this is not a limitation of the model *per se*. Firstly, the *spectral decomposition theorem* (Isham, 1995) implies that any measurement can be decomposed into a sum of projection operators. Secondly, more complex primes and cues can possibly be modeled through the use of a vector space approach that extracts the meaning of proceeding sentences, phrases and part word cues.

In summary, it seems likely that a broad class of systems which exhibit strong contextual dependencies among their subcomponents can be well modeled in this approach, and future work will seek to further clarify the conditions under which such systems become non-compositional.

## Acknowledgments

Supported by the Australian Research Council Discovery grants DP1094974, and DP0773341, and by the FP7 Marie Curie IRSES Project 247590 “QONTEXT”. Thanks also to the anonymous reviewers who’s comments strengthened this article considerably.

## References

- Aerts, D., Aerts, S., Broekaert, J., & Gabora, L. (2000). The Violation of Bell Inequalities in the Macroworld. *Foundations of Physics*, 30, 1387–1414.
- Bruza, P., Kitto, K., Nelson, D., & McEvoy, C. (2009). Is there something quantum-like about the human mental lexicon? *Journal of Mathematical Psychology*, 53, 362–377.
- Bruza, P., Kitto, K., Ramm, B., & Sitbon, L. (2012). The non-compositionality of conceptual combinations. (Under review)
- Bussemeyer, J. R., Pothos, E., Franco, R., & Trueblood, J. (2011). A Quantum Theoretical Explanation for Probability Judgment Errors. *Psychological Review*, 118(2), 193–218.
- Cereceda, J. (2000). Quantum mechanical probabilities and general probabilistic constraints for Einstein-Podolsky-Rosen-Bohm experiments. *Foundations of Physics Letters*, 13(5), 427–442.
- Fodor, J. (1998). *Concepts, Where Cognitive Science Went Wrong*. Oxford University Press.
- Frixione, M., & Lieto, A. ([In Press]). Representing concepts in formal ontologies: Compositionality vs. typicality effects. *Logic and Logical Philosophy*.
- Gagne, C. L. (2001). Relation and lexical priming during the interpretation of noun-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1), 236–254.
- Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. MIT Press.
- Hampton, J. (1997). Conceptual combination. In K. Lamberts & D. Shank (Eds.), *Knowledge, concepts, and categories* (p. 133–160). MIT Press.
- Isham, C. J. (1995). *Lectures on Quantum Theory*. London: Imperial College Press.
- Kitto, K. (2008). High End Complexity. *International Journal of General Systems*, 37(6), 689–714.
- Kitto, K., Ramm, B., Bruza, P. D., & Sitbon, L. (2010). Testing for the non-separability of bi-ambiguous words. In *Proceedings of the AAAI Fall Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes (QI 2010)*. AAAI Press.
- Kitto, K., Ramm, B., Sitbon, L., & Bruza, P. D. (2011). Quantum theory beyond the physical: information in context. *Axiomathes*, 21(2), 331–345.
- Laloë, F. (2001). Do we really understand quantum mechanics? Strange correlations, paradoxes, and theorems. *American Journal of Physics*, 69(6), 655–701.
- Murphy, G. (2002). *The big book of concepts*. MIT Press.
- Nelson, D., McEvoy, C., & Schreiber, T. (2004). The University of South Florida, word association, rhyme and word fragment norms. *Behavior Research Methods, Instruments & Computers*, 36, 408–420.
- Shimony, A. (1984). Contextual hidden variable theories and Bell’s inequalities. *British Journal for the Philosophy of Science*, 35, 25–45.
- Steyvers, M., & Tennenbaum, J. (2005). The large scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science*, 21, 41–78.
- Weiskopf, D. (2007). Compound nominals, context and compositionality. *Synthese*, 156, 161–204.

# Roles of Self Goal Setting in Insight Problem Solving

**Sachiko Kiyokawa (kiyo@isc.chubu.ac.jp)**

**Katsuyuki Hayashi (hw08075-2030@sti.chubu.ac.jp)**

Department of Psychology, College of Humanities, Chubu University  
1200 Matsumoto-cho, Kasugai-shi, Aichi, 487-8501, Japan

**Toshihiko Matsuka (matsuka.toshihiko@gmail.com)**

Department of Cognitive & Information Science, Chiba University  
1-33, Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522, Japan

## Abstract

Previous studies have shown that emphasizing the goal state could facilitate insight problem solving (e.g. Chronicle, MacGregor, & Ormerod, 2004). In these studies, the goal states were given by the experimenters and the participants were instructed to reach them. In the present study, we investigated whether the same facilitative effect could be obtained when the participants were forced to find the goal state by themselves. We used the 6-coin problem and compared the performance between the self goal setting condition and the control condition. The results showed that the participants in the self goal setting condition could solve the problem less often than those in the control condition when they were not allowed to reach the other goal state. It, however, slightly facilitated the insight problem solving if the participants were allowed to change the goal. The results indicated that self goal setting is effective in finding emergent goals.

**Keywords:** Insight problem solving; Goal setting, Plan

## Introduction

Problem solving is defined as an activity in which one tries to fill the gap between the initial and the goal state. It is well known that we adopt some heuristics in order to solve the problems. The hill climbing heuristic is one of the most common ones. It is the way of selecting operators so that the distance between the present state and the goal states can be minimized. To apply the heuristic, we need to know or at least to infer what goal state is. Although no one denies that goal plays a critical role in problem solving, it seems remain a matter of debate what roles the goal plays. In the present study, we investigate what roles the goal plays in insight problem solving.

## Importance of Goals in Insight Problem Solving

Although what processes underlies insight problem solving is still open (e.g. special process view vs. business as usual view), there is agreement that the goal plays an important role in insight problem solving. Kaplan and Simon (1990) applied the information processing framework to understand the process of insight problem solving. They argued that one uses some heuristics to narrow the problem space. MacGregor, Ormerod, and Chronicle (2001) have proposed the progress monitoring theory. They argued that hill climbing heuristic underlies the selection of moves to solve

the nine-dot problem. Ormerod, MacGregor, and Chronicle (2002) applied the theory to the 8-coin problem.

Hiraki and Suzuki (1998) proposed the dynamic constraint relaxation theory to explain the processes of insight problem solving. They hypothesized three types of constraints working during insight problem solving: object-level, relational, and goal. The object-level constraint is our natural tendency to encode objects at a basic level, although there are numerous other ways of interpretations. The relational constraint is a tendency to choose specific relations among innumerable alternatives. The word “relation” is defined as the manner in which objects are related to each other. The goal constraint is the ideal image, which provides feedback to the other two constraints by evaluating a match between the present and the desired states. They suggested that these constraints create an impasse at the earlier stage during insight problem solving and the incremental relaxation of the constraints driven by failures probabilistically causes qualitative transitions.

Wajima, Abe, and Nakagawa (2008) proposed the chaotic neural network model of the insight problem solving. Their model was implemented the goal orienting mechanism by which the model selects operators to minimize the gap between the present and the goal state. By comparing the models with and without the goal-orienting mechanism, they showed that the goal-orienting mechanism is necessary.

## Effects of Goals on Insight Problem Solving

The models mentioned above hypothesized that the goal plays a role as a criterion in evaluating the current states. When the goal state is explicitly shown, one can evaluate the present state easier and more accurately than when not. It can be expected that emphasizing the goal state facilitates the insight problem solving.

Suzuki, Miyazaki, and Hiraki (1999) examined whether emphasizing the goal state could be effective in solving the insight problem using the T-puzzle. The task was to arrange the four pieces such that they formed a T-shape. The goal state is essentially included and is explicitly shown in the original task instruction. In order to emphasize the goal state, they provided the T-shape template with the participants and asked them to match the pieces to it. The results showed that the solution rate in the template condition was higher than that in the control condition. It implied that reinforcing the goal constraint can be effective in insight problem solving.

Kojima, Ito, and Matsui (2008) investigated whether emphasizing the goal state could facilitate insight problem solving using the F puzzle. The F puzzle is to arrange the four pieces so as to make the F-shape. Along with the T-puzzle, the goal state is essentially included and is explicitly shown in the original task instruction. In order to emphasize the goal state, they provided the F-shape template with the participants and asked them to match the pieces to it. In addition to the template condition, they introduced the instruction condition, in which the participants were not provided any external aid and were required only to imagine the F-shape. The results showed that the solution rate in the template condition was higher than those in the instruction condition and in the control condition. Kojima et al. (2008) concluded that giving the template was effective in emphasizing the goal state and facilitating the top-down processing and that the top-down processing can be effective in insight problem solving.

Because these studies used the insight problem having a fixed goal, the participants had to reach it. However some insight problems, for example, the 6-coin problem, have more than one goal states. What roles does goal information play in solving the multi goal states problem? Chronicle, MacGregor, and Ormerod (2004) addressed the question using the 6-coin problem. They showed that the participants could reach the solution more often when they were given the visualized goal state than when were given only the original instruction.

### Purpose of the Present Study

Previous studies have shown that emphasizing the goal state could facilitate insight problem solving. In the previous studies, the information of the goal states were given by the experimenters and the participants were instructed to reach the goal state. In the present study, we investigate whether or not the same facilitative effect can be obtained when the participants are asked to find the goal state by themselves before performing the tasks. If emphasizing the goal state facilitated insight problem solving, we expected that self goal setting could be effective in insight problem solving as long as they set the goal state appropriately.

## Experiment 1

### Method

**Participants** Fifty-two undergraduates from Chubu University participated in the experiment and received a course credit following the completion of the experimental session. None have seen the 6-coin problem. They were randomly assigned to one of the two conditions: self goal setting and control. Twenty-eight participants were assigned to the self goal condition and 24 to the control condition.

**Task** The 6-coin problem (Chronicle et al., 2004) was used. The task was to rearrange the coins from the initial state

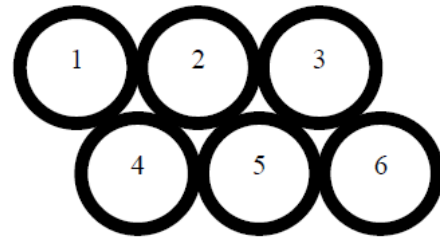


Figure 1: The initial state of the 6-coin problem cited from Chronicle et al. (2004). The numbers in the circles were not shown the participants.

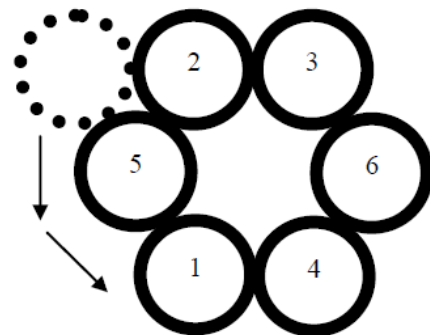
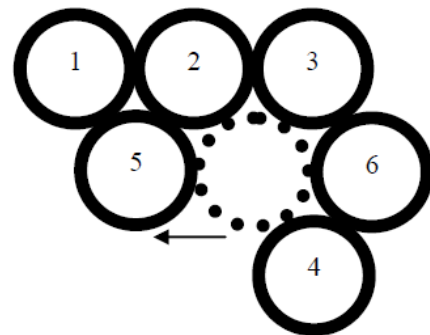
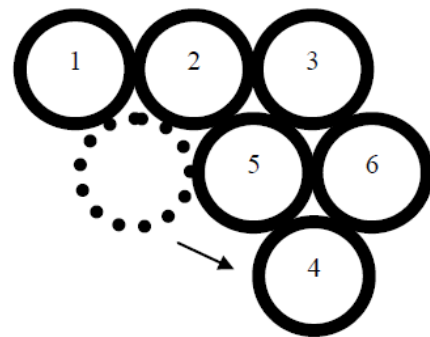


Figure 2: An example route to the ring goal state cited from Chronicle et al. (2004).

shown in Figure 1 such that each coin touched exactly two others following these four rules: (a) one can have three moves, no more and no fewer. (b) In each move, they have

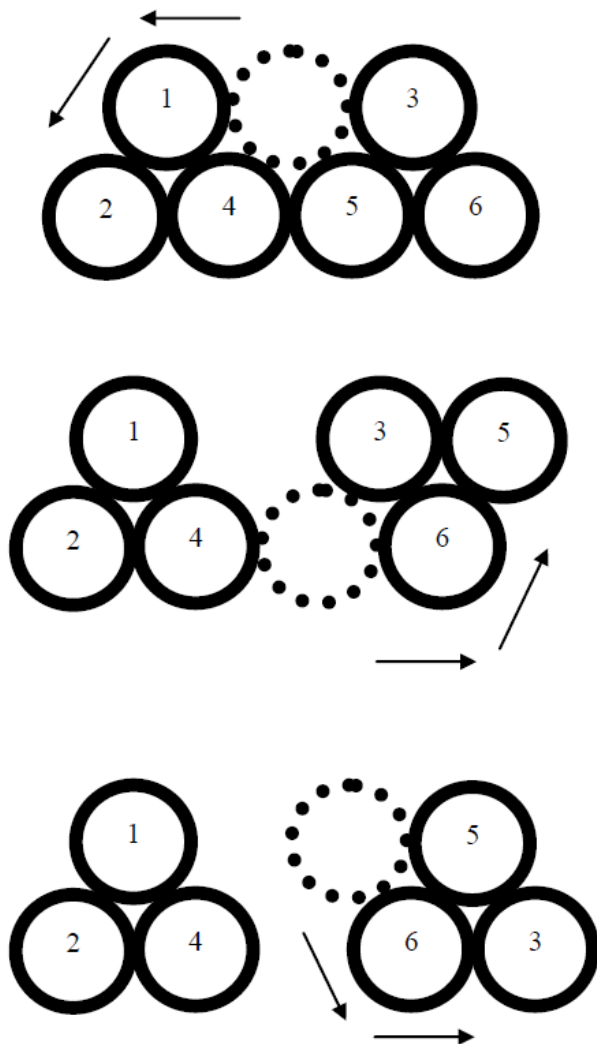


Figure 3: An example route to the 2-group goal state cited from Chronicle et al. (2004).

to slide one coin only. (c) When they slide a coin, it must not disturb any other coins. (d) At the end of each move, the moved coin must be touching two other coins. There could be the following two goal states: ring and 2-group. According to Chronicle et al. (2004), there are only two paths to the ring goal state and 176 to the 2-group one. An example solution path to the former goal state is shown in Figure 2 and the latter in Figure 3.

**Procedures** Participants were tested individually and their solution attempts were videotaped. For both conditions, participants were shown the initial state of the problem using 6 Japanese 500-yen coins. The participants in the self goal setting condition were asked to draw the goal state on a paper in three minutes and then to reach the goal state in 12 minutes. They were allowed only to reach the goal state they drew. The participants in the control condition were asked only to solve the problem in 15 minutes. The

experimental session was terminated when the participants found the solutions or when the designed time elapsed.

## Results and Discussions

Because a participant in the self goal setting condition was not able to draw the goal state within three minutes, the data was not included into analyses. As a result, 51 data was used for further analyses.

Firstly, we compared the performance between the self goal condition and the control condition. The performance in each condition is shown in Figure 4. The results showed that the participants in the self goal setting condition could solve the problem less often than those in the control condition (*Chi-square* ( $df=1$ ,  $N=51$ ) = 6.24,  $p < .05$ ).

Next, we examined the relationship between the goal states the participants depicted by themselves and the performance in the self goal setting condition. As shown in Table 1, 44.4% of the participants envisioned the inappropriate goals. Because the participants were restricted to the goal states they set, they could not reach the correct goal in principle. None of the participants who set the ring goal could reach the goal.

The participants in the self goal setting condition might not solve the problem because they set the inappropriate goals. Although the participants in the control condition might also search any paths to some inappropriate goals, they were able to change the goals if they wanted. On the other hand, those in the self goal setting condition were not allowed to change the goals even when they found the goal states inappropriate during problem solving. It might put them disadvantage situation.

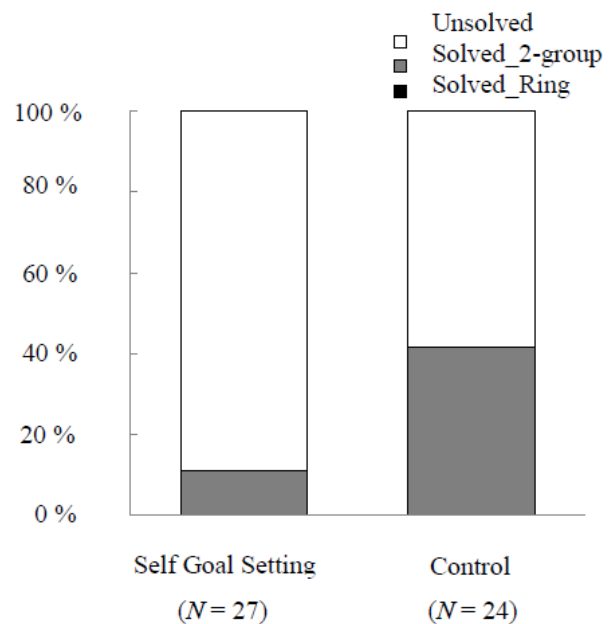


Figure 4: The performance in each condition.

Table 1: The relationship between the self set goals and the performance.

	Solved		Unsolved
	Ring	2-group	
Ring	0	-	10
2-group	-	3	2
Other	-	-	12

Another reason why those in the self goal setting condition could not solve the problem might be because it is more difficult to find any paths to the ring goal than the 2-group goal. Chronicle at al. (2004) have shown that there were only 2 routes to the ring goal whereas 176 to the 2-group goal. Thus, it was more difficult for those who could draw the ring goal state to find the routes to it.

## Experiment 2

Contrary to our expectation, in Experiment 1, self goal setting could not facilitate but disrupt the insight problem solving. Because some participants set the inappropriate goals and they were not allowed to change them, the situation might have negative effects on insight problem solving. If the inappropriate goal setting is cause of the disruptive effects, the disruptive effect will be diminished when the participants can change the goal state. In Experiment 2, we examine the effects of self goal setting on insight problem solving when the participants are allowed to change the goal state.

## Method

**Participants** Fifty-five undergraduates from Chubu University participated in the experiment. They received course credit for participation. None have seen the 6-coin problem. Twenty-seven participants were assigned to the self goal condition and 28 to the control condition.

**Task and Conditions** The task and conditions were the same as in Experiment 1, except that the participants in the self goal setting condition were allowed to reach not only the goals they set but also any other goals.

## Results and Discussions

Because a participant in each condition inappropriately finished the experimental session, these two data were excluded from the following analyses. As a result, 53 data was used for the analyses.

We compared the performance between the self goal condition and the control condition. The performance in each condition is shown in Figure 5. The results showed that the participants in the self goal setting condition could solve

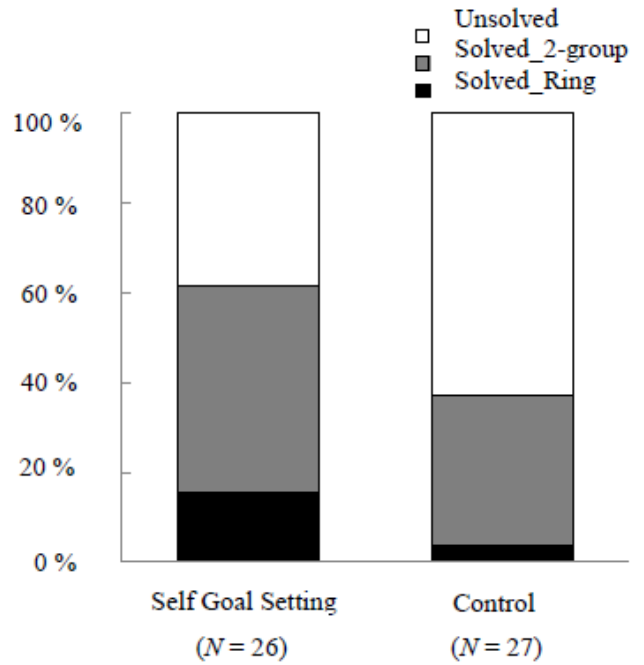


Figure 5: The performance in each condition.

Table 2: The relationship between the self set goals and the performance.

	Solved		Unsolved
	Ring	2-group	
Ring	4	4	4
2-group	0	5	1
Other	0	3	4

the problem more often than those in the control condition (*Chi-square* ( $df=1$ ,  $N=53$ ) = 3.18,  $p=.07$ ).

Next, we examined the relationship between the goal states the participants set by themselves and the goals they actually reached in the self goal setting condition. As shown in Table 2, 29.2% of the participants drew the inappropriate goal states. Unlike Experiment1, they were allowed to change the goals and 42.9% were able to reach the 2-group goal. Most of the participants who drew the 2-group goal state could find the paths to the 2-group goal. One-third of those who set the ring goal also reach the 2-group goal.

## General Discussion

In Experiment 1, self goal setting disrupted the insight problem solving contrary to our expectation. The results

were interpreted that because the participants could not set appropriate goal by themselves or because self goal setting itself disrupted the insight problem solving. In Experiment 2, self goal setting slightly facilitated the insight problem solving. However, many participants in the self goal setting condition reached the different goals from those they depicted. Self goal setting can be effective in insight problem solving when the goal works as a working hypothesis, but detrimental as a fixed criterion.

The theory of situated cognition predicted that we can find emergent solutions even when they search for a pre-defined goal. Suchman (1987), for example, argued that when a person takes a canoe in a rapid river, he/she may abandon his/her plan of how to go down in face of rapid currents, but the plan still has a role of orienting actions towards particular courses. It seems very similar to the processes observed in the self goal setting condition of Experiment 2. The participants who depicted the ring goal state at first tried to find the route to the ring goal and after some attempts they might found it too difficult to do. In the midst of the search, they might find another goal state, that is, the 2-group. As shown in Figure 2, the second step seems similar to the 2-group goal state. It might hint the participants that there can be another goal state and they might change the goal state. It can be said that the present study provided evidence supporting the notion the situated cognition theory pointed out.

The question to be addressed further is why the ring goal state set by themselves did not facilitate the insight problem solving whereas did when the experimenter gave the goal state in the Chronicle et al. (2004). The difference in effects of the goal on insight problem solving might be caused by source attribution effects. Several studies have shown that source of information has some effects on the performance. Schunn and Klahr (1993) investigated the effects of other-generated hypotheses on rule discovery. The results showed that giving the other-generated hypothesis led participants to investigate the plausibility of hypotheses more thoroughly and less false terminations with incorrect solutions. Kiyokawa, Ueda, and Okada (2004) experimentally clarified whether assessing other-generated hypotheses could facilitate hypothesis revision using a rule-discovery task. The results revealed that the participants who assessed the other-generated hypotheses before generating and assessing their own hypotheses performed better than those who generated their own hypotheses and assessed them thoroughly. Osman (2008) showed that seeing learning history of another participant facilitated transfer of acquired knowledge during the first task to the second one in implicit learning situation. In addition, she also showed that the facilitative effects was obtained even when the participants were provided with their own learning history in fact, only if they were told that they were derived from another participant. The results suggest that source attribution has the effect on transfer in implicit learning.

## Conclusion

The goal plays an important role in insight problem solving by directing the solvers' search. When it is not fixed, that is can be flexibly changed, emphasizing the goal state can facilitate insight problem solving. Even though the content of the goal state is the same, who set the goal can have effects on what effects is emerged: self or the other.

## Acknowledgments

This work was supported by KAKENHI (Grant-in-Aid for Scientific Research (C), 23500335). We thank Dr. Hajime Shirouzu for his helpful comment.

## References

- Chronicle, E. P., MacGregor, J. N., & Ormerod, T. C. (2004). What makes an insight problem? The roles of heuristics, goal conception and solution recoding in knowledge-lean problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 14-27.
- Garst, J., Kerr, N. L., Harris, S. E., & Sheppard, L. A (2002). Satisfying in hypothesis generation. *American Journal of Psychology*, 115, 475-500.
- Hiraki, K., & Suzuki, H. (1998). Dynamic constraint relaxation as a theory of insight. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, 5, 69-79.
- Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive Psychology*, 22, 373-419.
- Kiyokawa, S., Ueda, K., & Okada, T. (2004). The effects of other-generated hypotheses on scientific reasoning. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, 11, 228-238.
- Kojima, K., Ito, K., & Matsui, T. (2008). Effects of envisioning the goal state on insight problem solving. *Proceedings of the 25th Annual Meeting of the Japanese Cognitive Science Society*, (pp. 354-357).
- MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 176-201.
- Ormerod, T. C., MacGregor, J. N., & Chronicle, E. P. (2002). Dynamics and constraints in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 791-799.
- Osman, M. (2008). Positive transfer and negative transfer: Antilearning of problem solving skills. *Journal of Experimental Psychology: General*, 137, 97-115.
- Schunn, C. D., & Klahr, D. (1993). Self- vs. other-generated hypothesis in scientific discovery. *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 900-905). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Suchman, L. A. (1987). *Plans and situated actions: the problem of human-machine communication*. Cambridge University Press New York, NY.



- Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press New York, NY.
- Suzuki, H., Miyazaki, M., & Hiraki, K. (1999). Goal constraints in insight problem-solving. *Proceedings of the 2nd International Conference on Cognitive Science*, (pp. 159-164).
- Wajima, Y., Abe, K., & Nakagawa, M. (2008). A chaotic neural network model of insight problem solving with constraint. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, 15, 644-659.

# Differences in eye movements between same and other race face recognition.

**Eve K. Klama (eklama.enq@googlemail.com)**

Department of Psychology, University of Exeter  
United Kingdom

**Fraser Milton (f.n.milton@exeter.ac.uk)**

Department of Psychology, University of Exeter  
United Kingdom

## Abstract

Eye movements of twelve Caucasian participants were measured whilst they performed a recognition test of same (Caucasian) and other race (Indian) faces. We observed a standard other-race effect, with more items recognised correctly, fewer false alarms, and reduced reaction time to same-race than other-race faces. Additionally, a differential pattern of eye movements between races emerged. During the study phase, same-race faces were fixated more than other-race faces, whilst other-race faces resulted in a greater proportion of fixations to internal face features than same-race faces. At test, whilst no differences between races emerged in the number of fixations or in the proportion of fixations made to internal features, a significantly greater level of fixations were made to the left hemispace for other-race faces for both previously studied and lure faces. These differences in the pattern of fixation plausibly reflect the greater effort in the processing of other-race than same-race faces.

**Keywords:** other-race effect, eye-tracking, face recognition, same-race faces.

## Introduction

The other-race effect refers to the phenomenon that individuals are less proficient at recognising faces from a different race to their own. This is typically characterised by a 'mirror effect', whereby same-race faces attract a higher proportion of hits and a lower proportion of false alarms compared with other-race faces (Meissner & Brigham, 2001). As well as its theoretical importance, the other-race effect has considerable practical significance. In particular, it is well established that minority races in a community are more likely to be wrongfully convicted of a crime on the basis of erroneous eyewitness testimony (Meissner & Brigham, 2001). For instance, Scheck, Neufeld, & Dwyer (2000) found that most cases of mistaken eyewitness identification in the United States were of Caucasian victims misidentifying non-Caucasian suspects.

One explanation for the other-race effect is that individuals process other-race faces in a qualitatively different way to same-race faces. In particular, same-race faces are believed to rely on configural processes to a greater extent than other-race faces. Support for this comes

from the composite face effect (e.g., Michel, Rossion, Han, Chung, & Caldara, 2006) in which recognition of the upper half of a face was more disrupted by the bottom half of a different face for same-race than other-race faces. Furthermore, there is greater benefit for same-race than other-race faces from the whole face context when processing individual facial features (the whole/part effect, e.g., Tanaka, Kiefer, & Bukarch, 2004). Additionally, there is evidence that the face inversion effect is more pronounced for same-race than other-race faces (Rhodes et al., 1989; but see Valentine & Bruce, 1986). The greater exposure that people typically have to same-race faces (e.g., Chiroro & Valentine, 1995) may be one reason for the increased use of configural processes (Gauthier & Tarr, 2002). Similarly, perceptual learning (e.g., Gibson & Walk, 1956; McLaren & Mackintosh, 2000) may lead to same-race faces being less tightly clustered in multidimensional space than other-race faces, resulting in easier discrimination of same-race faces (as in "face-space" models, e.g., Byatt & Rhodes, 2004).

Eye movements provide an index to the allocation of visual attention towards facial features (Findlay & Gilchrist, 2003). As such, the pattern of eye movements can provide direct insight into cross-race processing differences. To date, however, eye tracking has seldom been used to explore the other-race effect. One exception is a study by Goldinger, He, & Pappas (2009) which examined eye movements to various features together with pupil dilation (an index of mental processing load in visual attention) for Caucasian and Asian faces, recruiting participants from both these races. Both Caucasian and Asian participants fixated more to same-race than other-race faces, and more to the eyes and hair for same-race faces and more to the nose and mouth for other-race faces. Pupil dilation was greatest for other-race faces, indicating the recruitment of greater resources. A separate study conducted by Blais, Jack, Scheepers, Fiset, & Caldara (2008) found a slightly different pattern of results, with Caucasians focussing on the eyes and East Asians focusing on the nose and mouth, regardless of the race of the faces viewed. As Goldinger et al. (2009) note, this discrepancy may be partially due to the fact that in their study they used faces with neutral emotions, whilst the faces presented in Blais et al. (2008) varied in expression.

Our study directly examined differences in the processing of internal features, relative to external features, for same-race compared to other-race faces. We also asked whether there would be cross-race differences in fixation

lateralization. These questions were not directly assessed by Goldinger et al. (2009) but have been investigated in studies tracking eye movements to famous/nonfamous faces (a familiarity benefit similar to the same-race advantage has been robustly established; see Johnston & Edmonds, 2009). For instance, Althoff and Cohen (1999) found that a greater proportion of fixations were delivered to nonfamous than famous faces in fame and emotion judgment tasks. Additionally, they showed that nonfamous faces evoked a greater proportion of fixations to the left hemispace than famous faces. Stacey, Walker, and Underwood (2005) showed that famous faces resulted in greater internal processing than nonfamous faces only under relatively restricted conditions (a matching-faces task). These findings are surprising given that behavioural work suggests that internal features are more important for familiar than unfamiliar face recognition (e.g., Ellis et al., 1979). Althoff and Cohen (1999) suggested that the greater processing of internal features for unfamiliar than familiar faces may reflect the necessity for more efficient sampling of information when viewing unfamiliar faces given that internal features are particularly useful for identifying people. They proposed a similar explanation for the greater left hemispace bias for nonfamous than famous faces - asymmetric viewing is more efficient due to the general symmetry of faces.

Our predictions were somewhat open-ended due to the lack of direct empirical investigation of these issues in the context of the other-race effect. However, given that internal features are regarded as more diagnostic and have greater involvement in configural processing than external features, more fixations to same-race than other-race faces would provide an explanation for the same-race recognition advantage. Nevertheless, this reasoning also applies to the face familiarity effect, yet Althoff and Cohen (1999) observed greater internal feature processing for nonfamous than famous faces. On these grounds, if one assumes that other-race faces can, on the whole, be regarded as less familiar than same-race faces, a greater reliance on internal features for other-race than same-race faces might be anticipated. For similar reasons, the results of Althoff and Cohen would suggest a greater reliance on left hemispace processing for other-race than same-race faces.

## Method

### *Participants*

Twelve Caucasian students (3 males, 9 females) from the University of Exeter, ranging in age from 18-39 ( $M = 22.92$ ,  $SD = 4.91$ ) participated. No participants had visited Asia for an extended period of time (i.e., over one month). Participants were tested individually in a testing cubicle.

### *Apparatus*

The experiment was run using E-Prime (Psychological Software Tools, 2002) on a Dell PC with a 22-inch color monitor and a standard computer keyboard. Participants sat 0.5 metres away from the screen.

The Eye link II system recorded movements in the right eye using a video-based eye tracker with a head movement compensation system connected to a Dell PC with a 17-inch TFT monitor. Eye movements were sampled on the recording computer at 500Hz. Pupil position was monitored via a miniature infrared CCD video camera mounted on an adjustable headband. The display computer initiated and terminated eye tracking recording on each trial.

### *Stimuli*

Stimuli consisted of forty colour photographs of male faces. Twenty Caucasian (taken from O'Toole et al., 2005) and twenty Indian faces (taken from Jain & Mukherjee, 2002) were used. All faces were full-face view, had neutral expressions, and no distinctive features (e.g., glasses, facial hair). Images were of comparable quality. Pictures were edited using Adobe Photoshop to achieve a resolution of 300-pixels wide; the height was constrained by the natural proportions of the face (average: 370.3-pixels). Faces were cropped, to remove the background of the image. The resulting images were presented centrally on a white background on a screen with a resolution of 800x600 pixels.

### *Procedure*

Our basic procedure was modeled on Experiment 2 of Stacey et al. (2005) with the difference that we manipulated the race of the faces rather than their familiarity. In the study phase, twenty faces were presented. Ten of these faces were Caucasian and ten Indian. Trials began with a black fixation cross for 500ms, followed by a blank screen for 500ms. A face was then presented for 5 seconds. No response was required but participants were instructed to remember the faces.

After a break of around 2 minutes, the test phase began. Here, forty faces were presented: the twenty studied faces (ten Caucasian, ten Indian) and twenty "lure" faces (ten Caucasian, ten Indian) which had not previously been seen. Faces were presented in a random order. Trials began with a black fixation cross lasting 500ms, followed by a blank screen for 500ms. A face was then displayed for 5 seconds. During this time, participants indicated whether they had seen the face previously (by pressing z on the keyboard) or if the face was new (by pressing m). If participants did not answer in time, no response was recorded and participants were encouraged to respond quicker.

In both the study and test phases, stimuli were presented in a random order. Eye movements were recorded in both phases, with corrections for drift conducted every 5 trials.

### *Analysis*

#### *Eye Movements*

Eye movements were analysed using EyeLink Data Viewer Software, which automatically detects saccadic eye movements and analyses these movements into individual fixations using a combined position/velocity/acceleration criterion (a saccade was defined as a period where eye velocity was greater than 30°/sec, eye acceleration was

greater than  $8000^{\circ}/\text{sec}^2$  and the eye had deviated at least  $0.1^{\circ}$  from its starting position). Fixations were defined as periods between saccades. Blink artefacts were automatically removed from the data.

Eye movements were analysed for the entire stimulus presentation period in the study phase. In the test phase, eye movements were analysed up until participants made their response. Using the EyeLink Data Viewer Software, Region of Interests (ROI) were created for internal (i.e., eyes, nose and mouth) and external (e.g., hair, ears) features. Furthermore, ROIs were created for the left and right hemispaces (including both internal and external features). ROIs were drawn separately for each face (c.f., Figure 1). The size of the ROI's between races was closely comparable. Fixations falling outside of the ROIs were excluded from subsequent analyses. In the analyses reported, for ease of exposition, we focus on the pattern of fixations; dwell measures showed a similar pattern.

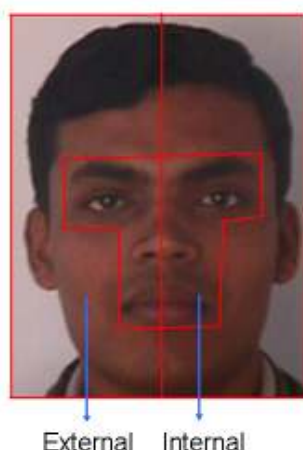


Figure 1. An illustration of how face regions were defined in terms of internal/external and right/left ROI for each face independently.

## Results

### Behavioural Results

#### Test Phase

A paired-samples *t*-test revealed that the hit rate for studied faces was significantly higher for same-race (Caucasian;  $M = .85$ ,  $SD = .14$ ) than other-race (Indian;  $M = .70$ ,  $SD = .18$ ) faces,  $t(11) = 2.51$ ,  $p = .029$ . Furthermore, the false alarm rate was higher for other-race ( $M = .35$ ,  $SD = .18$ ) than same-race ( $M = .08$ ,  $SD = .08$ ) faces,  $t(11) = 5.75$ ,  $p < .001$ . Response time was significantly higher for other-race ( $M = 1697.22$ ,  $SD = 478.65$ ) than same-race ( $M = 1407.15$ ,  $SD = 376.43$ ) faces,  $t(11) = 3.08$ ,  $p = .011$ . Time outs were minimal (on .004 of trials).

#### Eye movement data

##### Study Phase

Descriptive data of the pattern of eye-movement across the entire 5000ms study period is displayed in Table 1.

Paired-samples *t*-tests revealed that more fixations were made to internal than external features for both same-race,  $t(11) = 7.83$ ,  $p < .001$ , and other-race,  $t(11) = 9.51$ ,  $p < .001$ , faces. An additional paired-samples *t*-test, combining both internal and external features, showed that participants made significantly more fixations to same-race ( $M = 15.13$ ;  $SD = 3.71$ ) than other-race ( $M = 14.33$ ;  $SD = 3.68$ ) faces,  $t(11) = 2.78$ ,  $p = .018$ .

Table 1

*The Pattern of Eye Movements in the Study Phase Across the entire Study Period.*

	Same-race		Other-race	
	M	SD	M	SD
Number of fixations	15.13	3.71	14.33	3.68
Internal fixations	12.24	3.70	12.37	3.52
External fixations	2.88	1.31	1.97	1.25
Proportion of internal fixations	0.80	0.08	0.86	0.09
Right hemisphere fixations	7.25	3.45	6.52	3.46
Left hemisphere fixations	7.85	2.37	7.82	2.38
Proportion of left hemisphere fixations	0.53	0.16	0.56	0.16

Due to the significant difference between groups in terms of the mean number of fixations, we calculated the mean proportion of fixations to internal features (internal features/[internal features + external features]). Next, we divided the study interval into five separate 1000ms time bins to better characterise differences in the processing of other-race and same-race faces over time. This information is displayed in Figure 2. We then conducted a within-subject ANOVA with two factors, race (same-race and other-race) and time interval (0-1000ms, 1001-2000ms, 2001-3000ms, 3001-4000ms, and 4001-5000ms) to investigate this data. This yielded a significant effect of race,  $F(1,11) = 14.994$ ,  $p = .003$ ,  $\eta^2_p = .577$ , with the proportion of fixations to internal features significantly greater for other-race than same-race faces. There was, however, no significant effect of time interval,  $F(4,44) = .723$ ,  $p = .581$ ,  $\eta^2_p = .062$ , and no significant interaction between time period and race,  $F(4,44) = .070$ ,  $p = .991$ ,  $\eta^2_p = .006$ , indicating that the main effect of race remained consistent across time.

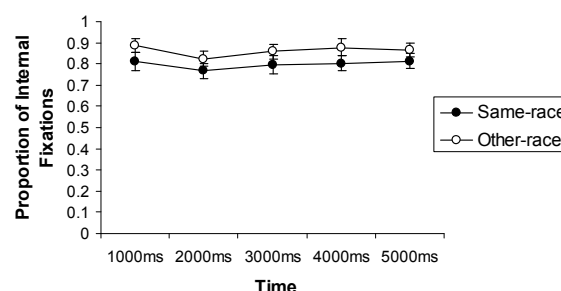


Figure 2. The mean proportion of fixations to internal face features across time intervals for same-race and other-race faces.

We calculated the mean proportion of fixations to the left hemisphere (left hemisphere/[left hemisphere + right hemisphere]) and conducted one-sample *t*-tests which

showed that fixations to the left hemispace did not differ from 0.5 (chance) for either same-race ( $M = .53$ ;  $SD = .16$ ),  $t(11) = .67$ ,  $p = .52$ , or other-race ( $M = .56$ ;  $SD = .16$ ),  $t(11) = 1.25$ ,  $p = .24$ , faces. Figure 3 displays the mean proportion of fixations to the left hemispace across the study interval for both same-race and other-race faces. A within-subject ANOVA with two factors, race (other-race, same race) and time interval (0-1000ms, 1001-2000ms, 2001-3000ms, 3001-4000ms, and 4001-5000ms) revealed that there was no significant effect of race,  $F(1,11) = 2.171$ ,  $p = .169$ ,  $\eta^2_p = .165$ . There was, however, a significant effect of time interval,  $F(1,11) = 5.783$ ,  $p = .001$ ,  $\eta^2_p = .345$ , with the proportion of fixations to the left hemispace decreasing across the study period. Furthermore, there was also a significant interaction between race and time interval,  $F(4,44) = 3.321$ ,  $p = .018$ ,  $\eta^2_p = .232$ . T-tests, assessing the nature of this interaction, indicated that other-race faces had a significantly greater proportion of fixations to the left hemispace than same-race faces for the first 1000ms time bin,  $t(11) = 2.702$ ,  $p = .021$ , but that there were no differences between face type for the other time periods (all  $Ps > .05$ ).

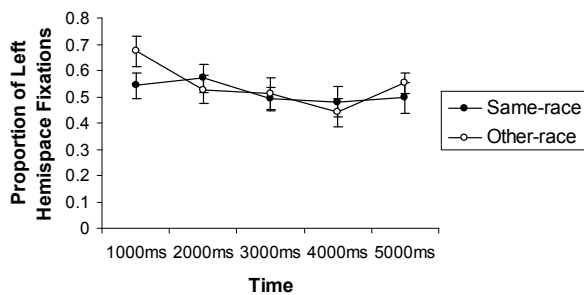


Figure 3. The mean proportion of fixations to the left hemispace across time intervals for same-race and other-race faces.

### Test Phase

Descriptive data for the test phase are shown in Table 2. As in the study phase, paired-samples t-tests revealed that there were no differences between same-race and other race faces for either studied,  $t = 1.399$ ,  $p = .189$ , or lure faces,  $t(11) = .466$ ,  $p = .651$ . There were more internal than external fixations for both “seen”,  $t(11) = 10.46$ ,  $p < .001$ , and “new”,  $t(11) = 8.16$ ,  $p < .001$ , same-race faces; similarly, there were more internal than external fixations for “seen”,  $t(11) = 8.37$ ,  $p < .001$ , and “new”,  $t(11) = 9.63$ ,  $p < .001$ , other-race faces.

As for the study phase, we calculated the mean proportion of fixations to the internal features. Unlike in the study phase, however, we did not partition the fixations into time intervals due to differences both within and between participants in the time spent viewing the faces. Table 2 shows the mean proportion of fixations to internal features for both other-race and same-race faces for both studied and lure faces. A 2x2 within subject ANOVA, with the factors

being race (other-race, same-race) and presentation type (studied faces, lure faces) found no significant effect of race,  $F(1,11) = .369$ ,  $p = .556$ ,  $\eta^2_p = .032$ , and no significant effect of presentation type,  $F(1,11) = .072$ ,  $p = .793$ ,  $\eta^2_p = .007$ . There was a marginal interaction,  $F(1,11) = 4.469$ ,  $p = .058$ ,  $\eta^2_p = .289$ , which did not reach significance.

Table 2

### The Pattern of Eye Movements in the Test Phase.

	Same-race				Other-race			
	Lure		Studied		Lure		Studied	
	M	SD	M	SD	M	SD	M	SD
Internal fixations	4.86	1.58	4.45	1.53	5.43	1.80	5.01	2.10
External fixations	0.66	0.78	0.66	0.49	0.77	0.68	0.55	0.64
Proportion of internal fixations	0.89	0.09	0.87	0.08	0.88	0.10	0.91	0.10
Right hemispace fixations	2.59	1.97	2.31	1.57	2.32	1.42	2.11	1.66
Left hemispace fixations	2.92	1.38	2.80	1.05	3.88	1.51	3.45	1.47
Proportion of left hemispace fixations	0.55	0.21	0.57	0.20	0.64	0.16	0.64	0.20

The mean proportion of fixations to the left hemispace for both studied and lure same-race and other-race faces are shown in Table 2. One sample t-tests showed that fixations to the left hemispace were reliably greater than chance (0.5) for both studied,  $t(11) = 2.551$ ,  $p = .027$ , and lure,  $t(11) = 2.960$ ,  $p = .013$ , other-race faces. There was no significant asymmetry, however, for either studied,  $t(11) = 1.200$ ,  $p = .256$ , or lure,  $t(11) = .768$ ,  $p = .459$  same-race faces. In terms of the proportion of fixations made to the left hemispace, a 2x2 within-subject ANOVA, yielded no significant main effect of presentation type (studied, lure),  $F(1,11) = .980$ ,  $p = .344$ ,  $\eta^2_p = .082$ , and no significant interaction between face type and race,  $F(1,11) = .148$ ,  $p = .708$ ,  $\eta^2_p = .013$ , but there was a significant main effect of race,  $F(1,11) = 8.619$ ,  $p = .014$ ,  $\eta^2_p = .439$ . Paired-samples t-tests revealed that the effect of race was significant for both studied,  $t(11) = 2.404$ ,  $p = .035$  and lure,  $t(11) = 2.274$ ,  $p = .044$ , faces with a greater proportion of fixations to the left hemispace for other-race than same-race faces.

## Discussion

We investigated the pattern of fixations of Caucasian participants while they performed a recognition task on a group of Caucasian (same-race) and Indian (other-race) faces. Consistent with previous studies, we found evidence for a “mirror-effect” (Meissner & Brigham, 2001); participants had fewer hits (recognising “seen” faces), more false alarms (incorrectly recognising “new” faces), and a longer response time for other-race than same-race faces.

In the study phase, we found, consistent with Goldinger et al. (2009), that participants made more fixations to same-race than other-race faces. Additionally, participants made a greater proportion of fixations to internal features for other-race than same-race faces. We divided the study period into five time periods to examine whether the nature of this effect changed over time. However, there was no indication of a time x face type interaction, which indicates that this

effect remained consistent throughout the study period. We also directly examined the lateralization of the fixations that participants made to same-race and other-race faces. Whilst there was no overall effect of race type, there was a significant effect of time with fixations becoming more equally divided between the left and right hemispace over time. Furthermore, we also found that there was a significant interaction between race and time, with a greater proportion of fixations made to the left hemispace for other-race than same-race faces for the first 1000ms time interval but not for the remaining four time intervals.

At test, stimuli were divided into those seen in the study phase, and “lure” stimuli which were only presented at test. For both types of faces there was no difference in the proportion of internal fixations to same-race or other-race faces. However, there was a significant effect of race for both studied and lure faces in terms of the proportion of fixations made to the left hemispace. Specifically, we observed a greater left hemispace bias for other-race than same-race faces.

The demonstration in the study phase that a greater proportion of fixations were made to internal features for other-race faces than same-race faces is consistent with previous work (e.g., Goldinger et al., 2009). In this regard, an explanation similar to that proposed by Althoff and Cohen (1999) to account for the greater proportion of internal fixations to nonfamous than famous faces seems applicable to our results. Specifically, Althoff and Cohen (1999) argued that there was greater need to effectively process the internal features, which are critical to face recognition, for nonfamous than famous faces. Our finding may, therefore, be due to the less efficient extraction of the internal features for other-race than same-race faces. The fact that participants made fewer fixations to other-race than same-race faces whilst trying to remember them supports the assumption that other-race face information is processed less easily (see also Goldinger et al., 2009). The more efficient processing of internal feature information for same-race faces would, consequently, provide greater opportunity to focus on external feature information, which still has informational value for recognition. This explanation is also consistent with Goldinger et al. (2009) who found greater pupil dilation for other-race than same-race faces, indicating greater processing effort for other-race faces.

Given this, it is striking that the internal feature bias for other-race faces appears relatively transient – this effect did not come close to reaching significance in the test phase. One might have reasonably expected a greater reliance on the more diagnostic internal features would also have been present in the test phase where the behavioural differences between same-race and other-race faces emerged. Instead, the effect was only detectable when participants were explicitly asked to encode the stimuli rather than when the requirement was to retrieve the stimuli. As such, this pattern of findings is in line with the idea that the greater bias to internal features for other-race than same-race faces reflects an encoding related perceptual process rather than a

retrieval-based process (for related behavioural evidence see Lindsay, Jack, & Christian, 1991; Tanaka et al., 2004; but see also Papesh, & Goldinger, 2009).

One caveat to the idea that the internal feature bias for other-race than same-race faces reflects an encoding rather than a retrieval process is that, as is common in face recognition studies (e.g., Goldinger et al., 2009; Stacey et al., 2005) the same picture of each face was shown at both study and test. This may have increased the reliance on pictorial codes during recognition rather than structural (abstracted memory representations) codes (Longmore, Liu, & Young, 2008), which are assumed to underlie face recognition outside the lab. This may, therefore, account for the difference between phases rather than the differing requirements of the study and test phases. However, as Longmore et al. (2008) note, if recognition was purely picture based, one might have expected equivalent recognition performance on other-race and same-race faces which was not the case in our study. Nevertheless, this potential issue could be addressed in future work by showing different photographs at study and test.

The pattern of fixations across the study and test phases was somewhat different to the internal feature effect. In the study phase, there was evidence of a greater bias to the left hemispace for other-race faces on first viewing the stimuli but this effect was not detectable over the remainder of the study period. In contrast, we found a left hemispace bias for other-race than same-race faces for both studied and lure faces in the test phase. Broadly speaking, therefore, these results indicate a greater lateralization asymmetry for other-race than same-race faces which was most marked in the test phase than the study phase. This may reflect that whilst the internal features effect appears to be due to encoding processes, the lateralization effect may reflect retrieval processes and is present only in the initial stages of encoding the stimuli. The demonstration that a greater proportion of fixations were made to the left hemispace for other-race than same-race faces during recognition is again similar to Althoff and Cohen’s (1999) finding of greater left hemispace viewing for nonfamous than famous faces. This finding can also be explained by postulating greater processing requirements for other-race than same-race faces. Under high processing demands, it appears efficient to focus primarily on the most diagnostic regions. This would mean a greater focus on one side of the face, given that faces are generally symmetrical (Althoff & Cohen proposed a similar explanation). We note that the limitations that Althoff and Cohen identified in their study concerning assessment of laterality effects (e.g., possible differences in texture or luminance between sides) similarly apply to our experiment. Future work could assess the generality of our effect by flipping one half of the face in a counterbalanced design such as in Rhodes (1985).

Previous work indicates that familiarity effects in face processing are influenced by task demands (Stacey et al., 2005). Our findings should, therefore, be generalised to different paradigms as well as to different races. Future

work should also include participants from different races to investigate cross-over interactions between recognition performance and the race of the participants. This would ensure that our findings are due to differences in cross-race face processing rather than factors such as the properties of the stimuli themselves. One issue with this sort of study, however, is that due to the extensive media exposure to Caucasian faces which means such faces are highly familiar to most populations, it might be preferable to carry out follow-up cross-over studies with non-Caucasian races. Nevertheless, our findings provide compelling evidence for processing differences between same-race and other-race faces. They also indicate that there may be overlap in the processes that underlie the same-race benefit and the familiarity advantage. Much remains to be understood, and this study should only be seen as a first step, but we hope that our experiment will help motivate future eye tracking work in this important area.

### Acknowledgments

This research was supported by the Great Western Research Initiative. We thank Chris Longmore for his assistance.

### References

- Althoff, R. R., & Cohen, N. J. (1999). Eye-Movement-Based Memory Effect: A Reprocessing Effect in Face Perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 997-1010.
- Blais, C., Jack, R.E., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture shapes how we look at faces. *PLoS ONE*, 3, e3022.
- Byatt, G., & Rhodes, G. (2004). Identification of own-race and other-race faces: Implications for the representation of race in face space. *Psychonomic Bulletin & Review*, 11, 735-741.
- Chiroro, P., & Valentine, T. (1995). An Investigation of the Contact Hypothesis of the Own-race Bias in Face Recognition. *Quarterly Journal of Experimental Psychology*, 45A, 879-894.
- Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition. *Perception*, 8, 431-439.
- Findlay, J.M., & Gilchrist, I.D. (2003). *Active Vision: The psychology of looking and seeing*. New York: Oxford University Press.
- Gauthier, I., & Tarr, M.J. (2002). Unravelling mechanisms for expert object recognition: Bridging brain activity and behaviour. *Journal of Experimental Psychology: Human Perception and Performance*, 28, 431-446.
- Gibson, E.J., & Walk, R.W. (1956). The effect of prolonged exposure to visually presented patterns on learning to discriminate them. *Journal of Comparative and Physiological Psychology*, 49, 239-242.
- Goldinger, S.D., He, Y., & Pappesh, M.H. (2009). Deficits in cross-race face learning: Insights from eye movements and pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 1105-1122.
- Jain, V. & Mukherjee, A. (2002). *The Indian Face Database*.  
vis. www.cs.umass.edu/~vidit/IndianFaceDatabase.
- Johnston, R.A., & Edmonds, A.J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, 17, 577-596.
- Lindsay, D.S., Jack, P.C., Jr., & Christian, M.A. (1991). Other-race face perception. *Journal of Applied Psychology*, 76, 587-589.
- Longmore, C.A., Liu, C.H., & Young, A.W. (2008). Learning faces from photographs. *Journal of Experimental Psychology: Human Perception and Performance*, 34, 77-100.
- McLaren, I.P.L. & Mackintosh, N.J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior*, 28, 211-246.
- Meissner, C. A., & Brigham, J. C. (2001). Thirty Years of Investigating the Own-Race Bias in Memory for Faces. *Psychology, Public Policy, and Law*, 7, 3-35.
- Michel, C., Rossion, B., Han, J., Chung, C-S., & Caldara, R. (2006). Holistic processing Is finely tuned for faces of one's own race. *Psychological Science*, 17, 608-615.
- O'Toole, A. J., Harms, J., Snow, S. L., Hurst, D. R., Pappas, M. R., Ayyad, J.H., & Abdi, H. (2005). A Video Database of Moving Faces and People. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 812-816.
- Pappesh, M.H., & Goldinger, S.D. (2009). Deficits in other-race face recognition: No evidence for encoding-based effects. *Canadian Journal of Experimental Psychology*, 63, 253-262.
- Psychological Software Tools. (2002). E-prime [Computer software]. Retrieved from <http://www.pstnet.com>
- Rhodes, G. (1985). Perceptual Asymmetries in Face Recognition. *Brain and Cognition*, 4, 197-218.
- Rhodes, G., Brake, S., Taylor, K. & Tan, S. (1989). Expertise and configural coding in face recognition. *British Journal of Psychology*, 80, 313-331.
- Scheck, B., Neufeld, P., & Dwyer, J. (2000). *Actual innocence*. New York: Random House.
- Stacey, P. C., Walker, S., & Underwood, J. D. M. (2005). Face processing and familiarity: Evidence from eye-movement data. *British Journal of Psychology*, 96, 407-422.
- Tanaka, J.W., Kiefer, M., & Bukach, C.M. (2004). A holistic account of the own-race effect in face recognition. *Cognition*, 93, B1-B9.
- Valentine, T. & Bruce, V. (1986). The effect of race, inversion and encoding activity upon face recognition. *Acta Psychologica*, 61, 259-273.



# Different Stable Patterns between Intra- and Inter-personal Systems: Experimental Study on Inter-limb Tapping Coordination

**Kentaro Kodama (kodamakentaro@nii.ac.jp)**

Department of Informatics, School of Multidisciplinary Science,  
The Graduate University for Advanced Studies, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan.

**Ryosaku Makino (ryosaku@nii.ac.jp)**

Department of Informatics, School of Multidisciplinary Science,  
The Graduate University for Advanced Studies, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan.

**Nobuhiro Furuyama (furuyama@nii.ac.jp)**

Information and Society Research Division,  
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan,  
Department of Informatics, School of Multidisciplinary Science, The Graduate University for Advanced Studies,  
and Graduate Schools of Tokyo Institute of Technology.

## Abstract

To reveal the differences between intra- and inter-personal systems in terms of the perceptual effect on the stability of inter-limb coordination, the present study conducted a finger-tapping experiment in in-phase and anti-phase mode. We investigated a between-subjects factor (the intra-/inter-personal condition), and a within-subject factor (the phase mode). In the intra-personal condition, participants bimanually tapped their index fingers, paced by a metronome, with the frequency gradually increasing from 1 Hz to 3 Hz. In the inter-personal condition, pairs of participants were asked to perform the same task, but to use their right or left index finger, while sitting side-by-side and looking at each other's fingers moving. Analysis showed that the average number of phase transitions, average time-to-transition and standard deviation of the relative phase differed between the intra-personal system and inter-personal system. Some results do not agree with the predictions of theoretical model proposed in previous studies on inter-limb coordination.

**Keywords:** Inter-limb coordination; Finger Tapping Task; Perceptual Effect; Dynamical Systems Approach

## Introduction

To reveal the differences between intra- and inter-personal coordination systems, a finger-tapping experiment was conducted in which two phase modes (in-phase and anti-phase) were manipulated as a factor. The present study had two objectives. One objective was to reveal the role of perceptual information in inter-limb coordination. For this, we conducted a finger-tapping experiment, comparing the intra-personal condition with the inter-personal condition. Because the tapping task required participants to utilize visual, auditory, and haptic information (e.g., looking at a moving finger, listening to auditory metronome stimuli as well as the sounds of tapping, and touching the surface of the desk), the effect of multi-modal perceptual information could be examined (Kodama and Furuyama, 2010). We

investigated not only the intra-personal system, which involves neural and mechanical coupling between limbs, but also the inter-personal system, which involves visual coupling through watching the other's movements. We could thus focus on the perceptual effect on inter-limb coordination. The other objective was to see how generally the existing model on human inter-limb coordination, called the Haken-Kelso-Bunz (HKB) model, could be applied to human inter-limb coordination in the intra-personal system and in the inter-personal system.

The first application of self-organization theory to human inter-limb coordination was attempted by Scott Kelso (Kelso, 1984) and his colleagues (Haken et al., 1985). Since then, research on this topic has developed tremendously. In this approach, called *the dynamical systems approach*, the general findings are that bimanual coordination is more stable in the in-phase mode than in the anti-phase mode; one of the most important observations is that phase transitions take place unidirectionally from the anti-phase mode to the in-phase mode when the required oscillation frequency reaches or exceeds a critical point. These findings led to the proposal of a theoretical model, called the Haken-Kelso-Bunz (HKB) model (Haken et al., 1985).

The above findings have been confirmed for inter-personal coordinated movement (e.g., swinging of pendulums or legs) (Schmidt et al., 1990, 1998). Phase transitions in inter-personal systems indicate that visual information involves coordinated movement because inter-personal systems do not involve mechanical or neural couplings between limbs, unlike intra-personal systems. Schmidt et al. (1998) suggested that the self-organization principle of intra-personal systems governs inter-personal systems as well but that the coupling strength between limbs is stronger in intra-personal systems.

However, most studies on coordinated movement have dealt with either intra- or inter-personal coordination of a pair of oscillators (fingers, legs, pendulums, etc.) wiggling or swaying in the air; that is, these studies did not address

the effect of haptic information in terms of contact on a surface of an environment.

On the other hand, in finger tapping studies, Mechsner et al. (2001) conducted an intra-personal four-finger tapping experiment in which the perceptual bias on bimanual coordination was investigated in terms of symmetry defined in visual, perceptual space. Although they found a perceptual bias on bimanual finger tapping, they did not reflect upon the implications of their findings on the HKB model. Some researchers (e.g., Takenaka and Ueda, 2003) conducted tapping experiments both in intra- and inter-personal conditions. These studies, however, did not compare the in-phase and anti-phase modes and did not investigate the frequency effect on the stability of movement.

As far as we can tell from a survey of the literature, no studies have compared intra-personal and inter-personal coordination in finger tapping from the perspective of dynamical systems as regards the in-phase mode and anti-phase mode and none has investigated the frequency effect as a *control parameter*. Therefore, we conducted a finger-tapping experiment to reveal the differences between intra-personal system and inter-personal system in terms of how multi-modal perceptual information affects the stability of tapping movements.

## Method

### Participants

A total of 30 healthy right-handed participants (15 males and 15 females) took part in the experiment, with 10 participants in the intra-personal condition and 10 pairs of participants (i.e., 20 participants) in the inter-personal condition. All participants were between 21 and 47 years of age (average age=26.9). The procedure was approved by the ethics committee at the National Institute of Informatics, where the experiment was conducted.

### Apparatus

A computer-generated metronome produced beeps, each lasting 85 msec. The metronome frequency was increased gradually from 1 Hz to 3 Hz over a 30-s trial after an initial 3-s period at 1 Hz. The metronome was run on a personal computer (Apple MacBook2130/13.3), and the beep sounds were conveyed to the participants through headphones (Sony MDR-NC600D) at a comfortable volume adjusted for each participant. A camcorder (TK-C1380; Victor), as part of the motion analyzer system DKH Frame-DIAS II, videotaped the movements of the participants' index fingers at 60 fields per second through the two-dimensional motion capture function of the Frame-DIAS II. The tapping movements and auditory stimuli were recorded on a hard disk drive (HDD) (Sony HVR-DR-60).

## Experimental Design and Procedure

The experiment was designed as a  $2 \times 2$  factorial with one between-subjects variable, i.e., the intra-/inter-personal condition (Figure 1), and one within-subject variables, i.e., the phase mode (in-phase and anti-phase) (Figure 2).

Participants were each seated at a desk in front of the camcorder. The task was to tap either in in-phase mode (two fingers tapping in synchronization) or anti-phase mode (two fingers tapping alternatively) at a pace dictated by the auditory metronome. The participants were instructed to keep their eyes open and watch their tapping movements during a trial, and to complete one full movement cycle, an extension-and-flexion cycle, for each beat of the metronome. They were also instructed to maintain the initial mode of coordination as much as possible, but not at the expense of losing pace with the metronome, and not to resist if they felt a change in the coordination pattern as a result of the increased tapping frequency. In the anti-phase mode, they tapped alternatively: the left finger was tapped in synchronization with the metronome beat while the right finger was in syncopation. Each condition was repeated four times, and the order of the trials was arranged randomly.

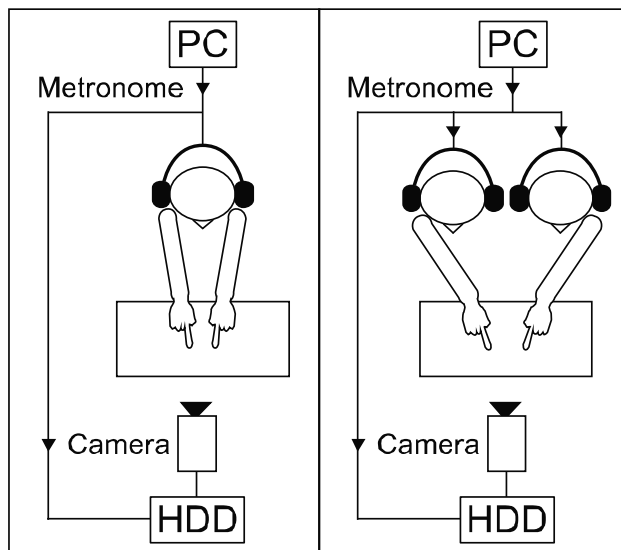


Figure 1: Experimental situation.  
(left; intra-, right; inter-personal condition)

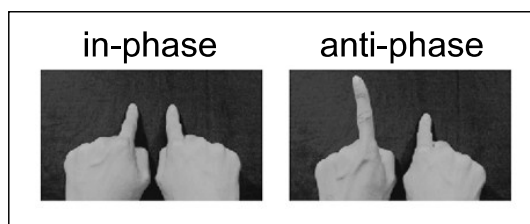


Figure 2: Phase mode.  
(left; in-phase mode, right; anti-phase mode)

## Data Analysis

In order to assess the stability of the tapping movement, we had previously analyzed the average number of phase transitions and average time-to-transition (Kodama and Furuyama, 2011). These indices, however, revealed only the total stability across a trial, i.e., *how many times* or *when* a phase transition occurred. In the present study, besides these two indices, we conducted a quantitative analysis of relative phase through a trial, to assess how progressively and how much fluctuation of movement increased or decreased. Relative phase is considered as an *order parameter* or, in other words, a *collective variable*, which represents the stability of a system. Thus, the phase transition should be described as a change in the relative phase from the perspective of the dynamical systems approach. In the present study, to assess the effect of frequency on performance, the relative phase was calculated and the standard deviation of the relative phase was obtained with respect to six 5-sec intervals.

**Average number of phase transitions** To measure the stability of movement in terms of temporally global stability, the average number of phase transitions was obtained as follows. When the right and left index fingers tapped together within a time window of 67 msec, derived from the video camera's frame rate, the tapping was categorized as being in the in-phase mode. When the time interval between one finger's peak of flexion (tap) and the other finger's peak of extension was within a time window of 67 msec, the tap was considered to be in the anti-phase mode. We considered that a phase transition would take place only if the taps in the opposite phase mode occurred at least five times in a row. The average number of phase transitions was obtained by counting the number of phase transitions in each of the phase mode conditions for each participant and each pair of participants.

**Average time-to-transition** We calculated the average time-to-transition in order to measure the temporally global stability, following a previous study's method (Riek and Wooley, 2005). The time-to-transition of a trial in which a phase transition occurred was defined as the interval from the start of the trial to the first tap in the opposite phase mode. The time-to-transition of a trial in which no transition occurred was defined as 30 s.

**Relative Phase Analysis** Besides two temporally global indices, we also calculated the relative phase between taps using standard procedures (e.g., Carson, 1995), in order to measure the temporally local stability of movement. To assess the effect of frequency on performance, each trial was separated into six equal time intervals of 5 s each (Riek and Wooley, 2005). The standard deviation of relative phase ( $SD \phi$ ) was calculated in each time interval.  $SD \phi$  reflects the stability of performance.

## Results

### Average number of phase transitions

Figure 3 shows the average number of phase transitions as a function of the intra-/inter-personal condition and phase modes. In the intra-personal condition, no transition was observed in the in-phase or anti-phase mode conditions; that is, there was no difference in the average number of phase transitions between two phase modes. In the inter-personal condition, no transition was observed in the in-phase mode condition, but it occurred an average of 2.5 times in the anti-phase one; that is, the transition occurred more often in the anti-phase mode than in the in-phase mode. In the anti-phase mode condition, transitions occurred more often in the inter-personal system than in the intra-personal condition.

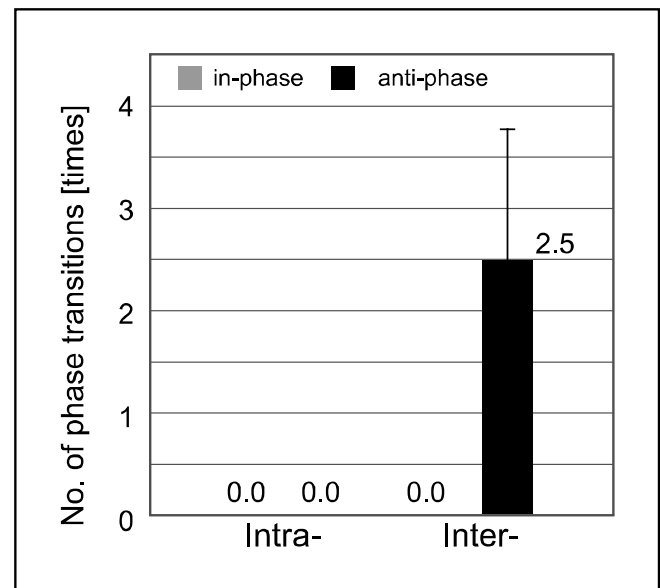


Figure 3: Average number of phase transitions.

### Time-to-transition

Figure 4 plots time-to-transition as a function of the intra-/inter-personal condition and phase modes. In the intra-personal condition, the time-to-transition was 30 sec in both modes; that is, the time-to-transition between the two modes had no difference. In the inter-personal condition, the time-to-transition was on average 29.9 s in the in-phase mode and 26.7 s in the anti-phase one; that is, it was shorter in the anti-phase mode than in the in-phase one. In the anti-phase mode condition, the time-to-transition was revealed to be shorter in the inter-personal system than in the intra-personal one.

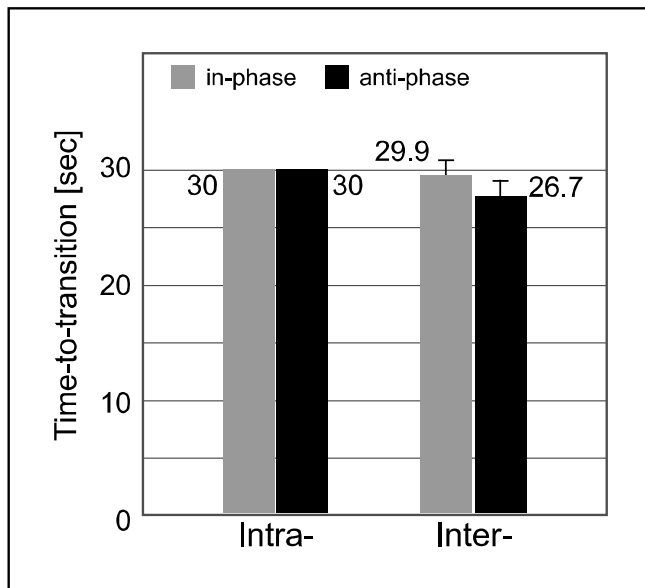


Figure 4: Average time-to-transition.

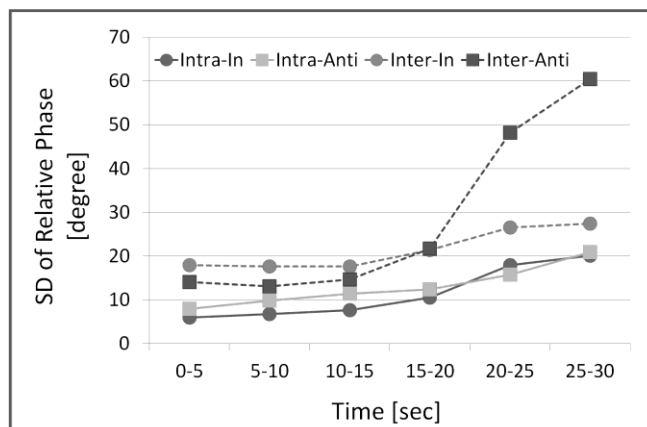


Figure 5: SD of relative phase.

## SD $\phi$

Figure 5 shows the standard deviation of the relative phase as a function of elapsed time of the trial (grouped into six five-second duration), equivalent to the frequency of movement. A three-way ANOVA (intra-/inter-personal condition (2)  $\times$  phase mode (2)  $\times$  frequency (6)) conducted on the SD  $\phi$  confirmed the main effect of the intra-/inter-personal condition ( $F(1,9)=48.186, p<.001$ ), phase mode ( $F(1,9)=34.097, p<.001$ ), and frequency ( $F(1,9)=100.841, p<.001$ ). It also revealed significant interactions: intra-/inter-personal  $\times$  phase mode ( $F(1,9)=14.338, p<.005$ ), intra-/inter-personal  $\times$  frequency ( $F(1,9)=16.996, p<.001$ ), phase mode  $\times$  frequency ( $F(1,9)=19.863, p<.001$ ), and intra-/inter-personal  $\times$  phase mode  $\times$  frequency ( $F(1,9)=28.746, p<.001$ ).

The simple main effect test for the intra-/inter-personal  $\times$  phase mode interaction revealed that there was no significant difference in SD  $\phi$  between in-phase mode and anti-phase mode in the intra-personal condition ( $F(1,9)=2.107, N.S.$ )

## Discussion

### Intra-personal condition

In the intra-personal condition, neither the average number of phase transitions, average time-to-transition nor SD  $\phi$  differed between the in-phase and anti-phase modes. This result did not agree with the prediction of the HKB model (Haken et al., 1985) that the stability of movement is higher in the in-phase mode than in the anti-phase one. It is possible that the range of the metronome frequency did not cover the critical frequency. But this is not the only possibility that contributed the result. We need to consider the contribution of 1) haptic feedback and 2) dynamical relation between inter-limb coordination systems and posture. On the basis of these two possibilities, we suggest two hypotheses to explain why such an unexpected result was obtained.

The first hypothesis concerns perceptual information (e.g., haptic information). Unlike the inter-limb coordination task that many previous studies conducted, e.g., wiggling fingers or swinging pendulums and legs, the participants in this study were required to tap on the surface of a desk, so haptic information was available at the time of the tapping. Kelso et al. (2001) reported that haptic information can stabilize finger extension-and-flexion movement; therefore, we suppose that this kind of haptic information may help to stabilize the anti-phase mode tapping movement.

The second hypothesis is that the posture may stabilize the finger tapping movement. In the tapping task, participants could support their postures by touching a desk, and cancel their body sway each time they tapped, so both the posture and limb movement might be less affected by each fluctuation. On the other hand, in the wiggling task, they could not cancel the body sway caused by the limb extension-and-flexion movement, and the instability of the posture might cause the phase transition in the anti-phase movement.

### Inter-personal condition

In the inter-personal condition, all indices of the average number of phase transitions, average time-to-transition, and SD  $\phi$  showed that the stability of movement was higher in the in-phase mode than in the anti-phase one. This result was in agreement with the results of the previous study (Schmidt et al., 1998); that is, visual information might involve organization and stabilization of coordinated finger tapping movements. Moreover, as in the previous study (Schmidt et al., 1998), one participant in the inter-personal

tapping experiment can always perceive the metronome beats as auditory information; that is, “on the beat” situation. On the other hand, the other participant must always tap at the midpoint between the metronome beats; that is “off the beat”. The possibility still remains that the “off the beat” participant might couple with not visual information, i.e., the partner’s finger’s motion, but rather auditory information, i.e., the metronome beats. Further experiments have to be done in order to reveal which kind of perceptual information (e.g., visual or auditory) is involved in organization and stabilization of the inter-personal coordination systems.

### Future direction

Recently, some researchers have referred to social coordination or joint-action in the inter-personal coordination paradigm from the cognitive perspective and the behavioral dynamics perspective (Schmidt et al., 2010). The cognitive perspective supposes there is a common coding of perception and action; that is, the same representations are used to perceive and perform an action. Additional evidence for such a common coding comes from discovery of mirror neurons in monkeys (Iacoboni et al., 1999), and mirror system in humans (Calvo-Merino et al., 2005). Although such mirror/common coding systems might be important for forming simple inter-personal coordination or imitating another’s behavioral patterns, it is difficult to explain how such a coordination or synchronization occurs *in time*. Even if perception and action coding occurs in mirror systems in the brain, we cannot share those representations directly, because the central nervous systems are not connected to each other. Therefore, we suppose that not only cognitive approaches based on neuroscience but also behavioral dynamics perspectives inspired by the dynamical systems approach are needed to reveal the coordination mechanisms in inter-personal systems. To do this, it will be important to identify which kind of perceptual information is involved in organization and stabilization of inter-personal coordinated movement. Recently, Ulzen et al. (2010) tried to apply the HKB model to inter-personal coordination during a treadmill walking task and confirmed that the dynamical model for rhythmic inter-limb coordination does not readily apply, at least not generically or robustly, to inter-personal coordination when people are walking side-by-side on a treadmill. These results suggested the possibility that the HKB model does not necessarily apply to every system.

As important as identifying the perceptual information available in the inter-personal system, it is also important to compare the intra-personal and inter-personal coordination systems. Coey et al. (2011) attempted to compare these two systems and evaluate the relationship between the stability of intra-personal coordination and the emergence of spontaneous inter-personal coordination. However, against their hypothesis that the stability of the intra-personal coordination patterns would affect the emergence of inter-

personal coordination in a pendulum-swinging experiment, the stability of the intra-personal coordination patterns did not affect the emergence of inter-personal coordination, and the emergence of inter-personal coordination did not affect the stability of the intra-personal coordination patterns (Coey et al., 2011). The present study did not clarify whether or not the stability of the intra-personal coordination patterns and the emergence of inter-personal coordination influence each other in the finger-tapping task. In the future, we should evaluate the possibility of such influences by means of analyzing the stabilities of both intra-personal and inter-personal systems.

### Conclusions

The present study investigated the differences between intra- and inter-personal systems in terms of the perceptual effect on the stability of inter-limb coordination in a finger-tapping experiment. A between-subjects factor (intra-/inter-personal system) and a within-subject factor (phase mode) were investigated. Standard deviation of relative phase revealed that the stability of the tapping movement differed for each factor and significant interactions among these factors. The stable pattern of the intra-personal system was different from that of the inter-personal system.

These findings require us to reconsider the perceptual effect on inter-limb coordination and its theoretical model. In order to classify complicated factors (e.g., the effect of haptic information on anti-phase tapping movement in the intra-personal system), we should conduct further experiments in the future.

### References

- Calvo-Merino, B., Glaser, D., Grezes, J., Passingham, R., & Haggard, P. (2005). Action observation and acquired motor skills: An fMRI study with expert dancers. *Cerebral Cortex*, 15, 1243–1249.
- Carson, R. (1995). The dynamics of isometric bimanual coordination. *Experimental Brain Research*, 105, 465–476.
- Coey, C., Varlet, M., Schmidt, R., & Richardson, M. (2011). Effects of movement stability and congruency on the emergence of spontaneous interpersonal coordination. *Experimental Brain Research*, 211, 483–493.
- Haken, H., Kelso, J., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, 51, 347–356.
- Iacoboni, M., Woods, R., Brass, M., Bekkering, H., Mazziotta, J., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, 286, 2526–2528.
- Kelso, J. (1984). Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology – Regulatory*, 15, R1000–R1004.

- Kelso, J., Fink, P., DeLaplain, C., & Carson, R. (2001). Haptic information stabilizes and destabilizes coordination dynamic. *Proceedings of The Royal Society of London*, 268, 1207-1213.
- Kodama, K., & Furuyama, N. (2010). The effect of auditory information on the stability of interpersonal tapping movement. *Proceedings of the 12th SICE SI2011 Annual Conference* (pp.1294-1297). Sendai, JP: Tohoku University.
- Kodama, K., & Furuyama, N. (2011). Comparing intra- and inter-personal coordination systems: perceptual effect on stability of finger tapping movement. *Proceedings of the 2011 IEEE/SICE International Symposium on System Integration*, E7-5. Kyoto, JP: Kyoto University.
- Mechsner, F., Kerzel, D., Knoblich, G., & Prinz, W. (2001). Perceptual basis of bimanual coordination. *Nature*, 414, 69-73.
- Riek, S., & Woolley, D. (2005). Hierarchical organisation of neuro-anatomical constraints in interlimb coordination. *Human Movement Science*, 24, 5-6, 798-814.
- Schmidt, R., Bienvenu, M., Fitzpatrick, P., & Amazeen, P. (1998). A comparison of intra-and interpersonal interlimb coordination: Coordination breakdowns and coupling strength. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 3, 884-900.
- Schmidt, R., Carello, C., & Turvey, M. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 2, 227-247.
- Schmidt, R., Fitzpatrick, P., Caron, R., & Mergeche, J. (2010). Understanding social motor coordination. *Human Movement Science*, 5, 834-45.
- Takenaka, T., & Ueda, K. (2003). Cognitive psychological approach to the temporal co-creation problem between self and others. *Proceedings of the 13th Japan Society of Mechanical Engineers Annual Conference*, 13, 17-18.
- Van Ulzen, N., Lamoth, C., Daffertshofer, A., Semin, G., & Beek, P. (2010). Stability and variability of acoustically specified coordination patterns while walking side-by-side on a treadmill: Does the seagull effect hold? *Neuroscience Letters*, 474, 79-83.

# How Can We Live with Overconfident or Unconfident Systems?: A Comparison of Artificial Subtle Expressions with Human-like Expression

**Takanori Komatsu (tkomat@shinshu-u.ac.jp)**

Faculty of Textile Science and Technology, Shinshu University,  
3-15-1 Tokida, Ueda 386-8567, Japan

**Kazuki Kobayashi (kby@cs.shinshu-u.ac.jp)**

Faculty of Engineering, Shinshu University,  
4-17-1 Wakasato, Nagano 380-8553, Japan

**Seiji Yamada (seiji@nii.ac.jp)**

National Institute of Informatics/ SOKEDAI/ Tokyo Institute of Technology,  
2-1-2 Hitotsubashi, Tokyo 101-8430, Japan

**Kotaro Funakoshi (funakoshi@jp.honda-ri.com)**

Honda Research Institute Japan Co., Ltd,  
8-1 Honcho, Wako 351-0188, Japan

**Mikio Nakano (nakano@jp.honda-ri.com)**

Honda Research Institute Japan Co., Ltd,  
8-1 Honcho, Wako 351-0188, Japan

## Abstract

Expressing the confidence level of a system's suggestions by using speech sounds is an important cue to users of the system for perceiving how likely it is for the suggestions to be correct. We assume that expressing the levels of confidence using human-like expressions will cause users to have a poorer impression of a system than if artificial subtle expressions (ASEs) were used when the quality of the presented information does not match the expressed level of confidence. We confirmed that this assumption was correct by conducting a psychological experiment.

**Keywords:** Artificial Subtle Expressions (ASEs), Human-like Expressions, Confidence, Users' Subjective Impressions

## Introduction

Human-machine communication using speech sounds is becoming more common (Cohen, Giangola, & Balogh, 2004; Nass & Brave, 2005) because users can obtain information while engaging in their primary tasks without facing nor manually operating the information providing systems (e.g., intelligent home appliances or car navigation systems). However, due to various reasons (for example, Benzeghibaa et al., 2007), such as noise in the sensors, the incompleteness of data, immaturity of technology, and the complexity of tasks, the reliability of such systems is often limited. Cai & Lin (2010) experimentally showed how expressing the levels of confidence for such systems to indicate whether the system's represented information is accurate or not to users plays an important role in improving both the user's performance and their impressions.

When intending to express a system's level of confidence, one can easily have the idea of using human-like verbal expressions such as "probably," "definitely," or "83% confident." However, expressing levels of confidence using such human-like expressions might frustrate users when the quality of the presented information does not match the expressed level of confidence. For example, users might feel frustrated with systems (like car navigation systems) that express a higher level of confidence like "you should follow my suggested route" or "I am 80% confident," but the represented information was wrong (this is the case of being "overconfident"). Since human-like expressions make users expect higher human-like abilities from the systems (for example, Sholtz & Bahrami, 2003; Kanda et al., 2008), such inconsistent behaviors eventually make them deeply disappointed (Aronson & Linder, 1965; Komatsu & Yamada, 2010; Komatsu, Kurosawa & Yamada, 2011).

Related to the above issue, we have proposed artificial subtle expressions (ASEs) as an intuitive methodology for notifying users of a system's internal state. Actually, the ASEs only have a complementary role in communication and should not interfere with communication's main protocol. This means that the ASEs themselves do not have any meaning without a communication context. In particular, we showed that ASEs implemented as beep-like sounds succeeded in accurately and intuitively conveying a system's confidence to the users (Funakoshi et al., 2010; Komatsu et al., 2010a; Komatsu et al., 2010b). Therefore, we assume that our proposed ASEs are suitable for expressing levels of confidence in comparison to human-like expressions.



The purpose of this study is then to confirm the above assumption that expressing levels of confidence using human-like expressions gives users a poorer impression of a system than by using those expressed by ASEs when the quality of the presented information does not match the expressed levels of confidence (in particular, where the system's suggestions are incorrect/correct even though the expressed confidence is high/low) by conducting a psychological experiment to comprehend the users' subjective impressions.

Such inconsistency between the represented information and the level of confidence is inevitable due to the immaturity of the current technology used in media terminals and due to the fact that the levels of confidence are just a probability indicating how accurate the represented information is. Therefore, this study should contribute to proposing a novel interaction technique on how to handle this inconsistency without frustrating the users.

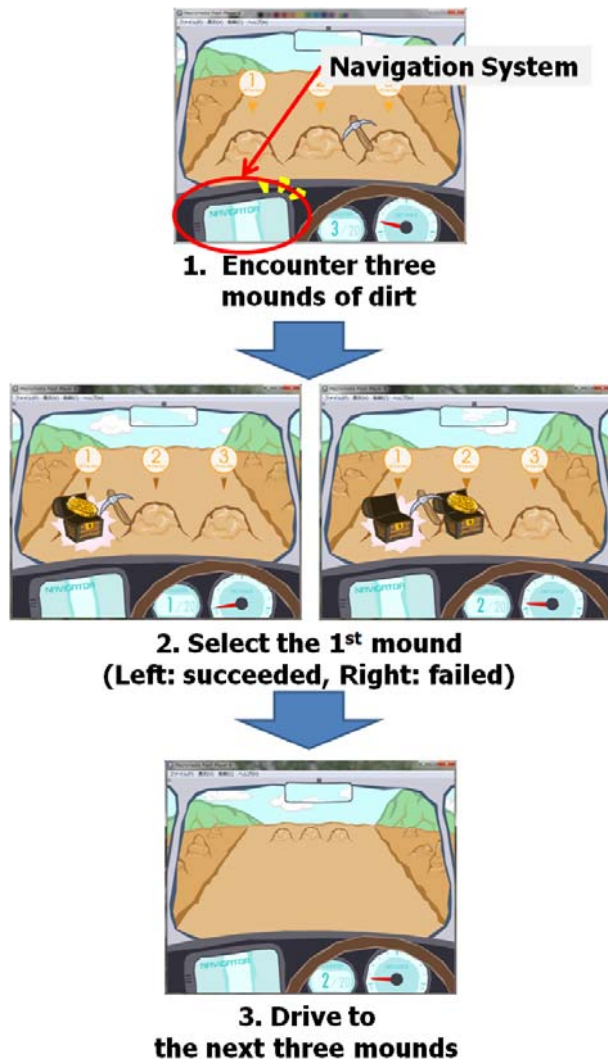


Figure 1: Driving and treasure hunting video game

## Experiment

### Environment

We used a “driving and treasure hunting” video game as our experimental environment for comprehending the participants' impressions of a system. In this game, a game image scrolls forward on a straight road as if the participant is driving a car using a navigation system with three small mounds of dirt appearing along the way. A coin is inside one of the three mounds, while the other two mounds contain nothing. The game ends after the participant encounters 20 sets of mounds (20 trials). The purpose of this game is to get as many coins as possible. The location of the coin amongst the three mounds was randomly assigned. In each trial, the navigation system next to the driver's seat (circle on top of Figure 1) told them which mound it expects the coin to be in by using speech. The participants could freely accept or reject the navigation system's suggestions. After the participant selected one mound among the three using a computer mouse, they could immediately know whether the selected mound contained the coin or not on the display (middle of Figure 1).

### Using Speech Sounds

In this experiment, the navigation system used Japanese speech sounds to suggest to the users the expected location of the coin; that is, “ichi-ban (no. 1),” “ni-ban (no. 2),” or “san-ban (no. 3).” These speech sounds were created by adding robotic-voice effects to the recorded speech sounds of one of the authors. These sounds were the main protocol (suggestion) of the navigation system. We then prepared the following three experimental stimuli (conditions) to express the levels of confidence of the main protocol.

- **ASE Condition:** One of the two ASEs was played 0.2 seconds after the speech sounds (Figure 2). These two ASEs were triangular wave sounds 0.5 seconds in duration with different pitch contours (Figure 3); that is, one was a flat ASE (onset F0: 400 Hz and end F0: 400 Hz) and the other was a decreasing ASE (onset F0: 400 Hz and end F0: 250 Hz). The suggestions with decreasing ASEs were able to inform users of the system's lower level of confidence in its suggestions while the ones with flat ASEs were to inform them of a higher level of confidence (Funakoshi et al., 2010; Komatsu et al., 2010a; Komatsu et al., 2010b).
- **Paralinguistic Condition:** As a kind of typical human-like expressions, we prepared two stimuli by modifying the paralinguistic information of the suggestions (“ichi-ban,” “ni-ban,” and “san-ban”), e.g., the rate of the utterances and intonation patterns; that is, one was an utterance with a faster rate with a falling intonation (“ichiban!”, Figure 4 (b)), while the other was a slower-rate utterance with a rising intonation (“i.chi.ba.n?”, Figure 4 (c)). We designed the latter stimulus (slower rate with rising intonation) to inform users of the system's lower level of confidence in the

form of a question, while the former stimulus (faster rate and falling intonation) informs them of a higher level of confidence.

- **Linguistic condition:** As another kind of typical human-like expressions, we prepared two stimuli by adding Japanese linguistic suffixes to the suggestions; that is, one with “desu (definitely)” 0.1 seconds after the suggestion, and the other with “dato omoi masu (I guess so).” We designed the suggestions with “dato omoi masu” to inform users of the system’s lower level of confidence, while the ones with “desu” to inform them of a higher level of confidence.

Among the 20 trials, the navigation system expressed the information with a higher level of confidence 10 times and with a lower one 10 times. The order of these two levels of confidence was counterbalanced across the participants.

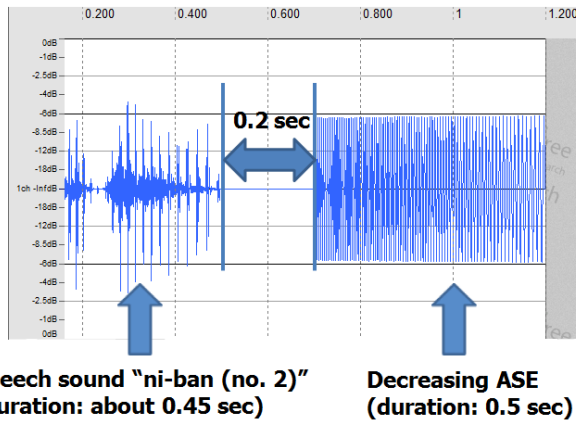


Figure 2: Speech sound “ni-ban (no.2)” and decreasing ASE

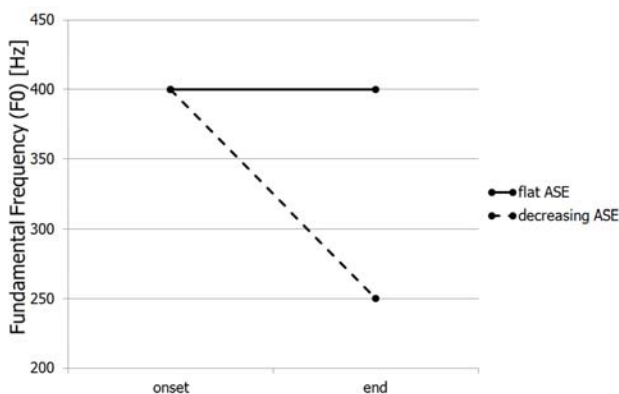


Figure 3: Flat and decreasing ASEs (duration: 0.5 second)

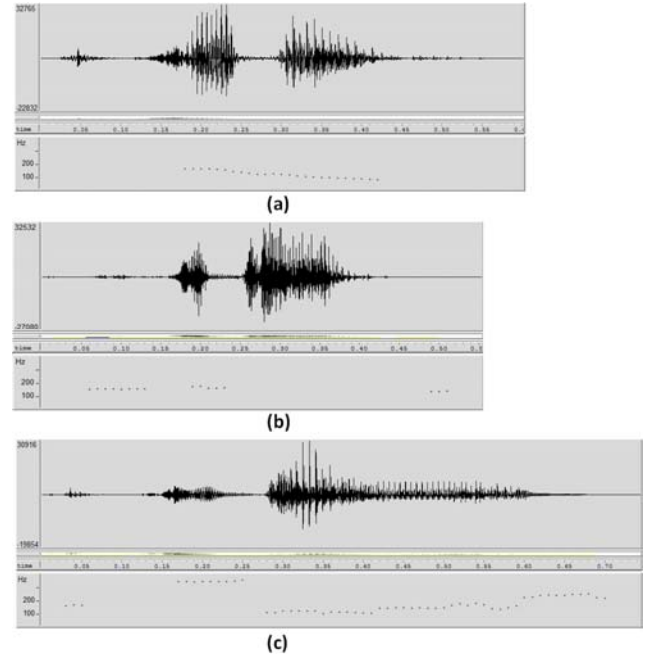


Figure 4: (a) Original wave form and pitch contour of “ichiban” used in ASEs and Linguistic conditions, (b) wave form and pitch contour of “ichiban!” and (c) wave form and pitch contour of “i..chi..ba..n?”

## Participants

Twenty Japanese university students (15 men and 5 women; 21 - 28 years old) participated. They were randomly divided into the following two experimental groups in terms of the accuracy of the navigation system’s levels of confidence.

- **Consistent Group (10 participants):** The participants in this group interacted with a system that expressed levels of confidence that were consistent with the correctness of the information it presented; that is, when the system expressed the information at a higher level of confidence, the rate of the suggested mound containing the coin was 100%, and when the system expressed the information with a lower level of confidence, the rate was 0%.
- **Inconsistent Group (10 participants):** The participants in this group interacted with a system that expressed levels of confidence that were inconsistent with the correctness of the information it presented; that is, when the system expressed the information with a higher level of confidence, the rate of the suggested mound containing the coin was 50%, and when the system expressed the information with a lower level of confidence, the rate was also 50%.

All the participants experienced all three experimental stimuli, so the experimental design was a  $2 \times 3$  mixed design; that is, the between-factor was consistent (consistent/inconsistent groups), while the within-factor was

the type of stimuli (ASEs/Paralinguistic/Linguistic conditions).

## Procedure

We used a web-based questionnaire system to comprehend the participants' impression of the navigation systems and their performances and behaviors using the treasure hunting video game in this experiment. First, the system displayed a consent form and the instructions for the experiment. Before starting the treasure hunting game, the participants were asked to listen to test sounds via a speaker or headphones and to adjust the sound volume to a comfortable level. Afterwards, they played the treasure hunting video game three times to experience all three conditions. The order of these conditions was counterbalanced among the participants.

After finishing each condition, the participants were asked to fill in a questionnaire on the navigation system, which consisted of 18 questions using a 7-point Likert scale (maximum evaluation: 7 points; minimum evaluation: 1 point). The summed points of these questions were used as the "participants' subjective impression scores" of this navigation system; that is, more points meant a better impression of the system (the highest score was 126 points and the lowest was 18). The questionnaire consisted of a modified love-liking scale (Rubin, 1970) and our original questions (Table 1: Cronbach's alpha: 0.86).

Table 1: 18 questions in the questionnaire

1. My feeling is the as usual even if I use this system.
2. This system has good adaptability.
3. This system deserves to work on responsible tasks.
4. I place complete reliance on this system.
5. This system makes favorable impressions on many people.
6. This system is always preferred among the similar systems.
7. I prefer this system because this system is similar to my way of thinking.
8. This system is human-like.
9. This system can offer good services.
10. I am satisfied with the services of this system.
11. I want to use this system again.
12. I cannot stand the mistakes made by this system.
13. This system is polite.
14. This system is a sufficiently reliable one.
15. This system is helpful for me.
16. This system is lovable.
17. I enjoy spending time with this system.
18. I feel tired when I use this system.

## Assumption

We assumed that expressing the levels of confidence using human-like expressions would give the users a poorer impression of the system than expressing these levels with ASEs when the quality of the presented information does not match the expressed levels of confidence; in particular, in the case where the systems' suggestions were incorrect/correct even though the confidence was high/low. That is, if we could observe that the participant's impression scores for the ASE condition were significantly higher than

those for the paralinguistic and linguistic conditions in the inconsistent group, we would be able to verify our assumption.

## Manipulation Check

We assumed that the types of experimental stimuli (ASEs/paralinguistic/linguistic conditions) did not affect the participants' performance and behaviors in this game but only their subjective impressions of the system. We then investigated the game scores, which indicated how many coins the participants acquired during the game (maximum: 20 coins) to clarify the relationship between the stimuli and their performance in this game (Table 2) and the acceptance rate, which indicated how many of the system's suggestions the participants accepted (maximum: 10 times for each confidence level) to clarify the relationship between the experimental stimuli and the participants' behaviors (Table 3).

The game scores were then analyzed with a  $2 \times 3$  mixed ANOVA (between independent variable: consistent/inconsistent groups, within independent variable: ASEs/paralinguistic/linguistic conditions, and dependent variable: game score). The results showed no significant difference on the main effects of the within independent variable [ $F(2,36)=0.04$ , n.s.]. The acceptance rates were analyzed with a  $2 \times 3 \times 2$  mixed ANOVA (between independent variable: consistent/inconsistent groups, #1 within independent variable: ASEs/paralinguistic/linguistic conditions, #2 within independent variable: suggestion with high/low confidence, and dependent variable: acceptance rate). The results showed no significant difference on the main effects of the #1 within independent variable [ $F(2,36)=0.04$ , n.s.].

Table 2: Game scores for each experimental condition

	Inconsistent	Consistent
ASEs	6.8	11.6
Paralinguistic	8.3	12.7
Linguistic	7	11.8

Table 3: Acceptance rate for each experimental condition according to confidence level

	Inconsistent	Consistent
ASE Low Confidence	5.4	2.4
ASE High Confidence	7.8	9.1
Paralinguistic Low Confidence	5.2	1.9
Paralinguistic High Confidence	7.8	9.4
Linguistic Low Confidence	5.1	2.5
Linguistic High Confidence	7.3	8.8

As a result of this manipulation check, we confirmed that the type of experimental stimuli did not affect the participants' performance and behaviors. Therefore, we can focus purely on the effects of the types of experimental stimuli on the participants' subjective impressions scores.

## Results

The users' subjective impression scores for each group and condition are shown in Figure 5. For the 10 participants in the consistent group, the average impression score for the ASE conditions was 58.2 (SD = 15.53), that for the paralinguistic ones was 75.1 (SD = 9.27), and that for the linguistic ones was 68.1 (SD = 11.85). For the 10 participants in the inconsistent group, the average impression score for the ASE conditions was 60.0 (SD = 11.40), that for the paralinguistic ones was 49.4 (SD = 11.94), and that for the linguistic ones was 48.6 (SD = 12.31).

These subjective impression scores were then analyzed using a  $2 \times 3$  mixed ANOVA (between independent variable: consistent/inconsistent group, within independent variable: ASEs/paralinguistic/linguistic, and dependent variable: users' subjective impression scores). The results of the ANOVA showed significant differences in the interaction effect [ $F(2,36)=11.50$ ,  $p<.01(**)$ , effect size:  $\eta^2=0.15$ ] and the main effect between independent variables [ $F(1,18)=9.99$ ,  $p<.01(**)$ ,  $\eta^2=0.22$ ]. The simple main effects of the between and within independent variables were then analyzed, and the results showed significant differences in the scores for the paralinguistic and linguistic conditions between the consistent and inconsistent groups [paralinguistic:  $F(1,18)=26.03$ ,  $p<.01(**)$ , linguistic:  $F(1,18)=11.71$ ,  $p<.01(**)$ ], and in the scores for the three experimental stimuli within both groups [consistent:  $F(2,36)=7.97$ ,  $p<.01(**)$ , inconsistent:  $F(2,36)=4.48$ ,  $p<.05(*)$ ]. A multiple comparison using an LSD test on the simple main effect of the within independent variables showed that the scores for the paralinguistic and linguistic conditions in the consistent group were significantly higher than those for the ASEs (MSe=90.4883, 5% level). In comparison, the scores for the paralinguistic and linguistic conditions in the inconsistent group were significantly lower than those for the ASEs (MSe=90.4883, 5% level).

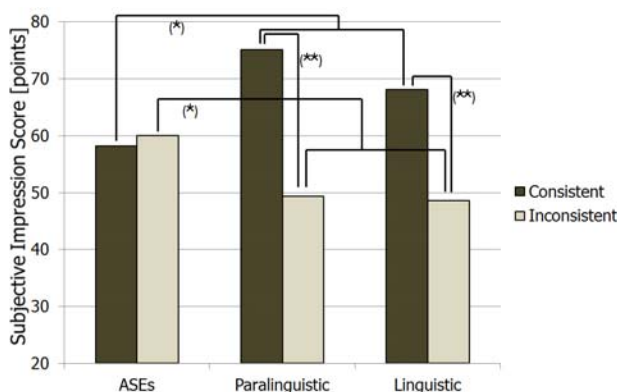


Figure 5. Subjective impression scores for each experimental condition and group

Therefore, we clearly observed that the users' impression scores for the ASE conditions were significantly higher than those for the paralinguistic or linguistic conditions in the

inconsistent group, so we were able to verify our assumption. Moreover, we found that the scores for the ASE conditions in both groups were almost the same, while the scores for the paralinguistic and linguistic conditions significantly differed. Therefore, this also implies that the users' subjective impressions for a system expressing ASEs were quite robust regardless of the consistency between the represented information and the levels of confidence.

## Discussion

The results of this study showed that expressing confidence using human-like expressions received a higher evaluation in comparison to using such expressions with ASEs when the accuracy of the system's levels of confidence was perfect (in consistent group). However, it is almost impossible to build a user interface that can always provide correct suggestions and levels of confidence to users based on the level of current technology. Therefore, establishing a concrete methodology for handling the inconsistency between the suggestion and the levels of confidence is indispensable and worthwhile for the HCI and cognitive science domains.

We now have to investigate whether the acquired results can be used in much more realistic applications, e.g., spoken dialogue systems like an actual car navigation system. If we succeed, we can strongly argue that expressing the levels of confidence of systems by using ASEs is a reasonable methodology for avoiding frustrating users, and will contribute to the proposals for a novel interaction technique for dealing with the inconsistency between the represented information and the levels of confidence.

In this experiment, we could not clarify which kinds of users' cognitive basis significantly affected the result of this experiment. We assume that several researches such as the gain and loss of esteem (Aronson & Linder, 1965), adaptation gap hypothesis (Komatsu & Yamada, 2010; Komatsu, Kurosawa & Yamada, 2011), uncanny valley (Mori, 1970) and human's cognitive nature for anthropomorphism (Reeves & Nass, 1996) are keys to find out the cognitive basis that could clearly explain the participants' behaviors observed in this experiment. To tackle with this issue, we are now planning to conduct a consecutive experiment to investigate the further abilities of ASEs; that is, whether the ASEs could inform users of not only the higher or lower level of confidence but the other kinds of meanings? or more detailed level of confidence? We believe that this consecutive study would clarify which users' cognitive basis affect their behaviors how they intuitively interpret the ASEs.

## Conclusions

Expressing the confidence level of a system's suggestions by using speech sounds is an important cue to users of the system for perceiving how likely it is for the suggestions to be correct. We assume that expressing the levels of



confidence using human-like expressions will cause users to have a poorer impression of a system than if the ASEs were used when the quality of the presented information does not match the expressed level of confidence. We confirmed that this assumption was correct by conducting a psychological experiment. In particular, the users' impression scores for the ASE conditions were significantly higher than those for the paralinguistic or linguistic conditions in the inconsistent group.

### Acknowledgments

This work was supported by KAKENHI (23700248) as Grant-in-Aid for Young Scientists (B) by The Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, and by joint research with Honda Research Institute Japan and with National Institute of Informatics, Japan.

### References

- Aronson, E., & Linder, D. (1965). Gain and loss of esteem as determinants of interpersonal attractiveness, *Journal of Experimental Social Psychology*, 1 (2), 156-171.
- Benzeghibaa, M., De Moria, R., Derooa, O., Dupont S., Erbesa, T., Jouveta, D., Fissorea, F., Lafacea, P., Mertinsa, A., Risa, C., Rosea, R., Tyagia, V., & Wellekensa, C. (2007). Automatic speech recognition and speech variability: A review, *Speech Communication*, 49, 763-786.
- Cai, H., & Lin, Y. (2010). Tuning Trust Using Cognitive Cues for Better Human-Machine Collaboration, *Proceedings of Human Factors and Ergonomics Society 2010* (pp. (5) 2437-2441).
- Cohen, M. H., Giangola, J. P., & Balogh, J. (2004). *Voice User Interface Design*, MA: Addison-Wesley.
- Funakoshi, K., Kobayashi, K., Nakano, M., Yamada, S., & Komatsu, T. (2010). Non- humanlike Spoken Dialogue: a Design Perspective, *Proceedings of 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 176-184).
- Kanda, T., Miyashita, T., Osada, T., Haikawa, Y., & Ishiguro H. (2008). Analysis of Humanoid Appearances in Human-robot Interaction, *IEEE Transactions on Robotics*, 24 (3), 725-735.
- Komatsu, T., & Yamada, S. (2010). Adaptation Gap Hypothesis: How differences between users' expected and perceived agent functions affect their subjective impression, *Journal of Systemics, Cybernetics and Informatics*, 9(1), 67-74.
- Komatsu, T., Yamada, S., Kobayashi, K., Funakoshi, K., & Nakano, M. (2010a). Artificial Subtle Expressions: Intuitive Notification Methodology of Artifacts, *Proceedings of the 28th international conference on Human factors in computing systems* (pp. 1941-1944).
- Komatsu, T., Yamada, S., Kobayashi, K., Funakoshi, K., & Nakano, M. (2010b). Artificial Subtle Expressions: Intuitive Notification Methodology of Artifacts, *Proceedings of the 32nd Annual Meeting of Cognitive Science Society* (pp. 447-452).
- Komatsu, T., Kurosawa, R., & Yamada, S. (2011). How Does the Difference Between Users' Expectations and Perceptions About a Robotic Agent Affect Their Behavior?, *International Journal of Social Robotics*, DOI=10.1007/s12369-011-0122-y.
- Mori, M. (1970). Bukimi no tani: The uncanny valley (K. F. MacDorman & T. Minato, Trans.). *Energy*, 7(4), 33-35. (Originally in Japanese).
- Nass, C., & Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*, MA: The MIT Press.
- Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, MA: The MIT Press.
- Rubin, Z. (1970). Measurement of romantic love, *Journal of Personality and Social Psychology*, 16(2), 265-273.
- Sholtz, J., & Bahrami, S. (2003). Human-Robot interaction: development of an evaluation methodology for the bystander role of interaction, *Proceedings of the 2003 IEEE System, Man and Cybernetics* (pp. 3212 - 3217).

# Effect of Social Skills on the Asymmetry in Facial Expressions

**Masashi Komori (komori@oecu.jp)**

Fac. Inf. Commun. Eng., Osaka Electro-Communication University  
18-8 Hatsucho, Neyagawa, Osaka, 572-8530, Japan

**Hiroko Kamide (kamide@arai-lab.sys.es.osaka-u.ac.jp)**

Faculty of Engineering Science, Osaka University  
1-2 Machikaneyama-cho, Toyonaka, Osaka, 560-8531 Japan

**Satoru Kawamura (s-kawamura@aist.go.jp)**

National Institute of Advanced Industrial Science and Technology Kansai  
1-8-31 Midorigaoka, Ikeda, Osaka, 563-8577, Japan

**Chika Nagaoka (nagaoka@educ.kyoto-u.ac.jp)**

Kokoro research center, Kyoto University  
46 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto, 606-8501 Japan

## Abstract

This study investigated the effect of social skills on the facial movement asymmetry in facial expressions. Three-dimensional facial landmark data of facial expressions (neutral, happy, and angry) were obtained from Japanese participants ( $n = 62$ ). After the facial expression task, each participant completed KiSS-18 (Kikuchi's Scale of Social Skills; Kikuchi, 2007). Through a generalized Procrustes method, facial landmark coordinates and their mirror-reversed versions were represented as points on a hyperplane. The asymmetry of each face was defined as Euclidian distance on the plane. Subtraction of the asymmetry level of a neutral face of each individual from the asymmetry level of a target emotion face was defined as the index of "movement asymmetry" of each emotion. Correlation coefficients of KiSS-18 scores and movement asymmetry scores were computed for both happy and angry expressions. Significant negative correlations between KiSS-18 scores and movement asymmetries were found for both expressions. The results indicate that symmetric facial expressions are higher with higher level of social skills.

**Keywords:** facial expression; facial asymmetry; social skills; landmark-based 3D shape analysis.

## Introduction

Facial expressions provide various signals for social interactions. Although human faces and facial expressions are somewhat symmetrical, numerous studies have focused on facial bilateral symmetry: the degree to which one half of a face is similar to the other half. Facial asymmetry derives from two sources: structural asymmetry and movement asymmetry (Schmidt, Liu & Cohn, 2006). Structural asymmetry derives from physical variation in laterality of facial structure, while movement asymmetry derives from lateralized facial muscle movement during facial expressions. In this study, we focused on movement asymmetry in creating emotional expressions.

The primary source of facial movement asymmetry is brain lateralization in emotion processing. Several studies have shown that emotions are expressed more intensely in the left hemiface (Sackeim, Gur & Saucy, 1978; Borod, Haywood & Koff, 1997), because most facial muscles, particularly those in the lower part, are innervated by the contralateral hemisphere (Borod, 1993). Thus the dominance of the left hemiface in facial expressions has been interpreted as supporting the hypothesis of brain lateralization of emotional processing (Schwartz, Davidson & Maer, 1975; Sackeim, Greenberg, Weiman, Gur, Hungerbuhler & Geschwind, 1982). More recent studies have asserted that both hemispheres process emotion, but each hemisphere is specialized for particular types of emotion (Fusar-Poli, Placentino, Carletti, Allen, Landi, Abbamonte, Barale, Perez, McGuire & Politi, 2009) such as positive-negative emotion (Davidson, 1992; Gur, Skolnick & Gur, 1994), or approach-withdrawal (Davidson, 1999).

While the laterality in facial expressions, human face perception mostly relies on facial information contained in the right hemiface (Gilbert & Bakan, 1973; Grega, Sackeim, Sanchez, Cohen & Hough, 1988, Kanwisher, McDermott & Chun, 1997; Sergent, Ohta & MacDonald, 1992). When asked to judge the facial expression of a briefly presented chimeric face image, perceivers tend to base their decision more frequently on the expression contained within the right side of the face, i.e., the left hemiface for the viewer. Thus, the lateralization in facial expressions can lead to failure in conveying the face's real emotions to an observer. Although the role of asymmetry of facial expressions in social interactions is still unclear, asymmetric facial expression is possibly an important variable. Facial asymmetry has been proposed as a signal of developmental stability that can indicate mate quality (Grammer & Thornhill, 1994; Kowner, 1996; Penton-Voak, Jones, Little, Baker, Tiddeman, Burt & Perrett, 2001). In general, the less asymmetric a face is, the more attractive it appears (Grammer, Fink, Moller &

Thornhill, 2003; Grammer & Thornhill, 1994). This asymmetry is believed to reflect past developmental stresses and to be related to the likely quality of the individual as a potential mating partner (Fink, 2004). Such preference for symmetry may also extend to the preference for movement symmetry.

This study investigates the relationship between social skills and facial movement asymmetry in emotional expressions. Social skills are generally defined as the set of skills that enable a person to interact and communicate with others in verbal and nonverbal forms of communication. If higher social skills are related to more symmetrical facial movements in emotional expressions, symmetrical facial expressions can be considered an appropriate approach to present the face owner's emotions to receivers in social interactions.

## Method

### Facial Expression Task

Japanese undergraduate and graduate students ( $n = 62$ : 20 men and 42 women; age: 19 to 26 years, mean age = 21.3,  $SD = 1.37$ ) provided three-dimensional facial shape data of neutral, happy, and angry expressions. First, the participants were instructed to show and maintain a neutral facial expression. Then, they were asked to recall their experiences in which they had felt the target emotions (happy or angry) and to describe the experience after taking each 3D image. The participants were not instructed to pose or maintain any expression of target emotion. The order of target emotions was counterbalanced among the participants. Three-dimensional (3D) shapes and textures of facial expressions of each face were captured using a 3D picture measurement device (TRiDY-S: JFE Techno-Research Corp.) based on pattern-projection method.

### Assessment of Participants' Social Skill

After the facial expression task, each participant completed KiSS-18 (Kikuchi's Scale of Social Skills; Kikuchi, 2007), an 18-item self-report measurement of social skills with higher scores indicating high level of social skills. This scale is based on six categories of social skills proposed by Goldstein (1980); basic skills, advanced skills, emotional management skills, stress management skills, offence management skills and planning skills. Basic skills include 'talking with others', 'maintaining a conversation' and 'introducing oneself'. Advanced skills include 'asking for help', 'giving instructions', 'obeying instructions', 'apologizing' and 'persuading'. Emotional management skills include 'managing fear', 'emotional expression' and 'managing others' anger'. Stress management skills include 'managing criticism' and 'managing a contradiction in message'. Offence management skills include 'helping others', 'conflict resolution' and 'managing trouble'. Planning skills include 'staying on target' and 'taking initiative'. The scale has demonstrated high reliability and validity in previous studies (Kikuchi 2007).

## Facial Shape Measurement

Thirty-six facial landmarks were selected on the basis of our previous studies (Kamide, Komori, Kawamura & Nagaoka, 2011; Figure1, Table 1). All 3D coordinates of the landmarks were visually measured using a computer program (Rapid Form 2004: INUS Technology) by referring to each of the 3D shape data and texture.

Table 1: Set of 36 facial landmarks.

No.	Location
1	hairline
2	forehead
3	forehead
4	outer corner of eyebrow
5	upper point of maximum width of eyebrow
6	lower point of maximum width of eyebrow
7	inner corner of eyebrow
8	inner corner of eyebrow
9	upper point of maximum width of eyebrow
10	lower point of maximum width of eyebrow
11	outer corner of eyebrow
12	root of nose
13	side of root of nose
14	side of root of nose
15	lateral angle of eye
16	center of upper eyelid
17	center of lower eyelid
18	medial angle of eye
19	medial angle of eye
20	center of upper eyelid
21	center of lower eyelid
22	lateral angle of eye
23	zygomatic
24	zygomatic
25	apex of nose
26	ala of nose
27	ala of nose
28	subnasal point
29	angle of mouth
30	upper lip (philtrum edge)
31	central upper lip
32	upper lip (philtrum edge)
33	angle of mouth
34	stomion
35	bottom of lower lip
36	chin





Figure. 1: Landmark locations. Photographs were formed by warping average facial texture.

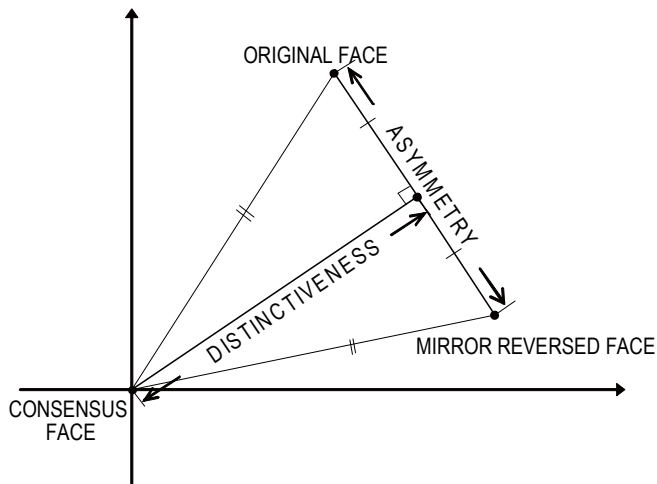


Figure 2: Schematic illustration of asymmetry. Each facial shape and its mirror-reversed version are represented as points on the tangent hyperplane. Only two axes are represented for ease of illustration. The “consensus face” is the origin of the space. Asymmetry is defined as the Euclidean distance from each original version to the mirror-reversed version.

### Facial Shape Standardization

Each face differed in location, size, and orientation. To standardize them, we performed a generalized Procrustes analysis (GPA) on the facial landmarks of all faces irrespective of the gender of the face. A GPA is an analytical method used for multivariate statistical analysis of landmark locations expressed in Cartesian coordinates. This method preserves information about the relative spatial relationships of landmarks throughout the standardization, and that has recently been applied to psychological research on human faces (Komori, Kawamura & Ishihara, 2009, 2011).

For the standardization of location and size, we used the centroid size technique (Bookstein, 1991). All facial shapes

were translated into the same origin (centroid) and scaled to the unit centroid size, which is the sum of the squared distances from the centroid to each landmark. For alignment of orientation, rotations around the centroid of the faces were performed (Dryden & Mardia, 1998) such that the sum of the squared distances among corresponding landmarks between samples was minimized. Using the GPA, each facial shape was represented as a point on a linear tangent hyperplane of 108 dimensions ( $36 \times 3$ ), which allowed us to treat the faces as multidimensional, normally distributed values.

The “Shapes” statistical package written by Dryden and Mardia (1998), which runs in an R statistical analysis environment, was employed for the analyses. In addition to the coordinates of 62 facial shapes, the mirror-reversed versions of the same faces were used in the facial shape analysis.

### Calculation of Facial Asymmetry

Through a generalized Procrustes method, each of the facial shapes and their mirror-reversed versions were represented as a point on the tangent hyperplane. We defined asymmetry (the converse of symmetry) of each facial shape as the Euclidean distance between the face and its mirror-reversed face on the hyperplane (Figure 2), according to our previous study (Komori, Kawamura & Ishihara, 2009). Furthermore, all original faces and their mirror-reversed faces were combined to create a consensus face. This was the average of all facial shapes and represented the origin of the tangent hyperplane. The distance from a given original face to the origin was the same as that from its mirror image to the origin. This distance can be regarded as an index of facial distinctiveness (the converse of facial averageness). Therefore, asymmetry and distinctiveness can be measured independently through this procedure; in other words, facial variations can be separated into distinctiveness and asymmetry.

### Calculation of Local Asymmetry

To investigate the degree of asymmetry in each facial part, such as eyebrows, eyes, and mouth, facial subspaces were constructed from the standardized landmark coordinates of eyebrows (from No. 4 to No. 12 of Table 1), eyes (from No. 15 to No. 22), and mouth (from No. 29 to No. 34). Asymmetry in each part of a face was defined as the Euclidean distance from the original version to the mirror-reversed version of each part in each subspace.

## Results

### Social Skill Score

Some studies have reported that males and females differ in facial shape (Little, Jones, Waite, Tiddeman, Feinberg, Perrett, Apicella & Marlowe, 2008) and facial muscle reactivity (Dimberg & Lundquist, 1990). It is possible that gender differences in social skills could be a potential confound in the analysis of the relationships between social

skills and facial expressions. However, the Kiss-18 scores were not significantly different between males and females ( $t(60) = -.57, p = .57$ ).

### Facial Asymmetry

The mean morphological asymmetry in each facial expression is shown in Figure 3. To examine the relationship between facial expressions (neutral, happy, and angry) and facial asymmetry levels, we conducted repeated measures ANOVA with facial asymmetry levels as the dependent variable. There was no significant effect of facial expression on the facial asymmetry levels ( $F(2,122) = .36, p = .72$ ).Local Asymmetries

The local asymmetries were calculated for eyebrows, eyes, and mouth (Figure 4). A repeated measures ANOVA was performed for each facial part, and a significant effect of

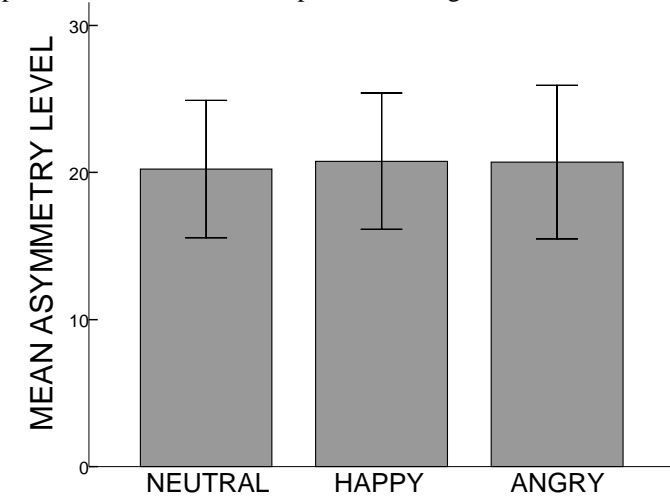


Figure 3: Mean facial asymmetry level. Error bars represent 1 S.D.

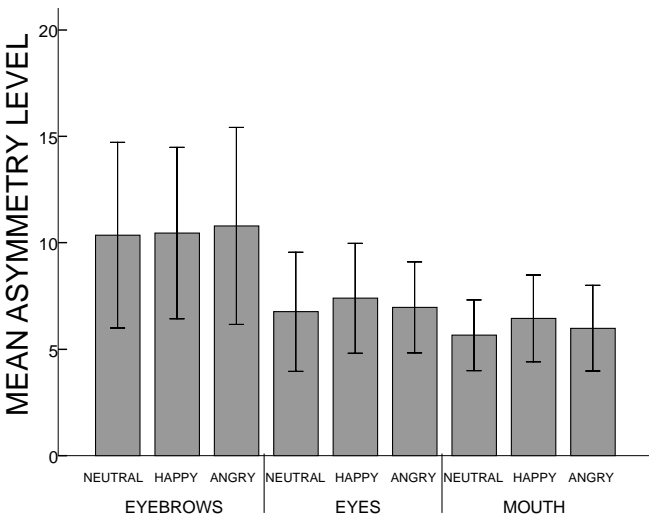


Figure 4: Mean facial asymmetry levels of facial parts. Error bars represent 1 S.D.

facial expressions on the asymmetry in mouth shape was found ( $F(2,122) = 4.10, p = .019$ ). However, there was no effect of facial expressions on the asymmetries in eyebrows or eyes (eyebrows:  $F(2,122) = .06, p = .94$ ; eyes:  $F(2,122) = .16, p = .21$ ).

### Relationship between Social Skills and Facial Asymmetry

Since there was no significant correlation between the Kiss-18 scores and facial asymmetry levels of neutral expression, facial structural asymmetry is considered potentially unrelated to social skills ( $r = .20, p = .13$ ). Thus, the subtraction of the asymmetry level of a neutral face of each individual from the asymmetry level of a target emotional face can be defined as the index of movement asymmetry that derives from facial muscle movement. Here we refer to the value as the “movement asymmetry score.”

To assess whether higher social skills are linked to facial movement asymmetry, correlation coefficients of the Kiss-18 scores and movement asymmetry scores were computed for both happy and angry expressions. Figure 5 shows the relationship between social skills and movement asymmetries. There was a significant negative correlation between the Kiss-18 scores and movement asymmetries for both expressions (happiness:  $r = -.30, p = .017$ ; angry:  $r = -.30, p = .018$ ), indicating that the higher a participant scored on the social skills test, the more symmetric their facial expressions were.

The partial correlation coefficients between social skills and movement asymmetries, using gender of the participants as control variables, were also significant for both expressions (happiness:  $r = -.30, p = .015$ ; angry:  $r = -.30, p = .017$ ). This suggests that the relationship between facial movement asymmetry and social skills was not caused by the gender differences in social skills.

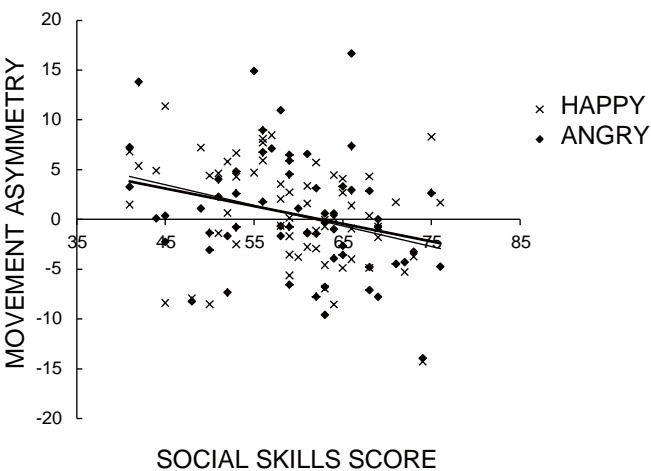


Figure 5: Relationship between social skill score and degree of facial asymmetry.

The correlation coefficient of the Kiss-18 score and movement asymmetry in each facial part was calculated for each target emotion. For a happy expression, movement asymmetry of none of the parts was significantly correlated with social skills (eyebrows:  $r = -.16$ ,  $p = .21$ ; eyes:  $r = -.13$ ,  $p = .33$ ; mouth:  $r = -.02$ ,  $p = .89$ ). On the other hand, for an angry expression, only movement asymmetry of mouth was found to be negatively correlated with social skills (eyebrows:  $r = .03$ ,  $p = .82$ ; eyes:  $r = -.17$ ,  $p = .19$ ; mouth:  $r = -.26$ ,  $p = .04$ ).

## Discussion

This study provides evidence that facial movement asymmetry in emotional expressions is linked to low social skills. Some studies have shown that a spontaneous smile is symmetrical, but a posed (voluntary) smile is asymmetrical (Gazzaniga & Smylie, 1990; Frank, Ekman & Friesen, 1993), suggesting a possibility that symmetrical facial expression is recognized as spontaneous facial expression derived from the facial owner's emotion. In fact, Ozono, Watabe, Yoshikawa, Nakashima, Rule, Ambady & Adams (2010) have reported that Japanese participants rated faces with greater smile symmetry as more trustworthy. Thus, the results of this study may reflect the connection between symmetrical facial expressions and trustworthiness.

The results also show that the relationship between low social skills and facial movement asymmetry is especially observed in the lower face region. Neurologically, movements of the lower face region follows voluntary muscle control, while upper face area movements follow automatic control (Rinn, 1994; Gazzaniga & Smylie, 1990). It is possible that the different effect of social skills on the asymmetry in the upper or lower facial areas is caused by such differential motor control.

The results of the study also suggest that the asymmetry quantification method of this study is an effective method for evaluating 3D facial asymmetry.

## Acknowledgments

A part of this study was supported by Grant-in-Aid for Scientific Research on Innovative Areas, "Face Perception and Recognition", from The Ministry of Education, Culture, Sports, Science and Technology (MEXT) (No. 23119724) and a Grant-in-Aid for Scientific Research (No. 16330126) from the Japan Society for the Promotion of Science.

## References

- Bookstein, F. L. (1991). *Morphometric tools for landmark data*. Cambridge: Cambridge Univ. Press.
- Borod, J. C. (1993). Cerebral mechanisms underlying facial, prosodic, and lexical emotional expression: A review of neuropsychological studies and methodological issues. *Neuropsychology*, 7(4), 445–463.
- Borod, J. C., Haywood, C. S., & Koff, E. (1997). Neuropsychological aspects of facial asymmetry during emotional expression: a review of the normal adult literature. *Neuropsychology Review*, 7(1), 41–60.
- Davidson, R. J. (1992). Anterior cerebral asymmetry and the nature of emotion. *Brain and Cognition*, 20(1), 125–151.
- Davidson, R. J., & Iwan, W., (1999). The functional neuroanatomy of emotion and affective style. *Trends in Cognitive Sciences*, 3(1), 11–21.
- Dryden, I. L., & Mardia, K. V. (1998). *Statistical shape analysis*. Chichester: Wiley & Sons.
- Fink, B. (2004). Second to fourth digit ratio and facial asymmetry. *Evolution and Human Behavior*, 25(2), 125–132.
- Frank, M. G., Ekman, P., & Friesen, W. V. (1993). Behavioral markers and recognizability of the smile of enjoyment. *Journal of Personality and Social Psychology*, 64(1), 83–93.
- Fusar-Poli, P., Placentino, A., Carletti, F., Allen, P., Landi, P., Abbamonte, M., Barale, F., Perez, J., McGuire, P. & Politi P. L. (2009). Laterality effect on emotional faces processing: ALE meta-analysis of evidence. *Neuroscience Letters*, 452(3), 262–267.
- Gazzaniga, M. S., & Smylie, C. S. (1990). Hemispheric Mechanisms Controlling Voluntary and Spontaneous Facial Expressions. *Journal of Cognitive Neuroscience*, 2(3), 239–245.
- Gilbert, C., & Bakan, P., (1973) Visual asymmetry in perception of faces. *Neuropsychologia*, 11(3), 355–362.
- Goldstein, A. P. (1980) *Skill streaming the adolescent: a structured leaning approach to teaching prosocial skills*. Research Press.
- Grammer, K., Fink, B., Møller, A. P., & Thornhill, R. (2003). Darwinian aesthetics: sexual selection and the biology of beauty. *Biological Reviews of the Cambridge Philosophical Society*, 78(3), 385–407.
- Grammer, K., & Thornhill, R. (1994). Human (*Homo sapiens*) facial attractiveness and sexual selection: The role of symmetry and averageness. *Journal of Comparative Psychology*, 108, 233–242.
- Grega, D. M., Sackeim, H. A., Sanchez, E., Cohen, B. H., & Hough, S. (1988). Perceiver bias in the processing of human faces: neuropsychological mechanisms. *Cortex*, 24(1), 91–117.
- Gur, R. C., Skolnick, B. E., & Gur, R. E. (1994). Effects of emotional discrimination tasks on cerebral blood flow: regional activation and its relation to performance. *Brain and Cognition*, 25(2), 271–286.
- Kamide, H., Komori, M., Kawamura, S., & Nagaoka, C. (2010) The Relationship between Social Skills and Morphology of Facial Expressions. *IEICE Technical Report*, 111(214), 7–12. (in Japanese)
- Kanwisher, N., McDermott, J., & Chun, M. M. (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception, *Journal of Neuroscience*. 17, 4302–4311.
- Kikuchi, K. (1988) Research on Science of Compassion: Psychology and Skill Orienting to Social Behavior, Kawashima-Shoten.

- Komori, M., Kawamura, S., & Ishihara, S. (2009). Averageness or symmetry: which is more important for facial attractiveness? *Acta Psychologica*, 131(2), 136–142.
- Komori, M., Kawamura, S., & Ishihara, S. (2011). Multiple mechanisms in the perception of face gender: Effect of sex-irrelevant features. *Journal of Experimental Psychology: Human Perception and Performance*, 37(3), 626–633.
- Kowner, R. (1996). Facial asymmetry and attractiveness judgment in developmental perspective. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 662–675.
- Little, A. C., Jones, B. C., Waitt, C., Tiddeman, B. P., Feinberg, D. R., Perrett, D. I., Apicella, C. A. & Marlowe, F. W. (2008). Symmetry is related to sexual dimorphism in faces: Data across culture and species. *PLoS one*, 3(5), 2106.
- Ozono, H., Watabe, M., Yoshikawa, S., Nakashima, S., Rule, N. O., Ambady, N., & Adams, R. B. (2010). What's in a Smile? Cultural Differences in the Effects of Smiling on Judgments of Trustworthiness. *Letters on Evolutionary Behavioral Science*, 1(1), 15–18.
- Penton-Voak, I. S., Jones, B. C., Little, A. C., Baker, S., Tiddeman, B., Burt, D. M., & Perrett, D. I. (2001). Symmetry, sexual dimorphism in facial proportions and male facial attractiveness. *Proceedings of the Royal Society B: Biological Sciences*, 268(1476), 1617–1623.
- Rinn, W. E. (1984). The neuropsychology of facial expression: a review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin*, 95(1), 52–77.
- Sackeim, H. A., Greenberg, M. S., Weiman, A. L., Gur, R. C., Hungerbuhler, J. P., & Geschwind, N. (1982). Hemispheric asymmetry in the expression of positive and negative emotions. *Archives of Neurology*, 39, 210–218.
- Sackeim, H. A., Gur, R. C., & Saucy, M. C. (1978). Emotions are expressed more intensely on the left side of the face. *Science*, 202(4366), 434–436.
- Schmidt, K. L., Liu, Y., & Cohn, J. F. (2006). The role of structural facial asymmetry in asymmetry of peak facial expressions. *Laterality*, 11(6), 540–561.
- Schwartz, G. E., Davidson, R. J., & Maer, F. (1975). Right hemisphere lateralization for emotion in the human brain: interactions with cognition. *Science*, 190(4211), 286–288.
- Sergent, J., Ohta, S., & MacDonald, B. (1992) Functional neuroanatomy of face and object processing. A positron emission tomography study, *Brain*, 115,15–36.

# Reasoning on the Raven's Advanced Progressive Matrices Test with Iconic Visual Representations

Maithilee Kunda\*, Keith McGregor\*, and Ashok Goel

Design & Intelligence Laboratory, School of Interactive Computing, Georgia Institute of Technology  
85 Fifth Street NW, Atlanta, GA 30332 USA

{mkunda,keith.mcgreggor}@gatech.edu, goel@cc.gatech.edu

*\*these authors contributed equally to this work*

## Abstract

Although the problems on Raven's Progressive Matrices intelligence tests resemble geometric analogies, studies of human behavior suggest the existence of two qualitatively distinct types of strategies: verbal strategies that use propositional representations and visual strategies that use iconic representations. However, all prior computational models implemented to solve these tests have modeled only verbal strategies: they translate problems into purely propositional representations. We examine here the other half of what may be a dual-process mechanism of reasoning in humans: visual strategies that use iconic representations. In particular, we present two different algorithms that use iconic visual representations to address problems found on the Advanced Progressive Matrices test, the best of which yields performances at levels equivalent to the 75th percentile for human test takers aged from 20 to 62 years-old. We discuss implications of our work for understanding the computational nature of Raven's and visual analogy in problem solving.

**Keywords:** Analogy; intelligence tests; knowledge representations; mental imagery; Raven's Progressive Matrices; visual reasoning.

## Introduction

The Raven's Progressive Matrices (RPM) test is a standardized intelligence test. The test consists of geometric analogy problems in which a matrix of geometric figures is presented with one entry missing, and the correct missing entry must be selected from a set of answer choices. Figure 1 shows an example of a matrix problem of this kind.

There are currently three published versions of the RPM: the original Standard Progressive Matrices (SPM), the Advanced Progressive Matrices (APM), developed as a more difficult test than the SPM for individuals in high IQ ranges, and the Colored Progressive Matrices (CPM), intended as a simpler test than the SPM to be used with children, the elderly, or other individuals falling into lower IQ ranges (Raven et al., 2003). The RPM tests are considered to be the single best psychometric measures of general intelligence, outside of multi-domain IQ tests like the Wechsler scales (Snow et al., 1984), and all three versions of the RPM are widely used in clinical, educational, occupational, and scientific settings.

Neuroimaging and behavioral studies suggest that humans recruit qualitatively different strategies on the RPM regarding what types of mental representations are used, specifically in terms of visual versus verbal strategies. Visual strategies use iconic mental representations rooted in the visual perceptual modality, such as mental imagery.

Verbal strategies use amodal propositional mental representations, such as linguistic description.

From factor analyses of both the SPM (Lynn et al., 2004; van der Ven & Ellis, 2000) and the APM (Dillon et al., 1981; Mackintosh & Bennett, 2005; Vigneau & Bors, 2005) as well as from fMRI data (Prabhakaran et al., 1997) comes evidence for various categories of RPM problems differentially eliciting from people either visual or verbal strategies. Studies of patients with focal brain lesions have also found linkages between brain regions associated with visual or verbal processing and successful performance on certain RPM problems (Berker & Smith, 1988; Villardita, 1985). Individuals with autism, who may exhibit a general bias towards using visual strategies over verbal ones (Kunda & Goel, 2007, 2011), tend to do particularly well on the RPM (Bölte et al., 2009; Dawson et al., 2007) and have been observed with fMRI to prefer predominantly visual strategies on the RPM (Soulières et al., 2009).

Despite this breadth of evidence for the existence of both visual and verbal RPM strategies, most computational RPM accounts have presumed to translate visual inputs into propositional representations, over which various kinds of reasoning then take place. One reason for this may be the general preponderance of propositional representations in computational accounts of cognition; in many models of visual reasoning across various task domains, visual knowledge too is represented using propositions (Carpenter et al. 1990, Lovett et al. 2010, Davies et al. 2008).

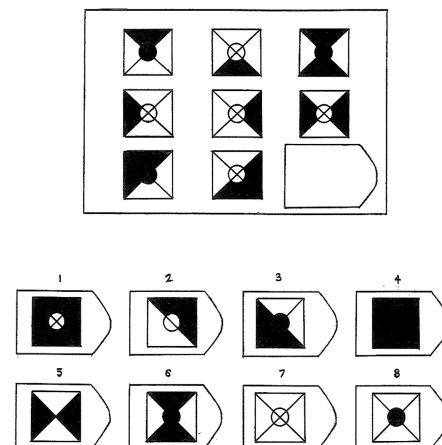


Figure 1: Example RPM Problem.

Another reason may stem from the practice of using verbal reporting protocols to study RPM problem solving. By their very nature, verbal reports are better suited to describing verbal strategies than visual strategies, which may introduce bias into the results of such protocols. Of even greater significance are findings across multiple task domains that the act of verbal reporting actually biases individuals towards using verbal strategies and/or impairs their use of visual strategies, a phenomenon known as “verbal overshadowing” (Schooler & Engstler-Schooler, 1990; Schooler et al., 1993). DeShon, Chan, and Weissbein (1995) found that a verbal reporting protocol on the APM significantly impaired accuracy on about half of the problems, and specifically on those typically solved using visual strategies.

The goal of our work is to develop computational models of a dual cognitive strategy that uses both verbal and visual representations. This first requires the development of computational models of the visual strategy itself. Once such computational models have been developed, they then may potentially be coupled with existing models of the verbal strategy. We have developed two such computational models of reasoning on the RPM using iconic visual representations. In earlier work, we tested these models against the SPM (Kunda, McGreggor, & Goel, 2010). In this paper, we apply these computational models to the APM.

In so far as we know, this work represents several firsts: it is the first report of any computational model addressing the entirety of the APM test, the first in which the problems are attempted using purely iconic visual representations, and the first to tackle the test using scanned images of each test page, without any re-rendering or representational change of inputs from those presented to a human test-taker.

## Computational Accounts of the RPM

Hunt (1974) proposed the existence of two different RPM strategies that varied primarily in how problem inputs were represented. The “Analytic” algorithm used propositions to represent problems as lists of features and logical operations to evaluate rules such as constancy and addition/subtraction. The “Gestalt” algorithm, akin to mental imagery, used iconic representations and perceptual operations like continuation and superposition. However, neither algorithm was actually implemented. All of the computational RPM models that have since been developed resemble Hunt’s Analytic algorithm in that they rely on a conversion of problem inputs into amodal propositional representations.

**Model 1** Carpenter, Just, and Shell (1990) used a production system that took hand-coded symbolic descriptions of certain problems from the Advanced Progressive Matrices (APM) test and then selected from a set of predefined rules to solve each problem. The rules were generated by the authors from a priori inspection of the APM. The rules were experimentally validated using a verbal reporting protocol, but the potential confound of a verbal overshadowing effect was not addressed. Differences between low- and high-scoring participants were modeled by developing two different versions (FairRaven and BetterRaven) of the production system; the more advanced system (BetterRaven) contained an increased vocabulary of

rules and a goal monitor. Both systems were tested against 34 of the 48 problems from the APM and solved 23 and 32 problems, respectively.

**Model 2** Bringsjord and Schimanski (2003) used a theorem-prover to solve selected RPM problems stated in first-order logic, though no specific results were reported.

**Model 3** Lovett, Forbus, and Usher (2010) combined automated sketch understanding with the structure-mapping analogy technique to solve SPM problems. Their system took as input problem entries sketched in Powerpoint as segmented shape objects and then automatically translated these shapes into propositional descriptions, using a sketch understanding system based on work by Biederman (1987). A two-stage structure-mapping process, following the theory of Gentner (1983), was then used to select the answer that most closely fulfilled inferred analogical relations from the matrix. This system was tested against 48 of the 60 problems on the SPM and solved 44 of these 48 problems.

**Model 4** The system of Cirillo and Ström (2010), like that of Lovett et al. (2010), took as input hand-drawn vector graphics representations of test problems and automatically generated propositional representations. Then, like the work of Carpenter et al. (1990), the system drew from a set of predefined patterns, derived by the authors from an a priori inspection of the SPM, to find the best-fit pattern for a given problem. This system was tested against 36 of the 60 problems on the SPM and solved 28 of these 36 problems.

**Model 5** Rasmussen and Eliasmith (2011) used a spiking neuron model to induce rules for solving RPM problems. This system took as input hand-coded vectors of propositional attribute-value pairs. While the system was said to correctly solve RPM problems, no specific results were reported.

## Our Approach

As mentioned above, despite considerable differences in architecture and problem-solving focus, all five of these computational models of the RPM have reasoned over amodal propositional representations of test inputs. We believe that Raven’s problems may be solved via computational models that use purely iconic visual representations of test inputs, and we present these models as a complementary view of reasoning on the RPM.

The two models that we have developed are the affine model and the fractal model, both of which use image transformations to solve RPM problems without converting the input images into any sort of propositional form. Previously, we described each of the models along with an analysis of their performance on all 60 problems from the SPM (Kunda, et al., 2010).

## Iconic Visual Reasoning

The affine and fractal methods differ in important ways, but share two intuitions: comparing images under a variety of transformations, and judging the similarity based upon features which arise from the images.

## Similitude Transformations

Each of our algorithms compares images (or fragments of images) under a variety of transformations. We use

similitude transformations, similarity-preserving transformations which are a subset of affine transformations. Similitude transforms are a linear composition of a dilation, an orthonormal transformation, and a translation. Our implementation presently examines images under eight orthonormal transformations, specifically dihedral group D4, the symmetry group of a square. The translation is determined as a consequence of the searching each algorithm performs. The affine method restricts dilation to a value of one, i.e. no scaling, whereas the fractal method uses a short sequence of progressively smaller dilation values. Thus, the fractal method's similitude transformations are contractive.

There is evidence that human visual processing can apply some of these types of transformations to mental images, or at least operations that are computationally isomorphic in some sense. In the theory of mental imagery proposed by Kosslyn, Thompson, and Ganis (2006), transformations of mental images include scanning (i.e. translation), zooming (i.e. scaling), and rotation, among others.

### A Model of Similarity

Our models must judge the similarity between images. The nature of this similarity may be determined by any number of means, many of which might associate visual or geometric features to points in a coordinate space, and compute similarity as a distance metric (Tversky 1977). Tversky developed an alternate approach by considering objects as collections of features, and similarity as a feature-matching process. We adopt Tversky's interpretation, and seek to derive a set of features for use in our matching process.

We desire a metric of similarity which is normalized, one where the value 0.0 means entirely dissimilar and the value 1.0 means entirely similar. We use the ratio model of similarity as described in (Tversky 1977), wherein the measure of similarity  $S$  between two representations  $A$  and  $B$  is calculated by the formula:

$$S(A,B) = f(A \cap B) / [f(A \cap B) + \alpha f(A-B) + \beta f(B-A)]$$

where  $f(X)$  is the number of features in the set  $X$ . Tversky notes that the ratio model for matching features generalizes several set-theoretical models of similarity proposed in the psychology literature, depending upon which values one chooses for the weights  $\alpha$  and  $\beta$ .

Although the same equation is used for similarity calculations, each of our models has its own interpretation of what constitutes a feature. In the affine method, a feature is defined to be a single pixel, and intersection, union, and subtraction operations are defined as the minimum, maximum, and difference of pixel values. This formulation assumes that pixels are independent features within the pixel sets represented by images  $A$  and  $B$ . While this notion of pixel independence is a strong simplification, it matches assumptions made by basic template theories of visual similarity that define similarity based purely on evaluations of the extent of overlapping figural units (Palmer, 1978), e.g. individual pixels. The fractal method uses features derived from different combinations of elements from the

fractal representation of the image comparison (McGreggor, Kunda, & Goel, 2010).

#### For each base transform $t_i$ :

- Apply  $t_i$  to image  $A$  to create image  $t_i(A)$ .
- Search all possible translation offsets between images  $t_i(A)$  and  $B$  to find single offset  $(x,y)$  yielding highest similarity between them.
- Calculate similarity  $s$  between images  $t_i(A)_{(x,y)}$  and  $B$
- For set-theoretic addition and subtraction, determine image composition operation  $\oplus$  and operand  $X$  as follows:
  - If  $\Sigma(A-B) = 0$ , then  $\oplus$  and  $X$  are null.
  - If  $\Sigma(A-B) = \Sigma(B-A)$ , then  $\oplus$  refers to image addition and  $X = B - t_i(A)_{(x,y)}$ .
  - If  $\Sigma(A-B) > \Sigma(B-A)$ , then  $\oplus$  refers to image subtraction and  $X = t_i(A)_{(x,y)} - B$ .

The composition transformation  $T_i$  is thus defined as precisely the transformation that changes image  $A$  into image  $B$ :

$$T_i(A) = t_i(A)_{(x,y)} \oplus X = B$$

#### Algorithm 1. Inducing a composite transform

### The Affine Model

Given a matrix problem, the affine model makes two basic assumptions: (a) that collinear elements are related by a composition of a similitude and/or set-theoretic transform, and (b) that parallel sets of elements share identical or analogous transforms. The model proceeds in three steps:

- 1) Induce a best-fit composite transform for a set of collinear elements in the matrix.
- 2) Apply this transform to the parallel set of elements containing the empty element; the result is a predicted answer image.
- 3) Compare this predicted image to the given answer choices for maximum similarity.

Algorithm 1 shows how, for a pair of images  $A$  and  $B$ , the "best-fit" composite transform is induced. The base unary transforms are the eight orthonormal symmetry transforms mentioned above (image rotations and mirrors), along with image addition (union of sets) and image subtraction (complement of sets). The base binary transforms are the five set operations of union, intersection, subtraction (both directions), and exclusive-or.

There are two places at which the affine model computes visual similarity, first in the induction of a best-fit composite transform, and second in the selection of the answer choice that most closely matches the predicted image. In addition to using Tversky's ratio model of similarity, as defined above, we also implemented a sum-squared-difference measure, which we converted to a measure of similarity (with minimum value of 0.0 and maximum value of 1.0) as:

$$SSDsimilarity = 1 / (1 + SSD)]$$

These two similarity measures exhibit different behaviors. The Tversky measure privileges matches that share more pixel content. In contrast, the SSD similarity measure



effectively ignores any pixel content that is shared; similarity is calculated only as a function of pixels that are different.

Once the transformation is found that maximizes similarity, the transformation is applied to the first entry or entries in the last row or column, as shown in Figure 2. The resulting image represents the algorithm's best guess as to the missing entry. This image is compared to the answer choices, and the best match is chosen as the final answer.

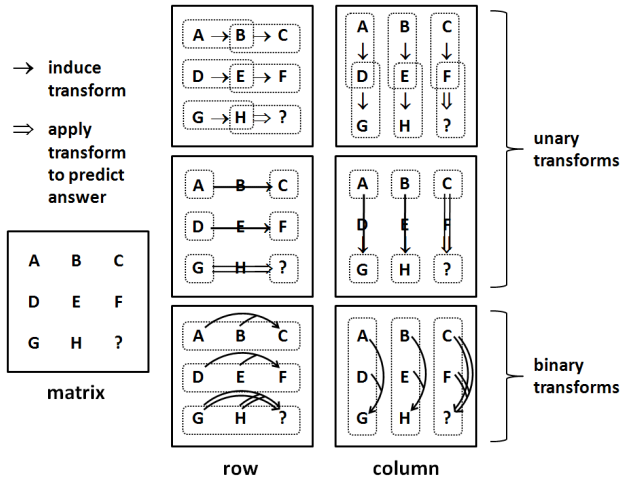


Figure 2. Sets of elements examined by the affine method

## The Fractal Method

Like the affine method, the fractal method seeks to find a re-representation of the images within a Raven's problem as a set of similitude transformations. Unlike the affine method, the fractal method seeks these representations at a significantly finer partitioning of the images, and uses features derived from these representations to determine similarity for each possible answer, simultaneously, across the bulk of relationships present in the problem.

For visual analogy problems of the form  $A : B :: C : ?$ , each of these analogy elements are a single image. Some unknown transformation  $T$  can be said to transform image  $A$  into image  $B$ , and likewise, some unknown transformation  $T'$  transforms image  $C$  into the unknown answer image. The central analogy in the problem may then be imagined as requiring that  $T$  is analogous to  $T'$ . Using fractal representations, we shall define the most analogous transform  $T'$  as that which shares the largest number of fractal features with the original transform  $T$ .

To find analogous transformations for  $A : B :: C : ?$ , the fractal algorithm first visits memory to retrieve a set of candidate solution images  $X$  to form candidate solution pairs in the form  $\langle C, X \rangle$ . For each candidate pair of images, we generate a fractal representation of the pairing from the fractal encoding of the transformation of candidate image  $X$  in terms of image  $C$ . We store each transform in a memory system, indexed by and recallable via each associated fractal feature.

**First, systematically partition  $D$  into a set of smaller images, such that  $D = \{d_1, d_2, d_3, \dots\}$ .**

**For each image  $d_i$ :**

- Examine the entire source image  $S$  for an equivalent image fragment  $s_i$  such that an affine transformation of  $s_i$  will likely result in  $d_i$ .
- Collect all such transforms into a set of candidates  $C$ .
- Select from the set  $C$  that transform which most minimally achieves its work, according to some predetermined metric.
- Let  $T_i$  be the representation of the chosen transformation associated with  $d_i$ .

**The set  $T = \{T_1, T_2, T_3, \dots\}$  is the fractal representation of the image  $D$ .**

Algorithm 2. Fractal Representation of  $D$  from  $S$

To determine which candidate image results in the most analogous transform to the original problem transform  $T$ , we first fractally encode that relationship between the two images  $A$  and  $B$ . Next, using each fractal feature associated with that encoding, we retrieve from the memory system those transforms previously stored as correlates of that feature (if any). Considering the frequency of transforms recalled, for all correlated features in the target transform, we then calculate a measure of similarity.

**Determining Fractal Similarity** The metric we employ reflects similarity as a comparison of the number of fractal features shared between candidate pairs taken in contrast to the joint number of fractal features found in each pair member (Tversky 1977). The measure of similarity  $S$  between the candidate transform  $T'$  and the target transform  $T$  is calculated using the ratio model. This calculation determines the similarity between unique pairs of transforms. However, the problems from the Raven's test, even in their simplest form, poses an additional concern in that many such pairs may be formed.

**Reconciling Multiple Analogical Relationships** In  $2 \times 2$  Raven's problems, there are two apparent relationships for which analogical similarity must be calculated: the horizontal relationship and the vertical relationship. Closer examination of such problems, however, reveals two additional relationships which must be shown to hold as well: the two diagonal relationships. Furthermore, not only must the "forward" version of each of these relationships be considered but also the "backward" or inverse version. Therefore for a  $2 \times 2$  Raven's problem, we must determine eight separate measures of similarity for each of the possible candidate solutions.

The  $3 \times 3$  matrix problems from the APM introduce not only more pairs for possible relationships but also the possibility that elements or subelements within the images exhibit periodicity. Predictably, the number of potential analogical relationships blooms. At present, we consider 48 of these relationships concurrently.

**Relationship Space and Maximal Similarity** For each candidate solution, we consider the similarity of each potential analogical relationship as a value upon an axis in a

Model	Representation	Input	Strategy	Set of APM	
				I (12 problems)	II (36 problems)
FairRaven	propositional	hand-coded propositions	prediction	7*	16*
BetterRaven	propositional	hand-coded propositions	prediction	7*	25*
Affine	iconic	scanned images	prediction	7	14
Fractal	iconic	scanned images	estimate / refine	12	26

\*out of 7 and 27 problems attempted, respectively

Table 1: Affine and fractal results on the APM compared with results of the Carpenter et al. (1990) model.

large “relationship space.” To specify the overall fit of a candidate solution, we construct a vector in this multidimensional relationship space and determine its Euclidean distance length. The candidate with the longest vector length is chosen as the solution to the problem.

The fractal method is described in more detail in McGregor, Kunda, and Goel (2010, 2011).

## Method

We tested our affine and fractal models on all 48 problems from the Raven’s Advanced Progressive Matrices test, 12 on Set I, and 36 on Set II. To obtain visual inputs, we scanned paper copies of each test at 200 dpi and manually corrected for small ( $\pm 3^\circ$ ) rotational misalignments. Thus, the input to the models was grayscale images in the PNG format, with each image containing a single problem (matrix and answer choices).

The models used a semi-automated procedure to extract individual sub-images from each problem image. Each 3x3 problem contained 8 sub-images (plus one target blank) for the matrix entries and 8 sub-images for the answer choices.

The models were run against two variations of the test inputs: raw inputs and quantized inputs. For the raw inputs, grayscale values were extracted directly from the original PNG images, and no color correction of any kind was performed. The raw inputs contained numerous pixel-level artifacts and some level of noise. For the quantized inputs, each grayscale value was rounded to be either white or black, thus turning the inputs into pure black-and-white images as opposed to grayscale.

In addition, each model considered multiple strategies when solving the problems. The affine method used two different similarity measures (Tversky and SSD). The fractal method used three different groupings of relationships (horizontal, vertical or both).

## Results

Across all input variations and strategies, the affine model correctly solved 7 of the 12 problems on Set I, and 14 of the 36 problems on Set II. These levels of performance generally correspond to the 25<sup>th</sup> percentile for both sets, for 20- to 62-year-olds (US norms) (Raven et al. 2003). Looking at input variations individually, the scores were 7 and 10 on each set for raw input, and 6 and 12 on each set for quantized input. Of the similarity measures used, the

best scores were achieved using the Tversky measure on the quantized set, with scores 6 and 12 on sets I and II respectively.

Likewise, the fractal algorithm correctly solved all 12 of the problems on Set I, and 26 of the 36 problems on Set II. This level of performance corresponds to the 95<sup>th</sup> percentile for set I, and the 75<sup>th</sup> percentile for set II, for 20- to 62-year-olds (Raven et al. 2003). Looking at input variations individually, the scores were 10 and 21 on each set for raw input, and 7 and 18 on each set for quantized input. Of the groupings used by the fractal method, the best scores were achieved by considering both horizontal and vertical groupings on raw input, at 7 and 17 on sets I and II respectively.

In comparison, Carpenter et al. (1990) report results of running two versions of their algorithm (FairRaven and BetterRaven) against a subset of the APM problems. Their results, and ours, are given in Table 2. On the ones not attempted by Carpenter et al. (1990), our methods score 4 and 5 on set I (of the 5 skipped), and 4 and 7 on set II (of the 9 skipped), for affine and fractal respectively.

## Discussion

We have presented two different models that use purely iconic visual representations and transformations to solve many of the problems on the Raven’s Advanced Progressive Matrices test. Our results align strongly with evidence from typical human behavior suggesting that multiple cognitive factors underlie problem solving on the APM, and in particular, that some of these factors appear based on visual operations. Additionally, in so far as we know, this work is the first report of any computational model addressing the entirety of the APM test, the first in which the problems are attempted using purely iconic visual representations, and the first to tackle the test using scanned images of each test page, without any re-rendering or representational change of inputs from those presented to a human test-taker.

This robust level of performance calls attention to the visual processing substrate shared by the affine and fractal algorithms: similitude transforms as a mechanism for image manipulation, and the ratio model of similarity as a mechanism for image comparison. Of course, there are many other types of visual processing that may or may not be important for accounts of visual analogy, such as non-similitude shape transformations or image convolutions, which certainly bear further investigation.

While it has been shown (Davies et al. 2008) that visuospatial knowledge alone may be sufficient for addressing many analogy problems, the representations used in that work were still propositional. In contrast, the methods described here use only visual representations. We believe the visual methods we have presented for solving the APM can be generalized to visual analogy in other domains, such as other standardized tests (e.g. the Miller's Geometric Analogies test), as well as to tests of visual oddity. We conjecture that these methods may provide insight into general visual recognition and recall. Cognitively, we hold that these strategies are a reflection of what Davis et al. (1993) referred to as the deep, theoretic manner in which representation and reasoning are intertwined.

## Acknowledgments

This research has been supported by NSF (RI) Grant #1116541, entitled "Addressing Visual Analogy Problems on the Raven's Intelligence Test," by the ONR through a NDSEG fellowship, and by the NSF GRFP fellowship program.

## References

- Biederman, I. 1987. Recognition-by-components: A theory of human image understanding. *Psych. Rev.*, 94, 115-147.
- Berker, E., & Smith, A. (1988). Diaschisis, site, time and other factors in Raven performances of adults with focal cerebral lesions. *Int. J. Neuroscience*, 38(3-4), 267-285.
- Bölte, S., Dziobek, I., & Poustka, F. (2009). Brief report: The level and nature of autistic intelligence revisited. *J. Autism and Developmental Disorders*, 39(4), 678-682.
- Bringsjord, S., & Schimanski, B. (2003). What is artificial intelligence? Psychometric AI as an answer. In *IJCAI* (Vol. 18, pp. 887-893).
- Carpenter, P., Just, M., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, 97(3), 404-431.
- Cirillo, S., & Ström, V. (2010). An anthropomorphic solver for Raven's Progressive Matrices (No. 2010:096). Goteborg, Sweden: Chalmers University of Technology.
- Davies, J., Goel, A., & Yaner, P. (2008). Proteus: A theory of visual analogies in problem solving. *Knowledge-Based Systems*, 21(7), 636-654.
- Davis, R. Shrobe H., and Szolovits, P. 1993. What is a Knowledge Representation? *AI Magazine* 14.1:17-33.
- Dawson, M., Soulières, I., Gernsbacher, M. A., & Mottron, L. (2007). The level and nature of autistic intelligence. *Psychological Science*, 18(8), 657-662.
- DeShon, R., Chan, D., & Weissbein, D. (1995). Verbal overshadowing effects on Raven's Advanced Progressive Matrices: Evidence for multidimensional performance determinants. *Intelligence*, 21(2), 135-155.
- Dillon, R., Pohlmann, J., & Lohman, D. (1981). A factor analysis of Raven's Advanced Progressive Matrices freed of difficulty factors. *Educational and Psychological Measurement*, 41, 1295-1302.
- Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7(2), 155-170.
- Hunt, E. (1974). Quote the raven? Nevermore! In L. W. Gregg (Ed.), *Knowledge and Cognition* (pp. 129-158). Hillsdale, NJ: Erlbaum.
- Kosslyn, S., Thompson, W., & Ganis, G. (2006). *The case for mental imagery*. New York, NY: Oxford U. Press.
- Kunda, M., and Goel, A.K. (2008). "How Thinking in Pictures can explain many characteristic behaviors of autism." In *Proceedings of the 7th IEEE International Conference on Development and Learning*, pp. 304-309.
- Kunda, M., & Goel, A. K. (2011). Thinking in Pictures as a cognitive account of autism. *J. Autism and Developmental Disorders*, 41(9): 1157-1177.
- Kunda, M., McGreggor, K., & Goel, A. K. (2010). Taking a look (literally!) at the Raven's intelligence test: Two visual solution strategies. In *Proc. 32nd Annual Conf. Cognitive Science Society*, pp. 1691-1696.
- Lovett, A., Forbus, K., & Usher, J. (2010). A structure-mapping model of Raven's Progressive Matrices. In *Proc. 32nd Annual Conf. Cognitive Science Society*.
- Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in Raven's Standard Progressive Matrices. *Intelligence*, 32(4), 411-424.
- Mackintosh, N., & Bennett, E. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence*, 33(6), 663-674.
- McGreggor, K., Kunda, M., & Goel, A. K. (2010). A fractal approach towards visual analogy. In *Proc. 1st Int. Conf. Comp. Creativity*, Lisbon, Portugal, January, 2010.
- McGreggor, K., Kunda, M., & Goel, A.K. (2011). Fractal analogies: Preliminary results from the Raven's test of intelligence. In *Proc. Second International Conference on Computational Creativity*, Mexico City, April 2011.
- Palmer, S.E. (1978). Structural aspects of visual similarity. *Memory and Cognition*, 6(2), 91-97.
- Prabhakaran, V., Smith, J., Desmond, J., Glover, G., & Gabrieli, J. (1997). Neural substrates of fluid reasoning: An fMRI study of neocortical activation during performance of the Raven's Progressive Matrices test. *Cognitive Psychology*, 33(1), 43-63.
- Rasmussen, D., & Eliasmith, C. (2011). A neural model of rule generation in inductive reasoning. *Topics in Cognitive Science*, 3(1), 140-153.
- Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. San Antonio, TX: Harcourt Assessment.
- Schooler, J., & Engstler-Schooler, T. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22(1), 36-71.
- Schooler, J., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *J. Experimental Psychology: General*, 122(2), 166-183.
- Snow, R., Kyllonen, P., & Marshalek, B. (1984). The topography of ability and learning correlations. *Advances in the Psychology of Human Intelligence*, 2, 47-103.
- Soulières, I., Dawson, M., Samson, F., Barbeau, E., Sahyoun, C., Strangman, G., et al. (2009). Enhanced visual processing contributes to matrix reasoning in autism. *Human Brain Mapping*, 30(12), 4082-107.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327-352.
- van der Ven, A. & Ellis, J. (2000). A Rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences*, 29(1), 45-64.
- Vigneau, F., & Bors, D. (2008). The quest for item types based on information processing: An analysis of Raven's Advanced Progressive Matrices, with a consideration of gender differences. *Intelligence*, 36(6), 702-710.
- Villardita, C. (1985). Raven's Colored Progressive Matrices and intellectual impairment in patients with focal brain damage. *Cortex*, 21(4), 627-634.

# Corpus-based metrics for assessing communal common ground

Roman Kutlak (r04rk9@abdn.ac.uk)

Kees van Deemter (k.vdeemter@abdn.ac.uk)

Chris Mellish (c.mellish@abdn.ac.uk)

Computing Science Department, University of Aberdeen  
Aberdeen AB24 3UE, Scotland, UK

## Abstract

This article presents the first attempt to construct a computational model of common ground. Four corpus-based metrics are presented that estimate what facts are likely to be in common ground. The proposed metrics were evaluated in an experiment with human participants, focussing on a domain of famous people. The results are encouraging: two of the proposed metrics achieved a large positive correlation between the estimates of how widely known a property of a famous person is and the percentage of participants who knew the corresponding property.

**Keywords:** Common Ground; Common Knowledge; Mutual Knowledge; Evaluation with human subjects; Web as corpus

## Introduction

Assessing other people's knowledge is crucial in many situations. Teachers, for example, do well to highlight information that their pupils do not know. Examples in other areas abound. Suppose, for example, we want to persuade you to reduce your intake of butter. We might do this by telling you "butter gives you high cholesterol". This argument only works if you, the hearer, know that cholesterol is bad for you, as is often assumed, for instance because it raises the likelihood of heart disease. The (presumed) fact that cholesterol is bad for you happens to be well publicised, and this might be what lies behind our assumption that you know it. Similar examples obtain in advertising, where companies might persuade you to buy a toothpaste by saying it contains fluoride, because they assume that many viewers know that fluoride is good for your teeth. It is often important to distinguish between knowledge and belief, but we will focus on cases where the distinction is less than crucial.

The difference between information assumed to be "given" (i.e., known by the hearer) and "new" (i.e., privileged information of the speaker) is crucial to philosophers, logicians and linguists (Frege (1892 (1952)); Strawson (1952); Van Eijck (1993), to mention but a few) and it is highly relevant to computational linguists working on Natural Language Generation (NLG) programs (Reiter & Dale, 2000), whose output is meant to mimic human language use. A central example is the generation of referring expressions, which has been studied extensively over the last 20 years (Krahmer & van Deemter, 2012). For example, an NLG program that aims to identify a person would do well to express properties that are likely to be known by the reader. For example, the expression "the former member of Led Zeppelin" would not be very informative to a hearer who has never heard of Led Zeppelin.

To the best of our knowledge, no general computational models exist for assessing what knowledge is likely to be known. In this paper, we examine a corpus-based strategy for building such a computational model. But, before we go into the details of our approach, there are some terminological and conceptual issues to be clarified. Common and mutual knowledge have been defined in different ways. In this paper, we shall follow the terminology of Vanderschraaf and Sillari, which the authors clarified with the following example.

Suppose each student arrives for a class meeting knowing that the instructor will be late. That the instructor will be late is mutual knowledge, but each student might think only she knows the instructor will be late. However, if one of the students says openly, "Peter told me he will be late again", then the mutually known fact is now commonly known. Vanderschraaf and Sillari (2009)

Thus, mutual knowledge is knowledge shared by a group of people. Common knowledge might be informally characterised as knowledge that is *publicly* shared by a group of people. Slightly more precisely, A and B have *mutual knowledge* of p if and only if A knows p and B knows p. They have *common knowledge* of p if they have mutual knowledge of p, and A knows that B knows p, and B knows that A knows p, and A knows that B knows that A knows p, and so on, *ad infinitum* (Lewis, 1969). Logicians and game theorists have proposed various precise definitions of common knowledge (including cases with more than two knowers), typically cast in epistemic logic, which formalise the "ad infinitum" (above) in different ways (Vanderschraaf and Sillari (2009)). For reasons that will become clear later, we use a third term that is often used in this connection, *common ground*, in a loose sense, when the distinction between mutual and common knowledge is irrelevant.

The psychologists Clark and Marshall observed that, in simple situations, common knowledge is enforced by "triple co-presence", where the speaker, the hearers and entities are physically present and the speaker believes that the hearers attend to the entities (Clark and Marshall (1981)). They contrast this simple situation (which they call *personal* common ground) with *communal* common ground, which arises not from physical co-presence but from being in a shared community (e.g., people living in Paris). Speakers are frequently able to distinguish between knowledge that is available to

members of such communities or to outsiders (Jucks, Becker, & Bromme, 2008; Nickerson, Baddeley, & Freeman, 1987). Personal common ground comes from joint personal experience of the agents; communal common ground derives from a range of sources, including the likelihood of common experience. For example, suppose Paris residents see the Eiffel Tower as they travel around their city. As they have no reason to believe that others do not see the same Eiffel Tower, they can exploit the knowledge shared by the residents of Paris namely, that the Eiffel Tower is in Paris. They can use this mutual knowledge as a shared basis for the communal common ground. In this case, the community are the Paris residents.

Our purpose was not to design a new theory of common ground, but to estimate what atomic facts are likely to be in (communal) common ground, using a corpus-based method. The method is based on metrics that use the frequency of information in a corpus to predict how widely known the information is. If it is successful, the method could be used to model different communities by studying different corpora. In other words, we offer a *parametrised* model, that has a corpus (or, equivalently, a community) as its parameter.

We focus in this paper on reference to famous people, because famous people are a prime instance of something people actually have common knowledge about. The proposed metrics will be used as one of a number of heuristics for selecting the content of a description of a famous person.

## Estimating Common Ground

The proposed heuristic simplistically hypothesises that facts that often co-occur in a document are more likely to be known because of frequency and repetition effects on memory (Atkinson & Shiffrin, 1968). Additionally, if a fact occurs frequently in a very large corpus (such as the world-wide web), which has many authors, then this implies that many people (i.e., many authors) know this fact. It seems plausible that people write about what they themselves know but, to our knowledge, no one attempted to examine how much can shallow corpus methods (i.e., methods not involving semantic analysis) tell us about mutual knowledge (Vanderschraaf & Sillari, 2009).

Assessing mutual knowledge is of substantial interest, and practical use, in its own right. So, how about common knowledge? If a fact occurs frequently in a corpus, is this evidence for common knowledge (as opposed to just mutual knowledge)? Clark (1996) essentially answered this question in the affirmative. He suggested that instead of thinking directly in terms of the knowledge of others, people use evidence as a basis for common ground (as in our example of the Eiffel Tower). Nickerson et al. (1987) and a series of studies performed by R. Krauss and S. Fussell (Fussell & Krauss, 1991; Krauss & Fussell, 1991) showed that people often use their own knowledge to estimate what others know. Given that it would not make much sense for computers to use “their own” knowledge, an alternative source of knowledge has to

be utilised, and we have chosen corpora as such a source of knowledge. Corpora in combination with measures of association were previously used in distributional models of semantic representation, where the main assumption is that if words appear in a similar context they have a similar meaning (Firth, 1957). Riordan and Jones (2011) examined in detail several distributional and feature-based models and concluded that they performed similarly well. Jurafsky (2003) argued that probabilistic modelling, as used by computational linguists, can effectively model some of the phenomena observed by psychologists. Given the past success of probabilistic models applied to various tasks including distributional semantics (Baroni & Lenci, 2010), we believed that the use of these techniques as a tool for estimating common ground was at least promising.

## Measures of association

Below, we list some of the main metrics that have been proposed for measuring the strength of association between words. These metrics assume that a *context* for the words has been defined. The context is frequently defined as a limited number of words before or after the target word or a short frame such as a paragraph in which the target word occurs. These contexts are not suitable for our purpose, because a fact about a person can be mentioned further away from the person’s name, especially if the name is pronominalised in consequent paragraphs. Instead, we will use an article as a context for our search. This can be, for example, a news article or a Wikipedia article.

**Frequency** The simplest measure of association between a person and a property (a fact about a person) is the frequency of occurrence of the name and the property together in a corpus. Taking a collection of documents as a corpus, frequency corresponds to the count of articles that contain the name and the property. This association is then the value of  $count(n, p)$  where  $n$  stands for the name of an entity and  $p$  is the property in question.

**Conditional Probability** A more sophisticated measure is conditional probability calculated as (1) and (2), where (1) measures the probability of the name given a property, and (2) measures the probability of a property given the name of the person. While the former measure normalises the results by the frequency of the property, the later measure takes into account how famous each person is.

$$assoc_{prob}(n, p) = P(n|p) = \frac{count(n, p)}{count(p)} \quad (1)$$

$$assoc_{prob}(p, n) = P(p|n) = \frac{count(p, n)}{count(n)} \quad (2)$$

**Pointwise Mutual Information** (PMI) (Fano, 1961) is a measure that compares how often two events  $x$  and  $y$  occur together. PMI exploits the fact that if two terms appear together often their joint probability ( $P(n, p)$ ) will be higher than if

they were independent ( $P(n)P(p)$ ). The value of PMI is positive for terms that co-occur and negative otherwise.

$$assoc_{PMI}(n, p) = \log_2 \frac{P(n, p)}{P(n)P(p)} \quad (3)$$

One problem with PMI is that infrequent words that only appear together achieve a disproportionately high score. In order for a property to be in common ground, it also has to be frequently mentioned. To mitigate the problem, (Hodges, Yie, Reighart, & Boggess, 1996) suggest multiplying each PMI score by  $count(n, p)$ . To reduce the big difference between the numbers of documents and to take into consideration the association as measured by PMI as opposed to the mere count, we multiply the PMI scores by the square root of the count. The final formula used for calculating the association is given by (4). Our pilot experiment showed better results with the adjusted PMI metric and any subsequent reference to PMI refers to (4).

$$assoc_{PMI}(n, p) = \sqrt{count(n, p)} * \log_2 \frac{P(n, p)}{P(n)P(p)} \quad (4)$$

## Search Engine as Corpus

How do we acquire the frequencies  $n$  and  $p$  in the metrics above? Turney (2001) successfully used the AltaVista search engine to measure association between words using a variation of PMI called PMI-IR, where he used numbers of hits returned by the search engine instead of real corpus probabilities. The number of hits corresponds to the number of documents on the Internet that contain the search term. The functionality of providing the number of hits is available from other search engines. Google does not respond to queries from programs other than web browsers but offers Google Custom Search which allows programmers to achieve the same functionality upon registration. Note that each of the search engines only searches a subset of all the documents available on the Internet and these subsets can differ substantially. This is also the case for Google accessed from a web browser and from a Google Custom Search.

There has been a debate as to whether to use search engines for research purposes (Kilgarriff, 2007; Pedersen, 2008). One of the arguments against using search engines was that the queries are optimised by performing morphological adjustments such as stemming and by looking up related words or synonyms. While these issues are pertinent to lexicography, they seem to be of less importance when it comes to establishing an association between a person and a property. A property can be described by different words and so such optimisations can in fact be very useful. There can be a problem with morphological changes to names but many search engines also offer search for exact phrases that are not morphologically or semantically manipulated and so we can avoid optimisations at places where they are undesirable. This is usually achieved by embedding the searched string in quotes.

Another problem that comes with the usage of search engines is the fact that we do not know the number of searched documents. This is necessary for calculating the probabilities used by the PMI metric. We have chosen a large constant  $N = 1.0e12$  for normalising the counts (the number of search results for the word *the* is about 25 billion on Google and about 10 billion on AltaVista and Bing and 2.4 billion on Google custom search).

## Experiment

We performed an experiment to evaluate how well the different heuristics perform. Given a person and a set of properties, our heuristics produce a set of  $\langle \text{property} : \text{score} \rangle$  pairs, where the *score* for each property is calculated by one of the described measures of association. Our goal is to assign scores so that they reflect the commonality of a particular property with regard to the name. This means that properties that are often associated with a name (e.g., Isaac Newton was a physicist) should get a higher score than properties that are less frequently associated with the name (e.g., Isaac Newton was the warden of the Royal Mint).

We used hearers' individual knowledge to assess how well the proposed heuristics perform. More specifically, the participants viewed statements such as "*Andy Warhol was American*" and "*Ernest Hemingway is the author of For whom the bell tolls*" and were asked to select one of the following statements: *true*, *false* or *don't know*. Our hypothesis is that **when a metric assigned a property higher score, a higher proportion of participants should give an affirmative answer** (i.e., state that the sentence involving the property is true). The success of the metric is measured as a Spearman correlation between the output of the metric and the percentages of affirmative answers assigned to the individual statements by the participants.

## Heuristic Options and Pilot

Aside from the choice of metric, several other choices had to be made. The first choice was which search engine to use. As we had no reason to believe that a particular search engine will perform better than others, our pilot tested the metrics on the three major search engines: AltaVista (Yahoo), Bing and Google.

The second choice is what search terms to choose. Most properties can be expressed as a combination of the attribute and a value extracted from sentences such as "Alfred Nobel was born in Stockholm." Choosing the value only would lead to a loss of information, because there would be no difference between properties such as  $\langle \text{bornIn} : \text{Stockholm} \rangle$  and  $\langle \text{diedIn} : \text{Stockholm} \rangle$  (since in both cases we would only search for Stockholm). On the other hand, attributes such as *actedIn* can be expressed by many similar expressions (e.g., *starred*). In such case, using both the attribute and the value might be too restrictive. As only empirical testing can show which option is better, we tested both. In the following tables, V stands for value only (e.g., "Stockholm") and AV stands for attribute and value (e.g., "born in Stockholm").

Thirdly, there is the question what to do with synonyms. While sometimes it might help to let a metric count all synonyms of a word, as people remember concepts rather than exact words, sometimes we would prefer to look for an exact phrase. This is especially the case when the value of the property is a proper name. This means that we had the option to quote the searched term to force the used search engine to look for an exact match. Again, our pilot tested both options (i.e., quoting and no quoting).

The choices described above left us with a large number of combinations. To minimise the likelihood of type II errors, we first performed a pilot experiment. The pilot uses a different set of stimuli than the real experiment. Based on the pilot, we then selected the most promising combinations. The setup and the procedure used in the pilot experiment were similar to the actual experiment (which is described in the following sections).

Table 1: Results of the pilot study: Spearman correlation between the heuristics and knowledge of hearers.

SE + Opt	Frequency	P(n   p)	P(p   n)	PMI
AltaVista V	0.27	0.25	0.30	0.32
Bing V	0.25	0.20	0.26	0.29
Google V	<b>0.47</b>	0.14	<b>0.37</b>	<b>0.51</b>
Google AV	<b>0.60</b>	0.23	<b>0.50</b>	<b>0.64</b>

Table 1 shows the Spearman correlations between the results of the individual metrics (unquoted option) and people's judgement. The options with quoted properties proved less useful so our final evaluation used unquoted properties. The best results were achieved by using Google and expressing properties as attribute and value. Field (2009) treats values around 0.1 as indicating small effects, values around 0.3 as medium effects and values around 0.5 as large effects. This standard terminology gives our PMI and Frequency based metrics a large (positive) correlation, and our P(p | n) metric a medium (positive) correlation. To validate our results, we selected **Frequency**, **P(p | n)** and **PMI** and evaluated them on a different set of properties using the Google search engine (table 1 shows the relevant numbers from the pilot study). The results of the final evaluation can be found in the section Results and Discussion.

## Participants

71 English speakers participated in our main experiment. 5 participants were discarded because they have not finished the experiment and further 5 participants were removed because the number of errors they made was more than 4 (mean + 2 \* std. dev). The total number of participants was 61; 30 females, 29 males and 2 unspecified.

## Materials

Ten people were selected for the experiment, each of whom was famous enough that their names occurred on the BBC

Historical Figures page <sup>1</sup>. Based on the pilot, we attempted to select the 10 in such a way that they varied maximally (i.e., spaced evenly) in terms of how well known they are. We created sentences of the appropriate form from facts concerning these people mentioned in Wikipedia and the BBC Historical Figures page. We also added properties that did not hold true of the person in question to keep our participants more focused and to make it less likely that a participant answered *true* to each statement without using their knowledge. Only the true statements were used in the analysis. The false statements were used as a measure of participant's effort. Participants who answered *true* to more than 4 false statements were discarded. We used 7 true properties and 5 false (control) properties for each person. This resulted in total of 120 statements. To make the task shorter, the statements were ordered alphabetically and then split into 5 groups of 24 statements (14 true, 10 false, 2 or 3 properties of each person in a group). Participants were randomly assigned to judge the statements in one of the groups. Figure 1 shows the names that were chosen for the evaluation and table 2 shows a sample of the properties that were judged by the participants along with the percentage of agreement answers.

- Admiral Nelson
- Alfred Nobel
- Andy Warhol
- Duke of Wellington
- Emperor Hirohito
- Ernest Hemingway
- Florence Nightingale
- Heinrich Himmler
- Louis Pasteur
- Plato

Figure 1: Famous people used in the evaluation experiment.

## Procedure

In order to find a large number of participants, the experiment was conducted online using the Amazon Mechanical Turk (MTurk). The use of MTurk can have some drawbacks, because it lets participants work from home, which makes it difficult to ensure that they are fully dedicated to the task; even worse, computer programs have occasionally been known to perform the task (instead of real people). Responses collected during the pilot experiment showed a large variability in the participants' effort, the amount of time taken to complete the experiment, and a large proportion of participants from non-English speaking countries. To mitigate some of these problems, to ensure a reasonable level of proficiency in English, and to avoid automatic responses generated by computers, participants had to successfully pass a cloze test which amounted to a very strict test of their English proficiency (Stubbs & Tucker, 1974). (Only native or highly fluent speakers tend to pass.) Furthermore, the final evaluation was advertised only to the US and UK population of the MTurk. In this way, we focussed on a particular cultural-linguistic community; the choice seemed natural

<sup>1</sup><http://www.bbc.co.uk/history/historic-figures/>



Table 2: List of properties of Ernest Hemingway, corresponding condition and the percentage of affirmative answers. Rank AV and Rank V show how the corresponding properties ranked according to the PMI metric using Google with unquoted properties.

Property	Condition	Percentage	Rank AV	Rank V
Ernest Hemingway was a writer.	true	100.0	1	2
Ernest Hemingway was American.	true	100.0	2	1
Ernest Hemingway received the Nobel Prize in Literature.	true	63.6	4	4
Ernest Hemingway is the author of For whom the bell tolls.	true	54.5	3	3
Ernest Hemingway committed a suicide.	true	50.0	6	5
Ernest Hemingway was British.	false	27.3	-	-
Ernest Hemingway was born in Oak Park.	true	25.0	5	6
Ernest Hemingway received the Italian Silver Medal of Bravery.	true	20.0	7	7
Ernest Hemingway is the author of A tale of two cities.	false	13.3	-	-
Ernest Hemingway invented dynamite.	false	0.0	-	-
Ernest Hemingway died in a plane crash.	false	0.0	-	-
Ernest Hemingway was born in Paris.	false	0.0	-	-

given that the searched pages must have been written in English in order to contain the searched terms. The inclusion of the pre-requisites (cloze test and country restrictions) greatly improved the results (e.g., less variation in the time taken to complete the experiment and fewer number of participants who made errors).

The first page showed the instructions on how to answer and how to navigate the website and also urged the participants to rely on their own knowledge and avoid using the Internet to answer the questions. The participants were then asked to fill in some information such as sex, age group and interests. The participants then viewed one statement at a time and were asked to select one of the three provided options (true, don't know, false). The participants could also provide a comment for each statement. After finishing the experiment they were given an opportunity to provide additional open comments.

The search engine queries were performed over December 2011 and January 2012. To ensure replicability of the experiment, we saved all the queries and the corresponding numbers of hits returned by the search engines. These files are available on our website <sup>2</sup>.

## Results and Discussion

Table 2 contains a sample of statements that were shown to the participants. Condition *true* means that it was a true statement and *false* means it was a false (control) statement. As previously mentioned, only the true statements were used in the analysis. The percentages of affirmative answers were correlated with the output of the metrics using Spearman correlation. All calculations were performed using the R statistical package (R Development Core Team, 2010).

Table 3 shows the final results of our experiment. We used the Google search engine and tested expressing properties as attribute and value (condition AV) and as value only (condition V). The properties were unquoted in both cases.

Our results show a large positive correlation between the PMI and the Frequency based metrics, and the knowledge of people and a medium positive correlation achieved by the

Table 3: Spearman correlation between the heuristics and the knowledge of hearers. All correlations were significant at  $p < 0.001$ .

Option	Frequency	P(p   n)	PMI
Google AV	0.639	0.437	0.664
Google V	0.632	0.475	0.662

P(p | n). This suggests that a heuristic for common ground that employs either the Frequency or the PMI metric, to large extent, agrees with the knowledge of general public.

The presented heuristic seems to work relatively well with the kinds of facts that appear in natural language generation systems. A natural question is whether the heuristic can give good results for facts that are so widely known that they are not explicitly stated. There are two kinds of these facts.

The first kind are facts such as "a person has a stomach." While it seems improbable that such a fact would be explicitly mentioned in a corpus, there are ways of implying it by statements such as "a person can get a stomach flu" or "a person can increase the risk of getting stomach cancer..." These statements include the words "person" and "stomach" and the heuristic can pick up these words without doing any semantic analysis. We tested the heuristic with a few of such statements and it seems to place them in to the well known part of common ground (i.e., assigns high scores to such facts). A proper evaluation would be needed to confirm this trend.

The second kind of facts are facts such as "Einstein had a stomach." This kind of facts requires inference, e.g. Einstein is a person and people have a stomach therefore Einstein had a stomach. As our heuristic works on the surface level of the text, it will not produce the expected results for such facts.

## Conclusion and further work

We set out to find a computation estimation of common ground, starting with mutual knowledge (in the sense of Vanderschraaf and Sillari (2009)). We hypothesised that standard co-occurrence measures could be used as an approximation of a solution to the problem and tested several of

<sup>2</sup><http://www.abdn.ac.uk/~r04rk9/cge.zip>

these measures against the knowledge of people in a particular community (cf. Clark and Marshall (1981)), as acquired in a new experiment with human participants. We consider these results to be highly encouraging. They suggest that the proposed heuristic (based on either Frequency or PMI, combined with Google search) are on the right track, at least in terms of estimating how widely known an atomic fact is (i.e., mutual knowledge); in section Estimating Common Ground we argued that this also makes it plausible that these heuristics could offer a reasonable approximation of *common* knowledge (i.e., the facts of which everyone in the community knows that everyone in the community knows them), but this was not directly investigated.

The community investigated in our experiment was comprised of native speakers of English in the UK and the USA with access to a computer. We believe it would be interesting to test to what degree different communities be modelled by different knowledge sources.

Our current work focuses on combining the metrics investigated here with other heuristics (including a discriminatory power heuristic for assessing the usefulness of a fact) to improve content selection algorithms for Natural Language Generation (e.g., Reiter and Dale (2000)).

### Acknowledgements

We would like to thank the members of the Natural Language Generation group of the University of Aberdeen and the anonymous reviewers for their valuable comments. This research is sponsored by the Scottish Informatics and Computer Science Alliance (SICSA).

### References

- Atkinson, R., & Shiffrin, R. (1968). Human memory: A proposed system and its control processes. In K. Spence & J. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). Academic Press, New York.
- Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Comput. Linguist.*, 36(4), 673–721.
- Clark, H. H. (1996). *Using language*. New York: Cambridge University Press.
- Clark, H. H., & Marshall, C. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). New York: Cambridge University Press.
- Fano, R. M. (1961). *Transmission of information: A statistical theory of communications*. New York: Wiley.
- Field, A. (2009). *Discovering statistics using spss*. SAGE publications Ltd.
- Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In *In studies in linguistic analysis*. Blackwell.
- Frege, G. (1892 (1952)). On sense and reference. In P. T. Geach & M. Black (Eds.), *Translations from the Philosophical Writings of Gottlob Frege*. Oxford: Basil Blackwell.
- Fussell, S. R., & Krauss, R. M. (1991). Accuracy and bias in estimates of others' knowledge. *European Journal of Social Psychology*, 21(5), 445–454.
- Hodges, J., Yie, S., Reighart, R., & Boggess, L. (1996). An automated system that assists in the generation of document indexes. *Natural Language Engineering*, 2(02), 137–160.
- Jucks, R., Becker, B.-M., & Bromme, R. (2008). Lexical entrainment in written discourse: Is experts' word use adapted to the addressee? *Discourse Processes*, 45(6), 497–518.
- Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In *Probabilistic linguistics* (pp. 39–96). MIT Press.
- Kilgariff, A. (2007, March). Googleology is bad science. *Comput. Linguist.*, 33, 147–151.
- Krahmer, E., & van Deemter, K. (2012, March). Computational Generation of Referring Expressions: A Survey. *Computational Linguistics*, 38(1), 173–218.
- Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition*, 9(1), 2–24.
- Lewis, D. K. (1969). *Convention: A philosophical study*. Cambridge, Massachusetts: Harvard University Press.
- Nickerson, R. S., Baddeley, A., & Freeman, B. (1987). Are people's estimates of what other people know influenced by what they themselves know? *Acta Psychologica*, 64(3), 245–259.
- Pedersen, T. (2008, September). Empiricism is not a matter of faith. *Comput. Linguist.*, 34, 465–470.
- R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. (ISBN 3-900051-07-0)
- Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. New York, NY, USA: Cambridge University Press.
- Riordan, B., & Jones, M. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2), 303–345.
- Strawson, P. (1952). Introduction to logical theory.
- Stubbs, J. B., & Tucker, G. R. (1974). The cloze test as a measure of english proficiency. *The Modern Language Journal*, 58(5/6), pp. 239–241.
- Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the twelfth european conference on machine learning (ecml-2001)*.
- Vanderschraaf, P., & Sillari, G. (2009). Common knowledge. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2009 ed.). <http://plato.stanford.edu/archives/spr2009/entries/common-knowledge/>.
- Van Eijck, J. (1993). The dynamics of description. *Journal of Semantics*, 10(3), 239–267.

# Getting off at the end of the line: the estimation of large numbers

David Landy (dlandy@richmond.edu)

Department of Psychology, 28 Westhampton Way  
University of Richmond, VA 23173 USA

Noah Silbert (nsilbert@umd.edu)

Center for the Advanced Study of Language  
University of Maryland. College Park, MD

Aleah Goldin (aleah.goldin@richmond.edu)

Department of Psychology, 28 Westhampton Way  
University of Richmond, VA 23173 USA

## Abstract

Despite their importance in public discourse, numbers in the range of one million to one trillion are notoriously difficult to understand. We examine magnitude estimation by adult Americans when placing large numbers on a number line and when qualitatively evaluating descriptions of imaginary geopolitical scenarios. Common conceptions of the number line suggest a logarithmic compression of the numbers (Dehaene, 2003). Theories of abstract concept learning suggest that in situations where direct experience is unavailable, people will use the structure of notation systems as a proxy for the actual system. (Carey, 2009; Landy & Goldstone, 2007).

Evaluations across two subject populations largely matched the predictions of the latter account. Approximately 40% of participants estimated *one million* approximately halfway between *one thousand* and *one billion*, but placed numbers linearly across each half, as though they believed that the number words “thousand, million, billion, trillion” constitute a uniformly spaced count list. Very brief training procedures proved partially successful both in correcting number line placement and in shifting participants’ judgments of geopolitical situations. These results reinforce notions of abstract concepts as grounded in external notation systems, as well as having direct implications for lawmakers and scientists hoping to communicate effectively with the public.

**Keywords:** number cognition, mathematical cognition, formal reasoning, human subjects experimentation

## Introduction

Large numbers<sup>1</sup> are interesting for both practical and theoretical reasons. Many arenas of public discourse rely on an understanding of large numbers, including debates about evolutionary biology, nanotechnology, and the reliability of DNA testing. The United States is currently involved in a heated conversation about the national budget and economy. The budget, the deficit and the debt are in the low trillions, while most proposed budget changes are in the millions and billions. Americans generally exhibit poor knowledge about spending on specific programs by the federal government (Gilens 2001), and it is likely that poor understanding of large numbers contributes to this ignorance.

Number systems covering this range are also an excellent example of an abstract system: magnitudes such as *one billion* are beyond our immediate experience and yet are clearly understood in part through abstracting the concrete

process of counting (Carey, 2009; Leslie, Gelman, & Gallistel, 2008). We experience large numbers primarily syntactically, and through associations with situations (e.g., claims that the U.S. deficit is \$1.4 trillion; Facebook has 700 million users; or the human body has 100 trillion cells).

One way we understand abstractions is by studying the properties of their concrete representations (Clark, 2006; Landy & Goldstone, 2007; Kirsh, 2010). For instance, Carey (2009) proposes that when learning to count, the memorized count list orients attention to appropriate features of the environment, so that the verbal label “eighteen” cues a learner that there is *something* that “eighteen” situations have in common. In addition to the simple presence or absence of labels, however, count lists have other structural properties: for instance, counting numbers are typically stated in sequence, with accompanied rhythmic hand motions, and are constructed on a semi-regular pattern. Here, we wonder how structural components of symbolic systems impact inferences made by reasoners.

## Structure in the numerals

A student learning the English counting system must master several different lists. In addition to the numbers from 1-9, one must learn the teen words, the tens words, and –most importantly for our purposes—is the *short scale*, used in the United States and Britain. In this system, one thousand million is “one billion”. This list “thousand, million, billion, trillion, quadrillion, ...” constitutes an effective count list, which after the initial “thousand”, bears an apparent sequential structure, and clearly derives from Latin number words. North American students typically learn the short scale up to “trillion” by around 7<sup>th</sup> grade (Skwarchuk and Anglin, 2002).

There are several common notations for understanding large numbers. In this paper, we focus on perhaps the most common one, which we will call the *hybrid notation*, because it combines number words and numerals. Examples of numbers in this from include “324 million”, “426”, or “5 thousand.”

We model large number understanding by combining two conceptually separate steps: the first involves the interpretation of a number word into an abstract numerical quantity (“abstract” because we are agnostic with respect to how people would actually estimate perceived quantities in the range of millions and billions—here we mean merely the interpretation can be treated as a metric), and the mapping

<sup>1</sup>Here, roughly those between 10<sup>5</sup> and 10<sup>13</sup>.

of a quantity into a response. For brevity, we blur these distinctions here.

Of the many plausible ways that people might extract quantities from number representations, the simplest is that people might roughly correctly estimate the relative values of large numbers. We will refer to this as the *linear* or *normative* model of large number understanding.

Second, if learners use the structure of the number notations—especially the short scale—as a guide to numerical size, then a different pattern is expected. Since the number words—millions, billions, trillions, are similar and uniformly spaced in their count list, people might evenly distribute the referred quantities. Since adults generally linearly estimate numbers from 1-1000 (Seigler & Opfer, 2003), this suggests a piecewise-linear pattern, in which (roughly) values like 1 thousand, 1 million, and 1 billion are separate units, which are spaced evenly on the line, and other values (such as 500 million) are linearly interpolated between these points. We will call this the *uniform spacing* or *piecewise linear* model.

Another plausible approach is based on developmental studies of line estimation with small number ranges (Siegler & Opfer, 2003). These studies have repeatedly demonstrated that number estimation errors tend to be highly compressed at the large end of the line. Traditionally, this compression has been modeled using a logarithmic function, and a fitted linear mapping from quantities to line positions (Booth & Siegler, 2008; Siegler & Opfer, 2003). We will call this combination the *log-linear* model.

Finally, it is naturally plausible that some people would either have no interpretation of the large numbers, or highly variable or non-monotonic interpretations.

## Empirical Methodology

We used two tasks to explore the number word interpretation: Number line estimation, and situation evaluation.

In typical *number line estimation* tasks, a participant is presented a line with labeled endpoints, and a stimulus numeral. The participant makes a mark indicating their estimate of the proportion of the line that corresponds to the proportion relating the stimulus number to the specified range. In the experiments reported here, the left end was always *1 thousand*, and the right end was *1 billion*. Prior to performing estimations, participants were shown a marked number line ranging from 1 to 10, and were instructed to likewise place their numbers in a linear manner.

In *situation evaluations*, participants made qualitative judgments about attempted government actions involving short-scale quantities. In each story, one number was selected as a goal, and a number to be evaluated was selected from the preceding element of the short scale. For instance, in one question a fictional country's government had a goal to eliminate their 1.1 trillion "taler" deficit, and proposed the solution cut 100 billion talers. Participants rated the quality of the attempted solutions on a 9-point scale from "very unsatisfactory" to "very satisfactory".

## Experiment 1

### Method

**Participants & Procedure** Partial course credit or monetary compensation was given to 67 participants recruited from the University of Richmond community. Three participants gave responses that were generally non-increasing across the number range, and were extremely variable; these participants' data were removed and replaced to yield our goal of 64 participants.

Participants made 108 number line estimates, on a line ranging from 1 thousand to 1 billion. Each stimulus number was the product of an integer strictly between one and one thousand, and either  $10^3$  or  $10^6$ .

Two between-groups differences were used to rule out possible confounds in our approach. First, in Experiment 1 half of all participants viewed numbers in the hybrid notation; for half all stimuli and endpoints were presented in the pure numeral format. Second, the range of the stimulus numbers was manipulated between participants, so that we could evaluate whether people shifted their placement to fit the distribution of observed numbers. Half of the participants saw numbers only in the millions; half estimated numbers which were evenly divided between those above and below 1 million. Neither manipulation affected results qualitatively or altered significance of contrasts; similar patterns were observed across all four groups; the slight differences will not be discussed here.

After completing the experiment, regardless of condition, participants filled out a paper form prompting them to generate the numerical form for each of one billion, one million, and one thousand. All but two participants did so correctly; one participant left the "one billion" mark blank, while the other made significant errors.

### Analysis

Our primary analysis compared linear and uniform spacing model fits with the log-linear, using a hierarchical Bayesian model fitting approach.

Since both models are linear above and below one million, the primary variable distinguishing the linear and uniform spacing models is the estimated position of one million on the line ( $M$ ).  $M$  was fitted at the individual subject level; since  $M$  ranges from 0 (extreme left) to 1 (right), the population was fitted as a uninformative beta distribution. Within this framework, the linear model is the special case when  $M = 0.001$ , pure uniform spacing is produced when  $M = 0.5$ . The prior on  $M$  was uniform between 0 and 1, and 0 elsewhere.

This *segmented linear* model was compared to a log-linear model,  $y = a \ln(x)$ ; this model also has one parameter, fixing the shape of the linear component. The left intercept of both models was fixed at 0. To capture variability in responses, both models assume truncated (at 0 and 1) normal distributed deviations from the model prediction.

A hierarchical mixture model mixing both components at the group level was fit to the data using JAGS through the

rJAGS package. In this model, each subject has some probability  $\pi$  of producing split linear responses and some probability  $1 - \pi$  of producing a log-linear responses. The model thus categorizes individuals as part of the fitting procedure. The model was simulated using MCMC, with 4 chains with 100 samples per chain, a burn-in of 30,000 iterations, and a thinning of 250 iterations per sample.

## Results

Figure 1 shows the fitted values of  $M$ . Qualitatively, nearly all participants were captured very well by the segmented linear model. The logarithmic model was selected as better fitting by the model for only one participant. Three other participants produced non-monotonic fits with wide variability, and were poorly fit by both models. The remaining 61 participants matched well the predictions of the segmented linear models. Figure 2 illustrates two typical patterns of response: one group of participants ( $n = 36$ ) were fit very well by the linear model, and thus had low  $M$  values; the other group had high values of  $\pi$  with typical  $M$  values centered around 0.4 ( $n=19$ ). The few participants with intermediate  $M$  values ( $n=6$ ) between 0.1 and 0.3 seemed to switch strategies, producing responses which were sometimes close to linear, and at other times very close to the uniform spacing model.

## Discussion

Experiment 1 demonstrates that there is not a general misunderstanding of large numbers, nor a logarithmic scaling of these numbers. Instead, a single, specific misconception of large numbers predominates errors: at least 85% of substantial deviations from linear responding involved a piecewise linear behavior, in which each of the ranges of “millions” and “billions” are linearly constructed, but are each of approximately identical size. Despite the prevalence of smooth, log-like functions in theories of economic and psychological utility functions and psychological magnitudes, evaluations of large numbers appear to no more than rarely approach logarithmic scaling.

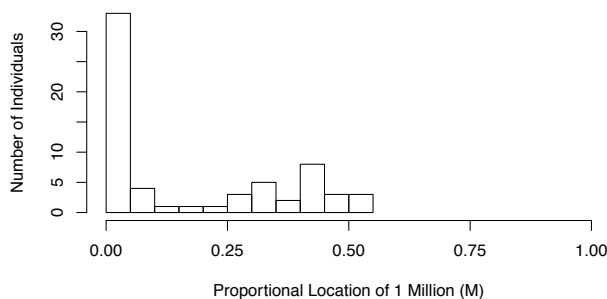


Figure 1: Histogram of individual fitted values of the position of 1 million ( $M$ ). The normative value is 0.001.

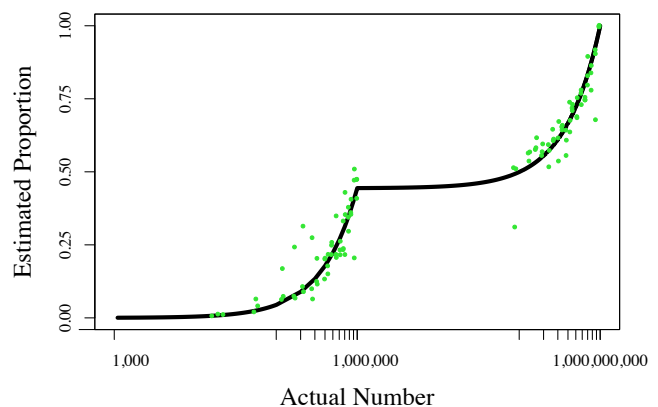
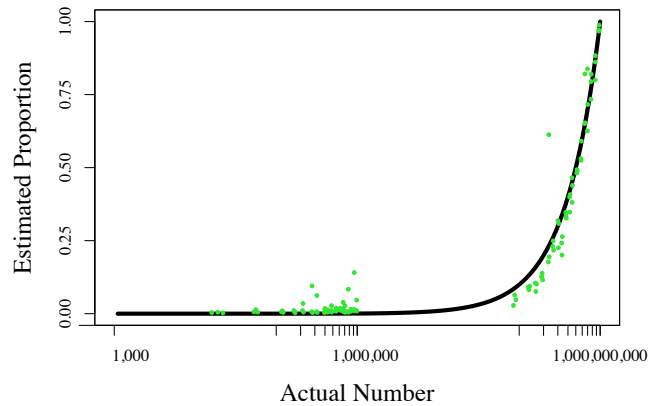


Figure 2: Number line estimates for a linear (top,  $M=0.0004$ ) and piecewise linear (bottom,  $M=0.44$ ) sample subject, along with predictions of the segmented linear model. The log-linear model predicts straight line responses on a log-scaled x-axis.

It is possible that participants in our study misconstrued the nature of the task, believing, for instance, that a segmented linear graph was requested. We believe this is unlikely for two reasons. First, although both linear and logarithmic number lines are fairly commonplace (for instance, as graph axes), segmented linear number lines—lines in which one linear number range lies adjacent to a linear range with a different unit—are vanishingly rare. Second, while piloting these materials we interviewed many individuals completing this task. While many made the error, none gave evidence having misunderstood the task. On the contrary, these individuals seemed very surprised when they realized or were told the normative location of one million.

## Experiment 2

The number line is an idiosyncratic task, involving visual and spatial components as well as number processing per se. It might be that the results of Experiment 1 result from idiosyncratic reasoning, and would not generalize well to

other kinds of number judgments. One purpose of Experiment 2, then, was to explore whether piecewise linear number line estimation would generalize to other tasks.

In studies involving smaller number ranges, learning of linear behavior can be strikingly sudden—with participants often becoming linear across an entire range from the presentation of just a single point (Opfer & Siegler, 2007). A second purpose of Experiment 2 was to explore whether similar approaches could lead to sudden reductions in misconceptions about large numbers, and such shifts in line estimation would generalize to evaluative judgments.

## Participants and Procedure

300 participants were recruited from Amazon's Mechanical Turk in exchange for small monetary remuneration. Mechanical Turk is a scalable workforce solution frequently used by psychologists to recruit subjects for online experiments (Mason & Suri, 2012). All tasks were completed remotely through a web interface.

Each participant first performed eight number line estimations (the *pretest*), followed by an intervention. Half of all participants saw an *encouragement* intervention, which simply thanked them for their hard work, and asked them to do their best on the rest of the experiment. The other half of participants saw a *training* screen, which reminded them that 1 billion was equal to 1,000 millions, and showed them the normative placement of 10 million on the number line from 1 thousand to 1 billion. Participants then completed eight more number line estimates (the *posttest*), followed by three situation evaluation questions.

In the *situation evaluation* task, participants read, in fixed order, three short narratives about how the governments of two fictional countries were dealing with various social challenges. The participants rated the quality of the attempted solutions on a 9-point scale from “very unsatisfactory” to “very satisfactory”. In each story, one number was selected as a goal, and a number to be evaluated was selected from the preceding element of the short scale. For instance, in question 3 (designed to match the U.S. budget for 2011) the goal was to eliminate the 1.1 trillion “taler” deficit, and the solution cut 100 billion “talers”. After both tasks were completed, participants reported their age, sex, and political affiliation, and briefly describing their problem-solving strategy. The strategy explanations provided an extra check that participants were in fact attempting the problems.

## Analysis and Results

**Number Line Estimation.** Estimates were modeled using a version of the model described in Experiment 1. Because the unimodal beta model at the family level did not capture the pattern of observed behaviors, in Experiment 2 data was fit only at the level of the individual. Further, the logarithmic model was not tested. Thus, the single model parameter was the estimated location of one million,  $M$ . Separate models were fit to the data before and after the intervention.

Figure 3 illustrates the shift in number line behavior before and after the intervention. An ANOVA evaluating  $M$  values as a dependent measure over time of estimation (pre vs. post intervention) and condition, indicated a significant interaction between the two ( $F(1, 298)=15.8, p<.01$ ). There was also a main effect of condition ( $F(1, 298)=4.4, p<.05$ ); considering only the pretest data, the difference was not significant ( $F(1,298)=.21, p>.0.5$ ).

As in Experiment 1, the empirical values of the  $M$  parameter were contrary to the predictions of the uniform spacing model. While the model predicts a mean value around 0.5 among the piecewise linear group, the actual mean fitted value was around 0.40.

**Situation Evaluations.** Evaluations were averaged across the three situations for analysis. These average responses were moderately normally distributed. An ANOVA of mean evaluation against pretest  $M$  and condition revealed significant effects of both ( $F(1, 298)=11.3, p<0.01$ , and  $F(1, 298)=4.3, p<0.05$ , respectively). Once behavior at posttest was included, however, it was the only significant predictor of situation evaluations ( $F(1, 298)=15.8, p<0.001$ ; see Figure 4); condition was no longer significant ( $F(1, 298)=2.4, p\sim.12$ ), suggesting that some of the effect of training on situation evaluation resulted from shifts in processes involved in number line estimation.

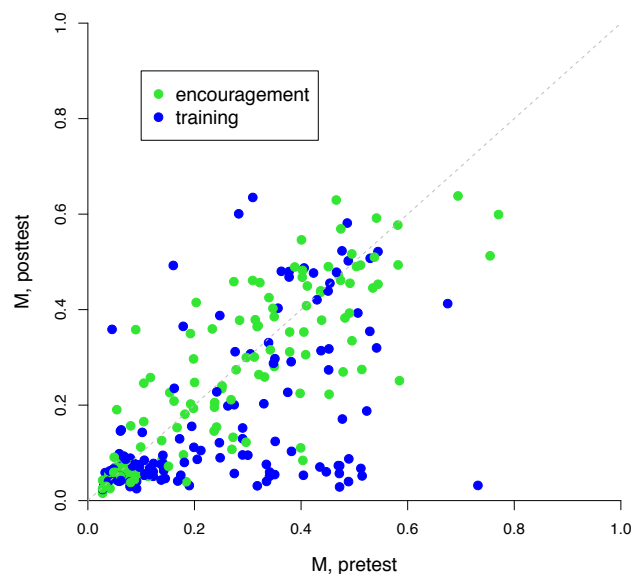


Figure 3: Best-estimated location of 1 million ( $M$ ) at pretest and posttest. The normative location is 0.001. The large preponderance of blue circles in the bottom right represents the efficacy of the training.



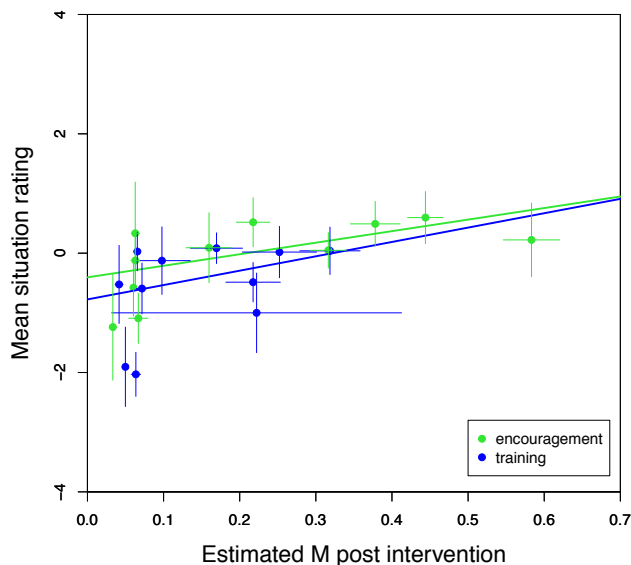


Figure 4: Situation evaluation against post-intervention placement of 1 million (M), binned into quantiles based on pre-intervention number line estimates. Errors reflect standard errors in the estimate of post-intervention values.

## Discussion

Experiment 2 demonstrated that the same strategies found in Experiment 1 are employed in a substantively different population: 45% of judgments were compatible with the piecewise linear account. However, participants were readily educable: just a single example of a normatively placed term sufficed to correct number line estimations in nearly half of error-prone estimators. Training seemed to be largely all-or-none: participants who shifted strategies halfway through the task gave responses nearly indistinguishable from those who had been following the final strategy from the beginning, judging both by number line estimations and situation evaluations.

Furthermore, Experiment 2 demonstrated that number line strategies are closely related to political judgments involving numbers in this range. People who estimate the number lines normatively are less optimistic about political situations involving numbers in this range. Furthermore, participants trained on the number line shifted their evaluations of political situations. Hence, the uniform spacing misconception does not result from reasoning specific to the number line task.

## General Discussion

Numbers picked out by the short scale—despite appearing frequently in educational contexts and public discourse—do not seem to be robustly understood by much of the population. While roughly half of our participants treated large numbers linearly, two experiments indicate that a large portion of the population—around 40 percent in the studies reported here—seems to evaluate large numbers based on the assumption that the number labels are roughly equally spaced as the numbers increase. Furthermore, people who

rely on an equal spacing heuristic when placing numbers on a line are more satisfied with poor resolutions to political problems involving comparable scales.

Currently, the people of the United States, along with many other countries, are deciding how best to handle economic debt and deficit crises. These conversations crucially involve the accurate assessment of numbers across the range of  $10^6$ - $10^{13}$ . The current results suggest that a substantial fraction of Americans are ill equipped to engage in these conversations. This conversation is of direct relevance to the practice of scientific research, which is often funded by grants in the low millions of dollars. Detractors of the government spending on science research and other programs often present funding information by contextualizing these amounts within the overall budget using short-scale labels.

Logarithmic number line behavior was rare or non-existent on this task, despite substantial prior research that has supported the hypothesis that unfamiliar number ranges are initially represented logarithmically (Siegler & Opfer, 2003; Dehaene, 2003). One possibility is that large numbers fall beyond the upper range of the approximate magnitude system (Izard & Dehaene, 2008). Another possibility is that the reasoning processes we find adults employing when estimating large numbers account for apparently logarithmic behavior in young children (Nuerk et al, 2001).

Although the hypothesis that people infer spacing on the number line from the structure of the short scale labels predicted the basic pattern of responses, it does not predict the observed structure perfectly. In particular, most people who erred in their estimate of the relative values of 1 thousand, 1 million, and 1 billion did not put 1 million halfway between the other two, but substantially close to 1 thousand. Anecdotally, people we have observed often placed 1 million more or less exactly in the middle, then ‘correct’ to approximately the 40% mark. One possibility is that this positioning reflects a compromise between uniform spacing and normative number knowledge, but the nature of such a compromise remains speculative.

These results are striking in that the actual numerical system of short scale words and place value notation is formally extremely simple, and the referent system—the natural numbers—is acquired fairly early in mathematical development. A simple induction suffices to suggest the referents of the large number words studied here, rather than a conceptual restructuring, as has been implicated in rational-number learning. These results emphasize that even when dealing with basic abstract material, accessible concrete structures play a key role in guiding the development of concepts and strategies (Carey, 2009; Goldstone & Landy, 2010). When dealing with large numbers, people rely heavily on number naming structures to fix the meaningful properties of particular number words. Instead of using the number labels as placeholders to an independently existing world, accessed via number principles, many people attend to the surface properties of number nomenclature to determine numerical properties. As



the four-year old daughter of the first author (who was at the time learning to read two digit numbers) put it “100 is just one more than 10. It’s three: one, two, three!” Magnitudes in this range are constructed by borrowing structure from the symbol systems used to represent them.

### Acknowledgments

Partial funds for this research came from an University of Richmond undergraduate research grant to the third author and Department of Education, Institute of Education Sciences grant R305A110060. . Thanks to Lisa Byrge, Iris Van Rooij, Erin Ottmar, and the Cave Lab for suggestions.

Experiment 1 is reported as Experiment 1 of Landy, Silbert, and Goldin (under review).

representations of numerical quantity. *Psychological Science*, 14, 237 – 243.

Skwarchuk, S. L., & Anglin, J. M. (2002). Children's acquisition of the English cardinal number words: A special case of vocabulary development. *Journal of educational psychology*, 94(1), 107.

### References

- Barth, H. C., & Paladino, A. M. (2011). The development of numerical estimation: evidence against a representational shift. *Developmental Science* 14, 125-135.
- Booth, J. L., & Siegler, R. S. (2008) Numerical magnitude representations influence arithmetic learning. *Child Development*, 79, 1016-1031.
- Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, 41, 189-201.
- Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.
- Clark, A. (2006). Material Symbols. *Philosophical Psychology*, 19(3), 291-307.
- Dehaene, S. (2003). The neural basis of the Weber-Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*, 7, 145-147. *Psychology*, 99(1), 1–17.
- Gilens, M. (2001). Political ignorance and collective policy preferences. *American Political Science Review*, 95(2), 379-396.
- Hollands, J. G., & Dyre, B. P. (2000). Bias in proportion judgments: The cyclical power model. *Psychological Review*, 107(3), 500-524
- Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, 106, 1221-1247.
- Kirsh, D. (2010). Thinking with external representations. *AI & Society*, 25(4), 441-454.
- Landy, D., & Goldstone, R. L. (2007). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(4), 720-733.
- Leslie, A. M., Gelman, R., & Gallistel, C. R. (2008). The generative basis of natural number concepts. *Trends in Cognitive Sciences*, 12(6), 213-218.
- Mason, W., Suri, S. (2012). Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior Research Methods*, 44, 1-23.
- Nuerk, H. C., Weger, U., & Willmes, K. (2001). Decade breaks in the mental number line? Putting the tens and units back in different bins. *Cognition*, 82(1), 25–33.
- Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple

# Tangent Point Orientation and Anticipated Trajectory Curvature - A Field Study on the Visual Control of High Speed Steering

Otto Lappi (otto.lappi@helsinki.fi)

Cognitive Science, Institute of Behavioural Sciences, PO Box 9  
00014 University of Helsinki

Esko Lehtonen (esko.lehtonen@helsinki.fi)

Traffic Research Unit, Institute of Behavioural Sciences, PO Box 9  
00014 University of Helsinki

## Abstract

An important visual strategy in the visual control of driving during curve negotiation is tangent point orientation – directing gaze to the inside road edge in the area of interest (AOI) around the tangent point. Yet, while the phenomenon has been replicated in many studies at a qualitative level, and several computational models have been proposed to explain it, there is no consensus on whether the actual gaze target is the tangent point itself – or some other road point in its vicinity- and what the functional significance TP targeting (looking at the tangent point) or TP orientation (looking at a point in the TP AOI, but not necessarily the tangent point) might be. We report here a previously unobserved dependence between gaze distribution on road curvature: gaze concentrates on the part of the road where the vehicle yaw rate (local curvature) will be highest. We therefore suggest this geometric property of the future path may act as a functionally salient visual reference for the driver, and that the oft-reported “tangent point orientation” may in some cases be a side-effect caused by the spatial contiguity of the tangent point and the point of maximal path curvature.

**Keywords:** visually guided behavior, driving, eye-movements, field studies, tangent point, optic flow.

## Introduction

One of the ultimate goals in modeling eye-movements during natural behavior would be to be able to predict the whole sequences of eye movements executed during the performance of a naturalistic task, such as reading, car driving or preparing a meal (Kowler, 2011; Land 2006, 2007). This goal may not be as far off as it may seem, for as research on eye movements in naturalistic tasks during the past two decades has shown, in naturalistic tasks (as opposed to many artificial laboratory tasks) eye-movements present a picture of surprisingly stereotypical patterns, where the subjects’ gaze behavior is closely bound to the task conditions, both in spatial terms (gaze is concentrated only on task-relevant gaze targets) and in temporal terms (gaze target selection is closely coupled to the execution of different phases of a complex task, picking out targets “just in time” – i.e. selecting targets whose state needs to be verified, or which are about to be manipulated in 1-2s, so that the use of short term visual memory can be minimized). (Ballard et al., 1995). Successful modeling would thus entail (i) identifying the relevant gaze targets, (ii) specifying their relevance to the task (phases) in terms of the cognitive com-

putations the information provided by these targets could support, and (iii) testing the differential predictions of models against data collected from participants performing actual naturalistic tasks.

In the domain of car driving, one pattern in particular has been the subject of continuous research and theoretical debate, namely, *tangent point oriented curve negotiation* (Land & Lee, 1994): directing gaze in the direction of inside road edge during curve driving (Figure 1). Yet, while the basic phenomenon has been replicated in many studies (Underwood et al, 1999; Land & Tatler, 2001; Chattington et al., 2007; Kandil et al., 2009, 2010), and several computational models have been proposed to explain it (Land & Lee, 1994; Boer, 1996; Wann & Swapp, 2000; Wann & Land, 2000; Wann & Wilkie, 2004), there is no consensus on whether the gaze target is the tangent point itself – or some other road point in its vicinity- or the functional significance of this gaze pattern.

There are several models that account for the basic pattern (see Table 1). Some assume that the tangent point itself is the relevant gaze target for visually controlled steering (its eccentricity acting as a feedback control parameter), others that it is points on the forward-planned future trajectory which the driver looks at (only falling near the tangent point due to geometric reasons).

As the models are all compatible with the general observation of “tangent point orientation”, but make subtly different predictions concerning the precise spatial and temporal dynamics of gaze, more detailed on-road data is required to arbitrate between the various models. We report here a previously unobserved dependence between gaze distribution on road curvature: gaze concentrates on the part of the road where the vehicle yaw rate (local curvature) will be highest.

We therefore suggest this geometric property of the future path may act as a functionally salient visual reference for the driver, and that “tangent point orientation” may in fact – at least in some cases - be a side-effect caused by the spatial contiguity of the tangent point and this other gaze target. (This highlights the inherent problems of using AOI based measures in naturalistic environments where the experimenter has no control over the overlap of the AOI’s of different gaze targets).

Table 1: “Tangent point orientation” models of visual control of steering.

i. Land & Lee (1994)	Drivers fixate the tangent point and use the visual angle of tangent point (or gaze) relative to the locomotor axis to judge the curvature of the bend. (This model makes no specific prediction about steering).
ii. Land & Lee (1994), Wann & Land (2000)	Drivers fixate the tangent point and actively steer so as to keep the visual angle of the tangent point (and gaze) at a <i>constant horizontal direction</i> .
iii. Wann & Land (2000), Wann & Wilkie (2004)	The driver fixates a target point on their future path (i.e. a point they wish to pass through) which is near the tangent point not necessarily the tangent point, and then steers so that <i>the fixated point sweeps from its initial offset to directly in front of the locomotor axis at a constant rate</i> ..
iv. Boer (1996)	A reference point (“next to the tangent point but slightly into the road”) is chosen, although not necessarily fixated directly. Steering and speed are controlled in the following manner: The driver observes the visual angle between the vehicle’s heading and the target point, estimates the vehicle’s speed and the geometric distance to the target point, and adjusts steering and speed so that the following constraints are satisfied (1) the visual angle to the target point shall reach zero in less time than it will take to traverse the distance to the target point (2) the trajectory minimizes the maximum required lateral acceleration (i.e. minimizes the maximum steering input), and (3) the furthest deviation from the road edge remains within lane boundaries. (Active model).
v. Kim & Turvey (1996), Wann & Swapp (2000)	Visual flow is used to steer the car on a linear or locally circular trajectory: the driver fixates a target point on the road she wishes to pass through - a reference point on the future trajectory that is at rest in the allocentric reference frame, but moves in the egocentric frames of reference due to optic flow – and steers so that the visual flow lines are straight rather than curved. What is more, all those flow lines that fall on the observer's future path will now be be vertical

## Subjects

Nineteen subjects participated in the experiment (11 male, 8 female, age range 23-52 years, mean 30, s.d. 7 years. All had held a valid driving license, and could be considered experienced drivers. Participants were recruited through university e-mail lists, some through personal contacts among students and staff. Condition for inclusion in the experiment was normal uncorrected vision (qualified to drive a car without correction) and sufficient driving experience (>20 000km). All participants were naïve to the purpose of the study (the tangent point hypothesis) and were given two cinema tickets as compensation for participating. All participants gave written informed consent, and the study was approved by the local ethics committee.

## Procedure

The test road was a 3.9 km low-standard two-lane rural road with a low traffic density and no lane markings (5.4 m pavement width). All drives were carried out in daylight. Participants drove the car to the test route, which was located 34 km from the campus. A few kilometers before arriving at the test road, the instrument panel was occluded. This was in done to reduce downwards glances to the speedometer during the test run, giving us the best opportunity to record road-directed gaze patterns. The participants did not express discomfort at having to drive without a speedometer. In addition to the participant who drove the car, a member of university staff acted as driving instructor on the front seat, giving route directions and ensuring safety.

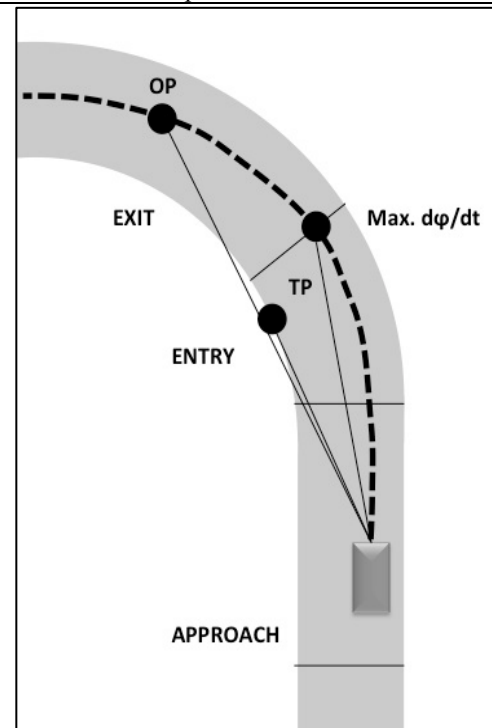


Figure 1: Illustration of some possible gaze targets in curve driving, discussed in the text. The future path of the vehicle is indicated by dotted line TP = tangent point; OP = occlusion point; point of maximum yaw rate is the point where the driver can begin to unwind the steering and accelerate out of the turn.

A research assistant acted as experiment instructor on the back seat on the driver's side. She was there to administer a cognitive secondary task on a different road section, the results of which are reported in Lehtonen, Lappi & Summala (2011). During the recording segments from which data is reported here the driver and the other persons in the car avoided any interaction. The participants drove the test route at their own pace. The driver was simply instructed to (i) drive as they normally would, but (ii) observe traffic laws and safety - in particular, they were explicitly instructed not to cut onto the lane of oncoming traffic in left-hand turns, if this was what they would do in normal driving. (This was deemed a necessary precaution because of the relatively high speed limit - 80 km/h - and the fact that many of the turns had a blind entry).

### Equipment and data preparation

The instrumented car was a model year 2007 Toyota Corolla 1.6 compact sedan with a manual transmission. The passenger side was equipped with brake pedals and extra mirrors for the driving instructor, as well as a computer display that allowed him to monitor vehicle speed, as well as the operation of the eye-tracker and the data-logging systems. The car was equipped with a two-camera (Smart Eye Pro versions 5.1 and 5.5 [www.smarteye.se](http://www.smarteye.se)) eye tracker operating at 60 Hz, a forward looking VGA scene camera, a GPS-receiver, as well as a forward looking infrared rangefinder (IBEO, [www.ibeo-as.com](http://www.ibeo-as.com)). Vehicle speed, the vehicle control signals (steering, throttle and brakes), as well as vehicle yaw rate were recorded directly from CAN-bus (all oversampled at 100 Hz). All signals were synchronized and time stamped on-line, and stored on a computer running custom MATLAB software, located in the rear luggage compartment. All subsequent data preparation, visualization and analysis was done with custom-made Python scripts, except for the final statistical analyses which were done with SPSS 18.

The data was segmented based on the time stamps corresponding to the GPS coordinates of the test route. To render different trials (drives) comparable, the data was given a distance-based representation. One trial, with no traffic or other "incidents" was chosen as a reference. The vehicle trajectory in an allocentric xy plane was computed by interpolating the GPS signal. This interpolated trajectory would then be used as the basis of a route-distance value. All participants' trials were then mapped onto this frame of reference, by first best-matching the observed GPS values to the reference trajectory, and then associating the rest of the data, with time-stamps matching the relevant GPS location.

The trajectory of the vehicle was also computed for individual trials by starting from point of gaze observation on the trajectory, and integrating the vehicle yaw rate and speed over time. This was used to estimate the point of gaze landing on the future trajectory, by finding first the point on the trajectory where the eccentricity of a point on the path-integrated trajectory corresponded with the visual eccentricity of gaze at the initial point - i.e. the point where the

line of sight would intersect the path-integrated trajectory. This gaze-landing point could then be assigned a route distance value based on the mapping of time stamps to route distances<sup>1</sup>

Tangent points were manually identified from still video frames from SmartEye's Scene camera (5 Hz frame rate), and the image coordinates of the mouse pointer and the eye-tracker coordinates were physically calibrated, using a calibration grid visible in the scene camera and the infrared rangefinder (whose coordinates were used as the native coordinate system for the car). The scene images were then associated with the rest of the data based on the time stamp of the video frame. If there was oncoming traffic in a turn (any vehicles or pedestrians visible in the forward-looking video camera), data for that turn was excluded from gaze-direction analyses (but included in the driving speed analyses). This was done in order to eliminate the effect of these potentially confounding visual targets in the road scene.

### Results

Four turns from the test route were selected for detailed analysis. The turns were chosen so that we would have two pairs of roughly similar turns, this way we could check whether any pattern of visual behaviour seen in a turn would also be seen in the other, similar, turn. The analysed turns comprised of two long left hand turns (hereafter denoted by T1 & T4), and two blind right-left sequences (T2/3 & T5/6)<sup>2</sup>. The blind sequences T2/3 and T5/6 probably resembled the roads used in Land and Lee (1994), while the faster turns T1 and T3, with a sighted approach phase, were probably more similar to the curves in Kandil et al. (2010).

#### Driver Gaze Behaviour in Relation to the Tangent Point (Tangent Point Orientation)

We first set out to replicate the commonly observed tangent point orientation. Based on the previous research reviewed above, we expected the drivers to direct their gaze towards the tangent point region, especially during the approach and

<sup>1</sup> Note that the gaze-landing point, as defined here, may not in all cases perfectly coincide with the driver's true gaze target, because it does not use pitch-information and effectively projects the trajectory and gaze directions onto a two-dimensional plane; it does, however, provide a good estimate when the road is flat - as in the present experiment. The reason we did not use the pitch information is the difficulty of measuring it reliably in the noisy environment.

<sup>2</sup> We refer to a turn as blind when both of the following conditions are met: (i) At no point during the approach phase - i.e. before the driver steers into the turn - is the exit of the turn or the exit of the entire sequence in the case of connected curves visible to the driver, and (ii) During the entire approach phase, the occlusion point falls within some angular threshold of the tangent point (the threshold used here was 10°). The *occlusion point* is defined the furthestmost part of the desired trajectory to which a continuous line of passage is visible, i.e. the point on the road where the driving line first disappears from view (Fig. 1, see Lehtonen, Lappi & Summala, 2011). Turns that are not blind by these criteria are said to be sighted.

turn entry phases. As table 2 shows, we could demonstrate a consistent pattern of tangent point orientation during curve approach and turn entry. This establishes for the curves analysed the basic pattern observed in many previous studies, which have reported drivers to direct their gaze into the tangent point region 60-70% of the time, when entering a bend (Land & Lee, 1994; Kandil et al., 2010).

Table 2: Estimated relative frequency of tangent point oriented gazes (% of all valid observations, mean and standard deviation) falling within three degrees horizontal of the tangent point during turn approach (before the driver turns the wheel) and entry into the turn (after the driver turns the vehicle). In connected curve sequences only the first turn of the sequence for which an approach phase can be defined is listed.

	<i>Approach</i>	<i>Entry</i>
T1	52 (23)	41 (28)
T2/3	72 (24)	67 (22)
T4	50 (25)	58 (24)
T5/6	71 (25)	60 (22)

### Tangent Point and Gaze Behavior in Relation to the Vehicle Frame of Reference

We next set out to investigate further the behaviour of the tangent point and the driver's gaze in vehicle coordinates. The Land & Lee (1994) model which predicts that the drivers steer so as to maintain the tangent point at a constant bearing angle. The Wann and Land (2000) model, in turn, predicts that tangent point eccentricity should decrease during turn entry (from a relatively eccentric value) when the driver begins to rotate the vehicle and to orient it towards a reference point (close to the tangent point) which he wishes to travel through.

Neither pattern was observed. Instead, as illustrated in figure 2, the observed horizontal positions of gaze and the eccentricity of the tangent point become progressively *more* eccentric during the turn-in phase of this right-left turn. The first vertical green line indicates the average point where the vehicle rotation rate exceeds one degree per second, the second indicates the point where the vehicle reaches its maximum rotation rate (the driver begins to turn the wheel back to the left) and the third green line indicates the point of maximum rotation-rate in the left hand turn (after which the driver begins to reach a maximum at towards the end of the turn-in phase).

Table 3 gives the values for tangent point displacement from the turn-in to the end of the entry phase. Although during the entry phase gaze is relatively concentrated in the TP region, the drivers nevertheless do not appear to be steering to compensate for the outward movement of the tangent point (and gaze) due to visual flow.

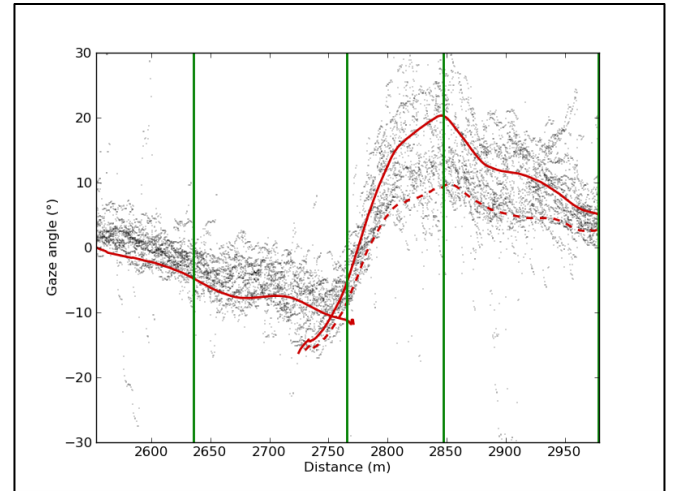


Figure 2: Tangent point and gaze in vehicle coordinates for the right-left turn sequence 5/6. The solid red line indicates the eccentricity of the tangent point (averaged across subjects), and the dotted line indicates the road centerline. Right is negative. Note that the tangent point (and gaze) are neither constant, nor do they sweep towards the vehicle centerline (zero eccentricity).

Table 3: Average displacement of the tangent point in the entry phase of each turn. Negative is to the right.

	N	mean
T1 entry	16	3.58
T2 entry	17	-8.68
T3 entry	16	25.1
T4 entry	16	0.66
T5 entry	15	-6.4
T6 entry	15	25.5

We could see the tangent point eccentricity (and gaze) was not constant through the turns but instead increased steadily, in step with the increase in vehicle yaw-rate as the car was entering the curve. As can be seen in Figure 2, the highest tangent point and gaze eccentricities appear occur systematically in the steepest parts of the turn (the green vertical lines indicating the parts of the trajectory where the yaw-rate values reach their maxima). This suggested that a possible relation might exist between the driver's gaze and the vehicle yaw-rate. We decided to investigate this relationship by looking at the estimated fixation density on the road, i.e. the distribution of gaze landing points where the line of sight would intersect the vehicle trajectory.

The relative frequency of gaze landings in each 10 m road segment through each curve was computed as the percentage of all gaze landings falling within the segment. This was computed first individually, and then averaged across

subjects. This distribution is shown by the histogram in figure 3 (bottom), and shows that gaze is not equally distributed on the road surface. This would happen if the drivers were, for example, fixating a point on their trajectory that would be some constant distance ahead. Instead, as shown in the top half of figure 3, while they are approaching a turn the drivers are looking further ahead – 100-150 meters up the road in this instance – but as they enter the turn, the gaze falls closer to the car. Also, when in a connected sequence of turns, the gaze moves discontinuously on the road. Here, the gaze is seen to jump further up the road (from right to left) just before the vehicle yaw rate has reached its peak and the left-hand part of the turn begins.

An unexpected pattern became evident when the relative frequency of estimated gaze landings in a specific part of the curve was compared with yaw rate at that part of the road (histogram in fig.3, bottom shows data for one curve). Gaze landings appeared to be concentrating those parts of the road where the vehicle yaw rate (and thus the road curvature as well) would be highest. I.e. there is a high correlation between the vehicle yaw rate (local road curvature) and the frequency of gaze landings. This correlation is shown graphically in figure 4, and the numerical values of the correlations are given in Table 4, showing that drivers preferentially look at segments of the road where they are going experience a high yaw rate, i.e. parts of the road with high curvature.

Table 4: Correlation (Spearman's Rho) between the median relative frequency (within a turn) with which a particular 10m road estimated as a potential gaze target, and the measured average yaw rate in that segment.

	n	Spearman's Rho
T1	41	0.811
T2/3	46	0.565
T4	53	0.717
T5/6	43	0.795

## Discussion

There are many different steering models available and all predict the same qualitative pattern of tangent point orientation (i.e. orienting gaze in the general direction of the tangent point, the TP AOI), but different predictions concerning the details of the gaze distribution pattern.

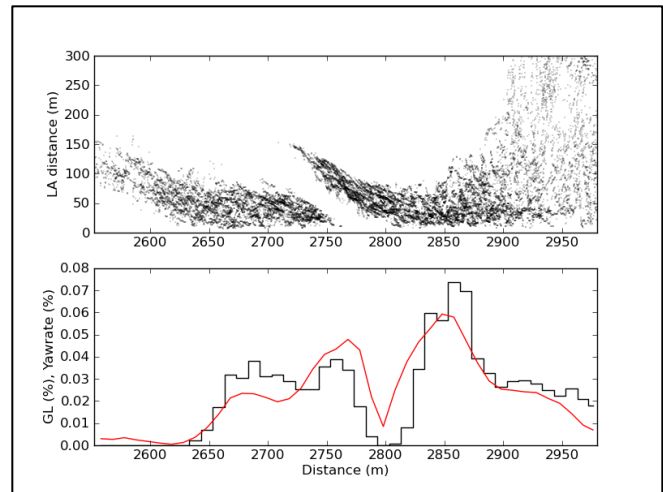


Figure 3: Horizontal axis location on the road (m from beginning of the route) Top: Estimated look-ahead distances through T5/6, i.e. distance of estimated gaze landing point from the current location, measured along the future path Bottom: Absolute value of vehicle yaw rate (continuous line) and the frequency histogram of gaze landings on each part of the curve (both variables scaled to sum to 100% within the analysed road section).

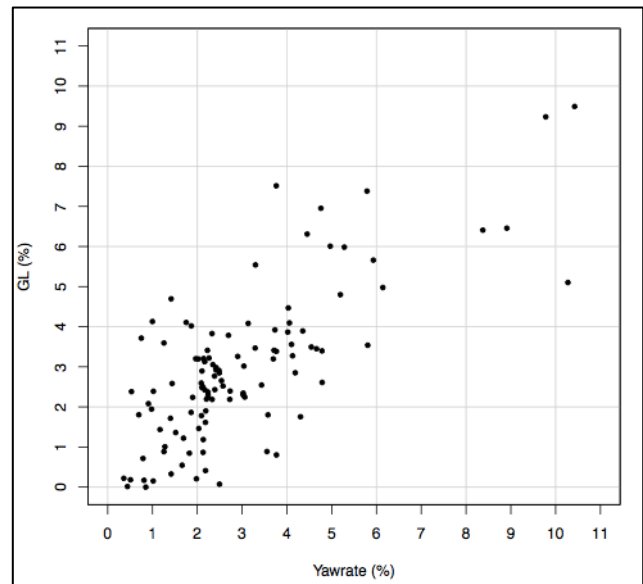


Figure 4: Top: Scatterplot of vehicle yaw rate and gaze landing density in the analysed turns. Each datapoint represents the across-subjects average for one 10m segment of all the curves analysed.

One model (Land & Lee 1994 / Wann & Land, 2000) would predict a constant value for gaze and tangent point eccentricity throughout a turn since, according to the model, the driver would make compensating steering movements (induce vehicle yaw) to cancel out the apparent horizontal motion of the tangent point due to optic flow. The prediction of the Wann and Land (2000) / Wann & Wilkie (2004) visual sweep model is that tangent point - and gaze - eccentricity should decrease during turn entry as the driver rotates the vehicle towards the reference point on the future path near the tangent point. Instead, we find that the tangent point - and gaze - become progressively *more* eccentric during the turn-in phase, without compensation or any apparent adverse effects to steering. We therefore conclude that these two steering models *which are based on the idea of using the tangent point as a visual target point which acts control parameter in set-point feedback loop* do not offer a general account for gaze/ steering behavior, at least on the kinds of turns analysed here (variable-radius curves on rural roads).

The differences of the remaining models - not based on simple negative feedback but on an *active trajectory planning strategy because they involve a predictive model of future path* – cannot be judged on the present data. It is thus not possible to pick one that would be the clear best fit for our data. However, the surprising finding that gaze was heavily concentrated on a few parts of the turn – namely, the locations where the vehicle achieved highest yaw-rate – offers some interesting possibilities. *Prima facie* none of the steering models appear to predict this high correlation between the yaw rate and probability of being selected as a gaze target, but we would consider that this pattern may be best interpreted as complementing the Wann & Swapp (2000) model or the Boer (1996) model. Both models predict that the driver should fixate *some* part of the road which he wishes to pass through, but neither specifies *which* part of the future trajectory the driver would or should look at. The location of highest anticipated trajectory curvature (max. yaw-dot) could perhaps serve as such a salient point of reference. (Note that is also behaviourally meaningful in terms of the sequencing of the driving task: this is the point where the curve/trajectory “opens up”, and the drive may begin to accelerate and unwind the steering).

There clearly exists a need for new and such accurate data on visual behaviour in curve driving – i.e. data that could speak the issue of which model or models offer the best fit to driver behavior. This study takes one steps in this direction, by presenting data that speaks directly to the *differential* predictions the models make beyond the *common* prediction of tangent point orientation (which was observed in all turns analysed). We also observed a surprising correlation between gaze landings on the road and measured vehicle rotation at that point, not predicted by any of the existing models.

It is suggested that the next generation of models be developed and empirically tested based on such detailed

quantitative basis - not only for the qualitative pattern of tangent point orientation.

## Acknowledgments

We are grateful to Mr. Juha Vepsäläinen for programming the annotation tool, and Mr. Henri Kotkanen and Mr. Harri Hiltunen for assistance in data collection.

## References

- Ballard D, M Hayhoe & J Pelz (1995). Memory representations in natural tasks. *J cogn neurosci* 7: 66-80.
- Boer E R (1996) Tangent point oriented curve negotiation. Intelligent Vehicles Symposium, 1996, Proceedings of the 1996 IEEE: 7-12.
- Chattington M, M Wilson, et al. (2007) Eye-steering coordination in natural driving. *Exp Brain Res*,180: 1-14.
- Kandil F I, A Rotter, et al. (2009) Driving is smoother and more stable when using the tangent point. *J Vis* 9: 1-11. <http://www.journalofvision.org/content/9/1/11> doi:10.1167/9.1.11
- Kandil F I, A Rotter, et al. (2010) Car drivers attend to different gaze targets when negotiating closed vs . open bends. *J Vis* 10: 1-11.
- Kowler, E. (2011). Eye movements: the past 25 years. *Vis Res* 51: 1457-1483.
- Land, M F (2006) Eye movements and the control of actions in everyday life. *Prog Ret Eye Res* 25: 296-324.
- Land M F (2007) Fixation Strategies During Active Behaviour: A Brief History. In: R P G Gompel, M H Fishcer, W S Murray and R L Hill (eds) *Eye Movements: A Window on Mind and Brain*. Elsevier. Pp. 77-95.
- Land M F and D N Lee (1994) Where we look when we steer. *Nature* 369: 742-744.
- Land M F and B W Tatler (2001) Steering with the head: The visual strategy of a racing driver. *Curr Biol* 11: 1215-1220.
- Lehtonen E, Lappi O & Summala H (2011) Anticipatory eye movements when approaching a curve on a rural road depend on working memory load. *Trans Res Part F Traffic Psychol* (2011).
- Underwood G, P Chapman, et al. (1999) The visual control of steering and driving: Where do we look when negotiating curves? In: A.G.Gale, I.D.Brown, C.M.Haslegrave and S.P.Taylor (eds) *Vision in Vehicles - VII*. Amsterdam, Elsevier.
- Wann J and M Land (2000) Steering with or without the flow: is the retrieval of heading necessary? *Trends Cogn Sci* 4: 319-324.
- Wann J P and D K Swapp (2000) Why you should look where you are going. *Nat Neuroci* 3: 647.
- Wann J P and R M Wilkie (2004) How do we control high speed steering? In: L M Vaina, S A Beardsley, S K Rushton (eds.): *Optic flow and beyond*. Kluwer Academic Publishers. Norwell, MA, USA. pp. 371-389.



# Arbitrary Category Labels Can Change Similarity Judgments of Human Faces

Frankie Lara (frankie.lara@jaguar.tamu.edu)

Amanda Hahn (achahn30@gmail.com)

Na-Yung Yu (nayungyu@gmail.com)

Takashi Yamauchi (takashi-yamauchi@tamu.edu)

Department of Psychology, Mail Stop 4235

Texas A&M University, College Station, TX 77843 USA

## Abstract

In two experiments, participants were presented with a triad of morphed White and Hispanic faces paired with pseudoword labels. The meanings of these labels were manipulated to represent categorical information about the face. Labels were said to represent either the person's belief, the food s/he ate, the disease s/he had, or the person's last name. The results indicated that categorical information affects our judgments of faces. Information categories such as belief, food, and diseases were particularly strong in modifying the participants' similarity judgment of faces, whereas information characterized with last names of faces were least powerful. Previous research focuses on race face perception being affected primarily by racial indicators or racial information. Our results provide that how we perceptually analyze faces is not confined to obvious racial cues, but by non-racial semantic information as well, suggesting that category-relevant information by itself provides a strong basis for inductive generalization.

**Keywords:** Labeling; Similarity; Categorization

## Introduction

In the perception of faces, there is a tendency to pay excessive attention to salient features such as race-specific (Eberhardt, Dasgupta, & Banaszynski, 2003; Levin, 2000; MacLin & Malpass, 2003). When we see racially ambiguous faces, we shift our attention to features that signal ethnicity such as hair-style or skin color and ignore other important information. This attention shift often results in undesired psychological effects such as cross-race face recognition deficit, i.e., faces that are categorized into "other race" are recognized less than the faces that are categorized into one's own race. (MacLin and Malpass, 2003) and erroneous impression formation (Kashima, 2000; Hamilton & Sherman, 1994; Macrae, Milne, & Bodenhausen, 1994).

Additional studies show that race-based categorization modifies perception of skin color and the width of faces and mouth (MacLin & Malpass, 2003). Research further suggests that holistic face-processing is more prevalent in same-race faces than in other-race faces (Michel, Rossion, Han, Chung, & Caldara, 2006b).

Social categorization based on in-group and out-group of an observer also yield a cross-race recognition deficit, implying that categorization itself can be a mediating factor changing face perception (Bernstein, Young, Hugenberg, 2007). As long as stimuli are grouped in a meaningful way, some modification in face perception is likely to occur. For

example, incremental training or labeling of faces into arbitrary categories generates a categorical perception effect, in which faces taken across a category boundary are recognized better than faces taken within the category boundary (Kikutani, Roberson, & Hanley, 2008). Taken together, these studies demonstrate that not just racial information per se, but category information plays a substantial role in modifying perception of faces.

This explains why social categorization often accompanies faulty generalization and stereotyping. Categories are generative in nature. Categorical labeling not only accentuates features that are central to the category (e.g., prototypical features), but also help generate new features by means of explanations and justifications (Kunda, Miller, & Claire, 1990). When confronted with contradictory attributes (e.g., a rich African-American businessperson), people often make up a subtype of the category (e.g., black entrepreneur) and use it to preserve their initial stereotypical belief (Macrae, Stangor, & Hewstone, 1996). Combinations of contradictory concepts such as "Harvard-educated carpenter" create new features such as "being rebellious" or "anti-social," which were not part of each separate concept – "Harvard-educated" and "carpenter" (Kunda, Miller, & Claire, 1990). When a person is characterized categorically ("Linda is a feminist" as opposed to "Linda believes in and support feminism"), people not only think that the person possesses prototypical attributes of the category ("Linda majored in philosophy in college") but also some unrelated features are deemed likely ("Linda likes Chinese food") (Yamauchi, 2005, 2008, 2009). Categorization also helps reframe people's attention. When geometric stimuli are grouped by categories, perceptual sensitivity within the category is reduced while the differences between categories are enhanced (Tajfel & Wilkes, 1963; Yamauchi & Yu, 2008; Yamauchi et al., 2002). On this basis, we think that when labels help form categories, labeling can modify people's perceived facial similarities. The two experiments tested this idea.

## Overview of the experiments

We employed a widely used triad task (Gelman & Markman, 1986; Sloutsky & Fisher, 2004; Yamauchi Markman, 2000; Yamauchi, Kohn, & Yu, 2007; Yamauchi & Yu, 2008; Yu et al., 2008, 2010; Waxman & Booth, 2001) in which we attached pseudoword labels (e.g., "Scrakies") to face pictures and examined how these labels

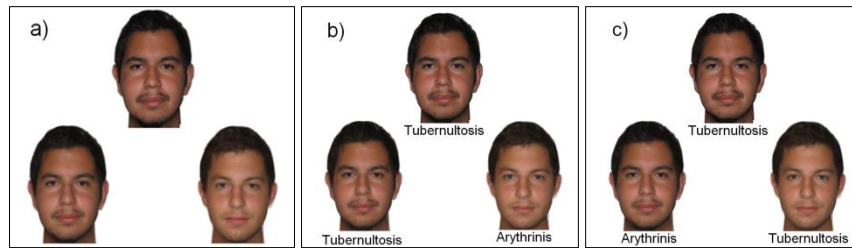


Figure 1: Otherwise identical, triads of faces shown (a) without labels, (b) with labels (different-label condition; Experiment 1), and (c) with labels (same-label condition; Experiment 2). Here, the dissimilar base picture appears on the right.

influenced participants' judgments of facial similarity when the same arbitrary labels represented different kinds of semantic information.

In our experiments, participants were shown a triad of human faces (Figure 1) and judged which face, bottom left or right, was more similar to the target face on the top. The target face (top face) was an original face that was either Hispanic or Caucasian. The bottom left and right faces were morphed faces. These faces were displayed either without a label (the control condition, Figure 1a) or with a label (the label condition, Figure 1b & 1c). Each label was an arbitrary pseudoword (such as "Scrakies"), but the meaning of the label was manipulated in the instructions that the participants read at the beginning of the experiment.

In the three experimental conditions, participants were told the arbitrary labels either represented the name of the food the person regularly eats (the food-label condition), the name of the disease that he has (the disease-label condition), or the name of belief that he follows (the belief-label condition). Note that these manipulations were introduced only in the instructions participants received and all participants received the same stimuli. These three conditions were contrasted to two control conditions, in which the labels were removed entirely from the stimulus frame (no-label condition, Figure 1a) or the arbitrary labels were characterized as the last name of the person (the last-name-label condition).

We measured the proportion of trials that participants chose the face that was physically dissimilar to the target image. For example, in Figure 1a, the left base image looks more similar to the target image when compared to the right base image. We measured the proportion of participants selecting the dissimilar face pictures (the right base image in Figure 1) when face pictures had no labels (Figure 1a), and when the target and dissimilar face pictures had different labels (Figure 1b – Experiment 1) or the same labels (Figure 1c – Experiment 2).

We predicted category information would affect similarity judgments while indexical labels would not to the same degree. We commonly classify people by their habit of eating (food-labels, e.g., vegetarians, ethnic-food lovers), the disease they have (disease-labels, e.g., people with high blood pressure, people with cancer, people with allergies) or the belief that people follow (belief-labels, e.g., Christians, positive thinkers). By categorizing people in this manner,

we obtain a sense of similarity and unity among category members. In contrast, a last name is indexical; it refers to a specific person. Groups, such as families, can be formed by last names, but they do not give us a sense of coherence. A name points to a specific entity within a category and therefore should no influence judgments of similarity as strongly as categorical labels. Thus, our hypothesis that categorical labeling helps modify perception of facial similarity leads to the prediction that arbitrary labels change the perception of face similarity in the food-label, disease-label, and belief-label conditions, but not in the last-name label condition.

When items do not share category membership, differences are emphasized. On the contrary, when items are in the same category, differences are diluted (Sloutsky, 2003; Waxman & Markow, 1995). Therefore, in Experiment 1, when the dissimilar picture does not share a label with the target, the proportion of participants selecting the dissimilar base picture should be considerably smaller in the food-label, disease-label, and belief-label conditions than the no-label condition. In Experiment 2, the dissimilar base picture shares a label with the target, so the proportion of participants selecting the dissimilar base picture should be larger than the no-label condition. The proportion of participants selecting dissimilar face pictures as "similar" should be indistinguishable between the no-label and last-name-label conditions in both Experiments

## Experiment 1

### Method

**Participants** A total of 191 undergraduate students participated in this experiment for course credit. They were randomly assigned to one of five conditions: no-label ( $n=39$ ), belief-label ( $n=39$ ), food-label ( $n=35$ ), disease-label ( $n=34$ ), and last-name-label ( $n=34$ ) conditions.

**Materials** Stimuli were triads of faces that either had no label or label attached to them (Figures 1a and 1b). The target was an original picture of either a Hispanic or Caucasian face, and the two base pictures were a morph of the original Hispanic and Caucasian face (Figure 2).

In total, we photographed five pairs of original Hispanic and Caucasian faces. All expressions are neutral and no faces contained any distinguishing features, e.g., none had moles or mustaches. These photographs were morphed into

five pairs of 20 images using Morph Man 4.0 (2003) software starting from the original Hispanic face and morphing towards the Caucasian face (Figure 2). Altogether there were 100 images (10 original faces and 90 morphed images) that had varying degrees of Hispanic and Caucasian facial features.



Figure 2. One real Hispanic face (far left) is morphed gradually with one real Caucasian face (far right). In the actual experiment, there were 18 morphed images between the two original faces.

From the 90 morphed images, base pictures were selected controlling for physical differences between stimuli. Specifically we developed three levels of physical difference— low, medium and high physical difference within conditions – based on the degree of merging two of the original face pairs. In the low physical difference

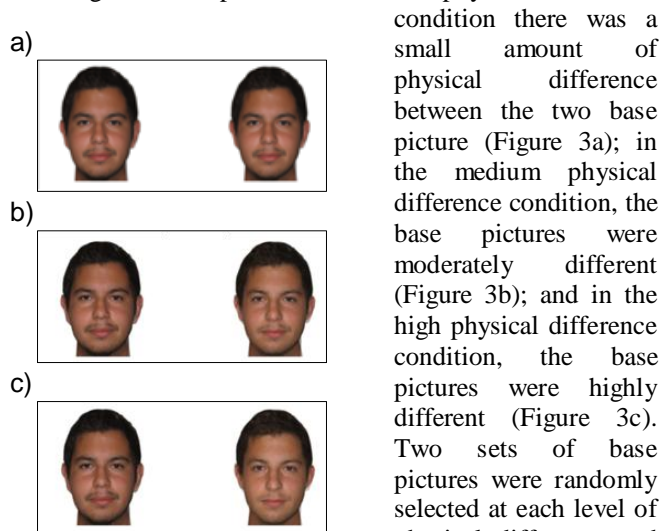


Figure 3: The three levels of physical difference— low (a), medium (b), and high (c) used in the similarity judgment task. In this example, the dissimilar picture is seen on the right.

**Procedure** Participants were shown 60 triads of pictures, one at a time, and judged which base picture within the triad was more similar to the target image by pressing either the right or left arrow key on the keyboard. The order of presenting the stimuli and the left-right location of placing the dissimilar base pictures were determined randomly. The experiment took approximately 15 minutes to complete.

**Design** The experiment had a 3 (Physical Difference; low, medium, and high; within-subjects factor)  $\times$  5 (Label Condition; belief-label, food-label, disease-label, last-name-label and no-label; between-subjects factor) mixed design. All participants in the five conditions (no-label, food-label, disease-label, belief-label, and last-name-label conditions) received the identical stimuli. The labels in each condition were physically the same, but the meaning attached to the labels was altered in the instructions.

## Results

Figure 4 summarizes the results in Experiment 1. There was a significant main effect of label condition:  $F(4, 176) = 4.49$ ,  $MSE = .03$ ,  $p = .002$ ,  $\eta^2 = .09$ . Individually, the belief-label ( $M = .10$ ), food label ( $M = .09$ ), and disease label ( $M = .11$ ) demonstrated a significant effect when compared to the no label condition ( $M = .18$ ): belief label vs. no label,  $t(76) = 3.76$ ,  $SE = .02$ ,  $p < .001$ ,  $d = .85$ ; food label vs. no label,  $t(72) = 5.66$ ,  $SE = .02$ ,  $p < .001$ ,  $d = 1.32$ ; and disease label vs. no label:  $t(71) = 4.13$ ,  $SE = .02$ ,  $p < .001$ ,  $d = .97$ . The last name label condition ( $M = .13$ ), however, showed no

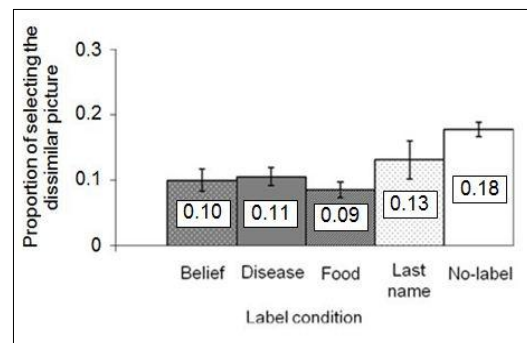


Figure 4. The proportion of participants selecting the dissimilar image of the base pair pictures according to label type in Experiment 1. The error bars represent two SE units calculated from each condition

significant difference when compared to the no label condition ( $M = .18$ ),  $t(71) = 1.59$ ,  $SE = .03$ ,  $p > .10$ ,  $d = .37$ .

Physical similarity played a role in judgments. Labels influenced similarity judgments of faces more in the low physical difference condition than the medium and high physical difference conditions. In the low physical difference condition, all labels produced a significant effect compared to the no-label condition,  $t_s > 2.33$ ,  $ps < .01$ ,  $d_s > .60$ . In the medium physical difference condition, the food label and disease labels produced significant effects when compared to the no label condition:  $t_s > 2.33$ ,  $ps < .01$ , while the belief and last name conditions did not,  $t_s < 2.33$ ,  $ps > .01$ . In the high physical difference condition, the disease condition was significantly different compared to the no label condition,  $t(65) = 3.16$ ,  $SE = .06$ ,  $p = .002$ ,  $d = .77$ , while other conditions were not,  $t_s < 2.33$ ,  $ps > .01$ .

## Discussion

Consistent with our hypothesis, labels attached to face pictures modified participants' judgments of similarity considerably. When the target and dissimilar face pictures had the different labels, the proportion of participants selecting the dissimilar face pictures decreased dramatically. The impact of the labels was particularly pronounced when these labels conveyed some categorical information, such as the types of food, disease, or belief that the people eat, have, or follow. When the labels represented the last names of the people, the effect of labels diminished considerably, supporting the view that the distortion of race perception occurs especially when labels are associated with categorical information.

Because such an effect was present primarily when labels conveyed categorical information, we suggest that the distortion of face perception is linked to the general mechanism of categorical perception. Experiment 2 tested this idea further.

## Experiment 2

The results from Experiment 1 suggest that labels attached to face pictures can modify people's perception of similarity. In Experiment 1, all dissimilar face pictures carried the different labels as the target picture (Figure 1b); as a result, the difference between the two face pictures (target and dissimilar base pictures) was exaggerated considerably. If, as hypothesized, the categorical labels attached to the face pictures are indeed responsible for the modified perception faces, then the labels can also create the perception of "sameness." In other words, if the dissimilar face pictures carry the same label as the target picture then the dissimilar face pictures should be perceived as more *similar* to the target picture. This was tested in Experiment 2. The only difference between Experiment 1 and 2 was the assignment of the labels. In Experiment 2, the labels of the base pictures were simply swapped so that the dissimilar face pictures and the target picture had the same label (Figure 1c). In Experiment 2, when compared to the no-label condition, the proportion of participants selecting the dissimilar face pictures should increase considerably when the dissimilar face pictures and the target picture have the same label. This phenomenon should occur primarily in the belief-label, food-label and disease-label conditions, but not when in the last-name condition.

## Method

**Participants** A total of 182 undergraduate students participated in this experiment for course credit. They were randomly assigned to one of five conditions: no-label ( $n=34$ ), belief-label ( $n=40$ ), food-label ( $n=38$ ), disease-label ( $n=33$ ), and last-name-label ( $n=37$ ) conditions.

**Materials and Procedure** The materials and procedure used in Experiment 2 were identical to those described in Experiment 1.

**Design** The design of Experiment 2 was the same as Experiment 1 except that the target face in each slide shared the same label as its least similar base face (Figure 1c).

## Results

Consistent with our hypothesis, our results show that category information, even though they were only indirectly related to race, can affect judgment of Hispanic and White faces.

Figure 5 summarizes the results of Experiment 2. There was a significant main effect of label condition:  $F(4, 177) = 4.50$ ,  $MSE = .16$ ,  $p = .002$ ,  $\eta^2 = .09$ . Specifically, the belief-label condition ( $M = .36$ ), the food-label condition ( $M = .33$ ) and the disease-label condition ( $M = .32$ ) demonstrated significant effects when compared to the no-label condition

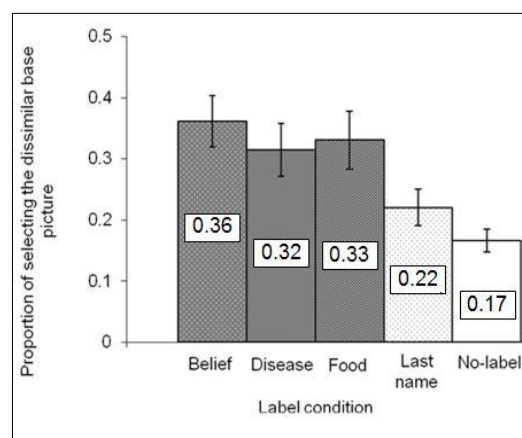


Figure 5. The proportion of participants selecting the dissimilar image of the base pair pictures according to label type in Experiment 2. The error bars represent two SE units calculated from each condition

( $M = .17$ ): belief-label vs. no-label,  $t(72) = 4.05$ ,  $SE = .05$ ,  $p < .001$ ,  $d = .95$ ; food-label vs. no-label,  $t(70) = 3.18$ ,  $SE = .05$ ,  $p < .005$ ,  $d = .75$ ; disease-label vs. no-label,  $t(65) = 3.11$ ,  $SE = .05$ ,  $p < .005$ ,  $d = .76$ . The last name label ( $M = .22$ ), however, showed no significance when compared to the no label condition ( $M = .17$ );  $t(69) = 1.84$ ,  $SE = .03$ ,  $p = .07$ ,  $d = .44$ . There was no interaction effect between the label condition and the physical difference:  $F(8, 354) = .74$ ,  $MSE = .01$ ,  $p = .66$ ,  $\eta^2 = .02$ .

The belief, food, and disease conditions all showed significance at each level of physical difference when compared to the no label condition:  $ts > 2.33$ ,  $ps < .01$ ,  $ds > .60$ . Just as in the between subject analysis, the last-name label produced null effects across all levels of physical difference:  $ts < 2.00$ ,  $ps > .05$ .

## Discussion

As in Experiment 1, labels attached to the face pictures in Experiment 2 modified participants' judgments of similarity considerably. Again, the impact of labels was particularly

pronounced when the labels conveyed certain categorical information such as the types of food, belief and disease that people eat, follow, and have. The effect of labels was reduced dramatically when the labels were associated with the names of people. In Experiment 2, the target and dissimilar face pictures had the same labels (Figure 1c). As a result, the proportion of participants selecting the dissimilar face pictures as similar to target pictures increased dramatically. These results indicate that the effect of the categorical labels is bi-directional. Categorical labels can create a sense of difference and a sense of proximity. These changes occurred primarily when labels were associated with the types of food, disease, and belief that people have (or eat) but not when labels were associated with the last names of the people.

Together, these experiments further support the view that the meaning attached to these labels, not labels themselves, modifies our perception of similarity both positively and negatively by enhancing the sense of similarity and difference depending on whether stimuli carry the same or different labels.

## General Discussion

Our results indicate that categorical information influences the participants' judgments of similarity. In Experiment 1, participants chose the dissimilar face significantly *less often* when the target and dissimilar face pictures had different labels. In Experiment 2, participants chose the dissimilar faces *more often* when the target and dissimilar face pictures had the same labels. The impact of the labels was negligible when the labels were associated with the last names of people. These results suggest that categorical information given to these labels were indeed responsible for the modified perception of similarity. This modified perception likely arose from some general mechanism underlying the categorical perception effect (Goldstone, 1994, 1995; Livingston, Andrews, & Harnad, 1998; Newell & Bulthoff, 2002; Roberson, Davies, & Davidoff, 2000).

Previous research has focused on race-specific cues to distort racial perception (Eberhardt, Dasgupta, & Banaszynski, 2003). Other research deliberately uses racially related facial features to distort race perception (MacLin & Malpass, 2003). Our study, which uses both White and Hispanic faces, extends previous results by demonstrating that perceptions of race-oriented faces can be distorted by attributing information that is not directly related to racial cues. Meaningful categorical labels can create a sense of similarity and difference depending on whether two stimuli have the same or different labels. Some category information seems to have greater impact than others. For example, the belief labels and the food labels were relatively stronger than the disease and last-name labels. Although the reason behind these differences cannot be determined based on these experiments, we speculate that some categories of information held stronger weight because of their behavior and preference implications

Gil Diesendruck and Heidi HaLevi (2006) point out that personality traits are the primary means by which categorical distinctions and inferences are made because these traits explain behavior and preferences (Yuill, 1992). Food-labels, disease-labels, and belief-labels explicitly refer to such behaviors and preferences, while last-name-labels do not suggest that two people will behave in the same manner.

This reaction to labels and assumptions based on these traits may be related to naïve theories that people form in everyday situations (Chao, Chen, Roisman, & Hong, 2007; Gelman, 2003). It is suggested that people tend to assume there is an essence underlying observed physical characteristics of people, animals, and things (Ahn, 2001; Chao et al., 2007; Gelman, 2003; Medin & Ortony, 1989; Murphy & Medin, 1985). Such essence can be biological characteristics (diseases or DNA, Medin & Atran, 2004), core beliefs (e.g., religion, Cairns, Jenworthy, Campbell & Hewstone, 2007), or behavioral habits (Gelman & Heyman, 1999). The results of our experiments demonstrated how easily people construct naïve theory and how powerful the influence of the naïve theory is. The simple label-meaning manipulation used on our experiments was powerful enough to alter their perception of people. Our results, combined with previous research, suggest categorical information is important when makes judgments about people.

## References

- Ahn, W. (2001). Dissociation between categorization and similarity judgment; differential effect of causal status on feature weights. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization* (pp. 87-107). New York: Oxford University Press.
- Bernstein, M. J., Young, S. G., & Hugenberg, K. (2007). The Cross-Category Effect: Mere social categorization is sufficient to elicit an own-group bias in face recognition. *Psychological Science*, 18, 709-712.
- Cairns, E., Jenworthy, J., Campbell, A., & Hewstone, M. (2007). The role of in-group identification, religious group membership and intergroup conflict in moderating in-group and out-group affect. *British Journal of Social Psychology*, 45(4), 701-716.
- Chao, M. M., Chen, J., Roisman, G., & Hong, Y.-y. (2007). Essentializing race: Implications for bicultural individuals' cognition and physiological reactivity. *Psychological Science*, 18, 341-348.
- Diesendruck, G., HaLevi, H., (2006). The Role of Language, Appearance, and Culture in Children's Social Category-Based Induction. *Child Development*, 77(3), 539-553.
- Eberhardt, J., Dasgupta, N., & Banaszynski, T. L. (2003). Believing is seeing: The effects of racial labels and implicit beliefs on face perception. *Personality and Social Psychology Bulletin*, 29, 360-370.
- Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. New York: Oxford University Press.



- Gelman, S. A., & Heyman, G. D. (1999). Carrot-eaters and creature-believers: The effects of Lexicalization on children's inferences about social categories. *Psychological Science*, 10(6), 489-493.
- Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, 23, 183-209.
- Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123(2), 178-200.
- Goldstone, R. L. (1995). Effects of categorization on color perception. *Psychological Science*, 6(5), 298-304.
- Hamilton, D. L., & Sherman, J. W. (1994). Stereotypes. In R. S. Wyer, Jr., & T. K. Srull (Eds.), *Handbook of social cognition* (2nd ed., Vol. 2, pp. 1-68). Hillsdale, NJ: Erlbaum.
- Kashima, Y. (2000). Maintaining cultural stereotypes in the serial reproduction of narratives. *Personality and Social Psychology Bulletin*, 26, 594-604.
- Kikutani, M., Roberson, D. & Hanley, J.R. (2008) What's in the name? Categorical Perception of unfamiliar faces can occur through labeling. *Psychonomic Bulletin & Review*, 15, 787-794
- Kunda, Z., Miller, D. T., & Claire, T. (1990). Combining social concepts: The role of causal reasoning. *Cognitive Science*, 14, 551-557.
- Levin, D.T. (2000). Race as a visual feature: Using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General*, 129, 559-574.
- Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(3), 732-753.
- MacLin, O. H., & Malpass, R. (2003). The ambiguous-race face illusion. *Perception*, 32, 249-252.
- Macrae, C.N., Milne, A.B., & Bodenhausen, G.V. (1994). Stereotypes and energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, 66, 37-47.
- Macrae, C.N., Stangor, C. & Hewstone, M. (Eds.) (1996). *Stereotypes and stereotyping*. New York: Guilford.
- Medin, D. L., & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review*, 111(4), 960-983.
- Medin, D., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 179 - 195). New York: Cambridge University Press.
- Michel, C., Rossion, H. J., Chung, C. H., & Caldara, R. (2006b). Holistic processing is finely tuned for faces of one's own race. *Psychological Science*, 17(7), 608-615.
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92(3), 289-316.
- Newell, F. N., & Bulthoff, H. H. (2002). Categorical Perception of familiar objects. *Cognition*, 85, 113-143.
- Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129(3), 369-398.
- Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*, 7, 246-251.
- Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, 133(2), 166-188.
- Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgment. *British Journal of Psychology*, 54, 101-114.
- Waxman, S. R., & Booth, A. E. (2001). Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive Psychology*, 43, 217-242.
- Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, 29(3), 257-302.
- Yamauchi, T. (2005). Labeling bias and categorical induction: Generative aspects of category information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 538-553.
- Yamauchi, T. (2008) Linking syntax and inductive reasoning: Categorical labeling and generic noun phrases. *Psychologia*, 51, 1-13.
- Yamauchi, T. (2009). Finding abstract commonalities of category members. *Journal of Experimental and Theoretical Artificial Intelligence*, 21 (3), 155-180.
- Yamauchi, T., Kohn, N., & Yu, N. Y. (2007). Tracking mouse movement in feature inference: Category labels are different from feature labels. *Memory & Cognition*, 35(5), 852-863.
- Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 28(3), 585-593.
- Yamauchi, T., & Markman, A. B. (2000). Learning categories composed of varying instances: The effect of classification, inference, and structural alignment. *Memory & Cognition*, 28, 64-78.
- Yamauchi, T., & Yu, N. Y. (2008). Category labels versus feature labels: Category labels polarize inferential predictions. *Memory & Cognition*, 36(3), 544-553.
- Yu, N., Yamauchi, T., & Schumatcher, J. (2008). Rediscovering symbols: The role of category labels in similarity judgment. *Journal of Cognitive Science*, 9, 89-109.
- Yu, N., Yamauchi, T., Yang, H., Chen, Y. & Gutierrez-Osuna, R., (2010). Feature selection for inductive generalization. *Cognitive Science*, 34, 1574-1593.
- Yuill, N. (1992). Children's conception of personality traits. *Human Development*, 35, 265 - 279.

# A Critical Look at the Findings of Sergent (1982)

**Lyuben D. Laskin (lyubenlaskin@gmail.com)**

Department of Cognitive Science and Psychology,  
New Bulgarian University, 21 Montevideo Street  
Sofia 1618, Bulgaria

**Meryl Varadinov (meryl.varadinov@gmail.com)**

Department of Cognitive Science and Psychology,  
New Bulgarian University, 21 Montevideo Street  
Sofia 1618, Bulgaria

## Abstract

It is widely believed that local and global levels of visual stimuli are better processed in the left and right cerebral hemispheres, respectively. One classic explanation for this observation is the spatial frequency hypothesis proposed by Sergent (1982), which states that the left hemisphere is more efficient at processing high spatial frequencies, whereas the right hemisphere is better with low spatial frequencies. Sergent tested this by measuring RTs for laterally presented stimuli (in the left and right visual fields) composed of high and low spatial frequencies and obtained results consistent with the hypothesis. We put Sergent's findings to the test by replicating her experiment; our first experiment was a direct replication of hers, while the second used the same procedure, but with different stimuli. Our results largely corresponded with those of Sergent, and the crucial interaction between visual field and spatial frequency was obtained in Experiment 1, but was qualitatively different from Sergent's. Possible explanations are discussed.

**Keywords:** spatial frequency; global/local processing; hemispheric differences; hierarchical stimuli; visual hemifield paradigm.

## Introduction

Neuropsychological case studies with brain-injured patients, neuroimaging studies, and experimental research with healthy participants have shown that there are certain functional asymmetries in the left and right hemispheres of the human brain. (Springer & Deutsch, 2001). When it comes to visual perception, one of the most studied phenomena in the field of hemispheric asymmetries concerns the difference between the left and right hemispheres in their ability to process (1) global vs. local aspects and (2) categorical vs. coordinate spatial relationships of visual stimuli. Studies show that the left hemisphere (LH) is better at processing the details of a visual stimulus, whereas the right hemisphere (RH) is superior for processing its overall shape, patterns formed by Gestalt principles, etc. (Han et. al., 2002; Hellige, 1996; Ivry & Robertson, 1998; Van Kleeck & Kosslyn, 1989). Similarly, research has shown a LH advantage in processing distance-independent categorical spatial tasks (e.g., "Is the dot to the left or right of the vertical line?"), and a RH advantage for spatial tasks that require relative- or absolute-distance spatial judgments (e.g., "Which dot is closer to the

vertical line?"; Hellige & Michimata, 1989; Jager & Postma, 2003; Kosslyn et. al., 1989, 1994).

An early attempt to explain some of the hemispheric asymmetries was made by Sergent (1982) who proposed that the global/local effect was due to hemispheric differences in the capacity to process different spatial frequencies.<sup>1</sup> According to this hypothesis, the LH advantage for local stimuli emerges because the LH is better at processing high spatial frequencies (HSF), whereas the RH global advantage is due to its efficiency in processing low spatial frequencies (LSF). Attempts to verify or falsify this hypothesis have been rather controversial. Some neuroimaging and neuropsychological studies support the LH-HSF/RH-LSF asymmetry (e.g., Han et. al., 2001; Mecacci, 1993; Peyrin et. al., 2004, 2006a; Woodhead, et. al., 2011), whereas others show results partially or entirely inconsistent with it (e.g., Grabowska et. al., 1989; Fink et. al., 1997, 1999). Studies with healthy participants which use the visual hemifield paradigm yield similarly mixed results, with some research consistent (Hübner, 1998; Peyrin et. al., 2006b; Proverbio et. al., 1997; Van Kleeck & Kosslyn, 1989) and other inconsistent (Blanca & Lopez-Montiel, 2009; Evert & Kmen, 2002) with the hypothesis.

Yovel, Yovel, & Levy (2001) present a good meta-analysis of studies that used the visual hemifield paradigm with hierarchical stimuli. These are mostly Navon-type letters (large letters composed of small letters; Navon, 1977). The basic idea behind those studies is that when such a stimulus is presented in the left (LVF) or right (RVF) visual field, the response times (RT) and error rates should show a LVF advantage when the response is given based on the large letter and a RVF advantage when it is given based on the small letters.<sup>2</sup> Despite the theoretical predictions, Yovel et. al. found that studies confirming them are outnumbered by those that don't find a significant visual field (VF)×stimulus type interaction (i.e., more studies failed to find one or both of the LVF-LSF/RVF-HSF advantages).

---

<sup>1</sup> This hypothesis has later been used to explain the categorical/coordinate spatial relationship asymmetry as well (Ivry & Robertson, 1998).

<sup>2</sup> Visual input to one of the VFs is initially received and processed by the contralateral hemisphere (Beaumont, 1983).



Of the studies reviewed by Yovel et. al. which used the divided attention task, only Sergent (1982) found the significant interaction mentioned above. Another interesting observation is that, as a whole, the studies that reported significant results used fewer participants than those that didn't. Furthermore, the authors of this paper have previously employed Sergent's procedure along with an additional manipulation, but failed to obtain positive results. Sergent (1982) has been one of the most cited studies in the literature of hemispheric asymmetries for low and high spatial frequencies, yet, to our knowledge, there have been no attempts to replicate its findings. For those reasons, we decided that there is value in conducting a study with the same methodology, but significantly more participants, in order to gain more insight into the validity of the original study's results.

### Sergent (1982) revisited

Sergent's study used Navon-type hierarchical letters projected to the LVF, RVF, and central visual field (CVF). The stimulus material consisted of 4 letters (for a total of 16 large-small letter combinations), 2 of which were targets and 2 non-targets. Twelve participants had to press a "yes" button if either the large or small letters were target, or a "no" button otherwise. The critical finding was that in the so-called "conflict conditions" in which a non-target large letter was composed of small target letters (L+S-), or vice versa (L-S+) a VF×stimulus type interaction was observed. That is, L+S- stimuli were responded to faster in the LVF/RH than in the RVF/LH, with the opposite result for the L-S+ condition (for the full results, see Fig. 2a). These results were interpreted in light of the spatial frequency hypothesis, i.e., that the LH is superior for HSF (small letters), whereas the RH shows an advantage for LSF (large letters).<sup>3</sup>

In Experiment 1, we use nearly the same procedure but with significantly more participants.

## Experiment 1

### Method

#### Participants

Forty-one volunteers and New Bulgarian University students (21 men and 20 women, aged 19-42) took part in the experiment for course credit. All were right-handed and with normal or corrected-to-normal vision. Their visual acuity was tested by presenting 4 small letters (the same as those used in the experiment) four times each in the RVF, CVF, and LVF, which they had to identify. All participants could identify most of the letters presented, deeming them fit to participate in the study.

<sup>3</sup> Sergent proposed that since the large and small letters are of equal complexity, they should differ only on the spatial frequency dimension.

### Stimulus Material and Apparatus

The stimuli were large letters composed of small letters. The letters used were H, C, Т и P from the Cyrillic alphabet (analogues of the English letters N, S, T, and R), the combinations of which add up to a total of 16 stimuli. The letters were chosen to visually resemble the ones in Sergent's study, which were H, L, T and F (L was replaced with C and F was replaced with P). In our study the target letters were H and C (Fig. 1).

Large letters subtended visual angles of  $2.08^{\circ} \times 1.36^{\circ}$  and small letters subtended angles of  $0.23^{\circ} \times 0.16^{\circ}$ . The stimuli were black and were presented against a white background using the E-Prime software package on a 19" monitor (refresh rate of 200 Hz) of a Samsung SyncMaster 959 NF. The stimuli were presented in the LVF, CVF, and RVF. In lateral presentations, the center of the stimulus appeared  $1.4^{\circ}$  to the left or to the right of the fixation cross.

### Design and Procedure

Participants were led into an experimental booth and introduced to the experiment by signing a consent form with general information about it. They were asked about handedness and tested for visual acuity. Instructions were given with emphasis on the importance of speed and accuracy of responses. Each trial began with the appearance of a fixation cross for 1500 msec, followed by a stimulus appearing for 150 msec in the LVF, CVF, or RVF, a 2000 msec response window (which terminated when participants pressed one of the response buttons), and a 2000 msec intertrial interval. Participants had to determine whether the stimulus contained either H or C or both by pressing a "yes" key if it did and a "no" key if neither level contained a target letter. The hand for response was counterbalanced across participants.

The experimental procedure began with a practice session, consisting of 36 trials with feedback on response accuracy, followed by the experimental session, consisting of five blocks of 72 trials for a total of 360 trials. The duration of the entire session was about 35 minutes, with 4 breaks in between experimental blocks, during which participants took several seconds to rest their eyes.

Both independent variables were within-subject: (1) VF (left, central, right); (2) stimulus type (6 types; see Fig. 1). An equal number of positive (containing a target at at least one level) and negative (containing only non-target letters) stimuli were randomly presented in each VF (each of the 4 negative combinations was presented 3 times as frequently as the remaining 12 combinations).

### Results and Discussion

Before performing any analyses on the RTs for the different conditions, we removed all incorrect responses. The average accuracy was 97% and no participants were excluded from further analysis based on low accuracy. We also removed all data points above or below 2.5 standard deviations from the mean for each participant (excluding 2.6% of the data). There was no main effect of the between-subject factors sex,

$t(39) = 0.175$ ,  $p = 0.862$ , and hand for response,  $t(39) = 0.212$ ,  $p = 0.83$ , so all further analysis was collapsed over them.

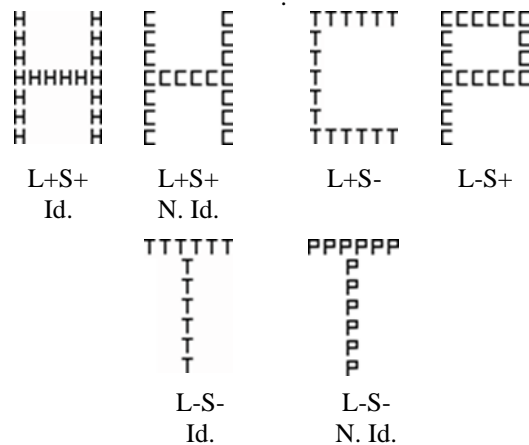


Figure 1: Sample stimuli used in Experiment 1.

Top left: positive, non-conflict stimuli with identical and non-identical large and small target letters; Top right: positive, conflict stimuli with a large target letter composed of small non-target letters, and a large non-target letter composed of small target letters. Bottom: negative stimuli with identical and non-identical non-target large and small letters. The C and P in our study replaced Sergent's L and F, respectively.

Figure 2b displays the mean RTs and standard errors for each condition. A repeated measures ANOVA revealed a main effect of VF,  $F(2, 39) = 15$ ,  $p < 0.001$ , with about 20 msec faster latencies in the LVF and CVF compared to the RVF. There was also a main effect of stimulus type,  $F(5, 36) = 37.984$ ,  $p < 0.001$ , with the L+S+ conditions having the shortest latencies, while the L-S- conditions had the longest latencies. Similar to Sergent (1982), we also found a main effect of identity,  $F(1, 40) = 41.977$ ,  $p < 0.001$ . That is, it took less time to respond to identical non-conflict stimuli (i.e., large targets made up of the same small targets and large non-targets made up of the same small non-targets), compared to their non-identical counterparts.

Next, we analyzed the conflict conditions, where we found a marginally significant difference in latencies. RTs were about 17 msec faster for L+S- than for L-S+,  $F(1, 40) = 3.548$ ,  $p = 0.067$ . This result is consistent with findings from the psychophysical literature, according to which LSFs are available to the visual system earlier than HSFs and are therefore processed faster (Breitmeyer, 1975; Breitmeyer & Ganz, 1977; Kulikowski & Tolhurst, 1973; Vassilev, Mihaylova, & Bonnet, 2001).

The conflict conditions were most interesting in regards to Sergent's hypothesis, namely, that there should be a LVH-RH advantage for the L+S- (LSF) condition and a RVF-LH advantage for the L-S+ condition (HSF). That is, when the decision is based on the large letter, there should be a RH

advantage, whereas the opposite result should hold when the decision is based on the small letters. Despite the overwhelming correspondence between Sergent's and our results, as seen in Figure 2a and 2b, we only found a LVF advantage for the large letters, but, critically, no RVF advantage for the small letters. In fact, the LVF was superior for both large and small letters. Nevertheless, we did find a significant two-way interaction between left/right visual fields and the L+S-/L-S+ conditions,  $F(1, 40) = 34.742$ ,  $p < 0.001$ . This is mainly due to the difference between the LVF and RVF being smaller for the L-S+ condition; when the decision was based on the small letters, participants were significantly slower to respond when the stimulus appeared in the LVF (but not slower than the RVF). However, we don't consider this observation to be good support for the hypothesis. First, the case may be that there is a ceiling effect when it comes to RVF RTs for the L-S+ stimuli. Second, it is very difficult to draw strong conclusions from RT interactions which don't involve change in sign of the slope of latency curves across conditions, given that the relationship between task complexity and RT is not always linear. In other words, the interaction could have simply been due to task difficulty (since small letters are more difficult to process) and not because of hemispheric asymmetries related to processing different spatial frequencies.

Given the discrepancy between our results and the results obtained by Sergent, we decided to analyze the data from individual participants separately. Another incentive for doing so was based on the controversial results obtained from other studies using similar procedures. We found 12 participants (from a total of 41) whose data had the pattern consistent with Sergent's results (i.e., shorter RTs in LVF than RVF for the L+S- condition, and shorter RTs for RVF than LVF for the L-S+ condition). Their results can be seen in Fig. 2d. A comparison of those and Sergent's results reveal remarkably similar patterns (coincidentally, she also used 12 participants).

Given the conflicting results of previous studies, combined with our findings, we consider the possibility that there might be a yet undiscovered dimension on which people differ. That is, there may be individual differences when it comes to hemispheric asymmetries for spatial frequency processing. Hence, it might have been the case that Sergent (1982) obtained positive results due to an unrepresentative sample of participants from the general population. We do not find this explanation to be particularly appealing, however, since it is done post-hoc, rather than based on empirical or theoretical reasoning. It is also possible, though less likely, given our larger sample size, that it is *our* sample that was unrepresentative. For those reasons, we decided to conduct a second experiment with different participants, using the same procedure, but slightly different stimuli. Instead of using hierarchical letters, we used hierarchical shapes. The purpose of this was also partly exploratory. It is possible that using verbal

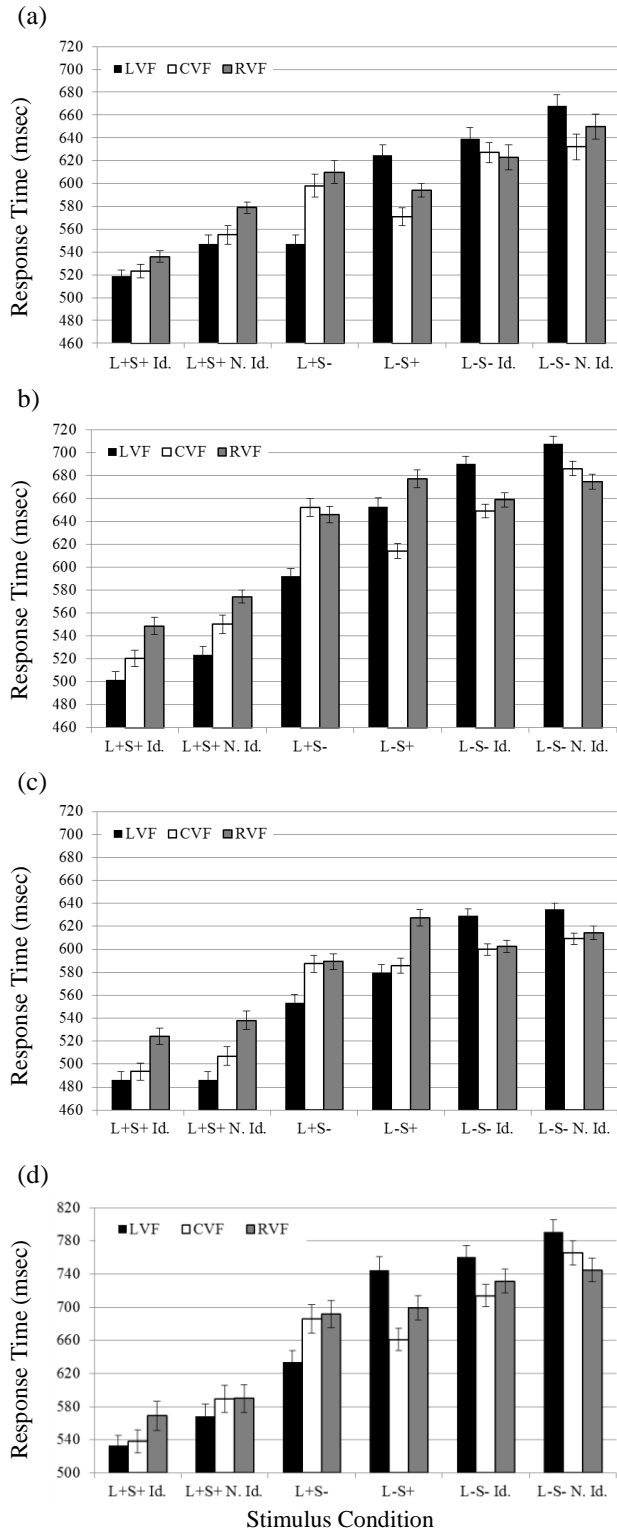


Figure 2: Mean RTs and standard errors for the six stimulus conditions for (a) Sargent (1982); (b) Experiment 1 of this study; (c) Experiment 2 of this study; (d) The data from 12 participants from Experiment 1 with results consistent with the spatial frequency hypothesis.

stimuli is a confounding factor, since it is a well-established fact that the LH plays a bigger role in language processing than does the RH (Springer & Deutsch, 2001). We wanted to see if using non-verbal stimuli would affect the results in any direction.

## Experiment 2

### Method

The method was identical to that of Experiment 1 with a few exceptions. Compound shapes were used, instead of compound letters (squares and triangles were target, whereas circles and crosses were non-target, see Fig. 3) and they subtended visual angles of  $1.94^{\circ} \times 1.94^{\circ}$  for the large shapes, and  $0.24^{\circ} \times 0.24^{\circ}$  for the small shapes. Thirty-nine right-handed volunteers and New Bulgarian University students (18 men and 21 women, aged 19-38) took part in the experiment for course credit (none of them had participated in Experiment 1).

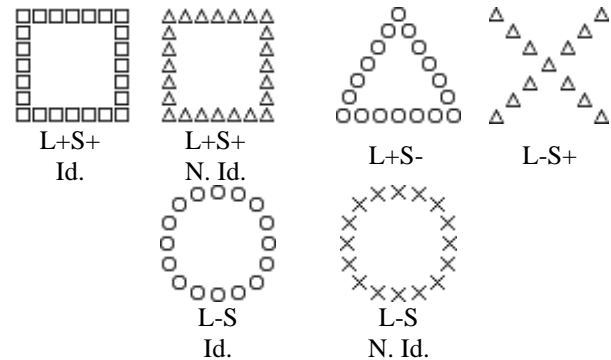


Figure 3: Sample stimuli used in Experiment 2

### Results and Discussion

As in Experiment 1, we excluded all incorrect responses from further analyses. The average accuracy was 97% but one participant was excluded because of a large error rate (26%). As before, we removed all data points above or below 2.5 SD from the mean (excluding 2.8% of the data). There was no main effect of the hand for response factor,  $t(36) = 1.036$ ,  $p = .307$ , but there was a marginally significant main effect of sex,  $t(36) = 1.879$ ,  $p = .068$ , with men having a small average RT (by 63 msec) than women. The latter factor and the hand for response factor did not interact with any others in a theoretically meaningful or statistically significant way, so we collapsed all further analyses over them.

Figure 2c displays the mean RTs and standard errors for each condition. A repeated measures ANOVA showed a main effect of VF,  $F(2, 36) = 32.054$ ,  $p < 0.001$ . As in Experiment 1, the latencies were about 20 msec faster in the LVF and CVF compared to the RVF. There was also a main effect of stimulus type,  $F(5, 33) = 60.144$ ,  $p < 0.001$ , with the L+S+ conditions having the fastest latencies, while the L-S- conditions had the slowest latencies. Similar to Experiment 1, we also found a main effect of identity,  $F(1,$

37) = 9.743,  $p = 0.003$ , and a large letter advantage,  $F(1, 37) = 16.605$ ,  $p < 0.001$ .

No significant visual field-stimulus type interaction was obtained,  $F(1, 37) = 1.25$ ,  $p < 0.269$ , but we still observed a LVF-RH advantage for both large and small shapes, and no RVF-LH advantage whatsoever for small shapes. As in the previous experiment, these results run counter to Sergent's findings. Unlike in the previous experiment, in this one only two of 38 participants exhibited similar patterns to the ones Sergent reported. Our results are also inconsistent with some of the studies reporting global/local or LSF/HSF interactions with VF presentations (Gier et. al., 2010; Hübner, 1998).

## General Discussion

Our experiments revealed several consistent findings. First, participants were faster as a function of both stimulus type and stimulus identity. That is, they were fastest to respond to a stimulus if both the large and small letters were target, slower when only one of the levels was target, and slowest when neither level was target; they were also faster to respond when the two levels were represented by the same letter or shape. In our interpretation, this is consistent with a model of parallel processing of the two levels where processing is facilitated when the global and local levels are composed of the same letters/shapes and when they're both either target or non-target. Therefore, when both levels are identical, it is more likely that at least one of them will evoke neural mechanisms that will lead to the proper response. Other studies have also shown evidence supporting the parallel processing of global and local levels of hierarchical stimuli (e.g., Hübner, 1997).

Another result of our experiments that is consistent with the literature is the idea that LSF are available to the visual system earlier than HSF. This phenomenon has been explained by the differences between the magnocellular and parvocellular visual pathways. They correspond to the *transient* and *sustained* channels described by Breitmeyer (1975), Breitmeyer & Ganz (1977), and Kulikowski & Tolhurst (1973). The magnocellular pathway is most sensitive to LSH and high temporal frequencies (HTF), whereas the parvocellular pathway is most sensitive to HSF and low temporal frequencies (LTF). Furthermore, the magnocellular pathway is more efficient when it comes to the speed with which it propagates information to the higher cortical structures, which explains the earlier availability of LSF.

When it comes to the central hypothesis that was tested in this study, our results are inconsistent with those of Sergent (1982). We did not observe the critical interaction between spatial frequency and the visual field of stimulus presentation. At the same time, it could be argued that our results are consistent with the general trend observed in studies using the hierarchical stimulus/visual hemifield paradigms of finding mixed results. A somewhat interesting finding of our study was that 12 participants showed RT patterns for the conflict conditions that were in line with the

hypothesis, but the remaining 29 participants did not. This might suggest that the conflicting results in the literature could be partially explained by individual differences on an unidentified dimension. Individual differences in processing time of the different levels (Evert & Kmen, 2002; Kimchi, 1992; Peyrin et. al., 2006b) are a possible candidate dimension. For example, Peyrin et. al. found that the classic hemispheric asymmetry for spatial frequency processing occurred only when the stimulus presentation time was 30 msec, whereas only a LVF/RH advantage emerged when the presentation time was 150 msec. Note that the latter is consistent with the results from our experiments. We propose that there may be between-subject, as well as within-subject, differences related to stimulus presentation time underlying these effects. Kimchi (1992) has done an overview of a variety of other factors that could influence the global/RH advantage in these studies, from the overall visual angle, sparsity, and number of local elements of the stimuli to goodness of form and attentional factors.

Finally, it is worth noting that some researchers argue that hierarchical stimuli are not a good way of manipulating spatial frequency (e.g., Peyrin et. al., 2003). Also, Yovel et. al. (2001) point to the salience of the hierarchical stimuli as a determining factor for obtaining the critical interaction. They observed a RH advantage for both global and local levels when stimuli were globally salient, as opposed to equally salient, in which case they observed a RH advantage for global letters and a LH advantage for local letters. The stimuli used in Sergent's and our study are of the globally salient type.

We conclude that, despite the large number of studies, there is still a significant amount of uncertainty surrounding the nature of hemispheric asymmetries for low and high spatial frequency processing. Further studies are needed to explore the underlying mechanisms, the different stimulus and procedural factors involved, as well as possible individual differences that are at play.

## Acknowledgments

We would like to thank our volunteers and Oksana Itkes for her assistance in finding many of the references. We also express our gratitude towards prof. Angel Vassilev and Ivan Vankov for their valuable advice and fruitful discussions.

## References

- Beaumont, J. G. (1983). Methods for studying cerebral hemispheric function. In A. W. Young (Ed.), *Functions of the right cerebral hemisphere*. London: Academic Press.
- Blanca, M.J., & López-Montiel, G. (2009). Hemispheric Differences for Global and Local Processing: Effect of Stimulus Size and Sparsity. *The Spanish Journal of Psychology*, 12 (1), 21-31.
- Breitmeyer, B. (1975). Simple reaction time as a measure of the temporal response properties of the transient and sustained channels. *Vision Research*, 15, 1411-1412.

- Breitmeyer, B., & Ganz, L. (1977). Temporal studies with flashed gratings: Inferences about human transient and sustained channels. *Vision Research*, 17, 861-865.
- Evert, D. L., & Kmen, M. (2003). Hemispheric asymmetries for global and local processing as a function of stimulus exposure duration. *Brain and Cognition*, 51, 115-142.
- Fink, G. R., Marshall, J. C., Halligan, P. W., Frith, C. D., Frackowiak, R. S., & Dolan, R. J. (1997). Hemispheric specialization for global and local processing: the effect of stimulus category. *Proceedings of the Royal Society of London B: Biological Sciences*, 264(1381), 487-494.
- Fink, G. R., Marshall, J. C., Halligan, P. W., & Dolan, R. J. (1999). Hemispheric asymmetries in global/local processing are modulated by perceptual salience. *Neuropsychologia*, 37, 31-40.
- Grabowska, A., Semenza, C., Denes, G., & Testa, S. (1989). Impaired grating discrimination following right hemisphere damage. *Neuropsychologia*, 27, 259-263.
- Gier, V., Kreiner, D., Solso, R., & Cox, S. L. (2010). The Hemispheric Lateralization for Processing Geometric Word/Shape Combinations: The Stroop-Shape Effect. *The Journal of General Psychology*, 137(1), 1-19.
- Han, S., Waver, J. A., Murray, S. O., Kang, X., Yund, E. W., & Woods, D. L. (2002). Hemispheric asymmetry in global/local processing: effects of stimulus position and spatial frequency. *Neuroimage*, 17(3), 1290-9.
- Hellige, J.B. (1996). Hemispheric asymmetry for visual information processing. *Acta Neurobiologiae Experimentalis*, 56, 485-497.
- Hellige, J.B., & Michimata, C. (1989). Categorization versus distance: Hemispheric differences for processing spatial information. *Memory and Cognition*, 17, 770-776.
- Hübner, R., (1997). The effect of spatial frequency on global precedence and hemispheric differences. *Perception & Psychophysics*, 59 (2), 187-201.
- Hübner, R., (1998). Hemispheric Differences in Global/Local Processing Revealed by Same-Different Judgements. *Visual Cognition*, 5 (4), 457-478.
- Ivry, R .B., & Robertson, L.C. (1998). The two sides of perception. *Cambridge, MA: The MIT Press*.
- Jager, G., & Postma, A. (2003). On the hemispheric specialization of categorical and coordinate spatial relations: A review of the current evidence. *Neuropsychologia*, 41, 504-515.
- Kimchi, R. (1992). Primacy of wholistic processing and global/local paradigm: A critical review. *Psychological Bulletin*, 2(1), 24-38.
- Kosslyn, S. M., Koenig, O., Barrett, A., Cave, C. B., Tang, J., & Gabrieli, J. D. E. (1989). Evidence for two types of spatial representations: Hemispheric specialization for categorical and coordinate relations. *Journal of Experimental Psychology: Human Perception & Performance*, 15, 723-735.
- Kosslyn, M. S., Anderson, A. K., Hillger, L. A., & Hamilton, S. E. (1994). Hemispheric differences in sizes of receptive fields or attentional biases? *Neuropsychology*, 8(2), 139-147.
- Kulikowski, J. J. & Tolhurst, D. J. (1973). Psychophysical evidence for sustained and transient detectors in human vision. *Journal of Physiology*, 232, 149-162.
- Mecacci, L. (1993). On spatial frequencies and cerebral hemispheres: some remarks from the electrophysiological and neuropsychological points of view. *Brain and Cognition*, 22(2), 199-212.
- Navon, D. (1977). Forest before trees: the precedence of global features in visual perception. *Cognitive Psychology*, 9, 353- 383.
- Peyrin, C., Chauvin, A., Chokron, S., & Marendaz, C. (2003). Hemispheric specialization for spatial frequency processing in the analysis of natural scenes. *Brain and Cognition*, 53, 278-282.
- Peyrin, C. Baci, M., Segebarth, C., Marendaz, C. (2004). Cerebral regions and hemispheric specialization for processing spatial frequencies during natural scene recognition. An event-related fMRI study. *Neuroimage*, 23, 698-707.
- Peyrin, C., Chokron, S., Guyader, N., Gout, O., Moret, J., & Marendaz, C. (2006a). Neural correlates of spatial frequency processing: A neuropsychological approach. *Brain Research*, 1073, 1-10.
- Peyrin, C., Mermillod, M., Chokron, S., & Marendaz, C. (2006b). Effect of temporal constraints on hemispheric asymmetries during spatial frequency processing. *Brain Cognition*, 62(3), 214-20.
- Proverbio, A. M., Zani, A., & Avella, C. (1997). Hemispheric Asymmetries for Spatial Frequency Discrimination in a Selective Attention Task. *Brain and Cognition*, 34, 311-20.
- Sergeant, J. (1982). The cerebral balance of power: Confrontation or cooperation? *Journal of Experimental Psychology: Human Perception and Performance*, 8, 253-272.
- Springer, S. P., & Deutsch, G. (2001). Left brain, right brain: Perspectives from cognitive science. *NY: W. H. Freeman and Company Worth Publishers*.
- Van Kleeck, M., & Kosslyn, S. M. (1989). Gestalt laws of perceptual organization in an embedded figures task. *Neuropsychologia*, 27(9), 1179-1186.
- Vassilev, A., Mihaylova, M., & Bonnet, Claude (2001). On the delay in processing high spatial frequency visual information: reaction time and VEP latency study of the effect of local intensity of stimulation. *Vision Research*, 42, 851-864.
- Woodhead, Z. V., Wise, R. J., Sereno, M., & Leech, R. (2011). Dissociation of sensitivity to spatial frequency in word and face preferential areas of the fusiform gyrus. *Cerebral Cortex*, 21, 2307-2312.
- Yovel, G., Yovel, I., & Levy, J. (2001). Hemispheric asymmetries for global and local visual perception: Effects of stimulus and task factors. *Journal of Experimental Psychology: Human Perception and Performance*, 27, 1369-1385.

# Neural Correlates of Episodic Memory Formation in Audio-Visual Pairing Tasks

**Chung-Yeon Lee (cylee@bi.snu.ac.kr)**

Brain Science Program, Seoul National University  
Seoul, 151-744, Republic of Korea

**Beom-Jin Lee (bjlee@bi.snu.ac.kr)**

Computer Science and Engineering, Seoul National University  
Seoul, 151-744, Republic of Korea

**Joon Shik Kim (jskim@bi.snu.ac.kr)**

Cognitive Science and Brain Science Programs, Seoul National University  
Seoul, 151-744, Republic of Korea

**Byoung-Tak Zhang (btzhang@snu.ac.kr)**

Computer Science and Engineering &  
Cognitive Science and Brain Science Programs, Seoul National University  
Seoul, 151-744, Republic of Korea

## Abstract

Understanding episodic memory formation of real-world events is essential for the investigation of human cognition. Most studies have stressed on delimiting the upper boundaries of this memory by using memorization tasks with conditional experimental paradigms, rather than the performance of everyday tasks. However, naturally occurring sensory stimuli are multimodal and dynamic. In an effort to investigate the encoding and retrieval of episodic memory under more naturalistic and ecological conditions, we here demonstrate a memory experiment that employs audio-visual movies as naturalistic stimuli. Electroencephalography measurements were used to analyze neural activations during memory formation. We found that oscillatory activities in the theta frequency bands on the left parietal lobe, and gamma frequency bands on the temporal lobes are related to overall memory formation. Theta and gamma power of the frontal lobes, and gamma power of the occipital lobes were both increased during retrieval tasks. Furthermore, subjects' memory retrieval performance on the query task was used to clarify our experimental results. Correlation between behavioral differences and neural activation was observed in the same regions. Our results extend the previous results of neurocognitive studies on memory formation via naturalistic stimuli, neural oscillations, and behavioral analysis combined.

**Keywords:** Memory formation; Naturalistic stimuli; EEG study; Neural oscillation; Reaction time.

## Introduction

Memory processing is one of the most prominent features of human cognition. Particularly, understanding how the experience of an episode in our everyday life can be transformed into a long-term memory is a central issue in human memory research.

Episodic memory formation is a complex neurocognitive process that involves many interacting brain regions, and

enables us to store contextual (spatial and temporal) information about individual events that can be later retrieved. In contrast, semantic memory consists of isolated facts that are decontextualized and not organized into a specific experience.

Research in cognitive neuroscience thus far has focused on the human memory system, and much knowledge has been acquired via the neuropsychological examination of brain-lesioned patients (Milner, Squire & Kandel, 1998; Squire, 2004). Furthermore, noninvasive neuroimaging studies in the last decade have revealed numerous functional roles of particular brain regions related to memory phenomena (Freeman, Dennis & Dunn, 2010; Spaniol et al., 2009).

However, the vast majority of experimental protocols used to study the neural mechanisms of episodic memory formation (Schacter & Addis, 2007; Vissers et al., 2008) deal with well-controlled experimental setups in which conditional static stimuli (e.g., single words or still images) are presented for memorization. They are typically dissimilar from what the human brain encounters in everyday life, since naturally occurring sensory stimuli are continuous and multimodal. (Bartlett, 1932; Jacobs & Shams, 2010; Tulving, 2002).

In an effort to investigate episodic memory encoding and retrieval under more naturalistic and ecological conditions, we performed a memory experiment employing audio-visual movies to mimic natural stimuli.

Movies are suitable for use as naturalistic stimuli because they reflect aspects of experiences in our daily lives by fusing multimodal sensory perception with emotional and cognitive overtones (Eisenstein, 1947; Hasson et al., 2007; Morin, 2005). In fact, cinematic materials have been used to explore memory since the early days of cinema (Boring, 1916), but only recently have some neuroimaging studies used audio-visual movies as natural stimuli for the examination of brain activation induced by memory

operations (Ben-Yakov et al., 2011; Furman et al., 2007; Green, Li & Bavelier, 2011; Hasson et al., 2007; Milton et al., 2011; Sestieri et al., 2011).

In contrast to recent studies investigating memory processes with functional magnetic resonance imaging (fMRI) during the presentation of video stimuli, here we aimed to examine brain activation using electroencephalogram (EEG) recordings. Neural oscillations have been linked to various cognitive phenomena in humans. Although there is no exact mapping of neural oscillatory rhythms to specific cognitive processes, neural oscillations in a specific frequency range in each brain region may function as their own cognitive process (Başar et al., 1999; Bechtel & Abrahamsen, 2010; Kahana, 2006; Klimesch et al., 2008).

We focused in this study on gamma and theta oscillations because recent research has suggested a functional role of theta (4–8 Hz) and gamma (30–50 Hz) oscillations in episodic memory (Doppelmayr et al., 1998; Klimesch et al., 2001, 2008; Osipova et al., 2006). Other rhythms such as delta (1–4 Hz), alpha (8–13 Hz), and beta rhythms (13–30 Hz) have not been discussed because they have not been consistently reported as being related to episodic memory, and they correlate not only with episodic memory but also with semantic memory (Düzel et al., 2005; Hanslmayr, Spitzer & Bauml, 2009; Weiss & Rappelsberger, 2000).

A secondary aim of this study was to investigate memory retrieval in consideration of the subject's behavioral factors. Because the human brain can be easily affected by numerous factors that disturb the expected neural response to experimental paradigms, there may not be a specific interpretation of the acquired neuroimaging results. To alleviate this problem, we designed a 2-way analysis of brain oscillation rhythms based on subjects' memory retrieval task performance. Using this approach enables the results of our oscillatory analysis during memory formation to be supported by behavioral analyses.

## Materials and Methods

Two goals were considered when constructing the task and methodology used in this study. First, a set of memory tasks was designed to look at 2 aspects of episodic memory: encoding and retrieval. Second, we obtained EEG data on the scalp while memory tasks were being performed by subjects in order to specify the anatomical locations of brain activations associated with episodic memory formation.

## Subjects

Ten neurologically healthy subjects (mean age,  $24.0 \pm 2.7$  years; 4 women) with normal or corrected-to-normal vision (mean eyesight in left and right eyes, 0.87/0.80) were recruited from Seoul National University (Seoul, Korea). Informed consent was obtained from all subjects in accordance with the guidelines of the institutional review board (IRB) at the Clinical Research Institute of Seoul National University Hospital.

## Stimuli

The experimental stimulus used in the encoding session was an audio-visual movie containing an episode of a television sitcom (duration, 27 minutes). The spoken language in the movie was American English, and subtitles were not displayed.

We extracted 20 video clips (each of 5 seconds), taking into consideration the current of its story, and captured 40 subsequent still images from the movie. These short video clips served as retrieval cues in the retrieval session of the experiment, and the captured images were used in the query tasks.

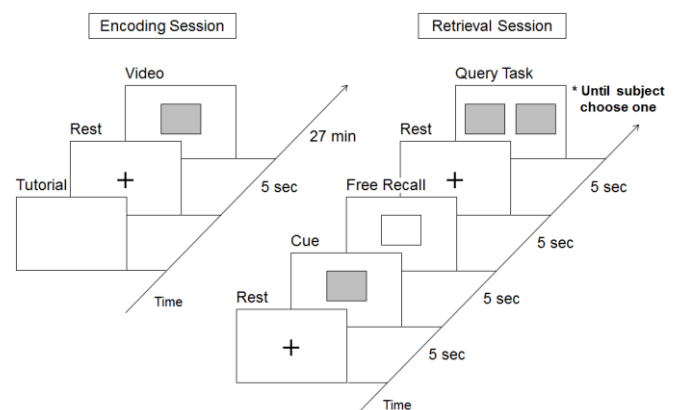
To make the experiment more accurate, we also developed a software program capable of displaying stimuli, including movies and images, sequentially following fixed time as input by an experimenter. Additionally, the program records subjects' responses and reaction times (RTs) during query tasks, so that these records could be used in EEG analysis. In the program, all the stimuli are 130 mm wide and 100 mm high, and were displayed with a black background on a 15-inch LCD monitor placed approximately 50 cm away from the subject.

## Experimental Paradigm

The entire experimental procedure is depicted in Figure 1. The memory experiment program begins with a tutorial task consisting of a concise example set of whole experiments in order to allow subjects to become accustomed to the experimental procedure.

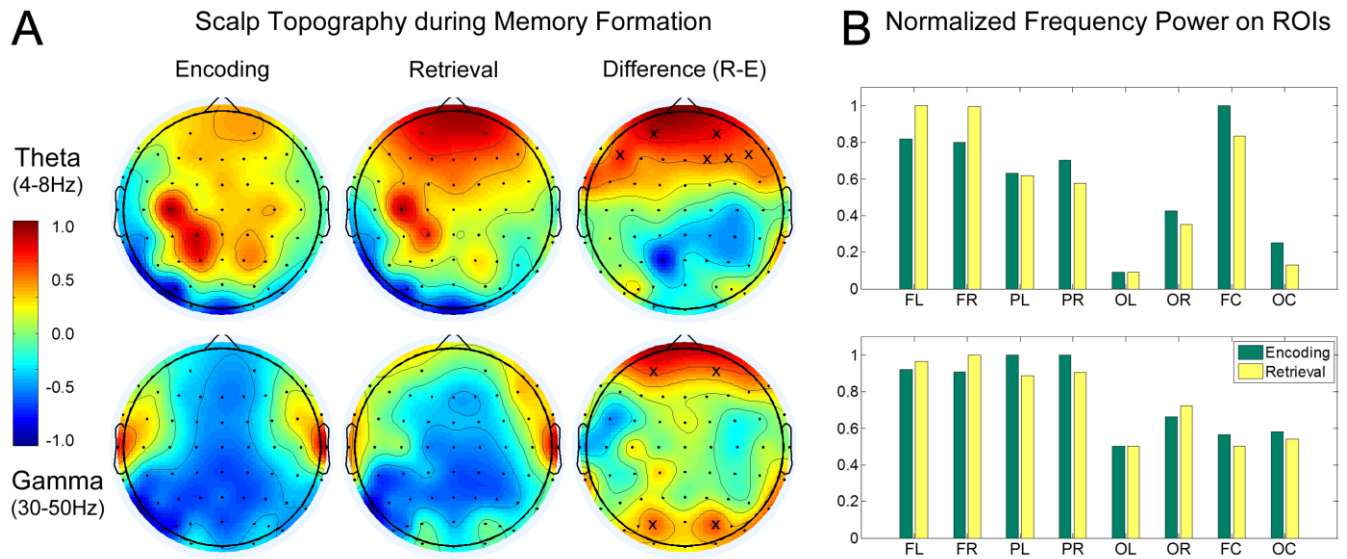
Following the tutorial, the encoding session starts, playing an episode of a television sitcom in its entirety. The subjects watched the movie without requiring any response. EEG measurement also began at this point. The EEG data were analyzed later by using the time stamps recorded by the program.

In the retrieval session, each subject performed video cue-based free recall tasks by imagining the scene immediately following the cue stimuli. Finally, as for the query task, 2 images contained in the cue video were presented in a random order. The subject was asked to decide whether the order of presentation is correct or incorrect, and press the "O



**Figure 1:** Schematic depicting the experimental paradigm.





**Figure 2:** Theta and gamma oscillatory activation during memory formation. (A) The first row indicates increased theta frequency power of the frontal lobes in the encoding task, and the second row indicates increased gamma frequency power of both the frontal and parietal lobes in the encoding task. Differences between retrieval and encoding tasks are depicted in the third column with ‘x’ markers which stand for locations showing the major differences (B) Normalized theta and gamma frequency power values on the 8 topographical regions of interest (frontal left, FL; frontal right, FR; parietal left, PL; parietal right, PR; occipital left, OL; occipital right, OR; frontal center, FC; and occipital center, OC) shows quantitative differences between encoding and retrieval sessions. In the retrieval session, theta activations on the left and right frontal regions were increased. Gamma activations on the frontal and parietal regions were increased, although other regions were decreased.

(correct)” or “X (incorrect)” button on a small keypad.

The experiment constituted 20 rounds of the retrieval task.

### EEG Measurement

Ongoing brain activity was recorded using Ag/AgCl electrodes mounted in a 128-channel Quik-cap, and a Neuroscan SynAmps amplifier (Neuroscan, El Paso, TX) in a dimly lit, soundproof, electrically shielded room at the Clinical Cognitive Neuroscience Center of Seoul National University Hospital. The ground electrode was located 10% anterior to FZ, with linked mastoids serving as references. Eye movements and blinks were monitored by a horizontal electrooculogram (hEOG) and a vertical electrooculogram (vEOG). Impedance was maintained at 5–10 kΩ or less. Throughout the experiment, EEGs were continuously obtained at a sampling rate of 1,000 Hz/channel.

### EEG Data Analysis

EEG data analysis was performed using the EEGLAB toolbox (Delorme & Makeig, 2004) developed at the Institute for Neural Computation in the University of California San Diego<sup>1</sup> using MATLAB 7.13 (MathWorks, Natick, MA). A baseline removal process was applied in order to eliminate some shift signals and to synchronize the zero levels of each channel. Eye movement artifacts were eliminated by excluding hEOG and vEOG components

extracted by Independent component analysis (ICA). All signals were then band-pass filtered between 4.0 and 50.0 Hz in order to exclude unnecessary frequencies. Some electrodes on the prefrontal area (AF7, FP1, FPZ, FP2, and AF8) were excluded to reduce any remaining chance of including ocular artifacts. A fast Fourier transform (FFT) in the frequency domain on the artifact-free EEG record was then performed. Frequency bandwidths were divided according to the following divisions: theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–50 Hz). From the computed FFTs, overall power averages were computed for each bandwidth.

## Experimental Results

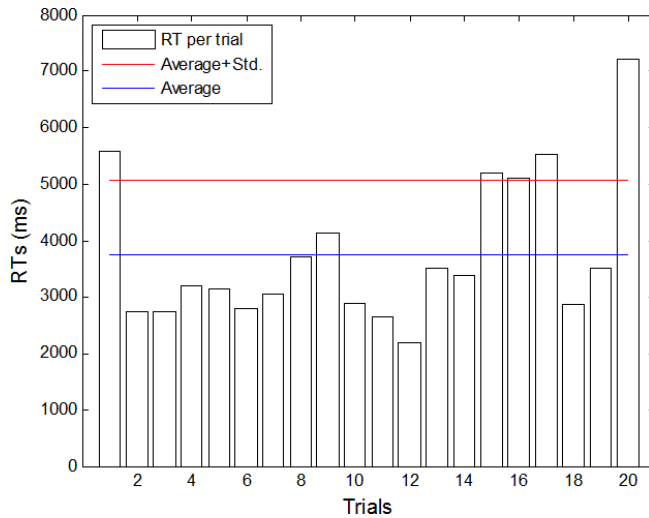
### Oscillatory Correlates of Memory Formation

2D brain maps showing the grand-average topographical distribution of theta and gamma frequency power during the 2 experimental conditions (memory encoding session and retrieval session) are presented in Figure 2(A).

First, we observed increased theta frequency activation on the left parietal lobe (C3 and CP1), and increased gamma activation on the temporal lobes (T7 and T8) during overall memory formation.

As seen in the topographical plots showing the difference between the encoding and retrieval sessions, theta and gamma power on the frontal lobes were more increased in retrieval sessions than in encoding sessions (indicated by

<sup>1</sup> <http://sccn.ucsd.edu/eeqlab>



**Figure 3:** RT histogram for one subject. The blue line stands for the average of RTs (3754ms) and the red line stands for the sum of the average and standard deviation of RTs (5053ms). From this subject, 5 trials (1st, 15th, 16th, 17th, and 20th) were chosen for the sustained RTs, and the remaining 15 trials were considered as the instant RTs.

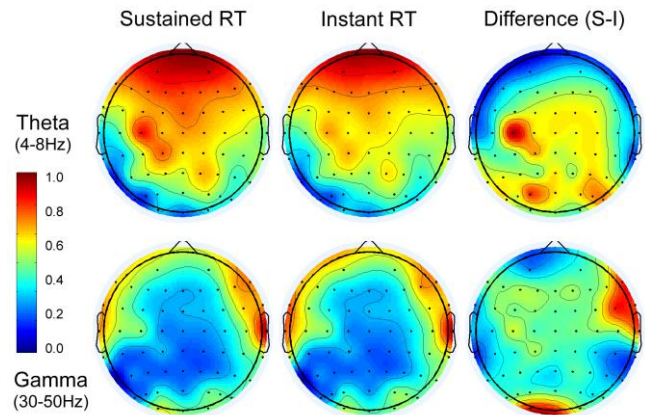
positive values of the difference between encoding and retrieval in Fig. 2(A)). Some of occipital theta powers in the retrieval session were also increased, but they were not significant differences. Contrarily, theta powers on the occipital center regions were decreased in retrieval sessions. Occipital gamma power was increased during the retrieval session. Independent *t*-tests verified that both differences are statistically significant for theta ( $F = 3.771$ ;  $p < 0.005$ ) and gamma frequency powers ( $F = 3.902$ ;  $p < 0.001$ ).

### Analysis based on Behavioral Factors

To clarify our finding regarding neural oscillations associated with memory formation, correlation between EEG and RT of the query task was analyzed. We classified the experimental results from 20 query tasks into 2 types: sustained RT and instant RT. Sustained and instant RT represent an RT that is longer and shorter, respectively, than the sum of the average and standard deviation of all RTs. Fig. 3 depicted RT histogram of one subject as an example. Theta and gamma oscillatory activations during the query task were then analyzed in consideration of the 2 different classes of responses on query tasks.

Interestingly, as shown in the topographical plots in Fig. 4, a correlation between neural activation and RT of subjects in query tasks were observed in the theta and gamma oscillations in some meaningful brain regions: increased theta power on the left parietal lobe, and increased gamma power on the right frontal and left occipital lobes during query tasks with sustained RT.

This indicates that memory formation, especially retrieval, is more activated when subjects make more effort on the query tasks. Additionally, this result supports the notion that regions with high theta and gamma power, including the left



**Figure 4:** Theta and gamma oscillatory activation during query task, classified according to the length of reaction time. Both theta and gamma were increased on the same regions with the former result of oscillatory activation during memory formation. In contrast to the tasks of instant RT, theta power on the left parietal lobe, gamma power on the right frontal and the left occipital lobes show greater activations in the RT-sustained tasks.

parietal lobe, right frontal lobes and left occipital lobes, are related to episodic memory retrieval.

As for other behavioral factors, the average correct response rate on the query task was 69.4% (about 14 questions were correct) and the average RT for each query task was  $7.1 \pm 3.3$  seconds. However, there was no significant effect observed in the frequency power analysis of factors such as true/false sequential order of presented pictures in the query task, correctly/incorrectly answered tasks, or the correlation between RT and these factors.

### Discussion

Recently, many functional neuroimaging studies using fMRI or positron emission tomography (PET) have shown that specific brain regions are significantly correlated with memory formation. For example, the temporal lobes have been long associated with memory retrieval (Nakamura & Kubota, 1996), and the medial prefrontal cortex (mPFC) and transverse occipital sulcus (transOCS) have been shown to play a significant role in subsequent memory retrieval (Hasson et al., 2007). Furthermore, neural oscillations in the gamma and theta frequencies have long been observed in cognitive tasks, such as those used to investigate episodic memory, and in focal lesion studies (Milner et al., 1985; Nyhus & Curran, 2010).

In consideration of using naturalistic stimuli, unlike traditional experiments that have consistently revealed memory effects for still images or words, we observed similar results in both the spatial and spectral domains. Moreover, our results link the frequency domain with each observed region (i.e., delta frequency on the frontal and left parietal lobe, and gamma frequency on the frontal and occipital lobes). This is similar to the results of several

studies that have shown greater gamma power in the posterior cortex for subsequently remembered items than for forgotten items (Hanslmayr et al., 2009; Osipova et al., 2006).

As for the individually different effects of the experimental stimuli, there is some evidence from previous studies showing the same effect in a previous experiment using similar movie narrative stimuli (Hasson et al., 2004; Wilson, Molnar-Szakacs & Iacoboni, 2008). Thus, a statistical evaluation showed that the observed brain activations in our results are significantly different from each other.

Our study has some limitations. It is not entirely unexpected that there was no significant correlation observed between correctly and incorrectly answered questions in the query task, because the subject might not care what the correct answer is at the time of the task. In addition, the paradigm of the query task is too ambiguous to determine which image has been shown earlier, because the presented images were both taken from the same scene. This problem could be addressed by choosing 2 images from different scenes, but such a paradigm might also evoke some semantic memory processes, with the subject cued by many possible different factors in the images or stories, such as actors' clothes and the scenery of each scene. Instead, RT seems to be a more appropriate factor for behavioral analysis because RT will be longer when subjects make more effort to retrieve stored memories, and thus, the brain region related to retrieval memory will be more activated at that time.

## Conclusion

In summary, we have demonstrated episodic memory formation using with an audio-visual movie as a naturalistic stimulus, and measured subjects' EEG signals throughout the task in order to reveal the neural responses involved in episodic memory operations. We also developed a memory task program designed to obtain more accurate experimental results. We identified oscillatory activities related to memory formation in the spatial and spectral domains. Different oscillatory activities in the retrieval session were also observed. A correlation between RTs and oscillatory activation during the query task was observed in the same regions. We expect that our naturalistic and behavioral factor-based analysis approach for memory investigation could help extend the range of neurocognitive research.

## Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2012-0005643, Videome; No. 2012-0005801, BrainNet), and the BK21-IT Program. We thank Clinical Cognitive Neuroscience Center of Seoul National University Hospital for technical supporting on EEG acquisition setup.

## References

- Bartlett, F. C. (1932). *Remembering: a study in experimental and social psychology*, Cambridge: Cambridge University Press.
- Başar, E., Başaar-Eroglu, C., Karakaş, S. and Schürmann, M. (1999). Are cognitive processes manifested in event-related gamma, alpha, theta and delta oscillations in the EEG?. *Neuroscience Letters*, 259, 165–168.
- Bechtel W. & Abrahamsen A. (2010). Understanding the Brain as an Endogenously Active Mechanism. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society.
- Ben-Yakov, A. and Dudai, Y. (2011). Constructing realistic engrams: poststimulus activity of hippocampus and dorsal striatum predicts subsequent episodic memory. *The Journal of Neuroscience*, 31, 9032–9042.
- Boring, E. G. (1916). Capacity to report upon moving pictures as conditioned by sex and age. *Journal of the American Institute of Criminal Law and Criminology*, 6, 820–834.
- Buckner, R. L., Logan, J., Donaldson, D. I., and Wheeler, M.E. (2000). Cognitive neuroscience of episodic memory encoding. *Acta Psychologica (Amsterdam)*, 105 127–139.
- Burgess, N., Becker, S., King, J. A., and O'Keefe, J. (2001). Memory for events and their spatial context: models and experiments. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 356, 1–11.
- Delorme, A. & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics. *Journal of Neuroscience Methods*, 134, 9–21.
- Doppelmayr, M., Klimesch, W., Schwaiger, J., Auinger, P. and Winkler, T. (1998). Theta synchronization in the human EEG and episodic retrieval. *Neuroscience Letters*, 257, 41–44.
- Düzel, E., Neufang, M. and Heinze, H. J. (2005). The oscillatory dynamics of recognition memory and its relationship to event-related responses. *Cerebral Cortex*, 15, 1992–2002.
- Eisenstein, S. (1949). *The Film Sense*, New York: Harcourt Brace & World, Inc.
- Freeman, E., Dennis, S., & Dunn, J. C. (2010). An examination of the ERP correlates of recognition memory using state-trace analysis. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society.
- Furman, O., Dorfman, N., & Hasson, U. (2007). They saw a movie: Long-term memory for an extended audiovisual narrative. *Learning and Memory*, 14, 457–467.
- Green, C. S., Li, R. and Bavelier, D. (2011). Perceptual learning during action video game playing. *Topics in cognitive science*, 2, 202–216.
- Hanslmayr, S., Spitzer, B. and Bauml, K.H. (2009). Brain oscillations dissociate between semantic and nonsemantic

- encoding of episodic memories. *Cerebral Cortex*, 19, 1631–1640.
- Hasson, U., Furman, O., Clark, D., Dudai, Y. and Davachi, L. (2007). Enhanced intersubject Correlations during movie viewing correlate with successful episodic encoding. *Neuron*, 57, 452–462.
- Jacobs, R. A. & Shams, L. (2010). Visual learning in multisensory environments. *Topics in Cognitive Science*, 2, 217–225.
- Kahana, M. J. (2006). The cognitive correlates of human brain oscillations. *The Journal of Neuroscience*, 26, 1669–1672.
- Klimesch, W., Doppelmayr, M., Yonelinas, A., Kroll, N. E., Lazzara, M., Röhm, D. and Gruber, W. (2001). Theta synchronization during episodic retrieval: neural correlates of conscious awareness. *Cognitive Brain Research*, 12, 33–38.
- Klimesch, W., Freunberger, R., Sauseng, P. and Gruber, W. (2008). A short review of slow phase synchronization and memory: evidence for control processes in different memory systems?. *Brain Research*, 1235, 31–44.
- Milner, B., Squire, L. R. and Kandel, E. R. (1998). Cognitive neuroscience and the study of memory. *Neuron*, 20, 445–468.
- Milton, F., Muhlert, N., Butler, C. R., Smith, A., Benattayallah, A., and Zeman, A. A. (2011). An fMRI study of long-term everyday memory using SenseCam. *Memory*, 19, 733–744.
- Morin, E. (2005). *The cinema, or, the imaginary man*, Minneapolis: University of Minnesota Press.
- Nakamura, K. & Kubota, K. (1996). The primate temporal pole: its putative role in object recognition and memory. *Behavioural Brain Research*, 77, 53–77.
- Nyhus, E. & Curran, T. (2010). Functional role of gamma and theta oscillations in episodic memory. *Neuroscience and Biobehavioral Reviews*, 34, 1023–1035.
- Osipova, D., Takashima, A., Oostenveld, R., Fernández, G., Maris, E. and Jensen, O. (2006). Theta and gamma oscillations predict encoding and retrieval of declarative memory. *The Journal of Neuroscience*, 26, 7523–7531.
- Sestieri, C., Corbetta, M., Romani, G. L. and Shulman, G. L. (2011). Episodic memory retrieval, parietal cortex, and the default mode network: functional and topographic analyses. *The Journal of Neuroscience*, 31, 4407–4420.
- Schacter D. L. & Addis D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 362, 773–786.
- Spaniol, J., Davidson, P. S., Kim, A. S., Han, H., Moscovitch, M. and Grady, C. L. (2009). Event-related fMRI studies of episodic encoding and retrieval: meta-analyses using activation likelihood estimation. *Neuropsychologia*, 47, 1765–1779.
- Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of Learning and Memory*, 82, 171–177.
- Tulving, E. (2002). Episodic memory: from mind to brain. *Annual Review of Psychology*, 53, 1–25.
- Vischers, C. T., Kolk, H. H., van de Meerendonk, N. and Chwilla, D. J. (2008). Monitoring in language perception: evidence from ERPs in a picture–sentence matching task. *Neuropsychologia*, 46, 967–982.
- Weiss, S. & Rappelsberger, P. (2000). Long-range EEG synchronization during word encoding correlates with successful memory performance. *Cognitive Brain Research*, 9, 299–312.
- Wilson, S.M., Molnar-Szakacs, I., and Iacoboni, M. (2008). Beyond superior temporal cortex: intersubject correlations in narrative speech comprehension. *Cerebral Cortex*, 18, 230–242.

# Human Cluster Evaluation and Formal Quality Measures: A Comparative Study

**Joshua M. Lewis**

josh@cogsci.ucsd.edu  
Dept. of Cognitive Science  
University of California, San Diego

**Margareta Ackerman**

mackerma@uwaterloo.ca  
Cheriton School of Computer Science  
University of Waterloo

**Virginia R. de Sa**

desa@cogsci.ucsd.edu  
Dept. of Cognitive Science  
University of California, San Diego

## Abstract

Clustering quality evaluation is an essential component of cluster analysis. Given the plethora of clustering techniques and their possible parameter settings, data analysts require sound means of comparing alternate partitions of the same data. When proposing a novel technique, researchers commonly apply two means of clustering quality evaluation. First, they apply formal Clustering Quality Measures (CQMs) to compare the results of the novel technique with those of previous algorithms. Second, they visually present the resultant partitions of the novel method and invite readers to see for themselves that it uncovers the correct partition. These two approaches are viewed as disjoint and complementary.

Our study compares formal CQMs with human evaluations using a diverse set of measures based on a novel theoretical taxonomy. We find that some highly natural CQMs are in sharp contrast with human evaluations while others correlate well. Through a comparison of clustering experts and novices, as well as a consistency analysis, we support the hypothesis that clustering evaluation skill is present in the general population.

**Keywords:** clustering; validity indices; psychophysics; visual perception; machine learning

## Introduction

Clustering is a fundamental data analysis tool that aims to group similar objects. It has been applied to a wide range of disciplines such as astronomy, bioinformatics, psychology, and marketing. Successful clustering often requires using a number of different clustering techniques and then comparing their output. The evaluation of clusterings is an integral part of the clustering process, needed not only to compare partitions to each other, but also to determine whether *any* of them are sufficiently good.<sup>1</sup>

As there is no universal clustering objective, there is no consensus on a formal definition of clustering. As a result, there are a wide variety of Clustering Quality Measures (CQMs), also known as internal validity indices, that aim to evaluate the quality of clusterings. To compare clusterings, researchers often select a CQM, which assigns a numerical value to a partition representing its quality.

Researchers rarely rely on CQMs alone. There is a deep implicit assumption running through the clustering literature that human judgment of clustering quality is quite good. Authors visually present the resultant partitions and invite readers to see for themselves that the new method performs well. To take one example, in their influential paper on spectral clustering Ng, Jordan and Weiss write, “The results are surprisingly good... the algorithm reliably finds clusterings consistent with what a human would have chosen.” (Ng, Jor-

dan, & Weiss, 2002) Up until now, clustering quality measures and human judgment were considered complementary approaches to clustering evaluation. Most papers that present novel clustering algorithms include these two types of evaluations separately.

Our study compares formal CQMs with human evaluations to determine how often they agree, and whether certain CQMs correlate better with human judgments than others. We also evaluate the consistency of human responses—if humans are very inconsistent, then it is unlikely that they are good judges of cluster quality (an ideal measure is stable on the same partition). Further, we separate our human subjects into expert and non-expert groups to determine whether clustering evaluation requires experience, and identify divergent strategies between the groups.

To sharpen our focus on a small set of CQMs, we construct a property-based taxonomy of CQMs that distinguishes them on grounds beyond their particular mathematical formulations. The CQMs selected for the study are diverse in that they each satisfy a distinct set of these properties.

Previous studies have investigated how humans choose the number of groups (Lewis, 2009) and partition data (Santos & Sá, 2005) in a clustering setting, but these approaches only show what humans think are the optimal partitions rather than how they judge partition quality in general. Our study uses a set of non-optimal partitions that humans partially order by quality, giving us more detailed quality judgments than in past work. Intuitively, in (Lewis, 2009) and (Santos & Sá, 2005) subjects took on the role of a *k*-choosing algorithm and a clustering algorithm (respectively), whereas in this study subjects are in the role of clustering evaluators.

Our main findings are as follows. Many CQMs with natural mathematical formalizations disagree with human evaluations. On the other hand, we identify CQMs whose evaluations are well correlated with those of humans. In particular, we find that Silhouette (Rousseeuw, 1987) and Calinski-Harabasz (Caliński & Harabasz, 1974) are highly correlated with human evaluations. Our findings also indicate that there is sufficient similarity between the evaluations of novices and experts to suggest that clustering evaluation is a task that does not require specific training (though it may benefit from training). This opens the door for using human computation resources such as Amazon’s Mechanical Turk to quickly solicit a large number of clustering quality judgments from novices as part of the data analysis process. Nevertheless, experts show much less sensitivity to the number of clusters and relate more closely to a greater range of clustering quality measures than novices, indicating a nuanced approach to the eval-

<sup>1</sup>If no good clusterings have been found the underlying dataset may have no good clustering (the data is not “clusterable”, see (Ackerman & Ben-David, 2009) for more on clusterability).



uation problem. Regarding consistency, we find that even novices are more consistent in their evaluations than our set of CQMs.

## Clustering quality measures

In this section we introduce the formal machinery describing the CQMs selected for our study.

Let  $X$  be a finite domain set. A *distance function* is a symmetric function  $d : X \times X \rightarrow \mathbb{R}^+$ , such that  $d(x, x) = 0$  for all  $x \in X$ . A  $k$ -*clustering*  $C = \{C_1, C_2, \dots, C_k\}$  of dataset  $X$  is a partition of  $X$  into  $k$  disjoint subsets (so,  $\cup_i C_i = X$ ). A *clustering* of  $X$  is a  $k$ -clustering of  $X$  for some  $1 \leq k \leq |X|$ . Let  $|C|$  denote the number of clusters in clustering  $C$ . For  $x, y \in X$  and clustering  $C$  of  $X$ , we write  $x \sim_C y$  if  $x$  and  $y$  belong to the same cluster in  $C$  and  $x \not\sim_C y$ , otherwise. Finally, a CQM is a function that maps clusterings to real numbers.

**Gamma:** This measure was proposed as a CQM by (Baker & Hubert, 1975) and it is the best performing measure in (Milligan, 1981). Let  $d^+$  denote the number of times that a pair of points that was clustered together has distance smaller than two points that belong to different cluster, whereas  $d^-$  denotes the opposite result.

Formally, let  $d^+(C) = |\{\{x, y, x', y'\} \mid x \sim_C y, x' \not\sim_C y', d(x, y) \leq d(x', y')\}|$ , and  $d^-(C) = |\{\{x, y, x', y'\} \mid x \sim_C y, x' \not\sim_C y', d(x, y) \geq d(x', y')\}|$ . The *Gamma* measure of  $C$  is  $\frac{d^+(C) - d^-(C)}{d^+(C) + d^-(C)}$ .

**Silhouette:** The Silhouette measure was defined by (Rousseeuw, 1987). Silhouette is the default clustering quality measure in MATLAB.

Let  $\text{dist}(x, C_i) = \text{avg}_{y \in C_i} d(x, y)$ . The *silhouette* of a point  $x$  with respect to clustering  $C$  is  $S(x, C) = \frac{\min_{j \neq i} \text{dist}(x, C_j) - \text{dist}(x, C_i)}{\max(\min_{j \neq i} \text{dist}(x, C_j), \text{dist}(x, C_i))}$  where  $x \in C_i$ . The *silhouette* of a clustering  $C$  is  $\text{sum}_{x \in X} S(x, C)$ .

**Dunn's Index:** Dunn's Index (Dunn, 1974) compares the maximum within-cluster distance to the minimum between-cluster distances. *Dunn's Index* of  $C$  is  $\frac{\min_{x \not\sim_C y} d(x, y)}{\max_{x \sim_C y} d(x, y)}$ .

**Average Between and Average Within:** The Average Between and Average Within measures evaluate the between-cluster separation and within-cluster homogeneity, respectively. The *average between* of  $C$  is  $\text{avg}_{x \not\sim_C y} d(x, y)$ . The *average within* of  $C$  is  $\text{avg}_{x \sim_C y} d(x, y)$ .

**Calinski-Harabasz:** The Calinski-Harabasz measure (Caliński & Harabasz, 1974) makes use of cluster centers. Let  $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$  denote the center-of-mass of cluster  $C_i$ , and  $\bar{x}$  the center-of-mass of  $X$ . Let  $B(C) = \sum_{C_i} |C_i| |c_i - \bar{x}|^2$  and  $W(C) = \sum_{C_i} \sum_{x \in C_i} |x - c_i|^2$ . The *Calinski-Harabasz* of  $C$  is  $\frac{n-k}{k-1} \cdot \frac{B(C)}{W(C)}$ .

**Weighted inter-intra:** The weighted inter-intra measure is proposed by (Strehl, 2002). It compares the homogeneity of the data to its separation. Let  $\text{intra}(C_i) = \text{avg}_{x, y \in C_i} d(x, y)$  and  $\text{inter}(C_i, C_j) = \text{avg}_{x \in C_i, y \in C_j} d(x, y)$ . The *Weighted inter-intra* of a clustering  $C$  is  $(1 - \frac{2k}{n}) \cdot (1 - \frac{\sum_i \frac{1}{n-|C_i|} \sum_{j \neq i} \text{inter}(C_i, C_j)}{\sum_i \frac{1}{|C_i|-1} \text{intra}(C_i)})$ ,

where  $n$  is the number of points in the dataset.

## Methods

We ran two groups of human subjects and a group of clustering quality measures on a partition evaluation task. Our human subjects were divided into a novice group with little or no knowledge of clustering methods and an expert group with detailed knowledge of clustering methods.

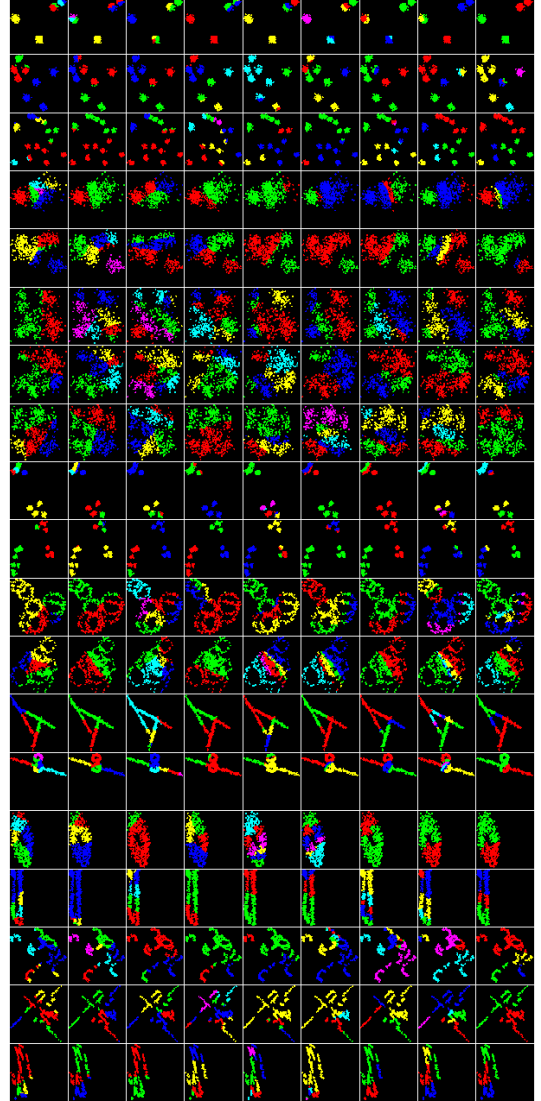


Figure 1: All stimuli. Datasets are in rows; partitions are in columns.

## Human subjects and stimuli

Twelve human subjects were recruited for this project as the novice group, 9 female and 3 male, with an average age of 20.3 years. The novice subjects have no previous exposure to clustering. The expert group consists of 5 people and includes the authors of this paper. All experts have studied clustering in an academic setting, and 4 have done research on the subject.

We used 19 different two dimensional datasets to generate our clustering stimuli, drawn from (Lewis, 2009), and chosen to represent a range of dataset types including mixtures of Gaussians and datasets with hierarchical structure. In order to maintain responsiveness of the stimulus presentation interface, we subsampled 500 points randomly from each dataset. We use synthetic datasets in order to better generate a wide range of stimuli, and our datasets are 2D to facilitate visualization.

Each dataset is randomly clustered nine times in the following manner. For each of the nine clusterings, we first draw the number of partitions,  $k$ , from a uniform distribution over the integers 2 to 6. Second we choose cluster centroids using two strategies: for four of the clusterings we randomly select  $k$  centroids from the original dataset, and for five of the clusterings we select  $k$  centroids from a Laplacian Eigenmap embedding of the data. Finally we color points based on the identity of their nearest centroid in the appropriate space. The goal of this approach is to create stimuli with varied clustering quality.

Each trial consisted of all nine different partitions of the same dataset randomly arranged per trial in a 3 by 3 grid (see Figure 1 for a visualization of all the stimuli). The datasets were shown as scatter plots with colored points on a black background to reduce brightness-related eye strain. For novice subjects, trials were organized into three blocks of 19, where each dataset appeared once per block and the order of the datasets within each block was randomized. Expert subjects were tested on one block of non-randomized datasets. We instructed subjects to choose the two best partitioned displays and the one worst partitioned display from the nine available on every trial (leaving six displays implicitly chosen as neutral).

## Analysis

We analyzed our novice subjects for internal consistency of their positive and negative classifications across blocks and found that even our least consistent subject performed well above chance. We did not exclude any subjects due to inconsistency and we did not analyze internal consistency for experts as they were only tested on one block.

To analyze consistency across subjects we use Fleiss'  $\kappa$  (Fleiss, 1971) and include neutral responses. Fleiss'  $\kappa$  measures the deviation between observed agreement and the agreement attributable to chance given the relative frequency of classifications and normalized for the number of raters. Neutral classifications are twice as frequent as non-neutral, and positive classifications are twice as frequent as negative classifications, so the compensation for relative frequency in Fleiss'  $\kappa$  makes it well-suited to our data. In addition, we perform a consistency analysis on the clustering quality measures by discretizing their classifications in a manner similar to the human data.

We analyze the relationship between novice classifications, expert classifications and clustering quality measures by calculating the Pearson's correlation coefficient,  $\rho$ , between

classifications. To make the responses as comparable as possible we normalize response vectors to a length of one within each dataset. Human subjects have to classify two positive and one negative partition per dataset, even if every partition is quite bad, so by normalizing within dataset we make the CQM responses similar in structure—partitions are judged only relative to other partitions within a dataset.

Because cluster centroids are chosen randomly, increasing  $k$  is likely to increase the chance of getting an undesirable partition (e.g. a partition with very few data points). Additionally, partitions with higher  $k$  require more effort to interpret, and therefore we might expect novice subjects to be biased towards a lower  $k$ . For these reasons our correlations control for  $k$  by partialing out a vector of  $k$  values for each partition. Geometrically this is equivalent to projecting each response vector onto the hyperplane orthogonal to the vector of  $k$  values.

## Results

### Correlation

Table 1 shows correlation coefficients between all measures for both expert and novice responses, with  $k$  factored out. The correlation between expert and novice human positive classifications is higher than the correlation between any CQM and either human positive classification. The negative human classifications have a similarly high correlation. The absolute values of the correlation coefficients between CQMs and expert classifications are strictly greater than or equal to those between CQMs and novice classifications, indicating a closer relationship between expert strategies and the dataset characteristics summarized by the CQMs when  $k$  is factored out.  $k$  itself correlates very strongly with the novices and less so with the experts. Silhouette provides the best overall correlation with expert classifications, and Avg Within provides the best overall correlation with novice classifications (save  $k$ ).

### Consistency

The most undesirable form of inconsistency across subjects or CQMs is both positive and negative responses to the same stimulus. For experts, stimuli with a number of positive classifications 3 or higher never receive a negative classification, and only once does this occur for stimuli with 2 positive responses. In contrast the CQMs exhibit much more disagreement and novices seem to fall somewhere in between. The quantitative measure  $\kappa$  bears this out: CQMs score 0.128, novices score 0.183 and experts score 0.213.  $\kappa$  ranges from  $-1$  to  $1$ , with  $-1$  representing complete disagreement,  $1$  representing complete agreement and  $0$  representing the amount of agreement expected by chance. While there is no standard significance test for differences in  $\kappa$ , the rating scale suggested by Landis and Koch (Landis & Koch, 1977) would characterize the CQM and novice rater groups each as in slight agreement, and the expert raters as in fair agreement. To test whether any one measure was significantly harming CQM consistency we left each out in turn from the analysis



Table 1: Correlation coefficients between human responses and CQMs with  $k$  factored out (except for the  $k$  column). Text in bold (excluding  $k$  column) if  $p < .0025$  after Bonferroni correction for  $n = 20$  comparisons per subject group and  $\alpha = .05$ .

$\rho$	Expert Positive	Expert Negative	Novice Positive	Novice Negative	Gamma	Silhouette	Dunn	Avg Within	Avg Btw	CH	W-Inter/Intra	$k$
Expert Pos	1	<b>-.35</b>	<b>.56</b>	-.19	-.15	<b>.46</b>	<b>.40</b>	<b>-.39</b>	<b>.34</b>	<b>.44</b>	.19	-.43
Expert Neg		1	-.13	<b>.44</b>	.09	<b>-.27</b>	-.12	<b>.44</b>	-.18	<b>-.36</b>	<b>-.30</b>	.32
Novice Pos			1	-.04	-.13	<b>.39</b>	<b>.40</b>	-.20	.23	<b>.30</b>	.04	-.73
Novice Neg				1	.08	<b>-.27</b>	.01	<b>.30</b>	-.07	<b>-.25</b>	<b>-.27</b>	.71

Table 2: A summary of the number of partitions for which a high degree of agreement was achieved by the raters. If a partition is classified as negative or positive by 80% - 100% of raters, it would be added to the top row, and similarly for the 60% - 79% bucket. The total possible number of agreed upon partitions is 57 (19 datasets \* 3 possible negative/positive responses to partitions per dataset).

% Majority	Experts	Novices	CQMs
80% - 100%	19	3	1
60% - 79%	20	11	7
Sum $\geq 60\%$	39	14	8

and found values ranging from 0.098 to 0.172, which is in line with the CQM consistency with no measure left out, and in every case less consistent than the novice subjects. Finally, we left out both Avg Within and Between, since they measure quality on intentionally simple and distinct dimensions, and found a  $\kappa$  of 0.110.

In Table 2 we summarize the consistency of experts, novices and cluster quality measures. It shows how often certain percentages of raters are able to agree on negative or positive classifications for particular stimuli. Experts agree over 60% of the time on more samples (39), than do novices (14) or CQMs (8).

## Discussion

### Comparing human evaluations with CQMs

Some natural quality measures have low correlation with human evaluations. Most notably, Gamma has low correlation with both positive and negative human classifications for both novices and experts. W-Inter/Intra has low correlation with the positive classifications of both subject groups. This shows that a natural mathematical formalization does not suffice to guarantee that the evaluations of clusterings produced using the CQM will seem natural to humans.

There are also CQMs that correlate well with human evaluations. Of these the most notable are CH and Silhouette. These two popular measures correlate well with both expert

and novice evaluations, on both the positive and negative classifications.

### Comparing experts with novices

Evaluations of experts and novices have a correlation score of 0.56, higher than the correlation of any CQM with any of the two subject groups. This suggests that a cluster evaluation skill is present in the general population.

On the other hand, we observe some interesting differences between the two groups of subjects. One of the most notable differences between experts and novices is that, while both groups prefer clusterings with fewer clusters, novices rely much more heavily on this heuristic.

Experts seem to use more, and more complex strategies than novices. Positive expert classifications correlate well with two more measures than positive novice classifications. No measure considered correlates better with novice classifications than with expert classifications, and in the great majority of cases the correlation is higher with expert classifications.

With a cover of at most six domain elements on any input dataset (see Definition 5 below), Dunn’s measure is (according to this measure of complexity) the simplest measure that we explore. While positive expert evaluations correlate well with five distinct measures, Dunn’s measure is one of three measures that correlate well with novice evaluations. This further illustrates that novices rely on fewer simpler strategies, which indicates that expert evaluations may be more sophisticated and reliable.

### Consistency

Given the difficulty of knowing whether humans or CQMs do a reasonable job of evaluating clustering quality, one might hope that at least they are consistent across individuals (or measures). Consistency indicates that some repeatable process is at work and that its repeatability is minimally affected by changes in input. Of course CQMs are perfectly consistent on a within measure basis—given the same partition they will always report the same quality—and one is tempted to suggest that between measure consistency is an unfair point of comparison; aren’t all the measures using quite different evalu-

ative procedures, and didn't we select them to be distinct? We did, but CQMs purport to evaluate clustering quality in general. Insofar as they evaluate this more nebulous property they should be consistent, even if their methods differ. As it turns out, they are somewhat consistent with each other, just not as consistent as humans. Further, the consistency story did not vary when we tested all the leave-one-out subsets of CQMs, indicating that CQM consistency is not being skewed by just one divergent measure.

Human experts are the most consistent group in this study. This lends empirical support to the common practice of seeking human visual evaluations of partition quality. Novices are less consistent, and as discussed above there is evidence that the evaluations they provide are less sophisticated. Despite the unfavorable comparison to experts, it is notable that subjects with no formal knowledge of cluster analysis are able to respond more consistently than a set of CQMs. This lends credence to the notion that our ability to evaluate partitions is acquired in the natural course of visual development.

### A Property-Based Taxonomy of CQMs

In the absence of formal guidelines for CQM selection<sup>2</sup>, in particular for selecting a versatile set of CQMs, we develop a property-based framework for distinguishing CQMs based on such a framework for clustering algorithms discussed in (Ackerman, Ben-David, & Loker, 2010b) (also see (Bosagh-Zadeh & Ben-David, 2009) and (Ackerman, Ben-David, & Loker, 2010a)). The framework consists of identifying natural properties of CQMs and classifying measures based on the properties that they satisfy. For the purposes of our study we use this framework to select meaningfully versatile CQMs. This taxonomy may have independent interest for choosing CQMs in other settings. Note that these properties are descriptive only, and not necessarily desirable.

Our taxonomy of CQMs follows a line of work on theoretical foundations of clustering beginning with the famous impossibility result by (Kleinberg, 2003), which showed that no clustering function can simultaneously satisfy three specific properties. (Ackerman & Ben-David, 2008) reformulate these properties in the setting of CQMs, and show that these properties are consistent and satisfied by many CQMs. We follow up on (Ackerman & Ben-David, 2008) by studying natural properties that can be used to distinguish between CQMs.

In Table 3, we present a taxonomy of our seven clustering quality measures. Each property, defined below, aims to capture some fundamental feature that is satisfied by some measures.

#### Normed clustering quality measures

A clustering quality measure  $m$  takes a domain set  $X$ , a distance function  $d$  over  $X$ , and a clustering  $C$  of  $X$ , and outputs a non-negative real number. Some quality measures are defined

<sup>2</sup>Although there are no formal guidelines for CQM selection, some interesting heuristics have been proposed, see, for example, (Vendramin, Campello, & Hruschka, 2009).

Table 3: A taxonomy of the seven quality measures used in the study.

	Gamma	Silhouette	Dunn	Avg Within	Avg Btw	CH	W-Inter/Intra
Order-consist.	✓	X	X	X	X	X	X
Sep-invariant	X	X	X	✓	X	X	X
Hom-invariant	X	X	X	X	✓	X	X
Bounded	✓	✓	X	X	X	X	X
Constant Cover	X	X	✓	X	X	X	X
Norm-based	X	X	X	X	X	✓	X

over normed vector spaces. *Normed CQMs* take a quadruple of the form  $(V, X, C, \|\cdot\|)$ , where  $V$  is a vector space,  $X$  a finite subset of  $V$ , and  $\|\cdot\|$  is a norm over  $V$ . Normed CQMs can rely on centers-of-mass of clusters that are not necessarily in  $X$ , but are part of the vector-space  $V$ . Observe that the centers-of-mass are not defined for un-normed CQMs. We define the properties for CQMs in general, but one can apply any property to a normed CQM by using the norm to define the distance function. That is, set  $d(x, y) = \|x - y\|$  for all  $x, y \in X$ .

#### Invariance and consistency properties

Invariance properties describe changes to the underlying data that do not affect the quality of a clustering. Consistency properties describe similarity conditions under which clusterings have similar quality. We propose two new invariance properties.

**Definition 1** (Separation Invariance). A CQM  $m$  is separation-invariant if for all  $X$  and distance functions  $d$  and  $d'$  over  $X$  where  $d(x, y) = d'(x, y)$  for all  $x \sim_C y$ ,  $m(C, X, d) = m(C, X, d')$ .

A separation invariant CQM is not affected by changes to between-cluster distances. Conversely, homogeneity invariant CQMs depend only on between-cluster distances, and are invariant to changes to within-cluster distances.

**Definition 2** (Homogeneity Invariance). A CQM  $m$  is homogeneity-invariant if for all  $X$  and distance functions  $d$  and  $d'$  over  $X$  where  $d(x, y) = d'(x, y)$  for all  $x \not\sim_C y$ ,  $m(C, X, d) = m(C, X, d')$ .

Observe that separation-invariance and homogeneity-invariance can also be viewed as consistency properties. An additional consistency property, order consistency, is an adaptation of an analogous property of clustering functions presented in (Jardine & Sibson, 1971). Order consistency describes CQMs that depend only on the order of pairwise distances.

**Definition 3.** A CQM  $m$  is order consistent if for all  $d$  and  $d'$  over  $X$  such that for all  $p, q, r, s \in X$ ,  $d(p, q) < d(r, s)$  if and only if  $d'(p, q) < d'(r, s)$ ,  $m(C, X, d) = m(C, X, d')$ .

## Domain and range properties

A bounded range can aid in interpreting the results of a CQM, in particular if the bounds are attainable by some clusterings.

**Definition 4** (Bounded). A CQM  $m$  is bounded if there exist datasets  $X_1$  over  $d_1$  and  $X_2$  over  $d_2$ , and clusterings  $C_1$  of  $X_1$  and  $C_2$  of  $X_2$ , so that  $m(C_1, X_1, d_1) \leq m(C, X, d) \leq m(C_2, X_2, d_2)$  for all  $C, X$ , and  $d$ .

Our next property describes the quantity of domain elements that effect the CQM. First, we introduce the notion of an  $m$ -cover of a clustering, a subset of the domain which has the same quality as the entire set. For clustering  $C$  of  $X$ , and  $X' \subseteq X$ , let  $C|X'$  denote the clustering  $C'$  of  $X'$  where for all  $x, y \in X'$ ,  $x \sim_{C'} y$  if and only if  $x \sim_C y$ .

An  $m$ -cover of clustering  $C$  of  $X$  is any set  $R \subseteq X$ , so that  $m(X, k) = m(R, C|R)$ . We define clustering quality measures that have a constant size cover for all clusterings.

**Definition 5** (Bounded Cover). A CQM  $m$  has bounded cover if there exists a constant  $r$  so that for every data set  $X$  and clustering  $C$  of  $X$ , there exists an  $m$ -cover of  $C$  of cardinality at most  $r$ .

CQMs that have a bounded cover search the domain space for some local features, ignoring most of the information in the dataset.

## Conclusions

We perform an empirical study comparing human evaluations of clustering with formal clustering quality measures. To select a versatile set of CQMs, we develop a theoretical property-based taxonomy of CQMs. Our study shows that some CQMs with seemingly natural mathematical formulations yield evaluations that disagree with human perception. On the other hand, we identify CQMs (CH and Silhouette) that have significant correlation with human evaluations.

Our consistency analysis reveals that even novices are at least as consistent as a broad set of CQMs, and perhaps more consistent. We also find significant correlations between the evaluations of expert and novice subjects. This lends support to the common practice of seeking human visual evaluations of partition quality. If one needs to evaluate a very large number of partitions it may be reasonable to use human computation via a service such as Mechanical Turk to rank partitions efficiently (or at least throw out the really bad ones). Finally, experts appear to use more sophisticated strategies than novices, indicating that training can improve human clustering evaluation performance.

## Acknowledgments

This work is funded by NSF Grant #SES-0963071, Divvy: Robust and Interactive Cluster Analysis (PI Virginia de Sa). Thanks to Cindy Zhang for valuable code contributions.

## References

Ackerman, M., & Ben-David, S. (2008). Measures of clustering quality: A working set of axioms for clustering. In *Advances in neural information processing systems*.

- Ackerman, M., & Ben-David, S. (2009). Clusterability: A theoretical study. *Proceedings of AISTATS-09, JMLR: W&CP*, 5, 1–8.
- Ackerman, M., Ben-David, S., & Loker, D. (2010a). Characterization of Linkage-based Clustering. In *Proceedings of colt*.
- Ackerman, M., Ben-David, S., & Loker, D. (2010b). Differentiating clustering paradigms: a property-based approach. In *Advances in neural information processing systems*.
- Baker, F., & Hubert, L. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70(349), 31–38.
- Bosagh-Zadeh, B., & Ben-David, S. (2009). A uniqueness theorem for clustering. In *Proceedings of the 25th conference on uncertainty in artificial intelligence, auai press*.
- Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, 3(1), 1–27.
- Dunn, J. (1974). Well-separated clusters and optimal fuzzy partitions. *Cybernetics and Systems*, 4(1), 95–104.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
- Jardine, N., & Sibson, R. (1971). *Mathematical taxonomy*. John Wiley and Sons, Inc., New York.
- Kleinberg, J. (2003). An impossibility theorem for clustering. In *Advances in neural information processing systems 15: Proceedings of the 2002 conference* (p. 463).
- Landis, J. R., & Koch, G. G. (1977, March). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Lewis, J. M. (2009). Finding a better k: A psychophysical investigation of clustering. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (p. 315–320).
- Milligan, G. (1981). A Monte-Carlo study of 30 internal criterion measures for cluster-analysis. *Psychometrika*, 46, 187–195.
- Ng, A. Y., Jordan, M., & Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (pp. 849–856). Cambridge, MA: MIT Press.
- Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53–65.
- Santos, J., & Sá, J. M. de. (2005). *Human clustering on bi-dimensional data: An assessment* (Tech. Rep. No. 1). INEB Instituto de Engenharia Biomédica, Porto, Portugal. Available from [http://www.di.ubi.pt/~lfbaa/entnetsPubs/JMS\\_TechReport2005\\_1.pdf](http://www.di.ubi.pt/~lfbaa/entnetsPubs/JMS_TechReport2005_1.pdf)
- Strehl, A. (2002). Relationship-based clustering and cluster ensembles for high-dimensional data mining.
- Vendramin, L., Campello, R., & Hruschka, E. (2009). *On the comparison of relative clustering validity criteria*. Sparks.

# Learning Cluster Analysis through Experience

Joshua M. Lewis

josh@cogsci.ucsd.edu

Department of Cognitive Science  
University of California, San Diego

Virginia R. de Sa

desa@cogsci.ucsd.edu

Department of Cognitive Science  
University of California, San Diego

## Abstract

The field of machine learning is constantly developing useful new techniques for data analysis, but they are often ignored by researchers outside the field due to unfamiliarity and the difficulty of keeping up with a large body of work. We propose a methodology for training researchers how algorithms work through experience, such that they gain an implicit, rather than explicit, understanding of their function. Thus we combine theory from discovery learning with advanced software and a more educated target population to foster such understanding. We have developed an open source application for exploratory data analysis called Divvy that lets users quickly and visually interact with a range of data analysis techniques. Using a simplified version of Divvy, we find that undergraduate subjects are generally able to learn machine learning concepts through experience, though they have only partial success in applying them.

**Keywords:** unsupervised machine learning; clustering; discovery learning; human computer interfaces

Machine learning has a PR problem. The field has developed many techniques that cluster, classify, or reduce the dimensionality of data, and most techniques could be profitably applied to scientific data sets. Researchers that are not machine learning experts face a daunting question, however—which techniques should I use to analyze my data? Authors proposing a new technique will focus on its strengths over its weaknesses, and most researchers do not want to spend a year reading math papers and becoming a machine learning expert in order to best analyze their data. So too often the analysis technique used is the convenient one (freely available online or as part of a software package), or the traditional one. Researchers miss out on the advances in machine learning, and the machine learning field is not as valuable as it could be to the broader scientific community.

There are two fundamental problems: expertise and access. Gaining expertise is difficult—if a researcher wants to find the right technique for the job, but is unwilling to engage in the time-consuming process of learning the details of every technique, how can they be trained to apply the best one? With the right tools, we believe discovery learning has substantial potential for training researchers. Software that provides direct and intuitive access to the behavior of machine learning algorithms can support the development of a pragmatic (not mathematical) understanding of the algorithms.

As an analogy, baseball players have an excellent idea of how baseballs behave. A baseball's behavior is, of course, governed by the laws of physics and an explicit description of that behavior might be quite complex when spin, deformation, wind and field texture are taken into account. Nevertheless, through extensive experience baseball players acquire an excellent pragmatic understanding of how baseballs

behave, an understanding that one might guess is founded on an implicit learned model of baseball behavior rather than the explicit model a physicist would give. We believe that interactive experience with machine learning techniques can give rise to a similar sort of practical and implicit model of algorithm behavior, and that researchers can use such a model to make informed decisions during data analysis.

Discovery learning is particularly compelling in this context because researchers often do not have the time or inclination to seek out traditional forms of instruction while analyzing data. Tools that support learning on the job are thus necessary and expedient. In this paper we test our hypothesis using a data analysis platform called Divvy that we've developed to provide such an experience that emphasizes speed and visualization.

Gaining access to a wide variety of data analysis techniques is also tricky—it might require technical knowledge (e.g. basic programming skills in whichever languages the techniques are in), or owning proprietary software like Matlab and formatting one's data for it. To that end Divvy is a free, open source project designed around a plugin architecture where machine learning researchers can package their algorithms with intuitive custom UIs that require no programming expertise from users.

In our experiment we give undergraduate subjects interactive experience with two clustering techniques,  $k$ -means (MacQueen, 1967) and single linkage (Johnson, 1967), labeled simply as method A and method B and without any explicit instruction as to their differences. We find that after training almost every subject learns a few relevant facts about A or B or their parameters, and that some subjects appear to be able to apply this knowledge to new analysis contexts.

## Divvy

Data analysis is often a laborious process. A researcher collects data, and then loads it into a software package such as Matlab or R. To apply an algorithm to his or her data, the researcher has to write a command or fill out a dialog box and then wait for processing to finish. Finally, the researcher will use other commands to visualize the algorithm's output. To change a parameter and see the impact it has, this process must be repeated. Some researchers might write a script that runs a set of different parameters and visualizations, and then go out for a coffee and come back to see if the whole endeavor bore any fruit.

This is a tenuous kind of interaction. A baseball, by virtue of being in the real world, provides critical instantaneous

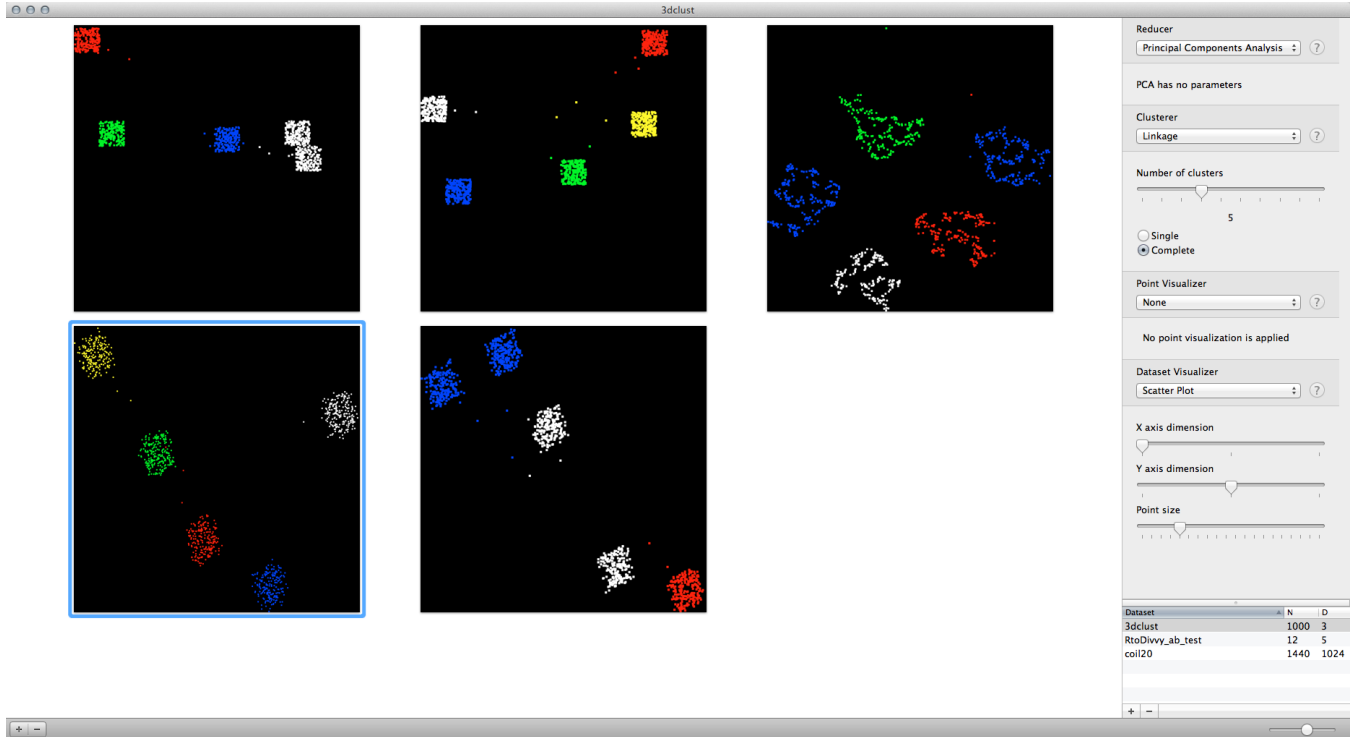


Figure 1: The full Divvy UI. Each visualization represents a different view of the same dataset (generated by combining a dimensionality reduction technique, a clustering technique and a dataset visualizer) and users can set the properties of each view using the tools to the right. A list of datasets resides in the bottom right, allowing the user to switch between them at any time, even while results are computing in the background.

feedback to those interacting with it. In the above process the algorithm does not, and the goal of the Divvy project is to close that gap and provide an interface where visualization happens instantaneously and researchers can tweak parameters and see their effect in real time. In a way, Divvy is providing the human analog to active learning (Cohn, Ghahramani, & Jordan, 1996), where learning algorithms choose which training samples to get based on what they predict to be the most informative. Divvy is similar in spirit to a tool called GGobi (*Ggobi data visualization system*, n.d.), which brings cutting edge methods in high-dimensional data visualization to a user friendly graphical interface but without the strong machine learning component Divvy provides.

Divvy supports four types of plugin: clusterers, reducers, point visualizers and dataset visualizers. Clusterers and reducers represent clustering and dimensionality reduction algorithms, respectively. Point visualizers represent single points in the dataset with visualizations, such as the image of a handwritten digit, and dataset visualizers represent the entire dataset, for example with a scatter plot. Each view of the dataset (of which a user can have a practically unlimited number) represents a combination of these four plugin types, so a user can compare, e.g.,  $k$ -means in the first two PCA dimensions with spectral clustering in the same embedding.

Divvy achieves real time responsiveness on many datasets

through parallel computing. Many personal computers (and all Macs) ship with multi-core processors (CPUs), as well as graphics processors (GPUs) that can be used for general purpose computation. High performance computing research has so far focused on how these hardware resources can make very large problems tractable (Raina, Madhavan, & Ng, 2009). With Divvy, we are using these technologies to make medium problems very fast—fast enough to feel real time, and to invite the exploratory interaction that we believe leads to learning. Even if an algorithm takes a while to run, users can continue to use Divvy to perform other analyses on the same dataset or even on others while they wait. Our UI design puts a focus on visualization, allowing users to simultaneously visualize many perspectives on their data. Algorithm parameters are controlled with standard UI elements (such as sliders or check boxes) rather than having to be specified with code. See Figure 1 for the full Divvy UI and Figure 2 for the simplified version of the UI we used in this experiment.

Divvy does not attempt to replace a user’s data analysis workflow, but rather to be a part of it. It can export data and visualizations in standard formats and import from other popular tools. Divvy, its source code, sample datasets, and R/Matlab data importers are freely available from <http://divvy.ucsd.edu> and on the Mac App Store.

## Discovery Learning

Our study represents a form of discovery learning (Bruner, 1961), also known as constructivist, inquiry or experiential learning. In discovery learning students learn material independently of explicit instruction by exploring environments, solving problems, or performing experiments. Several researchers have called into question the effectiveness of pure discovery learning, suggesting that active guidance from an instructor (Mayer, 2004), or a sufficient foundation of domain knowledge (Kirschner, Sweller, & Clark, 2006) are required for constructivist approaches to be successful.

Our target audience for Divvy differs from the traditional subjects used in studies of discovery learning. We intend for Divvy to be used by researchers such as faculty and graduate students who have a highly sophisticated understanding of their problem domain. Further, they are accustomed to self-directed learning. In this sense, though they do not have a detailed understanding of machine learning, they do have a foundation of domain knowledge with which they can determine whether the output of a machine learning algorithm is appropriate or not. In addition, Divvy provides some forms of active guidance. Divvy plugin UIs default to reasonable ranges for parameter settings and every plugin can specify a help link that takes users to a relevant resource on the web, such as a paper describing the method or a relevant Wikipedia article.

For these reasons we believe Divvy to be more likely to succeed than other examples of discovery learning that focus on elementary-, middle-, and high-school populations with less active guidance. In this study we use an undergraduate population that is generally less knowledgeable than our target population, representing a more challenging domain than that which Divvy will have in the wild. If undergraduates are able to learn machine learning concepts with Divvy then graduate students, postdocs, and faculty likely can as well.

As outlined above, we believe that guided learning is not necessarily practical or expedient for our target population. So while explicit instruction would certainly allow subjects to learn machine learning concepts, we do not compare Divvy to that form of learning in this paper. Here we focus on what, if anything, subjects are able to learn from a version of our more pragmatic approach to solving machine learning's PR problem.

## Methods

We recruited 22 undergraduate subjects for this experiment. Subjects received course credit for participation. One subject was excluded from the study after he indicated at the end during the interview segment that he must not have understood the instructions, and so we analyzed the data from a grand total of 21 subjects.

Each subject performed 36 trials, which were split into two 18 trial blocks, a training block and a testing block. In both blocks, subjects use the sliders to change the number of clusters,  $k$ , and the relative weighting of the horizontal and vertical

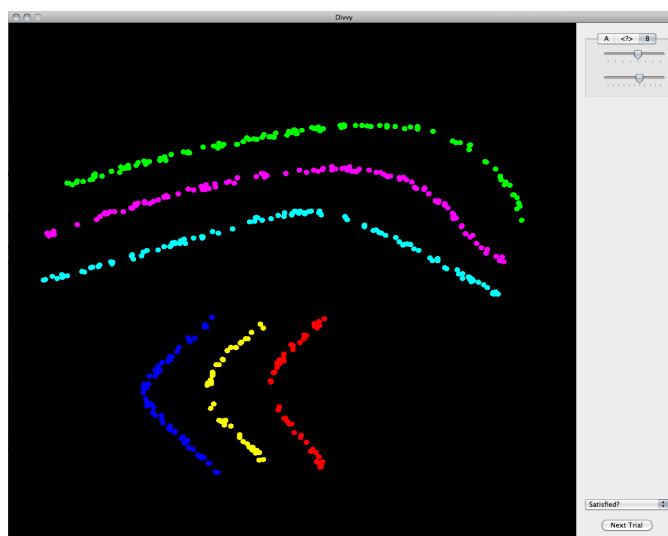


Figure 2: The Divvy UI used in this experiment. The tabs at the top right select method A ( $k$ -means) or B (single linkage), and the sliders below control the number of clusters and the relative weighting of the horizontal and vertical axes. Subjects indicate their satisfaction with a particular partitioning using the dropdown menu above the next trial button at the bottom right.

cal axes in order to best group the points in each stimulus (one stimulus per trial) and then indicate their satisfaction with the result (ranging from 1, not satisfied, to 7, very satisfied). In the training block, subjects use both A and B ( $k$ -means and single linkage, respectively) to group the points, and are required to arrive at a solution for each method. In the testing block, neither A nor B are initially selected and subjects must choose which method they want to use for that trial. Once the choice is made they cannot switch. We divided subjects into two groups of 10 and 11. One group's training set was the other's testing set, and vice versa. At the end of the two blocks, subjects filled out an interview form that assessed their knowledge. The eight interview questions were as follows (where circles means the individual data points):

1. What did you feel like method A was doing?
2. What organizations of circles was method A good for grouping?
3. What did you feel like method B was doing?
4. What organizations of circles was method B good for grouping?
5. Did you have a preference between A and B?
6. Why or why not?
7. What did the first (top) slider do?
8. What did the second (bottom) slider do?

We instructed subjects to do their best to learn what A, B and the sliders were doing in the first half of the experiment, as they would need to use that knowledge during the second half. We also made clear that not every stimulus could be ideally grouped with both A and B, and that if they did not like a solution they could just indicate dissatisfaction using the dropdown above the next trial button. We provided two helper images along with the instructions. One showed a well-separated mixture of Gaussians where each Gaussian had its own color. This was held up as a positive example. The second showed two circular groups split in half with color, which was considered a negative example. Beyond these very simple prompts (shown in Figure 3) we did not bias the subjects as to what a group should be.

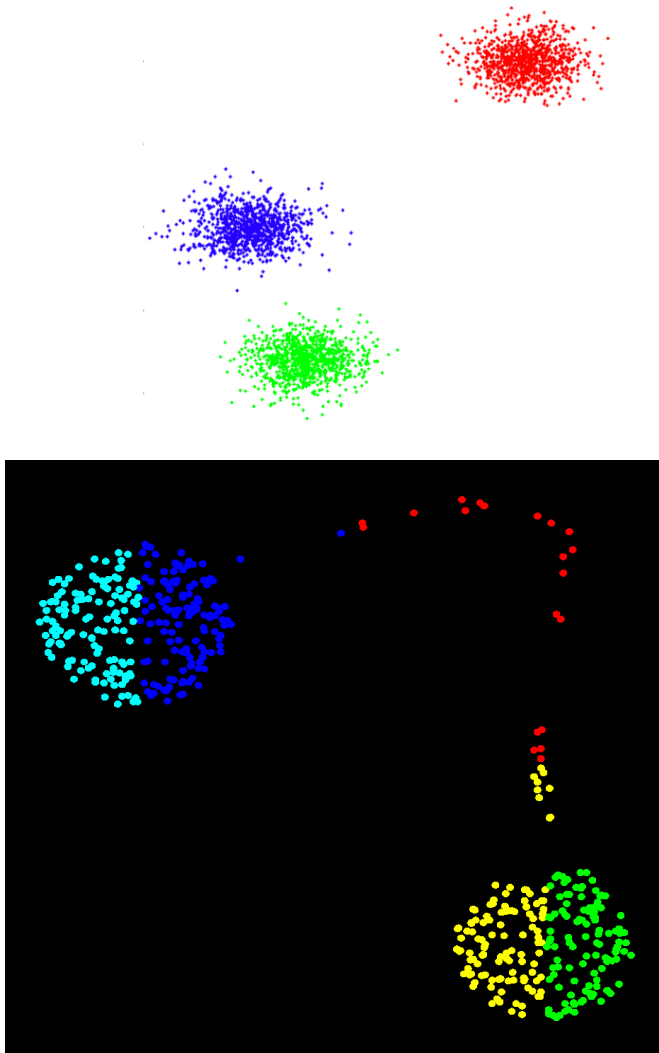


Figure 3: Sample images to give subjects basic guidance on good groups (top) versus bad groups (bottom).

The 36 stimuli fall into three categories, those where A is most effective (14), those where B is most effective (15), and

those where A and B are similarly effective (7). We created all 36 stimuli by hand in order to ensure that the first two categories had sufficient membership. Stimuli ranged from complex collections of lines, rings and spirals to connected and disconnected blobs to uniform noise. While these are not real data, so to speak, they provide us with a solid foundation on which to train and judge our subjects that real data would not necessarily provide. Additionally, most meaningful real data are more than two dimensional, and while the full version of Divvy uses dimensionality reduction techniques and multiple views to visualize such data, those techniques are not relevant to our core question in this experiment concerning cluster analysis.

Divvy records every method and parameter combination subjects try over the course of the experiment, including their final grouping and satisfaction. We use these data in concert with interview responses to determine what subjects were able to learn from their experience. From the Divvy records we extract two variables per subject, the total number of different algorithm and parameter settings queried in the training period (the number of “moves”), and the percent of correct method (A or B) choices (with any parameter choice) in the testing period, out of the stimuli for which there is a preferred method. From the interviews we code for understanding of seven possible concepts. The seven possible concepts are as follows:

1. The first slider controls the number of colors (i.e. clusters).
2. The second slider controls the orientation of the boundary between clusters.
3. *k*-means works well on blobs of points (compact regions).
4. Single linkage works well on extended shapes like lines or rings (non-compact regions).
5. *k*-means can work when there is no space separating clusters.
6. Single linkage works best when there is lots of space between clusters.
7. *k*-means tends to divide the points into evenly sized groups, whereas single linkage can make large and small groups.

We hypothesize that there will be a positive correlation between understanding a greater number of concepts and selecting the correct method. We also report correlations between these measures and the number of moves subjects take. To gain an understanding of the relative difficulty of learning the concepts we report in detail the concepts learned on a per-subject basis. Finally, we compare subject satisfaction when using the correct method on a stimulus versus the incorrect method. This test indicates whether subjects recognize when the partitions are not ideal. If the subjects cannot distinguish good partitions from bad given their intuition and the instructive samples, then there is not an opportunity for learning.



## Results

In Table 1 we summarize the contents of each subject's interview, using the seven concepts described above. Nineteen of 21 of the subjects learned at least one concept, and 15 of the subjects learned at least one concept excluding the simplest one (the function of the first slider). On average subjects learned 2.4 concepts over the course of the study.

Table 1: A summary of the concepts subjects learned. Subjects in bold chose the correct method for over 70% of stimuli in the test block.

Subjects	1st Slider	2nd Slider	k-means Blobs	Single Linkage Shapes	k-means No Separation	Single Linkage Separation	k Even vs SL Uneven	Sum
1								0
2								0
3	✓				✓	✓	✓	4
4	✓	✓		✓				3
5	✓				✓			2
<b>6</b>	✓	✓		✓				<b>3</b>
<b>7</b>	✓		✓	✓				<b>3</b>
<b>8</b>	✓			✓			✓	<b>3</b>
9	✓							1
10	✓						✓	2
11							✓	1
12	✓		✓	✓				3
<b>13</b>	✓		✓	✓				<b>3</b>
<b>14</b>	✓							<b>1</b>
15						✓	✓	2
<b>16</b>	✓				✓	✓	✓	<b>4</b>
17	✓							1
18	✓		✓	✓	✓	✓		5
<b>19</b>	✓		✓	✓			✓	<b>4</b>
20	✓	✓						2
21	✓				✓		✓	3
Sum	17	3	5	9	5	4	7	

The number of concepts learned correlates positively, but only as a trend, with both percent correct ( $p = .29$ ,  $p < .10$ ) and number of moves ( $p = .34$ ,  $p < .07$ ). Percent correct and number of moves are not correlated ( $p = -.22$ ,  $p < .84$ ). In Figure 4 we show scatter plots of the pairwise comparisons between these variables.

For stimuli with a correct answer where the subject used the correct method, we had 470 satisfaction ratings with  $\mu = 5.88$ ,  $\sigma = 1.37$ . For stimuli with a correct answer where the subject used the incorrect method, we had 444 satisfaction ratings with  $\mu = 4.94$ ,  $\sigma = 1.77$ . A t-test indicated a sig-

nificant  $p < .01$  effect of correct versus incorrect method on satisfaction, indicating that subjects were in general able to judge some difference between good and bad partitions.

## Discussion

Almost every subject learned about cluster analysis through their experience—over half learned three concepts or more. Giving researchers expertise and access through tools like Divvy promises to encourage and improve the application of machine learning techniques in other fields.

Nevertheless, some subjects had difficulty using the knowledge they acquired to make good data analysis decisions. Though subjects explored quite a bit during the training phase (an activity that showed a trending correlation with concept learning) they did not necessarily parlay that experience into better performance. So while we are pleased that subjects demonstrated concept learning in the interviews, we would like to investigate why they had trouble applying it. The subjects were overall less satisfied when using the incorrect method, which indicates that evaluative confusion was not the primary culprit.

Given that the core audience for Divvy is composed of graduate students, postdocs, and faculty, we would like to perform a follow-up study with that audience. While undergraduates serve as a useful lower bound, so to speak, for testing learning with Divvy, our target population is likely more motivated, more familiar with data analysis tasks, and in possession of greater domain knowledge.

The process of crystallizing the implicit knowledge gained during the experiment in the interview might help subjects make better decisions. To test this, a future experiment could place the interview between the training and test blocks. If this results in better performance, it would indicate that having to articulate knowledge assists concept crystallization and application, and that the subjects are in a sense still learning when they fill out the interview.

We do not think a comparison to traditional guided learning is useful since our target population will rarely have the time or inclination to seek out explicit instruction. However, we would be interested in comparing our results to other forms of discovery learning where the interaction between subject and software is modified. We believe that self-directed exploration with instantaneous feedback is valuable and we would like to compare our results with, e.g., simply showing subjects a set of partitions and their associated methods and parameter values without allowing them to choose parameters, or putting a delay between parameter changes and result visualization. These modifications would move the experimental context closer to traditional machine learning approaches where the training data are fixed (as opposed to the active learning paradigm mentioned earlier). It would also correspond to writing a script to run through a set of parameter settings and visualizations while one goes out for coffee, and then interpreting when one returns.

Our results provide compelling evidence that undergradu-

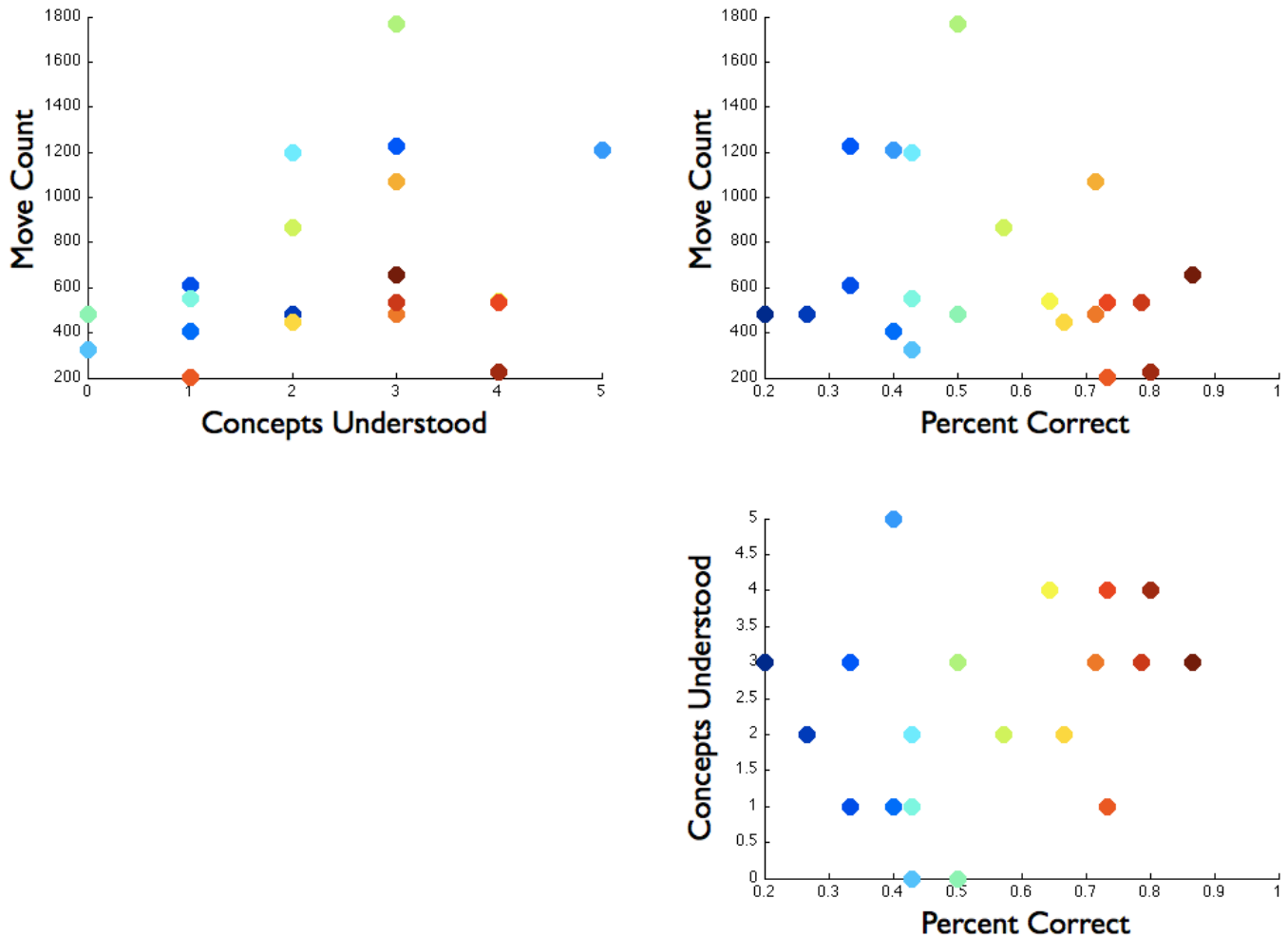


Figure 4: Scatter plots of the three main variables. The points are colored from dark blue to dark red based on percent correct.

ate subjects can learn useful concepts about machine learning algorithms just by interacting with them. This leads one to suspect that the target population for this work, practicing researchers, will be able to do so as well. Subjects do not reliably apply these concepts when tested, and additional study is required to determine why this is, and how to better support the discovery and application of machine learning concepts.

### Acknowledgments

This work is funded by NSF Grant #SES-0963071, Divvy: Robust and Interactive Cluster Analysis (PI Virginia de Sa). Thanks to Cindy Zhang for valuable code contributions.

### References

- Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review*, 31, 21–32.
- Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *CoRR*, cs.AI/9603104.
- Ggobi data visualization system. (n.d.). <http://www.ggobi.org>.
- Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3), 241–254.
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75–86.
- MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. Proc. 5th Berkeley Symp. Math. Stat. Probab., Univ. Calif. 1965/66, 1, 281–297.
- Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? the case for guided methods of instruction. *American Psychologist*, 59, 14–19.
- Raina, R., Madhavan, A., & Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning* (pp. 873–880). New York, NY, USA: ACM.

# The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains

Daniel Leyzberg (daniel.leyzberg@yale.edu)  
Samuel Spaulding (samuel.spaulding@yale.edu)  
Mariya Toneva (mariya.toneva@yale.edu)  
Brian Scassellati (scasz@cs.yale.edu)

Department of Computer Science, Yale University  
51 Prospect St., New Haven, CT 06511, USA

## Abstract

We present the results of a 100 participant study on the role of a robot's physical presence in a robot tutoring task. Participants were asked to solve a set of puzzles while being provided occasional gameplay advice by a robot tutor. Each participant was assigned one of five conditions: (1) *no advice*, (2) *robot providing randomized advice*, (3) *voice of the robot providing personalized advice*, (4) *video representation of the robot providing personalized advice*, or (5) *physically-present robot providing personalized advice*. We assess the tutor's effectiveness by the time it takes participants to complete the puzzles. Participants in the *robot providing personalized advice* group solved most puzzles faster on average and improved their same-puzzle solving time significantly more than participants in any other group. Our study is the first to assess the effect of the physical presence of a robot in an automated tutoring interaction. We conclude that physical embodiment can produce measurable learning gains.

**Keywords:** Robotics; Computer Science; Tutoring

## Introduction

What kinds of human-robot interactions benefit from the physical embodiment of a robot? For human-robot interactions that require manipulating the physical world, a physical robot is a necessity, but for those interactions where physical embodiment is optional, when is an embodied robot more useful than an on-screen agent?

In this study, we explore the differences in task performance of participants engaged in a cognitive learning task in which a robot acts as a tutor. Participants were asked to play a puzzle game while receiving strategy advice from either: a physically-present robot, a video of the same robot, its disembodied voice, a robot giving randomized advice, or no agent at all. We use the resulting data to draw conclusions about the effect of embodiment in robot tutoring tasks.

Previous work has investigated the social influence of a robot's embodiment. Does a robot engender more trust, more compliance, more engagement, or more motivation by its physical presence, more so than an on-screen agent or a video representation of a robot would? Such questions have been explored via two methodologies: self-report measures and task-performance measures. Using self-report measures, Kidd and Breazeal (2004) found that a physically-present robot was perceived as more enjoyable, more credible, and more informative than an on-screen character in a block-moving task. In Wainer, Feil-Seifer, Shell, and Mataric (2007), an embodied robot was rated as more attentive and more helpful than both a video representation of the robot

and a simulated on-screen robot-like character. Tapus, Tapus, and Mataric (2009) found that individuals suffering from cognitive impairment and/or Alzheimer's disease reported being more engaged with a robot treatment than a similar on-screen agent treatment.

Kiesler, Powers, Fussell, and Torrey (2008) used task-performance measures to find that participants who received health advice from a physically-present robot were more likely to choose a healthy snack than participants who received the same information in robot-video or on-screen agent conditions. In Bainbridge, Hart, Kim, and Scassellati (2008), a physically-present robot yielded significantly more compliance to its commands than a video representation of the same robot.

No previous work has investigated whether learning outcomes are affected by a robot's physical presence. The closest related work is in Intelligent Tutoring Systems (ITSs), which are educational computer programs that produce individualized lessons, advice, and questions usually in a workbook-style or quiz-style environment (Nkambou, Bourdeau, & Psyché, 2010). A parallel notion of embodiment called "the persona effect" exists in ITS research. (See Dehn and Van Mulken (2000) for an overview.) The persona effect is the impact, if any, that an on-screen character has on students using an ITS. The majority of research on the persona effect has shown no significant learning gains produced by on-screen agents, although many studies note that students find an ITS with an on-screen more engaging than one without (Moundridou & Virvou, 2002).

Our study is the first to assess the effect of the physical presence of a robot in an automated tutoring interaction. We use the task-performance measure of puzzle solving time in this work as well as several self-report measures.

## Methodology

### Participants

There were 100 participants in this study, between 18 and 40 years of age. The study was conducted in New Haven, Connecticut. Most participants were undergraduate and graduate students of Yale University. Each participant was assigned to one of five groups: (1) *no lessons*, (2) *randomized lessons from a physically-present robot*, (3) *personalized lessons from a disembodied voice*, (4) *personalized lessons from a video representation of the robot*, and (5) *personalized lessons from*

			1	1	2				1
			2	1	1	1	4	3	1
			1	3	3	2	3	3	3
1	1	3							
1	1	2							
1	1	2							
2	2	1							
1	1								
2	1	2							
7									
6	1								

(a) Sample nonogram puzzle, blank.

			1	1	2				1
			2	1	1	1	4	3	1
			1	3	3	2	3	3	3
1	1	3	■	■	■	■	■	■	■
1	1	2	■	■	■	■	■	■	■
1	1	2	■	■	■	■	■	■	■
2	2	1	■	■	■	■	■	■	■
1	1		■	■	■	■	■	■	■
2	1	2	■	■	■	■	■	■	■
7			■	■	■	■	■	■	■
6	1		■	■	■	■	■	■	■

(b) Sample nonogram puzzle, solved.

Figure 1: A sample nonogram puzzle. The objective of nonograms is, starting with a blank board (see left figure), to find a pattern of shaded boxes on the board such that the number of consecutively shaded boxes in each row and column appear as specified, in length and order, by the numbers that are printed to the left of each row and above each column (see right figure). For a more detailed explanation see the **Domain** section.

a physically-present robot. There were approximately 20 participants in each group. Exclusion criteria for participants were lack of English fluency or prior academic experience with robotics or artificial intelligence.

## Apparatus

In this experiment, participants were asked to solve a series of logic puzzles. In the four experimental conditions with a tutor, the tutor interrupted participants several times per puzzle to deliver puzzle-solving strategy lessons. The lessons themselves were pre-recorded audio and synchronized visual aids, between 21 and 47 seconds in length, that explained and gave examples of the use of a single puzzle-solving strategy. In the experimental conditions with *personalized lessons*, the order of the lessons was determined by a skill assessment algorithm that identified skills in which participants were weak; see the **Skills & Lessons** section. In the *randomized lessons* condition, the tutor chose a random lesson among the same ones used in the *personalized lessons* conditions, such that it was immediately applicable to the current state of the game-board. We compare the puzzle solving time performance between participants in these groups to evaluate the effect of the robot's physical presence on the effectiveness of the tutoring.

**Domain** To minimize the influence of prior experience, we chose a test domain to which participants likely had little previous exposure: a grid-based fill-in-the-blanks puzzle game called “nonograms” (or “nonogram puzzles”) that resemble crossword puzzles or Sudoku. Nonogram puzzles are a difficult cognitive task, one that requires several layers of logical inferences to complete. Solving a nonogram puzzle of arbitrary size is an NP-complete problem (Nagao, Ueda, Ueda, Sato, & Watanabe, 1996), meaning that no efficient computational solution is known.

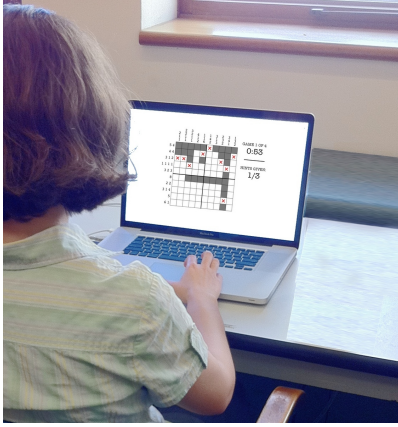
The objective of nonograms is, starting with a blank board, to shade in boxes on the board such that the number of consecutively shaded boxes in each row and column appear as

specified, in length and order, by the numbers that are printed to the left of each row and above each column. (See Figures 1(a) and 1(b) for a sample puzzle and solution.) For instance, a row marked as “4 2” must have 4 adjacent shaded boxes, followed by 2 adjacent shaded boxes—in that order, with no other boxes shaded, and with at least one empty box between the sets of adjacent shaded boxes. We refer to these contiguous sets of shaded boxes as “stretches” in this paper. For instance, the row described above requires two stretches, one of length 4, the other of length 2. One solves the puzzle when one finds a pattern of blank and shaded boxes such that all of the requirements for each row and column are satisfied.

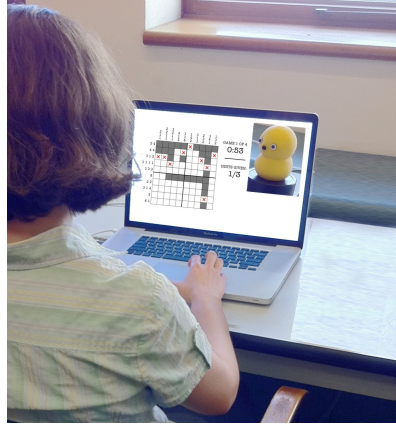
In a typical puzzle, one cannot solve many rows or columns independently. One must infer the contents of parts of rows or columns and use previous inferences as the basis of subsequent inferences. To that end, when a player has reasoned that in some box or boxes there should not be shading, they can mark such boxes with an ‘X’ for reference.

We created a full-screen nonograms computer program that participants used via mouse and keyboard. The user interface provided a timer and a count of how many lessons (called “hints” in the interface) the participant had received and how many they would receive; see Figure 2.

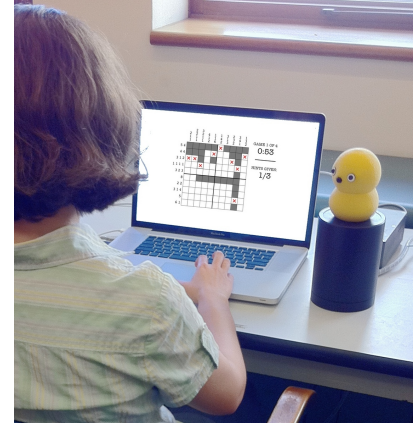
Participants were asked to play four puzzles on ten-by-ten grids with a time limit of fifteen minutes per puzzle. The puzzles themselves were the same across all participants. The fourth puzzle used the same board as the first, although disguised in the fourth puzzle by rotating the board 90° (such that the column stretch requirements were swapped with row stretch requirements). This means that the first puzzle and the last puzzle are of the exact same difficulty and require knowledge of the exact same set of skills to solve. This manipulation enables us to make within-subjects comparisons about the extent to which each participant improved their skills over the course of their participation in the study. There was no indication that any participant was aware of this manipulation.



(a) Experiment apparatus in the *no lessons* condition and the *personalized lessons* from a disembodied voice condition.



(b) Experiment apparatus in the *personalized lessons* from a video representation of the robot condition.



(c) Experiment apparatus in the *randomized lessons* from a physically-present robot condition and the *personalized lessons* from an embodied robot condition.

Figure 2: Experiment apparatus by condition.

**Skills & Lessons** In the four conditions with lessons, the tutor interrupted the participant three times per puzzle, paused the puzzle, and delivered a short lesson about nonograms. The lessons ranged from 21 seconds to 47 seconds in length and consisted of a voice recording and a set of animations presented on screen during the lesson as well as a set of coordinated robot motions specific to each lesson.

When beginning a lesson the tutor would turn to face the participant (in the *video* and *physically-present robot* conditions) and say “I have an idea that might help you,” or “Here’s another hint for you.” During the lesson, the tutor bounced subtly and looked back at the screen whenever, in the course of the lesson, it would make reference to the example presented on screen. For instance, when in the audio of the lesson the robot would say “Like in this example...” or “As you see here...,” the robot would turn briefly to the screen and then back to the participant.

Ten nonogram puzzle-solving skills were identified based on the subjective experience of the authors; they are not universally identified skills or rules for nonograms. Each skill is a set of row or column states in which one can logically fill in some of the remaining empty boxes. For example, a stretch of length 9 can fit in a blank row or column of 10 boxes in only two ways. Either it fills the first box and 8 more, or it fills those same middle 8 boxes and the last box. In either case, the middle 8 boxes are shaded. One of the ten skills in this experiment is that, for an empty row or column with just one stretch requirement of  $n$  where  $n > 5$ , the middle  $(2n - 10)$  boxes are shaded. See Figure 3 for examples and explanations of this skill and two others.

There was one recorded lesson for each skill. Three lessons were delivered per puzzle, for each of four puzzles. The number of lessons was constant for all participants regardless of how long they needed to finish the puzzle. Lessons were triggered either when a participant made no moves for 45 seconds or as he or she filled the 25<sup>th</sup>, 50<sup>th</sup> or 75<sup>th</sup> box on the board

(of 100). The user interface displayed the number of lessons remaining for each puzzle at all times.

In the *personalized lesson conditions* the lessons were chosen based on a skill assessment algorithm. For each skill, a weighted sum was calculated internally consisting of: (1) the number of recent demonstrations of that skill (weighted positively) and (2) the number of recent gameboard states in which a skill could have been applied but no action was taken (weighted negatively). These assessments were updated for each skill separately throughout the game, and the skill with the lowest assessment that was applicable to the current gameboard was the skill for which a lesson was selected. In this way, participants in the *personalized lesson* conditions received lessons based on their individual performance on the puzzles.

Alternatively, in the *randomized lesson condition*, lessons were chosen among the same ten lessons at random each time, such that the lesson chosen could be applied to the current state of the gameboard. This ensures that although the lessons were randomized, they would provide actionable information every time.

**Robot** The robot we used, Keepon, is a small yellow snowman-shaped robot; see Figure 2(c). Keepon has previously been used as an emotive non-threatening communication tool (Kozima, Nakagawa, & Yasuda, 2005; Leyzberg, Avrunin, Liu, & Scassellati, 2011).

The robot operated in one of three modes. First, it refereed the puzzle game: it welcomed participants when they started, told them when they had finished or when they had run out of time, and told them when the experiment was over. Second, it “observed” the board during gameplay: the robot frequently turned its head to face the location of the mouse cursor. Third, it delivered short gameplay lessons three times per puzzle: it “spoke” to the participant by turning to face him or her and “bouncing” its body subtly while playing one of several pre-recorded spoken messages. If a lesson needed to be repeated,

the robot would first apologize for repeating itself (i.e., “I’m sorry to repeat this hint but I think it might help.”).

To simplify the potential perception problems inherent in real-world measurements, the robot in this study received perfect knowledge of the state of the game. We did not use a robotic vision system to detect state changes.

## Procedure

Participants were first asked to watch a five-minute instructional video and read a two-page instruction manual describing the rules of nonograms and how to use the computer interface. In the video and in the text, participants were encouraged to use logical reasoning to make moves in the puzzle rather than making moves by guessing. Potential questions about the rules of the puzzle game were answered by the experimenter after the instructions.

During the experiment, participants were alone in a room with the computer, the robot in conditions including the robot, and a video camera positioned behind them; see Figure 2. Participants would choose when they were ready to start each new puzzle; each round would end either when the participant solved the puzzle or when fifteen minutes had elapsed, whichever happened first.

After the conclusion of the final puzzle, participants were asked to complete a survey consisting of five Likert-scale questions with open-ended follow-up questions for each. The questions were designed to assess whether the lessons were helpful, clear, and influential, as well as the user’s perceptions of the robot. We asked participants to rate: how relevant the lessons were, how much the lessons influenced their gameplay, how well participants understood the lessons, and how “smart/intelligent” and “distracting/annoying” they perceived the robot to be.

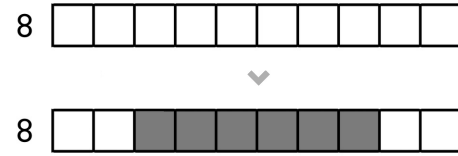
## Results

This study investigates the role of physical embodiment in a robot tutoring system. The behavioral measure is the time in which participants were able to solve each of the four puzzles. For the purposes of calculating a mean, puzzles in which participants ran out of time were evaluated as having solved the puzzle when time ran out, fifteen minutes from the start of each puzzle. This occurred in 12.4% of all puzzles.

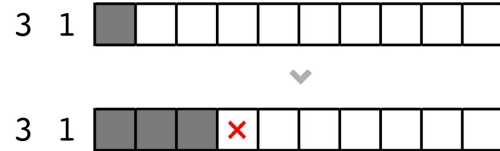
Table 1: Mean Solving Time

	Puzzle 1	Puzzle 2	Puzzle 3	Puzzle 4
<i>None</i>	13.6 ± 2.2	13.0 ± 2.3	12.3 ± 2.5	11.6 ± 2.7
<i>Rand.</i>	13.8 ± 1.4	12.5 ± 2.0	11.4 ± 2.3	10.3 ± 2.9
<i>Voice</i>	12.6 ± 2.4	10.7 ± 2.7	10.3 ± 3.3	9.1 ± 3.0
<i>Video</i>	12.8 ± 2.1	11.1 ± 2.6	9.9 ± 2.6	8.7 ± 2.4
<i>Robot</i>	12.7 ± 2.6	10.0 ± 3.5	9.4 ± 3.0	7.6 ± 3.1

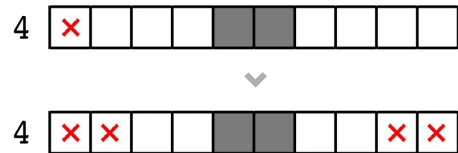
Participants in the *robot group* performed better, on average, on the second, third, and four puzzles than participants in any other group. See Table 1 for means and standard deviations and see Figure 4(a). In the forth puzzle, the



(a) In this row, there must be one long stretch. By the process of elimination one can infer that this stretch must occupy at least the middle six boxes, no matter where in the row it is placed.



(b) In this row, the first box is already shaded. Given that, and that the first stretch must be 3 boxes long, one can infer that the first three boxes must be shaded and the fourth must be crossed out.



(c) In this row, there is only one short stretch and some boxes are already shaded. One can infer that regardless of where that one stretch is placed, it cannot occupy the first two or the last two boxes in that row.

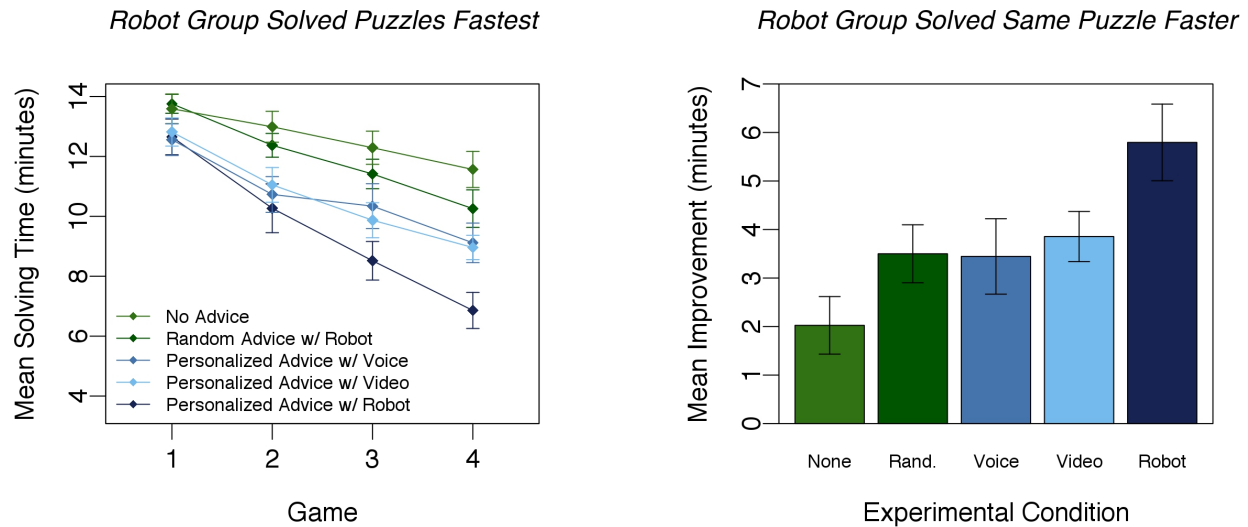
Figure 3: Examples of nonograms skills. Displayed are the contents of a row before and after each skill is applied. Although only rows are shown here, all nonograms skills apply to columns as well.

mean solving time in the *robot group* ( $M = 7.6$  minutes,  $SD = 3.1$ ) is significantly better than in the *video group* ( $M = 8.7$  minutes,  $SD = 2.4$ ),  $t(36) = 0.03$ , and in the *voice group* ( $M = 9.1$  minutes,  $SD = 3.0$ ) as well,  $t(36) = 0.02$ . These data indicate that the robot’s physical presence made a significant learning impact on participants greater than that of an disembodied voice and a video representation of a robot.

In this experiment, the first and fourth puzzles were 90° rotated variations of the same board. Thus they required exactly the same skills to solve and the difference in their solving time is a measure of the participants’ acquired knowledge over the course of the study. Participants in the *robot condition* improved ( $M = 5.8$  minutes,  $SD = 3.5$ ) their same-puzzle solving time significantly more than those in both the *video condition* ( $M = 3.9$  minutes,  $SD = 2.3$ ),  $t(36) = 0.048$  and *voice condition*, ( $M = 3.4$  minutes,  $SD = 3.5$ ),  $t(36) = 0.04$ ; see Figure 4(b). This data indicates that participants who received lessons from the robot learned more effectively than those who received only voice- or video-based lessons.

Survey results verify the following manipulation: participants in the three *personalized advice conditions* rated the lessons significantly more relevant ( $M = 6.0$ ,  $SD = 1.4$ )





(a) Mean solving time per puzzle. Participants in the *robot condition* solved each puzzle faster than participants in any other condition. In the fourth puzzle, significantly faster ( $p \leq 0.03$ ). See Table 1 for means and standard deviations.

(b) Mean improvement in solving time between puzzles #1 and #4. These two puzzles were variations of the same gameboard, disguised in the fourth puzzle by a 90° rotation. Participants in the *robot condition* improved their solving time significantly more than those in any other condition ( $p < 0.05$ ).

Figure 4: Behavioral measure results: (a) participants who received personalized lessons from an embodied robot solved every puzzle faster on average, the fourth significantly so ( $p \leq 0.03$ ) than participants in all other conditions; see Table 1. (b) *robot condition* participants also improved on their same puzzle solving time significantly higher more than participants in all other conditions ( $p < 0.05$ ).

than participants in the *randomized advice condition* ( $M = 3.9, SD = 1.1$ ),  $t(33) < 0.001$ . There was no significant difference in how highly participants rated their understanding of the lessons between groups: ( $M = 6.0, SD = 1.4$ ) in the *random condition*, ( $M = 6.6, SD = 1.2$ ) in the *voice condition*, ( $M = 6.6, SD = 1.5$ ) in the *video condition*, and ( $6.4, SD = 1.2$ ) *robot condition*; see Figure 5. These data indicate that whatever social effect physical embodiment has on this interaction, it does not influence the participants' perception of their understanding of the lessons, despite the fact that the behavioral measure indicates better learning in the *robot condition*.

## Discussion

Our results indicate that a physically-present robot tutor produces better learning gains than on-screen or voice-only tutors. Further work is needed to identify the underlying social factors and mechanisms that cause this effect.

One such factor may be the novelty of the stimulus. Robots are an uncommon stimulus in the present day; we may expect participants to be more attentive to the agent in the physically-present condition. However, novelty can also be a distraction. The physical presence of the robot during the game may divert the participant's attention from the puzzle solving task. More work is needed to identify what effect, if any, a robot's novelty has on interactions such as these.

Physical presence may imbue the robot with more per-

ceived authority than an on-screen agent. Earlier work in this area indicates that people are more likely to comply with commands given by a physically-present robot than an on-screen video of the same robot (Bainbridge et al., 2008). Embodiment may cause participants to take the robot tutor's advice more seriously. We are accustomed to receiving lessons from teachers and authority figures who have physical bodies. Perhaps a robot's physical presence increases its authority or social standing.

Participants, however, did not report having significantly more difficulty understanding the lessons in any of the three advice conditions. In fact, all four groups rated their level of understanding of the lessons fairly highly ( $M = 6.3, SD = 1.3$ ); see Figure 5(b). This may indicate that the embodiment effect is so subtle that the participants did not notice its effect on their learning.

Another social factor is the potentially increased sense of peer pressure during the performance of the task itself. The distinction between physically-present robot and on-screen agent may parallel the way we perform tasks when we think of ourselves as alone rather than in view of another person. In person-person interactions, social presence can lead to significantly worsened task performance, especially in cognitively-demanding tasks (Short, 1976). More work is needed to compare the potential effect of peer pressure caused by a physically-present robot tutor to the peer pressure exerted by a human tutor who observes as participants perform



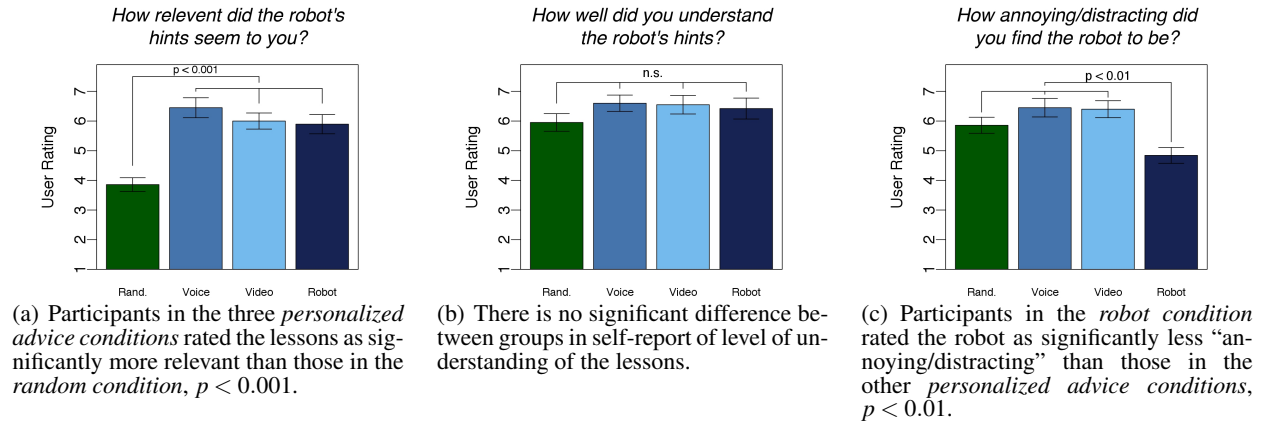


Figure 5: Results of self-report questionnaire measures completed after the interaction.

cognitively-demanding tasks.

Social presence effects may also be responsible for the survey result in which participants in the *physically-present robot condition* rated the robot ( $M = 4.7, SD = 1.8$ ) as significantly less “annoying/distracting” than participants in the other *advice conditions* ( $M = 6.1, SD = 1.3$ ),  $t(33) < 0.01$ ; see Figure 5(c). This may indicate that physical embodiment produces a significantly greater sense of social acceptance than an on-screen agent does.

Participants in the *robot condition* became better puzzle solvers than those in the other conditions. Further research is needed to identify the underlying social factors that contribute to this empirically-observed effect.

## Conclusion

This study investigates the role of physical embodiment of a robot tutor in a cognitive skill learning task. Participants who received personalized lessons from a physically-present robot outperformed participants who received the same kind of advice from a video representation of the same robot as well as participants who received the same kind of advice from a disembodied voice on the last three puzzles. Participants in the *robot condition* also improved their same-puzzle solving time significantly more than those in any other group, which is a direct measure of learning gains over the course of the experiment. From these data we conclude that physical embodiment can yield measurable learning gains in robot tutor interactions.

## Acknowledgments

This material is based upon work supported by grants from the National Science Foundation under contracts No. 1139078, No. 1117801, and No. 0835767.

## References

- Bainbridge, W., Hart, J., Kim, E., & Scassellati, B. (2008). The effect of presence on human-robot interaction. *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, 701–706.
- Dehn, D., & Van Mulken, S. (2000). The impact of animated interface agents: a review of empirical research. *International Journal of Human-Computer Studies*, 52(1), 1–22.
- Kidd, C., & Breazeal, C. (2004). Effect of a robot on user perceptions. *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, 4, 3559–3564.
- Kiesler, S., Powers, A., Fussell, S., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition*, 26(2), 169–181.
- Kozima, H., Nakagawa, C., & Yasuda, Y. (2005). Interactive robots for communication-care: a case-study in autism therapy. *IEEE International Symposium on Robot and Human Interactive Communication*, 341 - 346.
- Leyzberg, D., Avrunin, E., Liu, J., & Scassellati, B. (2011). Robots that express emotion elicit better human teaching. *6th International Conference on Human-Robot Interaction*, 347–354.
- Moundridou, M., & Virvou, M. (2002). Evaluating the persona effect of an interface agent in a tutoring system. *Journal of Computer Assisted Learning*, 18(3), 253–261.
- Nagao, T., Ueda, N., Ueda, N., Sato, C. P. T., & Watanabe, C. P. O. (1996). *Np-completeness results for nonogram via parsimonious reductions* (Tech. Rep.). Tokyo Institute of Technology.
- Nkambou, R., Bourdeau, J., & Psyché, V. (2010). Building intelligent tutoring systems: An overview. *Advances in Intelligent Tutoring Systems*, 361–375.
- Short, W. E. . C. B., J. (1976). *The social psychology of telecommunications*. London, England.
- Tapus, A., Tapus, C., & Mataric, M. (2009). The role of physical embodiment of a therapist robot for individuals with cognitive impairments. *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, 103–107.
- Wainer, J., Feil-Seifer, D. J., Shell, D. A., & Mataric, M. J. (2007). Embodiment and human-robot interaction: A task-based perspective. *IEEE Proceedings of the International Workshop on Robot and Human Interactive Communication*, 872–877.

# Designing Better Scaffolding in Simulation-Based Learning Environments Teaching Science Systems: A Pilot Study Report

**Na Li (nl2284@tc.columbia.edu)**

Teachers College, Columbia University, 525 W. 120<sup>th</sup> street  
New York, NY 10025 USA

**John Black (black@exchange.tc.columbia.edu)**

Teachers College, Columbia University, 525 W. 120<sup>th</sup> street  
New York, NY 10025 USA

**Mengzi Gao (mg3220@tc.columbia.edu)**

Teachers College, Columbia University, 525 W. 120<sup>th</sup> street  
New York, NY 10025 USA

## Abstract

Systems are an important part in today's science education. Computer simulations have many advantages in teaching science systems. Our research goal is to test whether a hierarchical instructional scaffolding framework embedded in simulation-based learning environments and retrospective mental modeling task could facilitate mental model construction in learning science systems. This pilot study was conducted with a sample of adult learners who didn't have strong science background. They were asked to learn a chemical system in a simulation-based environment. The results show that participants in the hierarchical scaffolding condition performed better than the non-hierarchical scaffolding condition. The retrospective mental modeling task could enhance learning only within the hierarchical scaffolding condition; while in the non-hierarchical scaffolding condition, the task was detrimental to learning. Based on the results from the pilot study, an 8 session curriculum teaching ideal gas laws to middle school students has been designed for our future study.

**Keywords:** Science systems; simulation-based learning environments; scaffolding.

## Research Background

Systems thinking skills have become increasingly important in today's science education. Scientific explanation of mechanism is usually difficult in learning a system (Jacobson & Wilensky, 2006). Studies have demonstrated that learning systems thinking skills go through several sequential stages before learners are able to grasp a network of mechanics-function relations (Assaraf & Orion, 2005); for example, studies on the reasoning processes of complex systems show that novices focus more on the structure of the system, while experts tend to reason around mechanism and functions of the system (Jacobson, 2001).

Structure-Behavior-Function framework (SBF) provides a language to describe experts' and novices' conceptual representations of systemic knowledge (Hmelo-Silver, Marathe & Liu, 2007). Structure refers to the elements of the system, behavior refers the mechanism of how the elements act and interact leading to certain outcomes; and function

refers to the roles of the elements or the outcomes caused by the elements' behaviors (Hmelo-Silver, Marathe & Liu, 2007). Explaining mechanism and causality is usually difficult for learners especially when the systems have hierarchical levels (Duncan & Reiser, 2007). One important pedagogical implication from these studies is to provide hierarchical instructional scaffolding based on the SBF framework to help learners iteratively modify their conceptual representations (Liu & Hmelo, 2009).

## Mental Models of Science Systems

Mental models are internalized representations of the structural and functional relations of the reality (Johnson-Laird, 1983). The constructivist perspectives imply that mental model construction goes through trajectories, and iterative mental model modification could be very effective (Vosniadou & Brewer, 1992). Learning science systems usually require learners to construct mental models with various entities and a network function relations among the entities, well scaffolded step-wise learning could produce better structured conceptual representations (Clement & Steinberg, 2002).

Active mental modeling, or active rule-driven visualization, involves cognitive processes such as mentally manipulate the visual information to solve a problem (Briggs & Bodner, 2005). Learners' knowledge of a system could still be fragmented after initial learning, retrospective mental modeling around system functions with "what" and "how" questions could facilitate internal information organization, and enhance reflective thinking.

## Using Computer Simulations to Teach Science Systems

Deep understanding of a system involves constructing a mental perceptual simulation for information retrieval and reasoning (Black, 2010). Computational modeling and visualizing technology makes it possible to show the otherwise invisible mechanism of systems (Wilensky & Resnick, 1999), which could provide rich perceptual information to ground the abstract concepts (Barsalou,

2008). Multiple dynamic representations at different abstract levels could provide complimentary information, constraining interpretation of any singular representation, and support deep understanding (Ainsworth & VanLabeke, 2004). Active integration of multiple structurally and conceptually mapped representations can potentially facilitate deep learning (van der Meij & de Jong, 2006; Plass, Homer & Hayward, 2009). At different learning stage, different representation could be used for different learning purposes. For example, a concrete graphical representation can be used to depict system phenomena, and an abstract flowchart representation can be used to model symbolic systemic mechanism after sufficient perceptual information has been delivered.

Although multiple dynamic representations in simulation-based environments have the potential to facilitate mental model construction in learning difficult science systems, the instructional scaffolding should be well designed to help learners make full use of the learning environments. Our research question in this pilot study and future studies is: How to design better scaffolding in simulation-based environments teaching science systems? Pedagogical research implies that mental model construction has a hierarchical nature and goes through stages, thus the scaffolding should support sequential and step-wise learning. Additionally, retrospective mental modeling around system function might facilitate internal organization of systemic knowledge.

## Hypotheses

H1: Hierarchical scaffolding based on the Structure-Behavior-Function framework produces better learning performance.

H2: Retrospective mental modeling task facilitates internal reconstruction of the system knowledge.

## Method

### Participants

Participants for this pilot study were 36 adult learners (Mean age: 29.7, SD=7.61) from a graduate school of Education with a diversity of ethnicity. 29 of them were females and 7 were males. Most of them majored in social sciences and humanities fields, and didn't have strong background in science. One case was dropped because the participant totally misunderstood the learning goals and didn't complete the posttest. Another pilot study earlier showed that reasoning across different levels of a complex system and scientific reasoning about causality was difficult even for this population, and that was why we conducted a pilot study in this population before designing the curriculum for the junior high population.

### Instrument

Two computer-based simulations teaching three Ideal Gas Laws were used in this study. Participants were asked to

learn how Temperature, Volume and Pressure of a certain amount of ideal gas interact and reason about the relationship between lower-level molecular activity and the emergent function. One simulation was a realistic model (see Figure 1) and the other was a conceptual flowchart model (see lower part of Figure 2). These two dynamic representations are structurally and conceptually mapped, depicting and describing the system knowledge at different abstract levels. The realistic model provides rich visual information of the system phenomena while the flowchart model emphasizes the mechanism and causality in the system. The function of a realistic graphic simulation is to provide rich perceptual information grounding the abstract symbolic concepts, and the function of a conceptual model simulation is to constrain the processing of the visual information, reinforce the symbolic level of understanding.

### Design

This study employed a 2x2 factorial design testing the effect of hierarchical scaffolding based on SBF conceptual framework (HS), the effect of the retrospective mental modeling task (RMM) and their interaction effect. Regarding the procedure of the experiment, the manipulation of "RMM vs. N-RMM" came after the manipulation of "HS vs. N-HS".

### Procedure

1. Participants signed the consent form
2. Participants read through powerpoint slides which gave them an introduction to what they were going to learn
3. Participants interacted with the simulations for a couple of minutes to get familiar with the interface
4. Learning stage: participants were randomly assigned to a condition, given the worksheet which guided them through the whole learning process. They were asked to think aloud as they were learning. Eight sessions were randomly selected to be videotaped and the verbal protocols transcribed.
5. Posttest

The whole session lasted around 70-80 minutes in total for each participant.

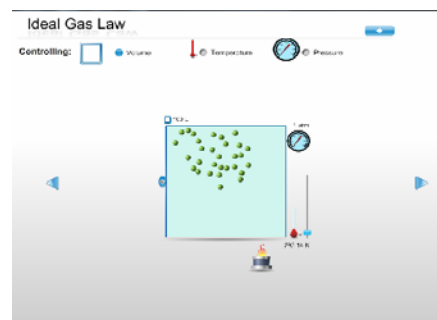


Figure 1: The realistic model simulation-an experiment

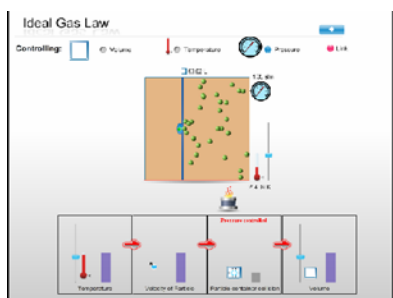


Figure 2: Two simulations displayed on the same page, and dynamically linked

## Manipulation

**Hierarchical Scaffolding.** The instruction was divided into three steps:

**Step 1.** Learners observed the higher level system function;

**Step 2.** Learners Described system lower level behaviors;

**Step 3.** Learners modeled the system causality around the system function.

In step 1, participants played with only the concrete graphic simulation (see Figure 1), and were asked to describe higher level phenomena for each ideal gas law (e.g.: how temperature and volume interact when pressure is constant). In step 2, they were asked to observe and describe lower level element behaviors (gas particle velocity, particle-container collision). In step 3, the instruction is function-centered with questions such as “when pressure is constant, why increasing temperature will lead to increased volume?” With both the concrete graphic simulation and the conceptual flowchart simulation (see Figure 2), participants were required to explain the lower level mechanism in a coherent matter for each ideal gas law. Simply speaking, in the HS condition, participants observed and described system structure and function, integrated fragmented behavior information, then connected the molecular behavior and the emergent T-V-P relationship, and explained the mechanism or causality in a coherent manner.

**Non-Hierarchical Scaffolding.** Indicated in a previous pilot study, when two simulations were displayed on the same page at the very beginning, participants tended to regard the flowchart concept model as complimentary fragmented behavior information, thus described each bar diagram separately, rather than using it as a modeling tool to explain the system causality. So for the no hierarchical scaffolding condition (N-HS), participants were given the combined simulations interface at the very beginning (see Figure 2); given the worksheet including all the questions asking about the system function and lower-level behaviors for each ideal gas law. The participants were asked to describe the structure, function and behavior knowledge for each ideal gas law, and were not guided to iteratively interrogate with the system. It was ensured that participants

in the N-HS condition had same amount of questions asking about the system functions and behaviors compared to the HS condition, while there were no structured progressive learning steps in this condition.

## Retrospective Mental Modeling & Control Condition.

There were three ideal gas laws for the participants to learn in this experiment. After a participant completed learning one ideal gas law, the other experimental variable (Retrospective mental modeling task) was manipulated. For the retrospective mental modeling condition (RMM), after learning each ideal gas law by interacting with the simulations, the participants were asked to close their eyes, describe the processes of how the phenomenon happens. For the no-retrospective mental modeling condition (N-RMM), there was simply no such a step.

## Measures

The posttest included four sections:

1. Comprehension task: participants were given three questions asking them to explain the mechanism for each ideal gas law phenomenon.
2. Four multiple-choice questions on problem solving
3. Explaining new diagrams: participants were given three new line diagrams representing the events happening from time A to time B, and they were required to visualize and describe the what happened in the system
4. Transfer task: participants were asked to explain everyday gas law problems

## Results

The posttest results indicate that the groups differ in their understanding of the lower-level molecular activity (the mechanism and processes of the system) but not higher-level structure and function of the system.

Task 1 (Comprehension task) and Task 3 (Explaining diagrams task) measured the understanding of lower-level behaviors and their functions (molecular activity). The answers for Task 1 (Comprehension task) were coded on the presence and absence of lower-level mechanism knowledge units (highest possible score: 7), by two raters blinded to the condition of the participants. The agreement was 93.2%, and the rest was resolved through discussion. The results of Task 1 are displayed in Table 1 and Figure 3.

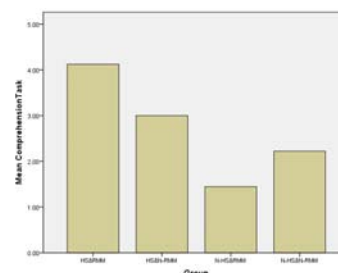


Figure 3. Comprehension task mean scores

Table 1. Comprehension task mean scores

Group	Mean	N	Std. Deviation
HS&RMM	4.1250	8	1.72689
HS&N-RMM	3.0000	9	1.00000
N-HS&RMM	1.4444	9	1.13039
N-HS&N-RMM	2.2222	9	1.09291
Total	2.6571	35	1.55190

Significance tests show that HS has significant effect on learning lower-level element behaviors and their functions,  $F(1, 31)=16.63$ ,  $p<.001$ , no main effect is found for RMM,  $F(1, 31)=.168$ , n.s., while the interaction of HS and RMM is significant,  $F(1, 31)=5.03$ ,  $p=.032<.05$ . Post-hoc tests show that HS&RMM performed significantly better than the N-HS&RMM and N-HS&N-RMM group, and the N-HS&RMM performed the worst, which indicates that without hierarchical scaffolding, the retrospective mental modeling will do no good but interfere with the learning.

Task 3 was also coded on the presence and absence of lower-level behavior and function knowledge units (possible highest score 8) by two raters blinded to the condition. The agreement is 95.9%. Participants who described more molecular activity in explaining the abstract line diagrams were believed to notice and appreciate the importance of behavior-function interdependence. The results of Task 3 (see Figure 4 and Table 2) show a similar pattern as in Task 1.

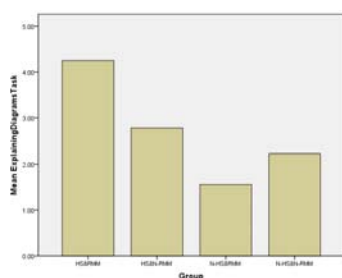


Figure 4. Explaining diagrams mean scores

Table 2. Explaining diagrams task mean scores

Group	Mean	N	Std. Deviation
HS&RMM	4.2500	8	2.81577
HS&N-RMM	2.7778	9	2.86259
N-HS&RMM	1.5556	9	1.50923
N-HS&N-RMM	2.2222	9	2.04803
Total	2.6571	35	2.46078

The main effect of HS is marginally significant,  $F(1, 31)=4.13$ ,  $p=.051$ ; although the interaction of HS and RMM is not statistically significant, it is mostly due to a small sample size.

No difference was found in Task 2 (multiple choice questions, highest possible score 4, see Table 3). It might be because the participants tended to do abstract rule-based reasoning rather than applying their mental models in solving the problems, as indicated in the interviews with some of the participants. e.g., one question is “If you want to maintain pressure at a constant level, which of the following combination would work?” Participants will tend to draw abstract rules (e.g., Pressure constant, Temperature increases, Volume increases) and then make the judgment for each choice, without visualizing the molecular activity and the processes of the system.

Table 3. Multiple choices task mean scores

Group	Mean	N	Std. Deviation
HS&RMM	2.7500	8	1.16496
HS&N-RMM	2.6667	9	.86603
N-HS&RMM	2.7778	9	1.09291
N-HS&N-RMM	2.8889	9	1.26930
Total	2.7714	35	1.05957

In task 4, although participants were originally expected to describe lower-level system behaviors to explain the everyday ideal gas law phenomena, many of them focused only on the higher-level structure and function of the system, so the answers were coded on the important system behavior and function knowledge units (both higher-level and lower-level, highest score: 8, see Table 4) The agreement between the two raters was 85.2%. In this task, the participants were only asked to explain the phenomena without explicit questions asking about the molecular activity. The data also imply that in order to help the learners to integrate the invisible lower-level system behaviors into their explanation, another level of scaffolding for transfer might be needed.

Table 4. Transfer task mean scores

Group	Mean	N	Std. Deviation
HS&RMM	4.2500	8	2.81577
HS&N-RMM	4.2222	9	2.10819
N-HS&RMM	4.2222	9	1.30171
N-HS&N-RMM	4.1111	9	2.42097
Total	4.2000	35	2.11159

### Some Qualitative Analysis

For better understanding of the effect of hierarchical scaffolding, eight verbal protocols (4 from HS condition, 4 from N-HS condition) were transcribed and analyzed. It was hypothesized that more clear and efficient trajectories of mental model construction could be found in the HS condition. The qualitative data does indicate that the participants in the HS condition were more likely to progressively modify their mental models.

Below are parts of a participant's verbal transcript (HS condition) which demonstrate how he gradually construct scientific causal model of the system.

When the participant was learning the structure and fragmented behavior knowledge (step 1):

*"As the temperature goes down, the volume decreases, the volume goes down, the temperature decreases I guess... yeah...ha...I've no idea how that works...but that's what the simulation tells me. Why would that happen?"*

*'cause the Temperature goes up, the Pressure goes up, the volume goes down, the pressure goes up...when pressure is constant...when the temperature goes down, volume goes down...hum..."*

Here the participant was dealing with the higher-level system function (when pressure is constant, temperature changes cause volume to change), he was curious about how that happens, which prepared him to actively integrate and connect the lower-level molecular behaviors knowledge. This also supports the idea of function-centered scaffolding.

The second part of the transcript indicates the participant was now trying to connect the two levels of information. He was trying to clarify the causal relationship among all the lower level and higher level elements.

*"so the temperature and velocity are clearly related, because as I bring the Temperature down, the velocity of particles move. If we wanna keep the pressure the same...So I am guessing, if I move this back, the pressure is probably gonna go...oh the pressure stays the same, the temperature will have to go up...yeah...so as volume increases...the temperature has to go up because...how can I explain that...so we have a constant pressure here, so that means...all of these have to collide at the same rate, that means when there is less space, they have to move a lot slower to*

*maintain the same pressure...yeah...now they have to move a lot faster, the temperature has to go up"*

After successfully integrating the information, when the participant was asked to answer the question "why temperature increases, volume increases" in the retrospective mental modeling stage, the participant was able to provide a very sophisticated answer while visualizing the system processes.

*"so if the pressure is constant, then as volume decreases, the temperature also has to decrease because the particles have to move at a slower rate in order to maintain the pressure in a smaller volume."*

Below are parts of a participant's verbal transcript (N-HS condition):

*"so pressure is gonna be controlled...and...temperature and volume...temperature affects volume...so...when you increase the temperature, you increase the volume, you decrease the temperature, you decrease the volume...the velocity also corresponds...and now container-collision is gonna go up...(confused)"*

Here the participant was trying to learn the Temperature-Volume relationship when pressure is controlled. She was given the combined simulations interface at the very beginning (see Figure 2.) and had to construct a hierarchical mental model without the progressive scaffolding. It could be seen that she was trying to integrate all the functional and behavior knowledge, but she struggled in trying to give a coherent explanation.

When she was asked to answer the question "why temperature increases, volume increases" in the retrospective mental modeling task, she failed in integrating lower-level behavior knowledge in her explanation, as can be seen in the following transcript:

*"The temperature increased, the volume increased. When the temperature decreased, volume decreased...Why? I have no idea."*

The qualitative data implies that learners might need to interrogate with the system progressively and iteratively in order to form deep understanding. Experiencing the system function and integrate the system behavior knowledge based on the system function could be very effective. Another implication is that modeling causality after learners have observed all the system behaviors lead to more compact and sophisticated mental models.

### Discussion

This pilot study demonstrates that hierarchical scaffolding (HS) could help learners better integrate the lower-level system behavior knowledge and learn the causality. The interaction between hierarchical scaffolding and retrospective mental modeling is interesting. It seems retrospective mental modeling could enhance learning only when the learning process itself is well scaffolded. One explanation is that learners need to internalize the knowledge in a well structured way before they can mentally reorganize the information in a coherent manner.



Without such a structure, mentally reorganizing the information might counter learning.

### Limitations and Future Study

This pilot study only included three-step scaffolding because we assumed the adult learners should be already familiar with the everyday ideal gas law phenomena. The total learning time was very short since this system was not too challenging to the adult learners. Based on the implications from this pilot, we have designed an 8 session curriculum teaching ideal gas laws for 7<sup>th</sup> and 8<sup>th</sup> graders. The hierarchical scaffolding is now operationalized into 5 steps, which help the junior high students gradually construct better mental models of the systems. Everyday ideal gas law problems are also incorporated into the simulation environment (e.g., students are able to manipulate the fire icon to make a soda can explode, etc). In our future study, learners' learning trajectory will also be recorded and analyzed to better answer our research question. To reduce the cognitive load in integrating the system behaviors and modeling system causality, two techniques are used in designing the simulations: a. the everyday ideal gas law simulation and the gas molecules simulation are dynamically linked for students to compare and analyze the different levels of the system (example, see Figure 5). b. Concrete icons instead of bar diagrams are used in the flowchart simulation which helps learners to model system causality (see Figure 6).

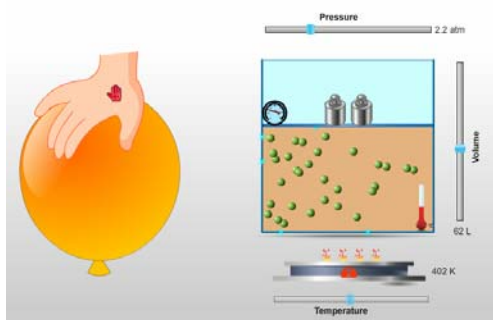


Figure 5. The everyday gas law problem and the molecule simulation dynamically linked

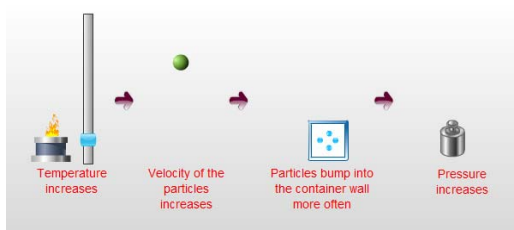


Figure 6. The flowchart simulation with concrete dynamic icons

### References

- Ainsworth, S. & van Labeke, N. (2004) Multiple forms of dynamic representation. *Learning and Instruction*, 14, 241–255.
- Assaraf, O. B., & Orion, N. (2005). Development of system thinking skills in the context of earth system education. *Journal of Research in Science Teaching*, 42(5), 518-560.
- Black, J. B. (2010). An embodied/grounded cognition perspective on educational Technology. In Issa Saleh (Ed.) *New science of learning: Cognition computers and collaboration in education*. Springer Publishing.
- Barsalou, L.W. (2008). Grounded cognition. *Annual Review of Psychology*, 59, 617-645.
- Briggs, M., & Bodner, G (2005): A Model of Molecular visualization. In J. K. Gilbert (Ed.): *Visualization in Science Education*, pp. 61-72. Springer.
- Clement, J., & Steinberg, M. (2002). Step-wise Evolution of Mental Models of Electric Circuits: A “Learning-Aloud” Case Study. *Journal of the Learning Sciences*, 11(4), 389-452.
- Duncan, R. G., & Reiser, B. J. (2007). Reasoning across ntologically distinct levels: Students' understandings of molecular genetics. *Journal of Research in Science Teaching*, 44(7), 938-959.
- Hmelo-Silver, C. E., Marathe, S., & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: Expert-novice understanding of complex systems. *Journal of the Learning Sciences*, 16(3), 307-331.
- Jacobson, M. J. (2001). Problem solving, cognition, and complex systems: Differences between experts and novices. *Complexity*, 6(3), 41-49.
- Jacobson, M. J., & Wilensiy, U. (2006). Complex systems in education: Scientific and educational importance and implications for the learning sciences. *The Journal of Learning Sciences*, 15(1), 11-34.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, interferences and consciousness*. Cambridge, England: Cambridge University Press.
- Liu, L., & Hmelo-Silver, C. E. (2009). Promoting complex systems learning through the use of conceptual representations in hypermedia. *Journal of Research in Science Teaching*, 46(9), 1023-1040.
- Plass J., Homer B., & Hayward E. (2009). Design factors for educationally effective animations and simulations. *Journal of Computing and Higher Education*, 21, 31-61.
- van der Meij, J., & de Jong, T. (2006). Supporting students' learning with multiple representations in a dynamic simulation-based learning environment. *Learning and Instruction*, 16, 199–212.
- Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of Conceptual Change in Childhood. *Cognitive Psychology*, 24, 535-585.
- Wilensky, U., & Resnick, M. (1999). Thinking in levels: A dynamic systems perspective to making sense of the world. *Journal of Science Education and Technology*, 8(1), 3-19.



# A New Angle on the EMPATH Model: Spatial Frequency Orientation in Recognition of Facial Expressions

**Rentao Li (reli@ucsd.edu)**

Department of Computer Science, 9500 Gilman Drive  
La Jolla, CA 92093 USA

**Garrison Cottrell (gary@eng.ucsd.edu)**

Department of Computer Science, 9500 Gilman Drive  
La Jolla, CA 92093 USA

## Abstract

Many have investigated the sensitivity of face processing to both spatial frequencies and face orientation, but few have researched the sensitivity of face processing to the orientation of spatial frequencies. One recent exception has been Yu, Chai, & Chung (2011), which investigated facial expression recognition in regards to the orientation of spatial filters and showed that most information is contained in the horizontal orientation. Here, we model the Yu, Chai, & Chung (2011) study using the EMPATH model, a feed-forward neural network that has been used to model facial expression recognition (Dailey, Cottrell, Padgett, & Adolphs 2002). We used the NimStim set of facial expressions, which were the basis for the Yu, Chai, & Chung (2011) experiment, and followed their method of filtering images through different spatial orientations. Our results show that this simple, biologically plausible model produces very similar results to that of human subjects in their study.

**Keywords:** emotions; facial expressions; spatial frequency; neural network; face recognition.

## Introduction

Many studies have been conducted regarding the role of spatial frequencies in human face recognition (Näsänen, 1999; Costen, Parker, & Craw, 1996; Gold, Bennett, & Sekuler, 1999), although the uniqueness of sensitivity of faces to spatial frequency has been debated (Williams, Willenbockel, & Gauthier (2009). However, few have explored how different *orientations* of spatial frequencies impact recognition of facial images. One such experiment was done by Yu, Chai, & Chung (2011), who passed facial images through orientation filters from -60 to 90 degrees in increments of 30 degrees. They found that the spatial information near horizontal (between -30 and 30 degrees) were the most important for normally-sighted human respondents to recognize facial expressions.

## Background: Yu, Chai, & Chung's Experiment (2011)

The aim of the Yu, Chai, & Chung (2011) experiment was to determine which spatial orientations on the face contained the most information for identifying emotions. The four emotions they tested were the closed-mouth forms of anger, fear, happiness, and sadness. Images were

obtained from the NimStim set of facial expressions (Tottenham, Tanaka, Leon, McCarry, Nurse, Hare, Marcus, Westerlund, Casey, & Nelson 2009), and were distorted with an orientation filter of bandwidth  $23^\circ$  in the Fourier domain, where the center of the filter ranged from  $-60^\circ$  to  $90^\circ$  in increments of  $30^\circ$ . Unfiltered images were used for comparison.

Their experiment consisted of having 15 normally-sighted human subjects try to recognize the expression displayed by each image under a four-way forced choice. The results indicate that the human observers had the most success with images filtered at orientations near the horizontal ( $-30^\circ$ ,  $0^\circ$ , and  $30^\circ$ ), suggesting that horizontal spatial information is most important for recognizing facial expressions. One modest exception to this trend is the fearful face; the human subjects tended to be significantly biased towards labeling a face as fearful as the orientation filter approached  $90^\circ$ , which seems to indicate that much of the information for fear is represented vertically.

The purpose of this current experiment is to determine if a neural network model can produce similar results as the human subjects, especially in regards to the increased recognition performance for horizontal orientations and the preference towards fear for vertical orientations. Such evidence would provide greater support for Yu, Chai, & Chung's (2011) findings and further validate EMPATH's flexibility and accuracy in modeling human face recognition.

## Methods

### The Model

The neural network used for this experiment closely followed the EMPATH model developed by Dailey et al. (2002), consisting of a biologically plausible, three-layer, feed-forward perceptron. EMPATH has been shown to have remarkable face recognition performance on aligned, grayscale images from Ekman and Friesen's POFA (1976). Without being tuned specifically to those images, the network classified the emotions Anger, Disgust, Fear, Happiness, Sadness, and Surprise with 90% accuracy on average, compared to 91.6% for human subjects (Dailey et al., 2002). For this experiment, we kept much of the settings (outlined below) identical to those of the original EMPATH

model, so the network was not tailored for the spatially filtered images or the NimStim dataset.

The first layer consisted of a set of model neurons based on the magnitude of Gabor filters, which have become a standard way to model complex cells in the early visual cortex (Daugman, 1985). In all, 40 different Gabor filters were used, in combinations of 5 scales and 8 orientations; filtering was done by passing the face images through a 29 by 35 “grid” of filters, resulting in 40,600 responses per image. Note that the orientations of the Gabor filters were the same as in the original EMPATH model, and were not changed to fit with the spatial filtering used in this study.

In order to reduce the dimensionality of the data set, we performed principal component analysis (PCA) on the Gabor filter outputs, producing 50 principal components (again based on the original EMPATH model). In this second layer, the principal components capture the distinguishing features of each facial expression but abstract away from details unique to each face; hence they allow the network to generalize to novel faces that are not part of the training set. As Dailey et al. stated, these components are similar to face cells in the inferior temporal cortex (2002).

Lastly, the principal components were fed into the third layer, consisting of a simple linear perceptron with six softmax outputs representing anger, fear, happiness, surprise, disgust, and sadness. This perceptron was trained using stochastic gradient descent with the cross-entropy error criterion. We used an “all-or-none” teaching signal that had “1” for each correct expression and “0” for the incorrect expressions. In order to replicate the four-way forced-choice employed by Yu, Chai, and Chang (2011), we only took the results of the four relevant emotions via a process described in “Training, Validating, and Testing.”

As stated in Dailey et al., we acknowledge that this perceptron is very simplistic (2002). However, since it was powerful enough to map the principal components to emotion categories, we did not feel that a non-linear classifier was needed.

## The Images

The images used in the testing set by Yu, Chai, & Chung (2011) consisted of morphs of images taken from the NimStim set (2009), which reduced variations among faces (such as race, gender, etc.). Without access to these exact morphs, however, we simply used some of the original NimStim images to create our training and testing sets. Given this difference, our results still closely matched those of Yu, Chai, & Chung (2011).

Our testing and training sets consisted of grayscale images of 30 different people (17 male, 13 female). Two of the images are shown in Figure 1. These images were judged to be the most frontally aligned, making them the most suitable for EMPATH. Each image was 240 x 292 pixels in size and was cropped closely about the face.

Both the testing and training sets contained images of six different emotions for each of the 30 people. These expressions were comprised of both open and closed-mouth

forms of anger, fear, happiness, and sadness; the open-mouth form of surprise; and the closed-mouth form of disgust.

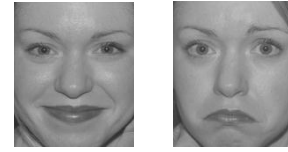


Figure 1: Two of the cropped images, corresponding to Happiness and Sadness

The additional expressions were chosen for the training set so that the network would have a more comprehensive exposure to the range of different emotions, making it more similar to the experience of the human subjects in the experiment. Although the testing set also contained six expressions, our method for producing the network’s output was able to emulate a four-way forced choice among the four expressions used by Yu, Chai, & Chung (2011), which effectively limited the output to only those four choices (detailed in “Training, Validating, and Testing”).

**Processing the training set:** In order to better replicate the images used by Yu, Chai, & Chung (2011), the 30 sets of images were closely cropped about the face using an oval mask so that only an oval-shaped portion of the face was visible. The parts that were cropped out were filled in with a uniform gray color of RGB value 127, and the entire image was adjusted to have a root-mean-square contrast value of 0.096, as per specifications given in Yu & Chung (2011). Examples of images used in the training set are shown in Figure 2.

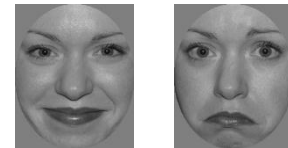


Figure 2: Two images from the training set

**Processing the testing set:** The process for creating the testing set was very similar to that of the training set. The 30 original sets of images were first cropped, aligned, and then processed using Yu, Chai, & Chung’s (2011) filters at six orientations from  $-60^\circ$  to  $90^\circ$  in increments of  $30^\circ$ , which selectively pass information at the specified orientations. Afterwards, the oval “mask” was applied to the images in the same way as that of the training set, and all of the images were again normalized to have the same root-mean-square contrast of 0.096. Examples of the test images are shown in Figure 3. These were then processed by the same Gabor filters as used in the training set, and the resulting filter responses were projected onto the 50 eigenvectors using the PCA that was computed on the training set.



Figure 3: All 6 filtering conditions shown horizontally from  $-60^\circ$  (top left) to  $90^\circ$  (bottom right).

### Training, validating, and testing

The last layer of the model was a 50-input, 6-output single layer perceptron with softmax outputs trained using cross entropy. This procedure leads to outputs that compute the conditional probability of the category given the inputs (Bishop, 1995). Hence the output of the network is a probability distribution over the facial expression categories.

Cross-validation and early stopping were used to prevent overfitting the network to the training set. Since there are thirty individuals in the training set, we performed thirty instances of cross validation, each time holding out a different individual in the training set to use for early stopping. This comprehensive cross-validation made the testing less prone to unevenness among the images. Overall, there were 30 independent test cycles; each time 1 set was chosen for testing, 1 chosen for validation, and the remaining 28 were used as the training set. The aggregate performance from these 30 sets of tests constituted the results for each filtering condition. The validation set was taken from the processed test set as a guide to know when to stop training (of course the validation and test images were never the same). Training was completed for each cycle once the cross-entropy error for the validation set was minimized using gradient descent, and the weights of the network were then used for the testing set.

The testing procedure involved computing the weighted sum of the 50-element test set using the weights from training, then again applying the softmax function. A simple max function was used to judge if the testing outputs matched the teaching signals, and to create the confusion matrix. However, since the weights were trained with six facial expressions, the max function was applied only among the four target emotions to create a four-way forced choice similar to what a human subject would have to perform. This is valid because the softmax function created outputs that were probabilities of each emotion being correct; thus, although the teaching signal was “all or nothing,” the outputs were not. We note that when humans undergo the task of selecting among four target emotions, it is entirely possible that the emotion they perceive is not among the four options, and thus they may have to answer with their second or third choice. This is essentially what we have emulated with our network.

### Results

Our model was able to fit the human data very well in several measures. Much of the data presented by Yu, Chai, & Chung (2011) is displayed in the form of confusion matrices, which pit the human responses against the actual targets. We used the same technique to display our data. Since Yu, Chai, & Chung (2011) presented their results on a poster, many of their figures lack numerical data. As such, much of our analysis will be dependent on comparing the visual presentations of data. The color spectrum of our confusion matrices were closely matched to that used by Yu, Chai, & Chung (2011) so that comparisons and conclusions can be made.

### Performance on Unfiltered Images

Figures 4a and 4b show the confusion matrix for the unfiltered images given by Yu, Chai, & Chung (2011) and by our model, respectively. In addition to the hits, the columns show false alarms and rows depict misses.

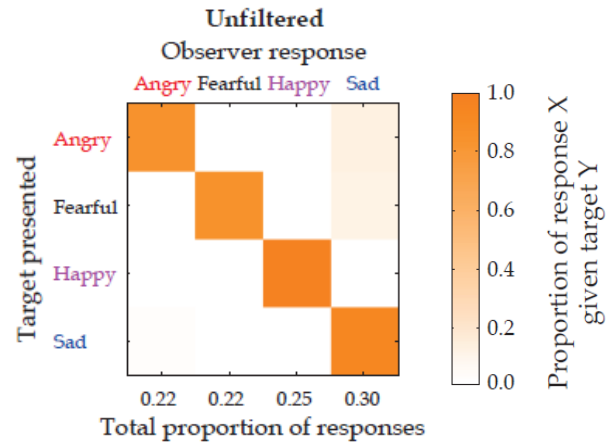


Figure 4a: Presented target vs. human responses (unfiltered) as presented in Yu, Chai, & Chung (2011).

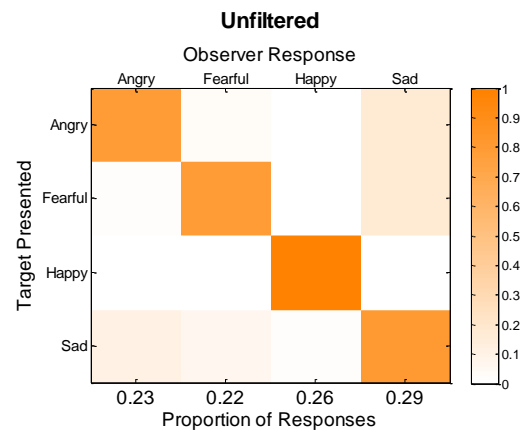


Figure 4b: Presented target vs. model responses, for unfiltered images.

Both the model and the human subjects demonstrated very good performance overall in recognizing the unfiltered images. The model also exhibited similar behavior as the human subjects in terms of having false alarms for sad faces when presented with Angry and Fearful faces. This is likely a characteristic of the closed-mouth sad faces in the NimStim set in general, since similar false alarms were present in Tottenham et al. (2009). The total proportion of responses was also similar between EMPATH and the human subjects, which suggests that the model was sensitive to many of the same facial features that the human subjects used for classification.

### Performance for individual filters

The results for the individual filters demonstrate that recognition performance decreased as the filter orientations approached 90°. Figures 5a and 5b illustrate the performances of humans and of our model, respectively.

Both sets of confusion matrices distinctly show greater occurrences of misses and false alarms at orientations near the vertical; i.e. 60°, -60°, and 90°. Some other general trends can be drawn from the data. For both humans and the model, sad faces tended to draw more false alarms and misses, regardless of the filter condition. Angry expressions tended to lose their uniqueness as the filters neared vertical, resulting in many misses, and few hits and false alarms.

One informative visualization of recognition performance is plotting the d prime calculations for each filter, which indicates how strong a signal is in relation to surrounding noise (Abdi, 2010). Hence, the d prime calculation for each filtering condition is proportional to how recognizable the expression is with that filter. Figures 6a and 6b depict graphs of d primes for each filtering condition normalized to the d prime of the unfiltered images (higher d primes still correlate to higher recognition performance).

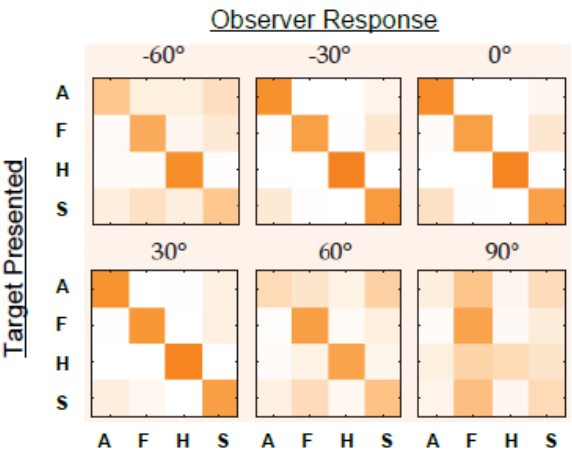


Figure 5a: Human performance for each filtering condition. From Yu, Chai, & Chung (2011).

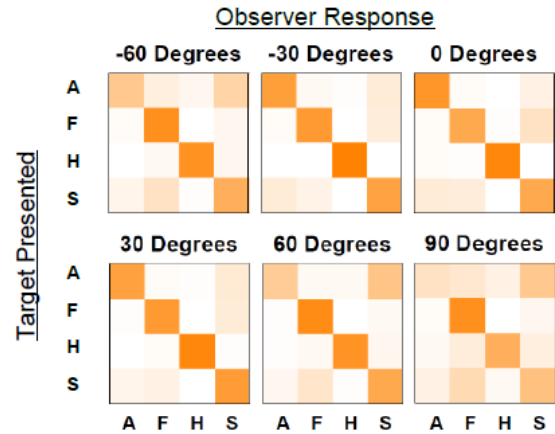


Figure 5b: EMPATH performances for each filtering condition.

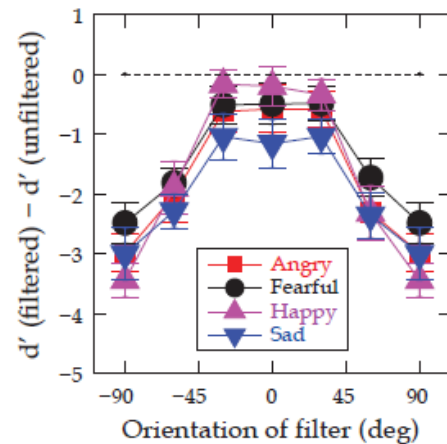


Figure 6a: Human data from Yu, Chai, & Chung (2011), showing d primes of each filtering condition, normalized to the unfiltered condition. Note that data for -90° was copied from data for 90°.

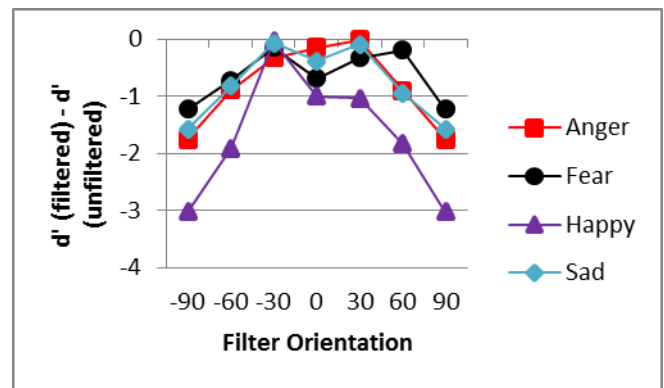


Figure 6b: Data from the EMPATH model showing d primes of each filter normalized to the unfiltered images.

Both d prime charts demonstrate lower recognition performance as the filters approached 90°. We note that

results from the EMPATH model are not as symmetrical as those from the human data (e.g. the discrepancy between the values for -30 and 30 degrees for happiness, and the M-shaped graph for fear). This is likely due to the fact that the original images are not vertically balanced; i.e. the positive and negative filters each obstruct slightly different features of the expressions. The resulting images were likely different enough to confuse the network. It would be worthwhile in the future to explore this phenomenon of asymmetry, especially since it was not apparent in the human data.

As noted earlier, fearful faces were less affected by the filter orientations as the other three emotions. Both d prime charts show that fear was the most easily recognized at the 90° orientation. The earlier confusion matrices (Figures 5a and 5b) likewise depict a relatively steady percentage of hits for fearful faces. Much of this is attributed to the fact that both human observers and EMPATH exhibited a significant bias towards fear at the vertical orientations, which increased the occurrences of both hits and false alarms. Figure 7 illustrates EMPATH’s high proportion of responses for fear at the vertical orientations. At the horizontal orientations, each emotion constituted close to 25% of the responses, but at the vertical orientation, responses in favor of fear approached 40%.

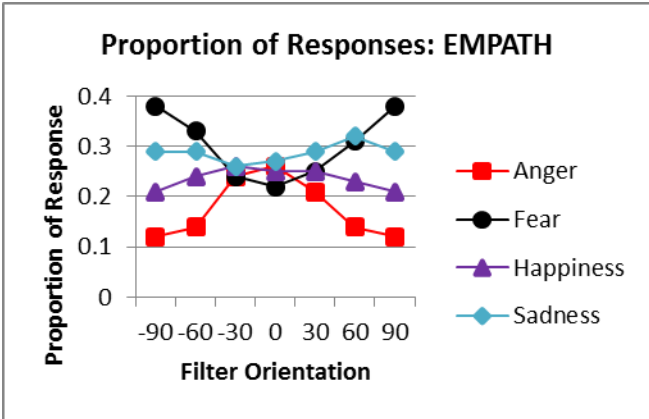


Figure 7: Total proportion of responses exhibited by EMPATH for each filter orientation.

### Aggregate Performance of 6 filters

EMPATH’s cumulative performance across all 6 filters is also similar to that of the human observers. Tables 1a and 1b show the overall performance of the human subjects and of EMPATH, respectively.

Firstly, the two tables show that with the exception of anger, EMPATH does significantly better in recognizing expressions than do the human subjects. Of course, the model can always be tailored to perform either better or worse, but we did not want to make adjustments just to suit these images. Secondly, the two tables depict many similar trends in the responses between the humans and the model. The overall proportion of responses shows that both the humans and the model were biased towards fear and

sadness, and that both were biased against anger. In both cases, anger was the most difficult expression to recognize and also the most difficult with which to be confused, based on the overall percentage of responses. The human subjects and EMPATH also had difficulty recognizing sad faces, but there was a high false-alarm rate as well. The Spearman rank correlation between the two matrices was very good, at  $r = 0.976$  ( $p < 0.001$ ) for the complete matrices. Since we were also interested in the misses and false alarms, we also calculated the rank just for the off-diagonals, which was very similar at  $r = 0.942$  ( $p < 0.001$ ).

Table 1a: Aggregate performance of human subjects for all 6 filter orientations.

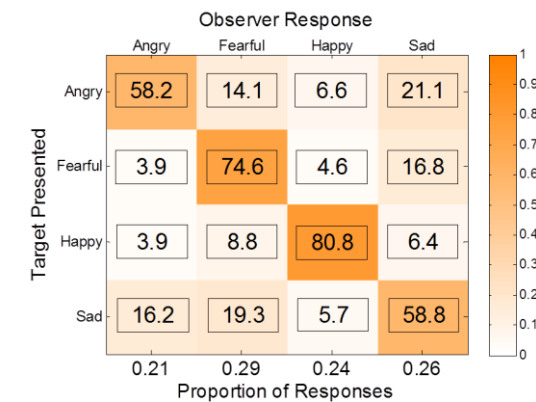
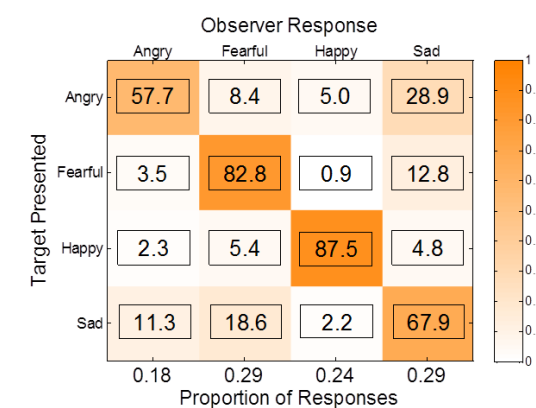


Table 1b: Aggregate performance of EMPATH for all 6 filter orientations



### Discussion

The aim of this present experiment was to model the experiment conducted by Yu, Chai, & Chung (2011) and determine if their results can be replicated using a neural network that was not specifically tuned to their images or their data. Our results demonstrate a strong similarity to the pattern of human responses, particularly in showing that information for facial expressions lies primarily on the horizontal orientation, with the modest exception of fearful faces, which solicited heavy bias from both human observers and EMPATH as the orientation approached



vertical. In particular, there was a very high proportion of hits and false alarms, suggesting that the vertical filter accentuated features in other expressions normally attributed to fear. Based on this data, it seems that much of the information for fear lies on the vertical, making it distinct from other expressions. It would be interesting to conduct further experiments with other image sets to determine if this phenomenon is a trait of the NimStim images or if it is more universal.

Some discussion regarding our use of  $d'$  prime is warranted. The procedure used by Yu, Chai, & Chung (2011) to calculate  $d'$  prime follows the standard guideline of  $d' = Z_H - Z_{FA}$ , where  $Z_H$  and  $Z_{FA}$  denote the inverse Gaussian distribution of hits and false alarms, respectively (Abdi, 2010). However, since this formula is typically used for two-way “Yes – No” tasks, the validity of using it for a four-way forced choice is debatable, since each emotion has one “Yes” response and three distinct “No” responses attached to it. Very little literature exist detailing  $d'$  prime calculations for multiple-way forced choice scenarios, but Alexander (2006) described an easily-computed approximation to the original version in Green & Swets (1966). Based on that, we have recalculated our graph of  $d'$  prime, which we depict in Figure 8. It should be noted that this approximation does not take false alarms into account. This resulted in a significantly higher  $d'$  prime for fearful expressions, which were actually greater for filtering conditions near vertical than for unfiltered images. Given that this is an approximation, the validity may of course also be debated, but we nonetheless present both calculations. This serves as a prediction of how the Yu, Chai, & Chung (2011) data will look if it were analyzed in the same way.

Given EMPATH’s demonstrated consistency in modeling human face recognition, another possible future experiment could be to determine which filtering orientations are ideal for recognition of each particular expression. It seems, based on this study, that the majority of expressions with the exception of fear would have an ideal filtering condition near the horizontal, but determining exact orientations would form testable hypotheses generated by the model.

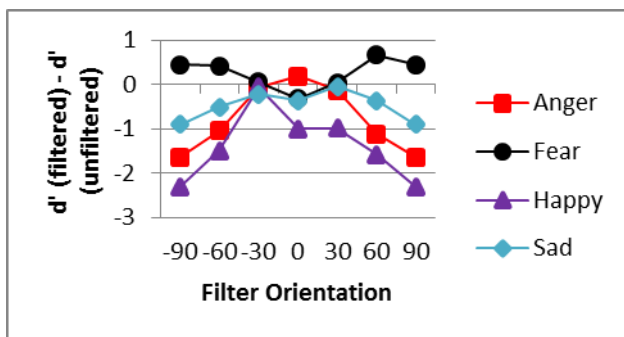


Figure 8: N-choice  $d'$  prime approximations, following procedure outlined by Alexander (2006) and normalized to the unfiltered images.

## Acknowledgements

We thank Dion Yu and Susana Chung for sharing their filtering code, and for providing us with ample clarification of their experiment. We thank Nim Tottenham for sharing the NimStim dataset with us. We also thank Gary’s Unbelievable Research Unit (GURU) for providing feedback on this project, and Matthew Tong in particular for overseeing part of the reconstruction of EMPATH. This work was supported in part by NSF grant #SBE-0542013.

## References

- Abdi, H. (2010). Signal detection theory (SDT). In Peterson, P.L. Baker, E., & B. McGaw (Eds.). *International Encyclopedia of Education*. New York: Elsevier.
- Alexander, J.R.M. (2006) An approximation to  $d'$  for n-alternative forced choice. University of Tasmania technical report, available at [eprints.utas.edu.au](http://eprints.utas.edu.au).
- Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.
- Costen N.P., Parker D.M., & Craw I (1996). Effects of high-pass and low-pass spatial filtering on face identification. *Percept Psychophys*, 58, 602–612.
- Dailey M.N., Cottrell G.W., Padgett C., & Adolphs, R. (2002). EMPATH: a neural network that categorizes facial expressions. *J. Cog. Neuro.*, 14, 1158-1173.
- Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2, 1160–1169.
- Ekman, P., & Friesen, W. (1976). *Pictures of facial affect*. Palo Alto, CA: Consulting Psychologist Press.
- Gold J., Bennett P.J., & Sekuler A.B. (1999). Identification of band-pass filtered letters and faces by human and ideal observers. *Vision Res*, 39, 3537–3560.
- Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Nasanen R. (1999). Spatial frequency bandwidth used in the recognition of facial images. *Vision Res*, 39, 3824–3833.
- Williams, N.R., Willenbockel, V. & Gauthier, I. (2009). Sensitivity to spatial frequency and orientation content is not specific to face perception. *Vision Res*, 49, 2353–2362.
- Tottenham N., Tanaka J.W., Leon A.C., McCarry T., Nurse M., Hare T.A., Marcus D.J., Westerlund A., Casey B.J., Nelson C. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Research*, 168, 242-249.
- Yu D., Chai A., & Chung S.T.L. (2011). *Orientation information in encoding facial expressions*. Poster presented at the Vision Sciences Society 2011 Annual Meeting, Naples, Florida.
- Yu D. & Chung S.T.L. (2011). Critical orientation for face identification in central vision loss. *Optometry and Vision Science*, 88, 724-732.

# Learning Image-Derived Eye Movement Patterns to Characterize Perceptual Expertise

**Rui Li (rxl5604@rit.edu)**

College of Computing and Information Science, 1 Lomb Memorial Drive  
Rochester, NY 14623 USA

**Jeff Pelz (pelz@cis.rit.edu)**

College of Imaging Science, 1 Lomb Memorial Drive  
Rochester, NY 14623 USA

**Pengcheng Shi (spcast@rit.edu) Anne R. Haake (arhics@rit.edu)**

College of Computing and Information Science, 1 Lomb Memorial Drive  
Rochester, NY 14623 USA

## Abstract

Experts have remarkable capability of locating, identifying and categorizing objects in their domain-specific images. Eliciting experts' visual strategies will benefit image understanding by transferring human domain knowledge into image-based computational procedures. In this paper, an experiment conducted to collect both eye movement and verbal description data from three groups of subjects with different medical training levels (eleven board-certified dermatologists, four dermatologists in training and thirteen novices) while they were examining and describing 42 photographic dermatological images. We present a hierarchical probabilistic framework to discover the stereotypical and idiosyncratic viewing behaviors exhibited within each group when they are diagnosing medical images. Furthermore, experts' annotations of thought units on the transcribed verbal descriptions are time-aligned with discovered eye movement patterns to interpret their semantic meanings. By mapping eye movement patterns to thought units, we uncover the manner in which these subjects alternated their behaviors over the course of inspection and how these experts parse the images.

**Keywords:** Eye movements; eye tracking; verbal description; multimodal data analysis; graphical model; user study; diagnostic reasoning

## Introduction

Perceptual expertise is considered to be the crucial cognitive factor accounting for the advantage of highly trained experts (Hoffman & Fiore, 2007). Experts generate distinctively different perceptual representations when they view the same scene as novices (Palmeri, Wong, & Gauthier, 2004 ; Smuc, Mayr, & Windhager, 2010). Rather than passively "photocopying" the visual information directly from sensors into minds, visual perception actively interprets the information by altering perceptual representations of the images based on experience and goals. By analyzing the whole sequences of fixation and saccadic eye movements from groups with different expertise levels, significant differences in visual search strategies between groups show human expertise plays a great role in medical image examination. In (Manning, Ethell, Donovan, & Crawford, 2006), the nature of expert performance of four observer groups with different levels of expertise was investigated . They compared multiple eye movement measures and suggested these distinctive

variations among the observations of the better performance from higher expertise level are due to the consequences of experience and training. In (Krupinski et al., 2006), an eye movement study was conducted on diagnostic pathology of light microscopy to identify distinctive viewing stereotypes for each level of experience . Their results suggest eye movement monitoring could serve as a basis for the creation of innovative pathology training routines.

In knowledge-rich domains, perceptual expertise is particularly valuable. Medical image understanding via manually marking and annotating become not only labor intensive for experts but also ineffective because of the variability and noise of experts' performance (Gordon, Lotenberg, Jeronimo, & Greenspan, 2009). For training and designing decision support systems, the basic perceptual strategies and principles of diagnostic-reasoning are also desired (Dempere-Marco, Hu, & Yang, 2011). To address this problem, it requires the ability of extracting and representing experts' perceptual expertise in a form that is ready to be applied. In this work, our contributions are: first, we discover and represent expertise-related eye movement patterns exhibited among multiple experts in an objective and unbiased way; second, to validate these patterns, we identify their semantic meanings by time-aligning them with standardized thought units annotated by additional experts. Third, we also characterize the eye movement patterns of three different expertise levels respectively which can be used to categorize users' expertise levels based on their visual inspection on medical images.

Human viewing behaviors are valuable yet effortless resources worth of exploiting. In specific domains experts perceptual expertise is considered to be more consistent and informative than their manual markings. Human vision is an active dynamic process in which the viewer seeks out specific information to support ongoing cognitive and behavioral activity (Henderson & Malcolm, 2009). Since visual acuity is limited to the foveal region and resolution fades dramatically in the periphery, we move our eyes to bring a portion of the visual field into high resolution at the center of gaze. Studies have shown that visual attention is influ-



enced by two main sources of input: bottom-up visual attention driven by low-level saliency image features and top-down process in which cognitive processes, guided by the viewing task and scene context, influence visual attention (Torralba, Oliva, Castelano, & Henderson, 2006 ; Loboda, Brusilovsky, & Brunstein, 2011). Growing evidence suggests that top-down information dominates the active image viewing process and the influence of low-level saliency guidance is minimal (Castelano, Mack, & Henderson, 2009). These theoretical outcomes provide us with the possibility to capture experts' cognitive strategies, perceptual expertise and expectations by investigating their stereotypical and idiosyncratic viewing behaviors, and decode their semantic meanings.

In our work we focus on medical images where domain knowledge and perceptual expertise are in demand. We elicit and model physicians' perceptual and conceptual expertise from their diagnostic reasoning process while inspecting medical images. Physicians examine medical images and verbally describe their thinking process as if teaching a trainee, and both their eye movements and verbal descriptions are recorded. In order to capture the stereotypical and idiosyncratic eye movement patterns exhibited among these physicians, we develop a hierarchical dynamic model. This model allows us to build a library of all the patterns exhibited by physicians' time-evolving eye movement series (scanpaths) and each eye movement pattern essentially corresponds to a particular statistical regularity of the temporal-spatial properties inferred from multiple eye movement series. Thus each physician's eye movement time series can be characterized by a particular combination of a subset of these patterns from this library. To investigate the relationships between visual and verbal conceptual processing by analyzing the verbal descriptions, additional experts annotate the transcribed verbal descriptions using standardized semantic labels (thought units) that describe the process of creating a differential diagnosis from their domain knowledge (Habif, Jr., Chapman, Dinulos, & Zug, 2005). After time-aligning these thought units with the eye movements patterns, we discovered significant correlations between them. This results indicate that the patterns we extracted from eye movement data possess distinct and specific semantic meanings in terms of human capabilities of image understanding.

## Experiment

Subjects recruited for the eye tracking experiment belong to three groups based on their training level including 11 board-certified dermatologists (attending physicians), 4 dermatologists in training (residents) and 13 undergraduate lay people (novices). We also recruited physician assistant students who served as "trainees" in order to motivate dermatologists to verbalize their diagnosis reasoning using the Master-Apprentice scenario, which is known to be effective in eliciting detailed descriptions.

A SMI (Senso-Motoric Instruments) eye tracking apparatus was applied to display the stimuli at a resolution of

1680x1050 pixels for the collection of eye movement data and recording of verbal descriptions. The eye tracker was running at 50 Hz sampling rate and has accuracy of 0.5° visual angle. The subjects viewed the medical images binocularly at a distance of about 60 cm. The experiment was conducted in an eye tracking laboratory with ambient light.

A set of 42 dermatological images, each representing a different diagnosis, was selected for the study. These images were presented to subjects on the monitor. Medical professionals were instructed to examine and describe each image to the students while working towards a diagnosis, as if teaching. The experiment lasted approximately 1 hour. The medical professionals were instructed not only to view the medical images and make a diagnosis, but also to describe what they see as well as their thought processes leading them to the diagnosis. The novice observers were instructed to examine the images and offer a detailed description as if describing to their doctors over the phone. Both eye movements and verbal descriptions were recorded for the viewing durations controlled by each subject. The experiment started with a 13-point calibration and the calibration was validated after every 10 images. Calibration is accepted if its variance is less than 0.5°. The audio recordings of the verbal descriptions from the dermatologists were transcribed and annotated.

An annotation study was conducted on the transcripts to investigate the semantic interpretations of the estimated eye movement patterns. During annotation two highly trained dermatologists identified 9 thought units. A thought unit is a single word or group of words that receives a descriptive label based on its semantic role in the diagnostic process. The thought unit labels are patient demographics (DEM), body location (LOC), configuration (CON), distribution (DIS), primary morphology (PRI), secondary morphology (SEC), differential diagnosis (DIF), final diagnosis (Dx), and recommendations (REC). Words not belonging to a thought unit were designated as 'None'. These two physicians annotated transcribed verbal descriptions with these thought units. The annotation were then time-aligned with eye movement patterns. Using this method, each unit of eye movement data, which is composed of a fixation and its successive saccade, receives two labels: one is its pattern indicator inferred by the model and the other is its time-aligned thought unit annotated through the consensus of multiple experts. This result allows us to interpret the eye movement patterns by measuring the correspondence between them and the thought units.

## Hierarchical Dynamical Model

A hierarchically-structured dynamical model was developed to capture both the common eye movement patterns shared among multiple expertise-specific groups of subjects and unique eye movement patterns exhibited by individuals. The hierarchical beta processes proposed by Thibaux et al.(Thibaux & Jordan, 2007) as a prior distribution of our model provides the flexibility of discovering more patterns as new eye movement data are observed. Since fixation and



$\{(\theta_{jk}, p_{ijk})\}$  as subject  $i$ 's personal subset of eye movement patterns given group  $j$ , as shown in Figure 1.

The transition distribution  $\pi_{ij} = \{\pi_{z_q(ij)}\}$  of the hidden Markov model at the bottom level governs the transitions between the  $i^{th}$  subject's personal subset of eye movement patterns  $\theta_{jk}$  of group  $j$ . It is determined by the element-wise multiplication between the eye movement subset  $\{p_{ijk}\}$  of subject  $i$  in group  $j$  and the gamma-distributed random variables  $\{e_{ijk}\}$ :

$$e_{ijk}|\gamma_j \sim \text{Gamma}(\gamma_j, 1) \quad (5)$$

$$\pi_{ij} \propto E_{ij} \otimes P_{ij} \quad (6)$$

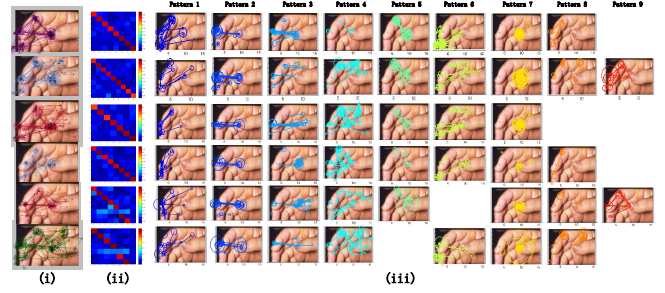
where  $E_{ij} = [e_{ij1}, \dots, e_{ijK_j}]$ . So the effective dimensionality of  $\pi_{ij}$  is determined by  $P_{ij}$ , which is inferred from observations.

We use Markov chain Monte Carlo sampler to do the posterior inference over this model. In one iteration of the sampler, each latent variable is visited and assigned a value by drawing from the distribution of that variable conditional on the assignments to all other latent variables as well as the observation. In particular, based on the sampling algorithm proposed in (Thibaux & Jordan, 2007), we developed a Gibbs sampling solution to the hierarchical beta processes part of the model.

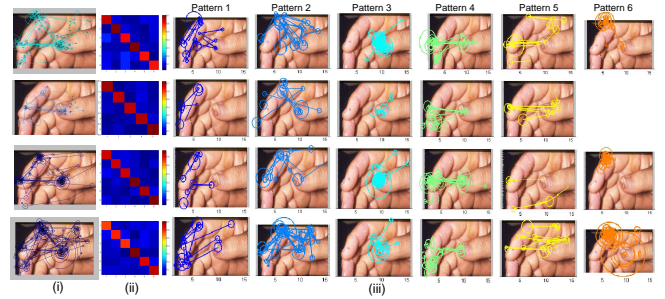
## Results and Discussion

In Figure 2, we illustrate one set of observed data and estimating processes from the framework of the 11 dermatologists diagnosing a case of a skin manifestation of endocarditis. In the medical image, there are multiple skin lesions spreading over the thumb nail and tip, the two parts of index finger and the middle finger as marked in (b) of Figure 2. A primary abnormality is on the thumb tip. The scanpaths in Figure 3a (i) indicate that dermatologists fixated on the primary abnormality heavily and switch their visual attention actively between and within the primary and secondary findings. The estimated patterns are color-coded as panels shown in (d) of Figure 2. These panels describe the time-evolving manner in which each individual alters eye movement patterns at the individual level.

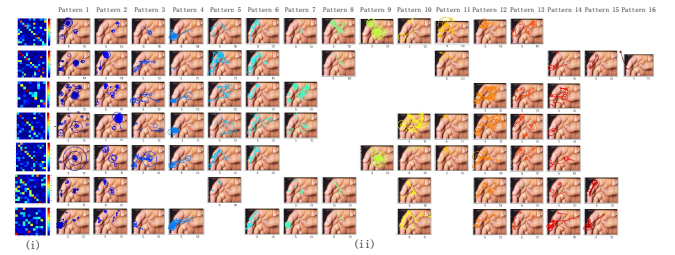
Pattern occurrence and thought unit alignment resulted in assignment of each fixation to a specific pattern and to a thought unit (or None). Initial integration of eye movement patterns with thought units was accomplished by calculating correspondence in Figure 4. Analysis shows, for example, that primary morphology (PRI) is closely related to the combination of two specific patterns: Pattern 2 is characterized by fixations switching between the primary and the different secondary abnormalities; and Pattern 7 by long fixations only on the primary abnormality. These patterns suggest dermatologists were seeking meaningful ways to integrate these two findings for some principled reasons, although these informative findings are separable in the sense that they are operationally defined and measured independently of one another. Pattern 7 has strong relationship to location (LOC) which ap-



(a) Nine inferred eye movement patterns from the 11 attendings. In (i) 6 attendings' scanpaths are super-imposed onto the image. (ii) shows the transition probability matrices of the nine eye movement patterns within the six scanpaths during diagnosis, which indicate the patterns are persistent. In (iii) the eye movement patterns are segmented from these corresponding scanpaths.



(b) Six inferred eye movement patterns from the 4 residents. In (i) the scanpaths of the 4 residents. (ii) shows the transition probability matrices of six eye movement patterns. In (iii) the eye movement patterns are segmented from these 4 scanpaths.



(c) Sixteen inferred eye movement patterns from the 13 novices. In (i) the transition probability matrices of the sixteen eye movement patterns, which suggest novices' visual behaviors are not persistent. In (ii) the patterns are segmented from these 7 scanpaths.

Figure 3: The inferred eye movement patterns of the three expertise-specific groups. Each observation unit of the eye movement sequences is composed of 4 components: fixation location (xy coordinate), fixation duration and saccade amplitude. We then apply our model on these sequential data to reveal the subtlety of the behavioral patterns varying over time. The inferred patterns were derived with 4 chains of 55000 sampling iterations. The color coding specifies the segments of each specific pattern.

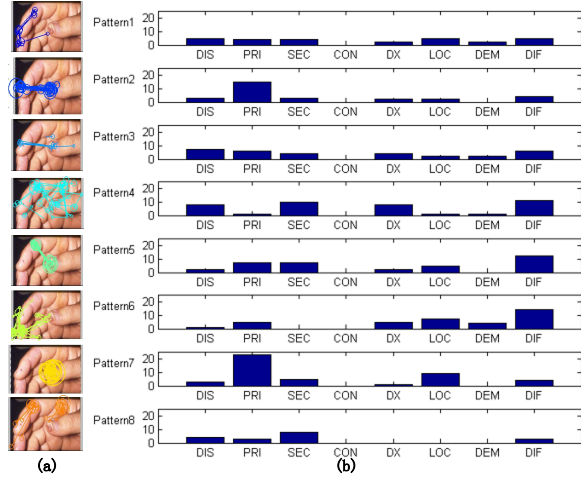


Figure 4: Correspondence between the 8 eye movement patterns of the 11 attendings (the rows) and their 8 thought units (the columns). (a) the representative patterns. (b) histograms show the corresponding relationship between discovered eye movement patterns and annotated thought units. For each pattern we plotted the counts of fixations which are labeled as the 9 thought units.

pears to correspond to the primary morphology location. Pattern 4 consists of scanpath segments which are characterized by shorter fixation durations and longer saccades. This scanning behavior strongly corresponds to thought units, including distribution (DIS), secondary morphology (SEC), diagnosis (DX) and differential diagnosis (DIF). Scanning pattern coupled with thought unit DX is possibly related to confirmation of secondary findings to support or rule out diagnostic hypotheses.

Some similar patterns also emerged in the resident group but is lacking in the novice group as shown in Figure 3b. This suggests that experts, equipped with domain knowledge organized in finer gradations of functional categories, can discriminate the significance of their findings in a particular context. In contrast, in Figure 3c the novices failed to do so, although they perceive the same abnormalities too. Compare Figure 3a (ii), Figure 3b (ii) and Figure 3c (i), the difference between the transition probability matrices of the three expertise-specific groups suggests professionals' eye movement patterns are more persistent than the novices'.

These results suggest that there exist structural regularities of experts' diagnostic-reasoning processes, and such perceptual and conceptual processing regularities can be captured and manifested through experts' eye movements. This is consistent with previous empirical studies (Patel, Arocha, & Kaufman, 2001). These discovered stereotypical eye movement patterns indicate that experts are able to rapidly invoke the appropriate specific knowledge and expertise, and initially detect a general pattern of disease. These capabilities lead them to a gross anatomic localization and narrow down the possible interpretations. On the other hand, novices have hard

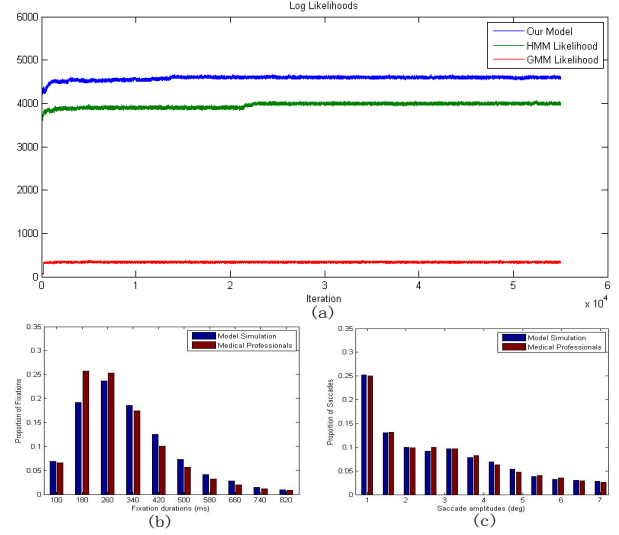


Figure 5: Quantitative performance evaluations. (a) The likelihood-value plots of a Gaussian mixture model, a hidden Markov model and our model after 55000 sampling iterations on our data-set. (b) The histogram of the fixation duration distributions of the 15 professionals (attendings and residents) and our model's simulations over 42 images. (c) The histograms of the saccade amplitude distributions of the 15 professionals and our model's simulations over 42 images.

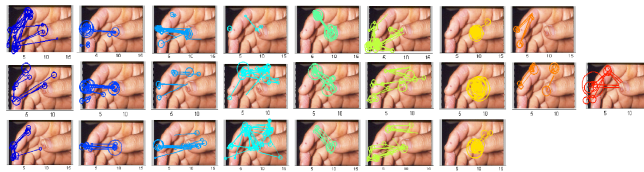
time to focus on the important structures and are more likely to maintain inappropriate interpretations.

To measure performance, we compared the log-likelihood values among our model, a hidden Markov model (HMM) and a Gaussian mixture model (GMM) as shown in Fig. 5 (a). To implement the HMM and GMM, we have to assume each eye movement sequence exhibits the same set of patterns. The log-likelihood values of our model and the GMM are 4000 vs. 300, which indicates our model fits the observation better. One possible cause is that the GMM makes a strong assumption that the eye movement data are independent which is hardly true. On the contrary, our model only assumes that the eye movement patterns are exchangeable in order. Additionally, our model and the HMM take sequential information of eye movements into account. We visualized the eye movement patterns from HMM and GMM in Fig. 6 (b)-(c) and make a comparison with our model's results in Fig. 6 (a). In Fig. 5 (b)-(c), our model generated 7356 fixation-saccade units to simulate the 15 professionals. It is worth noting that this result also validates the discovered eye movement patterns. Such simulation requires us to generate a set of realizations of eye movement patterns first from the hierarchical prior, simulate multiple possible sequences of these patterns, and then draw fixation-saccade samples from them.

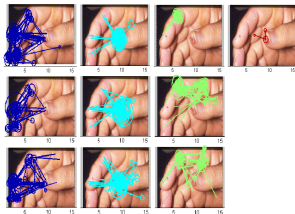
## Conclusions

Our approach identified and semantically interpreted both stereotypical and idiosyncratic expertise-specific eye move-

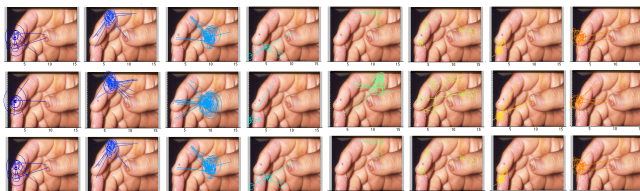




(a) 3 attendings' eye movement patterns inferred from our model.



(b) The same 3 attendings' eye movement patterns from the HMM



(c) The same 3 attendings' eye movement patterns from the GMM.

Figure 6: Illustrations of the attendings eye movement patterns from the three models.

ment patterns that only exist over time. In our future work, the discovered eye movement patterns will be related to image features by projecting the patterns from their temporal-spatial space into the image feature space. We will not only identify the most valuable image feature sets leading to a correct diagnosis but also uncover how a particular feature's importance changes over the course of the diagnostic reasoning process. These discoveries will provide training information to novices on how to look for relevant image features. Evaluation of a subject's expertise level is another future study. We can identify the expertise level given a subject's visual interaction with test images through calculating the model's posterior probability. Compared to simply calculating diagnosis error rates to evaluate expertise level, our approach can unveil which diagnostic reasoning steps lead to wrong diagnosis and the possible cognitive factors such as misconception, miscategorization and misperception, and form the basis of support systems.

## Acknowledgements

We thank M.D. Cara Calvelli and Dr. Cecilia Ovesdotter Alm for many helpful discussions. This work was supported by research grants from the National Science Foundation (IIS-0941452) and the NIH/NLM (R21 LM010039-01).

## Références

- Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing Task Influences Eye Movement Control during Active Scene Perception. *J. Vision*, 9(3), 1-15.
- Dempere-Marco, L., Hu, X., & Yang, G.-Z. (2011). A novel framework for the analysis of eye movements during visual search for knowledge gathering. *Cognitive Computation*, 3, 206–222.
- Gordon, S., Lotenberg, S., Jeronimo, J., & Greenspan, H. (2009). Evaluation of uterine cervix segmentations using ground truth from multiple experts. *J. Computerized Medical Imaging and Graphics*, 33(3), 205-216.
- Habif, T. P., Jr., J. L. C., Chapman, M. S., Dinulos, J. G., & Zug, K. A. (2005). Skin disease diagnosis and treatment. Elsevier Mosby.
- Henderson, J. M., & Malcolm, G. L. (2009). Searching in the Dark Cognitive Relevance Drives Attention in Real-world Scenes. *Psychonomic Bulletin and Review*, 16(5), 850-856.
- Hoffman, R., & Fiore, M. S. (2007). Perceptual (re)learning : a leverage point for human-centered computing. *J. Intelligent Systems*, 22(3), 79-83.
- Krupinski, E., Tillack, A., Richter, L., Henderson, J., Bhat-acharyya, A., Scott, K., et al. (2006). Eye-movement study and human performance using telepathology virtual slides. implications for medical education and differences with experience. *Journal of Human Pathology*, 37(12), 1543–1556.
- Loboda, T. D., Brusilovsky, P., & Brunstein, J. (2011). Inferring word relevance from eye-movements of readers. In *Proc. iui* (pp. 175–184). ACM Press.
- Manning, D., Ethell, S., Donovan, T., & Crawford, T. (2006). How do radiologists do it? the influence of experience and training on searching for chest nodules. *Journal of Radiography*, 12(2), 134–142.
- Palmeri, T. J., Wong, A. C.-N., & Gauthier, I. (2004). Computational approaches to the development of perceptual expertise. *TRENDS in Cognitive Sciences*, 8(8), 378–386.
- Patel, V. L., Arocha, J. E., & Kaufman, D. R. (2001). A primer on aspects of cognition for medical informatics. *J Am Med Inform Assoc.*, 8, 324–343.
- Smuc, M., Mayr, E., & Windhager, F. (2010). The game lies in the eye of the beholder: The influence of expertise on watching soccer. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 1631–1636). Austin, TX : Lawrence Erlbaum Associates.
- Thibaux, R., & Jordan, M. I. (2007). Hierarchical beta processes and the indian buffet process. *J. Machine Learning and Research*, 22(3), 25-31.
- Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, 113(4), 766–786.

# Multimodal Temporal Perception in Musicians: Evidence for Both Segregated and Supramodal Attentional Systems?

**Ahnate Lim (ahnate@hawaii.edu)**

Department of Psychology, University of Hawaii at Manoa  
2530 Dole Street, Honolulu, HI 96822 USA

**Scott Sinnett (ssinnett@hawaii.edu)**

Department of Psychology, University of Hawaii at Manoa  
2530 Dole Street, Honolulu, HI 96822 USA

## Abstract

Although musical training has been correlated with modulations of early perceptual and attentional processes, the majority of investigations neglect the possibility of cross modality enhancements. We investigated the effects of musical training by measuring spatial and temporal attention in a temporal order judgment task in auditory, visual, and crossmodal conditions with and without non-predictive cues. In Experiment 1, musicians had lower detection thresholds when compared to controls in all conditions (marginal in auditory). Experiment 2 showed mixed findings, with musicians demonstrating reduced capture from visual cues on the visual task compared to controls, and lower detection thresholds on the auditory task with visual cues. Adding spatial cues to the temporal order judgment tasks increased temporal thresholds for both groups, but only when they occurred within the same modality as the task, and not when presented in a different modality. The findings support both supramodal and segregated accounts of attentional resources.

**Keywords:** attention; perception; musicians; temporal order judgment; multisensory; visual; auditory; crossmodal

## Introduction

The human attentional system is impressively competent at processing information, considering how the efficiency and selectivity of attention facilitates perception and goal directed behavior amidst a constant plethora of stimuli. Interestingly, the neurological underpinnings of attention may change under certain conditions. This “plasticity” has been associated with compensations for losses in one sensory modality with enhancements in another modality (Röder et al., 1999). Furthermore, improved behavioral performance may also occur as a side effect of specific activities or hobbies such as video game playing (Granek, Gorbet, & Sergio, 2010; Green, Li, & Bavelier, 2010, but see Boot, Blakely, & Simons, 2011) and musical training (Hodges, Hairston, & Burdette, 2005; Lim & Sinnett, 2011).

Although the topic of non-musical cognitive benefits (e.g., mathematics, spatial-reasoning and linguistics) occurring as a result of musical training has been the focus of much research (for a summary, see Rauscher, 2003), there has been less emphasis on the effects of precise mechanisms of attention and perception. There is evidence from numerous studies conducted with musicians suggesting greater neuroplasticity when compared with non-musicians (see for example, Gaser & Schlaug, 2003; Münte, Altenmüller, & Jäncke, 2002). Although it should be noted

that these brain differences could equally be attributed to a predisposition that leads people to become musicians, rather than any specific training related enhancement. Even so, behavioral evidence from studies comparing musicians to non-musicians demonstrates improved perceptual abilities on various tasks in different modalities. These have included visual perceptual speed and discrimination (Helmbold, Rammsayer, & Altenmüller, 2005; Patston, Hogg, & Tippet, 2007) as well as auditory temporal discrimination (Hodges et al., 2005; Jones & Yee, 1997).

These studies also highlight an interesting possibility of training effects on attention: crossmodal enhancements (e.g., visual enhancements after auditory training). Some authors have suggested that the attentional system operates in a supramodal fashion, with all senses having access to a single reservoir of attentional resources (see Farah, Wong, Monheit, & Morrow, 1989; Pavani, Husain, Ládavas, & Driver, 2004; but see also Sinnett, Costa, & Soto-Faraco, 2006; C Spence & Driver, 1996; Wickens, 1984, for examples of a segregated attentional system). Thus, by testing musicians for enhanced attentional and perceptual capabilities in the visual modality, we can indirectly assess whether training in one sense (i.e., auditory musical training) leads to performance enhancements in another. This would provide support for a supramodal attentional system, and would closely align with recent investigations involving video game players, where auditory enhancements were observed despite the training being mostly visual based (Donohue, Woldorff, & Mitroff, 2010; Green, Pouget, & Bavelier, 2010).

The temporal order judgment (TOJ) task is an ideal tool to assess temporal processing differences between musicians and non-musicians. More importantly, the TOJ task can be presented under both unimodal and crossmodal conditions. The task requires participants to determine the correct order of subsequently presented stimuli, and allows for two measures of perceptual processing to be calculated: the just noticeable difference (JND), and the point of subjective simultaneity (PSS). The JND is a measure of the resolution or threshold of temporal discrimination, while the PSS is the time in which one stimulus can be presented before the other such that they are still perceived as occurring simultaneously (e.g., in a crossmodal task, it can indicate whether auditory or visual stimuli must be presented first for them to be perceived as simultaneous).

Humans are generally proficient at temporal discrimination. In studies examining within and cross-modal (visual, auditory, tactile) TOJs, Hirsh and Sherrick Jr. (1961) found that participants could discriminate temporal order between stimuli (JND) when presented as quickly as 20ms apart. Crucially, in crossmodal tasks (i.e., audiovisual presentations) the visual stimuli had to lead auditory stimuli by approximately 40-80ms for participants to perceive them as being presented simultaneously (PSS; see also Zampini, Shore, & Spence, 2003). Furthermore, research suggests that the resolution of temporal acuity is better in the auditory modality than in vision or touch (Chen & Yeh, 2009).

Given the efficacy at which humans can discriminate temporally, it is worth noting that significant gains or losses in TOJ performance can occur as a result of brain injury (Sinnett, Juncadella, Rafal, Azanon, & Soto-Faraco, 2007) or training (Donohue et al., 2010). This suggests that temporal perception is perhaps dependent on attentional mechanisms, and not purely a sensory-based process. Aside from studies showing *enhancements* on TOJs in video game players (e.g., Donohue et al., 2010), recent research has extended findings of better performance (lower JNDs) in the auditory modality to musical-conductors (Hodges et al., 2005) and in the visual modality to performing-musicians (Lim & Sinnett, 2011). Nevertheless, it is worth noting that research with musicians has not yet looked at crossmodal TOJs and the possibility that musical training, mostly auditory in nature, might have effects on visual TOJs.

Expert musicians were compared with non-musicians on a series of TOJ tasks that were presented under unimodal (visual or auditory), or crossmodal conditions. Given evidence from previous research, we expected to see lower JND scores for musicians when compared with controls in all conditions, but did not expect any differences in PSS scores.

## Experiment 1

### Participants

Twenty musicians (age =  $28 \pm 12$ ; 5 females) were recruited from the music department at the University of Hawaii at Manoa, local music studios, and through flyers. Musicians were required to have at least three years of formal training in music, and to have a regular practice schedule of at least six hours/week over the past six months. Control participants ( $n = 20$ ; age =  $22 \pm 5$ , 16 females. Note, pooled t-test comparisons showed no differences between males and females for any of the conditions, all  $p > .1$ ) were recruited from undergraduate courses, and had little or no training in music. All participants received either \$10 or course credit for their participation. Ethical approval was obtained from the University's Committee on Human Subjects.

### Stimuli and Apparatus

The basic TOJ task involves presenting participants with two stimuli separated by variable time intervals, referred to

as the stimulus onset asynchrony (SOA). The SOA length is manipulated to increasing or decreasing intervals that correspondingly makes the task easier or harder. A staircase approach was used in this experiment to adjust SOAs (see Stelmach & Herdman, 1991). The SOA started at 167ms and, for each successive trial, either decreased or increased (by 16.7ms) in a stepwise manner dependent on whether the participant answered the previous trial correctly. As the experiment progressed, each trial's SOA decreased making the order of occurrence difficult to determine. It can be inferred then, that as time progresses, changes in stepwise direction (up and down) will increase, reflecting increasing uncertainty in the participant. The task terminates once a total of twelve turning points have occurred.

Stimuli were presented on a 21" iMac using Bootcamp and DMDX software. Participants were seated at an eye to monitor distance of approximately 60cm. Prior to each trial a fixation-cross ( $0.5^\circ$ ) flanked by two square placeholders ( $1.4^\circ$ ) on the left and right was presented (see Figure 1). Stimuli for the visual task were horizontal and vertical lines ( $0.9^\circ$ ) and occurred centrally within the placeholders. For the auditory stimuli, processed samples of a dog and crow sound (350ms) were used (played at approximately 75db). In the crossmodal condition, the visual stimulus consisted of a black square ( $0.9^\circ$ ) within the placeholder, whereas the auditory stimulus was 50ms of white noise.

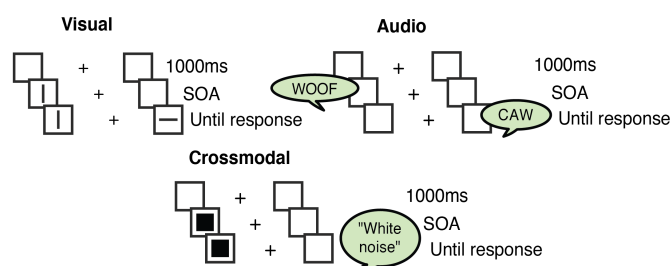


Figure 1. Stimuli for the three TOJ tasks in Experiment 1.

### Procedure

In all three conditions participants made unspeeded responses signaling which stimulus they believed had appeared first using one of two keyboard buttons. Onscreen instructions were presented first, followed by a short sequence of practice trials that included accuracy feedback. Presentation side (i.e., left or right) and stimuli order (e.g., horizontal or vertical line first) were randomized, as was the order of experimental conditions (e.g., audio, visual, crossmodal) for each participant.

### Results

Calculations of the JND and PSS were based on approaches used by previous studies (C. Spence, Baddeley, Zampini, James, & Shore, 2003; Stelmach & Herdman, 1991). Data from musicians and controls were pooled into separate groups. The average ratio of responses "horizontal line first" (e.g., for visual condition; for auditory condition ratio of crow sounds were used, etc.) was then plotted as a function



of the time in which the horizontal line preceded the vertical line. Data was then fit using a logistic function:

$$f(x, a, b) = \frac{1}{1 + \exp(-(x - a)/b)}$$

which was then used to obtain the JND and PSS estimates (similar to C. Spence et al., 2003). The PSS corresponds to parameter  $a$ , and is usually expected to fall at 0ms (or close) in unimodal conditions, as there is no reason to assume that a particular visual (or auditory) stimulus would be preferred over the other. It is more informative in the crossmodal condition as any shift would indicate whether auditory or visual events must precede the other for subjective simultaneity to be perceived. The JND relates to parameter  $b$ , which is adjusted to obtain the 75% JND as follow:

$$\text{JND}_{75} = \ln 3 \cdot b$$

Given that data was pooled within each group (musicians and controls), confidence intervals (95%) for each group's estimates and comparison p-values were calculated using a parametric bootstrap method with 999 replications (Efron & Tibshirani, 1993; for similar use of the bootstrap, see Azañón & Soto-Faraco, 2007).

**Auditory condition** Differences between musicians and controls were non-significant for PSS (2ms, CI = -8 to 12ms; vs. 9ms, CI = 1 to 16ms;  $p = 0.31$ ; respectively) and approaching significance for JND scores (43ms, CI = 34 to 53ms; and 56ms, CI = 45 to 68ms;  $p = 0.07$ ).

**Visual condition** The average PSS score for musicians' was significantly lower than controls by 10ms (-4ms, CI = -9 to 2ms; vs. -14ms, CI = -22 to 5ms;  $p = 0.037$ ; respectively), with negative PSS values indicating a possible bias in responses towards horizontal lines. The average JND score for musicians' was also significantly lower than controls by 18ms (29ms, CI = 23 to 35ms; vs. 47ms, CI = 37 to 56ms;  $p = 0.006$ ).

**Crossmodal condition** Differences between musicians and controls were non-significant for PSS (-43ms, CI = -60 to -25ms; and -63ms, CI = -93 to -30ms;  $p = 0.261$ ; respectively). It is worth noting that the negative PSS results indicate a bias in response towards the auditory modality (visual stimuli needed to be presented prior to auditory stimuli for simultaneity to be perceived). Musicians' average JND score was significantly lower than controls by 59ms (104ms, CI = 80 to 127ms; and 163ms, CI = 112 to 207ms;  $p = 0.021$ ).

## Discussion

There are two important findings that merit discussion. First, with the exception of the visual condition, no differences were observed for PSS between musicians and non-musicians. In the visual condition it is possible that there was a small bias towards horizontal lines for musicians, but note that performance hovered around zero as expected. The largest PSS differences were seen in the crossmodal condition. Specifically, visual stimuli had to precede auditory stimuli for both musicians and controls (by 43 and 63ms, see Figure 2) for them to be perceived as

occurring simultaneously (Hirsh & Sherrick Jr, 1961; Zampini et al., 2003).

Secondly, the temporal threshold for musicians was significantly lower in all conditions (visual, auditory, and crossmodal), although it should be noted that only marginal significance was observed in the auditory condition ( $p = 0.07$ ). Given that musical training is largely auditory in nature, a more robust difference in auditory JND scores was expected. It is possible that the "realistic" auditory stimuli used in our experiment may be more difficult than simpler tones, and therefore any effect might be somewhat masked. Furthermore, it might be possible that as the sounds were non-tonal, musicians may not have had a distinct advantage. Lastly, given that humans discriminate temporal events better in the auditory modality when compared to the visual modality, it is possible that performance was similar due to a ceiling effect.

Lastly, and also supported by Zampini et al. (2003), JND scores for the crossmodal condition increased nearly three-fold when compared to unimodal conditions, demonstrating that the task was more difficult.

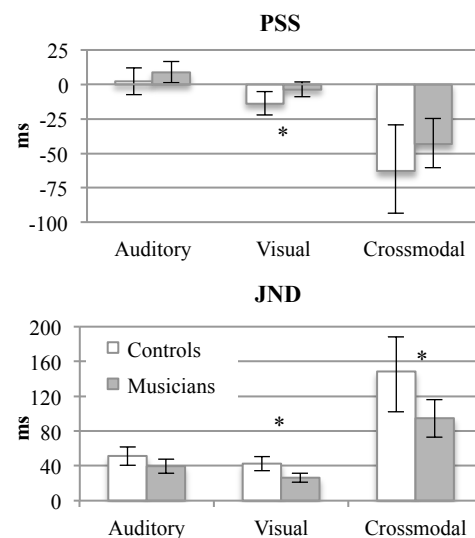


Figure 2. PSS and JND scores for Experiment 1. For PSS scores, positive indicates stimuli crow/horizontal/audio, and negative indicates dog/vertical/visual stimuli appearing first. Error bars indicate 95% CIs.

## Experiment 2

Spatial cues can also be incorporated into the TOJ tasks, allowing for a measure of how attention is oriented and captured. The presentation of exogenous cues prior to stimuli onset in a TOJ task creates a 'prior entry' effect, where attention is directed towards the cued side and subsequently affects performance on the task, regardless of whether or not the cue is predictive of location (see Shore, Spence, & Klein, 2001)

Exogenous orienting can occur from any stimulus that causes a reflexive or automatic capture of attention (e.g., bright flashes, loud sounds, etc.). By presenting an

exogenous cue in the TOJ task prior to the onset of stimuli, the cued side will be perceived as having occurred first. The PSS score then indicates how much in advance the uncued side must be presented before the cued for simultaneity to be perceived (see Shore et al., 2001). Thus, Experiment 2 included all of the unimodal conditions of Experiment 1 with the addition of within and crossmodal cues to determine whether spatial attention would differ between musicians and non-musicians. If musical training can improve spatial and temporal processing, it would be expected that musicians should have a smaller orienting effect, which would be manifested in lower PSS and JND scores than controls across all conditions. This would be indicative of improved temporal processing (JND) and less influence from peripheral distraction (smaller PSS). Furthermore, to our awareness this would be the first time multimodal cued TOJ tasks were conducted on musicians.

### Participants, Apparatus, and Procedure

The same participants from Experiment 1 also took part in Exp. 2 (all conditions from both experiments were interleaved and fully randomized). The discussion of the experiments is separated here for ease of understanding.

Stimuli and procedure were identical to those in Exp. 1, except for the addition of exogenous non-predictive cues in all conditions. In the visual condition, the cue was created by thickening the placeholder box of the respective side to a thickness of 4 pixels for 45ms. In the auditory condition, the cue was a laterally presented 500Hz sine wave lasting 45ms. The crossmodal condition consisted of two tasks: the first was an auditory TOJ task with visual cues, while the second was a visual TOJ task with auditory cues. All cues were randomly determined and had an equal chance of validly or invalidly cuing the target stimuli.

### Results

The JND and PSS scores were calculated using similar methods as in Exp. 1, by pooling musicians and control participants into two separate groups. For each of the four conditions, data from the two groups were fit to a weighted logistic function according to which stimulus was cued (e.g., horizontal/vertical bar, dog/crow sound, etc). The overall PSS value for each condition was computed as half the distance between each of the PSS values for the two curves. The average of the two JND values for each curve was used as the overall JND score. This approach essentially calculates the PSS for each type of stimulus cued, and averages the effect (see Shore et al., 2001). In order to gauge the influence of the cue, the two fitted curves were compared against one another. Logically, if the two curves were to map out on top of one another then the average would be 0 (PSS), as would be expected if the cue did not have any effect (assuming no bias for one stimulus type or the other). Thus, the larger the difference between the logistic fits for each cue, the larger the PSS, and by extension the greater effect that the cues had in general. This can similarly be applied to the calculation of JND, although

as the slope, the JND scores are expected to be similar for each stimulus type.

**Unimodal Cues Auditory condition:** The magnitude of the PSS shifts was not significantly different between musicians and controls (23ms, CI = 12 to 37ms; vs. 29ms, CI = 17 to 42ms;  $p = 0.259$ ; respectively, see Figure 3). Similarly, there were no differences in JND scores between the two groups (92ms, CI = 77 to 110ms; vs. 109ms, CI = 93 to 125ms;  $p = 0.106$ ). **Visual condition:** The magnitude of PSS shifts was significantly lower for musicians than controls by 29ms (30ms, CI = 10 to 46ms; vs. 59ms, CI = 48 to 71ms;  $p = 0.023$ ). On the other hand, JND scores for both groups were not significantly different (80ms, CI = 63 to 93ms; vs. 84ms, CI = 75 to 96ms;  $p = 0.29$ ).

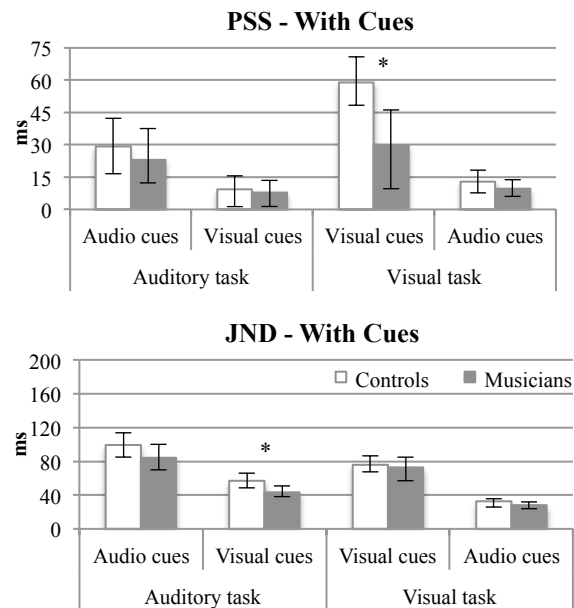


Figure 3. PSS and JND scores for Experiment 2. Asterisks indicate significant ( $p < .05$ ) between group differences. Error bars indicate 95% CIs.

**Crossmodal Cues Auditory TOJ with visual cues:** The magnitude of the PSS shifts did not significantly differ between musicians and controls (8ms, CI = 1 to 14ms; vs. 9ms, CI = 1 to 15ms;  $p = 0.39$ ; respectively). On the other hand, JND scores were significantly lower by 16ms for musicians compared to controls (47ms, CI = 42 to 56ms; vs. 63ms, CI = 53 to 73ms;  $p = 0.014$ ). **Visual TOJ with auditory cues:** The magnitude of the PSS shifts did not significantly differ between musicians and controls (10ms, CI = 6 to 14ms; vs. 13ms, CI = 8 to 18ms;  $p = 0.19$ ). Similarly, there were no differences in JND scores between the two groups (31ms, CI = 26 to 35ms; vs. 35ms, CI = 29 to 38ms;  $p = 0.088$ ).

**Cross experiment comparisons:** Further understanding of the cuing effects can be determined by comparing the results from the cued tasks in Experiment 2 to the no-cue unimodal tasks (auditory and visual) of Experiment 1. When doing so, JND differed for unimodal conditions but not for

crossmodal conditions. That is, the additional cues in Experiment 2 made the unimodal tasks harder for both musicians and non-musicians, as evidenced by longer temporal thresholds (JND) in both the auditory and visual modalities (all  $p < .01$ ). However, when the cues were presented in a separate modality (i.e., the crossmodal conditions of Exp. 2), JND scores were indistinguishable from the unimodal no cue conditions (Exp. 1) for both musicians and non-musicians (all  $p > .05$ ). Collectively, this may suggest that a difficult unimodal task can be made easier when presented as a crossmodal task (Sinnott et al., 2006; Sinnott et al., 2007; C Spence & Driver, 1996).

## Discussion

Robust findings from cross-experiment analyses broadly suggest that unimodal cues have detrimental effects on JND scores, whereas crossmodal cues do not. These results were similar for both musicians and controls. Excluding these cross experiment analyses, the only observed significant differences in Experiment 2 between musicians and controls were the lower PSS scores in the visual unimodal condition for musicians, and the lower JND score in the auditory-task/visual-cues condition for musicians. The lower PSS score indicates that musicians were captured less by the unimodal visual cues than non-musicians, while the lone JND difference seemingly suggests that crossmodal processing was easier for musicians, but only when judging temporal order for auditory targets that were cued visually.

## General Discussion

There are a number of important findings. To begin with, performance differences between musicians and controls were mixed in the auditory condition (musicians did have significantly lower JND scores in the auditory-task/visual-cues condition of Experiment 2, as well as marginally lower JNDs in the auditory condition in Experiment 1 [ $p = .07$ ], while the unimodal auditory condition of Experiment 2 was not significant). Thus, we do not see as strong a trend as Hodges et al. (2005), where auditory JND scores were significantly lower for musical conductors when compared to controls. This may be due to the use of different stimuli and experimental conditions. In the present experiment realistic sounds (dog and crow) were used, while auditory tones were used in Hodges et al.'s studies. Thus it is possible that pitch discrimination skills would not aid musicians in the auditory task used here. Furthermore, it is also possible that differences in auditory temporal processing may exist between conductors and performing musicians. It is worth noting however, that across all task types, JND scores for musicians were numerically lower than those for controls, although these differences were statistically significant in only four out of the seven conditions (Exp 1: auditory (marginal), visual, and crossmodal; Exp 2: Audio-task/visual cues).

Pertinent to the discussion is the tentative support for a supramodal account of attentional resources, supported by the fact that musicians outperformed controls on several

non-auditory related tasks, including smaller capture from visual cues, and lower JNDs for visual and crossmodal conditions (without cues). That is, it appears that musical training might have lead to improved visual processing. Having said that, as musical training involves much exposure to auditory stimuli, it was reasonable to expect enhancements in the auditory modality, although this was not consistently observed. Enhancements in the visual modality however, could be attributed to 1) better attentional resources, and/or 2) concomitant training in the visual modality from reading music while at the same time listening to and playing music, etc. Since we cannot rule out the second possibility however, these results can only be seen as tentative support for a supramodal account, pending further investigation with specific training conditions. Interestingly however, the robust findings of Experiment 2 where crossmodal PSS and JND scores were in fact lower than their unimodal counterparts (all  $p < .05$  and  $p < .001$ ; respectively), may provide stronger evidence for the exact opposite viewpoint: that is, a segregation of attentional systems (e.g., Sinnott et al., 2006; Wickens, 1984). Nevertheless, the current set of data makes it difficult to arrive at a decisive claim on either side of the debate, and may suggest a two-part attentional system that operates with both segregated and supramodal capacities. Indeed, it is likely that many previous findings supporting one theoretical account or the other may indeed be constrained by the varying methodologies used.

The segregated account is supported by the novel finding that was observed across both musician and control groups regarding the selective deficits in JND for only unimodal cues and not crossmodal cues. That is, when a within modality cue was added to the task, JND scores increased significantly for both musicians and control participants. However, when the cues were presented across modalities (i.e., a visual cue and an auditory TOJ task, or vice versa), performance was significantly better, and in fact did not differ from the no-cue conditions. This possibly suggests that the threshold of temporal detection may be robust to crossmodal distraction, while at the same time be vulnerable to distractions within the same modality.

As the between group differences in the auditory task in Experiment 1 were only marginally significant, this may suggest that auditory temporal acuity is less amenable to improvement through training (at least for the stimuli and task conditions used here), and that concomitant training effects are perhaps more robust in the visual domain. Importantly, the visual enhancements observed in JND in Experiment 1 lend support to the idea that attentional allocation, and therefore the improvement through training, may not be constrained within particular sensory modalities, but instead distributed to multiple modalities. Nevertheless, an important criticism of studies that used "trained" populations such as musicians and video game players, is the extent to which observed differences in experimental settings can actually be attributed to prior training. Boot et al. (2011), for instance, claimed that participants are often

aware of the purpose of the study as they are specifically recruited for their expertise, and that this awareness and potential motivational factor may very well influence performance. Unfortunately, our recruitment strategy for musicians did not allow us to keep them blind to the purpose of the study, and they may have been influenced by such knowledge. To this extent, our between group conclusions are largely speculative. Moreover, the nature of musical training in sighted individuals is in itself a multimodal experience, and further training studies would be better equipped to draw conclusions by controlling for the type of training each participant receives.

## References

- Azañón, E., & Soto-Faraco, S. (2007). Alleviating the crossed-hands deficit by seeing uncrossed rubber hands. *Experimental brain research*, 182(4), 537-548.
- Boot, W. R., Blakely, D. P., & Simons, D. J. (2011). Do Action Video Games Improve Perception and Cognition? *Frontiers in Psychology*, 2.
- Chen, K., & Yeh, S. (2009). Asymmetric cross-modal effects in time perception. *Acta psychologica*, 130(3), 225-234.
- Donohue, S., Woldorff, M., & Mitroff, S. (2010). Video game players show more precise multisensory temporal processing abilities. *Attention, Perception, & Psychophysics*, 72(4), 1120.
- Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap* (Vol. 57): Chapman & Hall/CRC.
- Farah, M., Wong, A., Monheit, M., & Morrow, L. (1989). Parietal lobe mechanisms of spatial attention: modality-specific or supramodal? *Neuropsychologia*, 27(4), 461-470.
- Gaser, C., & Schlaug, G. (2003). Brain structures differ between musicians and non-musicians. *Journal of Neuroscience*, 23(27), 9240.
- Granek, J., Gorbett, D., & Sergio, L. (2010). Extensive video-game experience alters cortical networks for complex visuomotor transformations. *Cortex*, 46(9).
- Green, C. S., Li, R., & Bavelier, D. (2010). Perceptual learning during action video game playing. *Topics in Cognitive Science*, 2(2), 202-216.
- Green, C. S., Pouget, A., & Bavelier, D. (2010). Improved Probabilistic Inference as a General Learning Mechanism with Action Video Games. *Current Biology*, 20(17), 1573-1579.
- Helmhold, N., Rammsayer, T., & Altenmüller, E. (2005). Differences in primary mental abilities between musicians and nonmusicians. *Journal of Individual Differences*, 26(2), 74-85.
- Hirsh, I., & Sherrick Jr, C. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology*, 62(5), 423-432.
- Hodges, D. A., Hairston, W. D., & Burdette, J. H. (2005). Aspects of multisensory perception: the integration of visual and auditory information in musical experiences. *Annals of the New York Academy of Sciences*, 1060, 175-185.
- Jones, M., & Yee, W. (1997). Sensitivity to time change: The role of context and skill. *Journal of Experimental Psychology: Human Perception and Performance*, 23(3), 693.
- Lim, A., & Sinnett, S. (2011). *Exploring Visual Attention in Musicians: Temporal, Spatial and Capacity Considerations*. Paper presented at the 33rd Annual Conference of the Cognitive Science Society.
- Münste, T., Altenmüller, E., & Jäncke, L. (2002). The musician's brain as a model of neuroplasticity. *Nature Reviews Neuroscience*, 3(6), 473-478.
- Patston, L., Hogg, S., & Tippett, L. (2007). Attention in musicians is more bilateral than in non-musicians. *Laterality*, 12(3), 262.
- Pavani, F., Husain, M., Ládavas, E., & Driver, J. (2004). Auditory Deficits in Visuospatial Neglect Patients. *Cortex*, 40(2), 347-365.
- Rauscher, F. (2003). Can Music Instruction Affect Children's Cognitive Development? *ERIC Digests* Retrieved 10 October, 2011, from <http://www.ericdigests.org/2004-3/cognitive.html>
- Röder, B., Teder-Sälejärvi, W., Sterr, A., Rösler, F., Hillyard, S., & Neville, H. (1999). Improved auditory spatial tuning in blind humans. *Nature*, 400(6740), 162-166.
- Shore, D. I., Spence, C., & Klein, R. M. (2001). Visual prior entry. *Psychological Science*, 12(3), 205-212.
- Sinnett, S., Costa, A., & Soto-Faraco, S. (2006). Manipulating inattention blindness within and across sensory modalities. *The Quarterly Journal of Experimental Psychology*, 59(8), 1425 - 1442.
- Sinnett, S., Juncadella, M., Rafal, R., Azanon, E., & Soto-Faraco, S. (2007). A dissociation between visual and auditory hemi-inattention: Evidence from temporal order judgements. *Neuropsychologia*, 45(3), 552-560.
- Spence, C., Baddeley, R., Zampini, M., James, R., & Shore, D. I. (2003). Multisensory temporal order judgments: When two locations are better than one. *Perception & psychophysics*, 65(2), 318.
- Spence, C., & Driver, J. (1996). Audiovisual links in endogenous covert spatial attention. *Journal of Experimental Psychology: Human Perception and Performance*, 22(4), 1005-1030.
- Stelmach, L. B., & Herdman, C. M. (1991). Directed attention and perception of temporal order. *Journal of Experimental Psychology: Human Perception and Performance*, 17(2), 539-550.
- Wickens, C. D. (1984). Processing Resources in Attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of attention* (pp. 63). New York: Academic Press.
- Zampini, M., Shore, D., & Spence, C. (2003). Audiovisual temporal order judgments. *Experimental brain research*, 152(2), 198-210.

# Reexamining Visual Orientation Anisotropies: A Bias Towards Simple Horizontal Stimuli on Temporal Order Judgments

Ahnate Lim (ahnate@hawaii.edu)

Department of Psychology, University of Hawaii at Manoa  
2530 Dole Street, Honolulu, HI 96822 USA

Scott Sinnett (ssinnett@hawaii.edu)

Department of Psychology, University of Hawaii at Manoa  
2530 Dole Street, Honolulu, HI 96822 USA

## Abstract

Although not currently a widely accepted notion, evidence suggests an anisotropy between horizontal and vertical orientations in visual processing. While there is evidence of an early neurological bias due to a greater number of cortical neurons tuned to the horizontal orientation, recent behavioral evidence suggests a “horizontal effect”, where performance for broadband horizontal stimuli is *worse* compared to vertical and oblique. Importantly however, this effect has only been observed for complex stimuli and is speculated to counterbalance for the greater occurrence of horizontal stimuli in the environment. In this experiment, we used a staircase temporal order judgment task in three spatial configurations (horizontal, vertical, and both) to test for 1) a bias towards either horizontal or vertical simple stimuli, and 2) whether performance would vary across different planes of stimuli presentation. A bias towards horizontal stimuli was observed, but only when presented in the horizontal plane. Theoretical implications are discussed.

**Keywords:** Horizontal bias; visual processing; anisotropy, stimulus orientation; temporal order judgment

## Introduction

Research has shown that in humans as well as many other species, visual (and tactile) stimuli are processed differently depending on their orientation. One of the more commonly observed biases in visual perception is a phenomenon termed the “oblique effect”, where stimuli presented in an oblique orientation are usually processed worse (i.e., in speeded detection, identification, resolution acuity and contrast sensitivity tests) than stimuli presented in the horizontal or vertical position (Appelle, 1972; Essock, 1980; for tactile, see Essock, Krebs, & Prather, 1997). Crucially, this phenomenon operates on at least two different levels. First, it has been linked causally to lower level vision, where it is attributed to differences in the number of cortical neurons in V1 tuned to stimulus orientation (Anzai, Bearse, Freeman, & Cai, 1995). Secondly, the effect also appears to be manifested in higher level cognitive processes such as memory, learning, and perception (for review, see Essock, 1980). The distinction between these two levels has in fact led to their classification as Class 1 and Class 2 oblique effects, respectively (Essock, 1980).

Apart from the well known oblique effect, studies that have attempted to compare anisotropies of horizontal and vertical orientations themselves against each other may not

have done so carefully (Hansen & Essock, 2004). It is worth noting that the current prevailing viewpoint is that horizontal and vertical stimuli are treated equally at the physiological level. Notwithstanding this dogma, there is considerable evidence suggesting that there is more neural circuitry in the visual cortex devoted to processing horizontal contours than vertical contours (Chapman & Bonhoeffer, 1998 (Figs. 1 and 2); Chapman, Stryker, & Bonhoeffer, 1996 (Figs. 1 and 2); Coppola, White, Fitzpatrick, & Purves, 1998; Mansfield, 1974; Mansfield & Ronner, 1978; Tiao & Blakemore, 1976). For instance, a study examining a large database of neurons in the cat’s striate cortex found that the largest population of cells are activated by orientations close to the horizontal position (Li, Peterson, & Freeman, 2003). Accordingly, it is curious that such a horizontal over vertical preference has not correspondingly been observed in behavioral tasks.

In fact, and despite the seeming neurological advantage for processing stimuli in the horizontal orientation, a study by Essock, DeFord, Hansen, and Sinai (2003) recently found diminished behavioral performance for horizontally presented stimuli (termed the “horizontal effect”). Furthermore, with complex “realistic” stimuli, they found that perceived orientation for broadband spatial content using horizontal, vertical, and oblique gratings was actually lowest for horizontal gratings, while oblique was instead seen best—a result seemingly contrary to the oblique effect, but solely at face value since the horizontal effect only appears to operate on complex stimuli. Interestingly, they explain these robust effects as being possibly due to a “whitening” mechanism that decreases the saliency of horizontal stimuli (which is argued to be most prevalent in natural scenes), thereby increasing the saliency of other broad-spectrum objects (such as predators, for instance).

Further research by Hansen and Essock (2004) replicated these findings in an experiment that used both simple and more complex “realistic” gratings. The classic “oblique effect” was seen with the simple gratings, whereas a “horizontal effect” (similar to Essock et al., 2003) was observed with the complex gratings. Additionally, the authors conducted an aggregate analysis of various natural scenes and found the prevalence of stimuli orientation in these scenes to be most prevalent in the horizontal orientation, then vertical, with the least prevalent being

oblique. They speculated that the horizontal effect may be a compensatory filter that at some level balances out for the greater abundance of such stimuli in the environment.

Hence, there appears to be evidence for both a bias towards, and a bias against horizontal oriented stimuli. That is, evidence at the physiological level suggests that the greater number of neurons tuned to horizontal orientations may lead to a bias in favor of horizontal detection. On the other hand, evidence also implicates the existence of a filter that may operate correctively against a bias towards greater occurrences of horizontal stimuli in the environment.

In light of these somewhat varying (but not mutually exclusive) viewpoints, several questions become relevant to the discussion. One is whether this “horizontal effect” is robust across all levels of perceptual processing. Recall that thus far the horizontal effect has only been observed with complex stimuli, therefore it is important to explore whether the same mechanism operates with simple stimuli, or if instead this mechanism only selectively operates in more complex “natural” scenes—as has been demonstrated in at least two studies (Essock et al., 2003; Hansen & Essock, 2004). Another issue is that, if this horizontal effect is somehow related to the prevalence of horizontal stimuli in natural scenes, might behavior change when presented with experimental layouts which contain more or less horizontal elements, and which are also holistically setup in a horizontal or vertical manner? Lastly, given that there appears to be a neurological bias towards detection of horizontally oriented stimuli as compared to vertically oriented (which to our awareness has never been demonstrated on behavioral measures), the question remains whether such a bias could in fact be detected at a behavioral level using simple stimuli? To better answer these questions, we designed an experiment consisting of a behavioral temporal detection task to test whether this bias exists with simple stimuli, using different experimental spatial layouts that contain varying elements of horizontal and vertical orientations.

The temporal order judgment task (TOJ) is an established psychophysical tool designed to assess the temporal processing of successively presented items. The task requires participants to determine the correct order of successively presented stimuli, and allows for two measures of perceptual processing to be calculated: the just noticeable difference (JND), and the point of subjective simultaneity (PSS). The former is a measure of the resolution or threshold of temporal discrimination, while the latter is the time in which one stimulus can be presented before the other such that they are still perceived as occurring simultaneously. Therefore, if a bias towards horizontal stimuli were to be observed, for instance, the PSS scores would indicate that the vertical stimuli must precede the horizontal (by a specific amount of time) for them to be perceived as occurring simultaneously.

It is worth noting that since humans are generally proficient at temporal discrimination (Hirsh & Sherrick Jr, 1961), the TOJ task is well suited for detecting small biases

in orientation processing. That is, using such a task would leave room for less error from extraneous variables such as task difficulty and interference from other cognitive processes that may come into play with other more complex stimuli and tasks.

For these reasons, the TOJ task appears to be particularly well suited for assessing threshold detection differences between horizontally and/or vertically presented stimuli. Subsequently, and to the best of our knowledge, this is the first time an adaptive step-function TOJ task has been used to investigate the anisotropy of stimulus orientation while also employing an adaptive staircase approach and the use of different experimental spatial configurations. The staircase approach will ensure that the majority of trials will occur at or close to threshold level.

In light of 1) existing neurological evidence for a horizontal bias, and 2) the lack of evidence for a countering “horizontal effect” for simple stimuli (the effect has only been observed for complex stimuli), we hypothesize that participants should be biased towards detecting horizontal stimuli better than vertical, although it is unclear whether the magnitude of this bias will be detectable here.

## Methods

### Participants

Participants ( $n = 33$ ; mean age =  $23 \pm 4$ ; 24 females) were recruited from undergraduate courses at the University of Hawaii at Manoa, and were offered course credit for their participation. All participants were naïve as to the purpose of the experiment and had normal or corrected to normal vision. Ethical approval was obtained from the University’s Committee on Human Subjects.

### Stimuli

Visual stimuli were presented on a 20”, Intel Core2Duo iMac using Bootcamp and DMDX software (Forster & Forster, 2003). Observers sat approximately 60 cm from the display. The targets in all tasks were vertical and horizontal lines that occurred within placeholder squares ( $2^\circ$  wide). These placeholders flanked a fixation cross in one of three different layouts (see Figure 1). These layouts corresponded to the three different tasks in this experiment.

### Procedure

Throughout each trial, the fixation cross and the two (in the horizontal and vertical layouts) or four (in the combined layouts) placeholders would remain on the display (see Figure 2). A target (either horizontal or vertical line, equiprobably) would appear in one of the place holders (also equiprobably) for a specified stimulus onset asynchrony (SOA) interval, followed by the other stimulus in the opposite place holder. The stimuli remained on the screen until participants then made an unspeeded forced choice response on the keyboard to indicate either “horizontal” or “vertical” first responses. An adaptation of Stelmach and Herdman’s (1991) step-function procedure was used to

determine the SOAs for each trial. Each trial began with an SOA of 167 ms. Depending on whether a correct or incorrect response was made, the SOA would respectively increase or decrease (by 16.7 ms) on the next trial. The experiment terminated after a total of 12 correct/incorrect reversals occurred.

In all three tasks, participants were first presented with onscreen instructions followed by a short sequence of practice trials, with accuracy feedback directly appearing after each trial. The experimenter monitored completion of the practice trials and ensured that participants understood the task requirements (repeating the practice session if necessary). Target presentation location (i.e., left, right, up or down) and stimuli order (e.g., horizontal or vertical line first) were randomized, as was the order of experimental tasks (i.e., horizontal, vertical, and combined) for each participant.

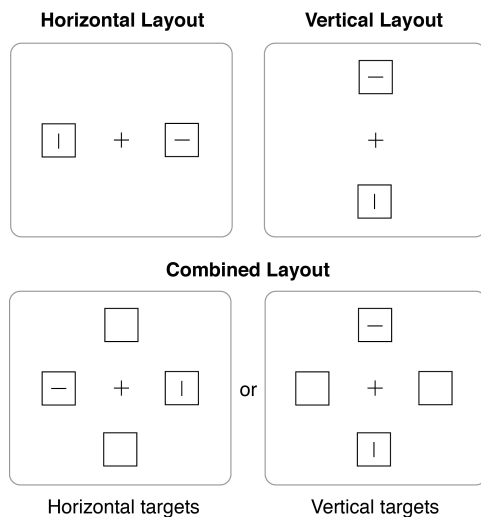


Figure 1. The three TOJ configurations. Each participant was tested on all three configurations. Note that on each trial in the combined layout, the task could occur on either the horizontal or vertical plane.

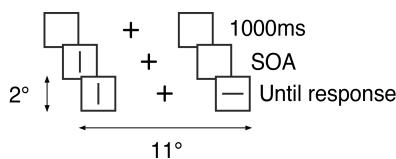


Figure 2. Example stimuli and time course for the TOJ task (horizontal layout displayed here; similar sequence occurred for the other layouts).

## Results

The results from the TOJ task can be analyzed to determine the point of subjective simultaneity (PSS). The PSS is the point in time in which one stimulus can be presented before the other such that they are still perceived as being

simultaneous. Note that this measure is usually expected to fall at 0 ms (or close to it, unless there is a bias in response). Additionally, the minimum amount of time that must separate two events such that they are still accurately perceived as occurring successively (and not simultaneously) can be measured. This is referred to as the just noticeable difference (JND) and is essentially a measure of the resolution or threshold of temporal discrimination. For this study we are more interested in the PSS than the JND, as the PSS can provide a measure of bias towards horizontal or vertical stimuli.

The calculation of both the PSS and JND was based on approaches used by previous research (for examples of other studies using similar methodologies and analyses, see Spence, Baddeley, Zampini, James, & Shore, 2003; Stelmach & Herdman, 1991). To begin with, the data from each of the three tasks were pooled together according to layout. The average ratio of responses "horizontal line first" was then plotted as a function of the time in which the horizontal line preceded the vertical line. For TOJ tasks, response rates typically follow a sigmoidal curve, from which data can be fit using the following logistic function:

$$f(x, a, b) = \frac{1}{1 + \exp(-(x - a)/b)}$$

where the response rate is mapped as a function of the SOA ( $x$ ), with two estimated parameters of central tendency ( $a$ ) and slope ( $b$ ; see C. Spence, et al., 2003).

Data was fit to this equation by minimizing the weighted sum of squares to obtain parameter estimates for  $a$  and  $b$ . The PSS, or SOA at which the participants considered the two stimuli to be simultaneous, corresponds to parameter  $a$ . The JND, or smallest interval between two stimuli giving a correct judgment probability 75% of the time, is directly related to parameter  $b$  (analogous to the slope of the central portion of logistic function). Here the relationship is that a steep slope will result in a smaller JND, and a shallow slope in a larger JND.

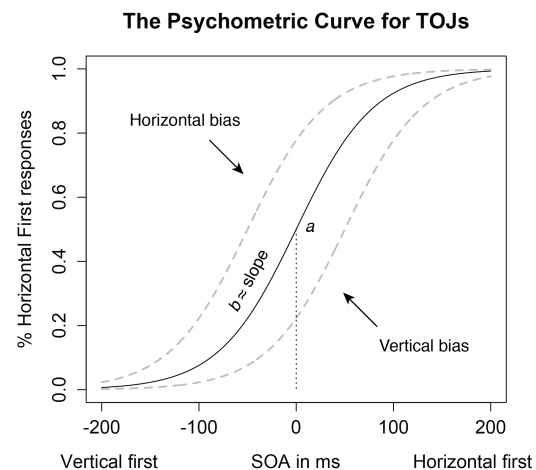


Figure 3. A typical TOJ response curve. Any bias in performance would be observed as a lateral shift of the curve, and correspondingly, the PSS score.



Confidence intervals (95%) for each group statistic were calculated using a parametric bootstrap method with 999 replications (for a similar bootstrap method employed in a TOJ study, see Azañón & Soto-Faraco, 2007; for an overview of the bootstrap, see Efron & Tibshirani, 1993). Given the unique nature of our dataset, we used a parametric bootstrap resampling approach for the statistical analyses due to particular benefits over more traditional means. That is, due to the varied number of trial observations and different response patterns resulting from the adaptive staircase paradigm, each individual's data points could vary significantly, and fitting the logistic function individually did not always converge or yield meaningful estimates. Thus, pooling data from all participants in each layout allowed for a better distribution of scores across all SOAs for the logistic fit from which we were able to extrapolate the overall PSS and JND values for each task using the above functions, and to subsequently estimate population parameters using bootstrapped confidence intervals<sup>1</sup>.

Furthermore, and in order to determine whether a PSS score, or bias towards a particular orientation was significant, we compared the results from each layout to a logistic function with identical parameters and characteristics, with the exception that the PSS was centered on 0 ms. This effectively allowed us evaluate the null hypothesis of whether the bias was significantly different from zero<sup>2</sup>.

### PSS scores

**Horizontal layout.** In the horizontal only configuration, there was a significant bias towards responding horizontal first ( $p < .05$ ). The magnitude of the PSS bias was 7ms, meaning that for horizontal and vertical lines to be perceived as occurring simultaneously, vertical lines had to precede horizontal lines by 7ms on average (CI = 1 to 12ms).

**Vertical layout.** In the vertical only configuration, the PSS was not significantly different than zero ( $p = 0.3$ ), with a 2ms bias towards horizontal first responses (CI = -3 to 8ms).

**Combined layout.** In the combined layout, when pooling the data across layouts, there was a significant PSS bias towards responding horizontal line first ( $p < .05$ ). The magnitude of this bias was 6ms (CI = 1 to 10 ms).

As the combined layout consisted of trials where the horizontal and vertical targets only occurred in either the horizontal or vertical plane (see Figure 1), we conducted a further analysis between these two sub-types to determine any differences in performance within the layout.

The PSS for the horizontal trials was significantly biased towards the horizontal stimulus ( $p < .05$ ). The magnitude of

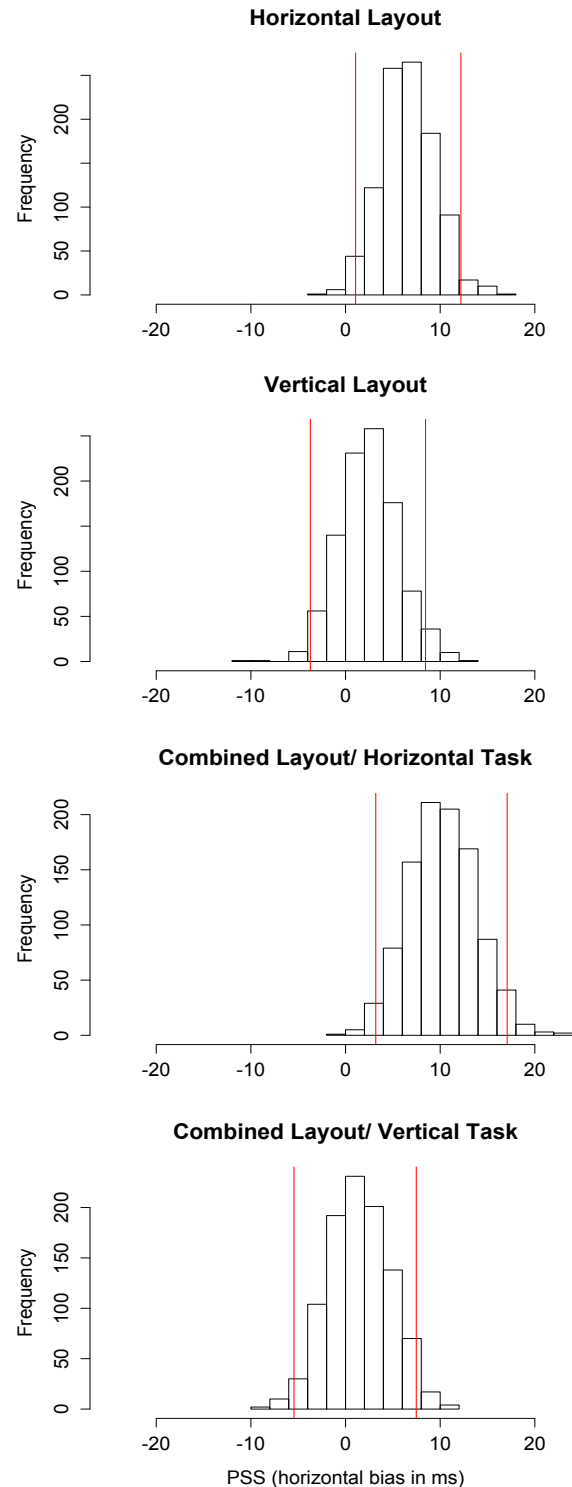


Figure 4. Parametric bootstrap resamples for each configuration (red lines denote 95% confidence intervals).

Positive PSS scores reflect a bias towards horizontal stimuli, whereas negative scores reflect a bias towards vertical. A zero PSS score would indicate lack of bias for either type of stimuli.

<sup>1</sup> Note that there is a growing consensus that certain exploratory techniques such as confidence intervals may be as useful as confirmatory ones (for a review, see Loftus & Masson, 1994). For details on bootstrap CIs, see Efron & Gong (1983).

<sup>2</sup> Given the existence of prior evidence for a bias towards horizontal orientations over vertical (see Introduction), a bootstrap comparison analogous to a one tailed  $t$ -test was used throughout.

this bias was 10ms (CI = 4 to 17 ms). In contrast, the PSS for the vertical trials not significantly different than zero ( $p = 0.4$ ) with a 1ms bias towards horizontal stimulus (CI = -4 to 7ms).

Given that both trials occurred within the same task configuration, we also ran a direct parametric bootstrap comparison test between the horizontal and vertical trials, and found that horizontal biases between the two trial types, as reflected by the magnitude of the PSS shifts (10ms vs 1ms) was significantly different ( $p < .05$ ).

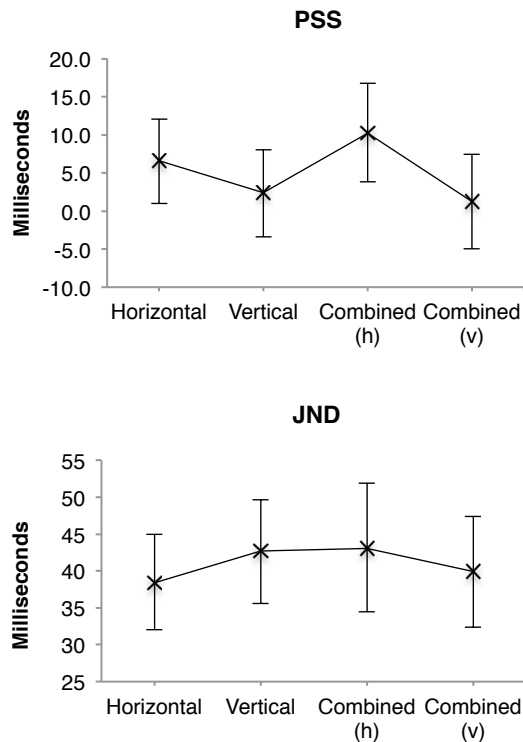


Figure 5. PSS (positive values indicate bias towards horizontal stimulus; negative towards vertical) and JND scores across the three tasks, with 95% confidence intervals.

### JND scores

Across all conditions, JND scores were 40ms on average, and as expected, scores did not significantly vary (all  $p > 0.1$ ). This confirms that the detection thresholds were similar in all configurations (see Figure 4), and that PSS differences across horizontal and vertical tasks were unlikely to be caused by extraneous variables such as the difficulty levels of the tasks.

### Discussion

Several novel findings were seen in this experiment. To our awareness, this is the first empirical investigation to examine performance among thousands of trials (the majority of them near threshold level SOAs) in a TOJ task to explore orientation bias between horizontal and vertical simple stimuli. We believe that this unique setup may have

allowed for the detection of orientation anisotropy between horizontal and vertical stimuli. Accordingly, an average bias of 9ms towards horizontal lines was observed during a horizontal task, suggesting that the vertical line must be presented on average 9ms before the horizontal line for simultaneity to be perceived. Thus, all things being equal, the horizontal orientation is preferred and appears to be processed more efficiently. Furthermore, this anisotropy was non-symmetrical in nature. That is, even in counterbalanced configurations with vertical placeholders, no similar bias towards vertical lines was observed (in fact, a very slight bias of 2ms on average towards horizontal lines was still observed in the vertical tasks, although this was not significant).

These findings are important for at least two reasons. First, in contrast to the prevailing view of equal treatment by the visual system for horizontal and vertical orientations, there is evidence that there may actually be an anisotropy between horizontal and vertical orientations. In fact, a bias towards the horizontal orientation has been observed at the neurological level in the visual cortex of several non-human animals (Chapman et al., 1996; Coppola et al., 1998; Li et al., 2003; Mansfield, 1974; Tiao & Blakemore, 1976) with visual systems expected to be ontogenetically analogous to humans. Thus our findings are the first to show a bias towards horizontally oriented simple stimulus, which speculatively may align with such findings of larger observed proportions of horizontally tuned cortical neurons. As stated by Essock et al. (2003), this result may not have been observed in the past due to the difficulty in obtaining large and unbiased samples. The use of the adaptive staircase design and parametric bootstrap analysis used here, however, offers a way of addressing this problem. Nevertheless, it is curious finding that we only observed this bias in the presence of a horizontal “plane”.

The fact that similar biases were only seen when performing the horizontal task (both in the horizontal and combined layouts) may suggest two possibilities: first, that the process of performing a horizontal task may facilitate the horizontal bias. The second possibility is that the process of performing a vertical task may inhibit the horizontal bias. Whether facilitation or inhibition (or both) is/are responsible for these results is beyond the scope of this paper. However, given the fact there we observed small (but non-significant) biases towards horizontal stimuli even for vertical tasks, we speculate that the inhibition argument may carry more weight.

Moreover, these results may dovetail with findings relating to the prevalence of horizons in natural scene layouts. Indeed, much of the world is sprawled out in a horizontal fashion due to the constraint of gravity, and it is thus conceivable for organisms to both have evolved visual systems that differentially process horizontal configurations for greater efficiency, and also to behaviorally adapt to such regularities in the natural world. Indeed, several examinations have been conducted on the statistical frequency of visual orientations in both naturalistic and

man-made environments, and have found greater occurrences of horizontally oriented stimuli (Baddeley & Hancock, 1991; Hansen & Essock, 2004; Keil & Cristóbal, 2000). To further add to this picture, learning may also play a vital role, as there is evidence of cross-cultural differences in visual anisotropies that can not be accounted for by mere exposure to a *carpentered* environment (Timney & Muir, 1976).

Interestingly, our findings can also be seen in a way to supplement the “horizontal effect”, which has been observed with complex broadband stimuli (Essock et al., 2003; Hansen & Essock, 2004). Specifically, in these studies the evidence for a “whitening” of (i.e., bias against) horizontally perceived orientations only occurred when complex broadband stimuli were used. From this study, we have confirmed that not only does the “horizontal effect” not apply to simple stimuli (lines), but also revealed the opposite: that in fact a bias towards simple horizontal stimuli can occur under conditions when a horizontal plane or task is present. Consequently, it is clear that comprehensive theoretical accounts of visual processing must ultimately reconcile and take into account these different phenomenological findings and the respective mechanisms responsible for such multi-level anisotropies.

## References

- Anzai, A., Bearse, M. A., Freeman, R. D., & Cai, D. (1995). Contrast coding by cells in the cat's striate cortex: monocular vs. binocular detection. *Visual Neuroscience*, 12, 77-93.
- Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological bulletin*, 78(4), 266.
- Azañón, E., & Soto-Faraco, S. (2007). Alleviating the crossed-hands deficit by seeing uncrossed rubber hands. *Experimental brain research*, 182(4), 537-548.
- Baddeley, R. J., & Hancock, P. J. B. (1991). A statistical analysis of natural images matches psychophysically derived orientation tuning curves. *Proceedings: Biological Sciences*, 219-223.
- Chapman, B., & Bonhoeffer, T. (1998). Overrepresentation of horizontal and vertical orientation preferences in developing ferret area 17. *Neurobiology*, 95(5), 2609.
- Chapman, B., Stryker, M. P., & Bonhoeffer, T. (1996). Development of orientation preference maps in ferret primary visual cortex. *The Journal of Neuroscience*, 16(20), 6443-6453.
- Coppola, D. M., White, L. E., Fitzpatrick, D., & Purves, D. (1998). Unequal representation of cardinal and oblique contours in ferret visual cortex. *Proceedings of the National Academy of Sciences*, 95(5), 2621.
- Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, 36-48.
- Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap* (Vol. 57): Chapman & Hall/CRC.
- Essock, E. A. (1980). The oblique effect of stimulus identification considered with respect to two classes of oblique effects. *Perception*, 9(1), 37-46.
- Essock, E. A., DeFord, J. K., Hansen, B. C., & Sinai, M. J. (2003). Oblique stimuli are seen best (not worst!) in naturalistic broad-band stimuli: a horizontal effect. *Vision research*, 43(12), 1329-1335.
- Essock, E. A., Krebs, W. K., & Prather, J. R. (1997). Superior sensitivity for tactile stimuli oriented proximally-distally on the finger: Implications for mixed class 1 and class 2 anisotropies. *Journal of Experimental Psychology: Human Perception and Performance*, 23(2), 515.
- Forster, K., & Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35(1), 116.
- Hansen, B. C., & Essock, E. A. (2004). A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. *Journal of Vision*, 4(12).
- Hirsh, I., & Sherrick Jr, C. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology*, 62(5), 423-432.
- Keil, M. S., & Cristóbal, G. (2000). Separating the chaff from the wheat: Possible origins of the oblique effect. *Journal of the Optical Society of America A*, 17(4), 697-710.
- Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique effect: a neural basis in the visual cortex. *Journal of Neurophysiology*, 90(1), 204-217.
- Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, 1(4), 476-490.
- Mansfield, R. (1974). Neural basis of orientation perception in primate vision. *Science*, 186, 1133.
- Mansfield, R., & Ronner, S. (1978). Orientation anisotropy in monkey visual cortex. *Brain Research*, 149, 229-234.
- Spence, C., Baddeley, R., Zampini, M., James, R., & Shore, D. I. (2003). Multisensory temporal order judgments: When two locations are better than one. *Perception & psychophysics*, 65(2), 318.
- Stelmach, L. B., & Herdman, C. M. (1991). Directed attention and perception of temporal order. *Journal of Experimental Psychology: Human Perception and Performance*, 17(2), 539-550.
- Tiao, Y., & Blakemore, C. (1976). Functional organization in the visual cortex of the golden hamster. *The Journal of comparative neurology*, 168(4), 459.
- Timney, B., & Muir, D. (1976). Orientation anisotropy: incidence and magnitude in Caucasian and Chinese subjects. *Science*, 193(4254), 699.

# A Bayesian Model of Rule Induction in Raven's Progressive Matrices

**Daniel R. Little** ([daniel.little@unimelb.edu.au](mailto:daniel.little@unimelb.edu.au))

School of Psychological Sciences, The University of Melbourne  
Parkville VIC 3010 Australia

**Stephan Lewandowsky** ([stephan.lewandowsky@uwa.edu.au](mailto:stephan.lewandowsky@uwa.edu.au))

School of Psychology, The University of Western Australia  
Crawley WA 6009

**Thomas L. Griffiths** ([tom\\_griffiths@berkeley.edu](mailto:tom_griffiths@berkeley.edu))

Department of Psychology, University of California, Berkeley  
Berkeley CA 94720-1650 USA

## Abstract

Raven's Progressive Matrices (Raven, Raven, & Court, 1998) is one of the most prevalent assays of fluid intelligence; however, most theoretical accounts of Raven's focus on producing models which can generate the correct answer but do not fit human performance data. We provide a computational-level theory which interprets rule induction in Raven's as Bayesian inference. The model computes the posterior probability of each rule in the set of possible rule hypotheses based on whether those rules could have generated the features of the objects in the matrix and the prior probability of each rule. Based on fits to both correct and incorrect response options across both the Standard and Advanced Progressive Matrices, we propose several novel mechanisms that may drive responding to Raven's items.

**Keywords:** Rule induction, Bayesian inference, Raven's Progressive Matrices

## Introduction

Raven's Progressive Matrices (Raven et al., 1998; Raven's from here on) is one of the most widely used assays of fluid intelligence, and much attention has focused on the underlying elemental cognitive processes. Raven's has arguably gathered more attention in the cognitive literature than any other psychometric measure of fluid intelligence, largely because it is an induction task *par excellence* that can be modeled computationally (see e.g., Carpenter, Just, & Shell, 1990; Verguts, De Boeck, & Maris, 2000). For example, Carpenter et al. (1990) presented a production-system model of Raven's to support a two-factor theory of Raven's with working memory capacity (WMC) as the first factor and a second factor related to the ability to abstract relations. This latter ability has been associated with several attributes including rule generation speed (Verguts & De Boeck, 2002), inference speed (Rasmussen & Eliasmith, 2011), and analogical comparison (Lovett, Forbus, & Usher, 2010; McGreggor, Kunda, & Goel, 2010).

These extant models of Raven's have focused on cognitive processes and mechanisms that underlie the inference of rules from the objects in the matrix. Further insight can be gained by exploring a computational-level analysis (Marr, 1982). As performance in Raven's relies primarily on rule induction, the task is conducive to instantiation within a Bayesian framework. For instance, Bayesian models of rule induction have

been successfully applied to similar tasks, such as numerical sequence prediction (i.e., which number follows in the sequence: 1, 2, 3, 5, 7, 11?; Austerweil & Griffiths, 2011) and rule-based categorization (Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Examining Raven's within the context of a Bayesian model allows exploration of questions about what people's priors (or in non-Bayesian terms, inductive biases) might be like for rules of the variety used in the Raven's test. Finally, the Bayesian formalism provides an extensible framework for using standard extensions to Bayesian models to capture other, more process-based interpretations of factors known to be relevant to performance on Raven's, such as memory and learning.

Here we present a Bayesian model of Raven's which interprets rule induction as Bayesian inference in which a set of rules with some prior probability are evaluated based on their ability to have plausibly generated the features of the items shown in the matrix. Rules are then sampled based on their posterior probability and Bayesian model averaging is used to predict which answers are most likely given the posterior distribution. Unlike extant models, which examine how successful the model is at predicting correct responses (e.g., Carpenter et al., 1990; Lovett et al., 2010; McGreggor et al., 2010), our model also makes predictions about the proportion of responses involving the various incorrect options.

## Bayesian Model of Raven's

Solving a Raven's problem can be conceptualized as a three-stage process involving feature extraction, rule-inference and prediction.<sup>1</sup> As illustrated in Figure 1, Raven's items have the following composition:

$$\begin{array}{ccc} O_{11} & O_{12} & O_{13} \\ O_{21} & O_{22} & O_{23} \\ O_{31} & O_{32} & ? \end{array} \quad (1)$$

where  $O_{ij}$  is the object in the  $i^{th}$  row and  $j^{th}$  column. Assuming the features of each object are extracted successfully,

<sup>1</sup>In the present model, we follow Carpenter et al. (1990) by hand-coding the features of the items. Several methods for extracting the features of Raven's items have been proposed (Lovett et al., 2010; McGreggor et al., 2010; Rasmussen & Eliasmith, 2011).

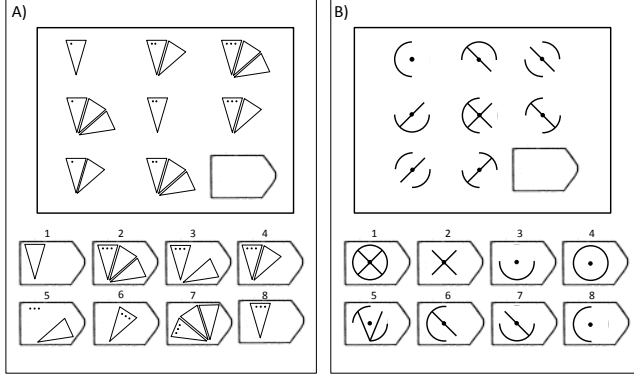


Figure 1: Two examples of matrices like those in the Raven's test. A: Example of an item containing a pairwise incremental rule, a constant rule and a permutation rule. B: Example of an item containing a constant rule and an XOR rule.

such that each object can be decomposed into  $N$  features,  $O_{ij} = \{f_1^{ij}, \dots, f_N^{ij}\}$ , then the goal is to infer the rule that generated the features of the last object in each row and column from the features of the first two objects in each row and column. By design, the rules that are applied to generate  $O_{13}$  from  $O_{11}$  and  $O_{12}$  are the same as the rules used to generate  $O_{23}$  from  $O_{21}$  and  $O_{22}$ . We refer to the set of rules that apply to each row  $G$ , where  $G = \{g_1, \dots, g_M\}$  is a collection of  $M$  rules for each feature. The third object in a row is assumed to have been generated by applying these rules to the features of the first two objects:  $O_{13} = G(O_{11}, O_{12})$  and  $O_{23} = G(O_{21}, O_{22})$ . We assume a separate set of rules,  $H$ , may apply to each of the features of the objects within a column,  $O_{31} = H(O_{11}, O_{21})$  and  $O_{32} = H(O_{12}, O_{22})$ . The column rules and the row rules may be different; however, because  $O_{33}$  can be predicted using either the rows or the columns, we restrict the following analysis to the row rules,  $G$ , but it also applies comparably to the column rules,  $H$ .

We assume that inferring a rule which generated a feature of the third object in a row or column can be conceptualized as finding the posterior probability of each possible rule applied to that feature:

$$p(g|f) = \frac{p(f|g)p(g)}{\sum_{i=1}^M p(f|g_i)p(g_i)} \quad (2)$$

where  $p(f|g)$  is the likelihood of generating feature,  $f$ , given the rule,  $g$ .

### Likelihood

In Raven's, all of the rules apply to individual features (which may be discrete or continuous valued; e.g., the three dots in Figure 1, panel A); hence, within a row, the likelihood will have a value of 0 or 1 depending on whether or not the rule successfully produces the features of the third object from the features of the first two objects in that row. To allow for miscalculations in the evaluation of a rule, we set the likelihood

equal to  $\epsilon$  whenever the rule could not have generated the features of  $O_{13}$  and  $(1 - \epsilon)$  whenever a rule could have generated the features of  $O_{13}$ , where  $\epsilon$  is a small number ( $\epsilon = .01$  in our simulations below). We make the further assumption that a rule may work within neither, only one, or both of the rows (or columns), in which case the probability of generating a feature given a rule across both rows is the product of the probabilities from each row separately:

$$p(f|g) = \begin{cases} (1 - \epsilon)^2 & \text{if the rule works for both rows} \\ \epsilon(1 - \epsilon) & \text{if the rule works for only one row} \\ \epsilon^2 & \text{if the rule works for neither row} \end{cases} \quad (3)$$

### Priors over rules

Carpenter et al.'s (1990) analysis of Raven's identified a taxonomy of rules used to create the Raven's problems. These rule types can be classified as involving transformations (e.g., a quantitative pairwise increment or decrement of a feature from one object in the matrix to the next, or a permutation of objects within a row or column), rules requiring logical operations (e.g., AND conjunctions, OR disjunctions and exclusive-or, XOR, relations between features; Matzen et al., 2010) and a constant rule in which features are maintained unchanged across items.<sup>2</sup> To provide a concrete example, Figure 1 presents two sample Raven's-like problems. The matrix in panel A contains a pairwise incremental rule (i.e., the dots increase across items from left to right) and a permutation rule (i.e., objects with 1, 2 and 3 triangles are permuted across rows and columns). The matrix in panel B contains a constant rule (i.e., the center dot appears in all items) and an XOR rule (i.e., features which appear in the first two objects do not appear in the third object and features which appear only in one of the first two objects also appear in the third object). Participants must infer these rules from the objects in the matrix and select the missing lower right object in each matrix from the set of possible response options below each matrix.

In total we used eight different rules derived from the taxonomy presented in Carpenter et al. (1990; see also, Matzen et al., 2010) and further analyses: 1) constant, 2) increment or 3) decrement, 4) permutation, 5) logical AND (i.e., maintain common features and delete unique features between objects), 6) logical OR (i.e., maintain unique features between objects), 7) logical XOR (see Figure 1, panel B), and 8) a Distribution of 2 rule.<sup>3</sup>

<sup>2</sup>Carpenter et al. (1990) refers to permutation rules as Distribution of 3 rules because the feature values appear once in the three objects within a row or column and to XOR rules as Distribution of 2 rules because the feature only appears in two of the three objects. Carpenter also refers to logical OR rules as addition and logical AND rules as subtraction.

<sup>3</sup>Six items that were generated using idiosyncratic rules were removed from the 72 Raven's Standard Progressive Matrices (RSPM) and Advanced Progressive Matrices (RAPM) items that we tested. We included the Distribution of 2 rule in our set because there are two items in the RAPM set which use a Distribution of 2 rule that is inconsistent with an XOR rule.

We tested three prior distributions on rules. First, we assumed that each rule had an equal prior probability (i.e., a *uniform* prior probability). Vodegel Matzen, van der Molen, and Dudink (1994) conducted a study using single rule matrices and found there was a clear order of rule difficulty, in which the easiest was constant in a row, followed by quantitative pairwise progression, permutation and logical rule operations, which were the most difficult. To capture this order of difficulty we developed a second prior which assumed that the probability of a rule was proportional to the ease with which that rule could be generated (e.g., the complexity of the rules, which is related to the mental effort necessary to infer and use a rule). For this prior (hereafter referred to as the *Carpenter* prior), we used the frequency with which each rule occurred in Carpenter et al.’s (1990) analysis as a proxy for the ease with which each rule could be generated and set the prior probabilities to be proportional to the presentation frequency. Finally, we assumed that the prior may be related to the accuracy with which items containing those rules could be solved. Again, the probabilities are also related to the ease or complexity with which a rule can be generated, but for this prior, we use the relationship between each rule and accuracy on items generated using that rule as a proxy for complexity. To compute this *accuracy-based* prior, we fit a logistic regression model using the rule profile of each item as the predictor variables (i.e., for each item,  $i$ , and for each rule,  $j$ , we set an indicator equal to 1 if item  $i$  was generated using rule  $j$  and to 0, otherwise) and the proportion correct for each item as the dependent variable. We then transformed the resulting exponentiated regression weights such that they ranged between 0 and 1 and summed to 1 across all of the rules. The actual probabilities for each of these priors are listed in Table 1.

Table 1: Prior probabilities for each rule.

Rule	Uniform	Carpenter	Accuracy-based
Constant	.125	.194	.150
Increment	.125	.223	.185
Decrement	.125	.223	.141
Permutation	.125	.058	.116
Logical AND	.125	.039	.167
Logical OR	.125	.058	.119
Logical XOR	.125	.165	.119
Distribution of 2	.125	.039	.081

### Predicting the response from the posterior

The perceptual complexity of the objects affects how easily rule inferences can be generated (Primi, 2001). For example, Meo, Roberts, and Marucci (2007) showed that performance was significantly worse when features within items were difficult to identify. We incorporate this finding into the current model by assuming that the response is based on the similarity between response options and objects predicted from the rules in the posterior, and for items which have features which are difficult to extract, the similarity does not need to be very high in order for the response options to match the object in the posterior.

Once the posterior probability of each rule is computed for each feature using Equation 2, we compute the missing objects as follows: For a given row rule,  $G_m$ , we predict the features of  $O_{33}$  by applying the rule to the features  $O_{31}$  and  $O_{32}$ . That is,  $\hat{O}_{33} = G_m(O_{31}, O_{32})$ . Response proportions are determined by computing the relative similarity of each response option,  $R_k$  to each object in the predictive posterior by  $s_{\{\hat{O}_{33}, R_k\}} = \exp(-c \times d_{\{\hat{O}_{33}, R_k\}})$ , where  $d_{\{\hat{O}_{33}, R_k\}}$  is the Euclidean distance between  $R_k$  and  $\hat{O}_{33}$  (i.e., the square root of the summed squared differences between the features of  $\hat{O}_{33}$  and  $R_k$ ) and  $c$  is a parameter which determines the steepness of the similarity gradient.

Similarities are weighted by their probability in the predictive posterior and normalized across response options to determine the probability that the model chooses each response option (see e.g., Rasmussen & Eliasmith, 2011). In our baseline model, we set  $c$  equal to 10 which results in strong responding to the response options which are represented most strongly in the posterior because the similarity gradient is quite steep when  $c \gg 1$ ; however, preliminary examination of the model fits to some items suggested that human responses were influenced by the similarity of some distractors to the correct response. For these items, we set  $c$  equal to 1, which implies a shallow similarity gradient and greater confusability between similar response options; we exhaustively tested each item to determine whether lowering  $c$  improved the fit for that item. We refer to this version of the model as the *Baseline + similarity model*. In this model, 20 of the 66 items had a lower  $c$  value than the other items (i.e., for these items,  $c = 1$ ).

Initial inspection of the model predictions additionally revealed a propensity for subjects not to choose response options which also appear as items in the matrix. To handle this, we introduced a heuristic into the model such that all objects that appeared in the matrix were removed from the posterior predictive distribution and from the response set before computing the response proportions. Through exhaustively testing each item, we determined that this heuristic improved the model’s predictions for 52 of the 66 items. We refer to this version of the model as the *Baseline + heuristic model*. We additionally tested a *full* model which incorporated both similarity-based responding and the response heuristic.

Table 2: Chi-square values for the fit to choice probabilities for Raven’s Standard Progressive Matrices and Advanced Progressive Matrices. The model that provides the best fit to the data for each prior is shown in bold.

Model	Uniform	Carpenter	Accuracy-based
Baseline Model	43603	37363	40699
Baseline + Heuristic	32642	32816	32871
Baseline + Similarity	23385	13884	19948
Full Model	<b>11761</b>	<b>12258</b>	<b>9010</b>

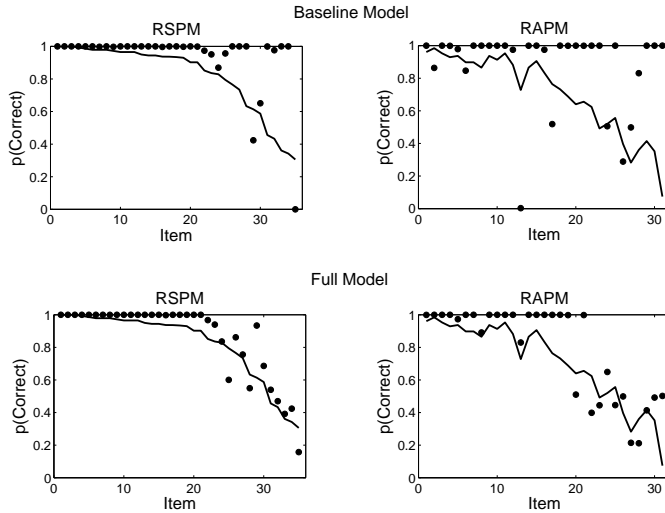


Figure 2: Baseline model and full model fits to proportion correct data from RSPM and RAPM (Little, Lewandowsky & Craig, 2012). Model predictions are shown by the black dots, the observed data are shown by the solid black line.

### Comparing the model to human performance

Descriptive statistics and accuracy data for the Raven's data were previously reported in Little, Lewandowsky, and Craig (2012); here, we are primarily concerned with how often each response option was chosen for each item. (Note that we have removed omissions from this data and look at the distribution of responses across the response options only for participants who actually gave a response). We fit the four versions of the model using the three different prior distributions to the response proportions from the RSPM and RAPM. Chi-square fit statistics are shown in Table 2. Based on these fit values, it is evident that adding similarity-based responding improves the fit over adding the response heuristic to the baseline model. Adding both modifications results in a substantial improvement when a uniform or an accuracy-based prior is used and a marginal improvement when the prior based on Carpenter's analysis is used. The overall best fit is found when the full model is used with an accuracy-based prior; consequently, we now focus on this model's predictions.

Figure 2 shows the accuracy predictions for the baseline and full models for the RSPM items (reordered according to accuracy rate observed in Little et al., 2012) and the RAPM items. The baseline model clearly predicts the correct answer for most of the Raven's items; however, this model overpredicts the propensity with which people choose the correct item for both RSPM and RAPM. By contrast, for the RSPM items, the full model accurately predicts the decrease in accuracy across the items. For the RAPM items, the full model predicts the decrease in accuracy for the hardest Raven's items, but still overpredicts the proportion correct for items in the middle difficulty range.

Figure 3 shows the (log transformed) predictions of the

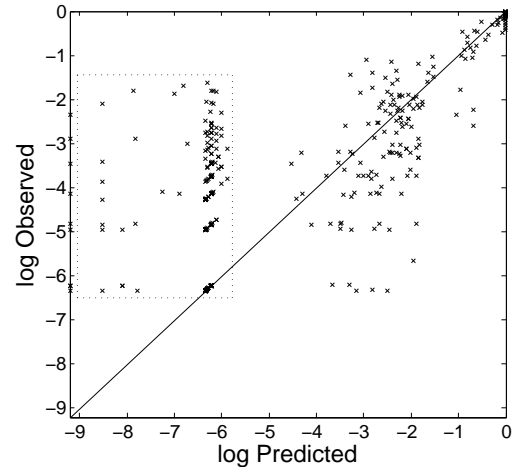


Figure 3: Predicted (full model) and observed response proportions for all response options from RSPM and RAPM. The dotted line surrounds options for which the model predicts near zero response proportions.

full model (with an accuracy-based prior) across all of the response options against the observed response proportions.<sup>4</sup> For a large proportion of response options, the model predicts the observed proportions correctly; however, the model also erroneously predicts a large number of response options near 0 (predicted log proportion less than -6). These items are also the items for which the model overpredicts the correct response in Figure 2. Examination of the response profiles for individual items reveals that for many items this overprediction is not detrimental to the qualitative pattern of results (see Figure 4). One possible explanation for the discrepant results is that people are guessing the answer with some small probability which would reduce the accuracy for some of the items and increase the proportion of false alarms to some of the distractors.

### Discussion

This paper has defined a Bayesian model of Raven's Progressive Matrices that provides an account of Raven's based on the idea that people infer rules by computing the posterior probability of those rules and using the rules to generate plausible responses. We considered three priors and two ways in which the model could be modified to accommodate human performance. Ultimately, a model incorporating an accuracy-based prior and both modifications provided the best fit to the data.

The success of the accuracy-based prior suggests that rules vary in how they contribute to accurate performance in Raven's. This relationship may reflect sensitivity to differing levels of complexity between the rules, and one way to handle this prior in a more principled way is to instantiate the rules

<sup>4</sup>Proportions equal to 0 or 1 were corrected by setting these proportions equal to  $1/(4N)$  or  $[N - (1/(4N))]/N$ , respectively.



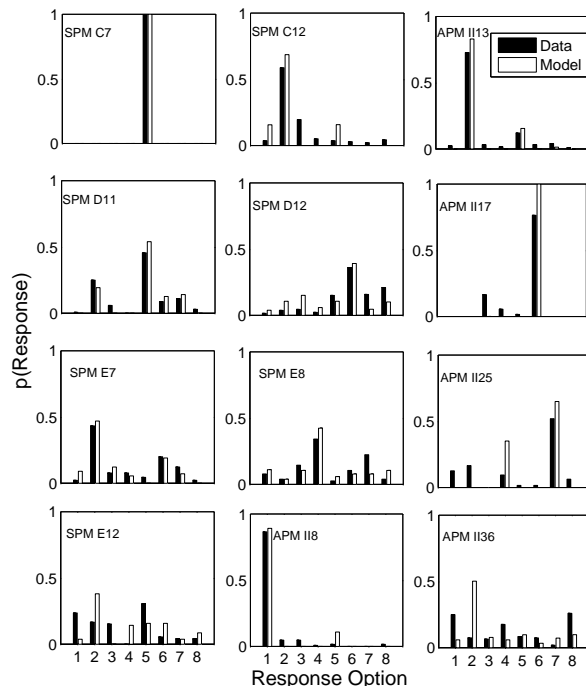


Figure 4: Model prediction profiles for a selection of items. Items SPM C12 (i.e. Standard Progressive Matrices item 12 from Set C) and APM II13 illustrate cases in which the model predicts response probabilities of 0 for responses that people occasionally select. Items SPM E12 and APM II25 illustrate two examples in which the model makes incorrect predictions. For item APM II36, the item with the highest error rate from either test, the model predicts that the correct response (option 2) should be selected most frequently, but humans prefer options 1 and 8.

using a common language with a formal definition of complexity, such as first-order logic (cf. Goodman et al., 2008).

The better fit produced by the similarity-based prediction modification suggests that people vary in how responses are generated to different Raven's items. For instance, for some items, responses are generated by comparing each response option to the possibilities in the predictive posterior; for these items, small differences in features of the response options do not result in a large difference in response prediction. For other items, the response must match the objects in the predictive posterior exactly. An aim for future research would be to identify what makes a feature hard to identify. This would allow appropriate *a priori* specification of the similarity-based prediction rather than the post-hoc approach adopted here. Finally, the heuristic mechanism suggests that people limit their responding to only the plausible response alternatives rejecting alternatives which are implausible because they duplicate items which appear as objects in the matrix.

One of the advantages of formulating a Bayesian model of this task is that we can make use of recent work that has explored how Bayesian models can be extended to cap-

ture different aspects of human cognition. One important aspect of performance on Raven's is that differences in WMC correlate highly with accuracy. Furthermore, the correlation with WMC increases as the items become more difficult (if the overall correlation between Raven's and WMC is large enough; Little et al., 2012). The model does not currently incorporate WMC; however, one possibility for extending the model is to represent the hypothesis space as a sampling distribution from the prior using importance sampling.

In an importance sampling scheme (Shi, Griffiths, Feldman, & Sanborn, 2010) samples from the prior are weighted by their likelihood to approximate the posterior distribution; more samples lead to a better approximation of the posterior. Differences in WMC could be modelled by varying the number of samples with high WMC participants having more samples from the prior. With more samples, the model is more likely to generate the correct answer. This idea is reminiscent of the difference between Carpenter et al.'s (1990) FAIRRAVEN and BETTERRAVEN models. The BETTERRAVEN model was given access to more rules than the FAIRRAVEN model and consequently was able to mimic the performance of participants with highly accurate Raven's performance and, by implication, higher WMC.

An alternative account of the relationship between WMC and Raven's is that higher WMC permits faster learning of what rules are likely to be necessary (Verguts & De Boeck, 2002). In support of this account, Carlstedt, Gustafsson, and Ullstadius (2000) found that WMC was correlated more strongly with homogenous intelligence test items (which all required the same rule to solve) than with heterogenous intelligence test items. Presumably, learning the relevant rule is easier for homogenous test items than for heterogenous items. In other related tasks, such as rule-based categorization, WMC is known to be correlated with learning rate (Lewandowsky, 2011; Sewell & Lewandowsky, in press). By this account, the rules at the end of the test are more diagnostic because they have had more time to be learned, thereby leading to greater divergence between low and high ability individuals. Learning in the Bayesian model of Raven's could be instantiated by using a special case of importance sampling known as particle filter sampling (Doucet, Freitas, & Gordon, 2001; Sanborn, Griffiths, & Navarro, 2010). In a particle filter model, a set of particles representing possible rules are drawn in proportion to their prior probabilities. As one progresses through the Raven's items, probabilities are updated in proportion to the success of each rule. Particles representing rules are maintained if they work, but are replaced with new samples from the prior if they do not. Again, higher WMC could be modelled using a larger number of samples. A particle filter model of Raven's would consider both Carpenter et al.'s (1990) and Verguts and De Boeck's (2002) accounts to be correct. That is, higher WMC allows for access to more rules by virtue of allowing more samples from the prior; higher WMC also allows for faster learning by allowing a larger number of particles to be updated from trial to trial.

Consequently, a particle filter model of WMC and Raven's provides a synthesis of these two approaches.

A limitation of Carpenter et al. (1990) and of our own work is that the inputs to the model are hand-coded (Lovett et al., 2010). Hand-coding ignores potentially important spatial representations between objects. Furthermore, Carpenter et al. (1990) did not model the process of rule discovery, but instead fixed the set of rules that were available to the model. The second criticism is less problematic because the rule set is comprehensive, covering the set of rules necessary to handle most of the items; rule discovery is couched in terms of updating the prior probability of each of the rules based on how well those rules work to explain the observed features.<sup>5</sup> Our Bayesian model is susceptible to the first criticism; however, in the present case, we argue that the model provides a good first step toward understanding how the features are used once they are extracted. It is possible that the feature extraction process might be modeled by introducing a prior over features (such as the Indian Buffet process prior, Austerweil & Griffiths, 2010; Griffiths & Ghahramani, 2011). We leave this as a prospect for future development.

**Acknowledgments** This work was supported by an ARC Discovery Project Grant DP120103888 and an Australian Professorial Fellowship to the second author.

## References

- Austerweil, J. L., & Griffiths, T. L. (2010). Learning invariant features using the transformed Indian Buffet process. *Advances in Neural Information Processing Systems*, 23.
- Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, 35, 499-526.
- Carlstedt, B., Gustafsson, J.-E., & Ullstadius, E. (2000). Item sequencing effects on the measurement of fluid intelligence. *Intelligence*, 28, 145-160.
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review*, 97, 404-431.
- Doucet, A., Freitas, N. de, & Gordon, N. (2001). *Sequential monte carlo methods in practice*. New York, NY: Springer.
- Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32, 108-154.
- Griffiths, T. L., & Ghahramani, Z. (2011). The Indian Buffet process: An introduction and review. *Journal of Machine Learning Research*, 12, 1185-1224.
- Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 37, 720-738.
- Little, D. R., Lewandowsky, S., & Craig, S. (2012). Working memory capacity and fluid abilities: The more difficult the item, the more more is better. *Manuscript submitted for publication*.
- Lovett, A., Forbus, K., & Usher, J. (2010). A structure-mapping mode of Raven's Progressive Matrices. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Marr, D. (1982). *Vision*. New York: W. H. Freeman and Company.
- Matzen, L. E., Benz, Z. O., Dixon, K. R., Posey, J., Kroger, J. K., & Speed, A. E. (2010). Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavior Research Methods*, 42, 525-541.
- McGreggor, K., Kunda, M., & Goel, A. (2010). A fractal analogy approach to the Raven's test of intelligence. In *AAAI workshops at the 24th AAAI conference on Artificial Intelligence* (p. 69-75). Atlanta, GA: Association for the Advancement of Artificial Intelligence.
- Meo, M., Roberts, M. J., & Marucci, F. S. (2007). Element salience as a predictor of item difficulty for Raven's Progressive Matrices. *Intelligence*, 35, 359-368.
- Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence*, 30, 41-70.
- Rasmussen, D., & Eliasmith, C. (2011). A neural model of rule generation in inductive reasoning. *Topics in Cognitive Science*, 3, 140-153.
- Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and vocabulary scales. Section 4: The Advanced Progressive Matrices*. Oxford, UK: Oxford University Press.
- Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational model: Alternative algorithms for category learning. *Psychological Review*, 117, 1144-1167.
- Sewell, D. K., & Lewandowsky, S. (in press). Attention and working memory capacity: Insights from blocking, highlighting and knowledge restructuring. *Journal of Experimental Psychology: General*.
- Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic Bulletin & Review*, 17, 443-464.
- Verguts, T., & De Boeck, P. (2002). The induction of solution rules in Raven's Progressive Matrices. *Journal of Cognitive Psychology*, 14, 521-547.
- Verguts, T., De Boeck, P., & Maris, E. (2000). Generation speed in Raven's Progressive Matrices test. *Intelligence*, 27, 329-345.
- Vodegel Matzen, L. B. L., van der Molen, M. W., & Dudink, A. C. M. (1994). Error analysis of Raven test performance. *Personality and Individual Differences*, 16, 433-445.

<sup>5</sup>We also tested a model with an expanded set of logical rules (e.g., NAND, NOR, etc) but this made no difference to the qualitative pattern of model fits.

# Easing and rising of tension from presence of others in player-observer turn-taking in a driving video game: A near-infrared spectroscopy study

Tao Liu (liu@cog.human.nagoya-u.ac.jp)

Hirofumi Saito (saito@is.nagoya-u.ac.jp)

Misato Oi (oi@cog.human.nagoya-u.ac.jp)

Matthew Pelowski (matthew@cog.human.nagoya-u.ac.jp)

Department of Cognitive Informatics, Graduate School of Information Science, Nagoya University  
Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

## Abstract

Social facilitation and social support literature, diverging with regards to increasing versus decreasing of an individual's tension, apprehend different aspects of "the presence of others." To examine the neural correlates of social presence effects, whether "the presence of others" increases or decreases an individual's tension, we measured prefrontal activation while participants performed a driving video game task using near-infrared spectroscopy (NIRS). Participants were divided into single and paired groups, and then sub-divided according to their game proficiency (high and low). The participant's task was to drive from start to goal with a default route map without an observer (single group) or under observation by an acquainted partner (paired group). The paired participants alternated their player-observer roles in a turn-taking style (Driver first and Observer second: D1-O2; Observer first and Driver second: O1-D2). The behavioral data demonstrated that, regardless of game proficiency, D1 in the paired group yielded fewer errors and longer driving time than single players, while no differences were found between D1 and D2. The tension evaluation scores in single players and D2 were higher than D1. In turn, the NIRS data revealed that, in low-proficiency players, single players and D2 who first observed D1's performance showed higher activation than D1, but neither did so in high-proficiency players. These results suggest that the presence of an acquainted partner (O1) functions positively to reduce an individual's (D1) tension in low-proficiency players. However, prior observation of another's performance may negate the positive social presence effect leading to an increase of tension in the subsequent task.

**Keywords:** presence of others; social facilitation; social support; individual difference; prefrontal cortex (PFC); near-infrared spectroscopy (NIRS).

## Introduction

Social cognitive neuroscience is a burgeoning interdisciplinary field combining the tools of cognitive neuroscience with questions and theories from various social sciences such as social psychology. Classical literature in social psychology has been primarily subsumed under two heads: direct interpersonal influence via interaction between persons and indirect interpersonal influence induced by the presence of others (Allport, 1920).

The latter is a fundamental, as well as the oldest, experimental research in social psychology. As put by Gordon Allport (1954), "the first experimental problem ... was formulated as follows: What change in an individual's normal solitary performance occurs when other people are present?" (p. 46). The present study considers this type of question, and aims to examine the effects of social presence on the individual's neural state in a player-observer dyadic situation.

Two main existing areas of research deal with different aspects of the presence of others. One is social facilitation that investigates how social presence affects one's performance in a general way. Another is social support that focuses on the issue of how other person present relaxes an individual in the stressful environments.

Social facilitation literature has revealed inconsistent effects of social presence on performance; both performance improvement and impairment are possible. For instance, Floyd Allport (1924) demonstrated positive influence from social presence, coining the term *social facilitation* to describe the increase of response merely from the presence of others. However, not all research shows positive effects. Sometimes the presence of others impairs an individual's performance (e.g., Pessin, 1933). To explain the seemingly conflicting results, Zajonc (1965) offered a predominant interpretation based on the Hull-Spence drive theory. According to Zajonc's arousal theory, the presence of others increases an individual's general arousal level, which in turn enhances the emission of dominant responses. In a simple task, appropriate responses are typically dominant, and accordingly the presence of others will improve performance; whereas in a complex task, appropriate responses are more typically not dominant, thus performance will be impaired.

There has been general agreement with this arousal-based explanation in the following social facilitation literature, with considerable debate mostly centered on the source of arousal itself—evolving several conceptualizations such as evaluation-apprehension theory (Cottrell, 1972), monitoring theory (Guerin & Innes, 1982), and distraction-conflict theory (Baron, 1986; for review see Aiello & Douthitt, 2001; Guerin, 1993; Uziel, 2007). These theories clearly differ in their explanations for performance effects of social presence. However, attempts to pinpoint a single exclusively accurate theory have been proven unsuccessful (Guerin, 1993), due

mainly to two reasons: 1) the existing theories are not mutually exclusive—"the theories are unable to predict performance effects in such a way that eliminates other possible explanations" (Aiello & Douthitt, 2001); 2) these theories all attempt to explain why simple task performance is improved and complex task performance is impaired in presence of others without objective criteria for determining the task complexity (Uziel, 2007).

Extensive literature on social support, however, has consistently shown that social presence not only functions to increase an individual's tension level, it also decreases an individual's tension as an emotional coping recourse (e.g., Cohen & Wills, 1985; Lazarus & Folkman, 1984; Lazarus, 1999). In light of stress and coping theory, when the individual evaluates an observer as non-supportive, social presence would cause stressful circumstances, whereas when the individual appraises the others as supportive, social presence would produce relaxation.

Therefore, the incongruent results in previous social facilitation literature may be concomitant, if we accept that the effects of presence of others may be changed positively or negatively according to the cognitive setting that an observer regards others such as a dynamically changing state of the observer. For instance, proficiency in performance of a player would be one of the most critical factors that may change the meaning of others for the player him or herself.

To better understand the functional formation and mechanisms underlying the above social presence effect, there has been a growing effort to explore these outcomes in the presence of others via activity changes in the brain. Using electroencephalography (EEG), Kim, Iwaki, Uno and Fujita (2005) reported larger error-related negativity (ERN) at three brain locations (Fz, Cz, and Pz) in children when they performed a go/no-go task under observation by a friend than when performed individually. The results suggest that social presence may increase one's tension level and accordingly affect behavior as well as attitudes and feelings.

In contrast, in a functional magnetic resonance imaging (fMRI) study, Karremans, Heslenfeld, van Dillen and Van Lange (2011) demonstrated that the presence of a supportive partner reduced prefrontal activation due to easing of tension when participants endured stress during a ball-tossing game. It should be noted that, however, the partner in this fMRI study was not really present, but only virtually so via imagination. One of the reasons stems from technical limitation of brain imaging such as fMRI that is unable to assess cortical function in ambulant participants in social environments.

Near-infrared spectroscopy (NIRS) is also a non-invasive method for studying functional activation by measuring changes in the hemodynamic properties of the brain. Unlike fMRI, NIRS has few physical constraints on participants and is tolerant to motion artifact permitting serial assessments of tasks in relaxed and realistic settings (Cui, Bryant, & Reiss, 2012). In particular, Liu, Saito, and Oi

(2012) have used a 2-channel NIRS unit named as PocketNIRS due to its portability (length: 100 mm; width: 61 mm; thickness: 18.5 mm, and weight: 100 g including the batteries), and mobility (transmitting the hemodynamic signals wirelessly via Bluetooth) to investigate intrapersonal and interpersonal cognitive processes during a driving video game. They assigned participants into one control and two experimental groups. The participant's task in the control group was to drive to goal with a route-map illustrating default turning points, while the memory group was instructed to drive the memorized default route without map (intrapersonal process), and the emergency group was asked to drive with route-map but to change the default route immediately by an extrinsically given "verbal command" (interpersonal process). The results demonstrated an instantly increased activation in prefrontal cortex (PFC) during an urgent turning maneuver resulting from the "direct" interpersonal influence via verbal command, but not from the intrapersonal process.

With respect to social presence effects (i.e., "indirect" interpersonal influence), using NIRS, Ito et al. (2011) have measured prefrontal activation when participants performed a working memory task with or without evaluative observation by experimenters. The participant's task was to observe a sequence of stimuli, and to judge whether a currently presented stimulus was identical with the one presented *n* trials previously. They found that the participants under observation by the experimenters yielded more errors and showed higher activation in both left and right PFC than those who performed without observation. The results demonstrate that the presence of others, for instance strange experimenters in their experiment, increases an individual's tension and influences the prefrontal activation.

Early studies of social presence effects have mainly employed strangers or friends as observers. In the present study, to sustain homogeneity between single and paired groups, the participants were recruited from new students who took a general course of psychology, and the participants in the paired group were matched to soften the extreme polarization of familiarity, and to keep impartial appraisal of the pairs of acquainted participants.

We aimed to extend from the existing literature on social presence effect—demonstrating both the positive and the negative aspects of social presence in one experiment. To address these issues, we measured bilaterally the prefrontal activation in participants when they performed the goal-achievement driving task used in Liu, Saito and Oi (2012) either without an observer (single group) or under observation by a partner (paired group). Participants in both the single and the paired groups were divided into two sub-groups depending on their game proficiency (high and low). The paired participants were asked to alternate their player-observer roles in a turn-taking style (D1-O2: Driver first and Observer second; O1-D2: Observer first Driver second), exploring the possibility that in the first driving task the presence of a partner (O1) may act as a supporter of D1 in

unfamiliar experimental environments, whereas in the second driving after observation of D1's performance, O2's presence may change its role into a source of stress (i.e., non-supporter).

We tested the following three hypotheses: first, the participants in the paired group (D1) would show lower prefrontal activation than those in the single group due to easing of tension resulting from presence of an acquainted partner (positive presence effect); second, D2 would show higher prefrontal activation in the subsequent driving than D1 due to rising of tension based on observation of preceding D1's error performance (negative presence effect); third, low-proficiency players would be somewhat more sensitive to the social presence than the high-proficiency players (task proficiency effect).

## Method

### Participants

Sixty-two right-handed students (53 males, 9 females, age:  $21 \pm 2.2$  years) from Nagoya University participated in the present study for the course credit. Participants were assigned to either single or (same-gender) paired groups, and subdivided according to their game proficiency (high and low). The pairs partnered with each other voluntarily, and their friendships—defined as the duration of their acquaintance—were assessed by self-report in the post questionnaire (friendship:  $1.7 \pm 1.4$  years). All participants had normal or corrected-to-normal vision. They were informed about the purpose and safety of the experiment, and written informed consent was obtained prior to participation. This study was approved by the local ethics committee.

### Materials and design

The same driving video game used in Liu, Saito and Oi (2012) was employed in the present study. During the experiment, players took a seat in front of a 32-in. monitor either individually in the single group or with a partner sitting beside in the paired group. The driving game was displayed on the monitor without sound, and the players controlled the game using a Sony game pad. Distance from the players to the monitor was set to 120 cm.

The participants were asked to obey the traffic rules and drive from start to goal with a default route-map without an observer in the single group or under observation by an acquainted partner in the paired group. Further, two instructions were given to participants in the paired group: 1) the player's performance would be evaluated by their partner as an observer, who needed to report the player's driving performance after the experiment; and 2) they would be asked to alternate their player-observer roles in a turn-taking style during the experiment. With respect to performance, in the present study we defined driving errors as that which lead to collision or driving on the pavement, however, this criterion was not explained to the participants.

### Procedure

Players practiced operating the game pad for 180 s, and then they drove two training trials followed by four experimental trials with distinct routes. A single trial consisted of a driving phase and two rest phases (20 s each) before and after the driving phase.

### Apparatus

The PocketNIRS (DynaSense Inc., Japan), operated at 735, 810 and 850 nm wavelengths, was used to measure the concentration changes of oxygenated hemoglobin (CoxyHb), deoxygenated hemoglobin and total hemoglobin. Two probes were attached to the forehead using double-sided adhesive sheets and centered on Fp1 and Fp2 positions, according to the international 10–20 system. Each probe consisted of one emitter optode and one detector optode located 3 cm apart. During the experiment two sets of PocketNIRS triggered by one signal were employed to measure the activation changes in paired player and observer simultaneously. The sampling rate for each channel was 10 Hz.

### Data analysis

The NIRS data which contained more than 10% non-near-infrared light signals was defined as noise data. All noise data, as well as data obtained from participants who did not follow the instructions, was excluded from further analysis. Complete data was obtained from 15 single participants (6 high-proficiency, 9 low-proficiency), and 18 pairs of participants (D1: 10 high, 8 low; D2: 8 high, 10 low).

We focused on CoxyHb during the driving phase in each group, since the oxygenated hemoglobin is the most sensitive parameter of regional cerebral blood flow (Hoshi, Kobayashi, & Tamura, 2001). A linear baseline correction was conducted on the NIRS raw data to remove longitudinal signal drift using the mean value of CoxyHb during the 5 s before the driving phase. Then *z*-scores were calculated using the mean value and the standard deviation of CoxyHb during the baseline period in four experimental trials and in both the left and the right hemispheres, independently. To eliminate influence of the errors made by the players during driving on brain activation changes, the data during the error periods was excluded from the NIRS dataset. The *z*-scores were averaged finally for the driving phase over all trials, and group-averaged *z*-scores for each group were obtained.

## Results

### Behavioral data

In the present study, we calculated the driving time and counted the number of errors in the driving phase as the performance indices. Statistical analysis was conducted by means of Statistical Package for the Social Sciences (SPSS) and the significant level was set at  $p < 0.05$ .

**Single group vs. paired group (D1)** Figure 1 illustrates the driving performance including the driving time and the number of errors in the single and the paired (D1) groups. To examine the effects of the presence of a partner as an observer (O1) on the player's (D1) performance, we separately performed a two-way analysis of variance (ANOVA) on the driving time and error numbers with social presence (single and paired) and game proficiency (high and low) as the between-participants factors.

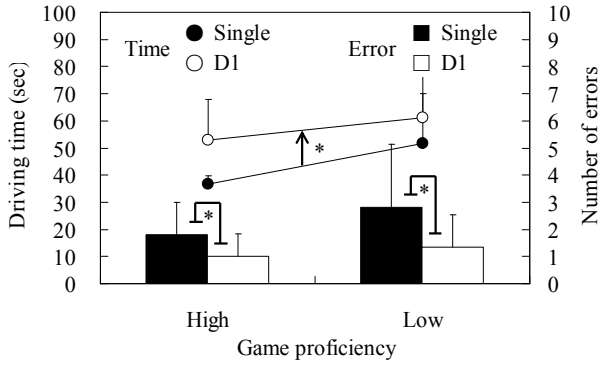


Fig.1 Mean driving time and number of errors in the single and the paired (D1) groups. D1 refers to the first driver in the paired group. Error bars represent standard deviation. \* indicates  $p < 0.05$ .

For both the driving time and the error numbers, analyses revealed significant main effects of social presence, respectively [ $F(1,29) = 4.49, p < 0.05, \eta_p^2 = 0.13$ ;  $F(1,29) = 4.36, p < 0.05, \eta_p^2 = 0.13$ ]. The participants in the paired group (D1) showed fewer errors and longer driving time than those in the single group. Neither the main effects of game proficiency nor the interactions were significant. These results indicate that participants performed better under observation by an acquainted partner than when alone, regardless of their individual game proficiency.

**Single vs. D2** To examine the social presence effects by O2 on D2's performance in the subsequent driving, we applied a two-way ANOVA [O2 presence (single vs. D2)  $\times$  game proficiency (high vs. low)]. For error numbers, no significant differences were found between single players and D2. For driving time, the analysis revealed a significant interaction [ $F(1,29) = 5.36, p < 0.05, \eta_p^2 = 0.16$ ]. In the simple main effect test, D2 showed a significantly longer driving time than the single high-proficiency players [ $F(1,12) = 7.35, p < 0.05, \eta_p^2 = 0.38$ ], but low-proficiency players did not. No significant differences were found between low- and high-proficiency players in either the single players or D2. These results suggest that after prior observation of D1's performance, the positive effect of social presence on performance disappeared in the subsequent driving of D2.

**D1 vs. D2 in the paired group** Figure 2 shows the driving performance in D1 and D2 within the paired group. To assess the effect of the prior observation of D1's performance on the subsequent driving of D2 under observation by O2, we performed a two-way ANOVA on the driving time and the error numbers independently with observation experience (D1 and D2) and game proficiency (high and low) as the between-participants factors. The result revealed no significant differences for both the driving time and the error numbers.

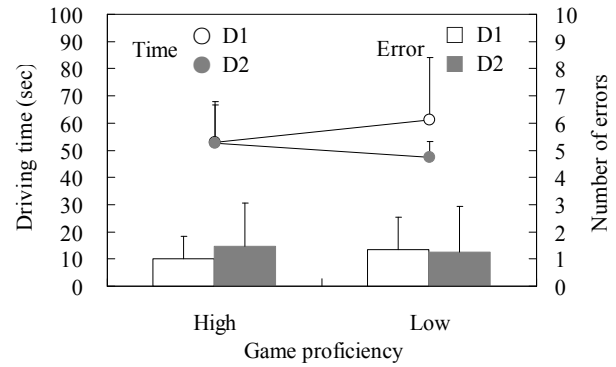


Fig.2 Mean driving time and number of errors in D1 and D2. D1 refers to the first driver in the paired group; D2 refers to the second driver in the paired group. Error bars represent standard deviation. \* indicates  $p < 0.05$ .

**Rating scores on participant's tension level** The tension scores were obtained through a questionnaire filled out by the participants after the experiment. The scores were on a 5-point scale (1 = not at all tense, 5 = extremely tense). The tension index shown represents the average and standard deviation of the participant's response in two domains (unsettled feeling and stress feeling). The tension index was  $1.5 (\pm 0.6)$  in the single players,  $1.1 (\pm 0.2)$  in D1, and  $1.4 (\pm 0.6)$  in D2, respectively. Paired t-test analysis revealed that the single players and D2 showed significantly higher tension than D1, respectively [ $t(17) = 2.06, p < 0.05, 1$ -tailed;  $t(21) = 1.90, p < 0.05, 1$ -tailed].

## NIRS data

**Single group vs. paired group (D1)** Figure 3 shows the average values of the z-score for CoxyHb in the driving phase in the single and the paired (D1) groups. To examine the social presence effect on prefrontal activation, we performed a two-way ANOVA [social presence (2)  $\times$  game proficiency (2)] in each hemisphere separately. In both the left and the right hemispheres, the analyses revealed significant main effects of game proficiency [ $F(1,29) = 8.75, p < 0.01, \eta_p^2 = 0.23$ ;  $F(1,29) = 7.29, p < 0.05, \eta_p^2 = 0.20$ , respectively], and interactions [ $F(1,29) = 11.10, p < 0.005, \eta_p^2 = 0.28$ ;  $F(1,29) = 6.24, p < 0.05, \eta_p^2 = 0.18$ , respectively].

In the simple main effect test, low-proficiency players showed significantly lower prefrontal activation in the paired group (D1) than those in the single group [ $F(1,15) = 11.83, p < 0.005, \eta_p^2 = 0.44$ ;  $F(1,15) = 7.44, p < 0.05, \eta_p^2 = 0.33$ , respectively], but high-proficiency players did not. In the single group no significant differences were found between high- and low-proficiency players. Whereas in the paired group, low-proficiency players showed significantly lower prefrontal activation than high-proficiency players [ $F(1,16) = 13.02, p < 0.005, \eta_p^2 = 0.45$ ;  $F(1,16) = 8.72, p < 0.01, \eta_p^2 = 0.35$ , respectively]. These results suggest that the presence of O1 decreased the tension level of D1 in low-proficiency players, but not in high-proficiency players.

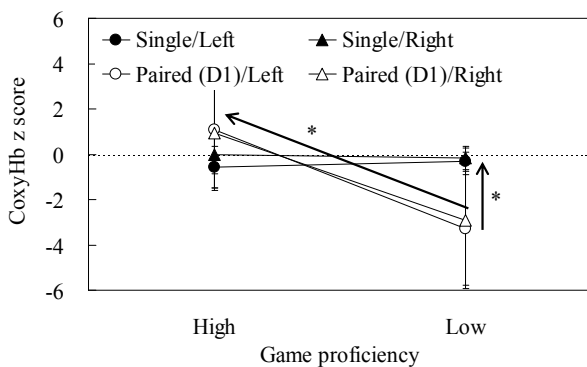


Fig.3 Average concentration changes of oxygenated hemoglobin (CoxyHb) in the driving phase in the single and the paired (D1) groups. D1 refers to the first driver in the paired group. Error bars represent standard deviation. \* indicates  $p < 0.05$ .

**Single vs. D2** To examine the effects of O2's presence on D2's prefrontal activation, we conducted a two-way ANOVA [O2 presence (2)  $\times$  game proficiency (2)]. In both the left and the right hemispheres, no significant differences were found between single players and D2. The results suggest that the positive presence effect by O2 disappeared in the second driving of D2 within the same player-observer pairs.

**D1 vs. D2 in the paired group** Figure 4 shows the average values of the z-score for CoxyHb in the driving phase in D1 and D2. To examine the effect of the prior observation of D1's performance on D2's prefrontal activation in the subsequent driving task, we conducted a two-way ANOVA [observation experience (2)  $\times$  game proficiency (2)] in both the left and the right hemispheres, respectively.

In the left hemisphere, ANOVA revealed a significant interaction between observation experience and game proficiency [ $F(1,32) = 9.22, p < 0.005, \eta_p^2 = 0.22$ ]. No significant main effects were found. In the simple main effect test, low-proficiency players showed significantly higher prefrontal activation in D2 than in D1 [ $F(1,16) = 6.14, p < 0.05, \eta_p^2 = 0.28$ ], but high-proficiency players did not. In D1 low-proficiency players showed significantly

lower prefrontal activation than high-proficiency players, but did not in D2.

In the right hemisphere, the results demonstrated a significant main effect of game proficiency [ $F(1,32) = 4.33, p < 0.05, \eta_p^2 = 0.12$ ]. Neither the main effect of observation experience nor the interaction was significant. These results suggest that after prior observation of D1's performance, the presence of O2 increased the tension level of D2 in the subsequent task.

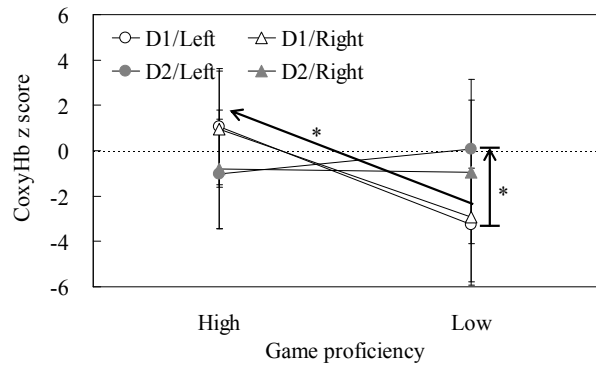


Fig.4 Average concentration changes of oxygenated hemoglobin (CoxyHb) in the driving phase in D1 and D2. D1 refers to the first driver in the paired group; D2 refers to the second driver in the paired group. Error bars represent standard deviation. \* indicates  $p < 0.05$ .

## Discussion

The present study was designed to examine the neural substrate of social presence effects in a natural player-observer environment. To achieve this goal, we measured prefrontal activation in participants without an observer (single group) and with an observer (paired group) during a driving video game using PocketNIRS. In this regard, we tested whether two paired groups (driver first D1 and driver second D2) manipulated in a player-observer turn-taking style consistently demonstrated lower prefrontal activation than the single players, regardless of prior experience of observation in D1 and D2.

Three main findings were obtained, and will be discussed in turn. First, the present data demonstrated lower prefrontal activation in the paired group (D1) than in the single group. The result is consistent with our hypothesis suggesting that the presence of others may serve as a supportive role relaxing an individual (positive presence effect).

Second, in the same social environment the present data revealed higher prefrontal activation in D2 than in D1. This result indicates that the supportive role of the observer may change to a non-supportive role, and increase an individual's tension (negative presence effect).

Third, as predicted, the above two effects were confirmed only in low-proficiency players, but not in high-proficiency players (task proficiency effect).



A unique aspect of the present study is that we demonstrated controversial aspects of social presence effects within one experiment: the presence of others may act positively to relax the individual as well as negatively to stress the individual, depending upon how the individual evaluates the role of the observer (supporter or non-supporter). Previous social facilitation studies have mostly emphasized the negative aspects of social presence leading to rising of tension. Social presence, however, is not just a major source of stress. Social support literature has also demonstrated the benefits of the presence of others to the individual's level of tension (e.g., Cohen & Wills, 1985; Lazarus & Folkman, 1984). Consistent with social support, the present study confirmed that the presence of others could reduce an individual's tension level. It is particularly interesting that after the prior observation of the partner's performance, the supportive effect of social presence disappeared; the supportive role of the observer may change to non-supportive role in the subsequent task.

The present study provides an important theoretical implication. The early social facilitation and social support literature has mainly focused on two distinct aspects of social presence effects, respectively (e.g., Cohen and Wills, 1985; Zajonc, 1965). The present study bridges a gap between them suggesting that research into social presence effects would benefit from combining the ideas of two theories and addressing the role of observer as an important moderating variable subject to subjective appraisal of the observer.

In conclusion, the present study suggests that research into social presence effect would be benefited by addressing individual differences, specifically how an individual evaluates the role of others, as well as the individual's task proficiency. Further study is needed to explore the neural correlates of the explicit role of the presence of others during cooperation and competition.

## References

- Aiello, J. R., & Douthitt, E. A. (2001). Social facilitation from Triplett to electronic performance monitoring. *Group Dynamics: Theory, Research, and Practice*, 5, 163-180.
- Allport, F. H. (1920). The influence of the group upon association and thought. *Journal of Experimental Psychology*, 3, 159-182.
- Allport, F. H. (1924). *Social psychology*. Cambridge, Mass.: Houghton-Mifflin.
- Allport, G. W. (1954). The historical background of modern social psychology. In G. Lindzey (Eds.), *Handbook of social psychology*, Vol. 1, Cambridge, Mass.: Addison-Wesley.
- Baron, R. S. (1986). Distraction-conflict theory: progress and problems. *Advances in Experimental Psychology*, 19, 1-40.
- Bond, C. F., & Titus, L. J. (1983). Social facilitation: a meta-analysis of 241 studies. *Psychological Bulletin*, 94, 265-292.
- Cohen, S., & Wills, T. A. (1985). Stress, social support, and the buffering hypothesis. *Psychological Bulletin*, 98, 310-357.
- Cottrell, N. B. (1972). Social facilitation. In C. G. McClintock (Eds.), *Experimental social psychology*, New York: Holt, Rinehart & Winston.
- Cui, X., Bryant, D. M., & Reiss, A. L. (2012). NIRS-based hyperscanning reveals increased interpersonal coherence in superior frontal cortex during cooperation. *NeuroImage*, 59, 2430-2437.
- Granados, C., & Wulf, G. (2007). Enhancing motor learning through dyad practice: contributions of observation and dialogue. *Research Quarterly for Exercise and Sport*, 78, 197-203.
- Guerin, B. (1993). *Social facilitation*. Cambridge, England: Cambridge University Press.
- Guerin, B., & Innes, J. M. (1982). Social facilitation and social monitoring: a new look at Zajonc's mere presence hypothesis. *British Journal of Social Psychology*, 21, 7-18.
- Hoffman, S. G., Moscovitch, D. A., Litz, B. T., Kim, H. -J., Davis, L. L., & Pizzagalli, D. A. (2005). The worried mind: autonomic and prefrontal activation during worrying. *Emotion*, 5, 464-475.
- Hoshi, Y., Kobayashi, N., & Tamura, M. (2001). Interpretation of near-infrared spectroscopy signals: a study with a newly developed perfused rat brain model. *Journal of Applied Physiology*, 90, 1657-1662.
- Ito, H., Yamauchi, H., Kaneko, H., Yoshikawa, T., Nomura, K., & Honjo, S. (2011). Prefrontal overactivation, autonomic arousal, and task performance under evaluative pressure: a near-infrared spectroscopy (NIRS) study. *Psychophysiology*, 48, 1562-1570.
- Karremans, J. C., Heslenfeld, D. J., van Dillen, L. F., & Van Lange, P. A. M. (2011). Secure attachment partners attenuate neural responses to social exclusion: an fMRI investigation. *International Journal of Psychophysiology*, 81, 44-50.
- Kim, E. Y., Iwaki, N., Uno, H., & Fujita, T. (2005). Error-related negativity in children: effect of an observer. *Developmental Neuropsychology*, 28(3), 871-883.
- Lazarus, R. S. (1999). *Stress and emotion: a new synthesis*. New York: Springer.
- Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*, New York: Springer.
- Liu, T., Saito, H., & Oi, M. (2012). Distinctive activation patterns under intrinsically versus extrinsically driven cognitive loads in prefrontal cortex: a near-infrared spectroscopy study using a driving video game. *Neuroscience Letters*, 506, 220-224.
- Pessin, J. (1933). The comparative effects of social and mechanical stimulation on memorizing. *American Journal of Psychology*, 45, 263-270.
- Uziel, L. (2007). Individual difference in the social facilitation effect: a review and meta-analysis. *Journal of Research in Personality*, 41, 579-601.
- Zajonc, R. B. (1965). Social facilitation. *Science*, 149, 269-274.

# Intentionality of Strong Anticipation in Motor Behaviors

Hsi-wen Daniel Liu (hwliu@pu.edu.tw)

Center for General Education, Providence University  
Shalu, Taichung 433, TAIWAN, R.O.C.

## Abstract

Pezzulo (2008) and Grush (1997, 2004, 2007) highlight, even insist, the role of representation and internal models in anticipatory systems, in contrast to the role of adaptivity in the so-called ‘mere adaptive systems’. The present paper argues against their claim, by alternatively arguing that the strong anticipation—anticipation without employing internal models—is primordial in the making of the anticipatory behavior, while internal models are supplementary in the light of efficiency. A novel notion of intentionality is raised for strong anticipatory behaviors, based on the history of on-line adjustments. The supplement of internal models on the top of a strong anticipatory system makes the resulting anticipatory system a Popperian machine, and consequently more flexible.

**Keywords:** Weak and strong anticipation; internal models; intentionality; motor behaviors.

## Introduction

The embodied and situated approach to cognition is a general revolt against the representationism, the thesis that cognition consists of representations. This thesis is challenged radically by the behavior-based robotics (e.g. Brooksian robotics, Gibsonian theory of vision) and dynamic systems approach to cognition. Wherein, the role of agent-environment interaction is highlighted in the making of cognition, and the notion of cognition without representation is radically raised. Within this approach, however, some contend that the role of representation should be preserved (Clark 1997, 2003; Keijzer 2001; Grush 1997), as is manifest in Clark’s notion of action-oriented representation. Later on, within the same approach does the need of representation scale up in the study of anticipatory systems. The role of representation (or, internal models) in such systems is highlighted, or even insisted, in contrast to the role of adaptivity in the so-called ‘mere adaptive systems’ (Pezzulo 2008; Grush, 1997, 2004, 2007). Model-based anticipation, thus, is seen as contrasted with adaptivity or reactivity, as is technically manifest in the contrast between weak and strong anticipation (Dubois 2003; Stepp and Turvey 2010).

According to Dubois (2003), anticipation is the determination of current states by taking account of future states. Weak anticipation predicts the future states with *models* of agents and the environment; whereas, strong anticipation predicts future states without such models. The strong anticipation employs the *system* itself (i.e. the agent and its immediate environment),<sup>1</sup> rather than an internal model (of agents and the environment). The role of models

<sup>1</sup> A slogan in favor of the strong anticipation is that the system itself is the best model (Stepp and Turvey 2010).

in anticipatory behaviors is well noticed (Pezzulo 2008)<sup>2</sup>, as anticipatory behaviors are maintained control with internal representations. By contrast, the role of strong anticipation seems to be underestimated. In fact, strong anticipation needs more study (Steep and Turvey 2010). That is, the role of adaptivity in the making of anticipation remains in need of research.

In what way do motor activities bear an intentional relation to the environment? This query is a bit hard to answer. The motor activities, on the one hand, are clearly not managed by thought. Nevertheless, such activities, on the other hand, do not seem to be completely meaningless, as manifested in a query of Wittgenstein’s: ‘When I raise my arm, my arm goes up. And the problem arises: what is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?’ (Wittgenstein 1953, 1, paragraph 621). Motor activities, later, are considered in Merleau-Ponty’s (1962) notion of *motor intentionality* as activities that are between reflexes and deliberate actions (Kelly 2000). Such activities, as Merleau-Ponty describes, are controlled by “a motor power, a ‘motor project’ (*Bewegungsentwurf*), a ‘motor intentionality’ in the absence of which the order remains a dead letter (Merleau-Ponty 2006: 126-127)”. But, the question remains as to what this “motor power” is. If it is something that makes motor activities intelligent, then the question, further, would be twofold. The first is a ‘what is’ question: what is it that makes motor activities intelligent? The other question relates to the way in which the term ‘motor intentionality’ can make sense: can that motor power be understood with a certain sense of intentionality? Those two questions have been responded since the end of the last century.

Regarding the first question, the topic of motor control is well noticed both in psychology (Kawato 1999; Jeannerod 2006; Desmurget and Grafton 2000, to cite only three) and in philosophy (Christensen and Hooker 2002; Clark 2002; Clark and Grush 1999; Dreyfus 2007; Grush 1999, 2004, 2007). Motor activities are well conceived of in terms of anticipatory agents, emulators, feed forward models and feedback loops, etc. Regarding the second question, it has

<sup>2</sup> Pezzulo (2008: 179) understands anticipatory behaviors in terms of internal representation: “we argue that the ability that characterize and defines a true cognitive mind, as opposed to a merely adaptive system, is that of building representations of the non-existent, of what is not currently (yet) true or perceivable, of what is desired”, and ... “[a] real mental activity begins when the organism is able to endogenously (i.e. not as the consequence of current perceptual stimuli) produce an internal representation of the world in order to select and guide its conduct goal-directed: the mind serves to coordinate with the future.”

for long been conceived (as aforementioned in the notion of motor intentionality) that to account for motor activities would there be a sense of intentionality, which is very different from Brentano's notion of intentionality, a notion based on a standing-for relation that is typically adopted in the orthodox AI. However, two recent accounts attempt to re-affirm the importance of the standing-for-based intentionality in explaining anticipatory behaviors. The first, Grush (1999, 2004, 2007) maintains a difference between representational anticipation systems and adaptive systems. He highlights the role of internal representation in emulators—standing in for the actual motor activities—without discussing whether there is another sense of 'representation' (in other words, vehicle of intentionality) that can explain motor behaviors. The other, Pezzulo (2008) advocates the role of internal representation in explaining the guidance of goal-directedness in the anticipatory behaviors. Common in Grush and Pezzulo are two proposals: firstly, the contrast between truly cognitive minds and mere adaptive systems; and secondly, explaining cognitive mind/systems in terms of internal representations of the *non*-actual activities. A system with such internal representations, which can vanish before they are actually carried out, is called a Popperian machine, as Popper (1996) says that it can "let its hypothesis die in its stead" (cited from Pezzulo 2008: 195).<sup>3</sup> That system, in Pezzulo's (2008) term, is a system that can act on its representations instead of acting on its reference. Here, being a Popperian machine is taken as a requirement of a system's being cognitive.

The present paper argues for the three-fold primordial role of strong anticipation in the making of the anticipatory behavior. As we will see in the following sections, the strong anticipation is made possible by means of agent-environment coupling and system's (internal) states, while the weak anticipation by means of modeling. Firstly, the strong anticipation is primordial in the making of the anticipatory behavior, while internal models are supplementary in the light of efficiency. It is strong anticipation rather than the internal models of the system that makes possible anticipation. Secondly, a novel notion of intentionality, in contrast to that on the basis of the standing-for relation, is raised for understanding strong anticipatory behaviors: the relation in which the internal states of a body *brings about* a pre-registered end-state in the environment. On grounds of that novel sense of intentionality, strong anticipatory systems are intentional, apart from its being intelligent. Thirdly, the supplement of internal models on the top of a strong anticipatory system makes the resulting anticipatory system a Popperian machine, which is qualified to be a full-blown cognitive agent. The present paper puts discussions in the context of motor activities.

## Anticipatory Behaviors

### Internal Models

Reaching movement is a simple but paradigmatic example of the anticipatory behavior. The inverse internal models provide the feed forward motor commands that are necessary to bring about the desired trajectory in relation to a goal (Kawato 1999). Thus, prior to the onset of the reaching movement is a motor plan assembled. In the course of reaching movement, the motor plan is updated continuously by internal feedback loops. Internal feedback loops are employed because the biological mechanism of sensory feedback loops produces significant delays (Kawato 1999; Desmurget and Grafton, 2000). The internal feedback loops rely on forward models that integrate the sensory inflow and motor outflow to evaluate the consequence of the motor commands sent to a limb. The forward internal model predicts sensory consequences of issued motor commands (the probable position and velocity of an effector) with negligible delays, or even predicts them in advance. Thus, the forward models make the strategy of feedback loops efficient in the light of making a *real time* response for the reaching movement (Desmurget and Grafton, 2000). The representational status of anticipatory behavior is considered based on the importance of internal models, that make anticipatory behavior sufficiently efficient and accordingly make it possible in the ecological niche (Grush, 1997, 2004, 2007; Pezzulo, 2008).

### Strong Anticipation

An anticipatory system is a system that takes account of future states. A strong anticipatory system is an anticipatory system without employing an internal model; otherwise, it is a weak anticipatory system (Dubois 2003; Steep and Turvey 2010).

Steep and Turvey (2008) understands the notion of reaction in the sense that the determination of systems' states only takes account of current or past states, without considering future states. In this sense, homeostasis can be seen as reactive, as it is seen as involving no explicit goals at all. Examples of systems with no explicit goals are as follows: the single-cell *Euglena* approaching to the sunlight (Goodale and Milner 2004: 40), a reflex, the Watt Governor (WG, van Gelder 1997), and the wall-following machine (Martaric 1990). The reflex as the resulting response is not pre-specified in the reflex mechanism of the muscle system. The WG is a device for the speed control. The state with a constant speed is the end-state, without being explicitly specified in the device. The end-state is an end-point of a purely physical chain. This makes the WG to be merely adaptive, although it is indeed an intelligent design.

By contrast, a thermostat, which is a goal-pursuing machine, embraces an explicit goal that is regarded as a pre-fixed state. The explicit goal is to be compared with a current state, and an error to be reduced is consequently derived. In addition, the outfielder discussed in Steep and

<sup>3</sup> A similar point is put in terms of a Popperian creature (Dennett, 1995, p. 375)—a creature capable of breaking their cycle of direct interactions with the local environment (Clark and Grush, 1999).

Turvey (2008) has an explicit end-state—catching the flying ball.

The strong anticipation preserves a role contributing to flexibility of motor behaviors. Such a role is two-fold. On the one hand, a strong anticipatory system approaches to the goal in the complex environment. For example, by lining up with the flying ball that becomes a fixed point (Clark 2003), the outfielder is likely to catch the ball successfully. On the other, the motor controlling system, as a strong anticipatory system, can further connect to various inner models and consequently makes the system at stake a *weak* anticipatory system. The former one, that achieves the goal, is the primary concern of an anticipatory system. Whereas, the latter, that is supported with inner models, makes the anticipatory system a Popperian machine and consequently more effective and efficient. As to the effectiveness, avoiding some likely impediment and thus making the way to achieve the goal more likely to success. As to the efficiency, it makes the way of achieving the goal passing various real-time constraints.

### **Adaptivity in Strong Anticipation**

This section considers two remarkable ways of adaptivity appearing in strong anticipatory systems: coupling and sensory feedback loops.

#### **Coupling Ensures Completing a Task**

The agent-environment coupling, as considered previously in the example of an outfielder, ensures the completion of a task. When such coupling arises, a task would be accomplished effectively and efficiently. Consider the aforementioned outfielder problem. When the outfielder moves to a position where the trajectory of the ball is seen to be a point heading toward him, the coupling arises, and catching the ball is straightforwardly ensured. The accomplishment of the task is sometimes determined by that coupling condition.

In addition, coupling is inherent in the Gibsonian theory of vision and action. Visual affordances of an object, say, a cup, provide opportunities of actions, for example, grasping the cup. Consider the simple tasks consisting of single actions, such as the outfielder's catching the ball as aforementioned, and the task of grasping a cup. A coupling relation arises between the affordances of an object and the consequently provided action. As the visual affordances provide opportunities of an action, the coupling relation ensures the completion of a corresponding task.

#### **The Sensory Feedback Control for the System's Performance**

The sensory feedback control need *not* rely on an internal model,<sup>4</sup> if the efficiency of motor movement is not taken

into account. The former part, the role of the sensory feedback control, is based on the dual model theory of motor movements, according to which a motor movement begins with a motor plan that rapidly transports the hand near to the target, and then depends on sensory feedback loops that slowly direct the hand to the target. The latter part, the non-efficiency of the sensory feedback control, is evident in the remarkable delay resulting from the sensory feedback control (Desmurget and Grafton, 2000). By contrast, a feed forward model, as manifest in emulators (Grush 2004), is an internal model that provides a simulation of motor actions. Based on a simulated motor action, when compared with the target, an error is derived internally.

To be noted, the derivation of an error and the following behavior of getting closer to the target on the basis of the sensory feedback control, is managed by the *system* without recourse to internal models. By contrast, the *internal* feedback loops (as opposed to the sensory feedback loops) rely on a forward model that makes the feedback strategy efficient enough for the real-time reaching movement. The resulting efficiency can be understood as it is a Popperian machine. However, as will be argued in next section, the *intentional* relation of the reaching movement is basically manifest in the system with sensory feedback loops, if we can provisionally put the consideration of efficiency aside. Such sensory feedback loops unfold on account of the system's actual activities and the sensory feedback control. The sensory feedback control of the system (as opposed to a model), hence, is primordial in the making of the motor control. The agent's way of relating to the target turn up primordially in the system's feedback loops.

### **Intentionality of the Strong Anticipatory System**

Do strong anticipatory systems bear intentionality? If they do, in what sense are they intentional? Let us put discussions in the context of motor movement. Such questions relate to two considerations. For the first consideration, such a notion of intentionality, if there is, seems to have something beyond the standing-for relation, given that motor movements are not completely thought-like. The goal-state is brought about as an end-state of a *causal* chain.<sup>5</sup> The question, more specifically, concerns the sense of intentionality which is born by causal activities. Then, the second consideration is that the motor movements, if they indeed bear intentionality in any sense, should be distinct from physical reactions. This section discusses the above two questions, centered on the notion of the goal.

---

<sup>4</sup> An internal model is a model of the system, including the agent and its immediate environment (Dubois 2003). A remarkable example is emulators (Grush 2004). The term 'internal' is

---

contrasted with 'external', including the system as an embodied agent and the environment.

<sup>5</sup> An end-state of a motor system is a state specifying the way in which the target relates to its environment and the body.

## An Account of Pragmatic Intentionality

To respond to the first consideration, a novel notion of intentionality—pragmatic intentionality—is raised: an agent's motor activities bear intentionality in the sense that internal states of the agent's body *bring about* an end-state explicitly pre-registered in her perception or perceptual imagery. If we consider how it is possible that those internal states could bring about a specific end-state of a causal chain, then we need to understand the evaluative nature of the goal. The end-state is taken as the goal on grounds of evaluation. There should be discrepancies marked as negative values against states away from it, and there should be machinery of reducing such negative values. Thus, pragmatic intentionality can be further characterized: internal resources (such as sensory feedback loops, bones and muscles, energy, etc.) of an agent's motor system evaluate the discrepancies between the current state and an explicitly pre-registered end-state of the motor task, and reduce those discrepancies with the result of bringing about that end-state in the environment.

The involving motor activities are authentically intentional, as the agent with its body's internal resources cashes out an explicitly pre-registered end-state in perception or perceptual imagery by bringing it about in the environment. Those bodily internal resources transform the pre-registered end-state in perception or imagery into an end-state in the environment, an end-state consisting of inter-relations between the body, the target object and their environment. This is a pragmatic sense of intentionality because internal resources of the body are *used for* bringing about a pre-registered end-state. As a note, the above consideration has nothing to do with an internal model with a standing-for relation. As an example, a thermostat has pragmatic intentionality as it consists of a feedback model in relation to a pre-registered end-state.

To respond to the second consideration, the present paper highlights the role of the aforementioned pre-registered goal-state in the motor system. The end-state of the motor processing initially turns up in perception or perceptual imagery, and later is registered in the motor system before the motor processing unfolds. This pre-registered status makes the motor processing be neither physical nor reactive. It promotes the processing of the motor system from the physical level to an intentional level. To be noted, the goal has a two-fold role in the motor system. Despite the standing-for status of the goal, it does not stand on its own as an internal model (such as an emulator) of the motor system. The goal has a standing-for relation as it is initially marked in the perception or perceptual imagery. Yet, once it is pre-registered as the end-state of the motor processing, it stands in the *causal* relation to states of the motor system. A goal, hence, can be a part of a strong anticipatory system. This two-fold role based on the pre-registered goal-state in the motor system makes the motor system be neither purely physical nor completely reactive to environmental circumstances.

## A Primitive Shooting System and a Coaching System

A weak anticipatory behavior involves three elements: the system itself (or, agent), its environment, and internal models, while a strong anticipatory behavior involves only the former two elements. Consider the cognitive role of a strong anticipatory system, which is an agent in its environment. Let us discuss in the context of a premature shooting system supervised by a coaching system. Suppose that shooting system is capable of hitting the target with the probability of 10 % per shooting, that is, in average, hitting the target once in ten shoots. Let us further consider a shooting system with a supporting condition: the coaching system serves as a model that simulates the shooter's ten shoots, selects the best one, and accordingly the model draws the shooting system to the shooting conditions of that selected (best) one. Thus, the shooting system can shoot with 100 % accuracy in average. Such a shooting system is an anticipatory system, as the model suggests the shooting system on the basis of (simulated) future conditions.

Then, in theory, that shooting system would be capable of hitting the target accurately more efficiently. The aforementioned shooting system provides an analogy of an accurate motor system, in the sense that the internal model stands as a coaching system of the motor system.

One may contend that a motor system without the internal modeling is but a non-cognitive adaptive system like homeostasis. The coaching system is definitely cognitive, as it is a Popperian machine. Yet, a shooting system itself cannot be cognitive without the support of a coaching system. That the motor behavior is cognitive is because of the representing role of the internal models.<sup>6</sup> Such a representing role is that the internal models simulate and accordingly stand for the motor system together with its environment.

## Intentionality, Representation, and Decouplability

### The Intentionality of Anticipatory Systems

An anticipatory system can be seen as bearing intentionality, in the sense different from the Brentano's notion of aboutness—the traditional sense of intentionality. Note that intentionality is a relation between the internal states of a system and the world. Consider the reaching behavior as an example. The system directs the body of a robot or a biological body, such as limbs, to the target in the environment. The achievement of a goal, hence, cannot appear internally. Rather, it should take place with the real body of the anticipatory system in its immediate environment. The achievement of a goal should be controlled internally in the anticipatory system, with or without the support of an internal model. The internal states

<sup>6</sup> The theme of the present paper is intentionality, but not cognition *qua* cognition, hence the question of defining cognition is simplified.

of that system are dynamic, which may optionally be supported with internal models (as weak anticipatory systems). Let us discuss in terms of the control theory. In a strong anticipatory system, those internal states include the making of a motor plan (as an inverse model of the goal), sensory feedback loops (producing an error by comparing the current position of the hand and the goal), and the hand movement with a view to minimizing the errors. A weak anticipatory system, by contrast, is further supported with internal models. Both in the strong and the weak anticipatory systems, the internal states subserve the body in the light of reaching the target.

As a reminder, we previously made a pragmatic definition of intentionality for strong anticipatory systems. The goal-directed agent processes resources of an agent's internal states and uses them such that the agent's performance reaches an explicitly pre-registered end-state of processing. The goal-directed agent is qualified to be intentionality-bearing since the processing of internal states brings about performance in the environment. Note that the goal-directed action is made by the anticipatory *system's* internal control (not by internal *models*). The internal control is manifest in the agent(hand)-environment coupling when the feedback loops are running. The internal control shows a tendency to reach the goal state.

How do the internal states of an anticipatory system work apart from the functionality of its internal models? In weak anticipatory system, internal models (e.g. emulators) represent the way in which the system will unfold in the environment. They represent on grounds of the aboutness relation, that the internal states of the system stand for the system's mechanisms and the environmental conditions. By contrast, apart from the internal models, the internal states of the anticipatory system *encode* the way in which the system manages to bring about the bodily reaching. How could the strong anticipatory system bear intentionality?

### Rowlands' Account of Representation in Deeds

The notion of strong anticipatory behaviors is closely addressed by Rowlands (2006) in terms of *deeds*: deeds consist of "an array of on-line, feedback-modulated adjustments that take place below the level of intention, but, collectively, promote the satisfaction of the antecedent intention (p. 103)." Rowlands claims that deeds are truly representational, where representation is characterized with the five conditions: (i) informational condition, (ii) teleological condition, (iii) decouplability condition, (iv) misrepresentation condition, and (v) combinatorial condition. Among them, what seems hardest to justify is the decouplability condition: "[i]tem *r* qualifies as representing states of affairs *s* only if *r* is, in an appropriate sense, decouple from *s*. The problem is that it seems to conflict with their *on-line* nature of feedback-modulated adjustments. Even if this problem is solved, a further problem is that simple adaptive devices, such as the WG, the thermostat, and homeostasis, would consequently appear to be qualified as representing.

Those two problems seem likely to be solved in Rowlands' justification of decouplability: the learning history of a motor system provides it with a proper function, which can run "off-drive" in the absence of environmental stimuli (p. 166). Repetitive practice of a motor action, say, catching a flying ball, results in a proper function of ball-catching. The inner states of the motor action are decouplable from the real stimuli of a flying ball, as such states can run in absence of those stimuli. The activating instances of a catching-ball device may differ while some of them might fail in a catching action. Yet, the learnt motor system has obtained the proper function of ball-catching.<sup>7</sup> In addition, animals can learn, whereas the WG, the thermostat, and a homeostatic system cannot. As a consequence, such adaptive devices cannot have a bearing of representing.

Rowlands' justification of decouplability, however, assumes intentionality instead of explaining it, as it resorts to the history of learning (in his term, practice). The present paper would justify the aforementioned decouplability without recourse to either practice or learning, but, instead, relies on the history of the system's on-line adjustments. A successful catching system may operate on different objects, and may even fail in some instances; yet, it is truly a system that represents the states of affairs of catching-a-flying-ball. Such a behavior does not causally depend on a specific kind of objects or events in the environment; hence, that motor system represents the capability of catching-a-flying-ball.

By contrast, the WG, the thermostat, and a homeostatic system, are not representing in the previous sense. The WG and the thermostat are (artificial) devices that do not retain a history of adjustments. In addition, a homeostatic system does not respond to the common living world, world in the literal sense. Here we preserve the term of representation to systems that are capable of responding to the common living environment. Homeostatic systems may be perfectly anticipatory system; yet, it seems to stand at the nebulous boundary of representation that anticipatory systems have.

### Conclusions

The present paper explains motor intentionality, by arguing that strong anticipation is primordial in the making of the anticipatory behavior, while internal models are *supplementary* in the light of efficiency, like a coach supporting a baseball team. A baseball team can perform without a coach despite its lacking good strategies for winning.

The strong anticipatory system bears intentionality because it transforms itself from a body registering a perceptual or imaginary end-state into a body which realizes that end-state in the environment. Such an intentionality is pragmatic because internal resources of the body are used

<sup>7</sup> Rowlands deems that this way of justification is "all the decouplability we can reasonably require for the deed (167)."

for bringing about an end-state in the environment. The pragmatic nature of such an intentionality makes clear two components of the goal-directedness: an explicit end-state pre-registered in perception or imagery, on the one hand, and an evaluative sub-system and a machinery that reduces the negative values, on the other. The former component is put intuitively as a goal and the latter is machinery that brings about the goal in the environment. Thus, we can succinctly put the pragmatic intentionality of strong anticipation in terms of the relation of *bringing-about*, that is, internal states of the body bringing about the end-state in the environment. A novel sense of intentionality, in contrast to that based on Brentano's notion of aboutness, is characterized in terms of a system's history of adjustments.

A notable difficulty in considering a novel sense of intentionality for motor activities is the question of how to distinguish between motor intentionality and mere adaptivity. Both of them are put in causal chains; yet, the former is intentional, while the latter remains purely physical (or biological). To resolve this difficulty, the present paper resorts to the history of (on-line) adjustments in the systems' response to the real environment. This is a feature unseen in merely adaptive systems such as homeostasis, the Watt Governor, and single-cell *Euglena* approaching to the sunlight. They do not register their history of (on-line) adjustments, and accordingly are not anticipatory systems.

Motor behaviors are anticipatory. The present paper explains why and how they are representing, in response to the questions addressed by Wittgenstein and Merleau-Ponty, as mentioned in the introduction. The supplement of internal models on the top of a strong anticipatory system makes the resulting anticipatory system a Popperian machine, which makes the anticipatory systems more flexible. Because they are representing, the strong anticipatory systems discussed in the present paper are not 'mere adaptive systems' as conceived of by Pezzulo (2008) and Grush (1997, 2004, 2007). Internal models in the anticipatory systems are supplementary, although they are necessary in the light of efficiency that is manifest in efficient motor behaviors.

### Acknowledgments

This research is supported by National Science Council, Taiwan, R.O.C., under grant NSC 100-2410-H-126-018.

### References

- Brentano, F. (1974). *Psychology from an Empirical Standpoint*. London: Routledge & Kegan Paul.
- Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again*. Cambridge, MA: MIT Press.
- Clark, A. (2002). Skills, spills and the nature of mindful action, *Phenomenology and the Cognitive Sciences*, 1: 385-387.
- Clark, A. (2003). *Natural-Born Cyborgs*. Oxford: Oxford University Press.
- Clark, A., and Grush, R. (1999) Toward a cognitive robotics. *Adaptive Behavior*, 7, 5-16.
- Desmurget, M. and Grafton, S. (2000). Forward modeling allows feedback control for fast reaching movements, *Trends in Cognitive Sciences*, 4 (11), 423-431.
- Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian, *Artificial Intelligence*, 171: 1137-1160.
- Dubois, D. M. (2003). Mathematical Foundations of Discrete and Functional Systems with Strong and Weak Anticipations. *Lecture Notes in Cur Science*, 2684: 110-132.
- Grush, R. (2007). Skill theory v2.0: Dispositions, emulation, and spatial perception, *Synthese*, 159(3):389-416.
- Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences* 27:377-442.
- Grush, R. (1997). The Architecture of representation, *Philosophical Psychology*, 10(1)5-23.
- Jeannerod, M. (2006). *Motor Cognition: What Actions Tell the Self*. Oxford: Oxford University Press.
- Kawato, M. (1999). Internal models for motor control and trajectory planning, *Current Opinion in Neurobiology*, 9, 718-727.
- Keijzer, F. (2001). *Representation and Behavior*, Cambridge, MA: MIT Press.
- Kelly, S.D. (2000). Grasping at straws: Motor intentionality and the cognitive science of skilled behavior. In M. Wrathall and J. Malpas (Eds.) *Heidegger, Coping, and Cognitive Science: Essays in honor of Hubert L. Dreyfus, volume 2*, Cambridge, MA: MIT Press.
- Martaric (1990). A distributed model for mobile robot environment-learning and navigation. Technical Report no. 1228, MIT Laboratory.
- Merleau-Ponty, Maurice, (1962). English translation by Colin Smith, *Phenomenology of perception*, London: Routledge.
- Pezzulo, G. (2008). Coordinating with the Future: The Anticipatory Nature of Representation. *Minds and Machines*, 18, 179-225.
- Popper, K. R. (1996). *Alles Leben ist Problemlösen. Über Erkenntnis, Geschichte und Politik*. München: R. Piper-Verlag.
- Rowlands, M. (2006). *Body Language: Representation in Action*. A Bradford book, MA: MIT Press.
- Stee, N. and M.T. Turvey (2010). On strong anticipation. *Cognitive Systems Research* 11: 148-164.
- van Gelder, T. (1997). Dynamics and Cognition, in *Mind Design II*, ed. by J. Haugeland, Cambridge, MA: MIT Press.
- Wittgenstein, L. (1953). *Philosophical Investigation*, Oxford: Blackwell.



# Is Embodied Cognition Infallible or Falsifiable?

## Investigating the Thesis as a Sound Scientific Theory

Katherine A. Livins & Leonidas A.A. Doumas (klivins; leonidas@hawaii.edu)

University of Hawaii at Manoa, Department Psychology  
2530 Dole St., Sakamaki C400, Honolulu, HI 96822

### Abstract

Embodied cognition is a growing area of research within cognitive science—one that it is often presented as a framework that may help us account for cognition as a whole. It is, however, a theory and, as such, it must live up to the requirements that all scientific theories do. Of particular importance is the degree to which it is falsifiable. This paper investigates this issue.

**Keywords:** embodied cognition, embodiment, falsifiability

Embodied cognition has been a growing theoretical framework within the field of cognitive science for the past twenty years. It has influenced research on topics from low-level perception (e.g., Meteyard *et al.*, 2008) to high-level reasoning (e.g., Casasanto, 2009) and, as it continues to grow, it is often put forth as an organizing explanation of cognition in general (e.g., Schubert & Semin, 2009). That said, it is a *theory*—specifically, a scientific one. As a result it must be evaluated as any theory would be. To date, the field has done a reasonable job: confirmatory evidence has rolled in, and its methodologies have been refined (e.g., Spivey, 2007). However, despite the field's efforts to do good science through the lens of embodiment, one fundamental evaluation factor seems absent: “Is it *falsifiable*?” remains an open question. Given the importance of falsifiability to scientific enterprise, answering the question is necessary. Addressing this question of falsifiability is the goal of this paper.

### Part 1: What is falsifiability?

Falsifiability is a cornerstone of scientific theory. As Popper (1963) argued, the difference between theories such as Freudian psychoanalysis and Newtonian physics lies not in whether it is possible to verify each theory, nor in whether they possess explanatory power (e.g., a way of explaining experimental results), but in their capacity to be debunked.

For a theory to be *falsifiable*, it is insufficient to only provide ways of supporting it. Instead, it must be possible to specify the types of evidence necessary for showing that the theory would *not* hold. Because it is impossible to directly observe a theory (since a theory is only an organizing principle built to explain a body of information or phenomena), defining only what would support it may allow for confirmation biases and the exclusion of important data. Thus, a scientific theory must involve making specific predictions about what must or should be observed if it holds, *and* what

must or should not be observed if it does not.

Take, for example, the Bohr model of atomic structure. While it is unnecessary to go into detail, the model generally specified a structure of the atom and the way that its nucleus could be related to its electrons. Despite its seeming simplicity, the theory made predictions, including what the energy levels of certain atoms should have been if the theory was correct. Thus, boundaries of what the model could and could not account for were established. When these predictions were violated (e.g., multi-electron atoms displayed energy levels different from what the model predicted) the theory was falsified.

Having the ability to falsify a theory becomes crucial when competing theories exist. If each theory is not specific enough to make predictions, then it becomes impossible to decide which one (if any) is correct. Furthermore, theories that do not make such predictions are often over-generalized (e.g., Freudian psychoanalysis) such that they can account for almost any sort of data. As a result, they can be warped to account for any phenomenon, data, or unique case, creating a confirmation bias. In other words, they become explanatory catch-alls with little real explanatory power. Falsifiability protects against exactly these sorts of problems and, as a result, should play an important part in the development of any theory.

### Part 2: Embodiment and Falsifiability

“Embodied cognition” is something of an umbrella term for work that is interested in the ways in which the body is involved in cognitive processing. In fact, there are so many definitions under the umbrella its boundaries are somewhat contentious. Take, for example, the quantitative issue of how many types of embodiment there are: Shaprio (2011) outlines three types; Wilson (2002) presents six, and there are more accounts still (e.g., Anderson, 2003; Ziemkie, 2003).

This lack of agreement could be considered a complication with regard to evaluating the falsifiability of embodiment; after all, if there exists multiple theories, then each one may require separate analyses. However, the heart of the differences usually centers on what it means *philosophically* for the body to be involved in cognition. For example, the *Enactive Approach* seems premised on the idea that there is no core self that resides inside a body, but that the self dynamically makes itself through interactions with the world. Furthermore, the world is not taken as pre-given,

instead being shaped through the actions of the individual (Shapiro, 2011). As a result, all cognition is enacted through a symbiosis of body and world. *Situated cognition* is similar but it does not break down the concept of “self”, instead maintaining a philosophic distinction between self and not-self (Shapiro, 2011). Cognition ends up being equally dependent upon (and shaped by) pressures from the world, though the self remains a meaningful unit of discussion and study.

Despite the many definitions that exist within the philosophic literature, cognitive scientists and experimental researchers, generally define and operationalize embodiment as some version of the claim that the body affects cognition, or that the body plays a meaningful role in it (whether that be through changes in perception and attention, differences in behavior, or the activation of neural motor systems) (e.g., Markman & Brendl, 2005; de Koning & Tabber, 2011). Thus, while there do exist competing philosophic offshoots of this basic idea, this more basic theoretical claim underlies most of them; it is at the heart of how the theory is defined and used empirically. The goal of this paper is to look at embodiment within this type of empirical (read, practical or scientific, not philosophic) context; we are interested in the ways in which experiments are designed when inspired by it, and the ways in which results are interpreted in light of it. In other words, we are interested in embodiment as a true scientific theory. If it does not live up to this type of scientific standard, it is important to ask whether it should be used as the basis for empirical work, or alternatively, whether there is something that can be done to make it live up to scientific standards. To the point, we proceed with an analysis of existing work, with an eye specifically towards whether the current application of embodiment meets the requirement of falsifiability.

To start this analysis we begin by looking at some of the research that has been done thus far. The remainder of this section will analyze two papers that investigate the affects of the body on cognition, and which are often cited as quintessential exemplars of this type of research. Special attention will be paid to the interpretations of the findings and their potential for allowing embodied cognition to be falsified.

### Part 2.1

The paper “Action observation and acquired motor skills: An fMRI study with expert dancers”, by Calvo-Merino *et al.* presents the idea that motor history affects future cognitive processing. It has been cited approximately six hundred times<sup>1</sup>, often as evidence for a meaningful relationship between mind and body (e.g., Spivey, 2007; Barsalou, 2008).

The study addresses whether one’s motor history affects one’s ability to observe movements in others. For this project, a number of expert dancers (skilled in either ballet or capoeira), along with an untrained control group, were recruited. All participants were asked to watch videos of choreographed sequences from both dance styles. As participants watched the videos, neural activity was monitored with fMRI. Of interest were the premotor and parietal cortices, along with the superior parietal lobe and the superior temporal sulcus—areas previously associated with human action, and observation of action in others (Calvo-Merino *et al.*, 2004; Grafton *et al.*, 1996).

The hypothesis was that activity in these areas would be strongest when watching a dance sequence that one had enacted firsthand. This result was observed. Calvo-Merino *et al.* argued that it must be the case that watching physical activity in others activates some sort of sensory-motor representation (Calvo-Merino, 2004), furthering the idea that bodily experiences affect cognition (as embodied cognition theory posits).

In order to address the degree to which this design offers the chance for falsification of embodied cognition, let us consider other possible outcomes of the study, and how Calvo-Merino *et al.* could have interpreted them. First, the dancers (and/or the control group) could have shown more neural activation when watching dance sequences that they had not engaged in firsthand. Embodiment could be supported by such a result equally well by arguing that the finding was due to a need for learning: motor areas activated more for understanding movements that could not, immediately, be made sense of by referencing past experiences. In fact, this interpretation would be consistent with other data, since it has been shown that the same motor areas that Calvo-Merino *et al.* were interested in demonstrate increased activation during visuomotor learning (Ghilardi *et al.*, 2000). Consequently, whether or not the dancers showed more or less activation than controls, the embodied cognition thesis would still be supported given that bodily experiences would be shown to affect cognitive processing in either case.

Alternatively, the same activation could have been seen between the dancers and the control group, irrelevant of the observed sequences. However, this finding would not necessarily contradict embodiment either. Perhaps the *use* of representations does not require activation of the motor-systems, but encoding is entirely dependant upon them and, as a result, the body still affects cognition. Or, perhaps more radically, one might suggest that such an outcome implies that humans do not possess representations at all. In fact, this is a popular claim among some embodied cognition theorists (Varela, *et al.*, 1992; Smith, 2005) and it could be that no difference was found between groups because there are no static representations stored in the

---

<sup>1</sup> Per GoogleScholar.

brain to be activated in the first place, leaving only motor areas to encode the video information.

While the possible interpretations of the results of the present study certainly indicate that embodiment theory is confirmable, it does not bode well for its ability to be falsified. In fact, it would be equally possible to explain Calvo-Merino *et al.*'s findings by adopting a competing perspective. A disembodied framework could be supported by suggesting that the dancers had encoded static representations of the sequences, which included the muscle groups necessary for performing them. This would account for why the motor areas of their brains lit up in response to watching them, without a need for participants to "simulate" anything. This explanatory flexibility is not a new idea; Mahon and Caramazza (2008) explicated the ways in which many findings that are supposed to support embodied cognition (such as Calvo-Merino *et al.*'s) may also be used to support a disembodied theory of cognition. It is troubling, though, that the theory's explanatory flexibility would exist no matter how the experiment turned out.

One could argue that the problems with this study are a function of how easily neuroimaging can be misapplied, and not of embodied theory itself. However, neuroimaging is a tool and, as such, there will always be ways to use it well and ways to use it poorly. It is quite possible to use this tool well (e.g., Engel *et al.*, 1994), but doing so requires the production of specific predictions so that the meaning of the scans can be clearly interpreted. A failure to do indicates a problem with the theoretical framework underlying the scans. In the case of the current study, the problem lays with the fact that any scan result could have been used to support embodied cognition. This is a problem of the theory, not of the technology.

#### Part 2.2

As another example, consider Tucker and Ellis' "On the relations between seen objects and components of potential actions" (1998). With almost as many citations<sup>2</sup> as Calvo-Merino *et al.*'s work, this study looks at whether the presentation of objects also potentiates the affordances<sup>3</sup> that they allow from the human body. The primary concern of the study is how perception is affected by one's body, and the ways that one can interact with the world through it.

More specifically, the study looks at the relationship between the hand movements used when dealing with everyday objects and the ways that those objects are visually perceived. Tucker and Ellis argued that if an object's affordances are part of its representation, then

it should be easier or faster to respond to the object with a physical movement that is somehow aligned with the affordance. They addressed this question by conducting three experiments.

First, they investigated whether participants would be faster to identify an object if it were presented in congruency with the hand responding to the presentation. For example, if a picture of a knife were presented in the direction that it would be grasped with the right hand, then the interest would be in whether a participant would press a response button more quickly with that right hand since movement in the right hand would be primed. However, an obvious problem is that this design does not determine whether participants are simply sensitive to left/right orientation, or handedness specifically. Thus, the second study asked participants to respond with two fingers on the same hand; the idea being that if participants were sensitive to handedness, instead of left/right orientation, then greater efficiency should not be demonstrated on a single hand.

Results indicated that participants responded more quickly to objects when they were presented in congruency with the hand that was pressing the button (i.e., with the hand that would be used to interact with the object as it was presented). Furthermore, this effect was not confused with simple left/right orientation or response since one-handed responses showed no difference in response times between left and right finger responses.

At first glance, these findings support the idea that an object's affordances are a part of how it is perceived and responded to. Consequently, they seem to support the claim that worldly interactions are embodied; however, if the other possible ways that the study could have turned out are analyzed in the same way that Calvo-Merino *et al.*'s study was, it becomes improbable that such a claim could have been denied.

First, Tucker and Ellis could have found that participants responded more slowly when a stimulus was presented to match the side of the grasping hand (i.e., the experiments' results had come out the opposite way). An obvious explanation for such a finding is that participants were asked to push a button instead of actually grasping – given that these are two distinct movements, it could have been that pushing buttons did not represent the affordances properly, causing interference. In fact, Tucker and Ellis realized this, and offered a third experiment to address it.

In the third condition, participants were given a new response method: responding involved the same wrist rotation and hand positioning that would be required for interacting with the presentation objects. Interestingly, Tucker and Ellis did not find a strong effect between hand positioning and response times, however, they concluded that their findings suggest that more than hand selection (left versus right) is involved in seeing

<sup>2</sup> Per GoogleScholar.

<sup>3</sup> A concept developed by Gibson (e.g., 1950), affordances are the potential ways in which one can interact with a given object (i.e., what sorts of bodily movements that object affords the human body).

and responding to objects (Tucker & Ellis, 1998).

Even if Tucker and Ellis had found a strong effect in their third condition, it would be difficult to argue with (i.e., falsify) any embodied interpretation. First, it could have been that participants responded to a stimulus with convergent affordances much more slowly than they responded to a stimulus with divergent affordances. While the exact meaning of such a finding is not obvious, it would suggest that there is some systematic relationship between the ways that we think about an object and the ways we can interact with it. Perhaps it could be argued that when a stimulus is in congruence with the response action, we begin to activate the next step of movement necessary for interacting with the object and, as a result, move more slowly. For example, if we are presented with a teacup, perhaps we begin preparing our mouths for tea, or our throats for swallowing. Such an interpretation would be supported by other experimental data, which suggests that visually controlled grasping motions (especially in children) are preceded by a series of planning motions, sometimes causing longer response times if the plan is non-ideal (Hofsten & Rönqvist, 1988). Alternatively, embodied cognition often posits that representations (if they exist) are dynamically updated to include experiences (Smith, 2005). It could be argued that more processing time is required to recode a perceived thing based on new data.

Finally, it could have been the case that Tucker and Ellis found no relationship at all. Such a null result would likely be interpreted as puzzling and could be explained away by pointing out that this experiment (like any) required a commitment to a specific methodology and, therefore, specific types of movement and stimuli. One could argue that the movements chosen by Tucker and Ellis were simply non-ideal for demonstrating a relationship (perhaps a knife is more about chopping or sharpening than grasping, and perhaps tea cups are more about sipping, pouring, washing, dunking or even swallowing).

In short, the experiment, although motivated by a prediction of the embodied theory of cognition, provided no easy means for falsifying that theory without running conditions with every possible movement ever associated with a given object. However, even then, research into perceptual dominance could be mobilized to explain such null results. For example, it appears that some sensory modalities are more dominant than others. Vision is often considered the most dominant of human modalities (Sinnott *et al.*, 2007). As a result, any study that is interested in the embodied relationship between vision and bodily movement, and that involves visual stimuli while looking for manual or tactile responses, may be considered flawed in that it could be the case that simply showing a participant a visual stimulus activates more visual experiences or representations

(embodied or otherwise) than hand-oriented experiences or representations. If it is the case that visual responses dominate motor responses, then any null results from a methodology employing visual to manual responses could be criticized and dismissed.

Importantly, we are not suggesting that anyone would actually use some of these alternative outcomes to demonstrate that embodied cognition *is* the case (e.g., that some absence of proof is proof for the thesis). The point is that all of these outcomes *can* be dealt with within an embodied framework. Consequently, it does not appear that this experiment (like the last) offers much by way of an opportunity for falsification.

### Part 2.3

We are not lodging any sort of specific attack on the researchers whose work we have reviewed (these papers were selected for their popularity and clarity). Our concern is that given such experiments are used as support for embodied cognition *qua* a scientific theory (e.g., for Calvo-Merino's paper, Gallese, 2008; Barsalou, 2008; de Konin, 2011; and for Tucker and Ellis' paper, Wilson 2002; Semin & Smith, 2008). It is problematic that such experiments could accommodate, or even explicitly support, an embodied account, no matter how they turned out.

Likewise, we are not trying to discredit any of the subtheories or philosophic work that has been inspired by embodied cognition; certainly, their subtheories, and empirical work based on those theories, that do make and test predictions (e.g., Gray *et al.*, 2006; Hommel *et al.*, 2001). However, embodied cognition itself is not just an inspirational tool—it is posited as a theory in and of itself, and at this point, it is one that seems remarkably susceptible to methodologies that suffer from confirmation biases.

### Part 3: Can We Fix The Problem?

Just because something has not been done, does not mean it cannot be. Thus, it seems important to look at embodied cognition and to determine whether it is theoretically *possible* to falsify. We begin by making the requirements of falsification explicit.

For a theory to be useful scientifically, it cannot simply lend itself to supporting research, but must also be sensitive to the ways that it could be shown *not* to hold. This requires specificity: “theory *x* would be untrue if *y* (and *z* and [...]) were to happen”, where *y* and *z* are observable, measurable and definitive. Doing so ensures that everyone (both supporters and dissenters) can recognize when it is time to abandon the theory. Furthermore, we must accept the possibility (no matter how small) that the theory will be falsified. Falsification would not be such a bad thing though—after all, it would mean that we get an opportunity to build a new, more complete theory of cognition.

Currently, embodied cognition does not seem to have

these sorts of specific boundaries. To date, the closest attempt seems to be a set of claims explicated by Wilson (2002). Unfortunately, even they seem too vague (i.e., unobservable, unquantifiable, indefinite) to act as falsifying predictions. To see this vagueness in action, we consider the first three.

First, Wilson argues “cognition is situated”. That is, cognition takes place in an environment and, therefore, cannot take place without perception and action. It is studied by looking for the ongoing impact of perceptual input during cognitive tasks (Wilson, 2002). However, if this claim is a concrete prediction (open to falsification), then embodied cognition fails whenever such tasks are completed without perception or action. However, Wilson herself posits that this prediction may not hold for all types of cognition; she points out that some things do not always rely on the intake of new information (her examples include planning, reasoning, etc.). Thus, the claim does not involve a clear list of conditions under which it does or does not hold. As a result, it seems more like a “sometimes” prediction than a definitive one.

Second, embodied cognition is supposed to support the idea that cognition is “time pressured” (i.e., that, when pressured to work quickly, cognition can differ). This claim does not make any more forward momentum than the first. If it is a prediction, then falsification should occur any time that cognition does not change based on time-pressure. While, of course, many cognitive activities do, Wilson herself provides us with examples of “time-locked” cognition, such as skilled hand movements (Wilson, 2002). As a result, if this claim is a crucial prediction, the thesis is already in trouble. Alternatively, if it is a “sometimes” prediction like the last claim, then it is simply not specific enough.

Third, embodied cognition expects that we “off-load” cognitive work onto the environment. This claim is evidenced by the fact that people make use of items from the world to solve problems (thus lessening the work required of processes such as working memory). Obviously though, there are times that tasks are not off-loaded, even within the realm of a single task. For example, when determining where furniture should go in a room, it is possible to move it around or draw a diagram, or to visualize different patterns that may be used. It is also possible to mentally map the room instead. Embodied cognition is not definitive enough to predict the “why”, or enough of the “how”, to account for anything more specific than that it *can* happen. This seems incomplete and, again, not usefully predictive.

Because of space constraints, we cannot engage in a similar analysis of the other three claims. As they are, though, they offer no opportunity for falsifiability as they also reduce to variations of “the body is important for cognition under some conditions”. While it is true that such a claim can be turned into experimental work,

it is not enough to fill out the statement, “theory *x* would be untrue if *y* (and *z* and [...]) were to happen” with observable, measurable and/or definitive claims. Without the capacity to produce such claims, embodied theory remains unable to produce sufficiently specific predictions to be falsified.

Basing empirical work on a theory with an unclear foundation can be problematic. Any experiment (like those in Section Two) will have confusing results. No matter how such a study turns out, it can account for the data within the embodied framework, and any null result can be dismissed. This lack of definitiveness leaves researchers that are already invested in embodiment frozen in unsolvable debates without a real possibility of coming to a conclusion, let-alone unifying cognition. Similarly, researchers outside the debate may have trouble finding application for embodiment within their own domains.

Newell (1973) argued that Cognitive Science is often guilty of discovering a phenomenon, doing a plethora of experiments to explore it, and never moving on to think about what the research “means”. It seems that this “plethora” point is where embodied cognition is now—at the crux of exploration and synthesis, with a real opportunity to move forward. The collection of confirmatory data definitely suggests that the body is likely important for cognition *in some way*, however, we have yet to answer what this way is in a specific or systematic sense, and this is what needs to be changed.

We are not suggesting that there is a quick fix, however, we believe this issue of falsifiability can guide future research in a meaningful way: Researchers need to start asking a few questions every time they invoke embodiment. First, asking, “Under what specific conditions do we expect this cognitive process to be embodied?” in combination with, “What cognitive functions are most penetrable to bodily or action-based manipulations (and which are not penetrable at all)?” will allow for the boundaries of the theory to be reigned in. It is unlikely that the answer to every cognitive mystery is “embodiment”, so we need to commit to specific claims about when we think embodiment is meaningful and to what processes we think it makes an important difference. Second, asking “What would it mean for embodiment, as a theory, if my results came out differently?” will force attention to the issues raised in this paper. If this answer cannot possibly be “it would suggest that process *x* is not embodied”, then the methodology may need to be revised.

We are certainly not saying that embodiment should be dismissed (quite the opposite, actually). However, to the extent that embodied cognition is going to be used as a theory, limiting cases need to be defined. Any research that does this is a step in the right direction, and would be useful, not only for the embodiment community, but also for interested outside observers.

We cannot offer a definitive end point for such efforts—we wish we could provide a fast, or easy answer, however developing a comprehensive theory that is aimed at organizing cognition in a deep and meaningful way is not a simple task. Our goal is simply to point out what we see as an important problem. It is our hope that those dedicated to embodied cognition research will respond by becoming more sensitive to the issues raised here, and that, as a result, these researchers will start making their claims more explicit and, more importantly, begin establishing limiting cases of the theory.

### Works Cited

- Anderson, M.L. (2003). Embodied cognition: a field guide. *Artificial Intelligence*, 149, 91-130.
- Barsalou, L.W. (2008). Grounded Cognition. *Annual Review of Psychology*, 59, 617-45.
- Calvo-Merino, B., Glaser, D.E., Grezes, J., Passingham, R.E., & Haggard, P. (2004). Action Observation and acquired motor skills: An fMRI study with expert dancers. *Cerebral Cortex*, 15.8, 1243-1250.
- Casasanto, D. (2009). Embodiment of abstract concepts: Good and bad in right- and left-handers. *Journal of Experimental Psychology: General*, 138.3, 351-367
- Descartes, R. (1641). Meditations on first philosophy. In R. Ariew (Ed.), *René Descartes: philosophical essays and correspondence* (pp. 97-141). Indianapolis: Hackett Publishing Company Inc.
- Engel, S.A., Rumelhart, D.E., Wandell, B.A., Lee, A.T., Glover, G.H., Chichilnisky, E.J., & Shadlen, M.N. (1994). fMRI of human visual cortex, *Nature*, 369, 525.
- Gallese, V. (2008). Empathy, embodied simulation and the brain: Commentary on Arago and Zepf/Hartmann. *Journal of the American Psychoanalytic Association*, 56, 769-781.
- Gibson, J.J. (1950). The perception of the visual world. Oxford: Houghton Mifflin
- Ghilardi, M.F., Ghez, C., Dhanwan, V., Moeller, J., Mentis, M., Nakamura, T., Antonini, A., Eidelberg, D. (2000). Patterns of regional brain activation associated with different forms of motor learning. *Brain Research*, 871, 127-145.
- Grafton S.T., Arbib, M.A., Fadiga, L., & Rizzolatti, G. (1996). Localization of grasp representations in humans by positron emission tomography: Observation compared with imagination. *Experimental Brain Res*, 112. 103-111.
- Gray, W.D., Sims, C., Fu, W.T., & Schoelles, M.J. (2006). The soft constraints hypothesis: A rational analysis of resource allocation for interactive behavior. *Psychological Review*, 113.3, 461-482.
- von Hofsten, C., & Ronqvist, L. (1988). Preparation for grasping an object: A developmental study. *Journal of Experimental Psychology*, 114.4, 610-621.
- Hommel, B., Musseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework of perception and action planning. *Behavioral and Brain Sciences*, 24, 849-937.
- de Koning, B.B., & Tabbers, H.K. (2001). Facilitating understanding of movement in dynamic visualizations: an Embodied Perspective. *Educational Psychological Review*, 23, 501-521/
- Mahon, B.Z., Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology – Paris*, 102, 59-70.
- Markman, A.B., Brendl, C.M. (2005). Constraining theories of embodied cognition. *Psychological Science*, 16.1, 6-10.
- Meteyard, L., Zokaei, N., Bahrami, B., & Vigliocco, G. (2008). Visual motion interferes with lexical decision on motion words. *Current Biology*, 18.17.
- Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W.G. Chase (Ed.), *Visual information processing* (pp. 283-308). New York: Academic Press.
- Popper, K. (1963). Science as Falsification. In *Conjectures and Refutations*. New York: Routledge
- Semin, G.R., & Smith, L. (2008). *Embodied grounding: Social, cognitive, affective and neuroscientific approaches*. Cambridge: Cambridge University Press.
- Schubert, T.W., & Semin, G.R. (2009). Embodiment as a unifying perspective for psychology. *European Journal of Social Psychology*, 39.7, 1135-1141
- Shapiro, L. (2011). *Embodied cognition*. New York: Routledge
- Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: The Colavita Effect Revisited. *Perception and Psychophysics*, 69.5, 673-686.
- Smith, L.B. (2005). Cognition as a dynamic system: Principles from embodiment. *Developmental Review*, 25, 278-298.
- Spivey, M. (2007). *The continuity of mind*. New York: Oxford University Press.
- Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance*, 24.3, 830-846.
- Varela, F., Thompson, E., & Rosch, E. (1992). *The embodied mind: Cognitive science and human experience*. Cambridge: MIT Press.
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9.4, 625-636.
- Ziemke (2003). What's that thing called embodiment? In: Alterman & Kirsh (Eds.) *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 1134-1139). Mahwah, NJ: Lawrence Erlbaum

# Natural language – no infinity and probably no recursion

**Erkki Luuk (erkkil@gmail.com)**

Institute of Computer Science, University of Tartu, Liivi 2  
50409 Tartu, Estonia

**Hendrik Luuk (hendrik.luuk@gmail.com)**

Department of Physiology, University of Tartu, Ravila 19  
Tartu 50411, Estonia

## Abstract<sup>1</sup>

We question the need for recursion in human cognitive processing by arguing that a generally simpler and less resource demanding process – iteration – is sufficient to account for human natural language and arithmetic performance. We claim that the only motivation for recursion, the infinity in natural language and arithmetic competence, is equally approachable by iteration and recursion. Second, we submit that the infinity in natural language and arithmetic competence reduces to imagining infinite embedding or concatenation, which is completely independent from the ability to implement infinite computation, and thus, independent from both recursion and iteration. Furthermore, we show that natural language is a finite rather than infinite set.

**Keywords:** recursion; iteration; language; brain; infinity; embedding; arithmetic.

## Recursion and embedding

An influential line of thought claims that a hallmark of human cognitive processing is recursion (e.g. Fitch, Hauser, & Chomsky, 2005; Hauser, Chomsky, & Fitch, 2002; Premack, 2007). Hauser et al. (2002) drew a distinction between the whole language faculty, including the aspects shared with other species or faculties (the faculty of language in the broad sense) and the unique aspects of the language faculty (the faculty of language in the narrow sense). They hypothesized that the unique aspects of the language faculty comprise "only the core computational mechanisms of recursion as they appear in narrow syntax and the mappings to the Sensory-Motor and Conceptual-Intentional interfaces" (Hauser et al., 2002, p. 1573). Lately, this hypothesis has been vigorously challenged (e.g. (Jackendoff & Pinker, 2005; Pinker & Jackendoff, 2005). Interestingly, none of the challenges question the infinity in natural language, and recursion is contested only in Lieberman (2008) and Bickerton (2009) from a neuroscientific and linguistic perspective, respectively.

---

<sup>1</sup> An earlier version of this paper "The redundancy of recursion and infinity for natural language" appeared in *Cognitive Processing* (2011, 12 (1)). As compared to the earlier version, some inaccuracies have been corrected, loosely relevant parts omitted and new arguments added to the present paper.

## Recursion

**Defining recursion** Rogers Jr. (1987, pp. 5-6) gives the following description of a Gödelian recursive definition: "A recursive definition for a function is, roughly speaking, a definition wherein values of the function for given arguments are directly related to values of the same function for "simpler" arguments or to values of "simpler" functions." For example, in the recursive function for defining Fibonacci numbers (for integers  $n > 1$ ),  $Fib(n)$  is directly related to  $Fib(n-1)$  and  $Fib(n-2)$ :  $Fib(n) = Fib(n-1) + Fib(n-2)$ . For the present discussion, the most important aspect of recursion lies in its ability to describe an infinity of (input,output) pairs by a finitely definable set of operations in an effectively computable manner (see Odifreddi, 1992, pp. 34-36, for details).

Recursion differs from iteration (another form of repetition) in two essential respects. First, definitions employing iteration do not involve self-reference. Second, without self-reference, every (input,output) pair needs to be defined explicitly, rendering it impossible to define infinite sets by finite means other than by a control flow loop. Computationally, there is a clear difference between a procedure that invokes another instance of itself (recursion) and a procedure that repeats itself either mechanically or with a control flow loop (iteration). Recursion and iteration are the only computational solutions for handling repetition.

As a rule, implementations of recursive functions are slower than those of iterative because recursive functions must allocate memory for their multiple instances. In Abelson et al. (1996), the difference between recursive and iterative **processes** is captured as follows. Recursive processes are characterized by a chain of deferred operations and require that the interpreter keep track of the operations to be performed later on. An iterative process is one whose state can be summarized by a fixed number of state variables, together with a fixed rule that describes how the state variables should be updated as the process moves from state to state. All this being said, recursive **procedures** (or definitions) tend to be formally and notationally more elegant than iterative ones. The difference between process and procedure is explained below.

The influential paper by Hauser et al. (2002) was the first to explicitly formulate the view that recursion is a component of the language faculty (regrettably, no definition of recursion was given in the article). After



reviewing all statements about recursion in the paper the following definition emerges (numbers in brackets refer to the pages in Hauser et al. (2002)): recursion is a neurally implemented (p. 1574) computational mechanism (p. 1573) that yields a potentially infinite array of discrete expressions (pp. 1570, 1571, 1574) from a finite set of elements (p. 1571). Crucially, a computational mechanism with finite input and potentially infinite output, described here, does not imply recursion, as it is possible to generate a potentially infinite output from a finite set of elements without recursive operation by implementing  $n \rightarrow \infty$  operations iteratively. Hence, it is not clear whether the mechanism described by Hauser et al. (2002) is recursive or iterative. While it is plausible that the notion of recursion as applied by Hauser et al. (2002) refers to the 'recursion' in the Minimalist Program (Chomsky, 1995), the latter allows for a range of interpretations. Tomalin (2011, p. 308) has, usefully, distinguished between nine different interpretations of 'recursion' in formal sciences, with 'inductive definition' as the most broad one (and thus the safest for syntactic theory). Tomalin's (2011) theoretic variants and computational equivalents of recursion ( $\lambda$ -definability, Turing computability et al.) are more or less on the same level of abstraction but recursion can also be defined in five different levels (in the order of decreasing abstractness):

Table 1. 'Recursion' in five different levels.

1. Inductive (or recursive) definition: A definition with a base rule, specifying that certain "simple" objects are in the class, and an inductive (recursive) rule, stating that if certain objects are in the class, then so are certain other objects formed or derived from them (Minsky, 1972).
2. Recursive definition for a function: A definition wherein values of the function for given arguments are directly related to values of the same function for "simpler" arguments or to values of "simpler" functions (Rogers Jr., 1987).
3. Recursive function: A function that is defined recursively (see 2).
4. Recursive procedure: A procedure that, when executed, invokes another instance of itself (Abelson, Sussman, & Sussman, 1996).
5. Recursive process: An execution of a recursive procedure (see 4).

We suggest that, for cognitive and neural modelling, the procedural level (4) is of central importance. For the following discussion, it is crucial that the recursion we are looking for is implemented in the brain (Hauser et al., p. 1574), i.e. it is not something that is posited at the level of computational theory only (Marr's (1982) level 1 of his three levels of information processing). However, the situation is still very confusing, as it is possible to have neurally implemented recursion of level 1 that is

implemented non-recursively (i.e. by iteration) at levels 4-5! In the next section we will give an example of this. It is also important to note that level 1 in Tab. 1 corresponds to Marr's level 1 (computational theory) and levels 4 and 5 in Tab. 1 correspond to Marr's level 2 (algorithm and input/output representation) of information processing. The following section examines whether and how any of the levels in Tab. 1 can be connected to Marr's level 3 (hardware implementation) in the brain.

**Recursion, iteration, and inductive definition** All open-ended sets (e.g. language expressions,  $\mathbf{N}$ ) can be defined inductively, i.e. recursively in the broadest sense. For example, one can have the following inductive definition of 'bear': (a) *Ted is a bear*; (b) *All entities that share at least 98% of Ted's genome are bears*. Observe that, although the set of potential 'bears' is open-ended and inductively defined, no recursion is computationally necessary to determine its contents. An iterative process that compares Ted's genome to that of potential 'bears' would do the job. Importantly, the difference between iteration and recursion pertains to levels 4-5 only. Thus, even if recursion were used in levels 1-3, the involvement of a recursive process or procedure would not be implied, as it can be implemented with a purely iterative computational process (e.g. on a Turing machine).

Below is a strip of iterative pseudocode (1) that defines the infinite set of finite strings  $\dots[X[X[X[XY]]]]$  that is also defined by the recursive strip (2):

```
(1)  $Y \rightarrow XY$  (iteration) :
    s=Y           //assign Y to s
    while true:   //infinite loop:
        rw(Y,XY,s) //rewrite Y as XY in s

(2)  $Y \rightarrow XY$  (recursion) :
    s=Y           //assign Y to s
    rec(s)        //declare function rec(s)
    {             //start definition of rec(s)
        rw(Y,XY,s) //rewrite Y as XY in s
        rec(s)     //call function rec(s)
    }             //end definition of rec(s)
```

As one may observe, (1) and (2) are computationally equivalent – at the level of computational theory, both are described by the rewrite rule  $Y \rightarrow XY$ . Incidentally, this also means that rewrite rules can be "recursive" only in the sense of "recursive definition" (level 1 in Tab 1., which corresponds to Marr's level 1).

From the viewpoint of effective calculability, general recursion and Turing computability are equivalent (Kleene 1952, p. 300), and a universal Turing machine can enumerate any recursively enumerable formal language (an infinite set of finite strings) as can general recursion (Sipser, 1997). Turing machine is based on iteration, whereas general recursion is based only on recursion. Over finite

outputs, recursion and infinite iteration are computationally equivalent (Turing-equivalent).

Crucially, as there is no neural model of recursion (of whatever level), one is unable to identify it in the brain and, accordingly, unable to verify its existence. On the other hand, Lieberman (2008, p. 527) has recently suggested that "neural circuits linking local operations in the cortex and the basal ganglia confer reiterative capacities, expressed in seemingly unrelated human traits such as speech, syntax, adaptive actions to changing circumstances, dancing, and music", thus obviating the need for a neurally implemented recursion. Of course, the distinction between neurally implemented recursive and iterative processes is rather opaque for present-day methods, i.e. the possibility of a neurally implemented recursion cannot be ruled out. We argue for iteration only as a simpler and equipotent computational alternative to recursion.

In sum, there would be no sense in (and no obvious way of) implementing level 1 (in Tab. 1) in the brain separately from levels 4 and 5. As for levels 2 and 3, it would be outlandish to assume that the brain somehow (and apparently redundantly) implements **equations** of the processes it carries out. Thus, implementations of these levels can be ruled out as well. We have also argued that implementations of levels 4 and 5 would be impossible to identify in the brain with present-day knowledge and methods.

**Recursion, induction and self-embedding: a confusion** In computer science, on the procedural level, recursion denotes the syntactic property that a procedure definition refers to the procedure itself (Abelson et al., 1996). In Chomsky's (1956, 1971 [1957]) phrase structure grammar, recursion is a property of rewrite rules (all that is on the left side of the rewrite arrow repeats on the right side of the arrow, e.g.  $A \rightarrow AB$ ) (Chomsky, 1956, 1971 [1957]). Essentially, this notion of recursion reduces to inductive definition. For some other theorists, recursion is a structural property: a situation where an instance of an item is embedded in another instance of the same item (e.g. Heine & Kuteva, 2007; Jackendoff & Pinker, 2005). For clarity, let us call the three recursion, induction and self-embedding, respectively. Recursion and self-embedding are logically independent for the following reasons. First, a self-embedded structure (an NP within an NP, a box within a box etc.) does not have to be recursively generated. Jackendoff and Pinker (2005) submit a picture of a rectangular form within another rectangular form as an example of 'recursion in visual grouping'. Obviously, this has no bearing whatever on recursion. It would be outlandish to assume that recursion is necessary to put a box in a box, or for understanding that a box is in a box. Yet, for conspicuous (but nonetheless insufficient) reasons, this assumption is held with syntactic categories like sentence and NP. The reasons are, of course, the inductive rewrite rules of generative grammar (e.g.  $NP \rightarrow A NP$ ), and they are insufficient as the type of induction can be generated iteratively as well (cf. (1)). Furthermore,

iteration is defined as the repeated application of a transformation (Weissstein, 2003), which is something that Chomsky's (1956, p. 113) description of his early transformational version of generative grammar explicitly incorporates: "/---/ phrase structure is limited to a kernel of simple sentences from which all other sentences are constructed by repeated transformations". The confusion with recursion can be traced back to Chomsky (1971 [1957], p. 24; 1956, p. 115), who refers to loops as 'recursive devices'. The source that the formalism is taken from refers to the loops as 'circuits' ("a closed series of lines in the graph with all arrows on the lines pointing in the same orientation" – Shannon & Weaver, 1964 [1949], p. 47). The circuits pertain to a graphic representation of finite-state Markov chains, and to call them 'recursive' is jumping to the conclusion, as Markov chains do not prescribe an algorithmic realization for the circuits – Markov chains are confined to Marr's level 1 just like inductive definitions. In Chomsky (1959, p. 143), 'recursive function' is used as a synonym for 'Turing computable function'. Again, this is confusing, as computational equivalence does not imply algorithmic equivalence (which is absent in this case). In sum, the Chomskian notion of "recursion" is a case of confusing Marr's levels of information processing (1 and 2).

Fitch (2010) claims that iterative functions are inadequate for generating center- and/or self-embedding. As an example of the superiority of recursion over iteration, he presents a recursive center-embedding program generating  $A^n B^n$ . Below is an iterative pseudocode that does the same:

```
C = ""           //evaluate C to the empty string ""
for i = 0 to n do: //loop n times
  C = concatenate("A",C,"B") //concatenate "A", C and
                             //"B", and assign the result to C
```

The program embeds  $C$  between "A" and "B"  $n$  times, with  $i$  indicating the depth of embedding in each cycle of the loop. It is true that "recursive functions take their own past output as their next input" (Fitch 2010, p. 75) but this feature is not unique to recursion – in our above examples, concatenate() coupled with iteration does the same.

As for confusing recursion with self-embedding (characteristic to most linguists but not to Chomsky), the two are already in principle very different. Recursion pertains to a process or procedure, self-embedding pertains to a structure. A recursive process or procedure is something that, more often than not, cannot be directly observed. Self-embedding, on the other hand, is usually salient and a subcase of a cognitive phenomenon we term 'hierarchical interpretation'. A defining difference between hierarchical and non-hierarchical interpretation is that only the former allows the same unit to be interpreted simultaneously as a type (i.e. category) and as a token (i.e. instance), hence implying additional interpretative correlates not present in the input. The type/token distinction is a precondition for self-embedding, where tokens are embedded under the same type (e.g. NP or clause). An example of hierarchical

interpretation is natural language. Linguistic interpretation is compounding, merging smaller units that are per se meaningful in the code (Chomsky, 1995; Hauser et al., 2002). As far as we know, linguistic code is unique among natural communication systems in stipulating semantic compositionality, whereby meaningful units are combined into diversely meaningful higher-order units (e.g., words into phrases, sentences and compound words, phrases into sentences and higher-order phrases, etc.).

As an illustration that self-embedding is possible in hierarchical interpretation only, consider the following example: the inductive center-embedding rule  $AB \rightarrow AAB$  generates the strings  $AAB$ ,  $AAAB$  etc. It is impossible to tell by looking at these strings whether their generation procedure (or process) was recursion, iteration or neither (cf. Fitch, 2010, and the example above). Furthermore, it is impossible to tell whether the strings exhibit self-embedding. Without any a priori assumptions about the generative mechanism (e.g. stipulation of a certain phrase structure grammar), it is undecidable whether a string  $\dots AAB \dots$  is embedded, concatenated, or elementary (assuming that different generative mechanisms may allow for different elementary strings).

## Embedding

Embedding is a situation where an item is embedded in any item (with infinity not implied). Embedding is logically independent from recursion (i.e. there can be one without the other). First, embedding does not have to be generated by a recursive rule. It can be created iteratively or by any other function with relevant output. Second, a recursive process or procedure does not have to yield (relevant) output. Assuming that we cannot witness a recursive process or procedure in situ (e.g. in the brain), two conditions must be met for attesting it: (1) it must generate output, and (2) there must be a one-to-one correspondence between the values of the recursive procedure and its output. Logically, self-embedding is a situation where an instance of an item is embedded in another instance of the same item (with infinity not implied); thus, self-embedding is a proper subset of embedding. The fact that embedding is hierarchical has frequently raised speculations about a putative underlying recursive process or procedure (or more unfortunately, resulted in confusing embedding with recursion). As explained above, a hierarchical or embedded structure is insufficient to decide on its generative mechanism.

## Infinity in natural language and arithmetic competence

The central claim of Hauser et al. (2002) and Chomsky (2010) is that a neurally implemented recursive process introduces infinity to natural language and arithmetic. An example of (potential) infinity in natural language and arithmetic competence is the **knowledge** that one can add 1 to  $n$ , append a natural language expression to text or embed clauses indefinitely. Of course, we are incapable of

**performing** infinitely in any of these tasks (hence the famous competence/performance distinction – Chomsky, 1995).

Chomsky's derivation of neurally implemented recursion for operating on  $\mathbf{N}$  is as follows. A) Any formal definition of the set of natural numbers  $\mathbf{N}$  incorporates recursion by means of the successor function, where  $1 = S(0)$ ;  $2 = S(S(0))$  etc. B) We have knowledge of the properties of  $\mathbf{N}$  (i.e., given enough time and space, we can compute the sum of two numbers, and distinguish the right from the wrong answer). From premises A and B he conjectures that neurally implemented recursion is required for operating on  $\mathbf{N}$  (e.g. for adding 4555 to 7884). Thus, we have the following: (a) infinity in natural language and arithmetic competence (to motivate neurally implemented recursive process in the first place), (b) neurally implemented recursion is required to operate on  $\mathbf{N}$ , and (c)  $\mathbf{N}$  is an offshoot of the language faculty. From these premises, it follows that (d) neurally implemented recursion underlies both  $\mathbf{N}$  and natural language.

On the face of it, the above argument for neurally implemented recursion is consistent and logically sound. However, premises (a)-(b) are false, and in section "Arithmetic performance" we argue in more detail that neurally implemented recursion is not necessary to operate on  $\mathbf{N}$ . As explained below, infinity in natural language and arithmetic competence reduces to **imagining** infinite embedding or concatenation, and thus does not qualify as an output of a recursive process or procedure (as there is no reason to assume that conceptualizing infinity requires recursion). The concept of neurally implemented recursion is largely motivated by the 'discrete infinity' property of natural language (Chomsky, 1995; Hauser et al., 2002). In fact, the whole distinction between the broad and the narrow language faculties, as originally proposed by Hauser et al. (2002), can be derived from this property. Importantly, the infinity of natural language has been always taken as axiomatic and never proven. The last instance that Chomsky appeals to in this question is Wilhelm von Humboldt (1999 [1836]) who simply states the infinity of natural language as a fact.

One might start from the observation that the maximum possible natural language "corpus" – everything that has ever been and will be processed – is not infinite but a finite, just physically uncountable set. We propose that this is precisely the nature of language as it should be accounted for. In fact, the very spacetime that can support physical computational systems is finite (Krauss & Starkman, 2000). This substantial correction (physically uncountable finity instead of infinity) is suggested for the sake of unambiguity and exactitude. Physically uncountable sets can be finite or infinite. Set-theoretically, potential and actual infinity (Moore, 1990) are proper subsets of physical uncountability. The evidence that Hauser et al. (2002, p. 1571) submit for discrete infinity covers also physically uncountable finity: "There is no longest sentence (any candidate sentence can be trumped by /---/ embedding it in "Mary thinks that ..."). It

would be a contradiction to assume that the size of a finite, physically uncountable array can be compared to the size of all others, or that such array can be embedded (the mere fact that we can imagine embedding such an array does not account for its capacity of being embedded). "There is no non-arbitrary upper bound to sentence length." This is as true for an infinite as it is for a finite, physically uncountable array.

The finity of natural language can be also derived logically, without invoking physically instantiated computation:

1. Natural language has a limit which is either infinite or finite.
2. Natural language computation takes time.
3. From 1 and 2 it follows that, for any given moment in time, there is an infinite number of finite limits that are never reached.
4. Assuming that the cardinalities of natural language and  $\mathbf{N}$  are equal<sup>2</sup>, there is only one infinite limit that is never reached.
5. For any finite limit that is never reached, the probability of natural language having it is  $> 0$ .
6. From 3-5 it follows that the probability of natural language having the infinite limit is 0.

### Arithmetic performance

If recursion were involved in conceptualizing numbers, our brain would execute something like a successor function  $\dots S(S(S(0)))\dots$  for natural numbers and maybe also  $n*n*n*n\dots$  for base- $n$  integer exponents (since we normally use base-10 numeral system,  $n$  would normally equal 10). While  $n*n*n*n\dots$  can be coded and implemented recursively as well as iteratively, it is unlikely that anything approximating  $\dots S(S(S(0)))\dots$  or  $n*n*n*n\dots$  would be run in our brains for conceptualizing numbers and performing arithmetic on them. If it were, our arithmetic performance should be significantly better than it tends to be. If, on the other hand, our inferior arithmetic skills are down to general performance limitations and/or penalties for (other) arithmetic operations, there would be no apparent use for running these procedures for our mathematical capacity. The only remaining justification for  $\dots S(S(S(0)))\dots$  and  $n*n*n*n\dots$  would be recursion in the language faculty. However, for this concession to make sense, there would first have to be some evidence for recursion in the language faculty. As we have argued at length above, at present we have merely conjectures built on invalid premises (see section "Infinity...").

It is easy to demonstrate that conceptualizing a principle (recursion) for producing a pattern of output ( $\mathbf{N}$  or self-

embedding) does not entail (1) that the principle is necessary for producing the pattern (as  $\mathbf{N}$  or self-embedding can be also produced iteratively), and (2) that the principle itself must be neurally implemented for us to be able to conceptualize it. For example, we can conceive that all natural numbers are derived from the number 20098 by  $\pm 1$  operations, i.e. each time we conceptualize a natural number  $x$  that is less than 20098, we subtract 1 from 20098 until we get  $x$  and each time we conceptualize a natural number  $y$  that is greater than 20098, we add 1 from 20098 until we get  $y$ . We can conceive this principle. Does it follow that the "20098  $\pm 1$ " principle must be neurally implemented for us to be able to conceive it in the first place? Surely not. Observe that the situation with the "20098  $\pm 1$ " principle is similar to the recursive one: we can conceptualize the principles but both are at odds with human arithmetic performance. To circumvent the latter problem in the neurally implemented recursion hypothesis, the competence/performance distinction has been called into effect. However, a competence/performance distinction could be also invoked for explaining why our performance is at odds with the "20098  $\pm 1$ " principle. Besides, the distinction raises a non-parsimonious psychological duality as to the conceptualization of relevant syntactic and arithmetic properties – we can conceptualize that the properties are given to us by a recursive principle, but we do not seem to follow the principle neither in linguistic nor arithmetic processing/performance. Furthermore, it seems inconsistent to explain the apparent discontinuity between competence and performance in arithmetic and language by e.g. limitations in primary memory – what potential advantage could a neurally implemented recursive principle bestow if its effects are subject to so severe constraints?

### Conclusion

We conclude with the following points. First, both recursion and iteration allow for finite definitions of infinite sets. Moreover, iterative solutions are frequently less resource demanding than recursive ones (cf. section "Defining recursion"). Second, three logically independent notions of "recursion" are being conflated and confused in linguistics (e.g. Chomsky, 1956, 1971 [1957]; Heine & Kuteva, 2007; Jackendoff & Pinker, 2005): (A) recursive algorithm (Marr's level 2), (B) recursive (or better, inductive) definition (Marr's level 1), and (C) an instance of an item embedded in another instance of the same item. We suggest a terminological way out of the confusion, by reserving 'recursion' for (A) for which there are no alternative terms, and designating (B) and (C) 'induction' and 'self-embedding', respectively. Third, the technical preciseness of the notion of recursion makes it next to impossible to find evidence for it in the brain with the present-day methods, and there is no reason to assume neurally implemented recursion by default (see below). Fourth, contrary to Chomsky (Chomsky, 1995; Hauser et al., 2002) and many others, we argue that a property of natural language is not discrete infinity but physically uncountable finity. Fifth, we

<sup>2</sup> Since we are interested in an upper bound of computation/processing time,  $\aleph_0$  is sufficient. It is difficult to develop the argument here but the very fact that time intervals seem to exist suggests that time is not infinitely divisible ( $\aleph_1$  and beyond). Other indications of this are e.g. Achilles and the tortoise paradox and Planck time.

reject the received opinion, articulated by Chomsky et al. (Chomsky, 2010; Fitch et al., 2005; Hauser et al., 2002), that neurally implemented recursion is necessary to explain natural language and arithmetic competence and performance. The only motivation for neurally implemented recursion is infinity in natural language and arithmetic competence (e.g. the knowledge that one can add 1 to n, append a natural language expression to text or embed clauses indefinitely). We claim that infinity in natural language and arithmetic competence reduces to imagining infinite embedding or concatenation, which is completely independent from an algorithmic capacity for infinite computation, and hence, completely independent from neurally implemented recursion or iteration. In sum, there is no infinity in natural language and arithmetic processing, but even if there were, iteration would be sufficient for generating it.

### Acknowledgments

We thank Noam Chomsky and Margus Niitsoo for extremely helpful, thorough and critical discussions, and Märt Muts, Lameen Souag, Lutz Marten, Tania Kuteva, Michael Corballis, Panu Raatikainen, Tim Gentner, Geoffrey Pullum and the anonymous reviewers for their comments and suggestions. All remaining mistakes are our own. Erkki Luuk was supported by the target-financed theme No. 0180078s08, the National Programme for Estonian Language Technology project "Semantic analysis of simple sentences 2", the Alexander von Humboldt Foundation, and the European Regional Development Fund through the Estonian Center of Excellence in Computer Science, EXCS.

### References

- Abelson, H., Sussman, G. J., & Sussman, J. (1996). *Structure and Interpretation of Computer Programs* (2nd ed.). Cambridge, MA: MIT Press.
- Bickerton, D. (2009). Recursion: core of complexity or artifact of analysis? In T. Givón, & M. Shibatani (Eds.), *Syntactic Complexity: Diachrony, Acquisition, Neuro-cognition, Evolution*. Amsterdam: John Benjamins.
- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2, 113-124.
- Chomsky, N. (1959). On certain formal properties of grammars. *Information and Control*, 2, 137-167.
- Chomsky, N. (1971 [1957]). *Syntactic Structures*. The Hague: Mouton.
- Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.
- Chomsky, N. (2010). Some simple evo-devo theses: how true might they be for language? In R. K. Larson, H. Yamakido, & V. Deprez (Eds.), *Evolution of Human Language: Biolinguistic Perspectives*. Cambridge: Cambridge University Press.
- Fitch, W. T. (2010). Three meanings of "recursion": key distinctions for biolinguistics. In R. K. Larson, V. Deprez, & H. Yamakido (Eds.), *The Evolution of Human Language: Biolinguistic Perspectives*. Cambridge: Cambridge University Press.
- Fitch, W. T., Hauser, M. D., & Chomsky, N. (2005). The evolution of the language faculty: clarifications and implications. *Cognition*, 97(2), 179-210; Discussion 211-125.
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, 298(5598), 1569-1579.
- Heine, B., & Kuteva, T. (2007). *The genesis of grammar: a reconstruction*. New York: Oxford University Press.
- Jackendoff, R., & Pinker, S. (2005). The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky). *Cognition*, 97(2), 211-225.
- Krauss, L. M., & Starkman, G. D. (2000). Life, the universe, and nothing: Life and death in an ever-expanding universe. *The Astrophysical Journal*, 531, 22-30.
- Lieberman, P. (2008). Cortico-striatal-cortical neural circuits, reiteration, and the "narrow language faculty". *Behavioral and Brain Sciences*, 31, 527-528.
- Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman and Company.
- Minsky, M. (1972). *Computation: Finite and Infinite Machines*. London: Prentice-Hall International.
- Moore, A. W. (1990). *The Infinite*. London: Routledge.
- Odifreddi, P. (1992). *Classical Recursion Theory: The Theory of Functions and Sets of Natural Numbers* (Vol. 125). Amsterdam: Elsevier.
- Pinker, S., & Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition*, 95(2), 201-236.
- Premack, D. (2007). Human and animal cognition: Continuity and discontinuity. *Proceedings of the National Academy of Sciences of the United States of America*, 104(35), 13861-13867.
- Rogers Jr., H. (1987). *The Theory of Recursive Functions and Effective Computability*. Cambridge, MA: MIT Press.
- Shannon, C. E., & Weaver, W. (1964 [1949]). *The Mathematical Theory of Communication*. Urbana: The University of Illinois Press.
- Sipser, M. (1997). *Introduction to the Theory of Computation*. Boston, MA: PWS Publishing Company.
- Tomalin, M. (2011). Syntactic structures and recursive devices: a legacy of imprecision. *Journal of Logic, Language and Information*, 20(3), 297-315.
- Weisstein, E. W. (2003). *CRC Concise Encyclopedia of Mathematics* (2nd ed.). Boca Raton: Chapman & Hall/CRC.
- von Humboldt, W. (1999 [1836]). The diversity of human language-structure and its influence on the mental development of mankind, *Wilhelm von Humboldt: On Language*. Cambridge: Cambridge University Press.

# Modelling the IAT: Implicit Association Test Reflects Shallow Linguistic Environment and not Deep Personal Attitudes

**Dermot Lynott (dermot.lynott@manchester.ac.uk),**

**Himanshu Kansal (himanshu.kansal@postgrad.mbs.ac.uk)**

Decision and Cognitive Sciences Research Centre, Manchester Business School, University of Manchester  
Booth Street West, Manchester M15 6PB, UK

**Louise Connell (louise.connell@manchester.ac.uk)**

School of Psychological Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, UK

**Kerry O'Brien (kerry.o'brien@monash.edu)**

Behavioural Studies, Monash University, PO Box 197, Caulfield East VIC 3145, AUSTRALIA  
and School of Psychological Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, UK.

## Abstract

People often have thoughts, attitudes and biases that are not themselves consciously aware of or that they would rather not share with others. To assess such attitudes, researchers use paradigms like the Implicit Association Test (IAT) that do not rely on explicit responding to determine the level of bias a person holds towards a particular target concept (e.g., race, gender, age). Responses in the IAT are assumed to reflect deeply held beliefs and attitudes, and not shallow, superficial associations. However, as linguistic distributional information has been shown to serve as a viable heuristic in many cognitive tasks, we investigated whether it could be used to predict the level of bias established by the IAT. We used a large corpus of language (Web 1T) and data from 16 IAT studies ( $N = 1825$ ) to examine whether the degree of linguistic co-occurrence for target concepts and attributes reflected the size of bias observed in human behavioural data. We found that the effect size of the linguistic biases corresponded strongly with the effect sizes from the behavioural data. We suggest that language reflects prevalent cultural attitudes which are captured by tasks such as the IAT, suggesting that the IAT may reflect shallow, linguistic associations rather than deeper conceptual processing.

**Keywords:** linguistic distributional information; implicit association test; IAT; attitudes; model.

## Introduction

If we openly asked people questions like "are you sexist" or "are you racist", we would probably expect people to be reluctant to respond, if we got any response at all. When asking for judgements on controversial topics and divisive issues, people have a strong desire to provide socially acceptable responses that may be contrary to their true beliefs (e.g., Furnham, 1986; Paulus, 1991). As such, there is often a disconnect between what people say and what they do. In order to avoid tasks that require explicitly thinking about a particular issue or that permit strategic responding by participants, researchers in social cognition have instead developed paradigms that try to tap into people's attitudes in a more implicit manner (Fazio, Jackson, Dunton & Williams, 1995; Greenwald, McGee & Schwartz, 1998). The most frequently used of these paradigms is the

Implicit Association Test, or IAT. The IAT is essentially a categorisation task, similar to many priming paradigms used across the cognitive sciences, designed to capture the degree of bias or prejudice that an individual has towards a particular concept. (e.g, race, age). We describe the task in more detail below.

A search using Google Scholar reveals that the IAT is referenced in over 4000 papers in the last 10 years alone. In spite of its widespread use, there is ongoing disagreement regarding what the IAT is actually measuring (Blanton et al., 2009; Fazio & Olsen, 2003; Greenwald, Poehlman, Uhlman & Banaji, 2009). Nonetheless, the creators and most proponents of the paradigm maintain that "the IAT assesses the strengths of associations between concepts" (p18., Greenwald, Poehlman, Uhlman & Banaji, 2009) and it is assumed to reflect deep, underlying, unconscious biases. However, when one situates the IAT within the broader context of cognitive research examining the structure of the conceptual system, such a claim is ambiguous.

Several researchers have described the conceptual system as comprising two distinct but interrelated components; a linguistic system and a simulation system (Barsalou, Santos, Simmons & Wilson, 2008; Connell & Lynott, 2011; 2012; Louwerse & Jeunieux, 2008). The linguistic system reflects language usage, and captures the distributional patterns (or statistical regularities) of words and phrases, making this system best suited for "quick and dirty" heuristic processing (Lynott & Connell, 2010). The simulation system, on the other hand, captures perceptual, affective and motor information from our environmental experience and is better suited to deep, slow, precise processing. Thus, performance in the IAT may reflect responses from one of these two systems, raising two alternative hypotheses. The first, is that the IAT indeed reflects personal attitudes emerging from deep-rooted, affective and conceptual processing in the simulation system. For example, an intelligence/obesity prejudice (O'Brien et al, 2007) would take the form of conceptual retrieval of a "fat person" automatically evoking associated concepts of "stupidity" and a negatively valenced affective association of "badness". This perspective is summarised by Nosek, Banaji and Greenwald (p112., 2002)

who argue that implicit attitudes "reveal the *deep* influence of the immediate environment and the broader culture on internalized preferences and beliefs".

The second option is that IAT scores reflect much shallower processing of the socio-cultural environment, specifically the token-to-token statistical patterns of the linguistic system. For example, people may often encounter the word "fat" in close proximity to the word "stupid" in conversations they hear or in texts they read, resulting in the automatic activation of the word "stupid" every time the word "fat" is encountered. Importantly, activating a word like "stupid" does not require full conceptual retrieval (Louwerse & Connell, 2011). Rather, linguistic associations like these operate at a shallow, superficial level that can produce a response to a given task without recourse to deeper conceptual or affective processing.

There is good reason to believe that the IAT may reflect the latter shallow linguistic associations rather than the former deeper, affective, conceptual attitudes. Several studies have shown that the linguistic system is used as a shortcut to provide a "good enough" response to conceptual tasks, whenever possible (e.g., Connell & Lynott, 2012; Louwerse & Connell, 2011). In particular, when processing demands are shallow and the participant is placed under time pressure the linguistic system provides a useful heuristic for responding without recourse to the greater computational expense of full, perceptual, affective and motor simulation of the concept. For example, conceptual tasks such as property-verification (e.g., making true/false judgments regarding object properties - *apple can be green*) can be successfully completed solely on the basis of the word-to-word associations of "apple" and "green"; these words frequently appear in close proximity and therefore it is a reasonable heuristic to assume that this property belongs to this concept. When participants respond quickly (Louwerse & Connell, 2011) or when the set of items is poorly constructed (Solomon & Barsalou, 2004) their responses are based on these linguistic associations and not on deeper conceptual representations. For example, using response time data from a property-verification task, Louwerse and Connell (2011) demonstrated that measures of distributional patterns from the linguistic system could be used to predict the faster responses of participants, but not their slower responses. Conversely, measures of the simulation system could predict slower responses, but not faster responses. Evidence that the IAT does not engage deeper processing is provided by a recent study by Foroni and Semin (2012). Foroni and Semin had two groups complete the IAT; one group completed the task as normal, while the other completed the task with facial feedback being inhibited by holding a pen between the lips during the task. Holding the pen in this position leads to sustained activation of the zygomaticus major muscle (used in frowning), which is normally activated following the presentation of a valenced stimulus. However, inhibition of this muscle during the IAT made no difference to the level of bias observed. This suggests that IAT does not engage

the affective system in a way that would be expected if the task required processing in the simulation system.

Given that the linguistic system is capable of providing quick and dirty responses in a variety of seemingly complex tasks and given that responses in the IAT may be of a superficial nature (i.e., not requiring the deeper processing of the simulation system), we considered whether IAT biases could be predicted by the statistical distributional associations in language. While the linguistic associations and simulation systems are closely related, they are not exact replications of each other because each system gains experience from a different source. Just because two words share a linguistic association in the socio-cultural environment, because they are sometimes juxtaposed, it does not mean that their referent concepts are tightly bound in a personal, affective/conceptual attitude. If IAT responses are predicted by linguistic associations then it suggests that the IAT itself is a shallow measure of the language structure to which an individual has been exposed and not necessarily a reflection of deeper biases. We describe below the IAT paradigm in more detail before outlining the current study.

Condition	Word belongs to...?	"stupid"	"smart"
Congruent	Bad OR Fat	fast	[incorrect]
	Good OR Thin	[incorrect]	fast
Incongruent	Bad OR Thin	slow	[incorrect]
	Good OR Fat	[incorrect]	slow

Table 1: Schematic of response patterns in an IAT on obesity prejudice. The first column describes category as congruent or incongruent pairings. The second column indicates the two judgements participants must consider for each target word. Third and fourth columns describe the predicted patterns for two target concepts "stupid" and "smart". "Incorrect" indicates a wrong answer (e.g., "stupid" should not be "good" or "thin")

### The Implicit Association Test

The IAT represents one of the most frequently used paradigms for examining implicit attitudes (e.g., Greenwald, McGee & Schwartz, 1998), with hundreds of studies already published using this approach (see e.g., Greenwald, Poehlman, Uhlman & Banaji, 2009). The IAT is used to give an insight into people's automatically activated biases and prejudices and is designed to overcome the issues of strategising and socially desirable responding by participants. The IAT achieves this by requiring extremely rapid and accurate responses from participants to tap into automatic associations between some target concept and an attribute. For example, O'Brien and colleagues (O'Brien, Hunter & Banks, 2007) examined people's anti-fat prejudices using the IAT to see whether people associated obesity with negative concepts like stupidity. The IAT contrasts performance for a *congruent* pairing of targets and attributes (e.g., obesity-bad; thinness-good) to an



*incongruent* pairing of targets and attributes (e.g., obesity-good; thinness-bad). The participant's task is to categorise target stimuli as they appear on screen using one of these two pairings. In a congruent block, if the word "stupid" appeared onscreen, the participant would press the key indicating they belonged to the "fat OR bad" category, while if the word "smart" appeared they would press the key to categorise it as belonging to the "thin OR good" category. In this way, each target attribute has an identifiably correct response. Table 1 presents a schematic of the responses in both congruent and incongruent conditions. Every participant completes both categorisation pairings in a counterbalanced fashion.

The key question is, which pairing do participants respond most quickly to. Once all responses have been made, a bias score can be calculated for each participant and then an overall bias can be calculated for the entire sample. If participants are generally faster in their responses for the congruent condition in the obesity IAT example above, this would indicate a negative bias towards obesity related concepts. In analysing response times, the IAT scoring algorithm calculates the difference in average response latency between the congruent and incongruent conditions and dividing by the standard deviation of all latencies for both conditions (Nosek, Greenwald & Banaji, 2007). For paper-based versions of the IAT the difference is calculated based on the number of correct responses in a 20 second period in the congruent/incongruent conditions (e.g., O'Brien et al, 2007). If there is a large difference between the categorisation conditions in terms of response times or number of correct responses, this will result in a larger bias score. The difference between congruent and incongruent response times or accuracy reflects the extent to which people believe that fat people are stupid and thin people are smart. The IAT is thus assumed to offer a window into deeply-rooted beliefs and prejudices that are otherwise difficult to impossible to access explicitly.

The IAT has been used to uncover and measure biases in a wide range of domains, such as attitudes towards race (Greenwald, McGee & Schwartz, 1998), gender stereotyping (Rudman & Kilianski, 2000), alcohol (Wiers et al, 2011) and doping among athletes (Brand et al, 2011) to name but a few examples. What's more, the IAT has been shown to be predictive of people's overt behaviors, underscoring its practical utility. In a meta-analysis of 156 studies, Greenwald and colleagues (Greenwald, Poehlman, Uhlmann, & Banaji, 2009) found that IAT measures correlated significantly with explicit measures of behaviour. In some cases, IAT scores are better predictors of behaviour than more explicit measures. For example, Asendorpf, Banse, and Mücke (2002) found that shyness in individuals was better predicted by a shyness-oriented IAT than by explicit self-ratings of shyness. In some ways the IAT seems to capture attitudes and beliefs that we hold, but that which find it difficult to consciously and explicitly access ourselves.

## The Current Study

We have described the IAT paradigm and suggested how it may draw on processing from either the shallow linguistic system or be reliant on deeper processing from the simulation system. To examine whether IAT performance is predicted by linguistic associations we used behavioural data from several published IAT studies and linguistic data extracted from the World Wide Web, using the Web 1T 5-gram corpus (Brants & Franz, 2006). The Web 1T is a snapshot of web pages indexed by Google in 2006 and contains over 1 trillion tokens, making it one of the most representative corpora of language available. Our aim was to examine whether co-occurrence patterns in the linguistic data could predict the effect sizes observed in the behavioural data. We expected that if implicit attitudes (as captured by the IAT) reflect the distributional patterns of language in the linguistic system, then we should see a significant fit between the two sets of data using regression analyses. On the other hand, if the IAT relies on deeper processing in the simulation system, we would not expect such a relationship to exist.

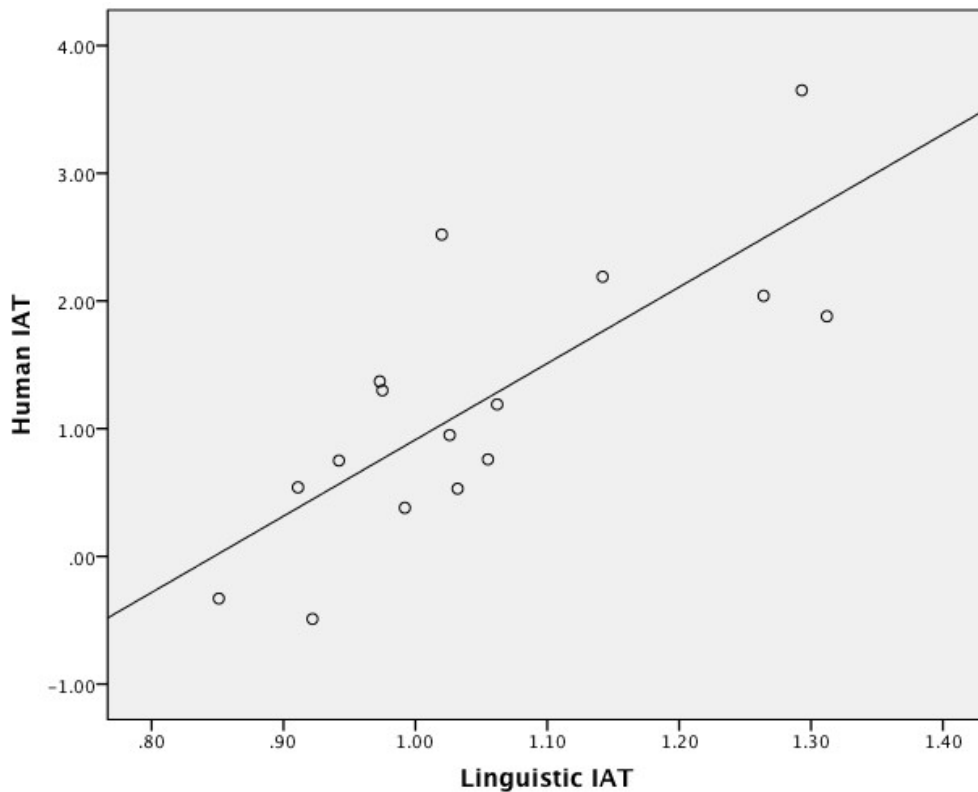
IAT Topic	Reference	N
Race (2)	Rudman & Ashmore, 2007	128
Flowers vs Insects	Greenwald et al., 1998	32
Instrument vs Weapon	Greenwald et al., 1998	32
Japanese vs Korean (2)	Greenwald et al., 1998	64
Alcohol	Wiers et al., 2011	108
Doping (2)	Brand et al., 2011	102
Alcohol and Sport	O'Brien & Lynott, in prep	120
Obesity (3)	O'Brien et al, 2007	1032
Gender (3)	Rudman & Kilianski, 2000	207

Table 2: Sources of the 16 Implicit Association Tests used in the study, with number of studies from each source in brackets, and the total N participants for each source.

## Method

**Materials** We selected 16 IATs from journal articles that used lexical (as opposed to pictorial) stimuli and provided a full list of materials used. Table 2 lists the topics of investigation for each of the IATs, the source reference and the total sample sizes from the original studies; 15 of the studies come from published journal articles and one from a paper in preparation (O'Brien & Lynott, in prep). The IATs cover a broad range of topic areas, including studies of racial stereotypes, obesity and gender roles. Each IAT consists of a set of category concepts (e.g., male, female, good, bad) and a set of target attributes that are of positive or negative valence (e.g., reliable, pleasant, terrible, nasty). The number of attributes used by the IATs ranged from 3 to 25, with a total of 324 target stimuli.

Figure 1: Scatterplot, with line of best fit, for IAT scores from human behavioural data plotted against scores derived from log-transformed linguistic data.



**Human Behavioural Data** The 16 IATs represent data from 1825 participants. From each IAT we extracted the overall effect size of the bias found based on the human responses ( $D_H$ ). The IAT effect size is closely related to Cohen's  $d$ , a popular measure of statistical effect size.

**Linguistic Model** The aim of the model is to calculate the size of the linguistic bias ( $D_L$ ) for each IAT, based on the specific terms used in each task. In order to approximate the linguistic distributional information, we carried out a corpus analysis using the Web 1T 5-gram corpus. Using the corpus we are able to calculate the difference in the strength of associations for each categorisation condition in each IAT. For example, for an IAT examining preference for flowers versus insects, we calculate the level of association between one categorisation pairing (e.g., co-occurrences for "flower" and all positive attributes + co-occurrences for "insect" and all negative attributes) and the inverse, "incongruent" categorisation pairing (co-occurrences for "flower" and all negative attributes + co-occurrences for "insect" and all positive attributes).

The strength of association is calculated by summing the frequency of co-occurrences of category terms (e.g., flower, insect) with each of the attribute terms (e.g., nice, horrible, etc.). For each category-attribute pairing (e.g., flower-nice), we calculated the cumulative 5-gram frequency of forward and backward co-occurrences between the category word and attribute (i.e., the summed count of occurrences of

[flower ... nice] and [nice ... flower] in the corpus with zero, one, two and three intervening words: for a similar approach, see Louwerse & Connell, 2011). Because of the large number of calculations this required (>10,000), we developed a semi-automated tool to take each set of IAT terms and output the collocation frequencies from the Web 1T corpus. Using these summed frequencies we then calculated a linguistic effect size using two models; one model based on raw frequency counts and one model based on the log-transformed frequencies (using the natural log). Ji (2010) discusses the improvement in the distribution curves of datasets of seven sub-corpora after having undergone log transformation as the transformation mitigates the effects of extreme values (i.e., highly frequent terms). Thus, both linguistic models represent the ratio of the frequencies for the congruent categorisation condition compared to the frequencies of the incongruent categorisation condition. Finally, we conducted separate linear regression analyses for the two linguistic models using the linguistic bias ( $D_L$ ) as a predictor variable and the behavioural bias ( $D_H$ ) as the dependent variable.

## Results & Discussion

The effect sizes in the human data ranged from -.49 to 3.65 ( $M = 1.2$ ,  $SD = 1.07$ ), while the effect sizes in the linguistic data ranged from .2 to 7.3 ( $M = 2.02$ ,  $SD = 2.1$ ) for the raw frequency model and from .85 to 1.31 ( $M = 1.05$ ,  $SD = .14$ )

for the log-transformed model. Using linear regression analyses we found significant relationships between the effect sizes calculated from the human data,  $D_H$ , and the effect sizes calculated from the linguistic models,  $D_L$ . This positive relationship indicates that the larger the effect in the linguistic data, the larger the predicted effect in the human data. Figure 1 illustrates the relationship between the biases predicted from the log ratio linguistic model and those derived from the human data. The regression model for the raw frequency ratio model was significant ( $r^2 = .612, p < .001, n = 16$ ) resulting in a  $\beta$ -coefficient of .782 for the linguistic predictor ( $t = 4.696, p < .001$ ). The regression model for the log frequency ratio model was also significant ( $r^2 = .596, p < .001, n = 16$ ) resulting in a  $\beta$ -coefficient of .772 for the linguistic predictor ( $t = 4.544, p < .001$ ). This indicates that both models reflect approximately 60% of the variance in the human IAT scores.

## General Discussion

This study investigated whether linguistic distributional information can be used to predict levels of implicit attitudes as measured by the Implicit Association Test. We observed significant relationships between effect sizes from human behavioural data and effect sizes calculated from linguistic distributional data. This finding suggests that performance on the IAT may not reflect the deeply-rooted biases and beliefs held by individuals and groups, but instead reflects shallower linguistic associations they have encountered in their environment. While the present results are promising, there are of course some caveats we need to be aware of. We discuss some of these limitations and avenues for future research below.

An obvious question to ask is, if the IAT reflects only shallow linguistic processing, then how is IAT performance predicting overt behaviours? There are two issues here. The first is that the while IAT is successful in predicting outcomes in certain sub-domains (e.g., political preferences), it poorly reflects outcomes in others (e.g., sexual orientation; Greenwald et al., 2009). The second is that even where IAT performance is claimed to predict other behavioural outcomes, the claims may not stand up to closer scrutiny. It is important when comparing overt measures to IAT performance that one uses implicit tasks that have clear behavioural outcomes. Good examples of this are patient treatments and using realistic CVs/resumés for assessing job candidates. Green et al (2007) found that doctors' levels of implicit bias towards black patients did not always tally with their decision to offer treatment using thrombolysis to remove blood clots. Although doctors with higher anti-black bias were less likely to treat black patients than white patients using thrombolysis, those doctors with low anti-black bias (but not a *pro*-black bias) were actually more likely to treat black patients than white patients. In re-analysing data looking at racially discriminating behaviour in job candidate selection, Blanton and colleagues (2009) found that the IAT failed to predict any discriminatory behaviour when factors such as rater reliability and outlier removal were taken into account. In one case, previous

evidence of anti-black prejudice was actually reversed revealing a pattern of pro-black bias. However, the process of candidate evaluations can be broken down further. For example, selection and shortlisting of candidates can be viewed as more of a heuristic process, while providing specific grades to individual candidates can be seen as more deliberative. Blommaert, van Tubergen and Coenders (2011) distinguished between these two aspects of the candidate assessment process while examining the effects of implicit attitudes towards ethnicity and gender. They found that only explicit measures predicted people's grading of candidates, but that both implicit (IAT) and explicit measures predicted people's shortlisting of candidates. Thus, it is important to take into account both the task domain and the nature of the task (i.e., heuristic or deliberative) to have a clearer idea of whether the IAT and therefore linguistic distributional information may have a role to play in predicting behavioural outcomes.

A limitation of the current approach is that it does not discriminate between different groups and different task contexts and how this would affect performance on a given IAT. For example, we would not necessarily expect a group of American students and a group of Chinese students to show the same type of bias judging American and Chinese faces paired with positive or negative attributes. One possibility would be to extend the model to incorporate additional domain terms that would calculate co-occurrence frequencies but limited to specific contexts to attempt to approximate these contextual effects.

Although it may be argued that language reflects our cultural beliefs and social norms, it is difficult to establish a causal relationship between implicit attitudes and the linguistic data. It may be that the linguistic distributional information is a) driving the formation of these biases, b) is the behavioural outcome of these biases or c) part of a self-sustaining cycle of biases influencing language influencing biases and so on. However, it is clear that exposure to socio-cultural attitudes does impact our own attitudes and behaviours, as developmental changes are evident in IAT performance. For example, older subjects tend to show larger IAT effects than younger subjects (Hummert, Garstka, O'Brien, Greenwald, & Mellott, 2002). As language is one of the key methods for the transmission of socio-cultural information, this underscores the possible role for language exposure in the formation of implicit attitudes (Nosek, Greenwald & Banaji, 2007).

In conclusion, we present the first model of implicit attitudes based on linguistic data extracted from the world wide web. We found that linguistic models revealed a strong correspondence with human behavioural data. We see language as a primary means for transmission of attitudinal information and agree with Uhlman and colleagues (in press) that "implicit attitudes reveal the power of cultures to reproduce themselves in individual minds". However, our findings also suggests that such implicit attitudes may not represent deeply rooted beliefs as has previously been assumed. Ongoing work is exploring the predictive power of the model by using linguistic data to predict attitude

effect sizes in advance of behavioural studies, providing a strong test whether our hidden beliefs can be revealed in the patterns of our language use.

## References

- Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology*, 83, 380–393.
- Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, & A. C. Graesser, A. (Eds.), *Symbols, embodiment, and meaning*. Oxford, UK: OUP.
- Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009). Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT. *Journal of Applied Psychology*, 94, 567–582.
- Blommaert, L., van Tubergen, F., & Coenders, M. (2011). Implicit and explicit interethnic attitudes and ethnic discrimination in hiring. *Social Science Research*, 41, 61–73.
- Brand, R., Melzer, M. & Hagemann, N. (2011). Towards an implicit association test (IAT) for measuring doping attitudes in sports. Data-based recommendations developed from two recently published tests. *Psychology of Sport and Exercise*, 12, 250 – 256.
- Brants, T., & Franz, A. (2006). *Web 1T 5-gram Version 1*. Philadelphia: Linguistic Data Consortium.
- Connell, L., & Lynott, D. (2011). Modality switching costs emerge in concept creation as well as retrieval. *Cognitive Science*, 35, 763–778.
- Connell, L., & Lynott, D. (2012). Principles of Representation: Why You Can't Represent the Same Concept Twice. Under Review.
- Fazio, R. H. & Olsen, M. A. (2003). Implicit measures in social cognition research: Their Meaning and Use. *Annual Reviews in Psychology*, 54, 297–327
- Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, 69, 1013–1027
- Foroni, F. & Semin, G. R. (2012). Not all implicit measures of attitudes are created equal: Evidence from an embodiment perspective. *Journal of Experimental Social Psychology*, 48, 424–427
- Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, 7, 385–400.
- Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., ... (2007). Implicit Bias among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients. *Society of General Internal Medicine*, 22, 1231–1238.
- Greenwald, A.G., McGhee, D.E., & Schwartz, J.L.K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Greenwald, A.G., Poehlman, T.A., Uhlmann, E.L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97, 17–41.
- Hummert, M. L., Garstka, T. A., O'Brien, L. T., Greenwald, A. G., & Mellott, D. S. (2002). Using the implicit association test to measure age differences in implicit social cognitions. *Psychology and Aging*, 17, 482–495.
- Ji, M. (2010). A corpus-based study of lexical periodization in historical Chinese. *Literary and Linguistic Computing*, 25, 199 – 213.
- Louwerse, M. M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, 35, 381–398.
- Louwerse, M. M., & Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. In M. de Vega, A. Glenberg, & A. C. Graesser (Eds.), *Symbols, embodiment, and meaning*. OUP.
- Louwerse, M.M. & Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. *Cognition*, 114, 96–104.
- Lynott, D., & Connell, L. (2010). Embodied conceptual combination. *Frontiers in Psychology*, 1:216, 1–14.
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting Implicit Group Attitudes and Beliefs From a Demonstration Web Site. *Group Dynamics: Theory, Research, and Practice*, 6, 101–115.
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes* (pp. 265–292). New York: Psychology Press.
- O'Brien, K. S., Hunter, J. A., & Banks, M. (2007). Implicit anti-fat bias in physical educators: Physical attributes, ideology, and socialisation. *International Journal of Obesity*, 31, 308– 314.
- O'Brien, K. S., Hunter, J. A., Halberstadt J., & Anderson J. (2007). Body image and explicit and implicit anti-fat attitudes: the mediating role of physical appearance comparisons. *Body Image*, 4, 249–256.
- Paulhus, D. L. (1991). Measurement and control of response bias. In: Robinson JP, Shaver PR, Wrightsman LS (eds). *Measures of Personality and Social Psychological Attitudes*. Academic Press: New York, 1991, pp 17–59.
- Rudman, L.A. & Kilianski, S.E. (2000). Implicit and Explicit Attitudes Toward Female Authority. *Personality and Social Psychology Bulletin*, 26, 1315 – 1328.
- Solomon, K.O., & Barsalou, L.W. (2004). Perceptual simulation in property verification. *Memory & Cognition*, 32, 244–259.
- Uhlmann, E.L., Poehlman, T.A., & Nosek, B. A. (in press). Automatic Associations: Personal Attitudes or Cultural Knowledge? To appear in Jon D. Hanson (Ed.) *Ideology, Psychology, and Law*. New York, NY: OUP.
- Wiers, R.W., Eberl, C., Rinck, M., Becker, E.S. & Linenmeyer, J. (2011). Retraining Automatic Action Tendencies Changes Alcoholic Patients' Approach Bias for Alcohol and Improves Treatment Outcome. *Psychological Science*, 22, 490 – 497.

# External Working Memory and the Amount of Distributed Cognition

Naoki Maeda

Teachers College, Columbia University, 525 W. 120<sup>th</sup> Street  
New York, NY 10027 USA

## Abstract

While processing of a large number of (visuo-spatial) items are oftentimes necessary for ongoing cognitive activities, the biological working memory can process only about four items of information. Then it is a mystery how we cope with complex world situations. This is the paradox of working memory. This paradox is solved once we view the external features of the world as constituting part of working memory. Part of working memory is externally distributed if the external features of the world constitute part of material supervenience base of working memory. Tversky's Spractions (2010), or actions onto the world, are the key to offload of cognition, because they redirect the attention at the working memory level only to relevant aspects of the world. To see how people use spractions to offload working memory load, subjects were asked to build a Lego block in front of a camera. Using cognitive ethnography, it was observed that they all relied on spractions to cognize. From the fact that the biological working memory can process only about four items of information, the amount of working memory based distributed cognition can be calculated.

**Keywords:** paradox of working memory; external working memory; spractions

## Introduction

Working memory (hereafter WM) is a limited capacity system to temporarily maintain, access, and update information necessary for ongoing cognitive activities (Baddeley, 2000, 2003; Awh et al., 2006; Jonides et al., 2008). Traditionally it is conceived of as short-term memory (STM) buffer, characterized by the firing of neurons; it can hold information for a couple of seconds. Since STM does not involve structural change of neural networks, information stored in STM is transient. STM, and hence WM, is one of the core components of cognition. Hardly any cognitive task can be completed without the involvement of STM. For example, when you add some numbers, you have to create a temporal mental representation for those numbers. Biological WM is a limited capacity system; it can maintain, access, and update only limited amount of information simultaneously. In his seminal paper "*The magical number seven, plus or minus two*," George Miller (1956) argued that the capacity of WM is limited to about seven items of information. However, as later pointed out, Miller's magical number seven was inflated due to the confound effect of linguistic chunking, a strategy to group small items of information into an integrated representation (discussed more below). According to a more accurate estimate (Sperling, 1960; Landman et al., 2003; Cowan, 2001; Jonides et al., 2008; Hauser et al., 2000; Block, 2007), the capacity of biological

(visuo-spatial) WM of human adults is limited to about four items of information. That is, WM can selectively attend to only about four items of information simultaneously.

While the capacity of biological WM is limited to about four items of information, the world around us is full of complexity, rich in detail, and oftentimes cluttered (i.e. there are usually more than four items of information in the world, and they are oftentimes relevant to ongoing cognitive activities). Thus, there is an overflow of information at the WM level (Kessell and Tversky, 2010; c.f. Rowlands, 1999; Block, 2007, 2011). Although we have to cognize quickly in response to stimuli in the world to survive (c.f. Cruse, 2003; Kirsh and Maglio, 1994), given the complexity of the world and the limited WM capacity, it is not at all clear how we can do so. Nevertheless, we are almost entirely unaware of the limitation of biological WM in daily life except for some minor occasions such as remembering a telephone number for the first time, and cope with high cognitive tasks day by day. It is a mystery how a severely constrained WM can cope with the complexity of the world. I call this the *paradox of working memory*.

A traditional strategy to overcome this paradox is the aforementioned chunking. Chunking is a way to enlarge a representational unit of attention so that more items of information can be processed with the same WM capacity. For example, although a random sequence of alpha-numerical letters are difficult to remember and process due to the limited capacity of biological WM, once they are chunked into a meaningful sequence, they can be remembered easily. Thus, a seemingly meaningless sequence of "CIAUCLAKGB" can be remembered easily by means of chunking; they are chunked into "CIA," "UCLA," and "KGB". Chunking thus enlarges the conceptual unit of attention by means of LTM. Although linguistic chunking is well known, it is not the only chunking. Information can be *visually* chunked if visual information is stable over time (Magnuson et al., 1998). Such expanded STM capacity with the assistance of non-STM, such as LTM, is called *compound STM capacity* (Cowan, 2001). Visuo-spatial information can be chunked both verbally and visually. Chunking itself is a partial solution to the paradox of working memory.

The question is "is chunking sufficient to overcome the limitation of WM?" It does not seem so, first because the world contains a lot of visual complexity that cannot be verbally chunked and second because the world is not stable enough to enable visual chunks to be formed. This can be seen in change blindness (Simons et al., 2005). Change blindness refers to the surprising difficulty observers have in detecting a significant change to their visual field. In the

experimental setting, usually two different pictures flicker quickly. Although there is a significant difference between them, a large portion of subjects, to their surprise, does not detect the change. In the change blindness cases, the reason why participants cannot detect the change is precisely due to the instability of the scene. And they do not seem to linguistically chunk the visual stimuli. If they could chunk visual information perfectly in order to account for every visual stimulus into four items of information, they would detect every single change. Although whether this is a case of “blindness” characterized by lack of experience is a controversial issue (Block, 2007, 2011), it is relatively clear that there is no cognitive access to the change (because we are not “aware” of the change). Thus, it seems that chunking alone does not solve the paradox of working memory. We are then brought back to the paradox.

### External Working Memory and Spractions

How can we solve the paradox, then? I argue that the solution to the alleged paradox is to view the external world – space, gesture, body, action, and so on – as constitutive part of material supervenience base of WM. Once we view the human agent and its immediate surrounding environment as a coupled cognitive system (Clark and Chalmers, 1998), the external features of the world *in the coupled system* can be regarded as constituting part of material supervenience base of WM (Rowlands, 1999). The external features of the world can temporarily maintain, allow for access, and update information necessary for ongoing cognitive activities. As a consequence, the external features of the world can functionally augment the limited capacity of biological WM. Because of the external features of the world, I argue, we can cope with complex real world situations, even if our biological WM capacity is severely limited. To cognize efficiently, in other words, we are naturally exploiting the external features of the world as a material supervenience base of WM function. There is no problem with chunking per se; I am proposing that it is only a partial solution to the problem. Once theoretically conjoined with the external WM, they together solve the paradox of working memory.

Biological WM has been believed to be augmented by LTM, the external features of the world, and such; this functional whole has been called *compound WM* (Cowan, 2001). Under the concept of compound WM, however, components other than biological WM, such as LTM and the external features of the world, are considered mere *causal part* of WM. That is, while the external features of the world, in which we are embedded, are important aids to WM, according to this view, they are not themselves constituents of WM (c.f. Rupert, 2004). I argue that components other than biological WM, especially the external features of the world, are indeed *constitutive part* of WM. The coupled system of biological WM and the external features of the world together constitutes functional WM; WM is actually extended into the external world (c.f. Hutchins, 1995).

Although it might seem trivial, the difference between ‘causal’ and ‘constitutive’ is important. Roughly stated, constitutive part of something is part of what it is to be that something, while causal part of something is not. Block (2007) illustrates this point as follows; “cerebral blood flow is *causally necessary* for consciousness, but activation of the upper brainstem is much more plausibly a *constitutive condition*, part of what it is to be conscious” (p.482; emphasis mine). The distinction of causal/constitutive in cognitive science is captured by the debate between extended cognition and embedded cognition (Rupert, 2004; also discussed in Clark, 2008). The hypothesis of extended cognition (dubbed as HEC in the literature) asserts that the external features of the world constitute part of material supervenience base of cognition; the hypothesis of embedded cognition (dubbed as HEMC in the literature) holds that the external features of the world are causally relevant to cognition but do not themselves constitute part of cognition.

	External WM	Compound WM
External World is ...	Constitutive part of WM	Causal part of WM
Hypothesis of ...	Extended cognition (HEC)	Embedded cognition (HEMC)

Table 1. The conceptual difference between external WM and compound WM.

The original idea of external WM comes from Rowlands (1999). Using George Miller, Rowlands argues that biological WM is enormously limited, unstable, and unreliable so the main locus of WM should actually be external information-bearing structures. Challenging Rowlands, Rupert (2004) claims that external WM is not plausible, while compound WM is, based on the fact that the nature of the contributions of the biological WM and external features of the world are profoundly different. As Clark (2008; also Clark and Chalmers, 1998) argues, however, externalism does not demand fine-grained functional similarities of the inner and outer contributions. While precisely how WM is offloaded is debated (Gray et al., 2004; Gray et al., 2006), the general upshot, then, is that WM is externally distributed if the external features of the world constitute part of material supervenience base of WM, even if functional similarities are not fine-grained.

According to computational cognitive science, the basic function of cognition is largely accounted for by two main factors; computation and representation (c.f. Horst, 2011). The concept of computation and representation naturally applies to WM as well. By means of representation and computation, WM can store, update, and access information necessary for ongoing cognitive activities. Then, there are two ways how we externalize WM; by externalizing computation and by externalizing representation. Having said so, it is important to note that computational function and representational function do overlap (McClelland et al., 1986; Clark, 1989, 1993) so that the distinction is merely ideal-typical.

We seem to be naturally offloading complex computation onto the external world if it is an available option (Gray et al., 2004; Gray et al., 2006; Wilson, 1994). For example, although we can rotate objects mentally (Shepard et al., 1971), it is more efficient (faster and more accurate) to do so physically. The ubiquity of physical rotation as computational action is found by Kirsh and Maglio (1994). It seems that we use external computation when WM load is heavy (Kirsh et al., 1994). Also, we offload WM representational function by exploiting the stability of the world. That is, by leaving information in the external world, we reduce the WM load the biological WM has to process, as the external world is too complex to process in the biological WM. In a way, we use the world as its best model, as roboticist Rodney Brooks once put (1991). In a block-copying-task (Ballard et al., 1992), subjects are asked to replicate a model shown in the model box in the workspace, using blocks in the resource box. Eye-movement tracking reveals that subjects look at each box many times, the same pattern found in the eye-movement tracking of the change blindness experiment. A natural interpretation is that we do not construct detailed internal representations of the external world, because the world is reliably there and representing the external world accurately exceeds the WM capacity.

From the ‘load theory of attention’, it is known that appropriately directing attention requires the active maintenance of stimulus priority in WM (De Fockert et al., 2001). Under high WM load conditions, then, it is difficult to maintain stimulus priority. As a result, more distracters are processed in WM. In other words, as WM load increases, we get more confused. This is a dilemma, since at the perceptual level (i.e. early selection), heavier (visual) load, or more visual information processing, reduces distracters (Lavie, 1995). When the world is visually complex, however, there is likely a heavy WM load and overflow of information at the cognitive level (i.e. late selection). Tversky (2010) argues that gestures, use of tools, and reconfiguration of the space will help us cognize, because they abstract, schematize representations, and facilitate our attention. That is, by means of abstracting and schematizing, attention is directed only to important aspects of the world. Tversky calls such abstracting/schematizing actions *spractions* (space-abstraction-action). Actions onto the external world, such as gestures, use of tools, and manipulations of the world, facilitate directing WM level attention only to relevant aspects of the world to the task at hand. That is, via *spractions*, WM load is offloaded onto the world (c.f. epistemic actions of Kirsh and Maglio; 1995). The hypothesis entailed is that, as WM load increases (as the world gets visually complex), people offload it onto the world rather than process it internally, although it is in principle possible to process it internally. Consequently more *spractions* (or epistemic actions) are likely to be observed.

## Experiment

To test whether/how we are offloading WM onto the world, subjects were asked to build Lego blocks and the way they used the space – use of *spractions* – was analyzed. Lego was chosen because Lego block assembly consists of pattern matching, planning, decision-making, and problem solving, all of which rely on WM. As the model used in the experiment targets young children, WM load is assumed relatively light. If offload of WM load is observed in this experiment, it can be generalized to many of daily situations, which have higher WM load.

## Method

The basic methodology used here is generally called cognitive ethnography (c.f. Ball et al., 2000; Hollan et al., 2000; Kirsh, 1999; Kessell and Tverksy, 2010). It differs from the traditional ethnography in that it emphasizes specificity, purposiveness, and confirmation (Ball et al., 2000). Rather than observing a field without prior knowledge or theory, cognitive ethnography relies on small-scale data collection based on representative time slices of the domain of interests that is confirmable. Instead of thick description (Geertz, 1973), participants’ activities were videotaped to be analyzed. To guarantee the objectivity of analysis (and consequently confirmation), codes were devised, and Cohen’s Kappa ( $0 \leq K \leq 1$ ) was calculated. Codes were devised so as to pick up *spractions*. Cohen’s Kappa is an indicator of inter-coder agreement; as Kappa is higher, coders interpret the same data more similarly, and thus analysis is considered more objective. It turned out that Cohen’s Kappa was 1.

## Participants

Total 6 female students from the same graduate school participated in the experiment on a voluntary basis. All participants agreed on being videotaped. They were all in their twenties when the experiment was conducted. They were all naïve as to the purpose of the study. They all signed an informed consent approved by the University Institutional Review Board (IRB).

## Material

Lego Technic 8065 (target age 7-14) was used for the experiment. It was selected for a pragmatic reason. It can be easily completed within one hour. Subjects were asked to build Lego Technic 8065 based on the instruction manual while being videotaped. Two different models can be built out of Lego Technic 8065. All were given the instructions for one of the models. Although some instructions instruct to sort blocks in advance, this one does not.

## Results

Although it is possible to process information internally, it was observed that participants constantly engaged in one or more of *spractions* over the videotaped session. That is, they



constantly used the external space as external WM. Out of six participants, five did sorting regardless that the instructions did not say to do so (some Lego instructions instruct to sort blocks in advance). Two did sorting in advance only; one did sorting on demand only; two did both. One did not do sorting at all; she made a significant number of mistakes. Although the sample size is too small to make a generalization, sorting seemed to help participants to think. As sorting is time-consuming, if viewed purely from the pragmatic perspective, it is disadvantageous (c.f. Kirsh and Maglio, 1994). Also, in principle it can be done in the head. Regardless, the participants did sorting. All the participants separated assembled blocks from the resource pool. This pattern was consistent. When there was more than one assembled block, they were grouped together and placed separated from not-yet-assembled blocks.

All of them looked at the instruction and/or model after picking up a piece. Although it was difficult to follow participants' eye movements, it was a consistent pattern that all the participants looked at the instruction and the model many times after picking up a desired piece. In most cases, they first looked at the instruction to pick a piece. Once they picked up a desired piece, they again looked at the instruction to see where it fit. Although it is possible to process both types of information simultaneously (i.e. which piece and where it goes), looking at the instruction once, it does not seem cost-effective given the stable world is out there and given that making detailed mental representation seems time consuming. It seems that use of external representation was commonplace. This finding is consistent with the theory of the limited WM capacity and previous experiments, such as block-copying task.

All of them did physical rotation and alignment following the instruction. When the model in the instruction was flipped, participants flipped their model as well. Although the instruction instructs to rotate, it does not instruct to align the model to the instruction. All the participants consistently aligned the instruction and the model under assembly. Such actions (alignment and rotation) are disadvantageous if they are taken purely as physical actions (c.f. Kirsh, 1995), but clearly have epistemic advantage. Although the Lego Technic 8065 is relatively simple (target age is 7-14), it still is too complex to mentally manipulate accurately. Physical rotation and alignment are clear cases of spraction.

One participant counted the number of holes by using another piece as a counting tool. She had to connect two parts by putting two bars into holes; there were thirteen holes, and bars had to be connected to the fourth and sixth holes respectively. She counted the number of holes on the instruction booklet with the piece. All the participants compared a piece with the booklet by placing the piece on the instruction, at least once during the assembly. Length is oftentimes overestimated or underestimated (Jones et al., 2006). It is thus difficult to accurately represent length mentally (WM load is heavy). The accurate length is printed on the instruction (obviously for measuring purpose).

Consequently, all the participants compared the length of a piece with the instruction by placing it on the instruction.

## Discussion

The world is full of complexity and we have to survive in such a complex world. The complexity of the world easily overwhelms the capacity of biological WM (Kessell and Tversky, 2010). Biological WM alone, then, does not seem to suffice for us to live a normal, smooth daily life. People exploit the external world as the material supervenience base of WM by means of spractions. People gesture, arrange the world, and make symbols and artifacts. Spractions and their consequences, such as reconfigured space, augment the limited biological WM capacity. WM then is not an equivalent concept to biological WM but it consists of biological WM and the external features of the world (and perhaps more, such as LTM). Both the brain and the world can serve as the material supervenience base of WM (c.f. Rowlands 1999). There is no qualitative difference between the external features of the world and the biological WM.

To observe how external WM plays out in reality, participants of the experiment were asked to assemble Lego blocks in front of a video camera. The analysis of the videotaped session revealed how they used the external world as external WM. They externally did what they in principle could do mentally. For example, they looked at the instruction after they picked up a desired piece to see where it is assembled. In principle, one gaze suffices to construct a mental representation of the external world. However, they referred back and forth between the piece and the instruction diagram. Similarly, they sorted blocks before assembly. In principle, sorting of pieces can be done purely mentally (if you have a photographic memory, you can in principle memorize all the patterns and locations of pieces on the table and classify them according to some manner). However, as the capacity of biological WM is limited to about four items of information, and Lego block assembly requires processing of more than four items of information, it seems participants externalized (offloaded) their cognition onto the world. Overall, spractions were observed constantly over the videotaped session. As the model used in the experiment target children between 7 and 14 years old, WM load is assumed relatively light. As many of daily situations are assumed to have higher WM load, it is inferred that offload of WM functions is ubiquitous in daily life.

The idea of externalization of WM function might be challenged on the ground that some people can do tremendous amount of information processing in the head alone without externalizing WM function. For example, some expert abacus users can multiple large numbers within a minute. Rumelhart et al. (McClelland et al., 1986, chap. 14) speculate that the ability to do information processing that seems too difficult to do in the head derives from the ability to do so externally; they are merely visually imagining what we do externally. Frank et al. (2011) confirmed this. That is, mental calculation by abacus users involves visual manipulation of imagined abacus.

Furthermore, they demonstrated that the amount of visual information abacus users process in the head cannot exceed the capacity limit of the biological visuo-spatial WM (in their case, 3). Thus, the fact that some people can do tremendous amount of information processing in the head without relying on the external world does not seem to constitute a counterexample.

We can calculate the amount of distributed WM-based cognition. The amount of externalized WM-based cognition is equal to the relevant amount of information for a given cognition minus four chunked items, the items of information the biological WM can process, or

$$y = z - \sum_{i=1}^4 x_i,$$

where  $y$  is the amount of distributed cognition (measured in the number of items),  $z$  is the number of items (cognitive load) demanded for a task at hand, and  $x_i$  is chunked items of information processed in the biological WM.

### Acknowledgments

I would like to thank professor Barbara Tversky of Stanford University and Teachers College, Columbia University, professor Hope Jensen Leichter of Teachers College, Columbia University, and Will Geluk for comments on earlier drafts.

### References

- Awh, E., Vogel, E., and Oh, S. (2006). Interactions between attention and working memory. *Neuroscience*, 139(1), 201-208.
- Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in cognitive sciences*, 4(11), 417-423.
- Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, 4(10), 829-839.
- Ball, L., and Ormerod, T. (2000). Putting ethnography to work: The case for a cognitive ethnography of design. *International Journal of Human-Computer Studies*, 53(1), 147-168.
- Ballard, D., Hayhoe, M., Li, F., Whitehead, S., Frisby, J., Taylor, J., et al. (1992). Hand-eye coordination during sequential tasks [and discussion]. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 337(1281), 331-339.
- Behrmann, M., and Tipper, S. (1999). Attention accesses multiple reference frames: Evidence from visual neglect. *Journal of Experimental Psychology: Human Perception and Performance*, 25(1), 83-101.
- Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, 9:261-325.
- Block, N., et al. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30(5), 481-498.
- Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences*, 15(12), 567-575.
- Brooks, R. (1991). Intelligence without representation. *Artificial intelligence*, 47, 139-159.
- Chalmers, D. (1996). Does a rock implement every finite-state automaton? *Synthese*, 108(3), 309-333.
- Chalmers, D. (1997). *The conscious mind: In search of a fundamental theory*. New York, NY: Oxford University Press.
- Chase, W. and Simon, H. (1973). Perception in chess. *Cognitive Psychology*, 4, 55-81.
- Clark, A. (1989). *Microcognition: philosophy, cognitive science, and parallel distributed cognition*. Cambridge, MA: MIT Press.
- Clark, A. (1993). *Associative engines: Connectionism, concepts, and representational change*. Cambridge, MA: MIT Press.
- Clark, A. (1998). *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.
- Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. New York, NY: Oxford University Press.
- Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. New York, NY: Oxford University Press.
- Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7-19.
- Clark, A. and Toribio, J. (1994). Doing without representing? *Synthese* 101, 401-431.
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1), 87-114.
- Cruse, H. (2003). The evolution of cognition—a hypothesis. *Cognitive Science*, 27(1), 135-155.
- De Fockert, J. W., Rees, G., Frith, C. D., & Lavie, N. (2001). The role of working memory in visual selective attention. *Science*, 291, 1803-1806.
- Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., and Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, 284(5416), 970-974.
- Frank, M. C. and Barner, D. (2011). Representing exact number visually using mental abacus. *Journal of Experimental Psychology: General*, 141(1), Feb 2012, 134-149.
- Geertz, C. (1973). *The interpretation of cultures*. New York, NY: Basic Books.
- Goldman, A. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. New York, NY: Oxford University Press.
- Gray, W., and Fu, W. (2004). Soft constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science*, 28(3), 359-382.
- Gray, W., Sims, C., Fu, W., and Schoelles, M. (2006). The soft constraints hypothesis: A rational analysis

- approach to resource allocation for interactive behavior. *Psychological Review*, 113(3), 461-482.
- Hauser, M., Carey, S., and Hauser, L. (2000). Spontaneous number representation in semi-free-ranging rhesus monkeys. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267(1445), 829-833.
- Hollan, J., Hutchins, E., and Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 7(2), 174-196.
- Horst, S. (2011). The Computational Theory of Mind. *The Stanford Encyclopedia of Philosophy*.
- Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.
- Jones, L and Lederman, S. (2006). *Human Hand Function*. New York, NY: Oxford University Press.
- Jonides, J., Lewis, R., Nee, D., Lustig, C., Berman, M., and Moore, K. (2008). The mind and brain of short-term memory. *Annual Review of Psychology*, 59, 193-224.
- Kessell, A. and Tversky, B. (2010). Thinking with hands and papers. Submitted.
- Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence*. 73(1-2), 31– 68.
- Kirsh, D. (1999). Distributed cognition, coordination and environment design. *Proceedings of the European conference on Cognitive Science*, 1–11.
- Kirsh, D. (2005). Multi-tasking and Cost Structure: Implications for Design. In, Bruno G. Bara, L. Barsalou, and M. Bucciarelli Mahwah (Eds.), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*. 1143-1148.
- Kirsh, D. (2006). Distributed cognition: A methodological note. *Pragmatics and Cognition*, 14, 249–262.
- Kirsh, D. (2010). Thinking with External Representations. *AI and Society*. 25: 441–454.
- Kirsh, D. and Maglio, P. (1994). On Distinguishing Epistemic from Pragmatic Actions. *Cognitive Science*, 18: 513-549.
- Landman, R., Spekrijse, H., and Lamme, V. (2003). Large capacity storage of integrated objects before change blindness. *Vision Research*, 43(2), 149-164.
- Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 451-468.
- Lavie, N. (2006). The role of perceptual load in visual awareness. *Brain Research*, 1080(1), 91-100.
- Lavie N., and De Fockert J. W. (2005). The role of working memory in attentional capture. *Psychonomic Bulletin and Review*, 12, 669–674.
- Levin, J. (2010). Functionalism. *The Stanford Encyclopedia of Philosophy*.
- Luck, S. J. and Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature* 390: 279 –281.
- Magnuson, J. S., Bensinger, D. G., Hayhoe, M., and Ballard, D. (1998). Learning to form visual chunks: On the structure of visuo-spatial working memory. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 645-650.
- McClelland, J.L., Rumelhart, D.E., and the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*, Cambridge, MA: MIT Press.
- Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2), 81-97.
- Noë, A. (2004) *Action in perception*. Cambridge, MA: MIT Press.
- Rowlands, M. (1999). *The body in mind: Understanding cognitive processes*. Cambridge, MA: Cambridge University Press.
- Rupert, R. (2004). Challenges to the hypothesis of extended cognition. *The Journal of Philosophy*, 101(8), 389-428.
- Shepard, R.N. and Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science* (171), 701–703.
- Simons, D., and Rensink, R. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, 9(1), 16-20.
- Smart, P., Engelbrecht, P., Braines, D., Strub, M., and Giammanco, C. (2010). The network-extended mind. *Network Science for Military Coalition Operations: Information Extraction and Interaction*, 191-236.
- Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74(11), 1-29.
- Tversky, B. (1981). Distortions in memory for maps. *Cognitive Psychology*, 13(3), 407-433.
- Tversky, B. (2010). Visualizing thought. *Topics in Cognitive Science*, 3(3), 499-535.
- Wilson, R. (1994). Wide computationalism. *Mind*, 103(411), 351-372.

# Experimental Investigation of Relationship between Complacency and Tendency to Use Automation System

**Akihiro Maehigashi (mhigashi@cog.human.nagoya-u.ac.jp)**

**Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)**

**Hitoshi Terai (terai@is.nagoya-u.ac.jp)**

Graduate School of Information Science, Nagoya University, Japan

**Kazuaki Kojima (koj@aoni.waseda.jp)**

Faculty of Human Sciences, Waseda University, Japan

**Junya Morita (j-morita@jaist.ac.jp)**

School of Knowledge Science, Japan Advanced Institute of Science and Technology, Japan

## Abstract

In this study, we experimentally investigated the relationship between complacency, defined as missing automation malfunctions, anomalous conditions, or outright failures, and the human tendency to prefer either automation or manual operation. We experimented using two different tasks with human participants to evaluate their individual tendencies to lapse into complacency and to use automation. The result indicated that the participants who prefer manual operation also tend to lapse into complacency. We assume that participant vigilance against automation stability might link these two phenomena.

**Keywords:** Human-automation system interaction; Complacency; Automation usage; Vigilance

## Introduction

Progress in technology has provided many opportunities for people to use automation systems, including on airplanes, ships, and in automobiles (Parasuraman, Molloy, & Singh, 1993; MacFadden, Giesbrecht, & Gula, 1998; Rajaonah, Tricot, Anceaux, & Millot, 2008). Parasuraman and Riley (1997) defined automation as technology that performs actions for humans. Human workload can be reduced by automation; however, the automation performance is often degraded by sudden environmental changes and automation malfunctions. Therefore, using automation is not always efficient. In using such automation, complacency is one major problem encountered by people. Complacency is defined as missing automation malfunctions, anomalous conditions, or outright failures caused by inadequate monitoring (Parasuraman & Manzey, 2010). Complacency causes fatal accidents since users do not detect or are slow to detect automation failures (Parasuraman et al., 1993).

Parasuraman and Manzey (2010) described why complacency happens. They stated that in multi-task situations where users allocate one task to automation and perform the other tasks by themselves, they need to manually conduct their tasks while monitoring the allocated automated task performance. In such situations, the manual tasks compete with the automated task for the user attention, and users tend to concentrate on their manual tasks instead of the automated task. Therefore, the frequency of monitoring automation is lowered, and users often fail or become slow to detect automation failures. Moreover, Parasuraman and Manzey

(2010) showed that there are two types of complacency: fixation and attention failures. Fixation failure occurs when the automation failure is out of the users' sight. Attention failure occurs when the automation failure is within the users' sight, but out of their attention.

Many studies about complacency have experimentally investigated fixation failure with multi-task situations where the automated and manual tasks were displayed separately. The participants conducted their own manual tasks while simultaneously monitoring the automated task. Automation breakdown occurred during the experiments. When automation failure was detected, the participants had to push a particular button on their computer keyboards. Such task situations showed the following strong complacency effects: (1) when the automation capability is stable rather than variable (Parasuraman et al., 1993), (2) when the workload is high rather than low (Metzger & Parasuraman, 2005), and (3) when the arousal level is low rather than high (Singh, Molloy, & Parasuraman, 1993).

On the other hand, a few studies about complacency have experimentally investigated attention failure. Duley, Westerman, Molloy, and Parasuraman (1997) set up the same multi-task situation as in the previous studies about fixation failure, but they superimposed the automated task on the manual task on a single display. Automation breakdown also occurred during the task. In their experiment, complacency occurred on the superimposed display. They showed that even when the automated task is in the users' sight, complacency occurs because they do not focus on the automated task.

The previous studies of human-automation system interaction, which experimentally investigated the human preference to use automation or to conduct manual operation, indicated individual differences in the selection. Lee and Moray (1992, 1994) showed that users who have a tendency to use automation, i.e., automation-oriented users, tend to trust automation. Rajaonah et al. (2008) showed that manual-oriented users tend to perceive higher workload and greatly decrease their vigilance while monitoring automation more than automation-oriented users. Additionally, Maehigashi, Miwa, Terai, Kojima, and Morita (2011) showed that automation-

oriented users tend to select whether to use automation or manual operation by reacting more sensitively to the changes of automation capabilities than manual-oriented users.

As indicated above, many studies have experimentally investigated the nature of human complacency in automation usage and its orientation. However, no experimental investigations have focused on the relationship between these two phenomena. In this study, we experimentally investigated the relationship between complacency and a preference of automation usage. We drew the following hypotheses:

- Hypothesis 1: A relationship exists between complacency and a preference for automation usage.

If Hypothesis 1 is supported, two detailed hypotheses are raised. Lee and Moray (1992, 1994) showed that users who tend to use automation also tend to trust it. Parasuraman and Manzey (2010) argued that overtrust in automation may lower the frequency of monitoring it, causing complacency. Therefore, it is predicted that automation-oriented users will detect automation failures more slowly than manual-oriented users. Therefore, Hypothesis 2a is as follows:

- Hypothesis 2a: Users who prefer to use automation tend to lapse into complacency.

On the other hand, Rajaonah et al. (2008) showed that users who tend to conduct manual operation tend to greatly decrease vigilance while monitoring automation more than automation-oriented users. Vigilance is the ability to sustain attention, and a lack causes complacency (Molloy & Parasuraman, 1996). Therefore, it is predicted that users who tend to conduct manual operation will tend to detect automation failures slower than automation-oriented users. Therefore, Hypothesis 2b is as follows:

- Hypothesis 2b: Users who prefer manual operation tend to become complacent.

## Experimental task

We used two different experimental tasks. The first was the auto-manual selection task used by Maehigashi et al. (2011). We evaluated participant preferences to use automation based on their performances in this task, where they tracked a line that scrolls downward past a circle vehicle. When the circle vehicle veers off the line, the performance score is reduced as operational error. The participants were allowed to switch to either auto mode (operation completely performed by the program) or manual mode (operation performed by participants using left and right arrow keys) by pressing a selector on the keyboard. We manipulated the auto and manual capabilities with five levels, and the auto and manual capabilities changed independently. The participants had to compare the auto and manual capabilities to select the mode that shows higher task performance (see Maehigashi et al. (2011) for details). It is preferable to select the auto mode when the auto capability is higher and the manual mode when the manual capability

is higher. However, deviation from the normative behavior is caused by the participant tendencies to use automation. We evaluated their preferences to use automation based on the percentage of using the auto mode in the auto-manual selection task.

The second was a supervisory control task (Figure 1). We evaluated participant tendencies to lapse into complacency based on their performances on a dual task shown on a single display. One was a search task in which the participants looked for target stimuli (L) among distracter stimuli (T) that scroll downward. When the target was found on the screen, the participants pressed a selector on the keyboard while the target is inside the double line (detection area) at the display's bottom. If the target is successfully detected, the color of the target letter changes to red. When the participants missed the target or gave a false alarm, the performance score was reduced as operational error. We manipulated the number of target and distracter stimuli for high and low workload conditions in the experiment.

The other supervisory control task was a monitoring one. The participants monitored an auto that operates a circle vehicle to track a line. The line scrolls downward past the circle vehicle. When the circle vehicle veers off the line, the performance score is reduced as operational error. Basically, the auto perfectly performs the line tracking. However, in specific timing, auto failures occur and the circle vehicle stops tracking the line during the task. When the participants detect the auto failures, they need to manually operate the circle vehicle by pressing the left and right arrow keys. In the supervisory control task, the participants simultaneously conducted these two search and monitoring tasks. We evaluated their tendency to become complacent based on their reactions to the auto failures.

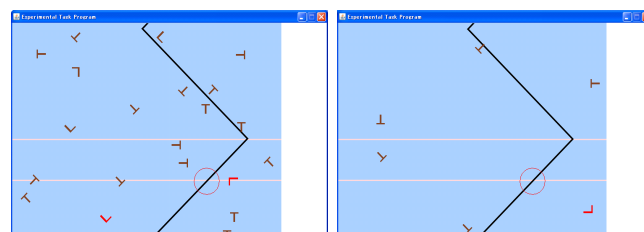


Figure 1: Supervisory control task: high (left) and low (right) workload conditions.

## Experiment

### Method

**Participants** Thirty-five university students participated in our experiment.

**Factorial design** In the supervisory control task, the experiment had a two-factor within participants design: (1) workload (high and low) and (2) type of failure (sudden and gradual). For the workload factor, 100 target and 400 distracter stimuli in the high workload condition and 25 target and 100 distracter stimuli in the low workload condition emerged during the task. For the type of failure factor, in the sudden

failure condition, the auto that has perfectly operated the circle vehicle suddenly stops because its capability suddenly becomes 0. The breakdown continues for 10 seconds. The circle vehicle that tracks the line suddenly stops moving and immediately veers off the line. In the gradual failure condition, the auto gradually stops operating because its capability gradually decreases to 0 for 40 seconds. The breakdown continues for 10 seconds. The circle vehicle that tracks the line gradually slows down and finally veers off the line.

**Procedure** Prior to the supervisory control task, we conducted an auto-manual selection task whose procedure was basically the same as in Maehigashi et al. (2011). First, the participants practiced conducting the task in two training trials. Next we conducted an experimental task that consisted of 25, 40-second trials. The auto and manual capabilities randomly changed among five levels during the task. When one trial ended and another began, the display showed “capabilities changed” at the center of the screen. Throughout the experiment, the values of the auto and manual capabilities were not displayed on the screen. The participants were required to achieve as high a score as possible. They were given a five-minute break after the auto-manual selection task was completed.

After the break, we conducted the supervisory control task. In the training trials, the participants first separately practiced conducting the search and monitoring tasks and then experienced both simultaneously. Each training trial lasted one minute. After the training trials, we conducted the experimental task that consisted of two blocks of nine trials each. One block was conducted in the high workload condition and the other in the low workload condition. The order of the workload conditions was counterbalanced among the participants. The auto failures occurred three times in each block. Sudden failure occurred 40 seconds after the second trial began. Gradual failures occurred from the beginning of the fifth (or sixth) and the ninth trials. The participants were required to operate the circle vehicle manually when auto failure occurred. However, participants were not informed about either types, the frequency, or the timing of the auto failure. When the auto capability recovered after the auto failures, the display showed “capability recovered” at the center of the screen. The participants were instructed to delegate the operation back to the auto after the recovery. Throughout the experiment, the manual capability was stable and sufficient to manually track the line.

## Result

**Manipulation check** In the auto-manual selection task, for all participants, the mean percentage of using the auto mode was 44.20%, and the mean task performance (the percentage the circle vehicle was on the line) was 75.02%. In the search task of the supervisory control task, the mean number of missed targets was 0.28 in the high workload condition and 0 in the low workload condition. The mean number of false alarms was 0.67 in the high workload condition and 0.31 in the low workload condition. Since the numbers of missed targets and false alarms were quite low, the participants conducted the search task almost perfectly. We conducted t-tests on the numbers of missed targets and false alarms in the high and low workload conditions. The number of missed targets was significantly higher in the high workload condition ( $t(34) = 2.53, p < .05$ ), and the number of false alarms was marginally higher in the high workload condition ( $t(34) = 1.77, p = .09$ ). These results confirmed that the workload was higher in the high workload condition than in the low workload condition.

**Evaluation index** We evaluated the participant preferences to use automation based on the percentage of using the auto mode in the auto-manual selection task. We evaluated the participant tendencies to lapse into complacency in the supervisory control task based on three evaluation indices: reaction time, distance, and accumulated distance. The reaction time was the time (msecs) from when the auto failure began to when the manual operation was first conducted. We utilized the time from when the auto breakdown occurred for the sudden failure and the time from when the auto capability started to decrease for the gradual failure. The distance (pixels) was measured between the circle vehicle and the line when the manual operation was first conducted during the auto failure. A distance over 30 pixels means that the circle vehicle is out of the line. The accumulated distance (pixels) is the distance accumulated from when the auto failure began through when the manual operation was first conducted. Table 1 shows the mean reaction time, the distance, and the accumulated distance for all participants in each condition of the supervisory control task.

**Consistency of tendency to lapse into complacency** Prior to verification of the hypotheses, we investigated the consistency of the participant tendencies to lapse into complacency. We conducted correlation analyses on the individual

Table 1: Mean reaction time, distance, and accumulated distance for all participants in each condition of supervisory control task. Values in parentheses show standard deviations.

	High workload		Low workload	
	Sudden failure	Gradual failure	Sudden failure	Gradual failure
Reaction time (msecs)	1051.74(425.16)	23350.63(9795.37)	1137.46(450.49)	24674.71(10222.56)
Distance (pixels)	19.83(7.44)	9.09(9.27)	20.96(7.75)	9.94(8.75)
Accumulated distance (pixels)	655.77(517.82)	1926.23(2165.69)	738.93(549.87)	1910.13(2025.02)

reaction times, distances, and accumulated distances across the four conditions in the supervisory control task (Table 2). The results showed similar correlations in each index. First, there were correlations between the high and low workload situations both in the sudden and gradual failure conditions. The results suggest that the participants who reacted faster to the auto failure in the high workload condition also reacted faster in the low workload condition. This consistency was observed only within the same type of auto failure. Moreover, there was a correlation between the sudden and gradual failure situations in the high workload condition, suggesting that the participants who reacted faster to the sudden failure also reacted faster to the gradual failure but only in the high workload condition.

The results of the correlation analyses in the supervisory control task showed the consistency of the participant tendencies to lapse into complacency. However, we only found consistency in the participant reactions to the different types of auto failures in the high workload condition. In the low workload condition, the number of search stimuli was low. The participants may have allocated enough attention to monitoring the auto operation to easily detect the auto failure. On the other hand, in the high workload condition, the participants probably had difficulty concentrating on the monitoring activities. Therefore, individual differences in complacency became salient only in the high workload condition.

### Relationship between complacency and tendency to use auto mode

To verify the hypotheses, we conducted a correlation analysis on the relationship between the individual reaction times, the distances, and the accumulated distances in the supervisory control task and the individual percentage of using the auto mode in the auto-manual selection task (Table 3). The results showed a correlation between each index in the sudden failure situation of the supervisory control task and the percentage of using the auto mode in the auto-manual selection but only in the high workload condition (Figure 2). The participants who had a tendency to conduct manual operation in the auto-manual selection task also tended to react slowly to the auto failure in the supervisory control task. The result supports Hypotheses 1 and 2b, but only in the sudden failure situations when the workload was high.

## Discussion

As a result of our experiments, we found a relationship between complacency and the tendency to prefer automation or manual operation. The participants who tended to conduct manual operation in the auto-manual selection task also tended to react to the automation failure slowly: they tended to lapse into complacency. This result supported Hypotheses 1 and 2b, but only in the high workload and sudden failure condition.

For the tendency to use automation, Rajaonah et al. (2008)

Table 2: Correlation matrices that show correlations on individual reaction times, distances, and accumulated distances among four conditions. Values are correlation coefficients ( $r$ ).

Reaction time		High workload		Low workload	
		Sudden failure	Gradual failure	Sudden failure	Gradual failure
High workload	Sudden failure	1			
	Gradual failure	.54**	1		
Low workload	Sudden failure	.54**	.27	1	
	Gradual failure	-.16	.40*	.09	1

Distance		High workload		Low workload	
		Sudden failure	Gradual failure	Sudden failure	Gradual failure
High workload	Sudden failure	1			
	Gradual failure	.48**	1		
Low workload	Sudden failure	.56**	.19	1	
	Gradual failure	-.02	.36*	.23	1

Accumulated distance		High workload		Low workload	
		Sudden failure	Gradual failure	Sudden failure	Gradual failure
High workload	Sudden failure	1			
	Gradual failure	.50***	1		
Low workload	Sudden failure	.59**	.26	1	
	Gradual failure	.15	.55*	.29	1

\* $p < .05$ , \*\* $p < .005$ , \*\*\* $p < .001$



Table 3: Correlation matrices that show correlations among individual reaction times, distances, and accumulated distances in four conditions of supervisory control task and individual percentage of using auto mode in auto-manual selection task. Values are correlation coefficients ( $r$ ).

		Supervisory control task					
		High workload					
		Sudden failure			Gradual failure		
		Reaction time	Distance	Accumulated distance	Reaction time	Distance	Accumulated distance
Auto-manual selection task	Percentage of using auto mode	-.49**	-.50**	-.56**	-.21	-.12	-.09

		Supervisory control task					
		Low workload					
		Sudden failure			Gradual failure		
		Reaction time	Distance	Accumulated distance	Reaction time	Distance	Accumulated distance
Auto-manual selection task	Percentage of using auto mode	-.28	-.26	-.30	.32	.26	.20

\*\*  $p < .005$

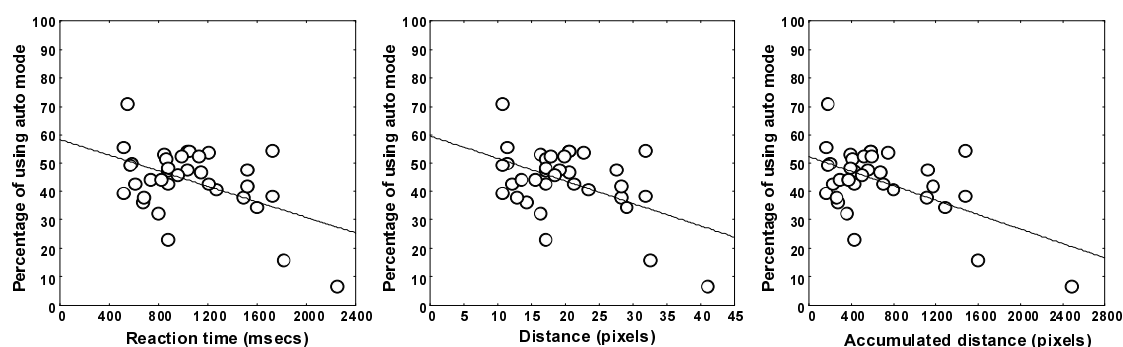


Figure 2: Correlation between individual reaction times, distances, and accumulated distances in high workload and sudden failure condition of supervisory control task (x-axis) and individual percentage of using auto mode in auto-manual selection task (y-axis).

showed that users who prefer manual operation tend to perceive higher workload and greatly decrease their vigilance while monitoring automation more than users who prefer to use automation. They discussed the possibility that manual-oriented users might try to avoid high workload and vigilance decrement without using automation. As a result, the manual-oriented participants in our study might tend not to use automation even when using it was efficient.

Next, for the effect of workload on complacency, the relation of the tendency to use automation in the auto-manual selection task and the reaction to the automation failure was detected only in the high workload condition of the supervisory control task. Molloy and Parasuraman (1996) showed that vigilance decrement is greater in the high workload condition than in the low workload condition. In our experiment's low workload condition, the number of search stimuli was relatively low so the participants could allocate enough attention to the auto operation. Therefore, the manual-oriented participants might be able to successfully manage vigilance

and quickly detect automation failure. On the other hand, in the high workload condition, the vigilance decrement became greater for the manual-oriented participants, resulting in slower detection of automation failure. That is the reason that we could only detect a significant correlation in the high workload condition.

Moreover, for the effect of automation failure types on complacency, the consistency of the tendency to use automation in the auto-manual selection task and the reaction to automation failure was detected only in the sudden failure situation of the supervisory control task. Endsley (1995) stated that individual differences in situation awareness influence individual decision making and the performances of actions. Endsley (1996) also indicated a relationship between vigilance and situation awareness decrements. In our experiment, in the gradual failure situation, the circle vehicle gradually slowed down and finally veered off the line. Some participants quickly shifted to manual operation after they noticed the irregular movement of the circle vehicle. Others might

continue to monitor the auto operation until just before the auto breakdown, anticipating that the circle vehicle would veer off the line after a certain amount of time. In such a situation, we assume that the participant reactions were influenced not only by individual differences in the awareness of the auto failure but also such factors as decision making strategy and action selections followed by awareness. By contrast, in the sudden failure situation, the circle vehicle suddenly stopped tracking the line and immediately veered off the line. In such a situation, the individual differences in the awareness of the auto failure directly influenced the participant reactions. As a result, there is only a relationship between the tendency to use automation and the reaction to the automation failure in the sudden failure condition.

Finally, Lee and Moray (1992, 1994) showed that users who tend to use automation tend to trust it. An overtrust in automation might lower the frequency of monitoring it and cause complacency (Parasuraman & Manzey, 2010). Therefore, we predicted that automation-oriented users would slowly detect automation failures. Contrary to our prediction, however, we found that manual-oriented users tended to slowly detect the automation failures, rejecting Hypothesis 2a. In our experiment, to evaluate the participant tendencies to lapse into complacency, we set up a situation where attention failure—not fixation failure—was induced using a superimposed display. Perhaps in such a situation, individual differences in vigilance rather than trust in automation link complacency and the tendency to use automation.

In this study, we investigated the relationship between complacency and the tendency to select whether to use automation or conduct manual operation. We evaluated complacency with a supervisory control task in which attention failure was induced. We evaluated the preference to use automation with an auto-manual selection task. Our experiment indicated that users who tend to conduct manual operation tend to lapse into complacency. However, such a relationship was found only in the high workload situation where sudden automation failure occurs. We assume that individual differences in vigilance link complacency and the tendency to use automation.

## References

- Duley, J. A., Westerman, S., Molloy, R., & Parasuraman, R. (1997). Effects of display superimposition on monitoring of automation. In *Proceedings of the 9th international symposium on aviation psychology* (pp. 322–328). Columbus, OH: Association of Aviation Psychology.
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37, 32–64.
- Endsley, M. R. (1996). Automation and situation awareness. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications*. Mahwah, NJ: Lawrence Erlbaum.
- Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35, 1243–1270.
- Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operator's adaptation to automation. *International Journal of Human-Computer Studies*, 40, 153–184.
- MacFadden, S. M., Giesbrecht, B. L., & Gula, C. A. (1998). Use of an automatic tracker as a function of its reliability. *Ergonomics*, 41, 512–536.
- Maehigashi, A., Miwa, K., Terai, H., Kojima, K., & Morita, J. (2011). Selection strategy of effort control: allocation of function to manual operator or automation system. In L. Carlson, C. Hoelscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 1977–1982). Austin, TX: Cognitive Science Society.
- Metzger, U., & Parasuraman, R. (2005). Automation in future air traffic management: effects of decision aid reliability on controller performance and mental workload. *Human Factors*, 47, 35–49.
- Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: vigilance and task complexity effects. *Human Factors*, 38, 311–322.
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Human Factors*, 52, 381–410.
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced “complacency”. *The International Journal of Aviation Psychology*, 3, 1–23.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Human Factors*, 39, 230–253.
- Rajaonah, B., Tricot, N., Anceaux, F., & Millot, P. (2008). The role of intervening variables in driver-acc cooperation. *International Journal of Human-Computer Studies*, 66, 185–197.
- Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Individual differences in monitoring failures of automation. *The Journal of General Psychology*, 120, 357–373.
- Lee, J. D., & Moray, N. (1992). Trust, control strategies

# Inductive reasoning in the courtroom: Judging guilt based on uncertain evidence

**Ann M. Martin (Ann.Martin@unsw.edu.au)**  
School of Psychology, University of New South Wales  
Sydney, 2052, Australia

**Brett K. Hayes (B.Hayes@unsw.edu.au)**  
School of Psychology, University of New South Wales  
Sydney, 2052, Australia

## Abstract

Most legal systems require jurors to consider all the evidence presented at trial. Hence when there is uncertainty over aspects of evidence this should be factored into juror judgments. Two experiments examined how mock jurors used uncertain information in their ratings of defendant guilt and final verdicts. Participants read scenarios where an eyewitness expressed uncertainty about the identity of a critical piece of evidence (e.g. the object a defendant was holding could have been a knife or a mobile phone). The respective probability of these alternatives was varied, as was their association with the alleged crime. When the probability of the alternatives was varied between subjects (Experiment 1) there was only weak evidence that jurors considered both alternatives. When probability was varied within-subjects (Experiment 2), jurors did consider both alternatives in their guilt judgments. The implications for theories of reasoning with uncertain information and forensic practice are discussed.

**Keywords:** Inductive reasoning; Probabilistic reasoning; Category-based induction, Forensic judgment

## Introduction

The process of deciding on the guilt or innocence of a defendant in a criminal trial could be seen as form of complex inductive inference (Lagnado, 2011). Induction involves drawing probabilistic inferences from given information. When the category membership of an object is known with certainty (e.g., a man at a robbery crime scene was carrying a *knife*) the process of inductive inference is relatively straightforward (e.g., the knife was likely being used as a weapon in the robbery).

However, the evidence presented in criminal cases is typically complex and fraught with uncertainty (e.g., an eyewitness may not be certain about the identity of the object that the defendant was holding). In such cases induction involves the consideration of multiple possible object categories that may have different implications for judgments about the defendant's guilt. For example, if a defendant was thought to be carrying a knife then a juror may judge the defendant as likely to be guilty of committing a crime. But if there is some chance that the object in the defendant's hand was something less incriminating (e.g., a metallic-colored mobile phone) then this may reduce belief in the defendant's guilt.

So just how do jurors respond to such uncertain alternatives when making inferences about guilt or

innocence? The legal answer to this question is straightforward and prescriptive. Most criminal jurisdictions specify that jurors should consider *all* evidence presented at trial when determining defendant culpability (e.g., Attorney General's Department of New South Wales, 2007). Laboratory studies of inductive reasoning with uncertain categories however, suggest a more complex answer (see Hayes, Heit & Swendsen, 2010 for a review).

Bayesian approaches to inductive reasoning, such as Anderson's (1991) Rational model, generally agree with the legal ideal, suggesting that reasoners incorporate information about all category alternatives when making inferences (hereafter referred to as *multiple-category reasoning*). To illustrate, let us assume that an eyewitness believes that the probability that the object the defendant was holding was a knife is 0.7 (which we will refer to as the *primary category*), with a 0.3 probability that the object was instead a metallic mobile phone (*the secondary category*). Further assume that these alternatives are associated with different conditional probabilities of guilt. If the person was holding a knife then the probability of them being guilty is high (e.g.,  $p(\text{guilt} | \text{knife}) = 0.9$ ), but if they were holding a mobile phone the probability of guilt will be much lower (e.g.,  $p(\text{guilt} | \text{phone}) = 0.2$ ). Applying Bayes' theorem, the Rational model combines the probabilities from the primary and secondary categories to give an estimate of the probability that the defendant was guilty, given they were seen with a metallic object in their hand ( $p(\text{guilt} | \text{metallic object}) = (0.7 \cdot 0.9) + (0.3 \cdot 0.2) = 0.69$ ). Note that this probability estimate is considerably lower than would be the case if the uncertainty over object identity was ignored. If the juror assumed that the object in the defendant's hand was a knife then the guilt estimate would be 0.9.

Unfortunately (from a legal point of view), empirical studies have so far found little evidence of multiple-category reasoning. Malt, Ross and Murphy (1995) for example, presented vignettes in which the category membership of a target character was uncertain, and asked participants to make various inferences about the targets. For each scenario two possible category identities were suggested, a more likely primary category and a less likely but plausible secondary category. The primary category was held constant across conditions (e.g., the vignette always made it clear that the target character was most likely a realtor). The secondary category however was varied; in one condition the target was most likely a *cable repairman*, in

another it was most likely a *burglar*. These alternatives have different implications for inferences such as “how likely is it that the man will pay attention to the sturdiness of the doors on the house?” Such behavior seems more likely if the target was a burglar than if they were a realtor or cable repairman. If people do consider multiple categories in induction, then their inferences should differ across the conditions with different secondary categories. Malt et al. (1995) however, found that participants tended to ignore the secondary category when making inductive inferences. Predictions were predominantly based on consideration of the primary category alone (also see Ross & Murphy, 1996).

Such “single-category reasoning” seems pervasive in non-forensic domains, having been demonstrated with a wide variety of artificial and natural categories (see Murphy & Ross, 2007, 2010 for reviews). Murphy and Ross (2007) suggest that single-category reasoning can be viewed as a cognitive heuristic that reduces the complexity of deriving inductive predictions from multiple uncertain alternatives.

The pervasive nature of single-category reasoning in previous studies leads to a negative prognosis for forensic reasoning, suggesting that jurors are also likely to use the single-category heuristic. Such a prediction is consistent with reports of juror “satisficing” where the juror focuses on aspects of trial evidence that are consistent with a single coherent story, ignoring contradictory evidence (Kuhn, Weinstock, & Flaton, 1994; Pennington & Hastie, 1992).

On the other hand, multiple-category reasoning may be more common in forensic situations because of their particular motivational demands. Motivational factors play an important role in determining the depth and complexity of reasoning (Gilovich & Griffin, 2010; Kunda, 1990). Forensic judgments like guilt or innocence are widely recognised as having profound consequences for a defendant and for the wider community (Bornstein & Greene, 2011). Hence jurors may be more likely to consider uncertain alternatives when making highly consequential judgments. Some support for this prediction comes from Hayes and Newell (2009) who found that multiple-category reasoning was more likely when neglect of the secondary category could lead to a highly negative outcome (e.g. when the primary category was a common but easily treatable disease and the secondary category was a rare but potentially terminal disease).

The main aim of the current studies therefore was to examine whether mock jurors would show multiple-category reasoning in cases where there was uncertainty about the identity of forensically relevant evidence in a criminal trial.

## Experiment 1

Experiments 1 and 2 were patterned after those of Malt et al. (1995) and Ross and Murphy (1996), using a design where probability estimates of guilt were compared across conditions in which the primary category was held constant and the secondary category varied (see Table 1). Participants in all conditions were shown written vignettes

which described criminal cases where there was uncertainty about the identity of a critical piece of evidence. An eyewitness testified that they observed the defendant carrying an object, which they believed to be a particular item (the primary category). However, they acknowledged that there was a lower probability the object may have been something else (the secondary category). In two Primary Related conditions (Comparison 1 in Table 1), the primary category consistently implicated the defendant in the crime (e.g., in a robbery the primary category was “knife”). However, the less likely secondary category varied such that it indicated that the defendant was guilty in one condition only (e.g., a stolen watch in one condition and a mobile phone in the other).

After reading the vignettes participants judged which object they thought the defendant was actually carrying (object categorization) and made inferences about the defendant’s guilt. Whether participants used single- or multiple-category reasoning to arrive at these inferences could be determined by comparing guilt estimates in the conditions where the secondary categories were varied (see Comparisons 1 & 2 in Table 1).

Table 1: Summary of the experimental design (with examples of critical object alternatives from the Robbery vignette)

	Secondary Category	
	Related (e.g., stolen watch)	Unrelated (e.g., mobile phone)
<b>Comparison 1</b>		
<b>Primary Category</b>	Primary = <i>knife</i>	Primary = <i>knife</i>
<b>Related</b>	Secondary = <i>stolen watch</i>	Secondary = <i>mobile phone</i>
(e.g., knife)		
<b>Comparison 2</b>		
<b>Primary Category</b>	Primary = <i>keys</i>	Primary = <i>keys</i>
<b>Unrelated</b>	Secondary = <i>stolen watch</i>	Secondary = <i>mobile phone</i>
(e.g., keys)		

A subsidiary aim of Experiment 1 was to re-examine the Ross and Murphy (1996) finding that multiple-category reasoning is more likely when the secondary category is closely linked to the prediction being made and the primary category is not. In a variant on the realtor/cable guy/burglar task, Ross and Murphy (1996) asked participants to make predictions that were more strongly associated with the secondary category of burglar but not with the primary category of realtor (e.g., the prediction “how likely is it that the man will try to find out if the householder keeps her windows locked?”). Participants were more likely to consider both categories when making such predictions.

Hence in the current study, we also ran a Primary unrelated condition where the primary category was not strongly associated with culpability and the association of the secondary category with culpability was varied. In the Primary unrelated, Secondary related condition the

secondary category was associated with guilt (e.g., a stolen watch), whereas in the Primary unrelated, Secondary unrelated condition, neither category was associated with the crime. Consideration of the secondary category should lead to higher guilt estimates in the Primary unrelated, Secondary related condition than in the Primary unrelated, Secondary unrelated group.

## Method

**Participants** Ninety-eight undergraduate students participated for course credit. The majority were female ( $n = 72$ ), and the mean age was 19.95 years ( $SD = 1.84$ ). All were Australian citizens aged 18 years or older, in accordance with Australian juror selection criteria.

**Design and Materials** The experiment followed a  $2 \times 2$  factorial design with the first factor being whether the *primary* category related to the crime and the second factor being whether the *secondary* category related to the crime (see Table 1). In Experiment 1 both factors were manipulated between subjects, with approximately equal numbers allocated to each experimental condition.

These factors were operationalized using written vignettes presented as brief criminal trial summaries. Each vignette was approximately 290 words in length and described a case in which an eyewitness reported seeing the defendant at the crime scene carrying an object whose identity was critical for evaluating defendant guilt (see Appendix for an example). Two category possibilities were provided for this object: a primary category, which was described as the most likely identity of the critical object (with an explicit likelihood of 70%), and a secondary category, which was described as having “a small chance” of being the identity of the object. Two vignettes with this structure were developed. One described a criminal trial for assault and robbery, and the other described a trial for arson.

Our assumptions about the relatedness of the various critical objects with the crime were confirmed in a pilot study. Fifty nine participants who did not take part in the main experiments read versions of the vignettes in which the eyewitness provided only one category for the critical object. Participants were then asked to rate the likelihood that the defendant was guilty on a 100-point scale (1= not at all likely, 100 = very likely). Defendants seen carrying “related objects” (*knife, stolen watch*) were rated as more likely to be guilty ( $M = 62.85$ ) than those carrying “unrelated objects” (*mobile phone, keys*), ( $M = 35.37$ ),  $F(1, 55) = 27.24$ ,  $p < .001$ .

**Procedure** Participants were told that they were to play the role of a juror in determining the guilt of a defendant in a criminal trial. Each vignette was then presented on a computer screen with the eyewitness evidence about the critical object alternatives written in bold type. After reading the vignette participants clicked an on-screen button which started a series of questions (with the vignette no longer visible). The first was a categorization question that asked

participants to rate the percentage likelihood that the item the defendant was carrying was (a) the primary category, (b) the secondary category, or (c) some other item, with the restriction that the three estimates must sum to 100 percent (see Murphy & Ross, 2010 for a similar procedure). This was followed by a filler question asking participants to recall the general location of the crime as described in the vignette. The final two questions required participants to infer defendant culpability based on the trial evidence. Participants first estimated the likelihood that the defendant was guilty on a 100-point scale (1=not at all likely, 100=very likely). They then rendered a binary verdict (‘guilty’ or ‘not guilty’) by clicking on one of two forced choice buttons.

Previous research (e.g., Harris & Hahn, 2009) indicates that mock jurors often give less weight to evidence that they perceive to be inconsistent. In our experimental vignettes all eyewitnesses expressed some inconsistency about the identity of the critical object. However, it is possible that the effects of this inconsistency on juror confidence in eyewitness evidence may have differed across experimental conditions (e.g., jurors may give more weight to evidence when both objects are positively related to the crime, than when one is related and the other is not). To check on this possibility all participants were also asked to rate their confidence in the reliability of the eyewitness on a seven-point scale (1= ‘very low confidence’, 7= ‘very high confidence’).

Within each condition participants completed both robbery and arson vignettes with order of vignette administration counterbalanced.

## Results and Discussion

**Preliminary Analyses.** The experimental predictions are based on the assumption that participants believe that the critical object was more likely to belong to the primary than the secondary category. The object categorization data identified 12 participants who did not rate the primary category as more likely. Consequently they were excluded from further analyses.

Preliminary analyses confirmed that the percentage likelihood ratings for membership of the critical object in the primary ( $M = 69\%$ ) and secondary categories ( $M = 26\%$ ) were close to those stated or implied in the vignettes. Notably participants in all conditions rated the likelihood of secondary category membership as well above zero (group means ranged from 21% to 30%). Hence participants in all conditions grasped the uncertainty about the category membership of the critical object.

Preliminary analyses confirmed that there were no significant differences between the robbery and arson vignettes in overall guilt estimates or ratings of reliability of eyewitness evidence. Consequently, in both experiments analyses of guilt judgments were collapsed across vignettes.

A two-way analysis of variance revealed a significant main effect of secondary category association on ratings of eyewitness reliability,  $F(1,82) = 6.36$ ,  $p = .01$ . Witnesses in

the secondary related conditions were rated as more reliable ( $M = 4.1$ ) than those in the secondary unrelated conditions ( $M = 3.52$ ). Hence, eyewitness reliability ratings were included as a covariate in analyses of guilt judgments.

**Guilt Likelihood Ratings and Guilt Verdicts.** Both variates were analyzed in  $2(\text{Primary category status}) \times 2(\text{Secondary category status})$  analyses of covariance with eyewitness reliability entered as a covariate. Figure 1 shows the adjusted guilt likelihood ratings for each condition. Not surprisingly the defendant was rated as more likely to be guilty when the primary category for the critical object was closely linked to the crime than when it was unrelated,  $F(1,81) = 32.97, p < .001$ . The more critical question was whether the status of the secondary category affected guilt estimates. There was no main effect of secondary category status on guilt likelihood ratings,  $F(1,81) = 0.02, p = .89$ , but there was a significant interaction between the status of the primary and secondary categories,  $F(1,81) = 5.05, p = .027$ . However, when Tukey's HSD tests were applied, there was no significant effect of secondary category association in the critical comparison between the two Primary Related conditions ( $q = 2.45, p > .05$ ) nor between the two Primary Unrelated conditions ( $q = -2.16, p > .05$ ).

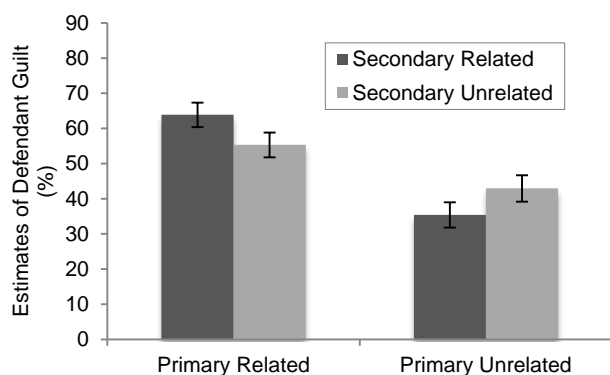


Figure 1. Mean likelihood ratings of defendant guilt (with standard error bars).

The proportion of guilty verdicts was higher when the primary category was related to the crime ( $M = 0.6$ ) than when it was unrelated ( $M = 0.21$ ),  $F(1,81) = 28.80, p < .001$ . However, there was no main effect or interaction involving the status of the secondary category ( $p$ 's  $> .15$ ). In other words there was no evidence that participants considered the secondary category when deciding on guilt verdicts.

In sum, this study found minimal evidence for multiple-category reasoning in a forensic context. Note however, that there was considerable variance in guilt judgments *within* conditions. For binary judgments for example, the mean standard error across experimental conditions was 0.073 (in a scale ranging from 0.0-1.0). This high level of variability is consistent with previous findings that individual jurors presented with the same evidence often give very different absolute estimates of guilt (e.g., Solana, García, & Tamayo,

1998). Such levels of variability are likely to have reduced the sensitivity of our tests of multiple-category reasoning. Experiment 2 addressed this issue by manipulating the status of the secondary category *within* subjects.

## Experiment 2

Experiment 2 aimed to provide a more sensitive test of whether mock jurors use multiple-category reasoning when faced with uncertain alternatives. The design was similar to the first study except that all participants completed two versions of each trial vignette; one where the secondary category was related to the crime, and one where it was unrelated. Multiple-category reasoning would be indicated if different judgments about guilt are given in these conditions.

## Method

**Participants** Forty-two undergraduate students participated for course credit. The majority were female ( $n = 29$ ) and the mean age was 20.69 years ( $SD = 4.98$ ).

**Design and Procedure** The design of Experiment 2 followed the description given in Table 1. Unlike the previous study however, the status of the secondary category (related or unrelated to the crime) was manipulated within subjects. All participants completed two versions of each of the robbery and arson vignettes; one with a crime-related secondary category, and one with an unrelated secondary category. The order of presentation of these alternate versions was counterbalanced across participants. To reduce sequencing effects alternate versions of the same vignette were never presented consecutively. Before completing the study participants were warned that they would sometimes be reading summaries with similar details and were instructed to "do your best to evaluate the trial summaries independently".

As in the previous study the status of the primary category was manipulated between subjects. Equal numbers were randomly allocated to conditions in which the primary categories were related to the crime or were unrelated. In all other respects the procedure was identical to Experiment 1.

## Results

**Preliminary Analyses** Eleven participants failed to assign the highest categorization rating to the primary category in at least one scenario, and were excluded from further analyses. Preliminary analyses again found that for the remaining participants, ratings of the likelihood that the critical object belonged to the primary and secondary categories closely matched the probabilities stated or implied in the vignettes. Notably, there was no evidence of sequencing effects on guilt judgments for the secondary related and secondary unrelated versions of each vignette. The guilt estimates for these alternate versions were unaffected by which version was presented first ( $F < 2.0$ ). As in the previous study however, eyewitness reliability was rated higher in secondary related than the secondary

unrelated conditions,  $F(1,29) = 4.23, p < .05$ . Consequently, eyewitness reliability scores were again included as covariates in guilt analyses.

**Guilt Likelihood Ratings and Guilt Verdicts** Both variates were collapsed across scenarios and the binary verdicts were coded as per Experiment 1. Both data sets were entered into separate 2 (primary category status)  $\times$  2 (secondary category status) multivariate analyses of variance, with repeated measures on the second factor and eyewitness reliability scores entered as covariates. Figure 2 shows adjusted mean guilt likelihood ratings. Once again there was a significant main effect of primary category status,  $F(1,28) = 27.22, p < .001$ , with higher guilt estimates when the primary category was incriminating than when it was not. More importantly, in this case there was also a robust effect of secondary category status,  $F(1,28) = 36.46, p < .001$ , but no primary  $\times$  secondary category interaction, ( $F < 1$ ). When the secondary category was unrelated to the crime, ratings of guilt likelihood were lower than when that category incriminated the defendant. This shows that participants were factoring both primary and secondary categories into their guilt estimates.

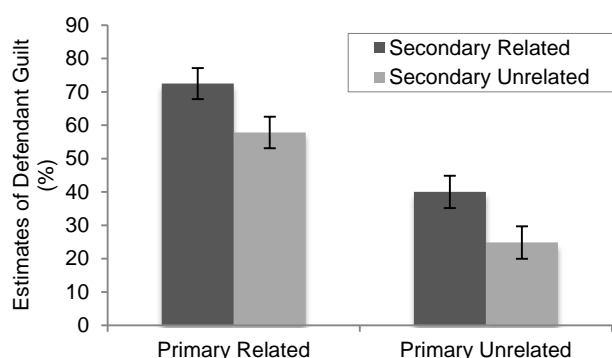


Figure 2. Mean likelihood ratings of defendant guilt (with standard error bars).

A similar pattern was observed for binary guilt verdicts. The proportion of guilty verdicts was higher when the primary category was related to the crime ( $M = 0.63$ ) than when it was unrelated ( $M = 0.13$ ),  $F(1,28) = 22.83, p < .001$ . Critically there was also a robust effect of secondary category status,  $F(1,28) = 8.33, p = .01$ . The proportion of guilty verdicts was significantly lower when the secondary category was unrelated to the crime ( $M = 0.30$ ) than when it was crime related ( $M = 0.46$ ). There was no primary  $\times$  secondary category interaction, ( $F < 1.0$ ).

These data indicate that mock jurors were considering both category alternatives when inferring guilt.

## General Discussion

The goal of these experiments was to examine whether mock jurors consider more than one alternative scenario when inferring defendant guilt on the basis of uncertain evidence. In our trial vignettes there was always some

uncertainty about the category membership of a critical piece of evidence, with a more likely (primary) and a less likely (secondary) category alternative. Mock jurors in both studies acknowledged this uncertainty, recognizing that the object *could* have belonged to the secondary category.

In Experiment 1 we found minimal evidence that jurors factored these uncertain alternatives into their guilt judgments but analyses of the critical comparisons did not find these differences to be reliable. Notably though when within-condition variance was reduced by manipulating secondary category status within-subjects (Experiment 2), robust evidence of multiple-category reasoning was found. When there was a possibility that the defendant was holding an innocuous rather than an incriminating object, participants reduced their ratings of guilt likelihood and were less like to return a verdict of guilty.

This is an important result because most previous work (e.g., Malt et al., 1995; Murphy & Ross, 2007, 2010; Ross & Murphy, 1996) has failed to find evidence that people consider more than one category when making predictions about objects whose category membership is uncertain. Ross and Murphy (1996) reported some evidence of multiple-category reasoning but only when the secondary category was more highly associated with the prediction than the primary category (equivalent to our primary unrelated condition). In Experiment 2 however, we found multiple-category reasoning in both primary related and primary unrelated conditions. This result is particularly interesting because it shows that consideration of the secondary category can lead to either decreases in the probability of a given prediction (in the Primary related conditions) or increases in prediction probability (in the Primary unrelated conditions).

One concern is that our strongest evidence of multiple-category reasoning was found when participants completed both the secondary related and secondary unrelated versions of the vignettes. Under these conditions participants could conceivably compare the structure of the two scenarios and this may have increased their sensitivity to the role of the secondary category in determining guilt. In other words, the strong evidence of multiple-category reasoning may have been an artifact of the repeated measures design. To examine this possibility we looked at participant guilt judgments the first time the vignettes were presented and then the second time the vignettes were presented as two separate sets of data. This order-artifact account predicts that we should only see evidence of multiple-category reasoning the second time that a participant sees the vignettes as it is the exposure to the first version that directs attention to the secondary category. Contrary to this account however, the effect of the secondary category on guilt ratings was robust on the first presentation on both the guilt likelihood ratings,  $F(1,26) = 7.70, p < 0.05$ , and the proportion of guilty verdicts,  $F(1,26) = 16.02, p < 0.01$ . In other words, the first time participants read the vignettes



they factored category uncertainty into their guilt ratings.<sup>1</sup>

So just why did we succeed in finding robust evidence in multiple-category reasoning in Experiment 2 when a majority of previous studies have failed to do so? Further research will be required to give a complete answer to this question. As noted earlier however, the highly consequential nature of forensic decisions may lead participants to be more reflective in their consideration of uncertain alternatives. An analogous finding is that expert clinicians have been shown to consider multiple uncertain categories but only when making clinically relevant predictions (Hayes & Chen, 2008). When required to make predictions about nonclinical materials they ignored category uncertainty.

Overall these studies provide qualified support for the conclusion that jurors can use multiple-category reasoning when making inferences about guilt. Clearly much further work is needed to be able to generalize these results to more realistic trial contexts where, for example, the explicit probabilities of different alternative scenarios are unlikely to be given. Nevertheless our findings suggest that, at least in some contexts, jurors can satisfy the legal imperative to consider multiple uncertain aspects of evidence.

### Acknowledgments

This work was supported by Australian Research Council Grant DP0770292 to the second author. We would like to thank Melissa Lim for her assistance with programming.

### References

- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98, 409-429.
- Attorney General's Department of NSW. (2007). *A guide for jurors*. Sydney, NSW: Court Services, Attorney General's Department of NSW.
- Bornstein, B. H., & Greene, E. (2011). Jury decision making: Implications for and from Psychology. *Current Directions in Psychological Science*, 20, 63-67.
- Gilovich, T. D., & Griffin, D. W. (2010). Judgment and decision making. In T. Susan & D. T. Gilbert (Eds.), *Handbook of social psychology, Vol 1 (5th ed.)* (pp. 542-588). Hoboken, NJ: John Wiley & Sons Inc.
- Lagnado, D. A. (2011). Thinking about evidence. *Proceedings of the British Academy*, 171, 183-223.
- Harris, A. J. & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 1366-1373.
- Hayes, B. K. & Chen, T-H. J. (2008). Clinical expertise and reasoning with uncertain psychodiagnoses. *Psychonomic Bulletin & Review*, 15, 1002-1007

<sup>1</sup> It should be noted that although this first presentation is similar in design to Experiment 1, different results were obtained. There is no clear explanation for this discrepancy. One possibility is that this was due to the new instructions used in Experiment 2, which strongly encouraged participants to attend the details of the vignettes.

- Hayes, B. K., Heit, E., & Swendsen, H. (2010). Inductive reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1, 278-292.
- Hayes, B. K. & Newell, B. R. (2009). Induction with uncertain categories: When do people consider the category alternatives? *Memory & Cognition*, 37, 730-743.
- Kuhn, D., Weinstock, M., & Flaton, R. (1994). How well do jurors reason? Competence dimensions of individual variation in a juror reasoning task. *Psychological Science*, 5, 289-296
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108, 480-498.
- Malt, B. C., Ross, B. H., & Murphy, G. L. (1995). Predicting features for members of natural categories when categorization is uncertain. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21, 646-661.
- Murphy, G. L., & Ross, B. H. (2007). Use of single or multiple categories in category-based induction. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches* (pp. 205-225). New York, NY: Cambridge University Press.
- Murphy, G. L., & Ross, B. H. (2010). Uncertainty in category-based induction: When do people integrate across categories? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 263-276.
- Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of Personality and Social Psychology*, 62, 189-206.
- Ross, B. H. & Murphy, G. L. (1996). Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 736-753.
- Solana, E. F., García, J., & Tamayo, I. (1998). Some individual differences in perception of the evidence and the verdict choice. *Psychology, Crime and Law*, 4, 361-373.

### Appendix

#### Excerpt from the Robbery Vignette used in Experiments 1-2 showing the primary categories (in bold) and secondary categories (in italics).

Ms. Kelly advised that the defendant Mr Bell was carrying a metallic looking object in his hands. She believed that the object was most likely a **KNIFE/ KEYS**. When asked to estimate her certainty Ms. Kelly advised that she was "reasonably certain, at least 70%". She advised that there was "a small chance" that the object was the *STOLEN WATCH/ a MOBILE PHONE*. In summary, Ms. Kelly testified that there was a small chance that the defendant was carrying the **STOLEN WATCH/ a MOBILE PHONE** but she believed the defendant was most likely carrying a *KNIFE/ KEYS*. Note that Ms. Kelly was sure that the defendant was only carrying one object.

# Elaborate Descriptive Information in Indoor Route Instructions

Vivien Mast (viv@tzi.de)

Cui Jian (ken@informatik.uni-bremen.de)

Desislava Zhekova (zhekova@uni-bremen.de)

I5-[DiaSpace], SFB/TR8 Spatial Cognition, University of Bremen  
Cartesium, Enrique-Schmidt-Straße 5, 28359 Bremen, Germany

## Abstract

The following paper presents the enhancement of indoor route instructions with descriptive generation strategies. We consider the latter to be highly important for the quality and helpfulness of automatically generated indoor route instructions. We conducted an experiment showing that participants receiving route instructions enriched with elaborate descriptive information instead of step-by-step procedural information for crucial route segments performed better in objective and subjective measures than those receiving only basic prescriptive route instructions. Based on the gained knowledge, we conclude that descriptive strategies are an important part of indoor route instructions and should be actively considered in system development.

**Keywords:** indoor route instructions; descriptive information; wayfinding; spatial cognition; navigation

## Introduction

Both navigation in indoor and outdoor environments profit from the use of landmarks since they are distinctive, easily recognizable and highly memorable (Sorrows & Hirtle, 1999). Humans select landmarks for their distinguishing characteristics (Presson & Montello, 1988). Although the importance of landmarks in route instructions is well established (Allen, 1997; Denis, 1997; Richter, 2007; Raubal & Winter, 2002), most research in automatic generation of route instructions focuses on one aspect of landmarks, namely to indicate the location at which a reorientation should take place in a network of paths. The main assumption of this approach is that good route instructions contain tightly coupled descriptive and prescriptive information. Therefore, current systems rely almost entirely on what Denis (1997) classified as Type 2 utterances – utterances coupling an action with a landmark. This leads to highly concise route instructions, but also limits the amount of descriptive information for each reorientation point to the mentioning of one landmark, possibly locating it with respect to the user.

While this approach is particularly useful for car navigation (Brenner & Elias, 2003) which occurs in network space, i.e. along a street network where clearly identifiable nodes (intersections) are connected by edges (streets), in pedestrian navigation the case is different. Pedestrian navigation includes many areas that belong to scene space: open areas which are characterized by the absence of clearly identifiable nodes and edges (Rüetschi, 2007; Schuldes et al., 2011). In network space, wayfinding consists mainly of selecting a path at each decision point, whereas in scene space, wayfinding is characterized by activities such as searching, exploring, and

matching. There are no clear paths to choose from, but large spaces, where piloting between landmarks is necessary. Oriented search might be used if the expected landmark cannot be seen (Allen, 1999). In such areas, route graph representations, and the resulting procedural information do not correspond very well to the needs of the wayfinder, as the function of landmarks changes from identifying a turning point to more vague orientational aid. Indoor navigation has elements of both network and scene spaces. In addition, indoor spaces are characterized by a very limited amount of different landmark types and a lack of highly salient landmarks. Usually landmarks consist mainly of doors, corridors and staircases, only very few of which are highly distinctive in comparison to outdoor landmarks which can be very diverse (a church, a petrol station, multiple intersections of different types, etc.). For this reason, the central roles of landmarks, i.e. signaling where actions should take place, as well as confirmation, are difficult to obtain in indoor scenarios. Additionally, this increases the difficulty of memorization, as it leads to instructions which contain a series of highly similar utterances.

A possible solution for these problems is the integration of more elaborate descriptive information into indoor route instructions. This can be realised by basing instructions on a scene space representation of space, and using a descriptive strategy for generating route instructions for those areas that can be characterized as scene space: Instead of superimposing abstract network representations onto open space areas, thereby producing a number of turning points and paths for an area which is viewed by a wayfinder as a coherent whole, this scene is described as one entity, and the location of the scene exit is described with respect to the scene. We assume that by introducing more elaborate descriptive information into indoor route instructions we can gain configurations of landmarks that can serve as highly salient landmarks, where simple landmarks will yield no sufficient differentiation. Moreover, we expect that the scene descriptions will enable more efficient localization of scene exits in the descriptions, minimizing the number of prescriptive statements. In contrast, the imposition of abstract networks onto open spaces will yield extra turns. We expect route instructions which integrate the descriptive approach to make it easier for participants to build up a mental image of the route in advance, leading to better memorization and increased confidence. In addition, mixing scene descriptions with prescriptive statements should yield more diverse route instructions, thereby additionally support-

ing memorization.

While our route instruction system has already been successfully evaluated (Cuayáhuatl, Dethlefs, Richter, Tenbrink, & Bateman, 2010), the goal of the present study is to explore the boundaries and potential for further development of the system by using a particularly difficult route which contains areas for which we consider the current (standard) approach lacking. For this purpose we first conduct and present an experiment which compares the wayfinding performance of participants receiving instructions, based either on solely procedural strategies as used by the system, or on a systematic mixture of procedural and descriptive strategies which the system currently does not provide. The results are presented and discussed with respect to the insights that can be gained for the development of route direction systems using natural language generation. In conclusion we propose directions that future research in the area could take.

### Experimental Setup

The experiment was conducted in GW2, a building at the University of Bremen which is notorious for its complexity. Each of the four floors has a different layout consisting of one or two main areas. The route (figure 1) was specifically chosen to be long and difficult, contain many turns and lead through a large portion of the 3rd level of the building. Secondly, it should contain two areas (A and B in figure 1) characterized by scene space rather than by network space. In both cases a diagonal crossing of the open area was necessary.

In our experiment the participants made use of an indoor route direction system called Infokiosk, developed as part of the I5-DiaSpace<sup>1</sup> project. Infokiosk (Cuayáhuatl et al., 2010) is a multimodal interactive spoken dialogue system for indoor wayfinding in complex buildings. It was developed based on a general computational dialogue system architecture and framework named DAISIE (Ross & Bateman, 2009), and can be described with the following three key components: 1) Dialogue management with a formal unified dialogue modeling approach combining information state update theories with generalized dialogue models (Shi, Jian, & Rachuy, 2011). 2) Route instruction generation with a combined computational model for generating unambiguous high-level context-specific route instructions (Richter, 2007). 3) Natural language generation with the probabilistic context-free representational underspecification framework (Belz, 2008) and the KPML natural language generation system (Bateman, 1997).

In the basic condition, the participants received route instructions generated automatically by Infokiosk. The instructions contained only procedural sentences in imperative mood, directly linking body turn actions to landmarks. An artificial route graph based on network space was superimposed onto the two open areas (figure 1, dotted grey lines), and they were described accordingly. Example (1) shows an instance of the instructions generated by the system in the basic condition for area A in figure 1:

<sup>1</sup><http://www.diaspace.org/>

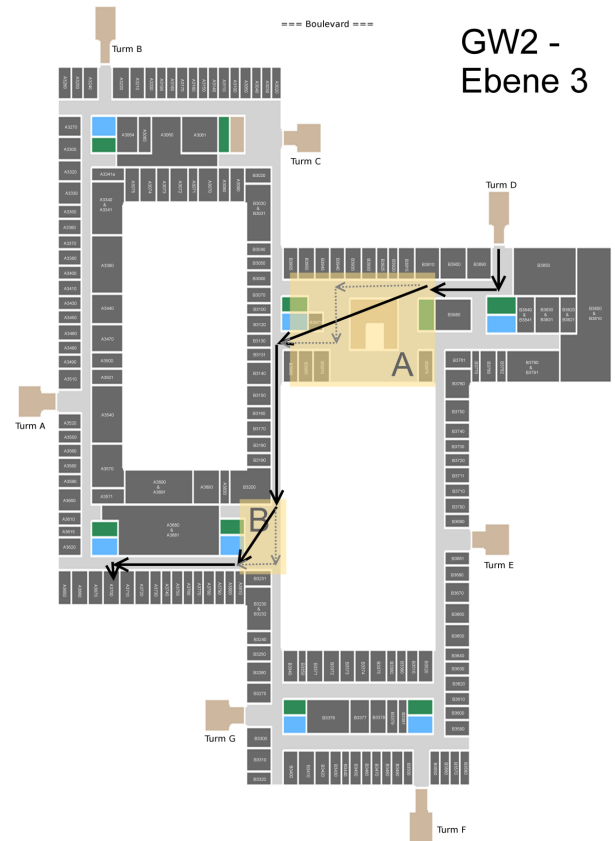


Figure 1: The selected route for the experiment. Grey dotted lines show the superimposed route graph for the basic condition.

- (1) "... go straight on until the third turning point on the left. Turn left, and go straight until the first glass door on the right. Turn to the right,..."

At the current stage, representations appropriate to the descriptive strategy, and the generation of the corresponding utterances are not supported by the Infokiosk. The integration of that capability in the system is not straightforward and should only be approached if such elements do show to improve wayfinding in indoor environments. Thus, in the descriptive condition, the participants received hand-crafted instructions by interaction with the dialogue system. In these instructions, imperative and declarative clauses were used, and the two scene space areas were described with full scene descriptions containing an introduction to the general structure of the area, one or more intermediary landmarks, or configurations of landmarks, and a localization of the goal exit with respect to the intermediary landmarks. Example (2) shows the instructions received for area A in figure 1 in the DC.

- (2) "... straight on until you reach a big hall area. In the middle of the hallway there is a staircase. Behind the

*staircase there are two glass doors which are partly hidden by concrete pillars. Go through the left door."*

It must be emphasized that the difference between the two conditions is not merely quantitative in nature, i.e. including a higher amount of descriptive information in the instructions. There is a qualitative difference, as the basic condition relies on superimposing an artificial route graph on areas characterized as scene space, giving a high amount of procedural information which precisely specifies the path for crossing an open area while reducing descriptive information to the minimum. In the descriptive condition, however, scene space areas are considered as holistic scenes with an entry point and an exit point. The scenes are described as seen from the entry point, and the position of the exit point is identified with respect to the described scene without any procedural information mentioning the path to take to get there. The descriptive instruction is only slightly more verbose (98 words, as compared to approx. 85 in the basic condition), but a different kind of information is chosen for verbalization.

## Participants

32 participants were tested. Two participants had to be excluded from the overall evaluation, because they found the goal by accident due to its location close to their position in the route. These participants explicitly stated that they expected the goal to be at a different position and were not using a search strategy, but saw the final goal by mere coincidence. The remaining 30 participants were used for evaluation, 15 in each of the two conditions (basic vs. descriptive). The participants were mostly first semester students at the University of Bremen. All had native or near-native competence of German and were between 19 and 31 years old (mean: 22). There were 21 female and 9 male participants.

The participants had little or no prior knowledge of the building. On a 7-point Likert scale, scores for the basic condition ranged from 1 to 4 (median 2; mean 2.3; standard deviation (sd) 0.88) while in the descriptive condition the range was from 1 to 3 (median 2; mean 1.6; sd 0.63). Scores for the basic condition were significantly higher than for the descriptive condition. (Wilcoxon rank sum test (one-sided):  $W = 161, p = 0.02$ )<sup>2</sup>.

The experiment was conducted with respect to a route on the 3rd level of the building. Most participants had been to the 3rd level either never, or less than five times before the experiment with no difference between the two groups (Wilcoxon rank sum test (two-sided):  $W = 123.5, p\text{-value} = 0.64$ ). Five participants (3 in the basic condition, 2 in the descriptive condition) reported having been to that particular floor more than five times.

There was no significant difference for spatial abilities between both conditions, as measured by the Questionnaire on Spatial Strategies (Münzer & Hölscher, 2011) ("Global self-

confidence, related to egocentric strategies", Wilcoxon rank sum test (two-sided):  $W = 88, p = 0.32$ ); "Survey strategy" (Wilcoxon rank sum test (two-sided):  $W = 99.5, p = 0.60$ ); "Knowledge of cardinal directions" (Wilcoxon rank sum test (two-sided):  $W = 74.5, p = 0.10$ ).

## Procedure

The overall procedure consisted of several steps that we describe in the current section.

First, participants were brought to the starting point via a route that did not cross the target route – they reached the starting point by entering the floor directly from the elevator.

Then, all participants were asked to fill in a short pre-questionnaire concerning age and prior knowledge of GW2.

After the pre-questionnaire the experimenter instructed the participants. Their goal was to find the room of a given person and for that they were only allowed to use the help of the Infokiosk. The participants were strongly encouraged to acquire the information solely using natural language. Handling of the microphone was explained in a short briefing. At this stage, participants were told that they should follow the route given to them by the system from their own memory, as far as this was possible. They were not informed that they would be able to recheck the instructions en-route. In this way, we enforced that they attempt full comprehension and memory of the route instructions in advance.

Right after the dialogue with the system, the participants were asked to answer 3 questions regarding perceived helpfulness of the system, confidence that they would find their goal, and how well they could visually imagine the route.

In the next step of the procedure they were instructed that they should follow the given route as closely as possible from memory, but that they could recheck a printout of the instructions as often as they wanted. The experimenter informed them that in case of doubt it was preferable to recheck the instructions than take a wrong turn.

The participants followed the route accompanied by the instructor. The experimenter did not answer any clarification questions (except for the initial perspective from which the instructions were given, i.e. "the system has explained the route as from the position in which you were seated when you received the instructions"), but whenever participants were indicating they were unsure about the route to follow, asked clarification questions, or explicitly stated that they were lost or had forgotten the instructions, the experimenter informed the participants that it was no problem at all to recheck the instructions as often as they wanted. Whenever participants explicitly stated doubts that they were going to be able to find the goal, the experimenter informed them that it was possible for them to give up if they wanted to.

At last, after finding the goal (or giving up), the participants were asked to give a retrospective report of their wayfinding and any doubts or problems that had occurred at the decision points. They were brought back to the starting point to fill in a final questionnaire about their performance and the perceived

<sup>2</sup>As the data was not normally distributed, a non-parametric test was chosen. The same applies for all the following statistical tests where non-parametric methods were chosen.

Table 1: Participants' task success in both conditions.

	participants	
	BC	DC
failed	7	0
succeeded	8	15

helpfulness of the system, and the Questionnaire on Spatial Strategies by Münzer and Hölscher (2011).

## Hypotheses

With respect to objective performance measures we expected the descriptive condition to improve task performance, yielding a higher task success rate, a lower number of wrong route segments travelled and a lower number of total route segments travelled. We also expected the descriptive condition to improve memorization, leading to less frequent consultation of instruction printouts.

With respect to subjective performance measures, the descriptive condition was expected to lead to higher self-ratings for confidence of being able to find the goal, subjective helpfulness of the instructions, and mental imagery.

## Annotation

In order to evaluate task success, the experimenter followed the participants in the wayfinding task and made a protocol of the path travelled and any instruction consultations. With respect to the used annotation scheme, the path was divided into segments, where each segment consists of the path between two decision points. At the end, every wrong segment that the participant travelled was counted. Only those segments were counted where the participant either travelled a segment that was not part of the intended route, or walked in the wrong direction along a segment that was part of the intended route. Wrong segments that were travelled several times in a task were counted several times accordingly. In order to be able to count the number of instruction consultations, participants were only given the instructions if they explicitly requested this and were not allowed to move while holding the instruction sheet. If participants had not moved at all between two consultations, this was counted as one consultation.

## Results

### Task Success

Task success is an objective performance measure indicating the number of participants that managed to find the target. As indicated in table 1, in the basic condition 7 out of 15 participants for that setting did not find the goal at all. Yet, in the descriptive condition all 15 participants managed to find the goal. The resulting difference is significant (Pearson's Chi-squared test with Yates' continuity correction:  $X^2 = 6.7081$ ,  $df = 1$ ,  $p = 0.01$ ). This result strongly supports our hypothesis that enriching route instructions with descriptive information significantly improves wayfinding success.

### Wrong Segments Traveled

Another objective performance measure that we considered is the number of wrong segments traveled during the wayfinding process. In the descriptive condition there were overall 82 wrongly traveled segments, with single participants travelling from 0 to 31 wrong segments (mean: 5.47, sd: 8.51), while in the basic condition there were 169 wrongly travelled segments altogether, single participants' error scores ranging from 0 to 27 (mean: 11.27, sd: 7.34).

In the descriptive condition, one participant contributed over 35% of overall wrong segments (31 out of 82) due to an exceptional misunderstanding that was not reproduced by any of the other participants. But even including this exceptional case, there remains a significant difference, indicating that participants receiving the basic instructions travelled more wrong segments than those receiving the descriptive instructions (Wilcoxon rank sum test one-sided, with continuity correction:  $W = 166.5$ ,  $p = 0.01$ ).

### Number of Instruction Consultations

The number of instruction consultations during wayfinding is as well an objective performance measure. It indicates how easy the given system instructions were to understand and memorize. The participants in the basic condition consulted the instructions en-route 74 times overall, ranging from 1 to 10 consultations (mean: 4.93, sd: 2.76) while the participants in the descriptive condition rechecked the instructions only 31 times, individual scores ranging from 0 to 6 (mean: 2.07, sd: 1.53). Participants in the basic condition consulted the instructions significantly more often than those in the descriptive condition (Wilcoxon rank sum test one-sided, with continuity correction:  $W = 180.5$ ,  $p = 0.002$ ).

### Confidence

As subjective performance measure we considered the participants' confidence. Immediately after they had finished the dialogue with the system, participants were asked how confident they felt that they would find their goal. This was before they were informed that they would be able to consult a printout of the instructions en-route. For this reason, confidence levels were generally fairly low, reflecting the difficulty of the task. In the basic condition, they ranged from 1 to 5 on a 7-point Likert scale (median: 3, mean: 3.13, sd: 1.30), while ranges in the descriptive condition were from 1 to 6 (median: 4, mean: 4.13, sd: 1.41). Confidence in the basic condition was significantly lower than in the descriptive condition (Wilcoxon rank sum test with continuity correction, one-sided:  $W = 70$ ,  $p = 0.04$ ).

### Mental Imagery

The level of mental imagery across all participants in both conditions shows the effect of the descriptions in both conditions on the participants' capability to envision the environment. Asked directly after the dialogue with the system, how well they could visually imagine the described route, participants in the basic condition gave scores from 1 to 5 on a

7-point Likert scale, (median: 3, mean: 2.93, sd: 1.39). Participants in the descriptive condition gave scores from 1 to 7 (median: 5, mean: 4.6, sd: 1.59). The difference is highly significant, indicating that participants in the descriptive condition could visually imagine the route better than those in the basic condition (Wilcoxon rank sum test, one-sided:  $W = 49.5$ ,  $p\text{-value} = 0.004$ ).

### Perceived helpfulness of the instructions

The scores participants gave for the helpfulness of the instructions given by the system show an interesting effect. Both conditions were perceived as equally helpful directly after receiving the instructions - in both conditions, scores ranged from 2 to 7 on a 7-point Likert scale with a median of 5 (basic condition: mean: 4.6, sd: 1.55; descriptive condition: mean: 4.73, sd: 1.58; Wilcoxon rank sum test with continuity correction, two-sided:  $W = 110$ ,  $p\text{-value} = 0.47$ ). After navigating the route, however, this changed. In the final questionnaire, scores for helpfulness in the basic condition ranged from 1 to 7 with a median of 3 (mean: 3.54, sd: 1.81), constituting a significant drop in comparison with pre-navigation scores (Wilcoxon rank sum test with continuity correction, one-sided:  $W = 155.5$ ,  $p\text{-value} = 0.036$ ).

For the descriptive condition, on the other hand, they stayed at the same high level, ranging from 3 to 7 with a median of 5 (mean: 5.27, sd: 1.49). Thus, the perceived helpfulness after navigation is significantly higher in the descriptive condition than in the basic condition (Wilcoxon rank sum test with continuity correction, one-sided:  $W = 51$ ,  $p = 0.005$ ).

## Discussion

The results clearly show that descriptive strategies can improve wayfinding in indoor environments. Participants in the descriptive condition had a higher success rate and walked the route with less wrong segments traveled than those in the basic condition. They also needed to consult the instructions less often. This is most likely due to the fact that the different structure of the environment, as compared to street networks, leads to differences in wayfinding strategies, and therefore different needs with respect to route instructions.

An important finding of this experiment is, that descriptive elements not only improve objective performance measures, but also subjective ones. The improvement of participants' confidence and mental imagery in the descriptive condition is a factor that is important for cognitively ergonomic route instructions. Humans should not only find their goal with automatically generated instructions, they should also feel comfortable and secure while doing so. The scores for perceived helpfulness of the system show an interesting effect: before wayfinding participants rate the instructions as equally helpful in both conditions, which is in contrast to the other subjective measures. After wayfinding the values change, resulting in a significantly higher value for the descriptive condition, matching objective performance and the other subjective measures. This might be due to the fact that before performance participants were not as secure about their quality

judgement as after, and possibly answers were influenced by their wish to be polite.

There are several reasons that might account for the better performance of subjects that were given the descriptive instructions. Firstly, the significantly higher values for mental imagery before setting off suggest that better mental imagery might be one of the factors that helped participants find their way more easily. Visuo-spatial imagery is an important factor in understanding and memorizing route directions (Denis & Fernandez, submitted). In addition, successful mental imagery involves deep semantic processing and the formation of a coherent situation model which have been shown to improve memory performance (Craig & Tulving, 1975; Kintsch, 1994). The greater difficulty of participants in the basic condition to visually imagine the route in advance indicates that these participants were not able to construct as good a situation model as those in the descriptive condition.

Another central aspect that was verified by statements of several participants in the retrospective reports is that configurations of landmarks can improve error-recovery and confidence en-route, acting as substitutes for highly-salient landmarks which rarely exist in indoor environments.

Finally, it is probable that the highly repetitive style which results from generating only prescriptive utterances yields a Ranschburg effect: The occurrence of several tokens of the same type in the input within a short time is known to have a negative effect on memorization (Jahnke & Bower, 1986; Kanwisher, 1987, compare). The Ranschburg effect has mainly been studied in series of unrelated numbers or words, but it is highly probable that the underlying mechanisms have an effect on the memorization of a series of highly similar sentences containing repeated instances of certain words, as seen in the basic condition of this experiment. The more varied linguistic structure and semantic content of the descriptive instructions neutralize this effect, thereby improving memorization.

## Conclusion and Future Work

Our work shows that the use of elaborate descriptive information into indoor route instructions can significantly improve the quality of automatically generated instructions. The reported results indicate that both objective and subjective performance measures rank the use of descriptive strategies higher than the condition in which only the prescriptive strategy was used. It needs to be shown, however, that the improvement remains significantly large when using computer-generated instructions based on the descriptive approach.

Also, buildings differ with respect to their structure. While this approach may be very useful for buildings that contain a high proportion of open spaces, it may not be necessary for buildings that consist entirely of long and narrow corridors with clearly identifiable intersections and can therefore be represented sufficiently by network space. It would be insightful to compare the two different approaches over a wider variety of routes in order to investigate how the two strategies

can best be combined, and how they interrelate with issues of conciseness: How much descriptive information is necessary, and at which points in a route should this type of information be provided? How do route length and dominance of scene space characteristics interact to favor one or the other type of instructions? It should also be examined whether the findings hold for pedestrian navigation in general.

Although we have hinted at some mechanisms that might underlie the performance improvements, a more detailed analysis of these mechanisms should be undertaken, in order to be able to clearly distinguish which aspects of the descriptions improve comprehension and memorization in which ways.

Natural language route direction systems for indoor (and pedestrian) navigation should take these results into account and find ways of modeling spatial information that allow for a more flexible combination of prescriptive and descriptive information.

### Acknowledgements

This research was supported by the SFB/TR 8 Spatial Cognition (Deutsche Forschungsgemeinschaft, DFG). We would also like to thank the I5-[DiaSpace] project group, and especially Thora Tenbrink for support and insightful discussions.

### References

- Allen, G. (1997). From Knowledge to Words to Wayfinding: Issues in the Production and Comprehension of Route Directions. In S. Hirtle & A. Frank (Eds.), *Spatial Information Theory A Theoretical Basis for GIS* (pp. 363–372). Berlin, Heidelberg: Springer.
- Allen, G. (1999). Spatial Abilities, Cognitive Maps, and Wayfinding: Bases for Individual Differences in Spatial Cognition and Behavior. In R. G. Golledge (Ed.), *Wayfinding Behavior* (pp. 46–80). Baltimore, London: John Hopkins University Press.
- Bateman, J. A. (1997). Enabling Technology for Multilingual Natural Language Generation: the KPML Development Environment. *Journal of Natural Language Engineering*, 3(1), 15–55.
- Belz, A. (2008). Automatic Generation of Weather Forecast Texts Using Comprehensive Probabilistic Generation-Space Models. *Natural Language Engineering*, 1, 1–26.
- Brenner, C., & Elias, B. (2003). Extracting Landmarks for Car Navigation Systems Using Existing GIS Databases and Laser Scanning. In H. Ebner, C. Heipke, H. Mayer, & K. Pakzad (Eds.), *Proceedings of Photogrammetric Image Analysis* (pp. 131–136).
- Craik, F. I. M., & Tulving, E. (1975). Depth of Processing and the Retention of Words in Episodic Memory. *Journal of Experimental Psychology: General*, 104, 268–294.
- Cuayáhuít, H., Dethlefs, N., Richter, K.-F., Tenbrink, T., & Bateman, J. (2010). A Dialogue System for Indoor Wayfinding Using Text-Based Natural Language. *International Journal of Computational Linguistics and Applications*, 1(1-2), 285–304.
- Denis, M. (1997). The Description of Routes: A Cognitive Approach to the Production of Spatial Discourse. *Cahiers Psychologie Cognitive*, 16(4), 409–458.
- Denis, M., & Fernandez, G. (submitted). The processing of Landmarks in Route Directions. In T. Tenbrink, J. Wiener, & C. Claramunt (Eds.), *Representing Space in Cognition: Interrelations of Behavior, Language, and Formal Models*.
- Jahnke, J. C., & Bower, R. E. (1986). Are There Two Ranschburg Effects? *The American Journal of Psychology*, 99(2), 275–288.
- Kanwisher, N. G. (1987). Repetition Blindness: Type Recognition Without Token Individuation. *Cognition*, 27, 117–143.
- Kintsch, W. (1994). Text Comprehension, Memory, and Learning. *American Psychologist*, 49, 294–303.
- Münzer, S., & Hölscher, C. (2011). Entwicklung und Validierung eines Fragebogens zu räumlichen Strategien (Development and Validation of a Self-report Measure of Environmental Spatial Strategies). *Diagnostica*, 57(3), 111–125.
- Presson, C. C., & Montello, D. R. (1988). Points of Reference in Spatial Cognition: Stalking the Elusive Landmark. *British Journal of Developmental Psychology*, 6(4), 378–381.
- Raubal, M., & Winter, S. (2002). Enriching Wayfinding Instructions with Local Landmarks. In M. Egenhofer & D. Mark (Eds.), *Geographic Information Science* (p. 243–259). Berlin, Heidelberg: Springer.
- Richter, K.-F. (2007). A Uniform Handling of Different Landmark Types in Route Directions. In *Proceedings of the 8th International Conference on Spatial Information Theory* (pp. 373–389). Berlin, Heidelberg: Springer.
- Ross, R. J., & Bateman, J. A. (2009). Daisie: Information State Dialogues for Situated Systems. In V. Matouek & P. Mautner (Eds.), *Text, Speech and Dialogue* (pp. 379–386). Berlin, Heidelberg: Springer.
- Rüetschi, U.-J. (2007). *Wayfinding in Scene Space: Modelling Transfers in Public Transport*. phd-thesis, University of Zürich.
- Schuldes, S., Boland, K., Roth, M., Strube, M., Krömker, S., & Frank, A. (2011). Modeling Spatial Knowledge for Generating Verbal and Visual Route Directions. In A. König, A. Dengel, K. Hinkelmann, K. Kise, R. J. Howlett, & L. C. Jain (Eds.), *KES(4)* (pp. 366–377). Springer.
- Shi, H., Jian, C., & Rachuy, C. (2011). Evaluation of a Unified Dialogue Model for Human-Computer Interaction. *International Journal of Computational Linguistics and Applications*, 2(1-2).
- Sorrows, M. E., & Hirtle, S. C. (1999). The Nature of Landmarks for Real and Electronic Spaces. In C. Freksa & D. Mark (Eds.), *Spatial Information Theory* (pp. 37–50). Berlin: Springer.



# Connectionist Model Accounting for Retardation of Cognitive-Dissonance Reduction Caused by Attention-Focus Switching

Takao Matsumoto (matsumoto@c.dendai.ac.jp)

Tokyo Denki University  
5 Senju-Asahi-cho, Adachi-ku, Tokyo 120-8551, Japan

## Abstract

A novel connectionist model accounting for cognitive dissonance is described, in which the concepts of self and attention are considered. The model makes it possible to use mathematical formulas to represent the cognitive-dissonance process. Analysis reveals that the model fits experimental data of major paradigms in cognitive dissonance theory and that attention-focus switching causes building-up of cognitive dissonance and retardation of its reduction.

**Keywords:** cognitive dissonance; connectionist model; attention; self-concept; mathematical analysis

## Introduction

Cognitive dissonance theory insists that dissonance is a psychological state of tension that people are motivated to reduce (Festinger, 1957). Dissonance causes feelings of discomfort, unhappiness, or distress. Any two cognitions are dissonant when one of them follows from the obverse of the other. To reduce dissonance, people add consonant cognitions or change evaluations for one or both cognitions to make them more consistent.

Cognitive dissonance theory makes a clear prediction when a firm expectancy is involved as one of the cognitions in question (Aronson, 1969). A well-known example of this is the famous Aesop's fable "The fox and the grapes." In the story, a fox wanted to get some grapes hanging high on vines and leaped with effort, but couldn't get them. Walking away, the fox said, "The grapes are surely sour, and I do not need them." Since the expectation and experience were inconsistent, the fox had cognitive dissonance, which he reduced by convincing himself that the expectation was not appropriate.

Shultz and Lepper (1996) proposed a connectionist model accounting specifically for the mechanism of cognitive dissonance. A constraint satisfaction neural network model was used to simulate data from the several major cognitive dissonance paradigms (Shultz, Leveille, & Lepper, 1999). In it weights between nodes are fixed and activations of units are changed. Dissonance is defined by a formula that is a function of activations of units and weights applied to links in the network. Networks tend to settle into a less dissonant state as activations of units are changed according to update rules. Another connectionist model was proposed by Van Overwalle and colleagues (2002, 2005). They represented attitudes in a feed-forward neural network with the delta-learning rule in which weights are allowed to change. Input nodes represent the features of the environment and two

output ports represent behavior and affect. Dissonance is defined as the discrepancy between expected and actual outcomes. They also simulated the experimental results of major cognitive dissonance paradigms. Several other computational models have been reported that deal with attitude phenomena through simulation using constraint-satisfaction or non-constraint-satisfaction networks (Mosler et al., 2001; Petty & Cacioppo, 1986; Read & Miller, 1994; Spellman, Ullman, & Holyoak, 1993).

People are motivated to prioritize to protect their self-system. Self-consistency theory (Aronson, 1969; Thibodeau and Aronson, 1992) emphasizes that self is involved in dissonance arousal and that not only cognitions but also self-concept need to be considered in discussing dissonance. Judgment and assessment for cognitions performed by self possibly become motives for arousal of cognitive dissonance.

On the other hand, attention is an important phenomenon of information processing in cognitive systems (Pisapia, Repovs, & Braver, 2008). It is a function for selecting and enhancing a limited area of information, while suppressing other areas. Cognitions are included in these areas of information.

To the author's knowledge, connectionist models for cognitive dissonance taking the concepts of self and attention into account have not been presented. In this paper, a novel model considering these concepts is described and reduction of cognitive dissonance based on the model is discussed.

## Connectionist Model

Figure 1 shows our connectionist model accounting for cognitive dissonance. In accordance with the cognitive dissonance theory (Festinger, 1957), two cognitions are adopted in the model, which are depicted as units R and I. We assume that unit R is a reality-based cognition such as the cognition of behavior, experience, or actual situation, while unit I is an imagination-based cognition such as the cognition of expectancy, hope, or belief. In the case of the previously mentioned fable, the imagination-based cognition held by the fox is "The grapes can stave off my hunger" and the reality-based cognition is "The grapes do not stave off my hunger."

Attention plays an important role in cognitive systems. It selects and enhances limited cognitions and suppresses others. If we consider two cognitions alone, when one of them is selected and enhanced, the other is rejected and

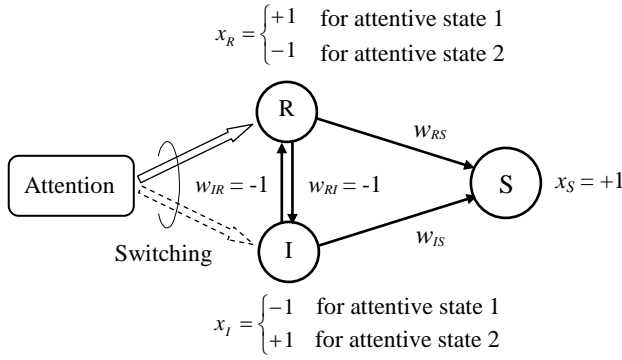


Figure 1: Connectionist model adopting the concepts of attention and self.

suppressed. Thus, the units of these cognitions perform a bistable operation. In order to achieve this operation, the units must be bidirectionally connected through links having negative weights as shown in Figure 1 (Rojas, 1996). In our model, to simplify the discussion, it is assumed that the activation level of units R and I is “+1” or “-1”, which correspond respectively to excited and inhibited state, and the weights between the units, i.e.,  $w_{RI}$  and  $w_{IR}$ , are -1. The bistable operation is determined by inputs to the unit pair applied through some additional links connected to the unit pair. The inputs’ condition depends on the attentive state. Since attention is usually focused on the cognition of unit R, the bistable operation is not perfectly symmetrical. There might be some bias difference between the inputs to the unit pair. In order to pick up the elements closely related to cognitive dissonance, such additional links are excluded from Figure 1.

Aronson (1969) emphasized that dissonance theory can make the clearest prediction of cognitive dissonance when we deal with the self-concept and cognitions about some behavior. Self is regarded as the key for arousing cognitive dissonance. In accordance with this view, we introduce a unit corresponding to self in our connectionist model, which is depicted as unit S in Figure 1.

Self is a complex system in which cognitive and affective elements are integrated. It is extremely difficult to rigorously represent the self in a simple connectionist model. Here, we assume that self is characterized by static equilibrium, in which there is resistance to incoming information that would change the status of elements (Nowak et al., 2000), and that, as pointed out in many studies of cognitive dissonance, people have a high or positive self-concept. Thus, we boldly introduce a single unit for self whose activation level is constant and takes a value of +1.

Cognition is composed of several elements having different attributes, each of which is related to the self with some evaluation given by the person. Cumulative evaluation

of all elements for a cognition is assumed to be the evaluation of the cognition. In our connectionist model, such an evaluation of cognition is represented by the weight of the link between the unit of the cognition and that of self. If the cognition is attractive for self, the evaluation or weight is positive, and if not, negative. The links are assumed to be unidirectional with directions from units R and I to unit S. The weight of the link between units R and S is  $w_{RS}$  and that between units I and S is  $w_{IS}$ . These weights take values between -1 and +1.

When we respectively represent the activation levels of units R, I, and S by  $x_R$ ,  $x_I$ , and  $x_S$ , the model’s cognitive dissonance can be described according to the definition given by Shultz and Lepper (1996) as follows:

$$CD = (2x_R x_I - w_{RS} x_R x_S - w_{IS} x_I x_S) / 4. \quad (1)$$

Focus of attention determines the state of the bistable operation of the unit pair composed of units R and I. We define attentive state 1 as the state in which attention is focused on the cognition of unit R and  $(x_R, x_I) = (+1, -1)$ , while attentive state 2 is defined as the state in which attention is focused on the cognition of unit I and  $(x_R, x_I) = (-1, +1)$ . Using Equation (1) and the previously-mentioned assumption that  $x_S = +1$ , cognitive dissonances for the two states are described as

$$CD = \begin{cases} (w_{IS} - w_{RS} - 2) / 4, & \text{for attentive state 1,} \\ (w_{RS} - w_{IS} - 2) / 4, & \text{for attentive state 2.} \end{cases} \quad (2)$$

## Analysis of Evaluation Change

In neural networks, the weight of a connection between two neurons changes according to the activation condition of the neurons. When the two neurons are excited simultaneously, the weight of the link between them increases and when they are not decreases. We assume that the weights in our connectionist model perform similarly to those of neural networks. The modified Hebbian learning rule presented by Oja (1982) incorporates the saturation characteristics of neurons into the original Hebbian rule. It represents the changes in weight as a function of the activation levels of input and output units, the weight between them, and a constant representing the learning rate during a time interval (O’Reilly & Munakata, 2000).

When we assume that the time interval is infinitesimal, the rule can be represented by the following differential equation:

$$\frac{dw}{dt} = \varepsilon' (xy - y^2 w), \quad (3)$$

where  $\varepsilon'$  is a constant representing the learning rate during a unit time interval.

In order to consider the behavior of  $w_{RS}$  we substitute  $w_{RS}$  for  $w$  in Equation (3). Since  $x = x_R = +1$  and  $y = x_S = +1$  for

attentive state 1 and  $x = x_R = -1$  and  $y = x_S = +1$  for attentive state 2, we obtain general solutions of Equation(3) as follows:

$$w_{RS} = \begin{cases} K_1 e^{-\varepsilon t} + 1 & , \text{ for attentive state 1,} \\ K_2 e^{-\varepsilon t} - 1 & , \text{ for attentive state 2.} \end{cases} \quad (4)$$

For  $w_{IS}$ , similar to the above, we can derive following solutions:

$$w_{IS} = \begin{cases} K_3 e^{-\varepsilon t} - 1 & , \text{ for attentive state 1,} \\ K_4 e^{-\varepsilon t} + 1 & , \text{ for attentive state 2.} \end{cases} \quad (5)$$

$K_1, K_2, K_3$ , and  $K_4$  are constants determined by initial conditions. Since the weights represent evaluations as previously mentioned, Equations (4) and (5) represent the time dependence of evaluations for cognitions R and I.

### Comparison with Experimental Results

To confirm the validity of our connectionist model and analysis, here we take up the free-choice paradigm as the first example, and compare theoretical results of our analysis and experimental results reported in the literature. In an experiment carried out by Brehm(1956), subjects were asked to rate each of a variety of items on desirability. They were next required to make a difficult choice, i.e., a choice between two items that they had rated high, or an easy choice, i.e., a choice between one item they had rated high and one they had rated low. The chosen items were given to the subjects who then rated them again. The experimenter measured the differences between the first and second ratings.

Applying our model to the experiment, we regard units R and I as the chosen and rejected items, respectively. This is because after a choice is made, its result is the reality for the subject and the chosen item becomes the element of the reality-based cognition, while the rejected item becomes that of the imagination-based cognition. Weights  $w_{RS}$  and  $w_{IS}$  are evaluations of the two items. We assume that the choice is carried out at  $\varepsilon t = 0$ . Weights' initial values at  $\varepsilon t = 0$  used in our theoretical examination are presented in Figure 2. The theoretical values are determined by making their maximum and minimum possible values, i.e., +1.0 and -1.0, correspond to the experimental maximum and minimum evaluations used in rating, i.e., 8 and 1, respectively. The values in the experiment and those in the theory are linearly related, and their correspondence is schematically shown in Figure 2. Since we consider the situation on the basis of reality, transitions of the weights or evaluations are calculated for attentive state 1.

Figure 3 shows our theoretical results at  $\varepsilon t = 0.7$  as well as the experimental data reported by Brehm(1956). Our theoretical results show that for both the difficult and easy

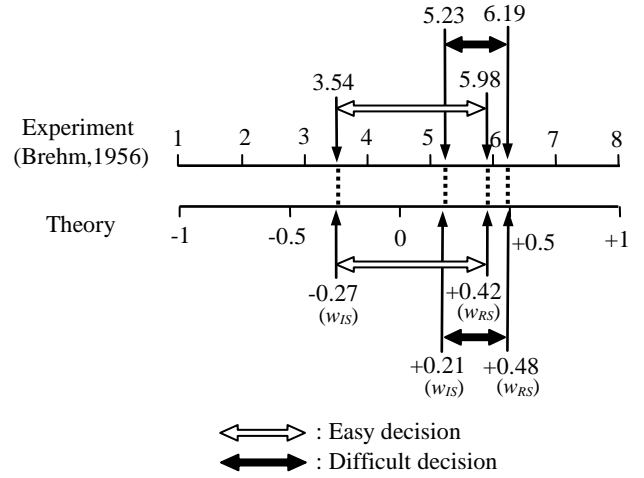
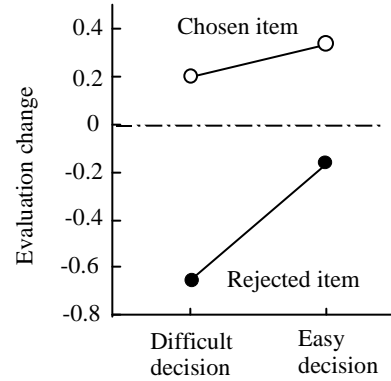
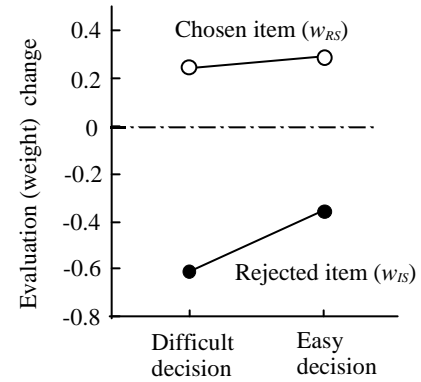


Figure 2 : Correspondence of the initial values of evaluation.



(a) Experiment (Brehm, 1956)



(b) Theory based on our model at  $\varepsilon t = 0.7$

Figure 3: Evaluation change in free-choice paradigm.

choices,  $w_{RS}$  increases and  $w_{IS}$  decreases, i.e., the separation between them increases, with the increase of time after the choice. The degree of separation increase is substantial for the difficult choice. Our results in Figure 3(b) are similar to the experimental data depicted in Figure 3(a).

We take up the insufficient-justification paradigm as the second example. In an experiment carried out by Freedman (1965), school children were forbidden to play with a desirable toy under either a mild or severe threat and the experimenter either stayed in or left the room while the children played. Actual play with the previously forbidden toy later indicated that derogation was greater under the mild than under the severe threat conditions only when there was no surveillance.

Applying our model to this experiment, we assume that units R and I are respectively the cognitions “I do not play with the toy” and “I play with the toy”. This is because, since the children are forbidden to play with the toy and are ordered to obey the directions, their reality-based cognition is “I do not play with the toy.” Weights  $w_{RS}$  and  $w_{IS}$  are respectively evaluations for not playing with and playing with the toy.

Here, we assume that weight  $w_{IS}$  is a compound of three subweights  $w_0(t)$ ,  $w_t$ , and  $w_s$  as shown in Figure 4. The intrinsic evaluation given by the subject for the cognition of unit I is  $w_0(t)$  ( $-1 \leq w_0(t) \leq 1$ ). Since the evaluation may change, it is represented by a function of  $t$ . The subweights  $w_t$  and  $w_s$  are additional weights caused respectively by the effects of threat and surveillance. Since the effects are independent of time,  $w_t$  and  $w_s$  are constants.

As mentioned above, we assume that the weights are bounded and take values between -1 and +1. Thus, we represent  $w_{IS}$  as follows:

$$w_{IS} = \begin{cases} w_0(t) + w_t + w_s, & \text{for } -1 \leq w_0(t) + w_t + w_s \leq +1, \\ -1, & \text{for } w_0(t) + w_t + w_s < -1, \\ +1, & \text{for } w_0(t) + w_t + w_s > +1. \end{cases} \quad (6)$$

Considering the situation on the basis of reality, we assume the attentive state is 1. Since threat and surveillance are unpleasant for the subject and thus function as negative elements of the cognition represented by unit I, we assume

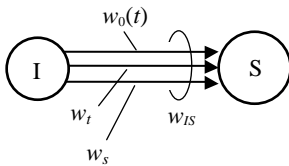


Figure 4: Weights in the connection between units I and S for insufficient-justification paradigm.

Table 1: Weights used in the theory for insufficient-justification paradigm.

	Non-surveillance	Surveillance
Mild threat	$w_0(0) = +1.0$	$w_0(0) = +1.0$
	$w_t = -0.2$	$w_t = -0.2$
	$w_s = 0$	$w_s = -2.0$
Severe threat	$w_0(0) = +1.0$	$w_0(0) = +1.0$
	$w_t = -1.8$	$w_t = -1.8$
	$w_s = 0$	$w_s = -2.0$

$w_t < 0$  and  $w_s < 0$ . Applying Equation (6) to the theory described in the previous section, we obtain

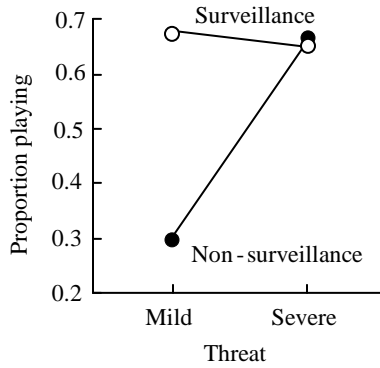
$$w_0(t) = \begin{cases} w_0(0), & \text{for } w_0(0) + w_t + w_s < -1, \\ (w_0(0) + w_t + w_s + 1)e^{-\varepsilon t} - w_t - w_s - 1, & \text{for } w_0(0) + w_t + w_s \geq -1. \end{cases} \quad (7)$$

Table 1 presents the values of the weights (evaluations) used in our theoretical examination. Since the toy might be very attractive for the subjects, we assume that  $w_0(0)$  takes the maximum value, i.e., +1.0. As surveillance seems to be effective for forbidding and induce highly negative feeling in the subject, the absolute value of  $w_s$  under surveillance is assumed to be large so that  $w_{IS}$  takes the minimum value, i.e., -1.0.

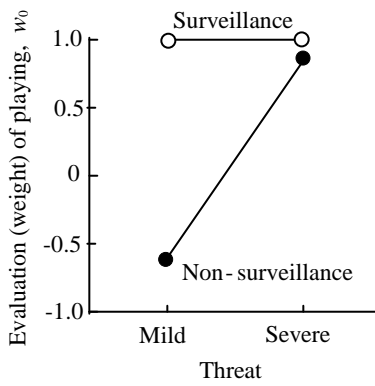
Transitions of the weights or evaluations are calculated for attentive state 1. Figure 5 shows our theoretical results for  $w_0(t)$  at  $\varepsilon t = 2$  as well as the experimental data reported by Freedman (1965). The theoretical result shown in Figure 5(b) indicates that derogation is greater under the mild than under the severe threat conditions only when there was no surveillance, which is similar to Freedman's experimental result.

## Attention-focus Switching

Here, following the Aesop's fable, we anticipate a case in which a person has a certain expectation concerning something, and he fails to realize it in spite of his effort. The object of expectation is the cognition of unit I in Figure 1, and the result of failure is the cognition of unit R. Since the expected object is attractive for the person and causes positive feeling, the cognition of unit I is positively evaluated and thus we assume  $w_{IS} > 0$ . In contrast, since the failure result is disagreeable and causes negative feeling, the cognition of unit R is negatively evaluated and thus we assume  $w_{RS} < 0$ . Since cognitive dissonance occurs when we experience the failure and thus attention is focused on the cognition of unit R, we consider here the situation in



(a) Experiment (Freedman, 1965)



(b) Theory based on our model at  $\varepsilon't = 2.0$

Figure 5: Evaluation change in insufficient-justification paradigm.

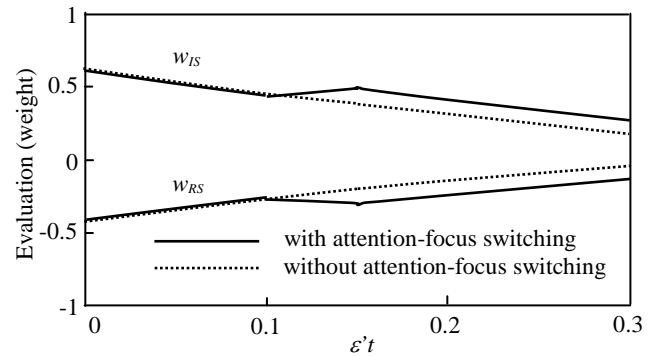
attentive state 1. Assuming that, as an example,  $w_{IS} = +0.6$  and  $w_{RS} = -0.4$  are the initial conditions at  $\varepsilon't = 0$ , we can obtain  $w_{IS}$ ,  $w_{RS}$ , and cognitive dissonance ( $CD$ ) at  $\varepsilon't \geq 0$  by using Equations (2), (4), and (5), which are shown by the dotted lines in Figure 6. The abscissa is  $\varepsilon't$ , which is a parameter proportional to the time passed after the person made the trial and failed.

The decrease of  $w_{IS}$  and increase of  $w_{RS}$  with the increase of time shown in Figure 6(a) indicate adaptation or rationalization of the person under cognitive dissonance. In accordance with these changes,  $CD$  is reduced with the increase of time as shown in Figure 6(b). Such changes appeared in  $w_{IS}$ ,  $w_{RS}$ , and  $CD$  corresponds to the phenomenon predicted by the cognitive dissonance theory (Festinger, 1957).

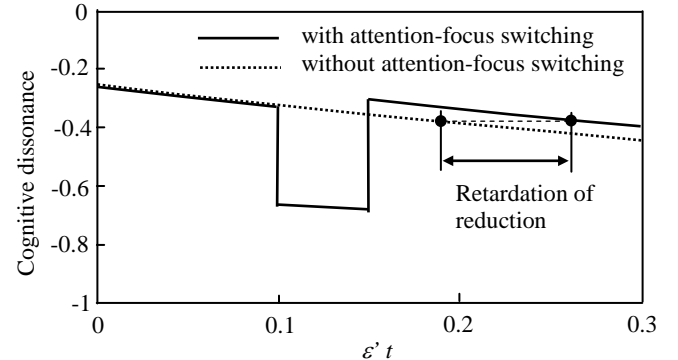
Conventional research on cognitive dissonance does not seem to have amply discussed the effect of attention. Since the occurrence of cognitive dissonance induces unpleasant feeling, the person mentioned above might strive to reduce it. He might recall the expectation and imagine the result that could be obtained in success of his trial. The fox in the

fable might say after a short time, "I have gotten hungrier. If I could get the grapes now, even if they were not ripe, I might eat them and be feeling full now". In this way, the focus of attention is switched from reality-based to imagination-based cognition. However, since one cannot live on imagination only and must act on the basis of reality, it is not long before the focus of attention is switched back to reality-based cognition. Thus, we assume here that the focus of attention is switched and the attentive state is changed such that  $1 \rightarrow 2 \rightarrow 1$ .

Transitions of evaluation and cognitive dissonance under such switching are calculated and depicted by the solid lines in Figure 6. In the first period of  $0 \leq \varepsilon't < 0.1$ , where the attentive state is 1, cognitive dissonance is monotonically reduced with the increase of  $\varepsilon't$ . At  $\varepsilon't = 0.1$ , the focus of attention is switched to the cognition of unit I and cognitive dissonance is reduced stepwise. During the period of  $0.1 \leq \varepsilon't < 0.15$ , where the attentive state is 2, reduction of cognitive dissonance continues with a small reduction rate. At  $\varepsilon't = 0.15$ , the focus of attention is switched back to the cognition of unit R. Cognitive dissonance is then built up stepwise and takes an amount greater than would be taken when switching is not performed at all. After that, attentive state 1 is retained and monotonical reduction of cognitive



(a) Change in evaluation with increase of time.



(b) Change in cognitive dissonance with increase of time.

Figure 6: Changes in evaluation and cognitive dissonance under attention-focus switching.

dissonance continues. Consequently, attention-focus switching causes retardation of cognitive-dissonance reduction as shown in Figure 6(b), which induces lingering of unpleasant feeling or discomfort. If the switching is frequently repeated, cognitive dissonance remains for a long time.

The well-known saying “What’s done is done” might imply that it is not worth worrying about an unfavorable situation caused by past behavior. It might also warn of the building-up of cognitive dissonance caused by attention-focus switching and suggest that attention not be focused on imagination-based cognition. Another well-known saying, stated by Dante Alighieri, is that “There is no greater grief than to recall a time of happiness when in misery.” This implies that to obtain peace of mind, it is important to focus one’s attention on the reality and discard the imagination even if one finds the reality unpleasant. The insistence common to these sayings might support the results shown in Figure 6(b).

## Conclusion

This paper described a novel connectionist model accounting for cognitive dissonance, in which the concepts of self as well as attention-focus switching are adopted. The model was investigated not with the computer simulation widely used in conventional research, but with a mathematical analysis based on a differential equation. Predictions based on the model were confirmed to coincide with experimental data reported in the literature. It was shown that attention-focus switching between reality-based and imagination-based cognition causes building-up of cognitive dissonance and retardation of its reduction. This coincides with the implication of well-known sayings suggesting ways to keep the mind away from feelings of suffering or discomfort.

## References

- Aronson, E. (1969). The theory of cognitive dissonance: A current perspective. In L. Berkowitz (Ed.), *Advances in experimental social psychology*, Vol.4, 1-34. NY: Academic Press.
- Brehm, J. W. (1956). Postdecision change in the desirability of alternatives, *Journal of abnormal and social psychology*, 52, 384-389.
- Festinger, L.(1957). *A Theory of Cognitive Dissonance*. Evanston, IL: Row, Peterson.
- Freedman, J. L. (1965). Long-term behavioral effects of cognitive dissonance. *Journal of Experimental Social Psychology*, 1, 145-155.
- Mosler, H., Schwartz, K., Ammann, F., & Gutsher, H. (2001). Computer simulation as a method of further developing a theory: Simulating the elaboration likelihood model, *Personality and Social Psychology Review*, 5(3)201-215.
- Nowak, A., Vallacher, R. R., Tesser, A., & Borkowski W. (2000). Society of self: The emergence of collective properties in self-structure. *Psychological Review*, 107, 39-61.
- Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15, 267-273.
- O'Reilly, R. C., & Munakata, Y. (2000). Computational explorations in cognitive neuroscience. Cambridge, MA: MIT Press.
- Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion*. NY: Springer-Verlag.
- Pisapia, N. D., Repovs, G., & Braver, T. S. (2008). Computational models of attention and cognitive control. In R. Sun (Ed.), *The Cambridge handbook of computational psychology*, NY: Cambridge University Press.
- Read, S. J., & Miller, L.C. (1994). Dissonance and balance in belief systems: The promise of parallel constraint satisfaction processes and connectionist modeling approaches. In R. C. Schank & E. Langer (eds.), *Beliefs, reasoning, and decision-making: Psycho-logic in honor of Bob Abelson*. Hillsdale, NJ: Lawrence Erlbaum.
- Rojas, R. (1996). Neural networks: A systematic introduction, Berlin: Springer-Verlag.
- Shultz, T. R., & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review*, 103, 219-240.
- Shultz, T. R., Leveille, E. & Lepper, M. R. (1999). Free choice and cognitive dissonance revisited: Choosing “Lesser Evils” versus “Greater Goods”. *Personality and Social Psychology Bulletin*, 25, 40-48.
- Spellman, B. A., Ullman, J. B., & Holyoak, K. J. (1993). A coherence model of cognitive consistency: Dynamics of attitude change during the Persian Gulf War. *Journal of Social Issues*, 49(4), 147-165.
- Thibodeau, R. & Aronson, E. (1992). Taking a closer look: Reasserting the role of the self-concept in dissonance theory. *Personality and Social Psychology Bulletin*, 18, 591-602.
- Van Overwalle, F., & Jordan, K. (2002). An adaptive connectionist model of cognitive dissonance. *Personality and Social Psychology Review*, 6, 204-231.
- Van Overwalle, F., & Siebler, F. (2005). A connectionist model of attitude formation and change. *Personality and Social Psychology Review*, 9, 231-274.

# Investigation of effects of working memory capacity on rule discovery process using eye movement data

Miki Matsumuro (muro@cog.human.nagoya-u.ac.jp)

Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)

Graduate School of Information Science, Nagoya University, Fro-cho, Chikusa-ku, Nagoya, Japan

## Abstract

Many studies have investigated the process of rule discovery. However, the data utilized in these studies, such as performance and verbal protocol data, were course-grained. In this study, we designed a new experimental method using eye movement data to observe the detailed process of rule discovery. In the proposed method, we corresponded the task display and a rule space in the participants' minds to understand how they consider the rules and observe instances by eye tracking. Then, we compared the process of rule discovery by people with high and low working memory capacities. The results of the experiment revealed that those with high working memory capacity tried to consider one or similar rules from multiple instances. On the other hand, those with low working memory capacity tended to consider various rules from one instance.

**Keywords:** Rule discovery; eye tracking; working memory capacity; search strategy

## Introduction

It is one of the most important activities to find regularities not only in science but also in many aspects of daily life. In this study, we tackle two goals in order to understand the process of rule discovery. Our first goal is to propose a new experimental method so as to observe the detailed process of rule discovery. The second goal is to investigate the relation between working memory capacity (WMC) and strategies of rule discovery.

The origin of this study is traced back to the dual space search theory proposed by Simon and Lea (1974). They suggested that the process of rule discovery develops through the interaction between two types of searches in two spaces, a rule space and an instance space. Problem solvers state rules by searching in a rule space while generating and observing instances in an instance space, and modify the rules or propose new rules based on their observations.

Although their theory can successfully explain the process of rule discovery, when conducting experiments based on this theory, there are limitations. One is that it is difficult to observe the detailed process of search in a rule space because the thought process to state rules cannot be monitored directly. To investigate this process, the researchers have utilized the protocol analysis method (e.g., Haverty, Koedinger, Klahr, & Alibali, 2000). The participants' think-aloud protocol data were used for this analysis. However, the data were coarse-grained on the time scale, and the participants mentioned only their conscious thoughts. Therefore, the analysis of the detailed patterns of search in the two spaces from protocol data often faced essential limitations. For these reasons, our first purpose is to propose a new experimental method to observe the detailed process of rule discovery.

The second goal is to investigate the relation between WMC and strategies of rule discovery. Dougherty and Hunter (2003) showed that people with high WMC maintained more alternative hypotheses in their mind when they engaged in hypothesis generation. However, their task was a probability judgment task and their analysis was based on participants' performance of the thought-listing task that was even more coarse-grained than the protocol data. Other studies also analyzed the effects of WMC based on the score of the reasoning tasks (e.g., Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002). In the current study, we focus not on outputs as results but on the process of rule discovery by using eye movement data.

In summary, the first purpose of our study is to propose a new experimental method by which we obtain detailed data about the process of rule discovery. The second purpose is to investigate how participants' WMC affects such a process. Our method using eye tracking was established based on the SDDS (Scientific Discovery as Dual Search) model of Klahr and Dunbar (1988). The reason we use eye tracking is that eye movement data are more fine-grained than verbal reports and obtained directly without participants' conscious effort (cf. Rehder & Hoffman, 2005). With this method, we examine differences in search strategies affected by WMC, selecting those who have high or low WMC based on screening test scores.

## Experimental Method

### Search in Each Space

We designed an experimental method by means of eye tracking. We utilize the idea of the structure of a hypothesis space in the SDDS model. It extends the dual space search theory for investigating scientific discovery. The hypothesis space in the SDDS model corresponds to a rule space in dual space search. The hypothesis space includes all available hypotheses, each of which is connected with others through search paths. The search path does not connect hypotheses that have different schema. Therefore, similar hypotheses with an identical schema construct a subspace of the hypothesis space. Based on this idea, in our method, we manipulate the structure of a rule space by giving a different function to each rule. Since a different function evokes a different schema, only rules that have the same function are connected, and they construct a subspace in a rule space.

We obtain search patterns in the rule and instance spaces by means of eye tracking. First, we define search in each space as follows. The search in a rule space is the process where



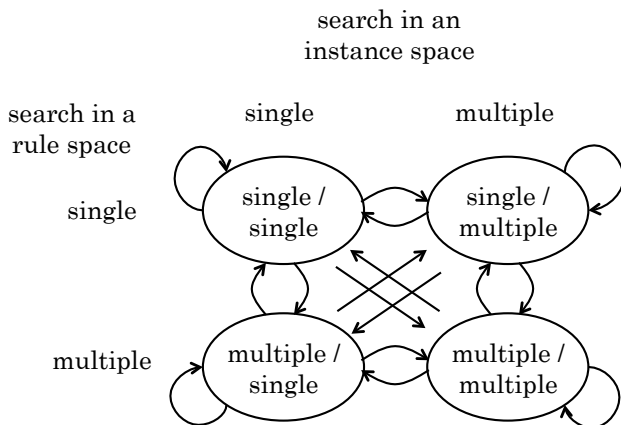


Figure 1: Four search statuses of the rule and instance spaces and transition patterns among the statuses.

participants consider rules from instances. On the other hand, the search in an instance space is defined as the activities where participants generate or observe instances to consider rules. In actual processes of rule discovery, timing to switch between searches in the rule and instance spaces is not clear. Therefore, in this method, we collect eye movement data as the search in a rule space and observations of instances as the search in an instance space.

There are two modes in the search process (Figure 1): the *single-* and *multiple-subspace search* in a rule space and the *single-* and *multiple-instance search* in an instance space. First, we define two modes of search in the rule space. The single-subspace search is the search mode in which participants search only in one subspace. In other words, they consider rules in the same subspace that are characterized by an identical function. On the other hand, in the multiple-subspace search mode, they try to decide which subspace they should search. Whereas the single-subspace search is a “search in a subspace,” the multiple-subspace search is a “search for a subspace” to be focused on.

For search in the instance space, we also define two modes of search. Participants in the single-instance search mode focus on observing one instance. They consider multiple rules from a single instance. In contrast, they observe various instances for stating rules in the multiple-instance search mode. They compare multiple instances to obtain cues for rule discovery.

Each participant’s search status is categorized into one of the four search statuses in Figure 1. The single/single status is one in which participants observe one instance and consider rules with an identical function. When participants consider a complex process, they do so in this status because they may concentrate their attention on a single event. Participants in the single/multiple status also consider rules in an identical subspace, but they observe and compare multiple instances to obtain the cues for rule discovery. On the other hand, in the multiple/single status, participants consider rules with various functions across multiple subspaces while focusing on examining one instance. Last, the participants in the multiple/multiple status also consider various rules across multiple

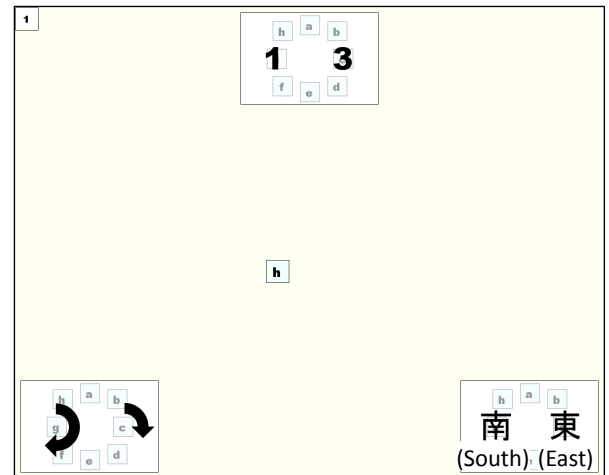


Figure 2: Example screenshot of task display in this study. The upper panel is the number panel, the lower-left panel is the arrow panel, and the lower-right panel is the compass panel. The directions were presented in Chinese characters. The target alphabet is presented in the center of the display.

subspaces while observing multiple instances in a short time. This status seems not to be suitable for rule searching because they simultaneously vary both factors, rules and instances, preventing efficient searches for rule discovery. To identify each search status (a circle in Figure 1) and each transition among the search statuses (an arrow in Figure 1), we systematically designed a rule discovery task as follows.

### Rule Discovery Task Using Eye Tracking

The task display consists of three panels (arrow, compass, and number panels) and eight letters (a to h) in the center of the display as shown in Figure 2. The eight letters are placed in circle. The presented letter (we call this the “target”) is determined based on the objects displayed on one of the three panels. The participants are asked to find a rule determining which panel relates to the target and how the target is determined by the objects in the related panel. Only one of the three panels relates to the target and the other two panels have no relation to it. The participants are required to find a rule as quickly as possible.

The experimental procedures are as follows. Before the experiment, an experimenter prepares six instances as stimuli. In the example instance in Figure 2, the rule may be: “the target is the letter (h) in the opposite position (North-West) of the combination of two directions (South-East) on the compass panel.” In this case, the experimenter selects five other instances so as to be instances consistent with the same compass rule. The participants are instructed that only one panel is related to the target in advance. The participants observe all six instances one-by-one using right- and left-arrow keys on the keyboard. The panel that has the same function is always presented in the same position on the display. Additionally, the six instances are presented in a cyclic manner; the participants can observe instances as many times as they want. When the participants think of a rule, they press the space key and report their rule to the experimenter. When the

Table 1: Functions and example rules of each panel

panel	function	how to decide target	example rule (see example in Figure 2)
number	order	a letter corresponds to a number in alphabetical order (“a” is 1)	Target is the sum of numbers in alphabetical order (e.g., $1 + 3 = 4 \rightarrow d$ )
arrow	rotation	a letter corresponds to an angle of arrows	Target is a shifted letter from “a” by an angle of left object (e.g., 180 degrees $\rightarrow e$ )
compass	position	a letter corresponds to a direction (based on map)	Target is pointed at by right object (e.g., East $\rightarrow c$ )

correct rule is discovered, the experiment is terminated. Every five minutes, the calibration for recording eye movement is performed. Table 1 shows functions and example rules in each panel. Before starting the task, the participants learn these functions sufficiently.

### Data and Search in Each Space

Figure 3 shows example data obtained in the experiment. For example, from 520 to 530 seconds in Figure 3, this participant focused her attention on the arrow panel, meaning that she considered only rules that use arrows; that is, she searched in the subspace of “rotation” rules in a rule space. The shift of fixation from the arrow to the number panel at around 530 seconds means that her search subspace shifted to the “order” rule subspace. In the same manner, the behavior after 530 seconds is interpreted that she expanded her attention toward the three panels, meaning that she searched all subspaces in a rule space simultaneously. Soon after broadening the search, she fixed her attention on the compass panel. In the final part, she shifted her search subspace to the “order” rule without broadening the search. On the other hand, for the search in an instance space, the participant seemed to focus on a single instance before around 560 seconds. Then, she moved to observe multiple instances during a short time after around 570 seconds.

With reference to Figure 3, around 520 seconds, her search mode was in the single/single status. Then, she searched in multiple subspaces in a rule space with an identical instance at around 530 seconds, meaning that her search status shifted to the multiple/single. Soon, she came back to the single/single status, and this status continued for a while. After around 570 seconds, she observed multiple instances one-by-

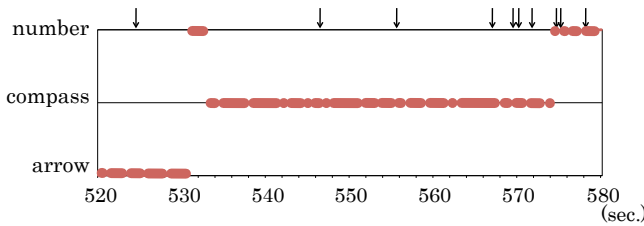


Figure 3: Timeline graph of collected data.

Each red horizontal line indicates a panel that the participant focused on. Labels of the vertical axis indicate three panels. Arrows on the top of the graph mean timing at which the participant shifted an instance to the previous or next one.

one, meaning that she moved to the single/multiple status. As shown in this example, it is possible to obtain detailed search processes in the rule and instance spaces by our experimental method.

We conducted an experiment with this method to observe the process of searches in the rule and instance spaces to understand the effect of WMC on the rule discovery process.

## Experiment

First, we measured the participants’ spatial span capability as a screening test.

### Screening Test

Fifty-seven undergraduates engaged in the spatial span test by Shah and Miyake (1996). In the task, a capital letter was presented on a computer screen that was not in the upright orientation. The participants were asked to indicate whether the letter was normal or mirror-imaged and were required to store the orientation of each letter for a subsequent recall test. One participant whose judgment score was around the chance level was excluded from analysis. Fifty-six participants’ data were scored based on the partial-credit unit scoring by Conway et al. (2005). The mean score was 0.576 ( $SD = 0.200$ ), the highest score was 0.920, and the lowest score was 0.121.

### Rule Discovery Task

#### Method

**Participants** The twelve participants with the highest WMC scores participated as the high-WMC group. Their mean score was 0.825 ( $SD = 0.068$ ). The twelve participants with the lowest WMC scores participated as the low-WMC group. Their mean score was 0.290 ( $SD = 0.091$ ). The mean WMC score of the high group was significantly higher than that of the low group ( $t(22) = 16.318, p < .001$ ).

**Apparatus** We presented the task display on a 17-in. monitor with a resolution of  $1280 \times 1024$  pixels. The participants were seated approximately 60 cm away from the monitor. The size of panels was approximately  $5.25^\circ \times 7.82^\circ$  of visual angle. Each panel was placed as shown in Figure 2 on the upper-center, lower-left, and lower-right of the task display. The participants’ eye movements were recorded using the Tobii T60 eye tracker at 60 Hz. The participants were allowed to move their heads naturally.

Table 2: Rules used in this study

	panel	rule
task 1	compass	Target is the opposite position of combination of two directions on the panel
task 2	arrow	Target is a shifted letter by sum of an angle of two arrows on the panel and 135 degrees
task 3	number	Target is the sum of the difference between two numbers and the bigger one on the panel

**Procedures** Approximately one month after the screening test, the rule discovery task was conducted individually. First, the participants were instructed on the task and learned the function of each panel sufficiently through practice. Task 1 was preliminarily performed to let the participants establish their own strategies. In task 1, the participants needed to find a simple rule in ten minutes. Soon after task 1 ended, task 2 was conducted with the same procedure as that in task 1. The participants needed to find a relatively complex rule in twenty minutes in task 2. The data of task 2 were used for analysis. As an exception, the participants who found the rule before fifteen minutes passed were led to engage in another task (task 3) with the same procedure as that in task 2. In this case, the data of task 3 were used for analysis. Table 2 shows the rules used in the experiment.

## Results

One participant in the low group was excluded from analysis because we did not obtain the data for more than fifteen minutes. We analyzed only the participants from whom more than 60% of eye movement data were recorded correctly. Based on the criterion, six participants in the high group and two participants in the low group were excluded; therefore, the data of six participants in the high group (WMC score  $M = 0.858, SD = 0.058$ ) and nine participants in the low group (WMC score  $M = 0.289, SD = 0.100$ ) were used for analyses. The average WMC score of the high group was still significantly higher than that of the low group ( $t(13) = 12.551, p < .001$ ).

**Search Status** Fixations longer than 100 msec were analyzed, and fixations outside the panels were excluded from analysis. First, we traced each participant's data along a timeline to acquire on which panel the fixations were observed and when an instance was shifted, as shown in Figure 3. We defined that the fixation shift happened when the participant's fixation point was observed on a different panel from the preceding one. Similarly, we defined that the instance shift happened when the participants pressed the arrow keys to observe another instance. Based on the median of the fixation time in one panel (8.961 seconds) and the mean of observation time in one instance (7.531 seconds), we segmented the timeline

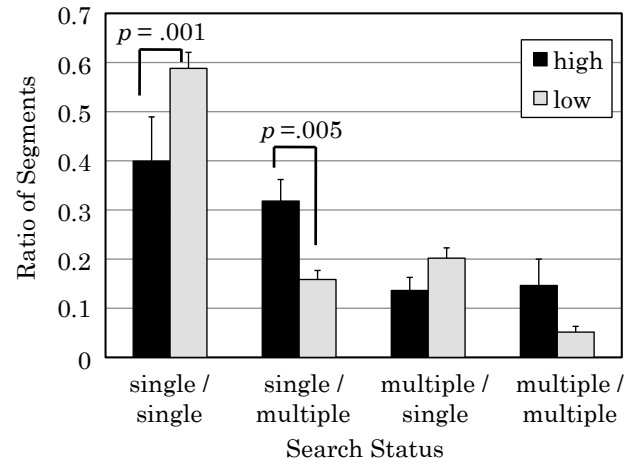


Figure 4: Ratio of segments in four search statuses in each WMC group (bars show standard errors).

every 7.5 seconds.

Then, we counted how many fixation and instance shifts happened in each segment. We also performed the same analysis with the segmentations at six, seven, and eight seconds, and similar results were confirmed. We defined that the participants were in the multiple-subspace search mode in a rule space when the fixation shifts were observed more than once in each segment. Similarly, for search in an instance space, in each segment when more than one instance shift were observed, the participants were defined to be in the multiple-instance search mode. Based on the definitions, we classified the participants' status in each segment into four categories shown in Figure 1.

Figure 4 presents the ratio of segments categorized into each search status in each WMC group. Mixed ANOVA with WMC (high and low) as between-subject factor and search status (four statuses) as within-subject factor was performed on the ratio of segments. The interaction between the WMC and the search status reached significance ( $F(3, 39) = 6.328, p = .001$ ). The ratio in the single/single status in the high-WMC group was significantly higher than that in the low-WMC group ( $F(1, 52) = 12.116, p = .001$ ). On the other hand, the ratio in the single/multiple status in the high-WMC group was significantly lower than that in the low-WMC group ( $F(1, 52) = 8.663, p = .005$ ). These results show that the participants in the high-WMC group compared multiple instances more frequently than those in the low-WMC group when they were in the single-subspace search mode in a rule space search. On the other hand, the participants in the low-WMC group searched in the single subspace in a rule space with fewer shifts of instances. When they searched in a rule space with the multiple-subspace search mode, there was no significant difference in search status between the two WMC groups.

**Transition Probability among Each Search Status** Next, we calculated the transition probability of shifting from one to another or being in the same search status. There are sixteen transition patterns shown in Figure 1. Figure 5 shows

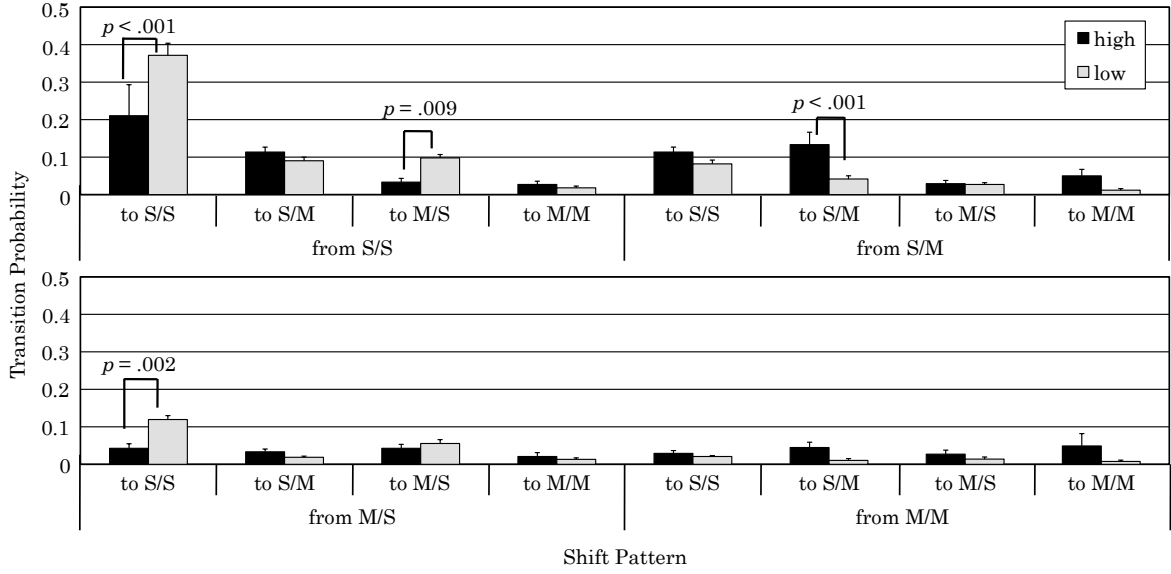


Figure 5: Probability of each transition pattern in each WMC group (bars show standard errors). S/S = single/single, S/M = single/multiple, M/S = multiple/single, M/M = multiple/multiple.

the transition probability of each transition pattern in each WMC group. Mixed ANOVA with WMC (high and low) as between-subject factor and transition pattern (sixteen patterns) as within-subject factor was performed on the transition probability. The interaction between the WMC and the transition pattern reached significance ( $F(15, 195) = 5.323, p < .001$ ). Four comparisons reaching significance are shown in Figure 5: the transition probability of single/single to single/single in the low-WMC group was significantly higher than that in the high-WMC group ( $F(1, 208) = 43.283, p < .001$ ); the transition probability of single/single to multiple/single in the low-WMC group was significantly higher than that in the high-WMC group ( $F(1, 208) = 6.955, p = .009$ ); the transition probability of single/multiple to single/multiple in the high-WMC group was significantly higher than that in the low-WMC group ( $F(1, 208) = 14.012, p < .001$ ); and the transition probability of multiple/single to single/single in the low-WMC group was significantly higher than that in the high-WMC group ( $F(1, 208) = 9.819, p = .002$ ). Other comparisons did not reach significance.

These results indicate that when they searched in a single subspace of a rule space, the participants in the high-WMC group tended to continue the multiple-instance search in an instance space, and those in the low-WMC group tended to continue to the single-instance search. On the other hand, the participants in the low-WMC group tended to shift to the search in multiple subspaces of a rule space focusing on a single instance more frequently than those in the high-WMC group, even though their multiple-subspace search did not continue very long.

## Discussion

Our first purpose was to propose a new experimental method by which we could capture the detailed process of rule discovery. This purpose was achieved by manipulating a structure of a rule space and using eye tracking. We successfully

observed the search statuses in the rule and instance spaces and the transition patterns among those in detail. The second purpose of this study was to compare the process of rule discovery by people with high or low WMC. Our method detected the different processes of the two types of participants.

## Advantages of New Method

The protocol analysis has two main limitations; as a result, we could not capture the detailed process of search in a rule space. The first difficulty is the grain size of data. Reporting one's thoughts cannot happen at the speed of thought, making it impossible for participants to report all of their thoughts. Participants would omit reporting their short-term thoughts. The use of eye tracking has the potential to solve this problem. We can directly obtain which direction participants focus their attention in by eye tracking. Furthermore, the sampling rate is very fine, 60 Hz in this study.

The second limitation of verbal protocols appears when participants have difficulty putting their thoughts into words. Participants sometimes have "no idea" during an experiment. In this study, all except one participant reported in the post-task interview that there were periods when they came up with no idea. In such a situation, the participants were usually confused and had difficulties in putting their thoughts into words. Moreover, the verbalization of confused thoughts would further muddle their thoughts. Eye movement data are always recorded throughout the task, even if participants cannot verbalize their thoughts.

We designed our task based on the SDDS model by Klahr and Dunbar (1988) to maximize the advantages of eye movement analysis. The task display as the observable externalized space consistently corresponds to the subspaces in a rule space as the internal representation. Due to this design, we could analyze search in a rule space from eye movement data. These features in our experimental method enabled us to capture each search status and the transition among the statuses

in detail.

### Differences between Each WMC Group

As a result of the experiment, different strategies were observed according to participants' WMC.

The analysis of the occurrence ratio of each search status and each transition pattern suggested that the participants in the high-WMC group tended to search in a rule space focusing on one subspace compared to those in the low-WMC group. Additionally, when they considered the rules that have an identical function, they observed and compared multiple instances more continuously. These results suggest that the participants with high WMC considered complex rules with an identical function while comparing multiple instances, meaning that they preferred the depth-first search in a rule space. Note that they did not necessarily fix their search to one single subspace because they actually shifted the search from one to another subspace at their own pace.

On the other hand, quick shifts of fixation among multiple panels, i.e. being in the multiple-subspace search mode in a rule space, were more frequently observed in the low-WMC group. However, this mode did not continue, and they soon came back to the single-subspace search mode. These results indicate that the participants with low WMC searched in a rule space switching the single- and multiple-subspace search modes alternatively. For the search in an instance space, they did not observe multiple instances as the participants in the high-WMC group did. This implies that they tried to consider rules from one instance, whereas the participants in the high-WMC group tried to consider one rule from multiple instances. These results suggest that they valued the breadth-first search. These two strategies, depth-first and breadth-first searches, were also observed in our previous study (Matsumuro & Miwa, 2011).

We suggest a possible reason that the search strategy was different according to their WMC based on the studies of category learning. Two strategies have been shown (DeCaro, Thomas, & Beilock, 2008; Rehder & Hoffman, 2005): the rule-based strategy where participants conduct hypothesis testing explicitly, and the information-integration strategy where rules are learned implicitly by integrating stimuli across multiple dimensions. The participants in the high-WMC group in our study observed multiple instances while searching in a single subspace in a rule space. They would generate hypotheses by comparing instances and tested these hypotheses as with the rule-based strategy. The explicit strategy was suitable for the participants with high WMC because it relies heavily on working memory. By contrast, the participants in the low-WMC group searched in multiple subspaces with a single instance. They would have gathered information from multiple dimensions and tried to figure out the relation between the target and the objects on the three panels implicitly, as with the information-integration strategy. The implicit strategy is processed without conscious control. Therefore, this strategy is suitable for the participants with low WMC because it does not require working memory load.

Given these points, it may be possible that our participants in each group selected a strategy suitable for each WMC. Note that the search in multiple subspaces by the participants in the low-WMC group did not continue. They had to conduct explicit hypothesis testing because all participants were required to find a rule that could be reported in words.

To investigate this possibility, we should analyze verbal protocols along with the eye movement data recorded in this study. Additionally, in future work, we need to investigate the relation between the search strategies and the discovery rate, and interaction of such a relation and individual differences.

### References

- Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, 12(5), 769–786.
- DeCaro, M. S., Thomas, R. D., & Beilock, S. L. (2008). Individual differences in category learning: Sometimes less working memory capacity is better than more. *Cognition*, 107(1), 284–294.
- Dougherty, M. R. P., & Hunter, J. E. (2003). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, 113(3), 263–282.
- Haverty, L. A., Koedinger, K. R., Klahr, D., & Alibali, M. W. (2000). Solving inductive reasoning problems in mathematics: Not-so-trivial pursuit. *Cognitive Science*, 24(2), 249–298.
- Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, 12(1), 1–48.
- Matsumuro, M., & Miwa, K. (2011). An investigation of search strategies for hypothesis generation using eye movement data. In L. Carlson, C. Hoelscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 3424–3429). Boston, MA: Cognitive Science Society.
- Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, 51(1), 1–41.
- Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General*, 125(1), 4–27.
- Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In L. W. Gregg (Ed.), *Knowledge and cognition*. Hillsdale, NJ: Lawrence Erlbaum.
- Süß, H. M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability - and a little bit more. *Intelligence*, 30(2), 261–288.

# How many Neurons for your ‘Grandmother’ ?

## Three Arguments for *Localised* Representations

Julien Mayor (julien.mayor@unige.ch)

FPSE, University of Geneva  
1211 Genève 4, Switzerland

Kim Plunkett (kim.plunkett@psy.ox.ac.uk)

Department of Experimental Psychology, University of Oxford  
Oxford OX1 3UD, United Kingdom

### Abstract

In a recent article, Bowers (2009) argues that local representations are more consistent with neuro-biological data than distributed representations, as typically generated in Parallel Distributed Processing (PDP) models. We present three reasons why *localised* neural representations are good candidates for supporting mental representations, as they provide a solution to the trade-off between combinatorial arguments that favour fully-distributed representations and metabolic arguments which favour localist representations.

**Keywords:** distributed representations, local representations, self-organising maps, synaptic pruning, brain metabolism

### Introduction

Over the last thirty years, hypotheses concerning the nature of mental representations have essentially been polarised to two interpretations: some researchers argue that brain representations are distributed (among them, proponents of the Parallel Distributed Processing (PDP) approach: e.g., Rumelhart, McClelland, & the PDP Research Group, 1986; Seidenberg & McClelland, 1989; McClelland & Rogers, 2003; Plaut & McClelland, 2010), while others suggest that local representations fit neuro-physiological data more accurately (e.g., Page, 2001; Bowers, 2009). Most of the arguments in favour of distributed representations fall into one of the following two categories: high combinatorial power and robustness with respect to lesions. In contrast, Bowers (2009) reviews neuro-biological evidence for relatively sharply tuned neurons reminiscent of localist representations and argues that distributed approaches fail to provide unambiguous representations under superposition.

We attempt to clarify the role of combinatorial power, in light of the superposition problem. We then introduce two metabolic arguments to the debate. We suggest that a potential solution to the debate relies on *localised* representations, capitalising on robust representations that span only a limited number of neurons, thereby minimising the energy expenditure associated with mental representations. Finally, we discuss the implications of this proposal and highlight examples already using localised representations.

### The combinatorial argument

The idea that distributed representations can code many more patterns than localist coding scheme is well established. Traditional binary coding, in which a neuron is either active or

silent, emphasises this difference. We will therefore reiterate this combinatorial argument using binary activation levels and comment on its validity in the context of decoding superposed patterns. The extension to continuous encoding will then be discussed.

### The case of binary encoding

**The coding advantage** Elementary calculus shows that  $2^n$  patterns can be encoded over  $n$  neurons, corresponding to the case of fully-distributed representations (see Fig. 1). With localised representations, the combinatorial power decreases rapidly. Suppose, for example, that each pattern can use at most  $n$  neurons out of a total system of  $N$  neurons. The number of patterns that can be stored in  $n$  neurons is then  $2^n$  in each of the subset of  $n$  neurons picked from the total pool of neurons. If  $n = N/2$ , for example, there are two subsets of  $n$  neurons, each coding  $2^n$  patterns. The total number of patterns  $p$  with a degree of localisation of  $n$  stored in  $N = 2n$  neurons would then be  $p = 2 \cdot 2^n$ . Fig. 1 depicts the number of patterns  $p$  that can be stored among  $N$  neurons (maximum 20 neurons in the simulation), for different levels of localisation  $n$ . A purely localist encoding ( $n=1$ ) can only store as many patterns as there are neurons. At the other end of the spectrum, fully-distributed representations can store  $2^n$  patterns. In between, the number of patterns one can store is directly related to the number of neurons that are involved in the coding of an individual pattern. As a consequence, localist representations have a limited capacity to store only as many separate representations as the number of neurons. Orders of magnitude can be gained by coding each patterns over a few neurons. With only 30 neurons, a fully-distributed approach would be able to store more than a billion different representations, a number that exceeds by orders of magnitude the likely capacity for human mental representations: “even if the distinguishable visual items are larger than the number of the different types of objects ( $< 100000$ ) that humans are able to discriminate, cortical visual neurons are certainly so numerous that there would be enough sets of them to represent each single object (or property)” (Pareti & De Palma, 2004, p.45). On the other hand, localised encoding ( $n > 1$ ) can rapidly reach the combinatorial power required to represent a very large number of different representations.

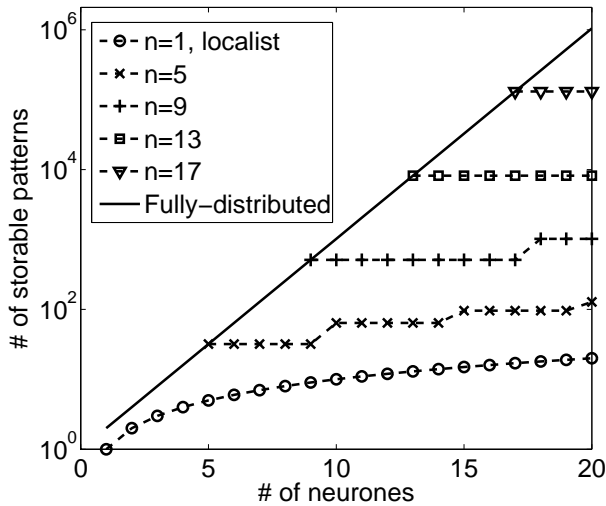


Figure 1: Number of patterns that can be stored as a function of the total number of neurones, using binary encoding. The different curves correspond respectively to a purely localist encoding, different levels of localisation and a fully-distributed encoding.

**The decoding problem** Encoding many representations is a necessary requirement for human cognitive performance. In many situations, multiple representations need to be encoded simultaneously. This leads to the potential ambiguity introduced by a superposition of representations. For example, visual scenes will contain a configuration of independent objects. A difficult task for the brain is therefore to be able to decode superposed representations unambiguously. Patterns of activation in the brain can be construed as multi-dimensional vectors. Elementary linear algebra constrains the number of linearly independent vectors to the number of dimensions represented in the system. Any additional vector can be expressed as a linear combination of other vectors. Consequently, there can be at most  $N$  independent patterns unambiguously coded across  $N$  neurones, on the assumption that each neurone encodes a separate dimension. Beyond that limit, additional representations can be misconstrued as a superposition of one or more other representations *even for fully-distributed representations*. If the network's task is to encode multiple patterns that can be superposed in the same neural substrate, the combinatorial advantage of distributed representations is compromised.

### The case of continuous encoding

Activation levels in neurones do not need to be restricted to a binary coding scheme. For example, simple rate coding models make use of a range of activation levels to encode different stimuli. More complex, and more realistic models of neurones make use of the rich dynamics of neuronal firing. A strict and complex mathematical analysis of the coding and decod-

ing capacities in both local and distributed representations for continuous coding schemes is beyond the scope of the present article. Nevertheless, it is worth commenting some implications of this approach.

A first observation is that coding is further enriched by the increased range of values any neurone can take. In fact, a single neurone could encode as many different patterns as needed, as long as the decoder has a resolution that is fine enough. The combinatorial advantage of distributed coding is still present for a decoder with a fixed resolution (e.g., the ability to detect subtle differences between relevant, and different, neural activation levels). However, as single neurones can encode many more patterns with continuous encoding schemes than with binary coding, fewer neurones are required to encode the same number of patterns. For example, 10 binary coding neurones are needed to represent 1000 patterns ( $2^{10} = 1024$ ), but if continuous activation levels can be detected with greater accuracy so that each neurone could take 10 distinct values each, only 3 neurones would then be required ( $10^3 = 1000$ ).

A second observation is that decoding subtle differences between different activation levels of neurones is a non-trivial problem. Presence of noise would limit the decoder's resolution and the more neurones needed to encode a representation, the more difficult it will be to decode that information and the more neurones required to act as decoders (e.g., see Földiák (2003) for a discussion of the advantage of representations that span fewer neurones (sparse representations) than fully-distributed representations for decoding).

The limited resolution of the decoder effectively reduces the case of continuous encoding to a simple extension of binary coding, where each neurone can have two distinct activation levels, to a case of  $N$ -coding, in which each neurone can take  $N$  distinct values. Small values for  $N$  magnify the problem of superposition of representations for distributed representations while reducing the problem of combinatorial limitations for localised representations. Large values for  $N$  would furthermore undermine the claim that localised (or even localist) representations cannot encode a sufficiently large number of different representations, while increasing the complexity and vulnerability of a decoder network that requires an increasingly large number of neurones.

### The metabolic argument

Let us turn now onto a consideration that can be made independently from the nature of the neural coding itself. It is often claimed that the resource needed to encode  $N$  patterns is less using fully-distributed representations than localist codes, thereby minimising metabolic expenses, because fewer neurones are required for distributed representations.

However, consider a brain structure with a given number of neurones required to represent a number of different patterns. The metabolic expense of the brain structure is, to a first approximation, proportional to the number of neurones that participate in the representation of the pattern(s) present



in a scene (or maintained/sustained in that brain structure). As a rule of thumb, if less neurons are required to participate in the representation of a pattern, the less the energy required for that task.

Fig. 2 depicts the energy consumption (as indexed by the number of neurons that participate in the representation) as a function of the number of patterns (or objects) that need being represented in a network of 20 neurons. Different curves

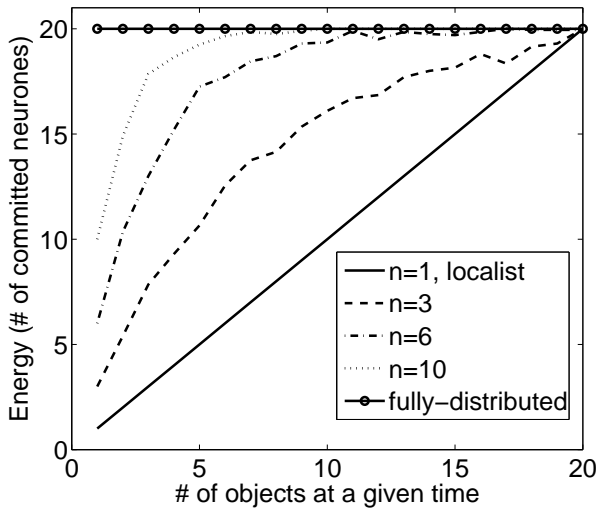


Figure 2: Energy as indexed by the number of neurons required to represent a number of objects simultaneously. The different curves correspond respectively to a purely localist encoding, different levels of localisation and a fully-distributed encoding. Localised representations use less energy than fully-distributed representations for a given architecture (in the present simulation, total number of neurons=20).

correspond to different localisation levels: in a purely localist coding, each pattern is represented by a single neuron. In this case, the energy consumption is proportional to the number of objects represented simultaneously. As the number of neurons involved in the representation of a given pattern increases, the energy expenditure increase for any number of patterns (or objects) being represented at a given time. Ultimately, fully-distributed representations require all neurons to participate in the representation of even a single pattern. Unless the system is working at full capacity at all times, energy consumption is minimised for localist representations but maximised for fully-distributed representations. Localised representations consume intermediate levels of energy.

### The synaptic pruning argument

Neural resources involve not only the cell bodies of neurons but also the connections between them. Associations between representations require appropriate connections or synapses.

Synaptic maintenance is also a contributor of energy consumption. For example, neural mappings between an object representation and its corresponding label require appropriate cross-modal synapses between visual and auditory areas. The number of cross-modal synapses required to form the mapping between the different brain structures depends on the degree of localisation of each representation in both structures. Figure 3 depicts the number of cross-modal synapses needed to maintain an appropriate mapping between representations in different neural structures, as a function of the degree of localisation of the representations in each structure (which have been chosen to be identical for the sake of simplicity). The number of synapses needed increases with the number of objects that are encoded in each modality. Note

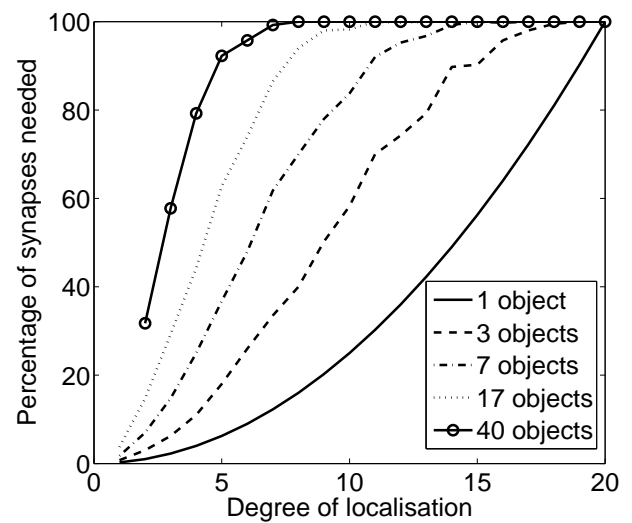


Figure 3: Percentage of cross-modal synapses required to maintain the mappings between uni-modal representations without degradation, as a function of the degree of localisation. The different curves correspond to different network loads in terms of the number of objects that need to be represented in each modality. Note that pruning can only be achieved with reduced levels of localisation.

that fully-distributed representations (where the degree of localisation equals the number of neurons, 20) require the full set of cross-modal synapses ( $20^2 = 400$ ) even when each structure is only required to encode a single object. A lower number of synapses can only be achieved for reduced levels of localisation for the neural representations.

It is important to note that the number of synapses is not constant during brain development. After an early proliferation of synapses, their number remains approximately constant, before environmentally induced synaptic pruning reduces the total number of synapses (see Huttenlocher, 2002). The observed synaptic pruning mechanism is usually associated with either an improvement in cognitive skills (Miller, Keller, & Stryker, 1989; Chechik, 1998) and/or an optimi-

sation in metabolic resources (Roland, 1993; Feinberg, Thode Jr, Chugani, & March, 1990), potentially leading to an internally-driven reorganisation of neural representations. So that synaptic pruning can operate without being detrimental to the task, representations benefit from a reduced degree of localisation.

### Discussion: Solutions to the trade-offs

Mental representations can have different levels of localisation, as defined by the number of neurons that are required to take part in the representation of an individual pattern. Many constraints may impact this degree of localisation, such as the number of different patterns a neural structure can code, the capacity to decode a superposition of patterns, robustness arguments and metabolic constraints at the level of neurons and synapses. The exact solution to the trade-off between these constraints remains elusive. While there are many advantages in having more than one neuron involved in the representation of a given pattern (robustness, combinatorics), there are at least as many in restraining the number of neurons taking part in a representation; metabolic minimisation, synaptic pruning, simpler decoding. We suggest that computational approaches to neuroscience and psychology may need to adopt the perspective that patterns are represented in a localised fashion; not localist (only one neuron per pattern) nor fully-distributed. The degree of localisation can, of course, be modulated according to the structure under consideration or, if the function is highly abstract, according to the task.

Self-Organising Maps (SOMs) offer an approach in which degree of localisation is discovered from exposure to the input structure (Kohonen, 1984). SOMs form topographically organised maps of neurons, such that neighbouring neurons respond to similar input. The resources on the map dedicated to a particular pattern or category is determined by many factors such as the number of different patterns that a SOM has been exposed to, the number of neurons on the SOM, the frequency with which a given category of patterns is presented, and the magnitude of the pattern variations in each category. After exposure to a structured environment, SOMs display a partitioned map from a representational perspective: each pattern creates a unique pattern of *localised* neural activity and each category of patterns would tend to solicit the same group of neighbouring neurons in order to represent different patterns that belong to the same category.

The organisation of a SOM after learning mimics cortical maps observed throughout many different cortical areas. SOMs have been very successful at modelling the architecture of the primary visual cortex (Miikkulainen, Bednar, Choe, & Sirosh, 2005) where neighbouring neurons are responsive to similar orientations of the visual scene (Hubel & Wiesel, 1962). Topologically-organised maps have also been found in the human auditory cortex (Romani, Williamson, & Kaufman, 1975; Pantev et al., 1995), in the human frontal and prefrontal cortex (Hagler & Sereno, 2006) and in parietal cortex (Sereno & Huang, 2006).

Beyond mimicking the neuro-anatomical organisation of cortical maps, SOMs sustain representations that possess interesting properties from a psychological perspective. For example, categories are formed in an unsupervised way, similarly to infant's capacities to form categories in the absence of supervision (Younger, 1985) and discrepancies between a pattern and its representation provide an accurate index of looking behaviour of young infants during categorisation tasks (Gliozzi, Mayor, Hu, & Plunkett, 2009). When input patterns possess a family resemblance structure (e.g., basic level categories of objects, Rosch & Mervis, 1975), representations on the SOM are warped in a manner that mimic categorical perception (Mayor & Plunkett, 2010). Since a single pattern activates a localised pattern of neural activity on the map, only a limited number of neurons contribute to the pattern representation. However, when the number of patterns represented exceeds the number of neurons on the map, a single neuron must participate in the representation of multiple patterns from the same category. Consequently, some neurons are maximally active when an average of a few patterns is presented to the map. This provides a representation advantage for central tendencies, thereby implementing at a representational level the advantage of prototypes over atypical members of a category (Rosch, 1973; Mervis, 1984). Interestingly, the fact that multiple neurons contribute to the representation of different members of the same category of patterns maintains a sensitivity to within category variations, as observed in speech perception (McMurray, Tanenhaus, & Aslin, 2002).

Mayor and Plunkett (2010) have also evaluated the impact of synaptic pruning in a model of early word learning, consisting of two SOMs connected by cross-modal Hebbian synapses. Synaptic pruning was shown to enhance the quality of word-object mappings, once stable representations of objects and labels were achieved on the maps. The localised representations of individual objects and labels permitted high levels of pruning so as to associate objects categories and their corresponding labels in a one-to-one mapping. Synaptic pruning of any one-to-one mapping between cortical representations (or thalamo-cortical projections) would also benefit from such localised representations. In contrast, high levels of pruning would be detrimental to highly distributed representations. The presence of high levels of synaptic pruning from mid-childhood would seem to favour the formation of these relatively localised mental representations.

It is noteworthy that any representations requiring a relatively small number of neurons also satisfy the conditions for metabolic constraints and synaptic pruning. However, these constraints do not require that neurons supporting the representation of a given pattern need to be neighbours. Examples of sparse coding (Quiroga, Kreiman, Koch, & Fried, 2008; Quiroga & Kreiman, 2010), in which only a small subset of neurons is active for a pattern have been shown offer decoding advantages (Földiák, 2003) as well as minimising metabolic demand. SOMs offer sparse coding in which the few neurons

taking part in the representation of a pattern are proximate, thereby providing additional advantages in terms of mimicking cortical maps that are found across the human brain (Hubel & Wiesel, 1962; Romani et al., 1975; Pantev et al., 1995; Hagler & Sereno, 2006; Sereno & Huang, 2006) and constraining the need for long distance connections (Durbin & Mitchinson, 1990). SOMs may also provide a potential advantage in terms of decoding the information, as representations of different patterns that belong to the same category tend to be similar. As a consequence, SOMs, localised representations in general, should lead to enhanced robustness in the presence of noise.

## Conclusion

The resources needed for mental representation are constrained by many different, and often opposing, pressures. A solution to the trade-off between robustness and combinatorial power, which favour representations with many neurons, and metabolic and synaptic pruning constraints, which favour fewer neurons, is to limit the number of neurons needed to represent a pattern. Sparse, localised representations provide an elegant alternative to purely localist representations and fully-distributed ones. Self-Organising Maps provide a natural, and unsupervised, approach for forming localised representations which mimic cortical maps found throughout the human cortex. The topographical structure of these SOMs also permit efficient pruning mechanisms to operate, maximising metabolic efficiency and providing accurate models of human cognitive performance and development.

## Acknowledgments

This work is supported by the Swiss National Science Foundation grant 131700 awarded to Julien Mayor and by the Economic and Social Research Council Grant RES-062-23-0194 awarded to Kim Plunkett.

## References

- Bowers, J. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review*, 116(1), 220–251.
- Chechik, G. (1998). Synaptic Pruning In Development: A Computational Account. *Neural Computation*, 10(7), 1759–1777.
- Durbin, R., & Mitchinson, G. (1990). A dimension reduction framework for understanding cortical maps. *Nature*, 343, 644–647.
- Feinberg, I., Thode Jr, H., Chugani, H., & March, J. (1990). Gamma distribution model describes maturational curves for delta wave amplitude, cortical metabolic rate and synaptic density. *Journal of Theoretical Biology*, 142(2), 149–61.
- Földiák, P. (2003). Sparse coding in the primate cortex. In M. Arbib (Ed.), *The handbook of brain theory and neural networks*. MIT Press, Cambridge, MA.
- Gliozzi, V., Mayor, J., Hu, J.-F., & Plunkett, K. (2009). Labels as features (not names) for infant categorisation: A neuro-computational approach. *Cognitive Science*, 33(4), 709–738.
- Hagler, D., & Sereno, M. (2006). Spatial maps in frontal and prefrontal cortex. *Neuroimage*, 29(2), 567–577.
- Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160(1), 106–154.
- Huttenlocher, P. (2002). *Neural Plasticity: The Effects of Environment on the Development of the Cerebral Cortex*. Harvard University Press.
- Kohonen, T. (1984). *Self-organization and associative memory*. Berlin: Springer.
- Mayor, J., & Plunkett, K. (2010). A neuro-computational model of taxonomic responding and fast mapping in early word learning. *Psychological Review*, 117(1), 1–31.
- McClelland, J. L., & Rogers, T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4(4), 310–322.
- McMurray, B., Tanenhaus, M., & Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86(2), 33–42.
- Mervis, C. (1984). Early lexical development: The contributions of mother and child. In C. Sophian (Ed.), *Origins of cognitive skills*. Hillsdale, N.J.: Lawrence Erlbaum.
- Miikkulainen, R., Bednar, J., Choe, Y., & Sirosh, J. (2005). *Computational Maps In The Visual Cortex*. Springer.
- Miller, K., Keller, J., & Stryker, M. P. (1989). Ocular dominance and column development: analysis and simulation. *Science*, 245, 605–615.
- Page, M. (2001). Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences*, 23(04), 443–467.
- Pantev, C., Bertrand, O., Eulitz, C., Verkindt, C., Hampson, S., Schuierer, G., et al. (1995). Specific tonotopic organizations of different areas of the human auditory cortex revealed by simultaneous magnetic and electric recordings. *Electroencephalography and clinical Neurophysiology*, 94(1), 26–40.
- Pareti, G., & De Palma, A. (2004). Does the brain oscillate? The dispute on neuronal synchronization. *Neurological Sciences*, 25(2), 41–47.
- Plaut, D., & McClelland, J. (2010). *Psychological Review*, 117(1), 284–288.
- Quiroga, R., & Kreiman, G. (2010). Measuring sparseness in the brain. *Psychological Review*, 117(1), 291–297.
- Quiroga, R., Kreiman, G., Koch, C., & Fried, I. (2008). *Trends in Cognitive Sciences*, 12(3), 87–91.
- Roland, P. (1993). *Brain activation*. Wiley-Liss New York.
- Romani, G., Williamson, S., & Kaufman, L. (1975). Tonotopic organization of the human auditory cortex. *Psychiatry*, 132, 650.
- Rosch, E. (1973). On the internal structure of perceptual

- and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language*. New York: Academic Press.
- Rosch, E., & Mervis, C. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573–605.
- Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1: Foundations). Cambridge, Massachusetts: The MIT Press.
- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523–568.
- Sereno, M., & Huang, R. (2006). A human parietal face area contains aligned head-centered visual and tactile maps. *Nature neuroscience*, 9(10), 1337.
- Younger, B. (1985). The segregation of items into categories by ten-month-old infants. *Child Development*, 56, 1574–1583.

# Why blame Bob?

## Probabilistic generative models, counterfactual reasoning, and blame attribution

John McCoy<sup>1\*</sup> (jmccoy@mit.edu), Tomer Ullman<sup>1\*</sup> (tomeru@mit.edu), Andreas Stuhlmüller<sup>1</sup> (ast@mit.edu), Tobias Gerstenberg<sup>2</sup> (t.gerstenberg@ucl.ac.uk) & Joshua Tenenbaum<sup>1</sup> (jbt@mit.edu)

<sup>1</sup>Department of Brain and Cognitive Sciences, MIT

<sup>2</sup>Cognitive, Perceptual and Brain Sciences, University College London, UK

\*These authors contributed equally to the paper.

### Abstract

We consider an approach to blame attribution based on counterfactual reasoning in probabilistic generative models. In this view, people intervene on each variable within their model and assign blame in proportion to how much a change to a variable would have improved the outcome. This approach raises two questions: First, what structure do people use to represent a given situation? Second, how do they choose what alternatives to consider when intervening on an event? We use a series of coin-tossing scenarios to compare empirical data to different models within the proposed framework. The results suggest that people sample their intervention values from a prior rather than deterministically switching the value of a variable. The results further suggest that people represent scenarios differently when asked to reason about their own blame attributions, compared with the blame attributions they believe others will assign.

**Keywords:** counterfactuals; blame attribution; probabilistic models; causal reasoning

### Introduction

Alice and Bob play a coin-tossing game. If their coin tosses match, they win. Alice goes first and tosses heads, Bob goes second and tosses tails, and hence they lose. Who, if anyone, will be blamed? Counterfactually, what would have happened if Alice had tossed heads? One intuition is that how much someone will be blamed for an outcome is closely related to how strongly they affected the outcome (cf. Spellman, 1997). Through counterfactual thinking, people can reason how a change in the past would have affected the present and use such reasoning for cognitive tasks including social judgments, causal attribution, problem solving, and learning (see Roese, 1997; Byrne, 2002, for reviews). But how do people reason counterfactually? And what is the relationship between counterfactual thinking and blame attribution?

Psychological research on counterfactual reasoning has revealed factors that influence which events attract counterfactual thoughts, including unusual events (Kahneman & Miller, 1986), early events in a causal chain (Wells, Taylor, & Turtle, 1987), and late events in a temporal chain (Byrne, Segura, Culhane, Tasso, & Berrocal, 2000). There have also been formal accounts which aim to explain the empirical findings in terms of principled mental operations that do not depend on event features (Spellman, 1997; Byrne, 2002; Chockler & Halpern, 2004; Rips, 2010; Petrocelli, Percy, Sherman, & Tormala, 2011). Some of these formal models have been

separately tested against empirical data (Sloman & Lagnado, 2005; Gerstenberg & Lagnado, 2010).

Kahneman and Tversky (1982) suggest that people reason counterfactually by using a “simulation heuristic”, whereby they mentally alter events and run a simulation of how things would have gone otherwise given these changes. In this paper, we use a computational-level framework that formalizes the spirit of this suggestion: when attributing blame, people mentally alter each possible event in turn, consider the consequences for the outcome, and blame an event in proportion to how much the change would have improved the outcome.

We model this computation of counterfactual consequences using interventions on causal models (Pearl, 2000). We explore what causal models people use to represent the games in our experiments and how they choose alternatives when intervening on a particular event.

The plan for the paper is as follows. We first describe the formal framework this work is based on and the space of models we explore. We then report results of experiments in which we varied aspects of the coin-tossing game described above, and suggest a possible explanation for these results within our framework. We conclude by discussing implications and limitations of this account, and possibilities for future research.

### Formal framework

We assume that, when reasoning counterfactually, people represent the situation they are reasoning about using a probabilistic generative framework. Probabilistic models have been used to explain many aspects of high-level cognition, including perception, prediction, decision making and social reasoning (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). In this paper, we use causal Bayes nets and the functional equations they are derived from as the underlying probabilistic generative framework (Pearl, 2000). Other representations are possible—see, for example, Gerstenberg and Goodman (in prep) for an approach to counterfactual reasoning based on probabilistic programs.

We model people’s reasoning about blame as follows. First, consider each event in the situation—represented by a variable in a causal Bayes net—and intervene on it, i.e., consider a counterfactual value for this event (‘do’ in Pearl, 2000). Each such intervention results in a distribution over

counterfactual worlds, where a counterfactual world is an assignment of values to variables. Next, compare these distributions to the actually observed world in order to assign blame. The variable being intervened on is assigned a degree of blame proportional to the difference in expected utility between the counterfactual world and the actual world. When interventions are chosen stochastically, we take the expectation over intervention values.

In our experiments, there are two possible outcomes, a win (1) and a loss (0), and the actual outcome is described as a loss. In this setting, our model of blame judgments is equivalent to assigning a degree of blame to a variable in proportion to the probability of reaching a counterfactual world in which the game is won after intervening on the variable. This combines the idea of using the ‘do’ operator as a psychologically plausible basis for counterfactual thinking (e.g., Sloman & Lagnado, 2005) with the idea of assigning causality by considering how each of the events in a given scenario affects the outcome (e.g., Spellman, 1997).

The framework as described presents at least two open questions about how people reason counterfactually to attribute blame.

**What generative structure do people use to represent a situation?** Even for simple scenarios, there exists a rich space of possible representations. Consider the coin-tossing game described in the introduction. Previous work suggests that people sometimes believe themselves or others capable of control over events such as coin tosses that are in fact random (Langer, 1975; Shafir & Tversky, 1992). Based on this work, we consider two simple generative models for the coin-tossing game, a *no-control* model that represents the coin tosses as independent, and a *control* model that represents the players as having some control over their tosses, assumes that they have knowledge of the previous player’s toss, and that they try to match this toss (see Figure 1). This ‘control’ is captured in the following way: If the coin tossed by the first player came up 1 (‘heads’), the bias for the second player’s coin getting 1 is now  $\alpha > 0.5$ . If the first player tossed 0 (‘tails’), then the bias for the second player’s coin getting 0 is now  $\alpha > 0.5$ . When we compute counterfactuals in the ‘control’ setting, we use the following functional equations that reflect this idea:  $u_1 = \text{Bernoulli}(\theta_1)$ ,  $u_{2a} = \text{Bernoulli}(\alpha)$ ,  $u_{2b} = \text{Bernoulli}(1 - \alpha)$ ,  $C_1 = u_1$ , and  $C_2 = u_{2a}$  if  $C_1 = 1$ , otherwise  $u_{2b}$ . We later discuss a perspective-dependent model, which combines the ‘control’ and ‘no-control’ models.

**What new value do people assign to the intervened-upon variable?** The ‘do’ operator produces counterfactual worlds when given a variable to intervene on and a value to set the variable to, but there is a question about what value people use. One possibility is that people set the intervened-upon variable to be different from what it was before the intervention. The idea of *only* considering alternatives to reality when reasoning counterfactually has intuitive appeal. This approach is taken by Gerstenberg, Goodman,

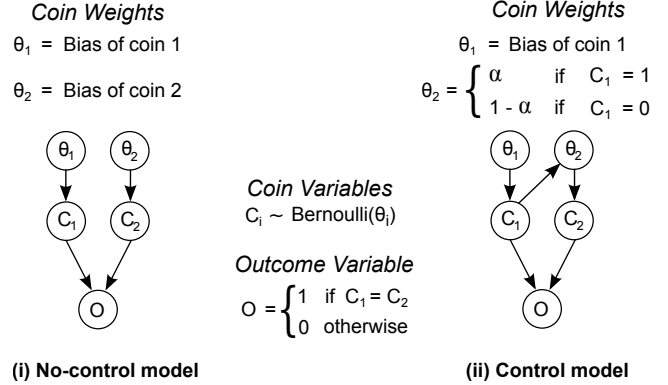


Figure 1: Different generative structures for the coin-tossing game. Coin tosses are drawn from a Bernoulli distribution with the coin bias as its parameter. ‘heads’ and ‘tails’, and ‘win’ and ‘lose’ are replaced by 1 and 0. (i) In the ‘no-control’ version the bias is simply the unchanged bias of the coin. (ii) In the ‘control’ version players are represented as having some influence over the coin, as is formalized in the text.

Lagnado, and Tenenbaum (2012) in reasoning counterfactually about physical events. In the case of binary variables, as in the current study, this involves simply switching the observed value, hence we refer to this way of choosing intervention values as *intervene-switch*. The generalization to non-binary variables is not immediate and is not considered here.

A second possibility is to draw the intervention value from the (conditional) prior over the variable being intervened upon. This allows an assessment of how unlikely the counterfactual event is, which has been stressed as an important factor by Petrocelli et al. (2011). We refer to this possibility for setting intervention values as *intervene-prior*. Figure 2 illustrates these two possibilities. A third possibility is that people choose an intervention value optimally, in such a way as to maximize the expected utility of counterfactual worlds. In our models for the coin-tossing domain, this possibility coincides with *intervene-switch*, hence we do not discuss it separately.

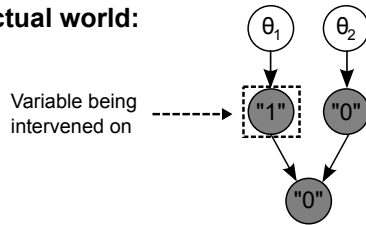
Each combination of these two factors—which causal structure to use, and how to choose an intervention value—results in a different model of blame attribution within our general framework. There are other factors which remain the subject of future research and which will almost certainly be necessary to predict subjects’ judgments in richer situations. However, even these two factors offer a space for exploring how people reason counterfactually and assign blame.

### A simple example of predicting blame judgments

Consider again the coin-tossing game with a fair coin, in which the player going first ( $C_1$ ) tosses heads and the player going second ( $C_2$ ) tosses tails, resulting in a loss. Who is to blame for the loss? We examine the predictions of four different models. In this scenario, some of the models make the same predictions; other scenarios used in our experiments provide additional discriminatory power.

1. *no-control/intervene-prior*: Coin tosses are independent

## 1. Actual world:



## 2. Intervention:

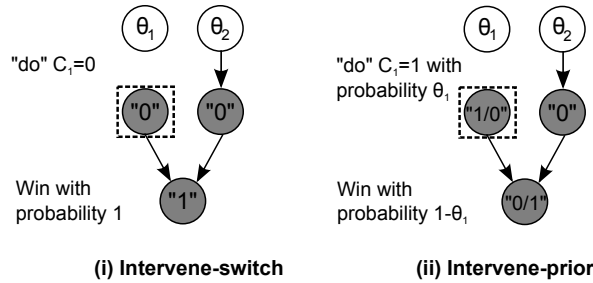


Figure 2: How to choose a new value for the intervened-upon variable, using the coin-tossing game as example. In the actual world the first player tossed heads (1) and the second player tails (0), and they thus lost. In intervening upon the first player's coin toss there are two possibilities: (i) *intervene-switch*: the first player's coin toss is changed to be the opposite of what it was (i.e. from 1 to 0), (ii) *intervene-prior*: intervention values are chosen from the prior and so there is a  $\theta_1$  chance of setting the first player's toss to heads and a  $1 - \theta_1$  chance of setting the first player's toss to tails.

and values for interventions are chosen by drawing them from the variable prior. We begin by considering an intervention to the variable  $C_1$ . Because the coin is fair, with probability 0.5 we choose the intervention 'tails', resulting in the outcome variable being assigned a 'win', as both players now tossed matching coins (counterfactually speaking). With probability 0.5 we choose the intervention 'heads', resulting in a 'loss'. So, by intervening on  $C_1$  we improve the odds of winning from 0 (actual observation) to 0.5, and the amount of blame assigned to  $C_1$  is 0.5. What about the second player's blame? For  $C_2$  the process is the same as  $C_1$ , thus the blame for  $C_2$  is also 0.5. Hence, this model predicts no difference in blame between  $C_1$  and  $C_2$ .

2. *control/intervene-prior*: Players are represented as having some control over their coin (Figure 1(ii)) and interventions are selected in the same way as before. The intervention on  $C_1$  works as described above, resulting in a 0.5 amount of blame. However, in intervening on  $C_2$ , we draw from a prior skewed towards 'heads' (the coin now has a bias  $\alpha > 0.5$  towards heads). With probability  $\alpha$  we choose the intervention 'heads' for  $C_2$ , and with probability  $1 - \alpha$  we choose 'tails'. This results in a 'win' with probability  $\alpha > 0.5$ , meaning  $C_2$  will be blamed more than  $C_1$ .
3. *no-control/intervene-switch*: Coin tosses are independent and the value used for an intervention is different from the observed value of the variable. Because  $C_1$  was observed to be 'heads', an intervention always sets it to 'tails', resulting in a 'win' with probability 1. This model also predicts no difference in blame between the players.
4. *control/intervene-switch*: Players are represented as having

some control over their coin (Figure 1(ii)) and the value used for an intervention is different from the observed value of the variable. Using this model, the control players have does not make a difference, as the intervention on  $C_1$  is always 'tails' (switching it from 'heads'), and the intervention for  $C_2$  is always 'heads'. In a way similar to the previous variant, the probability for a 'win' resulting from intervening on either  $C_1$  or  $C_2$  is 1, and thus both players receive the same blame.

## Experiment

### Procedures and methods

One hundred subjects per condition were recruited on Mechanical Turk. Approximately twenty subjects were dropped in each condition for failing comprehension questions. We presented descriptions of simple scenarios involving blame attribution in the aforementioned coin-tossing game. All scenarios share the following:

- (1) An introduction describing the game: Each person tosses a coin in turn. If all coins land the same, the players each receive \$1000, and otherwise receive \$0.
- (2) Subjects were told the order of play, and the result of each coin toss (e.g., "Bob tosses his coin first. It comes up heads. Then you toss a coin. It comes up tails.").
- (3) The end result was a loss.
- (4) Subjects were asked, depending on condition, how much blame they attributed to the players, or how much blame they believed the players would attribute.
- (5) Subjects responded using a discrete 1-7 scale, with 1 marked *minimal blame* and 7 marked *maximal blame*.

### Results and discussion

We use the space of models we have discussed to examine predicted differences in blame attribution, and test these predictions using one-sided Student *t*-tests.

**1. Same room, subject not involved** Subjects read descriptions of Player 1 and Player 2 playing the coin game. Player 1 tosses heads, then Player 2 tosses tails. Subjects were asked how much Player 1 would blame Player 2, and how much Player 2 would blame Player 1. The results of these and subsequent experiments, as well as the model predictions, are shown in Figure 3. This first experiment replicates the temporality effect (Miller & Gunasegaram, 1990) in the sense that subjects believe that Player 1 will blame Player 2 more than Player 2 will blame Player 1 ( $p < 0.0001$ , cf. Figure 3a). The only one of the four simple models under consideration which accords with these results (rows 3-6 in Figure 3) is the model which assumes causal control and draws intervention values from the prior. However, experimental results to be discussed shortly imply that the situation is more subtle. We first test the sensitivity of assumptions about causal control to knowledge about epistemic access (i.e. knowing the result of the the other player's coin toss).



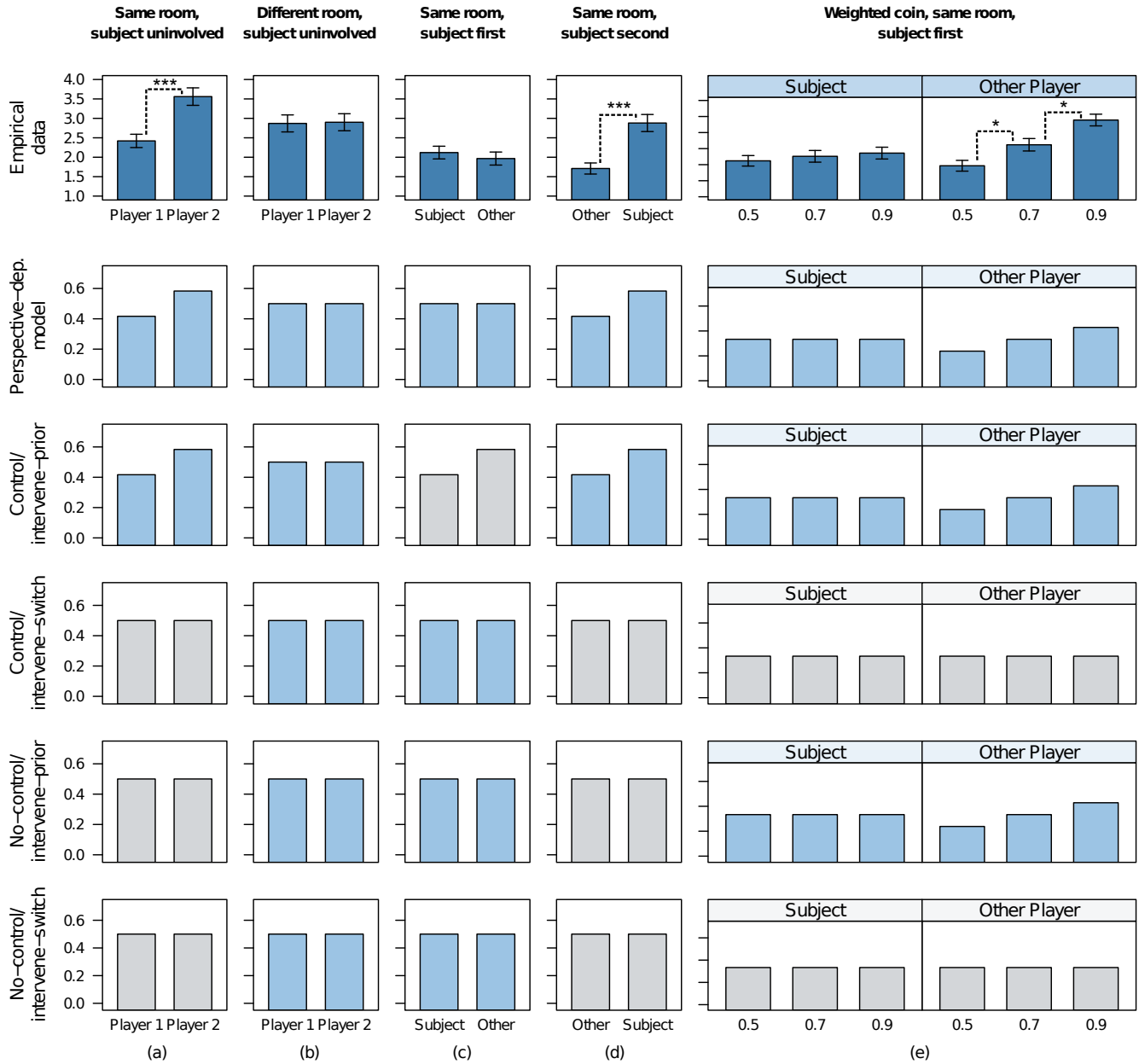


Figure 3: **Empirical results and model predictions for the coin-tossing game.** Columns represent separate experiments, with the empirical data shown in the top row. The y-axis is the mean amount of blame subjects believe will be assigned to the player on the x-axis, by the other player in the game. For example, in Experiment 1 subjects believe that Player 1 is assigned a 2.42 blame rating by Player 2, while Player 2 is assigned a 3.56 blame rating by Player 1. In each sub-figure, the player on the left side of the x-axis is the one going first. For example, in Experiment 1 Player 1 tosses the coin first. Model predictions are colored blue if they qualitatively match the ordinal blame judgments in the empirical results, and are grey otherwise. The amount of blame predicted by the models is normalized. We set  $\alpha = 0.7$  in the above, but this makes no difference to the qualitative results.

**2. Different rooms, subject not involved** Subjects read descriptions of a game that only differed from the one in the previous experiment by players being in different rooms, such that Player 2 was unaware of the result of Player 1's coin toss. In this experiment, subjects predict that Player 1 will blame Player 2 significantly less than in the 'same room' scenario ( $p < 0.05$ , cf. Figure 3b vs. a). There is no significant difference in the amount of blame attributed by the first and second player.

One explanation of these findings is that causal control is only believed possible when the second player is aware of the first player's coin toss. That is, subjects have a sophisticated model of the situation that treats the other players as agents that have the capacity for control, but that requires epistemic access for the agents to make use of this capacity.

Lack of epistemic access is not the only way to explain the results of this experiment. The two models based on setting intervention values by switching the observation (*intervene-switch*) are also consistent with these results. We will shortly provide independent evidence for sampling from the prior over switching, but first we manipulate whether subjects report their own blame judgments or their predictions about the judgments of others.

The 'classic' temporality effect replicated in Experiment 1 may strike some readers as odd. Surely one person is not to blame more than another in a purely random game? To examine this intuition, which is supported by previous research (Mandel, 2003), we now compare first-person and third-person blame judgments.

**3. Same room, subject involved** The game was described as in Experiment 1, but the subject was described as playing the game with another player (denoted 'Other Player' in Figure 3). One group of subjects was told that they tossed first, while another group was told that they tossed second. In both cases the player going first tosses 'heads' and the player going second tosses 'tails'. Subjects were asked how much they themselves would blame the other player, and how much they believed the other player would blame them. When subjects are asked about how much they think the other player will blame them, the temporality effect is replicated. That is, subjects believe that the other player will blame them significantly more when subjects flip second rather than first ( $p < 0.001$  cf. blame to 'Subject' in Figure 3c and d).

Crucially, however, when comparing the amount of blame subjects attribute to the other player, we find no effect of position (cf. blame to 'Other' in Figure 3c and d). That is, the temporality effect does not exist when subjects are asked about the amount of blame they themselves attribute. As in Experiment 2, the *no-control/intervention-prior* model is consistent with the lack of a temporal effect, as are the two models in which intervention values are set by switching the observation. One hypothesis is thus that subjects do not themselves attribute causal control when assessing blame, but believe that other people do so (this possibility is represented by the 'perspective-dependent' model shown in the second row

of Figure 3). It is also possible, however, that people think about choosing the intervention values differently depending on whether they are taking a first or third person perspective. That is, when considering how other people model a situation, they draw interventions from a prior, but when reasoning about their own perspective, they draw interventions by switching variables. The latter suggestion appears overly complex, but is not ruled out by the evidence presented so far.

We thus consider an experiment aimed specifically at examining whether people do set intervention values by sampling from the prior.

**4. Biased coin, subject involved** Subjects were told that they participated in the game with one other player. In one case, they were told that the other player used a coin biased 70% towards heads, in another case biased 90% towards heads. In both cases, subjects were told that they went first and got heads, and the other player tails. Subjects were asked how much they blamed the other player. We find that the greater the bias of the coin, the more the other player is blamed. The other player is blamed significantly more when the bias is 0.7 than when the coin is fair ( $p < 0.05$ , data from Experiment 3 were used for this comparison), and significantly more when the bias is 0.9 ( $p < 0.05$ ; cf. blame to 'Other' in Figure 3e). Subjects' ratings about how they themselves will be blamed were not affected by the bias of Other's coin (cf. blame to 'Subject' in Figure 3e). For modeling this situation, we make the simplifying assumption that the second player's coin is entirely dependent on the specified bias, rather than on any causal control.

The models which are consistent with the effect of coin bias are those where values for the intervention are drawn from the prior. The results are analogous to experiments showing the tendency of people to focus on more unusual events when asked to reason counterfactually (Kahneman & Miller, 1986).

Given the small space of models we consider, one parsimonious account of the empirical data is that both when attributing blame themselves and when reasoning about how others attribute blame, people draw values for the intervention from the prior. Our results further suggest that people may assume that other people believe in causal control of random events, but that they themselves do not. This model of causal control seems to be sensitive to factors such as the other player's state of knowledge.

## General discussion

We have explored aspects of a psychological framework for counterfactual reasoning, focusing in particular on its use for blame attribution. We have assumed that people represent the situation using probabilistic generative models and that they assign blame to an event by determining the counterfactual consequences of intervening on this event.

Within this framework, we have investigated what causal structure people use to represent a given situation and how people assign values to intervened-upon variables in the do-

main of coin-tossing games. In this domain, psychological effects such as the temporality effect and the tendency to focus on unusual events seem to arise naturally from counterfactual reasoning using probabilistic models.

We have presented data that suggests that people sample the values of their interventions from a prior rather than deterministically switching the variable to an alternate, unobserved value. Experimental evidence suggests that people may represent situations involving random events differently when reasoning about their own judgements compared to their predictions about the judgements made by others. People may model others as believing that such situations involve causal control, but they themselves may not believe in such control. One possible explanation for this perspective-dependent representation is that people model others' views in less detail than their own. For example, people may have a default assumption of causal control, but in situations such as coin games, they may be able to suppress this assumption. This suppression may take additional resources which, in general, may not be available or used when predicting how others represent the same situation.

This perspective-dependent difference in representation or computation provides an intriguing avenue for future research. For example, does there exist a similar difference in games of skill where control is the correct assumption? Beyond perspective-dependent differences, it is almost certain that different people model identically described situations in different ways, which suggests a per-subject analysis in addition to the aggregate approach taken in this paper.

There are many other ways in which our modeling and experimental results can be extended, even in the simple coin-tossing game. First, we do not explicitly take into account the prior probability of winning a game. Second, we do not explore situations in which the value of multiple variables would need to change to result in a win, for example a situation in which six people were playing the game with two people tossing tails and four people tossing heads. Third, while our model predictions are quantitative, we have restricted ourselves here to a qualitative analysis. Fourth, while we have modeled agents in some situations as having causal control, we did not give a full account of agents as having — and reasoning about—intentionality, foresight, and complex epistemic states, which are known to affect blame attribution (Lagnado & Channon, 2008). To capture the subtlety of human blame attribution and counterfactual thinking, richer models which include a more sophisticated representation of agents and their beliefs will be necessary.

## Acknowledgments

This work was supported by ARO (W911NF-08-1-0242), ONR (N00014-09-0124), an NSF Graduate Fellowship (TDU) and a doctoral grant from the AXA research fund (TG).

## References

Byrne, R. M. J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Science*, 6(10),

- 426–431.
- Byrne, R. M. J., Segura, S., Culhane, R., Tasso, A., & Berrocal, P. (2000). The temporality effect in counterfactual thinking about what might have been. *Memory and Cognition*, 28(2), 264–281.
- Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22, 93–115.
- Gerstenberg, T., & Goodman, N. (in prep).
- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Gerstenberg, T., & Lagnado, D. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, 115(1), 166–171.
- Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93(2), 136–153.
- Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). Cambridge University Press.
- Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, 108(3), 754–770.
- Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32(2), 311–328.
- Mandel, D. R. (2003). Judgment dissociation theory: An analysis of differences in causal, counterfactual and covariational reasoning. *Journal of Experimental Psychology: General*, 132(3), 419–434.
- Miller, D. T., & Gunasegaram, S. (1990). Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of Personality and Social Psychology*, 59(6), 1111–1118.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.
- Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of personality and social psychology*, 100(1), 30–46.
- Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, 34(2), 175–221.
- Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, 121(1), 133–148.
- Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, 24(4), 449–474.
- Sloman, S., & Lagnado, D. (2005). Do we “do”. *Cognitive Science*, 29, 5–39.
- Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, 126(4), 323–348.
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285.
- Wells, G. L., Taylor, B. R., & Turtle, J. W. (1987). The undoing of scenarios. *Journal of Personality and Social Psychology*, 53(3), 421–430.

# Modeling online word segmentation performance in structured artificial languages

**Stephan Meylan**

smeylan@stanford.edu  
Department of Psychology  
Stanford University

**Chigusa Kurumada**

kurumada@stanford.edu  
Department of Linguistics  
Stanford University

**Benjamin Börschinger**

benjamin.borschinger@mq.edu.au  
Department of Computing  
Macquarie University

**Mark Johnson**

mark.johnson@mq.edu.au  
Department of Computing  
Macquarie University

**Michael C. Frank**

mcf Frank@stanford.edu  
Department of Psychology  
Stanford University

## Abstract

Lexical dependencies abound in natural language: words tend to follow particular words or word categories. However, artificial language learning experiments exploring word segmentation have so far lacked such structure. In the present study, we explore whether simple inter-word dependencies influence the word segmentation performance of adult learners. We use a continuous testing paradigm instead of an experiment-final test battery to reveal the trajectory of learning and to allow detailed comparison with three computational models of word segmentation. Adult performance on languages with dependencies is equal or lower to those without. Of the models tested, all perform worse on languages with dependencies, though a novel particle filter-based lexical segmentation model produces learning curves most similar to human subjects.

**Keywords:** Word segmentation; statistical learning; bigrams; computational modeling; dependency structures.

## Introduction

Human learners can use distributional information to segment an unbroken speech stream into individual words after a short, ambiguous exposure (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). Past artificial language learning experiments have typically generated words according to a uniform word frequency distribution and randomly concatenated word types to create the artificial languages. The process of learning to segment such synthesized languages may deviate significantly from learning to segment natural language, which contains asymmetric word frequency distributions, systematic dependency structure, and correlated variation in the lengths and frequencies of words.

In natural language, word-by-word transitions are governed by dependency relationships between word categories. For example, English prepositional phrases have the internal structure  $PP \rightarrow P + NP$ , and an NP typically consists of a determiner and a noun. Since prepositions and determiners are lexical classes containing a relatively small number of short words, many instances of PPs often result in collocations used frequently in discourse (e.g., *in the house* or *on a map*).

Does such collocation structure make segmentation easier or more difficult? Both possibilities have some support in the literature. Frequent phrases are known to be problematic for segmentation mechanisms, especially for algorithms that rely on transitional probabilities (TPs). Due to high internal TPs, these phrases are often segmented as one unit, rather than separated into the multiple words that they contain (Goldwater,

Griffiths, & Johnson, 2009). Thus, dependency structure has the potential to reduce segmentation performance.

In principle, however, there may also be the potential for increased performance in segmenting structured languages, especially if the learner is able to learn the dependency structure of the language along with the structure of individual words. For this reason, Goldwater et al. (2009)'s bigram model outperformed other segmentation models. Modeling dependency structure might also provide synergistic gains in learning: Johnson and Tyler (2010) found that an ideal learner that acquired words and word-object correspondences simultaneously was far more successful at both than each independently, but only when the learner assumed a rich collocation structure in the language.

Our current experiments test the relationship between dependency structure and segmentation performance for both human learners and computational models. We created languages with varied levels of category size asymmetry and test adult subjects' performance in word segmentation based on these languages. Two-alternative forced-choice tests administered after a discrete training phase, as used by past studies, do not produce an interpretable time course of learning. Thus, in Experiment 1 we use a new experimental paradigm that provides us with a time course of learning by testing subjects throughout the duration of exposure to stimuli (Kurumada, Meylan, & Frank, under review). In Experiment 2, we corroborate the results of the first experiment using a classic two-alternative forced choice paradigm. Both experiments, and a set of simulations with three computational models, show that asymmetric word-category sizes support adult segmentation learning considerably better than symmetric category sizes, but that performance on languages with dependencies is generally worse than performance on languages with randomly ordered words.

## Experiment 1

To test the effects of dependency structure on word segmentation, we created two classes of artificial lexicons, one consisting of 12 word types and another of 8, and concatenated them to make languages with different dependency structures. Figure 1 is a diagram of the grammars that we used to produce the 12 word languages. Each sentence had four words. Three of the language types (which we refer to as "1515," "2424,"

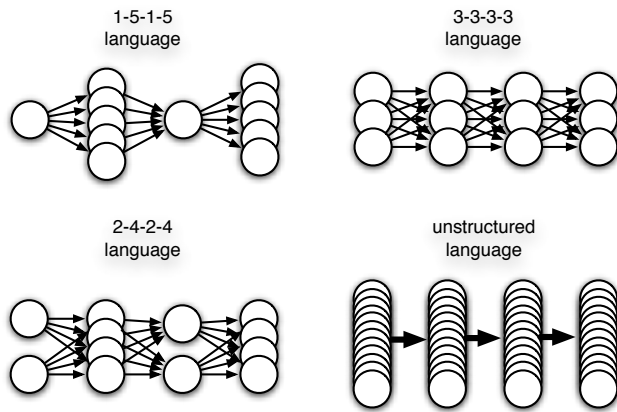


Figure 1: Schematic representations of the 4 types of languages used in Experiment 1. Circles represent individual words, arrows represent stochastic transitions.

and “3333”) were generated via a simple finite state grammar, while the fourth (unstructured) language type was generated by uniformly concatenating words with no notion of sentence position.

Each of the structured languages had a category structure such that sentences followed the form *ABCD*, where each category had a unique and non-overlapping set of words in it. The 1515 language, for example, had one word in the first category, five in the second, and so forth. We refer to such a language throughout as having “asymmetric” category sizes. In contrast, a 3333 language has “symmetric” category sizes. The 8 word languages were generated similarly, but only included three conditions (“1313,” “2222,” and unstructured).

We used a sentence-by-sentence segmentation paradigm (Kurumada et al., under review) to test adult learners’ segmentation performance and gather detailed information about learning trajectories in these languages. Learners listened to a set of sentences presented in a Flash web applet; after each sentence they were asked to click between syllables in an orthographic transcription to indicate where they thought word boundaries fell.

## Methods

**Participants** For the 12 word languages, we posted 128 separate assignments on Amazon’s Mechanical Turk (AMT) crowdsourcing platform and received 124 assignments from distinct individuals. For the 8 word languages, we posted 96 assignments and received 95. Subjects were paid \$1.00 for completing the task and were told that they would be paid an additional \$1.00 if they performed in the top quartile of participants. The mean task duration was 14 minutes and 17 seconds. Subjects needed to transcribe a common English word at the beginning of the task to access the task to confirm that they were English speakers.

**Stimuli** We created two classes of lexicons differing in the total number of word types (8 and 12). Words were made

by concatenating 2 – 4 syllables from a randomly selected inventory of consonant-vowel syllables. Stimuli were synthesized with the MBROLA synthesizer (Dutoit, Pagel, Pierret, Bataille, & Van Der Vrecken, 1996) at a constant pitch of 100Hz with 225ms vowels and 25ms consonants. Each syllable appeared in only one word type in each lexicon.

For the 12-type condition, 4 distinct languages were created by manipulating how many of 12 types (distinct word forms) appeared in each of four sentences positions (Figure 1): 1515, in which the words in the first and third sentence position were drawn from categories of a single type and the second and the fourth from categories with five types each, 2424, in which the words in the first and third sentence position were drawn from categories consisting of two types and the second and fourth from categories with four types, 3333 in which the word in each sentence position was selected from a category with three types, and unstructured, in which four words were selected uniformly at random from a single category of 12 types (without replacement, to avoid in-sentence repetition).

To reflect the relationship between frequency and word length seen in natural language (Zipf, 1965), the category assignment of lexical items ensured word length and frequency were inversely correlated (the shortest words appeared in the categories with the fewest types). Within each condition, 16 language variants were made using different phonemic inventories to control for unwanted phonological effects. The total number of word tokens per subject was 240 and the number of sentences was 60 in all languages. The total token frequencies of words were 60-12-60-12 in the 1515 condition, 30-15-30-15 in the 2424 condition, 20-20-20-20 in the 3333 condition. In the randomized condition each word appeared in each sentence position 5 times. Note that there was no discrete testing phase separate from the training phrase; rather, subjects were tested in the continuous paradigm on their knowledge of the language as they learned it.

Stimuli for the 8-type condition were similar to those in the 12-type condition except that we created only 3 languages: 1313, with the first and third word position drawn from categories with a single type and the second and fourth position drawn from categories with three types in each, 2222, in which each word position is selected from a category with two types, and randomized, where each sentence was composed from four words randomly selected from all 8 word types. 32 phonetic variants of each language were generated to control for phonological effects. The per-subject exposure was 240 tokens over 60 sentences, and the total token frequencies were 60-20-60-20 in the 1313 condition, and 30-30-30-30 in the 2222 condition; in the unstructured condition each word appeared 7 or 8 times in each position.

**Procedure** Before the experimental trials began, participants were instructed to listen to and transcribe a short, common English word to confirm that their computer’s audio system was working. Participants were then instructed that they would be presented with 60 consecutive sentences in a novel

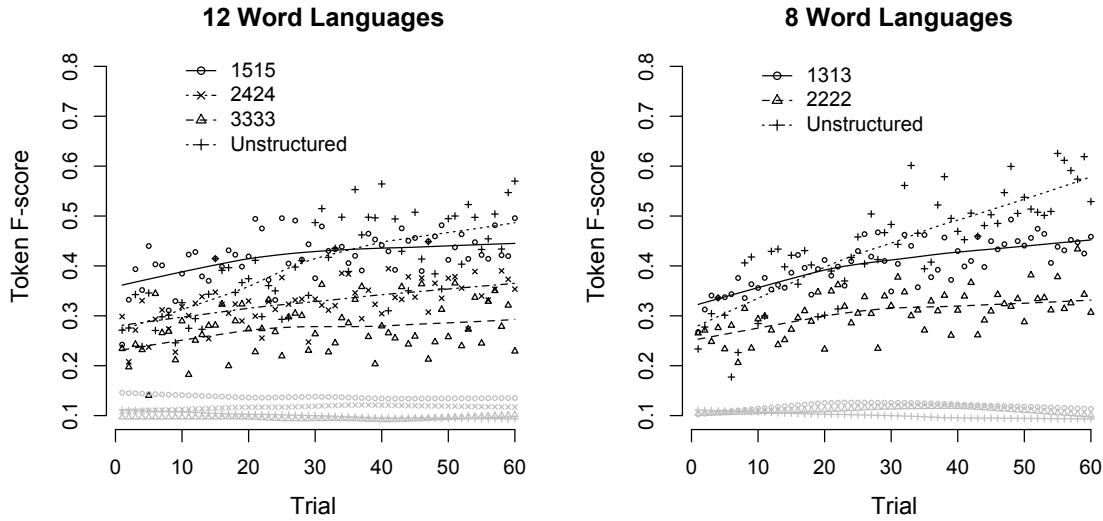


Figure 2: Token F-scores (a measure of segmentation performance for individual words) plotted for Experiment 1. Symbols represent mean performance for each condition (see legend) on a single trial, across participants. Black lines show a smoothed estimate of performance (Lowess curves with a smoothing span of .5). Gray lines at bottom show permutation baselines based on mean token F-score from reshuffling participant responses 1000 times.

artificial language, and that from the beginning they would need to click the boundaries between syllables presented on screen to indicate what they thought the boundaries between words were in each trial. While they would not know the words at first, they were told that they would be able to discern patterns and recognize at least some of the words as the experiment progressed. Each trial contained both an audio presentation and an orthographic transcription with selectable boundaries between each syllable. After clicking segments in a sentence subjects clicked a button marked “next” to proceed to the next trial.

## Results and Discussion

To assess subjects’ segmentation performance, we calculated the precision, recall, and F-score (the harmonic mean of precision and recall, as described in Goldwater et al., 2009). This metric was computed by subject and trial, both for boundary decisions and tokens.<sup>1</sup> Figure 2 shows token F-score aggregated across subjects for each language and condition.

All conditions showed some learning. Unstructured languages were learned best, and learning was faster in the 8 word than the 12 word languages. Trajectories of the structured conditions maintained a relatively constant ordering:

<sup>1</sup>In our example sentence (“indiangorillaseatbananas”), we compute these measures for a participant who gave the segmentation “indian|gorillas|eatbana|nas.” Computing word boundaries, the participant would have 2 hits, 1 miss, and 1 false alarm, leading to precision of .66 (hits / (hits + false alarms)), and recall of .66 (hits / (hits + misses)), for an F-score of .66. On the other hand, for word tokens, the participant would have 2 hits (“indian” and “gorillas”), 2 misses (“eat” and “bananas”) and 2 false alarms (“eatbana” and “nas”), for precision of .5, recall of .5, and F-score of .5.

the more asymmetrical the categories in the language were, the better participants learned. In both the 8-type and the 12-type, the slope of the unstructured condition is steepest, demonstrating the greatest amount of learning of the latent structure.

We analyzed token-by-token segmentation performance separately for the 8 word and the 12 word languages using mixed-effects logistic models (Gelman & Hill, 2006; Jaeger, 2008). The dependent variable in these models was whether a particular token had been segmented correctly. In the 12 word languages, we found a strong positive main effect of the log input frequency of that token ( $\beta = .19, p < .001$ ) and a negative effect of word length ( $\beta = -.56, < .001$ ), as well as significant effects of the 3333 test condition ( $\beta = -.69, p < .05$ ) and the 2424 test condition ( $\beta = -.62, p < .05$ ) with respect to the unstructured condition. The 1515 condition also had a negative effect on token segmentation ( $\beta = -.29$ ), though this was not reliably different than the unstructured condition ( $p > .3$ ). As in the 12 word languages, in the 8 word languages there was a positive main effect of the log input frequency of that token ( $\beta = .26, < .001$ ) and negative effect of word length ( $\beta = -.68, < .001$ ). There were also significant effects of the 1313 condition ( $\beta = -.71, p < .05$ ) and the 2222 condition ( $\beta = -.95, p < .01$ ).

To summarize: in both the 12 word and the 8 word languages we observed the best performance in the unstructured condition and the worst performance in the condition where types are symmetrically distributed across categories/sentence positions.

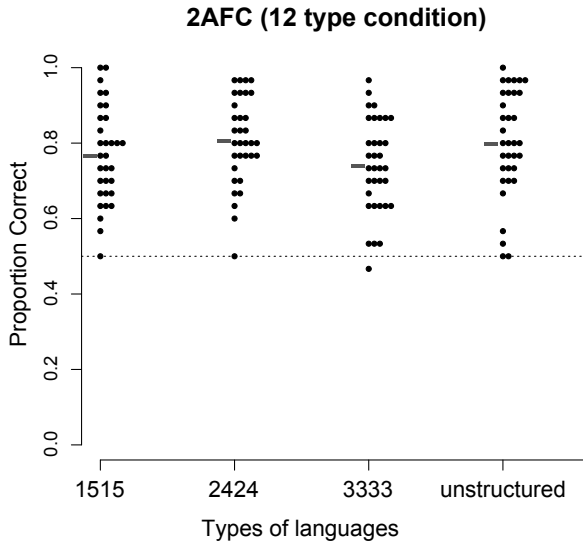


Figure 3: Proportion of correct 2AFC trials for the languages in the 12-type conditions in Experiment 2. Each point represents a participant. The black bar shows the condition mean and the dotted line at-chance (50%) performance.

## Experiment 2

Experiment 2 was conducted to ensure that the effects of dependency structure could be captured in a classic two-alternative forced choice task, where subjects were asked to distinguish between a word from the language and a distractor.

### Methods

**Participants** For the 12 word languages, 144 assignments were posted on AMT, of which we received 133 completed by distinct individuals. For the 8-type condition, 96 assignments were posted, of which we received 89. The same payment method as in Experiment 1 was used.

**Stimuli** The process of generating stimuli was nearly identical to the procedures presented in Experiment 1. For the training phase, 150 input sentences (totaling 600 tokens) were generated. A discrete test phase consisted of 30 pairs of a target word from the language and a length-matched distractor word, composed of syllables randomly selected from that language’s syllabic inventory. Such test trials were intended to test if subjects had reliably learned which words belonged to the language.

**Procedure** Participants were instructed that they would listen to 150 sentences in a novel artificial language, then take a short test on what they had learned about the language. Training sentences were presented aurally; when the audio file finished playing subjects needed to press “next” to advance to the next trial. In the test phase they were asked to choose which of the two options “sounded like it came from the lan-

guage.” For the test trials, the two options were presented aurally and subjects could replay the audio if desired.

## Results and Discussion

In the 8 word languages, performance was at ceiling, with all condition means above 90% in the forced choice test. As such there was no meaningful variance across languages and we do not discuss these conditions further. In the 12 word languages, the pattern of mean performance across conditions corroborated our findings in Experiment 1 (Figure 3). Although there was considerable variation across participants, performance was higher in the unstructured and asymmetric conditions; performance was lowest in the symmetric condition. We analyzed trial-by-trial 2AFC performance for the 12 word languages using a mixed-effects logistic model. There was a weak negative effect of trial number ( $\beta = -.02$ ,  $p < .001$ ; recall all learning happened in this experiment after the trial phase) and word length, though unlike in the first experiment *longer* word length facilitated segmentation in the 2AFC ( $\beta = .40$ ,  $p < .001$ ). The 3333 condition had a significant negative effect on performance ( $\beta = -.38$ ,  $p < .05$ ). Thus, 2AFC results support the negative effect of symmetrical category sizes seen in Experiment 1.

## Online Segmentation Models

Our goal in modeling human performance in Experiment 1 was to understand whether proposed ideal observer models are affected by structural complexity in the input in the same way as human subjects. We thus selected a range of models that have been suggested by previous literature to fit human performance: an incremental version of a transitional probability (TP) learner (Frank, Goldwater, Griffiths, & Tenenbaum, 2010); PARSER, a heuristic, memory-based model (Perruchet & Vinter, 1998); and a new online implementation of a probabilistic segmentation model (Börschinger & Johnson, 2011) using a particle filter. All models took unsegmented strings of characters representing syllables as input and produced segmented output sentence-by-sentence.

### Models and Parameters

**TP-based model** The transitional probability model is a boundary-finding approach that segments on the basis of statistically less likely transitions from syllable to syllable under the premise that within-word syllable TPs are higher than those at word boundaries (Saffran, Aslin, & Newport, 1996). In the present study, we calculate syllable bigram counts at the end of each sentence and calculate TP as

$$p(a|b) = \frac{c(a,b)}{\sum_{y \in V} c(a,y)} \quad (1)$$

where  $a$  and  $b$  are syllables,  $c(a,b)$  is the count of the bigram  $ab$ , and  $V$  is all bigrams observed up to that point in the corpus. Sentence boundaries, which contain potentially useful information, were limited by a special symbol and treated as syllables. We systematically varied the threshold (0 – 1 in



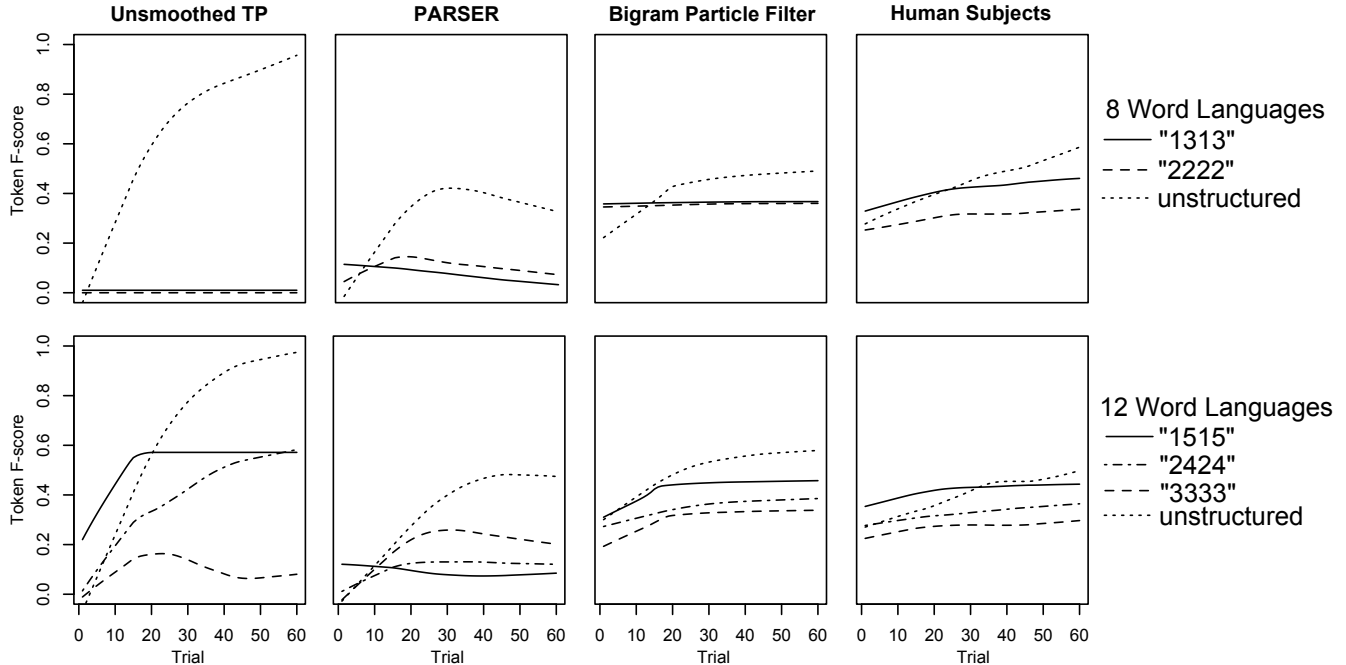


Figure 4: Lowess curves of mean token F-scores (smoothing span = .5) for the best fitting (using RMSE) parameter setting for each of the three models, plotted alongside human performance from Experiment 1.

.1 increments) at which a TP was low enough to constitute a word boundary, and placed boundaries between all syllables where TP was below the threshold.

**PARSER** PARSER (Perruchet & Vinter, 1998) is a lexicon-finding model that makes use of a persistent collection of segments whose weights increase when encountered in the input and decrease with exposure to new data lacking such segments. Found items similar to ones that are already in the collection also prompt a decrease in the weights of similar segments. The PARSER model has a high number of adjustable parameters, including initial weight, max segment length for consideration, the decay rate of material in the stored lexicon, the interference weight, a threshold to determine whether a parse should be considered a new word, and the reactivation gain. We adapted the model to output sentence-by-sentence segmentations, taking the initial parse of a new sentence (guided by the weights of the collection in memory) as the segmentation decisions for each sentence. For our simulations here, decay rate was most important, so we systematically manipulated this parameter. Among the other parameters, we used a max segment length for consideration of 3, an initial weight of 1, an interference rate of .005, a reactivation gain of .005, and a threshold of consideration as a new segment of 1.

**Bigram Model with Particle Filter Inference** The bigram model with particle filter inference is a new version of the Bayesian bigram model of Goldwater et al. (2009) that uses a particle filter rather than a Gibbs sampler to estimate the pos-

terior distribution over segmentations (Börschinger & Johnson, 2011). As in the original Bayesian bigram model, the lexicon for the particle filter is generated using a Dirichlet process, enforcing a probability distribution which gives higher probability to smaller lexicons with shorter words, and to a small number of high-probability collocations. Unlike the original version of the lexical model, however, the particle filter inference algorithm allows the model to be run incrementally through a corpus in a single pass, producing sequentially segmented sentences for our evaluation. In a particle filter, each particle constitutes a different set of segmentation hypotheses (see Börschinger & Johnson, 2011 for details). In our current study, we manipulated the number of different hypotheses that the model could track, with fixed concentration parameters of  $\alpha_0 = 12$  and  $\alpha_1 = 24$  in the 12 word languages and  $\alpha_0 = 8$  and  $\alpha_1 = 16$  in the 8 type languages.

## Model Results

For each model, we fit a single free parameter to human data, choosing one value that maximized Pearson's  $r$  and the one that minimized RMSE (threshold to treat as a boundary in the TP model, the decay rate of observed chunks in PARSER, and the number of particles for the particle filter bigram model). Figure 4 shows performance at the best-fitting parameter setting (according to RMSE) for each of the three models alongside the human data; Table 1 shows the RMSE and Pearson's  $r$  with the associated parameter setting in parentheses.

The bigram particle filter was the best-fitting model on both Pearson's  $r$  and RMSE. The particle filter also captured the

Model	Pearson's $r$	RMSE
Unsmoothed TP	.59 (.3)	.29 (.3)
PARSER	.70 (.05)	.23 (.008)
Particle Filter	<b>.71</b> (4)	<b>.07</b> (8)

Table 1: Two metrics of model fit for the three models

ordering of conditions observed in the human data, though the 8 type structured conditions were minimally distinguished from one another. PARSER performed well on  $r$ , but the forget rate (the fit parameter) with the highest correlation on  $r$  resulted in relatively low absolute performance (in the range of 0 - .2). Fitting PARSER to RMSE resulted in the selection of a much lower forget rate. The unsmoothed TP model performed poorly on both metrics. Though condition ordering for the TP model matched human results in the 12 type, both structured categories in the 8 type displayed no learning while the unstructured condition significantly outperformed human learners. Fit to RMSE, all three of the tested models demonstrated best performance on the unstructured input; fit to  $r$  there is a clear advantage for the unstructured condition, the only exception to this pattern being higher scores on PARSER for the 1515 condition.

### General Discussion

We began by noting a difference between natural languages and the artificial languages that have previously been used to study the phenomenon of ‘statistical word segmentation.’ Natural languages have complex inter-word dependencies, while previous experiments have purposefully created languages that lacked such dependencies. While this initial simplification was useful, it is uncertain whether dependency structure would be positive or negative for segmentation performance. Our experiments suggest that adults learn better from a language *without* dependencies than one with them. These results suggest that structural variability can be more helpful for the purpose of learning word segmentation than structural regularity, though whether this holds with larger languages remains to be examined.

Our method of using learning curves provided for a higher-resolution examination of the dynamics of both human word segmentation and of models instantiating hypotheses about the mechanisms of segmentation. The measurement of learning across time allows for a richer investigation of the performance for each model, revealing the baseline from which models initialize, their rate of learning, and their final attainment after having been exposed to the full set of stimuli.

In our estimation of whether the mechanisms of ‘statistical learning’ can scale to the task of learning the lexicon of a natural language, we must take into account the difficulties posed by dependency structure. Nevertheless, our studies—both experimental and computational—showed that there was an advantage for languages with an asymmetrical dependency structures. Languages with variability in the number of types assigned to categories facilitated segmentation, perhaps by

providing frequent tokens that served as anchors for segmentation (Kurumada et al., under review). Thus, as suggested by our previous work, it may be the case that high frequency material facilitates language learning by promoting segmentation of adjacent material.

### Acknowledgments

Thanks to the members of the Language and Cognition Lab for valuable discussion.

### References

- Börschinger, B., & Johnson, M. (2011). A particle filter algorithm for bayesian word segmentation. *Proceedings of the Australasian Language Technology Association Workshop*, 10–18.
- Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van Der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non-commercial purposes. In *Proceedings of the fourth international conference on spoken language* (Vol. 3, pp. 1393–1396). Philadelphia, PA.
- Frank, M., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, 117(2), 107–25.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.
- Goldwater, S., Griffiths, T., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112, 21–54.
- Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446.
- Johnson, E., & Tyler, M. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, 13(2), 339–345.
- Kurumada, C., Meylan, S. C., & Frank, M. C. (under review). Zipfian frequency distributions facilitate word segmentation in context.
- Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, 39(246–263).
- Saffran, J. R., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926.
- Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35(4), 606–621.
- Zipf, G. (1965). *Human behavior and the principle of least effort: An introduction to human ecology*. New York, Hafner.

# Tradeoff between Problem-solving and Learning Goals: Two Experiments for Demonstrating Assistance Dilemma

Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)

Hitoshi Terai (terai@is.nagoya-u.ac.jp)

Graduate School of Information Science, Nagoya University

Ryuichi Nakaike (nakaike@educ.kyoto-u.ac.jp)

Graduate School of Education, Kyoto University

## Abstract

Recent intelligent tutoring systems give participants various types of supports. We hypothesize that a high level of support activates participants' orientation to problem-solving goals but reduces the priority of attaining learning goals; as a result, higher problem-solving performance is attained, but the learning effect is reduced. We tested this hypothesis by using two relatively largely different experimental tasks: Tower of Hanoi puzzle as a simple problem solving task and Natural Deduction learning as a more complex learning task. Overall results supported our hypothesis and were discussed from the viewpoint of the assistance dilemma.

**Keywords:** Problem solving goal; Learning goal; Assistance dilemma.

## Introduction

Recently, highly interactive intelligent tutoring systems have been developed whose design principles come from cognitive science theories. A series of cognitive tutors has been constructed based on the ACT-R theory (Anderson, Corbett, Koedinger, & Pelletier, 1995). Intelligent tutoring systems give participants various feedback such as verification, correct response, try again encouragement, error flagging, and elaboration messages (Shute, 2008). In the interaction design between a tutoring system and learners, feedback to learners is a central issue.

In this context, the assistance dilemma has been recognized. Koedinger and Aleven pointed out a crucial question (Koedinger & Aleven, 2007): How should learning environments balance assistance giving and withholding to achieve optimal learning? High assistance sometimes provides successful scaffolding and improves learning, but at other times it elicits superficial responses without consideration from students. On the other hand, low assistance sometimes encourages students to make a large effort, but other times results in enormous errors and interferes with effective learning.

We reformulate the assistance dilemma as a tradeoff of selecting either the problem-solving goal or the learning goal. In a representative situation, participants learn while solving instance problems given by a tutoring system. Attaining the problem-solving goal means solving such instance problems as accurately and rapidly as possible. However, the learning goal requires another attainment that is usually more essential. A primary objective is not to solve instance problems but to learn by solving instances. Dweck classified two types of goals: learning and performance (Dweck, 1986; Ames, 1992). Highly motivated children tend to set learning goals to

increase their competence to understand or master something new rather than just solving problems. Comparing the learning and problem-solving goals in our current study corresponds to Dweck's learning and performance goals. Achieving a problem-solving goal is measured by the solution time and the error ratio for solving problems in the learning phase. The learning goal is usually measured by a posttest after the learning phase.

Another important difference between the two goals is that the problem-solving goal can be achieved with the support of a tutoring system, but the learning goal should be reached without supports of a tutoring system. Achieving the learning goal is usually measured in a setting without tutoring system support because learners should solve problems by themselves without external support from a tutoring system. The need for support means that participants do not complete the learning.

Participant goal setting may be influenced by the feedback information from a tutoring system. One perspective for characterizing the feedback is directive and facilitative (Black & William, 1998). Directive feedback tells participants what needs to be fixed in the next step. Such feedback tends to be more specific than facilitative feedback, which provides participants with comments and suggestions directly relating to the problem-solving. When participants are solving a problem, directive feedback may guide them to focus on the problem-solving goal.

Another perspective for characterizing feedback is its timing. Researchers have addressed whether feedback should be delivered immediately or delayed. Delayed means that it occurs minutes, hours, or even weeks later. Mathan and Koedinger reviewed various studies and concluded that timing effects emerge interactively with other factors such as task difficulty and individual student needs or characteristics (Mathan & Koedinger, 2002). Immediate feedback may facilitate problem-solving goals because participants are repeatedly given indications for determining what to do next when solving a problem.

In the context of the investigation of the assistance dilemma issue, we control the levels of support (LOS) in the following experiments. A high level of support means that more direct and immediate feedback is given. Our hypothesis is that a high level of support activates participants' orientation to problem-solving goals and reduces the priority of attaining

learning goals. This hypothesis predicts that in the high level support condition, the problem-solving performance is higher than in the low level support condition, but the learning effect was reduced; therefore in the posttest where no supports are given, participants who learned in the high level support condition score lower than those in the low level support condition.

We tested this hypothesis using two experimental tasks. In Experiment 1 we used the Tower of Hanoi (TOH) puzzle, which is one representative experimental task widely used in problems-solving studies. In Experiment 2, the participants engaged in a natural deduction (ND) task.

TOH is a simpler task. The problem space is systematically organized and is not so large. Problem solving is achieved by only one operator that corresponds to disk movement. The knowledge and strategies for the solution are represented by less than ten production rules. ND is a more complex task. Its problem space is much larger than that of TOH. To solve problems, since participants must acquire many kinds of inference rules and solution strategies, a complete model for solving ND problems consists of around a hundred production rules. In addition, TOH is basically a problem-solving task, but ND is a learning task. The participants in Experiment 1 joined the experiment in a laboratory setting; those in Experiment 2 engaged in it in a learning context. We confirmed our hypothesis using two relatively largely different experimental tasks.

## Experiment 1

### Task

The six disks TOH puzzle was used as an experimental task.

### Experimental system

The participants individually engaged in the task using an experimental environment established on a personal computer. Figure 1 shows an example screenshot of the experimental system. The participants selected one of the possible disk movements by clicking a button with a mouse. A production system model was mounted on the system to solve TOH by the perceptual strategy. The model infers the next step, the next five steps, and the next nine steps for reaching the goal state through the minimum steps and presents the participants the best next state, the best state after five, and nine steps as a hint.

The LOS was manipulated by the presented hints. In the highest LOS condition, the participants were presented the next step at every problem-solving trial. In other conditions where the best step after five or nine steps was presented, the participants were given such hints at every five or nine problem-solving trials. Higher supports mean direct and immediate feedback; therefore, the participants in these two conditions were given lower levels of support than those in the next step condition. Additionally, in the lowest LOS condition, the participants were given no hint information.

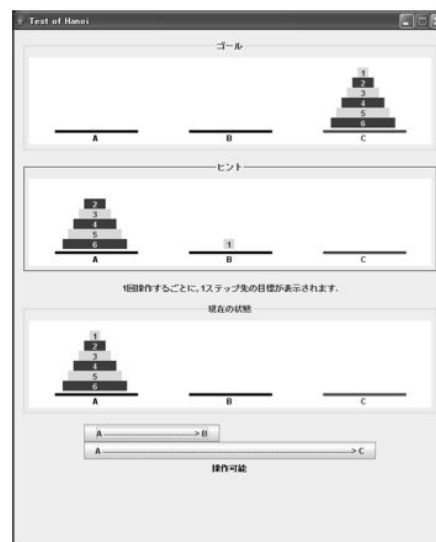


Figure 1: Example screen shot of experimental system for TOH. The upper and lower windows show the goal state and the current state. The middle window presents hint information; in this case the next one step is presented.

### Participants and Procedure

Seventy-one participants joined our experiment. 17, 19, 17, and 18 participants were assigned to the one step, five steps, nine steps, and no hints conditions, respectively. The experiment lasted 90 minutes. The participants were instructed to learn strategies for solving TOH and informed that after the learning session, a posttest would be performed to test their degree of skill acquisition.

In the initial stage of the experiment, the participants learned the constraints of the disk movements and how to use the experimental system. In the learning phase, they solved various types of six-disk-TOH problems in 40 minutes in one of the four experimental conditions. When one problem was completed, the next was given. After the learning phase, a posttest was performed in which the participants solved a test problem by themselves without hint information.

### Result

As a problem-solving performance measure, we used normalized steps for the solution in the learning phase. Figure 2<sup>1</sup> shows the average steps for the solution where the index indicated in the vertical axis was normalized by dividing the solution steps that the participants actually needed to follow by the minimum steps for reaching the goal state from the initial state in each problem. The value, 1.0, means the completed solution, and larger values indicate a poorer solution. The normalized steps needed for the solution were fewer in the one step, five, and nine step conditions where

<sup>1</sup>Note that in figures 2, 4, and 6, the value of the vertical axis is reversed to compare those with Figure 9 in conclusion.

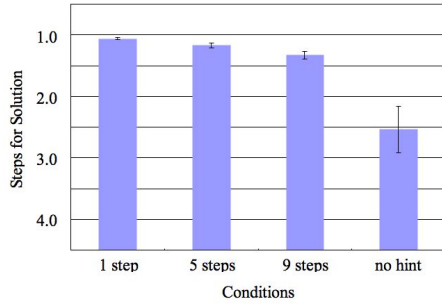


Figure 2: Average steps for solving problems in learning phase (TOH). The value of the vertical axis is reversed to represent higher values as high problem-solving performances, i.e., value 1.0 means the optimal problem solving. An ANOVA indicates that the main effect of Conditions factor was significant ( $F(3, 67) = 12.5, p < 0.01$ ). Fisher's LSD analysis shows significant differences between one step and no hints ( $p < 0.01$ ), five steps and no hints ( $p < 0.01$ ), and nine steps and no hints ( $p < 0.01$ ).

hint information was presented than those in the no hint condition. Our prediction was confirmed because the result of the problem-solving performance was worst in the lowest LOS condition. However, we did not detect statistically significant differences among the three hint conditions. The normalized steps in the three conditions almost reached 1.0. This means that the hint information was sufficient for reducing the trial and error behavior of the participants, even in the nine step condition.

Next, to investigate to what degree each participant thoroughly considered rational actions in each problem-solving step, we calculated the average time to decide each disk movement. We assumed that the priority of the problem-solving goal over the learning goal reduces this consideration time. Figure 3 shows the time that passed between one disk movement and the next. The time in the one step condition was shorter than in the five and nine step conditions, confirming our prediction. However, in the no hit condition, the time was also shorter than the five and nine step conditions, contradicting our prediction.

Next, as a learning performance measure, we used the normalized steps for the solution in the posttest. Figure 4 shows the result. The average steps in the nine steps condition were fewer than those in the one step and no hint conditions. The graph shows that in the three conditions (one, five, and nine step conditions) where hint information was presented, as lower LOSs were given, the learning effect increased, confirming our prediction. However, in the no hint condition, the performance was poorer than that in the higher LOS (nine steps) condition, contradicting our prediction.

Contradictory to our prediction, in the no hint condition, the time for deciding the next disk movement was shorter, and the learning effect was poorer. Perhaps in the learning phase,

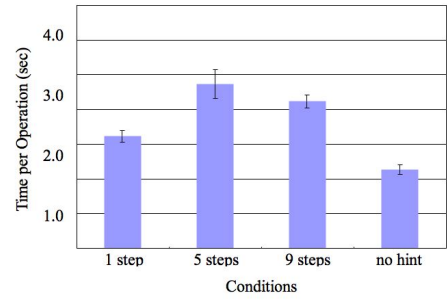


Figure 3: Average time for deciding disk movement (TOH). An ANOVA indicates that the main effect of Conditions factor was significant ( $F(3, 67) = 18.0, p < 0.01$ ). Fisher's LSD analysis shows significant differences between one and five steps ( $p < 0.01$ ), one and nine steps ( $p < 0.01$ ), five steps and no hints ( $p < 0.01$ ), and nine steps and no hints ( $p < 0.01$ ).

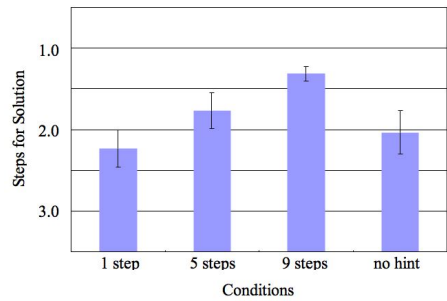


Figure 4: Average steps for solving problems in posttest (TOH). The value of the vertical axis is reversed to represent higher values as high problem-solving performances, i.e., value 1.0 means the optimal problem solving. An ANOVA indicates that the main effect of Conditions was significant ( $F(3, 67) = 3.25, p < 0.05$ ). Fisher's LSD analysis shows significant differences between one and nine steps ( $p < 0.01$ ) and nine steps and no hints ( $p < 0.05$ ).

it was difficult for the participants to learn strategies without hint presentation. This point will be mentioned below in the discussion and conclusion.

## Experiment 2

### Task

Natural deduction (ND) is a kind of proof calculus: e. g., inducing a proposition  $\neg Q \rightarrow \neg P$  from a premise  $P \rightarrow Q$ . Logical reasoning is expressed by inference rules closely related to a natural way of reasoning. They learned nine basic rules and five formal strategies, all of which are fundamental knowledge in ND. Most problems can be solved using this knowledge.

### Experimental system

The experimental system used in Experiment 2 was developed as a tutoring system for teaching ND to university un-

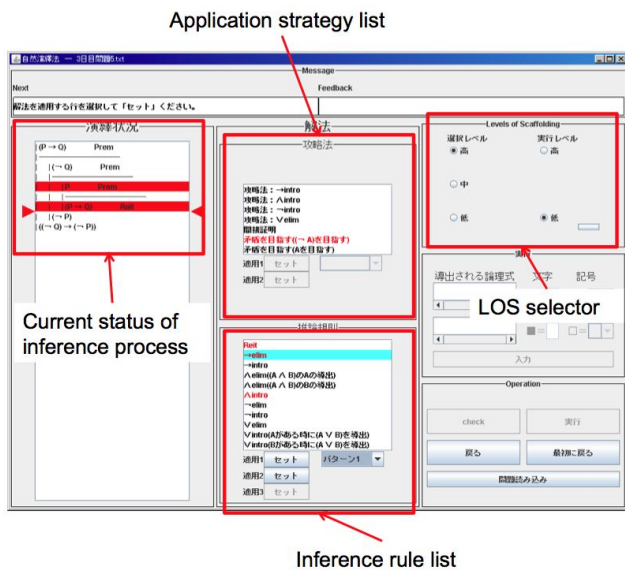


Figure 5: Example screen shot of ND tutoring system. The left side window shows a current status of inference processes where the red items are propositions to which the selected inference rule should be applied. The upper and lower center windows show a strategy list and an inference rule list where the red items are applicable candidates. The right upper window shows the LOS selector where the levels of support are controlled. In the current experiment, LOS was fixed based on the experimental manipulation.

dergraduates. Figure 5 shows an example screenshot of the tutoring system, which provides participants with lists of the inference rules and strategies. After users select one from the lists, the system automatically runs the rules and presents the inference result. The system scaffolds the participants by giving help information about the selection of the rules and strategies and presents the candidates of the rules that should be applied in a given situation.

The LOS was controlled by the presented hint information. In the high LOS condition, the system presented both applicable candidates (rules and strategies) and propositions to which the rules should be applied. In the low LOS condition, the system only presented a set of inference rules; no candidates were presented. The participants had to select an adequate rule from the list by themselves without system support. A high LOS means that the participants were given direct help that guides them to a determined behavior.

### Participants and Procedure

Twenty-nine participants joined our experiment. 13 and 16 were assigned to the high and low LOS conditions, respectively. The experiment was performed over three weeks in an introductory cognitive science class.

In the first week, the participants learned the basics of formal inference systems and ND as an example of the systems.

In the second week, the first half of the learning phase was performed where the participants learned the four basic inference rules. In this class session, all participants learned in the high LOS condition.

In the third week, the latter half of the learning phase followed where LOS was manipulated. The participants solved relatively complex problems for which a sub-derivation process with sub-goal setting was needed. The instructor demonstrated the solutions of two problems, and then the participants solved Problems 1 to 4 with the tutoring system. In the class, the participants were divided into two groups: high LOS and low LOS. After the learning phase, two posttests were performed. Posttest 1 was identical to Problem 2, which they solved in the learning phase, and Posttest 2 as a transfer problem was a new challenge for the participants.

### Result

The optimal steps for a solution are determined in TOH. However, in the solutions of some ND problems, various reasoning paths are rational; therefore we used the average time for solving each problem in the second half of the learning phase as a problem-solving performance measure. Figure 6 shows the result. The solution time was shorter in the high LOS group than in the low LOS group when solving Problems 3 and 4. This result is consistent with our prediction.

Figure 7 shows the time for deciding and implementing an inference rule to forward reasoning. The decision time was shorter in the high LOS group than in the low LOS group when solving Problems 3 and 4. This result is consistent with our prediction.

Next, we used the scores, i.e., the ratio of successfully solved tests, in the posttest as a learning performance measure. Figure 8 shows the result, indicating that when solving Posttest 2, more participants in the low LOS group reached the solution than in the high LOS group, confirming our prediction. This effect was only observed in solving the transfer problems, but not in the repeated problems. This result is consistent with earlier experimental studies, confirming that delayed and lowering supports make positive effects, especially in solving transfer problems (Schroth, 1992, 1997).

### Discussion and Conclusions

The assistance dilemma hypothesizes an optimum point of learning effects as a function of cognitive load. Koedinger et al. (2008) indicated two dimensions of assistance: the practice spacing dimension and the example-problem dimension (Koedinger, Pavlik, McLaren, & Aleven, 2008). They demonstrated a reverse U-shape learning curve on the two dimensions. We conceptualized such a learning effect curve with a problem-solving performance curve (Figure 9). As the support level increases, the problem-solving performance is gradually promoted; however the learning effect reaches maximum at a specific support level and decreases from the point.

When comparing this framework with the results of our two experiments, note that in Experiment 1, the results indi-

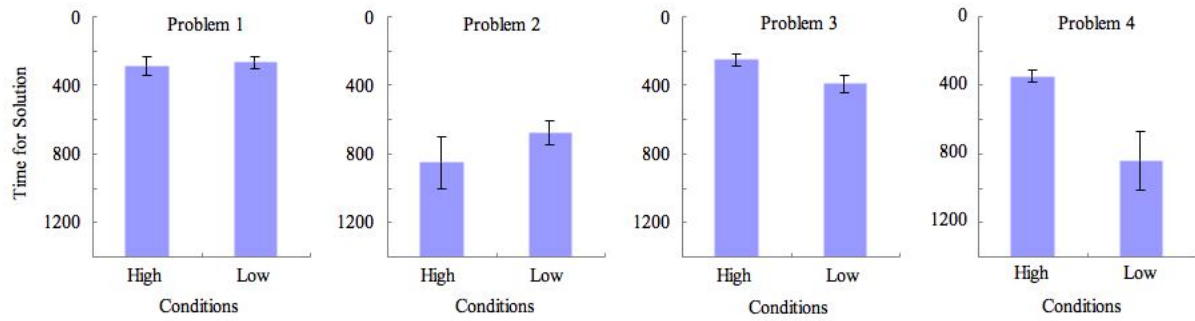


Figure 6: Average solution times for solving problems in learning phase (ND). The value of vertical axis is reversed to represent higher values as high problem-solving performances. T-tests show a marginal significant difference between high and low conditions in Problem 3 ( $t(22) = 1.92, p = 0.07$ ) and a significant difference in Problem 4 ( $t(15) = 2.70, p < 0.05$ ), but no differences in Problems 1 and 2 ( $t(25) < 1$ , n.s.;  $t(24) = 1.14$ , n.s.).

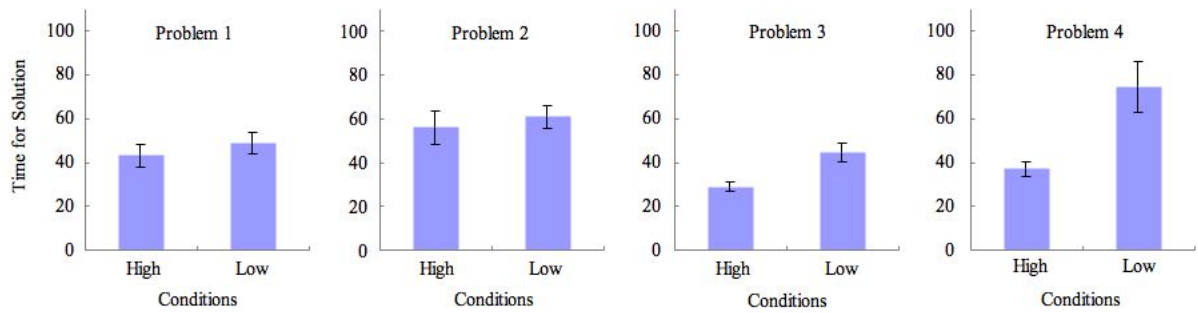


Figure 7: Average time for deciding inference rule (ND). T-tests show a marginal significant difference between high and low conditions in Problems 3 ( $t(22) = 2.55, p < 0.05$ ) and 4 ( $t(25) = 2.98, p < 0.01$ ), but no differences in Problems 1 and 2 ( $t(25) < 1$ , n.s.;  $t(24) < 1$ , n.s.).

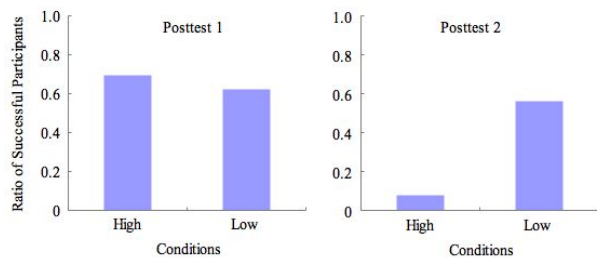


Figure 8: Ratio of successful participants (ND). Chi square tests show a significant difference between high and low conditions in Posttest 2 ( $\chi^2(1) = 7.49, p < 0.01$ ), but no difference in Posttest 1 ( $\chi^2(1) < 1$ , n.s.).

cated in Figures 2 and 4 demonstrated this pattern, controlling the support level from the highest to the lowest on the diagram. But in Experiment 2, the results, as indicated in Figures 6 and 8, only demonstrated the left side of the pattern: control from the highest to the mid level. Our hypothesis was that a high level of support activates participant orientation to

the problem-solving goal and promotes the problem-solving performance, but reduces the priority of attaining the learning goal and decreases the learning effects. This hypothesis is consistent with the left part of Figure 9. On the right side, meaning no or a very low LOS, both the problem-solving performance and the learning effect decreased. This pattern suggests two interpretations. One is that in the right side situation, even if the participants set their goal to learning in the learning phase, they might not be able to decide what to do next and may make enormous errors, resulting in low learning effects. The other interpretation is that the participants give up the attainment of the learning goal because they face difficulties in learning without support. In the current study, we could not decide which explanation is better. Future work will address this issue.

We discussed a tradeoff between problem-solving and learning goals with the degree of support in the learning phase. This tradeoff issue appears in various research fields. For example, the effect of goal specificity has been investigated (Burns & Vollmeyer, 2002; Sweller, 1988). Participants often neglected to consider the theories or rules behind phenomena when they aimed for a specific goal. That is,



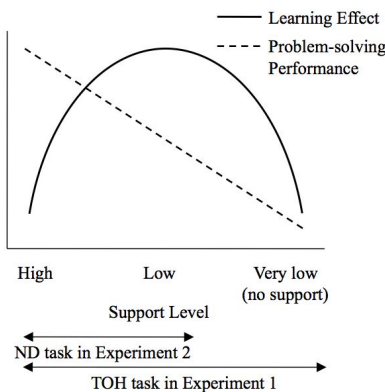


Figure 9: Conceptualized diagram of assistance dilemma and relation of problem-solving and learning goals.

they tended not to search in a hypothesis space, because they concentrated on a search in an instance space to achieve the goal. Consequently, they could not find any rules or underlying mechanisms of the task. Goal specificity studies have indicated that this tendency can be controlled by manipulating goal specificity. Specific goals are direct and immediate; therefore the degree of goal specificity is considered consistent with the levels of support in the current study. The results of our two experiments are consistent with the findings observed in many experiments conducted in the context of goal specificity studies.

Similar to our definition of the levels of support, in studies on automation usage, Sheridan & Verplank indicated ten grades of automation levels (Sheridan & Verplank, 1978). The highest level, Level 10, means that the computer acts entirely autonomously, and the lowest level, Level 1, means that a human does everything. Lin, et al. conducted a micro-world experiment where the participants controlled an atomic power plant under various levels of automation (Lin, Yenn, & Yang, 2010). Medium levels of automation maximized the participants' operations. The most important objective in the use of automation systems is stable manipulation while using the systems; therefore, the tradeoff issue of the two goals might not appear. However, researchers have pointed out the possibility that the continuous usage of automation systems may sometimes cause serious damage when the automation support is removed. For example, Parasuraman et al. argued that stable automation systems produce automation-induced complacency (Parasuraman, Molloy, & Singh, 1993; Molloy & Parasuraman, 1996). They demonstrated in a laboratory setting that operator detection of automation failures was substantially worse for constant-reliability than for variable-reliability (unstable) automation. As we mentioned, in the learning context, the problem-solving performance was measured with learning system support; on the other hand, the learning effect was usually tested without external support in posttests. In this sense, the complacency problem that

emerges when automation breaks down can be understood based on the tradeoff issue of the two goals.

## References

- Ames, C. (1992). Classrooms: Goals, in structure, and student motivation. *Journal of Educational psychology*, 84, 261-271.
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4, 167-207.
- Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5, 7-74.
- Burns, B. D., & Vollmeyer, R. (2002). Goal specificity effects on hypothesis testing in problem solving. *The Quarterly Journal of Experimental Psychology*, 55.
- Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, 41, 1040-1048.
- Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19, 239-264.
- Koedinger, K. R., Pavlik, P., McLaren, B., & Aleven, V. (2008). Is it better to give than to receive? the assistance dilemma as a fundamental unsolved problems in the cognitive science of learning and instruction. In *Proceedings of the 30th annual conference of the cognitive science society* (p. 2155-2160).
- Lin, C. J., Yenn, T., & Yang, C. (2010). Automation design in advanced control rooms of the modernized nuclear power plants. *Safety Science*, 48, 63-71.
- Mathan, S. A., & Koedinger, K. R. (2002). An empirical assessment of comprehension fostering features in an intelligent tutoring system. In *Lecture notes in computer science (proceedings of 6th international conference on intelligent tutoring systems)*, 2363 (p. 330-343).
- Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors*, 38, 311-322.
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology*, 3, 1-23.
- Schroth, M. L. (1992). The effects of delay of feedback on a delayed concept formation transfer task. *Contemporary Educational Psychology*, 17, 78-82.
- Schroth, M. L. (1997). The effect of different training conditions on transfer in concept formation. *Journal of General Psychology*, 124, 157.
- Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. Cambridge, MA: MIT Man-Machine Laboratory.
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78, 153-189.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12, 257-285.

# A Quantum Probability-theoretic account of human judgment using Positive-Operator-Valued Measures

Takayuki Miyadera (miyadera@nucleng.kyoto-u.ac.jp)

Department of Nuclear Engineering, Kyoto University,  
Kyoto 606-8501 JAPAN

Steven Phillips (steve@ni.aist.go.jp)

Mathematical Neuroinformatics Group, National Institute of Advanced Industrial Science and Technology (AIST),  
Tsukuba, Ibaraki 305-8568 JAPAN

## Abstract

People make logically inconsistent probability judgments. The “Linda” problem is a well-known example, which often elicits a conjunction/disjunction fallacy: probability of constituent event  $A$  ( $B$ ) judged more/less likely than their conjunction/disjunction. The Quantum Judgment model (QJM, Busemeyer et al 2011) explains such errors, which are not explainable within classical probability theory. We propose an alternative axiomatic approach in the framework of quantum probability theory that employs *positive* operators representing the set of general queries, in contrast to QJM which uses *projection* operators. Like QJM, our model accounts for conjunction/disjunction fallacies, averaging type errors, and unpacking effects, suggesting that it provides a viable model of judgement error. Further differences between our model and QJM are also discussed.

**Keywords:** Probability judgment; Quantum Probability theory; conjunction/disjunction fallacy; “Linda” problem.

## Introduction

People make probability judgments that are logically inconsistent with classical probability theory. The “Linda” problem is a well-known example: Participants are told that Linda was a philosophy student and an anti-nuclear supporter, and asked to judge her most likely current situation as either (a) feminist supporter, (b) bank teller, (c) feminist and bank teller—conjunction, (d) feminist, but not bank teller, or (e) feminist or bank teller—disjunction. Judging (b) as more likely than (c) is a *conjunction error* (fallacy), since by classical probability  $Prob(A \text{ and } B) \leq Prob(B)$ ; judging (a) as more likely than (e) is a *disjunction error*, since  $Prob(A) \leq Prob(A \text{ or } B)$ . These fallacies are well-known (Tversky & Kahneman, 1983; Bar-Hillel & Neter, 1993), but they are not explained by classical models (Busemeyer, Potheos, Franco, & Trueblood, 2011).

Busemeyer et al. (2011) proposed an alternative model based on *quantum probability theory* (QPT) (Peres, 1993). Their quantum judgment model (QJM) uses the properties of *quantum coherence* and *quantum interference* to explain conjunction and disjunction errors, respectively. An explanation of QJM follows (Appendix A summarizes QPT).

Using the Linda problem as an example, QJM assumes that beliefs about states of the world (e.g., *Linda is a feminist*) are represented as vectors  $\Psi$  in a *Hilbert space*  $\mathcal{H}$  where, e.g., the basis vectors represent feature combinations (e.g., non-/feminist, young/old, gay/straight). An event  $E$  (e.g., corresponding to the proposition *Linda is a bank-teller*) is

a *projection operator*, which projects the belief vector onto a subspace representing a possible outcome: e.g., *Yes* is a possible outcome (subspace) for *Linda is a bank-teller*. A *projection operator*  $E$  applied to  $\Psi$ , written  $E\Psi$  (i.e. vector/matrix multiplication), returns the probability of belief in that outcome, computed as  $Prob(E) = \langle \Psi | E \Psi \rangle$ , where  $\langle \cdot | \cdot \rangle$  is *Dirac notation* for the inner product. For events corresponding to conjunctions of propositions “ $E$  and  $F$ ” (e.g., *Linda is a feminist and Linda is a bank-teller*), the belief in an outcome is computed as: if  $Prob(E) > Prob(F)$ , then  $Prob(E \text{ and } F) = \langle \Psi | E F E \Psi \rangle$ , else if  $Prob(E) < Prob(F)$ , then  $\langle \Psi | F E F \Psi \rangle$ .<sup>1</sup>

In this paper, we provide an alternative formulation of human probability judgment within the framework of quantum probability theory. The most general class of queries is represented by the space of *positive operators*, which includes projection operators. Motivated by this observation, we propose a set of axioms to define a positive operator corresponding to the conjunction of a pair of general propositions. We provide an example that is consistent with this set of axioms. Moreover, we also show how this reformulation accounts for the conjunction and disjunction fallacies, averaging type errors, and unpacking effects.

## Quantum formulation using positive operators

A quantum system is described by a Hilbert space (see Appendix A). In quantum theory, a general query (or event) is represented by an operator  $A$  satisfying  $0 \leq A \leq 1$ . It should be noted that a projection operator also satisfies this condition. Thus an observable which takes a value in a set  $\Omega$  is represented by a positive-operator-valued measure (POVM)  $\{A_a\}_{a \in \Omega}$  on  $\Omega$ . Roughly, a POVM can be regarded as a “fuzzy” version of projection-valued measure (PVM). Thus a POVM is often called an unsharp observable.

**Assumption 1** A person’s belief state is described by a state of a quantum system.

We denote a Hilbert space by  $\mathcal{H}$ .

**Assumption 2** An event that has a family of possible outcomes  $\Omega$  is described by a family of positive operators  $E = \{E_i\}_{i \in \Omega}$  on  $\mathcal{H}$  that satisfies  $\sum_{i \in \Omega} E_i = 1$ . (Such a family of positive operators is called a positive-operator-valued measure (POVM).)

<sup>1</sup> $Prob(E \text{ and } F)$  is undefined when  $Prob(E) = Prob(F)$ .

In the above formulation, an operator corresponding to a proposition “A and B” is not specified. To give quantitative predictions, however, this operator needs to be specified. We take an axiomatic approach to identify a suitable operator. Let us denote the operator corresponding to “A and B” by  $\Lambda(A, B)$ . We assume that for a pair of POVMs  $\{A_a\}$  and  $\{B_b\}$ , an operator corresponding to “ $A_a$  and  $B_b$ ” does not depend on  $A_c$ ’s ( $c \neq a$ ) and  $B_d$ ’s ( $d \neq b$ ). That is,  $\Lambda$  is defined as a map  $\Lambda : E_+(\mathcal{H}) \times E_+(\mathcal{H}) \rightarrow E_+(\mathcal{H})$ , where  $E_+(\mathcal{H}) := \{A \mid \mathbf{0} \leq A \leq \mathbf{1}\}$ . It is natural to suppose that this  $\Lambda$  satisfies the following conditions:

- (o.b)  $\Lambda(A, B)$  satisfies  $\mathbf{0} \leq \Lambda(A, B) \leq \mathbf{1}$  for any  $A, B \in E_+(\mathcal{H})$ .
- (i.b) For any POVMs  $\{A_a\}$  and  $\{B_b\}$ , it holds that  $\sum_{a,b} \Lambda(A_a, B_b) = \mathbf{1}$ . (Thus,  $\{\Lambda(A_a, B_b)\}$  becomes a POVM.)
- (ii.b)  $\Lambda(A, A) = A$  for any projection  $A$ .
- (iii.b)  $\Lambda(A, \mathbf{1}) = A$  for any  $A \in E_+(\mathcal{H})$ .
- (iv.b)  $\Lambda(UAU^*, UBU^*) = U\Lambda(A, B)U^*$  for any  $A, B \in E_+(\mathcal{H})$  and any unitary operator  $U$ .

Some comments are helpful to understand each condition. Condition (o.b) is necessary to guarantee that the framework is closed under conjunction and well-defined. Condition (i.b) means that summation of the probabilities “ $A_a$  and  $B_b$ ” for running  $a$  and  $b$  is 1. Condition (ii.b) represents a trivial requirement. A proposition “Linda is a feminist and Linda is a feminist” is equivalent to “Linda is a feminist”. A restriction of  $A$  in condition (ii.b) may seem strange. However, even in a classical system, confirming that a fuzzy query is true does not guarantee that the same query is true. Therefore we impose a weaker condition than the one above.  $\mathbf{1}$  in condition (iii.b) represents a trivial proposition such as “Linda is Linda”. This condition implies that the proposition “Linda is a feminist and Linda is Linda” is equivalent to “Linda is a feminist”. Condition (iv.b) may need a detailed explanation. It means that an operator corresponding to “A and B” should be determined only by the interrelationship between  $A$  and  $B$ . In quantum theory, the relationship between  $A$  and  $B$  is exactly the same as that between  $UAU^*$  and  $UBU^*$  because unitary operation  $U$  can be interpreted as something like a “coordinate transformation”. Thus  $\Lambda(UAU^*, UBU^*)$  should be written as a function of  $\Lambda(A, B)$  and  $U$ . This function  $f(\Lambda(A, B), U) := \Lambda(UAU^*, UBU^*)$  must satisfy  $f(\Lambda(A, B), UV) = f(f(\Lambda(A, B), V), U)$ . Now we have  $f(\Lambda(U^*AU, \mathbf{1}), U) = \Lambda(UU^*AU^*U, \mathbf{1}) = \Lambda(A, \mathbf{1})$ . Using condition (iii.b), we obtain for any  $A$  and  $U$ ,  $f(U^*AU, U) = A$ . Setting  $U^*AU = B$  for an arbitrary  $B$ , we obtain

$$f(B, U) = UBU^*.$$

Thus it holds that  $\Lambda(UAU^*, UBU^*) = f(\Lambda(A, B), U) = U\Lambda(A, B)U^*$ . Note that these conditions are rather weak. For

instance, we do not require “A and B” to be equivalent with “B and A”.

Before showing the existence of a  $\Lambda$  satisfying these requirements, we show a proposition easily derived from them.

**Proposition 1** Suppose  $\Lambda$  satisfies the requirements (o.b) - (iv.b). It holds that for any projections  $P$  and  $Q$  satisfying  $P + Q \leq \mathbf{1}$ ,  $\Lambda(P, Q) = \mathbf{0}$ , and for any  $A$  with  $\mathbf{0} \leq A \leq \mathbf{1}$ ,  $\Lambda(A, \mathbf{0}) = \mathbf{0}$ .

**Proof:** Let us begin with  $\Lambda(P, Q) = \mathbf{0}$  for projections  $P$  and  $Q$  with  $P + Q \leq \mathbf{1}$ . We can define a PVM  $\{A_0, A_1, A_2\} := \{P, Q, \mathbf{1} - P - Q\}$ . Considering  $\sum_{a,b} \Lambda(A_a, A_b) = \mathbf{1}$ , we obtain

$$\begin{aligned} & P + Q + (\mathbf{1} - P - Q) + \Lambda(P, Q) + \Lambda(Q, P) \\ & + \Lambda(P, \mathbf{1} - P - Q) + \Lambda(\mathbf{1} - P - Q, P) \\ & + \Lambda(Q, \mathbf{1} - P - Q) + \Lambda(\mathbf{1} - P - Q, Q) = \mathbf{1}, \end{aligned}$$

where we used Conditions (i.b) and (ii.b). It concludes  $\Lambda(P, Q) = \mathbf{0}$ .

Consider a POVM  $\{A, \mathbf{1} - A\}$  and  $\{\mathbf{1}, \mathbf{0}\}$ . Condition (i.b) and (iii.b) are used to show

$$\begin{aligned} & \Lambda(A, \mathbf{1}) + \Lambda(A', \mathbf{1}) + \Lambda(A, \mathbf{0}) + \Lambda(A', \mathbf{0}) \\ & = A + A' + \Lambda(A, \mathbf{0}) + \Lambda(A', \mathbf{0}) = \mathbf{1}. \end{aligned}$$

It concludes  $\Lambda(A, \mathbf{0}) = \mathbf{0}$ . ■

To illustrate the existence of  $\Lambda$ , let us consider the following example.

**Example 1** Fix  $0 \leq p \leq 1$ . For any  $A, B$  satisfying  $\mathbf{0} \leq A \leq \mathbf{1}$  and  $\mathbf{0} \leq B \leq \mathbf{1}$ , we define  $\Lambda_p(A, B)$  by

$$\Lambda_p(A, B) = pA^{1/2}BA^{1/2} + (1-p)B^{1/2}AB^{1/2}.$$

Using  $\mathbf{0} \leq B \leq \mathbf{1}$ , one can show  $\mathbf{0} \leq A^{1/2}BA^{1/2} \leq \mathbf{1}$ . Thus  $\mathbf{0} \leq \Lambda_p(A, B) \leq \mathbf{1}$  holds and condition (o.b) is satisfied. Let us examine condition (i.b). Consider a pair of POVM  $\{A_a\}$  and  $\{B_b\}$ . We obtain

$$\begin{aligned} & \sum_{a,b} \Lambda_p(A_a, B_b) \\ & = \sum_{a,b} \left( pA_a^{1/2}B_bA_a^{1/2} + (1-p)B_b^{1/2}A_aB_b^{1/2} \right) \\ & = p \sum_a A_a^{1/2} \sum_b B_bA_a^{1/2} + (1-p) \sum_b B_b^{1/2} \sum_a A_aB_b^{1/2} \\ & = p \sum_a A_a + (1-p) \sum_b B_b = \mathbf{1}. \end{aligned}$$

Condition (ii.b) is satisfied because  $P^{1/2}PP^{1/2} = P$  holds for a projection  $P$ . Condition (iii.b) also follows immediately. In addition, it holds that

$$\begin{aligned} & \Lambda_p(UAU^*, UBU^*) \\ & = pUA^{1/2}U^*UBU^*UA^{1/2}U^* \\ & + (1-p)UB^{1/2}U^*UAU^*UB^{1/2}U^* \\ & = U\Lambda_p(A, B)U^*, \end{aligned}$$

where we used  $U^*U = \mathbf{1}$ . Thus condition (iv.b) is satisfied.

Thus we proved the following theorem.

**Theorem 1** *There exists  $\Lambda$  satisfying Conditions (o.b) - (iv.b). (This  $\Lambda$  is not uniquely determined.)*

### Conjunction and Disjunction Fallacies

The remaining task is to show that there exists a  $\Lambda$  that accounts for the conjunction and disjunction fallacies. We take  $\Lambda_{1/2}$  introduced in Example 1.

Let us consider a model described by a two-dimensional Hilbert space  $\mathcal{H} = \mathbb{C}^2$  which has an orthonormalized basis  $e_0$  and  $e_1$ . A pair of PVMs  $A = \{A_0, A_1\}$  and  $B = \{B_0, B_1\}$  are defined as  $A_n = |e_n\rangle\langle e_n|$  and  $B_n = |f_n\rangle\langle f_n|$  for  $n = 0, 1$ , where  $f_0$  and  $f_1$  are defined by

$$\begin{aligned} f_0 &:= \frac{1}{\sqrt{2}}(e_0 + e_1) \\ f_1 &:= \frac{1}{\sqrt{2}}(e_0 - e_1). \end{aligned}$$

Let us consider a pure state described by a vector

$$\Psi := \sqrt{\frac{9}{10}}e_1 - \sqrt{\frac{1}{10}}e_0.$$

The probability for each proposition is calculated as,  $Prob(A_1) = \frac{9}{10}$ ,  $Prob(A_1 \text{ or } B_0) = 1 - Prob(\Lambda_{1/2}(A_0, B_1)) = \frac{31}{40}$ ,  $Prob(A_1 \text{ and } B_0) = Prob(\Lambda_{1/2}(A_1, B_0)) = \frac{11}{40}$ , and  $Prob(B_0) = \frac{1}{5}$ . They satisfy

$$Prob(A_1) > Prob(A_1 \text{ or } B_0) > Prob(A_1 \text{ and } B_0) > Prob(B_0).$$

Thus this example shows both conjunction and disjunction fallacies. Note that conjunction and disjunction fallacies are supported by other choices of  $\Psi$ . It is an important future work to identify the relevant states.

In addition,  $\Lambda_p$  ( $0 < p < 1$ ) given by Example 1 is consistent with an observation of averaging type errors (Fantino, Kulik, & Stolarz-Fantino, 1997). Consider two general propositions  $A$  and  $B$ . Suppose that a state  $\Psi$  satisfies  $Prob(A) = \langle \Psi | A | \Psi \rangle > \langle \Psi | B | \Psi \rangle = Prob(B)$ . Then  $Prob(A) > Prob(A \text{ and } B)$  must follow. In fact, in our model it holds that

$$\begin{aligned} & Prob(A \text{ and } B) \\ &= \langle \Psi | \Lambda_p(A, B) | \Psi \rangle \\ &= p \langle \Psi | A^{1/2} B A^{1/2} | \Psi \rangle + (1-p) \langle \Psi | B^{1/2} A B^{1/2} | \Psi \rangle \\ &\leq p \langle \Psi | A | \Psi \rangle + (1-p) \langle \Psi | B | \Psi \rangle \\ &< \langle \Psi | A | \Psi \rangle = Prob(A). \end{aligned}$$

where we used  $B \leq \mathbf{1}$  and  $A \leq \mathbf{1}$ .

Unpacking effect, in its broad sense, is interpreted as a difference between  $Prob(A \text{ and } B) + Prob(A \text{ and } B')$  and  $Prob(A)$  (Rottenstreich & Tversky, 1997). That is, the law of classical (Kolmogorov) probability,

$$Prob(A \text{ and } B) + Prob(A \text{ and } B') = Prob(A)$$

is violated. We can show that this effect inevitably occurs between noncommutative sharp propositions no matter how we set  $\Lambda$ .

**Theorem 2** *Let  $P$  and  $Q$  be propositions represented by projection operators. If there is no state violating*

$$\begin{aligned} Prob(P \text{ and } Q) + Prob(P \text{ and } Q') &= Prob(P) \\ Prob(P \text{ and } Q) + Prob(P' \text{ and } Q) &= Prob(Q), \end{aligned}$$

*$P$  and  $Q$  commute with each other.*

**Proof:** If the above equations hold for arbitrary states,  $\Lambda$  satisfies

$$\begin{aligned} \Lambda(P, Q) + \Lambda(P, Q') &= \mathbf{1} \\ \Lambda(P, Q) + \Lambda(P', Q) &= \mathbf{1} \end{aligned}$$

and vice versa. These equations mean that PVMs  $\{P, \mathbf{1} - P\}$  and  $\{Q, \mathbf{1} - Q\}$  are jointly measurable. Hence,  $P$  and  $Q$  commute with each other (Miyadera, 2011). ■

Moreover, for general propositions, we have the following theorem.

**Theorem 3** *Let  $A$  and  $B$  be general propositions. If there is no state violating*

$$\begin{aligned} Prob(A \text{ and } B) + Prob(A \text{ and } B') &= Prob(A) \\ Prob(A \text{ and } B) + Prob(A' \text{ and } B) &= Prob(B), \end{aligned}$$

*their intrinsic ambiguities defined by  $V(A) = \|A - A^2\|$  satisfy*

$$V(A)^{1/2} V(B)^{1/2} \geq \frac{1}{2} \|[A, B]\|,$$

*where an operator norm  $\|\cdot\|$  is defined by  $\|A\| := \sup_{\Psi \neq 0} \frac{\|A\Psi\|}{\|\Psi\|}$ .*

This theorem was proved in Miyadera and Imai (2008).

### Discussion

In this paper, we provided an axiomatic formulation of human probability judgment within the general framework of quantum probability theory. A concrete instantiation was found that satisfies the axioms while accounting for the conjunctive/disjunctive fallacies, averaging type errors, and unpacking effects. We note, though, that QJM accounts for other effects that we have not yet addressed, e.g., *order effect*.

Here, we comment on some differences between our approach and QJM. In contrast to Busemeyer's model, our POVM formalism does not require computing  $Pr(A)$  (nor  $Pr(B)$ ) to obtain  $Pr(A \text{ and } B)$  because  $\Lambda(A, B)$  does not depend on a state. Also, in Busemeyer's formalism an exhaustive set of conjunctions may not sum to 1 (see Appendix B), whereas in our formalism the summation of probabilities is set to 1 (see Axiom (i.b)). Finally, our formulation can be naturally generalized to use mixed states. Further work is needed to explore the implications of these differences.

Proponents of QPT-based approaches divorce themselves from a commitment to the brain as a quantum device (Busemeyer et al., 2011). As a *descriptive* theory of human

judgment, one need not be committed to a quantum mechanical implementation. However, if a *causal* theory is sought—ultimately so for a science of cognition, then a theory based on QPT must be reconciled against the (lack of) evidence showing that the brain is indeed a quantum device (but, see Hameroff, 2002). An alternative to this predicament is to seek yet a further generalization of the quantum framework, which does not depend on quantum mechanics. *General operational probability* theory (Dvurecenskij & Pulmannova, 2000) and *category theory* (MacLane, 2000) are two possibilities for future investigation.

**Acknowledgments:** We would like to thank anonymous reviewers for valuable comments.

## Appendix

### Appendix A: Quantum Probability theory

A *quantum system* is described by a *Hilbert space*  $\mathcal{H}$ , which is a vector space (which we assume to be finite dimensional, i.e.,  $\dim \mathcal{H} < \infty$ ) over the complex field  $\mathbf{C}$  that is equipped with an *inner product*. The inner product  $\langle \cdot | \cdot \rangle$  defines a map  $\mathcal{H} \times \mathcal{H} \rightarrow \mathbf{C}$  satisfying: (i)  $\langle \phi | c_1 \psi_1 + c_2 \psi_2 \rangle = c_1 \langle \phi | \psi_1 \rangle + c_2 \langle \phi | \psi_2 \rangle$  for all  $c_1, c_2 \in \mathbf{C}$  and  $\phi, \psi_1, \psi_2 \in \mathcal{H}$ ; (ii)  $\langle \phi | \psi \rangle = \langle \psi | \phi \rangle^*$  for all  $\phi, \psi \in \mathcal{H}$ ; and (iii)  $\langle \phi | \phi \rangle \geq 0$ , and  $\langle \phi | \phi \rangle = 0 \Leftrightarrow \phi = 0$ . A family of vectors  $\{e_i\}_{i=1}^{\dim \mathcal{H}}$  is called an *orthonormalized basis* of  $\mathcal{H}$ , if it satisfies  $\langle e_i | e_j \rangle = 0$  for  $i \neq j$  and  $\langle e_i | e_i \rangle = 1$  for all  $i$ . A linear map  $A : \mathcal{H} \rightarrow \mathcal{H}$  is called an *operator*. Every operator  $A$  is associated with a unique operator  $A^*$  satisfying  $\langle A^* \psi | \phi \rangle = \langle \psi | A \phi \rangle$  (Riesz theorem);  $A^*$  is called a *conjugate operator* of  $A$ . An operator  $U$  satisfying  $UU^* = U^*U = \mathbf{1}$  is called a *unitary operator*. An operator  $A$  satisfying  $A = A^*$  is called a *self-adjoint operator*. A self-adjoint operator  $P$  satisfying  $P = P^* = P^2$  is called a *projection operator*. A self-adjoint operator  $A$  satisfying  $\langle \psi | A \psi \rangle \geq 0$  for all  $\psi \in \mathcal{H}$  is called a *positive operator*, written as  $A \geq \mathbf{0}$ , where  $\mathbf{0}$  denotes a null operator. Every projection operator is a positive operator. For a positive operator  $A$ ,  $A^{1/2}$  is defined as a unique positive operator satisfying  $A^{1/2} A^{1/2} = A$ .

State and observable are central notions in any physical theory. In quantum theory, a state is represented by a self-adjoint operator  $\rho$ , called a *density operator*, satisfying: (i)  $\rho \geq \mathbf{0}$ ; and (ii)  $\text{tr}(\rho) = \sum_i \langle e_i | \rho e_i \rangle = 1$  for any orthonormal basis— $\text{tr}$  is called *trace*. The set of all states is *convex*: i.e., any combination of two states  $p\rho + (1-p)\sigma$  (for  $0 \leq p \leq 1$  and states  $\rho, \sigma$ ) is also a state. A state  $\rho$  that does not have a nontrivial decomposition is called a *pure state*. A pure state is represented by a projection operator whose *rank* is 1, i.e., there exists a unit vector  $\psi$  ( $\|\psi\| = 1$ ) satisfying  $P\phi = \psi \langle \psi | \phi \rangle$ . ( $P$  is also written  $|\psi\rangle\langle\psi|$ .) This correspondence allows one to identify a unit vector with a pure state. A state that is not pure is called *mixed*.

An observable which takes a value in a set  $\Omega$  (assumed to be discrete set in this paper) is described by a family of positive operators  $\{A_a\}_{a \in \Omega}$  satisfying  $\sum_{a \in \Omega} A_a = \mathbf{1}$ . This is called a *positive-operator-valued measure* (POVM). The

probability of an outcome  $a \in \Omega$  in a state  $\rho$  is given by  $\text{Prob}(A_a) = \text{tr}(\rho A_a)$ . A POVM  $\{A_a\}$  is called a *projection-valued measure* (PVM) if  $A_a$  is a projection operator for each  $a$ . PVMs are often treated as more fundamental objects because each POVM can be represented as a PVM in an enlarged space (Naimark extension theorem). In fact, the space of projection operators can be regarded as a generalization of the Boolean algebra.

### Appendix B: Summation of the probabilities may not agree with one in Busemeyer et al. (2011).

Suppose that  $E$  and  $F$  (and their negations  $E' = \mathbf{1} - E$  and  $F' = \mathbf{1} - F$ ) satisfy for a state  $\psi$ ,  $\langle \psi | F \psi \rangle > \langle \psi | E \psi \rangle > \langle \psi | E' \psi \rangle > \langle \psi | F' \psi \rangle$ . Then we obtain  $\text{Prob}(E \text{ and } F) = \langle \psi | FEF \psi \rangle$ ,  $\text{Prob}(E' \text{ and } F) = \langle \psi | FE'F \psi \rangle$ ,  $\text{Prob}(E \text{ and } F') = \langle \psi | EF'E \psi \rangle$ ,  $\text{Prob}(E' \text{ and } F') = \langle \psi | E'F'E' \psi \rangle$ . Their summation may not agree with 1. In fact, let us consider  $\mathcal{H} = \mathbf{C}^2$  with an orthonormalized bases  $\{e_0, e_1\}$  and projection operators  $E$  and  $F$  defined by  $E = |e_1\rangle\langle e_1|$  and  $F = |f_0\rangle\langle f_0|$ , where  $f_0$  is defined by  $f_0 := \sqrt{\frac{1}{2}}e_1 + \sqrt{\frac{1}{2}}e_0$ . It can be shown that a state  $\psi = \sqrt{\frac{1}{4}}e_1 - \sqrt{\frac{3}{4}}e_0$  satisfies the above inequality, giving

$$\begin{aligned} & \langle \psi | FEF \psi \rangle + \langle \psi | FE'F \psi \rangle + \langle \psi | EF'E \psi \rangle + \langle \psi | E'F'E' \psi \rangle \\ &= 1 + \frac{\sqrt{3}}{4}. \end{aligned}$$

## References

- Bar-Hillel, M., & Neter, E. (1993). Extensional versus intuitive reasoning: The conjunctive fallacy in probability judgment. *Journal of Personality and Social Psychology*, 65, 1119–1131.
- Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, 118(2), 193–218.
- Dvurecenskij, A., & Pulmannova, S. (2000). *New trends in quantum structures*. Kluwer Academic.
- Fantino, E., Kulik, J., & Stolarz-Fantino, S. (1997). The conjunction fallacy: A test of averaging hypotheses. *Psychonomic Bulletin and Review*, 1, 96–101.
- Hameroff, S. R. (2002). Quantum computation in brain microtubules? the Penrose-Hameroff ‘Orch OR’ model of consciousness. *Proceedings of the Royal Society of London A Mathematical, Physical and Engineering Science*, 356, 1869–1896.
- MacLane, S. (2000). *Categories for the working mathematician* (2nd ed.). New York, NY: Springer.
- Miyadera, T. (2011). Uncertainty relations for joint localizability and joint measurability in finite-dimensional systems. *Journal of Mathematical Physics*, 52, 072105.
- Miyadera, T., & Imai, H. (2008). Heisenberg’s uncertainty principle for simultaneous measurement of positive-operator-valued measures. *Physical Review A*, 78, 052119.

- Peres, A. (1993). *Quantum theory, concepts and methods*. Dordrecht: Kluwer.
- Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review*, 104, 406–415.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunctive fallacy in probability judgment. *Psychological Review*, 90, 293–315.

# Comparison of Neural Responses between Exogenous and Endogenous Rule Shifting in Cued Switching Task; an ERPs study

Maki Miyajima ([miyajima.med@gmail.com](mailto:miyajima.med@gmail.com))

Atsuhito Toyomaki

Ichiro Kusumi

Tsukasa Koyama

Department of Psychiatry, Hokkaido University Graduate School of Medicine  
North15, West 7, Kita-ku, Sapporo, Japan

## Abstract

Task switching is a well-known cognitive paradigm to explore task-set reconfiguration processes such as rule shifting. In particular, endogenous task switching is thought to differ qualitatively from stimulus-triggered exogenous task switching. However, no previous study has examined the neural substrate of endogenous task switching. The purpose of the present study is to explore the differences between event-related potential responses to exogenous and endogenous rule switching at cue stimulus. We modified two patterns of cued switching tasks: exogenous (bottom-up) rule switching and endogenous (top-down) rule switching. In each task cue stimulus was configured in order to induce switching or maintaining rule. In Exogenous switching tasks, late positive deflection was larger in the switch rule condition than in the maintain rule condition. However, in endogenous switching tasks late positive deflection was unexpectedly larger in the maintain rule condition than in the switch rule condition. These results indicate that exogenous rule switching is explicit stimulus-driven processes whereas endogenous rule switching is implicitly parallel processes independent of external stimulus.

**Keywords:** Switching task; Event related potentials; rule shifting

## Introduction

Executive function is thought to be one of most important cognitive functions, and it is conceptualized as having four components: volition, planning, purposive action, and effective performance (Lezak, 1995). There are several neuropsychological tests that assess executive function, such as the Wisconsin Card Sorting Test (WCST), the Trail Making Test, and the Tower of London task. In WCST, the subject suppresses a no-longer-relevant task-set and replaces it with an appropriate new task-set; this process is called set shifting (Anderson, Damasio, Jones, & Tranel, 1991; Gold, Carpenter, Randolph, Goldberg, & Weinberger, 1997; Goldberg & Weinberger, 1994; Milner, 1963).

The task switching paradigm is a well-known and sophisticated paradigm that engages these cognitive processes and requires predictable or random alternation between two response selection tasks (Jersild, 1927; Rogers & Monsell, 1995). For example, participants have previously been instructed to switch between a letter classification task

(Task A) and a digit classification task (Task B) in a predictable sequence (e.g., AABB) or in a random sequence. Many studies using the task switching paradigm have observed that reaction time (RT) was reliably greater in switching trials than in maintain trials (no-switching trials) (Rogers & Monsell, 1995). Switch cost is estimated by subtracting the RTs in trials requiring no switch from those in trials requiring a switch; switch cost is thought to be an index of the extra difficulty associated with reconfiguring the active task-set.

Previous Event-related potential (ERP) and fMRI studies have identified the neural substrates of task switching processes. Some previous studies have implicated the lateral prefrontal cortex and the parietal cortex as being centrally involved in preparatory processing during task switching (Johnston, Levin, Koval, & Everling, 2007; Kamigaki, Fukushima, & Miyashita, 2009; Rowe J, 2008; Ruge, 2007). Savine and Braver (Adam C. Savine, 2010) reported that using incentives to modulate cognitive control specifically enhanced task-cue-related activation of the left dorsolateral prefrontal cortex. In an ERP study, Karayanidis et al. (Karayanidis, Coltheart, Michie, & Murphy, 2003) indicated that switch-related positivity and switch-related negativity were elicited by a target stimulus in switching trials. These ERP data reflected that switch-related positivity was unaffected by irrelevant task-cueing, whereas the amplitude and latency of switch-related negativity were modulated by the response-stimulus interval.

Many studies have used task switching, in which stimulus-triggered, bottom-up processes to shift between rules operate via external stimuli such as contextual stimuli, but few studies have focused directly on endogenous, voluntary rule shifting. Rogers and Monsell (Rogers & Monsell, 1995) used an endogenous task switching paradigm that required task-set alternation between trials; this paradigm places high demands on working memory. Participants were told to switch between “Task A” and “Task B” in a predictable sequence (AABB). Target stimuli were presented in one of four boxes continuously displayed on a computer screen, and stimulus position was rotated in a clockwise direction. Thus, on a given trial, the active task was cued by the position of the displayed stimulus. This paradigm seems to require a certain form of endogenous rule switching, because participants change tasks in a predictable sequence.



We believed that cognitive control of endogenous rule shifting is more difficult than control of exogenous rule shifting. For example, perseveration, which is a contextually inappropriate and unintentional repetition of response, is attributed to defective set shifting caused by frontal lobe lesion or cognitive dysfunction, and is often observed in patients with psychiatric disorders. Thus, people who display perseveration find it more difficult to move on from one idea to the next voluntarily than to follow instructions from others. It is likely that the neural substrates characterizing exogenous and endogenous rule shifting are different from one another.

No previous study has examined the neural substrate of endogenous rule shifting using tasks such as those reported by Rogers and Monsell. Their above-mentioned paradigm is simple, but the time point when participants switch rule voluntarily is unclear. Because contextual stimuli (four small squares occupying quadrants of the screen) are displayed continuously in their task and then there is no explicit cue which requires participants to switch rule endogenously. Therefore it seems to be difficult to measure neural responses to endogenous rule switching. It is worthwhile to examine differences in the neural activity elicited by exogenous and endogenous rule shifting in detail. We modified an existing cued switching task and configured the cue stimuli to induce two patterns of rule switching: exogenous (bottom-up) rule switching and endogenous (top-down) rule switching. Furthermore, we compared the neural responses elicited by exogenous rule shifting to those elicited by endogenous rule shifting using ERPs.

## Materials and methods

### Participants

Twelve right-handed student volunteers (12 male, aged 20–34 years, mean age 24.8 years) served as participants. All participants had normal or corrected-to-normal vision. None had a history of neurological or psychiatric disorders. All participants gave written informed consent, and the study was approved by the local research ethics committee.

### Procedure

We made novel cued switching tasks and configured cue stimuli in order to induce two patterns of rule switching: exogenous (bottom-up) rule shifting and endogenous (top-down) rule shifting. In practice trials, the participant learned to perform a two-choice response task accurately according to each of two response rules (see Figure 1). The first was an even vs. odd number classification rule, in which the participant was required to press a button with the left hand in response to an odd number or to press a button with the right hand in response to an even number. The second rule was a small vs. large number classification rule, in which the participant was required to press a button with the left hand in response to a number below five or to press a button with the right hand in response to a number above five. Figure 2 shows a schematic diagram of the cued switching task in

present study. In each trial, a cue stimulus was presented for 1000 ms and was followed by a target stimulus. The target stimulus remained on the screen until the participant pressed one of the two buttons. In the exogenous rule switching task, a white cue stimulus indicated that the subject should maintain the same rule as the previous trial; the white cue stimulus was presented in two or three consecutive trials, and then its color changed to red, indicating that the subject should switch to the other response rule. In the endogenous rule switching task, the cue stimulus was always white, and the participant was instructed to switch the response rule voluntarily every other trial, maintaining each response rule for two consecutive trials. Infrequently, an instruction stimulus showing another rule was presented instead of the cue stimulus; the instruction stimulus dictated that the participant had to switch the response rule immediately. This manipulation requires the participant to attend to the cue stimulus in every trial. If the participant sees the cue stimulus after two consecutive trials during which the same response rule was used, this means that the participant should switch to the other response rule endogenously. We thought that this procedure enables to measure neural response to endogenous rule switching effectively.

rule	response	
	left	right
	odd/even	odd even
	small/large	below 5 above 5

Fig. 1: Response rule in present task

### EEG recording and analysis

An electroencephalogram (EEG) (bandpass 0.16–30 Hz, digitized at 500 Hz) was recorded from 55 electrodes according to the international 10–10 system. Ag/AgCl electrodes were used, and impedance was kept below 10 kΩ. All electrodes were referenced to linked earlobes. An electrooculogram (EOG) was recorded from electrodes lateral to and below the left eye. The signals were digitized for an epoch of 800 ms starting 200 ms prior to the presentation of the cue and target stimulus respectively. During the present study, we recorded large volumes of ERP data, but we conducted a conventional ERP analysis using only a small number of channels and achieved satisfactory findings. We measured the mean amplitude of the ERP’s late positive component (LPC) over 250–300 ms to cue stimulus and LPC over 400–700 ms to target stimulus. A three-way repeated measures ANOVA with Greenhouse-Geisser correction was performed to compare mean LPC amplitudes on the basis of cue modality factor (exogenous vs. endogenous), response rule switching factor (maintain vs. switch), and electrode site factor (Fz, Cz and Pz).

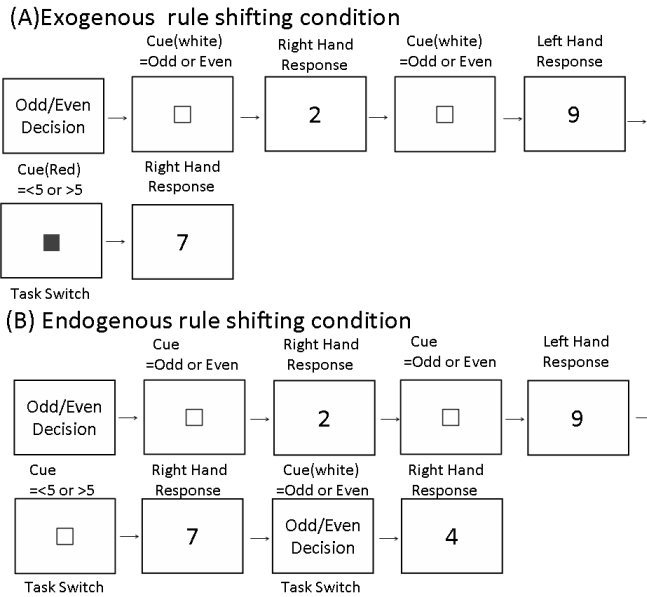


Fig. 2: Schematic diagram of cued switching task

## Result

### Behavioral data

We measured mean reaction times (RTs) and calculated switching costs (RT during trials in which the response rule switches minus RT during trials in which the response rule is maintained) for all conditions (see Figure 3). For RT data, the ANOVA revealed only one significant main effect, which was of switching,  $F(1, 11) = 35.25, p < .001$ . RTs in trials when the response rule was switched were significantly larger than RTs in trials when the response rule was maintained. A t-test on the switching cost results revealed no significant differences between the exogenous and endogenous switching conditions,  $T(11) = -0.76, p = .46$ .

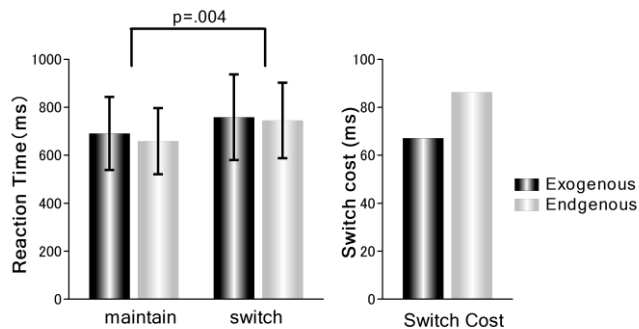


Fig. 3: Behavioral data

### ERPs data

Figure 4 shows the grand-averaged ERP waveforms elicited by the cue stimuli in our study. Figure 5 shows mean LPC amplitude by the cue stimuli at Pz. For the late positive component (LPC) mean amplitude data, the ANOVA revealed only one significant main effect, which was of electrode site factor,  $F(2, 11) = 5.43, p < .05$ . In addition, there was a significant interaction between cue modality factor and switching factor,  $F(1, 11) = 20.31, p < .01$ . LPC deflection in the exogenous task was larger in the switch condition than in the maintain condition, but this pattern was inverted in endogenous task; LPC amplitude in the endogenous task was larger in the maintain condition than in the switch condition.

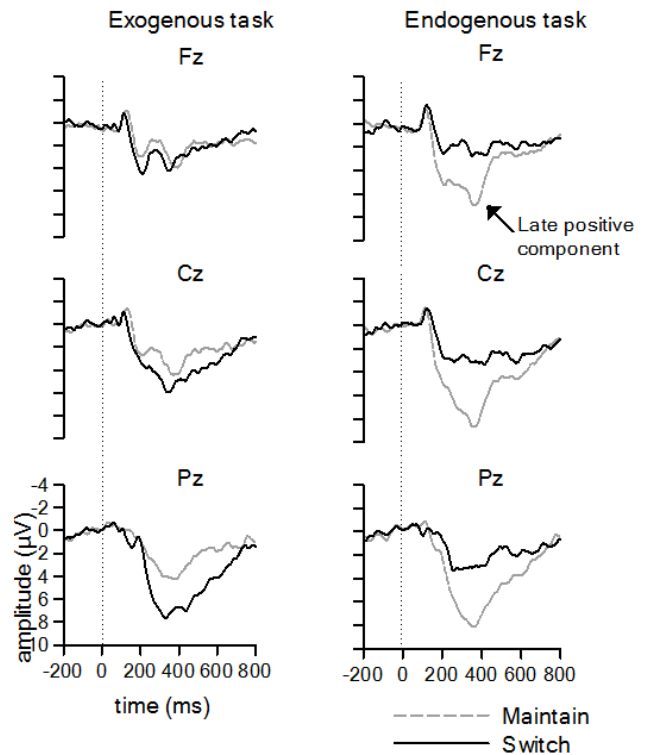


Fig. 4: Grand-averaged ERP waveforms in cue stimuli

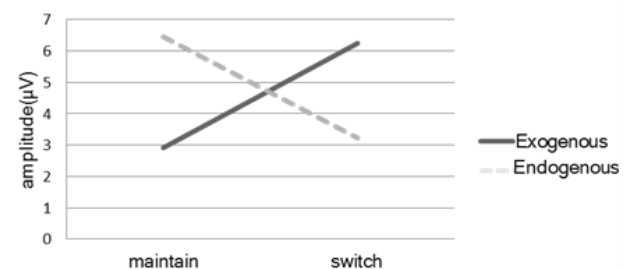


Fig. 5: Mean LPC amplitude in the cue stimuli at Pz

Figure 6 shows the grand-averaged ERP waveforms elicited by the target stimuli in our study. Figure 7 shows mean LPC amplitude by the target stimuli at Pz. For the late positive component (LPC) mean amplitude data, the ANOVA revealed significant main effect in electrode site factor,  $F(2, 11) = 14.90, p < .001$  and response rule switching factor  $F(2, 11) = 11.43, p < .006$ . In addition, there was a significant interaction between cue modality factor and switching factor  $F(1, 11) = 25.82, p < .001$ . In the target stimuli, LPC deflection in the endogenous task was larger in the maintain condition than in the switch condition but there was no significant difference in the exogenous task.

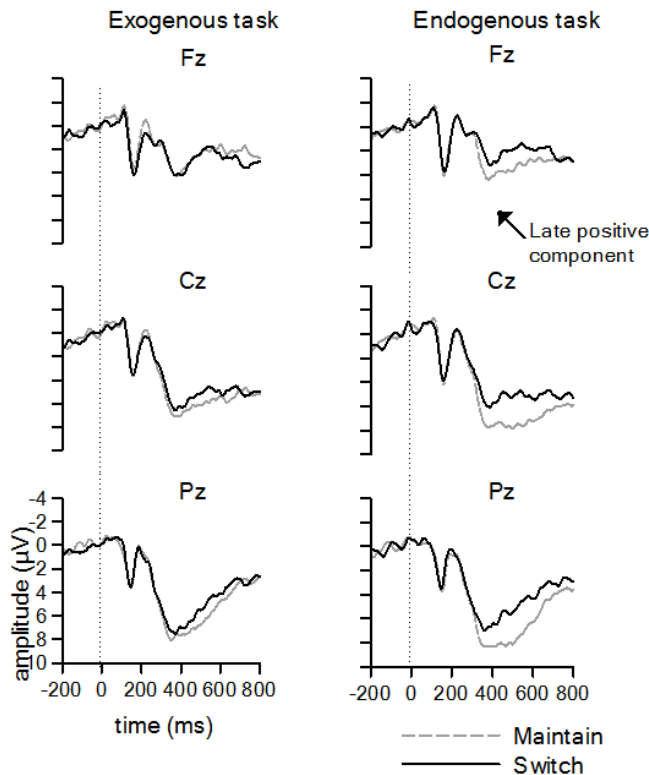


Fig. 6: Grand-averaged ERP waveforms in target stimuli

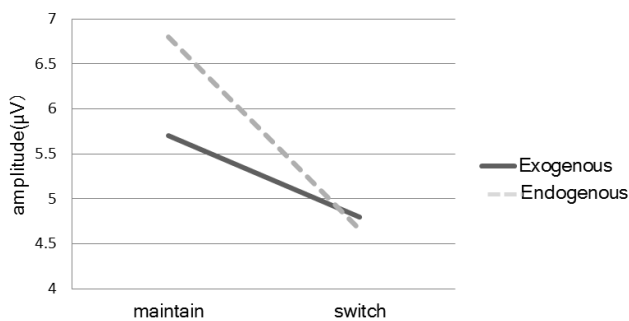


Fig. 7: Mean LPC amplitude in the target stimuli at Pz

## Discussion

The purpose of the present study is to explore the differences between ERP responses to exogenous and endogenous rule switching at cue stimulus onset. This study elucidated interesting ERP waveform changes elicited by the cue stimuli in various conditions. First we discuss the LPC amplitude at the cue stimuli. In the exogenous switching task, late positive component (LPC) amplitude was larger in the switching condition than in the maintaining condition, and thus, the ERP waveform pattern results ran parallel to the reaction time (RT) results. It is possible that changing the cue stimulus drives rule representation retrieval and working memory updates, and thus contributes to increases in positive deflection. Larger ERP deflection elicited by changing the cue stimulus is associated with the lengthening of RT upon presentation of the target stimulus. Surprisingly, the pattern of ERP deflection observed in endogenous task was reversed in the exogenous switching task. LPC amplitude was larger in the maintaining condition than in the switching condition when the switching criteria were endogenous, and this pattern contrasted with RT results.

This disparate pattern of ERP waveforms in the cue stimuli between the endogenous and exogenous tasks from the standpoints of reaction time (RT) measurements and task structure. We hypothesized that changing the cue stimulus in the endogenous switching task triggers voluntary top-down processing for response rule switching (such as memory retrieval and working memory updates) and therefore elicits a larger neural response, reflected by late positive component (LPC) amplitude. Contrary to our hypothesis, the results reflected a reduction of LPC amplitude in the response rule switching condition. Since the response rule was changed every two trials in the endogenous task, preparatory or anticipatory processes for rule switching might operate before the onset of the switching cue. In the present study's endogenous task, we occasionally presented instruction stimuli instead of cue stimuli, forcing participants to switch response rule at an unexpected time. Because of the inclusion of the instruction stimulus, participants could not switch the response rule thoughtlessly after two trials using the same response rule; participants still needed to identify the presentation of the switching cue stimulus carefully. We think that reduction of LPC amplitude in the switching condition indicates that the substance of the endogenous rule shifting process finishes before cue stimulus onset, and the cue stimulus is no more than a confirmation of the already-prepared new response rule. In addition, an increase in LPC amplitude during the ensuing maintaining trial might indicate the operation of proactive partial switching processes that anticipate the next trial. Since the maintaining condition is the last trial during which the participant uses an old response rule, the cue stimulus in maintain trials represents both the preservation of the current rule representation and its subsequent extinction. Because participants likely exercise considerable cognitive efficiency, processes of endogenous rule switching most likely take place at the earliest possible time. Therefore, the cue

stimulus in maintaining trials might cause the allocation of large attentional resources to subsequent rule switching and thus elicit large positive deflections in ERP. The switching cue stimulus presented during the following trial might only confirm an already-updated response rule and thus demand fewer resources, causing a smaller positive deflection. In the exogenous switching task, participants need to evaluate the content of the cue stimulus in every trial, and thus, preparatory or anticipatory rule shifting processes might not operate before the switch trials begin.

Secondly we discuss the LPC amplitude at the target stimuli. In the endogenous switching task, the late positive component (LPC) amplitude was larger in the maintain condition than in the switching condition, and this pattern contrasted with reaction time (RT). However in the exogenous switching task, there was no significant difference between the switching condition and maintaining condition (Fig.7). This result indicated discrepant pattern similar to ERPs of cue stimulus. It is likely that additional cognitive processes operates at target stimulus in maintain condition in endogenous task. In maintain condition, RT was smaller than in switch condition and thus additional process reflected by larger LPC might not be associated with reaction processes to target stimulus. We thought that this process is similar to results of cue triggered ERPs in endogenous task and associated with operation of proactive partial switching processes that anticipate the next trial.

Contrary to ERP results, switch cost, which is the difference in reaction time (RT) between switching and maintaining conditions, showed no significant differences between exogenous and endogenous switching tasks. A previous study conducted by Rogers and Monsell indicated that switch cost varies according to the response-stimulus interval (Rogers & Monsell, 1995), which corresponds to the cue-target interval in the present study. Although variation of switch cost is thought to be associated with processes of task-set reconfiguration (Karayanidis, et al., 2003), this is inconsistent with ERP results in present study. In endogenous task late positive component deflection elicited by switching cue stimulus was reduced but RT was larger than in maintain trial. Some previous studies indicate that switch cost reflects not only task-set reconfiguration before target response but also certain proactive interference (e.g., Karayanidis, et al., 2003). In present study the representation of the new response rule determined by the cue stimulus in both tasks might not be fixed perfectly in the participant's mind at the time of target stimulus onset. In addition participants' lack of experience in applying a new response rule just after applying an old response rule might contribute to this lengthening of RT.

## Conclusion

The ERP results suggest that participants' rule switching strategies vary implicitly according to task structure. To the best of our knowledge, no study has compared bottom-up and top-down task switching processes directly. We used a novel cued switching task to find differences between

neurophysiological responses during exogenous and endogenous rule shifting processes. We can say that exogenous rule switching is explicit stimulus-driven processes whereas endogenous rule switching is implicitly parallel processes independent of external stimulus.

## Acknowledgments

We obtain the knowledge of the difference of neural response between exogenous and endogenous rule shifting. We are grateful to our participants and would like to acknowledge the reviewers of this manuscript.

## References

- Adam C. Savine, T. S. B. (2010). Motivated Cognitive Control: Reward Incentives Modulate Preparatory Neural Activity during Task-Switching. *Neuroscience*, 30(31), 10294-10305.
- Anderson, S., Damasio, H., Jones, R. D., & Tranel, D. (1991). Wisconsin Card Sorting Test performance as a measure of frontal lobe damage. *J Clin Exp Neuropsychol*, 13, 909-922.
- Gold, J., Carpenter, C., Randolph, C., Goldberg, T., & Weinberger, D. (1997). Auditory working memory and Wisconsin cards sorting performance in schizophrenia. *Arch Gen Psychiatry* 54, 159-165.
- Goldberg, T. E., & Weinberger, D. R. (1994). The effects of clozapine on neurocognition: an overview. [Comparative Study Review]. *The Journal of clinical psychiatry*, 55 Suppl B, 88-90.
- Jersild. (1927). *Mental set and shift* (Vol. Whole No.89).
- Johnston, K., Levin, H. M., Koval, M. J., & Everling, S. (2007). Top-down control-signal dynamics in anterior cingulate and prefrontal cortex neurons following task switching. [Research Support, Non-U.S. Gov't]. *Neuron*, 53(3), 453-462. doi: 10.1016/j.neuron.2006.12.023
- Kamigaki, T., Fukushima, T., & Miyashita, Y. (2009). Cognitive set reconfiguration signaled by macaque posterior parietal neurons. [Research Support, Non-U.S. Gov't]. *Neuron*, 61(6), 941-951. doi: 10.1016/j.neuron.2009.01.028
- Karayanidis, F., Coltheart, M., Michie, P. T., & Murphy, K. (2003). Electrophysiological correlates of anticipatory and poststimulus components of task switching. *Psychophysiology*, 40(3), 329-348.
- Lezak, M. D. (1995). *Neuropsychological Assessment*. New York: Oxford University Press.
- Milner, B. (1963). Effects of different brain lesions on card sorting: the role of the frontal lobes. *Archives of Neurology* 9, 90-100.
- Rogers, R., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *J Exp Psychol Gen* 124, 207-231.
- Rowe J, H. L., Eckstein D, Owen AM (2008). Rule-selection and action-selection have a shared neuroanatomical basis in the human prefrontal and parietal cortex. *Cereb Cortex*, 18, 2275-2285.

Ruge, B. (2007). Neural mechanisms of cognitive control in task switching: rules, representations, and preparation. In W. J. Bunge SA, eds (Ed.), *The neuroscience of rule-guided behavior* (pp. 255-283). New York: Oxford UP.

# Children's understanding of hidden emotion, theory of mind, and peer relationship

**Ai Mizokawa (aimizokawa@gmail.com)**

Department of Psychology, Meiji Gakuin University  
1-2-37 Shirokanedai, Minato-ku, Tokyo 108-8636 Japan

**Masuo Koyasu (hgb03675@nifty.com)**

Graduate School of Education, Kyoto University,  
Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501 Japan

## Abstract

This study investigated correlations between understanding of hidden emotion and theory of mind. Five- and six-year-old children ( $N = 105$ , 48 boys and 57 girls) took hidden emotion tasks (TEC component 7), first- and second-order false belief tasks, and a vocabulary test. Teachers rated the children's social interactions in terms of peer relationships. Individual differences in children's understanding of first- and second-order false belief and understanding of hidden negative emotion were associated with differences in language ability. Individual differences in understanding of first-order false belief and understanding of hidden negative emotion were correlated, and this association remained after controlling age and language ability. The results also showed that children who were more advanced in understanding of first-order false belief are more likely to have fewer peer problems. These findings were discussed in terms of social and cognitive development.

**Keywords:** young children; hidden emotion; theory of mind; peer relationship.

## Introduction

Over the past 30 years, researchers have demonstrated that children's ability to understand mental states (desires, thoughts, belief, and emotions) develops throughout their childhood (Astington, Harris, & Olson, 1988; Pons, Harris, & de Rosnay, 2004). One such ability, theory of mind, which is narrowly defined as the understanding of other's mind, such as false belief, dramatically develops between the ages of 3 and 6 (see Wellman, Cross, & Watson, 2001), while there exist some cultural differences. For example, several studies have shown that Japanese children lag significantly behind American and British children on false belief tasks (Hughes, Ensor, Allen, Devine, De Rosnay, Koyasu, Mizokawa, & Lecce, 2011; Lewis, Koyasu, Oh, Ogawa, Short, & Huang, 2009; Naito & Koyama, 2006). Another important element of social competence, understanding of hidden emotion (Saarni, 1979, 1999), also emerges during the preschool period, between the ages of 4 and 6 (Gross & Harris, 1988; Harris, Donnelly, Guz, & Pitt-Watson, 1986). It has been found that there are no differences in age in regards to understanding of hidden emotion between American, British, and Japanese children (Gardner, Harris, Ohmoto & Hamazaki, 1988).

Children's theory of mind ability is traditionally measured by their performance in false belief tasks. The first-order false belief task measures children's ability to

attribute a first-order false belief to a story character (e.g., a mistaken belief about an object's identity or location). Most children pass the first-order false belief task by the ages 6 (Perner, 1991; Wimmer & Perner, 1983). They then pass the second-order false belief task between the ages of 6 and 9 (Perner & Wimmer, 1985). The second-order false belief task measures children's ability to attribute a second-order false belief (i.e. a mistaken belief about a belief) to a story character. Recently, a less demanding second-order false belief task has been created, making it easier to understand for young children (cf. Sullivan, Zaitchik, & Tager-Flusberg, 1994).

To measure children's understanding of emotion, including hidden emotion, Test of Emotion Comprehension (TEC; Pons & Harris, 2000) has been widely used for over the past 10 years. The TEC is an extensive measure of emotion comprehension, which evaluates nine components of emotion understanding. Component 7 assesses whether children understand that one can hide an underlying, or true, emotional state. Pons et al. (2004) found that 50% of 5-year-olds and 65% of 7-year-olds were able to distinguish expressed emotion from actual, felt (hidden) emotion, while only 5% of the 3-year-olds were able to understand the hidden emotion. The hidden emotion tasks include both hidden negative emotion tasks, where the protagonist is motivated to hide inner negative emotion, and hidden positive emotion tasks, where the protagonist is motivated to hide inner positive emotion. To distinguish between expressed apparent emotion and hidden inner emotion, it is necessary to identify the actual emotion and to keep it distinct from the apparent emotion. This ability is also required in false belief tasks. For example, in the unexpected displacement task, which is one of the most popular false belief tasks, participants have to retain the original placement of an object in their mind as well as its present location. In this light, the cognitive ability to understand incongruence between apparent and hidden emotion may relate to the understanding of others' false belief that is not based on reality.

Banerjee and Yuill (1999) addressed 4- to 6-year-old children's understanding of false emotion (hidden negative emotion: hiding inner negative emotion and expressing apparent happiness or neutral emotion) with self-presentational and prosocial motivation, and its association with second-order mental state understanding. They found that only an appreciation of self-presentational false

emotional expression was associated with children's performance on the second-order false belief task. Mizokawa and Koyasu (2007) revealed the association between 4- to 6-year-olds' understanding of pretend crying (hidden positive emotion: hiding inner positive emotion and expressing apparent sadness) and performance of both first- and second-order false belief tasks. These associations between understanding of hidden emotion and development of theory of mind stem from the nature of emotion hiding, which includes the potential that the person who looked at the false emotional expressions would be deceived and adopt a false belief.

As noted above, TEC component 7 assesses children's understanding of the fact that one can hide an inner emotional state. Although this component has been used as a unific index of understanding of hidden emotion, it seems that the tasks in component 7 include some different aspects of mental state understanding. One salient feature is that this component is made up of two hidden positive and two negative emotion stories. Note that there are some findings demonstrating that the expressers' motivation to hide emotion affects children's performance in understanding of hidden emotion (cf. Banerjee & Yuill, 1999; Mizokawa, 2007). However, the protagonist's motivation to hide inner emotion is not clearly specified in each task of TEC component 7.

Previous research suggests that understanding hidden positive emotion (i.e. revealing negative or neutral emotion when one actually has positive emotion) is more difficult for children than understanding hidden negative emotion (i.e. revealing positive or neutral emotion when one actually has negative emotion) (Mizokawa, 2007). That may be because, in some social situations such as receiving unwanted gifts (Saarni, 1979), parents want and instruct their children to mask their negative emotion and express positive emotion to protect another's (i.e. the sender's) feelings. Thus, hiding negative emotion would be internalized through such socialization processes in their early life. Meanwhile, children are rarely "taught" in society to hide their positive emotion as not to hurt others feelings (e.g., it is not appropriate to celebrate too much when one wins a game and another loses the game). Moreover, the expression of positive emotion is likely to be considered as generally "good" and "desirable" in the sense of strengthening social bonds. Children are rarely asked to inhibit or hide positive emotion except in special social situations, such as funerals. So, it may be relatively difficult for children to guess or infer the motivation for hiding positive emotion because they do not have much experience in hiding positive emotion by themselves or being taught to hide positive emotion. From this perspective, children understand hidden negative emotion (i.e. the discrepancy between hidden negative emotion and expressed positive or neutral emotion) somewhat automatically, that is, they can understand that people hide negative emotion without consciously thinking about the effect that others' emotional expressions have on recipients' mental states, what the recipient of the apparent

emotional expression feel, or whether the recipient has a false belief about expresser's actual emotion after witnessing the apparent emotion. They simply need to identify the inner emotion and to keep it distinct from the apparent expressed emotion. After witnessing hidden positive emotion, on the other hand, they might need to think more deeply and make guesses about what is going on in the expresser's and the recipient's mind. We expected that there would be different associations between understanding of hidden emotion and development of theory of mind according to positive-negative valence of emotion. To address these issues, we tested links between young children's understanding of false belief, and hidden positive and negative emotion individually.

In the present study, we also explored the relationship between mental state understanding (first- and second-order false belief and hidden emotion) and peer relations. It has been shown that the development of theory of mind is central to successful social interactions. Some studies have shown that the development of theory of mind is related to important aspects of children's social interactions (cf. Astington & Jenkins, 1995; Dunn & Cutting, 1999; Walker, 2005). As for emotion understanding, it has been revealed that children with greater emotion knowledge demonstrate more empathic and prosocial behaviors and popularity with peers (cf. Cassidy, Parke, Butkovsky, & Braungart, 1992; Denham, 1986; Garner, 1996; Walden & Field, 1990). However, there is little research addressing the link between understanding of hidden emotion and social interaction. We tested the association of individual measures of mental state understanding (first- and second-order false beliefs, hidden positive emotion, and hidden negative emotion) to peer problems. We hypothesized that children's theory of mind and their ability to understand hidden positive emotion would be negatively related to peer problems.

Overall, in the present study, we tested the difference between children's understanding of hidden positive emotion and understanding of hidden negative emotion, the association between hidden emotion and first- and second-order false belief understanding, and association between these mental state understandings and rated peer problems.

## Method

### Participants

One hundred and five Japanese children (48 boys and 57 girls, mean age = 6:1) and twelve teachers (11 women and one man) participated in this study. All were native Japanese speakers.

### Materials and Procedures

Tasks for children were administered individually by the experimenter (the first author) at their school in a quiet room. These children were also examined in a cross-cultural study by Hughes et al. (2011). The children's homeroom teachers rated each child's social interactions in terms of



peer relationships. The order of study tasks for children was counterbalanced across the participants.

**PVT-R** The children's language ability was assessed using PVT-R (Ueno et al., 2008), which requires them to select the picture named by the experimenter from an array of four pictures.

**False belief tasks** The children's understanding of first- and second-order false belief was assessed via four stories (cf. Hughes, Adlam, Happe, Jackson, Taylor, & Caspi, 2000). These included two first-order false belief tasks (Harris, Johnson, Hutton, Andrews, & Cooke, 1989), and two second-order false belief tasks (Sullivan et al., 1994).

In the first-order false belief tasks, participants were shown puppet-based stories that involved a nice surprise story and a nasty surprise story, and were required to answer the protagonist's first-order false belief question, a reality question, and control questions.

Following is an example of a story and questions used in the first-order false belief task (a nice surprise story): *I'm going to tell you a little story about Monty and his lunch box. Look, here's Monty. He wants Freddie to put an apple in his lunch box to take to school. But Freddie says there are no apples left, so he'll have to take a pear instead. Monty doesn't like pears at all. He really wanted an apple! He's so cross about the pear that he stamps all the way upstairs. [Q1 (control 1): How does Monty feel when he gets a pear? Does he feel happy or not happy?] [Q2 (control 2): How does Monty feel when he gets an apple? Does he feel happy or not happy?] But look, while Monty is out of the kitchen, Freddie finds one apple left in the cupboard. He decides to give Monty a nice surprise, and so takes out the pear, and puts an apple in Monty's lunchbox instead. Then he puts the lunchbox in Monty's bag. Monty comes back, picks up the bag and hurries off to school. So Monty doesn't see what's inside his lunchbox. Now it's lunchtime. Monty takes out his lunchbox. [Q3 (first-order false belief): What does Monty think is in the box, an apple or a pear?] [Q4 (reality): What is in the box really, an apple or a pear?]*

In the second-order false belief tasks, participants were shown picture-based stories and asked to answer first- and second-order false belief questions, a reality question, and memory control questions.

Following is an example of a story and questions used in the second-order false belief task: *This is Peter. Today is his birthday, and Peter's Mum is going to surprise him by giving him a puppy. She has hidden the puppy in the shed until it's time for Peter's birthday party. Peter says "I really hope you've got me a puppy for my birthday, Mum." But remember, Mum wants to surprise Peter. So instead of telling Peter she has got him a puppy, Mum says, "Sorry, I didn't get you a puppy, Peter. Actually, I've got you a really good toy for your birthday."* [Q1 (first-order false belief): *So, what did Peter think he was getting for his birthday?*] [Q2 (reality): *What was his Mum giving him really? (If child passes both Q1 and Q2, continue story. If child fails, stop story here.)* Now Peter decides to go outside to play.

*On his way out, he goes into the shed to get his bike, and he finds the birthday puppy! Peter says to himself, "Wow! Mum didn't get me a toy, she really got me a puppy for my birthday!" Mum didn't see Peter go to the shed, so she doesn't know he found the birthday puppy. Inside, the telephone rings. It's Peter's Granny, calling to find out what time the party is. Granny says to Mum, "What does Peter think you've got him for his birthday?" [Q3 (second-order false belief): What does Mum say to Granny?] [Q4 (memory control): Did Mum see Peter go into the shed?] [Q5 (memory control): What has Mum really got Peter for his birthday?]*

For the two first-order false belief tasks, the children's responses were judged as correct when they answered all the four questions correctly (a first-order false belief question, a reality question, and two control questions). Each of the two second-order false belief tasks included first- and second-order false belief questions, a reality question, and memory control questions. When children passed both a first-order false belief and a reality question, they got 1 point for first-order false belief understanding. When they passed a second-order false belief question and memory questions, they got 1 point for second-order false belief understanding. These tasks yielded two scores: the children's understanding of first-order false belief was indexed by summing the scores across the four tasks (scores ranged from 0 to 4), and their understanding of second-order false belief was indexed by summing the two tasks (scores ranged from 0 to 2).

**Hidden emotion tasks** The children's comprehension of hidden emotion was assessed by means of the TEC (Pons & Harris, 2000), component 7 (Hiding). The children were read four picture-based stories and asked to attribute an emotion to a character, who was motivated to hide his or her real emotion from another child and express a different emotion. These four stories were made up of two hidden positive emotion stories (hide inner positive emotion and express sad or neutral emotion) and two hidden negative emotion stories (hide inner negative emotion and express happy emotion). In the hidden positive emotion scenario, the story character gets a new bicycle, but he tries to hide how he feels inside in front of his friend, who does not have his own bicycle (positive 1), and the story character wins a game but she tries to hide how she feels inside in front of her friend who loses the game (positive 2). In the hidden negative emotion stories, the story character falls over in front of his friend, and he tries to hide how he feels inside (negative 1), and the story character is teased by her friend and tries to hide how she feels inside (negative 2). In TEC component 7, the children were asked to attribute an emotion to characters in each of the four stories. They were given 1 point if they attributed the appropriate emotion to each story character (ranging from 0-2).

**SDQ-Peer Problems subscale** Teachers completed a shortened version (in Japanese) of the Strengths and Difficulties Questionnaire (Goodman, 1997) that contained five items. For each item, they rated children's social

interactions using the 3-point Likert scale (not true, somewhat true, or certainly true). The five items were: (1) Rather solitary, tends to play alone; (2) Has at least one good friend; (3) Generally liked by other children; (4) Picked on or bullied by other children; (5) Gets on better with adults than with other children. 'Somewhat true' was scored as 1. 'Not true' was scored as 0 in item 1, 4, and 5, and scored as 2 in item 2 and 3. 'Certainly true' was scored as 2 in item 1, 4, and 5, and scored as 0 in item 2 and 3.

The total score is generated by summing the scores from the five items (scores ranged from 0 to 10). Note that a higher score indicates greater difficulties in the child's peer relationships.

## Results

There were no significant gender differences in any of the study measures. Table 1 shows the descriptive statistics for each study measure.

Table 1: Descriptive statistics for all measures

Measure	<i>M</i>	<i>SD</i>
Age (months)	73.29	3.96
PVT-R	28.87	3.63
HE positive	1.09	0.82
HE negative	1.18	0.83
FB (1st)	2.71	1.36
FB (2nd)	0.57	0.73
Peer problem	1.76	1.87

Note: HE= hidden emotion, FB = false belief.

### Hidden Positive Emotion vs. Hidden Negative Emotion

There was no significant difference between the mean scores of hidden positive emotion tasks and hidden negative emotion tasks (*n.s.*).

### Association between the Study Measures

Correlation and partial correlation coefficients between the different pairs of measures were calculated. Table 2 shows the correlation coefficients and partial correlation coefficients between language ability (PVT-R score), the hidden positive and negative emotion score, the first- and second-order false belief score, and the peer problem score.

Table 2: Correlation coefficients and partial correlation coefficients between study measures

	1	2	3	4	5	6
1 PVT-R	-	.09	.27**	.52**	.36**	-.15
2 HE positive	-	-	.25*	.08	.13	.11
3 HE negative	-	.23*	-	.32**	.11	-.10
4 FB (1st)	-	.05	.21*	-	.59**	-.24*
5 FB (2nd)	-	-.10	.01	.50**	-	-.11
6 Peer Problem	-	.11	-.08	-.19†	-.08	-

Note: The upper off-diagonal elements represent correlation coefficients, while the lower off-diagonal elements are partial correlation coefficients controlling age, and PVT-R score. \*\**p* < .01, \**p* < .05, † *p* < .10. HE = hidden emotion, FB = false belief.

**Language ability and mental state understanding (hidden emotion and false belief)** There were significant correlations between language ability (PVT-R score) and the understanding of hidden negative emotion (*r* = .27, *p* < .01). Children's language ability was clearly related to first- and second-order false belief understanding (first-order: *r* = .52, *p* < .01; second-order: *r* = .36, *p* < .01). There was no significant link between language ability and understanding of hidden positive emotion (*n.s.*).

**Hidden emotion and false belief** There was a significant correlation between understanding of hidden negative emotion and first-order false belief (*r* = .32, *p* < .01). Significant partial correlations were also found between the dyads when age, and PVT-R score were partialled out (*r* = .21, *p* < .05). There was no significant link between understanding of hidden positive emotion and understanding of first-order false belief, and between understanding of both positive and negative hidden emotion and understanding of second-order false belief (all: *n.s.*).

### Relation of Understanding of Hidden Emotion and False Belief to Peer Problems

Teachers rated 101 of 105 children's social interactions. Thus the data of 101 children-teacher pairs were used for the analysis. As shown in Table 2, there was significant negative correlation between the scores for peer problems and performance in first-order false belief tasks (*r* = -.24, *p* < .05). Significant partial correlations were also found between the dyads when age, and PVT-R score were partialled out (*r* = -.19, *p* = .05). That is, the children who were good at understanding other's first-order false beliefs have fewer difficulties in their peer relationships. There was no significant link between the scores for peer problems and any other mental understanding measures (second-order false belief, hidden positive emotion, and hidden negative emotion) (all: *n.s.*).

## Discussion

This study investigated differences between children's understandings of hidden positive and hidden negative emotions and examined associations between their understanding of hidden emotions and theory of mind. We also addressed the link between understandings of mental states and peer relationships. Individual differences in children's understandings of hidden negative emotions and first- and second-order false beliefs were clearly associated with differences in language ability. This result supported previous research demonstrating an important role for language in the development of theory of mind (Happé, 1995).

We found no significant differences between the mean scores for hidden positive emotion tasks and hidden negative emotion tasks, but did find a correlation between children's understandings of hidden positive and hidden negative emotions. The result was not consistent with that of a previous study, which reported that understanding hidden

negative emotions was more difficult than understanding hidden positive emotions (Mizokawa, 2007). This inconsistency may be attributable to differences between the tasks used in the previous study and those used in the current study, especially with respect to those used to address hidden positive emotions. In component 7 of the Test of Emotion Comprehension (TEC), each protagonist in the hidden positive emotion tasks had a motivation to “hide” happiness. On the other hand, because Mizokawa’s study focused on fake crying, each protagonist in the hidden positive emotion tasks had a motivation to “hide” happiness and “express” sadness. Given that the expression of negative emotions is generally construed to be undesirable, the TEC hidden positive emotion tasks used here may have been easier than those used in the previous study.

Although we found no significant difference between children’s scores on hidden positive and hidden negative emotion tasks, the two kinds of hidden emotions (positive and negative) differed in their relationships to aspects of theory of mind. However, the direction of this correlation was opposite to our expectation. We found an association between understanding of others’ minds and hidden negative emotions, whereas no significant link between children’s performances on hidden positive emotion tasks and false belief tasks was observed. One possible explanation of the association between these dyads may be that language ability facilitates children’s understanding of both first-order false beliefs and hidden negative emotions. However, this seems unlikely because these associations remained even after controlling for language ability and age. Although we had expected children to think deeply and make guesses about what was going on in the expresser’s and recipient’s minds when in the presence of someone else’s hidden positive emotion, the results revealed the opposite pattern. It is necessary to conduct further investigations into how children use their ability to understand what is going on in others’ minds in hidden emotion tasks to clarify the meaning of this finding.

In terms of children’s social interactions, the data indicated that children who were more advanced in their understanding of first-order false beliefs were more likely to have and be liked by peers. No association was found between understanding hidden emotions and peer problems. Children who understand first-order false beliefs may have more sophisticated communication skills based on their understanding of others’ minds. Interestingly, language ability *per se* was not associated with children’s peer relationships. These findings suggest that the development of theory of mind transforms and/or is transformed by children’s social interaction skills, as Hughes and Leekam argued (Hughes & Leekam, 2004). Moreover, these results also suggest the possibility that training related to understanding first-order false beliefs leads to fewer peer problems (cf. Ozonoff & Miller, 1995; Slaughter, Dennis, & Pritchard, 2002).

The findings of our study also suggest that understanding hidden emotions, as measured by component 7 of the TEC,

is not a unitary concept. That is, understanding hidden positive and hidden negative emotions should be viewed as related but distinct aspects of emotional understanding. In the hidden positive emotion tasks in component 7 of the TEC, children need to think more deeply and make guesses about another’s mind, which differs from the requirements of the hidden negative emotion tasks. We found no relationship between children’s performance on the hidden emotion tasks and their performance on the second-order false belief tasks or between peer problems and performance on the second-order false belief tasks. This may be due to a floor effect in the second-order false belief tasks.

Our study found a relationship between children’s understandings of first-order false beliefs and their understanding of hidden negative emotions. Although these measures are correlated with each other, only the understanding of first-order false beliefs was linked to peer relationships. When we observe children’s communication with peers, we can see that children who are good at understanding how people control emotional expression have complex relationships (e.g., those involving negotiation) with their peers. It may be expected that aspects of social interactions other than peer problems are linked with understanding hidden emotions. Future research is needed to reveal whether and how children’s understandings of mental states such as hidden emotions come to be reflected in their social interactions.

## Acknowledgments

This research was supported by a grant from MEXT to Masuo Koyasu (grant 70115658). We would like to express our thanks to the children who participated in this study and their parents and teachers.

## References

- Astington, J. W., Harris, P. L., & Olson, D. R. (Eds.) (1988). *Developing Theories of Mind*. New York: Cambridge University Press.
- Astington, J. W., & Jenkins, J. M. (1995). Theory of mind and social understanding. *Cognition and Emotion*, 9, 151-165.
- Banerjee, R. & Yuill, N. (1999). Children’s understanding of self-presentational display rules: Associations with mental-state understanding. *British Journal of Developmental Psychology*, 17, 111-124.
- Cassidy, J., Parke, R. D., Butkovsky, L., & Braungart, J. M. (1992). Family-peer connections: The roles of emotional expressiveness within the family and children’s understanding of emotions. *Child Development*, 63, 603-618.
- Cutting, A. L., & Dunn, J. (1999). Theory of mind, emotion understanding, language, and family background: Individual differences and interrelations. *Child Development*, 70, 853-865.
- Denham, S. A. (1986). Social cognition, prosocial behavior, and emotion in preschoolers: Contextual validation. *Child Development*, 57, 194-201.

- Dunn, J., & Cutting, A. L. (1999). Understanding others and individual differences in friendship interactions in young children. *Social Development*, 8, 201-219.
- Gardner, D., Harris, P. L., Ohmoto, M. & Hamazaki, T. (1988). Japanese children's understanding of the distinction between real and apparent emotion. *International Journal of Behavioral Development*, 11, 203-218.
- Garner, P. W. (1996). The relations of emotional role taking, affective/moral attributions, and emotional display rule knowledge to low-income school-age children's social competence. *Journal of Applied Developmental Psychology*, 17, 19-36.
- Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, 38, 581-586.
- Gross, D. & Harris, P. L. (1988). False belief about emotion: Children's understanding of misleading emotional displays. *International Journal of Behavioral Development*, 11, 475-488.
- Harris, P. L., Donnelly, K., Guz, G. R., & Pitt-Watson, R. (1986). Children's understanding of the distinction between real and apparent emotion. *Child Development*, 57, 895-909.
- Happé, F. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development*, 66, 843-855.
- Harris, P. L., Johnson, C., Hutton, D., Andrews, G., & Cooke, T. (1989). Young children's theory of mind and emotion. *Cognition and Emotion*, 3, 379-400.
- Hughes, C., Adlam, A., Happe, F., Jackson, J., Taylor, A., & Caspi, A. (2000). Good test-retest reliability for standard and advanced false-belief tasks across a wide range of abilities. *Journal of Child Psychology and Psychiatry*, 41, 483-490.
- Hughes, C., Ensor, R. A., Allen, L. L., Devine, R. T., De Rosnay, M., Koyasu, M., Mizokawa, A., & Lecce, S. (2011, April). *Theory of mind performance in British, Australian, Japanese and Italian children: Contrasts in culture or age of school entry?* Paper presented at the Society for Research in Child Development (SRCD) Biennial Conference, Montreal, Canada.
- Hughes, C., & Leekam, S. (2004). What are the links between theory of mind and social relations? Review, reflections and new directions for studies of typical and atypical development. *Social Development*, 13, 590-619.
- Lewis, C., Koyasu, M., Oh, S., Ogawa, A., Short, B., & Huang, Z. (2009). Culture, executive function, and social understanding. *New Directions in Child and Adolescent Development*, 123, 69-85.
- Mizokawa, A. (2007). Young children's understanding of false sadness. *The Japanese Journal of Developmental Psychology*, 18, 174-184. (in Japanese with English abstract).
- Mizokawa, A. & Koyasu, M. (2007). Young children's understanding of another's apparent crying and its relationship to theory of mind. *Psychologia*, 50, 291-307.
- Naito, M., & Koyama, K. (2006). The development of false-belief understanding in Japanese children: Delay and differences? *International Journal of Behavioural Development*, 30, 290-304.
- Ozonoff, S., & Miller, J. N. (1995). Teaching theory of mind: A new approach to social skills training for individuals with autism. *Journal of Autism and Developmental Disorders*, 25, 415-433.
- Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.
- Perner, J., & Wimmer, H. (1985). "John Thinks That Mary Thinks That. . ." Attribution of second-order belief by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, 39, 437-471.
- Pons, F., & Harris, P. L. (2000). *Test of Emotion Comprehension*. Oxford, UK: Oxford University Press.
- Pons, F., Harris, P. L. & de Rosnay, M. (2004). Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization. *European Journal of Developmental Psychology*, 1, 127-152.
- Saarni, C. (1979). Children's understanding of display rules for expressive behavior. *Developmental Psychology*, 15, 424-429.
- Saarni, C. (1999). *Developing emotional competence*. New York: Guilford.
- Slaughter, V., Dennis, M. J. & Pritchard, M. (2002). Theory of mind and peer acceptance in preschool children. *British Journal of Developmental Psychology*, 20, 545-564.
- Sullivan, K., Zaitchik, D. & Tager-Flusberg, H. (1994). Preschoolers can attribute second-order belief. *Developmental Psychology*, 30, 395-402.
- Ueno, K., Nagoshi, N. & Konuki, S. (2008). *PVT-R manual*. Nihon Bunka Kagakusha, Tokyo. (in Japanese)
- Walden, T., & Field, T. (1990). Preschool children's social competence and the production and discrimination of affective expressions. *British Journal of Developmental Psychology*, 8, 65-76.
- Walker, S (2005). Gender differences in the relationship between young children's peer-related social competence and individual differences in theory of mind. *The Journal of Genetic Psychology*, 166, 297-312.
- Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72, 655-684.
- Wimmer, H. & Perner, J. (1983). Belief about belief: Representation and constraining function of wrong belief in young children's understanding of deception. *Cognition*, 13, 103-128.

# The Effect of Visually and Phonologically Misleading Nonwords on Lexical Decisions of Native Japanese Readers

Rika Mizuno (mizunor@isc.chubu.ac.jp)

Takao Matsui (mat@psy.chubu.ac.jp)

Department of Psychology, Chubu University  
1200 Matsumoto-Cho, Kasugai-shi, 487-8501 JAPAN

## Abstract

Native Japanese readers were found to rely heavily on visual codes and scarcely on phonological codes in letter/word processing (Mizuno, Matsui, & Bellezza, 2007). This study aimed to determine if this processing feature of native Japanese readers influenced their process of lexical access by lexical decision tasks using visually misleading transposed-letter (TL) nonwords, phonologically misleading pseudohomophones, and standard nonwords. Lupker and Pexman (2010) found that the performance on a lexical decision task of native English readers was impaired by both TL nonwords and pseudohomophones. However, the results of two experiments in this study showed that the performance of native Japanese readers was impaired not by pseudohomophones but by TL nonwords. The results suggested that the processing features of native readers of various languages should influence their process of lexical access.

**Keywords:** lexical decision; nonwords; transposed-letter; pseudohomophone; native Japanese readers; native English readers

## Introduction

A lexical decision task (Meyer & Schvaneveldt, 1971) is a task in which participants decide whether presented letter strings are words or not. The lexical decision time of a word is considered to reflect the access time to lexical representation of the word, and the task has been used in many studies to explore the structure of lexical representation or the process of lexical access.

Various features of words have been found to influence lexical decision time. Such features include word frequency (e.g., Glanzer & Ehrenreich, 1979), neighborhood size<sup>1</sup> (Coltheart, Develaar, Jonasson, & Besner, 1977), semantic relation to primes (Neely, 1977), spelling-to-sound regularity (Parkin, 1982), and so on.

However, some research has shown that the effects of the features of nonwords on lexical decision time are also not negligible. For example, Shulman and Davidson (1977) found that pronounceable nonwords delayed the lexical decision time of words more than unpronounceable nonwords. Perea and Lupker (2004) found that the transposed-letter (TL) nonwords, which were made by transposing the two letters of words, delayed the lexical decision time for both words and nonwords. Lupker and Pexman (2011) compared the size of frequency effect in the TL, pseudohomophone, and standard nonword conditions. Frequency effect means that the lexical decision time of

more frequent words is shorter. They found that lexical decision time was longer and that frequency effect was greater in the TL and the pseudohomophone nonword conditions than in the standard nonword condition.

These findings about the effects of nonwords on lexical decision time not only contributed to the improved understanding of the process of lexical access but suggested that the features of nonwords should also be considered in using lexical decision tasks.

As described above, visually misleading TL nonwords and phonologically misleading pseudohomophones were confirmed to influence lexical decision time of words and nonwords. However, they were confirmed only with native English readers. Mizuno, Matsui, and Bellezza (2007) and Mizuno, Matsui, Harman, and Bellezza (2008) conducted several letter-matching experiments with native English and native Japanese readers, and found that native Japanese readers rely heavily on visual codes and not as much on phonological codes as native English readers do. Mizuno and Matsui (2012) also showed that visual similarity, rather than phonological similarity, between targets and distracters increased attentional blink of native Japanese readers, while Chun and Potter (1995) suggested that phonological similarity had a significant effect on the attentional blink of native English readers.

Consequently, we hypothesized that performance on a lexical decision task by native Japanese readers would not be impaired by phonologically misleading pseudohomophones because they rely scarcely on phonological codes. If this is verified, we will be able to not only indicate that their processing features of letters influence performance on lexical decision tasks but also alert many researchers using lexical decision tasks to take the processing features of their participants into account in choosing nonwords.

## Experiment 1

Experiment 1 compared lexical decision time and error rates of native Japanese readers across the TL nonword condition, the pseudohomophone condition, and the standard nonword condition. We predicted that their lexical decision time would be delayed and error rates for nonwords would be high only in the TL condition, and not in the pseudohomophone condition and the standard condition.

## Method

**Participants and Design** Thirty-six undergraduate students (14 women and 22 men) who were native Japanese

<sup>1</sup> The number of words that can be created by changing one letter while maintaining letter positions.

readers participated in return for course credit. Participants were assigned to all three conditions: the TL nonword, the pseudohomophone, and the standard nonword conditions.

**Equipment** The experiment was conducted on a personal computer (Fujitsu, FMV Esprimo D5350) running an experimental software (Cedrus Co., SuperLab2.0) with a 17-in. liquid crystal monitor (EIZO, FlexScan S1731). Responses were collected by a response box (Cedrus Co., RB-730). A chin support (Takei, T.K.K. 123i with 123j) was placed on the edge of the desk. The distance between participants' eyes and the screen was about 45 cm, and the height of the chin support was adjusted for each participant.

**Stimuli** All the stimuli were two-character and four-mora *Kanji* words. In total, 120 words and 120 nonwords (40 TL nonwords, 40 pseudohomophones, and 40 standard nonwords) were selected/created in the following manner: 240 words of frequencies between 15,000 and 100,000 were selected from the database (Amano & Kondo, 2003); 40 TL nonwords were made by transposing two *Kanji* characters (e.g., "盟連", from "連盟"), confirming that they had no homophones; and the remaining 200 words were divided into five sets of 40 words each. Three sets were assigned to word sets, another set was used for making pseudohomophones, and another was used for making standard nonwords. Forty pseudohomophones were made by replacing each *Kanji* character with another *Kanji* character with the same phone (e.g., "案低", from "安定"). The 40 standard nonwords were created by exchanging one of the two *Kanji* characters with one of the other words (e.g., "開税"), confirming that they were nonwords and had no homophones. The three nonword sets were combined with the three word sets to form six counterbalancing groups.

Japanese letters are typically written from left to right or from top to bottom but sometimes from right to left. The two letters, therefore, were written vertically from top to bottom lest the TL nonwords should be regarded as words. A two-letter stimulus presented on the monitor subtended 3 degrees of visual angles vertically and 1.5 degrees horizontally.

**Procedure** Six participants were allocated to each of the six counterbalancing groups. In each group, the orders of three nonword conditions were counterbalanced among the six participants.

Participants were tested individually. Each participant completed eight practice trials with standard nonwords followed by three blocks of 80 experimental trials. The order of the 80 trials was randomized. Participants were instructed to decide as quickly and accurately as possible whether the letter strings were a word or a nonword by pressing the right-most key if they were a word and the left-most key if they were not. In each trial, after a 1,100 ms interval, two asterisks written vertically were presented on the middle of the screen for 550 ms followed by the stimuli, which remained on the screen for three seconds or until the

participant responded.

Results

Trials involving latencies greater than 1,500 ms (1.2% of the word trials and 3.2% of the nonword trials) were removed from the following analyses according to Lupker and Pexman (2010).

**Word Lexical Decision Time** Means of correct lexical decision time for words in the three nonword conditions are shown in Figure 1. One-way repeated measures analysis of variance (ANOVA) revealed that the effect of the nonword condition was significant,  $F(2, 70) = 27.29$ ,  $MSE = 5,528.08$ ,  $p < .001$ . Multiple comparisons showed that the mean lexical decision time in the TL condition was significantly longer than that in the pseudohomophone condition and the standard nonword condition,  $ps < .01$ ,  $HSD = 53.04$ .

**Word Error Rates** Means and standard deviations of error rates for words in the three nonword conditions are shown in Table 1. ANOVA of arcsine transformed error rates revealed that the effect of the nonword condition was not significant,  $F(2, 70) = 1.62$ ,  $MSE = 37.97$ ,  $p = .20$ .

**Nonword Lexical Decision Time** Means of correct lexical decision time for nonwords in the three nonword conditions

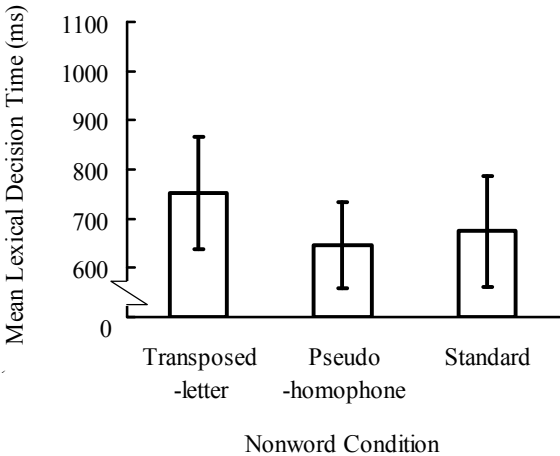


Figure 1: Mean and standard deviation of lexical decision time for words in each nonword condition in Experiment 1.

Table 1: Mean and standard deviation of error rates for words in each nonword condition in Experiment 1.

	Nonword Condition		
	Transposed -letter	Pseudo -homophone	Standard
Mean	0.061	0.048	0.060
SD	0.043	0.039	0.037

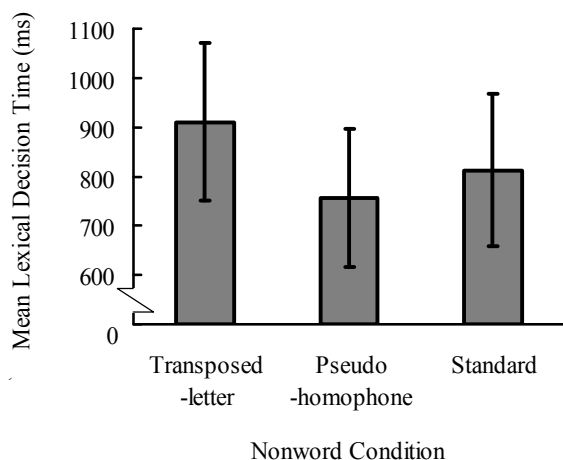


Figure 2: Mean and standard deviation of lexical decision time for nonwords in each nonword condition in Experiment 1.

Table 2: Mean and standard deviation of error rates for nonwords in each nonword condition in Experiment 1.

	Nonword Condition		
	Transposed -letter	Pseudo -homophones	Standard
Mean	0.097	0.042	0.060
SD	0.044	0.034	0.032

are shown in Figure 2. ANOVA revealed that the effect of the nonword condition was significant,  $F(2, 70) = 29.86$ ,  $MSE = 13,857.37$ ,  $p < .001$ . Multiple comparisons showed that the mean lexical decision time in the TL condition was significantly longer than that in the pseudohomophone condition and the standard nonword condition,  $ps < .01$ ,  $HSD = 83.97$ , and that the time in the pseudohomophone condition was shorter than that in the standard nonword condition,  $p < .05$ ,  $HSD = 66.71$ .

**Nonword Error Rates** Means and standard deviations of error rates for nonwords in the three nonword conditions are shown in Table 2. ANOVA of arcsine transformed error rates showed that the effect of the nonword condition was significant,  $F(2, 70) = 19.98$ ,  $MSE = 26.92$ ,  $p < .001$ . Multiple comparisons revealed that the error rate in the TL nonword condition was higher than that in the pseudohomophone condition and the standard nonword condition,  $ps < .01$ ,  $HSD = 3.70$ , and that in the pseudohomophone condition was lower than that in the standard nonword condition,  $p < .05$ ,  $HSD = 2.94$ .

## Discussion

The lexical decision time for words in the TL condition was longer than the lexical decision times for words in the pseudohomophone condition and in the standard condition.

This result was different from that of Lupker and Pexman (2010) with native English readers, which showed that lexical decision times for words in the TL condition and pseudohomophone condition were longer than the corresponding time in the standard nonword condition. This result suggests that lexical decision time for words of native Japanese readers, who do not rely so much on phonological codes, is not delayed by phonologically misleading nonwords. Error rates for words did not differ among nonword conditions, consistent with the result of Lupker and Pexman (2010).

As for nonwords, the results showed that lexical decision time in the TL condition was longer than in the other two conditions, and that error rates in the TL condition were higher than those in the other two conditions. However, Lupker and Pexman (2010) with native English readers showed that their lexical decision times and error rates in the TL and pseudohomophone conditions were greater than those in the standard nonword condition. These results indicated that the effect of phonologically misleading nonwords was scarce in the case of native Japanese readers.

Nonetheless, we did not expect that the lexical decision time in the pseudohomophone condition would be shorter than that in the standard nonword condition, or that the error rate in the pseudohomophone condition would be lower than that in the standard nonword condition. These results were inconsistent with the previously mentioned results indicating that phonological codes have a scarce effect on lexical decisions of native Japanese readers, and we considered it implausible to suppose that the phonological codes of pseudohomophones made lexical decision of nonwords easy.

Therefore, we reexamined the frequencies and stroke counts of all *Kanji* characters consisting of the nonwords in the three conditions. The stroke counts of *Kanji* characters (see Table 3) reflect their visual complexities. The means of frequencies and stroke counts (with standard deviations in parentheses) were, respectively, 456,281.6 (566,312.9) and 9.775 (4.40) for TL nonwords, 305,685.2 (392,268.0) and 8.91 (3.24) for pseudohomophones, and 497,443.0 (515,634.1) and 9.56 (3.73) for standard nonwords. The mean frequency and the mean stroke count of pseudohomophones were smaller than the others. A low frequency was likely to increase both lexical decision time and error rate. It could not unexpectedly decrease lexical decision time, or the error rate in the pseudohomophone condition. Therefore, we concluded that the low stroke counts were the real cause. Because native Japanese speakers tend to rely heavily on visual codes, it was

Table 3: Examples of stroke counts.

Stroke Count		
3	8	13
大	性	新



extremely plausible that the small mean of stroke counts of pseudohomophones made lexical decision time shorter and error rates lower than those of standard nonwords.

In Experiment 2, therefore, some of the *Kanji* characters consisting of pseudohomophones and standard nonwords were substituted so as to make the means of stroke counts and those of frequencies as even as possible in the three nonword conditions.

## Experiment 2

### Method

**Participants and Design** Thirty-six undergraduate students (13 women and 23 men) who were native Japanese readers participated in return for course credit. Participants were assigned to all three conditions: the TL nonword, the pseudohomophone, and the standard nonword conditions.

**Stimuli** The 120 words and 40 nonwords in the TL conditions were the same as those used in Experiment 1. Some of the *Kanji* characters in the pseudohomophone condition and the standard nonword condition used in Experiment 1 were substituted with other *Kanji* characters to make the means of stroke counts, along with the frequency of *Kanji* characters composing the nonwords, in the three conditions as equal as possible. The resultant means (with standard deviations in parentheses) of stroke counts and frequency of *Kanji* characters composing the nonwords in the TL condition were 9.78 (4.40) and 456,281.6 (566,312.9), respectively, those in the pseudohomophone condition were 9.76 (3.52) and 412,652.3 (369,488.1), and those in the standard nonword condition were 9.78 (3.72) and 422,790.2 (454,322.7).

**Equipment and Procedure** These were the same as in Experiment 1.

### Results

Trials involving latencies greater than 1,500 ms (2.9% of the word trials and 6.7% of the nonword trials) were removed from the following analyses as in Experiment 1.

**Word Lexical Decision Time** Means of correct lexical decision time for words in the three nonword conditions are shown in Figure 3. One-way repeated measures ANOVA revealed that the effect of nonword condition was significant,  $F(2, 70) = 12.47$ ,  $MSE = 5,572.14$ ,  $p < .001$ . Multiple comparisons showed that the mean lexical decision time in the TL condition was significantly longer than that in the pseudohomophone condition and the standard nonword condition,  $ps < .01$ ,  $HSD = 53.25$ .

**Word Error Rates** Means and standard deviations of error rates for words in the three nonword conditions are shown in Table 4. ANOVA of arcsine transformed error rates revealed that the effect of the nonword condition was not significant,  $F(2, 70) = 0.32$ ,  $MSE = 48.80$ ,  $p = .73$ .

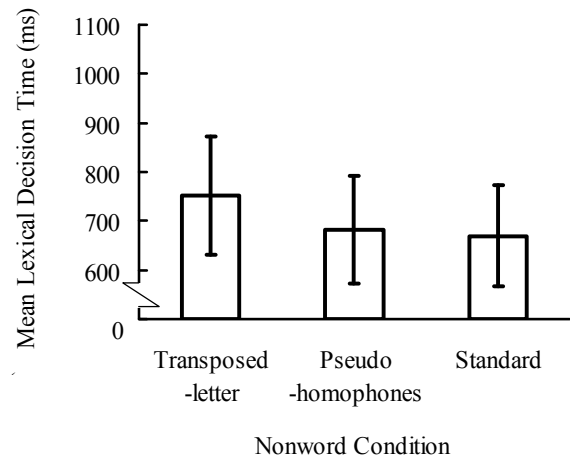


Figure 3: Mean and standard deviation of lexical decision time for words in each nonword condition in Experiment 2.

Table 4: Mean and standard deviation of error rates for words in each nonword condition in Experiment 2.

	Nonword Condition		
	Transposed -letter	Pseudo -homophone	Standard
Mean	0.051	0.059	0.048
SD	0.044	0.049	0.038

**Nonword Lexical Decision Time** Means of correct lexical decision time for nonwords in the three nonword conditions are shown in Figure 4. ANOVA revealed that the effect of the nonword condition was significant,  $F(2, 70) = 16.55$ ,  $MSE = 9,425.45$ ,  $p < .001$ . Multiple comparisons showed that the mean lexical decision time in the TL condition was significantly longer than that in the pseudohomophone condition and the standard nonword condition,  $ps < .01$ ,  $HSD = 69.25$ .

**Nonword Error Rates** Means and standard deviations of error rates for nonwords in the three nonword conditions are shown in Table 5. ANOVA of arcsine transformed error rates showed that the effect of nonword condition was significant,  $F(2, 70) = 5.36$ ,  $MSE = 32.03$ ,  $p = .007$ . Multiple comparisons revealed that the error rate in the TL nonword condition was higher than that in the pseudohomophone condition and the standard nonword condition,  $ps < .05$ ,  $HSD = 3.21$ , and no difference was found between the latter two conditions.

### Discussion

As expected, the lexical decision time for words in the TL nonword condition was longer than that in the pseudohomophone condition and the standard condition, and there was no difference between the latter two conditions.

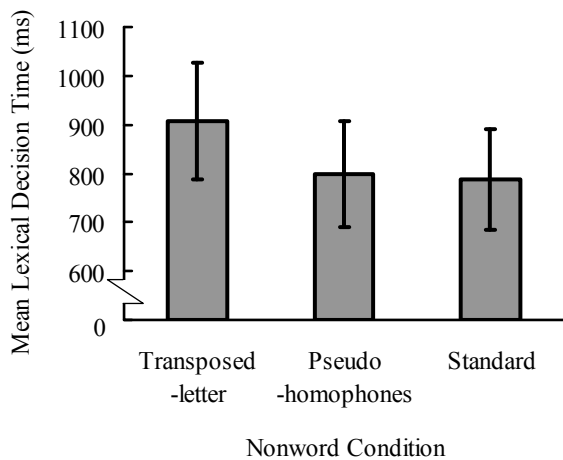


Figure 4: Mean and standard deviation of lexical decision time for nonwords in each nonword condition in Experiment 2.

Table 5: Mean and standard deviation of error rates for nonwords in each nonword condition in Experiment 2.

	Nonword Condition		
	Transposed -letter	Pseudo -homophones	Standard
Mean	0.101	0.064	0.068
SD	0.055	0.036	0.050

This result suggests that phonologically misleading nonwords have no effect on the lexical decisions of native Japanese readers. The error rates for words did not differ between the three nonword conditions, which is with the results of Lupker & Pexman (2010) with native English readers.

The lexical decision time for nonwords was the longest in the TL nonword condition, and there was no difference between that in the pseudohomophone condition and that in the standard nonword condition. This result was also expected and indicates that native Japanese readers rely scarcely on phonological codes. The error rate for words in the TL condition was higher than that in the standard nonword condition, and no other difference was found. Lupker and Pexman (2010) found that error rates in both the TL condition and the pseudohomophone condition were higher than the error rate in the standard nonword condition with native English readers. This result, therefore, suggests that native Japanese readers make fewer errors when deciding if phonologically misleading nonwords are nonwords because they scarcely rely on phonological codes.

## General Discussion

In Experiment 1, the lexical decision times for words and for nonwords were shorter and the error rate for nonwords was lower in the pseudohomophone condition than those in

the TL nonword condition. These results suggested that the processing feature of native Japanese readers relying heavily on visual codes and scarcely on phonological codes had influenced their performance on lexical decision tasks. However, it was inconsistent with their processing feature that the lexical decision time for nonwords was shorter and the error rate for nonwords was lower in the pseudohomophone condition than those in the standard nonword condition. This was considered due to the stroke counts reflecting the visual complexity of *Kanji* characters consisting of pseudohomophones more than those consisting of TL nonwords and standard nonwords used in Experiment 1.

In Experiment 2, the experiment was replicated after making the means of the stroke counts of *Kanji* characters consisting of nonwords in the three conditions as equal as possible. Consequently, lexical decision time was longer and error rate was higher in the TL conditions than those in the pseudohomophone condition and the standard nonword condition, and no difference was found between the latter two conditions. These results indicated that native Japanese readers rely scarcely on phonological codes but rely heavily on visual codes.

These results, in general terms, serve as evidence that processing characteristics of native readers of a certain language could influence performance on the lexical decision task. We hope that this will be a warning to all researchers who use this task.

We finally discuss the process of lexical decision of native Japanese readers in terms of some models of lexical access. According to Shimomura and Goryo (1998), the models of lexical access are roughly divided into two categories: single-route models and dual-route models. Single-route models suppose that phonological information processing always mediates lexical access (e.g., Van Orden, 1987). Dual-route models suppose that visual information processing and phonological information processing proceed in parallel (e.g., Coltheart, 1978), and some models also suppose that the two routes interact with each other (e.g., Ferrand & Grainger, 1994).

The results of this study indicated that phonologically misleading pseudohomophones had no effect on the performance of lexical decision tasks by native Japanese readers. They indicate that phonology does not necessarily mediate lexical access; this finding contradicts the single-route model, which supposes phonological mediation. At the same time, the results of this study contradict the dual-route model as well—the very reason Coltheart (1978) proposed a dual-route model was that pseudohomophones delayed the lexical decision time of native English readers.

However, a few recent studies support the validity of the dual-route model. Grainger, Muneaux, Farioli, and Ziegler (2005) examined the effect of visual and phonological neighborhood density on lexical decision time. They found that lexical decision time was short when both visual neighborhood and phonological neighborhood were dense or sparse because the target lexicon was likely to be the same, namely, because the cross-code consistency was high.

They also found that lexical decision time was prolonged when only one of the neighborhoods was dense because the target lexicon was likely to be different, namely, because cross-code consistency was low. They asserted that these results support the dual-route model, and these results were also found by Hino, Nakayama, Miyamura, and Kusunose (2011) with native Japanese participants.

We, therefore, felt that the results of this study should be explained by a dual-route model. We also considered that it is basically implausible to suppose that the processes of lexical access are disparate among native readers of different languages. How, then, could the results with native English readers and those with native Japanese readers be explained by the same dual-route model?

The Japanese language features an unparalleled number of homophones. This situation is similar to the situation when phonological neighborhood is dense and visual neighborhood is sparse—the cross-code consistency being very low. Native Japanese readers do know that the use of both visual and phonological codes, especially of phonological codes, delays their lexical access. Accordingly, they choose to rely heavily on visual codes and scarcely on phonological codes. On the other hand, the English language does not have as many homophones, and the cross-code consistency is high. Native English readers know the situation and do not reduce their reliance on phonological codes. Therefore, the use of pseudohomophones impaired their performance on lexical decision tasks. In brief, we believe that the results of this study are inconsistent with those of Coltheart et al. (1977) because the phonological processing features of native Japanese readers differ from those of native English readers, although their processes of lexical access are basically similar.

Human information processing must be highly efficient; it is unconceivable that people adhere to inefficient processing. We consider that it is most natural and reasonable to suppose that people change flexibly the weights of visual and phonological processing according to the features of their languages or situations, realizing the most efficient lexical access.

## References

- Amano, N., & Kondo, K. (2003). *NTT database series: Lexical properties of Japanese. Vol. 2. CD-ROM Version*. Tokyo: Sanseido.
- Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing*. New York: Academic Press.
- Coltheart, M., Develaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI*. Hillsdale, NJ: Erlbaum.
- Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 109-127.
- Ferrand, L., & Grainger, J. (1994). Effects of orthography are independent of phonology in masked form priming. *Quarterly Journal of Experimental Psychology*, 47A(2), 365-382.
- Glanzer, M., & Ehrenreich, S. L. (1979). Structure and search of the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, 18, 381-398.
- Grainger, J., Muneaux, M., Farioli, F., & Ziegler, J. C. (2005). Effects of phonological and orthographic neighborhood density interact in visual word recognition. *Quarterly Journal of Experimental Psychology*, 58A, 981-998.
- Hino, Y., Nakayama, M., Miyamura, S., & Kusunose, Y. (2011). Orthographic and phonological neighborhood size effects for Japanese Katakana words in a lexical decision task. *Japanese Journal of Psychology*, 81, 569-576.
- Lupker, S. J., & Pexman, P. M. (2010). Making things difficult in lexical decision: The impact of pseudohomophones and transposed-letter nonwords on frequency and semantic priming effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 1267-1289.
- Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90, 227-234.
- Mizuno, R., & Matsui, T. (2012). Difference due to native language in the determinant of target-distracter discriminability influencing attentional blink. *Tokai Journal of Psychology*, 6, 1-10.
- Mizuno, R., Matsui, T., & Bellezza, F. S. (2007). Difference between native English and native Japanese readers in the use of visual and phonological codes in processing phonograms. *Japanese Journal of Cognitive Psychology*, 5, 1-10.
- Mizuno, R., Matsui, T., Harman, J. L., & Bellezza, F. S. (2008). Encoding times of phonograms by English and Japanese readers: Eliminating the time for attention switching. *Japanese Journal of Cognitive Psychology*, 5, 93-105.
- Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited capacity attention. *Journal of Experimental Psychology: General*, 106, 226-254.
- Parkin, A. J. (1982). Phonological recoding in lexical decision: Effects of spelling-to-sound regularity depend on how regularity is defined. *Memory & Cognition*, 10, 42-53.
- Perea, M., & Lupker, S. J. (2004). Can CANISO prime CASINO? Transposed letter similarity effects with nonadjacent letter positions. *Journal of Memory and Language*, 51, 231-246.
- Shimomura, M., & Goryo, K. (1998). Shinteki jisyo gainen kara mita tango ninchi no mondai [The problems of word recognition in terms of lexicon]. In N. Osaka (Ed.), *Yomi: Nou to kokoro no jouhou shori* [Reading: Information processing in brain and mind]. Tokyo: Asakura.
- Shulman, H. G., & Davidson, T. C. B. (1977). Control properties of semantic coding in a lexical decision task. *Journal of Verbal Learning and Verbal Behavior*, 16, 91-98.
- Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound, and reading. *Memory & Cognition*, 15, 191-198.

# The Effects of Mental Imagery and Embodied Action on L2 Word Learning

**Laura M. Morett (morett@pitt.edu)**

Department of Psychology, University of Pittsburgh  
210 S. Bouquet Street, Pittsburgh, PA 15260 USA

**Raymond W. Gibbs (gibbs@ucsc.edu)**

Department of Psychology, University of California, Santa Cruz  
1156 High Street, Santa Cruz, CA 95064 USA

**Brian MacWhinney (macw@cmu.edu)**

Department of Psychology, Carnegie Mellon University  
5000 Forbes Ave., Pittsburgh, PA 15213 USA

## Abstract

Previous research has provided evidence that mental imagery and embodied action can facilitate word learning in a novel language. However, it is unclear *how* these factors interact—as well as *why* they play a role—in word learning. Through a set of four experiments, this research demonstrated that neither mental imagery nor embodied action directly promotes the acquisition of second language (L2) words by adult learners. Notably, both passive viewing of images and gestures, as well as active engagement in mental imagery and gesture enactment, were insufficient to enhance L2 word learning. These results suggest that adults are effective L2 word learners, and that, because of this, embodied action does not play an essential role in supporting L2 lexical acquisition.

**Keywords:** Second language acquisition, word learning, gesture, mental imagery, embodied cognition

## Introduction

When setting out to learn a new language, it is common to start by learning the meanings of its conceptually simplest units: words. When learning a word, learners must associate the acoustic and/or orthographic form of the word with a representation of the word's meaning. These representations are based on real-world interactions between the learner and the object to which the word refers; thus, they are a composite of the object's properties, as perceived and acted upon by the learner. As a result, word learning likely depends heavily upon both mental imagery and embodied action, in conjunction with text/speech processing and memory. The present research sought to investigate—and dissociate—the contributions of mental imagery and embodied action to second language (L2) word learning. This research was based on two major goals: (1) elucidate the cognitive processes responsible for L2 word learning in adulthood; (2) determine how to engage these processes, thereby enhancing L2 word learning.

Several lines of research indicate that mental imagery plays a central role in L2 word learning. In the keyword method of L2 word learning (Atkinson, 1975), learners are instructed to choose a word from their native language that is phonologically similar to the target L2 word, and to formulate a mental image of the referents of these words

interacting. For example, for the Spanish word *chico* (boy), a learner might choose the similar-sounding English word *chick*, and then could imagine a boy holding a chick. This method has been shown to be a more effective strategy for L2 word learning than verbal association of target words and translations (Atkinson & Raugh, 1975; Levin, McCormick, Miller, Berry, & Pressley, 1982; Raugh & Atkinson, 1975; Van Hell & Mahn, 1997). There is also evidence that, particularly in the early stages of L2 learning, the meanings of target words are learned more effectively via physical images, which cue mental imagery through isomorphism, than via verbal translation (Carpenter & Olson, 2011; Chun & Plass, 1996). These findings indicating that mental imagery promotes L2 word learning can be explained by dual coding theory (Paivio, 1990), which posits that long-term memory encoding can be enhanced via the simultaneous processing of visual and verbal information.

Aside from mental imagery, embodied action—action that learners perform using their bodies—also plays an important role in L2 word learning. Because learners' representations of word referents entail actions that learners have performed on these objects, the use of embodied action during word learning allows learners to create deeper, more meaningful word-referent associations. One method that has taken advantage of embodied action is the Total Physical Response Method (Asher, 1969), which has been successfully used to teach L2 words to novice learners (Asher, Kusudo, & Torre, 1974; Asher & Price, 1967). In this approach, the instructor teaches the target word by saying it while demonstrating its meaning using the body, and learners re-enact the instructor's actions using their own bodies in order to demonstrate comprehension of the words' meanings. Other evidence supporting embodied action comes from research showing that the enactment of gestures representative of word meanings (e.g., placing the hands together and opening them for *book*) during learning facilitates recall of target words to a greater degree than passively viewing pictures or representative gestures (Tellier, 2008). Taken together, these findings indicate that embodied action allows L2 learners to tap into their representations of word referents, creating robust,

multimodal associations between target words and the objects that they represent.

Although extant research suggests that mental imagery and embodied action both contribute separately to L2 word learning, it is unclear from this work if—and how—they interact. One way by which their respective contributions could be clarified is by investigating whether gesture, which typically accompanies and complements speech (McNeill, 2005), facilitates L2 word learning. Iconic (i.e., representative) gestures are created via embodied action and evoke mental imagery via their iconicity, which conveys visuospatial properties of the referent. Several studies (Allen, 1995; Kelly, McDevitt, & Esch, 2009; Tellier, 2008) have shown that L2 words are recalled more accurately over longer intervals when they are learned via representative iconic gesture than when they are learned via speech. Additionally, when L2 words are accompanied by non-representative iconic gesture (e.g., placing the hands together and opening them as for *book* while the word *drink* is presented), they are learned less effectively (Kelly et al., 2009). All of these findings indicate that the combination of mental imagery and embodied action in iconic gesture is a powerful tool for enhancing L2 word learning when it is synchronous with the meanings of the target words.

The objective of the current research was to clarify the independent contributions of mental imagery and embodied action—as well as their interactions—to L2 word learning. To this end, target words were presented in four conditions in which words were accompanied by stimuli that crossed these factors in a 2-by-2 design (see Table 1). These stimuli were designed to elicit either active or passive processing of mental imagery and embodied action, depending on the task instructions. Based on the research discussed above, it was predicted that iconic gesture viewing and enactment would result in the highest number of L2 words recalled, followed

by mental imagery, followed by meaningless embodied action and text/speech only. Furthermore, it was predicted that active learning conditions would allow participants to recall more words than passive conditions due to greater engagement of the sensorimotor system.

## Experiment 1

### Methods

Twenty-six undergraduate students (age:  $M = 20.25$ ;  $SD = 1.5$ ; sex: 11 males; 15 females) at a medium-sized public university in the US participated in return for partial course credit. All participants were fluent English speakers and had no knowledge of Hungarian.

Twelve Hungarian words and their English glosses were used in this research (see Table 2). Prior to this research, 15 English speakers who did not participate in this study were asked to rate the concreteness, imageability, and meaningfulness of the English glosses of 80 candidate words, and to gesture in a way that represented the meaning of each gloss. The 12 words with the most consistent responses from each of the categories were selected for the study. Videos of iconic gestures were created by recording a fluent Hungarian-English bilingual saying these words in each language while enacting the gestures produced by the most participants. Images were line drawings representing the English glosses of Hungarian target words from the International Picture Naming Project (Szekely et al., 2004). In order to control for possible vocal iconicity, audio of the pronunciation of Hungarian and English words was extracted from the iconic gesture videos and was played during presentation of referent images and text of words. Beat gestures—simple, non-iconic hand movements in time to speech prosody—represented embodied action without cuing mental imagery. Videos of these gestures were created by recording the Hungarian-English speaker saying the target words in each language while enacting beat gestures. These videos were presented with their own sound track in order to preserve prosodic speech-gesture synchrony.

The learning phase of this study consisted of three blocks comprising 12 trials apiece. In each learning trial, the

Table 1: Experimental design for Experiments 1-4.

	Mode	+ Mental Imagery	- Mental Imagery
+ Embodied Action	Passive	Iconic gesture	Beat gesture
	Active	Gesture enactment	Meaningless hand motion
- Embodied Action	Passive	Physical images	Text/Speech only
	Active	Mental imagery formation	Verbal repetition only

Table 2: Hungarian and English words from Expts. 1-4.

Hungarian	English
Betegség	Illness
Kalapács	Hammer
Kulcs	Key
Löni	To shoot
Mászni	To climb
Megütni	To hit
Orá	Watch
Öröm	Joy
Seprű	Broom
Tréfa	Joke
Unott	Bored
Varrni	To sew

following sequence of events was repeated twice: a randomly-selected Hungarian word was presented for 2000 ms., and after a 1000 ms. interstimulus interval, the corresponding English word was presented for 2000 ms. This study used a within-participants design; thus, for each experimental session, three Hungarian words and their English glosses were assigned at random to each of the following conditions: (1), iconic gesture, in which words were accompanied by video of a gesture representing their meaning; (2), beat gesture, in which words were accompanied by video of simple, non-iconic gestures made in time to speech; (3), image, in which words were accompanied by an image representing their meaning; (4), text, in which words were presented as text (see Materials section for a description of the stimuli used in each condition).

The test phase of this study consisted of a single block in which each Hungarian word that participants had learned was presented as text and speech. Participants responded by saying the corresponding English word or by saying “skip” if they could not remember it. In order to examine how learning conditions affected long-term L2 word recall, participants completed the test phase at three intervals following the learning phase: five minutes, one week, and one month.

## Results and Discussion

L2 word recall was quantified by scoring responses using a binary coding scheme (1 = correct, 0 = incorrect/skipped), and by converting scores into proportion correct for each participant and condition (in order to control for unscorable responses due to factors such as unintelligibility or technical errors in running the recall task). Proportional scores were submitted to repeated measures ANOVAs, in which participant and word were used as fixed factors.

This analysis revealed a main effect of recall interval,  $F_{pp}(2, 51) = 12.16, p < .001, \eta_p^2 = .36$ ;  $F_{word}(2, 18) = 6.42, p = .008, \eta_p^2 = .42$ . Bonferroni-corrected post hoc analyses showed that participants recalled more L2 words after five minutes than they recalled after one week ( $p_{pp} = .03$ ;  $p_{word} = .02$ ) and one month ( $p_{pp} = .001$ ;  $p_{word} = .06$ ). There was also a main effect of learning condition,  $F_{pp}(3, 77) = 7.70, p < .001, \eta_p^2 = .26$ ;  $F_{word}(3, 27) = 7.04, p = .001, \eta_p^2 = .44$ , see Figure 1. Post-hoc analyses showed that participants learned more words via text than they did via beat gesture ( $p_{pp} = .002$ ;  $p_{word} = .008$ ), iconic gesture (by participant;  $p_{pp} = .05$ ;  $p_{word} = .34$ ) and images ( $p_{pp} = .08$ ;  $p_{word} = .05$ ). However, the interval by condition interaction failed to reach significance. Together, these results suggest that orthographic representations of words may play a more integral role in L2 word learning than mental imagery or embodied action.

## Experiment 2

One possible explanation why L2 words learned via text were recalled better than words learned via iconic gesture, beat gesture, or images is because Hungarian words were

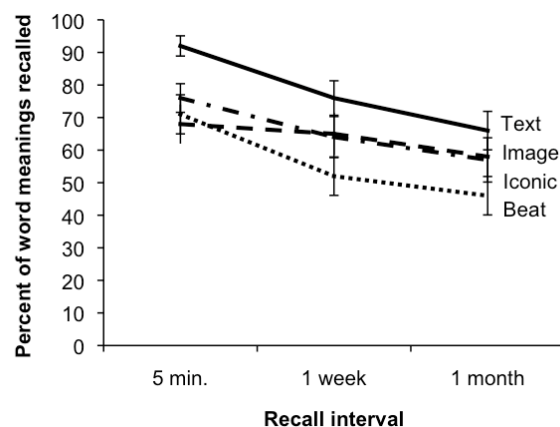


Figure 1: Percent of word meanings recalled by condition and recall interval for Expt. 1 (error bars represent *SEM*).

presented as text in test trials. This learning-test similarity may have resulted in transfer-appropriate processing, which occurs when similar cues are present at encoding and retrieval (Morris, Bransford, & Franks, 1977). In order to determine whether transfer-appropriate processing was responsible for the facilitatory effect of text observed in Experiment 1, recall trials were modified so that Hungarian words were presented as speech only, without any text. If, under these conditions, participants recalled similar numbers of L2 words presented as text, representational gesture, beat gesture, and speech, it could be concluded that the facilitatory effect of text observed in Study 1 was due to transfer-appropriate processing. Given a lack of conclusive reported evidence showing that L2 words learned as text are recalled better than words presented via other modalities (e.g., images), it was predicted that no significant advantage of text would be found in Study 2.

## Methods

Twenty-six undergraduate students (age:  $M = 20.64$ ;  $SD = 1.56$ ; sex: 12 males; 14 females) at a medium-sized public university in the US participated in return for partial course credit. All participants were fluent English speakers and had no knowledge of Hungarian. Additionally, participants had not participated in Experiment 1.

Learning conditions and trials were identical to those used in Experiment 1. Recall trials were also identical to those of Experiment 1, except that Hungarian words were presented as speech only, while the task instructions were displayed on the screen as text.

## Results and Discussion

As in Experiment 1, L2 word learning was quantified proportionally and was submitted to repeated measures ANOVAs with participant and word as fixed factors. This analysis revealed a main effect of recall interval,  $F_{pp}(2, 51) = 7.11, p = .003, \eta_p^2 = .31$ ;  $F_{word}(2, 10) = 47.14, p < .001, \eta_p^2 = .90$ , see Figure 2. Bonferroni-corrected post-hoc analyses showed that participants recalled more words after five

minutes than they recalled after one week ( $p_{pp} = .04$ ;  $p_{word} = .001$ ) and one month ( $p_{pp} = .05$ ;  $p_{word} = .004$ ). However, learning condition failed to reach significance, as did the interval by condition interaction. These results indicate that the superior recall for words learned via text observed in Experiment 1 was due to the presence of text in both learning and test trials.

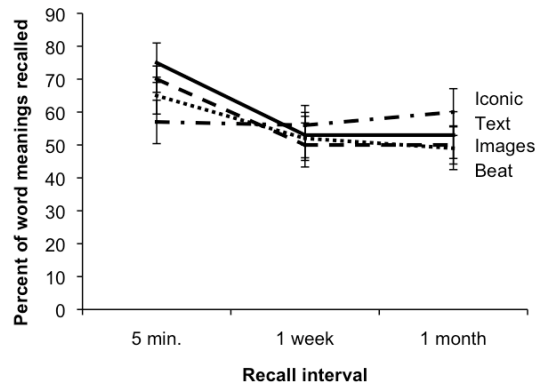


Figure 2: Percent of word meanings recalled by condition and recall interval for Expt. 2 (error bars represent *SEM*).

### Experiment 3

Contrary to the predictions, Experiments 1 and 2 failed to show that iconic gesture, mental imagery, or beat gesture promote L2 word learning via mental imagery and embodied action. One possible reason why iconic gesture may not have produced the facilitatory effect observed in other similar studies of L2 word learning is that participants may have encountered difficulties mapping gestures onto the words that they were intended to represent. Despite the care that was taken in selecting gestures to represent target words based on gesture production data from the norming study, it is possible that the gestures chosen may logically map onto the meanings of more than one word (e.g., the gesture representing to *climb* may also map onto *ladder*). Unclear gesture-word mappings such as these may cause confusion, unnecessarily increasing cognitive load upon presentation of “correct” English glosses, which in turn may negatively impact L2 word learning and recall. Experiment 3 attempted to control for gesture-meaning mismatches by presenting English glosses prior to Hungarian words, in order to ensure that iconic gestures were immediately associated with their intended meanings.

In addition to the reversal of language presentation order, the text condition was replaced with a speech only condition. The rationale for this replacement was that speech with no visual representation may be a more appropriate baseline condition (- embodied action, - mental imagery) for this research than speech with text.

### Methods

Twenty-seven undergraduate students (age:  $M = 21.05$ ;  $SD = 1.83$ ; sex: 13 males; 14 females) at a medium-sized public university in the US participated in return for partial course

credit. All participants were fluent English speakers and had no knowledge of Hungarian. Additionally, participants had not participated in Experiments 1 or 2.

Learning trials were similar to those of Experiments 1 and 2, except that the order of language presentation was reversed (2000 ms. English gloss, 1000 ms. ISI, 2000 ms. Hungarian word). Learning conditions were also similar to those of Experiments 1 and 2, except that the text condition was replaced with a speech only condition, in which words were presented as speech concurrently with a blank screen. Recall trials were identical to those of Experiment 2.

### Results and Discussion

As in Experiments 1 and 2, L2 word learning was quantified proportionally and was submitted to repeated measures ANOVAs with participant and word as fixed factors. This analysis revealed a main effect of recall interval,  $F_{pp}(2, 53) = 24.81$ ,  $p < .001$ ,  $\eta_p^2 = .61$ ;  $F_{word}(2, 14) = 36.59$ ,  $p < .001$ ,  $\eta_p^2 = .84$ , see Figure 3. Bonferroni-corrected post-hoc analyses showed that participants recalled more words after five minutes than they recalled after one week ( $p_{pp} = .04$ ;  $p_{word} = .002$ ) and one month ( $p_{pp} = .04$ ;  $p_{word} < .001$ ). However, learning condition failed to reach significance, as did the interval by condition interaction. Taken together with the findings of Experiments 1 and 2, these results confirm that, in the passive form, mental imagery and embodied action do not significantly enhance the acquisition of L2 words.

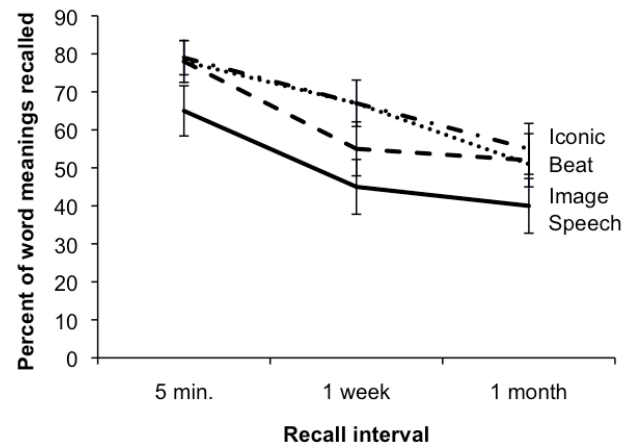


Figure 3: Percent of word meanings recalled by condition and recall interval for Expt. 3 (error bars represent *SEM*).

### Experiment 4

The objectives of this experiment were to examine the effects of the active engagement in embodied motion and mental imagery on L2 word learning, and to compare them to the effects of their passive counterparts, as documented in Experiments 1-3. For young children, only enactment—not mere viewing—of iconic gestures enhances word learning (Tellier, 2008). Thus, it is possible that L2 word learning



may be enhanced via *enactment* of iconic gestures and *visualization* of mental imagery, even though word learning was not facilitated via passive *viewing* of iconic gestures and physical images. Experiment 4 tested this possibility by examining the effectiveness of L2 word learning via the active equivalents of the learning conditions of Studies 1-3.

## Methods

Twenty-eight undergraduates (Age:  $M = 20.67$ ;  $SD = 1.56$ ; Sex: 12 males; 16 females) at a medium-sized public university in the US participated in return for partial course credit. All participants were fluent English speakers and had no knowledge of Hungarian. Additionally, participants had not participated in Experiments 1, 2, or 3.

As in Experiments 1-3, the learning phase of Experiment 4 consisted of three blocks of trials that varied by condition within participants. In learning trials, participants viewed video of a speaker saying an English gloss while producing a gesture representative of its meaning for 2000 ms., and after a 1000 ms. interstimulus interval, viewed video of the speaker saying a Hungarian target word while making the same gesture for 2000 ms. After one additional repetition of this sequence of events, participants repeated the English and Hungarian words aloud while performing the action corresponding to the condition to which the word had been assigned for that session. For words assigned to condition (1), gesture enactment, participants enacted the gesture that they had viewed in the video; for words assigned to condition (2), mental imagery formation, participants closed their eyes and visualized the words' meaning;<sup>1</sup> for words in condition (3), meaningless hand motion, participants made an X-shaped hand motion three times; for words in condition (4), repetition, participants repeated the words aloud while keeping their hands still. Recall trials were identical to those of Experiments 2-3.

## Results and Discussion

As in Experiments 1-3, L2 word learning was quantified proportionally and was submitted to repeated measures ANOVAs with participant and word as fixed factors. This analysis revealed a main effect of recall interval,  $F_{pp}(2, 56) = 12.90$ ,  $p < .001$ ,  $\eta_p^2 = .52$ ;  $F_{word}(2, 20) = 10.37$ ,  $p = .001$ ,  $\eta_p^2 = .51$ , see Figure 4. Bonferroni-corrected post hoc analyses revealed that participants recalled more words after five minutes than they recalled after one week ( $p_{pp} = .05$ ;  $p_{word} = .02$ ) or one month ( $p_{pp} = .03$ ;  $p_{word} = .01$ ). However, learning condition failed to reach significance, as did the interval by condition interaction. These results suggest that, although enactment facilitates L2 word learning by children, it does not seem to enhance L2 word learning by adults.

<sup>1</sup>The results of two participants who indicated in response to a question on the post-experimental survey that they did not follow the instructions regarding mental imagery were excluded.

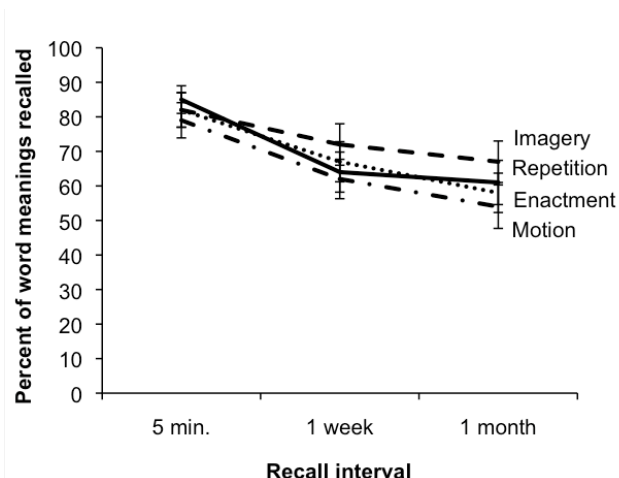


Figure 4: Percent of word meanings recalled by condition and recall interval for Expt. 4 (error bars represent *SEM*).

## General Discussion

The present research examined the independent roles and interactions that passive and active engagement in mental imagery and embodied action play in L2 word learning. The lack of differences in word learning as a function of condition suggest that neither of these factors plays an integral role in L2 word learning by adults. Given that adults are particularly effective word learners (Snow & Hoefnagel-Höhle, 1978), it is likely that the participants of this study were using alternative methods to associate target words with their meanings, such as phonological association or generation. One notable difference between this study and others (Kelly et al., 2009) is that, in the interest of ecological validity, the stimuli accompanying the target words in the current study were either congruent or neutral to the word meanings. As a result, the variation in word learning between conditions in the current study was more modest than that observed in previous studies. This greater similarity between conditions may explain the lack of significant differences between conditions, as well as why the results of the current study failed to replicate those of previous work (e.g., Kelly et al., 2009; Tellier, 2008).

Surprisingly, the results failed to show that meaningful embodied action, in the form of representative iconic gesture, facilitated L2 word learning and recall more effectively than mental imagery. This finding is inconsistent with several previous studies showing superior recall of L2 words learned via viewing or enactment of iconic gesture (Kelly et al., 2009; Tellier, 2008) or enactment of meaningful motion using the body (Asher, 1969; Asher et al., 1974). There are several possibilities why iconic gesture may have failed to facilitate L2 word learning in the current research. One possibility is that the gestures chosen to represent target words may not have been sufficiently iconic, and thus, may not have been as imagistic and meaningful as gestures used in other studies. Another possibility is that the learning phase may have been too brief to allow for associations between gestures and target words

to be formed. Future research should test these possibilities by examining the acquisition of L2 words via imagistically-rich gesture during an extended learning period, and by comparing the results to those of the current research.

Overall, the results of the current research suggest that neither mental imagery nor embodied action plays a key role in L2 word learning by adults. More specifically, the results indicate that the viewing of iconic gesture, images, and text during L2 word learning result in comparable recall of target words across both long and short learning-test intervals. Finally, the results demonstrate that L2 words are recalled more accurately over short (5 min.) learning-test intervals than longer intervals (1 week, 1 month). Taken together, these findings fail to replicate the results of work showing that representative iconic gesture viewing and enactment enhance L2 word learning (Kelly et al., 2009; Tellier, 2008), suggesting that the word learning techniques of adult L2 learners are already so effective that mental imagery and embodied action have a negligible impact on them.

### Acknowledgements

This research was supported by a National Defense Science and Engineering Graduate Fellowship and the Perlino Award to Laura M. Morett. The authors thank Jana Iverson for resources and helpful discussion.

### References

- Allen, L. Q. (1995). The effects of emblematic gestures on the development and access of mental representations of French expressions. *The Modern Language Journal*, 79(4), 521–529.
- Asher, J. J. (1969). The total physical response approach to second language learning. *The Modern Language Journal*, 53(1), 3–17.
- Asher, J. J., Kusudo, J. A., & Torre, R. de la. (1974). Learning a second language through commands: The second field test. *The Modern Language Journal*, 58(1/2), 24–32.
- Asher, J. J., & Price, B. S. (1967). The learning strategy of the Total Physical Response: Some age differences. *Child Development*, 38, 1219–1227.
- Atkinson, R. C. (1975). Mnemotechnics in second-language learning. *American Psychologist*, 30(8), 821–828. doi:10.1037/h0077029
- Atkinson, R. C., & Raugh, M. R. (1975). An application of the mnemonic keyword method to the acquisition of a Russian vocabulary. *Journal of Experimental Psychology: Human Learning and Memory*, 1, 126–133. doi:10.1037/0278-7393.1.2.126
- Carpenter, S. K., & Olson, K. M. (2011). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi:10.1037/a0024828
- Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *Modern Language Journal*, 80(2), 183–198.
- Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, 24, 313–334. doi:10.1080/01690960802365567
- Levin, J. R., McCormick, C. B., Miller, G. E., Berry, J. K., & Pressley, M. (1982). Mnemonic versus nonmnemonic vocabulary-learning strategies for children. *American Educational Research Journal*, 19(1), 121–136. doi:10.3102/00028312019001121
- McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.
- Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 16(5), 519–533. doi:10.1016/0022-5371(77)80016-9
- Paivio, A. (1990). *Mental representations: A dual coding approach*. Oxford University Press US.
- Raugh, M. R., & Atkinson, R. C. (1975). A mnemonic method for learning a second-language vocabulary. *Journal of Educational Psychology*, 67(1), 1–16. doi:10.1037/h0078665
- Snow, C. E., & Hoefnagel-Höhle, M. (1978). The critical period for language acquisition: Evidence from second language learning. *Child development*, 1114–1128.
- Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., Lu, C. C., et al. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, 51(2), 247–250. doi:10.1016/j.jml.2004.03.002
- Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, 8, 219–235.
- Van Hell, J. G., & Mahn, A. C. (1997). Keyword mnemonics versus rote rehearsal: Learning concrete and abstract foreign words by experienced and inexperienced learners. *Language Learning*, 47(3), 507–546. doi:10.1111/0023-8333.00018

# How the Hands Cue the Mind: The Effects of Iconicity and Enactment on Sign Language Acquisition

Laura M. Morett (morett@pitt.edu)

Department of Psychology, University of Pittsburgh  
210 S. Bouquet Street, Pittsburgh, PA 15260 USA

## Abstract

Iconicity is a powerful cue to symbolic meaning. However, it is unclear from previous research whether language learners benefit from iconicity. Prior research indicates that the motor system supports language acquisition, suggesting that iconicity expressed via this modality may be particularly salient. The present study investigates the effects of iconicity and enactment on the acquisition of American Sign Language by hearing adults. The results reveal that enactment enhances sign learning in general, but fail to show that iconic signs are learned more effectively than non-iconic signs. As such, they indicate that the motor system—but not iconicity—plays a key role in sign language acquisition.

**Keywords:** Second language acquisition, sign language, mental imagery, embodied cognition.

## Introduction

Sign language is the only type of natural language that is comprehended and produced exclusively in the visuospatial modality. Given that the visuospatial modality allows for greater isomorphism between symbols and their referents than the auditory modality, it follows that sign language should be more iconic than spoken language, and there is evidence that this is indeed the case (McNeill, 2005; O'Brien, 1999). Thus, although hearing speakers are accustomed to processing language in the auditory modality, they may be able to take advantage of this iconicity to expedite their learning of sign language. If iconicity plays a pivotal role in sign language acquisition, learners should be able to acquire sign languages more quickly and effectively than they learn spoken languages. Moreover, learners should be able to learn iconic signs and expressions more efficiently than lexical items that are not iconic.

Unlike spoken language, which is articulated primarily with the mouth and vocal tract, sign language is articulated with the hands and body. As such, another factor that may play an integral role in the acquisition of sign language is the engagement of the motor system. Theories of

embodied cognition posit that representations of language are inherently perceptual, and are encoded and retrieved via the body's sensorimotor system (Barsalou, 1999). Thus, these theories would predict that enacting signs—especially those that are iconic—allows sign language learners to tap directly into these perceptually-based representations, thereby facilitating their recall and comprehension. If the motor system does contribute significantly to sign language acquisition, learners should recall iconic signs better than non-iconic signs due to the isomorphism between the visuospatial properties of motor representations of signs and their referents.

## Iconicity and Language Acquisition

Meaningful hand movements, including gestures and signs, vary on the basis of several qualities, including conventionalization, semiosis, and relationship to speech. In order to show how different types of hand movements relate to one another on the basis of these characteristics, Adam Kendon and David McNeill (1992) developed a continuum, which is illustrated below in Figure 1. At one extreme of the continuum lies sign language, which is highly conventionalized, segmented and analytic, and occurs in lieu of speech. At the opposite extreme lies gesticulation, which is unconventionalized, global and synthetic, and occurs concurrently with speech. Although iconicity is not plotted on this continuum, it can be inferred that, due to its global and synthetic (i.e., holistic) nature, gesticulation is highly iconic, whereas sign is the least iconic of the hand motions.

It is important to note that iconicity varies within and between sign languages. Much of this variation can be explained by ontogenetic development. There is evidence that the home sign of individual deaf children as well as pidgin sign languages created by communities of deaf children are generally more iconic than conventionalized sign languages (Kendon, 1980; Senghas, Kita, & Özyürek, 2004). Moreover, even within highly-conventionalized sign

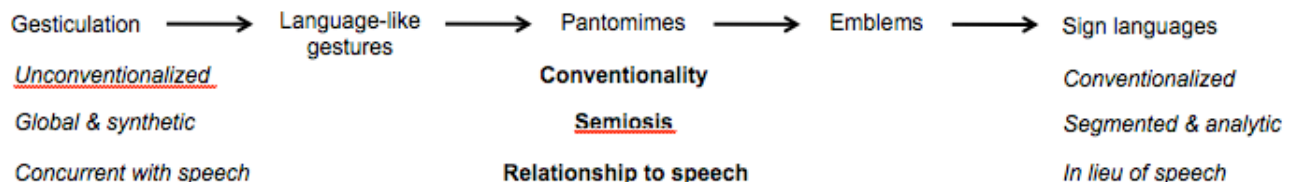


Figure 1: Kendon's continuum, as characterized by McNeill (2005).

languages, recently-coined signs are more iconic than signs that have been in the language longer (Frishberg, 1975). This is likely the case because signs are initially based on referents' affordances, which become obfuscated through inter-generational transmission of signs.

Obviously, it is quite plausible that iconicity may facilitate the acquisition of signs, due to its isomorphism with the visuospatial properties of the referent. Nevertheless, research has failed to provide conclusive evidence that children learn iconic signs more readily than they learn arbitrary signs. One study of the acquisition of American Sign Language (ASL) by congenitally deaf children showed that only 30% of these children's first 10 signs are iconic, and that this number increases to only 34% at 18 mos. (Orlansky & Bonvillian, 1984). Another study (Miller, 1987) showed that 3-year-old hearing children unfamiliar with ASL were unable to reliably select the correct referent of iconic signs on the Peabody Picture Vocabulary Test, a standardized, forced-choice measure of vocabulary development (L. M. Dunn & Dunn, 1997). Taken together, these findings suggest that both deaf and hearing children are unable to use signs' iconicity to associate them with their referents, thereby facilitating sign language acquisition.

Related work examining gesture comprehension has provided insight into the question of why young children are unable to associate iconic hand movements with their referents. One study showed that, by 26 months of age, children were able to associate iconic gestures with objects with similar affordances, even though they were unable to do so at 14 mos. of age (Namy, 2008). Another, more recent study demonstrated that 4-year-olds were better able than 2-year-olds to learn object labels associated with iconic gestures, but that both age groups learned object labels associated with arbitrary gestures at a comparable rate (Marentette & Nicoladis, 2011). Although this study also showed that 4-year-old children treated iconic gesture as an action associate rather than a label, 2-year-olds were not tested. The seeming inconsistency between the results of these two studies can be explained by the fact that the objects' affordances were demonstrated to children in the earlier study, but not the more recent study. Thus, this work demonstrates that children who were able to effectively associate iconic gestures with corresponding objects understand the relationship between them, allowing these children to use gesture as an embodied aid in word learning.

To date, no published research has examined whether the iconicity of sign language facilitates its acquisition by adult learners. However, there is evidence that adults unfamiliar with sign language can effectively guess the meanings of highly iconic signs, even when their referents are metaphorical (O'Brien, 1999). Work has also shown that adults are able to learn words from a novel spoken language accompanied by representative iconic gestures more effectively than words presented as speech only or words accompanied by non-representative iconic gestures (Allen,

1995; Kelly, McDevitt, & Esch, 2009). Given that adults understand the correspondences between object affordances and iconic hand movements—including signs—it follows that they should be better able to learn highly iconic signs than arbitrary signs.

## **Enactment and Language Acquisition**

Aside from being more iconic than spoken language, sign language is also more embodied than spoken language. Because the hands and parts of the body other than the vocal tract and face play a larger role in sign language than in spoken language, sign language engages the motor system to a greater degree than spoken language. This engagement of the motor system likely produces memory traces that are richer and more multimodal than those produced by spoken language, providing learners with additional recall cues.

Research examining recall of spoken language has provided evidence that engagement of the motor system during language processing enhances memory encoding and retrieval. For example, when presented with a series of instructions, adults recall more spoken instructions when they act them out than when they repeat them aloud (Svensson & Nilsson, 1989). Moreover, adults are more likely to recall spoken instructions for tasks that they have enacted for a longer time period (30 s.) than those that they have enacted for a brief time period (5 s.) (Cohen & Bryant, 1991), indicating that greater engagement of the motor system produces richer, more robust memories. A separate line of research has provided evidence that adults are more likely to produce sought-after words during speech disfluencies when they gesture than when they keep their hands still (Frick-Horbury, 2002; Frick-Horbury & Guttentag, 1998), indicating that gesture facilitates lexical access. Taken together, the results of all of this work suggests that the enactment of meaningful hand motions during language processing allows speakers to tap into their semantic representations more effectively, thereby promoting language encoding and recall.

There is also evidence that the motor system plays a key role in language acquisition. To this end, research has revealed a tight relationship between motor and language milestones in childhood, demonstrating that the onset of babbling is accompanied by repetitive motor movements (Iverson & Fagan, 2004), and that the transition to two-word speech is accompanied by gesture-word combinations (Capirci, Iverson, Pizzuto, & Volterra, 1996). Furthermore, several studies have shown that children are able to express symbolic meaning via hand motions before speech (Acredolo & Goodwyn, 1988; Bonvillian, Orlansky, & Novack, 1983; Iverson & Goldin-Meadow, 2005), and that children's iconic gesture production predicts their vocabulary development (Rowe & Goldin-Meadow, 2009a, 2009b; Rowe, Özçalışkan, & Goldin-Meadow, 2008). Finally, there is evidence that school-aged children are better able to learn the meanings of novel words from both

ASL Sign					
English	Adult	To answer	Box	Color	To count
Sign Type	Arbitrary	Metaphorical	Iconic	Arbitrary	Metaphorical

ASL Sign					
English	Country	Cup	Friend	Goal (objective)	Hammer
Sign Type	Arbitrary	Iconic	Metaphorical	Metaphorical	Iconic

ASL Sign					
English	To help	Message	Name	Pool (billiards)	To roll
Sign Type	Metaphorical	Metaphorical	Arbitrary	Iconic	Iconic

ASL Sign					
English	Sauce	Saw	To teach	Team	To twist
Sign Type	Iconic	Iconic	Metaphorical	Metaphorical	Iconic

Figure 2: ASL signs and English words used in study, listed by sign type.

their native language and unfamiliar second languages when they enact iconic gestures representing the words' meanings (Tellier, 2005, 2008).

It is important to note that the facilitatory effects of enactment observed in most studies discussed above stem from a combination of embodied action and mental imagery. Furthermore, there is experimental evidence that learning techniques incorporating mental imagery enhance second language vocabulary acquisition (Atkinson, 1975; Atkinson & Raugh, 1975). Aside from investigating the effects of iconicity and enactment on sign language acquisition, a secondary goal of the current study was to disambiguate the roles that embodied action and mental imagery play in sign enactment. As such, the study included conditions that were designed to elicit mental imagery and embodied action in combination, only mental imagery, only embodied action, and neither mental imagery nor embodied action.

To date, no published research has investigated whether enactment facilitates the acquisition of signed second languages by adults. On the basis of previous research, it was predicted that the enactment condition would result in ASL sign acquisition superior to that observed under the other conditions. This prediction stems from enactment's incorporation of both mental imagery and embodied action, and its resulting engagement of both the visuospatial and motor systems.

## Method

### Participants

Undergraduate students were recruited from the participant pool at the University of Pittsburgh, and received partial course credit in return for participation. All recruited individuals were fluent English speakers<sup>1</sup> and confirmed that they had no knowledge of American Sign Language (ASL) prior to the experiment. Additionally, all recruited individuals had normal hearing and normal or corrected-to-normal vision. 6 individuals were eliminated due to technical difficulties or failure to return for all three sessions, resulting in a final sample of 29 participants (age:  $M = 20.79$ ,  $SD = 1.65$ ; sex: 11 males; 18 females).

### Stimuli

Twenty ASL signs and their English glosses were used as stimuli for this research (see Figure 2). Each sign was classifiable into one of the following three types: Iconic, metaphorical, or arbitrary. Iconic signs depicted their

<sup>1</sup> Participants were not required to be native English speakers in order to participate, given that the English glosses of the signs were common words, and should thus be comprehensible to non-native undergraduate students, whose proficiency must be sufficient to comprehend academic English.



referent holistically or metonymically (e.g., pantomiming hammering for *hammer*); metaphorical signs represented the source domain of the conceptual metaphor structuring their referent (e.g., cupped hands moving forward three times, as if conveying an entity of information from the signer to the listener, for *to teach*); and arbitrary signs bore no structural resemblance to their referent (e.g., two fingers from both hands tapping one another repeatedly for *name*). The distinctions between these sign types were supported by empirical data collected from a separate group of participants unfamiliar with ASL (O'Brien, 1999), ensuring that they were applicable to the target population of the current study.

A female native signer of ASL was video recorded demonstrating the twenty signs used in this study. The signer was unaware of the goals of the study. Video footage of each sign was segmented and trimmed, yielding stimuli averaging 2.5 s. in duration. Additionally, ambient audio captured during video recording was expunged from the footage, yielding silent stimuli.

## Procedure

This experiment consisted of three sessions, the first of which included both a learning and test phase, and the second and third of which included only a test phase. In learning trials, participants were presented with video of a randomly-selected sign (~2500 ms.), and after a 1000 ms. interstimulus interval, were presented simultaneously with the corresponding English gloss as text and audio (2500 ms.). After one additional repetition of this sequence of events, participants performed one of four actions. For words presented in the enactment condition, participants enacted the sign with their own hands; as such, this condition included both mental imagery and embodied action. For words presented in the imagery condition, participants closed their eyes and visualized the sign's referent in their mind's eye without moving their hands; as such, this condition included mental imagery, but not embodied action. For words presented in the motion condition, participants made an X-shaped motion with their dominant hand three times; as such, this condition included embodied action but not mental imagery. For words presented in the comprehension condition, the learning sequence was repeated one additional time, and participants were not explicitly told to do anything; as such, this condition did not include either mental imagery or embodied action. Within each experimental session, each sign was randomly assigned to one of these four conditions in a within-participants design, such that five signs were presented in each condition for each participant. The learning phase consisted of 3 blocks comprising 20 trials apiece (one for each sign), yielding a total of 60 learning trials altogether.

Following the learning phase in the first session, participants were given a 5-minute break, and then completed the test phase. In test trials, upon being presented with English glosses as text and audio, participants were

asked to produce the corresponding ASL sign. Participants were instructed to try to recall the sign as best they could, but were told that they could say "skip" to move on if they could not recall a sign. During test trials, participants' signing was recorded by a video camera set approximately 45° to the left of their central viewing point. The test phase consisted of one block of 20 trials (one for each sign). Overall, the first experimental session lasted about 30 minutes.

In order to examine how long-term recall of signs varied by condition, participants returned to the lab for two follow-up sessions held one week and four weeks after the first session. In each of these sessions, participants completed the test phase in the manner described above. Each of the follow-up sessions lasted approximately 10 min. apiece.

## Results

Sign recall was quantified using a binary coding scheme (1 = correct; 0 = incorrect/skipped). Total number of signs recalled correctly for each participant and condition were converted into proportions, in order to control for unscorable responses caused by technical errors in running the recall task (which accounted for less than 5% of the data). In order for a sign to be coded as correct, it must have been performed using the same hand (dominant/non-dominant, as specified per participant on a post-experimental questionnaire), and must have had the same hand shape and movements as the correct ASL sign, as modeled by the signer.

To address the question of whether learning condition affects sign recall, proportional data were submitted to repeated measures ANOVAs, using participant and sign as fixed factors. These analyses revealed significant main effects of learning condition,  $F_{pp}(3, 87) = 7.16, p < .001, \eta_p^2 = .29$ ;  $F_{sign}(3, 45) = 14.07, p < .001, \eta_p^2 = .48$ , and recall interval,  $F_{pp}(1, 29) = 10.99, p = .004, \eta_p^2 = .38$ ;  $F_{sign}(1, 15) = 18.16, p = .001, \eta_p^2 = .55$ , but failed to reveal a significant condition-by-interval interaction,  $F_{pp} > 1$ ;  $F_{sign} > 1$ ; see Figure 3. Bonferroni-corrected post-hoc analyses showed greater recall accuracy for signs learned via enactment than via mental imagery ( $p = .04$ ), motion ( $p = .06$ ), and

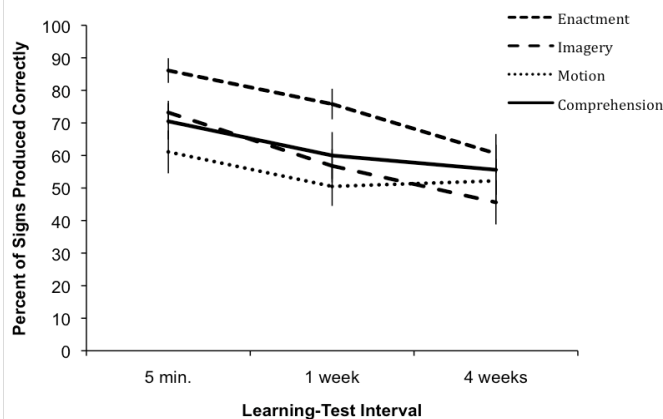


Figure 3: Percent of signs produced correctly by learning condition and recall interval (error bars represent SE).

Table 1: Mean number of signs produced correctly by sign type and recall interval.

Sign type	Recall interval		
	5 min.	1 week	4 weeks
Iconic	.74 (.36)	.66 (.24)	.62 (.28)
Metaphorical	.78 (.24)	.55 (.24)	.50 (.29)
Arbitrary	.60 (.26)	.56 (.28)	.44 (.27)

comprehension ( $p = .05$ ), as well as greater recall accuracy after an interval of 5 minutes than 1 week ( $p < .01$ ) and 4 weeks ( $p < .001$ ). These results indicate that enactment facilitates the acquisition of novel signs by hearing adult learners unfamiliar with sign language across both short and long learning-test intervals.

To address the question of whether iconicity affects the learning and recall of ASL signs, sign type (iconic, metaphorical, arbitrary) was entered into a repeated measures ANOVA, using sign as a fixed factor. This analysis failed to reveal a main effect of sign category on recall,  $F(1, 17) = 1.13$ ,  $p = .35$ ,  $\eta_p^2 = .12$ ; see Table 1. This result indicates that, similar to deaf children, hearing adult learners do not benefit significantly from iconicity when learning novel signs.

## Discussion

The current study investigated the roles of iconicity and enactment on the acquisition of ASL signs by hearing adult L2 learners. The results revealed that enactment facilitated sign learning more effectively than visualization of sign referents, performance of meaningless hand movements, or simple sign comprehension. However, the results failed to demonstrate that iconic signs are learned more effectively than arbitrary signs. Considered as a whole, these results suggest that enactment enhances ASL sign recall and production in hearing learners through the creation of motorically-rich lexical traces.

Unfortunately, the results of the current study do not provide insight into why adult L2 learners fail to benefit from the iconicity inherent in some signs, and in sign language in general. One possible explanation is that, like children, adults go through a developmental stage in the initial stages of language learning in which they are unable to associate the visuospatial properties of iconic signs with the affordances of their referents. Although adults are generally familiar with the affordances of common objects, it is possible that this inability to associate them with their corresponding signs derives from insufficient linguistic context, rather than from insufficient domain-general knowledge, which has been proposed to explain children's insensitivity to iconicity. When learning their first set of signs, adults unfamiliar with ASL are unable to relate their semantic and phonological properties to other similar signs, which may negate their ability to recognize iconicity. Alternatively, hearing adults' experience with spoken languages, in which iconicity is sparse, may lead them to

assume that language is not iconic, causing them to ignore any physical correspondences between signs and their referents. Finally, the novelty of processing language in the visuospatial modality may place a heavy cognitive load on adult L2 learners unfamiliar with sign language, negating any benefits that iconicity may have bestowed. Needless to say, future research is necessary to test between these possibilities and to clarify the cause of this null effect.

The advantage produced by enactment of signs during learning indicates that the motor system plays a key role in L2 lexical acquisition, particularly for sign language. Of note, only *meaningful* motion (i.e., sign enactment)—not arbitrary motion (i.e., X-shaped motions)—enhanced sign acquisition. This result is consistent with embodied theories of cognition, which maintain that the mental representations underlying language derive from meaningful interactions between our bodies and the world (Barsalou, 1999). It is also consistent with work showing that meaningless repetitive motion can disrupt the formation of visuospatial representations (Vandierendonck, Kemps, Fastame, & Szmalec, 2004). The observation that enactment is more effective at promoting sign learning than visualization of referents via mental imagery indicates that meaningful engagement of the motor system results in richer, more robust mental representations of signs, which are more likely to be retrieved successfully, particularly by inexperienced learners.

In conclusion, the results of this study demonstrate that adult L2 learners can take advantage of enactment, but not iconicity, to facilitate their acquisition of sign language. As such, they indicate that the hands cue the mind in sign language acquisition, rather than vice versa, demonstrating the power and depth of the body's cognitive capacity in relation to the acquisition of a novel second language.

## Acknowledgements

This research was supported by a National Defense Science and Engineering Graduate Fellowship and the Perlino Award to Laura M. Morett. The author thanks Ray Gibbs and Brian MacWhinney for helpful discussion.

## References

- Acredolo, L., & Goodwyn, S. (1988). Symbolic gesturing in normal infants. *Child Development*, 59(2), 450–466.
- Allen, L. Q. (1995). The effects of emblematic gestures on the development and access of mental representations of French expressions. *The Modern Language Journal*, 79(4), 521–529.
- Atkinson, R. C. (1975). Mnemotechnics in second-language learning. *American Psychologist*, 30(8), 821–828. doi:10.1037/h0077029
- Atkinson, R. C., & Raugh, M. R. (1975). An application of the mnemonic keyword method to the acquisition of a Russian vocabulary. *Journal of Experimental Psychology: Human Learning and Memory*, 1, 126–133. doi:10.1037/0278-7393.1.2.126



- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(04), 577–660. doi:null
- Bonvillian, J. D., Orlansky, M. D., & Novack, L. L. (1983). Developmental Milestones: Sign Language Acquisition and Motor Development. *Child Development*, 54(6), 1435–1445. doi:10.2307/1129806
- Capirci, O., Iverson, J. M., Pizzuto, E., & Volterra, V. (1996). Communicative gestures and the transition to two-word speech. *Journal of Child Language*, 23, 645–73.
- Cohen, R. L., & Bryant, S. (1991). The role of duration in memory and metamemory of enacted instructions (SPTs). *Psychological research*, 53(3), 183–187.
- Dunn, L. M., & Dunn, L. (1997). Peabody picture vocabulary test-III (PPVT-III). *Circle Pines, MN: American Guidance Service*.
- Frick-Horbury, D. (2002). The use of hand gestures as self-generated cues for recall of verbally associated targets. *The American Journal of Psychology*, 115(1), 1–20. doi:10.2307/1423671
- Frick-Horbury, D., & Guttentag, R. E. (1998). The effects of restricting hand gesture production on lexical retrieval and free recall. *The American Journal of Psychology*, 111(1), 43–62.
- Frishberg, N. (1975). Arbitrariness and Iconicity: Historical Change in American Sign Language. *Language*, 51(3), 696–719. doi:10.2307/412894
- Iverson, J. M., & Fagan, M. K. (2004). Infant Vocal–Motor Coordination: Precursor to the Gesture–Speech System? *Child Development*, 75(4), 1053–1066. doi:10.1111/j.1467-8624.2004.00725.x
- Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, 16(5), 367–371. doi:10.1111/j.0956-7976.2005.01542.x
- Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, 24, 313–334. doi:10.1080/01690960802365567
- Kendon, A. (1980). A description of a deaf-mute sign language from the Enga Province of Papua New Guinea with some comparative discussion. *Semiotica*, 32(1-2), 81–118.
- Marentette, P., & Nicoladis, E. (2011). Preschoolers’ interpretations of gesture: Label or action associate? *Cognition*.
- McNeill, D. (1992). *Hand and mind*. University of Chicago Press.
- McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.
- Miller, M. S. (1987). Sign iconicity: Single-sign receptive vocabulary skills of nonsigning hearing preschoolers. *Journal of Communication Disorders*, 20(5), 359–365. doi:10.1016/0021-9924(87)90024-4
- Namy, L. L. (2008). Recognition of iconicity doesn’t come for free. *Developmental Science*, 11(6), 841–846.
- Orlansky, M. D., & Bonvillian, J. D. (1984). The Role of Iconicity in Early Sign Language Acquisition. *J Speech Hear Disord*, 49(3), 287–292.
- O’Brien, J. (1999). Metaphoricity in the Signs of American Sign Language. *Metaphor and Symbol*, 14(3), 159–177. doi:10.1207/S15327868MS140301
- Rowe, M. L., & Goldin-Meadow, S. (2009a). Early gesture selectively predicts later language learning. *Developmental Science*, 12(1), 182–187.
- Rowe, M. L., & Goldin-Meadow, S. (2009b). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science*, 323(5916), 951–953. doi:10.1126/science.1167025
- Rowe, M. L., Özçalışkan, Ş., & Goldin-Meadow, S. (2008). Learning words by hand: Gesture’s role in predicting vocabulary development. *First language*, 28(2), 182–199.
- Senghas, A., Kita, S., & Özyürek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science*, 305(5691), 1779.
- Svensson, T., & Nilsson, L. G. (1989). The relationship between recognition and cued recall in memory of enacted and nonenacted information. *Psychological research*, 51(4), 194–200.
- Tellier, M. (2005). *How do teacher’s gestures help young children in second language acquisition?* Text presented at the International Society of Gesture Studies, Lyon, France. Retrieved from [http://gesture-lyon2005.ens-lyon.fr/article.php3?id\\_article=253](http://gesture-lyon2005.ens-lyon.fr/article.php3?id_article=253)
- Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, 8, 219–235.
- Vandierendonck, A., Kemps, E., Fastame, M. C., & Szmalc, A. (2004). Working memory components of the Corsi blocks task. *British Journal of Psychology*, 95(1), 57–79. doi:10.1348/000712604322779460

# Does thinking make you biased? The case of the engineers and lawyers problem

Kinga Morsanyi (km574@cam.ac.uk)

Department of Experimental Psychology, University of Cambridge  
Downing street, Cambridge, CB2 3EB, UK

Simon J. Handley (S.Handley@plymouth.ac.uk)

School of Psychology, University of Plymouth  
Drake Circus, Plymouth, PL4 8AA, UK

## Abstract

In this study we examined the cognitive processes involved in engineers and lawyers-type problems, using a novel method (i.e., asking for liking ratings). We were particularly interested in how participants process information about personality descriptions and base rates, which are provided in the problems. In line with previous research, we found that people detect the conflict between descriptions and base rates. Nevertheless, when instructed to reason logically, instead of relying on base rates, participants resolved the conflict by showing higher preference for description-based responses.

**Keywords:** conflict detection; dual-process theories; engineers and lawyers problem; heuristics and biases; instruction manipulation; liking ratings.

Dual process theories of reasoning and decision making (e.g., Evans & Over, 1996; Stanovich, 1999) propose that higher order cognition is based on two qualitatively different types of process. Type 1 (i.e., heuristic) processes are assumed to operate fast and automatically with little demand of cognitive capacity, whereas Type 2 processes are slow, conscious, and demanding of computational resources. The tasks used in the heuristics and biases literature (see e.g., Kahneman, Slovic & Tversky, 1982) can be answered by giving a heuristic-based response, or a response which corresponds to a normative rule of probability (although some theorists have questioned the assumption that giving probability-based responses to these problems is more normative than giving description-based responses – see e.g., Hertwig & Gigerenzer, 1999). Kahneman and Frederick (2005) used these tasks as illustrations for Type 1 and 2 processes at work. For example, consider the classic engineers and lawyers problem (Kahneman & Tversky, 1973; see Table 1.). In the original (conflict) problem, base rate information, which strongly favors lawyers, is presented together with a stereotypical description of an engineer. When participants are asked to decide if the person is more likely to be an engineer or a lawyer, they tend to give the response which corresponds to the description.

An interesting question is whether participants experience a conflict while they solve these tasks, or if they just give the first response that comes to mind. Dual-process theorists (e.g., Evans, 2006) assume that, as Type 1 processes operate quickly and automatically, all participants are inclined to give a Type 1 response by default.






Nevertheless, some individuals (usually the ones of higher cognitive ability – see e.g., Stanovich & West, 2000) are able to inhibit this initial response tendency, and give a response which is based on Type 2 computations. Thus, participants who eventually give a normative response are expected to experience a strong conflict between Type 1 and 2 response tendencies. However, what happens in the case of the majority of the participants who give a heuristic response (which is supposed to be delivered by Type 1 processes)? Do they experience any inner struggle, or do they simply give the first response that comes to mind without ever considering probabilistic information?

In a number of recent studies De Neys and colleagues (e.g., De Neys, Cromheeke & Osman, 2011; De Neys & Glumicic, 2008) used different versions of the engineers and lawyers problem (see Table 1 for illustrations). Besides the original version, they also developed a *non-conflict* task where base-rates and the description pointed to the same response, and they also used a *neutral* task where base rate information was presented together with a description which had no relevance to the choice options. Note that *neutral* problems typically elicit the response of “both options are equally likely” (which is considered a heuristic response). The reason that participants ignore base rates even when they are not provided with any other useful information is, presumably, that they try to base their response on the description (which they automatically assume to be relevant, although it is not). Thus, providing an irrelevant description is enough to draw participants’ attention from the base rates.

De Neys and Glumicic (2008) stated that even people who eventually give a description-based response show signs of conflict detection, although they are not consciously aware of this. They demonstrated that whereas in verbal protocols there was no mention of experiencing a conflict, less explicit measures showed signs of differential processing of base rates in conflict and non-conflict problems.

The purpose of the present study was to investigate further how people process base rate-information in the presence of base rate-congruent, incongruent, and neutral descriptions. In the experiment that we report below we used the problems developed by De Neys and Glumicic (2008) which we slightly modified to make them more appropriate for UK participants. However, instead of asking participants to generate a response, we provided them with a response (which we called a statement), and we asked them

Table 1: Different versions of the engineers and lawyers problem (based on De Neys & Glumicic, 2008).

<i>Conflict: Incongruent description and base rates</i>	
<p><b>(Part 1:)</b> In a study 1000 people were tested. Among the participants there were 5 engineers and 995 lawyers. Jack is a randomly chosen participant of this study.</p> <p><b>(Part 2:)</b> Jack is 36 years old. He is not married and is somewhat introverted. He likes to spend his free time reading science fiction and writing computer programs.</p>	
Statement (heuristic): <b>Jack is an engineer.</b>	Statement (non-heuristic): <b>Jack is a lawyer.</b>
    	
Don't like it at all.	Don't like it. Don't know. Like it. Like it very much.
<i>Non-conflict: Congruent description and base rates</i>	
<p><b>(Part 1:)</b> In a study 1000 people were tested. Among the participants there were 995 sixteen-year olds and 5 fifty-year olds. Ellen is a randomly chosen participant of this study.</p> <p><b>(Part 2:)</b> Ellen likes to listen to hip hop and rap music. She enjoys wearing tight shirts and jeans. She's fond of dancing and has a small nose piercing.</p>	
Statement (heuristic): <b>Ellen is sixteen.</b>	Statement (non-heuristic): <b>It is equally likely that Ellen is sixteen or that she is fifty.</b>
<i>Neutral: Base rates plus neutral description</i>	
<p><b>(Part 1:)</b> In a study 1000 people were tested. Among the participants there were 4 who live in Manchester and 996 who live in Liverpool. Chris is a randomly chosen participant of this study.</p> <p><b>(Part 2:)</b> Chris is 28 years old. He has a girlfriend and shares an apartment with a friend. He likes watching basketball.</p>	
Statement (heuristic): <b>It is equally likely that Chris lives in Liverpool or that he lives in Manchester.</b>	Statement (non-heuristic): <b>Chris lives in Liverpool.</b>

to evaluate the statement, using a 5-point rating scale of smiley faces, ranging from “don’t like it at all”=1 to “like it very much”=5 (see Table 1). This procedure was modelled on Topolinski and Strack (2009a). Liking ratings are sensitive to both conscious and unconscious influences (e.g., explicit preferences, affective priming, etc.), and they are ideal for detecting subtle changes in participants’ judgments (cf., Morsanyi & Handley, 2012). Thus, even if participants are unaware of being influenced by base rates / descriptions, these influences should be reflected in their liking ratings. Moreover, liking ratings convey more information than response choices. For example, it is possible that although a participant shows a strong preference for a certain response option, they also evaluate other options positively.

In order to explore the role of Type 1 and 2 processes in people’s judgments and in utilizing base rates and descriptions, we implemented an instruction manipulation (see e.g., Klaczynski, 2001). Half of the participants were asked to rely on their intuitions, whereas the rest of the participants were instructed to think logically. From a dual-process perspective, intuitive instructions should encourage Type 1 processing, whereas logical instructions should increase the influence of Type 2 processes. Indeed, previous research (e.g., Chiesi, Primi & Morsanyi 2011; Ferreira,

Garcia- Marques, Sherman & Sherman, 2006) showed that instructions affected participants’ susceptibility to reasoning biases. Thus, we expected that whereas in the intuitive condition participants would be strongly affected by descriptions, logical instructions should increase the tendency to rely on base rate information. Nevertheless, this should only happen if participants are consciously aware of the conflict, and if they judge that base rates are more relevant to making sound judgments than descriptions.

Another question that we wanted to investigate was whether heuristic responses are associated with higher liking ratings than non-heuristic responses. In a recent paper (Thompson & Morsanyi, 2012) we proposed that heuristic responses might be hard to resist, because of the positive affective valence that they carry. Specifically, we suggested that as heuristic responses are generated fluently and effortlessly, and fluent processing is associated with positive affect (see Topolinski & Reber, 2010 for a review), participants will prefer a heuristic mode of processing, because heuristic responses “feel right”. In order to test the assumption that liking ratings are closely related to participants’ actual choices, we presented the problems with an option, and they were asked to rate it according to how much they liked it. After performing a different task,

participants were (unexpectedly) presented with the problems once more, but this time they had to select a response from three options. Given the associations between heuristic (i.e., Type 1) processing, positive affect and confidence, we expected that initial liking ratings would be good predictors of subsequent choices.

In summary, the aim of the present study was to better understand the processes underlying performance on engineers and lawyers-type problems. Given the high rate of heuristic responses, we were particularly interested in whether participants experienced a conflict while solving the tasks. We employed a novel paradigm (i.e., asking for liking ratings) to investigate the effect of base rates and descriptions, and we combined it with an instruction manipulation, in order to explore the role of Type 1 and 2 processes in participants' judgments. Additionally, we wanted to test the assumption that heuristic responses are associated with positive affect. Finally, we also predicted that the affective valence of choice options would be closely related to how likely individuals will be to opt for a response when they are offered a choice between different responses.

## Methods

### Participants

The participants were 62 students (54 females, mean age 21 years 2 months) from the University of Plymouth, UK who participated in the study for ungraded course credit. Participants were randomly allocated either to the intuitive ( $n=32$ ) or to the logical ( $n=30$ ) instruction condition.

### Materials

The participants were presented with 12 problems: 4 *conflict* problems (i.e., where the description of the person was incongruent with base rates), 4 *non-conflict* problems (where descriptions and base rates were congruent), and 4 *neutral* problems (with irrelevant descriptions). The problems were presented in two parts (using a "moving window" procedure – see de Neys & Glumicic, 2008, Experiment 2). The base rates were presented first (together with the information that the person was randomly selected from a large sample – marked as Part 1 in Table 1), then participants had to press the space bar, and this information disappeared, and the description of the person (Part 2) appeared together with the statement about the person and the rating scale for liking ratings. Participants could review base rate information by pressing a radio button on the computer screen. The problems were presented in a random order, which was different for each participant. The statement that participants had to rate either corresponded to the base rates or to the description (or both), or it simply said that the person was equally likely to belong to either category (see Table 1 for examples). In order to reduce content effects, we created two task sets, where for the same problem participants were either offered a heuristic (i.e., description-based), or a non-heuristic response (see Table 1

for illustrations). Finally, in the second part of the experiment, participants were presented with the same problems again, using the same presentation format as in the first part. However, instead of providing liking ratings for one response option, participants had to choose from three responses (i.e., 1. the person belonged to one category – e.g., engineers; 2. the person belonged to the other category – e.g., lawyers; or 3. it was equally likely that the person belonged to either one or to the other category).

### Procedure

Participants solved the problems on the computer. First they were presented with instructions, and they were informed that they could review the first part of the problem. Additionally, in the intuitive condition participants were told: "*When you make your liking ratings, rely on your intuition and feelings. Give the first rating that comes to mind, without any conscious reflection, and do this as quickly as you can.*" In the logical condition the instructions ended like this: "*When you make your liking ratings, take the point of view of a perfectly logical person. Think about your answer very carefully. Don't rush. You can take as much time as you want.*" Subsequently, the participants were presented with a practice problem, and then they had to work through the 12 experimental problems. After this, they had to perform a different (unrelated) task for about 5 minutes. Finally, they were presented with the problems again. This time they had to choose from three options, rather than evaluating a response which was offered to them. In the second part of the experiment participants were instructed to consider the problems carefully, but they were not explicitly asked to reason intuitively or logically. This part also started with a practice problem.

### Results

First, as a manipulation check, we compared the average time that participants spent solving each problem across the intuitive and logical conditions (collapsed across all tasks). As expected, participants in the intuitive condition responded more quickly ( $M=17820$  ms,  $SD=4235$  ms) than participants in the logical condition ( $M=23872$  ms,  $SD=7223$  ms;  $t(60)=4.06$ ,  $p<.001$ ).

We also wanted to see whether we could replicate the pattern reported by De Neys and Glumicic (2008) regarding participants' inspection of base rates. Specifically, these authors reported that participants were more likely to opt for reviewing base rate information if the base rates were in conflict with the description of the person, as opposed to when there was no such conflict. In our analyses we included not only conflict and non-conflict problems, but also problems with neutral descriptions, in order to see whether problems with base rate-incongruent and neutral descriptions are processed differently. Finally, we were also interested in whether the tendency to review base rates differed across the two instruction conditions.

Participants in the intuitive condition reviewed on average 11% ( $SD=.22$ ) of the base rates in the case of

Table 2: Participants' liking ratings across the different types of task, and different statements.

	Incongruent		Congruent		Neutral	
	<i>Heuristic</i>	<i>Base rate</i>	<i>Heuristic/ base rate</i>	<i>Equally likely</i>	<i>Heuristic/ equally likely</i>	<i>Base rate</i>
intuitive	3.19 (.74)	2.66 (.76)	3.63 (.61)	3.14 (.95)	3.16 (.83)	3.27 (.84)
logical	3.47 (.82)	2.42 (.97)	3.60 (.93)	3.32 (.92)	3.68 (1.03)	3.22 (.85)

conflict, 8% ( $SD=.21$ ) in the case of non-conflict, and 9% ( $SD=.21$ ) in the case of neutral problems. The corresponding numbers in the logical group were 23% ( $SD=.24$ ), 14% ( $SD=.18$ ), and 28% ( $SD=.29$ ), respectively. A 3x2 mixed ANOVA with problem type (conflict/non-conflict/neutral) as a within-subjects factor and condition (intuitive/logical) as a between-subjects factor indicated a significant effect of problem type ( $F(2, 120)= 4.04, p=.020, \eta_p^2=.06$ ), and a significant effect of condition ( $F(1, 60)= 6.77, p=.012, \eta_p^2=.10$ ). The problem type by condition interaction was not significant ( $p=.114$ ). That is, in general participants in the logical condition were more inclined to review base rates. Follow-up analyses also showed that participants were more likely to review base rate information if descriptions were not in line with base rates, regardless of whether descriptions were conflicting or neutral. Indeed, the tendency to review base rates did not differ between conflict and neutral problems.

Next we analyzed participants' liking ratings (see Table 2). In order to gather further support for the claim that participants were sensitive to the conflict between the descriptions and base rates, we compared their liking ratings for description-based (i.e., heuristic) responses across problems where base rates and descriptions were congruent (non-conflict), and where these were incongruent (conflict). A 2x2 mixed ANOVA with condition (intuitive/logical) as a between-subjects, and problem type (congruent / incongruent) as a within-subjects factor indicated a significant effect of problem type ( $F(1, 60)= 5.95, p=.018, \eta_p^2=.09$ ). The effect of condition, and the condition by problem type interaction were not significant. That is, participants, regardless of condition, liked description-based responses more if these were not in conflict with base rates.

Another issue that we were interested in was whether participants' liking ratings were higher for heuristic responses than for non-heuristic responses. To investigate this question, we first collapsed ratings across *conflict* and *neutral* problems. As we described in the introduction, in the case of both types of task there is a general tendency for participants to disregard base rates. This is assumed to be the consequence of an automatic (i.e., heuristic) tendency to generate responses that correspond to (or take into account) the descriptions (cf. Kahneman & Frederick, 2005).

The average ratings for heuristic and base rate responses in the intuitive condition were  $M=3.17$  ( $SD=.51$ ) and  $M=2.96$  ( $SD=.63$ ), respectively. The corresponding ratings in the logical condition were  $M=3.58$  ( $SD=.60$ ) for heuristic,

and  $M=2.82$  ( $SD=.78$ ) for base rate responses. In line with our predictions, a 2x2 mixed ANOVA with response type (heuristic/base rate) as a within-subjects and condition (intuitive/logical) as a between-subjects factor indicated that participants liked heuristic responses more than base rate responses ( $F(1, 60)= 19.80, p<.001, \eta_p^2=.25$ ). Additionally, there was a significant interaction between response type and condition ( $F(1, 60)= 6.32, p=.015, \eta_p^2=.10$ ). Interestingly, this interaction showed that there was a greater difference between ratings for heuristic and base rate responses in the case of participants in the logical as compared to the intuitive condition. That is, participants who invested more time and effort into providing their liking ratings were more biased by the descriptions. Nevertheless, participants in both conditions provided higher liking ratings for responses which are supposed to be based on heuristic (i.e., Type 1) processing than for non-heuristic (i.e., Type 2) responses.

Finally, we wanted to investigate how closely the liking ratings were related to participants' actual response choices in the second part of the experiment (this analysis was conducted at the level of tasks; see Figure 1). The correlation between liking ratings and the probability that a participant selected a given response was significant both in the intuitive ( $r(384)=.20, p<.001$ ) and in the logical condition ( $r(360)=.38, p<.001$ ), and the association was significantly stronger in the logical condition, as indicated by a Fisher  $r$ -to- $z$  transformation ( $z=2.67, p=.008$ ).

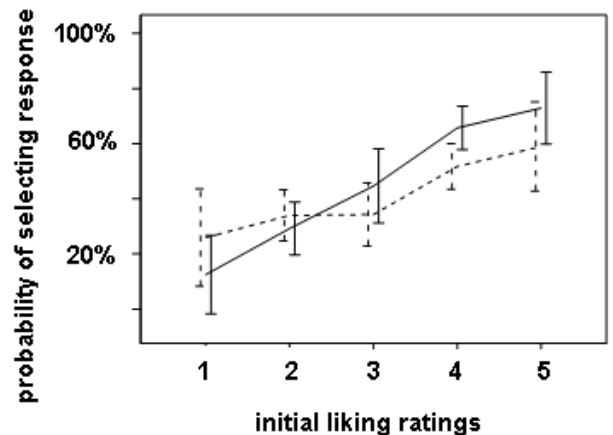


Figure 1: The probability of selecting a response as a function of liking ratings in the two conditions (broken line: intuitive condition).

## Discussion

In the present study we employed a new method (asking for liking ratings) to investigate the cognitive underpinnings of performance on engineers and lawyers-type problems. Liking ratings can be used to investigate both conscious and unconscious preferences, and they are very sensitive to subtle changes in participants' judgments.

Arguably, the most interesting finding is that although participants' probability judgments were biased by the person's description both in the intuitive and in the logical condition, this bias was stronger in the logical group. This is in contrast with earlier studies which generally reported a decrease in biases as a result of logical instructions (e.g., Chiesi et al., 2011; Ferreira et al., 2006). This finding is also in contrast with the assumption that in heuristics and biases tasks people automatically generate an initial heuristic response, which is either accepted without modification, or it is suppressed by conscious and effortful reasoning (e.g., Evans, 2006). Instead, it seems that, at least in the case of the engineers and lawyers problem, although most people show an initial (weak) preference for heuristic responses, this preference becomes significantly stronger when they invest more time and effort in the evaluation of the response options.

Indeed, a similar pattern has been observed in the case of the Wason selection task, using eye tracking methods (Ball, Lucas, Miles & Gale, 2003). In the selection task people tend to focus on a response (usually the one, which is considered the intuitive response) almost immediately after they are presented with the options, but then they spend a longer period considering this response before making their eventual choice. This pattern has been cited as evidence that although Type 2 processes are employed in the selection task, they are merely used to rationalize an initial, Type 1 response (Evans, 2006). Nevertheless, there is evidence that people do engage in a conscious reasoning process when they make their choices in the selection task, although this does not necessarily result in finding the normative solution (cf., Handley, Newstead & Neilens, 2011). Indeed, spending more time on evaluating a compelling response option might increase reasoners' confidence in the correctness of the response (see e.g., Thompson & Morsanyi, 2012).

The increased bias in the logical condition suggests that conscious reasoning processes play an active role in the engineers and lawyers problem. Instead of just approving intuitive response tendencies, they magnify the initial bias. Indeed, this process might involve the active rejection of base rates as a potential basis of judgment. In fact, participants not only rated the responses which corresponded to base rates lower than the responses which were in line with the descriptions, but their ratings for base rate responses were also slightly negative.

With regard to conflict detection, our results support earlier findings (e.g., de Neys & Glumicic, 2008) which suggested that participants experience a conflict when base rates and the description of the person cue different responses. However, we should note that providing a neutral

description resulted in similar levels of base rate-inspection as providing base rate-incongruent descriptions. Thus, base rate inspection could be taken as a sign of uncertainty or decreased processing fluency, rather than of "conflict detection". As we found no evidence for a difference between the intuitive and logical groups in "conflict detection" (as indexed by reviewing base rates, and the difference between liking ratings for description-based responses in conflict and non-conflict problems), it remains unclear if participants are conscious of the conflict. We could expect that offering a response which corresponds to the base rates in a conflict task makes base rate information more salient. Nevertheless, participants' evaluations of these options were slightly negative, which indicates that even if they were aware of the potential significance of base rates (and the conflict between base rates and descriptions), they still preferred to base their judgments on the descriptions. Thus, it is unlikely that the reason that only a small minority of participants give normative responses to the engineers and lawyers problem is that participants only detect the conflict unconsciously, and, as a result, their conscious responses remain unaffected by this.

As expected (cf., Thompson & Morsanyi, 2012), heuristic responses were liked more than probability-based responses. This corresponds to the general pattern that most participants select or generate a heuristic response when they are presented with the engineers and lawyers problem (see e.g., De Neys & Glumicic, 2008; Kahneman & Tversky, 1973). This pattern is also in line with the idea that heuristic processing is associated with positive affect, and this affective component might contribute to participants' tendency to accept these responses. Indeed, initial liking ratings were significantly related to participants' response choices. The finding that this relationship was stronger in the logical group suggests that these participants indeed considered the options more carefully than participants in the intuitive condition (given that response choices in the second part of the study were based on careful consideration).

Although affective reactions might contribute to both liking ratings and response choices, it is also possible that the liking ratings were unrelated to participants' affective states, and participants simply indicated with their liking ratings the extent to which they found a particular response correct or appropriate. Other studies (e.g., Topolinski & Strack, 2009b; Morsanyi & Handley, 2012) demonstrated through effective priming and emotion-misattribution manipulations that liking ratings are sensitive to participants' affective states. Nevertheless, future studies should seek to provide more direct evidence for the link between affect, liking ratings, and heuristic responses. One method which seems particularly suitable would be to measure the activation of facial muscles which are associated with smiling and frowning, using electromyography (see Topolinski, Likowski, Weyers, & Strack, 2009), while participants evaluate heuristic and non-heuristic response options.

In summary, our findings provide new insight into the cognitive processes involved in the engineers and lawyers problem. Most importantly, these results indicate that conscious thinking might contribute to the biases often observed in judgment and reasoning. Indeed, there is a growing body of evidence to indicate that responses which are assumed to be based on heuristic or automatic (i.e., Type 1) processing often require cognitive effort. Generating these responses might even be more effortful than producing other responses, which traditional dual-process approaches associated with effortful, Type 2 processing (see e.g., Handley, Newstead & Trippas, 2011; Morsanyi & Handley, 2008). These findings, together with criticism which is based on more theoretical considerations (e.g., Keren & Schul, 2009; Osman & Stavy, 2006), pose a challenge to dual-process theories of reasoning.

### Acknowledgments

The writing of this paper was supported by an ESRC Post-Doctoral Fellowship (ES/1038071/1) to K.M.

### References

- Ball, L. J., Lucas, E. J., Miles, J. N. V., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *Quarterly Journal of Experimental Psychology*, 56A, 1053-1077.
- Chiesi, F., Primi, C. & Morsanyi, K. (2011). Developmental changes in probabilistic reasoning: The role of cognitive capacity, instructions, thinking styles and relevant knowledge. *Thinking & Reasoning*, 17, 315-350.
- De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, e15954. doi:10.1371/journal.pone.0015954.
- De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of reasoning. *Cognition*, 106, 1248-1299.
- Evans, J. ST. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation, *Psychonomic Bulletin & Review*, 13, 378-395.
- Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.
- Ferreira, M. B., Garcia-Marques, L., Sherman, S. J., & Sherman, J. W. (2006). Automatic and controlled components of judgment and decision making. *Journal of Personality and Social Psychology*, 91, 797-813.
- Handley, S. J., Newstead, S. E. & Neilens, H. (2011). Matching bias requires analytic reasoning. In: Manktelow, D. Over & S. Elqayam (Eds.) *The Science of Reason: A Festschrift for Jonathan St.B.T. Evans*. (pp. 167-189). Psychology Press.
- Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 28-43.
- Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, 12, 275-305.
- Kahneman, D. & Frederick, S. (2005). A model of heuristic judgment. in K.J. Holyoak & R.G. Morrison [eds.] *The Cambridge Handbook of Thinking and Reasoning*. (pp. 267-293). Cambridge University Press.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237-251.
- Keren, G., & Schul, Y. (2009) Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, 4, 533-550.
- Klaczynski, P. A. (2001). Framing effects on adolescent task representations, analytic and heuristic processing, and decision making: Implications for the normative-descriptive gap. *Journal of Applied Developmental Psychology*, 22, 289-309.
- Morsanyi, K., & Handley, S. J. (2008). How smart do you need to be to get it wrong? The role of cognitive capacity in the development of heuristic-based judgment. *Journal of Experimental Child Psychology*, 99, 18-36.
- Morsanyi, K., & Handley, S. J. (2012). Logic feels so good -I like it! Evidence for intuitive detection of logicity in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory and Cognition*. (in press)
- Osman, M, Stavy, R (2006). Development of intuitive rules: evaluating the application of the dual-system framework to understanding children's intuitive reasoning. *Psychonomic Bulletin & Review*, 13, 935-953.
- Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.
- Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, 23, 645-726.
- Thompson, V.A. & Morsanyi, K. (2012). Analytic thinking: Do you feel like it? *Mind & Society*. (in press)
- Topolinski, S., Likowski, K. U., Weyers, P., & Strack, F. (2009). The face of fluency: Semantic coherence automatically elicits a specific pattern of facial muscle reactions. *Cognition & Emotion*, 23, 260-271.
- Topolinski, S., & Strack, F. (2009a). Scanning the "fringe" of consciousness: What is felt and what is not felt in intuitions about semantic coherence. *Consciousness & Cognition*, 18, 608-618.
- Topolinski, S., & Strack, F. (2009b). The architecture of intuition: Fluency and affect determine intuitive judgments of semantic and visual coherence and judgments of grammaticality in artificial grammar learning. *Journal of Experimental Psychology: General*, 138, 39-63.
- Topolinski, S., & Reber, R. (2010). Gaining insight into the "aha" experience. *Current Directions in Psychological Science*, 19, 402-405.



# Inferring aspectuality on French sentences: a minimalist approach

Damien Munch (Munch@Telecom-Paristech.fr) and  
Jean-Louis Dessalles (Jean-Louis.Dessalles@Telecom-Paristech.fr)  
INFRES, Institut Telecom ParisTech, 46 rue Barrault  
75013 Paris FRANCE

## Abstract

Current models of temporality in language are either inaccurate or too complex to be cognitively plausible. We present a cognitive model of the computation of aspect in French. Our approach emphasizes the importance of minimalism for cognitive plausibility: structures and computation are kept simple and combinatorial explosion is avoided. Though the model and its current implementation remain partial for now, our approach opens the way to a generic and cognitively plausible method for the determination of aspect.

**Keywords:** Cognitive minimalism; Natural Language Processing; Temporal aspect; Temporal reasoning

## Introduction

Humans are experts in the communication of temporal information. The coherence of discourse relies on the correct expression of time and aspect, both in narratives (e.g. to mark causality) and in argumentative discussions (think of an alibi). Though significant progress has been achieved in modeling temporal processing, current models are either inaccurate or too complex to be cognitively plausible. In the present paper, we stick to the idea that a plausible model should rely on a minimum number of principles. The paper presents new elements in that direction.

Linguists have established various categorizations of aspect, tense and modality (Vendler, 1967; Comrie, 1976; Veters, 1996, among others). They explain variations of meaning by postulating the existence of rich semantic structure stored in lexical entries. For example, Comrie (1976), based on (Vendler, 1967), associates binary attributes such as achievement, accomplishment, semelfactive or activity to verbs. The challenge is to infer aspect, such as repetition and perfectivity, and to predict semantic incorrectness from the combination of attributes when processing a sentence. One problem is to limit the number of attributes that lexical entries may instantiate in their structure. Another problem is to show that the chosen lexical attributes are sufficient.

In addition to fixed attributes attached to the lexical entries, some logicians and computer scientists introduced a *procedural* component in their models of temporal interpretation. To process *tense*, Reichenbach (1947) introduced a minimalist model based on three dynamical coordinates: Event, Reference and Speech. Despite its impressive description power, this model does not account for tense in nested clauses (Hwang & Schubert, 1992) and it fails to explain the behavior of some tenses in other

languages than English (Dowty, 1979; Comrie, 1985). Since then, Reichenbach's model has been steadily improved. The three coordinates have been changed for intervals and/or have been increased in number (Comrie, 1985; Gosselin, 1996; Elson & McKeown, 2010).

There have been attempts to process *aspect* in a minimalist way as well. Recent TimeML versions (Sauri & al., 2009) consider four attributes: Progressive, Perfective, Perfective\_progressive and None. Smith (1991) proposes a model based on only three viewpoints: Imperfective, Perfective, and Neutral. Ghadakpour (2004) uses only two viewpoints, called Figures and Grounds.

Lexical models, in which temporal knowledge is stored in static lexical attributes, face the problem of *attribute defeasibility*. For instance, the verb "to hit" is supposed to have the Punctual attribute; therefore, "she hit the wall for one minute before leaving" receives a repetitive interpretation (several knocks); however, "The small galaxy hit (collided with) the Milky Way for ten million years before collapsing" can receive a non-repetitive interpretation, in contradiction with the supposed Punctual attribute of the verb.

Procedural models, in which lexical entries are given computation power, are able to deal with context. For instance, in Gosselin's (1996) and Schilder's (2004) models, the function assigned to *en* (in French) or *in* (in English) checks whether the complement of the preposition involves duration. Schilder's model even checks whether the phenomenon happened in the past or not. Procedural models, however, are not parsimonious as long as they do not set limits to the computational power of words. For instance Person's (2004) implementation of Gosselin's model of French temporality associates a specific computing rule to each tense and each temporal marker (preposition, temporal adverbs, ...). Similar procedural approaches, in which temporal lexemes are given significant computation power, are proposed by others authors like Saussure (2003) and Schilder (2004). Though these models try to remain parsimonious *in fact*, they are not parsimonious *in principle*. Models in which words may have unlimited power (*i.e.* they may perform any computation like Turing machines) do not qualify as cognitive models, not only because they lack parsimony but also because they cannot explain how children learn the mechanisms of temporality of the surrounding language.

Models of temporality face another problem. The temporal meaning of a sentence cannot be deduced only from temporal information stored in lexemes. Moens and Steedman (1988) have shown that the mental

representations corresponding to events are not reducible to tense and aspect. They are closer to concepts such as causal sequences, preparatory processes, goals and consequent states. According to these authors, temporal attributes stored in the lexicon cannot capture the richness of interpretation that is accessible to humans. Temporal interpretation would involve causal relationships that lie beyond strict linguistic processing. Models such as Event Calculus (Shanahan, 1999), modeled by Mueller (2004), do take background knowledge into account. The problem, for such models, is to circumscribe the effect of context, not only to avoid unrealistic processing time, but also to keep the systematic character of some temporal phenomena.

Our aim is to design a cognitively plausible model of temporality that avoids the previously mentioned difficulties (attribute defeasibility, unlimited computational power and unlearnability, prohibitive processing time, loss of systematicity). We favor a minimalist approach, in which both structures and procedures are kept minimal. In what follows, we will first list a limited set of examples in French that we use as benchmark. Then we will see how Gosselin's and Schilder's models behave on such examples. We chose these two models because they use concepts similar to ours, such as *viewpoint*, *anchoring* and *granularity*. We will then describe our model and its single procedure: *tMerge*. We conclude with a discussion in which we assess the plausibility and the generality of our approach.

### Temporal correctness

Table 1 shows a few examples that have been tested for acceptability. We asked thirty-five French native speakers to answer multiple choices questions about forty-one sentences in French. These sentences were designed to test all combinations of tense, event type and time adverbials. They were proposed in random order and participants were allowed to stop whenever they want. We got an average of twenty answers by sentence. All sentences have the same form: verb phrase + prepositional phrase. Participants were asked the following questions:

- Is the sentence correct / incorrect ("one wouldn't say that") ?
- Does the event occur several times (repetition) / only once (possibly with breaks) ?
- Is the event finished / not finished / don't now ?
- Is the event taking place during the whole period indicated by the prepositional phrase / during only part of it / don't now ?

Participants could provide two sets of answers for a given sentence if they thought there were two meanings.

Table 1 shows some of the results. The binary values *unique/multiple*, *perfective/imperfective* and *whole/slice* refer to coded answers to questions b-d. Percentages for a given sentence correspond to participants who found the sentence correct. They may add up to more than 100% when several interpretations were given.

Table 1: Tested sentences

- |  |
|--|
| <p>(1) Elle mangera du gâteau en février.<br/> <i>She will eat (be eating) cake in February.</i><br/> (30%) <b>unique/perfective</b> and <b>slice</b> of February<br/> (80%) <b>multiple/imperfective</b></p> <p>(2) * Elle mangera du gâteau en 30 minutes.<br/> * <i>She will eat (be eating) cake (or from the cake) within 30 minutes.</i><br/> (30%) <b>unique/perfective</b></p> <p>(3) Elle atteindra le sommet en février.<br/> <i>She will reach the top in February.</i><br/> (76%) <b>unique/perfective</b> and <b>slice</b> of February</p> <p>(4) Elle mangera (à la cantine) pendant deux mois.<br/> <i>She will be eating (at the canteen) for two months.</i><br/> (100%) <b>multiple/imperfective</b></p> |
|--|

Table 2 shows classical sentences (examples 5-8) that were not included in the test.

Table 2: Sentences variations

- |  |
|--|
| <p>(5) Elle atteindra le sommet en 30 minutes.<br/> <i>She will reach the top within (the next) 30 minutes.</i><br/> <b>unique/perfective</b></p> <p>(6) * Elle atteindra le sommet pendant 30 minutes.<br/> * <i>She will be reaching the top for 30 minutes.</i></p> <p>(7) Elle atteindra le sommet pendant le prochain mois.<br/> <i>She will reach the top during the next month.</i><br/> <b>unique/perfective</b> and <b>slice</b> of the next month</p> <p>(8) Elle mangera du gâteau pendant les 30 prochaines minutes.<br/> <i>She will eat (be eating) cake during the next 30 minutes.</i><br/> <b>unique/perfective</b></p> |
|--|

The challenge is to account not only for the acceptability or incorrectness of sentences, but also for the judgements about repetition, perfectiveness and wholeness. The next section examines how Gosselin's and Schilder's models perform on this kind of examples.

### The computation of aspect

Gosselin's (1996) model represents perfectivity by considering intervals with two different types of boundaries: *intrinsic* and *extrinsic*. These boundaries are retrieved from the aspectual type of events (telicity, punctuality, dynamicity). For instance "manger du gâteau" ("to eat cake") will take extrinsic boundaries because its aspectual type is supposed to be *semelfactive* (this information is provided by some external cognitive processing).

Repetition appears during conflict resolution, when the granularities of two intervals are different. For instance in example (1), “manger du gâteau” and “février” (February) do not have the same granularity; the conflict is solved by iterating the interval of “manger du gâteau”.

Conflict resolution also involves instructions which can move one or both boundaries of an interval. This will lead to shrinking, expanding or moving one of the conflicting intervals. Slices in our examples would result from shrinking the interval of the adverbial phrases (“février”, “le prochain mois”).

In example (2) (“manger du gâteau en 30 minutes”), “manger du gâteau” is represented by an interval [B1,B2] with *extrinsic* boundaries, whereas “30 minutes” is represented by an interval [ct1,ct2] with *intrinsic* boundaries. Step b (Figure 1) succeeds, but step c fails because the two intervals have incompatible boundaries types.

Though Gosselin’s model seems to work fine, it is at the expense of simplicity. Specific instructions are assigned to ‘operators’, that is, to every lexeme with a temporal meaning, such as tenses and temporal prepositions. Figure 1 shows the instructions associated to the preposition “en” + *duration*. The problem is not only the actual complexity of such instructions, but also the fact that this complexity is not bound in principle.

- |   |
|---|
| <p>a) associate an interval [ct1,ct2] to the temporal adverbial</p> <p>b) <math>ct1 &lt; ct2</math> (non-ponctual adverbial, boundaries are dissociated)</p> <p>c) [ct1, ct2] CO [B1,B2] (adverbial coincides with the event)</p> <p>d) [I,II] ACCESS [B1,B2] (boundaries of the event must be ‘accessible’ by the reference interval ; <math>I \leq B1</math> and <math>II \geq B2</math>)</p> <p>The interval of the event [B1,B2] must be intrinsic (when “pendant” + duration need extrinsic boundaries).</p> |
|---|

Figure 1: instructions for “en” + *duration*, adapted from Gosselin (1996)

Schilder (2004) uses neither intervals nor boundaries in his model. Events are given one of the four aspectual values defined in TimeML (Sauri & al., 2009): Perfective, Progressive, Perfective\_progressive and None.

Schilder’s model can detect granularity incompatibilities, though it is not clear whether they are solved by operations like slice or repetition.

To deal with the examples of Tables 1-2, he proposes two different functions for each temporal preposition, depending on whether the complement is anchored or not. Figure 2 shows instructions for “in” (note that this function applies to the English or German “in”). In example (5), the event “reach the top” occurs at ‘timestamp’ TS, which is the ‘Document timestamp’ (DTS) *plus* the given duration (DUR) “30 minutes”. The granularity (G) of the event is given by

the document timestamp (Figure 2). Note that this computation does not seem to be always valid in English (for instance, in “She will defeat her opponent in 30 minutes”, meaning “She will play during 30 minutes and win”, the duration of the event should be DUR and not DTS).

<p><i>Anchored</i>: TSDUR</p> <p><i>Unanchored</i>: If Tense = Past then DUR else TSP1G where <math>TS = DTS + DUR</math> and <math>G = \text{gran}(DTS)</math></p>
---

Figure 2: function used by “in”, adapted from Schilder (2004)

Contrary to Gosselin, Schilder chooses to assign functions to all lexemes, not only the ‘temporal’ ones. On the other hand, processing is somewhat simpler, as it treats prepositions as unary instead of binary operators, as proposed by Pratt & Francez (2001). However, Schilder model has the same drawback as Gosselin’s: *each lexeme* is given a dedicated function. As long as there is no indication about how to limit the computational complexity of these functions, models cannot be considered minimalist.

## A minimalist model

In this paper, we present a model which is minimal in terms of structures, procedure and memory use. We rely on one single fixed-sized semantic structure, called *temporal Semantic Structure* (tSS), and one single non-recursive operation, called *temporal merge* (tMerge). Note that the use of the term *merge* (related to Chomsky, 1995) instead of *unification* is debatable (Jackendoff, 2005). To achieve this reduction, we decided to exclude several operations from the temporal processing proper, in line with (Moens & Steedman, 1988), considering that they required access to other cognitive modules (general knowledge and perception abilities, syntax, determination). These operations include:

- time location (whether a situation is located in time or not)
- temporal granularity (or order of magnitude) consistency checking
- causality and anteriority checking
- self-similarity checking
- phrase syntactic hierarchy

For our purposes, these operations need not be represented in a cognitively realistic way, as they are considered external to the model. We implemented them in a basic form in our perception module. We now describe the two central components of the model, tSS and tMerge.

## The Temporal Semantic Structures (tSS)

The tSS is the only structure processed in the model. A tSS is a non-recursive structure. It contains three attributes: an

*image reference* (ImageID) and two *switches* (*Viewpoint*, *Anchoring*).<sup>1</sup> These attributes may be uninstantiated at a given step of the processing.

**ImageID** The image identifier is a reference to a perceptive representation. The term 'image' includes any perceptive representation, either stored or constructed. The mere use of *ImageID* in the tSS allows us to grossly simplify several processes which are supposed to be performed in an external module. That module (called '*Perception*') is loosely defined as including anything which does not pertain to syntactic processing or to temporal processing proper (as defined in our model). This includes all forms of memory and all forms of knowledge, such as the granularity or the date of events. It also includes the ability to decide about anteriority and about self-similarity.

**Viewpoint** The *Viewpoint* switch may take two values, *figure* (f) and *ground* (g) (Ghadakpour, 2003). It may be defined in the lexical entry (for example, the French 'imparfait' (imperfect tense) requires a *ground*). However, it is most often determined during temporal processing. Viewpoints are a key element of our model. They provide information about how the speaker regards the temporal phenomenon at a given stage of processing: either 'from the outside' and considering the overall event (*figure*), or 'from within' (*ground*) (exactly as for space). These two values correspond to standard aspect values: perfective and imperfective. However, we consider the viewpoint as a *switch* that may change value during processing.

**Anchoring** The anchoring *switch* indicates whether *Perception* is able to provide some (absolute or relative) location for the perceptive image. The Anchoring *switch* may take two values: *anchored* (a) and *unanchored* (u). For example "(any) 30 minutes" will be unanchored, but "these 30 minutes" will be anchored. Some lexemes are ambiguous in this respect: "in February" may mean that we deal with a specific (anchored) or a recurring (unanchored) period. Anchoring bears a close relation to determination. "The concert" is likely to designate an anchored time period, whereas "a concert" refers to an unanchored time period.

### The temporal merge operator (tMerge)

The tMerge procedure (Figure 3) is launched whenever a phrase is recognized by the syntactic analysis. In other words, the syntactic and semantic analyses are fully synchronous. tMerge receives two tSS as input, plus the indication that one (the head) dominates the other (the complement). It returns a unique tSS as result. These elements appear as H, C and R respectively in Figure 3 (tSS are indicated in square brackets). Even if we deal here with temporal merge exclusively, we assume that the semantic merge operation performed in several other specialized modules (spatial relationships, determination, perception).

(1) The essential part of tMerge consists in a *basic merge* operation (bottom of figure 3): corresponding *switches* in the two input tSS are merely matched for compatibility to produce R.

(2) When basic unification succeeds, unification proceeds to the Perception module (see figure 3) where it generates a new image (this process, omitted from our model, merely concatenates images identifiers). The perceptive merge may apply *viewpoint* constraints to the resulting tSS depending of the nature of the phenomenon: *indivisible* events are bound to be *figures*, whereas *self-similar* events must be

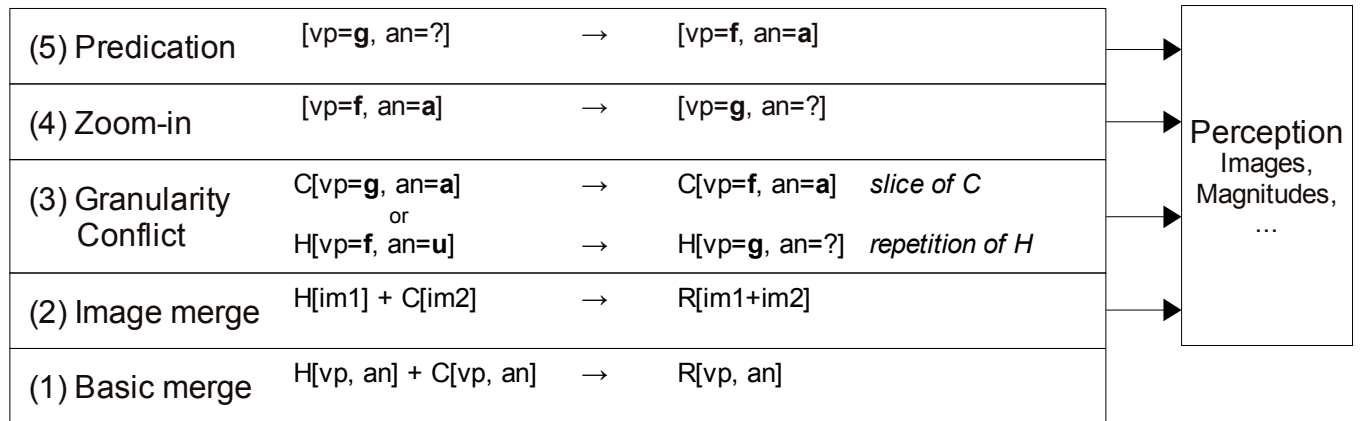


Fig 3: tMerge steps

<sup>1</sup> For the sake of simplicity, we omit *switches* related to tense.

*grounds*. The perceptive merge also checks for granularity compatibility. All the other operations of figure 3 aim at rescuing basic merge and perceptive merge in case of failure.

(3) The *Granularity conflict* rescue operation is triggered in examples like: “She will eat the cake in February”, where the orders of magnitude are hour *vs.* month. Depending on viewpoint and anchoring constraints, the complement element will be *sliced* (she will eat cake once at some point in February) or the head will be *repeated* (she will be eating cake repeatedly throughout February).

(4) The *Zoom-in* rescue operation may switch a viewpoint that is blocking unification from *figure* to *ground*. It can only apply if the input tSS is *anchored*, if it has an instantiated imageID (for example we can zoom-in on “this month” but not on “30 minutes”), and if Perception is able to create a zoomed image (as in “she reached the top in ten hours”, where one must imagine some definite ultimate climbing phase lasting ten hours).

(5) The last rescue operation is *Predication*. Its effect is to switch one tSS to an *anchored figure*. It requires that *imageID* be instantiated and it can be used only once in a sentence.

The model, characterized by the tSS and the tMerge operation, claims to be minimalist. tSS are not recursive (*i.e.* a tSS does not contain or refer to another structure of same nature, contrary to feature structures like those used in HPSG for instance). A tSS has a fixed length and cannot grow. Moreover, the tMerge operator is ‘amnesic’, which means that the input tSS are lost after the resulting tSS has been computed. This prevents the model from using unrealistic memory resources in uncontrollably growing structures. Many models are monotonic, which means that the structures they process can only grow in size and complexity during processing, becoming unrealistic for large inputs (Ghadakpour, 2003). Our model is non-monotonic and therefore avoids this problem.

Our model has been implemented in Prolog. The program provides all admissible solutions for an input sentence and signals incorrect sentences.

## Examples

The model and its implementation account for all sentences of our test, including the examples listed in Tables 1 and 2. It detects “incorrect” sentences; it correctly predicts repetition, slice and perfective and imperfectives aspects. Examples (1) and (2) are detailed below.

(1) Elle mangera du gâteau en février  
(*She will eat/be eating cake in February*)

The determiner “du” introduces a *ground* viewpoint. On the other hand, “en” is associated with a *figure*. Let’s consider the step where the two phrases “manger du gâteau” (“to eat cake”) and “en février” (“in February”) are to be unified.

Head: “manger du gâteau” (“to eat cake”)  
[im/i\_eat\_cake, vp/g, an/?]  
Complement: “en février” (“in February”)  
[i\_february, vp/f, an/a]

The *basic merge* (figure 3, (1)) detects a viewpoint conflict. The conflict could be solved either by zooming-in on “en février” (figure 3, (4)) or by predicating “manger du gâteau” (figure 3, (5)), but then the perceptive merge (figure 3, (2)) will detect a granularity difference. We must predicate “manger du gâteau” and zoom-in on “en février” in both case. This leaves us with two solutions.

In the first solution, the *figure* of the head is repeated, and a *ground-ground* merge becomes possible. In the second solution, the complement is sliced and a *figure-figure* merge becomes possible. Slicing is allowed by the fact that “February” is anchored (figure 3, (3)). We thus get the two following interpretations:

Result 1: “manger du gâteau (plusieurs fois) en février” (“to eat cake several times throughout February”)

[i\_eat\_cake\_february, vp/g, an/u]

Result 2: “manger du gâteau” (une fois) en (un moment de) février” (“to eat cake once at some point in February”)

[i\_eat\_cake\_february, vp/f, an/u]

(2) \* Elle mangera du gâteau en 30 minutes

(\* *She will eat (be eating) cake (or from the cake) within 30 minutes.*)

As previously, the tSS of “manger du gâteau” receives a *ground* viewpoint. By contrast with example (1), there is no granularity conflict, but the complement “en 30 minutes” is not anchored. Let’s consider the step where the two phrases “manger du gâteau” (“to eat cake”) and “en 30 minutes” (“within 30 minutes”) are to be unified.

Head: “manger du gâteau” (“to eat cake”)  
[i\_eat\_cake, vp/g, an/?]  
Complement: “en 30 minutes” (“within 30 minutes”)  
[i\_30\_minutes, vp/f, an/u]

There is no way to solve the viewpoint conflict: the complement cannot be zoomed-in because it is unanchored. Predication cannot be used to solve the *viewpoint* conflict, since it creates an *anchoring* conflict. The model returns an error, as expected.

## Conclusion

We have shown how some of the mechanisms of French aspectuality could be predicted using a minimalist model. We share various notions and mechanisms with Gosselin’s and Schilder’s models, including anchoring, granularity checking and dynamic conflict resolution. Our model departs from theirs by the fact that lexical structures are fixed instead of including algorithms. There is only one procedure in our model: tMerge, which is not attached to the

lexicon and can be synchronized with syntactic analysis. Our model is able to detect and solve aspectual effects such as repetition and slice, to identify the perfectivity and progressivity of events, and to detect incorrect sentences. The output of the model is congruent with the majority judgment among the participants we tested.

The notion of semantic *incorrectness* is relative, as a substantial number of participants considered these sentences as acceptable (e.g. 30% for example (2)). Acceptability seems to depend on several factors that are to be investigated: differences in the kind of computations performed in *Perception*, or differences in judging as correct sentences that wouldn't be uttered by a native speaker but that could still make sense. Another possible source of variation among participants may be judgments of relevance. For instance, "Elle mangera du gâteau en 30 minutes" (understood as: "She will eat cake within a period of 30 minutes") may be acceptable in a context in which any consumption of cake is supposed to require more than thirty minutes. We are currently investigating these phenomena.

Though we are confident in the fundamental principles and in the overall architecture of the model, we need to check its validity against a much larger variety of phenomena, not only in French but also in other languages. For instance, we are currently investigating how the English *progressive* "V-ing" (Deo, 2009) can be explained as a sub-categorization of the *ground* viewpoint depending of perceptive information. These investigations may bring us to adapt and augment the model, while hopefully keeping its minimalist character.

## References

- Chomsky, N. (1995) *The Minimalist Program*. Cambridge MA: MIT Press.
- Comrie, B. (1976). *Aspect*. Cambridge university press.
- Comrie, B. (1985). *Tense*. Cambridge university press.
- Deo, A. (2009). Unifying the imperfective and the progressive: partitions as quantificational domains. *Linguistics & Philosophy* 32 (5), 475–521.
- Dowty, D. (1979). *Word Meaning and Montague Grammar*. Dordrecht: D. Reidel.
- Elson, D., & McKeown, K. (2010). Tense and Aspect Assignment in Narrative Discourse. In *Proceedings of the Sixth International Conference in Natural Language Generation*.
- Ghadakpour, L. (2003). *Le système conceptuel, à l'interface entre le langage, le raisonnement et l'espace qualitatif: vers un modèle de représentations éphémères*. PhD Thesis, Ecole Polytechnique.
- Gosselin, L. (1996). *Sémantique de la temporalité en français*. Bruxelles: Duculot.
- Hwang, C.H., & Schubert, L. K. (1992). Tense Trees as the fine structure of discourse. *Proceedings of ACL'1992*, 232-240.
- Jackendoff, R. (2005). Alternative minimalist visions of language. *Chicago Linguistics Society (CLS)*, 41, 189-226.
- Moens, M., & Steedman, M. (1988). Temporal ontology and temporal reference. *Computational linguistics*. 14 (2), 15-28.
- Mueller, E. T. (2003). Story understanding through multi-representation model construction. In Hirst, G. and Nirenburg, S. (eds), *Text Meaning: Proceedings of the HLT-NAACL 2003 Workshop*, (pp. 46-53). East Stroudsburg, PA: Association for Computational Linguistics.
- Person, C. (2004). *Traitement automatique de la temporalité du récit: implémentation du modèle linguistique SdT*. PhD Thesis, Université de Caen Basse-Normandie.
- Pratt, J., & Francez, N. (2001). Temporal Generalized Quantifiers, *Linguistics and Philosophy* 24, 187–222.
- Reichenbach, H. (1947). *Elements of Symbolic Logic*. Free Press.
- Sauri, R., Goldberg, L., Verhagen, M., & Pustejovsky, J. (2009). *Annotating Events in English. TimeML Annotation Guidelines*. Brandeis University. Version TempEval-2010.
- Saussure, L. de (2003). *Temps et pertinence. Eléments de pragmatique cognitive du temps*. Bruxelles: Duculot.
- Schilder, F. (2004). Extracting meaning from temporal nouns and temporal prepositions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3 (1), 33-50.
- Shanahan, M. (1999). The event calculus explained. In Wooldridge, M.J., Veloso, M. (eds.), *Artificial Intelligence Today*. LNCS, 1600, (pp. 409–430). Berlin: Springer.
- Smith, C. S. (1991). *The Parameter of Aspect*. Dordrecht, NL: Kluwer.
- Vendler, Z. (1967). *Linguistics in Philosophy*. Cornell University Press.
- Vetters, C. (1996). *Temps, aspect et narration*. Rodopi.



# Does Analogy Facilitate Transitive Inference in Young Children?

Milena Mutaſchieva ([mmutaſchieva@nbu.bg](mailto:mmutaſchieva@nbu.bg))

Kristina Gotseva ([goceva@students.nbu.bg](mailto:goceva@students.nbu.bg))

Boicho Kokinov ([bkokinov@nbu.bg](mailto:bkokinov@nbu.bg))

Central and East European Center for Cognitive Science, Department of Cognitive Science and Psychology,  
New Bulgarian University, 21 Montevideo Street  
Sofia 1618, Bulgaria

## Abstract

Both transitive reasoning and analogy-making are present at very early stage of human development and the question arises whether the two developmental trajectories interact with each other. We are presenting an experiment with 4 years old children to test the hypothesis that the analogy-making capabilities can scaffold and facilitate the development of transitive inference and the empirical data support this hypothesis.

## Introduction

Transitive inference is the simplest form of deductive reasoning that combines two premises  $R(a,b)$  and  $R(b,c)$  and makes the conclusion that  $R(a,c)$ . For example, if John is stronger than Peter, and Peter is stronger than Bill, then necessarily, John is stronger than Bill. Piaget (1921) was the first psychologist who introduced the topic of transitive inference in the developmental framework. He and his collaborators claimed that the ability of transitive reasoning is developed relatively late during the stage of concrete operations (Inhelder & Piaget, 1958, 1964): first at the age of 8 about size relations like “bigger than”, and even later on about other relations like “heavier than”.

The developmental trajectory of transitive reasoning has been extensively studied and disputed in the literature and many psychologists found evidence that this ability is present much earlier in the development than it was assumed by Piaget. Thus since 1970s the research in the area was dominated by the debate about the age at which children can be considered to make transitive inferences. According to Trabasso (1977) and Halford (1993) this age depends on the specific procedure that tests the ability and under certain circumstances even children at the age of 4 can exhibit this capacity. McGonigle and Chalmers (1977) challenged the field by demonstrating that monkeys can also make transitive inferences. More recently it has been demonstrated that rats (Roberts & Phelps, 1994), pigeons (von Fersen et al, 1991, Lasareva & Wasserman, 2006), and even fishes (Grosenick, Clement & Fernald, 2007) do make transitive inferences. It has been argued that transitive inference is important for the survival because it allows the animals to determine social dominance of other animals without getting into mortal fights.

A very similar debate has also dominated the field of analogical reasoning development. Inhelder and Piaget (1958, 1964) claim that analogical reasoning is a type of

reasoning which develops during the formal operation stage i.e. after 11 years of age. However, like in the field of transitive inference, modern researchers have demonstrated that analogy is present at very early age if not from birth (Goswami, 1991, 2001, Kotovsky & Gentner, 1996). And again there is now some evidence that Chimpanzees (Gillan, Premack, & Woodruff, 1981; Oden, Thompson & Premack, 2001), Baboons (Fagot & Parron, 2010), Capuchin monkeys (Truppa et al, 2010, 2011), and even New Caledonian Crows (Taylor et al., 2007) can make analogies.

This research suggests that both transitive inference and analogy-making are very old evolutionary and are also present from very early age in human beings and the question arises whether the development of the two capabilities interact with each other.

Kokinov (1990, 1992) suggested that in fact deductive, inductive and analogical reasoning might be produced by the very same mechanisms. His proposal is that the analogical reasoning mechanisms are the basis and they are then used for deduction and induction as well. He obtained some experimental support for this proposal using adults. Halford (1993) suggested that deductive reasoning development is based on the already developed analogical reasoning capacities, but did not present empirical evidence for his claim. Goswami (1995) tried to find support for that idea by providing analogies to children in the class inclusion problem, but could not find an effect of the analogy. Based on the two proposals above Mutaſchieva and Kokinov (2008) designed an experiment to test whether analogy can actually help children make better transitive inferences and obtained positive results. However, the doubt remained whether this is really transitive reasoning. The current paper presents an experimental study that replicates and further extends these initial findings making sure that children are making proper transitive inferences.

Another topic of hot debate is the methodology of measuring transitive inference capabilities. Various authors have suggested various procedures and it might be that they measure different aspects of the transitive reasoning abilities. Inhelder and Piaget (1958, 1964) used the original definition of transitive inference with three objects where the premises are presented as two comparisons and the question is about the relation between the two objects unseen together. Later on, however, Bryant and Trabasso (1971) and McGonigle and Chalmers (1977) have used a form of the task with 5 items in a series with extensive training of some pairs and asking about others that are



unseen before. This procedure was meant to eliminate the potential memory problem that children can have remembering the premises. Pears and Bryant (1990) finally eliminated the training phase since there was severe criticism about the role of this training plays.

Sternberg (1980) introduced another version of the three object task, namely if A is shorter than B, and B is shorter than C, which is the shortest? We are using this version in our study. However, there is serious potential problem here. As Piaget emphasised crucial for the transitive inference is the understanding of the relativistic nature of the relations, i.e. that relations, unlike properties which are constant characteristic of an object (e.g. if the object is red it remains red during the whole experiment), are dynamically changing depending on the specific object you are comparing it to. For example, object B may be bigger than object A and at the same time be smaller than C, i.e. being “bigger” is not a property of the object itself. Inhelder and Piaget (1958, 1964) claim that this ability is mastered much later in life – in the stage of concrete operations, while prior to that stage children think that relations are absolute properties of the objects – properties like “larger” and “smaller” are mutually exclusive and it is not possible to attribute them to one and the same object.

In order to test whether children have actually obtained this relativistic understanding of relations and thus exhibit proper transitive inference we have designed a new procedure in which we first show children two separate relations like A is stronger than B, and B is stronger than C, then we ask which is the strongest. Let suppose that the child correctly responds that A is the strongest. Then we add a new object D and say that D is stronger than A and ask children which is the strongest object now. Alternatively in half of the trials we add a new object E and say that C is stronger than E and ask the child which is the strongest now. If children understand correctly the nature of the transitive inferences they have to switch to the new object D in the former case, but keep insisting that A is the strongest in the latter case. We believe this is a stronger and more conservative test of transitive inferences capacity and have introduced it in the current study.

There is another innovation in the test procedure. Since it is natural for children, we speak of how one animal is stronger than another one (e.g. this bear is stronger than this bear). If the animals were visually available than no inferences would be needed and children would visually find the strongest animal, that is why we are presenting the animals hidden in boxes and children have to remember that the bear in this box is stronger than the bear in that box. In addition, to make the task even more complicated we presented the three boxes not in a linear order, but in a triangular configuration. This complicates the task of children according to the “spatial model” (DeSoto, London & Handel, 1965, Huttenlocher, 1968) which assumes that transitive inferences are made based on a linear spatial mental model built in the process of understanding of the premises. As you will see, however, the triangular configuration is important for our manipulation and makes it possible to compare every two objects equally easy.

Finally, the manipulation in the experiment is the introduction of an analogy between the three (or four) boxes in the triangle (rectangular) configuration and a train that is taking turn. In the train the stronger animals are pulling the next ones. In this way we are trying to ground the abstract knowledge of children about “being stronger” with their familiar experience of being able to pull. This is further supported by using physical “draw-bars” between the wagons of the train. In both the experimental and control groups there is a second set of animals on the table which belongs to the experimenter and these animals are not hidden, i.e. the child can visually judge which is stronger than which and which is the strongest. We know from previous studies (Mutafchieva & Kokinov, 2007a, 2007b) that the child can possibly make a mapping between the two sets of animals (the experimenter’s and their own) and we assume that this should facilitate the inference process. The difference is that in the experimental group the train analogy is introduced, while in the control group it is not.

## Experiment

The goal of the present experiment was to find out whether an analogy with a familiar domain that is grounded in children’s physical experience (like the train analogy of pulling) would facilitate the transitive inference while we measure also children’s understanding of the relativistic nature of relations.

### Hypothesis

Our hypothesis was that when provided with the train analogy children would improve their performance on the transitive inference task and would demonstrate understanding of the relativistic nature of relations.

### Design

The experiment had a mixed design.

The *between-subject factor* had two levels:

- **Control condition:** one visible and one hidden set of objects were presented, but no analogy was provided;.
- **Analogy condition:** one visible and one hidden set of objects were presented, and each set was described as a train with wagons that are connected with draw-bars.

The *within-subject factor* was the **number of objects** participating in the object series:

- First measurement – after presenting the two sets of **three objects**.
- Second measurement – after adding a **fourth object** to each set.

Children in both conditions participated in both measurements of the within-subject factor.

The dependent variable was the number of correct responses to the transitive inference tasks and since there were five trials in each session the value could vary from 0 to 5.

## Stimuli

Eight animals of the same type were used in each trial: 8 bears, 8 swans, 8 mice etc. In the first measurement 6 animals were presented, divided in two sets of three animals of the same type – one set of 3 animals for the experimenter and one for the child. There was a big, medium and small animal in each set. The corresponding objects from the two sets were of different absolute sizes - for example the biggest mouse from the experimenter's set and the biggest mouse from the child's set had different sizes. In the second measurement a fourth animal of the same type was added to each set. In addition, in the Analogy condition six draw-bars were used to connect the wagons of the train – three for the experimenter's set and three for the child's set (Figure 1a and 1b, correspondingly).

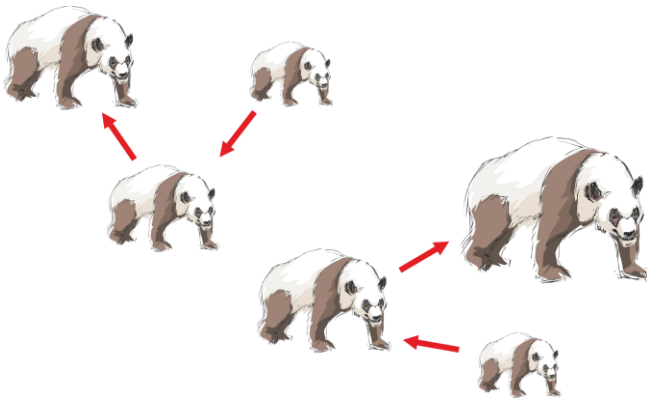


Figure 1a. Example of the stimuli used in the Analogy condition in the first measurement. The biggest object from the upper set was the same absolute size as the medium object from the lower set. The objects from the child's set were hidden under white boxes (Fig. 2)

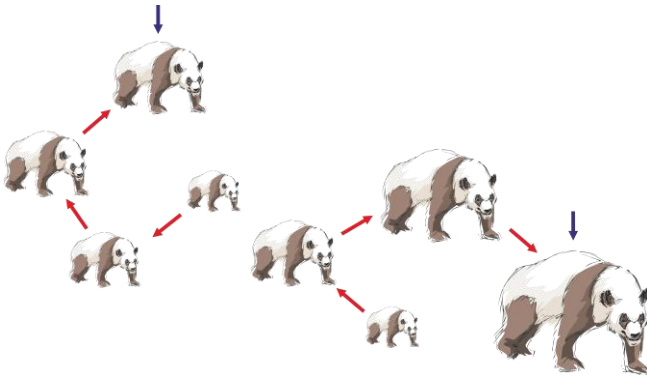


Figure 1b. Example of the stimuli used in the Analogy condition in the second measurement after adding the fourth object.

In this experiment the stimuli were presented in a triangle configuration. Different sets of stimuli were used in every trial with different spatial arrangements and absolute sizes as well as different animals. The objects from the child's set were covered with white boxes of equal size (Figure 2 and Figure 3, correspondingly).



Figure 2. Example of the stimuli used in the Analogy condition in the first measurement before adding the fourth object. The experimenter's set could be seen in the upper side of the table.



Figure 3. Example of the stimuli used in the Analogy condition in the second measurement after adding the fourth object, which in this case is smaller (weaker) than the smallest one.

## Procedure

Each child participated in two individual experimental sessions. Each session included three training trials and five test trials. In the training trials the experimenter gave the child feedback in order to make sure that he/she correctly understood the instruction.

The instruction for the Control group in the test trials of the first measurement session was (in Bulgarian language):

*"I have three bears and you have three bears. Out of these two of my bears this is stronger than this one (pointing to the biggest and the medium bear in the experimenter's set), and from these two of my bears this is stronger than this (pointing to the medium and the smallest bears from her set). Out of these two of your hidden bears, this one is*

stronger than this one, and this one is stronger than this one. Please tell me, where is your strongest bear hidden?"

The instruction for the second measurement for the Control group had two versions depending on whether we added a stronger or a weaker fourth animal. The first one was as follows:

*"Now close your eyes because I will make a trick. Now look, I added another bear to my animals which is stronger than this one (pointing to the formerly biggest animal from her set). I also added another bear to your animals. The new bear is hidden under this box and now this bear is stronger than this one (pointing to the formerly biggest animal from the child's set). Now, please tell me where is your strongest bear hidden?"*

The second version for the Control group in the second measurement was the same except the fact that the fourth added animal was weaker than the weakest one.

The corresponding instruction for the Analogy group in the test trials was the following:

*"I have three bears and you have three bears. Out of these two of my bears (pointing e.g. to the biggest and the medium bear in the experimenter's set) this one is stronger than this one and I will put this draw-bar in such a way that the stronger bear could pull the weaker one. Out of these two of your hidden bears, this one is stronger than this one. Please, put this draw-bar in such a way that the stronger bear could pull the weaker bear. Now, out of these two of my bears (pointing e.g. to the medium and the smallest bear from the experimenter's set) this one is stronger than this one and I will put the draw-bar in such a way that the stronger bear could pull the weaker one. Out of these two hidden bears this is stronger than this. Please, put this draw-bar in such a way that the stronger bear could pull the weaker one. Now look, my bears look like a train in a turn and your bears look like another train in a turn. Please tell me where is your strongest bear hidden?"*

The instruction for the two versions of the second measurement for the Analogy group was the same except that the added fourth animal was connected to the smallest or the biggest animal by a new draw-bar. We randomized the sequence in which the fourth strongest or weakest animal was added. An important fact about this experiment was that the child had never seen the objects in his/her set and it was not possible to solve the task by remembering the absolute sizes of the stimuli.

## Participants

49 children were studied in this experiment. 25 of them formed the Control group, 24 formed the Analogy group.

The average age of the children was 4 years and 5 months ranging from 4 years and 1 month to 4 years and 11 months.

## Results and discussion

The data are presented in Figure 4. The mean for the Control Group in the three objects trials is 2.56 (out of 5), and in the four objects trials is 2.04. There is a slight decrease in this group. The corresponding means for the Analogy group are 3.46 for the three objects trials and 3.58 for the four objects trials. The chance level for the three objects trials is 1.66 and for the four objects trials is 1.25.

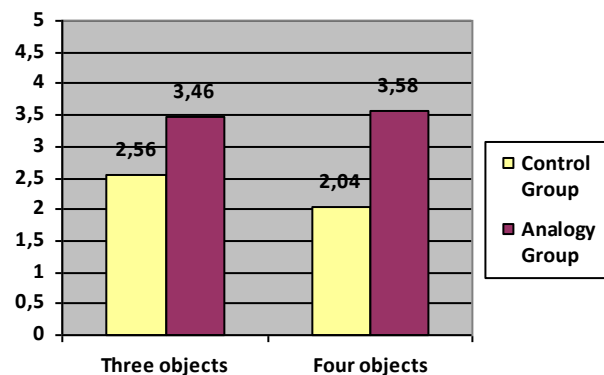


Figure 4. Mean scores (which vary from 0 to 5) for each measurement for each group. There is a clear Analogy effect  $F(1, 47) = 12.478$ , partial  $\eta^2 = 0.21$ ,  $p = 0.001$

A repeated measures ANOVA has been run with within group factor the number of objects and between group factor the presence or absence of analogy. There is an effect of the Analogy manipulation and the performance of the children in the Analogy group is significantly better than the performance of the children in the Control group ( $F(1, 47) = 12.478$ , partial  $\eta^2 = 0.21$ ,  $p = 0.001$ ). Pair-wise comparison shows that there is significant difference between the two groups in both the first measurement with sets of three objects ( $F(1,47)=4.142$ ;  $p=0.047$ ) and in the second measurement with sets of four objects ( $F(1,47)=22.548$ ;  $p<0.001$ ).

Analysis of the within subject factor shows that there is no difference between the two measurements of each child ( $F(1,47)=1.267$ ;  $p=0.266$ ) which means that there is no significant improvement or decrease of the results of each child after adding the fourth object. Children are equally good in solving transitive inference tasks with three and four objects. The interaction between the two factors is also not significant  $F(1, 47)=2.547$ ,  $p=0.072$ ).

Additional analysis shows that there is a significant difference between the children's performance in each measurement and the corresponding chance level in both groups. The difference between first measurement and chance level of 1,66 in the Control group is  $T(24)=7,709$   $p<0,001$ , a similar difference could be found when comparing the second measurement with the chance level of 1,25 in Control group ( $T(24)=12,134$ ;  $p<0.001$ ). For the Analogy group the results are similar ( $T(23)=11,985$ ;  $p<0.001$ ) when comparing the score of first measurement with the chance level of 1,66, and ( $T(23)=12,716$ ;  $p<0.001$ ) when comparing the score of the second measurement and chance level of 1,25). These results could be interpreted that in both measurements in both groups children are significantly better than the chance level in solving the transitive inference tasks.

## General Discussion

The results showed that both groups are significantly better than the chance level in both measurements which means that children are able to solve transitive inference

tasks to some degree. These findings are consistent with the results of Bryant and Trabasso (1971) and Pears and Bryant (1990) which showed that young children could solve transitive tasks when memory limitations were overcome or when the relations were visible for the child at the moment of decision making.

We can also claim that children showed an understanding of relativistic nature of relations because the results from the second measurement are also significantly better than the chance level in both groups. In three of out of 5 trials one and the same object is “stronger than” in the first measurement and is “weaker than” in the second measurement. So, the child had to reverse the answer in order to respond correctly. Children succeeded to overcome this difficulty and demonstrated an understanding that relations are not object’s attributes and have a relative nature.

Most importantly, there is a main effect of the Analogy manipulation and children in the Analogy condition demonstrated significantly better performance than the children in the Control condition (both with three and four objects). This shows that analogy making could play an important role in accumulating experience and development of the transitive reasoning ability. This is in accordance with Halford’s (1993) claim that deductive reasoning development is based on analogical reasoning and with Kokinov’s (1990, 1992) proposal that the same mechanisms could underlie both types of reasoning.

How can we explain the specific effect of the train analogy in this experiment? One possible explanation is that children know that the strongest car that pulls all other cars in the train is the locomotive which is the first wagon in the train. Thus when asked “Where is the strongest bear hidden?” they think of the configuration of boxes as a train and try to find where is the locomotive. Then they use the draw-bars in order to determine where the locomotive is. This explanation relies on their common knowledge of trains and the causal model of its movement. A second possible explanation is that children are mapping their hiding boxes to the experimenter’s set of visible animals; they determine perceptually which the strongest visible animal there is and map it back to the strongest animal hidden in their corresponding box. The question is why children cannot do this in the control condition, and the answer could be that the analogy with the train and the visible draw-bars in both “trains” help children in doing the mapping. This is in accordance with the results obtained in an earlier experiment showing that the train analogy and the use of draw-bars help children make the analogical mapping (Mutafchieva & Kokinov, 2007b). Finally, it could also be argued that the analogy does not play any role here and simply the presence of the draw-bars in both sets makes it easier for the children to do the mapping and find which box corresponds to the strongest animal in the experimenter’s set. Although we cannot rule out this possibility, we believe that this is not very probable since in this earlier experiment (Mutafchieva & Kokinov, 2007b) the results from a

condition with draw-bars without the introduction of the train analogy did not help children in doing the analogical mapping and this group was indistinguishable from the control group. Thus it seems that the draw-bars become meaningful and usable for the children only after they are considered as part of the causal model of the train.

It is also possible that the draw-bars simply facilitated children in remembering the relations between the four objects and these physical representations were visible for the child at the moment of decision making. As explained above, however, we believe that the draw-bars alone will not do the trick. Especially since the children in both groups seemed to successfully remember the premises which could be seen from their answers to the control questions asked. Of course, we need further experimentation in order to distinguish between all these (and other) explanations.

The main conclusion for the moment is that children in our experiment demonstrated an ability to make transitive inferences and understanding of relativistic nature of relations. In addition, the train analogy in combination with draw-bars as physical representations of relations facilitates this children’s ability to make transitive inferences. These first data suggest that analogy could be an effective mechanism for learning to do transitive inferences supporting Halford’s theory that deductive reasoning development is based on analogy-making development.

## Acknowledgments

The research reported in this paper has been supported by the ANALOGY project (6<sup>th</sup> FP, NEST program, contract 29088) and by the European Office for Aerospace Research and Development under grant FA8655-10-1-3061. We would like to thank Ana Doncheva and Margarita Pavlova for helping in the data collection process.

## References

- Ameel, E., Verschueren, N., Schaeken, W. (2007). The relevance of selecting what’s relevant: A dual process approach to transitive reasoning with spatial relations. *Thinking and Reasoning*, vol. 13 (2), 164-187.
- Breslow, L. (1981). Reevaluation of the literature on the development of transitive inferences. *Psychological Bulletin*, 89, 325-351.
- Bryant, P.E., & Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature*, 232, 456-458.
- DeSoto, C., London, M., & Handel, S. (1965). Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology*, 2, 513-521.
- Fagot, J; Parron, C (2010). Relational matching in baboons (Papio papio) with reduced grouping requirements. *JEP: Animal Behavior Processes*, 36(2), 184-193.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.
- Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes

- (Eds.), Perspectives on thought and language: Interrelations in development (pp. 225-277). London: Cambridge University Press.
- Gillan, D. J., Premack, D., & Woodruff, G. (1981). Reasoning in the Chimpanzee .1. Analogical Reasoning. *Journal of Experimental Psychology-Animal Behavior Processes*, 7(1), 1-17.
- Goswami, U. (1991). Analogical reasoning: What develops? A Review of Research and Theory. *Child Development*, 62, 1-22.
- Goswami, U. (1995). Transitive relational mappings in 3- and 4-year-olds: The analogy of Goldilocks and the Three Bears. *Child Development*, 66, 877-892.
- Goswami, U. (2001). Analogical reasoning in children. In Gentner, D., Holyoak, K., Kokinov, B. *The Analogical Mind: Perspectives from Cognitive Science*. A Bradford Book, The MIT Press. Cambridge.
- Goswami, U., Pauen, S. (2005). The effects of family analogy on class inclusion reasoning in young children. *Swiss Journal of Psychology*, vol. 64, n 2, pp. 115-124
- Grosenick, L., Clement, T. & Fernald, R. (2007). Fish can infer social rank by observation alone. *Nature*, 445, pp. 429-432.
- Halford, G.S. (1984). Can young children integrate premises in transitivity and serial order tasks? *Cognitive Psychology*, 16, 65-93
- Halford, G. (1993). Children's understanding: The development of mental models. Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey.
- Huttenlocher, I. (1968). Constructing spatial images: A strategy in reasoning. *Psychological Review*, 75(6), 550-560.
- Inhelder, B. & Piaget, J. (1958). The Growth of Logical Thinking from Childhood to Adolescence. New York: Basic Books.
- Inhelder, B. & Piaget, J. (1964). The Early Growth of Logic in the Child: Classification and Seriation. London: Routledge and Kegan Paul.
- Kokinov, B. (1990). Associative Memory-Based Reasoning: Some Experimental Results. In: *Proceedings of the 12th Annual Conference of the Cognitive Science Society*. MIT, Erlbaum, Hillsdale, NJ.
- Kokinov, B. (1992). Inference Evaluation in Deductive, Inductive and Analogical Reasoning. In: *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Erlbaum, Hillsdale, NJ.
- Kotovskiy L., Gentner D. (1996). Comparison and categorization in the development of relational similarity. *Child Dev.* 67:2797--2822.
- Lazareva, O.F., & Wasserman, E.A. (2006). Effect of stimulus orderability and reinforcement history on transitive responding in pigeons. *Behavioural Processes*, 72, 161-172.
- McGonigle B., Chalmers M (1977). Are monkeys logical? *Nature*, 267, 694-996.
- Mutafchieva, M. & Kokinov, B. (2007a). Does the Family Analogy Help Young Children To Do Relational Mapping? In: *Proceedings of the European Cognitive Science Conference*. Erlbaum, Hillsdale, NJ.
- Mutafchieva, M. & Kokinov, B. (2007b). Can Language be Replaced? Physical Representations of Relations Instead of Language Labels in Relational Mapping: Do They Help Young Children? In: *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. Erlbaum, Hillsdale, NJ.
- Mutafchieva, M. & Kokinov, B. (2008). Can Analogy Help Children Make Transitive Inference? In: *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Erlbaum, Hillsdale, NJ.
- Oden, D, Thompson, R., Premack, D. (2001). Can an Ape Reason Analogically? Comprehension and Production of Analogical Problems by Sarah, a Chimpanzee (Pan troglodytes). In: Gentner, D., Holyoak, K., Kokinov, B. (eds.) *The Analogical Mind: Perspectives from Cognitive Science*, Cambridge, MA: MIT Press, pp. 471-497.
- Pears, R., & Bryant, P. E. (1990). Transitive inferences by young children about spatial position. *British Journal of Psychology*, 81, 497-510.
- Piaget, J. (1921). Une forme verbale de la comparaison chez l'enfant. *Archive de Psychologie*, 18, 141 – 172.
- Piaget, J. (1971). *Biology and knowledge*. Edinburgh: Edinburgh University Press
- Rattermann, M.J., & Gentner, D. (1998). The effect of language in similarity: The use of relational labels improves young children's performance in a mapping task. In K. Holyoak, D. Gentner, & B. Kokinov. *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences*. New Bulgarian University, Sofia.
- Roberts, W.A. & Phelps, M.T. (1994). Transitive inferences in rats & A test of the spatial; coding hypothesis. *Psychological Science*, 5(6), 368-374.
- Sternberg, R.J. (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology: General*, 109, 119-159.
- Taylor, A., Hunt, G. Holzhaider, J., Gray, R. (2007). Spontaneous Metatool Use by New Caledonian Crows. *Current Biology*, 17, 1504-1507.
- Trabasso T. (1977). The role of memory as a system in making transitive inference. In R.V.Kail, Jr., & J.W. Hagen (Eds.), *Perspectives on the development of memory and cognition*. Hillsdale, NJ. Erlbaum.
- Truppa V, Garofoli D, Castorina G, Mortari E, Natale F, Visalberghi E (2010). Identity concept learning in matching-to-sample tasks by tufted capuchin monkeys. *Animal Cognition*, 13(6), 835-848.
- Truppa V, Piano Mortari E, Garofoli D, Privitera S, Visalberghi, E. (2011) Same/Different Concept Learning by Capuchin Monkeys in Matching-to-Sample Tasks. *PLoS ONE*, 6(8): e23809.
- von Fersen, L., Wynne, C.D.L., Delius, J.D., & Staddon, J.E.R. (1991). Transitive inference formation in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 17, 334-341.



# Cognitive Styles in Two Cognitive Sciences

James Myers (Lngmyers@ccu.edu.tw)

Graduate Institute of Linguistics, National Chung Cheng University  
Minhsiung, Chiayi 62102 Taiwan

## Abstract

Miller (1990) suggests that communication between linguistics and psychology is hampered for essentially cognitive reasons: linguists favor simplifying explanations while psychologists favor causal explanations. This paper reformulates this suggestion as three testable hypotheses. First, scientists vary in cognitive style from rationalist/nomological to empiricist/mechanistic. Second, linguistics is primarily a rationalist/nomological science, while psychology is primarily an empiricist/mechanistic one. Strikingly, even among nativists, linguistic and psychological research still contrast along the rationalist/empiricist dimension. Third, cognitive styles are relatively intractable, as suggested by empirical evidence that they are associated with intrinsic individual differences and by formal arguments that highlight their self-isolating nature.

**Keywords:** philosophy of science; epistemology; linguistics; psychology; individual differences.

## Introduction

Linguists and psychologists have often noted a persistent lack of mutual respect across their two disciplines (Johnson-Laird, 1987; Jackendoff, 1988; Miller, 1990; Carlson, 2003). Miller (1990, p. 321) ascribes this impasse to different ways of thinking, which we will call cognitive style:

What is holding up the free flow of ideas back and forth between linguists and psychologists? For what it is worth, my own view is that linguists and psychologists subscribe to different theories of explanation. Linguists tend to accept simplifications as explanations. [...] For a psychologist, on the other hand, an explanation is something phrased in terms of cause and effect, antecedent and subsequent, stimulus and response.

The present paper reformulates Miller's suggestion as the hypotheses in (1)-(3), then tests them.

- (1) Scientists vary in cognitive style from more rationalist and nomological to more empiricist and mechanistic.
- (2) Linguistics is primarily a rationalist/nomological science, while psychology is primarily an empiricist/mechanistic one.
- (3) Individual cognitive styles are relatively intractable.

By "cognitive style" we mean to cover both one's personal epistemology (theory of knowledge) and one's metaphysics (theory of reality). Rationalist epistemology emphasizes the role of reason in attaining knowledge, while empiricism emphasizes the senses. A nomological metaphysics explains

in terms of general laws, while a mechanistic explanation is expressed in terms of causal systems.

The overarching notion explored in this paper is that individual scientists tend to lean towards rationalism or empiricism, and towards nomological or mechanistic explanations, and that these leanings are reflected in their chosen scientific (sub)disciplines. Thus we expect to see reliable tendencies at both the individual and discipline level, but exceptions are not impossible.

Given the brevity of this paper and the ambitious nature of the hypotheses, this study cannot fully verify or falsify them. Obvious limitations include our focus on just two dimensions of scientific thinking (see, e.g., Hacking, 2002, for others), and the poorly understood relationship between individual and disciplinary styles (i.e., to what extent scientific styles are cognitive or social).

Each hypothesis is taken up separately. Support comes from published research in linguistics, psychology, and the history and philosophy of science.

## Cognitive Styles across the Sciences

Hypothesis (1) can be unpacked into three subhypotheses:

- (1a) Scientists vary from nomological to mechanistic.
- (1b) Scientists vary from rationalist to empiricist.
- (1c) Scientists in nomological sciences tend to be more rationalist, whereas those in mechanistic sciences tend to be more empiricist.

Regarding (1a), physicists favor nomological explanations, seeking the axioms underlying nature (exemplified in the Euclid-inspired style of Newton's *Principia*). Philosophers of physics sometimes write as if all science is inherently nomological (e.g., Scheibe, 2001), but in fact, physicists are atypical: chemists, biologists, geologists and other scientists generally concentrate on uncovering causal mechanisms, not general laws. As a typical example of a mechanism, Machamer, Darden, & Craver (2000) cite the transmission of chemical signals across a synapse, which involves entities (neurons, neurotransmitters) and activities (the releasing and binding of chemicals) in events that begin, progress, and end in space and time.

The ultimate reason for cross-disciplinary differences in explanatory metaphysics seems to be reality. The systems studied in fundamental physics are simple enough for simple laws to work well, and being fundamental, no lower-level causal explanation is possible. By contrast, chemical, biological, and geological phenomena merely need to be consistent with physics; historical contingences make them impossible to deduce from physics alone.

Yet there also seems to be a cognitive aspect to the preference for certain explanation types. Strikingly, scientists outside physics often demand mechanisms before accepting a pattern as scientifically significant, no matter how robust the observations. A classic example is the skepticism that greeted Alfred Wegener's theory of continental drift. This theory elegantly explained the fit between the coastlines of South America and Africa, among numerous other observations, but it was not accepted (revised as plate tectonics) until a plausible mechanism was discovered many decades later (Cohen, 1985).

As for (1b), cognitive scientists are used to thinking of rationalism as synonymous with nativism, but the focus here is instead on what Wole ski (2004) calls methodological rationalism (apriorism; nativism is genetic rationalism), as opposed to methodological empiricism (aposteriorism). Rationalism in this sense is associated with deduction from general axioms, while empiricism is associated with induction from specific tokens. More specifically, our focus is not on the views of professional philosophers, but the personal epistemologies (Muis, Bendixen, & Haerle, 2006) of practicing scientists.

Physicists are uncontroversially the most rationalist of scientists, having always tied their work to mathematics (as reemphasized by Galileo and Newton). Despite the physics-centric view of science among some philosophers, scientists in other disciplines view themselves quite differently. Surveys reviewed in Muis et al. (2006) find that while theoretical physicists group with mathematicians in scoring high for rationalist epistemology, chemists and biologists score high for empiricism.

The rationalist/empiricist contrast is not identical to the nomological/mechanistic contrast. Descartes claimed to use pure reason to deduce a theory of physics in which magnets attract via screw-like particles (Westfall, 1971). Similarly, natural selection has lawlike properties (Bock, 2010), even though arguments for it typically involve induction from masses of observations, as in Darwin's *Origin of species* (in sharp contrast to the deductive style of Newton's *Principia*).

Yet consistent with (1c), rationalists seem to have a natural affinity for laws and empiricists for mechanisms. Physicists since Newton have rejected mechanistic physics as question-begging, and biologists tend to see natural selection as a causal mechanism (Skipper & Millstein, 2005). More generally, while nomological approaches appeal to rationalists by being elegant and deductive, mechanisms appeal to empiricists because they are imageable as diagrams, posit causal interactions of the sort familiar from experience, and accommodate a diversity of induction-supporting observations through the cumulative elaboration of mechanical elements and architectures.

In a passage reminiscent of Miller (1990), Shapin (1996, p. 117) highlights the personal nature of cognitive styles:

Do you want to capture the essence of nature and command assent to representations of its regularities?  
Do you want to subject yourself to the discipline of

describing, and perhaps generalizing about, the behavior of medium-sized objects actually existing in the world? [...] The one is not necessarily to be regarded as a failed version of the other, however much partisans may defend the virtues of their preferred practice and condemn the vices of another.

Shapin illustrates this contrast with Newton and his contemporary Boyle, today known as the father of chemistry. When Newton first introduced his optical experiments to the Royal Society, he placed his observations in a deductive context: "I shall rather lay down the *Doctrine* first and then, for its examination, give you an instance or two of the *Experiments*, as a specimen of the rest" (quoted in Shapin, 1996, p. 114, italics in the original), adding that he had shown that the science of colors was "mathematical" with "as much certainty in it as any other part of optics," as "evinced by the mediation of experiments concluding directly and without any suspicion of doubt" (Shapin, 1996, p. 115). By contrast, Boyle's goals were simply to describe his experiments explicitly enough for others to experience them vicariously, if not replicate them and observe the results first-hand. Thus he wrote that he dared to "speak confidently and positively of very few things, except of matters of fact" (quoted in Shapin, 1996, p. 102). He even refused to express mathematically the gas law that today bears his name or to call it a law. Boyle also invented today's scientific reporting style, which clearly separates experimental description from interpretation.

### Cognitive Styles in Linguistics and Psychology

To paraphrase Shapin, we do not regard linguistic methodology as a failed version of psychological methodology, or vice versa. Rather, hypothesis (2) merely claims that there is a deep philosophical difference.

Miller (1990, p. 321) illustrates this difference as follows:

[A] grammarian who can replace language-specific rewriting rules with X-bar theory and lexicalization feels he has explained something: the work formally done by a vast array of specific rules can now be done with a simple schema. [...] To an experimental psychologist, X-bar theory is not an explanation; rather, if it is true, it is something to be explained.

More generally, linguistic explanations take the form of nomological grammars, which many linguists believe are themselves subject to universal laws. Evidence is slight by psychological standards, relying on a small number of clear cases; generality is assumed unless later shown otherwise. By contrast, psychological explanations take the form of processing systems, supported with empirical overkill by linguistic standards. Psychologists also follow Boylean reporting style, whereas linguists, like philosophers, freely interleave examples and analyses.

The contrasting philosophies underlying these practices are sometimes made explicit in rhetorical volleys aimed



across disciplinary borders. For example, Chomsky (2002, p. 102) claims that "the only field that has methodology courses, to my knowledge, is psychology," implying that psychologists still maintain the discredited notion of empirical discovery procedures. The psychologists Edelman & Christiansen (2003, p. 60) counter that linguists fail in their duty "to demonstrate the psychological (behavioral), and, eventually, the neurobiological, reality of the theoretical constructs," i.e., to demonstrate mechanisms.

Favoring mechanisms makes psychology an entirely typical science (outside of physics), but why does linguistics favor the nomological approach? Even though Chomsky is an easy target, linguists have always treated language nomologically. Bloomfield (1926) proposed a *Principia*-like axiomatic system for linguistic theory, and Chomsky (1966) links his own approach with Renaissance grammarians, who, in turn, built on the Aristotelian view of language as logic. Even before Aristotle and beyond Europe, Pāṇini developed an elaborate formal grammar of Sanskrit that presaged generative linguistics in important ways.

One motivation for nomological linguistics may be language itself: like physics, language is simple enough to be fruitfully described with simple laws. Yet unlike physics, a mechanistic approach to language is also fruitful, given that its processing involves causes and effects. Miller (1990, p. 321) tacitly admits the multifaceted nature of language when he addresses the question of why linguists and psychologists study it so differently, noting that "[s]ome have answered this question in terms of the competence-performance distinction: linguists and psychologists talk about different things," while "[o]thers have answered this question in terms of the structure-function distinction: linguists ask different questions of the same thing."

In fact, both of the alternative explanations he cites can be subsumed under cognitive style. Competence (grammar) is static linguistic knowledge, described in terms of structures; performance is language processing, described in terms of functions. Grammar is intrinsically atemporal, despite the confusion sometimes caused by the term "generative" (see Neeleman & van de Koot, 2010, for a computational argument that grammar must be atemporal). Hence grammar cannot be modeled mechanistically, since causal mechanisms necessarily operate in time (Machamer et al., 2000). This obligates grammarians to use nomological explanations, which, given hypothesis (1c), makes grammatical research particularly appealing to rationalists.

The central role of atemporality in nomological/rationalist approaches to language is highlighted by the turn that came with Saussure, the first modern linguist to explicitly advocate the primacy of synchronic grammar. The rise of structuralist linguistics in his wake was accompanied by a simultaneous rejection of psychological approaches to language, exemplified by the shift from the Wundtian philosophy of Bloomfield (1914) to the psychological agnosticism of Bloomfield (1933).

Since cognitive style is a property of individual scientists, methodological rationalists like Chomsky coexist in

linguistics with methodological empiricists like the corpus linguist Geoffrey Sampson. While Chomsky (2002, p. 102) praises "the Galilean move towards discarding recalcitrant phenomena if you're achieving insights by doing so," Sampson (2001, p. 1) asks linguists "to apply the same empirical techniques which have deepened our understanding of other observable aspects of the universe during the four centuries since Galileo." (Tellingly, both invoke Galileo: as noted by Shapin, 1996, scientists tend to take their personal cognitive style as defining science itself). Other non-rationalist approaches to linguistics include functionalism (e.g., Nichols, 1984), where causal mechanisms are central.

Psychologists also vary in cognitive style, with some attracted to nomological schools like rational analysis (Chater & Oaksford, 1999), or, earlier, Gestalt psychology. Yet other psychologists are quick to criticize such approaches for being insufficiently mechanistic, as seen in the peer commentary on Oaksford & Chater (2009), or in the comment by Bruce, Green, & Georgeson (2003, p. 127) that Gestalt laws have left us "with a set of descriptive principles, but without a model of perceptual processing."

The difference in cognitive style between (grammar-oriented) linguists and (most) psychologists is highlighted by the surprising fact that even when psycholinguists are genetic rationalists, they remain methodological empiricists. Compare Chomsky's nativism with that of the psychologist Steven Pinker. The poverty of the stimulus argument (Chomsky, 1986) deduces nativism from the premises that knowledge is rich but input to the learner is poor. This is a rational argument: Chomsky ascribes it to Plato, implying the irrelevance of two millennia of evidence. Moreover, contrary to a common misunderstanding (see, e.g., Chomsky & Katz, 1981), Chomsky does not induce universal grammar from grammatical universals.

By contrast, the prominence of these two arguments is exactly reversed in Pinker (1994). Pinker does rehearse a version of the poverty of the stimulus argument, but immediately afterwards he adds, "Chomsky's claim was tested in an experiment..." (p. 42). The remainder of Pinker's book is a long catalog of empirical evidence for nativism from a variety of sources, including cross-linguistic universals, in a style more like *Origin* than *Principia*.

The Chomskyan Jenkins (2000, p. 31) confirms that even if "no converging evidence at all of the kind Pinker detailed in his book [...] had turned up yet", we would still "be justified in accepting Chomsky's arguments that the 'basic design of language is innate,' to use Pinker's words", since "the argument from poverty of the stimulus [is] strong enough".

### The Intractability of Cognitive Style

Hypothesis (3) is supported by two independent arguments. For the benefit of different types of readers, one is empirical and one is rational. The empirical argument is based on evidence that cognitive style is associated with intrinsic individual differences. The rational argument is based on

formal analyses demonstrating the impossibility of choosing between cognitive styles either empirically or rationally.

### Individual Differences in Cognitive Style

The polarization of disciplinary styles can be explained in part by the social and cognitive factors driving group polarization more generally (Isenberg, 1986). Yet scientists do not merely absorb the cognitive style of their mentors and peers.

For example, the physicist Max Planck (quoted in Scheibe, 2001, p. 69) wrote:

What led me to my science and what fascinated me from a young age was the, by no means self-evident, fact that our laws of thought agree with the regularities found in the succession of impressions we receive from the external world, that it is thus possible for the human being to gain enlightenment regarding these regularities by means of pure thought....

Similarly, though Einstein described himself, thirty years after the fact, as having been changed "by the problem of gravitation" from "skeptical empiricism" into "a believing rationalist" (quoted in van Dongen, 2010, p. 57), his contemporary writings show that he had already rejected traditional empiricism before starting to work on gravitation (Howard, forthcoming).

Group dynamics also fail to explain why individual scientists can differ in cognitive style within a science. The physicist Dirac (1968) notes that "[w]hether one follows the experimental or the mathematical procedure depends largely on the subject of study, but not entirely so. It also depends on the man" (p. 22). Among his fellow theoretical physicists, he cites Heisenberg as exemplifying the former style, and Schrödinger (and himself) the latter.

Is cognitive style a matter of personality? Diamond & Royce (1980) explicitly argue that personal epistemology does correlate with personality traits, though unfortunately there has been little follow-up (cf. Muis et al., 2006). For example, Pashler, McDaniel, Rohrer, & Bjork (2009) cast serious doubt on the notion of intrinsic learning styles, but putative contrasts like visual versus verbal learning have nothing to do with empiricism versus rationalism.

More relevant are the many studies that reveal individual variation in reasoning. For example, in the Wason card selection task, where people are asked to test for violations of conditional propositions of the form if  $p$  then  $q$ , most fail to test cases of not- $q$ , even though finding not- $q$  given  $p$  would falsify the proposition. Oaksford & Chater (2009) review a series of experiments and mathematical models suggesting that most people approach the Wason task using Bayesian (i.e., essentially inductive) reasoning. This implies that the minority who give the normatively correct answer are using deductive reasoning. Since performing correctly is positively correlated with general intelligence (Stanovich & West, 2000) and working memory capacity (Copeland &

Radvansky, 2004), a facility with deduction seems to be an intrinsic individual trait.

Cognitive style may also be linked to another feature known to show individual variation: tolerance of uncertainty (Neuberg, Judice, & West, 1997). Deduction provides certainty (Euclid cannot be falsified), while induction does not (a black swan may be lurking around the next corner). Given this, it is unsurprising that Wilkinson & Migotsky (1994) found that the belief that knowledge depends on observations and data (which they label empiricism) is associated with the belief that knowledge is relative and context-dependent, whereas the belief that knowledge depends on logical and analytical thinking (which they label rationalism) is not associated with relativist tendencies.

When taken to an extreme, the intolerance of uncertainty is a hallmark of obsessive-compulsive disorder and autism spectrum disorders like Asperger syndrome (Sadock & Sadock, 2007). Indeed, autism is linked to mathematical (i.e., deductive) talent (Baron-Cohen, Wheelwright, Burtenshaw, & Hobson, 2007), and students with familial autism are more likely to choose a mathematics-intensive college major (Campbell & Wang, 2012). Complementarily, Péliissier & O'Connor (2002) found that participants diagnosed with obsessive-compulsive disorder had significantly more difficulty than matched controls in tasks that required induction, whereas both groups performed equally well with deductive tasks. Consistent with the above findings, the rationalists Newton (James, 2003) and Dirac (Farmelo, 2009) have been diagnosed (post hoc) with Asperger, whereas the more empiricism-inclined Darwin earned a very high score from historians for openness to experience (Shermer, 2002).

In short, scientists are unconsciously pushed in one or the other direction along the rationalism/empiricism continuum even before their entry into science. Cognitive styles are thus expected to be relatively immune to persuasion.

### Cognitive Styles and Logical Traps

Whether rationalists or empiricists, scientists all face the problem of induction. Since observations are always consistent with (infinitely) many theories, scientists rely on strategies like Ockham's razor to narrow their options. However, even when Ockham is satisfied, rationalists are free to favor laws and empiricists to favor mechanisms.

The implications of this problem can be investigated in the context of formal learning theory. This approach is most familiar to cognitive scientists in its application to language acquisition, but it also helps inform the philosophy of science (e.g., Kelly, 2000). One lesson is that the challenge faced by scientists is much harder than that faced by infants. Infants may have evolutionarily provided "inside help" guiding them to linguistic knowledge, but scientists are not so lucky. Even if scientific decisions are innately biased, these biases are not guaranteed to lead to scientific truth.

Because of this distinction between children and scientists, formal learning theory leads to opposite conclusions in the two cases. Gold's theorem (Gold, 1967) provides a simple

illustration of this. Any finite set of strings is compatible both with a finite language consisting solely of the attested strings and with a language generated by rules that generate the attested strings plus infinitely many more. However, if the learner is biased to posit only one language type, and the strings are generated by a previous learner with the same bias, successful language learning is guaranteed. Some creatures may achieve this by being innately biased towards finite languages, but presumably humans do so via an innate bias for infinite ones.

But now consider a different scenario. A scientist collects utterances from a human, which we assume has an infinite language, and from an alien with a finite language consisting of symbol strings up to length three. The human is observed to have many more string lengths than the alien, but like the child learner, the scientist can only ever have access to strings of finite length, hence finite sets of strings, even for human utterances. If the scientist is biased towards nomological explanations, an infinite language (e.g., the grammar  $a^*$ ) will be incorrectly posited for the alien as well. If the scientist is biased towards less analytical hypotheses (i.e., is empiricist), the human will be ascribed an alien-like language (i.e., a list of strings).

One may object that the scientist can escape from this dilemma by collecting evidence about what is ungrammatical. Unlike the child learner, the scientist may run an experiment, for example "asking" the alien if some string is "acceptable" (e.g., via some processing task). Suppose the alien accepts  $a$ ,  $aa$ ,  $aaa$ , but rejects  $aaaa$ ,  $aaaaa$ ; surely that demonstrates that it has a finite language. Yet as Johnson (2004) points out, even negative evidence can only come in finite sets. Rejection of  $aaaaa$  does not preclude the acceptance of  $aaaaaa$ .

Kelly (2007) illustrates his proof that Ockham's razor minimizes backtracking with a scenario in which an Ockham-obeying scientist posits all and only the particles that have been emitted by a machine so far, exactly parallel to our string-listing empiricist. Yet the nomological analysis is not only intuitively simple, but also requires no backtracking. If no string longer than three is uttered for a sufficiently long time, we can merely add a constraint to the grammar; we are not compelled to give up the infinite  $a^*$  component. Thus both approaches obey Ockham's razor.

Nothing in the above discussion depends on specific features of Gold's theorem; the problem arises from the general problem of induction. Thus even in more realistic scenarios, the two approaches cannot be distinguished either empirically or rationally.

## Conclusions

This paper has argued that linguists and psychologists talk past each other primarily because they have different epistemic and metaphysical commitments that are beyond conscious control.

Miller (1990, p. 322) predicts that it will be hard for linguists "to make clear to psychologists that simplifying explanations can be satisfying, once you grow accustomed

to them." Despite its tentative nature, the present study has provided ample reason to take Miller's prediction seriously. The predominance of the empiricist/mechanistic approach across the sciences suggests that it may be easier for linguists to appreciate psychological explanations than the other way around. Nevertheless, linguistics will always have a core of nomological rationalists. Given the multifaceted nature of language, this is perhaps just as it should be.

## Acknowledgments

Research was supported by the grants NSC98-2410-H-194-086-MY3 and NSC100-2410-H-194-109-MY3 from the National Science Council (Taiwan). Comments and other help came from Jenn-Yeu Chen, Ruey-Lin Chen, Tsung-Ying Chen, Leo Mos, Krista Muis, Gerry Rau, Gary Shyi, Jon Sprouse, James Tai, and four anonymous reviewers. The usual caveats apply.

## References

- Baron-Cohen, S., Wheelwright, S., Burtenshaw, A., & Hobson, E. (2007). Mathematical talent is linked to autism. *Human Nature*, 18, 125-131.
- Bloomfield, L. (1914). *An introduction to the study of language*. New York: Holt.
- Bloomfield, L. (1926). A set of postulates for the science of language. *Language*, 2, 153-164.
- Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart & Winston.
- Bock, W. J. (2010). Multiple explanations in Darwinian evolutionary theory. *Acta Biotheoretica*, 58, 65-79.
- Bruce, V., Green, P. R., & Georgeson, M. A. (2003). *Visual perception*, 4th ed. Psychology Press.
- Campbell, B. C., & Wang, S. S. (2012). Familial linkage between neuropsychiatric disorders and intellectual interests. *PloS ONE*, 7 (1), 1-4.
- Carlson, G. (2003). On the notion "showing something". In J. Moore & M. Polinsky (Eds.) *The nature of explanation in linguistic theory* (pp. 69-82). Center for the Study of Language and Information.
- Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, 3 (2), 57-65.
- Chomsky, N. (1966). *Cartesian linguistics*. New York: Harper and Row.
- Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. New York: Praeger.
- Chomsky, N. (2002). *On nature and language*. Cambridge, UK: Cambridge University Press.
- Chomsky, N., & Katz, J. J. (1981). What the linguist is talking about. In N. Block (Ed.) *Readings in philosophy of psychology*, vol. 2 (pp. 223-237). London: Methuen.
- Cohen, I. B. (1985). *Revolution in science*. Cambridge, MA: Harvard University Press.
- Copeland, D., & Radvansky, G. (2004). Working memory and syllogistic reasoning. *The Quarterly Journal of Experimental Psychology A*, 57 (8), 1437-1457.

- Diamond, S. R., & Royce, J. R. (1980). Cognitive abilities as expressions of three "ways of knowing". *Multivariate Behavioral Research*, 15 (1), 31-56.
- Dirac, P. A. M. (1968). Methods in theoretical physics. In *From a life of physics; Evening lectures at the International Center for Theoretical Physics, Trieste, Italy, 19-30*. A special supplement of the International Atomic Energy Agency Bulletin, Austria.
- Edelman, S., & Christiansen, M. H. (2003). How seriously should we take Minimalist syntax? *Trends in Cognitive Science*, 7 (2), 60-61.
- Farmelo, G. (2009). *The strangest man: The life of Paul Dirac*. London: Faber and Faber.
- Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10, 447-474.
- Hacking, I. (2002). "Style" for historians and philosophers. In I. Hacking (Ed.) *Historical ontology* (pp. 178-199). Cambridge, MA: Harvard University Press.
- Howard, D. (forthcoming). Einstein and the development of twentieth-century philosophy of science. In M. Janssen & C. Lehner (Eds.), *The Cambridge companion to Einstein*. Cambridge: Cambridge University Press.
- Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, 50 (6), 1141-1151.
- Jackendoff, R. (1988). Why are they saying these things about us? *Natural Language & Linguistic Theory*, 6 (3), 435-442.
- James, I. (2003). Singular scientists. *Journal of the Royal Society of Medicine*, 96, 36-39.
- Jenkins, L. (2000). *Biolinguistics: Exploring the biology of Language*. Cambridge, UK: Cambridge University Press.
- Johnson, K. (2004). Gold's theorem and cognitive science. *Philosophy of Science*, 71 (4), 571-592.
- Johnson-Laird, P. N. (1987). Grammar and psychology. In S. Mogdil & C. Mogdil (Eds) *Noam Chomsky: Consensus and controversy* (pp. 147-156). New York: Falmer Press.
- Kelly, K. T. (2000). The logic of success. In P. Clark & K. Hawley (Eds.) *Philosophy of science today* (pp. 11-38). Oxford: Clarendon Press.
- Kelly, K. T. (2007). Simplicity, truth, and the unending game of science. In S. Bold, B. Löwe, T. Räscher, & J. van Benthem (Eds.) *Foundations of the formal sciences V: Infinite games* (pp. 223-270). London: College Publications.
- Machamer, P., Darden, L. & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67, 1-25.
- Miller, G. A. (1990). Linguists, psychologists and the cognitive sciences. *Language*, 66, 317-322.
- Muis, K. R., Bendixen, L. D., & Haerle, F. C. (2006). Domain-general and domain-specificity in personal epistemology research: Philosophical and empirical reflections in the development of a theoretical framework. *Educational Psychology Review*, 18, 3-54.
- Neeleman, A., & van de Koot, H. (2010). Theoretical validity and psychological reality of the grammatical code. In M. Everaert, T. Lentz, H. De Mulder, Ø. Nilsen, & A. Zondervan (Eds.), *The linguistics enterprise: From knowledge of language to knowledge in linguistics* (pp. 183-212). Amsterdam: John Benjamins.
- Neuberg, S. L., Judice, T. N., & West, S. G. (1997). What the need for closure scale measures and what it does not: Toward differentiating among related epistemic motives. *Journal of Personality and Social Psychology*, 72 (6), 1396-1412.
- Nichols, J. (1984). Functional theories of grammar. *Annual Review of Anthropology*, 13, 97-117.
- Oaksford, M., & Chater, N. (2009). Précis of *Bayesian rationality: The probabilistic approach to human reasoning*. *Behavioral and Brain Sciences*, 32, 69-120.
- Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2009). Learning styles concepts and evidence. *Psychological Science in the Public Interest*, 9 (3), 105-119.
- Pélissier, M.-C., & O'Connor, K. P. (2002). Deductive and inductive reasoning in obsessive-compulsive disorder. *British Journal of Clinical Psychology*, 41, 15-27.
- Pinker, S. (1994). *The language instinct: How the mind creates language*. New York: William Morrow and Company.
- Sadock, B. J., & Sadock, V. A. (2007). *Kaplan and Sadock's synopsis of psychiatry* (tenth edition). Philadelphia: Lippincott Williams & Wilkins.
- Sampson, G. R. (2001). *Empirical linguistics*. London: Continuum.
- Scheibe, Erhard. (2001). Between rationalism and empiricism: The path of physics. In Brigitte Falkenburg (Ed.) *Between rationalism and empiricism: Selected papers in the philosophy of physics* (pp. 69-86). New York: Springer-Verlag.
- Shapin, S. (1996). *The scientific revolution*. Chicago: University of Chicago Press.
- Shermer, M. (2002). In *Darwin's shadow: The life and science of Alfred Russel Wallace: A biographical study on the psychology of history*. Oxford: Oxford University Press.
- Skipper, R. A., & Millstein, R. L. (2005). Thinking about evolutionary mechanisms: Natural selection. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2), 327-347.
- Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23, 645-726.
- van Dongen, J. (2010). *Einstein's unification*. Cambridge, UK: Cambridge University Press.
- Westfall, R. S. (1971). *The construction of modern science: Mechanisms and mechanics*. Cambridge, UK: Cambridge University Press.
- Wilkinson, W. K., & Migotsky, C. P. (1994). A factor analytic study of epistemological style inventories. *Journal of Psychology*, 128(5), 499-517.
- Wole ski, J. (2004). Analytic vs. synthetic and a priori vs. a posteriori. In I. Niiniluoto, M. Sintonen & J. Wole ski (Eds.) *Handbook of epistemology* (pp. 781-839). Dordrecht: Kluwer Academic Publishers.

# Process of Improvisational Contemporary Dance

**Yuko Nakano (qq096220@iii.u-tokyo.ac.jp)**

Graduate School of Interdisciplinary Information Studies, University of Tokyo  
Tokyo 113-0033, Japan

**Takeshi Okada (okadatak@p.u-tokyo.ac.jp)**

Graduate School of Education &  
Interfaculty Initiative in Information studies, University of Tokyo  
Tokyo 113-0033, Japan

## Abstract

The purpose of this study is to investigate the process of improvisational contemporary dance. To achieve this goal, we combine two types of methodology: analysis of data from interviews with dancers and analysis of their dance performances. Our findings reveal that while dancing, in order to create their movements, improvisational dancers interact with various stimuli that come from inside and outside of themselves (for example, images and feelings that they entertain during their dancing, and the music, space and audience of their dance performances). Through such interactions, dancers organize movements in their performances extemporarily, using various expressive techniques (for example, changing speed or image intentionally and seeing themselves from the viewpoint of a third person).

**Keywords:** Improvisation; Contemporary dance; Creativity

## Introduction

What happens in artists' minds when they generate their expressions or works? Recently research interest in artistic creation has been growing. There have been some empirical studies in psychology and cognitive science focused on artistic creation (e.g., Yokochi & Okada (2005) on Chinese ink painting, Tsuchikura (2010) on movie making, Tayanagi (2010) on jazz music, and Goan & Tujita (2007) on stage direction). Of the various forms of art, this study seeks to reveal the process of artistic creation empirically by focusing on dance, in particular improvisational contemporary dance.

Improvisation plays a critical role in the process of artistic expression. Sasaki, a scholar of aesthetics, has pointed out the importance of improvisation as follows: "In the exact moment of their generation, expressions in the artistic genres, such as fine arts, music, and drama are not generated based on predetermined plans, but based on impromptu activities" (Sasaki, 1995). Thus, it can be considered that improvisation exists as an essential part of the process of artistic expression.

Improvisational dance is one form of artistic expression in which such a process of improvisation saliently appears. Because improvisational dance presents audiences with the process of dance creation itself, making dance movements based on predetermined plans or repeating movements during dance performance is meaningless. Therefore, the

process of improvisational dance contains the essence of artistic creation and expression.

Previous studies in the domain of dance have pointed out the importance of improvisation in the creation of dance works, and the importance of impromptu expressions themselves (Fukumoto, 2007; 2009; Hosokawa, 2011; De Spain, 1997; Ribeiro & Fonseca, 2011; Soma & Hosokawa, 2007; Tsujimoto, 2010). Reviewing previous studies focused on how historically eminent dancers used improvisation to express themselves or create their dance works, Tsujimoto (2010) emphasized the significance of improvisation in dance. Tsujimoto (2010) also described an anecdote that when dancing extemporarily, the dancer's body instantly responded to stimuli from dance partners, the surrounding environment, and sensations born within the body. Through reviews of previous studies, Ribeiro & Fonseca (2011) also reported dancers' experience while dancing extemporarily. They suggested that improvisational dance is formed by "the interaction of the body with the environment and the affective and cognitive systems" (Ribeiro & Fonseca, 2011). Although these studies offer useful insights, these findings are based on reviews and anecdotal evidence, rather than on empirical evidence. They did not empirically investigate what stimuli dancers interact with in the process of improvisation or how dancers create their works.

In contrast, using information from reviews of a dancer's works, interviews with her, and field experiments, Hosokawa (2011) studied the process and skills of improvisational dance, focusing on Kei Takei, a famous contemporary dancer. On the basis of the results of the study, Hosokawa (2011) suggests that Kei Takei generates movement sequences intuitively based on her physical skill, and polishes movement sequences to develop her dance work in keeping with her dance philosophy. De Spain (1997) constructed a theoretical model about the process of solo dance improvisation through interviews and introspective reports with improvisers of dance. De Spain's model broke down the elements of improvisation into the categories of physical operands/operators, cognitive/affective operands/operators, determination, and attention. "The model also emphasized the importance of considering the existing state and flow of the improvisation at the moment of new action/interaction" (De Spain, 1997). Although Hosokawa (2011) and De Spain (1997) revealed

the process of improvisational dance, these studies focused only on the case of solo dance and their results are based mainly on verbal evidence such as interviews. In order to fully understand the process of improvisational dance in reality, it is necessary for us to analyze dance performance data in addition to interview data. Since the main medium for dancers to express what they want to express is their bodies, it is essential for researchers to analyze improvisational dance performances to understand how such a medium is used effectively. Therefore, we investigate the process of improvisational contemporary dance combining two types of methodology, analyses of data from interviews with dancers, and analyses of their dance performances.

Hence, this exploratory study aims at empirically investigating the improvisational process of contemporary dance by solo and duo dancers. Specifically, this study examines the relationship between dancers' internal processes and their dance performances, through: (1) interviews with dancers about what they pay attention to

while improvising; (2) fieldwork analyses of actual dance performances with introspective reports by the dancers.

## Study 1 Interviews with Dancers

In this interview study, we investigate what dancers think when they are dancing extemporarily.

### Methods

**Participants.** Ten professional contemporary dancers who have experience in improvisational dance participated in this study. (We use aliases A to J for these participants.)

**Procedure.** We conducted semi-structured interviews with the dancers about what they think when they are dancing extemporarily. The interviews were held between August and October 2010, with each interview lasting 90-120 minutes. All the interviews were recorded with a digital audio recorder, and the first author additionally took notes.

We asked the dancers about what they were thinking and what they valued when they were dancing extemporarily by

Table 1: Categories and definitions of dancers attention while dancing extemporarily

Category			Definition
Interaction with oneself	Internal experience	Message	To dance being aware of the message
		Feeling	To express the feeling that the dancer entertains while dancing
		Images	To dance retaining images that come into the dancer's mind while dancing
	Physical experience	Physical sensation	To enjoy the physical sensations that spring forth from the dancer's own body while dancing
Interaction with the outside	Stimuli from outside	Music	To dance being inspired by the music
		Space	To dance being inspired by the space
		Audience	Dialogue with the audience to touch their hearts or to surprise them
		Other dancers	To dance being inspired by other dancers
Means of expression	Confidence in oneself	Trust in inner experience	To move based on what the dancer feels without applying preformulated dance movements
		Trust in physical experience	To move following one's movements without conscious thought
	Expressive techniques	Switching	To change speed, rhythm, image or texture of movement intentionally
		Development	To develop movements or stories
		Seeing oneself from the viewpoint of a third person	To be conscious of how the dance looks to the audience
		Coordination with other dancers	To move together with other dancers, similarly, contrastingly or separately
	Personal decisions	Choice	To choose the most appropriate expression based on one's own feeling and the environment
		Continuation	Continue the current movement until satisfied

themselves and with other dancers. All the interviews were conducted in Japanese.

**Preparation for data analysis.** We analyzed the data by adapting the KJ method, a method to analyze the qualitative data. In detail, using the following procedure, the recorded interviews were prepared for analysis: 1. Transcription; 2. General understanding of the contents; 3. Identification of statements regarding attention or behaviour during extemporary dance; 4. Labelling of the statements with respect to dancers' attention or behaviour during extemporary dance; 5. Generation of categories by gathering similar labels together; 6. Consideration of the relationships between categories. We segmented the statements based on the speakers' turns, and coded all the segmented units based on what the dancer paid attention to while dancing extemporarily. In some cases, multiple codes are used for a single unit. Later, we assembled categories with similar meaning to create final categories.

During the coding process, several researchers and graduate students majoring in cognitive science checked the validity of the categories and revised inappropriate parts of labels and categories made by the first author.

## Results and Discussion

Table 1 is a collection of categories based on the interview data relating to what dancers pay attention to while dancing extemporarily. The statements were divided into three major categories, [interaction with oneself], [interaction with the

outside] and [means of expression]. Additionally, [interaction with oneself] was divided into <<internal experience>> and <<physical experience>>. The categories under <<internal experience>> are <message>, <feeling> and <images>. The category under <<physical experience>> is <physical sensation>. [Interaction with the outside] includes <<stimuli from outside>>, <music>, <space>, <audience> and <other dancers>. Topics associated with [means of expression] include the following three subcategories, <<confidence in oneself>>, <<expressive techniques>> and <<personal decisions>>. Subcategories under <<confidence in oneself>> are <trust in inner experience>, <trust in physical experience>; those under <<expressive techniques>> are <switching>, <development>, <seeing oneself from the viewpoint of a third person>, <coordination with other dancers>; those under <<personal decisions>> are <choice> and <continuation>. Table 2 shows which dancers referred to each category. In improvisational dance, dancers first pay attention to [interaction with oneself] and [interaction with the outside] and then, using [means of expression], they create a dance performance based on information from these interactions. Thus, when a dancer is dancing extemporarily, s/he interacts with factors such as <message>, <feeling>, <images>, <physical sensation>, <music>, <space>, <audience>, or <other dancers>. By combining this interaction with <<confidence in oneself>>, s/he creates physical movements as a dance performance.

Table 2: Dancers who referred to each categories

Category / dancers	A	B	C	D	E	F	G	H	I	J	total
Message	○	○		○	○		○	○		○	7
Feeling		○		○	○				○		4
Images									○		1
Physical sensation		○				○	○		○		4
Music		○	○	○	○	○	○	○	○	○	9
Space		○		○	○		○	○			5
Audiences			○		○		○	○	○		5
Other dancers	○	○	○	○	○	○	○	○	○	○	10
Trust in inner experience	○		○	○				○	○		5
Trust in physically experience	○	○	○	○	○	○	○	○	○	○	10
Switching		○					○	○	○		4
Development			○			○			○	○	4
Seeing oneself from the viewpoint of a third person	○	○	○	○	○	○	○	○	○	○	10
Coordination with other dancers	○	○	○	○	○	○	○	○	○	○	10
Choice									○		1
Continuation						○					1



Dancers choose such interactions flexibly by listening to their feelings and responding to their surroundings. Dancers continue the interaction that they chose until they are satisfied with their own decisions. Thus, in improvisational dance, dancers pay attention to many things that normally go unnoticed (their own feelings, physical sensations, the music, space, other dancers etc.) and they create dance movements by responding carefully to these.

In order not to let the interaction cease and so to create dance works that are fixed at a certain time and place, dancers use expressive techniques, such as <switching>, <development>, <seeing oneself from the viewpoint of a third person> and <coordination with other dancers>. Improvisational dancers present to audiences the process of dance creation as a dance work. Therefore, in addition to producing movements, it is also necessary to organize them into dance pieces.

In other words, while dancing, in order to create their dance, improvisational dancers interact with various stimuli that come from inside and outside of themselves (for example, images and feelings that they entertain during their dancing, and the music, space and audiences of their dance performances). Dancers use these interactions for their expression by responding to them sincerely and carefully. Through such interactions, dancers organize movements in their performances extemporarily, using various expressive techniques (for example, switching; changing speed or image intentionally and seeing themselves from the viewpoint of a third person).

The next important questions are how much interaction occurs and with which kind of stimuli in improvisational dance, and how these interactions affect the development of dance. In order to answer these questions, Study 2 examines dancers' internal processes and behavioural processes through introspective reports and analyses of performances.

## Study 2 Field Experiment

What exactly do dancers think and feel while dancing in a performance? In Study 2, we conducted a field experiment to capture the relationship between the introspective reports mentioned above and actual dance performances. The analysis of reflection focuses on which stimuli dancers pay attention to during the performance. Also, in the case of duo dance performance, a dance partner is an additional stimulus in the circumstance. We will analyze how a dance partner's movements affect the dancers' movements in duo dance performance.

## Method

**Participants and procedure.** We conducted a field experiment to analyze the improvisational dance performance of two expert dancers (H and M, both with more than 10 years of dance experience) on 5 November, 2010. After having them dance solo performances, we asked them dance together. We recorded the performance with four video cameras. Additionally, directly after the performance, we showed the video of the performance to

the dancers and asked them to reflect on how they felt, thought and moved during the performance.

The same music was used for both solo and duo performances. The theme H chose for her dance was, "Did you know that manifesto means to leave a handprint? To leave a handprint", and the theme M chose for his dance was "Inside a framed picture". H's solo ran 4 minutes 18 seconds. M's solo ran 4 minutes 17 seconds. Their duo piece ran for 4 minutes 30 seconds. In the duo performance, the dancers each maintained their own theme from their solo work. The only difference this time was that they had to dance with a partner who was expressing a different theme.

### Preparation for data analysis.

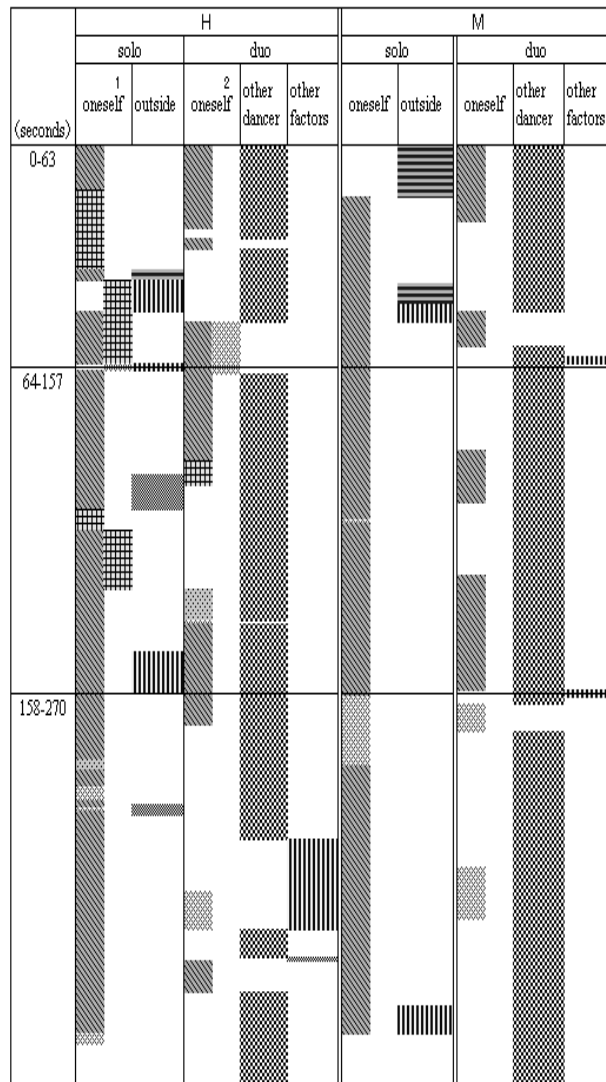
**Introspective reports:** After transcribing the introspective reports, we coded the statements based on what kind of stimuli and interactions occurred using the categories from Study 1.

**Performance analysis:** In order to objectively analyze the dance performance, we first focused on the vertical change of the centre of gravity of the dancers. Specifically, we coded each video frame (30 frames = 1 second) focused on the centre of gravity (1 lying down, 2 sitting, 3 kneeling, 4 standing, 5 stretching up, 6 jumping). We calculated the average score for the change in the centre of gravity, by dividing the score for the successive changes by the total time (in seconds).

## Results and Discussion

**Introspective reports.** After transcribing all of the introspective reports, we identified the contents of the performances that the verbal reports indicated. Then we identified which stimuli the dancers interacted with. We coded the data according to the categories used in Study 1. Because it was difficult to distinguish between "images" and "message", we coded the statements relating to both of these terms as "theme". Therefore, the categories for the statements are "theme", "physical sensation", "feeling", "dance technique", "music", "other dancers", "audience", and "space". Figure 1 shows how the objects of the dancers' attention change throughout the performance second by second. The columns in Figure 1 indicate "oneself" ("theme", "physical sensation", "feeling", "dance technique") and "outside" ("music", "audiences", "space") in the case of solo performance and "oneself" ("theme", "physical sensation", "feeling", "dance technique"), "other dancer", and "other factors" ("music", "audience", "space") in the case of duo performance.

The results show that H paid attention to "theme" in 81% of the solo performance time, and M did so in 86%. Therefore, it is clear that among the various stimuli, interaction with "theme" plays a central role in the development of the solo performance. Meanwhile, in duo performance, 80% of H's performance and 94% of M's performance were focused on the "other dancer", indicating that attention to the "other dancer" plays a central role in the development of the duo performance.



1: Whether the lines are right or left has no significance. This is only because the dancers referred to more than two categories at the same time.  
 2: Whether the lines are right or left has no significance. This is only because the dancers referred to more than two categories at the same time.

pattern	category
	theme
	physical sensation
	feeling
	dance technique
	music
	other dancer
	audience
	space

Figure 1: Change per second for the different categories

Table 3 : The average values for the position of the centre of gravity per second

	H		M	
	solo	duo	solo	duo
Part 1				
Sum of changing values	66	53	12	44
Duration of music (s)	64	64	64	64
<b>Average value</b>	<b>1.03</b>	<b>0.83</b>	<b>0.19</b>	<b>0.69</b>
Part 2				
Sum of changing values	43	45	45	54
Duration of music (s)	95	95	95	95
<b>Average value</b>	<b>0.45</b>	<b>0.47</b>	<b>0.47</b>	<b>0.57</b>
3 part				
Sum of changing values	53	91	41	44
Duration of music (s)	99	111	98	111
<b>Average value</b>	<b>0.54</b>	<b>0.82</b>	<b>0.42</b>	<b>0.40</b>

**Performance Analysis.** The music used in this performance was composed in such a way as to be easily split into three distinct sections. Therefore, by dividing the overall performance into three sections, we checked the change in the dancers' centre of gravity. The first section of the music (therefore the first section of the dance performance) took 64 seconds and the second took 95 seconds. The third section of the dance performance was determined by the time when the dancers finished the dance. The third section of H's solo took 99 seconds, and M's took 98 seconds. That of the duo performance took 111 seconds. The results of the centre of gravity analysis are shown in Table 3.

During the first section of the duo performance, the dancers somewhat adjusted their own movements to their partner's. H's score for the change in the centre of the gravity was higher (1.03) in the solo performance, and became lower (0.83) in the duo performance. M's score changed in the opposite direction (from 0.19 to 0.69). This result is consistent with the introspective data. According to the introspective reports, when paying attention to the dance partner, dancers sometimes adjusted their dance movements to the partner's and sometimes created a contrast during the first section.

In the second section, the two dancers' scores for the centre of the gravity were similar (0.47 for H and 0.57 for M). The introspective report is consistent with this finding

in the sense that both dancers thought that they were following their partner's movements. There was also little difference in the scores between solo and duo performances, which may be attributed to the slow music in this section.

In the third section, H's score was 0.82, while M's was 0.40. In the introspective report, each of the dancers stated that they had focused on the stimuli that s/he wanted to interact with, while paying attention to her/his dance partner. In particular, H was more influenced by the interaction with the music and the audience, while M interacted more with his own feelings. The fact that both of the dancers were concentrating on their own dance in this section seems to explain the difference between the dancers' scores.

These results suggest that dancers' dance movements in the duo performance are affected by the interaction with the other dancer's movements, as well as other stimuli such as their own feelings and music.

### General Discussion

In this study, based on interviews and introspective reports, we find that dancers take interactions with various stimuli seriously when improvising dance. Paying attention mainly to their own theme in solo performance and to the other dancers in duo performance, dancers use expressive techniques to construct an improvisational dance performance. In addition, from the performance analysis in Study 2, we find that dancing with another dancer changes a dancer's movements. From the data on changes in the centre of gravity, it is shown that other stimuli also affect the dance performance.

These findings are consistent with the claim by dance scholars Tsujimoto (2010) and Amagasaki (2004) that when dancing extemporarily, the dancer's body instantly responds to stimuli from dance partners, the surrounding environment and sensations born within the body. This study offers empirical evidence for such an anecdotal statement. In other words, through analyses of both verbal and performance data, this study reveals dancers' interaction with the internal process and the external environment in actual solo and duo performances.

This research also reveals that dancers use various expressive techniques to organize their dance movements as a dance piece. These findings are consistent with the claim by the dance scholar, Hosokawa (2011). However, this study offers concrete examples of these expressive techniques. Since there have not hitherto been any empirical studies focused on improvisational dance, our research makes a contribution to this field of research. Our findings suggest that new artistic expressions beyond a dancer's own repertory are born through interactions with the various stimuli at the actual moment of creation.

### Acknowledgments

We would like to express our gratitude to all the participants for having spent so much time on the interviews and dancing. We also received many insightful suggestions from

them on this research through discussions on dance and through dancing together.

### References

- Amagasaki, A. (2004). *Dance critique: Buyou no genzai / buyou noshintai. [Dance critique: Dance in the present day / the body in dance]* Tokyo: Keiso Shobo
- Fukumoto, M. (2007). The conception of "disorientation" in contact improvisation. *Journal of the Graduate School of Humanities and Sciences, Ochanomizu University*, 9, 51-60.
- Fukumoto, M. (2009). Studying on "arriving" in contact improvisation. *Annals of the Faculty of Art and Design, University of Toyama*, 3, 136-147.
- Goan, M. & Tujita, K. (2007). System dynamics on the creation of drama making processes. *Cognitive Studies*, 14, 509-531.
- Hosokawa, E. (2011). Buyou niokeru sokkyou kara sakuhinsousaku heno tenkai: Gendaibuyouka Kei Takei(1939-) wo jirei toshite [The process of dance creation using improvisation: A case study of Kei Takei]. In Y. Endo, E. Hosokawa, M. Takano, & Y. Uchikoshi (Eds.), *Buyougaku no genzai: Geijutsu, minzoku, kyouiku karano approach [The present dance study: Approaches from art, ethnology, and education]* (pp.77-98). Kyoto: Bunrikaku
- De Spain, K. S. (1997). Solo movement improvisation: constructing understanding through lived somatic experience. Diss. Temple University
- Ribeiro, M. & Fonseca A (2011). The empathy and the structuring sharing modes of movement sequences in the improvisation of contemporary dance. *Research in Dance Education*, 12, 71-85.
- Sasaki, K. (1995). *Dictionary of aesthetics*. Tokyo: University of Tokyo Press
- Soma, H. & Hosokawa, E. (2007). A study about the role of improvisation during the working process as to make a contemporary dance piece: From the case of a choreographer from Israel. *Annals of the Faculty of Education*, 56, 209-224.
- Tayanagi, E. (2010). Performance design and innovation in music: A case of improvisation and innovation in modern jazz. *Cognitive Studies*, 17, 459-473.
- Tsuchikura, E. (2010). The significance of plans in the creative process: A field study of the movie shooting. *Cognitive Studies*, 17, 713-728.
- Tsujimoto, S. (2010). Sokkyou surushintai. [The body in the improvisation]. In T. Kurihara, K. Yahagi, & S. Tsujimoto (Eds.), *Kuukan to katachi ni kannou surushintai [The body sympathizing with spaces and shapes]* (pp.63-88). Sendai: Tohoku University Press
- Yokochi, S. & Okada, T. (2005). Creative cognitive process of art making: A field study of a traditional Chinese ink painter. *Creativity Research Journal*, 17, 241-255.

# Learning Containment Metaphors

Sushobhan Nayak (snayak@iitk.ac.in)

Amitabha Mukerjee (amit@iitk.ac.in)

Indian Institute of Technology Kanpur

Kanpur, UP 208016 India

## Abstract

We present a computational approach that traces the developmental process, from containment image schemas to metaphors, in four phases: a) perceptual discovery of image schemas, b) associating perceptual arguments and the relation with linguistic units, c) discovering a linguistic structure encoding the schema, and finally d) enriching the semantics of the schema via extended language usage (via a corpus). In the first three phases, we use no prior knowledge about either the perceptual or language domains; in the corpus analysis, we use the WordNet ontology. Our input is an animation based on the Heider-Simmel video, together with a small corpus of transcribed commentaries. From the image sequence, we cluster the visual angle subtended by a landmark, and find that one cluster reflects containment. This is then correlated with the sentences from the adult commentaries uttered contemporaneously with containment situations, yielding strong object-nouns and relation-preposition associations. For discovering linguistic constructs, we use no knowledge of grammatical category or syntax but find recurring patterns using the approach of (Solan, Ruppin, Horn, & Edelman, 2002). Knowing the units involved, we can identify several phrasal patterns (e.g. “X moved into”, “in the Y”). We then search a corpus with the “in the Y” schema to identify container words. We find that the most common class involving containment is location (66%), followed by group membership (20%), time, and cognition (17% each). These may be thought of as language-based non-spatial enrichments for the image schema.

**Keywords:** containment metaphor; grounded concepts; selectional preference

## Introduction

Containment metaphors arise in infancy and may help organize the adult conceptual system (Mandler, 2010). The earliest structures (*image schema*) may arise initially in perception, but are then enriched by language in several ways, and extended to various non-spatial categories. Thus, a sentence such as “I put a lot of energy *into* washing the windows.” reflects the schema ACTIVITIES ARE CONTAINERS in the influential Conceptual Theory of Metaphor (CTM) (Lakoff & Johnson, 1980).

While such extensions of the initial spatial schema become conventionalized in a linguistic group, they retain the grounding. So while starting with the final text is not very useful, the grounded interpretation gives it much more flexibility. A computational study of the process would a) suggest mechanisms for understanding this process, and b) may itself be useful computationally - e.g. by providing an interpretation via simulation using the original grounding.

While much computational work has been done on metaphors, there appears no work that attempts such a vertical sweep from the initial perceptual schema to a language corpus. The emphasis within the NLP paradigm has been

on identifying and analyzing metaphors. The earliest *rule-based* attempts - e.g. (Fass, 1991) were based on hand-coded knowledge and metaphors were identified as a violation of selectional restrictions in a given context (e.g. “my car drinks gasoline”).

Other approaches use syntactic and co-occurrence statistics across large corpora to identify metaphors. We may call these attempts *corpus-driven*; work here may include Shutova, Sun, and Korhonen (2010) who demonstrate metaphor paraphrasing using noun-verb clustering, or Kintsch (2000) who effectively uses Latent Semantic Analysis to interpret metaphors like “My lawyer is a shark”. Cormet (Mason, 2004) is able to find mappings given separated datasets for two domains, e.g. it finds LIQUID  $\rightarrow$  MONEY once provided with LAB and FINANCE specific corpora to train from. Corpus-based approaches keep the metaphor mapping implicit, i.e. while the system can identify many metaphorical usages, the source domain has no grounding. Even distinguishing source from target domain is difficult, e.g. TIME co-occurs more often with SPEND than MONEY. Also, due to a primary reliance on verbs, it becomes difficult to treat ontological metaphors like CONTAINER that are more preposition dependent. Most importantly, purely linguistic approaches are hard to extend - e.g. container metaphors may invoke other attributes of the schema (e.g. ‘stir excitement’, or “the idea jumped out”).

A third category of work, which we may call *embodied modeling* (Narayanan, 1997), is more cognitively motivated. A model is learned from a tagged training set simulating pre-motor cortical representation of movement (Bailey, 1997); this is then mapped to other domains to interpret metaphoric usage such as “India releases the stranglehold on business”. The embodied approach is appealing and elegant, but is hard to scale up because new training data for learning the schema have to be manually created, and the syntactic structures require knowledge of the language.

In this work, we present a grounded model where initial image schemas are discovered from untagged video, and are then associated with textual commentary without using any prior language models. We focus on n-gram models, and on discovering merged paths through the sentence graphs (Solan et al., 2002). Once we have a basic construct, we can enrich the schema by exposing the learner to lots of language situations, which is simulated here by considering the 1-million word Brown corpus.

## Motivation

This work combines ideas of metaphorical extension from the seminal work of Lakoff and Johnson (1980) together with the

developmental ideas from Mandler (2010). Both suggest a strong role for spatiality in adult conceptual structures. Containment is discriminated by infants by the age of 2.5 months, and becomes “accessible” by 5.5 months, when it is used for multiple activities including visual and manual exploration (Spelke & Hespos, 2002). This may imply the presence of a mental structure incorporating arguments like a container and at least one trajector, and a function that given a configuration, accepts it as an instance of containment. This structure, which may be called an initial image schema, is eventually mapped to language, when containment is acquired before support (IN before ON before UNDER). This acquisition reflects an awareness not only of the preposition, but also for the linguistic argument structure that maps the image schema. But after this point, linguistic usage adapts the concept in ways that are specific to the linguistic-cultural context (Spelke & Hespos, 2002). Extensions emerge involving new structures that transfer the relationship to new domains, not only in language, but also in thought. Over increasing exposure, many of these extensions become conventionalized, many of which are listed in the CTM corpus.

To get a baseline check, we compiled 85 containment metaphors from Lakoff and Johnson (1980); of these, 65 involve prepositions in/into/out (IN a lot of trouble, INTO the century); the remaining 20 involve verbs explode, erupt, fill or adjectives full, empty - which profile other aspects of the containment schema (fullness, enclosed-ness etc).

Given this picture of the metaphor acquisition, extension, and conventionalization process, our goal is to try to model this embodied developmental process computationally, right upto the point where language affects and changes the image schema. In this process, we would like to minimize the domain knowledge available to the system; we assume only a large set of statistical learning tools, and a preference for smaller explanatory structures. Of course, we cannot model many important factors like social, interactive aspects.

The next section focuses on how an uninformed agent, with a capability for statistical learning, may acquire the containment schema as a cluster in its sensory space. The following sections discuss the discovery of linguistic units (and  $n$ -grams), the discovery of linguistic constructions associated with containment, and finally, the mapping to a large corpus.

## Learning Containment from Perception

Linguistic concepts are cognitively characterized in terms of *image schemas*, which are schematized recurring patterns from the embodied domains of force, motion, and space (Langacker, 1987; Lakoff & Johnson, 1980). The precise structure of an image schema remains quite unclear, with different authors using differing characterizations. In this work, we take an image schema to consist of two related structures. First is the list of arguments which participate in the associated relation or activity. The other is a characterization of the situation in terms of a function defined over some feature space, so that situations satisfying this function may

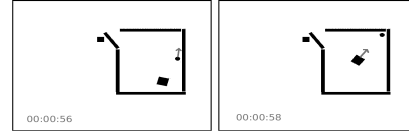


Figure 1: *Multimodal input: 2D video “Chase”*: Three shapes, big-square([BS]), small-square([SS]) and circle([C]) interacting with each other and the static box([box]).

be considered as instances of the image schema. We wish to learn such an image schema given a simple video as input (Figure 1), where three objects - a big square ([BS]), a small square ([SS]) and a circle ([C]) are moving around, interacting with each other and going in and out of a static [box] via a **door**. Though the objects deform a bit while rotating, and also occasionally overlap, it is relatively straightforward to segment them.

The linguistic database consists of a co-occurring narrative with 36 descriptions of the video. These narratives exhibit a wide range of linguistic variation in terms of linguistic focus, lexical choice and construction. In an earlier work in learning prepositions and nouns from the same multi-modal data, we used dynamic bottom-up attention to correlate objects seen in the video with their linguistic counterparts (*nouns*) (Mukerjee & Sarkar, 2007). In this work, we further consider the subset of utterances which have a temporal overlap with the frames during which a containment situation prevails.

## Acquiring Containment Prepositions

In spatial reasoning, there have been several attempts at defining spatial relations involving continuum measures defined over different geometric features on object pairs. Regier (1996), a seminal work in preposition grounding, uses angle measures and a connectionist network to correlate videos and prepositions. The work, however, is limited in the sense that Regier uses videos annotated with single words like IN, OUT, THROUGH etc. while we hope to learn these schemas by clustering the untagged video. Also, because his videos are tagged with prepositions, he never has to work to discover the preposition; we have to discover these units from the unconstrained unparsed narrative. Mukerjee and Sarkar (2007) use the same dataset as ours, but use a measure based on visual proximity - the *Stolen Voronoi Area* - to cluster space using Kohonen Self Organising Maps (SOMs). We initially tried these two approaches and find that in an unsupervised clustering task ( $k$ -means with 6 classes), these earlier models do not work well for distinguishing the inside and outside of irregular (L- or U-shaped) containers (1st row, Fig 2). In a supervised scenario they show good results training with sophisticated neural-nets over multiple epochs, but our goal is to try not to use supervision data.

Another feature implicated in place learning in animals is *visual angle* (Rolls et al., 1999) - the angle subtended by a landmark on the retinal image. We attempted to improve on the previous features by using a single feature - the total

angle subtended by a landmark at the object position. With this measure, we find that when the resulting feature space is clustered, one of the clusters works quite well for identifying the IN-schema. Computing this feature involves computing the angle that the landmark, **[box]**, would subtend at each point in the space; the result is measured and clustered using  $k$ -Means ( $k = 6$ ). We can see in Fig 2 (bottom row left) that one cluster completely covers what may be thought as the inside of **[box]**, whereas the the outside is graded between a number of clusters. If we accept this as a characterization for an image schema for containment, then the distribution of visual angle in this cluster will serve to represent this relation. To test whether this model, learned from the single **[box]** shape, really represents the *category* of containment relations, we generalize and evaluate it over a number of other shapes. The results of applying the same learned distribution to three novel shapes is shown in Fig 2(bottom row). We find that regions with varied levels of ‘IN-ness’ have been separately grouped, validating our choice of features. While for closed convex shapes the measure has a clear demarcation of ‘inside’(360° angle), it gives a more graded assessment for open figures as well, such as the open-top square in Fig 2(2nd row, 4th fig).

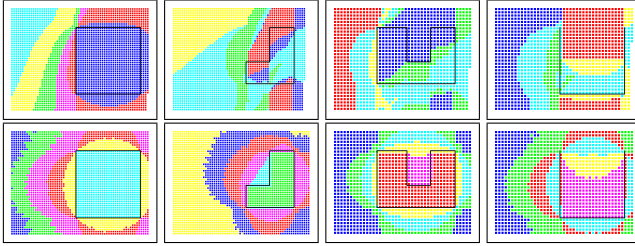


Figure 2: *k*-Means Clustering of Space. a) Voronoi and Angle features (top row) and b) Visual Angle feature (bottom row). The inside of all the containers has been clearly identified as a separate cluster only in the latter case.

At this stage, the system computes the visual angle subtended at an object position, for a landmark (the box or some other shape). The two arguments participating in this computation are the container and the trajector, though the system does not use these terms; these relationships are implicit in the feature computation. If the visual angle falls within the distribution associated with containment (the *IN-cluster*), it is accepted as an instance of this relation. Thus the system has both the arguments, and the acceptance function characterizing the image schema. This acquisition is pre-linguistic, from perceptual data alone. When a pre-discovered object, say **[BS]**, lies in IN-cluster, we have the argument structure **{[BS] IN [box]}** or **IN([BS],[box])**. After learning the unit IN, we can attempt to map this perceptual argument structure to linguistic syntactical structure, thereby discovering aspects of syntax.

## Linguistic Elements Describing Containment

We now detail the discovery of linguistic elements pertaining to containment through correlation of the commentary with the acquired schemata. We restrict our analysis to utterances occurring while **[BS]** is in the *IN-cluster* w.r.t. **[box]**. Since the commentaries are unconstrained and there is no syntactic information, every uttered word is a possible label. Given an uttered unit  $w_i$ , the probability that it refers to schema  $s_j$ , is given by:

$$P(s_j|w_i) = \frac{P(w_i|s_j)P(s_j)}{P(w_i)} \propto \frac{P(w_i|s_j)}{P(w_i)} = A_{ij}^r$$

This metric, the *relative association* ( $A_{ij}^r$ ), is prone to give erroneous results for infrequent units, while working well for high frequency words. For example, it gives an association value of 1 for a word that has been uttered only once in the whole commentary. To counter this trend, we also subscribe to mutual information between states  $s_j$  and words  $w_i$ , which eliminates the possibility of uninformative rare words being assigned a high score. The word-object association is then estimated using the product of mutual information of word  $w_i$  and state  $s_j$  with their joint probability,

$$A_{ij}^m = Pr(w_i, s_j) \log \frac{Pr(w_i, s_j)}{Pr(w_i)Pr(s_j)}$$

where  $A_{ij}^m$  is the *mutual association*. We use this measure because if  $W(= \cup_i w_i)$  and  $S(= \cup_j s_j)$  are two random variables then their Mutual Information  $I(W, S)$  would be

$$I(W, S) = \sum_i \sum_j Pr(w_i, s_j) \log \frac{Pr(w_i, s_j)}{Pr(w_i)Pr(s_j)} = \sum_i \sum_j A_{ij}^m$$

$A_{ij}^m$  is, thus, the contribution of each word object pair. The results are shown in Fig 3. Notice that *in*, *inside* and *into* emerge as the three dominant monograms (their frequencies in the containment subset are 28, 26 and 15 of 1100 words).

Before moving onto syntax discovery, we observe that the nouns corresponding to the three objects in the video, had been acquired earlier using an attentional correlation model (Mukerjee & Sarkar, 2007). These will be used in the next section: “big square” for **[BS]**, “little square” for **[SS]**, and “circle” for **[C]**.

## Deriving Syntactical Structure

Discovering syntactical structures of a containment preposition like IN will enable us to discover other labels for objects participating in containment. For example, if we know that only object  $A$  with label  $l_A$  is in container  $B$  with label  $l_B$ , the *perceptual* arguments of IN are **{trajector:A, container:B}**. Now, suppose at the same time we hear the utterance: ‘ $l_A$  goes into  $l_C$ ’. If we know the linguistic argument structure associated with IN, then we can have a high confidence from this single instance that  $l_C$  is most likely another label for the container  $B$ . Once internalized, this process would help the agent recreate context in a novel discourse, by simulating the action and

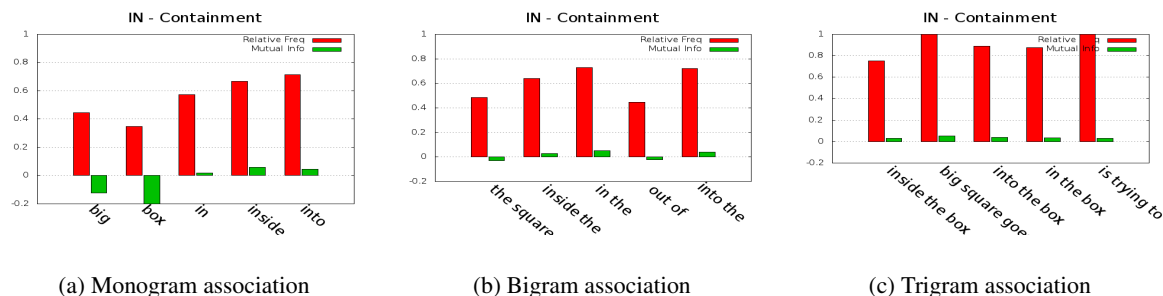


Figure 3: Association of words with the containment cluster. Both the association measures, as used previously for noun-label learning, are shown. The **red** bars indicate the *Relative Frequency* measure while the **green** bars are for the *Mutual Information* Measure.

identifying probable trajectors and landmarks via the syntactic argument structure. In the example above, if we don't know what 'goes' refers to, some idea of this may be formed by realizing that it is something that *A* and *B* may be participating in. This goes along with Siskind (1996)'s *constraining hypotheses with partial knowledge*: "When learning word meanings, children use partial knowledge of word meanings to constrain hypotheses about the meanings of utterances that contain those words."

We start our discovery of syntactic structure by analyzing bi- and tri-gram correlations, which are associated with the same metric as in the monogram case. We observe that the prominent bi-grams are `inside the`, `into the` and `in the`. The tri-grams that emerge are `inside the box`, `in the box` and `into the box`. These associations could help learn the label of not only the containment schema, but also the container itself. Note that the attention based model for learning nouns cannot learn the container/box, which is never dynamically salient. Thus, its label is not known. However it is prominent in these containment sentences, and discounting the frequent word `the` in trigrams such as "`{ inside | in | into } the box`", we may associate `box` with the **[BOX]**, treating it as a label for the container. We also note the presence of other fragments (`big square goes` and `is trying to`) as prominent trigrams; but these are present only in this context, and will be diluted as the agent looks at other containment situations. But despite some glimmers, the *n*-gram approach is not very illuminating regarding the construction encoding for containment.

A richer model of syntactic structure has been developed over many years by Edelman and his group, implemented as the tool ADIOS (Solan et al., 2002). In this approach, a graph called Representational Data Structure (RDS) is constructed from the morphologically segmented input sentences. It then repeatedly scans and modifies the RDS in an attempt to merge edges and come up with a more compact description of the input. In the process, a number of syntactic clusters and constructions are discovered without any prior knowledge of grammatical categories. Examples of some of the patterns

found are:

$$\left[ \begin{array}{c} \left[ \begin{array}{c} \textit{ball} \\ \textit{door} \\ \textit{box} \\ \textit{square} \end{array} \right] \\ \textit{the} \\ \left[ \textit{circle} \right] \end{array} \right] \rightarrow \left[ \begin{array}{c} \textit{move} \\ \textit{came} \\ \textit{got} \end{array} \right] \rightarrow \textit{into}$$

$$\left[ \begin{array}{c} \textit{in} \\ \textit{inside} \\ \textit{into} \end{array} \right] \rightarrow \textit{the} \rightarrow \textit{box}$$

Clearly, these structures are discovering some relevant patterns for containment, including the group `in | into | inside` which was also observed in the trigram model. The noun cluster - with `box` and `door` is also an impressive discovery. But most effective is the fact that `box` is identified as a label appearing after the IN while the trajector appears before.

One of the uses of this structure would be the discovery of synonyms. For example, consider the sentence `large square moves into the box`, which is being uttered as the agent perceptually knows that **[BS]** is moving into IN-cluster, activating schema `IN([BS],[box])`. By aligning this with the linguistic input, `large square` is learned as a referent for **[BS]**, i.e. a synonym for "big square". Also, an unit appearing frequently in the trajector position in the discourse is "it", which has no constant referent; it is possible now to discover that this is a unit which may refer to multiple objects. Further analysis may reveal that "it" is strongly correlated to the most recently observed trajector - and thus our system may begin its journey of understanding anaphora, another computationally promising domain in language.

## Metaphorical Mappings from Language Usage

We have alluded to learning metaphorical maps through association before. We are in a position where we have grounded concepts of agents taking part in an event, and the concept of containment, through a mapping for linguistic element IN. Consider the following occurrences:



1. What state is the project IN? IN-STATE (project, state)

2. He did it IN three minutes. IN-TIME (he, 3min)

Here, the abstract concepts of STATE, TIME etc. are understood in terms of a container, thanks to their syntactic association with linguistic instances of “IN a BOX (container)”, for which the learner has a physical basis. It’s therefore prudent to assume that metaphorical concepts would occur in similar lexico-syntactic environments in language usage. We have demonstrated the ability of the agent in discovering the ‘container’ through ADIOS; the question we would like to address now is, would it also be able to discover the object of containment in a novel context with sentences of myriads of different structure? We investigated it by running ADIOS on IN-containing sentences of Brown corpus. We find that it can, indeed, distinguish ‘containers’ from other elements of a sentence, as evidenced by the following pattern:

$$in \rightarrow the \rightarrow \begin{bmatrix} building & war & fight & car \\ group & death & woods & cellar \end{bmatrix}$$

While discovery of ‘container’ elements would be the first evidence of mapping of abstract concepts to the ‘container’, the mappings would be prominent only if further evidence abounds in language, so that learning due to false evidence is minimized. The *propensity* of a concept to be described as a ‘container’ can be gauged through the regularity of its occurrence in the object position of IN. In literature, **selectional preference (SP)**(Resnik, 1993) is used abundantly to measure regularities of a verb w.r.t. the semantic class (*subject, object* etc.) of its argument. It has been used previously for word-sense disambiguation(Resnik, 1993) and metaphor interpretation(Mason, 2004). While it has only been used for finding verb-preferences, we will adapt it to include prepositional preferences, so that we are able to learn containment metaphors. While verbs have different syntactic relations like *verb-object* or *subject-verb*, the prepositions we are considering, have only one relation to the trailing noun, that of Object of Preposition (*pobj*)(according to the Stanford Parser(Marneffe & Manning, 2006)). Therefore, the formulation from Resnik (1993) is slightly modified. The *selectional association* of class  $c$  for predicate  $p$ (IN) is defined as:

$$A(p, c) = \frac{1}{S(p)} P(c|p) \log \frac{P(c|p)}{P(c)}$$

where,

$$S(p) = D(P(c|p)||P(c)) = \sum_c P(c|p) \log \frac{P(c|p)}{P(c)}$$

and,

$$P(c|p) = \frac{freq(p, c)}{freq(p)} = \frac{\sum_{w \in c} count(p, w)}{freq(p)}$$

where  $count(p, w)$  is the number of time word  $w$  occurred, and  $classes(w)$  is the number of classes it belongs to.

Table 1: Selectional association strength of different classes

Class	SA	Class	SA
location	0.658	act	0.058
group	0.201	artifact	0.077
time	0.175	object	0.055
cognition	0.164	food	-0.030
state	0.145	animal	-0.042

WordNet (Feinerer & Hornik, 2011) is our knowledge-base for class  $c$ . WordNet was developed as a system that would be consistent with the knowledge acquired over the years about how human beings process language. To represent the early learner’s limited concept-repository, only top level classes of WordNet are considered. We use the Brown Corpus as the sample space to determine the selectional preferences. All the sentences involving the containment concepts, i.e. all 21,480 sentence-parts with words *in/into/inside* were extracted. The sentences with IN were converted to the functional form of IN(*pobj*) in a rather simple way: the first occurrence of a noun after IN in the tagged corpus was assigned to the concept. For example, the sentence fragment *into a hot cauldron* is converted to IN(cauldron).

The most occurring ‘container’ words were *world, way, order, years, case, states* etc. The resultant associations are shown in Table 1. Notice that Location class has the highest association for container schema, activating a LOCATIONS ARE CONTAINERS mapping. Group class also has a strong association to containers, representative of the notion that groups or teams are visualized as containers (*in a group, in a team*). Time, Cognition and State also show high associativity, while Food and Person class demonstrate a significantly negative mapping. The negative numbers only represent the weakness of mapping, and should not be treated as repudiating existence of the same. The association measures only demonstrate that some mappings are used more in language, and consequently, are stronger in our cognition than others. In fact, in the original metaphor list(Lakoff & Johnson, 1980), the most prominent mappings to container are those of Cognition(15%), State(14%), Location(7.3%), Group(8.6%), Time(5.4%) and Act(4.8%), somewhat representative of their strength acquired from the whole corpus. Similarly, the least occurring classes in the list too are Plant(0.3%), Animal(0.3%) and Food(1%). Some examples of sentences from both the Brown corpus and metaphor list are presented below:

- STATE/COGNITION AS A CONTAINER
  - Meredith began falling in love. (Brown)
  - We’re IN a mess. (Lakoff & Johnson, 1980)
- TIME AS A CONTAINER
  - We’re well into the century. (Lakoff & Johnson, 1980)
  - There comes a time in the lives of most of us when we want to be alone. (Brown)
- LOCATION AS A CONTAINER

- If you’ve travelled in Europe a time or two , ... (Brown)
- ...and begin to feel at home in the capitals of Europe... (Brown)
- ACT AS A CONTAINER
  - How did you get into window-washing as a profession? (Lakoff & Johnson, 1980)

## Conclusion

In this work, we proposed a plausible method for a primitive artificial agent with multi-modal input handling and feature extraction capability, to internalize linguistic concepts of containment. Containment metaphors are a primary way in which humans understand abstract concepts of state/emotion etc. and it’s therefore necessary for a cognitive system to be able to do so if it’s to acquire human level intelligence. Through a grounded system and selectional preference of IN, we also showed how the model can internalize conventional metaphors in vogue in English.

In the work, we have used some novel ideas, like that of modifying selectional preference to include prepositional bindings, which is, to our knowledge, unexplored in literature. Also, through simple methods of spatial feature extraction, unsupervised clustering and word-cluster association, we have been able to extract the idea of containment.

Nonetheless, this work can benefit from much improvisation. The image schema for containment that we provide, is a very crude one. While it’s possible to extend it to more general scenarios, as has been demonstrated, we haven’t investigated its limitations on non-convex shapes. That, however, hardly undermines our contribution here since in non-convex shapes, even adults find it difficult to separate regions into inside/outside, and it largely becomes an outcome of some context. Furthermore, our aim was to show that even with so meagre a sense-world (that of one video and associated commentaries), an artificial agent can get some semblance of human-like notions of containment and employ them to inculcate metaphorical mappings in its system. Instead of 81 seconds of learning, however, the human learner has days and months and years of exposure, and clearly this can lead to the construction of extremely rich and diverse schemata.

While we have handled concepts that easily fall into a container mould due to usage of IN, we have not been able to model other container metaphors that have different attributes, like the 20 out of our list of 85 that depend on verbs and adjectives related to substance. Simulating all that is indeed a Herculean task in this limited set-up. We intend to look into these aspects and also, to matters of finding more mappings using WordNet’s synsets (at present we are using only the lexical files – synset usage would lead to discovery of more maps, but it would also make the process noisy) in our future work.

## References

Bailey, D. (1997). *A computational model of embodiment in the acquisition of action verbs*.

- Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. In *Proc. 10th Annual Cogsci Conf.* (pp. 510–516).
- Fass, D. (1991). Met\*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1), 49–90.
- Feinerer, I., & Hornik, K. (2011). wordnet: Wordnet interface [Computer software manual]. Available from <http://CRAN.R-project.org/package=wordnet> (R package version 0.1-8.)
- Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, 6, 287–317.
- Kintsch, W. (2000, July). Metaphor comprehension: A computational theory. *Psychonomic Bulletin and Review*, 257–266.
- Lakoff, G., & Johnson, M. (Eds.). (1980). *Metaphors we live by*. University of Chicago Press.
- Langacker, R. (Ed.). (1987). *Foundations of cognitive grammar I: Theoretical prerequisites*. Stanford University Press.
- Mandler, J. M. (2010). The spatial foundations of the conceptual system. *Language and Cognition*, 2(1), 21–44.
- Marneffe, B. M. Marie-Catherine de, & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*.
- Mason, Z. J. (2004, March). Cormet: a computational, corpus-based conventional metaphor extraction system. *Comput. Linguist.*, 30, 23–44.
- Mukerjee, A., & Sarkar, M. (2007). Grounded acquisition of containment prepositions. In *Proc. ICON*.
- Narayanan, S. (1997). *Knowledge-based action representations for metaphor and aspect (karma)*. Unpublished doctoral dissertation, CS, UC Berkeley.
- Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. Bradford Books.
- Resnik, P. S. (1993). *Selection and information: A class-based approach to lexical relationships* (Tech. Rep.).
- Rolls, E., et al. (1999). Spatial view cells and the representation of place in the primate hippocampus. *Hippocampus*, 9(4), 467–480.
- Shutova, E., Sun, L., & Korhonen, A. (2010). Metaphor identification using verb and noun clustering. In *Proc. of COLING* (pp. 1002–1010).
- Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61, 39–91.
- Solan, Z., Ruppin, E., Horn, D., & Edelman, S. (2002). Automatic acquisition and efficient representation of syntactic structures. In *Proc. of NIPS*.
- Spelke, E., & Hespos, S. (2002). Conceptual development in infancy: The case of containment. *Representation, memory, and development: Essays in honor of Jean Mandler*, 223–246.

# Interaction of Word Learning and Semantic Category Formation in Late Talking

Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson

Department of Computer Science

University of Toronto

{aida,afsaneh,suzanne}@cs.toronto.edu

## Abstract

Late talkers (LTs) — children who show a marked delay in vocabulary learning — have also been shown to differ from normally-developing (ND) children with respect to the semantic organization of their learned vocabulary. We use a computational model of word learning to study how individual differences between LTs and NDs give rise to differences in abstract knowledge of categories emerging from learned words, and how this affects their subsequent word learning. Our results suggest that the vocabulary composition of LTs and NDs differ at least partially due to a deficit in the attentional abilities of LTs, which also results in the learning of weaker abstract knowledge of semantic categories of words.

## Introduction

Late talkers (LTs) are children with a marked delay in word learning at an early age, some of whom go on to exhibit specific language impairment (SLI). Early identification of LTs at risk for SLI is especially important, since early intervention can produce significant changes in the language development of these children (Desmarais, Sylvestre, Meyer, Bairati, & Rouleau, 2008). Many psycholinguistic studies have thus focused on understanding signs of late talking, as well as factors contributing to it (Paul & Elwood, 1991; Thal, Bates, Goodman, & Jahn-Samilo, 1997; Rescorla & Merrin, 1998; Ellis Weismer & Evans, 2002; Rowe, 2008; Stokes & Klee, 2009).

An important observation about late-talking children is that they seem to not only learn *more slowly* than their normally-developing (ND) peers, but also to learn *differently*. For example, the vocabulary composition of LTs shows greater variability, e.g., in terms of how consistently certain properties, such as shape, are associated with particular categories, such as solid objects (Jones & Smith, 2005; Colunga & Sims, 2011). More generally, the vocabulary of LTs has been shown to exhibit less semantic connectivity than that of NDs (Sheng & McGregor, 2010; Beckage, Smith, & Hills, 2010). The greater variability and the weaker connectivity in the vocabulary of LTs call for further investigation since they might be reflective of underlying cognitive deficits in these children.

Psycholinguistic evidence suggests that children’s word learning improves when they form some abstract knowledge about what kinds of semantic properties are relevant to what kinds of categories (Jones, Smith, & Landau, 1991; Colunga & Smith, 2005; Colunga & Sims, 2011). This abstract knowledge is argued to emerge by generalizing over the learned words. Stated otherwise, words that have been learned contribute to generalized abstract knowledge about word meanings and semantic categories, which then guide subsequent word learning. It is possible that because of the differences in the vocabulary composition of LTs and NDs, the two groups

of children also form different abstract knowledge of categories, which causes differences in their word learning (as suggested by Jones & Smith, 2005; Colunga & Sims, 2011).

We investigate this possibility by examining within a computational model the precise interaction between early word learning and knowledge of semantic categories of words. We do so by extending an existing model of cross-situational word learning (Fazly, Alishahi, & Stevenson, 2010). As in Nematzadeh, Fazly, and Stevenson (2011), we simulate the difference between ND and LT learners as a difference in the ability of the cross-situational learning mechanism to attend to appropriate semantic features for a word. Within this framework, we propose a new model that forms clusters of words according to their learned semantic properties, and that uses this knowledge in guiding the future associations between words and meanings. We show that the semantic clusters of words are qualitatively very different for our ND and LT models; moreover, the two learners exhibit striking differences in terms of the usefulness of their learned clusters for subsequent word learning. Through computational modeling, we thus suggest an interaction between the impaired ability of LTs to form informative abstract semantic categories, and the observed delay in their vocabulary acquisition.

## The Computational Model

### Overview of the Word Learning Model

The model of Fazly et al. (2010) is a cross-situational learner that incrementally forms probabilistic associations between words and their semantic properties. The input to a child is simulated as a sequence of utterances (a set of words), each paired with a scene representation (a set of semantic features, representing what is perceived when the words are heard):

**Utterance:** { *she, drinks, milk* }

**Scene:** { ANIMATE, PERSON, FEMALE, CONSUME, DRINK, SUBSTANCE, FOOD, DAIRY-PRODUCT }

Given such an input pair, the model adjusts its probabilistic representation of the meaning of each word. First, the model determines, *based on its current probabilistic knowledge of word–meaning associations*, which semantic features in the scene are more and less likely to be associated with each word in the utterance. Using that assessment of word–feature alignment in the current input, the model then updates its probabilistic representation of the meaning of each word.

In this way, the model uses cross-situational evidence to gradually improve its representation of the meaning of each word  $w$  as a probability distribution,  $p(\cdot|w)$ , over all semantic features: i.e.,  $p(f|w)$  is the probability of feature  $f$  being part

of the meaning of word  $w$ . At the heart of this process are two calculations which we briefly summarize here (see Fazly et al., 2010, for more detail). The *alignment probability* determines how strongly a word  $w$  and a feature  $f$  are associated in the current (multi-word) utterance  $U$  at time  $t$ , in proportion to the model’s current hypothesis of how likely the feature is part of the meaning of the word:

$$a_w^{(t)}(w|f) = \frac{p^{(t-1)}(f|w)}{\sum_{w' \in U_t} p^{(t-1)}(f|w')} \quad (1)$$

In order to collect this knowledge across all cross-situational uses of the word and feature, the model maintains an incrementally accumulated sum of these alignments that captures the overall strength of the association between  $w$  and  $f$ :

$$\text{assoc}^{(t)}(w, f) = \text{assoc}^{(t-1)}(w, f) + a_w^{(t)}(w|f) \quad (2)$$

The second key formula to the operation of the model is the *meaning probability* that uses the association scores to update the meaning of each word after processing an input pair:

$$p^{(t)}(f|w) = \frac{\text{assoc}^{(t)}(f, w) + \lambda(t)}{\sum_{f' \in \mathcal{M}} \text{assoc}^{(t)}(f', w) + \beta \cdot \lambda(t)} \quad (3)$$

where  $\beta$  is the number of expected distinct features,  $\mathcal{M}$  is the subset of those features that have been observed, and  $\lambda(t)$  is a smoothing function which we formulate in a way that captures the developing ability of the model to attend to input, as follows.

Research has shown that children’s ability to attend to relevant features of a perceived scene improve over time (e.g., Mundy et al., 2007). Moreover, LTs have been observed to show difficulty with the communicative abilities that enable children to direct appropriate attention on relevant aspects of a scene (e.g., Rescorla & Merrin, 1998). In recent work (Nematzadeh et al., 2011), we demonstrated that we can use the  $\lambda(t)$  function to simulate how quickly or slowly the attentional abilities of a learner develop over time. Specifically, the  $\lambda(t)$  function determines how much weight is given to unobserved word–feature pairs, with greater weight reflecting immature attentional skills in which the learner fails to focus on the observed (appropriate) meaning features. In the model,  $\lambda(t)$  is designed to decrease over time, to simulate gradually improving attentional processes that can appropriately focus on the observed word–feature pairs. We modeled the difference between ND and LT learners by having a  $\lambda(t)$  function for the latter that decreases much more slowly, corresponding to delayed development of appropriate attention to the input. Here we adopt that same formulation,<sup>1</sup> but extend the model as follows to consider the role of attention and its interaction with semantic category formation in word learning.

<sup>1</sup>Our ND and LT simulations here use the same settings for  $\lambda(t)$  as what we referred to as ND and LT<sub>5</sub> in our previous work.

## Learning Semantic Categories of Words

We extend the word learning model above by incorporating the ability to form clusters of words based on their learned semantics, and to use the resulting semantic categories in subsequent word learning.<sup>2</sup> These abilities represent a first step in integrating the model’s word learning with formation of conceptual categories. These extensions to the model are key to further examination of the cognitive mechanisms that might underlie the weaker semantic connectivity observed in the vocabulary of LTs. Specifically, while Nematzadeh et al. (2011) showed that learned words of their ND learner had greater semantic coherence than those in the LT learner, the model did not actually form semantic clusters of words, nor use semantic relations among words to help in word learning.

Our new model, at certain points in time (depending on the simulation), groups the words it has observed into clusters based on the similarity among their learned meanings. Given two words  $w$  and  $w'$ , we determine their degree of semantic similarity by treating their learned probability distributions over the semantic features,  $p(\cdot|w)$  and  $p(\cdot|w')$ , as input vectors to the cosine function. These cosine values guide the grouping of words using a standard unsupervised hierarchical clustering method. The clusters of semantically related words can then be analyzed to see how the factors that simulate ND and LT learners in the model contribute to different quality of semantic categorization, as observed by Sheng and McGregor (2010) and Beckage et al. (2010), among others.

Moreover, the semantic clusters enable us to build further on the explanation of late talking as arising from attentional differences in learners (as proposed in Nematzadeh et al., 2011). Specifically, we assume that learned semantic categories enable children to generalize their knowledge of related words, which can help focus subsequent word learning on relevant semantic features in the input. In our model, knowledge about the semantic category of a word can be used as an additional source of information about which semantic features are more likely to be aligned with the word in a given input. For example, features such as EDIBLE and FOOD should be more strongly aligned to a word referring to a kind of fruit than to a word referring to a kind of vehicle.

We achieve this in our model by aligning a word  $w$  and a feature  $f$  in an input utterance–scene pair according to both word-level and category-level information, the latter drawing on the incrementally created semantic clusters. We adopt the formulation used by Alishahi and Fazly (2010) to combine word and category information in the alignment probability:<sup>3</sup>

$$a^{(t)}(w|f) = \omega \cdot a_w^{(t)}(w|f) + (1 - \omega) \cdot a_c^{(t)}(w|f) \quad (4)$$

<sup>2</sup>We continue to refer to the clusters that our model learns both as *clusters*, to emphasize that they are learned in an unsupervised manner, and as *semantic categories*, to emphasize their connection to children’s knowledge of abstract categories.

<sup>3</sup>The approach of Alishahi and Fazly (2010) differs from ours: (1) They examine the role of syntactic categories (e.g., noun or verb) in word learning while we look at semantic categories. (2) They use predefined correct assignments of words to such parts of speech, but our clustering is based on the model’s learned semantic knowledge.

*apple*: { FOOD:1, SOLID:.72, ..., PLANT-PART:.22,  
PHYSICAL-ENTITY:.17, WHOLE:.06, ... }

Figure 1: Sample true meaning features & their scores for *apple*.

The first component of the above formula,  $a_w^{(t)}(w|f)$  is the word-based alignment, given in Eqn. (1) above. The second component,  $a_c^{(t)}(w|f)$ , is an analogous category-based alignment (described below). The  $\omega$  term is a weight (between 0 and 1) that determines the relative contribution of the two alignments; here we use a balanced weighting of 0.5.

Where the word-based alignment captures the association between a feature and a single word, the category-based alignment,  $a_c^{(t)}(w|f)$ , assesses the overall association between the feature  $f$  and the words in  $\text{cluster}(w)$ , the cluster assignment determined by the model for word  $w$ . This alignment is calculated by replacing occurrences of  $p(f|w)$  in Eqn. (1) with  $p(f|\text{cluster}(w))$ . We again follow Alishahi and Fazly (2010) in defining  $p(f|\text{cluster}(w))$  as the average of the meaning probabilities of the words in the cluster:

$$p^{(t)}(f|\text{cluster}(w)) = \frac{1}{|\text{cluster}(w)|} \sum_{w \in \text{cluster}(w)} p^{(t)}(f|w) \quad (5)$$

where  $|\text{cluster}(w)|$  is the number of words in the cluster.

## Semantic Representation in the Model

### The Representation of a Scene

The input data for our model consists of a set of utterances paired with their scene representations. As in Nematzadeh et al. (2011), the utterances are bags of lemmatized words, taken from the child-directed speech (CDS) portion of the Manchester corpus (Theakston et al., 2001, from CHILDES MacWhinney, 2000). The corpus is transcripts of conversations with 12 British children, ages 1;8 to 3;0. We use half the data as the development set, and the rest for final evaluations.

The corresponding scene representation for each utterance must be artificially generated, since no semantic annotation of the contextual scene exists for any large corpus of CDS. First, we create an input-generation lexicon containing the “true” meaning  $t(w)$  for each word  $w$  in our corpus:  $t(w)$  is a vector over a set of semantic features, each associated with a score. An example lexical entry is given in Figure 1; the creation of this lexicon is described below.<sup>4</sup> Next, to generate the scene  $S$  for an utterance  $U$ , we probabilistically sample an observed subset of features from the full set of features in  $t(w)$  for each word  $w \in U$ . This imperfect sampling allows us to simulate the noise and uncertainty in the input, as well as the uncertainty of a child in determining the relevant meaning elements in a scene. The scene  $S$  is the union of all the features sampled for all the words in the utterance.

### The Representation of Word Meaning

We focus on the semantics of nouns, since they are central to work on the role of category knowledge in word learning. Our previous work (Nematzadeh et al., 2011) used a

<sup>4</sup>It should be emphasized that the input-generation lexicon is not used for learning by the model; it is used only to create the input.

psycholinguistically-plausible set of features for this purpose (Howell et al., 2005); however, they were only available for a limited number of nouns. Here we develop an improved semantic representation for nouns that enables a more extensive test of our clustering method and associated processing involving semantic relatedness among words.

We construct the lexical entry  $t(w)$  for each noun  $w$  drawing on WordNet<sup>5</sup> as follows. For each synset in WordNet, we select one member word to serve as the semantic feature representing that synset. The initial representation of  $t(w)$  consists of the set of such features from each ancestor (hypernym) of the word’s first sense in WordNet.<sup>6</sup> We use the same features as in previous work to initialize  $t(w)$  for other parts of speech (Nematzadeh et al., 2011; Alishahi & Fazly, 2010).

To complete the representation of  $t(w)$ , we need a score for each feature which can be used in the probabilistic generation of a scene for an utterance containing  $w$ . We assume that general features such as ENTITY, that appear with many words, are less informative than specific features such as FOOD, that appear with fewer words. Hence, we aim for a score that gives a higher value to the more specific features, so that more informative features are generated more frequently.

We formulate such a score by forming semantic groups of words, and determining for each group the *strength* and *specificity* of each feature within that group; multiplying these components gives the desired assessment of the feature’s informativeness to that group of words.<sup>7</sup>

First, we form noun groups by using the labels provided in WordNet that indicate the semantic category of the sense; e.g., the first sense of *apple* is in category *noun.food*. (For words other than nouns, we form single-member groups containing that word only.) Next, for each feature  $f$  in  $t(w)$  for a word  $w$  in group  $g$ , the score is calculated by multiplying  $\text{strength}(f, g)$  and  $\text{specificity}(f)$ :

$$\text{strength}(f, g) = \frac{\text{count}(f, g)}{\sum_{f' \in g} \text{count}(f', g)}$$

$$\text{specificity}(f) = \log \frac{|G|}{|g : f \in g|}$$

where  $|G|$  is the total number of groups, and  $|g : f \in g|$  is the number of groups that  $f$  appears in;  $\text{strength}(f, g)$  captures how important feature  $f$  is within group  $g$  (its relative frequency among features within  $g$ );  $\text{specificity}(f)$  reflects how specific a feature is to a group or small number of groups, with larger values indicating a more distinctive feature. For each word  $w$ , each feature  $f$  in  $t(w)$  is associated with the score for  $f$  and  $g$  (where  $w \in g$ ); the resulting scores are then

<sup>5</sup><http://wordnet.princeton.edu>

<sup>6</sup>A native speaker of English annotated a sample of 500 nouns with their most relevant sense in our CDS corpus, revealing that the first WordNet sense was appropriate for 80% of the nouns. One regular exception was nouns with both ‘plant’ and ‘food’ senses, such as *broccoli*, which were predominantly referring to food. For these, we always use the ‘food’ sense.

<sup>7</sup>Our score is inspired by the tf-idf score in information retrieval.

re-scaled so that the maximum score is 1, to be appropriate for the probabilistic generation of the input scenes.

## Experimental Results

In our previous work (Nematzadeh et al., 2011), we showed in computational simulations that LT learners not only learn fewer words than an ND learner, but that the LTs also have a less semantically-connected vocabulary, a result in line with the findings of Beckage et al. (2010). Here, using our extended model with its improved semantic representation, we first analyze the learned clusters of words for our two learners, to confirm that the semantic category knowledge of the LT learner is of substantially poorer quality. We also investigate the differential effects of the learned clusters for the two learners in subsequent word learning. It is known that word learning in children is boosted by their knowledge of word categories (Jones et al., 1991). Here, we interleave the two processes of semantic clustering and word learning in our model, and examine the patterns of word learning over time, for the two learners, with and without category knowledge. Our hypothesis is that the ND learner not only forms higher quality semantic clusters of words compared to the LT learner, but that its (more coherent) category knowledge contributes to improved word learning over time.

### Analysis of the Learned Clusters

We examine the quality of the semantic clusters formed by each learner (ND and LT). We train the learners on 15K utterance–scene pairs, and perform a hierarchical clustering on the resulting learned meanings of all the observed nouns. To provide a realistic upperbound as a point of comparison for the two learners, we also cluster (using the same clustering algorithm and similarity measure) the true meanings of the nouns. These “TRUE” clusters indicate how well the nouns can be categorized by the clustering method on the basis of their true (in contrast to learned) meanings. In all cases, we set the number of clusters to 20, which is the approximate number of the actual WordNet categories for nouns.

To measure the overall goodness of each of the three sets of clusters (TRUE, ND, and LT), we compare the clustering to the actual WordNet category labels for the nouns, as follows. (The WordNet category labels reflect human judgments of semantic categories, since they are provided by manual annotation.) We first label each cluster  $c$  with the most frequent category assigned by WordNet to the words in that cluster, called  $\text{label}(c)$ . We then measure  $P(\text{recision})$ ,  $R(\text{ecall})$ , and their harmonic mean,  $F(\text{-score})$ , for each cluster, and average these over all clusters in a set. Given a cluster  $c$ ,  $P$  measures the fraction of nouns in  $c$  whose WordNet category matches the cluster label;  $R$  is the fraction of all nouns whose WordNet category is  $\text{label}(c)$  that are also in  $c$ . We report the average  $P$ ,  $R$ , and  $F$  scores for the TRUE, LT, and ND clusters in Table 1.

As expected, the  $F$  score is the highest for the TRUE clusters, which result from the same clustering algorithm but applied to noise-free semantic representations. In comparison, the ND learner has somewhat lower  $F$  scores, as well as  $P$  and

TRUE			ND			LT		
$P$	$R$	$F$	$P$	$R$	$F$	$P$	$R$	$F$
.77	.71	<b>.66</b>	.79	.53	<b>.51</b>	.88	.19	<b>.24</b>

Table 1: Average  $P$ ,  $R$ , and  $F$  scores (shown in boldface), for the TRUE, LT and ND clusters after processing 15K input pairs.

$R$  scores, compared to the TRUE clusters. By contrast, the LT clusters have a very low  $F$  score. These results confirm that, in contrast to the ND learner, the LT learner is unable to use its learned knowledge of word meanings to form reasonable categories of words, confirming that nouns in the vocabulary of the LT learner have less semantic coherence than those of our ND learner. Moreover, the unusual nature of the clusters formed by the LT learner (in contrast with ND) is further confirmed by its very high  $P$  and very low  $R$  scores compared to the TRUE clusters. Detailed examination of the clusters reveals that LT has learned a large number of small clusters (leading to high precision), but also a few large semantically-incoherent clusters (leading to very low recall).

### Incorporating Categories in Word Learning

Here we investigate the role of category formation in a naturalistic word learning setting. Specifically, we interleave the two processes by allowing the model to use its semantic clusters in word learning. To simulate the simultaneous learning of categories and word meanings, the model builds clusters from its learned noun meanings after processing every 1000 input utterance–scene pairs. It then uses these clusters when processing the next 1000 pairs (at which point a new set of clusters is learned). After the first 1000 input pairs, the model calculates the alignment probabilities using both word-based and category-based knowledge, as in Eqn. (4).

For each noun in an utterance, if it has been observed prior to the last clustering point, the model uses the cluster containing the noun to calculate the category-based alignment. But a novel (previously unobserved) noun has not yet been assigned to a cluster. However, it is recognized that children can use contextual linguistic cues to infer the general semantic properties of a verbal argument (Nation et al., 2003). For example, a child/learner knowing the verb *eat* might be able to infer that the novel word *dax* in “she is eating a dax” is likely referring to some ‘edible thing’. We assume here that a learner can use the context of a novel noun to identify its general semantic category. In our model, we simulate this inference process by giving the model access to the WordNet category label of the novel word. Recall that each noun sense in WordNet is assigned a category label that provides information about its general semantics. The model can then choose a learned cluster for the novel noun by identifying the cluster whose assigned label matches the WordNet category of the noun. If more than one cluster has the same label as the category of the novel word, the cluster with the highest precision is selected. If the learner does not have a matching cluster, no category information is used for the novel word.

We process 15K input pairs overall, and look at the aver-

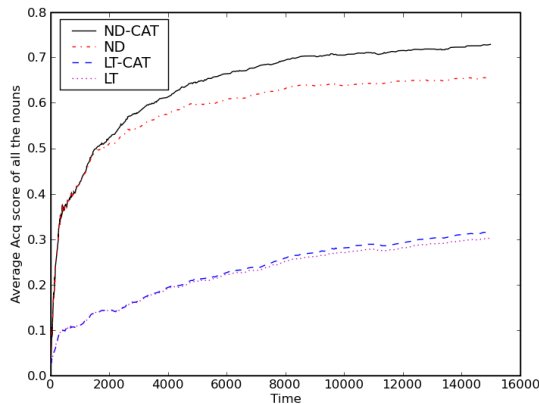


Figure 2: Change in the average Acq score of all nouns over time; ND-CAT and LT-CAT use category formulation during learning.

age acquisition score (Acq, defined below) of nouns for each learner, with and without category knowledge, as a function of time (the number of input pairs processed); see Figure 2. The Acq score for a word  $w$  shows how similar its learned meaning  $l(w)$  is to its true meaning  $t(w)$ :

$$\text{Acq}(w) = \text{sim}(l(w), t(w)) \quad (6)$$

where  $\text{sim}$  is the cosine similarity between the two vectors.

A comparison of the curves in Figure 2 reveals several interesting patterns. First, the use of category knowledge substantially improves the word learning performance of ND, whereas it has no effect at all on the (poorer) performance of the LT learner. These results further elaborate the findings of our analysis of the learned clusters: the clusters learned by the ND are a better match than those of the LT with the manually-annotated categories provided by WordNet; moreover, they are able to contribute helpful information to word learning, where the LT clusters are not.

Thus, the LT clusters are not only in principle of lesser quality, they are in practice less useful. Also, the positive effect of category knowledge for ND increases over time, suggesting that the quality of its clusters improves as the model is exposed to more input. This mutually reinforcing effect of semantic category formation with word learning underscores the importance of studying the interaction of the two.

### Category Knowledge in Novel Word Learning

Results of the previous section suggest that the ability of a learner to form reliable categories of semantically-similar words may be closely tied to its word learning performance. In particular, we expect category knowledge to increase the likelihood of associating a word with its relevant semantic features when there is ambiguity and uncertainty in the cross-situational evidence. For example, when a child hears “The wug will drink the dax” while observing an unknown animal and a bowl of liquid in the scene, the child must rely on information sources other than the cross-situational evidence to infer the possible meanings of the two novel words. (That is, the child must infer that *wug* as a drinker is more likely to

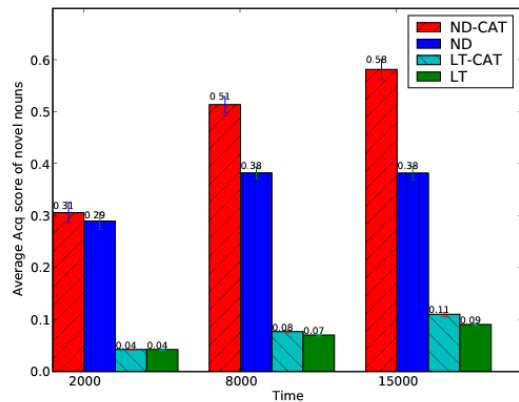


Figure 3: Changes in the novel word learning over time

be the unknown animal.) We predict a substantial benefit of category knowledge when observing a word for the first time, since this is when there’s the least cross-situational information available to a learner about the particular word and its features. Here we examine the effect of category knowledge on the learning of novel words over time, within the naturalistic setting of the utterance–scene pairs of our corpus, focusing on those inputs that include previously unseen words.

We train the model on 15K input pairs, but restrict evaluation to the learning of novel words. Specifically, we look at the difference in the Acq score of words at their first exposure, for the ND and LT learners, each with and without using category knowledge. To do this, we look at utterances containing at least two nouns, at least one of which is novel.<sup>8</sup> For each such input utterance, we record the resulting Acq score of all novel words in the utterance, and take their average. For each learner, we also examine the pattern of change in these average scores over time, as shown in Figure 3.

The results show that after 2K input utterances, there is no difference between using and not using categories for each of the learners (i.e., comparing ND-CAT and LT-CAT to ND and LT, respectively). This is because none of the learners has formed sufficiently good categories yet. After 8K utterances, ND-CAT performs much better than ND, showing the benefit of using category knowledge in learning novel words in an ambiguous setting. By contrast, for the LT learner, the Acq score of the novel nouns does not increase when using category information (LT-CAT) even with additional exposure to the input. Another interesting pattern is that for the ND learner, the average Acq score does not increase between 8K and 15K input utterances. However, when using categories (ND-CAT), this score increases over time. Although the ND model has learned additional words after 15K inputs, knowledge of more words alone does not result in improved learning of novel words. By contrast, the increasing semantic category knowledge in ND-CAT over time leads to greater improvements in learning the meaning of novel nouns.

<sup>8</sup>If the utterance only has 1 novel noun, the task is too easy because the features of nouns and other parts of speech do not overlap.



## Conclusions

One possible explanation for the language deficiencies of late-talking children is inadequacies in their attentional and categorization abilities (Jones & Smith, 2005; Colunga & Sims, 2011). In this paper, we have investigated (through computational modeling) two interrelated issues: (1) how variations in the development of attentional abilities in normally-developing (ND) and late-talking (LT) children may interact with their categorization skills, and (2) how differences in semantic category formation could affect word learning. We have extended a model of word learning that incorporates an attention mechanism (Nematzadeh et al., 2011) to incrementally cluster words, and to use these semantic clusters in subsequent word learning.

Psycholinguistic findings have noted that the vocabulary of LTs shows both a lack of appropriate generalization (Jones & Smith, 2005; Colunga & Sims, 2011), and less semantic connectivity (Beckage et al., 2010; Sheng & McGregor, 2010). We find here that the clusters formed by our LT model indeed show more inconsistency and less coherence compared to our ND learner. In addition, unlike our LT learner, our ND model can use its learned knowledge of word meanings to form semantically-coherent and informative categories, which in turn contribute to an improvement in subsequent word learning. Moreover, the LT learner has particular difficulties in learning novel words, while the ND learner gets increasingly better over time when it draws on category knowledge. The inability of an LT learner to form reasonable semantic clusters limits its ability to generalize its knowledge of learned words to new words. This could be a substantial factor in the LT's delayed vocabulary acquisition.

The model presented here treats semantic category learning and word learning as two interacting but independent processes. In particular, the mechanism for incorporating category knowledge into word learning simply adds this knowledge as another factor in guiding the formation of word-feature associations. Our ongoing work is exploring a unified mechanism in which category knowledge is integrated into the attentional mechanism of the word learning model. Such an approach will enable us to further explore how specific correlations between semantic properties and abstract categories (such as shape-solid object) emerge from the input (for LTs and NDs), and how these affect subsequent word learning.

## References

- Alishahi, A., & Fazly, A. (2010). Integrating syntactic knowledge into a model of cross-situational word learning. In *Proc. of CogSci'10*.
- Beckage, N., Smith, L. B., & Hills, T. (2010). Semantic network connectivity is related to vocabulary growth in children. In *Proc. of CogSci'10*.
- Colunga, E., & Sims, C. (2011). Early talkers and late talkers know nouns that license different word learning biases. In *Proc. of CogSci'11*.
- Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, 112(2), 347–382.
- Desmarais, C., Sylvestre, A., Meyer, F., Bairati, I., & Rouleau, N. (2008). Systematic review of the literature on characteristics of late-talking toddlers. *Int'l J. of Language and Communication Disorders*, 43(4), 361–389.
- Ellis Weismer, S., & Evans, J. L. (2002). The role of processing limitations in early identification of specific language impairment. *Topics in Language Disorders*, 22(3), 15–29.
- Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34(6), 1017–1063.
- Howell, S. R., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *J. of Memory and Language*, 53, 258–276.
- Jones, S., & Smith, L. B. (2005). Object name learning and object perception: a deficit in late talkers. *J. of Child Language*, 32, 223–240.
- Jones, S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development*, 62(3), 499–516.
- MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed., Vol. 2: The Database). Erlbaum.
- Mundy, P., Block, J., Delgado, C., Pomaes, Y., Hecke, A. V. V., & Parlade, M. V. (2007). Individual differences and the development of joint attention in infancy. *Child Development*, 78(3), 938–954.
- Nation, K., Marshall, C. M., & Altmann, G. T. (2003). Investigating individual differences in children's real-time sentence comprehension using language-mediated eye movements. *J. Experimental Child Psychology*, 86, 314–329.
- Nematzadeh, A., Fazly, A., & Stevenson, S. (2011). A computational study of late talkers in word-meaning acquisition. In *Proc. of CogSci'11*.
- Paul, R., & Elwood, T. J. (1991). Maternal linguistic input to toddlers with slow expressive language development. *J. of Speech, Lang., & Hearing Research*, 34, 982–988.
- Rescorla, L., & Merrin, L. (1998). Communicative intent in late-talking toddlers. *Applied Psycholing.*, 19, 398–414.
- Rowe, M. L. (2008). Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill. *J. of Child Language*, 35, 185–205.
- Sheng, L., & McGregor, K. K. (2010). Lexical-semantic organization in children with specific language impairment. *J. of Speech, Lang., & Hearing Research*, 53, 146–159.
- Stokes, S. F., & Klee, T. (2009). Factors that influence vocabulary development in two-year-old children. *J. of Child Psychology*, 50(4), 498–505.
- Thal, D. J., Bates, E., Goodman, J., & Jahn-Samilo, J. (1997). Continuity of language abilities: An exploratory study of late- and early-talking toddlers. *Developmental Neuropsychology*, 13(3), 239–273.
- Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *J. of Child Language*, 28, 127–152.

# Modeling dilution effects in perceptual load search tasks

Kleanthis C. Neokleous<sup>1,2</sup> ([Klneokl@Cs.Ucy.Ac.Cy](mailto:Klneokl@Cs.Ucy.Ac.Cy))

Marios N. Avraamides<sup>1</sup> ([Mariosav@Ucy.Ac.Cy](mailto:Mariosav@Ucy.Ac.Cy))

Christos N. Schizas<sup>2</sup> ([Schizas@Cs.Ucy.Ac.Cy](mailto:Schizas@Cs.Ucy.Ac.Cy))

<sup>1</sup>Department of Psychology, University of Cyprus

<sup>2</sup>Department of Computer Science, University of Cyprus  
P.O. Box 20537, 1678 Nicosia, Cyprus

## Abstract

A computational model of selective attention has been previously used to provide a concrete and comprehensive account for perceptual load findings in visual search tasks (Neokleous, Koushiou, Avraamides, & Schizas, 2009). Here, the same model was used to simulate findings from an experiment in which perceptual load effects were attributed to distractor dilution (Benoni & Tsal, 2010). By modeling at the neural level the continuous inhibitory interactions that take place among visual stimuli competing for cortical representation, the model reproduced successfully the behavioural pattern of results. The model thus offers a biologically-plausible way to reconcile findings that contradict Perceptual Load theory with those that support it.

**Keywords:** Computational Modeling, Spiking Neural Networks, Perceptual Load Theory, Dilution.

## Introduction

The *Perceptual Load* theory was proposed by Lavie and Tsal (1994; Lavie, 1995) to resolve the early vs. late debate concerning the locus of attention (e.g., Broadbent, 1958; Deutsch & Deutsch, 1963). It posits that selection of stimuli may take place early or late depending on the perceptual load of the task.

In a paradigmatic study of perceptual load Lavie and Cox (1997) had participants carry out a *high load* or a *low load* visual search task. In the high load task participants searched for two target letters (X and N) among 5 similarly-shaped letters arranged in a circular array. In the low load task, they searched for these targets among five instances of the letter O. In both conditions, a distractor letter, which participants were asked to ignore, was presented to the left or to the right of the array. Depending on condition, the distractor letter could either be congruent with the target (i.e., the same letter as the target), incongruent (i.e., the other target), or neutral (i.e., the letter “L”). Results revealed that in the low load task participants took longer to identify the target in the presence of an incongruent

distractor compared to when the distractor was congruent or neutral. In contrast, in the high load task, no difference between the three distractor conditions was found. Lavie and Cox (1997) argued that distractor interference was absent in the high load task because all attentional resources were consumed by the task leaving none to process the irrelevant distractor. In contrast, in the low load task only minimal resources were devoted to the task allowing spare resources to spill over to the processing of the distractor. Thus, the Perceptual Load theory posits that selection is early under high load conditions and late in low load conditions.

Despite its appeal, the Perceptual Load theory has been criticized on various grounds. First, a number of studies have provided findings that seem at odds with the theory (Eltiti, Wallace, & Fox, 2005; Johnson, McGrath, and McNeil, 2002; Torralbo & Beck, 2008). For example, Johnson et al. (2002) showed that cueing the target location with a 100%-predictive central cue in a low load visual search task eliminates distractor interference despite the fact that it does not alter the load of the task. Neokleous et al. (2009) reported the same result using an 80%-predictive peripheral cue. Second, Torralbo and Beck (2008) argued that the theory is unsatisfying because it does not provide a clear definition for perceptual load, and because the concept of exhaustive capacity cannot be easily reconciled with what is known about brain mechanisms.

To provide a more concrete formulation of the Perceptual Load Theory we have previously presented a biologically-plausible computational model, capable of simulating both the basic pattern of findings from Lavie and Cox (1997) and findings considered contradictory to the theory (e.g., Johnson et al., 2002). The model offered an explicit account for the possible neural mechanisms that give rise to perceptual load findings without relying on vague terms such as high and low load. The model simulated the data by modeling at the neural level the continuous inhibitory interactions that take place among visual stimuli competing for cortical representation. The strength of these inhibitory interactions is determined by the saliency of stimuli whereas top-down

signals are allowed to bias this competition by amplifying neural activity that matches the current goals.

Recently, Benoni and Tsal (2010) proposed a theoretical account for perceptual load effects that resembles the one implemented in our model. In their *Dilution* account, Benoni and Tsal (2010; see also Tsal & Benoni, 2010; Wilson, Muroi, & MacLeod, 2011) claim that the distractors in the visual search tasks employed by Lavie and Cox (1997) are processed regardless of load. However, distractor interference in the high load condition is eliminated due to diluting effects exerted by non-target letters in the search array towards the distractor.

Benoni and Tsal (2010) provided support for the Dilution account by showing that distractor interference is absent in a low load condition with high dilution (Exp.1). In this condition, participants searched for a red target in an array with three additional green letters or a green target in an array with three additional red letters, while ignoring a larger white distractor presented adjacently to the array. In the high load-high dilution condition the 4 letters of the search array, including the target, were displayed in the same color, either red or green. Finally, in a low load-low dilution condition the red or green target was presented without any accompanying letters in the search array.

Results showed that, as predicted by the Perceptual Load theory, (1) overall latencies were shorter for the low load-low dilution than the high load-high dilution, and (2) distractor interference was present in the low load-low dilution condition but not in the high load-high dilution condition. However, in contrast to the predictions of the Perceptual Load theory, no distractor interference was observed in the low load-high dilution condition despite the fact that latencies in this condition were as short as those of the low load-low dilution condition. Benoni and Tsal (2010) interpreted this finding as evidence that the perceptual load effects reported in the literature previously are caused by dilution.

Although the Dilution account aspires to offer a more concrete explanation of perceptual load effects than the Perceptual Load theory itself, it is also somewhat vague in some respects. For example, Benoni and Tsal (2010) argued that dilution requires the mere presence of non-target letters "...whose features are visually similar to those of the distractor" (p.1293). It is not very clear what constitutes a visually similar feature and how exactly a task can be categorized as high-dilution or low-dilution. Benoni and Tsal (2010) employed a low load task in which the target was presented alone in the search array thus no dilution was possible. However, Lavie and Cox (1997) have used a low load task in which the target is presented among flanking O's. Are O's expected to exert diminished dilution effects or none at all?

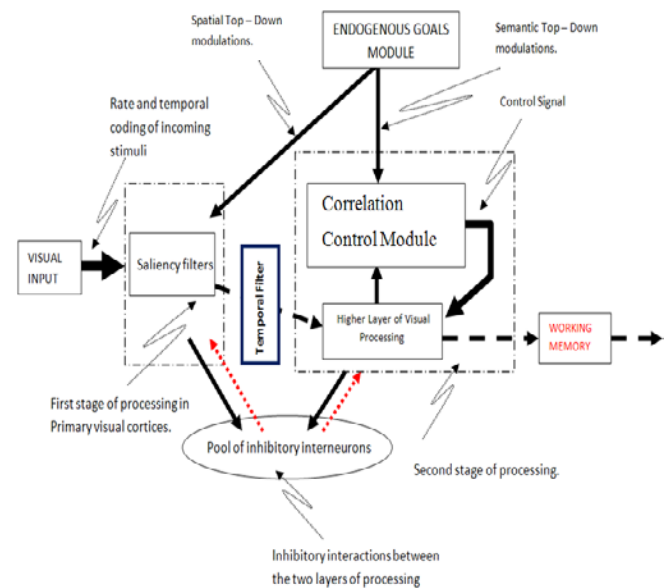
Here, we use the same computational model we described in Neokleous et al. (2009) to simulate the findings from Benoni and Tsal (2010; Experiment 1). The model requires neither a definition of load nor an explicit differentiation of tasks in terms of dilution. Also, in contrast to the Dilution

account in which inhibitory signals originate only from non-target letters in the search array and are directed only towards the distractor (i.e., the target is considered immune to inhibition), our model allows for inhibitory signals among all stimuli in the display.

## The computational model

The computational model has been previously used in similar form to simulate findings from the attentional blink phenomenon (Neokleous, Avraamides, Neokleous, & Schizas, 2009) and the relation between attention and consciousness (Neokleous, Avraamides, & Schizas, 2011). It is comprised of integrate-and-fire (I&F) neurons combined with coincidence detector (CD) neurons and simulates attention as a continuous stream of neural activity that is initially based on bottom-up information and gradually incorporates biases from top-down processes.

The model (Fig.1) involves two stages of processing implemented as spiking neural networks (SNN). The first stage involves the initial bottom-up competitive neural interactions among visual stimuli and corresponds to early visual areas in the occipital regions of the brain (e.g., V1, V2). The second stage of processing extends the neural pathway towards working memory and allows for relevant top-down information to exert an influence on neural activity.

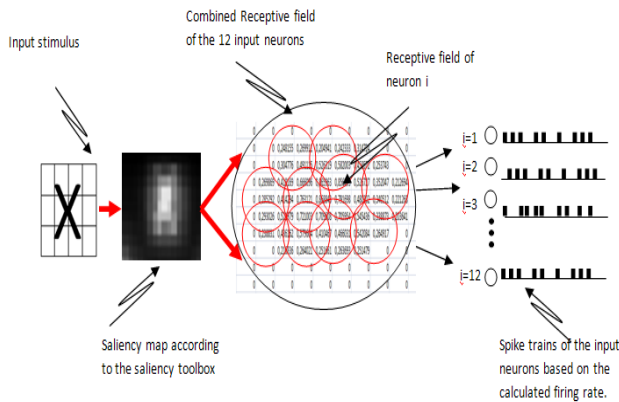


**Figure 1.** The modules of the computational model of visual selective attention.

In the first stage of processing, the initial representations of visual stimuli are created in the model on the basis of a saliency map. The modulation of visual activity by saliency in the early stages of visual processing is supported by neurophysiological findings that in area V1 of the visual cortex a neuron's response can be significantly suppressed

or enhanced by stimulation in the vicinity of its receptive field (Nothdurft, Gallant, & Van Essen, 1999; Wachtler, Sejnowski, & Albright, 2003; Shibata et al., 2008). In our model, we adopted a saliency map algorithm proposed by Koch and Ullman (1985). This algorithm was implemented by Walther and Koch (2006) into a Matlab toolbox (Saliency Toolbox - <http://www.saliencytoolbox.net>), that is used in the model to produce saliency values for spatial locations in the visual field. These values are produced based on a simple transformation algorithm that converts grayscale pixel values into frequency of spikes to establish the initial firing rates of the neurons that encode visual stimuli (Fig.2).

During the first stage of processing, neural activity can be modulated by spatial top-down factors. For example, when a cue is used to prime the location of a stimulus ahead of presentation, the neural activity corresponding to the stimulus is amplified. This implementation is based on findings from several studies showing that cues may enhance the neural activity of neurons that encode visual stimuli (e.g., Shibata et al., 2008; Silver, Ress, & Heeger, 2007).



**Figure 2.**Initial firing rate computations based on a saliency map algorithm.

The second stage of processing simulates the modulation of neural activity that represents visual stimuli by activity that maintains information about the targets as specified by the set of instructions (e.g., “Find x or y in the central array”). Support for such modulations stems from studies showing that neural activity in area V4 of the cortex influenced by top-down activity (e.g., Moran & Desimone, 1985; Reynolds & Desimone, 2003). The top-down effects in the second stage of processing are implemented in the model in a way that produces both rate amplification and synchronization of neural activity as suggested by neurophysiological evidence (e.g., Fries, Reynolds, Rorie & Desimone 2001; Gregoriou, Gotts, Zhou & Desimone 2009). That is, according to the model, attending a stimulus enhances the firing rates of neurons that correspond to that

stimulus and at the same time forces them to fire in a more synchronous rhythm. Similarly, the firing rates of neurons that correspond to unattended stimuli are suppressed.

The main components in the second stage of processing of the model have been inspired by Crick and Koch’s (1990) theoretical analysis on the role of attention and neural synchronization for the establishment of awareness. Crick and Koch (1990) based on neurophysiological findings showing that visual stimuli can elicit synchronized activity in the visual cortex, suggested that a prerequisite for the presence of neural synchronization is to have synchronous impulses in selected neuronal populations. Therefore, they proposed that visual selective attention may function in a way that it causes changes to the temporal structure of the neural spike trains that represent the information to be selected, and that this temporal structure may facilitate the transfer of the encoded information to working memory.

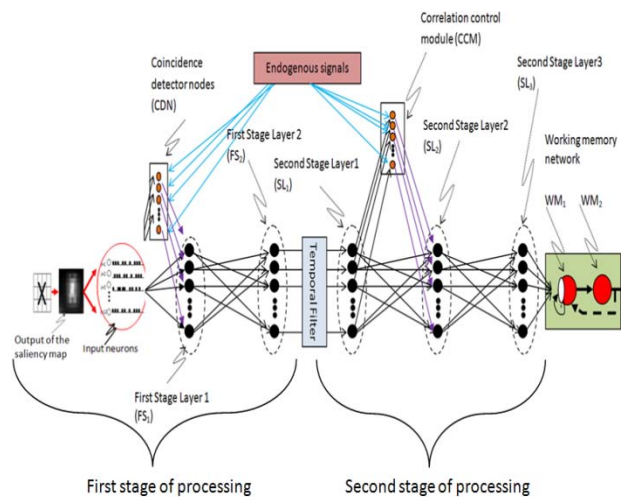
The idea presented by Crick and Koch (1990) was later supported by experimental evidence. In a comprehensive review, Womelsdorf and Fries (2007) presented evidence showing how attention selectively modulates the neurons that represent the attended stimulus feature or spatial location in a way that it synchronizes their responses. For example, Bichot, Rossi and Desimone (2005) recorded the neuronal spiking responses and LFPs in the visual area V4 of macaque monkeys and demonstrated that the allocation of attention towards a particular feature synchronizes the responses of selective sensory neurons, tuned to the attended feature. They suggested that feature salience is not only demonstrated with an increasing firing rate, but also by selectively synchronizing specific neuronal responses based on the similarity between feature preferences and the attended stimulus feature.

To incorporate these ideas in our model, templates that contain features of visual search targets are created and maintained in the endogenous goals module of the model and are used to evaluate the resemblance between any incoming visual input and a target. The evaluation of each stimulus takes place by computing the correlation between spike trains representing the stimulus and the spike trains maintaining target identity in the endogenous goals module. This is performed in the Correlation Control Module (CCM) of the model (Fig.1). However, before the neural activity of each incoming stimulus is processed in the CCM, it passes through a temporal filter that reorganizes the timing of spikes without altering the average firing rate. This mechanism is implemented in the model according to a pre-defined probability that reflects the degree of resemblance between the features of the incoming stimulus and those of a target. Thus, only the spike train patterns of a stimulus that shares features with the target will significantly change and become closer to the distinct spike train pattern of the target. The temporal filter mechanism used in the model is in line with Crick and Koch’s (1990) suggestion about the impact of selective attention on neural synchronization.

During the progression of neural activity through the two stages of processing, the encoded stimuli compete for access



to working memory (WM) through forward and lateral inhibitory interactions (from the pools of inhibitory interneurons), resulting into modulation of the strength of their neural response (Fig.3). This implementation is based on neurophysiological findings showing that competition for neural representation in visual areas V1 and V2 is initiated when two or more stimuli fall within the receptive fields of the same or nearby cells (Reynolds & Chelazzi, 2004; Reynolds & Desimone, 1999).

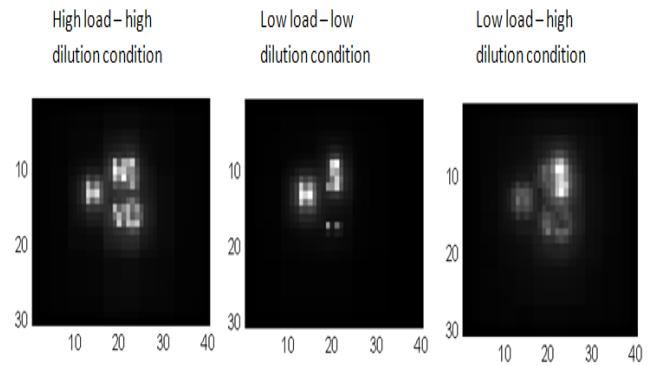


**Figure 3.**Top-down and bottom-up interactions during the progression of neural activity in the two stages of processing.

These interactions during the progression of neural activity produce enhancement and synchronization of neural activity that eventually lead to the selection of a particular stimulus for further processing.

## Computational Simulations

The computational model described in the previous section was used to simulate the pattern of findings reported by Benoni and Tsal (2010). Three aspects of the model are important for simulating the data: (1) spatial top-down signals enhance the neural activity of the neurons whose receptive fields fall within the area of the search array, (2) the saliency analysis produced different values for stimuli in each condition of the experiment (Fig.4) and led as a result to different initial firing rates, and (3) stimuli whose neural representation matches that of a target template held in the endogenous goals module are biased regardless of whether they appear, at a given trial, in the search array or as a distractor.



**Figure 4.**The three conditions and the output of the saliency map algorithm.

These aspects of the model allow the target to “win” the race to working memory but with different speed depending on the combination of load/dilution (low load-low dilution, low load-high dilution, high load-high dilution) and distractor compatibility (congruent vs. incongruent).

Fifty simulation trials were run for each of the combinations of load and compatibility. Median latencies from the model are shown in Fig.5. As seen in the figure, the model successfully produced the pattern of latencies reported by Benoni and Tsal (2010). Specifically, a compatibility effect (i.e., slower latency for incongruent vs. congruent distractors) was produced in the low load-low dilution condition only. Latencies were overall shorter in the low load-high dilution condition than in the high load-high dilution condition, but no difference between congruent and incongruent distractors was present in either condition. It should be noted that although the model successfully produced the patterns reported by Benoni and Tsal (2010), it was in all conditions slower than human participants by 150-200ms. However, it should also be pointed out that the simulations were run with exactly the same parameter settings that were previously used to simulate the findings of Lavie and Cox (1997; see Neokleous et al., 2009). That is, no effort was made to fit the behavioral data by tweaking the parameters of the model.

In the next section we discuss how exactly the model simulates the behavioral data.

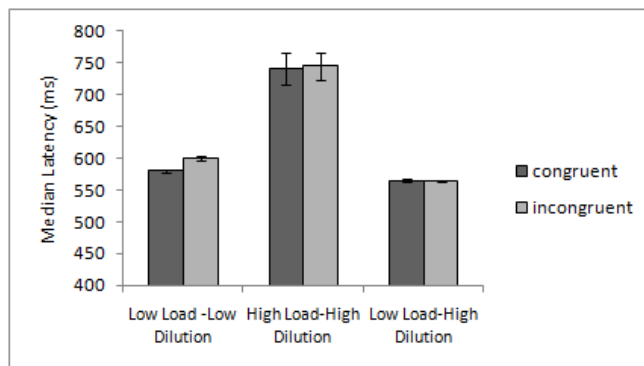
## How the model simulates the experimental data

### High Load-High Dilution Condition

In this condition the saliency analysis produced similar values for the target and the non-target letters in the search array. Saliency was somewhat higher for the distractor due to its larger size. As the task instructs participants to focus on the search array, the effects of spatial top-down signals were modeled by raising the firing rate for the four letters of the array. The higher neural activations for these letters resulted in greater inhibition from the letters of the search array (both target and non-target letters) towards the

distractor. As a result, the neural activity of the distractor was significantly reduced from the very early stages of processing, thus reaching the second stage of processing with low activation. Due to its low activation, the distractor, regardless of being congruent or incongruent with the target, could not influence much the response latency.

Besides inhibition exerted towards the distractor, the non-target letters of the search array produced strong inhibitory signals towards each other and towards the target. As a result, the neural activity of the target was reduced and the time needed to pass the set response threshold was increased. This accounts for the long response latencies observed in this condition.



**Figure 5.** Experimental data for the conditions reported by Benoni and Tsal (2010) compared to the simulation results from the computational model.

#### *Low Load -Low Dilution Condition*

Due to the absence of non-target letters in the search array of the low load-low dilution condition, the target and the distractor had about the same saliency. In fact, due to its larger size, the distractor initially had a somewhat higher saliency value than the target. The neural activity of the target was, however, amplified as it fell within the area that participants are instructed to attend. Although inhibitory interactions take place between the target and an incongruent distractor, both enter the second stage of processing with enough activation to produce a match with the goal templates held in the endogenous goals module. While a congruent distractor assists the target's processing, an incongruent distractor inhibits it. This results into (1) overall shorter latencies than in the high load condition, and (2) longer latencies for trials with incongruent than congruent distractors.

#### *Low Load-High Dilution Condition*

The low load-high dilution is the critical condition for differentiating the Dilution account from the Perceptual Load theory. The saliency analysis resulted in a higher value for the target letter than for the non-target letters of the search array because it was presented in a different color. In addition, the neural activity for all letters in the search array

was amplified to model top-down spatial effects. As a result, the target accumulated substantial activation which allowed it to exert strong inhibition towards the other elements of the display, including the distractor. As in the high load condition, the distractor reached the second stage of processing with low neural activation and was thus unable to exert strong influence on the processing of the target. The main difference between this condition and the high load condition, is that here, because of the initial amplification of the target's neural activity, the neural activations of the non-target letters in the search array were suppressed which allowed the target to be processed easier and faster.

## **Discussion**

The computational model of selective attention that was described in the present paper was previously implemented to account for the basic pattern of findings from the Perceptual Load paradigm (Lavie & Cox, 1997). The model was capable of simulating not only the basic pattern of load findings, but also findings that were considered contradictory to the theory (Johnson et al., 1992). Here, the same model, with no tuning whatsoever, was able to reproduce the pattern of findings from a study manipulating dilution (Benoni & Tsal, 2010).

Although the Dilution account of Benoni and Tsal (2010) resembles the functioning of the model we had presented earlier (Neokleous et al., 2009), its premises are not entirely in line with the way our model reproduces the behavioral data. According to the Dilution account, the representation of the distractor in high dilution conditions and in the typical high load conditions is degraded by inhibitory signals exerted from the non-target search array letters towards the distractor. Although we agree that such inhibitions take place, our model posits that in some cases, such as the in low load-high dilution of Benoni and Tsal (2010), the major source of inhibition on the distractor originates from the target. In contrast to the Dilution account, our model allows for inhibitory signals among all elements of the display. The amount of inhibition that any stimulus exerts on others depends on the strength of its neural activity, which according to the model, is based on its initial saliency and the biasing from top-down factors.

The computational model presented here is an attempt to provide a comprehensive and concrete account for Perceptual Load findings based on what is currently known about the neural mechanisms of selective attention. The way the model is implemented allows for modeling a wide range of empirical data related to perceptual load effects. Future empirical research will allow us to test predictions from the model and evaluate its validity.

## **Acknowledgments**

This research was supported by grant 0308(BE)/16 from the Cyprus Research Promotion Foundation.

## References

- Benoni, H., & Tsal, Y. (2010). Where have we gone wrong? Perceptual load does not affect selective attention. *Vision Research*, 50, 1292–1298.
- Bichot, P., Rossi, F., Desimone, R., (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science*, 308, 529-534.
- Broadbent, D. (1958). *Perception and Communication*. London: Pergamon Press.
- Crick, F., & Koch, C. (1990). Some reflections on visual awareness. *Cold Spring Harbor Symposia on Quantitative Biology*, 55, 953-962.
- Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, 70, 80-90.
- Eltiti, S., Wallace, D., & Fox, E. (2005). Selective target processing: perceptual load or distractor salience? *Perception & Psychophysics*, 67, 876-885.
- Fries, P., Reynolds, J. H., Rorie, A. E., & Desimone, R. (2001). Modulation of oscillatory neuronal synchronization by selective visual attention. *Science*, 29, 1560-1563.
- Gregoriou, G. G., Gotts, S. J., Zhou, H., & Desimone, R. (2009). High-frequency, long-range coupling between prefrontal and visual cortex during attention. *Science*, 324, 1207-1210.
- Johnson, D. N., McGrath, A., & McNeil, C. (2002). Cuing interacts with perceptual load in visual search. *Psychological Science*, 13, 284-28.
- Koch, C. & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4, 219-227.
- Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human, Perception, & Performance*, 21, 451-468.
- Lavie, N. & Cox, S. (1997). On the efficiency of attentional selection: Efficient visual search results in inefficient rejection of distraction. *Psychological Science*, 8, 395-398.
- Lavie, N., & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception & Psychophysics*, 56, 183-197.
- Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, 229, 782-784.
- Neokleous, K.C., Avraamides, M.N., Neocleous, C.K., & Schizas, C.N. (2011). Selective attention and consciousness: Investigating their relation through computational modeling. *Cognitive Computation*, 3, 321-331.
- Neokleous, K.C., Avraamides, M.N., Neocleous, C.K., & Schizas, C.N. (2009). A neural network model of the attentional blink phenomenon. *International Journal of Engineering Intelligent Systems*, 17, 115-126.
- Neokleous, K.C., Koushiou, M., Avraamides, M.N., & Schizas, C.N. (2009). A coincidence detector neural network model of selective attention. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*, Amsterdam, the Netherlands.
- Nothdurft, H. C., Gallant, J. L., & Van Essen, D. C. (1999). Response modulation by texture surround in primate area V1: correlates of "popout" under anesthesia. *Visual Neuroscience*, 16, 15-34.
- Reynolds, J. H., & Desimone, R. (2003). Interacting roles of attention and visual salience in V4. *Neuron*, 37, 853-863.
- Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque Areas V2 and V4. *Journal of Neuroscience*, 19, 1736-1753.
- Reynolds, J.H., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, 27, 611–647.
- Shibata, K., Yamagishi, N., Goda, N., Yoshioka, T., Yamashita, O., Sato, M. A., et al. (2008). The effects of feature attention on prestimulus cortical activity in the human visual system. *Cerebral Cortex*, 18, 1664-1675.
- Silver, M. A., Ress, D., & Heeger, D. J. (2007). Neural correlates of sustained spatial attention in human early visual cortex. *Journal of Neurophysiology*, 97, 229-237.
- Torralbo, A., & Beck, D. M. (2008). Perceptual-load-induced selection as a result of local competitive interactions in visual cortex. *Psychological Science*, 19, 1045-1050.
- Tsal, Y. & Benoni, H. (2010). Diluting the burden of load: Perceptual load effects are simply dilution effects. *Journal of Experimental Psychology: Human Perception and Performance*, 36, 1645–1656.
- Wachtler, T., Sejnowski, T. J., & Albright, T. D. (2003). Representation of color stimuli in awake macaque primary visual cortex. *Neuron*, 37, 681-691.
- Walther, D. & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, 19, 1395-1407.
- Wilson, D. E., Muroi, M., & MacLeod, C. M. (2011). Dilution, not load, affects distractor processing. *Journal of Experimental Psychology: Human Perception and Performance*, 37, 319-335.
- Womelsdorf, T. and Fries, P. (2007). The role of neuronal synchronization in selective attention. *Current Opinion in Neurobiology*, 17, 154-160.



# A Cultural Decision-Making Model for Negotiation based on Inverse Reinforcement Learning

Elnaz Nouri (nouri@ict.usc.edu)

Kallirroi Georgila (kgeorgila@ict.usc.edu)

David Traum (traum@ict.usc.edu)

Institute for Creative Technologies, University of Southern California  
12015 Waterfront Drive, Playa Vista, CA 90094, USA

## Abstract

We learn culture-specific weights for a multi-attribute model of decision-making in negotiation, using Inverse Reinforcement Learning (IRL). The model takes into account multiple individual and social factors for evaluating the available choices in a decision set, and attempts to account for observed behavior differences across cultures by the different weights that members of those cultures place on each factor. We apply this model to the Ultimatum Game and show that weights learned from IRL surpass both a simple baseline with random weights, and a high baseline considering only one factor of maximizing gain in own wealth in accounting for the behavior of human players from four different cultures. We also show that the weights learned with our model for one culture outperform weights learned for other cultures when playing against opponents of the first culture. We conclude that decision-making in negotiation is a complex, culture-specific process that cannot be explained just by the notion of maximizing one's own utility, but which can be learned using IRL techniques.

**Keywords:** cultural decision-making; negotiation; ultimatum game; inverse reinforcement learning.

## Introduction

Social scientists have often observed that people from different cultures behave differently in interactive situations (Camerer, 2003; Roth, Prasnikar, Okuno-Fujiwara, & Zamir, 1991). There are several different possible explanations for this, including

1. one culture is better than another at optimizing outcomes;
2. there is some kind of convention (Lewis, 1969) or equilibrium at work, such that people behave differently because the context is different, particularly their expectations about how others will behave. For example people in Japan or England drive on the left while people from America and Europe drive on the right, because that is the safest, most efficient way given how other drivers will behave, even though the goals of safety and efficiency are the same, and neither is innately better at achieving these goals;
3. the cultures have different goals, which lead to their optimizing different functions.

Most classical economic game-theory accounts of decision-making, e.g. (Neumann & Morgenstern, 1944), look at a monolithic notion of utility and maximizing

expected utility as the key to rationality. This, in effect, denies the third explanation above. For very simple games, where it is relatively easy to calculate the payoffs, the first possibility seems hard to believe, thus we are left with the hypothesis that differences in behavior are based on applying common utility principles to different problems. Others, e.g. (Gal, Pfeffer, Marzo, & Grosz, 2004), have claimed that there are many factors that contribute to the behavior of humans in social situations. This makes the third explanation plausible, if people from different cultures have different relative weights for the different factors. But this leads to a further question of how to determine those different weights. In (Nouri & Traum, 2011) we presented one such model of decision-making that culture-specific virtual agents were able to use to play the Ultimatum Game (see the following section) with each other or with people. The model used Hofstede's multi-dimensional model of culture (Hofstede, 2001) to determine the relative weights of different factors. However, in that work the weights were set manually using our intuitions about how to apply the literature, which involved a number of relatively arbitrary decisions.

In this paper we attempt to learn the weights using Inverse Reinforcement Learning (IRL) (Abbeel & Ng, 2004). To our knowledge no one has used IRL before in the Ultimatum Game or generally to learn patterns of behavior in negotiation. We also perform two experiments to try to get at the question above of what is the best explanation for the observed behavioral differences across cultures. On one account, it is the different goals that lead to different behavior. In this case we would predict that we learn different goals for different cultural patterns and that these goals would be better at generating observed behavior than other possible goals. On another account, we would expect the same set of goals to be satisfactory for any population, and differences in behavior to result from the different environments that are encountered. Our results show that the learned weights are better able to match observed distributions of culture-specific behavior than either arbitrary weights, a simple model based on economic gain, or in most cases the weights learned for other cultures. This suggests that cultures vary in goals, not just conventional circumstances but also that we can successfully use IRL techniques to learn population-specific goals for this type of game.

The structure of the paper is as follows. First we briefly present the Ultimatum Game and studies that show different behaviors for different culture groups. Then we describe our decision-making model and we present an overview of Reinforcement Learning (RL) and IRL. After that we talk about our experimental setup and present our results. Then we discuss our results and propose ideas for future work, and finally we conclude.

## Culture and the Ultimatum Game

We use the Ultimatum Game as a testbed for our model. The Ultimatum Game involves two players bargaining over a certain amount of money (in our experiments, \$100). One player, the proposer, proposes a division, and the second player, the responder, accepts or rejects it. If the responder accepts, each player earns the amount specified in the proposal, and if the responder rejects, each player earns zero. At perfect equilibrium, according to economic game theory, the proposer receives all or almost all of the money and the responder accepts all offers made to them. This classic experimental economics game has received a great deal of attention since the initial experiment by (Güth, Schmittberger, & Schwarze, 1982). Results from these studies often deviate from the predictions of game theory (Henrich, 2000; Camerer, 2003). In fact there is considerable variation of offers and rejection rates across studies (Henrich, 2000; Buchan, Croson, & Johnson, 1999), and it has been reported that people from different cultures behave differently in this game. For example (Roth et al., 1991) studied the Ultimatum Game in four countries (US, Japan, Israel, and former Yugoslavia). They found that the offers in US and Yugoslavia were higher than the offers in Japan which were higher than the offers in Israel. (Henrich, 2000) compared the behavior of 18-30 year old Machiguenga men of the Peruvian Amazon with UCLA students and found significant differences, i.e. the offers of the latter were higher than the offers of the former. (Buchan et al., 1999) studied the differences in comparable student populations in Pennsylvania and Tokyo and observed that the offers of the former were lower than the offers of the latter.

All the above studies clearly show that culture can play an important role in negotiation and in particular in the Ultimatum Game. The question however is what role: different goals, or different conventions, and whether we can learn to emulate culture-specific behavior.

## Our Decision-Making Model

Our decision-making model presented in (Nouri & Traum, 2011) considers a number of different metrics for evaluating a given situation, even for something as simple as division of money in an economic game such as the prisoner’s dilemma (Camerer, 2003) or the Ultimatum Game (Güth et al., 1982). Each of the metrics can be calculated from a basic payoff matrix. The metrics we considered for the Ultimatum Game include: *Self* (the agent’s own gain); *Other* (the gain of another); *Self/Other* (the relative gain of the negotiators); *Minimum* (lower bound of any participant - the aim of Rawls’

theory of justice (Rawls, 1971)). Each of these metrics can be given one or more valuations, choosing an optimum point and scale. The agent has a vector of weights, one per valuation, indicating the relative importance of that valuation. The total value for each choice is the sum of the product of values and weights for each valuation as shown in equation (1):

$$Value(Choice_i) = \sum_{j=1}^n (W_j * V_j(Choice_i)) \quad (1)$$

An advantage of this multi-valuation approach is that it can model an agent who cares (possibly to different extents) about different aspects of the situation, such as self-interest, collective interest, and fairness. In (Nouri & Traum, 2011) we also adapted this model to take into account Hofstede’s dimensions (Hofstede, 2001), i.e. *Individuality* (IDV), *Power Distance* (PDI), *Long Term Orientation* (LTO), *Masculinity* (MAS), *Uncertainty Avoidance* (UAI). Thus our generalized model shown in (2) breaks down the elements of the weight vector into one component per dimension, and thus an overall matrix of  $n$  valuations and  $m$  ( $=5$ ) dimensions.

$$Value(Choice_i) = \sum_{j=1}^n ((\prod_{d=IDV}^{UAI} W_{j,d}) * V_j(Choice_i)) \quad (2)$$

In this paper our focus is to learn the weights of (1) but our ultimate goal is also the learning of the weights of equation (2) that take into account Hofstede’s dimensions (see the discussion section).

## Reinforcement Learning and Inverse Reinforcement Learning

An agent’s policy is a function from contexts to (possibly probabilistic) decisions that the agent will make in those contexts. Reinforcement Learning (RL) is a machine learning technique used to learn the policy of an agent (Sutton & Barto, 1998). For an RL-based agent the objective is to maximize the reward it gets during an interaction. Because it is very difficult for the agent, at any point in the interaction, to know what will happen in the rest of the interaction, the agent must select an action based on the average reward it has previously observed after having performed that action in similar contexts. This average reward is called *expected future reward*. RL is used in the framework of Markov Decision Processes (MDPs). An MDP is defined as a tuple  $(S, A, P, R, \gamma)$  where  $S$  is the set of states (representing different contexts) which the agent may be in,  $A$  is the set of actions of the agent,  $P : S \times A \rightarrow P(S, A)$  is the set of transition probabilities between states after taking an action,  $R : S \times A \rightarrow \mathcal{R}$  is the reward function, and  $\gamma$  a discount factor weighting long-term rewards. At any given time step  $i$  the agent is in a state  $s_i \in S$ . When the agent performs an action  $\alpha_i \in A$  following a policy  $\pi : S \rightarrow A$ , it receives a reward  $r_i(s_i, \alpha_i) \in \mathcal{R}$  and transitions to state  $s_{i+1}$  according to  $P(s_{i+1}|s_i, \alpha_i) \in P$ . The quality of the policy  $\pi$  followed by the agent is measured by the expected future reward also called  $Q$ -function,  $Q^\pi : S \times A$

→  $\mathcal{R}$ . Details are given in (Sutton & Barto, 1998). There are several algorithms for estimating the  $Q$ -function and we use  $Q$ -learning (Sutton & Barto, 1998). However,  $Q$ -learning requires thousands of interactions between the agent and the environment in order to learn the optimal policy. In the case of a multi-party interaction, such as dialogue or the Ultimatum Game, the environment also needs to represent the decisions and actions of another participant. For this reason we need to build another agent, called a simulated user (SU) (Georgila, Henderson, & Lemon, 2006), that will behave as part of the environment and will interact with the policy for thousands of iterations to generate data in order to explore the search space and thus facilitate learning. Note that the SU generates a variety of actions for each state based on a probability distribution but does not learn from the interaction.

With RL, the reward function should be defined. Designing a good reward function is not trivial and not always possible. There are tasks where it is not clear what constitutes a good reward function. Inverse Reinforcement Learning (IRL) (Abbeel & Ng, 2004) aims to learn a reward function (not necessarily the true reward function) from a set of data recording interactions between the agent and the environment. This data is called *expert data*. The reward function  $R$  can be expressed as follows:

$$R_w(s, \alpha) = w^T \phi(s, \alpha) = \sum_{i=1}^k w_i \phi_i(s, \alpha) \quad (3)$$

where  $s$  is the state that the agent is in and  $\alpha$  the action that it performs in this state, and  $w^T$  is a vector of weights  $w_i$  for the feature functions  $\phi_i(s, \alpha)$ . Note that these feature functions are specified manually and the weights  $w_i$  are estimated by IRL.

In particular we use the imitation learning algorithm (Abbeel & Ng, 2004). The imitation learning algorithm is an iterative process. Initially we have a random policy  $\pi_i$  that by interacting with the SU generates data. Then this data is compared with the expert data and the weights  $w_i$  are calculated. Based on these weights a reward function is estimated and RL is performed to learn a new policy  $\pi_{i+1}$  which generates a new set of data by interacting with the SU. Then this new data is compared with the expert data and new weights are calculated, a new reward function is computed and so forth. The iteration stops when the distance between the data generated from the interaction of the latest policy with the SU and the expert data is lower than an empirically set threshold.

## Experimental Setup

We use data of the distribution of offers and acceptances or rejections for four different cultures (US, Japan, Israel, and former Yugoslavia) reported in (Roth et al., 1991). For each culture, we generate SU-proposers and SU-responders by using probability functions that match the reported data. (Roth et al., 1991) provide this data for the first and last round of the game. In our setup the game lasts 5 rounds. For the rounds in between we interpolate the first and last round values using

weights that vary depending on the round. For example, for round 4 we give a higher weight to the last round values and for round 2 a higher weight to the first round values. For each culture we generate “expert” data by having the SU-proposer interact with the SU-responder for that culture. We then apply IRL to learn weights of different motivational factors for each of these cultures and roles (proposer and responder), by iteratively playing against the appropriate SU. We then use the weights as a reward function, using RL, to learn policies for a proposer and responder for each culture. We evaluate success of the learned policies by how closely they match the expert data. We compare our learned policies with two baselines: RL models trained with either a random reward function or a reward function based on maximizing the wealth of the agent. We also compare the policies learned for a particular culture with the policies learned for the other cultures and the human expert data of the other cultures.

Our state definition includes information about the accumulated wealth gain of the agent (AccSelf), i.e. the wealth gain that the agent has gathered starting from the first round of the game, the accumulated wealth gain of the SU (AccOther), the wealth gain of the agent in the current round (Self), the wealth gain of the SU in the current round (Other), and also different representations of their relative gain (Self/Other) and the minimum gain (Min). We also take into account the round of the game. There are 11 actions that the proposer can perform (offer=0, offer=10, ..., offer=100). The initial context can be different for each round depending on the accumulated wealth of the agents, and the resulting reward is uncertain, depending on the action of the responder. For the responder, there are only two actions (accept, reject), but again there are many possible different start states to consider depending on the accumulated wealth of the agents (the reward is deterministic based on the state and action chosen).

The feature functions that we use are binary, i.e. the value of the feature function  $\phi_i(s, \alpha)$  is 1 when  $\phi_i$  is true for state  $s$  and action  $\alpha$ . So to form the feature functions  $\phi_i(s, \alpha)$  each feature is paired with all the available actions. Table 1 lists the features that we use to represent the type of context that we consider in each state. Thus for the proposer the feature function  $\text{Self} \geq 10 - \text{offer} = 10$  is 1 when the self gain of the proposer is  $\geq 10$  and the proposer has made an offer of 10, which means that this feature function is going to be 0 at the time of the offer (because at that point Self is always 0), 1 after this offer has been accepted, and 0 after this offer has been rejected. We also use additional features related to the accumulated wealth that are not depicted in Table 1 due to space constraints. In fact every possible value of AccSelf or AccOther can form a feature, e.g. AccSelf=150, AccOther=200, etc. Thus for the proposer the feature function AccSelf=150-offer=20 is going to be 1 when the accumulated wealth of the proposer is 150 and the proposer has made an offer of 20.

As we can see from the previous discussion our model is considerably different from the original SU model (human data) that just uses a probability distribution per round. First,

Table 1: Features used for IRL.

Self $\geq 0$	Other $\geq 0$	Self/Other $> 2$
Self $\geq 10$	Other $\geq 10$	Self/Other $> 1$
Self $\geq 20$	Other $\geq 20$	Self/Other $= 1$
Self $\geq 30$	Other $\geq 30$	Self/Other $< 1$
Self $\geq 40$	Other $\geq 40$	Self/Other $< 1/2$
Self $\geq 50$	Other $\geq 50$	Min(Self,Other) $= 0$
Self $\geq 60$	Other $\geq 60$	Min(Self,Other) $= 10$
Self $\geq 70$	Other $\geq 70$	Min(Self,Other) $= 20$
Self $\geq 80$	Other $\geq 80$	Min(Self,Other) $= 30$
Self $\geq 90$	Other $\geq 90$	Min(Self,Other) $= 40$
Self $= 100$	Other $= 100$	Min(Self,Other) $= 50$

our model is deterministic for each state but keeps track of additional state information, such as accumulated gains for each side. Thus we can still get a range of different offers and responses from our agents, depending on the learned policy for each state (including the accumulated gain) and the probability of those states. Second, our model for a specific culture includes a reward function, which is specific to that culture distribution. Third, the reward function could potentially be applied to other problems (see the discussion section), whereas we would have to collect human data to create a SU for a new problem.

We perform two experiments. The goal of the first experiment is to show that the reasoning behind the actions of the proposer is better modelled as a complex tradeoff of multiple goals, and cannot be explained merely by learning the behavior patterns of the partners. Thus for the proposer and the responder and the 4 cultures we learn 3 policies using RL; one based on a random reward function that assigns arbitrary weights (weak baseline), one where the reward function is based only on wealth (strong baseline), and one based on IRL. If only the data patterns mattered and not the reward function, we should see comparable performance between policies trained using the weak baseline reward functions and policies using the learned ones. Surpassing this weak baseline would be evidence that reward functions matter. The strong baseline follows classical economic game theory predictions. If everyone really does have this as a reward function and differences in behavior are due to learned differences in convention rather than goals, we should see this reward function able to match the observed behavior of different populations. On the other hand, if the IRL reward functions lead to better models than the strong baseline, that is evidence that multiple factors are taken into consideration.

The purpose of the second experiment is to show that the weights learned with IRL really are culture-dependent, i.e. that they work better for the culture that the weights were learned from than models learned for other cultures. To show that we use IRL to learn the reward function for the 4 cultures and then we use these reward functions to learn policies for each culture (for example for

the US culture we have policy-rewardUS-trainUS, policy-rewardJapan-trainUS, policy-rewardIsrael-trainUS, policy-rewardYugoslavia-trainUS). Then we test the 4 policies against SUs from the same culture that they were trained on (in this case, US). If the goals for different cultures really are different, then one would expect that policy-rewardUS-trainUS would better match the expert US data than policies learned using weights from other cultures.

To measure how closely the distributions generated with the 3 models match the human expert data we use Kullback-Leibler divergence.<sup>1</sup> The Kullback-Leibler (KL) divergence between two probability distributions  $P$  and  $Q$  is defined as follows:

$$D_{KL}(P||Q) = \sum_{i=1}^n P(i) \log_2 \frac{P(i)}{Q(i)} \quad (4)$$

where  $n$  is the number of points in the distribution that we consider. Because KL divergence is asymmetric we calculate  $D_{KL}(P||Q)$  and  $D_{KL}(Q||P)$  and then we take the average. The lower the KL divergence the closer the distributions.

## Results

In Table 2 we can see the KL divergences that we get when we compare our model and the two baselines with the human expert data for the proposer and responder policies of the 4 cultures. To avoid local optima or just being lucky with the random rewards, we ran both our model and the weak baseline (based on a random reward) multiple times and for each run we calculated the KL divergence. In Table 2 we report the median value of all computed KL divergences. In Figures 1 and 2 we can also see a graphical representation of our comparisons for the Japan proposer policy and the US responder policy. As we can see in all cases our IRL-based model outperforms both the weak and strong baselines. This verifies our hypothesis that decision-making is a complex process that cannot be attributed just to reacting to data or the sole factor of self-gain. It also shows the power of IRL for accurately modelling negotiation.

Table 2: KL divergences for IRL and the two baselines for all cultures and roles.

	Proposer			Responder		
	random	wealth	IRL	random	wealth	IRL
US	3.95	19.82	<b>2.84</b>	0.61	0.37	<b>0.10</b>
JP	4.01	4.86	<b>0.74</b>	0.64	0.25	<b>0.16</b>
IS	3.68	16.11	<b>1.29</b>	0.58	0.27	<b>0.13</b>
YU	9.28	3.49	<b>1.73</b>	0.57	0.26	<b>0.11</b>

The next question is whether our models are really capturing performance of people from the cultures that they were

<sup>1</sup>We also looked at Cartesian distance, but in all cases the best matching policy for the expert data was the same, so we report only KL-divergence, due to space restrictions.

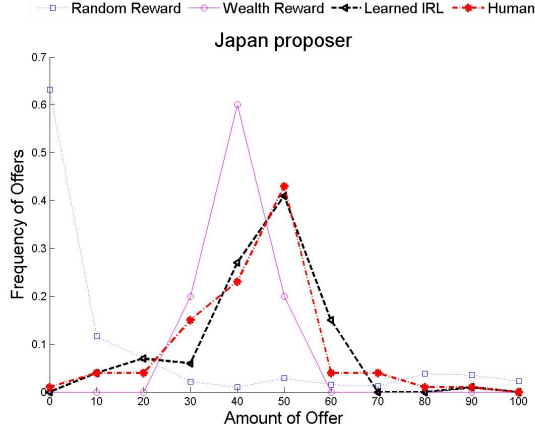


Figure 1: Comparison of random reward, wealth reward, IRL-based reward and human data for the Japan proposer policies tested with Japan SU-responders.

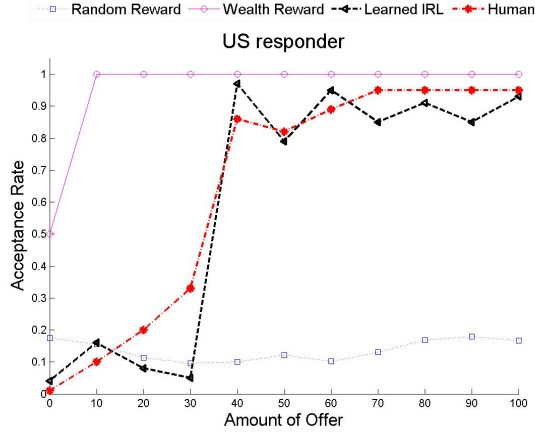


Figure 2: Comparison of random reward, wealth reward, IRL-based reward and human data for the US responder policies tested with US SU-proposers.

trained for. We examine this question in two ways. First we look at the KL divergences between all learned models and all original data sets. This is shown in Table 3. We can see that most of the time the model for each culture matches the data set from that culture better than other data sets. On the other hand there are several exceptions, for example, US proposers do better on Yugoslavia data than US data, and US responders perform well on all human data. We can also look at this table from a different perspective, as a way to compare various models (learned from data of different cultures) with the same human data. Here we can see that in most cases the data set is best modelled by the culture trained on it. However, there are a few exceptions, for example, Israel proposers are a better model of the Israel data than US and Yugoslav proposers, but a worse model of the Israel data than Japanese proposers. US proposers are not a very good model of the US data. US re-

sponders are the best model for US data, Japanese responders are a good model of the Japan data (equally good to US and Israel responders), Israel and Yugoslav responders are a good model of the Israel and Yugoslavia data respectively, but not as good as US responders. These results are encouraging and show that our models do not just beat the weaker baselines of wealth and random rewards, but also in most cases learn to model a culture better than models learned with different cultures. As we saw there are a few cases in which the results were not optimal. We believe that there could possibly be some convergence issues, even though our IRL algorithms ran for over 1000 iterations and our KL divergences are based on many runs, or perhaps, we need a larger set of features and constraints between features. Given that we take into account in our state accumulated wealth as well as rounds, our state space is fairly large. These are issues for further investigation.

Table 3: Cross-culture results, comparison with human data from different cultures (KL divergences). Best values are in bold (horizontally) and italics (vertically).

	Proposer				Responder			
	Human Data				Human Data			
	US	JP	IS	YU	US	JP	IS	YU
US	2.84	3.11	4.61	<b>2.71</b>	<i>0.10</i>	<i>0.13</i>	<i>0.08</i>	<b>0.06</b>
JP	<i>1.05</i>	<b>0.74</b>	<i>1.06</i>	1.96	0.27	<b>0.16</b>	0.18	0.24
IS	1.82	2.04	<b>1.29</b>	4.27	0.25	0.14	<b>0.13</b>	0.20
YU	2.21	2.83	5.76	<b>1.73</b>	0.15	0.27	0.20	<b>0.11</b>

In Table 4 we can see the results of experiment 2, where we use weights learned with one culture to learn policies by training on other cultures.<sup>2</sup> The results generally verify our hypothesis that the learned weights are culture-specific: with only two exceptions, the policy based on the reward function learned for that culture outperforms policies based on reward functions for all the other cultures. In the case of the US and Japanese responders, it appears that the policy trained with the Israel reward performs just as well as the policy using the learned reward function for US and Japanese responders, respectively. However the converse does not hold: the Japan and US reward functions do not work well for the Israel policies. These issues need to be investigated further.

## Discussion

Our results show clearly that there are various factors that may affect one’s decision and these factors may vary significantly depending on the culture of the decider. They also show the power of IRL for uncovering the decision-making mechanism of negotiators. (Turan, Dudik, Gordon, & Wein-gart, 2011) argue about the potential advantages of using IRL for learning the goals and motives of negotiation participants.

<sup>2</sup>We used only some of the many reward functions for each culture to learn policies for other culture data. In this table we show results for the reward functions that are closest to the median values reported in Table 2, but in some cases they are not identical.

Table 4: Cross-culture results, learning policies using rewards calculated from different cultures (KL divergences).

Policy/Role	Reward functions			
	US	JP	IS	YU
US Proposer	<b>2.84</b>	7.32	14.13	11.78
US Responder	<b>0.08</b>	6.77	<b>0.08</b>	19.10
JP Proposer	7.71	<b>1.58</b>	4.89	14.25
JP Responder	14.94	<b>0.11</b>	<b>0.11</b>	15.25
IS Proposer	3.92	5.93	<b>1.27</b>	19.52
IS Responder	7.62	6.95	<b>0.10</b>	13.08
YU Proposer	3.32	7.90	19.84	<b>1.73</b>
YU Responder	7.51	8.16	0.12	<b>0.06</b>

They use scenarios from group negotiation research and discuss how IRL could hypothetically be applied to such scenarios, but they have not actually used IRL for negotiation.

With IRL we calculated the weights for a number of features (see Table 1). These weights can be used in equation (1). However, in order to use equation (2) we need to find some kind of mapping between these weights and Hofstede’s dimensions. Our ultimate future goal is once we have Hofstede’s dimensions for a culture to be able to calculate these weights automatically. That would be very useful in cases where we do not have data to calculate the weights directly from but it is indeed a very ambitious goal. Other future work involves examining whether the reward function learned for one game or role can transfer to another. We also aim to experiment with larger numbers of RL and IRL iterations and runs, different exploration parameters, different state representations, and different features.

## Conclusion

We used IRL to learn a model for cultural decision-making in negotiation. This model takes into account multiple individual and social factors for evaluating the available choices in a decision set. Our model assigns different weights to these factors based on the modelled culture. We applied this model to the Ultimatum Game and we showed that weights learned from IRL surpass both a weak baseline with random weights, and a strong baseline that only seeks to maximize the agent’s own gain. Our model outperformed both baselines by generating behavior that was closer to the behavior of human players of the game in 4 different cultures. We also showed that the weights learned with our model for one culture outperform weights learned for other cultures when playing against opponents of the first culture.

Our results verify our hypothesis that decision-making in negotiation is a complex, culture-specific process that cannot be explained just by the notion of maximizing one’s own utility. We showed that cultures vary in goals, not just in conventional circumstances but also that we can successfully use IRL techniques to learn culture-specific goals.

## Acknowledgments

This work was funded by the NSF grant IIS-1117313 and a MURI award through ARO grant number W911NF-08-1-0301.

## References

- Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*.
- Buchan, N. R., Croson, R. T. A., & Johnson, E. J. (1999). Understanding what’s fair: Contrasting perceptions of fairness in ultimatum bargaining in Japan and the United States. In *Discussion paper, University of Wisconsin*.
- Camerer, C. F. (2003). *Behavioral game theory - Experiments in strategic interaction*. Princeton University Press.
- Gal, Y., Pfeffer, A., Marzo, F., & Grosz, B. J. (2004). Learning social preferences in games. In *Proceedings of the 19th National Conference on Artificial Intelligence* (p. 226-231).
- Georgila, K., Henderson, J., & Lemon, O. (2006). User simulation for spoken dialogue systems: Learning and evaluation. In *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH-ICSLP)*.
- Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4), 367-388.
- Henrich, J. (2000). Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *American Economic Review*, 90, 973-979.
- Hofstede, G. H. (2001). *Culture’s consequences: Comparing values, behaviors, institutions, and organizations across nations*. Thousand Oaks, CA: SAGE.
- Lewis, D. K. (1969). *Convention: A philosophical study*. Harvard University Press.
- Neumann, J. V., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.
- Nouri, E., & Traum, D. (2011). A cultural decision-making model for virtual agents playing negotiation games. In *Proceedings of the International Workshop on Culturally Motivated Virtual Characters*.
- Rawls, J. (1971). *A theory of justice*. The Belknap Press of Harvard University Press.
- Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review*, 81(5), 1068-95.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
- Turan, N., Dudik, M., Gordon, G., & Weingart, L. R. (2011). Modeling group negotiation: Three computational approaches that can inform behavioral sciences. In *E. A. Mannix, M. A. Neale, and J. R. Overbeck, eds., Negotiation and Groups (Research on Managing Groups and Teams)* (Vol. 14, p. 189-205).

# State effects of action video-game playing on visuospatial processing efficiency and attention among experienced action video-game players

Takashi Obana (a0068245@nus.edu.sg)

Maria Kozhevnikov (psymaria@nus.edu.sg)

Department of Psychology, National University of Singapore

## Abstract

Although researchers have speculated action video gaming might induce the state of “flow experience”, most previous experimental studies have focused primarily on the long-term (trait) effects of action video gaming, while overlooking possible short-term (state) effects characterizing the “flow” state. The goal of the current research was to investigate the state effects of action video games on visual-spatial processing efficiency and visual-spatial attention. We compared the baseline performance of experienced action video game players on two visual-spatial tasks and Attention Network Test with their performance on these tasks immediately after action video-gaming. The findings indicate half an hour of action video-game playing temporarily boosted participants’ performances on tasks that require visual memory, spatial transformations (mental rotation), and executive network of attention. The existence of such enhanced cognitive states implies the possibility of consciously accessing the latent resources of our brain and boosting our attentional and visual capacity upon demand.

**Keywords:** enhanced cognitive states, visual-spatial processing efficiency, attention, action video game

## Introduction

Phenomenological research suggests the existence of mental states in which overall mental functioning, as well as specific cognitive processes (e.g., attention, perception) can be enhanced for limited durations (Csikszentmihalyi, 1990; James & Marty, 1982; Maslow, 1999). Csikszentmihalyi (1997) termed such experiences “flow experiences” (p. 29), which are characterized by “complete focus” (p. 31) where “attention becomes ordered and fully invested” in the activity (p. 31). Despite the wealth of phenomenological evidence, these enhanced mental states are largely neglected in cognitive psychology research. Kozhevnikov, Louchakova, Josipovic, & Motes (2009) were the first to report experimental evidence on the existence of enhanced cognitive states as an aftereffect of focused meditation. In particular, they found that meditation that required holding the focus of attention on an internally generated image of a religious deity temporarily boosted participants’ performance in a number of visual-spatial working memory tasks. Kozhevnikov et al. (2009) suggested that a key characteristic of the induction of enhanced cognitive states, at least in the visual domain, is the intense voluntary focus of visual attention on a chosen object, which activates prefrontal-temporal and prefrontal-parietal connections in the brain, thus facilitating an enhancement of visual-spatial working memory.

A number of researchers have speculated that action video gaming might also induce the state of “flow experience”

(Chiang, Lin, Cheng, & Liu, 2011; Klasen, Weber, Kircher, Mathiak, & Mathiak, 2011; Sherry, 2004; Weber, Tamborini, Westcott-Baker, & Kantor, 2009) due to “cognitive absorption”, deep immersion, intense focus, and merging action with awareness (whereby awareness is only focused on activity) required during video-game play. Experimental studies, however, have thus far focused primarily on the trait effects that result from action video gaming (Bavelier & Green, 2003; Castel, Pratt, & Drummond, 2005; Dye, Green, & Bavelier, 2009; Green & Bavelier, 2006a, 2006b; Li, Polat, Makous, & Bavelier, 2009).

What characteristics of FPS enable these changes in cognitive performances to occur has been discussed in detail (Spence & Feng, 2010). However, despite its abundance in literature, playing FPS is seldom seen as a source of intense visual focus. While playing FPS might leave a lasting change in cognitive performances, the nature of cognitive states induced by intense visual focus remains to be unstudied. In the current study, action video game playing is seen as one way of inducing sustained intense visual focus. The aim of this study is to investigate the nature of cognitive states induced by intense visual focus by utilizing various psychological measurements.

Among the different cognitive processes affected by meditation which might be temporarily enhanced as a result of intense visual focus are different components of visual-spatial cognitive processing (Kozhevnikov, et al., 2009) as well as attentional components (Tang et al., 2007). In the present study we compared the baseline performance of experienced action video game players on two visual-spatial tasks (Mental Rotation Task and Visual Memory Task) as well as on the Attention Network Test measuring executive, orienting, and alerting components of attention with their performance on these tasks immediately after FPS video gaming.

## Method

Twenty-eight action video-game players (24 males) who have 4 to 20 years of experiences (Mean = 10), aged from 20 to 27 (Mean age = 23) were recruited for the study by advertising in National University of Singapore. One participant’s data was deleted due to procedural error. Two participants’ data were treated as outliers.

The participants of the current study played the action video-game *Unreal Tournament 2004* (referred to as FPS henceforth) by Atari. The video-game uses first-person point of view and require monitoring of the entire visual field (extent from fixation about 16° height × 29° width).



The participants were administered two computerized tasks assessing different aspects of visual processing: a visual memory task (VMT, MM Virtual Design, 2004) that assessed their ability to maintain images of complex static objects in visual working memory and a mental rotation task (MRT, Shepard & Metzler, 1971) that assessed their ability to dynamically transform and compare two spatial objects.

The VMT (MM Virtual Design, 2004) consisted of two parts. There were six test trials in the first part of the VMT. On each trial, participants were exposed to a single image (see Figure 1a) that appeared for 5 seconds. This display was replaced by an array of six images: five distractors and the previously shown image. Participants were asked to determine which image in the array was the previously shown image. There were 18 test trials in the second part of the VMT. On each trial, participants viewed an array of seven images (see Figure 1b) that appeared for 8 seconds. This array was replaced by another array of seven images: six of the previously studied images and one novel image. Participants were asked to judge which image in the second array was not present in the first.

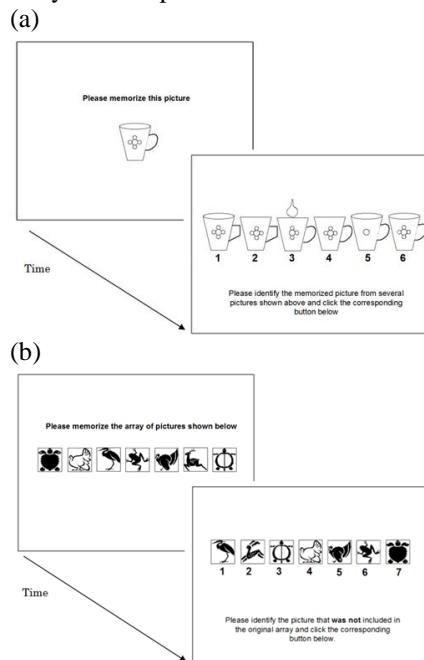


Figure 1: Examples of items from the visual memory test.

On each trial of the MRT, participants viewed a pair of two-dimensional pictures of three-dimensional forms (see Figure 2). The forms in each pair were rotated relative to each other around the x-, y-, or z-axis. Across trials, the amount of rotation ranged from 40° to 180°, in 20° increments. Participants judged whether the forms in the pair were the same or mirror-reversed. There were 36 test trials.

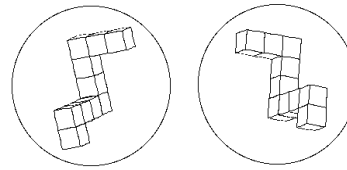


Figure 2: Example of a test trial from the Shepard & Metzler mental rotation task.

The Attention Network Test (ANT) was designed by Fan et al. (2002) to evaluate three attention networks: alerting, orienting and executive attention (see Figure 3). First, the alerting attention signifies the ability to make use of the presented cue. It represents individual's vigilance or alert state of preparedness to respond to the cue while attention is prompted to be diffused. Second, orienting attention is measured by the use of spatial cue. It signifies individual's ability to direct attention to specific location. Third, executive attention represents individual's ability to selectively attend to the significant stimuli while filtering out the distracting stimuli. It is measured by the incongruent targets.

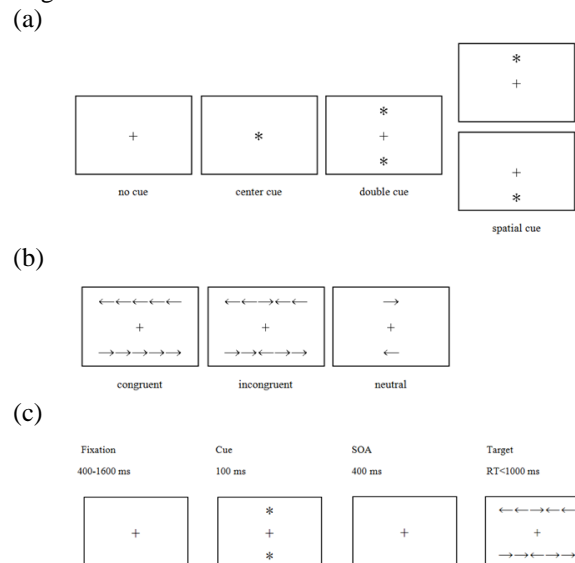


Figure 3: Attention Network Test (ANT). (a) The four cue conditions; (b) The three target conditions; and (c) an example of stimulus presentation sequence.

All the participants were tested individually, in a testing session lasting from 2.5 to 3 hours. First, as a pretest, all the participants completed the ANT, MRT, and VMT, the order of which was counterbalanced across participants.

After completing the pretest, all the participants were playing a video-game for 30 mins. Then, we randomly assigned the participants to the two following groups (see Figure 4).

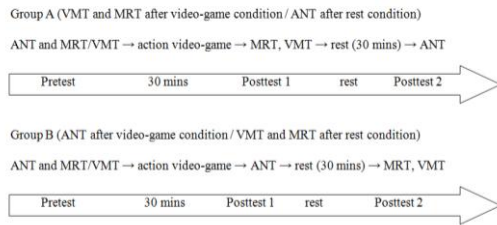


Figure 4: The time sequences of the different groups.

Above are the experimental procedures for group A (N=13) and B (N=12). In the beginning of the whole procedure, ANT, MRT, and VMT were given to all participants as pretests. The order of MRT and VMT was counterbalanced across participants. After the pretests, participants were asked to play FPS for 30 minutes. Right after that, the posttests 1 were administered. When participants are done with their posttests, 30 minutes of rest periods were given. After 30 minutes have elapsed, the whole sequence ended with posttests 2.

The informal interviews were conducted during the rest period. Participants were asked how they generally felt right after playing the FPS.

Posttest 1 was given to the participants to investigate the changes in their performance right after playing the action video-game that is to measure whether they exhibited enhanced cognitive state or to see if the action video-game has enhancing effect in short term. Posttest 2 was administered to the participants in order to see whether the state dissipates with time.

If to assume that the enhanced states do exist, they are of limited duration. The states dissipate quickly (about 20-25 min, according to the reports of meditators from Kozhevnikov et al. (2009) so it was impossible to give to the same participants all the visual and attentional tests in one session (overall time to complete all tests is more than 30 minutes) that is why some participants received only MRT and VMT (total duration; 10 to 15 minutes) after playing the videogame while others received only ANT (total duration; 25 to 30 minutes).

Second, we did not want to give to the same participants the same test three times (there is a possibility that they will remember some of the responses by the third time) in a row, due to a large practice effects. Thus for posttest 2 we used test(s) which was(were) not given in posttest 1.

## Results

First, we analyzed participants' descriptions of how they felt right after 30 mins of playing FPS by using the simplified version of phenomenological method (Giorgi, 1985). The recurring themes were "better focus", "faster reaction", "being more alert", and "heightened arousal level". Equally used description was feeling "tired" (7/28). Number of participants reported that playing FPS makes them psychologically alert but physically tired (6/28). It was reported that the state of alertness after playing FPS eventually dissipates in time. One participant who was very

observant to his inner state, reported he was able to react to ANT task much faster and more accurately right after playing FPS, but he could feel this state disappear upon finishing the first block of ANT (after 10 min). Thus, the state might be of short duration, approximately 10 min for 30 mins of FPS playing.

Second, we analyzed performance on the VMT and MRT. In order to avoid confounds arising from speed accuracy trade-offs, a measure of visuospatial processing efficiency for each imagery test was computed for mental rotation and visual memory tasks, similar to the previous literature (Kozhevnikov, Louchakova, Josipovic, & Motes, 2009).

Figure 5 presents the results for visual memory processing efficiency. A 2 (time: pretest vs. posttest)  $\times$  2 (Condition: after video-game vs. after rest) mixed-model ANOVA yielded a significant main effect of time,  $F(1, 25)=8.43$ ,  $p<.001$ , suggesting that there was a significant improvement in performance from pretest to posttest for all participants. The main effect of group was also significant,  $F(1, 25)=6.22$ ,  $p<.05$ , indicating that participants who took the VMT posttest after the videogame (VMT after video-game) performed significantly better than the ones who took posttest after the rest period (VMT after rest). The interaction between time and group was also significant,  $F(1, 25)=8.43$ ,  $p<.01$ . A follow-up ANOVA revealed a significant increase in efficiency from the pretest to the posttest for VMT after video-game,  $F(1, 13)=35.00$ ,  $p<.001$ , and a no significant increase for VMT after rest,  $F(1, 12)=2.52$ ,  $p=.14$ . These results indicate that there is an improvement in visual memory task performance, in posttest compared to pretest, only right after playing videogame. However, this improvement is no longer observable if the posttest is taken about half an hour after playing videogame.

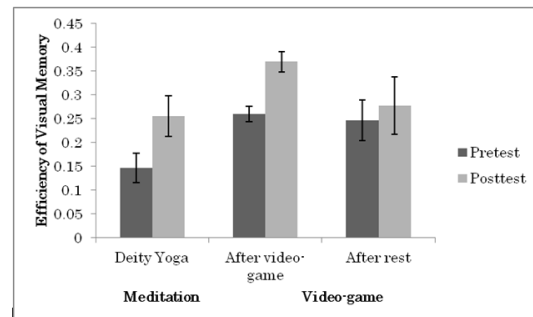


Figure 5: Processing of efficiency on the visual memory pre- and posttests as a function of group. Error bars show  $\pm 1$  SEM.

For comparison, on Figure 5 we also plotted the results for deity meditation group from Kozhevnikov (2009) study, where VMT was also given using pre-post test paradigm to different groups of meditators. Among all the groups of meditators, only *deity yoga* meditators (DY) showed the significant increase in VMT efficiency from pretest to posttest ( $F(1, 14)=26.41$ ,  $p<.001$ ). The comparison of effect size reveals participants in VMT after video-game ( $\eta^2=.73$ ) and DY ( $\eta^2=.65$ ) group are quite similar. Although the contents of focused objects are different, both playing FPS and DY involves the acts of intense visual focus.

Figure 6 presents the results for MRT efficiency. A 2 (time: pretest vs. posttest)  $\times$  2 (group: after video-game vs. after rest) mixed-model ANOVA yielded a significant main effect of time,  $F(1, 24)=16.54$ ,  $p<.01$  suggesting that there was a significant improvement in performance from pretest to posttest for all participants. However, main effect of group was not significant,  $F(1, 24)=.01$ ,  $p=.91$ . The interaction between time and group was marginally significant,  $F(1, 24)=3.23$ ,  $p=.085$ . A follow-up ANOVA revealed a significant increase in efficiency from the pretest to the posttest for MRT after video-game,  $F(1, 12)=13.87$ ,  $p<.01$ , and only marginal increase for MRT after rest,  $F(1, 12)=3.39$ ,  $p=.09$ . These results indicate that there is an improvement in mental rotation task performance, in posttest compared to pretest, only right after playing video-game. However, this improvement is no longer observable if the posttest is taken about half an hour after playing video-game.

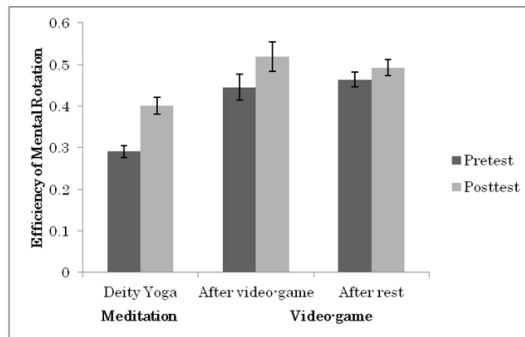


Figure 6: Processing of efficiency on the mental rotation pre- and posttests as a function of group. Error bars show  $\pm 1$  SEM.

For comparison, on Figure 6 we also plotted the results for deity meditation group from Kozhevnikov (2009) study, where MRT was also given using pre-post test paradigm to different groups of meditators. Among all the groups of meditators, only deity yoga meditators (DY) showed the significant increase in MRT efficiency from pretest to posttest ( $F(1, 14)=19.36$ ,  $p<.001$ ). The comparison of effect size again reveals participants in MRT after video-game ( $\eta^2=.54$ ) and DY group ( $\eta^2=.58$ ) are quite similar.

The result (interaction) for MRT comparing after video-game vs. after rest group was only marginally significant. This could be caused by the presence of ceiling effect. Since VGPs seem to exhibit exceptionally high spatial abilities from the beginning (Fig 6), the MRT was not able to reflect the effect of action video-game playing.

For ANT, three networks – executive, alerting and orienting – were calculated by subtracting the average RTs of specific condition from other condition (Fan et al., 2002).

ANT is composed of three test blocks, with each block taking approximately 10 minutes to complete. So that overall duration of the test is 30 min. However, as the qualitative data show, the duration of the enhanced states might be limited, and dissipate in 10 min as one of our participants reported. Since ANT took about 2 to 3 times longer compared to VMT or MRT (25 to 30 minutes vs. 10 to 15 minutes respectively), it is possible that by the time participants reached second block of the posttest, the

enhanced state started to dissipate. Thus, for each of the attentional network, we performed two analyses: 1) we compared ANT pretest with participants' performance on the first block on ANT posttest only, and 2) we compared performance on ANT pretest and ANT posttest (including all three blocks for both pretest and posttest).

Figure 7 presents the results for executive network. For better comparisons, the results for each of the three blocks of the ANT posttests are presented separately. A2 (time: pretest vs. posttest block 1)  $\times$  2 (Condition: after video-game vs. after rest) mixed-model ANOVA did not yield a significant main effect of time,  $F(1, 25)=0.48$ ,  $p=.50$  suggesting that there was no significant improvement in performance from pretest to posttest block 1 for all participants. The main effect of group was not significant either,  $F(1, 25)=1.48$ ,  $p=.24$  indicating that overall performance of participants' executive network did not differ from pretest to posttest block 1. However, the interaction between time and group was significant,  $F(1, 25)=9.84$ ,  $p<.01$ . A follow-up ANOVA revealed a significant increase in executive network efficiency from the pretest to the posttest block 1 for ANT after video-game,  $F(1, 12)=5.38$ ,  $p<.05$ , and a marginally significant decrease for ANT posttest block 1 after rest,  $F(1, 13)=4.36$ ,  $p=.06$ .

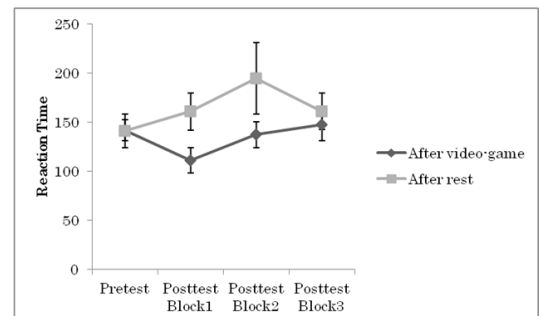


Figure 7: Executive network pre- and posttests as a function of condition. Higher reaction time denotes less efficient executive network. Error bars show  $\pm 1$  SEM.

Then, we conducted a 2 (time: pretest vs. posttest all 3 blocks)  $\times$  2 (Condition: after video-game vs. after rest) mixed-model ANOVA did not yield any significant main effects ( $p>0.2$ ). Similar to the previous case, only interaction between time and group was significant,  $F(1, 25)=7.73$ ,  $p<.05$ . A follow-up ANOVA revealed a significant decrease for ANT 3 blocks after rest,  $F(1, 13)=6.65$ ,  $p<.05$ , and no significant difference in executive network efficiency from the ANT pretest to the posttest all 3 blocks right after video-game,  $F(1, 12)=1.50$ ,  $p<.25$ .

These results indicate that there is an improvement in conflict resolution performance, in posttest compared to pretest, only about 10 minutes (during performance on block 1) right after playing video-game, and then the performance starts to degrade for both groups, as it could be seen from Figure 7.

For the orienting network and alerting network, there were no evidences that playing action video-game improved

network efficiency right after playing action video-game compared to after resting ( $p > 0.1$ ).

Taken together, the results of present study show that the enhanced cognitive states do exist. As the results of ANT (executive network) show, this state is transient and it dissipates quickly. The result shows that the residual effect of 30 min of action video-game playing can hold for at least 10 mins when the degree of visual focus intensity is strong enough.

## Conclusions

The findings of this study indicate that half an hour of FPS video game playing temporarily boosted participants' performances on tasks that require visual memory, spatial transformations (mental rotation), and executive network of attention. The effect of enhanced performance disappeared and returned to the baseline level for all the tasks after 30 minutes of rest. Furthermore, based on ANT data and participants' reports, these performance improvements might last no longer than 10 minutes after video-game playing. Thus we suggest that FPS action video game playing can give an access to relatively short temporary cognitive states characterized by drastically enhanced performance on visual-spatial and executive attention tasks.

Although some researchers (e.g., Weber, et al., 2009) speculated that the experience of flow might be related to the functioning of alerting and orienting networks during video-gaming (due to their importance in achievement/maintaining the alert attentional state as well as efficient orienting of attentional resources to spatial locations where the event takes place), our experimental results do not support this suggestion. The only attentional network which significantly improved during the enhanced state is the executive network. It is possible to infer that during enhanced states, attentional focus becomes "sharper" and of greater resolution, and thus, encoding of visual-spatial information as well as discrimination between target and flankers becomes more efficient. Meditation has a similar effect on attentional networks: Tang et al. (2007) showed an improved performance only in executive network as a result of short-term meditation.

The result of experiment 2 confirmed that intense visual focus induced by action video game playing enhances the efficiency of presented visual information processing.

Furthermore, our results show the significant increase in performance on such visual-spatial tasks as mental rotation and visual memory test during the enhanced cognitive state, similar to what has been reported by Kozhevnikov et al (2009) as a result of focused meditation. In our study, playing FPS required intense visual focus on many elements in the display such as "abrupt-onset events" (Spence & Feng, 2010, p. 93) and "significant objects" (p. 93) that must be discriminated and selected. Similar to DY meditation where the subject focuses on imagining himself/herself as a deity, focused visual attention seems to be crucial for inducing enhanced states. Indeed not all video-games seem to have a similar affect on visual attention (Bavelier & Green, 2003;

Feng, Spence, & Pratt, 2007; Green & Bavelier, 2006a, 2006b, 2007; Li, et al., 2009). For example, compared to FPS, puzzle video-games such as Tetris are shown not to have enhancing effect on spatial attention. Similarly, as it was shown by Kozhevnikov et al. (2009), not all types of meditation lead to enhanced cognitive states, but only those that require intense visual focus as DY. Other types of meditation, such as Open Presence which required "evenly distributed attention that is not directed toward any particular object experiences" (p. 646) did not produce an enhanced cognitive state. In addition focused attention, an interesting aspect of both FPS and DY is that both of them require egocentric (first-person perspective) spatial imagery. DY requires egocentric embodiment (visualizing oneself as a deity and not just focusing on an external image) and FPS "provides a natural egocentric compatibility between the visual input and the motor output" (Spence & Feng, 2010, p. 99). It is possible that a combination of focused attention and egocentric spatial imagery processes might be the key to inducing enhanced cognitive states. Future research should address this question by directly comparing the accessing of enhanced cognitive states using visual concentration on egocentric vs. allocentric visual-spatial images.

It should be noted, however, that although both DY meditation and videogaming seem to boost temporarily visual-spatial memory and attention, we should be very cautious to make any conclusions about similarity of enhanced cognitive states induced by meditation vs. video-gaming. First, it is known that meditation leads to emotional stability while video-gaming increases aggression and emotional desensitization (Anderson et al., 2010; Bushman & Gibson, 2011). Second, our participants in informal interview reported depletion of their attentional resources right after the enhanced states while monks reported "refreshed feeling" accompanying them long after meditation. Thus, while both focused meditation and action video-gaming playing might both boost visual-spatial cognitive processes, the nature and temporal dynamics of these enhanced states might be very different.

The fact that enhanced cognitive states do exist has significant practical implications for different domains of human performance. Csikszentmihalyi (1990) described the tremendous energy of enhanced cognitive states using a metaphor of atomic energy, which could be used for a variety of purposes. Knowing the means and the mechanisms behind cognitive enhancements would allow us to generalize related findings to help different cognitive problems (e.g. memory loss, attentional problems). Although it is transient, a temporary boost in visuospatial processing efficiency or attention can also greatly enhance the performance during the critical periods of our lives (e.g. training soldiers before going to battlefields). Furthermore, they can help us boost creativity (Csikszentmihalyi, 1996) or could be of particular use for learning.

## References

- Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., Saleem, M. (2010). Violent Video Game Effects on Aggression, Empathy, and Prosocial Behavior in Eastern and Western Countries: A Meta-Analytic Review. *Psychological Bulletin*, 136(2), 151-173. doi: Doi 10.1037/A0018251
- Broadbent, D. E. (1958). *Perception and communication*. New York,: Pergamon Press.
- Bartholow, B., Bushman, B., & Sestir, M. (2006). Chronic violent video game exposure and desensitization to violence: Behavioral and event-related brain potential data. *J Exp Soc Psychol*, 42, 532 - 539.
- Bushman, B. J., & Gibson, B. (2011). Violent Video Games Cause an Increase in Aggression Long After the Game Has Been Turned Off. *Social Psychological and Personality Science*, 2(1), 29-32.
- Chisholm, J. D., Hickey, C., Theeuwes, J., & Kingstone, A. (2010). Reduced attentional capture in action video game players. *Atten Percept Psychophys*, 72(3), 667-671.
- Csikszentmihalyi, M. (1990). *Flow : the psychology of optimal experience* (1st ed.). New York: Harper & Row.
- Csikszentmihalyi, M. (1997). *Finding flow : the psychology of engagement with everyday life* (1st ed.). New York: BasicBooks.
- Davidson, R. J., Kabat-Zinn, J., Schumacher, J., Rosenkranz, M., Muller, D., Santorelli, S. F., Sheridan, J. F. (2003). Alterations in brain and immune function produced by mindfulness meditation. *Psychosom Med*, 65(4), 564-570.
- Dye, M. W. G., Green, C. S., & Bavelier, D. (2009). Increasing Speed of Processing With Action Video Games. *Current Directions in Psychological Science*, 18(6), 321-326.
- Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the Efficiency and Independence of Attentional Networks. *Journal of Cognitive Neuroscience*, 14(3), 340-347.
- Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science*, 18(10), 850-855.
- Giorgi, A. (1985). *Phenomenology and psychological research*. Pittsburgh, Pa. Atlantic Highlands, N.J.: Duquesne University Press ; Distributed by Humanities Press.
- Green, C., & Bavelier, D. (2006). Enumeration versus multiple object tracking: the case of action video game players. *Cognition*, 101(1), 217-245.
- Green, C. S., & Bavelier, D. (2006). Effect of action video games on the spatial distribution of visuospatial attention. *J Exp Psychol Hum Percept Perform*, 32(6), 1465-1478.
- Green, C. S., Li, R. J., & Bavelier, D. (2010). Perceptual Learning During Action Video Game Playing. *Topics in Cognitive Science*, 2(2), 202-216.
- James, W., & Marty, M. E. (1982). *The varieties of religious experience : a study in human nature*. Harmondsworth, Middlesex, England ; New York, N.Y.: Penguin Books.
- Kozhevnikov, M., Louchakova, O., Josipovic, Z., & Motes, M. A. (2009). The Enhancement of Visuospatial Processing Efficiency Through Buddhist Deity Meditation. *Psychological Science*, 20(5), 645-653.
- Lazar, S. W., Kerr, C. E., Wasserman, R. H., Gray, J. R., Greve, D. N., Treadway, M. T., Fischl, B. (2005). Meditation experience is associated with increased cortical thickness. *Neuroreport*, 16(17), 1893-1897.
- Li, R. J., Polat, U., Makous, W., & Bavelier, D. (2009). Enhancing the contrast sensitivity function through action video game training. *Nature Neuroscience*, 12(5), 549-551.
- Lohman, D. F., & Nichols, P. D. (1990). Training spatial abilities: Effects of practice on rotation and synthesis tasks. *Learning and Individual Differences*, 2(1), 67-93.
- Lutz, A., Slagter, H., Dunne, J., & Davidson, R. (2008). Attention regulation and monitoring in meditation. *Trends in Cognitive Sciences*, 12(4), 163-169.
- Maslow, A. H. (1999). *Toward a psychology of being* (3rd ed.). cNew York: J. Wiley & Sons.
- Mevorach, C., Humphreys, G. W., & Shalev, L. (2006). Opposite biases in salience-based selection for the left and right posterior parietal cortex. [10.1038/n1709]. *Nature Neuroscience*, 9(6), 740-742.
- MM Virtual Design, L. (2004). *Imagery testing battery* [Computer software]. Newark, NJ: Author.
- Shepard, R. N., & Metzler, J. (1971). Mental Rotation of 3-Dimensional Objects. *Science*, 171(3972), 701-&.
- Siebert, A. (2000). My transforming peak experience was diagnosed as paranoid schizophrenia. *New Directions for Mental Health Services*, 2000(88), 103-111.
- Spence, I., & Feng, J. (2010). Video Games and Spatial Cognition. *Review of General Psychology*, 14(2), 92-104.
- Sutherland, A., & Crewther, D. P. (2010). Magnocellular visual evoked potential delay with high autism spectrum quotient yields a neural mechanism for altered perception. *Brain*, 133(7), 2089-2097.
- Tang, Y. Y., Ma, Y., Wang, J., Fan, Y., Feng, S., Lu, Q., . . . Posner, M. I. (2007). Short-term meditation training improves attention and self-regulation. *Proceedings of the National Academy of Sciences*, 104(43), 17152-17156.
- Wilson, C. (1972). *New pathways in psychology, Maslow & the post-Freudian revolution*. New York: Taplinger Publishing Company.

# Variation of Characteristics of Reading and Writing Difficulties in Japanese Children with Learning Disabilities

**Shino Ogawa (ogawa.shino.57n@st.kyoto-u.ac.jp)**

Primate Research Institute, Kyoto University  
Kanrin, Inuyama-city, Aichi, Japan

**Miwa Fukushima-Murata (miwa.fukushima.murata@gmail.com)**

Research Center for Advanced Science and Technology, The University of Tokyo  
4-6-1, Komaba, Meguro-ku, Tokyo, Japan

**Namiko Kubo-Kawai (namikokk@asu.aasa.ac.jp)**

Faculty of Psychology, Aichi Shukutoku University  
9, Katahira, Nagakute-cho, Aichi, Japan

**Tomoko Asai (t.asai.at@city.nagoya.lg.jp)**

Nagoya City Child Welfare Center  
4-16 Orido-cho, Showa-ku, Nagoya-city, Aichi, Japan

**Hiroko Taniai (h.taniai.67@city.nagoya.lg.jp)**

Department of Pediatrics, Nagoya Central Care Center for Disabled Children  
4-16 Orido-cho, Showa-ku, Nagoya-city, Aichi, Japan

**Nobuo Masataka (masataka.nobuo.7r@kyoto-u.ac.jp)**

Primate Research Institute, Kyoto University  
Kanrin, Inuyama-city, Aichi, Japan

## Abstract

There are conflicting hypotheses for the causes of Dyslexia in reading and writing difficulties, such as the phonological deficit hypothesis, double deficit hypothesis, magnocellular deficits hypothesis etc. The cause of the difficulties may vary between individuals. Moreover, most of these hypotheses consider only a single disability, despite the fact that factors related to reading and writing may affect the difficulty in various ways. We conducted this study to identify individual differences in the effect of Dyslexia. The participants were 12 Japanese children who were diagnosed with learning disabilities or suspected to be learning disabled. In this study, we considered how phonological awareness, visual perception, and phonological processing are related to reading and writing abilities in the Japanese language. In addition, we checked "handwriting ability." This study shows that reading and writing difficulties are caused by a variety of factors and that there are individual differences in the difficulties.

**Keywords:** Reading and writing Difficulties, Dyslexia, Individual differences, Japanese education

## Introduction

In Japan, official reports claim that 6.3% of elementary and middle school students enrolled in normal classes experience learning difficulties (MEXT Japan, 2002). This means that each class has two or more students with actual or potential learning problems, making learning disabilities an issue that should be urgently addressed to provide these students with special learning assistance. Students with

learning difficulties have more than one problem in reading, writing, listening, communicating, calculating, planning, and memorizing. In particular, support for reading and writing are very important. Difficulty with reading negatively affects all learning domains, thereby hindering academic performance in all subjects. A person's inability to read well can also generate an inferiority complex that results in the loss of his or her motivation to learn, which, in turn, may be linked to symptoms leading to juvenile delinquency (Kimberly & Richard, 2006; Siponmaa, Kristiansson, Jonson, Nyden et al., 2001). The inability to read also influences friendships outside of the classroom (Stanovich, 1986) and children's ability to process feelings of anger (Kazdin, Rodgers, Colbus, & Siegel, 1987; Moffitt & Henry, 1989). All of these factors suggest that addressing reading difficulties should be a priority for helping children with learning difficulties.

To support them, it is important to know the causes of the difficulty. Moreover, there are different manifestations of developmental dyslexia in different languages (Miles, 2000). Researchers (e.g., Landerl, Wimmer & Frith, 1997; Paulesu, McCrory, Fazio, Menoncello et al., 2000; Paulesu, Demonet, Fazio, McCrory, 2001; Wydell & Butterworth, 1999) argue that the discrepancy in the prevalence of reading impairments in different languages might be primarily due to inherent differences in the structure/characteristics of each orthography, specifically the way in which phonology is computed from it. In the alphabetic languages in which a



finer “grain” processing of orthography-to-phonology mapping is required, such as English or Danish, developmental dyslexia forms a large minority group. For these facts, to support Japanese children, it is necessary to know the characteristics of Japanese children’s reading and writing difficulties.

There are conflicting hypotheses for developmental dyslexia, reading and writing difficulties, such as the phonological deficit hypothesis (Shaywitz, 2003; Shaywitz & Shaywitz, 2005), double deficit hypothesis (Wolf & Bowers, 1999; Wolf & Bowers, 2000; Faust & Sharfstein-Friedman, 2003), magnocellular deficits hypothesis (Livingstone, Rosen, Drislane, & Galaburda, 1991), and so on. However, most of these hypotheses only consider a single disability. Other studies that discuss the issue with many factors don’t consider individual differences (e.g., Uno, Wydell, Haruhara, Kaneko et al., 2009). However, all factors related to reading and writing may affect the difficulty in various ways. This paper hypothesizes the influence of individual differences is suspected to add to the difficulties.

The core ability of reading and writing skills is phonological processing. Phonological processing is the ability to see or hear a word, break it down into discrete sounds, and then associate each sound with letter/s that make up the word. The prerequisite skills for phonological processing are the ability to analyze the phonological structure of sound and the ability to recognize its characters. According to the phonological model, the difficulty results from an impaired ability to segment spoken words into phonologic parts and link each letter to its corresponding sound (Shaywitz, 2003; Shaywitz & Shaywitz, 2005). Phonemes are small units of sound that can be conceptualized as the building blocks of words (for example, the word cat is comprised of three phonemes: k, aaaa, and t). That is the ability to analyze phonological structure. The Japanese language is based on a subsyllabic unit, the mora (Otake, Hatano, Cutler, & Mehler, 1993). According to magnocellular theory, the difficulty results from abnormalities of the magnocellular component of the visual system, which is specialized to quickly process temporal information (Stein & Walsh, 1997). That is the ability to recognize character. Furthermore, not only cognitive ability, but also the ability to correctly produce sound is necessary for reading. And handwriting abilities are necessary for writing letters correctly. Therefore, it is also important to possess these abilities. Moreover, in reading, there are two strategies, one is a lexical strategy based on whole word recognition and another a sub-lexical processing strategy based on a grapheme-to-phoneme conversion (Wydell & Butterworth, 1999). This means that if we want to know the ability of phonological processing in reading, we have to check not only word tests, but also non-word tests.

Additionally, Wydell & Butterworth (1999) established “the Hypothesis of Granularity and Transparency.” Through this hypothesis, they maintain that orthographies can be described by two dimensions: “transparency” and

“granularity” and argue that: (1) any orthography where the print-to-sound translation is one-to-one or transparent will not produce a high incidence of phonological dyslexia, regardless of the level of translation, i.e., phoneme, syllable, character, etc. This is the “transparency” dimension, and (2) even when this relationship is opaque and not one-to-one, any orthography whose smallest orthographic unit representing sound is coarse, i.e., a whole character or whole word, will not produce a high incidence of phonological dyslexia. This is the “granularity” dimension. Any orthography used in any language can be placed in the transparency-granularity orthogonal dimension described by this hypothesis. This is illustrated in Figure 1. The hypothesis argues that any orthography that falls into the shaded area in Figure 1 should not produce a high incidence of phonological dyslexia. Given the characteristics of Japanese orthography, both Japanese Kana and Kanji can be placed in the shaded area. For example, in Japanese Kana, the granularity of the smallest orthographic unit representing phonology is finer than the whole word, but coarser than the grapheme and its orthography-to-phonology translation relationship is at the level of syllables and one-to-one. For Kanji, on the other hand, the unit of granularity is much coarser, i.e., a character or a whole word and the relationship between orthography and phonology is very opaque, hence Kanji can be placed in the shaded area.

Granular Size

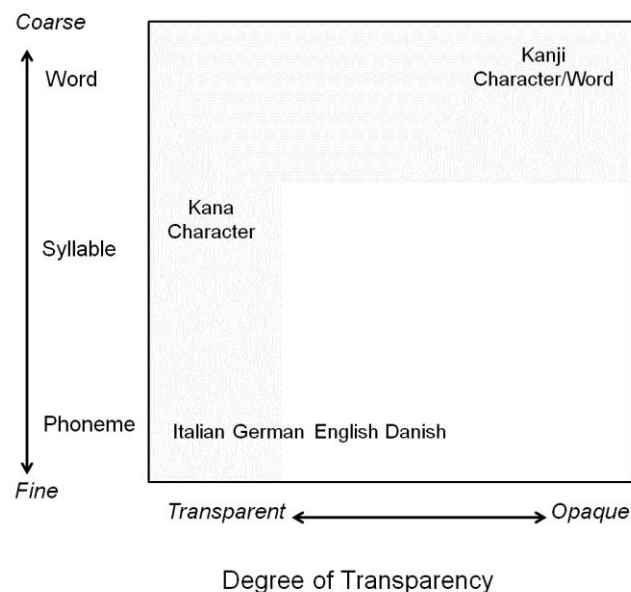


Figure 1: Hypothesis of granularity and transparency and orthography-to-phonology correspondence made by Wydell & Butterworth (1999).

In this study, we targeted the difficulties in Kana (Hiragana and Katakana), the most basic character in Japanese. We considered phonological awareness, visual



perception, and phonological processing which relate to reading and writing abilities in Japanese. In addition, we checked the ability of “expression of handwriting.” We conducted this study to bring out the effect of individual differences on the disability.

## Method

### Participants

The participants were 12 children (eight boys and four girls) who were recruited by the doctor of the Nagoya City Child Welfare Center and Nagoya Central Care Center for Disabled Children. The doctor believed that all had difficulty in reading and writing which harbored the possibility that they had developmental dyslexia. Table 1 presents participant profiles.

Table 1: Participant profiles.

Child	Grade	Gender	Dual diagnosis	
A	6	M	PDD	
B	5	M		AD/HD
C	4	M		
D	4	M	PDD	
E	4	M		AD/HD
F	3	M	PDD	
G	3	F	PDD	
H	3	F	PDD	
I	3	F	PDD	AD/HD
J	3	M		
K	3	M	PDD	AD/HD
L	3	F		

*Note.* M = Male, F = Female, PDD = Pervasive Developmental Disorder, ADHD = Attention Deficit/Hyperactivity Disorder.

In December 2010, seven children were enrolled in third grade, three in fourth grade, one in fifth grade, and one in sixth grade. All of them were enrolled in regular elementary school classes. Every child was either diagnosed as having learning disabilities or suspected of being learning disabled by the doctor. Five children exhibited symptoms that coexisted with pervasive developmental disorder, two with attention deficit/hyperactivity disorder, and two with both pervasive developmental disorder and attention deficit/hyperactivity disorder. The Wechsler Intelligence Scale for Children-Third Edition (Wisc-III) Full Scale IQ scores ranged from 79 to 112, with a mean IQ score of 97.8. Verbal IQ scores ranged from 72 to 120, with a mean IQ

score of 100.7. Performance IQ scores ranged from 80 to 120, with a mean IQ score of 95.1. No children stuttered and all could correctly produce sounds.

### Measures

**Phonological awareness task:** Each participant performed two phonological awareness tests: isolation of unvoiced sounds test and segmentation of choked sound test. Isolation of unvoiced sounds requires recognizing the individual unvoiced sounds in words, for example, “Tell me the second sound of the word you hear.” The experimental stimuli of isolation of unvoiced sounds test were 10 words of five characters like “ka-ta-tsu-mu-ri,” in Japanese “かたつむり.” The participants were asked to identify the second sound in three words, the third sound in four words, and the fourth sound in three words. The achievement scale was above 8/10. Segmentation of choked sound requires recognizing the number of sounds in words, for example, “Tell me how many sounds in the word you hear.” The experimental stimuli of segmentation of choked sound test were 10 words of six choked sound words and four unvoiced sounds words. These words had three to six characters. We analyzed only the choked sound words. The achievement scale was above 4/6. We performed these tests by checking references from the test performed by Hara (2001).

**Visual perception task:** Each participant performed three subtests of the Japanese version of the Developmental Test of Visual Perception (DTVP; Frostig, 1977): figure ground, position in space and spatial relations. Perceptual age was determined by the test. Scaled Scores (SS) of each test was determined using the DTVP. If the perceptual age was younger than the calendar age, SS was equal to or less than eight. Individual tests are not sufficiently different to measure separate abilities (Olson, 1968), and all of the tests are thought to be related to visual perception. In this study, participant passed the task if his or her score showed more than eight in every three scores of subtests. Otherwise, the participant failed the task.

**Phonological processing task:** Each participant performed six reading tests to examine their phonological processing ability. Four tests were subtests of the Screening Test of Reading and Writing for Japanese Primary School Children (STRAW; Uno, Haruhara, Kaneko, & Wydell, 2006): Hiragana character reading test, Katakana character reading test, Hiragana word reading test, and Katakana word reading test. These tests consist of 20 known words. The other two tests were the Hiragana non-word reading test and Katakana non-word reading test. These tests consist of 10 non-words. The achievement scale of each test was more than or equal to 90%. If a child’s score of the word reading test was above the achievement scale, but the character reading test or non-word reading test was below the achievement scale for at least one condition of Hiragana or Katakana, he or she can use the lexical strategy based on whole word recognition but

cannot use the sub-lexical processing strategy which is based on a grapheme-to-phoneme conversion. In this case, he or she was assumed to fail the phonological processing task. If a child's score of all the tests was above the achievement scale, he or she was assumed to have passed the phonological processing task.

**Handwriting ability task:** Each participant performed one subtest of the Japanese version of the Developmental Test of Visual Perception (Frostig, 1977): eye-motor coordination. Perceptual age was determined by the test. SS of the test was determined using the DTVP. If the perceptual age was younger than the calendar age, SS showed equal or less than eight. Participants passed the task if his or her score showed more than eight of the subtests.

## Procedure

Participants were tested individually in a quiet room at the Nagoya City Child Welfare Center. Each participant was seen more than five times over one week, each time for approximately 40 minutes. To exclude the factor of PDD or ADHD, we organized the physical environment (e.g., Treatment and Education of Autistic and related Communication handicapped Children). In each test errors were recorded. Also, each child's responses were videotaped for later reviewing. Children were told that these were not academic achievement tests, and that only the investigators would see their results. The data was collected from May 2010 to December 2010.

## Results

### Phonological awareness task

Every child passed the isolation of unvoiced sounds test. In the segmentation of choked sound test, only one child, F, failed. That means that only one participant had difficulty in phonological awareness.

### Visual perception task

Only two children, B and L, passed the task and the other 10 failed. This means that 10 children had problems in the recognition of characters in some way.

### Phonological processing task

Four children, A, B, H, and I passed all tests and therefore passed the phonological processing task.

Eight children failed the phonological processing task. Four children, C, D, E, and J failed only the Katakana non-word reading test, while L failed both the Katakana non-word reading test and the Katakana character reading test. They appeared to read the words of Katakana, but seemed to have a weak ability for phoneme-to-grapheme conversion in Katakana. Two children, G and K, failed both the Hiragana and Katakana non-word reading tests. They appeared to read the words of Kana (Hiragana and Katakana), but seemed to have a weak ability for phoneme-to-grapheme conversion in

Kana (Hiragana and Katakana). One child, F, passed only the Hiragana character reading test. She appeared to have read the words of Hiragana, but seemed to have a weak ability for phoneme-to-grapheme conversion in Hiragana. In addition, she couldn't complete all of the Katakana tests.

## Handwriting ability task

Four children, C, G, K and L, passed the task and eight failed. These eight children's writing movements may be related to their writing difficulties.

## Discussion

Table 2 presents the results of each test for each child and the type of characteristics in reading and writing difficulties. One child, F, is type 1 and failed all the tests. Three children, D, E, and J, are type 2 and passed only the phonological awareness test. Three children, C, G, and K, are type 3 and failed in visual perception and phonological processing. Three children, A, H, and I, are type 4 and failed in visual perception and handwriting ability. One child, B, is type 5 and failed only in handwriting ability. One child, L, is type 6 and failed only in phonological processing.

Table 2: Type of characteristics in reading and writing difficulties.

Type	PA	VP	PP	HA	Child	Grade
<b>1</b>	F	F	F	F	F	3
<b>2</b>	P	F	F	F	D	4
<b>2</b>	P	F	F	F	E	4
<b>2</b>	P	F	F	F	J	3
<b>3</b>	P	F	F	P	C	4
<b>3</b>	P	F	F	P	G	3
<b>3</b>	P	F	F	P	K	3
<b>4</b>	P	F	P	F	A	6
<b>4</b>	P	F	P	F	H	3
<b>4</b>	P	F	P	F	I	3
<b>5</b>	P	P	P	F	B	5
<b>6</b>	P	P	F	P	L	3

Note: PA = Phonological awareness, VP = Visual perception, PP = Phonological processing, HA = Handwriting ability, F = Failed, P = Passed

By examining individual levels for the four elements, as shown in Table 2, it becomes apparent that reading and writing difficulties are not caused by a single disability, but

rather by a combination of factors. Furthermore, the combination of individual elements is different. This means that students with learning disabilities need separate support even they have the same symptoms or reading and writing difficulties.

In the four elements, a higher percentage of children were considered to have problems with visual perception. Type 4 children passed the phonological processing task even though they failed the visual perception task. This result may be caused by Japanese education methods. The simultaneous oral spelling method is a good way for dyslexic children to acquire reading and writing skills (Thomson, 1996). Typical Japanese education methods, however, utilize simultaneous oral spelling techniques. Japanese has multiple characters that are similar to each other. Also, there are many strokes in Japanese Kanji. Therefore, we need to study the influence of visual perception on difficulties in reading and writing Japanese in the future.

The groundbreaking discovery of our study was that there were many children who have poor handwriting abilities. Stroke order is believed to very important in Japanese education.

However, if there is difficulty in handwriting, it may be hard for these children to write in handwriting stroke order. When considering the difficulty of writing, handwriting ability wasn't considered. However, from the viewpoint of quality of life, it is necessary to know a child's handwriting ability in an assessment. If a child has poor handwriting ability, he or she should be supported and taught that the stroke order is not necessarily important. In Japan, it is an accepted practice to learn characters from a set of reading and writing lessons. However, this method is not good for children who have poor writing abilities, in particular type 5 children, like child B, who have difficulty only with handwriting. These children need support in the form of separate reading and writing practice.

In this study, only child F failed the phonological awareness task. This supports the granularity dimension of "the Hypothesis of Granularity and Transparency." However, some children failed the phonological processing task because they failed the Katakana test. In particular, child L, a type 6, passed other tasks like the phonological awareness task, visual perception task and handwriting ability task. Why is there difficulty only in the Katakana phonological processing? Kana is a one-to-one from a character standpoint, but not a one-to-one transparent from a sound standpoint. This is illustrated in Figure 2.

Considering this, granularity and transparency and orthography-to-phonology correspondence of KANA will appear as presented in Figure 3.

For type 6 children, like child L, the Japanese syllabary table may be a good education support tool. The Japanese syllabary table may be utilized as a type of location map of phonemes (Seki et al., 2004). If children have already learned Hiragana, using the Katakana syllabary table may help them learn Katakana characters.

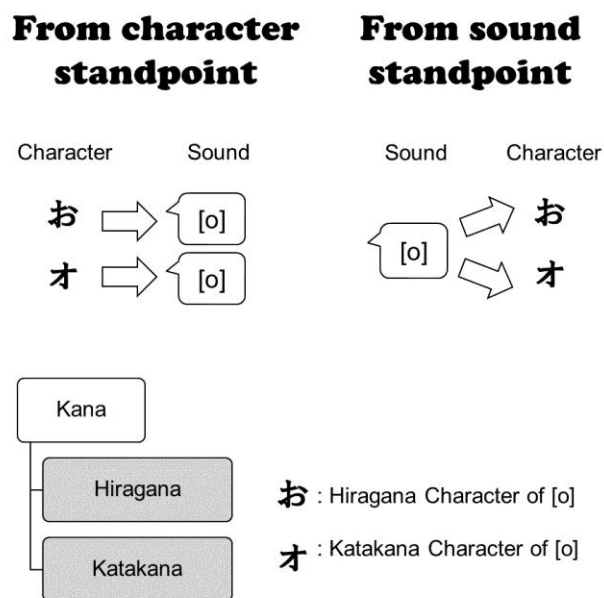


Figure 2: Transparency dimension of Kana from character and sound standpoints.

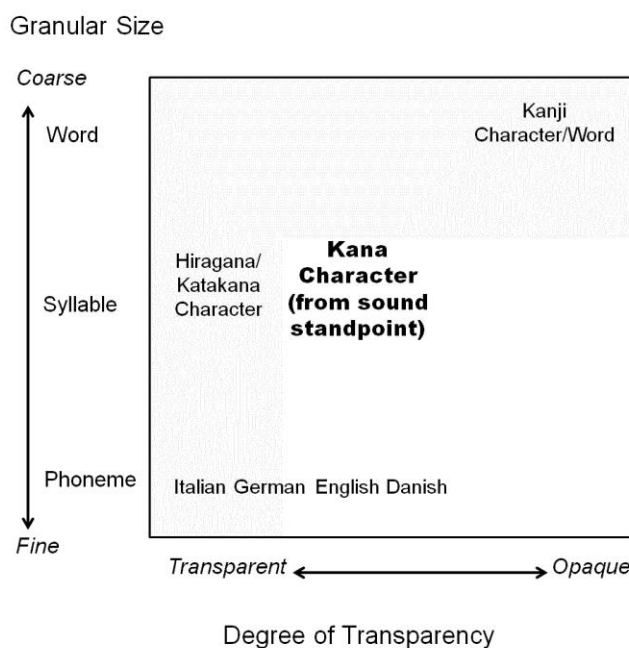


Figure 3: Granularity and transparency and orthography-to-phonology correspondence of KANA

In this study, we demonstrated that reading and writing difficulties in the Japanese language attributed to learning disabilities are caused by variety of factors and that there are individual differences in the difficulties. We also demonstrated that an assessment of handwriting ability is necessary to identify the proper types of support for

problems. Furthermore, we demonstrated that Kana is not one-to-one transparent from a sound standpoint. This research will have a large impact on education methods and techniques in Japan.

### Acknowledgments

This work was supported by JST RISTEX Implementation-Support Program, Grant-in-Aid for challenging Exploratory Research 20653076, consigned research fund from nagoya-city, and Grant-in-Aid for JSPS Fellows 235191. We thank Ms. Maki Ogawa for her enormous help with the experiments. We also express our gratitude to the participating children and their parents.

### References

- Kazdin, A. E., Rodgers, A., Colbus, D., & Siegel, T. (1987). Children's hostility inventory: measurement of aggression and hostility in psychiatric inpatient children. *Journal of Clinical Child Psychology*, 16, 320-328.
- Kimberly, A. M., & Richard, J. M. (2006). Disability and Juvenile delinquency: issues and trends. *Disability & Society*, 21, 613-627.
- Faust, M. & Sharfstein-Friedman, S. (2003). Naming difficulties in adolescents with dyslexia: Application of the tip-of-the-tongue paradigm. *Brain and cognition*, 53(2), 211-217
- Frostig, M. (1977). Developmental test of visual perception (K. Iibachi, Y. Suzuki, & S. Mogi, Trans.). *Palo Alto: Consulting Psychological Press*. (Original work published 1963).
- Hara, K. (2001). The development of phonological awareness in Japanese Children. *Japanese Journal of Communication Disorders*, 18(1), 10-18.
- Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German-English comparison. *Cognition*, 63, 315-334.
- Livingstone, M.S., Rosen, G.D., Drislane, F.W., & Galaburda, A.M. (1991) Physiological and anatomical evidence for a magnocellular defect in developmental dyslexia. *Proceedings of the National Academy of Sciences of the United States of America*, 88, 7943-7947.
- MEXT, Japan (2002). Nationwide Survey About Children Who Are Needed For Special Support Education Program In Enrolled In Ordinal School Education (in Japanese).
- Miles, E. (2000). Dyslexia may show a different face in different languages. *Dyslexia*, 6, 193-201.
- Moffitt, T. E., & Henry, B. (1989). Neuropsychological assessment of executive function in self-reported delinquents. *Development and Psychopathology*, 1, 105-118.
- Olson, V. (1968) Factor Analytic Studies of the Frostig Developmental Test of Visual Perception. *The Journal of Special Education*, 2, 429-433.
- Otake, T., Hatano, G., Cutler, A. & Mehler, J (1993) Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, 32(2), 258-278.
- Paulesu, E., Demonet, J. F., Fazio, F., McCrory, E., Chanoine, V., Brunswick, N., Cappa, S. F., Cossu, G., Habib, M., Frith C. D., & Frith, U. (2001). Dyslexia-cultural diversity and biological unity. *Science*, 291, 2165-2167.
- Paulesu, E., McCrory, E., Fazio, F., Menoncello, L., Brunswick, N., Cappa, S., Cotelli, M., Cossu, G., Corte, F., Lorusso, M., Pesenti, S., Gallagher, A., Perani, D., Price, C., Frith, C., & Frith, U. (2000). A cultural effect on brain function. *Nature Neuroscience*, 3, 91-96.
- Seki, A., Okada, T., Koeda, T. & Sadato, N. (2004) Phonemic manipulation in Japanese: an fMRI study, *Cognitive Brain Research*, 20, 261-272.
- Shaywitz, S. E. (2003). Overcoming Dyslexia. *Random House Inc.*, NY.
- Shaywitz, S. E. & Shaywitz, B. A. (2005). Dyslexia. *Biological Psychiatry*, 57(11), 1301-1309.
- Siponmaa, L., Kristiansson, M., Jonson, C., Nyden, A., & Gillberg, C. (2001). Juvenile and young adult mentally disordered offenders: The role of child neuropsychiatric disorders. *Journal of the American Academy of Psychiatry and the Law*, 29, 420-426.
- Stanovich, K. E. (1986). Matthew effects in reading: some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21, 360-407
- Stein, J. & Walsh, V. (1997) To see but not to read; the magnocellular theory of dyslexia. *Trends in neurosciences (Regulated.)*, 20(4), 147-152
- Thomson, M. (1996) The teaching of spelling using techniques of simultaneous oral spelling and visual inspection. *Australian Journal of Learning Difficulties*, 1(1), 12-14.
- Uno, A., Haruhara, N., Kaneko, M. & Wydell, T.N. (2006). *STRAW - Shougakusei no Yomikaki Screening Kensa [in Japanese] - Screening Test for Reading and Writing for Japanese Primary School Children*. Hiroshima, Japan: Saccuss Bell Publishers.
- Uno. A., Wydell, T.N., Haruhara, N., Kaneko, M. and Shinya, N. (2009) Relationship between reading/writing skills and cognitive abilities among Japanese primary-school children: normal readers versus poor readers (dyslexics). *Reading and writing*, 22, 755-789.
- Wolf, M., & Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology*, 91, 415-438.
- Wolf, M. & Bowers, P.G. (2000) Naming-Speed Processes and Developmental Reading Disabilities: An Introduction to the Special Issue on the Double-Deficit Hypothesis. *Journal of learning disabilities*, 33(4), 322-324.
- Wydell, T. N. & Butterworth, B. (1999) A case study of an English-Japanese bilingual with monolingual dyslexia. *Cognition*, 70, 273-305.

# Dynamic estimation of emphasizing points for user satisfaction evaluations

**Yoshimasa Ohmoto**(ohmoto@i.kyoto-u.ac.jp)

Department of Intelligence Science and Technology  
Graduate School of Informatics, Kyoto University  
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan

**Takashi Miyake** (miyake@ii.ist.i.kyoto-u.ac.jp)

Department of Intelligence Science and Technology  
Graduate School of Informatics, Kyoto University  
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan

**Toyoaki Nishida** (nishida@i.kyoto-u.ac.jp)

Department of Intelligence Science and Technology  
Graduate School of Informatics, Kyoto University  
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan

## Abstract

When many factors must be considered for decision-making, people dynamically change their emphasizing points, along with their understanding of these factors and the relationships between them. In previous work, we proposed a method to dynamically estimate emphasizing points (DEEP) based on utterances, physiological indices, and proposal selections. To evaluate this method in actual interactions, we conducted controlled WoZ (Wizard of Oz) experiments using Embodied Conversational Agents (ECAs), which interactively provide controlled information for decision-making. Using ECAs, we compare our method to an existing method, which estimates emphasizing factors through the “gradual method”. We confirm that our method can accurately estimate dynamic changes of emphasizing points, and that participants were more satisfied with the final proposal from the ECA that used DEEP.

**Keywords:** verbal and nonverbal behavior; physiological indices; preferential structure estimation.

## Introduction

When many factors must be considered for decision-making, we dynamically and interactively change the factors that we emphasize (which we call “emphasizing points”). We also change our understanding of these factors and relationships between them. For example, in travel planning, we have to synthetically consider factors, such as place, budget, members, and schedule. We often make such plans interactively with our friends and travel agency staff.

The interaction between conversational partners influences how we understand the factors and the relationships between them during the decision-making process. Therefore, their emphasizing points are often dynamically changed when faced with new information. However, the important factors may not only be the most recent points emphasized, but the process of interaction may also change the emphasizing points. People have to re-estimate the changes in their emphasizing points throughout the interaction.

In interactive decision-making with dynamic changes to emphasizing points, humans provide active demands and passive responses through verbal expressions, nonverbal reactions, proposal selection, and physiological state (Ohmoto,

Kataoka, Miyake, & Nishida, 2011). In the previous work, we analyzed the interaction process, and verbal and nonverbal behavior during the interaction to propose an estimation method of interaction using utterances, nonverbal behavior, physiological indices, and proposal selections.

The purpose of this study is to evaluate whether our proposed method based on dynamic estimation of emphasizing points (DEEP) is useful to participants in interactive decision-making and whether the proposed method can provide satisfactory proposals for participants. To test this method, we used Embodied Conversational Agents (ECAs), because it is difficult for human agents to achieve rigorously controlled interaction with participants based on our proposed method. Specifically, we conducted an experiment that compares the results of interactive decision-making with two types of ECAs; one provided proposals based on our method and another based on an existing method that gradually estimates emphasizing points based on verbal expressions and proposal selections.

The paper is organized as follows: Section 2 discusses work on interactive systems. Section 3 briefly explains DEEP, which dynamically estimates emphasizing points. Section 4 describes the experiment for comparing two types of estimation methods and then presents the results. Section 5 discusses the achievements and limitations of the proposed method. Section 6 concludes and discusses future work.

## Related work

Some researchers have developed systems that can provide proposals to satisfy user’s demands. These systems gradually estimate user’s demands throughout the interaction.

Kitamura et al. (Kitamura et al., 2008) developed the “Laddering” Search Service System that matches users queries with search targets by communicating with users throughout the interview. They assume that user’s emphasizing points do not change during the interaction.

Aydogan et al. (Aydogan & Yolum, 2007) proposed an architecture in which both consumers and producers use a

shared ontology to negotiate services. Through repetitive interactions, the provider accurately learns consumers' needs to provide better-targeted offers. The system learns consumers' needs over long-term interactions.

Kurata (Kurata, 2010) proposed a computer-aided tour planning system. The system provides several tour plans and asks the user to provide feedback. The feedback is utilized by the system for inferring the user's preferences and then for revising tour plans. This cycle is repeated until the user is satisfied with the final plan, with the hopes that this method gradually leads to a more satisfying experience of computer-aided tour planning. The system can then estimate user's emphasizing points. However, the user has to manually change emphasizing points when the user wants to change her/his emphasizing points during the interaction. Moreover, the user cannot modify their emphasizing points when he/she does not have knowledge about the planning.

Previous work revealed that user demands and needs could gradually be estimated through repetitive interactions. However, most of the research did not consider that user's demands and needs could change throughout the interaction. In contrast, we assume that emphasizing points can change over the interaction and we dynamically estimate these changes. We focus not only on active demands verbally expressed and proposal selections, but also on passive responses expressed by backchanneling, and nonverbal reactions.

It is, however, difficult to estimate human internal states through nonverbal information, especially when passively interacting with others. Therefore, we use physiological indices for estimating human internal states during interaction. There are various studies on estimating human internal states by measuring physiological indices (e.g. (Iwaki, Arakawa, & Kiryu, 2008)). There are also several studies that use these measured physiological indices for effective human-agent interaction.

Bosma et al. (Bosma & Andre, 2004) proposed a method that takes into account users' emotional state to disambiguate dialogue acts. They restrict to pedagogical agents that offer a text-based natural language interface for assisting the user in text communication. They estimated levels of arousal and valence by using physiological indices: skin conductivity response (SCR), heart rate, muscle activity, and respiration rate.

Prendinger and Ishizuka (Prendinger & Ishizuka, 2005) developed an interview agent which takes physiological data (skin conductance and electromyography) of users in real-time, interprets the data into emotions, and addresses the user's affective states in the form of emphatic feedback. In addition, they evaluated the agent by using SCR and heart rate. The empathic feedback has a positive effect on the interviewee's stress level while hearing the question.

As mentioned above, physiological indices are useful for estimating human internal states in interaction even when users passively interact with others. The proposed method uses physiological indices, SCR, electrocardiograms (LF/HF values), and skin temperature of fingers, to detect mental

stress, such as pleasure, excitement, and tension. The method estimates emphasizing points by using these physiological indices, as well as verbal expressions, and nonverbal responses.

We have discussed the achievements and limitations of previous work related to our objective of estimating emphasizing points for interactive decision-making. Because of the difficulty in detecting passive responses during interactions, most prior work estimated user demands and needs gradually through repetitive interactions that required active demands from users. Therefore, we propose a method that dynamically estimates emphasizing points by using physiological responses, which could detect human internal states even during a passive interaction, *in addition to* verbal expressions, and nonverbal responses. In this study, we apply the proposed method to actual interactions and experimentally evaluate whether proposals that use physiological responses are useful for participants' decision-making and for achieving satisfactory results in the interaction.

### **Dynamically estimating emphasizing points**

For our purpose, we conducted preliminary analyses to elicit useful information for dynamically estimating emphasizing points (DEEP) in human-human interaction (Ohmoto et al. 2011). As a result of the analyses of videos and physiological indices, we could suggest a method to DEEP which is explained next subsection. We proposed a method to DEEP based on the observation of human-human interaction in preliminary analyses, so, we think that the proposed method is one of methods realizing DEEP. In this section, we briefly explain the proposed method to DEEP based on verbal reactions, body movements, and physiological indices, when participants are given two proposals and asked for his/her selection and demands.

DEEP, in this paper, is applied to the situation in which many factors, including unknowns, for must be considered for decision-making. In this situation, a user interacts with a system based on DEEP and the system advises some useful proposals for user's decision-making. A proposition process in an interaction is as follows: First, the two most appropriate proposals at that point are explained from a DEEP system. After the proposition, the system asks the user what his/her demands were and which proposal is better. The DEEP system pays attention to the user's reactions and answers during the explanation and questions. The system then estimates the emphasizing points. The user repeats this process until one of the propositions satisfies the user's end goal.

### **Overview of DEEP**

The degree of emphasis for an emphasizing point is rated on a scale from zero to five. The rating is changed based on the following three factors during the explanation.

- **Verbal reactions**

Either of the two following reactions occurs.

- Listed words appear in answers or demands.

- The participant provides backchanneling phrases, which express acknowledgement, surprise, or understanding, such as “ah,” “oh,” “aha,” “I see,” and “I understand.”

- **Body movements**

The participant repeatedly nods three times or more.

- **Physiological indices**

Either of the two following responses occur (refer to (Miyata, 1998), (Lin, Omata, Hu, & Imamiya, 2005), (Iwaki et al., 2008) and (Nakazono, Hada, Ataka, Tanaka, & Nagashima, 2008)).

- SCR increases more than 10% compared to resting level.
- LF/HF value (electrocardiograph measurement) is more than 6.0.

Verbal reactions, body movements, and physiological indices, are used as criteria for determining when a new factor is discovered and should be emphasized, and for determining when a user’s degree of emphasis of a particular factor increases or decreases.

**Rules for changing estimated emphasizing points during explanation** The estimated emphasizing points are changed by the participant’s responses when a DEEP system explains the proposals.

- **Discovery of a new factor to be emphasized**

When any one of the three criteria appears during an explanation, the system decides that the factor should be slightly emphasized, and increases the degree of emphasis from zero to two. When any two or three criteria are present, the system increases the emphasis from zero to three.

- **Increasing or decreasing degree of emphasis**

When any one of the three criteria appears, the system decides that the factor should be emphasized, and increases the emphasis of the factor by one. When there are physiological reactions, but no verbal reactions, or body movements, the system decides that the factor should be emphasized less, and decreases the emphasis of the factor by one.

**Rules for changing estimated emphasizing points from active demands** The system asks whether a user has any demands. From the user’s response, the system determines what the user’s demands are and what changes there are to emphasizing points. The system uses assumed keywords in the user’s response to determine demands and changes to demands. Assumed keywords are words that express assumed emphasizing points, demands, and basic words necessary to capture demands. Words that are not expected to be included in answers are ignored.

- **Discovery of new factors to be emphasized**

When the emphasis degree of the discovered factor is zero, the system increases the degree of emphasis from zero to three.

- **Increasing or decreasing degree of emphasis**

When the emphasis of the discovered factor is greater than zero and the system decides that the factor should be increased, the system increases the degree by one. When the system decides that the emphasis of the factor should be decreased and the degree is greater than zero, the system decreases the degree by one.

**Deciding a better proposal by the user’s choice between the two proposals** Given two proposals, the system asks the user which is better. If the proposal satisfied the user’s end goal, that is the final proposal. If not, based on the answer, the system determines which proposal more satisfies the user or decides either that both proposals equally satisfy or that neither proposal is satisfactory. When the system determines that both proposals equally satisfy the user, the proposal in which the lowest skin temperature was recorded is regarded as better. When the system determines that neither proposal satisfied the user, the system does nothing.

### **Selecting the next step based on DEEP results**

According to the criteria mentioned above, changes to user’s emphasizing points are estimated after the proposals are given and data is collected from the user’s reactions and response. After the estimation, the next two proposals are selected based on the estimation results.

The next proposals are selected using a table of orthogonal arrays in advance. Orthogonal arrays are a special set of Latin squares, which can be used to estimate main effects using only a few experimental runs. From the table, the two proposals that most satisfy user’s emphasizing points are picked. When many proposals in the table can satisfy a user’s emphasizing points, the two proposals nearest to the best proposal for a user’s choice are selected. When neither proposal will satisfy the emphasizing points, the two proposals furthest from the previous proposition are selected. The distances of proposals are calculated by cosine similarity.

## **Experiment**

The purpose of this experiment was to investigate whether the DEEP method could accurately estimate emphasizing points in which many factors, including unknown factors, must be considered for decision-making. In the experiment, we used human-like virtual agents (ECAs) to strictly control the verbal and nonverbal expressions of the agent, which could affect user’s impressions of the proposals presented. The ECAs were operated by a WoZ (Wizard of Oz) interface because accurate voice recognition can be difficult. The proposed method was compared with the gradual method, which was discussed above, and is described in more detail below.

### **Task**

Participants were asked to design a mobile robot using a robot parts catalogue. Each participant interacted with an experimenter for two sessions, in which they designed a different robot that achieved different tasks. The participant could



change the design concept of the robot during the session without informing the change to the experimenter. The task had 23 criteria that the robot must meet and there were various ways to design robots that realize the same purpose. Examples purposes in Situation A were "taking photos of beautiful scenery" and "introducing old temples and shrines," while in Situation B, examples purposes were "a mountain climbing race" and "a city obstacle race."

### The gradual method compared with DEEP

We compared the DEEP method with gradual method. In the gradual method, the ECA provides the two proposals nearest to the best proposal of the user. When the user decides that neither proposal will suffice, the two proposals furthest from the last two proposals. This method only uses user's selection between the two proposals and gradually approaches a satisfactory proposal. The method does not pay attention to the dynamic changes of user's emphasizing points during the interaction. Therefore, only the user's actual choice is taken into account. This method can provide a better proposal than previous one in most cases. This is a better point than the DEEP method. This method was regarded as a modified version of work by Kurata (Kurata, 2010).

### Outline of WoZ

The experimenter entered into the system data that contained verbal reactions, body movements, and physiological indices, because we could not robustly capture this data in real-time. Each ECA generated verbal and nonverbal behavior that had been previously designed by the experimenter based on the expected reactions.

Both ECAs accepted the results of user's choice. In addition, the ECA with DEEP accepts data as was described in previous section. Verbal reactions and body movements are determined via visual observation. Physiological indices were automatically measured and the experimenter annotated which words or explanations may have triggered the physiological responses. Each ECA used the entered data to decide the proposals presented in the next proposition.

### Experimental settings

The experimental setting is shown in Figure 1. The participant sat in front of a 100-inch screen displaying the ECA. The experimenter sat out of view of the participant and entered the stimuli via a WoZ interface. Two video cameras recorded the participant's behavior; one was placed on the screen for recording the participant's behavior, and another was placed behind the participants for recording the screen. The participant's voice was recorded by microphones. Polymate was used to measure SCR, the electrocardiogram, and skin temperature of fingers. The experimenter instructed the participant to keep their left arm on an armrest.

### Participants

26 students (20 males and 6 females) participated in the experiment. They were undergraduate students from 18 to 25

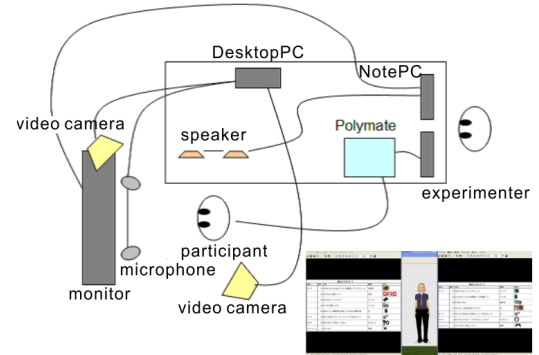


Figure 1: Experimental settings

Table 1: T-test results for accuracy in estimating emphasizing points

	proposed	gradual
average	2.1	1.0
standard deviation	0.69	1.0
t	2.49	
p	0.029*	

years old (an average of 20.6 years old). They did not know about robots but they were in science course. All of them interacted with both of ECA with DEEP and without DEEP.

### Procedure

After a brief explanation of the experiment, the experimenter began the experiment. Two sessions were conducted during the experiment. The experimenter randomly decided which ECA, DEEP or the gradual method, was used for the first session, and the other ECA was used for the second session. The participant repeatedly selected proposals provided by the ECA until he/she was satisfied his/her end goal for the robot. At the conclusion of each session, the participant completed a questionnaire regarding the ECA's evaluations.

### Results of accuracy in DEEP

We randomly picked seven participants before the experiment. These seven participants chose their top three emphasizing points out of 23 factors at the end of both session. The reason why we picked up a limited number of participants is that the choice of emphasizing points was very time consuming process because they had to understand the meanings of 23 factors and reflect on their decision-making. Therefore, we could gather a limited number of participants for the research. We then calculated concordance rates between the factors chosen by the user and the factors estimated by each ECA. We conducted a t-test to compare the concordance rates of DEEP with that of the gradual method. Results are shown in Table 1. Average values show the average number of matched factors.

Table 2: Chi-squared for the effect of method on dynamic changes

	changed	not changed
proposed	25	1
gradual	22	4
p	0.158	

Table 3: Sign-test for comparison of ECA method

	score (proposed > gradual)
average	1.0
standard deviation	1.9
p	0.013*

The results of a t-test confirmed that DEEP more accurately estimates emphasizing points than does the gradual method. We suggest that DEEP has sufficient performance for estimating emphasizing points because the average is high and the standard deviation is low. Therefore, by using verbal reactions, body movements and physiological indices, DEEP can correctly estimate the emphasizing points of each participant.

### Questionnaire results

The participants answered three rating questions on the ECA's behavior using a seven-point scale. The scale was presented as seven ticks on a black line without numbers, which we scored from -3 to +3.

Each of the three questionnaires contained two kinds of questions; one was on how much the ECA affected participant's thought ("how much" question), another was regarding which method had more affected participant's thought ("which" question).

**Changing emphasizing points and purpose of robot** Participants answered whether they dynamically changed their emphasizing points and purpose of the robot throughout the interaction ("how much" questions). We performed Chi-squared test to confirm that there was a significant difference between DEEP and the gradual method, and the results are presented in Table 2. Participants also answered which method caused more dynamic changes ("which" question); we performed sign-test to calculate the difference between the two methods, which is shown in Table 3 (when the gradual method caused most changes: -3 - when DEEP caused most changes: +3).

There is no significant difference between the "how much" scores, because both methods could cause dynamic changes during the interaction. This means that humans easily change their emphasizing points even when simple algorithms provide the proposal and explanation. Meanwhile, DEEP caused significantly more changes than did the gradual method. It is possible that participants pay attention to broader factors than contained in the mobile robot task because the proposed

Table 4: Wilcoxon signed-rank test results on user satisfaction of ECA's final proposal

	proposed	gradual
average	1.8	0.81
standard deviation	2.3	1.6
z	2.11	
p	0.035*	

Table 5: Sign-test results on which ECA provided the best proposal

	score (proposed > gradual)
average	1.1
standard deviation	2.3
p	0.038*

method was sensitive to changes in emphasizing points and modified subsequent proposals accordingly.

**Participant satisfaction of ECA's final proposal** Participants answered how satisfied they were with the ECA's final proposal ("how much" questions). The results of a Wilcoxon signed-rank test are shown in Table 4 (not at all: -3 - very much: +3). Participants also answered which method provided a more satisfactory proposal ("which" question). We performed a sign-test, and the results are shown in Table 5 (satisfy the final proposal of the ECA with gradual method: -3 - satisfy the final proposal of the ECA with DEEP: +3).

Both of Table 4 and Table 5 show that the ECA with DEEP provided a significantly more satisfactory proposal than the ECA with the gradual method. However, it is important to note that the standard deviation for the results of the ECA with DEEP in Table 4 and Table 5 are fairly large. We return to the implications of this result in the discussion.

**Naturalness of ECA's proposals** Participants answered how natural the sequence of proposals was ("how much" questions). We performed a Wilcoxon signed-rank test, and the results are shown in Table 6 (not at all: -3 - very much: +3). Participants also answered which method provided more natural proposals ("which" question). The results of a sign-test are shown in Table 7 (the ECA with gradual method provided most natural proposal: -3 - the ECA with DEEP provided most natural proposal: +3).

Both Table 6 and Table 7 show that the ECA with DEEP provided significantly more natural proposals than the ECA with gradual method. The each content of proposals were the same between the proposed method and gradual method. Therefore, naturalness must be attributed to presentation order and whether the proposals reflected their emphasizing points. The proposed method most likely provided more natural proposals because DEEP could quickly reflect changes in their emphasizing points.

Table 6: Wilcoxon signed-rank test results on naturalness of ECA proposals

	proposed	gradual
average	1.2	0.27
standard deviation	1.8	1.6
z	2.4	
p	0.015*	

Table 7: Sign-test results on which ECA provided more natural proposals

	score (proposed > gradual)
average	0.89
standard deviation	1.7
p	0.027*

## Discussion

In this study, we evaluated one method for estimating emphasizing points based on verbal and nonverbal information and physiological indices. As a result, we confirmed that our proposed method improved the accuracy of estimating emphasizing points, has more latitude in changing emphasizing points, is natural, and participants are more satisfied with the final proposal. In addition, we find evidence that people often change their emphasizing points and purpose of the task during the interactive decision-making process.

The proposed method considers changes of emphasizing points. Therefore, the proposed method often provided proposals that included new combinations of factors which the participant did not specially emphasize. One participant reported "I was often surprised at the dynamic changes of the proposals." The surprise sometimes causes uncomfortable feelings so we will have to consider proposal history and provide additional explanations for the change.

The standard deviations in proposed method are relatively large. This means that the effectiveness of the proposed method is different across individuals. One of the reasons was that some participants' demands could not be satisfied by the ECA. In those cases, the ECA did not provide any notification of impossibility or alternatives. In many possible cases, the ECA with DEEP quickly responded to participants' demands, so, in some impossible cases, the participants who had impossible demands felt disappointed, as would be expected. Future work should include notification capabilities.

## Conclusion

In this study, we evaluated whether our proposed method, which estimates dynamic changes of emphasizing points based on verbal reactions, body movements, and physiological indices, is useful for interactive decision-making and for selecting a proposal that satisfies the user's end goal. For this purpose, we conducted an experiment that compared two

methods: our method and an existing method that gradually estimates emphasizing points based on participants' proposal choice. As a result, we confirmed that DEEP improved estimation accuracy, user satisfaction, and naturalness of proposals. We propose that interactive decision-making be based on estimation of emphasizing points.

One important issue that should be explored in future work is more clearly define the criteria for noting verbal and non-verbal behavior. Physiological indices are very useful for estimating internal states of human but measuring these indices may not be natural in many cases. In future work, we will try to replace physiological indices with synthetic use of some verbal and nonverbal behaviors.

## References

- Aydogan, R., & Yolum, P. (2007). Learning consumer preferences using semanticsimilarity. In *Aamas '07: Proceedings of the 6th international joint conference on autonomous agents and multiagent systems* (pp. 1–8). New York, NY, USA: ACM.
- Bosma, W., & Andre, E. (2004). Exploiting emotions to disambiguate dialogue acts. In *Proceedings of the 9th international conference on intelligent user interfaces* (pp. 85–92).
- Iwaki, M., Arakawa, S., & Kiryu, T. (2008). *Influence on biosignal and working efficiency of sound environment in typewriting* (Tech. Rep.). IEICE technical report. ME and bio cybernetics.
- Kitamura, M., Shimohata, S., Sukehiro, T., Ikeno, A., Sakamoto, M., Orihara, I., et al. (2008). *Design and development of dialogue system for laddering search service* (Vol. 108; Tech. Rep.). IEICE technical report. Natural language understanding and models of communication.
- Kurata, Y. (2010). Interactive assistance for tour planning. *Spatial Cognition 2010 Lecture Notes in Artificial Intelligence*, 6222, 289–302.
- Lin, T., Omata, M., Hu, W., & Imamiya, A. (2005). Do physiological data relate to traditional usability indexes? In *Proceedings of the 17th australia conference on computer-human interaction* (pp. 1–10).
- Miyata, H. (Ed.). (1998). *The new physiological psychology (japanese)* (Vol. 3). Kitaohji-shobo.
- Nakazono, K., Hada, T., Ataka, E., Tanaka, H., & Nagashima, Y. (2008). *Workload evaluation of gaming task by physiological indices and psychological indices* (Vol. 107 - 553; Tech. Rep.). Technical report of IEICE. HIP.
- Ohmoto, Y., Kataoka, M., Miyake, T., & Nishida, T. (2011). A method to dynamically estimate emphasizing points and degree by using verbal and nonverbal information and physiological indices. In *The 2011 ieee international conference on granular computing 2011* (pp. 508–514).
- Prendinger, H., & Ishizuka, M. (2005). The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence*, 19(3-4), 267–285.

# The Vowel-Size Relationship Re-Examined Using Speeded Classification

Yuka Ohtake (4588208926@mail.ecc.u-tokyo.ac.jp)

Graduate school of Education, University of Tokyo  
7-3-1 Hongo Bunkyo-ku, Tokyo 113-0033, Japan

Etsuko Haryu (haryu@p.u-tokyo.ac.jp)

Graduate school of Education, University of Tokyo  
7-3-1 Hongo Bunkyo-ku, Tokyo 113-0033, Japan

## Abstract

The vowel-size relationship has been repeatedly reported: the vowels /a/ and /i/ elicit bigger/smaller images respectively. Previous studies reporting this relationship have required participants to make explicit decisions about the meaning of the target words including these vowels. In the present study, we attempted to re-examine the vowel-size relationship in two experiments using speeded classification tasks. The results of Experiment 1 indicate that participants associated the vowels with a bigger/smaller image even when they were not motivated to pronounce the vowels during the task. The results of Experiment 2 indicate that the proprioception of the absolute size of the mouth may not contribute to the vowel-size relationship. The process underpinning the vowel-size relationship is discussed.

**Keywords:** sound symbolism; vowel-size relationship; speeded classification; kinesthetic experience

## Introduction

The relationship between a word and its referent is said to be arbitrary, but many studies have reported relationships between them. This is referred to as “sound symbolism.” Among these studies, the vowel-size relationship has been repeatedly reported: the vowel /a/ is likely to make us imagine objects of bigger size whereas the vowel /i/ elicits images of smaller objects. Previous studies (e.g., Sapir, 1929; Newman, 1933; Tarte & Baritt, 1971) that reported this relationship have required participants to make explicit decisions about the meaning of target words including these vowels. For example, Sapir (1929) asked participants which table was bigger, /mal/ or /mil/, while Tarte & Baritt (1971) required participants to match CVC trigrams (e.g., /was/ or /wis/) with geometric figures of different sizes. However, these studies suffer from a weakness in the way the process underpinning the vowel-size relationship was investigated: participants had enough time to pronounce or simulate the target words during the task, which makes it difficult to determine which of the following two factors contributed to the vowel-size correspondence.

The first of these factors is the component formant frequencies of the vowels (Tarte, 1982). The second and third formants of the vowel /i/ are higher than those of the vowel /a/. Given that higher frequency sounds correspond to smaller images and lower frequency sounds correspond to bigger images (Gallace & Spence, 2006), frequencies of

vowels may explain why the vowel /a/ is likely to elicit bigger images and /i/ to elicit smaller images.

The second is the contribution of the kinesthetic experience of pronunciation (e.g., Newman, 1933). The vowel /a/ is pronounced with the mouth wide open and the tongue positioned low in the mouth. In contrast, the vowel /i/ is pronounced with the mouth slightly open, and the tongue positioned high in the mouth. Since the oral cavity is larger when pronouncing /a/ than when pronouncing /i/, the vowel /a/ is likely to elicit bigger images than the vowel /i/.

In the present study, we attempted to investigate the vowel-size relationship in a way that distinguished between these two factors. More specifically, in Experiment 1, we examined whether participants would associate the vowels /a/ and /i/ with bigger and smaller images, respectively, even when they were not motivated to pronounce the vowels during the task. In Experiment 2, we examined whether kinesthetic experience around the mouth (i.e., proprioception of the size of the oral cavity when pronouncing the vowels) on its own and without auditory experience could elicit bigger/smaller images.

To examine these problems, we used speeded classification tasks, which have been widely used in studies of cross-modal perception (e.g., Gallace & Spence, 2006). In this kind of task, participants have to discriminate between stimuli in one dimension while trying to ignore an irrelevant dimension, which enables us to see whether their response to the relevant dimension is influenced by the variation of the irrelevant dimension.

In Experiment 1, to investigate whether the vowels /a/ and /i/ elicit bigger/smaller images without the kinesthetic experience of pronunciation, we asked participants to judge the relative size of the target disk, while an irrelevant sound (the vowel /a/ or /i/) was presented simultaneously. If reaction times for judging the size of the target disk were influenced by the variation of the vowels, this would allow us to conclude that the acoustical features of vowels /a/ and /i/ elicit bigger/smaller images without kinesthetic experience.

In Experiment 2, to investigate whether the proprioception of the size of the oral cavity when pronouncing /a/ and /i/ could elicit bigger/smaller images on its own without the subject actually hearing any vowel sounds, we asked participants to judge the relative size of the target disk, while ensuring that they opened their mouths in the same way they would if pronouncing each

vowel. If reaction times for judging the size of the target disk were influenced by the variation in the way the participants opened their mouths, this would allow us to conclude that the kinesthetic experience alone elicits bigger/smaller images without auditory experience.

In the following, we will discuss the possible process underpinning the vowel-size relationship, taking both the above factors into account.

## Experiment 1

In Experiment 1, we attempted to investigate whether the vowels /a/ and /i/ elicit bigger/smaller images without the kinesthetic experience of pronunciation. In the experiment, participants were asked to judge whether a target disk was bigger or smaller than a standard disk. The target disk was presented following the standard disk. It was 10% or 20% shorter or longer in diameter compared to the standard disk. A task-irrelevant sound (/a/ or /i/) was sometimes presented simultaneously along with the presentation of the target disk. If it is the case that the vowel sounds (/a/ and /i/) elicit bigger/smaller images without the kinesthetic experience of pronunciation, the reaction times should have been shorter when the vowel-size relation is congruent (/a/ being presented when the target disk was bigger and /i/ being presented when it was smaller) than when it was incongruent.

## Method

**Participants** Thirty Japanese-speaking undergraduate students (14 males, 16 females; mean age, 22.2 years; range 20-36 years) took part in the experiment.

**Apparatus** The visual stimuli were presented on a laptop computer (Dell Inspiron 1526) with a 15.4-inch screen, or on a desktop computer (VAIO VGC-RA72P) with a 17-inch screen. Auditory stimuli were presented through headphones (Audio-Technica ATH-ANC7 or Sennheiser HDA200). The presentation of the stimuli and the recording of the participants' responses were controlled using Cedrus Superlab 4.0 software.

**Materials** The visual stimuli were the standard disk, the target disks, and the mask. The standard disk was gray and 3 cm in diameter, and the target disks were  $\pm 10\%$  and  $\pm 20\%$  of the diameter of the standard disk. The visual mask was a light-gray screen with dark-gray spray. Four different auditory stimuli were used for presentation of the vowels /a/ and /i/, respectively. The auditory stimuli were a recording of a Japanese female who had been asked to pronounce Japanese vowels. Her speech was recorded on a Roland R-09. The duration of each vowel was 300 ms. For the vowel /a/, the mean fundamental frequency was 240.3 Hz ( $SD = 2.16$  Hz), the mean first formant frequency was 780.3 Hz ( $SD = 46.1$  Hz), the mean second formant frequency was 1374.8 Hz ( $SD = 57.6$  Hz), the mean third formant frequency was 3077.5 Hz ( $SD = 241.7$  Hz), and the

mean intensity was 48.79 dB ( $SD = 2.48$  dB). For the vowel /i/, the mean fundamental frequency was 240.6 Hz ( $SD = 3.59$  Hz), the mean first formant frequency was 416.5 Hz ( $SD = 20.9$  Hz), the mean second formant frequency was 2712.3 Hz ( $SD = 67.6$  Hz), the mean third formant frequency was 3494.0 Hz ( $SD = 75.3$  Hz), and the mean intensity was 49.94 dB ( $SD = 2.43$  dB).

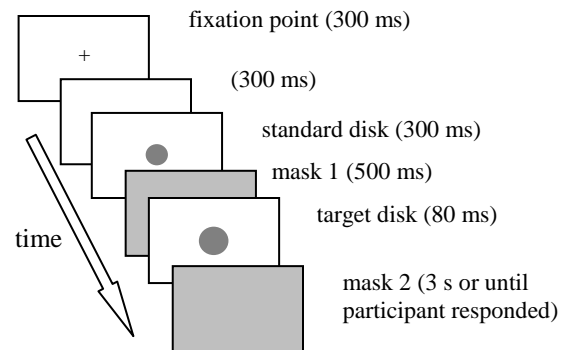


Figure 1: Illustration of the sequence of visual stimuli presented in each trial in Experiment 1 and 2.

**Procedure** The participants sat at a desk, 45 cm from the computer. It took about 10 minutes to complete the entire experiment.

Figure 1 illustrates the sequence of events in each trial. At the start of each trial, the word “Ready?” appeared at the center of the screen, and the participants could choose when to start by pressing the space key. At first, a fixation point was presented in the middle of the screen for 300 ms, followed by a blank white screen. After a 300-ms presentation of the blank screen, the standard disk was presented at the center for 300 ms, followed by the mask screen. The mask screen was presented for 500 ms and was followed by the target disk. The position of the target disk varied randomly (by up to  $\pm 0.3$  cm vertically and horizontally from the center of the screen) to prevent the participants from using superimposition cues to judge the relative size of the target disk. At the same time the target disk was presented, a vowel (/a/ or /i/) was presented in 20 trials for each vowel, and no sound was presented in the remaining 20 trials. The target disk was presented for 80 ms, followed by the mask screen. The mask screen stayed on the screen until the participant responded or until 3 seconds had elapsed, at which point the screen displaying the word “Ready?” appeared and the next trial was ready to begin.

The participants were asked to judge whether the target disk was bigger or smaller than the standard disk as rapidly as possible. The participants were instructed to indicate the relative size of the target disk by pushing “/” with the index finger of the right hand, or “\” with the middle finger of the right hand. Which key corresponded to “big” or “small” was counterbalanced across the participants.

Table 1: The means and standard errors of reaction times (in milliseconds) as a function of condition and size of the target disk in Experiment 1 and 2.

size	Condition					
	congruent		incongruent		control	
	RT		RT		RT	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Experiment 1						
+10%	408.7	29.6	423.9	25.8	397.9	23.7
+20%	375.6	18.2	410.2	20.2	381.3	21.9
-10%	469.8	29.4	492.8	41.0	482.3	36.6
-20%	387.5	18.6	423.1	26.7	397.5	21.6
mean	410.4	21.8	437.5	26.0	414.7	23.4
Experiment 2						
+10%	378.5	18.0	387.3	18.5	385.1	15.0
+20%	362.6	12.2	369.8	19.8	367.0	12.2
-10%	427.4	20.8	411.8	16.8	421.2	22.0
-20%	376.7	18.5	367.7	16.0	370.5	15.1
mean	386.3	15.5	384.1	15.2	385.9	14.6

The participants were informed that a task-irrelevant sound would sometimes be presented, but they were instructed to ignore it. The response times were calculated from the beginning of the second mask screen to the time of the decision. The participants completed 12 practice trials before the experiment to ensure that they clearly understood the task.

The experiment was composed of 60 trials, 15 trials for each size of the target disk ( $\pm 10\%$ ,  $\pm 20\%$ ). The order of trials was randomized for each participant. For each size of the target disk, five trials were presented with the vowel /a/, five trials were presented with /i/, and the remaining five trials were presented with no sound. Each of the trials was classified into three conditions, with 20 trials each: congruent condition (i.e., /a/ being presented when the target disk is bigger and /i/ being presented when it is smaller), incongruent condition (i.e., the opposite combination to the congruent condition), and control condition (i.e., no sound being presented along with the target disk).

## Results

The means and standard errors of reaction times as a function of the condition and size of the target disk are shown in Table 1. Because the error rate was quite low ( $M = 3.1\%$ ,  $SD = 2.9\%$ ), subsequent analysis was only performed on the reaction times.

**Reaction times** The reaction times for the wrong decision (3.1%) and above +3SD from the mean reaction times of each participant (1.2%) were excluded from the analysis.

We performed repeated measures of variance on the reaction times as a function of condition (3) and size of the target disk (4). The analysis revealed a significant main

effect of condition ( $F(2,58) = 7.08$ ,  $p = .002$ ), and a significant main effect of size of the target disk ( $F(3,87) = 10.89$ ,  $p = .001$ ),<sup>1</sup> but no significant interaction between condition and size ( $F(6,174) = .484$ ,  $p > .10$ ).<sup>1</sup>

A post hoc Bonferroni test of condition revealed significant differences between the congruent condition and the incongruent condition and between the control condition and the incongruent condition (all  $ps < .05$ ), with the slowest responses occurring in the incongruent condition. There was no significant difference between the congruent condition and the control condition. A post hoc Bonferroni test of size revealed significant differences between -10% and other sizes (all  $ps < .05$ ), with the slowest responses occurring in -10%. There were no significant differences between any pair of the remaining sizes (+10%, +20%, -20%).

In sum, reaction times were longer in the incongruent condition than in the congruent condition or control condition, and in -10% than in the other sizes.

## Discussion

In Experiment 1, we attempted to investigate whether participants would associate the vowels /a/ and /i/ with bigger and smaller images, respectively, without the kinesthetic experience of pronunciation, using a speeded classification task. The results indicate that they did. The participants responded more slowly in the incongruent condition than in the congruent condition or control condition. The vowel /a/ elicited bigger images and the vowel /i/ elicited smaller images without kinesthetic experience, which could interfere with the response of “big”

<sup>1</sup> A Greenhouse-Geisser adjustment was used to correct for violations of sphericity.

while hearing /i/ and with the response of “small” while hearing /a/. As Tarte (1982) pointed out, the component formant frequencies of vowels can explain the vowel-size relationship.

## Experiment 2

In Experiment 2, we attempted to investigate whether bigger/smaller images would be elicited only with the kinesthetic experience around the mouth when pronouncing the vowels /a/ and /i/ (i.e., the proprioception of the size of the oral cavity), without the subject actually hearing any vowel sounds, using the same task as Experiment 1. In the experiment, the participants completed the same speeded classification task with the kinesthetic experience of pronouncing vowels. We ensured that the participants opened their mouths in the same way they would if pronouncing each vowel by asking them to hold either of two types of solid object in their teeth: one was egg-shaped and the other was board-shaped. In order to hold the egg-shaped object with their teeth, participants had to open their mouth widely, and the resultant lip shape was similar to that when pronouncing the vowel /a/. On the other hand, holding the board-shaped object required participants to open their mouth slightly along the vertical axis and pull their lips sideways. This shape mimicked the lip shape when pronouncing the vowel /i/. If it is the case that the proprioception of the size of oral cavity elicits images of size without auditory experience, the reaction times should have been shorter when the participants were holding the egg-shaped object and the larger target disk was presented, and they are holding the board-shaped object and the smaller target disk was presented, compared with the opposite combinations.

## Method

**Participants** Twenty-four Japanese-speaking adults (13 males, 11 females; mean age, 26.8 years; range 22-42 years) took part in the experiment.

**Apparatus and Materials** The visual stimuli were presented on a laptop computer (Dell Inspiron 1526) with a 15.4-inch screen, controlled by Cedrus SuperLab 4.0. The visual stimuli were the same as Experiment 1. Two solid objects (egg-shaped and board-shaped) made from styrofoam were used to ensure the participants opened their mouths in the same way they would if pronouncing each vowel. The egg-shaped object was 5.5 cm in maximum diameter and 8 cm long, and the board-shaped object was 7.5 cm by 15 cm long and 0.5 cm thick. Twenty-four sets of the two objects were prepared so that each participant could use a new one.

**Procedure** As in Experiment 1, participants sat at a desk, and the experimenter instructed them to indicate the relative size of the target disk as soon as possible by pressing the keys. The sequence of the visual events in each trial was the same as Experiment 1.

The participants completed 12 practice trials before the experiment. The experiment was composed of six blocks of 72 trials, with a short break at the end of each block. Each block had 12 trials, three trials for each size of the target disk ( $\pm 10\%$ ,  $\pm 20\%$ ), and the order of the trials was randomized in each block for each participant. Six blocks were divided into three phases, which had two blocks each. In one phase, participants were instructed to open their mouth naturally, and hold the smaller side of the egg-shaped object in their teeth. In the other phase, participants were instructed to open their mouth slightly sideways, and hold the longer side of the board-shaped object in their teeth. In the remaining phase, participants were instructed to complete the task in the same way as the practice trials, i.e., to hold no object in their mouth. The order of the three phases was counterbalanced across participants. Each of the trials was classified into three conditions, 24 trials for each condition: congruent condition (i.e., the participants are holding the egg-shaped object and the larger target disk is presented, or they are holding the board-shaped object and the smaller target disk is presented, incongruent condition (i.e., the opposite combination to the congruent condition), or control condition (i.e., the participants are not holding anything when the target disk is presented).

## Results

The means and standard errors of reaction times and number of wrong decisions as a function of condition and size of the target disk are shown in Table 1. As in Experiment 1, because the error rate was quite low ( $M = 3.1\%$ ,  $SD = 2.9\%$ ), subsequent analysis was performed only on the reaction times.

**Reaction times** As in Experiment 1, the reaction times for the wrong decision (3.1%) and above +3SD from the mean reaction times of each participant (1.2%) were excluded from the analysis.

We performed repeated measures of variance on the reaction times as a function of condition (3) and size of the target disk (4). The analysis revealed a significant main effect of size of the target disk ( $F(3,69) = 14.8$ ,  $p < .001$ ),<sup>1</sup> but no significant main effect of condition ( $F(2,46) = .05$ ,  $p > .10$ ),<sup>1</sup> and no significant interaction between condition and size ( $F(6,138) = .438$ ,  $p > .10$ ).<sup>1</sup> A post hoc Bonferroni test of size revealed significant differences between -10% and the other sizes (all  $ps < .01$ ) with the slowest responses occurring in -10%, and a marginally significant difference between +10% and +20% with slower responses in +10% ( $p = .08$ ).

These results indicate that the condition did not affect the reaction times, although the size of the target disk affected them as in Experiment 1.

## Discussion

In Experiment 2, we attempted to investigate whether the proprioception of the size of the oral cavity could elicit



bigger/smaller images on its own without the subject actually hearing any vowel sounds, using the same task as Experiment 1. The reaction times did not differ significantly between in the congruent condition and in the incongruent condition. The results indicate that the proprioception of the size of oral cavity when pronouncing /a/ and /i/ may not, on its own, elicit the image of bigger/smaller sizes. However, it should be pointed out that in this experiment we controlled the absolute size of the oral cavity, in other words, we investigated the effect of the *static* kinesthetic experience of pronunciation. It is possible that the *dynamic* kinesthetic experience of pronunciation, that is, the temporal change of the relative size of the mouth, plays an important role in eliciting the image of bigger/smaller sizes.

It is also worth noting that the lack of uncertainty about the variation of stimuli in the irrelevant dimension may have weakened the effect of treatment (Gallace & Spence, 2006). In Experiment 1, there was an uncertainty about the variation of stimuli in the irrelevant dimension, induced by trial-by-trial variation. In contrast, in Experiment 2, the stimuli in the irrelevant dimension were fixed during each of the blocks.

In sum, the results in Experiment 2 indicate that the static kinesthetic experience (i.e., the proprioception of the absolute size of oral cavity) may not contribute to the vowel-size relationship, although it is possible that the dynamic kinesthetic experience could contribute to it. In addition, the lack of uncertainty about the variation of the irrelevant dimension may have weakened the effect of treatment.

## General Discussion

In the present study, we attempted to re-examine the vowel-size relationship in a way that distinguished between two possible factors, formant frequencies of the vowels (Experiment 1) and kinesthetic experience while pronouncing the vowels (Experiment 2), using the speeded classification paradigm.

The results of Experiment 1 indicate that the component formant frequencies of vowels on their own can explain the vowel-size relationship, and the results of Experiment 2 indicate that the static kinesthetic experience (proprioception of the absolute size of oral cavity) may not contribute to the vowel-size relationship.

However, in the results of Experiment 2, the possibility remains that the dynamic kinesthetic experience (the temporal change of the relative size of the mouth) might have elicited bigger/smaller images and had an influence on the results. Furthermore, we cannot completely eliminate the possibility that the dynamic kinesthetic experience may have affected the results of Experiment 1 from the viewpoint of motor theory (e.g., Liberman & Mattingly, 1985), which understands the perception of speech as vocal tract gestures. From this viewpoint, the dynamic kinesthetic experience automatically generated from hearing vowels may have affected the judgments of size, and supported the results of Experiment 1.

Taking the above into account, the vowel-size relationship can be mainly explained by the component formant frequencies and the static kinesthetic experience may not contribute to it, but the dynamic kinesthetic experience of pronunciation may play some role. Further research is needed to evaluate the role of component formant frequencies more exactly by controlling the kinesthetic experience more rigidly, and to investigate the role of the dynamic kinesthetic experience of pronunciation.

## References

- Gallace, A., & Spence, C. (2006). Multisensory synthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, 68(7), 1191-1203.
- Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21, 1-36.
- Newman, S. (1933). Further experiments in phonetic symbolism. *American Journal of Psychology*, 45, 53-75.
- Tarte, R. D., & Baritt, L. S. (1971). Phonetic symbolism in adult native speakers of English: Three studies. *Language and Speech*, 14, 158-168.
- Tarte, R.D. (1982). The relationship between monosyllables and pure tones: an investigation of phonetic symbolism. *Journal of Verbal Learning and Verbal Behavior*, 21, 352-360.
- Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology*, 12, 225-239.

# Building Conceptual Dictionary for Providing Common Knowledge in the Integrated Narrative Generation System

**Kensuke Oishi (g231g010@s.iwate-pu.ac.jp)**

Graduate School of Software and Information Science, Iwate Prefectural University, 152-52 Sugo Takizawa, Iwate 020-0193 Japan

**Yasunari Kurisawa (g031g054@s.iwate-pu.ac.jp)**

Faculty of Software and Information Science, Iwate Prefectural University

**Mami Kamada (g031i301@s.iwate-pu.ac.jp)**

Faculty of Software and Information Science, Iwate Prefectural University

**Itaru Fukuda (g031h134@s.iwate-pu.ac.jp)**

Faculty of Software and Information Science, Iwate Prefectural University

**Taisuke Akimoto (g236i001@s.iwate-pu.ac.jp)**

Graduate School of Software and Information Science, Iwate Prefectural University

**Takashi Ogata (t-ogata@iwate-pu.ac.jp)**

Faculty of Software and Information Science, Iwate Prefectural University

## Abstract

We explain the current version of a conceptual dictionary containing two hierarchies of verb concepts and noun concepts to be functioned in our narrative generation system. It is used for operating naturalness or validity of generated events and realizing or adjusting the intentional defamiliarization. Namely, this dictionary is a mechanism to be able to flexibly adjust a variety of generation from realistic narratives to fantastical narratives as well as the foundation for a narrative event and the elements. In the current version, verb concept dictionary has originally defined 5338 case frames and modified 1158 constraints and noun concept dictionary contains 121573 concepts including 5808 intermediate concepts.

**Keywords:** Narrative generation system; conceptual dictionary; verb/noun concept hierarchy; case frame; constraint.

## Introduction

This paper explains the development of a conceptual dictionary or hierarchical systems of concepts in the narrative generation system framework which is our main research theme. A basic unit for a narrative in the system is an event concept containing a verb concept and noun concepts and the information of these concepts is held in the conceptual dictionary. It is one of the central components in the system.

Narrative is the strongest method for organizing fragmentary knowledge human being has. We have been developing a narrative generation system as an intelligence tool for the creation of future literature & narrative (Ogata & Kanai, 2010). For digital art and entertainment such as computer game, perhaps narrative and story can become an important element in the same manner as traditional genres. Our research of narrative generation is done for the application to novel contents such as computer game and narrative generation based narrative or literature.

As a related work, Okada and Endo (1992) proposed a system to generate stories like Aesop fables. A story generated is a kind of simulation of the process that a main character or actor plots a sequence of planning actions. In contrast, our narrative generation system architecture is constructed as an organic

fusion of diverse narrative knowledge and techniques including planning, discourse structure, story grammar, script, discourse relation, and so on. Although the Aesop system has a feature as an application of conceptual dictionary research, our goal of the narrative generation system is pursuing the mechanism of narrative generation itself. Our extreme purpose is not developing conceptual dictionary itself, but creating narrative generation system. Therefore, a basic policy here is to use existing dictionaries as possible to customize and expand them according to the architecture and mechanism.

As a narrative generation study, Oz project (Bates, 1992) attempted the development of an interactive drama with dialogue and actions in autonomous agents. This system mainly focuses on the interactive techniques for the user's narrative experiments. In contrast, our system contains a variety of narrative and linguistic knowledge for generating deep and conceptual narrative structures. On the other hand, BRUTUS (Bringsjord & Ferrucci, 2000) is an interactive narrative generation architecture which has an integrative feature including story grammar, planning, and so on. However, it deals with only a specialized narrative theme, "betrayal". Whereas, we are intend to develop a more general mechanism for various types of narratives.

The goal of this paper is the proposition of a conceptual dictionary in the narrative generation system architecture we have been developing. The conceptual dictionary has two components of verb concept dictionary and noun concept dictionary and each system has a hierarchical structure based on single inheritance. A main issue in the development of conceptual dictionary is currently defining constraints, which means the knowledge for deciding the range of value for each case in an event as a basic unit in a narrative. A constraint is described in a verb concept and prescribes the possible range of noun concepts. In this paper, we describe the whole structure and some detailed parts of the conceptual dictionary by especially putting a focus on the description and role of constraints. Although this paper uses existing studies of conceptual dictionaries as a reference, in the combination with the domain of narrative generation system, a variety of novel and difficult issues emerges. For

example, realistic narratives and fantastical narratives respectively need different widths of conceptual constraint. This issue is directly treated in other papers (Zhang, Ono & Ogata, 2011, 2012). A characteristic of this study is an exploratory approach through the incremental system development, and it is hoped that models and theories progress through the repetition of design-implementation-experiment. The proposed conceptual dictionary in this paper provides a foundation for such incremental process toward more principled modeling.

## Narrative Generation and Conceptual Dictionary

In the macro design, the architecture of narrative generation system consists of conceptual generation phase and surface representation phase by natural language, visual media, and music. The former is divided into story as the narrative content to be narrated and discourse as the structure that a story is narrated. According to this framework, we have been developing various mechanisms and modules independently of each other. Currently, we are advancing the developing of an integrated narrative generation system in which a variety of modules are organically blended into a whole. The goal of this system is to execute the generation of narrative structures in some levels such as story, discourse, and surface representations in a unified method. Story and discourse are represented in formally same a tree structure description form. Each leaf node in the structure is corresponding to an event or state and each intermediate node is corresponding to a relation combining the child nodes including events, states, and the sub-structure. An event as a unit of conceptual representation is the most fundamental element of story and discourse. It is described in the form of case frame which is linked to a verb concept and some noun concepts in the conceptual dictionary. When story or discourse generation mechanism generates an event concept, the conceptual dictionary provides or constraints the semantic elements. A case frame for a verb concept is created and a constraint in the verb concept combining with noun concepts decides the semantic range of each noun concept in the case frame.

The proposed conceptual dictionary is actually positioned in a pilot version of integrated narrative generation system (Akimoto & Ogata, 2011b). It is implemented by Common Lisp with about 800 functions. The main macro modules are control mechanism, conceptual generation module, and surface representation module. The control mechanism calls each module according to a set of parameters as the goal of generation to automatically execute a generation process from story to expression. The conceptual generation module contains next modules to generate or transform narrative conceptual structure: a story generation mechanism including a single event generation mechanism using rhetorical techniques (Zhang, Ono & Ogata, 2011, 2012), a state-event transformation mechanism (Onodera & Ogata, 2012), an events sequence generation mechanism by story grammar based on Propp's narratology (Propp, 1969; Imabuchi & Ogata, 2011), and a discourse mechanism containing 13 kinds of narrative discourse techniques redefined Genette's narrative discourse theory (Genette, 1972; Akimoto & Ogata, 2011a). The surface representation mechanism consists of a sentence generation mechanism, an animated movie generation mechanism (Ogata, 2008), and a cyclical mutual transformation mechanism between conceptual narrative and music (Ogata, Akimoto & Seito, 2011).

Our narrative generation project is a longitudinal and exploratory study, and we have been employing both top-down approach relevant to the system's macro design and bottom-up one programming various modules in parallel. Although the above system is a kind of "bricolage" in the current state, the conceptual dictionary plays an important role in an organized integration of some modules developed independently. The standardizing of an event form and the reference of a common conceptual dictionary by various modules enables the combination as a whole narrative generation system. This is the most fundamental role of the conceptual dictionary and advanced topics of research for upgrading narrative generation emerge on the ground.

## The Structure of Conceptual Dictionary

As described above, the basic unit of a narrative is an event or an event concept described as a conceptual representation. A story and a discourse are represented as each tree structure (Figure 1). Each terminal node is corresponding to an event and each intermediate node is a relation for binding some the lower nodes. Each event is represented as a case frame with a verb concept and some noun concepts. We prepare next eight cases: agent (an subject in an action), counter-agent (an object in an action and a living thing), object (an object in an action and a no living thing), instrument (an tool used in an action), location (a place of an action), time (a pair of starting time and ending time of an action), from (a starting place of an action), and to (an ending place of an action). In addition, we prepare seven kinds of optional cases for treating such concepts which are not able to describe using above cases (Table 2).

The conceptual dictionary performs as a background for each event in narrative generation. It is divided into verb concept hierarchy and noun concept hierarchy. Both have a hierarchical structure based on is-a relations by the mechanism of single inheritance. A verb concept defines one or more case frames which become the templates of the event concept(s). In addition, it has the description of constraint condition (simply constraint) for limiting the range of each case's value. A constraint is defined by one or more noun concepts in noun concept hierarchy and means the range that an element in each case contained in a verb concept can refer inside the noun conceptual hierarchy. The current version of verb concept hierarchy has 5338 case frames and 4881 constraints. On the other hand, current noun concept hierarchy contains 121573 noun concepts including 5808 intermediate concepts.

In a verb concept, if noun concepts inside the range of constraint are used, a natural and possible event like "A knight fight an enemy" is formed. But, if noun concepts outside the range are used, an unnatural and impossible event like "A knight fight a windmill" is created. We show another example. In an event concept (eat (agent old-person) (object rice-ball)) ("an old-person eats a rice-ball"), when (eat (agent N1) (object N2)) is defined as the case frame corresponding to "eat" and the noun concept "food" is defined as the constraint of N2, the

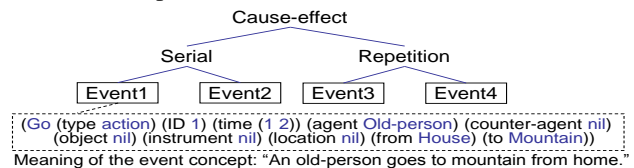


Figure 1: An example of narrative conceptual structure

noun concept to inserted into N2 is the subordinate concept of “food” in the noun concept hierarchy. Therefore, we can confirm that this event concept is constructed inside the constraint.

By the way, fantastical events or physically impossible ones appear in narratives normally. One of concerns for us in our narrative generation project is to be able to flexibly generate from more realistic narratives or events sequences to more fantastical ones. Although the objective of this paper does not discuss the topic, it has been treated in the study of advertising scenario generation system as an application of the narrative generation system in our group (for details, Zhang, Ono & Ogata, 2011, 2012). One of main objectives of this research is to adjust the semantic range in a single event using the concept and techniques of “defamiliarization”. Its idea was acquired from a rhetorical analysis of actual television commercial films. The analysis defined three types of standard rhetoric, which are for generating events corresponding to ordinary events, and nine types of irregular rhetoric, which are for generating events corresponding to extraordinary events. We call twelve types of rhetoric “product introduction rhetoric (or technique)”. In advertising narratives, events with a product are represented through the rhetoric. The former types of rhetoric draw the processes of “manufacturing”, “purchasing”, and “usage” of a product. In contrast, the latter types of rhetoric perform a kind of deviance by techniques of defamiliarization for “the actor’s action”, “the state of product”, “the background or place”, “the actor himself (herself)”, and so on. Defamiliarization means a literary technique for changing a familiar object into unfamiliar one to reinforce the impression. For example, the impression of a familiar product is reinforced by the application of defamiliarization techniques to the objects and agents. This idea will be able to generalize to narratives other than the advertising narrative.

We show the overview of the processing flow. First, user selects a product name from a products list prepared by the developer and the number (1-3) of a type of standard rhetoric. Based on the information, the system generates an ordinary event by acquiring a verb concept and some noun concepts within the standard rhetoric for the product. These concepts are prepared according to each product. Next, when the user designates an irregular rhetoric, the system rewrites the ordinary event to an irregular or extraordinary event by applying the corresponding rhetorical technique. Conceptual dictionary is used in the application of irregular rhetoric. Specifically, the system replaces an original concept in the target event with another concept in the different category by changing the reference region in the noun concept hierarchy. For example, to generate an event with a product “car” according to “the defamiliarization of location” rhetoric, first, the system generates an event like “(drive (agent woman) (instrument car) (location plateau))” using a noun concept “car” inside the constraint. Next, the system replaces a noun concept corresponding to the location with another concept outside the constraint using the defamiliarization processing to generate an event at the strange place like “(drive (agent woman) (instrument car) (location seabed))”. The condition of this processing is that the constraint in regular rhetoric is limited in the actual range. In this research, the conceptual dictionary basically defines the standard semantic range for each noun concept in an event from the viewpoint of actual possibility or physical possibility principally. On the other hand, extraordinary or irregular events are constructed by several types of defamiliarization rhetoric based on a kind of

constraint relaxation. Although a weak point of the current version is that the defamiliarization techniques are randomly applied, we succeeded at the experimental implementation of a simple framework of defamiliarization processing using the conceptual dictionary.

In addition, a story generation system by McIntyre and Lapata (2009) generates stories by using a knowledge base about compositions of an event sentence and chains of events extracted from a narrative corpus based on co-occurrence of words. Stories are generated by a kind of tree search of possible stories. Here, the system has several scoring criteria for pruning low scored branches that are strength of connection of words in an event and words between adjacent events, interestingness, and coherence.

## Existing Conceptual Dictionaries

Table 1 compares existing Japanese conceptual dictionaries. “Goi-Taikei, A Japanese Lexicon” (Ikehara et al., 1999) is the largest scale Japanese lexical and conceptual dictionary for Japanese-to-English machine translation, which hierarchically organizes semantic attributes based on single inheritance. We mainly referred to it as a starting point to construct the overall structures of two hierarchies. The lexicon is basically a lexical dictionary, and because the definition of constraints and the granularity of noun concepts’ categories are too rough to directly use in the narrative generation system, we considerably refined their organizations and uniquely added case frames for verb concepts. Especially, for the purpose of complementing the shortage of intermediate concepts in the noun conceptual hierarchy to a large degree, we referred to “Japanese WordNet” (Bond et al., 2009). Although we used only Japanese lexicon and Japanese conceptual dictionaries as a reference, narrative has an aspect which transcends linguistic difference and we also need to them.

## A Verb Concept Hierarchy

Figure 2 shows the overall structure of verb concept hierarchy. Although the Japanese lexicon contains about 12000 verb concepts under 36 categories, the current version of our verb concept hierarchy contains 4260 verb concepts in 6 categories to represent physical state changes (“physical transfer”, “possessive transfer”, “change of attribute”, “change of body”, “body motion”, and “generation”). A verb concept defines a sentence pattern (“A Japanese Lexicon” contains about 12000 sentence patterns), which is the template for a sentence that takes a verb as the predicate, one or more sets of case frames, the constraints, and the superordinate concept (by an “is-a” relation). The number of case frames is 5338 for 4260 verb concepts.

In Figure 3, we show the form by an example of a verb concept’s description by Common Lisp. The information relating to a verb concept is stored into a variable. In this example, next information is stored into the variable: (1) A corresponding sentence pattern; (“N1 eat N2”), (2) Two case frames; (“(eat(2) (agent N1) (counter-agent N2))” and “(eat(2) (agent N1) (object N2))” (“(2)” means the second meaning of “eat” covering some meanings), (3) Constraints for each case, and (4) the information that the superordinate concept is a verb concept category (“body motion”).

## Defining Case Frames

In the description of a case frame, we decided adequate cases corresponding to respectively a noun concept based on the

Table 1: Existing conceptual dictionaries

	<b>EDR Electronic Dictionary</b> (National Institute of Information and Communications Technology, 2001)	<b>Japanese WordNet</b> (Bond et al., 2009)	<b>Goi-Taikei, A Japanese Lexicon</b> (Ikehara et al., 1999)
<b>Number of concepts</b>	About 410,000	57,238	About 3,000
<b>Number of words</b>	About 270,000	93,834	About 300,000
<b>Structure of concept hierarchy</b>	Multiple inheritance	Multiple inheritance	Single inheritance
<b>Feature</b>	Bilingual dictionary, collocation dictionary, corpus, etc. are recorded.	The words are divided into group by synonym (synset).	About 12000 sentence patterns corresponding to the meaning of verb concept are recorded.

sentence pattern and the constraints. We show an example of the definition based on a sentence pattern, “N1 eat N2”. The constraints of N1 and N2 are respectively “(human, animal)” and “(food, life)”. This means that N1 takes the subordinate concepts of “human” or “animal” and N2 takes the subordinate concepts of “food” or “life”. In the N2, we distinguish “a living thing” from “a nonliving thing” and define two types of case frames, “(eat(2) (agent N1) (object N2))” and “(eat(2) (agent N1) (counter-agent N2))”. Here, the constraints of N1 in two frames are “human” and “animal”. However, in the N2, the constraints of two frames are respectively “food” and “life”. As mentioned above, a verb concept sometimes holds some different case frames.

Currently, case frames for 174 verb concepts are not defined because it is difficult to find an adequate case for a noun

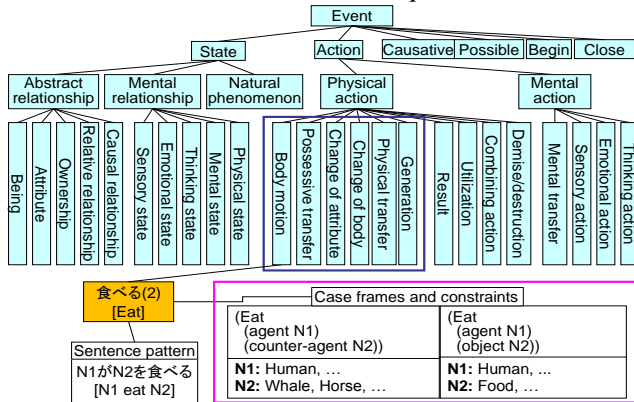


Figure 2: The overall structure of verb concept hierarchy

<b>Description format</b>
(set (intern <verb concept>) ((name (<verb concept's name>))
(sentence-pattern <sentence-pattern>)
(case-cons-set
((case-frame ((agent [<noun term> nil]) (counter-agent [<noun term> nil]) (location
[<noun term> nil]) (object [<noun term> nil]) (instrument [<noun term>
nil]) (from [<noun term> nil]) (to [<noun term> nil])))
(constraint ((<noun concepts>))))
(is-a (<superordinate concepts>))))
<b>An example</b>
(set (intern “食べる[eat](2)”) ((name (食べる[eat] (2)))
(sentence-pattern “N1が N2を 食べる”[N1 eat N2])
(case-cons-set
((case-frame ((agent N1) (counter-agent N2) (location nil) (object nil) (instrument nil)
(from nil) (to nil)))
(constraint ((“人<human>” “死人<dead person>” “人間<personal pronoun>”
“準人間<demi-human>”) (“鯨<whale>” “馬<horse>” “獣<cattle>”
“豚<pig>” “羊<goat>” “羊<sheep>” “鹿<deer>”
“猪<wild boar>” “兎<rabbit>” “鳥<bird>” “家禽<poultry>”
“鳥<game bird>” “魚<fish>” “魚<legendary>”
“たこ・いか・えび・かに<octopus・squid・prawn・crab>”))))
((case-frame ((agent N1) (counter-agent nil) (location nil) (object N2) (instrument nil)
(from nil) (to nil)))
(constraint ((“人<human>” “死人<dead person>” “人間<personal pronoun>”
“準人間<demi-human>”) (“食料<food>” “調味料<seasonings>”
“飲料・たばこ<drink・cigarette>”))))
(is-a (身体動作<Body motion>))))

Figure 3: The description of a verb concept

concept within the current cases. Such noun concepts are categorized into following seven types: “adverbial concept”, “possessive case”, “event”, “purpose of action”, “experiencer”, “target of comparison”, and “idiom”. We show an example of “purpose”. In a verb concept, “go(3)”, the sentence pattern is “N1 go to N3 from N2 for N4” and the constraint of N4 is “(abstraction)” which shows the purpose of “go”. For instance, “a man goes to a mountain from a house for mowing”. If we interpret it as “to”, we can not distinguish N4 from N3 (“(place, field)”). Table 2 set straight above seven cases to be expanded and the relevant studies. Okada (1991) and Takeuchi (2011) show other case frames. The former uses ten types of cases. And, the latter uses 71 types of cases to represent the argument structure of verbs which defines the relationship between a verb and nouns in a sentence. Because the description of case frames by the reference may become too long and complex, we tentatively describe seven types of case frames with “optional” sign.

In addition, for example, FrameNet (Fillmore & Baker, 2010) and Japanese FrameNet (Ohara, 2008) define the semantics of a word by a semantic frame, which means the structured knowledge about a typical scene, constructed with the frame elements. And, VerbNet (Kipper-Schuler, Dang, & Palmer, 2000; Kipper-Schuler, 2005) is a hierarchical verb lexicon in which each class is described by semantic predicates, thematic roles, and basic syntactic frames. Their resources and above Takeuchi (2011) are advanced semantic knowledge to deal with the complex narrative knowledge unit. In the modular approach of our narrative generation system architecture, a common and a large amount of conceptual dictionary treats the semantic knowledge which has a comparatively simple structure and other mechanisms handle more complex semantic processing. For example, Onodera and Ogata (2012) resolve an action into two states and the relations to generate sequences of states which form a basis of a story. The above resources are more directly related to the part of such complex semantic processing.

## Editing Constraints

The result of a previous evaluation and consideration shows that if we adopt “actual naturalness” as the criterion, the definition of constraints become too difficult or ambiguous (Oishi & Ogata, 2011). So, as a basic policy, we define constraints in the range of physical possibility (in our daily life in common sense). We think this criterion is comparatively clear though it may actually contain difficult problems. In addition, constraints contained in “A Japanese Lexicon” have originally the purpose and role of conflict resolution for Japanese-English translation. For the context of narrative generation, they are too wide to adequately limit the semantic range for each noun concept to be generated. For example, in the case of “eat(2)”, the subordinate concepts of a constraint “food” corresponding to the N2 contain inadequate concepts such as “drink”. And, the subordinate hierarchy of “life” also contains “the animal on a legend”. One of primary purposes of the extension of hierarchy is to avoid such ambiguity and inaccuracy.

The method is divided into two parts. First is by hyponymy concepts and “furniture” is substituted with “chair” and “bed”. Second is by the partial exclusion of hyponymy concepts and “dead person” is excluded from “human” using a minus sign such as “-dead person”. Oishi and Ogata (2011) have modified 300 constraints by the methods. For noun concepts defined as constraints of all terms in these case frames, we investigated



Table 2: Towards the expansion of deep case definition

problem	example of sentence pattern	Okada (1991)	Takeuchi (2011)	New case
1) adverbial concept	N1がN2をN3にN4縮める (N1 shorten N2 by N4 to N3)	OC: complement of attribute	numeral / correspond to adverb	Adverb
2) possessive case	N1がN2のN3にへ 潜る (N1 go under N2)			Possessive
3) event	N1がN2をN3で 破る (N1 beat N2 at/in N3)		situation	Situation
4) purpose of action	N1がN2からよりN3へ/にまでN4に 出かける (N1 go to N3 from N2 for N4)		purpose	Purpose
5) experiencer	N1がN2に 疲れる (N1 get tired of N2)		experiencer	Experiencer
6) target of comparison	N1がN2をN3の 倍にする (N1 make N2 twice as much as N3)	OS: souse		Souse
7) idiom	N1はN2(を)がN3から 遠のく (N1 go to N3 less frequently)		usage	Ideom

the subordinate concepts and edit them according to the methods. For instance, for the constraints of N3 in “N1 return to N3 from N2”, we extended the original constraints “(location place building)” to “(lodging housing area -area(scope) -area(human\_activity) -land -world foot\_of\_mountain mountain\_pass valley ground island/cape shore -bank farm site point\_of\_compass edge distance house(body)[housing])”. However, this work spent a lot of time.

To improve the work efficiently and organically, we prepared a lot of “sets” that describe a group of constraints to be used commonly in similar plural case frames. For instance, a common set can be applied to the case frames consisting of a place a person can sit down and the set can be used to the opposite concepts such as “stand up” too. Specifically, to a set of noun concepts used as constraints, we give a name that the set means as a whole to apply it to the constraints for other case frames. For example, we give a name, “inhabitable location” to a set of constraints, “(lodging housing area -area(scope) -area(human\_activity) -land -world foot\_of\_mountain mountain\_pass valley ground island/cape shore -bank farm site point\_of\_compass edge distance house(body)[housing])”. The set of noun concepts is commonly used for the constraint of N2 in other events like “N1 move to/into N2”, and so on.

In this time, we extracted 26 kinds of sets from the modified constraints in 300 case frames and applied the sets to other 5037 case frames. As a result, we could apply the sets to one or more elements in 4581 case frames. The number of cases that are adequate sets to all elements was 1158. For example, in “N1 take out N2 from N3”, we could apply a set, “a person of such age who can take care by herself or himself” to N1 and “transportable object” to N2.

## A Noun Concept Hierarchy

Figure 4 is the overall structure of noun conceptual hierarchy. The current version contains 5808 intermediate concepts and 115765 terminal concepts. Many of the higher level concepts are based on the information registered in “general semantic attributes system (hierarchy) for nouns” in “A Japanese Lexicon”. On the other hand, the hierarchy of lower intermediate concepts is basically organized referring to “Japanese WordNet”. The terminal concepts are based on the words in the Japanese lexicon.

Next Figure 5 is the format for describing the noun concept hierarchy and the actual example by Common Lisp. The upper list means the serial number showing registered intermediate concepts, and variables into which the information for each intermediate concept is stored. An intermediate concept has a list of hyponymy concepts, a number of the depth in hierarchy, the serial number of the superordinate concepts, and the range of serial numbers of the hyponymy concepts. In the example,

“(seal sea-lion fur-seal sea-animal dugong bakushia)” is defined as the subordinate concepts of “sea-animal”.

The method of constructing the noun conceptual hierarchy is as follows. The referred Japanese lexicon contains 141870 words in 2710 attributes in “general semantic attributes system (hierarchy) for nouns”. Here, each attribute is corresponding to a concept in our conceptual dictionary. These words have homonyms like “狸” (“tanuki” in Kanji) and “タヌキ” (“tanuki” in Katakana), which means commonly “raccoon dog” in Japanese. In the narrative generation system, we divide the generation process into a conceptual level to make a narrative content and a surface representation level to make the variations by language and other representation media. The former uses a conceptual dictionary and the latter uses a language dictionary. To realize the mechanism, after we integrated two or more homonymous words into a single concept, we viewed all words as concepts to store 115765 concepts into the noun concept hierarchy.

In addition, to be able to set more detailed and elaborated constraints, we added intermediate concepts and refined the classification referring to “Japanese WordNet”. In the step that we integrated homonymous words into a single concept to register them as concepts, many terminal concepts were directly combined with the intermediate concepts. For example, “seal” and “raccoon-dog” are respectively the subordinate concepts of “beast”, and more detailed categories do not exist. If we want to define “aquatic mammal” as a constraint, we have to list all the terminal concepts. On the other hand, in “Japanese WordNet”, the concept of “aquatic mammal” exists as a subordinate category of “aquatic mammal”. Therefore, we added an

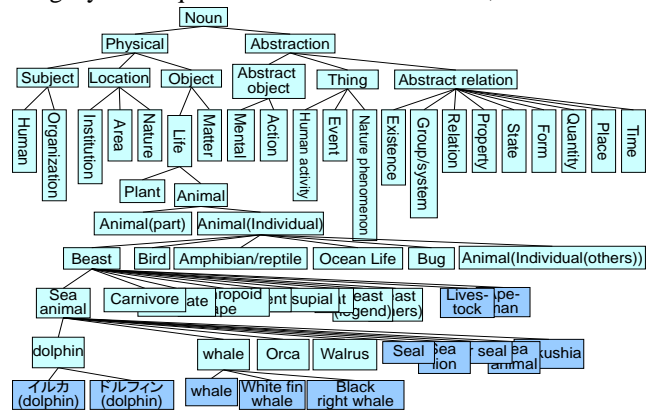


Figure 4: The overall structure of the noun concept hierarchy

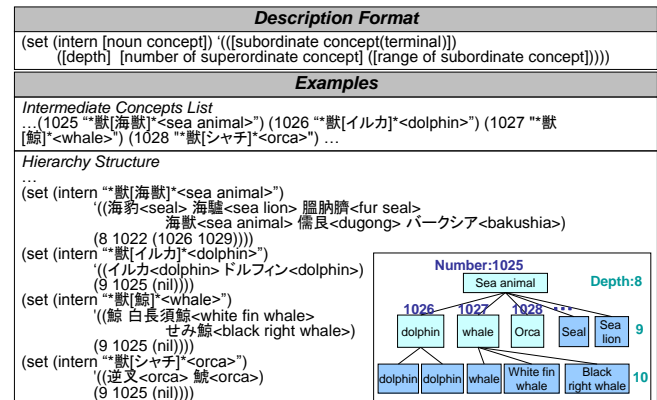


Figure 5: The format and an example of noun concept hierarchy

intermediate concept, “sea-animal”, under “beast”, and moved “seal” and “sea-lion” to the subordinate category.

### A Tentative Evaluation

For the concepts that constraint sets are applied, we attempted a confirmation of the validity of the constraints and the classification of noun concepts regarding to the constraints. The criterion is whether physically possible events can be generated by each constraint. For example, we determine that such event as “A salesgirl divorces her husband” is possible. In contrast, such event as “A spokesman is sailing on the paper” is impossible.

We prepared an experimental event generation program which selects a noun concept in the noun concept hierarchy for each case in the case frame at random according to the constraint. And, we generated 20 events by 100 case frames selected at random from 1158 case frames that all constraints were set. Three evaluators checked the generated 2000 events according to the criterion of possible/impossible. As a result, if 14/20 or more events were “physically possible”, we decided that the constraints in the case frame were comparatively adequate. When the results in the evaluators were different, we employed the result of majority. The success was 70 and the failure was 30. For 12 in the 30 failures, we could modify them by applying another set of constraint. For example, in the case of “N1 collect N2 from N3”, we may replace the set of “N2:object” from “goods, status, and notion to be able to send” to “goods to be able to send”. For the 15, we modified the constraints themselves. For example, in the case of “N1 sail N2”, we changed the constraint of “N2:location” to “(river waterway lake sea)” from “(river waterway lake sea sky)” which is corresponding to a set, “the place to be able to pass with ship and aircraft”. Other two ones are the description errors in case frames. Remaining 1 problem is the case that the verb concept itself is impossible like “N1 cast a spell on N2”. By applying the above criterion of evaluation, the difference of feelings by evaluators and the ambiguity of evaluation diminished more than previous criterion, natural/unnatural. This brings the simplicity of system development. However, there is also the difficulty that possible but ordinarily unnatural events can not be eliminated. Because it is difficult that the conceptual dictionary absorbs this sort of contextual knowledge, the solving needs to be given in the relation with the above-mentioned defamiliarization rhetoric. We are considering a basic framework that the conceptual dictionary gives comparatively simple and standard conceptual knowledge and a variety of narrative knowledge such as the defamiliarization operates and adjust more complex and advanced literary or artistic rhetorical techniques.

### Conclusions

We explained the current version of a conceptual dictionary containing two hierarchies of verb concepts and noun concepts to be functioned in our narrative generation system. It is used for operating naturalness or validity of generated events and realizing or adjusting the intentional defamiliarization. Namely, this dictionary is a mechanism to be able to flexibly adjust a variety of generation from realistic narratives to fantastical narratives as well as the foundation for a narrative event and the elements. In the current version, verb concept dictionary has originally defined 5338 case frames and modified 1158 constraints and noun concept dictionary contains 121573 concepts including 5808 intermediate concepts.

### References

- Akimoto, T., & Ogata, T. (2011a). Computational model of narrative discourse theory and reception theory in narratology and its implementation. *Proc. of the 13<sup>th</sup> Japanese Society for Language Sciences* (pp. 155-156).
- Akimoto, T., & Ogata, T. (2011b). A consideration of the elements for narrative generation and a trial of integrated narrative generation system. *Proc. of the 7<sup>th</sup> NLPKE* (pp. 369-377).
- Bates, J. (1991). Virtual reality, art, and entertainment. *The Journal of Teleoperators and Virtual Environments*, 1(1), 133-138.
- Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., & Kanzaki, K. (2009). *Proc. of the 7<sup>th</sup> Workshop on Asian Language Resources, ACL-IJCNLP* (pp. 1-8).
- Bringsjord, S. & Ferrucci, D. A. (2000). *Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS, a Storytelling Machine*. New Jersey: Lawrence Erlbaum.
- Genette, G. (1972). *Discours du récit. essai de méthode, Figures III*, Seuil: Paris. (Transl. Lewin, J. E. (1980). *Narrative discourse: an essay in method*. NY: Cornell University Press.)
- Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Ooyama, Y., & Hayashi, Y. (1999). *Goi-Taikai-a Japanese Lexicon CDROM*. Tokyo: Iwanami Shoten. (in Japanese)
- Imabuchi, S., & Ogata, T. (2011). A story generation system based on Propp combined with a conceptual dictionary. *Proc. of the 7<sup>th</sup> NLPKE* (pp. 359-362).
- McIntyre, N. & Lapata, M. (2009). Learning to tell tales: A data-driven approach to story generation. *Proc. of the 47<sup>th</sup> Annual Meeting of the ACL and the 4<sup>th</sup> IJCNLP of the AFNLP* (pp. 217-225).
- National Institute of Information and Communications Technology (2001). *EDR Electronic dictionary technical guide*. [http://www2.nict.go.jp/r/r312/EDR/J\\_index.html](http://www2.nict.go.jp/r/r312/EDR/J_index.html) (Last access: 1.11.2012).
- Ogata, T. (2008). Towards the movie construction in narrative generation system. In A. Kanai & Y. Niwa (Eds.), *Theory and practice of film editing*. Tokyo: Hosei University Press.
- Ogata, T., Akimoto, T., & Seito, A. (2011). A new version of circular mutual transformation system between music and narrative. *Proc. of the 25<sup>th</sup> JSAI (1H2-OS1-13in)*. (in Japanese)
- Ogata, T., & Kanai, A. (2010). *An introduction of informatics of narratology: on thought and technology of narrative generation*. Tokyo: Gakubunsha. (in Japanese)
- Oishi, K., & Ogata, T. (2011). Towards the development of conceptual dictionary for narrative generation system. *Proc. of the 7<sup>th</sup> NLPKE* (pp. 351-358).
- Okada, N. (1991). *Representation and strage of word concepts*. Tokyo: The Institute of Electronics, Information and Communication Engineers. (in Japanese)
- Okada, N., & Endo, T. (1992). Story generation based on dynamics of the mind. *Computational Intelligence*, 8(1), 123-160.
- Onodera, K. & Ogata, T. (2012). Sequence generation based on mutual relationship between state and action: As a mechanism in narrative generation system. *Proc. of the 4<sup>th</sup> DIGTEL* (pp. 159-161).
- Propp, V. (Попов, В. Я.) (1969). *Морфологиясказки*, Изд. 2е. Москва: Наука. (Transl. Scott, L. (1968). *Morphology of the Folktale*. Austin, TX: University of Texas Press.)
- Takeuchi, K. (2011). Construction of thesaurus of predicate-argument structure for Japanese verbs. *Proc. of the 25<sup>th</sup> JSAI (3H2-OS3-5)*. (in Japanese)
- Zhang, Y., Ono, J., & Ogata, T. (2011). An advertising rhetorical mechanism for single event combined with conceptual dictionary in narrative generation system. *Proc. of the 7<sup>th</sup> NLPKE* (pp. 340-343).
- Zhang, Y., Ono, J., & Ogata, T. (2012). Single event and scenario generation based on advertising rhetorical techniques using the conceptual dictionary in narrative generation system. *Proc. of the 4<sup>th</sup> DIGTEL* (pp. 162-164).



# Working memory's meager involvement in sentence repetition tests

Eve Okura (eveokura@hawaii.edu)

Department of Linguistics, 558 Moore Hall,  
Honolulu, Hawaii 96822 USA

Deryle Lonsdale (lonz@byu.edu)

Department of Linguistics & English Language, 4039 JFSB  
Provo, UT 84602 USA

## Abstract

Elicited imitation (EI) is a testing method for learners' oral language proficiency. One common criticism aimed at EI is that performance might not require linguistic knowledge, but mere rote memorization. This study explores the issue by administering two tests to the same group of students studying English as a second language: (1) a working memory test, and (2) an English EI test. Participants came from a range of English language proficiency levels. Our goal was to test whether scores from these two treatments (English EI scores and working memory scores) would correlate significantly. If not, this would suggest that there is some difference in what they measure. The results did fail to show a significant correlation between working memory and English EI scores. On the other hand, there was a significantly positive correlation between students' English EI scores and their placement level.

**Keywords:** working memory; elicited imitation; oral language testing; explicit/implicit linguistic knowledge

## Introduction

Numerous methods exist for assessing language competence, ranging from written grammar examinations to live interviews with a native speaker. Some discussion revolves around the different types of tests: what they actually measure, and whether they directly assess linguistic knowledge and ability versus some other capability. One such testing method is commonly referred to as *elicited imitation* (EI). EI is a testing method in which researchers present a series of sentences to a person who then imitates each sentence by repeating it as accurately as possible. These responses are recorded and subsequently analyzed. Since 1967 EI has been used in linguistic research, particularly for studying language acquisition (Slobin & Welsh, 1973). Many have further proposed that EI tests can assist in measuring language ability (Connell & Myles-Zitzer, 1982; Vinther, 2002; Erlam, 2006). Others have argued (Hamayan, Saegert, & Larudee, 1977) that EI cannot measure language ability, and that listeners can simply repeat what they hear, without involving any linguistic processing. Such arguments claim that a person could repeat the utterance without understanding what it means, and without having any ability in that language. In this paper we report on a study meant to contribute more human performance data to this debate.

Addressing the issue of whether or not EI test-takers rely solely on rote repetition in responses requires some consideration of memory. We assume here an admittedly simple but traditional tripartite view of memory: long-term, short-term, and working memory (Cowan, 1996). Furthermore,

we will focus on only the third component, working memory (WM). WM contains the information held momentarily, as it is needed to analyze, solve a problem, or perform a task.

Working memory is the part of memory that would determine whether high performance levels on EI tests are a result of mere parroting or whether linguistic knowledge and competency are also needed. An EI test-taker who did not know English would have to hear the stimulus utterance, retain it in WM, and then repeat the utterance exactly to get a perfect score on the item. If the item were beyond their WM capacity, they would be expected to get only part of it correct at best, all of it wrong at worst.

Clearly WM and second-language (L2) ability are not completely independent. Some WM capacity is necessary to even be able to respond in an EI test (Doughty & Long, 2003; Robinson, 2005); some WM capacity is also necessary in all analytical and linguistic tasks. In this paper we probe the degree of overlap between them with two EI tests, one targeting English ability and the other WM capacity.

## Background: WM, language, and EI

Early research in working memory includes Miller's famous discussion of *immediate memory span* positing a "magic number" of  $7 \pm 2$  (Miller, 1956), meaning that a person could generally hold up to seven completely unrelated items in mind simultaneously. Several researchers have subsequently advocated the number  $4 \pm 1$  instead as the relevant metric defining the scope of average WM capacity (Cowan, 2001). Miller also introduced the process of *recoding* now more commonly referred to as *chunking*. Chunking takes multiple separate items and agglomerates them into patterns, reducing the number of items to remember and speeding up processing.

Another thread of research has sought to differentiate WM, which is highly evanescent, from other types of memory that are longer-lasting. For example, Baddeley and Hitch (1986) experimented to see if storing several numbers in working memory affected the ability to carry out comprehension tasks. It did not, suggesting that memory is stored in a different "place" from where processing and problem-solving occur, as one did not impede the other. A distinction is typically drawn between WM and short-term memory (STM): the latter is a form of memory that lasts for up to 10-30 seconds after one receives a stimulus (Cowan, 1996). Discussions of how language interacts with WM and STM also involve the articulatory loop, a time-constrained buffer that temporarily

stores information when comprehending and producing language. Issues of activation, decay, and interference all complicate and enrich investigation of these areas; to the extent possible we will abstract away from or simplify these factors in the work reported in this paper.

STM involvement in conversational language includes managing spontaneity, multiple interlocutors, conversational turn-taking, contextual factors, meta-awareness, and explicit language knowledge. Evaluating L2 conversational speech abilities typically involves oral proficiency interviews (OPI's) (Lehman & Tompkins, 1998) which are costly and challenging to grade since it involves skilled human interlocutors.

Our study differs from natural speech in that spontaneity does not occur, as it would in a traditional OPI. In an EI test, as opposed to natural conversation, there is a response, but it is imitative rather than a spontaneous utterance invented by the individual. STM is what holds the stimulus information when the response is being planned (Cowan, 1996).

Language acquisition researchers distinguish between *acquisition* and *learning*. Acquisition is done intuitively, similar to a child developing her native language abilities (Krashen, 1982). Competence, the result of acquisition, is also subconscious: we are generally not consciously aware of the grammatical rules of the languages we have acquired. Instead, we have a "feel" for correctness, and errors do not "sound" or "feel" right, even if we do not consciously know what rule was violated. Krashen associates language acquisition with *implicit learning*.

On the other hand, Krashen defines *language learning* as a conscious awareness of grammar, vocabulary, syntax, etc., and refers to this as *explicit learning* (Krashen, 1982). Ellis echoes Krashen's use of the terms implicit and explicit and extends them beyond the learning process to linguistic knowledge itself (N. C. Ellis, 2008). Thus explicit and implicit knowledge of language are distinct and dissociated, they involve different types of representation, they are substantiated in separate parts of the brain, and yet they can come into mutual influence in processing. Explicit knowledge of a language is form-focused rather than meaning-focused, and involves conscious thought. Presumably much of the implicit and explicit knowledge of a language are stored away in long-term memory (LTM) and marshaled as necessary by STM when required.

EI testing is interesting for a variety of reasons. First, it is time-constrained; the responses must be immediate, so extensive deliberation is not possible. The time constraint also prevents use of the articulatory loop to enhance WM, and thus scores on the EI test. Second, the test involves repetition instead of full-fledged conversational interaction so that deliberation about of grammatical form is less salient. These two elements—time limit and avoiding meta-awareness of form—are thought to shift EI tests towards the assessment of implicit language knowledge instead of explicit knowledge (R. Ellis, 2005).

Some combination of WM, STM, and LTM is pressed into

service for language-based interactions including simultaneous listening and comprehension. STM, like WM, is limited in its capacity and reactivity, so processing language must involve optimally combining layers of representation of meaning (morphemes, lexemes, lexical items, phrases, sentences), that allow STM to group items together, thereby expanding its ability to deal with the situation at hand.

Meaning is another important consideration in memory, language, and learning (Doughty & Long, 2003). Robinson (2005) noted that phonological WM capacity in particular is associated with L2 speaking abilities. DeKeyser notes the difference between *meaningful* versus *nonmeaningful* associations, each in its relation to memory capacity: meaningful form-function mappings associate constituents in a sentence according to lexical and grammatical principles of the language. Establishing meaningful relationships between abstract entities draws more on insight, whereas associating nonmeaningful co-occurrence of concrete elements logically draws more on memory (DeKeyser, 2005).

Utterances are constructed as intonation units, and substantive units are fairly strongly constrained to have a typical length of about four words in English, indicative of the cognitive limits on how much information can be fully active in the mind at any one time (N. C. Ellis, 2002). At the syllable level most information has already been chunked up, reflecting the phonemic, morphological, and lexical properties of the language: speakers rarely deliberate at the syllable level in languages they are proficient in. On this basis, one would expect a nonce syllable test would measure memory, and a "meaningful" (i.e. language-based) test would draw less on memory than the nonce test, and more on linguistic ability.

Previous research has also shown that knowledge of semantics, grammar, syntax, etc., affects verbal STM performance. This suggests that such linguistic working memory chunking also occurs during the process of verbal repetition. The articulatory loop is also implicated in both vocal and subvocal repetition of a phrase, which helps to maintain it in short-term memory. The connection is complicated by factors including articulation rate, semantic properties, grammar class, word frequency or familiarity, and word sequencing. Since chunking can happen at these linguistic levels as well, an interesting question is the interaction of linguistic versus articulatory loop chunking; current thinking is that the former can supersede the latter (Morra, 2000).

It follows that, due to WM limitations, correctly repeating longer utterances in an EI test is only possible via linguistic recoding. Hence WM tasks involving processing of nonce syllables will be different from WM tasks enhanced by language ability, and the difference indicates to what degree the EI tests are measuring language.

Elicited imitation (EI) and sentence repetition testing (SRT) are the most common terms used to refer to the same testing method. Though a thorough review of the extensive history in EI testing is not possible here, we mention some of the strands of work most relevant to our current experiment.

EI is the “...repetition of a model sentence presented in a context calling for imitation, as opposed to...spontaneous imitation of...utterances” (Slobin & Welsh, 1973). Bley-Vroman and Chaudron (1994) suggest that EI performance requires both linguistic processing and short-term memory. They sketch a 4-step process of EI performance; the test-taker: (1) hears the utterance; (2) forms their version of it in their mind (a *representation*); (3) stores this representation in short-term memory; and then (4) verbalizes their representation of the initial utterance.

EI has been instrumental in measuring first language acquisition. By studying the EI of adult speech in children, Slobin and Welsh (1973) conclude that the process requires linguistic processing, particularly since children cannot repeat beyond what they could produce spontaneously, unless the utterances were short enough to leverage immediate memory. He argues that the same conclusions hold for adult L2 learners.

A well known basic asymmetry in language capabilities exists between comprehension and production; this also underlies EI testing. Vinther points out whereas some test subjects may be able to understand the stimulus but not reproduce it accurately, on the other hand “the opposite situation, that subjects should be able to produce well-formed imitations of sentences they have not understood and are not able to remember, is highly improbable” (Vinther, 2002).

Prior literature also provides guidelines for EI testing. Jessop et al. outlined the advantages and challenges of using EI as an L2 acquisition assessment, providing suggestions for using EI effectively. This paper addresses the central problem they delineate: “In the 1970s, EI’s validity was challenged: the major criticism being the possibility of rote repetition in response to stimuli (i.e., participants may be simply parroting what they hear)” (Jessop, Suzuki, & Tomita, 2007). They call for more study to validate the use of EI in measuring L2 performance.

EI tests thus measure verbal STM performance in the sense that they measure knowledge of the various linguistic elements discussed above, rather than rote memory ability. It is also apparent that WM is essential to perform well on a language EI test. The question is how much do language EI tests overlap with working memory tests in what they measure? If the two do not correlate perfectly, then whatever does not overlap implies that the remaining portion, the part of the EI test that does not overlap with WM, is measuring something else. As suggested by previous research, that “something else” is knowledge of grammar, vocabulary, syntax, and other linguistic features. To quote one researcher, “Elicited imitation goes beyond rote memory and repetition; rather, sentences are assumed to be ‘filtered’ through one’s grammatical system” (Gass & Mackey, 2007). This would predict that highly proficient L2 learners will do better at EI tests since their knowledge and expertise will help overcome memory capacity and time constraints.

Conversely, an L2 EI test would function more similarly to a WM test for individuals who do not yet have any ability

in the language. This is because without necessary language competence, the individual will not be able to “chunk” and use other opportunistic linguistic strategies; syllables would instead sound like nonce syllables to the student. For longer sentences mere rote repetition will not be possible since WM capacity will be exceeded.

If there were a significant WM/EI correlation across the board, one could conclude that memory is dominating the measurement of language ability. This would support arguments that EI tests are mere “parroting”, and hence offer no contribution to measuring language ability. If, however, there is minimal correlation between WM scores and EI language test scores, the tests cannot be primarily memory tests. We also predict that some students who have approximately the same WM capacity but differing language ability will score differently on the EI language test. In summary, we claim that EI language tests, when constructed properly, primarily measure linguistic ability, and demonstrably do not measure an individual’s working memory to a significant degree.

## Testing EI and WM

We administered two tests to a pool of students: an English EI test and a WM test. For the EI test, we used our own pre-existing English instrument: EI Form D. It contains 60 questions of varying lengths and complexity, carefully designed and engineered to target different levels of proficiency. Half of the items were also paired with a comprehension task.

The comprehension task involved two illustrations appearing side-by-side on the screen after an EI stimulus sentence was given, but before the test-taker recorded their response. One illustration depicted the meaning of the stimulus sentence, and the other picture was irrelevant to the stimulus. Test-takers were to click on the relevant picture. Only then were they allowed to record their audio response. These tasks were intended to divert attention away from form, and shifting the focus to meaning.

We chose to create our own working memory test—as opposed to using a previously existing one—in order to assure that its content, format, and delivery method integrated seamlessly with the English EI test. Our WM test uses nonce syllables (Cowan, 2005) drawn from a pre-existing repository of nonce words (McGhee, 2010). Meaningless monosyllabic words such as “kish” were used directly, and longer ones were split into nonce syllables. Caution was exercised to ensure that the individual syllables were not homophones of actual English words, either in isolation or in sequence; this would have enabled chunking and hence complicated WM measure interpretation. The 55 nonce syllables were arranged into ten items of increasing length: item 1 has one syllable, and item 10 has 10 syllables.<sup>1</sup>

Half (i.e. five) randomly selected WM items were recorded

---

<sup>1</sup>A reviewer has observed—and we concur—that a more commensurate WM test would have had sentences with (nonce) syllable totals equalling those in the EI test sentences. Unfortunately we no longer have access to the subject population for another within-subjects test.

by a male voice, and the other by a female voice, presented in randomized order. Syllables within a single item were separated by pauses of approximately one second in length. Care was also taken to assure flat prosody when pronouncing the syllables, avoiding any intonational contours that might make them easier to remember in sequence.

We administered the English EI test in January 2011 to incoming adult students at the BYU English Language Center (ELC) as a part of an initial placement assessment. The ELC in essence has 6 grade levels, Levels 0 through 5. Students are placed at these levels based on multiple assessments of their abilities, including grammar, writing, and listening comprehension tasks. Our test was given at the same time as other diagnostic language tests. Our EI test results were not used in placing students in their ELC levels.

The 94 students were randomly divided into two equally sized groups (A and B). Group A took the 60-item English EI test with 30 items paired with distractor comprehension tasks. Group B was given the same 60-item English test, but with the other half of the 60 items paired with comprehension tasks.

The presentation order of English EI items was randomized within each set (questions with tasks, and questions without tasks, respectively) to eliminate any possible item-order effects. In the test both groups were presented the section without the tasks first, and afterwards the section with the tasks. Time allowed for responses for individual items was proportional to their length. Due to technical difficulties (microphone malfunctions, web server problems, etc.) about half of the students' tests had to be discarded.

A few months later, in March and April, as many of the original students as could be located were invited to take the 10-item working memory test. This resulted in 67 students; some students had left the program, and others were unavailable due to scheduling problems. Again, various technical difficulties resulted in tests that were lost or unusable. The final pool of data consisted of 40 students for whom we had a complete set of English EI and WM scores.

Both tests were scored by a percentage method, i.e. the overall score on the test is a percent of the total number of syllables the student uttered correctly over the total number of syllables in the stimuli of the test. As with our prior work, we had trained human graders (all native English speakers for this project) score the test recordings via a web-deployed user-friendly interface. The rater listens to the audio recording, and marks each syllable that is present with a "1," and each syllable in the utterance that the test-taker missed with a "0." Six individuals graded results for the 5,430 English test items (89 test-takers times 60 items per test), and the 890 working memory test items. Individual EI item scores were aggregated across each test.

## Results and Discussion

To ensure that the WM test was actually testing memory, we analyzed the scores of all participants for each item. All WM

scores ranged from 27.27% to 52.73% (mean and mode at about 40%), with one outlier at 20%. As indicated in Figure 1, average accuracy across students on 1, 2, and 3-syllable items were almost 100% as expected. At 4 syllables, the average score dropped to just below 80%. At 5 syllables the score fell below 50% (49.55%), reminiscent of the  $4 \pm 1$  WM "magic number". These findings suggest that our WM test was measuring WM capacity as intended. In addition, the boxplot shows that skewness on both tests was minimal.

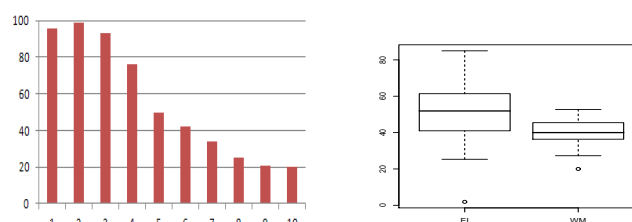


Figure 1: WM scores histogram (left) and boxplots for both tests' scores (right).

We then ran three Pearson correlations across individuals to quantify relationships of interest; see Table 1.

Test	EI	ELC Level
WM	$r=.249, n=40, p=.121$	$r=.130, n=40, p=.430$
EI		$r=.786, n=40, p=.000$

Table 1: Pearson correlations between WM and EI scores and placement level.

The correlations between WM and EI scores and between WM scores and ELC level do not reach significance, though the correlation between EI scores and ELC level does.

Consider the scatterplot graphs in Figure 2: each plot-point represents an individual student who took both tests, modulo overlap. Graph (a) shows that most students have average WM scores (around 40% overall), but have very divergent EI test scores, ranging vertically throughout almost the entire spectrum of English language ability, from 1.9% to 84.97%. We also see students with below-average WM scores but average English EI scores; on the other hand, some students who did not do as well on the EI test have above-average WM scores.

Scatterplot (b) shows that the student at the highest ELC level (5) had a WM score of 41.82%, but was outperformed on the WM test by a student at the lowest ELC grade (0), who had a WM score of 47.27%.

Scatterplot (c) reveals a much clearer correlation, the only significant positive correlation we found: EI versus ELC placement level. Most of the points align much more closely with the diagonal.

Of the three students in Level 0 (the ELC's lowest beginning level), one of them scored the very lowest on the English EI test, correctly repeating only 1.9% of the syllables.

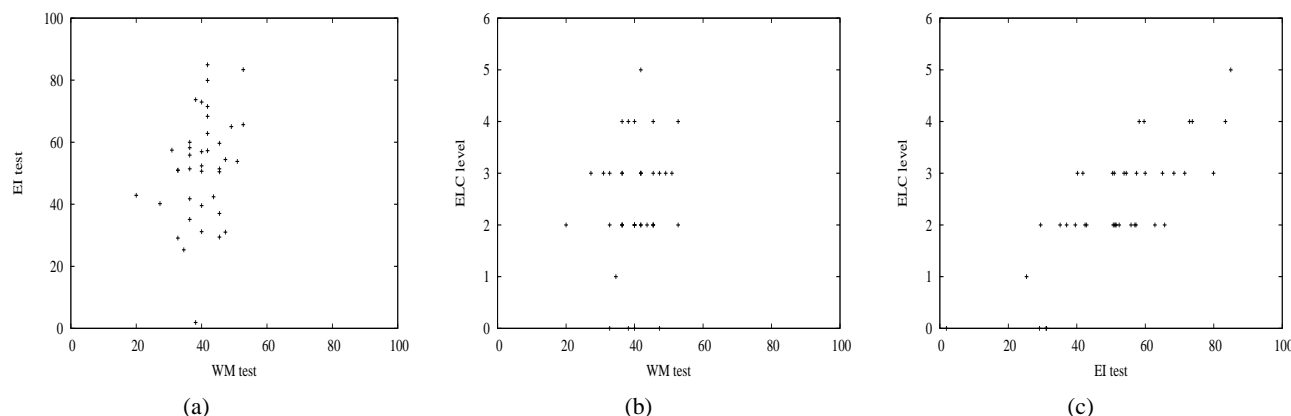


Figure 2: Scatterplots: (a) WM score vs. EI score; (b) WM score vs. ELC level; and (c) EI score vs. ELC level

This student's working memory score was perfectly fine, at 38.18%. Further, note that the lowest beginning level students could not break past the 30-40% barrier on the English EI test.

With their little-to-no knowledge of English, the lowest-scoring students were unable to repeat anything beyond their WM capacity. It seems reasonable to assume that below scores of roughly 30%, working memory tends to be the primary contributor to the test-takers' English EI results. Above 30%, knowledge of the language seems to have a greater influence, and working memory proportionately less so. This particular result follows our prediction—that the English EI test would primarily be an assessment of an individual's linguistic ability, and not an assessment of their working memory capacity—not as a rote memory test.

As mentioned above, most types of second language capacity tests assess explicit knowledge. Our test responds to Ellis' two criteria for measuring implicit knowledge: (1) the test must be time-constrained, and (2) the test must get at meaning, and, as much as possible, avoid focusing on form (R. Ellis, 2005). Our test was time-constrained: test-takers were only allowed a few seconds to respond. This did not give students a chance to analyze sentence structure or grammar, or indeed rehearse the item before repeating it.

Although we were not measuring actual language chunking in this test, if chunking by meaning enabled higher-level ELC students to perform better on the English EI test, then it may be that this test (or at least some of its items) allowed them to focus on meaning rather than form.

Shorter EI stimuli that fall within the constraints of working memory function more like a nonce syllable working memory test. This is especially true of students at Level 0, who have little knowledge of the language. But lengthier sentences are too long for WM capacity, and perhaps form-focused chunking does not sufficiently recode so many syllables into few enough chunks to retain in working memory. The longest sentences on the test can be remembered and uttered correctly more easily by those who can recode it into its

meaning, which accesses their implicit knowledge of the language. Otherwise these lengthier items become impossible for low-proficient learners to reproduce based on WM alone.

For example, one of the items on our English EI test, "Joe writes poetry," is short enough that working memory may have a greater correlation to scores on this particular item than the test overall. While due to its length it could be considered a working memory test item, its carefully designed grammatical features make it a test of at least some explicit linguistic knowledge.

Complex, lengthy items such as "When Jim entered the office he was immediately afraid of the uncommunicative boss." contain too many separate features for a person wholly dependent on explicit knowledge to maintain all the separate elements in STM and/or WM; a round-trip to implicit knowledge representations would be required. If such is the case, then specific items could be used to distinguish between working memory, explicit knowledge, and implicit knowledge of the language.

In summary, the lack of significant correlations between working memory and English EI scores and between working memory and ELC levels, and the significant correlation between English EI scores and ELC levels suggest that there is more to performance on EI tests than working memory capacity. Ellis proposed that EI tests could be developed to test for different types of knowledge, namely, explicit linguistic knowledge and implicit linguistic knowledge. He stated that time-limitations and a focus on meaning were necessary to construct an EI test that measured implicit linguistic knowledge. The test-takers' errors indicate that using items with specific grammatical features can distinguish between levels of ability. Beginning level test-takers had difficulty with even very short items, if they contained specific grammatical structures beyond their linguistic ability. Long sentences were also difficult or impossible for the beginning-level students, but doable for advanced speakers. These findings produce further research questions. Do utterances that are too long to repeat using average working memory capacity require chunking by

meaning to repeat correctly? If EI performance on longer sentences does access a translation into meaning, then according to Ellis' framework, these types of long utterances also test implicit knowledge. A further question still is, do test-takers even recall hearing those morphemes they were unable to accurately repeat, such as the word-final 3rd person singular "s" (e.g. "runs")?

### Conclusions and future work

The findings above indicate a lack of significant correlation between EI and WM scores for the same population. Though a direct causal relationship cannot be made to language use from this result, our measures do lend circumstantial support to the idea that WM testing does not primarily target linguistic ability, and hence directly or even significantly influence EI scores. This is in agreement with the findings of DeKeyser (2005) that elements without meaning (in this case, nonce syllables) draw on memory more than elements of meaning. Though some working memory is involved in language learning and production (Robinson, 2005), it was not shown here to be positively correlated with EI testing. This lends credence to the suggestion that the English EI test is testing linguistic knowledge.

There were a few limitations in the scope of this study, all of which can be overcome with further testing or analysis: (1) The sample size of forty test-takers is relatively small, though the largest we have seen yet for the task at hand. (2) We cannot guarantee that the nonce syllable sequences are meaningless in all of the non-English languages known by test-takers. (3) It would be informative to correlate other metrics (e.g. the participants' OPI scores) to working memory. (4) We have not yet analyzed the effect of the distractor tasks. (5) More detailed analyses of our results could be undertaken including such techniques as structured equation models or ANOVA analyses.

### Acknowledgments

We express our appreciation for support from the BYU Pedagogical Software and Speech Technology Research Group, the BYU Center for Language Studies, and the BYU English Language Center.

### References

- Baddeley, A. D., & Hitch, G. (1986). *Working memory*. Oxford: Oxford University Press.
- Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second language competence. *Research methodology in second-language acquisition*, 245-261.
- Connell, P. J., & Myles-Zitzer, C. (1982). An analysis of elicited imitation as a language evaluation procedure. *Journal of Speech & Hearing Disorders*(47), 390-396.
- Cowan, N. (1996). Short-term memory, working memory, and their importance in language processing. *Topics in Language Disorders*(17), 1-18.
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*(24), 87-185.
- Cowan, N. (2005). *Working memory capacity*. Hove, East Sussex, UK: Psychology Press.
- DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? a review of issues. *Language Learning*(55), 1-25.
- Doughty, C. J., & Long, M. H. (2003). Optimal psycholinguistic environments for distance foreign language learning. *Language Learning & Technology*(7), 50-80.
- Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*(24), 143-188.
- Ellis, N. C. (2008). The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. *The Modern Language Journal*(92), 232-249.
- Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*(27), 141-172.
- Erlam, R. (2006). Elicited imitation as a measure of 12 implicit knowledge: An empirical validation study. *Applied Linguistics*(27), 464-491.
- Gass, S. M., & Mackey, A. (2007). *Data elicitation for second and foreign language research*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Hamayan, E., Saegert, J., & Larudee, P. (1977). Elicited imitation in second language learners. *Language and Speech*(20), 86-97.
- Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *The Canadian Modern Language Review*(64), 215-238.
- Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon Press Inc.
- Lehman, M. T., & Tompkins, C. A. (1998). Reliability and validity of an auditory working memory measure: data from elderly and right-hemisphere damaged adults. *Aphasiology*(12), 771-785.
- McGhee, J. (2010). *Lexical decision reaction times: The effects of ambiguity and bilingualism*. (Unpublished)
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*(63), 343-355.
- Morra, S. (2000). A new model of verbal short-term memory. *Journal of Experimental Child Psychology*(75), 191-227.
- Robinson, P. (2005). Aptitude and second language acquisition. *Annual Review of Applied Linguistics*(25), 46-73.
- Slobin, D., & Welsh, C. A. (1973). Elicited imitation as a research tool in developmental psycholinguistics. In C. Ferguson & D. Slobin (Eds.), *Studies of child language development* (p. 485-489). New York: Holt, Rinehart and Winston Inc.
- Vinther, T. (2002). Elicited imitation: a brief overview. *International Journal of Applied Linguistics*(12), 54-73.

# Changing Discriminatory Norms Using Models of Conceptually-Mediated Cognition and Cultural Worldviews

Daniel Olsher (dan@intmind.com)

National University of Singapore

## Abstract

Cognitive modeling can provide immense social benefits, especially in the case of unconscious processes causing significant psychological and societal distress. One such process, discrimination against minorities, is typically grounded in unconsciously held stereotypes reflecting deep ignorance of the realities minorities face. Another, suicide terrorism, results in significant suffering for many communities. Related beliefs often arise from unquestionable values, norms and worldviews, however, making direct/conscious change appeals unworkable. Building upon cultural knowledge representation and cognitive modeling, this paper shows how change could be effected via cognitive operations on conceptual worldviews. After introducing a novel framework for modeling conceptually-mediated belief systems and techniques for guided dissonance reduction and network change, the paper applies these to anti-discrimination and terrorism reduction. Such work holds great potential for better aligning perception of minorities with the realities they face, reducing suffering, and disrupting psychological processes dependent on improper views of stigmatized minorities and other phenomena.

**Keywords:** Culture; Cognitive Models; Nuance; Knowledge Representation; Norm Change; Discrimination Reduction; Terrorism Reduction

## Introduction

Cognitive Science has the potential to provide immense social benefits, especially in shedding light on processes that are normally unconscious but which tend to cause significant psychological and societal distress.

One such process is discrimination, often grounded in stereotypes reflecting deep ignorance of the realities of what minorities are and what they face. Another is the process by which people decide to undertake violent action during conflict.

Because the beliefs underlying such processes often arise from norms, cultures, values, and other strongly unconscious, often nearly unquestionable, belief systems, direct conscious appeals are often unworkable and insufficient for effecting change.

In this paper, we consider how to construct procedures capable of achieving change in beliefs by modeling the beliefs people already hold (and the connections between them) and designing procedures tailor-made to effect change.

Incremental changes are most effective. The use of familiar concepts lowers fear, and as the status quo already appears to be 'reality', small changes to it are more likely to succeed. Direct/conscious attempts to change such beliefs often fall afoul of cognitive and social defense mechanisms, with excessive change causing dissonance, negative emotional outcomes, and ultimately, rejection.

This paper introduces a framework for modeling conceptually-mediated belief systems, a novel change

strategy entitled Directed Dissonance Reduction (DDR), a conceptually-mediated formulation of inoculation theory, and a brief discussion of techniques for modifying belief networks. Examples (including sample networks) are then given demonstrating how to apply these tools in the prejudice and terrorism reduction domains.

Beyond norm change, this research seeks to advance nuanced knowledge representation of worldviews, cultures, and norms as well as to advance our understanding of conceptually-mediated cognition.

## Core Representation Frameworks:

### COGVIEW and INTELNET

In this paper, the author's COGVIEW and INTELNET formalisms are used to model nuanced conceptual worldviews. Both formalisms are the subject of ongoing development, with this paper providing an early demonstration of their combined capabilities.

COGVIEW is a network-based framework for representing complex worldviews through 'fields' of interconnected concepts and energy flows. It serves as a base for simulating important psychological, conceptual, and cognitive semantics processes mediated by worldviews. COGVIEW reflects the insight that information originating from human cognitive processes is of a different type than that typically considered in the 'hard sciences'; namely, it is *deeply nuanced and inherently distributed*.

COGVIEW uses networks of concepts to store information. Networks are defined as collections of nodes connected with edges (or links). Energy originates from energy sources and flows across edges (capable of modifying the energy that flows across them). Each edge has an indicated direction in which energy traversing it will flow.

COGVIEW supports two levels of cognitive processing, the C(onscious)-Level and the U(nconscious)-Level. U-Level processing is not accessible to conscious awareness and is dominated by the effects of associational memory in that the primary activity taking place at this level is the spreading of energy between concepts and concept fields and the associative detection of congruence/similarity between various portions of the extended concept universe. Critically, 'rational' or 'logical' thought does not take place at this level and is therefore incapable of acting as a filter for change. Growing evidence (see Shermer, 2002) suggests that at this level, mere understanding may be cognitively tantamount to acceptance, requiring an active act of 'disbelief' to be overcome. 'Raw' concepts and a wide range of energies can be accessed directly here and immense amounts of energy easily shifted and deployed, as well as emotions triggered, all without conscious or rational intervention.



**Clashes** In COGVIEW, a *clash* occurs when energy flowing in one direction meets energy flowing in the other, an especially important occurrence when the positive or negative valences of clashing energies are opposed to one another.

Clashes represent the point at which conceptual incompatibilities become manifest, and when the ability and/or need to take some sort of meaningful action has been identified, that is, when the model makes a certain type of 'conclusion'.

The subnetwork within which a clash occurs will naturally be activated most when a result propagates to consciousness, and thus is the subnetwork from which the conscious knower will consider the 'insight' of the clash to have originated. Critically, this is true even (as is the common case) when the energy which caused the clash came from other concept fields. This phenomenon allows for persuasive processes in which the semantic/conceptual content of a clash is removed from its original context (usually one in which conscious processing would have caused the desired conceptual understanding to be rejected) and shifted to another context (i.e. another subportion of the conceptual cloud) wherein which the attribution of the clash output to that context is 'safer' (or more desired).

The sites of clashes often coincide with the most relevant and important (moral) 'issues' that a human would identify within particular conceptual fields. This phenomenon provides one source of evidence that specific COGVIEW networks are accurately reflecting certain real world semantics.

Arguably, clashes have a neural basis; for example, Lieberman, Schreiber, and Ochsner (2003, p. 689-690) suggest that "The C-system [conscious] (named for the "c" in reflection), consisting of the prefrontal cortex, anterior cingulate cortex, and medial temporal lobes, is recruited when the X-system [unconscious] fails to create coherent outputs from the different sources of input. ... According to this model, the C-system is usually involved only to the extent that the X-system fails to resolve the current set of inputs into a coherent output ... [I]f the conflict between different considerations is too large, the C-system will detect this tension in the X-system and become involved (Botvinick, Braver, Barch, Carter, and Cohen, 2000)."

## Tools for Effecting (and Inoculating Against) Belief Change

In this section we introduce tools for use both at the macro and COGVIEW network levels.

### Network-Level Tools

This section provides a very high-level listing (due to space limitations) of techniques for adjusting COGVIEW networks. While useful on their own, a high-level understanding of these techniques is also helpful in understanding Directed Dissonance Reduction.

1. **Break Link Between Two Concepts:** Disconnects energy sources from downstream concepts (can be accomplished via DDR).

2. Introduce new connections to compete for energy with pre-existing connections.
3. **'Attraction':** Prime concepts likely to attract energy and thus create more desirable energy flows. Priming of 'incompatible' concepts (those causing dissonance with other salient concepts) may cause confusion, weakening previous connections.
4. Use 'attractor cognitions', that is, new cognitions more pleasant than undesired ones. Protoypical examples: 'hot-button' concepts like 'family'. Energy from these may 'overwhelm' that of undesired concepts or paths.
5. **'Blocking cognitions'** capable of blocking flow along undesired paths. To accomplish, associate original path with something unpleasant, confusing, or fear-inducing. Highlight path nonoptimality by pointing out logical inconsistencies or suboptimal conclusions arising from the path.
6. **Create Links:** Repeatedly reference pairs of concepts together, point out semantic similarities, and/or create scenarios where concepts are always associated or required to be used together to accomplish a goal.
7. **Emphasize / De-emphasize Concepts:** De-emphasize through neglect - emphasize through repetition.
8. **Redirect Link:** Prime desired concepts and paths. Reinforce through Repetition and Need-To-Use. Repetition: use narratives or linguistic constructions continually invoking and refreshing the salience of desired concepts or paths. Need-To-Use: cognitions become more salient when they are necessary to accomplish important goals.
9. **Change Polarity** (what was once sending negative energy now sends positive, and vice versa): Represents significant changes in perspective. Overwhelm energy of one polarity with energy from a source of the opposite polarity.
10. **Re-Normalize:** Show that concept 'X' is actually much more like concept 'Y' than was originally believed to be the case.
11. **Stigma Disconnect:** Stigma is a highly potent source of negative energy. Stigma Disconnect breaks energy flows associated with stigma as early as possible and/or refits them with more positive energy sources.
12. **Concept Implantation and Reconstruction of Incorrect Concept Fields:** Entire concept fields may be incorrect, either in terms of content or in connections.

Associations may be reinforced if they cover entire concept chains - that is, from energy input nodes to output nodes, as this provides cognitive 'closure' and makes chains easier to process.

## Directed Dissonance Reduction

Dissonance reduction (Festinger, 1957) represents a powerful mechanism for effecting psychological change, capable of marshaling significant emotional and mental energies. In this section we reformulate this mechanism in terms of COGVIEW networks, and later demonstrate how it may be repurposed in service of both belief change and 'inoculation'.

DDR redefines dissonance reduction as a process of *positive energy maximization*, suggesting that people seek to maximize the most positive energy obtainable within their belief networks. Dissonance is understood as a threat to current energy maximization, and dissonance reduction as an attempt to create a new equilibrium. Motivation arises from discomfort, with motivation levels proportional to the potential amount of threatened energy change.

## Stages of Directed Dissonance Reduction

### Phase I. Design

Phase I analyzes subjects' COGVIEW networks and identifies clash nodes, important positive energy sources, and potential psychological defenses.

COGVIEW allows putative belief changers to identify where concepts and energy sources lie in relation to others and to determine what concepts to prime and interconnect in order to 'reach' desired energy sources.

### Phase II. Initiation (Build Initial Context)

Priming of concepts identified in Phase I takes place here. This phase may also seek to add new useful concepts or links, though strategies built significantly on pre-existing networks are likely to be more effective, reliable, and easier to implement in practice.

If subjects perceive that large negative changes in energy balance may result, denial may occur. In order to avoid this defense strategy, significant positive energy sources (such as FAMILY) may be identified and primed at this stage.

In some cases, negative energy sources may be primed, for example, if subjects believe that there is 'no problem' or that a potential dissonance-inducing stimulus is of little or no significance. A canonical example is the use of fear by terrorists to cause target populations to take their goals seriously. Another involves health appeals, in which subjects may not be convinced of the reality of possible threats.

### Phase III. Dissonance Introduction

In this stage, a stimulus - a single or small set of new links whose existence threatens the existing positive energy balance - is introduced, causing dissonance. These must be significant enough to threaten belief network energy maximization.

Care must be taken here to avoid resort to denial as a clash defense strategy; if the target is able to simply ignore the stimulus altogether, it will not be effective. Thus, stimuli should be factually accurate, important, and otherwise impossible to ignore.

### Phase IV. Directed Dissonance Reduction

Here, energy is introduced and links are adjusted so that dissonance will be resolved as desired. This stage attempts to direct the choice of energy maximization equilibrium; energy placed at clash sites is most beneficial.

## Phase V. Solidification

This phase solidifies energy at important clash points in order to enhance satisfaction and attitudes towards outcomes, as well as make it less likely that changes will be reversed or re-examined.

The more likely it is that the new energy balance will persist in future, the more anxiety will be reduced during the dissonance reduction process.

## Directed Dissonance Reduction Examples

We now provide two detailed examples of how DDR can be used in diverse domains.

### Coming Out

The first example illustrates the case of a son who comes out to his father, whose view of the situation is initially quite inaccurate. The goal of DDR is that the father maintain his positive associations and attitudes about his son.

Figure 1 provides a diagram of the father's COGVIEW network before DDR has taken place. Energy flows from the energy source nodes, following the arrows indicated in the diagram, with two properties: magnitude (amount) and valence (positive or negative). Whenever energy crosses a graph edge labeled 'NEG', the valence of that energy is reversed.

Energy derives from three sources. *Biological* energy arises from parental love, the need to guide children, parent-child and father-son bonding, the desire to raise well-adjusted children capable of participating in the wider world, and the psychological need to view oneself as a good parent. *Societal* energy arises from in-built needs to be seen positively by the community as a good parent and as someone who has raised a well-adjusted child. Lastly, *Cultural* energy draws on local imperatives to guide children, act as good parents, and raise moral children.

**Phase I (Design):** DDR goals are as follows: *break* the link between INCORRECT-BELIEFS-ABOUT-GAYS and GAY (Link *IncorrectBeliefs*) and *maintain* the link between IDEAL-SON and THE-SON (Link *IdealSon*). The unfavorable outcome is the opposite, that is, the breaking of link *IdealSon* and the maintenance of the *IncorrectBeliefs* link. Another undesired outcome could be the breaking of the link between THE-SON and FATHER - in other words, the father disowning the son.

Clash points occur at FAMILY, RAISING CHILDREN, and THE-SON. At FAMILY, positive energy (+50) comes from SOCIETY and significant negative energy from COMBINED-ENERGY (initially positive but reversed after crossing the NEG edge near INCORRECT-BELIEFS-ABOUT-GAYS). Without mitigation this will flow into IDEAL-SON, ultimately making this path one that contravenes DDR goals. If positive energy is introduced at this node through priming or statements such as 'your son wants to raise children and start a family', 'gay people want to start loving families', and/or associating FAMILY with significant positive energy, positive energy may 'overwhelm' the negative energy at the clash point. A similar technique may be applied at RAISING CHILDREN.

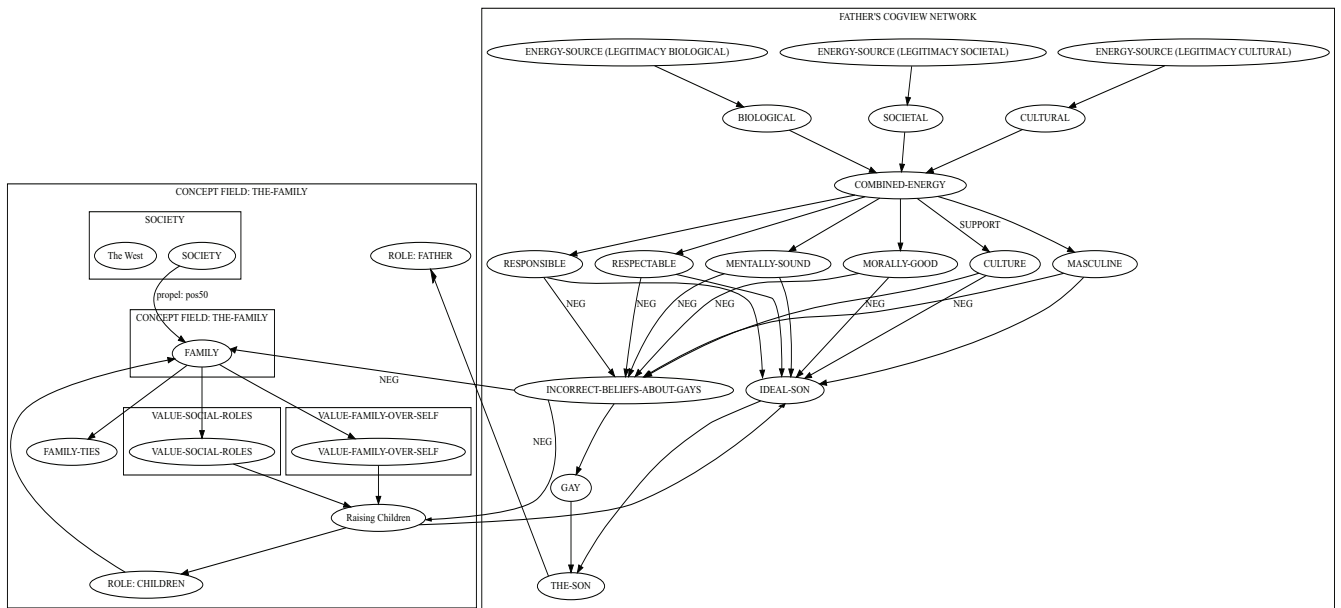


Figure 1: Father's COGVIEW Network before DDR

The clash located at THE-SON is particularly interesting, as it points to the exact dilemma faced by the father in this situation: either accept his son and break links to his incorrect beliefs or vice versa. Such correspondences can help verify how well COGVIEW networks fit the situations they represent.

Energy does not actually flow from INCORRECT-BELIEFS-ABOUT-GAYS to THE-SON until the dissonance-inducing cognition (linking GAY to THE-SON - the *ComingOut* link) is introduced, matching the reality of the problem being modeled.

Because THE-SON connects to FATHER, reminding the father that his son is a 'good' son will introduce positive energy into this node and make it more likely that he will not want to disown his son so that he may remain associated with this positive energy source.

Other positive energy sources include caring/compassion, parental attachment, the desire for enhanced well-being, the desire to focus on strengthening family and not break it apart, and so on. Other potential strategies could include creating new links between desirable nodes or changing the polarity of existing links between personal attributes (RESPONSIBLE, RESPECTABLE, etc.) and GAY or THE-SON.

**Phase II (Initiation):** In Stage II, the priming-related components of the Phase I design are carried out.

**Phase III (Dissonance Introduction):** In this stage, the link (from GAY to THE-SON), labeled as *ComingOut* above, is introduced by the son coming out to the father. Denial must

be avoided, which would appear as refusing to believe the son is gay.

**Phase IV: (Directed Dissonance Reduction)** At this stage, the full strategy designed at Phase I is implemented. Positive energy is inserted at FAMILY, RAISING CHILDREN, THE-SON, and so on. This stage makes use of link change strategies 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, and 12 (see above).

The goal here is to make it more painful for the father to configure his belief network in a manner contrary to the goals of DDR - that is, if he breaks or maintains non-preferred links, the maximum energy balance (equilibrium) he will achieve will ultimately be lower than that if he configures his network as desired. As an example, if the father breaks the link from IDEAL-SON to THE-SON, he will cease to receive the benefit of the positive energy entering the node THE-SON from IDEAL-SON. A similar outcome will arise if he breaks the link between IDEAL-SON and THE-SON.

**Phase V (Solidification):** Operations remain at Phase IV until sufficient change has been effected. To test, we may ask the father how he feels about his son, tallying the total number of positive and negative qualities.

Further positive energy may be added to THE-SON by focusing on the father's need that his son and other family members be associated with positive attributes, and re-priming may be performed on FAMILY, CHILDREN, and so on.

## Terrorism Reduction

DDR also has significant application in the domain of terrorism reduction. DDR can support public communications de-

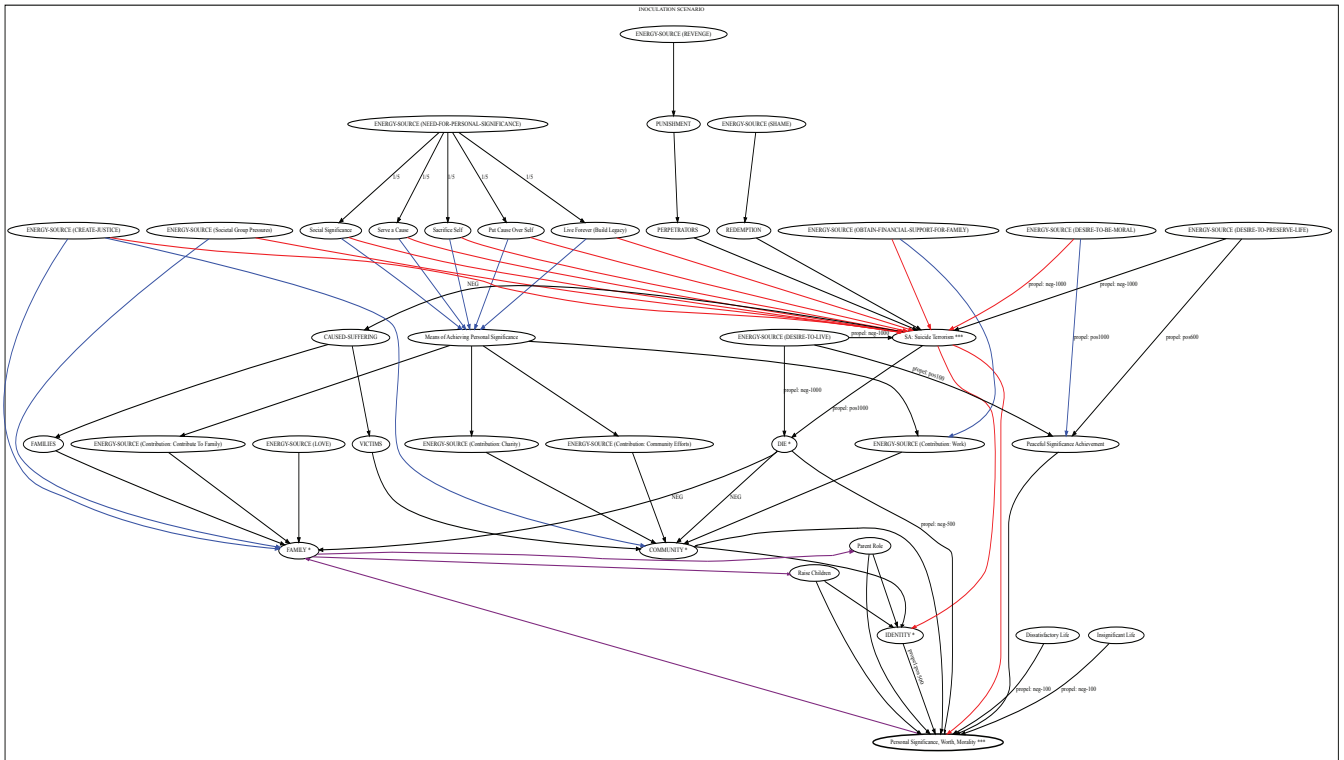


Figure 2: Citizen COGVIEW Network before Persuasion

signed to slow the adoption or 'taking root' of radical ideologies in susceptible individuals (termed 'inoculation'). Focus can also be placed on interrupting the flow from conception to the carrying out of acts.

We focus here on 'inoculation' against suicide terrorism, as it is a highly effective strategy capable of blocking entire event chains from occurring. Suicide terrorism's reliance on interlocking societal and personal value systems is an excellent fit for the COGVIEW+DDR approach.

We draw on Kruglanski et al. (2009)'s view of suicide terrorism as a quest for personal significance. We also draw on 'inoculation theory' (McGuire, 1961; Pfau et al., 1997), which suggests that subjects may be 'inoculated' against later attempts to change their beliefs by being exposed to weakened forms of counterarguments to their current positions. 'Beliefs' are understood here as COGVIEW links relevant to suicide terrorism and 'attempts to change beliefs' as attempts to convince others to undertake suicide terrorism.

Inoculation seeks to protect existing beliefs against change, assuming that people initially believe that suicide is wrong and that they should seek to make their life meaningful through appeal to family and activities rather than through terrorism.

Inoculation consists of two phases: *Threat* and *Refutational Preemption* (Pfau et al., 1997, 188-189). Threat motivates belief protection by demonstrating that beliefs (and energy maximization) can be threatened. Refutational Preemption inoculates against future threats by presenting weak-

ened forms of counter-arguments (which may ultimately be the same or different as what are eventually encountered, as the effect works in both cases.) (Pfau et al., 1997, 188-189) (McGuire, 1961, 330).

DDR assists (and induces) participants to 'create their own' counterarguments against persuasion attempts. As Pfau (1997, 189) suggests, "[i]t is the motivation provided to generate answers to potential counterarguments, as opposed to the specific information provided, that is responsible for [confering resistance]."

By highlighting anxiety related to energy equilibrium loss, DDR can help explain findings related to distraction and differential performance of inoculation as a strategy (see, for example Keating & Brock, 1974).

**Applying DDR** Threat occurs at Phases II and III of DDR, and Refutational Preemption at Phases IV and V.

**Phase I (Design - Threat):** Identify nodes expected to lose significant positive energy if persuasion occurs: FAMILY, COMMUNITY, IDENTITY, PERSONAL SIGNIFICANCE, WORTH, AND MORALITY and PEACEFUL SIGNIFICANCE ACHIEVEMENT. Also threatened is the *loop* between FAMILY and PERSONAL SIGNIFICANCE, WORTH, AND MORALITY. Loops such as these serve to simulate stability; the loss of stability represented by energy flowing within stable feedback loops is a significant loss in and of itself.

Simulate the introduction of negative energy into threatened

nodes (and loss of stability). This can be accomplished by calling attention to the following potential outcomes of energy loss:

**FAMILY:** Children without parents, not close to loved ones, loss of significance, judged as immoral, not supportive, broken family (leading to suffering, shame), ostracism, discrimination, prejudice, family suffering from revenge taken by others.

**COMMUNITY:** Chaotic communities, crime out of control, suffering, no rule of law, fear, bad example, children hurt, not growing up well, bad publicity for cause, martial law.

**IDENTITY:** Unclear about role in life, no personal meaning, unsure how to live life, deep instability, lose identity as good parent, good spouse, good member of community.

**PERSONAL SIGNIFICANCE, WORTH, AND MORALITY:** loss of face/worth to family, community, and friends, seen as wrong and immoral, family members feel societal shame, ashamed, unhappy due to lack of access when in need.

**PEACEFUL SIGNIFICANCE ACHIEVEMENT:** never-ending violence throughout the community, loss of life, family, and friends, international condemnation, loss of resources, great danger, great instability, fear throughout.

Loss of stability loop between **FAMILY** and **PERSONAL SIGNIFICANCE, WORTH, AND MORALITY**: great uncertainty, chaos, lack of predictability surrounding life.

**Refutational Preemption:** For this component we identify clashes at the following nodes: **SIGNIFICANCE ACHIEVEMENT: SUICIDE TERRORISM** (3 clashes), **PERSONAL SIGNIFICANCE, WORTH, MORALITY** (3 clashes), **FAMILY, COMMUNITY, DIE, and IDENTITY**.

Clash nodes are those most contested from a good-vs-bad moral perspective; at these sites worldviews may be modified significantly and immense shifts may occur. Clash points are marked with asterisks in Figure 2; the number of asterisks denotes the number of clashes occurring at any given node.

**Phases II and III** (Initiation and Dissonance Introduction) involve execution of the Threat component. Threat primes appropriate concepts and the existence of a threat in and of itself creates dissonance.

In Figure 2, blue links mark those that would be threatened and red links those that would be newly introduced should indoctrination succeed.

**Phase IV (DDR and Refutational Preemption):** The goal of DDR is to cause maintenance of blue links (those that are threatened) and rejection of red links (those that are indoctrination-related).

This can be accomplished by priming the clash nodes and significant energy sources that feed into them as follows: **PERSONAL SIGNIFICANCE, WORTH, MORALITY:** Prime **RAISE CHILDREN** and **PARENT ROLE** by highlighting importance of personal roles and positive emotions for all. Instinctual protecting, providing for. Prime **IDENTITY** through critical non-violent roles played by subject in the community.

**FAMILY:** Children with supportive parents in a stable home with a good reputation worthy of respect. The importance of the family as a unit *with the individual present to participate*

*within it.*

**COMMUNITY:** Stable communities conducive for leading a good, moral life, international support, and so on. Link strengthening, performed by simultaneously invoking **JUSTICE** and **COMMUNITY**, affecting the link between **CREATE JUSTICE** and **COMMUNITY**.

**DIE:** Prime **DESIRE-TO-LIVE** by cultural allusions to the importance of life, stories about what can be achieved during one's lifetime, meaning behind life, and so on. Mortality itself may be primed, but with undesirable follow-on effects.

Attempts may be made to introduce negative energy into **SIGNIFICANCE ACHIEVEMENT: SUICIDE TERRORISM**, though this is risky as it may be seen as a clumsy attempt to persuade, inducing change in an undesired direction.

The stability loop between **FAMILY** and **PERSONAL SIGNIFICANCE, WORTH, MORALITY**, will be automatically strengthened by introduction of positive energy to any constituent node, serving as further support for the status quo during DDR.

In the above, link-changing strategies 4,5, and 7 are employed.

## Conclusion and Next Steps

This paper has introduced strategies for modeling and changing beliefs and demonstrated examples of their application in the domains of prejudice and terrorism reduction. Further work will involve extended validation of the **COGVIEW** modeling framework and DDR technique, as well as development of further rubrics for using these techniques in critical real-world scenarios.

## References

- Festinger, L. (1957). *A theory of cognitive dissonance*. Row, Peterson.
- Keating, J., & Brock, T. (1974). Acceptance of persuasion and the inhibition of counterargumentation under various distraction tasks. *Experimental Social Psychology*(10).
- Kruglanski, A. W., Chen, X., Dechesne, M., Fishman, S., & Orehek, E. (2009). Fully committed: Suicide bombers' motivation and the quest for personal significance. *Political Psychology*, 30(3), 331–357.
- Lieberman, M. D., Schreiber, D., & Ochsner, K. N. (2003). Is political cognition like riding a bicycle? How cognitive neuroscience can inform research on political thinking. *Political Psychology*, 24.
- McGuire, W. (1961). Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *Abnormal and Social Psychology*, 63(2).
- Pfau, M., Tusing, K., Koerner, A. F., Lee, W., Godbold, L. C., Penaloza, L. J., ... Hong, Y.-H. (1997). Enriching the inoculation construct: The role of critical components in the process of resistance. *Human Communication Research*, 24(2), 187–215.
- Shermer, M. (2002). *Why people believe weird things* (2nd ed.). Holt.

# The Role of Comparison Processes in the Induction of Schemas for Design Styles

**Takanobu Omata (takanobu.omata@gmail.com)**

Department of Psychology  
University of California, Los Angeles  
Los Angeles, CA 90095 USA

**Keith J. Holyoak (holyoak@lifesci.ucla.edu)**

Department of Psychology  
University of California, Los Angeles  
Los Angeles, CA 90095 USA

## Abstract

Considerable evidence supports the effectiveness of close comparison of examples as a means to promote the induction of schemas that support generalization, especially to novel cases that require far transfer. The ease of comparison would appear to be maximized by presenting the to-be-compared cases in close spatial and temporal proximity. However, findings from a number of recent studies have been interpreted as evidence that induction is fostered not by presenting training cases for a single category together (massed practice), but rather by presenting them in an interspersed fashion (spaced practice). We address this apparent paradox in a study in which people are asked to learn the “styles” of furniture designs from a small number of examples of different products (e.g., a bed frame) and then classify examples of entirely different products (e.g., a chandelier). We contrasted a learning procedure based on comparison of examples presented simultaneously with procedures involving processing of individual items, either massed or spaced. Study time was minimized, and generalization was maximized, when learning was based on comparison. In a further study we use structural equation modeling to assess the content of the schemas for visual styles that are acquired by comparison processes. We propose that comparison fosters induction, whereas spacing facilitates retention and retrieval.

**Keywords:** induction; schema induction; design; spacing effect; structural equation modeling

## Introduction

Early work on learning and transfer based on analogy provided strong evidence that close comparison of two examples of a complex category (e.g., problems that can be solved by using converging weak forces) supports subsequent transfer to novel cases that exhibit a similar relational structure (Gick & Holyoak, 1983). The inductive benefit of comparison can arise either as a deliberate learning strategy or as a side effect of applying one solved source problem to an unsolved target problem (Novick & Holyoak, 1991; Ross & Kennedy, 1989). The positive impact of comparison has been demonstrated for both adults (Catrambone & Holyoak, 1989) and young children (Brown, Kane & Echols, 1986; Chen & Daehler, 1989; Holyoak, Junn & Billman, 1984; Kotovsky & Gentner, 1996; Loewenstein & Gentner, 2001; Namy & Gentner, 2002; Star & Rittle-Johnson, 2009). Comparison has been shown to guide schema formation in teaching such complex topics as negotiation strategies (Loewenstein, Thompson, & Gentner, 1999, 2003), and also

may play important roles in language learning (Gentner, 2010; Gentner & Namy, 2006). The dominant interpretation of these findings has been that comparison processes foster the induction of a schema for a class of situations, which in turn will facilitate subsequent transfer to additional examples (Gick & Holyoak, 1983).

It would be natural to assume that comparison, and hence induction, will be facilitated by presenting multiple examples simultaneously, or in close temporal proximity. However, this assumption has been challenged. The extensive literature on memory and retention provides robust evidence of an advantage for spaced over massed practice (e.g., Cepeda, Pashler, Vul, Wixted & Rohrer, 2006; Ebbinghaus, 1885/1964). Most of this research has focused on memory for specific items, such as words on a list. However, Kornell and Bjork (2008) showed that the advantage of spaced presentation over massed presentation of training examples extends to a task requiring induction of artistic styles. Classification of new examples was more successful when examples of paintings by different artists were intermixed during training (spaced condition) than when examples of paintings by an individual artist were presented in immediate succession (massed condition). These findings have been interpreted as evidence that spaced presentations actually facilitate participants’ generalization. Similar benefits of spacing have been observed in studies of children’s category learning (Vlach, Amkowsky & Sandhofer, 2012; Vlach, Sandhofer & Kornell, 2008).

On the face of it, the evidence for an advantage of spacing in fostering generalization poses a paradox. If comparison promotes schema induction, and is easier when the examples to be compared are presented in close proximity, it might seem that spaced presentation should hinder rather than help induction; i.e., one might expect spacing to be “the enemy of induction” (E. Z. Rothkopf, quoted by Kornell & Bjork, 2008, p. 585).

However, although simultaneous or massed presentation might be helpful or even necessary for comparison, the mere fact that examples are juxtaposed does not ensure that learners will engage in active comparison. Effective comparison typically is elicited by specific instructions to compare cases and write down commonalities (e.g., Gick & Holyoak, 1983). The benefit of such comparison instructions has been shown to greatly exceed that of simply providing two cases together (even on a single page) without comparison in-



structions (Gentner, Loewenstein & Thompson, 2003; Thompson, Gentner & Loewenstein, 2000; see also Kurtz, Miao & Gentner, 2001). In order to assess the impact of different presentation conditions, it is therefore important to include a condition in which learners are clearly instructed to perform active comparison of examples.

Accordingly, in Study 1 we directly contrasted a learning procedure based on comparison of examples presented simultaneously with procedures involving processing of individual items presented sequentially, either massed or spaced. We used a novel paradigm in which people attempt to learn realistic styles of furniture and related home décor. After being shown a small number of examples of home décor items of a specific style, participants were asked to judge whether examples of new décor items are of the same style (see Figure 1). This task involves far transfer, since the generalization items included different types of décor items than the training items (e.g., after seeing a dresser, bed frame, fabric and pillowcase set during training, a generalization item required judging the style of a chandelier). Although the relevant cues that might provide the basis for forming a schema are presumably visual, the schema is likely to be quite abstract, and not tied to any single décor type. In Study 2 we employ structural equation modeling in an effort gain insight into the nature of the style schemas acquired via comparison processes.

## Study 1

### Method

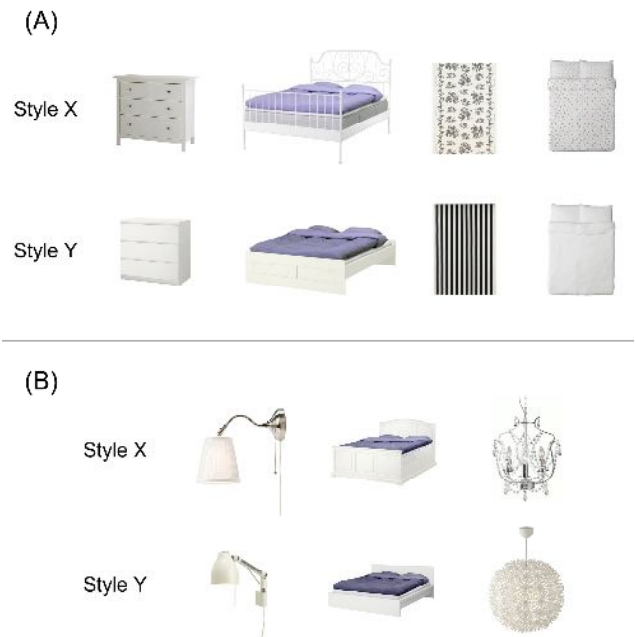
**Participants** A total of 147 participants (49 in each of three training conditions) were recruited online through Amazon Mechanical Turk (<http://www.mturk.com>). This system has been demonstrated to produce reliable data in many experimental studies (e.g., Paolacci, Chandler, & Ipeirotis, 2010; Mason, & Suri, 2011). Each participant was paid 60 cents for completing the study, which took about 10 minutes. At the conclusion of the study, information was collected about participants' age, nationality, and possible color blindness. Participants (52 male, 95 female) were all residing in the United States, and ranged in age from 18–61 years ( $M = 31.5$ ). We tracked IP addresses to ensure that participants were not repeatedly sampled. The participants described above excluded those who failed to complete all of the conditions in the experiment, or reported color blindness.

**Materials** The materials were 14 color pictures of IKEA furniture and other home décor items (7 each of two styles) printed in the catalog of IKEA 2012 (<http://info.ikea-usa.com/Catalog/>). The training items for each style (termed Style X and Style Y) consisted of a drawer, a bed frame, fabric, and a duvet set. The tests items used in the subsequent generalization phase were a wall lamp, a (novel) bed frame, and a chandelier (see Figure 1).

**Procedure and Design** Three conditions, Comparison, Massed and Spaced, were manipulated across participants. In the Comparison condition, the four items of each style, X and Y, were shown together on the screen (randomizing

position of items, and counterbalancing order of styles as well as assignment of the labels X and Y to styles). Participants were asked to write down three commonalities for each style. In the Massed and Spaced condition, instructions focused attention on individual items rather than commonalities. In the Massed condition each example of a given style was presented separately, but consecutively; whereas in the Spaced condition examples of the two styles were alternated. In both of the latter conditions participants were asked to write descriptions of each individual item of home décor. Participants advanced through the study phase at their own pace, and their total study time was recorded.

After the study phase, participants were presented with three pairs of new pictures of home décor items. For each pair one was a product of Style X, and the other a product of Style Y (positioned randomly on the left or right). Participants used an 8-point scale to rate which item was from Style X (where a rating of 8 indicated certainty that an item was from Style X).



*Figure 1.* (A). Four examples of each style presented in study phase of Study 1. (B). Three pairs of new pictures presented to the participants in the test phase. (NB: assignment of the labels X and Y was counterbalanced.)

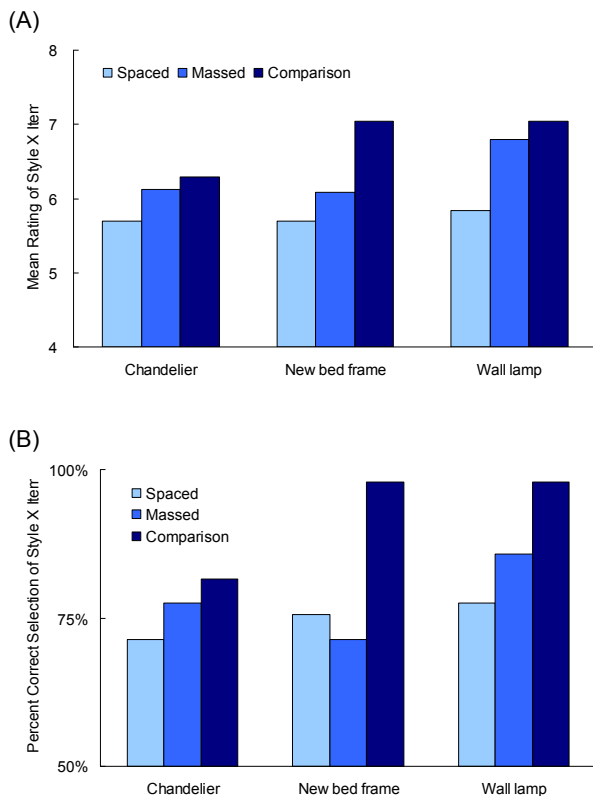
### Result and Discussion

Mean study times differed across the three conditions: 8.53 min (Spaced), 7.96 min (Massed), and 4.93 min (Comparison),  $F(2, 144) = 16.5$ ,  $MSE = 11.2$ ,  $p < .0001$ . Tukey tests showed that study time was lower for the Comparison condition than either the Massed or Spaced condition ( $p < .0001$ ), whereas the Spaced and Massed conditions did not differ reliably.

Figure 2A presents the mean ratings of the correct item representing Style X on the test phase, plotted such that



higher values (max = 8) indicate greater certainty that the correct item was indeed from Style X. A 3 x 3 mixed-factors ANOVA revealed a reliable effect of condition,  $F(2, 144) = 8.90$ ,  $MSE = 4.72$ ,  $p < .001$ . There was also a main effect of test item,  $F(2, 288) = 3.97$ ,  $MSE = 10.66$ ,  $p < .05$ , but the effect of condition did not vary significantly across items,  $F(4, 288) = 1.46$ ,  $MSE = 3.91$ ,  $p = .22$ . Collapsing over items, mean ratings were 5.74 (Spaced), 6.08 (Massed) and 6.79 (Comparison). Tukey tests revealed that generalization was more accurate for the Comparison condition than either the Spaced ( $p < .01$ ) or Massed condition ( $p < .05$ ), whereas the Spaced and Massed conditions did not reliably differ.



**Figure 2.** (A). Mean ratings of correct item from Style X for three pairs of new home décor items presented in test phase of Study 1. (B). Mean percentage of correctly selecting item from Style X for each of the three pairs.

We also calculated the percentage of correct choices by reducing the 8-point rating scale to a binary decision, breaking it at the midpoint (ratings of 1-4 vs. 5-8). As shown in Figure 2B, Comparison instructions yielded higher overall accuracy (93%) than either the Massed (78%) or Spaced (75%) condition. After aggregating across the three items, a one-way between-subjects ANOVA revealed a reliable effect of condition,  $F(2, 146) = 7.11$ ,  $MSE = 0.86$ ,  $p < .01$ . Tukey tests confirmed that generalization was more accurate for the Comparison condition than either the Spaced ( $p < .01$ ) or Massed condition ( $p < .05$ ), whereas the latter two

conditions did not reliably differ. The results of Study 1 thus provide clear evidence that comparison instructions designed to encourage induction of schemas for styles of home décor support more efficient learning and generalization than do presentation methods that focus attention on individual examples. The Comparison condition required less study time than either the Spaced or Massed conditions, yet yielded more successful generalization.

## Study 2

The results of Study 1 support the hypothesis that comparison instructions aid in inducing a schema for styles of furniture design, facilitating subsequent generalization. In Study 2 we examined the possible causal relationships between content of a style schema and generalization, using structural equation modeling to infer relevant latent variables. For modeling purposes only a single condition was run, with study, test, and evaluation phases. Participants first studied examples of multiple styles of home décor items, focusing on one particular style (called Style X), then judged which of new examples were from Style X, and finally rated descriptive words as to how well they fit their impression of Style X. The ratings of descriptors were used to estimate latent variables representing a schema for Style X, with the aim of predicting the generalization data from the test phase.

## Method

**Participants** A total of 175 participants (50 male, 125 female), ranging in age from 18–70 years ( $M = 32.4$ ), all residing in the United States, were recruited through Amazon Mechanical Turk. Participants were paid 60 cents for participating in the experiment, which took about 10–15 minutes.

**Materials** The materials were 23 pictures of IKEA home décor items, five representing each of five styles (labeled X, A, B, C, and D), selected from the same source as in Study 1. Style X (a specific style that was constant for all participants) was the focus of our modeling effort (see Figures 3–4). In addition, a list of 30 descriptive words was created for use in the evaluation phase (see Table 1). To create this list, 118 participants in an initial survey (also conducted on Mechanical Turk) were asked to generate words describing Style X (based on the same three examples used in the study phase of Study 2). Using these generated words as a starting point, we selected relatively high frequency and non-redundant words, and added several additional words (as distractors) that had *not* been generated to describe Style X.

**Procedure and Design** In the study phase, three pictures (a drawer, a bed frame, and a chandelier) were presented together for each style (see Figure 3). The study phase began with three steps that applied only to the examples of Style X (the style that was the direct focus of structural equation modeling). Participants (1) listed three common features of the examples of Style X; (2) described their “impression, feel, or sense of Style X” based on its common features; and (3) described the “personality” that Style X would represent

if it were thought of as a person. Next, participants saw all five examples of each type of home décor item (e.g., stools of each of the Styles X and A-D), and were asked to compare them and describe the difference between Style X and the other styles. This task was repeated for each of the three types of training items. The presentation order of pictures in Style X and the other styles was fully randomized.

During the subsequent test phase (see Figure 4), pairs of examples of four novel types of home décor items were presented (stools, wall lamps, duvet sets, and fabric). Each pair included one example of Style X and one example of some other style (Styles A-D). For each pair, participants rated which example was more likely to be from Style X, using the same 8-point scale as had been used in Study 1.

Finally, in the evaluation phase, participants were presented with the 30 descriptors in random order. Participants used a 4-point rating scale to evaluate how well word captured their impression of Style X (with a rating of 4 indicating that the descriptor “definitely” fit Style X).



Figure 3. Three examples of each style presented in study phase of Study 2.



Figure 4. Four pairs of new pictures presented to the participants in the test phase of Study 2. For each pair, the item in the left column is from Style X; that on the right is the foil, drawn from Styles A-D.

Table 1. Set of 30 descriptors used in evaluation phase of Study 2.

Words produced as descriptors of Style X	
feminine (25), old-fashioned (18), girly (17), traditional (9), plain (9), fancy (8), country (7), elegant (6), stylish (4), pretty (4), conservative (3), familiar (3), light (3), luxury (3), relax (3), cheery/active (2), romantic (2), soft (2), warm (2), modern (1), unisex (1)	
NB: number of respondents in initial survey (out of 118) who produced each descriptor in response to Style X is indicated in parentheses.	
Distractors (not produced in response to Style X)	
casual, cool, fashionable, formal, gorgeous, hard, masculine, natural, wild	

## Results and Discussion

Prior to conducting structural equation modeling, we checked descriptive statistics from the evaluation phase, identifying five descriptors with means ratings above 2.50 on the 4-point scale of applicability to Style X. We then performed exploratory factor analysis using these five scales, obtaining a promax rotation by maximum likelihood estimation using SPSS. Contribution ratios and eigenvalues were computed, and two factors were retained (eigenvalues > 1.0), which explained 60.0% of total variance. The inter-factor correlation between the two factors was 0.24.

We then performed structural equation modeling using EQS version 6.1 (<http://www.mvsoft.com/>) in order to identify apparent causal relationships between a style schema (based on the latent variables derived from ratings of descriptors) and the ability to discriminate examples of Style X on the test phase. The analysis identified a model with three latent factors that provided an excellent statistical fit to

the data,  $\chi^2(24) = 31.3, p = .14$ . The value of Root Mean Square Error of Approximation (RMSEA) was 0.04, and the value of the Comparative Fit Index (CFI) was 0.98.

The structural equation model is displayed in Figure 5. The factor labeled D (Discriminability) summarizes the ability to discriminate Style X from others on the test phase, based on the four items used to assess generalization of the knowledge about Style X acquired during the study phase. The D factor was in turn predicted by factors F (Femininity) and C (Classic), derived from the ratings of descriptors dur-

ing the evaluation phase. Specifically, factor F was associated with ratings of the words “feminine”, “fancy”, and “girly”, while factor C was associated with ratings of “traditional” and “old-fashioned”. The factor loadings for Factors F and C exceeded 0.50 and for factor D were over 0.20. Factors F and C had positive path coefficients between each other as well as to factor D. Intuitively, participants who associated Style X with descriptors indicative of a feminine and classic image were best able to discriminate examples of Style X from alternatives on the generalization test.

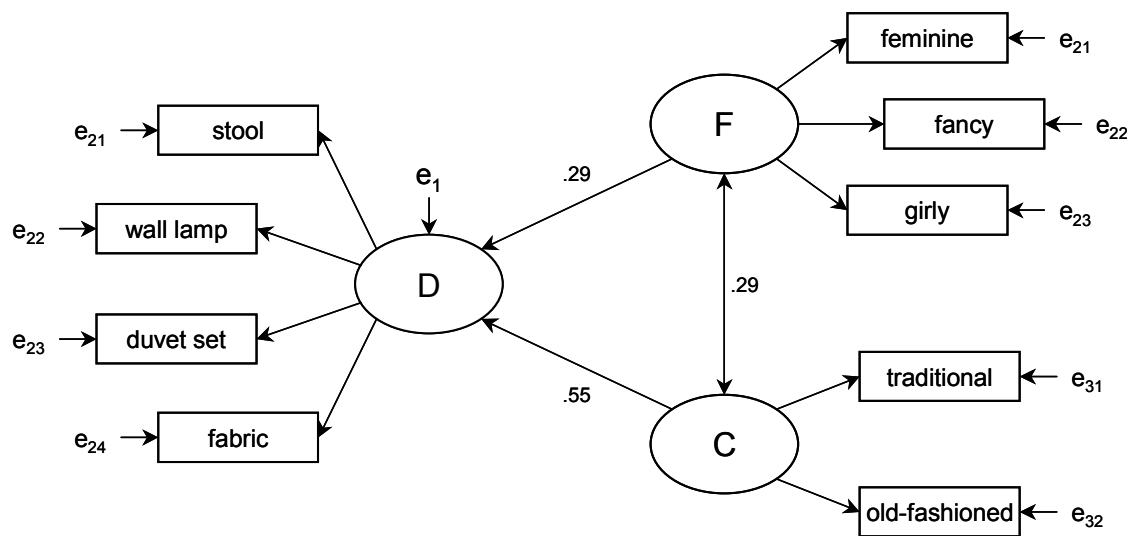


Figure 5. The model for predicting generalization of Style X to novel items on the test phase, derived by structural equation modeling in Study 2. Values of correlation coefficients are indicated. Error terms are denoted by subscripted *e*.

## General Discussion

Using realistic visual categories (styles of furniture and other home décor items) we showed in Study 1 that active comparison of examples of a category is more efficient (reduced study time and greater accuracy) in promoting subsequent generalization than is processing of individual examples, whether massed or spaced. These findings support the hypothesis that schema induction underlies successful generalization of a visual style. In Study 2 we attempted to assess the content of a design schema. Using structural equation modeling, we showed that after learning by comparison, generalization performance can be predicted by latent variables corresponding to relatively abstract concepts derived from descriptive terms.

Taken together, these findings support the usefulness of the concept of schema induction as a basis for generalization with complex categories. The results of Study 1, in particular, suggest that it would be a mistake to view spaced presentation *per se* as an optimal procedure for promoting induction and generalization. Not only did the Spaced condition yield poorer generalization (despite longer study time) than the Comparison condition, but it showed no advantage

(trend in wrong direction) relative to the Massed condition. Massed presentation, although not sufficient to reliably trigger comparison processes, may nonetheless make such processing more likely.

Based on other findings in the literature (e.g., Kornell & Bjork, 2008), there clearly are some conditions under which spacing improves later generalization performance, particularly in comparison to massed presentations that focus on encoding of individual items. In fact, comparison and spacing may well convey distinct and complementary benefits for categorization performance. In terms of the traditional stages of memory—encoding, retention and retrieval—schema induction can be viewed as a special type of encoding that focuses on abstraction of general cues to category membership, as opposed to encoding of individual training examples. We suggest that comparison facilitates schema induction during encoding, whereas spacing has its greatest impact on retention and retrieval. For complex categories of the sort used in the present studies, spacing likely makes schema induction more difficult; however, spacing may sometimes provide compensatory benefits in increasing retention and later retrieval of knowledge. This hypothesis suggests that generalization might be optimized by early

comparison of concurrently-presented examples, followed by subsequent spaced presentation of additional examples. Further research will be needed to explore the potential interactions between factors that guide induction of schemas and those that aid their retention and retrieval.

### Acknowledgments

Preparation of this paper was supported by a gift from Canon Incorporated, and by grant N000140810186 from the Office of Naval Research. We thank Airom Bleicher for assistance in running the studies.

### References

- Brown, A. L., Kane, M. J., & Echols, C. H. (1986). Young children's mental models determine analogical transfer across problems with a common goal structure. *Cognitive Development, 1*, 103-121.
- Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 1147-1156.
- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354-380.
- Chen, Z., & Daehler, M. W. (1989). Positive and negative transfer in analogical problem solving by 6-year-old children. *Cognitive Development, 4*, 327-344.
- Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger, C. E. Bussenius, & E. R. Hilgard, Trans.). New York: Dover. (Original work published 1885.)
- Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34*, 752-775.
- Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95*, 393-408.
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.
- Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 234-259). New York: Oxford University Press.
- Holyoak, K. J., Junn, E. N., & Billman, D. O. (1984). Development of analogical problem-solving skill. *Child Development, 55*, 2042-2055.
- Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*, 585-592.
- Kotovskiy, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67*, 2797-2822.
- Kurtz, K. J., Miao, C., & Gentner, D. (2001). Learning by analogical bootstrapping. *Journal of the Learning Sciences, 10*, 417-446.
- Loewenstein, J., & Gentner, D. (2001). Spatial mapping in preschoolers: Close comparisons facilitate far mappings. *Journal of Cognition and Development, 2*, 189-219.
- Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin and Review, 6*, 586-597.
- Loewenstein, J., Thompson, L., & Gentner, D. (2003). Analogical learning in negotiation teams: Comparing cases promotes learning and transfer. *Academy of Management Learning and Education, 2*, 119-127.
- Mason, W.A., & Suri, S. (2011). How to use Mechanical Turk for cognitive science research. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 66-67). Austin, TX: Cognitive Science Society.
- Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General, 131*, 5-15 10.
- Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 398-415.
- Paolacci, G., Chandler, J., & Ipeirotis, P.G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411-419.
- Ross, B. H., & Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 42-55.
- Star, J. R., & Rittle-Johnson, B. (2009). It pays to compare: An experimental study on computational estimation. *Journal of Experimental Child Psychology, 102*, 408-426.
- Thompson, L., Gentner, D., & Loewenstein, J. (2000). Avoiding missed opportunities in managerial life: Analogical training more powerful than individual case training. *Organization Behavior and Human Decision Processes, 82*, 60-75.
- Vlach, H. A., Ankowski, A. A., & Sandhofer, C. M. (2012). Same time or part in time? The role of presentation timing and retrieval dynamics in generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 246-254.
- Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition, 109*, 163-167.

# A State-Event Transformation Mechanism for Generating Micro Structures of Story in an Integrated Narrative Generation System

Kou Onodera (g231i007@s.iwate-pu.ac.jp)

Graduate School of Software and Information Science, Iwate Prefectural University, 152-52 Sugo  
Takizawa, Iwate 020-0193 Japan

Taisuke Akimoto (g236i001@s.iwate-pu.ac.jp)

Graduate School of Software and Information Science, Iwate Prefectural University

Takashi Ogata (t-ogata@iwate-pu.ac.jp)

Faculty of Software and Information Science, Iwate Prefectural University

## Abstract

This paper describes the current version of state-event transformation system to transform story lines and story worlds each other using a knowledge base and a conceptual dictionary. Moreover, we extend the basic framework to a circulative generation mechanism with a simple mutation function. Through the preliminary performance checks, we confirmed that the transformation of story worlds and story lines is approximately logically adequate and the circular process produces the diversity of stories. The proposed system is a module in an integrated narrative generation system and the pilot version is already implemented. In the context of the narrative generation system, the proposed system plays roles for expanding the variation of discourse to be generated and limiting possible narrative elements at any given time in a narrative generation process.

**Keywords:** Narrative generation system; story generation; state; action; conceptual dictionary.

## Introduction

In the context of natural language processing in the wide sense, the research of narrative generation system which aims at automatic generation of narrative texts by computer has been developing from 1960s. Meehan (1977) shows a classical approach and Bringsjord and Ferrucci (2000) is an example of the comparatively new result. Along with traditional literary genres, narrative and story will play important roles for digital entertainment genres such as computer game. The mechanism we propose in this paper is a part in an integrated narrative generation system we have studied as a project. The applicable goal is creating novel contents such as automatic generation game, which has not a fixed story, and narrative generation based narrative or literature, which is a form of novel containing narrative generation mechanisms. Moreover, narrative is the strongest method for organizing and structuring fragmentary information and a kind of collective knowledge in human being. The narrative generation system is also associated with a variety of issues such as the organic formation of fragmentary information, the diverse interpretation of an event or events, and so on (Ogata & Kanai, 2010).

In this paper, we deal with a part of mechanism for generating a story, which is a sequence of events to be narrated, in the narrative generation process. In concrete terms, we propose a system to make a correlation between state and event (action) which are main elements to construct

stories using a conceptual dictionary for noun/verb concepts and transformation rules. We analyze and classify the relationship between an action and the states in front and behind for 689 verb concepts to develop a mechanism that mutually transforms from an action to states or from states to an action and cyclically repeats the process. The action means an event in which a verb concept for an action is included as the central element. A temporal sequence of events is corresponding to a story. On the other hand, a collection of states means a static narrative knowledge supporting events.

## Narrative Generation and Proposed Mechanism

Three main modules of our narrative generation system are story generation mechanism, discourse mechanism, and surface expression mechanism. Story means a sequence of events to be narrated and discourse means the narrated structure of events, and both are described with conceptual representation. Expression contains surface representations including natural language, animated movie, music, and so on. The system has a conceptual dictionary and various narrative knowledge bases used mainly in story and discourse parts. Although our previous works was to develop the comparatively independent modules, we are currently starting to complete an integrated narrative generation system in which a variety of mechanisms are synthesized by standardizing data structure for event representation and constructing a conceptual dictionary to be used in a lot of modules commonly (Akimoto & Ogata, 2011). Figure 1 shows the overall structure of a pilot version of the system. The proposed system in this paper is corresponding to both “SL (story line)→SW (story world)” and “SW→SL” in the “Structural operation module”. As mentioned later in details, a story line is a sequence of events and a story world is a collection of states.

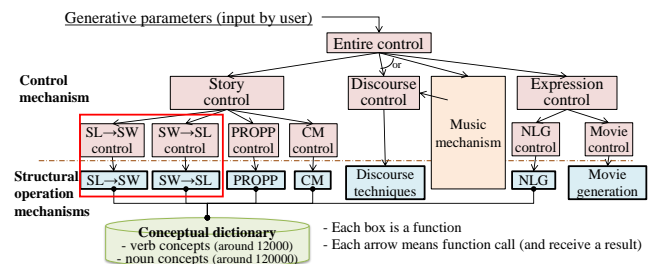


Figure 1: The overall structure of an integrated narrative generation system



A story is represented as a tree structure as shown in Figure 2. Three types of components in a story tree are relation, event, and state. An event forms a case frame with some elements such as agent, object, location, and so on. A relation semantically binds some events or the following part. The generation mechanism in this level is described in Ogata, Hori and Ohsuga (1994). The lowest layer of the tree is formed by states. A state means the precondition and the result of an event. Each state is represented with some kinds of frames describing the values of case frames in the corresponding event, namely it actually is a set of semantic data. We call the layer of states “story world” and the layer of events “story line”. An event produces two states and two states are corresponding to an event. This paper describes the detailed structures and proposes the mechanism of mutual transformation processing by refining and expanding previous studies (Nakashima & Ogata 2008; Onodera, Oishi & Ogata, 2011). In the context of narrative generation system, the primary roles of this mechanism are providing a function for generating narrative representations such as descriptions and explanations besides events sequences and a construction for limiting possible elements (like agent, object, and place) at any given time in a narrative generation process, using the detailed and explicit definition of states.

## Research Background

In the research of “the history of narrative” by philosophers (Noe, 2005), history is a kind of story to transform a chronicle as chronologically arranged events into a sequence of events sorted through the narrator’s filter. As described above, we divide a narrative generation process into the information to be narrated (story) and the narration itself (discourse and surface expression). It may be thought that the part of state and event in this paper is corresponding to the chronicle. One of the characteristics in our narrative generation system research is an interdisciplinary approach of the humanities such as narratology & literary theories and computer science such as AI. The materialization as a program of the conceptual idea and theory contributes to give the new tools of thought to such area. From this point of view, the problem is that what kinds of knowledge and techniques are used for executing the transformation from states as a chronicle to events as a story. Although we propose only the micro level’s technique in this paper, we are aiming at building the narrative generation system on such conceptual and philosophical foundation.

In AI, this paper is related to the function of planning that an agent automatically executes a goal through a goal oriented process. Planning technique, which is a strong foundation in many story or narrative generation systems (Mueller, 1990; Okada & Endo, 1992), generates stories or

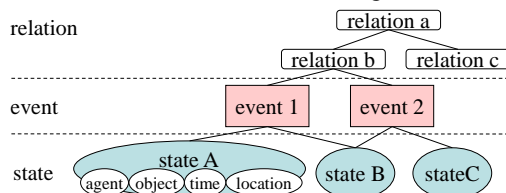


Figure 2: The structure of a story

plot lines to reach a goal through the recursive division of the goal. Although the planning is a comparatively macro framework for story generation, in this paper, we pay attention to the part of micro transformation process between states and events without a specific goal. Therefore, it is not planning in a strict sense. As our narrative generation system adopts a kind of modular approach in which a variety of elements are organically integrated, the aspect of goal-based planning can be incorporated into the system as a different module.

Next, conceptual dependency theory by Schank (1975) is a classical study regarding to the semantic categorization of concepts and categorizes verb concepts into 11 types of fundamental concepts. It is difficult to faithfully categorize real diverse concepts in accordance with the theoretical framework. However, we will be able to partially apply it or the idea as a reference, regarding to the part of especially abstract actions. A verb thesaurus by Takeuchi (2011) also shows a detailed system of semantic classification of verbs. It classifies verbs into basic events (states and activities) and complicated events (verbs for state change). Verbs for transitioning among states in this thesaurus are corresponding to verb concepts for changing states in this paper. Although our research shows more detailed classification about the categories of moving of agent and object, we may be able to refer to this thesaurus’s classification about other categories, namely, human relationships, body information such as personal appearance and physical condition, psychological information such as emotion, change of hardness and shape of object, etc. In addition, FrameNet (Fillmore & Baker, 2010) and Japanese FrameNet (Ohara, 2008) define the semantics of a word by a semantic frame, which means the structured knowledge about a typical scene, constructed with the frame elements. And, VerbNet (Kipper-Schuler, Dang, & Palmer, 2000; Kipper-Schuler, 2005) is a hierarchical verb lexicon in which each class is described by semantic predicates, thematic roles, and basic syntactic frames. Because this research’s purpose is not building a conceptual lexicon itself and the implementation from the viewpoint of the narrative generation system, we do not aim at the direct use or the direct revision of above resources currently. However, for revising the knowledge base mentioned later, the more direct reference will be needed in the future.

Moreover, the proposed system uses a conceptual dictionary to transform between states and events. As an example of story generation systems by the use of conceptual knowledge base, McIntyre and Lapata (2009) proposed a system which generates stories by using a knowledge base of compositions of an event sentence and chains of events. Stories are generated by a kind of tree search of possible stories. The system has several scoring criteria for pruning low scored branches. Although this is a completed story generation system, the proposed system is a mechanism in a big framework of narrative generation system to be able to add various other mechanisms such as a selection function based on some specific criteria on the foundation of a simple basic mechanism.

## An Overview of the Mechanism

Figure 3 draws an image of the state-event transformation mechanism. The state of A (a place in this case) changes or

does not change based on an event. A state includes some kinds of information such as a place, an agent's features, and object or objects, and so on. These elements in a state are described by each frame. For example, an agent's frame contains name, location, time, possession, and so on. The each element is really corresponding to an instance as the substantiation of each element in a conceptual dictionary. On the other hand, an event is described by a conceptual representation form shown in Figure 4. When the system recognizes a change in the states of agent, object, place, and time, it infers an event using "state-event transformation knowledge base". In contrast, the system can also infer two states from an event using the same knowledge base. We call a set of states "story world" and a sequence of events "story line". In summary, in a story line, each event is represented with a case frame including agent, object, location, time, and so on. In a story world, each state is represented with a set of four types of frames of agent, object, location, and time corresponding to above cases. And, each frame is constructed with slots and the values. A state means static knowledge and an event means dynamic one generated by the change of two states. The system can also repeat a circulative transformation process between states and events. The objective is to add various changes to the flows of stories through the cyclical process. However, as the state change and events are the relationship of one-to-many, in the case of simple circulation, the change of story worlds remains small compared with the comparatively large change of story lines. To add a larger change to story worlds, a kind of simple mutational function is introduced. This system is implemented by Common Lisp as with other main parts in the narrative generation system.

The conceptual dictionary is a hierarchical system for verb and noun concepts developed referring to "Goi-Taikēi: a Japanese lexicon" (Ikehara et al., 1999) and "Japanese WordNet" (Bond et al., 2009) mainly. It contains about 12000 verb concepts and about 120000 noun concepts. In the current version in this paper, we limitedly treat only verb concepts relevant to the "physical transfer" and "possessive transfer". The each verb concept is described as a case frame form

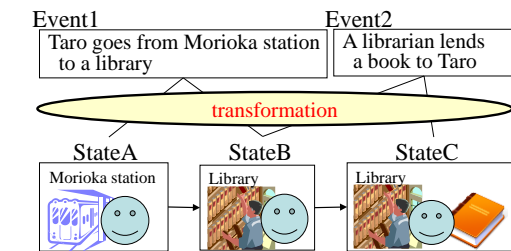


Figure 3: An image of the proposed mechanism

age1 goes loc2	Case	Function of case
(event 行く(1)[go(1)] (type action) (ID 1) (time (time1 time2) (agent age1) (counter-agent nil) (location loc1) (object nil) (instrument nil) (from nil) (to loc2))	type	Kind of event (action or happening) (Prince, 1987)
	ID	Identification of an event
	time	A pair of starting time and ending time of event
	agent	Agent of the action
	counter-agent	Object of the action (living thing)
	location	Occurrence of the event
	object	Object of the action (no living thing)
	instrument	Instrument or tool used by the action
	from	Starting position of the action
	to	Ending position of the action

Figure 4: The conceptual representation form of an event

shown in Figure 4. Oishi and Ogata (2011, 2012) propose the detailed explanation.

## State-Event Transformation Knowledge Base

This stores rules for transforming two states and an event mutually. As shown in Figure 5, a transformation rule consists of following three elements: (1) The pattern of the change of states, (2) a set of preconditions for the change, and (3) A set of verb concepts associated with the change. This example means "change of the place where a character exits". Above (2) is the condition to apply a transformation rule and this example means "an agent exits at a place, and the place and another place exists". In above (3), a group of verb concepts produce a same type of state change in common. In the current step of implementation, 138 "transformation rules" and 689 verb concepts are contained.

The transformation rules are hierarchically classified. The highest layer is divided into "the change of location, address, possession, health, durability, and posture" and "generation and disappearance" according to the changing element in each slot in agent, object, and place. Second layer shows the concrete way of changing in a slot and is classified into 27 kinds. For example, in the movement of place, there are the movement among different places, the vertical movement, and the high/low movement in a same place. And, one or more rules according to each movement can be defined. For example, in the movement among different places, an event satisfies both slots of "from" and "to" in common and another event satisfies only "from" or "to". Table 1 lists defined categories of state change. For example, the change of location slot in an agent frame or an object frame is corresponded to 11 types of categories. In this table, "large classification" means it is based on what (or which slot) is changed in a state and "small classification" means it is based on how it changes. For about 4200 kinds of "physical actions that an agent becomes the subject", about 2000 kinds of "other physical actions", and about 4100 kinds of "psychological actions", we are currently proceeding this expansion and revision by focus on what is changed in a state (we need to expand frame as necessary), how a state changes, and what kind of cases a verb concept has.

Type of state changes	
(Change of location: instance moves from X to Y)	
<b>Rule1: agent moves from X to Y</b>	
<b>Pattern of the change of states</b> (agentX's location changes from X to Y)	
<b>Precondition</b> (condition1 (there is agentX in locationX)) (condition2 (there is locationX)) (condition3 (there is locationY))	
<b>Set of verb concepts</b> (歩く(1)[walk(1)] 移動する(1)[move(1)] ...)	

Figure 5: A state-event transformation rule

Table 1: The classification of transformation rules (the details are omitted)

Large classification	Small classification
Change of location	Move, Move[up], Move[down], Move[out], Move[in], Move[near], Move[far], Carry, Go together, Near, Leave.
Change of posture	Stand, Sit, Lie.
Change of possession	Have, Relinquish, Give, Trade, Throw.
Change of health/durability	Death, Break, Damage, Heat.
Change of address	Migrate
Generation disappear	Generation, Disappear, Ingest



## Two Processes of Generation

The flows of transformation are from a story line to a story world (“story world generation”) and from a story world to a story line (“story line generation”). In addition, these transformations can be also repeated continuously or cyclically.

### Story World Generation Process

The process of story world generation is as follows (Figure 6): (1) The user specifies a story line and each value of agents, objects, and places in the story line. (2) The system searches a rule in a state-event transformation knowledge base according to a verb concept from first event in the story line. (3) The system selects a rule which has the verb concept in the group of verb concepts. (4) The system sets a state before the event which is the precondition in the selected rule. If a state already exists, the system simply overwrites by the new state. (5) The system refers the pattern of the change of states in the rule and makes a state after the event by changing the previous state set in above (4). (6) The system refers next event. If it exists, the process returns to above (2). If it does not exist, the process finishes.

### Story Line Generation Process

The flow of story line generation is as follows (Figure 7): (1) The user specifies two or more states in a story world according to the numerical order of ID. (2) The system compares the changed frame(s) and slot(s) in the first two states. (3) In the state-event transformation knowledge base, the system searches a transformation rule in which the first state & the precondition and the change in the second state & the content of change respectively match. (4) The system checks the constraints for each verb concept stored in the rule.

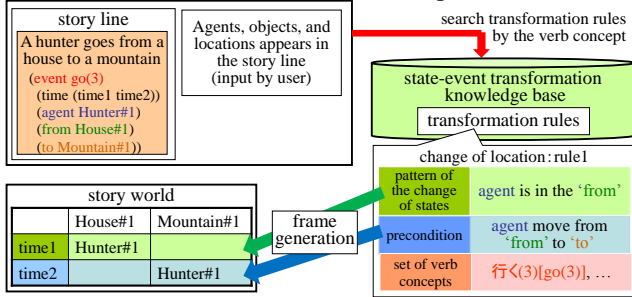


Figure 6: The flow of story world generation

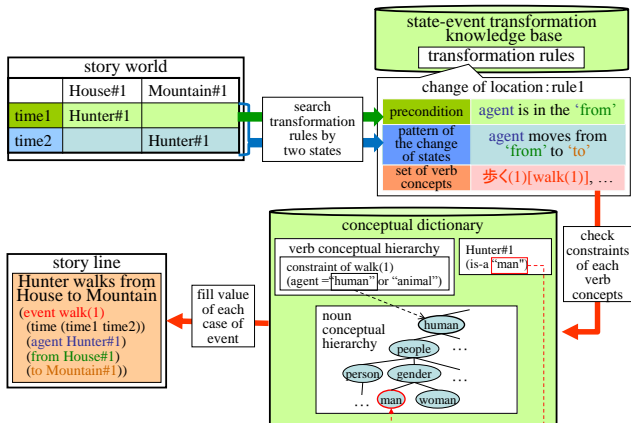


Figure 7: The flow of story line generation process

Specifically, the system checks whether the agents and objects are within the range of each constraint. And, the system allocates agents, objects, and places to the case frame for a verb concept that satisfies all constraints to generate an event. (5) If one or more events remain, the process returns to (2). And if no event, the process finishes.

### Circulative Generation Process

The system can repeat the transformation among story worlds and story lines. For example, an event such as “Taro goes to the park from the house” can be transformed into two states, “Taro is in the house” and “Taro is in the park”. Next, these states can be transformed into an event that is different from the former such as “Taro comes to the park from the house”. The system enables “circulative generation” in which two types of transformations are repeated each other. A verb concept in an event in a story line changes within a given range through such circulative process. However, a story world does not change from the initial story world because it is knowledge to prescribe the story content. Story world and story line have one-to-many relationship in the ordinary mechanism.

Moreover, we introduce a mechanism to produce larger changes in both story line and story world through the circulative generation process. Although we can imagine various methods to actualize it, we implemented a function using a simple mutation mechanism. In story world generation, the function randomly changes a state and the value of an agent or an object in a frame after the generation process. On the other hand, in story line generation, it randomly changes an event and a case’s value in an event concept after the generation process.

## Execution Examples

First, as the input information, we prepare the summary of a part of Mishima’s “The sailor who fell from grace with the sea” (1963) (Figure 8). Table 2 shows a generated story world. In addition, Figure 9 shows the result of a story line based on the story world. In these figures, we show the result by a simple sentence generation program. Figure 8 is a scene that Ryuji, Noboru, and a “don” move from a hill to a dry dock in Yokohama. In Figure 9, some changes occur. For example, Ryuji and Noboru move together in one event and the don moves as if he comes after Noboru.

Second, circulative generation was attempted based on a tale in Yanagita’s “The legends of Tono” (1955) (a story that a hunter forces a big priest to eat burned stones to punish the big priest who ate the hunter’s rice cakes without the permission).

E1: Ryuji#1 takes Sailor cap#1 in Ryuji#1's hand	E17: Ryuji#1 heads to Hill#2
E2: Ryuji#1 goes from Store#1 to Hill#1	E18: Noboru#1 heads to Hill#2
E3: Don#1 goes to Hill#1	E19: Don#1 heads to Hill#2
E4: Noboru#1 goes to Hill#1	E20: Ryuji#1 arrives at Dry dock#1 from Hill#2
E5: Ryuji#1 supplies Sailor cap#1 to Don#1	E21: Noboru#1 arrives at Dry dock#1 from Hill#2
E6: Noboru#1 buys Sailor cap#1	E22: Don#1 arrives at Dry dock#1 from Hill#2
E7: Noboru#1 delivers Sailor cap#1 to Ryuji#1	E23: Ryuji#1 sits down in Dry dock#1
E8: Ryuji#1 leaves Hill#1 for Sugita#1	E24: Noboru#1 sits down in Dry dock#1
E9: Noboru#1 leaves Hill#1 for Sugita#1	E25: Don#1 sits down in Dry dock#1
E10: Don#1 leaves Hill#1 for Sugita#1	E26: Ryuji#1 supplies Sailor cap#1 to Noboru#1
E11: Ryuji#1 goes around Slope#1	E27: Noboru#1 parts with Sailor cap#1
E12: Noboru#1 goes around Slope#1	E28: Noboru#1 takes Tea#1 in Noboru#1's hand
E13: Don#1 goes around Slope#1	E29: Noboru#1 supplies Tea#1 to Ryuji#1
E14: Ryuji#1 gets to Tunnel#1	E30: Ryuji#1 drinks Tea#1
E15: Noboru#1 gets to Tunnel#1	
E16: Don#1 gets to Tunnel#1	

Figure 8: Input information

Table 2: A part of generated story world

	Store#1	Hill#1	Dry dock#1	...
t1	Ryuji#1, Sailor cap#1		Noboru#1, Don#1, Tea#1	
t2	Ryuji#1 (have: Sailor cap#1), Sailor cap#1		Noboru#1, Don#1, Tea#1	
t3		Ryuji#1 (have: Sailor cap#1), Sailor cap#1	Noboru#1, Don#1, Tea#1	
t4		Ryuji#1 (have: Sailor cap#1), Noboru#1, Don#1, Sailor cap#1	Tea#1	
t5		Ryuji#1, Noboru#1, Don#1 (have: Sailor cap#1), Sailor cap#1	Tea#1	
t6		Ryuji#1, Noboru#1 (have: Sailor cap#1), Don#1, Sailor cap#1	Tea#1	
t7		Ryuji#1 (have: Sailor cap#1), Noboru#1, Don#1, Sailor cap#1	Tea#1	
...				

E1: Ryuji#1 gets Sailor cap#1	E13: Ryuji#1 carries Noboru#1 from Tunnel#1 to Hill#2
E2: Ryuji#1 escorts Sailor cap#1 to Hill#1	E14: Don#1 leaves Tunnel#1
E3: Noboru#1 follows Don#1 from Dry dock#1 to Hill#1	E15: Ryuji#1 follows Noboru#1 from Hill#2 to Dry dock#1
E4: Don#1 buys Sailor cap#1	E16: Don#1 chases Ryuji#1
E5: Noboru#1 buys Sailor cap#1	E17: Ryuji#1 sits down in Dry dock#1
E6: Ryuji#1 buys Sailor cap#1	E18: Noboru#1 sits down in Dry dock#1
E7: Ryuji#1 escorts Noboru#1 to Sugita#1	E19: Don#1 sits down in Dry dock#1
E8: Don#1 strolls around Sugita#1	E20: Ryuji#1 supplies Sailor cap#1 to Noboru#1
E9: Ryuji#1 escorts Noboru#1 to Slope#1	E21: Noboru#1 parts with Sailor cap#1
E10: Don#1 advances to Slope#1	E22: Noboru#1 wins Tea#1 in "lottery"
E11: Ryuji#1 moves Noboru#1 from Slope#1 to Tunnel#1	E23: Noboru#1 supplies Tea#1 to Ryuji#1
E12: Noboru#1 posts Don#1 at Tunnel#1	E24: Ryuji#1 drinks Tea#1

Figure 9: A generated story line

In this circulative generation, we implemented a simple mutation function. It randomly changes the generated values for frames of agents and objects or values for case frames of event concepts to other values for agents, objects, and places in the story. In this time, we attempted the mutation probability, 0.1. Figure 10 shows an input story line and two story lines of 4 times and 40 times. As a result, in 4th generation, a scene in which the hunter and the priest bite each other is inserted and various events occur over the rice cakes. In 40th generation, objects in the story are only two rice cakes and a scene in which the big priest dies is omitted. The reason of outstanding reduction of the number of events and the simplification of generated stories are basically by that elements which are introduced in the simple mutation are only ones appeared in the generated stories already. We plan to attempt another circulative experiment by the arbitrary introducing of elements that do not exist in the already generated stories.

### Performance Checks

We show some preliminary checks of the system's basic performance using the result of "The legends of Tono" as mentioned above. First, we analyzed 20 samples which were generated from a same story line to investigate the correctness and the diversity of generated states in the story world. Second, for the story line generation, we analyzed 20 samples which were generated from a same story world to confirm the possibility and the diversity of events. In both experiments, we did not consider the naturalness of context and adopted whether each event simply physically occur as the evaluation criterion. For example, a sample story has a scene in which a big priest eats a burned stone. In spite of actual unnaturalness, this scene is physically possible. In addition, in many stories, physically impossible events and fantastical events also occur. A basic policy of our narrative generation system is the

generation of realistically possible events, and unrealistic and fantastical events are generated on the basis of the advanced rhetorical techniques such as the "defamiliarization" processing by the relaxation and adjustment of conceptual constraints using a conceptual dictionary (Zhang, Ono & Ogata, 2011, 2012). Therefore, the proposed system aims at a mechanism for generating realistic events and to the extent possible. As this preliminary check, we confirmed that the system could generate comparatively adequate and diverse results in both generation mechanisms except that few impossible events are generated in the story line processing. To decrease the impossible events, the elaboration of the conceptual dictionary is required.

Next, for circulative generation, we executed 15 cycles from a story line to analyze the degree of change in the generated story lines. As a result, we observed several types of changes such as the appearance of new event(s) and the disappearance of event(s). Table 3 shows the result of comparison between the 5th story line and the 15th one.

At last, we investigated the naturalness of story lines. The naturalness simply means that a story has not any realistic contradiction. We analyzed 10 story lines by a circulative generation process. As a result, we discovered some types of

Input story line	
E1: Hunter#1 burns Rice cake#1	E12: Hunter#1 burns Rice cake#4
E2: Hunter#1 burns Rice cake#2	E13: Hunter#1 burns Stone#1
E3: Hunter#1 burns Rice cake#3	E14: Hunter#1 acquires Stone#1
E4: Hunter#1 gets Rice cake#1 from Patrol	E15: Hunter#1 abandons Stone#1
E5: Big priest#1 advances to House#1	E16: Big priest#1 chases Hunter#1
E6: Big priest#1 gets Rice cake#1 from Omitted letter	E17: Big priest#1 gets Rice cake#4
E7: Big priest#1 eats Rice cake#2	E18: Big priest#1 eats Rice cake#4
E8: Hunter#1 acquires Rice cake#3	E19: Big priest#1 acquires Stone#1
E9: Big priest#1 buys Rice cake#3	E20: Big priest#1 gnaws on Kiln#1
E10: Hunter#1 bites Big priest#1	E21: Big priest#1 parts with Stone#1
E11: Big priest#1 runs in Mountain#1	E22: Big priest#1 slips away from House#1
	E23: Big priest#1 loses Big priest#1's life
4th generation	
E1: Hunter#1 burns Rice cake#1	E15: Hunter#1 acquires Stone#1
E2: Hunter#1 burns Rice cake#2	E16: Big priest#1 abandons Rice cake#4
E3: Hunter#1 burns Rice cake#3	E17: Big priest#1 brings Rice cake#4 into Mountain#1
E4: Hunter#1 wins Rice cake#1 in award	E18: Big priest#1 confiscates Rice cake#4
E5: Big priest#1 gets to House#1	E19: Hunter#1 transports Rice cake#4 to House#1 by dump truck
E6: Big priest#1 collects Rice cake#2	E20: Big priest#1 abandons Rice cake#4
E7: Big priest#1 chews Hunter#1	E21: Big priest#1 collects Rice cake#4
E8: Hunter#1 abandons Rice cake#1	E22: Hunter#1 damages Kiln#1
E9: Hunter#1 buys Rice cake#3	E23: Big priest#1 parts with Stone#1
E10: Big priest#1 confiscates Rice cake#3	E24: Hunter#1 takes Big priest#1 out of House#1
E11: Hunter#1 chews Big priest#1	E25: Big priest#1 dies of road game
E12: Hunter#1 burns Rice cake#4	
E13: Hunter#1 abandons Rice cake#3	
E14: Hunter#1 burns Stone#1	
40th generation	
E1: Big priest#1 strolls around House#1	E8: Hunter#1 drifts in House#1
E2: Hunter#1 chews Rice cake#1	E9: Hunter#1 carries Rice cake#1 to Mountain#1
E3: Hunter#1 escapes from House#1	E10: Hunter#1 brings Rice cake#1 in House#1
E4: Hunter#1 chews Rice cake#2	E11: Big priest#1 escapes from House#1
E5: Hunter#1 leaves Mountain#1	E12: Big priest#1 runs the whole House#1
E6: Hunter#1 escorts Big priest#1 to Mountain#1	E13: Big priest#1 abandons Rice cake#1
E7: Big priest#1 travels around House#1	

Figure 10: An example of circular generation

Table 3: The change of story lines through circulation

	Input story line	5th story line	15th story line
Appearing agent	Hunter and Big priest are half-and-half	Hunter is two-thirds	Hunter is mostly
Appearing object	Stone and kiln appear once	Stone and kiln appear more than once	Stone is mostly
Location of agent (Hunter)	Stay in the House from beginning to end	Do a round trip to House and Mountain many times	Do a round trip to House and Mountain many times
Location of agent (Big priest)	Do a round trip to House and Mountain two times	Do a round trip to House and Mountain one times	Do a round trip to House and Mountain one times
Contradiction of the story	none	none	Hunter moves to the same location many times
Number of event	23	21	19

contradictions. In an example, an agent moved to the location X from location Y and again moves to a different place from the X. Such contradiction is brought from the random choice of a location by the mutation function. It is necessary to adjust the mutation function so that contradictions do not arise completely or if a contradiction occurs, a kind of complement mechanism rewrites state(s) and event(s) after the story line or the story world were made. We prepared two methods for such ex-post solution. First method is “the rewrite of information” and second one is “the complement of information”. The former is a method that when a contradiction among some events occurs, the description of a state is changed to generate a new story. In the latter, when a contradiction arises, a new state is inserted into the existing states to generate a new story. In the former, an event’s content changes but the story line’s length does not change. On the other hand, in the latter, as a new state and a new event are inserted, the story line’s length gets longer. As another contradiction, a dead agent acts. For such case, we partially alter the mechanism so that a dead man or a disappearing object is not chosen as an agent or an object by changing the category to which the noun concept is belonged. For example, in the event like “Taro (dead-man) walks”, the constraint of agent in verb concept “walk” was originally the category of “human” except for “dead-man”. Accordingly, the system can alter the category to which Taro belongs from “human” to “dead-man”. Of course, for narratives and stories, such contradiction or unreality is certainly necessary and rather very important and essential element. Again, our policy is that we regard the knowledge and mechanism prepared based on the realism or physical possibility in a narrow sense as a standard and basic method, and introduce the processing of rhetorical knowledge and techniques to adjust the reference range in the conceptual dictionary.

## Conclusions

This paper reported the current version of state-event transformation system to transform story lines and story worlds each other using a knowledge base and a conceptual dictionary. Moreover, we extended the basic framework to a circulative generation mechanism with a simple mutation function. Through the preliminary performance checks, we confirmed that the transformation of story worlds and story lines is approximately logically adequate and the circular process produces the diversity of stories. The proposed system is a module in an integrated narrative generation system and the pilot version is already implemented. In the context of the narrative generation system, the proposed system plays roles for expanding the variation of discourse to be generated and limiting possible narrative elements at any given time in a narrative generation process.

## References

Akimoto, T., & Ogata, T. (2011). A consideration of the elements for narrative generation and a trial of integrated narrative generation system. *Proc. of the 7<sup>th</sup> NLPKE* (pp.369-377).  
 Bringsjord, S. and Ferrucci, D. A. (2000). *Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS, a Storytelling Machine*. Mahwah, NJ: Lawrence Erlbaum.  
 Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., & Kanzaki, K. (2009). Enhancing the Japanese WordNet,

*The 7<sup>th</sup> Workshop on Asian Language Resources, ACL-IJCNLP 2009* (pp.1-8).  
 Fillmore, C. J. & Baker, C. (2010). A frames approach to semantic analysis. In Heine, B. & Narrog, H. (Eds.). *The Oxford Handbook of Linguistic Analysis*. 313-339. Oxford University Press.  
 Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., Ooyama, Y., & Hayashi, T. (1999). *Goi-Taikei: a Japanese lexicon CDROM*, Tokyo: Iwanami Shoten. (in Japanese)  
 Kipper-Schuler, K., Dang, G. T., & Palmer, M. (2000). Class-based construction of a verb lexicon. *Proc. of AAAI-2000* (pp. 691-696).  
 Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania.  
 McIntyre, N. & Lapata, M. (2009). Learning to tell tales: A data-driven approach to story generation. *Proc. of the 47<sup>th</sup> Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP* (pp. 217-225).  
 Meehan, J. R. (1977). Tale-spin, an interactive program that writes stories, *Proc. of the 5<sup>th</sup> IJCAI*, 1(5), 91-98.  
 Mishima, Y. (1963). *The sailor who fell from grace with the sea*, Tokyo: Kodansya. (in Japanese)  
 Mueller, E. T. (1990). *Daydreaming in humans and machines: a computer model of the stream of thought*, Norwood, NJ: Ablex.  
 Nakashima, M., & Ogata, T. (2008). The structure of a story, *Proc. of the 22<sup>nd</sup> JSAL* (1C2-3). (in Japanese)  
 Noe, K. (2005). *Philosophy of narrative*, Tokyo: Iwanami Shoten. (in Japanese)  
 Ogata, T., Hori, K., & Ohsuga, S. (1994). Towards narrative text generation based on narrative techniques and strategies. *Proc. of International Federation for Information and Documentation* (pp.296-300).  
 Ogata, T., & Kanai, A. (2010). *An introduction of informatics of narratology: on thought and technology of narrative generation*. Tokyo: Gakubunsha. (in Japanese)  
 Ohara, K. H. (2008). Lexicon, grammar, and multilinguality in the Japanese FrameNet. *Proc. of the 6<sup>th</sup> International Conference on Language Resources and Evaluation* (pp. 3264-3268).  
 Oishi, K. & Ogata, T. (2011). Towards the development of conceptual dictionary for narrative generation system. *Proc. of the 7<sup>th</sup> NLPKE* (pp. 351-358).  
 Oishi, K. & Ogata, T. (2012). The development of conceptual dictionary for narrative generation system: The structure and functions. *Proc. of the 4<sup>th</sup> DIGTEL* (pp. 168-170).  
 Okada, N., & Endo, T. (1992). Story generation based on dynamics of the mind. *Computational Intelligence*, 8(1), 123-160.  
 Onodera, K., Oishi, K., & Ogata, T. (2011). The generation of event sequences in story based on states-actions transformation and conceptual system. *Proc. of the 28<sup>th</sup> JCIS* (P2-42). (in Japanese)  
 Prince, G. (1987). *A Dictionary of Narratology*. Lincoln, Nebraska: University of Nebraska Press.  
 Propp, V. (Пропп, В. Я.) (1969). *Морфология сказки*, Изд. 2е. Москва: Наука. (*Morphology of the Folktale*. Austin, Texas: University of Texas Press. 1968.)  
 Schank, R. C. (1975). *Conceptual Information Processing*. Amsterdam, New York: Elsevier Science.  
 Takeuchi, K. (2011). Construction of a verb thesaurus for language processing considering synonym and difference of verbs in the level of argument structure. *The Proc. of the 25<sup>th</sup> LCCII in JCIS* (25G-05). (in Japanese)  
 Yanagita, K. (1955). *The legends of Tono*, Tokyo: Kadokawa Shoten. (in Japanese)  
 Zhang, Y., Ono, J., & Ogata, T. (2011). An advertising rhetorical mechanism for single event combined with conceptual dictionary in narrative generation system, *Proc. of the 7<sup>th</sup> NLPKE* (pp. 340-343).

# Teaching the Perceptual Structure of Algebraic Expressions: Preliminary Findings from the Pushing Symbols Intervention

Erin Ottmar (erin.ottmar@richmond.edu)

David Landy (dlandy@richmond.edu)

Department of Psychology, 28 Westhampton Way  
University of Richmond, VA 23173 USA

Robert L. Goldstone (rgoldsto@indiana.edu)

1101 E. 10<sup>th</sup> St., Indiana University  
Bloomington, IN 47405 USA

## Abstract

We describe an intervention being developed by our research team, Pushing Symbols (PS). This intervention is designed to encourage learners to treat symbol systems as physical objects that move and change over time according to dynamic principles. We provide students with the opportunities to explore algebraic structure by physically manipulating and interacting with concrete and virtual symbolic systems that enforce rules through constraints on physical transformations.

Here we present an instantiation of this approach aimed at helping students learn the structure of algebraic notation in general, and in particular learn to simplify like terms. This instantiation combines colored symbol tiles with a new touchscreen software technology adapted from the commercial *Algebra Touch* software. We present preliminary findings from a study with 70 middle-school students who participated in the PS intervention over a three-hour period.

**Keywords:** Algebra education; learning; perception; mathematical cognition

## Introduction

The core conceptual content of algebra is extraordinarily simple: it is largely exhausted by the properties of addition and multiplication over the real numbers, such as commutativity, associativity, and distributivity, together with basic properties of functions and equivalence relations over the same structure. This formal simplicity belies the great difficulty students have in mastering basic algebra content (NAEP, 2011) — and especially the notation universally used to express algebraic claims (McNeil, 2008; Koedinger & Alibali, 2008).

One way to explain the difficulty of algebra is that unlike number cognition, algebraic reasoning does not seem to fit neatly into a core conceptual domain (Dehaene, 1997; Carey, 2009). Children may then face the challenge of assembling new cognitive tools appropriate to algebraic interactions. This task is made more challenging because typical instruction in basic algebraic notation is often brief and involves an emphasis on memorization of abstract rules.

Algebraic literacy—the fluent construction, interpretation, and manipulation of algebraic notations—involves not just memorizing rules, but also learning appropriate perceptual processes (Goldstone, Landy, & Son, 2010; Kirshner, 1989; Landy & Goldstone, 2007, 2008, 2010; Kellman, Massey, & Son, 2010). Like other formal diagrammatic systems (such as, for example, Venn diagrams) algebraic notation aligns the structure of the content domain with automatic perceptual properties and necessary physical laws (Cheng, 1999; Landy, Allen, and Anderson, 2011; Landy, 2010). In

this way reasoning that is properly cognitive can be accomplished by perceptual-motor systems such as attention (Patsenko & Altmann, 2010) or perceptual organization (Landy & Goldstone, 2007; Novick & Catley, 2008). Although such transformation of cognitive work into perceptual processing may carry distinctive risk of mistaking perceptual properties of representations for content principles (Novick & Catley, 2007; Kirshner & Awtry, 2004), it may also be critical to reducing cognitive load in complex operations (Sweller, 1994).

Successful students often use perceptual and visual patterns available in notations to solve mathematical problems. Like many skills learned from long practices learning algebra involves perceptual training—learning to *see* equations as structured objects (Landy and Goldstone, 2007; Kellman et al., 2008; Kirshner & Awtry, 2004). For instance, people seem to group symbols into perceptual chunks and use these groups, rather than just calculation rules, to perform mathematics. Although in some cases the appropriate perceptual patterns are fairly easy to see (Kirshner & Awtry, 2004), in other cases understanding the visual forms requires that a learner internalize an appropriate way of seeing a piece of notation. Real-world motion, changes, and transformations are naturally memorable and easy to acquire, making these processes natural tools for helping students grapple with algebra (Landy, 2010). Some successful object-centered transformations, however, may not be as immediately obvious as others in traditional instruction. Therefore, training students to *see* the structure of algebra may be a promising approach to teaching algebraic ideas.

While this perceptual-motor understanding of algebraic forms is a potentially rich and powerful source of student understanding, it also stands as a barrier to learning if visual patterning is not taught in a controlled manner. While some students learn easily, others latch on to incorrect perceptions and, consequently, generalizations (Marquis, 1988; Kirshner, 1989; Nogueira de Lima & Tall, 2007). Our goal is to find instructional and pedagogical paths through which students can make use of the strength of perceptual patterns in algebraic notation without falling prey to misleading visual structures or overly procedural, low-level understandings.

## Pushing Symbols: Teaching the Structure of Algebraic Expressions

The purpose of the PS intervention is to explore an alternative method of algebra instruction that focuses



Perceptual Process	Formal Transformation	Illustration
Rigid Motion	Commutation Transposition Rearrangements	$x+a = a+x$
Splitting	Distribution	$a(b+c) = (ab+ac)$
Joining	Factoring, Canceling	$t+2x-2x+q = t+q$
Symmetric Creation/ Destruction	Equation Transformation, Simplifying Fractions	$x-8=3$ $x-8+8=3+8$

Figure 1. Algebraic Transformation Visualizations

student efforts on the visual structure of formalisms, both by directly presenting those visual patterns and by challenging students to maintain and explain them. This method is being instantiated in a pedagogical intervention (PS) consisting of a set of in-class discussions, activities, and a dynamic computer-based visualization method. The intervention allows students to physically and dynamically interact with algebraic expression elements, providing a potentially powerful source of perceptual-motor experiences. Rather than simply rewriting different static expressions, in PS learners directly interact with expression objects and transform them using dynamical laws. Because rigid motion is a powerful perceptual grouping mechanism (Palmer, 1999) it is anticipated that training in which students see correct algebraic structures in dynamic transformations may lead to improved understanding of algebraic concepts.

The PS intervention has several specific aims. First we aim to increase fluency and accuracy by improving the alignment between students' visual-motor processes and proper formal operations and transformations. (Figure 1). Second the PS program is designed to be engaging for students, which is intended to build efficacy in students and develop the attitude that algebra can be intuitive, predictable, and even fun.

### Algebra Structure Tiles

The Pushing Symbols manipulative system uses colored magnets and tiles to decompose the structure of algebraic expressions. There are 4 different colored tiles in a set (see Figure 2), and each color represents a specific mathematical object (number, variable, coefficient, symbol). Yellow tiles represent numbers (from  $\pm 1-9$ ), blue tiles represent symbols or mathematical operations. (+), red tiles represent x variables and coefficients (from  $\pm 1-9$ ), and green tiles represent y variables and coefficients (from  $\pm 1-9$ ). After modeling an expression, the tiles can be rearranged and simplified into equivalent expressions.

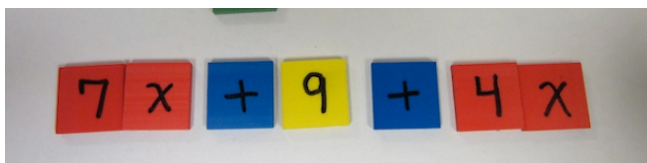


Figure 2: Algebra Structure Tiles

### The Algebra Touch Research (ATR) Software

The PS system uses a computer application developed in collaboration with Regular Berry software to teach students basic algebraic principles while richly engaging perceptual-motor systems (Figure 3). We describe software developed by Regular Berry software based on the *Algebra Touch* system, which instantiates the transformations specified by the PS intervention (We will call this *Algebra Touch: Research*, or *ATR*) In ATR students perform arithmetic functions by tapping on a sign and algebraic rearrangements are carried out by touching appropriate symbols and moving them into the desired location. ATR provides dynamical models of basic algebraic properties and transformations such as distributivity of multiplication over addition, commutativity, simplification of like terms through addition, and reduction of fractions to lowest terms.

ATR does not allow students to make mistakes; if they attempt to do something against the laws of mathematics, a brief side-to-side motion (a "shake") provides immediate feedback that their desired action was illegal. As a result, students immediately see how the rules result in legal transformations or manipulations in a way that is impossible with a traditional blackboard or overhead projector lesson.

Problems in ATR can be presented in either an untimed *list* mode or a *game* mode. In both modes the presentation and interaction with individual problems is identical. However, in the *game* mode problems are collected into *level*, and performance on any particular level is scored with a number of stars. Stars are based on the number of mistakes made during problem solution, and the speed with which a particular problem is solved. If too many mistakes are made or time runs out, the level is "failed" and must be restarted. At the end of each problem, the program provides immediate feedback to students about the number of errors they made and the speed to which they simplified the expression.

### Study Details

The PS approach has been instantiated in a single trial lesson covering combination of like terms. This lesson lasts approximately 90 minutes, and involves a large set of symbol tiles for teacher demonstrations on a whiteboard, smaller tiles used by students in pairs, and the ATR software.

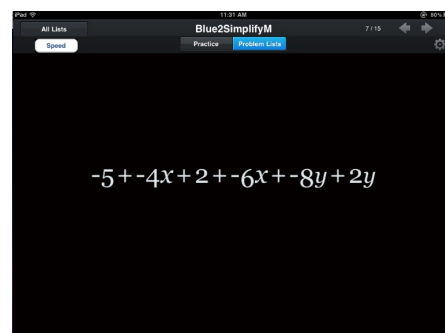


Figure 3: Algebra Touch Research software

We anticipated that the intervention would decrease the amount of structural errors that students made, and improve their overall understanding of simplifying expressions. Since the intervention did not explicitly address solving word problems, we did not anticipate a change in the number of word problems solved successfully. We also hypothesized that pre-test scores, self-efficacy, engagement, and performance on the iPad would positively contribute to post-test scores, while math anxiety would negatively contribute to post-test performance. We also predicted that the intervention would shift participation and engagement.

## Participants

Seventy eighth-grade students from an urban public middle school in the mid-east United States participated in this study during their regular mathematics instruction time. These students had never received instruction on like-terms or simplifying expressions before this intervention. Student assent and parental consent were obtained prior to participation in this study, in accordance with the directions of the University of Richmond Institutional Review Board.

## Study Procedures

The study took approximately 3 hours in total and occurred over three class periods. On the first day (90 minutes), students completed a pre-test on simplifying algebraic expressions and a Mathematics Self-Efficacy and Anxiety questionnaire. Next, students received a whole-group lesson on simplifying expressions. During this lesson, the teacher (the first author) led a series of discussions and used colored tiles to demonstrate algebraic structure. Students were then put into groups of 3 and used colored tiles to identify and combine like terms and simplify expressions. Third, students participated in a 20-minute exploration and training activity that provided students with an opportunity to learn how to use the iPad and *ATR* technology.

On the second day (90 minutes), students were each given an iPad, and were given 40 minutes to solve problems. Practice was divided into two phases. In the first phase, students simplified simple expressions involving no more than about 4 terms; in the second phase, more complex expressions involving up to 8 terms. Each 20 minutes phase was divided between an initial list of 10 untimed problems, followed by a set of 40 game problems.

Any pedagogical approach, especially those based on software interventions, must address the assistance dilemma (Aleven and Koedinger, 2002): how and how much help should be provided to learners, and when? *ATR* makes several fixed commitments: students cannot complete illegal transformations, for instance. In the current study, we also varied the amount of arithmetic support given to students. Participants were randomly assigned into 2 groups. In one group, students manually calculated the simple problems<sup>1</sup>,

<sup>1</sup> An example of manual and automatic calculation modes can be seen at <http://davidlandy.net/PushingSymbols/RPS--12-1-11-Like-Terms-Manual-1.mov> and <http://davidlandy.net/PushingSymbols/RPS--12-1-11-Like-Terms-Automatic-1.mov>

but arithmetic in structurally more complex problems was calculated automatically by the software; in the second condition, assistance pattern was reversed. There were no differences in structural understanding or success in word problems between the two groups, and, this manipulation will not be discussed further.

At the end of the intervention, students completed a questionnaire about their engagement during the intervention and a post-test. We also conducted student focus groups to receive feedback on what aspects of the intervention were most helpful and enjoyable. 2 weeks after the intervention, students completed a retention test.

## Measures

**Simplifying Expressions Assessments.** Each child completed an 18-item pre, post, and retention test on paper involving expression simplification. These tests assessed two major types of expression-related problem-solving skills: procedural facility with simplification (10-items), and expression construction and evaluation (word problems) (6 items). The problems on the pre, post, and retention tests were similar in form and difficulty.

We followed several steps to code the assessments. First, we coded each item on the assessment as incorrect, correct, or did not attempt. Next, to understand the source of the errors, we conducted error analyses on each item. Four error codes were used: 1) no error, 2) structural error; 3) addition or negative error; and 4) did not attempt. Structural errors include combining unlike terms, over-combination (simplifying the expression correctly and then combining un-like terms) or partial structural errors (moving around like terms but not completely simplifying the problem). Since the PS framework is designed to make structure concrete, naturally structural errors are particularly interesting for analysis. Addition and negative errors were coded when students used correct structure, but made an arithmetic error when combining terms. When a problem was left blank, we coded it as “did not attempt”. On average, students did not attempt to solve 25% of the pre-test problems, 16% of the problems on the post-test, and 20% on the retention test.

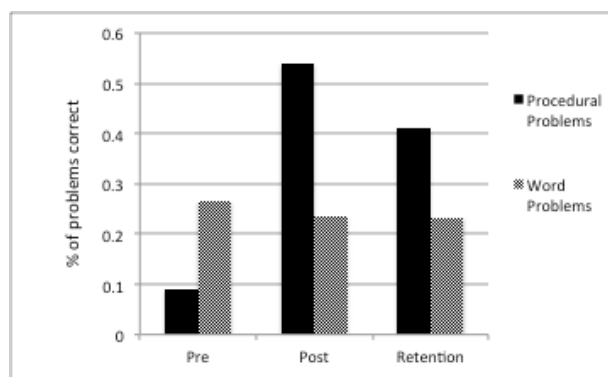


Figure 4: Proportion of Attempted Problems Solved Correctly (Free of structural errors)

Third, for each assessment (pre, post, and retention), we calculated 2 composite scores. 1) proportion of attempted procedural problems that were free of structural errors. 2) proportion of attempted word problems that were solved correctly. These two scores were used to measure student understanding of algebraic expressions in the analyses.

**ATR Performance.** iPad Performance was measured at 2 different levels, using the *Algebra Speed* game. Level 1 asked students to simplify a series of 36 *simple* expressions (ex.  $5+7+3$ ;  $x+2+6$ ). Level 2 asked students to simplify a series of 40 *complex* expressions (ex.  $7+2x+5x+4y+1+-2y$ ). Students could receive a maximum of 3 points for each problem solved. The points system accounted both the number of errors that they made and the speed to which they simplified the expression. At the end of each level students received a level performance score, which represented the total number of points received on the level. Total points on Level 1 (*simple*) and Level 2 (*complex*) were used as 2 measures of ATR performance.

**Mathematics Self-Efficacy and Anxiety Questionnaire.** Students were administered a set of 10-items pertaining to their self-efficacy and anxiety in mathematics. All 10 items were on a uniform 4-point scale (1=almost never, 2=sometimes, 3=most of the time, 4=almost all of the time). To assess students' math self-efficacy beliefs, 5 items were adapted from the Academic Efficacy subscale of the Patterns of Adaptive Learning Scales (Midgley et al., 2000) (e.g. "I know I can learn the skills taught in math this year") ( $\alpha=.82$ ). To measure students' feelings of math anxiety, 5 items were adapted from the Student Beliefs about Mathematics Survey (Kaya, 2008) (e.g. "I feel nervous when I do math because I think it's too hard") ( $\alpha=.61$ ). Scores for each construct were then averaged to create a mean math self-efficacy and mean mathematics anxiety composite.

**Student Engagement in Mathematics Questionnaire.** Student engagement during the lesson was measured using 18 items that were adapted from the Student Engagement in Mathematics Questionnaire (Kong, Wong, & Lam, 2003): (e.g. "Today I only paid attention in math when it was interesting."). All 18 items were on a 4-point scale (1=no, not at all true, 2=a little true, 3=often true, 4=yes, very true).

## Results

### Analysis 1: Does the Pushing Symbols Intervention improve student understanding of algebraic structure?

**Procedural Problems.** On average the intervention increased students' knowledge of algebraic structure (Figure 4). At pretest only 9.4% of problems were solved without structural errors. At post-test 54% of problems attempted were solved without structural errors (Improvement of 44.6%,  $t=10.48$ ,  $p<0.01$ ). At retention 41.4% of the problems were solved without structural errors (overall improvement of 32%,  $t=6.81$ ,  $p<0.01$ ). After 2 weeks students retained 72% of their structural learning.

**Word Problems.** As expected, the intervention did not appear to improve student understanding of word problems at post-test ( $t=-0.87$ ,  $p>0.05$ ) or retention ( $t=-0.07$ ,  $p>0.05$ ).

### Analysis 2: Relations between structural performance, efficacy, anxiety, engagement, and performance on ATR.

We conducted regression analyses to examine potential predictors of structural performance on the post-test. We included the following variables in the analysis: gender, math self-efficacy, math anxiety, engagement, pre-test performance, and iPad performance.

Correlations and descriptive statistics are reported in Table 1 and the regression results are presented in Table 2. Three main effects were found. First, results indicate that math efficacy was related to higher performance on the post-test (a 1 point increase in efficacy was related to a 1.27 point increase in performance). Second, successfully completing more problems (both simple and complex) on ATR was related to higher scores on the post-test. Further, students who reported being more engaged during the PS intervention performed higher on the post-test (for every 1 point increase in engagement, students performed 1.80 points higher on the post test). Interestingly, students' performance on the pre-test or levels of math anxiety did not predict performance at post-test.

Table 1: Means, Standard Deviations, and Correlations for Measures of Performance, Beliefs, and Engagement

Variable	Mean	SD	1	2	3	4	5	6	7	8	9	10
1. Performance on Post-test	5.40	3.70	-									
2. Gender	0.55	0.51	-0.10	-								
3. Math Self-Efficacy	2.95	0.58	0.29*	-0.12	-							
4. Math Anxiety	2.01	0.61	-0.27*	0.04	-0.37**	-						
5. Performance on Pre-test	3.09	0.58	0.45**	-0.25	0.19	-0.16	-					
6. AT Level 1- Simple expressions	0.94	1.57	0.30*	-0.21	0.02	-0.21	0.12	-				
7. AT Level 2- Complex expressions	63.36	37.85	0.14	-0.19	-0.14	0.12	0.01	0.14	-			
8. Math Engagement	51.44	46.50	0.47**	0.15	0.08	-0.32**	0.22	0.31*	-0.30**	-		
9. Scaffold Group	0.59	0.50	-0.10	-0.16	-0.06	0.18	-0.13	-0.15	.65**	-0.68**	-	
10. Performance on Retention test	4.14	4.02	0.69**	-0.09	0.01	-0.30*	0.25	0.25*	0.22	0.33**	-0.01	-



Table 2: Predictors of Algebraic Structure Performance on Post-Test

Variable	$\beta$	SE	t
Intercept	-7.04	3.76	-1.87
Gender	0.06	0.79	0.57
Math Self-Efficacy	0.21**	0.68	1.85
Math Anxiety	-0.07	0.69	-0.52
Performance on Pre-test	0.15	0.26	1.30
AT Level 1- Simple expressions	0.29**	0.01	2.73
AT Level 2- Complex expressions	0.44**	0.01	2.06
Math Engagement	0.30**	0.66	2.75
Scaffold Group	0.09	1.24	0.50

## Discussion

We have described an approach to algebra instruction that emphasizes perceptual and manual interactions with dynamically realized models of algebraic notation, as a vehicle for helping students become fluent with algebraic structure. Although our current results are quite preliminary and not experimental, they do demonstrate that a short intervention based on this framework may substantially improve student performance at simplifying expressions. Furthermore, this work adds to a small literature suggesting that touchscreen-based learning tools can successfully lead to student learning.

Although our results suggest that, on average, student performance increased substantially after receiving the intervention, not all students mastered the material. Many students still struggled with simplifying expressions or did not attempt many of the problems. It will be important to compare motion-based interventions such as this one with other methods of instruction in algebra notation in the future, to better understand the relative value of the AT system.

The current system contrasts with many popular algebra manipulative systems, such as Algebra Tiles and Hands-on Equations (Foster, 2007), in its emphasis on the structure of mathematical expressions rather than models of the concepts referred to by them. We certainly believe that connecting algebraic structure to relevant and intuitive examples has an important place in the teaching of algebra. However, given the clear demonstrations that students struggle to understand basic algebraic notation (Koedinger, Alibali, & Nathan, 2008), that closely connecting structure to content can impede learning (Kaminski, Sloutsky, and Heckler, 2006), and existing evidence linking teaching algebraic structure to improved student understanding of algebraic expressions (Banerjee & Subramaniam, 2011), we believe that there is good reason to pursue manipulative systems that expressly communicate algebraic structure through engaging perceptual and motor interactions.

The current findings also suggest that a hands-on approach to teaching the structure of algebra may benefit

students. Students reported that physically moving objects (the tiles and ATR) around helped them focus on the steps necessary to simplify expressions. There is also anecdotal evidence that students were applying the ideas of perceptual motion when solving problems on paper. We often observed students gesturing and moving the terms around with their fingers, as well as drawing lines or arrows to represent the legal moves and actions. These observations are consistent with research suggesting that gesturing or alternative ways to represent new ideas may improve student learning (Cook, Mitchell, & Goldin-Meadow, 2008). They also reported that this approach seemed to help them better understand previously taught concepts (such as commutative property, order of operations).

It is also worth noting that the intervention seemed to increase student interest, participation, and interactions. Both observational and student reported engagement during this intervention was high. Virtually all students reported that the intervention was engaging, and fun. In addition, virtually all students reported liking to solve algebra problems more in ATR than in more traditional approaches.

This study has several limitations that limit the conclusions that can be drawn from it. Although beyond the scope of this current study, future work utilizing a control group involving more traditional instruction and practice will better examine the efficacy of this intervention. It is also unclear how the learning from this intervention differs from learning that would occur from typical classroom instruction, and how such differences may impact learning of future topics (Schwartz and Black, 1996). The design of the study also does not allow us to tease apart which components of the intervention (classroom instruction, manipulatives, and/or practice on the iPad) are most useful in building student understanding. Although each of these components implemented the general framework and underlying cognitive principles, given the large current interest in technological interventions, it will be important in future work to distinguish the particular contributions of each of these components and their interaction.

The value of this research at its current stage lies in pointing the direction to a complex of ideas and practices that connect education, cognitive science, and interface design. As designed experiences become more ubiquitous and richly featured, it becomes increasingly possible to construct novel experiences that evoke abstract content in powerful, perceptually specific ways. The limit point of the approach we are pursuing is not just one in which problem solving is fun, game like, and perceptually powerful. Instead, this research represents a starting point toward a conception of formal learning in which the structures of mathematics are directly explorable—in which the abstract is rendered consistently concrete.

## Acknowledgments

This work was funded through a grant awarded by the Institute of Education Sciences, US Department of Education (Grant # R305A1100060).

The *Algebra Touch* software is owned and designed by Regular Berry software, and none of the authors is involved with the production or sale of *AT*. However, features of *AT* have been implemented in collaboration with the second author, with the express purpose of instantiating the *Pushing Symbols* framework.

## References

- Alibali, M. W. & Nathan, M. J. (2007). Teachers' gestures as a means of scaffolding students' understanding: Evidence from an early algebra lesson. In Goldman, R., Pea, R., Barron, B. J., and Derry, S. (Eds.) *Video Research in the Learning Sciences*. Mahwah, NJ: Erlbaum.
- Aleven, V., & Koedinger, K. R. (2002). An Effective Meta-cognitive Strategy: Learning by Doing and Explaining with a Computer-Based Cognitive Tutor. *Cognitive Science*, 26(2), 147-179.
- Banerjee, R., & Subramaniam, K. (2011). Evolution of a teaching approach for beginning algebra. *Educational Studies in Mathematics*.
- Bernardo, A. B., & Okagaki, L. (1994). Roles of symbolic knowledge and problem-information context in solving word problems. *Journal of Educational Psychology*, 86(2), 212-220.
- Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.
- Cheng, P.C.-H. Unlocking conceptual learning in mathematics and science with effective representational systems. *Computers and Education*, 33 (2-3), 109-130.
- Cook, S., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesticulating makes learning last. *Cognition*, 106(2), 1047-1058.
- Dehaene, S. (1997). *The number sense*. New York: Oxford University Press.
- Goldstone, R. L., Landy, D., & Son, J. Y. (2010). The education of perception. *Topics in Cognitive Science*, 2(2), 265-284.
- Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2006). Effects of concreteness on representation: An explanation for differential transfer. In R. Sun & N. Miyake (Eds.), *Proceedings of the XXVIII Annual Conference of the Cognitive Science Society*.
- Kellman, P. J., Massey, C. M., Roth, Z., Burke, T., Zucker, J., Saw, A., Aguero, K., & Wise, J. (2008). Perceptual learning and the technology of expertise: Studies in fraction learning and algebra. *Pragmatics & Cognition*, 16(2), 356-405.
- Kirshner, D. (1989). The visual syntax of algebra. *Journal for Research in Mathematics Education*, 20(3), 274-287.
- Kirshner, D., & Awtry, T. (2004). Visual salience of algebraic transformations. *Journal for Research in Mathematics Education*, 35(4), 224-257.
- Koedinger, K. R., Alibali, M. W. and Nathan, M. J. (2008). Trade-Offs Between Grounded and Abstract Representations: Evidence From Algebra Problem Solving. *Cognitive Science*, 32: 366-397.
- Kong, Q., Wong, N. & Lam, C. (2003). Student engagement in mathematics: Development of instrument and validation of construct. *Mathematics Education Research Journal*, 15(1), 4-21.
- Landy, Allen, & Anderson (2011). Conceptual discontinuity involves recycling old processes in new domains. *Behavioral and Brain Sciences*, 34, 136-137.
- Landy, D., & Goldstone, R. L. (2010). Proximity and Precedence in Arithmetic. *Quarterly Journal of Experimental Psychology*.
- Landy, D., & Goldstone, R. L. (2007). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(4), 720-733.
- Marquis, J. (1988). Common mistakes in algebra. In A. F. Coxford & A. P. Shulte (Eds.), *The ideas of algebra, K-12: National Council of Teachers of Mathematics Yearbook*. Reston, VA: National Council of Teachers of Mathematics.
- Martin, S. A., & Bassok, M. (2005). Effects of semantic cues on mathematical modeling: Evidence from word-problem solving and equation construction tasks. *Memory & Cognition*, 33(3), 471.
- McNeil, N. M. (2008). Limitations to teaching children 2 + 2 = 4: Typical arithmetic problems can hinder learning of mathematical equivalence. *Child Development*, 79., 1524-1537.
- Midgley, C., Maehr, M. L., Huda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., et al. (2000). *Manual for the Patterns of Adaptive Learning Scales*: University of Michigan.
- Nogueira de Lima, R. and Tall, D. (2008). Procedural embodiment and magic in linear equations. *Educational Studies in Mathematics*. 67 (1) 3-18.
- Novick, L. R., & Catley, K. M. (2007). Understanding phylogenies in biology: The influence of a Gestalt perceptual principle. *Journal of Experimental Psychology: Applied*, 13, 197-223.
- Palmer, S. E. (1999). *Vision Science: Photons to phenomenology*. Cambridge, MA: MIT Press.
- Patsenko, E. G., & Altmann, A. M. (2010). How playful is routine behavior? A selective-attention model of performance in the Tower of Hanoi. *Journal of Experimental Psychology: General*, 139, 95-116.
- Schwartz, D. L., & Black, J. B. (1996a). Analog imagery in mental model reasoning: Depictive models. *Cognitive Psychology*, 30(2), 154-219.
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning Instruction*, 4.

# Improving Representational Competence in Chemistry with Model-Based Feedback

Shamin Padalkar (shamin.padalkar@psych.ucsb.edu)

Mary Hegarty (mary.hegarty@psych.ucsb.edu)

Department of Psychological & Brain Sciences, University of California, Santa Barbara  
Santa Barbara, CA 93106 USA

## Abstract

Representational competence is an important component of learning Organic Chemistry. However, students are seen to be incompetent in translating from one kind of molecular diagram to another. An instructional method informed by spatial cognition research was designed and administered individually. The instruction involved having students check their solutions by attempting to match concrete models to their solution. The instruction helped students in the experimental group to identify their mistakes, understand the usefulness of concrete models and lead to large improvements in performance for the experimental group.

**Keywords:** concrete models; chemistry education; visualization.

## Introduction

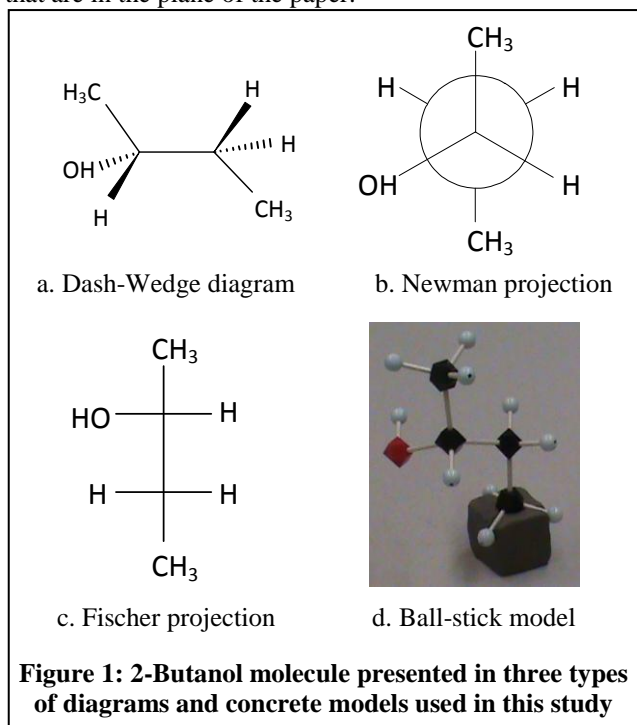
The literature in science education and chemistry education in particular, shows that interconnected cognitive skills, such as visualization, construction of mental models, model based reasoning, and representational competence are essential for acquiring mastery in the discipline (Kozma & Russell, 2005; Coll, 2006; Justi and Gilbert 2006; Treagust & Chittleborough, 2001). Kozma & Russell (2005) define 'representational competence' in the context of chemistry as 'a set of skills and practices that allow a person to reflectively use a variety of representations or visualizations, singly and together, to think about, communicate, and act on chemical phenomena in terms of underlying, a perceptual physical entities and processes'.

Representational competence is particularly important in organic chemistry. Organic chemists use several different representations of molecules, including different kinds of diagrams, models, and equations, for different purposes. For example, three kinds of diagrams are commonly used in organic chemistry and are introduced in the introductory college course on this topic. Mastering these diagrams is challenging, because they use different conventions to represent the three-dimensional (3-D) arrangement of atoms in the molecules in the two-dimensions of the printed page. They are also drawn from different orthogonal perspectives. This paper describes a study in which we examined students' ability to translate between these models, and tested an educational intervention that was designed to improve their representational competence using 3-D molecular models.

Examples of the three types of diagrams are given in Fig. 1 and their conventions and a brief description of each is given below.

## Diagrams Used in this Study

**Dash-Wedge Diagram** (Sometimes referred to, as perspective formula): In a Dash-Wedge diagram (Figure 1a), the molecule is oriented with the backbone carbons<sup>1</sup> at the two 4-way intersections of lines on the left and right of the diagram. Dashed lines represent bonds to atoms that are going into the page (below the plane of the paper). Wedge lines represent atoms that are coming out of the page (above the plane of the paper). Solid lines represent bonds to atoms that are in the plane of the paper.



**Figure 1: 2-Butanol molecule presented in three types of diagrams and concrete models used in this study**

**Newman Projections:** In a Newman projection (Figure 1b), the molecule is oriented with one backbone carbon in front of the other. The front carbon is located at the intersection of the 3 lines (noon, 4 o'clock and 8 o'clock around the circle). The substituents (atoms or groups of atoms) at the ends of these three lines are attached to the front carbon. The rear carbon is behind the circle. The substituents at the ends of the shorter lines connected to the circle (2 o'clock, 6 o'clock, and 10 o'clock around the circle) are attached to the rear carbon.

<sup>1</sup> Carbon backbone: longest series of covalently bonded carbon atoms in an organic compound.

**Fischer Projections:** In the Fischer projection (Figure 1c), the atoms at the right and left of the horizontal lines are coming out of the page (above the plane of the paper) and the atoms at the top and bottom of the vertical line are going into the page (below the plane of the paper). The two backbone carbons are located where the horizontal lines cross the vertical line. These carbons are in the plane of the paper.

The arrangement of atoms is of great importance in Chemistry, because even if the chemical formula is same, different arrangements of atoms result in different chemical properties. Dash-Wedge, Newman and Fischer diagrams serve different functions and hence chemists are often required to translate from one diagram to another. Ability to perform this representation translation task is also a measure of students' understanding of the 3-D structure of a molecule, as well as the conventions of a diagram, and prepares them for further problem solving. Therefore representation translation problems are included in typical assessments in organic chemistry classes.

Since it is difficult to visualize the 3-D structure of molecules from these diagrams, concrete, 3-D models (see Figure 1d) are sometimes used as pedagogic tool. A model represents the 3-D structure of the molecule directly, and therefore does not depend on remembering conventions for how the three dimensions are represented in a 2-D diagram. Furthermore, in translating between diagrams of molecules from different orientations, a student can rotate a physical model and observe the results, rather than having to perform difficult internal spatial transformations (mental rotation or perspective taking). This corresponds to what Kirsh (1997) referred to as a *complementary* action, that is, an action performed in the world that relieves the individual of the need to perform an internal computation. However chemistry instructors differ in their use of models. Some chemistry teachers use models while teaching and encourage their students to use them, but others rarely use models.

In recent studies (Stull, Hegarty, Dixon, & Stieff, submitted), undergraduate students were asked to translate between different kinds of diagrammatic representations of organic molecules and concrete models were made available to them. In different conditions across three experiments, students were encouraged to use the models and the correspondence between the models and diagrams was explicitly pointed out to the students. Students performed poorly on the representation translation task. When models were made available to them, many of the students did not use the models. However, those students who used the models performed significantly better on the diagram translation task. In conclusion, if models are used, they are extremely helpful in the translation task, but many students face a barrier to using them. Thus, just providing models is not enough; research is required to develop an appropriate instructional method for scaffolding the use of models.

## Exploratory study

In order to explore what strategies students use to solve the translation problems and how they interact with models, we first conducted a pilot study. Six undergraduate students were interviewed and asked to think aloud while solving six diagram translation problems. The students were familiar with the diagrams and their conventions. Most of the students used algorithms (rules) and/or internal visualization to solve the translation problems, rather than using models. Interestingly despite making many errors, these students were confident that they were performing the task correctly, which decreased their motivation for exploring the possibility of using models and improvement. That is, students had an illusion of understanding (cf. Rozenblit & Keil, 2002; Dunning, et. al., 2003). Specifically, they did not have clear understanding of the difference between stereoisomers and conformations<sup>2</sup>. Molecules that are stereoisomers have the same bond structure (in terms of which atoms are bonded with which other atoms) but different structures in terms of the relative locations of the substituents (atoms and groups of atoms) in 3-D space. An informal task in which they were asked to match the concrete model to their solution made students realize their mistake and to use the models effectively. From this study it was clear that (1) the participants need to know that they are making errors and what kind of errors they are making, and (2) they need to be guided to pay attention to the 3-D structure of the molecules as shown in the concrete models.

## Experimental Intervention Study

A short instruction, which required the participants use the model to check their solution, was designed and tested in an experiment. This instruction served two purposes: First it provided feedback to the participants. Second, it forced them to structurally align the model, therefore making them pay attention to the 3-D structure of the molecules, and revealed how the model could be used to help translate between the diagrams. We compared the accuracy of solutions of a group given this instruction (experimental group) to that of control group who performed the same representation translation tasks, but without the intervention. In addition to the accuracy of their solutions and demographic facts, we measured their spatial ability and general intelligence. Performance of the diagram translation task has been found to be correlated with spatial ability (Stull et al., submitted) but previous studies did not assess its relation to general intelligence. In addition, given that students are overconfident with their responses and many students do not spontaneously use models, we asked

<sup>2</sup> Stereoisomers have the same bond structure, but the different geometrical positioning of atoms and functional groups in space (e.g. switching the groups around one or more chiral (asymmetric) Carbon atom, which results into different chemical properties. However, rotation around C-C sigma bond results in a different conformation of the same molecules and it has the same chemical properties.

them to judge their levels of confidence and the usefulness of models before and after the interventions. We also videotaped the students while they performed the task and coded whether or not they used the models on each trial.

### Experimental Task

Students solved 18 problems (6 pre-test, 6 post-test and 6 transfer) in which they were provided one kind of diagram of a molecule (e.g., a dash-wedge diagram) and were asked to draw another kind of diagram (e.g. a Newman diagram) for the same molecule. The worksheet (8.5" x 11") included an instruction on the top and a diagram below it. Solution space on the work-sheet for the pre-test was divided into two equal parts by a horizontal line and participants were asked to draw their solution above the line. Post-test worksheets were not divided and participants were allowed to draw their solution wherever they wished.

### Research Design

The experiment followed a pre-test post-test design with control and experimental groups. Both experimental and control groups were first given basic instructions, which included the nature of the task, examples of three kinds of diagrams (see Figure 1) and their conventions (as described earlier). Participants were told that the instruction sheet would be kept, face down on the table and they could refer to it as necessary. They were also given a concrete (Ball & Stick) model<sup>3</sup> and reminded of the color codes for the different atoms in the models. The model was positioned in a clay stand. The experimenter demonstrated that the concrete model could be taken out of the stand and that it could be rotated in space and around the main carbon-carbon bond.

The pre-test consisted of six problems involving 4-Carbon molecules. It was followed by a short questionnaire on participants' level of confidence in their solutions and the usefulness of the concrete models. Then the experimental group went through a training intervention (described below) and the control group participants were given a 5-minute break. The post-test included a second set of six problems with 4-carbon molecules (enantiomers, or, mirror images) of the molecules in the pre-test problems and six 5-carbon problems. We refer to the set of 5-carbon problems as transfer problems, although they were very near transfer. The post-test was followed by a questionnaire which included questions about demographics and the same statements about confidence and usefulness of models as in the pre-test questionnaire. Finally, all participants completed the Vanderberg and Kuse (1978) Mental Rotation Test as a test of spatial ability (20 items administered in two 3 minute blocks) and Abstract Reasoning Test from the Differential

Aptitudes as a test of general reasoning ability (40 items, no time limit).

Participants were videotaped with their consent. The video camera was situated 2 feet above the table, usually on the left side of the participant. The experimenter sat to the left side of the participant, gave relevant instructions, provided relevant models and occasionally monitored the video camera. Participants saw only one model at a time, that is, the model of the molecule in the problem they were solving; the others were kept behind a screen. The assembly is shown in Figure 2.

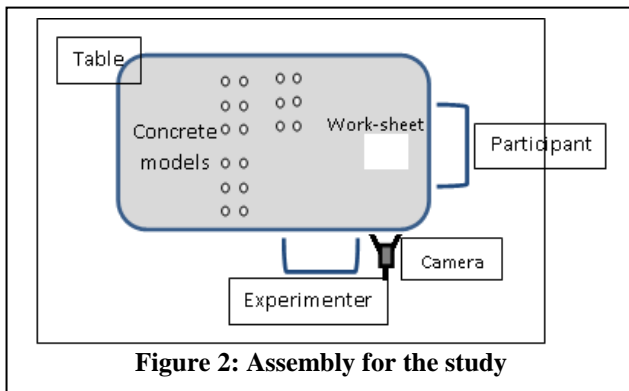


Figure 2: Assembly for the study

The intervention involved directions for participants to use models to check their own solutions and to draw correct solutions, if any of their solutions were found to be incorrect.

In the beginning of the intervention, participants were told "We are going to check the solutions". They were provided the concrete model for each problem and each problem was checked in three steps. Participants were provided help at each step if necessary. As a first step and were asked to match (i.e., structurally align) the model with the *given* diagram (cf., Gentner, 1983). This gave participants an opportunity to confirm that the given model indeed represented the given diagram and also gave practice in seeing the correspondence between model and diagram. In the second step of the intervention, participants were asked to align the model with their *solution* to the problem (which was drawn above the horizontal line). If the participant had drawn a correct solution, once s/he matched aligned the model with the solution, s/he was asked to move to the next problem. If the solution was incorrect, it would not be possible to structurally align the model with the solution. In this case, s/he realized that the model could not be matched and the solution was incorrect. The third step involved drawing a new corrected solution (below the horizontal line).

If the participant again drew an incorrect solution, Steps 2 and 3 were repeated. This was necessary for only 16 participants (on a total of 25 trials, i.e. an average of 1.56 trials per participant). If the participant drew an incorrect solution on the third attempt, he/she was told his/her mistake and was asked to go to the next problem. There was only one participant who could not draw the correct solution even after 3 cycles, and that was only on one trial.

<sup>3</sup> We used a 'Fundamental Organic Chemistry Set' manufactured by HGS Hinomoto Plastic Co., LTD (see Fig. 1d).



## Participants

The experimental group consisted of 30 participants (15 females) and the control group consisted of 24 participants (12 females), all undergraduate students at a research university. These students had completed at least one course in Organic Chemistry, in which they had been introduced to the three types of diagrams of organic molecules. The two groups did not differ in age (average = 20.3 years), spatial ability (average MRT scores = 35.67), general intelligence (average abstract reasoning test scores = 28.94), GPA (average = 3.15) or number of years in college (average = 3). The participants received course credit or \$20 for their participation.

## Coding of Diagrams

The data were coded in 2 ways:

**Number of correct solutions:** A score of '0' or '1' was assigned to each problem. A drawing had to be completely correct to receive a score of 1. The sum of correct solutions served as the total accuracy scores for the pre-test, post-test, and transfer problems.

**Level of accuracy:** Depending upon the type of error a level of 0 to 2.5 was assigned to each problem. In this scoring scheme, level 0 was assigned if a participant drew the wrong type of diagram or drew a diagram with missing or additional substituents. Level 1 was assigned when the diagram drawn was made up of the correct substituents, but these were incorrectly connected to the central carbon atoms. If the substituents were attached to the correct carbon atoms, but their 3-D spatial arrangement was incorrect, level 2 or 2.5 was assigned depending upon whether the mistake was made on both sides of the molecule (level 2) or only on one side (level 2.5). A fully correct diagram was assigned a level of 3. In addition to scoring the level of understanding for each problem, a student was assigned to a level of understanding (ranging from 0 to 3) if two-thirds of their solutions (4 of 6) were at or above this level of understanding for the pre-test, post-test, and transfer problems.

Data for 20 participants were coded independently by two researchers to establish the inter-rater reliability. Cohen's Kappa for the scores of pre-test, post-test and transfer problems together was 0.977.

## Coding of Model Use

Participants' use of the models was coded using the videos. Each trial was coded for whether or not participants moved the model in any way during each trial. Whenever participants moved the model it was coded as a use of the model. Pointing at the model (which happened extremely rarely) was not counted as using the model.

## Results

**Analysis of Correct Solutions:** Performance of the control and experimental groups on the pre-test, post-test and transfer problems is shown in Figure 3. The control group and experimental group had relatively poor performance on the pre-test, consistent with previous studies (Stull et al., submitted) and did not significantly differ on the pre-test,  $t(52) = 1.844$ ,  $p = .07$ . The experimental group performed significantly better after the intervention than before,  $t(29) = 9.344$ ,  $p < .001$ , and scored significantly better than the control group on the post-test,  $t(52) = 4.06$ ,  $p < .001$ . The average score for this group on the transfer problems is almost the same as for the post-test for the experimental group,  $t(29) = .162$ ,  $p = .87$ , indicating that what was learned from the intervention transferred to solving slightly more difficult problems.

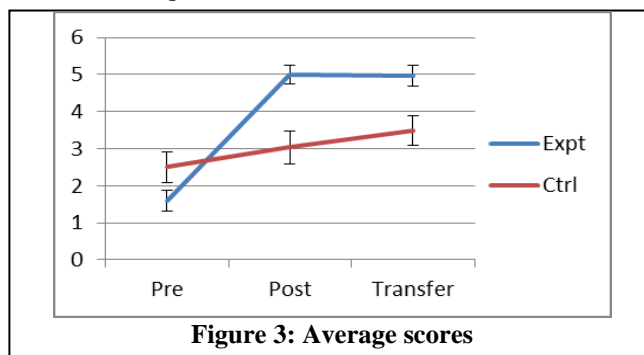


Figure 3: Average scores

The control group improved significantly from the pre-test to the post-test,  $t(23) = 2.18$ ,  $p < .05$ , and from the post-test to the transfer problems,  $t(23) = 2.41$ ,  $p < .05$ , thus making the difference between scores of the pre-test and transfer problems significant,  $t(23) = 3.72$ ,  $p = .001$ . Despite this, the experimental group outperformed the control group on the transfer problems,  $t(52) = 3.08$ ,  $p < .01$ .

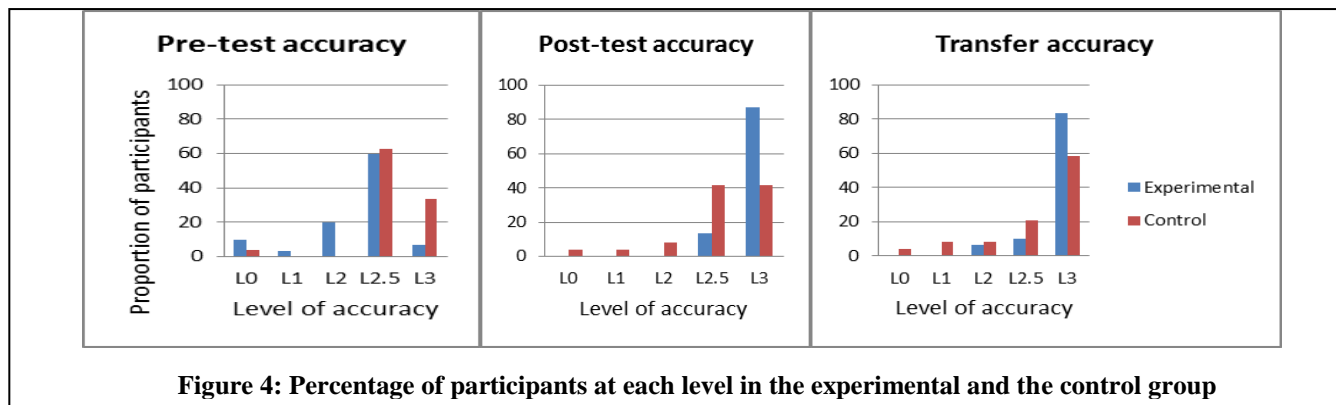


Figure 4: Percentage of participants at each level in the experimental and the control group

Thus, the intervention was successful and lead to very accurate performance (approximately 5 out of 6 problems solved correctly) on the post-test and transfer problems; although the control group spontaneously improved, they solved less than half of the post-test and transfer problems correctly.

**Analysis of Levels of Accuracy:** In the pre-test, the majority of students' drawings were at level 2.5 indicating that they understand the connectivity between the molecular substituents, but not their relative locations in 3-D space. In the post-test and transfer problems, the majority of students in the experimental group were at level 3, i.e., fully correct solutions. For the control group, the number of students performing at level 3 gradually increased (Figure 4). At the final (transfer) phase of the experiment 83% of participants in the experimental group and 58% of those in the control group were performing at Level 3.

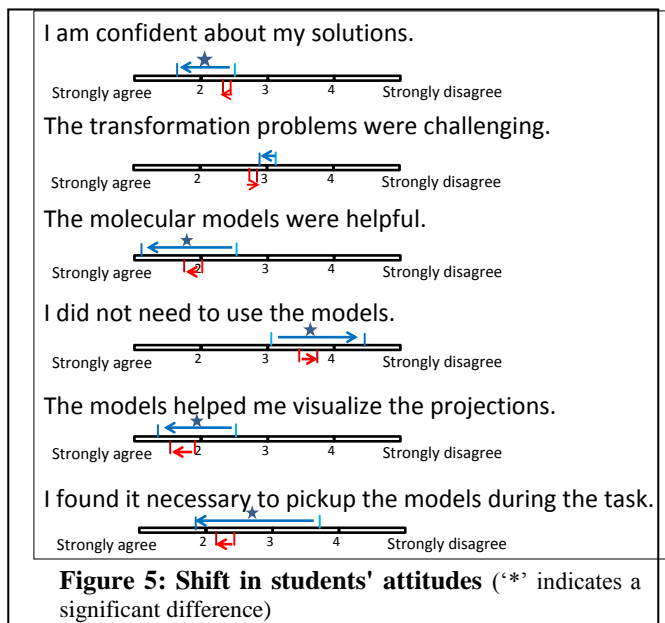
**Correlations:** As Table 1 shows, for the control group, scores on the pre-test, post-test and transfer correlated with each other. However this was not true for the experimental group. Participants in the experimental group performed well in post-test and transfer problems, regardless of their performance in the pre-test, which shows that the intervention was successful, irrespective of students' initial ability to do this task. As expected, scores on the pre-test and post-test correlated with scores on the mental rotation test (MRT) but this test was not significantly correlated with the transfer problems. The abstract reasoning test was correlated with performance only for the control group. Partial correlations of MRT with drawing performance, controlling for abstract reasoning, were significant for the pre-test (0.34\* pooling all participants). However, the correlations of MRT with the post-test and transfer problems were not significant. Thus, spatial ability is an important predictor of performance in the pre- and post-test but as the participants become familiar with the task, spatial ability becomes less important.

**Table 1: Correlations** ('\*' indicates a significant correlation)

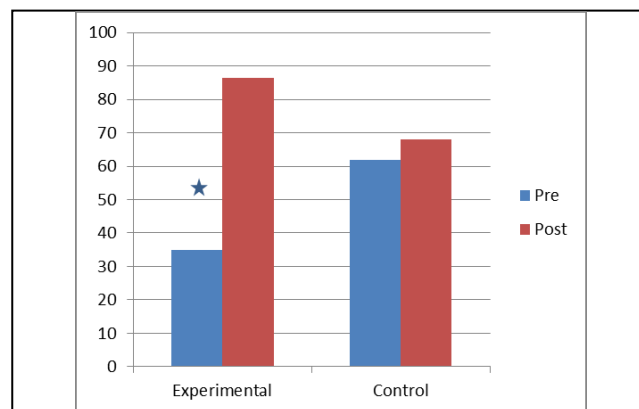
Scores		Pre-test	Post-test	Transfer
Post-test	Ctrl	0.832*		
	Expt	0.095	1	
Transfer	Ctrl	0.778*	0.902*	
	Expt	0.079	0.705*	1
MRT	Ctrl	0.480*	0.410*	0.275
	Expt	0.528*	0.361*	0.134
Abstract reasoning	Ctrl	0.519*	0.441*	0.410*
	Expt	0.238	0.311	0.130

**Students' perceptions about models and confidence level:** Students were given six statements (shown in Figure 5) to judge on scale of 1 to 5 after both the pre and post-test.

The arrows indicate the shift in their judgments from pre-test to post-test, with the experimental group shown in blue and the control group shown in red. Although overconfident in the pre-test, after the intervention, participants in the experimental group were significantly more and appropriately confident. They found models to be more helpful, and they recorded that they used the models more after the intervention.



**Participants' use of the models:** Participants in the experimental group moved the model on only 35% of trials during the pre-test but this percentage increased to 86% for the post-test and the transfer problems (Figure 6). On the other hand, participants in the control group moved the model on 62% trials in the pre-test and this percentage increased to 68% for the post-test. Thus the intervention was successful in inducing the experimental group to use the models. The tendency for the control group to use the model more in the pre-test appears to be due to sampling error, however it might also explain why these participants spontaneously improved on the problems, even without an intervention.



**Figure 6: Percentages of trials on which participants moved the models** ('\*' indicates a significant difference)



## Discussion

In summary, the intervention was successful. The accuracy of the experimental group increased from 27% to 83%. Although the control group underperformed the intervention group on the post-test, their accuracy steadily increased from 42% to 55% and their scores for the transfer problems were significantly higher than the pre-test scores. Spatial ability was an important predictor of performance on the pre and post-test, but it did not predict performance on the transfer problems, which suggests that spatial ability becomes less important with practice on the task.

Participants in the experimental group also showed an increase in confidence level, reported that they found models more useful, and reported that they used the models more often after the intervention. The latter result was validated by objective measures of their use of the models.

Analysis of students' levels of understanding (Figure 4) shows that the majority of participants were at level 2.5 in the pre-test which means that they switched the positions of two of the chemical groups around one of the central carbon atoms in the model. This mistake reflects a lack of understanding of the 3-D spatial relations between the chemical groups. There are two possible reasons for this error. First, participants may not realize the importance of the relative 3-D locations and hence draw a different molecule (an isomer of the correct molecule). Second they might understand the importance of the 3-D structure, but not be able to perform the required spatial transformation. The current experiment does not rule out either of these explanations, but it shows that once participants attempted to structurally align the models with the diagrams, they discovered their error and this in turn lead to increased use of the models and better performance in the post-tests.

If students do not understand the importance of the 3-D spatial relations, then giving them constructive feedback, that they drew an isomer, rather than the correct molecule, should be sufficient to improve performance. If they understand the importance of the 3-D relations, but are unable to perform the required spatial transformations, showing them how models map onto the diagrams should help them perform the correct spatial transformations. To identify which of these two reasons played an important role in students' inability to perform the diagram translation task, we are currently conducting a second experiment which compares a 'feedback' condition and 'model match' condition.

In any case, the current paper documents an intervention that was certainly useful. It took a short time (an average of 17 minutes to check the 6 pre-test problems) and could be accommodated in a laboratory or tutorial session in the context of an organic chemistry class. Further studies will be necessary to examine whether this type of intervention can lead to lasting gains in student performance, and whether the intervention leads to an understanding of the structure of molecules that can improve performance in a situation where students do not have access to models. Although experts often use more abstract rule-based

strategies to translate between diagrams (Stieff, 2007) understanding the 3-D structure of molecules is central to organic chemistry knowledge, and models appear to be an important stepping stone to reaching higher levels of understanding. Also, since models, in contrast to diagrams, are powerful representations having the unique quality of three-dimensionality, they have been an important tool in cutting edge research and hence students should be familiar with strengths and weaknesses of models through their own experience. The general approach of the intervention can be adopted and tested in the instructions of other disciplines such as geology, astronomy, architecture etc. in which three-dimensional structure and dynamic properties of the system are very important.

## Acknowledgments

This research was funded by National Science Foundation Grant DRL 1008650. We are thankful to Andrew Stull, Trevor Barrett, David Sanosa, Emily Steiner, Bonnie Dixon, Mike Stieff for their contributions to the research.

## References

- Coll, R. (2006) The role of models, mental models and analogies in chemistry teaching. In P. Aubusson, A. G. Harrison, S. Ritchie (Eds.) *Metaphor and analogy in science education*. Springer.
- Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, 12(3): 83-87. doi:10.1111/1467-8721.01235.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.
- Justi, R. and Gilbert, J. (2006) The role of analog models in the understanding of the nature of models in Chemistry. . In P. Aubusson, A. G. Harrison, S. Ritchie (Eds.) *Metaphor and analogy in science education*. Springer.
- Kirsh, D. (1997). Interactivity and multimedia interfaces. *Instructional Science*, 25, 79-96.
- Kozma, R. & Russell, J. (2005) Multimedia learning of chemistry. In R. Mayer (Ed.), *Cambridge Handbook of Multimedia Learning*, 409-428. New York: Cambridge University Press.
- Rozenblit, L. & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26, 521-562.
- Stieff, M. (2007). Mental rotation and diagrammatic reasoning in science. *Learning and Instruction*, 17, 219-234.
- Stull, A. T. Hegarty, M., Dixon, B., Stieff, M. (submitted) Representational Translation with Concrete Models.
- Treagust, D. F., Chittleborough, G. (2001), Chemistry: A matter of understanding representations, in J. Brophy (Ed.) *Subject-specific instructional methods and activities (Advances in Research on Teaching, Volume 8)*, Emerald Group Publishing Limited.

# Cognitive Biases in a Geospatial Intelligence Analysis Task: An ACT-R Model

Jaehyon Paik (jpaik@parc.com), Peter L. Pirolli (pirolli@parc.com)

Palo Alto Research Center  
3333 Coyote Hill Rd., Palo Alto, CA 94304

Christian Lebiere (cl@andrew.cmu.edu), Matthew Rutledge-Taylor (mattrt@andrew.cmu.edu)

Carnegie Mellon University, Psychology Department  
5000 Forbes Avenue, Pittsburgh, PA 15218

## Abstract

An ACT-R model of sensemaking in a geospatial intelligence task was developed based on Instance-Based Learning Theory (IBLT). The model (a) maintains hypotheses about the probability of attacks by insurgent groups, (b) seeks new information based on those hypotheses, and (c) updates hypotheses based on new evidence. The model provides a functional account of how these sensemaking processes are carried out in a cognitive architecture, and model performance can be compared to normative (Bayesian) standards. Simulations exhibit two well-known cognitive biases that are frequently identified as problems in intelligence analysis: (1) anchoring in the weighting of new evidence and (2) confirmation bias in seeking new information.

**Keywords:** ACT-R, cognitive biases, sensemaking

## Introduction

Sensemaking (Klein, Moon, & Hoffman, 2006a, 2006b; Pirolli & Card, 2005; Russell, Stefik, Pirolli, & Card, 1993) is a concept that has been used frequently in studies of intelligence analysis. The term suggests an active seeking and processing of information to achieve understanding. Sensemaking involves a set of processes aimed at seeking and filtering information, plus a set of processes that develop representational schemas (frames) that best fit the available evidence and provide a basis for understanding the data. In this paper we present the cognitive model of basic sensemaking processes for an intelligence analysis task. A major concern in the intelligence community is the impact of cognitive biases on the accuracy of analyses (Heuer, 1999). We present simulation results that exhibit *anchoring bias* in the evaluation of new evidence and *confirmation bias* in seeking evidence.

## The Geospatial Task

The geospatial task (Figure 1) is one of a set of challenge tasks developed as part of the IARPA ICARUS program to drive the development of integrated neurocognitive models of sensemaking. This specific task required reasoning based on a set of rules concerning the relation of observed evidence to the likelihood of attack by four different groups. A layered geospatial map is presented on a computer screen, with different layers presenting different forms of intelligence (INTs). The INTs include HUMINT (human intelligence), IMINT (image intelligence), MOVINT (movement intelligence), SIGINT (signal intelligence),

SOCINT (socio-cultural intelligence), and SIGACT (attack intelligence).

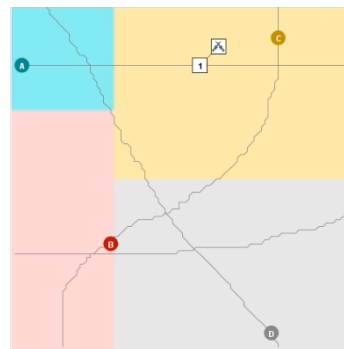


Figure 1: The screen shot of the geospatial task. The letters (A, B, C, D) indicate the center of the group location, and '1' surrounded by a box indicates the attack location.

The task begins with a given attack location (SIGACT) along with group centers (HUMINT), A, B, C, and D representing the center of activity for four possible insurgent groups. The first step is to report probabilities of attack by each group [A%, B%, C%, D%] based on the SIGACT and HUMINT (see Table 1)<sup>1</sup>. After that, the task is to iteratively choose among the four remaining INT layers (Table 1), up to a total of three INTs (layers), one at a time, in any order. Each INT layer provides unique evidence. Specifically, IMINT can reveal whether an attack happened on a government or military building, MOVINT provides evidence whether an attack occurred in dense or sparse traffic, SIGINT indicates electronic “chatter” or “silence” by different groups, and SOCINT indicates the group whose region the attack happened. At each stage, the selection of a particular INT provides evidence that can be used to update the probability distribution over the hypotheses about the responsibility of the four groups in producing the given attack. The rules specifying how evidence ought to update these probabilities is given in the PROBS rules in Table 1. After the last stage of INT selection, the task is to allocate resources (troops) to prevent further attacks.

<sup>1</sup> The new version of the task will provide the initial probabilities based on HUMINT

Table 1: Probabilistic rules provided to user for inferring beliefs about group attack likelihoods.

INTS	PROBS
HUMINT	If a group attacks, then the relative likelihood of attack decreases as the distance from the group center increases.
IMINT	If A or B attack then the attack is four times as likely to occur on a <i>Government</i> versus <i>Military</i> building. If C or D attack then vice versa.
MOVINT	If A or C attack then the attack is four times as likely to occur in <i>dense</i> versus <i>sparse</i> traffic. If B or D attack then vice versa.
SIGINT	If SIGINT on a group reports <i>chatter</i> , then attack by that group is seven times as likely as attack by each other group If SIGINT on a group reports <i>silence</i> , then attack by that group is one-third as likely as attack by each other group.
SOCINT	If a group attacks then that group is twice as likely to attack in its own versus other region.

### Anchoring and Confirmation Biases

Anchoring and confirmation biases have a long history of study in cognitive psychology and the intelligence communities (Heuer Jr, 1999; Klayman, 1995; Klayman & Ha, 1987; Nickerson, 1998; Tversky & Kahneman, 1974; Wason, 1960). Process models of these biases, especially in complex tasks, remain largely unexplored. In this paper we develop cognitively plausible process model of the geospatial task in the ACT-R architecture. We then compare this ACT-R model against a rational Bayesian model of the task to examine evidence of anchoring and confirmation biases.

#### Anchoring Bias and Anchoring and Adjustment Heuristic

Anchoring is a cognitive bias that occurs when individuals establish some belief based on some initial evidence, and then overly rely on this initial decision in their weighting of new evidence (Tversky & Kahneman, 1974). Human beings tend to anchor on some estimate or hypothesis and subsequent estimates tend to be adjustments that are influenced by the initial anchor point—they tend to behave as if they have an *anchoring+adjustment heuristic*. Adjustments tend to be insufficient in the sense that they overweight the initial estimates and underweight new evidence.

#### Confirmation Bias

Confirmation bias is typically defined as (for a survey, see Nickerson, 1998):

- The interpretation of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand (Nickerson, 1998)
- The tendency for people to seek information and cues that confirm the tentatively held hypothesis or belief,

and not seek (or discount) those that support an opposite conclusion or belief (Wickens & Hollands, 2000). The seeking of information considered supportive of favored beliefs (Nickerson, 1998).

Studies (Cheikes, Brown, Lehner, & Adelman, 2004; Convertino, Billman, Pirolli, Massar, & Shrager, 2008; Tolcott, Marvin, & Lehner, 1989) have found evidence of confirmation bias in tasks involving intelligence analysis, and there is a common assumption that many intelligence failures are the result of confirmation bias in particular (Chorev, 1996; Grabo & Goldman, 2004; Heuer Jr, 1999).

#### Biases in the Geospatial Task

The geospatial task might elicit anchoring and confirmation biases at multiple points in the process. Anchoring bias in weighing evidence might be found when participants revise their belief probabilities after selecting and interpreting a particular INT. The estimates of belief probabilities that were set prior to the new INT evidence could act as an anchor, and the revised (posterior) belief probabilities could be insufficiently adjusted to reflect the new INT (i.e., when compared to some normative standard).

Confirmation bias in weighing evidence can also be found in the hypothesis adjustment process. When applying a particular INT, such as IMINT (which supports multiple hypotheses), participants may only apply the adjustment to the preferred hypothesis while neglecting other groups also supported by evidence, or weight the evidence too strongly in favor of the preferred hypothesis.

Finding confirmation bias in seeking evidence in the task is somewhat more difficult since most INTS apply equally to all hypotheses. We used the SIGINT layer to identify this kind of bias because a single hypothesis has to be selected for that layer. SIGINT provides considerable gains to the selected hypothesis when chatter is detected (7 times more likely), so participants could get significant certainty. However, it loses considerable weight (3 times less likely) when silence is detected. Thus, a decision to choose the SIGINT layer too early (before a specific group has dominates the other in terms of relative likelihood) might be interpreted as confirmation bias in evidence seeking.

#### The ACT-R architecture

ACT-R (Anderson et al., 2004; Anderson & Lebiere, 1998) is a cognitive architecture that includes a declarative memory module that stores and retrieves information and a procedural module that coordinates the flow of information. Declarative knowledge in ACT-R is represented formally as *chunks* of information (Miller, 1956; Simon, 1974). Chunks are recalled from long-term declarative memory by an activation based retrieval process. Activation spreads from the current focus of attention, including goals, through *associations* among chunks in declarative memory. The spread of activation from one cognitive structure to another is determined by attentional weights on the associations among chunks. These weights determine the rate of activation flow among chunks. *Partial matching* is a

mechanism that allows for chunks in declarative memory that do not perfectly match a retrieval request to be retrieved. *Blending* is a memory retrieval mechanism that allows all chunks in declarative memory that match or partially match a retrieval request to blend together to create a new chunk representing an aggregate response (Lebiere, 1999).

Production rules are used to represent procedural knowledge in ACT-R. That is, they specify how to apply cognitive skill (know-how) in the current context, and how to retrieve and modify information in other modules. In ACT-R, each production rule has conditions that specify structures that are matched in limited-capacity buffers corresponding to information from the external world or other internal modules. Each production rule has actions that specify changes to be made to the buffers or requested functions in the associated modules.

### The Rational Model versus ACT-R Model

We developed an ACT-R model to perform the geospatial task, as well as a rational (Bayesian) model as a normative benchmark. The ACT-R model implemented a version of instance-based learning theory (Gonzalez, Lerch, & Lebiere, 2003), and the rational model employs a standard Bayesian approach for updating belief and selecting new evidence (INTs) based on an expected information gain metric.

### The Rational Model

From the PROBS rules discussed in table 1, we can extract specifications of the likelihoods of evidence,  $P(e|h)$ , where  $e$  is evidence (e.g., “chatter”) and  $h$  is a hypothesis (“group A attacks”). Bayes rule can be applied to compute the posterior likelihood of  $h$  given specific evidence  $e$  and prior probabilities  $P(h)$

$$P(h|e) = \frac{P(e|h)P(h)}{\sum_i P(e|i)P(i)}$$

where  $i$  iterates over all hypotheses.

For instance, in Figure 2, we assume some HUMINT data has been processed, a probability has been assigned to each of the hypotheses, and the goal is to evaluate the choice of an IMINT layer. The outcomes represent the estimates of government and military building attacks given the current hypotheses strengths. The posteriors are the updated probability distributions according to the outcomes.

The choice of INT layers can be evaluated by their effects on *expected information gain* (Austerweil & Griffiths, 2011). Information gain is defined as the reduction in entropy measured over the hypothesis probabilities that occur by acquiring additional evidence. Information gain is specified as

$$IG(D, e) = H(D) - H(D|e)$$

where  $H(D)$  is the entropy of the distribution of probabilities over hypotheses, and  $H(D|e)$  is the entropy of the distribution of posterior probabilities after some evidence  $e$  has been discovered.

$$H(D|e) = - \sum_i P(D|e) \log_2 P(D|e)$$

In Figure 2, the information gain for seeing an attack on a government building is .4 and an attack on a military building is .09. The *expected information gain* is calculated by weighting each of the possible outcomes of information gain by the probability of obtaining that outcome. Thus, the expected information gain for selecting the IMINT layer is  $(.56)(.4) + (.44)(.09) = .26$

Our rational model computed the expected information gain for all layers at each stage. The rational choices were compared to the selections made by ACT-R to identify biases.

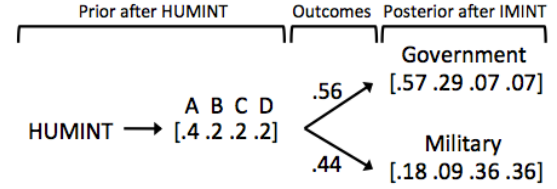


Figure 2: An example of the rational Bayesian hypothesis estimates for an IMINT layer selection.

### The ACT-R Model

We assume that an average person is not able to compute the expected information gain of all possible layers, because it involves substantial amounts of computation. We considered two cognitively plausible alternatives to develop an ACT-R model.

- *Difference reduction heuristics.* One cognitively plausible way to reduce complexity is to assume that people use a heuristic such as hill climbing to evaluate moves. Rather than focus on maximizing expected information gain, hill-climbing analysis could focus on achieving states that are closer to an ideal goal state (i.e., in this case, a state in which the attacks are unambiguously caused by Group A, or Group B, etc.). This would require some heuristic for evaluating differences (distances) from the goal state.
- *Memory-based move evaluation.* It is well known in the field of naturalistic decision making that experts invariably rely on vast amounts of declarative memory experience and well-practiced cognitive skill (Klein, 1998). We assume that participants store move outcomes in declarative memory, and that blended retrievals based on current states and possible moves can produce a blended retrieval of outcomes to those moves. This would be a weighted smoothing of gains that had been made by similar moves in the past. Although not precisely equivalent to the computation of rational expected information gains (a weighting over the gains achieved by possible layer selection outcomes), blending over memory of past INT outcomes and gains should produce similar effects.

Our ACT-R model of the geospatial task explored some plausible difference reduction heuristics in a memory-based move evaluation framework. The following weighted distance function assumes that the goal is to achieve certainty on one of the hypotheses (i.e.,  $p_i = 1$ ).

$$\sum_{i \in \text{hypotheses}} p_i(1 - p_i)$$

We assume that the model relies on the use of declarative chunks that represent hypothetical past experiences of selecting INT layers. This is intended to capture a hypothesized learning process whereby participants have attended to a current probability distribution, chosen a layer, revised their estimates of the hypotheses, and assessed the utility of the layer selection they just made. For instance, if a participant had experienced two situations in which they had assessed a probability distribution [.4 .2 .2 .2] and selected an IMINT layer, and had experienced a “government building” attack one time and a “military building” attack a second time (See figure 2). The model assumes the two chunks in its declarative memory.

(exp1	(exp2
isa layer-choice	isa layer-choice
prior-a 0.4	prior-a 0.4
prior-b 0.2	prior-b 0.2
prior-c 0.2	prior-c 0.2
prior-d 0.2	prior-d 0.2
layer IMINT	layer IMINT
outcomes government	outcomes military
utility 0.58)	utility 0.69)

where the utilities are computed by the weighted distance metric.

At a future layer selection point, a production rule will request a blended/partial matching retrieval from declarative memory like below:

```
+blending>
  isa layer-choice
  prior-a 0.45
  prior-b 0.15
  prior-c 0.15
  prior-d 0.25
  layer IMINT
  utility =utility
```

This retrieval will partially match against the experience chunks above, and will blend across the stored utilities for all experienced IMINT outcomes (i.e., both government and military building experiences in the past) to produce a kind of “expected” utility to match the =utility request.

### Hypothesis Probability Updating

Lebiere (1999) proposed a model of cognitive arithmetic that used retrieval of arithmetic facts to generate estimates of answers without explicit computations. The cognitive arithmetic model uses partial matching to retrieve facts related to the problem, and uses the blending mechanism to merge them together to issue an aggregate estimated answer. The model reproduced a number of characteristics of the distribution of errors in elementary school children, including both table and non-table errors, error gradients around the correct answer, higher correct percentage for tie

problems, and, most relevant here, a skew toward underestimating answers, as is common in anchoring and adjustment processes.

This approach was leveraged in the current model to account for how the PROBS rules (from table 1) are interpreted and applied to estimate the effects of the rules on the relative probabilities that the groups are responsible for the attack under examination. The ACT-R model’s memory was populated with a range of facts consisting of triplets: an initial probability, an adjustment factor, and the resulting probability. These chunks are derived from the PROBS rules shown in Table 1. For example, if the attack is found to occur on of road with dense traffic, the MOVINT rule specifies that groups A and C are 4 times as likely to have been responsible. When a layer of information is made available to the model, it adjusts the current set of probabilities by retrieving the relevant chunks and replacing the prior probabilities with the posteriors representing in the retrieved chunks. The results of this chunk based rule interpretation were then averaged over a thousand runs, given the variations in answers resulting from activation noise in the retrieval process. When provided with ratio similarities between probabilities (and factors), the primary effect is an underestimation of the adjusted probability for much of the probability range.

### Assessment

Biases can be defined as deviations from some norm (Jonathan D. Nelson, 2005; J.D. Nelson, McKenzie, Cottrell, & Sejnowski, 2010). In conjunction with producing the geospatial challenge tasks, the IARPA ICaRUS program has developed metrics for assessing cognitive biases. Anchoring bias or confirmation bias in weighing evidence is assessed by a *negative entropy* metric,  $N$  and confirmation bias in seeking information is assessed using a task-specific *confirmation metric*,  $C$ .

### Anchoring bias metric

Negative entropy is defined as

$$N = (H_{max} - H)/H_{max}$$

where  $H$  is the entropy of the distribution of probabilities over hypotheses and  $H_{max}$  is the maximum possible entropy.  $N$  increases with the certainty in a hypothesis (i.e., the “peakiness” of the distribution). At a given stage of updating belief probabilities [A%, B%, C%, D%] given some new INT evidence, we may assess the negative entropy,  $N_{ACT-R}$ , of the belief probabilities in ACT-R, and the negative entropy of the rational model,  $N_{Rational}$ . If  $N_{ACT-R} > N_{Rational}$  then the ACT-R model is exhibiting a confirmation bias in weighing evidence – i.e., over-weighting evidence that confirms the most likely hypothesis. Conversely, if  $N_{ACT-R} < N_{Rational}$  then the ACT-R model is exhibiting the anchoring bias.

### Confirmation bias metric

Confirmation bias in seeking evidence, is assessed by the fraction,  $C$ , of SIGINT choices requested about the

insurgent group that has been assigned highest probability of being the attackers.

$$C = \frac{\text{No. of SIGINT choices on the highest Prob. group}}{\text{Total no. of SIGINT choices}}$$

SIGINT provides considerable weight when “chatter” is detected, so selection of SIGINT for the highest probability group is interpreted as being confirmatory. It is assumed that if  $C > .5$  then the model exhibits confirmation bias in seeking evidence (random choice strategy be  $C = .25$ ).

## Results and Discussion

Each model was used to simulate 30,000 layer selections in 10,000 tasks. By using metrics that we explained in the previous section, we could identify that the ACT-R model exhibits anchoring and confirmation biases while conducting the task.

### Anchoring bias in weighing evidence

In the geospatial task, the ACT-R model revises its probability distribution over hypotheses after each layer selection, and this can be compared against the probability distribution of the rational model. As can be seen in figure 3, the ACT-R model is most often showing lower negative entropy than the rational model ( $N_{\text{ACT-R}} < N_{\text{Rational}}$ ). In other words, rather than showing a confirmation bias it is exhibiting a form of anchoring bias.

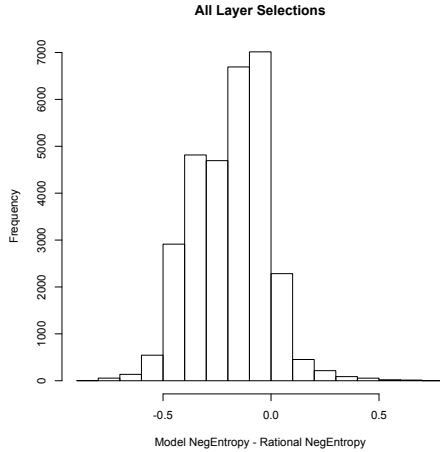


Figure 3: Difference negative entropy between the ACT-R model and rational model after each layer selections.

### Confirmation bias in seeking evidence

We analyzed the fraction of SIGINT choices for which the model requests SIGINT on the group with the highest probability. The result of the fraction for the ACT-R model is presented in table 2. The fraction of the model is greater than .5, so the ACT-R model is exhibiting confirmation bias in seeking evidence according to the  $C$  metric.

We also analyzed how the INT layers selected by the ACT-R model compared to the rational choice based on the expected information gain. The result is presented in figure

4. Note that the number of alternative choices varies within a task: The task begins with seven alternatives (IMINT, MOVINT, SOCINT, and four SIGINTs) available, and depending on the selection of the layer, the alternatives decrease within each trial.

Table 2: The results of the confirmation bias in seeking evidence for both models.

	SIGINT on the highest prob. group	Total No. of SIGINT	Fraction
ACT-R Model	6,191	9,044	.68

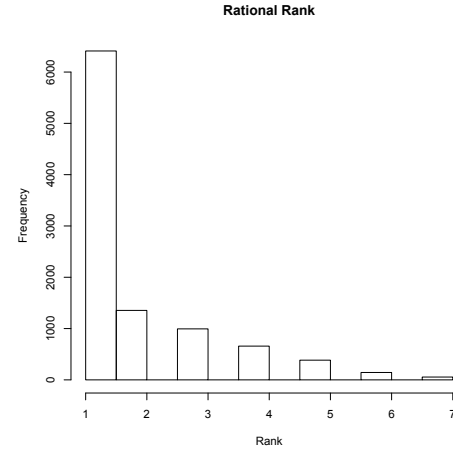


Figure 4: Frequency of the ACT-R model selecting the rational choice of Rank  $n$  (Rank 1 is the optimal choice).

Table 3 shows a confusion matrix that indicates the proportion of times the ACT-R model makes the same choice as the rational model. Although the ACT-R model agrees with the rational model at a level well above chance, it often differs from the rational. The rational model scarcely selects SOCINT layer (3 times among 30000), because the expected information gain for SOCINT is relatively low.

Table 3: Confusion matrix of the ACT-R model and rational model for layer selection.

		Rational Choice			
		IMINT	MOVINT	SIGINT	SOCINT
ACT-R Choice	IMINT	75%	15%	4%	0%
	MOVINT	18%	79%	4%	66%
	SIGINT	4%	3%	91%	0%
	SOCINT	3%	3%	1%	33%

Note that there is some interaction between the anchoring bias in evidence weighing and any biases that might emerge in choosing layers. If the ACT-R models (or participants) under-weight evidence and believe in a “less peaky”

probability distribution over hypotheses, then that can affect how far they believe that the current state or next state is from the goal, or how much more uncertainty can be reduced by a given layer choice. Biases in beliefs about the current situation will impact evidence-gathering choices.

The ACT-R model exhibits confirmation bias when evaluated against the ICArUS task-specific norm, *C*, which measures the propensity to use SIGINT to confirm the strongest current hypothesis. However, the selection of INT layers is generally highly consistent with the rational norm of seeking evidence that will produce the highest expected information gain. This illustrates how the notion of “bias” is dependent on the choice of norm, and how such norms do not always agree, especially in the case of “confirmation bias” (Jonathan D. Nelson, 2005). It has been shown (Austerweil & Griffiths, 2011) that confirmatory strategies are rational for a large class of tasks and people appear to approximate choices based on expected information gain.

### Acknowledgments

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract number D10PC20021. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

### References

- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036-1060.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.
- Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, 55(3), 499-526.
- Cheikes, B. A., Brown, M. J., Lehner, P. E., & Adelman, L. (2004). Confirmation bias in complex analyses MITRE Center for Integrated Intelligence Systems. Bedford: MA: MITRE.
- Chorev, M. (1996). Surprise Attack. The Case of the Yom-Kippur War. Fort McNair: WA: The industrial College of the Armed Forces.
- Convertino, G., Billman, D., Pirolli, P., Massar, J., & Shrager, J. (2008). The CACHE Study: Group Effects in Computer-Supported Collaborative Analysis. *Computer Supported Cooperative Work (CSCW)*, 17(4), 353-393.
- Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, 27, 591-635.
- Grabo, C. M., & Goldman, J. (2004). *Anticipating surprise: Analysis for strategic warning*. Washington, D.C.: Joint Military Intelligence College.
- Heuer Jr, R. J. (1999). Psychology of intelligence analysis: Center for the Study of Intelligence, Central Intelligence Agency (Washington, DC).
- Heuer, R. J. (1999). *Psychology of Intelligence Analysis*. Washington, D.C.: Center for the Study of Intelligence.
- Klayman, J. (1995). Varieties of confirmation bias. *Psychology of learning and motivation*, 32, 385-418.
- Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, 94(2), 211-228.
- Klein, G. (1998). *Sources of power: How people make decisions*. Cambridge, MA: MIT Press.
- Klein, G., Moon, B., & Hoffman, R. R. (2006a). Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems*, 21(4), 70-73.
- Klein, G., Moon, B., & Hoffman, R. R. (2006b). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, 21(5), 88-92.
- Lebiere, C. (1999). The dynamics of cognition: An ACT-R model of cognitive arithmetic. *Kognitionswissenschaft*, 8, 5-19.
- Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.
- Nelson, J. D. (2005). Finding Useful Questions: On Bayesian Diagnosticity, Probability, Impact, and Information Gain. *Psychological Review*, 112(4), 979-999. doi: 10.1037/0033-295x.112.4.979
- Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience Matters. *Psychological science*, 21(7), 960-969.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175.
- Pirolli, P., & Card, S. K. (2005). *The sensemaking process and leverage points for analyst technology*. Paper presented at the 2005 International Conference on Intelligence Analysis, McLean, VA.
- Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). *The cost structure of sensemaking*. Paper presented at the INTERCHI '93 Conference on Human Factors in Computing Systems, Amsterdam.
- Simon, H. A. (1974). How big is a chunk? *Science*, 183, 482-488.
- Tolcott, M. A., Marvin, F. F., & Lehner, P. E. (1989). Expert decision-making in evolving situations. *IEEE Transactions on Systems, Man and Cybernetics*, 19(3), 606-615.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, 12(3), 129-140.
- Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance* (3rd ed.). Prentice-Hall: Upper Saddle River, NJ.



# Can native-language perceptual bias facilitate learning words in a new language?

Bożena Pajak<sup>1</sup> (bpajak@ucsd.edu), Sarah C. Creel<sup>2</sup> (creel@cogsci.ucsd.edu), Roger Levy<sup>1</sup> (rlevy@ucsd.edu)

<sup>1</sup>Department of Linguistics, <sup>2</sup>Department of Cognitive Science, UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

## Abstract

Acquiring a language relies on distinguishing the sounds and learning mappings between meaning and phonetic forms. Yet, as shown in previous research on child language acquisition, the ability to discriminate between similar sounds does not guarantee success at learning words contrasted by those sounds. We investigated whether adults, in contrast to young infants, are able to attend to phonetic detail when learning similar words in a new language. We tested speakers of Korean and Mandarin to see whether they could use their native-language-specific perceptual biases in a word-learning task. Results revealed that participants were not able to fully capitalize on their perceptual abilities: only faster learners – as independently assessed by baseline trials – showed enhanced learning involving contrasts in phonetic dimensions informative in their native language. This suggests that attention to phonetic detail when learning words might only be possible for adults with better learning skills or higher motivation.

**Keywords:** word learning; non-native speech perception; second language acquisition

## Introduction

Humans are able to take advantage of many different resources available to them in the course of learning. For example, when learning a new language – whether in infancy or adulthood – humans actively search for regularities by analyzing the input in several alternative ways (e.g., examining either adjacent or non-adjacent dependencies; Gómez, 2002), and are able to simultaneously entertain multiple implicit theories about the input's underlying structure (e.g., Gerken, 2010). One of the complex features of learning a language is that listeners must perform concurrent analyses of the input at different levels of processing and integrate these multiple pieces of information at once. If, for example, we zoom in to the level of processing single words, one needs to encode phonetic cues and, at the same time, map the phonetic form onto meaning. This task may be particularly hard for beginner second language (L2) learners who are not yet familiar with the L2 sound system, especially when they are processing words with novel sounds that do not exist in their native language (L1). However, there is evidence that learners capitalize on whatever pieces of information are available to them to achieve this task: they might use lexical cues to make inferences about sound categorization (Feldman, Myers, White, Griffiths, & Morgan, 2011), and – conversely – take advantage of perceptual learning on sound categorization to help them make inferences about the lexicon (Perfors & Dunbar, 2010). The question we are addressing in this paper is another piece of this puzzle.

Specifically, we know that prior language knowledge is one of the starting points when learning a new language: L1-based perceptual biases can facilitate perception of novel sound contrasts that differ along phonetic dimensions informative in L1 (Pajak, 2010a, 2010b; Pajak & Levy, in prep.), and can

affect interpretation of distributional information from novel language input (Pajak & Levy, to appear). How efficiently, then, do adults capitalize on their L1-based phonetic generalizations when learning the lexicon in a new language?

Intuitively, it might seem that whatever perceptual abilities adults have, they should be able to use them both when distinguishing sounds and when learning novel words. That is, if they hear a distinction between sounds *b* and *p*, they should be able to easily distinguish between words like *ban* and *pan*. However, the picture emerging from prior research is far less clear. In fact, research with young infants suggests that the ability to discriminate perceptually between similar sounds does not in general guarantee immediately successful learning of words that are contrasted by those sounds. At 14 months, infants can easily discriminate the sounds *b* and *d*. However, when taught that a novel object is called a *bih*, but later on is referred to as a *dih*, infants do not notice this mispronunciation (Stager & Werker, 1997). The initial explanation proposed for this result was the limited resource hypothesis (Stager & Werker, 1997; Werker, Fennell, Corcoran, & Stager, 2002): since attending to fine phonetic detail while learning new words is computationally very demanding, young infants – who have limited attentional and cognitive resources – might have difficulty accessing all phonetic detail when focusing their attention on learning meaning. Subsequent research showed that 14-month-old infants succeed only with additional contextual information or under less demanding learning conditions (Fennell & Werker, 2003; Fennell, Waxman, & Weisleder, 2007; Rost & McMurray, 2009; Swingley & Aslin, 2002; Thiessen, 2007; Yoshida, Fennell, Swingley, & Werker, 2009).

Some evidence suggests that adults might have similar difficulties when learning words in a new language. In a study by Perfors and Dunbar (2010), native speakers of English were first trained on discriminating a non-native contrast between a prevoiced and a voiceless unaspirated stop ([gipur] vs. [kipur]), and then taught word-picture mappings using minimal-pair words distinguished by this non-native contrast. The results showed that while participants performed better than chance at learning similar words with the exact contrast they had been trained on ([gipur]-[kipur]), they were at chance at learning words contrasted by sounds with an analogous contrast ([bipur]-[pipur]). This was despite the fact that, after perceptual training on [g]-[k], participants were able to distinguish [b] and [p] perceptually. Thus, just like 14-month-old infants, adults had difficulty differentiating between similar words in a word-learning task, even though they could tell these words apart in a pure perceptual task.<sup>1</sup>

<sup>1</sup>Better performance on [gipur] and [kipur]) might have been due

However, the difficulty in learning similar-sounding words found by Perfors and Dunbar (2010) might have a purely perceptual basis. That is, learners' ability to discriminate a pre-voiced [b] vs. a voiceless unaspirated [p] might not have been sufficiently robust to be of any use in a word-learning task. This is similar to the intuition of Perfors and Dunbar, who point out that learners' representations of [b] and [p] categories might have been too fragile to see any advantage in word learning. If this reasoning is correct, then the comparison between 14-month-olds and adults in Perfors and Dunbar's (2010) study is less warranted because, at 14 months, infants have difficulty learning the words *bih* and *dih* despite easily discriminating between the sounds *b* and *d*.

In the study reported here we achieve a more direct comparison with the situation of 14-month-old infants by investigating how adults learn similar-sounding words that they can distinguish perceptually due to their L1-based phonetic generalizations. Specifically, we used two distinctions: *length* (e.g., [taja]-[tajja]) and *place* of articulation between alveolo-palatal and retroflex sounds (e.g., [gotɕa]-[gotɕa]). Our participants were native speakers of Korean and of Mandarin, who were previously shown to have differential perceptual sensitivity to these two distinctions, as illustrated in Figure 1 (Pajak, 2010a, 2010b; Pajak & Levy, in prep.). In particular, Korean speakers were shown to be better than Mandarin speakers at discriminating consonant length contrasts ([m]-[mm], [n]-[nn], [l]-[ll], [s]-[ss], [f]-[ff], [j]-[jj], [w]-[ww]), but the reverse was true for the alveolo-palatal vs. retroflex place contrasts ([ɕ]-[ʂ], [tɕ]-[ʈ], [ʑ]-[ʑ], [dʑ]-[dʑ]). This result was likely due to the fact that Korean has length distinctions, and Mandarin does not, but Mandarin has alveolo-palatal and retroflex sounds, while Korean does not (Lin, 2001; Sohn, 2001). Crucially, however, these perceptual sensitivities cannot be attributed to direct L1-to-L2 phonetic category transfer alone (as has been generally proposed for these types of perceptual patterns; Major, 2008) because not all of the tested speech sounds exist in Korean or Mandarin.<sup>2</sup> Consequently, Pajak hypothesized that perceptual advantages in non-native speech processing can arise from sensitivity to phonetic *dimensions* that are informative in L1, and not just sensitivity to specific L1 categories. In this study we test whether these perceptual advantages are exploited during word learning.

## Experiment

Participants learned novel word-picture mappings, where each word was in a minimal pair with either a *length* dis-

to the familiarity with these lexical items from perceptual training, which is consistent with infants also performing better on familiar words (Swingley & Aslin, 2002).

<sup>2</sup>Korean mostly uses length distinctions on vowels (e.g., [pu:l] 'fire' vs. [pu:l] 'blow'), but some long consonants ([ll], [nn], [mm]) arise from phonological assimilation processes (Sohn, 2001), and Korean tense obstruents ([p̚], [t̚], [k̚], [s̚], [t̚ɕ]) have sometimes been analyzed as long (Choi, 1995). As for Mandarin, alveolo-palatals and retroflexes exist as allophones of the same phonemic category, and voiced obstruents ([ʑ], [dʑ], [dʑ]) are entirely absent (except [ʑ]) since Mandarin has obstruent distinctions in aspiration, not voicing (Lin, 2001).

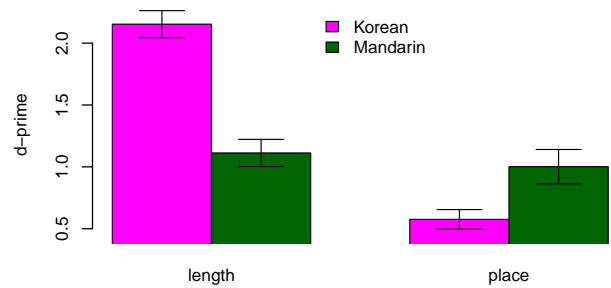


Figure 1: Perceptual discrimination results (Pajak, 2010a, 2010b; Pajak & Levy, in prep.).

inction or an alveolo-palatal vs. retroflex *place* distinction. We predicted that if adult L2 learners are able to attend to phonetic detail by using their L1-based resources when learning a new lexicon, then we should observe the same pattern in both the perceptual discrimination and the word-learning tasks: that is, Korean speakers should be better at learning *length* minimal pairs, and Mandarin speakers better at learning *place* minimal pairs.

## Method

**Participants** 54 undergraduate students at UC San Diego participated in the experiment for course credit or payment. Half were speakers of Korean, and the other half were speakers of Mandarin. Participants varied in terms of their length of residence in the US: some were born in the US, while others immigrated at some point after birth or were international students who arrived very recently. Consequently, they varied in English proficiency. Importantly, however, they all learned Korean or Mandarin from birth, reported high proficiency in those languages, and still used them regularly, predominantly with family. In most cases they had some high school and/or college exposure to Spanish or French. Some Mandarin speakers were also familiar with Taiwanese, mostly through family exposure. All participants reported no history of speech or hearing problems.

**Materials** The materials consisted of 16 nonce words of the form CVC(:)V, where each was in a minimal pair differing only in the middle consonant (a subset of contrasts tested by Pajak, 2010a, 2010b; Pajak & Levy, in prep.). There were 12 *length* words, with either a short or a long middle consonant, and 4 *place* words, with either an alveolo-palatal or a retroflex consonant (both pronounced as in Polish, which has a distinction similar to Mandarin), as illustrated in Table 1.

The materials were recorded in a soundproof booth by a phonetically-trained native speaker of Polish. There were 10 tokens recorded for each word. For *length* words, two tokens of each word with long consonants were chosen for the experiment. Subsequently, words with short consonants were created by shortening the tokens with long consonants in a way that, for each word and each recording, the naturally-recorded long consonant was reduced to half its duration so as

Table 1: Stimuli (in IPA).

LENGTH WORDS		PLACE WORDS	
<i>short</i>	<i>long</i>	<i>alveolo-palatal</i>	<i>retroflex</i>
taja	tajja		
tala	talla		
diwa	diwwa	gotça	gotṣa
difa	diffa	goṣa	goṣa
kema	kemma		
kena	kenna		

to maintain a constant 2:1 duration ratio (cross-linguistically, the long-to-short consonant ratio varies between 1.5 to 3; Ladefoged & Maddieson, 1996). For *place* words, two tokens each were chosen for the experiment with the goal of maximizing the similarity between the words in minimal pairs with regards to how vowels were pronounced, but at the same time choosing tokens with clearly enunciated middle consonants.

Each word was paired with a picture of a different kind of mushroom (see two examples in Fig. 2), which were chosen in order to include objects that were unfamiliar to our participants, but not so unfamiliar that participants would find them bizarre and hard to remember. We selected pictures that varied in shape and color so as to maximize visual differences between them. We created four different one-to-one word-to-picture mappings that were counterbalanced between participants in order to make sure that the results were not driven by any peculiarities in the mappings we chose.

**Procedure** Participants sat in front of a computer, and responded by using a mouse. They were instructed that in this experiment they would be learning a novel language, and, specifically, the language’s words for mushrooms. The experiment was completed in a single session. There were 4 training blocks (each with 128 trials, about 10-15min long) and 4 testing blocks (each with 64 trials, about 5min long), interleaved. Blocks were separated by self-terminated breaks. In each trial, two pictures were presented on a computer screen (see Fig. 2), and a word was played through headphones with a delay of 500ms. Participants were asked to click on the picture that they thought went with the word. In training, feedback was provided following the response in the form of the correct picture staying on the screen. A mouse click triggered the start of the next trial. Presenting feedback after participant’s response meant that the early responses were necessarily random. Participants were told to guess at first, and that through feedback they would eventually learn the correct word-to-picture mappings. In testing, no feedback was provided.

The training trial types consisted of picture pairs that were always associated with dissimilar word pairs (e.g., *taja-diwa*, *gotça-kemma*) so that participants were not directly alerted to the distinctions of interest. The testing trial types were always different from the training trials. There were four types of tri-



Figure 2: Example of a screen shot that participants saw throughout the experiment.

als in testing depending on the minimal-pair contrast that corresponded to the pictures, as illustrated in Table 2: (i) *length* (24 trials per block), (ii) *place* (8 trials), (iii) filler-dissimilar (16 trials), and (iv) filler-similar (16 trials). The critical trials consisted of the critical minimal-pair picture pairings (i.e., pairs of pictures whose corresponding words were a minimal pair): *length* pairs and *place* pairs. The filler trials consisted of *dissimilar* pairs, always differing in the first CV sequence, and *similar* pairs that shared the initial CV sequence. The picture position was counterbalanced. The trial order was pseudo-randomized: we created four randomized lists, and then altered them manually so that the same word was never repeated in two consecutive trials. Furthermore, the minimal-pair trials were always separated by at least two other trials. Each participant heard each list once, with a different list for each block. The block order was counterbalanced across participants.

Table 2: Trial types in testing.

CRITICAL PICTURE PAIRS		FILLER PICTURE PAIRS	
<i>type</i>	<i>example</i>	<i>type</i>	<i>example</i>
<b>Length</b>	taja-tajja	<b>Dissimilar</b>	tala-goṣa
<b>Place</b>	gotça-gotṣa	<b>Similar</b>	tala-taja

## Results

We analyzed accuracy scores from testing with mixed-effects logit models (Jaeger, 2008). We included random intercepts for participants and items, and random slopes for participants and items for all effects of interest that were manipulated within participants or within items. We controlled for participants’ nonverbal IQ, self-reported L1 proficiency, and current L1 exposure and use by adding them as fixed effects to the models. Although we present data for the four test blocks separately for purposes of visualization, in statistical analyses we collapse across test block.

As a sanity check, we expected that all participants, regardless of language background, should perform best on *filler-dissimilar* trials, slightly worse on *filler-similar* trials, and worst on *critical* trials. These overall results were borne out, as illustrated in Figure 3. In a model with fixed effects of TRIAL TYPE (*filler-dissimilar*, *filler-similar*, *critical*)

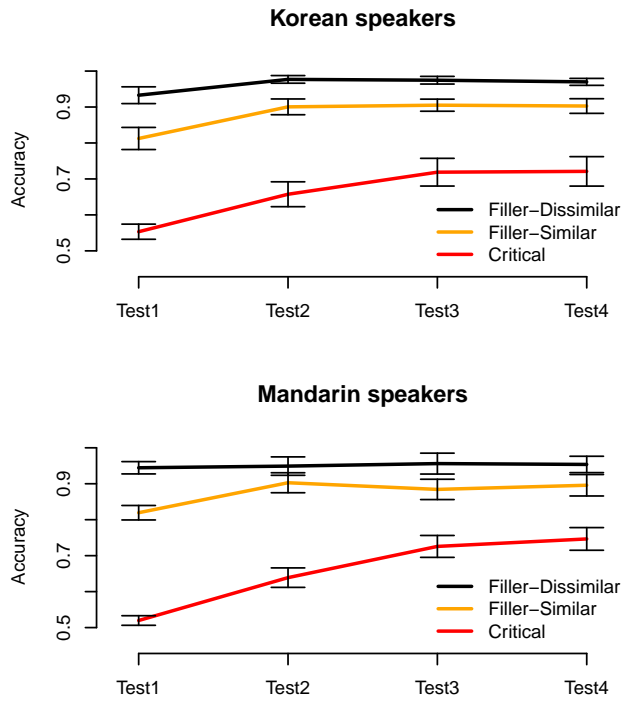


Figure 3: Proportion of correct responses on critical and filler trials (in all figures, error bars are standard errors).

and LANGUAGE (*Korean, Mandarin*), we found that the responses in the *filler-dissimilar* condition were significantly higher than in the *filler-similar* condition ( $p < .001$ ), which in turn were higher than in the *critical* condition ( $p < .001$ ). Neither LANGUAGE nor its interactions were significant in the model. Furthermore, a log-likelihood ratio test comparing the full model to a reduced version which did not contain LANGUAGE revealed no evidence that language background significantly contributed to the model ( $\chi(5) = 2.20; p = .82$ ), thus suggesting that there were no significant differences between the two language groups in overall response patterns.

Next, we compared Korean and Mandarin speakers on critical trials in a model with fixed effects of CRITICAL TRIAL TYPE (*length, place*) and LANGUAGE (*Korean, Mandarin*). If learners are able to capitalize on their L1-based perceptual generalization when beginning to learn new words, we should observe a difference in performance between the two language groups in line with their perceptual biases: Korean speakers should be more accurate on *length* pairs, and Mandarin speakers more accurate on *place* pairs. However, there was no significant interaction between CRITICAL TRIAL TYPE and LANGUAGE ( $p=.21$ ), indicating that Korean and Mandarin speakers did not differ in their accuracy when learning similar-sounding words that differed in either length or place. These results are illustrated in Figure 4. This is in striking contrast to perceptual discrimination results (Fig. 1; Pajak, 2010b, 2010a; Pajak & Levy, in prep.), where – using similarly constructed stimuli – Korean speakers clearly outperformed Mandarin speakers on perception of length con-

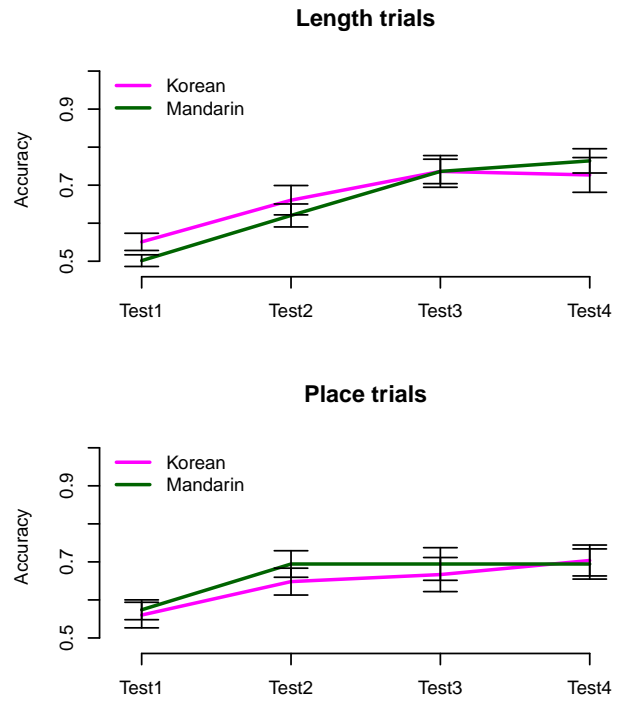


Figure 4: Proportion of correct responses on critical trials.

trasts, and the reverse was true for place contrasts.

However, we know that learners vary in their attention, motivation and learning skills. Thus, we asked whether only better learners are able to use their L1 resources and attend to fine phonetic detail in word learning. To answer that question we split the participants into two halves, top- and bottom-scoring on filler trials. We chose this way of doing the median split due to the fact that performance on fillers was a dimension independent from the variables of interest. The median score on all fillers combined was 94.5% accuracy. There were 7 participants who scored right at 94.5%, who were then split based on their performance on *dissimilar* fillers alone. The distribution of participants in terms of their language background was fairly equal in both groups: Korean=13 and Mandarin=14 in the top half, and Korean=14 and Mandarin=13 in the bottom half. The filler scores for both top-scoring and bottom-scoring participants are provided in Table 3 (next page). Both groups were highly accurate on filler pairs (at least 80% accuracy), but there was much more variability in the bottom-scoring group.

The results split by top-scoring vs. bottom-scoring participants are illustrated in Figures 5 and 6. Even by visual inspection alone, the results on critical trials look strikingly different in the top vs. bottom-scoring group: in the top half, participants were clearly improving in the course of the experiment (with the biggest jump from Test1 to Test2), while in the bottom half, participants' responses were close to chance throughout the experiment, with only minimal signs of learning (namely, Mandarin speakers seemed to improve on *length* trials toward the end of the experiment). We analyzed these

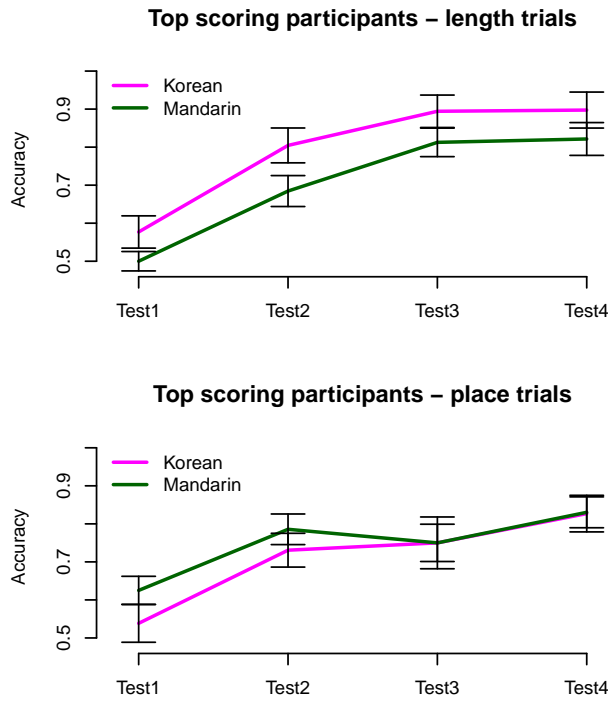


Figure 5: Faster learners: proportion of correct responses on critical trials.

Table 3: Proportion of correct responses on filler trials (standard errors in parentheses).

FILLER TYPE	TOP HALF		BOTTOM HALF	
	<i>Korean</i>	<i>Mandarin</i>	<i>Korean</i>	<i>Mandarin</i>
Test 1	.95 (.01)	.92 (.01)	.80 (.04)	.84 (.02)
Test 2	.98 (.01)	.98 (.01)	.90 (.02)	.87 (.05)
Test 3	.98 (.01)	.97 (.01)	.90 (.01)	.86 (.05)
Test 4	.99 (.00)	.99 (.01)	.89 (.02)	.86 (.05)

results with a model with fixed effects of CRITICAL TRIAL TYPE (*length*, *place*) and LANGUAGE (*Korean*, *Mandarin*), and an additional fixed effect of FILLER PERFORMANCE (*top*, *bottom*). We found a significant three-way interaction between CRITICAL TRIAL TYPE, LANGUAGE, and FILLER PERFORMANCE ( $p < .01$ ), indicating distinct response patterns for Korean vs. Mandarin speakers on *length* and *place* trials depending on their overall success rate in learning, as measured by their accuracy on filler trials. (A regression analysis treating filler performance as a continuous covariate yielded similar results).

For the top-scoring group, we found the pattern indicating that participants were taking advantage of their perceptual biases: Korean speakers more accurate on *length* trials than Mandarin speakers, but not on *place* trials, as indicated by a significant interaction between CRITICAL TRIAL TYPE and LANGUAGE ( $p < .01$ ). Furthermore, a model examining *length* trials only revealed a significant main effect of

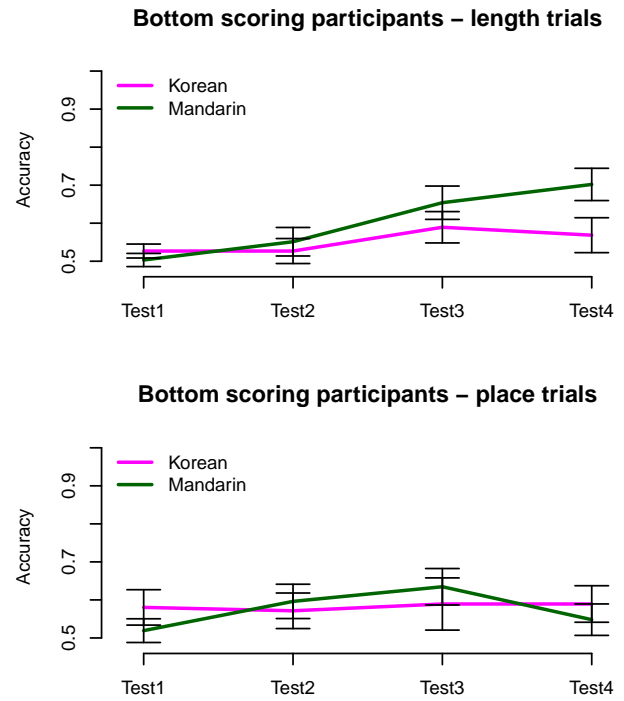


Figure 6: Slower learners: proportion of correct responses on critical trials.

LANGUAGE ( $p < .05$ ). On *place* trials, Korean and Mandarin speakers were not significantly different, but the numerical tendency was the opposite of that seen in the *length* trials: Mandarin speakers were slightly more accurate than Korean speakers.

For bottom-scoring participants, on the other hand, we found no significant interaction between CRITICAL TRIAL TYPE and LANGUAGE. There were also no significant differences between Korean and Mandarin speakers when only *length* ( $p = .21$ ) or only *place* trials ( $p = .99$ ) were examined. However, Mandarin speakers did seem to improve on *length* – but not *place* – trials toward the end of the experiment. It is unclear why Mandarin, but not Korean speakers showed this improvement, especially since – if anything – it was Korean speakers who seemed to perform slightly better on filler trials in the bottom-scoring group. As for the learning asymmetry – improvement on *length*, but not on *place* – it could be that *length* was simply easier to learn, perhaps due to it being a more salient cue (note that in the perception study, Fig. 1, *length* contrasts were obviously easier than *place*) and due to the fact that the stimuli included significantly more *length* words compared to *place* words, thus allowing more opportunity for perceptual learning of the *length* contrast.

Why did some learners succeed at using their perceptual abilities while others did not? The individual measures we collected indicate that the top-scoring and the bottom-scoring groups did not differ in nonverbal IQ nor L1 proficiency. However, the bottom-scoring participants did, on average, immigrate to the US later in life, and, consequently, had lower



English proficiency. This suggests that there might be an advantage for more balanced bilinguals (consistent with findings by Kaushanskaya & Marian, 2009), or perhaps students less accustomed to the US educational system have more overall difficulty performing these kinds of tasks in a laboratory setting.

## Discussion

Previous work (Pajak, 2010a, 2010b; Pajak & Levy, in prep.) has shown beneficial effects of L1 properties on L2 discrimination, but what about learning? For the populations as a whole there was no clear effect indicating that participants made effective use of their native-language resources, but there is evidence that the faster learners (as independently assessed by filler performance) were able to do so. This is in contrast to the discrimination results we cited, where L1-based perceptual advantages were observed for all participants. This result suggests that there is something inherently hard about the early stage of word learning that precludes attention to fine phonetic detail that is otherwise available during phonetic processing. This is even more surprising given that adults have well-developed attentional and cognitive capacities, but nevertheless fail to use them in this task. It is still an open question as to what exactly made the faster group of learners succeed at using their L1 resources and attending to fine phonetic detail. Perhaps they were more attentive throughout the experiment, more motivated, or had better learning skills.

Overall, the results reported here shed some more light on the interaction between sound perception and word learning in adults. In particular, they suggest that perceptual resources are not easily used when learning minimal-pair words. An intriguing possibility – consistent with results from infant studies (Thiessen, 2007; see also Feldman, Griffiths, & Morgan, 2009) – is that learners' initial strategy is to assume that similar sounding minimal pairs are homophones, and only phonetic evidence from non-minimal pairs (or explicit information about the number of categories) pushes them to revise that assumption. This kind of parsimony might benefit learning when there is uncertainty about phonetic category boundaries – a possibility that we plan to pursue in future research.

## Acknowledgments

For their helpful feedback the authors thank Eric Baković, Klinton Bicknell, Computational Psycholinguistics Lab at UC San Diego, and the audience of LSA 86. K. Michael Brooks helped with data collection. We are grateful to Eugene Carsey for permission to use his photographs of mushrooms (www.eugencarsey.com). This research was supported by NIH Training Grant T32-DC000041 from the Center for Research in Language at UC San Diego to the first author, and the UC San Diego Academic Senate Grant to the third author.

## References

Choi, D.-I. (1995). Korean "tense" consonants as geminates. *Kansas Working Papers in Linguistics*, 20, 25–38.

- Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2208–2213). Austin, TX: Cognitive Science Society.
- Feldman, N. H., Myers, E., White, K., Griffiths, T. L., & Morgan, J. L. (2011). Learners use word-level statistics in phonetic category acquisition. In *Proceedings of the 35th Boston University Conference on Language Development* (pp. 197–209). Somerville, MA: Cascadilla Press.
- Fennell, C. T., Waxman, S. R., & Weisleder, A. (2007). With referential cues, infants successfully use phonetic detail in word learning. In *Proceedings of the 31st Boston University Conference on Language Development*. Cascadilla Press.
- Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech*, 46(2-3), 245–264.
- Gerken, L. (2010). Infants use rational decision criteria for choosing among models of their input. *Cognition*, 115, 362–366.
- Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, 13(5), 431–436.
- Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed effects models. *Journal of Memory and Language*, 59, 434–446.
- Kaushanskaya, M., & Marian, V. (2009). The bilingual advantage in novel word learning. *Psychonomic Bulletin & Review*, 16(4), 705–710.
- Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford, UK; Cambridge, MA: Blackwell.
- Lin, H. (2001). *A grammar of Mandarin Chinese*. München: Lincom Europa.
- Major, R. C. (2008). Transfer in second language phonology: A review. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 64–94). Amsterdam/Philadelphia: John Benjamins.
- Pajak, B. (2010a). *Bilinguals generalize from known phonological contrasts in perception of a novel language*. St. Louis, MO. (Poster at the 51st Annual Meeting of the Psychonomic Society)
- Pajak, B. (2010b). Perceptual advantage from generalized linguistic knowledge. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 369–374). Austin, TX: Cognitive Science Society.
- Pajak, B., & Levy, R. (in prep.). *Inductive generalization over phonetic categories in perceptual reorganization: the case of segmental length*. (Manuscript)
- Pajak, B., & Levy, R. (to appear). Distributional learning of L2 phonological categories by listeners with different language backgrounds. In *Proceedings of the 36th Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.
- Perfors, A., & Dunbar, D. (2010). Phonetic training makes word learning easier. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1613–1618). Austin, TX: Cognitive Science Society.
- Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, 12(2), 339–349.
- Sohn, H.-M. (2001). *The Korean language*. Cambridge, UK: Cambridge University Press.
- Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388, 381–382.
- Swingle, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, 13(5), 480–484.
- Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, 56, 16–34.
- Werker, J. F., Fennell, C. T., Christopher T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: effects of age and vocabulary size. *Infancy*, 3(1), 1–30.
- Yoshida, K. A., Fennell, C. T., Swingle, D., & Werker, J. F. (2009). Fourteen-month-old infants learn similar-sounding words. *Developmental Science*, 12(3), 412–418.

# Seeing who sees: Contrastive access helps children reason about other minds

Kathie Pham, Elizabeth Bonawitz, & Alison Gopnik

{kathiepham, liz\_b, gopnik}@berkeley.edu

University of California, Department of Psychology, 3210 Tolman Hall  
Berkeley, CA 94720 USA

## Abstract

Does contrastive access help preschoolers succeed on traditional false-belief tasks? Three- and four-year-olds were presented with a modified version of the change-of-location story in which *two* characters are the focus of interest. In the contrastive access condition preschoolers observe that one character leaves the room while the other stays and witnesses the moving event; in the non-contrastive condition both characters leave the room and fail to observe the moving event. Despite having to track two different characters and their different knowledge states about the location of the toy, preschoolers were more likely to succeed on the task when the characters had contrasting access to the moving event. This result supports a previously unexplored qualitative prediction of the Goodman et al (2006) computational model of the false-belief task and also provides tentative support for the theory theory view of the false-belief transition.

**Keywords:** Cognitive development; theory of mind; False-belief task; Contrastive learning.

## Theory theory of mind

The ability to reason about other people's mental states, such as their beliefs and desires, their fears and aspirations, is often referred to as theory of mind. Having a theory of mind allows us to construct others as mental beings: entities much grander than their physical attributes or their observable actions. One result of this understanding is that as adults, we are able to not only consider our own beliefs, but the beliefs of countless others—diverging beliefs about a single reality, beliefs that may be mistaken.

Decades of research have suggested that three-year-olds tend to struggle with false-belief reasoning in a very specific way. Studies have shown that three-year-olds misinterpret minds systematically—when an agent's beliefs and reality diverge, they predict actions of that agent to be consistent with the reality, rather than the false-belief (Wimmer & Perner, 1983; Perner et al., 1987). One classic example that tests a child's false-belief understanding is the change-of-location task (Wimmer & Perner, 1983). A child is read a story about a character (e.g.) Sally, who stores her toy and then leaves the room. While she is away, a mischievous character moves the toy. Sally then returns to look for her toy and the child is asked, "Where will Sally first go look for her toy?" Three-year-olds often say that Sally will look where the toy actually is, consistent with the true state of the world, rather than the location consistent with the agent's false-belief. In contrast, older four-year-olds more often correctly answer that Sally will look in the place that the toy was initially left, successfully considering an agent's beliefs

(e.g. Baron-Cohen et al., 1985; Perner, Leekam, & Wimmer, 1987; Wimmer & Perner, 1983).

Despite decades of research replicating this finding, there is much debate about how and when knowledge about other's mental states develops, and in particular when children develop an understanding of false-belief. Some studies suggest that children go through a conceptual change around ages three to five—from systematically failing false-belief tasks to performing above chance (Wellman, Cross, & Watson, 2001). However, there have been compelling arguments for earlier developing theory of mind competence suggesting that as early as 10 to 15 months infants already have an awareness that actors act on the basis of their beliefs and false-beliefs (e.g., see Baillargeon, Scott, and He, 2010 for a review).

It is not yet clear how to best interpret these infant "false-belief" findings nor how to reconcile or integrate them with the preschool ones. Regardless, something definite and important is happening in children's theory-of-mind understandings in the preschool years, beyond earlier developments in infancy. There are likely to be contrasts between implicit predictive and explicit causal-explanatory knowledge. Furthermore, differences in false-belief understanding as measured in the preschool years predict several key childhood competences, such as how and how much children talk about people in everyday conversation, their engagement in pretense, their social interactional skills and consequently their interactions with and popularity with peers (Astington & Jenkins 1995; Lalonde & Chandler 1995; Watson et al. 1999). Furthermore, variability in preschool performance on theory of mind tasks overlaps with but is distinctively different from executive function and IQ (e.g., Carlson & Moses 2001). These findings are important for confirming theory of mind's significance and relevance during the preschool years as indexed by preschool theory of mind tasks (especially as researched thus far for false-belief tasks).

Though it is unclear what factors support success on looking-time measures in young infants, the research that will be presented here assumes a theory-like competence that, in particular, supports explanation (e.g. Gopnik & Wellman, 1992; Wellman & Liu, 2007). We take the idea that theory of mind is analogous to scientific theories, resulting in children's distinctive patterns of predictions and interpretations of evidence, which is often referred to as the theory theory account of theory of mind development (e.g. Gopnik, 1993; Gopnik & Wellman, 1992; Perner, 1991). What a theory-like understanding of mind permits is conceptual change—theory revision in the face of new



evidence, and beliefs that support verbal predictions, explanations, and counterfactual reasoning.

The explanatory value of the theory theory is limited in that current accounts do not define the specific mechanisms for change. However, advances in computational accounts of theory change and probabilistic models in particular naturally integrate with qualitative predictions of the theory theory (e.g. Schulz, Bonawitz, & Griffiths, 2007). In what follows we will briefly describe one such account of the false-belief transition and discuss a prediction about the role of contrast that falls out of this model. We then present a new empirical study designed to test this prediction. We conclude with a discussion of how these findings support a theory theory account of false-belief.

### Rational account of the false-belief transition

There have been a few computational accounts of false-belief transition (Berthiaume, Schultz, & Onishi, in review; Goodman et al, 2006; O’Loughlin & Thagard, 2000; Triona, Masnick, & Morris, 2002; Van Overwalle, 2010). Consistent with the idea that children’s changing proficiency on false-belief tasks are guided only by changes in executive function, O’Loughlin and Thagard (2000) have produced a connectionist model where the false-belief transition is driven by an increase in inhibition of the true belief location. We consider a different proposal consistent with the theory theory (Goodman et al. 2006, Figure 1). The model makes explicit the variables (concepts) that children before and after false-belief transition represent and provides an account of why the explanatory variables appealed to by passers and failers are different. Specifically, the model proposes that children incorporate a visual access variable (seeing the final location of the toy) into their theory of mind models. This is a critical variable in the change-of-location story; passers seem to understand that visual access influences an agent’s belief states and her subsequent actions: because Sally did not see the toy moved, she does not know that it is in the new location.

In support of the claim that an understanding of visual access is changing in a young false-belief reasoner, children’s explanations also reflect a shift in understanding the causal relation between an agent’s access and beliefs. In the failer’s model of theory of mind, explanatory power is reduced because of fewer variables available in the model. Goodman et al. (2006) found that children who successfully predicted an agent’s action in a false-belief task generate more belief and access explanations, whereas failers of the task appeal more to desires. For example, a passer of the task may explain why an agent went to a surprising location (where the toy is not): “Because she did not see it moved” (appealing to access), whereas a failer may explain why an agent went to a surprising location: “Because she wanted to go there” (appealing to desire). This is consistent with the proposed models in figure one—only in the passers’ model do children have access as a causally connected variable. In contrast, the failers’ model only has alternate desire available as an explanation of surprising behavior.

In this paper we explore one previously untested implication of the Goodman et al. (2006) model. If access is made more salient to children at the cusp of false-belief understanding, then success (correct predictions and explanations that appeal to “belief”) on the false-belief task should increase. Critically, our modification of the classic false-belief task makes the task more complicated for children to follow; if children’s success on these tasks is only dependent on development of executive function, then such a modification should *decrease* success. If instead children have a theory-like representation of mental states as sketched by the Goodman et al. (2006) model, then we should observe an increase in correct predictions (where Sally and Billy will look for their toy) and an increase in explanations that appeal to belief state.

### Learning by contrast

How might we make the access variable more salient to our young learners? One factor that might facilitate learning is contrast (for review, see Gentner, 2010). Comparisons help learners identify differences and similarities between concepts, making salient the relevant variables and causal connections between them (Gentner, Loewenstein, & Hung, 2007; Ming, 2009). For example, learning by contrast has helped adults discriminate between mathematical problem types, identify the deep underlying structure of the problem as well as the critical structural features required to solve it (Ming, 2009), and it has helped children learn quickly from small amounts of data (Gentner et al., 2009).

Contrastive learning has been largely unexplored in the theory of mind domain. To our knowledge, no task has looked at children’s understanding of false-belief and whether contrast can help children reason about other minds (though see Gershon & Woodward, 2012 for an example of how comparisons help children learn goals of tool use actions). If contrastive learning is robust, theory of mind understanding may be facilitated by presenting a situation in which agents have contrasting access (one agent sees the object moved, whereas one agent does not see the object moved). This may help children identify visual access as a critical variable mediating beliefs, and may tip the scales in favor of the passers’ model, thus increasing predictions characteristic of that theory<sup>1</sup>. Favoring the passers’ model will also lead to an increase in explanations that appeal to belief and access, rather than simply desire (as seen in Goodman et al., 2006 when comparing failers’ explanations to passers’ explanations).

### Extension to the “false-belief” task

While false-belief understanding is a hallmark of theory of mind, as it seems to indicate a critical appreciation of the distinction between mind and world, it may over-simplify

---

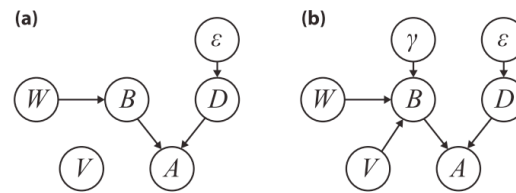
<sup>1</sup> The potential “boost” from contrast in recognizing the importance of visual access may only help children at the cusp of theory of mind understanding, who may already be (implicitly) weighing between both theories.

(and thus obscure) developmental change. Wellman and Liu (2004) designed a more nuanced conceptual scale that captures preschooler's developmental progression of theory of mind understanding. Thus, besides using age as a general marker, one way to gauge where a child lies on the spectrum of theory of mind understanding is to adopt Wellman and Liu's (2004) false-belief battery, and see where each participant scores on the developmental progression.

We adopt the full battery and use an element of the battery that tests false-belief understanding, the change-of-contents task, to control for false-belief ability when assigning participants to our experimental conditions and so we do not give children training on the change-of-location task used as our test. According to the Wellman, Cross, and Watson's (2001) meta-analysis, children who lack a coherent theory of mind have just as much trouble reasoning about their own false-beliefs as they do reasoning about others' false-beliefs. In the "self" change-of-contents tasks, a child is shown (e.g.) a band-aid box and then asked what they think is inside. Most children reply band-aids, as that is what it appears to contain. The box is then opened, revealing (e.g.) a key, rather than band-aids inside. The box is closed, and the child is asked "What did you think was inside the box before we opened it?" Three-year-olds often say key, rather than their true initial guess (band-aids), unsuccessfully ascribing false-beliefs to themselves. In the "others" version of this task, they also incorrectly ascribe the correct belief about the box contents to a new individual that should be naïve of the contents. This false-belief task, according to Wellman, Cross, and Watson's (2001) meta-analysis, is comparable to the change-of-location task—children who succeed in one will succeed in the other. The change-of-contents false-belief task (self and others) will provide an initial measure, independent of our modified test books, that we can use to classify children's starting false-belief understanding.

### Contrasting access in false-belief experiment

In this experiment, children at the cusp of the false-belief transition (three-and-a-half- to -five-year-olds) are first given a diagnostic assessment using Wellman and Liu's (2004) false-belief tasks, sans the change-of-location task. We then present children with one of two modified false-belief stories. In one version of the story, there are two characters (Sally and Billy) that together put their toy in a basket; one character leaves the room while the other stays and observes the toy being moved by a third character (Alex). The question posed to the children is where will Sally and Billy each look for the toy. After generating a prediction, children see that Sally goes to the original location and Billy goes to the new location and children are asked why each character looked there. The control task is nearly identical except both Sally and Billy leave the room and do not observe Alex moving the toy, and after the prediction phase both characters go to the original location and children are asked why Sally and Billy went there.



**Figure 1:** Goodman et al. (2006) model of (a) failers and (b) passers. Belief (B) & desire (D) determine action, but only in (b) is belief influenced by visual access (V). The parameter  $\gamma$  represents all the reasons, outside of the story, that an agent might change her mind.

If the models proposed by Goodman et al. (2006) accurately capture a relevant change (the addition of access in the passers' model) in children's theory of mind, then the contrasting access in the first book may make this variable more salient to children and may improve their performance on the task. Another possibility is that children's shifting ability in false-belief tasks is best explained by changes in executive functioning. If so, then increasing the task demands as with the contrasting access book (children have to track the multiple and different knowledge states of two characters) should make the task more difficult and lead to more incorrect predictions and explanations.

## Methods

### Participants

62 children (mean: 50 months, range: 38 to 65 months) were recruited from preschools and museums in the Berkeley area. Children were assigned to a *Contrast* or *Non-contrastive Control* condition based on their performance on the change-of-contents tasks, such that there were equal numbers of passers, failers, and children who answered 1 of 2 parts of the task correctly, assigned to each test condition.

### Procedure

Children were first given the diagnostic tests and then assigned to one of the two test conditions.

**Diagnostic tests.** We adapted the Wellman and Liu (2004) scaling of theory of mind tasks to include the following six tasks (in the order listed): Diverse Desires (i.e. can George like this even if you don't); Diverse Beliefs (i.e. if Linda thinks it is there but you think it's here, where will Linda look?); Knowledge Access (i.e. given this opaque box that Molly can't see inside, will she know what's inside?); change-of-contents self (i.e. what did you think was in the Band-Aid box when you first saw it?); change-of-contents other (i.e. what will this new person think is inside the Band-Aid box?); Belief-Emotion (i.e. will Sam be disappointed when he looks in the box?). Children were then assigned to one of the test conditions based on their performance on the change-of-contents self and change-of-contents other task such that equal numbers of passing and failing children were assigned to each condition.

**Test books.** The test books were the modified change-of-location story previously described. In the story, Billy and Sally have a stuffed bear. They hide their stuffed bear

underneath the lid of a basket. In the *Contrast* condition Sally has to leave the room. When Sally leaves the room, and Billy is in the room, a new character – Alex – is introduced. The experimenter says to the child, “Uh, oh. Here comes Alex. Alex is a troublemaker. Look, while Sally is away, Alex moves their toy from the basket to the box, and Billy sees Alex move their toy. See! When Alex moves the toy, Sally is not in the room, but Billy is in the room and sees!” There is a memory check: “Where is the toy now?” (box). Then the child is asked to predict where Sally will first go to look for her toy (if she went to look before Billy), and where Billy will first go to look for his toy (if he went to look before Sally); the order of questions was counterbalanced. Children were prompted to provide an answer or point to the location. The child is then shown where Sally and Billy actually go: in the *Contrast* condition Sally goes to the basket and Billy goes to the box, and then the child is asked to explain freely why Billy and Sally went to their respective locations, despite the true location of the toy. Once the child explains why Sally and Billy acted in the way that they did, the child must pass the final memory check in order to be included in the study: “Can you point to who was in the room, if anyone, when Alex moved the toy?” No child was excluded.

In the *Non-contrast* condition, Sally and Billy both leave the room. The experimenter then says, “Oh, oh. Here comes Alex. Alex is a troublemaker. Look, while Sally and Billy are away, Alex moves their toy from the basket to the box, and Billy and Sally don’t see Alex move their toy. See? When Alex moves the toy, Sally is not in the room, and Billy is not in the room, and they both do not see!”<sup>2</sup> Like the *Contrast* condition, there is a memory check, and then the child is asked to predict where Sally and Billy will first go to look for their toy. However, in the *Non-contrast* condition, the child is shown that both Sally and Billy go to the basket and then asked to explain freely why “they both went there.” In both conditions, if the child only provided an explanation for one character, he or she was asked to explain the behavior, in isolation, of the other character.

## Results

Children were divided into conditions such that there were 31 children in the *Contrast* condition ( $M = 50$  months,  $SD = 7.52$  months) and 31 children in the *Non-contrast* condition ( $M = 51$  months,  $SD = 8.15$  months) with no significant difference in ages between conditions  $t(61) = -0.34$ ,  $p = 0.73$ . Children were divided into a condition in order to control for false-belief ability: in each condition, there were three groups of children based on their change-of-contents performance—12 children who passed (passers) by correctly ascribing a false-belief to themselves and to another agent, 13 children who failed (failers) who were unable to ascribe a false-belief to either themselves or

another agent, and 6 children who we deem “liminal” because they were able to ascribe a false-belief in one case, but not in another case.

### Predictions by change-of-contents performance

In order to pass the predictions portion of the task, the child needed to correctly predict both where Sally will go to look for her toy and where Billy will go to look for his toy. Any other kind of prediction (e.g. correctly predicting where Billy goes, but incorrectly predicting where Sally goes) was scored as failing the task.

The children in the *Contrast* condition did significantly better in predicting both Sally and Billy’s actions than in the *Non-contrast* condition, Fisher exact  $p = .04$ , (see Figure 2a). In the *Non-contrast* condition: 13% incorrectly predicted where both Sally and Billy will go, 39% correctly predicted where both Sally and Billy will go, and 48% made one correct prediction about Sally or Billy (7 of 15 correctly predicting Sally). In the *Contrast* condition: 16% incorrectly predicted where both Sally and Billy will go, 65% correctly predicted where both Sally and Billy will go, and 19% made one correct prediction about Sally or Billy (half correctly predicting Sally).

### Predictions by complete diagnostic score

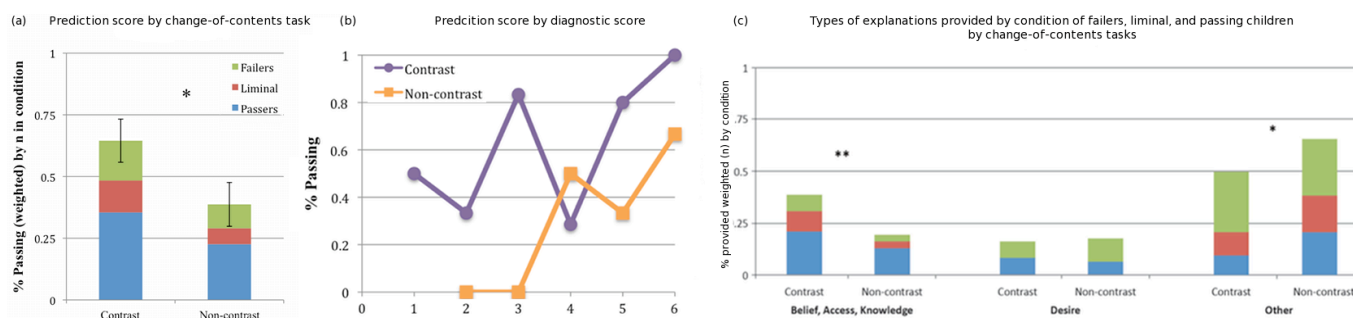
Participant responses were also analyzed as a factor of the number of tasks children initially passed on the six-task, diagnostic battery. In the *Contrast* condition, participant scores ranged from one to six, and in the *Non-contrast* condition participant scores ranged from two to six. Children in the *Non-contrast* condition had a marginally significant higher diagnostic score than the children in the *Contrast* condition,  $t(60) = -2.61$ ,  $p = .11$ , but this works against our hypothesis since children with higher scores on other theory of mind tasks should have higher false-belief performance.

There was an overall interaction between diagnostic score, condition, and passing,  $G^2(13) = 25.28$ ,  $p < .05$ , (see Figure 2b). Consistent with the predictions by change-of-contents performance, there was an interaction between condition and whether or not the child passed the predictions such that children in the *Contrast* condition were significantly more likely to correctly predict the agents’ actions,  $G^2(1) = 4.58$ ,  $p < .05$  (see Figure 3), even when the influence of the diagnostic score was removed  $G^2(5) = 15.58$ ,  $p < .05$ . There was also an effect of diagnostic score and passing; not surprisingly, children with higher diagnostic scores were significantly more likely to correctly predict where the agents would go in the false-belief tasks  $G^2(4) = 8.72$ ,  $p = .07$ .

### Explanations

Explanations were collapsed into three categories: “Belief, Access, Knowledge explanations” (which contained any explanation that appealed belief, access, or knowledge), “Desire explanations” (any explanation that appealed to desires), and “Other” (all non-mental explanations: an

<sup>2</sup> In both conditions, the Change-of-location story was modified to explicitly mention visual access (seeing) two times in order to emphasize who sees and who does not see the character move the toy.



**Figure 2.** Preschoolers' performance by condition based on (a) change-of-contents task (% predicting correctly), (b) diagnostic score (% predicting correctly), and (c) types of explanations generated by children based on change-of-contents task. So as not to conflate within group proportions, the percentages reported in (a) and (c) are the product of the percent responding in a particular group (e.g. 4/6 liminal Contrast children passed) weighted by the number of children in that group (e.g. 6/31).

explanation that appealed to the initial world, the final world, external information, or was unclassifiable). The first author coded all explanations and a research assistant reliability coded 82% of explanations. Coding was compared using a maximally conservative approach (using the 8 classification subdivisions) and both coders were blind to condition; reliability was high ( $\kappa = 0.79$ ). Analysis revealed an overall interaction between condition, explanation type, and initial false-belief performance,  $G^2(12) = 38.6$ ,  $p < .001$ . Consistent with Goodman et al.'s (2006) results, we also saw an interaction between explanation type and false-belief performance such that the better children performed on the diagnostic tasks, the more likely they were to appeal to belief, access, and knowledge in their explanations  $G^2(4) = 29.5$ ,  $p < .0001$ . Follow-up comparisons revealed that participants in the *Contrast* condition significantly more often appealed to beliefs, access, and knowledge than participants in the *Non-contrast* condition, Fisher exact  $p = .02$ . Furthermore, participants in the *Non-contrast* condition were more likely to appeal to "other" non-mental explanations, Fisher exact  $p = .05$ , (see Figure 2c).

## Discussion

Consistent with our predictions, results suggest that children are better able to predict and explain human behavior when observing agents who have contrastive visual access. Our modified false-belief tasks were more difficult in that they involved more characters, predictions, and explanations as compared to original change-of-location tasks. The *Contrast* condition was in a sense more complex than the *Non-contrast* condition because the child had to entertain two different perspectives, make two distinct predictions based on those perspectives, and then generate an explanation that accounts for these different perspectives and resulting behaviors; the *Non-contrast control* condition effectively required something much simpler: reason about Billy in the same way you reasoned about Sally. Nonetheless, predictive success in the *Contrast* condition was significantly higher than in the *Non-contrast* condition.

Explanatory success (where success in explanation requires appealing to higher level mental states such as beliefs and access) was also higher in the *Contrast* condition. These findings further support the claim that highlighting the access variable by contrast facilitates the child's ability to incorporate it into his or her model of other minds, and in turn, use it to explain behavior.

There are numerous alternative accounts for three-year-olds failures to pass false-belief tasks. Executive functioning involves planning, response inhibition, and cognitive flexibility (Zelazo, Carter, Reznik, & Frye, 1997). Thus, the ways in which executive functioning may potentially interact or interfere with theory of mind performance is vast (e.g. premature inhibitory control see Carlson, Moses, and Hix, 1998; for theory of mind Mechanism/Selection Processing see Scholl and Leslie, 2001). This kind of account is inconsistent with our findings. Because age and initial false-belief performance are controlled, executive functioning (such as inhibitory selection abilities) in the children should also be comparable between conditions. There is no reason to assume that the children in the *Contrast* condition developed greater inhibitory control than their *Non-contrast* counterparts. The *Non-contrast* condition did require the child inhibit the real world contents twice in order to pass the task – they must ascribe a false-belief to both Billy and Sally in the story. However, according to Wellman, Cross, and Watson's (2004) meta-analysis children given consecutive and equivalent false-belief tasks, which varied only in character and object, responded in highly consistent ways, giving identical responses 84% of the time. This finding suggests that asking a child to ascribe a false-belief twice, given equivalent situations, but differing in character, does not encourage switching answers. A child who is able to inhibit the real world contents once should be able to inhibit the real world contents a second time.

The study illustrates the helpfulness of presenting multiple and contrasting perspectives to a young learner with a developing theory of mind. Contrast has been shown to facilitate learning. However, until now, it has not been applied to the social domain. How contrast is helping may be, as Gentner (2010) proposes, a matter of highlighting a

variable and making it available for learning. Children appear to be incorporating the access variable into their working theory of mind, and contrast may be a means to facilitate this understanding. More generally, the results are consistent with a theory theory account of theory of mind: children seem to theorize others' inner workings, others' ambiguous actions and other's intangible beliefs, adjusting their theories in light of more information. The results support the theory that understanding other minds is learned just as any other theory is learned, and contrast can be applied to help interpret the ambiguous actions of others in the complex context of reality.

## References

- Atington J, Jenkins J. (1995). Theory of mind development and social understanding. *Cognition and Emotion*, 9:151–165.
- Baillargeon, R., Scott, R.M., He, Z. (2010) False-belief understanding in infants. *Trends in Cognitive Science*, 14(3), 110–118.
- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a 'theory of mind'? *Cognition*, 21, 37–46.
- Berthiaume, V., Shultz, T., & Onishi, K. (in review) A constructivist connectionist model of developmental transitions on false-belief tasks.
- Carlson, S. M. and Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, 72, 1032–1053.
- Carlson, S. M., Moses L. J., & Hix, H. R. (1998). The role of inhibitory processes in young children's difficulties with deception and false-belief. *Child Development*, 69, 672–691.
- Gentner, D., Loewenstein, J., & Hung, B. (2007). Comparison facilitates children's learning of names for parts. *Journal of Cognition and Development*, 8, 285–307.
- Gentner, D., Levine, S., Dhillion, S., & Poltermann, A. (2009). Using structural alignment to facilitate learning of spatial concepts in an informal setting. In B. Kokinov, K. Holyoak, & D. Gentner (Eds.), *Proceedings of the second analogy conference* (pp. 175–182). Sofia, Bulgaria: NBU Press.
- Gentner, D. (2010). Bootstrapping the Mind: Analogical Processes and Symbol Systems. *Cognitive Science*, 34(5), 752–775.
- Gershon, S. & Woodward, A. (2012). A claw is like my hand: Comparison supports goal analysis in infants. *Cognition*, 122, 181–192.
- Goodman, N.D., Baker, C.L., Bonawitz, E.B., Mansinghka, V.K., Gopnik, A., Wellman, H., Schulz, L., & Tenenbaum, J.B. (2006). Intuitive theories of mind: a rational approach to false-belief. *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.
- Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, 16, 1–14.
- Gopnik, A. & Wellman, H. M. (1992). Why the Child's theory of mind Really Is a Theory. *Mind & Language*, 7: 145–171.
- Lalonde, C.E., & Chandler, M.J. (1995). False-belief understanding in school. On the social-emotional consequence of coming early or late to a first theory of mind. *Cognition and Emotion*, 2, 167 – 185.
- Ming, N. (2009). Analogies vs. contrasts: A comparison of their learning benefits. In B. Kokinov, K. Holyoak, & D. Gentner (Eds.), *Proceedings of the second international conference on analogy*. (pp. 338–347). Sofia, Bulgaria: NBU Press.
- O'Loughlin, C., & Thagard, P. (2000). Autism and coherence: A computational model. *Mind & Language*, 15(4), 375–392.
- Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.
- Perner, J., Leekham, S., & Wimmer, H. (1987). Three-year-olds' difficulty with false-belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, 5, 125–137.
- Scholl, B.J., & Leslie A.M. (2001). Minds, Modules, and Meta-Analysis. *Child Development*, 72(3), 696–701.
- Schulz, L., Bonawitz, E.B., & Griffiths, T. (2007) Can being scared give you a tummy ache? Naïve theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*, 43(5), 1124–1139.
- Triona, L. M., Masnick, A. M., & Morris, B. J. (2002). What does it take to pass the false-belief task? An ACT-R model. *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (p. 1045). Mahwah, NJ: Lawrence Erlbaum Associates.
- Van Overwalle, F. (2010). Infants' teleological and belief inference: A recurrent connectionist approach to their minimal representational and computational requirements. *NeuroImage*, 52(3), 1095–1108.
- Watson, A. C., Linkie Nixon, C., Wilson, A., & Capage, L. (1999). Social interaction skills and theory of mind in young children. *Developmental Psychology*, 35, 386–391.
- Wellman, H. M., & Liu, D. (2004). Scaling of theory of mind tasks. *Child Development*, 75, 523–541.
- Wellman, H. M., & Liu, D. (2007). Causal reasoning as informed by the early development of explanations. In A. Gopnik & L.Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 261–279). New York, NY: Oxford University Press.
- Wellman, H. M., Cross, D. and Watson, J. (2001). Meta-analysis of Theory-of-Mind development: The truth about false-belief. *Child Development*, 72, 655–684.
- Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13, 103–128.
- Zelazo, P. D., Carter, A., Reznick, J. S., & Frye, D. (1997). Early development of executive function: A problem-solving framework. *Review of General Psychology*, 1, 198–226.



# Children and Pragmatic Implicatures: A Test of the Pragmatic Tolerance Hypothesis with Different Tasks

**Katrijn Pipijn (Katrijn.Pipijn@Ppw.Kuleuven.Be)**

Laboratory of Experimental Psychology, Tiensestraat 102 Bus 3711  
3000 LEUVEN, Belgium

**Walter Schaeken (Walter.Schaeken@Ppw.Kuleuven.Be)**

Laboratory of Experimental Psychology, Tiensestraat 102 Bus 3711  
3000 LEUVEN, Belgium

## Abstract

The pragmatic tolerance hypothesis (Katsos & Smith, 2010) was originated to explain the difference between children and adults concerning scalar implicatures. They introduced the use of a Likert-scale to test this hypothesis. We conducted a study with a within subjects design in which we compare children's binary and scalar responses to the same underinformative sentences. We also used two separate tasks to look at the effects of task difficulty on performance. The results show that the more difficult task, Euler circles, lead to less pragmatic responses compared to the easier task, drawings. Confirming the study by Katsos and Smith (2010; see also Katsos & Bishop, 2011) children choose the middle options on the scale more when they are confronted with underinformative sentences and they choose more extreme options for the control sentences. The comparison with the binary responses however, reveal that the link between the two measuring methods is not as straight forward as we would think.

**Keywords:** Scalar Implicatures; underinformative sentences; children; scalar responses pragmatic tolerance.

## Introduction

Communication is not always as straightforward as one might think. In 1989 Grice published his work on the cooperative principle that was meant to explain how our human interaction can be described. The cooperative principle expects a person to interact in a way that furthers the purpose of the conversation and indicates that a second person expects the first person to do so. The cooperative principle allows for implicatures to be used. When a person uses an implicature, the meaning of what that person says is not explicitly communicated, but can nonetheless be derived from what he says. The utterance is under-informative, more information could have been given but has not. For example when a wife asks her husband whether he'll be home for supper, and the husband answers that he has a meeting that will run late that day, then the husband is using an implicature. His wife will not expect him for dinner. One can assume that she accepts the meeting running late will be the reason, or at least a possible reason, that the husband will not be present at dinner. Nevertheless it is still possible that the husband will appear for dinner, for the implicature is cancellable. It is possible that the husband just meant he would be a little late for dinner, still he would not have lied in his earlier utterance.

One specific form of implicatures are scalar implicatures, which we will focus on in this paper. As the name implies, scalar implicatures consist of words that can be situated on a scale, known as Horn scales (see Horn, 1984). These words range from less informative to more informative, for example a scale containing words like <none>, <some> and <all>. Each word further on the scale contains more elements of a group. When a speaker uses a certain less informative word in an utterance, it is implicated that the more informative word is not applicable. When a person uses the word 'some', the word 'all' would not be appropriate. It is considered a mutual understanding between speaker and recipient that the speaker would have used the more informative word if it were suitable. Nevertheless he deliberately chose to use the less informative word on the scale therefore the more informative is not suitable. For example when the prime minister says 'Some banks are collapsing due to the financial crisis', a citizen can assume that 'not all' banks are collapsing due to this crisis, for the expression of 'some' implicates 'not all'. The citizen presumes that the prime minister would have said 'All banks are collapsing due to the financial crisis' if this were the case. If a few months later the prime minister makes the announcement 'All the banks have collapsed due to the financial crisis', this would not be a withdrawal of his earlier statement. Specific to implicatures is that they are cancellable in only one direction. When a speaker uses the weaker term 'some', it can later be easily corrected to 'all'. Yet when a speaker initially uses the stronger term 'all', it is not possible to change it to 'some' later on. At least not without admitting one was erroneous the first time. The stronger term 'all' entails the weaker term 'some' but not vice versa.

When a speaker uses the word 'some' in an utterance, there are two different ways to interpret this weak scalar term. The first way is the pragmatic way that was described above. A recipient might produce a scalar implicature and assume that the speaker meant 'some and not all' with the statement. Yet another way of interpreting the word 'some' is a purely explicit logical interpretation. The explicit meaning of the word 'some' is 'at least one and possibly all'. Both interpretation of the word are equally correct and it is the choice of the recipient on how he will interpret it.

Further in this article, we will refer to scalar implicatures as underinformative items or sentences.

We already know from different studies that children and adults interpret underinformative sentences in alternative ways. Noveck (2001) argues that a weak scalar term is understood in its explicit meaning first and will appear first in human development. Only later on the more complex pragmatic meaning will be incorporated. This argument is clearly demonstrated by the results of Noveck's study (2001). He found how children of 7-8 years old and 10-11 years old have acceptance rates of 89% and 85% for sentences that are logically true but pragmatically infelicitous. Adults on the other hand, accept these sentences in only 41% of the cases. This clearly demonstrated how for children the pragmatic meaning of these sentences is not incorporated. While for adults these pragmatic meanings are fully incorporated and are used as the principal criteria to accept or reject sentences.

The results also show how these differences between children and adults cannot be explained by the children's limited understanding of words like 'some' and 'all'. For all the different utterances that do not hold a conflict between the logical and the pragmatic meaning, the answering patterns of children and adults are very alike. The reason for the discrepancy between children and adults is not entirely clear. Noveck explains this by the posterior development of the pragmatic understanding of underinformative sentences. The processing of the pragmatic meaning of underinformative sentences is also cognitively much more demanding than the processing of the logical meaning (De Neys & Schaeken, 2007). Because of this, the pragmatic interpretation is harder to incorporate for children. Another factor that contributes to this is the nature of the task.

Pouscoulous et al. (2007) reported experiments in which they changed the nature of the task from verbal judgments to action-based judgments. Using small boxes that contained tokens, participants were asked to alter the setting of the tokens to match a statement. They were also allowed to leave a setting as it was. Within the experimental design, children's performance on producing implicatures was much higher than in experiments with verbal judgments. This increased implicature production was found for all ages (4-, 5-, and 7-year-olds as well as adults). Still, the developmental effect was present. These experiments show how the understanding of implicatures can be facilitated in young children by changing task features. Other studies have also showed how changing task features can facilitate children's performance (Guasti et al., 2005; Papafragou & Musolino, 2003; Papafragou & Tantalou, 2004).

Katsos and Smith (2010) did research on underinformative sentences in children and adults. They raised the pragmatic tolerance hypothesis to explain for differences between children and adults as well as differences between adults. The starting point of this hypotheses is that there are different degrees of violations. Several violations can lie within an utterance yet not every violation is equally grave. Participants can and will reject

utterances that are a grave violation of the logical truth. Yet they might accept or reject an utterance that only holds a violation of informativeness and thus is an infringement of the cooperative principle. There is no implicit rule on how to deal with pragmatically infelicitous utterances. The threshold of what is and what is not acceptable is individual for each person and is called pragmatic tolerance by Katsos and Smith (2010).

An obvious way to test this hypothesis was adopted by Katsos and Smith (2011, also see Katsos and Bishop (2011) and Katsos et al (2011)). Katsos and Smith (2010) introduced the use of a Likert scale to the research on underinformative sentences. A Likert scale is a bipolar psychometric scale on which a participant can indicate to what extent he agrees or disagrees with a certain statement. Katsos and Bishop (2011) made their participants indicate how much they agreed with utterances containing the words 'some' and 'all'. Both children and adults clearly rejected utterances that were inherently false and accepted utterances that had an optimal use of the words 'some' and 'all'. Interestingly, for the underinformative utterances, the answering patterns for children and adults were also very similar, as both groups chose the middle option on a 3-point Likert scale. This is in strong contrast with Noveck (2001) where the answering patterns for children and adults were much more distinct, notwithstanding the children in this study were older. Katsos and Smith (2011) explain this with the pragmatic tolerance principle. Children appear to be competent pragmatic comprehenders. They do sense the pragmatic violation when underinformative sentences are used. Yet due to their different tolerance levels, they do not experience this violation to be grave enough to be rejected. Therefore, when they are confronted with a two alternatives forced choice, they will not reject the violation while adults will.

In this paper, we want to explore these results more thoroughly and make three hypotheses. First of all, we will vary the task method. Pouscoulous et al. (2007) and others taught us that the nature of the task is of great importance. We expect that when we use different tasks, we will be able to make children reason more or less pragmatic, depending on the task difficulty. We will apply different methods than those used in Katsos and Bishop (2011) and Katsos and Smith (2010). Earlier research on underinformative sentences used different methods than the current ones. For example Newstead (1989, 1995) used Euler circles in his research. This abstract testing method should be difficult for children and thus induce more logical reasoning. We also developed a more child-friendly method using drawings which should induce more pragmatic reasoning in children.

Our second hypothesis concerns pragmatic tolerance. It seems obvious that this theory should be examined with a within subjects design in which children are confronted with a Likert scale as well as with the two alternative forced choice paradigm. We expand the testing method used in Katsos and Bishop (2011). Participants will be confronted with each underinformative sentence twice, once with the



option of responding on a Likert scale or once with a two alternative forced choice. With this research we expect to replicate Katsos and Bishops (2011) findings, namely that children do seem to detect a conflict when they are confronted with underinformative sentences. We expect that this conflict detection will be hidden when confronted with a two alternative forced choice but will become clear when they are confronted with the Likert scale. We will use children around the age of eleven, congruent with Noveck (2001). According to this study we expect children of this age to be still much more logical than adults.

Finally, we will look at consistency in children's answers. We expect that children that answer logically or pragmatically with the scalar measuring method, will answer in the same direction with the two alternative forced choice measuring method.

## Method

Twenty-two Dutch speaking children participated in this research (mean age 11,3 range 11-13).

The children received a pen and paper test. The test started with a cover-up story about a boy named Thomas. The children were told that Thomas was new in class and came from a foreign country. They were told he was still learning the Dutch language and the children were to indicate how precise his answers were. Children had to indicate their answers either by indicating right or wrong, or on a 5-point Likert scale. The ends of the Likert scale were illustrated with a happy smiley and a frowning smiley. On the scale, the children were to indicate how well they thought that the boy's answer was, going from completely wrong to completely right. They were also allowed to use the middle options when the answer was only a little right or wrong or evenly right and wrong.

Two different tests were used. Both tests had the same basic structure. We started each trial with a given situation. This situation was presented either by a figure or a drawing. Then the participants were given a statement about the situation. They were instructed to indicate how well the statement described the situation given above.

First was the Euler circles task. The circles for each figure were either completely overlapping, partially overlapping or completely disconnected. Each circle represented a group of blocks, for example 'red blocks', 'square blocks', which was written inside each circle. The participants received a statement about the blocks and had to judge how precise the statement described the circles setting. For an example of this, see Figure 1.

For the second task, we used a method which was more adapted to children, Drawings. For the given situation, the children were now shown a drawing of a real life setting, for example a few kids playing with a bow and arrows. Again the children had to judge a statement about the setting, e.g. 'Some arrows are shot in the bull's-eye'. Due to the more authentic stimuli, the task became much easier for children.

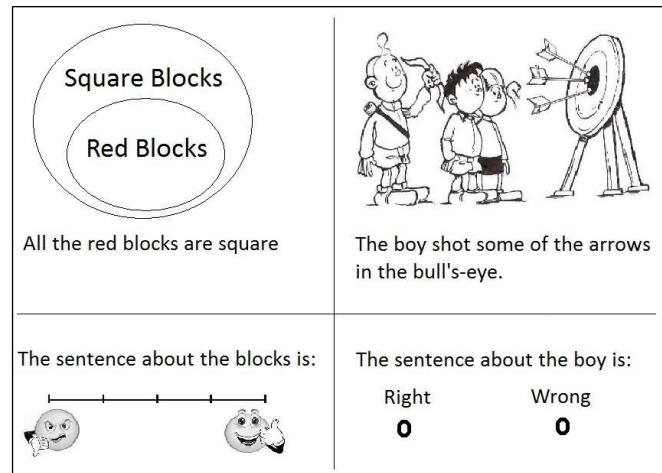


Figure 1: Example of Euler circles, drawings, scalar response option and binary response option.

## Results

We inverted all scores of the logically false items. This way, high scores on the control items, for both logically false items and optimal items, indicate competent reasoning. We also inverted answers on the underinformative items. Because of this, the maximal score of five points is an extreme pragmatic answer and the minimal score of one is an extreme logical answer. Finally we converted the binary zero and one scores to one and five scores to make them comparable with the scalar responses.

For the control items we found very high average scores, 4.72 (.20) for binary responses and 4.56(.34) for scalar responses. This means that the children understand the words 'some', 'all' and 'none' adequately. For the underinformative items, we found average scores of 3.93(.89) for the binary responses and 3.16(.98) for the scalar responses. For more detailed results, see Table 1.

Table 1: Mean ratings and standard error of the mean for Euler circles (EC) and Drawings (D)

	Binary	Scalar
EC – Control items	4.60 (.36)	4.46 (.36)
D – Control items	4.92 (.18)	4.65 (.41)
EC – Underinformative items	3.18 (1.51)	2.65 (1.05)
D – Underinformative items	4.70 (.57)	3.71 (1.11)

We ran a repeated measures design with three within factors with two levels each, namely measuring method, task and item type. We found three main effects. The two measuring methods levels, binary answers versus scalar answers, are significantly different from each other ( $F(1,21) = 9.46$ ,  $p < .01$ ). Binary responses are higher than scalar responses, as expected because binary responses only allow extreme answers. For the two tasks, Euler Circles seem to

be more difficult and lead to more logical answers than Drawings,  $F(1,21) = 54.07$ ,  $p < .00$ . For the item types, control items versus underinformative items, children answer more extreme for control items and more varied for underinformative items,  $F(1,21) = 54.72$ ,  $p < .00$ . We found two interaction effects. The interaction between measuring method and task was not significant but the other two interaction effects were, measuring method versus item ( $F(1,21) = 4.63$ ,  $p < .04$ ; see Figure 2) and task versus item ( $F(1,21) = 21.62$ ,  $p < .00$ ; see Figure 3). The three-way interaction was not significant.



Figure 2: Interaction between measuring method and item type.

We calculated the difference between the control items and the underinformative items for each measuring method. A paired-samples t-test on these values was significant ( $t(21) = 2.21$ ,  $p < .04$ ). This means that the interaction between measuring method and item type is explained by a difference in size of the effect of measuring method on item type.

The main effect of task and its interaction with item, mean that the Euler Circles were more difficult, especially for the underinformative items and thus lead to more logical answers. To confirm this, we calculated the difference between the control items and underinformative items for each task and analyzed with a paired t-test,  $t(21) = 4.65$ ,  $p < .00$ .

For the control items, 84% of the items were answered with an extreme answer of one or five on the scale. For the underinformative items, only 47% were answered with an extreme one or five. These two percentages were significantly different from each other ( $t(21) = 5.22$ ,  $p < .00$ ).

Finally we look at consistency of answers. We interpret being consistent between the two methods when a child gives an extreme answer of one or five on the scale and gives the equal binary response for the same item. For the control items, the children were fairly consistent between the two measuring methods. 80% of the children can be considered consistent under this rule. For the

underinformative sentences, children were much less consistent, only 33 % of them was consistent in their answers between the two methods. When we adopt a more flexible rule including also the two and four answers on the scale, which would also be acceptable, 87% and 57% of children can be considered consistent. For the underinformative items, 16% of the time the middle option of the scale was chosen.

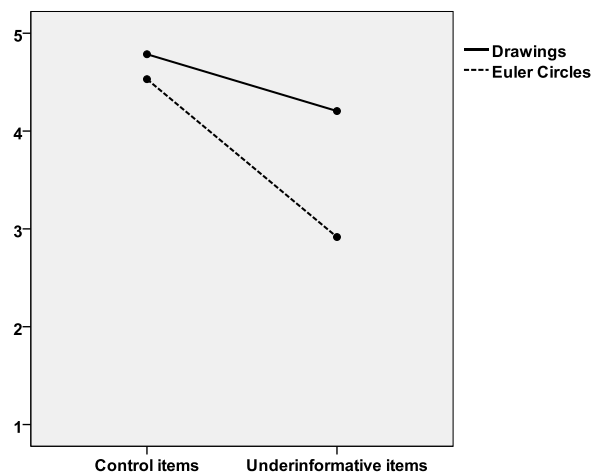


Figure 3: Interaction between task and item type.

## Discussion

In this study we examined three hypotheses. First of all, we expected that children's performance will depend on the task difficulty. More precisely, we expected the Euler circles to be more difficult than the Drawings task and to lead towards less pragmatic answers for the underinformative items. Next we expected to replicate Katsos and Bishops (2011) findings, namely that children answer extremely pragmatic or logical when confronted with control items but more doubtful when confronted with underinformative items and a scale. Finally we expected children to be consistent in their answers on the two different measuring methods.

For the first hypothesis, we can find confirmation in the main effects of task and the interaction between task and item type. The Euler Circles task is clearly more difficult than the Drawings task. For the control items this difference is small but significant. For the underinformative items, this difference becomes even larger. For the more difficult task, the Euler circles, this leads to more logical answers. For the easier task, the Drawings, children become more pragmatic. There still remains a significant difference with the control items though. We hereby can confirm what Pouscoulous et al. and others made us expect. Task features can influence children's pragmatic reasoning on underinformative sentences. We noted earlier that we expect task difficulty to be the determining factor here. Yet we acknowledge that another factor may be at work as well. The Euler Circles task is believed to rely on logical reasoning skills. It might

be possible that the logical interpretation is triggered by the general logical characteristics of the task. In this case, not task difficulty but the logical nature of the task would be the determining factor. More in depth research on the matter seems necessary. The tasks used in this study were also very adapted to usage with children. More grammatical approaches to the material might lead to different conclusions. If the grammatical view of scalar implicatures (e.g. Chierchia, 2006; Fox, 2007) is correct, then in principle the implicature-computing operator could also be inserted in embedded positions, thus giving rise to embedded scalar implicatures. Chierchia, Fox and Spector (de twee papers) argue that an implicature-computing operator can indeed be inserted in embedded positions. It would be interesting to see how our conclusions and those of Katsos and Smith (2010) and Katsos and Bishop (2011) could be incorporated into this grammatical approach.

Secondly, we found a significant effect of measuring method and an interaction with item type. The difference between binary answers and scalar answers for the control items is significant. But the difference between the methods becomes much larger for the underinformative items. This confirms our hypothesis and replicates Katsos and Bishop (2011). When confronted with a scale, children do feel that there is a conflict between the pragmatic and the logical interpretation of underinformative sentences. They tend to choose the middle options of the scale more often (53%) than when confronted with control items (16%). This rules out the possibility that children are just unfamiliar with the use of scales. They are adequate in using scales and it is a deliberate action to choose the middle options for the underinformative items and the more extreme options for the control items. This confirms the pragmatic tolerance hypothesis in that children use the scale to express that they feel the conflict between the logical and the pragmatic interpretation.

We do however find a difference with common literature. The children in this study seem to be much more pragmatic than reports from other studies, especially with the binary responses. One explanation for this is probably the children's ages. Much research on this topic used younger children than the ones used in this study. It is self-evident that the slightly older children used in this study would perform more pragmatically and adult-like. Moreover, the current study was conducted in Dutch. Previous unpublished research on underinformative sentences with Dutch speaking children, revealed that these children are more pragmatic than their English-speaking (Katsos and Bishop, 2011) or French-speaking (Noveck, 2001) counterparts. Dutch speaking children seem to be more comparable to Spanish speaking children for example. In a study by Katsos et al. (2011), Spanish-speaking children rejected pragmatically false underinformative statements in 87% of the cases. It seems that the Dutch word 'sommige' is not the exact equal of the English word 'some'. This will probably contribute to the high rate of pragmatic answers in Dutch-speaking children.

Finally we examined consistency. These results seem to differentiate from the earlier found results. The children were not very consistent in their answers. Especially for the underinformative items, children were consistent in only 57% of the cases and 16% they chose the middle option. This still leaves 27% of the cases where children were not consistent. This percentage seems rather high to us and it interferes with the pragmatic tolerance theory. In roughly one fourth of the times, children's binary responses and their responses on the scale are not related. On top of that and in contrast to the study by Katsos and Smith (2010), we found much larger variances for both the control items and the underinformative items. This all suggests that the link between binary answers and scalar answers is not a direct link. For control items and underinformative items, up to 19% of the answers were cases in which the children gave an exact opposite to answer the binary items and the scalar items. We can hypothesize that in these cases children just made a simple error and that this wasn't intentional or due to a lack of understanding. But there is no way to be sure of this and it is in contrast with high overall levels of performance.

In conclusion, our study mainly confirms the pragmatic tolerance hypothesis but it also questions some aspects of it. It is clear to us that the pragmatic tolerance hypothesis and the relationship between binary and scalar answers on underinformative sentences is not as straightforward and that more thorough research on the matter is necessary.

## Acknowledgments

This research was carried out with the financial support of the National Council for Scientific Research – Flanders, Belgium (FWO grant G.0634.09)

## References

- De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, 54, 128-133.
- Chierchia, G. (2006). Broaden Your Views: implicatures of Domain Widening and the "Logicity" of Language. *Linguistic Inquiry*, 37, 535-590.
- Chierchia, G., Fox, D., & Spector, B.. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. To appear in Portner, P., Maienborn, C. & von Stechow, K. (Eds.), *Handbook of Semantics*. New York, NY: Mouton de Gruyter.
- Chierchia, G., Fox, D., and Spector, B. (2009). Hurford's constraint and the theory of scalar implicatures: evidence for embedded implicatures. In Egré, P. & Magri G. (Eds.), *Presuppositions and implicatures. Proceedings of the MIT-Paris workshop*. MIT Working Papers in Linguistics 60.
- Fox, D. (2007). Free choice and the theory of scalar implicatures. In Sauerland, U. & Stateva, P. (eds.), *Presupposition and implicature in compositional*

- semantics*, 71–120. Houndmills, Basingstoke: Palgrave Macmillan.
- Grice, P. (1989). *Studies in the Way of Words*. Harvard University Press, Cambridge MA.
- Guasti, M.T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, 20, 667-696.
- Horn, L. (1984). Toward a new taxonomy for pragmatic inference. In Schiffrin, D. (ed.), *Meaning, Form and Use in Context: Linguistic Applications. Proceedings of GURT '84*. Washington D.C.: Georgetown University Press.
- Katsos, N., Andrés Roqueta, C., Estevan, R. A. C., & Cummins, C. (2010). Are children with Specific Language Impairment competent with the pragmatics and logic of quantification? *Cognition*, 119, 43–57.
- Katsos, N., & Bishop, D. V. M. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition*, 120, 67-81.
- Katsos, N., & Smith, N., 2010. Pragmatic Tolerance or a speaker–comprehender asymmetry in the acquisition of informativeness? In Franich, K., Iserman, K.M., Keil, L.L. (Eds.), *Proceedings of the 34th Annual Boston Conference in Language Development*, Cascadilla Press, MA, USA.
- Newstead, S. E. (1989). Interpretational errors in syllogistic reasoning. *Journal of Memory and Language*, 28, 78-91.
- Newstead, S. E. (1995). Gricean implicatures and syllogistic reasoning. *Journal of Memory and Language*, 34, 644-664.
- Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78, 165-188.
- Papafragou, A., & Musolino J. (2003). Scalar implicatures: experiments at the semantics/pragmatics interface. *Cognition*, 86, 253-282.
- Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition*, 12, 71-82.
- Pouscoulous, N., Noveck, I., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, 14, 347-376.

# Why is A Few Sometimes A Lot?

**Amanda Pogue (amanda.pogue@uwaterloo.ca)**

University of Waterloo, Department of Psychology,  
200 University Avenue West, Waterloo, ON N2L 3G1 Canada

**Adel Jalabi (ajalabi@uwaterloo.ca)**

University of Waterloo, Department of Psychology,  
200 University Avenue West, Waterloo, ON N2L 3G1 Canada

**Mathieu Le Corre (mlecorre@uwaterloo.ca)**

University of Waterloo, Department of Psychology,  
200 University Avenue West, Waterloo, ON N2L 3G1 Canada

## Abstract

It is not surprising to find that the quantity picked out by terms like *a few* and *a lot* is context dependent. We can easily accept that *a few* books might be 10 books, yet *a lot* of smartphones might only be 4 smartphones. The current paper posits that there are two hypotheses that can explain this context dependency: the Definite Number Hypothesis (DNH), and the Gradable Quantifier Hypothesis (GQH). The DNH suggests that the term *a few* corresponds to a definite range of values, and may pick out a larger quantity only if the range seems implausible for the given context. The GQH suggests that context-dependency is actually built into the meaning of *a few*. Experiment 1 supports the intuition that there is variability in the quantity that *a few* picks out based on context. The findings of Experiments 2 and 3 support the Gradable Quantifier Hypothesis.

**Keywords:** quantity-judgment; quantifiers; semantics; psycholinguistics; context

## Introduction

When we consider terms such as *a few* and *a lot*, we intuitively know that relative to each other *a few* will always be smaller than *a lot*; however, what does this tell us about how we should evaluate the quantity that they represent across different sets in the world? One potential strategy is to assume that *a few* is always a small quantity and that *a lot* is always a large quantity. However, this strategy may lead us astray. For example, in Experiment 1 we find that 19 English-speaking adults suggest that *a few* friends on Facebook® is about 76 friends, whereas *a lot* of houses owned by a single individual is approximately 4.3 houses. Given this finding, it seems highly unlikely that the aforementioned strategy is the best for describing how we evaluate these terms in the wild. How then might we be dealing with these terms?

There is no doubt that we must be using some sort of contextual information to guide our interpretations. But how do we do this? We constructed and consider two hypotheses. On the Definite Number Hypothesis, the meaning of *a few* corresponds to a definite range of values – namely, small subitizable quantities (2 to 4 or 5). It was predicted that this range might be determined by comparison to other similar quantity terms using pragmatics (see: Grice, 1989; Barner, Brooks, & Bale, 2011). Given

that adults could easily select other terms such as *a couple* or *a pair* to represent 2 items, and terms such as *several* or *a handful* to represent slightly larger quantities, it was predicted that people could interpret the term *a few* with relation to these quantity terms, and therefore give *a few* a definite range of 3 to approximately 5. However, when speakers use *a few* Xs in contexts where 3 to 5 Xs is an extremely atypical quantity of Xs (say, Facebook friends), hearers interpret what the speakers have said by making a post-hoc adjustment from the definite meaning of *a few* (i.e., 3 to 5) to the smallest plausible quantity, say about 70 for Facebook friends. On the Gradable Quantifier Hypothesis, the meaning of *a few* is inherently context-sensitive, much like the meaning of gradable adjectives like *small*, or *tall*. In other words, the range of quantities corresponding to *a few* cannot be fully specified in the abstract; it can only be specified given some particular context. For example, there is no sense in asking what degree of height corresponds to *small* in general. Rather, one must know what type of individual is being measured, and what is the typical height for that type of individual (see: Kennedy, 2007; Syrett, Kennedy, & Lidz, 2010). The same may be true for the range of quantities that correspond to *a few* (i.e., one may have to know what type of individual is being quantified, and what is the typical quantity in which such individuals are found).

Both hypotheses are consistent with the fact that the numerical value associated with *a few* varies with context. Therefore, that fact alone cannot adjudicate between them. However, they attribute this variability to different factors. According to the Definite Number Hypothesis, the variability of *a few* is a function of whether the quantities in the range of its meaning (i.e., 3 to 5) are plausible; if they are not, then one adjusts to the smallest plausible value for the context at hand. In contrast, according to the Gradable Quantifier Hypothesis, the likelihood of finding 3 to 5 individuals of a given type matters not at all. Rather, all that matters is the typical quantity in which these individuals cluster. In the current paper, we test these hypotheses in two ways. First, in Experiment 1 we confirm the intuition that there is variability in the quantities associated with *a few*. In Experiment 2, we ask whether one observes variability in the quantities associated with *a few*, even in contexts where

3 to 5 are highly plausible quantities. Finally, in Experiment 3, we ask whether speakers associate *a few* with various quantities when they have information about typical quantities, but no information concerning whether 3 to 5 are plausible quantities. Whereas the Definite Number Hypothesis predicts that there should be no variability in the quantities associated with *a few* in Experiments 2 and 3, the Gradable Quantifier Hypothesis predicts the opposite pattern of results.

## Experiment 1

In Experiment 1 we aimed to look at what quantities English-speaking adults attribute to the terms *a few* and *a lot* with regards to common every day scenarios. Participants were asked to estimate values of *a few* and *a lot* when they were used in context of various common real world situations. It was expected that participants' responses would be affected by context (the estimated average for each situation), such that the overall averages given for both *a few* and *a lot* would vary for each scenario based on the average.

### Methods

**Participants** Nineteen fluent English-speaking adults ( $M = 20.7$  years old, 15 females) were recruited from the University of Waterloo. Participants received partial course credit for participating in the study.

**Materials and Procedure** Participants were asked to fill out a questionnaire where they were required to estimate the values associated with a series of one-sentence scenarios. Twenty different scenarios were constructed. Fourteen of the scenarios were about types of things individuals might own / have (dogs, cats, cars, computers, bicycles, books, children, guitars, smartphones, houses, DVDs, shirts, pennies, and TVs), and the remaining 6 items were about things and events in the world (trees in a park, apples picked in an outing, goals scored in a season, emails received over a week, friends on Facebook, and photos tagged online). Half of the items were expected *a priori* to have very small averages (Small Quantity Items), and the other half were expected to have relatively higher reported averages (Large Quantity Items). Each of the scenarios were described by one sentence containing the target term *a few* (1), or *a lot* (2), and were paired with a question asking the participant to evaluate how many items were described in each scenario. Additionally, the questionnaire included items asking participants to evaluate what the average might be for each scenario (3).

- (1) Dan has a few friends on Facebook. How many friends do you think he has on Facebook?
- (2) Dan has a lot of friends on Facebook. How many friends do you think he has on Facebook?
- (3) What is the average number of friends that people might have on Facebook?

The questionnaire contained all 60 test items. Each of the terms (*a few*, *a lot*, and *average*) were presented in separate blocks. The order of the scenarios was counterbalanced across blocks, and the order of the blocks was counterbalanced across participants. Each block appeared on a separate page.

### Results

Participants were elicited to pick a number to answer each of the test items. Since they were not given specific numbers to choose from their responses were fairly variable. As a result, before conducting any analyses we ran a recursive outlier detector on the responses made by each participant per test item, and removed any data points that were deemed to be outliers by the PJ Outlier program (Van Selst & Jolicoeur, 1994). Using this measure we deleted 3.6% of the responses. We then conducted a 2x2 repeated measures ANOVA with Quantifier (*a few*, *a lot*, and *average*) and Test Item as the repeated measures, and Subject Gender and Order as the between subjects measures. The results indicated an effect of Test Item ( $F(19,19) = 3.055$ ,  $p < .001$ ), and an interaction of Test Item and Quantifier ( $F(38,38) = 2.655$ ,  $p < .03$ ). Indicating that the results differ based on the individual test items, and that the quantifier used plays a role in the results for each test item. There was no effect of Quantifier, Subject Gender, or Order, nor any other significant interactions ( $ps > .1$ ).

We conducted several planned comparisons to investigate the interaction. As predicted there was a significant difference between the Small Quantity Items ( $M = 1.48$ ), and the Large Quantity Items ( $M = 73.57$ ) when the participants were asked to estimate the means for the items ( $t(18) = 9.89$ ,  $p < .001$ ). Similarly, we found a difference between the Small Quantity Items ( $M = 2.26$ ), and the Large Quantity Items ( $M = 20.63$ ) when the participants were asked to estimate what was meant by *a few* ( $t(18) = 6.4$ ,  $p < .001$ ), and a difference between the Small Quantity Items ( $M = 4.96$ ), and the Large Quantity Items ( $M = 264.34$ ) when the participants were asked to estimate what was meant by *a lot* ( $t(18) = 6.69$ ,  $p < .001$ ).

## Experiment 2

Both the Definite Number Hypothesis and the Gradable Quantifier Hypothesis can explain the variability found in Experiment 1. In Experiment 2 we ask if we can find contexts where having 3 to 5 items is highly plausible, but where people interpret *a few* as picking out quantities larger than 5.

Experiment 2 provided participants with contexts where the use of *a few* could plausibly be used felicitously to describe 3 to 5 items, or could also be used to describe a potentially larger quantity based on the participants' world knowledge. Participants were also asked to give a plausibility rating for scenarios such as in (4) which incorporated the numbers 3 to 5 rather than using the terms *a few* or *a lot*. The Definite Number Hypothesis predicts a very low plausibility rating for all contexts where *a few*

picks out a quantity greater than 5. Alternatively, the Gradable Number Hypothesis is more agnostic about the role of plausibility, but might predict that the all of the contexts where *a few* picks out quantities larger than 5 should be considered plausible.

(4) Martha has four friends on Facebook.

## Methods

**Participants** Forty-four fluent English-speaking adults ( $M = 20.4$  years old, 25 females) were recruited from the University of Waterloo. Participants received partial course credit for participating in the study.

**Materials and Procedure** Participants were asked to fill out a questionnaire where they were required to estimate the values associated with a series of one-sentence scenarios (Estimation Survey). Seven new scenarios were constructed. These scenarios included: silver cars in a parking lot, trips to the mall / gym / library in a semester, upper year students in a first year class, movies watched in a semester, and instances of eating pizza in a year. These items were chosen because it was predicted by the authors that they were contexts where *a few* could pick out a quantity larger than 5, but that they could plausibly be used in scenarios with quantities between 3 to 5. Unlike in Experiment 1, each of the scenarios was used in just a single trial that used the target term *a few*. The order of the scenarios was counterbalanced across participants.

Each participant completed a second survey following the Estimation Survey. The second survey (Plausibility Survey) included the 7 new scenarios introduced in Experiment 2, and 9 of the items from Experiment 1 (specifically items where *a few* picked out larger quantities). Each context was turned into a one sentence scenario which incorporated a number between 3 to 5 (e.g., “There are three silver cars parked in the UW parking lot today.”). Participants were asked to rate the plausibility for each of these sentences on a scale of 1 to 5, where 1 is highly implausible, and 5 is very plausible. The Plausibility Survey was always completed after the Estimation Survey, and was always presented on a separate page. The order of the scenarios was randomized between old and new items, and was counterbalanced across participants.

## Results

Table 1 reports the average plausibility ratings for all the test items rated in Experiment 2, and the average quantity associated with *a few* for each of these items.<sup>1</sup> The items are listed in decreasing order of plausibility. Table 1 shows that subjects associated quantities larger than 5 with several of

the items that were rated as having at least medium plausibility (3 out of 5 or higher). Of the items analyzed, the number of friends on Facebook was the only one where subjects associated a large quantity with *a few*, but where quantities between 3 and 5 were judged as having low plausibility. This provides evidence that *a few* is genuinely gradable.

Table 1: Mean responses given for “a few X” in Experiment 1 (noted by\*) and Experiment 2, and the mean plausibility rating for each corresponding trial in Experiment 2

Test Item	Plausibility of 3-5	A Few
E-mails in a week	4.42	8.39*
Trips to the mall in a semester	4.42	4.02
Trips to the gym in a year	4.28	8.18
Trips to the library in a semester	4.28	5.73
Upper year students in a first year class	4.19	10.61
Movies watched in a semester	4.19	6
Apples picked in an outing	3.94	8.28*
Books	3.89	10.16*
Photos tagged online	3.78	27.16*
Pennies in a jar	3.78	25*
Shirts	3.75	10.89*
Pizzas eaten in a years	3.75	9.85
DVDs	3.67	8.28*
Silver cars in the parking lot	3.19	8.1
Trees in a park	3.08	18.5*
Friends on Facebook	2.5	76.26*

## Experiment 3

Experiments 1 and 2 provide evidence that adults use context to adjust their interpretations of *a few* and *a lot*. Experiment 2 further showed that it is unlikely that this effect is as a result of the participants interpreting the use of the term *a few* as being infelicitous.

Experiment 3 provides further evidence that *a few* must be gradable, as predicted by the Gradable Quantifier Hypothesis. In Experiment 3 participants are told about novel objects that people from Southern Mexico like to collect. It was assumed that the participants had little to no experience with people from Southern Mexico or the things that they like to collect, thus, the participants could not use any of their prior knowledge to make inferences. Instead, we provided the participants with information about the average number of each of the objects owned by individuals, and then asked them to estimate how many items are picked out by *a few* or *a lot* for each of the novel items. Since the participants have no prior experience with the objects, and since we did not provide them with any information about the plausibility of individuals only owning 3 to 5 of the items, participants must rely on the information provided to them.

<sup>1</sup> Note: Similar to Experiment 1 participants were asked to make personal judgments and the responses varied greatly. A test for outliers (Van Selst & Jolicoeur, 1994) suggested that 6.82% of responses be removed due to being significant outliers. The data reported in Table 1 excludes the outliers in Experiment 1 and Experiment 2.



The Definite Number Hypothesis predicts that in Experiment 3, all participants would say that *a few* is 3 to 5 items regardless of the provided average information. Conversely, the Gradable Quantifier Hypothesis predicts that the number that *a few* picks out in Experiment 3 should vary depending on the context.

## Methods

**Participants** Sixteen fluent English-speaking adults ( $M = 21.6$  years old, 10 females) were recruited from the University of Waterloo. Participants received partial course credit for participating in the study.

**Materials and Procedure** Participants were seated at a computer table beside an experimenter. The study was composed of three parts: two training tasks, and the main task. First, the participants were introduced to the format of the task. They were instructed that they would be told some short stories. Accompanying these stories would be pictures displayed on the computer screen. They were instructed that when the stories were completed that they would be asked questions about the stories. In order to respond to these questions the participants were asked to make a selection between three cards visible on the screen. Participants would indicate the answer they thought was correct. Additionally, they were told that only two of the cards would have visible answers, and if they did not think that the correct answer was on the visible cards that they should choose the blank card.

In the first training task participants were given three trials where they were asked to locate specific items (e.g., toy car, teddy bear, 2 monkeys) using the three card response system, and received feedback on their choices.

In the second training task participants were introduced to a picture of a novel object, and were told that it was a kind of object that people in Southern Mexico like to collect. They were then shown a slide that looked like the top left box in Figure 1, and told (5). For the second training task the number of items shown on the screen was either 2 or 3. They heard three stories with three different individuals, each with the same number of items. Participants were then asked to guess how many of the novel objects a fourth individual might own, and were given the choice between the same number of items from the previous slides (2 or 3), and another amount (1 or 5). The purpose of this training trial was to determine if the participants could abstract average information from three short stories. There were two trials in the second training task. No feedback was given on these trials.

(5) This is Sarah, and she has this many Xs.

The test trials were identical to the second training task, except the design of the question response slide. On each trial participants were told stories about a novel object that people in Southern Mexico like to collect. They were told about three individuals, and were shown how many objects

each of those individuals owned. Each participant heard 4 stories, of which half of the stories were about characters who owned an average of 5 of the novel stimulus item (Small Quantity Context), and the other half were about characters that owned an average 40 of the novel stimulus item (Large Quantity Context). The novel stimuli items were always presented inside of a box shaped like a rotated card. The size of the box, and the size of the items remained constant throughout the story slides, and in the question response slide. It was assumed that visual cues such as the density of the items should provide evidence to indicate a difference between the quantities presented in the story slides, and those on the question response slides.

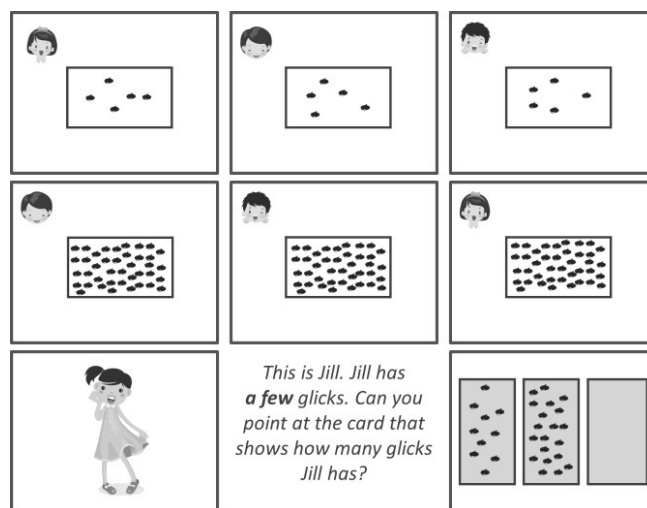


Figure 1: Example of Experiment 3 design. Row 1: an example of the slides that correspond with each of the three stories in the Small Quantity Context. Row 2: example of the slides that correspond with each of the stories in the Large Quantity Context. Row 3: the 4th character, the test question, and question response slide.

After hearing about how many items three characters had, participants were introduced to a fourth character and were told that the character “has *<a few / a lot>* of Xs.” After hearing this statement, participants were shown the question response slide with the three cards, and were asked to estimate how many of the novel objects they thought that the fourth character had. Participants had the option to pick between three cards: 10, 20, and blank, for each of the stories. If the participant chose the blank card they were asked to estimate how many items should be on the card. Each term was used once per context, and the order of the terms and contexts were counterbalanced across participants.

## Results

Figure 2 reports the percentage of response by card choice for each of the conditions in the task. As predicted by the Definite Number Hypothesis 93.75% of the participants chose the blank card in the Small Quantity Condition in the

*a few* trial, and reported that the quantity picked out for *a few* in this trial should be ~3.9. However, contrary to the Definite Number Hypothesis only 25% of the participants picked the blank card in the Large Quantity Condition, reporting that the quantity picked out for *a few* in this trial should be ~3.3. The remaining 6.25% and 75% respectively chose the small quantity card indicating that the quantity picked out for *a few* for these trials should be 10. Participants were significantly more likely to pick the blank card in the Small Quantity Condition for *a few* than in the Large Quantity Condition ( $\chi^2(1, N = 16) = 9.091, p < .01$ ). Additionally, participants were significantly more likely to pick the blank card in the Large Quantity Condition for *a lot* than in the Small Quantity Condition ( $\chi^2(1, N = 16) = 4.900, p < .03$ ).

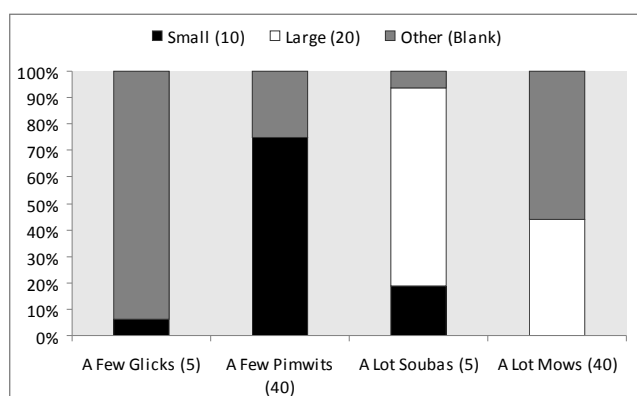


Figure 2: Percent per response type for each trial in Experiment 3.

## General Discussion

The results of the experiments presented in this paper provide direct evidence that there is variability in the quantities picked out by the term *a few*, and that this variability can be explained by context. This finding is highly unsurprising – no one would argue that context plays a role in the interpretation of all quantifiers. However, we set out to investigate more than just the effect that context has on the quantities picked out by *a few*, and endeavored to figure out how context is having this effect. We speculated that there are two hypotheses that could explain the role that context is playing for *a few*: 1) the Definite Number Hypothesis, and 2) the Gradable Quantifier Hypothesis.

According to the Definite Number Hypothesis, the variability of *a few* is a dependent on whether the quantities in the range of its meaning (i.e., 3 to 5) are plausible. If the quantities are not within the range of its meaning (like in the case of Facebook friends), then they must be adjusted to the smallest plausible value for the context at hand. Alternatively, according to the Gradable Quantifier Hypothesis, the likelihood of finding 3 to 5 individuals of a given type is not what matters, but rather the typical quantity in which these individuals cluster. These two

hypotheses made separate predictions regarding the outcomes of Experiments 2 and 3 in this paper.

In Experiment 2, the Definite Number Hypothesis predicted that participants would rate all of the contexts that for which quantities plausibly fall within its range as being highly plausible, and those for which quantities do not plausibly fall within its range as being implausible. The Gradable Quantifier Hypothesis makes no such prediction, and is not constrained in its meaning. Thus, it should have predicted that the plausibility ratings given by participants should not depend on the plausibility of the quantities falling within a specific range. The findings of Experiment 2 go against the predictions of Definite Number Hypothesis, as there was evidence for several items that do not fall within the range of *a few*'s definite meaning, yet were rated as being plausible scenarios. Though the prediction made by the Definite Number Hypothesis did hold true for the context involving friends on Facebook, but only for that one context. The results of Experiment 2 consequently seem to imply that people do not expect *a few* to have a definite range that requires post-hoc adjustments for larger contexts.

In order to further test these hypotheses, in Experiment 3 the only context information made available to participants was average information. Participants could not rely on their prior knowledge with the novel items to determine whether the definite range of quantities picked out by *a few* in the Definite Number Hypothesis is plausible for the novel contexts. Thus, the Definite Number Hypothesis predicted that participants should indicate that the quantity picked out by *a few* in all contexts should fall within the range of meaning (i.e., 3 to 5). On the other hand, the Gradable Quantifier Hypothesis predicted that *a few* would pick out a quantity relative to the typical quantity (i.e., average) for each context. Consequently, it would predict that the responses for the Small Quantity Condition would be different from the responses for the Large Quantity Condition. The results of Experiment 3 support the hypothesis predicted by the Gradable Quantifier Hypothesis, in that there was a significant difference in the responses given for the different context scenarios. Conversely the Definite Number Hypothesis only correctly predicted the responses for the Small Quantity Condition.

One potential limitation of Experiment 3 is that it is possible that while participants were not explicitly supplied with information that confirmed or denied the probability of 3 to 5 items being plausible quantities in the novel contexts, the participants may have inferred that the 3 to 5 would be implausible in the Large Context Condition. While this may be the case, when the results of Experiment 3 are considered with respect to the results of Experiment 2, which suggested that that we still see variability in the quantities picked out by *a few* despite 3 to 5 being plausible quantities for the given contexts, the Gradable Quantifier Hypothesis still provides a better explanation for the data.

The future direction of this research aims to investigate several questions left open by the findings in this paper. First, if the Gradable Quantifier Hypothesis is true, how

does context tell us what quantity is picked out by *a few*? The answer to this question can perhaps be found by determining the semantic form of *a few* and other gradable quantifiers with respect to the form of gradable adjectives. Secondly, it would be interesting to investigate whether all quantifiers are gradable like *a few*, or if it is the case that only some quantifiers are gradable in this way; it would be interesting to study what makes a specific quantifier gradable. Finally, the structure of gradable quantifiers seems like it would be very difficult to acquire. Given that both *a few* and *a lot* can be used in variable contexts to mean either a small quantity, or a relatively large quantity, how do children learn that what is important about gradable quantifiers is the context in which they are used, and not a definite number range?

The results of the present study support the Gradable Quantifier Hypothesis, suggesting that context-dependency is built into the meaning of terms such as *a few*, and rejects the Definite Number Hypothesis. This finding opens up a host of interesting questions about the gradability of other terms, and the acquisition of gradable quantifiers. Current research in the authors' lab aims to answer these questions with future and ongoing studies.

### Acknowledgments

Thanks to the members of the Developmental Science Lab and the members of the Developmental division in the Department of Psychology at the University of Waterloo for the useful feedback. A.P. would also like to thank Christie Haskell, Jonathan Fugelsang, and Hugh Rabagliati for the statistics advice. This research was supported by a grant to M.L. from the Natural Sciences and Engineering Research Council of Canada.

### References

- Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, 118, 87-96.
- Grice, P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.
- Kennedy, C. (2007). Vagueness and Grammar: The Semantics of Relative and Absolute Gradable Adjectives. *Linguistics and Philosophy*, 30, 1-45.
- Syrett, K., Kennedy, C., & Lidz, J. (2010). Meaning and Context in Children's Understanding of Gradable Adjectives. *Journal of Semantics*, 27, 1-35.
- Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size and skew on outlier elimination. *Quarterly Journal of Experimental Psychology*, 47A, 631-650.

# Toward machines that behave ethically better than humans do

Matthijs A. Pontier ([m.a.pontier@vu.nl](mailto:m.a.pontier@vu.nl))

Johan F. Hoorn ([j.f.hoorn@vu.nl](mailto:j.f.hoorn@vu.nl))

VU University Amsterdam, Center for Advanced Media Research Amsterdam (CAMErA),  
De Boelelaan 1081, 1081HV Amsterdam, The Netherlands

## Abstract

With the increasing dependence on autonomous operating agents and robots the need for ethical machine behavior rises. This paper presents a moral reasoner that combines connectionism, utilitarianism and ethical theory about moral duties. The moral decision-making matches the analysis of expert ethicists in the health domain. This may be useful in many applications, especially where machines interact with humans in a medical context. Additionally, when connected to a cognitive model of emotional intelligence and affective decision making, it can be explored how moral decision making impacts affective behavior.

**Keywords:** Cognitive modeling, Machine ethics, Medical ethics

## Introduction

In view of increasing intelligence and decreasing costs of artificial agents and robots, organizations increasingly use such systems for more complex tasks. With this development, we increasingly rely on the intelligence of agent systems. Because of market pressures to perform faster, better, cheaper and more reliably, this reliance on machine intelligence will continue to increase (Anderson, Anderson & Armen, 2005).

As the intelligence of machines increases, the amount of human supervision decreases and machines increasingly operate autonomously. These developments request that we should be able to rely on a certain level of ethical behavior from machines. As Rosalind Picard (1997) nicely puts it: “the greater the freedom of a machine, the more it will need moral standards”. Especially when machines interact with humans, which they increasingly do, we need to ensure that these machines do not harm us or threaten our autonomy. This need for ethical machine behavior has given rise to a field that is variously known as Machine Morality, Machine Ethics, or Friendly AI (Wallach, Franklin & Allen, 2010).

There are many domains where machines could play a significant role in improving our quality of life as long as ethical concerns about their behaviors can be overcome (Anderson & Anderson, 2008). This may seem difficult, and incorporating ethical behavior into machines is indeed far from trivial. Moral decision making is arguably even one of the most challenging tasks for computational approaches to higher-order cognition (Wallach, Franklin & Allen, 2010).

Moreover, with the increasing complexity of autonomous agents and robots, it becomes harder to predict their behavior, and to conduct it along ethical guidelines. Some may argue that this is a good reason not to let machines be responsible for making ethical decisions. However, the behavior of machines is still far easier to predict than the behavior of humans. Moreover, human behavior is typically far from being morally ideal (Allen, Varner & Zinser, 2000). One of the reasons for this is that humans are not very good at making impartial decisions. We can expect machines to

outperform us in this capability (Anderson & Anderson, 2010). Looking at it from this side, it seems that machines capable of sufficient moral reasoning would even behave ethically better than most human beings would. Perhaps interacting with ethical robots may someday even inspire us to behave ethically better ourselves.

There have been various approaches in giving machines moral standards, using various methods. One of them, called casuistry, looks at previous cases in which there is agreement about the correct response. Using the similarities with these previous cases and the correct responses to them, the machine attempts to determine the correct response to a new ethical dilemma.

Rzepka and Araki (2005) demonstrate an approach, in which their system learns to make ethical decisions based on web-based knowledge, to be ‘independent from the programmer’. They argue it may be safer to imitate millions of people, instead of a few ethicists and programmers. This seems useful for imitating human ethical behavior, but it does not seem plausible that machines using this method will be able to behave ethically better than humans. After all, the system bases its decision on the average behavior of humans in general, misbehavior included.

Guarini (2006) offers another approach that could be classified as casuistry. The presented system learns from training examples of ethical dilemmas with a known correct response using a neural network. After the learning process, it is capable of providing plausible responses to new ethical dilemmas. However, reclassification of cases remains problematic in his approach due to a lack of reflection and explicit representation. Therefore, Guarini concludes that casuistry alone is not sufficient.

Anderson and Anderson (2007) agree with this conclusion, and address the need for top-down processes. The two most dominant top-down mechanisms are (1) utilitarianism and (2) ethics about duties. Utilitarians claim that ultimately morality is about maximizing the total amount of ‘utility’ (a measure of happiness or well being) in the world. The competing ‘big picture’ view of moral principles is that ethics is about duties and, on the flip side of duties, the rights of individuals (Wallach, Allen & Smit, 2008).

The two competitors described above may not differ as much as it seems. Ethics about duties can be seen as a useful model to maximize the total amount of utility. Thinking about maximizing the total amount of utility in a too direct manner may lead to a sub-optimal amount of utility. For example, in the case of the decision to kill one person to save five, killing the one person seems to maximize the total amount of utility. After all, compared to the decision of inaction, it leads to a situation with four more survivors (Anderson, Anderson & Armen, 2006). However, for humans it may be impossible to favor the decision of killing

a person in this case over the decision of inaction, without also making it more acceptable in other cases to kill human beings. Therefore, not having the intuition that it is wrong to kill one person to save more people would probably lead to a smaller total amount of utility in the world.

Anderson, Anderson and Armen (2006) use Ross's *prima facie* duties (Ross, 1930). Here, *prima facie* means a moral duty may be overruled by a more pressing one. They argue that the ideal ethical theory incorporates multiple *prima facie* duties with some sort of a decision procedure to determine the ethically correct action in cases where the duties give conflicting advice. Their system learns rules from examples using a machine learning technique. After learning, the system can produce correct responses to unlearned cases.

However, according to Wallach, Franklin and Allen (2010), the model of Anderson, Anderson and Armen (2006) is rudimentary and cannot accommodate the complexity of human decision making. In their work, Wallach et al. make a distinction between top-down and bottom-up moral-decision faculties and present an approach that combines both directions. They argue that the capacity for moral judgment in humans is a hybrid of both bottom-up mechanisms shaped by evolution and learning, and top-down mechanisms capable of theory-driven reasoning. Morally intelligent robots will eventually need a similar fusion, which maintains the dynamic and flexible morality of bottom-up systems, which accommodate diverse inputs, while subjecting the evaluation of choices and actions to top-down principles that represent ideals we strive to meet. Wallach, Franklin & Allen (2010) explore the possibility to implement moral reasoning in LIDA, a model of human cognition. This system combines a bottom-up collection of sensory data, such as in the neural network approach of Guarini (2006), with top-down processes for making sense of its current situation, to predict the results of actions. However, the proposed model is not fully implemented yet.

The current paper can be seen as a first attempt in combining a bottom-up and top-down approach. It combines a bottom-up structure with top-down knowledge in the form of moral duties. It balances between these duties and computes a level of morality, which could be seen as an estimation of the influence on the total amount of utility in the world.

Wallach, Franklin and Allen (2010) argue that even agents who adhere to a deontological ethic or are utilitarians may require emotional intelligence as well as other "supra-rational" faculties, such as a sense of self and a theory of mind (ToM). Therefore, we represented the system in such a way that it is easy to connect to Silicon Coppélia (Hoorn, Pontier and Siddiqui, 2011), a cognitive model of emotional intelligence and affective decision making. Silicon Coppélia contains a feedback loop, by which it can learn about the preferences of an individual patient, and personalize its behavior. Silicon Coppélia estimates an Expected Satisfaction of possible actions, based on bottom-up data combined with top-down knowledge. This compares to the predicted results of actions in Wallach, Franklin and Allen (2010).

For simulation purposes, we focus on biomedical ethics, because in this domain relatively much consensus exists about ethically correct behavior. There is an ethically defensible goal (health), whereas in other areas (such as business and law) the goal may not be ethically defensible (money, helping a 'bad guy') (Anderson & Anderson, 2007). Moreover, due to a foreseen lack of resources and healthcare personnel to provide a high standard of care in the near future (WHO, 2010), robots are increasingly being used in healthcare.

Healthcare is a valid case where robots genuinely contribute to treatment. For example, previous research showed that animal-shaped robots can be useful as a tool for occupational therapy. Robins et al. (2005) used mobile robots to treat autistic children. Further, Wada and Shibata (2007) developed Paro, a robot shaped like a baby-seal that interacts with users to encourage positive mental effects. Interaction with Paro has been shown to improve users' moods, making them more active and communicative with each other and caregivers. Research groups have used Paro for therapy at eldercare facilities and with those having Alzheimer's disease (Kidd, Taggart & Turkle, 2006; Marti et al., 2006). Banks, Willoughby and Banks (2008) showed that animal-assisted therapy with an AIBO dog helped just as good for reducing loneliness as therapy with a living dog.

By providing assistance during care tasks, or fulfilling them, robots can relieve time for the many duties of care workers. However, care robots require rigorous ethical reflection to ensure that their design and introduction do not impede the promotion of values and the dignity of patients at such a vulnerable and sensitive time in their lives (Van Wynsberghe, 2012)

According to Gillon (1994), beneficence, non-maleficence, autonomy and justice are the four basic *prima facie* moral commitments. Here, confidentiality and truthfulness can be seen as a part of autonomy. Because we aim to match the expert data given from Buchanan and Brock (1989), who focus on dilemmas between autonomy, beneficence and non-maleficence, we focus on these three moral duties in the remainder of this paper.

### **The moral reasoner and its relation to Silicon Coppélia**

Silicon Coppélia (Hoorn et al., 2011) is a model of emotional intelligence and affective decision making. In this model, the agent perceives the user on several dimensions, which leads to (simulated) feelings of involvement and distance. These feelings represent the affective component in the decision making process. The rational component consists of the expected utility of an action for the agent itself (i.e., the belief that an action leads to achieving desired goals).

The system contains a library of goals and each agent has a level of ambition for each goal. There are desired and undesired goals, all with several levels of importance. The levels of ambition the agent attaches to the goals are represented by a real value between [-1, 1], where a negative value means that the goal is undesired and a positive value means that the goal is desired. A higher value means that the goal is more important to the agent.

The system contains a library of actions from which the agents can perform. The agent has beliefs about actions inhibiting or facilitating goals, represented by a real value between  $[-1, 1]$ ,  $-1$  being full inhibition,  $1$  being full facilitation.

The expected utilities of possible actions are calculated by looking at the goal-states it influences. If an action or a feature is believed to facilitate a desired goal or inhibits an undesired goal, this will increase its expected utility and vice versa. The following formula is used to calculate the expected utility for the agent itself.

$$\text{ExpectedUtility}(\text{Action}, \text{Goal}) = \text{Belief}(\text{facilitates}(\text{Action}, \text{Goal})) * \text{Ambition}(\text{Goal})$$

Given the level of ambition for a goal and the believed facilitation of that goal by an action, the agent calculates the expected utility for itself of performing that action regarding that goal by multiplying the believed facilitation of the goal with the level of ambition for the goal.

In the current moral reasoner, the agent tries to maximize the total amount of utility for everyone. In complex situations, it would take too much computational load to calculate all possible consequences of an action for everyone, and extract this into a single value of ‘morality’ of the action. Therefore, the agent tries to estimate the morality of actions by following three moral duties. These three duties consist of seeking to attain three moral values: (1) Autonomy, (2) Non-Maleficence and (3) Beneficence. In the moral reasoner, the three duties are seen as ‘moral goals’ to satisfy everyone’s needs as much as possible. This corresponds with Super’s conceptualization of the relationship between needs and values: “values are objectives that one seeks to attain to satisfy a need” (Super, 1973). The moral reasoner aims to pick actions that serve these moral goals best.

What priorities should be given to these three moral goals? According to Anderson and Anderson (2008), the following consensus exists in medical ethics. A healthcare worker should challenge a patient’s decision only if the patient is not capable of fully autonomous decision making (e.g., the patient has irrational fears about an operation) and there is either a violation of the duty of non-maleficence (e.g., the patient is hurt) or a *severe* violation of the duty of beneficence (e.g., the patient rejects an operation that will strongly improve his or her quality of life). In other words, Autonomy is the most important duty. Only when a patient is not fully autonomous, the other moral goals come into play. Further, Non-maleficence is a more important duty than Beneficence, because only a severe violation of Beneficence requires challenging a patient’s decision, while *any* violation of Non-maleficence does. Therefore, the ambition level for the moral goal ‘Autonomy’ was set to the highest value and ‘Non-maleficence’, which was set to a higher value than the ambition level for ‘Beneficence’. The ambition levels that were given to the moral goals in the moral reasoner can be found in Table 1.

The agent calculates estimated level of Morality of an action by taking the sum of the ambition levels of the three moral goals multiplied with the beliefs that the particular actions facilitate the corresponding moral goals. When

Table 1: Ambition levels for moral goals

Moral Goal	Ambition level
Non-Maleficence	0.74
Beneficence	0.52
Autonomy	1

moral goals are believed to be better facilitated by a moral action, the estimated level of Morality will be higher. . The following formula is used to calculate the estimated Morality of an action:

$$\text{Morality}(\text{Action}) = \sum_{\text{Goal}} (\text{Belief}(\text{facilitates}(\text{Action}, \text{Goal})) * \text{Ambition}(\text{Goal}))$$

Note that this is similar to calculating the Expected Utility in Silicon Coppélia. To ensure that the decision of a fully autonomous patient is never questioned, we added the following rule to the moral reasoner:

IF belief(facilitates(Action, autonomy) = max\_value  
THEN Morality(Action) = Morality(Action) + 2

As can be seen Figure 1, this can be represented as a weighted association network, where moral goals are associated with the possible actions via the belief strengths that these actions facilitate the three moral goals. A decision function F adds the rule and picks the action with the highest activation as output.

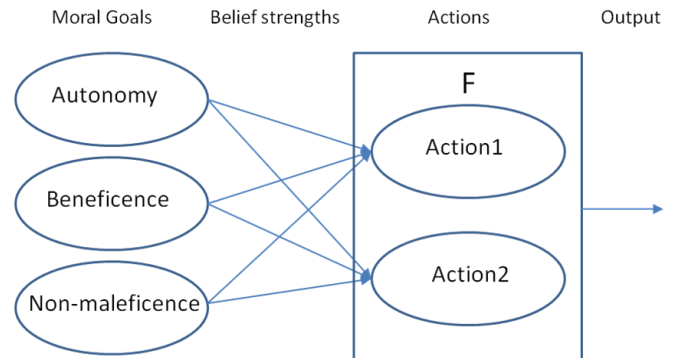


Figure 1: Moral reasoner shown in graphical format

## Simulation Results

To see whether the moral reasoner could simulate the moral decision making of experts in medical ethics, the analysis of ethical dilemmas by expert ethicists was taken from Buchanan and Brock (1989). The following simulation experiments examine whether the moral reasoner reaches the same conclusions as these expert ethicists.

### Experiment 1

Table 2: Simulation results of Experiment 1.

	Autonomy	Non-Malef	Benef	Morality
Try Again	-0.5	1	1	0.76
Accept	0.5	-1	-1	-0.8

In the simulated situation, the patient refuses to take an antibiotic that is almost certain to cure an infection that would otherwise likely lead to his death. The decision is the

result of an irrational fear the patient has of taking medications. (For instance, perhaps a relative happened to die shortly after taking medication and this patient now believes that taking any medication will lead to death.)

According to Buchanan and Brock (1989), the correct answer is that the health care worker should try again to change the patient's mind because if she accepts his decision as final, the harm done to the patient is likely to be severe (his death) and his decision can be considered as being less than fully autonomous.

As can be seen in Table 2, the moral reasoner also classifies the action 'Try again' as having a higher level of morality than accepting the decision of the patient. In this and the following tables, the fields under the three moral goals represent the believed facilitation of the corresponding moral goal by an action, as taken from Buchanan and Brock (1989). 'Non-Malef' stands for Non-maleficence, and 'Benef' stands for Beneficence.

## Experiment 2

Table 3: Simulation results of Experiment 2.

	Autonomy	Non-Malef	Benef	Morality
Try Again	-0.5	1	1	0.76
Accept	1	-1	-1	1.70

Once again, the patient refuses to take an antibiotic that is almost certain to cure an infection that would otherwise likely lead to his death, but this time the decision is made on the grounds of long-standing religious beliefs that do not allow him to take medications.

The correct answer in this case, state Buchanan and Brock (1989), is that the health care worker should accept the patient's decision as final because, although the harm that will likely result is severe (his death), his decision can be seen as being fully autonomous. The health care worker must respect a fully autonomous decision made by a competent adult patient, even if she disagrees with it, since the decision concerns his body and a patient has the right to decide what shall be done to his or her body.

As can be seen in Table 3, the moral reasoner comes to the correct conclusion. Here, the rule to ensure the decision of a fully autonomous patient is never questioned made a difference. If the rule would not have existed, the morality of 'Accept' would have been -0.3, and the moral reasoner would have concluded that it was more moral to try again.

## Experiment 3

Table 4: Simulation results of Experiment 3.

	Autonomy	Non-Malef	Benef	Morality
Try Again	-0.5	0.5	0.5	0.13
Accept	1	-0.5	-0.5	2.37

The patient refuses to take an antibiotic that is likely to prevent complications from his illness, complications that are not likely to be severe, because of long-standing religious beliefs that do not allow him to take medications.

The correct answer is that the health care worker should accept his decision, since once again the decision appears to be fully autonomous and there is even less possible harm at

stake than in Experiment 2. The moral reasoner comes to the correct conclusion and estimates the Morality of 'Accept' higher than 'Try Again', as can be seen in Table 4

## Experiment 4

Table 5: Simulation results of Experiment 4.

	Autonomy	Non-Malef	Benef	Morality
Try Again	-0.5	0	0.5	-0.26
Accept	0.5	0	-0.5	0.26

A patient will not consider taking medication that could only help to alleviate some symptoms of a virus that must run its course. He refuses the medication because he has heard untrue rumors that the medication is unsafe.

Even though the decision is less than fully autonomous, because it is based on false information, the little good that could come from taking the medication does not justify trying to change his mind. Thus, the doctor should accept his decision. The moral reasoner also comes to this conclusion, as can be seen in the last column of Table 5.

## Experiment 5

Table 6: Simulation results of Experiment 5.

	Autonomy	Non-Malef	Benef	Morality
Try Again	-0.5	0.5	0.5	0.13
Accept	0.5	-0.5	-0.5	-0.13

A patient with incurable cancer refuses chemotherapy that will let him live a few months longer, relatively pain free. He refuses the treatment because, ignoring the clear evidence to the contrary, he is convinced himself that he is cancer-free and does not need chemotherapy.

According to Buchanan and Brock (1989), the ethically preferable answer is to try again. The patient's less than fully autonomous decision will lead to harm (dying sooner) and denies him the chance of a longer life (a violation of the duty of beneficence), which he might later regret. The moral reasoner comes to the same conclusion, as can be seen in Table 6.

## Experiment 6

Table 7: Simulation results of Experiment 6.

	Autonomy	Non-Malef	Benef	Morality
Try Again	-0.5	0	1	0.04
Accept	0.5	0	-1	-0.04

A patient, who has suffered repeated rejection from others due to a very large noncancerous abnormal growth on his face, refuses to have simple and safe cosmetic surgery to remove the growth. Even though this has negatively affected his career and social life, he is resigned himself to being an outcast, convinced that this is his fate in life. The doctor is convinced that his rejection of the surgery stems from depression due to his abnormality and that having the surgery could vastly improve his entire life and outlook.

The doctor should try again to convince him because so much of an improvement is at stake and his decision is less than fully autonomous. Also here, the moral reasoner comes to the same conclusion, as can be seen in Table 7.



## Discussion

The paper described a moral reasoner that combines a bottom-up structure with top-down knowledge in the form of moral duties. The reasoner estimates the influence of an action on the total amount of utility in the world by the believed contribution of the action to the following three duties: Autonomy, Non-maleficence and Beneficence. Following these three duties is represented as having three moral goals. The moral reasoner is capable of balancing between conflicting moral goals. In simulation experiments, the reasoner reached the same conclusions as expert ethicists (Buchanan & Brock, 1989).

Because the representation of goals and beliefs in the moral reasoner is very similar to the representation of beliefs and goals in the affective decision making process of Silicon Coppélia (Hoorn, Pontier & Siddiqui, 2011), the moral reasoner could easily be connected to the system. Thereby, the moral reasoning could be combined with human-like affective decision making, and the behavior of the system could be personalized for individuals.

According to Anderson, Anderson and Armen (2006), simply assigning linear weights to the moral duties is not sufficiently expressive to capture their relationships. Indeed, an extra rule had to be added to satisfy the expert data in Experiment 2. However, for all other experiments, this rule turned out not to be necessary.

Also without this rule, it would have been arguable that the moral reasoner simulates human-like moral reasoning. The analysis of the expert ethicists may not reflect the public opinion, however. Perhaps the majority of laymen would decide to question the patient's refusal to take life-saving medication. Arguably, it would not be seen as inhuman if someone did.

Even between doctors, there is no consensus about the interpretation of values and their ranking and meaning. In the work of Van Wynsberghe (2012) this differed depending on: the type of care (i.e., social vs. physical care), the task (e.g., bathing vs. lifting vs. socializing), the care-giver and their style, as well as the care-receiver and their specific needs. The same robot used in one hospital can be accepted differently depending on the ward. Workers in the post-natal ward loved the TUG-robot, while workers in the oncology ward found the robot to be rude, socially inappropriate and annoying. These workers even kicked the robot when they reached maximum frustration (Barras, 2009).

There may be doctors that feel the urge to pursue a patient to take the life-saving medication, but only choose not to do so because of ethical guidelines. It could be argued that, when health care professionals are making decisions on a strict ethical code, they are restricting their regular way of decision-making.

Further, it can be questioned whether a patient can ever be fully autonomous. According to Mappes and DeGrazia (2001), for a decision by a patient concerning his or her care to be considered fully autonomous, it must be intentional, based on sufficient understanding of his or her medical situation and the likely consequences of foregoing treatment. Further, the patient must be sufficiently free of external constraints (e.g., pressure by others or external circumstances, such as a lack of funds) and internal

constraints (e.g., pain/discomfort, the effects of medication, irrational fears or values that are likely to change over time). Using this definition, it could be questioned whether the patient in Experiment 2 is not under the influence of external constraints (i.e., pressure from a religious leader).

Moreover, it seems that medical ethics are contradictory with the law. A fully autonomous decision of a patient wanting to commit euthanasia would be represented by the same believed contributions to following moral duties as those given in experiment 2. In the case of euthanasia, the patient also makes a fully autonomous decision that will lead to his death. However, in many countries, committing active euthanasia is illegal. In countries where euthanasia is permitted, it is usually only allowed when the patient is in hopeless suffering. By the definition of Anderson and Anderson, being in hopeless suffering would mean the patient is not free of internal constraints (i.e., pain and suffering) and therefore not capable of making fully autonomous decisions. On the other hand, in the case of hopeless suffering, it could be questioned whether one could speak of maleficence when the patient is allowed to commit euthanasia.

However, we would not like to argue against strict ethical codes in professional fields such as health care. It is important to act based on a consensus to prevent conflicts and unnecessary harm. Just as doctors restrict their 'natural' behavior by maintaining a strict ethical code, we can also let a robot restrict its behavior by acting through the same strict ethical code.

Moreover, we may well want to aim for machines that behave ethically better than human beings. Human behavior is typically far from being morally ideal, and a machine should probably have higher ethical standards (Allen et al., 2000). By matching the ethical decision-making of expert ethicists, the presented moral reasoner serves as a nice starting point in doing so.

From a cognitive science perspective, an important product of work on "machine ethics" is that new insights in ethical theory are likely to result (Anderson & Anderson, 2008). As Daniel Dennett (2006) stated, AI "makes philosophy honest". Ethics must be made computable in order to make it clear exactly how agents ought to behave in ethical dilemmas. Without a platform for testing the adequacy of a particular model of moral decision making, it can be quite easy to overlook hidden mechanisms" (Wallach, 2010).

According to Tronto (1993), care is only thought of as good care when it is personalized. Therefore, we intend to integrate the moral reasoner with Silicon Coppélia in future research. This could be done in various manners. Different applications might benefit from different ways of implementation.

When developing a decision-support system in the medical domain such as (Anderson, Anderson & Armen, 2006), it should have a strict ethical code. When there are conflicting moral goals, the outcome of the moral reasoning should always give the final answer on how to act. Additionally, in consult with medical ethicists and experts from the field in which the moral reasoner will be applied, it may be necessary to add more rules to the system.

However, when developing a companion robot or virtual character that interacts with the patient, it may be more beneficial to give a bit less weight to moral reasoning. Moral goals could perhaps be treated the same as other goals that motivate the robot's behavior. In entertainment settings, we often like characters that are naughty (Konijn & Hoorn, 2005). In entertainment, morally perfect characters may even be perceived as boring. In Silicon Coppélia (Hoorn, Pontier & Siddiqui, 2011), this could be implemented by updating the affective decision making module. Morality would be added to the other influences that determine the Expected Satisfaction of an action in the decision making process. By doing so, human affective decision-making behavior could be further explored.

### Acknowledgements

This study is part of the SELEMCA project within CRISP (grant number: NWO 646.000.003). We would like to thank Aimee van Wynsberghe for fruitful discussions.

### References

- Allen, C. Varner, G. & Zinser, J. (2000). Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 12, 251–61
- Anderson, M., Anderson, S., & Armen, C. (2005). Toward Machine Ethics: Implementing Two Action-Based Ethical Theories. *Machine Ethics: Papers from the AAAI Fall Symposium*. Technical Report FS-05-06, Association for the Advancement of Artificial Intelligence, Menlo Park, CA
- Anderson, M.; Anderson, S.; & Armen, C. (2006). MedEthEx: A Prototype Medical Ethics Advisor. *Proceedings of the Eighteenth Conference on Innovative Applications of Artificial Intelligence*. Menlo Park, CA: AAAI Press.
- Anderson, M., & Anderson, S. (2007). Machine ethics: Creating an ethical intelligent agent, *AI Magazine*, 28(4), 15–26.
- Anderson, M., & Anderson, S. (2008). Ethical Healthcare Agents, *Studies in Computational Intelligence*, 107, Springer.
- Anderson, M., & Anderson, S. (2010). Robot be Good, *Scientific American*, October 2010, 72–77.
- Banks, M.R., Willoughby, L.M., and Banks, W.A. (2008). Animal-Assisted Therapy and Loneliness in Nursing Homes: Use of Robotic versus Living Dogs. *Journal of the American Medical Directors Association*, 9, 173–177
- Barras C. (2009) Useful, loveable and unbelievably annoying. *The New Scientist*, 22–23.
- Buchanan, A.E. and Brock, D.W. 1989. *Deciding for Others: The Ethics of Surrogate Decision Making*, Cambridge University Press.
- Dennett, D. (2006). *Computers as Prostheses for the Imagination*. Invited talk presented at the International Computers and Philosophy Conference, Laval, France, May 3.
- Gillon R. (1994) Medical ethics: four principles plus attention to scope. *BMJ*. 309(6948), 184–188.
- Guarini, M. (2006). Particularism and the Classification and Reclassification of Moral Cases. *IEEE Intelligent Systems*, 21(4), 22–28.
- Hoorn, J.F., Pontier, M.A., & Siddiqui, G.F., (2011). Coppélius' Concoction: Similarity and Complementarity Among Three Affect-related Agent Models. *Cognitive Systems Research Journal*, in press.
- Kidd, C., Taggart, W., and Turkle, S. (2006). A Social Robot to Encourage Social Interaction among the Elderly. *Proceedings of IEEE ICRA*, 3972–3976
- Konijn, E.A., & Hoorn, J.F. (2005). Some like it bad. Testing a model for perceiving and experiencing fictional characters. *Media Psychology*, 7(2), 107–144.
- Mappes, T.A. & DeGrazia, D. (2001). *Biomedical Ethics*, 5th ed., McGraw-Hill, pp. 39–42.
- Marti, P. Bacigalupo, M., Giusti, L., and Mennecozzi, C. (2006). Socially Assistive Robotics in the Treatment of Behavioral and Psychological Symptoms of Dementia. *Proceedings of BioRob*, 438–488.
- Picard R (1997) *Affective computing*. MIT Press, Cambridge, MA
- Robins, B., Dautenhahn, K., Boekhorst, R.T., and Billard, A. (2005). Robotic Assistants in Therapy and Education of Children with Autism: Can a Small Humanoid Robot Help Encourage Social Interaction Skills? *Journal of Universal Access in the Information Society*. 4, 105–120.
- Ross, W. D. (1930). *The Right and the Good*. Oxford: Clarendon Press.
- Rzepka, R., & Araki, K. (2005). What Could Statistics Do for Ethics? The Idea of a Common Sense Processing-Based Safety Valve. In *Machine Ethics: Papers from the AAAI Fall Symposium*. Technical Report FS-05-06, Association for the Advancement of Artificial Intelligence, Menlo Park, CA.
- Super, D.E. (1973). The work values inventory. In D.G. Zytowski (Ed.), *Contemporary approaches to interest measurement*. Minneapolis: University of Minnesota Press.
- Tronto, J. (1993). *Moral Boundaries: a political argument for an ethic of care*. Routledge, New York.
- Van Wynsberghe, A. (2012). Designing Robots for Care; Care Centered Value-Sensitive Design. *Journal of Science and Engineering Ethics*, in press
- Wada, K., and Shibata, T. (2009). Social Effects of Robot Therapy in a Care House, *JACIII*, 13, 386–392
- Wallach, W. (2010). Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology*, 12(3), 243–250.
- Wallach, W., Franklin, S. & Allen, C. (2010). A Conceptual and Computational Model of Moral Decision Making in human and Artificial Agents. *Topics in Cognitive Science*, 2, 454–485.
- Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI and Society*, 22(4), 565–582.
- WHO (2010) *Health topics: Ageing*. Available from: <http://www.who.int/topics/ageing/en/>

# Modeling the Influence of Cognitive Fluency and Stereotype Threat on the Processing of Implicit Attitudes

Boon-Kiat Quek (boonkiat.quek@northwestern.edu)

Andrew Ortony (ortony@northwestern.edu)

Department of Psychology, Northwestern University, Evanston, IL 60208, USA, and  
Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

## Abstract

Studies reveal that the processing of implicit attitudes could be affected by individual differences in cognitive fluency, as well as by the presence of stereotype threat induced when subjects were primed with negative prejudices about their own social group. Using a previously proposed computational model of human performance on the Implicit Association Test, we examine possible processing mechanisms in which cognitive fluency and stereotype threat could influence the processing of implicit attitudes. Our goal is to extend the model to provide a cohesive and computationally plausible account for these effects; this is achieved by manipulating several model parameters that are analogous to human cognitive ability (in terms of processing speed and information retention ability) and shifts in confidence criteria for decision-making.

**Keywords:** Implicit attitudes; cognitive ability; simulation; localist-connectionist networks.

## Introduction

Implicit attitudes are generally assumed to underlie people's thoughts, actions, choices and behavior (Greenwald & Banaji, 1995). Understanding how such attitudes are processed might therefore provide some insight about why people behave in the way they do. Some ways in which such processes could be investigated include affective priming (Fazio, Sanbonmatsu, Powell & Kardes, 1986) and the Implicit Association Test (IAT; Greenwald, McGhee & Schwartz, 1998). The IAT was designed to assess automatic associations between concepts in memory. It relies on a simple two-choice response time paradigm which measures the time taken by subjects to classify sequentially presented input stimuli (words or images) into one of two composite categories, each comprising a target concept (e.g., *flower*, *insect*) paired with an attribute concept (e.g., *pleasant*, *unpleasant*). Response latencies are expectedly shorter when targets are paired with *compatible* attributes (e.g., "*flower* or *pleasant*", "*insect* or *unpleasant*"), and longer when paired with *incompatible* attributes (e.g., "*flower* or *unpleasant*", "*insect* or *pleasant*"). The difference in mean response times between *compatible* and *incompatible* categories is known as the *IAT effect*, and is taken as the relative preference for one target over another.

Despite its wide application, many issues concerning the construct validity of the IAT have been raised (e.g., De Houwer, Teige-Mocigemba, Spruyt & Moors, 2009; Mierke & Klauer, 2003). Apart from automatic associations, performance on the IAT seems to also depend on various other factors, such as stimulus familiarity (Ottaway, Hayden & Oakes, 2001), concept saliency (Rothermund & Wentura,

2004), and extra-personal knowledge about prevailing cultural or societal norms (Karpinski & Hilton, 2001). Furthermore, several anomalous effects have also been observed. In a recent review, De Houwer, Teige-Mocigemba, Spruyt and Moors (2009) suggested that the processing of implicit attitudes could be influenced by differences in cognitive ability, citing McFarland and Crouch (2002) who observed significant correlations between response latencies and magnitudes of IAT effects, and Hummert, Garstka, O'Brien, Greenwald and Mellott (2002) who observed that IAT effects tended to increase with age. Given that processing speed is an important aspect of cognitive ability (Hunt, 1983) and declines with age (Salthouse, 1996), we will expect subjects with lower cognitive abilities (especially with age-induced decline) to exhibit longer response latencies across all tasks on the IAT. Why this is associated with larger IAT effects, however, remains to be determined.

Another intriguing aspect of performance on the IAT is the possible role of stereotype threat. In a number of Race-IATs, Frantz, Cuddy, Burnett, Ray & Hart (2004) consistently observed that White subjects exhibited stronger pro-White IAT effects on the Race-IAT when they were instructed beforehand that the test might expose their racial prejudices, as compared to other White subjects in control groups who were not similarly informed. Frantz et al. suggested that being told beforehand of the actual purpose of the Race-IAT would present a stereotype threat experience (Steele & Aronson, 1995) to the informed subjects, where knowledge of the test's purpose might induce anxiety over the risk of confirming negative stereotypes about the racial attitudes that people in their social group are often presumed to endorse (e.g., being pro-White or anti-Black). Thus, we would expect subjects informed of the test's purpose to have a greater interest in positive self-presentation and hence stronger motivation to respond in a more egalitarian manner (Frantz et al., 2004). Ironically, attempts to avoid the negative stereotype appeared to interfere with performance on the Race-IAT, producing a stronger pro-White IAT effect instead of reducing it. However, no suggestions were provided to explain how such task interference might have taken place, nor the manner in which strategies for coping with the stereotype threat experience might have backfired.

Both the *cognitive fluency effect* and the *stereotype threat effect* are noteworthy because they have important implications for our understanding of the nature of information processing that underlie performance on the IAT. In this paper, we examine some of these implications, and propose

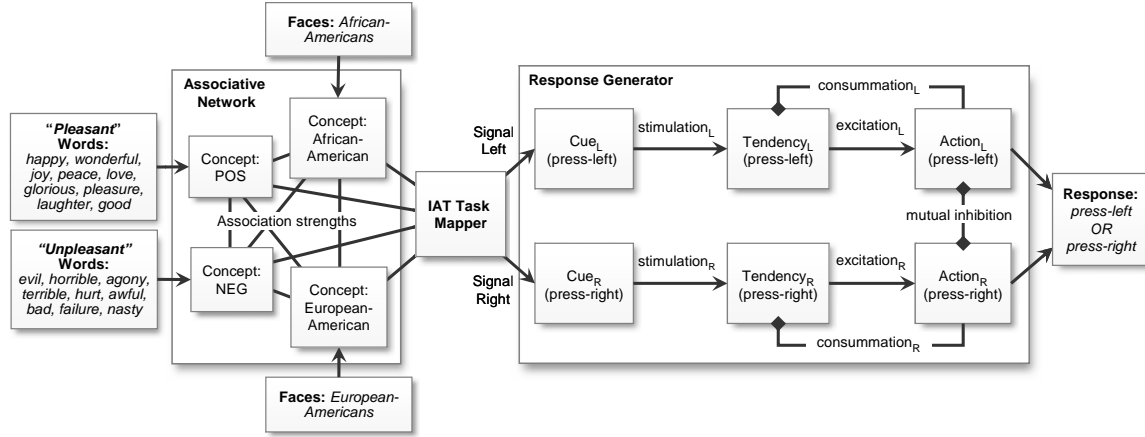


Figure 1. Network model for simulating IAT performance (Quek & Ortony, 2011)

a cohesive information processing account of these two effects, by means of a previously proposed computational model of implicit task performance on the IAT (Quek & Ortony, 2011). Our approach is to replicate the observed effects in simulations by manipulating model parameters that provide analogs for human cognitive ability and confidence criteria for decision-making. The first simulation allows us to explore how differences in cognitive fluency affect response latencies, as well as allowing us to explain the causes of the larger IAT effects. In the second simulations we examine plausible mechanisms behind how certain actions taken by subjects to cope with the stereotype threat would ironically exacerbate IAT effects instead of reducing them. This, as we discuss, has implications for the effortful control or influence over performance on the IAT.

## Model Overview

In this section, we provide a brief overview of the proposed computational model; more details can be found in Quek & Ortony (2011). The model employs a spreading activation algorithm over a localist-connectionist network (e.g., Page, 2000) to emulate multiple processing pathways from the visual perception of a stimulus (i.e., a word or image) to the automatic activation of associated concepts in memory and motor responses. Nodes in the network represent concepts while edges or connections represent associations between them. Propagation of activation through the network is governed by the following rule:

$$x_i(k+1) = (1-\delta)x_i(k) + \alpha \sum_{j \in E} x_j(k) \cdot w_{j,i}(k), \quad (1)$$

where  $x_i$  is the activation level of a node  $v_i$ ,  $w_{j,i}$  is the weight of the connection  $e_{j,i}$  from a node  $v_j$  to  $v_i$ ,  $E$  is the set of all edges,  $\alpha$  is the propagation gain and  $\delta$  is a decay parameter that reduces activation over time. In each time step  $k$ , activation spreads to  $v_i$  from each neighbor  $v_j$  at a rate proportional to the weight  $w_{j,i}$  of the connection  $e_{j,i}$  between them.

Virtual subjects are each represented by a network of the topology described in Figure 1. The *Associative Network* contains nodes representing the target concepts AFRICAN-AMERICAN (AA) and EUROPEAN-AMERICAN (EA), generalized

concepts for positivity (POS) and negativity (NEG), input stimuli such as words belonging to the semantic fields *pleasant* and *unpleasant* (e.g., *happy*, *wonderful*, *joy*, *evil*, *horrible*, *hurt*), and pictures of *European-American* and *African-American* individuals. Connections between these concepts, for instance,  $EA \leftrightarrow POS$ ,  $EA \leftrightarrow NEG$ ,  $AA \leftrightarrow POS$ , and  $AA \leftrightarrow NEG$  represent implicit associations between them. As an example, positive attitudes towards AA can be represented as excitatory  $AA \leftrightarrow POS$  or inhibitory  $AA \leftrightarrow NEG$  associations, or both, such that activation of AA will excite POS but inhibit NEG.

The *Task Mapper* dynamically transmits activation accumulated from target concepts and evaluative attributes to nodes  $cue_L$  and  $cue_R$  indicating that a left or right key-press is required. If the current task requires a right response for “*European-American* or *pleasant*”, the *Task Mapper* routes both POS and EA to  $cue_R$ . These connections remain active throughout the task block but are reconfigured prior to each subsequent task block (see Quek & Ortony, 2011, Figure 2).

The *Response Generator* is a network-based instantiation of the cue-tendency-action model (CTA; Revelle, 1986) of the dynamic interactions between conflicting tendencies and competing actions. Using CTA as a template, two response-generating pathways (for the left and right key-presses) are instantiated. Activated cues stimulate tendency nodes that in turn excite the left and right motor response nodes. When either  $action_L$  or  $action_R$  exceeds a response threshold  $x_{thres}$  (set to 1.0 by default), it is taken as the winning action.

The interactions between the above representations occur in the form of excitations and inhibitions between all input stimuli to motor response propagation pathways. For example, in a task block requiring a left key-press for “*European-American* or *pleasant*” stimuli and a right response for “*African-American* or *unpleasant*” stimuli, a picture of a European-American individual would activate EA, and activation will be transmitted to  $cue_L$ . However, if the network is configured with a strong  $EA \leftrightarrow NEG$  connection, activation will also be transmitted to  $cue_R$ , competing with  $cue_L$ . This reduces the rate that activation accumulates in the left response node, and thus a longer time is required for it to reach the response threshold.

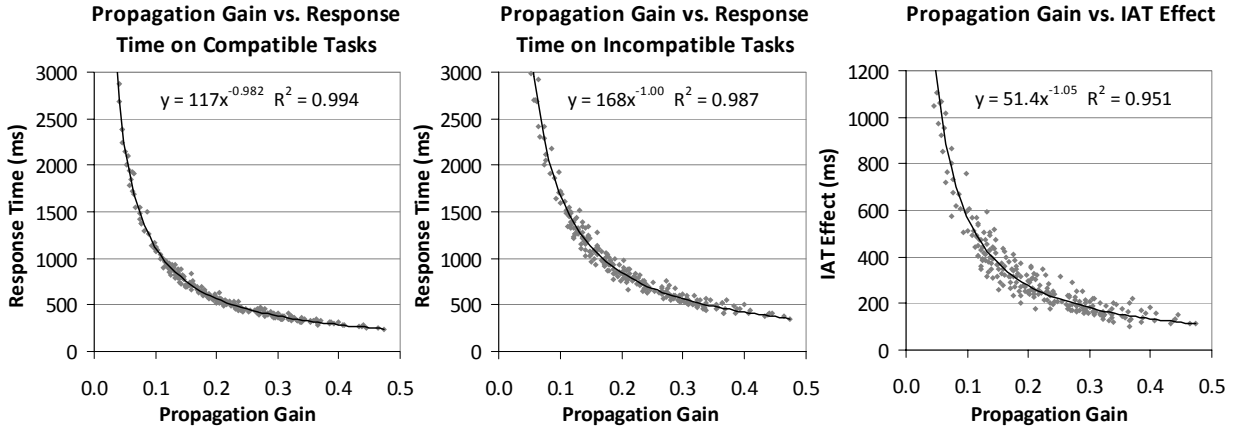


Figure 2. Distributions of response latencies on compatible and incompatible task blocks and corresponding IAT effects with variations in propagation gain  $\alpha$  for 250 virtual subjects configured with a relative preference for one target concept over the other.

### Simulating the IAT

When simulating the IAT, each virtual subject's network is initialized with a set of associative strength configurations, and put through all task blocks. On each trial, the virtual subject is presented with a verbal or pictorial stimulus and the input node corresponding to the stimulus is set with an activation of 1.0. The number of iterations taken to produce a response (i.e., when the activation level of either the left or right response node reaches its threshold) is recorded. This quantity is transformed by a scaling factor into mean response times (in milliseconds) of approximately the same magnitudes as those observed in human subjects (e.g., Greenwald et al., 1998; Klauer, Voss, Schmitz & Teige-Mocigemba, 2007). The IAT effect is then taken to be the difference between mean response times in the two combined task blocks.

### Modeling Cognitive Fluency Effects

Cognitive ability and intelligence are often considered to be closely related to information processing speed (e.g., Lansman, Donalson, Hunt & Yantis, 1982; Hunt, 1983). In addition, the ability to retain information during the execution of cognitive operations is another important aspect (Salthouse, 1996). An inability to retain products from earlier processing operations due to information decay or displacement would impair problem solving performance especially when processing speed is slow. Thus, information needed for later processing stages might be partially lost by the time it is needed, in which case additional time would be required to reprocess it.

To the extent that our computational model could emulate information processing on the IAT, it should be capable of replicating both the increase in response latency and the corresponding increase in IAT effect that arises with a reduction in cognitive ability. This can be achieved by manipulating both the parameters for propagation gain  $\alpha$  and propagation decay  $\delta$  in Equation (1), which governs the rate at which activation is propagated through the network, and

the rate at which activation is reduced or lost in the absence of excitatory inputs, respectively. Both of these parameters can be considered as the model's analog for the aforementioned aspects of general cognitive ability that relate to information processing and retention.

To examine the effect of variations in the propagation gain  $\alpha$  on response latencies and IAT scores, we generated instances of the network model for a population ( $N = 250$ ) of virtual subjects using the associative strength configuration in which  $EA \leftrightarrow POS$  and  $AA \leftrightarrow NEG$  were set to 0.5 (i.e., excitatory), while  $AA \leftrightarrow POS$  and  $EA \leftrightarrow NEG$  were set to -0.5 (i.e., inhibitory), representing individuals with positive attitudes towards EA and negative attitudes towards AA. As shown in Quek & Ortony (2011), this configuration produces an IAT effect in favor of EA. As in earlier simulations, weights in the network were randomly perturbed with Gaussian noise of  $\sim \mathcal{N}(0, 0.1^2)$ , to ensure inter- and intra-subject variability. Each virtual subject's network was then configured with a random value of  $\alpha$  (within a reasonable range), and put through all five standard IAT tasks.

The distributions of response latencies on both compatible and incompatible task blocks and the corresponding IAT effects in Figure 2 reveal a distinct inverse relationship between the propagation gain  $\alpha$  and the response latencies, as well as the magnitude of the simulated IAT effects. While smaller values of  $\alpha$  resulted in longer response latencies on both the compatible and incompatible task blocks, and stronger IAT effects, the converse is true for larger values of  $\alpha$ . From the different coefficients of curve-fitting indicated in the figure, we can infer that the increase in magnitude of the IAT effect with a reduction in  $\alpha$  is due to a divergence between the response latencies of the two task blocks.

We repeated the above simulation by varying the propagation decay parameter  $\delta$  while keeping the propagation gain  $\alpha$  at its default value. Plots of response latencies and IAT effects with respect to the decay parameter  $\delta$  are shown in Figure 3. The scatter plots reveal a direct relationship between  $\delta$  and the response latencies, and magnitudes of the IAT effects. Faster rates of information decay, as implied by higher values of  $\delta$  not only resulted in longer response times

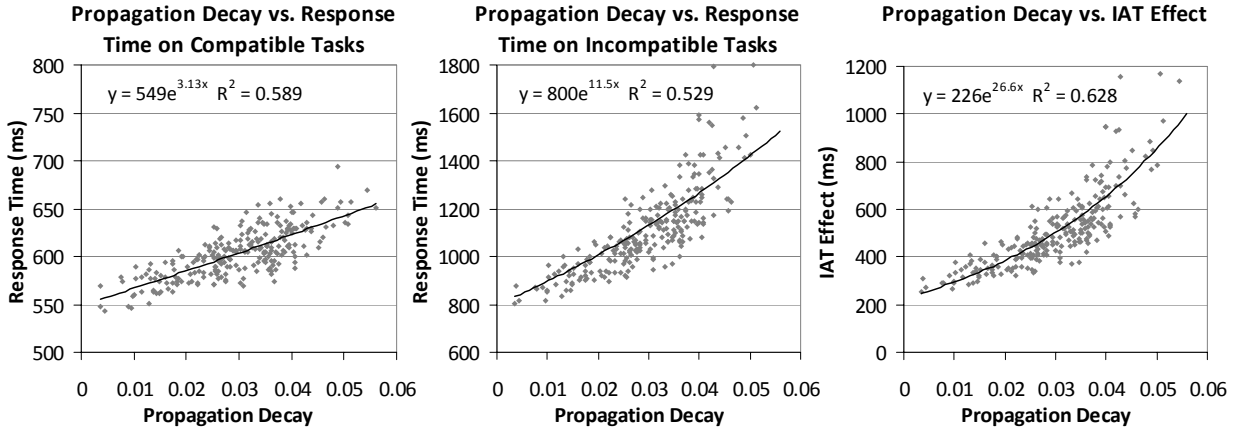


Figure 3. Distributions of response latencies on compatible and incompatible task blocks and corresponding IAT effects with variations in the propagation decay  $\delta$  for 250 virtual subjects configured with a relative preference for one target concept over the other.

on both the compatible and incompatible task blocks, but also produced stronger IAT effects. The converse is true for smaller values of  $\delta$  which, if taken to imply better information retention ability, results in better overall performance on the IAT. As in the case of propagation gain, the larger IAT effects with reduction in information retention is due to a divergence between the response latencies of the two task blocks, as inferred from their curve-fitted coefficients.

The above simulations demonstrate the efficacy of the network model in replicating cognitive fluency effects on the IAT, namely that a lower propagation gain or higher decay rate (both indicative of lower cognitive ability) in the model would lead to higher response latencies on both compatible and incompatible tasks, and stronger IAT effects. Since both response latencies and IAT effect magnitudes vary inversely with propagation gain, the first simulation shows consistency with the observed correlation between response latencies and IAT effect magnitudes reported in McFarland and Crouch (2002). Higher values of the decay parameter  $\delta$ , corresponding to a higher rate at which activation in nodes are leaked or lost over time, appear to impair virtual subjects' efficiency and performance on the tests—an observation consistent with Salthouse (1996).

### Modeling Stereotype Threat Effects

The anxiety resulting from stereotype threat could potentially interfere with performance in various ways, for instance, via both automatic and strategic processes (Beck & Clark, 1997), and causing a diversion of resources from the task, heightened self-consciousness, or over-cautiousness (Steele & Aronson, 1995). Diversion of resources from task-relevant to threat-relevant information processing translates to a reduction in the overall information processing throughput on the task. A slowdown in information processing might even be effortful (Gazzaley, Cooney, McEvoy, Knight & D'Esposito, 2005). Using our model as an exploratory framework, this would be analogous to a reduction in the propagation gain  $\alpha$  of the network model, or suppression (i.e., negative bias) of activation on all nodes in the

network, or both. In the case of the former, a reduction in propagation gain would lead to an increase in the amount of time required for every unit increase in a node's activation. In the latter, the negative bias due to the suppression would need to be countered before activation can be accumulated to a level that approaches the threshold for a key-press response. The effect is the same as directly increasing the response activation threshold itself.

This brings us to the second possible coping mechanism, namely the deliberate act of exercising greater caution in the completion of the tasks, wherein subjects might require of themselves a higher degree of confidence before committing to a response. This explanation is similar in spirit to Brendl, Markman, and Messner's (2001) suggestion that subjects might increase their response activation thresholds in response to an increase in the perceived difficulty of the task (i.e., with tasks in the incompatible task blocks being more difficult or demanding than those in other task blocks). In addition, the increase in confidence criteria is consistent with Beck and Clark's (1997) proposal that anxiety activates reflective modes of thinking. In our terms, the process of exercising additional caution or adopting more stringent response criteria is analogous to an upward shift in  $x_{thres}$ , which is the level of activation that  $action_L$  or  $action_R$  must reach before a motor action is performed.

We have already shown in the previous section that a decrease in propagation gain would be accompanied by longer response latencies as well as more pronounced IAT effects (see Figure 2). Without having to repeat this simulation, the same reasoning in explaining the relationship between cognitive fluency and performance on the IAT can be applied here to account for the increase in IAT effects due to slowdowns in information processing (as a result of task interference). Such reductions in the rate of information processing (or propagation gain, as shown in Figure 2) would have resulted in stronger IAT effects, confirming the observations in Frantz et al. (2004).

Our next and remaining task is to examine the impact that increasing the response threshold  $x_{thres}$  would have on IAT performance. So far,  $x_{thres}$  has been set to the maximum pos-

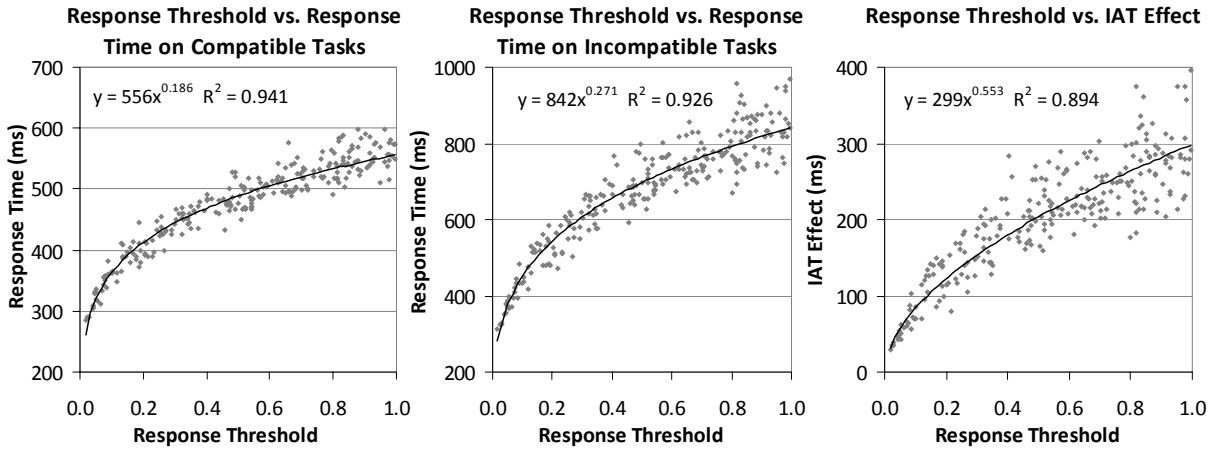


Figure 4. Distributions of response latencies on compatible and incompatible task blocks and corresponding IAT effects with variations in response threshold  $x_{thres}$  for 250 virtual subjects configured with a relative preference for one target concept over the other.

sible activation level of 1.0, but in this simulation, its value is varied within a reasonable range to determine how it affects response latencies on the combined tasks and the corresponding IAT scores. We begin by generating a population of 250 virtual subjects and initializing them with the same associative strengths as in the previous simulation, namely with  $EA \leftrightarrow POS$  and  $AA \leftrightarrow NEG$  set to 0.5 while  $AA \leftrightarrow POS$  and  $EA \leftrightarrow NEG$  were set to -0.5. Connections in each virtual subject's network were then randomly perturbed with Gaussian noise, before being put through all five standard tasks on the simulated Race-IAT.

Simulated response latencies on both the compatible and incompatible task blocks, and corresponding IAT effects are shown in Figure 4. We observe that higher response thresholds resulted in longer response latencies on both combined task blocks, as well as more pronounced IAT effects. In addition, the increase in response threshold is accompanied by higher variance in both response latencies and IAT effect. Thus, to the extent that an increase in the response activation threshold can be interpreted as the coping strategy of exercising caution or raising the confidence criterion, such a strategy could be responsible for a portion of the increase in IAT effect, as predicted by the simulation.

## Discussion

The fact that cognitive ability has a moderating effect on IAT performance has important implications for the validity of the IAT as a measure of the automatic associations and not some other construct. Suppose two subjects have IAT effect scores in the same direction but with different magnitudes. The fact that one of them has an IAT effect with a larger magnitude than the other would not necessarily imply that he or she definitely endorses a stronger positive implicit attitude towards the favored target concept, or a stronger negative implicit attitude towards the less preferred target concept, since the larger IAT effect could be due to differences (e.g., a reduction) in cognitive ability. In the case of this simulation, the associative strength configurations in all

250 subjects were the same (apart from their superposed random perturbations), yet they had a wide distribution of response latencies and IAT effects through just the manipulation of the processing gain and decay rate. For this reason, other means of calculating IAT effects, such as standardizing IAT scores with the “improved scoring algorithm” (Greenwald, Nosek & Banaji, 2003) should be utilized if a reasonable between-subject comparison is desired, although there is continued debate on this matter (cf. Blanton, Jaccard, Gonzales & Christie, 2006; Nosek & Sriram, 2006).

The computational model that we have employed provides some insight about reactions or strategies that might be adopted by subjects during the IAT to cope with stereotype threat. The idea that subjects actually engage in the effortful control of information processing (through either enhancement or suppression) is supported by empirical evidence from fMRI and EEG studies (Gazzaley, Cooney, McEvoy, Knight & D’Esposito, 2005). Furthermore, the ability to exercise greater caution in terms of focusing attentional resources and raising the response threshold or confidence criterion (Treisman & Faulkner, 1984; Petrusic & Baranski, 2009) is equally plausible, especially if the presence of stereotype threat increases the perceived difficulty of the tasks (cf. Brendl, Markman & Messner, 2001). The remaining research question lies in clarifying the nature of these mechanisms through experiment involving human subjects, especially with the advent of neuro-imaging techniques (e.g., Gazzaley et al., 2005; Stanley, Phelps & Banaji, 2008).

## Conclusion

Using a computational model of performance on the IAT, we examined the influence that cognitive fluency, and strategies for coping with stereotype threat could have on the processing of implicit attitudes. By varying several critical model parameters that are analogous to human cognitive ability (such as processing speed and information retention ability), the model accounts for the correlation between



longer response latencies and IAT effects that arises with lower cognitive ability. Furthermore, a reduction in information processing, and the adoption of a more conservative response criterion (which was modeled as shifts in response thresholds) was found capable of reproducing the exacerbated IAT effects that were empirically observed when stereotype threat was present.

## Acknowledgments

We wish to thank the anonymous reviewers for their valuable feedback and suggestions. B.-K. Quek is supported by a postdoctoral fellowship from the Agency for Science, Technology and Research (A\*STAR), Singapore.

## References

- Beck, A. T., Clark, D. A. (1997). An information processing model of anxiety: Automatic and strategic processes. *Behavior Research and Therapy*, 35, 49–58.
- Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the Implicit Association Test: Implications for criterion prediction. *Journal of Experimental Social Psychology*, 42, 192–212.
- Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology*, 81, 760–773.
- De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, 135, 347–368.
- Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, 69, 229–238.
- Frantz, C. M., Cuddy, A. J. C., Burnett, M., Ray, H., & Hart, A. (2004). A threat in the computer: The Race Implicit Association Test as a stereotype threat experience. *Personality and Psychology Bulletin*, 30, 1611–1624.
- Gazzaley, A., Cooney, J. W., McEvoy, K., Knight, R. T., D'Esposito, M. (2005). Top-down enhancement and suppression of the magnitude and speed of neural activity. *Journal of Cognitive Neuroscience*, 17, 507–517.
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem and stereotypes. *Psychological Review*, 102, 4–27.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85, 197–216.
- Hummert, M. L., Garstka, T. A., O'Brien, L. T., Greenwald, A. G., & Mellott, D. S. (2002). Using the Implicit Association Test to measure age differences in implicit social cognitions. *Psychology and Aging*, 17, 482–495.
- Hunt, E. (1983). On the nature of intelligence. *Science, New Series*, 219, 141–146.
- Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology*, 81, 774–788.
- Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology*, 93, 353–368.
- Lansman, M., Donaldson, G., Hunt, E., Yantis, S. (1982). Ability factors and cognitive processes. *Intelligence*, 6, 347–386.
- McFarland, S. G., & Crouch, Z. (2002). A cognitive skill confound on the Implicit Association Test. *Social Cognition*, 20, 483–510.
- Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the Implicit Association Test. *Journal of Personality and Social Psychology*, 85, 1180–1192.
- Nosek, B. A., & Sriram, N. (2007). Faulty assumptions: A comment on Blanton, Jaccard, Gonzales and Christie (2006). *Journal of Experimental Social Psychology*, 43, 393–398.
- Ottaway, S. A., Hayden, D. C., & Oakes, M. A. (2001). Implicit attitudes and racism: Effects of word familiarity and frequency on the implicit association test. *Social Cognition*, 19, 97–144.
- Page, M. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioral and Brain Sciences*, 23, 443–512.
- Petrusic, W. M., & Baranski, J. V. (2009). Probability assessment with response times and confidence in perception and knowledge. *Acta Psychologica*, 130, 103–114.
- Quek, B.-K., & Ortony, A. (2011). Modeling underlying mechanisms of the Implicit Association Test. In L. Carlson, C. Hoelscher, & T.F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp.1330–1335). Austin, TX: Cognitive Science Society.
- Revelle, W. (1986). Motivation and efficiency of cognitive performance. In D. R. Brown & J. Veroff (Eds.), *Frontiers of Motivational Psychology: Essays in honor of J. W. Atkinson*. Berlin: Springer.
- Rothermund, K., & Wentura, D. (2004). Underlying processes in the Implicit Association Test: Dissociating salience from associations. *Journal of Experimental Psychology*, 133, 139–165.
- Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, 103, 403–428.
- Stanley, D., Phelps, E., & Banaji, M. (2008). The neural basis of implicit attitudes. *Current Directions in Psychological Science*, 17, 164–170.
- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African-Americans. *Journal of Personality and Social Psychology*, 69, 797–811.
- Treisman, M., & Faulkner, A. (1984). The setting and maintenance of criteria representing levels of confidence. *Journal of Experimental Psychology*, 10, 119–139.

# Modeling the Effect of Evaluative Conditioning on Implicit Attitude Acquisition and Performance on the Implicit Association Test

Boon-Kiat Quek (boonkiat.quek@northwestern.edu)

Andrew Ortony (ortony@northwestern.edu)

Department of Psychology, Northwestern University, Evanston, IL 60208, USA, and  
Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

## Abstract

Using a previously proposed computational model of human performance on the Implicit Associations Test (IAT), we explore how evaluative conditioning could inform attitude acquisition and formation of automatic associations in memory, and demonstrate the effects of such learning on implicit task performance on the test. This is achieved by augmenting the model with a learning mechanism based on a modified Hebbian learning rule that adapts associative strengths between concepts depending on the temporal proximity of their activation. By manipulating the frequencies at which different stimuli are paired and presented as input to the network, we demonstrate how virtual subjects could acquire associative strengths that were subsequently reflected in simulated IATs as stronger relative preferences in favor of targets that were more frequently presented with positively-valenced stimuli. The model predicts that associations that are already strong have limited prospects for continued reinforcement.

**Keywords:** Hebbian learning; implicit attitudes; simulation; localist-connectionist networks.

## Introduction

Much discussion over the emergence of automatic associations between concepts and their evaluations in memory has taken place within the context of evaluative and classical conditioning (e.g., De Houwer, 2007; De Houwer, Baeyens & Field, 2005; Olson & Fazio, 2001; 2002). Evaluative conditioning is defined as a change in the extent of liking or disliking towards a stimulus that is caused by the frequent pairing of that stimulus with other liked or disliked stimuli (De Houwer, Baeyens & Field, 2005).

The interest in evaluative conditioning research is fueled by the fact that it has the potential to explain the emergence of attitudes and account for the ways in which people's attitudes and beliefs, and consequently their behavior, could be influenced. Thus, it has wide implications especially with regards to consumers' preferences, tastes, and purchasing habits. For instance, Gibson (2008) recently demonstrated the effect of evaluative conditioning in influencing implicit attitudes towards mature brands (e.g., Coke and Pepsi). It was shown that the consistent pairing of positive stimuli with a particular brand could help create and strengthen positive attitudes towards that brand, although the effect was observed only for subjects who had relatively neutral attitudes towards both brands to begin with. Olson and Fazio (2001; 2002) reported similar conditioning effects in which frequent pairings between novel conditioned stimuli (CS) and valenced unconditioned stimuli (US) could result in the acquisition of implicit attitudes towards novel target con-

cepts that were created *a propos* for the experiments, and consequently influence subjects' behaviors and responses on Implicit Association Tests (IAT; Greenwald, McGhee & Schwartz, 1998) involving those novel targets, even though subjects reported no explicit memory of the CS-US pairings.

However, the causal mechanisms by which the evaluative conditioning effect could emerge have yet to be satisfactorily uncovered, owing in part to conflicting empirical data about the conditions under which such effects might occur (De Houwer, Baeyens & Field, 2005). Many controversies revolve around whether associations were learnt as a result of automatic as opposed to conscious controlled processes, whether evaluative conditioning effects were due to a repertoire of processes (as opposed to a single mechanism) or contingent on subjects' awareness of stimuli pairing, and whether the learning is resistant to extinction (De Houwer, 2007; Walther, Weil & Dusing, 2011).

This paper represents our attempts at providing a computational account of the effect of evaluative conditioning on the acquisition of automatic associations between concepts in memory. Through simulations, we examine the impact that frequent pairing of target stimuli with various positively or negatively valenced stimuli would have on implicit task performance, such as on the Implicit Association Test. This is done with a number of goals in mind. First, to provide additional support for the cognitive plausibility of a previously proposed computational model of implicit task performance on the IAT (Quek & Ortony, 2011). Our approach is to augment the localist-connectionist model with a cohesive explanatory account of how automatic associations between concepts in memory could be formed or acquired through experience, a process analogous to how various attitudes are acquired throughout an individual's lifetime.

A second goal is to determine if we could make use of the computational model to address some of the research gaps identified by De Houwer, Baeyens and Field (2005), especially in view of what they see as a lack in the availability of detailed accounts for the processes and mechanisms that underlie evaluative conditioning, and the conditions under which it could occur. More generally, and as pointed out by Van Overwalle and Sieber, (2005), there appears to be limited theoretical advancement in the "understanding of the storage or strengthening of attitude-object associations in human memory." Before more empirical insights are made available, computational approaches such as modeling and simulation could provide an interim but effective means for understanding various candidate processes underlying attitude acquisition or formation (e.g., Eiser, Fazio, Stafford &

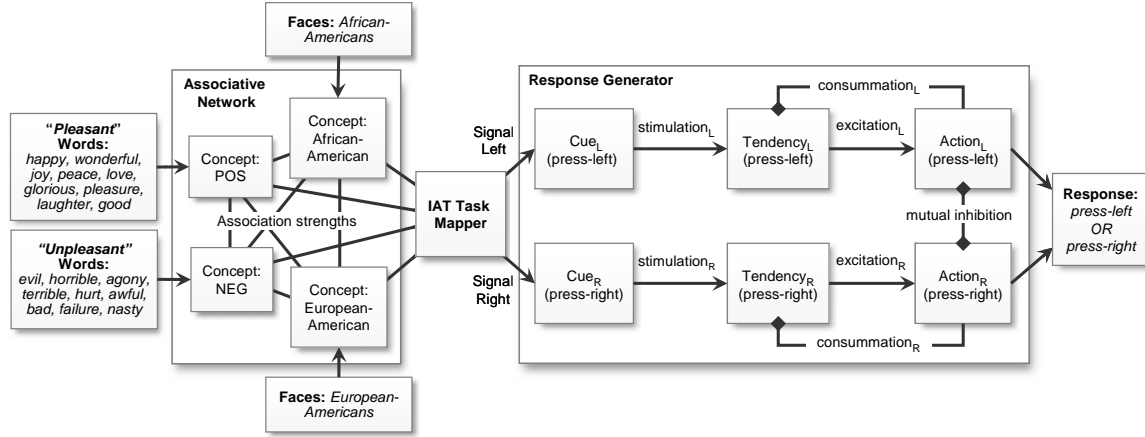


Figure 1. Network model for simulating performance on the IAT (Quek & Ortony, 2011)

Prescott, 2003; Van Overwalle & Sieber, 2005). In our case, a computational model that is demonstrably capable of replicating IAT effects on the basis of different associative strengths between concepts could serve as a platform on which various candidate learning mechanisms could be evaluated, by examining their impact on performance on the IAT. Doing so would also provide an example to demonstrate how learning mechanisms could be incorporated into localist-connectionist models, to fulfill a gap pointed out by some theorists that current associative models of attitudes lack mechanisms that could learn or update internal states and representations in response to information obtained externally from the world, as compared to connectionist models (Van Overwalle & Sieber, 2005). Finally, providing a psychologically plausible mechanism for how associative strengths in the network could be learnt would help allay potential criticisms and modeling concerns over the seemingly arbitrary manner in which associative weights in the earlier model were configured or initialized.

### Model Overview

In this section, we provide a brief overview of the computational model used (for more details, see Quek & Ortony, 2011). The model is a localist-connectionist network (e.g., Page, 2000) that emulates the multiple processing pathways from visual perception (i.e., a word or image) to the automatic activation of associated concepts in memory and motor responses. In general, nodes in the network represent concepts while connections represent associations between them. Information is processed in the network through the flow of activation from one node to another, a process governed by the following propagation rule:

$$x_i(k+1) = (1-\delta)x_i(k) + \alpha \sum_{j,i \in E} x_j(k) \cdot w_{j,i}(k), \quad (1)$$

where  $x_i$  is the activation level of a node  $v_i$ ,  $w_{j,i}$  is the weight of the connection  $e_{j,i}$  from a node  $v_j$  which is a neighbor of  $v_i$ ,  $E$  is the set of all edges,  $\alpha$  is the propagation gain (set to 0.2) and  $\delta$  is a decay parameter (set to 0.001) that reduces activation over time. In each time step  $k$ , activation spreads to  $v_i$  from each of its neighbors  $v_j$  at a rate proportional to the

weight  $w_{j,i}$  of the connection between them. Positive values of  $w_{j,i}$  are excitatory while negative values are inhibitory, while a value of zero implies a neutral or null connection.

### Model Components

The network comprises a few components (see Figure 1). The *Associative Network* contains nodes representing the target concepts AFRICAN-AMERICAN (AA) and EUROPEAN-AMERICAN (EA), attribute concepts for positivity (POS) and negativity (NEG), input stimuli such as a list of *pleasant* and *unpleasant* words (e.g., *happy*, *wonderful*, *joy*, *evil*, *horrible*, *hurt*), and pictures of *European-American* and *African-American* individuals. Connections between target-attribute concept node pairs (i.e.,  $EA \leftrightarrow POS$ ,  $EA \leftrightarrow NEG$ ,  $AA \leftrightarrow POS$ , and  $AA \leftrightarrow NEG$ ) are taken to represent implicit attitudes. For example, a positive attitude towards EA can be represented as excitatory  $EA \leftrightarrow POS$  or inhibitory  $EA \leftrightarrow NEG$  associations, or both, such that when EA is activated, POS will be similarly activated while NEG would be inhibited. Similarly, negative attitudes towards EA can be represented by excitatory  $EA \leftrightarrow NEG$  or inhibitory  $EA \leftrightarrow POS$  associations, or both, such that activation of EA would excite NEG but inhibit POS.

The *Task Mapper* is responsible for transmitting activation from target and attribute concepts to  $cue_L$  and  $cue_R$  which are nodes indicating that a left or right key-press is required. If the present task requires a right response for “*European-American* or *pleasant*”, both POS and EA would be routed to  $cue_R$ . These connections are reconfigured at the beginning of each task block, and during which they remain active (see Quek & Ortony, 2011, Figure 2).

The *Response Generator* implements Revelle’s (1986) cue-tendency-action model (CTA), which in turn is based on Atkinson and Birch’s (1970) dynamics of action theory. CTA describes the dynamic interactions between conflicting tendencies and competing actions. Using CTA as a template, we construct two response-generating pathways (for the left and right key-presses). When activated, response cue nodes will stimulate action-tendency nodes, which will activate response nodes representing the left and right motor responses. When either of the response nodes exceeds a certain activation threshold, it is taken as the winner.

The interactions between the above representations take place as excitations and inhibitions along different propagation pathways. For example, in a task block requiring a left key for “*African-American* or *unpleasant*” and a right key for “*European-American* or *pleasant*”, an African-American picture would activate  $AA$ , and activation will be transmitted to  $cue_L$ . However, if the network is configured with a strong  $AA \leftrightarrow POS$  connection, activation will also be transmitted to  $cue_R$ , competing with  $cue_L$ . This reduces the rate that activation will accumulate in the left response node, and thus a longer time is required for it to reach the response threshold.

### Simulating the Implicit Association Test

Each virtual subject’s network is first initialized with a set of associative strengths that represents its implicit attitudes, and put through the standard IAT task blocks. The network is provided with a simulated verbal or pictorial input in each trial. The number of iterations taken to produce a response is recorded, and then transformed by a scaling factor into a simulated response time (in milliseconds) of the same order of magnitude as those observed in human subjects (e.g., Greenwald et al., 1998; Klauer, Voss, Schmitz & Teige-Mocigemba, 2007). To compute the IAT effect, we take the raw difference between the simulated mean response times in the two combined task blocks.

### Simulating Evaluative Conditioning

To examine the effect that learning processes might have on IAT performance, it would be necessary to extend the localist-connectionist model with mechanisms that could modify its internal features in response to environmental input. While the use of learning is a mainstay of connectionist and parallel distributed processing models (e.g., Cohen, Dunbar & McClelland, 1990; McClelland & Rumelhart, 1986; Read et al., 2010), it is relatively uncommon in localist-connectionist models (Page, 2000).

A number of connectionist models for simulating the automatic acquisition of associations in memory have been proposed (e.g., Eiser, Fazio, Stafford & Prescott, 2003); these typically employ some form of error-correction learning (such as the ubiquitous *delta rule*) that adjusts weights to learn particular stimulus-to-response mappings such that the actual and expected outcomes will eventually converge over time. It is unclear if this is a realistic portrayal of the manner in which associations between concepts are learnt or formed, since the notion of what an *expected outcome* or *reward* ought to be, is ill-defined, or at best, arbitrary. For instance, frequent exposure to a pair of conditioned and unconditioned stimuli need not necessarily involve a motor response or behavioral outcome, though it can be accompanied by a change in state—which in this case would be an increase or decrease in the associative strength between concepts in memory, which can be taken as a change in the degree of liking or disliking for the said stimuli. Work by Herz, Sulzer, Kühn and van Hemmen (1989), and more recently Verguts and Notebaert (2008) employed Hebbian learning rules to learn such state changes.

Hebbian learning (or *plasticity*, Hebb, 1949) can be construed as a form of reinforcement learning in which connections between nodes that *fire* (in the context of neural networks) or are jointly activated within a temporally proximate timeframe would be strengthened over time, such that future joint activation of the associated nodes would co-occur with greater ease. Mathematically speaking, the Hebbian learning rule can be characterized as:

$$\Delta w_{i,j} = \lambda \cdot (x_i \cdot x_j) \quad (2)$$

where  $\lambda$  is a learning rate parameter,  $x_i$  and  $x_j$  are the activation levels of two nodes  $v_i$  and  $v_j$ , while  $w_{i,j}$  is the weight of the edge  $\varepsilon_{i,j}$  originating from node  $v_i$  and terminating at  $v_j$ . In neural networks,  $x_i$  and  $x_j$  are known as the pre- and post-synaptic activation levels of the connection between  $v_i$  and  $v_j$ , respectively. The product  $x_i x_j$  can be conceived as a measure of similarity between the activation levels of both nodes. The learning rule causes the connection weight between these two nodes to increase proportionately with respect to the degree in which both nodes are temporally activated together. However, this rule is known to be unstable in that connection weights will tend to increase without bounds over time if repeatedly reinforced, or saturate at their maximum and minimum boundaries. To enhance stability, we add a discounting term representing the portion of activation in  $v_j$  that is not due to  $v_i$ :

$$\Delta w_{i,j} = \lambda \cdot (x_i \cdot x_j) \cdot (x_j - x_i w_{i,j}). \quad (3)$$

Doing so ensures that  $w_{i,j}$  will be adapted in relation to only that portion of the activation in  $v_j$  that is not due to  $v_i$ , which prevents  $w_{i,j}$  from over-learning the joint activation between  $v_i$  and  $v_j$ . Thus, associations that are already strong to begin with will cease to increase without bounds. Our formulation of the Hebbian learning rule is similar to the simple but provably stable form proposed by Oja (1982):

$$\Delta w_{i,j} = \lambda \cdot x_j \cdot (x_i - x_j w_{i,j}). \quad (4)$$

The difference between the two formulations is that we have swapped the roles of  $x_i$  and  $x_j$  within the parentheses, and kept  $x_i$  in the product to preserve the role of the similarity term  $x_i x_j$ . Furthermore, we inserted a decay term to allow weights to gradually decay over time, in the absence of activation, to arrive at the following:

$$\Delta w_{i,j} = -\gamma \cdot w_{i,j} + \lambda \cdot (x_i \cdot x_j) \cdot (x_j - x_i w_{i,j}), \quad (5)$$

where  $\gamma$  is the weight decay rate. For implementation purposes, the learning rule is rewritten as an update function:

$$w_{i,j}(k+1) = (1-\gamma) \cdot w_{i,j}(k) + \lambda \cdot x_i(k) \cdot x_j(k) \cdot [x_j(k) - x_i(k) w_{i,j}(k)] \quad (6)$$

We further constrained the model to learn only the weights of associations between positively-activated concept nodes, while allowing associative weights between non-activated or negatively-activated (i.e., inhibited) node pairs to decay and eventually become extinct over time.

Prior to performing the simulation,  $\lambda$  and  $\gamma$  were set to 0.05 and 0.0005 respectively after an initial process of iterative search through parameter space to yield post-learning weights that had a large but unsaturated range.

## Simulations





To perform the simulations, we begin with a network configuration in which the weights of the associations  $EA \leftrightarrow POS$ ,  $EA \leftrightarrow NEG$ ,  $AA \leftrightarrow POS$ , and  $AA \leftrightarrow NEG$  are all initialized to zero. In each epoch, 100 pairs of input stimuli, each comprising an attribute concept exemplar (e.g., the word *wonderful*) and a target concept exemplar (e.g., a picture of a White or Black individual) were selected at random. Input nodes corresponding to both exemplars in each stimulus pair were set with an activation of 1.0. The learning rule in Equation (6) was then applied in tandem with the propagation rule defined in Equation (1). Propagation of activation through the network would activate the concept nodes corresponding to these input stimuli. At the same time, the learning rule is expected to enhance the connection weights between pairs of activated concept nodes, for instance, between EA and POS, or AA and POS, using the above example of the word *wonderful* and a picture of a White or Black individual.

By manipulating the frequency at which input exemplars are selected from each attribute and target concept pair, we can simulate situations in which the exemplars of specific target-attribute concept pairs co-occur more frequently than others. As an example, to produce the condition that EA and *pleasant* exemplars co-occur twice as often as EA and *unpleasant*, the frequency for the latter is set to half of the former's. We expect the learning rule to adapt association weights in a manner that will eventually reflect the patterns of distributions across the frequencies at which each target-attribute concept pair is presented.

In this first simulation, two learning conditions were defined, as shown in Table 1. In the first condition (a), the frequency distribution across the target-attribute concept pairs  $AA+POS$ ,  $AA+NEG$ ,  $EA+POS$ , and  $EA+NEG$  were set to 40%, 10%, 10%, and 40%, respectively. The second condition (b) was defined by the distribution 10%, 40%, 40%, and 10%, for target-attribute pairs in the same order. These represent the probability in which paired-stimuli are sampled from the respective concept pair, thus the absolute proportions themselves may vary. Virtual subjects in each condition were put through a pre-learning IAT, followed by the above learning phase during which 100 pairs of stimuli were presented for 100 epochs each. Finally, a post-learning IAT was administered to the virtual subjects. More details concerning the procedures in which the simulated IATs were conducted are found in Quek & Ortony (2011).

Figure 2 shows the evolution of target-attribute associative strengths over the course of learning for 25 virtual subjects in each condition, while the post-learning associative strengths are shown in Table 2. In condition (a), stronger  $AA \leftrightarrow POS$  and  $EA \leftrightarrow NEG$  associations emerged after learning, while  $AA \leftrightarrow NEG$  and  $EA \leftrightarrow POS$  increased but at a much slower rate. In (b), stronger associations were found for  $EA \leftrightarrow POS$  and  $AA \leftrightarrow NEG$ , while the remaining two increased but at a much slower rate. When put through both the pre-learning and post-learning IATs, condition (a) had a non-significant mean IAT effect of -0.03ms prior to learning,  $t(24) = -0.132$ ,  $p = 0.896$ , but a significant post-learning mean IAT effect of

Table 1: Presentation frequency of paired stimuli in two experimental conditions during the learning phase

Stimulus Pair	Prototypical exemplars	Presentation Frequency	
		Condition (a)	Condition (b)
$AA+POS$	 + “happy”	40%	10%
$AA+NEG$	 + “sorrow”	10%	40%
$EA+POS$	 + “laughter”	10%	40%
$EA+NEG$	 + “horrible”	40%	10%

Note: EA: European-American; AA: African-American; POS: Positivity; NEG: Negativity.

Table 2: Post-learning target-attribute associative strengths

Association	Condition (a)		Condition (b)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
$AA \leftrightarrow POS$	.865	.083	.287	.148
$AA \leftrightarrow NEG$	.289	.084	.862	.093
$EA \leftrightarrow POS$	.256	.120	.862	.091
$EA \leftrightarrow NEG$	.874	.061	.292	.117

Note: EA: European-American; AA: African-American; POS: Positivity; NEG: Negativity.

-124.8ms,  $t(24) = -19.0$ ,  $p < .001$ , which is indicative of an implicit preference for AA over EA. Similarly, condition (b) exhibited a non-significant pre-learning mean IAT effect of 0.24ms,  $t(24) = 0.771$ ,  $p = 0.448$ , but a significant post-learning mean IAT effect of 120.6ms,  $t(24) = 22.3$ ,  $p < .001$ , indicative of an implicit preference for EA over AA. Considering that each network began with non-significant pre-learning IAT test scores but expressed significant post-learning IAT effects, and since no other modifications were made to the network, we may conclude that the increase in IAT effect is due to the associations that were acquired over the course of learning. As expected, the emerging associative strengths in each condition (Table 2) showed a similar pattern to the distributions of presentation frequencies of the corresponding target-attribute pairs (Table 1).

To investigate the impact of different co-occurrence frequencies on the post-learning IAT effect, we repeated the above simulation for 250 virtual subjects, only this time varying the frequency distribution for each subject by interpolating randomly between 50%, 0%, 0%, 50%, and 0%, 50%, 50%, 0% for the respective target-attribute concept pairs  $AA+POS$ ,  $AA+NEG$ ,  $EA+POS$ , and  $EA+NEG$  that were presented during learning. When the proportions of both  $AA+POS$  and  $EA+NEG$  stimuli were reduced from 50% to 0%, the proportions of  $AA+NEG$  and  $EA+POS$  stimuli were increased from 0% to 50%, in a complementary manner, while ensuring that all four proportions add up to 100%.

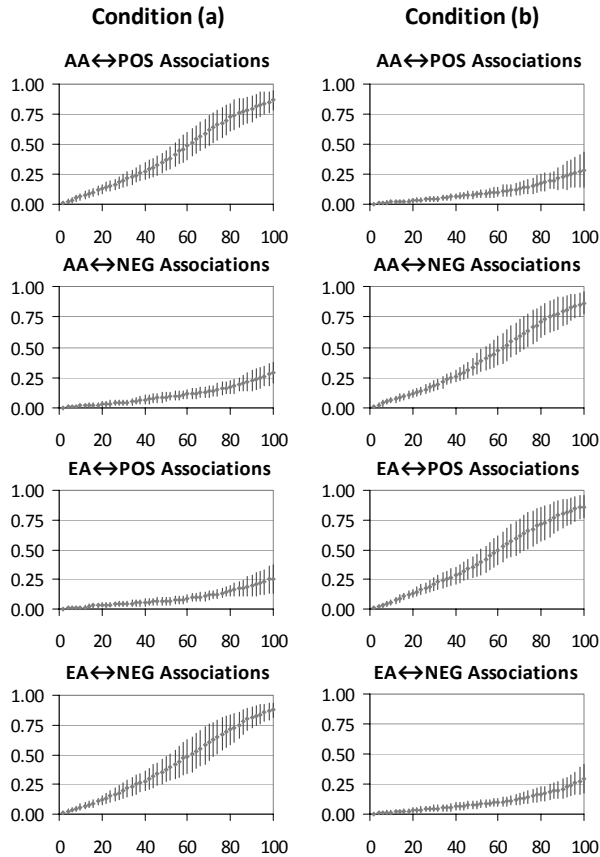


Figure 2. Evolution of associative strengths over the course of learning, for virtual subjects in two conditions. EA: European-American; AA: African-American; POS: Positivity; NEG: Negativity. Y-axis: associative strengths. X-axis: learning epochs. Error bars: standard deviations.

Plotting the post-learning IAT effect against the presentation frequencies for the four target-attribute concept pairs in Figure 3, we found that when a larger proportion of EA+POS and AA+NEG paired stimuli were presented to the model during the learning phase, the post-learning IAT subsequently produced larger IAT effects that were in favor of EA. Conversely, when more input stimulus pairs were selected from AA+POS and EA+NEG and presented to the model during learning, the post-learning IAT had larger IAT effects in favor of AA. When all input stimulus pairs were presented with about the same probability (i.e., keeping the proportions to 25% for each target-attribute concept pair), the post-learning IAT effect was close to zero.

## Discussion

With the computational model, we have demonstrated how automatic associations between target and attribute concepts could be acquired by repeated exposure to pairs of input exemplars—as similarly achieved in human subjects via evaluative or classical conditioning (De Houwer, 2007; Olson & Fazio, 2001). Stronger associations were acquired for target-attribute concept pairs whose input exemplars were presented together more frequently, and weaker associations

were learnt for other target-attribute concept pairs whose input exemplars were presented together less frequently.

These simulations have some important implications especially with regards to the malleability of implicit attitudes. First, the ability to influence or generate novel associations through consistent pairing of target and attribute stimuli supports the findings of Olson and Fazio (2001) and of Gibson (2008), particularly the latter's discovery that the effects of evaluative conditioning were observed only for subjects who initially had relatively neutral attitudes towards the targets, and not those who already possess a significantly stronger preference for one target over the other. In our terms, this could be explained by the longer amount of time required for stronger associative strengths to decay or weaken over time when the corresponding paired stimuli were no longer presented as frequently.

Second, the evolution of associative strengths over learning epochs in Figure 2 showed a gradual slowdown as they approached 1.0, suggesting that, as these associations increase in strength over the course of learning, the extent to which they can be further increased is limited. Thus, there is limited room for the continued positive reinforcement of associations whose strengths are already high, such that they become less susceptible to learning. Consistent with empirical observations (Gibson 2008; Joy-Gaba & Nosek, 2010), the model thus predicts that this would limit the impact that evaluative conditioning might have on attitudes that have already been firmly ingrained, and thus the continued malleability of attitudes through such means could be reduced. While it could be argued that this effect is largely a result of the modified Hebbian learning rule we devised in Equation (6) that limits the extent to which already-strong associations could continue to be increased, the weights will nonetheless be subject to the finite upper boundary even when the standard unconstrained Hebbian rule in Equation (2) were used instead, and give rise to the same observations.

Third, the simulation results so far are in agreement with Mitchell, Anderson and Lovibond's (2007) proposal that the IAT itself could be used as a means for detecting the occur-

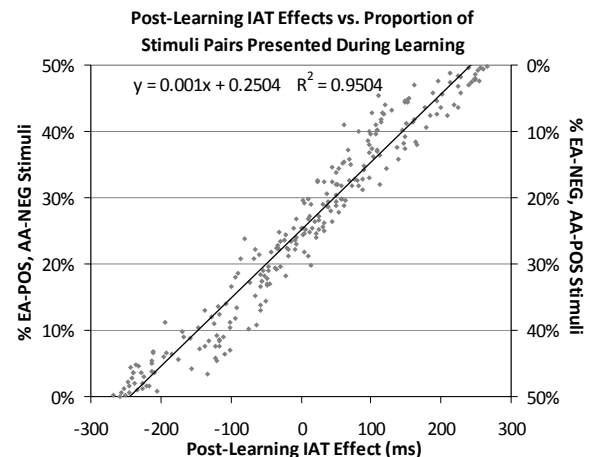


Figure 3. Post-learning IAT effects in virtual subjects (N=250) across presentation frequencies of input stimuli from each of the target-attribute concept pairs during the learning phase.

rence of evaluative conditioning, which, to be consistent with Gibson (2008), is to be expected only for target concepts that have yet to be strongly associated with any particular attributes. Finally, considering that the simulated mechanisms of learning are not specific to valenced attributes, they could be relevant not just for evaluative conditioning, but for explaining other more generic forms of conditioning or learning, such as the effectiveness of re-affirmations to enhance self-concept and self-esteem.

## Conclusion

In summary, we have augmented the cognitive plausibility of our computational model (whose purpose was to account for the emergence of IAT effects) by providing a cohesive and cognitively-plausible account of the manner in which implicit attitudes could be acquired through evaluative conditioning, as well as their subsequent effects on implicit task performance on a simulated IAT. This is achieved via a modified Hebbian learning rule that adapts associations between concept representations in memory relative to the different frequencies at which target stimuli are paired with other positively or negatively valenced stimuli. An additional contribution of the model is in demonstrating how localist connectionist models too are amenable to learning mechanisms, just like their connectionist counterparts (Van Overwalle & Sieber, 2005). Extending the simulations beyond the permitted scope of this paper to include additional learning conditions and a more comprehensive analysis of the viability of the learning algorithms presented (in comparison to possibly other candidates) would be a logical continuation of this work in future.

## Acknowledgments

We wish to thank the anonymous reviewers for their valuable feedback and suggestions. B.-K. Quek is supported by a postdoctoral fellowship from the Agency for Science, Technology and Research (A\*STAR), Singapore.

## References

- Atkinson, J. W., & Birch, D. (1970). *The dynamics of action*. New York: John Wiley.
- Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop Effect. *Psychological Review*, 97, 332–361.
- De Houwer, J. (2007). A Conceptual and Theoretical Analysis of Evaluative Conditioning. *The Spanish Journal of Psychology*, 10, 230–241.
- De Houwer, J., Baeyens, F., & Field, A. P. (2005). Associative learning of likes and dislikes: Some current controversies and possible ways forward. *Cognition and Emotion*, 19, 161–174.
- Eiser, J. R., Fazio, R. H., Stafford, T., & Prescott, T. J. (2003). Connectionist simulation of attitude learning: Asymmetries in the acquisition of positive and negative evaluations. *Personality and Social Psychology Bulletin*, 29, 1221–1235.
- Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research*, 35, 178–188.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley & Sons.
- Herz, A., Sulzer, B., Kühn, R., & van Hemmen, J. L. (1989). Hebbian learning reconsidered: Representation of static and dynamic objects in associative neural nets. *Biological Cybernetics*, 60, 457–467.
- Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology*, 41, 137–146.
- McClelland, J. L., & Rumelhart, D. E. (1986). A distributed model of human learning and memory. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. II. Psychological and Biological Models* (pp. 170–215). Cambridge, MA: MIT Press/Bradford Books.
- Mitchell, C. J., Anderson, N. E., & Lovibond, P. F. (2003). Measuring evaluative conditioning using the Implicit Association Test. *Learning and Motivation*, 34, 203–217.
- Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15, 267–273.
- Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, 12, 413–417.
- Olson, M. A., & Fazio, R. H. (2002). Implicit acquisition and manifestation of classically conditioned attitudes. *Social Cognition*, 20, 89–103.
- Page, M. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioral and Brain Sciences*, 23, 443–512.
- Quek, B.-K., & Ortony, A. (2011). Modeling underlying mechanisms of the Implicit Association Test. In L. Carlson, C. Hoelscher, & T.F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp.1330–1335). Austin, TX: Cognitive Science Society.
- Read, S. J., Monroe, B. M., Brownstein, A. L., Yang, Y., Chopra, G., & Miller, L. C. (2010). A neural network model of the structure and dynamics of human personality. *Psychological Review*, 117, 61–92.
- Revelle, W. (1986). Motivation and efficiency of cognitive performance. In D. R. Brown & J. Veroff (Eds.), *Frontiers of Motivational Psychology: Essays in honor of J. W. Atkinson*. Berlin: Springer.
- Van Overwalle, F., Siebler, F. (2005). A connectionist model of attitude formation and change. *Personality and Social Psychology Review*, 9, 231–274.
- Verguts, T., & Notebaert, W. (2008). Hebbian learning of cognitive control: Dealing with specific and nonspecific adaptation. *Psychological Review*, 115, 518–525.
- Walther, E., Weil, R., & Düsing, J. (2011). The role of evaluative conditioning in attitude formation. *Current Directions in Psychological Science*, 20, 192–196.



# Constraints, Inferences, and the Shortest Path: Which paths do we prefer?

Marco Ragni (marco.ragni@cognition.uni-freiburg.de)

Center for Cognitive Science,  
University of Freiburg, Germany

Jan M. Wiener (jwiener@bournemouth.ac.uk)

Psychology Research Centre,  
Bournemouth University, UK

## Abstract

How do we reason about incomplete spatio-temporal descriptions? How might a map influence formerly constructed preferred mental models? Little research so far focused on a combination of two central fields important for successful route planning: the way humans deal with constraint based reasoning (especially with some sort of spatio-temporal constraints) and the way in which humans plan with a given map (especially with problems inspired by typical Traveling Salesman Problems). This, however, becomes even more interesting in cases in which the spatio-temporal constraints allow for several solutions. Do the predictions of the preferred mental model theory still hold true in such situations? This article investigates the influence of maps on the generation of preferred models. The goal is to bring together the theory of (preferred) mental models and route planning.

**Keywords:** Spatial reasoning; preference effects

## Introduction

In everyday life we often reason with incomplete information or have to take constraints into account during reasoning. Cognitive processes involved in such reasoning about spatial relations and the construction of according mental models have recently been the subjects of interest in studies about spatial relations (Knauff, Rauh, Schlieder, & Strube, 1998; Rauh et al., 2005). However, the question of how external representations of space such as maps, or map-relevant knowledge influences and interacts with reasoning processes is widely unknown. The research communities concerned with how people use maps to solve spatial or navigational problems and how people solve reasoning problems are mostly distinct. There are, however, many situations in which people reason with maps or with map-like knowledge. In this paper we present and investigate two classes of problems.

**Path planning from maps.** Imagine planning a sightseeing trip through the downtown area of an unfamiliar city: you do have a map and you want to visit multiple sites of interest. Of course, you are interested in minimizing the distance you have to traverse along your tour. Problems of this kind are typically referred to as Traveling Salesperson

Problems (TSP): A salesman has to visit a number of cities and start from a specific location to which he will also return after visiting each city. The traveling salesman will aim for the shortest possible route and avoid any detours (Wiener & Tenbrink, 2008). Formally, TSP-Problems are NP-complete (Garey & Johnson, 1979).

Human performance and the cognitive strategies employed when solving TSPs have been investigated in real environments involving movement through space (e.g., Gärling & Gärling, 1988) as well as in more abstract visual or map-like versions of the TSP in which a number of dots are displayed on a computer screen which have to be connected such that the resulting tour is as short as possible (e.g., MacGregor & Ormerod, 1996). When planning actual site seeing trips, however, we often face additional constraints besides minimizing distances: some sites of interest may close before others and therefore have to be visited earlier. Or, you may want to be at a specific site at a particular time, for example, to have lunch. In addition, you are still striving to minimize path length. Similar challenges arise when planning shopping trips during which multiple stores have to be visited. Here, we also often face additional constraints besides minimizing distances: Frozen food or ice-cream, for example, is best be bought towards the end of the shopping tour to avoid defrosting before returning home. Moreover, in order to minimize the effort of carrying purchased goods, heavy items should be bought towards the end of the trip. Again, path length should be minimized. All these factors impose constraints on the path-planning problem and have to be taken into account when planning a trip. Below is an example of combined spatial optimization and reasoning problem:

- (1) Buy bread before ice-cream.  
Buy eggs after ice-cream  
Buy a gallon of water after eggs.  
Buy a chair after a gallon of water.

Problems (1) belongs to a class of problems that are referred to as *determinate problems*, as they allow only for a single solution:

bread ice-cream eggs water chair

Problem (2) belongs to a class of problems that are referred to as *indeterminate problems*, as they allow for multiple – three – solutions.

- (2) Buy bread before ice-cream.  
Buy eggs after ice-cream.  
Buy a gallon of water after ice-cream.  
Buy a chair after a gallon of water

Which is consistent with the following three models:

bread ice-cream eggs water chair  
bread ice-cream water eggs chair  
bread ice-cream water chair eggs

The key idea of the *mental model theory* is that reasoners translate these constraints into a mental model – an abstraction or analogical reflection – of the state of affairs and use this representation to solve the reasoning problem. An important finding is that when faced with indeterminate problems featuring multiple solutions, humans tend to construct only one initial model – the so-called *preferred mental model* (Rauh et al., 2005; Ragni, Fangmeier, Webber, & Knauff, 2007), which is easier to maintain in working memory than any other mental model (Ragni et al., 2007; Knauff, 2006). Preferred mental models have been initially identified for Allen's interval calculus (Knauff, Rauh, & Schlieder, 1995), a more detailed introduction of preferred mental models is given in the next section. What happens when reasoning about a problem – as the one described above – when the shortest path does not correspond to the preferred mental model? Is any influence measurable? Although this question is of high ecological validity, to the authors' knowledge, it has not yet been approached. In this paper we will present a first experiment to analyze from the perspective of a mental model theorist whether – and if so, how – preferred mental models can be “overridden” by external stimuli.

## Background

### The theory of preferred mental models

A central question in the context of incomplete information is: How are indeterminate problems such as Problem (2) processed? Are there preferred interpretations? The mental model theory (MMT), introduced by Johnson-Laird and Byrne (1991), suggests that people draw conclusions by constructing and inspecting a spatial array that represents the state of affairs described in the premises. It is a three-stage process consisting of a comprehension, description, and validation phase. In the comprehension phase, reasoners construct a mental model that reflects the information from the premises. If new information is encountered during the reading of the premises it is immediately used in the construction of the model. During

the description phase, this model is inspected to find new information that is not explicitly given in the premises. Finally, in the validation phase alternative models are searched that refute this putative conclusion. However, some questions remain open with regards to how people deal with multi-model problems. For example, which model is constructed first, and does this model construction adhere to certain principles? And, why do reasoners neglect some models? None of these questions are answered by the classical mental model theory. In contrast the preferred mental model theory (PMMT) has been developed to explain why humans in general tend to construct a preferred mental model (PMM). The PMM is the starting point for deriving a putative conclusion. In the model variation phase the participants tend to make local and continuous transformations starting from the PMM to search counter-examples (Rauh et al., 2005).

Several predictions of the PMMT about insertion principles as well as transformation strategies in spatial relational reasoning can be shown (Ragni et al., 2007). Assume we have two premises of the form

- (1) A is to the left of B and
- (2) A is to the left of C.

Humans tend to process these premises sequentially, i.e. first a model A B is generated and then object C is inserted into the model. There are two possibilities where C can be inserted, in-between A and B (first-fit principle) and to the right of B (first-free-fit principle). The latter principle has been empirically confirmed in small-scale descriptions (e.g., Ragni et al., 2007; Jahn et al., 2005). An interesting aspect, however, is how this might influence reasoning if a map is given?

**Path planning and Distance Optimzation.** Path planning and optimization with maps has primarily been investigated by means of visual versions of the TSP (e.g., Graham, Joshi, & Pizlo, 2000; Vickers, Butavicius, Lee, & Medvedev, 2001). In these experiment, participants are presented with a number of target locations on a computer screen – usually presented as identical black dots on a white background – and are asked to connect these locations with straight lines such that the resulting path is as short as possible. Results from these studies demonstrate that humans reach very good performance levels even with as many as a few dozen target locations. The strategies and heuristics applied are a matter of ongoing debate. The convex hull has been suggested to be part of the problem solving strategy (MacGregor & Ormerod, 1996), the crossing avoidance hypothesis states that participants avoid crossing tours, as they know that crossings lead to sub-optimal solutions (Van Rooij, Stege, & Schactman, 2003), and the hierarchical nearest neighbor strategy assumes that in a first step clusters of several neighboring dots are established, which are then sequentially linked into a tour, using the nearest neighbor algorithm (Vickers, Bovet, Lee, & Hughes, 2003).

Only few studies investigating TSPs with maps used richer environments in which different target locations could be visually distinguished requiring some form of memory. In a recent study, Tenbrink and Wiener (2009) presented participants with maps depicting a regular 5x5 grid of locations each of which could be identified by a unique symbol. Participants were given so-called shopping lists depicting the symbols of a start location and four to nine target locations. Their task was to identify the locations in the grid and then mark the shortest possible round trip from the start that visits all target locations in the map. By analyzing participants' planning performance, their chosen paths, as well as retrospective linguistic representations, a number of cognitive strategies applied when solving the TSPs could be identified. Most importantly, participants flexibly employed and connected a repertory of multifaceted strategies allowing them to simplify and structure the problem space across subtasks involved in solving the TSPs (for a navigational version of this paradigm, see Wiener, Ehbauer, & Mallot, 2009).

As mentioned before, path planning in every-day life often requires taking into account additional constraints besides minimizing distances. Hayes-Roth and Hayes-Roth (1988) presented one of the view studies investigating complex planning from maps with additional constraints (but see also the related Plan-A-Day paradigm, Nellen & Funke, 2002). Participants were given a map of a town depicting multiple shops and other locations along with a list of errands. These errands included buying vegetables at the grocery, buying a toy for a dog at the pet store (both purely spatial constraints), but also picking up a car at a certain time in a certain location (spatio-temporal constraint). Moreover, more errands were specified than the subject could possibly accomplish in the time available, which required him/her to sort out (less important) errands to formulate a realistic plan. Hayes-Roth & Hayes-Roth developed a general model of complex planning, assuming that the planning process comprises many distinct *specialists* contributing decisions to a tentative plan that is refined incrementally.

## Experiment – Reasoning, Route Planning, and Maps

In this experiment we investigated the connection between the construction of (preferred) mental models from a set of premises and the subsequent task of planning a trip consistent with the premises. In order to do so, participants were presented with determinate and indeterminate reasoning problems describing spatio-temporal relations between sets of destinations. After processing the premises and (possibly) constructing a (preferred) mental model, they were asked to draw a round trip into a map visiting the destinations in an order that is consistent with the premises. If the planning task in fact interfered with the constructed mental model, we expected performance differences depending on whether or not the round trips defined by the

premises were optimal or clearly sub-optimal with respect to path length.

### Participants.

Nineteen students from the University of Freiburg took part in this experiment (9 females,  $M = 23.3/22.1$ ,  $SD = 2.2/2.1$ ). They were paid for their participation or received course credits.

### Materials.

To investigate the impact of map like presentation of target locations on reasoning performance and the selection of preferred mental models, we generated four types of reasoning problems (see Fig. 1).

1. **Optimal determinate problem (D-optimal):** The correct solution to these reasoning problems always matches the shortest possible – optimal – route to visit all target destinations.
2. **Suboptimal determinate problem (D-sub-optimal):** The correct solutions to these reasoning problems were clearly suboptimal with respect to their length.
3. **Preferred optimal indeterminate problems (IP-optimal):** The preferred mental model to these reasoning problems matched the shortest possible – optimal – route. Two alternative correct solutions existed that were not identical with the preferred mental model and that were sub-optimal with respect to their length.
4. **Preferred suboptimal indeterminate problems (IP-suboptimal):** The preferred mental models to these reasoning problems were clearly suboptimal with respect to their length. Two alternative correct solutions existed, one of which was optimal with respect to metric length.

### Methods.

Each participant was presented with 16 reasoning problems, four of each type described above. To control for the influence of the specific configuration of start and target places, we used four different configurations and balanced the types of reasoning problems across the configurations. Each reasoning problem was presented on three pages: The first page contained the first two premises; the second page contained the third and fourth premises, and the third page contained a regular  $5 \times 5$  grid in which the 5 positions mentioned in the premises were marked (see Figure 1). Participants were instructed to read premises 1 and 2, to turn the page over, read premises 3 and 4, turn the page over, and

to connect the positions in the layout in order to mark a round trip that was consistent with the premises. They were instructed not to scroll back after having turned a page or to take any notes.

## Hypotheses & Predictions.

Given the specific procedures of the experiment, two competing hypotheses are conceivable: **First**, the external representation and the task of sketching the corresponding route do not influence the reasoning process. This is based on the assumption that the mental model is generated while reading and processing the premises. Hence, the external representation that is provided only after the last premises was processed does not influence the mental model. Participants would then simply sketch the tour that corresponds to their mental model. In case of determinate problems this would lead to identical performance (with respect to error rate) between the types of reasoning problems (D-optimal/D-suboptimal). In case of indeterminate reasoning problems, we expect that participants will select the preferred mental model, regardless of whether or not the according path was optimal or suboptimal (IP-optimal/IP-suboptimal). **Second**, the external representation influences the mental model, as the task of sketching a round trip for the shopping route implicitly requires choosing a short solution. In this case, we expect interferences between finding the correct solution to the reasoning problem and planning the shortest path. Such an interference would have a selective impact on performance for determinate reasoning problems of type *D-suboptimal*, for which the shortest (optimal) path and the correct solution to the reasoning problem were different, but not for determinate reasoning problems of type *D-optimal*, for which the optimal path and the correct solution to the reasoning problem were identical. The predictions for indeterminate problems are not as straight forward, as each indeterminate problem features three different solutions.

## Results.

Three out of the 19 participants were removed from the final data as their performance on finding the correct solution for determinate problems was clearly below 50% (12.5%, 12.5%, 37.5%). In addition, thirteen trials were removed from the final data set, as these solutions featured branching points – participants had drawn two arrows from one location.

The different spatial configurations had no influence on participants' performance ( $F(3, 46.35) = .27, p = .85$ ). For the remaining analyses we therefore pooled the four different configurations. On average, participants found a correct solution to 89.1% of the reasoning problems. A 2x2 ANOVA with the factors of *type of reasoning problem*

(determinate, indeterminate) and *solution* (optimal, suboptimal) was carried out. We did not observe a main effects for type of reasoning problem [ $F(1, 17.99) = 0.12, p = .91$ ] or for solution [ $F(1, 18.28) = .06, p = .81$ ]. However, the interaction type of reasoning problem x solution was significant [ $F(1, 16.72) = 8.96, p < .01$ ].

To further investigate the nature of this interaction, we performed post hoc *t* tests revealing that performance for determinate problems of type *D-optimal* was better than for determinate problems of type *D-suboptimal* (93.2% vs. 82.8%; *t*-test:  $t(15) = 2.24, p = .04$ , see Figure 2). For indeterminate reasoning problems, the pattern was different: surprisingly, participants performance was better for problems of type *IP-suboptimal* than for those of type *IP-optimal* (97.8% vs. 82.8%; *t*-test:  $t(14) = 2.38, p = .03$ , see Figure 2).

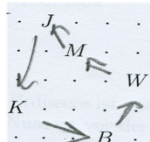
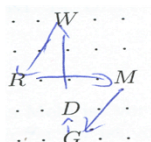
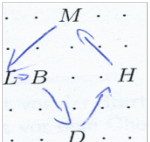
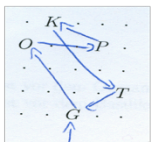
Determinate optimal	Determinate suboptimal
Premises 1. K is before B 2. B is before W 3. W is before M 4. M is before J 	Premises 1. G is before D 2. D is before W 3. W is before R 4. R is before M 
Indeterminate preferred = optimal	Indeterminate preferred = suboptimal
Premises 1. L is before B 2. B is before D 3. B is before H 4. H is before M 	Premises 1. G is before O 2. O is before P 3. O is before K 4. K is before T 

Figure 1: The four different types of reasoning problems along with exemplary data by participants. All participants received the premises (in German) with full names, e.g. der Trevibrunnen vor dem Kolosseum (the Fountain of Trevi before the Colosseum) instead of initials.

For correct solutions to indeterminate problems, we evaluated whether or not participants chose the preferred mental model. In 87.8% of the cases, they did choose the preferred mental model (*t*-test against chance level [with three possible solutions, chance level was 33.33%]:  $t(15) = 15.79, p < .001$ ).

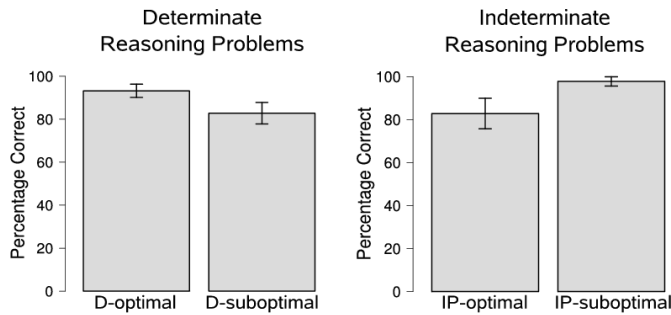


Figure 2: Results of the Experiment: **left:** the determinate problem description – allowing for one path solution only; **right:** the indeterminate problem description, allowing for three solutions.

Preference for the preferred mental model did not differ between types of indeterminate problems (IP-optimal: 87.2% versus IP-suboptimal: 85.0%;  $t$ -test:  $t(14)=0.29$ ;  $p=.78$ ).

## Discussion.

The findings for the determinate problems to which only a single correct solution exists, revealed a clear difference in performance. Specifically, participants showed better performances for problems in which the correct solution to the reasoning problem was identical to the shortest possible path (D-optimal) as compared to reasoning problems in which the correct solution and the optimal path were different (D-suboptimal). This finding suggests interference between the reasoning process and the task of planning a round trip. In other words, the map influenced the reasoning process.

Contrary to classical deduction tasks, indeterminate reasoning problems do not appear to be more difficult than determinate ones. A possible explanation for the lack of a systematic difference in the current paradigm comes from the fact that the higher number of possible solutions in indeterminate problems in the specific task allows for a higher error tolerance (in all cases the objects 3, 4, and 5 in the premises).

Some of the participants had drawn routes with branching points, i.e., they had drawn two arrows from one object. Such branching solutions were mostly found in indeterminate problem cases (14 out of the 17 cases in total). We had to remove these cases from the final data set as we were not able to extract a single unambiguous solution. However, these branching solutions clearly reflect a special type of errors, as they usually reflected the indeterminate nature of the problems. Note, however, that by removing these trials from the analysis, we artificially increased performance primarily for indeterminate problems, which might explain the surprisingly high performance in these problems.

An analysis of the chosen solutions for indeterminate problems clearly demonstrated that participants did not choose randomly between the three possible solutions, but preferred one over the others. The preferred solution was identical to the one generated by the first-free fit strategy, a preferred mental model generation strategy identified in previous experiments (Ragni et al., 2007) on small-scale scenarios. Then again, the constraints in this experiment were clearly spatio-temporal in their nature – the premises “the fountain of Trevi before the colosseum” refers to the sequence of the events. In that sense, it is not surprising that the identified preferences were similar to those identified in small-scale scenarios (Schaeken, Johnson Laird, & d’Ydewalle, 1996).

## General Discussion

In dealing with maps there is one important and new question: What is the influence of the (implicit) task of planning a short path using maps while taking into account spatio-temporal constraints? The way reasoners typically construct preferred mental models when reasoning about indeterminate problems has been identified in several experiments (cp. Ragni et al., 2007). The most prominent encoding strategy applied in such cases is the first-free-fit strategy. This strategy, however, does not allow for predicting how external constraints such as the length of the routes resulting from reasoning problems influence the reasoning process itself. In this study we combined reasoning about spatio-temporal constraints with the task of planning short paths with a map (without explicitly stating that the shortest path must be found). The planning task influenced the reasoning task: In determinate problems – in which only a single solution existed – participants showed better performance if that solution was identical to the shortest possible path. Furthermore, for indeterminate cases, we found strong preferences for the solution that corresponded to the first-free-fit strategy. Participants’ performance in finding a possible solution for IP-suboptimal problems was greater than for IP-optimal problems. This result is surprising and was not predicted, as the optimal solution to the route-planning problem – i.e., the shortest possible route – was identical to the preferred mental model for IP-optimal problems but not for IP-suboptimal problems. Future research will address this interesting effect.

## Acknowledgments

This paper presents work done in the project R8- [CSPACE] of the Transregional Collaborative Research Center SFB/TR 8 Spatial Cognition. Funding by the German Research Foundation (DFG) is gratefully acknowledged.

## References

- Gärling, T., & Gärling, E. (1988). Distance minimization in downtown pedestrian shopping. *Environment and Planning A*, 20, 547-554.
- Graham, S. M., Joshi, A., & Pizlo, Z. (2000). The travelling salesman problem: A hierarchical model. *Memory & Cognition*, 28 (7), 1191-1204.
- Hayes-Roth, B., & Hayes-Roth, F. (1988). A cognitive model of planning. In A. Collins & E. E. Smith (Eds.), *Readings in cognitive science: A perspective from psychology and artificial intelligence* (p. 496-513). San Mateo, CA: Kaufmann.
- Jahn, G., Knauff, M., Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. *Memory & Cognition*, 35, 2075-2087.
- Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.
- Knauff, M. (2006). Deduktion und logisches Denken. In J. Funke (Ed.), *Denken und Problemlösen. Enzyklopädie der Psychologie* (Vol. 8). Göttingen: Hogrefe.
- Knauff, M., Rauh, R., & Schlieder, C. (1995). Preferred mental models in qualitative spatial reasoning: A cognitive assessment of Allen's calculus. In (p. 200-205). Mahwah, NJ: Lawrence Erlbaum Associates.
- Knauff, M., Rauh, R., Schlieder, C., & Strube, G. (1998). Mental models in spatial reasoning. In *Spatial cognition* (p. 267-292).
- MacGregor, J. N., & Ormerod, T. C. (1996). Human performance on the traveling salesman problem. *Perception and Psychophysics*, 58, 527-539.
- Nellen, S., & Funke, J. (2002). The role of exploration and forward checking in human scheduling. Cognitive Science 2002 Conference, Fairfax, Virginia, August 8-10, 2002.
- Ragni, M., Fangmeier, T., Webber, L., & Knauff, M. (2007). Preferred mental models: How and why they are so important in human spatial reasoning. In C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, & T. Barkowsky (Eds.), *Spatial cognition v*. Berlin: Springer.
- Rauh, R., Hagen, C., Knauff, M., Kuss, T., Schlieder, C., & Strube, G. (2005). Preferred and alternative mental models in spatial reasoning. *Spatial Cognition and Computation*, 5, 239-269.
- Schaeken, W., Johnson Laird, P. N., & d'Ydewalle, G. (1996). Mental models and temporal reasoning. *Cognition*, 60, 205-234.
- Tenbrink, T., & Wiener, J. (2009). The verbalization of multiple strategies in a variant of the traveling salesman problem. *Cognitive Processing*, 10 (2), 143-161.
- Van Rooij, I., Stege, U., & Schactman, A. (2003, Mar). Convex hull and tour crossings in the Euclidean traveling salesperson problem: implications for human performance studies. *Memory & Cognition*, 31 (2), 215-220.
- Vickers, D., Bovet, P., Lee, M. D., & Hughes, P. (2003). The perception of minimal structures: performance on open and closed versions of visually presented Euclidean travelling salesperson problems. *Perception*, 32 (7), 871-886.
- Vickers, D., Butavicius, M., Lee, M., & Medvedev, A. (2001). Human performance on visually presented traveling salesman problems. *Psychological Research-Psychologische Forschung*, 65 (1), 34-45.
- Wiener, J., Ehbauer, N., & Mallot, H. (2009). Planning paths to multiple targets: memory involvement and planning heuristics in spatial problem solving. *Psychological Research*, 77 (5), 644-658.

# Think Outside the Box: The Effects of Cognitive Training on Creative Problem Solving

Jared T. Ramsburg (jramsb2@uic.edu)

University of Illinois at Chicago  
Department of Psychology (MC 285)  
1007 West Harrison Street  
Chicago, Illinois, 60607-7137

Robert J. Youmans (ryouman2@gmu.edu)

George Mason University  
Department of Psychology  
4400 University Drive, MSN 3F5  
Fairfax, VA 22030

## Abstract

Problem solving requires the use of higher mental functions, functions that can be improved with training (Brown, Ryan, & Creswell, 2007). The current study examines the effects of *meditation* on creative problem solving. Participants were undergraduate students ( $n = 81$ ) who were randomly assigned to meditate or rest. Next, Pp were asked to solve a problem: fishing out a small object from inside a box using one of four available tools. Two of the available tools were potentially useful, but the other two were intentionally designed to be useless (i.e., they were incapable of retrieving the object). There were no differences between meditators and non-meditators with respect to solution rates, tool switching behavior, or overall persistence. However, meditators spent more time with their first tool that they selected, and more time attempting to solve the task with the useful tools. Brief meditation training may promote certain cognitive strategies that are conducive to successful problem solving; implications are discussed.

**Keywords:** problem solving; cognitive training; meditation.

## Introduction

Recently, investigations have demonstrated that *cognitive training*, activities designed to improve performance, self-control, and brain physiology, provides an effective method for improving a variety of higher order mental functions including reasoning, decision-making, and task-set switching (Basak, Boot, Voss, & Kramer, 2008; Mestre, Dufresne, Gerace, & Hardiman, 1993; Willis et al., 2006). Research suggests that a broad range of cognitive training techniques is effective. Manger, Eikenland, and Asbjornsen (2002) were able to improve the social-cognitive functioning of female schoolchildren with a 9-month long social-cognitive training program. Basak et al. (2008) utilized 23.5 hours of training using a real-time strategy video game, and found that the training improved task switching, working memory, and reasoning. Smith et al.

(2009) found that a computerized cognitive training program improved attention and memory of aged adults. Indeed, these and many other studies demonstrate that cognitive training can improve cognitive performance in a variety of areas when used over the long term.

The study described here investigated a very specific form of cognitive training called focused *meditation*, a process where a person attempts to sustain attention on a selected thought, detect mind wandering, and return to focused attention. Meditation is a form of cognitive training, and studies have demonstrated the benefits of meditation on cognitive functions (Ramsburg & Youmans, 2011; Tang et al., 2007; Zeidan, Johnson, Diamond, David, & Goolkasian, 2010). Meditation techniques are also relatively simple to teach, and some research has demonstrated measurable improvements in cognitive function even with very little meditation experience. Ramsburg and Youmans (2011) found that initial periods of meditation that lasted only six minutes at the start of a class lecture improved academic performance amongst lower division college students. Tang et al. (2007) found that five 20-minute sessions of meditation reliably improved attentional functioning, while Zeidan et al. (2010) found improvements in attentional functioning, working memory, and visuo-spatial processing with only four sessions of meditation training.

## Problem solving and meditation?

General problem solving skills involve an ability to understand a problem, devise a plan of action, execute the plan, and examine the results (Newell & Simon, 1972). An important question in the problem solving literature is whether the individual steps within the problem solving process can be enhanced, and the answer it would seem is 'maybe.' For example, we know that experts in a given domain are often superior to novices at both understanding a problem and devising a plan of action. However, we also know that when the rules of a domain are compromised



(e.g., impossible chess positions) or when an expert in one domain is asked to perform a task in a different domain, so-called ‘experts’ often fail to *transfer* their expertise between the two domains (Ohlsson, 2011).

Although meditation has been shown to improve a variety of cognitive functions, relatively few studies have directly examined the influences of meditation training on problem solving (Dillbeck, 1982; Kindler, 1979; Raingruber & Robinson, 2007). Kindler (1979) found that group problem solving could be improved with meditation training, specifically, improving speed to solution and promoting effective teamwork. Raingruber and Robinson (2007) found using a qualitative approach that nurses engaged in meditation training reported improvements in their problem solving abilities, attention, and calm. These findings suggest that meditation *might* be an effective method for improving problem solving, but more study is clearly required.

If meditation were to enhance problem solving, we hypothesize here that its influence might be most pronounced on the actual execution of any given problem solving strategy. Whereas expertise is likely to act on the stages of problem solving that require accurate problem framing (i.e., understanding problems and devising solutions), meditation and other types of cognitive training might have positive effects on some of the key cognitive processes utilized while trying to execute a plan of action during problem solving (Lutz et al., 2009; Mayer, 1992; Tang et al., 2007; Zeidan et al., 2010). Ly and Spezio (2009) found via fMRI that meditation could improve decision-making by influencing neural circuits in an enduring manner for recruitment during the self-regulation of social cognitive processes. Other studies have demonstrated that mediators appear to display more conscientious decision making (Kirk, Downar, & Read Montague, 2011), or cognitive flexibility, a mental ability important in problem solving (Dillbeck, 1982; So & Orme-Johnson, 2001). These and other studies suggest a link between meditation and other forms of cognitive training, and the cognitive functions used during problem solving itself.

In this study, we examined the effect of meditation, a form of cognitive training, on a real-world, creative problem-solving task. Participants received either brief meditation training or rest, and then attempted to solve a novel creative problem. Although past studies have shown only moderate training benefits on problem solving tasks (e.g., Kershaw & Ohlsson, 2004), we hypothesized that meditation, which has been shown to improve cognitive functioning (So & Orme-Johnson, 2001; Tang et al., 2007) may be an effective method for promoting creative problem solving in general.

## Method

### Participants

Eighty-one California State University, Northridge students participated for course credit. Participants had an average age of 18.73 with a standard deviation of 1.32. There were

61 females and 20 males. Participants identified themselves as Latino/Hispanic (45.7%), Black/African-American (19.8%), Caucasian (19.8%), more than one race/other (6.2%), Asian or Pacific Islander (4.9%), and Middle Eastern (2.5%).

### Problem-solving task

The researchers developed a novel creative problem-solving task. The task was a physical problem that required the use of unique experimenter designed tools to solve the task. Specifically, the objective was to get a bolt out of a box and across a red line (six and a half feet away from the box) using the four experimenter designed tools.

### Materials

**Box and Bolt.** The box (see Figure 1) was made using 5 ½ by ¾ inch wood, a 16 inch in length, 3 inch high, and 1/16 inch thick piece of clear flexible plastic, and a 23 by 18 inch plywood base. The back face of the box measured 17 ¼ inches. The two wooded sides measured 23 inches in length. The clear plastic front piece with red tape lining its top was 3 inches high and was glued into place 16 ½ inches from the back side of the box, which left two 6 ¾ inch in length non encased walls. A modified 5/8<sup>th</sup> inch lag bolt was the target that participants were trying to retrieve. The bolt was placed standing upright inside the box at the beginning of each experiment. The bolt was 3 inches in length had seven 1 ¾ inch washers secured by two 5/8<sup>th</sup> inch nuts.

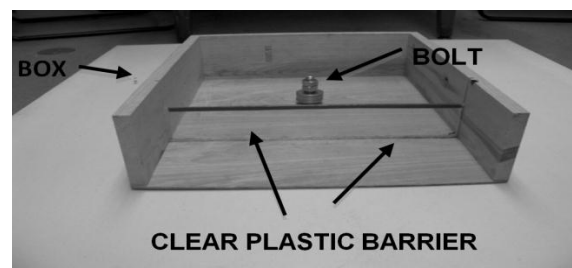


Figure 1. The Box with the Bolt in the Starting Position.

**Claw.** The claw was one of the two tools that were capable of retrieving the bolt (see Figure 3). It consisted of a modified metal grabber extension tool. The claw was 53 ¾ inches long. A piece of thin rope connected the “trigger” at the handle, and the grasping claw. The grasp of the claw was weak, making the straightforward method of using the tool to retrieve the bolt difficult, but not impossible.

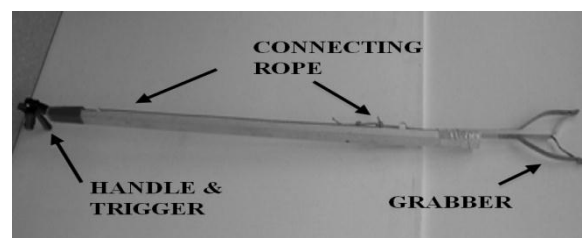


Figure 2. The Claw.

**Magnet.** The magnet was the second tool that was capable of retrieving the bolt (see Figure 4). It consisted of a 35 ½ inch long PVC pipe that was 1 ¼ inches in diameter that was connected to another PVC pipe that was 12 inches in length. The two pieces of PVC pipe were joined by a spring that was 5/6 inch diameter and 8 inches long. A rope was laced through the top that hung 34 ½ inches. At the end of the rope was a 2-inch diameter magnet. The magnet was too weak to lift the bolt outright, but it could be used to drag the bolt. Using the magnet and loose string in combination, it was difficult, but possible, to retrieve the bolt.

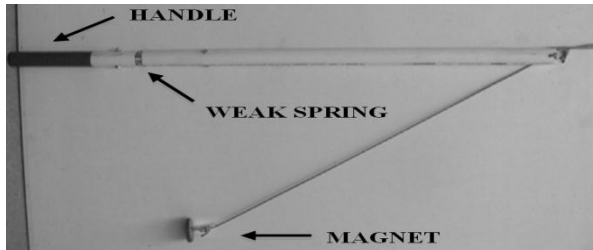


Figure 4. The Magnet.

**Spatula.** The spatula was one of two tools that were not capable of retrieving the bolt. It consisted of a 49-inch long piece of flexible PVC pipe that was 5/8 inches in diameter had a plastic paint scrapper fastened to one end by three screws (see Figure 5). The other end had a less flexible PVC pipe that was 11 ½ inches long and 5/8 inch in diameter attached. At each of these ends was a 23 ½ inch long PVC pipe that would serve as the handles for the tool. The flexibility of the tool made the tool useless.

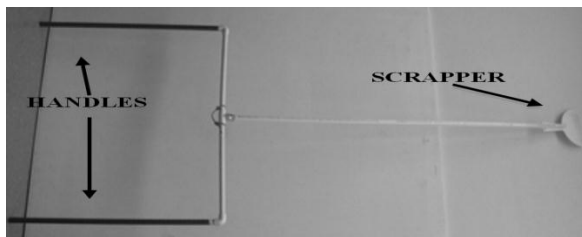


Figure 5. The Spatula.

**Ring.** The ring was the second tool that was not capable of retrieving the bolt. It consisted of a 48-inch long fiberglass rod that was 1/8 inches in diameter where at one end of the rod a thin rope was attached that was 44 inches in length (see Figure 6). At the end of the rope a 4 ½ inch long heavy-duty carabineer was tied; this carabineer served as the 'ring' that participants would use for 'hooking' the bolt. The shape of the ring made the tool useless.

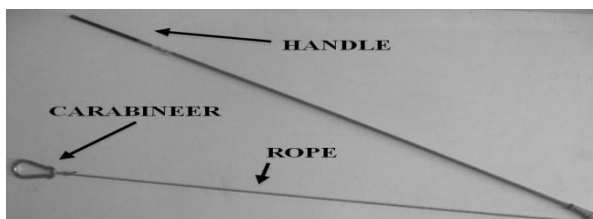


Figure 6. The Ring.

**Self Report Measures.** Six items administered pre and post examined feelings of tension, ease, anxiety, self-confidence, nervousness, and overexcitement measured on five-point likert scales (1 = Not at all to 4 = Very Much So). Additionally post measures for 'how difficult was the task?' (1 = very difficult to 5 = very easy) and 'how frustrating was the task?' (1 = very frustrating to 5 = not frustrating) were recorded.

**Video Camera.** A handheld digital-video camera on a stationary platform was used to record tool switch behavior and allow for an objective measure of time.

## Design & Procedure

The study was an experimental design. Participants entered the laboratory individually and were given consent forms to review, after which, they were given a brief mood questionnaire. This was followed with a 10-minute audio recording of eyes closed guided meditation or eyes closed rest.

Following the 10-minute training participants received instructions for the creative problem-solving task. Participants were informed that they would be participating in a creative problem-solving task with the objective to remove a bolt from a box and bringing it across a red line using four experimenter-designed tools. Importantly, participants had to obey the following four rules: 1) you must stay within the red box when using a tool 2) you can only touch the red and black parts of the tool when in use 3) you can only use one tool at a time 4) you can switch tools back and forth at your own discretion. Any participant that had questions or needed clarification was given further instruction if necessary for understanding the task and rules.

The participants were informed that the camera would be recording tool usage. When a participant either solved the task or quit the camera was turned off and the participant then filled out questionnaires assessing mood, perceived frustration of the task, perceived difficulty of the task, and demographics information. Finally, participants were debriefed and thanked for their time.

## Results

The hypothesis was that meditation training would result in successful completion of the task. For those that failed to solve the task meditation would result in greater initial persistence, overall persistence, and fewer tool switches. Tool use behavior was also examined in order to determine favoritism in tool usage. No significant differences were found for mood,  $F < 1$ , frustration,  $F(1, 78) = 1.17, ns.$ , or task difficulty,  $F(1, 79) = 1.3, ns.$  The results revealed no significant difference in success rates between meditation (36.59%) and rest (35%),  $\chi^2(81) = .02, ns.$  Overall tool success rates were as follows: claw (20.99%), magnet (13.58%), spatula (0%), and ring (0%; see Figure 7 for results by condition).

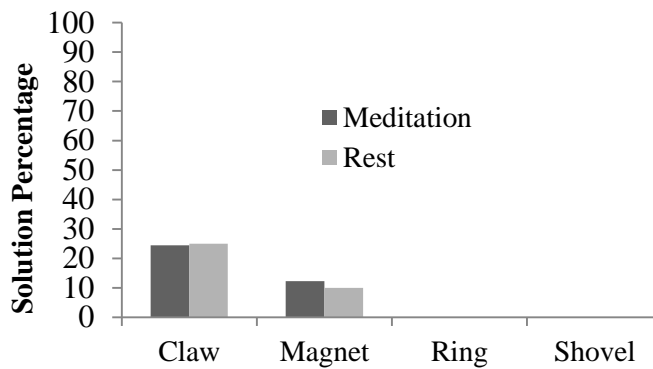


Figure 7. Success rates by tools, split by condition.

Of the remaining 52 participants that failed to solve the task, 51 were used in the other analyses. One was removed due to lost data. The meditation condition ( $M = 60.44$  seconds,  $SD = 29.21$ ) compared to the rest condition ( $M = 43.62$  seconds,  $SD = 20.79$ ) spent significantly more time with the first tool,  $F(1, 49) = 5.65$ ,  $p < .05$ . No reliable differences were found between the meditation condition ( $M = 722.64$  seconds,  $SD = 534.2$ ) and the rest condition ( $M = 790.69$  seconds,  $SD = 539.06$ ) for total time spent,  $F < 1$ . No reliable differences were found between the meditation condition ( $M = 7.52$ ,  $SD = 3.72$ ) and the rest condition ( $M = 7.73$ ,  $SD = 3.52$ ) for tool switch behavior,  $F < 1$ .

In a series of follow-up analyses, we examined tool use behavior. Specifically, the time participants spent with the useful tools (i.e., tools that were successfully used to solve the task, the magnet and claw) and the useless tools (i.e., tools that no one could solve the task with, the spatula and ring), in order to, determine whether participants favored the useful over the useless tools. Post-hoc Tukey's HSD tests showed that meditators spent more time with the useful tools ( $M = 436.88$  seconds,  $SD = 420.07$ ) than the useless ones ( $M = 227.13$  seconds,  $SD = 227.62$ ),  $p < .05$ , and that resters had no reliable differences between time with the useful tools ( $M = 372.82$  seconds,  $SD = 310.02$ ) compared to the useless ones ( $M = 276$  seconds,  $SD = 303.56$ ),  $p > .05$ ; see Figure 8). In sum, success rates did not differ between groups, nor did overall persistence, or tool switch behavior, but those that meditated had a greater initial persistence, and spent more time with the useful tools than the useless tools, where resters did not differ in tool use preference.

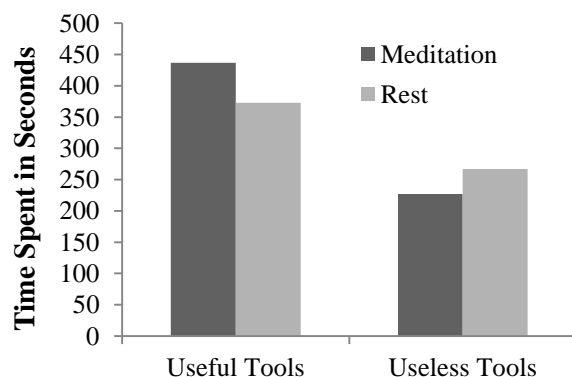


Figure 8. Time spent with tool types by condition.

## Discussion

Although our hypotheses were not fully supported, the findings suggest that meditation training might promote some important behaviors useful in problem solving. One could view the enhancement in initial persistence as taking the time to get to know the problem, and spending more time with the useful tools might suggest that meditators were recognizing what tools appeared to be getting them closer to the goal. Nonetheless, problem-solving success did not differ between the groups, but was higher than anticipated with  $1/3^{\text{rd}}$  of both conditions completing the task. However, the  $1/3^{\text{rd}}$  success rate may be low enough to demonstrate the difficulty of the task, where ratings of task difficulty did not differ between groups. One might expect that the meditation training might improve mood or reduce frustration with the task since meditation is often associated with better mood and less anxiety (Brown et al., 2007), but the brevity of the training as seen in other studies using limited training measures does not consistently result in improvements in mood (Ramsburg & Youmans, 2011).

In order to gage persistence we had to eliminate those that had solved the task because their times would reflect time to solution, and not persistence. When we examined persistence among those that failed to solve the task, we found no differences between the groups, which might suggest that brief meditation training does not enhance persistence. However, when examining initial persistence we found that those receiving the meditation training spent significantly more time with their initial chosen tool, perhaps, suggesting a need to become more familiar with the problem before attempting the task with other tools.

Notably, of the four tools available to the participants, two of the tools could be used to solve the task (claw and magnet) and two were useless (spatula and ring). We found that participants in both conditions did not differ in the number of switches they made, suggesting similar levels of flexibility. However, a closer inspection of the amount of time with the various tools revealed that meditators spent more time with the useful tools, whereas, the resters had no reliable difference in how they spent their time attempting to solve the task. The increase in time for useful tools among meditators may have resulted from recognition of their value for solving the task, perhaps as a result of improvements in cognitive functioning. Conceivably, attentiveness for usability of one tool over another might lead those with meditation training to use the tools that appear to be more effective given that meditation training is associated with improved attentional functioning (Tang et al., 2007). Nonetheless, more research is needed to determine what aspects of cognitive functioning might most benefit from meditation training when faced with a problem-solving task and whether success rates might increase with extensive training.

A problem with the present study that makes further interpretation of the results difficult stems from the finding that meditators did not improve in overall problem success.

One admittedly post-hoc explanation might be related to the observation that succeeding in this particular problem solving task actually required that participants not only select the tools that made the task possible, but use them in non-traditional ways to retrieve the bolt. Research suggests that chunk decomposition and constraint relaxation can account for success with creative or insight problems when faced with an impasse (see Knoblich, Ohlsson, Haider, & Rhenius, 1999). Chunk decomposition relates to the ability to unpack chunks of information to where detailed approaches can be vested. Constraint relaxation involves the ability to reduce the severity of constraints that may occur when faced with a problem. These processes appear to be activated via successive failures that elicit further decomposition and greater constraint relaxation. To be a successful problem solver in the present study, a participant would need to relax the constraints associated with how the tool can be used because solutions require participants to use the useful tools in a non-normal fashion in order to achieve the goal (e.g., the claw's grasp is ineffectual, but with some maneuvering the claw could be used like a shovel). Additionally, problem solving success may depend on decomposition of chunks associated with the possible strategies that when chunked are unsuccessful, but when decomposed may provide novel strategies (i.e., a chunked strategy can involve many steps that are thought of as one process, but when the steps are isolated novel divergent approaches could be adopted by combining different steps from different strategies). More broadly, a participant would need to think divergently, which is not a process influenced by meditation training, although thinking divergently is a process associated with some eastern philosophies that utilize meditation training (see Dogen, 2007). Our participants were only exposed to a brief 10-minute meditation exercise absent eastern philosophical approaches that expound divergent thinking. As such, in the absence of an enhancement in divergent thinking, neither group could be expected to think more creatively.

## Future Research

The present study presented findings that may suggest cognitive improvements with some aspects of problem solving. However, the present study did not adequately determine what underlying mechanisms cognitive training via meditation influenced. Future research might investigate what components of cognitive functioning, influenced by meditation, are responsible for enhancing problem solving. For instance, past research has shown that meditation can enhance creativity (Jedrczak, Beresford, & Clements, 1985; Travis, 1979), attention (Lutz et al., 2009), memory (Kozhevnikov, Louchakova, Josipovic, & Motes, 2009), and self-regulatory functioning (Brown, Ryan, & Creswell, 2007), but less is known about the applicability of these benefits to problem solving. Understanding how the training influences performance will help in determining how the training might be used and whether certain aspects of the training should be emphasized over other options.

The present study utilized a unique physical problem-solving task, with the objective of determining whether meditation training could improve creative problem solving performance. The results of the experiment left more questions than answers, demonstrating the breadth of studies that could follow, which might better clarify the processes responsible for the benefits seen in the present study. For instance, a brief ten-minute training exercise may have produced some benefits to problem solving; one might infer that more extensive training might further enhance performance, where deliberate practice has been known to enhance performance amongst novices and experts (Ericsson et al., 1993).

## References

- Basak, C., Boot, W. R., Voss, M. W., & Kramer, A. F. (2008). Can training in a real-time strategy video game attenuate cognitive decline in older adults? *Psychology and Aging, 23*, 765-777.
- Brown, K. W., Ryan, R. M., Creswell, J. D. (2007). Mindfulness: Theoretical foundations and evidence for its salutary effects. *Psychological Inquiry, 18*, 211-237.
- Dillbeck, M. C. (1982). Meditation and flexibility of visual perception and verbal problem solving. *Memory & Cognition, 10*, 207-215.
- Dogen, E. (2007). *Shobogenzo: The treasure house of the eye of the true teaching*. (H. Nearman, Trans.). Mount Shasta, CA: Shasta Abbey Press. (Original work published 1231-1253).
- Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363-406.
- Jedrczak, A., Beresford, M., & Clements, G. (1985). The TM-Sidhi program, pure consciousness, creativity and intelligence. *The Journal of Creative Behavior, 19*, 270-275.
- Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*, 63-85.
- Kershaw, T. C., & Ohlsson, S. (2004). Multiple causes of difficulty in insight: The case of the nine-dot problem. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 3-13.
- Kindler, H. S. (1979). The influence of a meditation relaxation technique on group problem-solving effectiveness. *The Journal of Applied Behavioral Science, 15*, 527-533.
- Kirk, U., Downer, J., & Montague, P. R. (2011). Interoception drives increased rational decision-making in meditators playing the ultimate game. *Frontiers in Neuroscience, 5*, 1-11.
- Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(6):1534-1555
- Kozhevnikov, M., Louchakova, O., Josipovic, Z., & Motes, M. A. (2009). The enhancement of visuospatial processing efficiency through Buddhist deity meditation. *Psychological Science, 20*, 645-653.
- Lutz, A., Slagter, H. A., Rawlings, N. B., Francis, A. D.,

- Greischar, L. L. & Davidson, R. J. (2009). Training enhances attentional stability: Neural and behavioral evidence. *The Journal of Neuroscience*, 29, 13418-13427.
- Ly, M. & Spezio, M. L. (2009). The effect of meditation on neural systems implicated in social judgments. *NeuroImage*, 47, S194.
- Manger, T., Eikeland, O.-J., & Asbjørnsen, A. (2002). Effects of social-cognitive training on students' locus of control. *School Psychology International*, 23, 342-354.
- Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). New York: Freeman.
- Mestre, J. P., Dufresne, R. J., Gerace, W. J., & Hardiman, P. T. (1993). Promoting skilled problem-solving behavior among beginning physics students. *Journal of Research in Science Teaching*, 30, 303-317.
- Newell, A., & Simon, H. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice Hall.
- Ohlsson, S. (2011). *Deep learning: How the mind overrides experience*. Cambridge, UK: Cambridge University Press.
- Raingruber, B., & Robinson, C. (2007). The effectiveness of tai chi, yoga, meditation, and reiki healing sessions in promoting health and enhancing problem solving abilities of registered nurses. *Issues in Mental Health Nursing*, 28, 1141-1155.
- Ramsburg, J. T., & Youmans, R. J. (2011). Cognitive training promotes academic success: An analysis of focused meditative practices on student quiz performance. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*.
- Smith, G. E., Housen, P., Yaffe, K., Ruff, R., Kennison, R. F., et al. (2009). A cognitive training program based on principles of brain plasticity: Results from the improvement in memory with plasticity-based adaptive cognitive training (IMPACT) study. *Journal of American Geriatric Society*, 57, 594-603.
- So, K. T., & Orme-Johnson, D.W. (2001). Three randomized experiments on the longitudinal effects of the Transcendental Meditation technique on cognition. *Intelligence*, 29, 419-441.
- Tang, Y., Ma, Y., Wang, J., Fan, Y., Feng, S., et al. (2007). Short-term meditation training improves attention and self-regulation. *Proceedings of the National Academy of Sciences*, 104, 17152-17156.
- Travis, F. (1979). Creative thinking and the Transcendental Meditation technique. *The Journal of Creative Behavior*, 13, 169-180.
- Pagnoni, G., Cekic, M. (2007). Age effects on gray matter volume and attentional performance in Zen meditation. *Neurobiol. Aging* 28,1623-1627.
- Willis, S. L., Tennstedt, S. L., Marsiske, M., Ball, K., Elias, J., et al. (2006). Long-term effects of cognitive training on everyday functional outcomes in older adults. *American Medical Association*, 296, 2805-2814.
- Zeidan, F., Johnson, S. K., Diamond, B. J., David, Z., & Goolkasian, P. (2010). Mindfulness meditation improves cognition. Evidence of brief mental training. *Consciousness & Cognition*, 19, 597-605.

# Changing Global Warming Beliefs with Scientific Information: Knowledge, Attitudes, and RTMD (Reinforced Theistic Manifest Destiny Theory)

Michael Andrew Ranney (ranney@berkeley.edu)<sup>1,2</sup>

Dav Clark (davclark@berkeley.edu)<sup>2</sup>

Daniel Lee Reinholz (reinholz@berkeley.edu)<sup>1</sup>

Sarah Cohen (sarahc51990@berkeley.edu)<sup>1</sup>

<sup>1</sup>Graduate School of Education, University of California, Berkeley, CA 94720

<sup>2</sup>Department of Psychology, University of California, Berkeley, CA 94720

## Abstract

Unlike peer nations' residents, Americans are less accepting of, and concerned by, (especially anthropogenic) climate change. Reinforced Theistic Manifest Destiny theory (RTMD; e.g., Ranney, 2012) explains many such "U.S.-exceptionalist" phenomena by combining geopolitical history with six belief constructs: afterlife, deity, nationalism, creation, evolution, and global warming. We assess predictions that climate change acceptance is increased by mechanism-explaining interventions. A 270-participant survey established widespread mechanistic ignorance, and an experiment with 149 other Americans (Californians and Texans) showed that a 400-word description of climate change's mechanism dramatically reduced ignorance and increased climate change acceptance. The mechanism, briefly, is: (a) Earth's surface absorbs (mostly visible) sunlight and subsequently emits *infrared* light, which (b) greenhouse gases selectively absorb and retain (because these molecules can become asymmetrical), so (c) heat energy leaves more slowly, warming Earth. Our intervention yielded desirable conceptual changes and science-coherent attitude changes. RTMD-predicted between-construct relationships were obtained and/or again replicated.

**Keywords:** Climate Change; Global Warming; Conceptual Change; Society; Science Education; Belief Revision.

In contrast to our exploding human population, many species are dwindling—often due to people's actions (e.g., hunting, degrading environments, and introducing non-native species). Still, many past effects pale compared to the threat of global climate change. (Nb. We will henceforth largely use the colloquial term, "global warming," although of course not *all* locations may exhibit warming due to human-enhanced greenhouse conditions.) In geologic time—on the order of 10,000 years or more—warming periods have consistently resulted in high levels of extinction (Mayhew, Jenkins, & Benton, 2008). Now, though, comparable warming is occurring over hundreds of years or less—posing a unique threat to many species' futures (cf. Harte & Harte, 2008), and direct threats to humans (particularly the poor)—such as increased risks of floods, droughts, and low crop yields (Kerr, 2007). Nothing, then, seems to exceed the importance of researchers finding ways to help people accept that anthropogenic global warming is (1) occurring, and (2) crucial to quickly address (Harte & Harte, 2008).

This urgent state is due to dramatic, human-caused atmospheric greenhouse gas increases from pre-industrial levels (about 260 years ago); for instance, methane is up by 150% and carbon dioxide is up by 40%. These levels are

*accelerating*, and may easily cause rapid mass extinctions (Malcolm et al., 2006), as with prior fast warmings. Fortunately, if humans act quickly, we may be able to conserve much of the current biosphere (Harte & Harte, 2008).

Sadly, U.S. attitudes clash with the 97% of actively publishing climate scientists who accept global warming's tenets (Anderegg, Prall, Harold, & Schneider, 2010). Leiserowitz, Maibach, and Roser-Renouf (2010) report that only 57% of the U.S. accepts global warming as occurring, and only 47% accepts it being "caused mostly by human activities." The U.S. accepts *both* less than do similarly developed "peer" nations. Indeed, among 33 peers *and* non-peers, only Indonesia, South Africa, and Nigeria rated global warming as less "serious" than the U.S. (Leiserowitz, 2007). Given global warming's potentially disastrous, irreversible effects, increasing Americans' global warming acceptance seems a worthy goal (Ranney, 2012). In the next subsection, we highlight a theory (RTMD) designed to explain U.S. exceptionalism regarding scientific, religious, and nationalistic affinities—especially the marked divergence in climate beliefs noted above. We then describe two empirical studies that test the RTMD-inspired notion that science instruction may powerfully rectify false beliefs about climate change.

## Theory: Reinforced Theistic Manifest Destiny

As Ranney (2012; Ranney & Thanukos, 2011) and other researchers have discussed, Americans are clearly outliers compared to peer nations' residents. Beyond global warming acceptance, Ranney describes other dimensions of American exceptionalism (e.g., regarding guns, murders, prisoners, military costs, executive salaries, income variability, teen pregnancies, infant deaths, health inefficiency, evolution acceptance, biblical literalism, piety, and beliefs in God and an afterlife). He also proposed the Reinforced Theistic Manifest Destiny (RTMD) theory to explain how a nation's collective theistic (and related) beliefs are reinforced—militarily, economically, etc. (Ranney, 2012; Ranney & Thanukos, 2011). RTMD focuses on beliefs and attitudes regarding the six inter-related constructs shown in Figure 1. The theory predicts that (1) acceptances of *creation*, *nationalism*, a *deity/dieties*, and an *afterlife* positively correlate, (2) acceptances of *evolution* and *global warming* positively correlate, and (3) constructs in (1) negatively correlate with those in (2)—partly because creation and evolution incohere. Among other offered explanations, RTMD explains the U.S.'s low acceptance of both evolution (Mil-

ler, Scott, & Okamoto, 2006) and global warming, compared to peer nations. At its heart, RTMD posits that unparalleled U.S. military and economic success—especially in WWI and WWII—has bolstered U.S. nationalism and theism, inhibiting American acceptance of evolution and global warming (Ranney, 2012). (In brief, Americans feel most reinforced for thinking “God is on our side.”)

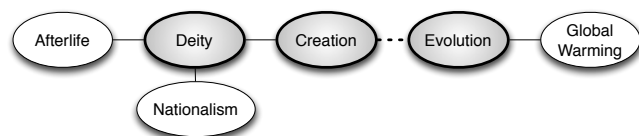


Figure 1. RTMD theory extends an often implicit “received view” (the three bold ovals; Ranney, 2012) of modest U.S. evolution acceptance with three extra constructs (non-bold ovals), such as global warming. Solid/dashed lines respectively represent coherent/conflicting conceptual links.

To date, RTMD has been assessed with seven data sets from the U.S. and one from Canada. RTMD predicts the directions of the 15 possible correlations among Figure 1’s six constructs; only the five main theoretical links are shown. Empirically, the 15 correlations virtually always exhibit the directions RTMD predicts (Ranney, 2012). RTMD theory also predicts that a change in one’s global warming acceptance may affect the other five variables—especially evolution acceptance. Analyses of such correlation-change predictions, though, exceed this piece’s space and scope.

## Mechanistic Knowledge and Climate Attitudes

RTMD notes the negative correlations between science-based and more faith-based constructs, and suggests that increased scientific knowledge may enhance positive attitudes towards evolution and climate change. Acceptance beliefs largely follow understanding, given variation in both. Thus, while evolutionary acceptance might be uniformly high or low in a certain classroom or campus, evolutionary biologists largely accept evolution more than do “Intro to Bio” students (Shtulman & Calabi, 2012). Therefore, low global warming acceptance may well be related to a lack of understanding. Consider humanity’s adoption of a heliocentric model of our solar system over geocentrism. Although geocentrism had (or has) its appeal, an understanding of gravity and orbits renders geocentric arguments hard to accept—or even “silly.” Similarly, we might expect that increasing individuals’ knowledge of the mechanism of global warming may help them accept the “less desirable” model of the world in which global warming is occurring (Ranney & Thanukos, 2011; cf. just world theory, e.g., Feinberg & Willer, 2011). Our approach to environmental conceptual and belief changes thus differs from most other efforts, which don’t focus as heavily on understanding the mechanism of the greenhouse effect (e.g., Al Gore’s Climate Project). Ours appears to be the first work to examine the extent to which one’s mechanistic global warming understanding affects relevant attitudes and support for climate policies.

Our research group is carrying out experiments and surveys that assess the hypotheses that (1) a proper understanding of global warming is rare, but (2) enhancing such understanding has desirable effects, such as increased global warming acceptance. There are many global warming education efforts, yet it is difficult to find an explanation of the basic physical/chemical mechanisms involved that is appropriately complete (see below) and yet not filled with too much extra detail (Ranney, Clark, Reinholz, & Cohen, 2012). We are not sure why this mechanistic pedagogical lack exists, but suggest that many would-be climate educators (1) do not adequately understand the mechanism themselves, and/or (2) fear that Americans are incapable of learning the basic scientific principles behind the greenhouse effect; others may (3) doubt that a scientific understanding would make much difference in our attitudes and policies towards global warming, and still others may (4) despair about the political, financial, or even agnotological (e.g., deceptive) elements of the quandary, suggesting that “the masses just can’t learn this.” But the aforementioned survey data suggest that peer nations’ residents accept (and fear—e.g., anthropogenic) global warming more than Americans do, and humanity has accepted other difficult ideas that were heavily suppressed by political or economic powers—such as heliocentrism and the links between tobacco smoke and severe illnesses. Likewise, we predict that a mechanistic warming explanation *may* help many people appreciate the soundness of climate change’s science—driving greater acceptance, concern, and imperative action.

Unless you read (and recall) this piece’s abstract, it is not likely that you can explain (even quite basically) the physical mechanism by which global warming occurs—as was true of most of our research team prior to our studies. Indeed, part of the present piece’s motivation stems from dozens of interviews the first author carried out with colleagues and acquaintances. Responses were often embarrassing and rarely accurate. So, imagine that you were chatting with a physicist/chemist, and she eyeballs you and asks, “How would *you* explain global warming’s mechanism?” Please take 30 seconds to answer this query before reading on.

Now that you’ve pondered global warming’s mechanism, please visit the abstract’s points (a) through (c). Did your explanation include these fundamental mechanistic aspects? In contrast to our abstract’s directly *mechanistic* explication (which is much abbreviated), many people articulate global warming’s *temporal precursors*, such as over-industrialization, rather than the fundamental mechanism that greenhouse gases are transparent to the sun’s incoming visible (i.e., “colors of the rainbow”) light, yet largely opaque to the infrared light that the Earth radiates. Other people often articulate global warming’s *effects*—such as increasing mean global temperatures, sea level increases, extinctions, or melting icecaps—but draw a blank mechanism-wise. Yet others focus on atmospheric features that don’t, or negligibly, explain global warming (e.g., ozone layer depletion). Many people who are familiar with both global warming’s potential precursors (e.g., more CO<sub>2</sub> emis-



sions) and effects cannot describe the *causal mechanism* between them. (Nb. Only one of the many dozens of interviewees knew that greenhouse gas molecules exhibit at least a transient electric dipole moment—e.g., via vibrations, bending, etc., to become asymmetrical—as CO<sub>2</sub> and CH<sub>4</sub> can, but O<sub>2</sub> cannot.)

Given the relations that RTMD theory coordinates, we infer that many Americans who *accept* global warming do so with sub-scientific rationales, and many who *reject* it do so by repressing the science due to religiously or nationalistically motivated reasoning. Thus, if Americans could grasp global warming's science, that might offer a better basis for shifting or strengthening their attitudes. Given that our dozens of informal interviews suggest the widespread lack of a mechanistic global warming understanding among U.S. residents, we predicted that this lack undermines U.S. global warming acceptance. Unfortunately, surveys about global warming knowledge (e.g., Leiserowitz, Maibach, & Roser-Renouf, 2010) rarely ask for mechanistic knowledge and often focus on recognition (thus likely overpredicting what is known by non-climatologists). Therefore, Study 1 below assessed our hypothesis that a broader, more representative sample than those in the informal interviews might also show a modest understanding of global warming's mechanism. Then, in Study 2, we assessed whether a brief explication of the mechanism might be successful in markedly enhancing both global warming knowledge and acceptance regarding anthropogenic climate change.

### Study 1: Gauging Global Warming Knowledge

Prior studies have documented numerous difficulties in understanding global warming (e.g., Shepardson, Niyogi, Choi, & Charusombat, 2011; Bord, Fisher, & O'Connor, 1998). Herein, though, we focus on less-studied difficulties in *mechanistic* understanding, three of which we believed were most critical to understanding the greenhouse effect (as highlighted in the abstract): (a) differentiating types of light/radiation, (b) understanding how the greenhouse effect depends on infrared light's selective absorption by greenhouse gases, and (c) understanding how that consequently warms the troposphere, water, and ground. To test the hypothesis that an understanding of global warming is indeed related to acceptance, we needed to determine whether the above conceptual difficulties were prevalent in a general adult population. Thus, we designed and administered a survey of global warming attitudes *and* understandings.

### Method: Participants, Design, Procedure, Materials

We collected 270 anonymous surveys from park visitors ( $n = 201$ ) and community college students ( $n = 69$ ) in San Diego. Random intercept sampling techniques were used for the park visitors; the community college students agreed to fill out the survey during a scheduled class break. A \$5 gift card compensated each participant. We refrain from offering contrived hypothesis tests below, simply reporting percentages—as is common with survey data. The full survey took 10-15 minutes to complete. We report here on a subset of

items: (1) 20 policy preference Likert items (e.g., “How much effort do you want the federal government to put into X?”), (2) two global warming belief items, (3) six short-answer global warming knowledge questions, (4) 13 items on possible causes of global warming (e.g., participants were to label Y as a major/minor/non- cause of global warming), and (5) four items gauging respondents' willingness to make personal sacrifices for specific climate policies. Short answers were coded and scored with a rubric that showed high inter-rater reliability (mean Cohen's  $\kappa = 0.74$ ).

### Results and Discussion

The data support our hypothesis that Americans rarely understand global warming's mechanism. When asked to explain “the basic physical, chemical, or biological mechanism of global warming,” only 32 participants (12%) referenced gases in the atmosphere (e.g., emissions, CO<sub>2</sub>, or pollution) trapping heat—which is merely a partial understanding. Of these 32, only four (1%) attempted to differentiate types of energy (or light). Not a single participant (0%) mentioned either correct absorption(s) or the difference (input/output asymmetry) between visible and infrared light, which is the crux of understanding the greenhouse effect. Notably, only eight participants (3%) even named the greenhouse effect. Many responses included possible *causes*, yet few included possible mechanisms—and the item's median score was 0.

Problematic conceptions were also prevalent. For instance, in answering our question about global warming's mechanism, 42 participants (16%) claimed that the destruction of the atmosphere or ozone layer was letting in more heat, thus causing global warming. This finding echoes the results of previous work (e.g., Bord, Fisher, & O'Connor, 1998). Indeed, on a subsequent “possible cause” item, 201 participants (74%) incorrectly believed that ozone depletion was a major cause of global warming, but only 81 (30%) knew that livestock are a major cause of global warming.

Despite this poor showing regarding global warming knowledge, many people were willing to accept global warming and its anthropogenic origins. In particular, when the responses “mildly agree” and “strongly agree” are combined, 217 people (80%) agreed with the statement, “I am certain that global warming (i.e., climate change) is actually occurring,” and 208 participants (77%) agreed that “human activities are a significant cause of global warming.” Although this willingness to accept global warming is higher than the averages found on most national surveys, there appears not to have been a prohibitive ceiling effect.

Crucially, experimenter-scored knowledge of the mechanism significantly correlated with peoples' willingness to accept global warming as both real ( $r = .22, p = .0002$ ) and anthropogenic ( $r = .17, p = .005$ ). Also importantly, anthropogenic climate change acceptance significantly predicted (via ordinal models) *all four* survey items about willingness to sacrifice ( $\chi^2(4) > 32, p < 0.001$ )—and one's knowledge score significantly predicted two of these ( $\chi^2(1) > 3.8, p < 0.05$ ). In addition, all 15 correlations among the six RTMD

constructs fell in the predicted directions—replicating previous findings—and 13 of the 15 were significantly different from zero at  $p < .01$ ; likewise, evolution/creation acceptance strongly predicted global warming knowledge and acceptance (as occurring *and* anthropogenic)—notably, even more strongly than political party.

In sum, these U.S. respondents clearly knew little about the mechanism of the greenhouse effect—the anthropogenic increase of which is the basis for global warming. This is true even of individuals who accept the reality of global warming, which ought give us pause. The mere acceptance of global warming, even absent knowledge of its basic science, appears to yield warranted climate policy attitudes. We predict, though, that skeptically evaluated knowledge of a basic mechanistic account should enhance that precursory global warming acceptance. (Consider someone who might accept evolution without even having a rudimentary understanding of how organisms procreate!) Scientific literacy ought to mean that people seek out causal explanations, just as those who deny global warming—should they be believed—ought to explain the mechanism by which our planet would be unaffected by massive additions of greenhouse gas emissions. More directly, Study 1 shows that, in crucial cases, veridical knowledge has a clear relationship to one’s willingness to sacrifice. It seemed incumbent upon us, then, to begin developing interventions meant to improve Americans’ understandings of the basic physical-chemical global warming mechanism—as described in Study 2.

## Study 2: Learning and Increased Acceptance

Drawing on past research on both physics cognition and the Numerically Driven Inferencing paradigm (NDI; e.g., Garcia de Osuna, Ranney, & Nelson, 2004), we hypothesized that a small amount of targeted information could yield dramatic conceptual changes—ultimately including changes in attitude and acceptance. In the NDI paradigm, people are asked to estimate the value of a quantity, and they are later told its true value. By having individuals “put their cards on the table” before receiving the true value, we inhibit hindsight bias and post-hoc rationalization, and the impact of the information is thus increased (Rinne, Ranney, & Lurie, 2006). Here we report on a similarly compact and empirically grounded intervention with a 400-word text that highlights the three key conceptual pieces noted in Study 1’s introduction (labeled a-c). See Ranney et al. (2012) for the full text. Our most recent work combines NDI and RTMD, utilizing misleading “anti-climate change acceptance” numeric quantities—yielding notable shifts in attitudes and self-rated knowledge—but this is outside the scope of this paper.

### Method: Participants, Design, Procedure, Materials

For Study 2, 103 University of California, Berkeley, and 46 University of Texas, Brownsville, undergraduates were randomly assigned to one of two groups: “sandwich” or “no-pretest.” Sandwich group participants: (1) both provided an explanation of the greenhouse effect (effectively “putting their cards on the table”) and filled out knowledge and atti-

tude surveys, (2) read a 400-word explanation of the mechanism of the greenhouse effect and gave a rating of experienced surprise, and (3) were re-tested on their knowledge and attitudes (with a posttest identical to the pretest). No-pretest (or “open-faced”) group participants completed only (2) and (3) above. Thus, (1) and (3) can be thought of as “bread” and (2)—the explanation—is the “jam” of our design. The no-pretest group offers a between-subjects contrast via their posttest, obviating test/re-test concerns about experimenter demand regarding the sandwich group. Just before leaving the experiment, all participants also filled out a demographic questionnaire. Surveys were again anonymous, as in Study 1.

Below, we report data from the 85 Berkeley and 41 Brownsville students who completed the survey as intended and had been U.S. residents for ten years or more (because we expressly consider *U.S.* exceptionalism/nationalism). Of the Berkeley data, we analyzed 43 no-pretest (open-faced) surveys and the pretest part of 42 sandwich surveys—but due to anticipated time constraints, only 30 sandwich post-tests could be completed/obtained. Of the Brownsville data, we analyzed 22 no-pretest and 19 sandwich surveys. To be conservative, all between-group t-tests were Welch-method adjusted for unequal variance and sample size. All hypotheses below were clearly stated as *a priori* ones and were replicated across our two samples except where noted.

The attitude survey used 12 items (on 9-point Likert scales) to assess the six RTMD constructs. True knowledge of global warming was assessed based on (1) three written responses by participants and (2) (on the posttest only:) two fill-in-the-blank items about the types of light (visible, infrared, etc.) involved in the greenhouse effect. *Self*-reports of knowledge were also reported on a 9-point Likert scale.

## Results and Discussion

### The Crucial Global Warming Mechanism Was Learned

Even our rather sophisticated samples initially exhibited incorrect or non-normative understandings of the greenhouse effect’s mechanism (e.g., on the roles of ultraviolet light, the ozone layer’s depletion, non-greenhouse-gas pollution, and the reflection of incoming light). Most notably, not a *single* pre-test explanation mentioned different light/radiation types or atmospheric retention time, despite an explicit prompt to explain any differences between the energy traveling toward and away from Earth. However, after reading the 400-word description, 61% of the Berkeley participants across both groups correctly answered that “infrared” light was emitted from Earth (in its fill-in-the-blank space), as did 55% of the Brownsville students who responded.

Beyond the blank-filling items, we statistically analyzed individuals’ *qualitative* explanations—creating scoring rubrics for three central concepts: (a) differentiating between the types of light entering and exiting the atmosphere, (b) atmospheric greenhouse gases’ interactions with radiation, and (c) the increased atmospheric retention time of energy. Inter-rater reliability was again high (weighted  $\kappa = 0.71$  based on about one-third of the Berkeley data;  $\kappa = 0.67$

across the full Brownsville dataset). Table 1 shows the percentages of all possible points: overall, we found dramatic knowledge increases (doublings, triplings, or more), which were significant for all subscales—both within-subjects for the sandwich condition, and between-subjects from the sandwich pretest to the no-pretest condition's posttest, ( $p < .05$  for all six improvement possibilities).

Table 1. The mean percentage scores for each of the three assessed global-warming constructs (with greenhouse gases = GHGs), for each test and sample (for California;Texas). All improved from pretest (\*:  $p < .05$ ; \*\*:  $p < .005$ ).

Group & Test (& means)	Light	GHGs	Energy
Sandwich Pretest (means <sub>CA;TX</sub> = 33%;11%)	33%; 7%	39%;16%	28%;11%
Sandwich Posttest (means <sub>CA;TX</sub> = 69%;37%)	78%;36% **,*	83%;39% **,**	47%;35% *,*
No-pretest Posttest (means <sub>CA;TX</sub> = 66%;49%)	66%;43% **,**	74%;54% **,**	57%;51% **,**

### Global Warming Acceptance Via Mechanistic Learning

It may seem quite remarkable, but participants' global warming acceptance increased dramatically after our brief intervention, as predicted. To assess this, we used all of the 73 Berkeley posttest ratings in a paired t-test, and used imputation for pretest scores for the no-pretest group. (In particular, the full set of 42 pretest ratings was used to avoid sampling bias.) We found a significant change in global warming acceptance on the posttests, as compared to pretest measures ( $t(72) = 2.28, p = .01$ ). This result was replicated with the Brownsville surveys ( $t(39) = 4.24, p < .0001$ ). In addition, although Study 2's statistical power was rather limited, the correlation matrices for the RTMD variables again largely supported RTMD theory—as was certainly found in Study 1 and all prior studies. The relationship between knowledge and attitudes was also reflected in Berkeley students' naïve pre-test data, in which participants' *self-perceived* ratings of their own global warming knowledge correlated significantly with their global warming attitudes ( $r = .39, p = .01$ ). This was not the case with Brownsville students ( $r = .15, p = .55$ ), which may be reflective of their overall lower self-perceived knowledge.

Please recall that we had also predicted a between-conditions difference in surprise ratings due to reduced hindsight biases among the sandwich participants. The difference for Berkeley students was at the significance borderline ( $t(42.08) = 1.65, p = .05$ ); the surprise ratings only reached "6" in the no-pretest condition (out of 9, with "5" being "somewhat surprising"), but were as high as "9" (i.e., "extremely surprising") in the sandwich condition. Among Brownsville students, surprise was uniformly higher, with a numerically similar difference between conditions, although this result was not significant ( $t(38.1) = 0.92, p = .18$ ).

### Conclusions from Study 2

This experiment replicates and extends the findings noted earlier (from prior interviews and Study 1), such that even

rather well-educated people initially held mostly non-normative understandings of global warming's mechanism. Only 400 words later, though (roughly the duration of a TV commercial break), dramatic increases were observed in (1) mechanistic knowledge and (2) global warming acceptance. (Further, the increases were found in divergent U.S. states and colleges.) Differences in surprise ratings between the sandwich and "no-pretest" ("open-faced") groups further support the notion that eliciting an explanation or theory prior to offering information *increases* surprise and *reduces* post-hoc rationalization and hindsight bias. (On surprise, see Clark & Ranney, 2010; Munnich, Ranney, & Song, 2007.)

## General Discussion

Of Study 1's 270 participants, none could fully explain that (a) visible light makes its way to Earth's surface where it is (mostly) absorbed and emitted later as infrared light, (b) this infrared light is largely (actually, 90%) absorbed by greenhouse gases before reaching outer space, and (c) this slows energy loss and warms Earth. (Ranney calls the unimpeded 10% the "Goldilocks tithe.") While others (e.g., Leiserowitz, 2007) have shown aspects of U.S. ignorance, we extend these results to mechanistic understandings of the greenhouse effect. If our sample even vaguely represents the U.S. public, then they rarely understand global warming's mechanism. Further, such knowledge *does* relate to policy preferences, willingness to sacrifice due to legislation, and beliefs about (anthropogenic) climate change's reality. We suggest that prior works' inattention to this mechanism may be because it is scientifically uncontroversial relative to the effects, mitigation strategies, and other causes re: climate change. However, it seems that mechanistic knowledge may play a key role in successful climate policy ventures.

Just as knowing "how reproduction works" supports evolutionary acceptance (cf. Shtulman & Calbi, 2012), our studies show that a mechanistic global warming understanding (e.g., a-c above) supports *its* acceptance. Study 2 showed that we increased students' acceptance by increasing global warming knowledge. Space prohibits a full treatment of this, but a new study further shows that, after providing people with misleading, cherry-picked facts, we caused them to discount climate change (dropping from 6.5 to 5.9 on a 9-point scale) with a dramatic, concurrent drop in their confidence in their knowledge (plummeting from 5.0 to 2.9, on a 9-point scale). Thus, as the nefarious are well aware, empirical data *do not always* increase acceptance, global coherence, and self-confidence in one's understanding.

In short, work spawned by the Reinforced Theistic Manifest Destiny theory (Ranney, 2012; Ranney & Thanukos, 2011) regarding concerns about U.S. exceptionalism led us to find a successful way to enhance wisdom about the greenhouse effect's mechanism. That is, we found that instruction focused on a mere 400 words of text dramatically increases undergraduates' global warming understandings *and* increases their mean acceptance of anthropogenic global warming. We suspect that our instruction is effective in that it addresses head-on the implicit mystery of how ener-

gy—as visible light—can easily get close to Earth’s surface and troposphere, yet has difficulty leaving that surface/troposphere (as absorbed, intercepted infrared light). Future research will determine our intervention’s longevity, among other attempts to better comprehend the landscape of the cognitions and emotions regarding global warming. As the studies above demonstrate, insights from cognitive science show much promise for tackling the challenges for climate-relevant education in the U.S. and abroad.

## Acknowledgments

We greatly thank Megan Beale, Amanda Cain, Roxana Farjadi, Jackie Felipe, Benji Walklet, and Jeff Wilson (as well as M. Crain, J. Fong, D. Gillingham, L. Goldwasser, A. Lazaris, L. Nevo, B. Rai, T. Ryan, J. Spector, and UC-B).

## References

- Anderegg, W. R. L., Prall, J. W., Harold, J., & Schneider, S. H. (2010). Expert credibility in climate change. *Proceedings of the National Academy of Sciences*, 107(27), 12107-12109.
- Bord, R. J., Fisher, A., & O'Connor, R. E. (1998). Public perceptions of global warming: United States and international perspectives. *Climate Research*, 11(1), 75-84.
- Clark, D., & Ranney, M. A. (2010). Known knowns and unknown knowns: multiple memory routes to improved numerical estimation. In K. Gomez, L. Lyons & J. Radinsky (Eds.), *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences* (Vol. 1, pp. 460-467): International Society of the Learning Sciences, Inc.
- Feinberg, M., & Willer, R. (2011). Apocalypse soon? Dire messages reduce belief in global warming by contradicting just world beliefs. *Psychological Science*, 22, 34-38.
- Garcia de Osuna, J., Ranney, M. A., & Nelson, J. (2004). Qualitative and quantitative effects of surprise: (Mis)estimates, rationales, and feedback-induced preference changes while considering abortion. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the Twenty-sixth Annual Conference of the Cognitive Science Society* (pp. 422-427). Mahwah, NJ: Erlbaum.
- Harte, J., & Harte, M. E. (2008). *Cool the Earth, save the economy: Solving the climate crisis is EASY*. Retrieved from <http://www.cooltheearth.us/>
- Kerr, R. A. (2007). Global warming is changing the world. *Science*, 316(5822), 188.
- Leiserowitz, A. (2007). *International public opinion, perception, and understanding of global climate change (Human Development Report 2007/2008)*: UNDP. [Online]. Available: [http://hdr.undp.org/en/reports/global/hdr2007-2008/papers/leiserowitz\\_anthony6.pdf](http://hdr.undp.org/en/reports/global/hdr2007-2008/papers/leiserowitz_anthony6.pdf) [2009, July 9].
- Leiserowitz, A., Maibach, E., & Roser-Renouf, C. (2010). *Climate change in the American mind: Americans global warming beliefs and attitudes in January 2010*. Yale University and George Mason University. New Haven: CT. Yale Project on Climate Change. Available at: <http://environment.yale.edu/uploads/AmericansGlobalWarmingBeliefs2010.pdf>.
- Malcolm, J. R., Liu, C., Neilson, R. P., Hansen, L., & Hannah, L. (2006). Global warming and extinctions of endemic species from biodiversity hotspots. *Conservation Biology*, 20, 538-548. doi:10.1111/j.1523-1739.2006.00364.x
- Mayhew, P. J., Jenkins, G. B., & Benton, T. G. (2008). A long-term association between global temperature and biodiversity, origination and extinction in the fossil record. *Proceedings of the Royal Society B: Biological Sciences*, 275, 47-53. DOI:10.1098/rspb.2007.1302
- Miller, J. D., Scott, E. C., & Okamoto, S. (2006). Public acceptance of evolution. *Science*, 313, 765-766.
- Munnich, E. L., Ranney, M. A., & Song, M. (2007). Surprise, surprise: The role of surprising numerical feedback in belief change. In D. S. McNamara & G. Trafton (Eds.), *Proceedings of the Twenty-ninth Annual Conference of the Cognitive Science Society* (pp. 503-508). Mahwah, NJ: Erlbaum.
- Ranney, M. A. (2012). Why don't Americans accept evolution as much as people in peer nations do? A theory (Reinforced Theistic Manifest Destiny) and some pertinent evidence. In K.S. Rosengren, S.K. Brem, E.M. Evans, & G.M. Sinatra (Eds.), *Evolution challenges: Integrating research and practice in teaching and learning* (pp. 233-269). Oxford: Oxford University Press.
- Ranney, M.A., Clark, D., Reinholz, D., & Cohen, S. (in press for 2012). Improving Americans’ modest global warming knowledge in the light of RTMD (Reinforced Theistic Manifest Destiny) theory. In *The Future of Learning: Proceedings of the 10th International Conference of the Learning Sciences*. International Society of the Learning Sciences, Inc.
- Ranney, M. A., & Thanukos, A. (2011). Accepting evolution or creation in people, critters, plants, and classrooms: The maelstrom of American cognition about biological change. In R. S. Taylor & M. Ferrari (Eds.), *Epistemology and science education: Understanding the evolution vs. intelligent design controversy* (pp. 143-172). New York: Routledge.
- Rinne, L., Ranney, M. A., & Lurie, N. (2006). Estimation as a catalyst for numeracy: Micro-interventions that increase the use of numerical information in decision-making. In S. A. Barab, K. E. Hay & D. T. Hickey (Eds.), *Proceedings of the 7th International Conference on Learning Sciences* (pp. 571-577). Mahwah, NJ: Erlbaum.
- Shepardson, D. P., Niyogi, D., Choi, S., & Charusombat, U. (2011). Students’ conceptions about the greenhouse effect, global warming, and climate change. *Climatic Change*, 104, 481-507.
- Shtulman, A., & Calabi, P. (2012). Cognitive constraints on the understanding and acceptance of evolution In K.S. Rosengren, S.K. Brem, E.M. Evans, & G.M. Sinatra (Eds.), *Evolution challenges: Integrating research and practice in teaching and learning* (pp. 47-65). Oxford: Oxford University Press.

# Evidence that Threatening Situations Enhance Creativity

**Sean N. Riley (research@seanriley.ca)**

Department of Psychology, University of British Columbia  
3333 University Way, Kelowna B.C V1V 1V7

**Liane Gabora (liane.gabora@ubc.ca)**

Department of Psychology, University of British Columbia  
3333 University Way, Kelowna B.C V1V 1V7

## Abstract

We tested the hypothesis that threatening situations enhance creativity. 60 participants viewed a series of photographs and rated them on level of threat. They then wrote two short stories: one based on the photograph they rated as most threatening, and the other based on the photograph they rated as least threatening. The stories were rated for level of creativity. Paired samples t-tests revealed that stories based on threatening pictures produced a higher degree of creativity than those based on non-threatening pictures. Theoretical frameworks consistent with these findings are discussed.

**Keywords:** Cognitive tuning, creativity, existential anxiety, inspiration, mood, narrative, story telling, threat.

## Introduction

Creativity can be defined as the ability to generate ideas, interpretations, or solutions that are both novel, and meaningful or appropriate (Sternberg *et al.*, 2010). Creativity is widely associated with personal fulfillment (May, 1975; Rogers, 1959), self-actualization (Maslow, 1959), and with maintaining a competitive edge in the marketplace. Creative therapies are useful in clinical contexts, including the assessment and resolution of conflict (Goldblatt *et al.*, 2011), dementia (Hannemann, 2006), self-esteem (Anzules, Haennl & Golay, 2007), and stress reduction (Curl, 2008). In addition, creative individuals tend to have higher levels of life satisfaction (Tan, Ho, Ho, & Ow, 2008), emotional intelligence (Noferesti & Alghorabaie, 2011), and intelligence in general (Batey & Furnham, 2009). These findings suggest that creativity is, generally speaking, a positive attribute with numerous constructive byproducts.

However, there are significant drawbacks to creativity (Cropley, Cropley, Kaufman, & Runco, 2010; Ludwig, 1995). Generating creative ideas is time consuming, and a creative solution to one problem often generates other problems, or has unexpected negative side effects that may only become apparent after much effort has been invested. Creative people often reinvent the wheel, and may be more likely to bend rules, break laws, and provoke social unrest (Cropley, Kaufman, & Cropley, 2003; Sternberg & Lubart, 1995; Sulloway, 1996). They tend to be more emotionally unstable and prone to affective disorders such as depression and bipolar disorder, and have a higher incidence of schizophrenic tendencies than other segments of the population (Andreason, 1987; Flaherty, 2005; F. Goodwin

& Jamison, 1990). Computational models suggest that there is a detrimental impact on society if either the ratio of creative to relatively uncreative individuals is too high, or if the creative individuals are *too* creative (Gabora & Firouzi, 2012; Gabora & Leijnen, 2009; Leijnen & Gabora, 2009).

There is also preliminary evidence that situations that are demanding, threatening, or involve conflict, put one in a more creative state of mind. For example, it has been shown that individuals who are in the midst of conflict set broader and more inclusive cognitive categories (De Dru, Carsten & Nijstad, 2008). Creativity is positively correlated with aggression (Tacher & Readdick, 2006), group conflict (Troyer & Younggreen, 2009) anxiety (Carlsson, 2002), and dishonesty (Gino & Ariely, 2011). Finally it has been found that negative affect leads to greater creative output (Akinola, Mendes, 2008).

This study further investigates the hypothesis that threatening situations put one in a more creative state of mind. Specifically, it assesses whether stories written in response to threatening stimuli are more creative than stories written in response to non-threatening stimuli.

## The Study

### Participants

Participants ( $n = 60$ ; 19 M, 41 F) were recruited for this study through the University of British Columbia (Okanagan campus) course credit for research participation program used in introductory psychology classes. Participants received 1% added to their overall course grade.

### Stimuli

Participants were shown 15 photographs depicting situations that had been independently classified as either threatening or nonthreatening by both experimenters. Classification was determined by assessing whether a normal individual would feel significantly at risk of harm or death if they were in the scenario depicted by the photograph. The photographs classified as 'threatening' were deemed to be threatening not directly, but indirectly, in the same sense that one feels threatened by a war movie or horror film.

Examples of the threatening and non-threatening photographs used to generate stories in this study are given in Figures 1 and 2.





Figure 1. Example of non-threatening photograph.



Figure 2. Two examples of threatening photographs.

## Procedure

The participants were asked to rate each photograph according to how threatening they believed it to be on a 7-point Likert scale. Each participant was then given 30 minutes to write two short stories: one about the photograph that was rated by that participant as most threatening, and one about the photograph that was rated by that participant as least threatening. If two photographs were rated as

equally threatening, the photograph that was presented first in the sequence was used. To guard against order effects, both story order and picture order were counterbalanced.

## Raters

Four raters were University of British Columbia undergraduates who were enrolled in an advanced psychology research methods and statistics course. They are referred to as *student raters*. They consisted of three females and one male.

The fifth rater was a well-known and extensively published fiction author who was not reimbursed for his participation. He is referred to as the *expert rater*. He was male.

## Rating

Raters were trained to evaluate participant responses using a previously developed rubric for assessing story creativity, the 'Creativity' portion of the Wisdom Intelligence Creativity Synthesized (WICS) rubric (Sternberg, 2005; Sternberg *et al.*, 2010, 2012), and a set of stories that had previously established creativity ratings. The WICS rubric assesses story creativity on the basis of evidence of the respondent's ability to provide novel yet meaningful ideas, narratives or interpretations of situations, or solutions to problems, or to view situations and events in a new and meaningful way.

Raters practiced on the sample stories until their ratings on these sample stories correlated significantly with these stories' previously established creativity ratings.

## Results

The creativity ratings provided by the expert rater were significantly positively correlated with the creativity ratings provided by the student raters, as shown in Table 1.

Table 1: Bivariate correlation between expert rater and each of the four student raters.

	Rater One	Rater Two	Rater Three	Rater Four
Non-threatening	.289*	.454**	.286*	.313*
Threatening	.376**	.156	.368**	.477**

\*Significant at < .05; \*\*Significant at < .01.

Intraclass correlation was used to assess the level of agreement amongst all five raters. They were significantly in agreement, as shown in Table 2.

Table 2. Inter-class correlation between all raters.

	ICC	Value	Sig.
Single Measures	.430	8.542	.001
Average Measures	.883	8.542	.001

We created two rater groups: the average of all the raters' scores, as well as the average of the two most central raters,

for each of the stories. The central raters were selected by dropping the raters with the highest and lowest means, and then selecting from the remaining three raters the two whose mean ratings were closest to one another. We conducted a paired samples t-test between stories based on threatening pictures, and stories based on non-threatening pictures, for both rater groups. The mean creativity scores for stories generated in response to non-threatening and threatening stimuli are shown in Figure 4.

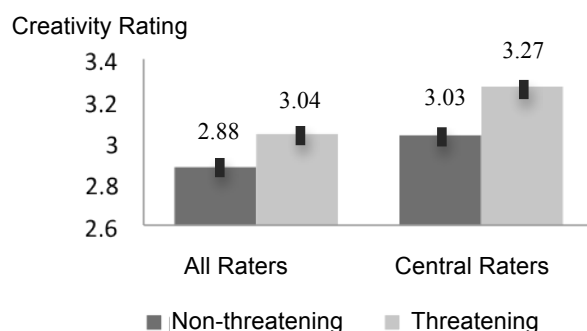


Figure 4. Mean creativity scores for stories generated in response to non-threatening and threatening stimuli.

The results indicate a high level of agreement between raters ( $p = .001$ ) and support for our hypothesis. As shown in Table 3, using both the average of all raters, as well as the two central raters, paired sample t-tests revealed that stories based on threatening pictures were rated as significantly more creative than those based on non-threatening pictures ( $p < .05$ ).

Table 3. Paired samples t-tests.

	t-value	SE	Cohen's <i>d</i>	Sig.
All Raters	2.08	.077	.269	.041
Central Raters	2.29	.101	.296	.026

## Discussion

These results support our hypothesis that stories written in response to threatening photographs elicit a higher degree of creativity than those written in response to non-threatening photographs. Several theoretical frameworks are consistent with and could help explain these findings.

One kind of explanatory framework comes from creativity research. It is widely assumed that the creative process involves *searching* through memory and/or *selecting* amongst a set of predefined candidate ideas. For example, computer scientists have modeled the creative process as a heuristic search (e.g., Simon, 1973, 1986). In psychology, there is much evidence for, and discussion of, the role of *divergent thinking* in creativity (Guilford, 1968; for a review see Runco, 2010). Divergent thinking is presumed to involve the generation of multiple, often unconventional, possibilities. When construed from this perspective, the creative process often goes hand-in-hand with the notion of selection, since if you come up with

multiple alternatives you eventually weed some of them out. Indeed, many well-known theories of creativity, such as the Geneplore model (Finke, Ward, & Smith, 1992), and the Darwinian theory of creativity (Simonton, 1999) involve two stages: the generation of possibilities, followed by the exploration and ultimately selective retention of the most promising of them.

However, there are strong neurobiological (Gabora, 2010a), experimental (Gabora, 2010b; Gabora, O'Connor & Ranjan, in press), and theoretical (Gabora, 2005, 2007), reasons to believe that the generation stage of creative thinking may be *divergent* not in that it moves in multiple directions or generates multiple possibilities, but in the sense that it produces a raw idea that is vague or unfocused, and that requires further processing to become viable. Similarly, the exploration stage of creative thinking may be *convergent*, not in ordinary sense that it entails selecting from amongst alternatives, but in the sense that it entails considering a vague idea from different perspectives until it comes into newly defined focus. In other words, the terms 'divergent' and 'convergent' may be applicable to creative thought in the sense of going from well-defined to ill-defined, and *vice versa*. Although a particular creative thinking process *may* involve search or selection amongst multiple well-defined possibilities, it *need* not, and moreover, that *selection* need not figure prominently in a general theory of creativity.

According to this alternative view of creativity, creative individuals wrestle with those issues or ideas that are, for them, in a state of *potentiality*. An artist might wrestle with how to capture the feeling of a particular landscape, and a writer might wrestle with depicting how events in an imaginary world would unfold. Over time, creative ideas come to assume a form that is more fully *actualized*, or well-defined, as they are considered from different perspectives in accordance with the constraints of the domain in which they are expressed. By giving form to that which exists in a state of potentiality, the individual gains a richer understanding and appreciation of it, as well as a sense of control or mastery over it.

Central to this view of creativity is the notion of a *worldview*. The term worldview is used to refer to one's internal model of the world, as well as one's values, predispositions, and habitual patterns of response (Gabora, 2000, 2008; Gabora & Aerts, 2009). Each idea the creator comes up with is construed as a different expression of the same underlying core network of understandings, beliefs, and attitudes. Thus an individual's outputs are inter-related, and potentially pave the way for one another. By adopting the notion of a worldview, our account places equal emphasis on external creative outcomes and the internal cognitive and emotional restructuring brought about by the creative process. The cognitive reorganization and personality dynamics (e.g., involving wellbeing, self-discipline, or self-discovery) are viewed as the internal, less readily measurable, but equally important, counterpart to external manifestations of the creative process. The



transformation that occurs on canvas or on the written page is mirrored by a sense of personal, cognitive transformation and self-discovery from within.

In the case of our study, this alternative theory suggests that the threatening stimuli created a dichotomy in terms of one's understanding of the world. On the one hand, we believe that the world is a just and fair world, and that we, as individuals, are deserving of just and fair treatment. However, the threatening stimulus confronts this conception by imposing upon us a negative reality that threatens our internal model of how the world operates. In response to this, we tap into our creative potential and hone in on a suitable explanation for the threat's existence. This is done in an attempt to reconcile the worldview dichotomy, and impose a sense of meaning and understanding as to why this negative reality exists, ultimately forging a new and cohesive worldview structure.

These findings are consistent with the previously mentioned literature on the dark side of creativity. The positive correlation between negative affect and creativity (Akinola & Mendes, 2008) gives credence to the notion that creativity can arise from worldview disequilibrium, that is, from a schism between one's reality and one's internal conception of reality. In the case of negative affect, one's reality is that one is experiencing unpleasant and unsettling emotions; however, the internal conception of reality is predicated on the notion that one should be happy. In response to this dichotomy, the creative process helps to create a new, cohesive worldview structure that reduces negative affect (e.g., Curl, 2008).

These findings can also be interpreted in terms of the cognitive-tuning hypothesis (Schwarz, 2002). This theory posits that our mood mirrors our surroundings, and mood-states in turn, affect how information from the environment gets processed. In the case of a negative mood-state, one narrows attention and focuses on the negative environment. One then becomes more motivated to appropriately cope with the negative mood, and engage in bottom-up processing of the environment to understand the factors underpinning the negative environment. Conversely, a positive-mood state does not require an in-depth assessment of the underpinning factors; thus, one engages in top-down processing, and uses heuristics to understand the environment. In terms of creative output, the bottom-up processing used in negative mood-states allows for a more in-depth analysis of the environment, which can lead to more novel interpretations of its underpinning factors. The positive-mood state does not necessitate this in-depth analysis; thus a common, less novel narrative is developed. In the study reported here, the threatening photographs may have induced a negative mood-state, and the non-threatening photographs may have induced a neutral or positive mood-state, with the ensuing impact on information processing and creative output.

To best of our knowledge, this paper is the first study to directly assess the role of threatening stimuli on creative output. However, the study has some weaknesses and

limitations. First, the photographs used were not drawn from the International Affective Picture System. Although the photographs used in the study did, on a qualitative level, hold a high degree of consistency in terms of arousal and valence, future studies should consider using standardized photographs. Furthermore, due to time constraints, participants were limited to 15 minutes of writing time per story. It is possible that, given adequate time, we could see a convergence of creative output. Despite these limitations, we believe the study yielded some highly intriguing preliminary findings concerning the role of threat on creative output, which pave the way for further research into this phenomenon.

## Acknowledgments

We would like to acknowledge grants to the second author from the National Science and Engineering Research Council of Canada, and the Fund for Scientific Research of Flanders, Belgium.

## References

- Akinola, M., & Mendes, W. B. (2008). The dark side of creativity: Biological vulnerability and negative emotions lead to greater artistic creativity. *Personality and Social Psychology Bulletin*, 34(12), 1677–1686.
- Andreason, N. C. (1987). Creativity and mental illness; prevalence rates in writers and their first degree relatives. *American Journal of Psychiatry*, 144, 1288–1292.
- Anzules, C., Haennl, C., & Golay, A. (2007). An experience of art therapy for patients suffering from obesity. *European Diabetes Nursing*, 4(2), 72–76.
- Batey, M., & Furnham, A. (2009). The relationship between creativity, schizotypy and intelligence. *Individual Differences Research*, 7(4), 272–284.
- Cropley, D. H., Cropley, A. J., Kaufman, J. C., & Runco, M. (2010). *Creativity in schools: Tensions and dilemmas*. Cambridge: Cambridge University Press.
- Cropley, D. H., Kaufman, J. C., & Cropley, A. J. (2003). Malevolent creativity: A functional model of creativity in terrorism and crime. *Creativity Research Journal*, 20, 105–115.
- Curl, K. (2008). Assessing stress reduction as a function of artistic creation and cognitive focus. *Art Therapy*, 25(4), 164–169.
- Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications*. Cambridge, MA: MIT Press.
- Flaherty, A. W. (2005). Frontotemporal and dopaminergic control of idea generation and creative drive. *Journal of Comparative Neurology*, 493, 147–153.
- Furnham, A., Batey, M., Booth, T. W., Pater, V., & Lozinskaya, D. (2011). Individual difference predictors of creativity in Art and Science students. *Thinking skills and creativity*, 6(2), 114–121.
- Gabora, L. (2000). Conceptual closure: Weaving memories into an interconnected worldview,” in G. Van de Vijver & J. Chandler (Eds.), *Closure: Emergent organizations and*

- their dynamics. New York: Annals of the New York Academy of Sciences.
- Gabora, L. (2002). Cognitive mechanisms underlying the creative process. In T. Hewett & T. Kavanagh (Eds.), *Proceedings of the fourth international conference on creativity and cognition* (pp. 126–133). Loughborough, UK: Loughborough University Press.
- Gabora, L. (2005). Creative thought as a non-Darwinian evolutionary process. *Journal of Creative Behavior*, 39(4), 65–87.
- Gabora, L. (2007). Cultural evolution entails (creativity entails (concept combination entails quantum structure)). *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium 8: Quantum Interaction*, March 26–28, Stanford University, pp. 106–113.
- Gabora, L. (2008). The cultural evolution of socially situated cognition. *Cognitive Systems Research*, 9(1–2), 104–113.
- Gabora, L. (2010a). Revenge of the ‘nerds’: Characterizing creative thought in terms of the structure and dynamics of memory. *Creativity Research Journal*, 22(1), 1–13.
- Gabora, L. (2010b). Recognizability of creative style within and across domains: Preliminary studies. *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 2350–2355). August 11–14, 2010, Portland, Oregon.
- Gabora, L., & Firouzi, H. (2012). Society functions best with an intermediate level of creativity. *Proceedings of the Annual Meeting of the Cognitive Science Society*. August 1–4, 2011, Sapporo Japan.
- Gabora, L. & Leijnen, S. (2009). How creative should creators be to optimize the evolution of ideas? A computational model. *Electronic Proceedings in Theoretical Computer Science*, 9, 108–119.
- Gabora, L. & Aerts, D. (2009). A mathematical model of the emergence of an integrated worldview. *Journal of Mathematical Psychology*, 53, 434–451.
- Gabora, L., O’Connor, B. & Ranjan, A. (in press). The recognizability of individual creative styles within and across domains. *Psychology of Aesthetics, Creativity, and the Arts*
- Gino, F. & Ariely, D. (2011). The dark side of creativity: Original thinkers can be more dishonest. *Journal of Personality and Social Psychology*. Advanced online publication. doi: 10.1037/a0026406
- Goldblatt, R., Elkins-Abuhoff, D., Gaydos, M., Rose, S. & Casey, S. (2011). Unlocking conflict through creative expression. *The Arts in Psychotherapy*, 38 (2), 104–108.
- Goodwin, F. K. & Jamison, R. (1990). Alcohol and drug abuse in manic-depressive illness. In: Goodwin, F. and Jamison, K., Eds. *Manic-depressive illness*, pp. 210–226. New York: Oxford University Press.
- Guilford, J. P. (1968). *Intelligence, creativity and their educational implications*. San Diego: Knapp.
- Hannemann, T. B. (2006). Creativity with dementia patients: Can creativity and art stimulate dementia patients positively? *Gerontology*, 52(1), 59–65.
- Leijnen, S. & Gabora, L. (2009). The tradeoff between degree of creativity and number of creators in a computational model of society. In B. Cooper & V. Danos (Eds.) *Proceedings of Developments in Computational Models: Computational Models from Nature (DCM 09)* -- A workshop in association with the 36th International Colloquium on Automata, Languages, and Programming (ICALP). July 11, Rhodes, Greece.
- Ludwig, A. M. (1995). *The price of greatness*. New York: Guilford Press.
- Maslow, A. H. (1959). Creativity in self-actualizing people. In H. Brothers (Ed.), *Creativity and its cultivation*. New York: McGraw-Hill.
- May, R. (1975). *The courage to create*. New York: Bantam.
- Nofresteri, A. & al-ghorabaie, F. M. (2011). Emotional intelligence and creativity in university students. *Journal of Iranian Psychologists*, 7(26), 175–186.
- Rogers, C. (1959). Toward a theory of creativity. In H. Anderson (Ed.), *Creativity and its cultivation*. New York: Harper & Row.
- Runco, M. (2010). Divergent thinking, creativity, and ideation. In (J. Kaufman & R. Sternberg, Eds.) *Cambridge handbook of creativity*. (pp. 413–446). Cambridge UK: Cambridge University Press.
- Schwarz, N. (2002). Situated cognition and the wisdom in feelings: Cognitive tuning. In L. F. Barret & P. Salovey (Eds.), *The wisdom in feeling: Psychological processes in emotional intelligence*. New York: Guilford Press.
- Simonton, D. K. (1999). Creativity as blind variation and selective retention: Is the creative process Darwinian? *Psychological Inquiry*, 10, 309–328.
- Simon, H. A. (1973). Does scientific discovery have a logic? *Philosophy of Science*, 40, 471–480.
- Simon, H. A. (1986). Understanding the processes of science: The psychology of scientific discovery. In T. Gamelius (Ed.), *Progress in science and its social conditions* (pp. 159–170). Oxford: Pergamon Press.
- Sternberg, R. (2005). WICS: A model of positive educational leadership comprising wisdom, intelligence, and creativity synthesized. *Educational Psychology Review*, 17, 191–262.
- Sternberg, R., Bonney, C. R., Gabora, L., Jarvin, L., Karlitz, T. M. & Coffin, L. (2010). Broadening the spectrum of undergraduate admissions. *College & University*, 86(1), 2–17.
- Sternberg, R., Bonney, C. R., Gabora, L., & Merrifield, M. (2012). WICS: A model for college and university admissions. *Educational Psychologist*, 47(1), 30–41.
- Sternberg, R. J., & Lubart, T. I. (1995). *Defying the crowd: Cultivating creativity in a culture of conformity*. New York: Free Press.
- Sulloway, F. (1996). *Born to rebel*. New York: Pantheon.
- Tan, A., Ho, V., Ho, E. & Ow, S. (2008). High school students’ perceived creativity self-efficacy and emotions in a service learning context. *The International Journal of Creativity and Problem Solving*, 18(2), 115–126.

## Appendices

Appendices A, B, C, and D provide examples of stories that scored high or low with respect to creativity, generated in response to non-threatening or threatening photographs. (The stories may end abruptly because they were asked to stop at the end of the allotted writing time.)

### Appendix A

The following is an example of a high scoring story produced in response to the non-threatening stimulus shown in Figure 1.

In Jaredsville, Fourth of July parade was something everyone in town looked forward to. The sleepy town in Missouri came together once a year to provide food, movies, festivities, and friendship to their fellow townsman. The streets are filled with people, some searing food on the grill while others paint children's faces into tigers and clowns.

But everyone came, of course, to see the Jaredsville professional dance exhibition. A few of the ladies in town had once been professional dancers, traveling across the world displaying their beauty and raw skills. They all began to slowly get in position with the start of the drums. The seguin on their clothes danced in the sunlight to the movement of their bodies. The drums began beating faster and faster and the women expertly kept with the beat, dancing faster and more radically until their bodies looked like surges of light. The effect was almost hypnotizing as the crowd had quieted down; became fixated by the dancing flashes of light.

And just as quickly as they had started, they finished. As they left the middle of the street, the crowd applauded them fiercely. Yet, something had changed; the crowd was no longer the large bustling rabble it once was. It had now been neutralized, as if soother by the dancers.

### Appendix B

The following is an example of a low scoring story produced in response to the non-threatening stimulus shown in Figure 1.

The town had a celebration today, an annual celebration that celebrated the culture and music of the country. The women were the center of the attention, they dressed in black or white like the men did, but they were skirts. The women danced to the music and paraded around the whole town. Some knew exactly what was going on and the moves to perform such a dance. Others had their eyes towards the 'leader', who perhaps was the one who made the routine.

The bright sunny day made it all the more amazing to watch the show, as old men were wearing matching hats and marching around with drums. There were people everywhere along the sides of the road, up in the balcony, just to get a glimpse of the festivities that were happening. Some cheered and waved around, and some took pictures. All of the stores in that main road seemed to be closed, probably because of how hectic and wild the place was getting!

It probably was a great success; the warm weather, the loud music, and cheering of people made the day so great! This celebration will most likely continue with similar events every year. After all, it is a celebration of culture and music to the country!

### Appendix C

The following is an example of a high scoring story produced in response to the threatening stimulus shown in Figure 3.

The man pushed Cheryl under the truck, "Where is he?" he was screaming at her. Cheryl could only sob more, trying to control the

emotion was impossible, an endless waterfall of emotions falling over her was all she could feel. "I said, where is he!" The unknown man yelling once again, for the first time in days she felt something, the clasp of his hand over her face. Only days earlier someone she loved had placed his hands in the same place, but this time there was no tenderness, no caress to show they cared. "I will put your body under the tire and back over you slowly, crushing one bone at a time 'till you tell me where the man I am looking for is!" She looked down still focusing on the hand on her face, how she missed him! A few days ago, her lover had left. Where, she had no idea, "Business to take care of" he had said, but what business would leave her alone so this man, the one with the strangling hold on her face, could find her. She had tried to run, but with everyone here who only cared for themselves (all they could care for really with so much sickness and death around), no one answered her screams, her frantic running and trying to escape this man. She ripped her shirt on one of the branches she hit and had the sensation of a bruise forming. She paid no heed to it, only a small discomfort when she compared it to the gaping hole she felt in her chest. Suddenly the man's coat began vibrating, he answered the phone with his hand still around her neck. "Mmmm ok, yes, dead? Good." Without warning he dropped her to the ground and stood looking over her. "Turns out we don't need you" he said leering, "but I do know some of the boys may want you". He picked her up by the hair and began dragging her to the truck "And who knows? Maybe we will let you see your man's body before we lay you down for us" Cheryl tried to force herself once more, with less gusto than before as a part of her heart was crying out what the bald man said to be true was not! She would believe he was alive until she laid her head on his still, dead chest and felt him one last time.

### Appendix D

The following is an example of a low scoring story produced in response to the threatening stimulus shown in Figure 3.

His hands were cold and gritty upon my face they began to turn white as he squeezed harder and threatened me for the money. All I wanted to do was run away, run away from this life in general. He looked at me with such disdain. I wonder how I even got to this point in my life. My heart raced while tears poured down my face, he was a vile man with the intent to kill. "Please" I repeated over and over, "give me one more chance. I'll get you your money." He spat on the ground under the bridge at my feet like I was a piece of scum. I wanted a different life but this grease bag was what I had to deal with every day. Constantly abusing every inch of me. The only thing that scared me more than staying with him was leaving him. He'd find me, he always did. I didn't just need to break away physically, but emotionally and for my own safety. The numbness of the drug he provided was my only gateway out of this hell hole. This could have been the last time I ever saw daylight. Daylight and his eyes, burning with power over every meek creature that would obey him.

# Automatic selection of eye tracking variables in visual categorization for adults and infants

Samuel Rivera<sup>1</sup>, Catherine A. Best<sup>2</sup>, Hyungwook Yim<sup>2</sup>, Aleix M. Martinez<sup>1</sup>, Vladimir M. Sloutsky<sup>2</sup>, Dirk B. Walther<sup>2</sup>

<sup>1</sup>Department of Electrical and Computer Engineering, 205 Drees Lab, 2015 Neil Avenue, Columbus, OH 43210

<sup>2</sup>Department of Psychology, 225 Psychology Building, 1835 Neil Avenue, Columbus, OH 43210

## Abstract

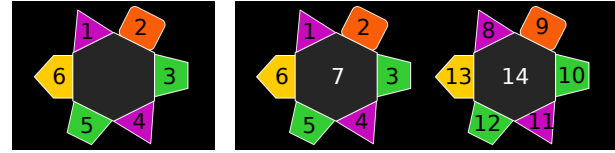
Visual categorization and learning of visual categories exhibit early onset, however the underlying mechanisms of early categorization are not well understood. The main limiting factor for examining these mechanisms is the limited duration of infant cooperation (10-15 minutes), which leaves little room for multiple test trials. With its tight link to visual attention, eye tracking is a promising method for getting access to the mechanisms of category learning. But how should researchers decide which aspects of the rich eye tracking data to focus on? To date, eye tracking variables are generally handpicked, often resulting in biases of the eye tracking data. Here, we propose an automated method for selecting eye tracking variables based on their usefulness to discriminate learners from non-learners of visual categories. We presented infants and adults with a category learning task and tracked their eye movements. We then extracted an over-complete set of eye tracking variables encompassing durations, probabilities, latencies, and the order of fixations and saccadic eye movements. We applied the statistical technique of ANOVA ranking to identifying those variables among this large set that covaried with the learner/non-learner label. We were able to identify learners and non-learners above chance level using linear SVM and the top eye tracking variables.

**Keywords:** eye tracking; category learning.

## Introduction

Categorization is the process of forming an equivalence class, such that discriminable entities elicit a common representation and/or a common response. This task has two components (a) category learning (i.e., forming a common representation for a set of items and/or learning a common response) and (b) categorization or classification (i.e., partitioning a stimulus set using this common representation). While category learning exhibits an early onset (Quinn, Eimas, & Rosenkrantz, 1993), relatively little is known about the underlying mechanism and the development of early categorization. One of the reasons is the limited duration of typical experimental paradigms with infants, yielding only a small number of data points per participant. These limitations have restricted researchers ability to answer fundamental questions about categorization in infants: How do infants learn a category? And what changes in the course of development?

Given that eye movements are tightly linked to visual attention (see Rayner, 1998, for a review), eye movements can provide critical information of how attention allocation changes during category learning. However, eye tracking yields a large amount of data, and it is usually not clear from the outset how to make sense of these data. These data could be converted into numerous variables, and there is no principled way of deciding which of these variables should be selected and why. The *key idea* of our approach is to develop algo-



(a) Category object

(b) Identical category pair

Figure 1: Image (a) is an example category object used in the eye tracking study, with the Areas of interest (AOI)s enumerated. Image (b) defines the AOIs for a pair of objects. The numbers were not shown to participants.

rithms that *automatically* select the eye tracking variables that are systematically linked to category learning.

Our approach was as follows. We extracted a large set of possible variables from the adult or infant gaze sequence during a categorization task (e.g. fixations, saccades, gaze sequences, etc). Some of the variables have been used in analyzing categorization experiments, whereas others are new. We used ANOVA to identify the eye tracking variables that best predict category learning in adults and subsequently in infants, and validated the variables using linear SVM. The significant contribution of this work is that it provides a methodology for identifying eye-tracking variables that are linked to category learning, thus allowing researchers to better understand categorization.

## Method

### Category Object

The category members were flower-like objects with six petals. An example category object is shown in Fig. 1(a), with the petals enumerated for clarity. The four different categories used were defined by a single petal having a distinguishing color and shape. Specifically, the category defining features were A (pink triangle at AOI 4), B (blue semi-circle at AOI 4), C (orange square at AOI 6), and D (yellow pentagon at AOI 6).

### Adult Experiment

To validate the efficacy of the approach before applying it to infants, adult participants were tested. Adult participants from an introductory psychology course at The Ohio State University underwent a series of category learning tasks while their eye gaze was tracked by a Tobii T60 eye tracker at 60Hz. Adults had normal or corrected-to-normal vision, and sat approximately 60 cm away from the display.

Adult participants were assigned to either a supervised or an unsupervised condition. In the supervised condition participant were advised to look for a single distinguishing fea-

ture prior to the start of the experiment. In the unsupervised condition, no hint was given. Previous research suggests that this hint has large consequences with respect to how quickly participants learn to classify the objects, especially when there are few overlapping features (Kloos & Sloutsky, 2008).

The experiment had 8 blocks, with each block consisting of a learning phase and a testing phase. The learning phase had 8 trials, and during each trial a category member was displayed on the center of the screen one at a time for 1.5 seconds each. The testing phase immediately followed the learning phase. There were 4 testing blocks, and on each testing block a category member of the to-be-learned category and a member of a new category were presented side by side. The images were displayed until the participant identified the learned category object by pressing a key. The position of the learned category member (left or right) was counter-balanced. In addition, in-between trials a randomly located fixation point (cross-hair) directed the participants gaze to a position on the monitor.

The to-be-learned category remained the same for the first 4 blocks. A second to-be-learned category was introduced in the final 4 blocks. If the experiment started with category A or B, the second category was either the C or D, and vice-versa. That is because while the A and B categories were defined by a relevant feature at AOI 4, categories C and D were defined by a relevant feature at AOI 6. This provided a mechanism to verify the reproducibility of the variables determined most important.

### Infant Experiment

The infant experiment was similar to the adult experiment, but was adapted for infants by using a familiarization paradigm. To aid infant learning, category exemplars were shown in pairs on each trial. This was also done so that the images in the learning and testing phases had an identical layout. Furthermore, there was only a supervised condition, in which the infants were presented with a pre-trial fixation video of synchronized sound and motion (e.g., looming with whistle sound) to draw the infants attention to the single category-relevant feature. Once the infant looked at the fixation video, the learning trial commenced. Infants had to accumulate 3 seconds of looking to the category exemplar pairs. Whenever an infant looked away, an attention-grabbing fixation was presented until the infant reconnected with the images on the screen. After accumulating 3 seconds of looking to the stimulus pair, the supervisory fixation video was again presented followed by another learning image pair. This procedure was repeated with 8 learning pairs per block.

The testing phase of the infant experiment used a paired preference test with one category member paired with one non-category member. The standard assumption is that the infant can discriminate between the category and non-category object if he or she consistently looks at one object significantly more than the other object (i.e., whether the infant displays a novelty or familiarity preference). There were two test trials per block, where an exemplar from the learned cat-

egory was paired with an exemplar from a novel category. Test trials were presented for a fixed duration of 6 seconds, and left/right position of familiar or novel category objects was counterbalanced.

### Filtering eye tracking data

The gaze data obtained with the Tobii eye tracker contains noise, missing data, and micro-saccades, which makes identifying true fixations and saccades difficult. Therefore, raw eye tracking data exported from every experimental block was filtered using a Kalman filter (Murphy, 2004) before extracting the variables of interest. The eye gaze data from both the left and right eye were filtered separately. The average of the filtered data from left and right eyes yielded the mean eye gaze data, which were used in the current analyses.

### Labeling the Data

The eye movement sequences during the *learning phase* of the experiment aid in understanding category learning, while the sequences during the *testing phase* aid our understanding of category use. Before applying our methodology to understand these processes, however, the eye tracking data from both the learning and testing phases of the experiments were labeled as *learner* (class 1), *non-learner* (class 0), or *indeterminate* (class 2). Indeterminate samples were not used.

**Adult Labels:** Intuitively, labels for adult data are readily identified based on the accuracy of the responses during the testing phase. An uninterrupted string of correct responses during the testing phase suggests that the participant has learned the category. Each adult experimental block yielded 12 eye movement sequences. These correspond to eye movements during the presentation of 8 exemplar images during the learning phase and 4 test images during the testing phase. Adult participants had 4 blocks of learning and discriminating the same category before switching to a new category. This amounted to 32 samples of the learning phase, and 16 samples of the testing phase for each category per participant. The 16 samples from the testing phase were associated with a 16 digit binary string, called the *response string*. This data structure shows performance over the first and last 4 blocks of the experiment. A one identifies a correct response, while a zero denotes an incorrect response on the associated test trial.

We expect a learner's response string to contain a series of ones beginning within the string and terminating at the end of the response string. This pattern indicates that at some point the participant learned the category and correctly discriminated the category from that point on. A participant who has not learned the category (non-learner) would select one of the two stimuli by chance in each trial. A non-learner could get lucky and achieve a series of correct guesses. In order to determine if a participant is a learner or a non-learner we need to establish a criterion that allows us to reject chance as the cause for a series of ones. The question that we need to answer is how many ones we should expect for a learner. We address this problem by assessing how likely it is that we see a sequence of  $M$  consecutive ones in a binary response string

of length  $R = 16$ . Under the null hypothesis, the participant does not know the category label and selects one of the stimuli by chance, giving her a 50% chance of correctly guessing the category member. Each sequence is equally likely given this assumption, so the probability of guessing at least  $M$  right in a row is the total number of sequences having  $M$  ones in a row ( $(R - M + 1) \times 2^{(R-M)}$ ) divided by the total number of binary sequences of length  $R$  ( $2^R$ ). This yields the probability  $p = (R - M + 1)/(2^M)$ . For  $R = 16$ ,  $M = 10$  is the minimum number that achieves a significance level of  $p < 0.01$  ( $p = 0.0068$ ). Therefore, we rejected the null hypothesis that a participant was guessing randomly when we identified a consecutive string of 10 correct responses.

We call the position of the first correct response in this string of correct responses the point of learning (POL). The *test phase* and *learning phase* samples before the POL were labeled as non-learner, while the samples after the POL were labeled as learner. The learning phase samples from the block associated with the POL were labeled as indeterminate, because it was unclear at exactly which trial during the block the category was learned.

If the learning criterion was not achieved, we then identified the remaining non-learning and indeterminate samples. We first labeled correct responses at the end of the respond string as indeterminate. Those samples did not meet the learning criterion, but might be attributed to learning late in the experiment. The remaining samples were labeled as non-learner. Approximately 8% of the adult eye track samples were labeled indeterminate.

**Infant Labels:** Obviously, infants are not able to respond by keyboard to identify a category object. Instead, we used a variant of the preferential looking paradigm to determine if an infant could discriminate between novel exemplars of a familiar category object and a novel category object. Recall that the preferential looking paradigm assumes that infants who consistently look more to one class of stimuli when shown two classes of stimuli are able to discriminate between the two classes. This means that if the infant consistently looks longer at the learned category object (or novel category object), then he or she is assumed to be discriminating between the familiar and novel categories.

Given this paradigm, we labeled each infant's gaze data by blocks. Each block consisted of two test phase samples. We determined novelty preference as the ratio of total looking time to the novel category object compared to the total looking time to the novel category plus the familiar category object. We sorted the mean of the novelty preference for each block according to the absolute difference from 0.5. A third of the blocks with mean novelty preference closest to 0.5 were labeled as non-learner. The third of the blocks with novelty preference furthest from 0.5 in absolute value were labeled learner. Otherwise, the samples were labeled indeterminate. Approximately 33% of the infant eye track samples were labeled indeterminate.

## Variables List

We compiled an over-complete list of eye tracking variables. We began with the fundamental variables, fixations and saccades. Fixations occur when eye gaze is maintained at a single position for at least 100ms. They were identified using the dispersion threshold algorithm of (Salvucci & Goldberg, 2000). Saccades are rapid eye movements that move the eye gaze between points of fixation. To be considered a saccade, the eye movement needed to exceed smooth pursuit velocity of  $30^\circ$  per second or  $0.5^\circ$  per sample at 60Hz (Stampe, 1993). The fixations and saccades were determined with respect to a specific Area of Interest (AOI) within an object. AOIs are regions of an object image or scene that can be grouped in some meaningful way, such as color uniformity, and are relevant or non-relevant based on their role in determining the object category.

These fundamental eye tracking variables were combined in various ways to derive a larger set of variables. Our variables list is defined as follows:

1. *AOI fixation percentage* describes the percentage of time fixated at the different AOIs during a trial. All non-AOI fixations were discarded in this and all of the variables defined. The fixation percentages were normalized so that they sum to 1, unless there were no fixations at AOIs. In that case, all percentages were set to 0.
2. *Relevant AOI fixation density* describes the percentage of time fixated at the relevant AOI(s).
3. *AOI fixation sequence* describes the sequence of AOI fixations during one trial. We limited this sequence to a fixed number of fixations, starting with trial onset (not counting fixations to the fixation mark). The number of fixations to consider as well as the start position were determined using cross validation (CV). In addition, the fixation sequence was represented as a sequence of relevant and non-relevant AOI fixations. The analysis showed that the latter representation was more informative in some cases.
4. *Duration of fixations in sequence* describes the duration of each fixation in the sequence described by variable 3.
5. *Total distance traveled by eye* is a scalar describing the total distance traveled by the eye gaze during a trial.
6. *Histogram of fixation distances to relevant AOI* describes how much time is spent fixated near or far from the relevant AOI(s). The number of bins was determined using CV.
7. *Number of unique AOIs visited* is a scalar describing the total number of unique AOIs fixated during a trial.
8. *Saccade sequence* is similar to variable 3, but describes the sequence of AOI saccades during one trial. All saccades whose targets were not to AOIs were discarded in this and all of the variables defined. The sequence was limited to a fixed number of saccades, starting at the first saccade. The

number of saccades to consider as well as the start saccade were determined using CV. In addition, the saccade sequence was represented as a sequence of saccades to relevant and non-relevant AOIs.

9. *Relative number of saccades to an AOI* is the saccade analogue of variable 1, and describes the relative number of saccades to the AOIs during eye movement.
10. *Fixation latency to relevant AOI* describes the delay before fixating at a relevant AOI during an eye movement. It is a scalar between 0 and 1, where 0 corresponds to fixating to a relevant AOI immediately and 1 describes a sequence where a relevant AOI is never fixated.
11. *Saccade latency to relevant AOI* is a scalar between 0 and 1 defining the delay before saccading to a relevant AOI.

Thus, eye movement was represented by a *feature vector*  $\mathbf{x} = (x_1, x_2, \dots, x_d)^T$  whose  $d$  entries correspond to the variables described. For clarity, *features* denote the entries of the feature vector which encodes the eye tracking variables, while *variables* correspond to the measures of the eye tracking enumerated above. Therefore,  $d$  is much larger than 11, because encoding certain variables requires multiple feature values. In addition, the information encoded by several of these features overlaps. This over-complete representation allows us to find the encoding that is best suited to describe the categorization task. To this end we performed variable selection on this over-complete set.

### Variable Selection

Our goal is to identify the subset of variables from the set defined above that can best separate the classes: category learners and non-learners. This was achieved using ANOVA feature selection by ranking. ANOVA feature selection relies on a standard hypothesis test on each feature of  $\mathbf{x}$ . Specifically, let  $x_i$  denote the  $i^{th}$  feature of  $\mathbf{x}$ . Using a dataset of eye tracking feature vectors and the associated class labels, we performed a two tailed  $t$ -test of the null hypothesis which states that samples of  $x_i$  coming from classes 1 and 0 are independent random samples from normal distributions with equal means,  $\mu_{i1}$  and  $\mu_{i0}$ , respectively. The alternative says that the class means are different. We calculated the test statistic and the corresponding  $p$ -value. A low  $p$ -value means the null hypothesis is rejected with confidence. Since the goal is to find the variables which best separate the classes, the feature with lowest  $p$ -value is ranked as best. The  $p$ -values were calculated for all features  $x_i, i = 1 \dots d$ , and they were ranked from best to worst according to increasing  $p$ -values. If we vectorize the indices of the  $t$  top ranked features as  $\mathbf{k} = (k_1, k_2, \dots, k_t)^T$ , then after feature selection  $\mathbf{x} = (x_{k_1}, x_{k_2}, \dots, x_{k_t})^T$ .

### Linear Classification

Once the important variables were identified, we used them to classify the gaze data as having originated from a learner

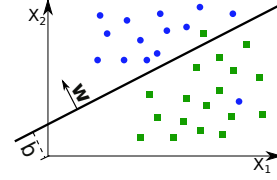


Figure 2: Illustration of a linear classifier.  $\mathbf{w}$  is the normal vector of the hyperplane which separates the feature space into two decision regions, and  $b$  is the distance from the origin to the hyperplane. The blue circles represent samples from class 1, while the green squares represent samples from class 0. All but one of the blue circles exists on the positive side of hyperplane, and are classified correctly.

or non-learner. This required that we train a classifier to distinguish between two classes of data. Recall that each eye movement results in a feature vector, or *sample*  $\mathbf{x}$ . A classifier defines a decision rule for predicting whether a sample is from class 0 or 1. A linear classifier was used because of its ease of interpretation (Martinez & Zhu, 2005) – the absolute model weights give the relative importance of the eye tracking variables. We illustrate in Fig. 2 with a 2-dimensional linear classifier model specified by  $\mathbf{w}$  and  $b$ .  $\mathbf{w}$  is the normal vector of the hyperplane which separates the feature space into two decision regions, and  $b$  is the distance from the origin to the hyperplane (i.e., the offset).

All samples  $\mathbf{x}$  above the hyperplane are assigned to class 1 while the samples below are assigned to class 0. Data samples  $\mathbf{x}$  existing on the boundary satisfy  $\mathbf{w}^T \mathbf{x} - b = 0$ . Therefore, samples are classified according to the sign of  $\mathbf{w}^T \mathbf{x} - b$ . In this example  $\mathbf{w} = (-.55, .83)^T$ , so the second dimension,  $x_2$ , is more informative for classification because it has a larger absolute value. Note that in our case the feature space has not two but up to 334 dimensions, depending on the cut-off for variable selection.

In this work we used the Support Vector Machine (SVM) classifier. SVM is a linear classifier which maximizes the margin between two classes of data (Burgess, 1998). In the case that the training samples are perfectly separable by a hyperplane, we can find  $\mathbf{w}$  and  $b$  such that the data satisfies the following constraints,

$$\mathbf{x}_i^T \mathbf{w} - b \geq 1 \text{ for } y_i = 1, \quad (1)$$

$$\mathbf{x}_i^T \mathbf{w} - b \leq -1 \text{ for } y_i = 0. \quad (2)$$

Essentially, these constraints specify that the samples from the different classes reside on opposite sides of the decision boundary. The margin between the classes, defined by  $\frac{2}{\|\mathbf{w}\|_2}$  where  $\|\cdot\|_2$  defines the L2-norm, is then maximized subject to the above constraints. The dual formulation of the constrained optimization problem results in a quadratic program for  $\mathbf{w}$  and  $b$ . In the case that samples from each class are not linearly separable, a penalty is introduced to penalize the amount that a sample is on the wrong side of the hyperplane. Again, the dual formulation results in a quadratic program for  $\mathbf{w}$  and  $b$ . We used the implementation of (Chang & Lin, 2001).

The classification accuracy used for adults was the leave-



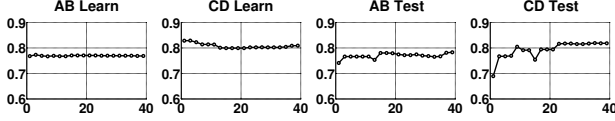


Figure 3: Leave-one-subject-out cross-validation accuracy for adult participants as a function of the number of top ranked variables used for classification. Classification was done with SVM while ANOVA was used for feature selection. AB and CD correspond to category object A or B and C or D respectively. In almost all cases, the classification accuracy was near the maximum after including very few features.

one-subject-out cross-validation (LOSO-CV) accuracy. In LOSO-CV, the samples belonging to one participant are sequestered, and the remaining samples are used to train the classifier. The sequestered samples are then classified with the learned classifier, and the procedure is repeated for every participant in the database. The total number of correctly classified samples divided by the total number of samples is the LOSO-CV accuracy. The classification accuracy used for infants was the leave-one-experiment-block-out cross-validation (LOBO-CV) accuracy. This alternative accuracy measure makes more effective use of the eye movement data when the sample size is very small. In LOBO-CV, the samples belonging to one experiment block are sequestered, and the remaining samples are used to train the classifier. The sequestered samples are then classified with the learned classifier, and the procedure is repeated for every block in the database. The total number of correctly classified samples over the total number of samples is the LOBO-CV accuracy.

## Results

### Adult Experiment

A total of 24 adults were tested in the supervised experiment while 46 adults were tested in the unsupervised adult experiment. This resulted in 728 learning class samples and 1256 non-learning class samples for the learning phase, and 473 learning class samples and 601 non-learning class samples for the testing phase in the A or B category learning condition. There were 496 learning class samples and 1568 non-learning class samples for the learning phase, and 323 learning class samples and 717 non-learning class samples for the testing phase in the C or D category learning condition. The indeterminate samples were not used in any of the experiments. After labeling the data, the eye tracking variables were extracted from each gaze sequence. Each labeled data sample resulted in a 182-dimensional feature vector for the learning phase samples, and a 334-dimensional feature vector for the testing phase samples.

SVM was applied to determine the LOSO-CV error as a function of the number of top features selected by ANOVA feature selection. The results are summarized in Fig. 3, and show that a very small set of variables yields a high classification rate. Adding more variables does not improve the accuracy.

The stable performance beyond just a few variables suggests that a small number of variables is sufficient for dis-

Table 1: The following variables were determined most relevant during the adult category learning and testing phases. The results suggest that looking at the relevant AOI is important during the learning phase. In addition, the first few fixations are very important during testing. The bold face entries show variables that were consistently determined most relevant across category conditions. We used shorthand notation for a few words: fixation (fix), saccade (sac), relevant (rel).

A or B Learning Condition	C or D Learning Condition
<b>1. Latency to relevant AOI fix</b>	<b>1. Latency to relevant AOI fix</b>
<b>2. Density of fix at AOI 4</b>	<b>2. Density of fix at AOI 6</b>
3. Distance to AOI 4, hist bin 2 of 30	3. Distance to AOI 6, hist bin 5 of 30
4. Look at AOI 4 on fix 2	<b>4. Look at AOI 6 on fix 1</b>
<b>5. Look at AOI 4 on fix 1</b>	5. Look at non-rel AOI on fix 1
A or B Testing Condition	C or D Testing Condition
<b>1. Look at non-relevant AOI on fix 3</b>	1. Look at non-relevant AOI on fix 4
2. Look at non-relevant AOI on fix 2	<b>2. Look at non-relevant AOI on fix 3</b>
<b>3. Look at non-relevant AOI on sac 2</b>	<b>3. Look at non-relevant AOI on sac 2</b>
4. Look at non-relevant AOI on fix 1	<b>4. number of unique AOIs fixated</b>
<b>5. number of unique AOIs fixated</b>	5. Look at non-relevant AOI on sac 3

criminating learners and non-learners. The top five variables are listed in Table 1. The bolded entries are present in the top five variables across both the A or B and C or D conditions. Note that AOI 4 for the A or B condition is equivalent to AOI 6 in the C or D test condition.

### Infant Experiment

A total of 16 infants ranging from 6 to 8 months of age were tested in the supervised infant experiment. One participant's data were discarded because the infant would not cooperate. This resulted in 135 learning class samples and 137 non-learning class samples for the learning phase, and 40 learning class samples and 40 non-learning class samples for the testing phase in the A or B category learning condition. For the C or D category learning condition this resulted in 139 learning class samples and 127 non-learning class samples for the learning phase, and 40 learning class samples and 40 non-learning class samples for the testing phase. The indeterminate samples were not used in any of the experiments. After labeling the data, the eye tracking variables were extracted from each gaze sequence. Each labeled data sample resulted in a 334-dimensional feature vector for the learning and testing phase samples.

SVM was applied to determine the LOBO-CV error as a function of the number of top features selected by ANOVA feature selection. The results are shown in Fig. 4, where we see that the accuracy varies when considering different features. Thus, while classification of learners and non-learners is still possible with infants, it is not as clear-cut as with adults. This is to be expected because of the amount of random movements typical of babies. The top five infant variables are shown in Table 2. There are no bold entries because no variables were consistently selected across the A or B and C or D conditions.

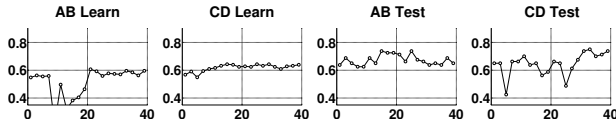


Figure 4: Leave-one-block-out cross-validation accuracy for infant participants as a function of the number of top ranked variables used for classification. Classification was done with SVM while ANOVA was used for feature selection. AB and CD correspond to category object A or B and C or D respectively. Chance level was at 0.5.

Table 2: The following variables were determined most relevant during the infant category learning and testing phases. There were no consistently relevant features across category conditions. We used shorthand notation for a few words: fix-ation (fix), saccade (sac), relevant (rel).

A or B Learning Condition	C or D Learning Condition
1. Den of fix at AOI 10	1. Distance to AOI 13, hist bin 5 of 35
2. Look at AOI 10 on fix 3	2. Distance to AOI 6, hist bin 21 of 35
3. Distance to AOI 11, hist bin 5 of 35	3. Density of fix at AOI 13
4. Density of saccade to AOI 10	4. Distance to AOI 13, hist bin 2 of 35
5. Distance to AOI 4, hist bin 20 of 35	5. Look at non-rel AOI on sac 3
A or B Testing Condition	C or D Testing Condition
1. Look at non-relevant AOI on fix 6	1. Density of fix at AOI 7
2. Distance to AOI 4, hist bin 20 of 35	2. Look at AOI 3 on sac 3
3. Distance to AOI 4, hist bin 8 of 35	3. Look at AOI 7 on fix 1
4. Look at AOI 7 on fix 4	4. Look at AOI 10 on fix 4
5. Density of sac to AOI 10	5. Look at AOI 7 on fix 3

## Comparing Infants to Adults

The above results raise a new question. How similar are the attention models of adults and infants? Specifically, since the infant data is so noisy, can we use the adult model to improve on the infant one? To test this, we used the adult classifier model trained with the top five variables to predict if infants were learners or non-learners. This was done only for the testing phase, because the testing phase images for adults and infants are similar so that the extracted variables correspond. Infants were classified with 49% accuracy in the A or B condition, and with 50% accuracy in the C or D condition with chance level at 50%. These findings suggest that infant category learners do not direct their attention like adult learners.

## Discussion and Conclusion

We have developed a methodology for automatically determining eye tracking variables that are relevant to understanding category learning and discrimination processes. Previous research has relied on ad-hoc techniques to determine which variables should be analyzed. Instead, we used statistical methods to find the important variables in an over-complete set of variables.

The efficacy of the approach was verified with an adult categorization study. The variables determined most relevant for adults emphasize looking at the relevant AOI(s) longer, and earlier during the categorization tasks. This result is satisfying for two reasons: 1) It is expected that category learners quickly focus their efforts on the relevant AOI(s), and 2) These variables coincide with the variables *proportion fixation time* and *relative priority* of previous eye-tracking category learning studies such as (Rehder & Hoffman, 2005). Finally, we demonstrated that the adult model does not predict

infant categorization. This is evidence of different attention processes for infants and adults during categorization.

Note that the important variables were verified by the *task* and *stimuli* described. Altering these parameters may result in different important variables. By comparing the important variables among different tasks and stimuli, we can further dissociate which eye tracking variables are linked to specific processes during categorization.

## Acknowledgments

This research was partially supported by NIH grant R01 EY-020834 to AM, NSF grant BCS-0720135 and NIH grant R01 HD-056105 to VS, and a Seed Grant by the Center for Cognitive Science (CCS) at OSU to DBW, VS, and AM. SR was partially supported by a fellowship from the CCS.

## References

- Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2, 121–167.
- Chang, C.-C., & Lin, C.-J. (2001). LIBSVM: a library for support vector machines [Computer software manual]. (Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>)
- Kloos, H., & Sloutsky, V. M. (2008). What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General*, 137(1), 52–72.
- Martinez, A. M., & Zhu, M. (2005). Where are linear feature extraction methods applicable? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(12), 1934–1944.
- Murphy, K. (2004). *Kalman filter toolbox for matlab*. Available from <http://www.cs.ubc.ca/~murphyk/Software/Kalman/kalman.html>
- Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception*, 22(4), 463–475.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372–422.
- Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, 51, 1–41.
- Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Etra '00: Proceedings of the 2000 symposium on eye tracking research & applications* (pp. 71–78). New York, NY, USA.
- Stampe, D. M. (1993). Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems. *Behavioral Research Methods, Instruments, & Computers*, 25(2), 137–142.

# Categorisation in High and Low Schizotypes

Aaron C. Roberts (A.Roberts.10@unimail.winchester.ac.uk)

Department of Psychology, University of Winchester, Winchester, SO22 4NR, UK

Nick Braisby (Nick.Braisby@winchester.ac.uk)

Department of Psychology, University of Winchester, Winchester, SO22 4NR, UK

## Abstract

Disorders provide an important source of information in developing theories of normal categorisation. Disruption in categorisation in individuals diagnosed with schizophrenia has been widely evidenced. However, findings are often contradictory and subject to significant confounds. In the present study 35 high schizotypes and 35 low scorers completed a semantic categorisation task and a borderline categorisation task, with measures of category membership judgment, similarity and response time being taken. Results revealed that high schizotypes made significantly fewer positive category membership judgments than low schizotypes ( $p = .003$ ) and suggest that different theoretical explanations may be required to explain the categorisation of high and low schizotypes. Explanations in terms of theories of normal categorisation are developed.

**Keywords:** Concepts; Categorisation; Schizotypy; Essentialism; Semantic Processing.

## Introduction

Disorders of categorisation are an important source of information in developing theories of normal categorisation. Work on category-specific impairment, for example, has led to the proposal that functional and appearance attributes are represented in concepts in different ways (cf. Strnad, Anzellotti & Caramazza, 2011).

Categorisation is known to be disrupted in other disorders, most notably dementia (Doughty et al., 2009), autism (Church et al., 2010), and schizophrenia (Doughty & Done, 2009), as well as in neuropsychological cases (Cohen, Johnston & Plunkett, 2000).

Schizophrenia has long been associated with the suggestion that categorisation is subject to over-inclusion, that category boundaries are shifted outward, incorporating into the category items that would normally be regarded as non-members (Lawrence, Doughty, Al-Mousawi, Clegg & Done, 2007). However, disorders also present difficulties for experimental investigations of categorisation. In the case of schizophrenia, for example, significant confounds are often present, such as hospitalisation, medication, psychosis, and attentional dysfunction. Perhaps unsurprisingly, and in spite of the prominence the suggestion of over-inclusion has received, relatively little is actually known of categorisation in schizophrenia. Moreover, the sparse evidence is often inconsistent or contradictory.

However, schizophrenia has been linked theoretically to schizotypy, a multidimensional construct that assumes healthy individuals may manifest subclinical presentations of schizophrenic symptoms. Schizotypy varies in degree

along a continuum from such psychologically healthy individuals to those diagnosed with schizophrenia (Claridge, 1985). Those high in schizotypy are described as psychosis-prone, and may experience oddities of belief, behaviour, eccentricities, idiosyncratic speech, peculiar ideas, and social awkwardness or aversion at a subclinical level (Siever, Kalus & Keefe, 1993). High schizotypes have been studied in order to gain understanding of vulnerabilities to schizophrenia (Peters, Pickering & Hemsley, 1994). Indeed, in the presence of sufficient environmental stressors, the contention is that high schizotypes may develop schizophrenia, and present with appropriate clinical symptoms.

High schizotypes therefore provide the researcher with a compelling opportunity to examine 'disordered' cognition in an otherwise healthy population. However, there have been relatively few studies of categorisation in high and low schizotypes, and what data there are appear inconsistent.

## Semantic Processing in Schizophrenia

Abnormalities in semantic processing are thought to be central to cognitive abnormalities in schizophrenia, with deficits reported on a wide variety of semantic processing tasks (Chen, Wilkins & McKenna, 1994). Further, semantic deficits are suggested to underlie disturbances in thought and language in schizophrenia, which might not only explain deficits observed in other cognitive domains, but also provide a cognitive explanation for common symptoms in schizophrenia, such as delusions (Rossell, Rabe-Hesketh, Shapleske & David, 1999) and thought disorder (Gouzoulis-Mayfrank et al., 2003).

## Semantic Processing in Schizotypy

Although impairments in attention (Lenzenweger, Cornblatt & Putnick, 1991) and executive functioning (Suhr, 1997; Tallent & Gooding, 1999) have been found in schizotypy, few studies have addressed the relationship between schizotypy and semantic processing. Those that have done so have used semantic priming tasks (Beech, McManus, Bayliss, Tipper & Agar, 1991; Morgan, Bedford & Rossell, 2006) and revealed differences between high and low schizotypes (Morgan et al., 2006).

However, there has been little research on other areas of semantic processing in schizotypy, Morgan, Bedford, O'Reagan & Rossell (2009) being a notable exception. Very few studies of schizotypy have included categorisation tasks and none have made categorisation a primary focus.

## Categorisation in Schizotypy

The majority of studies on categorisation in schizotypy have used fluency tasks whereby participants are asked to generate exemplars given a category label (Barrantes-Vidal et al., 2003; Duchene, Graves & Brugger, 1998; Kiang & Kutas, 2005). Fluency tasks are typically used to make estimates of the semantic distance between pairs of concepts. Although these provide a measure of the organisation of semantic information, they do not address the process of categorisation directly. Categorisation tasks, which require participants to make category membership judgments, have been used in only two studies (Kiang & Kutas, 2005; Morgan et al., 2009).

Kiang & Kutas (2005) presented their participants with a category definition followed by exemplars of varying typicality and required them to judge whether or not each was a category member. In their EEG study, no group differences were found in the N400 component, although they did report a negative correlation between schizotypy score and ERP amplitude differences between category members and non-members.

Morgan et al. (2009) asked participants to rate the category membership of exemplars of varying degrees of relatedness. Group differences were found only for low frequency items, with high schizotypes regarding low frequency exemplars as belonging less to the category than did low schizotypes. Differences in the same direction were reported as nearing significance for high frequency category members and borderline exemplars.

Surprisingly, the authors did not comment on the contrast between this apparent ‘under-inclusion’ and the over-inclusion reported in schizophrenia. Indeed, despite the suggestion that in schizophrenia categories are over-included, there has been very little focus on this question in relation to schizotypy. It is unclear whether differences in categorisation between high and low schizotypes are generally not reliable, or whether previous studies have not been sufficiently sensitive to detect them.

## Categorising Borderlines

Previous investigations of schizotypy have not examined the relative contribution of different attributes to categorisation. Yet characteristic and necessary features (cf. Rips, Shoben & Smith, 1973) allow for the creation of two borderline cases: exemplars with characteristic but not necessary features and exemplars with necessary but not characteristic features. Similar cases have been used in previous research, for example, in debates as to whether categorisation is similarity- or theory-based (e.g. Rips, 1989).

This study therefore consists of two tasks: a replication of the semantic categorisation task reported by Morgan et al. (2009) with the additional measures of response times, and similarity ratings to better gauge whether categorisation in schizotypy is characterised by over- or under-inclusion; and a borderline categorisation task using two types of borderline exemplars, as above, for artefact and natural kind categories, with measures of categorisation and similarity

judgments, and response times, in order to shed light on the ways in which category boundaries might be shifted.

## Experiment

### Method

#### Design

The study employed two tasks. In both, dependent variables were categorisation (member, non-member), similarity rating (1-7), and response time (ms).

**Semantic categorisation** This employed a 2 x 5 design with one between-participants factor (Group [low, high]) and one within-participants factor (Relatedness [high frequency, low frequency, borderline, related but outside of the category, unrelated]).

**Borderline categorisation** This employed a 2 x 2 x 2 x 3 design with one between-participants factor (Group [low, high]) and three within-participants factors (Appearance [+,-], Essence [+,-] and Category [artefact, food natural kinds, non-food natural kinds]).

#### Participants

Two hundred and seventy eight participants were screened using an online questionnaire, the Oxford-Liverpool Inventory of Feelings and Experiences (O-LIFE: Mason, Claridge & Jackson, 1995). Scores from the questionnaire were used to determine high and low groups: high group ( $n = 35$ ), above the 70th percentile ( $\geq 44$ ) and the low group ( $n = 35$ ), below the 30th percentile ( $\leq 29$ ). The high group consisted of 6 males and 29 females with a mean age of 25.83 ( $SD = 9.39$ ) and the low group consisted of 9 males and 26 females with a mean age of 27.54 ( $SD = 11.37$ ).

#### Materials

**O-LIFE** The O-LIFE is a 159-item questionnaire based on the Combined Schizotypal Traits Questionnaire (Bentall et al., 1989) and is used to measure schizotypy. The O-LIFE yields 4 factors: unusual experiences (e.g. ‘Do you think that you could learn to read other’s minds if you wanted to?’), cognitive disorganisation (e.g. ‘Are you easily confused if too much happens at the same time?’), introverted anhedonia (e.g. ‘Are there very few things that you have ever enjoyed doing?’), and impulsive nonconformity (e.g. ‘Do you at times have an urge to do something harmful or shocking?’).

**Semantic categorisation** The semantic categorisation task replicates and extends the task reported by Morgan et al. (2009). They had selected eighteen categories from the norms of Battig and Montague (1969) and of Hampton and Gardiner (1983): body parts, clothing, drinks, flowers, food flavouring, furniture, insects, instruments, mammal, metal, part of building, professions, reading material, sport, tools, type of cloth, vehicle, and weapon. For each category,

Morgan et al. (2009) identified 5 different exemplars of differing degrees of relatedness, resulting in 90 trials, as follows: (1) high frequency (e.g. leg for the category ‘body part’), (2) low frequency (e.g. thumb), (3) borderline (e.g. joint), (4) related but outside the category (e.g. wig), and unrelated (e.g. cricket). To extend the replication, response times (ms) and similarity ratings were also measured.

**Borderline categorisation** Based on Braisby (2004), this task employed three types of category (food natural kinds, non-food natural kinds and artefacts), with 4 different categories for each: apple, chicken, potato and salmon for food natural kinds; canary, dog, oak tree and rose for non-food natural kinds; and car, fork, piano and sailboat for artefacts. For each category, 4 different exemplars were described, defined by the presence or absence of appearance and essence properties, resulting in 48 trials. Exemplars were presented in scenarios and were defined by having: (1) appearance properties absent, essence properties absent (A-E-), (2) appearance properties absent, essence properties present (A-E+), (3) appearance properties present, essence properties absent (A+E-), and (4) appearance properties present, essence properties present (A+E+). The following is an example of how stimuli were presented in scenario form for the category ‘apple’, for the exemplar type (A+E-).

‘You have just acquired an apple. You discover that it has been genetically modified so that it has NONE of the genetic properties specific to apples. Upon examination, you find that it looks, feels, smells and even tastes JUST like an apple.’

For natural kind categories, essential properties were expressed in terms of possession of the genetic properties specific to the category; for artefacts, these were expressed in terms of the original intended function (cf. Bloom, 1996).

## Procedure

Participants completed both tasks and stimuli were presented and responses recorded using E-Prime (Schneider, Eschman & Zuccolotto, 2002). The order of tasks was counterbalanced.

**Semantic categorisation** Participants were required to read the exemplar and category names and make a similarity judgment, and then a category membership judgment (Yes or No). Practice examples were used.

**Borderline categorisation** The scenario appeared on screen and participants were required to make a similarity judgment, and then a category membership judgment (Yes or No). Practice examples were used.

## Results

For both semantic categorisation and the borderline categorisation task, category membership judgments were scored 1 for a ‘yes’ response and 0 for a ‘no’ response. For

the borderline categorisation task, category membership judgments, category membership judgment response times, similarity ratings and similarity rating response times were all averaged over the four categories belonging to each superordinate category.

A series of 2 x 5 ANOVAs were conducted to examine category membership judgments, category membership judgment response times, similarity ratings and similarity rating response times.

## Semantic Categorisation

Critically, no group differences were found for measures of category membership judgment, similarity judgment or response times, nor did any approach significance (all  $p > .3$ ). Thus Morgan et al.’s (2009) key finding of group differences for low frequency items was not supported.

Consistent with their findings, however, the main effect of relatedness was significant for category membership judgments [ $F(2.82, 191.80) = 1613.00, p < .001$ , partial  $\eta^2 = .96$ ] and similarity ratings [ $F(2.52, 171.27) = 1558.00, p < .001$ , partial  $\eta^2 = .96$ ], with both ratings increasing with semantic relatedness (see Figure 1).

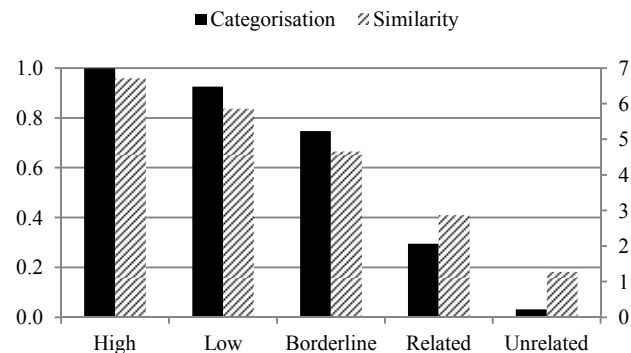


Figure 1. Mean proportion of positive category judgments (left axis) and mean similarity ratings (right axis) for each level of relatedness.

Examining response times, as expected, the main effect of relatedness was also significant for category membership judgments [ $F(2.74, 186.32) = 17.22, p < .001$ , partial  $\eta^2 = .20$ ] and similarity ratings [ $F(3.37, 229.10) = 43.91, p < .001$ , partial  $\eta^2 = .39$ ]. These results support previous findings with categorisation response times following an inverted V-shaped function: response times increase as the semantic distance between the category and the exemplar increases to the boundary, and decrease with increasing semantic distance (Rips et al., 1973).

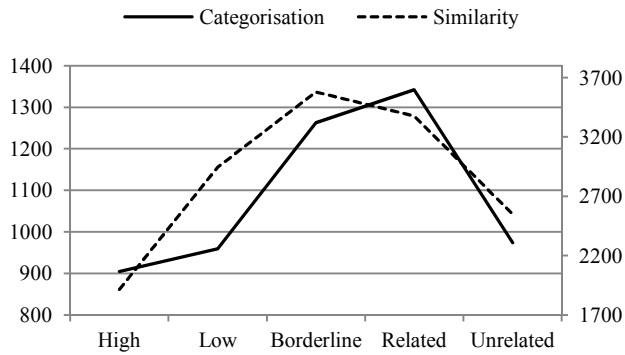


Figure 2. Mean category judgment (left axis) and similarity rating (right axis) response times (ms) for each level of relatedness.

### Borderline Categorisation

A series of  $2 \times 2 \times 2 \times 3$  ANOVAs were conducted to examine category membership judgments, category membership judgment response times, similarity ratings and similarity rating response times.

**Category membership judgments** Critically, the main effect of group was significant [ $F(1, 68) = 9.52, p = .003$ , partial  $\eta^2 = .12$ ], with high schizotypes providing significantly fewer positive category membership judgments ( $M = .43$ ) than low schizotypes ( $M = .52$ ).

The category by appearance interaction was significant [ $F(1.70, 115.31) = 12.03, p < .001$ , partial  $\eta^2 = .15$ , Huynh-Feldt corrected here and elsewhere] and contrasts revealed significant interactions when comparing artefacts to food natural kinds [ $F(1, 68) = 15.48, p < .001$ , partial  $\eta^2 = .19$ ] and artefacts to non-food natural kinds [ $F(1, 68) = 13.80, p < .001$ , partial  $\eta^2 = .17$ ]. The effect of appearance properties was greater for food (0.44) and non-food (0.43) natural kinds than it was artefact categories (0.29).

The category by essence interaction was also significant [ $F(1.68, 113.98) = 30.04, p < .001$ , partial  $\eta^2 = .31$ ]. Contrasts revealed significant interactions when comparing artefacts to food natural kinds [ $F(1, 68) = 41.50, p < .001$ , partial  $\eta^2 = .38$ ], and artefacts to non-food natural kinds [ $F(1, 68) = 31.31, p < .001$ , partial  $\eta^2 = .32$ ]. The effect of essence properties was greater for artefact categories (0.68) than both food (0.43) and non-food (0.46) natural kinds.

Similar effects were obtained for similarity ratings. Taken together, these results imply that overall participants essentialised artefact categories more strongly than they did the food or non-food natural kinds, for which the influence of appearance properties was equally strong.

There was a significant three-way interaction between appearance, essence and group [ $F(1, 68) = 8.02, p = .006$ , partial  $\eta^2 = .11$ ]. Pairwise comparisons revealed differences nearing significance between the high group ( $M = .895, SE = .038$ ) and low group ( $M = .974, SE = .015$ ) for [A+E+] exemplars [ $t(68) = 1.94, p = .059, r = .23$ ] and between the high group ( $M = .462, SE = .062$ ) and the low group ( $M =$

.621,  $SE = .059$ ) for [A-E+] exemplars [ $t(68) = 1.88, p = .064, r = .22$ ]. Pairwise comparisons between the high group ( $M = .352, SE = .056$ ) and low group ( $M = .462, SE = .062$ ) for [A+E-] exemplars were not significant [ $t(68) = 1.31, p = .196, r = .16$ ]. Thus, the low schizotypes tended to give a higher proportion of positive categorisation judgments than high schizotypes when essence properties were present. However, these conclusions must be tempered due to the possibility of a floor effect (i.e. in the [A-E-] condition).

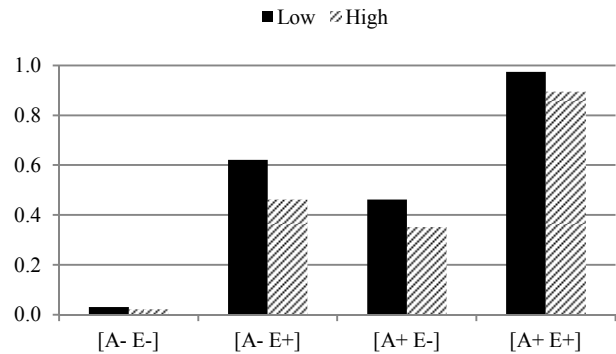


Figure 3. Three-way interaction between appearance, essence and group. Mean proportion of positive category judgments are shown for each exemplar type.

The three-way interaction between category, appearance and group was also significant [ $F(1.70, 115.31) = 3.84, p = .031$ , partial  $\eta^2 = .05$ ]. Contrasts revealed a significant difference between the high and low group when comparing artefacts to food natural kinds [ $F(1, 68) = 6.31, p = .014$ , partial  $\eta^2 = .09$ ]. High schizotypes (0.34) were more influenced by appearance properties than low schizotypes (0.24) when categorising artefacts. However, low schizotypes (0.49) were more influenced by appearance properties than high schizotypes (0.39) in categorising food natural kinds.

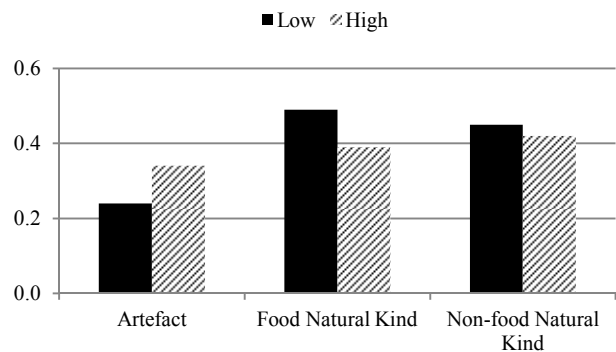


Figure 4. Effect of appearance properties (A+ minus A- category membership judgments) for high and low group, artefact, food natural kind and non-food natural kind categories.

**Category judgment response times** Importantly, the main effect of group was not significant [ $F(1, 68) = .54, p = .465$ , partial  $\eta^2 = .008$ ] suggesting that the difference in categorisation judgments does not stem from a speed-accuracy trade-off. The category by group interaction was significant [ $F(1.91, 129.53) = 3.40, p = .039$ , partial  $\eta^2 = .048$ ]. Contrasts revealed a significant interaction when comparing the high schizotypy group and low group response times for the artefact and the non-food natural kind categories [ $F(1, 68) = 5.27, p = .025$ , partial  $\eta^2 = .072$ ]. For the artefact category, response times were similar for the low group and the high group, however for the non-food natural kinds category response times decreased for the high group and increased for the low group. Pairwise comparisons revealed differences nearing significance between artefact categories ( $M = 1508, SE = 126.3$ ) and non-food categories ( $M = 1895, SE = 215.6$ ) for the low group [ $t(34) = -1.98, p = .056, r = .32$ ], while those between artefact categories ( $M = 1612, SE = 203.5$ ) and non-food categories ( $M = 1492, SE = 178.5$ ) for the high group were not significant [ $t(68) = 1.18, p = .247, r = .20$ ].

**Similarity ratings** The main effect of group was not significant [ $F(1, 68) = .008, p = .928$ , partial  $\eta^2 = .000$ ] nor were there any significant interactions involving group.

## Discussion

The current research aimed to explore over- or under-inclusion in categorisation in schizotypy by (a) utilising more sensitive measures of categorisation (i.e. similarity ratings and response times) than in previous studies and (b) utilising borderline exemplars.

### Semantic Categorisation

Kiang and Kutas (2005) used only high and low frequency, and unrelated exemplars, and found no group differences. Morgan et al. (2009) used similar stimuli, taken from similar norms (Battig & Montague, 1969; Hampton & Gardiner, 1983), as well as borderline and related exemplars, and found group differences. The present study found no group differences, and so it remains unclear where the locus of any group difference, if such exists, might lie.

### Borderline Categorisation

In contrast, the borderline categorisation task provides some limited support for the findings by Morgan et al. (2009). Consistent with their findings, high schizotypes provided significantly fewer positive category judgments than the low group – that is high schizotypes show evidence of under-inclusion. It is possible that the semantic categorisation task is too insensitive to reliably reveal group differences, and that the borderlines task, which focuses attention on cases whose categorisation is uncertain, is more sensitive.

That the group difference appears localised to exemplars possessing essential properties suggests that low schizotypes may be more essentialist, making positive category judgments more readily when essential properties are

present. For both groups, the effect of appearance properties was greater for natural kinds than it was for artefact categories, and the effect of essence properties was greater for artefact categories than for natural kinds. At least for these participants and these stimuli, the artefact categories appear to be more strongly essentialised.

Finally, it is interesting that no group differences emerged in response times on the semantic categorisation task. Morgan et al. (2009) reported that they may have revealed group differences as they used shorter SOAs than Kiang and Kutas (2005), thus increasing task demands. It is possible that increasing the demands of the semantic categorisation task would render it more sensitive to group differences.

The data support previous findings that have shown an inverted V-shaped function in categorisation times (Chen et al., 1994; Rips et al., 1973). Again, no group differences were noted, suggesting that group differences in categorisation in the Morgan et al. (2009) study were not due to high schizotypes requiring more time to categorise.

## Explaining Group Differences

As indicated, one possible explanation for the fact that high schizotypes provided fewer positive category judgments than the low group, and fewer positive category judgments for both types of borderline exemplar, is that they weighed less heavily the presence of essential properties. Alternatively, they may have operated with a stricter definition or higher threshold for category membership than low schizotypes, with the categorisation of high schizotypes being more similarity-based.

Essence properties were only present in the [A-E+] borderline exemplars and on this basis these exemplars should have received more positive category membership judgments if participants were essentialist. As noted above, it appears as though artefact categories were more strongly essentialised. It also appears as though the low group were more essentialist, as [A-E+] borderlines were categorised more positively by the low than by the high group.

Of course, as this is one of the few studies to directly examine the relationship between schizotypy and categorisation, these data are unlikely to be decisive as regards extant theories of concepts. Further work is needed to confirm these findings, possibly including meta-analytic studies that offer the prospect of considerably greater statistical power. Nevertheless, it is possible that the categorisation of high and low schizotypes may require somewhat different theories, with those of high schizotypes being more similarity-based, and those of low schizotypes being more essentialist. Although the current study is not able to adjudicate between different theoretical frameworks, it does suggest the promise of further studies of schizotypy and categorisation in helping to do so.

## Acknowledgements

Our thanks go to the many participants who gave their time to take part in this research, and to Mercè Prat-Sala for helpful comments on earlier versions of this research.



## References

- Barrantes-Vidal, N., Fananas, L., Rosa, A., Caparros, B., Dolors Riba, M., & Obiols, J. E. (2003). Neurocognitive, behavioural and neurodevelopmental correlates of schizotypy clusters in adolescents from the general population. *Schizophrenia Research*, 61, 293–302.
- Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph*, 80(3), 1–46.
- Beech, A., McManus, D., Baylis, G., Tipper, S., & Agar, K. (1991). Individual differences in cognitive processes: Towards and explanation of schizophrenic symptomatology. *British Journal of Psychology*, 82(4), 417–426.
- Bentall, R. P., Claridge, G., & Slade, P. D. (1989). The multidimensional nature of schizotypal traits: A factor analytic study with normal subjects. *British Journal of Clinical Psychology*, 28(4), 363–375.
- Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition*, 60, 1–29.
- Braisby, N. R. (2004). Deference and essentialism in the categorization of chemical kinds. In R. Alterman, & D. Kirsch (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 174–179). Lawrence Erlbaum Associates: Mahwah, NJ.
- Chen, E. Y. H., Wilkins, A. J., & McKenna, P. J. (1994). Semantic memory is both impaired and anomalous in schizophrenia. *Psychological Medicine*, 24, 193–202.
- Church, B. A., Krauss, M. S., Lopata, C., Toomey, J. A., Thomeer, M. L., Coutinho, M. V., ... Mercado, E. (2010). Atypical categorisation in children with high-functioning autism spectrum disorder. *Psychonomic Bulletin & Review*, 17(6), 862–868.
- Claridge, G. (1985). *Origins of mental illness*. Oxford, UK: Blackwell.
- Cohen, G., Johnston, R. A., & Plunkett, K. (2000). *Exploring cognition: Damaged brains and neural networks, readings in cognitive neuropsychology and connectionist modelling*. Hove, UK: Psychology Press.
- Doughty, O. J., & Done, D. J. (2009). Is semantic memory impaired in schizophrenia? A systematic review and meta-analysis of 91 studies. *Cognitive Neuropsychiatry*, 14(6), 473–509.
- Doughty, O., Lawrence, V., Al-Mousawi, A., Ashaye, K., & Done, D. (2009). Overinclusive thought and loosening of associations are not unique to schizophrenia and are produced in Alzheimer's dementia. *Cognitive Neuropsychiatry*, 14(3), 149–164.
- Duchene, A., Graves, R. E., & Brugger, P. (1998). Schizotypal thinking and associative processing: A response commonality analysis of verbal fluency. *Journal of Psychiatry and Neuroscience*, 23, 56–60.
- Gouzoulis-Mayfrank, E., Voss, T., Morth, D., Thelen, B., Spitzer, M., & Meinke, U. (2003). Semantic hyperpriming in thought disordered-patients with schizophrenia: State or trait? A longitudinal investigation. *Schizophrenia Research*, 65, 65–73.
- Hampton, J. A., & Gardiner, M. M. (1983). Measures of internal category structure: A correlational analysis of normative data. *British Journal of Psychology*, 74, 491–516.
- Kiang, M., & Kutas, M. (2005). Association of schizotypy with semantic processing differences: An event-related brain potential study. *Schizophrenia Research*, 77, 329–342.
- Lawrence, V. A., Doughty, O., Al-Mousawi, A., Clegg, F., & Done, D. J. (2007). Do overinclusion and distorted semantic category boundaries in schizophrenia arise from executive dysfunction? *Schizophrenia Research*, 94(1), 172–179.
- Lenzenweger, M. F., Cornblatt, B. A., & Putnick, M. (1991). Schizotypy and sustained attention. *Journal of Abnormal Psychology*, 100, 84–89.
- Mason, O., Claridge, G., & Jackson, M. (1995). New scales for the assessment of schizotypy. *Personality Individual Differences*, 18, 7–13.
- Morgan, C. J. A., Bedford, N., O' Reagan, A., & Rossell, S. L. (2009). Is semantic processing impaired in individuals with high schizotypy? *The Journal of Nervous and Mental Disease*, 197, 232–238.
- Morgan, C. J. A., Bedford, N., & Rossell, S. L. (2006). Evidence of semantic disorganisation using semantic priming in individuals with high schizotypy. *Schizophrenia Research*, 84, 272–280.
- Peters, E. R., Pickering, A. D., & Hemsley, D. R. (1994). "Cognitive inhibition" and positive symptomatology in schizotypy. *British Journal of Clinical Psychology*, 33, 33–48.
- Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and analogical reasoning*. New York, NY: Cambridge University Press.
- Rips, L., Shoben, E., & Smith, E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, 12, 1–20.
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh, PA: Psychology Software Tools.
- Siever, L. J., Kalus, O. F., & Keefe, R. S. (1993). The boundaries of schizophrenia. *Psychiatric Clinics of North America*, 16(2), 217–244.
- Strnad, L., Anzellotti, S., & Caramazza, A. (2011). Formal models of categorisation: Insights from cognitive neuroscience. In E. M. Pothos, & A. J. Wills (Eds.), *Formal approaches in categorisation*. Cambridge, UK: Cambridge University Press.
- Suhr, J. A. (1997). Executive functioning deficits in hypothetically psychosis-prone college students. *Schizophrenia Research*, 27, 29–35.
- Tallent, K. A., & Gooding, D. C. (1999). Working memory and Wisconsin Card Sorting Test performance in schizotypic individuals: A replication and extension. *Psychiatry Research*, 89, 161–170.

# Inferring Metaphoric Structure from Financial Articles Using Bayesian Sparse Models

**Martin Sälzle (saelzle\_martin@ceu-budapest.edu)**

Department of Cognitive Science, Central European University  
Frankel Leó útca 30-34, Budapest, 1023, Hungary

**Mark T. Keane (mark.keane@ucd.ie)**

School of Computer Science & Informatics, University College Dublin  
Belfield, Dublin 4, Ireland

## Abstract

Drawing from a large corpus (17,000+ articles) of financial news, we perform a Bayesian sparse model analysis of the argument-distributions of the UP and DOWN-verbs, used to describe movements in indices, stocks and shares. Previous work, by Gerow and Keane (2011a, 2011b, 2011c), has shown, using measures of overlap and k-means clustering, that metaphor hierarchies and antonymic relations can be found in this data; for instance, UP verbs have *rise* as a superordinate organizing a distinct set of subordinate verbs (*soar, jump, climb, surge, rebound, advance*). This work empirically realizes theories about the structuring of our conceptual systems with metaphors (Lakoff, 1992; Lakoff & Johnson, 1980) but does so using a distributional approach to meaning; namely, that words that occur in similar contexts have similar meanings (see Wittgenstein, 1953). However, Gerow and Keane's analysis does not show the overall structure of how these metaphors semantically relate to one another. In the present paper, we re-analyzed their data using a Bayesian sparse model (Lake & Tenenbaum, 2010) in order to infer this metaphor space as a uniform representation, based on the argument distributions. Therefore, we treated arguments as features of metaphors. Our model learned three dimensional graphs in an unsupervised manner as sparse representations of the metaphoric structure over all argument distributions, in parallel. Doing so, it also successfully indicates the metaphoric hierarchies and antonymy relations, that were found by the previous models. In conclusion, we discuss the benefits of this approach.

**Keywords:** Argument features; analogy; Bayesian inference; emergent structure; corpus analysis; metaphor hierarchies; semantic cognition; similarity; sparse representation; spatial metaphor; structure discovery; unsupervised learning.

## Introduction

In recent years, significant progress has been made in deriving meaning from statistical analyses of distributions of words (e.g., Gerow & Keane, 2011a; Landauer & Dumais, 1997; Turney, 2006; Turney & Pantel, 2010; Michel et al., 2010). This distributional approach to meaning takes the view that words that occur in similar contexts tend to have similar meanings (see Wittgenstein, 1953) and that by analyzing word usage we can recover meaning. For instance, Michel et al., (2010) argue that significant insights into human culture and behavior can be derived from analyzing very large corpora, such as the GoogleBooks repository.

Gerow and Keane (2011a-c; henceforth abbreviated as G&K) took such a distributional approach to understanding metaphorically-structured knowledge (in hierarchies and antonymic relationships) between "UP" and "DOWN" verbs from a corpus of financial news reports. Lakoff and Johnson (1980) have argued that metaphors are used to structure many domains of human experience and also many abstract conceptual domains (e.g., emotions). They specifically identified the use of the UP-DOWN metaphor opposition in accounts of wealth (e.g., WEALTH-IS-UP as in *high class*) and in the *rise* and *fall* of numbers (e.g., MORE-IS-UP; LESS-IS-DOWN).

G&K (2011a) build a corpus of 17,000+ financial articles covering a 4-year period, about the major world stock indices (Dow Jones, NIKKEI, FTSE-100) from the *Financial Times*, *NY Times* and *BBC* websites; the corpus contained over 10M words. After parsing the corpus, G&K selected all the sentential instances of the most commonly occurring UP and DOWN verbs (see G&K, 2011a, 2011b for details). Table 1 shows some of the most commonly used arguments found in the corpus, indicating the metaphoric usage of the selected verbs. G&K then analyzed the clustering in these distributions (using k-means clustering) and the overlaps between the distributions of different verbs (using the % overlap in each pair-wise comparison of verb arguments). This analysis threw up some striking regularities.

Table 1: The percentage of *rise*'s argument distribution covered each of the 10 most frequent arguments.

Rank	Argument Word	% of Corpus
1	Index	7.3
2	Share	5.6
3	Point	4.8
4	Percent	2.9
5	Price	2.4
6	Stock	2.0
7	Yield	1.9
8	Cent	1.3
9	Profit	0.9
10	Rate	0.9

## Metaphor Hierarchies

G&K (2011b) argued that if one verb-metaphor (e.g., that referred to by *rise*) was organizing another metaphoric verb (e.g., *soar*) then the argument distribution of the former should largely cover the latter, but the opposite would not be the case. They also argued that verb-metaphors at the same level of generality (e.g., a basic level), *sibling metaphors*, would have symmetrically overlapping argument distributions. Their coverage- and cluster analysis confirmed these types of structure. On coverage, they found that *rise* organized a group of metaphoric siblings (*soar, jump, climb, surge, rebound, advance*), set off from a set of other more outlying verb-metaphors (*increase, rally, recover, gain, alleviate, elevate*; see Figure 1, A). In the clustering, they found that *rise* was quite separate from all the other verbs that clustered together and that *gain* and *climb* were quite distinct (see Table 2). A similar pattern was found for *fall* and its subordinate-related verb-metaphors (*dip, retreat, sink, plunge, tumble, slide slip, slide, ease, drop, plummet*), with *decline, lose, decrease* and *worsen* being outliers (see Figure 1, B, and G&K, 2011b, for detailed results).

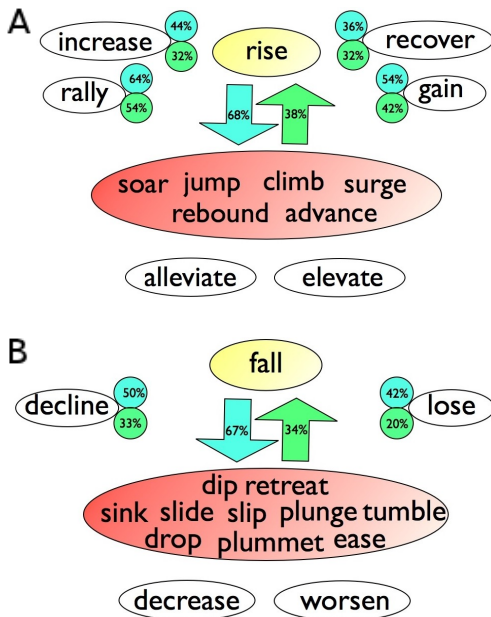


Figure 1: Argument coverage of (A) main UP-verbs and (B) main DOWN-verbs from G&K (2011b).

## Metaphoric Antonyms

G&K (2011c) also analyzed verb distributions for antonymic relations; arguing that preferred antonyms *rise-fall* should have more similar distributions than less preferred antonyms *rise-decrease* or *rise-lower*. G&K performed a psychological experiment to find the preferred antonyms between the UP and DOWN verbs and then formulated several different similarity measures (Euclidean

distance, cosine similarity, K-L divergence) on the argument distributions to determine which one best predicted the human choices. Given a set of 13 UP verbs and 15 DOWN verbs (as possible antonyms) people identified 114 unique antonym pairs. Of these, in 60% of cases, the cosine similarity of the argument distributions of pairs correctly identified the most preferred antonym-pair from the human ratings. This figure rose to 87% if we consider identifying the 1<sup>st</sup> or 2<sup>nd</sup> most preferred pairs (see G&K, 2011c for details). Table 3 lists some results of the human antonymy ratings.

Table 2: Top 5 clusters in k-means analysis of UP-verbs (\* rest = the remaining verbs).

Rank	Cluster Groups	% of Tot. (Freq.)
1	rise, rest*	62% (1451)
2	rise, gain, rest*	18% (702)
3	rise, [climb, gain], rest*	4% (36)
4	rise, [jump, climb, gain], rest*	3% (27)
5	all-verbs-as-one-group	2% (18)

Table 3: Some examples of people's verb antonymy ratings, conducted by G&K (2011c). Percentage measures indicate mean antonymy ratings over participants and sub-tasks (free generation and match the opposite).

Verb pair	Antonymy
rise-fall	57%
jump-fall	31%
drop-climb	13%
decline-rise	27%
slide-climb	23%
soar-plummet	17%

## Using Sparse Models Instead

G&K found a number of interesting regularities for hierarchical and antonymic relationships between the argument distributions of UP and DOWN verbs. However, their results were based on different approaches, rather than a unifying model, and do not indicate the semantic structure of the metaphoric corpus as a whole. Arguably, it is essential to understand the cognitive semantics of the corpus, as the meaning of individual concepts must depend on how they relate to one another (Kemp & Tenenbaum, 2008). Bayesian sparse models, also known as sparse graph codes (MacKay, 2003), appear to be good candidates for this task.

Bayesian sparse models basically infer an emergent structure in a probabilistic framework (Rogers & McClelland, 2004). Applied to semantics, they have been shown to perform particularly well at finding regularities for the clustering of features for very large numbers of words from different conceptual domains (Lake & Tenenbaum,

2010). These models assume that people learn a set of parameters that fit their observed data well.

Sparse models may be better at handling metaphoric structure than other structured probabilistic models for semantic cognition (e.g., Kemp & Tenenbaum, 2008). The latter generate structures as instances of forms and discover the structural instance of the form that best explains the underlying dataset; including, structural instances based on the graph grammar of trees, linear orders, multidimensional spaces, rings, dominance hierarchies, cliques, and other forms that are supposed to be the organizing principles for data of different cognitive domains. In this way, these models account for domain-specific inferences. The learned structures can then be used to model human inductive reasoning about novel properties of objects within those domains (Kemp & Tenenbaum, 2009).

However, considering a dataset of metaphors, we need to take into account that contemporary cognitive linguistics understands conceptual metaphor not as domain-specific inference but rather as mappings from one conceptual domain to another (Lakoff, 1987; Gibbs, 1994, 1996; Fauconnier & Turner, 1998, 2003). For example, mapping the directionality of movement to changes in quantity (e.g., “prices are rising”). A cognitive model of metaphoric structure would, therefore, not necessarily need to select between structural instances of domain specific forms. Since a metaphoric corpus is likely to consist of many mappings of many different conceptual domains, it would rather need to infer an emergent structure on the basis of a psychologically justified prior probability over the hypothesis space of possible structures. We think that sparseness would be a useful prior for such a model, as it accounts for the cognitive parsimony that is needed to mentally structure metaphors over the vast array of semantic domains (Lakoff, 1992); as well as for the trade off between cognitive effect and computational effort (Wilson & Carston, 2006).

### Sparse Model Analysis of Verb-Metaphors

Lake and Tenenbaum’s (2010) Bayesian sparse model was used on G&K’s verb-metaphor corpus, involving UP and DOWN verbs, extracted from the larger finance corpus (see Data Set). This metaphor corpus contained 9,700+ distinct sentence instances for these two sets of verbs. The sparse model should be able to learn and graph the structure of these verb-metaphors by determining how they covary with regard to the frequency of their argument features. Graphically, the verb-metaphors are represented as nodes in a weighted graph, where the strength of the link between two object-nodes is related to the weighted covariation of their features. The weights of the graph, denoted as the symmetric matrix  $W$ , are learned from data by optimizing an objective function that trades off the fit to the data with the sparsity of the graph. In the present study, the sparse model technique was used to build three different graphs: a graph for the UP verb-metaphors and their arguments (using a  $13 \times 386$  matrix), a graph for the DOWN verb-metaphors and their arguments (using a  $15 \times 456$  matrix), and a graph of

the combined set of UP and DOWN verb-metaphors (using a  $28 \times 605$  matrix).<sup>1</sup>

### Method

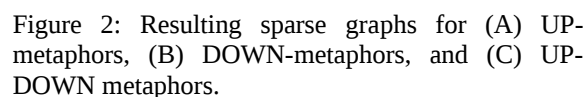
**Data Set** A total of 28 verbs were used, 13 UP-verbs with 386 distinct, unique arguments, 15-DOWN verbs with 456 distinct, unique arguments (based on those used by G&K, 2011b-c). There were 9,721 distinct sentence instances in the corpus (5803 sentences with UP verbs, 3918 sentences with DOWN verbs).

**Model Setup** The code for the model we used was written in MATLAB (provided by Brenden Lake, Department of Brain and Cognitive Sciences, MIT; see Lake and Tenenbaum, 2010, for a detailed description of the model). Formally, the undirected graph  $W$  defines a multivariate Gaussian distribution  $p(f^{(k)}|W)$  in the generative model, known as a Gaussian Markov Random Field (GMRF), where the  $n$  objects are the  $n$ -dimensions of the Gaussian. With a prior distribution on sparsity, the model then estimates the maximum a posteriori (MAP) parameters  $W$  as optimal structure based on data. Each data set  $D$  was cast in a  $n \times m$  matrix with  $n$  metaphors and  $m$  arguments. Therefore, the columns of  $D$ , denoted as arguments  $\{f^{(1)}, \dots, f^{(m)}\}$ , were assumed to be independent and identically distributed drawn from  $p(f^{(k)}|W)$ . With the  $n$ -dimensional Gaussian distribution, it is assumed that arguments vary smoothly over the graph. So, if two metaphors  $i$  and  $j$  happen to be connected by a large weight ( $w_{ij}$ ), they share similar application frequencies over arguments. As a result of sparsity, most metaphors are not directly connected in the learned graph (i.e.,  $w_{ij}=0$ ). The resulting weights allowed us to further apply a Markov Cluster Algorithm (MCA) to classify verb metaphors based on the covariation of their argument distributions. Inflation and pre-inflation settings for the MCA were held on standard (see Freeman et al., 2007; Theodoridis, Van Dongen, Enright, & Freeman, 2009).

### Results & Discussion

Figure 2 shows the resulting weight matrices illustrated as sparse graphs learned for the three different datasets: (A) UP verbs, (B) DOWN verbs and (C) UP-DOWN verbs combined (all graphs were drawn with BioLayout Express 3D; videos of rotating versions of respective graphs should be retrievable by clicking on them). In each graph, the labeled nodes represent verb-metaphors (e.g., *rise*, *fall*). The links show the connection weights and consequential distances between the nodes, denoting similarity over all

<sup>1</sup> Resulting weight matrices are available from the authors.



**Metaphoric Structure of UP Verbs** Figure 2 (A) shows the sparse graph for the UP verb-metaphors. Overall, it literally provides a much better picture of the semantic space of the metaphors with the relative distances between each clearly shown, compared to G&K’s (2011b, 2011c) analyses. First, note that the *rise* node stands out as being distinct and non-similar to most of the other nodes. Counter-intuitively, this occurs because though *rise* has arguments that cover many of the arguments of most other verbs *combined* (also see Figure 1, A), it has fewer arguments in common *individually* with any given verb (and, therefore, low similarity with each). Second, *rise*, *climb* and *gain* cluster separate to the remaining verb-metaphors (purple vs. green nodes). While we know that *rise* has asymmetric coverage regarding most other verbs, *climb* and *gain* have not (also see Figure 1, A). Therefore, the latter are two highly interconnected outliers.

**Metaphoric Verb Antonyms** Figure 2 (C) shows the sparse graphs for the combined UP and DOWN verb corpora. These graphs are slightly different because they deal with both categories of verbs. G&K's (2011c) analysis for antonymy worked on the basis that the key antonyms would be highly similar.



relative to other pairings across the two sets of verbs. Again, the sparse model shows this very clearly as notable key antonym pairs appear as close nodes: for instance, *rise-fall*, *gain-drop*, *climb-slip*, *gain-lose*. Further, how the verb-metaphors cluster (shown by node coloration) indicates semantic similarity in how they got applied. However, the antonymy ratings from the human subject experiment of G&K (2011c) correlate just weakly with the ones from the model (Pearson's  $r=0.4$ ; see Figure 3). This might have experimental- and model related reasons: first, the verb-metaphors from the corpus were applied by human speakers to describe financial changes. The experimental data, however, are abstract antonymy ratings of verbs, having neither applicational relation to the domains relevant to conceptualize finance, nor to any other cognitive domain. (Future experiments for metaphoric antonymy would need to take this into account.) Second, the model's antonymy ratings for the superordinates *rise* and *fall* had to be excluded, since they were 0 to all other metaphors, except to one another. Finally, in the graph are also some high-similar pairings within the same verb set, like *rally-rebound* and *slip-ease*, that are clearly “just similar” and not antonyms. The latter indicates that some prior categorization of what-are-known-to-be broadly opposite sets is required before such a merged model might be useful. Again, an additional coverage analysis is needed to isolate *rise* and *fall* as superordinates (see also Figure 1, A and B).

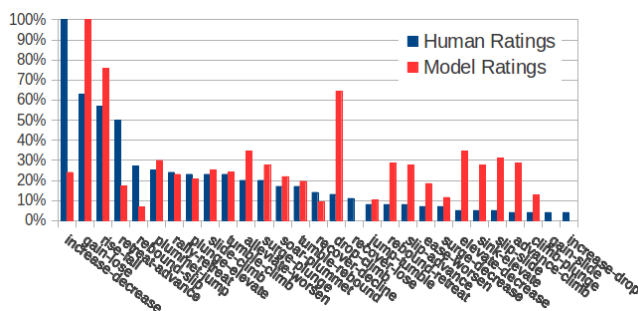


Figure 3: Model- versus human antonymy rating of verb-pairs in per cent. Human ratings reflect perceived antonymy of verbs (see G&K, 2011c); whereas model ratings reflect the computed antonymy of verb-metaphors used to describe financial changes. The latter are weighted entries of the model's weight matrix.

## Conclusion

The suggestion that significant parts of our conceptual systems are structured by metaphors has mainly received support from linguistic and anthropological analyses (see Lakoff & Johnson, 1980). However, cashing out these ideas

empirically in a systematic way has proven difficult. The promise of the present work is that these ideas can be empirically supported by a distributional analysis of verb arguments, with such metaphoric import. We have shown that sparse models can provide a rich and informative basis for relating these verb-metaphors together in a uniform metaphor space. We believe that this approach may be useful in modeling other cognitive tasks that rely on these metaphoric spaces (e.g., language comprehension, analogical thinking). For instance, in analogical thinking it has long been argued that conceptual slippage (Hofstadter, 1995) and re-description (Keane, 1996; Kurtz, 2006) are needed to account for human abilities: Bayesian sparse models provide a basis for allowing such slippage, assuming structural support for the slippage being considered.

However, our work has also indicated that the sparse models will still need a coverage analysis to isolate superordinate metaphors. And, because these are important for conceptually structuring the metaphor space, they should be implemented in the way sparse models generate and learn structure. This might be achievable by using hierarchical Bayesian sparse models (Chandrasekaran, Parrilo, & Willsky, 2010) that potentially discover organizing metaphoric concepts as hidden or latent variables, and further increase sparsity.

## Acknowledgments

This work was carried out as part of a self-funded MSc in Cognitive Science at UCD by the first author. We want to thank Brenden Lake for providing the code for the sparse model, Aaron Gerow for all the data, Anne Tamm for linguistic advice, and Máté Lengyel for technical suggestions.

## References

- Chandrasekaran, V., Parrilo, P. A., & Willsky, A. S. (2010). Latent variable graphical model selection via convex optimization. In *The 48th Annual Allerton Conference on Communication, Control, and Computing*, (pp. 1610–1613). IEEE.
- Fauconnier, G., & Turner, M. (1998). Conceptual integration networks. *Cognitive science*, 22(2), 133–187.
- Fauconnier, G., & Turner, M. (2003). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.
- Freeman, T. C., Goldovsky, L., Brosch, M., Van Dongen, S., Mazière, P.,...Enright, A. J. (2007). Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Computational Biology*, 3(10).
- Gerow, A., & Keane, M. T. (2011a). Mining the Web for the "Voice of the Herd" to track stock market bubbles. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence: Barcelona, Spain*.
- Gerow, A., & Keane, M. T. (2011b). Identifying metaphor hierarchies in a corpus analysis of finance articles. In

- Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*: Boston, MA.
- Gerow, A., & Keane, M.T. (2011c). Identifying metaphoric antonyms in a corpus analysis of finance articles. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*: Boston, MA.
- Gibbs, R.W. (1994). *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press.
- Gibbs, R. W. (1996). Why many concepts are metaphorical. *Cognition*, 61, 309–319.
- Hofstadter, D. R. (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. Together with the Fluid Analogies Research Group. NY: Basic Books.
- Keane, M. T. (1996). On adaptation in analogy. *Quarterly Journal of Experimental Psychology*, 49A, 1062–1085.
- Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31), 10687–10692.
- Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, 116(1), 20–58.
- Kurtz, K. J. (2005). Re-representation in comparison: Building an empirical case. *Journal of Experimental & Theoretical Artificial Intelligence*, 17(4), 447–459.
- Lake, B., & Tenenbaum, J. (2010). Discovering structure by learning sparse graph. In *Proceedings of the 33rd Annual Cognitive Science Conference*: Boston, MA.
- Lakoff, G. (1987). *Women, fire, and dangerous things. What/How Categories Reveal About the Mind*. Chicago: Chicago UP.
- Lakoff, G. (1992). *The contemporary theory of metaphor*. In A. Ortony (Ed.), *Metaphor and Thought* 2nd Edition. Cambridge University Press.
- Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), 211.
- MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.
- Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, ... Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176.
- Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.
- Theocharidis, A., Van Dongen, S., Enright, A. J., & Freeman, T. C. (2009). Network visualization and analysis of gene expression data using BioLayout Express3D. *Nature protocols*, 4(10), 1535–1550.
- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7), 309–318.
- Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3), 379–416.
- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1), 141–188.
- Wilson, D., & Carston, R. (2006). Metaphor, relevance and the 'emergent property' issue. *Mind & Language*, 21(3), 404–433.
- Wittgenstein, L., & Anscombe, G. E. M. (1953/2001). *Philosophical investigations: the German text, with a revised English translation*. Wiley-Blackwell.



# Conceptual Change through Socially Constructive Interaction in the Classroom

**Moegi Saito (saitomoegi@coref.u-tokyo.ac.jp)**

Consortium for Renovating Education of the Future, 7-3-1 Hongo, Bunkyo-ku,  
Tokyo, 113-0033 JAPAN

**Naomi Miyake (nmiyake@p.u-tokyo.ac.jp)**

Consortium for Renovating Education of the Future, 7-3-1 Hongo, Bunkyo-ku,  
Tokyo, 113-0033 JAPAN

## Abstract

While collaborative learning is common in schools, most research has focused on small group collaborative processes, in one shot practice. In order to investigate the nature and the mechanism of collaborative learning in a larger group, this paper presents analysis of a classroom discussion in which 21 third-graders (8- to 9-year-old students) collaboratively discussed predictions about the results of a series of experiments as a whole class over 12 course hours and became able to grasp a rudimentary scientific concept of atomic theory.

We analyzed students' levels of achievement, conversation patterns, selection sequences of predictions of the experimental results, and the contents of their utterances of two students who were most active. The results revealed that all the children succeeded in expressing their grasp of rudimentary atomic theory, yet their routes to this achievement were individualistic and diverse. A preliminary qualitative analysis of two children's utterances shows that while their models were similar during the first half of the course, their differences became more explicit toward the end, which resulted in intense discussion between the two. The diversity observed in the entire course of this class and these two children's explicit, focused dialogue could have contributed to the successful conceptual change for all class members.

**Keywords:** Collaborative conceptual change, HEI class, Socially constructive interaction,

## Introduction

While the amount of research on collaborative learning has increased sharply in the last 10 to 15 years, many studies deal with small groups of learners in mostly one-time practice. Studies of entire-class discussions, which are the norm in regular schools, are still few.

In regular elementary to middle school classes in Japan, it is quite common for 10 to 20 students to discuss their ideas about textbook descriptions and experiment demonstrations, under the teacher's guidance. In science classes, this is a common practice aimed at helping the students change their concepts from rules of thumb to more scientifically rigorous ones. Understanding what actually happens in these practices should be beneficial for developing practically feasible collaborative classrooms.

In the similar vein, in STEM education engaging the learners in demonstrations and experiments has been considered effective. In real classrooms, these activities

often take place in a series of course hours rather than in one time practice. Understanding how these cumulative activities may affect children's conceptual change would also be informative regarding the quality of such classes.

In order to address these needs, this paper presents analysis of the learning processes observed in a well-designed STEM class. In the class, 21 third-graders collaboratively discussed predictions of the results of a series of experiments and were able to grasp a rudimentary scientific concept of atomic theory. The children were encouraged to discuss alternative predictions of carefully ordered experiments as a whole class in order to realize that water cannot enter "where there is air and vice versa" with justifications ranging from heavily relying on their daily experiences to abstracting theory-like reasons based on accumulated observations of the results of previous experiments.

For this class, we tried to answer the following questions.

1. How well did the 21 children understand the concept at the end of the course?

2. What was the trajectory of understanding of the whole class qualitatively? More concretely, in terms of choices of the experiment results and the number of similar explanations by the children, did the class converge toward the "same" answer (as possibly expected by many teachers as well as some convergence-oriented theoreticians), or was divergence among the children maintained? Divergence here means that each individual child creates his or her own explanations of understanding, even at the very end of the course, with some different degrees of understanding underlying their seemingly or apparent "same successful" understanding (e.g., "everybody answers correctly"). It is important to make this distinction so that the teacher can create a class atmosphere that focuses on either divergence or convergence. In our own preliminary study, orientation toward more diversity has been identified as having more potential for successful collaboration.

3. What was the qualitative nature of the children's process of conceptual change, if it happened? In order to gain some insight, we analyzed the utterances of the two most active participants. We analyzed how they changed their expressions of their models as the course developed.

## Research context:

### Hypothesis-Experiment-Instruction

Our research context here is the series of science classes designed using the Hypothesis-Experiment Instruction (HEI) framework (Itakura, 1963) and the target content of the “Air and Water” unit (Itakura, 1970). The objective of this unit is to understand that water cannot enter where there is air and vice versa, to serve as the basis of a rudimentary grasp of atomic theory. HEI is a strategy to teach basic scientific concepts. An HEI “unit” consists of multiple such “problems,” or experiments, carefully ordered to guide the development of scientific concepts underlying the problem set. The teacher uses a problem set sheet for each experiment; the sheet explains the experiment and alternatives of possible answers. Each student chooses one alternative, and the result of their selections is written on the blackboard so that the students know the distribution. They then are encouraged to give reasons for their choices, to question others, and to discuss among themselves in order to make a better prediction. They are allowed to change their prediction before the experiment that will confirm the correct answer. At the end of the class, each student writes comments about the activities. By repeating this activity to cover the entire set of problems in a unit, each student in an HEI class is expected to integrate the results of the experiments in her/his own way in order to formulate an individualized “hypothesis,” or the student’s rudimentary scientific concept.

The “Air and Water” unit consists of the 11 problems explained in Table 1. These problems can be classified into two subsets, Problem 1 (P1) through Problem 6 (P6) and the rest. The first set deals with problems whose answers are justifiable with daily experiences. In contrast, situations of P7 to P10 do not occur often in children’s daily lives; thus, they are difficult for children to imagine.

Table 1: Wordings of the 11 problems in the “Air and Water” HEI unit

P1	When an empty glass is pushed into water upside down, will the water come into the glass?
P2	If you place a crumpled piece of paper in the glass and do the same as in Problem 1, will the paper get wet?
P3	An upside-down glass with water inside is in the water. When you lift it up through the surface of the water, what will happen to the water in the glass?
P4	What will happen when you suck air through a straw from an upside-down glass in the water?
P5	Which dropper sucks more water, one whose tip is deep in the water or one whose tip is shallow?
P6	Can water be sucked through a 1m straw?
P7	A can of juice has just one hole on top. When the can is turned upside down, will some juice come out of it?
P8	Will some juice come out of a can that has two holes on its top and is turned upside down?
P9	Suppose you put the can used in problem 8 deep into the water, keeping your finger tight on one of the holes. Will some water go into the can?
P10	What will happen to the can in problem 9 if you remove your finger?

P11	Will some soy sauce come out of its container if you put your finger onto the hole on its top?
-----	--

The latter problems require learners to rely on their newly formed “hypotheses,” from accumulating predictions and observations of the experimental results in the previous problems. The learners are expected to realize that water cannot enter where there is air and vice versa, with justifications starting from relying on their daily experiences to abstracting theory-like reasons. The last problem, P11, can be answered by relying on either daily experiences or newly learned understanding, or both. This is to confirm their achieved levels as well as to let the children connect the hypotheses to daily life, so that they may see that their clearer understanding is usable in everyday situations.

The targeted concept of this unit is “water cannot enter where there is air and vice versa,” according to Itakura (1970), the developer of this curriculum. From the perspective of modern science, we should use such concepts as “atmospheric pressure” and “surface tension” to fully explain the results of the P7 through P10 experiments; however, the learners were not expected to understand these concepts in this unit. Instead, the emphasis was on having the children experience “how to think scientifically.” The idea of “water cannot enter where there is air and vice versa” itself is not sufficient for high school education, but it is adequate to give a general justification to cover all of the 11 experiments. Thus, becoming able to predict the results of an experiment and to justify the prediction requires a change in concept, tying experience-based rules of thumb to more scientifically justifiable explanations.

In order to identify the levels of conceptual change discussed here, we use the four-level model that ties the children’s daily experiences and the scientific concepts (Miyake, 2009) listed in Table 2. The learner can create a rule of thumb based on one incidence of experience (e.g., coming too close to a heated stove lets the child create a useful rule of thumb of “red, warm, could be extremely hot, to be avoided”). Usually these experiences should be repeated numerous times so that the child can create a more stable rule of thumb and understand the world around her/him. These are individually created, experience-based “concept” levels of Level 1 (based on one incident) and 2 (repeated and summarized). The other side of the model, Level 4, includes the consensus reached by the scientific community, the state-of-the-art concepts shared by professionals. These are the concepts explained in textbooks and expected to be taught at school. In between, at Level 3, is a wide zone for learners needing to change their Level 2 basic experience-based rules of thumb to Level 4 scientifically community-shared concepts. Many instructional methods have been developed and tested to facilitate this conceptual change (Vosniadou, 2008). Some of these methods heavily utilize the power of social construction in the form of collaborative design (c.f., Roschelle, 1992; Howe, et.al. 2005; Vosniadou, et.al.2007; Miyake, 2008). HEI is such a method (Saito & Miyake, 2011).

According to this model, the targeted conceptual change for the “Air and Water” unit is to reach Level 3, starting from Level 2.

Table2 : Four-stage model of conceptual change

Theories constructed through collaboration	Lv.4	Scientific concept, created and shared in the scientific community
	Lv.3	A socially constructed, yet individually understandable “story” tying abstracted ideas created on their own as well as borrowed from others and the the rules-of-thumb accumulated in daily life.
Knowledge and rules from individual experience	Lv.2	A rule of thumb created by accumulating one’s own (yet many) experiences from different situations
	Lv.1	A rule -of- thumb based on one incidence

## Children’s conceptual change

### Data

The data come from the “Air and Water” unit in an HEI class conducted in May and June 2002. Twenty-one third graders participated in 12 lessons taught by a highly experienced HEI teacher, Yuko Saito, who voluntarily kept records of the distributions of students’ predictions before and after the class discussions, the students’ discussions using hand-written notes, and voice recordings. She also kept copied records of the notes taken by all the students during class. The transcribed voice recordings and the copied notes are the data we analyzed here.

### Predictions of answers to the problems

Fig. 1 plots the percentage of correct predictions made by the children upon reading the explanation of the experiment, prior to discussion.

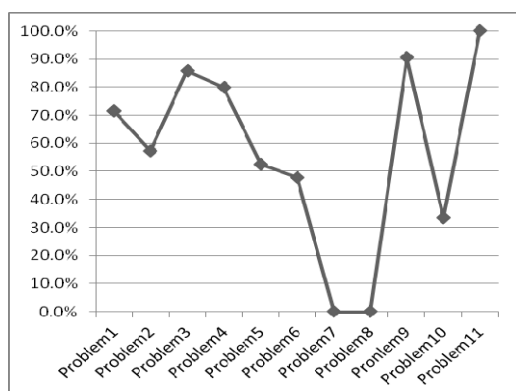


Figure 1: The percentage of correct predictions.

All the children made correct predictions for the last experiment, whose result was not readily obvious from daily

experiences. It could be assumed that the entire class somehow successfully formed a rudimentary concept of physical identity, at least with regard to whether or not air and water could share the same physical space.

However, this conceptual shift did not occur smoothly. For P7 and P8, the problem situation significantly deviated from the children’s daily experiences, thus making it particularly difficult for them to make correct predictions. After their complete failure on these problems, the children somehow recovered their performance over P9 and P10, and were able to correctly predict 100% on the transfer problem of P11.

### Shift of levels of the children’s concepts

Because the shift pattern of predictions indicated successful learning at least at the end, we could expect to observe a corresponding shift of concept levels in the children’s utterances and the note descriptions. We coded the contents of the students’ discussions during the classes and the written comments after the classes according to the levels described in Table2. The operational definition used for this coding and the corresponding example of each level are presented in Table3. The correspondence rate of coding by two coders was 94%.

Table3: Categorization for level of conceptual change in “Air and Water”

	Operational definition	Examples
Lv.3	Explanation based on understanding of “water cannot enter where there is air and vice versa, “which could lead the learners to correctly answer one or more problems.	“The seal stops the air. When the air can’t move, the water won’t move, either” for P11
Lv.2	Explanation based on generalized rules of thumb	“When the can has two holes, water can move” for P10
Lv.1	Explicit reference to a particular experience	“I have tried such an experiment with a wash bowl in the bath.” for P1

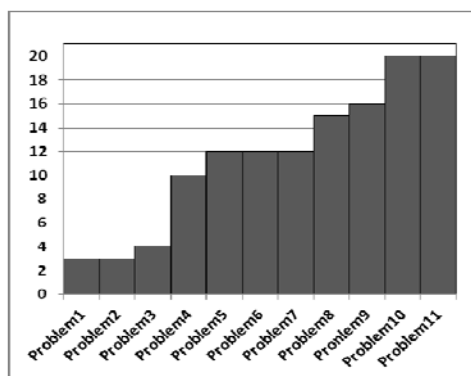


Figure 2: The accumulated number of Lv3 expressions

If the class successfully changed their concept, each student could express Lv 3 understanding at least once. Figure 2 plots the accumulated number of students whose justification fell at Lv 3 for each problem.

Twenty of the 21 children's expressions of reasons for their choices were coded as Lv 3 explanations at the end. Thus, we could conclude that the class was successful in helping each child construct a starting scientific model about air and water.

## Quantitative analysis of the children's collaborative learning paths

### Shift pattern of predictions of each student

What was the qualitative trajectory of this successful conceptual shift in the whole class? In terms of choices of alternatives of the experiment results and the similarities of explanations given by the children, did the class converge to the "same" answer, or did the individual students diverge?

Each student could choose one of three possible predictions for each problem. This allowed many possible paths of choices, but for 7 of the 11 problems, more than half of the students chose the same prediction (Fig. 1). If students had chosen the same prediction because they understood in a similar manner, there could not have been so many patterns of choices. When we asked teachers who often teach HEI classes how many courses there could be for an 11-problem unit, they typically said there would be only three or four paths in one class of 20 to 30 students. For analysis, we checked the alternative that each child chose for each problem. We then compared the sequence of such predictions for each child against that of the other children. Every child had a unique sequence of alternatives for the 11 problems; thus, we concluded that there were 21 paths of selections. For P1, P5, and P10, student's selections were evenly distributed among the three alternatives. Thus, they did not tend to converge to fewer selections toward the end of the unit. This diversity could have made the interaction among students constructive, promoting conceptual change in the class, as confirmed above.

### Similarity of models used in discussions

As for the similarity of explanations given by the children, did the class converge to the "same" answer, or did it diverge among them? A simple prediction could indicate convergence, as some previous research indicates (c.f., Roschelle), however, a detailed analysis of individual paths of understanding indicates the opposite, that they diverge as the discussion progresses (c.f., Miyake, 1986). In order to test this idea, we counted each child's utterances for each problem, as an indicator of convergence of their models. When their models converge, the learner who sufficiently explains one problem would not be motivated to repeat it for another, while a new member who comes to a definite understanding based on the model might wish to express it for a different problem. This would give us a relatively even

number of utterances among the children, as well as a relatively smooth decline of the number of utterances from the beginning problems to the ending ones.

Fig. 3 plots the number of utterances for each problem. We do not observe any clear decline of utterance frequencies across the problems. The children seem to have been motivated to speak up for some problems and not for others, even in the middle of the course, though a clear decrease is observed toward the end.

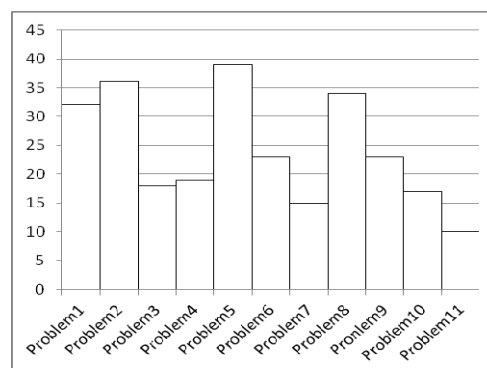


Figure 3: Number of utterances during discussion for each problem in the "Air and water" unit

When we counted the number of utterances for each child, we found a clear pattern of some kids talk more while some don't. The pattern may mean that the children could learn a lot while just being silent, attentively listening the others' talks and discussions (Saito&Miyake,2011).

## Qualitative aspects of the process of collaborative conceptual change among the core members

### Shift pattern of predictions

In order to investigate the qualitative aspects of the conceptual change process in this class, we selected the two most active children (A and B), and analyze the contents of their utterances. The purpose of this analysis is to understand the results described above. Quantitative analysis of the pattern indicates more diversity than normally expected by both teachers and researchers who believe in a convergent pattern of collaborative conceptual change. The qualitative aspects of the processes imply value in this diversity. We would like to know whether the most active two worked as a convergent target model for final success, or if their divergence contributed to their success. Were the two similar to each other in making predictions by choosing similar alternatives, sharing similar models for justification at the beginning, and gradually converging and uttering their understanding to compensate each other so that the rest of the class could use their model as a target and "learn" after them? Or were they different until the end to stimulate the thinking of the rest of the

class? For analysis, we examined their prediction paths from their chosen alternative for each problem, and inferred the cognitive models they could have adopted, based on their utterances.

First, we compared the prediction paths of A and B, and then inferred their models based on what they said and how they said it (Table 4).

Table 4 : Predictions and justifications of A and B for each problem

	A	B
P1	○ I tried this yesterday in the bath. There was no water at the top of the turned-over bucket; there could be air there.	○ I tried this also. The turned-over bucket is heavy to pull up. There may be air in it (to make it heavy).
P2	○ The paper would not get wet, when even half the cup is filled with air.	○ Air comes first, so the paper in the cup would not get wet.
P3	○ I tried this; the water was to the top of the glass, maybe because of the buoyancy force.	○ Just pulling up the glass would not do anything. The water inside stays there.
P4	○ When sucked there would be nothing in the glass, so the water should come up.	○ The air would escape if there is enough time (for the air to do so).
P5	○ The strength of push does not have anything to do with it.	○ I have tried this with a toy; when I pushed it hard, the water rushed out.
P6	○ Where there is no air, water should come up, even when the length is 1m.	○ Teacher said she would do her best, so I think she could do it.
P7	× If there were a hole on top, the juice would come out, but with a little hole on the bottom, the juice would only drip.	× When I made a hole in a small drink bottle, the juice dripped. When I made the hole bigger, then juice came out continuously.
P8	× (SEE TABLE 5 for detail)	× (SEE TABLE 5 for detail)
P9	○ Because the upper hole is covered, the water would not go in.	○ (no mention of justification)
P10	○ Now there is a hole on top, so the water pushed in, and pushed the air out.	× Water will come in about halfway. → ○ This time a large amount of water will come in.
P11	○ When some air goes in, it pushed the soy sauce. If you cover the hole, no air goes in, so no sauce would come out.	○ I tried a similar thing, and only a little soy sauce came out, so it would not come out.

N.B. ○ is the correct prediction, × indicates a wrong choice

Both students chose correct answers from the beginning to the middle (P1 to P6), indicating they possibly shared the model correctly enough to choose the correct answer. Yet in the latter part of the course, they did not answer correctly, and their choices differed for P8 and P10. This pattern could mean at least two things. First, they were similar to each other when they based their judgments on their experiences (i.e., they shared more or less the same set of experience-based rules of thumb), but their shifted concepts differed. Second, they may have held different models or rules of thumb from the very beginning, both of which were correct enough for each to choose the correct answer (for different reasons); and the difference became sharper as the problems

became more difficult and required more sophisticated use, or expressions of justification for their choices.

### Diversity of models

A closer look at the students' utterances indicates that for P1 through P6, though both students based their justifications on their daily experience, they apparently formed some generalizable model of "when air goes out, water gets in." This explanation became more sophisticated as the problem progressed. Both talked about their bath experiences in P1. However, for P5 student A mentioned "when there is the same amount of force to push, the amount of air coming out should also be the same," whereas student B said "because the same amount of air is lost, so should be the amount of water." These explanations were applicable not only to the problem at hand but also to previous problems, indicating increased abstraction levels.

Yet slight differences were observed even in these early-stage justifications. While A repeatedly used the word "top" in his explanation, B did not use it at all. This difference becomes more readily apparent in their justification explanations after P7, where the children's models stopped working.

For P7, A's repeated use of the word "top" indicated that he was trying to use the same model until P6. However, B began to consider the size of the hole, introducing a new factor in his model (he did not mention size before P7). For P8, their explanations clearly differed as they engaged in a lively discussion in front of the class. Table 5 presents details of the beginning of their explanations.

Table5: Example of utterances of A and B during P8 discussion

A	B
This time, there are two holes at the bottom of the can. I think, if one hole is above the other, it is easy for the air to enter from that hole and the juice would come out from the other hole below. But in this problem, both of the holes are on the bottom so no air could get in. That is, because the air is outside, no juice could come out, or no air could go in either. The two are separated at the bottom.	For me, it does not matter which hole lets the juice out, but the air goes in from here, it goes up to fill half of the can, then the juice would get pushed out from the other hole and drips out.

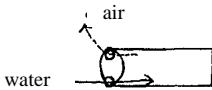
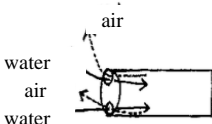
A distinguished between the positions of the two holes punched in the can. This distinction enabled him to explain "if one hole is above the other, it is easy for the air to enter from that hole," and "the juice would come out from the other hole below." From these utterances, we infer that A's model has components like "air is above, while water is below." His perspective is on the relative positioning of the air and water. We call this model the "positioning model." In contrast, B mentioned that "it does not matter which hole

lets the juice out.” Thus, we infer that he did not pay attention to the positions. Instead his explanation that “air goes in ...to push the juice out” indicates that he focused more on the interaction between air and water. We call this model the “interaction model.”

For the entire discussion of P8, A kept using the same model, emphasizing the significance of the relative positions. His last comment was “if there were any air in the can, it stays on top, so no juice would come out. In contrast, toward the end of this discussion, B began referring to the relative positions of the two holes in his model as “the air, when it somehow gets in, may move through there to reach the top, and the air on the top would let the juice come out from the bottom,” indicating his awareness of the possible integration of A’s model into his. Many questioned B’s choice, so B had to explain his position repeatedly. This could have contributed to B’s refinement of his model, while A could stay with his older model.

Preliminary analysis indicates that A and B preserved their differences and never converged to the same model. At the end of class, their explanations solicited by the teacher made this difference clear. Both drew their models on the blackboard, with explanations (Table 6).

Table 6: Last drawings and comments of A and B during P10 discussion

A	B
 <p>Though my choice is the same as B, the air becomes foamy when water comes, doesn't it? As the foam climbs up and up, stops on top and goes out. The bottom hole lets out not air, but water. The bottom is water first.</p>	 <p>Because there are holes on both, perhaps a lot goes in. Water goes in like this, and the water rushes out.</p>

This difference between A and B was maintained even at the very end of this course. We cannot deny the possibility that this difference kept both A and B engaged in the discussion, which remained the focus for the rest of the class and made everyone stay on task, thus fostering successful conceptual change for the whole class.

## Conclusion and Discussion

Analysis of the data gained from this HEI class on “Air and Water” indicated three patterns. Through the course of 12 classes covering 11 experiments or problems, the 21 children discussed their predictions of possible answers. This design successfully led them to change their experience-based rules of thumb into more scientific understanding of “air and water do not share the same space.”

Each child’s path of this conceptual change was unique, allowing each to create her/his own understanding. As indicated by qualitative analysis of the utterances of the two most active students, each could preserve her/his own model, possibly from the very beginning to the end. Yet this diversity among the children could be the source of prolonged discussion on the same topic for the length of this course, allowing both fast and slow learners to change their models. We plan to devise a better way to infer the cognitive models from these sporadic yet rich and complex utterances of children, in an effort to determine conditions for successful classroom discussion patterns that lead to conceptual change for every individual child in regular classes.

## References

- Clement, J. (2008). The Role of Explanatory Models in Teaching for Conceptual Change. Vosniadou, S. (Ed.), "Handbook of research on conceptual change", London, Taylor & Francis Group, 2008. pp.417-452
- Howe, C., McWilliam, D. & Cross, G. (2005). "Chance favours only the prepared mind: Incubation and the delayed effects of peer collaboration. *British Journal of Psychology* 96, 67-93.
- Itakura, K. (1963). "Kasetsu-Jikken-Jugyo no Teisho, (The proposition of Hypothesis-Experiment-Instruction), " *Rika Kyoshitsu (The Journal of Science Education)*, November, 1963
- Itakura, K. (1970). *The workbook of Hypothesis-Experiment instruction <Air and Water>*. Tokyo: Kasetsu-Sha. (Original version [in Japanese] 1970, English Version 2007)
- Miyake, N. (1986). "Constructive interaction and the iterative process of understanding," *Cognitive Science*, 10, 151-177.
- Miyake, N. (2008). Conceptual Change through Collaboration. In Vosniadou, S., ed. *International Handbook of Research on Conceptual Change*. London, Taylor & Francis Group.
- Miyake, N. (2009). *Conceptual change through collaboration, Paper presented at AERA 2009, San Diego*.
- Roschelle, J. (1992). Learning by Collaborating: Convergent Conceptual Change. *Journal of the Learning Sciences*, 2(3), 235-276
- Saito, M., & Miyake, N. (2011). "Socially constructive interaction for fostering conceptual change," *Proceedings of the 9th International Conference on Computer-Supported Collaborative Learning, (CSCL2011)*, Hong Kong, 96-103.
- Shirouzu, H., Miyake, N., & Masukawa, H. (2002). Cognitively active externalization for situated reflection, *Cognitive Science*, 26(4), 469-501.
- Vosniadou, S. (Ed.), (2008) "Handbook of research on conceptual change", London, Taylor & Francis Group.

# Strategy Changes in Causal Structure Learning: The Role of Task Complexity

Motoyuki Saito (m-saito@kwansei.ac.jp)

Department of Psychological Science, Kwansei Gakuin University  
Hyogo, 662-8501, JAPAN

Tsuneo Shimazaki (shimazaki@kwansei.ac.jp)

Department of Psychological Science, Kwansei Gakuin University  
Hyogo, 662-8501, JAPAN

## Abstract

Saito and Shimazaki (2012) found that people rely upon covariation information rather than temporal order information as cues to causal structure, whereas Lagnado and Sloman (2006) reported an opposite finding, indicating relatively greater influence of temporal order cues. The present research examines the hypothesis that such conflicting findings result from differences in task complexity. Specifically, it is proposed that covariation information becomes less influential as the number of variables increases. Experiment 1 investigated the relationship between the judgment strategy (i.e., covariation vs. temporal order) and the number of variables comprising a causal structure. As a result, people favored covariation cues primarily in tasks with simple causal structure. Experiment 2 used more complex causal structure. The results demonstrated that the tendency to emphasize covariation cues or to rely upon temporal order cues changes as a function of task complexity. These results were consistent with both previous findings and discussed in terms of causal Bayes net theories and heuristic models.

**Keywords:** causal structure learning; causal reasoning; covariation; temporal order; task complexity.

## Introduction

Many psychological studies have shown that both children and adults easily form representations of causal relations (Sloman, 2005; see also Holyoak & Cheng, 2011 for a review). Causal relations lead to associations of various kinds of events and they may form part of complex causal structures. Knowledge about the causal structure plays an important role in explanations, predictions, control, as well as decision making (Pearl, 2000). Despite its importance, in many cases, actual causal connections among constituent elements are often difficult to discern or to tease out of a complicated pattern of contingencies. For instance, if one hears a bit of gossip from colleague, X, and then hears same story from another colleague, Y, this might lead to the inference that X had passed the rumor to Y (i.e.,  $X \rightarrow Y$ ) based on the temporal order in which one receives this information. However, it is also possible that, earlier, Y had initially gossiped to X and then it is heard from X prior to seeing Y (i.e.,  $Y \rightarrow X$ ). Or, a third party, such as the boss, Z, may have spread this rumor (i.e.,  $X \leftarrow Z \rightarrow Y$ ). In light of this, how might be people acquire knowledge about causal structure?

As Hume (1739/2000) has argued that causal relations are unobservable and therefore must be induced from

observable events, covariation among observable events serves as a fundamental cue to learn causal structure. Covariation is formally represented as a joint probability distribution and is specifically explained as patterns of presence and absence for binary variables. When a causal relation exists, strong covariation between a cause and its effect will be expected except for the possibility that both variables are caused by a common cause. In contrast, the absence of covariation indicates that two variables are not related to each other—except for the effects of other variables. However, there are several limitations to the use of covariation cues. First, covariation information becomes more complex as the number of variables increases. With two binary variables, for example, covariation is represented by a  $2 \times 2$  contingency table of 4 data patterns; however, 32 data patterns result from five binary variables. In addition, covariation itself is inadequate for distinguishing a unique causal structure from models that represent the same joint probability distribution (i.e., Markov equivalent). When event X covaries with event Y, for instance, it is difficult to determine the precise cause. These examples suggest the difficulty of learning causal structure using only covariation cues.

In addition to covariation, another important cue to causal structure is temporal order in which people observe the states of variables. Because causes are often observed to happen prior to their effects, when event X precedes event Y, it is probable that X causes Y. However, the observed temporal order does not always serve as an accurate cue. First, temporal order may mislead people with regard to the direction of causal relations. In situations where people observe effects prior to their causes, temporal order information indicates the opposite causal direction. A second issue concerns spurious correlation. Even if event X precedes event Y, their co-occurrence might be the result of a hidden common cause Z. In this case, a temporal delay between two variables will result in a false belief that the earlier event causes the later, despite the fact that no causal relation exists. Thus, although temporal order cues can facilitate causal structure learning, they may also mislead causal inferences.

Combined with information indicating hidden causes are absent, both kinds of information become more useful. When event X covaries with event Y, three possible causal structures are presumed (i.e.,  $X \rightarrow Y$ ,  $X \leftarrow Y$ , or  $X \leftarrow Z \rightarrow Y$ ). The absence of hidden causes enables people to exclude the



possibility that both events are caused by a hidden cause. If event X occurs alone in this situation, then it is suggested that X causes Y. Since nothing happens without a cause (i.e., necessity), an event that occurs alone must be a cause variable but not an effect variable.

Previous studies on causal structure learning have provided conflicting evidence regarding the use of covariation cues and temporal order cues (e.g., Lagnado & Sloman, 2006; Saito & Shimazaki, 2012; White, 2006). On one hand, Lagnado and Sloman (2006) demonstrated that people preferred temporal order cues to covariation cues. In their experiment, participants were required to send messages from a master computer to one of four computers in a network (e.g., computer A), to observe whether other computers also received the messages (e.g., computer B, C, & D), and then to infer the structure of network. Participants observed the states of the computers in the order different from the causal order (e.g., temporal order:  $A \rightarrow D \rightarrow C \rightarrow B$ , causal order:  $A \rightarrow B \rightarrow C \rightarrow D$ ). Although participants were given instructions including information on the unreliability of temporal order and the absence of hidden causes, their judgments were based on temporal order cues rather than covariation cues. White (2006) reported similar results indicating that participants relied heavily on temporal order information, in spite of the fact that they received explicit instructions regarding the way in which causal structures are induced from covariation information.

On the other hand, Saito and Shimazaki (2012) showed the opposite results that people use covariation cues rather than temporal order cues. The experimental task was to observe occurrences of two types of bacteria and to infer their causal relationship. As in Lagnado and Sloman (2006), participants were instructed that temporal order cues were unreliable and that there were no hidden causes. In the condition in which covariation cues contradicted temporal order cues, participants heavily favored covariation information over temporal order information. Furthermore, these judgments were made after several observations.

A possible interpretation of these conflicting findings involves differences in task complexity, especially as this is reflected by number of variables. It is possible that task complexity modulates an individual's judgment strategy. A critical difference between the preceding experiments is the number of variables that constitute the causal structure. Whereas Lagnado and Sloman (2006) required participants to learn causal directions among four variables, and White (2006) adopted five variables in constituting a causal structure, the design of Saito and Shimazaki (2012) presented a causal structure involving only two variables—the minimum unit for causal structures. Since increasing the number of variables complicates covariation information, it therefore would be difficult to induce a complex causal structure based solely on covariation cues.

Several studies have revealed the role of task complexity in causal learning and inference (e.g., Marsh & Ahn, 2006; Reips & Waldmann, 2008). Marsh and Ahn (2006) demonstrated that the task complexity served as a

determinant of a primacy effect and a recency effect. When a few variables were observed, information presented earlier weighted more than information presented later (i.e., primacy effect); in contrast, information presented later was emphasized more than information presented earlier when many kinds of variables existed (i.e., recency effect). In a similar vein, Reips and Waldmann (2008) showed that accurate diagnostic inferences depended on the number of variables in the experimental task. These studies suggest the importance of task complexity in causal judgment.

The purpose of the present study is to investigate the relationship between task complexity and the judgment strategy in causal structure learning. In order to manipulate task complexity, the number of variables composing the causal structures was manipulated in Experiment 1. Experiment 2 employed different forms of causal structures. The hypothesis predicts that people use covariation cues rather than temporal order cues in learning simple causal structures whereas they rely more upon temporal order cues than covariation cues in inferring complex causal structures.

## Experiment 1

Experiment 1 was designed to investigate how strategies change as a function of the number of variables in learning a causal structure. The experimental task was to observe states of the variables and to infer causal relations among these variables. The number of variables in the causal structure was varied for the manipulation of task complexity. The authors predicted that when the number of variables was small covariation would be more emphasized rather than temporal order and that when the number of variables was large temporal order would be more reflected than covariation.

## Method

**Participants and design** A total of 24 undergraduates from Kwansei Gakuin University received course credit for taking part in this experiment. The number of variables (three, four, and five) was manipulated within participants. Each participant performed three causal learning tasks with different causal structures.

**Instructions** Participants received verbal and written instructions in Japanese. An English translation of outlines of the instructions was as follows:

Imagine that you are a scientist who is attempting to reveal causal relations among several newly discovered bacteria. Whether one bacterium propagates another bacterium, or whether they are irrelevant to each other are unknown to you. In order to investigate the relations among the bacteria, you put one type of bacterium into the 40 containers of liquid nutrient medium. Then you observe the states of other bacteria under the microscope. Before performing each task, you will be informed about which bacterium is put into the nutrient medium. (The bacterium serves as a first cause.)

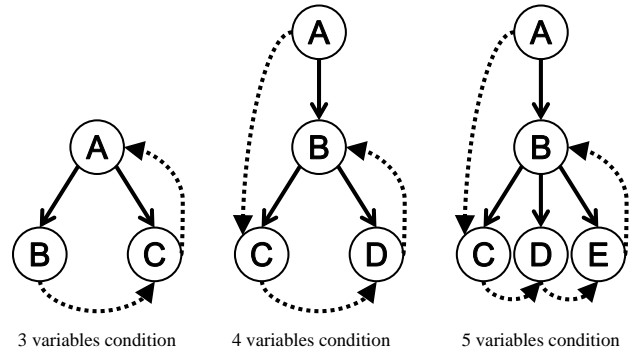
Additionally, the following three facts should help you consider causal relations among bacteria. First, it is not always true that one bacterium is certain to propagate other bacterium even if there is a causal relation between them. That is, a causal relation among bacteria is probabilistic. Second, there are no hidden causes as this is a controlled situation. Therefore, when a bacterium exists, except for the one originally put into the liquid nutrient medium by yourself (i.e., first cause), there is another bacterium that acts to propagate it. Third, the states of bacteria develop in order. This is either because it takes time for a bacterium to propagate another bacterium or because it takes time to identify the presence and absence of bacteria due to their invisibility without the use of microscope. Therefore, it could be that a bacterium has propagated another bacterium before it turned out to be present.

Your task is to observe the occurrences and non-occurrences of these bacteria and to infer causal relations among them. Note that the experimental task does not require any knowledge of biology. (The remaining instructions describe how to progress through the learning phase and test phase.)

After receiving the instructions, participants were asked whether they understood the instructions. In this cover story, the number of variables (three vs. four vs. five) corresponded to the number of bacteria participants observed in each task.

**Learning phase** At the beginning of the learning phase, participants were taught how many kinds of bacteria would appear (e.g., three, four, or five) and which bacterium would be put into the 40 containers of nutrient medium (i.e., first cause) in each condition. The learning phase consisted of 40 trials that presented information about the presence and absence of the bacteria. Participants were asked to observe the states of the bacteria and to consider their causal relations. First, a button labeled “NEXT” was displayed on a screen. After clicking the button, the shapes of the several number of bacteria labeled with question marks were shown (i.e., they remained unknown). The number of bacteria varied across conditions. There were three bacteria in the three variables condition, four bacteria in the four variables condition, and five bacteria in the five variables condition. Then, information about the state of the bacterium was given in order. The presence of bacteria was indicated by the appearance of bacteria; in contrast, the absence of bacteria was represented by the appearance of bacteria labeled with a cross mark. The inter-stimulus interval was 1s, and the screen was returned to its primary state (i.e., “NEXT”) 2s after the all bacteria appeared.

Figure 1 illustrates the causal order in which causes produced effects and the temporal order in which participants observed the states of variables. Covariation information was arranged based on these causal structures and the causal strength of the relation. As the instructions



3 variables condition 4 variables condition 5 variables condition

Figure 1: Causal structures in Experiment 1. Continuous lines represent causal relations, whereas dotted lines indicate temporal orders.

stated causal relations among bacteria were probabilistic, the probability that causes produced effects was 80 percent. In the three variables condition in which participants observed 40 cases, for example, 26 cases included the presence of bacteria A, B, and C. In 6 cases, bacteria A and B were present; bacteria A and C were present in additional 6 cases. The remaining 2 cases included the presence of bacterium A and the absence of bacteria B and C. While the first cause was always present, other variables did not occur unless their specific cause was present. As the number of variables increased, covariation information became more complex. As is evident in Figure 1, temporal order was inconsistent with causal order. Arranging the temporal order to be different from the causal order enabled assessment of the degree to which covariation cues and temporal order cues were used to infer causal structure.

**Test phase** After observing 40 cases, participants were told to infer causal structure in the test phase. Participants received a sheet in which bacteria were displayed in the same way as they were shown in the learning phase, with light gray lines between the bacteria. The instructions required participants to judge whether a casual relation existed and to draw an arrow from a cause to an effect on the line when a causal relation was assumed. Participants had to consider three lines in the three variables condition, six lines in the four variables condition, and ten lines in the five variables condition.

## Results and Discussion

To investigate judgment strategy in causal structure learning, the authors defined usage rates as measures of the degree to which participants used covariation cues and temporal order cues. The usage rate of covariation was calculated by dividing the number of links drawn by participants to be consistent with covariation cues by the number of all links suggested by covariation information. In the three variables condition, for example, the link from A to B and the link from A to C were supported by covariation cues (see Figure 1). If participants gave these two links as their answer, their covariation usage rates were 100 percent; in contrast, when they failed to answer both links, the usage rates were 0 percent. The usage rate of temporal order was calculated in

a similar manner. In the above example, temporal order cues sustained the link from B to C and the link from C to A respectively (see Figure 1). If participants responded either of the links, their temporal order usage rates were 50%. These indices represent the amount of use of each cue by participants.

Figure 2 shows the usage rates of two types of cues in each condition. When the number of variables was three or four, participants used covariation rather than temporal order; however, there seemed to be no difference in the five variables condition. A two-way repeated measures ANOVA with the type of cue (covariation vs. temporal order) and the number of variables (three vs. four vs. five) as within-participants factors revealed a significant main effects of the type of cue,  $F(1, 23) = 14.01, p < .01$ , and the number of variables,  $F(2, 46) = 4.43, p < .05$ . The interaction between the type of cue and the number of variables was also significant,  $F(2, 46) = 6.76, p < .01$ . Subsequent tests of the simple main effect of the type of cue were significant in the three and four variables condition,  $F(1, 69) = 21.03, p < .001, F(1, 69) = 10.11, p < .01$  respectively, but not in the five variables condition,  $F < 1$ , indicating the task complexity served as the modulator of the judgment strategy.

Although the judgment strategy varied across the number of variables, the result of the four variables condition was inconsistent with previous findings in which the temporal order information had a greater impact than covariation in learning the causal structure composed of four variables (Lagnado & Sloman, 2006). In their experiment, the 58% usage rate of temporal order cues was higher than the 39% usage rate of covariation cues when the causal order ( $A \rightarrow B \rightarrow C$  &  $D$ ) differed from the temporal order ( $A \rightarrow D \rightarrow C \rightarrow B$ ). The opposite results might stem from differences in cover stories. Whereas the present study dealt with the causal relations of several types of bacteria, the previous study used computer networks in which participants sent a message to one computer and observed whether other computers received. Prior knowledge and experience about computer message would give more

weight to temporal order cues. In fact, Saito and Shimazaki (2012) have reported that the use of temporal order cues depends on its reliability.

In summary, Experiment 1 showed that covariation cues were used more often than temporal order cues when participants learned causal relations among three or four variables. However, when the causal structure consisted of five variables, the preference for covariation disappeared. This pattern of results supports the claim that judgment strategy changes as a function of increasing the number of variables.

## Experiment 2

The results of Experiment 1 demonstrated that task complexity modulates participant's judgments about causal structures. However, these findings did not reveal a tendency to rely upon temporal order rather than covariation. The goal of Experiment 2 was to provide further evidence about the relationship between task complexity and the judgment strategy. In order to ascertain whether temporal order cues were more influential than covariation cues in learning complex causal structures, different forms of causal structures were used. Specifically, causal structures with multiple causal links were adopted. This is because increasing the number of causal links leads to more complicated covariation information. Again, the authors predicted that participants' judgments will be based more on covariation than on temporal order when the causal structure was relatively simple, whereas temporal order cues should be more influential than covariation cues when participants inferred the complex causal structure.

## Method

**Participants and design** A total of 24 undergraduates from Kwansei Gakuin University participated for course credit. None of them took part in Experiment 1. As in Experiment 1, the number of variables (three, four, and five) was varied within participants. Participants were asked to perform three causal learning tasks with different causal structures.

**Procedure** Each participant completed the tasks of observing states of the bacteria and inferring their causal relations. The procedure was identical to Experiment 1 with the exception that different forms of causal structures were used. Although the number of variables was constant across two experiments, the number of causal links in Experiment 2 was larger than that in Experiment 1 (see Figure 1 and 3). Increasing causal links resulted in more complex patterns of covariation information. For example, the four variables condition in Experiment 2 provided seven types of co-occurrence information, whereas there were five kinds of co-occurrence information in the four variables condition in Experiment 1.

Instructions explained the cover story and indicated to the participants that they were required to judge causal relations among bacteria. As in Experiment 1, participants were informed that causal relations were probabilistic, that there were no hidden causes, and that temporal order was

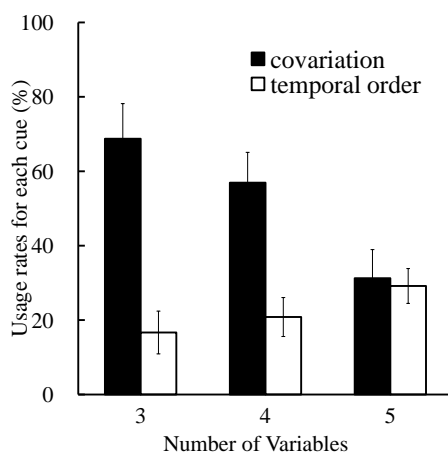


Figure 2: Usage rates for covariation cues and temporal order cues in Experiment 1. Error bars reflect standard errors.

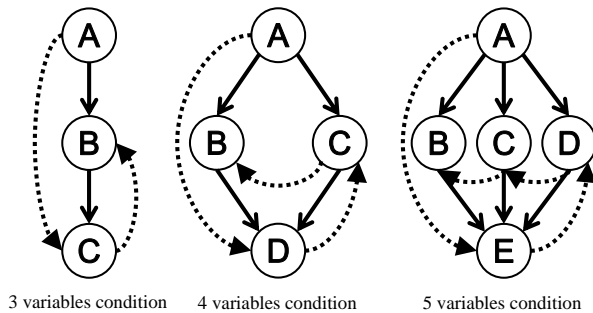


Figure 3: Causal structures in Experiment 2. Continuous lines represent causal relations, whereas dotted lines indicate temporal orders.

not always an accurate cue. In the learning phase, participants received information on 40 cases of bacteria states through observation. Each condition differed in the number of bacteria participants observed. As can be seen in Figure 3, there were three variables with two causal links in the three variables condition, four variables with four causal links in the four variables condition, and five variables with six causal links in the five variables condition. Moreover, the causal order differed from the temporal order in each condition, allowing assessment of the degree to which each cue was used. In the test phase, participants were told to infer the causal structure in the same way as in Experiment 1.

## Results and Discussion

Participants' responses were analyzed in a manner similar to Experiment 1. Figure 4 shows the usage rates of covariation cues and temporal order cues in each condition. The results of the three variables condition indicated that covariation cues were emphasized over temporal order cues, replicating this effect in Experiment 1. In contrast, participants in the four and five variables conditions based their judgment more upon temporal order than upon covariation. A 2 (the type of cue: covariation vs. temporal order)  $\times$  3 (the number of variables: three vs. four vs. five) repeated measures ANOVA yielded a significant main effect of the number of variables,  $F(2, 46) = 9.31, p < .001$ , and a significant interaction between the type of cue and the number of variables,  $F(2, 46) = 9.66, p < .001$ . To explore the interaction, an analysis of the simple main effect of the type of cue was conducted for each condition. The tendency to emphasize covariation rather than temporal order was marginally significant in the three variables condition,  $F(1, 69) = 3.36, p < .10$ . There was no significant difference in the four variables condition,  $F(1, 69) = 1.75, ns$ . In the five variables condition, however, the usage rate of temporal order cues was reliably higher than that of covariation cues,  $F(1, 69) = 7.94, p < .01$ . These results suggest that task complexity determines whether participants rely on covariation cues or temporal order cues.

In order to investigate effects of forms of causal relations on judgment strategy, a 2 (causal structure: Exp.1 vs. Exp.2)  $\times$  2 (the type of cue: covariation vs. temporal order)  $\times$  3 (the

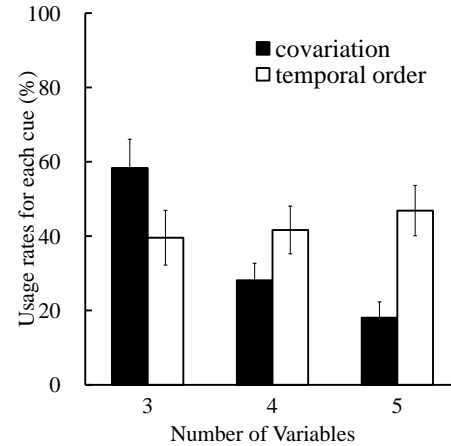


Figure 4: Usage rates for covariation cues and temporal order cues in Experiment 2. Error bars reflect standard errors.

number of variables: three vs. four vs. five) mixed ANOVA was performed, with causal structure as a between-participants factor and the type of cue and the number of variables as within-participants factors. As a result, a main effect of the number of variables,  $F(2, 92) = 12.03, p < .001$ , an interaction between causal structure and the type of cue,  $F(1, 46) = 11.21, p < .01$ , and an interaction between the type of cue and the number of variables,  $F(2, 92) = 15.11, p < .001$  were significant. To explore the interaction between causal structure and the type of cue in greater detail, the simple main effects of causal structure were tested. The usage rate of covariation cues in Experiment 1 was higher than that in Experiment 2,  $F(1, 92) = 5.92, p < .05$ . On the contrary, the usage rate of temporal order cues in Experiment 1 was lower than that in Experiment 2,  $F(1, 92) = 8.14, p < .01$ . These results indicate that forms of causal structures also modulate judgment strategy.

Taken together with the results of Experiment 1, Experiment 2 provides further evidence on the relationship between task complexity and the judgment strategy. Participants in the three variables condition emphasized covariation over temporal order; on the other hand, temporal order cues were given more weight than covariation cues in the five variables condition. These results bridge the gap between the findings about the preferential use of covariation (Saito & Shimazaki, 2012) and about the preferential use of temporal order (Lagnado & Sloman, 2006; White, 2006). In addition, the comparison between experiments demonstrates that judgment strategy is affected not only by the number of variables but also by the form of the causal structure.

## General Discussion

The present study clarifies conflicting evidence concerning the use of covariation cues and temporal order cues in causal structure learning. Lagnado and Sloman (2006) showed temporal order to be more influential on judgments than covariation; however, Saito and Shimazaki (2012)

reported that covariation was favored over temporal order. In the present study, the authors interpreted these results from the viewpoint of the task complexity and manipulated the number of variables which consisted of a causal structure. Experiment 1 demonstrated that covariation cues were carried more weight than temporal order cues in learning simple causal structure and that this preference disappeared as the number of variables increased. In addition, Experiment 2 investigated the relationship between the judgment strategy and the number of variables with different forms of causal structures. The results of Experiment 2 are consistent with both findings concerning the preferential use of covariation cues in a simple task (Saito & Shimazaki, 2012) and the preferential use of temporal order cues in a complex task (Lagnado & Sloman, 2006; White, 2006). These results show the task complexity, composed of the number of variables and the form of the causal structure, serves as a modulator of the judgment strategy in causal structure learning.

The results of the present study can be interpreted in terms of salience and validity in multiple-cue probability learning (Kruschke & Johansen, 1999). According to Kruschke and Johansen (1999), irrelevant cues have a deleterious effect on the use of valid cues and this effect becomes more apparent as the salience of irrelevant cues increase. In the present experiment, covariation served as valid cue whereas temporal order was high salient but less reliable cue. Increasing the number of variables in the causal structure brought lower salience of covariation and higher salience of temporal order, which resulted in a deleterious effect on the use of covariation cues. These similar findings imply the tight coupling between causal learning and category learning.

The present study has several implications for models in causal structure learning. The use of covariation cues is easily explained by constraint-based methods (Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004) and Bayesian methods (Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003) in causal Bayes nets and broken link heuristics (Mayrhofer & Waldmann, 2011). Whereas Constraint-based methods compute independence and dependence in bottom-up process, Bayesian methods make probabilistic inferences for each causal model using Bayes' theorem in top-down process. Broken link heuristics offers a simple explanation with a determinism bias and a sufficiency bias. In contrast, the use of temporal order cues is well accounted for by local computations (Fernbach & Sloman, 2009). According to this heuristic model, people focused not on covariation cues but instead on temporal order cues because of their quick accessibility and lower computational demands. The tendency to use temporal order cues is also explained by temporal strategy (Rottman & Keil, 2012), which induces causal directionality from temporal change over time. Although these models focus on either covariation cues or temporal order cues, the present results suggest the importance of both cues in causal learning. An intriguing question for future research concerns how people integrate

covariation with temporal order for inferring causal structure.

## References

- Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 678-693.
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111, 3-32.
- Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, 62, 135-163.
- Hume, D. (1739/2000). *A treatise of human nature*. Oxford, England: Oxford University Press.
- Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1083-1119.
- Lagnado, D., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 451-460.
- Marsh, J. K., & Ahn, W.-K. (2006). Order effects in contingency learning: The role of task complexity. *Memory & Cognition*, 34, 568-576.
- Mayrhofer, R., & Waldmann, M. R. (2011). Heuristics in covariation-based induction of causal models: Sufficiency and necessity priors. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3110-3115). Austin, TX: Cognitive Science Society.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, United Kingdom: Cambridge University Press.
- Reips, U.-D., & Waldmann, M. R. (2008). When learning order affects sensitivity to base rates: Challenges for theories of causal learning. *Experimental Psychology*, 55, 9-22.
- Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, 64, 93-125.
- Saito, M., & Shimazaki, T. (2012). *Rethinking the use of covariation in causal structure learning*. Manuscript submitted for publication.
- Sloman, S. A. (2005). *Causal models: how people think about the world and its alternatives*. New York: Oxford University Press.
- Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, 27, 453-489.
- White, P. A. (2006). How well is causal structure inferred from cooccurrence information? *European Journal of Cognitive Psychology*, 18, 454-480.

# The Comprehension of Adjective Metaphors Is Selectively Affected By Negative Meanings Associated With Adjectives As Vehicles

**Maki Sakamoto (sakamoto@inf.uec.ac.jp)**

Department of Informatics and Engineering, The University of Electro-Communications  
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

**Miho Sumihisa (m\_sumihisa@edu.hc.uec.ac.jp)**

Department of Informatics and Engineering, The University of Electro-Communications  
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

**Takuya Matsumoto (matsumoto@edu.hc.uec.ac.jp)**

Department of Human Communication, The University of Electro-Communications  
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

**Akira Utsumi (utsumi@inf.uec.ac.jp)**

Department of Informatics and Engineering, The University of Electro-Communications  
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

## Abstract

Previous metaphor studies have paid much attention to nominal metaphors and predicative metaphors and little attention has been given to adjective metaphors. The most adjective metaphor studies have only examined how the acceptability of adjective metaphors can be explained by the pairing of adjective modifier's and head noun's modalities. Sakamoto & Utsumi (2009) showed that adjective metaphors, especially those modified by color adjectives, tend to evoke negative meanings. Sumihisa et al (2011) examined whether evoking negative meanings is the unique feature of adjective metaphors through comparison among nominal metaphors and predicative metaphors for the Japanese language and revealed that meanings of metaphors are basically affected by meanings of vehicles, but when vehicles themselves had neutral meanings, negative meanings were evoked more frequently for adjective metaphors among the other types of metaphors. The purpose of this study, therefore, explores the reason why adjective metaphors evoke negative meanings more frequently than the other types of metaphors. For this purpose, we examined what kind of meanings associated with topics or vehicles affect the comprehension of metaphors. Our psychological experiments revealed that meanings associated from vehicles affect the comprehension of metaphors. And when metaphorical expressions have vehicles with positive or negative meanings, metaphorical expressions show the same meanings as the vehicles. On the other hand, when metaphorical expressions have vehicles with neutral meaning, only adjective metaphors evoke negative meanings. Our results suggest that the comprehension of adjective metaphors is selectively affected by the negative meanings associated with adjectives as vehicles.

**Keywords:** adjective metaphors; nominal metaphors; predicative metaphors; Japanese language; negative meanings.

## Introduction

Metaphor studies in the domain of cognitive science have paid much attention to nominal metaphors such as “*My job is a jail*” (e.g., Bowdle & Gentner, 2005; Glucksberg, 2001) and predicative metaphors such as “*He shot down all of my arguments*” (e.g., Lakoff & Johnson, 1980).

Previous metaphor studies, however, have paid little attention to adjective metaphors such as “*sweet touch*” and how they are comprehended. Some models have been proposed to explain the mechanism of metaphor comprehension in cognitive science. Glucksberg and his colleagues (Glucksberg & Keysar, 1990) propose categorization theory. This theory addresses mainly nominal metaphors and argues that people understand nominal metaphors by seeing the target concept as belonging to the superordinate metaphorical category exemplified by the source concept. As for the mechanism of adjective metaphors, Utsumi & Sakamoto (2007) propose a two-stage categorization theory and argue that the comprehension process of adjective metaphors could be explained as a two-stage categorization process.

Many studies focusing on adjective metaphors have examined how the acceptability of adjective metaphors can be explained by the pairing of adjective modifier's and head noun's modalities. Ullmann (1951), in a very early study on adjective metaphors, proposes a certain hierarchy of lower and higher perceptual modalities. His thesis of directionality asserts that a metaphor with a source domain lower in the hierarchy of sense modalities than the target domain should tend to be cognitively more accessible than a metaphor with the reverse direction of domains. Williams (1976) makes a more differentiated claim of directionality, in which a similar order of sense modalities is proposed. Recently, Yu (2003) highlights cross-linguistic differences when he makes different directionality claims for different languages (English as compared to Chinese). Werning, Fleischhauer, & Beşeoğlu (2006) explore the factors that enhance the cognitive accessibility of adjective metaphors for German. Very few studies, however, have attempted to explore meanings evoked by adjective metaphors.

Sakamoto & Utsumi (2009) is one of the few studies which have explored meanings evoked by adjective metaphors. They compare the actual semantic changes observed through their psychological experiments with the semantic changes predicted by Abstract Performance Grammar (APG) model. APG proposed by Osgood (1980) states the crucial rules to evoke semantic changes through fine semantic interactions in the processing of linguistic expressions.

158 Japanese adjective metaphors were used for their psychological experiment. Participants were asked to rate

the assigned expressions against 15 SD scales such as “uncomfortable – comfortable” and “dark – light”. The ratings were made on a 7-point scale ranging from -3 through 0 to +3. The value -3 was regarded as the negative semantic pole and the value +3 as the positive semantic pole. All the mean values of vehicles and topics rated on the 15 SD scales were classified into T=V, T<V, T>V (T : topics, V : vehicles). Using t-test (two-tailed, the alpha level .05), the cases which have no significant difference between the mean value of T and V were regarded as T=V.

The other codes such as T<V and T>V fall to the cases which have significant differences between the mean values of T and V. In order to compare the actual semantic changes resulting from their experiment with the semantic changes predicted by APG model, Sakamoto & Utsumi (2009) classified the actual semantic changes resulting from their experiment as show in Table 1. Using t-test (two-tailed, the alpha level .05), they regarded the cases which have no significant difference between the mean values of T and metaphors as ‘no change’ (0) and the cases which have significant differences between them as changes either to the negative pole (-) or to the positive pole (+). Table 1 shows the comparison between the predicted semantic changes and the actual semantic changes observed through their experiment.

Table 1: Comparison between predicted semantic changes and actual semantic changes

semantic intensity	predicted change	actual change			Sum
		0	+	-	
T=V	0	331	17	261	609
T<V	+	366	230	76	672
T>V	-	119	9	961	1089
Sum		816	256	1298	2370

numbers = cases of SD scales

In order to see the tendency for adjective metaphors to evoke positive or negative meanings, Sakamoto & Utsumi (2009) classified all the cases showing different changes from the APG prediction either into positive meaning or negative meaning. The cases showing no change as against the prediction of changing to - were regarded as evoking a weakly positive meaning, and were classified into the positive meaning category in the same way as those which changed to + against the prediction of changing to -. The cases showing no change against the prediction of changing to + were regarded as evoking weakly negative meaning, and were classified into the negative meaning category in the same way as those which changed to - against the prediction of changing to +. As a result, 848 cases which showed changes different from the APG prediction were classified into 145 positive meanings and 703 negative meanings. A Chi-square test showed that the cases showing negative meanings were significantly more frequent than those showing positive meanings,  $\chi^2 (1, N=848) = 367.175, p < .001$ . Based on this result, Sakamoto & Utsumi (2009) suggest that adjective metaphors tend to evoke negative meanings.

Sumihisa et al. (2011) examined whether evoking negative meanings is the unique feature of adjective metaphors. In the experiments, Sumihisa et al. (2011) first selected nouns as topics to make Japanese metaphorical expressions. They selected four nouns (e.g. *smell* ('nioi'), *moment* ('genzai'), *footstep* ('ashito'), and *pose* ('shisei')) with neutral meanings among 54 nouns by psychological experiment. They also conducted another psychological experiment in which participants were asked to rate meanings of vehicles only. Verbs, nouns and

adjectives were selected respectively as vehicles with positive meanings, neutral meanings and negative meanings. They combined the topics and the vehicles and made metaphorical expressions.

Sumihisa et al. (2011) conducted a psychological experiment in which participants evaluate the meanings of metaphors. Participants were asked to rate the assigned expressions against 9 SD scales (7 SD scales given in Table 2 and additional scales “difficult – easy” and “unfamiliar – familiar”). The ratings were made on a 7-point scale ranging from -3 through 0 to +3. They regarded the value -3 as the negative semantic pole and the value +3 as the positive semantic pole.

Table 2: List of SD scales used for the experiment

dislike – like	inelegant - elegant
ugly - beautiful	uncomfortable - comfortable
dark - light	bad - good
sad - glad	

Sumihisa et al. (2011) focused on the metaphorical meanings evoked by the semantic interaction between topics and vehicles. They classified metaphorical expressions into the cases showing no semantic change, those showing the change to the positive semantic pole or those showing the change to the negative semantic pole. They conducted t-test (two-tailed, the alpha level .05) to see semantic changes evoked by the semantic interaction between topics and vehicles. Since only the topics with neutral meanings were selected through the pre-experiment, metaphorical expressions which have no significant difference between their mean value and value 0 were regarded as metaphors showing no semantic change (0). And metaphorical expressions which have significant difference between their mean values and value 0 were classified into either metaphors showing the change to the positive semantic pole or those showing the change to the negative semantic pole.

As a result, when vehicles have positive or negative meanings, all types of metaphors tend to evoke positive or negative meanings. However, when vehicles have neutral meanings, although nominal metaphors tend to evoke neutral meanings, predicative metaphors and adjective metaphors tend to evoke negative meanings. Especially adjective metaphors tend to evoke negative meanings more frequently. They also classified the metaphors either into metaphors showing negative meanings or the others and compared among the three types of metaphors as shown as Table 3.

Table 3: Number of expressions showing negative meanings and the other meanings

	-	+ or 0	sum
nominal metaphors	7	19	26
predicative metaphors	8	19	27
adjective metaphors	17	11	28
sum	32	49	81

They revealed that adjective metaphors evoke significantly more frequently negative meanings than the other two types of metaphors,  $\chi^2 = (1, N = 54) = 6.234, p < .05$  for adjective metaphors vs. nominal metaphors,  $\chi^2 = (1, N = 55) = 5.357, p < .05$  for adjective metaphors vs. predicative metaphors.

Based on this result, Sumihisa et al. (2011) suggest that nominal metaphors and predicative metaphors basically tend to show neutral meanings, while adjective metaphors tend to show negative meanings.



This study explores the reason why adjective metaphors evoke negative meanings more frequently than the other types of metaphors. Utsumi & Sakamoto (2007) argue that the comprehension process of adjective metaphors can be explained as a two-stage categorization process. They speculate the comprehension process of “red voice” created from the neutral vehicle “red” as follows: the adjective “red” first evokes an intermediate category “red things” to which “blood”, “fire”, “passion”, “apple” and “danger” typically belong. Then exemplars relevant to the noun “voice” are selected and they evoke a final abstract category of property like “scary”, “screaming” and “dangerous”. In this way, adjective metaphors are understood by not be directly mapped onto the topics from ad hoc category of vehicles but mediating to an intermediate category. When meanings of adjective metaphors were processed in the two-stage categorization process, exemplars with negative meanings might be selected among various exemplars belonging to the intermediate category evoked by adjectives as vehicles.

In this study, therefore, we conducted a psychological experiment in which participants were asked to choose words related to meanings of adjective metaphors among those associated from vehicles and topics. We hypothesize that even if there were negative and positive exemplars in an intermediate category, exemplars with negative meanings tend to be selected to process meanings of adjective metaphors. As for nominal metaphors, on the other hand, prototypical exemplars associated with vehicles tend to be selected since the people understand nominal metaphors via the categorization process, namely by seeing the target concept as belonging to the superordinate metaphorical category exemplified by the source concept.

## Pre-experiment

### Topics and Vehicles

We decided to use 4 nouns as topics which were tested as having neutral meanings by Sumihisa et al. (2011); *smell* (“nioi”), *moment* (“genzai”), *footstep* (“ashioto”), and *pose* (“shisei”).

Candidates of vehicles of nominal, predicative and adjective metaphors were selected from the Japanese thesaurus (yamaguchi, 2003). We selected 50 adjectives, 50 nouns and 50 verbs to be used as vehicles.

In order to see the meanings of vehicles we conducted a psychological experiment. In the experiment, 15 Japanese males and females, aged 20 – 25, were asked to rate 150 words (50 adjectives, 50 nouns and 50 verbs) against the 9 SD scales; dark-light, dislike-like, inelegant-elegant, sad-glad, ugly-beautiful, uncomfortable-comfortable, bad-good, difficult-easy and unfamiliar-familiar. These SD scales were selected by a psychological experiment (Sumihisa et al., 2011) in which participants were asked to choose SD scales for which they can easily see one of semantic pole as positive and the other semantic pole as negative.

The ratings were made on a 7-point scale ranging from -3 through 0 to +3. We regarded the value -3 as the negative semantic pole and the value +3 as the positive semantic pole. We conducted t-tests (two-tailed, the alpha level .05) and regarded the words which have no significant difference between the mean semantic values of the words and “0” as words with neutral meanings. And the words which have significant difference between their mean value and value 0 were classified into either words with the positive meaning or those with the negative meaning.

We selected 5 vehicles with positive meaning, 5 vehicles with negative meaning and 5 vehicles with neutral

meaning to make nominal, predicative and adjective metaphors respectively.

As for nominal metaphors, nouns such as *fortune* (“kouun”), *freedom* (“jiyuu”), *justice* (“seigi”), *life* (“inochi”) and *dream* (“yume”) were selected as vehicles with positive meaning. Nouns such as *faith* (“shinkou”), *joke* (“joudan”), *patience* (“nintai”), *transient* (“mujou”) and *philosophy* (“tetsugaku”) were selected as vehicles with neutral meaning. And nouns such as *evil* (“aku”), *hell* (“jigoku”), *dissatisfied* (“fuman”), *self-preservation* (“hoshin”) and *downfall* (“metsubou”) were selected as vehicles with negative meaning.

As for predicative metaphors, verbs such as *appear* (“arawareru”), *believe* (“shinjiru”), *flutter* (“tokimeku”), *clear* (“hareru”) and *laugh* (“warau”) were selected as vehicles with positive meaning. Verbs such as *make merry* (“ukareru”), *dry* (“kawaku”), *cut fine* (“kizamu”) and *cry* (“naku”) were selected as vehicles with neutral meaning. And verbs such as *be irritated* (“iradatsu”), *doubt* (“utagau”), *remain* (“todomaru”), *betray* (“negaeru”) and *warp* (“yugamu”) were selected as vehicles with negative meaning.

As for adjective metaphors, adjectives such as *new* (“atarashii”), *sweet* (“airashii”), *cool* (“kakkoii”), *white* (“shiroi”) and *equal* (“hitoshii”) were selected as vehicles with positive meaning. Adjectives such as *black* (“kuroi”), *hard* (“katai”), *fine* (“komakai”), *long* (“nagai”) and *deep* (“fukai”) were selected as vehicles with neutral meaning. And adjectives such as *stinking* (“kusai”), *dull* (“nibui”), *worn-out* (“boroi”), *shabby* (“misuborashii”) and *disgraceful* (“mittomonai”) were selected as vehicles with negative meaning.

### Words associated from vehicles and topics

We examine what kind of meanings associated with topics or vehicles affect the comprehension of metaphors. In order to research the words associated from vehicles and topics, we conducted a pre-experiment. 30 Japanese males and females, aged 20 – 24, were asked to answer 3 or more words associated from 45 vehicles and 4 topics. We decided to use for the later experiment the words associated from vehicles and topics chosen by 2 or more participants. Then we conducted another pre-experiment to evaluate the meaning of the words associated from vehicles and topics. 60 Japanese males and females, aged 20 – 26, were classified into 2 groups. 107 or 108 words assigned to each group and participants were asked to rate the assigned words against 7 SD scales (in table 2). Based on the result of this experiment, we classified the words into the words with positive meaning, the words with negative meaning and the words with neutral meaning. We conducted t-tests (two-tailed, the alpha level .05) and regarded the words which have no significant difference between the mean semantic values of the words and “0” as words with neutral meanings. And the words which have significant difference between their mean value and value 0 were classified into either words with the positive meaning or those with the negative meaning.

## Experiment

### Metaphorical expressions

We combined vehicles and topics which were selected by pre-experiments and made nominal, predicative and adjective metaphors. Then we conducted a psychological experiment in order to see what kind of meanings associated with topics or vehicles affect the comprehension of metaphors. In the experiment, 60 Japanese males and females, aged 20 - 26, were assigned to 180 metaphorical expressions and were asked to choose words which they believe to be related to the

meaning of each metaphorical expression among those associated from vehicles and topics. Participants were also asked to rate meanings of the 180 metaphorical expressions respectively against 9 SD scales (7 SD scales given in table 2 and additional scales “difficult – easy” and “unfamiliar – familiar”).

## Results and Discussion

### Meanings evoked by metaphors

We classified metaphorical expressions into the metaphorical expressions which have vehicles with neutral meaning, positive meaning and negative meaning. Then, we analyzed the meaning of metaphorical expressions and the words associated from vehicles and topics. We conducted t-test (two-tailed, the alpha level .05) to see semantic changes evoked by the semantic interaction between topics and vehicles. Since only the topics with neutral meanings were selected through the pre-experiment, metaphorical expressions which have no significant difference between their mean value and value 0 were regarded as metaphors showing no semantic change (0). And metaphorical expressions which have significant difference between their mean values and value 0 were classified into either metaphors showing the change to the positive semantic pole or those showing the change to the negative semantic pole.

### Metaphors using vehicles with neutral meanings

Table 4 shows the number of 3 types of metaphors which show the positive, negative or neutral meanings when vehicles are neutral.

Table 4: Number of metaphors showing positive, negative and neutral meanings when vehicles are neutral

	positive	negative	neutral	sum
nominal metaphors	1	1	18	20
predicative metaphors	4	9	7	20
adjective metaphors	2	11	7	20
sum	7	21	32	60

As for the metaphors in which vehicles of their own have neutral meanings, the proportion of the metaphors showing the neutral meanings was the highest.

As for the total number, a Chi-square test was conducted among the expressions showing positive (+), negative (-), and neutral (0) meanings. As a result, there were significant differences between the number of nominal metaphors and that of predicative metaphors ( $\chi^2(1, N=40)=8.533, p<.05$ ) and also between the number of nominal metaphors and that of adjective metaphors, ( $\chi^2(1, N=40)=11.905, p<.05$ ). However, there was no significant difference between the number of predicative metaphors and that of adjective metaphors, ( $\chi^2(1, N=40)=0.400, p>.05$ ).

As for nominal metaphors, the result of Chi-square tests showed that the metaphorical expressions with neutral meaning were significantly more than the other expressions, ( $\chi^2(1)=15.211, p<.05(+ \text{ vs. } 0), \chi^2(1)=15.211, p<.05(- \text{ vs. } 0)$ ). As for predicative metaphors, there was no significant difference among the number of metaphorical expressions which showed positive meaning, negative meaning and neutral meaning, ( $\chi^2(1)=.181, p>.05(+ \text{ vs. } 0), \chi^2(1)=.250, p>.05(0 \text{ vs. } -)$ ,  $\chi^2(1)=1.923, p>.05(+ \text{ vs. } -)$ ). As for the adjective metaphors, there was significant difference among the number of metaphorical expressions which showed positive meanings and neutral meanings ( $\chi^2(1)=2.778, p<.05(+ \text{ vs. } 0)$ ), and the number of metaphorical expressions which showed positive

meanings and negative meanings, ( $\chi^2(1)=6.231, p<.05(+ \text{ vs. } -)$ ). However, there was no significant difference between the number of metaphorical expressions which showed neutral meanings and negative meanings, ( $\chi^2(1)=.889, p>.05(0 \text{ vs. } -)$ ).

These results show that nominal metaphors are basically affected by the meaning of vehicles and tend to show neutral meanings. Adjective metaphors show negative meanings, although meanings of vehicles are neutral.

### Metaphors using vehicles with positive meanings

Table 5 shows the number of metaphors which show the positive, negative and neutral meanings when vehicles are positive. As for the metaphors in which vehicles of their own have positive meanings, the proportion of the metaphors showing positive meanings was the highest. As for the total number, Chi-square tests were conducted among the expressions showing positive (+), negative (-), and neutral (0) meanings. As a result, there was no significant difference between nominal metaphors, predicative metaphors and adjective metaphors, ( $\chi^2(4, N=60)=4.034, p>.05$ ). The result shows that, as for vehicles with positive meanings, the three types of metaphors tend to show positive meanings.

Table 5: Number of metaphors showing positive, negative and neutral meanings when vehicles are positive

	positive	negative	neutral	sum
nominal metaphors	20	0	0	20
predicative metaphors	19	1	0	20
adjective metaphors	19	0	1	20
sum	58	1	1	60

### Metaphors using vehicles with negative meanings

Table 6 shows the number of metaphors which show the positive, negative and neutral meanings when vehicles are negative. As for the metaphors in which vehicles of their own have negative meanings, all the metaphors showed negative meanings.

Table 6: Number of metaphors showing positive, negative and neutral meanings when vehicles are negative

	positive	negative	neutral	sum
nominal metaphors	0	20	0	20
predicative metaphors	0	20	0	20
adjective metaphors	0	20	0	20
sum	0	60	0	60

### Words associated with metaphors

The results described so far showed that negative meanings were evoked more frequently for adjective metaphors among the other types of metaphors when vehicles were neutral. This section discusses the results of the psychological experiment in which participants were asked to choose words related to meanings of metaphorical expressions among those associated from vehicles and topics. We want to see, even if negative and positive exemplars were associated with vehicles or topics, exemplars with negative meanings tend to be selected to process meanings of adjective metaphors.

### Nominal metaphors

The second left column of Table 7 shows the total number of the words associated from vehicles or topics and the second right column the number of words which participants selected as those related to meanings of nominal metaphors.

Table 7: the number of the words which were associated from vehicles or topics

	all words	selected words	rate
vehicles	810	749	92.50%
topics	735	348	47.30%

A Chi-square test was conducted among the words which were associated from vehicles and topics. The result showed that words associated with vehicles were selected significantly more frequently than those with topics, ( $\chi^2(1, N=1545)=381.063, p<.05$ ).

We examined the frequency in which positive, neutral or negative words associated with vehicles or topics were selected by participants when they process the meanings of nominal metaphors created from vehicles with positive meaning. Table 8 shows the results. The result of Chi-square tests showed that the associative words with positive meaning were significantly more frequently selected than the others, ( $\chi^2(1, N=436)=76.926, p<.05 (+ vs. 0), \chi^2(1, N=431)=104.972, p<.05 (+ vs. -)$ ).

Table 8: the number of the associative words when vehicles have positive meaning

	all words	selected words	rate
positive	336	284	84.50%
neutral	100	41	41.00%
negative	95	30	31.60%

Table 9 shows the frequency in which positive, neutral or negative words associated with vehicles or topics were selected by participants when they process meanings of nominal metaphors created from vehicles with neutral meaning.

Table 9: the number of the associative words when vehicles have neutral meaning

	all words	selected words	rate
positive	242	195	80.60%
neutral	119	90	75.60%
negative	130	87	66.90%

As for the number of the associative words when vehicles have neutral meaning, there was no significant difference among each number of the associative words with positive, neutral and negative meaning, ( $\chi^2(1, N=361)=1.175, p>.05 (+ vs. 0), \chi^2(1, N=249)=2.292, p>.05 (0 vs. -), \chi^2(1, N=372)=8.598, p=.05 (+ vs. -)$ ).

When the vehicles have negative meaning, table 10 shows the result of the number of the associative words. The results of Chi-square tests showed that the associative words with negative meaning were significantly more frequently selected than the others, ( $\chi^2(1, N=348)=29.264, p<.05 (0 vs. -), \chi^2(1, N=392)=106.940, p<.05 (+ vs. -)$ ).

Table 10: the number of the associative words when vehicles have negative meaning

	all words	selected words	rate
positive	167	73	43.70%
neutral	123	85	69.10%
negative	225	206	91.60%

The results so far suggest that the comprehension of nominal metaphors is basically affected by the prototypical exemplars associated with vehicles.

## Predicative metaphors

The second left column of Table 11 shows the total number of the words associated from vehicles or topics and the second right column the number of words which participants selected as those related to meanings of predicative metaphors.

Table 11: the number of the words which were associated from vehicles or topics

	all words	selected words	rate
vehicles	715	644	90.10%
topics	735	403	54.80%

The result of Chi-square tests showed that words associated with vehicles were selected more frequently than those with topics, ( $\chi^2(1, N=1450)=224.275, p<.05$ ).

We examined the frequency in which positive, neutral or negative words associated with vehicles or topics were selected by participants in the same way as nominal metaphors. Table 12 shows that in the comprehension of predicative metaphors created from vehicles with positive meaning words with positive meanings were significantly more frequently selected than the others, ( $\chi^2(1, N=409)=44.675, p<.05 (+ vs. 0), \chi^2(1, N=412)=69.834, p<.05 (+ vs. -)$ ).

Table 12: the number of the associative words when vehicles have positive meaning

	all words	selected words	rate
positive	318	272	85.50%
neutral	91	48	52.70%
negative	94	41	43.60%

Table 13 shows the frequency in which positive, neutral or negative words associated with vehicles or topics were selected by participants when they process meanings of predicative metaphors created from vehicles with neutral meaning. The results of Chi-square tests showed that there was no significant difference among each number of the associative words with positive, neutral and negative meaning, ( $\chi^2(1, N=387)=.000, p>.05 (+ vs. 0), \chi^2(1, N=233)=.104, p>.05 (0 vs. -), \chi^2(1, N=384)=.131, p>.05 (+ vs. -)$ ).

Table 13: the number of the associative words when vehicles have neutral meaning

	all words	selected words	rate
positive	269	194	72.10%
neutral	118	85	72.00%
negative	115	85	73.90%

Table 14 shows the frequency in which positive, neutral or negative words associated with vehicles or topics were selected by participants when they process meanings of predicative metaphors created from vehicles with negative meanings. The results of Chi-square tests showed that words with negative meaning were selected significantly more frequently than the others, ( $\chi^2(1, N=245)=25.136, p<.05 (0 vs. -), \chi^2(1, N=352)=40.022, p<.05 (+ vs. -)$ ).

Table 14: the number of the associative words when vehicles have negative meaning

	all words	selected words	rate
positive	194	118	60.80%
neutral	87	56	64.40%
negative	158	143	90.50%

These results for predicative metaphors suggest that the

comprehension of predicative metaphors is basically affected by the exemplars associated with vehicles.

### Adjective metaphors

Table 15 shows the total number of the words associated from vehicles or topics and the number of words which participants selected as those related to meanings of adjective metaphors.

Table 15: the number of the words which were associated from vehicles or topics

	all words	selected words	rate
vehicles	724	627	86.60%
topics	735	444	60.40%

The result of Chi-square tests showed that words associated with vehicles were selected more frequently than those with topics, ( $\chi^2(1, N=1459)=128.193$ ,  $p<.05$ ).

Table 16 shows the frequency in which positive, neutral or negative words associated with vehicles or topics were selected by participants when they process meanings of adjective metaphors created from vehicles with positive meanings. Chi-square tests showed that words with positive meaning were selected significantly more frequently than the others, ( $\chi^2(1, N=400)=32.967$ ,  $p<.05$  (+ vs. 0),  $\chi^2(1, N=408)=141.638$ ,  $p<.05$  (+ vs. -)).

Table 16: the number of the associative words when vehicles have positive meaning

	all words	selected words	rate
positive	325	283	87.10%
neutral	75	44	58.70%
negative	83	19	22.90%

Table 17 shows the frequency in which positive, neutral or negative words associated with vehicles or topics were selected when participants process meanings of adjective metaphors created from vehicles with neutral meanings. Although vehicles were neutral, words with negative meaning were selected significantly more frequently than the others, ( $\chi^2(1, N=262)=7.162$ ,  $p<.05$  (0 vs. -),  $\chi^2(1, N=360)=9.089$ ,  $p<.05$  (+ vs. -)).

Table 17: the number of the associative words when vehicles have neutral meaning

	all words	selected words	rate
positive	227	161	70.90%
neutral	129	92	71.30%
negative	133	113	85.00%

Table 18 shows the frequency in which positive, neutral or negative words associated with vehicles or topics were selected by participants when they process meanings of adjective metaphors created from vehicles with negative meanings. Chi-square tests showed that words with negative meaning were selected significantly more frequently than the others, ( $\chi^2(1, N=277)=27.380$ ,  $p<.05$  (0 vs. -),  $\chi^2(1, N=378)=30.959$ ,  $p<.05$  (+ vs. -)).

Chi-square tests showed that words with negative meaning were selected significantly more frequently than the others, ( $\chi^2(1, N=277)=27.380$ ,  $p<.05$  (0 vs. -),  $\chi^2(1, N=378)=30.959$ ,  $p<.05$  (+ vs. -)).

These results suggest that adjective metaphors are different from nominal and predicative metaphors in the comprehension where words with negative meanings tend to be selected although vehicles themselves are neutral.

Table 18: the number of the associative words when vehicles have negative meaning

	all words	selected words	rate
positive	199	128	64.80%
neutral	98	61	62.20%
negative	179	159	88.80%

### Conclusion

This study explored the reason why adjective metaphors evoke negative meanings more frequently than the other types of metaphors. The results showed that exemplars with negative meanings among various exemplars tend to be selected to process meanings of adjective metaphors. This result suggests that, unlike nominal metaphors processed by the categorization theory, adjective metaphors are processed by the two-stage categorization theory (Utsumi & Sakamoto, 2007), in which exemplars with negative meanings are selected among various exemplars belonging to the intermediate category evoked by adjectives as vehicles. We still do not know why exemplars with negative meanings are used to process meanings of adjective metaphors.

### Acknowledgments

This study was supported by a Grant-in-Aid for Scientific Research B (No.23300098) from the Japan Society for the Promotion of Science.

### References

- Bowdle, B., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, 112(1), 193-216.
- Glucksberg, S. (2001). *Understanding figurative language: From metaphors to idioms*. Oxford: Oxford University Press.
- Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, 97, 3-18.
- Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. Chicago: The University of Chicago Press.
- Osgood, C. E. (1980). The cognitive dynamics of synesthesia and metaphor. In R. P. Honeck & R. R. Hoffman (Eds.), *Cognition and figurative language* (pp. 203-238). Lawrence Erlbaum Associates.
- Sakamoto, M., & Utsumi, A. (2009). Cognitive effects of synesthetic metaphors evoked by the semantic interaction. In *Proceeding of the 31<sup>st</sup> Annual Meeting of the Cognitive Science Society*, 1593-1598.
- Sumihisa, M., Tsukurimichi, H., Utsumi, A., & Sakamoto, M. (2011). Is Evoking Negative meanings the Unique Feature of Adjective metaphors?. In *Proceedings of the 33<sup>rd</sup> Annual Meeting of Cognitive Science Society*, 2655-2660.
- Ullmann, S. (1951). *The principles of semantics*. Oxford: Blackwell.
- Utsumi, A., Sakamoto, M. (2007). Computational evidence for two-stage categorization as a process of adjective metaphor comprehension. In *Proceeding of the 2<sup>nd</sup> European Cognitive Science Conference*, 77-82.
- Werning, M., Fleischhauer, J., & Beşeoğlu, H. (2006). The cognitive accessibility of synaesthetic metaphors. In *Proceedings of the 28<sup>th</sup> Annual Conference of the Cognitive Science Society*, 2365-2370.
- Yamaguchi, T. (2003). *Nihongo Dai-Thesaurus (Japanese Thesaurus)*. Tokyo: Taishukan Shoten. (in Japanese).
- Yu, N. (2003). Synesthetic metaphor: A cognitive perspective. *Journal of Literary Semantics*, 32(1), 19-34.



# Problem-Solving Strategy Selection in Relation to Formal Schooling

**Mennat-Allah Saleh (mennat-allah.saleh@student.guc.edu.eg)**

Department of Media Engineering and Technology

German University in Cairo

New Cairo City -Main Entrance Al Tagamoa Al Khames, Egypt

**Christian Sturm (christian.sturm@guc.edu.eg)**

Department of Media Engineering and Technology

German University in Cairo

New Cairo City -Main Entrance Al Tagamoa Al Khames, Egypt

## Abstract

A study of the literacy-generated cognitive cultural gap was carried out on subjects of different literacy background ranging from illiterate individuals to university students in different majors. The characteristics that aid literate and illiterate people in solving mathematical problems efficiently were identified and analyzed. A field research was carried out in the field of algorithmic problem solving and in the reasoning domain, followed by constructing a software cognitive model to represent the findings. Findings showed that in both domains cognitive ability did not improve with level of literacy, rather the formality of the problem solving strategy selected demonstrating a link between these two domains.

**Keywords:** Cognitive Psychology, Cognitive Modeling, Problem Solving, Literacy, Deductive Reasoning.

## Introduction

The interaction with illiterate people is a common experience in Egypt. More than one third of the adult population is not able to read and write (UNICEF, 2012). One could conclude that due to these every-day encounters, illiterates would be appreciated as they represent an important part of Egyptian's workforce and are preservers of Egyptian's rich cultural heritage. The Egyptian society, however, marks clearly its division not only based on economic power. Throughout the country, illiteracy is commonly associated with a lack of mental and cognitive capabilities that leads to a tremendous depreciation of this group by the literate part of the society (Hollingshead, 1975). Given the nature of this stereotype, the area of problem-solving strategies has been chosen for comparison. The motivation of the research is to explore this cultural gap and find ways to bridge it.

Several approaches to cultural differences in problem-solving have been explored in the past. Tedre et al. have shown how computation is done across different cultures and looked at the education of computer science students (Tedre, Sutinen, Kahkonen, & Kommers, 2003). Gerdes looked in a similar way at the cultural differences on math and mathematical problem solving (Gerdes, 2005). Several researches suggest that human reasoning is based on building personalized mental models; hence using mental models in attempting to formalize and represent reasoning is a valid approach (Knauff, Mulack, Kassubek, Salih, & Greenlee, 2002). In addition to formulating the models, the field of cognitive psychology also draws attention to the basic thinking principles

that are present in all individuals, but are used differently across cultures. It was used in combination with cross cultural psychology, which is the study concerned with the thinking principles that are generated from cultural differences (Adler & Gielen, 2001). This field has been used by many researches to study how education affects cognitive behaviors, memory, problem solving and logical reasoning (Segall, Dasen, Berry, & Poortinga, 1999).

Based on this previous work and the given circumstances in Egypt, it was decided to look at reasoning and algorithmic problem solving tasks together with the mental models created during the process for both literate and illiterate subjects including different educational domains.

## Problem Space

### Domain Selection

It was hypothesized that the level of a person's formal schooling is related to the level of formality regarding their approach to the selection of problem solving strategies. The functionally illiterate subjects in this context are compared to the formal schooling subjects as having a lower level of schooling formality (Kosmidis, Zafiri, & Politimou, 2011). Hence, domains that need a formal and strategic approach were selected for this research. The goal was to identify and quantify problem-solving strategies easily. Numeric problem solving, algorithmic problem solving and reasoning meet this requirement. As the relationship between literacy and numeracy has been widely studied in this context already, the two latter domains were selected.

### Algorithm Domain

The main focus while testing the algorithmic domain was to identify the subjects strategies in conjunction with the level and type of education while solving a procedural algorithmic problem. The subjects were asked for a self-report on their strategies. This report was contrasted with the steps they actually took in solving the problem. The Towers of Hanoi problem was used to test this domain due to its strong mathematical basis and the possibility to easily adapt it to different cultural contexts. This was specifically important for the illiterate subjects that are not used to any kind of formal testing situations encountered in the lab. The most important characteristic of the Towers of Hanoi consists in the variety of

correct strategies to solve it. Hence, this did not limit the subjects options to only one correct way, but would allow them to select the strategy they are most comfortable with based on their personal literacy backgrounds (Gunzelmann & Anderson, 2001). The expected problem solving strategies were: analogy, divide and conquer, mean-ends analysis, trial and error, random strategy, research and working backwards (Chiew & Wang, 2004). In addition to the problem solving strategy, the mental model created by the solver during the test was of special interest, too. It was expected that solvers would have one of the following mental models:

- Formal Representation: Using logic and formal mathematical representation for the problem.
- Previously-Prepared Mental Model: Solver solves the entire problem in their mind before solving on board.
- On-the-go Mental Model: Subjects build an image of the disks in their mind and use it to solve the problem virtually before physically, a few steps at a time.
- No Representation

A dynamic, additional element in form of a random disk was introduced by the test facilitator during the test in order to be able to distinguish between subjects that would apply a previously prepared model and subjects that would create the models on-the-go. The subject was asked to proceed with the solution using the additional disk. This procedure provided the opportunity to check, in addition, the degree of agility of the subjects strategy as well as their understanding of the problem. The point in time as well as the type of disk added was determined by the facilitator during the test based on the subject's performance.

## Reasoning Domain

Reasoning was used as a problem solving domain in this research based on the previous work of Tulviste et al. (1978). They have shown that reasoning is a skill that improves strongly with the level of literacy (Tulviste, Riikliku, & Toimetised, 1978). Several types of reasoning were considered for this study. Syllogistic deductive reasoning was selected as studies have shown that it is a skill that comes with formal schooling and evolves to be used in everyday life on different types of problems. In addition, the mental proof theory shows that syllogistic deductive reasoning is approached by solvers using one of the following techniques (Knauff et al., 2002; Rips, 1994): spatial reasoning, visual reasoning or reasoning using formal logic, entailing a variety of strategies with different levels of formality that relates to the level of the subjects formal schooling. The Zebra puzzle was suggested as a research test question (Stangroom, 2010). The approach needed to solve the problem, however, requires that the subjects use a formalized written schedule. This would have forced them to select one strategy over the others for correctness instead of their own personal preference. Moreover, it would have meant that illiterate subjects would have

never been able to solve it as they are not able to use pen and paper for their solution. Therefore, a smaller version of the puzzle with only 3 instead of 5 variables was created as stated below:

A street has 3 houses, each house has a different color, and in each house the owner has a different nationality and owns a different pet. Given these following clues, what is the color of the fish's house?

- The cat lives in the center house.
- The green house is on the left.
- The French lives in the blue house.
- The German owns a dragon.
- The Egyptian lives in the red house directly to the right of the dragon.

Subsequently, the subjects were asked to answer the following question: what is the color of the fish's house?

The spatial memory is used for solving the Towers of Hanoi Problem while deductive syllogistic reasoning tasks tend to make use the verbal memory (Handley, Capon, Copp, & Harper, 2002). Studies have shown that the performance of both types of memory is independent from each other. Measuring the performance, however, was not the main objective to measure in this study but rather the strategy selection. This selection, as mentioned before, is hypothesized to depend on the subjects formal schooling independent from the type of memories. Therefore, it is expected that the subjects show a similar selection of strategies for both domains chosen here.

## Hypothesis and Test Design

### General Hypothesis

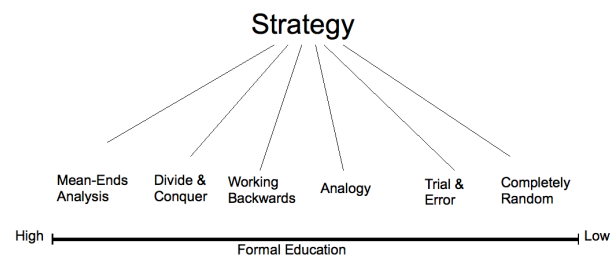


Figure 1: Towers of Hanoi strategy selection tree

It was hypothesized that the strategy selected for the Towers of Hanoi would be based on the formality of the subject's education as shown in figure 1. In addition, it was hypothesized that the subjects may move between different strategies when solving the problem. This shift, however, would be minimal as the general tendency is expected to be clear and stable.

## Test Design

The following profiles were defined for the test subjects in order to test the hypotheses mentioned above:

- 20 computer science university students, who received training in formal logic.
- 20 Illiterate subjects, who are now in their early stages of literacy classes. The short class-room experience equipped them with the necessary preparation and tolerance towards participating in the experiments.
- 20 applied arts university students, who have only received formal schooling but no training in formal logic. They have, however, strong spatial reasoning skills.

The content of the test has been adapted to Egyptian illiterate. The deductive reasoning problem became the narration of a story replacing nationalities with common names and pets with farm animals. The Towers of Hanoi problem was contextualized by representing the disks with water buckets of different sizes and the pegs with floor tiles. The explanation of the problem was based on a story to justify the reason for moving the buckets.

During the test, the subjects' age, study-major (if applicable) and duration in literacy class (if applicable) were captured. This data defined the independent variables as follows: Level of literacy, type of education and training in formal logic. In addition, the subjects were asked questions regarding their approach, mental representations and changes in strategies. The strategy was classified as either: none, on-the-go or prepared beforehand. The following dependent variables were recorded: completion of task (y/n), time for task completion, number of attempts in deductive reasoning, approach to deductive reasoning, Towers of Hanoi strategy selection, Towers of Hanoi strategy change, Towers of Hanoi adaptation to dynamic disk addition.

## Subject-Specific Hypothesis

Figure 2 and tables 1 and 2 represent the three hypotheses of each subject background.

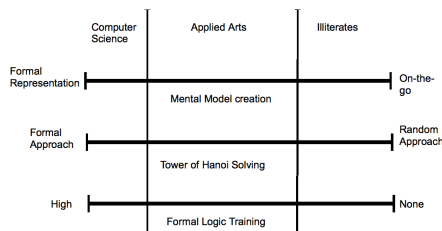


Figure 2: Towers of Hanoi Hypothesis

## Test Results and Discussions

### Pretest

A pretest was conducted using the above test design on engineering students who have studied the Constraint Program-

Subject	Mental Representation
Computer Science	Numerical and Formal
Applied Arts	Visual model
Illiterates	Visual model or none

Table 1: Towers of Hanoi Mental Model Hypothesis

Subject	Approach	Memory
Computer Science	Formal Logic	Verbal Memory
Applied Arts	Spatial Reasoning	Spatial Memory
Illiterates	Intuition	N/A

Table 2: Deductive Reasoning Hypothesis

ming course, applied arts students and Illiterate subjects. No modifications were recommended for applied arts and illiterate tests. . All of the engineering students, however, were already familiar with the solution to the Towers of Hanoi problem. Therefore, another control group with a lower level of formal training was added. This new group consisted of computer science students in their sophomore year. The results recorded for the deductive reasoning question were observed by asking the subject questions about their strategy and deducing their approach from the paper they used in solving. After the test, it was concluded that there is no observed difference when using spatial or visual reasoning and that both strategies may have different mental representations, but they use the same inference technique.

## The Main Test Results

The Towers of Hanoi problem was analyzed in two steps: the observation and the inference. The observed results are the strategies recognized by the facilitator. They were recorded because it was difficult for some of the subjects to properly explain their own approach. They included:

- The Random strategy was recorded when the subject showed no learning pattern and moved the disks haphazardly often reaching no correct solution.
- The Trial and Error strategy was recorded when the subject first started moving disks randomly and then appeared to recognize incorrect moves and avoided repeated them, this was often observed when the board reaches a similar state to a previously encountered stage and the subject seemed to recall their previous error.
- The Wrong Towers strategy was recorded when the subject used the right pattern of movements but choose the wrong destination for the movement of the first disk.
- The Correct strategy was recorded when the subject used the right pattern of movements from the first move without any errors.

The inferred results were reported by the subjects' own explanation of their strategy. They were used in addition to the observed strategies for the final analysis and included:



- Random strategy was recorded when the subject says that they did not understand the game and they just moved the disks in any fashion attempting to reach a solution.
- Trial and Error strategy was recorded when the subject says that at first they started without knowing how to proceed, but with time they understood the game and learned from their errors.
- Pattern Recognition occurs when the subject identifies patterns. A pattern is a group of disks stacked in a certain way on the board. The pattern recognized by subjects were either the 3-disk pattern, disks stacked on top of each other or the flat pattern. The flat pattern means that three disks are placed on the board, while each disk is placed on a different tower. The subjects would always follow a learned set of moves upon identifying any of the patterns.
- Mean Ends Analysis was recorded when the subject says they solved the problem by breaking it down and attempting to move the largest disk to the destination first, which required moving all the disk stacks above it to the intermediate tower and so forth. This strategy was recorded when the subjects identified the recursive nature of the problem.

The following are the observed and inferred results for each group of test subjects:

- Computer science students trained in constraint programming were almost equally divided between using the formal approach (12) and spatial reasoning (8) for the deductive reasoning problem as opposed to the original hypothesis of a high tendency for formal approach. Those who choose a formal approach tended to use mean-ends analysis to solve the Towers of Hanoi before and after adding the dynamic disk. The ones that used spatial reasoning were divided between trial and error and mean-ends analysis before adding the disk and continued to use their selected strategy after addition of the disk.
- Engineering subjects tended to choose spatial reasoning (17) over formal approach (3) for solving the deductive reasoning problem, as opposed to the hypothesis that they would have a tendency for a formalized approach. For the Towers of Hanoi, most of them used mean-ends analysis before disk addition. Only a few used pattern recognition, too. The majority that used mean-ends continued after the disk addition with mean-ends, whilst only a few applied Trial and Error. Those who selected pattern recognition were equally divided upon the four strategies after disk addition.
- Applied Arts subjects mostly selected spatial reasoning (18) as hypothesized. A surprising 10%, however, approached the deductive reasoning problem with a formal strategy (2). The applied arts subjects were the most diversified in their strategy selection showing that their education did not seem to limit their approach. They were

equally divided upon all strategies before and after disk addition except for the unselected random strategy.

- As hypothesized, 100% of the illiterate subjects (20) used spatial reasoning to solve the deductive reasoning problem. The subjects were evenly divided between all the four strategies of the Towers of Hanoi before disk addition. This contradicted the initial belief that all illiterates would approach the problem in a random way. Upon dynamic disk addition, most of the subjects would either continue using their current strategy or use a strategy that is one degree less formal according to the hierarchy shown in figure 3. The results clearly indicate that illiterates were affected by the disk addition process.

### Statistical Results Analysis

The different strategies were numbered according to the pyramid shown in figure 3. In order to determine the significance of the results, the Kruskal Wallis test was used.

Subject	N	Min	Max	Median
Constraint Programming	20	4.0	2.0	4.0
Arts	20	2.0	4.0	2.0
Engineering	20	3.0	4.0	4.0
Illiterates	20	1.0	4.0	2.0

\*p < .01

Table 3: The statistical analysis of all groups before addition

H=12.65\*

Subject	N	Min	Max	Median
Constraint Programming	20	4.0	2.0	4.0
Arts	20	2.0	4.0	2.0
Engineering	20	2.0	4.0	4.0
Illiterates	20	2.0	4.0	2.0

\*p < .01

Table 4: The statistical analysis of all groups after addition

H=12.30\*

Both tests were statistically significant at a 1% level of significance. Therefore, we can state that the literacy and formal schooling background significantly affects the selection of problem-solving strategies when solving the Towers of Hanoi problem. In addition and in line with the initial assumption, there was no correlation found between the task performance and the level of formality of the selected strategy ( $p = .08$ ). This means that the use of formal methods did not lead to a faster performance in comparison to using spatial reasoning. Finally, the effect of literacy on the type of mental model created in the Towers of Hanoi problem was tested. The results showed no statistical significance ( $p = .94$ ). This shows that all four subject domains created spatial, beforehand or no mental models depending on a different parameter other than their

educational backgrounds. It also showed that strategy selection in the Towers of Hanoi problem is not entirely dependent on the mental model created as opposed to the original hypothesis, since the selection of the strategy proved to be dependent on the subject's educational background.

The original hypothesis of the tree of problem selection strategies was not confirmed. It was rather observed that the strategy selection was done using the pyramid shown in figure 3. The strategies are ranked according their proximity to the most optimum Towers of Hanoi algorithm with one being the least optimum. It was observed that subjects would start at a level of the above pyramid depending on their level and type of formal education, their working background and their understanding of the problem and rules. Subjects go up the pyramid but never down. No subject was observed to move 2 levels up the pyramid, they only move up 1 level per game or none at all. Once an additional disk was introduced, however, most subjects would move down the pyramid considering the new situation a new problem. Some of the subjects did not even capitalize on the previous knowledge they obtained before the addition of the new disk.

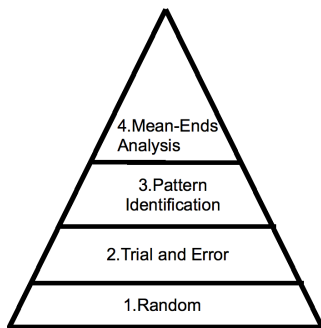


Figure 3: Towers of Hanoi Observations

## Software Model

As a first step towards possible future prediction of cognitive behaviors of subject profiles, a software model was used to represent the strategies observed. It was implemented to formalize and represent the test results and experiment with new strategies. It was also implemented as a guidance for predicting problem solving strategy selection and execution based on a given subject profile. Having such a tool can help educational institutes in understanding the mindset of their target groups as well as guide them through their cognitive development needs which will assist in curricula formation. Four agents were created, one for each subject background and the agents navigated between the strategies implemented below according to the percentage of subjects within that background recorded to use that strategy. The model was created on Java as its object-oriented nature facilitated the implementation and its high-level characteristic allowed enough abstraction to represent only the details examined during the tests.

## Modeling Reasoning

The reasoning problem strategies were modeled separately for formal representation and spatial reasoning.

Formal reasoning was modeled using Constraint Programming. Constraint Programming (CP) is a recent paradigm based on artificial intelligence and declarative programming (Rossi, Beek, & Walsh, 2006). Its advantage over other programming methodologies is its abstraction. The houses are represented as a data domain from 1 to 3 and the pets, nationalities and colors as sets of variables that associate with these domains. By placing the puzzle restrictions above on the representation, the CP solver finds the house that has no pet and allocates the fish to it.

The spatial reasoning model was simulated the way the solvers described their approach. A list of possible streets was kept in the subject's brain at a time. Each street had its unique configuration of houses. Upon presenting the subject with an additional clue from the puzzle, the subject attempts to add this to his current streets list. A clue can either be added to an existing street, if it can be merged with one, delete an existing street, if it presents a contradiction showing the initial street was not correct, or create a new street, if that clue cannot be represented sufficiently within the existing context. At the end of the puzzle, the subjects have eliminated all incorrect streets and are left with one street.

## Modeling Towers of Hanoi

The following four strategies are modeled separately:

- **Random Strategy:** Each move entails moving a randomly selected disk to a random tower.
- **Trial and Error:** The trial and error strategy implementation was divided into two modules: representing memory and using learning, and representing how the subject learned from their movements. Only the first module was implemented and the second was left as future work. The approach represented the subject's memory as a pre-set number of movements that they have executed. The memory starts as initially empty. Before undergoing any movement, the agent checks if the current game uses a configuration that the subject has already learnt, in this case it switches to the pattern identification strategy below. If the game represents a pattern that the subject did not learn yet, it marks the current move. The subject selects a disk using a random generator in this case. After the second module mentioned above will be implemented, however, it will be constricted by learning. Once a move is executed, it is saved into the subject's memory. If there is no space in the memory, the oldest move is deleted. Now, the agent checks upon the last sequence of moves starting from the marked one. If they have solved the pattern correctly up to the pattern threshold, then the subject learned the pattern. A pattern threshold is the initial number of difficult moves the subject needs to do to understand a pattern. The rest of the moves then come easily as described and observed during tests.

- Pattern Identification: The 3-pattern and flat pattern as described in Test Results and Discussions section are implemented. The first disk destination is selected randomly.
- Mean-Ends Analysis: The algorithm moves all disks smaller than the  $n$ th disk to the intermediate tower. The  $n$ th disk is then moved to the destination tower. Finally, all the disks at the intermediate tower are moved on top of the  $n$ th disk at the destination using the same fashion.

## Conclusion

This research has tested the hypothesis that *cognitive ability improves with literacy* is a common misconception. It is rather the case that the selection of given strategies is affected by the literacy background.. The hypothesis was proven correct in both the domain of reasoning and algorithmic problems. The relationship between these two domains was also examined in this research. It was hypothesized that individuals who select a formal problem solving strategy in one of them would select a formal one in the other. This formality would be directly traced back to their level of formal schooling. The hypothesis was confirmed. It was shown that, despite the fact that deductive reasoning is located in a different part of the working memory than algorithmic problem solving and that both cognitive processes are completely independent, the level of formal schooling equally impacts both.

A software model was based on the findings of the research and represented agents from the following literacy backgrounds: Illiterates, Computer Science, Constraint Programming and Applied Arts students. The agents used the different cognitive approaches of each subject domain to solve the Towers of Hanoi problem and a deductive syllogistic reasoning problem adapted from the Zebra puzzle.

## Future Work

The software model can be enhanced by adding the second element of the Trial and Error learning system using reinforcement learning technique. It may also be broadened after adapting more tests to present a general thinking pattern in the domains of algorithmic problem solving and reasoning by the agents. This would mean that the model, given any newly introduced problem within these two domains, would be able to simulate the behavior of a given subject's profile. Finally, the model may be built on a pre-existing cognitive framework which will improve its accuracy and make it more representative of the human brain.

This model can be used to enhance the understanding of the different cognitive profiles in the Egyptian society. It can be used by Education Scientists to understand the cognitive gap generated by the literacy cultural gap and make use of the country's "cultural capital" as referred to by Pierre Bourdieu (Bourdieu, 1984). This capitalization can best be practiced by analyzing the reason behind these cognitive gaps and adapting the school, university as well as literacy classes curricula to make use of this information. In addition, the newly improved model can be used to test how students would think

and adapt to different problems presented to them to predict their understanding and cognitive behavior.

## References

- Adler, L., & Gielen, U. (2001). *Cross-cultural topics in psychology*. Westport, CT: Praeger.
- Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste*. Harvard University Press (Cambridge, Mass.).
- Chiew, V., & Wang, Y. (2004, aug.). Formal description of the cognitive process of problem solving. In *Cognitive informatics, 2004. proceedings of the third IEEE international conference on* (p. 74 - 83).
- Gerdes, P. (2005). Ethnomathematics as a new research field, illustrated by studies of mathematical ideas in african history. *Philosophia Mathematica III*, 13, 135161.
- Gunzelmann, G., & Anderson, J. R. (2001). An ACT-R model of the evolution of strategy use and problem difficulty. In *Proceedings of the fourth international conference on cognitive modeling* (p. 109-114).
- Handley, S. J., Capon, A., Copp, C., & Harper, C. (2002). Conditional reasoning and the tower of hanoi: The role of spatial and verbal working memory. *British Journal of Psychology*, 93(4), 501-518.
- Hollingshead, A. B. (1975). Four factor index of social status. *Unpublished manuscript, Department of Sociology, Yale University, New Haven, CT.*
- Knauff, M., Mulack, T., Kassubek, J., Salih, H. R., & Greenlee, M. W. (2002). Spatial imagery in deductive reasoning: a functional MRI study. *Cognitive Brain Research*, 13(2), 203 - 212.
- Kosmidis, M. H., Zafiri, M., & Politimou, N. (2011). Literacy versus formal schooling: Influence on working memory. *Archives of Clinical Neuropsychology*, 26(7), 575-582.
- Rips, L. (1994). *The psychology of proof*. MIT Press, Cambridge, MA.
- Rossi, F., Beek, P. v., & Walsh, T. (2006). *Handbook of constraint programming (foundations of artificial intelligence)*. New York, NY, USA: Elsevier Science Inc.
- Segall, M., Dasen, P., Berry, J., & Poortinga, Y. (1999). *Human behavior in global perspective*.
- Stangroom, J. (2010). *Einstein's riddle: Riddles, paradoxes and conundrums to stretch your mind*. Allen & Unwin.
- Tedre, M., Sutinen, E., Kahkonen, E., & Kommers, P. (2003, aug.). Appreciating the knowledge of students in computer science education in developing countries. In *Information technology: Research and education, 2003. proceedings. itre2003. international conference on* (p. 174 - 178).
- Tulviste, P., Riikliku, T., & Toimetised, U. (1978). On the origins of theoretical syllogistic reasoning in culture and in the child. , 474, 3-22.
- UNICEF. (2012, 30 January). Unicef statistics [Computer software manual]. [http://www.unicef.org/infobycountry/egypt\\_statistics.html](http://www.unicef.org/infobycountry/egypt_statistics.html).

# Shared Book Reading between Mother and Infant Facilitates The Frequency of Joint Attention

Ayumi Sato (ayusatotenjin@gmail.com)

Department of Psychology, Graduate School of Letters  
Doshisha University, Japan

Ichiro Uchiyama (iuchiyam@mail.doshisha.ac.jp)

Faculty of Psychology, Doshisha University, Japan

## Abstract

This study examines the effect of shared book reading on mother-infant joint attention interactions in infancy. In experiment 1, pairs composed of 9-month-old infants and their mothers ( $N = 10$ ) were observed in three conditions: the shared-book, toy-play, and no-material condition. The results indicate a frequency of passive joint and coordinated joint attention in the shared book context than in others. Experiment 2 longitudinally investigated the effect of increasing the time of shared book reading on the frequency of passive and coordinated joint attention. Twenty-eight pairs of 9-month-old infants and their mothers were randomly assigned to one of two groups; the first was a shared book reading condition ( $N = 11$ ), in which mothers were asked to share books every day and were given picture books regularly from the first observation (at 9 months) until the second observation (at 12 months). In the control group ( $N = 11$ ), mothers were given no instruction. The results show that increasing shared book reading increases the frequency of passive joint attention. Therefore, it is suggested that shared book reading increases joint attention episodes and that repeated shared book reading increases it in other contexts.

**Keywords:** shared book reading; joint attention; mother-infant interaction; infancy; longitudinal study

## Introduction

Shared book reading at home has been advocated as a method of enhancing children's language abilities. In 1992, Bookstart was started by Booktrust in Birmingham, U.K. This program distributes picture books and booklets for shared book reading to infants and their mothers. One of its goals is promoting pre-school child literacy; thus, follow-up surveys on literacy changes in children whose mothers received Bookstart packs were conducted to evaluate the project. For example, the timing of start of the shared book reading was reported as the strongest predictor variable in the literacy of two-year-olds (DeBaryshe, 1993). It was suggested that the earlier mothers and children started sharing picture books, the better the children were at comprehension and speaking (Payne, Whitehurst, & Angell, 1994). Wade & Moore (1998) reported that children whose mothers received Bookstart packs were better at not only reading and writing skills but also mathematics than were children whose mothers had not. Many studies agree that shared book reading promotes children's language skills.

The mechanism of shared book reading's enhancements has not been clear, but one factor is thought to be joint

attention. Tomasello (1995) argues that joint attention is characterized by the coordination of attention among the self, the other, and some external object or event. Shared book reading is thought to be an exceptional opportunity for the occurrence of joint attention (e.g., Karrass, VanDeventer, & Braungard-Rieker, 2003). It is known that joint attention plays an important role in children's language acquisition (e.g., Tomasello & Farrar, 1986). Repeated joint attention between mothers and infants makes mother-infant interactions predictable for infants and renders mapping words on the world easier (Bruner, 1985). As shared book reading provides more episodes of joint attention, it is thought to promote child literacy.

However, whether joint attention episodes occur more often during shared book reading than in other contexts has not been clear. Although many studies have used the shared book context to observe joint attention (e.g., Fletcher, Perez, Hooper, & Claussen, 2005) and have used both picture books and toys at once (e.g., Bakeman & Adamson, 1984; Mundy & Gomes, 1998), few studies have discussed shared book reading separately or compared its joint attention effects with those in other contexts (e.g., toy play).

The few extant studies involve infants older than one year. For example, Yont, Snow, & Vernon-Feagans (2003) compared mothers' utterances to their 12-month-old infants during shared book events with those during toy play and found that the number of utterances about objects made during the mothers' shared attention with their infants was higher in the shared book context than in the toy play one. Another study found that mothers and their 18-month-olds pointed more often in the shared book context than in the wooden block play one (Sugai, Akita, Yokoyama, & Nozawa, 2010). Therefore, shared book reading increased mother-infant joint attention interactions when the infants were older than 12 months.

However, whether shared book reading promotes the number of joint attention episodes in infancy hasn't been confirmed. It is in infancy, especially at around 9 months, that a child's joint attention capacity develops most rapidly (Dunham & Moore, 1995). It is thus necessary that the frequency of joint attention in infancy be compared to shared book reading and other play contexts. Furthermore, an important question is whether continued shared book reading changes the joint attention in mother-infant interaction. If shared book reading facilitated children's

joint attention abilities, joint attention frequency would increase in the other contexts (e.g., a toy-play context) through repeated shared book reading. To investigate this possibility, we need to increase the shared book reading time and examine its effect on joint attention frequency in the other contexts.

Therefore, this study investigates the effect of shared book reading on joint attention frequency by comparing shared book reading and other contexts in experiment 1. Next, we intervene in the shared-book condition group by increasing its shared book reading time; then, we longitudinally compare its joint attention frequency in a toy play situation with that of the control group.

## Experiment 1

To investigate whether shared book reading increases joint attention episodes more than other play contexts, we compared the number of joint attention episodes during the shared book reading, toy play, and no material contexts, all common for 9-month-old infants and their mothers. To measure mother-infant joint attention frequency in these free play situations, we used indexes defined by Bakeman & Adamson (1984). Bakeman & Adamson (1984) performed longitudinal observations and indexed descriptions of the joint attention of infants from 6 to 18 months and their mothers in free play situations at home (using picture books and toys) and found that mothers' behavior is important for the occurrence of joint attention in infancy. Bakeman & Adamson (1984) divided joint attention episodes into two components; passive joint and coordinated joint. Passive joint occurs when mothers actively draw the infants' attention to an object, and coordinated joint occurs when infants voluntarily become involved in play and coordinate their attention to both mothers and objects. In passive joint, mothers need actively to lead the infants' attention; once coordinated joint begins, the mother must follow the infant's attention. Bakeman & Adamson (1984) suggest that mothers' coordinated behavior promotes children's joint attention development.

As Yont et al. (2003) and Sugai et al. (2010) show, mothers seem to draw their infants' attention to an object more frequently in the shared book context than in other contexts, and passive joint appears to occur in that context more frequently than in others. On the other hand, coordinated joint behavior may decrease in the shared book context because it does not need mothers' active behavior but rather coordinated and relatively passive behavior.

## Method

**Participants** Ten 9-month-old infants and their mothers participated in this study. The mothers were recruited from an official health center when bringing their children in for checkup.

**Procedure** Each mother and infant pair individually visited a university laboratory. First, they played freely in a play room for a few minutes and were told that we would videotape their interactions. When they seemed to relax and

play actively, they moved into the next room for the recording sessions.

**Conditions** All mothers and infants participated in three conditions: (a) in the shared-book condition, the infant and mother were observed while playing on the floor with a set of picture books we provided. We used the word "play" rather than "read" when we instructed them to prevent mothers from feeling they had to read the materials consecutively; (b) in the toy-play condition, the infant and mother played with a set of toys we provided; (c) in the no-material condition, the infant and mother played without using any materials.

Before each session, the mother was instructed to play freely, as if at home. After the instructions, an experimenter moved into a curtained area and sat calmly out of sight of the pair.

Each condition session lasted for about ten minutes and was conducted on different days. The second condition session wasn't conducted until at least three days after the first one, and the third didn't begin until at least three days after the second. The order of the conditions was counterbalanced.

**Materials** In two conditions, the mother and child used the materials we provided: (a) in each shared-book condition session, a set of picture books was used, consisting of ten books (*Harapeko-Aomushi* [*The Very Hungry Caterpillar*], *Jya-Jya Biri-Biri*, *Kingyo-ga Nigeta*, *Otsukisama-Konbanha*, *Kutsu-Kutsu-Aruke*, *Gatan-Goton Gatan-Goton*, *Nenai-ko Dareda*, *Wanwan-Nakunoha-Dare*, *Kuttuita*), all in Japanese. They were ranked in the top 10, by a bookstore near the university when surveyed on the bestselling picture books for 9-month-olds; (b) in each toy-play condition session, a set of ten toys was used: a ball containing a bell and a small doll, a handkerchief that makes paper sounds, a rattle, a toy telephone, a cloth bar that makes a funny sound, a toy trumpet, a roly-poly, a pacifier, a toy car containing a bell, and a toy tambourine. These had also been ranked in the top 10 by a toy store near the university when surveyed on the bestselling toys for 9-month-olds.

**Coding** After the above sessions, the videotaped mother-infant interaction was coded for ten minutes by an experimenter sitting out of sight of the pair. Six categories of engagement, as defined by Bakeman & Adamson (1984), were used to code the interactions: (a) unengagement, in which the infant appears uninvolved with the mother, object, or activity, although he or she might be scanning the environment; (b) onlooking, in which the infant is observing the mother's activity, often quite intently, but not taking part in the activity; (c) persons, in which the infant is engaged with just the mother (typically involving face-to-face or individual play, as, for example, when infants giggle and coo when their mothers place their face close to theirs and tickle them); (d) objects, in which the infant is involved in playing with objects alone, attending only to the books, toys, or whatever is at hand; (e) passive joint, in which the infant and mother are actively involved in the same object but the infant evidences little awareness of the mother's

involvement or even presence (mothers often attempt to induce this state by manipulating objects in ways that seem designed to capture their infants' attention and make the objects "come alive" for them); and (f) coordinated joint, in which the infant is actively involved in and coordinates his or her attention to both mother and the object the mother is involved with (as, for example, when the infant pushes a toy car the mother has been pushing and then looks back and forth between the mother's face and the toy car). If the infant gazes the object only by mother's attention drawing behavior and gazes mother's face only by mother's voice but doesn't keep gazing just one of them and doesn't voluntarily shift attention between them, the episode was coded into passive joint. On the other hand, if the infant alternately shift attention between them on his or her own initiative, the episode was coded into coordinated joint.

Fifty percent of the sessions were coded independently by two experimenters. The degree of agreement was gauged using Pearson's product-moment correlation coefficient ( $r$ ). The  $r$ s ranged from .91 to .98.

## Results

The average durations of the six engagements (unengagement, onlooking, persons, objects, passive joint, and coordinated joint) during each condition (shared-book, toy-play, and no-material) are presented in Table 1. These durations were analyzed with the conditions using a repeated-measures analysis of variance (ANOVA), run separately for each engagement category. The results of ANOVA indicated significant differences in unengagement, persons, objects, passive joint, and coordinated joint ( $F(2,18) = 22.76, p < .01, \eta_p^2 = .72$ ;  $F(2,18) = 53.07, p < .01, \eta_p^2 = .92$ ;  $F(2,18) = 20.47, p < .01, \eta_p^2 = .69$ ;  $F(2,18) = 29.04, p < .01, \eta_p^2 = .76$ ;  $F(2,18) = 13.02, p < .01, \eta_p^2 = .59$ , respectively) and no significant difference in onlooking. Multiple comparison tests (Bonferroni,  $p < .05$ ) were conducted in the former five categories.

In unengagement and persons, the mean duration in the no-material condition was significantly longer than in the shared-book and toy-play conditions. There was no significant difference between the shared-book and toy-play conditions. In objects, the mean duration in the shared-book condition was significantly longer than in the no-material

and that in the toy-play was significantly longer than in the shared-book. In passive joint and coordinated joint, the mean duration in the toy-play condition was significantly longer than in the no-material, and that in the shared-book was significantly longer than in the toy-play.

## Discussion

Passive joint occurred more often in the shared book context than in other contexts, as we had predicted. These results indicate that mothers' efforts to draw infants' attention to an object seem stronger in the shared book context than in other contexts. Therefore, not only infants older than 12 months but also younger infants follow the passive joint pattern with their mothers more often in the shared book context.

One reason for this is that shared book reading requires adults. Picture books consistent with the codes (like pictures and letters) require adults to scaffold the children's understanding, as they can't read letters (Karpov, 2005). Therefore, mothers spontaneously increase active drawing attention behavior more in shared book reading context.

On the other hand, coordinated joint also increased more in the shared book reading context than in other contexts. This result shows that although the mothers actively draw their child's attention, once the coordinated joint event begins through the infant in the shared-book reading context, the mother follows the infant's attention, which coordinates their behavior and attention.

One of the reasons for this is the difference in attractiveness between picture books and toys. Many of the toys used in this study were common among 9-month-old infants and made sounds. Nine-month-olds prefer objects that make sounds and toys are more attractive than picture books. The duration results in the object category show that duration was significantly longer in the toy-play than in the shared-book context. It is known that mothers are appeased more readily when their children maintain their focus on an object (Harman, Rothbart, & Posner, 1997). When faced with unattractive objects, infants are put into bad moods, and mothers must then struggle to produce a good mood. Therefore, mothers may become sensitive about their child's attention or intent.

Table 1: The durations of the engagement states in the shared-book, toy-play, and no-material contexts.

State of engagement	Shared-Book Context		Toy-Play Context		No-Material Context	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Unengaged	141.15	63.19	84.30	19.35	292.00	111.80
Onlooking	53.90	53.92	50.30	52.85	13.35	15.61
Persons	30.30	25.72	14.40	10.12	226.90	122.84
Objects	168.85	83.40	341.60	59.38	45.45	47.81
Passive Joint	182.35	83.38	98.75	70.68	18.30	29.77
Coordinated Joint	23.60	22.23	10.95	18.91	4.00	10.36

Note: the numerical values are presented in seconds.  $N = 10$ .

## Experiment 2

In experiment 1, passive joint and coordinated joint increase more in the shared book reading context than in other contexts. As previously indicated, if shared book reading facilitates child literacy by enhancing joint attention ability, increasing the shared book reading time may promote joint attention frequency not only in the shared book reading context but also in the others. To explore this possibility, we investigated the effect of increased shared book reading time on joint attention frequency by creating a shared-book condition group and the control condition group.

A child's joint attention ability develops rapidly at around 9 months of age (Dunham & Moore, 1995). Visual joint attention (Butterworth, 1991), the act of following the direction of another's gaze (like passive joint) develops quickly between 10 and 12 months (Corkum & Moore, 1995). Moreover, active joint attention, drawing someone's attention to an object (like coordinated joint) develops at around 12 months, later than passive joint (Lempers, 1979; Leung & Rheingold, 1981). Therefore, timing our intervention (i.e., increasing the shared book reading) at between 9 and 12 months seems appropriate.

As Bakeman & Adamson (1984) suggest, coordinating the mother's behavior to her infant's is important in producing and developing the child's joint attention ability. Experiment 1 revealed how, in the shared book context, passive and coordinated joint occurred more often than in the other contexts, and the mothers' coordinate behavior occurred more frequently. Therefore, conducting shared book reading seems to facilitate the frequency of both passive and coordinated joint.

## Method

**Participants** Twenty-two 9-month-old infants and their mothers participated in this study. Recruiting occurred as in experiment 1. They were divided into two condition groups of 11 pairs each, based on the infant's age, in the first of two recorded free play sessions (for children of 9 and 12 months). One of the groups (with 6 boys and 5 girls) was the shared-book condition. The other (with 7 boys and 4 girls) was the control condition.

**Conditions** (a) In the shared-book condition, after the first recorded free play session for the 9-month-olds, each mother was instructed to share picture books with her infant at least once a day (unless something prevented it) and to visit the laboratory again when the infant was 12 months of age. The experimenter sent each mother two picture books every four weeks when the infant was between 9 and 12 months of age. The mothers received six picture books in total, chosen by her from the eight books we provided after the first recorded free play session. The aim was to motivate the mother to share books with her infant; thus sending books the mothers had already used was avoided. The eight books we provided were suggested by Bookstart Japan (2006): *Jya-Jya Biri-Biri*, *Otsukisama-Konbanha*, *Kingyoga Nigeta*, *Kudamono*, *Tamago-no-Akachan*, *Pyo-nn*, *Shirokumachan-no-Hotcake*, *Rhythm*. (b) In the control

condition, the mother was given no instructions beyond being asked to visit the laboratory again when the infant was 12 months.

**Procedure** Each mother and infant pair visited a university laboratory. First, they played freely in a play room for a few minutes and were told that we would be videotaping their interactions. When they seemed to relax and play actively, they moved into the next room for the recording sessions. After the sessions, the mother was asked to fill out a questionnaire asking her the amount of time she had spent sharing books and toys respectively with her infant in the week prior to coming for the session for both the first and second laboratory visits.

**The recorded free play session** All mothers and infants participated in the session when the infants were 9 and 12 months. They were observed while they played on the floor with a set of toys we provided. The set of toys comprised a plastic boat, a stuffed monkey, a puzzle, a set of wooden blocks, a set of shape sorting cubes, and a drawing board, as was used by Stipek, Recchia, & McClintic (1992).

Before beginning the session, the mother was instructed to play freely, as at home. After the instructions, the experimenter moved into a curtained area and sat calmly out of sight of the pair. The session lasted for about ten minutes.

**Coding** After the above sessions, the videotaped mother-infant interaction in the recorded free play session was blindly coded for just ten minutes by the experimenter sitting out of sight of the pair. In experiment 2, two of the six categories (Bakeman & Adamson, 1984), passive joint and coordinated joint, were used to code the mother-infant interactions.

Twenty-five percent of the sessions were coded independently by two experimenters. The degree of agreement was gauged with Pearson's product-moment correlation coefficient ( $r$ ). The  $r$ s of passive joint and coordinated joint were .81 and .87, respectively.

## Results

**1 Analysis of the time spent on shared book reading and interaction between mother and infant** Three data sets were missed because the mothers had no time to fill out the questionnaire after the free play session. The average amount of time spent in shared book reading and interaction between mothers and infants in each group with infants of 9 and 12 months is shown in Table 2. The time of interaction means the total time during which mother shared books or toys with her infant. Data on the mean amount of time were analyzed in a 2 (age: 9 and 12 months of age)  $\times$  2 (group: the shared-book and the control group) mixed analysis of variance (ANOVA), in which age was a within-subject and group a between-subject variable, after a logarithmic transformation. In shared book reading time, the result of ANOVA yielded a significant age  $\times$  group interaction ( $F(1, 17) = 5.79, p < .05, \eta_p^2 = .25$ ). The results of subordinate tests indicated a significant simple main effect of age in the shared-book condition group ( $F(1, 34) = 5.30, p < .05$ ), meaning that mothers shared books with their infants more



often when their infants were 12 months of age than when they were 9 months, and a significant simple main effect of group when infants were 12 months of age ( $F(1,17) = 5.02, p < .05$ ), meaning that mothers in the shared-book condition shared books with their infants when they were 12 months of age more often than in the control condition.

In interaction time, the result of ANOVA revealed no significant difference anywhere.

Table 2: The Amount of Time of Shared Book Reading and Interaction in The Shared-Book and Control Groups.

Group	9 months of age		12 months of age	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Shared-Book ( <i>N</i> = 11)				
Shared book reading	60.32	100.02	97.82	73.77
Interaction	1312.14	1005.88	1141.00	769.66
Control ( <i>N</i> = 8)				
Shared book reading	60.63	68.94	35.00	35.04
Interaction	1976.88	1013.97	1302.50	501.02

Note. There numerical values are presented in minutes per a week

**2 Analysis of the number of times of joint attention episode** The average number of times of passive joint and coordinated joint in each group when infants were 9 and 12 months of age are presented in Figure 1 and Figure 2, respectively. The mean numbers of times data were analyzed in a 2 (age: 9 and 12 months of age)  $\times$  2 (group: the shared-book and the control group) mixed analysis of variance (ANOVA) where age was within-subject and group was between-subject variables.

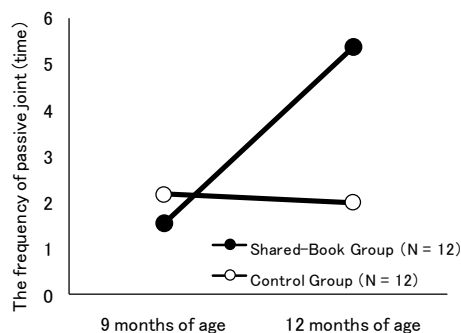


Figure 1 The Frequency of Passive Joint in Both Groups.

In passive joint, the results of ANOVA indicated a significant age  $\times$  group interaction ( $F(1, 20) = 4.59, p < .05, \eta_p^2 = .19$ ). The results of the subordinate tests indicated a significant simple main effect of age in the shared-book condition group ( $F(1, 20) = 8.37, p < .01$ ), meaning that, in that group, the frequency of passive joint was higher when

the infants were 12 months than when they were 9 months, and a significant simple main effect of group when infants were 12 months ( $F(1, 40) = 5.67, p < .05$ ), meaning that the frequency in the shared-book condition was higher than in the control condition when infants were 12 months. The results of ANOVA also yielded a significant main effect of age ( $F(1, 20) = 3.80, p < .10, \eta_p^2 = .16$ ), meaning that the frequency of passive joint significantly increased from when infants were 9 months to when they were 12 months.

In coordinated joint, the results of ANOVA indicated a significant main effect of age ( $F(1, 20) = 13.56, p < .01, \eta_p^2 = .40$ ), meaning that the coordinated joint frequency significantly increased from when infants were 9 months to when they were 12 months.

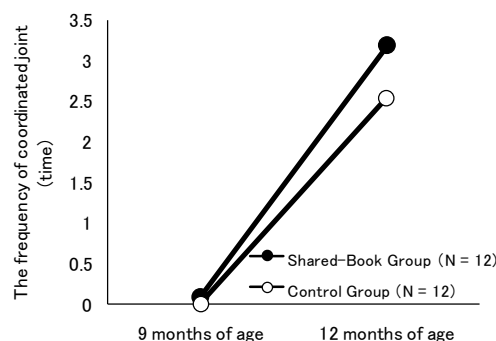


Figure 2 The Frequency of Coordinated Joint in Both Groups.

## Discussion

The results in Table 2 show that, in the shared-book group, the amount of time spent in shared book reading significantly increased and that this time was greater for the 12-month-olds in the shared-book group than in the control group. The results in Table 2 also show that the amount of time of whole interaction was no different between groups. Therefore, our intervention through increasing the shared book reading time seems to have been successful, but it doesn't mean just increased interaction time.

As Figure 1 indicates, passive joint frequency significantly increased in the shared-book group for the 12-month-olds, and the shared-book group's frequency was higher than the control group's. Therefore, it was suggested that increasing the shared-book reading time increases the passive joint dimension.

One reason for this is that the child's joint attention ability was promoted by the shared book reading. The occurrence of passive joint allows the child to find and follow his or her mother's gaze. As indicated in Experiment 1, passive joint occurred more often in the shared book context than in other contexts. The children in shared book reading group had more opportunities to have passive joint experience and to be aware of the existence of his or her mother's gaze or attention than those in control group, which is foundation of the ability of joint attention. A child's ability to follow the direction of people's gazes (as in passive joint) develops

rapidly from 10 to 12 months (Corkum & Moore, 1995). The children in shared book reading group were promoted to be aware the existence of others' attention through their shared book reading activities at home, and it promoted the development of passive joint and the success of passive joint was higher at 12 months old.

The results in Figure 2 show that the pairs in both groups increased in coordinated joint but that there was no difference between groups. The reason involves the maturation factor in a child's coordinated joint attention ability. As mentioned, passive joint attention ability develops remarkably between 10 and 12 months (Corkum & Moore, 1995), but coordinated joint attention ability develops at around 12 months (Lempers, 1979; Leung & Rheingold, 1981). Therefore, infants in both groups developed their abilities dramatically before reaching 12 months, and the study masked the effect of shared book reading on coordinated joint. In future, we must measure coordinated joint at younger and older ages than 12 months to eliminate the influence of the maturation factor.

## Conclusion

The results of experiment 1 show that, in the shared book context, passive and coordinated joint occurred more often than in the other contexts. This indicates that, in that context, mothers actively draw their child's attention to objects but coordinately follow their child's gaze and behavior. The results of experiment 2 suggest that repeated shared book reading increased the frequency of passive joint attention not in the shared book context but in the other context. This indicates that the shared book reading experiment facilitated the child's joint attention ability and/or promoted the mother's skill at drawing the child's attention.

## References

- Bakeman, R., & Adamson, L. B. (1984). Coordinating attention to people and objects in mother-infant and peer-infant interaction. *Child Development*, **55**, 1278-1289.
- Bookstart Japan (2006). The recommended picture books. <[http://www.bookstart.net/fr\\_news.html](http://www.bookstart.net/fr_news.html)> (25<sup>th</sup> of December, 2006)
- Bruner, J. S. (1985). The role of interaction formats in language acquisition. In J. P. Forgas (Ed.), *Language and social situations*. New York: Springer-Verlag.
- Butterworth, G. (1991). The ontogeny and phylogeny of joint attention. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development, and simulation of everyday mindreading*. Oxford, England: Blackwell. pp. 223-232.
- Corkum, V., & Moore, C. (1995). Development of joint visual attention in infants. In C. Moore & P. J. Dunham (Eds.), *Joint attention—Its origins and Role in Development*. Hillsdale: Lawrence Erlbaum Associates. pp. 15-28.
- DeBaryshe, B. D. (1993). Joint picture-book reading correlates of early oral language skill. *Journal of Child Language*, **20**, 455-461.
- Dunham, P. J. & Moore, C. (1995). Current themes in recent research on joint attention. In C. Moore & P. J. Dunham (Eds.), *Joint attention—Its origins and Role in Development*. Hillsdale: Lawrence Erlbaum Associates. pp. 15-28.
- Fletcher, K. L., Perez, A., Hooper, C., & Claussen, A. H. (2005). Responsiveness and attention during picture-book reading in 18-month-olds to 24-month-old toddlers at risk. *Early Child Development and Care*, **175**, 63-83.
- Harman, C., Rothbart, M. K., & Posner, M. I. (1997). Distress and attention interactions in early infancy. *Motivation and Emotion*, **21**, 27-43.
- Karpov, V. Y. (2005). *The neo-Vygotskian approach to child development*. Cambridge: Cambridge University Press.
- Karrass, J., VanDeventer, C. M., & Braungart-Rieker, M. J. (2003). Predicting shared parent-child book reading in infancy. *Journal of Family Psychology*, **17**, 134-146.
- Lempers, J. D. (1979). Young children's production and comprehension of nonverbal deictic behaviors. *The Journal of Genetic Psychology*, **135**, 93-102.
- Leung, E. H. L., & Rheingold, H. L. (1981). Development of pointing as a social gesture. *Developmental Psychology*, **17**, 215-220.
- Mundy, P., & Gomes, A. (1998). Individual differences in joint attention skill development in the second year. *Infant Behavior & Development*, **21**, 469-482.
- Payne, A. C., Whitehurst, G. J., & Angell, A. L. (1994). The role of home literacy environment in the development of language ability in preschool children from low-income families. *Early Childhood Research Quarterly*, **9**, 427-440.
- Stipek, D., Recchia, S., & McClintic, S. (1992). Self-evaluation in young children. *Monographs of the Society for Research in Child Development*, **57**.
- Sugai, Y., Akita, K., Yokoyama, M., & Nozawa, S. (2010). A developmental study of pointing during joint picture book reading: A longitudinal study comparing picture book reading and building block construction settings. *The Japanese Journal of Developmental Psychology*, **21**, 46-57.
- Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & P. Dunham (Eds.), *Joint attention: Its origins and role in development*. Hillsdale, NJ: Erlbaum. pp. 103-130.
- Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child Development*, **57**, 1454-1463.
- Wade, B., & Moore, M. (1998). An early start with books: Literacy and mathematical evidence from a longitudinal study. *Educational Review*, **50**, 135-145.
- Yont, K. M., Snow, C. E., & Vernon-Feagans, L. (2003). The role of context in mother-child interactions: an analysis of communicative intents expressed during toy play and book reading with 12-month-olds. *Journal of Pragmatics*, **35**, 435-454.

# A Computational PDP Model for Explaining Automatic Imitation

Matthias Scheutz (mscheutz@cs.tufts.edu)

Department of Computer Science, 165 College Avenue  
Medford, MA 02155 USA

Bennett I. Bertenthal (bbertent@indiana.edu)

Department of Psychology, 10<sup>th</sup> Street  
Bloomington, IN 47405 USA

## Abstract

Recent evidence suggests that automatic imitation is mediated by an observation-execution matching system that cannot be reduced to the same processes responsible for other stimulus-response (S-R) compatibilities. A computational model is developed with different patterns of connectivity for imitative and spatial compatibilities, and it is successful in simulating the results from three different S-R tasks. Variations of the model with identical connections for mediating the two compatibilities reveal a significantly poorer fit. These results provide converging evidence that imitative and spatial compatibilities are mediated by different processes.

**Keywords:** automatic imitation, S-R compatibility, connectionist modeling, direct matching hypothesis

## Introduction and Background

The tendency to unintentionally and unconsciously mimic actions performed by others has long been noted. Charles Darwin, for example, commented that at leaping matches spectators would move their own feet as if imitating the athletes. More recently, Dijksterhuis and Bargh (2001) noted that we tend to whisper or speak louder when others do, scratch our heads upon seeing someone else scratch, or cycle faster after seeing a cycling race on TV. During social interactions, mimicry translates into a greater desire to want to cooperate and affiliate with those individuals imitating our gestures (Chartrand and Bargh, 1999). In spite of the frequency and significance of these behaviors, our understanding of the underlying mechanisms responsible for these automatic tendencies remains incomplete at best.

The prevailing hypothesis for explaining spontaneous mimicry or automatic imitation is that the perception of some actions automatically activates corresponding motor programs. There are by now more than 75 experimental studies investigating automatic imitation, and most of the evidence is based on stimulus response compatibility paradigms, in which both stimuli and responses involve human movements (Heyes, 2011). In this paradigm, faster responding is observed when stimuli and responses correspond along some perceptual, structural or conceptual dimension than when they do not (referred to as a “compatibility effect”). When both the stimuli and responses involve human movements, it is often assumed that automatic imitation is involved. One problem with this interpretation is that the pattern of results for automatic imitation and all other S-R compatibility effects is exactly the same (i.e., faster response times for compatible than

incompatible responses to the stimuli). As a consequence, these results beg the question as to whether the processes mediating automatic imitation are specialized or instead are the same processes involved in other S-R compatibility tasks. In order to resolve this issue, it is necessary to find a paradigm where the results for automatic imitation and other S-R compatibility tasks are predicted to be different.

We recently provided such evidence by comparing imitative and spatial compatibilities in two experiments (Boyer, Longo, and Bertenthal, in press). The first tested for spatial compatibility with an imitative cue as the imperative stimulus, and the second tested for imitative compatibility with a spatial cue as the imperative stimulus. The stimulus consisted of a left or right hand with fingers spread apart and appeared on a computer screen from a third person perspective. Participants were instructed to respond to either the left-right spatial location or the anatomical identity of the index or middle finger tapping downward (Figure 1).

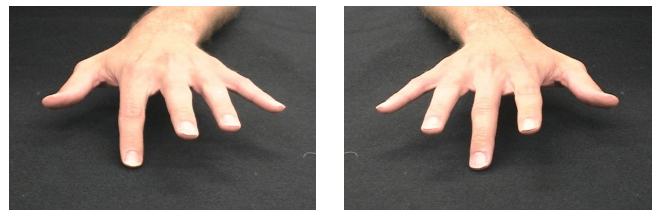


Figure 1. Left panel depicts the left hand stimulus with the index finger tapping down. Right panel depicts the right hand stimulus with the middle finger tapping down.

Responses consisted of pressing a key with the index or middle fingers on the right hand. In the standard S-R compatibility task (henceforth abbreviated “S-R task”), **the responses were compatible with a task-irrelevant spatial stimulus when the left hand was presented** (see Figure 1). For example, participants instructed to respond to the spatial cue would press a key with their index finger when responding to the left tapping finger. In this condition, both the stimulus and response are index fingers, and thus the response is facilitated via automatic imitation. Likewise, participants instructed to respond to the imitative cue would, for example, press a key with their middle finger when responding to the middle finger tapping. In this condition, both the stimulus and response correspond to the right side, and thus the response is facilitated via spatial compatibility. **When the stimulus corresponded to a right hand, the**

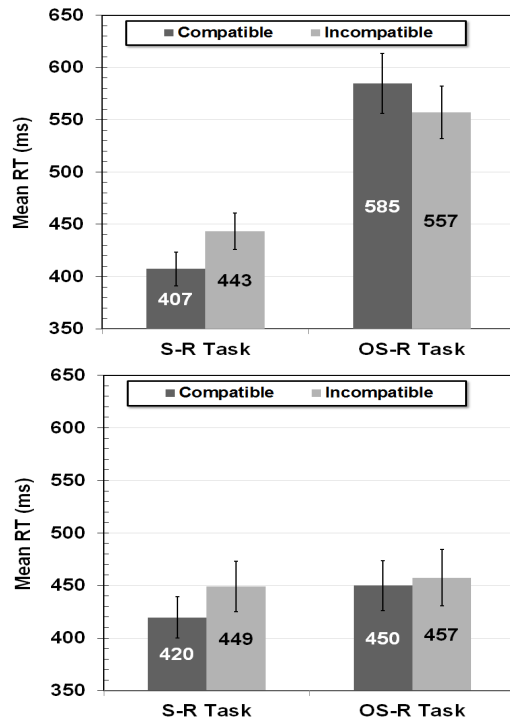


Figure 2. *Upper panel.* Mean response times (ms) to spatially compatible and incompatible stimuli as a function of task in Experiment 1. *Lower panel.* Mean response times (ms) to imitatively compatible and incompatible stimuli as a function of task in Experiment 2. (Error bars represent  $\pm$  standard error of the mean.)

**responses were not compatible with a task-irrelevant stimulus.** In the Opposite S-R compatibility task (henceforth referred to as “OS-R task”), the task involved responding to the stimulus not presented on that trial (e.g., participants responded to a index finger with their middle finger). **Responses were compatible with a task-irrelevant stimulus that corresponded to the right hand, and incompatible with a stimulus corresponding to the left hand** (see Figure 2).

For the S-R task, Boyer et al. (in press) predicted similar results when the irrelevant stimulus was either spatially or imitatively compatible. By contrast, they predicted different results for the OS-R task. Switching instructions from a spatial S-R to a spatial OS-R task was first investigated by Hedge and Marsh (1975), who reported a “reverse compatibility effect”. Although there is no consensus concerning the underlying mechanism, most hypotheses suggest that the recoding of the stimulus generalizes to the task irrelevant stimulus responsible for spatial compatibility and hence facilitates responding to the incompatible stimulus. By contrast, Boyer et al. (in press) hypothesized that imitative compatibility involves a direct and independent mapping between the task-irrelevant imitative stimulus and the response as suggested by recent neurophysiological evidence (Rizzolatti and Craighero, 2004). Thus, no reverse compatibility effect is predicted.

The results from the two experiments supported these predictions. In the S-R compatibility tasks, participants’ response times were faster to the spatially and imitatively compatible stimuli than to the incompatible stimuli (see Figure 2). By contrast, in the OS-R tasks, participants’ response times to the spatially incompatible stimuli were faster than to the spatially compatible (i.e., reverse compatibility effect), whereas response times to the imitatively compatible stimuli were faster than to the imitatively incompatible stimuli (although this difference was not significant). The results from this latter experiment thus suggested that imitative and spatial compatibilities are not mediated by the same domain-general process.

Although the preceding results suggest that spatial and imitative compatibilities are mediated by different processes, it remains an empirical question as to whether the response differences are a function of different neural architectures or more simply a function of differences in the parameterization of the same architecture. The standard processing model for explaining S-R compatibility effects is a dual route model whereby responses are activated by an intentional route as well as an automatic route (Zorzi and Umiltà, 1995; Zhang, H., Zhang, J., and Kornblum, 1999). If the automatic route activates the same response as the intentional route, the response is facilitated. If, however, the automatic and intentional routes activate different responses, then the response is slowed down. In spite of sharing this general processing assumption, all dual route models are not the same. Some models are designed with some or all inputs mapped directly to the outputs, and other models are designed with a middle decision layer that selects the response (e.g., Sausser and Billard, 2002). In the current research, we hypothesized that the spatial compatibility task is mediated by the latter model, whereas the imitative compatibility task is mediated by a hybrid model (i.e., the automatic imitative route involves a direct mapping between input and output, but the controlled route involves a task-based mapping between the input and a middle decision layer). This latter model is consistent with recent theories suggesting that automatic imitation is due to a shared representation between the observation and execution of actions.

The purpose of the current investigation was to test this hypothesis with a computational model that was designed to simulate the empirical results from the previous study by Boyer et al. (in press).

### A PDP Model of Spatial and Ideomotor Compatibility Effects

We started with our previous computational modeling efforts (Boyer et al. 2009) and develop a new three-layered (input-hidden-output) connectionist network, with nodes at each layer representing the stimulus input, the S-R translation, and the response, respectively. We use simplified interactive activation and competition connectionist units (Rumelhart and McClelland, 1986) with change in activation over time is given by  $\Delta act/\Delta t = netin-$

$act(netin+decay)$ , where  $act \in [0,1]$  is the *activation* of the unit,  $netin \in [0,1]$  is the *sum of the weighted inputs* to the unit and  $decay \in [0,1]$  is a constant *decay factor*. The model consists of eight units: two input units, called *finger units*, representing the perceived index (“I”) and middle (“M”) input fingers; two input units, called *location units*, representing the left (“L”) and right (“R”) location of the perceived input fingers (depending on the stimulus hand); two *output units* representing the index finger in the left location (“IL”) and the middle finger in the right location (“MR”) of the right hand; and two hidden units (or decision units), called *SR units*, affecting the S-R translation between inputs and outputs (“SR-IL” and “SR-MR”). As in (Boyer et al. 2009), we start with a base model which shows the participant’s state before any task-based instructions and the task models which show the participant’s condition after a task-based instruction. The base model consists of automatic connections between input and hidden, and hidden and output units. The input finger (“I” or “M”) and spatial location (“L” or “R”) are mapped onto the requisite hidden unit (“SR-MR” or “SR-IL”) via the connections  $I \xrightarrow{a} SR-IL$ ,  $M \xrightarrow{a} SR-MR$ ,  $L \xrightarrow{a} SR-IL$ , and  $R \xrightarrow{a} SR-MR$ , respectively. These connections correspond to the compatible S-R translations between the imperative stimulus and the response, which we hypothesize are processed automatically by the hidden units presumably because they are overlearned and automatized. In addition to these automatic connections, we assume direct connections between input and output fingers that reflect the hypothesized direct matching pathways:  $I \xrightarrow{i} IL$  and  $M \xrightarrow{i} MR$  (note that there are no direct connections between input and output spatial locations). Lastly, hidden units are mapped onto corresponding output units via connections  $SR-IL \xrightarrow{a} IL$  and  $SR-MR \xrightarrow{a} MR$ .

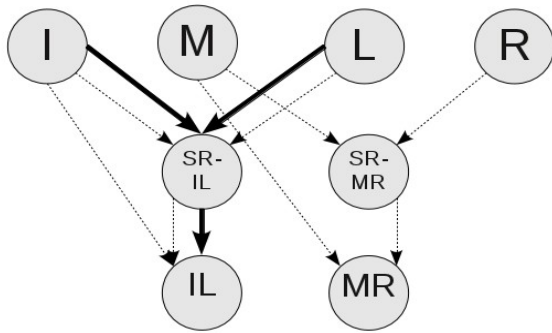


Figure 3. The proposed neural network model: the base model consists of only the dashed connections; the bold connections depict the S-R compatible task-based connections for the index finger.

From the base model, we construct the task models by adding additional connections that depend on the task instructions. Specifically, we add connections between

input and hidden units: for the spatial compatibility models, we add task-based connections  $I \xrightarrow{t} SR-IL$  and  $M \xrightarrow{t} SR-MR$  reflecting that the imperative stimulus is the anatomical identity of the finger, and for the imitative compatibility models, we add task-based connections  $L \xrightarrow{t} SR-IL$ , and  $R \xrightarrow{t} SR-MR$  reflecting that the imperative stimulus is the left-right location of the finger. Furthermore, for the standard (S-R) mapping models we add task-based connections between hidden and output units  $SR-IL \xrightarrow{t} IL$  and  $SR-MR \xrightarrow{t} MR$ , while for the opposite (OS-R) mapping models we add task-based connections between hidden and output units  $SR-IL \xrightarrow{t} MR$  and  $SR-MR \xrightarrow{t} IL$  in which the controlled connection from the hidden unit crosses to the opposite output unit. This crossing is necessary given that the task instruction requires participants to select the opposite response to that selected in the standard mapping condition (i.e., either responding to the finger with the opposite identity or spatial location). Different from (Boyer et al. 2009), the current model has only excitatory connections (i.e., connections with positive values).

Inputs are applied to the model by fixing the netinput of the respective input units (e.g., “I” + “L”) at a particular value to indicate, for example, the perceived index finger in the left position). The input is applied on each cycle of the trial because participants are able to observe the stimulus finger until they respond. The state of the model is updated in discrete time steps (“cycles”) that correspond to 10 ms of real-time. Response selection is achieved whenever an output unit reaches the *action threshold* (i.e., the activation needed to perform a motor response by the finger). As such, the number of cycles computed from the introduction of the input (i.e., moving finger) until the action threshold is reached can be used to simulate the response times directly (e.g., 30 cycles correspond to 300 ms).

To minimize the number of free model parameters that can be used to fit models to the empirical data, we fix all base model parameters based on the study by Boyer et al., 2009. Specifically, we assume that all automatic connections  $\rightarrow_a$  and direct mapping connections  $\rightarrow_i$  have the same strength in all models, and they are set to a very low value of 0.001 (which is too low to generate any actions without task instructions, even if all input units are activated together). Moreover, we assume that the same decay factor of 0.05 for all computational units and also fix both the external input and the action threshold at 0.5. With those base model parameters fixed, we are left with the task-based connections as *free parameters* that can be used to fit the models to the empirical data.

## Model Fitting and Simulation Results

We start with spatial and imitative compatibility in the S-R condition. There are four free parameters in each condition



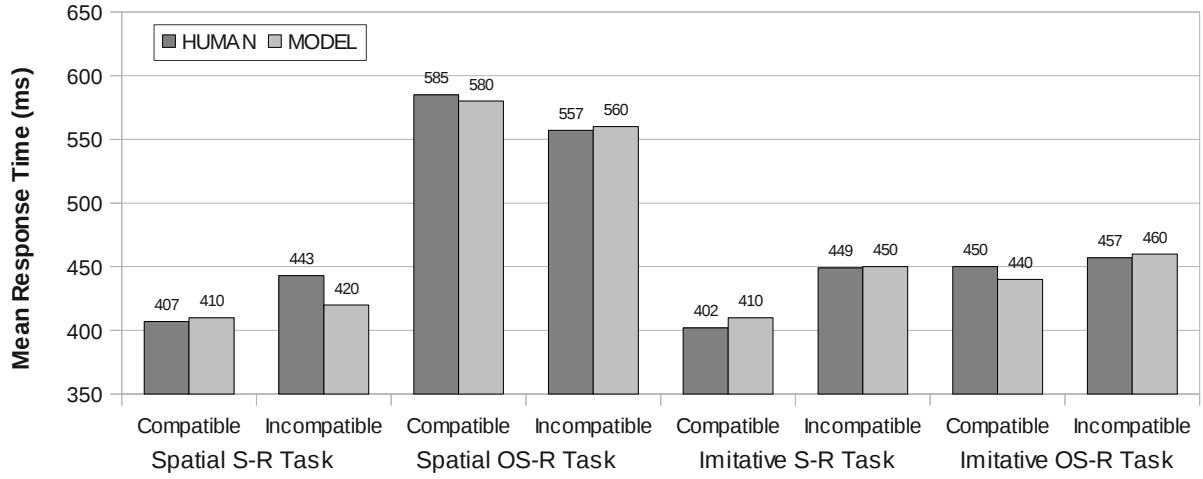


Figure 4. The simulation results of the proposed model for both compatibilities in the S-R and OS-R conditions.

corresponding to the input-to-hidden and hidden-to-output task connections, and we make the simplifying assumption that all task-based connections in the standard mapping models have the same strength. This assumption is plausible because both finger and location inputs require similar levels of encoding and integration followed by a similar S-R translation process regardless of whether the task instruction is to respond to the anatomical identity or location of the finger (Boyer et al., in press). Accordingly, we searched for a single positive value for all task-based connections that minimizes the *root mean squared error* for each model (*RMSE*). We found  $\rightarrow_t = 0.086$  to be such a value with *RMSE* = 12.28 ms, which is just above the model resolution of 10 ms necessary for simulating a compatibility effect.

The next step was to use the model parameters in the S-R condition to predict the results in the OS-R condition. As previously described, the hidden-to-output connections for the task-based processes reverse between these two layers because the correct response is opposite to the one cued by the stimulus. We assume that, because only the hidden-to-output connections are different from the S-R models, the OS-R models should use the same connection strength (0.086) as the S-R models for the input-to-hidden connections. Moreover, we assume that both hidden-to-output connections have the same strength for both outputs; but based on the results of Boyer et al. (in press) showing longer response times for recoding the spatial cue, we introduce different strengths for imitative vs. spatial compatibility conditions. Hence, we are left with only one free parameter  $SR-IL \rightarrow_t MR=SR-MR \rightarrow_t IL$  in each of the two opposite mapping conditions.

In order to fit this free parameter, we begin with the OS-R condition for imitative compatibility. Unlike the results for the spatial compatibility condition which showed a reverse compatibility effect, the empirical results for this condition were similar to those for the S-R condition. Given this similarity, we predicted that the same value of 0.086 should work for  $SR-IL \rightarrow_t MR$  and  $SR-MR \rightarrow_t IL$  in the OS-R

condition for imitative compatibility. Consistent with our hypothesis, this simulation was very successful with *RMSE* = 7.38 ms, which is less than the model resolution of 10 ms.

The situation was somewhat different for fitting the spatial compatibility model in the OS-R condition. Recall that this condition differed from the other three conditions in two ways: (1) the results revealed a reverse compatibility effect, i.e., response times were faster to the incompatible than to the compatible stimulus; and (2) the response times in this condition were significantly higher than in the comparable imitative compatibility condition (i.e., 571 ms vs. 454 ms). Hence, in order to model these two results, we made two predictions: (1) the reverse compatibility effect is a function of the model architecture, and therefore it should be unnecessary to change the connection strength of 0.0856 for  $SR-IL \rightarrow_t MR=SR-MR \rightarrow_t IL$  from the connection strengths used in the other conditions; and (2) a lower connection strength (less than 0.086) is required for  $SR-IL \rightarrow_t MR=SR-MR \rightarrow_t IL$  to fit the model to the significantly longer response times with a small *RMSE*.

In the first simulation of this condition, we did not change the hidden-to-output connection weight (set  $SR-IL \rightarrow_t MR=SR-MR \rightarrow_t IL=0.086$ ) which yielded a response time of 460 ms for the incompatible condition and a response time of 470 ms for the compatible condition. These results thus confirm our first prediction because they are consistent with a reverse compatibility effect. In the second simulation we lowered  $SR-IL \rightarrow_t MR=SR-MR \rightarrow_t IL=0.083$  by about 3%, which yielded a response time of 560 ms for the incompatible condition and a response time of 580 ms for the compatible condition. In contrast to the previous simulation, these results revealed not only a reverse compatibility effect but also a very small *RMSE* = 4.12, thus supporting our second prediction.

Several points are worth noting about the above modeling results. First and foremost, the current models are capable of capturing the reversed spatial compatibility effect in the

empirical data that Boyer et al. (2009) failed to capture. At the same time, the current model is simpler than the one presented in Boyer et al. (2009) because it has fewer nodes and fewer connections, does not use any inhibitory connections, and thus has overall fewer parameters. And

finally, the fits we obtained here are better than the fits in Boyer et al. (2009). The fact that spatial and imitative compatibilities in the OS-R condition required separate models is theoretically significant. This result thus provides an important source of evidence for concluding that the two compatibilities are not mediated by the same processes.

### Testing the Generalizability of the Model

Given that the model succeeded at capturing the data from Boyer et al. (in press), we sought to test its generalizability as a means of providing further evidence that spatial and imitative compatibilities are not mediated by the same processes. Recently, Catmur and Heyes (2010) conducted a related study testing the effects of spatial and imitative compatibilities on response times. In this study, participants responded to a discriminative cue with an abduction of either the index or little finger of their right hand. On each trial, a left or right hand was displayed initially in a neutral position with fingers spread apart and the outline of a small white circle appearing equidistant between the tips of the index and little fingers. Participants were instructed to respond as quickly as possible to the circle changing to orange or purple by abducting their index or little finger depending on the task instructions. Simultaneous with the appearance of the discriminative cue, the index or middle finger of the stimulus hand was abducted. By varying whether the stimulus corresponded to the left or right hand, it was possible to independently manipulate imitative and spatial compatibility, such that both compatibilities were present, only one, or neither.

To allow our current model to simulate the Catmur and Heyes (2010) task, we add two additional color nodes (“O” for “orange” and “P” for “purple”) together with task-based connections  $O \rightarrow SR-IL$  and  $P \rightarrow SR-MR$  to the two hidden nodes “SR-IL” and “SR-MR”, respectively, in the base model to reflect the requirement to abduct the little finger (“little finger”) for the orange stimulus, and the index finger (“index finger”) for the purple stimulus or vice versa (note that we can re-use the “middle finger” node for the “little finger”). We hypothesize that our model should also be able to simulate the results from this study with only minor adjustments. Keeping the task-based connections while increasing the automatic connections (to 0.003) and lowering the direct mapping connections (to 0.0001), we get a result (with a small RMSE of 7.07ms) that very closely captures the Catmur and Heyes (2010) data (see Figure 5). This result confirms that our model is not limited to simulating results from only one specific experiment. It should be noted, however, that we do not simulate all results from the Catmur and Heyes (2010) experiment, because they also investigate the time course of the compatibility effect during

the trial. In order to model these within trial timing effects, it would be necessary to design a stochastic model which is currently in development, but it is premature to report any results from this model.

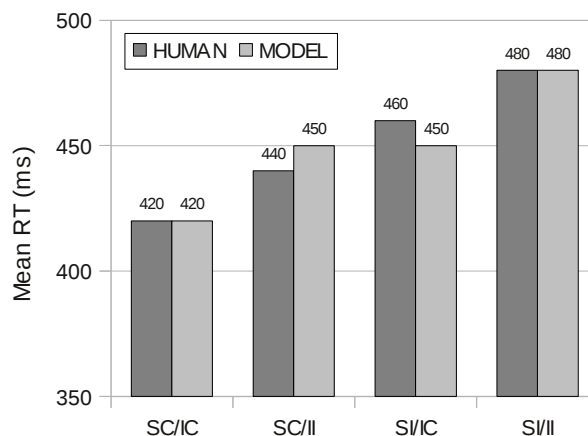


Figure 5. The simulation results of the model for the Catmur and Heyes (2010) experiment (“SC”=“spatial compatible”, “IC”=“imitative compatible”, “SI”=“spatial incompatible”, and “II”=“imitative incompatible”).

### General Discussion

There are three principal conclusions from this investigation: (1) A computational model was presented that was successful in simulating the effects of spatial and imitative compatibilities in three separate conditions. (2) Unlike previous dual route models for explaining stimulus-response compatibilities, spatial and imitative compatibilities did not conform to the same architecture (i.e., imitative compatibility included a direct input-output connection, whereas spatial compatibility did not include this direct connection). (3) The finding that spatial and imitative compatibilities were modeled differently provides converging evidence that these two compatibilities are mediated by different processes.

A legitimate question is, therefore, whether the model would have been equally successful if we reversed which stimulus dimension was simulated with direct connections. Figure 6 shows the model with direct input-output connections for spatial, but not for imitative compatibilities. As can be seen, the model is just the mirror image of our previous model in the S-R conditions and thus fits the empirical results almost as well (with a slightly larger RMSE=15.52). In the OS-R condition, however, the model differs significantly from the empirical results, showing the exact opposite relations (as indicated by the two ovals in Figure 6). Not surprisingly, the RSME=38.87 for the OS-R condition is significantly higher than for the previous model with the direct imitative connections (RSME=7.12). It thus appears that these direct connections are necessary for predicting the imitative compatibilities in the OS-R condition. Given these results, we also investigated whether adding to our model direct connections for spatial compatibility improved the fit.



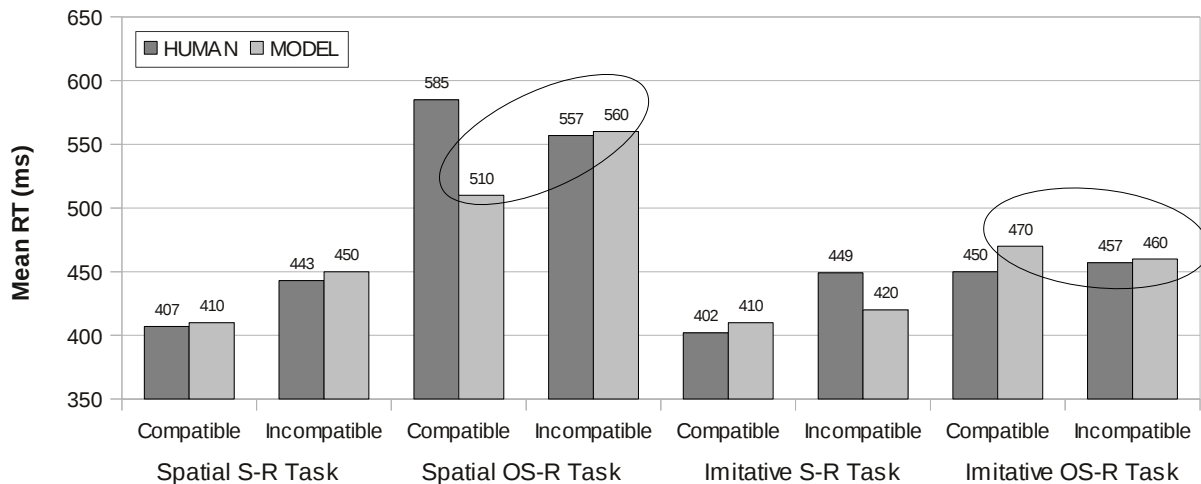


Figure 6. The simulation results of the model with no direct imitative, but direct spatial connections for both S-R and OS-R conditions. The ovals depict the two conditions in which the model predicts the opposite effects from the empirical data.

Preliminary experiments show that the addition of direct connections between the location and output nodes results in does even worse (without lowering of the task-based connections) in the S-R condition (with RSME=21.27 compared to RSME=13.99 for our previous model). However, as with the model including only direct spatial connections, this new model failed to match the empirical data in the reverse spatial compatibility condition, with an overall RSME=37.89 for both OS-R conditions. Hence, neither models with only direct spatial nor models with direct spatial and direct imitative connections are capable of matching the empirical data in the OS-R conditions as well as our proposed model under the given parameter assumptions.

## Conclusions

In this paper, we introduced the first computational PDP model that can be fit to human data dissociating spatial from imitative compatibilities *and* can be used to make predictions about related tasks. While the results are constrained by the chosen constants and the modeling process, the comparison with alternative model architectures together with the model's ability to predict performance in a related, yet different task, are an encouraging step towards a full-fledged investigation of model parameters and model architectures that might be able to account for the empirical differences between imitative and spatial compatibility effects in a great variety of experimental paradigms.

## References

- Boyer, T. W., Scheutz, M., and Bertenthal, B. I. (2009). Dissociating ideomotor and spatial compatibility: Empirical evidence and connectionist models. In N. Taatgen, H. van Rijn, and L. Schomaker (Eds.) *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2280-2285). Austin, TX: Cognitive Science Society.
- Boyer, T.W., Longo, M.R., and Bertenthal, B.I. (in press). Is automatic imitation a specialized form of stimulus-response compatibility? Dissociating imitative and spatial compatibilities. *Acta Psychologica*.
- Catmur, C., and Heyes, C. (2011). Time course analyses confirm independence of imitative and spatial compatibility. *Journal of Experimental Psychology: Human Perception and Performance*, 37, 409-421.
- Chartrand, T. L., and Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76, 893-910.
- Dijksterhuis, A., and Bargh, J.A. (2001). The perception-behavior express-way: Automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology*, 33, 1-39.
- Hedge, A., and Marsh, N. W. A. (1975). The effect of irrelevant spatial correspondence on two-choice response-time. *Acta Psychologica*, 39, 427-439.
- Heyes, C. (2011). Automatic imitation. *Psychological Bulletin*, 137, 463-483.
- Rizzolatti, G., and Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169-192.
- Rumelhart, D. E. and McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I*. Cambridge, MA: MIT Press.
- Sausser, E. L., and Billard, A. G. (2006). Parallel and distributed neural models of the ideomotor principle: An investigation of imitative cortical pathways. *Neural Networks*, 19, 285-298.
- Zhang, H., Zhang, J., and Kornblum, S. (1999). A parallel distributed processing model of stimulus-stimulus and stimulus-response compatibility. *Cognitive Psychology*, 38, 386-432.
- Zorzi, M., and Umiltà, C. (1995). A computational model of the Simon effect. *Psychological Research*, 58, 193-205.

# Taking Development Seriously: Modeling the Interactions in the Emergence of Different Word Learning Biases

**Savannah M. Schilling** ([Savannah.Schilling@Colorado.Edu](mailto:Savannah.Schilling@Colorado.Edu))

Department of Electrical, Computer, & Energy Engineering, 425 UCB  
Boulder, CO 80309-0425 USA

**Clare E. Sims** ([Clare.Holtpatrick@Colorado.Edu](mailto:Clare.Holtpatrick@Colorado.Edu))

Department of Psychology & Neuroscience, 345 UCB  
Boulder, CO 80309-0345 USA

**Eliana Colunga** ([Eliana.Colunga@Colorado.Edu](mailto:Eliana.Colunga@Colorado.Edu))

Department of Psychology & Neuroscience, 345 UCB  
Boulder, CO 80309-0345 USA

## Abstract

Development is about change over time. Computational models have provided insights into the developmental changes seen in different cognitive phenomena, including within the domain of word learning. The present paper uses a computational model to investigate the interdependencies between the emergence of different word learning biases. This model allows investigation of how the emergence of the shape bias influences novel noun generalization to two other types of items. The results suggest that the emerging shape bias for solids can either strengthen or weaken other types of biases depending on the strength of the cues to solidity or non-solidity; further, these results make predictions about children's biased word learning over time.

**Keywords:** Computational models; neural networks; trajectories; word learning; shape bias.

## Introduction

Computational models have proven to be an important tool for investigating many issues within cognitive development (e.g., Munakata & McClelland, 2003). Such models can provide insights about the mechanisms that underlie learning patterns seen across childhood. In the domain of word learning, various models have been used to investigate fast mapping and the taxonomic shift (Mayor & Plunkett, 2010), variability effects in learning phonetically similar words (Apfelbaum & McMurray, 2011), task effects in novel noun generalization (Samuelson, Schutte, & Horst, 2009), and word learning at different levels of abstraction (Xu & Tenenbaum, 2007). In this paper we focus on using connectionist models to examine developmental trajectories in word learning. This approach has the potential to guide novel and testable predictions about children's language development.

This paper employs a computational modeling approach to investigate the emergence of word learning biases that support early language acquisition. This approach allows us to analyze in detail how different word learning biases interact and influence one another over the course of word learning. For example, does a later emerging bias build onto and benefit from an earlier bias, or is there a period of conflict as new knowledge is assimilated with prior

knowledge? Our results indicate that different biases do interact, and the emergence of one bias can either strengthen or weaken other biases depending on the strength of cues provided in the learning context. These results allow us to make predictions about the timing of children's word learning and generalization as biases emerge.

## Word Learning Biases

One of the reasons that children are such skilled language learners is because of biases. In the context of word learning, biases are constraints on the range of things that children will consider in deciding what a new word refers to. Rather than assuming that any word can be used to label any item, children exhibit principled patterns of behavior in the ways in which they learn words. The main constraint we will focus on in this paper is found within the domain of noun learning: the shape bias. The shape bias refers to young children's tendency to generalize newly learned nouns to other objects based on similarity in shape (Landau, Smith, & Jones, 1988). That is, if a child is taught a novel name for a novel solid object, he or she will extend that name to other objects that match the original in shape, even if that shape match differs in texture, color, or size. Children show a reliable shape bias by 2 years of age (Samuelson & Smith, 1999).

A related phenomenon in noun learning is the material bias. While the shape bias is seen in children's generalization of labels to solid objects, the material bias concerns the labeling of non-solid substances. The material bias has been found using the same novel noun generalization (NNG) paradigm typically used in studies of the shape bias. Children taught a novel name for a novel non-solid substance tend to generalize that name to other non-solids that match the original in material rather than to non-solids matching in features like shape and size but made out of a different material (e.g., Soja, 1992; Soja, Carey, & Spelke, 1991). The material bias is typically seen slightly later than the shape bias, at 3 years of age (Yoshida & Smith, 2005).

Altogether, the evidence suggests that over the first years of life children develop preferential attention to different

features of items in noun learning, first to shape in naming solid objects and then to material in naming non-solid substances. This raises the question of whether and how these biases interact with each other. Does development of the shape bias earlier on have any impact on the material bias? Research looking at naming and generalization of other, ambiguous kinds of items hints at this possibility. For example, deformable items usually have a characteristic shape, but are also often categorized as being similar to each other in material (e.g., an item such as *towel*; Samuelson, Horst, Schutte, & Dobbartin, 2008). Previous research shows that young children categorize deformable items based on similarity in material when they are not labeled, but categorize based on similarity in shape when the items are labeled (Samuelson & Smith, 2000). Samuelson and colleagues (2008) hypothesize that this behavior represents an overgeneralization of the shape bias. This suggests that the emergence of the shape bias can influence how children learn and generalize names for other, more ambiguous kinds of items as well.

### The Emergence of Biases

Although there is some debate over the origin of word learning biases (e.g., see Samuelson & Bloom, 2008), there is strong evidence to suggest the link between vocabulary growth and the emergence of word learning biases. This has been explored especially in relation to the shape bias, with several pieces of evidence suggesting interactions and feedback between attending to shape in the context of learning names for solid objects and overall word learning.

First, there is evidence that the emergence of the shape bias influences subsequent word learning. For example, Smith, Jones, Landau, Gershkoff-Stowe, and Samuelson (2002) intensively trained 17-month-old children on labels for novel shape-based categories. The children exposed to this training not only developed a shape bias earlier than is typically seen, they also showed a dramatic increase in vocabulary size compared to a control group. These results suggest that the development of the shape bias accelerates children's learning of object names outside of the lab.

Evidence for a feedback relationship between word learning and the emergence of the shape bias comes from a study by Gershkoff-Stowe and Smith (2004). In this study, children were longitudinally tested on their attention to shape in generalizing a novel label. The researchers also collected diaries tracking children's vocabulary growth. The results showed that children's attention to shape increased as the number of nouns in their vocabularies increased.

Together these studies suggest an interesting pattern of interactions between language acquisition and the emergence of word learning biases. As children add more nouns to their growing vocabularies, they show an increasing preference to attend to shape in the context of naming solid objects. This preference to attend to shape in turn facilitates subsequent word learning. This leaves open the question of how different biases may interact over the

course of language development. Modeling word learning offers a unique way to investigate this question.

### Modeling Word Learning & Biases

Computational models of word learning have been used to investigate the conditions that support word learning biases. For example, Colunga and Smith (2005) trained a network with a vocabulary structure of half solid objects characterized by shape and half nonsolid objects characterized by material—a vocabulary structure which should directly promote the development of shape and material biases. Results of the virtual analog of the NNG task confirmed this prediction in that the network showed a shape bias for solids and a material bias for nonsolids. In a second simulation, they trained the same network on a realistic early vocabulary, structured like that of a typical 30-month-old. The earliest nouns that children typically learn are dominantly comprised of solid objects characterized by shape (e.g., *ball*, *spoon*), and include fewer non-solid substances characterized by material. The typical early noun vocabulary also includes types of items that can be characterized by both shape and material. When trained on this more complex vocabulary structure, the network again treated novel test items in ways consistent with shape and material biases. Colunga and Smith (2005) also reported behavioral data with young children confirming the predictions of this network. More recently, Colunga and Sims (2011) used the same kind of network to successfully predict differences in novel noun generalization patterns between early- and late-talker children.

These studies show that computational modeling is a powerful tool for exploring the emergence of word learning biases. However, no one has yet used this approach to investigate the relationships between different biases as they emerge over the course of word learning. Our approach offers a new perspective by modeling different word learning biases together on a developmental timescale.

### Approach and Overview

Our approach is to train a network on a typical early child vocabulary and then test it on generalization of three different kinds of items over the course of word learning. The goal is to see how the shape bias emerges over word learning and how the emergence of the shape bias impacts generalization performance on other kinds of items. To test for a shape bias, we implement a virtual NNG task by exposing networks to novel solid items and seeing if they treat those alike in shape more similarly than those alike in material (as in Colunga & Smith, 2005). We also test the network on two types of novel non-solid items: simply-shaped, clearly material-based non-solids and complex-shaped, ambiguous non-solids. An example of a simply-shaped nonsolid would be a glob of paint. While it can take on slightly different shapes, it always maintains its “blob-like” shape. Simply-shaped non-solids are inherently material-based because there is more variation in the material than in the shape, making the material of an item a

more useful cue in classification. Complex-shaped, ambiguous non-solids, on the other hand, can take on virtually any shape, e.g. paint smeared in the shape of a peace symbol. In this case, shape and material have equal variability and either one could be a cue to classifying an item. We explore whether the network develops a material bias for these types of non-solids, and whether that development depends on the emergence of the shape bias for solids.

## Simulation

### Method

Computational models use the Leabra algorithm (Local, Error-driven and Associative, Biologically Realistic Algorithm), which combines Hebbian and error-driven learning. Weights are adjusted based on correlations between activation units and feedback of errors through back propagation (O'Reilly et al., 2012).

**Architecture** The architecture is adapted from Colunga and Smith (2005) and is implemented as shown in Figure 1. Words are represented discretely and are input on the Word Layer. Referents are represented as distributed patterns over several dimensions on the Perceptual Layer. For example, the shape and material of an object (say the roundness of a particular ball and its yellow rubbery material) are represented by an activation pattern along the Perceptual Layer, with 12 units for shape and 12 units for material. Solidity is represented discretely; one unit stands for Solid and another for Non-Solid. Finally, there is a 25 unit Hidden Layer that is connected to all the other layers and to itself. The Hidden Layer serves as the bridge between the Word Layer and the Perceptual Layer and it is where learning occurs. Learning progresses as internal representations, or weights, update and form links between the other two layers.

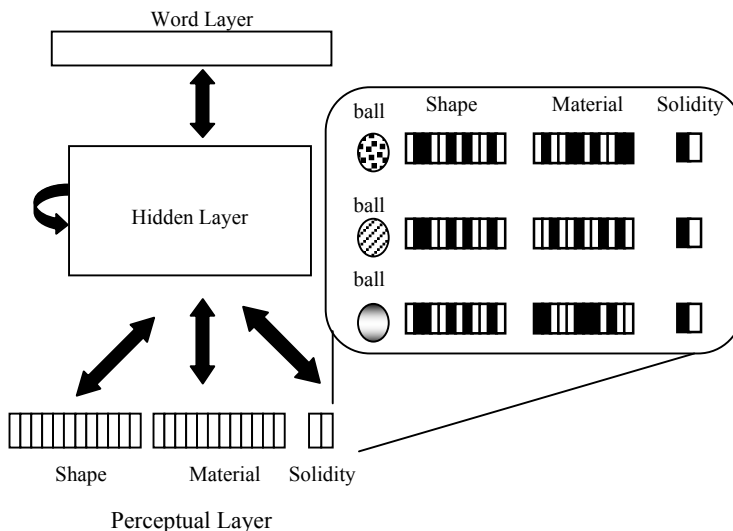


Figure 1: Architecture of the network and example input patterns.

Table 1: Noun category proportions used to create the input vocabulary structure. Beneath each proportion is an example noun belonging to that category.

	Shape	Material	Both
Solid	52% <i>ball</i>	10% <i>cheese</i>	12% <i>crayon</i>
Non-solid	4% <i>bubble</i>	16% <i>glue</i>	6% <i>jeans</i>

**Input Patterns** The input patterns consisted of training and testing patterns. The training patterns were structured to capture the same correlational structure as the vocabulary of a typical 30-month-old child (Fenson et al., 1993). The structure of this training input was based on that used in Colunga and Smith (2005). Using adult judgments, nouns were categorized by both solidity (either solid or non-solid) and characteristic feature (either shape, material, or both). The structure of the typical early noun vocabulary could then be expressed as proportions of each type of category. See Table 1 for the 6 categories and proportions used in the current study. The network in this study was trained to learn 50 noun representations, designed to have approximately the category structure of the typical 30-month-old's vocabulary. Note that these 50 noun representations are word inputs to the network and do not necessarily correspond to 50 nouns that a child learns.

The test input patterns consisted of three kinds of items represented along the Perceptual Layer: solids, simply-shaped non-solids, and complex-shaped non-solids. Test patterns were made up of triplets of novel items: an exemplar and two choice items.

For solid test items, this pattern was instantiated by manipulating activation across all shape and material units of the network. For clear non-solids, half of the shape units were kept constant to capture the fact that non-solid substances are typically seen in simple shapes like smears and splashes.<sup>1</sup> That is, these simply-shaped, clear non-solids had some variability in the shapes they could take on, but also had some imposed limitations. Finally, the ambiguous non-solid test patterns were represented by manipulating activation across all of the shape and material units. This type of non-solid test pattern is called ambiguous because, unlike other non-solids the network was exposed to, they can take on more complex shapes and thus be construed as having a characteristic material or a characteristic shape. Therefore, testing involved generalization to novel solid and simply-shaped, non-solid items as well as generalization to a new, ambiguous type of non-solid item.

**Progression of Word Learning** To chart the course of bias development, we tested the network at multiple points throughout word learning. Weights were recorded at initialization, every five words from 0 to 30 words learned, and every ten words from 30 to 50 words learned. The

<sup>1</sup> This method for simplifying the shapes of non-solids was also used in the training input patterns.

endpoint of learning was recorded as either asymptotic performance of learning all 50 words, or at the end of 500 epochs of training. This resulted in ten checkpoints along the trajectory of word learning. Although other measures such as duration of training could capture how much a network has learned, we used number of words learned because this is the key factor driving the development of biases.

**Training.** On each trial of training, a word was paired with a referent. The patterns associated with each word were determined based on which noun category that word was meant to represent. For example, a word for a solid item characterized by shape (like a ball; see Figure 1) should be used to label things that are like each other in shape but differ from each other in material. To simulate this pattern, we randomly selected an input vector to represent, for example, ball shape. On individual training trials, we paired that shape pattern with the label *ball* and a randomly selected material pattern. Therefore over multiple training trials, a word for a solid item characterized by shape would be represented by the same shape but different material patterns. We did this for each of the 50 nouns in the training set. This part of the simulation was intended to put into the network the lexical knowledge that a typical child would bring to the laboratory NNG task.

**Testing** Following training, the next step was to identify what sort of word learning biases each run of the network had developed. We addressed this question with a virtual version of the NNG task. We presented the network with three NNG tasks, one for each of the three types of items: solids, clear non-solids, and ambiguous non-solids. On each test trial of these tasks, we presented the network with a triad of novel entities (one at a time) on the perceptual layer. The triad consisted of an exemplar and two choice items, one matching the exemplar in shape only and one matching in material only. The only difference between the trials for the clear non-solids and those for ambiguous non-solids was that for clear non-solids, the material-matching choice item had a simplified pattern along the shape layer, as discussed in the input patterns section. For each of these three inputs, we recorded the resulting pattern of activation on the hidden layer. This is a measure of how the network represents these items. If the network emphasizes the shape of the item, then the similarities of the internal representations for the exemplar and its shape matching choice should be *greater* than the similarity of the internal representations for the exemplar and the material matching choice. If this same relationship is *less*, then the internal representations highlight the material of the items. We used these similarities along with Luce's choice rule (Luce, 1963) to calculate probability of choice in order to predict performance in the NNG task.

## Results

We averaged over 10 runs with different initial random weights. First we looked at the network's test output across the entire course of learning. As shown in Figure 2, the network preferred the shape match choices at test to different extents depending on both the solidity of the item presented and the size of its vocabulary at that point. For example, before training, the network showed no preference for either the shape match or the material match test choice for solid items, tending to choose the shape match about half the time ( $M = .50$ ,  $SD = .02$ ). In contrast, by the end of training the network chose the shape match for solid items about three quarters of the time at test ( $M = .78$ ,  $SD = .04$ ).

On the other hand, while the network started at a similar state for complex-shaped, ambiguous non-solids, it showed a different preference by the end of training. Specifically, the network began with no preference for shape or material, but developed a shape preference for these non-solids, albeit to a lesser extent than it did with solid items ( $M = .62$ ,  $SD = .03$ ). This pattern shows an overgeneralization of the shape bias to this particular kind of non-solid item.

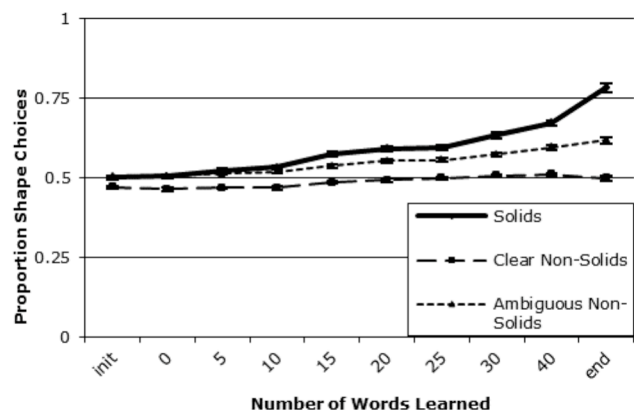


Figure 2: Mean proportion of shape match choices at test for each kind of test item. Error bars represent standard error.

Still another pattern was seen in how the network treated simple-shaped, clear non-solids at test. In this case, the network actually began with a preference for material, choosing the shape match test choices slightly but significantly less than 50% of the time ( $M = .47$ ,  $SD = .01$ ,  $t(9) = -8.31$ ,  $p < .001$ ). This inherent material bias was present at initialization due to the input structure of the clear non-solids items. Recall that for this type of item, half of the shape units were kept constant across all of the input. Because of this, at the time of initialization the networks had less information about variations in shape on which to base representations. Instead, the networks harnessed the relatively richer material information immediately available about clear non-solid items, and thus showed a slight preference for this feature initially. However, by the end of learning the initial preference for material was gone ( $M = .50$ ,  $SD = .02$ ). What caused the network to lose this early preference for material in the context of clear non-solids? And more specifically, did it have anything to do with what



was concurrently being learned about solid items? To get at this question, we next focused our analysis on the time window in which the shape bias emerged in learning.

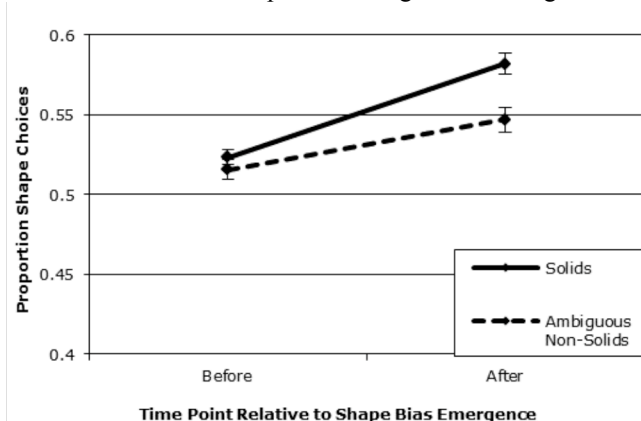


Figure 3: Mean proportion of shape match choices at test for solids and ambiguous non-solids immediately before and after the emergence of the shape bias for solids.

Identifying the point in learning at which the network first developed a shape bias involved several steps. First we examined the output of each of the ten runs and picked out when in learning the network chose the shape match test choice for solid items greater than 55% of the time.<sup>2</sup> This happened on average around the time the network had learned 15 words from the training set. The next step was to identify how this point of emergence of the shape bias in solids affects the network's behavior concerning the two kinds of non-solids. We isolated the closest time points in learning both preceding and following the emergence of the shape bias. Using this focused time window, we examined how the network treated each type of non-solid item in relation to solid items as the shape bias for solids emerged.

To examine the interaction between the emerging shape bias and the network's choices among complex-shaped, ambiguous non-solid test items, proportion of shape choices were submitted to a 2 (test item type: solid or ambiguous non-solid)  $\times$  2 (time point in learning: before or after the emergence of the shape bias) mixed design analysis of variance (ANOVA).<sup>3</sup> This analysis yielded a significant main effect of time point,  $F(1, 30) = 13.41, p = .001$ , with no other significant effects. This shows that as the shape bias for solids emerges in the course of learning, a similar bias develops for ambiguous non-solids (see Figure 3), suggesting an overgeneralization of the shape bias to ambiguous non-solids. Further, this shape bias for both types of items increases over the time window of interest.

Next we examined the interaction between the emerging shape bias and the network's choices among simply-shaped, clear non-solid test items. Proportion of shape choices were

submitted to a similar 2 (test item type: solid or clear non-solid)  $\times$  2 (time point) ANOVA. This analysis yielded main effects of both test item type ( $F(1, 30) = 25.47, p < .001$ ) and time point ( $F(1, 30) = 11.14, p < .01$ ). These main effects are qualified by a significant interaction between test item type and time point in learning,  $F(1, 30) = 8.73, p < .01$ . As can be seen in Figure 4, as the shape bias for solids emerges, the network's preference for shape choices for clear non-solids also increases, however the nature of these changes differs between item types. In fact, the network shows a preference for material test choices for clear non-solids, even just after the emergence of the shape bias for solids. Although this preference for material diminishes somewhat over the time window in question (as shown by an increase in proportion of shape choices), it does not do so at a rate proportional to the growth of the shape bias for solids. This suggests that the emergence of the shape bias may have a slight diminishing influence on the material bias for clear non-solids.

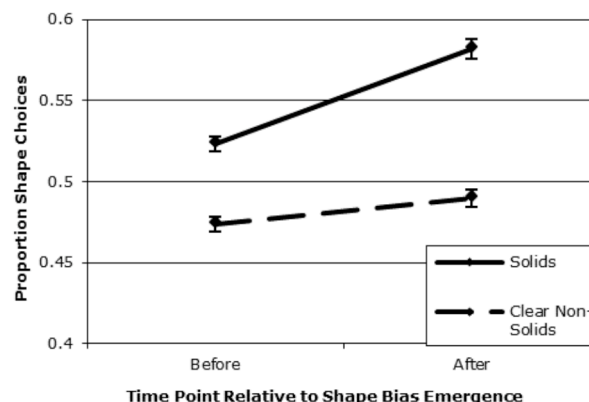


Figure 4: Mean proportion of shape match choices at test for solids and clear non-solids immediately before and after the emergence of the shape bias for solids.

Finally, to check that these patterns were specific to the time window surrounding the emergence of the shape bias, we ran the same analyses for the immediately following time window. The only effect that reached significance was a main effect of item type between solid and clear non-solid test items,  $F(1, 32) = 37.62, p < .001$ . As in the above analysis, this reflects an ongoing difference in the extent of shape choices that the network made in the context of these two types of items. The fact that no other effects were significant supports the argument that the preceding results are specific to the time window in learning immediately surrounding the emergence of the shape bias.

## Discussion

The current study investigated the dynamics involved in the development of the shape bias for solid items and learning about other kinds of items in early language acquisition. We found that the emergence of the shape bias for solids led to either an overgeneralization of the shape bias or a diminishment of the material bias, depending on the

<sup>2</sup> 55% was chosen as the threshold for emergence because at this value the network chose shape matches reliably greater than chance across the ten runs,  $t(9) = 12.54, p < .001$ .

<sup>3</sup> This type of analysis was chosen following Colunga & Smith (2005) in their analyses of similar networks.

strength of the cues to solidity or non-solidity provided by the test items, depending on the type of test item used.

First, for complex-shaped, ambiguous non-solid test items, we observed an overgeneralization of the shape bias. This is consistent with findings that 3-year-old children overgeneralize the shape bias in extending names to deformable items (Samuelson et al., 2008). This suggests that for ambiguous categories, the word learning bias that is already established, typically the shape bias, takes precedence and guides generalization.

Second, for simply-shaped, clear non-solid test items, the material bias diminished as the shape bias for solids developed, although it did so at a slower rate compared to the growth of the shape bias. This suggests that the material bias for clear non-solids was more resistant to the influence of the developing shape bias, perhaps in part due to the inherent material preference for these items. This inherent bias was due to the structure of the input itself, but a similar early material bias has also been seen in children if they are tested with simply-shaped non-solid substances on the NNG task (Colunga & Smith, 2005).

The mechanisms driving the observed changes in bias learning and development in our networks speak directly to possible mechanisms in children. The current work suggests that the network can mainly develop one bias at a time, which would explain why as the shape bias for solids takes off, the network also shows a growing preference for shape for the two other types of items. If this is the case, we would predict that as the network continued to train on more words, the material bias for non-solids would come online (as seen in children) and cause a dip in the established shape bias for solids. However, by this time the shape bias for solids would be well established and thus largely resilient against the network's shift to focusing on material. After enough exposure to the vocabulary structure, we would expect the network to have acquired both a shape bias for solids and a material bias for non-solids, as is seen in children later in language development.

This work could be extended to make predictions about and map to specific places of bias emergence in individual children's word learning. By using longitudinal MCDI data from individual children, one could use networks to analyze the development of language biases in early- and late-talkers (as done in Colunga and Sims 2011), looking specifically at how the emergence of certain word learning biases affects other biases over time. These biases could show different patterns of interaction among early- and late-talkers. In sum, this work opens the door for further modeling and can make novel, testable predictions about the development of children's word learning.

### Acknowledgments

Research supported by an award from the John Merck Fund and NICHD R01 HD067315 to Eliana Colunga.

### References

- Apfelbaum, K.S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, 35, 1105-1138.
- Colunga, E., & Sims, C.E. (2011). Early talkers and late talkers know nouns that license different word learning biases. In L. Carlson, C. Hoelscher, & T. Shipley (Eds.), *Proc. of the 33<sup>rd</sup> Annual Conference of the Cognitive Science Society* (pp. 2550-2555). Austin, TX.
- Colunga, E., & Smith, L.B. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, 112, 347-382.
- Fenson, L., Dale, P., Reznick, J.S., Thal, D., Bates, E., Hartung, J., et al. (1993). *The MacArthur Communicative Developmental Inventory: User's guide and technical manual*. San Diego, CA: Singular Publishing.
- Gershkoff-Stowe, L. & Smith, L.B. (2004). Shape and the first hundred nouns. *Child Development*, 75, 1098-1114.
- Luce, R.D. (1963). Detection and recognition. In R.D. Luce, R.R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 1-41). New York: Wiley.
- Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*, 117, 1-31.
- Munakata, Y., & McClelland, J.L. (2003). Connectionist models of development. *Dev. Science*, 6, 413-429.
- O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., (2012). *Computational Cognitive Neuroscience*. Wiki Book, 1st Edition. URL: <http://ccnbook.colorado.edu>
- Samuelson, L.K. & Bloom, P. (2008). Special section: What counts as an explanation of development: The shape bias as a case study. *Developmental Science*, 11, 183-184.
- Samuelson, L.K., Horst, J.S., Schutte, A.R., & Dobbertin, B.N. (2008). Rigid thinking about deformables: Do children sometimes overgeneralize the shape bias? *J. of Child Language*, 35, 559-589.
- Samuelson, L.K., Schutte, A.R., & Horst, J.S. (2009). The dynamic nature of knowledge: Insights from a dynamic field model of children's novel noun generalization. *Cognition*, 110, 322-345.
- Samuelson, L.K., & Smith, L.B. (2000). Children's attention to rigid and deformable shape in naming and non-naming tasks. *Child Development*, 71, 1555-1570.
- Smith, L.B., Jones, S.S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L.K. (2002). Creating a shape bias creates rapid word learners. *Psychological Science*, 13, 13-19.
- Soja, N.N. (1992). Inferences about the meaning of nouns: The relationship between perception and syntax. *Cognitive Development*, 7, 29-45.
- Soja, N.N., Carey, S., & Spelke, E.S. (1991). Ontological categories guide young children's inductions of word meanings: Object terms and substance terms. *Cognition*, 38, 179-211.
- Xu, F., & Tenenbaum, J.B. (2007). Word learning as Bayesian inference. *Psych. Review*, 114, 245-272.
- Yoshida, H. & Smith, L.B. (2005). Linguistic cues enhance the learning of perceptual cues. *Psych. Science*, 16, 90-95.



# Interactions between abstract actions and apparent distance

Kathryn Sears ([kathryn.sears@richmond.edu](mailto:kathryn.sears@richmond.edu))

David Landy ([dlandy@richmond.edu](mailto:dlandy@richmond.edu)),

Jessica Lesky ([jessica.lesky@richmond.edu](mailto:jessica.lesky@richmond.edu))

Department of Psychology, University of Richmond  
Richmond, VA 23173

## Abstract

Perceptions influence the way we act in our environment based upon judgments assessing required efforts to perform an action and the availability and demand for immediate action on an object (Proffitt, 2006B). Social and physical anxiety has been shown to distort perceptions of depth and perceptions of object size (Stefanucci et al., 2008; Cañal-Bruland et al., 2010). Relatively little work, however, has explored the potential role of depth perception in abstract reasoning tasks (Landy & Linkenauger, 2010). In Experiment 1, the relationship between depth perception and the order of actions taken to simplify arithmetic expressions was investigated by manipulating apparent distances of arithmetic operations of high and low syntactic precedence. When the high precedence operations appeared to be closer to the participant, expressions were solved more quickly than when low precedence expressions appeared to be closer. Experiment 2 explored the whether the affordance of abstract actions conversely impacted perceived distance by asking participants to make distance judgments to multiplication and addition operations. Experiment 2 found no impact of anxiety about mathematics on perceived distance. However, effects resulting from condition assignment were found to influence perceived distances, as well as solving strategy. We interpret results in terms of attention, which we speculate plays a key role moderating both ordering behavior and perceived distance.

**Keywords:** Perception; Anxiety; Attention

## Introduction

The basic purpose of perception is to guide sensible action. Specifically, perception functions as a tool that informs and guides actions (Proffitt, 2006A). Perception involves consideration of the necessary efforts required to perform an action, which in turn may bias judgments of object size and proximity. For instance, perceived distance of an object is more than simple metric distance. Participants carrying a heavy backpack estimate the slant of hills to be steeper than those not wearing backpacks (Proffitt, 2006B). Likewise, simply intending to act on an object could make it seem closer Witt & Proffitt (2005). Witt & Proffitt (2005) found that athletes playing well reported a softball ball as being bigger than its actual size and therefore closer. Their ability to hit the ball was correlated with their batting average, demonstrating the relationship between perceptions of object size and performance when acting on that object. Perceptual phenomena like these indicate that perception relates the body and goals to the opportunities and costs of acting in the physical environment.

Although the connection between concrete physical action capabilities and depth perception is well documented, much less work has explored the relationship between perceived depth and non-concrete behaviors. There is evidence that anxiety also influences perception (Proffitt, 2006A). For instance, participants with a fear of heights judge hills to be steeper than do those who are less afraid (Stefanucci et al, 2008). According to Teachman et al (2008), fear of heights is associated with perceptual biases in judging heights, implying that an individual's emotional state influences what is seen, perhaps because it affects perceived costs, such as the cost of falling down a steep hill.

Anxiety not only distorts perception in the sense that distances appear farther and angles steeper, but also by altering the appearance of an object's size. Cañal-Bruland, Pijpers, & Oudejans (2010) studied the relationship between anxiety and depth perception, while also taking into consideration perceptions of object size. Participants were asked to throw darts at a target from a position on a rock wall. Cañal-Bruland, et al. found that the low anxiety group performed better and saw the target as bigger. However, these findings are limited to perceptions related to actions in the physical world.

Even less is known about cases in which the action itself is abstract or non-physical. In the case of abstract calculation, since all actions are in principle equally easy to perform, the logic of perceived energetic cost per se does not predict any relationship between depth and intentions to act. However, mathematical rules specify that certain operations are to be performed before other others. Particularly relevant for the studies reported here are the order of precedence rules, which require that multiplications be executed before additions. As a result, some operations demand action before others, making it possible to vary the relative availability of actions. Furthermore, since the actions do not depend on reach distance, any action lies within the action boundary (Fajen, 2005). Landy & Linkenauger (2010) found a relationship between the availability of computational actions in compound arithmetic expressions and judgments of depth. The study explored judged depth of terms, and indeed found that participants in a forced-choice task preferred to align depth and precedence; furthermore, in a face-vase illusion in which sub-expressions of mathematical equations were superimposed onto a face-vase illusion, when the times sign was over the vases, participants reported seeing the faces less often than when the times sign was over the faces. These perceptual effects can be explained by the

affordances of immediate action associated with the times sign as a result of taking precedence when using order of operations in solving arithmetic expressions. The current work builds on Landy & Linkenauger (2010) by considering whether perceived (rather than judged) distance interacts with the availability of concrete action

Visual attention also plays a significant influence on action-specific perception and performance. In their study, Cañal-Bruland, Zhu, van der Kamp, & Masters (2011) explored this relationship through a golf-putting task that manipulated the target-directed visual path. They found that participants receiving full visual access to the target and who putted more successfully estimated the target circle to be bigger than their less successful counterparts. Thus, the results of this study have shown that visual attention influences perceptions object size, but only for objects of an intended action. The relationship between anxiety and depth perception is abstract when attentional influence is taken into consideration. The current study aims to further explore attention as a moderating variable in the relationship between math anxiety and depth perception.

Math anxiety is defined by a strong tendency to avoid math, which leads to lower competency levels in math compared to those without math anxiety (Ashcraft, 2002). Hoffman (2010) posits that anxiety is a common impediment to learning in college students. Anxiety has shown to impede working memory processes involved with problem-solving efficiency, especially for women. Implications from prior research suggest that more needs to be done in order to understand the perceptual difficulties associated with math anxiety, especially if research can identify attention as a target for future interventions.

The purposes of the current studies were as follows: first, to establish whether the effect on judgments of apparent distance in pictures reported by Landy & Linkenauger (2010) generalized to perception in physical situations and second, to evaluate the influence of anxiety in distorting perceptions of depth in abstract situations.

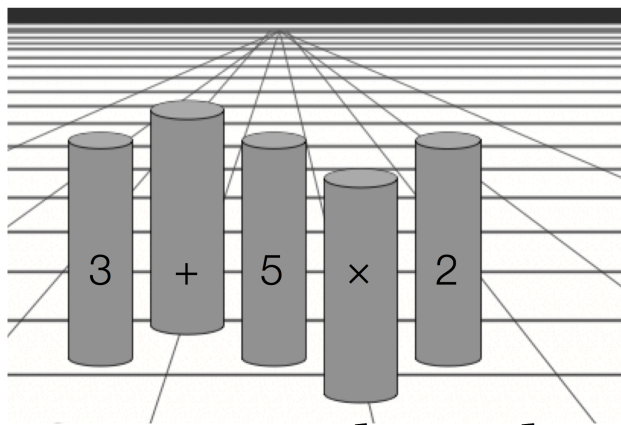


Figure 1: The stimuli used in Experiment 1. In this congruent problem, the times sign (which is high precedence) appears to be closer to the subject.

## Experiment 1

### Participants & Procedure

Thirty members of the University of Richmond community were given partial course credit in exchange for participation.

Participants sat in front of a computer. Participants were shown 128 simple arithmetic problems, and were instructed to state the solution to the problems out loud. Before beginning, participants were reminded of the order of operations through direct instruction and an example, and were explicitly instructed to ignore any irrelevant images or visual structure. Responses were recorded by microphone and analyzed using CheckVocal (Protopapas, 2007).

All trials involved single digit operands, and were of one of two forms:  $a+b \times c$  or  $a \times b+c$ . Correct solution values ranged from 13 to 76.

The first 8 trials were practice trials. In these trials, problems were presented against a white background. In the remaining 120 trials, problems were presented against a background image intended to affect the apparent depth of the operators and operands. In this image (see Figure 1), stimulus problems appeared to be placed on a set of pillars, which varied in whether the left operator appeared closer (*left-closer* condition), or the right (the *right-closer* condition).

We hypothesized that problems would be solved more easily when the high-precedence operation appeared to be closer to the participant (called *congruent* problems). Following comparable results in the manipulation of spacing cues (Landy & Goldstone, 2010) we expected the alignment of precedence and distance cues to selectively influence order of operation reversals and operation errors (e.g., performing a multiplication instead of an addition), and to affect correct-trial response time.

### Results

Because trial RTs were substantially non-normal the median response time for each condition was calculated for each subject, and subjected to a standard ANOVA. Pillar and operation structure functioned as independent variables (see Table 1). Neither main effect was significant (pillars:  $F(1,28)=0.04$ ,  $p \sim 0.85$ , operations:  $F(1, 28)=1.2$ ,  $p \sim 0.3$ ); the interaction was significant and in the predicted direction ( $F(1,28)=10.4$ ,  $p < 0.01$ ).

Mean accuracy was 0.90 (Min=0.79, max=1.0). Participants made very few order of operations errors and operation errors: together, these made up just 2.9% of responses. There was no difference between problems that aligned pillars and operation structure and problems that misaligned them. The error rate was 0.1 (SE=0.01) on incongruent problems, and 0.06 on congruent problems (SE=0.01) on congruent problems. A logistic regression over operation order and pillar structure revealed neither a main effect (operations:  $z=0.4$ , pillars:  $z=0.16$ , all  $p$ 's  $> 0.5$ ), nor an interaction between the two ( $z=0.5$ ,  $p \sim 0.6$ ). Results

Table 2: Mean response time for accurate trials in Experiment 1 (Errors are standard errors).

Leftmost operation	Closer Pillar	
	Left	Right
Plus	2619 (127)	2516 (111)
Times	2465 (129)	2554 (131)

were nearly identical when just order and operation errors were considered: The mean rate of operation errors and order errors was 0.03 (SE=0.007) and 0.028 (SE=0.007) for congruent and incongruent problems, respectively; there was no difference between these by a logistic regression over pillar structure and operation (pillar  $z = -1.1$ , operation order = -0.1, interaction  $z = -0.5$ , all  $p$ 's > 0.61).

## Discussion

Apparent distance did, as predicted, influence the execution of basic arithmetic problems. Despite the equal availability and readability of all terms in these problems (i.e., all problems were presented on identical local backgrounds, at identical sizes), problems in which high precedence problems were apparently nearer to the participant were solved more quickly than those in which the reverse was true. This suggests that reasoners use cues about physical structure when making abstract decisions such as operation ordering. Experiment 1 indicates a bidirectional relationship between apparent distance and arithmetic ordering.

In other similar work (e.g., Landy & Goldstone, 2010), we have typically observed both accuracy and response time effects. One difference between those and these was that the rate of errors in this study was very low overall (just 10%), and the rate of errors expected to be related to operation order was even lower (just 2.9% overall). The population of students in the current study was generally efficient at basic arithmetic. This fact, together with the direct reminder to follow the order of operations, may have shielded participants from making many direct strategic errors when solving the expressions.

## Experiment 2

Experiment 2 explored the relationship between perceived distance, action affordance, and experienced anxiety in an abstract domain--arithmetic. Participants judged the distance to specific symbols inside an arithmetic expression, which either did or did not afford immediate action. We hypothesized, first, that participants who knew the order of operations would judge times signs as closer than plus signs, since in unparenthesized expressions with multiple operations multiplications afford more immediate action than additions do. We explored whether demonstrations that anxiety distorts perceptions of action affordances would generalize to abstract actions and situations.

## Participants

The 32 participants of this study were from the Richmond community between the ages of 18 and 30. They were recruited through flyers posted around the University of Richmond campus, as well as, through weekly online campus announcements.

## Apparatus and Materials

**Experimental Room.** The room was set-up with two identical tables (137 cm x 76 cm) joined at a corner, forming an L-shape. The participant's chair was positioned at the corner of these tables facing the wall. Both tables were positioned approximately 23 cm from the wall. The front legs of the chair were positioned 12 inches from the front outer-legs of the table.

**Boards and Measuring Key.** Expressions were displayed on boards made of a white foam poster board, cut into rectangles (45.7 cm x 30.5 cm) and supported by 30.5 cm easel backs. The boards angled slightly, so that the top was slightly further from the participant than the bottom.

A total of eight boards were made, one for each arithmetic expression. The boards were placed at distances both within and out of reach to the participant. Measured from the back of the chair, expressions were placed at four distances: 80cm, 40cm, 110cm, and 50cm.

**Questionnaires.** Participants completed the Abbreviated Math Anxiety Scale (Hopko et al., 2003) and WRAT-3 assessment (Roberston, 2010), as well as a questionnaire assessing each participant's particular math background and interests.

## Procedure

Participants sat in a chair facing the corner of the two tables. To familiarize the participant with the procedure, a warm up trial consisting single row of five dots was presented on the right-hand table at a distance of 80cm. Participants were informed that following the warm-up, the board would display arithmetic expressions instead of dots. Because we were interested in the potential long-term influence of operation ordering practice on distance perception, participants were not instructed in a particular solving method. In particular, participants were not reminded of the order of operations, but were simply asked to solve the problems as they usually would, and to do their best.

Participants were instructed to estimate the distance from their sternum to the center dot. Participants indicated their estimate by extending the tape measure along the left-hand table, with the blank side of a tape measure up. This method allowed the participant to have full control over the estimation process, while blinding them to the actual measurement. When providing their estimate, participants were not permitted to use the tape measure on the table displaying the board, nor were they allowed to reach out and touch the board.

During actual experimentation, the procedure followed in the same manner as the warm-up, substituting arithmetic expressions for the row of dots. Participants viewed four expression pairs displayed on boards set at one of four distances (80cm, 40cm, 110cm, 50cm) from the edge of the table closest to the participant. We repeated the procedure for each participant and with variance in the complexity of the math problems: four multiplication and four addition, with two easy and two complicated in each. Problems were paired, so that for each problem a participant solved while focusing on a multiplication sign, they later solved a problem identical except for the substitution of additions for multiplications. For example, if given the problem  $5+4 \times 3+7$ , we asked them to focus on the multiplication sign when giving their estimate. Immediately following their estimate, they proceeded to solve the expression aloud, step-by-step.

The order of distances was fixed, but the problem presentations were counterbalanced in a Latin Square design, leading to a total of eight conditions (see *Table 1*).

Table 2: Stimuli for Experiment 2 (Condition 1)

Expression	Center Operation	Distance (cm)	Difficulty
$3 \times 5 + 2 \times 7$	Add	80	Easy
$1/3 \times 4 + 2/3 \times 1/2$	Add	40	Difficult
$5 \times 8 + 2 \times 3$	Add	110	Easy
$1/4 \times 1/6 + 5 \times 2/3$	Add	50	Difficult
$3 + 5 \times 2 + 7$	Mult	80	Easy
$1/3 + 4 \times 2/3 + 1/2$	Mult	40	Difficult
$5 + 8 \times 2 + 3$	Mult	110	Easy
$1/4 + 1/6 \times 5 + 2/3$	Mult	50	Difficult

Each expression pair was set at the same distance. Thus, if  $5+4 \times 3+7$  was set at 50 cm, then  $5 \times 4+3 \times 7$  was also set at 50 cm. Between trials the participant was given a packet of mazes and instructed to complete as much of it as they could while the experimenter set up the next trial.

Following the distance perception session, the math anxiety level and abilities of each participant was assessed with a basic mathematics proficiency test, WRAT-3. Each participant was given 10 minutes to complete as much of the packet as they could, without returning to the problems they skipped. After the WRAT-3, participants completed the Abbreviated Math Anxiety scale.

## Results

Distance estimates (see *Table 3*) were subjected to an ANOVA using operation structure and problem difficulty, within-participants factors, and math anxiety and

precedence behavior as between-participant factors. Participant precedence behavior was coded by the experimenter as either the correct use of order of operations or the incorrect use. Participants who did not use the order of operations correctly tended to use other strategies, such as adding from left to right, or computing sums before products. It was typically difficult to discern what strategy a participant had used, and some participants tended to shift strategies over the course of a trial. To be clear: differences in operations based on precedence strategy are expected only among subjects who apply correct order of operations. There was no significant main effect of math anxiety on distance estimates ( $F(1,30)=.02, p \sim .9$ ) or in the interaction between math anxiety and problem difficulty on distance estimation ( $F(1, 30) = 1.10, p > .05$ ). There was, however, a significant interaction between the operation in focus and precedence behavior ( $F(1, 30)=5.0, p<0.05$ ), such that operations which were treated as high precedence yielded lower distance estimates. A follow-up analysis considering just participants who correctly used order of operations revealed a significant main effect of focal operation ( $F(1, 17) = 4.87, p < .05$ ). There was no significant interaction for who did not obey standard precedence rules; numerically, this group tended to estimate the plus signs as closer than the times signs ( $F(1, 12)=3.1, p \sim 0.11$ ).

Post-hoc analyses revealed a possible confound of the counterbalance condition, in that the condition in which the participant was randomly assigned seemed to influence whether or not they estimated the plus or times signs to be closer and the strategy used to solve the expressions. Specifically, those who received the conditions presenting the times sign as the operation in focus first tended to use order of operations when solving the expressions (87.5% correct precedence behavior), whereas those receiving the plus sign as the operation in focus tended not to (31% correct precedence). The effect of first operation on precedence behavior was significant by Fisher's exact test ( $p<0.01$ ); however, since this effect was not predicted ahead of time, it should be interpreted cautiously. To ensure that the main analyses were not affected by an overall linear trend in distance judgments across trials, the data were reanalyzed using problem order as a covariate; results were in all ways similar to those reported above.

Table 3: Mean (standard error) distance underestimated distance to central operations in Experiment 2.

Central operation	Precedence Behavior	
	Correct	Reversed or Left/Right
Plus	14.2 (1.7)	17.2 (1.4)
Times	16.8 (1.9)	15.6 (1.4)

## Discussion

Although our hypothesis that math anxiety would lead to overestimation of distances was not supported, our findings support the prediction that those who knew the order of operations would perceive the times sign to be closer. However, it is quite possible that null findings could be a result of such a small sample size. Taken together with the results from Experiment 1, high precedence operations afford immediate action and therefore, may appear to be closer. The effect of the experimental conditions suggest that attention may serve as a moderating variable between depth perception, anxiety, and the perceived efforts to enact an action.

## General Discussion

Two experiments verified and extended the basic findings of Landy & Linkenauger (2009), demonstrating a bidirectional relationship between perceived distance and effective arithmetic syntax. While prior results indicated a metaphorical relationship between distance and precedence, Experiment 2 here demonstrated that participants who correctly apply the order of operations also estimate the actual physical distance to high precedence sign as smaller than that to a low precedence sign. Furthermore, these are the first results to demonstrate that (simulated) physical distance affects the application of abstract formal operations.

Both of these phenomena are familiar in interactions with physical objects. Intentions and object affordances both impact perceived distance (Proffitt, 2006B). We know of no research directly exploring the impact of apparent distance on action selection, but it seems quite likely that actions with physical objects are selected in part on the basis of perceived proximity. In the current experiments, however, the relevant action is itself abstract, and the ease of adding and multiplying does not depend in any obvious way on physical proximity. The interpretation of interactions between apparent distance and abstract actions is thus less clear than with concrete objects. A dramatic reading might hold that explicit distance perception tracks the perceived difficulty of engaging in behavior—that is, that explicit perceptions of distance are fundamentally less concrete and more abstract than has previously been supposed. The fact that we did not find any sign of a relationship between math anxiety, problem difficulty, and perceived distance speaks against such a dramatic conclusion.

A more plausible interpretation is that abstract procedures, such as calculation are executed via systems normally devoted to perception and action (Landy & Goldstone, 2007, 2009; Goldstone, Landy, and Son, 2010; Landy, Allen, and Anderson, 2011). On this “Rigging Up Perceptual Systems” account, learning to engage in formal operations, such as operation ordering often involves adapting a pre-existing perceptual-motor system that already performs computations roughly appropriate to the to-be-learned content. Previously identified systems include the use of perceptual grouping and attention to perform

operation ordering (Landy & Goldstone, 2007, 2010; Goldstone et al 2010); it may be that some individuals implement operation ordering via distance perception mechanisms, learning to treat sub-expressions which should be ignored as farther away, and so leveraging powerful machinery that produces distance judgments to automatize routine computation.

It also seems possible that differences in perceived distance may simply be a result of focused attention: that is, it is possible that simply attending to an object increases its apparent proximity. Though we know of no direct demonstration of this possibility, attention is thought to affect other aspects of perception, such as contrast (Treue, 2004), apparent speed of motion (Turatto et al, 2007), and apparent size (Anton-erxleben & Treue, 2007; but see Schneider, 2008). Attentional focus may also influence perceived distance. This explanation provides a natural account for the influence of precedence judged distance and figure and ground perception in the face-vase illusion previously demonstrated by Landy & Linkenauger (2009). Furthermore, attention is thought to be necessary for action-specific effects on perceived distance (Cañal-Bruland et al., 2011) Finally, though it should be interpreted with caution, the unpredicted relationship between initial focal operation and both distance and precedence behavior is also compatible with an attention-based interpretation. It may be that asking people to attend to a particular sign influenced both ordering behavior and perceived depth.

## Conclusions

In our experiments, we did not find any effect of perceived difficulty on estimated distance, either when problem difficulty varied, nor based on personal skill or arithmetic self-efficacy. Of course, null results such as these must be interpreted with caution; nevertheless, there is little indication here of a very tight analogy between estimations of abstract difficulty and perceptions of physical distance.

On the other hand, we found substantial bidirectional influences between order of operations and perceived depth, suggesting that the relationship between action ordering and depth is not restricted to concrete behaviors, but is also involved in abstract actions as well.

## Acknowledgments

This research was partially funded by Department of Education, Institute of Education Sciences grant R305A110060, as well as an undergraduate research grant from the University of Richmond. Beth Crawford provided helpful comments on the design of Experiment 2, and Sally Linkenauger on the design of Experiment 1.

## References

- Anton-Erxleben, K., Henrich, C., & Treue, S. (2007). Attention changes perceived size of moving visual patterns. *Journal of Vision*, 7(11):5, 1-9.

- Cañal-Bruland, R., Pijpers, J., & Oudejans, R. (2010). The influence of anxiety on action-specific perception. *Anxiety, Stress, & Coping*, 23(3), 353-361.
- Cañal-Bruland, R., Zhu, F., van der Kamp, J., & Masters, R. (2011). Target-directed visual attention is a prerequisite for action-specific perception. *Acta Psychologica*, 136, 285-289.
- Fajen, B. R. (2005). The scaling of information to action in visually guided braking. *Journal of Experimental Psychology: Human Perception and Performance*, 31, 1107-23.
- Goldstone, R.L., Landy, D., & Son, J. Y. (2010). The Education of Perception. *Topics in Cognitive Science*, 2(2), 265-284.
- Hoffman, B. (2010). "I think I can, but I'm afraid to try": The role of self-efficacy beliefs and mathematical anxiety in mathematics problem-solving efficiency. *Learning and Individual Differences*, 20(3), 276-283.
- Hopko, D.R., Mahadevan, R., Bare, R.L., & Hunt, M.K. (2003). Abbreviated math anxiety scale (AMAS): construction, validity, and reliability. *Assessment*, 10(2), 178-182.
- Landy, D., Allen, C., & Anderson, M. L. (2011). Conceptual discontinuity through recycling old processes in new domains. *Behavioral and Brain Sciences*, 34(3), 136-137.
- Landy, D., & Goldstone, R. L. (2007). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(4), 720-733.
- Landy, D., & Goldstone, R. L. (2010). Proximity and precedence in arithmetic. *Quarterly Journal of Experimental Psychology*, 63(10), 1953-1968.
- Landy, D., & Linkenauger, S.A. (2010). Arithmetic Notation...now in 3D! *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Austin, TX, USA.
- Proffitt, D. R. (2006A). Distance Perception. *Psychological Science*, 15(3), 131-135.
- Proffitt, D.R. (2006B). Embodied perception and the economy of action. *Perspectives on Psychological Science*, 1(2), 110-122.
- Protopapas, A. (2007) CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods*, 39(4), 859-862.
- Roberston, G.J. (2010). Wide-range achievement test. *Corsini Encyclopedia of Psychology*, 1-2.
- Schnall, S., Harber, K., Stefanucci, J., & Proffitt, D. (2008). Social support and the perception of geographical slant. *Journal of Experimental Social Psychology*, 44(5), 1246-1255.
- Schneider, K. A. (2008). Attention biases decisions but does not alter appearance. *Journal of Vision*, 8, 1-10.
- Stefanucci, J. K., & Proffitt, D. R. (2009). The roles of altitude and fear in the perception of heights. *Journal of Experimental Psychology: Human Perception & Performance*, 35, 424-438.
- Stefanucci, J. K., Proffitt, D. R., Clore, G., & Parekh, N. (2008). Skating down a steeper slope: Fear influences the perception of geographical slant. *Perception*, 37, 321-323.
- Teachman, B.A., Stefanucci, J.K., Clerkin, E.M., Cody, M.W., & Proffitt, D.R. (2008). A new mode of fear expression: Perceptual bias in height fear. *Emotion*, 8(2), 296- 301.
- Treue, S. (2004). Perceptual enhancement of contrast by attention. *Trends in cognitive sciences*, 8(10), 435-7.
- Turatto, M., Vescovi, M., & Valsecchi, M. (2007). Attention makes moving objects be perceived to move faster. *Vision research*, 47(2), 166-78.
- Witt, J.K., & Proffitt, D.R. (2005). See the ball, hit the ball: Apparent ball size is correlated with batting average. *Psychological Science*, 16(12), 937-938.

# Vection (self-motion perception) alters cognitive states, cognition of time, mental number line and personality

**Takeharu Seno (seno@design.kyushu-u.ac.jp)**

Faculty of design, Kyushu University  
4-9-1, Shiobaru, Minami-ku, Fukuoka, Japan

**Shuichiro Taya (s.taya@qmul.ac.uk)**

School of Biological and Chemical Science, Queen Mary Collage,  
University of London, Mile End Road, London E1 4NS

**Yuki Yamada (yamadayuk@gmail.com)**

Faculty of Human-Environment Studies, Kyushu University  
6-19-1 Hakozaki, Higashi-ku, Fukuoka Japan

**Keiko Ihaya (ihayakk@gmail.com)**

Faculty of Human-Environment Studies, Kyushu University  
6-19-1 Hakozaki, Higashi-ku, Fukuoka Japan

**Hiroyuki Ito (ito@ design.kyushu-u.ac.jp)**

Faculty of design, Kyushu University  
4-9-1, Shiobaru, Minami-ku, Fukuoka, Japan

**Shoji Sunaga (sunaga@ design.kyushu-u.ac.jp)**

Faculty of design, Kyushu University  
4-9-1, Shiobaru, Minami-ku, Fukuoka, Japan

## Abstract

We examined the relationship between cognitive states and visually-induced self-motion perception, i.e. “vection” (latency, duration and magnitude). It is often anecdotally reported that time experienced in return travel (back to the start point) seems shorter than time spent in outward travel (travel to a new destination). Here, we report the first experimental results showing that return travel time is experienced as shorter than the actual time because of perceiving vection. Secondary, we explore how numbers are represented in depth in our mental space, we asked participants to sequentially speak random numbers while they observed forward/backward vection. We found that participants tended to generate larger numbers when they perceived backward self-motion. Finally, We found that all the measures of vection correlated negatively with the degree of narcissistic traits of participants.

**Keywords:** vection, time perception, mental number line, personality

## Introduction

Self-motion perception as determined by vision alone is called ‘vection’ (e.g. Fischer & Kornmüller, 1930). Stimulus attributes for effective vection induction have been extensively studied (Seno et al., 2009). Recently, the relationships between vection and cognition were examined. Vection and attention (Seno et al., 2011a), time perception (Seno et al., 20011b), cognitive bias (Palmisano & Chan,

2004), quantity perception (Seno et al., 2011c) and personality (Seno et al., 2011d) have been reported. In this article we introduce our three examples that vection alters cognitive states. Those examinations have not been conducted before. Our experiments were the first challenges of vection and multi dimensional human cognition.

## Time perception and vection

A number of factors have been known to modulate the subjective duration of the interval in time, e.g., attention to time passage (e.g. Zakay & Block, 1997), subjective event number in the stimulus presentation period (Poynter, 1989), whether the task is prospective or retrospective (e.g. Doob, 1971) and the boredom impatience and anticipation (Brown, 1985). Adding a dual task (in addition to the evaluation of the interval time) increases the errors and decreaseS the accuracy of evaluation (Brown, 1997 review). As a new factor of determining experienced time, we show that a return travel is perceived shorter in time than an outward travel. We succeeded in showing that vection strength modulate the shrinkage of the return travel. This was a very new finding.

## Method

We presented participants with virtual travel from Fukuoka, Japan, to a world-famous city, such as Paris, and examined subjective durations of stimulus movies, i.e. expanding-



optic-flow or dynamic-random-dot (DRD). We presented expanding flow during both the outward and return travels under the optic-flow condition. A round trip was assigned a cover story. Before the movie presentations, subjects understood that they would be asked to estimate the time durations of the stimulus presentation. We stated “Now we will go to Paris from Fukuoka” before the first stimulus presentation (outward trip). After the first stimulus presentation and before the second stimulus presentation, we stated “Now we will go back to Fukuoka from Paris” (return trip). Stimulus images were presented on a display with  $1,024 \times 768$  pixel resolution and at 75 Hz refresh rate. Each optic-flow display consisted of 16,000 randomly placed white dots on a black background. The dots were uniformly distributed within a simulated cube which subtended  $72$  (horizontal)  $\times$   $57$  (vertical) deg in visual angle. The stimuli were displayed on a 50-inch plasma display, with a viewing distance of 57 cm. The optic-flow simulated the forward or backward self-motion in constant-velocity (16 m/s). The DRD was refreshed at 75 Hz and the numbers of the dots were also 16,000. Twenty undergraduate students participated in each condition. The physical duration of stimulus presentations was 40 sec. Subjective durations were orally reported.

There was additionally the static-plane condition. A static random-dot plane was virtually placed 30 cm farther than the optic-flow plane (Figure 1). The farther dot plane effectively inhibited self-motion perception (e.g. Seno et al., 2010).

There were three stimulus conditions, optic flow, DRD and the static-plane conditions. The expected vection strengths were strong, medium and nothing for optic flow, static-plane and DRD conditions respectively. The destination of the trip was randomly chosen from ten very famous world cities (e.g. New York, Tokyo, etc.). The expected vection strengths for the three conditions were estimated from our previous study (Seno et al., 2010). Those subjective strengths were 60, 40, 0 for optic flow, Static plane, and DRD respectively (100 was very strong vection and 0 was no vection).

## Results and discussion

In the optic-flow condition, return travel was perceived as 5 sec shorter than the 40-sec physical presentation duration ( $Z=6.49$ ,  $p<0.01$ ). In the static-plane condition, the estimated duration in return travel was slightly longer than that; however, it was still shorter than 40 sec ( $Z=3.02$ ,  $p<0.05$ ). In the DRD condition, there was no shrinkage of the return travel.

We plotted the differences in subjective durations between outward and return-travel trials (Figure 2). Positive values indicated subjective time shortening in return travel for the round-trip conditions. Only in the optic flow and static-plane conditions, those values were significantly larger than 0 ( $z=11.97$ ,  $4.83$ , respectively,  $p<0.01$ ). That is, return travel was perceived as significantly shorter than outward travel in the two conditions.

Perceived time shrinkage was induced by perceiving vection. The degree of the shrinkage was correlated to the strength of vection. Miles et al. (2010) reported that, depending on the vection direction, daydreaming was oriented to the future or the past. Considered together with our results, vection seems to have some power to alter cognition.

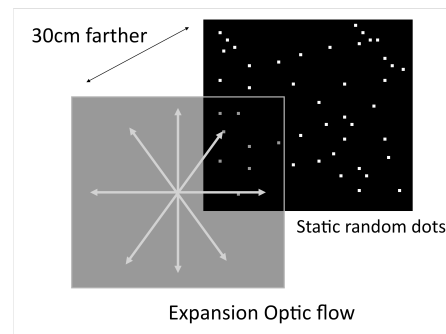


Figure 1. A schematic illustration of the farther-plane condition.

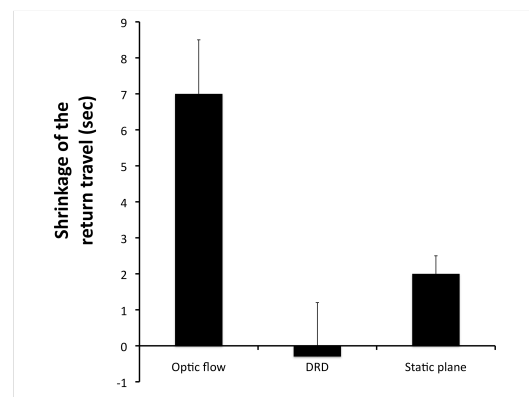


Figure 2. The shrinkage of the return travel. Error bars indicate SEM.

## Mental number line and vection

Previous studies have revealed a close connection between the representation of numbers and space (see Hubbard et al 2005 for a review); numerical magnitudes are represented in a mental number line (MNL) in an ascending order from left to right (e.g. Dehane et al 1993) and from bottom to top (e.g. Schwarz and Keus 2004).

The representation of MNL is tightly linked with our body motion. For example, Loetscher et al (2008) asked participants to generate random numbers orally whilst turning their head, and found that participants generated ‘small’ numbers more frequently when they turn their head to the left than to the right. A correlation between the magnitude of generated numbers and the direction of saccades has also been reported (Loetscher et al 2010): participants made rightward saccades just before they generated the larger numbers while they made leftward saccades when generated the smaller numbers.

Previous studies have only examined the MNL on the 2D plane. The goal of this study was to explore space-number interaction in the front-rear direction by using induced vection. Specifically, we asked participants to complete the random number generation task whilst they perceived the forward/backward body motion induced by expanding/contracting optic flows. This challenge was quite new and there has never been such study.

## Method

In each trial, whilst observing the optic-flow, the participants had to report orally four different numbers. Instruction was given as follows: “Please speak four different numbers in the interval between 0 to 100 as random as possible.” To make sure of the occurrence of vection, there was a 10 s interval between stimulus onset and the start of the task (for the delay of vection induction). Ten participants generated 8 numbers (4 x 2 trials) and the other ten participants generated 16 numbers (4 x 4 trials) for each direction of the optic-flow. The expanding and contracting trials were randomized. Twenty participants observed optic-flows in a dark chamber.

The perception of optic flow-induced vection was verified in a separate session after conducting the number generation session, when the magnitude and duration of vection with the same visual stimuli was measured.

## Results and discussion

The average of the generated numbers was significantly larger with the contracting motion than with the expanding motion (two-tailed paired t-test:  $t_{19} = 2.83$ ,  $p < .05$ ) (Figure 3). We found no bias such that observers generated numbers in an ascending or a descending sequence. The results showed that the sensation of self-motion could bias the magnitude of generated numbers, suggesting that, together with the previous results (Loetscher, et al. 2008, 2010), the representation of numerical magnitudes is tightly linked with our body motion. The present results suggest that the smaller numbers are represented in a front space while the larger numbers are represented in a rear space.

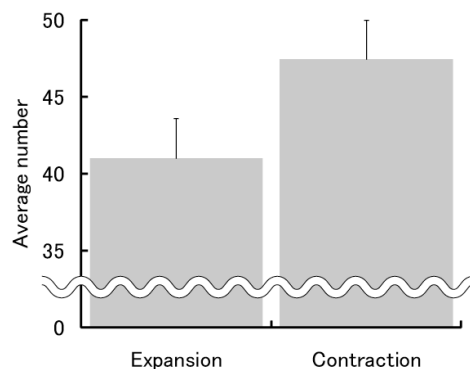


Figure 3. The average of generated numbers. Error bars indicate SEM.

## Personality

In our previous studies, we noticed that there were considerable individual differences between vection latency, duration and magnitude. Moreover, these individual differences appeared to be consistently exhibited, regardless of the stimulus conditions, i.e. an observer who perceived longer and stronger vection in one experiment also perceived longer and stronger vection in other vection experiments. These observations suggest that vection correlates with long-lasting characteristics of the participants. It has been reported that some cognitive abilities correlate with properties of personality (DeYong et al., 2005). For example, the performance of mental rotation (Ozer, 1987) or the magnitude of the attentional blink (McLean & Arnell, 2010) correlated with personality. A positive affectation influences performance in a variety of cognitive tasks (Ashby et al., 1999). The ‘Big Five’ factors of personality (openness, conscientiousness, extraversion, agreeableness, neuroticism) have also been found useful in predicting some everyday behaviors (Paunonen and Ashton, 2001). Therefore we thought that vection might also correlate with the personality traits of participants. We believe that the personality is one important aspect of human cognition. Thus this examination was one new challenge for vection and cognition.

## Method

We used an optic flow of expansion as vection stimulus. The duration of the stimulus was fixed at 40 seconds. Thirty adult volunteers participated in the experiment. Eight trials were conducted and the observers were asked to press a button when they perceived self-motion. At the end of each trial, the observers were also instructed to rate the subjective strength of vection using a magnitude estimation on a scale from 0 (no vection) to 100 (very strong vection).

After the vection task, the observers completed three questionnaires. The first included scales of public and private self-consciousness, which were Japanese versions of Fenigstein’s original index (Sugawara, 1984; Fenigstein et al., 1975). The second was the Narcissistic Personality Inventory Short Version (NPI-S; Oshio, 1999), based on the Narcissistic Personality Inventory developed by Raskin and Hall (1979). The third was related to the Big Five personality factor scales (Japanese version by Saito et al., 2001).

## Results and discussion

The results of the vection study were consistent with those found in our previous studies (the average vection latency, duration and magnitude were 12.25 seconds, 21.98 seconds, and 38.74, respectively), confirming the validity of the vection measures. Vection latency and total-NPI-S score were positively correlated ( $r^2 = 0.47$ ,  $p < 0.0001$ ), whereas duration or magnitude and total-NPI-S score were negatively correlated ( $r^2 = 0.28$ ,  $p < 0.003$ ;  $r^2 =$

0.19,  $p < 0.02$ , respectively) (Figure 4). No other comparisons between the three vection measures and personality scale scores showed any significant correlations.

We found that the more narcissistic the observer was, the weaker the perception of vection was.

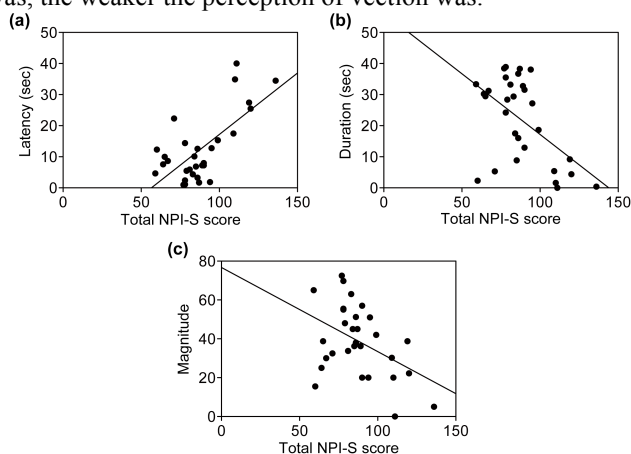


Figure 4. Correlations between vection measures (a: latency; b: duration; c: magnitude) and total-NPI-S score.

## Conclusion

We here showed three examples that vection alters our cognitive states. They were all new findings. Vection induced shrinkage of time of the return travel. Vection revealed our mental number line in depth. Vection and personality were correlated. Further examinations of the relationship between vection and cognitive states should be done in the future.

## Acknowledgments

The first author was aided by Japan Society for Promotion of Science.

## References

- Ashby FG., Isen Am., Turken AU. 1999 "A Neuropsychological Theory of Positive Affect and Its Influence on Cognition" *Psychological Review* **106** 529-550
- Brown, SW. 1985 "Time perception and attention: The effects of prospective versus retrospective paradigms and task demands on perceived duration." *Perception and Psychophysics* **38** 115-124
- Brown, SW. 1997 "Attentional resources in timing: Interference effects in concurrent temporal and nontemporal working memory task" *Perception and Psychophysics* **59** 1118-1140.
- Doob, LW. 1971 "Patterning of time" New Haven: Yale University Press.
- Fenigstein A, Seheier M E, Buss A H, 1975 "Public and private self-consciousness: Assessment and theory" *Journal of Consulting and Clinical Psychology* **43** 522-
- Fischer M H, Castel A D, Dodd M D, Pratt J, 2003 "Perceiving numbers causes spatial shifts of attention" *Nature Neuroscience* **6** 555-556
- Hubbard EM, Piazza M, Pinel P, Dehaene S, 2005 "Interactions between number and space in parietal cortex" *Nature Review Neuroscience* **6** 435-448
- Loetscher T, Schwarz U, Schubiger M, Brugger P, 2008 "Head turns bias the brain's internal random generator" *Current Biology* **18** R60-62
- Loetscher T, Bockish C J, Nicholls M E R, Brugger P, 2010 "Eye position predicts what number you have in mind" *Current Biology* **20** R264-265
- Miles, LK., Karpinska, K., Lumsden, J., Macrae, CN. 2010 "The meandering mind: Vection and mental time travel" *PLoS One* **5**:e10825
- Oshio A. 1999 "Narcissistic personality and friendship in high-school students" *The Japanese Journal of Personality* **8** 1-11 (in Japanese)
- Ozer, DJ. 1987 "Personality, Intelligence, and Spatial Visualization: Correlates of Mental Rotations Test Performance" *Journal of Personality and Social Psychology* **53** 129-134
- Paunonen SV., Ashton MC. "Big five factors and facets and the prediction of behavior" *Journal of Personality and Social Psychology* **81** 524-539
- Palmisano, S, Chan, AY. 2004 "Jitter and size effects on vection are immune to experimental instructions and demands" *Perception* **33** 987-1000
- Poynter, D. 1989 "Judging the duration of time intervals: A process of remembering segments of experience. Time and human cognition: A life-span perspective" *Advances in Psychology* **59** 305-33
- Raskin RN., Hall CS. 1979 "A narcissistic personality inventory" *Psychological Reports* **45** 590
- Zakay, D. Block, RA. 1997 "Temporal cognition" *Current Directions in Psychological Science* **6** 12-16.
- Dehaene S, Bossini S, Giraux P, 1993 "The mental representation of parity and numerical magnitude" *Journal of Experimental Psychology: General* **122** 371-396
- Saito, T., Nakamura, T., Endo, T., & Yokoyama, M. 2001 "Standardization of Big Five scales using the adjective check list" *Kyushu University Psychological Research* **2** 135-144 (in Japanese)
- Schwarz W, Keus I M, 2004 "Moving the eyes along the mental number line: comparing SNARC effects with saccadic and manual responses" *Perception & Psychophysics* **66** 651-664
- Seno, T., Ito, H. & Sunaga, S. 2009 "The object and background hypothesis for vection" *Vision Research* **49** 2973-2982
- Seno, T., Ito, H., Sunaga, S., 2010 "Vection aftereffect from expanding/contracting stimuli" *Seeing & Perceiving* **23** 273-294
- Seno T, Ito H, Sunaga S, 2011a "Attentional load inhibits vection" *Attention, Perception & Psychophysics*.
- Seno T., Ito H., Sunaga S. 2011b "Self-motion perception compresses time experienced in return travel" *Perception* **40** 497-499
- Seno T., Taya S., Ito H., Sunaga S. 2011c "Mental number line in depth revealed by vection" *Perception*, **40**, 1241-1244.
- Seno, T., Yamada, Y. & Ihaya K. 2011d "Narcissistic people cannot be moved easily by visual stimulation" *Perception*, **40**, 1390-192
- Sugawara K. 1984 "An attempt to construct the self-consciousness scale for Japanese" *Japanese Journal of Psychology* **55** 184-188 (in Japanese)

# Conscious and unconscious thought preceding complex decisions: The influence of taking notes and intelligence.

Aline Sevenants (aline.sevenants@ppw.kuleuven.be)

Dieter Daniëls (dieter.daniëls@student.kuleuven.be)

Leen Janssens (leen.janssens@ppw.kuleuven.be)

Walter Schaeken (walter.schaeken@ppw.kuleuven.be)

Department of Psychology, University of Leuven, Tiensestraat 102  
B-3000 Leuven, Belgium

## Abstract

For many years, research has been done to find the best way to make decisions. Dijksterhuis and Nordgren (2006) formulated the Unconscious Thought Theory (UTT), stating that when making complex decisions it is better not to think consciously, but to direct your attention elsewhere, letting the unconscious make the decision. However, a wealth of research has found evidence against the predictions of UTT. Thorsteinson and Withrow (2009) found that participants, who were allowed to take notes during the information intake stage, made better decisions thinking consciously. The current study is a replication of Thorsteinson and Withrow (2009), being a four conditions design (immediate decision, unconscious thought, conscious thought or conscious thought with notes) with the addition of intelligence as a variable. The conclusion of Thorsteinson and Withrow (2009) is supported: The best complex decisions are made when participants take notes and use them while thinking consciously. Moreover, it is shown that intelligence is positively correlated with better decisions.

## Introduction

When you buy a new house or car, you face a complex decision with many choice options that have different advantages and disadvantages. There are several ways to make this decision. You could try to list up all the different attributes of all the choice options, and think deeply about which option best suits your needs. Another strategy would be to make sure you are well informed about the different options, but not to decide immediately. After a good night of sleep, a gut feeling will arise, a preference for one of the options, even though you don't know where it came from. These are two completely different ways of making a complex decision, and throughout the years, there has been a lot of discussion about the intriguing question which of these strategies results in the best decisions.

For a long time, decision-making has been seen as a matter of rationality, objectivity and reflection. According to this view, a good decision can be made by breaking down the decision into small amounts of information, which have to be evaluated separately (e.g., Edwards, 1961; Dawes & Corrigan, 1974).

Later, this view has been challenged. Dijksterhuis, Bos, Nordgren, and van Baaren (2006) formulated the *deliberation-without-attention* hypothesis, stating there is a trade-off between the complexity of a decision and the usefulness of conscious thought when making the decision. To make an easy decision, it is better to think consciously, whereas unconscious thought should be used to solve more complex, broader decision problems. Unconsciousness is

believed to have an equal performance, no matter the difficulty level of the choice. Consciousness, in contrast, is especially good at making easy choices, even better than unconsciousness. But as decisions get more complex, consciousness has more problems with decision making, thereby performing worse than unconsciousness (Dijksterhuis et al., 2006). The deliberation-without-attention hypothesis is drawn from the Unconscious Thought Theory (UTT; Dijksterhuis & Nordgren, 2006), which explains the different characteristics of conscious and unconscious thought.

UTT and the deliberation-without-attention effect seemed to explain findings in earlier research (e.g., Wilson & Schooler, 1991; Wilson et al., 1993; Halberstadt & Levine, 1999). Research by other authors, however, led to conclusions that cannot be explained by UTT. In Experiment 1 of Thorsteinson and Withrow (2009), participants had to recall as many attributes as possible, before or after judging the different choice options. Only when judgement preceded recall, results provided evidence for the deliberation-without-attention hypothesis. The authors argued, however, that by recalling attributes, the participants in the unconscious-thought condition engaged in a form of conscious thought. It was also argued that the weighting principle of UTT is based on a weighted-additive model (WADD), but that a TALLY-model is used in research by Dijksterhuis (2004) and Dijksterhuis et al. (2006) to measure the quality of the choice (Newell et al., 2009). A weighted-additive model calculates the quality of a choice by the weight of every attribute, for example, if a cup holder is less important in a car than a good mileage, the cup holder should not get as much weight in the calculation of the quality of the cars. A TALLY-model calculates the options by simply adding the number of positive attributes. A cup holder thus has as much influence on the car's score as a good mileage.

In order to clarify the contradicting findings in the literature, a meta-analysis was conducted by Acker (2008), which showed a large heterogeneity between different studies. It revealed a small but unconvincing advantage in favour of unconscious thought. Since different studies led to other conclusions, more research is needed to clarify under which conditions unconscious thought can be useful.

Previous research showed that unconscious thought does not necessarily perform better under certain circumstances. One specific condition under which unconscious thought seems to lose its advantages is when participants are not obligated to rely on their memory when

making a decision. Some research already exists on this topic, but the current study will try to further elaborate some missing parts. Of special interest for the current study, is Experiment 2 of Thorsteinson and Withrow (2009). In this experiment, participants were given the opportunity to overcome their memory limitations, thereby changing the outcomes of the deliberation-without-attention effect. Laboratory studies by Dijksterhuis and colleagues (Dijksterhuis, 2004; Dijksterhuis et al., 2006) follow the same paradigm. Participants are given information about different choice options (e.g., cars, apartments or roommates). After having consciously read all the attributes of the different choice options, they have to make a decision immediately (immediate-condition), after a few minutes of conscious thought (conscious-thought condition) or after a few minutes of distraction (unconscious-thought condition). Thorsteinson and Withrow (2009) included a fourth condition in this paradigm, the conscious-thought-with-notes condition. Participants in this condition were allowed to take notes of the attributes during the presentation period, and use these notes while thinking consciously about their decision. Memory limitations would not influence the quality of the choices of these participants. Participants in the conscious-thought-with-notes condition turned out to make better decisions than participants in the other conditions. The mean unconscious score did not differ significantly from the mean score in the conscious-thought condition without notes. This study thus was unable to replicate the findings of Dijksterhuis et al. (2006), but showed that, when overcoming memory limitations, conscious thought seems to be beneficial for complex decisions. Newell et al. (2009) also included a conscious-thought condition with information in their second experiment. In this condition, participants were provided with an information sheet containing all the attributes, while thinking consciously. Participants in this condition made better choices, but the difference was only significant compared with the immediate condition, not with the unconscious-thought condition. Rey et al. (2009) found that participants in their conscious-thought condition, who also had access to the information while thinking consciously, performed worse. The effect of using a memory aid does not seem clear yet. Too little research has been done to date to draw a clear conclusion. The current study aims to provide a valuable addition on this topic.

Also important is the method Thorsteinson and Withrow (2009) used to measure the quality of the choice. Participants were asked to rate the attributes on importance for them. The description of the four choice options, in this study: non-existing cars (materials from Dijksterhuis et al., 2006), is formulated in bipolar attributes, for example, a car does or does not have a cup holder. Therefore, a score for each car can be calculated for each participant by multiplying the importance score of an attribute with either minus one (if the attribute is negative for this car) or one (if the attribute is positive for this car). As a dependent measure, Thorsteinson and Withrow (2009) calculated the difference between the score of the car with the highest score and the score of the chosen car. A participant thus had a score of zero if the chosen car was the best possible car

for this participant, and higher than zero if he/she chose another car. This way, a WADD-model was used to measure the quality of a choice. The current research will include both a TALLY and two WADD calculations (a dichotomous WADD measurement and a continuous variant), in order to find the model that represents best human decision-making.

Despite the diversity of phenomena related to IQ, few have attempted to understand – or even describe – its influences on judgment and decision making (Frederick, 2005). Studies on time preference, risk preference, probability weighting, ambiguity aversion, endowment effects, anchoring, and other widely researched topics rarely make any reference to the possible effects of cognitive abilities (or cognitive *traits*). The majority of studies approach the deliberation-without-attention effect as a universal effect, being applicable to all participants. However, it could be possible that one mode of thought would be more beneficial for some participants, but not necessarily for others. Therefore, individual differences in intelligence will be studied in the current research. The short version of Raven's Advanced Progressive Matrices (Bors & Stokes, 1998) will be used to measure intelligence. To our knowing, no research on the influence of intelligence on the deliberation-without-attention effect has been done to date. However, it might influence the decision quality. A higher intelligence level could be an advantage when processing the information. Furthermore, when using conscious thought to make a decision, the quality of the analysis of the information could be influenced by intelligence.

## Experiment

### Method

#### Design and Participants

A total of 341 participants were taken over different educational levels. Participants were recruited from university students and students in secondary education. The sample thus contained participants engaging in academic education ( $n = 213$ ), general secondary education ( $n = 37$ ), technical secondary education ( $n = 41$ ), and vocational secondary education ( $n = 24$ ). University students participated in exchange for course credits, secondary educational students did so voluntarily. Ages of the participants ranged from 16 to 30, of which 99.1% was 24 or younger. Participants were randomly assigned to one of the four conditions (immediate, unconscious-thought, conscious-thought, and conscious-thought-with-notes condition).

#### Procedure and Material

As in most research on UTT, the paradigm of Dijksterhuis (2004) was used. All participants were presented 48 attributes from four different, nonexistent cars (12 attributes per car). These stimulus materials were taken from Dijksterhuis et al. (2006), the same as used in Thorsteinson and Withrows (2009) second experiment. Each attribute was formulated either positive or negative. All cars were described on the same 12 features. The Hatsdun had 75% positive attributes, whereas the Nabusi was characterized by only 25% positive features. Two more

neutral cars were included, with the Kaiwa having 58% and the Dasuka 50% positive attributes.

Because participants were also recruited under the national legal age for a driving licence, it could possibly be hard for them to imagine buying a car. To make sure that these younger participants would perceive the materials the way intended by the researchers, a pre-test was conducted with a small sample of secondary education students, none of whom participated in the real experiment ( $n = 13$ ). The materials used by Thorsteinson and Withrow (2009) in their first experiment, which contained attributes from apartments, were tested at the same time, to find out which materials suited best the needs of participants their age. Half the sample were given the attributes in a positive formulation (e.g., "The car has good mileage."), and were asked to rate the importance of the attributes. The other half received the same attributes, but negatively formulated (e.g., "The car has poor mileage.") and were asked to rate how bad they felt about the car not having the feature. A mean score for each choice option was calculated, by multiplying the mean score of each attribute by either minus one (if the choice option did not have the feature) or one (if the choice option did have the feature), and summing these. The results showed that, for both sets of materials, features that were intended to be positive, were perceived as positive, and those intended to be negative, were perceived as negative. The choice options were ranked in the same order as intended. Since both sets were suitable, the car-materials were used to make comparison with Thorsteinson and Withrows (2009) second experiment easier.

In the real experiment, participants in the conscious-thought-with-notes condition were instructed that they were allowed to write down whatever they wanted during the presentation of the information, whereas participants in the other conditions were just instructed to pay attention to the attributes presented. They all knew they would have to make a decision on which car they preferred, resulting in impression forming (Lassiter et al., 2009). The 48 attributes were presented in a random order, for eight seconds each, using E-prime software (Psychological Software Tools, Pittsburgh, PA).

After the presentation of the information, participants in the conscious-thought conditions had four minutes to think carefully about the different cars. After four minutes, they had to choose the car they preferred. Then they were asked to rate the 12 different attributes on importance, on a scale from one, meaning "no importance at all", to seven, meaning "very important". This made it possible to calculate the subjective preferences of each participant. In the unconscious-thought condition, participants were distracted for four minutes by solving anagrams, before they had to make their decision and rate the attributes. In the immediate condition, participants had to choose their preferred car immediately after the presentation of the information, followed by the rating of the attributes, and were then instructed to solve the anagrams as a filler task in order to obtain the same the experiment duration in all conditions.

After these tasks, all participants were asked to fill in the short version of Raven's Advanced Progressive Matrices (Bors & Stokes, 1998). During the last decade, this has been one of the most widely used instruments by researchers interested in participants' inductive or analytic reasoning capacities or fluid intelligence (Cattell, 1963). Raven's Advanced Progressive Matrices (Raven, Court, & Raven, 1988) or APM is a version of these matrices intended for use with people above average aptitude and designed to reliably differentiate among those in the top 25% of the population (Bors, & Stokes, 1998). In the present study a short version of the APM (Bors, & Stokes, 1998) has been used. It is a selection of 14 items with increasing difficulty drawn from Set 2 of the original APM (item 3, 10, 12, 15, 16, 18, 21, 22, 28, 30, 31 and 34). For all 14 items, participants had to indicate which was the missing segment required to complete a 3x3 matrix, which takes about 20 minutes. All parts of the experiment were presented in Dutch.

## Results

First, the proportions of participants choosing the best car according to a TALLY-model, were compared between the different conditions, following Dijksterhuis (e.g., Experiment 2 in Dijksterhuis, 2004). As shown in Figure 1, participants in the conscious-thought-with-notes condition performed best, with a proportion of 0.578 choosing the Hatsdun, which had the most positive features. Participants in the conscious-thought condition, who did not take notes, performed only slightly worse (.54). Participants in the unconscious-thought condition made the worse choices, with a proportion of only .32 choosing the Hatsdun. Those who couldn't engage in any form of thought, in the immediate condition, scored in between the conscious and unconscious-thought conditions, with a proportion of .48 choosing the objectively best car. An ANOVA showed that the main effect was statistically significant:  $F_{(3, 337)} = 4.64$ ,  $p < .01$ ,  $MSE = .24$ . Tukey contrasts showed that the differences between both conscious-thought conditions and the unconscious-thought condition were significant, with  $p < .01$  for the conscious-thought-with-notes condition and  $p < .01$  for the conscious-thought condition.

Since UTT is based on a WADD-model, and the proportion of participants choosing the Hatsdun is a TALLY measurement, the best weighted subjective choice options were calculated. For each participant, a unique rating for each car was calculated according to a WADD-model, by summing the importance ratings of each attribute, with the ratings of attributes that the particular car did not have, being counted negative.

It turned out that the Hatsdun was, subjectively, not the best car for each participant. For 23.2% of the participants, another car than the Hatsdun had a higher or equal subjective score, so not choosing the Hatsdun was not necessarily a bad decision for them. Therefore, the proportions of participants choosing their subjectively highest rated car were compared between the four conditions, making it a dichotomous WADD measure (see Figure 2).

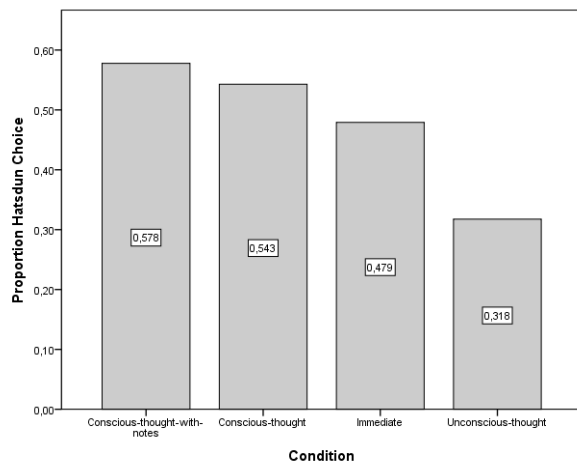


Figure 1. Proportion of participants choosing the best car according to a TALLY-model (the Hatsdun) under different conditions.

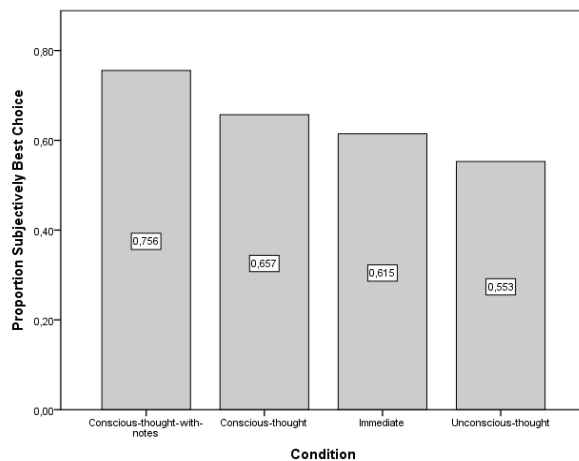


Figure 2. Proportion of participants choosing their best option according to a dichotomous WADD-model under different conditions.

In all conditions, the proportions of participants choosing their subjectively best car were higher than the proportions of those choosing the Hatsdun, indicating that participants in all conditions followed a weighted decision model. The pattern of differences between the conditions changed, but the ranking remained the same. Unconscious thinkers scored relatively better with this weighted measurement, with a proportion of 0.553 choosing their subjectively best car. However they still performed worse than participants in all other conditions. Participants in the immediate condition chose their subjectively best choice option with a proportion of .62, in the conscious-thought condition with a proportion of .66 and in the conscious-thought-with-notes condition with a proportion of .76. An ANOVA revealed that this effect was significant,  $F_{(3, 337)} = 2.83$ ,  $p < .01$ ,  $MSE = .23$ . Post-hoc Tukey contrast showed that participants in the conscious-thought-with-notes condition made significantly better decisions than those in the unconscious-thought condition,  $p < .01$ .

It could be argued that not choosing the car that suits one's subjective needs best, does not mean the choice was necessarily a bad one, since for some participants the subjective scores of some choice options didn't differ

much. For example, one participant had chosen the Kaiwa, with a subjective score of 22, but the best option for this person would have been the Hatsdun, with a score of 24. This choice was not as bad as the choice of another participant who had also chosen the Kaiwa, with a subjective score of -6 as opposed to his best option, the Hatsdun, with a subjective score of 18. Therefore, to measure the real quality of a participant's choice, the continuous WADD calculation of Thorsteinson and Withrow (2009) was used. For each participant, the subjective score of the chosen car was subtracted from score of the car with the highest subjective score. Participants that chose the car that suited them best thus had a difference score of zero. The higher the difference score, the worse the choice made by the participant. This continuous WADD measurement seemed to fit participants' decision patterns best, because 14.2% of those who did not choose their best option according to the dichotomous WADD-model, had a difference score of only two. With difference scores ranging up to 68 in the total population, these choices were not necessarily bad ones. When using the difference score as the dependent variable, the pattern remained the same but the relative difference between the unconscious-thought and the conscious-thought-with-notes condition increased. Participants engaging in unconscious thought had a mean difference score of 6.87, thereby performing only slightly worse than participants in the immediate (6.60) and conscious-thought condition (5.10), but remarkably worse than those in the conscious-thought-with-notes condition, with a mean difference score of only 2.467 (see Figure 3). An ANOVA showed that this effect was significant,  $F_{(3, 327)} = 2.99$ ,  $p < .01$ ,  $MSE = 122.29$ . Post-hoc Tukey contrasts indicated that the difference between the conscious-thought-with-notes condition and the unconscious-thought condition was significant,  $p < .01$ . The mean difference score of the conscious-thought-with-notes condition was also marginally significantly lower than the mean score of the immediate condition,  $p < .05$ .

Finally, intelligence seems to influence decision quality. In the total population<sup>1</sup>, a positive correlation<sup>2</sup> was found between the scores on the short version of Raven's Advanced Progressive Matrices and the proportion of participants making the best decision according to a dichotomous WADD model (.17). A negative correlation with the difference scores (-.10) also provided evidence that a higher intelligence level leads to better decisions. Within the different conditions, no correlations with the difference scores were found, but the dichotomous WADD scores correlated positively with the intelligence measure in the immediate (.21) and unconscious-thought conditions (.27). In the immediate condition, a correlation with the TALLY measurement (.20) was found.

<sup>1</sup> To make sure the different levels of intelligence were equally spread over the different conditions, t-tests were conducted. No significant differences were found.

<sup>2</sup> All reported correlations are significant at the .05 level except for the correlation with the dichotomous WADD model in the total population, which is significant at the .01 level.



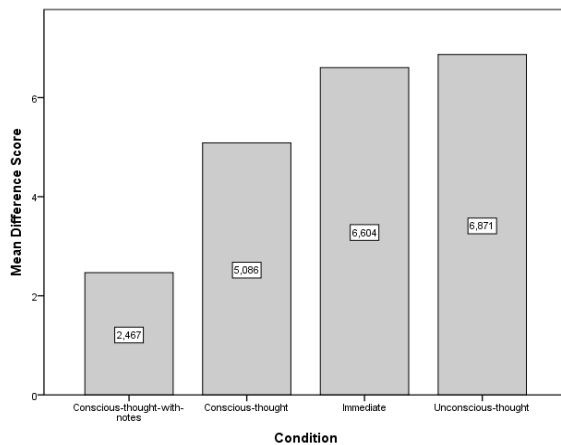


Figure 3. Means of difference scores (continuous WADD-model) under different conditions.

## Discussion

The results of this experiment did not support the predictions of UTT. Using three different methods to measure decision quality, including the measurement used by Dijksterhuis (2004), no evidence for a beneficial effect of unconscious thought was found. In contrary, participants in the unconscious-thought condition performed worst of all participants, significantly worse than participants in the conscious-thought-with-notes condition. It thus seems that overcoming memory limitations is enough to make conscious thought a better decision strategy. But even without notes, participants engaging in conscious thought outperformed those engaging in unconscious thought. Even though this difference was only significant when a TALLY-model was used to measure the decision quality, the results of the other measurement methods still show the opposite pattern as predicted by UTT. The results of Thorsteinson and Withrow (2009) were confirmed.

Another important finding is that participants seemed to follow a WADD-model to make their decisions. In all conditions participants chose the car that was best for them, as calculated with a weighted additive model, more often than the car that was best according to a TALLY-model. A continuous WADD-model measured decision quality even better. This finding adds up to the findings of other research in support of using a WADD-model to measure decision quality (Newell et al., 2009; Thorsteinson & Withrow, 2009). Also from a theoretical point of view, a WADD-model should be used, since it reflects the weighting principle of UTT better. However, even when using this weighted measure, conscious thought seems to outperform unconscious thought when notes can be taken. These participants had an advantage since they were able to structure the information and select what to write down, writing only down what they consider important for their decision. These advantages made weighting easier. Under these circumstances, it thus seems that conscious thought is beneficial.

Individual differences also seemed to influence decision quality, but not under all circumstances. For the

overall population, intelligence seemed beneficial for the decision quality, but in the conscious thought conditions, no correlation could be found. This might be due to the overall better decisions made in these conditions.

## References

- Acker, F. (2008). New findings on unconscious versus conscious thought in decision making: Additional empirical data and meta-analysis. *Judgment and Decision Making*, 3(4), 292-303.
- Bors, D. A., & Stokes, T. L. (1998). Raven's advanced progressive matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement*, 58, 382-398.
- Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81(2), 95-106.
- Dijksterhuis, A. (2004). Think different: The merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology*, 87(5), 586-598.
- Dijksterhuis, A., Bos, M. W., Nordgren, L. F., & van Baaren, R. B. (2006). On making the right choice: The deliberation-without-attention effect. *Science*, 311, 1005-1007.
- Dijksterhuis, A., & Meurs, T. (2006). Where creativity resides: The generative power of unconscious thought. *Consciousness and Cognition*, 15, 135-146.
- Dijksterhuis, A., & Nordgren, L. F. (2006). A theory of unconscious thought. *Perspectives on Psychological Science*, 1(2), 95-109.
- Dijksterhuis, A., & van Olden, Z. (2006). On the benefits of thinking unconsciously: Unconscious thought can increase post-choice satisfaction. *Journal of Experimental Social Psychology*, 42, 627-631.
- Edwards, W. (1961). Behavioral decision theory. *Annual Review of Psychology*, 12, 473-498.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19, 25-42.
- Halberstadt, J. B., & Levine, G. M. (1999). Effects of reasons analysis on the accuracy of predicting basketball games. *Journal of Applied Social Psychology*, 29(3), 517-530.
- Newell, B. R., Wong, K. Y., Cheung, J. C. H., & Rakow, T. (2009). Think, blink or sleep on it? The impact of modes of thought on complex decision making. *The Quarterly Journal of Experimental Psychology*, 62(4), 707-732.
- Rey, A., Goldstein, R. M., Perruchet, P. (2009). Does unconscious thought improve complex decision making? *Psychological Research*, 73, 372-379.
- Thorsteinson, T. J., & Withrow, S. (2009). Does unconscious thought outperform conscious thought on complex decisions? A further examination. *Judgment and Decision Making*, 4(3), 235-247.
- Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, 60(2), 181-192.

# When Choice Effects Compete: An Account by Extended EBA Model

Kenpei SHIINA (shiina@waseda.jp)

Department of Educational Psychology, Waseda University, Tokyo, Japan

## Abstract

A modified version of EBA is proposed to account for choice set effects (similarity, attraction, and compromise) and their interactions. The new model has ingredients of search-for-dominance-structure theory and counter race model, highlighting conflict resolution and deliberation in decision making. Simulation results show that the model can reproduce the choice set effects and predict the interactions between them.

**Keywords:** EBA; Choice Set Effects; Dominance Structuring ; Race Model.

## The EBA Model Revisited

The elimination by aspects model (EBA, Tversky, 1972a, b) is a classical and well-known model, which remains attractive because it was one of early models that provided a clear processing assumption for choices with mathematical rigor. EBA model asserts that an aspect is probabilistically selected and options that do not have this aspect are discarded. (In this paper, attribute, aspect, and feature are used interchangeably). Another aspect is then selected and options that do not have the second aspect are discarded. Proceeding in this way, the EBA process terminates when only one option remains. Although EBA is a simple model that does capture an important facet of decision making, there are many empirical effects and theoretical notions that the model cannot explain. This paper describes an updated version of EBA model that can account for choice set effects (e.g., similarity, attraction, and compromise effects) and their interactions.

The paper consists of four parts: The first section briefly reviews properties of EBA in view of theories and findings after its proposal in 1972. In the second section, a modified EBA called REGAL model is proposed (Shiina, 1994, for an earlier version): the model incorporates the ideas of dominance structuring and race model. In the third section, a simulation study is presented that shows how the new model explains choice set effects and their interactions by producing probability topographies. Finally, implications for further research on choice and decision making are addressed.

## Original EBA and its Properties

We use the following symbols and notations (Figure 1a).

$T = \{X, Y, Z, \dots\}$ : The finite total set of choice alternatives.

$\Omega = \{\alpha, \beta, \gamma, \rho, \dots\}$ : The finite total set of features (aspects).

$x', y', z', \dots \subseteq \Omega$ : Subsets of  $\Omega$  representing  $X, Y, Z, \dots$

For example, if  $X$  has features,  $\alpha, \rho, \omega$  and  $\theta$ , then

$x' = \{\alpha, \rho, \omega, \theta\}$ . It is assumed that  $\Omega = \bigcup_{x' \in T} x'$ .

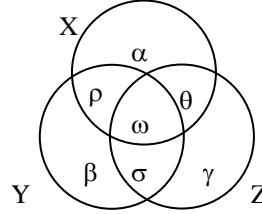


Figure 1a

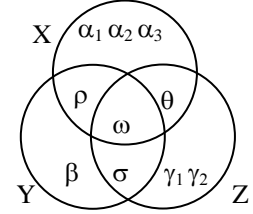


Figure 1b

## Common Aspects

Aspects that are common to all the alternatives ( $\omega$  in Figure 1a) should be ignored in EBA whereas there is evidence that common aspects play an important role both in decisions and choice satisfaction (Chernev, 1997). It seems very unnatural, moreover, that we should intentionally ignore the important features, common or not, that come to mind. If we consider a version of EBA that permits common aspect selection (Common aspect EBA or CEBA), it can be proved that CEBA does not change choice probabilities and can produce decision time predictions. Suppose that  $T = \{X, Y, Z\}$ ,  $x' = \{\alpha, \rho, \theta, \omega\}$ ,  $y' = \{\beta, \rho, \sigma, \omega\}$ , and  $z' = \{\gamma, \sigma, \theta, \omega\}$  (Figure 1a) with the understanding that  $\alpha = u(\alpha)$  etc., where  $u()$  is a utility (value) function. In binary case :  $T = \{X, Y\}$ , we have

$$P(X \leftarrow \{X, Y\}) = (\alpha + \theta) / (\alpha + \theta + \beta + \sigma)$$

$$P(X \leftarrow \{X, Y\}) = (\alpha + \theta) / L + (\rho + \omega) / L \times P(X \leftarrow \{X, Y\})^{CEBA},$$

$$\text{where } L = \alpha + \theta + \beta + \sigma + \rho + \omega$$

Solving the second expression for  $P(x \leftarrow \{x, y\})^{CEBA}$ , we have

$$P(X \leftarrow \{X, Y\})^{CEBA} = \frac{(\alpha + \theta) / L}{1 - (\rho + \omega) / L} = \frac{(\alpha + \theta) / L}{\{L - (\rho + \omega)\} / L}$$

$$= (\alpha + \theta) / (\alpha + \theta + \beta + \sigma) = P(X \leftarrow \{X, Y\})^{EBA}$$

This equivalence also holds true for trinary or more numerous choices and thus Tversky's exclusion of common features is well grounded. For example, in trinary case :

$$T = \{X, Y, Z\},$$

$$P(X \leftarrow \{X, Y, Z\})^{CEBA} = P(X \leftarrow T)^{CEBA} =$$

$$1 / K \times \{\alpha + \rho P(X \leftarrow \{X, Y\}) + \theta P(X \leftarrow \{X, Z\}) + \omega P(X \leftarrow T)^{CEBA}\}$$

$$= \frac{\alpha}{K} + \frac{\rho}{K} \left\{ \frac{\alpha + \theta}{\alpha + \theta + \beta + \sigma} \right\} + \frac{\theta}{K} \left\{ \frac{\alpha + \rho}{\alpha + \rho + \gamma + \sigma} \right\} + \frac{\omega}{K} P(X \leftarrow T)^{CEBA}$$

where  $K = \alpha + \beta + \gamma + \rho + \sigma + \theta + \omega$ . This expression is identical to EBA if common feature  $\omega$  is ignored. Setting

$$L = \alpha + \rho(\alpha + \theta) / (\alpha + \theta + \beta + \sigma) + \theta(\alpha + \rho) / (\alpha + \rho + \gamma + \sigma)$$

we can derive a simple recursive formula:

$$P(X \leftarrow T)^{CEBA} = L / K + \omega / K \cdot P(X \leftarrow T)^{CEBA}$$

$$= L/K + \omega/K \left[ L/K + \omega/K \cdot P(X \leftarrow T)^{CEBA} \right] = \dots$$

$$= \frac{L}{K} + \frac{L}{K} \frac{\omega}{K} + \frac{L}{K} \left( \frac{\omega}{K} \right)^2 + \frac{L}{K} \left( \frac{\omega}{K} \right)^3 + \dots + \left( \frac{\omega}{K} \right)^m P(X \leftarrow T)^{CEBA}$$

Therefore when  $m \rightarrow \infty$ , we can derive an expression that is identical to the trinary EBA, because

$$P(X \leftarrow T)^{CEBA} = \frac{L}{K} \frac{1}{1 - \omega/K} = \frac{L}{K - \omega}$$

$$= \frac{\alpha + \rho(\alpha + \theta) / (\alpha + \theta + \beta + \sigma) + \theta(\alpha + \rho) / (\alpha + \rho + \gamma + \sigma)}{\alpha + \beta + \gamma + \rho + \sigma + \theta} = P(X \leftarrow T)^{EBA}$$

The generalization to  $n$ -option situation is straightforward, which gives reason for the idea that common features are processed but are indifferent to the final choice, and how the CEAB interpretation produces decision time predictions, because this interpretation yields a geometric distribution which permits rough estimates of decision times.

### Conjunction of Features and Dominance

It appears that EBA uses attributes one at a time. A careful reading of the original paper reveals, however, that an aspect can be an aggregate or conjunction of aspects. The aspect  $\alpha$  in Figure 1a, for example, may itself be a set of sub-aspects  $\{\alpha_1, \alpha_2, \alpha_3\}$  as in Figure 1b. In Figure 1, the choice of  $\alpha$  automatically leads to the choice of X. This paper interprets this property as “X was chosen because X dominated other options on  $\alpha$ .” If  $\alpha = \{\alpha_1, \alpha_2, \alpha_3\}$ , the interpretation will be “X was chosen because X dominated the other options on  $\{\alpha_1, \alpha_2, \alpha_3\}$ ”.

Dominance or dominance structuring is another key concept in the present paper. According to Montgomery & Willen (1999, p.148), “the decision maker attempts to find a dominance structure, that is, a cognitive structure in which the to-be-chosen alternative dominates other alternatives on relevant attributes.” It seems natural and promising to interpret EBA within the framework of search for dominance structuring (SDS) theory (Montgomery, 1989; Montgomery & Willen, 1999).

Because SDS model asserts that dominance should be established on a *bundle* of relevant attributes, a to-be-chosen alternative should be dominant on the conjunction of relevant attributes. A comparison of EBA and SDS in this respect gives a novel perspective; EBA is a very limited SDS in the sense that EBA process is a type of dominance structuring based upon a single attribute or a bundle of unique attributes. The key ideas linking EBA and SDS models are that dominance structuring is performed on a *conjunction* of attributes and the conjunctive set is *sampled* from the total set of attributes and thus changes over time. The conjunction of attributes is called an *evaluation set* or an *aspect lineup* in the model to be presented. An evaluation set is a set of attributes, so it is totally different from the reconsideration set, which is a set of alternatives.

### Reconsideration, Deliberation, and Deferral

Decision deferral and deliberation are two sides of the same coin because, in both cases, the decision maker cannot resolve decisional conflict at hand and thus deliberation evolves over time. Whereas it is often said that there are no widely accepted definition of conflict (Tversky & Shafir,

1992), we adopt the simple definition of Coombs and Avrunin (1988, p.222) that conflict arises in “a situation in which a choice must be made in the absence of dominance”.

It is often argued that EBA is suitable for relatively easy everyday choices, but may lead to less than optimal decisions. Janis and Mann (1977, p.32) pointed out, for example, that the decision maker may run out of relevant aspects and/or alternatives before reaching a decision, or may end up taking an alternative that is inferior to those eliminated.

If a pure EBA processing is used, it is almost inevitable that we may arrive at a much inferior choice with some probability. In that case, it will be very hard to justify both the final outcome and the choice process leaving a strong feeling of regret. We may start over, reconsidering discarded attributes and alternatives, whereas the original EBA does not have the mechanism to allow for such reprocessing. It is obvious that we should deliberate when the decision problem at hand is very important (buying expensive items, choosing a spouse or a job, etc.). Deliberation is a time-developing process and can continue for months or even years, during which time we think about the choice repeatedly. In this regard, an EBA model reinforced by SDS seems very appealing, because deliberation and justification are the key concepts of the SDS theory. Further, the combination is consistent with the current consensus that preferences are actively constructed, not merely revealed (Bettman, Luce, & Payne, 1998).

### Conflict Resolution and Counter Model

Tversky himself later maintained that decision making is a type of conflict resolution and that justification for the decision is necessary (Simonson, 1989; Simonson & Tversky, 1992; Tversky & Shafir, 1992; Shafir, Simonson, & Tversky, 1993). Dominance structuring is one method of conflict resolution and, if successful, it makes the choice self-evident (Montgomery & Willen, 1999, p.148). Although there are easy decisions that can be made almost automatically, decision makers facing an important decision should repeat consideration. Therefore, a counter or race model (Smith & Van Zandt, 2000) may be hypothesized in which the confidence for each alternative is accumulated during reconsideration, as the decision problem is addressed from many perspectives over time. This reconsideration process, possibly with repeated generation of evaluation set by attribute sampling, would be necessary to make the decision satisfactory or satisficing.

### REGAL Model

The REGAL (REpeated-Generation of Aspect Lineup) model integrates concepts from EBA, SDS, and the race (counter) models and tries to extend EBA in the following ways. First, reconsideration of aspects and alternatives is allowed, as this is an undeniable aspect of real-world decision-making. Second, REGAL permits the generation of conjunctive aspects, because a decision maker often has a bundle of minimum but unstable requirements that can be represented by a fluctuating conjunction of aspects. Third,

REGAL does not ignore aspects that are common to the alternatives, because there is evidence that common aspects do affect choice by enhancing the satisfaction. Finally, the concepts of counter and criterion (threshold) are adopted from race model to allow for decision time predictions.

In short, our revised EBA incorporates a) *reconsideration and deliberation processes*, b) a more *flexible aspect selection* that permits conjunctions, c) the processing of *common features*, and d) decision time generation processes.

### The Flow of the REGAL Model

The processing flow of the model is as follows (see also Figure 2 and Appendix):

(1) Let  $T$  be the set of alternatives and  $\Omega$  be the set of attributes used to represent the alternatives. The set  $x' \subseteq \Omega$  is the feature representation of Alternative  $X$ .

(2) A decision maker probabilistically samples a set of attributes  $\Psi$  from  $\Omega$ , called an *evaluation set* or an *aspect lineup*. The evaluation set is repeatedly regenerated during each cycle of REGAL process as shown schematically by the loop from (6) back to (2) in Figure 2. The reconsideration processes are represented in the loop.

(3) The degree of satisfaction for each alternative is evaluated on the *current evaluation set* determined in (2). The degree of satisfaction  $0 \leq S_\Psi(x) \leq 1$  is defined as

$S_\Psi(x) = \text{Goodness of Option } X \times \text{Structural dominance of } X \text{ over the other alternatives on evaluation set } \Psi$ .

The “Goodness” part says when both the evaluation set and  $X$  have many features and thus are rich then the function tends to output a larger satisfaction value. The “Dominance Part” outputs how dominant  $X$  is over the other alternatives in the attribute structure induced by  $\Psi$ .

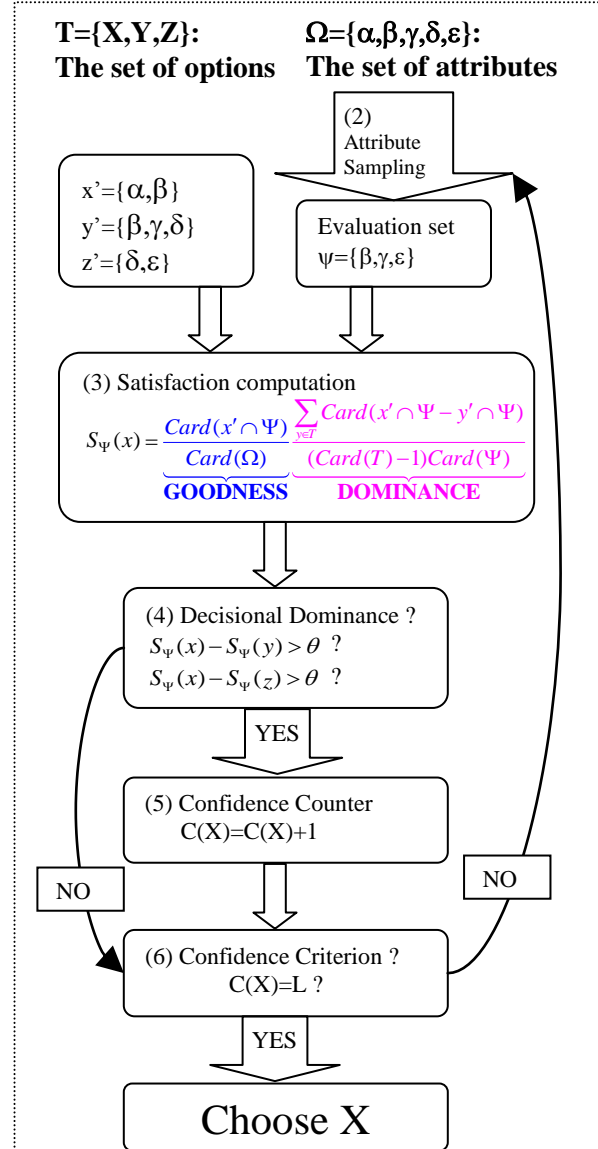
(4) A *decisional criterion*  $\theta$  is assumed. If  $S_\Psi(x) - S_\Psi(y) > \theta$  for all  $Y \in T - \{X\}$  then  $X$  is in the state of *decisional dominance*. If no option is dominant the process jumps to (6). Decisional dominance means that an option dominates the others in satisfaction at the moment.

(5) For each alternative, there is a *confidence counter*  $C(X)$ . If  $X$  becomes decisional dominant in (4), then a unit increment is added to  $C(X)$ . Race model is incorporated around (5) and (6).

(6) If  $C(X)$  reaches a *confidence criterion*  $L$ , then  $X$  is chosen else the process resumes from (2), where the regeneration of evaluation set  $\Psi$  is performed. The confidence criterion or threshold  $L$  represents decision quality: If the decision is momentous,  $L$  will be a high value, whereas if the decision is less important,  $L$  will be smaller. The depth of reconsideration processes are partially moderated by  $L$  and thus a merit of REGAL is that it encompasses the notion of decision quality.

When only one attribute is very important in decision, EBA, SDS, and REGAL tend to produce the same final choice.

**Counter Model** The loop between (6) and (2) constitutes a counter model (Smith & Van Zandt, 2000). The exact probability of choice is given in Appendix (Section 5).



**Figure 2: Outline of REGAL Model**

**Relation to Previous Models** Based on Manski (1977), several researchers have proposed conjunction models to make consideration sets (Andrews & Manrai, 1998; Gilbride & Allenby, 2004, 2006). *Subset Conjunction* proposed by Jedidi and Kohli (2005) is similar to the *evaluation set* in this paper. REGAL does not make consideration sets explicitly, but when  $y' \cap \Psi = \emptyset$  alternative  $Y$  is virtually excluded from the set of choice options and the remaining alternatives form a temporal consideration set. The major difference between REGAL and these previous models is that the consideration set is always temporary in REGAL.

### Simulation: Choice Set Effects

This section demonstrates how REGAL explains choice set effects (Roe, Busemeyer, & Townsend, 2001) and their interactions. Choice set effects have been studied on 2-dimensional continuous attribute spaces (Figure 3). Suppose

there are two options : A and B. Choice set effects (Similarity, Attraction, and Compromise effects) occur as a function of the location of new option C.

Both EBA and REGAL prefer discrete features and thus a special arrangement is needed to deal with continuous dimensions (See, Gensch & Ghose, 1992 for another approach). Basically, the continuous dimension is divided into segments and each segment is reinterpreted as a feature. For example, let  $A'=\{\alpha, \gamma, \delta\}$  and  $B'=\{\alpha, \beta, \gamma\}$  before the placement of C. If we place a new option  $C_1$ , the definition of segments and thus the features are changed accordingly (Figure 3) :  $A'=\{\alpha_1, \alpha_2, \gamma, \delta_1, \delta_2\}$ ,  $B'=\{\alpha_1, \alpha_2, \beta, \gamma\}$ , and  $C'=\{\alpha_1, \gamma, \delta_1\}$ , in this particular case.

**The Probability of Choosing a Feature** Let  $P(\tau), \tau \in \Omega$  be the probability that feature  $\tau$  is included in  $\Psi$ .  $P(\tau)$  is a function of feature importance or salience. In the present simulation, the length of feature would represent the salience, so that it is simply assumed :

$$P(\tau) = 1 - \exp(\text{length}(\tau) \times K_1) \quad K_1: \text{constant}$$

which is sub-additive (Rottenstreich & Tversky, 1997) in the sense that  $P(\alpha) \leq P(\alpha_1) + P(\alpha_2)$  when  $\alpha = \{\alpha_1, \alpha_2\}$ .

**Value Function and Satisfaction Function** The value (utility) of feature  $\tau$  is again assumed to be a function of the length of feature :

$$v(\tau) = K_2 \times \log(\text{length}(\tau)) \quad K_2: \text{constant.}$$

Using this function, the original binary version of satisfaction function (Equation (A.2) in Appendix) is converted into a continuous dimensional version:

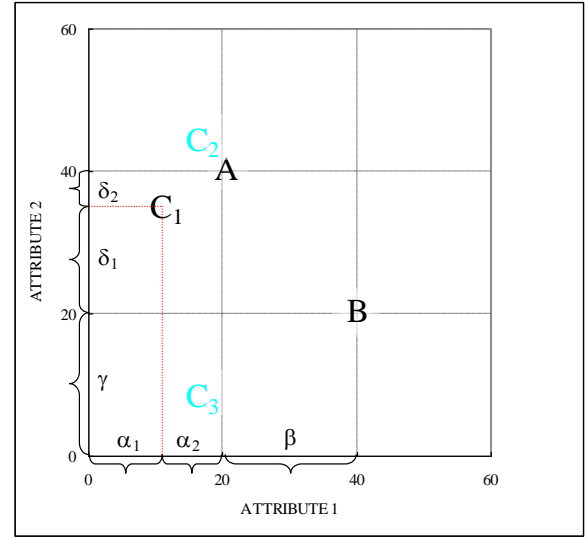
$$S_\Psi(x) \equiv \frac{\sum_{\tau \in x \cap \Psi} v(\tau)}{\underbrace{\text{Card}(\Omega)}_{\text{GOODNESS}}} \frac{\sum_{\tau \in x \cap \Psi - y \cap \Psi} v(\tau)}{(\underbrace{\text{Card}(T) - 1}_{\text{DOMINANCE}}) \sum_{\tau \in \Psi} v(\tau)}$$

**Interactions of Choice Set Effects** Suppose that new Option C is placed at  $C_1$ . Attraction effect predicts  $P(A)$  will be larger while Similarity effect predicts  $P(A)$  will be smaller. If C is place at  $C_2$ , Similarity effect predicts  $P(A)$  smaller while Compromise effect predicts  $P(A)$  will be larger. If C is placed at  $C_3$ ,  $P(A)$  may become larger because C is dominated by both A and B but more strongly dominated by A than B, and there may be a slight similarity effect between C and B as well.

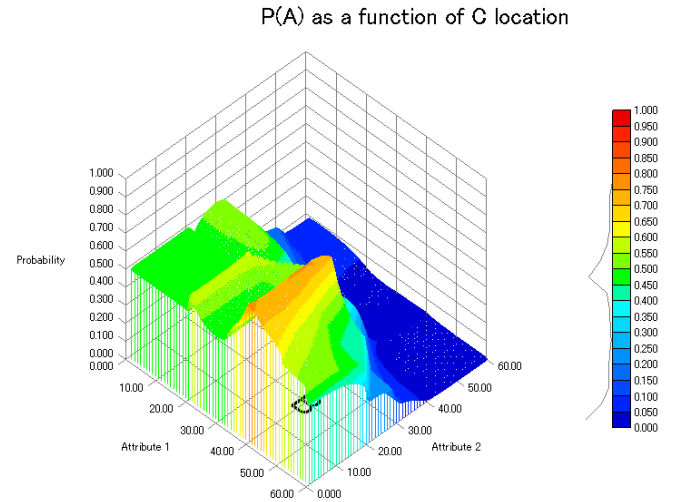
Apparently, in these cases choice set effects compete or collaborate and the choice probabilities should be determined as a non-linear function of *forces* that try to push up or down probabilities. In choice effect studies, *pure* conditions in which the effect of a single *force* becomes observable have been used. A next challenge should be to clarify the joint effects of choice set effects and the present study is the first such attempt.

## RESULTS

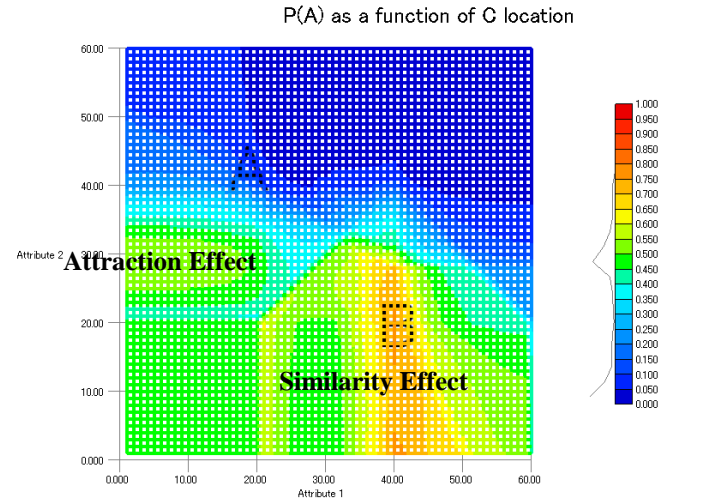
By moving Option C on the space (Figure 3), we can observe how the probabilities for Options A, B, and C change. REGAL can probe the joint effects by generating *probability topographies* (Figures 4, 5, and 6) calculated from Equation (A.3) in Appendix. Several detailed values for parameters are also shown in Appendix.



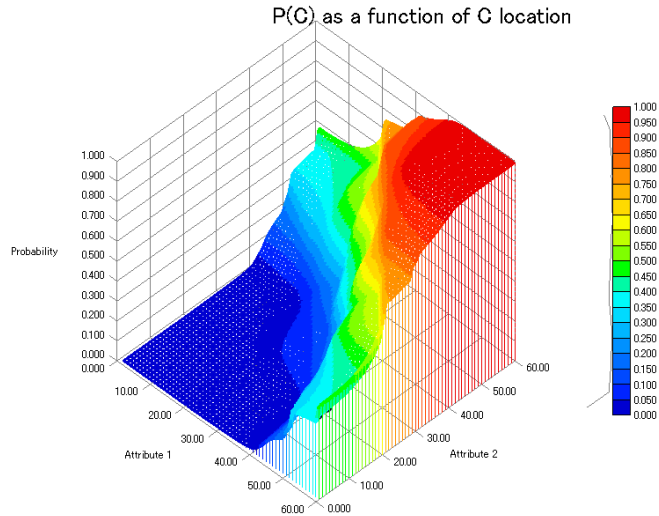
**Figure 3: Option representation on 2-dimensional attribute space.**



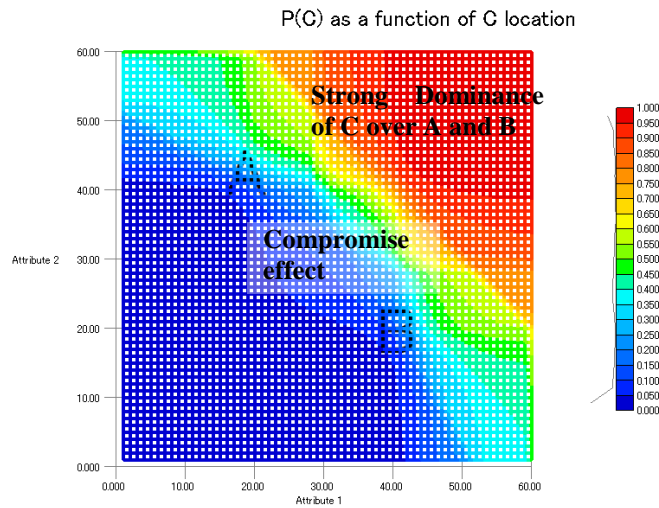
**Figure 4a: P(A) as a function of C location**



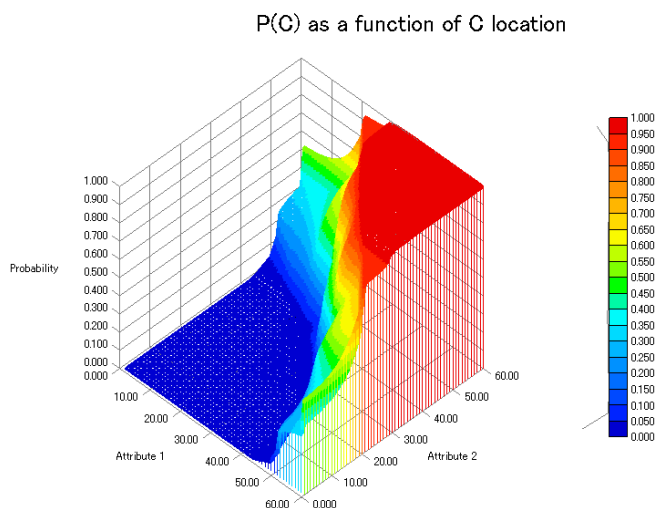
**Figure 4b: P(A) as a function of C location**



**Figure 5a:  $P(C)$  as a function of  $C$  location**



**Figure 5b:  $P(C)$  as a function of  $C$  location**



**Figure 6:  $P(C)$  as a function of  $C$  location when  $L$  is large ( $L=20$ ).**

Figures 4a (bird's eye view) and 4b (contour map) show the same topographic probability information. The probabilities are theoretical predictions in an ideal space

The reading of Figure 4a and 4b is a little confusing: they are showing  $P(A)$ , the probability that Option A will be chosen, *as a function of  $C$  location*. Therefore, the right lower high region, for example, should be read that “if you place  $C$  around here,  $P(A)$  will be high.” Figures 5a and 5b show  $P(C)$  as a function of  $C$  location. In this case, the figures give a natural interpretation. The graph for  $P(B)$  is omitted because it is an exact mirror image of Figure 4.

Major observations are as follows. 1) Strikingly high  $P(C)$  is obtained when  $C$  strongly dominates  $A$  and  $B$  (Figure 5b, upper right) and very low  $P(C)$  when  $C$  is dominated by both  $A$  and  $B$ . 2) Similarity effect is stronger than attraction effect (Figure 4b) and compromise effect is very weak (Figure 5b). Attraction effect was weak possibly because attraction and similarity effects compete in Figure 4b. 3) Increasing  $L$  made large probabilities larger and small probabilities smaller. As a result, the slope in Figure 6 became steeper. Psychologically, this would mean that deeper deliberation changes probabilistic choice into something akin to logical judgment, decreasing the chance of taking inferior options. Of course, these observations depend upon the present configuration of options, definitions of features, the shape of the value function, and the parameter values. More intensive search is needed to validate the model. Due to space limitation, RT predictions will be presented elsewhere.

## Discussion

It is well-known that MDFT model (Roe, et.al, 2001) and LCA model (Usher & McClelland, 2001) are able to mimic the three choice set effects. The present model is distinct both in architecture and in processing assumptions and can deal with, at least in theory, any number of options and attributes. Further, it can produce predictions for choice probabilities and decision times. This paper showed only qualitative validity of REGAL in an ideal theoretical space and empirical tests will be necessary in the future study.

## References

- Andrews, R.L., & Manrai, A.K. (1998). Feature-based elimination: model and empirical comparison. *European Journal of Operational Research*, **111**, 248-267.
- Bettman, J.R., Luce, M.F., & Payne, J.W. (1998). Constructive consumer choice processes. *Journal of Consumer Research*, **25**, 187-217.
- Chernev, A. (1997). The effect of common features on brand choice: Moderating role of attribute importance. *Journal of Consumer Research*, **23**, 304-311.
- Coombs, C.H., & Avrunin, G.S. (1988). *The Structure of Conflict*. Hillsdale, NJ: Erlbaum.
- Gensch, D.H., & Ghose, S. (1992). Elimination by dimensions. *Journal of Marketing Research*, **29**, 417-429.
- Gilbride, T.J., & Allenby, G.M. (2004). A choice model



- with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science*, **23**, 391-406.
- Gilbride, T.J., & Allenby, G.M. (2006). Estimating heterogeneous EBA and economic screening rule choice models. *Marketing Science*, **25**, 494-509.
- Janis, I.L., & Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice and commitment*. Free Press.
- Jedidi, K., & Kohli, R. (2005). Probabilistic subset-conjunctive models for heterogeneous consumers. *Journal of Marketing Research*, **42**, 483-495.
- Manrai, A.K., & Sinha, P.K. (1989). Elimination-by-cutoffs. *Marketing Science*, **8**, 133-152.
- Montski, C.F. (1977). The structure of random utility models. *Theory and Decision*, **8**, 229-254.
- Montgomery, H. (1989). From cognition to action: The search for dominance in decision making. In H. Montgomery & O. Svenson (Eds.), *Process and structure in human decision making* (pp. 23-49). New York: Wiley.
- Montgomery, H., & Willen, H. (1999). Decision making and action: The search for a good Structure. In P. Juslin & H. Montgomery (Eds.), *Judgment and Decision Making. Neo-Brunswikian and Process-tracing Approaches* (pp. 147-173). Mahwah, NJ: Erlbaum.
- Roe, R., Busemeyer, J.R., & Townsend, J.T. (2001). Multi-alternative decision field theory: A dynamic connectionist model of decision-making. *Psychological Review*, **108**, 370-392.
- Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review*, **104**, 406-415.
- Shafir, E.B., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition*, **49**, 11-36.
- Shiina, K. (1994). Hesitation, indecision, and status quo: A Generalization of the EBA model. *Japanese Psychological Review*, **37**, 250-264. (In Japanese)
- Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research*, **16**, 158-174.
- Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, **29**, 281-295.
- Smith, P.L., & Van Zandt, T. (2000). Time-dependent Poisson counter models of response latency in simple judgment. *British Journal of Mathematical and statistical Psychology*, **53**, 293-315.
- Tversky, A. (1972a). Elimination by aspects: A theory of choice. *Psychological Review*, **79**, 281-299.
- Tversky, A. (1972b). Choice by elimination. *Journal of Mathematical Psychology*, **9**, 341-367.
- Tversky, A., & Shafir, E.B. (1992). Choice under conflict: The dynamics of deferred decision. *Psychological Science*, **3**, 358-361.
- Usher, M., & McClelland, J.L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, **108**, 550-592.

## Appendix: Technical Details of REGAL

This part should be read with reference to Figure 2.

### (1) Initial setting

$T$ : the set of alternatives,  $\Omega$ : the set of attributes

### (2) Evaluation set construction

$\Psi$ : Evaluation feature set

$P(\tau)$ : The probability that feature  $\tau$  is included in  $\Psi$ .  $P(\tau)$  is an increasing function of feature importance or salience.

The probability of generating evaluation set  $\Psi$  is

$$P(\Psi) \equiv P(\Psi \leftarrow 2^\Omega - \phi) = \frac{\prod_{\tau \in \Psi} P(\tau) \prod_{\tau \notin \Psi} (1 - P(\tau))}{1 - \prod_{\tau \in \Omega} (1 - P(\tau))} \quad (A.1)$$

Important features tend to stay in  $\Psi$  and common features are not forcefully excluded.

(3) **Satisfaction function**  $S_\Psi(x)$  measures the degree of satisfaction and is a function of  $\Psi$  and  $X$ .

$$S_\Psi(x) \equiv \frac{\underbrace{\text{Card}(x' \cap \Psi)}_{\text{GOODNESS}}}{\underbrace{\text{Card}(\Omega)}_{\text{GOODNESS}}} \frac{\sum_{y \in T} \text{Card}(x' \cap \Psi - y' \cap \Psi)}{(\text{Card}(T) - 1) \text{Card}(\Psi)} \quad (A.2)$$

$\text{Card}()$  is the cardinality of a set. *Goodness* part takes a value in  $[0,1]$ : the value of 1 is obtained when  $x' = \Psi = \Omega$ , that is, alternative  $X$  is perfectly dominating the other alternative in the sense it has all relevant attributes, and the value 0 is obtained when  $x' \cap \Psi = \phi$ , that is,  $X$  has no relevant attributes under the current evaluation set. *Dominance* part represents the structural dominance of  $X$  over the other alternatives. This part also takes a value in  $[0,1]$ : the value of 1 is obtained when  $x' = \Psi$  and  $y' \cap \Psi = \phi$  for all  $Y$  except  $X$ , that is, alternative  $X$  is perfectly dominating the other alternative with respect to the current evaluation set, and the value of 0 is obtained when  $x' \cap \Psi - y' \cap \Psi = \phi$ , that is,  $X$  is totally dominated by the other alternatives under the current evaluation set.

(4) **Probability of decisional dominance** The probability that Alternative  $X$  becomes dominant,  $M_X$ , is defined as

$$M_X \equiv \sum_{\Psi \in 2^\Omega - \phi} P(X \text{ is decisional dominant} | \Psi) P(\Psi) \\ = \sum_{\Psi \in 2^\Omega - \phi} P(S_\Psi(X) - S_\Psi(Y) > \theta, \forall Y \in T - X | \Psi) P(\Psi)$$

where  $\theta \sim N(\mu, \sigma^2)$ : Decisional criterion that may fluctuate.

In the text, the variance is set to 0 and thus  $\theta$  is a constant.

(5) **Confidence counter  $C()$** . The loop between (6) and (2) in Figure 2 can be captured by a race model and an alternative that first reaches  $L$  is chosen, where  $L$  is a confidence criterion. From the standard result of Poisson counter model (Smith and Van Zandt, 2000), the final choice probability is given in closed-form by:

$$P(X) = \sum_{j_1=0}^{L-1} \sum_{j_2=0}^{L-1} \cdots \sum_{j_{X-1}=0}^{L-1} \cdots \sum_{j_{n-1}=0}^{L-1} \frac{1}{\prod_{k=1}^n j_k!} \prod_{k=1}^n \left( \frac{\lambda_k}{\sum_{i=1}^n \lambda_i} \right)^{j_k} \left( \frac{\lambda_X}{\sum_{i=1}^n \lambda_i} \right)^{j_n} \quad (A.3)$$

where  $n$  is the number of attributes and  $\lambda_k$ 's are Poisson strength parameters. Without loss of generality, we can set  $\lambda_k = M_k$ . The free parameters are  $L, \mu, \sigma^2$  and  $P(\tau), \tau \in \Omega$ . By adjusting them, we can examine whether REGAL can mimic the three choice-set effects. In the simulation,  $L=5$ ,  $\mu=0$ , and  $\sigma^2=0$  were used.



# Creative Process of Improvised Street Dance

Daichi Shimizu (tothefuture0415@yahoo.co.jp)

Graduate School of Education, University of Tokyo  
Tokyo 113-0033, Japan

Takeshi Okada (okadatak@p.u-tokyo.ac.jp)

Graduate School of Education, &  
Interfaculty Initiative in Information Studies, University of Tokyo  
Tokyo 113-0033, Japan

## Abstract

This paper presents findings from our empirical study of the creative process of improvisation, which has rarely been the subject of research in cognitive science. In this study, battle scenes in street dance were selected as an example of improvised performances. We conducted an experiment to investigate real-time cognitive processes. The results indicated three features: 1) Dancers mainly used well-practiced patterns, and discovered new patterns of dance; 2) In the process of discovering new patterns, dancers often utilized errors in their performance; 3) The processes of discovery were different in the performance of one dancer (solo scene) and the performance of two dancers (battle scene). In solo performance, dancers discovered new patterns by concentrating on their patterned dance. In battle performance, dancers discovered new patterns by utilizing stimuli from the situation (e.g. the music, their opponent) and using errors as an opportunity to loosen the constraints of their well-practiced patterns.

**Keywords:** Improvisation; Street dance; Personal discovery; Utilizing errors; Battle scenes

## Introduction

Improvisations such as jazz or improvisational dance or drama are complicated human activities which seldom become research subjects in cognitive science. However, improvisations are thought to be the origin of many activities in the arts (see Bailey, 1980). The moment when a person gains new ideas is considered to be related to improvisational patterns (Pressing, 1984). Based on these suggestions, improvisations are thought to be a core element in human creativity.

## Features of improvisational activities

Most previous studies dealing with improvisation have investigated jazz music (e.g., Mendonça & Wallace, 2004; Tayanagi, 2010; Weisberg et al., 2004). Mendonça & Wallace (2004) investigated the duo performance of jazz musicians, and suggested that musicians use some fixed patterns in improvisation. They also suggested that a musician utilizes the music of the other musician as a guideline for his own musical performance.

Weisberg et al. (2004) examined records of the improvisations of professional jazz musicians, and

suggested that they often utilized specific formulas (50-90% of each performance) in their performances.

Tayanagi (2010) investigated the literature and the biographies of professional jazz musicians theoretically, and claimed that accepting inevitable errors in performance and utilizing these errors is very important for innovation and the production of new patterns in jazz music. This suggestion is consistent with the claim of Pressing (1984).

Bailey (1980) investigated the cognitive process of improvisation by interviewing professional musicians in many genres of music. Based on anecdotal evidence, he suggested that there are differences in music between improvisations by one person and improvisations by multiple persons.

From these suggestions, we could summarize the features of improvisations as follows. 1) Performers use fixed-patterns. 2) Performers utilize the errors which are inevitably generated to make new patterns. 3) The number of person participating in the improvisation makes some difference.

## Personal discoveries of new ideas in improvisations

In this study, we investigated the process of improvisational activities, paying special attention to “personal discoveries”. “Personal discovery” is defined as the discovery of new ideas, expressions or techniques occurring in creative activities, which the performers (creators) did not envisage prior to these activities. This concept mostly focused on the cognitive process of creators, and from this point of view, personal discovery is strongly related to Psychological Creativity (creativity which results in processes or products that are new and useful to the creators themselves), as Boden (1991) suggested. The personal discovery in dance is a movement which may not be new in a general sense, but is new to the dancer performing it. Many researchers have suggested that unpredicted findings like personal discoveries play important roles in creating new products or making scientific discoveries (e.g., Dunbar, 1993; Suwa & Tversky, 1997). In the case of improvisations, personal discoveries also play important roles when creating new products, expressions or techniques which performers did not envisage in advance (Bailey, 1980). In this sense, improvisation involves personal discovery as its core.

## Breakdance as an improvisational activity

This study deals with the battle scenes of breakdance (a major genre in street dance) as an example of improvisation. Breakdance first appeared in Manhattan in the late 1970s, and has spread widely around the world. This dance consists of four patterns: entry (dance in a standing position), footwork (dance performed on the floor), power moves (dance with acrobatic movements like rolling), and freeze (dance poses held in acrobatic positions) (OHJI, 2001). In the battle scenes, dancers stand facing one another and perform their improvisational dance for 30-40 seconds in turns. Dancers in break dance have to perform while listening to unfamiliar music, communicating with an opponent, and responding to the dance of the opponent. Hence, the battle scenes of breakdance are highly improvisational. Therefore, it is appropriate to use battle scenes as the object of research into improvisational activities.

### Purposes of this study

This study investigates the cognitive processes of dancers in battle scenes of breakdance, which are considered to be an example of an improvisational activity. Specifically, we focus on three questions based on the findings of previous studies: 1) How often are fixed patterns of dance used? In previous studies, it has been hypothesized that fixed patterns are used in improvisation more than 50% of the time. 2) Do dancers utilize the errors which are inevitably generated in improvisational dance to find new patterns of dance movements? If so, how do they utilize these errors? 3) Are the improvisational activities of a solo dancer different from the improvisational activities of multiple dancers? In order to answer these questions, we conducted an experiment with dancers.

## Methods

### Participants

Fourteen semi-expert dancers participated in this experiment (mean age 24.5 ( $SD=3.8$ ), mean experience 6.5 years ( $SD=4.4$ )). The level of skill of the dancers was evaluated from the following two aspects: the acquisition and use of the basic skills required for the four patterns of dance, and the acquisition and use of advanced skills relating to power moves (highly skilled movements). The evaluation was conducted by the first author using a videotape recorded during the experiment. As a result of this evaluation, we found that all the subjects had advanced levels of skill in addition to the basic skills of breakdance and were able to perform various patterns of movements in breakdance.

### Procedure

In this experiment we set two conditions, the solo scene condition and the battle scene condition. The only difference between the solo scene and the battle scene was that there was no opponent (dance partner) and so no dance by an opponent in the solo scene. This solo condition was

set to investigate question 3, relating to differences arising from the number of persons. Both scenes used the same music (Bomb the Bass, "Megablast"). The experiment was conducted in one room of the gymnasium of the university. The room size was 14.4 x 14.5 meters. The performances were recorded on video.

This experiment consisted of three different sessions: 1) Preparative session (explanation of experimental procedure and warm up); 2) Performance session; 3) Post-performance session (dancers' reflections on their own movements and thoughts during the dance performances).

- 1) We explained to the participants the outline of this experiment, i.e., the design of the experiment (two independent variables, solo scene and battle scene), and the resting time between the two sections. Then, we told the subjects to take about 30 minutes to warm up.
- 2) Each dancer performed the solo or the battle scene. For each scene, the dancers performed for about 30-40 seconds and then took a 30-40 seconds interval (in the solo scene, they just waited without dancing, and in battle scene, they watched the opponent dancing). They repeated this set three times. Music was continually playing during each scene. Just before the performance, we explained to the dancers the details of each scene (three sets of dances and intervals), and asked them to perform naturally as in a usual battle scene. For the solo scene, we instructed the dancers to perform as if it were a battle scene, pretending there was an opponent.
- 3) We asked the dancers to reflect on their dance performances and report their thoughts during the dance (Figure 2). First, the dancers watched videos of their dance performances, and they segmented their dance movements into meaningful units. Then the dancers evaluated each dance segment using a creativity score (novelty and dexterity), and reported what they were thinking while dancing each segment.

We conducted these three sessions for one scene (solo or battle), took a break of about an hour, then repeated sessions 2 and 3 for the other scene. The order of the scenes was counterbalanced.

### Outline of analyses

In this study, we analyzed the processes of improvisation with three sets of data: 1) Creativity score of dancers (self-evaluation); 2) Self-report of cognitive process by dancers (report of thoughts); 3) Categorization of dance movements based on the usage of the four types of movements (categorization of dance movements).

1) We used the data from the creativity scores of dancers. Through the use of these data, we aimed at investigating the features of dance movements from the dancers' own viewpoints. The objects of the creativity score (novelty<sup>1</sup> and

<sup>1</sup> This consists of 3 rating scores: Dance 1) is well practiced; 2) is not well practiced, but has been performed before; 3) has never been performed. We used these scores because in the preliminary interviews with other dancers, the dancers told us that to judge the

Table 1. Mean number and percentage of dance movements corresponding to each novelty score (sum of three trials)

<i>Scene</i>	<i>Score 1 (well-practiced)</i>	<i>Score 2 (not well-practiced, but has been performed before)</i>	<i>Score 3 (has never been performed)</i>
Solo	8.2(4.64) 66%	3.1(2.88) 25%	1.1(1.41) 9%
Battle	8.1(4.95) 66%	3.0(1.66) 25%	1.1(1.29) 9%

Table 2. Mean number and percentage of dance movements corresponding to each dexterity score (sum of three trials)

<i>Scene</i>	<i>Score 1 (very poor)</i>	<i>Score 2 (poor)</i>	<i>Score 3 (moderate)</i>	<i>Score 4 (good)</i>	<i>Score 5 (very good)</i>
Solo	1.64(1.60) 13%	3.36(1.22) 27%	4.64(3.05) 37%	2.29(2.27) 18%	0.57(0.94) 5%
Battle	2.00(2.18) 16%	3.50(2.03) 29%	4.29(2.92) 35%	2.00(2.25) 16%	0.43(0.85) 3%

dexterity<sup>2</sup>) were based on previous studies of creativity (e.g., Finke et al., 1992).

We summed up the data of the creativity scores and conducted statistical analyses. In addition, we identified dance movements with high creativity scores (2 or 3 for novelty and 4 or 5 for dexterity), and analyzed the data. These high-scoring dance movements reflect the dancers' personal discoveries, because they reported the new and useful movements that they had "discovered". By analyzing them, we were able to investigate the features of "personal discoveries" in each scene.

2) We used the answers to the question, "What were you thinking while you were dancing these particular movements?" in the report on cognition. In the analyses, we categorized the focus of consideration of the dancers while dancing and classified each statement according to the category. By analyzing what the dancers were giving their consideration, we were able to investigate the points about which the dancers thought deeply in each scene. In addition, we identified the statements about high-scoring dance movements which were thought to reflect their personal discovery, and analyzed them using these categories. By analyzing them, we were able to investigate the focus of consideration of the dancers when generating new patterns.

3) We used the data from the performances of the dancers (dance movements in performance sessions), and categorized them into the four types of breakdance. By comparing the number of movements of each type between the solo scene and battle scene, we were able to investigate the nature of dance movements in each scene objectively.

## Results and Discussion

Before analyzing the details of the data, we compared the basic features of both scenes (the time of performance, the number of dance segments). We conducted a paired *t*-test on

these data and found that there was no statistical difference between the two scenes (solo: 95.4 (23.26)<sup>3</sup> seconds, battle: 86.3 (16.52) seconds,  $t(13) = 1.68$ ,  $p = .12$ ) (solo: 12.5 (5.07), battle: 12.2 (4.93),  $t(13) = 0.75$ ,  $p = .75$ )<sup>4</sup>

### Creativity score of dancers

#### Novelty score (Table 1)

Using a sign test, we conducted a contrast analysis of each novelty score (score 1 - score 3) in the solo scene and the battle scene. As a result, we found that there were no differences between the two scenes (the *p*-values of scores 1, 2, 3 were  $p = .79$ ,  $p = .58$ ,  $p = 1.00$ ). Then we examined the number and percentage of each novelty score in each scene to determine which scores frequently appear. As shown in Table 1, there were high degrees of appearance of score 1 in both scenes. The percentages of score 1 are 66% in mean rate in both scenes. These results show that dancers mainly use well-practiced, somewhat patterned dance movements in improvisational activities.

#### Dexterity score (Table 2)

Using a sign test, we conducted a contrast analysis of each dexterity score (score 1 - score 5) in the solo scene and the battle scene. As a result, we found that there were no differences between the two scenes (the *p*-values of scores 1, 2, 3, 4, 5 are  $p = .79$ ,  $p = .58$ ,  $p = 1.00$ ,  $p = 1.00$ ,  $p = .63$ ). Then we

<sup>3</sup> In this study, we used the mean score which sums up the three trials in the performance session in each scene.

<sup>4</sup> These data have high degrees of *SD* and are thought to be out of Gaussian distribution. We conducted a sign test to eliminate the influence of individual differences. Analyses show the same results as the *t*-test (time of performance:  $p = .12$ , number of dance movements:  $p = .75$ ). The reason why high degrees of *SD* appear seems to be as follows. Each dancer performs a trial using a subjective time scale acquired through his/her dance experience. Each one may have a different subjective time span. The sizes of chunks of dance that dancers think of as a dance unit may differ individually. Based on this supposition, we employed a statistical test that utilizes the comparison of each individual (like a sign test).

practice level, a 3-point rating was much more suitable than a 5-point rating.

<sup>2</sup> This consists of five rating scores: Dance is: 1) very poor; 2) poor; 3) moderate; 4) good; 5) very good).

Table 3. Definition of the categories and mean number (sum of three trials)

<i>Higher category</i>	<i>Lower category</i>	<i>Definition</i>	<i>Solo</i>	<i>Battle</i>
A: Consideration of their own dance	a: Well-practiced dance movements	Dancers consider well-practiced dance	3.1 (2.57) 27%	2.4 (2.03) 21%
	b: New patterns	Dancers give consideration to new patterns of dance	0.6 (0.93) 4%	0.5 (0.76) 4%
	c: Dance composition	Dancers give consideration to the composition of their dance	1.6 (1.09) 10%	0.8 (0.70) 6%
B: Consideration of information about the situation	d: Music	Dancers give consideration to the music	2.5 (2.14) 19%	2.6 (2.34) 23%
	e: Opponent (partner)	Dancers give consideration to their partner	0.2 (0.43) 2%	1.7 (1.20) 14%
	f: Physical position	Dancers give consideration to their physical position	1.4 (1.28) 10%	2.1 (1.46) 16%
C: Consideration of other factors	g: No specific consideration	Dancers give no consideration to anything specific	2.3 (1.98) 21%	1.5 (2.35) 11%
	h: Other factors	Dancers give consideration to other factors	0.9 (1.10) 7%	0.5 (0.65) 5%

examined the number and percentage of each dexterity score in each scene. The results showed that scores 2 and 3 frequently appeared in both scenes. Hence, we are able to suggest that dancers mainly use dance movements which show similar dexterity to well-practiced dance movements.

#### **Dance movements corresponding to personal discoveries**

We identified high creativity scoring dance movements (2 or 3 for novelty and 4 or 5 for dexterity) and examined their rates of appearance in both scenes.

The results show that there are 14 high-scoring dance movements (8% of all the dance movements) in the solo scene, and 17 high-scoring dance movements (10% of all the dance movements) in the battle scene. Even in the short-time performances (80-100 sec.) in this experiment, dancers found new patterns and made personal discoveries. The result that there are high-rated uses of patterned dance movements indicates that dancers in improvisation mainly use patterned dance movements and gradually find new patterns through improvisation. To compare the rate of appearance between each scene, we conducted a sign test and found that there was no statistical difference ( $p=.51$ ).

#### **Consideration of dancers in performances**

##### **Analyses of statements about all dance movements (Table 3)**

The  $\kappa$  coefficient was calculated by the first author and a researcher who did not know the purpose of this study, using about 20% of all the data, 70 dance movements, to check the reliability of the rating. The  $\kappa$  coefficient was 74.1%, which guarantees the reliability of the ratings. Using a sign test, we conducted a contrast analysis to compare the number of each category in the solo and the battle scenes. The results show that there were statistical differences in c)

Consideration of dance composition, e) Consideration of the opponent, f) Consideration of the dancer's own physical position ( $p=.039$ ,  $p=.003$ ,  $p=.065$ ). In the solo scene, dancers often think about the composition of whole dance movements. In contrast, in the battle scene, the dancers consider information about the situation (opponent, physical position).

We also compared the numbers and percentages in each category in each scene to determine frequently appearing categories. As shown in Table 3, in the solo scene, a) Consideration of well-practiced dance movements, d) Consideration of the music, g) No specific consideration, and in the battle scene, a) Consideration of well-practiced dance movements, d) Consideration of the music, e) Consideration of opponent, f) Consideration of the physical position appeared more frequently than other categories. Thus, we conclude that in the solo scene, dancers think about well-practiced dance movements or the music, and construct performances considering the whole composition of their dance movements. In contrast, in the battle scene, dancers consider the situation (music, opponent, physical position) more closely than their own movements. The reason why these differences were shown was as follows. Since in the solo scene with no opponent, dancers did not have to communicate with the opponent, they could concentrate on their own performance. However in the battle scene, dancers need to communicate with the opponent and to deal with changes in the situation (OHJI, 2001), so they concentrated on information about the situation. We describe the details of these processes below.

##### **Analyses of statements about dance movements corresponding to personal discoveries**

Table 4. Example statement about the process of personal discovery in a solo scene

S171: Why did you rate this dance as novelty score 2?
G173: What should I say? I did an uprock in this dance. Usually, I don't perform this movement a lot.
S172: Is that this movement? (Watching the video).
G174: Yes. Because of this movement, I rated this dance as score 2.
S173: Why did you suddenly sit down? When you performed this movement, what were you thinking?
G175: I thought that in trial 1 or 2, I had danced in a standing position a lot, and I didn't do a move like sitting down. So I did this sitting movement in trial 3.
S174: So was this a movement which you tried to do intentionally?
G176: Yes, I decided on it just before the movement.

In this section, we focus on the objects of consideration of dancers in personal discoveries by analysing the data of high creativity score dance movements. Because of the low number of corresponding dance movements, we could not find a statistical difference between the two scenes. However, the results suggest that c) Consideration of composition, d) Consideration of the music, and g) No specific consideration frequently appeared in the solo scene (numbering 4 dance movements, 5 dance movements, 4 dance movements, out of a total of 14 dance movements). In the battle scene, d) Consideration of the music, e) Consideration of the opponent, f) Consideration of physical position frequently appeared (respectively, 8 dance movements, 3 dance movements, 4 dance movements, out of a total of 17 dance movements). From these results, we conclude that dancers consider their dance movements and make personal discoveries in the solo scene, while dancers in the battle scene consider information about the situation more closely and make personal discoveries.

Besides these implications, two statements (Table 4, 5) about personal discoveries suggest that in the solo scene, dancers considered the context of each dance movement, and intentionally made new patterns of dance movements. In contrast, in the battle scene the dancers tried to consider the situation, and deal with changes in the situation. However, they were able to make use of only limited patterns of well-practiced dance movements. Failure in dynamical dance movements such as power moves (one of the four core patterns in breakdance) leads to a loosening of the restrictions of the patterned dance movements, and the dancers are able to find new patterns.

#### Differences between the solo scene and the battle scene

We investigated the reasons why differences between the solo scene and the battle scene existed. One participant clearly mentioned how the two scenes differed (Table 6). In

Table 5. An example statement about the process of personal discovery in a battle scene

S141: Do you usually find new patterns in a battle scene?
B144: I usually don't use only fixed patterns. When performing, I just think what techniques (movements) I should use next. So, that was it. I just wanted to do a short one. I also thought that I would use free and flexible dance movements in the rest of the performance. I'm always ready for freeze movements anytime when it's necessary.
S142: So, do you dance with flexible combinations of movements when you dance freely?
B145: Yes, I always use flexible combinations, maybe. However, even in those combinations, I might have a tendency to use some particular combinations of fixed patterns.
S143: What do you think about this dance in terms of your tendency?
B146: This dance is not in keeping with that tendency. It goes against the tendency. T (the opponent) might have thought that this dance looked great.
S144: Why do you think you performed dance in this way?
B147: Hmm, my physical position after doing Trax <sup>5</sup> in that situation was probably a little different from the usual one. I didn't think of anything when performing.
S145: You didn't think of anything during the performance, but the physical position was different from usual.
B148: It's different, but maybe the music is one of the factors that caused it.

the solo scene, which had no opponent, dancers tended to perform well-practiced dance movements, not to fail and to arouse the audience, and they concentrated on their own performance. In contrast, in the battle scene, the dancers had to consider the improvisational communications with their opponent, which were thought to be an important factor in the battle scene, and they tried to think about the information (music, opponent) and to perform dynamical, impressive dance movements such as power moves.

#### Features of dance movements in each scene

We conducted a contrast analysis to compare the features of dance movements in solo and battle scenes in terms of the frequency of the four types of movement. The results of the sign test show that there were statistical differences in entry (dance movements in a standing position) (solo scene: 42.4 (22.18) seconds, battle scene: 33.3 (12.18) seconds) and power moves (dance with acrobatic movements like rolling) (solo scene: 15.1 (13.5) seconds, battle scene: 17.5 (10.55) seconds) ( $p=.057$ ,  $p=.092$ ). These results suggest that dancers perform dynamical movements (like rolling or

<sup>5</sup> One of the dance movements which is categorized as a power move.

Table 6. An example statement about the difference between the two scenes

M94: In the battle scene, I usually compose my dance movements, taking the situation of the place into account, while watching the dance of opponent. But in the solo scene, since I'm used to performing in public, I use well-practiced and skilled dance movements.
S94: You use mainly well-practiced dance movements?
M95: Yes, it was so in the solo scenes. In the battle scene, I wanted to pay more attention to my partner.
S95: What do you think makes this difference between the two scenes?
M96: Partners are an essential part in battle scenes. Communication with the partner is important and an interesting aspect of the battle scene. A dancer who is good at that communication looks cool, I think. In the solo scene, however, to be applauded is important, and I want to give a skilled performance to accomplish it. So I tend to use well-practiced dance movements.

jumping) more frequently in the battle scene. This result matches with the inference of the previous section, which suggested that dancers considered the opponent and tended to perform dynamical dance movements more frequently in the battle scene.

### General Discussion

This study has investigated the cognitive processes of dancers in improvisational activities such as the battle scene of breakdance, focusing on personal discoveries. The results have shown the following three findings. 1) Dancers mainly used fixed patterns of dance movements (about 60-70% of the whole dance) and gradually found new patterns of dance movements in improvisational activities. 2) By failing in dynamical movements, they were able to loosen the constraints of fixed patterns of dance movements, and found new patterns. 3) The processes of personal discovery (finding new patterns) varied with the presence of an opponent (partner). With reference to point 3, the following two processes have been revealed. In the absence of an opponent, the dancers thought about their own dance movements, and found new patterns by considering carefully the composition of their dance movements. In presence of an opponent, dancers considered the information about the situation (such as the music, opponent), and tended to perform dynamical dance movements more frequently. Then, when failing in these dynamical movements, they had to continue their performance from the present physical position that was different from their dominant (fixed) patterns, and they were able to find new patterns that were beyond fixed patterns.

This study has contributed new and clear findings about the features of improvisation based on the findings of previous studies (e.g., Mendonça & Wallace, 2004; Pressing, 1984; Tayanagi, 2010). Through a concrete example, we

describe the process by which dancers utilized errors to make new movement patterns, relaxing the constraints of their fixed patterns. In addition, we have focused on the original aspect of "the differences between improvisational activity of one person and that of multiple persons". The fact that there are differences between solo and collaborative activities has been suggested in many domains, especially in the domain of creativity (e.g., Okada & Simon, 1997, in scientific discovery). However, there has been no clear suggestion of these differences in the domain of improvisation. This study contributes original insight into the domain of improvisation.

In order to acquire more detailed understandings of improvisation, further studies are needed to solve problems such as the problem of generalization and research method.

### Acknowledgments

We would like to express our gratitude to all the participants for having spent so much time on our research and for giving us extremely useful suggestions about breakdance.

### References

- Bailey, D. (1980). *Improvisation: Its Nature and Practice in Music*. Buxton: Moorland Publishing.
- Boden, M. A. (1991). *The creative mind: Myths and mechanisms*. New York: Basic Books.
- Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, 17, 397-434.
- Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications*. Cambridge, MA: MIT Press.
- Mendonça, D. & Wallace, W. A. (2004). Cognition in Jazz Improvisation: An Exploratory Study. *Proceeding of the 26<sup>th</sup> Annual Meeting of the Cognitive Science Society* (pp. 5-10). Chicago, IL: Lawrence Erlbaum Associates.
- OHJI. (2001). *ROOTS OF STREET DANCE*. Tokyo: Bunkasha.
- Okada, T., & Simon, H. A. (1997). Collaborative discovery in a scientific domain. *Cognitive Science*, 22, 107-130.
- Pressing, J. (1984). Cognitive processes in improvisation. In Crozier, W. R., & Chapman, A. J. (Eds.), *Cognitive processes in the perception of art*. Amsterdam: North Holland.
- Suwa, M., & Tversky, B. (1997). What do architects and students perceive in their design sketches? A protocol analysis. *Design Studies*, 18, 385-403.
- Tayanagi, E. (2010). Performance design and innovation in music: A case of improvisation and innovation in modern jazz. *Cognitive Studies*, 17, 459-473.
- Weisberg, R. W., Brinkman, A. R., Folio, C. J., Dick, A., Fleck, J. I., Niederberg, B., & Barrett, F. (2004). Towards a cognitive analysis of creativity: Improvisation in jazz. In R. Parncutt, & A. Kessler, & F. Zimmer, (Eds.), *Proceedings of the Conference on Interdisciplinary Musicology (CIM04)*. Graz/Austria: University of Graz.

# Knowing When to Abandon Unproductive Learning

**Thomas R. Shultz (thomas.shultz@mcgill.ca)**

Department of Psychology and School of Computer Science, McGill University  
1205 Penfield Avenue, Montreal QC, Canada H3A 1B1

**Eric Doty (eric.doty@mail.mcgill.ca)**

Department of Psychology, McGill University  
1205 Penfield Avenue, Montreal QC, Canada H3A 1B1

**Frédéric Dandurand (frederic.dandurand@gmail.com)**

Department of Psychology, Université de Montréal, 90 ave. Vincent-d'Indy  
Montréal, QC H2V 2S9 Canada

## Abstract

Autonomous learning is the ability to learn effectively without much external assistance, which is a desirable characteristic in both engineering and computational modeling. We extend a constructive neural-learning algorithm, sibling-descendant cascade-correlation, to monitor lack of progress in learning in order to autonomously abandon unproductive learning. The extended algorithm simulates results of recent experiments with infants who abandon learning on difficult tasks. It also avoids network overtraining effects in a more realistic manner than conventional use of validation test sets. Some contributions and limitations of constructive neural networks for achieving autonomy in learning are briefly assessed.

**Keywords:** autonomous learning; abandoning learning; constructive neural networks; SDCC.

## Introduction

Autonomous learning is the ability to learn effectively without much external assistance. As such, autonomy is a desired quality in fields such as machine learning and artificial intelligence where the effectiveness of learning systems is seriously compromised whenever human intervention is required. It is likewise a desired feature in cognitive science where a goal is to understand the adaptive functioning of human and other biological agents in their natural environments. An important characteristic of autonomous learners is that they can shape their own learning and development, in large part by choosing what problems to work on. Such choices include selecting a problem to learn and deciding whether to continue learning on the selected task or abandon it in favor of something else.

## Knowing When to Quit

Knowing when to stop learning has two obvious components – quitting when the problem has been mastered and when it is unlikely to be mastered. In the constructive neural networks that we favor, victory is declared, and learning terminated, when the network is correct on all training examples, in the sense of producing outputs that are within some score-threshold of their target values (Fahlman & Lebiere, 1990; Shultz, 2003).

Cessation of learning without mastery is considerably more problematic, despite being an important component of autonomous learning in biological agents. It may be useful to analyze such early quitting in terms of costs and benefits. The total cost of learning can be conceptualized as energy expenditure (of the learning effort) plus opportunity cost (the value of the best alternative not chosen, whether other learning or exploitation of resources):  $Cost_{Total} = Energy_{Learn} + Cost_{Opportunity}$ . Then the net payoff of learning is the benefit of successful learning minus the total cost of learning:  $Payoff_{Net} = Benefit_{Learn} - Cost_{Total}$ . In continuing to work on an unlearnable problem, there would be a large negative payoff, cost without benefit. Having started to learn such a difficult problem, it could be sensible to abandon it when lack of progress becomes evident.

## Previous Work on Abandoning Learning

Recent computational modeling does suggest that a key factor in deciding to abandon learning early is whether learning progress is being made (Schmidhuber, 2005, 2010). In that work, learning progress is monitored by tracking the first derivative of error reduction to identify intrinsic rewards, while a reinforcement-learning module selects actions to maximize future intrinsic rewards. These models curiously conflate novelty with learning success, but it seems more correct to base novelty on initial error, and compute learning success as recent progress in error reduction. These models also include a reinforcement-learning controller that selects actions, and an external network to track learning progress. It seems simpler to continue learning by default until lack of progress is detected, perhaps in terms of stagnation in error reduction.

In an idealized learning model, infant looking was modeled by information-theoretic properties of stimuli (Kidd, Piantadosi, & Aslin, 2010). The negative log probability of an event (corresponding to the number of bits of information conveyed by a stimulus) was conditioned on observing previous events. The larger the negative log probability, the more surprising the current event. As predicted, 7- to 8-month-old infants were more likely to look away from either highly informative or uninformative events. The authors dubbed this the *Goldilocks* effect as



infants prefer to work on tasks that are neither too easy nor too difficult, but just about right in terms of complexity. Although interesting and consistent with an idealized statistical model, these findings are not tied to any neural computational mechanisms. Also, this model is presumably restricted to repeated sequences of events.

Other recent experiments reported that 17-month-olds attend longer to learnable versus unlearnable artificial-language grammars, taking more trials and more time on grammars in which a valid generalization over input utterances could be made (Gerken, Balcomb, & Minton, 2011). Thus, there is now independent evidence that infants may have an implicit metric of their learning progress and can direct their attention to more learnable material.

### Constructive Artificial Neural Networks

Constructive artificial neural networks (CANNs) grow a network topology while learning, inspired by principles of brain function and statistical mechanics. Among the attractive features of CANNs are graded knowledge representations, capacity for change and self-organization, and neurological plausibility. CANNs such as cascade-correlation (CC) grow by recruiting new hidden units whose activity correlates with network error (Fahlman & Lebiere, 1990). An extension, sibling-descendant cascade-correlation (SDCC), dynamically decides whether to install a newly recruited unit on the current highest layer (as a sibling) or on its own higher layer (as a descendant), thus optimizing the network topology for the problem being learned (Baluja & Fahlman, 1994). Unit recruitment corresponds roughly to processes of neurogenesis and synaptogenesis in the service of learning (Shultz, Mysore, & Quartz, 2007). Such CANNs have been used to simulate many cognitive, linguistic, and social phenomena while addressing important and longstanding issues about development and learning (Shultz, 2003; Shultz & Fahlman, 2010). They have also yielded testable predictions, many of which have been confirmed in psychological research. Moreover, CANNs have also made considerable progress on several aspects of autonomous learning, including network construction in which new abilities are built on top of earlier achievements.

In the present work, we extend SDCC to abandon learning that is failing to make progress. This is a natural extension for SDCC, which already is able to change phases when it detects lack of progress. Both CC and SDCC operate in two phases: output phase, in which connection weights entering output units are adjusted to reduce network error, and input phase in which weights entering hidden units are adjusted in order to increase the covariance between candidate-unit activation and network error, which ends up recruiting the candidate that best tracks network error. Output phase ends when error reduction stagnates, whereas input phase ends when the covariances between candidate activation and network error stop changing.

We hypothesized that, if error stagnation continues even after recruitment, this could additionally signal that the problem might be unlearnable. This would be the case, for

example, on problems with a random structure and insufficient regularities. Of course, some potentially learnable problems are so difficult that their patterns may only seem random. In either case, learning may be frustratingly slow and thus signal to stop and turn to something else more feasible. Here, we apply our extended algorithm to learning problems of varying randomness, discuss its potential to cover the infant experiments just reviewed, and briefly assess the overall ability and limitations of CANNs to learn autonomously.

## Method

### Algorithm Extension for Abandoning Learning

As noted, each of the two phases in CC and SDCC assesses progress within a phase. We define a learning cycle as an input phase, which recruits a hidden unit, followed by an output phase, which employs the new recruit to help reduce network error. (The first learning cycle has only an output phase, and no input phase.) To assess learning progress across learning cycles, we implemented a new, outside loop to assess progress at the end of each output phase, according to the following algorithm, in which a counter is initialized to 0:

If first learning cycle, then record current error and continue to input phase

Otherwise, compare current error to previous error as absolute difference

If absolute difference > threshold  $\times$  previous error, then reset counter to 0 and continue to input phase

Otherwise,

If counter = patience, then abandon learning

Otherwise, increment counter by 1 and continue to input phase

This algorithm is analogous to the progress-assessing loops already used in the output and input phases of CC and SDCC, which compute an absolute difference between a current and previous measure (network error for output-phase and learning-cycle loops, covariance for input-phase loops), and test if this difference is greater than a threshold proportion of the previous value. If the absolute difference exceeds this product, learning continues. If it does not exceed this product, then there is a check to determine if a patience parameter value has been reached. If patience has been exceeded, then the current loop is terminated; otherwise the patience counter is incremented by 1 and learning continues. Resetting the counter to 0 whenever the threshold proportion is exceeded insures that the number of cycles without exceeding the threshold proportion must be consecutive rather than sporadic. Although we rarely alter the threshold and patience parameters for output and input phases, here we do explore some parametric variation for assessing progress across learning cycles.

### Continuous XOR

We tested our extended algorithm on a continuous version of the exclusive-or (XOR) problem. This is a well

understood problem in which the simplicity of binary XOR is replaced by a more complex continuous version (Shultz & Elman, 1994; Shultz, Oshima-Takane, & Takane, 1995). Starting from 0.1, input values are incremented in steps of 0.1 up to 1, producing 100  $x, y$  input pairs that are partitioned into four quadrants of the input space, as illustrated in Figure 1. There is a single output unit with a sigmoid activation function. Values of  $x$  up to 0.5 combined with values of  $y$  above 0.5 produce a positive output target (0.5), as do values of  $x$  above 0.5 combined with values of  $y$  up to 0.5. Input pairs in the other two quadrants yield a negative output target (-0.5). These constitute the training patterns for conditions that are completely learnable.

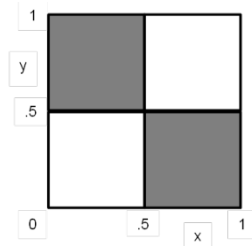


Figure 1. Schematic drawing of the continuous-XOR problem. Gray sectors yield a positive output while white sectors a negative output.

To implement problems of different levels of difficulty, we vary learnability, defined as the percentage of target outputs that are not randomly selected: 0, 25, 50, 75, 80, 85, 90, 95, or 100. If a fresh random number in the range  $[1, 100] \geq$  the particular learnability percentage, then the output target (-0.5 or 0.5) is selected by a .5 chance.

Generalization test patterns are generated by incrementing  $x$  and  $y$  values by 0.1 to .94 starting from 0.14. There are 81 such test patterns, all with correct outputs.

In preliminary simulations, it became apparent that learning results were also sensitive to variation in the threshold parameter, so we varied threshold systematically (.05, .1, .15, .2, and .3), while holding patience at 2.

## Results

We do not present all of our results here, but only those needed to make important points about basic principles.

### Learning Threshold of .15

Typical training-error results are plotted in Figure 2 for two networks, one exposed to patterns with 50% learnability and the other exposed to patterns with 100% learnability. Learning threshold is here set to .15. The diamonds just above the error curves indicate the particular output-phase epochs at which a hidden unit is recruited. As is typical for all threshold values, error is reduced much further with full learnability than with 50% learnability. Moreover, as is typical for thresholds of .1 and higher, learning is abandoned much earlier with 50% learnability than with 100% learnability. These results suggest that the extended

algorithm is effective at detecting lack of progress in learning and show what underlies grouped results to follow.

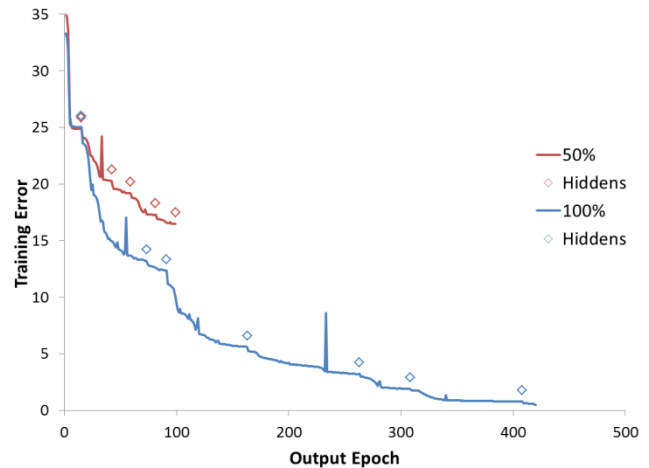


Figure 2. Training error in two networks. With 50% learnability, learning stops at 99 epochs. With 100% learnability, learning stops at 420 epochs.

To see a more general picture, mean per-pattern training error for 20 networks under each learnability condition is plotted in Figure 3, again for a learning threshold of .15. Per-pattern error for a network is computed by dividing total network error by the number of patterns. Each curve is cut off at the mean number of output-phase epochs to abandon learning for that level of learnability, even though some networks surpass this number. Figure 3 provides a more complete demonstration that error reduction is greater with higher learnability and that the extended algorithm is effective at detecting lack of learning progress. Generally, the lower the learnability, the earlier learning is abandoned, at least up to 90% learnability.

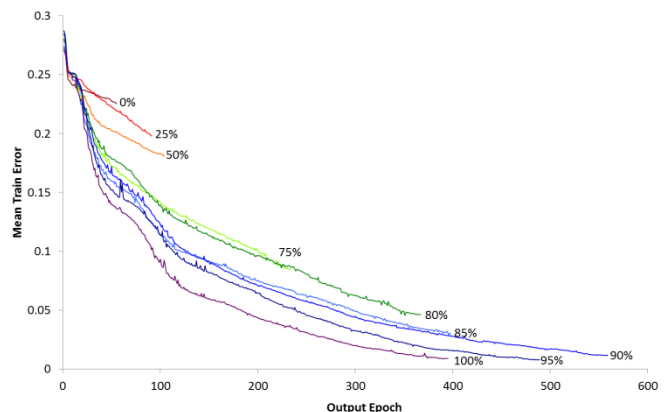


Figure 3. Mean per-pattern training error for 20 networks under each learnability condition over learning cycles.

Mean learning times are shown in Figure 4, which plots the mean output epochs and SE bars for the same 20 networks. This shows more abstractly that low levels of learnability lead to early abandonment of learning.

Moreover, the inverted U-shaped curve reveals a substantial Goldilocks effect wherein networks show more sustained learning for problems of moderate difficulty, peaking at 90% learnability.

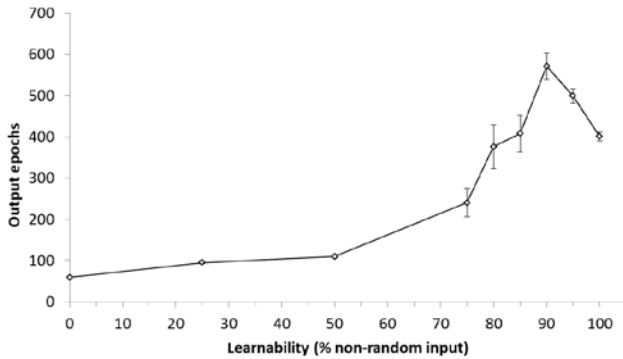


Figure 4. Mean output epochs (with SE bars) for 20 networks with learning threshold of .15.

However, in these simulations, the prolonged learning characteristic of the Goldilocks peak does not often yield superior performance. This is illustrated for these same networks in Figure 5, which plots mean per-pattern test error for each learnability condition. Notice the rise in error on test patterns for learnability conditions in the 50-90% range. Such increases in test error over training suggest that networks are over-fitting the training patterns and starting to memorize the random training patterns instead of abstracting a function to account for the examples. Their earlier success in bringing error down is presumably due to abstracting the continuous-XOR function. But from then on, their only recourse is to start memorizing the random patterns. At 0% learnability, it is impossible to abstract even a basic idea of the exclusive-or problem.

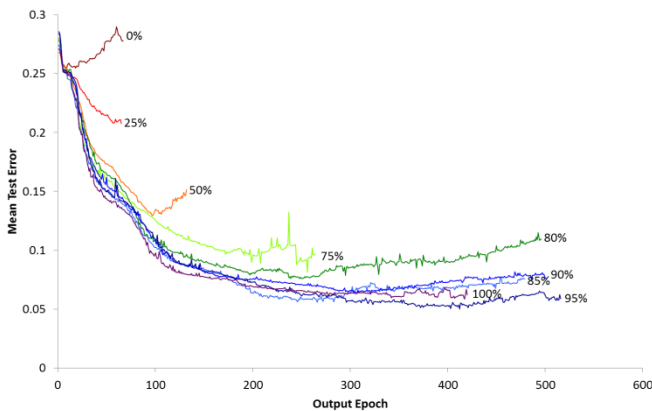


Figure 5. Mean per-pattern test error averaged across 20 networks under each learnability condition.

A rise in test error in what is typically called the validation test set is conventionally used by programmers to determine when to stop network learning. This can be particularly important when using static networks with

back-propagation, which have no natural stopping point and where there is no a priori idea of how many hidden units with which to equip a network. Such validation test sets are ordinarily unnecessary for CANNs, which start small and keep growing until the problem is learned. With substantial numbers of random patterns to be memorized, as here, it can be beneficial to also use test error as a training aid, even for CANNs. With a learning threshold of .15, the extended SDCC algorithm was unable to detect, from training error alone, that learning was not progressing, in the sense of generalization ability. Although validation test sets are useful for programmers, they are unrealistic for autonomous learners. Whenever target values and the resulting error signals are available, it is likely that learners would use them to adjust connection weights, thus effectively eliminating such examples from the validation test set.

### Learning Threshold of .3

This raises the question of whether other, less sensitive learning-threshold values could be used to curtail learning investment in unproductive tasks like our 50-90% learnability conditions. The answer, as revealed in Figure 6, is yes for a learning threshold of .3. In this case, there are no general increases in test error, except at 0% learnability.

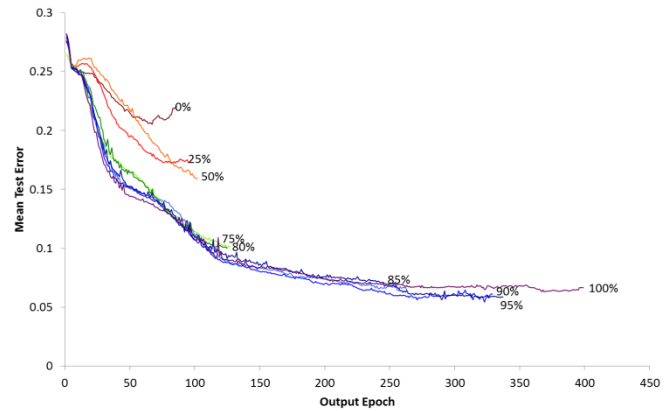


Figure 6. Mean test error averaged across 20 networks under each learnability condition.

However, the Goldilocks effect for these same networks disappears, as revealed in Figure 7. The learning-time peak is now at 100% learnability as all other conditions have abandoned learning earlier. More generally, we find a trade-off between the Goldilocks effect and avoidance of rising test error. As learning threshold increases, the likelihood of finding a Goldilocks effect drops.

## Discussion

### Interpretation of Results

Our results show that monitoring progress across learning cycles can be used to abandon learning that is unlikely to be successful. This is both realistic and adaptive because, with many problems and domains to learn, it is wasteful to

devote time and energy to learn tasks that are too difficult or impossible. In an abstract sense, on an admittedly different task, our simulations show the ability to capture results like those in two new experiments on learning in human infants. Infants spend more time learning artificial grammars that are possible to learn than they do on grammars that are impossible to learn (Gerken, et al., 2011). Similarly, our neural networks abandon learning impossible tasks, but not tasks that are possible to learn. Further, the network results show that the more difficult the task, the earlier that learning is abandoned, a finding that could serve as a prediction for new human experiments.

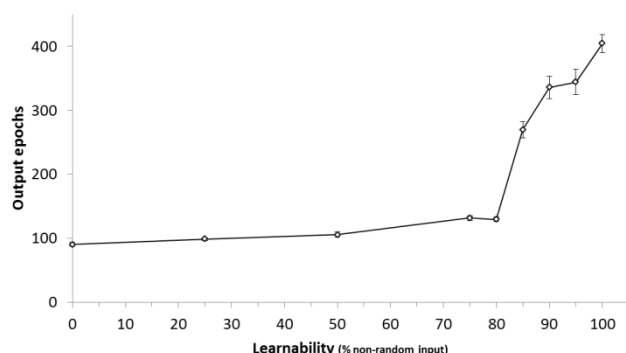


Figure 7. Mean output epochs (with SE bars) for 20 networks, with learning threshold of .30.

Another infant experiment showed a Goldilocks effect in the sense of spending more learning effort on problems of moderate difficulty than on problems that are too easy or too difficult (Kidd, et al., 2010). Our networks show this effect as well, but add a qualification that the Goldilocks effect diminishes at higher levels of a learning-threshold parameter. This offers another prediction to test in human experiments. The psychological equivalent of learning threshold could be sensitivity to changes in error.

In our model, easy tasks are discarded because they have been learned, whereas overly difficult tasks are abandoned because learning has stalled. This identifies two different explanations for turning away from a learning task, one based on success and the other on failure. In contrast, learning may continue as long as some detectable progress is being made.

Our model offers a plausible neural mechanism for such phenomena that allows for further theoretical exploration and extensions. We plan to apply our algorithm to alternate tasks and problems, including those used in psychology experiments and those that vary on dimensions of difficulty other than the proportion of random training patterns.

Our model predicts that learners need to have learning experience with a problem in order to determine whether to continue with it or not. At least with inexperienced learners, there is no shortcut to avoid actually trying to learn. Supporting this idea, we found that amount of first-trial error does not predict learnability on the problems we studied here. Learners may need to give it a serious try

before being able to predict whether they might succeed. It would be interesting to see if this is also true of biological learners. If learners exhibit shortcuts to avoid attempted learning, this would imply generalizing across learning content due to previous experience, as when learning shuts down in the presence of mathematical equations.

We also found that overtraining effects can be eliminated with high learning thresholds. This is more realistic for autonomous agents than is monitoring error increases on a validation set of test patterns. Moreover, we find that the Goldilocks and overtraining effects tend to occur in the same circumstances, at relatively low rather than high learning thresholds. Goldilocks peaks are due to the increased learning times caused by low learning thresholds.

There is, of course, more to autonomous learning than abandoning unsuccessful learning. There is also, for example, the choice of which problems to try to learn. We hypothesize that novelty detection, characterized by high initial error, plays a role in choosing learning problems. Abandoning fruitless learning is an essential component of autonomous learning because, as noted, it frees the learner to search for and work on more appropriate problems.

## Achieving Autonomy in Learning

Our results show that a small extension to SDCC can provide a useful mechanism for detecting lack of progress in learning, which is an essential component of autonomous learning. In this context, it is worth considering how CANNs such as SDCC fare in terms of other aspects of autonomous learning (Douglas & Sejnowski, 2007). Although there are no completely autonomous artificial learning creatures yet, it is also true that CANNs have made considerable progress in increasing autonomy in learning.

In terms of network construction, SDCC, unlike algorithms for human-designed networks, autonomously designs and builds a network topology that is well suited to the problem being learned. The emerging topology can be flat or deep or anything in between, and learning stops when the problem has been mastered.

Unlike the ordered hierarchies of some static network topologies, SDCC implements a potentially deep, heterarchical topology in which increasingly higher-level, more abstract concepts are composed of simpler ones. Each new hidden unit in SDCC receives signals from input units and any existing lower level hidden units, thus continually building on existing knowledge. Knowledge-representation analysis shows that the first hidden units learn to represent the most obvious and superficial aspects of a problem domain, whereas later hidden units refine and abstract that knowledge (Shultz, 2003). This componential structure is further enhanced in knowledge-based CC (KBCC), where whole, previously learned sub-networks compete to be recruited (Shultz & Rivest, 2001; Shultz, Rivest, Egri, Thivierge, & Dandurand, 2007).

With regard to data selection, like many other artificial neural networks, SDCC focuses on inputs that predict its output, quickly ignoring inputs that are not predictive.

Although such non-predictive inputs are rarely included in practice, it is important to note that, when they are included, they are rapidly and functionally eliminated by learning of near-zero connection weights. It would be feasible to eliminate such detected irrelevant inputs from training patterns altogether, effectively allowing learning to focus attention on what is important while creating a more efficient network.

Among the issues that remain challenges for CAANs, as well as for other network learning algorithms, are single-trial learning, temporal spacing effects, the wake-sleep cycle, synaptic meta-plasticity, relations between brain structure and function, real-time learning in a changing world, and social learning (Douglas & Sejnowski, 2007).

The role of supervision of learning is a complex topic deserving more extended discussion than we can provide here. Suffice it to say that CANNs can learn without a teacher.

For more genuine and more complete autonomy in learning, we believe that it will be important to examine the evolution of learning methods and to implement computational models in robots, with pressures for real-time behavior in fluid environments. Evolution through natural selection is the most plausible natural source of learning mechanisms in both biological and artificial agents (Dunlap & Stephens, 2009). Based on the cost-benefit analysis we presented in the Introduction, it might be possible to show that abandonment of learning itself is favored by natural selection in evolution simulations. And, of course, robotic applications pose a particularly challenging test of learning autonomy.

## Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada, with an operating grant to TRS and a fellowship to FD. Planning this work benefitted from discussions with LouAnn Gerken, Nick Chater, and Scott Fahlman. We are also grateful to Vincent Berthiaume for relevant pilot work, Simon Reader for pointers to papers on evolution and learning, and Caitlin Mouri for helpful comments on an earlier draft.

## References

- Baluja, S., & Fahlman, S. E. (1994). Reducing network depth in the cascade-correlation learning architecture. In Pittsburgh, PA: School of Computer Science, Carnegie Mellon University.
- Douglas, R., & Sejnowski, T. (2007). Final workshop report: future challenges for the science and engineering of learning, National Science Foundation (USA). Washington, DC.
- Dunlap, A. S., & Stephens, D. W. (2009). Components of change in the evolution of learning and unlearned preference. *Proceedings of the Royal Society B-Biological Sciences*, 276, 3201-3208.
- Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in neural information processing systems 2* (pp. 524-532). Los Altos, CA: Morgan Kaufmann.
- Gerken, L. A., Balcomb, F. K., & Minton, J. L. (2011). Infants avoid 'labouring in vain' by attending more to learnable than unlearnable linguistic patterns. *Developmental Science*, 14, 972-979.
- Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2010). The Goldilocks Effect: Infants' preference for stimuli that are neither too predictable nor too surprising. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2476-2481). Austin, TX: Cognitive Science Society.
- Schmidhuber, J. (2005). Self-motivated development through rewards for predictor errors / improvements. In *Developmental Robotics 2005 AAAI Spring Symposium*. Stanford University, CA.
- Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation. *IEEE Transactions on Autonomous Mental Development (1990-2010)*, 2, 230-247.
- Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge, MA: MIT Press, (Chapter Chapter).
- Shultz, T. R., & Elman, J. L. (1994). Analyzing cross connected networks. In J. D. Cowan, G. Tesauro & J. Alspector (Eds.), *Advances in Neural Information Processing Systems 6* (pp. 1117-1124). San Francisco, CA: Morgan Kaufmann.
- Shultz, T. R., & Fahlman, S. E. (2010). Cascade-Correlation. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning, Part 4/C* (pp. 139-147). Heidelberg, Germany: Springer-Verlag.
- Shultz, T. R., Mysore, S. P., & Quartz, S. R. (2007). Why let networks grow? In D. Mareschal, S. Sirois, G. Westermann & M. H. Johnson (Eds.), *Neuroconstructivism: Perspectives and prospects* (Vol. 2, pp. 65-98). Oxford, UK: Oxford University Press.
- Shultz, T. R., Oshima-Takane, Y., & Takane, Y. (1995). Analysis of unstandardized contributions in cross connected networks. In D. Touretzky, G. Tesauro & T. K. Leen (Eds.), *Advances in Neural Information Processing Systems 7* (pp. 601-608). Cambridge, MA: MIT Press.
- Shultz, T. R., & Rivest, F. (2001). Knowledge-based cascade-correlation: Using knowledge to speed learning. *Connection Science*, 13, 1-30.
- Shultz, T. R., Rivest, F., Egri, L., Thivierge, J.-P., & Dandurand, F. (2007). Could knowledge-based neural learning be useful in developmental robotics? The case of KBCC. *International Journal of Humanoid Robotics*, 4, 245-279.

# Listening to Thematic Music Prior to a Generation Task Causes Thematic Elements to Be Included in a Story Generation Task.

Cynthia M. Sifonis (sifonis@oakland.edu)

Department of Psychology  
Rochester, MI 48309-4401 USA

William C. Fuss (wcfuss@oakland.edu)

Department of Psychology  
Rochester, MI 48309-4401 USA

## Abstract

The current study examines whether thematic music (e.g., battle music) can activate related concepts in memory (e.g., weapons, death) and whether the activated concepts are more likely to be included in a story generation task. Participants listened to one of two 90-sec excerpts of thematic music (Baby theme or War theme) either before or after engaging in a story generation task. Their stories were examined to determine the degree to which the thematic elements of the music were included in the stories. Across two experiments, there was evidence that listening to war or baby themed music before engaging in a generation task increased the likelihood that elements associated with that theme would be incorporated into the story. Evidence also existed that the music theme interacted with the story theme to influence the degree that thematic elements would be incorporated into the story.

**Keywords:** concepts; categories; knowledge representation; music; creativity; generation.

## Introduction

We use music to celebrate events, entertain, motivate and inspire us. Hearing a song can bring back memories of a college romance. The musical score of a movie warns us when something unpleasant is about to happen. Given the pervasiveness of music in our lives, surprisingly little research has been directed at determining whether music is represented in semantic memory and how the musical associations stored in semantic memory affect performance on a variety of tasks

People's musical knowledge develops throughout their lives via exposure to such things as lullabies, music education, background music in television and film, and musical entertainment. Because people live in a musically rich environment, they become attune to the complexities of musical structure at an early age. Even three-year-old children demonstrate an understanding of the relationship between music in a major key being associated with a happy mood and music in a minor key being associated with a sad mood (Kastner & Crowder, 1990). This type of knowledge affects five to six-year-old children's interpretation of a story such that "happy" music playing in the background causes the children to form positive interpretations of a neutral story whereas "sad" music playing in the

background causes them to form negative interpretations (Ziv & Goshen, 2006).

Exposure to certain types of music becomes associated with particular events throughout a lifetime of musical experiences (e.g., weddings, circuses). It stands to reason that the link between music and event becomes strong enough that knowledge of the music will join object and event information in the concepts and schemas with which it is associated (e.g., "wedding music," "circus music"). Empirically demonstrating the existence of these associations is just beginning though.

Daltrozzo and Schön (2009) argued that music conveys concepts in the same way that images and words convey concepts. They supported this claim by demonstrating a larger N400 component of the event-related brain potentials to 1-second musical excerpt targets following a conceptually unrelated compared to a conceptually related linguistic context. Though this is fairly direct evidence of conceptual processing of musical information, it is left to other researchers to explain how such a short excerpt of music carries enough information to activate conceptual information in memory. Stronger (albeit indirect) evidence for the conceptual representation and processing of musical information is provided by North, Hargreaves, & McKendrick (1997) who demonstrated that ethnic background music can affect consumer decisions. When French music was playing in a wine shop, consumers were more likely to buy French wine than German wine. When German music was playing, the opposite was true, Boltz (2001) also demonstrated the effect of music on actions by demonstrating that musical soundtracks could activate a schematic framework thus affecting the interpretation of an ambiguous movie scene. Boltz (2001) had participants watch an ambiguous scene paired with either "positive" or "negative" music. She found that when a scene was paired with positive music it activated a positive schema resulting in positive interpretations of the events occurring in the scene (e.g., the man following the woman is a long lost lover) and subsequent memory for positive objects (e.g., flower bouquet) and events in the scene. When a scene was paired with negative music it activated a negative schema resulting in negative interpretations of the events occurring in the scene (e.g., the man following the woman is her brother and plans to kill her) and subsequent memory for

negative objects (e.g., human skull) and events in the scene. An interesting component of the paradigm Boltz (2001) used to demonstrate music acting as a schematic influence on the cognitive processing of film events is that she had participants extrapolate the film's ending as a means of examining the effect of schema activation by looking for elements of those schemas in the extrapolations. In demonstrating the influence of the schemas activated by the music via the events and objects described in participant's extrapolations of the film's ending, Boltz (2001) also demonstrated that music can activate conceptual knowledge which then affects performance in a generation task.

Those studying creative cognition, frequently use generation tasks to examine the influences of concepts and categories on performance. In doing so, they gain knowledge about the contents and organization of conceptual knowledge and how this knowledge is applied when generating new ideas. For example, Marsh, Binks and Hicks (1999) demonstrated that participants shown examples sharing the conceptual feature of hostility (e.g., weapons, fangs) were more likely than those not experiencing the examples, to generate novel exemplars with hostile features. They also demonstrated that simply activating the concept of hostility was sufficient to influence performance on a generation task. Participants who unscrambled hostile sentences were more likely to generate novel exemplars with hostile features compared to participants who had unscrambled conceptually neutral sentences.

If people represent thematic music in memory and that representation is linked to the objects and activities with which it is associated, it should be possible to activate that representation and its associated concepts by having people listen to the music. For example, if the concept of "war" is associated with a certain type of music, then having people listen to "war music" should activate the war concept and the elements associated with war such as "marching" and "weapons." Once those concepts are activated, they should affect performance in a subsequent generation task by increasing the likelihood that conceptual elements will be incorporated into the novel product.

## Experiment 1

As discussed, music experienced during goal-directed activity (shopping, watching a movie) has the ability to affect purchasing decisions as well as story interpretation and generation in what appears to be a conceptually congruent fashion. What is less clear is whether our representation of thematic music is rich enough that listening to such music in the absence of other stimuli is sufficient to activate more complex concepts such "fighting in a war" and "tucking a baby into bed" much less instantiate those concepts in a generation task. The current experiment seeks to demonstrate the ability of music to activate complex concepts in memory and affect the content incorporated into stories generated in a story generation task.

Participants will listen to one of two types of thematic music ("tucking a baby into bed" (baby) theme, "going off to war" (war) theme) either before or after a story generation task. The stories generated by participants will be examined for the presence of elements associated with either one of the two music themes.

We hypothesize that participants who listen to the war themed music prior to engaging in the generation task will include more war themed concepts than baby themed concepts in their stories than people who listened to the war themed music after engaging in the generation task. Participants who listen to the baby themed music prior to engaging in the generation task will include more baby themed concepts than war themed concepts in their stories than people who listened to the war themed music after engaging in the generation task.

## Methods

### Pilot Testing

**Participants, Stimuli and Procedure.** Thirty-one Oakland University students participated in exchange for experimental credit in the song-rating task.

Participants wearing headphones were seated at a computer and interacted with a Flash program that provided instructions, randomized song presentation and collected participants' responses to the songs. The 14 songs presented to participants during pilot testing satisfied the criteria of 1) being strongly thematic and containing at least 90 seconds of music devoid of 2) environmental sounds (e.g., barking dog, crying baby) or 3) lyrics.

On each trial, participants listened to a 90 sec. song clip, then listed the "first three things the song made them think of" and rated the familiarity, pleasantness, and liking of the song on a 7 point Likert scale (7 = most familiar, pleasant, liked) (See Table 1).

Responses to each song were content analyzed to identify the three most salient concepts (listed by > 30% of sample) that each song brought to mind (MusicConcepts).

The two songs chosen for use (DeLaTerra and Carousel) were selected because they 1) both elicited consistent thematic responses from participants ("war" and "baby" respectively), 2) the themes were extremely different from each other and 3) the songs were equally unfamiliar.

It was thought that the activation of a concept by listening to the music might spread to associated concepts and affect performance on the generation task. Consequently, a different group of 18 students enrolled in a Cognitive Psychology course engaged in a feature listing task, listing features for the three Carousel and three DeLaTerra MusicConcepts. Nine conceptually related features (MusicAssociates) were listed by at least 33% of the students in response to the Carousel MusicConcepts and seven features were listed by at least 33% of the students in response to the DeLaTerra MusicConcepts.



participant.

## Stimuli and Procedure

Participants were seated at a computer with headphones on and interacted with a Flash program in the browser that provided instructions, presented stimuli and recorded participants' responses. All participants spent four minutes engaged in a nine-item Remote Associates Task (included to disguise the study's purpose), listened to 90 seconds of music, and spent 15 minutes writing a story with the theme "My Adventure on an Alien Planet." Conditions different only in terms of which of the two songs participants experienced and whether the songs were presented immediately before or immediately after the story generation task.

## Coding

Raters blind to condition coded the stories generated by the participants for the presence of the MusicConcepts and MusicAssociates associated with the two songs. This resulted in four dependent variables: the proportion of Baby MusicConcepts, the proportion of Baby MusicAssociates, the proportion of War MusicConcepts, and the proportion of War MusicAssociates included in the story written by the

## Results

Four 2 X 2 ANOVAs were conducted examining the effects of music theme (Baby, War) and position (Before, After) on the proportion of MusicConcepts and MusicAssociates included in the stories.

The analyses revealed no main effects or interactions on the tendency for participants to include the Baby MusicConcepts or MusicAssociates in their stories,  $p > .05$ . The same was also true for the tendency to include War MusicConcepts in their stories. However, there was a significant main effect of position on the proportion of War MusicAssociates incorporated into the story,  $F(1, 137) = 4.34$ ,  $p < .05$ ,  $\eta^2 = .032$ . This main effect was moderated by a non-reliable interaction,  $F(1, 137) = 3.15$ ,  $p < .10$ ,  $\eta^2 = .023$ . Participants exposed to the War theme music before the story generation task included a greater proportion of War MusicAssociates into their stories ( $M = .04$ ,  $SE = .01$ ) than those who were exposed to War music after the generation task ( $M = .00$ ,  $SE = .01$ ) or those who were exposed to the Baby music either before ( $M = .02$ ,  $SE = .01$ ) or after ( $M = .02$ ,  $SE = .01$ ) the generation task.

Table 1: Characteristics of 90-second Song Samples and Evoked Concepts (Concepts)

Theme	Music		Music Primes	Familiarity		Pleasantness		Liking	
	Song	Artist		M	SD	M	SD	M	SD
Church	Dante's Prayer	Loreena McKinnett	Church, Sadness, Peace	3.4	1.77	4.75	2.01	4.34	1.81
Evil	Toccatta and Fugue in D Minor	Bach	Church, Organ, Horror	4.91	1.85	3.91	1.72	3.89	1.86
	Carmina Burana Introduction	Orff	Choir, Movies, Marching Band	6.00	1.60	4.53	1.56	4.26	1.89
Indian	Ahini-Lalita	Ravi Shankar	Middle East, India, Desert	3.51	2.09	4.29	1.53	3.89	1.60
Arabia	Lawrence of Arabia	Henry Mancini	Movies, Broadway	4.18	1.95	5.15	1.48	4.53	1.50
	Marco Polo	Loreena McKinnett	Desert, India, Egypt	3.80	1.73	4.91	1.63	4.57	1/60
Latin	Latin Quarter	Big Lazy	Spanish, Salsa, Tango	3.60	2.04	4.97	1.46	4.37	1.54
Asian	Asian-Thai-Classic Chinese Folk Music	Chinese Folk Music	China, Asia, Oriental	3.60	2.16	5.03	1.54	4.40	1.42
Circus	Cirque De La Mort	Vernian Process	Circus, Carnival, Scary	2.88	1.92	2.97	1.57	2.74	1.66
	Death of a Doll Maker	Creature Feature	Video game	2.69	2.34	3.57	1.82	3.4	2.02
War	De La Terre a La Lune	Vernian Process	Marching, Marching band, War	3.06	2.10	3.5	1.24	3.0	1.37
	Your Betrayal	Bullet for My Valentine	Rock, Rock Concerts	3.67	2.04	3.74	1.83	4.24	1.89
Hostile	Agitated Screams of Maggots	Dir en Grey	Anger, Rock, Screamo	2.69	2.08	2.06	1.63	2.09	1.67
Baby	Se Lest	Sigur Ros	Baby, Sleep, Music Box	2.32	1.97	4.15	1.69	3.38	1.48
	Carousel on a Slide Projector	Lullatone	Sleeping Baby, Lullaby	3.09	2.15	5.00	2.15	4.09	1.88

## Discussion

The hypothesis that thematic music experienced prior to engaging in a story generation task increases the likelihood thematic elements associated with that music will be incorporated into the story is partially supported. Participants exposed to music that brings to mind concepts associated with fighting in a war are more likely to incorporate concepts such as weapons, death and blood into the stories they write immediately after listening to the music. However, exposure to war themed music did not increase their tendency to incorporate the specific terms brought to mind by the music (as revealed in pilot testing) into their stories. Exposure to baby themed music did not appear to affect whether or not concepts associated with tucking a baby into bed were included in the stories either.

So why did exposure to war themed music affect performance and exposure to baby themed music did not? Perhaps because baby themed elements were incorporated into the stories at higher rates than war themed elements across all conditions. Specifically, baby themed elements related to sleeping and dreaming were commonly included in the stories generated by participants.

We believe participants faced with the task of describing their “adventure on an alien planet” seemed compelled to explain how they ended up on that alien planet. A common solution was to say they fell asleep in bed and dreamt that they were on an alien planet. If this is a valid explanation for the lack of influence of baby themed music on performance in a generation task, then perhaps a more plausible setting will decrease the use of literary devices such as sleeping and dreaming in the story generation task.

## Experiment 2

Experiment 2 tests the proposal that the story generation task scenario given to participants in Experiment 1 hid the influence of baby themed music by causing them to incorporate items associated with that theme (sleeping, dreaming, bed) into their stories to rationalize being on an alien planet.

We predict that providing a more “down-to-earth” scenario will allow the effects of the baby themed music to be manifested in the story generation tasks by decreasing the inclusion of the specific baby-themed concepts of sleeping and dreaming across all conditions as a literary device. We maintain the hypotheses that experiencing war themed music prior to engaging in a story generation task will increase the proportion of war themed items included in the story compared to when the music is experienced after the generation task. Similarly, experiencing baby themed music prior to the generation task will increase the proportion of baby themed items included in the story compared to when the music is experienced after the generation task or compared to when war themed music is experienced prior to the generation task.

## Methods

### Participants

One-hundred and seventeen Oakland University students participated in exchange for experimental credit and were randomly assigned to one of four conditions: 29 listened to Carousel before generation (Baby Before), 33 after generation (Baby After), 28 participants listened to DeLaTerra before generation (War Before) and 27 after generation (War After). None of the participants provided data for Experiment 1.

### Stimuli and Procedure

The stimuli and procedure were identical to Experiment 1 with the exception that participants were now asked to spend 15 minutes writing a story with the theme “An Adventure to a Foreign, Undiscovered, yet Inhabited Land.”

## Results

Four 2 X 2 ANOVAs were conducted examining the effects of music theme (Baby, War) and position (Before, After) on the proportion of MusicConcepts and MusicAssociates included in the stories.

Analyses revealed a significant music theme X position interaction on the incorporation of Baby MusicConcepts incorporated into the story in the generation task,  $F(1, 117) = 5.41, p < .05, \eta^2 = .05$ . Participants exposed to the baby themed music prior to the generation task ( $M = .10, SE = .02$ ) were more likely to incorporate Baby MusicConcepts items into their stories than people who were exposed to the music after the generation task ( $M = .03, SE = .02$ ) or people who were exposed to the war themed music before the task ( $M = .02, SE = .02$ ). In fact, those exposed to the war themed music before the generation task were actually less likely to incorporate Baby MusicConcepts into their stories than those exposed to the war themed music after the generation task ( $M = .06, SE = .03$ ).

Further analyses revealed no main effects or interactions on the tendency for participants to include the Baby MusicAssociates, War MusicConcepts, or War MusicAssociates in their stories,  $p > .05$ . This suggests the effects of thematic music on incorporating thematic elements into the story is more complicated than simply being the effect of whether or not thematic music is heard prior to the generation task. It appears as if incorporating thematic concepts into a story depends on the types of concepts activated by the music AND the context of the story generation task. The disappearance of the effect of War themed music on the incorporation of War MusicAssociates suggests that those War MusicAssociates are more compatible with people’s understanding of the types of adventures that are possible on an alien planet than they are with adventures to an unexplored, yet inhibited foreign land.

To statistically examine the effects of story scenario on the tendency to include Baby or War themed elements in a

story generation task, we conducted post-hoc analyses comparing Experiments 2 results to those in Experiment 1. We believe this is a valid comparison because the two experiments differed only in terms of the theme participants were asked to write a story about. Also, participants were drawn from the same subject pool during the same year for both experiments.

Because we are interested in how the story theme interacts with the placement and the theme of the music to affect the tendency to incorporate Baby themed or War themed items into the story, we combined the Baby MusicConcepts and Baby MusicAssociates variables into a single variable (BabyAll) that measures the proportion of both Baby MusicConcepts and Baby MusicAssociates (12 items total) included in the story. The same was done for the War MusicConcepts and War MusicAssociates, resulting in WarAll (10 items total).

Table 2. Effects of Music Theme, Position and Story Theme on Proportion of Total Baby and War Items in Stories						
			Story Theme			
			Alien		Foreign	
BabyAll	Music	Position	M	SE	M	SE
	Baby	Before	.07	.01	.06	.01
		After	.07	.01	.03	.01
	War	Before	.07	.01	.04	.01
		After	.06	.01	.04	.02
WarAll	Baby	Before	.02	.01	.01	.01
		After	.02	.01	.02	.01
	War	Before	.04	.01	.01	.01
		After	.02	.00	.00	.00

A 2 X 2 X 2 ANOVA examined the effect of music theme (Baby, War), position (Before, After) and story theme (Alien, Foreign) on the inclusion of baby themed items and war themed items in the story generation task.

The analyses revealed a significant main effect of story theme on the inclusion of baby themed elements (BabyAll) in the story,  $F(1, 254) = 5.91, p < .05, \eta^2 = .02$ . Participants are more likely to include BabyAll items when writing a story with an alien theme ( $M = .07, SE = .01$ ) than when writing a story with a foreign land theme ( $M = .04, SE = .01$ ).

Analysis of the tendency to incorporate War themed elements (WarAll) in the generation task also revealed a significant main effect of story theme,  $F(1, 254) = 4.09, p < .05, \eta^2 = .02$ . Participants are more likely to include the WarAll items when writing a story with an alien theme ( $M = .02, SE = .00$ ) than when writing a story with a foreign land theme ( $M = .01, SE = .00$ ).

There was also a non-reliable Music Theme X Position interaction  $F(1, 254) = 2.87, p < .10, \eta^2 = .01$ . Participants are more likely to incorporate WarAll items into their stories when they hear war themed music before engaging in the generation task ( $M = .03, SE = .01$ ) compared to hearing it after the task ( $M = .01, SE = .01$ ) or hearing baby themed

music before ( $M = .02, SE = .01$ ) or after ( $M = .01, SE = .01$ ) the task.

## Discussion

The hypothesis that a story scenario set on Earth would allow the effect of listening to baby themed music on the inclusion of baby themed items in the generation task to be visible was supported. Participants who listened to baby themed music prior to engaging in a generation task, were more likely than those who listened to the music after writing their story to include items that pilot testing indicated were specifically activated by the baby themed music into their stories. The effect of listening to war themed music on performance in a generation task was weak enough that it was not visible in the Experiment 2 data. However, combining the Experiment 1 and 2 data provided enough power to demonstrate listening to war themed music rather than baby themed music before engaging in a generation task does increase the number of war themed components that are included in the generation task.

To explain the results of the Experiments 1 and 2, we would like to propose that the story scenario that participants were writing about interacts with the type of concept activated by the music to influence the degree to which participants incorporate thematic elements of the music into the story. For both story themes it was demonstrated that whether or not music concepts were incorporated into the story was influenced by the scenario for which they were writing the story with both baby themed music elements and war themed music elements being more likely to be incorporated into the novel stories when writing stories set on an alien planet. Interestingly, the reason why each set of elements was more likely to be incorporated differs between the two scenarios and themes.

Demonstrating an effect of listening to baby themed music on the tendency to include baby themed items into the story in Experiment 2 but not in Experiment 1 is probably due to participants feeling a need to rationalize the visit to the alien planet in Experiment 1 causing them to rely on the plot device of falling asleep and dreaming that they visited an alien land. This was evident in the fairly high rates of baby themed items across all four conditions in Experiment 1. Because sleep, dreams, and bed were all “tucking a baby into bed” themed items, the plot device washed out any effects that hearing the baby themed music might have had on the stories participants generated. The plot device was not necessary for stories about an adventure in an undiscovered foreign land so only the participants exposed to baby themed music prior to the generation task included baby themed elements into their story.

In contrast, the greater tendency to include war themed elements in the stories in Experiment 1 compared to Experiment 2 probably was due to the concept of war being more congruent or plausible with a visit to an alien planet than a visit to a foreign land. This suggests participants’ knowledge of alien planets and foreign lands is influencing

which thematic elements activated by the music get incorporated into the story or whether they are incorporated at all.

The proposal that the thematic elements activated by listening to thematic music interacts with the knowledge of the story scenario to affect performance in a generation task is consistent with some of the previous work on conceptual expansion and concept activation. Sifonis (1995) (as discussed in Ward, Smith & Finke, 1999) demonstrated that participants asked to describe a restaurant for a race of bird-like aliens integrate their category knowledge of birds with their schema of restaurants to generate a novel exemplar of “restaurants for bird-like aliens.” These novel restaurants retained some of the salient features of restaurants on Earth (e.g., having tables and serving food). However, in the process of replacing the human customer in the restaurant schema with birds in the context of an alien planet, the novel restaurants incorporated features of all three types of knowledge: restaurants, birds, alien planet (i.e., the customers sat on perches at tables and were served worm burgers but didn’t have to pay for their meals because the civilization had advanced beyond the need to exchange money for services). Perhaps the themes activated by the music are, in fact, activating a complex schema rather than just the individual concepts and their associates that participants reported during pilot testing. If this is the case, then the effects of the music on story generation might be more evident when looking for schematic elements associated with the theme (e.g., baby theme activating schema elements associated with children and parents and school).

Another explanation for a process underlying performance in the story generation task is that the specific scenario participants asked to generate stories for activates a specific form of the thematic element. Solomon and Barsalou (2001) have argued that properties associated with a concept are represented differently in different concepts. For example, the property of “red” is represented differently for “hair” and “wine” and “blood.” The dominant representation of the property at the time of retrieval varies on the basis of context as well as other factors. Consequently, the specific conceptual properties that are activated by the thematic music and incorporated into the generation task might depend on the story context/scenario of the generation task. If this is the case, then coding the stories for the features associated with the specific combination of music theme and story scenario (e.g., features of an “undiscovered foreign land in which there are a lot of babies”), should increase the ability to observe the effects of the thematic music on performance in the generation task.

Additional directions for future experiments include finding music that more strongly activates a particular theme. Using familiar rather than unfamiliar music should increase the effect of that music on the performance in the generation task over the low levels observed in the current study. Alternately, perhaps music associated with a

particular geographical region (e.g., Arabia, China) would be more likely to result in those features being included in a story about a visit to a distant land than music with a theme that is not geographical in nature.

In summary, the research described in this paper provides preliminary data suggesting that music can be associated with concepts and that listening to the music can activate those concepts and affect performance in a generation task. Because of the pervasiveness of music in our lives and its use to help enhance our performance, moods or thoughts, developing a greater understanding of the influence music has on those thoughts and behaviors will allow us to increase the manner in which music benefits us. It will also increase our understanding of how music is represented in memory and how that representation is manifested in performance,

## Acknowledgments

Special thanks to the students of my Advanced Experimental Design class, especially Stephanie Camarata, Angela Hasman and Caitlin Kleist for developing the initial ideas that resulted in this research. I would also like to thank Scott Niewinski, Emily Olthoff and Candice Lambert for data collection and coding. Reviewer number two also receives my thanks for their insightful and useful comments that helped me improve this abstract.

## References

- Bolt, M.G. (2001). Musical soundtracks as a schematic influence on the cognition processing of filmed events, *Music Perception*, 18, 427–54. doi:10.1006
- North, A. C., Hargreaves, D. J., and McKendrick, J. (1997). In-store music affects product choice. *Nature*, 390, 132. doi: 10.1038/36484
- Daltrozzo, J. & Schön, D. (2009). Conceptual processing in music as revealed by N400 effects on words and musical targets. *Journal of Cognitive Neuroscience*, 21, 1882–1892. doi: 10.1162
- Kastner, M. P. & Crowder, R.G. (1990). Perception of the major/minor distinction: Emotional connotation in young children, *Music Perception*, 8, 189–201.
- Marsh, R.L., Bink, M.L., & Hicks, J.L. (1999). Conceptual priming in a generative problem-solving task. *Memory & Cognition*, 27, 355–363.
- Orgs, G., Lang, K., Dombrowski, J., & Heil, M. (2008). N400-effects to task-irrelevant environmental sounds: Further evidence for obligatory conceptual processing. *Neuroscience Letters*, 436, 133–137. doi:10.1016
- Solomon, K.O. & Barsalou, L. W. (2001). Representation properties locally. *Cognitive Psychology*, 43, 129–169.
- Ziv, N. & Goshen, M. (2006). The effect of ‘sad’ and ‘happy’ background music on the interpretation of a story in 5 to 6-year-old children. *British Journal of Music Education*, 23, 303–314. doi:10.1017/S0265051706007078
- Ward, T.B., Smith, S.M. & Finke, R.A. (1999). Creative cognition. In R.J. Sternberg (Ed.) *Handbook of Creativity*. New York: Cambridge University Press.

# Maps in the Head and Maps in the Hand

**Kenny Skagerlund (kenny.skagerlund@liu.se)**

Department of Behavioral Sciences, Linköping University  
SE-581 83 Linköping, Sweden

**David Kirsh (kirsh@ucsd.edu)**

Cognitive Science Department, University of California, San Diego  
La Jolla, CA 92093, USA

**Nils Dahlbäck (nils.dahlback@liu.se)**

Department of Computer and Information Science, Linköping University  
SE-581 83 Linköping, Sweden

## Abstract

Using the perspective of situated cognition we studied how people interact with a physical map to help them navigate through an unfamiliar environment. The study used a mixture of cognitive ethnography and traditional experimental methods. We found that the difference between high and low performing navigators showed up in the speed they completed their task and also in the way they use maps. High performers plan routes using a survey method whereas low performers use a route strategy. We suggest that when people are given a task that does not match their cognitive style they try to transform the task to better suit their cognitive abilities and cognitive style.

**Keywords:** Map use, navigation, wayfinding, situated cognition, spatial cognition.

## Introduction

Interest in human spatial cognition and navigational capacity has a long history, ranging from the pioneering work of Siegel & White (1975) to contemporary contributions by Montello (1998; 2005) and Hegarty et al. (2002; 2006). Spatial cognition is concerned with how people represent space and navigate through it. (Montello, 2005). In the “classical” view, knowledge, from the perceived environment, is represented as a *cognitive map* (Tolman, 1948, Galotti, 2008). Siegel & White (1975) distinguished three types of knowledge involved in forming and using cognitive maps: i) *landmark knowledge*, ii) *route knowledge*, and iii) *survey knowledge*. Landmark knowledge is information about the particular features at a location. Route knowledge is information about specific pathways for moving from one location to another; it may be coded procedurally or declaratively. Survey knowledge is metric information about the relative location and estimated distances between landmarks, the very thing captured in a standard map, showing the location of all paths and features in a Euclidean plane. All this work investigates the representational architecture of *internal* spatial representations, focusing on questions such as whether cognitive maps are map-like in nature or more like nodes in a graph representation.

Lawton (1994; 1996) found that people tend to report using one either an *orientation strategy* or a *route strategy* when navigating, but not both. Orientation strategies are cognitive processes that use survey knowledge, the umbrella term for world-centric relations. When a subject thinks in an allocentric reference frame using global attributes of a terrain such as cardinal directions, and Euclidean positioning of landmarks, they are using orientation or *survey strategies* (Prestopnik & Roskos-Ewoldsen, 2000) for wayfinding. Route strategies, by contrast, are based on an egocentric frame of reference, where paths are defined as those throughways available from where the subject is at the moment.

Research on spatial knowledge acquisition and navigation has mostly been confined to strict laboratory settings and virtual environments (e.g. Allen, Kirasic, Dobson, Long, & Beck, 1996; Montello, 1993), or to environments where the subject is led along a fixed route in an urban area (Kato & Takeuchi, 2003). In these studies maps have been used primarily as a diagnostic tool to reveal the subject’s internal representation. For instance, a subject might be asked to sketch the route she followed, marking down all the landmarks she can recall. (Liben, 2010). Little or no attention has been paid to the actual practices of subjects when they use maps to navigate.

A map, if properly used, is an artifact that extends a person’s *survey knowledge* (Montello, Hegarty and Richardson 2004). It behaves in the same way as an internal map except that it is external. Because we interact with internal and external representations differently, however, it is worth examining in detail the diverse ways that people interact with maps. Do all subjects rotate maps? When, why? How do they gesture? Do they point on the map and then to the world? How often do they glance at a map? When?

To study the practices of map use we videoed subjects using a map of UCSD campus as they found their way from a starting location to a goal location. In the analysis we divided our subjects into two groups – route-based navigators, and survey-based navigators – using the well-known measures developed by Lawton (1994). We report here on our findings and offer an explanation of the results

in line with the ideas of situated, distributed and embodied cognition.

Clark (2008) has coined a term, or principle, to label this type of interaction in which epistemic actions (Kirsh & Maglio, 1994) can be incorporated – *The Principle of Ecological Assembly* (PEA). This principle states that “*the canny cognizer tends to recruit, on the spot, whatever mix of problem-solving resources will yield an acceptable result with a minimum of effort*” (Clark, 2008, p.13). But, given the existence of individual differences in cognitive styles, it is not obvious that “a minimum of effort” means the same thing for all people. A specific instance of this is the difference between persons who have a preference for a route strategy or a survey strategy. Another aim of the present work is to study how map use might differ between people depending on their preferred navigation strategies.

Since we are bridging or trying to relate two different research traditions, we address our research questions using a combination of the experimental methods traditionally used in research on navigation and wayfinding, with cognitive ethnography used in research on situated and distributed cognition.

## Method

The study was undertaken on the University of California, San Diego campus. UCSD is sufficiently complex and covers a large enough area to be challenging for most navigators unfamiliar with the campus. It can also be considered representative of an urbanized area.

17 participants were recruited using Craigslist, which is an online ad-service where ads can be placed for a fee. The participants were between 20 and 58 years,  $Mean = 32.1$  ( $SD = 13.23$ ), 8 female, 9 male; they were unfamiliar with the UCSD campus. To eliminate vision as a factor in performance they had to have 20/20 vision – with or without corrective lenses or glasses.

Participants were asked to find their way from a starting point to goal location. Three different start-goal pairs were used. These pairs were chosen and evaluated during a pilot study, where they were determined to be equally hard. Criteria for hardness were the number of salient landmarks, the density of buildings throughout the area, length (air distance), visual access. By using start-goal pairs that overlapped and crossed through the campus center, the environmental features and vistas were as equivalent as possible, leading us to infer they were equally complex.

## Materials

The materials used in this study included the official visitor map of the UCSD campus, which was handed to the participants and used throughout the navigation task.

Several recording tools were used, including the handheld video camera – Canon Vixia HG21 – and a head-mounted video camera – ContourHD 1300 LED 1080p Headcam – that captured the behavior of the participants. The motion pattern of the participants was recorded via a GPS – Victory Corp. Columbus V-900 Multifunction GPS data logger.

The Santa Barbara Sense of Direction Scale (SBSOD) was used to measure the participants’ sense of orientation, or the awareness of location or orientation. This instrument is a self-report measure which has been found to predict objective measures of these abilities, such as dead reckoning (Hegarty et al., 2002). This instrument has proved to be internally consistent and has sufficient test-retest reliability. The SBSOD is highly correlated with measures of spatial knowledge acquired from direct experience in the environment, and Hegarty et al. (2002) has shown that it is related to knowledge that involve orienting oneself within the environment.

The Wayfinding Strategy Scale (Lawton 1994) is a survey that measures to what extent a person depends on strategies relying on survey knowledge or route knowledge respectively. The survey contains 14 items of the sort of propositions that participants have to grade the degree of agreement along a 5-point Likert-type scale.

To measure dead reckoning a pointing task was used. At a number of places the participants were asked to point in the direction of an unseen landmark; a traditional compass was used to assess the participants’ error in this task.

## Procedure

The study itself was divided into two separate sessions. The first session was a pretest, where participants filled out electronic counterparts to the physical instances of SBSOD and the wayfinding strategy scale over the internet. The surveys could be completed at any time the participants wished from the moment of agreement of participation in the study to the day when the experiment session began. The questionnaire was filled out prior to the experiment session, which was of vital importance as to ensure validity. If it would have been completed after the experiment trial, there would have been a possible risk that participants took into account their recent navigational performance, and thus affecting the self-assessment.

On arriving for the second session, a consent form was filled out by each of the participants, a parking permit was paid for and given to them if needed, and they were then told to step into a car for transportation to another location. From this moment on, the experimental session had officially started and they were instructed to try to pay attention to where they are located in the world from that point onwards. The participant was dropped off at one of the marked drop-off locations where they were picked up by another experimenter. On site of the drop-off point the equipment was set-up, which included mounting the headcam on the participant and getting a stable GPS signal. The participant was then told to estimate and point into the direction of the meeting point, the experimenter then used the compass to derive the correct azimuth which was then communicated to the participant.

After having been given the correct direction, the participant was led non-linearly to the actual starting point of the navigation task. The starting point was located approximately 100 meters away, occluded from the drop-off

point. Another dead reckoning task was performed, where the participants were told to estimate and point into the direction of the *drop-off point* this time, after which it was time to initiate the primary task of the experiment session – the navigation task. At this point, the participants were given the campus map and told which building they were standing next to at the starting point. The participants were given time to find the building on the map, after which they were then told what destination they would be finding their way to. In similar manner, they were given time to locate the building on the map. Now, they were told to navigate to the destination by foot preferably using the shortest path. After any contingent questions and uncertainties had been mitigated, and after they had been instructed to try to verbalize their thoughts navigational strategies out loud, they were given the signal that it was OK for them to begin the navigation trial.

Throughout the navigation trial, the experimenters where filming the participants with the handheld camera while at the same time interviewing them according to a stipulated script, while given the freedom to *ad lib* when interesting observations were made. During the navigation task, the participant where given two instances of the dead reckoning-task where the azimuths were jotted down by the experimenter. Finally, when the participants reached their destination, they performed one last dead reckoning-task which concluded the navigation trial.

### Coding procedure

Three experimenters were coding the video material, and although no formal kappa value was calculated to establish the inter-rater reliability, the experimenters were trained simultaneously and looked at each other's code at the outset and very beginning of the coding to establish a consensus. The coders also consulted each other whenever any phenomenon raised any doubts concerning how to code it.

The coding scheme included a time stamp for each observation, a high level transcription of the think aloud

verbalizations, as well as gestures and other bodily actions such as body turns and visual references of the environment. Of specific interest was how participants interacted with the map. Thus, physical actions and manipulations of the map were pertinent to incorporate and code for in the coding schema, such as map rotations and folding of the map, in conjunction non-physical interactions with the map (e.g. coding for glances on the map), in order to reveal regularities of map use with respect to preferences of map interaction. In addition, gestures such as pointing on the map, or putting a thumb on the current location on the map, or running a finger across the map was coded for as well.

## Results

In the first part we will report on the quantitative measures used, to set the ground for the second part where we will present detailed observations of the use of the map as well as other orientation strategies used by our participants.

### Quantitative results

The results from the measures on The Wayfinding Strategy Scale, in the table called Orientation score, SBSOD score, dead reckoning error, number of map alignments and map consultation frequency, as measured by number of glances per hour, are presented in table 1 Note that the Orientation score test was introduced after the first 6 subjects participation.

All these three measures show considerable variation, SBOD from 34 to 94, navigation time from 14 to 58 minutes, and map consultation from 77 to 292 glances per hour, i.e. a ratio of approximately 4:1; the number of map alignments show an even higher variation, from 0 to 10.

There is a clear dependency between these measures. Dead reckoning error is negatively related to sense of direction as measured by SBSOD ( $r = -.43$ ,  $p < .05$ ), similar to results by Hegarty et al (2002). Also, as predicted, there

Table 1: Overview of results on performance measures and variables

Subject	Gender	Orientation Score	SBSOD-Score	# Map Alignments	Navigation Time	Dead Reck. Err.	Tot glances	Glances/hour
Subject 1	F	N/A	53	3	28	69	40	85,7
Subject 2	M	N/A	37	4	15	11	36	144
Subject 3	M	N/A	94	8	48	52,33	73	91,3
Subject 4	M	N/A	23	4 N/A		131	79	90
Subject 5	F	N/A	70	3	13	56	21	96,9
Subject 6	F	N/A	70	8	15	21	65	260
Subject 7	F	24	34	10	51	69,33	162	190,6
Subject 8	M	24	74	8	24	49,33	55	137,5
Subject 10	M	28	50	1	22	63	117	292,5
Subject 11	F	20	86	10	58	77,5	190	196,6
Subject 15	F	22	60	5	15	26	75	300
Subject 17	F	22	64	9	25	34,25	98	235
Subject 21	F	29	55	5	24	15	103	257,5
Subject 23	M	25	82	3	15	17,25	31	124
Subject 24	M	31	92	0	14	12,5	18	77,1
Subject 25	M	32	61	4	21	57,66	47	134,3
Subject 26	M	26	74	0	14	12	61	261,4



was a marginally significant negative correlation between SBSOD and map consultation frequency ( $r = -.52, p < .051$ ).

To investigate the existence of any potential difference between people who claim to use orientation strategies in contrast to people who primarily rely on route knowledge when engaged in navigation, a median split was performed on the sample on the orientation score, creating two groups here named. “Orienters” ( $n = 5, \bar{x} = 29.2, SD = 2.39$ ) and “Non-Orienters” ( $n = 5, \bar{x} = 22.4, SD = 1.67$ ). The difference between the two groups was significant  $t(8) = 5.22, p < .001$ .

It was noted above that the frequency of map use varied considerably between the participants, and that this was correlated with sense of direction. The question is then if the difference in map use is just a quantitative difference, or if there also is a qualitative difference in *how* the maps are used. One of the most obvious differences in how subjects use maps turns on alignment and registration: whether subjects prefer to leave the map in its native upright position determined by the orientation of labels, or do they rotate it so that the features on the map align with what they see. Using this criterion there is a clear difference in map use between the two groups. Map alignment frequency is significantly lower for Orienters ( $\bar{x} = 2$ ) than Non-Orienters ( $\bar{x} = 8.4$ ) ( $t(8) = 3.15, p < .01$ ).

To get a deeper understanding of map use differences we turn to a detailed analysis of the way subjects interact with maps and the various orientation strategies used.

## Navigation strategies

Our participants all make use of Siegel and White’s (1975) three basic kinds of navigation information, survey, path and landmarks. But they do not do so in the same way. They use different cognitive strategies within each of these broad categories. When they extract information outside of their preferred mode they try to transform that information into their preferred form. To highlight these differences, we present excerpts illustrating the strategies used by high and low performers as measured by their orientation score. In the present study, we did not note any significant differences between high and low performers in the use of landmarks. But there are differences in the use of survey and path information.

### Survey information strategy

As many authors have suggested, maps can be seen as an external form of survey knowledge. Our argument is that external representations still have to be interpreted and often a map user will physically interact with a map to facilitate interpretation. High and low performers interact differently.

Take the case of participant S24. He was a top performer on all the quantitative measures, one of two subjects who *never* rotated or aligned his map throughout the entire navigation task. S26 was the other. He kept his map in its canonical label reading position, that is fixed in a north-up position. Shortly after he began the task he was asked whether he has a particular strategy in mind. He answers:

*“I think so, I mean the idea is that I’m just gonna go right down here [pointing and tracing downwards along a depicted walkway on the map] and probably take a left on Voigt [traces with his finger to the right along Voigt Dr] and go south on Hopkins Lane (...)”*

In the video we see that S24 leaves the map in a north-up position and slightly tilts it for the experimenter to see the map. S24 runs his finger quickly down the map – that is, in a southerly direction – and then hastily makes a perpendicular turn with his finger and traces rightwards on the map while saying “*and probably take a left on Voigt*”. When he runs his finger southward on the map while simultaneously claiming that he will go “*down here*” his motion is in the same direction as he is. The map is egocentrically aligned. But when he runs his finger to the right (east) while saying “left”, however, he is breaking the egocentric view. Arguably S24 is making an inference based on an imagined egocentric perspective. He *imagines* himself, or rather *projects* (Kirsh, 2008) his future location ‘down there’ onto the external representation – the map – and quickly translates between egocentric and allocentric perspectives. What is striking about this particular incident is the ease and speed with which he performs this multi-layered action.

By comparison one of the low performers, S11, uses the map in a very different way. She frequently stopped throughout the navigation task to look at the map and subsequently tried to align the map to correspond to the surrounding environment in order to extract and assimilate information of where to go next. In the following excerpt, she has a vague idea of *where* she is, but she is not sure of her bearings in terms of cardinal directions and exactly where she should go.

*“...I’m trying to find that way. [pointing on the map] I’m gonna look at it upside down so I can see where I...then I know ‘cause we were on Voigt [Drive] before...”*

Her ambition is to walk south on Voigt, but she is uncertain about her bearings. She previously saw a sign for Voigt Drive and has a rough sense of her self-location. She then rotates the map to align it with her view in her current position. In contrast to S24, who showed an impressive management of directionality in the map and world, presumably through internal computation of the relation between world-centric and egocentric information, S11 is unable (or unwilling) to make these internal transformations and instead rotates the map, adapting the map artifact to fit her internal representations. This allows her to deduce whether she should go left or right on Voigt Drive, i.e. thinking in terms of a path strategy, instead of thinking about the world in survey knowledge manner.

### Route information strategy

Low-performers preferred to travel on straight roads and paths. Curving paths make it harder to keep track of one’s cardinal direction. They also avoided travelling on paths with high visual complexity – such as dense buildings. High visual complexity makes it harder to identify one’s

preferred landmarks on a map; the more buildings the more visual distractors. Again, high-performers had no such aversions. They are sufficiently proficient in using cardinal directions that they do not hesitate to use curved paths, or take opportunistic short-cuts that twist, even if this means going through dense buildings.

Subject 11 explained why she preferred to walk along a path that clearly diverted from the direction of the goal destination and optimal solution:

*"I could have weaved through [the buildings] but I think it would have taken me much longer (...) this is a straight way, so it's good, I can avoid going through all the cluster of buildings, it's less complex (...) I get lost very quickly, to me it's a lot easier to navigate than having to go through and around buildings..."*

S24 has the opposite attitude. In this excerpt, he is walking on a pre-planned path eastbound but suddenly decides to take another path:

*"I'm gonna head up this path, and I didn't go that way [pointing on the map on the original path he was supposed to take] because it was directly east, and this path kinda branches southeast, which is the direction we're going towards, so I figure I'd take that one."*

In contrast to S11, S24 never hesitated to take a diagonal path, as long as it was a shorter path. He seemed indifferent to visual complexity, taking routes that required navigating through buildings located at the very center of UCSD campus where several large commercial stores and buildings are located. The whole time he managed to stay oriented and solved the navigation task with the least glances to the map and nearly the least travel time. It seems that high performers can use the world better. They can find paths using cardinal directions as a grid, while low performers instead develop and use paths that will minimize the need to orient using cardinal directions.

### Transforming the information to suit the style

One way of describing the differences between high and low orienteers is that high orienteers think mostly in a world-centric manner whereas low orienteers think in a map-centric manner. For a high orienteer navigation involves keeping track of one's bearing. Because they have much better dead reckoning skills they always have a reasonable idea of where they are and where they have to go. When they look at a map it may be to see what is coming up next, but it is more likely to verify that they are where they think they are. They want to update their location relative to where they must go. They can do this by finding their location on the map in the orientation they have been holding it. They have no problem tracking themselves going south on a map facing north. They sense they are going south. They get too little cognitive saving from reorienting it to pay the price of the harder readability that comes from inverting labels.

Low orienteers, however, do their reasoning on the map. They orient the map so that it is aligned with their currently perceived view of the world. They put it in correspondence with the features that are in view so that they can then trace

where they must go on the map to reach the destination. Since the two, map and world are in alignment, they can then take their bearing straight off the map. They need to walk in the same direction in the world as the map. The two can be laid on top of each other. Straight paths are to be preferred because each revision of direction requires re-checking the map and that is a cognitive effort. The cognitive cost exceeds the physical cost of walking farther. Similarly, it is easier to avoid high feature areas, where buildings are close together, because the more visual clutter there is the harder it is to align the map since it requires checking more buildings and more angles.

### Strategy choice and performance

It may seem that low orienteers must be slower than good orienteers. Whereas in general this is true it is by means necessary. An efficient map-user can keep the map reasonably aligned by taking long straight paths. This saves them the cognitive effort of re-aligning the map, and it saves time too because realignment can be time-consuming. It is quite possible that the map-using time saved by taking a straight path more than compensates for longer distance.

An interesting case is S15, who has a very low score on the sense of direction test, but also has one of the shortest navigation times (15 min). S15 seems very aware of her preferences, as illustrated in the excerpt below.

*"I'm gonna go straight [pointing with her arm and hand in forward direction] and then straight [pointing with the same arm in an orthogonal direction to the left] (...) I don't like diagonals."*

When we look at S15's details we see that she glanced at the map 300 times, the most of any participant. But her glances were brief because she stayed on a map-aligned course. Given the straight line route she chose this meant that she could rotate the map just a few times (5).

If we assume that S15's performance is solid evidence that she is a good navigator then it does not follow that navigation ability always correlates with spatial ability. In previous work on navigation (Hegarty et al, 2006), navigation ability was found to correlate well with measures of spatial ability and with self-assessment of sense of direction. But the case of S15 illustrates that while this may be true in general there are outliers for whom it is not true. We hypothesize that people who know their own strengths and weaknesses in spatial understanding develop interactive strategies that compensate. They develop techniques for coordinating map use with route features to minimize time and cognitive effort.

### Discussion

The results presented here show that how maps are used differs between different people, depending on their navigation abilities. Navigators with a high orientation score, keep the map in the same upright position regardless of how well this matches their current view of the environment. They have no problem in mapping the view of the map to their current view of the environment using internal or mental transformations. Navigators with a low

orientation score, on the other hand, find mental transformations effortful, and instead prefer to externally manipulate the map to align it with their current view. These two also seem to differ in how they plan their route through the environment. Navigators with a low sense of direction prefer to walk along straight lines and in the open. Even though their paths are not the most direct, in effect requiring them to take detours, they still prefer them as long as they make the navigation task simpler. Navigators with a high sense of directions invariably prefer the shortest path even when that involves cutting across visually complex areas, following paths that wind. They even take opportunistic short-cuts on narrow pathways whenever they can.

Another way of looking at the difference between these two groups is that good navigators, comfortable with using a survey strategy have no problem using the terrain information in allocentric form, i.e. as a map. The other group is more comfortable with a route strategy, they tend not to use survey strategies and not surprisingly they prefer terrain information that is presented in egocentric form. They do this both by initially planning a simple route with no curves, and they manipulate the map to be able to read off bearing directly from the map without having to perform transformations from cardinal to egocentric direction.

The disparities between the two groups suggest that they confront their navigation task operate in different ways. Survey strategists have good sense of direction. They maintain a strong sense of where they came from and where they have to go. A map for them is a tool to help see the future but they consult it primarily to get confirmation that their sense of bearing is correct. The map is more for feedback than pure planning. Route strategists rely much more heavily on maps. They plan every step of their route on the map, and they make point wise decisions about where they are and where they must go next by orienting the map. It seems that route strategists do as much computation on the map as possible, whereas survey strategists do as much computation on their internal representation of the world.

## References

- Allen, G. L., Kirasic, K. C., Dobson, S. H., Long, R. G., & Beck, S. (1996) Predicting environmental learning from spatial abilities: An indirect route. *Intelligence*, 22, 327-355.
- Clark, A. (2008) *Supersizing the Mind: Embodiment, Action and Cognitive Extension*. Oxford University Press.
- Darken, R. P., & Sibert, J. L. (1996) Wayfinding Strategies and Behaviors in Large Virtual Worlds. *The International Journal of Human-Computer Interaction*, 8(1), 49-72.
- Galotti, K. (2008) *Cognitive Psychology: in and out of the laboratory (4<sup>th</sup> edition)*, Wadsworth Publishing.
- Gärbling, T., Lindberg, E., Carreiras, M., & Böök, A. (1986) Reference systems in cognitive maps. *Journal of Environmental Psychology*, 6(1), 1-18.
- Hegarty, M. Richardson, A. E., Montello, D. R., Lovelace, K & Subbiah, I. (2002) Development of a Self-Report Measure of Environmental Spatial Ability. *Intelligence*, 30, 425-447.
- Hegarty, M., Montello, D. R., Richardson, A. E., Ishikawa, T. and Lovelace, K. (2006) Spatial Abilities at Different Scales: Individual Differences in Aptitude-Test Performance and Spatial-Layout Learning. *Intelligence*, 34, 151-176.
- Hutchins, E. (1995) How a Cockpit Remembers Its Speeds. *Cognitive Science*, 19(3), 265-288.
- Kato, Y., & Takeuchi, Y. (2003) Individual differences in wayfinding strategies. *Journal of Environmental Psychology*, 23, 171-188.
- Kirsh, D. (1996). Adapting the Environment Instead of Oneself. *Adaptive Behavior*, 4(3-4), 415-452.
- Kirsh, D. (2009) Projection, Problem Space and Anchoring. In N.A. Taatge & H. van Rijn (Eds.), *Proceedings of the 31<sup>st</sup> Annual Conference of the Cognitive Science Society* (pp. 2310-2315). Austin, TX: Cognitive Science Society.
- Kirsh, D., & Maglio, P., (1994) On Distinguishing Epistemic from Pragmatic actions. *Cognitive Science*, 18(4), 513-549.
- Lawton, C.A. (1994) Gender Differences in Way-Finding Strategies: Relationship to Spatial Ability and Spatial Anxiety. *Sex Roles*, 30 11/12, 765-779.
- Lawton, C.A. (1996) Strategies for indoor wayfinding: The role of orientation. *Journal of Environmental Psychology*, 16, 137-145.
- Liben, L.S. (2010) Identifying Locations and Direction on Field and Representational Mapping Tasks: Predictors of Success. *Spatial Cognition & Computation*, 10, 105-134.
- Montello, D.R. (1998) A new framework for understanding the acquisition of spatial knowledge in large-scale environments. In M.J. Egenhofer & R.G. Golledge (Eds.), *Spatial and temporal reasoning in geographic information systems* (143-154). New York: OUP.
- Montello, D.R., Hegarty, M., Richardson A.E. (2004) Spatial Memory of Real Environments, Virtual Environments, and Maps. In: Allen, G.L (Ed.), *Human spatial memory: Remembering where*, 251-285. Mahwah, NJ: Lawrence Erlbaum Associates.
- Montello, D.R. (2005) Navigation. In Miyake, A and Shah, P (Eds.) *The Cambridge Handbook of Visuospatial Thinking* (Cambridge, Cambridge University Press), pp. 257-294.
- Prestopnik, J.L., & Roskos-Ewoldsen, B. (2000) The relations among wayfinding strategy use, sense of direction, sex, familiarity, and wayfinding ability. *Journal of Environmental Psychology*, 20, 177-191.
- Rupert, R. (2010) *Cognitive Systems and The Extended Mind*. Oxford University Press.
- Siegel, A.W., & White, S.H. (1975) The development of spatial representations of large-scale environments. In H. W. Reese (Ed.), *Advances in child development and behavior* (Vol. 10, 9-55). New York: Academic Press.
- Tolman, E.C. (1948) Cognitive maps in rats and men. *Psychological Review*, 55, 189-208.

# When Students Don't Benefit From Attention Guidance in Animations: The Role of Working Memory in Learning From Animations

Irene T. Skuballa (skuballa@psychologie.uni-freiburg.de)

Rolf Schwonke (schwonke@psychologie.uni-freiburg.de)

Alexander Renkl (renkl@psychologie.uni-freiburg.de)

University of Freiburg,  
Department of Educational and Developmental Psychology, Engelbergerstr. 41,  
79085 Freiburg, Germany

## Abstract

The present study examined how students' working memory capacity influences learning from animations with or without guidance. We tested three different conditions: visual guidance, instructional guidance, and no guidance. The results show that especially visual guidance was perceived as being helpful for making references between narration and display of an animation. However, students without guidance outperformed both groups of students with guidance on a domain-specific knowledge test. A significant interaction between type of guidance and working memory capacity revealed that visual guidance impeded learning in students with high working memory capacity, whereas instructional guidance impeded learning in students with low working memory capacity. Our results suggest that working memory capacity is an important learner variable that should be taken into account to understand intervention effects and to customize learning environments to learners' needs.

**Keywords:** learning; working memory capacity; animation.

## Introduction

Animations can make unseen movements, interrelationships, and interdependencies or "difficult-to-see" particles and components in a system visible and thus accessible to comprehension. Animation can be defined as "a pictorial display that changes its structure or other properties over time and which triggers the perception of a continuous change" (Schnotz & Lowe, 2008, p. 304). This definition also pertains to dynamic visualizations, for example, presentations of how a technical device works or how a complex object is assembled. After a long line of research, nowadays there is no doubt that well-designed animations are helpful tools for fostering learning and transfer in different domains (Höffler & Leutner, 2007; Linn, Chang, Chiu, Zhang, & McElhaney, 2011).

Disadvantages of animations in the context of learning are grounded in their transitory and simultaneous nature. First, the presentation of entities in an animation is time-limited and subject to transience. This can hamper processing of important pieces of information, especially when the learner has not paid immediate attention to the relevant animated parts. Second, the simultaneity that characterizes one of animations' advantages for learning is potentially also a pitfall. When a series of events takes place at the same time, learners' limited capacities may be overwhelmed. Hence,

meaningful learning that requires learners to actively select and organize relevant information in order to integrate it into existing schemata in long-term memory can be impeded.

Motivated by possible disadvantages of animations, design factors have been proposed that aim at guiding learners' visual attention (Ayres & Paas, 2007). We tested two promising ways of fostering attention guidance to relevant information in animations, namely instructional guidance by giving verbal instruction prior to the presentation of the animation and visual guidance by blurring out irrelevant information in the animation. In addition, we investigated the influence of working memory (WM) capacity on learning from animations with these two types of attention guidance.

## Guidance in animations

### Instruction

Providing instructions on how to select and integrate information that is presented in different modes can have a positive effect on learners' attention allocation and, thus, on learning processes. Instructions can be given before rather than during the presentation of a certain learning environment in order to avoid interference with the display of the learning contents during the actual learning phase. On the other hand, processes of recalling and maintaining the instructions during learning may "bind" WM capacities.

In the context of multiple external representations, instructing learners on the functional relationships between representations can foster learning outcomes by guiding visual attention (Schwonke, Berthold, & Renkl, 2009). Gopher, Weil, and Siegel (1989) argue that mere prolonged exposure to a complex and dynamic task does not necessarily improve a learner's performance. Instead, a complex task should be decomposed into subcomponents, and the focus of attention should be changed according to these predefined subcomponents. Computer game players who received instructions to focus on single sub-tasks—for example, first ship control and then mine handling—outperformed players without any instructions to change their focus (Gopher et al., 1989). When following these instructions, "by a systematic manipulation of emphasis on different task subelements, subjects were led to explore a

wider range of attention strategies and improved their ability to cope with the high load of tasks” (Gopher, Weil, & Bareket, 1994, p. 389). Moreover, trainee pilots who first adopted strategies in attention allocation according to the emphasis change method in a computer game were better at actual piloting of an airplane than trainees who did not (Gopher et al., 1994). The emphasis change method works by way of external instructions prior to the learning phase. It is based on change of focus on components of a complex task and feedback (Gopher et al., 1989). However, it remains open whether such a method fosters only sensomotoric skills or also (meaningful) learning from animations.

### **Cueing**

Cueing and signaling to highlight key information offer a more apparent and invasive way of attention guidance (Ayres & Paas, 2007). In general, cueing “refers to the addition of design elements that direct the learner’s attention to important aspects of the learning material” (Plass, Homer, & Hayward, 2009, p. 39). Learners are not required to remember prior instructions. Cueing can be achieved by adding attention-directing objects such as arrows, circles, or colors to make relevant parts more noticeable. Another possibility is to reduce the luminance or the clarity of irrelevant parts in the visual display so that the important parts attract attention (“spotlight display”; Jarodzka et al., 2010). This technique makes animations less complex by directing learners’ attention to relevant information, thereby reducing the search space and freeing capacities for meaningful learning (De Koning, Tabbers, Rikers, & Paas, 2010; Mautone & Mayer, 2001). Whereas arrows and colors run the risk of delivering too much new information, a change of luminance or acuity creates a spotlight and may be perceived as less distractive. The latter methods minimize the visual display to the most important parts and events while preserving a holistic view. The advantages of cueing should fit especially the needs of learners with low WM capacity and those who are easily distracted by simultaneity.

In a study on learning a perceptual task (i.e., diagnosing seizures in infants), cueing was used to guide the learners’ visual attention in a tutorial video. A spotlight display was superior to a circle display that was supposed to direct attention and to a control condition without visual guidance (Jarodzka et al., 2010). In line with these results, De Koning, Tabbers, Rikers, and Paas (2007) reported encouraging findings on the superiority of cued animations over non-cued versions in terms of comprehension and transfer performance. Unfortunately, these results could not be replicated. Cueing in an animation on the cardiovascular system did not lead to better learning outcomes than non-cueing, although eye tracking data revealed that learners’ visual attention was guided more frequently to cued than non-cued contents (De Koning et al., 2010). In sum, it is unclear why learners do not always benefit from cueing, even when their attention was successfully directed to the

relevant regions in the animations. Considering learner variables such as WM capacity may help in providing adequate answers to this open question.

### **Working Memory in Multimedia Learning**

Learning and comprehension are dependent on learners’ ability to allocate and regulate attention. Before information can be stored in long-term memory it has to be processed in WM (Baddeley, 2003). Given the limited capacity of WM, only a small amount of the perceived information can be actively processed in order to acquire knowledge in the form of schemas. In their review, Schüler, Scheiter, and van Genuchten (2011) showed that WM capacity is a stable construct that affects the processing of static multimedia learning material such as texts and graphics. Because of its constraints, capacity likely plays a prominent role in learning from animations which can put high demands on learners.

Animations are often complemented by narrations. Consequently, in addition to information presented in visual mode (i.e. display) learners have to integrate information presented in auditory mode (i.e., narration). This leads to simultaneous demands on different components of WM. Auditory information is processed in WM’s phonological loop, while visual information from the animated visual display is processed in WM’s visuo-spatial sketchpad. Both types of information have to be temporarily stored and integrated in the episodic buffer (Baddeley, 2003).

WM measures reflect a domain-free ability to hold and process several information chunks “actively” while ignoring irrelevant information through the control of attention. This ability varies between individuals and influences the task performance. Hence, higher WM capacity facilitates not only processing of multiple information but also suppression of distracting information. Individuals with high WM capacity outperformed individuals with low WM capacity on visual selective attention tasks thanks to their flexibility in allocating their attention to visual stimuli (Bleckley, Durso, Crutchfield, Engle, & Khanna, 2003). The same results apply to the auditory channel. In a replication of the cocktail party phenomenon, individuals with low WM capacities detected their name in an irrelevant message more often than individuals with high WM capacities did, indicating that low WM individuals are more susceptible to distraction; at the same time, high WM individuals outperformed low WM individuals on a shadowing task (Conway, Cowan, & Bunting, 2001).

The results demonstrate that WM capacity is of vital significance in the process of attention allocation to visual and auditory information. Thus, WM capacity should have an influence on learning from animations. Furthermore, it is reasonable to assume that learners with different capacities may require different types of attention guidance for successful learning. By implication, learners with different levels of WM span may react differently to the same

instructional design, such as verbal instruction or visual guidance.

## Hypotheses

In our approach, we tested the effects of two types of guidance in an animation depicting the processes within a technical device (i.e., parabolic trough power plant system). We tested three conditions: a visual guidance group, who watched an animation with a clear spotlight on the relevant information while the visual clarity of irrelevant parts was reduced; an instructional guidance group, who received an instruction prior to the animation to make references between the narration and the visual display; and a no-guidance group, who did not receive any guidance on how to process the animation. We tested the following hypotheses:

- (1) As part of a manipulation check, we expected that subjectively perceived difficulty to make references between narration and visual display would be higher in the no-guidance group than in the groups with guidance.
- (2) With regard to the learning outcome, we expected the guidance groups to outperform the no-guidance group on a domain-specific posttest.
- (3) Besides a general effect (“main effect”) of WM capacity on the learning outcomes, we assumed an interaction between guidance and WM capacity. Participants with low WM capacity should benefit more from guidance than would participants with high WM capacity.

## Method

### Participants

The participants were  $N = 81$  (62 female) students from the University of Freiburg (age  $M = 22.14$ ,  $SD = 3.18$ ). Participants were randomly assigned to one of three conditions: visual guidance, instructional guidance, or no guidance. Each condition comprised of 27 participants.

### Materials

**Prior knowledge test** A pretest on prior knowledge of solar energy and the parabolic trough power plant system consisted of 30 items. Knowledge about technical devices entails being able to describe their structures, processes, and functions (Kalyuga & Hanham, 2011): Structures are the components an object consists of and their relationships; processes describe what happens in the system and how the device operates; functions characterize the purpose of the device and its sub-components and “provide” the answer to the question of what it is designed for. The prior knowledge test thus required participants to answer questions on solar energy in general but also on the structures, processes, and functions of the system (Cronbach’s  $\alpha = .79$ ).

**Learning performance** In our animation structures, processes, and functions were specified visually and

verbally by the visual presentation and narration. Hence, the posttest also comprised questions on the structures, processes and functions of the parabolic trough power plant. The learning outcome was assessed with 40 items. The overall reliability of the posttest was good (Cronbach’s  $\alpha = .90$ ).

**Test of WM capacity: Letter-Number Sequencing test (LNS)** WM span was assessed by the Letter-Number Sequencing test measuring especially the WM and attention span (adapted from the German version of the Wechsler Adult Intelligence Scale, WAIS-III; von Aster, Neubauer, & Horn, 2009). Participants listened to a sequence of letters and numbers (e.g., T-9-A-3) and reproduced them afterwards, but were asked to place the numbers in numerical order and the letters in alphabetical order, (e.g., 3-9-A-T). The level of complexity was defined by the number of elements, namely letters and numbers. The test started with two elements, and the level of complexity gradually increased by adding one element at a time up to a final sequence of eight elements. For each correctly announced sequence participants received one point. All points were summed up to a total score (between 0 and 21 points).

**Animation** The animation was colored and lasted about 5 minutes. It depicted how a parabolic trough power plant and its three cycles (i.e., oil cycle, water-steam cycle, and salt cycle) work. Each cycle is characterized by unique structures and serves a specific role in the conversion of solar energy to electric power. The solar radiation as well as the direction and flow of the different fluids in the system were animated.

Corresponding to our three conditions, we developed three versions of the animation. The visual guidance version included cueing by blurring out the cycles of the system that were not in the focus of the narration. In this way, a holistic view of the animation was preserved, while the relevant parts were made more salient by “spotlights.” The purpose of cueing was to visually guide participants’ attention through the animation and to assist them in making references between narration and the animated visual display. In the instructional guidance version, participants had to read an instruction prior to the animation on how to make references between the narration and the visual display. They were informed that several things would happen simultaneously and that it was thus crucial to follow the narration and map it to the animation. A third version involved no guidance at all, neither visual nor instructional.

### Procedure

Participants were tested in individual sessions approximately 60 minutes in length. After being explained the procedure, participants were asked to complete a short questionnaire on demographic data. They were then seated in front of a 22” computer monitor screen that was set at an operating distance of 60 to 80 cm. Next, prior knowledge was assessed and participants were asked to give subjective

evaluations of their knowledge about the system (ten-point Likert scale from 1 = no knowledge to 10 = very good knowledge). The assessment of WM capacity followed (Letter-Number Sequencing), after which participants watched the animation. They were then asked to rate their knowledge about the system once again (after watching the animation), and asked how difficult it was for them to map between the narration and the animated visual learning content, again on a ten-point Likert scale. Finally, learning performance was assessed by a domain-specific posttest.

## Results

First, we conducted a one-way ANOVA on the pretest. The conditions did not differ with respect to prior knowledge,  $F(2,78) = 0.00$ ,  $p = .998$ ,  $\eta^2 = .00$ . Before and after the animation, participants had to rate their knowledge about parabolic trough power plants. Participants in each condition showed a significant increase in their subjectively perceived knowledge about the system after having seen the animation (no guidance:  $M_{before} = 1.30$ ,  $SD_{before} = 0.87$ ;  $M_{after} = 6.12$ ,  $SD_{after} = 2.41$ ;  $t(25) = 9.81$ ,  $p < .001$ ,  $\eta^2 = .79$ ; instructional guidance:  $M_{before} = 1.26$ ,  $SD_{before} = 0.71$ ;  $M_{after} = 5.35$ ,  $SD_{after} = 2.45$ ;  $t(24) = 9.05$ ,  $p < .001$ ,  $\eta^2 = .77$ ; visual guidance:  $M_{before} = 1.44$ ,  $SD_{before} = 1.31$ ;  $M_{after} = 5.00$ ,  $SD_{after} = 2.74$ ;  $t(25) = 6.50$ ,  $p < .001$ ,  $\eta^2 = .62$ ). There were no differences between groups with respect to their perceived knowledge after watching the animation,  $F(2,74) = 1.73$ ,  $p = .184$ ,  $\eta^2 = .05$ .

As part of our manipulation check, participants were asked to rate how difficult it was to map between the visual display and the narration during the learning phase. Because the assumption of homogeneity of variance was violated, we conducted a Kruskal-Wallis test. The type of guidance (experimental condition) significantly affected the perceived difficulties in making references between visual display and narration,  $H(2)=7.57$ ,  $p = .021$ . Mann-Whitney tests were used to follow up this finding (Figure 1). There was no difference between the no-guidance group ( $M = 5.19$ ,  $SD = 2.47$ ) and instructional guidance group ( $M = 4.22$ ,  $SD = 1.93$ ;  $U = 282$ ,  $p = .149$ ,  $r = -.20$ ), but participants in the no-guidance group reported significantly more difficulties in making references than did participants in the visual guidance group ( $M = 3.48$ ,  $SD = 1.63$ ;  $U = 217$ ,  $p = .009$ ,  $r = -.35$ ).

Nor were there any differences in WM capacity (LNS) between the conditions,  $F(2,78) = 1.42$ ,  $p = .248$ ,  $\eta^2 = .035$ . Overall, WM capacity was positively correlated with the learning outcomes,  $r = .32$ ,  $p = .004$ . To test our next two hypotheses, we performed a general linear model in which we predicted learning outcomes by condition, WM capacity (as continuous variable), and the respective interaction term. Condition had a significant effect on learning outcomes,  $F(2,78) = 5.78$ ,  $p = .005$ ,  $\eta^2 = .133$ . Pairwise comparisons revealed a significant difference between the no-guidance group and the instructional group,  $p = .039$ , as well as between the no-guidance and the visual guidance group,  $p = .043$ . The no-guidance group (adjusted  $M = 26.20$ , 95% CI

[23.63, 28.78]) outperformed both the instructional guidance group (adjusted  $M = 21.54$ , 95% CI [18.95, 24.12]) and the visual guidance group (adjusted  $M = 21.48$ , 95% CI [18.77, 24.20]).

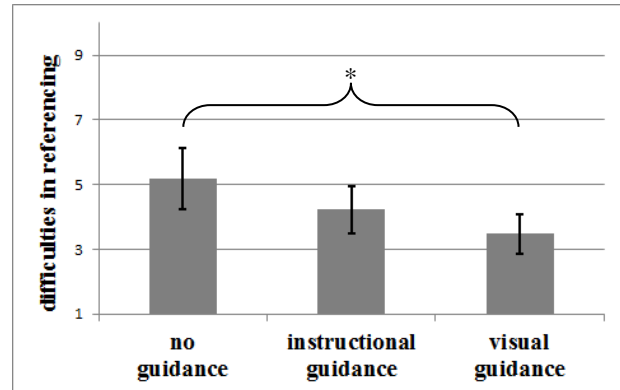


Figure 1: Perceived difficulties in referencing between visual display and narration (1 = none, 10 = many; 95% CI).

The effect of WM capacity on learning outcomes failed to reach statistical significance,  $F(2,78) = 3.73$ ,  $p = .057$ ,  $\eta^2 = .05$ . However, there was a significant interaction effect between WM capacity and type of guidance (experimental condition),  $F(2,78) = 5.24$ ,  $p = .007$ ,  $\eta^2 = .12$ . Students with low WM capacity were hindered by instructional guidance and students with high WM were hindered by visual guidance (Figure 2). Overall, no guidance was a good fit for learners with low as well as high WM span.

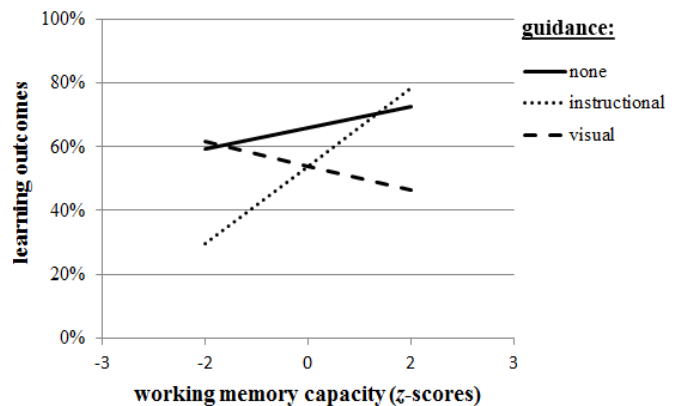


Figure 2: Interaction between working memory capacity (z-scores) and learning outcomes (%).

## Discussion

The present study tested whether learners benefited from guidance in an animation on how a technical device works, and whether different types of guidance led to different learning outcomes. Furthermore, it investigated the interaction between learners' WM capacity and type of guidance. Learners in all conditions indicated a significant increase in their knowledge after watching the animation. In accordance with our first hypothesis, the no-guidance group reported the highest level of difficulties with making



references between the visual display and the narration, followed by the instructional group and the visual group, who reported the fewest difficulties. Nevertheless, because only the difference between the no-guidance and visual guidance group was statistically significant, we consider our hypothesis only partially confirmed. Contrary to our expectations expressed in the second hypothesis, the no-guidance group outperformed both guidance groups on learning outcomes. Our third hypothesis was partially confirmed, given that WM capacity did affect learning from animations. In our experimental conditions, visual guidance had a detrimental effect on learners with high WM capacity, whereas instructional guidance had a detrimental effect on learners with low WM capacity. In general, the no-guidance group performed best, despite more perceived difficulties in making references. It follows that low WM learners benefited from no and visual guidance while high WM learners benefited from no and instructional guidance.

Contradicting conclusions drawn by Ayres and Paas (2007) that animations are more effective when key information is cued or signaled, our findings suggest that cueing does not necessarily have a positive effect on learning outcomes although it can reduce the level of perceived difficulties. Cueing as well as instruction aim at reducing cognitive load by directing learners to the relevant information in the learning content. Making references between different sources of information (auditory and visual) can be assumed to be an indicator of cognitive load. In this respect, visual guidance accomplished its purpose by synchronizing highlights on visual information with narration and hereby facilitating mapping. However, we assume that this might have led learners to invest less effort in active learning, after perceiving the content as (too) easy to comprehend. The framework of desirable difficulties offers an explanation as to why visual attention guidance does not always lead to better learning outcomes in the field of dynamic visualizations (De Koning et al., 2010): learners may be “lulled into a false sense of understanding” that makes them overestimate their understanding of the learning content (Linn et al., 2011, p. 239). Hence, an animation that is designed to make comprehension and processing subjectively too easy can mislead learners about the necessary effort. One might deduce from this that some degree of perceived difficulty can challenge learners and make them invest more effort in active learning processes. When learners perceive the stimulus as being more demanding, they may try to compensate for that by expending more effort in understanding the material (Salomon, 1984). Based on these findings, we propose that in order to promote active integration of learning materials, learners need to be given some challenges. More research is needed to find the balance between promoting effort and overload.

Another explanation for the poor performance of both guidance groups may be unfamiliarity with the chosen types of guidance. Blurring out the irrelevant parts can irritate learners. Visual guidance is an invasive form of alteration to

the original display. It restricts learners’ natural exploration behavior to the highlighted parts of the learning environment; spotlights expose only the parts that are important for the immediate moment. As a consequence, a deeper holistic integration of past and present information may be disrupted. Instructional guidance, on the other hand, is a less invasive type of support, at least with respect to the visual display. However, it requires learners to keep the instruction in mind while processing the animation. In light of limited WM capacities, learning processes and recall of instruction may conflict; especially learners with low WM capacity may suffer from this type of guidance. High WM learners, by contrast, may be able to follow the strategy they were instructed to apply while simultaneously blocking irrelevant and distracting information through the course of learning. Furthermore, guidance, whether invasive or not, can interfere with already established strategies and, thus, with self-regulatory processes. In contrast to the method of emphasis change (Gopher et al., 1989), we did not offer any feedback to our participants in the instructional guidance condition, neither on their performance nor on their attention allocation. As a suggestion for further research we propose a real-time feedback on learners’ eye movements.

Narration per se influences a learner’s attention. It can evoke expectations and provide knowledge directly prior to the processing of visual information. Consequently, prior expectations and knowledge can affect attention allocation to a visual display and influence the integration of new information (Kriz & Hegarty, 2007). Narration can therefore function as a top-down guidance of visual attention (Kriz & Hegarty, 2007; Lowe & Boucheix, 2008). Students in the no-guidance condition were guided by the narration but still had enough room to explore the whole display and thus integrate diverse information. In sum, the no-guidance condition seems to be a perfect fit for learners who can self-regulate their needs according to their resources, for example, their WM capacity and prior knowledge. At this point it should be stressed that our participants were students in a highly selective psychology program who already had thirteen successful years of school education and thus may be considered highly experienced in learning from multiple external representations and dynamic visualizations.

Based on our findings, we suggest that future multimedia research should place more emphasis on learner variables such as WM capacity to shed light on intervention effects of instructional designs. From our point of view, WM capacity could play a severe role in learning processes that could be comparable to the significance of prior knowledge in this context (Kalyuga, Ayres, Chandler, & Sweller, 2003).

## Acknowledgments

We thank our technical consultants and supporters: Adrian Skuballa for programming computer-standardized tests, Michael Kutz for bringing our storyboard to life, and Clara Pieck for giving our animation her professional voice. Furthermore, we thank our student assistants: Nicole

Dillner, Caroline Fortunski, and Sabina Panetta. This research was supported by the ScienceCampus Tübingen "Informational Environments" (Adaptable and Adaptive Multimedia Systems).

## References

- Ayres, P., & Paas, F. (2007). Can the cognitive load approach make instructional animations more effective? *Applied Cognitive Psychology*, 21, 811–820. doi:10.1002/acp.1351
- Baddeley, A. H. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, 4, 829–839. doi:10.1038/nrn1201
- Bleckley, M. K., Durso, F. T., Crutchfield, J. M., Engle, R. W., & Khanna, M. M. (2003). Individual differences in working memory capacity predict visual attention allocation. *Psychonomic Bulletin & Review*, 10, 884–889. doi:10.3758/BF03196548
- Conway, A. R. A., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, 8, 331–335. doi:10.3758/BF03196169
- De Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2007). Attention cueing as a means to enhance learning from an animation. *Applied Cognitive Psychology*, 21, 731–746. doi:10.1002/acp.1346
- De Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2010). Attention guidance in learning from a complex animation: Seeing is understanding? *Learning and Instruction*, 20, 111–122. doi:10.1016/j.learninstruc.2009.02.010
- Gopher, D., Weil, M., & Bareket, T. (1994). Transfer of skill from a computer game trainer to flight. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 36, 387–405. doi:10.1177/001872089403600301
- Gopher, D., Weil, M., & Siegel, D. (1989). Practice under changing priorities: An approach to the training of complex skills. *Acta Psychologica*, 71, 147–177. doi:10.1016/0001-6918(89)90007-3
- Höfler, T. N., & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction*, 17, 722–738. doi:10.1016/j.learninstruc.2007.09.013
- Jarodzka, H., Balslev, T., Holmqvist, K., Nyström, M., Scheiter, K., Gerjets, P., & Eika, B. (2010). Learning perceptual aspects of diagnosis in medicine via eye movement modeling examples on patient video cases. In S. Ohlson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1703–1708). Austin, TX: Cognitive Science Society.
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, 38, 23–31. doi:10.1207/S15326985EP38\_01\_4
- Kalyuga, S., & Hanham, J. (2011). Instructing in generalized knowledge structures to develop flexible problem solving skills. *Computers in Human Behavior*, 27, 63–68. doi:10.1016/j.chb.2010.05.024
- Kriz, S., & Hegarty, M. (2007). Top-down and bottom-up influences on learning from animations. *International Journal of Human-Computer Studies*, 65, 911–930. doi:10.1016/j.ijhcs.2007.06.005
- Linn, M. C., Chang, H.-Y., Chiu, J. L., Zhang, Z. H., & McElhaney, K. (2011). Can desirable difficulties overcome deceptive clarity in scientific visualizations? In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting. A festschrift in honor of Robert A. Bjork* (pp. 235–258). New York, NY: Psychology Press.
- Lowe, R., & Boucheix, J.-M. (2008). Learning from animated diagrams: How are mental models built? In G. Stapleton, J. Howse, & J. Lee (Eds.), *Diagrammatic representation and inference. 5th International Conference, Diagrams 2008, Herrsching, Germany, September 19-21, 2008; proceedings* (pp. 266–281). Berlin: Springer.
- Mautone, P. D., & Mayer, R. E. (2001). Signaling as a cognitive guide in multimedia learning. *Journal of Educational Psychology*, 93, 377–389. doi:10.1037/0022-0663.93.2.377
- Plass, J. L., Homer, B. D., & Hayward, E. O. (2009). Design factors for educationally effective animations and simulations. *Journal of Computing in Higher Education*, 21, 31–61. doi:10.1007/s12528-009-9011-x
- Salomon, G. (1984). Television is "easy" and print is "tough": The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of Educational Psychology*, 76, 647–658. doi:10.1037/0022-0663.76.4.647
- Schnotz, W., & Lowe, R. (2008). A unified view of learning from animated and static graphics. In R. Lowe & W. Schnotz (Eds.), *Learning with animation. Research implications for design* (pp. 304–356). Cambridge, New York: Cambridge Univ.
- Schüler, A., Scheiter, K., & Van Genuchten, E. (2011). The role of working memory in multimedia instruction: Is working memory working during learning from text and pictures? *Educational Psychology Review*, 23, 389–411. doi:10.1007/s10648-011-9168-5
- Schwonke, R., Berthold, K., & Renkl, A. (2009). How multiple external representations are used and how they can be made more useful. *Applied Cognitive Psychology*, 23, 1227–1243. doi:10.1002/acp.1526
- Von Aster, M., Neubauer, A., & Horn, R. (Eds.). (2009). *Wechsler-Intelligenztest für Erwachsene WIE* (2nd ed.). Frankfurt am Main: Pearson Assessment & Information GmbH.

# The Inductive Potential of Religion Categories in Northern Ireland

**Kirsty Smyth (ksmyth26@qub.ac.uk)**  
**Conor Pendergrast (cpendergrast01@qub.ac.uk)**

**Aidan Feeney (a.feeney@qub.ac.uk)**  
School of Psychology, Queen's University Belfast  
University Road, Belfast BT7 1NN, Northern Ireland

**John D. Coley (j.coley@neu.edu)**  
**R. Cole Eidson (eidson.r@husky.neu.edu)**  
Department of Psychology, Northeastern University  
360 Huntington Ave., MS 0125 NI  
Boston, MA 02115-5000

**Ulrike Niens (u.niens@qub.ac.uk)**  
School of Education, Queen's University Belfast  
69/71 University Street, Belfast BT7 1HL, Northern Ireland

## Abstract

People often behave as if category members share an essence, and essentialising a category in this way promotes inductive inference. Although natural kind categories have been predominantly studied, some social categories are essentialised and here we consider the inductive potential of religion categories for children in Northern Ireland. We asked seven-, nine- and eleven-year olds in Catholic-maintained, State-controlled and Integrated schools to decide whether pairs of children shared a property. We manipulated the degree of shared membership in religion, gender and musical categories. Overall, religion was much more inductively potent than gender, although older children and children from the Catholic-maintained school were more likely to use additional information about other categories when evaluating inferences. These results suggest that religion categories are a powerful basis for inference even in children as young as seven-years, and that religion categories may be essentialised in Northern Ireland.

**Keywords:** Essentialism, inductive reasoning, category-based induction, cognitive development

## Introduction

One of the central functions of categories is to support inductive inference (see Murphy, 2004). For example, knowing that someone is a member of the Rotary Club may help us to predict their behavior on the basis of previous encounters with other members of that category. The strength of an inductive inference from one member of a category to another is, in part, determined by the degree to which the category is essentialised. Members of essentialised categories are believed to share an underlying, unobservable essence (Gelman, 2003; Medin & Ortony, 1989), which we often can't define, and so is sometimes thought of as a place holder (Medin & Ortony, 1989). Essentialised categories are assumed to have innate potential, fixed boundaries, deep causal properties, the capacity to be informative, and stable membership across

time (Demoulin, Leyens & Yzerbyt, 2006; Gelman, 2003). Essentialist reasoning has been found very early in childhood; preschoolers make category-based inferences even when perceptual similarity conflicts with category-membership (Gelman & Markman, 1986).

According to Gelman (2003) inductive inference is an indirect measure of essentialism; the more a category is essentialised, the greater inductive potential it will have. Although there is a very large social psychological literature on essentialised social categories (e.g. Haslam, Bastian, Bain & Kashima, 2006), there has been relatively little work on how we make inductive inferences from social categories. In this paper we will describe a large developmental study designed to investigate the inductive potential of religious categories in Northern Ireland.

## Essentialised Social Categories

Evidence is accumulating to suggest that children essentialise certain socially-constructed categories (Birnbaum et al., 2010; Deeb et al., 2011; Diesendruck & haLevi, 2006; Gil-White, 2001; Hirschfeld, 1995; Kinzler & Dautel, 2012; Rhodes & Gelman, 2009; Taylor et al, 2009). Perhaps of most relevance to our goal of investigating religion categories is previous work on race and ethnicity. The degree to which race categories are essentialised has been studied in a number of ways. For example, Deeb et al. (2011, Study 3) presented Arab and Jewish participants with a questionnaire designed to investigate children's essentialised beliefs about such categories, whereas Rhodes and Gelman (2009) asked children whether categorization decisions made by an alien, most of which were inconsistent with the participants' social categories, were correct. Kinzler and Dautel (2012) asked participants whether a target child's 'race' or the language they spoke was most likely to persist into adulthood.

Although different methodologies lead to different conclusions about the developmental trajectory of

essentialist beliefs about race categories, it is clear that culture plays a role. For example, Rhodes and Gelman (2009) observed an effect of whether adolescents were from a rural or an urban background on how likely they were to treat 'race' as a natural kind. Deeb et al. (2011) showed that although essentialist beliefs about race categories appear to recede by late childhood, culture and experience play a role in determining the age at which they first begin to wane. Kinzler & Dautel (2012) found that European-American children do not treat 'race' as more predictive than language until the age of 10; whereas 5 year old African-American children privilege 'race' over language.

Other studies have examined the inductive potency of race categories. For example, Hirschfeld (1995) showed that children as young as three found race categories more inductively powerful than occupation or body build categories. In a study conducted in Israel with Jewish and Arab children, Diesendruck and haLevi (2006) compared the inductive potential of personality traits and a variety of social categories including ethnicity and gender. They found that five-year-old children treated social categories as having more inductive potential than personality traits, while adults found personality traits more powerful. However, when the children and adults made inductive inferences based on social categories only, ethnicity had the greatest inductive potential for both groups. Birnbaum et al. (2010) showed that ethnicity categories (Arab vs. Jew) had the most inductive potential for religious Jewish children by age 5 and up until 11 years of age. Birnbaum et al's study also highlights the effect of culture as neither secular Jewish children nor Arab children showed the same effect.

## Religion Categories

In this study we examined the development of sensitivity to the religious categories, *Catholic* and *Protestant*, in Northern Ireland. We are by no means the first to study the psychological effects of the history of sectarian conflict between Catholics and Protestants in Northern Ireland (for a relevant reviews see Trew, 2004). However, a study of the inductive potency of religion categories in Northern Ireland has the potential to add significantly to our understanding of essentialised social categories. Most obviously, whereas there are visual cues to racial or gender category membership, there are no such cues for religion category membership in Northern Ireland. Gil-White (2001) has suggested that essentialised social categories possess visual cues to category membership. In seeking to establish that social categories for which there are no visual cues can be highly inductively potent, our work will be a test of that hypothesis.

In addition, given the link that has been found between essentialism and stereotyping, prejudice and negative intergroup relations (e.g. Haslam, Bastian, Bain & Kashima, 2006; Howell, Welkum & Dyck, 2011; Pauker et al, 2010; Prentice & Miller, 2007), any tendency to essentialise religion categories may have important social implications for individuals in Northern Irish society. A review by

Prentice and Miller (2007) highlighted how essentialist beliefs, such as innate potential, immutability of category membership, stability of category membership over time, and the belief that deep, hidden properties of a category give rise to its observable properties can lead individuals to stereotype another group, feel less inclined to attempt to cross category boundaries that are apparently immutable, and maintain prejudicial attitudes towards an outgroup. All of these consequences of essentialised thinking could be said to pertain in Northern Ireland, yet there has been no study of how religion-category-based reasoning develops.

## The Current Study

This study aimed to examine the inductive potential of religion categories for seven-, nine-, and eleven-year-old children in Northern Ireland. The lower age group has been selected because of findings that by six to seven years of age, *Catholic* and *Protestant*, are meaningful social labels for children (Trew, 2004). Our first concern was the extent to which children are prepared to base inferences about individuals on information about the social categories to which they belong, and how that tendency changes over development. However, we were also concerned with the effects of the environment on inference-making from religion categories. One obvious environmental difference between Northern Irish children is the kind of school that they attend as in Northern Ireland there are different types of school: state-controlled (often attended primarily by children from a Protestant background); Catholic-maintained (attended primarily by children from a Catholic background); and integrated (attended by children from both communities). So as to examine effects of educational environment on inferences, we included school type as a variable in the study.

On each trial of our task we showed participants a line drawing of a child said to be a member of two categories, for example a male Protestant. Next we told participants that the child possessed a blank property (e.g. likes to play a game called *badlage*), and asked them whether they thought another male Protestant, a male Catholic, a female Protestant, and a female Catholic would also possess the feature. Unlike standard triad tasks, our method does not force participants to choose between categories as the basis for an inference. Work on ethnic categories (Deeb et al, 2011) has shown that younger and older children spontaneously mention and remember information about ethnic categories or use them as the basis for inference. Accordingly, we predicted that to the extent that religion categories are essentialised, all of the children in this study would base inferences on religion category membership. Other work (Birnbaum et al, 2010) has shown that older children take additional information into account when making inferences. Accordingly, we predicted that with development, children would use the control categories in our task as an additional basis for inference. Because Deeb et al. (2011) found that children from integrated schools were more sensitive to ethnicity categories, we predicted

that children in such schools in Northern Ireland might show greater sensitivity to religion category somewhat earlier in development than children attending other schools.

## Method

**Participants** One hundred and thirty five (67 females and 68 males) primary school children in Northern Ireland participated. Participants were aged 6-7 (P3), 8-9 (P5) and 10-11 (P7). Participants were recruited from one integrated (33 children: 9 in P3; 10 in P5; 14 in P7); one state-controlled (45 children: 12 in P3; 11 in P5; 22 in P7); and one Catholic-maintained school (57 children: 18 in P3; 19 in P5; 20 in P7). Children were tested for ten minutes each in a quiet corner of the school.

**Design** The study had a 3 (School: Integrated; State-controlled; Catholic-maintained) x 3 (Year: P3; P5; P7) x 4 (Trial Type: both categories overlap; religion category only overlaps; other category only overlaps; no overlap) mixed design. Trial Type was manipulated within subjects, and school and year were between subject variables.

On each trial participants received information about two children's membership in two categories. In four trials the categories were religion and gender, in four others they were religion and musical instrument, and in a final four they were gender and musical instrument. There were two religion categories, Protestant and Catholic, two gender categories, male and female, and two musical instrument categories, piano player and guitarist. Trials were blocked by category pair and block order was counterbalanced.

**Materials** Each participant attempted 12 trials. A trial consisted of a base picture and a target picture. The same base picture was used for four target pictures. In one of the base pictures, information was presented about the religion and gender categories of the child. In another, information was presented about religion category and which musical instrument the child could play, and in the third base picture, information was presented about gender and musical instrument. The sets of four target pictures presented with each base depicted four different children who either shared memberships of the same two categories as the base child, shared membership of only one of the categories, or shared membership of neither category. Example pictures and trials are presented in Figure 1.

In all cases information about category membership was related orally when a picture was presented. Religion category membership was visually represented by a picture of a familiar Catholic or Protestant cathedral in Northern Ireland. Gender category was clearly discriminable from the pictures because of visual cues such as dress and hair length, and musical instrument category was depicted by the inclusion of a picture of a guitar or a piano. When gender was not varied in a trial, the child was presented in silhouette.

Three unfamiliar, novel properties were invented: 'likes to play a game called badlage'; 'knows how to use a mixtle';

'will go brooping at the weekend'. The same property was used for each of the trials associated with a particular base picture, and the order in which properties were mentioned was fixed. However, the order in which base pictures were presented was fully counter-balanced so that each property was presented with each base picture an equal number of times. We also varied the base pictures so that, for example, the base child defined by membership of a religion and a gender category was depicted equally often as a Catholic boy, a Catholic girl, a Protestant boy or a Protestant girl.

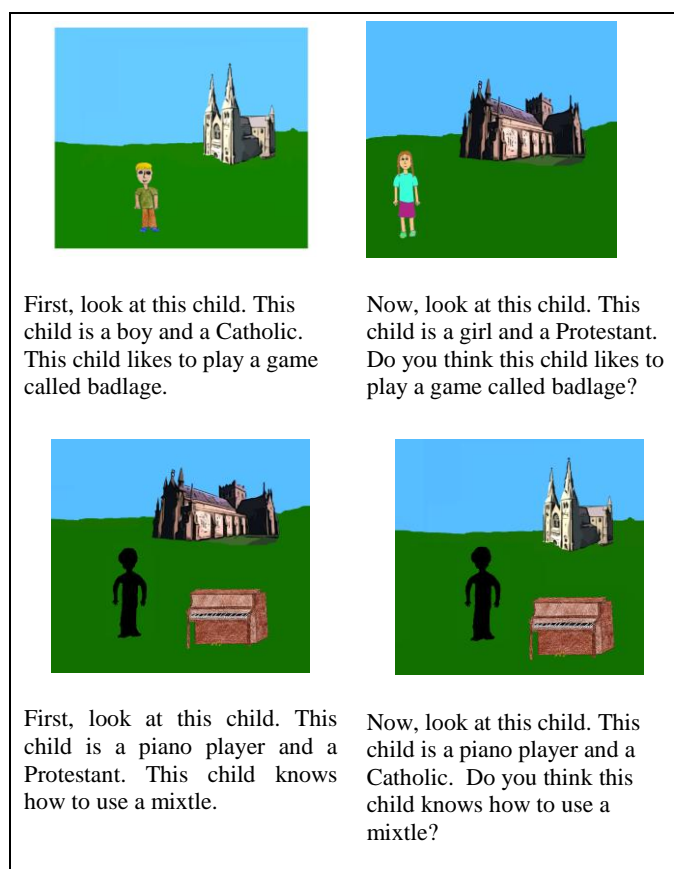


Figure 1: Example stimulus materials. The top panel illustrates how gender and religion categories were depicted. The bottom panel illustrates how religion and musical instrument categories were depicted.

**Procedure** Participants were tested individually at a quiet location in their school. On each trial participants indicated whether the novel property was possessed by both children or by the base child only.

## Results

We present a number of different analyses of our results. First, we present an analysis of the eight trials where information was presented about the religion category to which the base child belonged. In four of these trials, the other category was gender, and in the remaining four the other category was musical instrument. These eight trials

may be divided into four types. In two *REL+/GM+* trials, the base and target children belonged to the same two categories. In two *REL+/GM-* trials, base and target belonged to the same religion category, but to different gender or musical instrument categories. In two *REL-/GM+* trials, base and target belonged to different religion categories but to the same gender or musical instrument categories. Finally in two *REL-/GM-* trials, base and target belonged to entirely different categories.

We carried out a 4(Trial Type) x 3(Year) x 3(School) mixed design ANOVA on the number of times participants endorsed each type of inference. As may be seen in Figure 2, the results of this analysis contained a highly significant main effect of trial type,  $F(3, 378) = 56.38, p < .001$ . Post hoc tests revealed that each of the means involved in this effect differed significantly from every other mean. The main effect was qualified by significant interactions between School and Trial Type,  $F(6, 378) = 2.27, p < .05$ , and between Year and Trial Type,  $F(6, 378) = 2.46, p < .05$ . These interactions are shown in Figures 3 and 4 respectively.

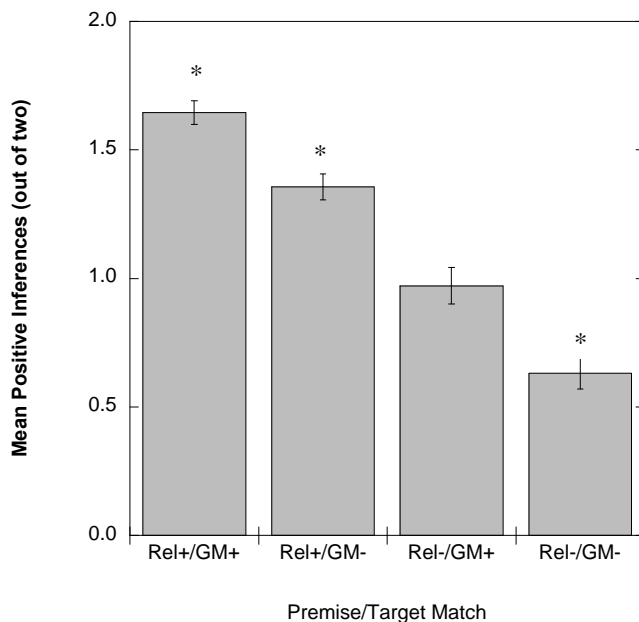


Figure 2: Mean rates of inference for each trial type. (Note that here and throughout, bars marked with an asterisk differ significantly from chance.)

An examination of Figure 3 suggests that regardless of the type of school attended, children are more willing than not to project properties once the base and target children share religion category membership. However, only children attending the Catholic-maintained school were as willing to project properties in the *REL-/GM+* condition as in the *REL+/GM-* condition. These observations are supported by t-tests comparing the mean number of properties projected by children in each school on *REL+/GM-* and *REL-/GM+* trials. The difference was not significant for children attending the Catholic maintained school but was significant

for the other two groups of children. Thus, the finding here is of differences due to educational context in the potency of categories other than religion.

Examination of Figure 4 suggests that something similar underlies the significant interaction between Trial Type and Age. Whereas religion appears to be the only inductively potent category for seven- and nine-year-old children, eleven-year-olds are almost as likely to project the property in *REL-/GM+* trials as in *REL+/GM-* trials. The results of paired t-tests for each age group, comparing rates of

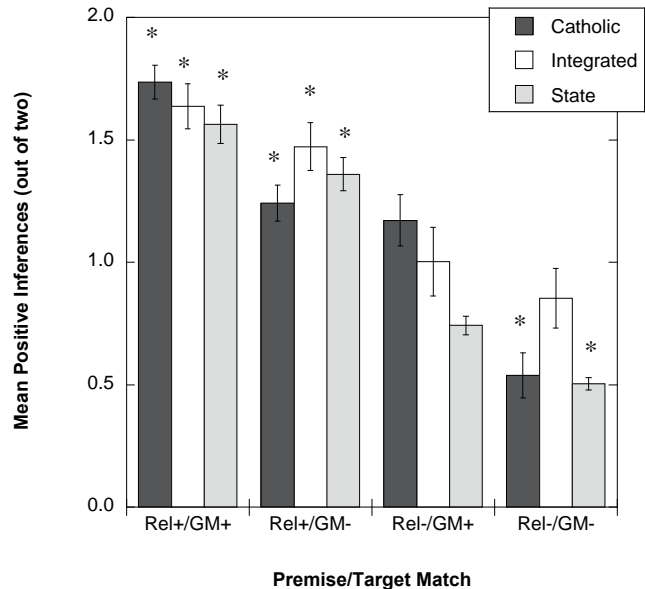


Figure 3: Interaction between school and trial type. (Bars marked with an asterisk differ significantly from chance.)

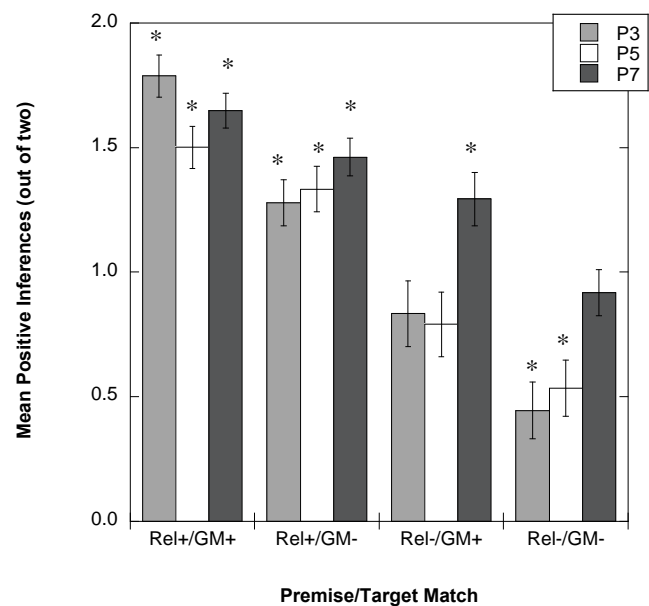


Figure 4: Interaction between year and trial type. (Bars marked with an asterisk differ significantly from chance.)



inference in each of these trial types, tended to support this interpretation. It is also apparent from Figure 3 that the oldest participants were more likely to project properties in REL-/GM- trials, but we have no explanation for this unexpected finding.

Thus far, we have collapsed gender and musical instrument categories together. It is clear from the analysis we have just described that overall, when information about religion category membership is present, religion categories are inductively more powerful than either musical instrument or gender categories. However, these analyses do not include trials where information about religion category was not present.

To examine the relative potency of all three types of category manipulated in this experiment, we calculated an index for each one reflecting the number of times a property was projected when there was overlap on that category type or relation. For example, the Gender index included G+/REL+, G+/REL-, G+/M+, and G+/M- trials. A 3(Relation) x 3(Year) x 3(School) mixed design ANOVA on scores on these indices revealed a highly significant main effect of Relation,  $F(2, 250) = 40.43$ ,  $p < .001$ , and a significant interaction between Relation and School,  $F(4, 250) = 3.05$ ,  $p < .02$ . The mean score on the Religion index ( $M = 3.00$ ) was just higher than the mean score on the Music index ( $M = 2.96$ ), but not significantly so. Both of these means were significantly higher than mean scores on the Gender index ( $M = 2.21$ ), and scores on all three indices were significantly greater than chance ( $p < .005$ ).

The interaction with School is displayed in Figure 5 where it appears that music is more inductively potent for children in the Catholic maintained school than is religion, whereas these two categories are equally potent for children in the Integrated school, and religion is more potent than music for

children in the State controlled school. A t-test confirmed that the difference between music and religion categories was statistically significant for children in the Catholic school. However, the difference between music and religion was not significant ( $p = .11$ ) for children in the State-controlled school. Scores on the Gender index were significantly lower than scores on both of the other indices for all three groups.

## Discussion

These results show that by age seven, children in Northern Ireland use religion categories—or at least the labels *Catholic* and *Protestant*—to guide inferences about non-obvious properties. These categories promote stronger inferences than other social categories (musicality, gender), and do so for children across a diverse range of school environments. Together, these findings suggest religion is essentialised by school children in Northern Ireland.

Although religion-based inferences were universal, we did see differences in the use of other social categories. Specifically, the musical instrument played by a child was a more potent guide for inferences for children in the Catholic school than the Integrated school, and more so for the Integrated school than the State school. Additionally, children in the State-controlled school were less likely than others to base inferences on gender. These findings suggest that culture or school may emphasize certain social categories over others, leading to differences in inductive potential for those categories. To children in Catholic-maintained schools, information about multiple social categories may be important, whereas to children in State-controlled schools, religion categories may be most salient. An alternative explanation is that ability levels differed between the schools, and that the use of information about more than one category to evaluate an inference requires additional cognitive resources. Further experimental work will need to be carried out in order to discriminate between these possibilities.

In addition to the universal potency of religion categories, we observed a developmental increase in the inductive potency of other social categories; 11-year-olds were more likely than younger children to base inferences on shared membership in non-religious categories. One interesting methodological point about the pattern of results that we have just described is that it would not have been detected by a standard triad task. To illustrate, Birnbaum et al. (2010) presented participants with trials where, for example, a child was said to belong to two social categories (e.g. a male Arab child) and participants were asked whether this base child was more likely to share a property with a male Jewish child or a female Arab child. Using this method, one can underestimate the age at which a certain type of social category begins to become important as well as the period for which it is important. Because our method did not force participants to choose between categories as a basis for their inference, we were able to observe that religion categories continued to be important at eleven-years of age but that

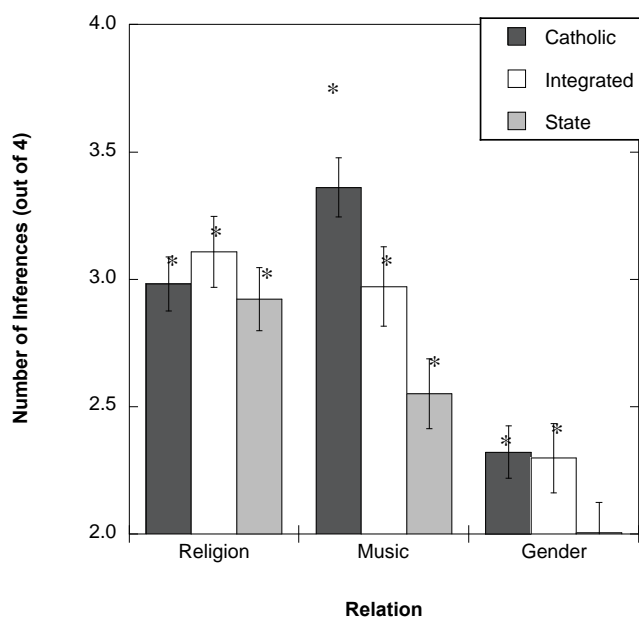


Figure 5: Interaction between Relation and school. (Bars marked with an asterisk differ significantly from chance).



other categories had also become more important by that age. Had we used a triad task we might have observed a tail-off in the apparent use of religion categories by older participants.

Of particular interest in this study is the performance of children from the Integrated school; by virtue of their contact with children from the “other” religious category they might display sensitivity to religion category earlier than children attending other schools. Deeb et al. (2011) reported such an effect for Jewish children attending integrated schools in Israel. We did not find an equivalent effect of school type here, although because we did not ask participants to disclose their religion, we cannot perform the same analysis as Deeb et al. (2011). Perhaps Catholic and Protestant children attending the same Integrated school behave differently, and we are currently carrying out a study designed, in part, to examine whether this is the case.

Consideration of children attending Integrated schools is useful when thinking about the relationship between a category’s inductive potency and the degree to which that category might be essentialised. One possible relationship is that social categories are inductively potent to the degree that they are essentialised, as argued by Gelman (2003). Another possibility is that a social category could be inductively strong even when it is not strongly essentialised. For example, Deeb et al (2011) reported a negative correlation between scores on a questionnaire designed to measure the extent to which children essentialise ethnic categories, and children’s tendency to recall information about ethnic category membership from an early description. Participants who essentialised ethnic categories less, recalled more information about ethnicity. This analysis collapsed across school type, so the result may have been driven entirely by children from the integrated sector. In any case, the result raises the possibility that children from different school sectors might perform similarly with respect to religion categories on an inductive inference task, but essentialise the categories to different extents. Members of a society divided on religious lines may not believe that members of the same religion category share an essence, but may still recognize the usefulness of information about religion category membership when making predictions about individuals.

Thus far, our work has concerned children growing up in a society that is still divided along religious lines. To properly assess the inductive potency of religion categories and the effects of culture on potency, there is a need for cross-cultural work comparing children in Northern Ireland to children from a society where religion is less divisive. For example, had we run this study in such a society we would have expected to find that gender categories are more inductively potent for children than are religion categories. We are currently collecting data to test this hypothesis.

We are at the beginning of a large project designed to investigate the extent to which religion categories support inferences differently across culture, across environment and across development. The results we have presented here

suggest that for children developing in a culture where religion categories are highly important, regardless of educational environment, those categories are inductively potent relatively early in development and continue to be so at least until late childhood.

## References

- Birnbaum, D., Deeb, I., Segall, G., Ben-Eliyahu, A. & Diesendruck, G. (2010) The development of social essentialism: The case of Israeli children’s inferences about Jews and Arabs. *Child Development*, 81, 3, 757-777.
- Deeb, I., Segall, G., Birnbaum, D., Ben-Eliyahu, A. & Diesendruck, G. (2011) Seeing isn’t believing: The effect of intergroup exposure on children’s beliefs about ethnic categories. *Journal of Personality and Social Psychology*, 101, 6, 1139-1156.
- Demoulin, S., Leyens, J. P. & Yzerbyt, V. (2006) Lay theories of essentialism. *Group Processes & Intergroup Relations*, 9 (1), 25-42.
- Diesendruck, G. & haLevi, H. (2006) The role of language, appearance, and culture in children’s social category-based induction. *Child Development*, 77, 3, 539-553.
- Gelman, S. A. (2003) The essential child: Origins of essentialism in everyday thought. New York: Oxford University Press.
- Gelman, S. A. & Markman, E. M. (1986) Categories and induction in young children. *Cognition*, 23, 183-209.
- Gil-White (2001) Are ethnic groups biological ‘species’ to the human brain? Essentialism in our cognition of some social categories. *Current Anthropology*, 42, 4, 515-554.
- Haslam, N., Bastian, B., Bain, P. & Kashima, Y. (2006) Psychological essentialism, implicit theories, and intergroup relations. *Group Processes & Intergroup Relations*, 9 (1), 63-76.
- Hirschfeld, L. (1995) Do children have a theory of race? *Cognition*, 54, 209-252.
- Howell, A. J., Welkum, B. A., Dyck, H. L. (2011) Psychological essentialism and its association with stigmatization. *Personality and Individual Differences*, 50, 95-100.
- Kinzler, K. D. & Dautel, J. B. (2012) Children’s essentialist reasoning about language and race. *Developmental Science*, 15: 1, 131-138.
- Medin, D. & Ortony, A. (1989) Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical processing* (pp. 179-195). New York: Cambridge University Press.
- Murphy, G. L. (2004) The big book of concepts. Massachusetts: MIT Press.
- Rhodes, M. & Gelman, S. A. (2009) A developmental examination of the conceptual structure of animal, artefact, and human social categories across two cultural contexts. *Cognitive Psychology*, 59, 244-274.
- Pauker, K., Ambady, N. & Apfelbaum, E. P. (2010) Race and essentialist thinking in racial stereotype development. *Child Development*, 81, 6, 1799-1813.
- Prentice, D. A. and Miller, D. T. (2007) Psychological essentialism of human categories. *Current Directions in Psychological Science*, 16, 4, 202-206.
- Trew, K. (2004) Children and socio-cultural divisions in Northern Ireland. *Journal of Social Issues*, 60, 3, 507-522.

# Does number interference occur during sentence processing?

Katja Suckow (k.suckow@dundee.ac.uk)

School of Psychology, The University of Dundee  
Dundee, DD1 4HN, UK

Roger P.G. van Gompel (r.p.g.vangompel@dundee.ac.uk)

School of Psychology, The University of Dundee  
Dundee, DD1 4HN, UK

## Abstract

Models of interference in sentence processing claim that object relative clauses are harder to process than subject relatives due to interference between the subject and object noun phrase. The interference effect for object relatives at the verb should be more pronounced when the two noun phrases retrieved from memory are similar. To test this, two eye tracking experiments manipulated whether the number feature of the noun phrases (singular or plural) was either the same or different. Both experiments showed the well-known relative clause effect. However, in Experiment 1 the effect of number congruency was in the opposite direction from that predicted by interference. Experiment 2 showed the interaction predicted by similarity based interference at sentence wrap-up, but because this interaction was observed later than the relative clause effect and only occurred in Experiment 2, it suggests that retrieval interference due to cue overlap is a weak effect that might be the result of a checking procedure in syntactically complex sentences.

**Keywords:** sentence processing; similarity based interference; working memory.

## Introduction

There is much evidence that object relative clauses such as (1) are more difficult to understand than subject relative clauses such as (2) (e.g., King & Just, 1991). It is generally believed that this difficulty stems from limitations in working memory (Gibson, 1998; Gordon, Hendrick, & Johnson, 2001; Lewis, 1996; King & Just, 1991).

1. The banker that the accountant helps counted the money.
2. The banker that helps the accountant counted the money.

For example, King and Just (1991) found that readers with low working memory experienced more difficulty with object than subject relatives. More recently, Lewis (1996) and Van Dyke and Lewis (2003) have proposed an account that assumes that working memory demands increase when two similar linguistic items (e.g., two noun phrases) need to be retrieved simultaneously, slowing down sentence processing. Lewis and Vasishth (2005) proposed an ACT-R account that explains how the processing of subject and object relatives is affected by the similarity of the noun phrases.

Consistent with this, Gordon, Hendrick, and Johnson (2001) found that object relatives are easier to process when the embedded noun phrase is a name (*Joe*) and the head noun a definite noun phrase (*the barber*) compared to when both are definite noun phrases.

However, Van Dyke and Lewis's (2003) and Lewis and Vasishth's (2005) similarity based interference account claims

that interference is not just due to similarity in the type of noun phrase but also due to similarity of other features. Van Dyke and Lewis (2003) described how the degree of overlap between a retrieval cue and two similar linguistic items affects the retrieval of these items. Two items (the subject and object noun phrase) have to be retained in memory in object relatives until they can be integrated with the verb. At the point of integration at the verb, retrieval cues are used to identify the target item from memory. When items in memory share retrieval cues it is difficult to identify this target. Thus, a similarity based interference effect arises for items with retrieval cue similarity. The account of Lewis and Vasishth (2005) predicts that a similarity based interference effect arises at the embedded verb in object relatives because two items have to be retrieved from memory instead of only one for subject relatives. This interference effect at the verb should be larger when these two items share retrieval cues.

For example, in object relatives like (1) the two subject and object noun phrases (*the accountant* and *the banker*) need to be simultaneously retrieved at the embedded verb (*helps*). On the other hand, only one noun phrase (*the banker*) needs to be accessed at the verb (*helps*) in subject relatives like (2). The similarity based interference effect that occurs with object relatives should be particularly strong when the two noun phrases that need to be retrieved are similar. Thus, because both the difficulty of object relatives and the difficulty of retrieving items that share retrieval cues are effects of interference at the embedded verb, they should occur at the same time.

Van Dyke and Lewis (2003) claim that number information is one of the retrieval cues for identifying a target. In relative clauses, congruency in the number feature of the subject and object noun phrase may have a strong effect on interference, because the subject has to agree in number with the verb: The retrieval cues of the target item (the subject noun phrase) have to match the number cue of the search probe (the verb). If the object has the same number as the subject and therefore also matches the number feature of the verb, this should result in interference when the subject and object are integrated with the verb. Thus, according to the retrieval cue based interference account, retrieval difficulty at the verb (e.g., *helps*) in object relatives should be more pronounced when the two noun phrases in memory share the same number than when they do not.

To test this account the current study investigated whether object relatives are harder to process when the two noun phrases in memory share the same number retrieval cue than when the number retrieval cues are different.

In sum, similarity-based interference models make the following three predictions for the experiments: (1) object relatives should be harder to process than subject relatives; (2) object relatives should be particularly hard to process when the subject and object noun phrase have the same number; (3) because both the relative clause and number effect predicted in (1) and (2) are due to interference when the noun phrases are retrieved at the verb (*helps*), the effects should occur simultaneously.

## Experiment 1

We conducted an eye movement reading experiment that contrasted sentences containing an embedded object relative clause with an embedded subject relative. The two initial noun phrases were either the same or different in number.

### Participants

Experiment 1 had 40 participants. All participants were non-dyslexic English native speakers and members of Dundee University. They received course credits in exchange for their participation. Participant treatment was in accordance with the ethical standards. The study was approved by the ethics committee at the University of Dundee.

### Materials and Design

Table 1 shows a sample item in all conditions and the areas of interest for the eye movement analyses. Thirty-two critical sentences were created in eight different conditions. The experiment had a 2x2x2 design with the factors (1) relative clause type (subject vs. object relative), (2) number congruency (subject and object noun phrases same vs. different in number) and (3) counterbalancing of number information (NP2 singular vs. plural). Because the number counterbalancing variable was not of theoretical interest, we collapsed across it in the analyses.

Eight lists were created. Each list contained 32 critical items, with four items in each of the eight conditions. One condition of each item appeared in each list. Five participants were randomly assigned to each list. In addition to the 32 experimental items, 85 filler sentences were presented and yes/no comprehension questions were presented after each sentence.

### Apparatus and Procedure

The experiment was carried out using the Experiment Builder Program from SR Research on a PC. An Eyelink 1000 Desktop Mount recorded participants eye movements at a 1000Hz sampling rate. The experiment was controlled by the Experiment Builder software on a separate PC. DataViewer (SR research) as well as R (R 2.13.1 foundation for statistical computing) were used for data analysis.

## Results

Three different eye-tracking measures were analysed for each region. *First pass duration* is the duration from entering an area of interest for the first time until leaving it into any direction. This measure does not include fixations that occurred after readers had fixated a subsequent region (i.e. the region of interest was skipped). *Regression path duration* is the sum of fixations from entering the area from the left for the first time until the first fixation outside to the right of the region occurs. That means that there should be no fixation on a right-bound region before entering the region in regression path duration. *Total reading time* is the sum of all fixations in an interest area.

Figure 1 shows regression path duration and total reading time in the different interest areas for Experiment 1. Since there were no differences between the conditions in first pass duration, we omit the plot for this measure here.

We conducted both analyses of variance with subjects (F1) and items (F2) as a random variable. Relative clause type and number congruency were treated as within subject and within item fixed variables. In addition, subject group was a fixed between subject variable in the by-subject analyses and item group a between item variable in the by-item analyses.

In the following results section we only report analyses of the variables that showed significant effects by subjects or by items.

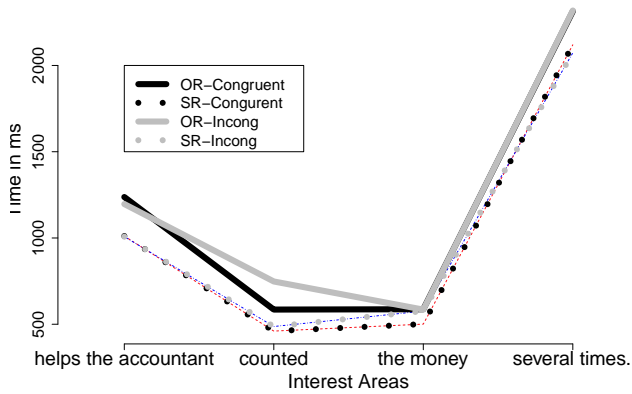
**Critical Region (*helps the accountant*).** The analyses of variance showed a main effect of relative clause for regression path duration:  $F(1,32) = 33.03, p < .005$ ;  $F(1,24) = 35.09, p < .01$ . Reading times for object relatives were longer than for subject relatives. Total reading time at the critical region also showed an effect of relative clause:  $F(1,32) = 30.95, p < .01$ ;  $F(1,24) = 31.86, p < .01$ , indicating that object relatives took longer to read than subject relatives.

**Spillover 1 Region (*counted*).** Analyses of variance showed a main effect of relative clause type for the regression path duration measure,  $F(1,32) = 32.68, p < .01$ ;  $F(1,24) = 57.42, p < .01$ . Mean reading times for object relatives were longer than for subject relatives. The analyses also showed a main effect of number congruency for regression path duration:  $F(1,32) = 9.75, p < .01$ ;  $F(1,24) = 14.22, p < .01$ . Reading times were longer when the noun phrases were different than the same in number. The analyses of regression path duration also showed an interaction effect between relative clause type and number congruency,  $F(1,32) = 5.80, p < .05$ ;  $F(1,24) = 5.40, p < .05$ . Simple effect analyses for the object relatives showed that they took longer to read when the noun phrases were different than the same in number:  $F(1,32) = 8.94, p < .01$ ;  $F(1,24) = 12.22, p < .01$ . In contrast, simple effect analyses for subject relatives showed that there was no difference between the same and different conditions ( $F_s < 1$ ). For the total reading time measure, there was a main effect of relative clause type,  $F(1,32) = 25.11, p < .01$ ;  $F(1,24) = 15.87, p < .01$ . Reading times for object relatives were longer than for the subject relatives. In addi-

Table 1: Example material and areas of interest for Experiment 1

	critical region	spillover 1	spillover 2	wrap-up
<i>Subject relative/same number</i>				
The banker that	helps the accountant	counted	the money	several times.
The bankers that	help the accountants	counted	the money	several times.
<i>Object relative/same number</i>				
The banker that	the accountant helps	counted	the money	several times.
The bankers that	the accountants helps	counted	the money	several times.
<i>Subject relative/different number</i>				
The bankers that	help the accountant	counted	the money	several times.
The banker that	helps the accountants	counted	the money	several times.
<i>Object relative/different number</i>				
The bankers that	the accountant helps	counted	the money	several times.
The banker that	the accountants help	counted	the money	several times.

Regression Path Duration



Total Reading Time

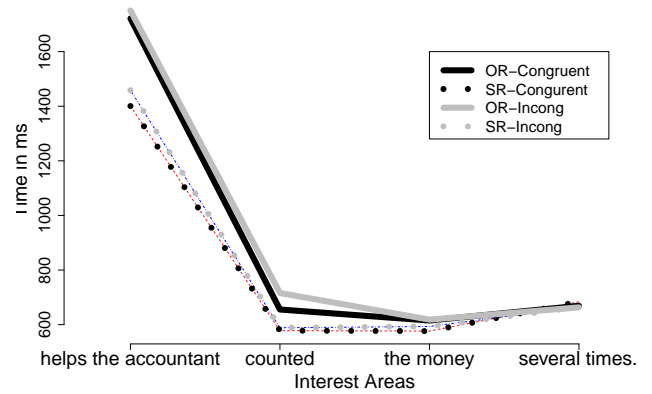


Figure 1: Regression Path and Total Reading Times in Experiment 1

tion, the analyses of the total reading time measure indicated a marginal effect of number congruency,  $F(1,32) = 3.41$ ,  $p = .07$ ,  $F(1,24) = 2.96$ ,  $p = .09$ , with the different condition being read slower than the same condition.

**Spillover 2 Region (*the money*).** There were no effects in the spillover 2 region.

**Wrap-up Region (*several times*).** An analysis of variance of the regression path duration measure showed a main effect of relative clause type in the wrap-up region that was significant by subjects,  $F(1,32) = 7.01$ ,  $p < .05$ ; but not by items,  $F(2,24) = 4.18$ ,  $p = .05$ , with object relatives being slower than subject relatives.

## Discussion

Experiment 1 showed that object relatives were harder to process than subject relative clauses in regression path and total reading times for the critical region (*help the accountant*) and the spillover 1 region (*counted*). There was an effect of number congruency in total reading times in the spillover 1 region, with the different number conditions being harder than the same number conditions. A similar effect was also found in regression path duration in the same region, though it only occurred with object relatives. The direction of these congruency effects is opposite to that predicted by similarity based interference models, which claim that the same number condition should be more difficult. The reversed effect may be due to priming of number information from the first to the second noun. Most important, this suggests that similarity-based interference due to number congruency does not appear to have a strong effect on the processing of subject or object-relative clauses.

## Experiment 2

Experiment 1 showed an effect of relative clause type but no effect of number interference. It is possible that number interference does occur, but that in contrast to what is predicted by the similarity based interference account, it is a later effect. Number interference may occur after initial syntactic analysis of the relative clause, when readers check whether the verb also agrees with noun phrases other than the subject. This sort of mechanism is similar to the account put forward by Sturt (2003) for a delayed effect of *unaccessible antecedents* in anaphor resolution. Sturt (2003) argues that processes associated with this effect are not part of initial anaphor interpretation but later procedures like recovery and sentence wrap-up.

Because the region after the relative clause was relatively long in Experiment 1, the retrieval interference effect due to cue overlap may be spread out over the region and difficult to detect. Therefore, the final region after the relative clause was shortened in Experiment 2. The same materials as in Experiment 1 were tested without the final region (e.g., *several times*).

## Participants

The number of participants and the selection criteria for these participants were the same as in Experiment 1.

## Materials and Design

The materials in this experiment were the same as in Experiment (1) with one difference. Instead of two spillover regions and a wrap-up region after the critical region, there was only one spillover region followed by a sentence wrap-up region. Table 2 shows a sample item with the interest areas for this experiment.

## Apparatus and Procedure

The procedure and the apparatus were the same as in Experiment 1.

## Results

Figure 2 shows the plots for first pass duration and total reading time and Figure ?? the plot for regression path duration in the different interest areas. We analysed the results from first pass duration, regression path duration and total reading time in the same way as in Experiment 1.

**Critical Region (*helps the accountant*).** In regression path duration, we observed a main effect of relative clause type  $F(1,32) = 17.81$ ,  $p < .01$ ;  $F(1,24) = 16.91$ ,  $p < .01$ . Object relatives had longer reading times than subject relatives. The total reading time measure also showed a main effect of relative clause type,  $F(1,32) = 46.07$ ,  $p < .01$ ;  $F(1,24) = 31.80$ ,  $p < .01$ . Reading times for object relatives were longer than for subject relatives.

**Spillover Region (*counted*).** First pass duration showed a main effect of relative clause,  $F(1,32) = 8.95$ ,  $p < .01$ ;  $F(1,24) = 8.65$ ,  $p < .01$ : Object relatives took longer than subject relatives. In regression path duration, there was an effect of relative clause type that was significant by subjects  $F(1,32) = 6.23$ ,  $p < .05$  but marginal by items  $F(2,24) = 3.81$ ,  $p = .06$ . Analyses of total reading time for this region showed a significant main effect of relative clause  $F(1,32) = 39.00$ ,  $p < .01$ ;  $F(1,24) = 26.94$ ,  $p < .01$ . Mean reading times for object relatives were longer than for subject relatives.

**Sentence Wrap-Up region (*the money*).** In first pass duration, there was a marginal effect of relative clause by subjects:  $F(1,32) = 3.41$ ,  $p < .10$  but no effect by items  $F(2,24) = 1.97$ . There was also an effect of number congruency by subjects  $F(1,32) = 4.42$ ,  $p < .05$  but not by items  $F(2,24) = 2.51$ ,  $p < .15$ , with a longer first pass duration when the noun phrases shared the same number cue than when they did not. Most importantly, there was an interaction between relative clause type and number congruency in first pass duration  $F(1,32) = 4.10$ ,  $p = .05$ ; and  $F(1,24) = 4.91$ ,  $p < .05$ . Simple effect analyses for object relative clauses showed that they had longer reading times when the noun phrases were the same than different in number:  $F(1,32) = 6.20$ ,  $p < .05$ ;

Table 2: Areas of interest for the materials used in Experiment 2.

	critical	spillover	wrap-up
<i>Subject relative</i>			
The banker(s) that	help(s) the accountant(s)	counted	the money.
<i>Object relative</i>			
The banker(s) that	the accountant(s) help(s)	counted	the money.

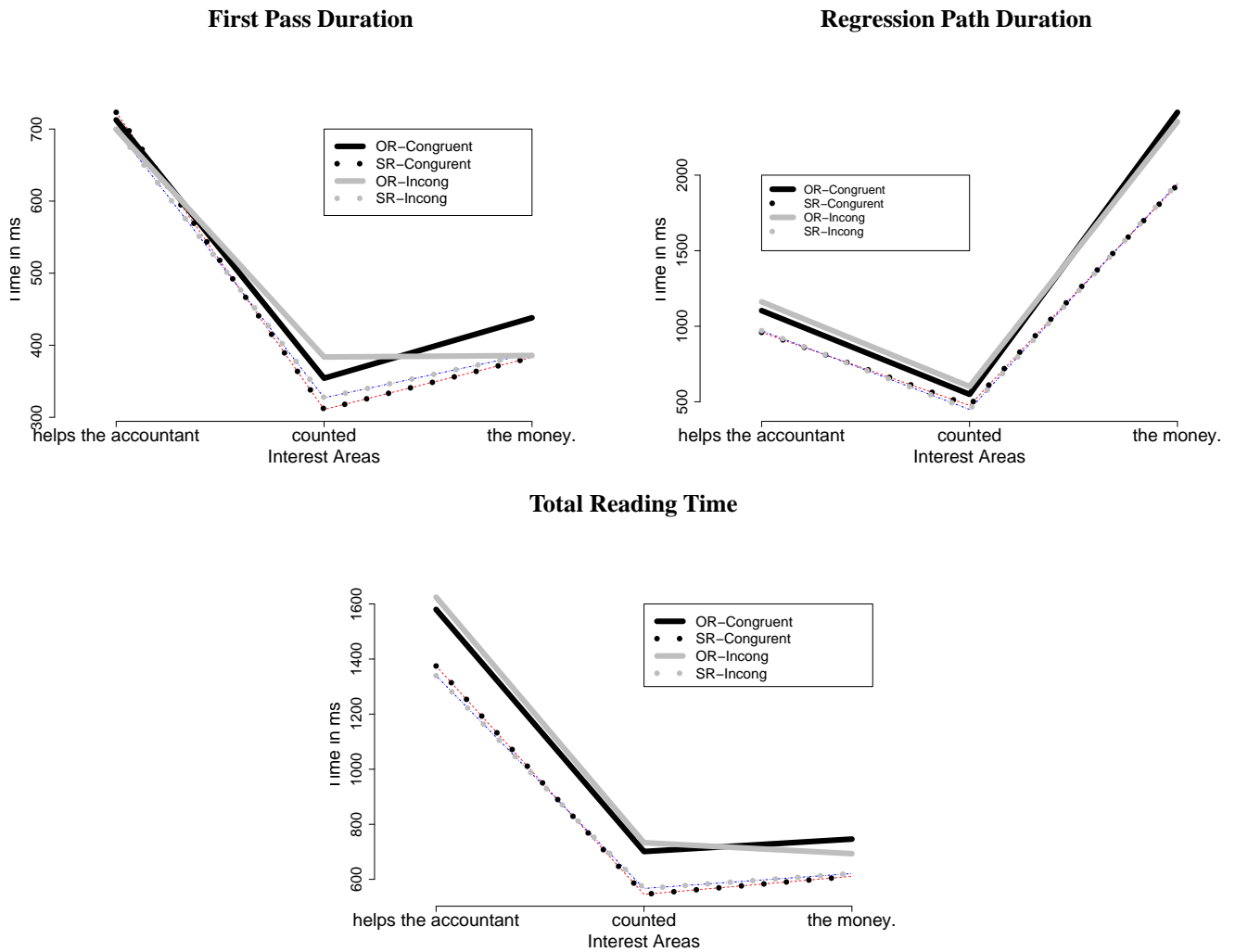


Figure 2: First Pass, Regression Path and Total Reading Times in Experiment 2

$F(1,24) = 5.97, p < .05$ . Simple effect analyses for subject relatives showed that there was no difference between the same and different noun phrases ( $F_s < 1$ ). The analyses of regression path duration showed a main effect of relative clause type  $F(1,32) = 22.76, p < .01$ ;  $F(1,24) = 10.70, p < .01$ . Object relatives had longer reading times than subject relatives. In the total reading time measure, we observed an effect of relative clause  $F(1,32) = 18.63, p < .01$ ;  $F(1,24) = 16.88, p < .01$ . Object relatives took longer to read than subject relatives.

## Discussion

Object relatives were more difficult than subject relatives relatives in regression path duration and total reading time (in all three regions) and in first pass duration (in the spillover region). Most interesting, there was an interaction between relative clause type and number congruency in first pass time for the final wrap-up region (*the money*): object relatives were more difficult when the noun phrases had the same number than a different number, whereas there was no difference for subject relatives. Thus, the number congruency effect appeared later than the object relative difficulty. This is surprising given the predictions made by the interference models mentioned in the introduction. Both the relative clause effect and the number congruency effect are claimed to be due to memory interference that occurs when the subject and object noun phrase are retrieved at the embedded verb. Thus, the difficulty with object relatives and the number interference effect should have occurred in the same regions and measures.

## General Discussion and Conclusion

Both Experiments 1 and 2 showed that object relatives are more difficult than subject relatives, which is in line with previous studies (Gibson, 1998; Gordon et al., 2001; Lewis, 1996; King & Just, 1991). This effect was found in the relative clause region (in regression path and total reading time in Experiment 1 and Experiment 2). However, neither Experiment 1 nor Experiment 2 found evidence for number interference in early measures. In Experiment 1, sentences were easier when the nouns had the same number than when their number was different. This is contrary to the predictions of similarity based interference accounts. In Experiment 2, there was evidence for number interference in object relatives, but this occurred during later processing, in the final region.

The observed delay of the number congruency effect in comparison to the relative clause effect in Experiment 2 is not consistent with memory interference models. They claim that difficulty with object relatives (compared to subject relatives) is due to interference that occurs when the subject and object noun phrase are retrieved at the embedded verb. Therefore, they predict that the number congruency effect should occur in the relative clause region, simultaneously with the slow-down with object relative clauses.

The results suggest that the interference effect is a later process that does not arise during structure building. One possibility is that it is due to a checking process that occurs

when a sentence is structurally complex as is the case with object relatives. Because this checking process occurs after the relative clause has been constructed, it may have been spread out over the regions following the relative clause in Experiment 1. Because the part of the sentence following the relative clause was long in Experiment 1, it may have been impossible to detect the later interference effect. When the main clause predicate was shortened in Experiment 2, we did find evidence for later number interference.

Together, the experiments suggest that the effect of retrieval cue overlap is weak. In addition to being weak, it occurs later than the relative clause effect, indicating that the initial difficulty with object relatives is not due to interference. These findings are more in agreement with theories that assume that processing difficulty with object relatives is not due to interference, such as the locality account (Gibson, 1998), experienced-based accounts (e.g., Wells, Christiansen, Rae, Acheson, & MacDonald, 2009) and expectation models (Levy, 2007; Hale, 2006).

## Acknowledgements

The research presented here was supported by the ESRC (Economic and Social Research Council).

## References

- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68, 1-76.
- Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(6), 1411-1423.
- Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4), 609-642.
- King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30, 580-602.
- Levy, R. (2007). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126-1177.
- Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, 25(1), 93-115.
- Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29, 375-419.
- Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48, 542-562.
- Van Dyke, J., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, 49, 285-316.
- Wells, J. B., Christiansen, M. H., Rae, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, 58, 250-271.



# The effect of text continuity on spatial representation

**Masashi Sugimoto (sugimoto.masashi.85m@st.kyoto-u.ac.jp)**

Department of Cognitive Psychology in Education,  
Graduate School of Education, Kyoto University  
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501 Japan

**Takashi Kusumi (kusumi@educ.kyoto-u.ac.jp)**

Department of Cognitive Psychology in Education,  
Graduate School of Education, Kyoto University  
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501 Japan

## Abstract

Two experiments examined the hypothesis that constructing spatial representation and making inference from it with route description requires text continuity. Participants read the spatial text and answered true/false questions about it. In Experiment 1, we transposed sentences in a spatial text, and in Experiment 2, we inserted irrelevant tasks into a spatial text. The results showed that performance in a route perspective decreases when text has lost its continuity. This decrease in performance was not found in a survey perspective. These results indicate the continuous nature of route perspective, not only at the surface level of description but also at the level of cognitive processing.

**Keywords:** route perspective; survey perspective; spatial mental models, spatial representation, text continuity

## Introduction

When we think about a space or when we are trying to follow directions, we construct spatial representations and infer spatial information from them. Taylor & Tversky (1992) defined two types of perspective in the input and output of spatial representation—route perspective and survey perspective. In route perspective, terms such as “front,” “back,” “left,” and “right” are used to give directions from the perspective of an imagined viewer (e.g., “When you get out of the building, you can see a supermarket in front of you.”). Survey perspective, however, includes terms such as “north,” “south,” “east,” and “west” to give directions, taking a bird’s eye view (e.g., “The building is north of the supermarket.”).

Many studies have pointed out the difference between these two perspectives. Of specific importance to the present investigation, some studies (Pazzaglia & Cornoldi, 1999; Pazzaglia, Meneghetti, De Beni, & Gyselinck, 2010) focused on the two components of visuo-spatial working memory (VSWM) in route perspective. They divided VSWM into two components: the spatial sequential process and the spatial simultaneous process. Spatial sequential tasks require participants to recall the order of the stimulus

presentation, while spatial simultaneous tasks require participants to recall the visual configuration of the presented stimulus (Pazzaglia et al., 2010). Their results showed that the spatial sequential process is more involved in processing route description, whereas the spatial simultaneous process is more involved in processing survey description. This implies that the ability to process information sequentially is an essential factor for descriptions in route perspective.

More support for this idea comes from a study using children with learning disabilities (Mammarella, Meneghetti, Pazzaglia, Gitti, Gomez, & Colnoldi, 2009). Children with nonverbal (visuo-spatial) learning disability (NLD), reading disability (RD), or no disability participated in the experiment. They listened to route, survey, and non-spatial descriptions. After that, they performed a verification task and a location task. Although their performance was no different in the verification task with the non-spatial description, children with NLD showed decreased performance on the verification task, especially with regard to the survey description. In the location task, children with NLD had decreased performance more on the survey than on the route description (though this difference did not reach significance). Mammarella et al. (2009) showed that NLD children can form mental models from route description. They indicated that this is due to the “serial nature of language” involved in the route perspective.

Previous studies have shown the importance of spatial-sequential ability in route perspective processing. This ability belongs to the participants, and not to description itself. Therefore, in this study, we focused on the nature of description that whether each sentence had strong connections between itself and the previous/following sentence. There are two reasons why we emphasized continuity of route description. First, in route description, the directional terms are relative. Therefore, it is important to be aware of where one has come from and which direction he or she is facing. If no attention is given to it, one can easily get lost because the directional terms must be defined in relation with the imagined viewer. Second, the subject of the route description is “you.” In addition to

actually moving around, the subject of the description cannot warp to a distant place. They must move step by step, continuously.

In this study, we focused on the text continuity itself. Our hypothesis was that effective route description requires text continuity. If route perspective description is truly continuous, sentence order (i.e., text continuity) is important for it. In contrast, survey perspective does not need continuity and sentence order is less important. We manipulated the text continuity in two ways. In Experiment 1, we changed the order of sentences in a spatial text, while in Experiment 2, we inserted an interference task into a more complex spatial text. The novel point of this study is that we focused not on the traits of participants (Brunyé & Taylor, 2008b; Pazzaglia, et al., 2010), but on the traits of the text.

## Experiment 1

In Experiment 1, we tried to examine the effect of continuity of sentence order on comprehension of route description. Although previous studies (Pazzaglia & Cornoldi, 1999; Pazzaglia, et al., 2010) have shown the importance of the spatial-sequential ability of learners in processing route description, the effect of the sentence itself has not examined. If the processing of route perspective is actually continuous, a sentence in the description must be connected to the previous and following sentences. Therefore, when this connection is broken, one faces considerable trouble learning information from route descriptions. We did not expect this effect in survey learning, because survey perspective is simultaneous and not sequential (Pazzaglia et al., 2010).

## Method

**Participants** 35 Japanese graduates and undergraduates (19 males and 16 females) participated in Experiment 1 for a monetary reward. Mean age was 22.5 (range 18-28,  $SD = 2.6$ ). We excluded three males from the analysis for not following instructions. Half of the participants studied all descriptions in the survey perspective, and the rest in route perspective.

**Experiment design** The design was  $2 \times 2 \times 2$  with learning perspectives (survey vs. route) as a between subjects factor, text continuity (continuous vs. discontinuous), and test perspective (survey vs. route), as within subjects factors.

**Materials** Twenty-eight spatial texts were prepared. Each text consisted of four sentences and described one environment where four landmarks (landmarks A, B, C, and D) appeared along a straight road. The first sentence referred to the position of one landmark (landmark A) in relation to the road. The second sentence referred to the spatial relationship between landmarks A and B. The third

and fourth sentences referred to the relationships between landmarks B and C, and landmarks C and D, respectively. In discontinuous condition, the order of the third and fourth sentences was reversed. Therefore, the third and fourth sentences referred to the relationships between landmarks C and D, and landmarks B and C, respectively.

Each text had six verification tasks that asked about the relationships between two landmarks. Half were correct descriptions and the rest were incorrect.

**Procedure** We instructed participants to read the spatial text as fast as possible. After participants had finished reading the text, they answered true/false verification tasks about the environment that they had just learned about. Continuous and discontinuous texts were presented in random order. Halfway between the trials, participants took a rest. All stimuli were presented on a PC screen.

## Results

Trials that included reading time beyond  $\pm 2$  SD or under one second were excluded from the analysis below. According to this criterion, 77.7% of all trials were used.

Fig. 1 shows the accuracy of the verification question, which asked about the spatial relationships between the landmarks appearing in the third sentence. We chose only the third sentence because it is the initial sentence that differs according to the text continuity. In both continuous and discontinuous condition, the first and the second sentence are same.

The results of ANOVA showed a significant interaction ( $F(1, 30) = 4.997$ ,  $p = .03$ ,  $\eta_p^2 = .14$ ) between text continuity  $\times$  test perspective. All other interactions did not reach significance. In the survey test conditions, the accuracy did not show a significant difference ( $F(1, 30) = 1.400$ ,  $p = .25$ ,  $\eta_p^2 = .05$ ). In the route test conditions, however, the accuracy was higher in the continuous condition than the discontinuous condition ( $F(1, 30) = 4.608$ ,  $p = .04$ ,  $\eta_p^2 = .13$ ). In addition, the accuracy was higher in the route test condition than in the survey test condition when text was continuous ( $F(1, 30) = 10.343$ ,  $p = .00$ ,  $\eta_p^2 = .26$ ). This difference, however, was not found in the discontinuous condition ( $F(1, 30) = 0.146$ ,  $p = .70$ ,  $\eta_p^2 = .00$ ).

## Discussion

As we predicted, the accuracy was higher in the continuous condition than in the discontinuous condition when participants used route perspective during the test. In contrast, performance did not show significant difference between continuous condition and discontinuous condition when they used survey perspective during the test. These results show that participants need text continuity when they recall spatial representation in route perspective. When participants recall the spatial relationships, they rely onto spatial representations which they had constructed before.

And whether the construction of the spatial representation was continuous or discontinuous, affect the spatial

representation itself.

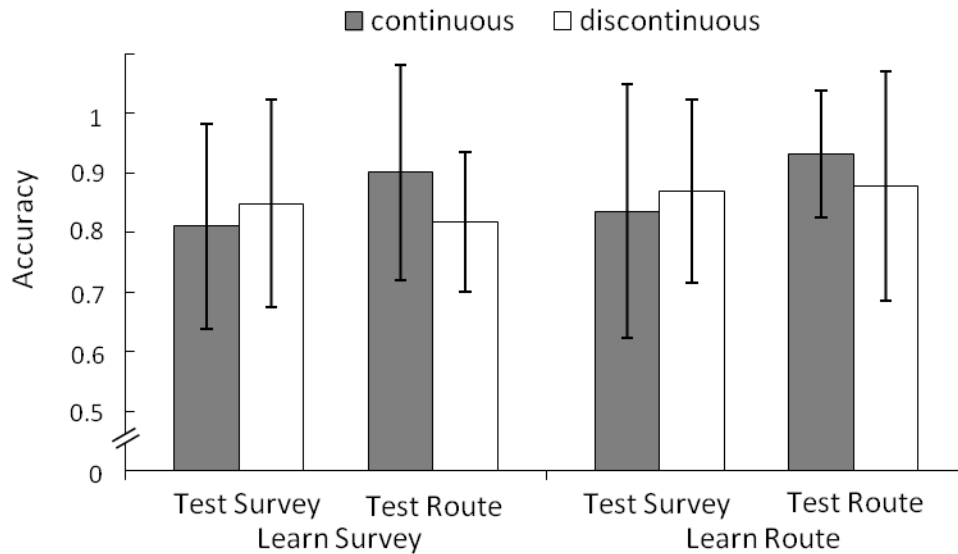


Fig. 1: Mean accuracy to the questions about the third sentence in Experiment 1 (bar means *SD*)

These differences however, appeared according to test perspective but did not according to learning perspective. This is not in line with our prediction. Although text continuity was a factor in learning, the performance differed according to the test perspective, rather than the learning perspective. We can think, however, that learning perspective has some effect on spatial representation. If spatial representations were the same regardless of learning perspectives, these differences would not appear because text continuity affects before the construction of spatial representation, not after. Therefore, constructed spatial representations should differ by the time participants construct it. One possible idea is that some factor lacks when participants learned the text in discontinuous conditions. They have to recall spatial information without the factor. When they recall in route perspective, the lack of the factor is make participants to have trouble in recall. When participants recall in survey perspective, however, the feature of survey perspective covers the lack of the factor. It is possible the factor is continuity of spatial representation.

One puzzling result is that participants showed better performance in route test than in the survey test when text was continuous. Previous studies showed superiority of survey perspective than route perspective in performance (Brunyé & Taylor, 2008a; Brunyé & Taylor, 2008b, Shelton & Gabrieli, 2002). We could not find this superiority of survey perspective in Experiment 1. There are two possibilities account for this tendency. One is that sentence order continuity works as a facilitator for route perspective, not that sentence order discontinuity works as an inference for route perspective. The other is that the studied

environments were too simple. It is possible that in a simple environment, participants need not to form abstract spatial representations, and it covers up the difference in learning perspective.

To solve these problems, in Experiment 2, participants studied a more complex text than that of Experiment 1. A complex text makes participants better infer spatial relationships according to their spatial representations. Therefore, the verification task would reflect their spatial performance more accurately.

## Experiment 2

In Experiment 2, as in Experiment 1, we examined the hypothesis that effective route description requires sentence continuity. We used the spatial texts used in Taylor & Tversky (1992). These texts were more complex than those used in Experiment 1, and participants had to make an inference about the environment. We manipulated text continuity by inserting irrelevant questions into the text. In the continuous condition we inserted short tasks that did not relate to the main text, yet still made participants conduct spatial inferences (such as “How many windows do you have in your room?” or “Which city is in the north, Kyoto or Nagoya?”). In the discontinuous condition, we inserted a simple counting task not to let participants rehearsal the text (“ $200 - 7 = ?$ ” or “ $100 + 8 = ?$ ”).

We made two predictions about the results. First, when participants learned the text from a survey perspective where the text did not need to be continuous, recall performance did not differ between the two conditions of

text continuity. However, when participants learned text from a route perspective where the text does need to be continuous, performance did differ between the two conditions.

## Method

**Participants** 67 Japanese graduates and undergraduates (34 males and 33 females) participated to Experiment 2 for a monetary reward. Mean age was 21.1 (range 18-25,  $SD = 1.8$ ). 33 participants (17 males and 16 females) studied all descriptions with a survey perspective, and 34 (17 males and 17 females) studied all with a route perspective.

**Experiment design** The design was  $2 \times 2 \times 2$  mixed, with learning perspectives (survey vs. route) as a between subjects factor, with text continuity (continuous vs. discontinuous) and test perspective (survey vs. route) as within subjects factors.

**Materials** Three tasks were conducted. Spatial text learning, Corsi blocks (Corsi, 1972) and the pathway span test (Mammarella, Cornoldi, Pazzaglia, Toso, Grimoldi, Vio, 2006). All tasks are conducted on the PC screen.

We used two spatial texts (town and convention center) from Taylor & Tversky (1992). Each text described an environment from two perspectives—survey and route. Each text had 28 True/False verifications: four questions were non-locative recognition, four were non-locative paraphrased, four were survey recognition, four were route recognition (survey and route recognition questions required inferences when study perspective and test perspective differed), six were survey inference, and six were route inference. In each category, three statements were true and the rest were false.

When participants learned the text, we inserted irrelevant tasks in every three sentences. In discontinuous condition as an experimental condition, we inserted spatial questions which are irrelevant to the main spatial text. Participants had to infer spatial relationships or to recall spatial alignment of objects which does not appear in the main text. We instructed participants to answer in five seconds. After five seconds passed, they return to the learning of the main text independently of the fact they answered to the inference questions or not, and the answer is correct or not.

In continuous condition as a control condition, we inserted simple counting tasks. We used counting tasks to prevent participants from rehearsal of the main spatial text. We instructed participants to repeat answering by five seconds passed (e.g. 93, 86, 79 ...). After the five seconds passed, they return to the learning of the main text.

In the Corsi blocks task, participants memorized the order of the position where a dot appeared. The number of stimuli in one trial was from four to seven, and there were twelve trials. This task measured spatial-sequential ability. Studies

have found positive relationships between this task and route perspective performance (Mammarella et al., 2006; Pazzaglia & Cornoldi, 1999; Pazzaglia, et al., 2010).

In the pathway span task, participants were told to follow movement in a five by five matrix according to the direction instructions. The number of instructions in one trial ranged from four to seven, with twelve trials in total. This task also measured spatial-sequential ability.

## Procedure

First, participants conducted the Corsi blocks task. Then they were allocated to either the survey study condition or the route study condition, as performance on the Corsi block task did not differ between conditions. Next, participants read two spatial texts. One text was read for the continuous conditions and the other for the discontinuous conditions. After they read one spatial text, they answered 28 true/false questions about each text. Finally, they conducted the pathway span test. All stimuli were presented on a PC screen.

## Results

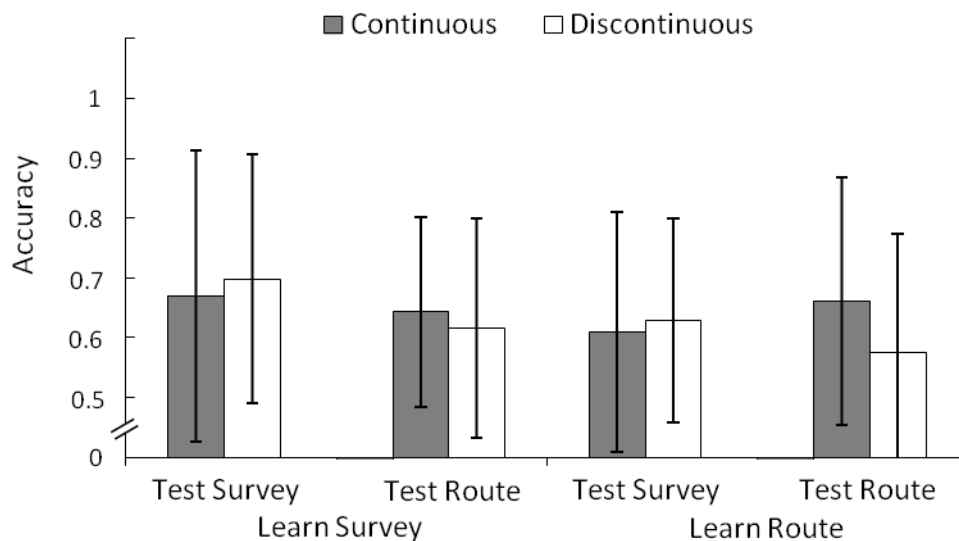
One male was excluded from the analysis because he did not follow instructions. Another male was excluded because his verification performance in one condition was much lower (19.4%), although chance level was 50%. Both of them studied in the survey perspective. Verbatim questions that included reaction times beyond  $\pm 2$  SD were excluded from the analysis. According to this criterion, 97.0% of all trials were used.

Fig. 2 shows the accuracy of the verification task where participants needed spatial inferences. The results of the ANOVA showed a marginally significant interaction between text continuity  $\times$  test perspective ( $F(1, 65) = 3.769, p = .06, \eta_p^2 = .05$ ). In the continuous condition, the accuracy in survey and route condition did not showed significant difference ( $F(1, 65) = 0.219, p = .64, \eta_p^2 = .00$ ). In the discontinuous condition, participants showed lower accuracy in the route test than in the survey test ( $F(1, 65) = 5.614, p = .02, \eta_p^2 = .08$ ). In the survey test condition, the accuracy in the continuous condition and discontinuous condition did not showed significant difference ( $F(1, 65) = 0.663, p = .41, \eta_p^2 = .01$ ). In route test condition, participants showed lower performance in the discontinuous condition. The effect size of this difference is not small. This difference, however, did not reach to significance ( $F(1, 65) = 2.428, p = .12, \eta_p^2 = .04$ ). We conducted ANCOVA which controlled for the scores either Corsi blocks task or pathway span test or both of them. Those analyses, however, showed no statistical significance.

## Discussion

As predicted, when text lost its continuity, participants in the route test condition decreased performance compared to

requires text continuity was partially supported in Experiment 2. The discontinuous text lowered the accuracy



the survey condition. Our hypothesis that route description

when participants were tested in the route perspective. The

Fig. 2: Mean accuracy to the spatial inference questions in Experiment 2 (bar means *SD*)

effect was, however, limited and some predicted results did not reach to significant level.

As in the Experiment 1, the effect of text continuity appears according to the test perspective. We have a good reason, however, to believe text continuity affect how we learn the spatial information as we stated in the discussion of Experiment 1.

Previous research (Pazzaglia & Cornoldi, 1999; Pazzaglia, et al., 2010) has shown the importance of the spatial-sequential ability of learners in processing route description. This study shows that not only the spatial-sequential ability of the readers but also the text continuity itself is important to route description. This result reveals the more continuous nature of the route perspective than the survey perspective. In route description, it is clear that one must remember from where he or she has come, and which direction he or she is facing. The result of this study indicates that text continuity is essential, not only to description itself but also to the mental processing of route description.

## General Discussion

We conducted two experiments and examined the effect of text continuity in route perspective. In Experiment 1, we manipulated text continuity by transposing the sentences. In the test phase, text continuity increased performance of the route perspective. In Experiment 2, we manipulated the continuity by inserting irrelevant tasks into the more complex spatial descriptions than Experiment 1. Performance on the route test decreased in the discontinuous condition where participants inferred spatially. These results

support our hypothesis that route perspective needs text continuity.

Previous studies have found that spatial sequential ability is necessary for learning route description (Pazzaglia & Cornoldi, 1999; Pazzaglia, et al., 2010). In this study, we found that not only spatial-sequential ability but also text continuity is essential for understanding route descriptions. This dependence on continuity seems to relate to the nature of route description, as readers must continuously update changes in their local environment (Shelton & Gabrieli, 2002).

This study revealed that text continuity affects the retrieving of spatial information in route perspective. It remains unrevealed, however, when text continuity affect route perspective. There can be two possibilities about it. There was a significant interactions between test perspective  $\times$  text continuity. There is no doubt, therefore, text continuity affect at the test. In this case, spatial representation formed in discontinuous condition lacks continuity regardless of the learning perspective. When one uses route perspective to retrieve information from that discontinuous representation, a problem occurs. He or she can't rely onto continuity of the representation, has to navigate in his/her spatial representation discontinuously and the performance decreases. When one uses survey perspective at the test, whether the spatial representation is continuous or discontinuous does not matter. He or she can successfully retrieve information from his/her spatial representation even when it is discontinuous. Although the interaction between study perspective and text continuity was not statistically significant, this does not necessarily

denies the possibility of the effect of text continuity at learning. It is possible that the effect of text continuity at learning exists, however, is too weak to affect at the learning.

We refer to three remaining problems in this study. First, we used spatial text to present stimuli to participants in this study. Text and languages are naturally continuous and the results of this study might appear only when one studies one's environment using language. Therefore, we must confirm these results through other forms of studying, such as navigation in reality or watching videos involving specific locations. Second, we have to solve problems about the types of two inference tasks in Experiment 2. We regarded spatial question as an inference task and counting tasks as a control one. This distinction, however, is relatively arbitrary. It is possible that these tasks differed in types of inference, not in continuity. To solve this problem, we may need another control condition which is different from ones in continuity. In addition to that, forming spatial representation and extract information from that is quite complex process and are thought to be affected by individual difference. Participants take many strategies and their abilities differ quite a large way (Kato & Takeuchi, 2003). Therefore, it is possible that individual difference and strategy preference affect the results and covers up the effects of some factors. In Experiment 2, almost all results showed the same directions with the predictions. The statistical analysis, however, showed only few of them. Therefore, the next step is to take into account individual difference and the strategy preference.

### Acknowledgement

This research was supported by Grant-in-Aid for JSPS Fellows.

### Reference

- Brunyé, T. T., & Taylor, H. A. (2008a). Extended experience benefits spatial mental model development with route but not survey descriptions. *Acta Psychologica*, **127**, 340-54.
- Brunyé, T. T. Taylor, H. A. (2008b). Working memory in developing and applying mental models from spatial descriptions. *Journal of Memory and Language*, **58**, 701-729.
- Corsi, P. M. (1972). Human memory and the medial temporal region of the brain. Unpublished doctoral dissertation, McGill University, Montreal.
- Kato, Y., & Takeuchi, Y. (2003). Individual differences in wayfinding strategies. *Journal of Environmental Psychology*, **23**, 171-188.
- Mammarella, I. C., Cornoldi, C., Pazzaglia, F., Toso, C., Grimoldi, M., & Vio, C. (2006). Evidence for a double dissociation between spatial-simultaneous and spatial-sequential working memory in visuospatial (nonverbal) learning disabled children. *Brain and Cognition*, **62**, 58-67.
- Mammarella, I. C., Meneghetti, C., Pazzaglia, F., Gitti, F., Gomez, C., & Cornoldi, C. (2009). Representation of survey and route spatial descriptions in children with nonverbal (visuospatial) learning disabilities. *Brain and Cognition*, **71**, 173-179.
- Pazzaglia, F., & Cornoldi, C. (1999). The role of distinct components of visuo-spatial working memory in the processing of texts. *Memory*, **7**, 19-41.
- Pazzaglia, F., Meneghetti, C., De Beni, R., & Gyselinck, V. (2010). Working memory components in survey and route spatial text processing. *Cognitive Processing*, **11**, 359-369.
- Shelton, A. L., & Gabrieli, J. D. E. (2002). Neural correlates of encoding space from route and survey perspectives. *The Journal of Neuroscience: the official journal of the Society for Neuroscience*, **22**, 2711-2717.
- Taylor, H. A., & Tversky, B. (1992). Spatial Mental Models Derived from Survey and Route Descriptions. *Journal of Memory and Language*, **29**, 261-292.

# Stress assignment in the development of reading aloud: Nonword priming effects on Italian children

**Simone Sulpizio (simone.sulpizio@unitn.it)**

Fondazione De Vincenzi ONLUS

Department of Cognitive Science and Education, University of Trento, corso Bettini 31  
38068 Rovereto (Tn), Italy

**Magali Boureux (magali.boureux@univr.it)**

Department of Philosophy, Pedagogy and Psychology, University of Verona, Lungadige Porta Vittoria 17,  
37129 Verona, Italy

**Cristina Burani (cristina.burani@istc.cnr.it)**

Institute of Cognitive Sciences and Technologies, CNR, Via S. Martino della Battaglia 44,  
00185 Roma, Italy

**Chizuru Deguchi (chizuru.deguchi@unitn.it)**

Department of Cognitive Science and Education, University of Trento, corso Bettini 31  
38068 Rovereto (Tn), Italy

**Lucia Colombo (lucia.colombo@unipd.it)**

Department of General Psychology, University of Padua, Via Venezia 8  
35131 Padua, Italy

## Abstract

Two experiments investigated the development of two aspects related to stress assignment in reading. First, we tested whether the role of distributional knowledge concerning stress changes with the development of the reading system; second, we tested whether stress information is computed independently of phonemic information since the first stages of reading acquisition. We ran two identical experiments in Italian, one with children of two age levels (second and fourth grades) and one with adults. Results showed that older children behave similarly to adults, but younger children do not. Differently from the advanced readers, younger children use more general distributional knowledge about stress and are not able to compute stress information apart from phonemes. Taken together, our results suggest that the stress subsystem, and in particular the mechanisms working at the level of the phonological buffer are not fully developed during the first stages of reading.

**Keywords:** lexical stress; stress neighborhood; reading development; pathway priming.

## Introduction

The process of stress assignment has recently become a central issue in the reading aloud literature. In languages like English, Dutch, or Italian, with a lexicon composed of polysyllabic words and with unpredictable stress, understanding how people read words aloud implies, among other things, understanding how readers assign stress. Research on stress assignment in reading has mainly focused on two issues: First, how readers assign stress to words and nonwords (*e.g.*, Colombo, 1992; Rastle &

Coltheart, 2000); Second, how suprasegmental information is represented in the reading system (Colombo & Zevin, 2009; Sulpizio, Job & Burani, in press b).

Let us consider the mechanisms for stress assignment first. Although stress assignment may not cause problems in reading words – readers may retrieve stress information as part of their lexical knowledge – it causes difficulties when stimuli are unknown words or nonwords. Thus, one issue is how readers are able to assign stress with no reference to lexically-stored information (Sulpizio, Arduino, Paizi, & Burani, in press a). Within a connectionist view of reading, studies in English and Italian have highlighted that two different types of distributional knowledge drive readers in stress assignment to unknown stimuli. First, the distribution of the stress patterns in the lexicon – *e.g.*, in Italian, 80% of three-syllabic words bear penultimate stress (maTIta<sup>1</sup>, pencil), while 18% bear antepenultimate stress (BIbita ‘drink’) (Thornton, Iacobini, & Burani, 1997)<sup>2</sup>. This may induce readers, following a default stress bias, to assign the most common pattern to unknown words (Colombo, 1992; Rastle & Coltheart, 2000). Second, some orthographic/phonological units work as cues for stress assignment, with word ending being a strong predictor of the stress pattern (*e.g.*, Arciuli, Monaghan, & Ševa, 2010; Kelly, Morris, & Verrechia, 1998). Consider Italian as an

<sup>1</sup> Capital letters indicate stressed syllable.

<sup>2</sup> The remaining 2% of three-syllabic words bear stress on the final syllable, and in this case stress is graphically marked (*e.g.*, colibri ‘hummingbird’)



example: Most of the words ending in -ola bear antepenultimate stress (PENTola, 'pot'; BAMbola, 'doll' *etc.*), that is, -ola has a stress neighborhood composed of many antepenultimate stress friends. Following this statistical tendency, when readers see an unknown word or a nonword ending in -ola, they will be prone to assign antepenultimate stress according to its stress neighborhood (Burani & Arduino, 2004; Colombo, 1992; Sulpizio *et al.*, in press a). Thus, readers have two sources of information to assign stress, *i.e.*, the lexical knowledge and the distributional information driven by their implicitly acquired statistical knowledge.

The representation of stress within the reading system has been investigated only recently. Two studies on Italian have shown that suprasegmental information may be partially independent from segmental information: Using a priming methodology, both studies found that the word's stress pattern can be primed independently of its segmental level (Colombo & Zevin, 2009; Sulpizio *et al.*, in press b). This finding is in line with the word production literature (Levelt, Roelofs, & Meyer, 1999) – where stress is part of an abstract metrical representation including the number of syllables and stress position – and with the view that speech production and reading aloud may share, at least in part, the last stages of processing, *i.e.*, phonological and phonetic word encoding (Roelofs, 2004). Thus, stress information would be partially independent of phonemic information and readers could compute the former independently from the latter.

But what about young readers? When a child starts to read, her/his reading system is not fully developed, her/his lexicon may be relatively small, and her/his knowledge of the statistical properties of the language may be relatively limited. How can the development of the stress system be characterized? Does knowledge of stress properties and its application to reading require time to develop? The issue concerning the development of distributional knowledge for stress has received little attention (but see Arciuli *et al.*, 2010, discussed below). To our knowledge, few studies investigated how a metrical representation, autonomous from segmental information, develops in young readers. Colombo, Deguchi and Boureux (submitted) found that young (7-years old) children were little affected by the stress pattern of the priming context nonwords in reading nonword targets, while priming was significant in older children.

In the present paper we further investigated the issue of whether children, when starting to read, are able to use the autonomous representation of stress as adult readers and thus may show stress priming effects. Differently from Colombo *et al.* (submitted), we used real words as targets. Words have a lexical representation which includes stress position. Consequently, if the words' stress pattern is retrieved from lexical memory, this information may be automatically available before information from the prime has any effect, particularly in children whose reading processes are relatively slow. On the other hand, Sulpizio *et*

*al.* (in press b) found significant stress priming for words in adults. Thus the question of whether and in which conditions we may find stress priming for words in adults and children is still open.

We ran two identical experiments, one with children of two age levels (II and IV grade) and one with adults. We adopted the "pathway priming" paradigm (Colombo & Zevin, 2009; Zevin & Balota, 2000) to test the possibility of inducing stress priming not only in adults, but also in young readers. In this paradigm, each target is preceded by five nonword primes with the same stress pattern that act as a small list context for the target. These micro-lists were included in a larger list in which all primes were homogeneous for stress. Participants have to read all stimuli aloud and they are not aware that some stimuli are primes and others are targets. By manipulating the congruency between primes' stress and target's stress – the target could have the same stress as the five preceding primes or a different one – we investigated whether a priming effect would occur, with participants being facilitated when reading a target in the congruent stress condition (when the target had the same stress as its five preceding primes), compared to the incongruent stress condition (when the target had a different stress than its five preceding primes). We expected that the stress priming effect would be stronger in adult readers and (perhaps) in older children, than in younger children.

Colombo and Zevin (2009) showed that stress priming effects are stronger within a sub-lexical context. Consequently, we only used nonwords as primes. In this way, participants were strongly encouraged to rely on sub-lexical reading. The use of nonword stimuli allowed us to further test the second issue: How readers develop distributional knowledge for stress assignment. Arciuli and colleagues (2010) ran a nonword reading experiment with English-speaking children. They found that younger (5/6-year-old) children were more affected by the main distribution of stress patterns in English – they assigned stress to the initial syllable more frequently – and were not affected by the final part of the nonwords. By contrast, in older children (7/8-year-old) the bias toward initial stress became weaker and children were influenced by specific orthographic cues, such as the nonwords' ending. A similar pattern of results was found in Italian. While young children were more prone to assign dominant stress to nonwords and less affected by orthographic neighborhood than older children and adults (Colombo *et al.*, submitted), Italian adults and older children (11 years old) assigned stress to nonwords and low-frequency words on the basis of stress neighborhood, showing very weak evidence for a bias toward the penultimate (dominant) stress in reading words aloud (see also Burani & Arduino, 2004; Colombo & Zevin, 2009; Paizi, Zoccolotti, & Burani, 2011). If these findings reflect cognitive constraints holding for different languages, we might expect similar trajectories in the development of distributional knowledge for stress in English and Italian: Older children and adults should be more affected by stress

neighborhood, while younger children might be more influenced by the distributional bias toward the dominant stress.

In summary, the present study investigated two issues related to the development of stress assignment in reading. First, we investigated whether stress assignment is fully developed already in young readers so that, when reading a word aloud, they are able to exploit the prosodic information available from the context. Second, we investigated whether the trajectory of the development of stress assignment in Italian is similar to what Arciuli *et al.* (2010) found for English, *i.e.*, that distributional knowledge of stress moves from a general distributional bias to more subtle statistical properties, such as stress neighborhood. To test these issues we ran two reading experiments in Italian, a language in which stress is not predictable by rule (Krämer, 2009) and there are two main stress patterns asymmetrically distributed, *i.e.*, penultimate stress – which is the dominant pattern and appears in 80% of words – and antepenultimate stress – which appears only in 18% of words. The two experiments were identical, except that one was run with children of different grades and one with adults.

## Methods

### Experiment 1 - Children

**Participants.** Two groups of elementary school children took part in the experiment: The first group included 20 second graders (13 males, mean age: 7.1, sd: 0.3); the second group included 18 fourth grade children (10 males, mean age: 9.2, sd: 0.4). All participants were native Italian speakers, with normal or corrected-to-normal vision.

**Materials & Method.** Thirty three-syllabic antepenultimate-stress words were used as target stimuli. All words had a low frequency (mean frequency: 15.93, sd: 27.96, out of 1.5 million occurrences, Barcelona Corpus, Istituto di Linguistica Computazionale, 1989, unpublished manuscript). The phonological lexical representation of low frequency words is less likely to be automatically retrieved, and thus priming should be easier to obtain. Two sets of three-syllabic phonologically legal Italian nonwords were used in order to create two stress priming contexts. They were constructed in such a way that one set should receive penultimate stress and the other one should receive antepenultimate stress according to the stress neighborhood consistency. Thus, nonwords having the nucleus of the penultimate syllable and the last syllable typical of words stressed with the penultimate (dominant) stress pattern (*e.g.*, -ato in geLAto, “ice cream”) were defined as penultimate-stress primes, whereas nonwords having the nucleus of the penultimate syllable and the last syllable typical of words stressed on the antepenultimate syllable (*e.g.*, -olo in TAvoLo “table”) were defined as penultimate-stress primes. To examine the efficacy of stress neighborhood consistency based on word-ending, we carried out a pre-test, by presenting to adult readers the nonwords in a random order

in a word naming paradigm. The pre-test showed that stress was assigned to the nonwords according to their ending: 72% of nonwords with penultimate-stress neighborhood received penultimate stress and 78% of nonwords with antepenultimate-stress neighborhood received antepenultimate stress. Penultimate- and antepenultimate-stress nonwords were matched on: Length in letters (mean: 6.2 [sd: 0.5] vs. 6.3 [sd: 0.6]); number of consonant clusters (mean 0.3 [sd: 0.4] vs. 0.4 [sd: 0.5]) and initial phonemes.

Fifty penultimate- and 50 antepenultimate-stress nonwords were selected as primes. Target words were divided into 3 sets (10 each). Ten targets were assigned to the penultimate prime list, 10 targets to the antepenultimate prime list, and 10 targets were paired with a set of two simple geometric figures, square and triangle, each repeated 5 times. The latter condition was included to preserve interest and attention to the reading task in children, and was kept similar in the experiment with adults to have a perfectly matched control experiment. Stimuli preceded by figure primes were considered fillers, and not analyzed. Each target was preceded by 5 primes, as in Colombo & Zevin (2009; *cf.* Zevin & Balota, 2000). All target words had antepenultimate stress and they were congruent with the antepenultimate prime list, whereas they were incongruent with the penultimate prime list. The three different target lists were presented between-participants.

**Apparatus & Procedure.** The monitor was in VGA color. A voice key connected to the PC's real-time clock collected response latencies. The experiment was run using E-Prime software (Psychology Software Tools, Pittsburgh, PA). Participants were tested individually. They were instructed to read aloud each stimulus as fast and as accurately as possible. Stimuli were presented on the computer screen. In each trial, a fixation point was presented for 300 ms, followed by the stimulus in black color. At the start of articulation the letter string turned in red when the voice key responded. Response time was measured from the onset of the stimulus to the onset of articulation. The stimulus remained on the screen until the experimenter coded each trial as correct or as an error (stress or phonemic error) by pressing one of the keys of the keyboard. Advancement of trials was made by the experimenter, as soon as response coding was done. When the letter string disappeared, the next trial started immediately. Participants were audio-recorded to allow a further verification of experimenter's evaluation. Each participant received the two priming lists in separate blocks in a counterbalanced order; half of the participants received the penultimate-stress prime list first, while the other half was presented the antepenultimate-stress prime list first. The experimental blocks were preceded by a practice session with stimuli not included in the experimental trial.

## Experiment 2 – Adults

**Participants.** Twenty-four participants (6 males, mean age: 22.6 sd: 1.3) took part in the experiment. They were all Italian native speakers and had normal or corrected-to-normal vision.

**Materials, Method, Apparatus & Procedure.** The same as in Experiment 1.

## Results

### Experiment 1 - Children

**Prime analysis.** Overall nonword primes were read consistently with their stress neighborhood and this was true for both second graders and fourth graders (Table 1). However, younger children assigned penultimate (dominant) stress significantly more often than older children, especially when reading nonwords with antepenultimate-stress neighborhood (younger children assigned penultimate stress to 56% of the nonwords whereas older children assigned penultimate stress to 52% of the nonwords) and the difference was significant ( $\chi^2 = 4.16$ ,  $p < .05$ ).

Table 1: Children. Percentages of nonwords read with each stress pattern for each class.

	II grade		IV grade	
	Penult. stress	Antepen. stress	Penult. stress	Antepen. stress
Penultimate stress neighborhood	74%	26%	79%	21%
Antepenultimate stress neighborhood	37%	63%	23%	77%

#### Target analysis.

Responses shorter than 250 ms or longer than 3500 ms (5.8% of all data points) were excluded from the analyses. Naming times and errors were both analyzed using mixed-effects models, with class (II grade vs. IV grade) and stress congruency (primes and target sharing the same stress vs. primes and target with different stress) as fixed factors. Participants and items were treated as random factors. The models were fitted using the *lmer* function (*languageR* package, Baayen, Davidson, & Bates, 2008) in R software (version 2.11). Results for errors are reported in Table 2.

**Naming times.** The mixed-effects model was run with naming latencies as dependent variable and class and stress congruency as predictors. The model showed a main effect of class ( $\beta = -0.33$ , st. err. = 0.11,  $t = -2.89$ ,  $p < .01$ ), with fourth graders faster than second graders. No other effect reached significance (stress congruency:  $t = 1.5$ ,  $p > .1$ ; class by stress congruency interaction:  $t < 1$ ,  $p > .5$ ).

**Errors.** Two analyses were run, one considering stress errors and the other one considering phonemic errors. Both analyses were performed with accuracy as dependent

variable and class and stress congruency as predictors. When considering stress errors, the mixed-effects model showed a significant interaction between class and stress congruency ( $\beta = 1.36$ , st. err. = 0.4,  $z = 3.39$ ,  $p < .01$ ), with the effect of stress congruency being significant only for fourth graders. No other effect reached significance (both  $z$ ,  $< 1$ ). Differently, when considering phonemic errors, only the effect of stress congruency approached significance ( $\beta = 0.57$ , st. err. = 0.3,  $z = 1.89$ ,  $p < .1$ ), with more errors in the incongruent stress condition. No other effect reached significance (class:  $z = 1.3$ ,  $p > .1$ ; class x prime interaction:  $z = -1.4$ ,  $p > .1$ ). Inspection of means shows that the congruency effect was mainly due to second-graders (Table 2).

Table 2: Children. Mean error percentages for the targets in the two stress-prime conditions.

	Stress errors		Phonemic errors	
	II grade	IV grade	II grade	IV grade
Congruent stress prime	28%	13%	16%	17%
Incongruent stress prime	29%	31%	25%	18%

### Experiment 2 – Adults

**Prime analysis.** Participants read nonword primes according to their stress neighborhood, with no tendency to overuse the penultimate stress (penultimate and antepenultimate stress were assigned 53% and 47% of the times, respectively).

#### Target analysis.

Responses shorter than 250 ms or longer than 1500 ms (4.1% of all data points) were excluded from the analyses. Naming times and errors were both analyzed using mixed-effects models, with stress congruency (primes and target sharing the same stress vs. primes and target with different stress) as fixed factor. Participants and items were treated as random factors.

**Naming times.** The mixed-effects model was run with naming latencies as dependent variable and stress congruency as predictor. The effect was not significant ( $t = -1.1$ ,  $p > .1$ ).

**Errors.** Stress errors were few, consequently they were analyzed together with phonemic errors. A mixed-effects model was performed with response accuracy as dependent variable and stress congruency as predictor. The effect of stress congruency (congruent stress prime = 4% errors; incongruent stress prime = 10% errors), was significant ( $\beta = 1.38$ , st. err. = 0.50,  $z = 2.76$ ,  $p < .01$ ): Readers were more accurate when primes and target shared the same stress than when primes and target had a different stress.

## Discussion

In two reading experiments, we tested what kind of distributional information younger and older children use

for stress assignment, and whether stress priming affects both children and adults in reading Italian aloud. The results show two main findings: First, while both adults and older children assign stress to non-words (prime stimuli) according to their stress neighborhood, younger children do exploit stress neighborhood, but they also show a tendency to overgeneralize penultimate (dominant) stress in assigning stress. Second, participants were more accurate to read a target when it was preceded by a set of primes with the same stress (both primes and target received antepenultimate stress), than when it was preceded by a set of primes with a different stress (primes received penultimate stress and target antepenultimate stress). This pattern was found with older children and adult readers, but not with younger children. One might argue that reading is a very different task for younger and older children and that this fact might be sufficient to explain the data. Although we do not exclude any effect of task familiarity, we believe that a better explanation may refer to how the reading system develops and what lexical/sub-lexical strategies children adopt (*cf.* Peressotti, Mulatti, & Job, 2010).

In interpreting this pattern of results, we should consider the development of distributional knowledge first. Previous research has shown that stress neighborhood can be considered the main factor able to drive stress assignment in adult readers (Arciuli *et al.*, 2010; Burani & Arduino, 2004; Kelly *et al.*, 1998; Protopapas, Gerakaki, & Alexandri, 2006) and in connectionist simulations of Italian (Pagliuca & Monaghan, 2010). Some studies have suggested that also the bias toward the dominant stress pattern in the language might play a role in stress assignment, but only when readers heavily rely on a sub-lexical procedure (Colombo & Zevin, 2009; but see also Protopapas *et al.*, 2006, for a different perspective in languages other than Italian). Similar to what was found by Arciuli *et al.* (2010) with English-speaking children, our study suggests that readers may use two types of information for stress assignment, but their relative employment changes during the acquisition of reading. Differently from older children and adults, who assign stress on the basis of stress neighborhood alone, Italian young children assign stress to nonwords not only on the basis of stress neighborhood, but also of the distributional bias toward the dominant (penultimate) stress in their native language. This might be due to at least two reasons: First, children are more prone to use sub-lexical reading (Ziegler & Goswami, 2005), thus increasing the chance to apply the distributional general bias. Second, in developing their lexicon and the ability to analyze orthographic and phonological information, children might develop their distributional knowledge by discovering more subtle correspondences between word orthography and stress pattern. Thus, the development of distributional knowledge would follow a trajectory that goes from the more general bias toward the dominant stress to the more specific stress neighborhood, which may require time to develop.

Let us now consider the stress priming effect. Our results are in line with previous research that found stress priming effects in reading (Colombo & Zevin, 2009; Sulpizio *et al.*, in press b). The fact that the stress pattern of a word can be primed confirms the idea suggested in word production models that, when reading a word aloud, participants compute suprasegmental information independently of segmental information (Levelt *et al.*, 1999) instead of retrieving the word's prosodic information from the lexicon. Accordingly, we can assume that these two types of information are first computed separately by means of specific mechanisms, and then assembled together prior to word articulation. However, while both older children and adults show stress priming effects, younger children do not. This difference could be explained by assuming that, although the prosodic system of younger readers is fully developed (Juszyk, Houston, & Newsome, 1999), it is not yet able to exploit prosodic information from the context, namely, from the primes' stress pattern. A similar interpretation is supported by the pattern of phonemic errors made by young readers, which decrease in the congruent prime condition: This pattern suggests that the prime affects the target's computation, but only at the segmental level; differently, younger readers are not able to use stress information driven by the primes, as no effect of the primes' congruency emerged at the level of stress errors.

The CDP++ model of reading aloud (for English words; Perry, Ziegler, & Zorzi, 2010) can account for our data fairly easily. The model assumes that stress information can be both lexically retrieved and sub-lexically computed. At the sub-lexical level a connectionist network maps graphemes onto phonemes and the orthographic input onto a stress pattern. Moreover, the model assumes that the phonological output buffer includes two different components, namely the *stress output nodes* and the *phonemic output nodes*: The former are responsible for stress assignment and the latter for phoneme activation. During training the network may learn to associate specific orthographic cues with a specific stress (*e.g.*, the final sequence *-ola* with antepenultimate stress). Thus, the probability that a pseudoword will receive a certain stress depends on the strength of the connections established between the orthographic cues and stress position: The more frequent the association between an orthographic cue and a stress pattern, the stronger the connection between stress and orthography.

The assignment of stress according to stress neighborhood may emerge at the sub-lexical level, with the word final sequence driving stress assignment. This mechanism might work less efficiently in young children who are learning to read, because they have to learn which orthographic sequences may work as a strong cue for stress. Thus, young children might assign stress to nonwords on the basis of a more general distributional tendency such as the bias toward the dominant stress pattern in the language. The stress priming effect may occur at the level of the *stress output nodes*: When planning the target's articulation, readers

might be affected by the repeated pre-activation of the primes' metrical structure, which can be congruent or incongruent with the target. Thus, stress may be primed at the level of the phonological output buffer, when readers assemble the phonological unit that has to be articulated. Finally, the absence of a stress priming effect in younger children might be due to the absence of a fully developed stress system in the first stages of reading.

To conclude, the present study has shown two important aspects of the developmental trajectory of stress assignment in reading. First, when reading a stimulus aloud, readers make use of their distributional knowledge to assign stress and they do it since the first stages of reading development. However, the type of distributional knowledge exploited by readers changes developmentally: While younger readers are more prone to use general knowledge about the dominant stress pattern in the language, older readers are more affected by more specific distributional knowledge, namely stress neighborhood. Second, and more important, stress information can be computed separately from phonemic information, but young children are more likely to use information that has been extensively acquired - (*i.e.*, the dominant pattern in the language), rather than information gathered from the context (*i.e.*, the priming list). Taken together, the present findings suggest a final conclusion: At the early stages of reading, the word stress assignment subsystem appears partially underdeveloped.

### Acknowledgments

The study was carried out with the support of grant PRIN 2009 from the Italian Ministry of Education and Research, and grant "Progetti di Ateneo Bando 2007" to L. Colombo. We thank Josephine Trippi for running the experiments.

### References

- Arciuli, J., Monaghan, P., & Ševa, N. (2010). Learning to assign lexical stress during reading aloud: Corpus, behavioral, and computational investigations. *Journal of Memory and Language*, 63, 180-196.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.
- Burani, C., & Arduino, L. S. (2004). Stress regularity or consistency? Reading aloud Italian polysyllables with different stress patterns. *Brain and Language*, 90, 318-325.
- Colombo, L. (1992). Lexical stress effect and its interaction with frequency in word pronunciation. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 987-1003.
- Colombo, L., Deguchi, C., & Boureau, M. (submitted). Stress priming in Italian nonword reading.
- Colombo, L., & Zevin, J. (2009). Stress Priming in Reading and the Selective Modulation of Lexical and Sub-Lexical Pathways. *PLoS ONE*, 4, e7219.
- Istituto di Linguistica Computazionale, CNR, Pisa, Italy (1989). *Corpus di Italiano scritto contemporaneo*. Unpublished manuscript.
- Juszyk, P. W., Houston, D., & Newsome, M. (1999). The beginning of word segmentation in English-learning infants. *Cognitive Psychology*, 39, 159-207.
- Kelly, M. H., Morris, J., & Verrekia, L. (1998). Orthographic cues to lexical stress: Effects on naming and lexical decision. *Memory and Cognition*, 26, 822-32.
- Krämer, M. (2009). *The phonology of Italian*. Oxford: Oxford University Press.
- Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, 1-75.
- Paizi, D., Zoccolotti, P., & Burani, C. (2011). Lexical stress assignment in Italian developmental dyslexia. *Reading & Writing*, 24, 443-461.
- Pagliuca, G., & Monaghan, P. (2010). Discovering large grain-sizes in a transparent orthography: Insight from a connectionist model of reading aloud for Italian. *European Journal of Cognitive Psychology*, 22, 813-835.
- Peressotti, F., Mulatti, C., & Job, R. (2010). The development of lexical representations: Evidence from the position of diverging letter effect. *Journal of Experimental Child Psychology*, 106, 177-183.
- Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology*, 61, 106-151.
- Protopapas, A., Gerakaki, S., & Alexandri, S. (2006). Lexical and default stress assignment in reading Greek. *Journal of Research in Reading*, 29, 418-432.
- Rastle, K., & Coltheart, M. (2000). Lexical and nonlexical print-to-sound translation of disyllabic words and nonwords. *Journal of Memory and Language*, 42, 342-364.
- Roelofs, A. (2004). Seriality of phonological encoding in naming objects and reading their names. *Memory & Cognition*, 32, 212-222.
- Sulpizio, S., Arduino, L. S., Paizi, D., & Burani, C. (in press a). Stress assignment in reading Italian polysyllabic pseudowords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Sulpizio, S., Job, R., & Burani, C. (in press b). Priming lexical stress in reading Italian aloud. *Language and Cognitive Processes*.
- Thornton, A. M., Iacobini, C., & Burani, C. (1997). *BDVDB. Una base di dati sul vocabolario di base della lingua italiana*. Roma: Bulzoni.
- Zevin, J. D., Balota, D. A. (2000). Priming and attentional control of lexical and sublexical pathways during naming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26, 121-135.
- Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, 131, 3-29.

# Doppel Teleoperation System: Isolation of physical traits and intelligence for personality study

Hide Nobu SUMIOKA<sup>1</sup> (sumioka@atr.jp)

Shuichi NISHIO<sup>1</sup> (nishio@ieee.org)

Erina OKAMOTO<sup>2</sup> (okamoto.erina@irl.sys.es.osaka-u.ac.jp)

Hiroshi ISHIGURO<sup>1,2</sup> (ishiguro@is.sys.es.osaka-u.ac.jp)

<sup>1</sup> Hiroshi Ishiguro Laboratory, ATR 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan

<sup>2</sup> Graduate School of Eng. Science, Osaka Univ., Machikaneyamacho 1-3, Toyonaka-shi, Osaka, 560-0043 Japan

## Abstract

We introduce the “Doppel teleoperation system”, which isolates several physical traits from a speaker, to investigate how personal information is conveyed to other people during conversation. With the Doppel system, one can choose for each of the communication channels to be transferred whether in its original form or in the one generated by the system. For example, the voice and body motion can be replaced by the Doppel system while the speech content is preserved. This will allow us to analyze individual effects of physical traits of the speaker and content in the speaker’s speech on identification of personality. This selectivity of personal traits provides us with useful approach to investigate which information conveys our personality through conversation. To show a potential of this proposed system, we conduct an experiment to test how much the content of conversation conveys the personality of speakers to interlocutors, without any physical traits of the speakers. Preliminary results show that although interlocutors have difficulty identifying their speakers only by using conversational contents, they can recognize their acquaintances when their acquaintances are the speakers. We point out some potential physical traits to convey our personality.

**Keywords:** social cognition, android science, human-robot interaction, personality psychology, personal presence

## Introduction

Where does personality come from? Do we characterize other people from what they are saying or from how they behave? These issues about personality have been long studied in cognitive psychology. Recent progress has provided us with dimensions of personality to measure human personality (McCrae, Zonderman, Costa, Bond, & Paunonen, 1996) and cognitive models (Brunswik, 1956). Thanks to the establishment of such methodologies, personality studies have been gaining attention not only in cognitive science but also from the viewpoint of design of human-computer/-robot interaction (Fong, Nourbakhsh, & Dautenhahn, 2003; Nass, Moon, Fogg, & Reeves, 1995).

Many studies on personality have been devoted to clarifying what information conveys personality traits of an individual. They have revealed that there exists a strong relationship between physical traits and personality. For example, some studies reported, using criterion measures based on self and peer reports, that a person’s appearance, including facial expression (Berry, 1990, 1991; Little & Perrett, 2006) and clothing style (Naumann, 2009), enables other persons to judge the person’s personality accurately. While these studies were based on photographs of the face or full-body, other studies have shown that body movement (Kenny, Horner,

Kashy, & Chu, 1992) and voice (Scherer & Scherer, 1981; Borkenau & Liebler, 1992) also provide useful information for judging personality traits, especially extraversion.

Although these studies showed several communication channels in which personal traits are presented, the experimental setting was limited to the case where a judge observes a person: there was no conversation between them, although the contents of a conversation would likely be the most informative. A crucial difficulty in examining the relationship between physical traits and personality during a conversation is to isolate physical traits of an individual person from the conversation and to control their effects. Such isolation and control would allow us to investigate not only independent effects of physical traits and personal thought but also mutual interaction among them on identification of personality.

Interactive artificial agents might help us overcome this difficulty since they have been utilized as controllable “humans” to understand the cognitive mechanism of human adults or infants (Itakura, 2008; Yoshikawa, Shinozawa, & Ishiguro, 2007). In this context, some studies have addressed the problems of the behavior and appearance of the agents as contribution to both cognitive science and robotics, using a robot that has a very human-like appearance, called an android (Ishiguro, 2007). While typical androids are controlled as stand-alone agents, a teleoperated android, called a “geminoid”, which has a very similar appearance to a living individual (Sakamoto, Kanda, Ono, Ishiguro, & Hagita, 2007; Nishio, Ishiguro, & Hagita, 2007) has been developed as a telecommunication medium to address several issues on telepresence and self-representation (Straub, Nishio, & Ishiguro, 2010). This system enables an operator to have nonverbal and physical interaction, including body touch, gesture and facial expression, as well as verbal one with other people, by operating an android that might have a different appearance from the operator, remotely.

Although the geminoid system provides us with a way to isolate physical appearance from personality traits, it still transfers not only conversational content but also many other physical traits of its operator such as body movement, facial expression, and speech features. We solve this problem by assuming a speaker who gives content and an operator who acts as a “mediator”, which might distort speech features of the speaker as well as control the geminoid’s movement. The assumption of the mediator enables us to eliminate physical

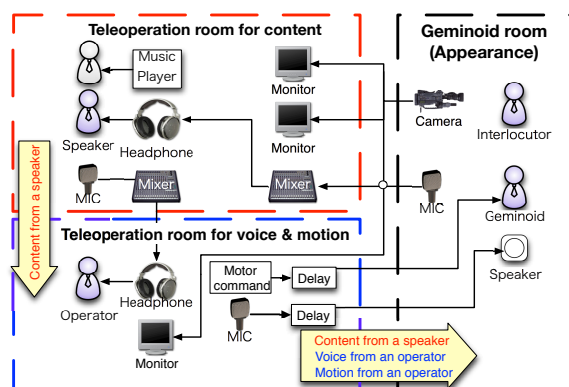


Figure 1: Overview of Doppel teleoperation system. Arrows with yellow show communication channels to be conveyed to the operator or the interlocutor and their sources.

traits of the speaker in speech features such as voice sound and accent from the conversation: interlocutors receive the content from speakers and the others from the mediator and the geminoid. As a result, personal information in conversation is separated into physical traits (appearance, body movement, and speech features) and content of speech (personal thought). Such a system to isolate physical traits will allow us to convey some of personal information of the speakers and to replace the others with ones belonging to a geminoid and its operator selectively.

In this paper, we propose a teleoperation system called “Doppel”, which isolates several physical traits from conversation. This system allows us to analyze individual effects of physical traits of a speaker and content in the speaker’s speech on identification of personality by controlling the physical traits to be conveyed to an interlocutor. To show a potential of the system for investigating how personalities of speakers are conveyed to interlocutors, we report an experiment where identification of the speakers during conversation are tested.

In the rest of the paper, we first describe the proposed system. Next, we report an experiment that we conducted to verify how much content of conversation provides personalities of speakers for their interlocutors. Preliminary results show that although interlocutors have difficulty identifying speakers only using conversational content, they can recognize that they are talking with strangers or their acquaintance. Finally, we discuss what information might provide personalities of speakers for their interlocutors during conversation.

## Doppel Teleoperation System

Figure 1 shows an overview of the proposed system, called the “Doppel Teleoperation System”. The system is based on the telecommunication system for a teleoperated android and uses a “geminoid” that resembles a living individual (Sakamoto et al., 2007; Nishio et al., 2007). The existing system is used for an operator to communicate with remote people. Unlike a video conference system where we only

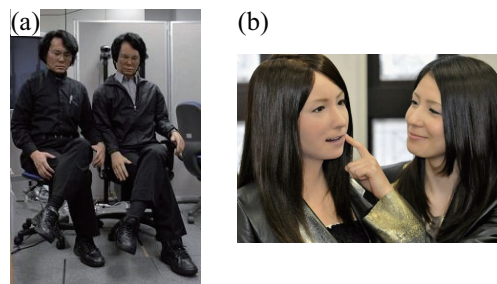


Figure 2: Geminoids. (a) Geminoid HI-1 (right) and the model (left). (b) Geminoid F (left) and the model (right).

provide visual and voice information, it is expected to convey the presence of the operator. We extend this system to isolate individual communication channels by separating the teleoperation system into two subsystems: one for a speaker to have a conversation with an interlocutor and the other for an operator to control voice and motion of the geminoid. In the proposed system, the speaker communicates with the interlocutor through the geminoid, hearing what the interlocutor says and talking into a microphone in another room. The operator hears what the speaker says and then repeats the speaker’s words in her/his way of speaking in another teleoperation room. Therefore, the system allows us to eliminate physical traits of a speaker during conversation: appearance from a geminoid, vocal information and motion one from the operator, and the content of conversation from the speaker. In the following section, we provide more detailed information about the system.

## Appearance:Geminoid

Appearance of a speaker is replaced with the interface between the speaker and an interlocutor. The interface should have human-like appearance to investigate the influence of physical traits on personality identification in human-human interaction. This is achieved by using a geminoid (Figure 2), which resembles an existing individual. Geminoid HI-1 is designed so that its appearance resembles a living male (Figure 2(a)). It has 50 degrees of freedom (DoFs) including 13 DoFs for facial expression. Geminoid F has a similar appearance to a living female (Figure 2(b)). Most of 12 DoFs are used for facial expression.

Both geminoids have two different controllers: a conscious behavior controller and an unconscious one (Sakamoto et al., 2007). While the conscious behavior controller is driven by command from an operator to change behavior of a geminoid based on a set of preprogrammed body motions, subtle expressed motions such as breathing, blinking, and trembling are added by the unconscious behavior controller to maintain the naturalness of the geminoid’s behavior. In addition to such semi-automatic control, lip movements of the geminoid are synchronized with those of its operator. This is realized by a facial feature tracking software through the camera in front of the operator.



### Content of conversation: Speaker

A speaker decides what a geminoid says, monitoring the conversation between the geminoid and an interlocutor. The words of the speaker to the interlocutor is conveyed not to the interlocutor but to the operator, who hears it in an operation room, through a microphone.

### Voice and motion: Operator

The operator controls a geminoid to convey verbal and non-verbal information about a speaker and the operator him/herself to an interlocutor. The operator repeats what the speaker says in the operator's way of speaking in front of a microphone, which is connected to the sound system located behind the geminoid. One might argue for using a system for speech information processing instead of a human operator. Due to the limitations of current technology for speech recognition, we decided to use a human operator.

In addition to conveying the words of the speaker, the operator controls how much s/he provides the interlocutor with physical traits of the speaker such as speed of speaking and accent, and movement. For example, if the operator repeats the speaker's words, mimicking the speaker's way of speaking, speech features of the speaker will help the interlocutor identify the speaker. As a result, the geminoid's voice and movement are presented to an interlocutor as a mixture of verbal information and non-verbal one from a speaker and an operator.

### Experiment: personal identification based on conversational content

The proposed system allows us to isolate communication channels from a speaker and design a new experimental setting that is difficult for existing methodologies. As a first step to verifying how we identify personality traits of other persons during conversation, we investigated whether people can identify a person using only conversational content and how much physical traits affect the identification of the person.

In the following experiments, we used geminoid F. To avoid the speaker being identified due to not conversational content but personal information, an operator was asked to replace the speaker's dialect and specific words to identify the speaker (e.g., the speaker's nickname) with standard dialect and general words (e.g., you), respectively though the content of what a speaker said was preserved.

Since it is difficult for ordinary people to make such replacement, we assigned a female actor as the operator. The lip and head movements of the geminoid F were synchronized with ones of the operator, while other body movements and facial expressions were ignored except for eye blinking, which was realized by the unconscious controller. The communication channels and their sources are summarized in Table 1.

### Working hypothesis and prediction

Although physical appearance, motion, and voice include personality traits, content of conversation should also provide

Table 1: Sources of communication channels during conversation in the experiments

Channel	Source
Appearance	Geminoid F
lip motion	Operator
Voice sound	Operator
Speaking speed	Operator
Accent	Operator
Conversational content	Speaker

much information to identify personality traits because it includes person's thoughts, opinions, and feelings. It will convey more personal information if speakers are acquaintances. Therefore, we verify whether the following hypotheses are established or not.

Hypothesis 1 (H1): people can identify a speaker by using only content of conversation.

Hypothesis 2 (H2): people can correctly identify more speaker by using only content of conversation in case where the speaker is an acquaintance than in case where the speaker is a stranger.

We conducted experiments with two different conditions to verify these hypotheses: *stranger* condition where a speaker and an interlocutor do not know each other and *acquaintance* condition where a speaker and an interlocutor know each other well. We will verify the H1 by evaluating accuracy of a guess of a speaker from among four possible candidates. H2 will be tested through comparison between the accuracy of the guess in the *stranger* condition and one in the *acquaintance* condition.

### Participants

Since the geminoid F has a female physical appearance, only female participants were recruited to eliminate the possibility that gender difference makes it easier for an interlocutor to guess actual speaker. Seventy-six Japanese females participated in the experiment. We made nineteen pairs of two persons who do not know each other for the *stranger* condition while there were nineteen pairs of close friends for the *acquaintance* condition. We assigned one of each pair as a speaker and the other as an interlocutor. The average age of all participants was 25.3 (SD = 6.7).

### Procedure

A subject as an interlocutor was asked to chat about a given topic with a speaker and to guess the speaker from among four possible speakers: the parted subject as a speaker ( $S_s$ ), the model of the geminoid F ( $S_g$ ), the operator of the geminoid F ( $S_o$ ), and the assistant of an experimenter ( $S_a$ ). The last three persons were fixed through all experiments and we confirmed that the interlocutor did not know them. The model and operator of the geminoid F have never been selected as

actual speaker in the experiment. Therefore, the selection of  $S_g$  or  $S_o$  by the interlocutor is assumed to be caused not by conversational content but by other physical traits of the geminoid F and the operator. It is implied that, while her/his guess was based on the appearance of geminoid F if the interlocutor selected  $S_g$ , s/he guessed from the movement and voice of the operator if the interlocutor selected  $S_o$ .

Each experiment consisted of six three-minute sessions. An experimenter selected a topic to be discussed and actual speaker from between  $S_s$  and  $S_a$  before each session. The topic was chosen from two different kinds of topics: common topics and delicate ones. The common topics were the topics which people have more chance to talk about (e.g., “how do you want to enjoy your life after your retirement?”). Some topics were related to personal histories such as Christmas gift that speakers got as a child or personal preference such as favorite type of man. The delicate topics were about what people have less chance to discuss (e.g., “should we revoke elder persons’ driving licences to obviate car accidents?”). The selected speaker was told to discuss the given topic through the geminoid with the interlocutor while the person who was not selected was told to listen to music with headphones so as not to hear the conversation between the actual speaker and the interlocutor. Three consecutive selections of the same speaker were avoided not to make the interlocutor recognize the speaker because of long conversation.

Before each experiment starts, each possible speaker was asked to talk about two different topics provided by an experimenter. The talk was videotaped for two minutes per topic. An interlocutor watched the videos of all talks to discern personalities of all speakers. After that, she was asked to rate their personalities with the Japanese Property-Based Adjective Measurement questionnaire (Hayashi, 1978), which has high correlation between its three components and the extraversion, openness and agreeableness components of the Big Five Model (McCrae et al., 1996).

After rating the personalities of all speakers, the interlocutor was led to the experimental room where the geminoid F was located. The operator and speakers ( $S_s$  and  $S_a$ ) were separated into different rooms, respectively. After a brief explanation about the specifications of geminoid F, the number of sessions and the duration of each session, the experimenter informed the interlocutor that actual speaker could change for each session. It was also noted that the geminoid was controlled by the operator whom the interlocutor saw in the video and she would talk based on her own thought or what one of the other speakers was saying. During a session, the actual speaker and the interlocutor asked each other questions about a given topic and responded to each other. After each session concluded, the interlocutor was asked to guess who the speaker was and to provide the reason for her guess. She also rated personality of the speaker with the questionnaire (Hayashi, 1978). After the experiment finished, the interlocutor was debriefed about the experiment.

Table 2: Average of accuracy rate of guessing in *stranger* condition and *acquaintance* condition

condition	accuracy rate
<i>stranger</i>	0.28
<i>acquaintance</i>	0.31
total average	0.29

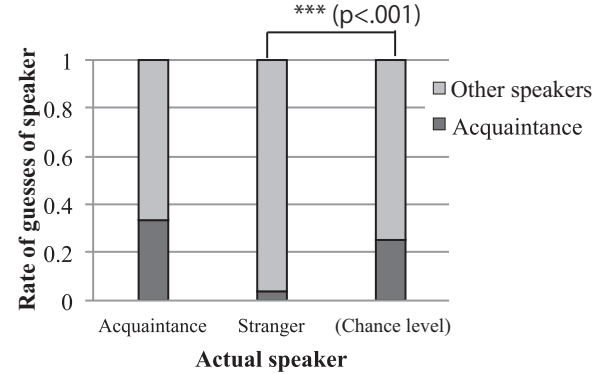


Figure 3: Rate of guesses on each actual speaker in acquaintance condition.

## Evaluation

The performance of an interlocutor was evaluated with how often an interlocutor guessed right on actual speaker. Due to limited space, the analysis of personalities rated by interlocutors with the Japanese Property-Based Adjective Measurement questionnaire is not reported here.

## Result

Table 2 shows average accuracy rates of guessing actual speaker for two conditions and total average across subjects. Although the total average rate is slightly higher than the rate expected by chance (0.25), no significant difference was found between them ( $p = 0.13 > .05$  by binomial test). This result indicates that it is hard to guess who is talking from conventional content, rejecting our first hypothesis. We also tested the second hypothesis by comparing average accuracy rates between two conditions. However, there was no significant difference between them ( $p = 0.85 > .05$  by Wilcoxon test) although the rate in the *acquaintance* condition is slightly higher than one in the *stranger* condition. This result suggests that the difference between two conditions does not support the our second hypothesis.

In the *acquaintance* condition, an interlocutor talks with not only an acquaintance but also a stranger (i.e. the assistant). If she identifies their acquaintances more correctly than the stranger, the second hypothesis is supported within the *acquaintance* condition. Therefore, we compared performance of guessing the actual speaker when the acquaintance is the speaker with one when the stranger is the speaker in the *acquaintance* condition.

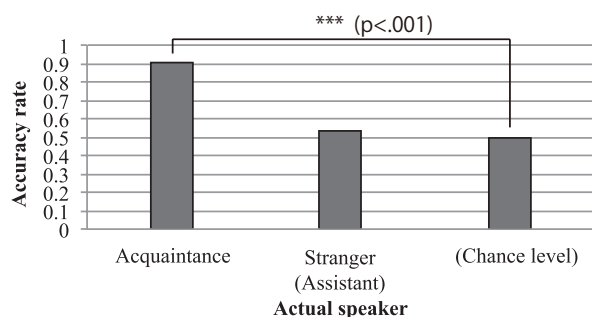


Figure 4: Accuracy rates of guessing on different actual speakers in the acquaintance condition.

Figure 3 shows the rate of guesses of speakers on each actual speaker in the acquaintance condition. The left bar in the figure shows the rate for acquaintances as actual speaker while the middle bar represents the one for the stranger ( $S_a$ ) as actual speaker. The bar on the right shows the rate in case where the guesses occurred by chance. The black part in each bar stands for the rate of guesses of acquaintances. As we can see, the guessing performance for acquaintances as actual speakers is slightly higher than the rate expected by chance. However, there was no significant difference between them ( $p = 0.17 > .05$  by binomial test). Interestingly, the guessing performance is significantly different when the stranger was the actual speaker ( $p = 0.00 < .001$  by binomial test). These results indicate that, while it is hard for the interlocutors to identify acquaintances as the actual speakers, they can recognize that the actual speakers are not their acquaintances.

The results shown in Figure 3 suggest that our second hypothesis is not supported. However, the low accuracy rate of the acquaintances might be caused by the strong conservative bias of the interlocutors when making judgment of actual speaker as their acquaintances. To distinguish accuracy from bias effects, we computed  $A'$  and  $B''$  scores (Grier, 1971) from hit rate (i.e., the rate of the guess of the acquaintance given acquaintances as actual speakers) and false-alarm rate (i.e., the rate of guess of acquaintance given the stranger as actual speaker). The scores showed that the interlocutors are sensitive to their acquaintance ( $A' = 0.80$ ) though they have strong conservative bias against guessing their acquaintances ( $B'' = 0.70$ ). Both scores were higher than ones in case of guessing the stranger ( $S_a$ ) as actual speaker ( $A' = 0.54, B'' = 0.036$ ). In fact, it was revealed that the interlocutors identify their acquaintances significantly when we calculate the accuracy rate given that the interlocutors answered that speakers were their acquaintances. Figure 4 shows the accuracy rates of guessing the actual speakers in *acquaintance* condition. As can be seen, the guessing performance for their acquaintances as actual speaker is much higher than the performance expected by chance. Actually, there was significant difference between them ( $p = 0.00 < 0.001$  by binomial test). We cannot find such difference for the guessing performance for the assistant as actual speaker (see the middle bar in Figure 4). These re-

Table 3: Selection probability of possible speakers

speakers	selection probability
the model of geminoid F	0.285
the operator	0.241
the assistant	0.263
the subject as a speaker	0.206

sults imply that the interlocutors identify their acquaintances as the actual speaker, supporting the second hypothesis.

We also calculated probabilities of guessing each possible speaker as the actual speaker to test how much physical traits can affect the identification of speakers (Table 3). Although the probabilities of selecting the geminoid model or the operator are slightly higher than that of selecting subjects as speakers, we were able to see no significant difference among them. This might imply that no physical trait is much stronger than others when several traits are presented.

## Discussion

The results revealed that it is difficult for people to identify a person without her/his physical traits: physical appearance, body movement, and speech features. In fact, after the experiments, some interlocutors reported that they felt as if the geminoid had another new personality, not one of possible speakers. Even though some results did not support our hypotheses, the results gave us fruitful insights. Especially, it is interesting why it was easy for the interlocutors to recognize that actual speaker is not their acquaintances. Exclusion of some physical traits presented in the experiment will reveal what information provides interlocutors with enough information to make the judgment.

The accuracy and response bias scores suggested that the low accuracy rate of guessing acquaintances might be caused by the conservative bias of the interlocutor for the guess. This finding was supported by the high accuracy rate of guessing when the interlocutors guessed their acquaintances as the actual speaker as shown in Figure 4. Since our second hypothesis was partially supported not between conditions but in the acquaintance condition, further verification is needed.

The accuracy rates shown in Table 3 tell us that there is no significant effect of physical traits on personal identification. This might implies that the identification of personality during conversation results from mutual interaction among physical traits and conversational content. The investigation of such interaction seems difficult for existing approaches because they needs to extract single modality from all modalities and exclude the others. Our system is useful for such investigation because it allows us to examine not only single effect of the physical traits and conversational content but also the mutual interaction among them by controlling the presented information selectively. We will conduct experiments with different combinations of physical traits to investigate how physical traits and conversational content interact

with other traits as a future work.

One concern in this system is the influence of the geminoid on interlocutors' judgment. Previous studies have reported that people respond to an android as they respond to a human if it shows human-like behavior (Shimada & Ishiguro, 2008). Since we designed the geminoid so as to resemble human in appearance and movement, it is expected that the subjects consider the geminoid as a "human". However, it is well-known as uncanny valley (Mori, 1970) that even small lacks of human likeness affect human perception of androids. Therefore, how human likeness of androids affect human judgmental process should be addressed in the future.

We should point out that the interlocutors still extracted some physical traits of their speakers through conversation even though we tried to eliminate this possibility in the conversation. More precisely, they used some speech features to guess actual speakers: timing of speech, duration of speech, and expression of feedback to the interlocutors' comments like "Really?", "Exactly", and "No way!". In addition, it turns out that interlocutors might use speed of speech and accent to guess actual speakers in preparatory experiments. A detailed investigation of physical traits including such speech features is also valuable as future work.

## Conclusion

We introduced the "Doppel teleoperation system", which isolates several physical traits from conversation, to investigate how personal information is conveyed to other people during conversation. With the Doppel system, one can choose for each of the communication modalities to be transferred whether in its original form or the one generated by the system. This will allow us to analyze individual effects of physical traits of the speaker and content in the speaker's speech on identification of personality. We tested how much content of conversation conveys the personality of speakers for interlocutors, without any physical traits of the speakers. Preliminary results showed that although interlocutors have difficulty identifying their speakers only by using conversational content, they could recognize that they were talking with their acquaintances. We hope that this system helps us understand our cognitive mechanism of our personality.

## Acknowledgments

This research was supported by Grant-in Aid for Scientific Research (S), KAKEN (20220002) and JST, CREST. HS thanks M. Shimada for valuable comments.

## References

- Berry, D. (1990). Taking people at face value: Evidence for the kernel of truth hypothesis. *Social Cog.*, 8, 343–361.
- Berry, D. (1991). Accuracy in social perception: Contributions of facial and vocal information. *Journal of Personality and Social Psycho.*, 61, 298–307.
- Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero-acquaintance. *Journal of Personality and Social Psychology*, 62, 645–657.
- Brunswik, E. (1956). Perception and the representative design of psychological experiments (2d ed.).
- Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3), 143–166.
- Grier, J. (1971). Nonparametric indexes for sensitivity and bias: computing formulas. *Psycho. Bulletin*, 75(6), 424.
- Hayashi, F. (1978). The fundamental dimensions of interpersonal cognitive structure. *Bulletin of the Faculty of Education of Nagoya University*, 25, 233–247. (in Japanese)
- Ishiguro, H. (2007). Android science—toward a new cross-interdisciplinary framework. *Robo. Res.*, 28, 118–127.
- Itakura, S. (2008). Development of mentalizing and communication: From viewpoint of developmental cybernetics and developmental cognitive neuroscience. *IEICE TRANSACTIONS COMMUNICATIONS E SERIES B*, 91(7), 2109.
- Kenny, D. A., Horner, C., Kashy, D. A., & Chu, L. (1992). Consensus at zero-acquaintance: Replication, behavioral cues, and stability. *Journal of Personality and Social Psychology*, 62, 88–97.
- Little, A. C., & Perrett, D. I. (2006). Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology*, 98, 111–126.
- McCrae, R., Zonderman, A., Costa, P., Bond, M., & Paunonen, S. (1996). Evaluating replicability of factors in the revised neo personality inventory: Confirmatory factor analysis versus procrustes rotation. *Journal of Personality and Social Psychology*, 70(3), 552.
- Mori, M. (1970). Bukimi no tani (the uncanny valley). *Energy*, 7(4), 33–35.
- Nass, C., Moon, Y., Fogg, B., & Reeves, B. (1995). Can computer personalities be human personalities?. *International Journal of Human-Computer Studies*.
- Naumann, L. (2009). Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin*, 35(12), 1661–1671.
- Nishio, S., Ishiguro, H., & Hagita, N. (2007). Geminoid: Teleoperated android of an existing person. *Humanoid robots-new developments. I-Tech*.
- Sakamoto, D., Kanda, T., Ono, T., Ishiguro, H., & Hagita, N. (2007). Android as a telecommunication medium with a human-like presence. In *Proc. of the acm/ieee int. conf. on human-robot interaction* (pp. 193–200).
- Scherer, K., & Scherer, U. (1981). Speech behavior and personality. *Speech evaluation in psychiatry*, 115–135.
- Shimada, M., & Ishiguro, H. (2008). Motion behavior and its influence on human-likeness in an android robot. In *Proc of the annual conf. of the cog. sci. society* (pp. 2468–2473).
- Straub, I., Nishio, S., & Ishiguro, H. (2010). Incorporated identity in interaction with a teleoperated android robot: A case study. In *Proc. of int. sympo. on robot and human interactive commu.* (pp. 119–124).
- Yoshikawa, Y., Shinozawa, K., & Ishiguro, H. (2007). Social reflex hypothesis on blinking interaction. In *Proc. of the annual conf. of the cog. sci. society* (pp. 725–730).

# Individuals' Process of Metaphor Interpretations and Interestingness Cognition

**Tomohiro Taira** (cogpsy.t.taira@gmail.com)

Center for Research and Development of Higher Education, Osaka City University  
3-3-138 Sugimoto Sumiyoshi-ku, Osaka-shi 558-8585, Japan

**Takashi Kusumi** (kusumi@educ.kyoto-u.ac.jp)

Graduate School of Education, Kyoto University  
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

**Akira Utsumi** (utsumi@inf.uec.ac.jp)

Department of Informatics and Engineering, The University of Electro-Communications  
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

## Abstract

In this paper, we investigated the process of interestingness cognition in metaphor comprehension. We did this from the point of view that the interestingness of a metaphor (e.g., “*life is like a gamble*”) is related to its interpretative diversity. Two studies were conducted to assess this phenomenon: Study 1 (interpretation-production) and Study 2 (interpretation-presentation study). In Study 1, we observed that a greater number of interpretations were produced from a metaphor that was interesting and easy to understand as compared to one that was less interesting and difficult to understand. In Study 2, we observed that a metaphor was more interesting when more information on simile interpretation was presented. On the basis of these results, we discuss the relationship between the process of metaphor comprehension and metaphor evaluation.

**Keywords:** metaphor/simile comprehension; interpretive diversity; interestingness.

## Introduction

Sentences such as “*life is like a gamble*” and “*marriage is like a refrigerator*” include comparative senses. Such sentences, consisting of a paired topic and vehicle, which we refer to as a “metaphor” (strictly a “simile”), indicates similar points between two words: *life* is like *a gamble* (both unpredictable and implying risk) and *marriage* cools a relationship or keeps it fresh, as does a *refrigerator* for its contents. Similarity is very important for metaphor comprehension. Recent studies have discussed similarity cognition or the factors that affect similarity cognition in metaphor comprehension. In fact, most studies discuss the relationships between similarity cognition and the process of metaphor comprehension (e.g., Gentner, 1983; Ortony, 1979; Tversky, 1977).

## The Process of Metaphor Comprehension

Similarity cognition in metaphor comprehension is described simply as “the similarity between the topic and vehicle.” The question of similarity involves two forms of nuance: the qualitative sense and the quantitative sense. As illustrated in aptness views (e.g., Chiappe & Kennedy, 1999; Chiappe, Kennedy, & Smykowski, 2003; Jones &

Estes, 2006), the former refers to the degree (“goodness” and “adequacy”) to which the topic and vehicle are similar. As defined in Chiappe and Kennedy (2001), goodness and adequacy indicate the extent to which a comparison captures the important features of the topic. For example, *gamble* includes features salient in, and applicable to, the nature of *life*: in gambling’s sense of “unpredictability,” *gamble* adequately represents an important aspect of *life*. Likewise, both *a refrigerator* and *marriage* cool something, but *marriage* is comparatively difficult to express with a *refrigerator*. Previous studies have shown that this type of similarity affects the process of metaphor comprehension. Jones and Estes (2006) experimentally revealed that the strength of metaphor aptness predicts metaphor/simile preference, reading time for a metaphor, and the ease of interpretation of a metaphor. An apt relationship between the topic and vehicle creates a preferential metaphorical (categorical) expression, is read faster, and is rated as easier to understand than a less apt relationship.

According to the quantitative view, similarity cognition is based on the number of features shared by both the topic and the vehicle. If this number is large, similarity cognition between the topic and vehicle is strong. In the process of metaphor comprehension, these shared features are generated as metaphor interpretation: the metaphor with the most shared features is predicted to produce the most interpretations. In previous studies, simulation results have shown reliable evidence that the productivity of metaphor interpretation, such as interpretative diversity (Utsumi & Kuwabara, 2006; Utsumi, 2007), is more closely related to the process of metaphor comprehension than to the goodness of similarity (i.e., metaphor aptness). Thus, the topic-vehicle relationship that produces several interpretations is the preferred metaphor or simile.

## The Process of Metaphor Evaluation

As described above, the similarity cognition of a metaphor plays an important role in the process of metaphor comprehension. On the other hand, some previous studies have suggested that similarity cognition is also related to the process of metaphor evaluation, such as the rhetoric effect and how funny and interesting a metaphor is.

A metaphor is understood through its cognitive effect, which is not only the enhancement of word meaning (Blasko & Connine, 1993; Gernsbacher, Keyser, Robertson, & Werner, 2001; Taira & Kusumi, 2011) but also its pragmatic effect (Sperber & Wilson, 1994). The former has been revealed to be affected by the strength of the similarity cognition, such as aptness (Blasko & Connine, 1993; Taira & Kusumi, 2011); the latter appears to be related to similarity cognition. For example, Roberts and Kreuz (1994) show that all figurative expressions have some discourse goal. Among them, a metaphor (e.g., “*life is a gamble*”) and a simile (e.g., “*life is like a gamble*”) have different pragmatic goals. One difference is that the simile is used as a humorous expression, while the metaphor is not. Previous studies have shown that the simile is a comparative expression based on similarity cognition, unlike the metaphor (Bowdle & Gentner, 2005; Jones & Estes, 2006) so that similarity cognition is related to humor.

In consideration of the above, we examined the relationship between the simile comprehension process and its evaluation process in a previous study (Taira, Nakamoto, & Kusumi, 2006). The aim of that study was to examine the process of interestingness cognition through correlations between factors affecting the process of simile comprehension. We studied 75 undergraduate native Japanese speakers and employed 30 Japanese similes (e.g., “*life is like a gamble*,” “*marriage is like a refrigerator*”). Through a simple rating task, the ease of comprehension, similarity, familiarity, unpredictability, and interestingness of each simile were measured. In addition to these ratings, the number of interpretations for each simile was collected in another study where participants were required to write out their interpretations of the simile.

Table 1: Correlations between the factors of metaphor comprehension in Taira, Nakamoto, and Kusumi (2006)

Factors	2	3	4	5	6
1. Ease of Comprehension	.960	.947	.938	.740	.347
2. Similarity	-	.933	.927	.705	.382
3. Familiarity	-	-	.967	.696	.302
4. Unpredictability	-	-	-	.712	.300
5. Interestingness	-	-	-	-	.533
6. Number of Interpretations	-	-	-	-	-

$N = 30$

Correlations between the metaphor factors are shown in Table 1, and the results of a path analysis based on the correlation data are shown in Figure 1. The results indicate that both similarity and familiarity, considered factors related to similarity cognition (Chiappe & Kennedy, 2001), are related directly to ease of comprehension. Furthermore, the ease of comprehension and the number of interpretations directly affect simile interestingness: the more easily the simile is understood and the more interpretations the simile produces, the simile is interpreted as more interesting.

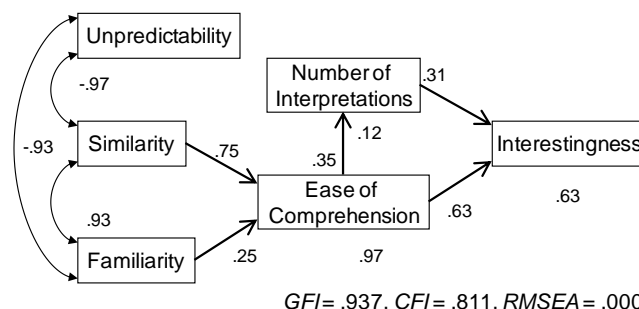


Figure 1: The process of interestingness cognition (Taira, Nakamoto, & Kusumi, 2006)

## The aim of our research

Our previous studies suggest that the similarity cognition of a metaphor, especially quantitative similarity, is related not only to the process of metaphor comprehension but also to the process of metaphor evaluation (i.e., interestingness cognition). However, such results are only suggested by correlational data; it is unknown whether a metaphor that is interesting and easy to understand really produces several interpretations and whether metaphor interpretation directly affects interestingness cognition. It is unclear whether the interestingness of a metaphor is based on the metaphor’s properties or an individual’s action. In this paper, we investigated the process of interestingness cognition in metaphor comprehension.

In Study 1, we examined the number of interpretations for a simile. Our previous study did not reveal the process of interpretation production in simile comprehension; thus, we did not determine whether an individual could produce several interpretations from a simile that is interesting and easy to understand. In Study 1, we examined the number of interpretations for various metaphors with different levels of interestingness and ease of comprehension.

In Study 2, we examined whether the interpretation itself increases the interestingness cognition. As in Study 1, interestingness cognition is inferred through correlational relationships. If this prediction is correct, a metaphor will be judged more interesting when more interpretations of the metaphor are presented. For Study 2, we provide experimental data on the relationship between the interpretation and the interestingness of metaphors.

## Study 1

The aim of Study 1 was to confirm that an individual produces more interpretations for a more comprehensive and interesting metaphor, and vice versa.

## Method

**Participants** 800 participants were recruited from an internet research company. All were native Japanese speakers.

**Materials** Thirty-six Japanese similes were selected from the materials used in Taira and Kusumi (2009); some were also selected from materials used in our previous study

(Taira, Nakamoto & Kusumi, 2006). For these similes, Taira and Kusumi (2009) examined interestingness and ease of comprehension using 5-point scales (1 = “not at all interesting or easy to understand” to 5 = “very interesting or easy to understand”). They were clustered within three simile types: 12 similes that were seen as highly interesting and very easy to understand (e.g., “*life is like a gamble*.” interestingness,  $M = 3.21$ , ease of comprehension,  $M = 4.04$ ), 12 similes seen as moderately interesting and easy to understand (e.g., “*a husband is like jewelry*.” interestingness,  $M = 2.81$ , ease of comprehension,  $M = 3.38$ ), and 12 similes seen as less interesting and difficult to understand (e.g., “*marriage is like a refrigerator*.” interestingness,  $M = 2.38$ , ease of comprehension,  $M = 2.36$ ). The correlation between interestingness and ease of comprehension was very strong ( $r(36) = .88$ ). This result is similar to results obtained in Taira, Nakamoto, and Kusumi (2006); thus, the material selection in Study 1 was appropriate. In this paper, we defined each type of simile within a high-, middle-, and low-rating group.

**Procedures** This study was part of an omnibus internet survey that measured higher-order literacy. The monitors participated in the survey on the internet. They were required to access the website described by the internet research institute and to answer questions relevant to our study. Three similes had been selected from each category. Participants were required to provide as many interpretations of each simile as possible. The interpretations were typed into a textbox on the webpage.

## Results and Discussion

Between 57 and 86 participants produced interpretations for each simile. Data were coded and clustered. Through this procedure, the number of interpretation units for each simile was examined. We defined an interpretation unit as the component included within the participant’s text with an independently important sense for the metaphor’s interpretation. For example, if one participant produces the interpretation “it is unpredictable and followed with any risk. It does not describe what will happen next” for “*life is like a gamble*,” two interpretation units are produced because the second sentence includes the same unit that appears in the first sentence.

Table 2: Mean number of interpretation units

	low-rating	middle-rating	high-rating
interpretation	1.25	1.37	1.51
unit (SD)	(.73)	(.73)	(.92)

$N=800$

There were strong correlations between the interpretation unit and ease of comprehension ( $r(800) = .498$ ) and interestingness ( $r(800) = .404$ ) in Taira and Kusumi (2009). These results suggest that participants produced more interpretations for similes that were more interesting and easy to understand. In addition, the mean of the interpretation unit per participant is shown in Table 2. The

mean data were analyzed through one-way ANOVAs with participants ( $F_p$ ) and items ( $F_i$ ).

The main effect of rating group was significant ( $F_p(2, 1598) = 36.86$ ,  $\eta^2 = .02$ ;  $F_i(2, 22) = 5.14$ ,  $\eta^2 = .24$ ;  $ps < .001$ ). Multiple comparisons revealed significant differences between the low and middle rating groups ( $t(1598) = 4.00$ ,  $r = .15$ ), low and high rating groups ( $t(1598) = 8.58$ ,  $r = .28$ ), and middle and high rating groups ( $t(1598) = 4.58$ ,  $r = .16$ ).

The results show that participants produced different numbers of interpretations according to their ease of comprehension and interestingness. This is somewhat consistent with results from Taira, Nakamoto, and Kusumi (2006). However, both ease of comprehension and interestingness in Study 1 were defined through data from our previous studies (Taira & Kusumi, 2009). Results from Study 1 did not indicate whether participants really conceived the metaphor as interesting and easy to understand. This problem was addressed in Study 2.

## Study 2

Study 1 revealed relationships between the ease of comprehension/interestingness of a metaphor and its number of interpretations. From these results, however, we cannot ascertain whether interestingness cognition is followed by metaphor interpretation or whether interestingness cognition follows metaphor interpretation. In Study 2, we controlled the number of metaphor interpretations and investigated the effect of interpretation on interestingness cognition.

## Method

**Participants** Fifty-four participants took part in Study 2. All were native Japanese speakers and had not participated in Study 1.

**Materials** From Study 1, the 12 similes that were defined within the high-rating group (e.g., “*life is like a gamble*”) and the 12 similes that were defined within the low-rating group (e.g., “*marriage is like a refrigerator*”) were selected. For each simile, three relevant simile features (e.g., for “*life is like a gamble*,” “*unpredictable*,” “*followed with any risk*,” and “*needing strategy*”) were applied. The three relevant features were selected from the first, second, and third most popular interpretation units produced in Study 1.

**Procedure** Study 2 was composed of three tasks: a rating task, a reading span task (RST), and a re-rating task. These tasks were performed in aforementioned order.

The rating task was a simple rating task in which participants were required to rate the ease of comprehension, interestingness, and unpredictability of the similes. Each factor was rated on 7-point scales (1 = “very difficult to understand,” “not at all interesting,” and “very predictable” to 7 = “very easy to understand,” “very interesting,” and “very unpredictable”).

For the RST, a standardized procedure of the Japanese RST (Osaka & Osaka, 1994) was performed. For this task, 2 to 5 sentences with one word underlined were presented in order; participants were required to read aloud each



sentence. After all the sentences were presented and read, participants were required to read all the underlined words without the sentence. The task included 22 trials: the first two were practice trials and the remaining 20, true trials. The RST was used only as a filler task between the rating and re-rating tasks.

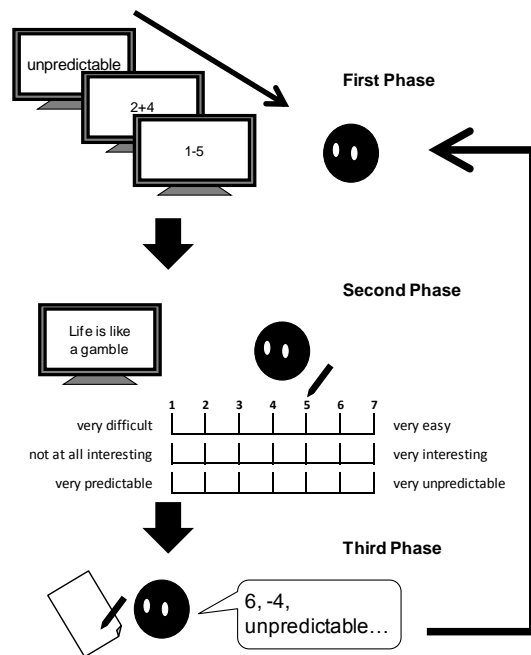


Figure 2: Design of the re-rating task in Study 2

The re-rating task was composed of three phases (see Figure 2). During the first phase, three information units were presented to participants. The information unit was either the feature (e.g., “unpredictable,” “followed with any risk,” or “needing strategy”) relevant to the simile (e.g., “life is like a gamble”) or a simple addition-subtraction calculation formula (e.g., “2 + 4” or “1 - 5”). The composition of the information units included three features without a calculation formula, one feature and two calculation formulas, and no features and three calculation formulas. Participants were required to comprehend the information units because they would perform a recall task after this phase. During the second phase, participants were required to rate the ease of comprehension, interestingness, and unpredictability of the similes in the same manner as during the rating task. Participants were instructed to re-rate the similes based on their current impression (not based on their previous rating). During the third phase, participants were required to recall features and calculation formulas learned during the first phase. After participants finished the third phase, the next trial began. This task included 26 trials: the first two were practice trials.

## Results and Discussion

The scores for ease of comprehension, interestingness, and unpredictability in the rating and re-rating tasks were examined. Mean scores for ease of comprehension,

interestingness, and unpredictability for the high rating group are shown in Figure 3, and the scores for the low rating group are shown in Figure 4.

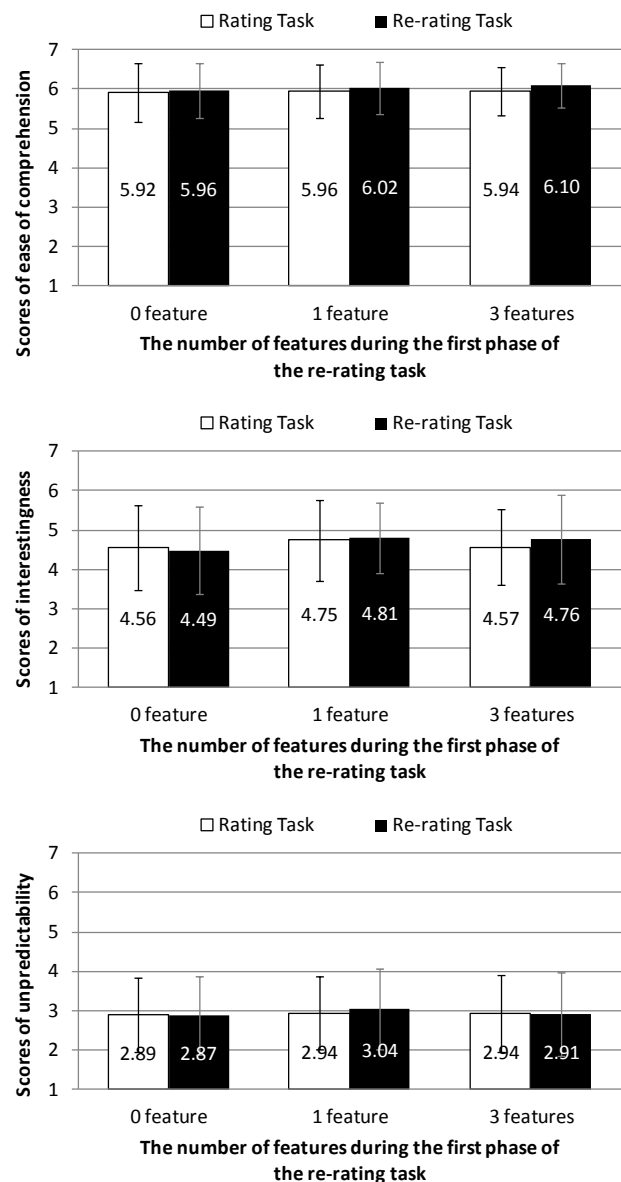


Figure 3: The mean scores (SD) of the high-rating group

**High-rating group’s results** For the high-rating group, the ease of comprehension score was very high, and a series of 2 (task type: rating/re-rating) x 3 (the number of feature: 0/1/3) repeated-measures ANOVAs revealed a significant main effect of task type ( $F(1, 53) = 4.53, p < .005, \eta^2 = .00$ ) but no main effect of the number of features ( $F(2, 106) = .47, \eta^2 = .00$ ) or any interactions ( $F(2, 106) = .87, \eta^2 = .00$ ). Likewise, ANOVAs were conducted on the interestingness and unpredictability scores. The unpredictability result revealed no significant main effects of task type ( $F(1, 53) = .06, \eta^2 = .00$ ), the number of features ( $F(2, 106) = .51, \eta^2 = .00$ ), or any interactions ( $F(2, 106) = .51, \eta^2 = .00$ ). The

result of interestingness also showed no significant main effects of task type ( $F(1, 53) = .50, \eta^2 = .00$ ) and the number of features ( $F(2, 106) = 2.96, p < .10, \eta^2 = .01$ ), and no significant interaction ( $F(2, 106) = 1.48, \eta^2 = .00$ ).

If the prediction that metaphor interpretation directly affects and increases interestingness cognition is correct, the results from the high rating group suggest that the simile of the high rating group originally produced several interpretations (from Study 1); thus, the scores for each rating task factor were the same as the scores in the re-rating task where interpretations were presented.

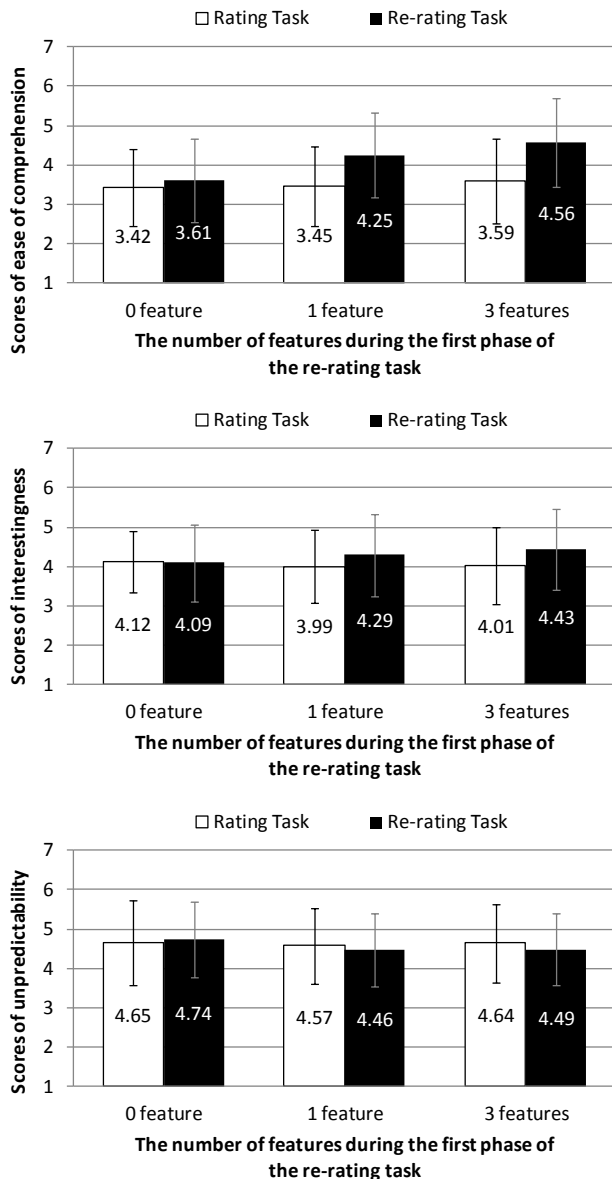


Figure 4: The mean scores (SD) of the low-rating group

**Low-rating group's results** For the low-rating group, a series of 2 (task type: rating/re-rating) x 3 (the number of features: 0/1/3) repeated-measures ANOVAs revealed significant main effects of task type ( $F(1, 53) = 56.89, p$

$< .001, \eta^2 = .08$ ) and the number of features ( $F(2, 106) = 9.46, p < .001, \eta^2 = .04$ ) in the ease of comprehension scores, as well as a significant interaction ( $F(2, 106) = 15.05, p < .001, \eta^2 = .02$ ). To deconstruct the interaction, Ryan's multiple comparisons test indicated simple main effects of task type on both the 1-feature and 3-feature conditions ( $F_s(1, 159) = 42.79, 63.04, p_s < .001, \eta^2 = .06, .08$ ). Simple main effects of the number of features on the re-rating task were also significant ( $F(2, 212) = 20.96, p < .001, \eta^2 = .07$ ): in the re-rating task, the scores in both the 1-feature and 3-feature conditions were higher than the 0-feature condition ( $t_s(106) = 2.59, 4.32, p_s < .05, .001, r = .28, .40$ ). On the other hand, scores for the 3-feature condition were not significantly higher than the 1-feature condition ( $t(106) = 2.08, r = .14$ ).

For the interestingness scores, there were also significant main effects of task type ( $F(1, 53) = 8.74, p < .005, \eta^2 = .01$ ) and interactions ( $F(2, 106) = 3.39, p < .05, \eta^2 = .01$ ) but no significant effects for the number of features ( $F(2, 106) = .50, \eta^2 = .00$ ). Ryan's multiple comparisons test also indicated simple main effects of task type in both the 1-feature and 3-feature conditions ( $F_s(1, 159) = 5.54, 10.62, p_s < .05, .005, \eta^2 = .01, .02$ ), but no simple main effect of the number of features for the re-rating task ( $F(2, 212) = 2.58, p < .10, \eta^2 = .01$ ). Conversely, for the unpredictability scores, there were no significant main effects (task type:  $F_s(1, 53) = .51, \eta^2 = .00$ ; number of features:  $F_s(2, 106) = 1.06, \eta^2 = .01$ ), or interactions ( $F_s(2, 106) = 1.80, \eta^2 = .00$ ).

These results suggest that the presentation of metaphor interpretation, which is related to similarity cognition, affects the process of metaphor comprehension: the interestingness of a metaphor might be increased through interpretations. This is consistent with the prediction that interpretative action significantly affects interestingness cognition. Our results also confirm previous studies suggesting that metaphor appreciation is based on the resolution of incongruity (Utsumi, 2002; Utsumi, 2005). However, the solution of unpredictability was not detected by results from Study 2. One possible interpretation is that unpredictability might be attributed not to the simile but to the interpretation itself. The low-rating similes are generally difficult to comprehend and produce its interpretations (from Study 1) so that the presented interpretations in Study 2 can be also unexpected to the participants. If some participants confounded this cognitive process with the task judgment that required the evaluation of the simile itself, results from the low-rating group are probable. This problem needs to be addressed in future research by using more strict instructions and experimental paradigms.

## General Discussion

The current studies have provided experimental evidence of metaphor comprehension/evaluation. Previous studies have only revealed relationships between these constructs and were unable to fully determine whether evaluation results are based on the metaphors' properties or individuals' inner processes.

Our results suggest that metaphor evaluation is based on interpretative action. Moreover, our results indicate that metaphor comprehension is strongly affected by whether the connection between two different concepts is discovered. Thus, our results support the quantitative view of metaphor comprehension (Utsumi, 2007). However, our results do not fully discount the qualitative view given that the number of interpretations observed depends on the context, the saliency of interpretation, and an individual's cognitive ability. Our task paradigm, especially that of Study 2, shows incongruence between the interpretation during the task and the interpretation that the individual produces. We usually produce metaphor interpretations when reading or listening to them and unaided by any relevant information. We typically are unable to refer to adequate interpretations, as were participants in Study 2. In future research, we will examine the relationship between metaphor interpretation and metaphor evaluation through a task requiring participants to produce interpretations of metaphors.

Previous studies have discussed the relationship between the process of comprehension and an individual's cognitive ability, such as working memory (e.g., Chiappe & Chiappe, 2007; Pierce & Chiappe, 2009; Pierce, McLaren, & Chiappe, 2010). However, there are few studies examining the relationship between evaluation processes, such as interestingness, and working memory. Future research will need to examine the working memory factor, which is predicted to affect the process of both comprehension and evaluation.

### Acknowledgments

Part of this research was supported by a Grant-in-Aid for Scientific Research (B) (No. 23300098) from the Japan Society for the Promotion of Science.

### References

- Blasko, D., & Connine, C.M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 295-308.
- Bowdle, B., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, 112, 193-216.
- Chiappe, D., & Chiappe, P. (2007). The role of working memory in metaphor production and comprehension. *Journal of Memory and Language*, 56, 172-188.
- Chiappe, D., & Kennedy, J. (1999). Aptness predicts preference for metaphors or similes, as well as recall bias. *Psychological Bulletin & Review*, 6, 668-676.
- Chiappe, D., & Kennedy, J. (2001). Literal bases for metaphor and similes. *Metaphor and Symbol*, 16, 249-276.
- Chiappe, D., Kennedy, J., & Smykowski, T. (2003). Reversibility, aptness, and the conventionality of metaphors and similes. *Metaphor and Symbol*, 18, 85-105.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.
- Gernsbacher, M. A., Keyser, B., Robertson, R. R. W., & Werner, N. K. (2001). The role of suppression and enhancement in understanding metaphors. *Journal of Memory and Language*, 45, 433-450.
- Jones, L., & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, 55, 18-32.
- Ortony, A. (1979). Beyond literal similarity. *Psychological Review*, 86, 161-180.
- Osaka, M., & Osaka, N. (1994). Working memory capacity related to reading: Measurement with the Japanese version of reading span test. *The Japanese Journal of Psychology*, 65, 339-345.
- Pierce, R., & Chiappe, D. (2009). The roles of aptness, conventionality, and working memory in the production of metaphors and similes. *Metaphor and Symbol*, 24, 1-19.
- Pierce, R., & Chiappe, D. (2009). The role of working memory in the metaphor interference effect. *Psychonomic Bulletin & Review*, 17, 400-404.
- Roberts, M., & Kreuz, J. (1994). Why do people use figurative language? *Psychological Science*, 5, 159-163.
- Taira, T., & Kusumi, T. (2009). The cognition of the topic and vehicle and aptness of metaphor. *Proceedings of the 9th Annual Meeting of the Japanese Cognitive Linguistics Association* (pp. 465-471).
- Taira, T., & Kusumi, T. (2011). The topic comprehension process in simile sentences. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 2156-2161).
- Taira, T., Nakamoto, K., & Kusumi, T. (2006). The effects of ease to understand and interpretative diversity on interestingness of metaphor. *Proceedings of the 70th Annual Convention of the Japanese Psychological Association* (240).
- Taira, T., Nakamoto, K., & Kusumi, T. (2007). Metaphor familiarity and interpretation diversity. *Cognitive Studies*, 14, 322-338.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327-352.
- Utsumi, A. (2002). Toward a cognitive model of poetic effects in figurative language. *Proceedings of 2002 IEEE International Conference on Systems, Man and Cybernetics* (SMC2002).
- Utsumi, A. (2005). The role of feature emergence in metaphor appreciation. *Metaphor and Symbol*, 20, 151-172.
- Utsumi, A. (2007). Interpretive diversity explains metaphor-simile distinction. *Metaphor and Symbol*, 22, 291-312.
- Utsumi, A., & Kuwabara, Y. (2005). Interpretive diversity as a source of metaphor-simile distinction. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 2230-2235).

# Transmission of Rumor and Criticism in Twitter after the Great Japan Earthquake

**Yuko Tanaka (Yuko.Tanaka@stevens.edu)**

Howe School of Technology Management, Stevens Institute of Technology  
Castle Point on Hudson, Hoboken, NJ 07030 USA

**Yasuaki Sakamoto (Yasuaki.Sakamoto@stevens.edu)**

Howe School of Technology Management, Stevens Institute of Technology  
Castle Point on Hudson, Hoboken, NJ 07030 USA

**Toshihiko Matsuka (matsukat@muscat.L.chiba-u.ac.jp)**

Department of Cognitive and Information Science, Chiba University,  
1-33, Yayoicho, Inage Ward, Chiba-shi, Chiba, 263-8522 JAPAN

## Abstract

The purpose of this study was to examine psychological factors that affect the transmission of rumor and criticism in social media during disasters. 40 students at Chiba University evaluated 10 rumor tweets and corresponding 10 criticism tweets that were posted in Twitter after the Japan March 11 Earthquake. Among some psychological factors, only importance was related to intended transmission of rumor. Surprisingly, accuracy and anxiety were not predictors of any transmission. Estimated transmission of criticisms was higher when its importance was high, while that of rumor did not vary according to importance. Interestingly, although participants estimated that criticisms were spread more than rumor, they intended to transmit rumors as much as criticisms.

**Keywords:** Rumor; criticism; disaster; social media; technology; communication

## Introduction

A 9.0 magnitude earthquake hit northeastern Japan on March 11, 2011. The Great East Japan Earthquake triggered powerful tsunami waves and a series of aftershocks, which caused a devastating damage to the country.

During the disasters, social media played an important role in obtaining and transmitting information to understand the situation. An example is Twitter, which enables its users to send and read text messages of up to 140 characters, known as “tweets,” and to forward a message by re-tweeting a tweet to followers through a single click.

Communications during disasters increasingly relies on social media like Twitter. One reason is that social media allow immediate and interactive information transmission. This advantage, for example, led to the discovery and rescue of some individuals who were isolated in a disaster area after the Japan Earthquake. Although social media can play an important role in sharing information and coordinating disaster response, social media can also facilitate the dissemination of false information, potentially creating widespread panic. After the Japan Earthquake, for example,

Twitter was immediately filled with tweets about the disaster that included not only useful information but also false rumors (Ogiue, 2011; Tachiiri, 2011). The spread of false rumors about the disaster became a major social problem, and the Japanese government called attention to false rumors on the Internet.

Given the growing use of social media in people’s everyday life, cognitive science research needs to examine how people process information using social media technologies. This work contributes to this need by analyzing the transmission of rumor and criticism in Twitter after the Japan Earthquake. During disasters, factors such as time pressure and psychological stress come into play, and each individual’s decision and action can have significant impact. For instance, the immediate and far-reaching spread of false information can be detrimental. Thus, it is important to study how users interact with information in social media.

Although Twitter is a new technology started in 2006, the spread of false rumors during disasters is not a new phenomenon (e.g., Prasad, 1935; Sinha, 1952). For example, Prasad (1935) categorized rumors after the great Indian earthquake of 1934. He found that the same types of rumors about earthquakes appear again and again in different locations during the past 1,000 years (Prasad, 1950).

Rumor study caught attention after World War II (see Rosnow & Foster, 2005). Rumor was defined as “*unverified and instrumentally relevant information statements in circulation that arise in contexts of ambiguity, danger, or potential threat and that function to help people make sense and manage risk*” (DiFonzo & Bordia, 2007, p.13). It was distinguished from gossip, defined as an evaluative statement about someone’s private lives.

Past rumor studies revealed psychological factors that affect rumor behavior, such as accuracy, anxiety, and importance of rumors, and examined rumors in different situations, including universities, organizations, and communities (Anthony, 1973; DiFonzo & Bordia, 2000; Rosnow, 1991; Rosnow, et al., 1988; Walker & Beckele,

1987). However, few of the past work have examined rumors in social media, which allow users to communicate with a large number of people who are physically distant. Thus, it is unclear whether we can apply the findings from past studies to rumor transmission in Twitter.

In addition to the abundance of rumors in Twitter, we noticed that many people tried to stop the spread of false rumors by criticizing the rumor tweets. It was not only the government and organization but also many individual Twitter users who posted criticism tweets. A number of studies have shown that refutation decreases the level of belief in rumors (e.g., Allport & Postman, 1947; Bordia, et al., 2000; Iyer & Debevec, 1991). Thus, criticism tweets could minimize the impact of false rumors by making users critical. Even if false rumors spread widely in Twitter, the negative effect of false rumor would be curbed if criticisms also spread. For this reason, examining how criticisms spread in Twitter deserves attention.

In the current study, we examine the psychological factors that affect rumor and criticism transmission using actual tweets posted after the Japan Earthquake. Specifically, we study whether perceived accuracy and importance of tweets and anxiety arising from the tweets relate to the intended and estimated transmission of the tweets.

## Method

### Participants

Forty students (18 male, mean age 20 years) from Chiba University in Japan participated for course credit. In addition, they received a gift card in the amount of 500 Japanese yen (about \$6.5). The experiment was conducted from October 19 to November 1, 2011. Chiba University is located in one of the areas affected by the disasters.

### Stimuli

We collected 10 rumor tweets related to the disasters following the Japan Earthquake and 10 criticism tweets that criticized the corresponding rumor tweets (see Appendix). Each tweet was posted in Japanese on Twitter between March 11 and September 7, 2011. Each of the 20 tweets was converted to a 700×162 pixels image in the PNG format (see Figure 1). The user name associated with each tweet was generated by randomly combining alphabet and number. The image also contained the actual date when the original tweet was posted. We created each criticism tweet by adding the word “RT” (an abbreviation for Re-Tweet), the user name of the corresponding rumor tweet, and part of the rumor tweet to the criticism (see Figure 1, bottom). The maximum number of characters in each tweet image was 140 in Japanese.

### Design and Procedure

The within-subject factors were tweet type (rumor vs. criticism) and transmission (intended vs. estimated). Participants accessed the experiment through the Internet using computer. They were instructed to answer all

questions within 50 minutes. The experiment consisted of the four phases in the following order:

**1. Rumor tweet** Participants answered the following eight questions about each rumor tweet: (1) Familiarity – Have you heard this information? (Yes, No); (2) Anxiety – How anxious did you feel when you heard this information? (1 Not at all, 7 Highly anxious); (3) Importance – How important do you think this information is? (1 Not at all, 7 Highly important); (4) Intended receiver – Who should know this information? (Family, Friend, Victims, Many Japanese, Many people abroad, Anyone, Other); (5) Intended transmission – How many people do you think should know this information?; (6) Self-accuracy – How accurate do you think this information is? (1 Not at all, 7 Highly accurate); (7) Estimated transmission – How many people do you think have already known this information at present?; (8) Others-accuracy – How accurate would others think this information is? (1 Not at all, 7 Highly accurate). Each tweet was presented in a random order.

**2. Criticism tweet** The design and procedure were the same as those of the rumor tweet phase.

**3. Demographic information** There were demographic questions and other questions about the degree of damage experienced and familiarity with Twitter and media.

**4. Debriefing** Each participant was explained the purpose of the experiment. It was emphasized that the tweets in the experiment might be false, and that the spread of false rumor was becoming a social problem after the disaster. In addition, for further information, we recommended useful books and websites that examined the false rumors related to the disaster.

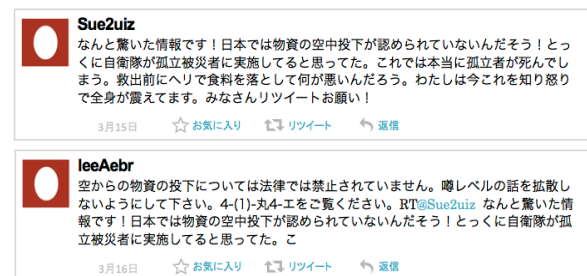


Figure 1: Top) A rumor tweet – “Air drop of supplies is not allowed in Japan! I though it has already been done by the Self-Defense Forces. Without it, the isolated people will die! I’m trembling with anger. Please retweet!” Bottom) A corresponding criticism tweet – “Air drop of supplies is not prohibited by the law. Please don’t spread rumor. Please see 4-(1)-丸 4-エ.”.

## Results and Discussion

On average, the tweets we used were relatively unfamiliar to the participants. The range of proportion of participants who were familiar with 10 rumors was from 0% to 38% ( $M = 22.5\%$ ). That of 10 criticisms was 0% to 43% ( $M = 9.8\%$ ). The overall results are shown in Appendix.

Table 1: Pearson's Coefficients of Correlation in Rumor and Criticism tweet

	Rumor tweet					Criticism tweet				
	1	2	3	4	5	1	2	3	4	5
1. Self-accuracy										
2. Others-accuracy	.62**					.79**				
3. Anxiety	.64**	.35*				.29	.14			
4. Importance	.80**	.44**	.71**			.73**	.59**	.57**		
5. Intended transmission	.29	.22	.02	.32*		.24	.28	.08	.26	
6. Estimated transmission	-.05	.14	-.23	-.11	.31*	.13	.16	.26	.03	.47**

Note. <sup>†</sup> $p < .10$ , \* $p < .05$ , \*\* $p < .01$

We removed outliers of intended and estimated transmissions using the Smirnov-Grubbs test. Analyses were repeated until no additional outliers were observed.

We examined the relationship between the four psychological factors (self-accuracy, others-accuracy, anxiety, and importance) and transmission (intended and estimated). Each mean of these factors was shown in Appendix. Table 1 shows Pearson's coefficients of correlation among these factors.

**Self-accuracy and Others-accuracy** Self-accuracy was positively correlated with the other three psychological factors: others-accuracy, anxiety, and importance ( $r = 0.62$ ,  $0.64$ ,  $0.80$ ,  $p < .001$ , respectively) in the rumor condition. In the criticism condition, it was positively correlated with others-accuracy and importance ( $r = 0.79$ ,  $0.73$ ,  $p < .001$ , respectively), while it was not significantly correlated with anxiety. The relationships between others-accuracy with the other factors showed the same pattern as self-accuracy.

In order to examine the relationship between self-accuracy and others-accuracy in detail, we performed a repeated-measures ANOVA on accuracy rate as the dependent variable, familiarity (familiar vs. unfamiliar) as the between subject factor, tweet type (rumor vs. criticism), and accuracy (self vs. others), as the within-subject factors. All three main effects were significant ( $F(1, 796) = 24.7$ ,  $21.7$ ,  $70.3$ , respectively,  $p < .001$ ): The familiar condition ( $M = 4.8$ ,  $SD = 1.7$ ) were more accurate than the unfamiliar condition ( $M = 4.2$ ,  $SD = 1.7$ ), the criticism condition ( $M = 4.7$ ,  $SD = 1.5$ ) were more accurate than the rumor condition ( $M = 3.9$ ,  $SD = 1.7$ ), and others-accuracy ( $M = 4.6$ ,  $SD = 1.4$ ) was higher than self-accuracy ( $M = 4.0$ ,  $SD = 1.8$ ).

The interaction between tweet type and accuracy was also significant ( $F(1, 796) = 66.3$ ,  $p < .001$ , Figure 2). Simple main effect of tweet type was significant only on self-accuracy ( $F(1, 1592) = 61.7$ ,  $p < .001$ ): The rumor condition ( $M = 3.4$ ,  $SD = 1.8$ ) was less accurate than the criticism condition ( $M = 4.7$ ,  $SD = 1.7$ ). Simple main effect of accuracy was significant only on rumor tweet ( $F(1, 796) = 136.6$ ,  $p < .001$ ): Self-accuracy ( $M = 3.4$ ,  $SD = 1.7$ ) was lower than others-accuracy ( $M = 4.5$ ,  $SD = 1.5$ ).

These results show that participants evaluated rumor tweets less accurately than criticism tweets; however, they estimated that others would evaluate rumor tweets as accurate as criticism tweets. This result indicates that participants think that they can detect unreliability of a

rumor tweet but that others cannot. There is a tendency for participants to underestimate the ability of others to evaluate accuracy of a rumor.

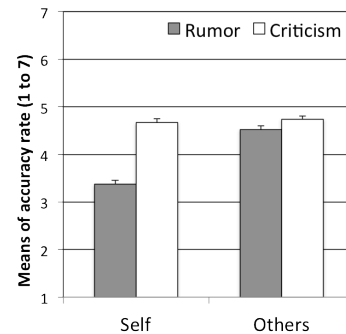


Figure 2: The analysis revealed that the bar of self-accuracy in the rumor condition was lower than the other three bars. Each bar shows the mean of 40 participants and 10 tweets. Error bars represent standard errors.

**Anxiety** The means of anxiety in the rumor and the criticism conditions were  $4.0$  ( $SD = 1.9$ ) and  $3.0$  ( $SD = 1.8$ ), respectively. In the rumor condition, anxiety was positively correlated with self-accuracy, others-accuracy, and importance ( $r = 0.64$ ,  $0.35$ ,  $0.71$ ,  $p < .001$ ,  $p = 0.05$ ,  $p < .001$ , respectively). On the other hand, anxiety was positively correlated with importance ( $r = 0.57$ ,  $p < .001$ ) in the criticism condition, while there was no correlation between anxiety and any accuracy. Thus, we found that the more accurate participants perceived rumors were, the more anxious they became. In contrast, anxiety was not correlated with accuracy in criticism tweet. In other words, if a tweet adopts a writing style in which it criticizes other tweet, participants' perceived accuracy of the tweet does not influence anxiety.

The inconsistency in accuracy and anxiety between the present study and past studies may be due to the differences in the measurement of transmission. An index to measure rumor transmission used in many studies was proportion assessed by dividing the number of rumor passed along by the number of rumor heard (e.g., Rosnow, et al., 1986; Rosnow, et al., 1988; DiFonzo & Bordia, 2000). By contrast, in the present work, the transmission was measured by asking to how many people a participant intended to transmit a rumor.

**Importance** The means of importance in the rumor condition and the criticism condition were 4.0 ( $SD = 1.9$ ) and 4.3 ( $SD = 1.9$ ), respectively. Importance was positively correlated with the other three psychological factors. It was also positively correlated with intended transmission in the rumor condition ( $r = 0.32, p = 0.04$ ). On the other hand, there was no significant correlation in criticism tweet.

**Intended and Estimated Transmission** Intended transmission was correlated with no psychological factors except for importance in the rumor condition. Estimated transmission was not correlated with any psychological factors. It was positively correlated only with intended transmission in both the rumor condition and the criticism condition ( $r = 0.31, 0.47, p = 0.05, 0.002$ , respectively).

The nonsignificant relationship between anxiety and rumor transmission is inconsistent with the past finding that anxiety was related to rumor spread (e.g., Anthony, 1992; Jaeger, et al., 1980; Prasad, 1935; Pezzo & Beckstead, 2006; Rosnow, et al., 1988; Walker and Beckerle, 1987). As the mean anxiety of rumor in this study was approximately at the center of the 7-point scale, it would be unlikely that we observed a floor or ceiling effect as Pezzo & Beckstead (2006) pointed out.

Of all psychological factors examined in the current study, importance was the only factor that was related to rumor transmission. The more important participants evaluated a rumor, the more they intended to transmit it. This result is congruent with the past studies showing that a positive relationship of importance to transmission (see DiFonzo & Bordia, 2007). Unlike the past work, in which the relationship between importance and transmission did not reach statistical significance (e.g.,  $r = 0.12, p > 0.30$  in Rosnow et al., 1988), the current study found a strong relationship between importance and transmission. This relationship was found only in intended transmission, but not in estimated transmission. Thus, participants distinguished these two questions of transmission; subjective importance of rumor had an effect on subjective intention of transmission, not on expected behavior of others.

### Rumors vs. Criticisms

A tweet type (rumor vs. criticism) by transmission (intended vs. estimated) analysis of variance was conducted, with means of 10 tweets (Table 2) in each tweet type as a dependent variable. The main effect of transmission was significant,  $F(1, 40) = 16.7, p < .001$ : Intended transmission was higher than estimated transmission. The main effect of tweet type and interaction between two factors did not reach statistical significance. This is, we think, because familiarity in both rumor and criticism was very low. Approximately only 16% of participants were familiar with the tweets. Most tweets were new to participants, and, thus, they estimated that a tweet had not been spread as much as they wanted to transmit it.

As a next step, each response of intended transmission and estimated transmission was classified into two conditions according to importance rate: low importance (1

to 4) and high importance (5 to 7). We performed this classification in the rumor condition and the criticism condition separately. There was no main effect of tweet type on intended transmission as the dependent variable by analysis of variance (ANOVA), with tweet type (rumor vs. criticism) and importance (high vs. low) as the between subject variables,  $F(1, 779) = 0.6, p = 0.4$ , while the main effect of tweet type was significant on estimated transmission as the dependent variable,  $F(1, 717) = 38.0, p < .001$ : Estimated transmission rates in the criticism condition ( $M = 1,036,225.4, SD = 2,789,937.5$ ) were higher than the rumor condition ( $M = 115,900.4, SD = 284,913.2$ ).

Table 2: Means and standard deviations for intended transmission and estimated transmission in the rumor condition and the criticism condition.

	Intended transmission		Estimated transmission	
	Rumor	Criticism	Rumor	Criticism
Mean	8,451,611	10,414,968	123,811	1,125,707
(SD)	(13,242,046)	(18,640,311)	(121,392)	(1,815,517)

Note. Each mean is of 40 participants.

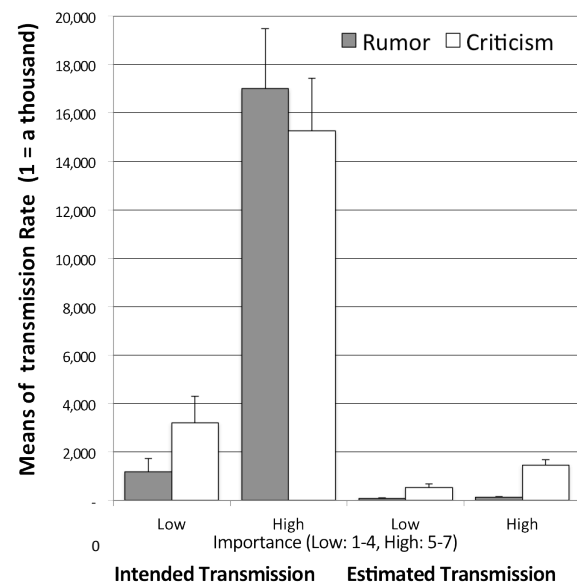


Figure 3: There was no significant difference between rumors and criticisms on intended transmission, while the two-way ANOVA on estimated transmission revealed the interaction between tweet type and importance. Error bars represent standard errors.

The interaction between tweet type and importance was also significant on estimated transmission  $F(1, 717) = 8.3, p = .004$  (see right four bars in Figure 3). The simple main effect of importance was significant only in the criticism condition,  $F(1, 375) = 10.4, p = 0.001$ : Estimated transmission rate of the criticism condition was higher in



high importance condition ( $M = 1,455,451.9$ ,  $SD = 3,250,638.1$ ) than low importance condition ( $M = 536,565.9$ ,  $SD = 2,011,841.9$ ). There was no significant difference in estimated transmission of the rumor condition between high and low importance conditions. The simple main effect of importance was significant on intended and estimated transmission,  $F(1, 779) = 62.3$ ,  $F(1, 717) = 11.4$ ,  $p < .001$ , respectively: transmission rate in importance high condition (Intended transmission:  $M = 144,816.0$ ,  $SD = 312,746.4$ , Estimated transmission:  $M = 1,455,451.9$ ,  $SD = 3,250,638.1$ ) was higher than importance low condition (Intended transmission:  $M = 94,578.9$ ,  $SD = 261,279.8$ , Estimated transmission:  $M = 536,565.9$ ,  $SD = 2,011,841.9$ ).

In intended transmission, there was no significant difference between rumor and criticism tweet. Regardless of tweet type, the more important a tweet was, the higher the participants' intended transmission was. This result shows a simple relationship between importance and intended transmission. On the other hand, importance had an effect on estimated transmission by interacting with tweet type. While the degree of importance of rumor tweets did not influence on estimated transmission, that of criticism tweets did influence on estimated transmission; the more important a criticism tweet was, the higher the participants' estimated transmission was. These results show a discrepancy between intended and estimated transmissions in terms of tweet type. If people actually could transmit rumor and criticism as they intended, rumor and criticism should be spread at the same rate in Twitter. Thus, they should estimate a similar pattern of transmission of rumor and criticism. However, participants estimated that important criticisms were more spread than important rumors. This may imply that estimated transmission, in the present study, reflects participants' wish that criticisms should be shared more than rumors.

With regard to criticism tweet, it was operationally defined, in the current study, as a tweet that criticized a rumor tweet by citing it. While accuracy of rumor was correlated to anxiety, accuracy of criticisms was not correlated to anxiety. This result indicates that accuracy of a criticism does not increase anxiety. One interpretation is that the perceived accuracy of criticisms reduces anxiety compared to rumors. Similarly, Bordis, DiFonzo, & Haines (2005) showed that rumor denials reduced anxiety by interacting with personal relevance and source credibility.

Besides, criticism tweets were evaluated more accurate and more important than rumor tweets. There are two important points here. First, approximately 90% participants were unfamiliar with the criticisms on average. Second, participants were not given any strong evidence to show that each criticism was truthful. Based on the definition of rumor (DiFonzo & Bordia, 2007, p.13), the criticisms presented in the current study are also considered a kind of rumor. In this perspective, the difference between a criticism tweet and a rumor tweet in this study was whether it adopts criticizing form citing an original tweet. Why were criticism tweets evaluated as more accurate and more important than rumor

tweets? One possible explanation is that, if a tweet has a criticizing form citing an original tweet, the accuracy and importance of the criticizing content are raised relatively to the accuracy and importance of the cited original tweet.

## Limitation and Future Research

The measurement of transmission is one limitation of this study. In terms of ethical consideration, tweet transmission was measured by asking intended and estimated transmission instead of measuring how participants actually spread tweets. The actual spread of tweet is also affected by the number of follower a user has. Thus, even if a participant intend to transmit a tweet to a few million people, if the user has only a few follower, the tweet will not be transmitted so much. We need to pay attention to the difference between intended or estimated transmission and actual transmission.

The order of rumor tweet and criticism tweet was fixed in the current study: All participants were given rumor tweets first, and then criticism tweets. Thus, the possibility, which responses to criticism tweets were influenced by the responses to rumor tweets, was not excluded. What if we receive criticism tweets first? Do psychological responses to criticisms help minimizing the spread of false rumor? Further research is needed to examine this possibility by comparing the experiment presented in the present study as a condition with another conditions with the different order of rumors and criticisms.

## Conclusions

The major contributions of the present study were (1) to reveal the relationship between psychological factors and information transmission using real tweets posted after the earthquake; and (2) to compare rumor tweet and criticism tweet. As a general conclusion, perceived importance seems to be a main predictor of tweet transmission. Better understanding of users' behavior through continued work in this area can help the design and use of social media systems, which in turn will help minimize the spread of false information during disasters, and enhance people's social media literacy.

## Acknowledgments

This research was supported by the National Science Foundation under grant IIS-1138658.

## References

- Allport, G. W., & Postman, L. J. (1947). *The psychology of rumor*. New York, NY: Holt, Rinehart, & Winston.
- Anthony, S. (1992). The influence of personal characteristics on rumor knowledge and transmission among the deaf. *American Annals of the Deaf*, 137, 44-47.
- Bordia, P., DiFonzo, N., Haines, R., & Chaseling, E. (2005). Rumors Denials as Persuasive Messages: Effects of Personal Relevance, Source, and Message Characteristics. *Journal of Applied Social Psychology*, 35(6), 1301-1331.

- Bordia, P., DiFonzo, N., & Schulz, C. A. (2000). Source characteristics in denying rumors of organizational closure: Honesty is the best policy. *Journal of Applied Social Psychology, 11*, 2301-2309.
- DiFonzo, N., & Bordia, P. (2000). How top PR professionals handle hearsay: Corporate rumors, their effects, and strategies to manage them. *Public Relations Review, 26*, 173-190.
- DiFonzo, N., & Bordia, P. (2007). *Rumor psychology: Social and organizational approaches*. American Psychological Association, Washington.
- Iyer, E. S., & Debevec, K. (1991). Origin of rumor and tone of message in rumor quelling strategies. *Psychology and Marketing, 8*, 161-175.
- Jaeger, M. E., Anthony, S., & Rosnow, R. L. (1980). Who hears what from whom and with what effect: A study of rumor. *Personality and Social Psychology Bulletin, 6*, 473-478.
- 荻上チキ (2011). 検証 東日本大震災の流言・デマ. 光文社新書 [Ogiue, K. *An examination of rumor and false rumor during the great east Japan earthquake*. Kobunsha, Tokyo]
- Pezzo, M., & Beckstead, J. (2006). A Multilevel Analysis of Rumor Transmission: Effects of Anxiety and Belief in Two Field Experiments. *Basic and Applied Social Psychology, 28*, 91-100.
- Prasad, J. (1935). The psychology of rumor: A study relating to the great Indian earthquake of 1934. *British Journal of Psychology: General Section, 26*, 1-15.
- Prasad, J. (1950). A comparative study of rumours and reports in earthquakes. *British Journal of Psychology: General Section, 41*, 129-144.
- Rosnow, R. L. (1991). Inside rumor: A personal journey. *American Psychologist, 46*, 484-496.
- Rosnow, R. L., Esposito, J. L., & Gibney, L. (1988). Factors influencing rumor spreading: Replication and extension. *Language & Communication, 8*, 29-42.
- Rosnow, R. L., & Foster, E. K. (2005). Rumor and gossip research. *APA Online: Psychological Science Agenda, 19*.
- Rosnow, R. L., Yost, J. H., & Esposito, J. L. (1986). Belief in rumor and likelihood of rumor transmission. *Language & Communication, 6*, 189-194.
- Sinha, D. (1952). Behaviour in a catastrophic situation: A psychological study of reports and rumors. *British Journal of Psychology, 43*, 200-209.
- 立入勝義 (2001). 検証 東日本大震災: そのときソーシャルメディアは何を伝えたか? ディスカバー・トゥエンティワン [Tachiiri, K. *An examination of the Great East Japan Earthquake: What did social media report during the disaster?* Discover Twenty One, Tokyo.]
- Walker, C. J., & Beckerle, C. A. (1987). The effect of anxiety on rumor transmission. *Journal of Social Behavior and Personality, 2*, 353-360.

Appendix. Rumor and criticism tweets and familiarity, accuracy, anxiety, importance, and transmission rate of the tweets.

Tweet type	Summary of tweet	Familiarity	Accuracy		Anxiety	Importance	Transmission	
			Self	Others			Intended	Estimated
1	R My friend at a insurance company said "Cancer insurance commercials stopped after the nuclear accidents."	25%	3.6	4.6	3.9	4.3	8,455,462	100,210
	C If you put it that way, sure. But that's a false rumor.	0%	4.9	5.0	2.7	4.0	9,018,592	356,047
2	R According to my friend, a radioactive material was detected from urine after he ate sushi.	38%	3.3	4.5	3.7	3.8	5,928,793	206,139
	C It's unclear the radioactive material was caused by the fish.	3%	4.7	4.6	2.8	4.2	7,573,040	38,752
3	R Toxic substance will drop with rain due to an explosion at Cosmo oil company.	8%	3.0	4.1	3.6	3.4	3,489,061	55,591
	C That's a definitely false rumor. NHK denied it.	43%	5.2	5.0	2.5	5.1	8,973,592	2,128,116
4	R Medical license is deprived by MEXT if a doctor gives a certificate of being exposed to radiation.	35%	3.6	4.8	4.5	4.1	8,508,545	141,873
	C The license cannot be deprived easily by MEXT or MHLW.	0%	4.8	4.8	2.9	4.0	13,904,373	582,892
5	R Robberies and rapes occurred during the Kobe earthquake.	3%	3.1	4.2	4.0	4.0	11,334,509	99,545
	C Few robberies and rapes occurred. Victims helped each other orderly. Why do you spread lies and false rumors? Stop it.	8%	4.7	4.8	3.5	4.9	7,057,882	1,284,337
6	R It was denied strongly, but after all, the meltdown occurred.	8%	3.7	4.9	4.3	4.2	15,360,119	108,083
	C Has the possibility of a meltdown been pointed out, hasn't it?	8%	4.3	4.1	3.2	4.1	16,840,780	1,531,965
7	R Air drop of supplies is not allowed in Japan!	30%	3.2	4.3	4.3	3.9	8,428,024	155,370
	C Air drop is not prohibited by the law.	0%	4.8	5.0	2.6	4.5	5,186,518	967,029
8	R Tokyo Electric Power Co.'s workers run and left. They were drinking in other city.	28%	3.0	4.2	3.8	3.4	2,876,451	49,975
	C Tokyo Electric Power Co. "The workers were found dead."	13%	4.2	5.0	4.4	4.3	12,279,433	1,561,848
9	R Did anyone watch "Senior vice transport minister Tsujimoto protested against the rescue operation by US army" on NHK?	33%	3.8	5.0	4.2	4.4	4,237,578	138,179
	C There's no source but the tweet, so it would be a rumor.	0%	4.5	4.6	2.9	3.7	7,497,978	392,898
10	R Chubu, Kansai, and Kyusyu Electric Power companies are beginning to transfer electricity to Kanto. Please cooperate!	20%	3.5	4.8	4.1	4.2	14,685,952	111,262
	C Transfer is impossible because of the difference in frequency.	25%	4.8	4.7	2.7	4.2	8,799,558	1,562,686

Note. R = rumor, C = criticism. Accuracy, Anxiety, Importance, Transmission = the means of 40 participants. Familiarity = the proportions of participants who answered that they heard the rumor.

# Ontological Properties of Animals in a Children's Dictionary With and Without Common-Sense Knowledge

**Julia M. Taylor (jtaylor@purdue.edu)**

Computer and Information Technology & CERIAS, Purdue University  
401 N. Grant Street, W. Lafayette, IN 47907-2021 USA

**Victor Raskin (vraskin@purdue.edu)**

Linguistics & CERIAS, Purdue University  
656 Oval Drive, W. Lafayette, IN 47907-2086 USA

**Christian F. Hempelmann (chempelm@purdue.edu)**

Linguistics and CERIAS, Purdue University  
656 Oval Drive, W. Lafayette, IN 47907-2086 USA

## Abstract

The paper applies a limited version of the resources of Ontological Semantic Technology to the descriptions of animals in the American Heritage First Dictionary and constructs a partial ontology from them. The explicitly mentioned properties in the descriptions are then supplemented by common-sense knowledge that the descriptions assume available to their young readership, and the output is compared to the previous one. The results, albeit modest, shed some interesting light on the most similar and dissimilar pairs of animals, as described in text.

**Keywords:** common-sense knowledge; Ontological Semantic Technology; children's dictionary; animal dataset; similarity.

## Introduction

The paper explores the common-sense knowledge that is necessary to fully understand the definitions/descriptions (henceforth, just descriptions) of around 100 animals in the 2007 edition of the American Heritage First Dictionary (AHFD 2007) aimed at children in grades K-2 (ages 5-8). It does not address common-sense reasoning. The purpose is to get a grasp on how implicit information affects the structure of perceived knowledge, in this case of animal descriptions, and similarity among entities.

The AHFD contains about 2,000 entries, claimed, almost entirely correctly, to be written with a controlled vocabulary so that that every description contains only words that also have entries in the dictionary. What is challenging in this design for our task is the possible implication of self-sufficiency, that is, of a much reduced dependency on the child's knowledge of the world, unstated explicitly in the natural language descriptions but (unconsciously) assumed to be present for full comprehension.

Most AFHD definitions for animals follow the "genus proximum, differentia specifica" format that is common for dictionaries, in particular for the classification of animal species: "Goldfish is a kind of *fish* [genus]. Goldfish are usually *small* and *orange* [differentia]." Apart from the differentiating specifics (small, orange) and the few properties that are specifics for the genus (in the AHFD, fish

live in water, have tails, can swim well) and are thus inherited, the remaining knowledge necessary to understand such definitions remains implicit, because it is presumed to be common-sense of different kinds. There is, for example, no mention of a fish's gills or fins.

Capturing common-sense knowledge is a daunting task. We are assuming that descriptions of the (animal) world of this dictionary requires less common-sense knowledge simply because this world is more restricted than that of an adult and the dictionary was obviously designed to accommodate that. If that is so, then getting a grasp of that knowledge may be more feasible than in case of a common, unlimited, adult-level natural language communication<sup>1</sup>.

The goal is, then, to illustrate how much information is lost when common-sense knowledge is not made explicit. Using the methods of computational semantics, specifically our Ontological Semantic Technology, we are taking advantage of the unique design of the dictionary to identify the required common-sense knowledge for a reasonably full comprehension of its animal descriptions. In this way, we aim to get a sense of its common-sense knowledge dependency. As a result, we also hope to clarify some issues concerning the very nature of common-sense knowledge and the feasibility of its computational acquisition and use, which is, as a matter of act, our primary and real concern.

In Section 1, we introduce the notions of 'hard' and 'soft' common-sense knowledge and explore its relation to underdetermination of reality by language and to saliency and, then, to ontology and natural language meaning, contingency, and instantiation.. In Section 2, we will briefly survey pertinent prior work. Section 3 will sketch out the Ontological Semantic Technology, our research tool as we applied it to the material. Section 4 compares the worldview on the animals that the descriptions define with the one complemented by the common-sense knowledge necessary to understand them. Section 5 discusses the results,

---

<sup>1</sup> The distinction between children as "novices" who know less about many domains than "expert" adults is well established (e.g., Carey 1985)-for better or worse.

identifies the strengths and weaknesses of our approach, and discusses the future lines of research.

## Kinds of Common-Sense Knowledge

### Hard and Soft Common-Sense Knowledge

We are introducing this new pair of terms to differentiate between two kinds of common-sense knowledge that a reader of the AHFD must possess to fully comprehend a description. If that reader does not understand a word in the description and that word has its own AHFD entry, he or she may access the required knowledge from that entry, where it is explicitly stated. So, from the point of the initial entry, this information is implied but from the point of the dictionary, it is explicitly stated. We call this information the ‘soft common-sense knowledge.’ If, on the contrary, some information that is needed for a full comprehension of the entry is not stated explicitly anywhere in the dictionary, we refer to it as the ‘hard common-sense knowledge.’ The paper focuses on the latter.

Starting with the randomly selected AHFD description of *snake*, we line up (see graph in [http://web.ics.purdue.edu/~vraskin/snake\\_new\\_label.pdf](http://web.ics.purdue.edu/~vraskin/snake_new_label.pdf)) the lexical chains underlying the entry dependency: if entry  $E(x)$  uses word  $y$  in the description of word  $x$ , and  $y$  is not previously mentioned in the chain,  $E(x)$  leads to  $E(y)$ . If  $E(i)$  does not evoke any new words in its description, it becomes a terminal in the dependency line. The longest, 10-node dependency line holds, starting with the topmost leftmost node and ending with the rightmost node at the bottom of the picture: SNAKE is-a REPTILE is-a ANIMAL is-a-not PLANT agent-of MAKE has-agent BEE agent-of FILL result-in FULL precondition-of-not HOLD unspecified ROOM. What this perfectly representative branch illustrates is that there is no consistent or predictable semantic dependency in the chain and that the vagaries of lexicographic connection can traverse the domain of knowledge, common-sense and other, in all directions, with some connections not easily explained.

Altogether, the knowledge required to understand every word in the description of *snake* as well as every word in the descriptions of those words, and, in turn, every word in the descriptions of those words, and so on to the end of the chain, is expressed in 86 entries. Realistically speaking, no 5-year-old will read all the entries: much more likely, they will have the requisite knowledge of the words. Nevertheless, this information is made available by the AHFD compilers, perhaps similarly to the glosses, footnotes, and explanatory appendices in adult-level materials. Its availability makes it not quite common-sense knowledge—so we refer to this explicit, but remote information, as weak common-sense knowledge.

### Underdetermination and Saliency

It is known that language underdetermines reality (see, for instance, Barwise and Perry 1983; Nirenburg and Raskin 2004): no matter how fine-grained or verbose the description of an event, there will be tons of details about

the situation that will remain unmentioned. If two men walk into the room, a report of that may include what they look like, what they wear, the speed of their movement, etc. But it will mention nothing about their places of birth, parents’ names and occupations, what cars they drive, what they had for breakfast, etc.

Now, all that knowledge exists, and common-sense knowledge includes that these people have a birth place, have parents, likely drive cars (especially if they’re Americans), etc. What is essential, however, is that most of the existing but implicit information is not prominent: much more likely, the prominence goes with the purpose of those people’s entrance into the room, whether there is any cause for alarm or displeasure, etc. The amount of prominent, or salient common-sense knowledge is much more limited in any situation.

Unfortunately, saliency (see Giora 2003: 13-38 and references there) is dynamic and fluctuates very rapidly. In AHFD, however, saliency may be conveniently seen as deliberately delimited by the availability of entries for words, thus reflecting the compilers’ notion of the mental model for a five-year-old’s world.

### Instantiation and contingency

In Ontological Semantic Technology (OST), the ontology consists of concepts and relations between them that are determined by properties. The concepts anchor lexical senses that are defined in the separate lexicon. Thus, one sense of the word *cat* is anchored in the ontological concept CAT. In a sentence, *A cat can jump from the floor to the top of a bookcase*, CAT is what the word *cat* means, i.e., a generic, any member of the class.

In the sentence, *Kisa the cat can jump from the floor to the top of a six-foot tall bookcase*, however, it is no longer a generic cat, but a specific instance of the concept, and the relationship between the meaning of the word *cat* and the ontological concept CAT is no longer that of generic anchoring. This instantiation makes the sentence contingent on a number of indices, such as the identities of the speaker and hearer, time, place, etc. (see Lewis 1972—cf. Bar Hillel’s 1954 comment on rare non-contingent sentences, such as, *Ice floats on water*).

We understand common-sense knowledge as non-contingent and involving concepts, not their instances. It is about what exists in the world, not what we know about particular objects or events. Our common-sense knowledge includes the fact that houses may be painted in various colors; it does not include the fact that Tom’s house is grey with burgundy trim.

So the common-sense knowledge left implicit by the AHFD is strong, non-contingent, and definitely less salient than the knowledge explicitly supplied by the AHFD in its descriptions.

## Prior Pertinent Work on Common-Sense Knowledge

Distinguishing common-sense knowledge from other implicit types of knowledge has been an issue in approaches to knowledge engineering, and while it always is a central one, it often remains implicit. Knowledge-based NLP has (re-)matured enough both to be able to need as well as to accommodate the type of “deep” knowledge that overlaps with the varying notions of common sense.

McCarthy (1959) is often cited as the earliest mention of common sense in the literature, but Bar-Hillel’s (1954) well-known example, “Little John played in his pen,” is already a clear indication of the necessity and importance of the common-sense knowledge—in this case, about relative sizes of objects.

Prominently, Lenat (1990) started an early large-scale systematic project on acquisition of common-sense knowledge, CyC. His method was hand-coding by a large number of research engineers, with a high turnaround and no well-defined acquisition methodology, which affected results and rendered them unusable for the NLP community.

Gordon and Schubert’s overview (2010) classifies current approaches to common-sense knowledge acquisition as: hand-authoring of rules, as in CyC; abstracting from clusters of propositions (e.g., Van Durme 2009); and directly interpreting general statements, such as glosses in dictionaries (e.g., Clark et al. 2008), akin to the approach of the present paper. Other researchers have used tagging, annotating, and/or generic machine learning techniques for automatically extracting implied common-sense knowledge from explicit text on the Web, about which Lin et al. (2004) have legitimate reservations, because explicit statements on the Web do not necessarily express common-sense knowledge.

Finally, we need to mention the area of research on common sense dedicated to children’s development of such knowledge, not least related to their overall linguistic-cognitive development. In particular, children’s knowledge about animals is one of the applications. Results that inform our present approach include that children focus on external features rather than internal organs, on habitats, on behavior relevant for humans (dangerous, edible) rather than cladistic accuracy (“Is a camel an ungulate?”), and that children’s knowledge is derived from observations as much as instructions, parents, or media (see Prokop et al. (2007), Tunnicliffe et al. (2007), Byrne et al. (2010)).

In our own previous work (Taylor et al. 2011a), we include in the common-sense knowledge rules of a separate resource the knowledge-of-the-world information that is not already contained in the ontology and lexicon (see next section). In the experiment there, we processed text with our system, and as part of routine quality assurance, added the necessary common-sense knowledge wherever we failed to interpret the text correctly because of the unavailability of this information in our resources (after we have excluded other, more banal reasons for the failure, such as an error in the resources or a bug in the software). Thus, we identified

as missing, for example, size classes necessary to understand spatial relations between physical objects, such as the understanding that a containing object should have greater dimensions than the (solid) object it contains.

In contrast to previous work, which addressed the identification and acquisition of common-sense knowledge by OST for the general purpose of processing text, this paper applies an appropriately limited version of our resources to a very limited corpus of a specific genre in an attempt to compare the ontological information following from the AHFD descriptions only with the ontological information arising from the descriptions supplemented by the common-sense knowledge that the descriptions imply in their readership.

## Brief Introduction to OST

Charniak’s (1972) often (mis)cited children’s story is used primarily to discuss inferencing and, hence, reasoning. It is even more suitable for exemplifying (in square brackets) the most common common-sense knowledge that OST has to deal with in order to fulfill its function of representing the meaning of natural language text accurately and comprehensively.

*Jane was invited to Jack’s birthday party.* [One brings presents to a birthday party. Presents are often purchased. To purchase something, one needs money.] *She wondered if he would like a kite.* *She went into her room and shook her piggy bank.* [Piggy banks contains money, usually coins. Coins make noise when shaken] *It made no sound.* [Coins make noise.] → (either there was no money in the piggy bank or just no coins but rather bills → in the former case, Jane may have lacked the money to buy the present)].

The italicized part is the original story; our formulation of the common-sense knowledge is in square brackets; the parenthesized part following the first arrow represents our formulation of inferences in reasoning, and while definitely pertinent to common-sense knowledge, it will be left out of this paper. It is noteworthy that the reasoning statements are contingent on the story while common-sense knowledge is generic.

The first and essential function of OST is to interpret the text of the story. The OST processor reads each sentence linearly and looks it up, word by word, in the OST English lexicon. Every sense of every (non-auxiliary, non-parametric) word in the lexicon is anchored in an ontological concept, with its properties and fillers, and the fillers can be restricted by the sense. The OST ontology, unlike its lexicons, is language-independent (see Nirenburg and Raskin 2004 for the basic theory of Ontological Semantics, and Raskin et al. 2010, Hempelmann et al. 2010, Taylor and Raskin 2011, Taylor et al. 2010, 2011a,b, for the much revised OST).

To use a greatly simplified example, the sense of the English word *invite* will be anchored in the ontological concept, probably also labeled “INVITE.” The label does not contain any but distinguishing information for the computer

and can be any ASCII combination—it is there just for the convenience of the human acquirer.

INVITE		
is-a	communicative event	
agent	human	
beneficiary	human	
theme	social-gathering	
purpose	entertainment	
invite		
Invite		
agent	[preceding NP]	
beneficiary	[following NP]	
theme	[to NP]	
...		

And the text meaning representation (TMR) of the first sentence of the story will result from matching the meaning of the NPs in the appropriate EVENT slots. The reality is, of course, harder, with more complex syntax, ambiguity, etc. The unenhanced-OST problem with the story is still more advanced: while TMR for each sentence is not hard to produce, the system will not be able to relate the sentences to each other, and the text will lack cohesiveness.

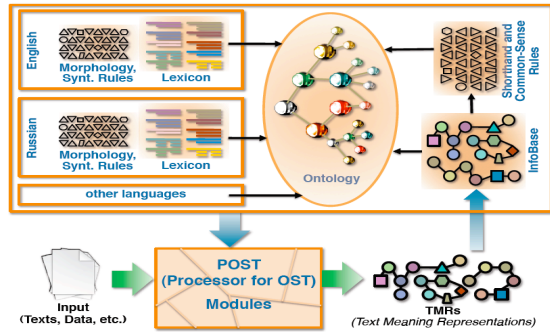


Figure 1: OST Architecture

In OST, the information processed prior to computing the TMR of the current sentence is used to clarify, complement, and disambiguate the current representation process. In this case, that information would be helpless and useless because, other than Jane as the agent, no previous sentence in the story even mentions objects in the following sentences, and it is the common objects (or events) that the anaphora/coreference resolution establishes as bridges between and among sentences. Jane will emerge from the story, as interpreted by the unenhanced OST, as performing three unrelated actions. It is the common-sense knowledge statements in the square brackets above that have to provide such common objects to make OST processing possible: the bridge words are underlined in the story above., and the common-sense knowledge enhanced text can be processed by OST normally.

This is why we recently added to the OST architecture (Figure 1 above) the common-sense knowledge resource (Taylor et al. 2011a) and the methodology of adding to it when the TMRs fall short of the (often hypothetical) gold standard (cf. Allen et al. 2008).

## Ontology of Descriptions and Ontology With Common-Sense Knowledge

In general, we are interested not only in reading and understanding a text, but also in structuring information that this text contains, as well as enhancing our ontology when newly acquired information requires. We are using information about animals from AHFD to see whether such task is possible. We then check whether supplying additional information (common-sense knowledge left implicit in the dictionary) would help with the task (cf. Perfors et al. 2005; Kemp et al. 2006).

Typically, a hierarchy is perceived as one of the most important properties in ontology construction. All animal descriptions of the dictionary provide such information. Unfortunately, sometimes a word is used that may have multiple senses (such as *cat* being a domestic cat or feline) thus creating a flawed hierarchy. One of the goals, then, is to identify such descriptions.

The proposed measure is conditional on an accepted membership assumption. If we assume the veracity of “B is A” as a reference point, which gives us a certain amount of knowledge about B in terms of its properties, we estimate the extent to which “C is B” is (dis)confirmed as

$$\frac{\sum_i 2^{-n} * w * hier_{P_i(C,B)}}{num(i)} \text{ where } w \text{ is a property weight, and}$$

$$hier_{P_i(C,B)} = \begin{cases} 1, P.b \in D_B \text{ \& } P.c \in D_C \text{ \& } b = c \\ -1, P.b \in D_B \text{ \& } P.c \in D_C \text{ \& } b \neq c \\ 0.1, P \in D_C \text{ \& } P \notin D_B \\ 0, \text{ otherwise} \end{cases}$$

Placement in the hierarchy as well as concepts’ properties (may) affect similarity between concepts. For the purposes of this paper, we assumed that the properties that are taken into account are all equally weighted. We measure similarity

of two concepts as  $\frac{\sum_i 2^{-n} * w * sim_{P_i(A,B)}}{num(i)}$  where  $sim_{P_i(A,B)}$  is

defined as:

$$sim_{P_i(A,B)} = \begin{cases} 1, P.a \in D_A \text{ \& } P.b \in D_B \text{ \& } a = b \\ -1, P.a \in D_A \text{ \& } P.b \in D_B \text{ \& } a \neq b \\ 0.1, P \in D_A \text{ \& } P \notin D_B \text{ \& } P \notin D_A \text{ \& } P \in D_B \\ 0, \text{ otherwise} \end{cases}$$

## Results and Conclusion

We first wanted to see what kind of structure we would get from the descriptions without the use of common sense. We calculated pair-wise similarity measurements for all animals with AHFD descriptions. The similarities ranged from -1.25 to 0.78. It is possible for the similarity to be -N where N is the number of properties in both descriptions and all properties in the descriptions match but their fillers do not. Having calculated the mean and standard deviation, we looked at the results that were at least 3 standard deviations away from the mean as most similar cases and most dissimilar ones. The dissimilar pairs were: ant/chicken, ant/crocodile, ant/pony, ant/whale, bee/chicken, beetle/chicken,



bug/chicken, bug/shark, bug/whale, butterfly/chicken, caterpillar/chicken, caterpillar/crow, caterpillar/whale, chicken/cricket, chicken/fly, chicken/mosquito, chicken/moth, chicken/whale, cricket/whale, crocodile/whale, mosquito/whale, moth/shark, moth/whale, turtle/whale.

It should be noted that, with the exception of the chicken/whale, turtle/whale, and crocodile/whale pairs, the dissimilar pairs contain insects. One member of the pair is (typically) a bird or a mammal that is somehow different from the rest of its class, thus deserves an explicit clarification, such as a whale being a mammal. For some reason, insects also received a fairly large amount of description and thus were easy to contrast with other animals.

The similar pairs are: ape/monkey, bear/panda, bee/moth, beetle/butterfly, beetle/cricket, beetle/fly, bug/caterpillar, butterfly/cricket, butterfly/fly, camel/giraffe, caterpillar/cricket, caterpillar/moth, cricket/fly, donkey/zebra, eagle/hawk, fox/wolf, goose/turkey, horse/pony, leopard/lion, lion/tiger. Again, (an expected) a pattern can be noticed here: those animals that received a lot of similar descriptions are being selected.

There were 7 animals or categories in the dictionary that were used in the is-a relations other than to indicate an offspring of an animal. These categories were: animal, insect, bird, fish, reptile, cat, and horse. Mammal got an entry in the dictionary but was not used in any of the descriptions. We excluded entries that indicated *a young animal*, such as *a kitten is a young cat*. We calculated the mean and standard deviation of each animal relative to the above 7 categories using the *hier* metric described above. We assumed that if an entry had a description that X is Y, and  $\text{hier}(X, Y)$  was lower than the mean for that overall category, the definition should be questioned and should not be used for hierarchy construction. The following entries were so affected: bat is-a animal, crab is-a animal, goat is-a animal, hippo is-a animal, sheep is-a animal, whale is-a animal, ostrich is-a bird, tiger is-a cat, lion is-a cat.

There are several explanations for the results: cat is defined as a domestic animal, and thus, of course, cannot be a parent of wild animals. Crab has more features that puts it next to fish, and so does whale, including the description of the habitat. Hippo is mostly described swimming in lakes and rivers. Bat is similar in its description to a bird. Goat and sheep created a puzzle for us. However, we considered it to be a success to have only 2 problematic entries.

Interestingly, contradicting the dictionary, the metric suggests that donkey and zebra should be types of horse; dog and hamster should be types of cat. These entries suggest that there is not enough differentiation between the affected animals for them to be correctly classified.

We therefore wanted to see if the ratio changes when the omitted common-sense knowledge is added to the descriptions and if some puzzling results are corrected. The common-sense knowledge consisted of a number of additional animal properties, explicitly stated in some

descriptions but omitted from others, with clearly implied values, so we added that information directly to the ontology as it emerged from the descriptions. The addition to common-sense knowledge solved the hierarchy problem of animal in the previous experiment not being an animal, and did not introduce any additional problems.

The distribution of the resulting similarity is shown in Figure 2. As seen there, listing results that are 3 standard deviations away from the mean proved to be impractical, although that was done in the first experiment,. Thus, the results below reflect the same number of dissimilar pairs as the first experiment: ape/duck, ape/swan, duck/snail, crocodile/snail, chicken/snail, crow/snail, alligator/snail, ape/goose, beaver/snail, eagle/snail, hawk/snail, goose/monkey, fox/snail, hamster/snail, duck/monkey, ape/snail, deer/swan, ape/crab, goose/snail, camel/snail, jellyfish/monkey, ape/penguin, bear/snail, goat/snail. As with previous results, there is a concept that is most dissimilar to others (snail), and the dissimilarity looks plausible (all below 0).

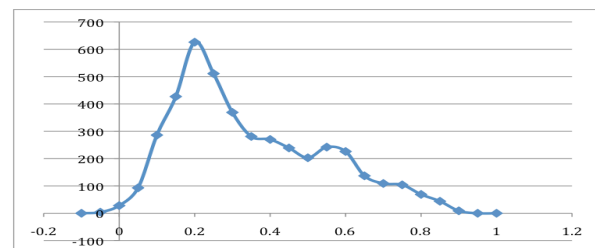


Figure 2. Distribution of similarity of animals

The pairs that are most similar are again several insects and birds, as well as eagle/hawk, crab/lobster, hippo/rhinoceros, bull/cow, dolphin/whale, horse/pony, cow/sheep, cow/pony, cow/sheep, frog/toad, pig/pony, lion/tiger, mouse/rat, jellyfish/octopus, donkey/zebra, spider/worm. As expected, the similarity results look (more?) reasonable with common-sense knowledge (Figure 3).

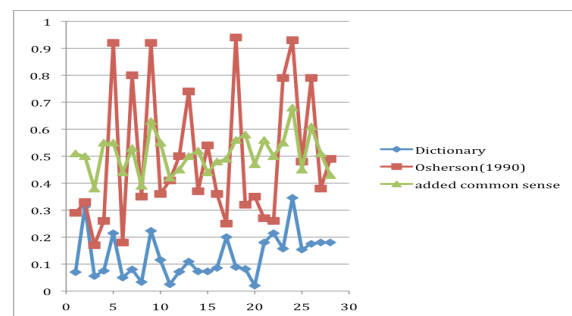


Figure 3: Pair-wise similarity of 7 animals.

While the most similar and least similar results look good, the middle section will need to be improved. Figure 4 shows pair-wise comparisons of perceived similarity between cow, dolphin, elephant, horse, mouse, rhino, seal, and squirrel. We anticipate these results to improve when weights are added to properties.



## Conclusion

We have demonstrated, on a very limited corpus of animal descriptions intended for a very young audience, that it is possible to detect semantic structure in natural language descriptions as well as pointing to flawed descriptions. The results improve with the addition of common-sense knowledge omitted from but implied by the descriptions. The specific material, from a children's dictionary that was designed to limit the amount of world knowledge that the young reader could be counted on contributing, helped us delimit the common-sense knowledge. It is clear, of course, that this method of defining this elusive resource is not useful outside of this artificially restricted environment but the convenient handle to it was too tempting to resist.

## References

- AHFD (2007). The American Heritage First Dictionary. Boston-New York: Houghton Mifflin.
- Allen, J. F., Swift, M., & de Beaumont, W. (2008). Deep semantic analysis of text. *Proc. Step '08*. Venice.
- Bar-Hillel, Y. (1954). Indexical expressions. *Mind* 63, 359-379. Reprinted in his *Aspects of Language*, Jerusalem: Magnes, 1970, 69-88.
- Barwise, J., & Perry J. (1983). *Situations and Attitudes*. Cambridge, MA: MIT Press-Bradford.
- Byrne, J., Dale-Tunncliffe, S., Patrick, T., & Grace, M. (2010). Children's understanding of animals in their everyday life in the UK and USA. *Proceedings of the 7th Conference of European Researchers in Didactics of Biology ERIDOB*.
- Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.
- Charniak, E. 1972. Toward a Model of Children's Story Comprehension. Artificial Intelligence Technical Report Number 266, Department of Computer Science, Massachusetts Institute of Technology, Cambridge, MA.
- Clark, P., Fellbaum, C., & Hobbs, J. (2008). Using and extending WordNet to support question answering. *Proceedings of the 4th Global WordNet Conference*, 111-119.
- Giora, R. (2003). *On Our Mind: Salience, Context, and Figurative Language*. New York: Oxford University Press.
- Gordon, J. M., & Schubert L. K.. (2010). "Quantificational Sharpening of Commonsense Knowledge", in: Havasi et al., 27-32.
- Havasi, C., Lenat, D. B., & Van Durme, B. (Eds.). (2010). Commonsense Knowledge: Papers from the AAAI Fall Symposium. Menlo Park, CA: AAAI Press, 2010.
- Hempelmann, C. F., Taylor, J. M., & Raskin, V. (2010). Application-guided Ontological Engineering. *Proceedings of the International Conference on Artificial Intelligence*, Las Vegas, NE.
- Kemp, C., Tanenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. 2006. Learning systems of concepts with an infinite relational model. *Proc AAAI-06*.
- Lenat, D. B. (1990) CYC: Toward programs with common sense, *Communications of the ACM* 33:8, 30-49.
- Lewis, D. (1972). General semantics. In Davidson, D. and Harman, G. (Eds.), *Semantics of Natural Language*, Dordrecht - Boston: Reidel.
- Liu, H., & Singh, P. (2004). ConceptNet—a practical commonsense reasoning tool-kit, *BT Technology Journal*, 22:4. 211-226.
- McCarthy, J. (1959). Programs with common sense. *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, 75-91. London: Her Majesty's Stationary Office.
- Nirenburg, S., & Raskin, V. 2004. *Ontological Semantics*. Cambridge, MA: MIT Press
- Opfer, J. E. & Siegler, R. S. (2004). Revisiting preschoolers' *living things* concept: A microgenetic analysis of of conceptual change in basic biology. *Cognitive Psychology* 49, 301-332.
- Osherson, D. N., Wilkie, O., Smith, E. E., López, A., and Shafir, E.1990. Category-based induction. *Psychological Review*, vol. 97, no. 2, 185-200.
- Perfors, A., Kemp, C., Tenenbaum, J. B. 2005. Modeling the acquisition of domain structure and feature understanding. *Proc. CogSci-05*.
- Prokop, P., Prokop, M., Dale-Tunncliffe, S., & Diran, C. (2007). Children's ideas of animals' internal structures. *Journal of Biological Education*, 41-2, 62-67.
- Raskin, V., Hempelmann, C. F., & Taylor, J. M. (2010). Guessing vs. knowing: The two approaches to semantics in natural language processing, *Annual International Conference Dialogue 2010*, 642-650, Bekasovo (Moscow), Russia.
- Taylor, J. M., & Raskin, V. (2011). Understanding the unknown: Unattested input processing in natural language, *Proc. FUZZ-IEEE-11*, Taipei, Taiwan.
- Taylor, J. M., Hempelmann, C. F., & Raskin, V. (2010a). On an automatic acquisition toolbox for ontologies and lexicons in Ontological Semantics, *International Conference on Artificial Intelligence*, Las Vegas, NE.
- Taylor, J. M., Raskin, V. & Hempelmann, C. F. (2011a). From disambiguation failures to common-sense knowledge acquisition: A day in the life of an ontological semantic system. In Gilof, R., (Ed.), *Proceedings of the Web Intelligence and Intelligent Agent Technology Conference*. IEEE.
- Taylor, J. M., Raskin, V., & Hempelmann, C. F. (2011b). Towards computational guessing of unknown word meanings: The Ontological Semantic approach, *Proc of CogSci-11*, Boston, MA.
- Tunncliffe S. D., Boulter, C., & Reiss, M. (2007). Pigeon – friend or foe? Children's understanding of an everyday animal. *Proceedings of the British Educational Research Association Annual Conference*.
- Van Durme, B., Michalak, P., & Schubert, L.K. (2009). Deriving Generalized Knowledge from Corpora Using WordNet Abstraction. *Proceedings of EACL*, 808-816.

# An Experimental Examination of Emergent Features in Metaphor Interpretation Using Semantic Priming Effects

Asuka Terai (asuka@nm.hum.titech.ac.jp)

Global Edge Institute, Tokyo Institute of Technology,  
2-12-1, Ookayama, Meguro-ku, Tokyo, 1528550 JAPAN

Robert L. Goldstone (rgoldsto@indiana.edu)

Psychological and Brain Sciences, Indiana University,  
Bloomington, IN 47405 USA

## Abstract

In comprehension of the metaphor “*TOPIC is VEHICLE*,” emergent features in the interpretation of metaphors are characteristic neither of the topic nor the vehicle. An experiment examines the hypothesis that new features emerge as metaphoric interpretations through association with non-emergent features connected with the topic, vehicle, or both. In the experiment, participants were presented with a non-emergent feature as a prime, a metaphor, and an emergent feature, sequentially. Participants were then asked to respond as to whether the emergent feature is an appropriate interpretation of the metaphor. The results showed that primed non-emergent features derived from the vehicle facilitate the recognition of emergent features. The results support an account in which new features emerge through two processes – non-emergent features are recognized as interpretations of the metaphor and then these non-emergent features facilitate the recognition of emergent features. **Keywords:** Metaphor comprehension; Feature emergence; Interaction.

## Introduction

In this research, we examined the process of feature emergence, which is realized in comprehension of metaphors taking the form of “*TOPIC is VEHICLE*”, such as “Education is a gateway.” In previous papers, interpretations (features) of this kind of metaphor were classified into four types: common features, topic features, vehicle features and emergent features (Becker 1997; Gineste, Indurkha & Scart, 2000; Nueckles & Janetzko 1997). When an interpretation is thought of in relation to both the topic and the vehicle, it is regarded as a common feature. When an interpretation is thought of as a characteristic of the topic (or of the vehicle), it is referred to as a topic feature (or a vehicle feature). The common, topic and vehicle features are regarded as non-emergent features. Finally, emergent features are not typically thought of in relation to either the topic or the vehicle alone, but do come to mind when the topic and vehicle enter into a metaphoric comparison. For example, for the metaphor “Ideas are diamonds,” the feature “come in a flash” is a topic feature, “beautiful” is a vehicle feature, “precious” is a common feature because it is listed as a feature when people are given either “ideas” or “diamonds” by themselves, and “unique” is an emergent feature because it is not listed for either word by itself, but is listed when the words are paired.

Previous research (Gineste et al., 2000) made a list of features and reported that over 60% of metaphoric interpretations are emergent features. Emergent features are thus

prevalent and play an important role in metaphor comprehension. Furthermore, the authors conducted an experiment using priming effects. In their experiment, emergent, topic-term or vehicle-term features were presented and participants judged whether the feature was related to the primed metaphor (topic-/vehicle-term) or not. Emergent features required a longer response time to be regarded as a feature of the prime than topic or vehicle features, when the features were tested with topic-term or vehicle-term primes. When tested with the metaphor as the prime, the topic and vehicle features required longer response times than did the topic-term or vehicle-term as the prime. However, the emergent features did not change their response times from one prime condition to another. As a result of these results are consistent with the interaction theory of metaphor (Black 1979), which suggests that metaphor comprehension is a product of an interaction between the target and the vehicle concepts.

However, there is also evidence that links emergent interpretations asymmetrically with topics and vehicles. Becker listed interpretations of metaphors in an experiment. She reported that altering a metaphor’s vehicle (e.g. “A smile is a knife” vs. “A smile is a pearl”) produced greater changes in emergent content than did altering the topic (e.g. “A smile is a knife” vs. “Teeth are knives”). This suggests that emergent features are influenced primarily by one’s representation of the vehicle. Nueckles and Janetzko (1997) introduce the idea that metaphor comprehension proceeds in analysis-based and synthesis-based stages. According to their idea, there is first an analysis of the lexical meanings of the topic and vehicle during the analysis-based stage. If the topic and vehicle have sufficient similarity, the metaphor comprehension does not proceed to the synthesis-based stage. For cases in which the topic-vehicle similarity is not sufficiently high, a shift to synthesis-based processing occurs. In the later case, the metaphor comprehension is achieved through a construction of new components of meaning by synthesis of the topic and the vehicle. It is during this second phase that emergent features would be generated.

Previous computational models of metaphor comprehension have been constructed under the assumption that emergent features are emphasized more than non-emergent features through interactions among features in metaphor comprehension (Utsumi, 2000; Terai & Nakagawa 2007, 2008,

2010). All of these models function to increase activation of emergent features beyond that of non-emergent features by incorporating interactions among features. Terai and Goldstone (2011) reported that emergent features require more time to be recognized during a metaphoric interpretation than do non-emergent features. Conversely, when a relatively long time period was allowed for metaphor processing, then recognition of non-emergent features was diminished. This suggests that non-emergent features that are true of one metaphor term but not the other have reduced activation as metaphor processing continues. The results support the kind of positive and negative interaction among features assumed by the computational models above, and also supports Nueckles and Janetzko's (1997) two-process assumption. In particular, the meanings of the topic and the vehicle are emphasized as non-emergent features, and then with ongoing interactions among features, emergent features are discovered as valid interpretations. However, empirical evidence is still lacking to support the details of this mechanism.

Some previous research used a priming paradigm to investigate the roles played by the topic and the vehicle in metaphor comprehension and reported that metaphor comprehension was facilitated by presentation of either a vehicle or topic concept (Wolff & Gentner, 2000; McGlone & Manfredi, 2001). McGlone and Manfredi used a sentence ascribing a metaphor-irrelevant or metaphor-relevant property to a topic or a vehicle as a prime. They found that all of these presentations, including the presentation of the sentences ascribing metaphor-irrelevant properties to topics, facilitated metaphor comprehension with the exception of sentences ascribing metaphor-irrelevant properties to vehicles.

Thus, we conducted an experiment employing priming effects of non-emergent features (common, topic and vehicle features) in order to investigate the role played by these features in processing emergent features and test the two-process assumption of feature emergence. If the two-process assumption is correct, non-emergent features should activate emergent features. An unresolved issue concerns the kinds of non-emergent features that most influence activation of emergent features.

## Experiment Method

In this experiment, we examined the priming effect of non-emergent features on processing of emergent features in metaphor comprehension.

### Participants

134 undergraduates participated in this experiment. All participants were native English speakers.

### Materials

We selected 39 metaphors of the form "*TOPIC is VEHICLE*" from Becker (1997) as "target metaphors." Becker (1997) asked participants to list features of individual words and interpretations of metaphors involving those words. She categorized the resulting features into four types: emergent, com-

mon, topic, and vehicle features. Based on her categorization and feature listings, for each of these 39 metaphors, 1 to 4 emergent features and 1 to 4 non-emergent features were selected. 114 emergent features, 26 common features, 29 topic features and 32 vehicle features were selected. The types of the non-emergent features (common, topic or vehicle feature) selected for a given metaphor differed from each other. Three native English speakers checked these selected features to ascertain whether a feature and the word is an appropriate interpretation of the metaphor or not. For the selected items, at least one judge recognized the feature as an apt interpretation of the metaphor.

In addition, another 39 metaphors of the form "*TOPIC is VEHICLE*" from previous research (Gentner & Clement 1988, McGlone & Manfredi 2001) were used as "irrelevant metaphors" in a baseline condition.

### Procedures

The procedures are shown in Figure 1. In the prime condition, participants were first presented with a non-emergent feature as a prime on a screen for 2 seconds. In the no-prime condition, no prime was presented. In both conditions participants were then asked to interpret a metaphor that was presented on the screen for 3 seconds. After presentation of the metaphor, an emergent feature was presented as a target word. The participants were asked to respond "Yes" or "No" depending on whether the word was related to the metaphor or not. Participants responded by pressing the "p" key ("Yes") or "q" key ("No") within 6 seconds. They were asked to respond as fast as possible without sacrificing accuracy. If they could not respond within 6 seconds, the feature disappeared and the text "Your response is too slow" appeared on the screen. The fixation point was presented between trials. For each metaphor, the combination of the prime and emergent features was randomized. The target metaphor and baseline conditions were presented equally often, but the presentation frequencies of feature conditions were dependent on the number of features.

In order to distinguish between the relationship between just a prime (non-emergent features) and target words (emergent features) and the interaction between them in metaphor understanding, we used irrelevant metaphors as a baseline. For example, if presentation of the common feature "beautiful" as a prime for the metaphor "Stars are diamonds" facilitates recognition of the subsequent emergent feature "amazing," there are two possible cognitive mechanisms. One is that "beautiful" influences the metaphor understanding process and facilitates recognition of "amazing" as an interpretation of the metaphor. The other is that "amazing" directly relates to "beautiful" much in the same way that "doctor" is related to "nurse," and so presentation of "beautiful" facilitates "amazing" regardless of the metaphor understanding process. Thus, we also employed irrelevant metaphors as a baseline. If presentation of "beautiful" does not influence judgments of the interpretation of an irrelevant metaphor (e.g. "Crime is a disease") but does influence judgments of the related metaphor, then this will be taken as evidence that not only the

relationship between “beautiful” and “amazing” is relevant, but also the interaction between “beautiful” and “amazing” have a role in metaphor understanding. Therefore, in the target metaphor condition, prime and target words are listed as interpretations of an intervening target metaphor. In the baseline condition, an irrelevant metaphor is presented between prime and target words, which do not relate to it. Therefore, there were two conditions of metaphors (target metaphor and baseline conditions) and four conditions of primes (common, topic, and vehicle features, and no-prime). After all trials, participants were asked to evaluate aptness and conventionality of the metaphors on the scale of 1 (highly inappropriate or rare metaphor) to 5 (highly appropriate or common metaphor).

## Results

The data of two participants were removed because they responded “Yes” to more than half of the items in the baseline condition. We analyzed the remaining data obtained from 132 participants. The average rate at which the participants responded within the time limit was 99.4%.

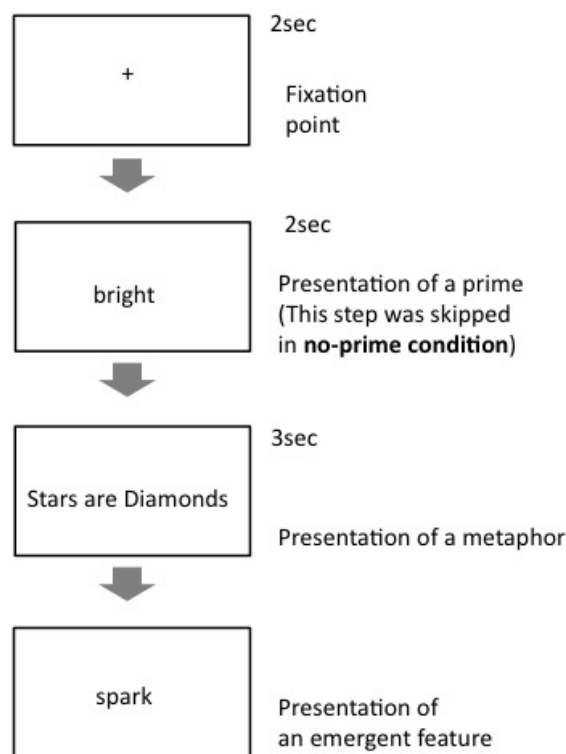


Figure 1: Time course of a trial in the experiment. In the target metaphor condition, prime and target words are interpretations of the metaphor which was presented between them. In the baseline condition, the primes and targets do not relate to the metaphor.

## Analysis of all data

The averages of response times are shown in Table 1. In the target metaphor condition, the response time is longer than in the baseline condition ( $F(1, 131) = 6.43, p < .05$ ). There is also a main effect of prime kind ( $F(3, 393) = 7.38, p < .01$ ). Combining across difference of kinds of prime and metaphor conditions, primes made participants respond faster. The results show that the prime facilitated responding to target words (emergent feature) and made judgment to them faster in both conditions. The results indicate that there is a relationship between a prime (non-emergent features) and target words (emergent features).

We analyzed the difference between response times depending on the response. Average response times when the participants respond “Yes” or “No” are shown in Table 2. Thirty participants responded “Yes” (“No”) for all emergent features with any kind of prime in the target metaphor condition (in the baseline condition) and these data were removed for this analysis. There was an interaction between metaphor condition and responses ( $F(1, 101) = 86.1, p < .01$ ). Participants responded “Yes” significantly faster than “No” in the target metaphor condition. Conversely, they responded “No” faster than “Yes” in the baseline condition. In the target metaphor condition, when participants responded “Yes,” the response is correct in the sense that the materials were designed so that the target would be an apt interpretation of the metaphor. Similarly, in the baseline condition, a “No” response would be correct. This suggests that people might be making errors not through fast guessing, but because of close competition between “Yes” and “No” responses. And, when participants made errors, there was a possibility that they could not interpret the metaphor. Thus, the response times confirmed an influence of the primes to “the correct responses.” So, we tested the “correct” responses.

## Analysis of correct responses

The average response times when the participants responded with the “correct” response are shown in Figure 2. The results show that the primed non-emergent features (common, topic, vehicle features) facilitated processing of emergent features ( $F(3, 393) = 8.12, p < .01$ ) in both conditions.

Therefore, we tested proportions of correct responses. The proportions of “Yes” responses in the target metaphor and “No” responses in the baseline condition are shown in Figure 3.

The proportions were analyzed using a two-way ANOVA<sup>1</sup>. The proportion of correct responses in the baseline condition is higher than in the target metaphor condition ( $F(1, 131) = 15.7, p < .01$ ). The response times and the accuracy rates indicate that it was more difficult to recognize an emergent feature as an interpretation of the metaphor than to find that

<sup>1</sup>When an arcsine transformation was applied to the proportion data because of the restriction of these data to a 0-1 range and the proportions were analyzed using a two-way ANOVA after transformation, the results show the same tendency that are indicated when the non-transformed proportions are analyzed.

Table 1: Averages of the response times (milliseconds) for all data. Standard deviations are shown in parentheses.

		Metaphor (condition)	
		Target metaphor (Target metaphor condition)	Irrelevant metaphor (Baseline condition)
Kinds of prime	Common feature	1769.0 (957.0)	1724.2 (905.4)
	Topic feature	1759.8 (933.1)	1716.5 (927.2)
	Vehicle feature	1756.6 (877.8)	1718.2 (908.6)
	No prime	1848.7 (952.2)	1797.5 (934.8)

Table 2: Averages of the response times (milliseconds) depending on their responses. Standard deviations are shown in parentheses.

		Target metaphor condition	
		Response	
		“Yes”	“No”
Kinds of prime	Common feature	1704.3 (914.2)	1919.8 (1035.2)
	Topic feature	1726.6 (922.3)	1830.2 (952.3)
	Vehicle feature	1701.3 (891.6)	1861.6 (983.3)
	No prime	1816.3 (929.9)	1921.0 (997.1)
		Baseline condition	
		Response	
		“Yes”	“No”
Kinds of primes	Common feature	1839.3 (1080.4)	1687.9 (840.0)
	Topic feature	1821.8 (1010.5)	1684.1 (830.3)
	Vehicle feature	1851.9 (1042.3)	1679.6 (862.6)
	No prime	1848.9 (1056.8)	1781.8 (894.1)

the emergent feature was not an interpretation of the irrelevant metaphor. Furthermore, there is a two-way interaction ( $F(3, 393) = 4.16, p < .01$ ). There are no significant differences among the proportions correct in the baseline condition, however, the accuracy with the vehicle primes is significantly lower than with the common primes or without a prime in the target metaphor condition at a  $p < .05$  level. This means that presentation of a non-emergent feature as a prime affected the metaphor understanding process and that vehicle primes inhibited recognition of emergent features as interpretations of the metaphor. The difference of the results in the two conditions suggests an interaction between non-emergent and emergent features in metaphor comprehension.

Furthermore, there is no significant difference between proportions of correct responses with and without a com-

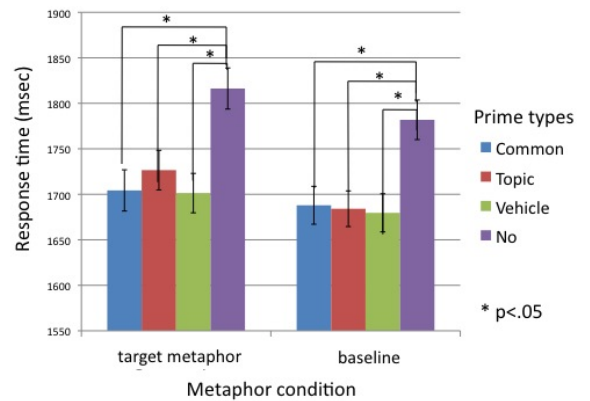


Figure 2: Average response times for correct responses. (The response time when they responded “Yes” in the target metaphor condition and when they responded “No” in the baseline condition.)

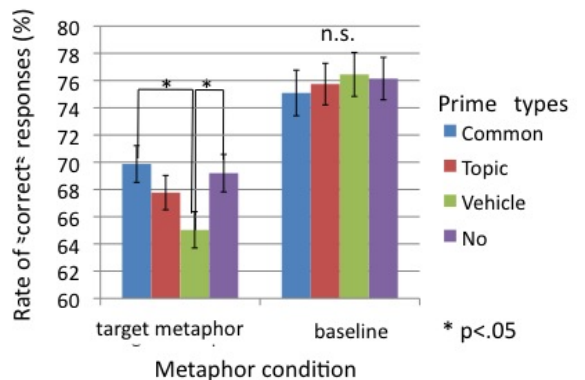


Figure 3: Average proportions of correct responses. (Average proportions of “Yes” responses in the target metaphor condition and that of “No” responses in the baseline condition.)

mon prime in the target metaphor condition. However, the proportion with common primes is slightly higher than without a prime. That is, the presentation of the common feature might make the emergent features recognized more easily as the metaphor interpretations.

## Discussion

The response time results show that participants gave faster correct responses with a prime than without a prime. Even in the baseline condition, the response times with a prime are shorter. Therefore, there is a robust relationship between the primed feature (non-emergent feature) and the target word (emergent feature), regardless of the condition. The proportions of correct responses with a vehicle prime were significantly lower than that with a common prime or with no prime in the metaphor condition but there was no significant difference in the baseline condition. The results of the proportion correct suggest the existence of an interaction between non-emergent and emergent features in metaphor comprehension.

McGlone and Manfredi (2001) found that the primed metaphor-irrelevant properties of vehicles inhibited metaphor comprehension. They concluded that metaphor-irrelevant property primes led their participants to initially consider the inappropriate literal sense of the vehicle, rather than retrieve the metaphoric category that the vehicle exemplified, their results are consistent with the interactive property attribution model (Glucksberg, McGlone, & Manfredi 1997). This latter approach models metaphor comprehension as a process of interpreting the topic as a member of the metaphoric category of the vehicle. This model can account for the inhibition of the emergent interpretation of the metaphor by presenting the vehicle feature prime. For example, the vehicle “gateway” can be interpreted either as a literal arch or as a metaphorical opening. If the vehicle prime “swings” is presented, this may be expected to inhibit the emergent metaphoric feature of “way to reach desired destination.” Our results provide evidence for active competition between literal and figurative interpretations of vehicles.

Particularly when a metaphor is not conventional, the metaphoric category of the vehicle may not simply be retrieved. For example, in the conventional metaphor “the lawyer is a shark,” the metaphoric category of “shark” (*e.g.* vicious, dangerous) may be stable and reliably activated. However, for the metaphor “marriage is a joyride,” because “joyride” is not only fun but also dangerous, the metaphoric category of “joyride” could be retrieved as a product of the features of either “fun” or “dangerous.” That is, the category of “joyride” might be unstable and flexibly represented. In this case, the presentation of a vehicle feature “fun” might inhibit the recognition of “frightening” which is associated with “dangerous.” In this experiment, the average rating of conventionality of the target metaphors is 2.74 (standard deviation is 1.73). Because this value is between 2 (slightly rare) and 3 (neutral), these metaphors should probably be considered non-conventional. As such, the priming of vehicle features could have inhibited recognition of the emergent features, many of which would not have been automatically triggered by the vehicle.

There was no significant difference between the proportions of correct responses with topic primes and without/with other primes in the metaphor condition. These results show

that features derived from the vehicle might affect feature emergence more than that from the topic, bearing in mind that the effect of vehicle features on the interpretation of emergent features is negative. This assumption is consistent with Becker’s (1997) suggestion that altering a metaphor’s vehicle produced greater changes in emergent content than did altering the topic.

The proportion of correct responses with the common feature primes is slightly higher than that without a prime, and the response times with primed common features are significantly shorter than without a prime. Priming with common features apparently made participants recognize emergent features quickly and easily. The common features are also, naturally, present in the vehicle. Thus, having a feature that is shared by both the topic and vehicle affects feature emergence in a very different manner than when the feature is only possessed by the vehicle. Given that the common features also come from the topic, variations in them are more limited than that of the vehicle features. In fact, Nueckles and Janetzko (1997) reported that different people agree on the same common features as interpretations of the metaphor. Furthermore, Becker (1997) showed that the common features are judged to be most important for metaphor interpretation. Therefore, priming with the common feature might give a well-constrained, general direction of the interpretation to the reader. As a result, it is likely that new features for a metaphor emerge through associations with common features. This is consistent with the findings that emergent features are relatively creative reactions to metaphors (Gineste et al., 2000), because the common features might be more active when the metaphor is primed than when the vehicle or topic term only is primed.

From these results, there are apparently interactions, both facilitative and inhibitory, among features in metaphor comprehension. Furthermore, because the feature derived from the vehicle had the greatest impact on emergent feature interpretation, it supports the interactive property attribution model. The property attribution model assumes that the literal level of abstraction is appropriate only for the topic term. The vehicle term is understood at a higher level of abstraction, specifically, a category. If the assumption is correct, then the primed features derived from the vehicle have a stronger affect on metaphor comprehension.

Previous research has shown that emergent features required more time to be recognized as interpretations of the metaphor than non-emergent features (Terai & Goldstone, 2011). Nueckles and Janetzko (1997) suggested the idea that metaphor comprehension proceeds in two stages. Based on these previous results and the findings in this experiment, we may speculate that metaphor comprehension proceeds in two processes. In the first process, a reader tries to interpret a metaphor based on the non-emergent features. If this does not produce a sufficiently high aptness evaluation, then the reader attempts to find emergent features, a longer and cognitively more taxing process. The first process could be ac-

completed according to the interactive property attribution model (Glucksberg et al., 1997). That is to say, at first, non-emergent features are discovered as interpretations through a process in which the vehicle is understood to be referring to a metaphoric category that includes the topic's literal referent as a member. Subsequently, new features become associated with non-emergent features.

This assumption has been incorporated into a simulation model but not validated by an experiment. The model (Terai & Nakagawa 2008, 2010) simulates feature emergence under the two-process assumption which consists of a categorization process followed by a dynamic interaction process. The categorization process is based on the interactive property attribution model and the dynamic interaction process represents interaction among features. The currently reported results support the two-process assumption. However the results did not provide unambiguous evidence for the two-process assumption. Therefore, more examination is required to verify it.

To elucidate the mechanism of feature emergence during metaphor interpretation, we used non-emergent features as primes. The experiment was conducted in order to explain feature emergence during the comprehension of metaphors taking the form of "TOPIC is VEHICLE." For this type of metaphor, the interaction could explain feature emergence. However, when "emergence" is explained in different types of metaphor comprehension (e.g. predicative metaphor comprehension), it might not be sufficient. Indurkha (2006) proposed an idea to explain how new representations emerge through a cognitive agent's interaction with the environment from the viewpoint of "interaction" and "Gestalt perception."

### Acknowledgments

We are deeply grateful to Dr. Angela H. Becker for allowing us to use their valuable data. This research is supported by MEXT's program "Promotion of Environmental Improvement for Independence of Young Researchers," KAKENHI Grant-in-Aid for Young Scientists (B) 23700160. We want to thank to Alex Nay for collecting data.

### References

- Becker, A. H. (1997). Emergent and common features influence metaphor interpretation. *Metaphor and Symbol, 12*, 243–259.
- Black, M. (1979). More about metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 19–45). Cambridge: Cambridge University Press.
- Gentner, D., & Clement, C. (1988). Evidence for relational selectivity in the interpretation of analogy and metaphor. In G. H. Bower (Ed.), *The psychology of learning and motivation, advances in research and theory* (pp. 307–358). New York: Academic Press.
- Gineste, M., Indurkha, B., & Scart, V. (2000). Emergence of features in metaphor comprehension. *Metaphor and Symbol, 15*, 117–135.
- Glucksberg, S., McGlone, M., & Manfredi, D. (1997). Property attribution in metaphor comprehension. *Journal of Memory and Language, 36*, 50–67.
- Indurkha, B. (2006). Emergent representations, interaction theory and the cognitive force of metaphor. *New Ideas in Psychology, 24*, 133–162.
- McGlone, M. S., & Manfredi, D. A. (2001). Topic-vehicle interaction in metaphor comprehension. *Memory and Cognition, 29*, 1209–1219.
- Nueckles, M., & Janetzko, D. (1997). The role of semantic similarity in the comprehension of metaphor. In *Proceedings of the 19th annual meeting of the cognitive science society* (pp. 578–583).
- Terai, A., & Goldstone, R. L. (2011). Processing emergent features in metaphor comprehension. In *Proceedings of the 33rd annual meeting of the cognitive science society* (pp. 2043–2048).
- Terai, A., & Nakagawa, M. (2007). A neural network model of metaphor understanding with dynamic interaction based on a statistical language analysis; targeting a human-like model. *International Journal of neural systems, 17*, 265–274.
- Terai, A., & Nakagawa, M. (2008). A corpus-based computational model of metaphor understanding incorporating dynamic interaction. In V. Kurkova et al. (eds): *Proceedings of icann 2008, part2, lncs 5164* (pp. 443–452). Berlin Heidelberg Springer-Verlag.
- Terai, A., & Nakagawa, M. (2010). A computational model of metaphor understanding based on a probabilistic concept structure-using statistical analysis of Japanese corpus (in Japanese). *Cognitive Studies, 17*, 129–142.
- Utsumi, A. (2000). Hiyu no ninchi / keisan moderu (cognition of metaphors / computational model). *Computer Today, 96*, 34–39.
- Wolff, P., & Gentner, D. (2000). Evidence for role-neutral initial processing of metaphors. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*, 529–541.



# An fMRI Investigation of Feature-Emergence-related Activation within Metaphor Comprehension

**Asuka Terai (asuka@nm.hum.titech.ac.jp)**

Global Edge Institute, Tokyo Institute of Technology,  
2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8550 JAPAN

**Naoko Kuriyama (kuriyama@hum.titech.ac.jp)**

Graduate School of Decision and Science Technology, Tokyo Institute of Technology,  
2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8550 JAPAN

**Masanori Nakagawa (nakagawa@hum.titech.ac.jp)**

Graduate School of Decision and Science Technology, Tokyo Institute of Technology,  
2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8550 JAPAN

**Kimihiko Yamagishi (yamagishi@hum.titech.ac.jp)**

Graduate School of Decision and Science Technology, Tokyo Institute of Technology,  
2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8550 JAPAN

**Takashi Kusumi (kusumi@educ.kyoto-u.ac.jp)**

Graduate School of Education, Kyoto University,  
Yoshidahonmachi, Sakyo-ku, Kyoto, 606-8501 JAPAN

**Koji Jimura (jimura@cns.pi.titech.ac.jp)**

Precision and Intelligence Laboratory, Tokyo Institute of Technology,  
4259 Nagatsuta-cho Midori-ku Yokohama, 226-8503, JAPAN

## Abstract

Metaphor comprehension involves the generation of novel semantic attributes, especially when a metaphor emphasizes a shared but atypical characteristic of the relevant concepts. The present functional magnetic resonance imaging (fMRI) study explores neural activation during the process of attribute generation known as feature emergence. The participants judged whether a presented semantic feature was an appropriate interpretation of a primed metaphor sentence. Two types of features were evaluated: emergent features that are not applicable to the respective concepts and only become salient in a metaphorical context and non-emergent features which are typical characteristics. In contrast to non-emergent features, processing of emergent features mainly involved prefrontal regions of the right hemisphere, including the precentral gyrus. The present results suggest that feature emergence necessitates a shift of semantic attention that drives a novel metaphor interpretation beyond the semantic elaboration implicated within the left prefrontal cortex.

**Keywords:** Metaphor comprehension; Feature Emergence; functional MRI.

## Introduction

It is widely accepted that the comprehension of metaphorical expressions entails greater semantic elaboration than for literal sentences, because the relevant concepts are not directly connected in terms of literal similarity, even though an expression is understandable (Ortony, 1979). Reflecting this salient nature of metaphors, it is also known that metaphor usage enhances recognition memory performance within reading tasks (Reynolds & Schwartz, 1983) and that it is more

persuasive when logical reasoning is less effective (Sopory & Dillard, 2002).

The primary form for metaphors consists of two concepts combined by the “*be*” verb, such as “*A is B*”, where (*A*) and (*B*) are known as the *topic* and the *vehicle*, respectively. This type of metaphor is believed to be interpreted through four types of features depending on whether the feature represents a typical characteristic of the concepts. In the first type of “common feature”, the topic and vehicle share a typical feature. For example, in the metaphor of “*ideas are fireworks*”, “*brilliant*” is a common feature, because both *ideas* and *fireworks* can be regarded as being brilliant. In the second type of “vehicle feature”, the feature is only typical of the vehicle concept. Taking the same metaphor example again, “*momentary*” would be a vehicle feature because it is only relevant to “*fireworks*”. In contrast to a vehicle feature, a topic feature is only a typical characteristic of the topic concept. “*Flash*” would be a topic feature in this example. Although these three feature types minimally involve a typical characteristic of either the topic or the vehicle concepts, the last type of “emergent feature” does not involve a typical characteristic of neither the topic nor the vehicle. However, non-emergent features only appear for interpretations of metaphorical context. The feature “*sudden*” is interpretable within the metaphor example, but it is not a typical characteristic of either the “*ideas*” topic or the “*fireworks*” vehicle. Gineste and colleagues (2000) claim that more than 60% of metaphoric interpretations involve emergent features, which suggests that

emergent features play a major role within metaphor comprehension.

Although previous cognitive neuroscience studies have explored the neural regions involved in metaphor comprehension, they have failed to distinguish between the interpretations of emergent and non-emergent features (e.g., Rapp et al. 2004; Shibata et al. 2007a, 2007b; Stringaris et al. 2007). Some studies indicate that literal and metaphorical sentence comprehension activates multiple cortical regions, including the ventro lateral prefrontal cortex near Brodmann areas (BA) 47 and 44/45, mainly in the left hemisphere (LH), and probably reflects normal semantic processing. In contrast, some other studies have reported fronto-temporal activation within the right hemisphere (RH) during the processing of verbally presented figurative expressions compared to literal expressions (e.g., Marshal et al. 2007; Stringaris et al. 2006). In particular, Marshal and colleagues (2007) found that the processing of word pairs forming novel metaphors, compared to conventional ones, elicited stronger activation within the posterior superior temporal sulcus and the right inferior frontal gyrus mainly in the RH, and Stringaris and colleagues (2006) found right-lateralized prefrontal activation when participants search for a wider range of semantic relationships between a metaphoric sentence and a word. However, others have reported the opposite pattern of lateralization (e.g., Rapp et al. 2004, Stringaris et al. 2007, Shibata et al. 2007b). Rapp and colleagues (2004) report that recognition of metaphors only elicits prominent brain activity in the LH—inferior frontal (BA45/47), inferior temporal (BA20) areas, and the posterior medial/inferior temporal (BA37) gyrus—with no RH activation.

The present study investigates the pattern of cortical involvement during metaphor comprehension. One possibility is, as Marshal and colleagues (2007) suggest, that the processing of novel metaphors involves the right hemisphere. Another possibility is that RH activation reflects some decision criteria processing of presented expressions. Shibata and colleagues (2007a, 2007b) suggest that right frontal regions are more activated when the metaphorical aspect is emphasized for judgments. Thus, they argue that metaphor comprehension requires association search over a wide range which involves RH activation.

The present study hypothesizes that feature emergence activates RH frontal regions during metaphor comprehension. In our previous studies, we have argued that feature emergence is enhanced with longer comprehension times, which suggests that feature emergence may involve interaction spreading out over the semantic network (Terai & Goldstone 2011). We apply the same paradigm within this fMRI study to explore the neural regions activated in feature emergence. We first develop a set of metaphors with features for which feature emergence occurs. In addition to non-emergent features, this set of metaphors was judged by participants while fMRI imaging was conducted.

## Materials and Method

### Participants

All participants ( $N = 10$ ; mean age = 23.3 years; range 21–29 years; 5 male, 5 female) were healthy, right-handed, native Japanese speakers. The current study was approved by the ethical committee of Tokyo Institute of Technology. All participants gave their informed consent before the experiment, and they were compensated for their participation (2000 JPY/hour).

### Materials

Ten metaphorical Japanese sentences of the style “*topic is vehicle*” (e.g., “Ideas are fireworks”) were created based on a behavioral experiment. The behavioral experiment used three feature types (emergent, non-emergent, and filler features). For each metaphor, an emergent (e.g., “sudden”), a non-emergent (e.g., “sparkle”) and a filler feature (e.g., “pray”) were created based on results of a behavioral study. One hundred and fifty five participants evaluated 16 metaphors and six features for each metaphor. The metaphors and participants were divided into four groups. Each participant evaluated the relationship between four metaphors and their features on a 7-point scale. They also evaluated the relationships between eight concepts (the topics and vehicles of the four metaphors) and the features of the metaphors, and rated the characteristics of metaphors (e.g., comprehensibility and novelty). Ten metaphors with higher mean ratings for comprehensibility (above 4) were selected as metaphors that can be easily interpreted. Each metaphor was then assigned with the three types of features, based on comprehensibility ratings of more than 5 (slightly comprehensible) from participants who could interpret a metaphor<sup>1</sup>. Features that fulfilled the following two criteria were classified as emergent features; 1) the mean adequacy rating was greater than three (neutral) and 2) the mean typicality rating was lower than that for the metaphor<sup>2</sup> [mean ratings ( $SD$ ) for metaphor, 5.01 (1.80), topic, 4.60 (2.10), and vehicle, 4.38 (1.99)]. In contrast, features with mean ratings for the metaphor of more than 3 and with higher mean ratings for the vehicle than for the metaphor were classified as non-emergent features [mean ratings ( $SD$ ) with the metaphor, 5.18 (1.81), topic, 4.85 (1.89), and vehicle, 5.88 (1.67)]. If the mean rating for a feature’s relationship with the metaphor was less than 3, the feature was classified as a filler feature [mean ratings ( $SD$ ) with metaphor, 1.80 (1.29), topic, 1.95 (1.43), vehicle, 2.05 (1.66)].

<sup>1</sup>The average proportion of participants who could interpret a metaphor was 33.0%.

<sup>2</sup>We expect that the relationships of an emergent feature with both the vehicle and with the topic are less than that with the metaphor. However, generally, metaphor interpretations can be regarded as a topic characteristic if it is not a typical characteristic. Participants tend to rate vehicles’ relationships with the metaphor highly. Thus, we consider the difference between the relationships with the vehicle and with the metaphor as being important.

## Procedures

fMRI scanning consisted of two sessions. For the first session (concept term condition), at the beginning of each trial, a fixation point “+” was presented for 900 msec on the computer screen. Then, participants were presented with a concept term, which was either a vehicle or a topic, for 3900 msec. They were asked to read the term covertly. Next, the fixation point “+” was presented for 900 msec again. Then, a feature (either an emergent, a non-emergent or a filler feature) was displayed on the screen for 4000 msec. The participants were instructed to respond concerning whether the feature is a typical characteristic of the concept (the vehicle or the topic) term during the presentation period. They held a response box in each hand and responded by pressing the right button (“Yes”) or the left button (“No”). The trials were pseudo-randomly ordered. Fixation trials were also intermixed pseudo-randomly.

In the second, metaphor session, the procedure was as follows: At the beginning of each trial, the participants were presented with a metaphorical sentence on the computer screen for 3900 msec. They were asked to silently read and interpret the metaphor. Then, the metaphor disappeared and a fixation point “+” was presented for 900 msec, followed by the presentation of a feature for 4000 msec. The participants were asked to respond “Yes” or “No” depending on whether the feature represented an adequate interpretation of the metaphor by pressing the right button (“Yes”) or the left button (“No”) during the presentation period. The trials were presented pseudo-randomly, such that the same metaphor never appeared in succession. Fixation trials were intermixed pseudo-randomly.

Participants were given instructions prior to performing the tasks. They were asked to think of nothing during presentation of the fixation point. We conducted a follow-up survey with a questionnaire. The participants were asked to evaluate the characteristics of the metaphors (comprehensibility, novelty, etc.) on 7-point scales.

## Imaging procedures

fMRI scans were conducted with the 3T GE SignaHDxt scanner (General Electric, Milwaukee, WI) at Tokyo Institute of Technology, Japan. High-resolution anatomical images were obtained using an FSPGR T1-weighted sequence [repetition time (TR) = 7.712 msec; echo time (TE) = 2.88 msec, flip angle (FA) = 11 deg, slice thickness = 1 mm; in-plane resolution =  $1 \times 1 \text{ mm}^2$ ]. Functional images were obtained using a GE-EPI sequence [TR = 2.0 sec, TE = 30 msec, FA = 90 deg, slice thickness = 3.0 mm, in-plane resolution =  $3.75 \times 3.75 \text{ mm}^2$ , slice gap = 1.0 mm].

## Imaging analysis procedures

Imaging analysis was performed using SPM8 (<http://fil.ion.ucl.ac.uk/spm/>). All the functional images were temporally aligned across the brain volumes, spatially registered to a MNI template, resampled

into 2-mm isotropic voxels, and then spatially smoothed with an 8-mm FWHM Gaussian kernel.

The present imaging analysis focused on the metaphor condition. For each participant, a voxel-wise GLM analysis was first performed to estimate parameter values for the MR signal magnitudes. The trial period for the metaphor condition was modeled by two independent regressors, one coding metaphor-sentence presentation, and the other coding feature-word presentation that involved a feature judgment, convolved with a canonical HRF (together with time and dispersion derivatives). The metaphor regressor lasted from presentation onset to offset of the sentence, while the judgment regressor lasted for 1000 msec, which was determined based on the mean RTs for each condition for each participant. The estimated parameters during feature judgment were then contrasted between the emergent and non-emergent conditions.

Parameter estimates for each participant were submitted to a group analysis using a voxel-wise random-effect model. Whole-brain exploratory analysis was performed to identify the brain regions showing activity during feature judgment in the emergent condition relative to the non-emergent condition. Due to the small size of the sample in the present study, significance was assessed by a relatively liberal threshold of ten or more continuous voxels above  $p < .001$  ( $Z = 3.3$ ) uncorrected.

## Results

### Behavioral results

The mean reaction times in the concept term condition are shown in Figure 1. A two-way ANOVA with reaction times for the feature conditions and concept types as factors showed a significant main effect for feature types [ $F(2, 18) = 46.35$ ,  $p < .01$ ]. A post-hoc t-test showed that reaction times for the emergent features were longer than reaction times for non-emergent features [ $t(9) = 4.99$ ,  $p < .01$ ]. These results suggest that emergent features are more involved in evaluation-related processes, which is consistent with previous studies (Gineste et al. 2000).

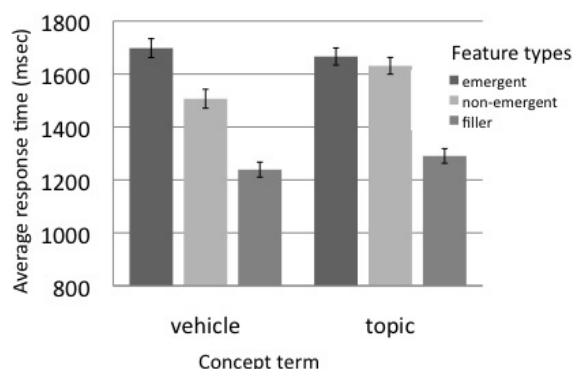


Figure 1: Mean reaction times ( $\pm SEs$ ) for emergent, non-emergent and filler features (in concept term condition)

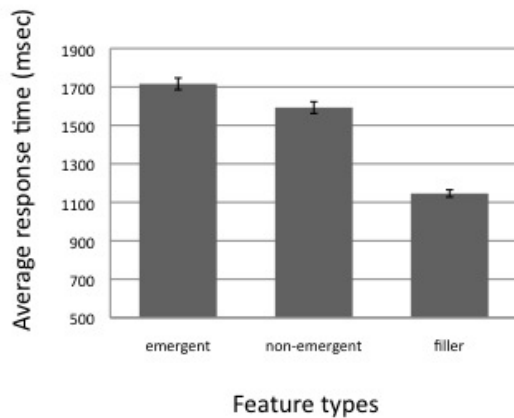


Figure 2: Mean reaction times ( $\pm SE$ s) for emergent, non-emergent and filler features (in the metaphor condition)

Figure 2 shows the mean reaction times for the metaphor condition. A one-way ANOVA for reaction times in the feature condition revealed that significant main effect of feature type [ $F(2,18) = 29.20, p < .01$ ]. The reaction times for the emergent and the non-emergent features were significantly longer than those for the filler features at the 1% level. However, there was no significant difference between reaction times for emergent features and for non-emergent features ( $p=.19$ ). These results indicate that emergent features are activated within metaphor processing as much as non-emergent features.

The average proportions of “Yes” responses are shown in Table 1. All participants responded “No” for almost all the filler features. The proportions of “Yes” responses for emergent features in the metaphor condition (51.0%) were significantly lower than for the non-emergent features (67.0%) [ $F(1,9) = 12.5, p < .01$ ]. In the concept term condition, when the target was the vehicle, the average proportion for non-emergent features (85.0%) was higher than that for emergent features (34.0%) [ $F(1,9) = 157.1, p < .01$ ]. When the target was the topic, the average proportion for emergent features was 45.0%, while for non-emergent features, it was 30.0% [ $F(1,9) = 5.55, p < .05$ ]. Collectively, these findings suggest that emergent features were not easily understood compared to non-emergent features, which is consistent with previous studies (Terai, Goldstone 2011).

These behavioral results suggest that there are distinct differences between the processing of emergent and non-emergent features. In particular, the results indicate that non-emergent features are recognized as a typical characteristic of the vehicle but emergent features are not.

Post-scan evaluations of the metaphor characteristics revealed that the mean rating of the metaphors’ comprehensibility was 4.00 on the 7-point scale (1: incomprehensible - 7: comprehensible), suggesting that the participants understood the presented metaphors. The mean rating for the metaphors’ novelty was 4.29, suggesting that the participants did not re-

Table 1: Mean proportions for emergent and non-emergent (%). For almost all of the filler features, the proportions were almost 0%.

	Concept Term Condition		Metaphor Condition
	Vehicle	Topic	Metaphor
Emergent	34.0	45.0	51.0
Non-emergent	85.0	30.0	67.1

gard the metaphors as being very conventional.

### Imaging results

In order to investigate the difference between the processing of emergent and non-emergent features within metaphor comprehension, we analyzed the differential contrast between the activation of emergent features and for non-emergent features in the metaphor condition (Figure 3 and Table 2). The emergent feature condition showed significant activation in the precentral gyrus (BA6) in the RH and the cingulate gyrus. In contrast, in the concept term condition, no activation was observed in the emergent condition compared to the non-emergent condition.

### Discussion

The present study has explored the pattern of neural activation during feature emergence within metaphor comprehension. Emergent features elicited higher levels of activations in the right precentral gyrus (BA6) and the cingulate gyrus in contrast to non-emergent features in the metaphor condition.

Higher levels of activation have been observed in the right precentral gyrus and the cingulate gyrus during the reading of anomalous metaphors compared to literal sentences and in the right precentral gyrus compared to conventional metaphors (Ahrens et al. 2007). The metaphors were presented as a sentence (e.g. “The theory framework of this theory is very loose”) and the participants were asked to read the sentence and press a button when they finished. In such a situation, participants may search for the relationships between the topic (e.g. “the theory framework”) and the feature (e.g. “loose”) presented in the sentence. In particular, in the anomalous metaphor sentence condition (e.g. “Their capital has rhythm”), they may search for a wider range of associations between the topic (e.g. “their capital”) and the feature (e.g. “have rhythm”). One may speculate that the precentral gyrus is more activated when participants search for a relationship between a metaphor and a novel feature. Accordingly, emergent features may require additional cognitive processing in searching for a wider range of semantic relationships for the metaphor.

Furthermore, these right frontal regions have been associated with stimulus-driven attentional reorientation (Corbetta & Shulman 2002). It is thus feasible to assume that the presentation of emergent features shifts the participants’ seman-

Table 2: Activation contrasts obtained for emergent features versus non-emergent features in the metaphor condition. (Activations listed here were obtained at a voxel level of two-tailed  $p < .001$ , uncorrected, clusters of 10 or more)

Z value	Cluster size	MNI(x)	MNI(y)	MNI(z)	BA	Side	Cerebral Region
4.40	10	52	-2	52	6	R	precentral gyrus
4.10	54	-2	-24	24	24	L/R	cingulate gyrus

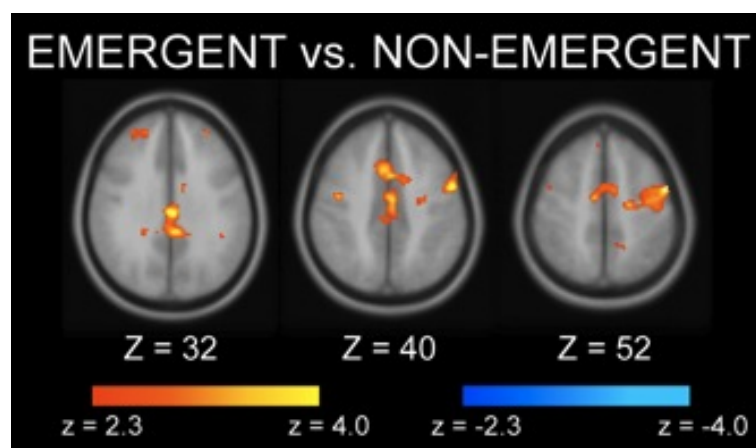


Figure 3: Activation patterns for the differential contrasts between emergent features versus non-emergent features in the metaphor condition (two-tailed  $p < .001$ , uncorrected, clusters of 10 or more)

tic attention to words that are triggered by the presentation of an emergent feature. In other words, these results suggest that the presentation of an emergent feature prompts participants to modify their comprehension processes to incorporate interaction between the topic and the vehicle. That would be consistent with experimental results (Gineste et al. 2000) which indicate that the processing of emergent features is facilitated by interaction between the target and the vehicle concepts.

Previous studies have suggested that searching for a wider range of semantic relationships within metaphor comprehension elicits activation of the right inferior frontal gyrus (Stringaris et al. 2006, Shibata et al. 2007a). However, the results from the present study failed to demonstrate that searching for a wider range of semantic relationships led to activation of this area. In Stringaris et al. (2006), the word was presented with the metaphor and the participants judged the relationship. In contrast, in the present study, the feature was presented after the metaphor had disappeared and we have analyzed data obtained during the interval between feature presentation onset and the participants' responses. Thus, the results from the present experiment only reflect the recognition processing of the feature. The present study's result observed no significant activation differences for the right inferior frontal gyrus.

Moreover, in the present study, emergent features were not spontaneously generated within the interpretative processing of the metaphors. Accordingly, the results only represent a part of the mechanisms of feature emergence within metaphor comprehension. It should also be noted that only ten individ-

uals participated in the present study which only consisted of ten stimulus item for each condition. In order to obtain more robust results, further experiments are needed with more participants and more items.

Nonetheless, the present results indicate that the right frontal area is involved in the processing of emergent features. Accordingly, it is not possible to account for the involvement of the RH merely in terms of metaphor novelty (Marshall et al. 2007, Ahrens et al. 2007), for it may rather reflect feature emergence. If RH involvement can explain differences between the cognitive processing of figurative and literal sentences, feature emergence may be related to metaphor recognition when reading sentences.

## Acknowledgments

We are deeply grateful to Dr. Noburu Hidano, Dr. Hiroyuki Akama and Dr. Atsushi Terao for helping our research. This research is supported by MEXT's program "Promotion of Environmental Improvement for Independence of Young Researchers", KAKENHI Grant-in-Aid for Young Scientists (B) 23700160.

## References

- Ahrens, K., Liu, H., Lee, C., Gong, S., Fang, S., & Hsu, Y. (2007). Functional mri of conventional and anomalous metaphors in madarin chinese. *Brain and Language*, 100, 163-171.
- Becker, A. H. (1997). Emergent and common features influ-

- ence metaphor interpretation. *Metaphor and Symbol*, 12, 243–259.
- Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Neuroscience*, 3, 201–215.
- Gineste, M., Indurkha, B., & Scart, V. (2000). Emergence of features in metaphor comprehension. *Metaphor and Symbol*, 15, 117–135.
- Marshall, N., Faust, M., Hendler, T., & Jung-Beeman, M. (2007). An fmri investigation of the neural correlates underlying the processing of novel metaphoric expressions. *Brain and Language*, 100, 115–126.
- Ortony, A. (1979). Beyond literal similarity. *Psychological Review*, 86, 161–180.
- Rapp, A. M., Leube, D. T., Erb, M., Grodd, W., & Kircher, T. T. J. (2004). Neural correlates of metaphor processing. *Cognitive Brain Research*, 20, 395–402.
- Reynolds, R. E., & Schwartz, R. M. (1983). Relation of metaphoric processing to comprehension and memory. *Journal of Educational Psychology*, 75, 450–459.
- Shibata, M., Abe, J., Terao, A., & Miyamoto, T. (2007a). Neural bases of metaphor comprehension an fmri study (in Japanese). *Cognitive Studies*, 14, 339–354.
- Shibata, M., Abe, J., Terao, A., & Miyamoto, T. (2007b). Neural mechanisms involved in the comprehension of metaphoric and literal sentences: An fmri study. *Brain Research*, 1166, 92–102.
- Sopory, P., & Dillard, J. P. (2002). The persuasive effects of metaphor: a meta-analysis. *Human Communication Research*, 28, 382–419.
- Stringaris, A. K., Medford, N. C., Giampietro, V., Brammer, M. J., & David, A. S. (2007). Deriving meaning: distinct neural mechanisms for metaphoric, literal, and non-meaningful sentences. *Brain and Language*, 100, 150–162.
- Stringaris, A. K., Medford, N. C., Giora, R., Giampietro, V., Brammer, M. J., & David, A. S. (2006). How metaphors influence semantic relatedness judgments: The role of the right frontal cortex. *Neuro Image*, 33, 784–793.
- Terao, A., & Goldstone, R. (2011). Processing emergent features in metaphor comprehension. In *Proceedings of the 33rd annual meeting of the cognitive science society* (pp. 2043–2048).

# Explanation Reconstruction through Reinterpretation of Key Facts

Hitoshi Terai (terai@is.nagoya-u.ac.jp)

Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)

Shota Matsubayashi (s.matubayashi@cog.human.nagoya-u.ac.jp)

Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku,  
Nagoya, Aichi, 464-8601, Japan

## Abstract

Reconstructing explanations is crucial for the progress of science. We focused on the transition of interest in a key fact that contradicts the preceding explanation and has a central role in its reconstruction. We used a short story as an experimental material in which the participants first constructed a naïve explanation and reconstructed it. First, when the naïve explanation was rejected, a new explanation was required, after interest in the key fact was inhibited. Second, hypothesized premises not inconsistent with the naïve explanation were sought to protect the naïve explanation. Third, interest in the key fact was recovered through the process of the explanation reconstruction. Last, we facilitated the explanation reconstruction by having the participants focus on the key fact.

**Keywords:** eye movement analysis; scientific explanation; key fact; insight; naïve concept.

## Introduction

Scientific activities aim to understand the world by two ways: descriptive and explanative (Simon, 2000). Descriptive understanding grasps the nature and characteristics of phenomena by observations and experiments; explanative understanding grasp the mechanisms behind the phenomena and the reasons why such phenomena appear.

Through the history of science, descriptive understanding is usually established first and then explanative understanding is investigated. For example, Kepler's law described the orbit of the planets, and then Newton's law explained why they moved in such orbits. Science has developed while pursuing such explanative understanding about phenomena. The construction of explanations is crucial for science.

As historical facts, we can confirm many cases where the explanation for a certain phenomenon was completely changed because the structures of the explanation and the concepts of objects were essentially shifted. Such cases are generally observed in the history of science: e.g., the shift from the caloric theory to the oxygen theory and the transition from Newton's traditional theory to Einstein's relative theory.

As an example, consider the change of the caloric theory to the oxygen theory. Initially, in the caloric theory, burning was explained as the release of caloric. After an inconsistency was observed about the caloric theory, the weight increase after burning, a new explanation was required. In the current study, we call such an instance that contradicts a preceding explanation and must be interpreted by a new explanation a "key fact." The change of the explanation from the caloric theory, i.e., burning released caloric, to the oxygen theory, i.e., burning is connected with oxygen, was es-

tablished by reinterpreting the key fact, i.e., the increase of weight by burning.

Note that there are two ways of understanding a key fact. One is by local modification and slight expansion of a previous explanation. The other is understanding by an essential change of a previous explanation. An interpretation about key facts completely changes between the two types of understanding: a completely different interpretation about a key fact is given in each of the two theories. In the oxygen theory, the key fact is explained by the connection with oxygen, but in the caloric theory, it is explained by the release of phlogiston that has negative weight.

Difficulties exist in such essential reconstruction of explanations. Interesting processes are often observed that prevent such reconstruction. One is the stubborn refusal to abandon explanations. People prefer to protect an established explanation by modifying and adding new reservations than shifting to a new one. In such a case, to protect the old explanation, people may focus on other irrelevant facts and arbitrarily proposed premises that are not inconsistent with the previous explanation. They sometimes add secondary explanations as protection. For example, in the caloric theory, a premise, phlogiston might have negative weight, was hypothesized and investigated.

Now we summarize the problems we address in this paper. We investigate a situation in which fact F that cannot be interpreted by explanation A is observed, and new explanation B is required. For the transition from explanation A to B, a mental leap is needed, meaning that fact F must be reinterpreted. We call fact F a key fact and investigate how it is processed through a reconstruction of the explanation.

The research questions and hypotheses are drawn in the following. We hypothesize that reinterpretation of key facts is crucial for the transition to a new explanation. However, people tend to pursue unrelated facts or arbitrarily hypothesized premises to protect old explanations, and such reinterpretation of key facts may be postponed. As a result, the interest in key facts is inhibited, and the reconstruction of explanation is impeded.

We propose two hypotheses:

**Hypothesis 1** When a previous explanation is rejected and a new explanation is required, interest in a key fact may be temporarily inhibited.

**Hypothesis 2** With the inhibition of interest in a key fact, other facts and hypothesized premises that are not inconsistent with a previous explanation are searched for, and



Introduction	Taro was driving to Las Vegas by rental car. His car broke down in a small town. He decided to get a haircut while the car was repaired. There are only two barbershops in the town: Alf's shop and Bally's shop. He is considering which to select.
Filler (Place)	Alf's shop is on the ground floor of a building located in the east area of town. In the building, there is a stationary shop. Bally's shop is along a street running in the west area of town. There is a supermarket near it.
Key Fact	Alf's hair is unkept, and the nape of his neck is messy. Bally's hair is beautifully cut, and the nape of his neck is neat.
Filler (Barbershop)	The windows of Alf's shop are light blue, and natural scene are pictured on the cover pages of the books in the shop. The windows of Bally's shop are light green, and various letters are written on the cover pages of the books in the shop.
Filler (Time)	Alf's shop is open until late. He often eats dinner at his favorite restaurant near the shop. Bally's shop is open early. He usually walks around the shop in the morning.

Figure 1: Barber task (used in Experiment 1).

the old explanation is protected.

We are also interested in the transition process from a former to a new explanation. We confirm that the reinterpretation of a key fact has a central role in the transition of the explanation. Two additional hypotheses are drawn.

**Hypothesis 3** The recovery of interest in a key fact is observed through the process of the explanation reconstruction.

**Hypothesis 4** We may promote explanation reconstruction by having participants focus on a key fact.

### Concern for a key fact

We used a short story as an experimental material because people understand a text by unifying meanings while adding implicit information and inferences about omitted and un-presented sentences (Rumelhart & Ortony, 1977; Seifert & Robertson, 1985).

We used a text that was modified from a barber task (Gardner, 1978). Figure 1 shows the material. In this material, (1) a naïve explanation is initially constructed by a key fact; (2) a new explanation is required where a shift of perspective is needed; (3) a rational explanation is constructed by reinterpreting the key fact.

In the story, in a town with only two barbershops, a character is looking for a barber and must to select either barber A whose staff has unkept hair or barber B whose staff's hair is beautifully cut. Initially, participants may select barber B where the following naïve explanation is given: "a barber with beautiful hair is very skilled." However, a new explanation is required after being informed that the character selected barber A. The fact, "barber A's hair is messy and barber B's hair is neat," contradicts the initial explanation. Therefore, in the story, the key fact that must be reinterpreted is:

"barber A's staff has unkept hair and barber B's staff has neat hair." The reconstructed explanation from Gardner (1978) is: "each does the other's hair because there are only two shops in town; therefore, barber A's staff who did barber B's staff's hair is more skilled." In the reconstructed explanation, the key fact becomes evidence for selecting barber A but in the naïve explanation, it is evidence for selecting barber B. The meaning of the key fact has completely shifted with the transition from the initial to the reconstructed explanation.

In the text, other unrelated facts than the key fact are described; therefore, other secondary additional explanations may be possible to protect the initial naïve explanation. However, if participants construct such an explanation based on other facts than the key facts, the contradiction remains unsolved: the character selects a messy barber. The shift to the reconstructed explanation by reinterpreting the key fact is required for consistently understanding the story structure.

## Experiment 1

We confirmed the validity of the barber's story as an experimental task to examine our hypotheses. Experiment 1 confirmed whether most participants initially constructed the naïve explanation. Additionally, to confirm Hypothesis 2 preliminary, participants reconstructed their explanations after the naïve explanation was rejected.

The definitions of the naïve and reconstructed explanations are described below.

**Naïve explanation** A barber who has beautiful hair is very skilled.

**Reconstructed explanation** Each does the other's hair because there are only two shops in the town; therefore, barber A's staff who did barber B's staff's neat hair is more skilled.

## Subjects

Fifty-three undergraduate students participated in Experiment 1.

## Task

Figure 1 shows the barber task used in Experiment 1.

## Procedure

Experiment 1 was constructed of two phases: the initial explanation phase and the reconstruction phase. In the initial explanation phase, the participants read the story while thinking about which barbershop to select and their task was to construct an explanation for their decision. The initial explanation phase was followed by the reconstruction phase in which the naïve explanation was rejected, and they were required to reconstruct their explanation of the story.

## Results

Forty four of the 53 participants initially constructed the naïve explanation. A binomial test revealed that they primarily constructed naïve explanations (two-sided:  $p < .01$ ). Moreover, the reconstructed explanations by the 44 participants were classified into four types (Table 1). Three other explanations than the reconstructed explanation were based on such facts about place as “shop B is located near the repair shop,” and about time such as “shop B is open until late,” and hypothetical information not included in the story. A chi-square test revealed a significant difference in the numbers of these explanations ( $\chi^2(3) = 10.8, p < .05$ ), and a multiple comparison using Ryan’s method showed that the explanations based on place and time were constructed significantly more than the reconstructed explanation ( $p < .01, p < .01$ ).

Table 1: Produced second explanations.

Explanation	#	Example of description
Target (reconstructed explanation)	3	Each does the other's hair because there are only two shops in the town; therefore, Bally's staff who did barber Alf's beautiful staff's hair is more skilled.
Place	15	Alf's shop was near the car repair shop.
Time	18	Taro wanted to get a haircut late in the evening.
Misc	11	-

Some descriptions were classified into multiple categories because they included multiple facts.

These results confirmed the validity of the barber task as an experimental task for our study. Additionally, we confirmed that the participants tended to add secondary explanations based on the facts about place and time to protect the naïve explanation, preliminarily supporting Hypothesis 2.

## Experiment 2

We confirmed both the inhibition of interest in the key fact (Hypothesis 1) and resumption of interest (Hypothesis 3) us-

ing eye movement analysis to capture the transition of interest.

## Subjects

Twenty-one undergraduate students participated in the experiment.

## Task

The story was displayed on a computer screen. The filler (barbershop) part of the text (Figure 1) was removed due to limitations of the display size.

## Procedure

The experiment was conducted individually, and participant eye movements were recorded using a Tobii T60 eye tracker.

As in Experiment 1, in the initial explanation phase, the participants were required to explain the story. Their fixation ratios of the key fact during the initial explanation phase were used as the baseline for analysis of the subsequent reconstruction phase. The fixation ratio of each fact was normalized by the number of letters that were included in each part.

After the initial explanation phase, the participants were reconstructed their explanations as answers to a quiz. When the participants found an idea, they reported it. They did the experiment at their own pace. When they gave another explanation than the reconstructed explanation as the target, they were told that it was not correct and were told to reconsider. This phase was continued for 30 minutes; when each participant constructed the reconstructed explanation, it was terminated.

## Results

**Inhibition of fixation on key fact** To examine the inhibition of interest in the key fact after the naïve explanation was rejected, we analyzed the fixation ratio of the key fact part in the initial stage of the reconstruction phase. First, we examined whether the fixation ratio of the key fact was less than the baseline obtained in the initial explanation phase and the other facts (place and time).

Figure 2 shows the fixation ratio of each of the facts (place, key, and time) during the first minute and the subsequent minute in the reconstruction phase.

In the first minute (0–60 sec), a t-test indicated no significant differences between the fixation ratio of the key fact and the baseline ( $t(20) = 1.62, n.s.$ ). A one-way ANOVA showed a significant main effect of the three facts (place, key, and time) ( $F(2, 40) = 8.97, p < .001$ ), and a multiple comparison using Ryan’s method showed that the fixation ratio of the fact about place was significantly higher than those of the key fact and the fact about time ( $p < .05, p < .05$ ).

Next, we conducted the same analysis on the subsequent minute (60–120 sec). A t-test indicated a significant difference between the fixation ratio of the key fact and the baseline ( $t(20) = 2.86, p < .01$ ). A one-way ANOVA showed a significant main effect of the three facts (place, key, and time) ( $F(2, 40) = 4.36, p < .05$ ), and a multiple comparison

using Ryan's method showed that the fixation ratio of the key fact was significantly lower than the place and time facts ( $p < .05$ ,  $p < .05$ ).

In the subsequent minute, we confirmed that the fixation ratio of the key fact was significantly lower than the baseline and the place and time facts. These results support that interest in the key fact was inhibited when the naïve explanation was rejected.

On the other hand, in the first minute, our prediction was not observed; the fixation ratio of the fact about place was substantially higher. This result might be affected by the order of the three facts, which were arranged as place, key, and time (see Figure 1). The participants probably read the story in this order, reflecting the result in the first minute.

The overall results of Experiment 2 supported Hypothesis 1.

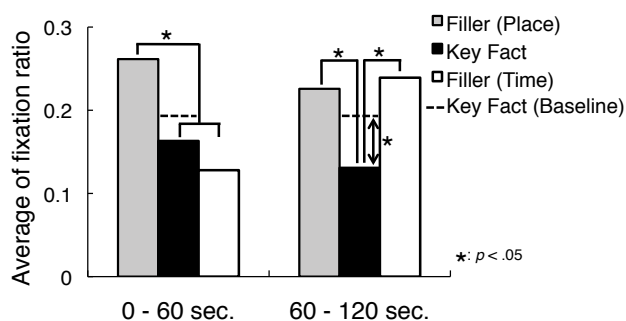


Figure 2: Transition of fixation ratio of each fact.

**Resumption of fixation on key fact** Next, we analyzed the process of recovering interest in the key fact to build reconstructed explanation (Hypothesis 3). We focused on the 11 of 21 participants who successfully constructed the reconstructed explanation and analyzed the transition process of their fixation ratio of the key fact part.

Figure 3 shows the transition with the progress of three phases: the first 60 seconds, the last 60 seconds before reaching the reconstructed explanation, and the residual between them.

A one-way ANOVA showed a significant main effect of the three phases ( $F(2, 20) = 7.62$ ,  $p < .005$ ), and a multiple comparison using Ryan's method showed that the fixation ratio of the key fact in the last phase was significantly higher than those in the first and middle phases ( $p < .05$ ,  $p < .05$ ). The interest in the key fact gradually improved even though the ratio in the middle phase was not greater than in the first phase, partially supporting Hypothesis 3.

### Experiment 3

In Experiment 2, we observed the inhibition of interest in the key fact after the naïve explanation was rejected (Hypothesis 1).

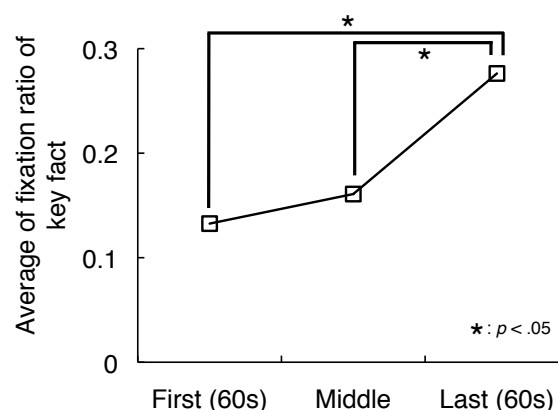


Figure 3: Transition of fixation ratio of key fact in successful group.

This means that the participants searched for unimportant facts that did not contradict the naïve explanation.

In Experiment 3, we examined whether forced recovery of interest in the key fact could facilitate the reconstruction of their explanations (Hypothesis 4) by externally controlling interest in the key fact.

### Subjects

Forty-one undergraduate students participated in this experiment.

### Procedure

This experiment was conducted in small groups on personal computers for the stimulus presentation and data acquisition.

The participants were divided into two experimental conditions: highlighted (21 participants) and non-highlighted (20 participants). In the highlighted condition, the key fact was colored red to facilitate interest in it; the participants were instructed that the highlighted sentences were crucial for finding the right explanation. There was no such highlight in the non-highlighted condition.

In the initial explanation phase, the participants constructed explanations about the story as in Experiments 1 and 2. Then, in the reconstruction phase, they were also required to reconstruct their explanations and report them by computer keyboard. After reporting their explanations, they received a message: "since there is another reasonable explanation acceptable to all, please reconsider." They were told that there was an evaluator in another room connected by the Internet, even though no such evaluator existed, and the same message was always returned. The maximum time of the reconstruction phase was 15 minutes, and the data were analyzed until they reached reconstructed explanations.

### Results

The explanations that were constructed in the initial explanation phase were mostly naïve explanations (19 of 21 in the

highlighted condition and 18 of 20 in the non-highlighted condition).

Next, we analyzed the facts to which the participants referred until they reached the reconstructed explanation in the reconstruction phase. The referenced facts were identified by their description about the explanations. The ratio of each of the referenced facts in the generated explanations is shown in Figure 4.

A two-way ANOVA was conducted with the experimental conditions (highlighted and non-highlighted) as a between-participant factor and the facts (place, key, and time) as a within-participant factor. There was neither a significant main effect of the experimental conditions nor of the facts ( $F(1, 33) = .16, n.s.$ ;  $F(2, 66) = 1.15, n.s.$ ), but there was significant interaction between the conditions and the facts ( $F(2, 66) = 8.55, p < .001$ ). A multiple comparison using Ryan's method showed differences between the highlighted and non-highlighted conditions of the key and place facts ( $p < .05, p < .05$ ). There were also significant differences between the key and place facts in the highlighted condition ( $p < .05$ ), and the key fact and the place and time facts in the non-highlighted condition ( $p < .05, p < .05$ ).

These results suggest that the participants in the highlighted condition attempted to reinterpret the key fact. On the other hand, in the non-highlighted condition, there was little mention of the key fact. When there was no facilitation of interest in the key fact, the participants tended to modify their explanation based on other facts, supporting Hypothesis 2 that was preliminary supported in Experiment 1.

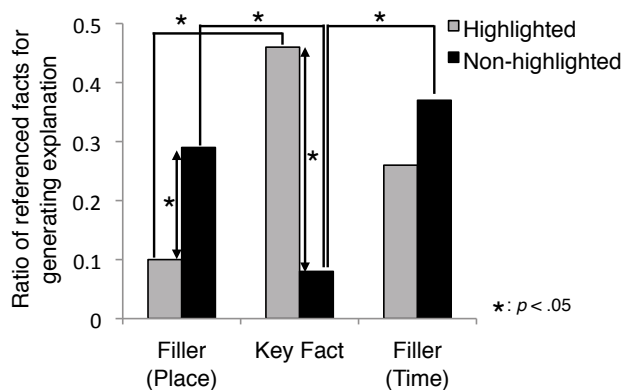


Figure 4: Ratio of referenced facts for generating explanation.

Finally, the ratios of the accumulated number of the participants who reached the reconstructed explanation in the reconstruction phase are shown in Figure 5. A chi-square test showed significant differences between the two experimental conditions at 10 and 15 minutes ( $\chi^2(1) = 6.26, p < .05$ ;  $\chi^2(1) = 7.79, p < .01$ ).

These results indicate that enhancing interest in the key fact facilitated the reconstruction of the naïve explanation, and

shifting to the reconstructed explanation supporting Hypothesis 4.

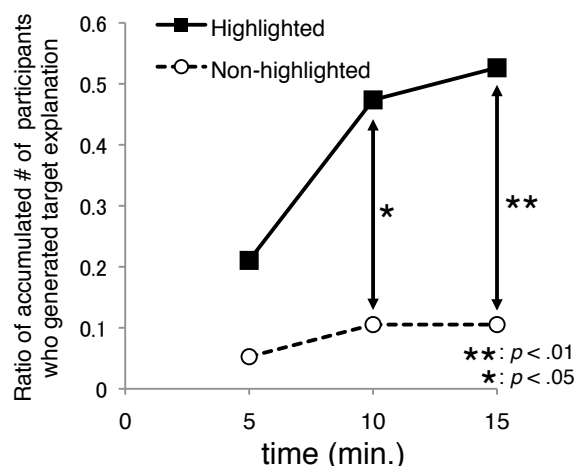


Figure 5: Ratio of accumulated number of participants who generated consistent explanations.

## Discussions and Conclusions

In our study, we focused on reinterpretation of a key fact for explanation reconstructions and examined the inhibition of interest in it and the improvement of interest in it by eye-movement analysis. The following is a summary of our experiment.

- When the naïve explanation was rejected by contradictions with the key fact and a new explanation was required, interest in the key fact was inhibited (Hypothesis 1 was supported).
- In such a situation, other facts and hypothesized premises not inconsistent with the naïve explanation were searched for and focused on, and the naïve explanation was protected (Hypothesis 2 was supported).
- Interest in the key fact was recovered through the process of the explanation reconstruction, especially before reaching the solution (Hypothesis 3 was partially supported).
- We facilitated the explanation reconstruction by having the participants focus on the key fact (Hypothesis 4 was supported).

In studies of hypothesis generation and testing, some human biases have been confirmed. For example, confirmation bias gathers positive instances to confirm hypotheses (Wason, 1960), and established hypotheses are maintained against anomalous data (Chinn & Brewer, 1998; Mason, 2001).

In the educational psychology domain, it has also been confirmed that naïve concepts that are not supported by related evidences are held strongly (McCloskey, Washburn, & Felch, 1983; Watts & Zylberstan, 1981). Watts and Zylberstan (1981) studied junior high-school students' naïve concepts about the inertia law and concluded that even if they received accurate knowledge about the law from lectures, they repeated the naïve explanation for events related to it.

In our experiments, we set up an experimental situation in which the participants reconstructed the naïve explanation. Theories in scientific activities were established based on the accumulated results of experiments through history in which the fixation to an explanation may be much stronger. Note that similar phenomena were observed using a short story in the laboratory setting of our current study.

The situation requiring a shift in explanations in our experiments seems similar to the settings dealt with in studies of insight problem solving. Here, mental constraints arising from the perceptual features of a problem and past experiences create an impasse and prevent problem solvers from finding the new relations required to solve the problem. These mental constraints are gradually relaxed unconsciously in some cases even if problem solvers meet an impasse, where they often ignore key evidence that leads to a solution. With activities that do not follow these mental constraints and acceptance of such crucial instances for solutions, problem solvers gradually reach a solution (Knoblich, Ohlsson, Haider, & Rhenius, 1999; Knoblich, Ohlsson, & Raney, 2001; Ohlsson, 1992; Terai & Miwa, 2003). In our experiment, we also observed such a recovery process that of focused on key facts.

It is difficult to manage interest in facts without the influence of hypotheses and concepts that were previously constructed (Kaplan & Simon, 1990; Luchins & Luchins, 1950; Wason, 1960). For example, Bilalic, McLeod, and Gobet (2008) used a rule discovery task that required participants to search for a better solution than the one they had already found. Their study indicated that the participants were unconsciously prevented from searching for facts that were unrelated to the existing solution, even if they were instructed to seek alternatives. This result also indicates that human behavior is largely constrained by constructed hypotheses and concepts.

In Experiment 3 of our study, even though we controlled the participant interest in the key fact that contradicted the naïve explanation by highlighting it, there was no significant difference between the highlighted and non-highlighted conditions during the first five minutes. Moreover, only half of the participants constructed the reconstructed explanation, even if such external stimuli were given. This suggests that the inhibition of interest in the key fact that contradicted the naïve explanation remained even after deep consideration for 15 minutes. We must study the interaction between conscious and unconscious activities by combining verbal protocols and eye movement analysis to understand the process of shifting explanations in more detail.

## References

- Bilalic, M., McLeod, P., & Gobet, F. (2008). Why good thoughts block better ones: The mechanism of the pernicious Einstellung (set) effect. *Cognition*, 108, 652–661.
- Chinn, C. A., & Brewer, W. E. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *JOURNAL OF RESEARCH IN SCIENCE TEACHING*, 35, 623–654.
- Gardner, M. (1978). *Aha! insight*. New York: W. H. Freeman & Co.
- Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive Psychology*, 22, 374–419.
- Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(6), 1534–1555.
- Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition*, 29(7), 1000–1009.
- Luchins, A. S., & Luchins, E. H. (1950). New experimental attempts at preventing mechanization in problem solving. *Journal of General Psychology*, 42, 279–294.
- Mason, L. (2001). Responses to anomalous data on controversial topics and theory change. *Learning and Instruction*, 11, 453–483.
- McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9, 436–648.
- Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In M. T. Keane & K. J. Gilhooley (Eds.), *Advances in the psychology of thinking* (pp. 1–44). Upper Saddle River, NJ: Prentice-Hall.
- Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99–135). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Seifert, C. M., & Robertson, S. P. (1985). Types of inferences generated during reading. *Journal of Memory and Language*, 24, 405–422.
- Simon, H. A. (2000). Discovering explanations. In R. A. Keil, F. C. & Wilson (Ed.), *Explanation and cognition* (pp. 21–59). Cambridge, MA: MIT Press.
- Terai, H., & Miwa, K. (2003). Insight problem solving from the viewpoint of constraint relaxation using eye movement analysis. In *proceedings of the 4th international conference of cognitive science* (pp. 671–676).
- Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12, 129–140.
- Watts, D. M., & Zylberstan, A. (1981). A survey of some children's ideas about force. *Physics Education*, 16, 360–365.

# Automatic Detection of Metonymies using Associative Relations between Words

**Takehiro Teraoka (teraoka@sfc.keio.ac.jp)**

Graduate School of Media and Governance, Keio University,  
5322 Endoh, Fujisawa, Kanagawa, Japan

**Ryuichiro Higashinaka (higashinaka.ryuichiro@lab.ntt.co.jp)**

NTT Cyber Space Laboratories, NTT Corporation,  
1-1 Hikarinooka Yokosuka, Kanagawa 239-0847 Japan

**Jun Okamoto (juno@kaetsu.ac.jp)**

Department of Business Innovation, Kaetsu University,  
2-8-4 Hanakoganeiminamicho, Kodaira, Tokyo 187-0003, Japan

**Shun Ishizaki (ishizaki@sfc.keio.ac.jp)**

Graduate School of Media and Governance, Keio University,  
5322 Endoh, Fujisawa, Kanagawa, Japan

## Abstract

It is crucial for computers to detect metonymic expressions because sentences including them may have different meanings from literal ones. In previous studies, detecting metonymies has been done mainly by rule-based and statistical approaches. The problem of current metonymy detection is that using syntactic and semantic information may be not enough to detect metonymic expressions. In this study, we propose an approach for detecting them with associative information between words. We evaluated our method by comparing it with a baseline that uses syntactic and semantic information. As a result, our method showed significantly better accuracy (0.84) of judging words as metonymic or literal expressions than that of the baseline.

**Keywords:** Metonymy; Association Experiment; Associative Concept Dictionaries; Verbs; Nouns

## Introduction

Metonymy is a figure of speech, where one item's name represents another which usually has a close relation with the first one. The metonymic relation, as shown in Table 1 (Lakoff & Johnson, 1980; Taniguchi, 2003; Yamanashi, 1988), has different patterns which are classified predominately into two types: spatial contiguity and temporal contiguity (Taniguchi, 2003). Below is a Japanese example for 'Container for Content':

*kare-ga isshoubin-wo nomihoshita*  
(He drank up a 1.8-liter bottle.)

The Japanese sentence above means literally that he drank up the bottle. Of course, it does not mean that he drank or ate the bottle itself, but its content, usually Japanese sake. Japanese sake is generally in a large bottle made from glass, and called *bin* in Japanese. It has a capacity of 1.8 liters, *isshou*. Therefore, the above example sentence where *isshoubin* is a metonymic expression means that he drank up Japanese sake in a 1.8-liter bottle. Since a sentence including metonymy is grammatically correct on a literal level, it is difficult for computers to grasp its true meaning as humans do.

Table 1: Metonymic expressions with spatial contiguity and temporal contiguity.

Metonymic patterns	Examples of sentences (metonymic reading)
-spatial contiguity-	
Container for Content	<i>kare-ha glass-wo nonda</i> 'He drank the glass (liquid).'
Producer for Product	<i>kare-ha Mahler-wo kiita</i> 'He listened to Mahler (symphony).'
Controller for Controlled	<i>Nixon-ga Hanoi-wo bakugekishita</i> 'Nixon (government) bombed Hanoi.'
Object Used for User	<i>gakuseifuku-ga aruiteiru</i> 'The school uniform (student) is walking.'
Material for Product	<i>kare-ha caffeine-wo nonda</i> 'He drank caffeine (soft drink).'
Others	<i>riiron-ga sore-wo jishshoushita</i> 'The theory (proposer) claimed that.'
Metonymic patterns	Examples of sentences (metonymic reading)
-temporal contiguity-	
Result for Cause	<i>kanojo-ga sekimensuru</i> 'She is blushing.' (She is ashamed)
Cause for Result	<i>kare-ga sakazuki-wo katamukeru</i> 'He is tipping the sake cup.' (He is drinking the Japanese sake)

In English metonymy detection, most previous studies have taken mainly rule-based and statistical approaches. The rule-based approach uses semantic networks and hand-crafted rules to detect metonymies (Bouaud, Bachimont, & Zweigenbaum, 1996; Fass, 1991; Iverson & Helmreich, 1992). The representative work of statistical approach used corpus-based metonymy resolution on location names (Markert & Nissim, 2003). Moreover, by using syntactic, semantic, encyclopedic, or collocation information as machine learning features, some conventional studies for detecting metonymic expressions were suggested (Markert & Nissim, 2007; Nastase & Stube, 2009). Their methods are effective, but they only dealt with metonymies on country names and companies. When considering the variety of metonymic patterns in Table 1, it is desirable to be able to detect various



Table 2: Semantic relations used in experiments.

Semantic relation	Content
Agent	Subject of an action
Object	Object of an action
Source	Source of an action
Goal	Goal or end of an action
Duration	Time or term of an action
Location	Location or space during an action
Tool	Tool or material of an action
Aspect	Aspect, degree or frequency of an action
Reason	Reason or cause of an action
Purpose	Purpose of an action

metonymies. In Japanese, although with small data sets, the manually constructed case frame dictionary and Goi-Taikai—A Japanese Lexicon (Ikehara et al., 1999), which consist of syntactic and semantic information, have been used for detecting various metonymies (Murata, Yamamoto, Kurohashi, Isahara, & Nagao, 2000; Suga & Ishizaki, 2006).

The problem of current metonymy detection is that using syntactic and semantic information may not be enough to detect metonymic expressions because in our daily conversations and readings we understand metonymic expressions in sentences by using associative relations between words unconsciously. As Yamanashi described (Yamanashi, 1987, 1988), metonymic relations relate to psychological association; we consider that computers also need associative information to improve the accuracy of metonymy detection.

By using our associative concept dictionaries for verbs and nouns (hereinafter referred to as Verb-ACD and Noun-ACD, respectively) (Okamoto & Ishizaki, 2001; Teraoka, Okamoto, & Ishizaki, 2010), our previous study proposed an approach to metonymy detection with associative information and showed its effectiveness (Teraoka, Okamoto, & Ishizaki, 2011). In this study, we focus on detecting only metonymic expressions of the spatial contiguity type as our first step, and enhance our approach by using decision tree learning.

### ACD Construction

In this section, we describe the Verb-ACD and the Noun-ACD that we use to extract associative information for detecting metonymic expressions.

#### Verb-ACD

The Verb-ACD (Teraoka et al., 2010) consists of the following three elements: stimulus words, associated words from the stimulus words with semantic relations, and word distances among them. The stimulus words are basic verbs with semantic relations that corresponded to deep cases. We quantify word distance between the stimulus word and the associated one.

**Association Experiments** To collect associative information on verbs, we conducted large-scale association experiments on the web. The stimulus words were verbs from Japanese elementary school textbooks, and we prioritized 200 of them that were entry words in a basic Japanese dictionary

(Morita, 1989). We prepared ten semantic relations shown in Table 2: Agent, Object, Source, Goal, Duration, Location, Tool, Aspect, Reason, and Purpose. The experiment participants were requested to give associated words of the stimulus words with these semantic relations.

**Quantification of Word Distances** We used the linear programming method to calculate distances between stimulus words and associated ones. As shown in Eq. (1), the distance  $D(x, y)$  between the stimulus word  $x$  and the associated word  $y$  is expressed with the following formulae:

$$D(x, y) = \frac{7}{10}IF(x, y) + \frac{1}{3}S(x, y) \quad (1)$$

$$\text{where } IF(x, y) = \frac{N}{n(x, y) + \delta}, \quad (2)$$

$$\delta = \frac{N}{10} - 1 (N \geq 10), \quad (3)$$

$$S(x, y) = \frac{1}{n(x, y)} \sum_{i=1}^{n(x, y)} s_i(x, y). \quad (4)$$

The distance consists of the inverse of frequency of an associated word  $IF(x, y)$  in Eq. (2) and the average of the associated word order  $S(x, y)$  in Eq. (4). Each coefficient was obtained by using the Simplex Method. Let  $N$  denote the number of participants in the experiments, and  $n(x, y)$  denote the number of participants who responded with the associated word  $y$  to the stimulus word  $x$ . Let  $\delta$  in Eq. (3) denote a factor which limits  $IF(x, y)$  to a certain numerical level when  $N$  increases. Let  $s(x, y)$  denote the associated word's order of each participant.

Each semantic relation of two words is expressed by each distance where the smaller the distance is, the closer two words are. For example, when a stimulus verb is the Japanese word, *aruku* 'walk' and the semantic relation is Source, one of the associated words is *ie* 'home' of which the distance is 1.38. Meanwhile, the distance between walk and *kaisha* 'office' is 9.92. The relation of these distances thus expresses a degree of association from the verb with the semantic relation.

Currently, there are 345 stimulus verbs in the Verb-ACD and the number of all participants is approximately 1,300. The participants were undergraduates and graduate students of Keio University. Each stimulus verb was presented to 40 participants. There were approximately 135,000 associated words. When all overlapping words were eliminated, there were 30,000 associated words.

#### Noun-ACD

The Noun-ACD consists also of stimulus words, i.e., nouns, associated words with semantic relations, and word distances among these words (Okamoto & Ishizaki, 2001). Table 3 shows the semantic relations and examples when the stimulus word is a Japanese word *jisho* 'dictionary'. Currently, the number of the stimulus words in the Noun-ACD is 1,100 and



Table 3: Examples of associated words in the Noun-ACD when the stimulus word is ‘dictionary’.

Semantic relation	Examples of associated words
Hypernym	<i>shuppanbutsu</i> ‘Publication’, <i>hon</i> ‘Book’
Hyponym	<i>waeijisho</i> ‘Japanese-English dictionary’
Part / Material	<i>midashigo</i> ‘Entry word’
Attribute	<i>muzukashii</i> ‘Difficult’, <i>yasashii</i> ‘Easy’
Synonym	<i>jiten</i> ‘Encyclopedia’
Action	<i>yomu</i> ‘Read’, <i>shiraberu</i> ‘Investigate’
Situation	<i>toshokan</i> ‘Library’, <i>honya</i> ‘Book store’

the number of participants is 50. The total number of associated words is approximately 280,000. When all of overlapping words are eliminated, the number of associated words is about 64,000.

### Proposed Method for Detecting Metonymies

To detect metonymic expressions in sentences, we use associative information between words in the Verb-ACD and the Noun-ACD. Our proposed method extracts attribute values of input sentences and detects metonymic expressions with decision tree learning. We first describe our basic idea, and then, the attributes of decision tree learning.

#### Basic Idea for Metonymy Detection

Semantic relations between metonymic expressions and their predicates seem to be more unnatural than that of literal expressions and their predicates. Hence, it is natural for humans to associate more literal expressions from predicates than metonymic ones. Our basic idea therefore is that the degree of word distances in the Verb-ACD and the Noun-ACD can express the measures of judging expressions as ‘Metonymic’ or ‘Literal’.

A method based on the basic idea is detecting metonymic expressions with associative information by using relations of two paths of synset nodes in the Japanese WordNet (Isahara, Bond, Uchimoto, Utiyama, & Kanzaki, 2008). One is the path from synsets of associated words to their hypernym synsets. The other is from synsets of each word in a sentence to their hypernym synsets. If there is a shared synset node between these two paths, the word in the sentence is regarded as a literal expression. On the other hand, it is possible to be a metonymic expression if there is no shared synset. Our system outline consists of four steps:

1. **Morphological and Syntactic Analyses.** The system analyzes an input sentence morphologically and syntactically by using McCab and CaboCha, respectively.
2. **Extraction of Associative Information.** From the results of morphological and syntactic analyses, the system extracts a predicate in the sentence and its modification relations. When the predicate is a verb or a verbal noun followed by *suru*, e.g., *taiho-suru* ‘arrest (verb)’ where *suru* added to *taiho* ‘arrest (noun)’, the shortest and the second-shortest associated words from a pair of the predicate verb

and a particle corresponding to the semantic relation in Table 2 are extracted from the Verb-ACD. If the sentence has more than one particle, the system extracts associated words from each noun with the particle. If the predicate is anything except a verb, two stimulus words of the noun as an associated word with the semantic relation Attribute in Table 3 are extracted from the Noun-ACD. In the same manner as the case with the predicate verb, these word distances are the shortest and the second-shortest ones between the predicate, i.e., the associated and the stimulus word.

3. **Extraction of Noun Information.** The system extracts synsets and hypernym synsets of all nouns in the sentence from the Japanese WordNet. These hypernym synsets are all synsets which the system obtains from nouns in the sentence to each third upper level for the synset hierarchy. If there are proper nouns in the sentence, it extracts each synset of properties which are from the result of the morphological analysis because the Japanese WordNet does not have enough synsets of proper nouns. For example, if one of the proper nouns in the sentence in Table 1 is *hanoi* ‘Hanoi’, the system extracts synsets and hypernym synsets of *chiiki* ‘LOCATION’ which is a property from the result of morphological analysis.
4. **Confirmation of Shared Synset.** By comparing synsets and hypernym synsets of the associated words with those of nouns or the properties of proper nouns in the sentence, the system confirms whether a shared synset node is between both paths of synset nodes. If there are one or more shared synsets, the system judges the noun as ‘Literal’. On the other hand, if there is no shared synset, the system judges it as ‘Metonymic’.

The system thus decides on the correct category, ‘Metonymic’ or ‘Literal’, of every noun in input sentences and can detect metonymies with associative information.

#### Metonymy Detection using Decision Tree Learning

We prepared attributes shown in Table 4 for the decision tree learning. These attributes are all factors obtained in the basic idea.

*Semantic\_relation* represents semantic relations corresponding to particles with nouns in sentences. In addition, one of its values ‘Noun’ was used when the predicate was not a verb. *Distance\_1st\_candidate* and *Distance\_2nd\_candidate* were the shortest word distance and the second one between the predicate and the associated word, respectively. *Number\_A\_synset* and *Num\_A\_hyponym* were the number of synsets of the associated words and the sum of hypernym synsets from the synsets for three upper levels, respectively. *Num\_N\_synset* and *Num\_N\_hyponym* were also the number of synsets of nouns in the sentence and the sum of hypernym synsets for three upper levels. *Num\_HN\_synset* and *Num\_HN\_hyponym* were the number of synsets of the noun’s hypernyms and the sum of hypernym synsets of the

Table 4: Attributes and values with decision tree learning.

Attribute	Description	Value
<i>Semantic_relation</i>	Semantic relations corresponding to particles with nouns in a sentence	Agent, Object, Source, Goal, Location, Tool, Noun
<i>Distance_1st_candidate</i>	The shortest word distance between the predicate and associated words	Continuous
<i>Distance_2nd_candidate</i>	The second shortest word distance between the predicate and associated words	Continuous
<i>Number_A_synset</i>	The number of synsets of associated words	Continuous
<i>Number_A_hyponym</i>	The sum of hyponym synsets from the associated words for three upper levels	Continuous
<i>Number_N_synset</i>	The number of synsets of nouns in a sentence	Continuous
<i>Number_N_hyponym</i>	The sum of hyponym synsets from the nouns for three upper levels	Continuous
<i>Number_HN_synset</i>	The number of synsets of hyponyms of nouns in a sentence	Continuous
<i>Number_HN_hyponym</i>	The sum of hyponym synsets of hyponyms of the nouns in a sentence	Continuous
<i>Match_node</i>	The degree of linked nodes from each synset of the associated words and the nouns in a sentence to a shared synset	None, Near, Middle-Near, Middle, Middle-Far, Far

hyponyms for two upper levels to equalize hyponym levels from initial synsets as above, i.e., three upper levels. Let *Match\_node* denote the degree of linked synset nodes from each synset of the associated words and the nouns in the sentence to the shared synset. By using the sum number of linked nodes, this degree was separated to the following six levels: ‘None’, ‘Near’, ‘Middle-Near’, ‘Middle’, ‘Middle-Far’, and ‘Far’. ‘None’ means that there was no shared synset, i.e., the noun was judged as ‘Metonymic’. ‘Near’ means that either of the synset of the associated word or that of the noun in the sentence was just the shared synset at least, i.e., the sum of linked nodes was 0 or 1. ‘Middle-Near’ means that the average of each node was 1, i.e., the sum of linked nodes was 2. ‘Middle’ means that the sum of linked nodes was 3. ‘Middle-Far’ means that the average of each node was between 2 and 3. ‘Far’ means that the average of each node was more than 3, i.e., the sum of linked node was more than 6.

## Experiment

To evaluate our method, we prepared a baseline system where the Goi-Taikei—A Japanese Lexicon (Ikehara et al., 1999) was used to automatically detect metonymies. We prepared test sentences with literal and metonymic expressions and evaluated our method by comparing its recall, precision, and F-measure rates with those of the baseline. In this section, we describe the baseline, test sentences, and the evaluation results.

### Baseline System

The baseline system consisted of syntactic structures and noun properties in the Goi-Taikei, which was used for detecting metonymies (Murata et al., 2000). It first selects a syntactic type of the predicate using its syntactic information in the Goi-Taikei after morphological and syntactic analyses of an input sentence. It employs the highest priority order of syntactic information in each predicate verb because this order

indicates an order of preference in the Goi-taikei. The preference order was defined in order to translate from Japanese to English or from English to Japanese (Shirai, Ooyama, Ikehara, Miyazaki, & Yokoo, 1998). The syntactic information on each verb is a set of syntactic type and noun properties, and expresses that each verb has nouns with a part of speech. The baseline system then obtains nouns in the syntactic information and their properties. These noun properties consist of some nouns and are expressed by the hyponyms and hyponyms in the noun semantic hierarchy. Finally, the system judges the word as ‘Metonymic’ if each word in the sentence does not belong to the noun’s hyponyms in the hierarchy.

### Test Sentences

We prepared 90 test sentences which consisted of 45 ones with metonymic expressions and 45 ones with literal expressions. As shown in Table 5, most of the former sentences were extracted from the previous studies (Murata et al., 2000; Yamanashi, 1988). The latter were extracted from newspaper corpora of the Mainichi Newspaper (‘93–’95 and ‘03–’04) and included words used in the metonymic sentences. In 90 test sentences, there were 113 nouns which both our method and the baseline judged as ‘Metonymic’ or ‘Literal’.

### Results and Discussion

To judge each noun as ‘Metonymic’ or ‘Literal’, we extracted attributes from 90 test sentences and constructed 113 cases. We trained 112 cases, tested the other case with the training data, and repeated this procedure in a round-robin fashion. By running 113 folds, each case was judged as ‘Metonymic’ and ‘Literal’. From Table 6, we can see that our method judged correctly 95 cases and the baseline system did 81 cases correctly. Our method showed higher accuracy (0.84) than that of the baseline. There was significant difference ( $p < 0.05$ ) between them. Here, the statistical difference was determined by McNemar’s test. The evaluation measurements

Table 5: Examples of test sentences (in Japanese).

Metonymic sentence (English translation)	Literal sentence (English translation)
<i>isshoubin-wo nonda</i> (Someone drank the issho-bottle.)	<i>isshoubin-wo saidan-ni oita</i> (He places the issho-bottle on the altar.)
<i>kasetsu-ga genri-wo setsumei-suru</i> (The hypothesis explains the elements.)	<i>kankeisha-ga setsumei-shita</i> (People involved explained that.)
<i>shirobai-ga ihansha-wo taiho-shita</i> (The police motorcycle arrested the criminals.)	<i>keisatsukan-ga hanzaisha-wo taiho-shita</i> (The police man arrested the criminals.)
<i>shikisha-ha sono-clarinet-wo waratta</i> (The conductor laughed at the clarinet.)	<i>jibun-wo waratta</i> (Someone laughed about oneself.)
<i>kao-wo soru</i> (Someone shaves own face.)	<i>hige-wo soru</i> (Someone shaves a beard.)
<i>atama-wo karu</i> (Someone clips own head.)	<i>tanbo-de ine-wo karu</i> (Someone mows rice plants in the paddies.)

Table 6: Accuracy in judging whether metonymic expressions or literal meanings. Asterisk indicates statistical significance over baseline. (\*  $p < 0.05$ )

	Baseline	Proposed method
Accuracy	0.72 (81/113)	0.84 (95/113)*

Table 7: Precision, recall, and F-measure rates in detecting metonymic expressions.

	Baseline	Proposed method
Precision	0.63 (31/49)	0.85 (33/39)
Recall	0.69 (31/45)	0.73 (33/45)
F-measure	0.66	0.79

were recall, precision, and F-measure calculated by using the numbers of correct detections above. Our method expressed higher recall (0.73), precision (0.85), and F-measure (0.79) than those of the baseline system as shown in Table 7.

The two main reasons for our method’s superiority are as follows. First, there were differences between our method and the baseline in the way that knowledge was used. As described previously, the baseline used the highest priority order of syntactic information in each predicate. The priority order in the Goi-Taikai was defined as preference to translate, so it seemed to express the order of frequency of its usage (Shirai et al., 1998). From these, the baseline system used the highest frequency of syntactical information of the predicates. On the other hand, information on the predicates which our method used was short word distances between them and their associated words in the Verb-ACD and the Noun-ACD. From the results, it seemed to be more suitable to use the associative information of predicates. The second reason is that separating stages of *Match\_node* was a good way to detect metonymies. Here, to investigate the detail of our method, we show the result of the decision tree learning in training 113 cases in Figure 1. As shown in the figure, *Match\_node* in ‘None’ or ‘Far’ was judged as ‘Metonymic’ and that in ‘Near’ or ‘Middle’ was done as ‘Literal’. As mentioned previously, ‘Far’ means that the average of each node is more than 3. There

```

Match_node in {None, Far}: Metonymic (32/5)
Match_node in {Near, Middle}: Literal (43/6)
Match_node = Middle-Near:
...Distance_2nd_candidate <= 2.74: Metonymic (3)
: Distance_2nd_candidate > 2.74: Literal (8)
Match_node = Middle-Far:
...Number_S_hyponym <= 19: Literal (22/4)
: Number_S_hyponym > 19: Metonymic (5)

```

Figure 1: Result of decision tree learning in 113 cases.

are more synsets of abstract nouns in higher levels hence it is natural to be judged as ‘Metonymic’ in ‘Far’ where the matching synset is at higher levels in the mean. On the other hand, it is also natural to be judged as ‘Literal’ in ‘Near’ or ‘Middle’. From these, the sum of both the synset node from associated words and that from nouns indicates the measures of detecting metonymies.

Given an example of the results, when an input Japanese sentence was *shikisha-ha sono-clarinet-wo waratta* ‘The conductor laughed at the clarinet.’ in Table 5, our method judged ‘clarinet’ as ‘Metonymic’ while the baseline could not. In the Verb-ACD, the associated words whose distances were especially short were *hito* ‘human’ and *telebi-bangumi* ‘TV program’. Therefore, it extracted these associated words, their synsets, and hypernym synsets from Japanese WordNet. It then compared them with ‘clarinet’ and its synset expressed by music instruments. Since the extracted words and their synsets did not match ‘clarinet’ and or its synset, the expression was judged as ‘Metonymic’. Meanwhile, the baseline extracted syntactic information of the following predicate verb ‘laugh’ from the Goi-Taikai: “N1 laughs at N2” where noun properties of “N1” and “N2” were *hito* ‘human’ and *asterisk* ‘all properties’, respectively. The property of ‘clarinet’ was *gakki* ‘instrument’ and belonged to “N2” whose property was *asterisk* ‘all properties’. As a result, the baseline system judged ‘clarinet’ as ‘Literal’. In general, we usually understand the meaning ‘The conductor laughed at the clarinet player’ when we read the sentence. Of course, it is not wrong syntactically that the conductor laughed at the instru-

ment of clarinet, but it is unnatural in daily conversations. Our method was closer to our associations in daily conversations and more appropriate to detect metonymies than the baseline. We therefore conclude that using associative information can improve computer's ability to detect metonymies as humans do.

However, our method incorrectly judged some literal expressions as 'Metonymic'. The reason was that some associated words in the Verb-ACD and those in their synsets in the Japanese WordNet were metonymies. Our method incorrectly judged some metonymic expressions as 'Literal' because the variety of associated words with the short word distances was sometimes too restricted. This small variety within the group of associated words could have led to a smaller range in the search space of the Japanese WordNet, leading to the tendency to detect too many metonymies.

### Summary and Future Work

We used the Verb-ACD and the Japanese WordNet to detect metonymic expressions in sentences with associative information. We found that our method has a higher accuracy of judging 'Metonymic' or 'Literal', recall, precision, and F-measure of detecting metonymies than those of the baseline that only uses syntactic and semantic information.

Future work includes detecting metonymies for the temporal contiguity and constructing a system for interpreting metonymic expressions. We would like to integrate them into our current detection method to improve our analysis of metonymy.

### Acknowledgments

This work has been partially supported by the Graduate School Doctorate Student Grant Aid Program 2011, Keio University. We would like to thank the students at Shonan Fujisawa Campus of Keio University for their participation in the association experiments.

### References

- Bouaud, J., Bachimont, B., & Zweigenbaum, P. (1996). Processing metonymy: a Domain-Model Heuristic Graph Traversal Approach. In *Proceedings of the 16th International Conference on Computational Linguistics* (Vol. 1, pp. 137–142).
- Fass, D. (1991). met\*: A Method for Discriminating Metonymy and Metaphor by Computer. *Computer Linguistics*, 17(1), 49–90.
- Ikehara, S., Miyazaki, M., Shirai, S., Yokoo, A., Nakaiwa, H., Ogura, K., et al. (1999). *Goi-Taikei: A Japanese Lexicon CD-ROM*. Iwanami Shoten.
- Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., & Kanazaki, K. (2008). Development of Japanese WordNet. In *Proceedings of the 6th International Conference on Language Resources and Evaluation* (pp. 2420–2422).
- Iverson, E., & Helmreich, S. (1992). Metallel: An Integrated Approach to Non-Literal Phrase Interpretation. *Computational Intelligence*, 8(3), 477–493.
- Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.
- Markert, K., & Nissim, M. (2003). Corpus-Based Metonymy Analysis. *Metaphor and Symbol*, 18(3), 175–188.
- Markert, K., & Nissim, M. (2007). Semeval-2007 task 08: Metonymy resolution at semeval-2007. In *Proceedings of the 4th international workshop on semantic evaluations (SemEval-2007)* (pp. 36–41).
- Morita, Y. (1989). *A Dictionary of Basic Japanese*. Kadokawa Gakugei Shuppan Publishing.
- Murata, M., Yamamoto, A., Kurohashi, S., Isahara, H., & Nagao, M. (2000). Metonymy Interpretation Using the Examples, "Noun X of Noun Y" and "Noun X Noun Y". *Journal of Japanese Society for Artificial Intelligence*, 15(3), 503–510. (in Japanese)
- Nastase, V., & Stube, M. (2009). Combining Collocations, Lexical and Encyclopedic Knowledge for Metonymy Resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 910–918).
- Okamoto, J., & Ishizaki, S. (2001). Construction of Associative Concept Dictionary with Distance Information, and Comparison with Electronic Concept Dictionary. *Journal of Natural Language Processing*, 8(4), 37–54. (in Japanese)
- Shirai, S., Ooyama, Y., Ikehara, S., Miyazaki, M., & Yokoo, A. (1998). Introduction to Goi-Taikei: A Japanese Lexicon. In *Ipsj sig notes* (pp. 47–52). (in Japanese)
- Suga, T., & Ishizaki, S. (2006). Construction of Metonymy Understanding System Using Associative Concept Dictionary. In *Proceeding of Annual Meeting of the Natural Language Processing Society of Japan (nlp2006)* (pp. 817–820). (in Japanese)
- Taniguchi, K. (2003). *New Developments of Cognitive Semantics -Metaphor and Metonymy-*. Kenkyusha. (in Japanese)
- Teraoka, T., Okamoto, J., & Ishizaki, S. (2010). An associative concept dictionary for verbs and its application to elliptical word estimation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (pp. 3851–3856).
- Teraoka, T., Okamoto, J., & Ishizaki, S. (2011). Detecting metonymic expressions with associative information from words. In *Proceedings of the 12th Pacific Association for Computational Linguistics Conference*, #13.
- Yamanashi, M. (1987). Metonymic interpretation and associative processes in natural language. In M. Nagao (Ed.), *Language and artificial intelligence* (pp. 77–86). Elsevier Science Publishers B.V. (North-Holland).
- Yamanashi, M. (1988). *Metonymy and Understanding*. University of Tokyo Press. (in Japanese)

# Flexible sequence learning in a SOM model of the mirror system

**Serge Thill (serge.thill@his.se)**

Interaction Lab, School of Humanities and Informatics  
University of Skövde  
54128 Skövde, Sweden

**Josef Behr (jbehr@uos.de)**

Institut für Kognitionswissenschaft, Neurokybernetik  
University of Osnabrück  
49069 Osnabrück, Germany

**Tom Ziemke (tom.ziemke@his.se)**

Interaction Lab, School of Humanities and Informatics  
University of Skövde  
54128 Skövde, Sweden

## Abstract

We present initial work on a biologically and cognitively inspired model that may allow embodied agents to autonomously learn sequences of action primitives (forming an overall behaviour). Specifically, we combine a flexible model of sequence generation with a model of parietal mirror neuron activity. The main purpose is to illustrate that the approach is viable. Although further work is needed to improve the results sketched out here, the concept is sound and relevant both to efforts in modelling mirror neuron activity and enabling artificial embodied agents to autonomously learn sequences of action primitives.

**Keywords:** Behavioural sequence learning; Ordinal node model; Self-organising maps; Mirror neurons

## Introduction

We are concerned with the problem of generating sequences of action primitives which are flexible with respect to the precise time it takes to execute the different components (primitives) of the same sequence at different times. A thorough discussion of the issue is given, for instance, by Sandamirskaya & Schöner (2010). In a nutshell, part of the problem is that one cannot simply chain together the different primitives through, for example, simple Hebbian learning. Rather, mechanisms must exist for keeping track of the current location in the sequence, including ways of verifying that the current action has successfully completed or failed to complete. Sandamirskaya & Schöner (2010) describe a general framework which can address these issues and we briefly sketch the main points in the next section.

Overall, the aim of the work in the present paper is to combine said framework with a model of parietal mirror neuron activity (Thill et al., 2011) and to illustrate that such an approach is, in principle, viable. Importantly, since the mirror neuron model used here autonomously organises itself, the work proposed here may be relevant and helpful in designing artificial embodied agents that should autonomously learn sequences of actions and use them to predict actions of others.

## Sequencing via ordinal nodes and conditions of satisfaction

The gist of the framework by Sandamirskaya & Schöner (2010) is the existence of *ordinal nodes* which essentially count through the sequence. These nodes are implemented via coupled dynamical systems (see *Methods*), designed so that only one node can be active at a time. Upon completion of the element of the sequence represented by the active node, activation is passed onto the next node in the sequence. In their work (e.g. Sandamirskaya & Schöner, 2010; Sandamirskaya et al., 2011), the action primitives forming the sequence exist in the sensorimotor representation of an embodied agent, implemented using techniques from Dynamic Field Theory (Schöner, 2009; Spencer et al., 2009). This has the advantage that the sensorimotor representations of these primitives are stable (since they are essentially stable fixed-point attractors), which makes it particularly simple to link specific locations in the dynamic fields representing the sensorimotor space of the agent to specific ordinal nodes. Part of the challenge of the work presented in the present paper is to illustrate that the ordinal node system could also be attached to a representation with more noise and less stability than dynamic fields.

The decision that a given action primitive has completed is implemented a separate system (also exploiting dynamic fields) that checks for a *Condition of Satisfaction* (CoS). One of the open challenges here is the question of how to best learn the CoS for specific primitives (including identifying that the primitive has, for whatever reason, failed). It is not the purpose of the present work to address the open issues regarding the CoS - rather, we focus on combining the ordinal node model with a model of mirror neuron activity discussed in the next section.

## Mirror system sequences

One example of sequencing in biology is given by the hypothesised functioning of the mirror system. Without entering the debate on what higher-level cognitive abilities mirror

neurons may or may not be useful/essential for (see for instance Hickok, 2008; Rizzolatti & Sinigaglia, 2010, for such a debate), it appears that parietal mirror neurons in macaque monkeys organise into pools of neurons responding to specific motion primitives (*e.g.* a *reach* or a *grasp* but not both; Fogassi et al., 2005). It has then been hypothesised (*e.g.* Chersi et al., 2006) that these pools of neurons can be chained together to form sequences of simple, often-encountered actions (such as *reach-grasp-bring-to-mouth* for eating). Models on the basis of this hypothesis have proven useful, for instance, in putting forward theories unifying apparently conflicting results on interference and facilitation in action language processing (Chersi et al., 2010).

A particular model that specifically addresses the development of parietal mirror neurons has been previously presented by some of us (Thill et al., 2011). This model uses a self-organising map (SOM) to illustrate how a “blank” structure, through the organisational principles of SOMs can autonomously form an organisation whose activity resembles that of parietal mirror neurons.

The inputs to the model represent an arbitrary encoding of observed (or executed) motion primitives (*e.g.* based on changes in position per time step) and contextual information (including, for instance, affordances in the perceived scenery). These two components are sampled from two distinct spaces (of arbitrary dimensionality) and concatenated into a single input vector as required by standard SOM implementations (Kohonen, 1997). The model is trained on repeated presentations of all combination of motion primitives and contexts. After training, the model can be run on-line by continuously feeding it input vectors and some plasticity (allowing, for instance, the learning of new primitives) can be retained by not reducing the learning rate to 0 (albeit keeping it at a low level, see Thill & Ziemke, 2010).

The trained maps organise in a fashion remarkably similar to that of parietal mirror neurons (Fogassi et al., 2005): Within the map, different areas encode different action primitives (which could represent motions such as *reaching*, *grasping* or *bring-to-mouth*, similar to *e.g.* Chersi et al., 2006). Within the area encoding one such primitive, some nodes are active whenever the model input encodes that primitive. Others are active only if the action input additionally encodes a specific context in which the primitive is observed (usually sufficient to specify the most likely goal of the action, see Thill et al., 2011). The proportion of context-independent nodes is a direct consequence of the way inputs are represented (specifically, of the ratio between the maximal variability in encoding the primitives and contextual information respectively, called  $\beta$  in the model). Exploring how  $\beta$  (for which values between 1 and 5 cover most aspects of interest) affects the organisation of the maps revealed that, for  $\beta \approx 3.5$ , the proportion of context-independent nodes is similar to the corresponding neurophysiological data observed in the parietal mirror area of macaques (Fogassi et al., 2005).

## Combining models

Previous models of parietal mirror neuron activation tend not to address the timing aspect of the chains in much detail, focussing instead on merely linking the different pools forming a chain through hard-coding (Chersi et al., 2006, 2010) or, for instance, Hebbian learning (Erlhagen et al., 2007). With the exception of Chersi et al. (2010), these models do not take into account that the pools encoding the same primitive under different goals are not entirely distinct (Fogassi et al., 2005). Thill et al. (2011), whose main focus is the exact nature of this overlap between populations, do not specifically address chain formation at all.

The present paper therefore presents an augmented version of the model from Thill et al. (2011). Specifically, we now implement the learning of chains of primitives, using the approach of Sandamirskaya & Schöner (2010). This new model then allows us to address a number of open issues: to what extent is the ability to activate the correct (and only the correct) sequence of events (given the first element) affected by the overlap between neural populations? When observing an action primitive in an unknown context, is it possible to predict all possible chains this action could be part of?

These issues are relevant, both for our understanding of (in particular) sequences in mirror neuron activity and for the ability to endow artificial agents with similar abilities. If one subscribes to the hypothesis that mirror neuron activity helps us understand the actions of others (see Rizzolatti & Sinigaglia, 2010, for a thorough review and discussion), then the ability to predict the likely outcome of an action given the initial movement based on the resulting mirror neuron activity is a desirable ability. This includes the ability to autonomously learn sequences of actions as well as the ability to both correctly identify a sequence if the context is clear and predict all possible sequences if the context is ambiguous (for instance, a familiar gesture observed in a completely new context).

## Methods

### Overall model design

The model (Fig. 1) is composed of a self-organising map which is meant to represent parietal mirror neuron activation (Thill et al., 2011) and an ordinal node model for sequence learning (Sandamirskaya & Schöner, 2010). The activity over time in the SOM is used (1) to train the sequence learning model, (2) to activate learned sequences and (3) to provide the input necessary to move from one sequence element to the next. It therefore combines the idea of chaining pools of neurons (*e.g.* Chersi et al., 2006) with the flexible execution of sequences provided by the ordinal nodes model of Sandamirskaya & Schöner (2010).

### Self-organising maps as a mirror system

The self-organising maps used in this paper are in essence identical to those used by Thill et al. (2011) and are trained in the same manner. The only difference is that the previous maps explicitly dedicated part of their space to the theoretical

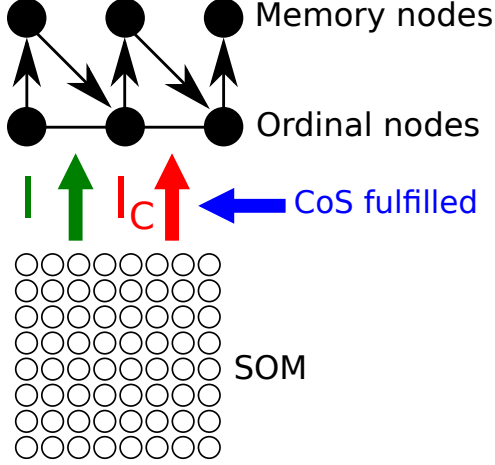


Figure 1: **Overall model architecture.** Activation in the SOM feeds into the ordinal nodes, both to activate the sequence (green) and to move between sequence elements (red) if the CoS is fulfilled (blue). Connections between the SOM and the nodes are bidirectional; node activity can thus also be used to activate regions in the SOM (omitted in the figure for clarity).

possibility of learning motion primitives from a second limb (see Thill & Ziemke, 2010, for details). Since that is irrelevant here, the present maps are trained assuming the need to represent just one limb. The trained maps therefore represent, as before, five motion primitives observed under two different contexts. They behave as described in the introduction: input vectors consisting of a concatenation of observed/executed motion encoding and contextual information are continuously fed to the map. Depending on the previously discussed ratio  $\beta$ , some nodes of the map will be active regardless of the contextual information whereas others will be sensitive to the latter (see Thill et al., 2011, for a complete discussion of the definition of activity).

### Ordinal node model

The ordinal node model used here largely follows Sandamirskaya & Schöner (2010) and is described by the following equations:

$$\begin{aligned} \tau \dot{d}_i(t) = & -d_i(t) + h_d + c_0 f(d_i(t)) \\ & - c_1 \sum_{i' \neq i} f(d_{i'}(t)) + c_2 f(d_{i-1}^m(t)) \\ & - c_3 f(d_i^m(t)) - c_{CoS} I_C(t) + c_{in} I \end{aligned} \quad (1)$$

$$\begin{aligned} \tau \dot{d}_i^m(t) = & -d_i^m(t) + h_m + c_4 f(d_i^m(t)) \\ & - c_5 \sum_{i' \neq i} f(d_{i'}(t)) + c_6 f(d_i(t)) \end{aligned} \quad (2)$$

where  $d_i$  refers to the activation of the  $i$ th ordinal node (and  $d_i^m$  is the associated memory node needed for proper

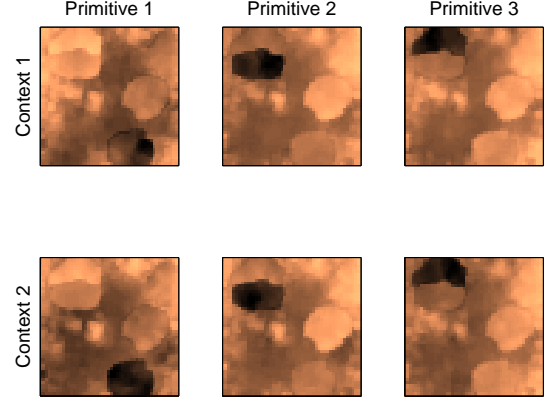


Figure 2: **Sequences in the SOM.** Shown are the activation of three primitives (columns) seen under two different contexts (rows) for a map with  $\beta = 3$ . Dark regions indicate most activity and are clearly in different locations for different primitives. For the same primitives in different contexts, similar regions are activated but the overlap between the most active neurons in each context is limited (see Thill et al., 2011, for a thorough discussion).

functioning, see Sandamirskaya & Schöner (2010) for details),  $f(\cdot)$  is a sigmoidal nonlinearity and the constants in the present implementation are chosen as:  $c_0 = 7.2$ ,  $c_1 = 3.6$ ,  $c_2 = 4.8$ ,  $c_3 = 0.8$ ,  $c_4 = 4$ ,  $c_5 = 2$ ,  $c_6 = 2.6$ ,  $c_{in} = 0.1$ ,  $c_{CoS} = 0.2$ ,  $h_d = -5$  and  $h_m = -2$ . A detailed discussion of the functioning of the model is given by Sandamirskaya & Schöner (2010). We deviate in two minor aspects: (1) The term  $c_{in}I$  is added and provides an external input (obtained from the activity in the SOM described above). This is only used at the beginning of a sequence to activate the first ordinal node. (2) We simplify the Condition-of-Satisfaction (CoS) aspect. In the original model, this is given by an additional dynamic field which is able to “perceive” that the CoS has been reached. Here, the inhibitory activation is obtained from the same SOM that would provide  $I$  for the activation of the first node, which simplifies the design of the model. Since the model is not actually implemented in an agent, there is also no point in devising a sophisticated “perception” of the CoS here. Rather, the CoS is presumed fulfilled after a randomly chosen number of time-steps and the inhibitory activation released to the ordinal model, thus moving the model onto the next element of the sequence. This is acceptable for the present purposes since the point here is to illustrate the learning of sequences, not the ability to autonomously detect that an element of a sequence has completed (or failed to complete). An implementation of this model in an agent would of course need to address this aspect in more detail.



## Task and learning

For each  $\beta$  value between 1 and 5 (in increments of .5), 100 maps have been generated. Each map is activated manually with a series of input vectors which simulate a sequence of 3 motion primitives being executed first in one context and then in another (see Fig. 2 for an example of two sequences). Two sets of ordinal nodes are used to learn these two sequences. Learning is achieved during manual activation of the map by clamping the relevant ordinal node to an active state and then using simple Hebbian learning to train weights between this node and all neurons in the map (with normalised activation). After training, any weights below a threshold of 0.5 are set to 0 to allow only the SOM nodes with the strongest activation to connect with the relevant ordinal nodes.

Of particular interest are the following questions: Will both sets of ordinal nodes correctly activate if the SOM activity is that of the first element of their respective sequences? Also, will a set of ordinal nodes trained on the first sequence remain *inactive* if the SOM activity represents the first element of the second sequence (and vice versa)? Illustrating these behaviours would confirm good performance of the model given that sequences are correctly activated if and only if the map activity corresponds to their first element. It should be remembered at this point that map activity is noisy and fluctuates over time - the task is therefore not trivial.

An additional interest is the behaviour of the model in case of ambiguous contextual information. As discussed in the introduction, this could correspond to observing a familiar primitive in an unfamiliar context and predicting what the likely outcome of the action could be. It is of course a matter of debate what the exact behaviour of the model should be in this case; one could for instance argue that it should depend on how similar the unfamiliar context is to previously encountered ones. Here, we simply investigate the behaviour if the vector encoding contextual information is truly ambiguous, namely by corresponding to the point in the input space whose coordinates are equidistant from the subspaces encoding all known contexts. In other words, the ambiguous context encoding vector cannot be uniquely assigned to any previously encountered case. We simply postulate that, in the absence of any information that could favour either of the chains, the desirable behaviour of the model is to activate both, essentially predicting that both behaviours are equally likely.

## Results

### Correct activation/non-activation

For each value of  $\beta$ , 100 sets of 2 sequences have been learned. Per set, the sequences differ only in the context in which they have been executed. As  $\beta$  increases, the proportion of neurons active in one but not both of the contexts decreases (Thill et al., 2011). It can therefore be expected that the basic task of correctly activating a sequence if the map activity corresponds to its first element (and not activating said sequence if the contextual information is that of the sec-

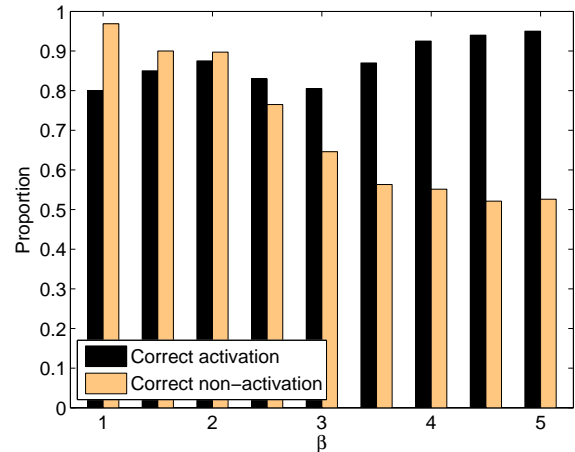


Figure 3: **Correct activation/non-activation.** Dark bars indicate the proportion of cases in which presenting the first element of a sequence correctly triggered the sequence. Light bars indicate the proportion of the cases that correctly trigger which correctly remain silent if the first element presented has the same motion primitive but different contextual information.

ond sequence) becomes harder as  $\beta$  increases. This is indeed what we find (see Fig. 3). Specifically, it is possible, in most cases, to correctly activate a sequence by presenting its first element in map activation (although it does fail on occasion, likely due to the noisy map activity). Importantly, this is independent of  $\beta$ , which is expected. The light bars in Fig. 3 then show how many (proportionally) of the sequences correctly activated by their own first element also remain silent when the first element of the second possible sequence is presented instead. As expected, this number decreases over time but remains over .5 in all cases.

However, this measure iterates over sequences that are correctly activated (or not); it does not measure the number of maps for which both sequences are correctly activated (or not). The evolution of this proportion is shown in Fig. 4 (black bars) and is decreasing more dramatically as  $\beta$  increases. At the same time, it should be noted that for e.g.  $\beta = 4$ ,  $\approx 60\%$  of nodes in the SOM encoding a given primitive are active independent of context (leaving only 20% capable of uniquely identifying each of the contexts).

### Correct behaviour under ambiguous context

The second interesting question was whether both sequences would be activated by the first motion primitive shown in a perfectly ambiguous context. Considered independently of the performance on the previous task, we find that a large number of models indeed activate both sequences given an ambiguous context. In particular, we find that this proportion increases with  $\beta$  (from 0.4 to  $> 0.9$ ), likely due to the increasing number of neurons which are active irrespective of con-

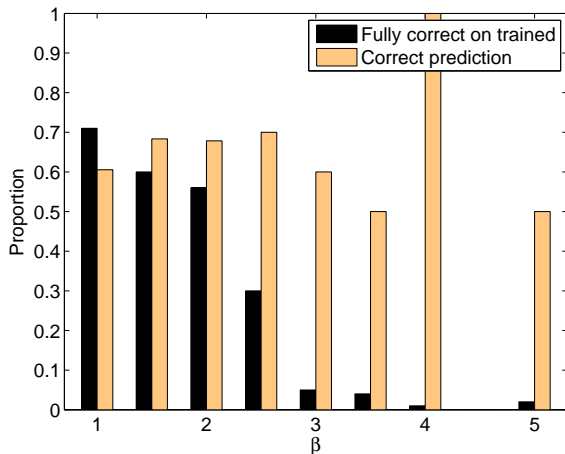


Figure 4: **“Perfect” models**. Black bars indicate proportion of cases in which the same model activates the correct sequence for each of the two learned sequences if presented with the first element of each sequence and does not activate the wrong sequence. Light bars indicate the proportion of cases fulfilling the first condition which also correctly activate both sequences if the contextual information is ambiguous

text.

This illustrates the expected effect of  $\beta$ : as the distinction between contexts diminishes, activation of both sequences is facilitated. However, this measure can again be seen as being a bit too general since there is not necessarily anything special about activating both sequences if the same model failed to *not* activate a sequence when primed with the first element (including the contextual information) of the second sequence. The more interesting question is therefore simply how many of the models that correctly behave given the learned sequences (black bars in Fig. 4 *also* behave as expected given the first primitive under an ambiguous context (within the context of this work, we can call these models “perfect”, since they fulfil all the expectations set out to them). Surprisingly, this proportion appears to be independent of  $\beta$  (light bars in Fig. 4), although it has to be kept in mind that for  $\beta \geq 3$ , the number of models which fulfill the first condition is rather low.

In other words, if a model is capable of correctly activating the relevant sequence (and only that sequence) given a full first element of that sequence, it is likely to *also* activate both sequences if given the first primitive under an ambiguous context. This is the most significant result in the present paper: although it is increasingly difficult to find a model which will correctly activate its sequences given the first element as  $\beta$  increases, it is then much easier to find a model which can also activate both sequences in the case of perfectly ambiguous contextual information.

## Discussion

### Insights from the model

Most of the results shown have their main purpose in illustrating that the model works as expected, including the increasing difficulty in obtaining “perfect” models as the  $\beta$  values of the underlying maps increase. The exact extent of this increase in difficulty is hard to judge from the work presented here as several aspects can be improved. First and foremost, the parameters for the ordinal nodes model are set independently of  $\beta$ , even though  $\beta$  has a rather significant effect on the input into the ordinal nodes and therefore the behaviour of the sequences. The fact that it was possible at all to create successful models across the entire range of  $\beta$  values underlines the potential of the approach. In future work, however, the focus would have to be on  $\beta$  values around 3 – 3.5, since these are the values for which the activity in the SOM most resembles that observed by Fogassi et al. (2005) in parietal mirror neurons (Thill et al., 2011).

The self-organising maps themselves are randomly generated; nonetheless it seems that some are more suited for a combination with an ordinal nodes model than others (since some “perfect” models were found even for  $\beta = 5$ , although the number was very low) and more work would be needed to investigate what features of these maps, if any, facilitate the task. Insights into this question could prove very valuable in more general future work combining the ordinal node model with sequences that are generated in systems which do not offer the “nice and clean” activation patterns of dynamic fields.

The connections between the activity in the map and the ordinal nodes are learned with a simple general Hebbian learning approach and the only transformation of the map activity consisted of a simple normalisation. Again, this is about the simplest approach imaginable and it is likely that improvements, including possible non-linear transformations of map activity, can lead to a higher proportion of “perfect” models for larger values of  $\beta$  (in particular of course for  $\beta \in [3, 3.5]$ ).

The most interesting result in the present paper was that the largest difficulty resided in finding models which perform as expected when started with a first element from either learned sequence and not, as one might have expected, in finding such a model that *also* perform correctly on the prediction task. This is encouraging as it illustrates that the concepts of using a combination of our previous SOM models of mirror neurons and the ordinal node model has potential, not just for generating the sequences one wishes to generate but also for predicting what sequences observed actions can be part of; this both in the case where the contextual information strongly favours one of the learned sequences and when the contextual information is perfectly ambiguous.

Again, there is a need for future work in this aspect. It seems reasonable (for the purposes of predicting likely sequences an observed primitive could belong to) to expect that a perfectly ambiguous context should activate all candidates but it is less clear - and beyond the scope of what can be

achieved in this space - what should happen if the context is merely ambiguous but closer in input space to some known contexts than others. Should the model simply activate the most likely sequence or would one prefer a mechanism that could attach a confidence value - indicated for instance by the time it takes the first ordinal nodes of all candidate sequences to activate - to indicate most and least likely sequences?

### Overall relevance

The work presented here is relevant for at least two areas. First is the modelling of mirror neurons as it is one of the first attempts to explicitly include the idea that executing the same action primitive at different points in time can lead to different durations, thus going beyond simple Hebbian-type associations directly between the primitives forming an overall action sequence (e.g. Chersi et al., 2010). Second, by modelling the specific organisation of parietal mirror neurons (which can develop autonomously, see Thill et al., 2011) and using that as an input to the ordinal node system, the model may provide a way for an artificial agent to learn sequences of primitives online and autonomously, which is still an open challenge (Sandamirskaya & Schöner, 2010).

The practical future applications are thus primarily in the design of future artificial cognitive systems; however all aspects of the model are inspired by biology; any implementation of the model could thus also be relevant to improve our understanding of the analogous biological systems.

### Conclusion

We presented an initial implementation of a mirror system activity model augmented with a framework for generating sequences. The main purpose was that it is in principle feasible to use the ordinal node framework to this effect. Although further work is needed to improve the quality, it was possible to show that the model can learn sequences based on the noisy SOM activity as well as correctly predict the likely sequence an observed initial primitive can belong to (including predicting both if both are equally likely). Since the SOM autonomously organises, the model presented here may be a viable candidate for autonomous sequence learning using the ordinal node framework (Sandamirskaya & Schöner, 2010).

### Acknowledgments

This work was supported by the European Commission FP7 project *NeuralDynamics*, (A neuro-dynamic framework for cognitive robotics: scene representations, behavioral sequences, and learning), Grant agreement no. 270247.

### References

Chersi, F., Mukovskiy, A., Fogassi, L., Ferrari, P. F., & Erlhagen, W. (2006). A model of intention understanding based on learned chains of motor acts in the parietal lobe. In *Proceedings of the 15th annual computational neuroscience meeting*. Edinburgh, UK.

Chersi, F., Thill, S., Ziemke, T., & Borghi, A. M. (2010). Sentence processing: linking language to

motor chains. *Frontiers in Neurobotics*, 4(4), DOI:10.3389/fnbot.2010.00004.

Erlhagen, W., Mukovskiy, A., Chersi, F., & Bicho, E. (2007). On the development of intention understanding for joint action tasks. In *Proceedings of the 6th IEEE international conference on development and learning* (p. 140-145). Imperial College London.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science*, 308, 662-667.

Hickok, G. (2008). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience*, 21(7), 1229-1243.

Kohonen, T. (1997). *Self-organizing maps*. Heidelberg: Springer.

Rizzolatti, G., & Sinigaglia, C. (2010). The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. *Nature Reviews Neuroscience*, 11(4), 264-274.

Sandamirskaya, Y., Richter, M., & Schöner, G. (2011). A neural-dynamic architecture for behavioral organization of an embodied agent. In *IEEE 10th international conference on development and learning (icdl), frankfurt*.

Sandamirskaya, Y., & Schöner, G. (2010). An embodied account of serial order: how instabilities drive sequence generation. *Neural Networks*, 23, 1164-179.

Schöner, G. (2009). Toward a unified theory of development. In J. P. Spencer, M. S. C. Thomas, & J. L. McClelland (Eds.), (p. 25-48). Oxford.

Spencer, J. P., Perone, S., & Johnson, J. S. (2009). Toward a unified theory of development. In J. P. Spencer, M. S. C. Thomas, & J. L. McClelland (Eds.), (p. 86-118). Oxford.

Thill, S., Svensson, H., & Ziemke, T. (2011). Modeling the development of goal-specificity in mirror neurons. *Cognitive Computation*, 3(4), 525-538.

Thill, S., & Ziemke, T. (2010). Learning new motion primitives in the mirror neuron system: A self-organising computational model. In S. Doncieux et al (Ed.), *Sab 2010, Inai 6226* (p. 413-423). Heidelberg: Springer.

# Fuzzy Memory Theory and its Use in Cognitive Science

Chris Thornton  
Informatics  
University of Sussex  
Brighton  
BN1 9QH  
UK  
c.thornton@sussex.ac.uk

## Abstract

Fuzzy memory theory extends fuzzy set theory to the case of imperfectly performing memory devices. In fuzzy set theory, the key concept is that of graded set membership. The degree to which an item belongs to a set is specified by a continuous function. Fuzzy memory theory is organized around the analogous concept of *graded recall*. Items stored in a fuzzy memory are associated with cues, such that each item is recalled by provision of the corresponding cue. But unlike conventional memory (where cues are typically addresses) the recall process may vary in its degree of error. The item produced may embody missing information. The capacity of a fuzzy memory is then measured in terms of the net information content of recalled items. The theory has potential applications for new forms of technology, but also for the study of cognition. In particular, it can be the means of formalizing the properties of error-prone natural memory mechanisms. It can also supply a non-circular explanation for similarity-based category formation.

## Introduction

Since its innovation nearly half a century ago (Gottwald, 2010), fuzzy set theory has become an essential tool of analysis in a wide range of disciplines (cf. Zadeh, 1965, 1976, 1982). In essence, the theory is a generalization of the classical theory of sets, in which set membership becomes a continuous rather than all-or-nothing criterion. The framework is particularly of use in contexts where set membership has a probabilistic character. There have been many applications in cognitive science. For example, the theory has been used for analysis of concept combination and feature emergence (Osherson and Smith, 1981, 1982; Zadeh, 1965, 1976, 1982). In this approach, concepts are deemed to represent fuzzy sets of objects, on which basis combinational concepts can be seen to represent the (fuzzy) intersections of the sets associated with the constituents (Murphy, 2002).

The present paper extends the theory of fuzzy sets to the case of memory. In fuzzy set theory, we assume an item can be a member of a set to a greater or lesser degree. In the proposed extension, it is the degree to which items are recalled that is variable. Information theory (Shannon, 1948; Shannon and Weaver, 1949) provides the means of measurement. The imperfection with which a particular item is recalled can be related to the amount of missing information (i.e., uncertainty) the recalled item exhibits.

Formally, a fuzzy memory device is considered to be a function from cues to items. But the items produced by the memory have to be distinguished from the reference items that are deemed to be stored. The behaviour of a fuzzy memory device is thus characterized as

$$f : C \rightarrow X'$$

where  $C$  is the set of cues,  $X$  is the set of reference items that are considered to be stored and  $X'$  is a set that replaces each member of  $X$  with the item recalled.

The storage capacity of the device then depends on the information content of the recalled items. This is denoted  $I(X')$ . But in calculating capacity, we must also take account of  $C$ , the set of cues. If this is not done, a function that simply copies its argument has the potential to exhibit an arbitrarily large storage capacity. The capacity of a fuzzy memory device is thus defined to be the *net* information content of recalled items. This is the information content of the reference items less the content of the keys and the information that is lost in recall. Formally, the capacity of fuzzy memory  $f$  is

$$I(X') - I(C)$$

where  $C$  is the set of cues used to elicit members of  $X'$  by  $f$ .<sup>1</sup>

In fuzzy set theory, the constituents of a set are characterized using a continuous function. This allows any degree of set membership to be specified. Extending the idea to the case of memory, we require a probabilistic model for the variables on which reference items are constructed. These are termed *base variables* below. A model that imposes a distribution on each variable will not suffice, since it cannot accommodate information losses involving specific variable combinations. A fuzzy memory is therefore considered to be a *composite* of distributions, in which each distribution applies either directly or indirectly to one or more base variables. Information losses arising in the recall of base-variable combinations are then accommodated.

This constituency can be formalized in a recursive way. Letting a distribution be considered *contributory* if it

---

<sup>1</sup>The word ‘memory’ refers to fuzzy memory in all cases.

applies directly to a base variable, or to a variable whose values themselves designate contributory distributions, we can characterize the constituents of memory  $f$  thus:

$$f = \{P \mid \text{contrib}(P, f)\}$$

In this formula,  $\text{contrib}(P, f)$  is true just in case distribution  $P$  applies either directly or indirectly to any base variable of  $f$ . Variables that mediate indirectly applying distributions are termed contributory variables. This distinguishes them from the base variables on which reference items are constructed.

A storage criterion for fuzzy memory can then be formalized. Fuzzy memory  $f$  is deemed to store some item  $x$  if  $x$  can be probabilistically reconstructed from  $f$ 's distribution composite. The informational requirements are as follows. The reconstruction must reconstitute the original item with measurable (but potentially zero) loss of information, and this loss must be less than the information required to specify the cue for the reconstruction. An item is deemed to be stored, then, just in case it can be reconstructed with a net gain of information. The degree of recall is the net gain obtained.

The degree to which an item is recalled by a fuzzy memory is analogous to 'grade of membership' in fuzzy set theory. The evaluation is formalized as follows. The grade of recall for some item  $x$  given some cue  $c$  is

$$r(x, c) = I(x') - I(c)$$

where  $x'$  is the device's probabilistic reconstruction of  $x$ , and  $x'$  satisfies the requirement of being derivable from  $x$  by elimination of information. The definition of  $I(c)$  depends on how cues are constituted (see below). Given there are  $k$  bits of content in each base-variable instantiation, the value of  $I(x')$  is

$$I(x') = \sum_i k - H(x'_i)$$

where  $H(x'_i)$  is the entropy of the distribution derived for the  $i$ 'th variable of  $x'$ .

## Illustrations

To illustrate the use of fuzzy memory, it is convenient to look at the case of 1-bit devices. These are simple assemblies in which base and contributory variables are all binary. Variable values are the digits 1 and 0, each of which has an information content of 1.0 bit. An advantage of this type of device is that it is particularly easy to represent as a tree diagram (cf. Figure 1). It also allows a simple cueing protocol, in which each cue comprises some subsequence of the variable evaluations required to render a fully deterministic reconstruction. In this context, a cue is a sequence of reconstruction constraints.

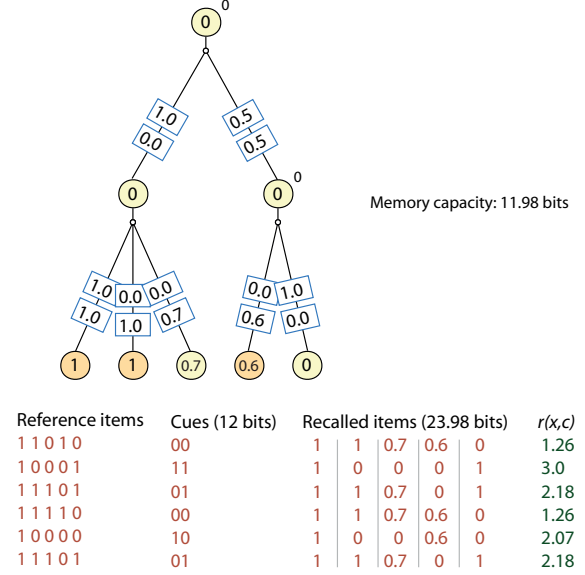


Figure 1: Recall performance of a 1-bit memory.

Consider Figure 1. This is a 1-bit fuzzy memory constructed for the reference items shown in the lower, left part of the figure. The tree structure in the upper part of the figure represents the composite of distributions. Notice this has two levels of construction. These are termed level 1 and level 2. Circles in the bottom row represent base variables. Circles elsewhere represent contributory variables. The digit seen within a circle is the value (or more generally distribution) the variable acquires in recall of the first reference item. Finally, the arcs leading down from each circle show which distributions are designated by which values of the relevant contributory variable.

Consider the leftmost contributory variable at the first level of the composite. By following the attached arcs downwards, we see that each value of this variable designates distributions applying to the three leftmost base variables. All distributions being over the values 0 and 1, only the probability applying to 1 is specified. Thus a value such as '0.6' is shorthand for the distribution  $\langle 0.4, 0.6 \rangle$ , i.e., the distribution that gives probability 0.4 to a value of 0 in the designated variable, and probability 0.6 to a value of 1.

Also of note is the way the distributional values are arranged. Over each set of arcs, we have two rows of boxes. Lower boxes contain distributions designated by a value of 0 in the contributory variable; upper boxes represent distributions designated by a value of 1. Applying these conventions, it should be possible to interpret all the distributions of Figure 1. For example, looking at the bottom, left part of the structure, we see that a value of 0 in the rightmost variable at level 1 designates a distribution on the third base variable which gives a probability



0.7 to the value 0.

The listing in the lower part of the figure portrays the behaviour of the device for the six reference items. (Distributions on variables are those obtained during recall of the initial item.) Each row shows the degree to which a particular item is recalled by its cue, with the associated recall grade  $r(x, c)$  appearing on the far right of the figure. As noted, each cue in a 1-bit memory is some subsequence of the disambiguating values required to produce a fully determined reconstruction. For the initial reference item  $\langle 1\ 1\ 0\ 1\ 0 \rangle$ , we have the cue 00. The initial digit in this sequence resolves the initial ambiguity in the reconstruction: i.e., it supplies the value 0 for the root contributory variable. The second digit resolves the next ambiguity arising. This affects the rightmost contributory variable at level 1. This value is subject to the equiprobable distribution developed in the previous step. The variable acquires the value 0.

With the cue now exhausted, the reconstruction continues in an un-cued way. The first, second and fifth base variables then acquire implicitly deterministic values, i.e., distributions that embody no loss of information. Regarding the third and fourth base variables, we have the distributions  $\langle 0.3, 0.7 \rangle$  and  $\langle 0.4, 0.6 \rangle$  respectively. Adopting the previous convention for representing binary distributions, the recalled item is then  $\langle 1\ 1\ 0.7\ 0.6\ 0 \rangle$ . There is of recall of 1.26 bits. This is the information obtained from the five bits of the reference item after deducting the two bits of the cue, and the 1.74 bits eliminated by the two information-losing distributions. Summing recall values for all six reference items, the total capacity of the memory is found to be 11.98 bits.

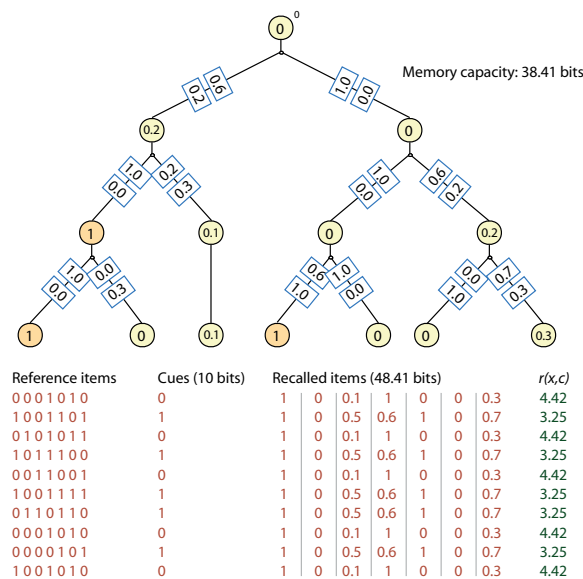


Figure 2: 1-bit memory with a capacity of 38.41 bits.

A more complex illustration of the 1-bit memory is provided by Figure 2. This is a fuzzy memory for the 10 reference items shown on the left. In this example, reference items are defined in terms of seven base variables rather than five. Contributory variables exist at three levels of construction rather than two. The cueing protocol remains the same. But, here, cue sequences comprise a single digit, meaning they supply values for the root variable only. The main part of the construction process thus proceeds in an un-cued way, with the result being a noticeable increase in the uncertainty of what is recalled. However, with the informational cost of cues at a low level, the memory capacity remains relatively substantial.

## Applications in cognitive science

Fuzzy memory theory formalizes a statistical form of memory in a way that reflects fuzzy set theory. Applications of a technological nature are one possibility. But might there also be applications in cognitive science? One way the theory could be used involves natural forms of memory. Human memory is notoriously error-prone, perfect recall being more the exception than the rule. There may be potential, then, for using the framework as a way of theoretically modeling human and other biological forms of memory.

Another possible application involves modeling development of categorical and conceptual representations. It is widely believed that such behaviours are at the heart of cognition (e.g. Harnad, 2005) and that categories are constructed so as to group entities by similarity (Machery, 2009). But it remains a considerable challenge to explain why this should be the case (Murphy, 2002). The temptation is to say that categories are formed as a *result* of the ways in which similarities are identified. Unfortunately, this makes no sense if the way we identify similarities depends on the category representations we bring to bear. With similarity being used to explain both why an entity is assigned to a certain category, and also why that category exists in the first place, such theories are placed ‘in the perilous position of using explanations which presuppose the very notions that they attempt to explain’ (Hahn and Chater, 1997, p. 84).

Fuzzy memory theory has the potential to address this dilemma. The critical factor affecting performance in fuzzy memory is the degree to which recalled items resemble reference items. The more closely each recalled item approximates its referenced counterpart, the less information is lost and the greater the capacity of the memory. But notice that recall loses less information if distributions deploy more extremal probabilities (i.e., probabilities closer to 1.0 or 0.0). At the same time, distributions must fulfil the function of modeling the referenced items; i.e., they must be the means of constructing valid approximations.

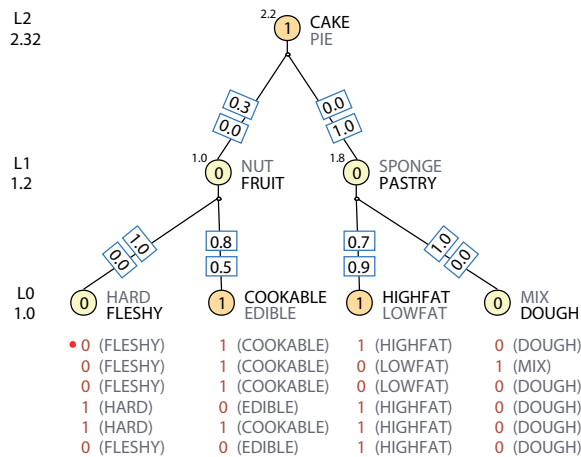


Figure 3: 1-bit fuzzy memory interpreted as a structure of categories.

Combining these two observations, we see how an implicit similarity preference can arise. The distribution-composite of a fuzzy memory divides the reference data into implicit groupings based on variable subsets. Each of these groupings is then implicitly subdivided into two parts by the (binary) evaluations of the contributory variable involved. More memory capacity is obtained if within-group similarity is maximized, since this is the basis for achieving more extreme probabilities. Distribution composites that group by similarity thus yield relatively greater memory capacity. On the assumption that human memory can be formalized under the fuzzy memory model, the disposition to construct similarity-based categories would then be explained in a non-circular way. It would be seen as a strategy for *increasing* memory capacity.

The schematic of Figure 3 illustrates the idea in more intuitive terms. (The corresponding informational assessments appear in Figure 4.) This memory is constructed to represent the reference items shown in the lower, left part of the figure as usual. But all the variables of the memory are here seen as representing categories. The assumed domain is that of food concepts, such as COOKABLE and DOUGH. Each contributory variable has two labels. These name the categories that are represented by the two values of the variable, with the category represented by a value of 1 appearing above the category represented by a value of 0. The value 0 in the leftmost contributory variable at level 1, for example, represents the FRUIT category, while the value 1 in this variable represents NUT.

The distributions in the composite fulfil their key function of modeling the reference items. But notice how they also mediate similarity-based groupings. FRUIT can be seen to represent a group that includes all FLESHY items, some of which are COOKABLE.

Reference items	Cues (9 bits)	Recalled items (13.75 bits)				$r(x,c)$
0 1 0 1	0	0	0.5	0.2	1	1.27
0 1 1 0	0	0	0.5	0.2	1	1.27
0 1 1 1	0	0	0.5	0.2	1	1.27
1 0 1 1	11	1	0.4	0.2	1	0.3
1 1 1 1	11	1	0.4	0.2	1	0.3
1 0 1 0	11	1	0.4	0.2	1	0.3

Memory capacity: 4.75 bits

Figure 4: Graded recall using implicit food categories.

SPONGE represents a group that includes all cases of MIX, most of which are HIGHFAT. These fuzzy groupings then become the means of constructing higher-level categories such as PIE. At this level too, the structure can be viewed as exploiting similarities, although here they reflect the way instances of PIE incorporate (manifestations of) PASTRY and FRUIT.

On this interpretation, the distributions of the composite are akin to the prototype representations envisaged by Rosch and others (e.g. Rosch, 1973; Hampton, 2000). The difference is that here they are deployed at multiple levels of description, and in a way that gets around the problem of prototype compositing (Osherson and Smith, 1981; Prinz and Clark, 2004; Connolly et al., 2007). The construction of similarity-based groupings comes to be seen as a way of increasing the capacity to remember certain data. Exploitation of similarity is explained as a way of increasing the capacity of memory.

## Concluding comment

Inspiration for the idea of graded recall comes from the idea of graded set membership. This is the key concept in fuzzy set theory. What lies behind the operationalization of the idea is more obviously information-theoretic in nature: assessing degree of recall involves measurement of information loss. Fuzzy memory theory thus combines two distinct areas of theoretical analysis into a hybrid framework. This is found to have applications of a cognitive nature when note is taken of the ways the framework can model error-prone memory, and the development of categorical forms of representation.

## References

- Connolly, A. C. , Fodor, J. A. and Gleitman, L. R. (2007). Why stereotypes dont even make good defaults. *Cognition*, 103 (pp. 122).
- Gottwald, S. (2010). An early approach toward graded identity and graded membership in set theory. *Fuzzy Sets and Systems*, 161, No. 18 (pp. 23692379).
- Hahn, U. and Chater, N. (1997). Concepts and similarity. In Lamberts and Shanks (Eds.), *Knowledge, Concepts and Categories* (pp. 43-92). Cambridge, Mass: The MIT Press.



- Hampton, J. A. (2000). Concepts and prototypes. *Mind and Language*, 15 (pp. 299-307).
- Harnad, S. (2005). To cognize is to categorize: cognition is categorization. In Cohen and Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science* (pp. 19-43). Elsevier.
- Machery, E. (2009). *Doing without Concepts*. Oxford: Oxford University Press.
- Murphy, G. L. (2002). *The Big Book of Concepts*. London, England: The MIT Press.
- Osherson, D. N. and Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition: International Journal of Cognitive Psychology*, 9 (pp. 35-58).
- Osherson, D. N. and Smith, E. E. (1982). Gradedness and conceptual combination. *Cognition*, 12 (pp. 299-318).
- Prinz, J. J. and Clark, A. (2004). Putting concepts to work: some thoughts for the 21st century. *Mind And Language*, 19 (pp. 57-69).
- Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In Moore (Ed.), *Cognitive Development and the Acquisition of Language*. Academic Press.
- Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana, Illinois: University of Illinois Press.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27 (pp. 379-423 and 623-656).
- Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8 (pp. 338-353).
- Zadeh, L. (1976). A fuzzy-algorithmic approach to the definition of complex or imprecise concepts. *International Journal of Man-Machine Studies*, 8 (pp. 249-291).
- Zadeh, L. (1982). A note on prototype theory and fuzzy sets. *Cognition*, 12, No. 3 (pp. 291-298).

# From Head to Toe: Embodiment Through Statistical Linguistic Frequencies

**Richard Tillman (r.tillman@memphis.edu)**

Department of Psychology / Institute for Intelligent Systems, University of Memphis  
400 Innovation Drive, Memphis, TN 38152 USA

**Vivek Datla (vvdatta@memphis.edu)**

Department of Computer Science / Institute for Intelligent Systems, University of Memphis  
400 Innovation Drive, Memphis, TN 38152 USA

**Sterling Hutchinson (schthns@memphis.edu)**

Department of Psychology / Institute for Intelligent Systems, University of Memphis  
400 Innovation Drive, Memphis, TN 38152 USA

**Max Louwerse (mlouwerse@memphis.edu)**

Department of Psychology / Institute for Intelligent Systems, University of Memphis  
400 Innovation Drive, Memphis, TN 38152 USA

## Abstract

Recent literature in the cognitive sciences has demonstrated that cognition is fundamentally embodied. For instance, various studies have shown that semantic knowledge about the human body correlates with spatial body representations, suggesting that such knowledge is embodied in nature. An alternative explanation for this finding comes from the Symbol Interdependency Hypothesis, which argues that perceptual information is encoded in language. We demonstrated that the findings that can be explained by an embodied cognition account can also be explained through statistical linguistic frequencies. Co-occurrence frequencies of names for common body parts correlated with experimental findings from adults and children. Moreover, the position of the body parts was predicted on the basis of statistical linguistic frequencies. These findings suggest that language encodes embodied information.

**Keywords:** embodiment; statistical linguistic frequencies; symbolic cognition; embodied cognition; conceptual processing; symbol interdependency.

## Introduction

Over the last decade the notion that cognition is fundamentally embodied has dominated the cognitive sciences (Glenberg, 1997; Goldstone, & Barsalou, 1998; Barsalou, 1999; Lakoff & Johnson, 1999; Zwaan, 2004; Pecher & Zwaan, 2005; Semin & Smith, 2008). The central argument in theories of embodied cognition is that our minds co-evolved with our bodies, especially the sensory motor system, and that cognitive processes therefore heavily rely on perceptual simulations. This argument is in sharp contrast with a computational symbolic approach to cognition. Views of symbolic cognition suggest that meaning can be derived from linguistic context (Landauer & Dumais, 1997). In other words, instead of mental reenactment, mental representations can be seen as internal

structures of symbolic concepts and do not necessarily have a direct relation to perceptual states (Fodor, 1975; Pylyshyn, 1984).

There is a large body of literature that finds evidence that cognition is embodied. Studies have shown that processing within modalities is faster than having to map across modalities (e.g., Marques, 2006; Pecher, Zeelenberg & Barsalou, 2003; Spence, Nicholls & Driver, 2000). Language comprehension seems to be influenced by action representations primed in experimental tasks (e.g., McCloskey, Klatzky, & Pellegrino, 1992; Zwaan, Stanfield & Yaxley, 2002), and visual representations get activated during language comprehension. Perceptual feature characteristics that have affected language comprehension include orientation (Stanfield & Zwaan, 2001), temporality (Zwaan, Madden & Whitten, 2000), visibility (Rapp & Horton, 2003), spatial configuration (Louwerse, 2008; Zwaan & Yaxley, 2003), modality (Louwerse & Connell, 2011; van Dantzig et al., 2008), direction (Glenberg & Kaschak, 2002; Kaschak et al., 2005), or location (Šetić & Domijan, 2007).

Several embodied cognition studies have shown a relation between the meaning of words and their spatial configuration when presented on the screen. For instance, when words for concepts in the air, such as birds and insects, are presented in the upper half of a screen, participants respond faster than when the same words are presented in the bottom of the screen, with a reverse effect for words referring to concepts on land or in the ocean (Šetić & Domijan, 2007; Pecher, Van Dantzig, Boot, Zanolie, & Huber, 2010). Similarly, when word pairs such as *attic* and *basement* are presented vertically, one above the other, iconic pairs are processed faster than reverse iconic pairs, presumably because comprehenders perceptually simulate the position of these concepts (Zwaan & Yaxley, 2003).

Other studies have demonstrated that the vertical configuration of words on the screen and the meaning of those words can be extended to concepts we literally embody, such as body parts. For instance, understanding parts of our body is directly linked to the spatial representation of the human body, and that representation contains veridical information about the relative distance between body parts (Smeets et al., 2009; Struiksmā, Noordzij, & Postma, 2011; Van Elk & Blanke, 2011). When participants were presented with combinations of concepts that represent body parts, such as *head-neck*, processing time was considerably faster when the embodied distance of those concepts was small, compared to concepts for which the distance is large, such as *head-toe*. Studies like these yet again show that embodiment explains cognition.

However, the question can be raised as to what extent the relation between body semantics and spatial body representations can only be explained by an embodied cognition account. This is an important question, particularly if other accounts are complementary to the embodied cognition account.

We have argued for one such account in a number of studies. The Symbol Interdependency Hypothesis argues that language comprehension is both perceptual and linguistic in nature (Louwerse, 2008, 2011; Louwerse & Connell, 2011; Louwerse & Jeuniaux, 2010). That is, language comprehension is linguistic through statistical interdependencies between linguistic units and is perceptual through the references linguistic units make to perceptual representations. The Symbol Interdependency Hypothesis thereby makes an important prediction: language has evolved to become a communicative shortcut for language users and it encodes relations in the world. Accordingly, it is hypothesized that the findings attributed to an embodied cognition account can also be explained through statistical linguistic frequencies.

In a number of studies, we have shown that language indeed encodes perceptual information. Louwerse, Cai, Hu, Ventura, and Jeuniaux (2006) and Louwerse and Zwaan (2009) aimed to determine if language encodes geographical information by comparing city latitude/longitude with how often those cities appeared in a corpus. Louwerse, Cai, and Hutchinson (in press) have shown that these predictions are not limited to English, but can also be found in Chinese (predicting cities in China) and Arabic (predicting cities in the Middle East). Louwerse and Benesh (in press) have recently shown how using the *Lord of the Rings* trilogy the longitude and latitude for cities in the fictional Middle Earth can be predicted. The physical distance between cities was accurately estimated based upon statistical linguistic frequencies of cities, thus suggesting that language does encode (perceptual) geographical information.

The encoding of perceptual information in language goes well beyond geography. Louwerse and Connell (2011) have shown that the modality of a word (e.g., *sour*, *soft*, *loud*) can be predicted on the basis of statistical linguistic frequencies. That is, computational estimates on the modality of a word were less precise (visual/tactile, olfactory/taste, auditory) but equally as accurate as human estimates on the modality of words.

In addition to geographical predictions and modality predictions, Louwerse (2008) investigated whether iconicity of words can be predicted. Analogous to binomials such as *top and bottom*, *high and low*, and *up and down*, this study found that the iconic order of concepts such as *flower-stem* could indeed be predicted by simply looking at the order of the words.

It is relevant here to address the question whether these statistical linguistic cues are in fact used by comprehenders. Louwerse (2008) tested whether word pairs like *flower-stem*, presented vertically, yielded faster response times because participants were perceptually simulating the word pairs, or because of the word order (a linguistic factor). The findings demonstrated that the frequency of word pairs such as *flower-stem* (a perceptually realistic order) is significantly higher than word pairs *stem-flower* (a perceptually unrealistic order), and that linguistic frequencies explained response times at least as well as perceptual ratings.

The effect of perceptual and linguistic factors on cognitive processes is modulated by stimulus, cognitive task, and by duration of processing. Louwerse and Jeuniaux (2010) showed that linguistic factors best explained semantic judgments of word pairs, whereas perceptual factors best explained iconicity judgments of picture pairs. Furthermore, they concluded that linguistic factors dominated when participants were involved in shallow cognitive processes, and that perceptual factors dominated in deeper cognitive tasks. Louwerse and Connell (2011) extended these findings, showing that faster response times were best explained by linguistic factors, and slower response times were best explained by perceptual factors. These findings suggest that the relative employment of linguistic or perceptual representations changed as a function of the task, duration of the task, or stimulus.

In the following study, we determined whether embodied information – information about the distance between body parts – was also encoded in language. To test for this possibility, we conducted a computational linguistic study in which we calculated the co-occurrence of body part names and compared the statistical linguistic frequencies with the existing experimental data. We thereby hypothesized that body parts that are perceptually close together are placed in similar linguistic contexts, thereby allowing for accurate computational estimates on the position of the body part.

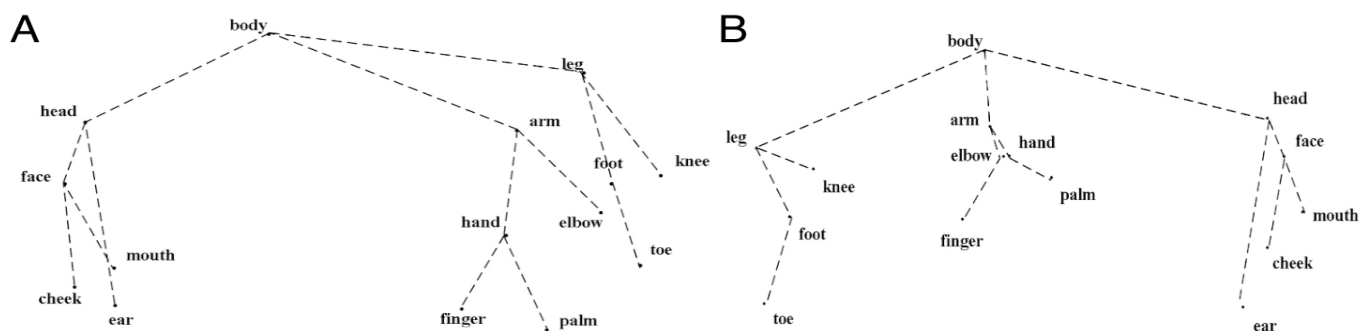


Figure 1. Body part similarity ratings of adults (A) and children (B) (Jacobowitz, 1973).

### Study 1

In previous research, Jacobowitz (1973) explored the development of language by comparing body part similarity ratings of five-year-old children, and adults. The 15 body parts used were: *Arm, body, cheek, ear, elbow, face, finger, foot, head, hand, knee, leg, mouth, palm, and toe*. Jacobowitz conducted four replicated multi-dimensional scaling analyses (RMDS), which simultaneously analyzed multiple matrices. The dimensional scaling illustrated that the five-year-olds grouped the head items, arm items, and leg items more similarly (see Figure 1B). The adults, on the other hand, grouped head terms together, but the other extremities were grouped by function (e.g., arm and leg (limbs) were more similar, finger and toe (digits) were more similar, etc.) (see Figure 1A). Jacobowitz found that the similarity ratings for body parts were hierarchical for both the children and adults.

In the current study, Jacobowitz's (1973) data was compared with findings from statistical linguistic frequencies. We calculated the frequency of first-order co-occurrences in the *Web 1T 5-gram* corpus (Brants & Franz, 2006). This corpus consists of one trillion word tokens (13,588,391 word types) from 95,119,665,584 sentences. The volume of the corpus allows for an extensive analysis of patterns in the English language. The frequency of co-occurrences of the 15 words was computed in bigrams, trigrams, 4-grams and 5-grams. For instance, the frequency of the words {*head, toe*} was determined by considering these words next to one another {*head, toe*}, with one word in between {*head w1 toe*}, with two {*head w1 w2 toe*} or with three intervening words {*head w1 w2 w3 toe*}. This method is identical to the one used in Louwerse (2008), Louwerse and Jeuniaux (2010), and Louwerse and Connell (2011).

The result of these computations was a 15 x 15 matrix of raw frequencies of co-occurrences, from which log frequencies were obtained. This matrix was submitted to an MDS analysis using the ALSCAL algorithm (see Young, Takane, & Lewyckij, 1978). For purposes of mapping the relative location of body parts, it is insufficient to simply obtain the co-occurrence frequencies in the Google corpus. The frequencies must be converted to x and y coordinates, and then a mathematical analysis performed to find the relative spatial location of the body parts. Multidimensional scaling (MDS) is a series of mathematical operations that can illuminate patterns within data that may not be

immediately recognizable with standard numerical output (Kruskal, & Wish, 1977; Blake, Schulze, & Hughes, 2003). MDS has been utilized to not only analyze similarity, but also to provide a graphical representation of those similarities. A Euclidean distance measure transformed the semantic similarities into dissimilarities, such that the higher the value, the longer the distance. Default MDS criteria were used with an S-stress convergence of .001, a minimum stress value of .005, and a maximum of 30 iterations. The fitting on a two-dimensional scale was moderate, with a Stress value = .21 and an  $R^2 = .86$ .

To do justice to the geometry of the 2D variables in Jacobowitz (1973), we used bidimensional regression analyses to compare the participants' estimates with the actual coordinates of the body parts. Tobler (1994) and Friedman and Kohler (2003) introduced bidimensional regressions in order to compute the mapping of any two planes under consideration. Whereas in a unidimensional regression each data point is shifted by intercept and slope, each actual and predicted value of the dependent variable are presented by a point in space, whereby vectors represent intercept and slope.

A bidimensional regression yielded a significant correlation between the frequency estimates and Jacobowitz's (1973) loadings on a two-dimensional plane for both the adult study,  $r = .66$ ,  $p < .01$ ,  $n = 15$ , and the child study,  $r = .63$ ,  $p = .01$ ,  $n = 15$ . To ascertain that these findings could not be attributed to accidental pairings of coordinates, we conducted a Monte Carlo simulation, randomly sampling each dataset 1000 times. The findings solidified the results, with no bidimensional relation between the statistical linguistic frequencies and Jacobowitz's (1973) adult data, average  $r = .23$  ( $SD = .12$ ),  $n = 15$  or child data, average  $r = .24$  ( $SD = .12$ ),  $n = 15$ . These findings suggest that statistical linguistic frequencies can explain data obtained from human participants.

In addition to the comparison between Jacobowitz's (1973) two-dimensional fitting, we compared a one-dimensional solution, using the first dimension of the MDS solution, with the location of the body part terms. The correlation between the location of the body part words and the computational estimates was again high,  $r = .6$ ,  $p < .001$ ,  $n = 15$ . The linear fitting between the computational estimates and the actual position is presented in Figure 2.

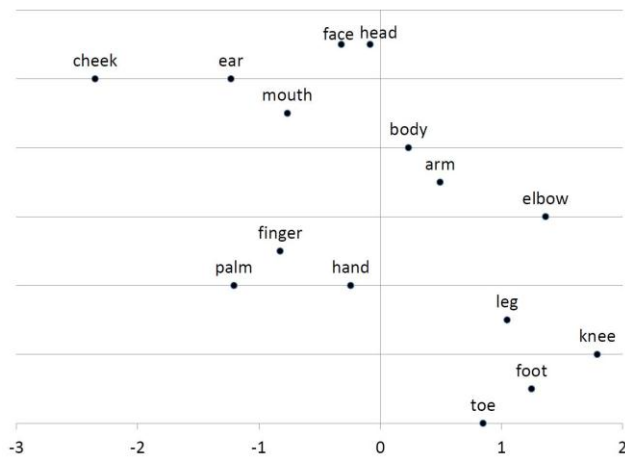


Figure 2. Multidimensional Scaling of 15 body parts from Jacobowitz (1973).

## Study 2

Van Elk and Blanke (2011) established a relationship for spatial position of body parts as well as the relative distance between them for native French speakers (see Table 1). In experiment 1, 38 body parts were assigned to nine categories dependent upon the distance from each other on the body (e.g., forehead/toe = 9; forehead/eye = 1). The words were then presented vertically in a congruent or incongruent spatial position (forehead/toe; toe/forehead). Subjects demonstrated increased RTs for larger distances, while position congruency did not seem to have an effect. Experiment 2 consisted of an iconicity judgment also using relative distance and congruency. However, in this experiment the words were not in the center of the screen as in Experiment 1, but arranged in varying distances from each other. There were significant main effects, as well as

an interaction, for the error rates. The RTs revealed there were main effects of congruency and distance, but no interaction was found.

We computed the log frequency of all combinations of the English body part words and compared the data with the Van Elk and Blanke (2011) distance data. Because the algorithm functions best with single words in a 2-5 gram window, we removed all words that require two words in English (*under arm*, *ring finger*, *index finger*, and *middle finger*). Moreover, no frequencies were found for *instep* and *pinkie* combinations, these words were removed from the analysis.

The correlation of the 32 x 32 word pair frequencies and the distances was significant,  $r = .35$ ,  $p < .001$ ,  $n = 1024$ , with higher frequencies yielding lower physical distances. This finding suggests that embodiment is encoded in language, such that the relative location of body parts can be estimated using statistical linguistic frequencies.

Next, we conducted analyses similar to the first study, whereby we did not use the raw frequency comparisons, but instead entered the  $n \times n$  matrix in an MDS algorithm and used the loadings of the body parts names as a comparison. To do justice to the one-dimensional plane Van Elk and Blanke (2011) used, the MDS solution was restricted to a one-dimensional solution. The fitting was moderate,  $Stress = .47$ ,  $R^2 = .50$ . When the loadings of the 32 body parts were compared with their physical distances, a strong correlation was found,  $r = -.76$ ,  $p < .001$ ,  $n = 32$ .

To determine whether these findings could in any way be attributed to accidental pairings of variables, we again conducted a Monte Carlo simulation, whereby correlations of the 1000 randomizations of the data were computed. The average correlation did not come close to the correlation obtained for the actual data, average  $r = .15$ ,  $p = .41$ ,  $n = 1000$ . As before, we plotted the position of the body parts and their corresponding words (Figure 3).

Table 1: Body part positions and categories (Van Elk & Blanke, 2011).

Word	Position	Loading	Cat.	Word	Position	Loading	Cat.	Word	Position	Loading	Cat.
hair	1	0.79	1	back	3	-1.15	10	palm	6	1.05	13
eye	1	1.01	4	shoulder	3	-0.42	9	thigh	7	-0.83	17
ear	1	1.13	4	chest	3	0.31	10	leg	7	-0.56	18
forehead	1	1.35	2	elbow	4	-0.81	11	knee	8	-0.95	19
eyebrow	1	2.09	3	wrist	5	-0.82	13	calf	8	-0.88	20
neck	2	-0.14	8	forearm	5	-0.75	12	ankle	9	-1.36	21
throat	2	0.91	8	butt	5	-0.58	16	shin	9	-1.24	20
chin	2	0.92	7	thumb	5	0.13	14	heel	9	-1.09	22
nose	2	1.24	5	stomach	5	0.86	15	foot	10	-0.9	21
lip	2	1.35	6	hand	6	-0.89	13	toe	10	-0.61	23
cheek	2	1.56	5	hip	6	-0.7	16	palm	6	1.05	13

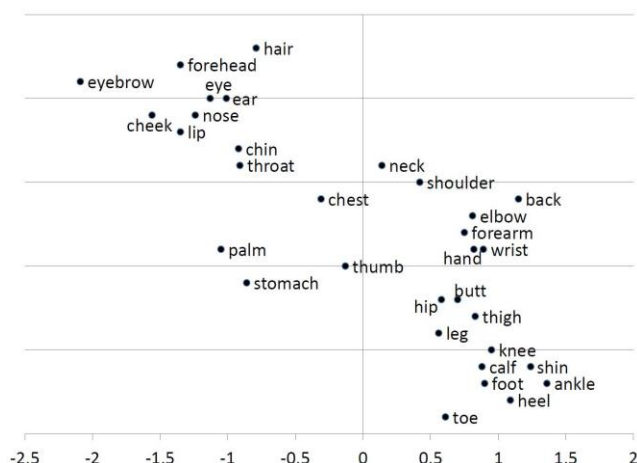


Figure 3. Multidimensional Scaling of 32 body parts (Van Elk & Blanke, 2011).

## Discussion

Recent literature has shown that perceptual information, such as geographical locations, modalities, and iconicity, is encoded in language. The current paper extended these findings by addressing the question whether language encodes (literally) embodied information: whether statistical linguistic frequencies can explain the relative location of different parts of the body. Results from two computational studies showed that such frequencies indeed can estimate the relative location of body parts. In study 1, we demonstrated that computationally derived values can explain human similarity estimates of body-parts. Study 2 similarly found that word frequencies can estimate physical distances between body parts. Both of these studies support the claim that language encodes body information.

We conclude that language inherently contains body part information, such that experimental results can be approximated computationally. This is in line with previous research that has demonstrated that language encodes geographical information (Louwerse & Zwaan, 2009), that it encodes modality specific information (Louwerse & Connell, 2011), spatial information (Louwerse, 2008), and social relations (Hutchinson, Datla, & Louwerse, in press). The current study adds to this literature and suggests that cognition is indeed both embodied and symbolic.

## References

- Barsalou, L. (1999). Perceptual symbol systems. *Behavior and Brain Sciences*, 22, 557-660.
- Blake, B., Schulze, S., & Hughes, J. (2003). *Perceptual mapping by multidimensional scaling: A step by step primer*. Research Reports in Consumer Behavior. Cleveland, OH: Cleveland State University.
- Brants, T., & Franz, A. (2006). *Web 1T 5-gram version 1*. Philadelphia: Linguistic Data Consortium.
- Fodor, J. (1975). *The Language of Thought*. New York, NY: Crowell.
- Friedman, A. & Kohler, B. (2003). Bidimensional regression: A method for assessing the configural similarity of cognitive maps and other two-dimensional data. *Psychological Methods*, 8, 468-491.
- Glenberg, A. (1997). What memory is for. *Behavior and Brain Sciences*, 20, 1-55.
- Glenberg, A., & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin and Review*, 9, 558-565.
- Goldstone, R., & Barsalou, L. (1998). Reuniting perception and conception. *Cognition*, 65, 231-262.
- Hutchinson, S., Datla, V., & Louwerse, M. M. (in press). Social networks are encoded in language. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Jacobowitz, D. (1973). *Development of semantic structures*. Unpublished dissertation. University of North Carolina-Chapel Hill, NC.
- Kaschak, M., Madden, C., Theriault, D., Yaxley, R., Aveyard, M., Blanchard, A., & Zwaan, R. (2005). Perception of motion affects language processing. *Cognition*, 94, B79-89.
- McCloskey, B., Klatzky, R. L., & Pellegrino, J. (1992). On rubbing your stomach while tapping your fingers: Interference between motor planning and semantic judgments. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 948-961.
- Kruskal, J. & Wish, M. (1977). *Multidimensional Scaling*. Beverly Hills, CA: Sage Publications.
- Lakoff, G. & Johnson, M. (1999) *Philosophy in the Flesh*. New York: Basic Books.
- Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
- Louwerse, M. (2011). Stormy seas and cloudy skies: conceptual processing is (still) linguistic and perceptual. *Frontiers in Psychology: Cognition*, 2, 1-4.
- Louwerse, M. (2008). Embodied representations are encoded in language. *Psychonomic Bulletin and Review*, 15, 838-844.
- Louwerse, M. & Benesh, N. (in press). Representing spatial structure through maps and language: Lord of the Rings encodes the spatial structure of Middle Earth. *Cognitive Science*.
- Louwerse, M., Cai, Z., Hu, X., Ventura, M., & Jeuniaux, P. (2006). Cognitively inspired natural language based knowledge representations: Further explorations of Latent Semantic Analysis. *International Journal of Artificial Intelligence Tools*, 15, 1021-1039.
- Louwerse, M., Cai, Z., & Hutchinson, S. (in press). The Chinese route argument: Predicting the longitude and latitude of cities in China and the Middle East using statistical linguistic frequencies. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Louwerse, M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the

- modality of words. *Cognitive Science*, 35, 381-398.
- Louwerse, M. & Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. In M. de Vega, A. Glenberg, & C. Graesser (Eds.). *Symbols, embodiment, and meaning* (pp. 309-326). Oxford University Press: New York, NY.
- Louwerse, M. & Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. *Cognition*, 114, 96-104.
- Louwerse, M. & Zwaan, R., (2009). Language encodes geographical information. *Cognitive Science*, 33, 51-73.
- Marques, J. (2006). Specialization and semantic organization: Evidence for multiple semantics linked to sensory modalities. *Memory & Cognition*, 34, 60-67.
- Pecher, D., Van Dantzig, S., Boot, I., Zanolie, K., & Huber, D. E. (2010). Congruency between word position and meaning is caused by task induced spatial attention. *Frontiers in Cognition*, 1, 1-8.
- Pecher, D., van Dantzig, S., Zwaan, R., & Zeelenberg, R. (2009). Language comprehenders retain implied shape and orientation of objects. *The Quarterly Journal of Experimental Psychology*, 62(6), 1108-14.
- Pecher, D., & Zwaan, R. (Eds.). (2005). *Grounding cognition: The role of perception and action in memory, language, and thinking*. New York: Cambridge University Press.
- Pylyshyn, Z. (1984). *Computation and cognition: Towards a foundation for cognitive science*. Cambridge, MA: MIT Press.
- Horton, W. & Rapp, D. (2003). Out of sight, out of mind: Occlusion and the accessibility of information in narrative comprehension. *Psychonomic Bulletin & Review*, 10, 104-110.
- Semin, G., & Smith, E. (Eds.). (2008). *Embodied grounding: Social, cognitive, affective, and neuroscientific approaches*. New York, NY: Cambridge University Press.
- Šetić, M., & Domijan, D. (2007). The influence of vertical spatial orientation on property verification. *Language and Cognitive Processes*, 22, 297-312.
- Smeets, M., Klugkist, I., van Rooden, S., Anema, H., & Postma, A. (2009). Mental body distance comparison: a tool for assessing clinical disturbances in visual body image. *Acta Psychologica*, 32, 157-165.
- Spence, C., Nicholls, M., & Driver, J. (2001). The cost of expecting events in the wrong sensory modality. *Perception & Psychophysics*, 63, 330-336.
- Stanfield, R., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, 12, 153-156.
- Struiksmā, M., Noordzij, M., & Postma, A. (2011). Reference frame acceptability in haptics differs for the blind and sighted in the horizontal but not in the vertical plane. *Perception*, 40, 725-738.
- Tobler, W. (1994). Bidimensional regression. *Geographical Analysis*, 26, 186-212.
- Van Dantzig, S., Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2008). Perceptual processing affects conceptual processing. *Cognitive Science*, 32, 579-590.
- Van Elk, M. & Blanke, O. (2011). The relation between body semantics and spatial body representations. *Acta Psychologica*, 138, 347-358.
- Young, F., Takane, Y., & Lewyckij, R. (1978). ALSCAL: A nonmetric multidimensional scaling program with several different options. *Behavioral Research Methods and Instrumentation*, 10, 451-453.
- Zwaan R. A. (2004). The immersed experience: toward an embodied theory of language comprehension. In B. H. Ross, Ed. *The Psychology of Language and Motivation*, 44. New York, NY: Academic Press.
- Zwaan R., Madden C., & Whitten S. (2000) The presence of an event in the narrated situation affects its activation. *Memory and Cognition*, 28, 1022-28
- Zwaan, R., Stanfield, R., & Yaxley, R. (2002). Language Comprehenders Mentally Represent the Shapes of Objects. *Psychological Science*, 13(2), 168-171.
- Zwaan, R., & Yaxley, R. (2003). Spatial Iconicity Affects Semantic Relatedness Judgments. *Psychonomic Bulletin and Review*, 10(4), 954-8.



# Evidence for Modality-Specific Processes in Approximate Numerical Comparison

Midori Tokita (tokita.midori@ocha.ac.jp)  
Akira Ishiguchi (ishiguchi.akira@ocha.ac.jp)  
Ochanomizu University, Otsuka Bunkyo-ku,  
Tokyo, 112-8610 Japan

## Abstract

It has been claimed that a genuinely abstract number representation exists and is capable of representing the numerosity of any set of discrete elements irrespective of whether they are presented in visual or auditory modality. To test whether adults can compare large numerosities cross-modally as accurately as intra-modally, we measured Weber fractions and a point of subjective equality of numerical discrimination in the visual, auditory, and cross-modal conditions with use of a carefully controlled experimental procedure. Results showed distinct differences between the performances of the visual and the auditory condition in such way that numerical discrimination of the auditory sequence is more precise than that of visual sequence. Moreover, the performance of cross-modal trials differed among participants, with the exception that they were all worse than the auditory condition and that the number of visual stimuli was overestimated. Taken together, our findings implied that numerical discrimination of the auditory and visual stimuli mediates the modality-specific processes, suggesting that the numerical representation process can be complex of multiple stages.

**Keywords:** numerical discrimination; sensory modality; cross-modal comparison

## Introduction

Many studies supported the idea that humans possess innate neural mechanisms that generate approximate, not precise, numerical representations. Results from studies of numerical competence in infants, young children, and nonhuman animals have shown that the approximate numerical system is evolutionally old and is equipped early in human development (e.g., Cantlon & Brannon, 2006; Feigenson, Dehaene & Spelke, 2004; Hauser, Tsao, Garcia, & Spelke, 2003; Whalen, Gallistel & Gelman, 1999). Furthermore, converging empirical findings from several areas of cognitive neuroscience argue for biological determined mechanisms for approximate number representation (e.g., Nieder & Dehaene, 2009; Piazza, 2010). At the same time, certain researchers have prompted extensive investigation over the processes of number representation in the behavioral and neurophysical field (e.g., Kadosh, Lammertyn, & Izard 2008; Kadosh & Walsh, 2009).

One of the claims made by the proponents of abstract numerical representation is that the processing of approximate numerical representation is independent of

sensory modality. They argued that abstract numerical representation could genuinely be capable of representing the numerical of any set of discrete elements, whether they were presented in the visual or auditory condition (Barth, Kanwisher, & Spelke, 2003; Gallistel & Gelman, 1992; Jordan & Brannon, 2006; Piazza, 2010). In these studies, it has been demonstrated that there was no cost of comparing numerosities across versus within visual and auditory stimulus sets. They claimed that the comparison across presentation modality was not performed using modality-specific numerical representations but rather using the true abstract numerical representation system. Evidence for modality-independent numerical representation ability has also been claimed in infants (e.g., Jordan & Brannon, 2006; Kobayashi, Hiraki & Hasegawa, 2005) and animals (Jordan, Brannon, Logothetis & Ghazanfar, 2005).

It has, however, remained unclear whether these approximate numerical representations are truly modality-independent. Three primary reasons exist for doubting the modality-independence of the approximate numerical representation. First, some evidence has shown that there were significant differences in the performance of numerical judgments for visual, auditory, and tactile senses (e.g., Riggs, Ferrand, Lancelin, Fryziel, & Dumur, 2006; Lechelt, 1975; Philippia, van Erp, & Werkhoven, 2008). For example, in the rapid counting experiment, Lechelt (1975) compared adult performance in numerosity judgment of visual, auditory, and tactile stimuli and demonstrated that perceived numerosity differed among modalities. Philippi, van Erp, & Werkhoven (2008) demonstrated that the stimuli with a short interstimulus interval (ISI) are underestimated and the tendency is stronger for visual than for auditory stimuli. Second, it is known that the processing of temporal information is much more efficient in the auditory than in the visual modality (Penny, Gibbon, & Meck, 2000; Ivry, 2008). For example, in time related tasks such as duration discrimination and empty interval estimation, the performance in the auditory presentation is significantly better than that in the visual and the tactile presentations (e.g., Grondin, 2010). As it has suggested that the temporal information affects the numerical discrimination (Tokita & Ishiguchi, 2011), there is the possibility that numerical discrimination among modalities differed when the experimental conditions are rigorously controlled. Third, limitations may exist within experimental procedures of empirical studies that claimed modality-independence of numerical representation in terms of control of stimuli,

precision in measurement, and numbers of items tested. For example, Barth et al (2003) used a cross-modal comparison task and found that accuracy on these tasks was comparable to those on intramodal tasks, suggesting that non-numerical cues did not play a substantial role even in intramodal tasks. Numerical contrasts in their studies were, however, quite large such as Weber fraction of .50 or greater. With this level of measurement precision, the difference in the performance of each task could remain undetected. More to it, in infant and animal studies, the number of items tested was smaller than four. Because it remains unclear whether a system for representing small numbers of objects is distinct from that for representing larger numbers of objects, it is necessary to test whether the effects of sensory modality differ among a variety of numerosities.

In this study, we tested whether and how the numerical comparison of visual, auditory, and cross-modal presentation would differ under the adequate control of the concerns discussed above. We measured Weber fractions of discrimination task to assess the difference in the precision. Many studies have shown that both behavioral and neuronal tuning functions obey the Weber law (i.e., discriminability depends on the ratio of the numerical to be compared) over a broad range of numerosities (e.g., Burgess & Barlow, 1983; Nieder & Dehaene, 2009; Tokita & Ishiguchi, 2011; Whalen, Gallistel, & Gelman, 1999). We also measured a point of subjective equality (PSE) to test the accuracy of numerical comparison. Importantly, we involved rigid stimuli controls so that other properties such as stimuli duration and interval duration would not be confounded with the number of elements.

In Experiment 1, we compared the performance of numerical discrimination between the visual and the auditory presentation. In Experiment 2, we compared the performance of the visual, auditory, and cross-modal numerical comparison to examine how the numerical information in the different modality may integrate.

## Experiment 1

We examined the precision of approximate numerical comparison in two sensory modalities: visual and auditory. The schematic view of stimuli presentation is shown in Figure 1.

In a visual condition, elements in a set were consisted of sequences of flashes, while in an auditory condition, elements in a set were presented in a tone sequence. We employed two levels of standard event numbers (i.e., standard number), 10 and 20, to test whether and how precision across presentation conditions would differ among standard numbers. To examine the precision, we obtained Weber fractions that indicate the participant's variance of numerical comparison. In deriving the Weber fractions, we used the method of constant stimuli in which participants in each trial decided which stimuli—standard or comparison—had more events.

## Method

**Participants** Five participants participated in the experiment. All had normal or corrected-to-normal hearing and vision. All participants had no prior experience in numerical comparison tasks.

**Design** Two independent variables were examined in the experiment: the sensory modality (i.e., visual and auditory) and standard number (i.e., 10 and 20). The numbers in the comparison stimuli for the standard number of 10 and 20 were “8, 9, 11, 12” and “16, 18, 22, 24”, respectively. Trials in the visual and auditory conditions were separated and each constituted trial blocks. Two experimental conditions were presented among participants in a counterbalanced order. Trials in all the standard number sets were intermixed within a block. Each condition had 320 trials (40 repetitions  $\times$  4 comparison levels  $\times$  2 standard numbers), resulting in 640 trials in total. Each block had 64 trials, with 10 blocks in total. Participants performed three or four blocks in each experimental session, which took three days in total. Intermissions of approximately three minutes were given between blocks. Sequence of the trials was completely randomized within a block. Standard stimuli came first in half the trials and second in the remaining ones. Participants were given 16 practice trials before the actual experiment began.

**Stimuli** In the visual condition, two sequences of light gray dots appeared in a dark gray display region. Luminance of the dot was approximately 8 cd/m<sup>2</sup>. In the auditory condition, two sequences of tone were presented with the built-in-speaker of the desktop computer at the intensity of about 60 dB (Sound pressure level). Auditory stimuli were 700 Hz pure sinusoidal sounds generated by Macintosh's computer.

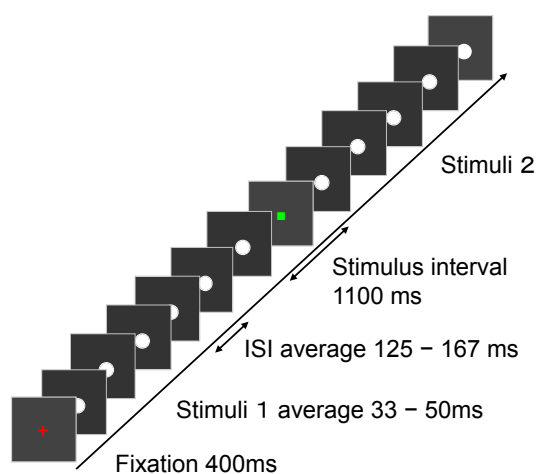


Figure 1: Schematic view of stimuli used in Experiment 1 and 2. The pair of events was sequentially presented in random orders.

In both conditions, we carefully controlled the stimulus duration and the inter-stimulus interval (ISI) so that the time for a sequence and presentation rate of stimuli would not be confounded with the number of elements. All element in a particular sequence had the same duration, but the durations varied from sequence to sequence between 33 to 50 ms. In half of the trials in a block, the average ISI was 125 ms in both standard and comparison sequences. In the remaining half, average ISI in the comparison sequences were carefully controlled so that average total interval for the standard sequence and that for the comparison sequence would be approximately equal. Thus, the number of events would be the only cue for numerical judgments. Many studies have provided evidence that the minimum ISI between two successive stimuli for correctly reporting their temporal order is about 40 ms and that this order threshold is invariant for auditory visual stimuli (e.g., Poppel, 1997; Kanabus, Szélag, & Poppel, 2002). Thus, sets of events in this experiment were perceived as successive independent of the sensory modality. To make the sequence aperiodic, we randomly added temporal jitter (−24, −17, −8, 0, 8, 17, or 24 ms) to each ISI so that the temporal rate would not constitute a rhythmic pattern.

Importantly, ISI were carefully determined so that the participants would not make judgments based on the verbal counting and/or temporal patterns. To make verbal counting impossible, the longest stimulus interval was set to be less than 250 ms, as previous studies have proved that participants could not rely on verbal or sub-verbal counting within that duration (e.g., Piazza, Mechelli, Price & Butterworth, 2006; Tokita & Ishiguchi, 2011).

**Measurements** The PSE and Weber fractions were measured using the method of constant stimuli. First, the number of events in comparison stimuli was plotted on the x axis and the proportion of “greater” response for each comparison stimulus was plotted on the y axis. The plotted data points constructed the psychometric function approximated by a cumulative Gaussian function, on which the difference threshold was obtained. This difference threshold was defined as the smallest amount of the element number change, for which a correct response rate of 75% was achieved. Weber fractions were obtained by dividing the difference thresholds by the standard numbers. The PSEs were obtained as the value of the location on the psychometric function at which the standard and comparative choice probabilities were equal to 50%. In this experiment, we obtained the standardized PSE, dividing the PSE by the standard number.

**Procedure** Participants sat in a darkened room at a distance of approximately 115 cm from the presentation screen. A numeric keypad was placed directly in front of the participants. The participants made responses by pressing the “1” or “3” key.

Each trial started with a red fixation cross for 400 ms followed by the first sequence. Pairs of sequences—standard and comparison sequences—were shown in

succession in random order. The two sequences were separated by a stimulus interval of 1100 ms.

The participant’s task was to choose which sequence, the first or second, contained more elements. Feedback with a short beep sound was given when participants made an incorrect choice. At the beginning of each session, the participants were explicitly instructed to attend to the number of elements presented and to discriminate on the basis of the numerical they felt, and not by verbal counting. They were also instructed to see the center of monitor in the auditory condition as well.

A Macintosh G4 computer was used to generate the display and the sound, and to record the data. Stimuli were presented on a color monitor at a refresh rate of 120 Hz (SONY Color Graphic Display Model GDM-F400).

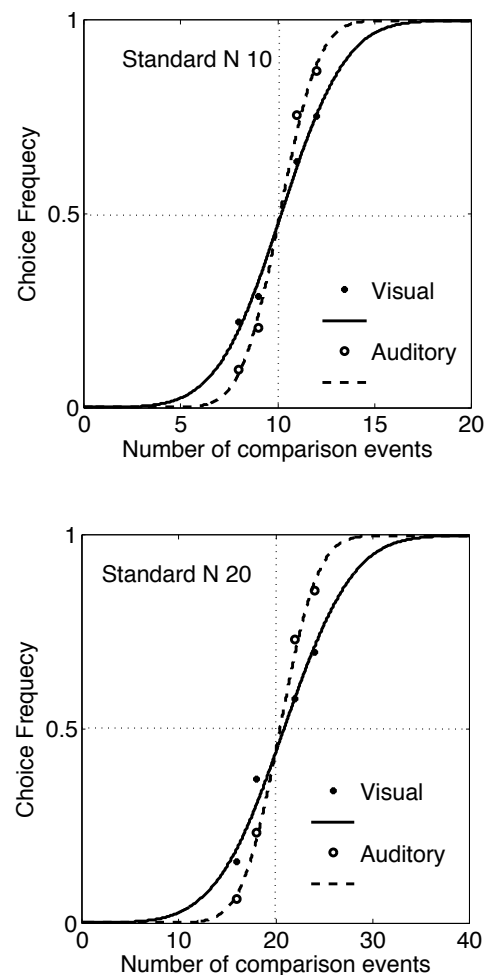


Figure 2: Average psychometric functions for the each presentation condition (a) standard number of 10 and (b) standard number of 20.

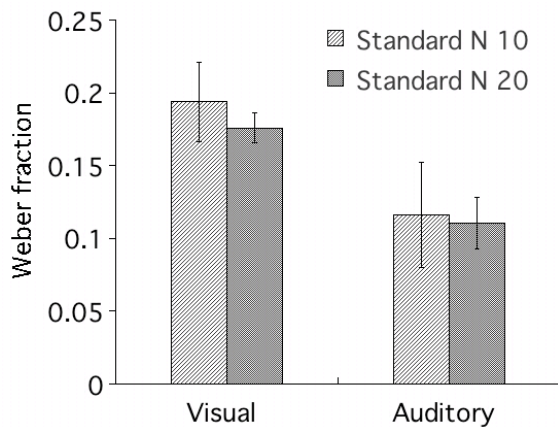


Figure 3: The means of Weber fractions in the visual and the auditory conditions at standard number of 10 and 20. Error bars represent standard deviations.

## Results and discussion

Figure 2 shows the average psychometric functions for each standard number. Figure 3 shows the mean Weber fraction in each condition. The fits of data points to psychometric functions were generally good, and the Pearson moment correlation coefficient exceeded 0.9 in all cases with the exception of the visual condition at the standard number 10 and 20 for one participant. The data of the participant were excluded while those of the remaining participants were used for further analysis.

To test whether and how precision in numerical comparison differs between the visual and the auditory conditions, a 2 modality (visual and auditory)  $\times$  2 standard numbers (10 and 20) repeated measures analysis of variance (ANOVA) was conducted on individual Weber fractions. This yielded a significant main effect for presentation modality [ $F(1, 3) = 90.38, p < .01$ ]. Weber fractions in the auditory condition were significantly smaller than those in the visual condition, suggesting that the numerical judgment in auditory modality was more precise than that in visual modality. No significant effect of the standard numbers was observed [ $F(1, 3) = .48, p > .1$ ], suggesting that the precision of numerical judgment was not affected by the number of elements within the numerical range tested in this experiment.

## Experiment 2

We tested the precision of approximate numerical comparison in three presentation conditions (i.e., visual, auditory, and cross-format). Since no systematic difference was observed between the standard numbers, we only use one standard number 10 in this experiment. Stimuli presentations of the visual and the auditory conditions were the same as those in Experiment 1. In the cross-modal condition, elements in one set were presented in the visual

sequence and those in the other set were presented in the auditory sequence. To examine the precision, we obtained Weber fractions that indicate the participant's variance of numerical comparison. To test the accuracy of the numerical comparison, we derived the point of subjective equality.

## Method

**Participants.** Newly recruited five participants participated in the experiment. All participants had no prior experience in numerical comparison tasks. All had normal or corrected-to-normal hearing and vision.

**Design** We compared three presentation conditions: the visual, the auditory and the cross-modal condition. The cross-modal condition had two sub-conditions: the cross-modal condition 1 and the cross-modal condition 2. In cross-modal condition 1, standard stimuli were visual sequence and comparison stimuli were auditory sequence. In cross-modal condition 2, standard stimuli were auditory sequences and comparison stimuli were the visual sequences. The numbers in the comparison element for the standard number at 10 were "7, 8, 9, 11, 12, and 13".

Trials in the visual, auditory, and two cross-modal conditions were separated and each constituted trial blocks. Three experimental conditions were presented among participants in a pseudo-counterbalanced order. Each condition had 192 trials (32 repetitions  $\times$  6 comparison levels), resulting in 768 trials in total. Each block had 48 trials, with 16 blocks in total. Participants performed five to six blocks in each experimental session, which took three days in total. Intermissions of approximately three minutes were given between blocks. Sequence of the trials was completely randomized within a block. Standard stimuli came first in half the trials and second in the remaining ones. Participants were given 12 practice trials before the actual experiment began.

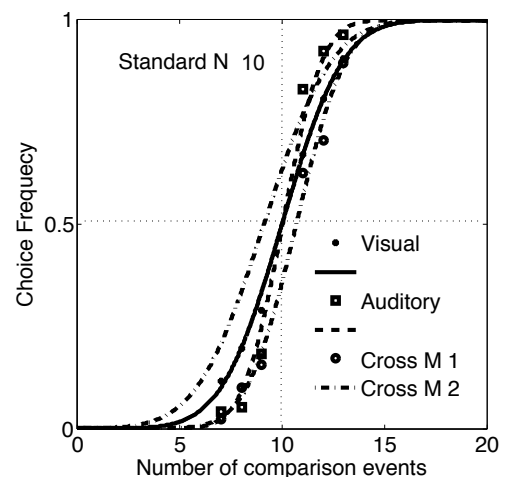


Figure 4: Average psychometric functions for the each presentation condition.

The stimuli, measurement, and procedures were the same as those in Experiment 1, with following exception. In the cross-modal condition, the auditory stimuli and the visual stimuli were shown in succession in random order.

## Results and discussion

Figure 4 shows the average psychometric functions for each condition. The fits of data points to psychometric functions were generally good, and the Pearson moment correlation coefficient exceeded 0.9 in all cases. Figure 5 shows the mean of Weber fractions and that of standardized PSEs of all participants. We averaged over Weber fractions for two cross-modal conditions for all participants and used the data for further analysis.

In order to test whether and how precision in numerical comparison differs between the visual, the auditory, the cross-modal conditions, a 3 condition repeated measures ANOVA was conducted on individual Weber fractions. There was a significant main effect for presentation modality [ $F(2, 4) = 9.43, p < .01$ ], and a Bonferroni post hoc analysis revealed that the Weber fractions in the visual and the cross-modal conditions were significantly larger than those in the auditory condition, indicating that precision in the visual and the cross-modal conditions was substantially worse than that in the auditory condition the same as the results in Experiment 1. The results suggested that the performance of the cross-modal trials would lie between that of the visual and auditory trials.

In order to test how cross-modal comparison affected the accuracy of numerosity comparison, we conducted a one-sample  $t$  test to compare the mean of the PSEs of the cross-modal condition 1 and that of the cross-modal condition 2 with the PSE of 0, respectively. The mean of PSEs in the cross-modal condition 1 was significantly larger than 0 [ $t(4) = 3.54, p < .05$ ] and the mean of PSEs in the cross-modal condition 2 was significantly smaller than 0 [ $t(4) = -5.43, p < .05$ ]. The results showed that the number of visual stimuli was overestimated relative to that of auditory stimuli in both cross-modal conditions.

## Discussion

We investigated whether and how precision in approximate numerical judgment between visual, auditory, and cross-modal presentations would differ. Our results demonstrated three significant findings. First, precision for numerical comparison of auditory sequence was significantly higher than that of visual sequence across two standard numbers. Second, precision in the visual and the cross-modal conditions was substantially worse than that in the auditory condition. Third, the number of visual elements was overestimated relative to that of auditory elements. Taken together, our results imply the existence of modality-specific processes in numerical comparison of the visual and auditory stimuli.

Our results are consistent with the previous studies that have shown the difference in counting precision across

modalities (e.g., Lechelt, 1975). Lower precision in the visual presentation is also consistent with the results of those studies. It is noteworthy that the similar effects were observed between the counting task and numerical comparison task.

What is the source of difference in the precision in numerical representation between visually and auditory presented stimuli, and how does the discrepancy in precision occur? In any modality, or cross-modal condition, stimuli need to be successively enumerated across time when the items of a set are presented sequentially. In this condition, the cardinal value of stimuli can be represented by the last numerical quantity. Common aspects of those numerical judgment is that they are time related irrespective to the sensory modality. In other time related tasks such as duration discrimination and empty interval estimation, it is known that the performance in the auditory presentation is significantly better than that in the visual and the tactile presentations. Thus, it is predicted that the temporal resolution may cause the superiority of auditory modality in numerical judgments. Further investigations are necessary to explore the possibility.

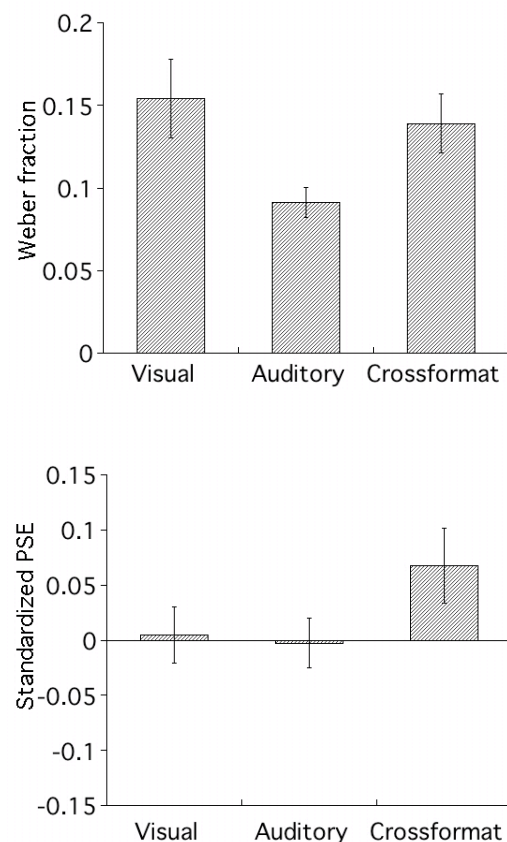


Figure 5: The means of Weber fractions and the means of PSEs of each modality condition. Error bars represent standard deviations.

As the performance of the cross-modal trials seemed to lie between that of the visual and auditory trials, it could be predicted that the convergent system could integrate the information from the auditory and visual numerical processing to form the higher abstract numerical presentations.

Another novel finding from this study is the overestimation of visual stimuli in the cross-modality comparison. Why did observer overestimate the number of visual stimuli relative to that of auditory stimuli? One possibility is that observers may overestimate the number of events with greater uncertainty (i.e., visual stimuli) in the decisional process. Another possibility is that the observers may perceive the events at the faster rate more numerous; since the time estimation for visual stimuli is shorter than the auditory stimuli, the visual stimuli appear with faster rate when they are presented at the identical rate. To test this possibility, we need to examine how human compare numerosities across modality in further investigation.

In conclusion, this study provided evidence for modality-specific processes in approximate numerical representation in human adults. Although many studies support the idea that human adults as well as infants and non-human animals share the modality-independent numerical system, it remains unknown how numerical information from the modality-specific system is combined at the judgment stage. Our findings imply that the process of approximate numerical representation is complex and involves multiple stages.

## References

- Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition*, 86, 201–221.
- Burgess, A., & Barlow, B. H. (1983). The precision of numerical discrimination in sequences of random dots. *Vision Research*, 23, 811–820.
- Cantlon, J. F., & Brannon, E. M. (2006). Shared system for Ordering Small and Large Numbers in Monkeys and Humans. *Psychological Science*, 17, 401–406.
- Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, 8, 307–314.
- Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, 44, 43–74.
- Groudin, S. (2010). Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception, & Psychophysics*, 72, 561–582.
- Hauser, M. D., Tsao, F., Garcia, P., & Spelke, E. S. (2003). Evolutionary foundations of number: spontaneous representation of numerical magnitudes by cotton-top tamarins. *Proceedings for the Royal Society of London*, 270, 1441–1446.
- Ivry, R. B., & Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. *Trends in Cognitive Sciences*, 8, 273–280.
- Jordan, K. E., & Brannon, E. M. (2006). The multisensory representation of number in infancy. *Proceedings for the national academy of sciences*, 103, 3486–3489.
- Jordan, K. E., Brannon, E. M., Logothetis, N. K., & Ghazanfar, A. A. (2005). Monkeys match the number of voices they hear to the number of faces they see. *Current Biology*, 15, 1034–1038.
- Kanabus, M., Szelag, E., Rojek, E., & Roppel, E. (2002). Temporal order judgment for auditory and visual stimuli. *Acta Neurobiologiae Experimentalis*, 62, 263–270.
- Kadosh, R. C., & Walsh, V. (2009). Numerical representation in the parietal lobes: Abstract or not abstract? *Behavioral and brain Sciences*, 32, 313–373.
- Kadosh, R. C., Lammertyn, J., & Izard, V. (2008). Are numbers special? An overview of chronometric, neuroimaging, developmental and comparative studies of magnitude representation. *Progress in Neurobiology*, 84, 132–147.
- Kobayashi, T., Hiraki, K., & Hasegawa, T. (2005). Auditory-visual intermodal matching of small numerosities in 6-month-old infants. *Developmental Science*, 8, 409–419.
- Lechelt, E. C. (1975). Temporal numerosity discrimination: Intermodal comparisons revisited. *British Journal of Psychology*, 66, 101–108.
- Nieder, A., & Dehaene, S. (2009). Representation of number in the brain. *The Annual Review of Neuroscience*, 32, 185–208.
- Penny, T. B., Gibbon, J., & Meck, W. H. (2000). Differential effects of auditory and visual signals on clock speed and temporal memory. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 1770–1787.
- Philippia, T. G., van Erp, J. B. F., & Werkhoven, P. J. (2008). Multisensory temporal numerosity judgment. *Brain Research*, 1242, 116–125.
- Piazza, M. (2010). Neurocognitive start-up tools for symbolic number representations. *Trends in Cognitive Science*, 14, 542–551.
- Pöppel, E. (1997). A hierarchical model of temporal perception. *Trends in Cognitive Science*, 1, 56–61.
- Riggs, J. K., Ferrand, L., Lancelin, D., Fryziel, L., & Dumur, G. (2006). Subitizing in Tactile Perception. *Psychological Science*, 17, 271–272.
- Tokita, M., & Ishiguchi, A. (2011). Temporal information affects the performance of numerical discrimination: Behavioral evidence for a shared system for numerical and temporal processing. *Psychonomic Bulletin & Review*, 18, 550–556.
- Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in Human: The psychophysics of Number Representation. *Psychological Science*, 10, 130–137.



# Embodied Communication Practices in Instructive Interaction between Musicians

Jackson Tolins (jackson.tolins@colorado.edu)

Department of Linguistics, Hellems 290, 295 UCB

Boulder, CO 80302 USA

## Abstract

Musicians, in the discussion and teaching of their art, commonly make use of vocalizations in order to demonstrate a particular melody or musical phrase. In the present study, we consider the use of these vocalizations as part of embodied depictions, and the role that these embodied communication practices play in music instruction. Data drawn from video recordings of private lessons between a clarinetist and his instructor demonstrate that these enactments are used in order to (re-)represent the experience of both performing and listening. The music instructor makes use of these embodied depictions for a number of actions central to teaching the art, including assessment, direction, and displays of understanding. In considering body-based communicative practices as an instructive tool, we consider both simulation and social action based cognitive perspectives through application in the analysis of goal-oriented interaction.

**Keywords:** Instruction; Embodied Cognition; Non-lexical Speech; Social Actions; Gesture; Body; Music Education.

## Introduction

For musicians, the ability to talk to each other about what is essentially a non-linguistic domain (music) plays a vital role in the teaching, rehearsing, and performing of their art. The requirements of the structured activity of music instruction motivate the musicians to make frequent use of ‘nonsense’ syllables during interaction; for example ‘*it sounds really effective, but it sounds a little bit urrrllliiaa*’. These non-lexical vocalizations appear to be a crucial and commonly used resource, for both teacher and student, in music instruction, despite the large and well-established technical vocabulary available. From a traditional linguistic view, these vocalizations are semantically null, and cannot be used as arbitrary and conventional referencing symbols (Goodwin, Goodwin & Olsher 2002). Yet in the situated activity of the music lesson, much of the instruction is achieved through their use. How is this accomplished? Understood within a framework of embodied cognition, which emphasizes the role of modal representations (Barsalou 2008), these ‘nonsense’ syllables become much more; they represent a means of sharing direct perceptual experience, providing the participants of the interaction the ability to make use of the mental machinery involved in perception and simulation for the purposes of learning and development.

The purpose of the current paper is to demonstrate that the interactive goals of the lesson, namely the teaching and learning of music performance, are achieved by and motivate the use of these embodied vocalizations as perceptually parallel direct experiences of the music under discussion. Combined with gestures, body movements,

prosodic contours, and interaction with the physical environment, the nonsensical speech sounds depict the music under discussion. By taking the framework of embodied cognition, through which behavior is understood in relation to the perception- and action-based systems of the body (Barsalou, 2003; Barsalou, 2008), the vocalizations can be understood as a means of grounding the interactive goals of the lesson in direct experience.

## Embodied Cognition and Instruction

Theories of embodied cognition, which have been invoked in a wide range of fields, focus on the explaining cognitive phenomena through an understanding of agent-environment dynamics. In contrast to cognitivist and computational paradigms, embodied cognition replaces the study of amodal symbolic representation and processing with the study of mental simulations and situated action as the basis of cognition. They provide an especially appropriate foundation for the study of the instruction and education of a non-verbal art such as music. In many types of instructional settings, body-based communication techniques anchor complex ideas and increase comprehension (Iverson & Goldin-Meadow, 1998; Goldin-Meadow, Kim, & Singer, 1999). For a paradigm in which cognition is situated socially and cognitive work is off-loaded onto the environment (Wilson, 2002), the use of multimodal communication practices in an educational setting aides in establishing appropriate learning and problem-solving behaviors (Nathan, 2008). For example, both children who are instructed to manipulate physical objects while reading mathematical story problems, and those instructed to imagine the manipulation have higher gains in problem solving ability than their non-imagining counterparts (Glenberg, Jaworski, Rischal, & Levin, 2007).

The paradigm of embodied cognition has been applied successfully to the study of body-based and situated mathematics education (see e.g. Lakoff & Nuñez, 1997; Nuñez, Edwards, & Matos, 1999). These researchers have focused on the embodied conceptual system from which mathematical reasoning arises, as well as the role played by the situated social contexts in which learning occur. These studies demonstrate that difficult concepts in a domain such as mathematics can be explained, in a manner beneficial to improving curriculum, through an understanding of the relationship between our perceptual experiences and our conceptual structures.

Similar to math education, an instructive setting such as a music lesson provides researchers interested in grounded cognition theories a naturalistic source of data through which to study both embodied communicative practices and



the social actions achieved through their use. Music performance requires a number of physical, sequential, and emotive skills, the learning of which takes place over many years of instruction. The ethnographic study of music instruction offers a window into the role features of embodied cognition play in the development of non-verbal skills in the interaction between student and mentor.

## **Music and Language**

In the data collected for the present study, the musicians are clarinetists. For a clarinet, sound is produced by the vibration of a wooden reed fitted to the underside of the mouthpiece. In order to produce this vibration on the clarinet, the clarinetist forces air through the instrument with their breath in a manner that mirrors the production of speech. In a manner also similar to speech, the tongue is used in order to stop the vibration of the reed and thus produce gaps of various sizes between notes. By changing the way in which the tongue touches the reed, thereby stilling it, different styles and strengths of separation between notes can be achieved. The use of vocalizations in the embodied depictions can be viewed, then, as mimicking the sensations of performing on the clarinet, from the exhalation of air from the lungs to the use of the tongue at various locations and with various pressures throughout the mouth.

Recent research has also demonstrated that the online processing of music and language overlaps (Patel, 2008; Fedorenko, Patel, Casasanto, Winawer, & Gibson, 2009), indicating that listening to a vocalized depiction of music may be a more similar experience to directly listening to instrumental music than has been previously thought. This would allow for the musicians to substitute the direct experience of music played on the clarinet with the non-lexical speech sounds of the vocalizations.

## **Vocalizations as Overt Simulation**

Previous work on non-conventional, non-lexical speech sounds has focused almost exclusively in two arenas: non-linguistic quotes and the use of non-lexical speech sounds by aphasics. Goodwin et al (2002) demonstrated how an aphasic man was able to utilize nonsense syllables in order to communicate effectively with those around him. Similar work on the use of non-lexical vocalizations as a communicative tool for agrammatic aphasics has been done by Wilkinson and colleagues (2010), who showed that the use of embodied enactment is an effective interactional means to achieve social actions with limited lexical and grammatical resources. Non-linguistic quotations, on the other hand, ("the car went *vrrrm*"), are typically described in particular syntactic constructions in the literature (Hudson, 1985), commonly as non-linguistic demonstrations in which the speaker depicts, rather than describes, their referents (Clark & Gerrig, 1990).

Given the similarities between both experiencing and performing music and experiencing and performing speech sounds, it becomes possible for the musicians to take

advantage of this similarity for the rehearsal of their art. Indeed, many music teachers will recommend mental rehearsal through making use of any opportunity, such as a walk across campus, to mentally play through a piece. Research into the role of simulation and imagined actions in the planning and execution of actions, such as playing an instrument, has flourished within the last 20 years (Jeannerod, 2001; Barsalou 2010). These studies demonstrate that the covert rehearsal of actions takes on many of the same characteristics as overt action, including temporal characteristics key to music production and experience (Decety, Jeannerod, & Prablanc, 1989). Drawing on the cognitive processes involved in simulation of the voice and music perception can also increase memory of rhythmic patterns (Pich 2000).

The enactments considered below can be described as overt, shared simulation; simulation with an interactional purpose. The overlapping cognitive substrate activated for both the perception and mental imagery of speech sounds and music provides the means by which embodied depictions of music may be substituted for the direct experience, to the benefit of the goals of the situated activity.

## **Vocalizations as Social Actions**

In many activities where the focus of the interaction is a non-verbal domain, the use of embodied communication is invaluable. Particularly in an instructive environment, the embodied actions considered here play an integral role in the development of expertise (Lave & Wenger, 1991; Firth & Wagner, 2007; Melander & Sahlstrom, 2009). By analyzing how it is that the vocalizations are used to achieve the specific interactional goals of the music lesson, namely the development of the abilities of the student, we provide strong motivation for including the study of embodied communicative practices, as part of the study of education and instruction. In the data collected for this study three distinction actions were found that make use of the vocalizations; assessment, direction, and response. Most commonly, the embodied depictions are used as a means of conveying a critique of a previously experienced bit of music. The depictions are also frequently used as part of a directive, commanding the student to follow with a translation of the vocalization on his instrument. Lastly, the musicians use the vocalizations in response to a bit of talk in order to display an understanding of what has just been described.

## **Assessment**

Consider the non-lexical vocalization that occurs in the following transcript. In depicting the music played by the student, the instructor makes use of multimodal and embodied communication practices to highlight the quality he wishes to assess, emphasizing and exaggerating the particular feature in a way that allows the student to perceive the issue in the music as well. In this example, the depiction is quotative in that it indexes previously

performed music and gains its interactional meaning through this reference as part of the larger action of assessment and evaluation. Prior to the start of the transcript, the student has just finished performing a passage from Ned Rorem's *Poems by Sylvia Plath*, which ends with a run of short fast descending notes. The instructor, after giving a general positive assessment, highlights this run of notes and attempts to correct the way the student performs them. In the transcriptions, "I" is used to indicate lines spoken by the instructor, "S" for the student.

#### Transcript 1

- I: Now here  
**(points to sheet)**<sup>1</sup>  
 I: umm, make sure-  
 I: make sure that we hear  
 S: **(blows air through instrument)**  
 I: the fours. Right now it sounds really-  
 effective but it sounds a little bit  
 → *urrrllllliiaa*, youknowwhatI'msaying  
 → **(Hand across)**  
 S: mm hmm

The instructor uses the speech sounds that make up the vocalization to depict the lack of clarity between the notes. Unlike almost all other vocalizations found in the data, in which the nonsense syllables typically consist of an open consonant-vowel structure, one syllable for each note, the instructor uses a combination of vowels and liquid consonants to emphasize the slurred nature of his student's performance. This choice in speech sounds allows for an embodied experience of the indistinct quality that the instructor observed, and indeed the vocalization is the only source of the negative assessment in the interaction. The transition from the high vowels of [u] and [i] to the low vowel [a] can also be seen to mirror the downward motion of jaw required for the move from the high to low register of the clarinet. The manipulation of the voice, in combining speech sounds into nonsense syllables of particular pitch, loudness, and tempo, is a key feature of the depiction.

The action achieved by the use of the non-lexical vocalization is strengthened by the simultaneous use of gesture. While the instructor is making the '*urrrllllliiaa*' sound, he draws his hand across the space in front of him in a slowly downward moving motion before rotating his pinky finger upward in release (see figure 1). The smooth motion of the gesture mirrors the emphasized characteristic of the quotation, with both the vocal and spatial modalities working together to depict the nature of the music being critiqued (see also Bräm & Bräm 2004 for an analysis of gesture in music-oriented interactions). His whole body also takes part in the reenactment; while he maintains orientation towards the sheet music, his torso and shoulders sway and dip across to the left, following and emphasizing the motion

<sup>1</sup> **(Bold)** is used to represent physical actions co-occurring with the line directly preceding.

of his hands. Thus three different modalities - gesture, prosody, and speech - are being integrated in the embodied depiction.

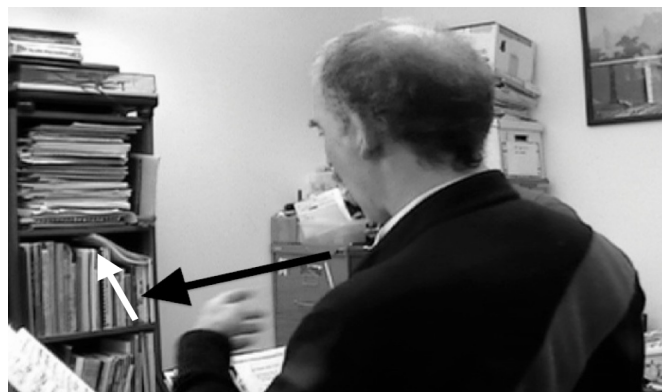


Figure 1: Gesture co-articulated with vocalization. The black arrow represents the first movement path, the white arrow the second.

The distinct features of the vocalization as used for assessment are a) the location within an instructive action sequence, wherein the student's previous performance is the subject matter, and b) its use as a quotative device to both pinpoint the trouble source in the previous performance and to simultaneously provide a critique. Importantly the embodied depiction itself is the only spoken negative critique, contrasted with '*very effective*' as part of an explanation of '*make sure we hear the fours*'. The role of the student in listening to such a depiction is to take on the role of the expert, re-experiencing his performance from a less direct standpoint that allows him to critically assess the facets of his performance emphasized by the instructor's quotation, which he may then affirm, as he does in the last line of the transcript.

#### Direction

Beyond assessment, the embodied depictions are used to directly change a future performance of a piece of music. Rather than indexing a previous experience, these directives specify exactly how a subsequent portion of music should be played. These directives may be used for a number of reasons during instruction, including taking apart a difficult passage at a slower speed for familiarization, practicing various facets of the technical skills, and prescribing changes to performances, as the following transcript demonstrates. Here, the instructor presents a hypothetical performance in his vocalization, one that does not necessarily index any previous experience. Rather, by making use of the depiction, the teacher offers the student an overt simulation of the music and in doing so requests a like performance by the student. The student may then draw on this perceptual experience as the basis of his own playing, which he does directly subsequently in translating what he experienced in the vocalized simulation into playing the clarinet.

### Transcript 2

- I: Just play the F, make a little  
bit of a diminuendo and then think  
→ baadaadaaDAA  
→ I: leading! yapupuuPAA  
(student inhales and then plays)

These directives are used for a number of purposes in the activity of the lesson. As in the example above, they are used to make specific changes to the performance style of the student. They may also be in forming a deeper understanding of the music through directing the student to play in an unfamiliar manner, such as playing a passage with a particular subdivision of the rhythm, or at a reduced tempo. Through presenting the student with a new perspective on the music through the vocalization, and directing the student to then mimic the vocalization with his instrument, the instructor develops both the student's perceptual and action-based understanding of the music.

### Response

Finally, a third social action makes use of the embodied depictions in the instructional setting between the musicians. These act as responsive displays of understanding, and so work to depict what has just previously been described by the co-participant. In transcript (3) below, the student is attempting to describe the rhythmic breakdown of a passage. The instructor responds by vocalizing the passage under discussion as his interpretation of the student's description. By embodying the music in this manner, the instructor displays his understanding of his student's talk in a way that allows the student to provide uptake, which he does positively directly after the depiction.

### Transcript 3

- S: So right now I'm thinking more of a six  
(pause)  
S: I'm um going to-  
I: yaadaadaadaadaadaadaa  
(pause)  
S: yea. diyadadadada  
I: So you're going for the C  
S: Um, I mean, yea I think I'm trying to

Here, the instructor's vocalization enacts his understanding of his student's explanation, allowing the student to experience the music directly and confirm that the instructor has indeed understood what was meant. That the student then double-checks the accuracy of this vocalized depiction by repeating it to himself demonstrates the role these vocalizations play as perceptual anchors for their understanding of the music.

## Conclusion

The means by which the musicians make use of cognitive parallels between the perception of their voices and the perception of music is not easily understood in terms of a semantics based on amodal and symbolic processing. Instead, applying an embodied perspective we have demonstrated that the role of simulation in perception and action may be exploited by the interactants for the purposes of the situated activity of instruction. An ethnographic study of the use of embodied communicative practices in education drives home the need for a larger investigation into the role that embodied information processing plays in the development of procedural and technical skills. In the music lesson, the instructor makes use of the perceptually grounded comprehension of auditory phenomena, such as music and the voice, in order to commit a large number of distinct social actions essential to developing musicality on the part of the student. This includes the quotation of previously shared experiences to simultaneously point out the trouble spot and provide a critique, depicting novel musical expressions to be translated from voice to instrument, and displaying understanding of some bit of talk. In contrast to previous studies that focus on the role of embodied experience to develop abstract concepts, these few examples demonstrate that the instructor is able to manipulate the perceptual-experiential systems of simulation *in situ* to rapidly and accurately share perspectives that will lead to the development of his student. Music is but one of many non-verbal domains in the arts and other technical skills in which the study of embodied communication practices could lead to the development of improved education.

## Acknowledgments

I would like to thank Barbara Fox and the members of CU Language Project for many helpful conversations and comments, and the musicians for their willingness to participate.

## References

- Barsalou, L.W. (2003) Situated simulation in the human conceptual system. *Lang. Cogn. Process.*, 18, 512-62  
Barsalou, L.W. (2008) Grounded cognition. *Annu. Rev. Psychol.*, 59, 617-645  
Barsalou, L.W. (2010) Grounded cognition: Past, present, and future. *Topics in Cog. Sci.*, 2, 716-724  
Bräm, P. & Bräm T. (2004) Expressive gestures used by classical orchestra conductors. In: Muller, C. & Posner, R. (Eds.) *The semantics and pragmatics of everyday gestures*, Weidler Buchverlag  
Clark, H. & Gerrig R. (1990) Quotations as demonstrations. *Language*, 66, 4, 764-805  
Decety, J., Jeannerod, M., & Prablanc, C. (1989) The timing of mentally represented actions. *Behav. Brain Res.*, 34, 35-42

- Goodwin, C., Goodwin, M. H., & Olsher, D. (2002) Producing sense with nonsense syllables: Turn and sequence in conversations with a man with severe aphasia. In: Ford, C. E., Fox, B., & Thompson, S. A. (Eds.) *The language of turn and sequence*, Oxford University Press
- Firth, A. & Wagner, J. (1997) On discourse, communication, and (some) fundamental concepts in SLA research. *The modern language journal*, 81, 3, 285-300
- Glenberg, A. M., Jaworski, B., Rischal, M., & Levin, J. R. (2007). What brains are for: Action, meaning, and reading comprehension. In D. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* Mahwah, NJ: Lawrence Erlbaum
- Goldin-Meadow, S., Kim, S., & Singer, M. (1999) What the adult's hands tell the student's mind about math. *Journal of educational psychology*, 6, 138-43
- Hudson, R. (1985) The limits of subcategorization. *Linguistic analysis*, 15, 233-55
- Iverson, J. & Goldin-Meadow, S. (1998) Why people gesture when they speak. *Nature*, 396, 228
- Jeannerod, M. (2001) Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage*, 14, S103-S109
- Lave, J. & Wenger, E. (1991) *Situated learning: Legitimate peripheral participation*. New York: Cambridge University Press
- Lakoff, G. & Nunez, R. (1997) The metaphorical structure of mathematics: Sketching out cognitive foundations for a mind-based mathematics. In: L. English (ed.) *Mathematical reasoning: Analogies, metaphors, and images*, Erlbaum, Hillsdale, NJ
- Melander, H. & Sahlstrom F. (2009) Learning to fly: The progressive development of situation awareness. *Scandinavian journal of educational research*, 53, 2, 151-166
- Nathan, M. J. (2008). An embodied cognition perspective on symbols, gestures and grounding instruction. In M. DeVega, A. M. Glenberg, & A. C. Graesser (Eds.), *Symbols, embodiment and meaning*. Cambridge: Oxford University Press.
- Nunez, R., Edwards, L., & Filipe Matos, J. (1999) Embodied cognition as grounding for situatedness and context in mathematics education. *Educational studies in mathematics*, 39, 45-65
- Pich, J. (2000) The role of subvocalization in rehearsal and maintenance of rhythmic patterns. *Spanish journal of psychology*, 3, 63-67
- Wilkinson, R., Beeke, S., & Maxim, J. (2010) Formulating actions and events with limited linguistic resources: Enactment and iconicity in agrammatic aphasic talk. *Research on language and social interaction*, 43, 1, 57-84
- Wilson, M. (2002) Six views of embodied cognition. *Psychon. Bull. Rev.*, 9, 625-36

# Viewing and performing actions can change what you see

Alexia Toskos Dils (atoskos@stanford.edu)

Stephen J. Flusberg (sflus@stanford.edu)

Lera Boroditsky (lera@stanford.edu)

Stanford University, Department of Psychology  
Jordan Hall, 450 Serra Mall, Building 420, Stanford, CA 94305 USA

## Abstract

Previous research has demonstrated a tight link between object perception and action: viewing an object primes the action needed to interact with it, while priming an action can affect the speed and accuracy with which we perceive the object. However, it is not yet known whether motor information can qualitatively change what object we actually perceive. We investigated this issue by having participants view or perform an action before viewing an ambiguous object. Results showed that viewing an action (a picture of a hand displaying a power or precision grasp) biased participants to interpret the ambiguous object as congruent with the action prime (Experiments 1 and 2). Conversely, performing an action (moving small or large balls from one tray to another) biased participants to interpret the object as incongruent with the motor action. Together, these results suggest viewing and performing actions can actually change what we see.

**Keywords:** Object perception; Action; Embodiment

## Background

Can our actions influence how we perceive the world and affect the very contents of our visual awareness? Though perception and action have traditionally been studied independently in the cognitive sciences, in our everyday experience of the world they are dynamically linked. For example, many of the objects we look at are also the objects we grasp and manipulate. More generally, our movements and actions in the environment alter what perceptual information we have access to, and these changes in perceptual stimulation consequently influence how we traverse our surroundings and what actions we choose to take. For reasons such as these, ecologically orientated psychologists have argued that we perceive the world in terms of how it affords action (Gibson, 1979).

In recent years, researchers have gathered evidence in support of this view, showing tight links between perception and action across a wide range of cognitive and behavioral tasks (e.g., Witt & Proffitt, 2005; Bhalla & Proffitt, 1999; Witt, Proffitt, & Epstein, 2004). Other researchers have examined the role that motor actions play in object perception (e.g., Borghi et al., 2007; Bub et al., 2008; Chao & Martin, 2000; Helbig, Graf, & Kiefer, 2006; Tucker & Ellis, 1998, 2001; Witt & Brockmole, in press; Witt, Kemmerer, Linkenauer, & Culham, 2010). For example, Tucker and Ellis conducted a series of studies to test whether people automatically generate a motor representation in response to the visual presentation of an object, even when there is no intention to act on the object (Tucker & Ellis, 1998; 2001). In one experiment,

participants made a left or right-handed button press to indicate whether an image of an object on the screen was upright or inverted. The objects were chosen to have a clear right or left-handed affordance (e.g., a frying pan with a handle oriented to the left affords a left-handed grasp). Participants responded faster and made fewer errors when their responding hand was congruent with the (task-irrelevant) affordance of the object on the screen.

Additional work has found that the relationship between motor actions and object perception is *functional* and not merely epiphenomenal. For example, Borghi et al. (2007) found that participants were faster to respond a picture of an object when it was preceded by a picture of a hand displaying an action that was congruent with the object. The authors concluded that visually priming an action facilitates object recognition (see also Helbig et al., 2006, Witt & Brockmole, in press). This suggests that preventing someone from engaging in an action should impair object recognition in a parallel fashion. Indeed, Witt et al. (2010) showed that participants were slower and less accurate when responding to a picture of a tool if the handle in the picture was oriented towards the participant's hand that was busy squeezing a rubber ball.

Taken together, these studies suggest that motor information can play a significant role in object perception by affecting the speed and accuracy with which we perceive an object. However, it is unclear just how deeply motor information can penetrate into our visual perception of objects. For instance, can viewing or performing a particular action *qualitatively* affect this perceptual process and change what object we actually see?

We investigated this possibility across three experiments. In Experiments 1 and 2, participants first viewed an image of a hand depicting a particular action (one of two specific grasp types). They then saw an image of an ambiguous object and had to indicate what they perceived it to be. Participants were biased to interpret the object as congruent with the action prime.

What cognitive mechanisms might underlie this effect? One possibility is that viewing the hand action prime led participants to imagine or simulate performing that action themselves (Parsons, 1987; Rizzolatti & Craighero, 2004). Then, when they viewed the ambiguous image, participants saw the object they were prepared to interact with because of this active motor state (Hommel et al., 2001). On this view, perceived events and planned actions share a common representational medium to the extent that they share common (abstract) features. Alternatively, this effect may

have simply been a result of purely visual or semantic priming due to the association between certain grasp types and certain objects.

To distinguish these possibilities, in Experiment 3 participants engaged in an actual manual motor action (moving small or large balls from one tray to another) while naming pictures displayed on a computer screen, including the ambiguous object used in Experiment 2. A visual priming account would predict that performing an action should have no effect on this task as long as participants cannot see their own hands as they engage in the action. A semantic priming account would predict that no matter how the action concept is activated (e.g. viewing an action, talking about an action, performing an action), the results should yield the same facilitation effect observed in Experiments 1-2. Conversely, the common coding approach would predict that performing an action should actually *interfere* with a participant's ability to perceive an action-congruent object, which will therefore lead them to perceive an action-*incongruent* object (Hommel et al., 2001). In this study, participants were actually biased to interpret the object as incongruent with the motor action they performed. This suggests that viewing and performing actions are supported by the same underlying representations.

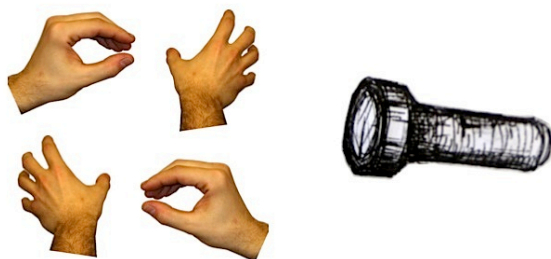
## Experiment 1

Can viewing an action change what object we see?

### Methods

**Participants** 815 individuals were recruited to participate in this study through the amazon.com Mechanical Turk website in exchange for payment.

**Stimuli & Procedure** The stimuli for this experiment consisted of four photographs of hands and an ambiguous object line drawing created by the authors (Figure 1). The four hand photographs showed either left or right hands in either a power or precision grasp. Pilot testing suggested that the ambiguous object could be interpreted as an object that afforded a power grasp (*flashlight*) or as an object that afforded a precision grasp (*screw/bolt*). The drawing could also be interpreted as an object that afforded a right-handed functional grasp (e.g., the *flashlight* as oriented in Figure 1) or as an object that afforded a left-handed functional grasp (e.g., the *screw/bolt* as oriented in Figure 1).



**Figure 1.** In Experiment 1, participants viewed one of the four hand images on the left, and then viewed the ambiguous object on the right.

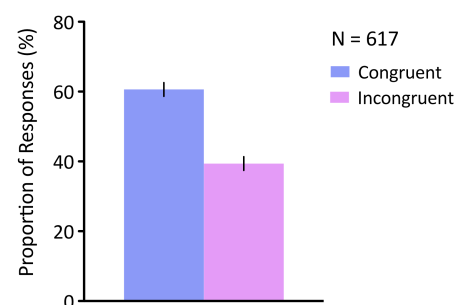
One of the four hand images was randomly selected for each participant and displayed on the screen for three seconds. Next, the ambiguous object drawing was displayed at 56% the size of the hand for three seconds. The left/right orientation of the drawing was counterbalanced across participants. After this, participants were asked to identify the object in the line drawing that they had just seen. They were then asked to identify whether the hand they had seen was a left or right hand. Finally, they were asked if they had any additional interpretation of the object and indicated whether they were left-handed, right-handed, or ambidextrous.

### Results

The data from 179 participants were removed from analysis because they failed to respond to the test questions appropriately (e.g., did not provide an interpretation of the ambiguous object), because they took the survey more than once, or because they responded incorrectly to the question of whether the hand prime they saw was a left or right hand. This last question was used as manipulation check to ensure that participants were looking at and paying attention to the experimental stimuli.

*Testing for effects of grasp type.* For the remaining 636 participants, we coded their initial ambiguous object interpretation as *congruent* if it matched the hand prime they saw (i.e., power grasp hand and flashlight interpretation or precision grasp hand and bolt/screw interpretation). Responses were coded as *incongruent* if they did not match the hand prime (i.e., power grasp hand and bolt/screw interpretation or precision grasp hand and flashlight interpretation). 19 participants came up with both interpretations for the ambiguous object and were therefore removed from further analysis.

Of the 617 participants in this final set of data, nearly 61% (N=374) gave *congruent* responses, while 39% (N=243) gave *incongruent* responses (Figure 2). A chi-square goodness of fit test revealed that this difference was highly significant,  $\chi^2 = 27.4, p < 0.0001$ .



**Figure 2.** Results from Experiment 1, showing proportion of congruent and incongruent object interpretations. Error bars represent the standard error of the proportion.

One possible explanation for these results is that our study design was transparent and therefore our participants simply



told us what they thought we wanted to hear. If our results were caused by this demand characteristic, then participants presumably perceived *both* interpretations for the ambiguous object and selected the interpretation they believed would make us happy. We tried to account for this possibility by asking participants if they had any additional interpretation of the object as one of our follow-up questions. In fact, 126 of our 617 participants provided additional interpretations of the object. Among the remaining 490 participants who only perceived one object interpretation, the results mirrored our previous analysis: nearly 61% (N=297) gave *congruent* responses, while 39% (N=193) gave *incongruent* responses. A chi-square goodness of fit test revealed that this difference was highly significant,  $\chi^2 = 21.7, p < 0.0001$ .

*Testing for effects of orientation.* We also asked whether the laterality of the action prime or participants' own handedness biased perception of the ambiguous object. To test these possibilities, participants' interpretations were coded as *leftward* if they saw the object whose handle (i.e., the head of the "bolt" or the barrel of the "flashlight") pointed to the left, and *rightward* if they saw the object whose handle pointed to the right. Neither the laterality of the action prime (51% congruent, N=249; 49% incongruent, N=241;  $\chi^2 = 0.1, p > 0.5$ ), nor the handedness of the participant (51% congruent, N=238; 49% incongruent, N=237;  $\chi^2 < 0.01, p > 0.5$ ), predicted whether subjects made a leftward or rightward interpretation of the object.

## Discussion

In this experiment we asked whether viewing an action would influence what participants saw when they looked at an ambiguous object. We found that when participants were primed with a hand displaying a power grasp they were more likely to interpret an ambiguous drawing as an object that required a power grasp (*flashlight*). Conversely, when they were primed with a hand displaying a precision grasp, they were more likely to interpret the drawing as an object that required a precision grasp (*screw/bolt*). These results remained even after we removed participants who provided multiple interpretations of the ambiguous object, which helps to rule out an explanation based on demand characteristics. These findings suggest that viewing an action can qualitatively affect our perception of an object.<sup>1</sup>

However, manual actions are complex, and grasp type is just one dimension out of many that might affect object

perception. In Experiment 1, we also tested whether priming an action with a right or left hand, irrespective of whether it displayed a power or precision grasp, would influence whether people perceived a leftward or rightward-facing object. We also reasoned that action simulations might be constrained by the idiosyncrasies of an individual's own motor system, so we tested whether the handedness of each participant, irrespective of the action prime, caused them to see a leftward or rightward-facing object. In our task, neither the laterality of the action prime nor the handedness of the participant affected what the ambiguous object appeared to be.

Why might the type of grasp displayed by a hand affect object perception, but not the laterality of the grasp or handedness of the participant? Perhaps some features of actions become privileged over others because they are more reliably associated with specific objects. Whether an object requires a power or precision grasp, for example, depends largely on the object's size, and for artifacts like flashlights and bolts, size is relatively constant across instances. The hand we use to grasp these objects, however, varies considerably depending on what we intend to do with the object and what else our hands are busy doing. By pitting various features of manual action against one another, we might have limited the likelihood that weaker effects of laterality and handedness would materialize. Exploring this possibility with objects that are ambiguous on one dimension only is the subject of future work.

It is also worth noting that the object we used in Experiment 1 was a *tool* under all possible interpretations, which might further limit our ability to generalize the effects of viewing actions to all graspable objects. Would the patterns we found in Experiment 1 with the flashlight/bolt image extend to graspable objects whose *primary* affordance is related to eating and not grasping (e.g., fruit)? Furthermore, the flashlight/bolt image remains an abstract, ambiguous, unrealistic line drawing. Would a photorealistic image in which the ambiguity of the object was less obvious show similar effects from viewing actions? To test these possibilities, we replicated this study in Experiment 2 using a photorealistic image of an object that could either be seen as an apple or a cherry.

## Experiment 2

In Experiment 1 we found that viewing an action influenced what participants saw when they looked at an ambiguous object. However, it remains unclear whether these results will generalize to more realistic-looking objects that are not in the tool category. To address this issue, we created a new ambiguous object, the photorealistic image depicted in Figure 3 that can be interpreted as an apple (power grasp affordance) or a cherry (precision grasp affordance). We then replicated Experiment 1 using this new object.

## Methods

<sup>1</sup> The results of Experiment 1 replicate a pilot version of this study reported at an earlier meeting of the Cognitive Science Society conference (Flusberg, Toskos Dils, & Boroditsky, 2010). Though the main effect in that study was nearly the same as in Experiment 1, the nature of the ambiguous object we used (a line drawing that could be perceived as a football or a nut) limited how we could interpret the results. First, this object elicited much more varied interpretations than the stimuli used in the present set of studies, suggesting that it may have been perceived as an extremely abstract figure rather than a concrete object. Second, a greater proportion of participants had multiple interpretations of the object than we see in the present study.



**Participants** 353 individuals were recruited to participate in this study through the amazon.com Mechanical Turk website in exchange for payment.

**Stimuli & Procedure** The stimuli and procedure for this experiment were identical to Experiment 1, with the exception of the ambiguous object, which was the cherry/apple picture depicted in Figure 3 presented at 29% the size of the hand.



**Figure 3.** The ambiguous object created for Experiment 2. It can be interpreted as an apple, which affords a power grasp, or a cherry, which affords a precision grasp.

## Results

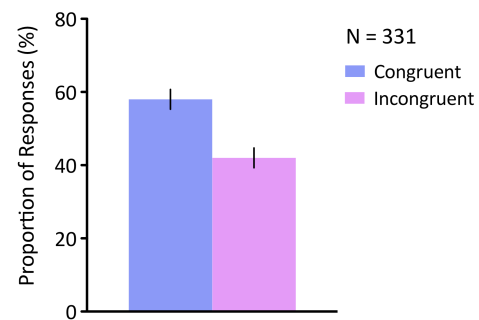
The data from 22 participants were removed from analysis because they failed to respond to the test questions appropriately, because they took the survey more than once, or because they responded incorrectly to the question of whether the hand prime they saw was a left or right hand. For the remaining 335 participants, we coded their initial ambiguous object interpretation in the same way we did in Experiment 1. Four participants came up with both interpretations for the ambiguous object and were therefore removed from further analysis.

Of the 331 participants in this final set of data, 58% (N=192) gave *congruent* responses, while 42% (N=139) gave *incongruent* responses (Figure 4). A chi-square goodness of fit test revealed that this difference was highly significant,  $\chi^2 = 8.16$ ,  $p < 0.005$ . Once again, we used responses to our follow-up question to help rule out a demand characteristic account of these results. 110 participants provided additional interpretations of the object. Among the remaining 221 participants who only perceived one object interpretation, the results mirrored our previous analysis. Nearly 62% (N=136) gave *congruent* responses, while 38% (N=85) gave *incongruent* responses. A chi-square goodness of fit test revealed that this difference was highly significant,  $\chi^2 = 11.32$ ,  $p < 0.001$ . The pattern of results produced by the *cherry/apple* in Experiment 2 did not differ reliably from the pattern produced by the *flashlight/bolt* from Experiment 1,  $\chi^2 = 0.03$ ,  $p > 0.5$ .

## Discussion

The results of Experiment 2 replicated what we found in Experiment 1 using a photorealistic ambiguous object that was in a very different category from the tool image used in the previous study. Taken together, these experiments

demonstrate that viewing an action can qualitatively change how people perceive an object.



**Figure 4.** Results from Experiment 2, showing proportion of congruent and incongruent object interpretations. Error bars represent the standard error of the proportion.

What cognitive mechanisms might underlie this effect? One possibility is that the hand action prime led participants to simulate performing that action themselves (Parsons, 1987; Rizzolatti & Craighero, 2004). Then, when shown the ambiguous image, participants saw the object they were prepared to interact with because of this active motor state (Hommel et al., 2001). On this view, perceived events and planned actions share a common representational medium to the extent that they share common (abstract) features. Alternatively, this effect may have simply been a result of visual or semantic priming due to associations between grasp types and objects. Importantly, these accounts make three distinct predictions about how performing an action when participants cannot see their hands should affect object perception. A visual priming account would predict that performing an action should have no effect on this task as long as participants cannot see their own hands as they engage in the action. A semantic priming account would predict that no matter how the action concept is activated (e.g. viewing an action, talking about an action, performing an action), the results should yield the same facilitation effect observed in Experiments 1-2. Conversely, the common coding approach would predict that performing an action should actually *interfere* with a participant's ability to perceive an action-congruent object, which will therefore lead them to perceive an action-*incongruent* object (Hommel et al., 2001). Experiment 3 was designed to differentiate among these possibilities by having participants perform a manual motor action while they observed the ambiguous object used in Experiment 2.

## Experiment 3

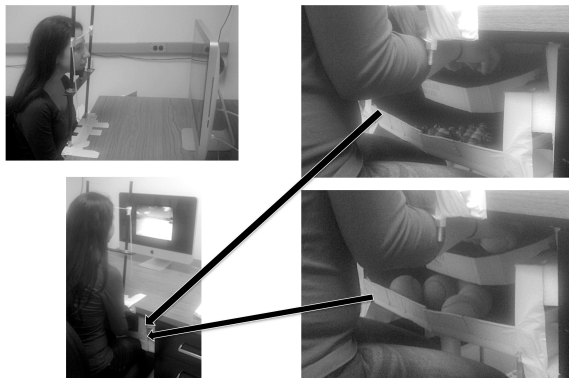
Can performing an action change what you see in the same way that observing an action does?

## Methods

**Participants** 102 individuals were recruited to participate in this experiment from the Stanford community in exchange for course credit or five dollars.

**Stimuli & Procedure** When participants entered the lab, they were told they would be partaking in a study of multitasking. They were then seated at a desk and positioned with their head in a chin rest facing a computer screen (Apple iMac, 20" monitor). At this point they were given detailed instructions for how to proceed in the task.

The motor action participants performed consisted of picking up and moving balls located in a tray underneath the desk they were seated at (Figure 5). Participants picked up one ball in each hand from the lower tray and moved them simultaneously to the upper tray whenever they heard a beep coming from the computer. The apparatus was designed so that balls placed in the upper tray would fall back down into the lower tray. Importantly, with their heads in the chinrest, participants could not see this action as they performed it. Half of the participants were randomly assigned to pick up tennis balls, which require a power grasp action, while the remaining participants picked up small bouncy balls, which require a precision grasp action.



**Figure 5.** The laboratory setup used for Experiment 3. Half of participants moved bouncy balls in each hand (upper-right) while the remaining participants moved tennis balls in each hand (lower-right).

When the experiment began, the screen was black. A beep was played every 1.25 seconds, and each time it played participants engaged in the ball moving action. After 12.5 seconds, pictures started appearing on the screen one by one, each one remaining on the screen for 2 seconds, with an inter-stimulus interval of 500 milliseconds. While these pictures were appearing, the beeps kept playing at a rate of one every 1.25 seconds (twice per image).

Participants were instructed to name aloud the image on the screen as quickly as possible. There were 12 images in all, and the first 11 depicted objects or scenes that did not afford a particular manual grasp action (e.g., beach, house, etc.). The final image was the ambiguous cherry/apple

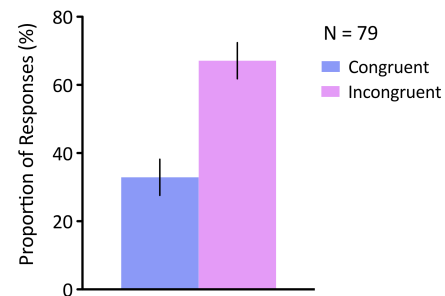
object used in Experiment 2. The pictures were presented in the same order for all participants.

## Results

The results from 2 participants were removed because they failed to complete the experimental task (i.e., they did not name every picture that appeared on the screen).

For the remaining 100 participants, we coded their response to the final picture (the ambiguous cherry/apple) as *congruent* if it matched the action they were performing (moving tennis balls and said apple, or moving bouncy balls and said cherry), and *incongruent* if it did not match the action they were performing (moving tennis balls and said cherry, or moving bouncy balls and said apple). 21 participants said both cherry and apple and were therefore removed from further analysis.

Of the 79 participants in this final set of data, 33% (N=26) gave *congruent* responses, while 67% (N=53) gave *incongruent* responses (Figure 6). A chi-square goodness of fit test revealed that this difference was highly significant,  $\chi^2 = 8.56, p < 0.005$ . This pattern differed reliably from the patterns observed in Experiment 1, ( $\chi^2 = 20.87, p < 0.0001$ ), and Experiment 2,  $\chi^2 = 18.06, p < 0.0001$ .



**Figure 6.** Results from Experiment 3, showing proportion of congruent and incongruent object interpretations. Error bars represent the standard error of the proportion.

## Discussion

In this experiment we asked whether performing an action would influence what participants saw when they looked at an ambiguous object. We found that when participants engaged in power grasp action (moving tennis balls in each hand), they were biased to perceive an ambiguous object that was *incongruent* with that action (i.e., a cherry, which affords a precision grasp). Similarly, when they engaged in precision grasp action (moving small bouncy balls in each hand), they were also biased to perceive an ambiguous object that was incongruent with that action (an apple, which affords a power grasp).

Therefore, it seems that performing an action can change what people see when they look at an ambiguous object, and the direction of the effect suggests that it arises from overlapping representations between perception and action. Indeed, as predicted by the common coding account (and not by visual or semantic priming mechanisms), the specific pattern of results in Experiment 3 shows the *opposite*

pattern from what we observed in Experiments 1 and 2 (when participants simply observed an action). In those experiments, viewing an action resulted in a *priming* effect, such that participants were biased to perceive an object that was congruent with the action they observed. In Experiment 3, on the other hand, performing an action resulted in an *interference* effect, such that participants were biased to perceive an object that was incongruent with the action they were engaged in.

However, there is one key difference between the experiments that may also have contributed to these divergent results. While participants in Experiments 1 and 2 only observed a single visual hand prime, participants in Experiment 3 engaged in a repetitive series of manual actions, moving balls from one tray to another 34 times. Behavioral repetition of this sort has been known to cause *adaptation* effects such that performance on a subsequent task is biased in the opposite direction of repeated behavior (e.g., Cattaneo et al., 2011). We are currently working on a new series of laboratory studies designed to tease apart the different possible mechanisms that may underlie the divergent patterns of results observed in these experiments.

### General Discussion

We began this paper by asking whether viewing or performing an action could qualitatively affect how people perceive an object. That is, does our current motor state change how we see the world?

In Experiments 1 and 2, participants first viewed an action hand prime and then viewed an ambiguous object. They were biased to perceive the object as congruent with the preceding hand image. When the hand prime displayed a power grasp, participants were more likely to see an object that afforded such a grasp, like a flashlight or an apple. When the hand prime displayed a precision grasp, participants were more likely to see a bolt or cherry, which afford the same grasp type. In Experiment 3, participants performed a manual motor action while they interpreted the ambiguous object. When they were picking up tennis balls, which afford a power grasp, they were more likely to see an object that afforded a precision grasp (i.e., cherry). Similarly, when they were picking up small bouncy balls, which afford a precision grasp, they were more likely to see an object that afforded a power grasp (i.e., apple).

These results demonstrate that viewing or performing an action can in fact qualitatively change what an object is perceived to be, and the pattern of results across experiments suggests that this shift is subserved by shared representations between perception and action.

### Acknowledgments

The authors would like to thank Laura Malkiewich for creating our cherple stimulus and help with earlier versions of these studies. We would also like to thank all members of the Cognition Lab. This research was supported by a McDonnell scholars grant & NSF BCS #1058119 to LB.

### References

- Bhalla, M. & Proffitt, D. R. (1999). Visual-motor recalibration in geographical slant perception. *JEP: Human Perception and Performance*, 25, 1076-1096.
- Borghi, A.M., Bonfiglioli, C., Lugli, L., Ricciardelli, P., Rubichi, S., Nicoletti, R. (2007). Are visual stimuli sufficient to evoke motor information? Studies with hand primes. *Neuroscience Letters*, 411(1), 17-21.
- Bub, D. N., Masson, M. E. J., & Cree, G. S. (2008). Evocation of functional and volumetric gestural knowledge by objects and words. *Cognition*, 106, 27-58.
- Cattaneo, L., Barchiesi, G., Tabarelli, D., Arfeller, C., Sato, M. & Glenberg, A.M. (2011). One's motor performance predictably modulates the understanding of others' actions through adaptation of premotor visuo-motor neurons. *Social Cognitive & Affective Neuroscience*, 6(3), 301-310.
- Chao, L. L. & Martin, A. (2000) Representation of manipulable man-made objects in the dorsal stream. *Neuroimage*, 12, 478-484.
- Flusberg, S. J., Toskos Dils, A., & Boroditsky, L. (2010). Motor affordances in object perception. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2105-2110). Austin, TX: Cognitive Science Society.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Lawrence Earlbaum: Hillsdale, NJ.
- Helbig, H. B., Graf, M., & Kiefer, M. (2006). The role of action representations in visual object recognition, *Experimental Brain Research*, 107(2), 221-228.
- Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, 24, 840-937.
- Parsons, L. M. (1987). Imagined spatial transformation of one's body. *JEP: General*, 19, 178-241.
- Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. *Annual Reviews Neuroscience*, 27, 169-192.
- Tucker, M. & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *JEP: Human Perception and Performance*, 24(3), 830-846.
- Tucker, M. & Ellis, R. (2001). The potentiation of grasp types during visual object categorization. *Visual Cognition*, 8(6), 769-800.
- Witt, J. K., & Brockmole, J. R. (in press). Action alters object identification: Wielding a gun increases the bias to see guns. *JEP: Human Perception and Performance*.
- Witt, J. K., Kemmerer, D., Linkenauger, S. A., & Culham, J. (2010). A functional role for motor simulation in naming tools. *Psychological Science*, 21, 1215-1219.
- Witt, J. K. & Proffitt, D. R. (2005). See the ball, hit the ball; Apparent ball size is correlated with batting average. *Psychological Science*, 16, 937-939.
- Witt, J.K., Proffitt, D.R., & Epstein, W. (2004). Perceiving distance: A role of effort and intent. *Perception*, 33, 577-590.

# Honoring Different Ontological Boundaries: The Role of Language in Category Formation

**Duc N. Tran (dntran2@uh.edu)**

Department of Psychology, 126 Heyne Bldg.  
Houston, TX 77204 USA

**Hanako Yoshida (yoshida@uh.edu)**

Department of Psychology, 126 Heyne Bldg.  
Houston, TX 77204 USA

## Abstract

The present study examines the different ways in which language structure marks individuation and cue early attention in a novel noun generalization task. Results in the present study extends the Boundary Shift Hypothesis, suggesting that the linguistic boundary between individuals and nonindividuals influences the perceptual boundaries and the correlational patterns formed overtime between ontological kinds. The results provide a new perspective on the facilitative role of linguistic markers in category formations, rather than strictly in boundary formations. This demonstrates the potential degree of cognitive processing among different language learners and lends support towards a mechanistic explanation of the role of language in categorial formations.

**Keywords:** category learning, categorical structure, ontology, ontological boundaries, linguistic structure, individuation

## Introduction

Language is a symbol system that maps the world's categories. It allows people to efficiently shape their world through abstract representations. Words have long been viewed as a vital unit that allows one to shape abstract information to promote and refine clusters of information for representational use. Such examinations allow one to group information into arbitrary categories that assist in later retrieval (with less effort and less cost). Categories are essential in all general learning and have become a key device in allowing children (Katz, 1963) and adults (Goldstone, Lippa, & Shiffrin, 2001) to effectively make sense of their surrounding world. Previous research concerning ontological boundaries provided evidence that suggests linguistic cues available in one's environment enhances the way world categories are distinctively perceived. The present mechanistic proposal hypothesizes that if a language supports both boundaries, then the cluster of correlations between perceptual and linguistic cues available in such language should readily aid in honoring all ontological categories. The present research addresses this question by examining how the Vietnamese language may honor different ontological boundaries and, more importantly, category formation.

## Categorical Structure and Ontological Distinctions

There are three different types of ontological categories in the world: Animate, Inanimate (Discrete), and Substances.

Distinctions between these categories depend on the magnitude of individuation. In the Individuation Continuum described by Lucy (1992), individuation occurs when an entity is conceptualized as bounded and discrete. In this continuum, animates lie at one end (more individualized) and substances at the other end (less individualized) of a continuous spectrum, with inanimates comprising the middle (i.e., animates——inanimates——substances). The likelihood that a particular entity is conceptualized as an individual varies systematically across the continuum from animates to substances (Lucy, 1992).

Learning associations for categorization that ties labels and meanings together facilitates relational judgments to be transferred to novel stimuli (Lupyan, Rakison & McClelland, 2007). For instance, when different labels were provided among the same exemplars, discrete differences are highlighted between the stimuli that affect one's judgment to separate items into different categories (Lupyan, Rakison & McClelland, 2007). A traditional approach to this issue has asked whether language is simply a symbol system that maps to all the relevant categories found in the world. Another view is that language creates and shapes human cognition (Whorf, 1956). People can interpret the world quite differently if they come from different language backgrounds and such differences have significant effects on the level of cognitive processing among these language learners (Cook, 1977; Bent, 2006). Research on the Japanese and English language, for instance, posits the importance of linguistic cues as a vital factor that couples available perceptual cues in the environment to facilitate perceptual regularities in the world (Imai & Gentner, 1997; Yoshida & Smith, 2005).

**Individuation in English** In English, individuation is frequently demonstrated by the count/mass distinction. Count nouns are nouns that can take the plural form (e.g. cups, cats), usually denoted with an *-s* after the noun. Thus, count nouns are conceptualized as discrete entities that are bounded and individualized. Mass nouns, however, are not pluralized (e.g., milk, water), but instead take continuous quantifiers (e.g., some, much). Thus, mass nouns are conceptualized as continuous entities that are unbounded and massed. For example, "*My cats (count noun) drank some (continuous quantifier) milk (mass noun) from the*

bowl” would be more grammatically correct than “My *many cat drank milks from the bowl.*” Although nouns may take both count and mass forms (e.g., “Would you like *some muffins?*”), English generally treats animates and objects as individuals, while substances are treated as masses. The likelihood of treating an object as an individual, therefore, drops markedly between objects and substances in English (Soja, Carey & Spelke, 1991; Soja, 1992; Imai & Gentner, 1997; Yoshida & Smith, 2003). See Figure 1 for an illustration. Further, solidity proves to be an important factor that highlights the contrastive nature between substances among other objects (Colunga & Smith, 2000). The key point here is that the English language privileges substances as continuous masses.

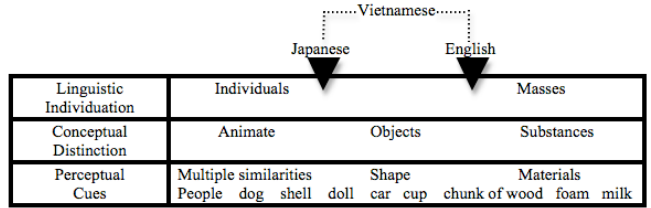


Figure 1: Three mutually dependent layers: linguistic, conceptual, and perceptual organization of the Boundary Shift Hypothesis. Linguistic individuation marks descriptive linguistic functions (e.g., lexical semantics/animacy, classifiers) available in each language—that is, *-iru* and *-aru* Japanese distinctions for animates; *mass/count* English distinctions for substances. Conceptual distinctions indicates the ontological categories—animates, inanimates/objects, and substances. Perceptual cues indicate the typical features associated with each ontological category provided in the real world. Different ontological distinctions for each language (Japanese and English; Yoshida & Smith, 2003) and the predicted boundaries for Vietnamese are illustrated.

**Individuation in Japanese** Japanese lexical and syntactic devices relevant to individuation are different from those in English (Yoshida, 2001). Japanese nouns that refer to multiple entities are not necessarily pluralized (e.g., the same expression can mean the same thing—“*there was a dog*” and “*there were many dogs*”). A particular plural suffix *-tachi* is never used with inanimate nouns (and is optional with animate nouns). There are unique quantifiers for animates, but those used for objects and substances form an overlapping set. The Japanese language also have separate ‘exists/is located’ verbs for animates and inanimates (*-iru* and *-aru* respectively). Thus, the likelihood of treating an object as an individual drops markedly between animates and objects in Japanese. See Figure 1. The key point here is that the Japanese language privileges animates as individuals (Yoshida & Smith, 2003).

In both cases, this can be considered as the consequence of different correlational patterns among the types of linguistic cues available in each language. Recent studies have taken such measure by providing a mechanistic approach in exploring the role of language in category

formations through the Boundary Shift Hypothesis (e.g., Yoshida & Smith, 2003, 2005; Hidaka & Saiki, 2004).

### The Boundary Shift Hypothesis

A mechanistic approach towards the formation of categorical organizations could, perhaps, be explained by the Boundary Shift Hypothesis (Yoshida & Smith, 2003). As introduced by Yoshida & Smith (2003), “ontological partitions” individuates the boundaries through specification of categorical concepts among the three distinct psychological forms (i.e., different kinds of existence) that serve as a foundation for human category learning (i.e., animals/animates, object/inanimates, and substance). Each category has its own set of perceived characteristics and children are able to categorize novel objects based upon its perceptual traits (Landau, Smith & Jones, 1988). Based on previous studies, it has been suggested that when children are presented with an object with eyes and/or limbs and a novel name, they are likely to select different objects that have the same shape and texture, thereby strengthening a category based on animate features (Yoshida & Smith, 2003). However, children are likely to form categories based on the same shape when objects are solid, angular, and made-up of multiple parts (Yoshida & Smith, 2003). The Boundary Shift Hypothesis explains ontological partitions by advocating the view that the language one learns influences or shifts the boundaries of the ontological space of objects and substances (Yoshida & Smith, 2003). Namely, this view suggests that categorization may be due to the correlational structure presented in the world (Samuelson & Smith, 1999). The cluster of correlations between perceptual and linguistic cues relevant to individuation enhances the perceptual characteristics of individualized entities and support formation of ontological categories (Yoshida & Smith, 2003). See Figure 2 and 3 for an illustration. Correlations among these perceptual cues and category structure, then, are systematically generalized by each language and differ accordingly among different language systems (i.e., consequences of different correlational patterns).

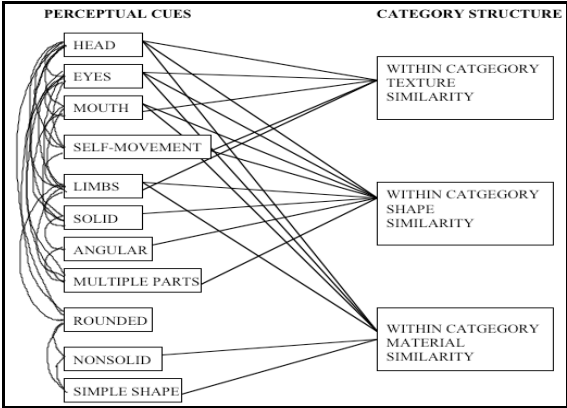


Figure 2: Illustration of associations between perceptual cues and category structure available in the world (Yoshida & Smith, 2003).



In this sense, language is viewed as a functional aspect that encompasses clusters of associations (i.e., weight of the correlations) that modifies the way we conceptualize categories. As demonstrated by Imai & Gentner (1993) and Yoshida & Smith (2003), psychological forms are conceptualized differently in English and Japanese. Thus, ontological categories are the products of learned correlations among the perceptual and linguistic cues.

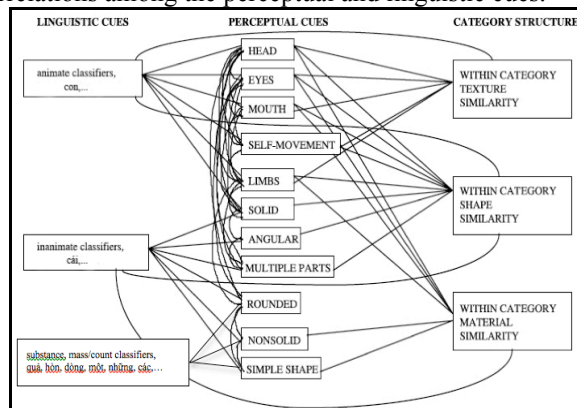


Figure 3: Associations among perceptual cues, category structure and linguistic cues available to learners of Vietnamese.

As a simple mapping system, language would honor the existing category structure and not influence category development. Thus, category development would be similar between children speaking different languages. The big question is, if language structure plays a vital role in shifting the boundaries between different ontological categories, would a language that have clear distinctions in representing different ontological structures demonstrate clear ontological partition (thus supporting the Boundary Shift Hypothesis)? All such theories predict that boundaries are formed from clusters of correlations; however, categorical knowledge is vast and may require more sophisticated ideas in explaining multiple categories. The purpose of the present study examines whether this type of boundary shifting (i.e., creating one boundary) is necessary for all children, or if children can reflect the reality of the three distinct categories. That is, the study explores the different ways in which language structure marks individuation by examining the Vietnamese language, which has implications in category formation through utilization of explicit classifiers to represent different ontological categories. If language does indeed influence category structure, we would expect to find a difference between the category formations of young children speaking different languages.

### Vietnamese Classifiers

Why Vietnamese? Vietnamese is a rich language that encompasses multitudes of explicit classifiers for speakers to conceptualize, classify, and describe spatial characteristics (shape, size, position) of objects in the surrounding world (Ly, 1999). Vietnamese classifiers are groups of nouns that have grammatical/syntax (Nguyen,

1963, 1975), semantic (Ly, 1999), and some implications in cognitive foundations (Lakoff, 1986; Friedrich, 1970). In the Vietnamese language, there are more than 40 different types of classifiers.

Vietnamese classifiers have two main functions: (1) singling out objects from different classes and, (2) help partition the world categories into various types (Ly, 1999). For example, in the sentence “con mèo” or “a cat” (CL+cat), the cat here is perceived as an individual animate object because of the classifier, whereas “cái ghế” or “a chair” (CL+chair) will be used to denote an individual inanimate object. In the Vietnamese language, the classifier “cái” is used most often for inanimate objects, while the classifier “con” indicates general animacy. Furthermore, classifiers describe explicitly the spatial characteristics of objects through the notion of salience and meaning. For instance, in English, the spherical feature is included in the meaning of the noun “ball” only implicitly. In Vietnamese, the same feature receives explicit expression by means of the classifier “quả/trái” (fruit/round-like), such as “quả/trái banh” or “a ball” (CL+ball).

Moreover, there are two types of classifiers: (1) Numerical (or non-descriptive) and (2) Descriptive. In numerical classifiers, an example would be “cái ghế” or “a chair” (CL+chair), which demonstrates that the classifier “cái” is indicating one chair. “Con”, however, may be used to describe inanimate and/or substances that are volitional in nature, such as “con sông” or “a river” (CL + river); “con dao” or “a knife” (CL + knife). When paired up with numerals, both classifiers may indicate count nouns. Without the use of “cái”/“con” preceding the noun, mass nouns would be implicated.

Additionally, there are certain distinctions for mass/count nouns in the Vietnamese language. For count nouns, singular forms are determined whether there is definite/limited size (e.g., “một cái bánh” or “a (one) piece of cake” numeral+CL+noun) or indefinite in size through the deletion of numerals (e.g., “cái bánh” or “piece of cake”; CL+noun). In a similar vein, plural count nouns are also dependent on definite/limited in size (e.g., “những cái bánh” or “some cake”; limited plural+CL+noun) and indefinite/maximal in size (e.g., “các cái bánh” or “every/all cake”; unlimited plural+CL+noun)—both of which can be viewed in parallel to the -s suffix that is added at the end of nouns in the English language. For mass or non-count nouns, however, it is not dependent on the size. Cao (1999) notes that mass nouns in Vietnamese are neutral to definiteness or non-definiteness. Where, in contrast, the zero article is used with a non-count noun (e.g., “bánh” or “cake” (zero or no CL+noun).

Given the richness of the descriptive language structure in Vietnamese, where would the Vietnamese language stand in regards to the Boundary Shift Hypothesis? That is, the linguistic boundary between individuals and nonindividuals perceptual boundaries between ontological kinds. Further, how do children come to understand the type of items or objects that are organized in different ways? Where does the

knowledge of different kinds of things emerge (i.e., animates, inanimates/objects, substances)? Japanese and English demonstrate homogenous differences in honoring two different ontological boundaries, is one or the other maximized or are both ontological distinctions present in the Vietnamese language? The present study hypothesizes that Vietnamese children should behave similarly to Japanese and English children—that is, the richness of the Vietnamese language should allow children to build distinct categorical formations for all ontological boundaries. See Figure 1. To test children’s knowledge of ontological categories, an adaptation of the Novel Noun Generalization (NNG) task was used (Soja, 1992). NNG tasks have been used to provide insight into children’s systematic expectations about how nouns map to distinct categories.

## Method

### Participants

Thirty monolingual Vietnamese participants with ages ranging from 23.85 to 33.22 months ( $M=29.59$ ,  $SD=2.91$ ) from Vietnam participated in the present study. Of the 30 participants, 20 completed the entire task and were therefore included in the analysis (attrition rate= 33.33%; 7 due to fussiness, 3 due to fatigue). Participants were recruited at a local preschool in Đồng Nai, Việt Nam. Prior to participation, all children were screened to ensure that Vietnamese was the only language they were regularly exposed to.

**Control.** Nine monolingual Japanese participants with ages ranging from 23.71 to 40.39 months ( $M=31.72$ ,  $SD=5.78$ ) from the USA (recently immigrated; temporary residents) participated in the present study for comparison results. Of the 9 participants, 8 participants completed the entire task (1 due to fussiness). Participants were recruited at a local Japanese daycare in Houston, TX. Primary caretakers were monolingual Japanese. Prior to participation, all children were screened on English and Japanese to ensure that Japanese was the only language they were regularly exposed to.

### Measurement Tools

A basic demographic questionnaire on language exposure and a parent checklist on productive vocabulary were used to ensure and control for homogeneity among the participants. To assess the children’s vocabulary, parents were asked to complete an adapted Vietnamese version of the MacArthur–Bates Communicative Development Inventories (MCDI; Fenson, Dale, Reznick, Bates, Hartung, Pethick, & Reilly, 1993). The Vietnamese version of the MCDI was developed by translating the American English (Fenson et al., 1993) and the Japanese MCDI (Ogura & Watamaki, 1997; see also Ogura, Yamashita, Murase, & Dale, 1993). Adult native speakers of Vietnamese translated and modified the documents. For the control group, the Japanese MCDI was used. The MCDI was used to control for vocabulary development among the participants.

### Procedure

Children sat at a comfortable distance from the computer screen in a quiet room at the preschool. A native Vietnamese experimenter sat next to the child and administered the task. A 5-minute break was implemented after the 27th trial (of 54 total trials) to reduce fatigue. Responses were recorded in-session by the experimenter.

### Task

The task was administered as a flash demonstration on a 15” HP laptop. There were a total of 54 trials consisting of 18 exemplars—6 animates, 6 inanimates, 6 substances—with 3 presentations each exemplar. Trials were presented in 6 blocks (i.e., 9 trials per block).

**Familiarization trials.** Children were presented with flash demonstrations of novel entities that were animate, inanimate, and substance. The objects were mixed and orders were randomized. All flash demonstrations began with the appearance of a novel object, followed by an animation of a hand acting upon the object.

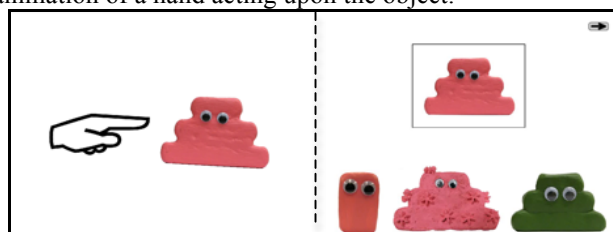


Figure 4: Animate. Testing choices matched on Color+Texture (CO+TX), Shape+Color (SH+CO), and Shape+Texture (SH+TX) from left-to-right respectively.



Figure 5: Inanimate. Testing choices matched on SH, TX, and CO from left-to-right respectively.

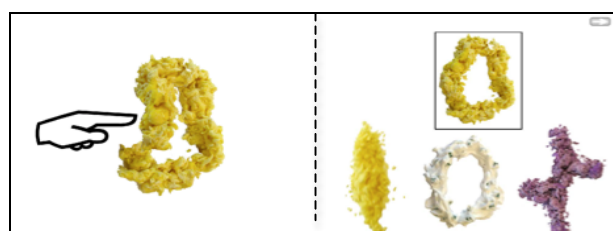


Figure 6: Substance. Testing choices matched on CO, SH, and TX from left-to-right respectively.

For animates, the novel entity had animate characteristics such as blinking eyes and volitional mannerisms (i.e., moved out of the screen from the incoming hand). For inanimates, novel objects were depicted without animate



cues (i.e., angular, curvature, solid blocks) and demonstrated static mannerisms (i.e., incoming hand moved the object out of the screen). For substances, the non-solidity feature was illustrated by manipulations from the hand (i.e., incoming hand changed the shape of the substance). See Figure 4-6 for an example of the animate, inanimate, and substance stimuli used. During each flash demonstration, the experimenter would introduce a new novel label attached to each entity (e.g., “*Này là Phoom. Em thấy không? Đây là Phoom đó!*”/“*This is a Foom (novel label). See? This is a Foom (novel label)!*”). Instructions were given in a neutral manner (i.e., no classifiers were given) to avoid biasing the child’s response.

**Testing trials.** After each demonstration, children were shown three testing choices and were asked to identify which of the new testing choices presented is called by the same label (e.g., “*Em chỉ cho chị, nào là Phoom?*”/“*Can you point to the Foom (novel label)?*”). Again, questions were given in a neutral manner to avoid biasing the child’s response. For animates, testing choices were matched on Shape-Texture (SH+TX), Shape-Color (SH+CO), and Color-Texture (CO+TX). For inanimates and substances, testing choices were matched on Shape (SH), Texture (TX), and Color (CO). See Figure 4-6 for an example of the testing choices. According to Jones & Smith (2002), adult judgment indicates that the expected answer choice for animates should be organized by similarities based on SH+TX, SH for inanimates, and TX for substances.

## Results

Replicating previous results (Imai & Gentner, 1997; Yoshida & Smith, 2001, 2003), Japanese monolingual participants (control group) significantly chose feature matched on SH+TX ( $t(7)=2.986$ ,  $p<.05$ ) for animates, SH ( $t(7)=1.225$ ,  $p<.05$ ) for inanimates, and SH ( $t(7)=3.666$ ,  $p<.05$ ) for substances.

As predicted for the Vietnamese participants, results demonstrate that they honored all ontological distinctions—animates, inanimates, and substances—suggesting the role of language in category structure.

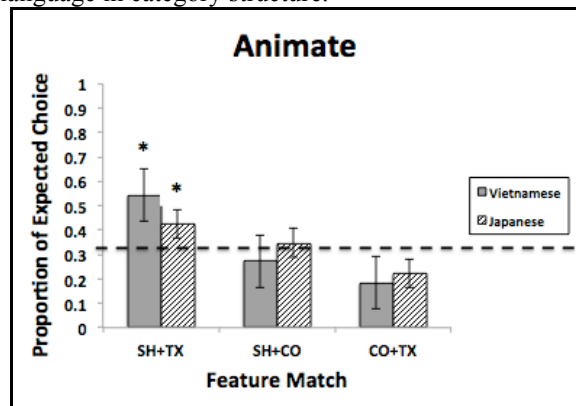


Figure 7: Proportion of expected choice for Animates matched on SH+TX, SH+CO, and CO+TX.

That is, Vietnamese participants significantly chose feature matched on SH+TX for animates, SH for inanimates, and TX for substances. Proportion of expected choices were performed against chance ( $p=.33$ ) using multiple t-tests. All expected choices were significantly above chance as illustrated by the star and dotted line (See Figure 7-9).

Specifically, in the *animate trials*, Vietnamese participants chose feature matched on SH+TX (expected) 53.89% of the time,  $t(19)=4.597$ ,  $p<.001$ , versus 28.61% for SH+CO match,  $t(19)=-.913$ ,  $p=.373$ , and 17.5% for CO+TX match,  $t(19)=-5.220$ ,  $p<.001$ . See Figure 7.

For the *inanimate trials*, Vietnamese participants chose feature matched on SH (expected) 56.38% of the time,  $t(19)=4.514$ ,  $p<.001$ , versus 20% for TX match,  $t(19)=-3.739$ ,  $p<.001$ , and 23.61% for CO match,  $t(19)=-2.021$ ,  $p=.058$ . See Figure 8.

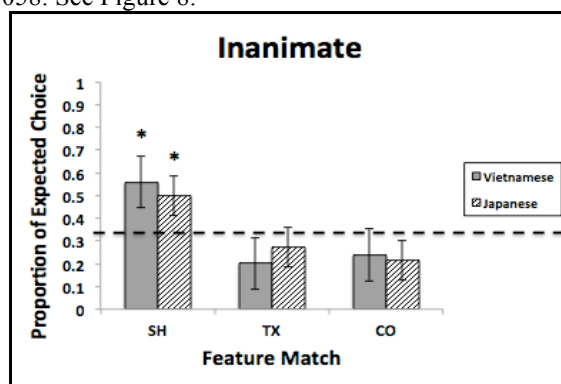


Figure 8: Proportion of expected choice for Inanimates matched on SH, TX, and CO.

Finally, in the *substance trials*, Vietnamese participants chose feature matched on TX (expected) 48.89% of the time,  $t(19)=3.620$ ,  $p<.01$ , versus 26.67% for SH match,  $t(19)=-1.468$ ,  $p=.158$ , and 25% for CO match,  $t(19)=-2.010$ ,  $p=0.59$ . See Figure 9.

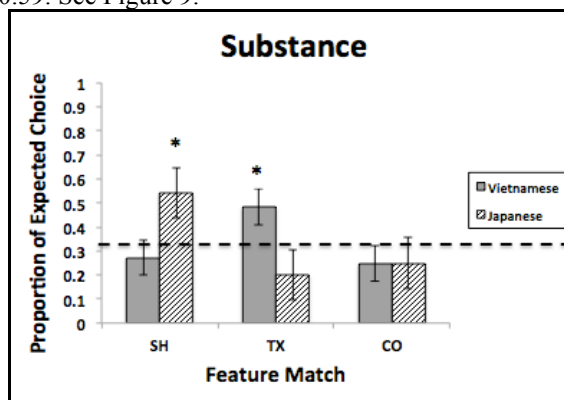


Figure 9: Proportion of expected choice for Substances matched on SH, TX, and CO.

## General Discussion

Results from the present study supports previous research (i.e., Boundary Shift Hypothesis) suggesting that the way in which children form categories depends largely on the language they are learning and the correlational patterns

they develop given the perceptual cues and regularities available in their environment. In particular, individuation among the perceptual boundaries between ontological kinds in the Vietnamese language is highly influenced by the availability and use of explicit classifiers within the language. Specifically, classifiers that highlight animates (i.e., *con*), inanimates (i.e., *cái*), and substance (i.e., numerical+CL inclusions for count nouns and deletion of classifiers for mass nouns) among the variety of classifiers help Vietnamese children to identify the discrete differences among different entities. Overtime, such regularities are produced to create clusters of correlations that allow children to form discrete ontological boundaries. This suggests that categorization is highly dependent on the structure of the language being learned, perceptual cues and regularities available in the environment, and the correlational pattern over time. The current results indicate that this phenomena is robust across tasks, regardless of task variations (Soja, Carey & Spelke, 1991; Soja, 1992; Imai & Gentner, 1997; Yoshida & Smith, 2003), among a variety of languages that foster distinct ontological boundaries. Therefore, we expect that English monolingual children should behave similarly from previous literature (Soja, Carey & Spelke, 1991; Yoshida & Smith, 2003). In sum, how children form categories may depend largely on the language they are learning and, in particular, on the way that language individuates kinds.

### Acknowledgments

Support for this research was provided in part by the National Institutes of Health grant (R01 HD058620), the Foundation for Child Development, and the University of Houston's Enhance and Advance Research (GEAR) program. We would like to extend our appreciation to the parents, teachers, and children who participated in the study. Special thanks are also due to Lauren Baker who created and designed the stimuli used in the present study.

### References

- Bent, T., Bradlow, A.R., & Wright, B.A. (2006). The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 32(1), 97–103.
- Cao, X.H. (1999). *Tieng Viet: May van de ngu am, ngu phap, ngu nghia*. Hanoi: Nha xuất bản Giáo dục.
- Colunga, E. & Smith, L. (2000) Committing to an Ontology: A connectionist Account. Talk presented at the 22nd Annual Meeting of the Cognitive Science Society.
- Cook, V.J. (1977). Cognitive processes in second language learning. *International Review Applied Linguistics*, XV/1, 73-90. Reprinted in D. Nehls (ed.), *Studies in Language Acquisition*. Julius Groos, 1980.
- Goldstone, R.L., Lippa, Y., & Shiffrin, R.M. (2001). Altering object representations through category learning. *Cognition*, 78, 27–43.
- Hidaka, S. & Saiki, J. (2004). A mechanism of ontological boundary shifting. Talk at the 26th Annual Meeting of the Cognitive Science Society.
- Fenson, L., Dale, P., Reznick, J.S., Bates, E., Hartung, J., Pethick, S., & Reilly, J. (1993). *MacArthur Communicative Development Inventories*. San Diego, CA: Singular.
- Friedrich, R. (1970). Shape in grammar. *Language*, 46, 379–407.
- Imai, M., & Gentner, D. (1997). A cross-linguistic study of early word meaning: universal ontology and linguistic influence. *Cognition*, 62, 169–200.
- Katz, P. A. (1963). Effects of labels on children's perception and discrimination learning. *Journal of Experimental Psychology*, 66, 423–428.
- Lakoff, G. (1986). Classifiers as a reflection of mind. In *Noun Classes and Categorization*, ed. C. Craig. Amsterdam.
- Landau, B., Smith, L.B. & Jones, S.S. (1988). The importance of shape in early lexical learning, *Cognitive Development*, 3, 299–321.
- Lucy, J.A. (1992). *Language diversity and thought: A reformulation of the linguistic relativity hypothesis*. Cambridge: Cambridge University Press.
- Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Labels facilitate learning of novel categories. *Psychological Science*, 18(12), 1077–1083.
- Ly, T. T. (1999). Representation of space in Vietnamese classifiers. *Mon-Khmer Studies*, 29, 71–80.
- Nguyen, T.C. (1963). *K coprosu o klassifikatorax vo Cietnamskom jazyke*. Filologii stran Vostoka. Leningrad.
- Nguyen, T.C. (1975). *Tu loai danh tu trong tieng Viet hien dai*. Ha Noi.
- Samuelson, L., and Smith, L. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition*, 73, 1–33.
- Soja, N.N., Carey, S., & Spelke, E.S. (1991). Ontological categories guide young children's inductions of word meaning: object terms and substance terms. *Cognition*, 38, 179–211.
- Soja, N. (1992). Inferences about the meanings of nouns: the relationship between perception and syntax. *Cognitive Development*, 7, 29–46.
- Whorf, B.L. (1956). "The Relation of Habitual Thought and Behavior to Language." In J.B. Carroll (ed.) *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf* (pp. 134–159). Cambridge, MA: MIT Press.
- Yoshida, H. & Smith, L.B. (2001). Early noun lexicons in English and Japanese. *Cognition*, 82, 63–74.
- Yoshida, H. & Smith, L.B. (2003). Shifting ontological boundaries: How Japanese- and English- speaking children generalize names for animals and artifacts. *Developmental Science*, 6, 1–34.
- Yoshida, H. & Smith, L.B. (2003). Correlation, concepts and cross-linguistic differences. *Developmental Science*, 6(1), 30–34.
- Yoshida, H. & Smith, L.B. (2005). Linguistic cues enhance the learning of perceptual cues. *Psychological Science*, 16(2), 90–95.

# Writing facilitates learning to read in Chinese through reduction of holistic processing: A developmental study

Ricky Van Yip Tso ([richie13@hku.hk](mailto:richie13@hku.hk))

Terry Kit-fong Au ([terryau@hku.hk](mailto:terryau@hku.hk))

Janet Hui-wen Hsiao ([jhsiao@hku.hk](mailto:jhsiao@hku.hk))

Department of Psychology, University of Hong Kong  
604 Knowles Building, Pokfulam Road, Hong Kong SAR

## Abstract

Holistic processing has been identified as an expertise marker of face and object recognition. In contrast, the expertise marker of recognizing Chinese characters is reduced holistic processing (Hsiao & Cottrell, 2009), which is driven by Chinese writing experiences rather than reading ability (Tso, Au, & Hsiao, 2011). Here we investigate the developmental trend of holistic processing in Chinese character recognition and its relationship with reading and writing abilities by testing Chinese children who were learning Chinese at a public elementary school in Hong Kong on these abilities. We found that the holistic processing effect of Chinese characters in children was reduced as they reached higher grades; this reduction was driven by enhanced Chinese literacy rather than age. In addition, we found that writing performance predicts reading performance through reduced holistic processing as a mediator. We thus argue that writing hones analytic processing, which is essential for Chinese character recognition, and in turn facilitates learning to read in Chinese.

**Keywords:** Chinese character recognition, holistic processing, reading, writing, copying

## Introduction

Holistic processing (HP)—the ability to process separate features as a single whole unit—is an expertise marker for face and object recognition (see e.g., Bukach, Gauthier, & Tarr, 2006; Gauthier & Bukach, 2007; Richler, Wong, & Gauthier, 2011; though some argue that it is specific to faces, e.g., McKone, Kanwisher, & Duchaine, 2007). Chinese characters share many visual properties with faces. In contrast to words in most alphabetic languages that are linear in structure and consist of letter series of varying length, the Chinese writing system is logographic—The configuration of Chinese characters is more homogenous and square, and each character is a grapheme that represents a morpheme (Shu, 2003; Wong & Gauthier, 2006). The basic units of a Chinese character are strokes that combine to create more than a thousand different stroke patterns in the Chinese writing system; these stroke patterns form the characters (Hsiao & Shillock, 2006). A typical literate recognizes more than 3,000 individual Chinese characters regardless of variations in font (Hsiao & Cottrell, 2009). This is similar to face recognition in which faces are recognized individually regardless of variations in facial expressions (Hsiao & Cottrell, 2009; Wong & Gauthier, 2006). Despite the similarity between Chinese characters and faces, the expertise marker for Chinese character recognition is reduced HP (Hsiao & Cottrell, 2009).

Experienced Chinese readers employ less HP than novices in perceiving Chinese characters; this may be because Chinese readers are more sensitive to the internal constituent components of Chinese characters. They can readily ignore some configural information, such as exact distances between features, which are unimportant for character recognition (Ge, Wang, McGleery, & Lee, 2006). In contrast, these internal constituent components may not look easily separable to novices as they are less able to distinguish individual features and components in Chinese characters (Chen, Allport, & Marshall, 1996; Ho, Ng, & Ng, 2003; Hsiao & Cottrell, 2009). In addition, reduced holistic processing (i.e., analytic processing) of Chinese characters is enhanced by Chinese writing experiences (Tso, Au, & Hsiao, 2011). In Tso and colleagues' (2011) study, two groups of Chinese readers were tested: Chinese literates who could read and write (i.e., Writers), and Chinese literates who had limited writing exposure and thus had reading performance far better than writing performance (i.e., Limited-writers). Limited-writers had reading performance comparable to Writers', yet Limited-writers had far poorer performance than Writers in a dictation task (i.e., recall and write down a Chinese word when instructed). Writers perceived Chinese characters less holistically than Limited-writers—this between-group difference in HP could mainly be explained by dictation (writing) performance when the reading and copying variables were statistically controlled.

In Hong Kong, elementary schools do not explicitly place emphasis in its curriculum on teaching the Chinese character radicals (i.e., character components that consist of one or more identifiable stroke patterns); yet, children become more aware of the internal orthographic components in Chinese characters as they progress to higher grades (Ho et al., 2003). This may be explained by motor programming through extensive copying and writing as a requirement in Chinese lessons at school (Guan, Liu, Chan, Ye, & Perfetti, 2011; Tan, Spinks, Eden, Perfetti, & Siok, 2005). Reading performance is significantly predicted by copying ability (Chan, Ho, Tsang, Lee, & Chung, 2006; McBride-Chang, Chung, & Tong, 2011; Tan et al., 2005), as well as dictation performance (McBride-Chang et al., 2011; Tse, Kwan, & Ho, 2010). Writing experience may enhance reading ability because children may consolidate knowledge of graphomotor memory of character strokes as they copy the stroke patterns (Tan et al., 2005; Tse et al., 2010). Learning to write was experimentally shown to strengthen Chinese

character recognition (Guan et al., 2011). Neuroimaging studies also suggested that writing experience plays an important role in shaping reading-specialized neural representations (James & Atwood, 2009; Longcamp, Anton, Roth, & Velay, 2003; Siok, Perfetti, Jin, & Tan, 2004). All these results suggest a close relationship between sensory-motor integration development through writing practice and the development of reading skills, particularly for recognizing Chinese characters.

Although previous studies have suggested a close relationship between Chinese writing and reading performance in children, the underlying mechanism remains unclear (e.g., Guan et al., 2011; McBride-Chang et al., 2011)<sup>1</sup>. As reduction in HP of Chinese characters marks expert-level character recognition, here we hypothesize that writing experience enhances character recognition performance by modulating the perceptual system, allowing readers to identify Chinese characters more analytically (Tso et al., 2011). The modulating effect of writing experience on our perception of Chinese characters has never been studied before in Children who are learning to read and write Chinese characters. Here, we investigate whether children in upper grades perceived characters less holistically than children in lower grades in an elementary school where the Chinese language is taught. We also examine their Chinese reading and writing performance to see what can predict Children's reduced HP of Chinese characters. We predict that upper-grade children (who should have better Chinese literacy than lower-grade children) will process Chinese characters more analytically (i.e., less holistically) than children in lower grades. Because of the direct relationship writing experiences have with reduced HP (Tso et al., 2011), we hypothesize that children's reduction in HP can be predicted by writing performances across grades. Since writing performance also strongly correlates with reading abilities in Chinese (see e.g., Guan et al., 2011; McBride-Chang et al., 2011; Tan et al., 2005), we hypothesize that HP mediates between Chinese reading and writing abilities in children. More specifically, we predict that writing experience leads to reduced holistic processing in Chinese characters, which in turn enhances reading abilities in Chinese (Fig. 1)

Tso and colleagues (2011) also suggested that writing experience may facilitate reading Chinese characters in an unfamiliar font, as Limited-writers had difficulty reading words in the Feng font (a font that mimics handwriting and was unfamiliar to the participants) whereas this effect in Writers was minimal. Hence here we also examine the possible effect of enhanced Chinese writing proficiency on naming characters and words in familiar and unfamiliar fonts.

<sup>1</sup> Although some studies have proposed that writing enhances graphomotor memory of Chinese characters, which in turn facilitates reading (e.g., Tan et al., 2005; Tse et al., 2010). Yet, this hypothesis has not been statistically tested.



Fig. 1. Predicted mediation effect of reduced holistic processing between Chinese writing and reading performance.

## Methods

### Participants

56 first grade (mean age = 5.88 years, SE = .051), 73 third grade (mean age = 7.90, SE = .056), and 88 fifth grade (mean age = 9.89, SE = .047) Chinese children from an elementary in Hong Kong participated in our study. They were all Cantonese native-speaking and were all receiving regular Chinese language curriculum at school. All of them had normal or corrected-to-normal vision.

### Procedures

#### Test for holistic processing

To test for HP effects, we adopted procedures from Hsiao and Cottrell (2009). 160 pairs of medium to high frequency Chinese characters in Ming font were chosen; 80 pairs had a top-bottom configuration and 80 pairs had a left-right configuration<sup>2</sup> (Fig. 2). The frequency information of the Chinese characters was obtained from Ho and Kwan (2001). The top-down and left-right characters were matched in stroke number and frequency. In each trial, children were presented with two characters and instructed to attend to only half (either top or bottom for top-bottom characters, or left or right for left-right characters) of each character and judge whether they were the same or different. Forty pairs were presented in each of the four conditions (Fig. 3a): *same in congruent trials*, *different in congruent trials*, *same in incongruent trials*, and *different in incongruent trials*. The complete composite paradigm (Gauthier & Bukach, 2007) was adopted so that in congruent trials, the attended and irrelevant halves corresponded to the same response (i.e., both were the same or different) while in incongruent trials, the attended and irrelevant halves corresponded to different responses (e.g., the top halves were the same while the bottom halves were different). We adopted this paradigm to avoid response biases that may occur in the partial composite design in which the irrelevant halves would always be



Fig. 2. Examples of Chinese characters with left-right configuration (left) and top-bottom configuration (right).

different (see Gauthier & Bukach, 2007; Robbins & McKone, 2007; Richler, Cheung, & Gauthier, 2011).

<sup>2</sup> Both left-right and top-bottom are common Chinese character structures. Left-right is a more dominant structure than top-bottom in the Chinese lexicon.



In each trial, after 1,000 ms of central fixation, participants were cued with a symbol that indicated which half of each character they should attend to. The pair of characters was then presented, with one above and one below the initial fixation point, followed by a mask. During the 500ms presentation time, children looked at each character once and responded as quickly and accurately as possible by pressing corresponding buttons to judge if the character parts were the same or different (Fig 3b). Accuracy was collected. We measured participants' discrimination sensitivity  $A'$  as:

$$A' = 0.5 + \left[ \text{sign}(H - F) \frac{(H - F)^2 + |H - F|}{4 \max(H, F) - 4HF} \right]$$

$H$  and  $F$  are the hit rate and false alarm rate, respectively.  $A'$  is a bias-free nonparametric measure of sensitivity; we did not use  $d'$  because response biases may affect its measurement when assumptions of normality and homogeneity of variance are not met (Stanislaw & Todorov, 1999). The  $A'$  difference between incongruent and congruent trials (i.e., Holistic  $A'$ ) measures HP—a more positive value marks a stronger HP effect.

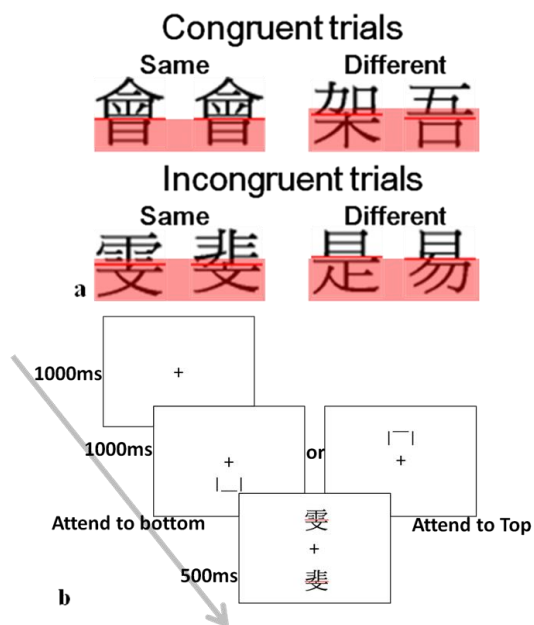


Fig. 3. Illustration of stimulus pairs in the complete composite paradigm and trial sequences (for top-bottom characters). In (a), the attended components are shaded in red. In (b), a 1,000 ms central fixation cross precedes each trial, followed by a cue either below or above the cross to indication which half (top or bottom) of the characters participants should attend to in the following display.

#### Tests for reading and writing performance:

Four tests were administered: 1. Character naming task, 2. Word naming task, 3. Character copying task, and 4. Word dictation task. Tasks 1 and 2 assessed participant's reading ability, while Tasks 3 and 4 assessed their copying and word recalling/writing ability respectively.

#### Reading tests:

##### 1. Character naming task:

Children were presented with 60 Chinese characters one at a time. Half of the stimuli were presented in Ming font (a commonly used font type; Fig. 4a) and the other half were presented in Feng font (an unfamiliar font type that simulates handwriting; Fig. 4b). They were instructed to read aloud the characters as quickly and as accurately as possible. The characters were arranged from high to low frequency (frequency information for primary students was obtained from Leung & Lee, 2001). The trials stopped after 5 consecutive errors made. Each trial started with a central fixation cross for 500ms, followed by the character presentation. The screen turned blank after a child had responded and the experimenter pressed a button to record the accuracy and to start the next trial. Their response time was measured as the time difference between the stimulus onset and the onset of the pronunciation, detected by a microphone.

##### 2. Word naming task:

Children read aloud 30 two-character words arranged from high to low frequency (frequency information for primary students was obtained from Leung & Lee, 2001) as quickly and accurately as possible. Half of the stimuli were presented in Ming font and the other half were presented in Feng font. The procedure was the same as that of the character naming task. The response time was measured as the time difference between the stimulus onset and the onset of the pronunciation of the first character.



Fig. 4. Examples of a Chinese character in Ming font (a) and Feng font (b).

#### Writing tests:

##### 3. Character copying task:

Children copied 30 characters (10 real characters, 10 pseudo-characters, and 10 Korean characters) as quickly and as accurately as possible. The Chinese characters were randomly selected from the characters used in task 1. The pseudo-characters were orthographically legal but non-sense characters. Each trial started with a central fixation cross for 500 ms, followed by the character presentation. Each time, after a child had copied a character, the experimenter pressed a button immediately and the screen turned blank to start the next trial. Their response time was recorded.

##### 4. Word dictation task:

Children wrote down 30 two-character words (the same words used in task 2) as quickly and as accurately as possible when they heard each word said in a female voice presented by a computer. Two-character words were used instead of characters to reduce ambiguity due to the many homophonic characters in the Chinese lexicon. Each trial started with the words "Get ready" on the screen for 500 ms. After hearing the word, participants pressed corresponding buttons to

indicate whether they could recall the word or not, before they started writing. After they finished writing, the experimenter pressed a button to indicate accuracy and to reveal the next word. Accuracy rate was recorded.

These experiments were all conducted using E-prime v2.0 (Psychology Software Tools, Pittsburgh, PA).

## Results

Repeated-measures ANOVA was used to investigate HP effects (congruency: congruent vs. incongruent trials x grade: Grade 1 vs. Grade 3 vs. Grade 5). We found a significant effect of grade ( $F(2, 213) = 25.090, p < 0.001$ ), a significant effect of congruency ( $F(1, 213) = 268.319, p < 0.001$ ), and an interaction between congruency and grade ( $F(1, 213) = 11.376, p < 0.001$ ). The main effect of congruency suggests that across grades, children process Chinese characters holistically. While the main effect of grade showed that the performance level increased with grade, the interaction between grade and congruency suggests that children processed Chinese characters with varying levels of congruency effect across grades (Fig. 5).

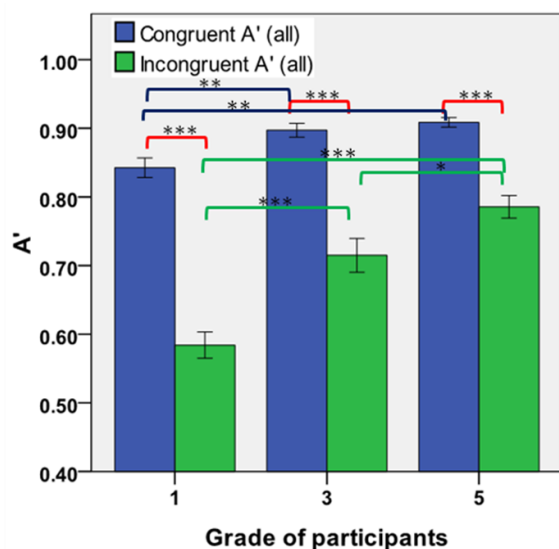


Fig. 5. A' of congruent and incongruent trials for first, third and fifth graders in the holistic processing task (\* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ ).

Pairwise post-hoc t-tests showed that A' in congruent trials was larger than in incongruent trials in first ( $t(55) = 12.01, p < .001$ ), third ( $t(72) = 7.838, p < .001$ ) and fifth graders ( $t(86) = 8.439, p < .001$ ). In congruent trials, first graders had a smaller A' than third ( $t(127) = 3.226, p < .01$ ) and fifth graders ( $t(141) = 4.576, p < .001$ ) while third and fifth graders did not differ statistically in A' ( $t(158) = .994, n.s.$ ). In incongruent trials, first graders had a smaller A' than third ( $t(127) = 3.991, p < .001$ ) and fifth graders ( $t(141) = 7.878, p < .001$ ), and third graders had a smaller A' than fifth graders ( $t(158) = 2.450, p < .05$ ). We also conducted pairwise post-hoc t-tests on the A' difference between incongruent and

congruent trials (i.e., Holistic A') between children in the 3 grades. We found that first graders had a larger Holistic A' than third graders ( $t(127) = 2.334, p < .05$ ) and fifth graders ( $t(141) = 5.401, p < .001$ ), while third graders had a larger Holistic A' than fifth graders ( $t(158) = 2.230, p < .05$ ). These results suggest that HP of Chinese characters in children was reduced as they reached higher grades (see Fig. 5).

## Chinese proficiency tests

Pearson's correlation regression analysis showed a positive correlation between grade and character-naming accuracy ( $r^2 = .602, p < .001$ ), word-naming accuracy ( $r^2 = .512, p < .001$ ) and dictation accuracy ( $r^2 = .707, p < .001$ ); and a negative correlation between grade and character-naming response time ( $r^2 = .269, p < .001$ ), word-naming response time ( $r^2 = .347, p < .001$ ) and character-copying response time ( $r^2 = .620, p < .001$ ). These results suggest that children had better Chinese reading and writing proficiency as they reached higher grades.

## Font familiarity effect on character and word naming

Repeated-measures ANOVA (font: Ming vs. Feng x grade) was used for the analysis on character-naming response time. We found a main effect of font ( $F(1,181) = 16.9, p < .001$ ) and a main effect of grade ( $F(2,181) = 37.01, p < .001$ ), but no interaction effect was found between font and grade ( $F(2,181) = .348, n.s.$ ).

For word-naming response time, we found a main effect of font ( $F(1,177) = 17.1, p < .001$ ) and a main effect of grade ( $F(2,177) = 59.6, p < .001$ ). We also found an interaction

Table 2. Hierarchical regression analysis among holistic processing and reading and copying performance

Predicted Variable: Holistic A'		
Age vs. Character Naming (RT & Accuracy)		
Steps	Variables	$\Delta r^2$
1	Age	.110***
2	Character Naming	.037*
1	Character Naming	.141***
2	Age	.005
Age vs. Word Naming (RT & Accuracy)		
1	Age	.099***
2	Word Naming	.056**
1	Word Naming	.147***
2	Age	.008
Age vs. Copying (RT)		
1	Age	.104***
2	Copying	.001
1	Copying	.067***
2	Age	.038**
Age vs. Dictation (Accuracy)		
1	Age	.155***
2	Dictation	.014 <sup>1</sup>
1	Dictation	.118***
2	Age	.011

<sup>1</sup>  $p < 0.1$  \* $p < 0.05$  \*\* $p < 0.01$  \*\*\* $p < 0.001$

effect between font and grade ( $F(2,177) = 3.894, p < .05$ ). We then performed pairwise post-hoc t-tests on the response time difference between Ming font words and Feng font words among children in the 3 grades. We found that first graders had a larger font effect in response time than third graders ( $t(109) = 1.941, p = .055$ ) and fifth graders ( $t(117) = 2.192, p < .05$ ) while third and fifth graders did not differ statistically ( $t(128) = .135, n.s.$ ). These results suggest that word-naming performances depended greatly on font familiarity in lower grades than in upper grades. This font-grade interaction was not found in character-naming.

#### Hierarchical Regression Analysis

We investigated how HP could be uniquely predicted by literacy level by partialing out the variance due to age. As summarized in Table 2, HP was predicted uniquely by reading and writing performance and vice versa. The variance of HP can be significantly explained by reading and dictation performances, but not copying performance, when partialing out the variance due to age.

#### Mediation Analysis

We conducted a mediation analysis to test the mediation effect of HP between dictation accuracy and word naming accuracy<sup>3</sup>. Regression analysis revealed that Holistic A' predicted word naming accuracy ( $\beta = -.265, p < .01$ ). A Sobel mediation test showed that Holistic A' significantly mediated the relationship between dictation accuracy and word naming accuracy ( $z = 2.403, p < .05$ ). The mediator effect of Holistic A' was partial as the direct effect of dictation accuracy and word naming accuracy remained statistically significant,  $\beta = .635, p < .001$  (Fig. 6).

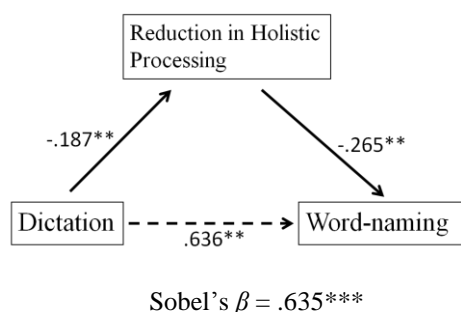


Fig 6. Partial mediation effect of reduced holistic processing on dictation and word naming performances (\* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ ).

### **Discussion**

Our study showed that children in upper grades processed Chinese characters less holistically than children in lower grades. Further analyses showed that this effect could not be solely accounted for by age. Consistent with Hsiao and Cottrell's (2009) findings, better Chinese reading and writing proficiency strongly predicted reduced HP. Because

elementary schools in Hong Kong do not explicitly teach children Chinese character radicals (Ho et al., 2003), their reduction in HP in upper grades is more likely the result of enhanced Chinese literacy.

Previous studies had reported a close relationship between Chinese writing and reading performance. Yet, the underlying mechanism remained unclear (e.g., Guan et al., 2011; McBride-Chang et al., 2011; Tan et al., 2005; Tse et al., 2010). In our study, the mediation analysis suggests that HP is a significant mediator of dictation accuracy and word naming accuracy (Fig. 6). This result is consistent with our hypothesis that the HP effect in Chinese character recognition is predicted by writing performance (i.e., the ability to recall and write down Chinese characters), and in turn predicts word reading performance (Fig. 1). Perhaps writing experience enhances the ability to analyze the orthographic structures and components of Chinese characters in children, which leads to reduced HP (i.e., more analytic) as suggested in Tso and colleagues' (2011) study. Consistent with previous findings that suggest our sensorimotor learning influences our perception (He et al., 2003; James & Atwood, 2009; Longcamp et al., 2003), we show how writing performance can be associated with reduced HP in Chinese character recognition. In addition, we show that this change in the perception of Chinese characters (reduced HP) can in turn modulate reading performance. This study is also the first to report on the developmental trajectory of HP of Chinese characters in children.

Our results also revealed that children were better at recognizing Chinese characters and words in a familiar font (Ming) than in an unfamiliar font (Feng); this effect of font familiarity was more prominent in first graders than third and fifth graders. ANCOVA analysis showed that the group difference in the font familiarity effect became insignificant when dictation accuracy was set as a covariate ( $F(2,172) = 1.84, n.s.$ ). This suggests that writing experience, measured as the ability to recall and write characters here, facilitates reading words in an unfamiliar font (Tso et al., 2011). Further investigation, however, is needed to see whether this effect is due to a particular dimension or general improvement in literacy.

As future work, we will obtain information of HP of Chinese characters from non-Chinese speaking children that do not receive the local Chinese curriculum as a control group. We speculate that the pattern of their holistic processing of Chinese characters will not drop as much as Chinese-learning children as they do not learn to write Chinese explicitly.

In conclusion, our study provides an information processing account of the relationship between writing and reading ability in Chinese children. We show that children learning to read and write Chinese characters processed Chinese characters with reduced HP as they reached upper grades; this reduction in HP may be facilitated by writing experience and may in turn enhance word reading ability.

<sup>3</sup> This mediation pathway with word-naming accuracy as the first and dictation accuracy as the final step was not significant.



## Acknowledgments

We are grateful to the Research Grant Council of Hong Kong (project code: HKU 745210H to J.H. Hsiao) and the HKU Foundation (Seed Grant #10401359 to J.H. Hsiao).

## References

- Bukach, C. M., Gauthier, I., & Tarr, J. M. (2006). Beyond faces and modularity: The power of an expertise framework. *Trends in Cognitive Sciences*, 10, 156-166.
- Chan, D. W., Ho, C. S.-H., Tsang, S.-m., Lee, S.-h., & Chung, K. K. H. (2006). Exploring the reading-writing connection in Chinese children with dyslexia. *Reading and Writing*, 19, 543-561.
- Chen, Y. P., Allport, D. A., & Marshall, J. C. (1996). What are the functional orthographic units in Chinese word recognition: The stroke or the Stroke pattern? . *Q. J. Exp. Psychol-A*, 49, 1024-1043.
- Gauthier, I., & Bukach, C. (2007). Should we reject the expertise hypothesis? *Cognition*, 103, 322-330.
- Ge, L., Wang, Z., McGleery, J. P., & Lee, K. (2006). Activation of face expertise and the inversion effect. *Psychological Science*, 17, 12-16.
- Guan, C. Q., Liu, Y., Chan, D. H. L., Ye, F., & Perfetti, C. A. (2011). Writing strengthens orthographic and alphabetic-coding strengthens phonology in learning to read Chinese. *Journal of Educational Psychology*, 103(3), 509-522.
- He, A. G., Tan, L. H., Tang, Y., James, A., Wright, P., & Eckert, M. A. (2003). Modulation of neural connectivity during tongue movement and reading. *Human Brain Mapping*, 18, 222-232.
- Ho, C. S., & Kwan, T. W. (2001). Hong Kong, Mainland China & Taiwan: Chinese character frequency-A trans-regional, diachronic survey. Retrieved December 30, 2011, from Character Frequency: <http://arts.cuhk.edu.hk/Lexis/chifreq/>
- Ho, C. S., Ng, T., & Ng, W. (2003). A radical approach to reading development in Chinese: The role of semantic radicals and phonetic radicals. *Journal of Literacy Research*, 35, 849-878.
- Hsiao, J. H., & Cottrell, G. (2009). Not all visual expertise in holistic, but it may be leftist: The case of Chinese character recognition. *Psychological Science*, 20(4), 455-463.
- Hsiao, J. H., & Shillock, R. (2006). Analysis of a Chinese phonetic compound database: Implications for orthographic processing. *Journal of Psycholinguistic Research*, 35, 405-426.
- James, K. H., & Atwood, T. P. (2009). The role of sensorimotor learning in the perception of letter-like forms: Tracking the causes of neural specialization for letters. *Cognitive Neuropsychology*, 26, 91-110.
- Leung, M. T., & Lee, A. W. Y. (2001). The Hong Kong corpus of primary school Chinese. Paper presented at the 9th meeting of the International Clinical Phonetics and Linguistics Association. Hong Kong.
- Longcamp, M., Anton, J. L., Roth, M., & Velay, J. L. (2003). Visual presentation of single letters activates a premotor area involved in writing. *NeuroImage*, 19, 1492-1500.
- McBride-Chang, C., Chung, K. K. H., & Tong, X. (2011). Copying skills in relation to word reading and writing in Chinese children with and without dyslexia. *Journal of Experimental Child Psychology*, 110, 422-433.
- McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences*, 11, 8-15.
- Richler, J. J., Cheung, O. S., & Gauthier, I. (2011). Beliefs alter holistic face processing ... if response bias is not taken into account. *Journal of Vision*, 11(13):17, 1-13.
- Richler, J. J., Wong, Y. K., & Gauthier, I. (2011). Perceptual expertise as a shift from strategic interference to automatic holistic processing. *Current Directions in Psychological Science*, 20(2), 129-134.
- Robbins, R., & McKone, E. (2007). No face-like processing for objects of expertise in three behavioural tasks. *Cognition*, 103, 34-79.
- Shu, H. (2003). Chinese writing system and learning to read. *International Journal of Psychology*, 38, 274-285.
- Siok, W. T., Perfetti, C. R., Jin, Z., & Tan, L. H. (2004). Biological abnormality of impaired reading is constrained by culture. *Nature*, 431, 71-76.
- Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments & Computers*, 31, 137-149.
- Tan, L. H., Spinks, J. A., Eden, G. F., Perfetti, C. A., & Siok, W. T. (2005). Reading depends on writing, in Chinese. *PNAS*, 102, 8781-8785.
- Tse, H. M., Kwan, D. P., & Ho, C. S. (2010). *Should dyslexic children stop copying? Exploring the relationship between word reading, copying and dictation*. Paper presented at the Annual International Academy for Research in Learning Disabilities Conference.
- Tso, R. V. Y., Au, T. K., & Hsiao, J. H. (2011). *The influence of writing experiences on holistic processing in Chinese character recognition*. Paper presented at the Thirty-Third Annual Conference of the Cognitive Science Society, Boston.
- Wong, A. C., & Gauthier, I. (2006). An analysis of letter expertise in a levels-of-categorization framework. *Visual Cognition*, 15, 854-879.

# Temporal Dynamics of Action Perception: The Role of Biological Appearance and Motion Kinematics

**Burcu Aysen Urgan** ([burgen@cogsci.ucsd.edu](mailto:burgen@cogsci.ucsd.edu))

Department of Cognitive Science, University of California, San Diego  
9500 Gilman Drive, La Jolla, CA 92093 USA

**Markus Plank** ([mplank@ucsd.edu](mailto:mplank@ucsd.edu))

Institute for Neural Computation, University of California, San Diego  
9500 Gilman Drive, La Jolla, CA 92093 USA

**Hiroshi Ishiguro** ([ishiguro@ams.eng.osaka-u.ac.jp](mailto:ishiguro@ams.eng.osaka-u.ac.jp))

Department of Adaptive Machine Systems, Osaka University  
Suita, Osaka, Japan

**Howard Poizner** ([hpoizner@ucsd.edu](mailto:hpoizner@ucsd.edu))

Institute for Neural Computation, University of California, San Diego  
9500 Gilman Drive, La Jolla, CA 92093 USA

**Ayşe Pinar Saygin** ([saygin@cogsci.ucsd.edu](mailto:saygin@cogsci.ucsd.edu))

Department of Cognitive Science, University of California, San Diego  
9500 Gilman Drive, La Jolla, CA 92093 USA

## Abstract

We studied action perception and the role of visual form and visual motion kinematics of the observed agent using a stimulus set of human and humanoid robot actions and electroencephalogram (EEG). Participants viewed 2s. videos of three agents (Human, Android, Robot) performing recognizable actions: Human had biological form and motion, Android had biological form and non-biological motion, and Robot had non-biological form and non-biological motion. Early in processing (P200), Robot was distinguished from the other agents, likely due to low-level visual properties of the stimuli. We found a right temporal N170, which was most pronounced for Human, indicating possible modulation of this face- and body-sensitive ERP component by biological motion. There was a centro-parietal negativity (N300) that was most pronounced for Robot, and a later one (N400) for Human and Android. In the same time period (N300), Android was distinguished in the frontal channels from the other agents. A late positivity (P600) distinguished Human, again in frontal channels. These results highlight differential spatiotemporal cortical patterns during action perception depending on the viewed agent's form and motion kinematics.

**Keywords:** action perception; body perception; biological motion; social robotics; artificial agents; neuroimaging; EEG, ERP; uncanny valley

## Introduction

Successfully perceiving and understanding others' body movements is of biological significance, from hunting prey and avoiding predators, to communication and social interaction. The functional properties of the

neural systems that support action and body movement perception is currently an active research area in cognitive science and neuroscience.

Artificial agents such as robots can perform recognizable body movements, but can have varying degrees of biological appearance (form) and motion. As such, they provide us with an opportunity to study the specificity of neural responses to the seen agent's form and motion (as well as mismatches between the two). A prominent idea in action perception is *simulation theory*, whereby others' actions are understood via an internal sensorimotor simulation of the seen action in our own body representations (Barsalou, 2009). Supporting this, neural activity for action perception shows modulation by the degree of similarity between the observed action or actor, and the observers' own body (Buccino et al., 2004; Calvo-Merino, Grezes, Glaser, Passingham, & Haggard, 2006; Rizzolatti & Craighero, 2004). In terms of artificial agents such as robots, one might thus predict that increasing human-likeness engages simulation mechanisms more effectively.

On the other hand, human resemblance is not necessarily always a positive feature in artificial agent design. According to the *uncanny valley theory*, as an agent is made more human-like, the reaction to it becomes more and more positive and empathetic, until a point is reached at which the agent becomes oddly repulsive (Mori, 1970), an effect well-known in robotics and animation. Despite anecdotal evidence, there is little scientific data to characterize the uncanny valley (MacDorman & Ishiguro, 2006; Saygin,

Chaminade, Ishiguro, Driver, & Frith, 2011; Steckenfinger & Ghazanfar, 2009).

Previous studies on the perception of actions of humanoid robots have not found consistent results for or against simulation theory (Chaminade & Cheng, 2009). In a recent fMRI study, a more complex relationship between neural responses and the human-likeness of the observed agent was observed (including potential neural signals related to the uncanny valley), suggesting that focusing on simulation theory may be too narrow (Saygin, Chaminade, & Ishiguro, 2010). Furthermore, the specific role of biological appearance or biological motion in action processing have not been sufficiently explored in previous work, but is an area of interest in both social robotics and cognitive neuroscience (Chaminade, Hodgins, & Kawato, 2007; Kanda, Miyashita, Osada, Haikawa, & Ishiguro, 2008; Saygin, Chaminade, Urgen, & Ishiguro, 2011).

Although fMRI studies have identified the brain areas that are involved in action observation, much less is known about temporal aspects of body movement processing (Hirai, Fukushima, & Hiraki, 2003; Jokisch, Daum, Suchan, & Troje, 2005; Krakowski et al., 2011; Press, Cook, Blakemore, & Kilner, 2011). Since action processing is a naturally temporally unfolding event, it is important to further study its neural dynamics.

In the present study, we manipulated the form and the motion of the observed agent and recorded neural activity in the human brain using high-density electroencephalography (EEG), which allows us to investigate neurophysiological processes on a millisecond time scale. We used a unique stimulus set of well-matched human and humanoid robot actions (Saygin, Chaminade, Urgen et al., 2011). The stimuli consisted of videos of three agents: Human, Android, and Robot (Figure 1). Human had biological form and motion, Android had biological form and non-biological motion, and Robot had non-biological form and non-biological motion. The latter two were actually the same robot videotaped in two different appearances, but with identical kinematics. Another dimension of the stimuli was the congruence in the form and movement kinematics of the agents. Whereas Human and Robot had congruence in their form and movement kinematics (both being biological or non-biological, respectively), Android had incongruence in its form and movement kinematics as it had a biological appearance but non-biological movement kinematics.

Our goal is to study the temporal dynamics of action perception and its modulation by the seen agent's form and motion in relation to current theories in the field. Neural signals that may index simulation process would be expected to show some specificity to the Human condition. If the simulation process is driven primarily by appearance, responses to the Android are expected to be similar to the Human. If on the other

hand, biological motion is important for engaging simulation, Android responses are instead expected to show the same pattern as the Robot. As for the uncanny valley theory, we would expect neural responses for the Android to be distinct from the other conditions. Of course, the simulation theory and the uncanny valley theory are not mutually exclusive, and there may be evidence for both, possibly at different brain regions and in different time periods.

## Methods

### Participants

Twelve adults participated in the study. Participants were recruited from the student community at the University of California, San Diego (3 females, mean age: 24.4). All participants were right-handed, had normal or corrected-to-normal vision and no history of neurological disorders. Participants were either paid \$8 per hour or received course credit for their participation. They were informed about the nature of the study and signed consent forms in accordance with the UCSD Human Research Protections Program.

### Stimuli and Procedure

The experimental stimuli consisted of 2-second videos of three agents performing recognizable actions: A Human, an Android, and a Robot (Figure 1).



Biological Motion	No	No	Yes
Biological Appearance	No	Yes	Yes
Congruent Motion and Appearance	Yes	No	Yes

Figure 1. Still frames from a drinking action for Robot, Android and Human agents and the experimental features of interest (form and motion).

The Android was Repliee Q2 (Ishiguro, 2006), and the Robot condition was the same robot in a modified appearance (Saygin, Chaminade, Ishiguro et al., 2011). We recorded EEG as participants watched video clips of the 3 agents carrying out five different upper body actions (drinking, picking an object, hand waving, talking, nudging). The experiment consisted of 15 blocks of 60 trials with equal number of videos of each agent.

The stimuli were displayed on a 22' Samsung monitor at 60 Hz. In order to prevent an augmented visual evoked potential at the beginning of the movie

onset that might occlude subtle effects between conditions, we displayed two consecutive gray screens (700-1000 ms and 500-700 ms, respectively) before each video clip. In order to minimize eye movement artifacts, subjects were instructed to fixate a fixation cross at the center of the screen. In order to control for subjects' attention throughout the experiment, every random 6-10 trials, a comprehension question was displayed (e.g., Drinking? Yes/No) and subjects responded with a bimanual key press.

### EEG Recordings and Analysis

EEG was recorded at 512 Hz from 64 ActiveTwo Ag/AgCl electrodes (Biosemi, Inc.) following the International 10-10 system. The electrode-offset level was kept below 25  $\mu$ V. Four additional electrodes were placed above and below the right eye, and lateral to the eyes to monitor oculomotor activity. The data were analyzed with MATLAB and the freely available EEGLAB toolbox (Delorme & Makeig, 2004). Data was high-pass filtered at 1 Hz, low-pass filtered at 50 Hz, and re-referenced to average mastoids. Atypical epochs of electromyographic activity were removed from further analysis by semiautomated epoch rejection procedures as implemented in EEGLAB. In order to discard eye-related artifacts, the data were decomposed by extended infomax ICA using binica as implemented in EEGLAB. The data were epoched time-locked to the onset of the video clips ranging from 200 ms preceding onset to 2000 ms after onset. Data was explored both qualitatively and quantitatively. Grand Average Event-related potentials (ERP) were computed using the BrainVision Analyzer 2 software package (BrainVision, Inc.). For display purposes, ERPs were low-pass filtered at 25 Hz.

Scalp topographies for the different conditions were generated. We identified specific channels and time periods for statistical analysis. For an unbiased analysis of differences between conditions, temporal regions of interest were determined from the mean grand average ERP activity across all conditions by visual inspection of all channels. The specific time window for each component was chosen to be the narrowest time window that was common to all channels that featured the respective component. This led to the selection of six time windows: 75-150 ms, 155-205 ms, 210-260 ms, 270-370 ms, 430-540 ms and 630-800 ms from stimulus onset. Not all channels had visible components in the ERP plots, but the temporal regions were chosen to be inclusive of all possible components of interest. Within each time window, we applied paired t-tests to compare individual mean amplitudes between conditions (Robot, Android, Human). The rationale of applying paired t-tests instead of ANOVA was because the former provide a test of our experimental hypotheses without considering

irrelevant comparisons. Since our design was not a full 2x2 factorial design with form and motion (lacking the non-biological form and biological motion condition) the main effect/interaction structure of a conventional ANOVA does not correspond to the experimental comparisons of interest (the effect of form, of motion, and of congruence of form and motion). Four of the analysed time windows showed the following 5 ERP components that significantly differed between experimental conditions: An occipital P200 (155-205 ms), a central temporal N170 (155-205 ms), a centroparietal and frontal N300 (270-370 ms), a frontal N400 (430-540 ms), as well as a central and frontal P600 (630-800 ms). Where we presented data from selected channels, these were chosen as representative channels among those in the same region (as evident in the scalp distributions in Figure 2) for distinguishing one of the agents (Human, Android or Robot), thus showing the modulation of the respective component by form, motion, or congruence of form and motion. The reported p-values have been corrected for multiple comparisons unless stated otherwise (at alpha level 0.05).

### Results

EEG scalp topographies of the three conditions differed both spatially and temporally. Early on, the processing

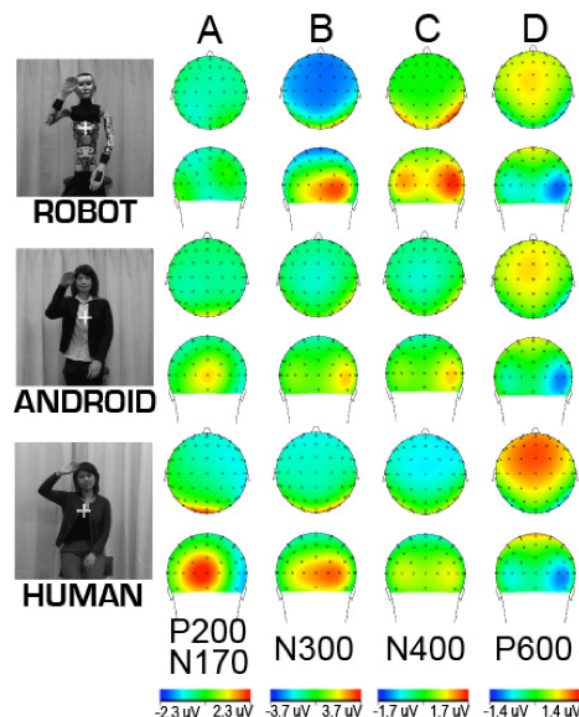


Figure 2. Scalp topographies corresponding to A) 155-205 ms (P200/N170), B) 270-370 ms (N300), C) 430-540 ms (N400), D) 630-800 ms (P600).

of Robot was distinguished from Human and Android, with an increased positivity across occipital regions for

the latter two agent conditions (Fig. 2A). Then, Robot was distinguished from Human and Android with a stronger negativity across frontal, central, and centro-parietal areas (Fig. 2B). Later, Robot was again distinguished from the other two agents with an increased positivity in centro-parietal regions, and Human was distinguished in the frontal regions (Fig. 2C). In a later stage, Human was distinguished with a stronger positivity in frontal regions (Fig. 2D).

The ERPs were then quantitatively compared across conditions to explore the role of biological form and biological action processing. Figure 3 shows ERP plots from representative channels in which the component of interest showed statistically significant amplitude modulations across conditions.

In the time window between 155-205 ms, we observed an occipital positivity (P200) that was stronger for Human and Android as compared to Robot ( $p < 0.05$ ). Although Human elicited an increased P200 than Android, Human and Android did not differ significantly, indicating a form-based modulation of this component (Fig. 3, Iz). The same time window also showed an N170 in right centro-temporal channel T8, which showed a motion-sensitive amplitude modulation (Fig. 3, T8): Here, Human (featuring biological motion) elicited increased negative amplitude compared to Android and Robot ( $p < 0.05$  and  $p < 0.01$ , respectively); the latter conditions did not differ significantly.

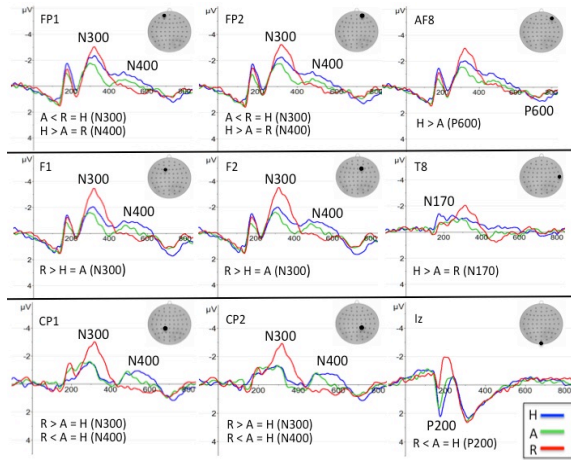


Figure 3. ERP Plots for selected channels depicting the condition effects for each component. (A: Android, H: Human, R: Robot)

Between 270 ms and 370 ms, there was a centro-parietal and frontal N300 (Fig. 3, CP1, CP2, F1, F2, Fp1, Fp2). Robot elicited a more pronounced negativity compared to Human and Android in frontal and centro-parietal channels bilaterally (Fig. 3 CP1:  $p < 0.01$ ; CP2:  $p < 0.01$ , F1:  $p < 0.01$ ; F2:  $p < 0.01$ ), indicating form-sensitive modulation. The N300 amplitudes of Human and Android did not differ. The same time window

showed less pronounced negativity for Android compared to Human and Robot in the most anterior-frontal channels bilaterally, possibly indicating a modulation by the (in)congruence of form and motion (Fig. 3, Fp1:  $p < 0.001$ ; Fp2:  $p < 0.05$ ). The responses for Human and Robot did not differ.

Between 430 ms and 540 ms, we observed a comparable negative amplitude in centro-parietal channels for Human and Android, which was absent for the Robot condition (Fig. 3 CP1, CP2), resulting in significant differences (Fig. 3 CP1 and CP2:  $p < 0.05$ ). In the same time window, in frontal channels, Human elicited an increased negativity compared to Android and Robot (Fig 3. Fp1:  $p < 0.05$ ; Fp2:  $p < 0.01$ ).

Finally, between 630 ms and 800 ms, we observed a late positivity peaking in frontal channels, which was increased for Human vs. Android (Fig. 3 AF8:  $p < 0.01$ ). The responses for Android and Robot did not differ in this time interval.

## Discussion

We investigated the temporal characteristics of neural activity during the perception of actions using a unique stimulus set of well-matched human and humanoid robot actions to manipulate the visual form and visual motion kinematics of the observed agent as we recorded electrical brain potentials (EEG). We found that neural activity during action perception is modulated differentially by the appearance and motion of the agent being observed, allowing us to observe the unfolding of perceptual and cognitive processes during action perception.

### P200

We found that an early stage of visual processing of the actions between 155-205 ms showed a form-sensitive modulation, where Robot (non-biological appearance) was distinguished from the other two agents (biological appearance, Figure 3 Iz). This is consistent with previous research on the P200 component, which is generally associated with early visual processing and is known to be sensitive to physical properties of visual stimuli (Luck & Hillyard, 1994). Since Robot had a distinct appearance from Human and Android, including low-level differences such as higher contrast and spatial frequencies, we interpret this effect as indicative of early perceptual differences, sensitive to the visual appearance of the agent being observed.

### N170

The early negative component N170, especially in the right hemisphere, has been associated with face and body processing in previous ERP research (de Gelder et al., 2010). In our study, the agents had different levels of anthropomorphism in their faces and bodies (i.e. biological vs. non-biological both in form and motion).



The Robot had a mechanical looking face with no movement, the Android and Human had similar facial appearance, but the Human face also featured biological motion (even though the actions used here did not feature prominent facial expressions and were upper body movements). We found that the amplitude of the N170 was modulated by the anthropomorphism of the agent, as manifested by a larger N170 for Human compared to the other agents. Since previous work on the N170 used static faces and bodies, our result may indicate that dynamic (biological) facial/bodily motion also elicits the N170. Another possibility is that the amplitude of the N170 might be differentially modulated depending on the presence of a context, as in our case the face was perceived together with the body during the performance of an action, whereas in previous work, still faces and bodies were shown as stimuli. As such, our results offer possible new studies to understand the functional significance of the N170 component.

### **N300/N400 complex**

The N300/N400 complex with an anterior distribution has been associated with the mapping of visual input onto representations in semantic memory (Sitnikova, Holcomb, Kiyonaga, & Kuperberg, 2008). The increased centro-parietal negativity that we found for the robot condition in the 270-370 ms time interval (Figure 3, CP1 and CP2) over anterior regions may reflect a difficulty in mapping visual input onto existing semantic representations, since robots are currently not very familiar, certainly not in the context of actions such as those in our stimuli (e.g., drinking from a cup). If this interpretation is correct, we can also deduce this process being driven primarily by the form of the agent, for if motion was a factor, the Android were equally, if not more difficult to match to semantic memory. There was also a significant effect in the same time range in frontal channels, where Android differed from the other two agents (Figure 3, Fp1, Fp2). Given the Android represents a mismatch between form and motion being potentially linked to the uncanny valley phenomenon (Ishiguro, 2006; Saygin, Chaminade, Ishiguro et al., 2011), this could be a potential component to explore in future studies on the uncanny valley, or on congruence of form and motion more generally.

### **P600 (late positivity)**

In previous work, a late positivity or P600 has mostly been studied in the domain of language and is most commonly associated with syntactic processing (Friederici, 2004). Few studies have interpreted the P600 in other domains (Sitnikova et al., 2008). In our data (Figure 3, AF8), we found that this component was elicited most strongly by the Human condition. This can lead us new avenues of research to understand the

functional significance of the ERP components observed in action perception.

### **Implications for Action Processing**

Although action processing has been an active area of study in cognitive neuroscience, most work to date has used fMRI rather than electrophysiology. More specifically, a number of studies have focused on the perception of human and robot agents with fMRI, with inconsistent support for the simulation theory (Saygin, Chaminade, Ishiguro et al., 2011). Here, we add new ERP results to this literature, providing information about the role of humanoid form and humanoid motion during the course of action perception.

The stimuli used here were previously utilized in an fMRI repetition-suppression study in which brain activity did not show evidence for form-based or motion-based simulation per se, but instead was most significantly affected by form-motion incongruence (Saygin, Chaminade, Ishiguro et al., 2011). Here, with a more time-resolved method, we found distinct stages of processing during which neural responses differed based on both the form and the motion of the seen agent. These effects were likely lost due to the temporal insensitivity of fMRI, highlighting the importance of using multiple, complementary techniques.

A well-known face-sensitive component, the N170, was elicited by our stimuli. Previous work on this ERP signature of face processing has used static face stimuli, as opposed to movies including the body as we did here. Our data suggest new possible ways in which the N170 can be modulated. Specifically we hypothesize that either biological motion of the face and/or the context provided by the body are modulators of the N170.

Our data did not reveal patterns of activity that can be linked straightforwardly to simulation theory. There was some selectivity for the Human (for whom simulation theory would predict differential effects, whether driven by form or motion) for the frontal N400 and P600, but there is little prior literature on actions for these components, and no link to sensorimotor simulation that we are aware of. The uncanny valley theory also cannot account for all of the patterns in our data, although the frontal N300 response could be interpreted as biomarker for the uncanny valley. These components should be viewed as possible indices related to each theory, to be tested in new studies.

Overall, in this first ERP study of action perception with human and humanoid agents, we highlight the complexity of action processing that can be revealed using more time-resolved methods. We found distinct neural signatures of the viewed agent's form and motion in different time periods, both early (perceptual) and late (cognitive) in processing. These results do not globally fit into either simulation or uncanny valley frameworks, although a focus on specific components

such as the N170 and N300/400 in upcoming studies might help better understand the mechanisms of action perception and its neural basis. Work on neural dynamics of action processing can not only shed light on the cognitive neuroscience of action perception, but also to inform the burgeoning field of social robotics (Saygin, Chaminade, Urgen et al., 2011).

## Acknowledgements

This research was supported by the Kavli Institute for Brain and Mind (APS), California Institute of Telecommunications and Information Technology (Calit2) (BAU, APS), NSF (SBE-0542013, Temporal Dynamics of Learning Center), and ONR (MURI Award # N00014-10-1-0072, HP). The authors would like to thank Marta Kutas, Arthur Vigil, members of the Intelligent Robotics Laboratory (Osaka University) for the creation of the stimuli, and Joe Snider (Poizner Lab) for his help in the experimental setup.

## References

- Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society of London B*, 364(1521), 1281-1289.
- Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., et al. (2004). Neural circuits involved in the recognition of actions performed by nonconspecifics: an fMRI study. *Journal of Cognitive Neuroscience*, 16(1), 114-126.
- Calvo-Merino, B., Grezes, J., Glaser, D. E., Passingham, R. E., & Haggard, P. (2006). Seeing or doing? Influence of visual and motor familiarity in action observation. *Current Biology*, 16(19), 1905-1910.
- Chaminade, T., & Cheng, G. (2009). Social cognitive neuroscience and humanoid robotics. *Journal of Physiology Paris*, 103(3-5), 286-295.
- Chaminade, T., Hodgins, J., & Kawato, M. (2007). Anthropomorphism influences perception of computer-animated characters' actions. *Social Cognitive and Affective Neuroscience*, 2(3), 206-216.
- Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1), 9-21.
- de Gelder, B., Van den Stock, J., Meeren, H.K.M., Sinke, J.B.A., Kret, M.E., Tamietto, M. (2010). Standing up for the body: Recent progress in uncovering the networks involved in processing bodies and bodily expressions. *Neurosci. Biobehav. Rev.*, 34, 513-527.
- Friederici, A. D. (2004). Event-related brain potential studies in language. *Current Neurology Neuroscience Report*, 4(6), 466-470.
- Hirai, M., Fukushima, H., & Hiraki, K. (2003). An event-related potentials study of biological motion perception in humans. *Neurosci Lett*, 344(1), 41-44.
- Ishiguro, H. (2006). Android science: conscious and subconscious recognition. *Connection Science*, 18(4), 319-332.
- Jokisch, D., Daum, I., Suchan, B., & Troje, N. F. (2005). Structural encoding and recognition of biological motion: evidence from event-related potentials and source analysis. *Behavioral Brain Research*, 157(2), 195-204.
- Kanda, T., Miyashita, T., Osada, T., Haikawa, Y., & Ishiguro, H. (2008). Analysis of humanoid appearances in human-robot interaction. *IEEE Transactions on Robotics*, 24(3), 725-735.
- Krakowski, A. I., Ross, L. A., Snyder, A. C., Sehatpour, P., Kelly, S. P., & Foxe, J. J. (2011). The neurophysiology of human biological motion processing: A high-density electrical mapping study. *Neuroimage*, 56(1), 373-383.
- Luck, S. J., & Hillyard, S. A. (1994). Electrophysiological correlates of feature analysis during visual search. *Psychophysiology*, 31(3), 291-308.
- MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, 7(3), 297-337.
- Mori, M. (1970). The uncanny valley. *Energy*, 7(4), 33-35.
- Press, C., Cook, J., Blakemore, S. J., & Kilner, J. M. (2011). Dynamic modulation of human motor activity when observing actions. *Journal of Neuroscience*, 31(8), 2792-2800.
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169-192.
- Saygin, A. P., Chaminade, T., & Ishiguro, H. (2010). The perception of humans and robots: Uncanny hills in parietal cortex. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2716-2720). Portland, OR: Cognitive Science Society.
- Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. F. (2011). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience*.
- Saygin, A. P., Chaminade, T., Urgen, B. A., & Ishiguro, H. (2011). Cognitive neuroscience and robotics: A mutually beneficial joining of forces In L. Takayama (Ed.), *Robotics: Systems and Science*. Los Angeles, CA.
- Sitnikova, T., Holcomb, P. J., Kiyonaga, K. A., & Kuperberg, G. R. (2008). Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. *Journal of Cognitive Neuroscience*, 20(11), 2037-2057.
- Steckenfinger, S. A., & Ghazanfar, A. A. (2009). Monkey visual behavior falls into the uncanny valley. *Proceedings of the National Academy of Sciences of the United States of America*, 106(43), 18362-18366.



# Effects of Discourse Goals on the Process of Metaphor Production

Akira Utsumi (utsumi@inf.uec.ac.jp)

Kota Nakamura (kota@utm.inf.uec.ac.jp)

Maki Sakamoto (sakamoto@inf.uec.ac.jp)

Department of Informatics, The University of Electro-Communications  
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

## Abstract

Only a few attempts have so far been made at exploring the process of metaphor production, although a large number of studies have addressed metaphor comprehension. Therefore, in this paper, we address the problem of how people generate metaphors or identify an apt vehicle for a given topic of metaphors. Specifically, we examine how the process and product of metaphor production differ between two discourse goals of metaphor, namely an explanatory purpose (e.g., to clarify) and a literary purpose (e.g., to aesthetically pleasing). Experiment 1 analysed the metaphors (or vehicles) generated in the metaphor production task, and demonstrated that people identified more prototypical exemplars of the property attributed to the topic as a vehicle for explanatory metaphors than for literary metaphors. In addition, it was found that explanatory metaphors were more apt and conventional, and had high topic-vehicle similarity than literary metaphors, while literary metaphors were more familiar and imageable than explanatory metaphors. Experiment 2 used a priming paradigm to assess the online availability of prototypical and less prototypical members of the topic property during metaphor production. The result was that both prototypical and less prototypical members were activated in producing literary metaphors, while neither members were activated in the production of explanatory metaphors. These findings indicate that the process of metaphor production is affected by discourse goals of metaphor, and suggest that only prototypical members of the category are rapidly searched for a vehicle during the production of explanatory metaphors, while both prototypical and less prototypical members are searched to generate literary metaphors.

**Keywords:** Metaphor production; Discourse goal; Priming

## Introduction

Metaphor has been a main topic of research in cognitive science, because metaphorical expressions are frequently observed in our everyday use of language. Hence, a large number of studies have been made on how people comprehend metaphors (e.g., Bowdle & Gentner, 2005; Gibbs, 2008; Glucksberg, 2001; Utsumi, 2011). In contrast, only a few studies have addressed the process of metaphor production (for a notable exception, see, for example, Chiappe & Chiappe, 2007; Katz, 1989). This paucity of research on metaphor production is especially problematic, given that by its very nature a metaphor is an intentional, linguistic device employed to convey ideas that might be otherwise inexpressible. To ameliorate this situation, in this paper, we empirically explore the process of metaphor production.

Previous studies on metaphor production analysed the products of metaphor production (i.e., metaphorical expressions produced) in terms of the qualitative dimensions of metaphors and/or individual differences. Concerning the qualitative dimensions of metaphor products, Fainsilber and Ortony (1987) demonstrated that descriptions of emotional states

contained more metaphorical language than did descriptions of actions. Katz (1989) examined the properties of metaphor vehicles by asking participants to choose, from a set of alternatives, a vehicle that completes a given sentence frame (e.g., Chemistry is the \_\_\_\_\_ of science) as comprehensible and apt metaphors. The result was that participants were likely to choose the vehicles that were moderately distant from the topic and referred to concrete domains. Concerning individual differences, it was demonstrated that the quantity and quality of metaphor products were affected by individual differences such as writing experience (experienced or novice) (Williams-Whitney, Mio, & Whitney, 1992), reasoning and imagery ability (Katz, 1989), gender (male or female) (Hussey & Katz, 2006), and working memory capacity (Chiappe & Chiappe, 2007). For example, Chiappe and Chiappe (2007) demonstrated that people with high working memory capacity produced more apt metaphors than did low capacity individuals.

Although shedding light on the specific aspects of metaphor production, these studies did not address one important aspect of metaphor production, namely *discourse goals of metaphor*. Because metaphors (and other figurative language) are intentionally used to accomplish certain communication goals (Roberts & Kreuz, 1994), it is obviously crucial to explore the effects of discourse goals on the process of metaphor production. Discourse goals that are accomplished by the use of metaphorical expressions can be classified broadly into two classes: explanatory and literary purposes (Steen, 1994; Utsumi, 2005). These two goals are quite different and sometimes incompatible with each other; explanatory metaphors are used to clarify certain properties of the topic, while literary metaphors are used to evoke an aesthetically pleasing feeling by enriching the meanings conveyed by the metaphors. It naturally follows that discourse goals are likely to affect the process of generating metaphors, or more specifically choosing the vehicles of metaphors. This paper aims at examining how the process and product of metaphor production differ between these two discourse goals.

In metaphor production, people often have in mind a topic that they want to express and some properties that they intend to attribute to the topic. They must identify an appropriate or apt vehicle to convey the intended meaning (i.e., the information that the topic has the property). Hence, metaphor production essentially involves the process of searching for or retrieving an apt vehicle. According to the categorization (or attributive category) theory of metaphor (Glucksberg, 2001; Glucksberg & Keysar, 1990), apt vehicles must not only have the intended property but also be a prototypical exemplar of

that property. For example, consider the topic “rumor” and the property “quickly spreads from person to person.” People have to identify an appropriate vehicle to express by metaphor that a rumor quickly spreads from person to person. This property is true of a number of things such as “virus,” “swine flu,” and “louse,” but a virus is a prototypical exemplar of the category “things that quickly spread from person to person.” Hence, the metaphor “The rumor is a virus” seems to be more apt than “The rumor is a swine flu” and “The rumor is a louse.”

The research question to be answered here is how people identify or select a vehicle that accomplishes their communication purposes. This question can be rephrased as how the process of searching a set of things (i.e., search space) for a vehicle differs according to whether explanatory metaphors or literary metaphors are intended. To tackle this problem, we consider the observed differences between explanatory and literary metaphors. Some studies have found that semantic aptness (Steen, 1994; Utsumi, 2005), clarity (Gentner, 1982), and interpretive richness (Gentner, 1982; Utsumi, 2005) are distinctive properties for distinguishing between explanatory and literary metaphors. Explanatory metaphors are more apt (or appropriate) and clearer than literary metaphors, while literary metaphors are interpretively richer than explanatory metaphors. These findings suggest that semantically apt and clear vehicles may be preferably searched for in producing explanatory metaphors, while less apt vehicles that enrich the metaphorical meaning may be searched for in producing literary metaphors. According to the attributive category theory, vehicles of apt metaphors are prototypical members of the category characterized by the intended property of the topic (Glucksberg & Keysar, 1990). On the other hand, for a metaphor to be semantically rich and involve a number of metaphorical interpretations, its vehicle must be less prototypical because highly prototypical vehicles evoke only the intended property.

From these discussions, we can derive the following hypothesis about the process of metaphor production:

**Hypothesis about process:** *Only prototypical members of the category are searched for a vehicle during the production of explanatory metaphors, while less prototypical members are also considered to generate literary metaphors.*

This hypothesis presupposes that people first search prototypical members of the category before searching less prototypical members, regardless of discourse goals (e.g., Giora, 2003; Rosch & Mervis, 1975). Hence, this hypothesis implies that when producing explanatory metaphors, they do not have to search less prototypical members, because prototypical members are sufficient for apt vehicles. However, when producing literary metaphors, people have to search less prototypical members after prototypical members are activated.

The hypothesis about process implies another hypothesis about the metaphors (or vehicles) produced.

**Hypothesis about products:** *More prototypical and apt vehicles are chosen for explanatory metaphors than for literary metaphors.*

Considering that prototypicality and other related properties such as conventionality and familiarity can be classified into a more general notion of salience (Giora, 2003), it is also predicted that conventional and familiar vehicles are preferably chosen for explanatory metaphors. In addition, metaphor aptness refers to the extent to which the vehicle’s metaphoric category captures an important feature of the topic, and thus it reflects the similarity between the vehicle and topic of a metaphor (e.g., Chiappe & Kennedy, 1999). Hence, we also predict that vehicles chosen for explanatory metaphors are more similar to the topic than those chosen for literary metaphors.

In this paper, we test these hypotheses through two experiments, namely, an offline generation experiment and an online priming experiment. In the metaphor generation experiment (i.e., Experiment 1), we examined the validity of the hypothesis about products by analysing the vehicles generated for explanatory or literary metaphors. We compared the vehicles for explanatory metaphors and literary metaphors in terms of several factors: vehicle prototypicality, metaphor conventionality, metaphor aptness, topic-vehicle similarity, vehicle familiarity, vehicle word frequency, and vehicle imageability. In the online priming experiment (i.e., Experiment 2), we used a priming paradigm to test the hypothesis about the internal process of metaphor production, particularly to examine what words (or concepts) are activated during identifying an appropriate vehicle. In the priming experiment, participants were presented with a topic (i.e., an incomplete metaphorical sentence) and its property as a prime and asked to determine a vehicle appropriate for a given discourse goal. Afterwards, a target word was presented and participants were asked to make a lexical decision about it. The target conditions were a word that was highly prototypical of the category made up of the property, a word that was less prototypical of the category, and a control word unrelated to the property.

It must be noted here that the previous studies mentioned earlier analysed only the products of metaphor production using an offline paper-and-pencil experiment paradigm; they did not directly examine the online processes of metaphor production. To the best of our knowledge, this is the first study to apply an online priming paradigm to the study of metaphor production.

## Experiment 1

In Experiment 1, we tested the hypothesis about the product by collecting metaphor vehicles in a metaphor generation task and the ratings of their properties in a vehicle rating task.

### Method

**Participants** Forty undergraduate and graduate students participated as volunteers. All participants were native speakers of Japanese.

**Materials** Twenty pairs of a topic and a property to be attributed to the topic were used for the experiment. Topic words were selected from Japanese abstract nouns comprising two kanji characters, and their properties were expressed in a Japanese short phrase referring to a salient feature of the

topic. In order to eliminate an undesirable effect of the abstractness of the topic words on the process of categorization (Glucksberg & Keysar, 1990), we equalized the degree of abstractness of the topic words by choosing them from abstract categories of words at almost the same depth (i.e., depth of 7 or 8) in the hierarchical structure of the Japanese thesaurus. For example, the topic word “plan” (“*keikaku*” in Japanese) was paired with the property “does not always go as scheduled” (“*yotei doori-ni ikanai*” in Japanese).

**Procedure** This experiment comprised two tasks, namely a metaphor generation task and a vehicle rating task.

The metaphor generation task was conducted by 20 participants. Each participant was assigned all the 20 topic-property pairs, one half of which were used to generate explanatory metaphors and the other half of which were used to generate literary metaphors. Topic-vehicle pairs were counterbalanced across conditions so that each pair appeared 10 times in both conditions. The presentation order of the pairs in each group was randomized for each participant. Participants, who were run individually, were seated in front of a computer screen. They were first given an overall instruction of the experiment and presented with four practice trials (two for explanatory metaphors and two for literary metaphors) followed by two groups of 10 experimental trials. In the explanatory metaphor condition, participants were instructed to generate a vehicle that clearly explains the given property of the topic, while in the literary metaphor condition they were instructed to generate a metaphor aesthetically pleasing enough to use in literary works. On each trial, they were presented with a sentence frame including a topic (e.g., “A plan is (like) a \_\_\_\_\_”) and a property (“does not always go as scheduled”) in the center of the screen, and asked to generate an apt vehicle that completes the sentence at their own pace. When participants came up with a suitable vehicle, they indicated it by pressing the appropriate key on the keyboard. Afterwards, they typed the vehicle as quickly as possible. Reaction times were measured from the onset of the topic-property pair until the appropriate key was pressed (metaphor production time), and from the key press until the input of the vehicle was completed (vehicle typing time).

In the vehicle rating task, another 20 participants were presented with all the 20 topic-property pairs and their vehicles generated (i.e., metaphors and their property attributed to the topic) in the metaphor generation task. They were asked to rate each metaphor or vehicle on the following four 7-point scales: vehicle prototypicality (7 = *prototypical*, 1 = *not at all prototypical*), metaphor conventionality (7 = *conventional*, 1 = *novel*), metaphor aptness (7 = *apt*, 1 = *not at all apt*), and topic-vehicle similarity (7 = *similar*, 1 = *dissimilar*).

## Results and Discussion

First, we analysed the metaphor production time for explanatory and literary metaphors in the metaphor generation task. We eliminated from the analysis extreme outliers (i.e., production times shorter than 2s or longer than 40s, and production times of the trial whose vehicle typing times were longer than 30s), and averaged the remaining production times. As

Table 1: Means (*M*) and standard deviations (*SD*) of seven factors of explanatory and literary metaphors in Experiment 1

Factor	Explanatory		Literary	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Vehicle prototypicality***	5.14	0.64	4.70	0.43
Metaphor conventionality***	4.21	0.90	3.31	0.60
Metaphor aptness*	5.16	0.67	4.86	0.42
Topic-vehicle similarity***	3.98	0.43	3.60	0.59
Vehicle familiarity*	5.90	0.33	6.09	0.28
Vehicle word frequency	3.14	0.52	3.17	0.54
Vehicle imageability*	5.05	0.33	5.32	0.44

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

a result, literary metaphors ( $M=77.39s$ ,  $SD=50.18s$ ) took longer to generate than explanatory metaphors ( $M=55.82s$ ,  $SD=29.74s$ ). This difference was marginally significant in the participant analysis,  $F_p(1, 19) = 4.18$ ,  $p = .055$ , although not significant in the item analysis,  $F_i(1, 19) = 2.13$ ,  $p > .1$ . (Note that, in all the ANOVAs reported in this paper, the data were analyzed by participants  $F_p$  and by items  $F_i$ .) This result is consistent with the hypothesis about the production process, in that people must search less prototypical members of the category after searching prototypical members, and consequently require more time to generate literary metaphors than explanatory metaphors. Note, however, that this result may not be reliable enough to warrant the hypothesis, because we had no efficient methods for confirming that participants really spent all the time producing metaphors. (Hence, Experiment 2 was conducted to collect reliable data on the internal process of metaphor production.)

We then compared the explanatory and literary metaphors generated in terms of four factors (i.e., vehicle prototypicality, metaphor conventionality, metaphor aptness, and topic-vehicle similarity). Table 1 shows the mean rating values of these four factors for explanatory and literary metaphors. Explanatory metaphors were rated as significantly higher on all the four factors than literary metaphors,  $F_p(1, 19) = 44.24$ ,  $p < .001$ ,  $F_i(1, 19) = 9.44$ ,  $p < .01$  for vehicle prototypicality;  $F_p(1, 19) = 1256.76$ ,  $p < .001$ ,  $F_i(1, 19) = 16.76$ ,  $p < .001$  for metaphor conventionality;  $F_p(1, 19) = 4.76$ ,  $p < .051$ ,  $F_i(1, 19) = 3.70$ ,  $p = 0.70$  for metaphor aptness; and  $F_p(1, 19) = 23.53$ ,  $p < .001$ ,  $F_i(1, 19) = 8.49$ ,  $p < .01$  for topic-vehicle similarity. These results are entirely consistent with the hypothesis about products, confirming that more prototypical and apt vehicles are chosen for explanatory metaphors than for literary metaphors.

Furthermore, we exploratorily analysed three additional factors of vehicle words — i.e., vehicle word familiarity, vehicle word frequency, and vehicle word imageability — that may differ between explanatory and literary metaphors. These values were derived from the database of Japanese lexical properties “Nihongo No Goi Tokusei.” In this database, the familiarity and imageability of a word were given as the mean rating scores on a 7-point scale ranging from 1 to 7, and the frequency of a word was given as the number of times the word occurred in a newspaper corpus. In the analysis, these

scores were used for vehicle familiarity and imageability, and the common logarithm of the frequency score was used for vehicle word frequency. The last three rows of Table 1 show the mean values of these three factors across vehicle words. Two out of three factors were significantly different; the vehicles of literary metaphors were more familiar and imageable than those of explanatory metaphors,  $F_i(1, 19) = 5.25, p < .05$  for vehicle familiarity; and  $F_i(1, 19) = 5.52, p < .05$  for vehicle imageability. The finding on vehicle imageability is intuitively acceptable, because literary works often evoke mental imagery, although Katz, Paivio, and Marschark (1985) found an opposite result that the vehicles of poetic metaphors were less imageable. On the other hand, the finding of vehicle familiarity is seemingly surprising, but can be explained by the assumption that the use of familiar words as a vehicle may enrich the interpretation of metaphors. Note that this finding is consistent with the finding of Katz et al. (1985) that poetic metaphors were rated as being more familiar.

## Experiment 2

In Experiment 2, we tested the prediction about the metaphor production process using a priming paradigm, in which an incomplete metaphorical sentence was presented first with the property to be attributed to the topic, and the task was to make a lexical decision about a target word presented after the metaphorical sentence. The target conditions were a word highly prototypical of the category characterized by the property (HPT), a word less prototypical of the category (LPT), and a control target (CNT) that is unrelated to the metaphor.

Faster lexical decisions in comparison with the CNT indicate online activation. If only the prototypical members of the category are searched for a metaphor vehicle, the HPT would be faster to make a lexical decision than the CNT, but the LPT would not be faster. If less prototypical members are also searched, both the LPT and HPT would be faster than the CNT. Hence, if our hypothesis about the process is right, only the HPT would be facilitated when explanatory metaphors are intended, while both the HPT and LPT would be facilitated when literary metaphors are generated.

## Method

**Participants** Fifty-four undergraduate and graduate students participated as volunteers. All participants were native speakers of Japanese.

**Materials** Thirty Japanese pairs of a topic and a property attributed to the topic (including all the 20 pairs used in Experiment 1) were used as primes. For each prime pair, three target words (i.e., HPT, LPT, and CNT) comprising two kanji characters were prepared. The HPTs and LPTs were selected on the basis of ratings collected from independent groups of participants in a norming study. The CNTs were selected randomly from a dictionary so that they were unrelated to both the topic and the property, and their word frequency was approximately equal to that of the HPTs and LPTs.<sup>1</sup> For exam-

ple, the pair of the topic “all-night activity” (“*tetsuya*”) and its property “unhealthy” (“*karada ni aku eikyo wo ataeru*”) was combined with the HPT “drug” (“*mayaku*”), the LPT “edacity” (“*oogui*”), and the CNT “speech” (“*enzetsu*”). In addition, another 20 topic-property pairs were prepared and used as filler sentences for nonword targets.

**Norming study** For each of the 30 topic-property pairs, six words (comprising two kanji characters) were prepared that referred to an object or a concept with that property but different degree of prototypicality. Ten undergraduate students rated these words on the 7-point scale of prototypicality ranging from 1 (*not at all prototypical*) to 7 (*prototypical*). For each topic-property pair, the word with the highest rating was selected as an HPT, and the word rated as the lowest of the words with the prototypicality degree of 4 (i.e., the midpoint) or higher was selected as an LPT. The prototypicality rating of the HPTs ( $M = 5.91, SD = 0.51$ ) was significantly higher than that of the LPTs ( $M = 4.37, SD = 0.35$ ),  $F(1, 29) = 211.23, p < .001$ , confirming that LPTs and HPTs were appropriately selected.

**Procedure** A within-participants design was used with each participant processing all the 50 topic-property pairs (i.e., 30 pairs with word targets and 20 pairs with nonword targets) under all conditions. The 50 pairs were equally divided into two groups, each of which comprised 15 pairs with word targets and 10 pairs with nonword targets. One group was used to generate explanatory metaphors and another group was used to generate literary metaphors. Topic-vehicle pairs with word targets were counterbalanced across all conditions (i.e., two conditions of discourse goal and three target conditions) so that each pair appeared an equal number of times in all conditions. The presentation order of two groups of discourse goals (i.e., whether explanatory or literary metaphors were generated first) and the order of the pairs in each group were randomized for each participant.

Participants, who were run individually, were seated in front of a computer screen. They were first given an overall instruction of the experiment and then presented with four practice trials (two for explanatory metaphors and two for literary metaphors) followed by two groups of 25 experimental trials. On each trial, they were presented with a sentence frame with a topic in the subject position (“A plan is (like) a \_\_\_\_\_”) and a property (“does not always go as scheduled”) as a prime in the center of the screen for 7000 ms and asked to consider an appropriate vehicle word to fill in the blank. A target word (HPT, LPT, CNT, or nonword) was then presented 500ms after the offset of the topic-property pair. Participants were asked to decide whether the target word was a word or a nonword as quickly as possible; they indicated decision by pressing the appropriate key on the keyboard. Finally, they typed the vehicle that occurred to them before the lexical decision. Reaction times were measured from the onset of the target word until the appropriate key was pressed.

## Results and Discussion

Only reaction times of correct decision were used in the analysis. In addition, reaction times greater than 10,000ms were

<sup>1</sup>The mean logarithms of word frequency score were 3.41 for the HPT, 3.62 for the LPT, and 3.60 for the CNT. They were not significantly different,  $F(2, 58) = 0.492$ .

Table 2: Means (*M*) and standard deviations (*SD*) of correct lexical decision times in milliseconds for Experiment 2

Metaphor condition	HPT (High Prototypicality)			LPT (Low Prototypicality)			CNT (Control)	
	<i>M</i>	<i>SD</i>	<i>DIF</i>	<i>M</i>	<i>SD</i>	<i>DIF</i>	<i>M</i>	<i>SD</i>
Explanatory	1450	550	51	1474	584	27	1501	727
Literary	1431	475	155	1303	420	283	1586	743

Note. *DIF* = difference from control target.

eliminated from the analysis.

Table 2 shows mean lexical decision times and standard deviations for the correct “yes” responses. The time difference (*DIF*) from the CNT indicates the extent of a priming effect. In the explanatory metaphor condition, the HPT produced a moderate priming effect (51ms faster than the CNT), but the LPT showed only a small priming effect (27ms faster). On the other hand, in the literary metaphor condition, both targets showed a much larger priming effect. In particular, the priming effect of the LPT (283ms) was larger than that of the HPT (155ms), indicating that less prototypical exemplars of the category were activated during the production of literary metaphors. These results are entirely consistent with our hypothesis that only prototypical members of the category are searched for a vehicle during the production of explanatory metaphors, while less prototypical members are also considered to generate literary metaphors.

To confirm these differences statistically, we conducted a two-way ANOVA of Target (HPT, LPT, or CNT)  $\times$  Discourse goal (explanatory or literary) on lexical decision times. These two factors were within participants and within items. First of all, the main effect of Discourse goal was not significant in either analyses, but the main effect of Target was significant in the participant analysis,  $F_p(2, 106) = 4.54, p < .05$ . Post-hoc pairwise comparisons showed that the priming effect of the LPT was significant ( $p < .05$ ) and the priming effect of the HPT was marginally significant ( $p = .051$ ). This result indicates that discourse goals did not affect the overall processing time of lexical decision and the priming effect was observed in this experiment. It confirms that this priming experiment was successful and the priming methodology can be applied effectively to the study of metaphor production.

The most important result was that the interaction between two factors was significant in the participant analysis,  $F_p(2, 106) = 3.40, p < .05$ , although not significant in the item analysis. The nature of this interaction was that, the simple main effect of Target was significant in the literary metaphor condition,  $F_p(1, 159) = 4.57, p < .05$ , but not significant in the explanatory metaphor condition. Post-hoc pairwise comparisons ( $p < .05$ ) revealed that both the LPT ( $M = 1303$ ms) and HPT ( $M = 1431$ ms) were significantly faster than the CNT ( $M = 1586$ ms). Again, these results are consistent with the hypothesis of this paper, although the absence of a significant priming effect of the HPT in the explanatory metaphor condition was not consistent with the hypothesis.

## General Discussion

The two experiments reported in this paper provided empirical evidence in favor of our view that discourse goals of metaphor affect the process of metaphor production. Specifically, we focused on two types of metaphors — i.e., explanatory metaphors and literary metaphors — that accomplish different discourse goals, and demonstrated that the production of literary metaphors required activation of both prototypical and less prototypical members of the category characterized by the topic property, while the production of explanatory metaphors did not. This processing difference leads to the finding on the products that explanatory metaphors generated in the metaphor production experiment were more prototypical than literary metaphors. In addition, it was found that explanatory metaphors were more conventional and apt, and had higher topic-vehicle similarity, while the generated vehicles for literary metaphors were more familiar and imageable.

The finding that less prototypical members of the category were activated during the processing of literary metaphors and as a result less prototypical vehicles are selected for literary metaphors is especially interesting, because it indicates that some metaphors cannot be explained by the attributive category theory of metaphor (Glucksberg, 2001; Glucksberg & Keysar, 1990). We have demonstrated elsewhere that the comprehension of predicative metaphors (i.e., figurative expressions that involve the metaphorical use of a verb or an adjective) cannot be explained by the attributive category theory of metaphor, and proposed an indirect categorization theory of metaphor (Utsumi & Sakamoto, 2007, 2011). The finding of this study suggests a possibility that some nominal metaphors (in particular, literary metaphors) may be processed by an indirect categorization or other mechanisms.

On the other hand, the finding that prototypical members were not activated during the production of explanatory metaphors is not consistent with the hypothesis. One possible explanation of this result would be that the process of identifying a vehicle for explanatory metaphors was so rapid that prototypical members were no longer activated 7500 ms after the onset of the prime. To test this possibility, we must repeat a priming experiment by varying the stimulus-onset asynchrony (SOA) of prime and target, which is left for future research.

The discourse goals of metaphorical or figurative expressions are affected by their grammatical form, i.e., whether figurative comparisons are expressed in metaphor form “An X is a Y” or in simile form “An X is like a Y.” Glucksberg and Keysar (1990) argued that metaphors are more forceful

than similes and alert an addressee that a specific set of properties is intended. Roberts and Kreuz (1994) demonstrated that metaphors and similes differ in that metaphors are used to add interest, but similes are used to deemphasize. If the discourse goals really affect the process of metaphor production as demonstrated in this paper, these pragmatic differences between metaphor and simile imply that the grammatical form has an important influence on the process of metaphor production. Furthermore, a number of studies have revealed that several properties addressed in Experiment 1 (e.g., aptness, conventionality, topic-vehicle similarity) determine the preferred form of figurative comparison (e.g., Bowdle & Gentner, 2005; Chiappe & Kennedy, 1999; Jones & Estes, 2006). Hence, the finding of Experiment 1 may serve to explain the relation between the grammatical form and the production process.

Computational modeling is also an efficient methodology for the study of metaphor production. Recently, computational studies based on a semantic space model such as latent semantic analysis (LSA) have addressed metaphors and shed new light on the process of metaphor comprehension (e.g., Kintsch, 2000; Utsumi, 2011; Utsumi & Sakamoto, 2007). In particular, the predication algorithm in a semantic space model proposed by Kintsch (2000) can embody the attributive category theory, and has been shown to achieve good performance of simulating the process of metaphor comprehension. It follows that the study of metaphor production can also benefit from the same computational modeling framework. For example, the predication algorithm computes the meanings of metaphors by combining neighbors of the vehicle (i.e., word vectors similar to the vehicle vector) with the topic and the vehicle. The process of metaphor production can also be modeled in the same way; an appropriate vehicle can be selected from the neighbors of the vector for the metaphorical meaning expressed by a given topic-vehicle pair, and a search space can be manipulated by varying the number of and quality of neighbor vectors. Indeed, Chiappe and Chiappe (2007) attempted to explain their findings on metaphor production using the predication algorithm.

These research topics for advancing the study of metaphor production are worth pursuing in further research.

## Acknowledgments

This study was supported by a Grant-in-Aid for Scientific Research B (No.23300098) from the Japan Society for the Promotion of Science.

## References

- Bowdle, B., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, 112(1), 193–216.
- Chiappe, D., & Chiappe, P. (2007). The role of working memory in metaphor production and comprehension. *Journal of Memory and Language*, 56, 172–188.
- Chiappe, D., & Kennedy, J. (1999). Aptness predicts preference for metaphors or similes, as well as recall bias. *Psychonomic Bulletin & Review*, 6, 668–676.
- Fainsilber, L., & Ortony, A. (1987). Metaphorical uses of language in the expression of emotions. *Metaphor and Symbolic Activity*, 2(4), 239–250.
- Gentner, D. (1982). Are scientific analogies metaphors? In D. Miall (Ed.), *Metaphor: Problems and perspectives* (pp. 106–132). Sussex, England: The Harvester Press.
- Gibbs, R. W. (Ed.). (2008). *The cambridge handbook of metaphor and thought*. Cambridge University Press.
- Giora, R. (2003). *On our mind: Salience, context, and figurative language*. Oxford University Press.
- Glucksberg, S. (2001). *Understanding figurative language: From metaphors to idioms*. Oxford University Press.
- Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, 97, 3–18.
- Hussey, K. A., & Katz, A. N. (2006). Metaphor production in online conversation: Gender and friendship status. *Discourse Processes*, 42(1), 75–98.
- Jones, L., & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, 55, 18–32.
- Katz, A. (1989). On choosing the vehicles of metaphors: Referential concreteness, semantic distances, and individual differences. *Journal of Memory and Language*, 28, 486–499.
- Katz, A., Paivio, A., & Marschark, M. (1985). Poetic comparisons: Psychological dimensions of metaphoric processing. *Journal of Psycholinguistic Research*, 14(4), 365–383.
- Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, 7(2), 257–266.
- Roberts, R., & Kreuz, R. (1994). Why do people use figurative language? *Psychological Science*, 5(3), 159–163.
- Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573–605.
- Steen, G. (1994). *Understanding metaphor in literature: An empirical approach*. Longman Publishing Group.
- Utsumi, A. (2005). The role of feature emergence in metaphor appreciation. *Metaphor and Symbol*, 20(3), 151–172.
- Utsumi, A. (2011). Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive Science*, 35(2), 251–296.
- Utsumi, A., & Sakamoto, M. (2007). Computational evidence for two-stage categorization as a process of adjective metaphor comprehension. In *Proceedings of the Second European Cognitive Science Conference (EuroCogSci2007)* (pp. 77–82).
- Utsumi, A., & Sakamoto, M. (2011). Indirect categorization as a process of predicative metaphor comprehension. *Metaphor and Symbol*, 26(4), 299–313.
- Williams-Whitney, D., Mio, J., & Whitney, P. (1992). Metaphor production in creative writing. *Journal of Psycholinguistic Research*, 21(6), 497–509.

# Modeling Efficient Serial Visual Search

**Bella Z. Veksler (bellav717@gmail.com)**

Air Force Research Laboratory, Wright-Patterson Air Force Base  
Dayton, OH 45431 USA

**Wayne D. Gray (grayw@rpi.edu)**

Cognitive Science Department, Rensselaer Polytechnic Institute  
Troy, NY 12180 USA

## Abstract

Humans perform visual search fairly efficiently, finding targets within only a few fixations. Data from eye-tracked participants was subjected to a fixation by fixation analysis to pinpoint why participants tended to make fewer fixations than would be expected by chance. The goal of this paper is to present a computational model that performs visual search as efficiently as humans. The model varied several components that may have aided visual search: memory, search strategy, and degree of parafoveal vision. Two dependent measures were used to evaluate the model: number of fixations to find the target and the distribution of saccade amplitudes. The best fitting model suggested that the biggest contribution to efficient search came from larger parafoveal vision. Search strategy, however, accounted for the distribution of saccade amplitudes.

**Keywords:** visual search; model; memory; parafovea

## Introduction

Visual search is ubiquitous. Whether we are locating an item in the grocery store, trying to find our car in a busy parking garage, or looking for an important piece of information on a web page, visual search is involved in most every task we perform. In this paper we discuss two critical components of efficient serial visual search, *the number of fixations taken to find a target* and the *strategy* used to move the eyes around the screen. Our emphasis is on *active vision* (Findlay & Gilchrist, 2003) to examine the search strategies used by people as they search for items in their environment. The goal of the current work was to devise a computational cognitive model that was capable of reproducing human visual search efficiency. A set of process models varied different cognitive capacities theorized to affect search efficiency (i.e., deliberate strategy, memory size and parafovea size) to explore the parameter space associated with serial search efficiency.

Visual search as a paradigm has been studied meticulously for the better part of the last 50 years. In that time several notable models of visual search have been proposed (Duncan & Humphreys, 1989; Treisman & Gelade, 1980; Wolfe, 1994). The paradigm itself consists of the detection of a target among a varying number of distractors. Search time has been found to be influenced by number of distractors (set size), similarity of distractors and targets, and number of features used to define a target (Davis & Palmer, 2004; Wolfe, 2003). The ease with which a target can be detected is often varied and response time data is typically used as the dependent measure.

While knowing how quickly visual information is found is important, understanding *how* that information is found is just as important—“vision is a tool, not the task” (Pelz &

Canosa, 2001, p. 3588). For this reason, the task our participants performed was not visual search, but a decision making task that, like grocery shopping or finding the car in the parking garage, just happened to require visual search. The vast majority of visual search studies have largely ignored the process of visual search, with a few notable exceptions (Zelinsky, Rao, Hayhoe, & Ballard, 1997; Araujo, Kowler, & Pavel, 2001; Unema, Pannasch, Joos, & Velichkovsky, 2005; Zelinsky, 2008). This is problematic because “visual search is more than the time taken by an observer to detect a target and press a button. It is instead a richly complex behavior having both a spatial and temporal dynamic” (Zelinsky et al., 1997, p. 448). By relying on only response time data, visual search paradigms have essentially thrown out the spatiotemporal contingencies that propel the search process. In recent years, however, a considerable effort has been put forth to connect eye movements with the underlying cognitive process (Liversedge & Findlay, 2000).

Zelinsky (2008) analyzed eye movements from participants who searched for common household items on a tabletop. The display was limited to six stationary locations where objects could appear and on each trial there was either one, three or five items to search through. Results demonstrated that eye movements were directed towards geometric centers of progressively smaller groups of objects. It should be noted that due to the limited search display (only six possible locations and up to five items visible on any given trial) the eye movement sequences were relatively short and, in practice, limited to the first three fixations. Thus, a study which has a more complex object structure and which examines longer sequences of fixations may better elucidate the visual search process. One study that looked at longer fixation traces found that fixations and saccades progress in a coarse-to-fine strategy whereby fixation durations increase while saccade amplitudes decrease as search continues (Over, Hooge, Vlaskamp, & Erkelens, 2007). Over et al. (2007) found that participants initially attended to general properties of the search environment (i.e., the lay of the land) but, as the trial progressed, gradually paid attention to specific, detailed information.

One question we can ask is whether the layout of the display facilitates and/or guides the visual search process. Others have found that external landmarks aid visual search by reducing the number of refixations on previously viewed items (Peterson, Boot, Kramer, & McCarley, 2004; Myers & Gray, 2010). In previous work, we found that segmenting



the visual search display into perceptual clusters provides a starting point for understanding where the eyes may go (Veksler & Gray, 2011). The modeling work presented here utilizes the perceptual segmentation found in our previous work to explore the efficiency of serial visual search within this paradigm. Furthermore, two metrics are used to compare human and model data: number of fixations to locate the target and the distribution of saccades around the screen during search.

The role of memory within visual search has also been greatly debated. In some instances, researchers have inferred from response time data that memory is not utilized during search (Horowitz & Wolfe, 2003; Melcher & Kowler, 2001). In other instances, it has been shown that visual search is guided by memory for previously viewed items (Korner & Gilchrist, 2007; Peterson, Beck, & Wong, 2008, 2001), that there is some memory for the search path (Dickinson & Zelinsky, 2007), and that more new locations are searched as opposed to old (Beck, Peterson, Boot, Vomela, & Kramer, 2006; McCarley, Wang, Kramer, Irwin, & Peterson, 2003). The current work also explores the role of memory within visual search, by varying the number of previously seen items that the model avoids re-fixating during search.

In summary, we use human data and computational modeling to explore the combination of components that contribute to efficient serial visual search. The components explored include search strategy, amount of memory for previously seen items, and the effective field of view. Previewing our conclusions, the major contribution to search efficiency comes from a larger parafovea. Memory plays an important role in this task, though not as an important role as we might have expected. Search strategy was explored as human data indicated that participants did not move their eyes around the screen in a random fashion, but rather transitioned across clusters of items on the screen.

## Experiment

We explore the allocation of attention during visual search when search is a subtask of a larger decision making task. The larger task was composed of the following on each trial,

1. 20 targets (represented as two-digit numbers) appeared on a *radar* screen at random locations (left-side of Figure 1). Each two-digit target subtended a  $0.62^\circ$  of visual angle.
2. Participants were provided with a list of six targets (right-side of Figure 1) and told to determine which target had the highest threat value.
3. Participants had to locate each one of the targets on the radar screen (visual search) and click on it with the mouse.
4. The target's threat value appeared next to the target. The delay between clicking and appearance varied between groups – 1, 2, 4, or 8 seconds.

5. Participants held the number and threat value of the target with the “highest threat value so far” in memory as they continued to locate other targets in the list of six.
6. When they decided that they had found the target with the highest threat value (usually, but not always, after an exhaustive search), they selected that target (with the mouse) in the list on the right hand side, and clicked on the Choose button (at the bottom right of Figure 1).

Although participants searched through the display for multiple targets on any given trial, for purposes of this paper, only the first search through the display (until the first search target is found) will be reported and modeled.

## Method

Participants were divided into four conditions which varied the duration of how long they had to wait before information (threat value of target) appeared (1, 2, 4, or 8 s). All other aspects of the task remained the same across all participants.

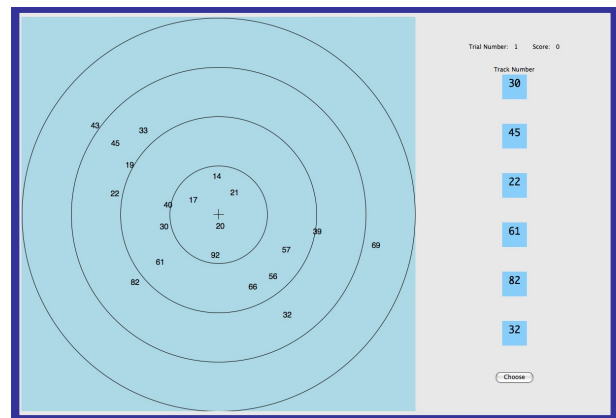


Figure 1: Task environment.

**Participants** A total of 88 participants from Rensselaer Polytechnic Institute were run during the study. Of those, 12 were excluded because their valid eye data fell below 90%, and two were excluded because their accuracy scores fell below 3 standard deviations of the group mean resulting in 74 participants included in the analyses (57 males). There were 19 participants in two of the conditions and 18 in the other two. The mean age of all participants was 18.8,  $SD=0.85$ .

**Apparatus** The experiment was presented using a computer running Mac OS X on a 17 inch flat-panel LCD monitor set to 1024x768 resolution,  $39^\circ \times 25^\circ$  of visual angle at the distance at which participants sat from the screen. The software used for the experiment was written in LispWorks 5.0. An LC Technologies eye tracker was used to collect eye data during the study at a rate of 120Hz. A chin rest was used to help ensure the accuracy of recorded eye data. Eye data quality was checked after every block of 10 trials to ensure the eye tracker was functioning and participants remained calibrated.

**Procedure** Participants were run separately. After signing informed consent forms, participants were given task instructions, calibrated to the eye-tracker and asked to keep their chin in the chinrest throughout the duration of the experiment. They also had to fixate a fixation cross prior to each trial to ensure the eye-tracker’s accuracy.

Participants completed six blocks of 10 trials (60 trials total). A mandatory 60s break was included halfway through the study. A practice block of 5 trials was included prior to the 60 experimental trials during which time the experimenter remained in the room to ensure that participants understood how to do the task and that eye data remained valid. The experiment took  $\approx 40$  minutes to complete. Each trial proceeded as described in the beginning of the Experiment section.

## Results

The majority of participants tended to search for targets in the order presented on the right hand side of the display (top to bottom). Participants tended to locate the first target they were searching for after  $\approx 8$  fixations on radar items. Since the first search in a trial was not biased by any memory effects of having found a target on a previous search, it will be used for comparison to simulation model results.

**Number of Fixations to Find Target** Table 1 summarizes the average number of fixations to locate the target, by condition in the study. A one-way ANOVA was conducted and indicated that there was not a significant effect of condition on either the total number of fixations to find the target,  $F(3, 69) = 1.27, p = .29$ , or the number of unique fixations to find the target,  $F(3, 69) = 0.63, p = .60$ . Importantly, of the fixations shown in Table 1, roughly one target is fixated twice. This pattern suggests that participants were not necessarily maintaining all of the searched items in memory.

Table 1: Average number of total and unique fixations on radar items prior to finding first target in a trial.

Condition	N	Mean Total (SD)	Mean Unique (SD)
1	18	8.38 (1.6)	7.66 (0.86)
2	19	7.57 (1.06)	7.24 (0.75)
4	19	8.2 (1.48)	7.47 (0.93)
8	18	8.0 (1.19)	7.48 (0.87)

Collapsing over conditions, Figure 2 plots the cumulative probability of finding the first target selected. A two-way ANOVA (number of fixations as a repeated factor) was run to determine whether the lockout condition influenced search efficiency. There was a significant main effect of number of fixations,  $F(43, 2967) = 2847.23, p < .001$ , no interaction,  $F(129, 2967) = 0.95, p = .63$ , and no main effect of condition,  $F(3, 69) = 0.54, p = .66$  on the probability of finding the target within that number of fixations. Therefore all data from the different conditions was collapsed to be used for compar-

ison with the models.

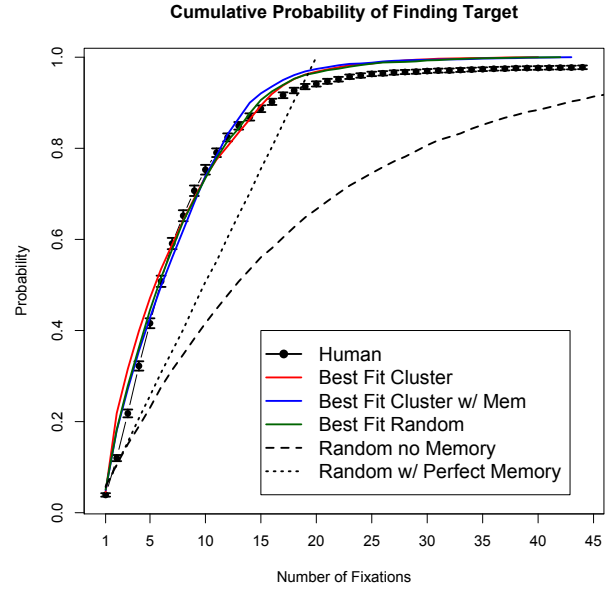


Figure 2: Cumulative Probability of Finding the Target Within Number of Fixations.

Figure 2 shows the cumulative probability of finding the initial target within  $N$  fixations, aggregated across all 74 participants. This figure also suggests that about 50% of the time participants located the target within 8 fixations. For comparison purposes, Figure 2 also shows what would be expected by chance in a model that randomly searched the radar with either no memory (dashed line) for previously seen items or perfect memory (dotted line). As can be seen, participants find the target in fewer fixations (8 on average) than would be expected by chance (10 or half of the items on the screen). This suggests that accounting only for the amount of items held in memory (so as not to refixate them) during search is insufficient to model this efficiency.

**Eye Movements and Clusters** In prior work, we derived a perceptual clustering algorithm which utilized human judgments of clusters to segment the display (Veksler & Gray, 2011). Participants in that study judged items to be part of the same cluster if they were within  $3.28^\circ$  of visual angle of each other. The algorithm adds items to a single cluster if they are less than  $3.28^\circ$  of visual angle apart, further adding more items that fall within  $3.28^\circ$  of all the items in the cluster already until no more can be added. The segmented displays generated using this algorithm were then used to determine whether search is based on clusters of targets.

Figure 3 illustrates the probability within the human eye data of a participant remaining in any given cluster given the size of that cluster. Given the eye fixation transitions, three equations were derived to fit the transition probabilities in the human data and to be later used in the model that moves its

eyes around the screen.

- Likelihood of staying in a cluster given the size of the cluster is:

$$P(\text{Stay In Cluster}) = .3292 * \ln(\text{clustersize}) - .0266 \quad (1)$$

- If participants stay within a cluster, the likelihood of them looking at the closest item to the current fixation within the cluster is:

$$P(\text{Go To Closest In Cluster}) = 1.4324 * (\text{clustersize})^{-.776} \quad (2)$$

- If participants move their gaze outside of the cluster, the likelihood of them looking at the closest item outside of the cluster is:

$$P(\text{Go To Closest Outside Cluster}) = .1888 * e^{(.0577 * \text{clustersize})} \quad (3)$$

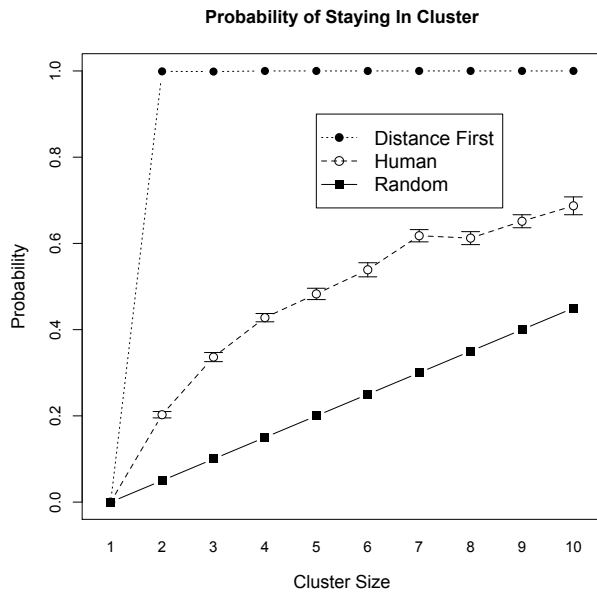


Figure 3: Probability of staying in the cluster on subsequent fixation. Distance First: prediction if participants always saccaded to closest item to current. Random: prediction if participant randomly saccaded around the screen.

**Distribution of Saccade Amplitudes** In addition to looking at the number of fixations that participants made to find the target, we also looked at the distribution of saccades (distances traveled by the eye between fixations). Figure 4 illustrates the distribution of saccade amplitudes in the human data (solid black line). As can be seen, the majority of saccades span about  $2.26^\circ$  of visual angle indicating participants moved their eyes to locations fairly close to each other. There is however, a smaller second mode around  $17^\circ$  indicating that

participants occasionally swept their eyes across larger areas of the screen. It is beyond the scope of this paper to address these larger sweeps or when they tended to occur.

## Model

Several visual search models were explored and simulated in order to model the efficiency of human serial visual search. There were three parameters that were manipulated in the modeling of the visual search process in this task. The first was the degree to which memory for previously seen items was used in the search process. The memory component essentially avoids shifting gaze to a target if it has been previously fixated within the last  $N$  fixations. The number of items held in memory was varied between 1(no memory)-19(perfect memory). It should be noted that even though we only looked at the first search within the trial, there may still be memory operating during search, particularly for previously searched locations.

The second parameter that was explored was the effective Field of View (FOV) that the model has. The model is able to shift its gaze to the target it is searching for if it notices it within its parafovea, typically about 2 to 6 degrees of visual angle around the current fixation (Reis & Judd, 2000). While the fovea is the high acuity region of the retina, up to about  $2^\circ$  of visual angle, the parafovea is a region in which acuity is not as high, with decreasing acuity as the eccentricity from the fovea increases. We explored values of 1,2,2.5, and 3 degrees of visual angle around the current fixation point providing an effective fovea+parafovea region (FOV) of 2, 4, 5, and 6 degrees, respectively.

The final manipulation had to do with the actual search strategy used. Three search strategies were explored: cluster-based, cluster-based with memory for clusters and random.

The cluster-based search model first segments the screen into several clusters based on prior empirical work (Veksler & Gray, 2011). These clusters are then used to guide the model's eye movements based on the cluster transition probabilities as per Equations 1-3. The model decides on each fixation whether or not it wants to shift attention away from the current cluster. It then decides with a certain probability to shift its gaze to either the closest item within the cluster or the closest item outside of the current cluster. The cluster-based model with memory for clusters also maintained memory for clusters it has already searched. Thus, when transitioning out of a cluster, it avoided looking to targets within previously searched clusters.

The random model search strategy is used as a baseline model. This model disregards the placement of the items on the screen and randomly chooses a target from the set of targets in the radar. The memory component was varied from a random model with no memory to one with perfect memory for targets already seen. One limitation of this model is that because the model disregards placement of items on the screen, its shifts of gaze can span long distances resulting in inefficient eye movements.

There were  $19(\text{memory store}) \times 4(\text{FOV angle}) \times 3(\text{strategies})$  models run on the radar targets used by participants in the study. Each model was run on each of the trials of human data and the number of fixations along with saccade amplitudes that were made prior to finding the first target were recorded to be compared with human data from the same set. In all, each model was run on 4559 trials.

## Results

The simulations were run to determine which models could find the targets in the radar using the same number of fixations that participants used. The cumulative likelihood of finding the first target in a trial within  $N$  fixations was derived for each model and the human data and then compared. As an additional dependent measure, fixation transitions were recorded for each model and the distribution of saccade amplitudes was compared with human data.

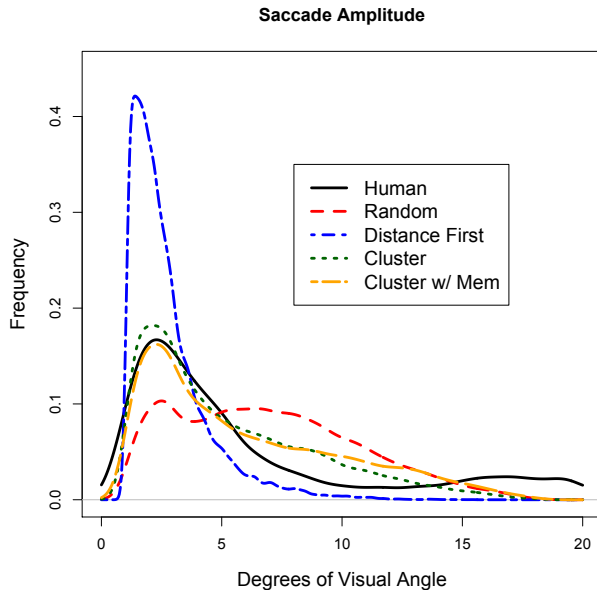


Figure 4: Distribution of saccade amplitudes across all eye data for humans and models. Models depicted are best fitting.

Search efficiency was greatly improved by the inclusion of a parafovea in all of the models (a FOV of 4, 5 or 6 degrees of visual angle). Without a parafovea (57 models), the best fit that can be achieved between human and model data has an  $\text{RMSE}=0.11$  and an  $\text{R}^2=.92$ . The model that achieves this is the cluster search model with cluster memory and memory for 16 individual targets. If we include a parafovea, 74% of the parafovea-included models surpass this fit. Therefore, the models that were next compared all had varying degrees of a parafovea.

Based on RMSE, the top 15 models all had an effective FOV of 5 degrees. The top cluster-based search model that utilized cluster memory had a memory of 4 items,  $\text{RMSE}=0.018$  and an  $\text{R}^2=.99$ . For comparison, the top

Table 2: Best fitting simulation model in each search strategy, comparing cumulative number of fixations. FOV: effective field of view in degrees of visual angle.

Search Strategy	Memory	FOV (°)	RMSE	$\text{R}^2$
Random	1	2	0.19	0.83
Random	19	2	0.04	0.90
Random	14	5	0.017	0.99
Cluster	15	6	0.025	0.99
Cluster w/ Mem	4	5	0.018	0.99

Table 3: Simulation models' results comparing saccade amplitude distributions. FOV: effective field of view in degrees of visual angle.

Search Strategy	Memory	FOV (°)	RMSE	$\text{R}^2$
Random	13	6	0.0009	.44
Distance First	16	6	0.0015	.75
Cluster	4	6	0.0004	.90
Cluster w/ Mem	8	6	0.0004	.86
Random	14	5	0.0010	.36
Cluster	15	6	0.0004	.88
Cluster w/ Mem	4	5	0.0005	.82

cluster-based search model that did not utilize cluster memory needed to remember 15 items and required a FOV of 6 degrees to attain good fit,  $\text{RMSE}=0.0252$  and an  $\text{R}^2=.99$ . The top random search model needed a memory for 14 items and a FOV of 5 degrees,  $\text{RMSE}=0.017$ ,  $\text{R}^2=.99$ . Table 2 summarizes the results of the model comparisons along with baseline comparison to the two models depicted in Figure 2. For conciseness only the best fitting models are reported.

These results suggest that for a model to be able to search as efficiently as human participants, it needs to have some amount of a parafovea and either a large memory for individual items or a small memory for individual items along with some memory for clusters searched.

Next we compared the distribution of saccade amplitudes over the course of the search in each of the models. As an added baseline, a distance-first model was run to show what would happen if the model always saccaded to the closest item to its current point of gaze. Figure 4 depicts the human data along with the best fitting models using each of the search strategies. Table 3 provides statistics for both the best fitting models (top panel) as well as the best fitting models from the cumulative number of fixations comparison (bottom panel). In terms of modeling the distribution of saccade amplitudes, both of the cluster-based search models fit the human data well. The random search and distance first model, however, have much poorer fits.

## Discussion

This work was intended to provide a computational model of the efficiency of serial visual search found in humans. Two dependent measures were used to evaluate the models generated: efficiency of search (number of fixations to locate a target) and the distribution of saccade amplitudes (how far the eye moved between fixations). It was found that incorporating a larger parafovea contributed a great deal to the efficiency with which the model was capable of finding the target. The inclusion of a memory for clusters allowed the model to have less of a need for a larger memory store for individual items searched. The cluster-based search model was also much better able to reproduce the distribution of saccade amplitudes found during human visual search, suggesting the efficacy of a search strategy based on segmentation of a display into clusters.

One limitation of the current cluster-based model and direction for future work is accounting for the longer spanning saccades as when human participants transition out of a cluster (i.e. moving to the opposite side of the screen). Another is addressing the discrepancy between the best fitting models according to the two dependent measures.

## Acknowledgments

The work was supported, in part, by grant N000141010019 to Wayne Gray from the Office of Naval Research, Dr. Ray Perez, Project Officer. Preparation of this document was performed while the main author held a National Research Council Research Associateship Award at Air Force Research Lab.

## References

- Araujo, C., Kowler, E., & Pavel, M. (2001). Eye movements during visual search: the costs of choosing the optimal path. *Vision Research*, 41(25-26), 3613–3625.
- Beck, M. R., Peterson, M. S., Boot, W. R., Vomela, M., & Kramer, A. F. (2006). Explicit memory for rejected distractors during visual search. *Visual Cognition*, 14(2), 150–174.
- Davis, E. T., & Palmer, J. (2004). Visual search and attention: an overview. *Spatial Vision*, 17(4-5), 249–255.
- Dickinson, C. A., & Zelinsky, G. J. (2007). Memory for the search path: evidence for a high-capacity representation of search history. *Vision Research*, 47(13), 1745–1755.
- Duncan, J., & Humphreys, G. W. (1989). Visual-search and stimulus similarity. *Psychological Review*, 96(3), 433–458.
- Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision*. Oxford Univ. Press.
- Horowitz, T. S., & Wolfe, J. M. (2003). Memory for rejected distractors in visual search? *Visual Cognition*, 10(3), 257–298.
- Korner, C., & Gilchrist, I. D. (2007). Finding a new target in an old display: evidence for a memory recency effect in visual search. *Psychonomic Bulletin & Review*, 14(5), 846–851.
- Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, 4(1), 6–14.
- McCarley, J. S., Wang, R. X. F., Kramer, A. F., Irwin, D. E., & Peterson, M. S. (2003). How much memory does oculomotor search have? *Psychological Science*, 14(5), 422–426.
- Melcher, D., & Kowler, E. (2001). Visual scene memory and the guidance of saccadic eye movements. *Vision Research*, 41(25-26), 3597–3611.
- Myers, C. W., & Gray, W. D. (2010). Visual scan adaptation during repeated visual search. *Journal of Vision*, 10(8.).
- Over, E. A. B., Hooge, I. T. C., Vlaskamp, B. N. S., & Erkelens, C. J. (2007). Coarse-to-fine eye movement strategy in visual search. *Vision Research*, 47(17), 2272–2280.
- Pelz, J. B., & Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Research*, 41(25-26), 3587–3596.
- Peterson, M. S., Kramer, A. F., Wang, R. X. F., Irwin, D. E., & McCarley, J. S. (2001). Visual search has memory. *Psychological Science*, 12(4), 287–292.
- Peterson, M. S., Boot, W. R., Kramer, A. F., & McCarley, J. S. (2004). Landmarks help guide attention during visual search. *Spatial Vision*, 17(4-5), 497–510.
- Peterson, M. S., Beck, M. R., & Wong, J. H. (2008). Were you paying attention to where you looked? the role of executive working memory in visual search. *Psychonomic Bulletin & Review*, 15(2), 372–377.
- Reis, H., & Judd, C. (2000). *Handbook of research methods in social and personality psychology*. Cambridge Univ Press.
- Treisman, A. M., & Gelade, G. (1980). Feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136.
- Unema, P. J. A., Pannasch, S., Joos, M., & Velichkovsky, B. M. (2005). Time course of information processing during scene perception: the relationship between saccade amplitude and fixation duration. *Visual Cognition*, 12(3), 473–494.
- Veksler, B. Z., & Gray, W. D. (2011). A tale of two problems: human judgments of visual clusters and data collection via the web vs. paper. In *Proceedings of the Human Factors and Ergonomics Society 55th Annual Meeting*. Human Factors and Ergonomics Society. Las Vegas, NV.
- Wolfe, J. M. (1994). Guided search 2.0 - a revised model of visual-search. *Psychonomic Bulletin & Review*, 1(2), 202–238.
- Wolfe, J. M. (2003). Moving towards solutions to some enduring controversies in visual search. *Trends in Cognitive Sciences*, 7(2), 70–76.
- Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, 115(4), 787–835.
- Zelinsky, G. J., Rao, R. P. N., Hayhoe, M. M., & Ballard, D. H. (1997). Eye movements reveal the spatiotemporal dynamics of visual search. *Psychological Science*, 8(6), 448–453.



# A Conceptual Network-Based Approach to Inferring the Cultural Evolutionary History of the Baltic Psaltery

**Tomas Veloz (tomas.veloz@ubc.ca)**

Department of Mathematics, University of British Columbia  
3333 University Way, Kelowna BC, V1V 1V7, Canada

**Ilya Tëmkin (itemkin@nvcc.edu)**

Biology Department, Northern Virginia Community College, Annandale, VA, 22003  
and Department of Invertebrate Zoology, National Museum of Natural History, Washington, DC 20013, USA

**Liane Gabora (liane.gabora@ubc.ca)**

Department of Psychology, University of British Columbia  
3333 University Way, Kelowna BC, V1V 1V7, Canada

## Abstract

The application of conventional phylogenetic techniques for inferring cultural history is problematic due to differences in the nature of information transmission in biological and cultural realms. In culture, units of transmission are not just measurable attributes, but communicable concepts. Therefore, relatedness amongst cultural elements often resides at the conceptual level not captured by traditional phylogenetic methods. This paper takes a cognitively inspired approach to analyzing material cultural history. We show that combining data for physical attributes of cultural artifacts with conceptual information can uncover cultural influences among different ethnolinguistic groups, and reveal new patterns of cultural ancestry. Using the Baltic psaltery, a musical instrument with a well-documented ethnographic and archaeological record, we recovered a previously unacknowledged pattern of historical relationship that is more congruent with geographical distribution and temporal data than is obtained with other approaches.

**Keywords:** archaeology; artifacts; cladistics; cultural evolution; material culture; network model; phylogeny

## Introduction

The artifacts we put into the world reveal much about the minds that conceived them. The evolutionary history of human artifacts tells the story of how our thoughts, beliefs, and understanding of the world we live in, has unfolded over the ages. Using tools and techniques that include insights from cognitive science, we are starting to piece this exciting story together.

Phylogenetic approaches to reconstructing evolutionary patterns and processes, applied routinely in systematics, are increasingly applied not just to linguistics, but also elements of material culture, such as textiles, weapons, and musical instruments (*e.g.*, Collard, Shennan & Tehrani, 2006; Forster & Toth, 2003; Mace & Holden, 2005; Shennan, 2008; Whiten *et al.*, 2011). Originally developed in biology for inferring historical relationships among groups of organisms, phylogenetics makes use of assumptions about how information is organized and transmitted that reflects peculiarities of the biology world.

The direct transfer of methodology from biology to culture has raised the question about the extent to which meaningful parallels can be drawn between the processes of change in the two domains (*e.g.*, Eldredge, 2000; Gabora, 2006). Application of phylogenetics to material culture assume that the same (or analogous) causal processes operate in culture and nature. However, what is transmitted through culture is not just the objects themselves, but rather communicable perspectives and concepts, such as notions of complementarity (*e.g.*, between a mortar and pestle, which share no attributes but clearly are related), or competition for the same cultural niche (*e.g.*, spear, gun, rope, and so forth) that may or may not be reflected in the artifact design. Indeed, some claim that the differences between biological and cultural evolution are so insurmountable that insights obtained from biology are completely irrelevant in a cultural context (Moore, 1994; Dewar, 1995; Terrell, 2001). Others such as ourselves take a more moderate stance, arguing that well there are significant parallels as well as differences between biological and cultural systems, phylogenetic techniques have limited application to culture, and it is necessary to either significant modify existing approaches, or develop altogether new ones (Eerkens, Bettinger, & McElreath, 2005; Borgerhoff-Mulder *et al.*, 2006; Gabora, 1998, 2006, 2008; Nunn *et al.*, 2006; Tëmkin & Eldredge, 2007).

In a previous paper (Gabora *et al.*, 2011) we put forward a graph theory-based approach to modelling the evolution of cultural artifacts, and applied it to a well-studied set of artifacts: early projectile points from the Southeastern United States. This data set had previously been modelled using a phylogenetic approach (O'Brien, Darwent, & Lyman, 2001), and using an earlier version of the network-based approach, upon which our model is based (Lipo, 2005). The model included reticulate relationships as well as hierarchical groupings, and incorporated conceptual information to complement physical attribute data. We showed that incorporating conceptual information that is not typically captured by the phylogenetic analysis can significantly alter the inferred pattern of historical relationships amongst artifacts.

The current paper reports on new developments of the model, most notably, a means of evaluating the relative contributions of different types of data to historical inference. In addition, we apply the approach to a very different domain, thus demonstrating its generalizability.

## The Data Set

The experimental data set is a representative selection of Baltic psalteries, a traditional plucked stringed musical instrument distributed among Baltic, Finnic, and Slavic peoples of Northeastern Europe. Until recently, the Baltic psaltery remained an integral part of secular and ritual life, and has become a national symbol for every ethnic group that has it. The origin and historical development of the Baltic psaltery has been a controversial subject for over a century (reviewed by Raynolds, 1984) and remains so to this day (Povetkin, 1989; Haas, 2001; Tëmkin, 2004).

The data on psalteries consist of extensive descriptions of structural and ornamental features, and documented (or inferred) playing styles for 13 ethnographic (dated by 17-20 centuries) and two archaeological (dated by late 10-13 centuries) artifacts, representing major pertinent ethnolinguistic groups. The data set includes two Estonian (EST), two Finnish (FIN), three Latvian (LAT), three Lithuanian (LIT), three Russian (RUS), and two presumably Slavic archaeological instruments from Novgorod, northwestern Russia (NVG).

## The Conceptual Network Approach

In this section we outline our approach. We begin by summarizing how it models attributes and concepts. We then move on to the conceptually new contribution of this paper, the use of ‘perspectives’ to bias the network structure in culturally meaningful ways.

## The Structure of Attributes and Concepts

Following convention, concepts are indicated by all capital letters (PSALTERY), whereas an actual artifact, or instance of a psaltery is indicated with all small letters (psaltery).

The more superficial level of conceptual structure consists of what Rosch (1978) refers to as *basic level concept*, such as PSALTERY, which mirror classes of objects that share a broad range of perceivable attributes. These basic level concepts can be recursively differentiated. For example, a psaltery's attribute “strings” can be differentiated into “metal” or “nylon” depending on the type of the material the strings are made of. Each of these subordinate attributes can be further resolved by introducing their respective attributes, such as, for instance, “metal type,” “nylon type,” or “color.” Some attributes of the second degree may be shared by those of the first: both metal and nylon strings have color, but only metal strings are made from a specific metal type. Hence, a basic level concept can be represented as a root of a graph and its attributes arranged by levels of descriptive resolution. Because attributes at a given level can be connected to multiple attributes of levels above and below, the resultant structure contains both hierarchical and reticulate aspects.

Basic level concepts are generalized at a more abstract level as instances of superordinate concepts, such as MUSICAL INSTRUMENT. Superordinate concepts typically refer to multiple basic level categories (e.g., MUSICAL INSTRUMENT consists of both PSALTERY and CORNET).

## Conceptual Structure and its Representation

Each artifact is represented by a network of attributes consisting of reticulated hierarchies. The attributes can be physical or non-physical (conceptual). The total network of all available attributes constitutes the conceptual structure (Figure 1). Each artifact is represented by a subnetwork of the conceptual structure. (In a sense this is conceptually similar to phylogenetic approaches where all taxa and their attributes are described as arrays of specific character states in a character state data matrix.)

The conceptual structure introduced here is based on the state context property (SCOP) theory of concepts (Aerts & Gabora, 2005a, 2005b; Gabora & Aerts, 2002) and is equivalent to a simplified ontology (Sowa, 2000). We

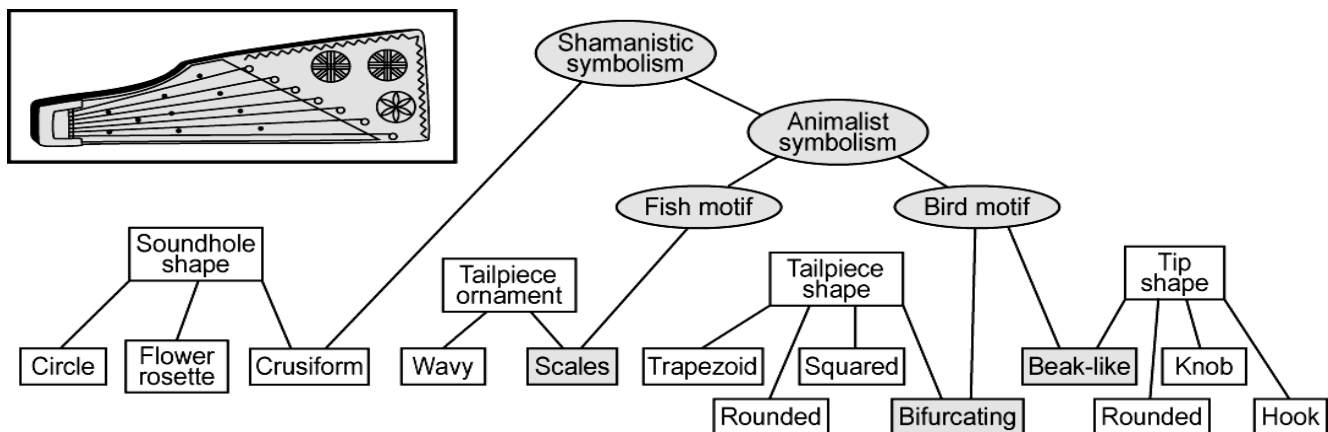


Figure 1: A segment of the conceptual structure used in the Baltic psaltery analysis. Elements in square boxes represent physical attributes of artifacts. Elements in oval boxes correspond to conceptual attributes. Shaded boxes designate attributes associated with symbolic significance, i.e., those relevant to the “Symbolism” perspective (see text for details). Note that this perspective includes both conceptual and physical attributes. Inset shows an example of a Baltic psaltery.



incorporate the notion of context by introducing the notion of *perspective* (see next section), and use graph theory instead of operator theory to develop similarity metrics.

### Incorporation of Perspectives

Given the heterogeneity of attributes, (structural vs. ornamental, physical vs. conceptual, and so forth), it may be useful to be able to explore the effect of different sets of attributes on the evolutionary pattern. Depending on the perspective from which a given artifact is considered, the emphasis is placed on a particular set of attributes. For example, a musical instrument can be viewed as an art object, a product of craftsmanship, a sound-producing device, or a sacred symbol. Given a particular perspective, a set of attributes may include physical features (e.g., presence of a soundhole), conceptual non-physical descriptors (e.g., ritualistic function), or a mixture of both. Because physical and non-physical attributes represent different culturally relevant aspects of the artifact, including both types of data in the analysis can potentially produce a more comprehensive, and presumably, more realistic pattern of evolutionary history for a collection of artifacts as a whole.

This is accomplished by defining a *perspective*, a part of the conceptual structure that includes a predefined subset of attributes that may or may not be directly linked to each other. Multiple perspectives can be defined for a given conceptual structure, and they may or may not have attributes in common. In the analysis of the Baltic psalter, we defined three perspectives: (1) *Physical Attributes*, containing structural or decorative features of actual artifacts; (2) *Performance*, containing concepts related to music performance styles; and (3) *Symbolism*, containing concepts or physical attributes associated with sacred symbolism and ritualistic significance. Perspectives in the conceptual network approach are, in a sense, analogous to character partitions in biological phylogenetics.

### Reliability and Similarity

In the proposed approach, similarity among artifacts is assessed by pairwise comparison of their network representations. The comparisons can be made between entire sub-networks corresponding to complete representations of the two artifacts, or with respect to a particular perspective, or set of perspectives. To formalize this notion of similarity, we assume graph representations of two artifacts,  $a$  and  $a'$ , and a perspective  $p$ . We introduce two functions:

$$O(a, a', p) = |a \cap a' \cap p| \quad (1)$$

$$D(a, a', p) = |a \cap p| + |a' \cap p| - 2O(a, a', p) \quad (2)$$

where  $|x|$  is the number of vertices of  $x$ , and  $x \cap y$  is the set of common vertices between  $x$  and  $y$ .  $O(a, a', p)$  is the *overlap* and  $D(a, a', p)$  is the *divergence* between artifacts  $a$  and  $a'$  with respect to  $p$ . The overlap and divergence

account for the number of attributes included in  $p$  that are shared or non-shared, respectively, by  $a$  and  $a'$ .

Accounting for both the overlap and divergence is critical to determine the similarity of two artifacts. Overlap alone can lead to an overestimation of overall similarity in some situations, such as in a trivial case where one of two artifacts possesses just one attribute included in a particular perspective and this attribute is shared by another artifact that has a greater number of attributes included in the same perspective. In this case, the failure to account for divergence will erroneously interpret the complete overlap as absolute similarity between the artifacts with respect to the chosen perspective.

Some perspectives capture more of the overlap and divergence between artifacts than others. Indeed, a (non empty) perspective may include none of the attributes of the artifacts being compared, or it may include them all. In general, some portion of the total overlap and divergence between two artifacts is captured by a perspective. We define the *reliability*  $R(p, a, a')$  as the proportion of the overlap and divergence between artifacts  $a$  and  $a'$  given perspective  $p$  as follows:

$$R(p, a, a') = (O(a, a', p) + D(a, a', p)) / (|a| + |a'|) \quad (3)$$

Note that the entire conceptual structure can also be considered as a perspective. In this case, its reliability is equal to 1 for any given pair of artifacts. This is because it contains all possible concepts used to represent each artifact. Hence, the whole conceptual structure, and more inclusive perspectives in general, have greater reliability. However, the notion of reliability in itself may not be sufficient as an estimation of the perspective's effect on the similarity between artifacts. There may exist a small portion of the conceptual structure such that if considered as a perspective it would have small reliability, but which may nevertheless be vital for establishing the similarity of some artifacts. Therefore, given a set of perspectives  $P = \{p_1, \dots, p_n\}$ , we introduce a *perspective weight vector*,  $V = \{v_1, \dots, v_n\}$ , which gives the relative degree of importance of  $p$  in  $P$ , and defines the *similarity* between two artifacts  $a$  and  $a'$  with respect to  $P$  by the following formula:

$$S(P, V, a, a') = \sum_{i=1}^n v_i R(p_i, a, a') (O(a, a', p_i) - D(a, a', p_i)) \quad (4)$$

The *similarity*  $S(P, V, a, a')$  takes into account the overlap and divergence between their graph representations summed over all perspectives, each weighted by their respective reliability values and perspective weight vectors. Note that the greater the reliabilities of the perspectives in  $P$ , the greater the similarity between  $a$  and  $a'$ .

The approach provides a means of exploring the effect of individual perspectives on evolutionary inference. There are several ways of going about this, ranging in objectivity from fully automated, unbiased naive models to sophisticated expert-specified models that allow for incorporation of background information: (1) a *uniform*

*weights* model, which weighs all perspectives equally; (2) an *implied weighting* model, which weighs each perspective proportional to its reliability; (3) a *sensitivity* model, which explores a range of weights allowing for identification of most and least stable relationships; and (4) an *expert choice* model, which enables the user to specify unique weights (including removing a selected perspective from the analysis) to each perspective.

## Similarity Graph and Cultural History

A pairwise similarity matrix based on comparing all pairs of artifacts with respect to a chosen perspective (or set of perspectives), is used to compute a *similarity graph* where each vertex (node) corresponds to an artifact, and an edge implies that the extreme vertices of the edge are similar. The edge is labelled by the similarity weight between the two artifacts. The graphical representation of the similarity graph can subsequently be interpreted as a historical pattern of relationships amongst included artifacts. It is important to emphasize that the similarity graph is *not* a cultural phylogeny, but a mere representation of similarity among artifacts: it does not incorporate explicit actual cultural transmission models but provides an independent framework for establishing historical hypotheses that can corroborate or disagree with existing models of cultural change. We restrict ourselves to the *maximal similarity graph*, which connects each artifact to only those that have the highest similarity to it (in most cases, for each artifact there is only one most similar artifact). The maximal similarity graph provides an approximation to the artifact's true cultural history.

## Computational Implementation

The program we use to infer cultural lineages was developed using the object-oriented Java platform with extension packages for working with networks (JUNG). It allows for the creation of a conceptual structure by adding nodes (concepts) and edges (conceptual relationships). Perspectives and artifacts can be generated as well. One can specify the entries of the perspective weight vector using an array of sliders (one slider is automatically created for each perspective the user creates). Other software functions allow the user to export and import these structures for later use. The currently implemented default weighting scheme is the *implied weighting* model, which weighs each perspective proportional to its reliability. By modifying the perspective weights, the user can recompute the similarities, and visualize the resulting changes to the similarity graph. This enables exploration of the resulting similarity graphs found in different regions of perspective weight space.

## Results

We present the patterns of relationship obtained for the Baltic psalteries with respect to two perspectives: *Physical Attributes* and *Symbolism*. When only physical attributes of the psalteries are considered, the resulting similarity graph recovers clusters of instruments corresponding to

ethnolinguistic affinities with the exception of the Baltic instruments that appear to be more dispersed (forming three lineages; Figure 2A). This is consistent with previous results based on maximum parsimony analysis (Tëmkin, 2004; Tëmkin and Eldredge, 2007). On the other hand, when only symbolic aspects are taken into account, such sharp delineation based on the linguistic affinity becomes less evident and novel relationships emerge (Figure 2B). For example, in the former scenario, the Baltic instruments were linked to the Estonian (Finnic) ethnographic instruments, whereas in the latter they have no connection with the Estonian instruments, and display a novel connection with the Slavic instruments.

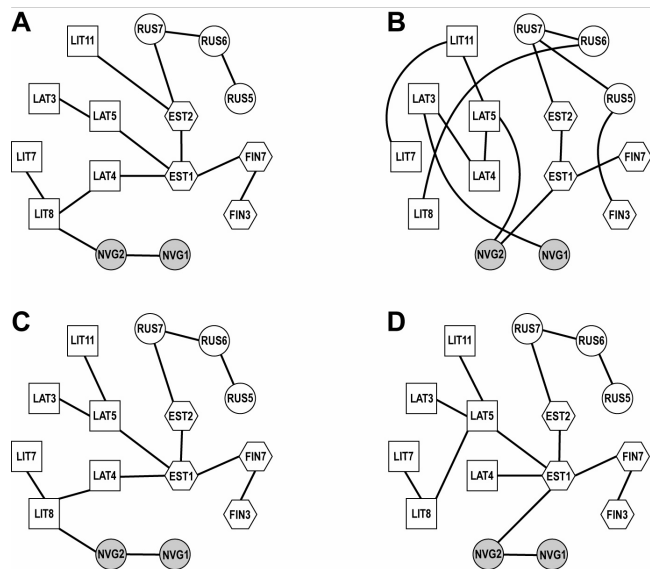


Figure 2. Similarity graphs based on the conceptual network analysis of Baltic psalteries under different perspective weighting schemes. (A) *Physical attributes*; (B) *Symbolism*; (C) *Physical attributes* and *Symbolism* (equal weights); (D) *Physical attributes* (25% weight) and *Symbolism* (75% weight). Each node corresponds to a single artifact. The node shapes indicate ethnolinguistic groups: Slavic (circle), Finnic (hexagon), and Baltic (square). Shaded nodes designate archaeological instruments (10-13 cc.); the remaining nodes correspond to ethnographical instruments (17-20 cc.).

Such incongruence in similarity inference between the two perspectives suggests that the constructive principles of the psalteries were regionally constrained and possibly insured by master and apprentice relationship which has a strong linguistic component. Similarity in symbolic elements across cultures, however, appears to correspond more closely to the geographic proximity and, possibly, some symbolic features spread as decorative designs without affecting structural aspects of local musical instrument making traditions.

When both perspectives are analyzed simultaneously under equal weights, the resulting similarity graph is similar to those based on the analysis of physical attributes alone, but results in greater similarity amongst the Baltic

instruments (Figure 2C). This congruence between the two graphs can be accounted for by a much greater number of physical (80) than symbolic (20) attributes.

To reveal the impact of symbolic attributes given the situation that they were outnumbered by physical attributes, the data set was re-analyzed with 25% and 75% weights for the *Physical Attributes* and *Symbolism* perspectives respectively. The resulting similarity graph was identical to the results of the analysis under equal-weight with respect to the relationships among the Baltic, Finnic, and Slavic ethnographical instruments with two significant differences (Figure 2D). First, the relationships among the Baltic instruments were stronger (as five out of 6 instruments formed a single cluster). Second, the connection of most ancient, archaeological instruments shifted from the Baltic to the Finnic instruments.

The most comprehensive cladistic analysis of the Baltic psaltery failed to unequivocally resolve which of the three groups of instruments (Baltic, Finnic, or Slavic) were more closely related to the root of the tree, the medieval artifacts from Novgorod in northwestern Russia (Tëmkin & Eldredge, 2007). The recovery of the Lithuanian (Baltic) psaltery as most basal agrees with a presumed northward diffusion of the instrument (the second wave of dispersal) in (Tëmkin, 2004). The alternative scenario, in which the Novgorodian instruments bear greater similarity to the Finnic instruments (largely attested by shared symbolic significance) suggests an intriguing hypothesis for interpreting the instrument's history as it is more consistent with archaeological data which indicates that medieval Novgorod, where most ancient Baltic psalteries were discovered, had a substantial proportion of Finnic population (Tönurist, 1977).

## Discussion

By fusing physical information about artifacts with the conceptual information our ancestors were using to *create* these artifacts, we arrive at a more accurate picture of the evolutionary trajectories by which the artifacts evolved, and by which our human understanding of the world took shape. This paper develops the conceptual and mathematical foundation for a novel approach to reconstructing patterns of cultural evolutionary history based on graph theory. It uses algorithms for constructing and displaying similarity graphs that can be biased by conceptual knowledge from different domains. This approach circumvents the limitations of traditional phylogenetic approaches by (1) allowing for simultaneous analysis of cognitive information and physical character data, (2) providing the means for evaluating relative contributions of different types of data to historical inference, and (3) expanding hierarchical approach to include reticulate relationships.

We tested the utility of the conceptual network approach for inferring historical patterns of relatedness amongst artifacts by applying it the analysis of the Baltic psaltery, a stringed musical instrument unique to northwestern Europe. Not only did the approach capture the essential features of

the instrument's history inferred previously using other methods, it also provided new insights that invite a novel interpretation of the instrument's evolution. To achieve this, it was necessary to distinguish between hierarchically organized conceptual attributes (largely pertaining to sacred symbolic imagery), and physical characteristics (such as elements of the instrument's construction). Although this was readily accomplished with the current approach, it cannot be done with other approaches. The approach is not limited to a specific data types; it can be extended to include linguistic information, or any other discrete character information.

## Future Directions

Although limited in its present formulation, the conceptual network approach provides large number of avenues for future theoretical and mathematical developments. We are currently developing methods for computing the expected reliability of perspectives and determining the range of weights (parameter space) in which a given perspective becomes significant in the similarity computation. Other immediate plans include developing sophisticated software for analysis of the conceptual network.

In this investigation, we only explored the maximal similarity graph, which connects each artifact to the artifact that has the highest similarity to it according to the chosen perspective weight vector (first neighbours graph). In future investigations, we will consider the second or further most similar artifacts ( $n^{\text{th}}$  neighbour similarity graph), and establish a mechanism to automatically split the perspective space according to different criteria of similarity graph equivalence. This will enable us to explore the conglomeration of similarity graphs obtained from the different sets of perspective weight vectors. To further refine the approach, we will construct similarity graphs not only considering the most similar connections of each artifact, but connections above a certain threshold. This could reveal different similarity ranges where the parameter space is split in qualitatively different ways.

Finally, we plan to investigate the applicability of Kemp and Tenenbaum (2008) structure discovery models.

## Acknowledgments

We would like to acknowledge grants to Liane Gabora from the *Social Sciences and Humanities Research Council of Canada*, the *Natural Sciences and Engineering Research Council of Canada*, and the *Concerted Research Program of the Fund for Scientific Research of Belgium*.

## References

- Aerts, D., & Gabora, L. (2005a). A state-context-property model of concepts and their combinations I: The structure of the sets of contexts and properties. *Kybernetes*, 34(1&2), 167–191.



- Aerts, D., & Gabora, L. (2005b). A state-context-property model of concepts and their combinations II: A Hilbert space representation. *Kybernetes*, 34(1&2), 192–221.
- Borgerhoff Mulden, M., Nunn, C. L., & Towner, M. C. (2006). Macroevolutionary studies of cultural trait transmission. *Evolutionary Anthropology* 15, 52–64.
- Coley, J. Hayes, B., Lawson, C., & Moloney, M. (2004). Knowledge, expectations, and inductive inferences within conceptual hierarchies. *Cognition* 90, 217–253.
- Collard, M., Shennan, S. L., & Tehrani, J. J. (2006). Branching, blending, and the evolution of cultural similarities and differences among human populations. *Evolution and Human Behavior* 27, 169–184.
- Dewar, R. E. (1995). Of nets and trees: Untangling the reticulate and dendritic in Madagascar prehistory. *World Archaeology* 26, 301–18.
- Eerkens, J. W., Bettinger, R. L., & McElreath, R. (2005). Cultural transmission, phylogenetics, and the archaeological record. In Lipo, C. P., O'Brien, M. J., Collard, M., & Shennan, S. (Eds.) *Mapping Our Ancestors: Phylogenetic Approaches in Anthropology and Prehistory*. New York: Aldine de Gruyter.
- Eldredge, N. (2000). Biological and material cultural evolution: Are there any true parallels? *Perspectives in Ethology* 13, 113–53.
- Forster, P. & Toth, A. (2003). Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proceedings of the National Academy of Sciences USA* 100(15), 9079–9084.
- Gabora, L. (1998). Autocatalytic closure in a cognitive system: A tentative scenario for the origin of culture. *Psycoloquy*, 9(67). [[adap-org/9901002](http://adap-org/9901002)]
- Gabora, L. (2006). The fate of evolutionary archaeology: Survival or extinction? *World Archaeology* 38(4), 690–696.
- Gabora, L. (2008). The cultural evolution of socially situated cognition. *Cognitive Systems Research*, 9(1-2), 104–113.
- Gabora, L., Leijnen, S., Veloz, T., & Lipo, C. (2011). A non-phylogenetic conceptual network architecture for organizing classes of material artifacts into cultural lineages. *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 2923–2928). July 20–23, 2011, Boston MA.
- Gabora, L., & Aerts, D. (2002). Contextualizing concepts using a mathematical generalization of the quantum formalism. *Journal of Experimental and Theoretical Artificial Intelligence*, 14(4), 327–358.
- Haas, A. (2001). Intercultural conflict and the evolution of the Baltic psaltery. *Journal of Baltic Studies* 32, 209–250.
- Holden, C. J., & Mace, R. (2003). Spread of cattle led to the loss of matriline in Africa: a co-evolutionary analysis. *Proceedings of the Royal Society B: Biological Sciences* 270(1532), 2425–2433. ISSN: 0962-8452
- Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*. 105(31), 10687–10692.
- Lipo, C. P. (2005). The resolution of cultural phylogenies using graphs. In Lipo, C. P., O'Brien, M. J., Collard, M., & Shennan, S. J. (Eds.) *Mapping our Ancestors*. New Brunswick: Transanunnction Publishers.
- Mace, C. J., & Holden, R. (2005). A phylogenetic approach to cultural evolution. *TRENDS in Ecology and Evolution*, 20(3), 116–121.
- Moore, J. H. (1994). Putting anthropology back together again: The ethnogenetic critique of cladistic theory. *American Anthropologist*, 96, 925–48.
- Nunn, C. L., Borgerhoff Mulden, M., & Langley, S. (2006). Comparative methods for studying cultural trait evolution: A simulation study. *Cross-Cultural Research* 40(2), 177–209.
- O'Brien, M. J., Darwent, J., & Lyman, R. (2001). Cladistics is useful for reconstructing archaeological phylogenies: Paleoindian points from the Southeastern United States. *Journal of Archaeological Science* 28, 1115–1136.
- Povetkin, V. I. (1989). O proiskhozhdenii guslei s igrovym oknom: iz opyta vosstanovitelnykh rabot. In *Istoriya i Kultura Drevnerusskogo Goroda*. Moscow: Moscow University Press.
- Raynolds, S. (1984). Whence the Baltic psaltery? *Paper presented at the Ninth Conference on Baltic Studies*, (pp. 1–22). June 14–16, 1984, Montréal, Canada.
- Rosch, E. (1978). Principles of categorization. In Rosch, E. & Lloyd, B. (Eds.) *Cognition and categorization*. Hillsdale, NJ: Lawrence Erlbaum.
- Shennan, S. 2008. Evolution in archaeology. *Annual Review of Anthropology*, 37, 75–91.
- Sowa, J. (2000). Ontology, Metadata, and Semiotics, ICCS'2000, Darmstadt, Germany, August 14, 2000. Published in B. Ganter & G. W. Mineau, eds., *Conceptual Structures: Logical, Linguistic, and Computational Issues*, Lecture Notes in AI #1867, Springer-Verlag, Berlin, 2000, 55–81.
- Tëmkin, I. (2004). The evolution of the Baltic psaltery: A case for phyloorganology. *The Galpin Society Journal* 57, 219–30.
- Tëmkin, I., & Eldredge, N. (2007). Phylogenetics and material cultural evolution. *Current Anthropology* 48, 146–153.
- Terrell, J. E., Hunt, T. L., & Gosden, C. (1997). The dimensions of social life in the Pacific: Human diversity and the myth of the primitive isolate. *Current Anthropology* 38, 155–95.
- Terrell, J. E. (2001). Introduction. In J. E. Terrell (Ed.) *Archaeology, language, and history: Essays on culture and ethnicity*. Westport, Conn.: Bergin and Garvey.
- Tõnurist, I. (1977). Kannel vepsamaast setumaani. In I. Rüütel (Ed). *Soome-Ugri Rahvaste Muusikapärandist*. Tallinn: Eesti Raamat.
- Whiten, A, R. A. Hinde, K. N. Laland, & C. B. Stringer. 2011. Culture evolves. *Philosophical Transactions of the Royal Society B*, 366, 938–948.

# Factors influencing children's display of surprise

Mandy Visser<sup>a</sup> (Mandy.Visser@uvt.nl)

Emiel Krahmer<sup>a</sup> (E.J.Krahmer@uvt.nl)

Marc Swerts<sup>a</sup> (M.G.J.Swerts@uvt.nl)

<sup>a</sup>Tilburg centre for Cognition and Communication (TiCC), School of Humanities, Tilburg University,  
PO Box 90153, 5000 LE Tilburg, The Netherlands

## Abstract

Earlier studies found a discrepancy between the display and feeling of surprise. Therefore, we assessed what factors influence the display of surprise in children of two age groups: 8- and 11-year-olds. We manipulated the social setting (children either competed or collaborated), and the cause of surprise (a surprisingly positive or negative event). We found that children used more features to express negatively caused surprise, compared to positively caused surprise and that 11-year-olds used more facial features than 8-year-olds. In a subsequent perception study, adults judged video clips with surprised and neutral reactions, for the degree of surprise that was displayed. We found higher ratings of surprise for negatively vs. positively surprised children, competing vs. collaborating children, and 11-year-olds vs. 8-year-olds. These results confirm that in addition to the feeling of surprise, its cause, the social setting, and age also affect the display of surprise.

**Keywords:** Emotions; facial expression; surprise; collaboration; competition; social development.

## Introduction

According to Ekman (1997), surprise is prototypically displayed through a combination of three facial features: people who are surprised raise their eyebrows, open their mouth and widen their eyes. However, when researchers try to elicit the expression of surprise with participants, this prototypical display is rarely shown (e.g., Reisenzein, Bördgen, Holtbernd & Matz, 2006). There appears to be a low emotion-facial display ratio, which means that when participants indicate they feel surprised, they do not frequently use the complete set of Ekman's features to express their emotion. We focus on two possible reasons for this.

First, earlier research often assumes that there is only one sort of surprise (e.g., Scherer, Zentner, and Stern, 2004). However, some studies have shown that surprise can be differently expressed, depending on its cause (e.g., Shepperd & McNulty, 2002). For example, when someone is surprised by a positive event, like giving an unexpectedly correct answer to a difficult question, his or her facial expression would differ from the facial expression of someone who is surprised by a negative event, like giving an unexpectedly incorrect answer to an easy question. This could be a reason for not finding a general full facial expression of surprise. Therefore, in our studies, we look at

different causes of surprise and its accompanying facial expressions.

Second, there may be contextual effects as well, in the sense that the expression of surprise could depend on the setting in which it is elicited, something which is often ignored in earlier studies. Research on surprise usually takes place in a nonsocial situation, sometimes even deliberately, to get a "clean" view on the expression of surprise (Reisenzein et al., 2006). However, we think that the social setting might be an important indicator for the expression of surprise. People tend to exaggerate, minimize, neutralize and fake expressions, depending on the social situation they are in (Matsumoto, Hee Yoo, Hiramaya & Petrova, 2005). For example, people probably express their surprise about an incorrect answer differently in the company of a teammate than in the company of an opponent. Such social factors may represent a second reason for a low emotion-facial display ratio. Hence, in this research, we examine the possible effect of the social situation on the facial expression of surprise more closely. We created a quiz game that could be played in two conditions, namely a collaborative or competitive setting. In this way, the participants' goals and interests varied in each condition and thus a different social context was created.

Adjusting behaviour and expressions to a social context is something people learn gradually. Children need time to acquire the social display rules, which means that younger children are not as skilled in that respect, compared to older children (Piaget, 1950). Earlier studies on children and surprise involved mainly perception and understanding of the emotion based on the theory of mind (e.g., Hadwin & Perner, 1991); they rarely concerned children's expression of surprise. Research that did study at children's expression of surprise involved merely infant participants (e.g., Scherer, Zentner, and Stern, 2004). We think it is important to study the expressions of surprise with older, more socially skilled, age groups as well. Saarni (1979) showed that children's adjusting behavior to social contexts doubles between the ages of 8 and 11. Therefore, we included both age groups (8 and 11 years old) in our studies.

In sum: the aim of this research is to study children's expressions of surprise, caused by different events (positive and negative events), in different social situations. To assess the influence of social setting and age on facial expressions of different sorts of surprise, we conducted three studies. The aim of the first study was to examine whether

participants actually experienced the different kinds of surprise differently. In our second study, we wanted to know whether features of the full facial display of surprise appeared, and whether there was an effect of age and social setting. In our third study, we focused on the perception of the facial expressions of surprise.

### Study 1: Production of surprise

We wanted to elicit different kinds of surprise using a natural elicitation procedure. Therefore, we created a game-based experiment in which two participants simultaneously had to play a knowledge quiz. In this quiz, we manipulated various questions to induce situations in which a quiz partner's answer was unexpectedly correct, or unexpectedly incorrect, in order to elicit either a surprised feeling with a positive cause, or a surprised feeling with a negative cause.

### Method

**Participants.** In total, 90 children participated in this study. We selected participants from two age groups; 8-year-old children (42 children in total, 45% girls) and 11-year-old children (48 children in total, 56% girls). The participants had to play a knowledge quiz in self-selected pairs. These pairs were randomly divided across two experimental conditions; half of the pairs played the game in a competitive setting and half of them in a collaborative setting. The experiment was conducted in two primary schools in Zoetermeer, the Netherlands. Beforehand, we informed parents about the experiment and asked for their signed permission for their child to participate.

**Stimuli.** The knowledge quiz consisted of 30 questions, which participants had to answer by taking turns, such that each of them responded to 15 questions. Both participants saw a question on their respective screens, but only one participant had to give an answer, while the other just listened to the response. For the next question in the list, they changed roles so that the other participant would answer a question, and vice versa. These questions were selected from the children's edition of the game *Triviant Pursuit* and a Dutch version of the "Wechsler Intelligence Scale for Children". We made sure that both easy and hard questions were included, in order to elicit both correct and incorrect answers. An example of an easy question is "Which month follows March?", an example of a difficult question is "What is glass made of?"

The participants were asked to sit behind two separate computer screens, which were arranged in such a way that they were not able to see each other or each other's computer screen (see Figure 1), but they were able to hear each other's answers. Participants were led to believe that they both saw the same list of questions on their computer screens. However, unknown to the participants, in order to elicit a surprise reaction, the questions posed were different for the two participants. In doing so, we could manipulate various questions to create situations in which the speaking participant's answer was unexpectedly correct, or

unexpectedly incorrect, according to the knowledge of the listening participant. More specifically, we aimed to elicit reactions of two types of surprise.

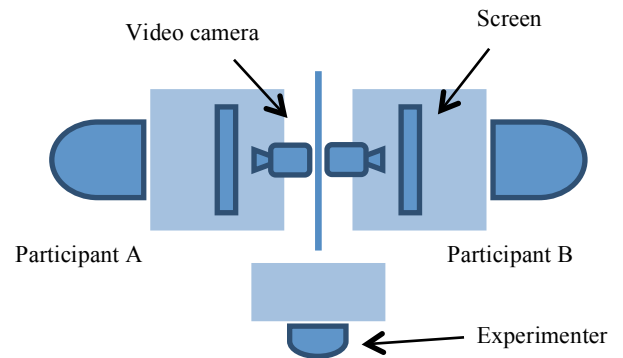


Figure 1. *Experimental setting*

First, we manipulated questions in such a way that participants were positively surprised. We showed the answering participant a question that was easy to answer, while the listening participant saw a question that was extremely difficult to answer. For example, the answering participant was given the question: "Which year follows 1933?", a question that is likely to be very easy to answer. However, simultaneously, the listening participant saw the question: "In which year was the city Tilburg established?" which is a difficult question. So to the listening participant, it would probably come as a positive surprise that his/her partner would give a quick and confident sounding response to this complex question.

Second, we also tried to elicit what we call surprise with a negative cause. We showed an easy question to both answering and listening participants, but these questions were not similar. For example, the answering participant was given the question: "Which animals live on a farm and roll in the mud?" while the listening participant saw the question: "Which animals live in an aquarium?" So to the listening participant, it would probably come as a negative surprise that his/her partner would give an incorrect answer to a relatively easy question.

For each pair of participants, we manipulated four questions to elicit surprise with a positive cause, and four questions to elicit surprise with a negative cause. This means that each participant answered two positively manipulated questions and two negatively manipulated questions, and listened to two positively manipulated answers and two negatively manipulated answers.

**Procedure.** Before the start of the quiz game, the pairs of participants were randomly assigned to a competitive or collaborative condition. Participants were told that they were going to play a knowledge quiz together and that they had to take turn in answering the questions that appeared on a screen. They were told to answer as many questions correctly as possible together (collaborative setting), or that they were playing against each other, and that they had to

compete to get the most correct answers (competitive setting). To emphasize this social setting, participants wore same colored T-shirts in the collaborative setting, and T-shirts with different colors in the competitive setting. Apart from the color of the T-shirts and the introduction given by the experimenter, the procedure was exactly the same for both conditions.

The participants' face and upper body were filmed by a video camera. After each answer, both participants had to indicate how certain they were about its correctness. In this way, we could see whether children indeed thought that the answers given by their opponent or team member were correct or incorrect, and check whether our manipulations worked properly. Participants had to indicate this certainty of correctness on a five-point Likert scale, by pointing out specific facial representations of the items to the camera. For example, a very unhappy face (corners of the mouth pulled down) represented a score of 1 (very uncertain about the correctness), and a very happy face (corners of the mouth pulled up) represented a score of 5 (very certain about the correctness). These facial representations of Likert scales are fairly standard for studies involving children (e.g., Lockl & Schneider, 2002) and our participants acknowledged that they are easy to use.

All pairs of participants began the experiment with a training part to ensure they were familiar with the quiz and the social setting they were in. This training phase consisted of ten questions with different levels of difficulty (five for each participant, without using any manipulations). To stimulate participants to try their best and to emphasize the social setting pairs were in (competition or collaboration), they were told that (depending on the condition) the best individual or the best team of the class would receive a prize. In addition, after participating in the experiment, all participants received a pencil and eraser as appreciation for their contribution.

## Results

We first checked whether our manipulations to elicit different types of surprise had worked by computing a difference score from the certainty scores of both answer-giving and listening participants. We expected these difference scores to diverge, in such a way that with a negative manipulation, the answering participant was sure that his/her answer was correct, and the listening participant was sure the answer was incorrect. For positive manipulations, we expected the answering participant to believe that his/her answer was correct, and the listening participant not to know the correct answer, which means that the listening participants would have to be less certain about the correctness. In the baseline condition, we expected both participants' certainty scores to be approximately the same.

We used analysis of variance (ANOVA) with the surprise manipulation (baseline, positive manipulation and negative manipulation) as between-subjects factor and the difference score as dependent variable. As shown in figure

2, there is an effect of the surprise manipulation, as reflected in the differences between speaker's and listener's certainty scores,  $F(2,43) = 80.101$ ,  $p < .001$ . A Bonferroni post hoc test showed that for the baseline condition ( $M = 0.22$ ,  $SD = 0.77$ ), the difference in speaker and listener's certainty score is significantly smaller, than with both the surprise manipulations. Moreover, the difference score for positive manipulations ( $M = 1.60$ ,  $SD = 1.12$ ) is in turn significantly smaller than for negative manipulations ( $M = 3.07$ ,  $SD = 1.32$ ).

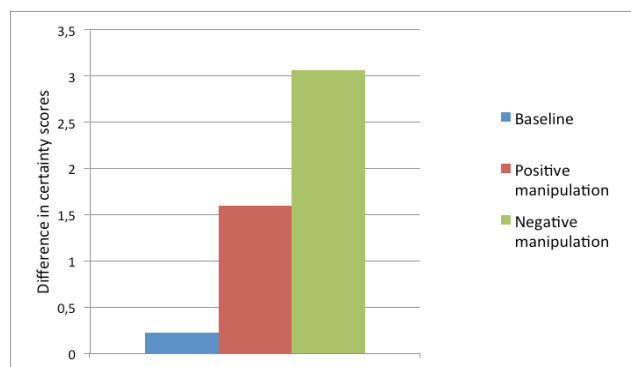


Figure 2. Differences in certainty scores for surprise manipulations.

## Discussion

While participants played a knowledge quiz in pairs, we tried to elicit expressions of surprise by manipulating the situation in which a partner's answer would be unexpectedly correct or unexpectedly incorrect. From comparing the certainty scores for both manipulated questions and regular questions, we can presume that the manipulations generated situations that differed regarding the experienced degree of surprise. For negative manipulations, there was a large difference in certainty scores, positively caused surprise resulted in a smaller discrepancy between participants' certainty scores, and finally, certainty scores for baseline answers hardly differed from each other.

Next, we wanted to know whether the type of surprise had any effect on the facial expressions of the participants. Therefore, in the next part of this research, we investigate how children express the two differently caused emotions of surprise and whether this differs for the two age groups and social settings.

## Study 2: Expression of surprise

The aim of the second study is to gain insight into children's facial expression of surprise caused by different events. We investigate the presence of Ekman's full facial display of surprise in the data collected from our game based experiment.

## Method

**Stimuli.** We selected 96 video clips of children listening to the answer to a question by the other participant during the



quiz game, with an equal distribution of positively and negatively caused surprise manipulations and baseline condition (no manipulation), and with an equal distribution across age and social settings. This gave a 3 x 2 x 2 design (surprise x age group x social setting). We randomly selected three questions per couple for labelling purposes. For the positively caused surprise manipulation, we used all video clips with reactions to the answer to the “1933” question (in which the answering participant was given the question: “Which year follows 1933?”, a question that is very easy to answer. Simultaneously, the listening participant saw the question: “In which year was the city Tilburg established?” which is a difficult question) and for the negative surprise manipulation, we used the reactions to the answers to the “pigs in an aquarium” question (in which the answering participant was given the question: “Which animals live on a farm and roll in the mud?” while the listening participant saw the question: “Which animals live in an aquarium?”). As a baseline condition, we used the reactions to a third, easy, question without manipulation. For thirteen pairs the manipulation did not work properly, according to the comparison of the feeling of correctness scores of both participants, therefore we used data from 32 out of 45 pairs.

The selected video clips contained the listening participants’ reactions to the speaking participants’ answers, from the moment the question appeared on the screen until the next question was shown. For the purpose of labeling, all certainty scores presented by the children on the smileys were blurred and the video clips were presented without any sound.



Figure 3. Stills illustrating the three labeled features (left: eyebrow movement, middle: eye widening and right: mouth opening)

**Labeling and annotation.** Two independent labellers, who were blind for experimental condition (age, manipulation and social setting), manually coded all selected clips of listening children. Following an explicit procedure, they labeled the presence or absence of the features that represent the full facial display of surprise. According to Ekman (1997), this full facial display of surprise consists of three features: 1) moving the eyebrows, 2) dropping jaw or opening the mouth and 3) widening the eyes. For representative examples of the labeled features, see figure 3. Before labeling, coders had a short training phase, to make sure both coders labeled the video clips in the same way. All Kappa’s indicated acceptable inter coder agreement (Kappa’s were .64 for brow movements, .70 for eye opening

and .69 for mouth opening). Inconsistent labels were discussed until consensus was reached.

## Results

We used an ANOVA for analyzing the appearance of features belonging to Ekman’s (1997) full facial display of surprise in the video clips, with our surprise manipulations, age and social settings as between subject factors. Analysis of the full facial display of surprise (by counting up scores of separate features) shows an effect of the sort of surprise,  $F(2,96) = 27.836$ ,  $p < .001$ . A posthoc test (Bonferroni method) reveals a significant difference in the appearance of the full facial display of surprise between all three conditions (Baseline:  $M = .19$ ,  $SD = .134$ ; Positive manipulation:  $M = 1.00$ ,  $SD = .134$ ; Negative manipulation:  $M = 1.59$ ,  $SD = .134$ ).

When we take a closer look at the features, they all appear to be affected by the surprise manipulation. Table 1 shows that both brow movement and mouth opening are significantly more present in manipulated conditions than in the baseline condition, while a similar trend can be observed for eye widening.

Table 1. Percentages of appearance features in baseline condition, positive and negative manipulations.

	Baseline	Positive	Negative	Chi <sup>2</sup>
Brow movement	6.2%	40.6%	68.8%	$p < .001$
Eye widening	6.2%	21.8%	28.1%	$p = .069$
Mouth opening	6.2%	37.5%	62.5%	$p < .001$

We also found an effect of age on the overall appearance of the facial display of surprise,  $F(1,96) = 5.255$ ,  $p < .05$ . Older children ( $M = 1.10$ ,  $SD = .109$ ) use more features to express their surprise than younger children ( $M = .75$ ,  $SD = .109$ ). We found no effect of social setting on the overall appearance of the facial display of surprise,  $F(1,96) = .017$ , *ns*.

## Discussion

We found a significant difference in the frequency of use of the facial features between the three conditions. Participants who were surprised by a negative cause showed most facial features, compared with the other conditions. A closer look at the features reveals that mainly opening of the mouth and brow movement are used more with negatively caused surprise than with positively caused surprise. It seems that there is no difference in the features that are used in our created sorts of surprise, although there is difference in their relative frequency. We also found that older children used more features for showing surprise. An explanation for this could be that 11-year-old children show more surprise because they are more aware of the social situation (Saarni, 1979). We did not find a significant difference between the social settings. However, it is conceivable that the social

setting is important for expressing surprise, since *perception* of surprise is important for the function of self-presentation in a social setting. According to Feldman Barrett, Mesquita and Gendron (2011), facial features might carry affective information, but emotional meaning is rather contingent on context. Therefore, we decided to conduct a perception test using the same video clips as in study 2.

### Study 3: Perception of surprise

The third study consisted of a rating experiment to test how video clips from study 2 are perceived in terms of different sorts of surprise as a function of social setting and age group.

#### Method

**Participants:** Thirty students from Tilburg University (16 female) participated as judges in the perception experiment (age range: 18- 48 years old,  $M = 22.07$ ,  $SD = 5.42$ ).

**Stimuli:** The same 96 video clips that were labeled for the presence or absence of surprise features in study 2 were used in the perception test as stimuli.

**Procedure:** All 96 video clips were shown to the participants in one of two random orders. First, the identification number of the stimulus was presented (1 through 96), followed by the actual stimulus. During an inter-stimulus interval of three seconds the screen turned black, and participants were asked to rate the child's level of surprise, on a five-point Likert scale. To ensure that participants were familiar with the perception task, the experiment was preceded by a short training phase.

#### Results

We conducted a  $3 \times 2 \times 2$  ANOVA with sort of surprise, age and social setting as within-subject factors and the perceived level of surprise as dependent variable.

We found a main effect of social setting on the perceived level of surprise,  $F(1,29) = 72.023$ ,  $p < .001$ . Competing children ( $M = 3.93$ ,  $SD = 0.48$ ) were rated as more surprised than collaborating children ( $M = 3.48$ ,  $SD = 0.51$ ). Age did not have an effect on the perceived surprise level,  $F(1,29) = 3.494$ ,  $ns$ .

We also found an effect of sort of surprise on the perceived level of surprise,  $F(2,28) = 132.9$ ,  $p < .001$ . A Bonferroni post hoc test showed that children in the baseline condition ( $M = 3.01$ ,  $SD = 0.54$ ) were perceived to be less surprised than the children in both surprise manipulation conditions. Children with a negatively caused surprise ( $M = 4.54$ ,  $SD = 0.64$ ) were perceived to be more surprised than children with a positively caused surprise ( $M = 3.57$ ,  $SD = 0.64$ ).

We found two interaction effects involving sort of surprise. First, there is an interaction between age and sort of surprise,  $F(2,28) = 43.476$ ,  $p < .001$ . After running split analyses, were we looked at the perception of surprise for both age groups separately, we did not find a significant

difference in perceived surprise for the 8-year-old children between the baseline condition and the positive manipulation, but only a difference between these two conditions and the negative manipulation. For the perceived surprise of 11-year-old children there was a significant difference between all three conditions, see figure 4.

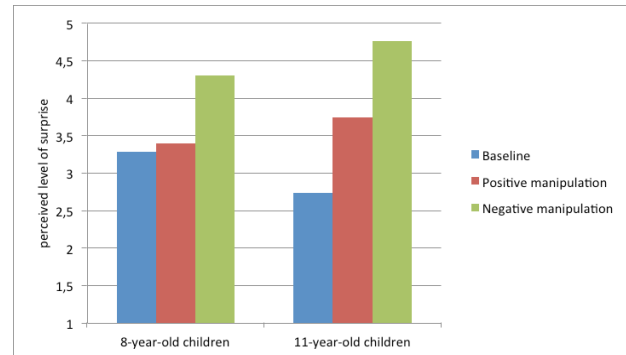


Figure 4. The interaction effect between age and type of surprise, on perceived level of surprise.

Second, we found an interaction effect between social setting and type of surprise on the perceived level of surprise,  $F(2,28) = 26.190$ ,  $p < .001$ . Post hoc analyses (Bonferroni method) reveal that in collaboration, there is a larger difference between the perception of positive and negative surprise than in competition. In the competitive setting, children are, overall, perceived to be more surprised than in the collaborative setting, see figure 5.

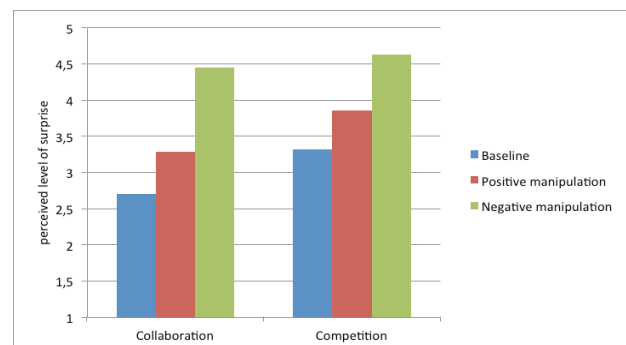


Figure 5. The interaction effect between social setting and type of surprise, on perceived level of surprise.

#### Discussion

The perception test showed that surprise was perceived differently between the three types of surprise. We found that positive surprise was perceived as less prominent than negative surprise. Furthermore, we found that 11-year-old children's expressions of surprise were perceived to be more distinct among the different conditions. This could mean that children express their surprise more accurately as they grow older. We also found that competing children are perceived to show more surprise than collaborating children, at least for the baseline condition and positive caused surprise.

## General discussion and conclusion

Earlier research has had difficulties in extracting a full facial display of surprise with participants (Reisenzein et al., 2006). The aim of the studies presented in the current paper was to examine two possible reasons for this; firstly the false assumption that there is only one kind of surprise and secondly the lack of taking contextual factors into account.

First, we assumed that surprise could be expressed differently, depending on its cause. We manipulated various questions in the knowledge quiz to extract reactions of surprise, either with a positive or negative cause. Analyzing the difference scores of participants' certainty judgments showed that the manipulations worked as intended (study 1). After annotating the facial features (study 2) and conducting a perception test (study 3), we can conclude that we found significant differences in expressing and perceiving surprise. Participants who were surprised by a negative cause showed most features of the full facial display of surprise, compared to the other conditions. Although we did find that manipulation affected the frequency in use of features, the different kinds of surprise were not related to certain specific features. It seems that the cause of the emotion leads to different *degrees* in surprise expressions, instead of a *different* surprise expression. Possibly, negative events cause more surprise than positive events. We must note, however, that the created social settings may have interfered with the concept of positive or negative surprise. For example, in competition, an incorrect answer on an easy question probably does not evoke an absolutely *negative* feeling of surprise with the opponent, as an error of the opponent is actually good for the other player. However, we still think that although the naming of the two kinds of surprise (positive versus negative) might be somewhat inaccurate, their difference in cause remains. In our research, we elicited feelings of surprise with two distinctive causes, and results show us that the cause of the surprise affects the expression of the emotion. However, future research should consider these causes in respect of the social settings participants are in.

Second, we wanted to study if the expression of surprise could depend on this social setting. Therefore, our participants played the knowledge quiz in either a collaborative setting, or a competitive setting. From study 3 we can conclude that children in competition were perceived to be more surprised than collaborating children. This may be due to the fact that children in competition are more aware of their social environment, because self-presentation is more important in this setting. Expressing surprise could be beneficial for the players' progression in the game. We can conclude that the social setting appears to be important for the expression of surprise.

Since 11-year-old children appear to adjust to social situations far more than 8-year-old children (Saarni, 1979), we also expected to find age to affect expressions of surprise. This was confirmed by our data analyses: 11-year-old children were more expressive and were perceived as more surprised than 8-year-old children. It seems that older

children can make a more accurate distinction between expressions of surprise with different causes than younger children.

We can conclude that the expression of surprise is affected by several factors, like age, social setting and the cause of the surprise. Therefore, we think future studies should consider these factors when studying surprise. Our data suggest that the expression of surprise is more than a mere reflex to an unexpected stimulus, and that it can be moderated by contextual factors.

## Acknowledgments

Many thanks to the children and teachers from third and sixth grade at primary schools Florence Nightingale and De Tjalk in Zoetermeer for their corporation. Thanks to Martijn Goudbeek and Marieke Hoetjes for their useful comments, and to Hans Westerbeek and Lisanne van Weelden for their assistance.

## References

- Ekman, P. (1997). Expression or communication about emotion. In N. L. Segal, G. E. Weisfeld, & C. C. Weisfeld (Eds.), *Uniting psychology and biology: Integrative perspectives on human development* (Vol. 48, pp. 384–392). Washington, DC: American Psychological Association.
- Feldman Barrett, L., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, 20, 286–290.
- Hadwin, J., & Perner, J. (1991). Pleased and surprised: Children's cognitive theory of emotion. *British Journal of Developmental Psychology*, 9(2), 215–234.
- Lockl, K., & Schneider, W. (2002). Developmental trends in children's feeling-of-knowing judgments. *International Journal of Behavioural Development*, 26, 327–333.
- Matsumoto, D., Hee Yoo, S., Hiramaya, S., & Petrova, G. (2005). Development and validation of a measure of display rule knowledge: The display rule assessment inventory. *Emotion*, 5(1), 23–40.
- Piaget, J. (1950). *The Psychology of Intelligence*. San Diego: Harcourt Brace Jovanovich.
- Reisenzein, R., Bördgen, S., Holtbernd, T., & Matz, D. (2006). Evidence for strong dissociation between emotion and facial displays: The case of surprise. *Journal of Personality and Social Psychology*, 91(2), 295–315.
- Saarni, C. (1979). Children's understanding of display rules for expressive behaviour. *Developmental Psychology*, 15, 424–429.
- Scherer, K. R., Zentner, M. R., & Stern, D. (2004). Beyond surprise: The puzzle of infants' expressive reactions to expectancy violation. *Emotion*, 4, 389–402.
- Shepperd, J. A., & McNulty, J. K. (2002). The affective consequences of expected and unexpected outcomes. *Psychological Science*, 13, 85–88.

# Estimating Semantic Transparency of Constituents of English Compounds and Two-Character Chinese Words using Latent Semantic Analysis

Hsueh-Cheng Wang (hchengwang@gmail.com)

Li-Chuan Hsu (lchsu@mail.cmu.edu.tw)

Yi-Min Tien (tien@mercury.csmu.edu.tw)

Marc Pomplun (marc@cs.umb.edu)

Department of Computer Science, University of Massachusetts at Boston,  
100 Morrissey Boulevard, Boston, MA 02125 USA

## Abstract

The constituents of English compounds (e.g., butter and fly for butterfly) and two-character Chinese words may differ in meaning from the whole word. Furthermore, the meanings of the words containing the same constituent (e.g., butter in “butterfingers”, or “buttermilk”) may or may not be consistent. Estimating semantic transparency of a constituent is usually difficult and subjective because of these uncertainties and ambiguities. It is rather unexplored why a constituent is considered transparent/opaque by raters, and how its polysemy correlates to its transparency. We propose a computational method for predicting semantic transparency based on Latent Semantic Analysis. We computed the primary meaning of a constituent by a clustering analysis and compared it to the whole-word meaning. The proposed method successfully predicted participants’ transparency ratings, and may explain the cognitive processes in raters when classifying semantic transparency of English compounds and two-character Chinese words.

**Keywords:** compound words; semantic transparency; latent semantic analysis; Chinese; clustering.

## Introduction

A compound word is a word composed of at least two free lexemes that refer to a new concept. Compound words with two constituents are defined as semantically transparent (transparent-transparent, referred to as TT, see Frisson, Niswander-Klement, & Pollatsek, 2008) when the whole word meaning can be grasped through its individual constituents, such as *cookbook*. Compound words are regarded as semantically opaque (opaque-opaque, OO), when their meaning cannot be fully derived from its constituents, e.g., *cocktail*. Some compound words are considered partially opaque (opaque-transparent, OT, or transparent-opaque, TO) when the primary meaning of one of the constituents is related to the meaning of the compound, such as *butterfly* or *staircase*, respectively.

Typically, transparency ratings are the most common method to obtain transparency information. Transparency rating experiments have used target words that differed substantially in their estimated transparency by researchers or a group of participants. For instance, Pollatsek and Hyönä (2005) selected 80 compound words, 40 of which they assumed to be semantically transparent, and the other 40 to be opaque. They asked eight participants to rate these words regarding their transparency using a 7-point scale (1 for

totally transparent and 7 for totally opaque), and the ratings were clearly lower for the supposedly transparent sets than for the supposedly opaque ones. Similarly, Frisson et al. (2008) asked 40 participants to rate transparency in terms of appropriate categories (e.g., opaque-transparent (OT), transparent-opaque (TO), opaque-opaque (OO), and transparent-transparent (TT)), and there was good agreement between the participants’ choices and the predefined classification by Frisson et al. (2008). The proportion of participants’ choices agreeing with the predefined classification was 65% for OO, 71% for OT, 65% for TO, and 86% for TT. Moreover, the proportion of participants classifying at least one of the constituents as opaque for the predefined opaque words was very high: 95% for OO, 93% for OT and 95% for TO. Inhoff et al. (2008) selected “headed” and “tailed” compound words, i.e., compound words whose meaning was primarily defined by their first or second constituents, respectively. They had 13 participants rate 390 compound words using an 11-point scale ranging from 0 to 10, where 0 indicated that the meaning of the compound was solely associated with the meaning of the first constituent, while 10 denoted that the meaning of the compound was solely associated with the one of the second constituent. Compounds with mean ratings below 4 (mean: 3.34) or above 6 (mean: 7.18) were considered to be headed or tailed, respectively. It is important to notice that the definition of headed and tailed compound words might not equal the TO and OT conditions discussed above. For example, the second constituent of a headed compound may be opaque or transparent, as long as its meaning is less closely related to the compound than the meaning of the first constituent is.

Two-character Chinese words, similar to English compound words, differ in how the meanings of the first and second characters relate to the meaning of the word. Some two-character Chinese words are semantically transparent, i.e., both characters are transparently related to the meaning of the whole word. Other words are fully opaque, i.e., the meaning of neither constituent is related to the meaning of the compound, or partially opaque. Table 1 lists some examples of transparent, opaque, and partially opaque Chinese words.

According to the estimation of Zhou and Marslen-Wilson (1995), 74% of Chinese words are made up of two characters, although some words consist of only one

character and some consist of three or more characters. A Chinese character is a writing unit which has a single syllable and meaning(s). It is approximately equal to a morpheme in most cases. However, unlike English and other alphabetic writing systems, Chinese words are written without spaces in a sequence of characters. The concept of a word is not as clearly defined in Chinese as it is in English, which means that Chinese readers might somewhat disagree where word boundaries are located (see Rayner, Li, & Pollatsek, 2007, for a review). According to the segmentation standard by Huang, Chen, Chen, and Chang (1997) used by the Academia Sinica Balanced Corpus (ASBC; Academia Sinica, 1998), not all characters constitute one-character words. Furthermore, a Chinese character might be shared by many words, but the meaning of the character and those words may not be consistent.

Table 1. Examples of transparent, opaque, and partially opaque Chinese words. The meaning of the whole word and the primary meanings of 1st and 2nd characters are shown in parentheses.

	Whole Word	1st Character	2nd Character
TT	球場 (ball court)	球 (ball)	場 (court)
OO	壽司 (sushi)	壽 (age)	司 (in charge of)
TO	智商 (I.Q.)	智 (Intelligent)	商 (commerce)
OT	開水 (boiled water)	開 (open)	水 (water)

Early studies of the morphological processing of Chinese polymorphemic words asked how compound words are represented in the mental lexicon and how their lexical processing in visual or auditory word recognition is performed. Recent studies investigated semantic composition (Mok, 2009) and frequency effects (see Zhou, Ye, Cheung, & Chen, 2009, for a review). In Mok (2009), the experimenter pre-defined semantic transparency on a 6-point scale, where 1 is opaque and 6 is transparent. A constituent was classified transparent if the rating was equal to or greater than 4, and opaque otherwise. Five participants were then provided the 6-point scale again for each constituent. Constituents with an average rating greater than 3.5 were classified as transparent, and the others were categorized as opaque.

There are also a few unpublished studies attempting to estimate semantic transparency of Chinese two-character words by researchers or human raters such as Tsai (1994), Lee, C. Y. (1995), and Lee, P. J. (2007). For example, a five-point scale ranging from 1 to 5 was used in the study by Lee (2007), and words were considered transparent when the average score was below 2 while opaque when the average score was greater than 4. Tsai (1994) categorized opaque words into OT and TO conditions, but Lee (1995) and Lee (2007) generalized OT, TO, or OO conditions as opaque words (referred to as idiomatic words).

Unfortunately, estimates of semantic transparency are often subjective and vary across raters, and sometimes even

the meaning of transparent compounds cannot be unambiguously determined from the meanings of their constituents (see Frisson et al., 2008). Inhoff et al. (2008) pointed out that a semantic relationship often exists between an opaque lexeme and its compound, for example, even though “jailbird” typically refers to a person rather than an animal, it can convey useful semantic information, such as being caged or wishing to fly free. This topic was also studied in the literature on conceptual combination (e.g., Wisniewski, 1996; Costello & Keane, 2000), which indicated that one part of a compound has an *exocentric interpretation* such as *shape* (“seahorse” is a fish whose head is the *shape* of a horse’s head) or the head concept (in the “seahorse” case the diagnostic predicate being *shape*). Participants might be able to interpret constituents being defined as opaque to a meaning related to the compound according to some kinds of relation (e.g., *shape*) or the polysemy of the constituent and compound. This subjectivity and variability also occurs in characters of Chinese two-character words. Therefore, a computational model may be a way to average across subjective differences of estimating semantic transparency.

## Predicting Transparency using Latent Semantic Analysis

This study proposes a computational method for estimating transparency using Latent Semantic Analysis (LSA). LSA is a method to represent the meaning of words by statistical computations applied to a text corpus (Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007). Typically, terms are words, and a term-to-document co-occurrence matrix is established from a corpus. Then a mathematical method, singular value decomposition (SVD), is used to reduce the dimensions of the original matrix (see Martin & Berry, 2007). The meaning of each term is represented as a vector in *semantic space*. One can compute the semantic similarity values for any two terms in a given language using the LSA cosine value, which ranges between -1 and 1, but rarely goes below 0. Randomly chosen pairs of words have a mean of 0.03 and a standard deviation of approximately 0.08 (see Landauer et al., 2007). An LSA web site is freely available (<http://lsa.colorado.edu/>, accessed September, 2010; see Dennis, 2007).

LSA has been successful at simulating judgments of semantic similarity, word categorization, discourse comprehension, essay quality (see Landauer & Dumais, 1997; Landauer et al., 2007; Jones & Mewhort, 2007, for a review). LSA has also been used to investigate morphological decomposition; for example, Rastle, Davis, Marslen-Wilson and Tyler (2000) investigated morphologically complex words with semantically transparent embedded stems (e.g., “depart” vs. “departure”) and opaque embedded stems (e.g., “apart” vs. “apartment”). Furthermore, Diependaele, Dunabeitia, Morris and Keuleers

(2011) used LSA to estimate transparency between full words and constituent-embedded stems, which yields “viewer” vs. “view” as being highly transparent and “corner” vs. “corn” as highly opaque.

One possible method (abbreviated as C2W) of measuring semantic transparency is, similar to Diependaele et al. (2011), to compute the LSA cosine values between the compound word and each of its constituents. For example, the LSA cosine value between “staircase” and “stair” is 0.57 while the one between “staircase” and “case” is 0.07. Since the constituent “stair” and the compound word “staircase” result in a clearly higher cosine value, “stair” is considered semantically transparent, while “case” is considered opaque. However, this computation for English words may or may not reflect how a Chinese rater classifies a constituent as transparent/opaque for two-character Chinese words.

One possible solution is to access the primary meaning of a constituent. The first step of our proposed idea is to find words containing a constituent that a rater possibly activates. Since a constituent may have several meanings, the *primary meaning* of the constituent is computed by a hierarchical clustering algorithm. Since LSA cosine values rarely go below 0 in high-dimensional spaces, we use one minus the absolute value of the LSA cosine as distance function and a given threshold. The selection of this threshold is discussed below in the Reanalysis of Previous Data and General Discussion sections. Since word frequency is important for word recognition and reading (see Rayner et al., 2007), the cluster with the highest sum of word frequency is considered the primary meaning. For example, the transparency of constituent “butter” in “butterfly” is determined as follows. Using the text corpus “general reading up to 1<sup>st</sup> year college,” the LSA cosine values among “butter”, “butterfly”, “buttercup”, “butterfingers”, “buttermilk”, “butterscotch”, “butterfat”, and “butterwick” are shown in Table 2. Based on the LSA cosine values, semantic relationships can be visualized by multi-dimensional scaling (MDS) as presented in Figure 1. The results of the cluster analysis are demonstrated in Figure 2. The group of “butter”, “buttercup”, “buttermilk”, “butterscotch”, “butterfat”, and “butterwick” are considered primary meaning for their highest sum of frequency. According to a threshold 0.9, “butterfly” and “butterfingers” are clustered individually. We applied “document to term” comparison to compute the LSA cosine value between the primary meaning cluster (i. e., a string of “butter buttercup buttermilk butterscotch butterfat butterwick”) and “butterfly”, and 0.04 is obtained. This approach is abbreviated as M2W.

The M2W approach takes the polysemy of a constituent into account and works even when a constituent is not a stand-alone word. M2W is especially useful for the Chinese language in which many characters do not exist as one-character words in the corpus (described below).

Since the LSA-based method may be able to estimate transparency of English compounds, it could possibly be applied to Chinese two-character words in a similar manner.

Table 2. The LSA cosine values among “butter”, “butterfly”, “buttercup”, “butterfingers”, “buttermilk”, “butterscotch”, “butterfat”, and “butterwick”.

butter	-fly	-cup	-fingers	-milk	-scotch	-fat	-wick
1							
0.04	1						
0.09	0.09	1					
0	-0.1	-0.1	1				
0.44	-0	0.12	0.01	1			
0.45	0.05	-0	0.02	0.35	1		
0.12	-0	0.04	0	0.11	0.16	1	
-0	0.01	0.12	-0	0.09	0.03	0.04	1

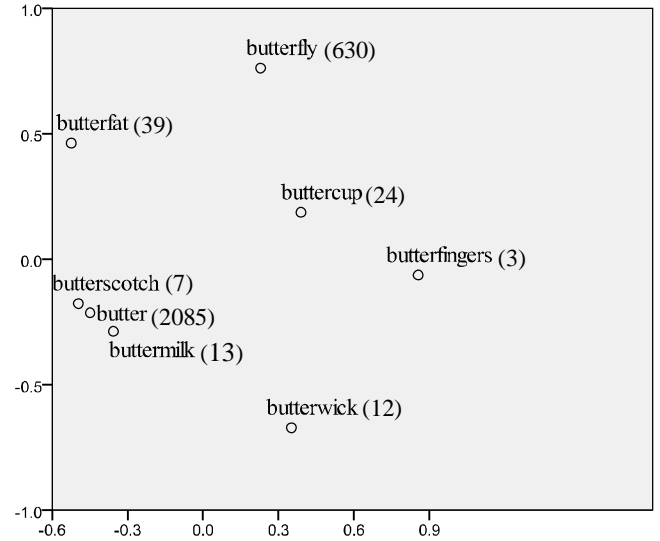


Figure 1: The MDS result for an example of semantic relationships for “butter” and words containing “butter”. The frequency for each word in British National Corpus (BNC) is shown in parentheses. The x and y axis represent dimensions 1 and 2, respectively, of the abstract, two-dimensional Euclidean output space of the MDS algorithm.

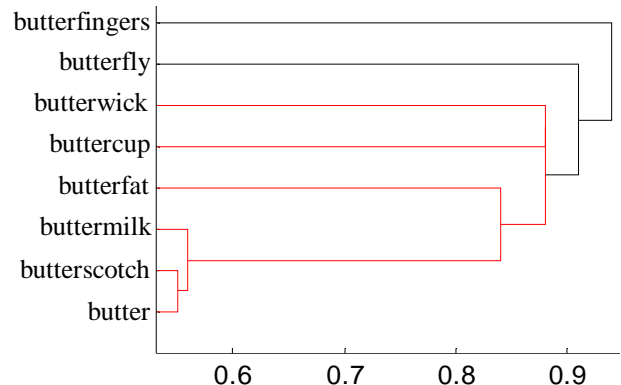




Figure 2. The results of hierarchical clustering of the example word “butter”.

Following the principle of creating semantic spaces (Quesada, 2007), our previous studies (Wang et al., 2010; Chen, Wang, & Ko, 2009) built an LSA semantic space of Chinese (abbreviated as SP-C) from ASBC which contains approximately 5 million words (or 7.6 million characters). Texts in ASBC were collected from different topic areas and classified using five criteria: genre, style, mode, topic, and source, in order to make ASBC a representative sample of modern Chinese language. Word segmentation was performed manually according to the standard by Huang et al. (1997). For representatives of words in the corpus, words that occurred less than 3 times per 5 million were excluded. A 49021 x 40463 term-to-document co-occurrence matrix was then established. SP-C has been shown to successfully estimate word predictability (see Wang et al., 2010) and word association in Chinese language (see Chen, Wang, & Ko, 2009).

The term-to-document co-occurrence matrix of SP-C was established using the unit of words, which may be one or more Chinese characters. The C2W approach requires Chinese two-character words to have their constituent characters appear as single-character words more than 3 times in the corpus. Within the 49021 words available in SP-C, 31,637 are two-character words. For 3,921 out of these 31,637 two-character words, either the first or second characters are unavailable due to the frequency restriction. Nevertheless, the M2W approach can still compute the primary meaning of characters despite this characteristic of Chinese. It is even possible that a Chinese reader does not have a single-character representation in his or her mental lexicon for non-stand-alone characters. The polysemy of a character might be involved and the primary meaning might be obtained during lexical access.

Table 2 shows examples of the polysemy of character “馬” (horse). The whole-word meanings of words such as “馬背” (horse back) and “馬鞍” (saddle) are close to “馬” (horse), while the ones of words, e.g., “馬虎” (careless) and “馬桶” (stool) are not. The character “馬” in the word “馬來” (Malaysian, pronunciation: ma-lai) and “馬國” (Malaysia) refers to the abbreviation of Malaysia/Malaysian because of its pronunciation. Figure 3 demonstrates the clustering of character “馬” - the meaning of “馬” is “horse” in words 1 to 4, and is related to “Malaysia” in words 8 and 9. Since the sum of frequency in ASBC for the group of words 1 to 4 is the highest, the group of words 1 to 4 is considered the primary meaning of character “馬”.

It is necessary to verify the proposed computational method by comparing its results with human transparency ratings. We evaluated how LSA estimates transparency of English compounds using the materials of Frisson et al. (2008). The evaluation for two-character Chinese words was conducted by re-analyzing the materials of Tsai (1994) and Lee (2007).

Table 2. A list of character “馬” as one-character word and the two-character words beginning with character “馬”. C2 Meaning is the primary meaning of the second character. WFreq is whole-word frequency in ASBC.

	Word	Whole-Word Meaning	C2 Meaning	WFreq
1	馬	horse		342
2	馬背	horse back	back	14
3	馬鞍	saddle	saddle	4
4	馬車	carriage	car	37
5	馬虎	careless	tiger	13
6	馬桶	stool	tub	23
7	馬腳	a clue of	foot	4
8	馬來	Malaysian	come	11
9	馬國	Malaysia	country	12

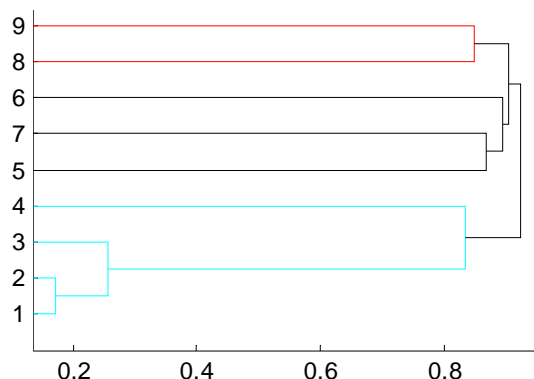


Figure 3. The results of hierarchical clustering of example “馬” with numbers referring to Table 2.

## Reanalysis of Previous Data

Ten opaque-opaque, 14 opaque-transparent, and 10 transparent-opaque compounds defined in Frisson et al. (2008) were estimated by our classifiers using C2W and M2W. A receiver operating characteristic (ROC) analysis was performed and the area under the curve (AUC) was used as measurement. Figure 4 illustrates the ROC curves for C2W and M2W (threshold = 0.1, 0.8, and 1), and the AUCs are 0.82, 0.74, 0.82, and 0.75, respectively. A threshold too low may generate too many groups, while a threshold too high only produces one group and therefore causes more false alarm cases. We found that when a compound is high-frequent and its constituent is opaque and low-frequent, the primary meaning of the constituent might be taken over by the compound and therefore the constituent is incorrectly considered transparent. We suggest that C2W could be used when the constituent is low-frequent, and an item-level human judgment should be performed for further analysis.



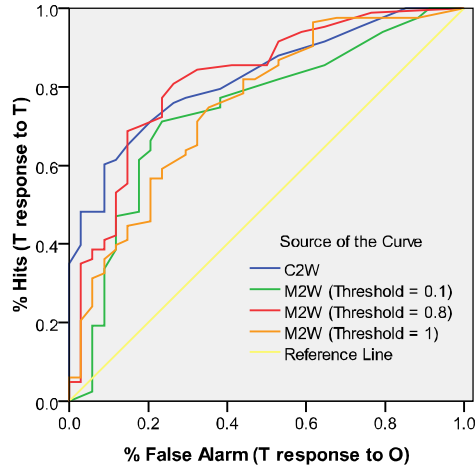


Figure 4: ROC curves for C2W and M2W (threshold = 0.1, 0.8, and 1) for English readers.

For two-character Chinese words, we pre-defined 160 characters selected from the materials of Tsai (1994), Lee, C. Y. (1995), and Lee, P. J. (2007) as either transparent (T) or opaque (O), then those characters were rated by eleven students who completed a college degree in Taiwan participated. All participants were native speakers of Chinese (traditional script). Participants were presented with the two-character word, and asked to respond either “T” or “O” for each constituent. The measure of human rating of each constituent was calculated as the probability with which participants responded “T” to the constituent, e.g., 0.91 for 10 out 11 participants responding “T.” The means and standard deviations of human rating (Human), C2W, and M2W are shown in Table 3, where there are 99 transparent and 50 opaque constituents available for C2W. Figure 5 illustrates the results of the ROC analysis, and the AUCs of human rating, C2W, and M2W are 0.99, 0.76, and 0.85, respectively. The Spearman rank correlations (a non-parametric test) between Human and C2W and between Human and M2W are 0.48 and 0.53, respectively. These results suggest that M2W not only overcomes the constraint that C2W is unable to compute transparency when constituents are unavailable in SP-C, but also outperforms C2W in ROC and correlation analyses. As mentioned above, the concept of a word is not as clearly defined in Chinese as in English, and Chinese readers might learn the polysemy of characters implicitly from polymorphemic words. We suggest that M2W may be a better approach than C2W for predicting transparency of constituents of two-character Chinese words.

Table 3. The means and standard deviations (in parentheses) of human rating (Human), C2W, and M2W.

	Human	C2W	M2W
T	0.79 (0.19)	0.22 (0.17)	0.35 (0.27)
O	0.13 (0.14)	0.08 (0.06)	0.07 (0.10)

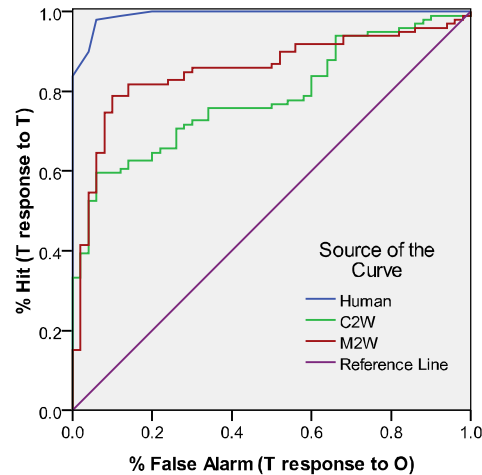


Figure 5: ROC curves for human rating (Human), C2W and M2W for Chinese readers.

## General Discussion

The most important outcome of the current study is its proposed computational method of using LSA to estimate semantic transparency, which may reflect the polysemy of constituents and how raters access meanings. Corroborating evidence from two different languages was presented by testing the method with English compounds used in prior compound word study (including Frisson et al., 2008) and two-character Chinese words in the transparency judgment in this study.

The results could be adapted to further Chinese reading research using eye movements. For example, it is still being debated how Chinese words are accessed by readers. Yan, Tian, Bai, and Rayner (2006) investigated the effect of word and character frequency on word processing, and they suggested that when a two-character word is frequent and has been seen quite often in print, it is accessed as a single entity in the mental lexicon of Chinese readers, whereas when it is infrequent, the word needs to be accessed via its characters (and hence an effect of character frequency emerges). However, some studies have argued for the priority of characters over words (e.g., Chen, Song, Lau, Wong, & Tang, 2003). Therefore, it is still unclear how opaque and transparent words are processed during natural reading. It would be valuable to address these issues using semantic transparency and eye-movement analysis.

The current limitations of the proposed method in Chinese might be the relatively small corpus size. Cai and Brysbaert (2010) published SUBTLEX-CH based on a larger corpus (47 million characters) of film and television subtitles, and they suggested that SUBTLEX-CH is a good estimate of daily language exposure and captures much of the variance in word processing efficiency. It is possible that a Chinese LSA semantic space could be established based on this larger corpus as long as the corpus provides enough information in terms of “documents”, i.e., a set of words that relate to the same topic in a document. It is important to

notice that the size of a corpus is not its only criterion of being representative, but the selection of texts covering different varieties in a corpus should also be taken into account. Furthermore, there are traditional and simplified scripts of Chinese, and it is important to test whether the semantic space built by traditional Chinese is compatible with simplified Chinese.

In addition to the selection of the threshold, since it is related to the distance function of the clustering algorithm and the LSA values, we suggest that an optimization test should be performed for each semantic space. We imply that a threshold might be involved in the transparency judgments by human raters and that each participant might have a different threshold for the “cut-off” of opacity. It should be clear that the use of the proposed computational method is not intended to replace the standard measures that are based on human raters, but that it offers a different perspective and an opportunity to examine the lexical processing for estimating semantic transparency.

## Acknowledgments

Portions of the data were presented at the Asia-Pacific Conference on Vision (APCV) 2010. Thanks to Keith Rayner, Jinmian Yang, and Marc Brysbaert for helpful comments on this study.

## References

- Academia Sinica. (1998). *Academia Sinica balanced corpus* (Version 3) [Electronic database]. Taipei, Taiwan.
- Cai, Q. & Brysbaert, M. (2010). SUBTLEX-CH: Chinese Word Frequencies Based on Film Subtitles. *PLoS ONE* 5(6): e10729. doi: 10.1371/journal.pone.0010729.
- Chen, H.-C., Song, H., Lau, W. Y., Wong, K. F. E., & Tang, S. L. (2003). Developmental characteristics of eye movements during reading. In C. McBride-Chang & H. C. Chen (Eds.), *Reading development in Chinese children* (pp. 157–169). Westport, CT: Praeger.
- Dennis, S. (2007). How to use the LSA website. In T. Landauer, D. McNamara, S. Dennis & W. Kintsch Eds. *Handbook of Latent Semantic Analysis*. Erlbaum, 57-70.
- Diependaele, K., Dunabeitia, J. A., Morris, J., & Keuleers, E. (2011). Fast morphological effects in first and second language word recognition. *Journal of Memory and Language*, 64(4), 344-358.
- Frisson S., Niswander-Klement E., & Pollatsek A. (2008). The role of semantic transparency in the processing of English compound words. *British Journal of Psychology*, 99, 87-107.
- Inhoff, A. W., Starr, M. S., Solomon, M. P., & Lars, P. (2008). Eye movements during the reading of compound words and the influence of lexeme meaning. *Memory & Cognition*, 36(3), 675-687.
- Jones, M. N. & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite orthographic lexicon. *Psychological Review*, 114, 1-37.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240.
- Landauer, T. K., McNamara, D. S., Dennis S., & Kintsch W. (2007). *Handbook of Latent Semantic Analysis*, Lawrence Erlbaum Associates.
- Lee, C. Y. (1995). The representation of semantically transparent and opaque words in mental lexicon. Unpublished master's thesis, National Chung Cheng University, Chia-Yi, Taiwan. (in Chinese)
- Lee, P. J. (2007). The representation of semantically transparent and opaque words in mental lexicon: evidence from eye movements. Unpublished master's thesis, National Chung Cheng University, Taipei, Taiwan. (in Chinese)
- Martin, D. I. & Berry, M. W. (2007). Mathematical foundations behind latent semantic analysis. In T. Landauer, D. McNamara, S. Dennis & W. Kintsch Eds. *Handbook of Latent Semantic Analysis*. Erlbaum, 35-55.
- Mok, L. W. (2009). Word-superiority effect as a function of semantic transparency of Chinese bimorphemic compound words. *Language and Cognitive Processing*, 24 (7/8), 1039-1081
- Pollatsek, A. & Hyönä, J. (2005). The role of semantic transparency in the processing of Finnish compound words. *Language and Cognitive Processing*, 20 (1/2), 261-290.
- Quesada, J. (2007). Creating Your Own LSA Spaces. In T. Landauer, D. McNamara, S. Dennis & W. Kintsch Eds. *Handbook of Latent Semantic Analysis*. Erlbaum, 71-88.
- Rastle, K., Davis, M. H., Marslen-Wilson, W. D., & Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition: A time-course study. *Language and Cognitive Processes*, 15(4/5), 507-537.
- Rayner, K., Li, X., & Pollatsek, A. (2007). Extending the E-Z Reader model of eye movement control to Chinese readers. *Cognitive Science*, 31, 1021–1033.
- Tsai, C.-H. (1994). Effects of semantic transparency on the recognition of Chinese two-character words: Evidence for a dual-process model. Unpublished master's thesis, National Chung Cheng University, Chia-Yi, Taiwan. (in Chinese)
- Yan, G, Tian, H., Bai, X., & Rayner, K. (2006). The effect of word and character frequency on the eye movements of Chinese readers. *British Journal of Psychology*, 97, 259-268.
- Wang, H. C., Pomplun, M., Ko, H. W., Chen M. L., & Rayner, K. (2010). Estimating the effect of word predictability on eye movements in Chinese reading using latent semantic analysis and transitional probability. *Quarterly Journal of Experimental Psychology*, 63, 1374-1386.
- Zhou, X., Ye, Z., Cheung, H., Chen, H.-C. (2009). Processing the Chinese language. *Language and Cognitive Processing*, 24 (7/8), 929-946.

# Visual Attention is Attracted by Text Features Even in Scenes without Text

Hsueh-Cheng Wang (hchengwang@gmail.com)

Shijian Lu (slu@i2r.a-star.edu.sg)

Joo-Hwee Lim (jooHwee@i2r.a-star.edu.sg)

Marc Pomplun (marc@cs.umb.edu)

Department of Computer Science, University of Massachusetts at Boston,

100 Morrissey Boulevard, Boston, MA 02125 USA

Institute for Infocomm Research, A\*STAR, Singapore

1 Fusionopolis Way, Singapore 138632

## Abstract

Previous studies have found that viewers' attention is disproportionately attracted by texts, and one possible reason is that viewers have developed a "text detector" in their visual system to bias their attention toward text features. To verify this hypothesis, we add a text detector module to a visual attention model and test if the inclusion increases the model's ability to predict eye fixation positions, particularly in scenes without any text. A model including text detector, saliency, and center bias is found to predict viewers' eye fixations better than the same model without text detector, even in text-absent images. Furthermore, adding the text detector – which was designed for English texts – improves the prediction of both English- and Chinese-speaking viewers' attention but with a stronger effect for English-speaking viewers. These results support the conclusion that, due to the viewers' everyday reading training, their attention in natural scenes is biased toward text features.

**Keywords:** real-world scenes; text detector; eye movements; visual attention.

## Introduction

When inspecting real-world scenes, human observers continually shift their gaze to retrieve information. Viewers' attention has been found to be biased toward visually salient locations, e.g., high-contrast areas, during scene viewing or search (Itti & Koch, 2001) or toward the center of the screen when viewing scenes on computer monitors (Tatler, 2007). Since it is also known that viewers pay a disproportionate amount of attention to faces (Cerf, Frady, & Koch, 2009), Judd, Ehinger, Durand, and Torralba (2009) equipped their model of visual saliency with a face detector (Viola & Jones, 2004) and a person detector (Felzenszwalb, McAllester, & Ramanan, 2008). In those images that contained depictions of people, their model with all features combined outperformed models trained on typical saliency features such as color, orientation, intensity, and contrast. Cerf et al. (2009) refined the "standard" saliency model by adding a channel of manually-defined regions of faces, texts, and cellphones, and demonstrated that the enhancement of the model significantly improved its ability to predict eye fixations in natural images.

Besides depictions of people, texts in natural scenes are usually important pieces of information, which could be shown on depictions of signs, banners, advertisement billboards, license plates, and other objects. Human text

detection in natural scenes is critically important for people to survive in everyday modern life, for example, by drawing attention to traffic signs or displays showing directions to a hospital or grocery store. Our previous studies (Wang & Pomplun, 2011; under revision) suggested that attention seems disproportionately attracted by texts but that the specific visual features of texts, e.g., edge density, rather than typically salient features such as color, orientation, intensity, or contrast, are the main attractors of attention. This finding was in line with the results in Baddeley and Tatler (2006) that high spatial frequency edges, not contrasts, predict where we fixate.

Automatic text detection has been a hot topic in the fields of computer vision and pattern recognition for its practical applications. The special features of texts, e.g., the small variation of the stroke width (see Epshtein, Ofek, & Wexler, 2010; Jung, Liu, & Kim, 2009) or edge density (Lu, submitted) have been used to develop text detectors. Although many text detection techniques, i.e., texture-based, region-based, and stroke-based methods, have been reported, many non-text objects, such as windows, fences, or brick walls, easily cause false alarms (see Lu, submitted; Ye, Jiao, Huang, & Yu, 2007, for a review). Furthermore, many established text detectors are restricted under commercial patents. Therefore, only few text detectors are freely available or have been tested in visual attention studies.

Lu, Wang, Lim, and Pomplun (submitted) developed specialized text features, e.g., histograms of edge width and edge density, trained with Support Vector Machine (SVM) classifiers. The study reported better performance compared with earlier studies (e.g., Epshtein, et al., 2010; Jung, et al., 2009) on public text-detecting competition datasets (ICDAR2003 and ICDAR2005). In the present study, we used the automatic text detector developed by Lu et al. (submitted) to test whether it can improve the prediction of viewers' fixations. This detector employs contrast of strokes over background, width of strokes, joints of horizontal and vertical strokes, and stroke structure as key variables

Although manually-defined regions of texts were shown to improve the prediction of eye fixations in text-present images (Cerf et al., 2009), it is unclear if viewers' attention is biased toward any non-text objects which share some features of texts, particularly in text-absent images. In the present study, two eye-movement datasets obtained in our previous investigations (Wang & Pomplun, under revision)

are re-analyzed. The goals of the present study are (1) to investigate the contribution of the automatic text detector to the prediction of eye fixations in real-world scenes, and (2) to verify the hypothesis that viewers' text detection skills are "trained" through exposure to language and affect attentional control even in text-absent scenes

## Experiment 1: Unconstrained Texts

We superimposed unconstrained texts onto real-world scenes, i.e., placed them in unexpected locations, in front of either homogeneous background, i.e., in regions with the lowest luminance contrast in the image before placing the text parts, or inhomogeneous background, i.e., those areas with the highest luminance contrast, and found that texts attracted more attention than non-text objects. This dataset is chosen for re-analysis in the present study since the stimuli contain both text-present and text-absent images. Two models, both including saliency and center-bias maps (channels), but one with and one without text-detector map are compared in order to determine whether the inclusion of the text detector improves the prediction of fixations, particularly in text-absent images.

## Method

**Participants.** Twelve students from the University of Massachusetts at Boston participated. All had normal or corrected-to-normal vision and were between 19 and 40 years old. Each participant received 10 dollars for a half-hour session.

**Apparatus.** Eye movements were recorded using an SR Research EyeLink Remote system with a sampling frequency of 1000 Hz. Subjects sat 65 cm from an LCD monitor approximately 34 x 25 degrees of visual angles. A chin rest was provided to minimize head movements. After calibration, the average error of visual angle in this system is 0.5°. Stimuli were presented on a 19-inch Dell P992 monitor with a refresh rate of 85 Hz and a screen resolution of 1024x768 pixels. Although viewing was binocular, eye movements were recorded from the right eye only.

**Stimuli.** Two hundred natural-scene images were selected from the LabelMe dataset (Russell, Torralba, Murphy & Freeman, 2008). Eighty out of these images were randomly selected to be superimposed with one text and one line drawing. The other 120 images were presented without any modification. For the placement of texts and line drawings, two different items (items A and B in Table 1) were chosen for each scene, and their addition to the scene was performed under four different conditions: either (1) a word describing item A (e.g., "sled" as shown in Table 1) and a drawing of item B, (2) a word describing item B (e.g., "yoyo") and a drawing of item A, (3) a scrambled version of a word describing item A (e.g., "dsle") and a drawing of item B, and (4) a scrambled version of a word describing item B (e.g., "yyoo") and a drawing of item A. All four conditions of text-drawing pairs were presented in a between-subject design, i.e., each participant only viewed one of these conditions. Half of the words (object labels)

were placed in front of homogeneous background and the other half were placed on inhomogeneous background. Figure 1 shows an example of all four conditions with words and drawings on homogeneous background. The eccentricity of the text or the drawing was randomly assigned and varied between 200 and 320 pixels (average: 253 pixels). The minimum polar angle, measured from the screen center, between the text and the drawing in each image was set to 60 degrees to avoid crowding of the artificial items. All texts and drawings were resized to cover approximately 2500 pixels.

Table 1: Examples of texts (words and scrambled words) and object drawings used in Experiment 1.

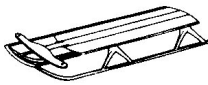

	Item A	Item B
Texts	sled (dsle)	yoyo (yyoo)
Object Drawing		



Figure 1. An example of 4 conditions of stimuli for low-frequency words drawn on homogeneous background. (a) Word of Item A (sled) vs. drawing of Item B, (b) word of Item B (yoyo) vs. drawing of Item A, (c) scrambled word of Item A (dsle) vs. drawing of Item B, and (d) scrambled word of Item B (yyoo) vs. drawing of Item A.

**Procedure.** Equal numbers of subjects freely viewed stimuli from conditions 1, 2, 3, and 4 in a counter-balanced design (described below), and each stimulus was presented for 5 seconds. The free viewing task has been widely used in previous studies (e.g., Judd et al, 2009; Cerf et al., 2009). The software "Eyetrack" developed by Jeffrey D. Kinsey, David J. Straczuzi, and Chuck Clifton, University of Massachusetts Amherst, was used for recording eye movements.



**Analysis.** Two eye movement measures were taken: *correlation (R)* and *Receiver Operating Characteristic (ROC)*. The Pearson correlation coefficient  $R$  between two maps is computed according to sampling points taken every 10 pixels along the  $x$  and  $y$  axes, and then the correlation coefficient between saliency/center-bias/text-detector and attentional maps (described below) are obtained. An example of a stimulus image and its attention, saliency, center-bias, and text-detector maps is shown in Figure 2. The computation of the ROC measure is described in Hwang, Higgins & Pomplun (2009). If a map had higher correlation or ROC values with regard to the subjects' fixations, the map was considered a better predictor of visual attention. The chance level is 0.5 for ROC and 0 for  $R$ .

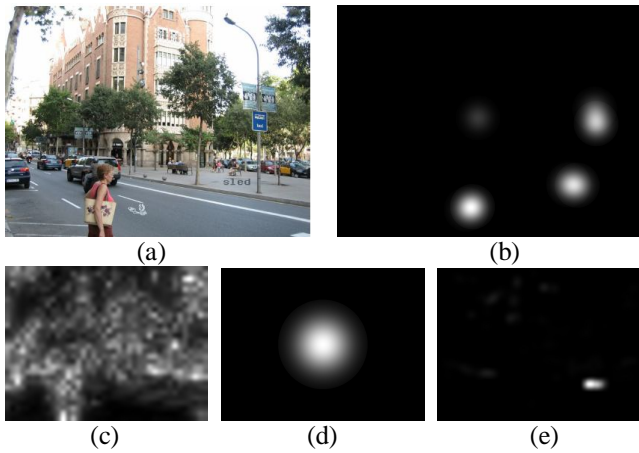


Figure 2. An example of (a) stimulus image, (b) attention (3-second viewing) (c) saliency, (d) center-bias, and (e) text-detector maps.

Saliency was calculated by the freely available computer software “Saliency Map Algorithm” using the standard Itti, Koch, and Niebur (1998) saliency map based on color, intensity, orientation, and contrast. A center-bias map was obtained using a two-dimensional Gaussian distribution at the center of the screen with 3 degrees of visual angle (90 pixels in our experiment setting). The text-detector maps were computed using the automatic text detector which analyzes features such as variation of edge width and edge density.

For the attentional map, we excluded the initial center fixation and included all other fixations within a given viewing duration. The attentional map was built according to each fixation in an image by a two-dimensional Gaussian distribution centered at the fixation point, where the standard deviation was one degree of visual angle to approximate the size of the human fovea. Then we simply summed up these Gaussian distributions for fixations weighted by their durations (see Pomplun, Ritter, & Velichkovsky, 1996).

We computed the attentional maps for each image inspected by each viewer for the initial 1.5, 2, ..., 5 seconds.

The averages of correlations and ROC values for each viewer were calculated for all, text-present, text-absent, text in front of homogeneous (H-BG), and text in front of inhomogeneous backgrounds (INH-BG) images, and an ANOVA and paired t-tests were performed to analyze the differences between these values

## Results and Discussion

**Models with and without Text-Detector Maps.** The average  $R$  and ROC values of all 12 viewers are shown in Table 2. Text-detector maps overlap attentional maps the best when the images contain text in front of homogeneous background, and the worst in text-absent images. These results are consistent with the finding by Judd et al. (2009) that object detectors by themselves do not predict attention well when the objects are absent and therefore should be used in conjunction with other features.

Table 2: The average  $R$  and ROC of saliency (Sali), center-bias (Center), text-detector (TextDet), saliency combined with center-bias (SC), and all combined (SCT) maps as predictors of the attentional maps for 3-second viewing. H-BG represents images in front of homogeneous background, and INH-BG represents images on inhomogeneous background.

	Sali	Cen	TextDet	SC	SCT
R - All	0.14	0.16	0.15	0.18	0.20
Text-Present	0.11	0.12	0.20	0.14	0.16
H-BG	0.09	0.10	0.24	0.10	0.12
INH-BG	0.14	0.15	0.15	0.17	0.19
Text-Absent	0.15	0.19	0.12	0.21	0.22
ROC - All	0.65	0.63	0.63	0.69	0.72
Text-Present	0.61	0.61	0.66	0.64	0.70
H-BG	0.55	0.60	0.67	0.58	0.67
INH-BG	0.67	0.62	0.64	0.70	0.72
Text-Absent	0.67	0.64	0.62	0.72	0.73

One-way ANOVAs with the factor “predictor” showed that the performances of Sali, Cen, TextDet, SC, and SCT maps differed significantly in all, text-present, H-BG, INH-BG, and text-absent images for  $R$ , all  $F_s(4; 55) > 3.64$ ,  $ps < .05$ , and ROC, all  $F_s(4; 55) > 11.17$ ,  $ps < .01$ . SC (without text-detector) obtained significantly lower measures than SCT (with text-detector maps) for all, text-present, H-BG, INH-BG, and text-absent images for  $R$ , all  $t_s(11) > 3.93$ ,  $ps < .01$ , and ROC, all  $t_s(11) > 7.68$ ,  $ps < .001$ . The results indicate that the text detector improved the prediction of viewers' visual attention. It is interesting to see that the SCT obtained higher  $R$  and ROC than the SC even in text-absent images. One plausible explanation is that some non-objects containing text-like features catch a disproportionate amount of attention.

**Text-Present vs. Text-Absent and H-BG vs. INH-BG Images.** The five predictors were analyzed in one-way ANOVAs with the factor “image type,” and the results

demonstrate that both R and ROC values significantly differed in all, text-present, text-absent, H-BG, and INH-BG images, all  $F_s(4; 55) > 4.91$ ,  $p_s < .01$ , and all  $F_s(4; 55) > 4.72$ ,  $p_s < .01$ , respectively, except ROC for Cen,  $F(4; 55) = 0.92$ ,  $p > .4$ . The text detector (TextDet) performed better for text-present images than text-absent ones with regard to R,  $t(11) = 10.67$ ,  $p < .001$  as well as ROC,  $t(11) = 5.66$ ,  $p < .001$ . Homogeneous background images obtained higher values than inhomogeneous background images for both R,  $t(11) = 7.31$ ,  $p < .001$ , and ROC,  $t(11) = 3.94$ ,  $p < .01$ .

**Visual Attention over Time.** SCT outperformed SC (without text detector) for all viewing durations for R and ROC in both text-present images, both  $t_s(11) > 9.68$ ,  $p_s < .001$ , and text-absent ones, both  $t_s(11) > 3.93$ ,  $p_s < .01$ . The difference between SCT and SC was larger in text-present images than in text-absent ones. In text-present images, the R of TextDet initially dominated but decreased over time, while the R of Sali increased (see Figure 3a). These data suggest that texts are typically detected early during the inspection process and receive sustained attention while the viewers are reading them, thereby elevating the occurrence of text features near fixation. Later in the process, viewers tended to be guided more strongly by saliency as defined by the Itti and Koch algorithm. In text-absent images, the R of Sali, Cen, and TextDet increased over time, indicating that the corresponding mechanisms became more important during the later – likely more focused and fine-grained (Unema, Pannasch, Joos, & Velichkovsky, 2005) – stages of inspection. Clearly, Sali and Cen played more important roles when texts are absent.

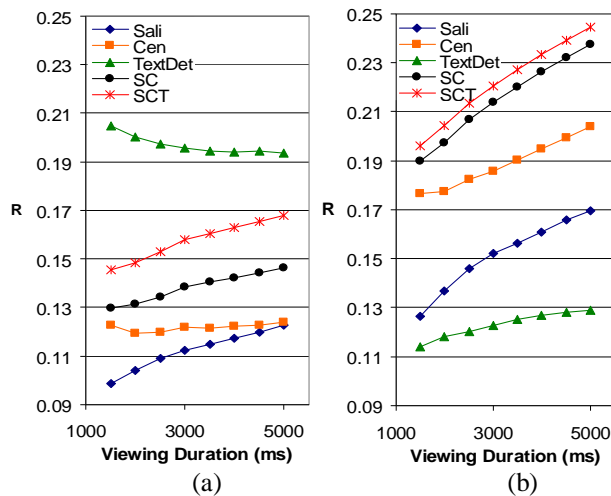


Figure 3. Correlations for 1.5-, 2-, ..., and 5-second viewing of (a) text-present and (b) text-absent images.

## Experiment 2: English vs. Chinese Texts and Native Speakers

In Experiment 1, we showed that the addition of a text-detector map to saliency and center-bias maps makes the model a better predictor of viewers' visual attention. Our

hypothesis is that viewers have developed a “text detector” because they are exposed to texts everyday and become sensitive to text-patterns. Wang and Pomplun (under revision) found that native speakers of English and Chinese-speakers were both attracted by English and Chinese texts in real-world scenes but were attracted more strongly by the texts of their native languages. The reason might be that English and Chinese texts share some common features, such as the histogram of edge width, but also contain their unique features, e.g., Chinese texts usually contain vertical, horizontal, and diagonal strokes but fewer “curves” (such as in “O” or “G” in English). In Experiment 2, the dataset in Wang and Pomplun (submitted) was reanalyzed and our expectation was that the text detector (Lu, submitted) designed for English texts will perform better prediction of gaze fixations for English-speaking viewers than for Chinese-speaking ones.

## Method

**Participants.** In the group of non-Chinese English speakers, 14 students from the University of Massachusetts at Boston participated. All of them were native speakers of English, and none of them had learnt any Chinese or had participated in Experiment 1. For the group of Chinese speakers, 16 native speakers of Chinese were recruited at China Medical University, Taiwan. Each participant received 10 US dollars or 100 Taiwan dollars, respectively, for participation in a half-hour session. All had normal or corrected-to-normal vision.

**Apparatus.** At both sites, the experiment setup was identical to Experiment 1.

**Stimuli.** As shown in Figure 4, the original texts were either rotated by 180 degrees or replaced by Chinese texts. The rationale for using upside-down English texts was to keep the low-level features such as regular spacing and similarity of letters but reduce possible influences of higher-level processing such as meaning. Figure 4a illustrates C1, in which half of the original texts were rotated and the other half was replaced with Chinese texts. In C2, as demonstrated in Figure 4b, the upside-down texts in C1 were replaced with Chinese texts, and the Chinese texts in C1 were replaced with the original, but upside-down, English texts.



Figure 4. Example of Chinese and upside-down English texts used in Experiment 2. (a) Condition C1 (b) Condition C2.

**Procedure.** The procedure was identical to Experiments 1 except that half of the subjects viewed condition 1 (C1) stimuli and the others viewed condition 2 (C2) stimuli in a between-subject counter-balanced design.

**Analysis.** The analyses were identical to Experiment 1.

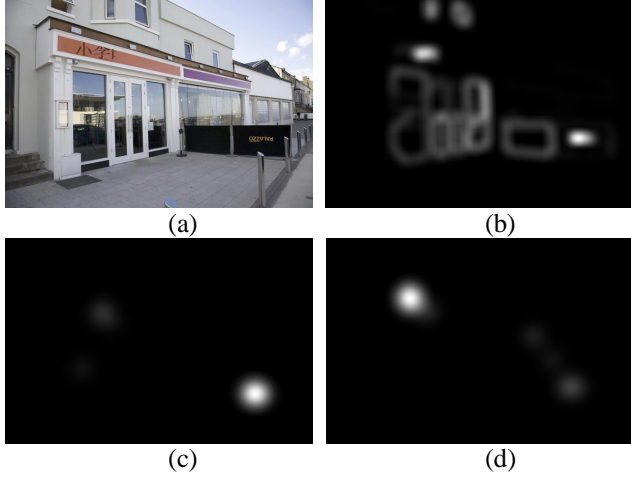


Figure 5. An example of (a) stimulus image, (b) text-detector map, (c), attentional map of an English-speaking viewer (5-second viewing), and (d) attentional map of a Chinese-speaking viewer (5-second viewing).

## Results and Discussion

**Models with and without Text-Detector Maps.** The average R and ROC of all 14 English-speaking and 16 Chinese-speaking viewers are shown in Table 3. For English-speaking viewers, one-way ANOVAs showed that the Sali, Cen, TextDet, SC, and SCT maps performed differently in all, text-present, and text-absent images for R, all  $F_s(4; 65) > 8.47$ ,  $ps < .01$ , and for ROC, all  $F_s(4; 65) > 53.78$ ,  $ps < .001$ . SCT predicted attentional maps better than SC in all, text-present, and text-absent images for R, all  $ts(13) > 3.49$ ,  $ps < .01$ , and ROC, all  $ts(13) > 6.61$ ,  $ps < .001$ . For Chinese-speaking viewers, similar results were obtained - the performances of Sali, Cen, TextDet, SC, and SCT maps significantly differed for both R, all  $F_s(4; 75) > 33.91$ ,  $ps < .001$ , and ROC, all  $F_s(4; 75) > 22.86$ ,  $ps < .001$ . SCT yielded better prediction of attentional maps than SC for both R, all  $ts(15) > 4.85$ ,  $ps < .001$ , and ROC, all  $ts(15) > 5.29$ ,  $ps < .001$ . The results of SCT are consistent with Experiment 1 in that the text detector improved the prediction of viewers' visual attention, even in text-absent images.

**Text-Present vs. Text-Absent Images.** For English-speaking viewers, TextDet performed better in text-present images than in text-absent ones for both R,  $t(13) = 6.41$ ,  $p < .001$ , and ROC,  $t(13) = 5.58$ ,  $p < .001$ . For Chinese-speaking viewers, similar results were found: text-present images obtained higher R and ROC than text-absent ones,  $t(15) = 4.97$ ,  $p < .001$ , and  $t(15) = 7.35$ ,  $p < .001$ , respectively.

**English vs. Chinese-Speaking Viewers.** As shown in Figure 6, TextDet predicted English-speaking viewers' attention better than Chinese-speaking viewers' attention for all viewing durations in both text-present images,  $t(7) = 23.12$ ,  $p < .001$ , and text-absent images,  $t(7) = 5.38$ ,  $p < .01$ . These results indicate that the text detector that was designed for English texts performed better at predicting the allocation of attention for English-speaking viewers than for Chinese-speaking ones.

Table 3: The average R and ROC of saliency (Sali), center-bias (Cen), text-detector (TextDet), saliency combined with center-bias (SC), and all combined (SCT) maps as predictors of attentional maps for 5-second viewing. En represents English-speaking viewers, and Ch means Chinese-speaking viewers.

	Sali	Cen	TextDet	SC	SCT
R (En)	0.17	0.17	0.14	0.20	0.21
Text-Present	0.15	0.16	0.16	0.19	0.21
Text-Absent	0.18	0.17	0.12	0.21	0.22
R (Ch)	0.17	0.16	0.12	0.19	0.20
Text-Present	0.15	0.15	0.14	0.18	0.19
Text-Absent	0.18	0.17	0.11	0.20	0.21
ROC (En)	0.69	0.61	0.60	0.72	0.73
Text-Present	0.68	0.62	0.63	0.71	0.73
Text-Absent	0.69	0.61	0.59	0.72	0.73
ROC (Ch)	0.68	0.60	0.60	0.70	0.71
Text-Present	0.67	0.61	0.62	0.69	0.71
Text-Absent	0.68	0.60	0.58	0.70	0.70

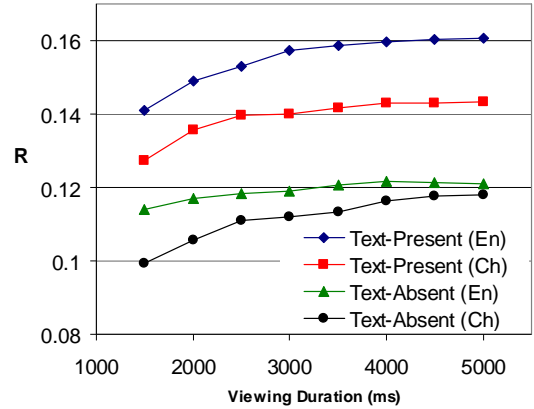


Figure 6. The R values of TextDet for 1.5-, 2-, ..., and 5-second viewing of text-present and text-absent images by English-speaking (En) and Chinese-speaking (Ch) viewers.

## General Discussion

In Experiment 1, we found that adding a text detector to an attention model improved its prediction of viewers' visual attention, even in text-absent images. Our results suggest that non-text objects whose features resemble those of texts (such as high spatial frequency edges) catch a disproportionate share of attention. Based on the current



data, it seems that the viewers' "biological text detectors" are somewhat similar to the artificial system and influence the viewers' distribution of attention when viewing real-world images. From a time-course analysis, it appears that the biological text detector influences the allocation of attention particularly strongly during later stages of image inspection when viewers are increasingly likely to attend to detailed local structures (see Unema et al., 2005) for semantic interpretation of perceived text.

Whereas the results of Experiment 1 could have been caused by the text detection algorithm being sensitive to visual features that generally attract attention, such as edge density, this interpretation becomes implausible given the results of Experiment 2. We found that the text detector designed for English texts predicted English-speaking viewers' attention better than Chinese-speaking viewers', supporting the hypothesis that viewers have developed a "text detector" that is sensitive to text patterns they are familiar with. It is interesting to see that the way we learn to read influences our allocation of visual attention in everyday life, even when there are no texts presented and we are not specifically looking for any texts.

While the present study has demonstrated the influence of language on visual attention in real-world scenes, further research needs to identify the visual features that underlie this effect. This could be achieved by using text detection algorithms for different writing systems and test their individual components as predictors of native and non-native speakers' attention in natural scenes. Besides a more comprehensive understanding of attentional control in humans, such studies may also result in technological advances. Human viewers can easily locate texts in natural scenes, performing clearly better than current text-detection techniques even when the texts are degraded by noise, rotated, distorted, or shown from unusual perspectives. Consequently, the results of this line of research, such as analyzing what features or local structures are actually learned by the biological text detector, might contribute to the development of more effective automatic text detectors, which could, for example, make a great difference to visually challenged people's lives.

### Acknowledgments

Preparation of the article was supported by Grant R01EY021802 from the National Eye Institute to Marc Pomplun.

### References

Baddeley, R. J. & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research*, 46(18), 2824-2833.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, 9(12):10, 1-15.

Epshtein, B., Ofek, E., & Wexler Y. (2010). Detecting text in natural scenes with stroke width transform. *Computer*

*Vision and Pattern Recognition (CVPR)*, San Francisco, USA, 2963-2970.

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A Discriminatively Trained, Multiscale, Deformable Part Model. *Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, USA, 1-8.

Hwang, A. D., Higgins, E. C., & Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision*, 9(5), 1-18 (25).

Itti, L., Koch, C., & Niebur, E. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans Pattern Analysis and Machine Intelligence* 20 (11): 1254-1259.

Itti, L., & Koch, C. (2001). Computational Modeling of Visual Attention. *Nature Reviews Neuroscience*. 2(3):194-203.

Jung, C., Liu, Q., and Kim, J. (2009). A stroke filter and its application for text localization. *Pattern Recognition Letters*, 30(2):114-122.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look, *IEEE International Conference on Computer Vision (ICCV)*, Kyoto, Japan, 2106 - 2113.

Lu, S., Wang, H.-C., J.-H. Lim, & Pomplun, M. (submitted). Learning Text Saliency for Automatic Text Detection in Natural Scenes.

Pomplun, M., Ritter, H., & Velichkovsky B., (1996). Disambiguating Complex Visual Information: Toward Communication of Personal Views of a Scene, *Perception*, 25, 8, 931-948.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation, *International Journal of Computer Vision*, 77, 1-3, 157-173.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4, 1-17.

Torralba, A., Oliva, A., Catelhano, M., & Henderson, J.M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113, 766-786.

Unema, P. J. A., Pannasch, S., Joos, M., & Velichkovsky, B.M. (2005). Time course of information processing during scene perception. *Visual Cognition*, 12(3), 473-494.

Viola, P. & Jones, M. (2004) Robust real-time face detection. *International Journal of Computer Vision*, 57(2), 137-154.

Wang, H. C. & Pomplun M. (2011). The attraction of visual attention to texts in real-world scenes. *The Annual Meeting of the Cognitive Science Society (Cogsci2011)*, 2733-2738.

Ye, Q., Jiao, J., Huang, J., & Yu, H. (2007). Text detection and restoration in natural scene images. *Journal of Visual Communication and Image Representation*. 18, 504-513.

# Implicit theories of the causes of weight gain in adults

Nicole Ware

School of Psychology, Charles Sturt University, Bathurst, NSW 2795, AUSTRALIA

Dr Rachel Dryer (rdryer@csu.edu.au)

School of Psychology, Charles Sturt University, Bathurst, NSW 2795, AUSTRALIA

## Abstract

This study sought to explore the range of beliefs about weight gain and whether these beliefs varied according to personal weight management history. A questionnaire specifically designed for the study was completed by 376 participants (94 males, 282 females; mean age 43.65 years,  $SD=13.24$ ). Principal component analysis identified five causal attribution factors which were interpreted as Lack-of-Self-Control, Lifestyle-Limitations, Psychological, Biological/Medical, and Modern-Living. The level of endorsement for these causal attribution factors suggested an acknowledgement of the multiple causes to weight gain. However, the most highly endorsed factor, Lack-of-Self Control, reflected the “commonsense” view of weight gain being a matter of overeating, under-exercising and lacking in self control. Personal weight management history was found to be associated with variations in beliefs with the more effort one had applied to weight management; the more highly they endorsed causes both within and outside of individual control.

**Keywords:** implicit theories, causal explanations, weight gain, obesity.

## Introduction

Obesity is considered a worldwide epidemic (World Health Organization [WHO], 2000) with over half the adult population in many countries classified as either overweight or obese (Bovbjerg, 2008; Lim, Norman, Clifton, & Noakes, 2008; Thorburn, 2005). Overweight and obesity are widely recognized as being associated with enormous psychological, health, and economic costs, both on individual and national scales (Stroebe, 2008). More recently it has been recognized that such costs occur with any level of weight gain. Increases in body weight of less than five kilograms have been found to be associated with increased disease load, with the associations occurring even within the healthy BMI range (Lim, et al., 2008; WHO, 2000). This has contributed to the WHO’s (2000) decision to advocate the prevention of weight gain in adults as the key strategy in managing obesity. This strategy aims to prevent initial weight gain in normal weight adults, as well as preventing further weight gain in those who are already overweight or obese. At an individual level, prevention of weight gain is quantified as gain of less than five kilograms across adulthood (B. A. Swinburn, et al., 2004).

Prevention of weight gain has now become a focus of public discussion and debate, with numerous theories of causes and solutions in the public arena (Faith, Fontaine, Baskin, & Allison, 2007). Individual changes are often asserted as solutions (Lombard, et al., 2009) along with a range of population scale strategies which include education

campaigns to raise community awareness about the importance of healthy eating, healthy weight and physical activity; changes to urban design and land use to encourage physical activity; and reshaping the food supply to increase access to healthy food, restrict access to unhealthy foods and a regulation of the sugar, fat, and salt contents of foods (NOTF Obesity Working Group, 2009).

Understanding the beliefs and attitudes regarding weight gain held within the community is extremely important for the acceptance and successful implementation of population-level interventions aimed at reducing weight gain (Lombard et al., 2009). Personal beliefs and attitudes about a psychological problem constitute an implicit theory (or explanatory model). An implicit theory contains the individual’s understanding of the causes of the problem, the expected course and prognosis. Implicit theories arise from the individual’s personal experiences, but are primarily mediated by the individual’s social and cultural environment (Furnham, 1988). Many reasons for exploring implicit theories have been identified, particularly within a public health framework. The theories held by lay people may contain elements that have been previously overlooked by the scientific community. In addition, the experience of lay people may place them in a position to identify flaws and shortcomings in current scientific theories or models (Entwistle, Renfrew, Yearley, Forrester, & Lamont, 1998; Popay & Williams, 1996). Addressing such differences between lay knowledge and scientific research also helps to increase the community’s perception of the relevance of research and its resulting policy, and hence the acceptance of interventions (Popay & Williams, 1996).

Various studies have examined the beliefs about obesity amongst lay and professional groups such as doctors, nurses, dieticians, and teachers. These studies have generally found that most people did recognize multiple factors as contributing to obesity, but were more likely to recognize factors within the individual’s control. For example, Ogden and Flanagan (2008) in their comparison of general practitioners and lay people found that behavioral causes of obesity (e.g., not enough exercise, eating too much, and too many unhealthy foods) were most strongly endorsed in both groups. Social causes of obesity such as a lack of education, and biological causes, were not strongly endorsed by either group.

Okonkwo and While (2010), in a study involving University students, found reduced physical activity and the promotion and low cost of fast foods to be strongly endorsed, with genetics receiving the lowest endorsement.

However, they found that participants who were overweight and obese were more likely to endorse genetics and the high costs of fruit and vegetables as causes of obesity. This suggests those who have experienced weight gain are more likely to have greater awareness of factors outside of individual control.

Physical inactivity, eating too much of the wrong foods, and mood changes leading to overeating were again found as the most strongly endorsed causes of obesity in a study of beliefs held by British dieticians (Harvey, Summerbell, Kirk, & Hills, 2002), with biological factors again being least endorsed. This study also sought to determine differences in beliefs about obesity compared to overweight. It was found that the dieticians held similar causal beliefs for both overweight and obesity, but that obese people were seen as more responsible for their weight than were overweight people.

These studies provide insight into beliefs about obesity amongst both lay and professional populations. However, obesity is a recognizable medical condition that refers to an excess of body fat (C. L. Ogden, Carroll & Legal, 2003). Obesity has also often been conceptualized as a biological deviation from the 'normal' healthy state (Jutel, 2006). In contrast, weight gain is less visibly recognizable, is susceptible to fluctuations over the lifespan and affects a larger proportion of the population. Furthermore, individuals may fail to recognize their own weight gain over time and fail to recognize their weight problems (Ziebland, Thorogood, Fuller & Muir, 1996). Consequently, implicit theories about weight gain may be different to those of obesity.

One of the few studies to examine the issue of weight gain was conducted by Jackson, Ball & Crawford (2001) who examined the beliefs about the causes of personal weight gain and loss. The causes were assessed through open answer responses. However, categories for response coding were limited so that a full assessment of causal theories was not possible. Despite this, they found that over one third of participants had gained weight over the previous 12 months, and fewer than half of these acknowledged changes in the amount of food or activity alone as a cause of their weight gain. Other causes given were changes in food type, medical conditions, growth, ageing, and "no special reason".

Paxton and Sculthorpe (1999) examined the issues of weight and weight gain by using the Dieting Beliefs and health locus of control scales and found that beliefs about weight varied according to socioeconomic status and weight. They found that the low SES group was more likely to recognize the influence of factors outside the individual's control (e.g., luck, genes) and environmental factors on weight compared to those in high SES. The authors partly attributed this finding to the limited access to resources faced by the low SES group (Paxton & Sculthorpe, 1999). Consistent with Okonkwo and While's (2010) study, overweight participants were more likely to endorse factors outside of the individual's control as well as environmental factors compared to normal weight participants. However,

they were also more likely to hold the belief that weight is internally controlled. The authors suggest that this higher endorsement of both internally controlled and externally controlled factors may result from both an increased sensitivity to the individual focus of weight loss campaigns, as well as an unsuccessful dieting history, although long term history was not examined.

The current study sought to conduct a more comprehensive examination of the range of beliefs/attitudes about weight gain in adults by using questionnaire items generated from both the general public and the literature on obesity and weight gain. Previous research have used limited number of items and/or predetermined summed categories/scales imposed by the researchers, thereby limiting the ability of these studies to fully explore causal beliefs held by the general community. The current study also examined whether implicit theories of weight gain differed on the basis of personal weight management history. Studies on weight loss intervention have reported that overweight individuals viewed their weight problem as arising from their own motivation and physical shortcomings or as a response to specific issues or challenges in their lives (Greener et al., 2010).

## Method

### Participants

The participants (N= 376; 94 males, 282 females; mean age = 43.25, S.D. = 13.64) in the main study were recruited from regional (e.g., Cobar, Dubbo, Parkes,) and metropolitan areas of Australia (e.g., Adelaide, Melbourne, Sydney) through a snowball sampling approach and random distribution of the questionnaire in shopping areas in a major regional centre in central western New South Wales (e.g., Bathurst, Orange).

### Materials

The items to be included in the questionnaire were developed from both a pilot study and a literature review. Twenty participants (11 females, 9 males), took part in the pilot study. The age range was 18-74 years, (mean=38.00, SD=14.51), with participants from both regional and metropolitan areas. Each participant was interviewed individually and asked to provide possible causes of weight gain in adults. Any causal belief identified by two or more participants were phrased into a questionnaire item and included in the final questionnaire. The resultant items were supplemented by items drawn from the literature including government publications and policy documents (e.g., NOTF, 2006; NOTF Obesity Working Group, 2009; National Preventative Health Taskforce, 2010; Smith, et al., 2005; WHO, 2000; WHO, 2002); previous studies that have explored beliefs about weight management, weight gain, and obesity (e.g., J. Ogden & Flanagan, 2008; Okonkwo & While, 2010); and current literature that examine scientific theories of weight gain and obesity (e.g., Eby & Colditz, 2008; Faith, et al., 2007; Greener, et al., 2010; Lombard, et

al., 2009; Stroebe, 2008; B. Swinburn & Egger, 2004; B. A. Swinburn, et al., 2004).

The final questionnaire listed 42 causal items of weight gain. Participants were asked to rate the importance of each causal item on a six point scale (not at all important to extremely important). Demographic information regarding gender, age, location, education level, current weight and height were also obtained. Weight gain was defined in the questionnaire as a gain of more than 5kg above the participant's usual body weight. Participants were also asked about unplanned weight gain, years spent on weight management, degree of effort in weight management.

## Procedure

Participants recruited using the snow-ball sampling approach were given the choice of paper-based or electronic questionnaires. Envelopes were provided with the option of returning directly to the researcher, or through prepaid post. Participants recruited in shopping areas were given the option of completing the questionnaire at that time or at a later time. Those completing the questionnaire at the shopping area were provided with a sealable envelope to ensure anonymity and confidentiality. Reply-paid envelopes were supplied to shoppers who chose to complete their questionnaire at a later time. Questionnaires were distributed over a three week period across a variety of days and times in an effort to include shoppers from a variety of backgrounds. The questionnaire took approximately 15-20 minutes to complete. Return of the completed questionnaire was taken as indication of consent. The overall return rate of paper-based questionnaires was approximately 47%.

## Results

### Preliminary analyses.

Participant's postcode and suburb were used to classify participants as located in a major city, inner-regional, or outer-regional/very remote according to the Australian Standard Geographical Areas - Remoteness Structure developed by the Australian Bureau of Statistics (ABS). Socioeconomic status was calculated according to postcode percentile rankings within Australia using the ABS Socioeconomic Indexes for Areas (SEIFA) data cube 2006. Current BMIs were calculated using reported height and weight. The planned tests were quite robust to violations of distribution, however where skewness was severe appropriate transformations were conducted and transformed data used for analyses. No significant differences were obtained between the participants using electronic and paper versions of the questionnaire with regards to mean factor ratings.

### Main Analyses.

A principal component analysis was conducted on the ratings of the 42 causal items. A Velicer's minimum average partial test (MAP) (Zwick & Velicer, 1986) was used to determine the number of components to be

extracted. Varimax rotation was applied to determine orthogonal factors and enhance interpretability. Only those items with factor loadings greater than 0.30 were included. To further enhance the uniqueness of the factors, items with similar loadings ( $\pm .20$ ) on multiple factors were excluded from the factors and any further interpretation. Mean factor scores were calculated using a sum of scores by factor divided by the number of items (DiStefano, Zhu, & Míndrilă, 2009).

A five-factor solution was extracted from the data which accounted for 50.57% of the variance. Fifteen items were excluded from further analysis due to similar loadings on two or more factors (see Table 1). The first factor, labeled *Lack-of-Self-Control* (Cronbach's  $\alpha=.81$ ), accounting for 12.19% of the variance, consisted of 8 items relating to a lack of control of diet and exercise. Labeled *Lifestyle-Limitations*, the second factor accounted for 11.37% of the variance and included 5 items (Cronbach's  $\alpha=.76$ ). These items reflect the impact of the higher cost of healthy eating, the influence of long and irregular work hours. This factor also included a lack of awareness of the effects of current lifestyle on weight gain. Explaining 10.63% of the variance, the third factor consisted of 6 items (Cronbach's  $\alpha=.85$ ). This factor was labeled *Psychological* to reflect the content of the items (i.e., depression, stress and low self-confidence). This component also included the ageing item (i.e., Normal part of growing older). The fourth factor was labeled *Biological/Medical* and its items related to hormonal, metabolic and medication-related causes. Consisting of 4 items (Cronbach's  $\alpha=.80$ ) it explained 8.27% of the variance. The final factor, explaining 8.11% of the variance, was labeled *Modern-Living* (Cronbach's  $\alpha=.72$ ). This consisted of 4 items reflecting the reduction in physical activity through the use of cars, modern appliances, and electronic entertainment as well as the recent surge in the "diet" food industry.

Mean factor scores were calculated and are presented in Table 1 with mean item ratings (and standard deviations), the corresponding overall item ranking, and the rotated component loadings. Higher mean scores reflect a greater degree of endorsement in causing weight gain. *Lack-of-Self-Control* was regarded as the most important causal factor with its eight items being the top eight ranked items based on means. Although still acknowledged as important, lower means were found for the other causal attribution factors. Pair-wise comparisons were conducted with a Bonferroni adjustment for the ten possible comparisons resulting in a critical  $\alpha=.005$ . The comparisons confirmed that *Lack-of-Self-Control* was rated as significantly more important compared to the remaining factors (all  $t_s > 20.69$ ,  $p < .0001$ ). The ratings of the remaining causal attribution factors did not differ from each other (all  $t_s < 1.92$ ,  $p > .05$ ).

Mean ratings for each factor were compared according to demographic and weight history. A Bonferroni correction was applied to reduce family-wise error rate across the five causal attribution factors resulting in a critical  $\alpha=.05/5=.01$ . Independent samples t-tests showed that females rated the

importance of *Lifestyle-Limitations* higher than did males  $t(365)=-3.06$ ,  $p=.002$ ,  $d=0.39$ . Females also rated the factors of *Psychological* and *Biological/Medical* as more important as causes of weight gain than did males  $t(372)=-3.56$ ,  $p<.001$ ,  $d=0.43$ ; and  $t(361)=-3.38$ ,  $p=.001$ ,  $d=0.41$ , respectively.

A one-way ANOVA showed an effect for location for the factor of *Modern-Living*<sup>1</sup>,  $F(2, 367)=9.88$ ,  $p<.001$ ,  $\eta^2=.05$ . A Tukey's post-hoc analysis (adjusted for uneven group sizes—Tukey-Kramer) with a critical  $\alpha=0.1/3=.003$  to adjust for the three possible pair-wise comparisons, showed that those residing in major cities rated *Modern-Living*<sup>1</sup> as significantly less important than did those residing in either inner-regional or outer-regional/remote areas. A significant difference was also found for location for the *Lack-of-Self-Control*<sup>1</sup> component,  $F(2, 367)=5.75$ ,  $p=.003$ ,  $\eta^2=.03$ . Post hoc analysis showed that those residing in inner-regional areas rated this factor as being significantly more important than those in major cities.

Correlations (all two tailed) showed that age was weakly associated with *Modern-Living*  $r(243)=.180$ ,  $p<.001$ ,  $R^2=.03$ ; with increasing age reflecting increased importance placed on *Modern-Living* as a cause of weight gain. A greater degree of effort in weight management was related to greater endorsement of the *Lack-of-Self-Control* factor  $r(304)=.26$ ,  $p<.001$ ,  $R^2=.07$ . Increasing amount of effort was also associated with increasing importance attributed to the *Lifestyle-Limitations* factor;  $r(300)=.21$ ,  $p<.001$ ,  $R^2=.04$ ; and the *Psychological* factor,  $r(304)=.23$ ,  $p<.001$ ,  $R^2=.05$ . Similarly, the longer the amount of time spent actively managing weight the greater importance attributed to the *Psychological* factor,  $r(244)=.20$ ,  $p=.002$ ,  $R^2=.04$ .<sup>1</sup>

## Discussion

The aim of the current study was to explore the implicit theories held about weight gain and how these vary according to demographics and personal weight management history. Five factors were obtained to explain causes of weight gain. These were *Lack-of-Self-Control*, *Lifestyle-Limitations*, *Psychological*, *Biological/Medical*, and *Modern-Living*. Overall, the level of endorsement for these factors indicated that they were recognized to some degree as being important; however, the *Lack-of-Self-Control* factor was regarded as the most important in causing weight gain. This factor reflects the “commonsense” view of weight gain with its individual focus of eating too much, not exercising enough, and being lazy or lacking in self-control. The high importance attributed to this factor is consistent with obesity studies in which the most endorsed causes are those regarded to be under individual control (Harvey, et al., 2002; J. Ogden & Flanagan, 2008; Okonkwo & While, 2010).

The *Modern-Living* and *Lifestyle-Limitations* factors recognized social and environmental changes that have impacted on current lifestyles. The *Modern-Living* factor incorporated recent increases in the use of technology such as modern appliances, electronic entertainment, and cars as well as the recent increase in availability and consumption of “diet” foods. The *Lifestyle-Limitations* factor acknowledged the contributions to weight gain through time difficulties associated with long and/or irregular working hours, high costs of healthy food relative to unhealthy foods and lack of awareness of the effects of current lifestyle. The final two factors of the current study reflected individual level causes, but those generally recognized as being outside of individual control. The *Psychological* factor included the effects of emotional issues such as depression and stress, as well as ageing on weight gain. Hormonal and metabolic issues and medication effects were expressed in the *Biological/Medical* factor.

The current study also examined how the beliefs about weight gain varied according to demographics characteristics. Females compared to males, regarded the factors of *Lifestyle-Limitations*, *Psychological*, and *Biological/Medical* as being more important in causing weight gain. Increasing age was found to be associated with increasing endorsement for the *Modern-Living* factor as a cause. Unlike previous studies on obesity, no differences were found according to SES or education level.

Some differences were found in the level of endorsement for the causal attribution factors on the basis of location. Those participants living in major cities rated *Lack-of-Self-Control* as less important compared to those living in inner-regional areas. Those in major cities also rated *Modern-Living* of lower importance than did those living in regional and remote areas. These differences may reflect differences in lifestyle and/or limited resources/options available to those living in regional areas of Australia. For example, limited public transport services and centralization of services in regional areas may have increased the reliance on cars and other technologies leading to a greater awareness of the impact of modern living amongst this population.

The current study also sought to examine whether beliefs about weight gain would be associated with the personal experiences of weight management, with those who have experienced unplanned weight gain and unsuccessful weight management more likely to recognize the role of factors involved in weight management that are outside the individuals control than are those who have not experienced such weight difficulties. However, endorsement of the causal attribution factors did not differ on the basis of whether or not one had experienced unplanned weight gain or on whether one had actively managed their weight. Current BMI was also not associated with the levels of endorsement for any of the causal attribution factors. Instead, this study found that greater time spent (in years)

<sup>1</sup> These factors suffered skewness. These were transformed using logarithms, inverses and SQRT as appropriate for statistical tests with t and F statistics reported for transformed data.

Table 1. Rotated factor item loadings, means, standard deviation and rankings for the five causal attribution factors of weight gain.

Factor labels and items	Mean	SD	Rank	1	2	3	4	5
<b>(1)Lack-of-Self-Control</b> (Cronbach's $\alpha=.81$ )	4.22	0.66						
Eating the wrong types of foods.	4.49	0.87	2	<b>.68</b>				
Eating more food than you need	4.50	0.86	1	<b>.63</b>				
Not enough physical activity/exercise.	4.42	0.92	3	<b>.61</b>				
Lack-of-Self-Control.	4.09	1.07	6	<b>.59</b>				
Eating too many convenience foods/take away.	4.13	1.10	5	<b>.58</b>				
Enjoying high fat/high sugar "bad" foods.	4.30	1.01	4	<b>.54</b>				
Too much snacking.	3.84	1.05	8	<b>.54</b>				
Being lazy.	3.97	1.19	7	<b>.54</b>				
<b>(2)Lifestyle-Limitations</b> (Cronbach's $\alpha=.76$ )	3.12	0.98						
Lack of awareness of problems with current eating/exercise habits.	3.13	1.29	20		<b>.63</b>			
Working long hours.	3.16	1.41	19		<b>.58</b>			
Low price of high fat/ high sugar foods compared to fruit and vegetables.	3.24	1.35	16		<b>.57</b>			
Shift work/irregular working hours.	2.92	1.44	25		<b>.56</b>			
High costs of healthy foods (e.g., fruits, vegetables, grains, lean meat).	3.16	1.42	18		<b>.51</b>			
<b>(3)Psychological</b> (Cronbach's $\alpha=.85$ )	3.20	1.00						
Poor self-confidence	2.94	1.31	24			<b>.70</b>		
Loneliness/social isolation.	3.31	1.40	14			<b>.60</b>		
Low self-esteem.	3.30	1.32	13			<b>.59</b>		
Depression.	3.39	1.40	12			<b>.58</b>		
Stress.	3.47	1.27	11			<b>.57</b>		
Normal part of growing older (i.e., aging)	2.80	1.25	26			<b>.53</b>		
<b>(4)Biological/Medical</b> (Cronbach's $\alpha=.80$ )	3.12	1.08						
Medical conditions – e.g. thyroid problems.	3.25	1.43	15				<b>.75</b>	
Side effect of medications.	3.05	1.38	22				<b>.73</b>	
Hormonal/pregnancy related changes in metabolism.	3.19	1.39	17				<b>.70</b>	
Slow metabolism.	3.00	1.28	23				<b>.60</b>	
<b>(5)Modern-Living</b> (Cronbach's $\alpha=.72$ )	3.19	0.95						
Increased use of modern appliances rather than manual labor e.g. ride on mowers, remote controls	3.11	1.30	21					<b>.69</b>
Increased use of cars over walking/cycling.	3.49	1.20	10					<b>.69</b>
Increased participation in sedentary leisure activities (e.g. TV, computers & electronic games)	3.67	1.21	9					<b>.64</b>
Eating too much of 'diet' 'low fat' 'fat free' foods.	2.50	1.43	27					<b>.43</b>

Fifteen items excluded from analysis due to similar loadings on two or more factors:

Emotional 'comfort' eating	Too much soft/fizzy drinks	Too much alcohol
Larger portion sizes.	Increased consumption of refined/processed foods	A lack of nutritional knowledge
Poor family eating habits	Confusing other cues with hunger (e.g., boredom, thirst)	Disruptive life-events (e.g., divorce, grief)
Genetic factors	Giving up smoking	Lack of time for meal planning
Lack of physical activity at work	Advertising and marketing of unhealthy foods	Eating too little of 'diet', 'low' fat, 'fat free' foods

managing weight was associated with more importance being attributed to the *Psychological* factor. Greater effort on weight management was also associated with higher endorsement of the *Lack-of-Self-Control*, *Lifestyle-Limitations*, and *Psychological* factors. This suggests that increased effort in weight management is associated with increased recognition of a wider range of causes to weight gain which can be both within and outside of the individual's control. This is consistent with the findings reported by Paxton and Sculthorpe (1999) in overweight/obese women and by Greener et al. (2010) in participants with an unsuccessful dieting history. Both of these studies reported that these individuals attributed their weight problem to personal short-comings but were also aware of environmental pressures outside of the individual's control. The current finding also suggests that the amount of effort expended on weight management may provide a better account of relevant weight history than actual weight gain or loss.

The current findings suggest that, in general, the community as a whole needs greater levels of education about the contribution of factors outside the control of the individual in causing weight gain. Educating the general public of the multiple contributing factors to weight gain would also lead to greater acceptance of population-level strategies that are not specifically targeted towards those who are already overweight or obese. This is particularly relevant given the consistency between the current findings about weight gain and beliefs about obesity reported in previous studies.

It should be cautioned that the current findings do not reflect causality. Other limitations of the current study include unequal group sizes within the location, SES and education categories which may have impacted on the number of significant differences obtained. For example, only 25% of the sample was male despite the researchers' efforts at recruiting more male participants. This study was also based on self-report, possibly tapping into a social-desirability bias. However, the anonymous nature of the questionnaire should have assisted in reducing this bias.

## References

- Bovbjerg, V. E. (2008). The epidemiology of obesity: Causal roots - Routes of course. In E. M. Blass (Ed.), *Obesity: Causes, mechanisms, prevention, and treatment* (pp. 19-72). Sunderland, MA: Sinauer Associates, Inc.
- DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Practical Assessment, Research & Evaluation*, 14(20). Retrieved from <http://pareonline.net/pdf/v14n20.pdf>
- Eby, J. G., & Colditz, G. A. (2008). Obesity/Overweight: Prevention and Weight Management. In H. Kris (Ed.), *International Encyclopedia of Public Health* (pp. 602-609). Oxford: Academic Press.
- Entwistle, V. A., Renfrew, M. J., Yearley, S., Forrester, J., & Lamont, T. (1998). Lay perspectives: advantages for health research. *BMJ*, 316(7129), 463-466.
- Faith, M. S., Fontaine, K. R., Baskin, M. L., & Allison, D. B. (2007). Toward the reduction of population obesity: Macrolevel environmental approaches to the problems of food eating, and obesity. *Psychological Bulletin*, 133(2), 205-226.
- Greener, J., Douglas, F., & van Teijlingen, E. (2010). More of the same? Conflicting perspectives of obesity causation and intervention amongst overweight people, health professionals and policy makers. *Social Science & Medicine*, 70(7), 1042-1049. doi: <http://dx.doi.org/10.1016/j.socscimed.2009.11.017>
- Harvey, E. L., Summerbell, C. D., Kirk, S. F. L., & Hills, A. J. (2002). Dietitians' views of overweight and obese people and reported management practices. *Journal Human Nutrition & Dietetics*, 15, 331-347.
- Jackson, M., Ball, K., & Crawford, D. (2001). Beliefs about the causes of weight change in the Australian population. *International Journal of Obesity*, 25(10), 1512-1516. doi: <http://dx.doi.org/10.1038/sj.ijo.0801728>
- Jutel, A. (2006). The emergence of overweight as a disease entity: measuring up normality. *Social Sciences & Medicine*, 63(9), 2268-2276.
- Lim, S. S., Norman, R. J., Clifton, P. M., & Noakes, M. (2008). Losing weight through lifestyle modification: A focus on young women. In A. B. Turley & G. C. Hofmann (Eds.), *Life style and health research progress* (pp. 155-181). Hauppauge, NY: Nova Biomedical Books; US.
- Lombard, C. B., Deeks, A. A., & Teede, H. J. (2009). A systematic review of interventions aimed at the prevention of weight gain in adults. *Public Health Nutrition*, 12(11), 2236-2246.
- National Obesity Task Force. (2006). *Healthy weight for adults and older Australians: A National action agenda to address overweight and obesity in adults and older Australians 2006-2010*. Commonwealth of Australia.
- National Obesity Task Force Obesity Working Group. (2009). *Australia: the Healthiest Country by 2020 Technical Report 1 Obesity in Australia: a need for urgent action: Including addendum for October 2008 to June 2009*. Canberra: Commonwealth of Australia.
- National Preventative Health Taskforce. (2010). *Taking Preventative Action – A Response to Australia: The Healthiest Country by 2020 – The Report of the National Preventative Health Taskforce*. Commonwealth of Australia.
- Ogden, J., & Flanagan, Z. (2008). Beliefs about the causes and solutions to obesity: A comparison of GPs and lay people. *Patient Education and Counselling*, 71, 72-78.
- Okonkwo, O., & While, A. (2010). University students' views of obesity and weight management strategies. *Health Education Journal*, 69(2), 192-199.
- Paxton, S. J., & Sculthorpe, A. (1999). Weight and health locus of control beliefs in an Australian community sample. *Psychology and Health*, 14, 417-431.
- Popay, J., & Williams, G. (1996). Public health research and lay knowledge. *Social Science & Medicine*, 42(5), 759-768. doi: 10.1016/0277-9536(95)00341-x
- Smith, A. M., Lopez-Jimenez, F., McMahon, M. M., Thomas, R. J., Wellik, M. A., Jensen, M. D., & Hensrud, D. D. (2005). Action on Obesity: Report of a Mayo Clinic National Summit. *Mayo Clinic Proceedings*, 80(4), 527-532. doi: 10.4065/80.4.527
- Stroebe, W. (2008). *Dieting, overweight, and obesity: Self-regulation in a food-rich environment*. Washington, DC: American Psychological Association; US.
- Swinburn, B., & Egger, G. (2004). The runaway weight gain train: too many accelerators, not enough brakes. *BMJ*, 329(7468), 736-739. doi: 10.1136/bmj.329.7468.736
- Swinburn, B. A., Caterson, I., Seidell, J. C., & James, W. P. T. (2004). Diet, nutrition and the prevention of excess weight gain and obesity. *Public Health Nutrition*, 7(1A), 123-146.
- World Health Organisation [WHO]. (2006). BMI Classifications. *Global database on body mass index*, 2011, from [http://apps.who.int/bmi/index.jsp?introPage=intro\\_3.html](http://apps.who.int/bmi/index.jsp?introPage=intro_3.html)
- World Health Organization [WHO]. (2000). *Obesity : preventing and managing the global epidemic*. Geneva: World Health Organization.
- World Health Organization [WHO]. (2002). *The world health report 2002: Reducing risks, promoting healthy life*. Geneva.
- Ziebland, S., Thorogood, M., Fuller, A., & Muir, J. (1996). Desire for the body normal: body image and discrepancies between self reported and measured height and weight in a British population. *Journal of Epidemiological Community Health*, 50, 105 – 106.
- Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for detremining the number of components to retain. *Psychological Bulletin*, 99, 423-442.



# Relationship between Phonemes and Tactile-emotional Evaluations in Japanese Sound Symbolic Words

**Junji Watanabe (watanabe.junji@lab.ntt.co.jp)**

NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation  
3-1, Morinosato-Wakamiya, Atsugi, Kanagawa, 243-0198 Japan

**Yuuka Utsunomiya (u0816007@edu.hc.uec.ac.jp)**

Department of Informatics and Engineering, The University of Electro-Communications  
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

**Hiroya Tsukurimichi (h.tukurimichi@edu.hc.uec.ac.jp)**

Department of Informatics and Engineering, The University of Electro-Communications  
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

**Maki Sakamoto (sakamoto@inf.uec.ac.jp)**

Department of Informatics and Engineering, The University of Electro-Communications  
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

## Abstract

Many languages have a word class whose speech sounds are linked to sensory experiences (sound symbolism). Here we investigated sound symbolism in touch. Specifically, we performed psychophysical experiments to study the relationship between phonemes of Japanese sound symbolic words and emotional evaluations of objects in contact. Participants expressed the sensation obtained from touched materials using Japanese sound symbolic words and evaluated the comfort of tactile sensation. Our results show the existence of unique associations between the phonemes of the words for expressing the sensation and the evaluations of comfort in touch. Next, we compared the results with those for a condition in which the evaluations of comfort on tactile materials were made without expressing the sensations. We found that when certain phonemes are used for expressing the sensation, the evaluations can be biased.

**Keywords:** Sound symbolism; Tactile sensation; Emotional judgment; Onomatopoeia

## Introduction

Against a classical notion in linguistics that speech sounds and meanings of words are independent, the existence of synesthetic associations between sounds and sensory experiences (sound symbolism) has been demonstrated over the decades (e.g., Fox, 1935; Jespersen, 1922; Köhler, 1929, 1947; Newman, 1933; Sapir, 1929; Taylor, 1963; Werner & Wapner, 1952; Wertheimer, 1958, for early studies) and, more or less, in the languages of the world (e.g., Brown, Black, & Horowitz, 1955; Davis, 1961; Emeneau, 1969; Enfield, 2005; Hinton, Nichols, & Ohala, 1994; Klank, Huang, & Johnson, 1971; Kovic, Plunkett, & Westermann, 2010; Nuckolls, 1999; Voeltz & Kilian-Hatz, 2001). The characteristics and universality of such sensory-sound correspondence have been studied to provide a clue for understanding the development of language abilities (Imai, Kita, Nagumo, & Okada, 2008; Kantartzis, Imai, & Kita,

2011; Maurer, Pathman, & Mondloch, 2006; Westbury, 2004) and language evolution (Ohala, 1997; Ramachandran & Hubbard, 2001, 2003).

It is also known that the sensory-sound correspondence can be found not only in words referring to visual shapes, which were demonstrated in the landmark studies (e.g., mal/mil and buba/kiki for round and sharp shapes in Sapir, 1929 and Ramachandran & Hubbard, 2001, respectively), but also in those referring to tactile, smell, and taste sensations. However, the majority of studies in the area of sound symbolism have been limited to visual-sound correspondence. Consequently, we are investigating the sound symbolic associations in touch, specifically the association between the phonemes of Japanese sound symbolic words (onomatopoeia) for expressing tactile sensations and subjective evaluations of comfort/discomfort for touched objects.

We focus on tactile-emotional evaluations, because compared with other languages, Japanese is known to have a large number of onomatopoeic words for tactile sensations, and because associations between the phonemes of Japanese onomatopoeic words and typical categories of tactile *sensations* (not *emotion*) can be observed. For example, onomatopoeic words expressing a sense of smoothness often use the consonant /s/ in the first syllable as in "sara-sara" [(a) in Table 1], while those expressing roughness often use /z/ in the first syllable as in "zara-zara" [(b) in Table 1]. Similarly, characteristic first consonants are observed in each of the tactile categories, such as /p/ for softness as in "puru-puru" [(c) in Table 1] and /k/ for hardness as in "kachi-kachi" [(d) in Table 1]. More importantly, tactile sensations are strongly connected to changes in the states of comfort and discomfort (Gallace, & Spence, 2010), and although it has been reported that touching objects can evoke distinct emotional states according to the textures of touched objects (Ramachandran,

& Brang, 2008), the associations between speech sound for expressing tactile sensations and touch-induced emotional states are unclear.

Table 1: Typical Japanese onomatopoeic words for expressing tactile sensations.

Sounds	Meanings
(a) Sara-sara	Smoothness
(b) Zara-zara	Roughness
(c) Puru-puru	Softness
(d) Kachi-kachi	Hardness

In the current study, we performed two psychophysical experiments. In the first experiment, when participants touched an object, they were asked to express the tactile sensation using Japanese onomatopoeic words [sound symbolic words (SSWs)] and then rate the comfort of the touched object with the semantic differential (SD) method (referred as the SSW+SD condition). This condition was aimed at specifying the systematic association between phonemes of Japanese onomatopoeic words and tactile-emotional evaluations. Our results demonstrate for the first time the existence of unique associations between them.

Next, we compared the results for the SSW+SD condition with those for a condition in which participants made only the tactile-emotional evaluations (SD-only condition). Since recent brain-imaging studies suggests that processing of SSWs could activate corresponding sensory areas (Arata, Imai, Okuda, Okada, & Matsuda, 2010; Hashimoto, Usui, Taira, Nose, Haji, & Kojima, 2006; Osaka, Osaka, Morishita, Kondo, & Fukuyama, 2004), it is expected that speaking them might affect speaker's subjective evaluation of the touched objects. The results demonstrate that when certain phonemes are used for expressing the sensation, the evaluation can be biased.

## Materials and Methods

### Participants

Thirty naïve participants, aged between 19 and 26 years old, took part in the experiments. Fifteen of the 30 (ten males and five females) performed the experiment in the SSW+SD condition; the other fifteen (ten males and five females) performed the experiment in the SD-only condition. They were unaware of the purpose of the experiments, and they had no known abnormalities of their verbal or tactile sensory systems or any particular skills with respect to touch. They visited a laboratory at the University of Electro-Communications for one day to conduct trials. Informed consent was obtained from the participants before the experiment started. Recruitment of the participants and experimental procedures were approved by the University of Electro-Communication Research Ethics Committee and were conducted in accordance with the Declaration of Helsinki.

### Apparatus and Materials

We selected 120 types of tactile materials for the experiments, including fabrics, papers, metals, leathers, rubbers, woods, sand, rocks, and plastics. Preliminarily, we confirmed that onomatopoeic words for expressing the 120 materials would cover the major phonemes of Japanese onomatopoeic words in touch. When feasible, samples were cut to a size of 6 cm x 6 cm and stacked in layers to 2-mm thickness. The rocks and sand were loose in a container. As illustrated in Fig. 1, participants sat in front of a box with an 8 cm x 10 cm hole in it (the materials box) and placed the index finger of the dominant hand into the box through the hole to touch a material; they could not see a material while they were touching it.

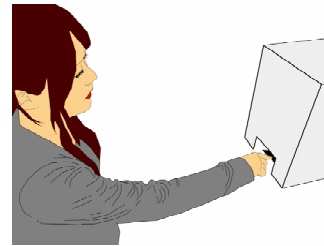


Figure 1: Participant touching a material.

### Procedure

A trial in the SSW+SD condition comprised a describing period, in which participants reported an onomatopoeic word to express the tactile feeling while touching one of the 120 materials, and a rating period, in which they evaluated the comfort/discomfort of the touched material on a seven-point scale (Very comfortable +3, Comfortable +2, Slightly comfortable +1, Neither 0, and three levels, -1 to -3, for uncomfortable feeling) while touching it. In the describing and rating periods, they could freely run their fingers and push on the surface of the material, and the time for answering was not limited. A trial in the SD-only condition comprised only the rating period. In the experiments, the experimenter placed one material in the material box, and the participant touch it and respond, and then replaced it with another after the participant's response. Tactile materials were presented in a random order.

### Results for SSW+SD condition

The results for the SSW+SD condition produced 1800 sets of onomatopoeic words and subjective evaluations of comfort level (120 materials x 15 participants). In 87.1% of all trials (1569 cases), the onomatopoeic word had a two-mora repetition form (e.g. "sara-sara"). We therefore analyzed the relationship between the phonemes of onomatopoeic words in two-mora repetitions form and evaluations of comfort/discomfort using the 1569 instances. The average of the ratings obtained across the 1569 cases was 0.08. This value suggests that half of trials yielded a rating in a comfortable level, and the other half a rating in an uncomfortable level (nonbiased for participant's response opportunities).

We took particular note of the vowels and consonants of the first syllable, which show strong sound symbolic associations (Hamano, 1998). The averages of ratings across trials in which the same phonemes were used in the first syllable were statistically compared with the average of the 1569 cases (0.08) (t-test comparing the average with constant value). Of the vowels, as shown in Table 2, only /u/ had a statistically significant relationship to comfort. Vowels /i/ and /e/ were not used often, but they were deeply related to discomfort (/i/ was marginally significant, and /e/ was significant). Among the consonants, /m/, /h/, and /s/ were related to comfort, while /g/, /z/, /j/, /b/, and /n/ were significantly related to discomfort.

Next, we combined the vowels and consonants of the first syllables and did a similar statistical analysis for the values

(e.g., an analysis based on the value of /sa/ rather than the values of /s/ and /a/ separately). Table 3 shows the results of combinations between the vowels and consonants, which are listed in Table 2. The vowel /u/ was significantly related to comfort with many consonants (e.g., /h/ as in "huwa-huwa", /s/ as in "sube-sube", /t/ as in "tsuru-tsuru", and /p/ as in "puru-puru"). Conversely, /e/ was related to discomfort with any consonant (e.g., /p/ as in "pecha-pecha", /n/ as in "neba-neba", and /b/ as in "beta-beta"). Consonants /m/, /h/, and /s/ were related to comfort regardless of the vowel, while /g/, /z/, /j/, /b/, and /n/ were related to discomfort regardless of the vowel. Vowels /o/, /a/ and /i/ and consonants /p/ and /t/ were related to different levels of comfort depending on the consonant or vowel they were combined with, suggesting that as phonemes, they have only a weak relation to comfort level.

Table 2: Numbers (Num.) and average of ratings (Ave.) of the first vowel and consonant. Only phonemes whose numbers were more than 16 (1 % of 1569 cases) are listed. P-values of statistical analysis to the overall average (0.08) are also shown (+:  $p < 0.1$ , \*:  $p < 0.05$ , \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ ). Red and blue shading indicate that the phoneme is significantly different in positive and negative values, respectively.

First vowel				First consonant			
	Num.	Ave.	$p$ value		Num.	Ave.	$p$ value
/u/	528	0.63	0.00 (***)	/m/	38	1.32	0.00 (***)
/o/	215	0.10	0.88	/h/	132	1.11	0.00 (***)
/a/	504	0.02	0.32	/s/	243	0.65	0.00 (***)
/i/	102	-0.21	0.05 (+)	/t/	189	0.54	0.14
/e/	220	-0.95	0.00 (***)	/k/	74	0.18	0.47
				/p/	209	0.07	0.87
				/g/	148	-0.18	0.02 (*)
				/z/	218	-0.33	0.00 (***)
				/j/	35	-0.47	0.03 (*)
				/b/	204	-0.96	0.00 (***)
				/n/	38	-1.37	0.00 (***)

Table 3: Averages of ratings for combinations of first consonants and vowels, which are listed in Table 2. Significant differences from the overall average (0.08) are also shown (+:  $p < 0.1$ , \*:  $p < 0.05$ , \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ ). Red and blue shading indicate that the combination is significantly different in positive and negative values, respectively.

	/u/	/o/	/a/	/i/	/e/
/m/		1.31 (***)			
/h/	1.24 (***)				
/s/	0.63 (***)		0.66 (***)	0.61 (+)	
/t/	0.83 (***)			-0.65 (*)	
/k/			0.04		
/p/	0.42 (+)	0.75 (**)	-0.13	0.14	-0.41 (***)
/g/	-0.09	-0.02	-0.32 (*)		
/z/			-0.31 (***)		
/j/			-0.24		
/b/	-0.15	-0.58 (***)			-1.46 (***)
/n/					-1.15 (***)

## Results for SD-only condition

We compared the results for the SSW+SD condition with those for the SD-only condition to examine the influence of expressing the sensations with sound symbolic words on the emotional evaluation (comfort level). First, we calculated averages of ratings for each tactile material (120 materials) in the SSW+SD (using 1569 cases) and SD-only (using 1800 cases) conditions. The analyses using 240 averaged values (120 materials x 2 conditions) were performed in two domains, in which the averaged values in the SD-only condition were positive (comfort materials) or negative (discomfort materials). For comfort materials (62 materials), the means of ratings across 62 materials were 0.70 in SSW+SD and 0.73 in SD-only. They were not significantly different (paired t-test,  $t(61) = 0.39$ ,  $p = 0.70$ ). On the other hand, for discomfort materials (58 materials), the means across 58 materials were -0.48 in SSW+SD and -0.73 in the SD-only. There were significant differences between the values ( $t(57) = 3.41$ ,  $p < .001$ ). These statistical analyses indicate that expressing tactile sensations with sound symbolic words while touching materials does not affect the emotional evaluation if the material is originally comfortable. However, when a touched material elicits discomfort, the emotional evaluation can be attenuated by expressing the sensation with sound symbolic words.

To clarify the systematic association between phonemes of onomatopoeic words and their influence on the evaluation, we made distribution maps of 120 materials as shown in Figs. 2 and 3, respectively. The phonemes represent each material in the maps. The representing phonemes were determined from the first syllables used for expressing the sensations most in the trials in the SSW+SD condition. Ones in the right-side area in the maps are comfortable materials and vice versa. In addition, ones in the upper side in the maps indicate materials whose comfort levels are improved (Left side: attenuated discomfort. Right side: enhanced comfort) and vice versa. Numbers of phonemes in the upper and lower right side areas are almost identical in the maps. On the other hand, more phonemes are located in the upper left area (attenuated discomfort) than in the lower left area (enhanced discomfort). This means that the comfort levels were systematically biased toward comfort for the discomfort materials, which agrees with the previous statistical analyses.

The differences in the effect of attenuation or enhancement according to the phonemes were statistically examined. We collected phonemes whose numbers in the comfort or discomfort area were more than six (5% of 120 materials) and calculated the means of differences of rating values in the SSW+SD condition from those in the SD-only condition across their corresponding materials. Figure 4 shows the trends for the comfort materials. For vowels, /u/ slightly enhances the comfort (marginally significant effect), but /a/ attenuates the comfort. For consonants, none of the phonemes affected comfort levels for the comfort materials. Figure 5 shows the trends for the discomfort materials.

When the phonemes, /u/ in vowels, /z/ and /p/ in consonants, were used for expressing the tactile sensation, the level of discomfort can be attenuated. Some negative values in comfort level (discomfort materials) became positive (perceived as comfort).

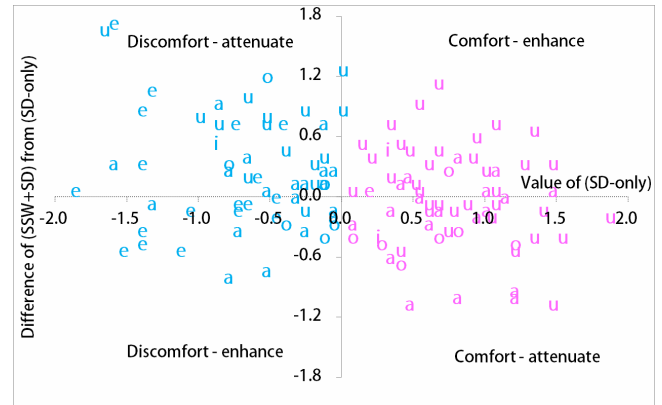


Figure 2: Distributions of vowels according to the values in the SD-only condition (horizontal axis) and difference of SSW+SD from SD-only (vertical axis).

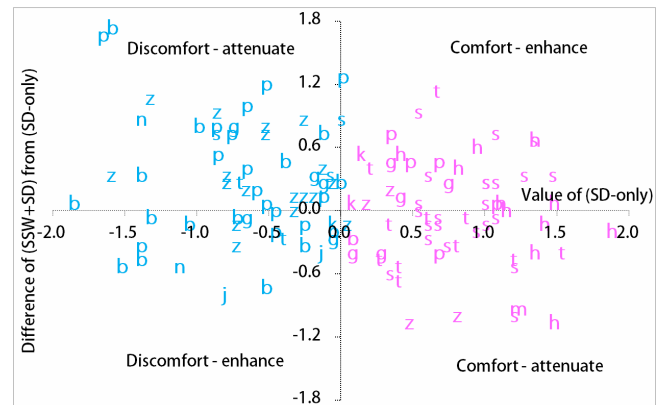


Figure 3: Distributions of consonants according to values in the SD-only condition (horizontal axis) and difference of SSW+SD from SD-only (vertical axis).

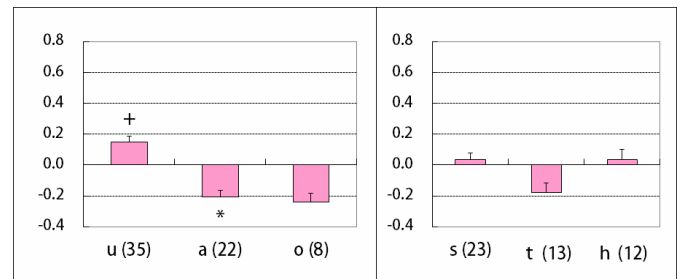


Figure 4: Difference of ratings in the SSW+SD condition from those in the SD-only condition for the phonemes in the comfort area. The data for vowels (left) and consonants (right) are shown with the number and statistical differences from zero. +:  $p < 0.1$ , \*:  $p < 0.05$ .

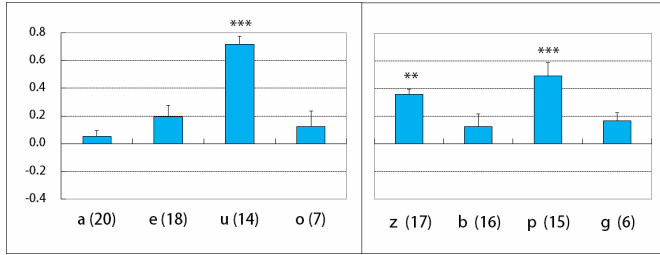


Figure 5: Difference of ratings in the SSW+SD condition from those in the SD-only condition for the phonemes in the discomfort area. The data for vowels (left) and consonants (right) are shown with the number and statistical differences from zero. \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ .

## Discussion

### Sensory Category and Sound Symbolism

Humans categorize sensory inputs using words, and words are important indexes in investigating such sensory categories. Previous touch studies identified major factors in object recognition through touch and established standards of classification based on physical properties of materials or in terms of sets of adjectives (Hollins, Bensmaïa, Karlof, & Young 2000; Hollins, Faldowski, Rao, & Young, 1993; Picard, Dacremont, Valentin, & Giboreau, 2003; Tiest, & Kappers, 2006). However, the studies did not discuss sensibilities, such as like and dislike or comfort and discomfort, and to our knowledge, there has been no research regarding the association between touch-induced emotional states and sound symbolic words.

In the experiment in the SWS+SD condition, we found a strong association between comfort and the vowel /u/ and consonants /m/, /h/, and /s/ in the first syllables. If similar sensory categories are expressed using onomatopoeic words with similar phonemes, it will be possible to clarify the categories of comfort/discomfort, which is difficult to study directly. Since the tactile sensations described with the phonemes /m/, /h/, and /s/ seem different, as with "moko-moko" and "huka-huka" for squishiness and "sara-sara" for smoothness, these phonemes might be originated from different categories of comfort. Similarly, for discomfort, the vowel /i/ as in "chiku-chiku" for pointiness and /e/ and consonants /n/ and /b/ as in "neba-neba" and "beta-beta" for stickiness, and consonants /j/, /z/, and /g/ as in "jori-jori," "zara-zara," and "gori-gori" for roughness may represent different uncomfortable categories.

### Uniqueness of Tactile Sound Symbolism

Phonemes related to tactile comfort or discomfort, together with the meanings of general sound symbolism in Japanese onomatopoeic words, are summarized in Table 4 (see Hamano, 1998). The association between the vowel /u/ and comfort has not been identified in the general sound symbolism. For vowel /i/, tactile sensations can be seen as analogous to general sound symbolism, but the association with discomfort seems to be particular to tactile sensation. The association of /e/ with discomfort can be seen as

analogous to general sound symbolism. For consonants, no association between comfort and consonants /m/, /h/ and /s/ has been suggested. For discomfort, the tactile perception and comfort level agree with general sound symbolism for /n/. But it is difficult to analogize the comfort level for /b/ from general sound symbolism, and the association of /z/, /j/, and /g/ with discomfort also appears to be specific to the tactile sense.

As for consistency between modalities, there seems to be similar trends in the sound symbolic relationships between phonemes and perceptual experiences regarding touch and tastes. According to Sakamoto and Chiba (2005), vowel /u/ and consonants /s/ and /h/ are associated with positive evaluations of tastes, while vowels /i/ and /e/ and consonants /n/, /z/, /j/, /g/, and /b/ are associated with negative evaluations. In addition, as for the generality of our results, although the sound symbolism is observed in the languages of the world, whether the trends obtained in our experiments are universal or not is an issue awaiting further investigation.

Table 4: Phonemes related to tactile comfort or discomfort, together with the meanings of general sound symbolism in Japanese onomatopoeic words.

	Comfort	General sound symbolism	
/u/	+	Small round holes/ projections	Specific
/i/	-	Lines, straight elongations	Specific
/e/	-	Coarse, vulgar, unsuitable	Analogous
/m/	+	Indistinct, unclear	Specific
/h/	+	Softness	Specific
/s/	+	Smoothness	Specific
/n/	-	Sticky, uncomfortable	Agree
/b/	-	Taut	Specific
/j/	-	Abrasive	Specific
/z/	-	Abrasive	Specific
/g/	-	Touching a hard surface	Specific

### Affection of Sound Symbolic Words on Evaluation

We found that when certain phonemes are used for expressing sensations for the touched material, the evaluations of comfort levels can be biased. Specifically, when /u/ was used, the bias was always toward comfort, which agrees with the association between phonemes and emotional evaluations in Table 2. But neutral phonemes /a/ and /p/ affect the evaluation. In addition, consonant /z/, which is associated with discomfort, attenuated the discomfort. Considering that emotional states can be evoked by the sensations automatically (Ramachandran, & Brang, 2008) and how words are articulated can be a possible mechanism relevant to sound symbolism (Ohala, 1983), the sound symbolism of comfort and discomfort for the tactile sense may be affected by phonetic (acoustic) comfort or discomfort, as observed in our experiment. But explaining the association is a problem for future research.



## Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research on Innovative Areas (No. 22135007) from the Ministry of Education, Culture, Sports, Science and Technology in Japan.

## References

- Arata, M., Imai, M., Okuda, J., Okada, H. & Matsuda, T. (2010). Gesture in language: How sound symbolic words are processed in the brain. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp.1374-1379).
- Brown, R. W., Black, A. H., & Horowitz, A. E. (1955). Phonetic symbolism in natural languages. *The Journal of Abnormal and Social Psychology*, 50, 388-393.
- Davis, R. (1961). The fitness of names to drawings: A cross-cultural study in Tanganyika. *British Journal of Psychology*, 52, 259-268.
- Emeneau, M. B. (1969). Onomatopoeics in the Indian linguistic area. *Language*, 45, 274-299.
- Enfield, N. J. (2005). Areal linguistics and mainland southeast Asia. *Annual Review of Anthropology*, 34, 181-206.
- Fox, C. W. (1935). An experimental study of naming. *The American Journal of Psychology*, 47, 545-579.
- Gallace, A., & Spence, C. (2010). The science of interpersonal touch: An overview; *Neuroscience and Biobehavioral Reviews*, 34, 246-59.
- Hamano, S. (1998). *The sound symbolic system of Japanese*. Stanford, CA: CSLI Publications; Tokyo: Kuroshio.
- Hashimoto, T., Usui, N., Taira, M., Nose, I., Haji, T., & Kojima, S. (2006). The neural mechanism associated with the processing of onomatopoeic sounds. *Neuroimage*, 31, 1762-1770.
- Hinton, L., Nichols, J., & Ohala, J. (Eds.). (1994). *Sound symbolism*. Cambridge, UK: Cambridge University Press.
- Hollins, M., Bensmaïa, S., Karlof, K. & Young, F. (2000). Individual differences in perceptual space for tactile textures: Evidence from multidimensional scaling. *Perception and Psychophysics*, 62, 1534-1544.
- Hollins, M., Faldowski, R., Rao, S., & Young, F. (1993). Perceptual dimensions of tactile surface texture: A multidimensional scaling analysis. *Perception and Psychophysics*, 54, 697-705.
- Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism between a word and an action facilitates early verb learning. *Cognition*, 109, 54-65.
- Jespersen, O. (1922). The symbolic value of the vowel i. *Philologica*, 1, 1-19.
- Kantartzis, K., Imai, M., & Kita, S. (2011). Japanese Sound-Symbolism Facilitates Word Learning in English-Speaking Children. *Cognitive Science*, 35, 575-586.
- Klank, L. J. K., Huang, Y. H., & Johnson, R. C. (1971). Determinants of success in matching word pairs in tests of phonetic symbolism. *Journal of Verbal Learning and Verbal Behavior*, 10, 140-148.
- Köhler, W. (1929) *Gestalt Psychology*. New York: Liveright Publishing Corporation.
- Köhler, W. (1947) *Gestalt Psychology (2<sup>nd</sup> Ed.): An Introduction to New Concepts in Modern Psychology*. New York: Liveright Publishing Corporation.
- Kovic, V., Plunkett, K., & Westermann, G. (2010). The shape of words in the brain. *Cognition*, 114, 19-28.
- Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound-shape correspondences in toddlers and adults. *Developmental Science*, 9, 316-322.
- Newman, S. S. (1933). Further experiments in phonetic symbolism. *The American Journal of Psychology*, 45, 53-75.
- Nuckolls, J. (1999). The case for sound symbolism. *Annual Review of Anthropology*, 28, 225-252.
- Ohala J. J. (1983) The origin of sound patterns in vocal tract constraints. In: MacNeilage (ed.) *The production of speech*. Berlin: Springer-Verlag.
- Ohala, J. J. (1997). Sound Symbolism. *Proceedings of 4th Seoul International Conference on Linguistics* (pp. 98-103).
- Osaka, N., Osaka, M., Morishita, M., Kondo, H., & Fukuyama, H. (2004). A word expressing affective pain activates the anterior cingulate cortex in the human brain: an fMRI study, *Behavioral Brain Research*, 153, 123-7.
- Picard, D., Dacremont, C., Valentin, D., & Giboreau, A. (2003). Perceptual dimensions of tactile Textures. *Acta Psychologica*, 114, 165-184.
- Ramachandran, V. S., & Brang, D. (2008). Tactile-emotion synesthesia. *Neurocase*, 14, 390-399.
- Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia -A window into perception, thought, and language. *Journal of Consciousness Studies*, 8, 3-34.
- Ramachandran, V. S., & Hubbard, E. M. (2003). Hearing colors, tasting shapes. *Scientific American*, 288, 43-49.
- Sakamoto, M. & Chiba, A. (2005). Sound symbolic analyses of onomatopoeic words expressing tastes. *Proceedings of 130<sup>th</sup> Annual Meeting of Linguistic Society of Japan*, 306-311 (in Japanese) .
- Sapir, E. (1929). A study of phonetic symbolism. *Journal of Experimental Psychology*, 12, 225-239.
- Taylor, I. K. (1963). Phonetic symbolism re-examined. *Psychological Bulletin*, 60, 200-209.
- Tiest, W. M. B., & Kappers, A. M. L. (2006). Analysis of haptic perception of materials by multidimensional scaling and physical measurements of roughness and compressibility. *Acta Psychologica*, 121, 1-20.
- Voeltz, F. K. E., & Kilian-Hatz, C. (Eds.). (2001). *Ideophones*. Amsterdam: John Benjamins.
- Werner, H., & Wapner, S. (1952). Toward a general theory of perception. *Psychological Review*, 59, 324-38.
- Wertheimer, M. (1958). The relation between the sound of a word and its meaning. *The American Journal of Psychology*, 71, 412-415.
- Westbury, C. (2004). Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and Language*, 93, 10-19.

# Actor-Observer Asymmetries in Judgments of Intentional Actions

**Sarah Wellen (swellen@andrew.cmu.edu)**

Department of Philosophy, 135 Baker Hall, Carnegie Mellon University  
Pittsburgh, PA 15213 USA

**David Danks (ddanks@cmu.edu)**

Department of Philosophy, 135 Baker Hall, Carnegie Mellon University  
Pittsburgh, PA 15213 USA; and  
Institute for Human & Machine Cognition, 40 S. Alcaniz St.  
Pensacola, FL 32502

## Abstract

Much recent empirical research has explored the influence of moral evaluations on judgments about the intentionality of foreseeable side-effects of actions. Research on this ‘Side-Effect effect’ (also called the ‘Knobe effect’) has relied almost exclusively on vignette-based surveys, which have serious limitations when used in isolation. We present a novel behavioral methodology that tests the Side-Effect effect in two previously unexamined contexts: (i) judgments of real (rather than hypothetical) actions, and (ii) judgments about one’s own actions. The results suggest that judgments about one’s own actions tend to show a reverse Side-Effect effect: actors judge that (real) positive side-effects of their own actions are intentional whereas negative ones are not. The use of non-hypothetical situations also appears to attenuate the standard Side-Effect effect, which raises interesting challenges for standard theoretical accounts. These results provide preliminary evidence that the Side-Effect effect is driven by the same mechanisms underlying other asymmetries in causal attribution.

**Keywords:** Intentional action; Social cognition; Moral psychology; Actor-observer bias; Side-effect effect; Knobe effect; Vignettes

## Introduction

Notions of ‘intentionality’ and ‘causation’ are crucial in the way we understand our own and others’ mental lives, as well as a range of social interactions (e.g. Gergely, Nadasdy, Csibra, & Biro, 1994; Gopnik, *et al.*, 2004; Leslie & Keeble, 1987; Sloman, 2005; Woodward, Sommerville, & Guajardo, 2001). There has recently been a surge in research on folk understandings of these notions, much of it by experimental philosophers (e.g. Adams & Steadman, 2004a, 2004b; Knobe, 2003, 2004, 2006; Machery, 2008; Nadelhoffer, 2006; Nado, 2008; Uttich & Lombrozo, 2010). This research has revealed that folk judgments often display some surprising patterns. Perhaps the canonical example is the so-called ‘Side-Effect effect’ (also called the ‘Knobe effect’): experimental participants judge that a foreseeable side-effect of an action is more intentional when the side-effect is morally bad than when it is morally good. This paper develops a novel experimental method to investigate the Side-Effect effect that avoids some of the limitations of previous methods. Moreover, we report findings that

suggest that the Side-Effect effect is an instance of a much more general pattern of actor-observer biases, rather than a distinct phenomenon in its own right.

## The Side-Effect Effect

Consider a company chairman who acts to maximize profit, even though that action has the (foreseeable) side-effect of harming the environment. Suppose also that the chairman does not care about the environment; the occurrence of this side-effect is irrelevant (in any way) to his decision. Knobe (2003) found that 85% of participants judged that the chairman had nonetheless *intentionally* harmed the environment. When the program had the side-effect of helping the environment, though, only 23% of participants judged that the chairman had intentionally helped the environment (Knobe, 2003). In both vignettes, the chairman’s explicit goals (i.e., his explicit intentions) and actions remain the same, yet people’s judgments about whether he intentionally brought about the side-effect change significantly depending on the valence of the side-effect. This influence of outcome valence was present cross-culturally (Knobe & Burra, 2006), in four-year-old children (Leslie, Knobe, & Cohen, 2006), and for a range of mental state ascriptions besides simply intentionality (Pettit & Knobe, 2009). While there is general agreement that the intentions of the actor should influence our moral evaluations, it is less clear why the converse should hold. Moral concerns seemingly ought not to have a pervasive impact on what appear to be straightforward judgments about an actor’s mental states (though see Bratman, 1987 and Uttich & Lombrozo, 2010, for defenses of this influence).

Two different types of theories have been offered to explain the Side-Effect effect. Side-effect-centered theories hold that the different responses arise because of properties of the side-effect: some feature of negative or positive side-effects leads the actions to be judged as intentional or unintentional (respectively), or both. For instance, perhaps negative side-effects are always thought to be brought about intentionally (Knobe, 2004, 2006; but see also Pettit & Knobe, 2009, for an updated view), or negative cases require the actor to judge trade-offs (Machery, 2008). Alternately, the fact that an action violates the generalized



norm against bringing about negative outcomes might (rationally) be informative about the actor's mental state, since norm-violation typically requires additional reasons or intentional action (Uttich & Lombrozo, 2010).

In contrast, process-centered theories hold that our judgments would be symmetric in the two conditions, except that one of the situation-types (or both) elicits a mental process that alters or biases our judgment away from our typically symmetric intuitions. Various influencing processes have been suggested: a negative emotional reaction to the chairman (Nadelhoffer, 2006); a desire to blame based on conversational pragmatics of intentional language (Adams & Steadman, 2004a, 2004b); an asymmetric responsibility judgment (Wright & Bengson, 2009); or a distinct "moral mechanism" (Nado, 2008).

The Side-Effect effect has been studied almost exclusively using vignette-based surveys, which arguably have serious limitations when used in isolation. Vignettes focus on hypothetical situations in which many important (in the real world) properties and conditions are left unspecified. Exactly what information is included in (or omitted from) a vignette has been shown to have a substantial effect on people's judgments (Gugliermo & Malle, 2010; Mele & Cushman, 2007; Phelan & Sarkissian, 2008). Moreover, vignettes describe hypothetical situations, and previous research has found that judgments can differ substantially when provided in response to real rather than hypothetical situations. For instance, experimental participants reveal different utility functions when they are given real and hypothetical scenarios (List & Gallet, 2001; Murphy, Allen, Stevens, & Weatherhead, 2005; Neill, Cummings, Ganderton, Harrison, & McGuckin, 1994). Vignettes also require (almost always) that the experimental participant be an observer, not the actor. Thus, it is an open question whether the Side-effect effect occurs when an individual judges the intentionality of the side-effects of her own actions.

### Actor-Observer Asymmetries

Actor-observer asymmetries have been found in a wide range of domains. For instance, actors and observers largely attend to different aspects of a social interaction (Malle & Pearce, 2001). Actors are willing to engage in riskier behaviors than observers will condone (Fernandez-Duque & Wifall, 2007) and we appear to hold ourselves to different moral standards than we hold others to (Nadelhoffer & Feltz, 2008). One classic asymmetry is the Actor-Observer hypothesis in causal attribution (Jones & Nisbett, 1971): we tend to emphasize internal dispositional factors (e.g., attitudes, personality traits) as the causes of other people's actions, and aspects of the external situation (e.g., social constraints, situational factors) as the causes of our own actions. Recent research suggests the classic formulation is an oversimplification; instead of a simple internal/external distinction, actors tend to emphasize their reasons for acting (the beliefs, desires, and values that contributed to their decision to act), whereas observers tend to cite the causal

history of these reasons, including the attitudes, personality traits, and upbringing of the actor (Knobe & Malle, 2002).

A recent meta-analysis found little evidence for a general Actor-Observer asymmetry, but did find evidence for one that was mediated by the valence of the action (Malle, 2006). Observers tend to use more internal explanations than actors when explaining a negative action, but *fewer* internal explanations when explaining a positive action:

Table 1: Influences on causal attribution (adapted from Malle, 2006)

	Negative Valence	Positive Valence
<b>Overall</b>	$\text{diff}_A < \text{diff}_O$ <sup>1</sup>	$\text{diff}_A > \text{diff}_O$
<b>Real Situations</b>	$\text{diff}_A < \text{diff}_O$	$\text{diff}_A > \text{diff}_O$
<b>Hypothetical Situations</b>	$\text{diff}_A < \text{diff}_O$	$\text{diff}_A \approx \text{diff}_O$

Malle (2006) found that the valence-modified actor-observer bias was magnified when participants explained real events (middle row of Table 1). For hypothetical events, however, there was no difference between actors' and observers' explanations of positive actions. This suggests that, at least when explaining our own actions, it makes a difference whether the action is real or hypothetical.

This paper presents a novel experimental method in which participants judged real actions from either an actor or an observer role. We used a  $3 \times 3$  design that varies role {Actor, Observer, Motivated Observer} between-participant and side-effect valence {Neutral, Positive, Negative} within-participant. Side-effect-centered theories largely predict that the actor-observer manipulation should make no difference, since actors and observers should share the same conceptual asymmetries. The exception is the "Rational Scientist" view of Uttich & Lombrozo (2010), which arguably predicts that there should be no Side-Effect effect in the actor condition, regardless of outcome, since one presumably already knows one's own mental state. Most process-centered theories do not make a determinate prediction, since they leave unspecified whether and how the process that causes the asymmetry applies to judgments of one's own actions. Finally, if participants exhibit the asymmetric pattern of judgments characteristic of the valence-modified Actor-Observer hypothesis, then it is likely that the Side-Effect effect is due to the same mechanisms driving asymmetries in causal attributions.

## Experiment

### Participants

46 participants from the McGill, Concordia, and Carnegie Mellon University communities were divided into the Actor

<sup>1</sup> 'diff<sub>A</sub>' and 'diff<sub>O</sub>' refer to the difference between the number of internal and external reasons given by actors and observers, respectively, in the different conditions.

(N = 16), Observer (N = 15), and Motivated Observer (N = 15) conditions. Participants were mostly undergraduates; all were fluent in English and without cognitive deficits.

## Method

The experiment involved a computer game in which the participants (in the Actor condition) generated their own actions, discovered the consequences of these actions, and judged whether they had brought about those consequences intentionally. The game involved a computer interface where each movement of a joystick (up, down, left, or right) led to one or more colored balls being displayed on the computer screen. One of these balls (the red one) was the 'goal ball' and each time the actors got this ball they were rewarded with 10 tokens. Participants in the Actor and Motivated Observer conditions were told that they would redeem the tokens for money at the end of the experiment (but were not provided with the 'exchange rate'); participants in the Observer condition were simply told that collecting tokens was the goal of the game.

The experiment had four phases. In the *Practice* phase, participants in every condition (Actor/Observer/Motivated Observer) played the computer game. During this phase, a joystick movement up deterministically produced a red ball (and so 10 tokens), while movements in all other directions produced a different color ball (and so no tokens). Each action thus produced a single consequence. Participants continued until they generated the red ball on six consecutive trials. At this point, the connections were shuffled (without notice) so that each movement produced a different color ball than before, and participants continued until they again generated the red ball six consecutive times by discovering and then moving in the new rewarded direction. The Practice phase continued until the movement-ball connections had been shuffled three times.

**Actor Condition** Participants in the Actor condition then moved (without notice) to the *Neutral* phase. This phase was identical to the practice phase, except that participants now received *two* balls after each joystick move. A move in the "rewarded" direction always produced both a red ball and a white ball, and every other move resulted in a random pair of unrewarded balls (white, yellow, blue, or green). The white ball thus played the role of a foreseeable side-effect: participants quickly learned that it always appeared with the red ball, but were indifferent to its occurrence. The white ball was introduced after the Practice phase to ensure that participants were motivationally neutral towards it.

The rewarded direction was randomly selected at the start of the Neutral phase. Once participants had made six consecutive moves in that direction, the rewarded direction would change and participants would have to rediscover it. The Neutral phase ended when the participant discovered the rewarded direction after the third change of ball-movement connections.

Participants then entered either the *Negative* or *Positive* phase (order counterbalanced between participants). In these

phases, participants were informed that generating the white ball now had consequences: a randomly chosen other participant in the experiment (i.e., *not* the participant) would either gain (Positive) or lose (Negative) three tokens for each white ball that the participant generated (no others were actually helped or harmed, however). As in the Neutral phase, participants continued until there had been four different rewarded positions. After completing one of these phases, Actor participants were informed of the change in the valence of the white ball, and then performed the other phase.

Throughout all three non-practice phases (Neutral, Positive, and Negative), the following two questions were asked at regular intervals:

Action question: You moved [direction]. How intentional was this?

Color question: You got a [color] ball. How intentional was this?

The Action question was always presented directly after a joystick move, and the Color question was presented after the ball display screen. We are particularly interested in responses about the white (side-effect) ball after participants had already discovered the currently rewarded direction. In this case, the participants presumably could foresee that they would get the white ball, so the white ball is a foreseen side-effect of their action. Answers were reported by clicking on a (540-point) rating scale with left-middle-right anchor points of "not at all intentional", "somewhat intentional", and "completely intentional".

**Observer Condition** In the Observer condition, participants performed the practice phase, but then watched a video of an actor playing the Neutral phase instead of playing it themselves. Participants were told that the other individual was receiving real money when she got red balls. The actor's hand, joystick, and computer screen were shown, but not her face. Whenever the actor was asked an Action question, participants observed her answer "completely intentional." But instead of observing the actor answer for the Color question, the video paused and participants were asked:

Color question (Observer): Julie got a [color] ball. How intentional was this?

Participants were able to replay the video as much as desired before providing an answer. After the Neutral phase, participants observed the same actor in the Positive and Negative phases. Participants were told the relevant valence information, and the video order of the Positive and Negative phases was counterbalanced between participants. As a comprehension check, participants were also given a follow-up questionnaire asking whether the actor was trying to get the white ball in the Positive and Negative phases.

**Motivated Observer Condition** The Motivated Observer condition was identical to the Observer condition, except that participants were informed that they were the individual who had been randomly selected as the 'other participant' to

gain (Positive) or lose (Negative) three tokens whenever the actor got the white ball. They were reminded that tokens were redeemable for actual money.

## Results

There were no order effects, so different orderings were pooled for further analyses. Although the questions used a 540-point rating scale, responses largely concentrated around the three labeled points, indicating that many participants interpreted the scale as categorical. We thus transformed the data into a five-point scale, where each category corresponded to an equal-length segment of the original line. This transformation captured the generally categorical use of the line while preserving information about those few participants who used between-label points in their answers. Intentionality ratings for the goal (i.e., the red ball) were all at ceiling. In contrast, mean intentionality ratings for the foreseen side-effect (i.e., the white ball) are shown in Figure 1 (error bars indicate confidence intervals).

In the Actor condition, a planned, one-way within-participants ANOVA revealed a significant effect of phase valence ( $F(2, 42)=6.09, p<.05$ ). Post hoc Tukey HSD comparisons revealed that the Positive phase ( $M=3.18, SD=1.34$ ) led to significantly higher ratings than the Neutral ( $M=2.10, SD=1.36$ ) and Negative ( $M=2.09, SD=1.25$ ) phases,  $p<.05$ ; the latter two were not significantly different. Participants in the Actor condition thus rated the foreseen side-effect as significantly more intentional when it had a positive valence than a negative one, as found in standard actor-observer biases in causal attribution. A planned, one-way within-participants ANOVA did not reveal a significant effect of valence in the Observer ( $F(2, 42)=0.36, p=.69$ ) or Motivated Observer ( $F(2, 42)=0.48, p=.62$ ) conditions. The conditions were designed to also permit a statistical comparison of ratings from different conditions. One-way between-participants ANOVAs revealed significant effects of participant role for the Neutral ( $F(2, 42)=9.33, p<.05$ ) and Negative ( $F(2, 42)=6.97, p<.05$ ) phases, but not for the Positive phase ( $F(2, 42)=1.3, p=.28$ ). Actors had significantly different ratings than both the Observers and the Motivated Observers, except when the side-effect was positive.

Since every participant experienced all three valence phases, we can also analyze the rating patterns of each individual. We define ‘POS’ participants to be those who gave significantly higher intentionality ratings for the Positive phase compared to the Negative phase. As predicted by the valence-modified actor-observer hypothesis, the frequency of POS participants was significantly greater in the Actor condition (8/16 participants) than in the Observer condition (4/15 participants),  $\chi^2(1, N=30)=.96, p<.05$ , or Motivated Observer (2/15 participants) conditions,  $\chi^2(1, N=30)=1.0, p<.05$ . We can also approximate forced-choice probes (common in Side-Effect effect experiments) by further discretizing intentionality ratings into a binary variable (“Side effect was intended” iff rating  $\geq 3$ ). For the Actor

condition, the positive side-effect was significantly ( $p<.05$ ) more likely to have been rated as intentional than the negative one (75% vs. 38%). There was no corresponding difference in the Observer (86% vs. 73%;  $p>.05$ ) or Motivated Observer (73% vs. 86%;  $p>.05$ ) conditions.

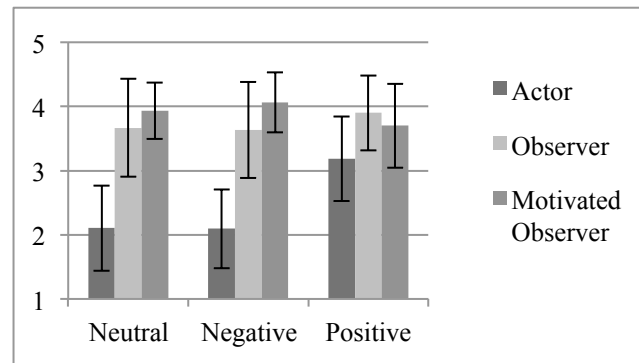


Figure 1: Mean intentionality ratings for foreseen side-effect (95% confidence intervals)

## Discussion

When asked to judge their *own* actions, participants were systematically more likely to judge those actions as intentional when the foreseen side-effect was positive rather than negative; that is, they exhibit something like a reverse Side-Effect effect. This pattern of results closely mirrors the valence-modified actor-observer hypothesis in causal attribution, where actors are more likely to cite internal dispositional factors as the causes of their positive actions than their negative ones. This suggests that the mechanisms driving the Side-Effect effect are plausibly the same ones driving asymmetries in causal attribution, although the nature of these mechanisms is still an open question.

Process-centered accounts focus primarily on the motivations of the observers. While these accounts have not made explicit predictions about the actor case, it is clear that in moral scenarios actors and observers often have different motivational pressures. Actors may have an incentive to minimize their responsibility (blameworthiness) for negative side-effects while overestimating their responsibility (praiseworthiness) for positive ones. Observers do not have these same concerns, and may even have incentives that tend in the opposite direction. However, we found that actors and observers differed in their intentionality judgments even for neutral side-effects, a result that cannot be explained by process-centered accounts. This difference is instead plausibly explained in terms of the epistemic differences between actors and observers. Actors have greater access to their own intentions, and thus are more aware that the side-effect (regardless of valence) was not among their goals. The observers do not have direct access to the mental states of the actor and thus they may rely more heavily on the behavior of the actor and the norms of the situation as a guide to her intentions.

This epistemic explanation fits nicely with the ‘Rational Scientist’ view, wherein epistemic differences between

norm-conforming and norm-violating cases drive the asymmetries in observer judgments. However, these epistemic considerations alone are unable to explain the reverse side-effect for actors, because actors have (at least perceived) direct access to their own mental states. It seems likely, then, that a combination of epistemic and motivational factors is necessary to explain the full range of results. Further research should explore this possibility.

Interestingly, we did not find a significant difference between judgments in the Positive and Negative phases in the Observer condition, as has typically been found using vignette studies. There are several possible reasons for this null result. One explanation is that we used visually presented situations involving real people rather than hypothetical vignettes. The abstract nature of vignettes potentially magnifies various distorting effects; for example, participants are perhaps less likely to believe the hypothetical chairman's reported intentions and more likely to be swayed by a moral evaluation of his actions. Another (not mutually exclusive) possibility is that the side-effect in the present experiment was not sufficiently negative or positive from the perspective of the observers. This explanation seems unlikely, however, since no significant difference was found even in the Motivated Observer condition where the side-effect directly affected the participant. Similarly, Feltz, Harris, & Perez (2010) found no significant Side-Effect effect for observer conditions using realistic settings, suggesting that our finding is not simply an artifact of our experimental design. Finally, an explanation could be given from the perspective of the 'Rational Scientist' view: the norms of computer games in psychological experiments differ from the norms of corporations affecting the environment, and it might be that the observers did not believe the actors in the experiment were violating any norms. This explanation is in some tension, however, with the fact that individuals were observed to violate the generalized norm not to harm others.

The results we report do not provide a definitive connection between the Side-Effect effect in vignette studies and judgments in realistic settings. The vignette results may plausibly generalize in cases where the side-effect valence is strongly negative or positive, when bringing about the negative side-effect involves violating a norm, or when there is greater psychological distance between the participant and the person whose actions are being judged (e.g., when reading a news article or listening to a description), although we currently have no evidence that this is the case. It is clear, however, that the Side-Effect effect should be situated within a more general class of valence-mediated actor-observer biases, and that the connections between these biases and folk psychological judgments warrant further investigation.

### Acknowledgments

Andrew Reisner, Lesley Fellows and Nathalie Camille provided valuable advice and assistance. We also thank Joshua Knobe, Thomas Nadelhoffer, and other participants

at the 2010 SPP Conference for helpful comments on an earlier version of this paper. This research was partially supported by the James S. McDonnell Foundation.

### References

- Adams, F., & Steadman, A. (2004a). Intentional action and moral considerations: Still pragmatic. *Analysis*, 64(3), 268-276.
- Adams, F., & Steadman, A. (2004b). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*, 64(2), 173-181.
- Bratman, M. (1987). *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.
- Fernandez-Duque, D., & Wifall, T. (2007). Actor/observer asymmetry in risky decision making. *Judgment and Decision Making Journal*, 2(1), 1-8.
- Feltz, A., Harris, M., & Perez, A. (2010). Perspective in intentional action attribution: Reversing the side-effect effect. Manuscript Submitted for Publication.
- Gergely, G., Nadasdy, Z., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, 56, 165-193.
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T. & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111, 3-32.
- Gugliermo, S., & Malle, B. F. (in press). Can unintended side-effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*. In Press.
- Jones, E. E., & Nisbett, R. E. (1971). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. E. Kanouse, H. H. Kelly, R. E. Nisbett, S. Valins & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior*. Morristown, NJ: General Learning Press.
- Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63(3), 190-194.
- Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis*, 64(2), 181-187.
- Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, 130(2), 203-231.
- Knobe, J., & Burra, A. (2006). The folk concepts of intention and intentional action: A cross-cultural study. *Journal of Cognition and Culture*, 6(1-2), 113-132.
- Knobe, J., & Malle, B. F. (2002). Self and other in the explanation of behavior. *Psychologica Belgica*, 42, 113-130.
- Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25, 265-288.
- Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science*, 17(5), 421-427.
- List, J. A., & Gallet, C. A. (2001). What experimental protocol influence disparities between actual and

- hypothetical stated values? *Environmental & Resource Economics*, 20(3), 241-254.
- Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind & Language*, 23, 165-189.
- Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, 132(6), 895-919.
- Malle, B. F., & Pearce, G. E. (2001). Attention to behavioral events during social interaction: Two actor-observer gaps and three attempts to close them. *Journal of Personality and Social Psychology*, 81, 278-294.
- Mele, A. R., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. *Midwest Studies in Philosophy*, 31, 184-201.
- Murphy, J. J., Allen, P. G., Stevens, T. H., & Weatherhead, D. (2005). A meta-analysis of hypothetical bias in stated preference valuation. *Environmental & Resource Economics*, 30(3), 313-325.
- Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for jury impartiality. *Philosophical Explorations*, 9(2), 203-220.
- Nadelhoffer, T., & Feltz, A. (2008). The actor-observer bias and moral intuitions: Adding fuel to Sinnott-Armstrong's fire. *Neuroethics*, 1(2), 133-144.
- Nado, J. (2008). Effects of moral cognition on judgments of intentionality. *The British Journal for the Philosophy of Science*, 59, 709-731.
- Neill, H. R., Cummings, R. G., Ganderton, P. T., Harrison, G. W., & McGuckin, T. (1994). Hypothetical surveys and real economic commitments. *Land Economics*, 70(2), 145-154.
- Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgement. *Mind & Language*, 24(5), 586-604.
- Phelan, M. T. & Sarkissian, H. (2008). The folk strike back; Or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies*, 138, 291-298.
- Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. Oxford: Oxford University Press.
- Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, 116, 87-100.
- Woodward, A., Sommerville, J., Guajardo, J. (2001). How infants make sense of intentional action. In B. F. Malle, L. J. Moses, D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition*. Cambridge, MA: The MIT Press.
- Wright, J., & Bengson, J. (2009). Asymmetries in judgments of responsibility and intentional action. *Mind & Language*, 24(1), 24-50.

# Learning Causal Structure through Local Prediction-error Learning

**Sarah Wellen (swellen@andrew.cmu.edu)**

Department of Philosophy, Baker Hall 135  
Pittsburgh, PA 15213 USA

**David Danks (ddanks@cmu.edu)**

Department of Philosophy, Baker Hall 135  
Pittsburgh, PA 15213 USA; and  
Institute for Human & Machine Cognition, 40 S. Alcaniz St.  
Pensacola, FL 32502 USA

## Abstract

Research on human causal learning has largely focused on strength learning, or on computational-level theories; there are few formal algorithmic models of how people learn causal structure from covariations. We introduce a model that learns causal structure in a local manner via prediction-error learning. This local learning is then integrated dynamically into a unified representation of causal structure. The model uses computationally plausible approximations of (locally) rational learning, and so represents a hybrid between the associationist and rational paradigms in causal learning research. We conclude by showing that the model provides a good fit to data from a previous experiment.

**Keywords:** Causal learning; causal Bayes nets; prediction-error learning; algorithmic level

## Introduction

From a young age, we spontaneously, and often effortlessly, come to understand the causal structure of the world, and then use that knowledge to both predict what might happen in the future and also design actions that will achieve our goals (e.g., Gopnik, *et al.*, 2004; Sloman, 2005). Our focus here is causal learning from covariational data: how do people learn the causal structure of the world from a sequence of observations or interventions of that world?

Causal learning can usefully be separated into the related-but-distinct problems of representation and dynamics—*what* is learned and *how* is it learned. In this paper, we develop a novel account of causal learning that, at a high level, uses quasi-associationist processes to learn directed graph-like causal representations. It is thus a hybrid of the standard rationalist vs. associationist approaches to causal learning.

## Representations of Causal Structure

The development of causal Bayesian networks prompted a major advance in our understanding of causal knowledge. A causal Bayes net has two components: (i) a directed acyclic graph (DAG) whose nodes represent variables and directed edges represent direct causal relations (see Figure 1); and (ii) a probability distribution that encodes how causes influence their effects. These two elements represent qualitative and quantitative causal structure, respectively,

and are connected by a pair of assumptions (Markov and Faithfulness) that capture the ways in which causal structure manifests in observed data. Sloman (2005) and Spirtes, Glymour, & Scheines (1993) provide useful expositions of the causal Bayes net framework.

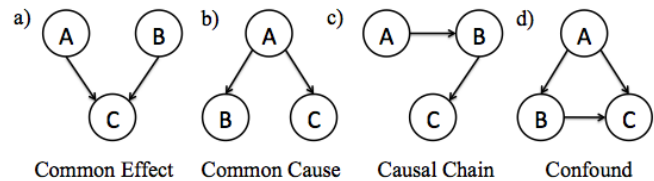


Figure 1: Prototypical 3-variable causal Bayes nets

There is substantial evidence that the type of structural knowledge captured by a causal Bayes net—or at least, the directed graphical model part—is necessary to account for many causal reasoning abilities. One hallmark of *causal* reasoning, rather than correlational, is that cases involving observations vs. interventions are treated differently (Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005). For example, one can infer, from an observation of a professor's gray hair, that she likely has many publications. No such inference follows if she instead intervened to dye her hair gray. Causal Bayes nets can straightforwardly account for this difference, as interventions are represented by 'graph surgery,' where a variable that is intervened upon is separated from its typical causes (Spirtes, *et al.*, 1993). This surgery changes the informational relations, and so one's inferences can be different in the two situations.

Some aspects of causal knowledge are not easily represented by this formalism (e.g. the spatiotemporal relations between causes and effects), but it seems to provide a good account of people's representations of causal structure. Thus, we aim to develop a theory of causal learning in which people learn a directed graph (perhaps acyclic, though we will allow for cyclic structures).

## Dynamics of Causal Structure Learning

Theories about how people use covariation to learn directed graph representations can be divided roughly into *rational* and *heuristic* accounts of causal learning. Rational accounts

model causal learning as rational inference. These include constraint-based algorithms (e.g., Glymour, 2003; Gopnik, *et al.*, 2004), and those based on Bayesian inference (e.g., Steyvers, *et al.*, 2003; Griffiths & Tenenbaum, 2005). They are usually intended at the computational level of analysis, as they show how the cognitive system's performance solves the problem faced by that system, but do not attempt to characterize the underlying cognitive processes. There have been some recent attempts to develop algorithmic (i.e. process) models of causal learning based on approximations of Bayesian inference (e.g., Bonawitz, *et al.*, 2011). These models have so far only addressed causal strength learning, and it is not clear how to extend them (in a computationally tractable manner) to structure learning.

Heuristic accounts of causal learning propose that people use various cues to suggest and modify causal hypotheses in a not-necessarily-rational (though presumably sensible) manner. Causal model theory (Waldmann, 1996) proposes that learners use cues such as covariation, temporal order, and spatial proximity to select an initial causal structure and adjust it in the face of inconsistent data (Lagnado, Waldmann, Hagmayer, & Sloman, 2007). Causal model theory has never been entirely formally specified, though some parts have received formal treatment.

The local computations model (Fernbach & Sloman, 2009) attempts to explain how learners use data from interventions to learn a causal structure. The key idea is that, when a variable is intervened upon and other variables change, the learner infers that the intervened-upon variable caused those other variables. Critically, all learning in this model is local, as people evaluate individual causal relations rather than entire graphs. The model we present here adopts this important insight and extends it to all covariation-based structure learning, including learning from observations.

The single-effect learning model (Waldmann, *et al.*, 2008) also assumes that people focus on evaluating single causal relations. It is a model of learning from observations, and proposes that learners estimate the causal power (Cheng, 1997) of each potential cause of an effect. If a variable has sufficient (estimated) causal power, then the learner accepts the causal relation and integrates it with her previous causal knowledge. This model has found some empirical support in both humans and rats (Waldmann, *et al.*, 2008).

Our model adopts the single-effect learning model's focus on causal power, and the integration of these individually learned relations into a unified causal structure. However, the standard causal power theory is a computational theory that makes no commitment to underlying processes. Danks, Griffiths, & Tenenbaum (2003) provided a prediction-error-based model of causal strength learning whose equilibrium states are causal powers, and so their model can be viewed as an algorithmic implementation of the causal power theory. Moreover, its basis in prediction-errors is consistent with neuroscientific evidence that the right lateral prefrontal cortex encodes prediction-error signals during causal learning (Corlett, *et al.*, 2004; Turner, *et al.*, 2004).

Another lacuna in the single-effect learning model is that it does not explain how the learner uses a causal power estimate to determine whether a link actually exists. We thus provide a decision procedure for causal relation acceptance based on both the learner's point estimate and her confidence in that estimate. This addition allows us to model the dynamics of learning for directed graphs that are more complex than the single-effect structure.

## The LPL Model

The Local Prediction-error Learning (LPL) model aims to explain how observations and interventions are used to learn causal structure when one has relatively little prior knowledge. We do not model many other relevant sources of information, including verbal communication, reasoning, or spatiotemporal information. The model does assume that the learner knows the functional form of the causal relations and (when relevant) the expected temporal delay between causes and effects.

The LPL model begins with an initial causal structure hypothesis: a directed graph representing the individual's prior beliefs, where an edge indicates an *a priori* belief that there is a causal connection, and absence indicates only agnosticism.<sup>1</sup> For typical experiments in which participants have little prior knowledge, this will be an empty graph. The model alters this causal structure hypothesis by adding or removing single edges, thereby reducing the structure learning problem to the simpler task of evaluating individual causal relations. Multiple experimental results suggest that learners focus primarily on single causal relations (e.g., Gopnik *et al.*, 2004; Waldmann, *et al.*, 2008), presumably because of the computational complexity of evaluating larger structures.

Figure 2 shows a high-level overview of the LPL algorithm. The key pieces to be explained are the Causal Strength Estimates, and how the Decision Procedure changes the Causal Structure Hypothesis.

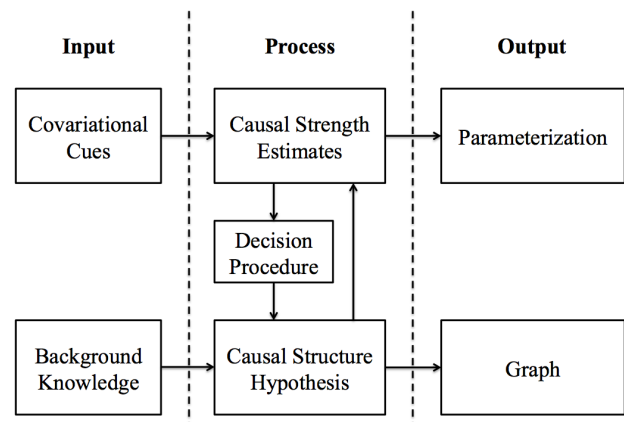


Figure 2: A high-level description of the LPL model

<sup>1</sup> The model can also encode *a priori* belief of definite edge absence, though we omit this complication for reasons of space.



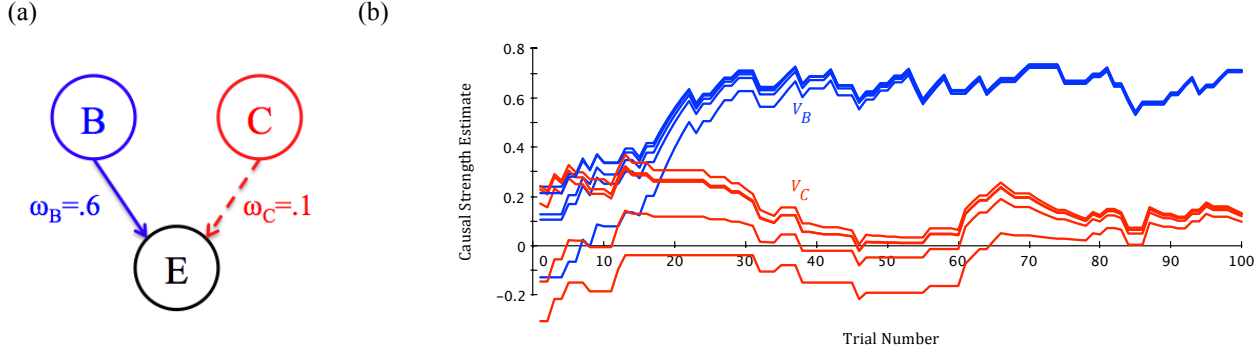


Figure 3: (a) Example local learning context (solid / dashed arrows indicate known / potential causal relations); (b) Causal strength estimates of  $B$  (blue) and  $C$  (red) using five particles per edge

### Causal Strength Estimates

The LPL model generates causal strength estimates for each possible cause-effect pair that is not ruled out *a priori*. That is, for each pair of variables ( $A$ ,  $B$ ), the learner estimates the causal strength of  $A \rightarrow B$  and  $B \rightarrow A$  unless she has prior knowledge about potential edge direction. We assume here that the correct functional form for causal relations is noisy-OR, and so causal strengths are causal powers (Cheng, 1997; Griffiths & Tenenbaum, 2005), though this can change based on background information.

Like the single-effect learning model, we assume that the appropriate scope for learning causal strength includes the potential cause, the effect, and any other definite causes of that effect. For instance, consider Figure 3(a), where the learner believes that  $B$  causes  $E$  and is trying to determine whether  $C$  also causes  $E$ . Unlike the single-effect learning model, however, causal strength estimates for  $C$  are generated by a mechanism similar to particle filters in approximate Bayesian inference, though our “particles” move by associationist learning.

The LPL model initially draws  $n$  particles for each possible causal relation from a prior strength distribution determined by the learner’s background knowledge. The learner’s current beliefs about whether  $C$  causes  $E$  are represented by these particles  $\{V_C^1, \dots, V_C^n\}$ . There is a corresponding set  $\{V_B^1, \dots, V_B^n\}$  of particles for  $B$ . The use of multiple particles enables the model to capture both strength estimates and confidence in those estimates. The mean particle value,  $\bar{V}_C = \frac{1}{n} \sum_{i=1}^n V_C^i$ , is the point estimate of  $C$ ’s causal strength. The average squared deviation of the particles,  $D_C = \frac{1}{n} \sum_{i=1}^n (V_C^i - \bar{V}_C)^2$ , is the learner’s (lack of) confidence: low values of  $D_C$  indicate high confidence.

We define a *layer*  $i$  of particles for an effect  $E$  as the  $i$ -th particle from each known and potential cause of  $E$ . In Figure 3(a), for example, layer  $i$  would be  $\{V_B^i, V_C^i\}$ . A layer of particles is a specific hypothesis about the strengths of all known and potential causes of  $E$ . Each layer is updated

independently after each data point by prediction-error learning. Such learning can be represented schematically as:

$$V_C^{i,t+1} = V_C^{i,t} + \alpha (\text{observed} - \text{expected})$$

The learning rate ( $\alpha$ ) is a free parameter, and *observed* has the value 1 if the effect occurs and 0 if it does not. The value of *expected* is typically the expected value of the effect variable, calculated using the functional form for the cause-effect relation. Many associationist learning models fit this schema, including the classic Rescorla-Wagner model and the causal power estimator of Danks, *et al.* (2003).

The *expected* value is computed separately for each potential cause in a layer. The current structure hypothesis has an influence because *expected* is based on only definite, known causes and the particular target potential cause for that update; other variables are ignored. This restriction reduces the computational demands on the learner, and fits real-world contexts where the learner cannot simultaneously attend to all the potential causes in her environment. If the causes combine as causal powers, then the expected value of  $E$  (for layer  $i$  and potential cause  $C$ ) is:

$$\text{expected} = \prod_{K=\text{present}} (1 - V_K^{i,t-1}) \left( 1 - \prod_{J=\text{present}} (1 - V_J^{i,t-1}) \right)$$

where  $J(K)$  is the set of  $E$ ’s generative (preventive) causes.

Figure 3(b) shows how initial causal strength estimates can change over time. Data were generated by Figure 3(a) with a noisy-OR functional form. At first, the particles are spread widely around zero, representing the learner’s uncertainty in her estimate. As the learner observes more data points, prediction-error learning brings the particles closer to the true parameter values. The layers of particles that are further from the true values will generally have greater errors and thus will shift more towards the true values during learning. As a result, the estimates in different layers converge,<sup>2</sup> representing the learner’s increasing confidence. This process gives no account of structure learning, however, so we turn to that now.

<sup>2</sup> Though they only stabilize around equilibrium values. If the learning rate is based on the learner’s current (lack of) confidence, then true convergence is possible.

## Causal Structure Judgments

The LPL model has a single, definite structure hypothesis at each point in time, which can then be modified by either adding or removing an edge. These modifications are based on a decision procedure applied to the causal strength estimates after each update.

Since an edge with a causal strength of zero is equivalent to no edge, the decision procedure uses a t-test on each set of particles with the null hypothesis that the particles are drawn from a distribution with mean  $\mu = 0$ . The outcome of this test depends on both the particles' mean and deviation. A free parameter  $p_{critical}$  guides the decision procedure. If there is no edge in the graph and the t-test rejects the null hypothesis (i.e., the  $p$ -value  $p$  of the test statistic is less than  $p_{critical}$ ), then an edge is added. If there is an edge present and the t-test does not reach significance (i.e.,  $p > p_{critical}$ ), then the edge is removed from the graph.

If a  $C \rightarrow E$  edge is added or removed, future calculations of *expected* change for *other* potential causes of  $E$ , as those involve only the known causes of  $E$ . Crucially, this form of causal structure learning satisfies: the learner accepts the most plausible structure as a working hypothesis rather than representing and evaluating all possible structure hypotheses (as in standard Bayesian models).

## Other Factors

Temporal information and the data source can influence the interpretation of covariational data, and so are also incorporated into the LPL model.

**Interventions** Given an observation about  $C$  and  $E$ , the LPL model updates the causal strength estimates for both  $C \rightarrow E$  and  $E \rightarrow C$  whenever the model does not yet know which direction the causal influence flows (if any). If  $C$ 's value is instead set by intervention, then one knows that  $C$  is severed from its normal causes. Thus, one should not update causal strength estimates for potential causes of  $C$ . Operationally, if given data about an intervention on  $C$ , the LPL model updates only the  $C \rightarrow E$  particles, and not the  $E \rightarrow C$  ones.

**Temporal Information** Temporal delays between the cause and effect influence contingency learning, though mediated by the learner's expectations (Buehner & May, 2003; Buehner & McGregor, 2006). The LPL model compares the observed temporal difference  $d_{E-C}$  between a potential cause and the effect to the expected temporal difference  $d_{typ}$ . If the learner expects the delay to always be  $d_{typ}$ , then the causal strength estimates update only when that delay occurs. If the learner expects the timeframe of the causal mechanism to be noisy, then the model reduces the salience of  $C$ —captured in the learning rate  $\alpha$ —as a potential cause of  $E$  in proportion to  $d_{err} = d_{E-C} - d_{typ}$ . We define a learning rate  $\alpha'$

that decreases exponentially as  $d_{err}$  increases:  $\alpha' = \alpha e^{-\frac{d_{err}}{s}}$ , where  $s$  is a scaling parameter that determines how sharply  $\alpha'$  drops off as  $d_{err}$  increases.

## Evaluating the LPL Model

### Data

We evaluate the LPL model using data from Lagnado & Sloman (2006). In this experiment, participants had to discover the causal connections between four computers by sending 100 text messages to computer A and observing whether those messages were sent on to other computers. The true causal system is shown in Figure 4, where the arrows represent noisy causal relations. Messages always reached computer A, and the probability of a message being transmitted from one computer to the next was 0.8. Messages never spontaneously occurred.<sup>3</sup> Trial order was randomized both for participants and for modeling.

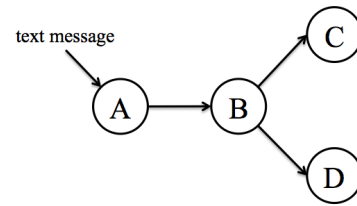


Figure 4: Causal structure from Lagnado & Sloman (2006)

The original experiment contrasted temporal and covariational information, so there were four conditions that varied the temporal order in which messages appeared. Condition 1 involved no timing information, but conditions 2-4 did (with different delays<sup>4</sup>).

### LPL Model

Participants had no prior knowledge of causal structure, so the initial model was the empty graph (i.e., agnosticism). Connections between the computers were clearly generative, so the model only considered causal strength estimates between 0 and 1. For each possible edge, five particles were drawn from a truncated Gaussian ( $\mu = 0$ ,  $\sigma^2 = .2$ ).

The LPL Model has four other free parameters. The expected temporal delay  $d_{typ}$ , and the temporal scaling parameter  $s$  are not used with simultaneous occurrences (as in condition 1). We thus first determined the values for the learning rate  $\alpha$  and the critical significance level  $p_{critical}$  by maximizing model fit (via a grid search) for condition 1 only. Model fit was based on  $R^2$  values<sup>5</sup> for the proportions, over all possible causal relations  $CR$ , of (a) 1000 model runs that yielded  $CR$ , and (b) experimental participants that

<sup>3</sup> The resulting case distribution ( $N = 100$ ) was: 51 cases with ABCD; 13 AB-CD; 13 ABC-D; 3 AB-C-D; and 20 A-B-C-D.

<sup>4</sup> The messages always appeared in the same order within conditions: A-B-D-C in Condition 2, A-D-C-B in Condition 3, and A-B-CD (C and D simultaneous) in Condition 4.

<sup>5</sup>  $R^2 = 1 - (SS_{err} / SS_{tot})$ , where  $SS_{err}$  and  $SS_{tot}$  are the sum of squared differences between the participant endorsement frequencies and the model proportion or mean endorsement, respectively.

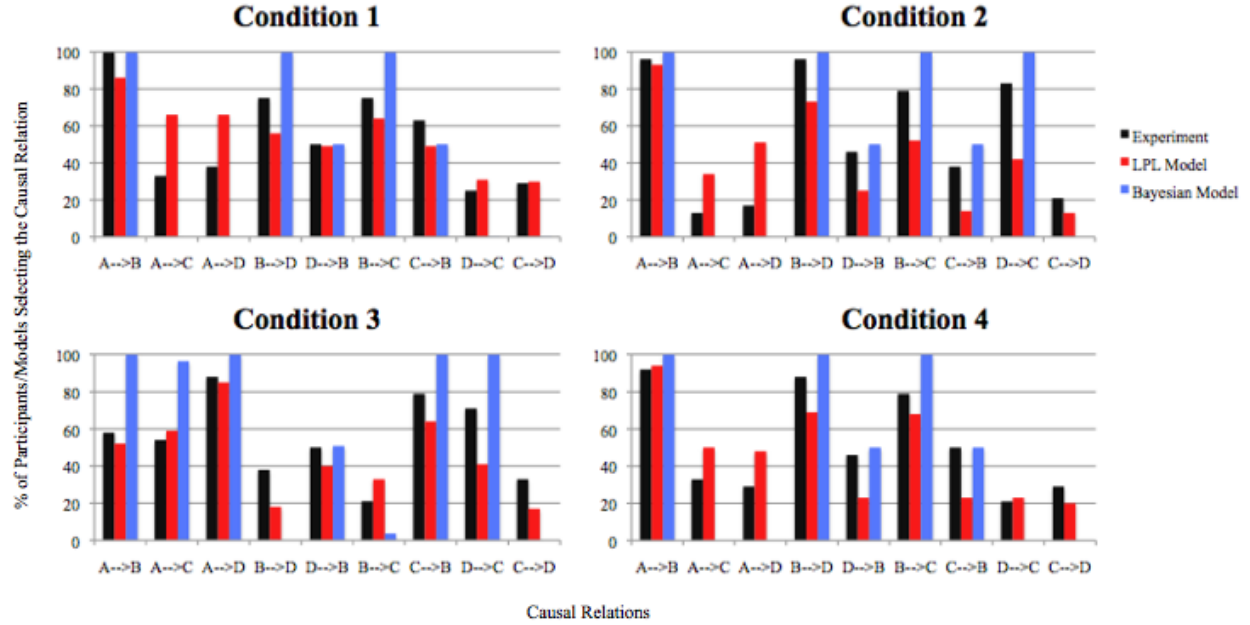


Figure 5: Proportion of causal relation endorsements by the LPL model, Bayesian model, and experimental participants

endorsed  $CR$ . The optimal model fit ( $R^2 = .47$ ) was with  $\alpha = 0.1$  and  $p_{critical} = 7 \times 10^{-5}$ .

These parameter values were used for all subsequent simulations. Model results for conditions 2-4 thus provide cross-validation for those parameter values. We set  $d_{typ} = 1$ , as the natural temporal delay between a computer sending a text message and one receiving it would be one time-step. We then searched and found that  $s = 7$  optimized model fit across conditions 2-4 ( $R^2 = .47$ ).

### Bayesian Model

We compare the LPL model to a standard Bayesian model of causal structure learning. The model used a uniform prior over all possible graphs (cyclic and acyclic) over the four variables. The posterior probability of a graph  $H$  given the  $j$ -th datapoint is:

$$P(H_i | d_j) = \frac{P(d_j | H_i, do(A)) P(H_i)}{P(d_j)}$$

If  $t_V$  denotes the time of  $V$ , then the likelihood is given by:

$$\begin{aligned} P(d_j | H_i, do(A)) &= P(b, c, d, t_B, t_C, t_D | H_i, do(A)) \\ &= P(b, c, d | H_i, do(A)) P(t_B, t_C, t_D | b, c, d, H_i, do(A)) \end{aligned}$$

Participants were told the true parameterization, so we use that distribution to calculate  $P(b, c, d | H_i, do(A))$ . For temporal sequences, the Bayesian model also assumed that

delay probabilities followed an exponential decay function:

$$P(d_{E-C}) = \frac{1}{2s} e^{-\frac{|d_{err}|}{s}} \quad .6$$

This adjustment introduces a new free parameter,  $s$ , that was estimated by maximizing model fit across conditions 2-4 ( $s = 2$ ,  $R^2 = .23$ ). To determine Bayesian model predictions, we assumed that people probability match: the proportion of “Bayesian endorsements” for each causal relation  $CR$  was simply the posterior probability of  $CR$ .

### Results and Discussion

Figure 5 shows the LPL and Bayesian model predictions, as well as the actual participant data.  $R^2$  values for the models for each condition are shown in Table 1.

Table 1:  $R^2$  values for the models

	LPL Model	Bayesian Model
Condition 1	.47	-.03 <sup>7</sup>
Condition 2	.40	.81
Condition 3	.46	-1.01
Condition 4	.59	.36
Overall	.47	.23

The LPL model explains roughly half the variance in participant responses across all conditions, whereas the Bayesian model fit varies widely. Moreover, the Bayesian model does much worse than the LPL model in Condition 1

<sup>6</sup> The probability of a temporal sequence is complicated for cyclic graphs, as one must consider multiple ways to generate a temporal sequence. Technical details are available upon request.

<sup>7</sup> If  $R^2 < 0$  then the mean predicts more variance than the model.

(i.e., with no temporal information), suggesting that the modification of the Bayesian model to allow for temporal delays does not explain the poor fit.

At the same time, both models provide good qualitative fits to the data: the model-participant correlations are  $r = .74$  for the LPL model and  $r = .97$  for the Bayesian model. However, only the LPL model predicts the appropriate variability in the participants' responses. For instance, the data are sufficient in Condition 1 for a Bayesian learner to determine the true causal structure (except for  $D \rightarrow B$  and  $C \rightarrow B$ , about which it is indifferent), and so even probability matchers should exhibit relatively little variation. However, many experimental participants select causal relations that are not part of the true structure, and some omit relations that are. Participants do not seem to be fully rational learners, and the LPL model is able to explain the types of errors that occur.

## Conclusion

The LPL model aims to provide a formal algorithmic model of the mechanisms underlying covariation-based causal structure learning. It provides a computationally well-specified dynamical model that learns directed graphs, and so potentially captures the cognitive mechanisms underlying causal learning. Moreover, this model predicts some of the sub-optimal learning behaviour exhibited by participants. Open questions remain about, for example, the suitability of the t-test-based decision procedure. But the LPL model provides a model that bridges the gap between associationist and rational models of causal learning.

## Acknowledgments

This research was partially supported by a Scholar Award from the James S. McDonnell Foundation.

## References

- Bonawitz, E., Denison, S., Chen, A., Gopnik, G., & Griffiths, T. L. (2011). A simple sequential algorithm for approximating Bayesian inference. *Proceedings of the Thirty-third Cognitive Science Society*.
- Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *The Quarterly Journal of Experimental Psychology: Section A*, 56(5), 865-890.
- Buehner, M. J., & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning*, 12(4), 353-378.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological review*, 104(2), 367-405.
- Corlett, P. R., Aitken, M. R. F., Dickinson, A., Shanks, D. R., Honey, G. D., Honey, R. A. E. (2004). Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. *Neuron*, 44(5), 877-888.
- Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 67-74). Cambridge, MA: MIT Press.
- Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3), 678.
- Glymour, C. (2003). Learning, prediction and causal Bayes nets. *Trends in Cognitive Sciences*, 7(1), 43-48.
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111(1), 1-31.
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51(4), 334-384.
- Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(3), 451-460.
- Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. In A. Gopnik, & L. Schultz (Eds.), *Causal learning: Psychology, philosophy, and computation*, 154-172.
- Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. New York: Oxford University Press.
- Sloman, S., & Lagnado, D. A. (2005). Do we 'do'. *Cognitive Science*, 29, 5-39.
- Spirtes, P., Glymour, C. N., & Scheines, R. (1993). *Causation, prediction, and search*. Cambridge, MA: The MIT Press.
- Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, 27(3), 453-489.
- Turner, D. C., Aitken, M. R. F., Shanks, D. R., Sahakian, B. J., Robbins, T. W., & Schwarzbauer, C. (2004). The role of the lateral frontal cortex in causal associative learning: Exploring preventative and super-learning. *Cerebral Cortex*, 14(8), 872-880.
- Waldmann, M. R. (1996). Knowledge-based causal induction. *Psychology of Learning and Motivation*, 34, 47-88.
- Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: A minimal rational model. In N. Chater, & M. Oaksford, *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. Oxford: Oxford University Press.
- Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2), 216-227.

# Understanding each other: Defining a conceptual space for cognitive modeling

Robert West (robert\_west@carleton.ca) &

David Pierre Leibovitz (dpleibovitz@ieee.org)

Institute of Cognitive Science, 2201 DT, Carleton University  
1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6 Canada

## Abstract

Cognitive modeling is a complex endeavor so it is not surprising that the goals and intentions of modelers are often misunderstood, even by other modelers. To try to clarify this we have attempted to map out the various philosophical and theoretical commitments that one makes when creating a cognitive model or architecture. The goal of this is to avoid misunderstandings between the adherents of different modeling systems and between cognitive modelers and the rest of the scientific community.

**Keywords:** cognitive modeling.

## Introduction

In the 1990s there was movement to contrast mainstream cognitive modeling, which was labeled as *cognitivist*, with alternative approaches which were asserted to represent a fundamentally different paradigm. These alternatives included *situated* cognition, *distributed* cognition, *dynamicism*, *embodied* cognition and *subsumption* architectures. However, Vera and Simon (1993) argued that these theories represented progress and innovation but not an alternative approach. They did this by arguing that critics of the mainstream view had mistakenly assumed that the practices, strategies, and short cuts of mainstream modelers represented their actual philosophical and theoretical commitments. We argue that Vera & Simon's argument was a legitimate response and more generally that philosophical and theoretical commitments cannot be determined solely by analyzing systems and practices associated with a method of modeling. Full understanding requires explicating the philosophical and theoretical commitments of the modeler.

Today all of the alternatives to mainstream modeling discussed in Vera and Simon (1993) are accepted in the main stream. That is to say, they are broadly recognized as important contributions. At this time it would be fair to say that there is no widely accepted, "main stream" approach to modeling. However, despite efforts to understand different modeling systems as alternative approaches, each with their own strengths (e.g., McClelland, 2009) there are still numerous attempts to vilify modeling systems by critics who do not fully understand the goals and intentions of the system creators and users. In our view, the idea that a modeling system can be dismissed based on a principle or a philosophical argument is, in fact, a philosophical mistake.

We argue that pigeon holing modeling systems according to broad philosophical and theoretical distinctions (e.g., *cognitivist* vs. *anti-cognitivist*, *computational* vs. *non-computational*, *representational* vs. *anti-representational*,

etc.) is misleading and counterproductive. In place of this, we advocate a multidimensional approach to characterize modeling systems along numerous dimensions, including the beliefs and motivations of the modeler. Thus in our system, two modelers can use the same computational code but actually have very little in common. Likewise two modelers could use very different codes (e.g., the ACT-R symbolic/subsymbolic code, J. R. Anderson & Lebiere, 1998; and the NENGO spiking neuron code, Eliasmith & Anderson, 2003) and still be completely on the same page. To illustrate this approach we will use the controversial example of the "symbol" throughout the paper, although each dimension can be applied to any modeling construct. Due to limited space we have focused on dimensions that we believe are important.

## AI and Useful fictions

The strong AI hypothesis says that if the functions of the human mind can be correctly simulated on a computer then there will be no difference between the human mind and the computer mind. It is important to note that this hypothesis is silent about the level of abstraction or embodiment required for success. It could involve high level algorithms realized as software, or it could involve a brain made of highly realistic mechanical neurons embodied in a lifelike humanoid robot and raised as a human infant. Therefore, if we apply strong AI to symbols it means that the symbols in a model are a valid way of representing what is taking place in the brain. Alternatively, symbols can be viewed as *useful fictions*. That is, the brain does not process symbols but there is something about the processing of symbols that is analogous to how a brain works and it is therefore useful or expedient to model it in this way.

## Metaphysics

In mainstream western philosophy there are substances and processes that act on the substances. For example, logical formalisms can be used to act on symbolic representations about the state of the world. We will refer to this as substance philosophy. In process philosophy (see Bickhard, 2010; Whitehead & Griffin, 1931) there are no modular substances, only interacting processes. What appear to be substances are temporary emergent properties of ongoing processes. According to Quantum Physics, process philosophy is true for physical objects. For example, the chair across the room is a temporary quantum process not an independent object. Likewise, when you close your eyes and

remember the chair, that memory is a temporary neural process not an object.

Often psychological and linguistic constructs are discussed as if they were actual objects, which leads to confusion because it could mean they are meant as objects (i.e., substance philosophy) or that they are meant as a simplification standing in for a process (i.e., process philosophy). For example, the use of symbols in a model could signify that symbols exist, or it could signify that there is a process that acts as though symbols exist. More generally, a process philosophy view implies that psychological or linguistic constructs in a model should be regarded as abstract proxies standing in for interacting processes – not informationally encapsulated modules as suggested by Fodor (1983). The issue then becomes the relative stability of the processes underlying psychological constructs. Process philosophy is silent on this – neural processes giving rise to psychological constructs could be very stable, resulting in a relatively crisp, well defined constructs; or they could be noisier, resulting in fuzzy and possibly temporary constructs. Determining this, in our opinion, is not a philosophical issue. However, deciding if constructs, such as symbols, actually exist is a philosophical issue.

### Divide and conquer versus unification

The simplicity principle (Chater & Vitányi, 2003) refers to the idea that cognitive phenomena are best modeled in the simplest way. This goal, related to Occam's Razor, is not contentious but becomes less clear when the scope of the phenomena is considered. Newell, in his famous (1973) paper, argued that the phenomena to be explained is the whole brain. Newell (1990) distinguished between micro models (i.e., independent models of different phenomena) and architectural models (i.e., models constrained by the use of a cognitive architecture aimed at describing the whole brain). The goal with micro models is to make them as simple as possible but the goal for models built in an architecture is more complex. The model should be as simple as possible given the constraints of the architecture but the actual goal is to produce an architecture that is as simple as possible across numerous models of different phenomena (including neural phenomena, (J. R. Anderson, 2007a)).

Therefore, from an architectural point of view, *having lots of incommensurate micro models is not useful*, regardless of how simple they are individually. One way around this is to argue that the micro model represents a distinct cognitive/neural module that encapsulates and operates on a particular kind of information (Fodor, 1983). For example, by arguing that there is a distinct symbol based language module, one can ignore issues or problems concerning the viability of using symbols to model other functions of the brain. Therefore, the form of a particular model could reflect the goal of creating a unified architecture, of understanding a distinct module, or of creating the simplest model for a specific phenomenon.

### Reverse and forward Engineering

Reverse engineering involves testing a system, the brain in this case, and working backwards to discover how it functions. Unfortunately, having a model that performs similarly to a human does not confirm that it is a valid model because other future tests may disconfirm it.

However, modeling can also move forward through forward engineering. Forward engineering involves designing a system with the goal of achieving certain functions. Therefore, the goal is to achieve the same functionality that humans have without worrying about doing it in the same way as the brain does. If successful the result would be a system that is roughly isomorphic to how the brain behaves, but does not give insight into how the brain does it. For example, the use of symbols in a model could be due to the belief that symbols behave isomorphically to neural representations.

The difference between backward and forward engineering is also important when evaluating how a model has been evaluated. Generally speaking, attempts to reverse engineer involve careful comparisons to experimental data while attempts to forward engineer involve showing that a model can produce certain functionality. It is important to note that a modeler may also iterate between reverse and forward engineering.

### Epistemic commitments

Epistemic commitment refers to the mechanisms used to build the model. Specifically, we mean it to refer to a commitment to a particular way of understanding and modeling the brain. The debate between proponents of symbol systems and proponents of neural networks is an example of an epistemic debate. The idea motivating such debates is that it is necessary to first get the way of modeling right, otherwise the resulting models will be misleading and ultimately dead ends. This issue has fueled a lot of debate within cognitive science. Examples of different systems are: *symbol systems, neural networks, holographic systems, dynamic systems, spiking neuron models, Bayesian networks, logical systems, grammatical systems* etc. However, it is also possible to view these as tools rather than competing theories, in which case the choice of a particular way of modeling would reflect a pragmatic choice rather than a principled one. Another approach is to view different modeling systems as different lenses for viewing a phenomena in different ways (McClelland, 2009).

Related to this, it is important to note that the word *architecture* is used in two ways. As noted above, it can refer to a system meant to be a unified model of the whole brain, or it can refer to mechanisms for building models that are able to model the whole brain. For example, ACT-R (J. R. Anderson, 1993) is an attempt at creating a unified architecture but it is meant to be a hybrid system and therefore does not embody an epistemic commitment. ACT-R is often described as a production system but this is incorrect as ACT-R has numerous modules that use numerous mechanisms. The use of a production system



module to coordinate the other modules in ACT-R is a commitment to a theory about unification; it is not an epistemic commitment (i.e., a way of understanding the whole brain). In contrast, NENGO (Eliasmith & Anderson, 2003) is a system for building spiking neuron models according to a specific theory about spiking neurons, so the use of NENGO can be seen as an epistemic commitment (i.e., that the whole brain can be modeled in this way).

### **Ontological commitments**

Ontological commitment refers to the way a model is divided into functional parts and their connectivity. There are two reasons why ontological commitments are important. The first has to do with creating unified cognitive architectures. Simply put, a valid cognitive model of the whole brain requires that the functional parts of the model map onto the functional parts of the brain. Although the results of experimental psychology are good for testing models, they may be misleading in terms of telling us what the parts are because their ontologies are defined primarily to make experimentation possible on different psychological phenomena. Unfortunately this does not necessarily tell us what the actual parts are. For example, we remember facts and we remember episodes and these can be treated separately for experimental purposes, but we still do not know if we have separate semantic and declarative memory systems or if they are both products of a single long term memory system.

Another very important issue related to ontologies is cognitive re-use (see M. L. Anderson, 2010). This refers to whether or not our cognitive ontology corresponds to a dedicated neural area. Much of the neural localization work taking place today explicitly or implicitly assumes that it does. However, the cognitive re-use hypothesis is that higher-level cognitive mechanisms and functions can be created by re-using and recombining lower level cognitive mechanisms and functions. If this is true then there are two important consequences: (1) specific brain areas are not dedicated to specific cognitive functions, and (2), the ontology that we should be looking for is at a lower level.

Therefore, modelers may believe that the modules of their system correspond to actual cognitive functions in the brain, and they may further believe that these functions map to dedicated areas of the brain. But having a module in a model does not necessarily mean that they believe either of these things. For example, following the cognitive re-use hypothesis, a module could also represent a function that is created through the interaction of lower level functions under specific conditions. Symbols, or any other construct, can be thought of in either way.

### **System levels**

Allan Newell (1980) proposed that the brain is constructed in the way that computers are engineered, according to system levels. The reason why natural systems would develop distinct hierarchical levels was developed by Simon (1962) but a discussion of this is beyond the scope of this

paper. A system level occurs when the behavior of a complex lower level system can be understood in terms of less complex higher level constructs. For example, in the theory of thermodynamics, the complex interactions of atomic particles can be understood through higher level concepts such as heat and pressure. So a systems level is a real thing (in as much as heat and pressure are real things) but it is important to note that a system level can be weak or strong depending on the relative reduction in complexity produced by the emergent level. A weak system level is leaky, meaning that it is sometimes affected by system levels below it (e.g., Saunders, Kolen, & Pollack, 1994).

The cognitive level is theorized to exist as a systems level above the neural level but there is considerable controversy over whether it exists and if it does, what form does it take? The symbol system hypothesis asserted that the cognitive level is based on processing symbols. Likewise, Chomsky (e.g., Chomsky, 1995) argued that for understanding language, symbols could be divorced from the underlying system that produces them. However, it is instructive to look at exactly what was meant by, "symbol." For Chomsky a symbol is a word, but Newell defined a symbol in terms of distal access (Newell, 1990). Distal access refers to using information that is not local, i.e., information that is transported from another part of the brain. The form of the information or the way it is transferred is not important, therefore Newell's commitment to symbols is completely different from Chomsky's commitment to symbols.

### **Level of Analysis**

Level of analysis is different from system level. Level of analysis refers to analyzing a system at a particular level (e.g., neural, neural groups, networks, symbols). Using a level of analysis may or may not indicate a belief that the level is a systems level. So the use of symbols in a model may indicate a commitment to the symbol system hypothesis, but it could also occur because that level of analysis is useful, without any commitment to the existence of an actual systems level. Also, it is possible to test a model constructed at a higher systems level using a lower level of analysis if there is a theory about how the lower level is related to the higher level. For example, ACT-R models can be tested using a neural level of analysis with an fMRI scan (J. R. Anderson, 2007b). Choice of a level of analysis reflects beliefs about the most effective way of testing a model.

### **Consciousness**

Explaining consciousness is a special case of the strong AI issue that deserves its own section. The question is, could a properly constructed cognitive architecture actually have conscious experiences. From a strong AI point of view the answer is yes, but many people reject this position because they find it hard to imagine. This seems to be due mainly to our subjective experience of qualia.

Qualia refer to the various phenomenal feelings of our conscious experiences. From a modeling point of view



qualia creates a potential problem because thought, emotion, and different types of perception do not feel the same to us; they feel qualitatively different. However, from a cognitive science perspective, and a neuroscience perspective, all qualia arise from information processing that is ultimately realized through the firing of neurons. Since we do not understand what consciousness is or how it creates different qualia from the same underlying mechanism, most cognitive models simply ignore the issue or focus on the correlates of consciousness (e.g., awareness, wakefulness, report ability, etc.).

However, the concept of qualia is important for modeling because it cuts across the board and separates the issue of how information is processed from how it is subjectively experienced. By setting aside the issue of qualia we are implicitly adopting the view that qualia is an epiphenomena; that is, we can model the brain without considering qualia because qualia has no functional significance (Dennett, 1991). This is very convenient since it allows us to model all aspects of the brain as information processing and ignore or put off the problem of explaining why different brain functions feel qualitatively different from each other.

Alternatives to understanding consciousness as an epiphenomena arising from information processing are scarce. Searle (1980) makes his arguments against strong AI by arguing that it leads to absurd consequences or conclusions (the Chinese room is his most famous example), but he does not offer an alternative explanation. Hameroff & Penrose (1996) argue that normal information processing is inadequate to model human cognition and consciousness. They propose that the brain is capable of quantum computing and therefore a valid simulation would require a quantum computer. Although this view is not popular it should be noted as quantum computing is so far the only scientific alternative to normal computing, although, as Penrose concedes, it is still a type of information processing. Chalmers (2010) has argued that if you reject that consciousness arises from information processing, the only option is to adopt some form of dualism.

## Philosophy of science

Some people define science with Popper's (1935) notion of falsifiability. However, although it is in theory possible to falsify cognitive models, it is often the case that the failure of a model leads to changes in the model rather than a rejection of the model. With unified architectures the problem of falsification is trickier because in order to test the architecture, it must be used to build a model of a task, therefore, if it fails, it is unclear if the architecture has been falsified or just the model. Newell (1990) realized this and argued that Lakatos' definition of science (1970) was more appropriate than Popper's for understanding architectures. Essentially, Lakatos defines science in terms of making progress over time, therefore if an architecture or model is improved through testing and refinement so that it explains

more, it can be considered scientific (for a detailed discussion see Cooper, 2007).

It is interesting to note that although some of the criticism directed at testing models comes from Experimental Psychology, Experimental Psychology also fails to follow Popper's model. Specifically, most experiments in Experimental Psychology test for significant differences predicted by a theory, therefore, falsifying the theory would mean showing no significant difference, which would mean accepting the null hypothesis, which is not allowed in the ANOVA or t-test statistics that are generally used. Like modeling, theories in Experimental Psychology are generally altered and not rejected. In both cases it is possible to construe theories that are falsifiable; it is just not very common. The criticism of modeling coming from Experimental Psychology has more to do with statistics. Specifically, Experimental Psychology has a clear definition for defining when two conditions are significantly different. In contrast, the goal for a model is to show that it is significantly similar to a set of data and there is not an agreed upon standard for this (e.g., Roberts & Pashler, 2000), although there are statistical ways to tackle the issue (e.g., Stewart & West, 2010).

Another issue arises from comparisons with Computer Science or Engineering where it is common to evaluate algorithms against each other according to some clear criterion or test set. According to this approach, cognitive models should be compared to see which one explains the data best. This can be done when the models are specifically designed to model the same problem (see Erev et al., 2010 for an example). However, it is not commonly done because models are designed with different goals in mind, therefore a good test set for one might be an inappropriate or poor test set for another. It all depends on the goals and the theoretical framework of the modeler, which is why it is important to be clear about these.

## Conclusion

We have outlined a number of dimensions on which modelers can take different views. Most of them are binary so it is possible to say agree, disagree, or agnostic. This list is not exhaustive, but being aware of where we stand on these issues can potentially avoid a lot of misunderstanding and provide a richer view of the whole modeling enterprise.

We have tried to be neutral in terms of laying out this list but we acknowledge that some people may feel that some of the choices we have presented are invalid. For, example, one could argue that there is no such thing as system levels in the brain. Our point is that we should separate that argument from the evaluation of modeling systems that appear to embody systems levels.

Another issue is the relationship between the different dimensions that we have laid out. People tend to associate sets of beliefs with the use of different modeling systems. Possibly some of the dimensions we described are correlated and logically go together. However, arguments about whether certain dimensions are conceptually related

or conceptually independent should be separated from subjective impressions concerning the co-occurrence of dimensions across the users of different modeling systems.

In order to progress in understanding the various cognitive modeling spaces, the impacts of these and other dimensions need to be further deliberated.

## References

- Anderson, J. R. (1993). *Rules of the Mind* (p. 336). Lawrence Erlbaum.
- Anderson, J. R. (2007a). Using Brain Imaging to Guide the Development of a Cognitive Architecture. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 49-62). New York, NY: Oxford University Press.
- Anderson, J. R. (2007b). *How Can the Human Mind Occur in the Physical Universe?* *Science* (p. x-290). Oxford University Press.
- Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought* (p. 504). NJ: Erlbaum.
- Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, 33(4), 245-66.
- Bickhard, M. H. (2010). Does Process Matter? An Introduction to the Special Issue on Interactivism. *Axiomathes*, 21(1), 1-2.
- Chalmers, D. J. (2010). *The Character of Consciousness. Consciousness and Cognition* (p. xxvii-596). Oxford University Press.
- Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in cognitive sciences*, 7(1), 19-22.
- Chomsky, N. (1995). Language and Nature. *Mind*, 104(413), 1-61.
- Cooper, R. P. (2007). The Role of Falsification in the Development of Cognitive Architectures: Insights from a Lakatosian Analysis. *Cognitive Science*, 31(3), 509-533.
- Dennett, D. C. (1991). *Consciousness Explained* (p. 528).
- Eliasmith, C., & Anderson, C. H. (2003). *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems* (p. 376). Cambridge, MA: MIT Press.
- Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., Hertwig, R., et al. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, 23(1), 15-47.
- Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology* (p. 154).
- Hameroff, S., & Penrose, R. (1996). Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. *Mathematics and Computers in Simulation*, 40(3-4), 453-480.
- Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge* (pp. 91-196). Cambridge, UK: Cambridge University Press.
- McClelland, J. L. (2009). The Place of Modeling in Cognitive Science. *Topics in Cognitive Science*, 1(1), 11-38.
- Newell, A. (1973). You Can't Play 20 Questions with Nature and Win: Projective Comments on the Papers of this Symposium. In W. G. Chase (Ed.), *Visual Information Processing: Proceedings of the 8th Symposium on Cognition* (pp. 283-308). New York: Academic Press.
- Newell, A. (1980). Physical Symbol Systems. *Cognitive Science*, 4(2), 135-183.
- Newell, A. (1990). *Unified Theories of Cognition* (pp. 1-530). Cambridge, MA: Harvard University Press.
- Popper, K. R. (1935). *The logic of scientific discovery* (English tr.). New York, NY: Basic Books.
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107(2), 358-367.
- Saunders, G. M., Kolen, J. F., & Pollack, J. B. (1994). The Importance of Leaky Levels for Behavior-Based AI. In D. Cliff, P. Husbands, J.-A. Meyer, & S. W. Wilson (Eds.), *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior (Complex Adaptive Systems) SAB94*. MIT Press.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-457.
- Simon, H. A. (1962). The Architecture of Complexity. *Proceedings of the American Philosophical Society*, 106(6), 467-482.
- Stewart, T. C., & West, R. L. (2010). Testing for Equivalence: A Methodology for Computational Cognitive Modelling. *Journal of Artificial General Intelligence*, 2(2), 69-87.
- Vera, A. H., & Simon, H. A. (1993). Situated Action: A Symbolic Interpretation. *Cognitive Science*, 17(1), 7-48.
- Whitehead, A. N., & Griffin, D. R. (1931). Process and Reality. *Economica*, (32), 251.

# Reading direction is sufficient to account for the optimal viewing position in reading: The case of music reading

Yetta Kwailing Wong (yetta.wong@gmail.com)

Janet Hui-wen Hsiao ([jhsiao@hku.hk](mailto:jhsiao@hku.hk))

Department of Psychology, University of Hong Kong  
604, Knowles Building, Pokfulam Road, Hong Kong

## Abstract

The Optimal viewing position (OVP), the position where word recognition is the best, is biased to the left for English words. Several explanations have been proposed to account for this phenomenon, including the left hemispheric dominance for language, asymmetric information structure of words, and reading direction. However, it is unclear which factor(s) is necessary or sufficient to cause an asymmetric OVP. Using music reading, which shares only the reading direction but not the other two factors with word reading, we show that the OVP for three-note sequences is significantly biased to the left only for expert readers but not for novices. The degree of asymmetry in the OVP curve for music readers increases with individual reading skill, suggesting that their OVP is gradually shifted to the left during the development of reading skills. These suggest that habitual reading direction is sufficient to account for a biased OVP to the left.

**Keywords:** optimal viewing position, word reading, music, expertise, visuospatial bias

## Introduction

It has been well documented that where we look within a word or a sentence determines our reading performance. For example, we recognize English words the best when we fixate to the left of the middle of the words, i.e., the optimal viewing position (OVP) for English words is on the left (also called ‘convenient viewing position’; O’Regan, 1984; Brysbaert & Nazir, 2005). This cannot be explained by the acuity function of our eyes, which is the highest at the fovea but drops symmetrically in the left and right visual periphery (Bouma, 1970). Why is the OVP for English word asymmetric and biased to the left, but not to the right?

Multiple factors have been proposed to account for this phenomenon. The first factor is related to the cerebral hemispheric dominance for language processing (Brysbaert & Nazir, 2005). When we fixate at the left part of a word, most of the letters falls onto the right visual field, where information is initially projected to the left hemisphere. As the language center for most people is in the left hemisphere, word recognition is more efficient when we fixate at the left part of a word as compared with when we fixate at the right part of a word (where most of the word falls onto the left visual field and is initially projected to the right hemisphere). Supporting this account, individuals with

right-hemisphere-dominant language functions have a shifted OVP more towards the end of a word compared with the left-hemisphere-dominant individuals (Brysbaert, 1994; Hunter et al., 2007).

Second, the OVP for words is affected by the information structure of the words. For example, the OVP shifts to the informative position of the words in terms of word identity or meaning, both when the informative part is at the word beginning (e.g., the left part of an English word) or at the end (e.g., the right part of an English word; O’Regan et al., 1984; Deutsch & Rayner, 1999). Also, adding a prefix shifts the OVP towards the word end while a suffix shifts the OVP towards the word beginning (Farid & Grainger, 1996). Since the initial letters are in general more informative about the identity of the word than the last letters for English, the OVP for English words is on the left (Brysbaert & Nazir, 2005; Farid & Grainger, 1996).

Third, the OVP for words can be explained by reading direction. In left-to-right scripts, since the newly arriving information and the next eye movement is on the right, attention is directed more to the right visual field. With years of reading training, perceptual span for reading (the region around fixation from which useful information is extracted) extends further to the right compared with the left (Deutsch & Rayner, 1999). The OVP for English words is on the left because a left fixation leaves most of the word in the right visual field where English readers learn to recognize the word better (Brysbaert & Nazir, 2005). Prior work shows that the OVP for right-to-left scripts (e.g. Arabic) have a more symmetrical OVP (Farid & Grainger, 1996).

While many factors can modulate the OVP for words, which one(s) is necessary and/or sufficient for an asymmetric OVP to occur? In word reading, it is impossible to isolate and test the effect of each factor. Here, we tested whether reading direction alone is sufficient to lead to a left-biased OVP with the domain of music reading. While music reading shares the left-to-right reading direction with English reading, it does not involve strong hemispheric lateralization as experts learn to recruit both hemispheres for music reading (Wong & Gauthier, 2010). In addition, music notation does not follow as strict morphological/

orthographical rules as English text does. Therefore it is unlikely that music sequences in general have an asymmetric information distribution as that in English words<sup>1</sup>. Therefore, music reading allows us to test whether reading direction is sufficient to cause a left-biased OVP.

Here we used three-note sequences and single notes (i.e., the shortest note sequences) as our stimuli. A sequential matching task which did not require music knowledge was used so that we were able to measure the OVP in both experts and novices. In addition, we took advantage of the wide range of music reading ability among the participants to examine how the OVP changes with reading skills. We hypothesized that the OVP is gradually shifted to the left (for left-to-right scripts) when one's reading skill improves. To test this hypothesis, we examined the relationship between the degree of asymmetry of the OVP curve and individual music reading fluency. The hypothesis predicts that the degree of asymmetry of the OVP curve should increase with individual reading fluency.

## Methods

### Participants

Forty-two participants completed the experiment for cash payment or course credits. All participants were right-handed (according to the Edinburgh Handedness Inventory; Oldfield, 1971) except three participants (one intermediate and two novice readers) who were subsequently excluded from data analysis. Twenty-six participants had been formally trained in music reading and were further divided into the expert and intermediate group according to their performance in the perceptual fluency test (see below). The thirteen experts included 12 females and 1 male ( $M_{\text{age}} = 20.2$ ,  $s.d. = 1.69$ ) with 13.4 years of music reading experience on average (ranging from 10-20 years). The twelve intermediate readers included 11 females and 1 male ( $M_{\text{age}} = 21.8$ ,  $s.d. = 4.36$ ) with 9.3 years of experience reading music on average (ranging from 2-17 years). The thirteen novices reported that they could not read music,

<sup>1</sup> There is no consensus and no formal study (to our best knowledge) on the information structure of music sequences. However, probable combinations of sequences (e.g. melodies) are defined by specific music pieces without general morphological, orthographical or phonological structure applicable to all pieces. In this experiment, no musical context, key signatures or accidentals (e.g. sharps or flats) were provided and the sequences only varied along the most common C major scale. In this case, all combinations of the notes are highly probable such that the notes are unlikely more predictable by the left or right part of the sequences. Although some pitch pairs may be more frequent than others in general (e.g. tonal pitch pairs such as 'C' and 'E' are used more frequently compared with tritone pairs such as 'C' and 'F#'), such predictiveness of tone pairs should be largely symmetrical (e.g. 'C' is unlikely followed by 'F#', and 'F#' is also unlikely followed by 'C'). As a result, there is presumably no information structure biased to the left or right for music sequences, at least under the current context.

with 8 females and 5 males ( $M_{\text{age}} = 22.4$ ,  $s.d. = 5.42$ ) and 0.31 years of music reading experience (ranging from 0-3 years). All reported normal or corrected-to-normal vision and gave informed consent according to the guidelines of the Ethics Committee of the University of Hong Kong.

### Stimuli and Design

The experiment was conducted on PCs with the Eyelink 1000 eyetracker (SR Research Ltd, Canada), and Matlab using the Psychtoolbox and the Eyelink Toolbox extension. The eyetracker was positioned on the desk and sampled pupil location at 500 Hz. The tracking mode was pupil and corneal reflection. The standard nine-point calibration procedure was administered at the beginning of the task; the procedure was repeated whenever the drift correction error was larger than one degree of visual angle during the experiment. The acceleration threshold was 8000 degree/s<sup>2</sup> and the threshold for saccade velocity was 30 degree/s. Participants viewed the stimuli at 62 cm from the monitor using a chin rest.

The stimuli were generated with Matlab. 400 three-note sequences were randomly generated, with the constraint that there were no repeated notes within each sequence and no repeated sequences within the set. The sequences subtended about 3° x 3°. Each sequence was paired with a distractor sequence, in which one of the notes was shifted for one step up or down (counterbalanced). Single notes included 11 quarter notes from the note below the bottom staff line (D4) to the note above the top line (G5). They subtended about 1.6° x 3.2° in visual angle. The contrast of the single note stimuli was reduced to half to avoid ceiling performance. The distractor of each single note was the note either one step up or down (counterbalanced).

A sequential matching task was used (Figure 1a). Each trial started when a central fixation was confirmed by the eyetracker. Then, a target stimulus was presented (for 600 ms for sequences and 80 ms for single notes) while participants maintained a central fixation. If the eyetracker detected an eye fixation away from the center, the trial was aborted and an error message was presented to the participant. Next, a second stimulus was presented in the upper or lower visual field at 3.6° from the central fixation. Participants were instructed to saccade to this image and judged whether the two stimuli were identical or not by key press as fast and as accurately as possible.

The critical manipulation was the position of the first target stimulus such that participants fixated at different viewing positions. For sequences, the target was presented at 2° left, 1° left, 0°, 1° right, or 2° right from the central fixation such that participants' central fixation fell onto the far-right, right, center, left, or far-left part of the sequences respectively (Figure 1b). For single notes, the target was presented at 2.5° left, 0° or 2.5° right from fixation. The

dependent measure was the sensitivity ( $d'$ ) and response time (RT).

Trials with different fixation positions were randomized. For sequences, there were 400 trials with 80 trials for each fixation position. For single notes, there were 180 trials with 60 trials for each fixation position. Participants were tested with single notes before the sequences. For each type of stimulus, 20 practice trials with feedback were provided before testing (without feedback).

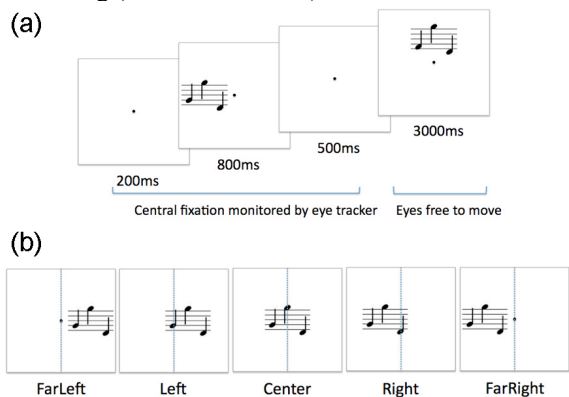


Figure 1. The sequential matching task (a) and the five fixation positions relative to the note sequences (b). Participants kept central fixation indicated by the black dot. The blue line was marked for illustration purposes and was not actually presented during the test.

### Measure of perceptual fluency

We assessed fluency in music reading with a sequential matching paradigm and used this as an indicator of individual music reading ability since it is more direct and objective compared with other measures such as years of experience and self-rated ability (Wong & Gauthier, 2010). On each trial, a central fixation was presented for 200 ms, followed by a 500 ms pre-mask, and a four-note sequence for a varied duration. After a 500 ms post-mask, two four-note sequences appeared side-by-side, one identical to the first sequence, and the other with one of the notes shifted by one step (with up/down shifts counterbalanced). The task was to select the matching sequence by key press. The presentation duration threshold for 80% accuracy was estimated four times, each with 40 trials, using the QUEST algorithm (Watson & Pelli, 1983). Sequences were randomly generated using notes ranging from the note below the bottom line (a 'D' note) to the note above the top line (a 'G' note). Contrast for all the stimuli was lowered by about 60% to avoid a ceiling effect.

To control for individual differences not specifically tied to expertise with notes, perceptual fluency for four-letter strings was measured in an identical procedure. The strings were randomly generated with 11 letters: b, d, f, g, h, j, k, p, q, t, and y. These letters were selected because they contain parts extending upward or downward, similar to musical notation. To create distractor strings, one of the four letters

was selected (counterbalanced across stimuli) and replaced by a different letter randomly drawn from the set. The string was shown at the same lowered contrast as the sequences.

## Results

One novice and one intermediate reader were excluded from data analyses because their perceptual fluency for notes was  $> 3$  s.d. away from the mean of the rest of the group. Therefore, thirteen experts, twelve intermediate and twelve novice readers were included.

### OVP for sequences

We observed a left-biased OVP for note sequences in experts, which was not found in the other two groups. A  $3 \times 5$  ANOVA with Group (Experts, Intermediates, Novices) and Fixation Position (Far Left, Left, Center, Right, Far Right) on  $d'$  revealed a significant main effect of Group,  $F(2, 34) = 10.2$ ,  $p = .0003$ , in which experts performed better than the other groups in general (LSD tests,  $p < .05$ ). A main effect of Fixation Position was significant,  $F(4, 136) = 18.8$ ,  $p \leq .0001$ , which marginally interacted with Group,  $F(8, 136) = 1.89$ ,  $p = .066$  (Figure 2a).

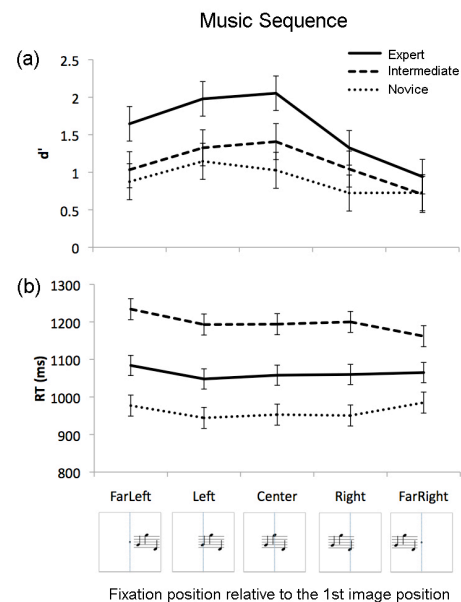


Figure 2. Matching performance for three-note sequences in  $d'$  (a) or RT (b) with different fixation positions relative to the first presented images.

To increase statistical power, we limited our analyses within the expert and novice groups, as the OVP function for intermediate readers was similar to the other two groups (Figure 2a). Results were similar to the above, except that the Group  $\times$  Fixation Position interaction reached significance,  $F(4, 92) = 3.47$ ,  $p = .011$ . Sheffé tests ( $p < .05$ ) revealed that  $d'$  was similar across positions for novices, suggesting that none of the viewing positions was 'optimal'. For experts, in contrast,  $d'$  was similar between the Far Left, Left and Center positions, while  $d'$  for the Center position

was better than the Right and Far Right positions. Importantly,  $d'$  for the Left position was better than the Right, and that for the Far Left position was better than the Far Right, suggesting the OVP for three-note sequences was biased to the left for experts.

Within the intermediate readers, we did not observe any clear pattern for the OVP function. A one-way ANOVA with Fixation Position on  $d'$  was significant,  $F(4,44) = 4.69$ ,  $p = .003$ . Sheffé tests ( $p < .05$ ) revealed that  $d'$  at the Center position was better than the Far Right but no different from the Far Left. However,  $d'$  was similar between Left and Right positions, and between Far Left and Far Right positions. Therefore we could not conclude that the OVP function for the intermediate readers was biased to either side of the sequences.

For RT, the  $3 \times 5$  ANOVA with Group (Experts, Intermediates, Novices) and Fixation Position revealed a main effect of Group,  $F(2, 34) = 7.36$ ,  $p = .002$ , in which intermediate readers responded significantly slower than the other two groups (LSD tests,  $p < .05$ ; Figure 2b). A main effect of Fixation Position was significant,  $F(4, 136) = 3.08$ ,  $p = .018$ , in which performance at the Far Left position was slower than the Left in general (Sheffé tests,  $p < .05$ ). The interaction between Group and Fixation Position did not reach significance ( $p > .2$ ). When the intermediate readers were excluded, only the main effect of Fixation Position was significant in a similar manner as the above.

### OVP for single notes

A left OVP was observed in intermediate readers but not in experts or novices. A  $3 \times 3$  ANOVA with Group (Experts, Intermediates, Novices) and Fixation Position (Left, Center, Right) on  $d'$  revealed a significant main effect of Group,  $F(2, 34) = 5.41$ ,  $p = .009$ , in which the only group difference was that experts performed better than novices in general (LSD tests,  $p < .05$ ). A main effect of Fixation Position was observed,  $F(2, 68) = 48.8$ ,  $p \leq .0001$ , in which performance was better at the Center than the Left positions and at the Left than the Right positions (LSD tests,  $p < .05$ ). The interaction between Group and Fixation Position was significant,  $F(4, 68) = 4.26$ ,  $p = .004$  (Figure 3a).

We subsequently analyzed the effect of Fixation Position for each group separately. The main effect of Fixation Position was significant in each group, all  $ps < .004$ . For experts and novices, performance was the best at the Center, while performance at the Left and the Right position was similar (LSD tests,  $p < .05$ ), suggesting that the OVP curve was largely symmetrical. However, for intermediate readers, performance at the Center was better than the Left position, which was in turn better than the Right position. In other words, we observed a left OVP with single notes only for the intermediate readers but not for experts or novices.

For RT, the main effect of Fixation Position was significant,  $F(4, 68) = 8.14$ ,  $p = .0003$ , with faster responses at the Center than the other two positions (LSD tests,  $p < .05$ ; Figure 3b). The main effect of Group and its interaction with Fixation Position was not significant ( $F_s < 1$ ).

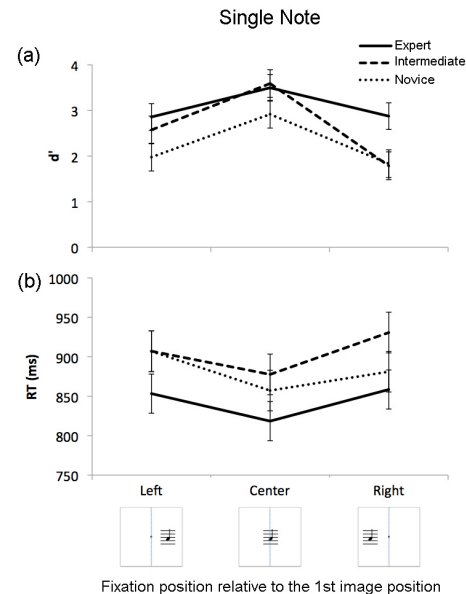


Figure 3. Matching performance for single notes in  $d'$  (a) or RT (b) with different fixation positions relative to the presented images.

### Perceptual fluency

As expected, experts had the highest perceptual fluency for notes, followed by the intermediate group and then by the novices. A one-way ANOVA for Group (Experts, Intermediates, Novices) on duration threshold for notes revealed a significant main effect of Group,  $F(1, 34) = 19.8$ ,  $p \leq .0001$ , where the performance for each group was significantly different ( $M_{Exp} = 316.5$  ms;  $M_{Int} = 680.4$  ms;  $M_{Nov} = 930.9$  ms; LSD tests,  $p < .05$ ). In contrast, duration threshold for letters was similar for all groups ( $M_{Exp} = 186.4$  ms;  $M_{Int} = 207.5$  ms;  $M_{Nov} = 233.9$  ms;  $F < 1$ ), suggesting that experts have a higher perceptual fluency for notes, which cannot be explained by a general perceptual advantage.

### Predicting the degree of asymmetric OVP with perceptual fluency with notes

Does the degree of asymmetry of the OVP curve increase with one's reading ability? We addressed this question by computing the degree of asymmetry of the OVP curve for note sequences using the measure  $d'_{Left} - d'_{Right}$  in each music reader (novices were excluded in this analysis). A significant correlation was observed between the degree of asymmetry and individual perceptual fluency,  $r = -.48$ ,  $p = .015$ ,  $df = 23$  (Figure 4a). A similar trend was observed at



far positions ( $d'_{\text{FarLeft}} - d'_{\text{FarRight}}$ ), though it did not reach significance ( $r = -.27, p = .19$ ). These suggest that the left viewing position of sequences becomes more optimal during the development of music reading skills.

For single notes, the correlation between the degree of asymmetry of the OVP curve ( $d'_{\text{Left}} - d'_{\text{Right}}$ ) and individual perceptual fluency approached significance in an opposite direction ( $r = .37, p = .066, df = 23$ ; Figure 4b). The advantage of left viewing position gradually diminished with better music reading skills.

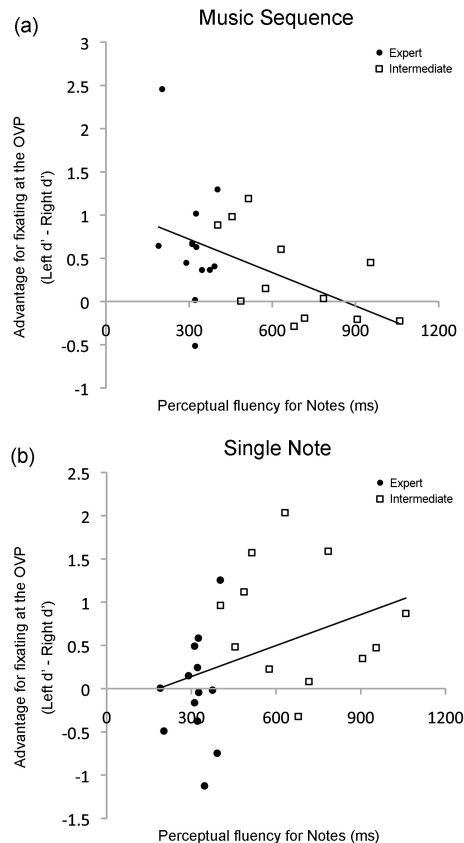


Figure 4. Scatter plots between perceptual fluency for notes and individual degree of asymmetry of the OVP for note sequences (a) and single notes (b).

## Discussion

### For three-note sequences

For three-note sequences, performance was in general the best at the center position, consistent with the highest acuity at fovea. Importantly, we observed an OVP biased to left in music reading experts but not in intermediate or novice readers. Since music reading shares a left-to-right reading direction with word reading but not the hemispheric dominance or asymmetric information distribution, our results suggest that extensive reading experience in the left-to-right reading direction is sufficient to lead to a left-biased OVP in reading.

Our results also suggest that a biased OVP is gradually developed through reading training. For novices, recognition performance is similar across viewing positions and none of the viewing positions is ‘optimal’. When music reading skills develop, the OVP is gradually shifted to the left, suggested by the correlation between the degree of asymmetry in the OVP curve and individual music reading ability. Note that our results cannot be explained by the reading habits of other languages (e.g., Chinese and English for our participants), since all of the participants had the same left-to-right reading habit, while only experts produced a left OVP for music sequences.

While our results suggest that reading direction is a major factor leading to a left-biased OVP in reading, the OVP may also be modulated by other factors, such as the left-hemispheric lateralization for language functions (Brysbaert, 1994; Hunter et al., 2007) and an asymmetric information structure of words (Deutsch & Rayner, 1999; Farid & Grainger, 1996). It is worth noting that different types of word information may become important depending on the OVP task, such as word naming, identification, lexical decision, or word matching tasks (e.g., O’Regan, 1984; Deutsch & Rayner, 1999; Nazir et al., 2004; Farid & Grainger, 1996; Stevens & Grainger, 2003). To evaluate the effect of general information structure of words on the OVP, one should consider whether any observed OVP pattern is solely determined by the characteristics of the specific sets of word stimuli, especially for the distribution of information important for the testing task. In any case, even without an asymmetric information distribution or hemispheric dominance, as in the case of isolated music sequences in the current study, a left OVP can still be observed. It suggests that these are not necessary factors leading to a biased OVP.

### For single notes

For single notes, we observed a left-biased OVP among intermediate readers but not in experts or novices, and the left bias of the OVP decreased with enhanced music reading fluency. There are multiple ways to interpret these findings. First, the performance for experts approached ceiling for all viewing positions (the mean  $d'$  was larger than 3 and the mean accuracy was larger than 90% for all viewing positions) such that potential differences across viewing positions failed to emerge. Indeed, within the experts whose average accuracy for the Left and Right positions < 90%, a left-viewing advantage emerged numerically ( $d' = 2.39$  for Left;  $d' = 2.08$  for Right;  $N = 5$ ), supporting the idea that a ceiling effect prevented a left OVP to be observed among experts. According to this explanation, the OVP for both music sequences and single notes are both biased to the left. Another possible explanation is that the asymmetric OVP effect for single notes simply becomes weaker as in the case of word reading that the asymmetric OVP effect was weaker in short words than long words; Hunter et al., 2007; Ellis,



Young, & Anderson, 1988). This may be caused by a weakened influence from reading direction on short sequences as experts are able to skip them during reading, and such tendency may become larger with better music reading skills.

It has been proposed that reading direction may partly underlie visuospatial asymmetry effects observed in the processing of some visual stimuli, such as identity or affect judgments for faces (Vaid & Singh, 1989; Brady, 2011), or bisection of straight lines (Chokron & Imbert, 1993; see also Kazandjian & Chokron, 2008). Our current results suggest that the visual field asymmetry caused by habitual word reading direction is not generalizable to all domains of object recognition. Specifically, a left-biased OVP for English words is presumably shared by all of our participants who are either English or Chinese-English bilingual readers (O'Regan et al., 1984), while the visual field asymmetry for musical notation varied across groups. In particular, our novices, who did not have music reading experience and thus were most vulnerable to potential transfer effects from word reading habits, did not show a bias that was consistent with the asymmetry observed in word processing. Further studies should investigate why the visuospatial biases stemmed from reading direction generalize to faces and line bisection but not to musical notes.

## Conclusions

In this study, we demonstrate with the case of music reading that a left-to-right reading direction is sufficient to lead to a left-biased OVP in expert reading. The OVP for music sequences may gradually shift to the left in the course of music reading training as reading skills improve. Our failure of observing a left-biased OVP in music sequence processing in novices suggests that the asymmetry effect created by word reading habits is not generalizable to all domains of object recognition. In contrast, it may be a result of learning changes during the development of reading expertise.

## Acknowledgments

We are grateful to the Research Grant Council of Hong Kong (project code: HKU 745210H to J.H. Hsiao).

## References

- Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, 226, 177.
- Brady, N. (2011). Understanding spatial bias in face perception and memory. In *Spatial dimension of social thought*. T. W. Schubert & A. Maass (Ed.). Mouton de Gruyter.
- Brysbaert, M. (1994). Interhemispheric transfer and the processing of foveally presented stimuli. *Behavioural Brain Research*, 64, 151-161.
- Brysbaert, M., & Nazir, T. (2005). Visual constraints in written word recognition: Evidence from the optimal viewing position effect. *Journal of Research in Reading*, 28, 216-228.
- Chokron, S. & Imbert, M. (1993). Influence of reading habits on line bisection. *Cognitive Brain Research*, 1, 219-222.
- Deutsch, A., & Rayner, K. (1999). Initial fixation location effects in reading Hebrew words. *Language and Cognitive Processes*, 14(4), 393-421.
- Farid, M., & Grainger, J. (1996). How initial fixation position influences visual word recognition: A comparison of French and Arabic. *Brain and Language*, 53, 351-368.
- Hsiao, J. H., & Cottrell, G. W. (2009). Not all expertise is holistic, but it may be leftist: The case of Chinese character recognition. *Psychological Science*, 20(4), 455-463.
- Hunter, Z. R., Brysbaert, M., & Knecht, S. (2007). Foveal word reading requires interhemispheric communication. *Journal of Cognitive Neuroscience*, 19:8, 1373-87.
- Kazandjian, S., & Chokron, S. (2008). Paying attention to reading direction. *Nature Reviews Neuroscience*, 9, 965.
- Nazir, T. A., Ben-Boutayab, N., Decoppet, N., Deutsch, A., & Frost, R. (2004). Reading habits, perceptual learning, and recognition of printed words. *Brain and Language*, 88, 294-311.
- O'Regan, J. K., Lévy-Schoen, A., Pynte, J., & Brugailière, B. (1984). Convenient fixation location within isolated words of different length and structure. *Journal of Experimental Psychology: Human Perception and Performance*, 10(2), 250-257.
- Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9(1), 97-113.
- Stevens, M., & Grainger, J. (2003). Letter visibility and the viewing position effect in visual word recognition. *Perception & Psychophysics*, 65(1), 133-151.
- Vaid, J. & Singh, M. (1989). Asymmetries in the perception of facial affect: Is there an influence of reading habits? *Neuropsychologia*, 27, 1277-87.
- Watson, A. B., & Pelli, D. G. (1983) QUEST: a Bayesian adaptive psychometric method. *Perception & Psychophysics*, 33 (2), 113-20.
- Wong, Y. K., & Gauthier, I. (2010). A multimodal neural network recruited by expertise with musical notation. *Journal of Cognitive Neuroscience*, 22:4, 695-713.

# Effects of Learning Order and Previous Language Experience in Novel Word Learning

**Elizabeth A. Woods (ewoods@uh.edu)**

Department of Psychology, University of Houston, 126 Heyne Building  
Houston, TX 77204-5022 USA

**Hanako Yoshida (hyoshida@uh.edu)**

Department of Psychology, University of Houston, 126 Heyne Building  
Houston, TX 77204-5022 USA

## Abstract

Although bilingual language research has increased considerably over the past few decades, there is still much controversy regarding the mechanisms of bilingual language acquisition. The present work aims to provide insights into potential mechanisms of bilingual language learning by utilizing a novel word-learning paradigm in which monolingual and bilingual adults (Experiment 1) and children (Experiment 2) are taught two novel languages either simultaneously or sequentially. Results suggest that bilingual language learning may be occurring in a domain specific manner dependent on learning order, previous language experience, and the participant's age.

**Keywords:** novel word learning; order effects; second language acquisition; bilingualism

## Introduction

Given the increasing prevalence of bilingualism in the world, it is crucial to understand how individuals learn multiple languages and what method of language instruction facilitates learning. Although bilingual language research has increased considerably over the past few decades, there is still much controversy regarding the underlying mechanisms of bilingual language acquisition (BLA). The present paper seeks to shed light on potential mechanisms of BLA by comparing monolingual and bilingual adults (Experiment 1) and children (Experiment 2), who are taught two novel languages either simultaneously or sequentially. This is one of the first papers that experimentally manipulates order of acquisition (OoA) by directly comparing simultaneous and sequential language learning. In addition, the present paper examines the effect of previous language experience on novel word learning by directly comparing monolingual and bilingual participants. Recent research suggests that growing-up bilingual has measurable positive effects on cognitive flexibility (Bialystok, Barac, & Blaye, 2010; Bialystok & Martin, 2004; Bialystok & Viswanathan, 2009; Carlson & Meltzoff, 2008; Kovács & Mehler, 2009; Mezzacappa, 2004) and possibly even novel word learning (Kaushanskaya & Marian, 2009; Yoshida et al., 2011). However, the specific nature of the bilingual cognitive advantage has not been fully addressed. The present work provides insights into these theoretical issues by utilizing a novel word-learning

paradigm in which monolingual and bilingual adults (Experiment 1) and children (Experiment 2) are taught two novel languages either simultaneously or sequentially.

## Previous Language Experience & Word Learning

There is increasing evidence to suggest that monolinguals and bilinguals may employ different strategies or have different biases for word learning. In particular, one such bias is the Mutual Exclusivity (ME) principle. The standard artificial word-learning task that is used to test ME presents a well-known object with a well-known name (e.g., a cup) and a novel unnamed object and then asks participants to select an object, using a novel label (e.g., "Find the dax"). Monolinguals consistently map the novel label to the novel object (Markman, 1989; Markman & Wachtel, 1988). However, bilinguals do not perform as consistently. Many studies have reported that ME may develop later or to a lesser extent in bilinguals (Au & Glusman, 1990; Byers-Heinlein & Werker, 2009; Davidson, Jergovic, Imami, & Theodos, 1997; Davidson & Tell, 2005; Houston-Price et al., 2010; Merriman & Kutlesic, 1993). Others suggest that by 30 months old, bilinguals show ME effects as strong as those of monolinguals (Frank & Poulin-Dubois, 2002). Clearly, there are many important questions that remain in the study of early word learning in monolinguals and bilinguals.

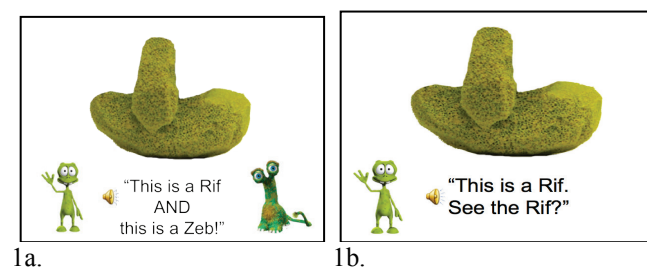
## Order of Acquisition

It is well established that the timing of language exposure can have a substantial impact on how well a language is learned and maintained. Although there are differences of opinion as to the exact age, the general consensus is that there is a sensitive period sometime before puberty, after which language acquisition becomes more difficult (Flege & MacKay, 2004; Johnson & Newport, 1989; Kuhl et al., 1992; Weber-Fox & Neville, 1996; Werker & Tees, 1983; Yamada, 1995). Based on this notion, bilinguals are often classified as being an early or late bilingual, depending on the age of acquisition (AoA) of the second language. There is also evidence regarding AoA effects in experimental studies (Cortese & Khanna, 2007; Morrison & Ellis, 2000; Pérez, 2007) and connectionist networks (Ellis & Lambon Ralph, 2000; Lambon Ralph & Ehsan, 2006; Monaghan &

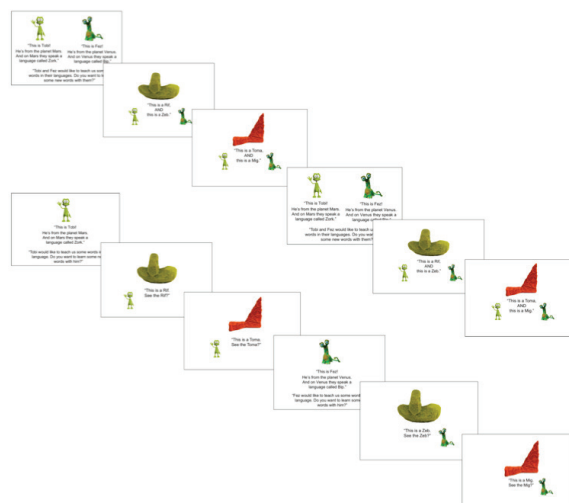
Ellis, 2002) to suggest that early-acquired information is processed faster than later acquired information. Although research regarding AoA and sensitive periods overwhelmingly suggests that learning earlier is better, there is still much controversy regarding the mechanism behind this early advantage.

One explanation that has not been given much consideration is that a more basic mechanism such as order of learning may be responsible for this early advantage (Burling & Yoshida, 2011). Recent experimental paradigms have found that early-learned items are retained better than later-learned items, even when difficulty and age and amount of exposure are controlled (Izura et al., 2011; Stewart & Ellis, 2008). This suggests that AoA effects could actually be due to more basic order effects. The present work further examines this possibility of order effects by comparing simultaneous and sequential learning of two novel languages.

The bilingual literature readily discusses OoA; bilinguals who acquire both languages from birth are referred to as simultaneous bilinguals and those who learn one language before the other are referred to as sequential bilinguals. Simultaneous and sequential bilingualism are even conceptualized as two different mechanisms – Bilingual First Language Acquisition and Second Language Acquisition respectively (for a review see Genesee & Nicoladis, 2007; Paradis, 2007) – suggesting that simultaneous and sequential language learning are qualitatively different. Although these studies provide an important step towards understanding the different types of bilingualism retrospectively, the majority do not experimentally manipulate simultaneous and sequential dual language learning. In addition, BLA is often confounded by the fact that all simultaneous bilinguals are early bilinguals, and many sequential bilinguals are late bilinguals. Given that early learning is thought to be superior, it is important to compare simultaneous and sequential language learning when both occur early. The present work addresses these difficulties by experimentally testing adults (Experiment 1) and children (Experiment 2) who are in the active process of acquiring two new languages in two different mock bilingual environments.



**Figure 1.** Example training trials from the simultaneous condition (a) and sequential condition (b).



**Figure 2.** Sample training design for two word-object pairs in the simultaneous condition (on the top) and the sequential condition (on the bottom) for Experiment 1 and 2.

## Experiment 1

The purpose of Experiment 1 was to determine how the two factors of order of acquisition (OoA) and previous language experience influence word learning in adults. To accomplish this, we utilized a novel word-learning paradigm that imitated a bilingual environment by exposing monolingual and bilingual adults to two novel languages. Participants were randomly assigned to learn these two novel languages either one after the other (Sequential Condition) or at the same time (Simultaneous Condition). Participants were instructed that they would be learning new words from two friendly aliens that were from different planets and spoke different languages. They were then introduced to the aliens and underwent 16 training trials followed by 16 testing trials.

## Participants

Eighty adults ( $M_{\text{age}} = 24.06$  years,  $SD = 4.20$ ) were recruited from the University of Houston in Houston, Texas and were asked to fill out a consent form and demographic questionnaire that included questions about their language history. Based on their responses, participants were identified as bilingual or monolingual and were randomly assigned to either the simultaneous or sequential condition. This yielded four groups of adults: Simultaneous Bilingual ( $N = 21$ ), Simultaneous Monolingual ( $N = 19$ ), Sequential Bilingual ( $N = 20$ ), and Sequential Monolingual ( $N = 20$ ).

Consistent with the Houston population, monolinguals were monolingual English speakers, and bilingual participants represented a variety of language groups, including Arabic-English, French-English, German-English, Japanese-English, Malayalam-English, Mandarin-English, Spanish-English, and Vietnamese-English. Similarly, bilinguals included those who learned both languages from birth, both languages early in life but sequentially, and

second language learners who learned their second language long after their first. This allows the results of the present study to be extended to a broader bilingual population.

## Materials

Sixteen novel word-object pairs were created by assigning each of 16 novel words to one of eight novel objects so that each object had two labels – one in each language. Objects were digital photographs of three-dimensional items created in our lab from craft materials; words were selected from a non-word database (Horst, 2009), produced and recorded by a female native English speaker, and presented auditorily to participants. Sounds were intentionally selected to be as similar as possible across the two languages in order to increase experimental control for the particular question of order effects. Words in each language were matched for length and number of syllables. Alien characters were three-dimensional digital cartoon figures selected from an online graphics database.

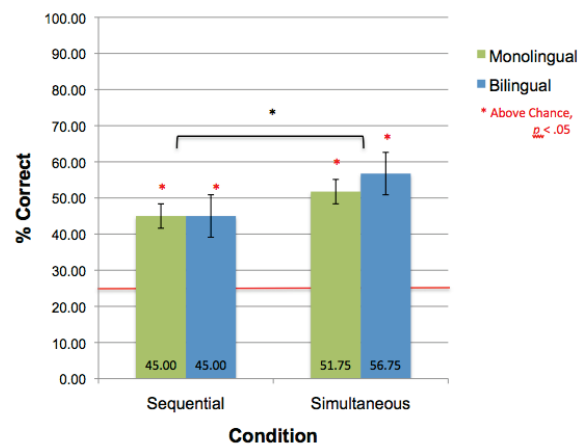
## Training

**Sequential Condition** Training Phase 1 included 8 training trials, each of which presented a single novel object on the screen paired with an auditory phrase (i.e. “This is a Dax! See the Dax?”). Each trial lasted 5 seconds with both the auditory phrase and the object presented at trial onset, and the object remaining on the screen for the entire trial. This was immediately followed by Training Phase 2, which continued in the same manner, but with a set of 8 new words for the same objects. By the end of these two phases of training, each of the 8 objects had been on the screen for 10 seconds, and each of the 16 words had been heard twice.

**Simultaneous Condition** Training Phase 1 occurred in the same manner as the sequential condition, except that on each training trial the object was paired with two novel names within the same auditory phrase (i.e. “This is a Dax and this is a Mig!”). Training Phase 2 was an exact replication of phase 1 for this condition. By the end of these two phases of training, each of the 8 objects had been on the screen for 10 seconds, and each of the 16 words had been heard twice.

## Testing

Testing included 16 trials, one for each word-object pair presented during training. Each testing trial consisted of a four-choice forced alternative test that instructed the participant to select a particular object (i.e. “Point to the Dax!”). All target and non-target stimuli had been presented during training. Testing trials were randomized so that word-object pairs from phases 1 and 2 were tested intermixed. Participants received positive or negative auditory feedback on each testing trial based on their response.



**Figure 3.** Experiment 1 results of adults’ learning of novel word-object pairs.

## Results

Overall, adults learned approximately half ( $M = 49.69\%$ ,  $SD = 17.37$ ) of the word-object pairs, and all four groups of adults performed significantly above chance level. See Figure 3 for more details. To determine the influence of order of language acquisition and previous language experience on learning, a fixed effects ANOVA with factors of order (sequential, simultaneous) and language status (monolingual, bilingual) was conducted. Only a main effect of order was found,  $F(1,76) = 5.95$ ,  $p = .017$ , with an advantage for simultaneous learning over sequential learning. No main effect of language status or an interaction of order with language status was found. These findings suggest that regardless of previous language experience, adults learned the languages better when they were presented simultaneously.

## Experiment 2

Results from Experiment 1 suggest that when two languages are taught simultaneously, adults retain a greater number of words than when taught sequentially. In addition, previous language experience did not influence adults’ performance. However, given that children process and learn information differently from adults (Hudson, Kam, & Newport, 2005; Hudson, Kam, & Newport, 2009; Ramscar et al., 2011), and may use different word learning strategies than adults, this raises the question of how order of acquisition and previous language experience influence word learning in children. To accomplish this, we utilized a shortened version of the novel word-learning paradigm in Experiment 1.

## Participants

Seventy-nine children ages 36 to 72 months ( $M_{\text{age}} = 51.78$  months,  $SD = 11.9$ ) were recruited from the local communities and preschools in Houston, Texas. Parents completed a consent form and demographic questionnaire that included questions about their child’s language history. Based on their responses, participants were identified as

bilingual or monolingual and were randomly assigned to either the simultaneous or sequential condition. This yielded four groups of children: Simultaneous Bilingual ( $N = 17$ ), Simultaneous Monolingual ( $N = 23$ ), Sequential Bilingual ( $N = 17$ ), and Sequential Monolingual ( $N = 22$ ). Similar to Experiment 1, bilingual participants represented a variety of language groups and language backgrounds. However, since all participants were 6 years of age or younger, both languages were learned relatively early in life.

## Materials

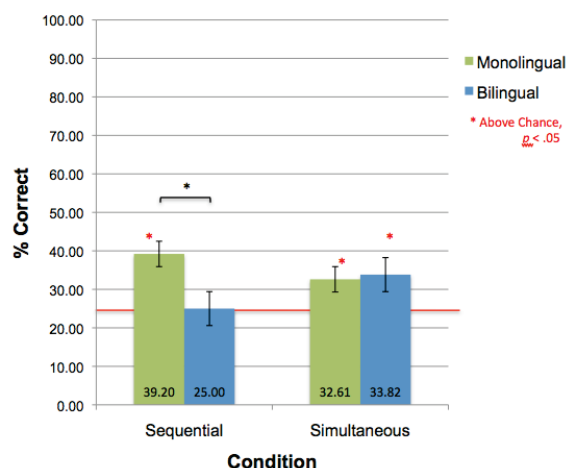
Stimuli were a subset of the stimuli from Experiment 1 and included 8 novel words and 4 novel objects (each object had two labels – one in each language).

## Procedure

Procedures followed in the same manner as that of Experiment 1, but with a total of 8 training trials instead of 16 and 8 testing trials instead of 16.

## Results

Overall, children learned one-third ( $M = 33.07\%$ ,  $SD = 16.63$ ) of the word-object pairs, and three of the four groups of children performed significantly above chance level. See Figure 4 for more details. To determine the influence of order of acquisition and previous language experience on learning, a fixed effects ANOVA with factors of order (sequential, simultaneous) and language status (monolingual, bilingual) was conducted. No main effects were found, but an interaction effect of order by language status was found,  $F(1,75) = 4.40$ ,  $p = .039$ . This suggests that the effect of order of acquisition on novel word learning was dependent on previous language experience. Post hoc simple main effects analyses revealed that monolinguals scored significantly higher than bilinguals in the sequential condition, but there were no differences between monolinguals and bilinguals in the simultaneous condition.



**Figure 4.** Experiment 2 results of children's learning of novel word-object pairs.

## General Discussion

These findings have implications for our understanding of the mechanisms underlying word learning in general, and how those mechanisms might differ in monolinguals and bilinguals, and adults and children. In particular, our results suggest that word learning may be occurring in a domain specific manner dependent on order of learning, previous language experience, and participant's age.

In regards to the effect of order of learning, adults learned a greater percent of the words in the simultaneous condition than in the sequential condition. This is consistent with previous research regarding sensitive periods (Flege & MacKay, 2004; Johnson & Newport, 1989; Kuhl et al., 1992; Weber-Fox & Neville, 1996; Werker & Tees, 1983; Yamada, 1995), AoA effects (Cortese & Khanna, 2007; Ellis & Lambon Ralph, 2000; Lambon Ralph & Ehsan, 2006; Monaghan & Ellis, 2002; Morrison & Ellis, 2000; Pérez, 2007), and OoA effects (Izura et al., 2011; Stewart & Ellis, 2008), all of which suggest that learning information earlier is better. This simultaneous advantage could also potentially be a result of more distributed learning (see Cepeda, et al., 2006) since information in the simultaneous condition was repeated across two training sessions instead of repeated within a single training session.

However, the effect of order was different for children. In the case of Experiment 2, learning order interacted with children's previous language experience. Monolingual children scored significantly higher than bilingual children in the sequential condition, but there were no differences between monolinguals and bilinguals in the simultaneous condition. Learning order still mattered for children, but the ideal type of acquisition was dependent on their previous language experience.

The effect of previous language experience also differed in children and adults. Experiment 1 revealed no differences between monolingual and bilingual adults in terms of the percent of words learned. However, in Experiment 2, monolingual children performed better than bilingual children in the sequential condition. These results seem inconsistent with previous research that would suggest a possible bilingual advantage for language learning (Kaushanskaya & Marian, 2009; Yoshida et al., 2011). However, the current work essentially entailed associative learning of word-object pairs, whereas previous studies that have found a bilingual cognitive advantage typically involved attentional shifting or competition resolution. Thus, it may be that any bilingual cognitive advantages are highly task specific, and that the current task did not tap into the cognitive skills for which bilinguals typically show an advantage. Future research should continue to evaluate monolinguals' and bilinguals' performance on various types of learning tasks.

Given that the effects of learning order and previous language experience had differential effects in adults and children, it may be that different learning mechanisms are driving bilingual language learning in adults and children.

This is consistent with previous research suggesting that adults and children process information in qualitatively different ways (Hudson, Kam, & Newport, 2005; Hudson, Kam, & Newport, 2009; Ramscar et al., 2011). Additional developmental work is needed to determine how and why these learning mechanisms may change across development.

The present work highlights the potential importance of order in learning and the differential effects that it may have for language learning in monolinguals and bilinguals, as well as in adults and children. As seen here, bilingual language learning may be highly dependent on the specific type of learning environment as well as other factors that were not directly addressed in the present work, such as the relationship between the two languages and the individual's specific linguistic background (Bialystok, McBride-Chang, & Luk, 2005; Cummins, 1979; Müller, 1998; Müller & Hulk, 2001). Although the current stimuli were designed to be similar across languages for purposes of experimental control, this is not akin to a natural language learning environment in which languages are phonologically distinct from one another. Bilingual children in particular may be highly sensitive to such linguistic cues given their experience with these cues signaling which language is being used. These matters merit further investigation to advance general learning theories as well as theories of bilingual language acquisition.

### Acknowledgments

This research was supported in part by a National Institutes of Health grant (R01 HD058620), the Foundation for Child Development, and the University of Houston's Grants to Enhance and Advance Research (GEAR) program. We thank the children and parents who participated in this study.

### References

- Au, T. K., & Glusman, M. (1990). The principle of mutual exclusivity in word learning: To honor or not to honor? *Child development*, 61, 1474-1490.
- Bialystok, E., Barac, R., & Blaye, A. (2010). Word mapping and executive functioning in young monolingual and bilingual children. *Journal of Cognition and Development*, 11(4), 485-508.
- Bialystok, E., & Martin, M. M. (2004). Attention and inhibition in bilingual children: Evidence from the dimensional change card sort task. *Developmental Science*, 7, 325-339.
- Bialystok, E., McBride-Chang, C., & Luk, G. (2005). Bilingualism, language proficiency, and learning to read in two writing systems. *Journal of Educational Psychology*, 97 (4), 580-590.
- Bialystok, E., & Viswanathan, M. (2009). Components of executive control with advantages for bilingual children in two cultures. *Cognition*, 112, 494-500.
- Burling, J., & Yoshida, H. (2011). A Developmental Perspective on Order and Learning: Temporal Effects on Cued Attention. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2937-2942). Austin, TX: Cognitive Science Society.
- Byers-Heinlein, K., & Werker, J. F. (2009). Monolingual, bilingual, trilingual: Infants' language experience influences the development of a word learning heuristic. *Developmental Science*, 12(5), 815-823.
- Carlson, S. M. & Meltzoff, A. N. (2008). Bilingual experience and executive functioning in young children. *Developmental Science*, 11(2), 282-298.
- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132, 354-380.
- Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2, 342 words. *Quarterly Journal of Experimental Psychology*, 60, 1072-1082.
- Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, 49, 222-251.
- Davidson, D., Jergovic, D., Imami, Z., & Theodos, V. (1997). Monolingual and bilingual children's use of the mutual exclusivity constraint. *Journal of Child Language*, 24, 3-24.
- Davidson, D. & Tell, D. (2005). Monolingual and bilingual children's use of mutual exclusivity in the naming of whole objects. *Journal of Experimental Child Psychology*, 92, 25-45.
- Ellis, A. W., & Lambon Ralph, M. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 1103-1123.
- Fllege, J. E., and MacKay, I. R. A. (2004). Perceiving vowels in a second language. *Studies of Second Language Acquisition*, 26, 1-34.
- Frank, I., & Poulin-Dubois, D. (2002). Young monolingual and bilingual children's responses to violation of the Mutual Exclusivity Principle. *International Journal of Bilingualism*, 6(2), 125-146.
- Genesee, F., & Nicoladis, E. (2007). Bilingual acquisition. In E. Hoff & M. Shatz (eds.), *Handbook of Language Development*, 324-342. Oxford, Eng.: Blackwell.
- Houston-Price, C., Caloghris, Z., & Raviglione, E. (2010). Language experience shapes the development of the mutual exclusivity bias. *Infancy*, 15(2), 125-150.
- Hudson Kam, C. L. & Newport E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language*



- Learning and Development*, 1, 151-195.
- Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, 59, 30-66.
- Izura, C., Pérez, M., Agallou, E., Wright, V. C., Marín, J., Stadthagen-González, H., & Ellis, A. W. (2011). Age/order of acquisition effects and the cumulative learning of foreign words: A word training study. *Journal of Memory and Language*, 64, 32-58.
- Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21, 60-99.
- Kaushanskaya, M., & Marian, V. (2009). The bilingual advantage in novel word learning. *Psychonomic Bulletin & Review*, 16, 705-710.
- Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255, 606-608.
- Kovács, Á. M., & Mehler, J. (2009). Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 6556-6560.
- Lambon Ralph, M. A., & Ehsan, S. (2006). Age of acquisition effects depend on the mapping between representations and the frequency of occurrence: Empirical and computational evidence. *Visual Cognition*, 13, 928-948.
- Markman, E. M. (1989). *Categorization and Naming in Children: Problems of Induction*. MIT Press.
- Markman, E. M. & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20, 121-157.
- Merriman, W. E. & Kutlesic, V. (1993). Bilingual and monolingual children's use of two lexical acquisition heuristics. *Applied Psycholinguistics*, 14, 229-249.
- Mezzacappa, E. (2004). Alerting, orienting, and executive attention: Developmental properties and sociodemographic correlates in an epidemiological sample of young, urban children. *Child Development*, 75, 1373-1386.
- Monaghan, J., & Ellis, A. W. (2002). What, exactly, interacts with spelling sound consistency in word naming? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 183-206.
- Morrison, C. M., & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology*, 91, 167-180.
- Müller, N. (1998). Transfer in Bilingual First Language Acquisition. *Bilingualism: Language and Cognition*, 1 (3), 151-171.
- Müller, N. & Hulk, A. (2001). Crosslinguistic Influence in Bilingual Language Acquisition: Italian and French as Recipient Languages. *Bilingualism: Language and Cognition*, 4 (1), 1-21.
- Paradis, J. (2007). Second language acquisition in childhood. In E. Hoff and M. Shatz (Eds.), *Handbook of Language Development* (pp. 387-406). Oxford: Blackwell.
- Pérez, M. A. (2007). Age of acquisition persists as the main factor in picture naming when cumulative word frequency and frequency trajectory are controlled. *The Quarterly Journal of Experimental Psychology*, 60, 32-42.
- Ramscar, M., Dye, M., Klein, J., Ruiz, L.D., Aguirre, N. & Sadaat, L. (2011). Informativity versus logic: Children and adults take different approaches to word learning. In L. Carlson, C. Hoelscher, & T.F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3326-3331). Austin, TX: Cognitive Science Society.
- Stewart, N., & Ellis, A. W. (2008). Order of acquisition in learning perceptual categories: A laboratory analogue of the age of acquisition effect? *Psychonomic Bulletin & Review*, 15(1), 70-74.
- Weber-Fox, C. M., & Neville, H. J. (1996). Maturation constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience*, 8, 231-256.
- Werker, J. F., & Tees, R. C. (1983). Developmental changes across childhood in the perception of normative speech sounds. *Canadian Journal of Psychology*, 37, 278-286.
- Yamada, R. A. (1995). Age and acquisition of second language speech sounds. Perception of American English /r/ and /l/ by native speakers of Japanese. In *Speech Perception and Linguistic Experience: Issues in Cross-language Research*. W. Strange York, Timonium, MD, pp. 305-320.
- Yoshida, H., Tran, D. N., Benitez, V., & Kuwabara, M. (2011). Inhibition and Adjective Learning in Bilingual and Monolingual Children. *Frontiers in Developmental Psychology*, 2:210.



# **Inference and culture : The distinction between low context culture and high context culture as a possible explanation for cultural differences in cognition**

**Hiroshi Yama (yama.hiroshi1204@gmail.com)**

School of Literature and Human Sciences, Osaka City University

Sumiyoshi, Osaka 558-8585, JAPAN

**Norhayati Zakaria (norhayatizakaria@uowdubai.ac.ae)**

Faculty of Business and Management, University of Wollongong, Dubai

1-02, Block 15, Knowledge Village, Dubai, U.A.E.

## **Abstract**

Nisbett et al. (2001) claim that Easterners are more likely to use holistic thinking to solve problems, whereas Westerners are more likely to use analytic thinking. This distinction in cognitive behaviors has often been explained by using a framework based on the fact that Western culture favors independent self-construal (individualist culture) and Eastern culture favors interdependent self-construal (collectivist culture). However, we propose another possible cultural explanation in the distinction between Western low context culture and Eastern high context culture (Hall, 1976). We particularly focus on the difference between the rule-based inference more common in low-context Western cultures and the dialectical inference more common in high-context Eastern cultures, and we argue that rule-based inference using global rules is more adaptive in low context cultures.

Keywords: Culture; Psychology; Reasoning; Cross-cultural analysis.

## **1. Cultural Differences in Cognition**

Nisbett (2003; Nisbett, Peng, Choi, & Norenzayan, 2001) reviewed previous studies on cultural differences in cognition and described the differences in terms of a distinction between analytic and holistic cognition. He argued that individuals from Western cultures are more likely to engage in analytic cognition, whereas individuals from Eastern cultures are more likely to engage in holistic cognition. According to his definition, analytic cognition involves detachment of the object from its context, a tendency to focus on attributes of the object to assign it to a category, and a preference for using rules about the categories to explain and predict the object's behavior. In contrast, holistic cognition is oriented towards context or the field as a whole, attention to relationships between a focal

object and the field, and a preference for explaining and predicting events on the basis of such relationships.

The distinction between analytic and holistic can be described in terms of four dimensions: context-dependent/independent, dispositional/situational attribution, rule-based/dialectical, and stable/changeable views. In terms of the third dimension (rule-based vs. dialectical), people from Western cultures are more inclined to employ rule-based thinking, whereas people from Eastern cultures are more apt to employ dialectical thinking (Buchtel & Norenzayan, 2008; Peng & Nisbett, 1999; Spencer-Rogers, Boucher, Mori, Wang, & Peng, 2009). For example, Norenzayan, Smith, Kim, and Nisbett (2002) reported that, when being asked which group an object should belong to (categorization task), Americans tended to focus on a single property (rule-based inference), whereas Koreans tended to respond based on family resemblance (intuitive inference). Peng and Nisbett (1999) proposed that the cognitive style of Chinese was dialectical, whereas the cognitive style of Americans was rule-based.

## **2. Nisbett's Explanations for the Cultural Differences in Cognition**

Nisbett et al. (2001) explained the distinction between Western analytic and Eastern holistic cognition by using the cultural value dimensions that underlines the individualist and collectivist cultures (Triandis, 1995). They discussed how each style is adaptive within its own cultural type. We regard culture as a hypothetical construct to explain people's behavior as well as to describe social patterns. In the long history of cultural studies, it has been claimed that Western societies have established individualist cultures, whereas Eastern societies have developed collectivist cultures

(Triandis, 1995). The distinction between individualist and collectivist culture is a hypothetical concept proposed to explain the observed differences in behavior, such as that people from Eastern cultures have a stronger preference for sociability and interdependence than do people from Western cultures. Markus and Kitayama (1991) connected this distinction to two kinds of selves. They postulated that, in general, Western cultures foster and favor an independent self, whereas Eastern cultures foster and favor an interdependent self. This distinction refers to differences in how people view themselves: people from Western cultures are likely to view themselves as individualistic, ego-centric, and discrete from society, whereas people from Eastern cultures are more inclined to view themselves as collectivistic, socio-centric, and related to others and to their society.

Nisbett (2003, Nisbett et al., 2001) argued that in a collectivist culture it is adaptive to attend not only to an object itself but also to its context in order to keep the harmony, hence Eastern cultures' holistic cognition is practiced and facilitated. More recently, Nisbett has said he prefers an explanation based on the personal level, in other words on the concept of self-construal (e.g., Varnum, Grossman, Kitayama, & Nisbett, 2010).

The both explanations of Nisbett's are compatible with the results of something called "cultural priming." As already mentioned, it is assumed that Western cultures foster development of an independent self, whereas Eastern cultures promote development of an interdependent self (Markus & Kitayama, 1991). Cultural priming is the mechanism that makes either independent or interdependent self-construal accessible, and the accessible self-construal in turn affects the style of cognition. For example, Kühnen, Hannover, and Schubert (2001) reported that participants who were asked to point out the differences between themselves and their friends or parents (primed as independent self-construal) showed a tendency to process stimuli unaffected by the context (analytic cognition), whereas those who were asked to point out the similarities between themselves and their friends or parents (primed as interdependent self-construal) were more apt to do context-bounded thinking (holistic cognition).

For the distinction between rule-based inference and dialectical inference, Nisbett (2003) adds the importance of cultural tradition. The Western style of thinking has been heavily influenced by the philosophy of Ancient Greece, whereas the Eastern style of thinking grew out of the traditions of Taoism, Confucianism, and Buddhism.

Aristotle's logic was accepted by many Western cultures as it is abstract and universal, whereas Eastern cultures preferred ideas that encouraged and reinforced the harmony of their society. For example, the dual concept of *yin* (negative aspects of the world) and *yang* (positive aspects of the world) form the central essential of Taoism, describing how polar opposites or seemingly contrary forces are in reality interconnected and interdependent. It reflects the tradition of Chinese ontology that the world is constantly changing and shifting, like the balance between *yin* and *yang*, and is full of contradictions. Nisbett concludes that the Chinese view the world as easy to change (e.g., Ji, Nisbett, & Su, 2001), hence abstract rules are not useful for predicting future events or guiding behavior. Nisbett postulates that this is why the Chinese (and thus Easterners) are less likely to use rule-based inference.

In short, these results support the view of cultural psychologists who assume that mind and culture are inseparable. In Western societies, people live in an individualist culture, develop independent self-construal, and thus are more likely to demonstrate analytic cognition, whereas people in Eastern societies live in a collectivist culture, develop interdependent self-construal, and are more apt to demonstrate holistic cognition. This view is summarized as the social orientation hypothesis (Varnum et al., 2010).

However, we see some problems with Nisbett's (2003) explanation. The first is the alleged adaptive nature of Eastern cultures' attention to contextual information. It may well be adaptive to pay attention not only to a target person (object), but to all in-group members (context) in order to maintain in-group harmony in a collectivist culture. However, strictly speaking, this cognitive style is adaptive only in the field of person cognition in a collectivist culture. How can this person cognition be transferred to objects and their context?

Secondly, if Eastern cultures view the world as changeable, the question is whether they try to predict those changes using rules such as *yin* and *yang*. However, Nisbett (2003) takes his interpretation of the Eastern view as fact, and infers that the concept of *yin* and *yang* reflects this view.

### **3. Low Context and High Context Cultures**

#### **3.1 Hall's (1976) Definition**

In order to resolve the problems above, we propose an explanation based on the distinction between low context and high context culture. Hall (1976) introduced a dominant cultural dimension called context to explore the relationship

between culture and communication. His interest was built upon the need to understand the factors that facilitate or inhibit effective communication between individuals from different cultural backgrounds. In explaining this key cultural dimension, Hall and Hall (1990) integrated three main concepts: context, information, and meaning. These three concepts encapsulate context as a system of meaning for information exchanges between groups of people or within a group of people. They further argued that context is embedded in information for the purpose of creating meaning in a message. In other words, without information or context, a message is deemed to be without meaning, therefore insignificant.

With this understanding, Hall's context dimension provides a framework that enables people to comprehend communication forms ranging from the purely non-verbal — hand gestures, body language, facial expressions, and tone of voice (all of which are situational and important in high context cultures) — to the purely verbal, such as written text or spoken words (all of which are informational and important in low context cultures), in order to achieve meaning as the ultimate goal. Zakaria and Cogburn (2010) summarized it thus: high context is known as 'content independent', while low context is known as 'context independent.' There is some evidence that, generally speaking, Western cultures are low context whereas Eastern cultures are high context. For example, Ishii, Reyes, and Kitayama (2003) reported that Americans spontaneously attend more to verbal content than to vocal tone, whereas Japanese attend more to vocal tone than to verbal context. This evidence suggests that Japanese prefer indirect and implicit communication while Americans prefer direct and explicit communication. In their analysis of websites, Würtz (2006) found that websites created by Japanese, Chinese, and Koreans, who are from presumed high context cultures, adopted the visual effects offered by the Internet to convey their messages efficiently to a greater degree than did sites created by Germans, Americans, and Northern Europeans, who are from presumed low context cultures.

In this paper, we propose that the distinction between high context and low context cultures (the L/H context dimension) (Hall, 1976) offers a better explanation for cultural differences in cognition than the distinction between individualism and collectivism (the I/C dimension).

People in a high context culture can interpret messages from others without full descriptions, because implicitly shared information is available to assist the interpretation. For example, if people implicitly share the

idea that a diamond is very expensive and normally used for special occasions, then given the statement 'A presented a diamond ring to B' they may infer that A is proposing marriage to B. Therefore, a speaker can notify you of A's proposal just by saying that A presented a diamond ring to B without further information on marriage. On the other hand, people in a low context culture need explicitly expressed words for communication, because they have little or no implicitly shared information to draw on. Hence, they rely on communication which rests upon direct and explicit communication.

As for the problem of the transference from person cognition to objects, the explanation using the H/L context distinction does not need to rely on transference. According to Ishii et al. (2003) using the H/L context distinction can explain the degree of contextualization, and the degree of how people attend contextual or situational information, which are the first two aspects of the dimension between analytic and holistic cognition. In a high context culture, people's attention is attuned to contextual information because they are accustomed and encouraged to use this information for communication, whereas in a low context culture, people's attention is directed towards the target they want to identify. It is highly plausible that this cultural training affects people's cognition in each culture.

### **3.2 The H/L Context Distinction as an Explanation for Cultural Differences in the Usage of Rules**

The H/L context distinction can also provide an explanation for cultural differences in the usage of rules. The outline of our argument is as follows;

- (1) A global rule is needed when a local rule becomes useless.
- (2) A local rule becomes useless when natural laws and/or social customs are variable.
- (3) Social customs are more variable in low context (Western) cultures than in high context (Eastern) cultures.
- (4) Eastern high context cultures' dialectical inference is not based on global rules but on local rules, while the opposite is true of Western low context cultures.
- (5) Therefore, Western low context cultures are more inclined to use rule-based inference than Eastern high context cultures.

This explanation resolves the problem of why the Chinese, for example, are less inclined to use rules to describe changes that they perceive in the world: because by and large they encounter less variability in their local world and therefore local practice remains useful.

Why have scientific theories been needed for humans? Rules are used to describe the world in terms of natural laws, and to predict consequences. Although they do not give direct suggestions for human action, they are useful in gaining resources (benefits) or avoiding hazards. For example, it is adaptive for people to learn the follow law in a hunter-gatherer society:

*If you go to the river in autumn, you can catch salmon.*

People learn that they can catch salmon every autumn and thus can smoke salmon for eating through the winter. A scientific theory may give an explanation for why one can catch salmon in autumn. If the law is always true, and you can count on the appearance of the salmon every year, theories are not necessary. However, theories which describe the biological mechanisms, habits or behavior of salmon become useful when there are no salmon one autumn. These theories may explain why this situation has occurred and give people some idea how to deal with it: give up fishing, move to another place, or clean the river. Therefore, theories are needed when the environment is not stable, and its natural laws are irregular.

However, we do not assume that the cultural differences in the tendency to seek for a global rule arise from differences in environmental variability between West and East. Rather, we focus on any rule which is used as a cultural coordination device. It takes the form of a deontic conditional, which codifies obligation, permission, and inhibition. For example, Tom fell in love with a girl whose name was Anne when he lived in her country. He wants to marry her, hence he presents her with a diamond ring based on the following belief:

*If Anne wishes to marry Tom, she accepts the diamond which Tom presents her.*

However, it is quite possible that this rule cannot be applied in another culture where people do not share the common belief that a diamond is a marriage gift. If Anne lives in such a society, she may not accept the diamond even if she wants to marry Tom because the rule that Anne knows is as follows:

*If Anne wishes to marry Tom, her father accepts an amount of money from Tom.*

Tom may find out or figure out this rule, give money to

Anne's father, and marry Anne. However, it is more adaptive for them both to know the reasons for the two different rules: that is, the principle of a marriage gift in order to have a happier life. If they know the reasons, they can abandon the old local rules and create new, more global, rules when their child gets married.

We propose that local rules are less useful in a low context culture than in a high context culture. The case of the marriage of Tom and Anne is a typical example: the variability between their backgrounds means that their local rules differ, and acting on them leads to miscommunication. In a nutshell, a global or fundamental rule is necessary when a local rule becomes useless. Cross-cultural studies indicate that people raised in Western cultures prefer more global rules than those raised in Eastern cultures. For example, the results of Norenzayan et al. (2002) cited above showed that Americans preferred to categorize based on formal rules, whereas Koreans inferred based on family resemblance. We propose that the rules used by the Americans are more global than the family resemblance used by the Koreans. Family resemblance consists of set of local rules, and each rule is not true for all members of a category. Spencer-Rodgers, Boucher, Mori, Wang, and Peng (2009) claim that Eastern cultures' dialecticism is naïve, by which they mean that Eastern cultures are more likely to retain some local rules which are contradictory of each other, without resolving the contradiction. Therefore, Eastern cultures' inference is also more local than Western.

The relationship between environmental variability and the necessity of local and global rules are shown in Figure 3. A global rule is not necessary if natural laws or social customs are completely stable. It is needed if the local rule based on natural law or social customs becomes useless. Therefore, the lower the utility of local rules, the higher the necessity of global rules. However, in a completely chaotic situation a global rule is not useful either. Hence, the need for global rules describes an A-shaped curve, as shown in Figure 1. In this figure, focusing on the H/L context distinction, we consider the variability of social customs on the horizontal axis and necessity of a rule on the vertical axis. Social customs are stable in a high context culture, hence it can be located on the left, whereas a low context culture can be located in the middle where social customs are variable to some extent (but not enormously so). However, Nisbett (2003; Nisbett et al., 2001) argues that the Chinese view the world as more changeable than Westerners do, hence he locates the Chinese culture further to the right, where the environment is not stable. This is contradictory with the idea

that Chinese culture is high context, and may therefore be wrong. In short, we cast doubt on Nisbett's argument that Chinese dialectical thinking is based on their view that the world is easily changeable.

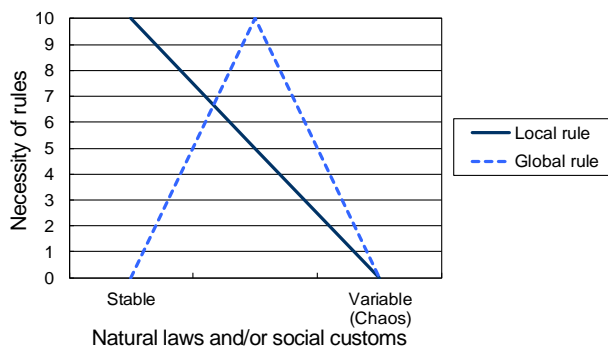


Figure 1: Relationship between Variability and Necessity.

### 3.3 Reinterpretation of the Experimental Results on Cultural Priming

How can the results of cultural priming be explained based on the H/L context distinction? Studies on cultural priming by Kühnen et al. (2001) are one reason for considering an explanation based on the I/C dimension. They assumed that participants who were asked to consider the differences between themselves and their parents or friends activated their independent self-construal, whereas participants who were asked to consider the similarities between themselves and their parents or friends activated their interdependent self-construal.

However, it is possible that their experimental manipulation changed the degree of their participants' feeling of shared context with other people. Asking people to consider differences between themselves and their parents or friends may activate their belief that others are different from themselves, and thus they are not able to rely upon the information which they share with others for communication. On the other hand, asking them to consider similarities between themselves and their parents or friends may activate their belief that others are similar to themselves. In other words, the former brought participants to a low context situation whereas the latter brought them to a high context situation.

### 3.4 Accounting for Cultural Diversity

Our more ambitious aim is to connect the L/H context distinction to the explanation for cultural diversity using natural, ecological, and geographical factors. In other words, not only to explain cultural differences in cognition using the L/H context distinction, but to explain the distinctions themselves using natural, ecological, and geographical

factors. The ecological bases for individualism and collectivism have been intensively discussed. The person level of social independence can be intermediate between their group life style and their analytic cognitive style. These studies are known as a socioecological approach (Oishi & Graham, 2010).

We do not deny these discussions. However, the L/H context distinction may be explicable by natural factors as is the I/C distinction. This problem is too large to fully discuss here; we simply point out some factors contributing to the difference between low context and high context culture, which lead people to either rule-based inference or dialectical inference respectively.

The concept of the L/H context distinction is often employed by researchers in human communication. When people perform intercultural communication, both parties are in effect in a low context situation because they share fewer implicit assumptions than when they communicate with someone from their own culture. This idea is a developed version of Langer (1989), who argued that mindful communication is needed for intercultural communication. His concept of 'mindful' communication can be interpreted as explicitly deliberate and careful communication in which people read others' minds when there is a lack of shared implicit assumptions.

Since a low context situation arises when people engage in intercultural communication, a low context culture is more likely to develop in a multicultural environment (one in which people from different cultures keep their own culture but interact with each other). This situation also creates an environment wherein local rules become useless more often.

A geographical factor that reinforces a multicultural environment is when there is no spacious plain which can become the place for a large culturally unified society; societies must therefore remain geographically separated. In order for multicultural conditions to arise and persist, however, these different cultural societies must interact with each other -- for example, if each society is not economically self-sufficient, and can prosper only if it trades with other societies. An ecological factor that enhances the likelihood of trade is an unbalanced distribution of resources among these societies. One place that satisfied all these conditions was Ancient Greece.

## 4. Conclusion: A New Framework

The primary goal of this paper is to propose a possible explanation for cultural differences in cognition, specifically

the analytic cognition practiced by Western cultures and the holistic cognition practiced by Eastern cultures, using the distinction between low context culture in the West and high context culture in the East instead of the distinction between Western individualist culture and Eastern collectivist culture.

Summarizing these points, we propose a framework as shown in Figure 2. In order to explain cultural diversity naturalistically, we give the primary role to geographical and ecological factors. People need rule-based inference in a low context culture, but whether a low context culture (multicultural environment) arises depends on these factors. Our framework is contrasted with that of cultural psychologists (e.g., Nisbett, 2003; Nisbett et al., 2001), who assume that culture and mind are inseparable and emphasize the role of self-construal in culture and cognitive style. By contrast, we propose that culture and *context* are inseparable and, as such, that context has a strong connection to the types of information required in order to draw effective meanings or sense-making into the thinking process.

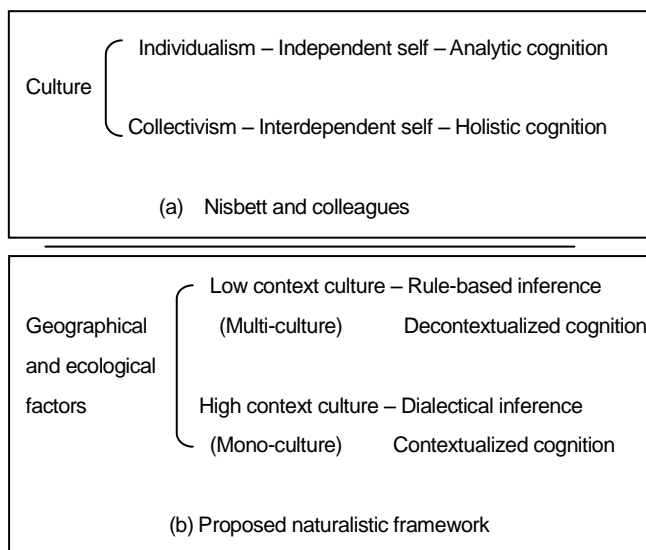


Figure 2: Nisbett's Framework vs Proposed New Framework

## References

- Buchtel, E. E. & Norenzayan, A. (2008). Which should you use, intuition or logic? Cultural differences in injunctive norms about reasoning. *Asian Journal of Social Psychology*, 11, 264-273.
- Hall, E. T. (1976). *Beyond culture*. Garden City, NJ: Anchor Books/Doubleday.
- Hall, E. T., & Hall, M. R. (1990). *Understanding cultural differences*. Yarmouth, ME: Intercultural Press.
- Ishii, K., Reyes, J. A., & Kitayama, S. (2003). Spontaneous attention to word content versus emotional tone: Differences among three cultures. *Psychological Science*, 14, 39-46.
- Ji, L., Nisbett, R. E., & Su, Y. (2001). Culture, change, and prediction. *Psychological Science*, 12, 450-456.
- Kühnen, U., Hannover, B., & Schubert, B. (2001). The semantic-procedural interface model of the self: The role of self-knowledge for context-dependent versus context-independent modes of thinking. *Journal of Personality and Social Psychology*, 80, 397-409.
- Langer, E. (1989). *Mindfulness*. Reading, MA: Addison-Wesley.
- Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98, 224-253.
- Nisbett, R. E. (2003). *The geography of thought: How Asians and Westerners think differently...and why*. New York: The Free Press.
- Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review*, 108, 291-310.
- Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science*, 26, 653-684.
- Oishi, S., & Graham, J. (2010). Social ecology: Lost and found in psychological science. *Perspectives on Psychological Science*, 5, 356-377.
- Peng, K., & Nisbett, R. E. (1999). Culture, dialectics, and reasoning about contradiction. *American Psychologist*, 54, 741-754.
- Spencer-Rogers, J., Boucher, H. C., Mori, S. C., Wang, L., & Peng, K. (2009). The dialectical self-concept: Contradiction, change, and holism in East Asian Cultures. *Personality and Social Psychological Bulletin*, 35, 29-44.
- Triandis, H. C. (1995). *Individual and collectivism*. Boulder, CO: Westview Press.
- Varnum, M. E. W., Grossman, I., Kitayama, S. & Nisbett, R.E. (2010). The origin of cultural differences in cognition: The social orientation hypothesis. *Psychological Science*. 19, 9-13.
- Würtl, E. (2006). Intercultural communication on web sites: A cross-cultural analysis of web sites from high-context cultures and low-context cultures. *Journal of Computer-Mediated Communication*, 11, 274-299.
- Zakaria, N. & Cogburn, D. L. (2010). Context-dependent vs. content-dependent: An exploration of the cultural behavioural patterns of online intercultural communication using E-mail. *International Journal of Business and Systems Research*, 4, 330-347.

# “What” versus “How” in Nonvisual Whole-Body Movement

Naohide Yamamoto (n.yamamoto@csuohio.edu)

Dale A. Hirsch (d.a.hirsch@csuohio.edu)

Department of Psychology, Cleveland State University

2121 Euclid Avenue, Cleveland, OH 44115, USA

## Abstract

Dissociable processes for conscious perception (“what” processing) and guidance of action (“how” processing) have been identified in visual, auditory, and somatosensory systems. The present study was designed to find similar dissociation within whole-body movements in which the presence of vestibular information creates a unique perceptual condition. In two experiments, blindfolded participants walked along a linear path and specified the walked distance by verbally estimating it (“what” measure) and by pulling a length of tape that matched the walked distance (“how” measure). Although these two measures yielded largely comparable responses under a normal walking condition, variability in verbal estimates showed a qualitatively different pattern from that in tape-pulling when sensory input into walking was altered by having participants wear a heavy backpack. This suggests that the “what” versus “how” dissociation exists in whole-body movements as well, supporting a claim that it is a general principle with which perceptual systems are organized.

**Keywords:** Perception; action; somatosensory system; vestibular sense; walking

## Introduction

It has been well documented that perceptual systems contain two separable modes of information processing (Milner & Goodale, 1995; Ungerleider & Mishkin, 1982): One is to consciously recognize a stimulus (so-called “what” processing) and the other is to locate it in space and guide action toward it (so-called “how” or “where” processing). For example, a neurological patient who suffered from visual form agnosia was not able to verbally report the orientation of a slot presented in front of her, but was nevertheless able to put a card in the slot in a normal manner (Goodale, Milner, Jakobson, & Carey, 1991). Such dissociation between “what” and “how” (or “where”) has been most clearly established in the visual system, but similar distinctions have also been made in the auditory system (Anourova et al., 2001; Berlin & Zatorre, 2000; Kaas & Hackett, 1999; Maeder et al., 2001; Rauschecker, 1998; Romanski et al., 1999) and somatosensory system (Aglioti, Beltramello, Bonazzi, & Corbetta, 1996; Halligan, Hunt, Marshall, & Wade, 1995; Kammers, van der Ham, & Dijkerman, 2006; Marcel, 2003; Paillard, 1999; Paillard, Michel, & Stelmach, 1983; Reed, Klatzky, & Halgren, 2005; Rossetti, Rode, & Boisson, 1995; Sathian et al., 2011; Sittig, Denier van der Gon, Gielen, & van Wijk, 1985; Van Boven, Ingeholm, Beauchamp, Bikle, & Ungerleider, 2005; Westwood & Goodale, 2003). These converging findings suggest that separate processing of conscious perception and action guidance (or stimulus location) is a general principle with which perceptual systems are organized. However, in virtually all of the previous studies concerning this dissociation within the somatosensory system,

actions were carried out only with body parts, while the body itself remained stationary (e.g., hand or finger movement; for a review, see Dijkerman & de Haan, 2007). In the present study, we explored whether similarly dissociable processes underlie whole-body movements (i.e., walking).

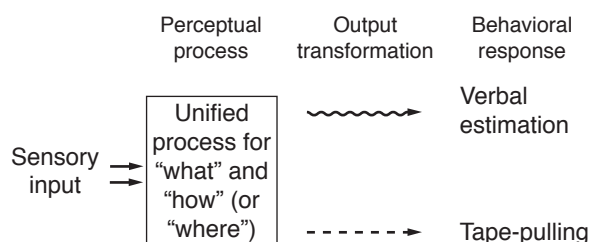
Whole-body movements such as walking present a unique perceptual condition because they encompass not only somatosensory information about motion of each body part but also vestibular information about acceleration and velocity of the entire body. It has been shown that somatosensory perception can be altered in the presence of vestibular inputs (Bottini et al., 1995; Ferrè, Bottini, & Haggard, 2011), which may be due to the fact that somatosensory and vestibular information are processed in an integrated fashion in largely overlapping areas of the brain (Bottini et al., 1994; Fasold et al., 2002; Guldin & Grüsser, 1998; Schwarz & Fredrickson, 1971). For example, Ferrè et al. demonstrated that sensitivity to tactile stimuli can be increased by caloric vestibular stimulation, and they also showed that this perceptual enhancement was specific to the somatosensory system. Findings like this indicate that vestibular inputs affect the operation of the somatosensory system, suggesting that somatosensory processes underlying whole-body movements are not identical to those subserving partial-body actions. Thus, it should not be assumed that similar dissociation between conscious perception and action guidance would be found in whole-body movements as well. Rather, it is an open question that should be addressed empirically.

To address this issue, we conducted two experiments in which blindfolded participants walked along a linear path and indicated the walked distance by using two types of measures: One was driven by a motoric response in which participants pulled a length of tape that matched the walked distance (Philbeck, Woods, Kontra, & Zdenkova, 2010). The other was verbal estimation of the walked distance, which required conscious awareness of how far they had walked. Although we hypothesized that these two measures are based on dissociable processes (i.e., “what” process for verbal estimation and “how” process for tape-pulling), we did not simply look for different patterns of response from them. Even if information about the walked distance was processed in a unified manner for verbal estimation and tape-pulling, they could still yield distinct patterns of data because the post-perceptual transformation required to translate the internal representation of the (already processed) distance information into a behavioral output might be carried out differently for each mode of response (Figure 1A). Thus, a stronger test on whether dissociable processes exist in whole-body movements can be



done by altering sensory input into nonvisual walking and observing how the patterns of response are modulated relative to baseline patterns observed under a normal sensory condition (Foley, 1977; Philbeck & Loomis, 1997). If verbal estimation and tape-pulling were subserved by two separate processes (Figure 1B), it would be more likely that these processes were affected differently by the alteration of sensory input. As a consequence, patterns of response in the two measures would also change differently from the baseline. On the other hand, if a sole process underlay both verbal estimation and tape-pulling (Figure 1A), the altered sensory input would cause some common change in both measures (e.g., both verbal estimates and lengths of tape pulled doubled). These possible changes from the baseline can be observed even if responses in verbal estimation and tape-pulling were generated by different output transformations, because there is no logical ground to postulate that these post-perceptual transformations should also be modified by the sensory alteration; rather, it would be more reasonable to assume that they should remain unchanged.

#### A. Single-process model



#### B. Dual-process model

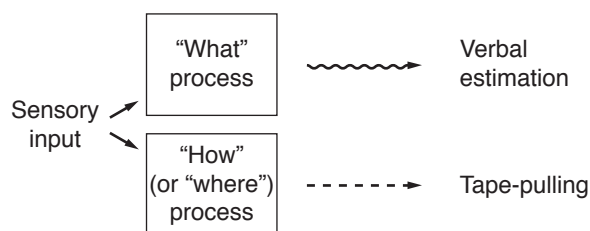


Figure 1: Schematic diagrams describing two theoretical models of perceptual processing for whole-body movements. (A) A unified process underlies whole-body movements. Responses in verbal estimation and tape-pulling are both controlled by the single process. Different shapes of the arrows representing output transformation indicate the possibility that the same output from the perceptual process can be transformed differently into verbal estimation and tape-pulling. (B) Two separate processes subserve whole-body movements. Responses in verbal estimation and tape-pulling are based on outputs from “what” and “how” processes, respectively.

## Experiment 1

In Experiment 1, healthy adult participants walked without vision under a normal walking condition. It has been shown that verbal estimation of visually specified distance and motoric responses to indicate it (such as tape-pulling) tend to yield similar, if not identical, patterns of response when they are performed by neurologically intact individuals under a normal viewing condition (Philbeck & Loomis, 1997; Philbeck et al., 2010). Thus, it was predicted that similar patterns of data would be observed in the two response modes. They would form a basis with which changes caused by altering sensory input into nonvisual walking (implemented in Experiment 2) were evaluated.

### Method

**Participants** Twelve students (6 males and 6 females, 18–27 years of age) at Cleveland State University volunteered in return for extra credit in psychology courses.

**Materials and Design** Participants walked linear distances of 1–6 m without vision at their own natural pace. A sighted experimenter walked with them while supporting their arm for safety reasons, but no assistance was provided for walking. Participants first walked 3- and 5-m distances, once apiece. They then walked distances of 1, 2, 4, and 6 m six times apiece in random order. These 26 trials were performed in one session. First two trials were used only to acquaint participants with the experimental procedure and therefore excluded from analyses. A long (at least 15 m) and loose measuring tape was used for tape-pulling. For each walked distance, participants made both a tape-pulling response and a verbal estimate. Thus, the experiment utilized a 2 (gender)  $\times$  2 (response mode)  $\times$  4 (walked distance) factorial design. Both response mode and walked distance were within-subject variables. Participants were run individually.

**Procedure** Prior to the experiment, participants were given an opportunity to practice pulling the tape. While standing still, they held one end of the measuring tape and the experimenter extended it so that it was parallel to the floor. A paper clip was attached to the tape at an arbitrary distance (generally in the range of 1–6 m) and participants pulled the tape until the paper clip reached their hands. They used a hand-over-hand motion to pull the tape, rather than pulling it by one hand while holding accumulated tape by the other hand. Participants viewed how the paper clip approached them as the tape was reeled in. This was repeated a few more times with different distances to the paper clip until participants felt acquainted with the tape-pulling procedure.

Following the practice session, participants wore a blindfold and hearing protectors (noise reduction rating: 21 dB) to obscure their vision and hearing. They were then guided to a nearby hallway in which the experiment was conducted. They were disoriented before taken to the hallway, and thus had no clear idea about where they were in the building during the experiment. This manipulation was included because prior

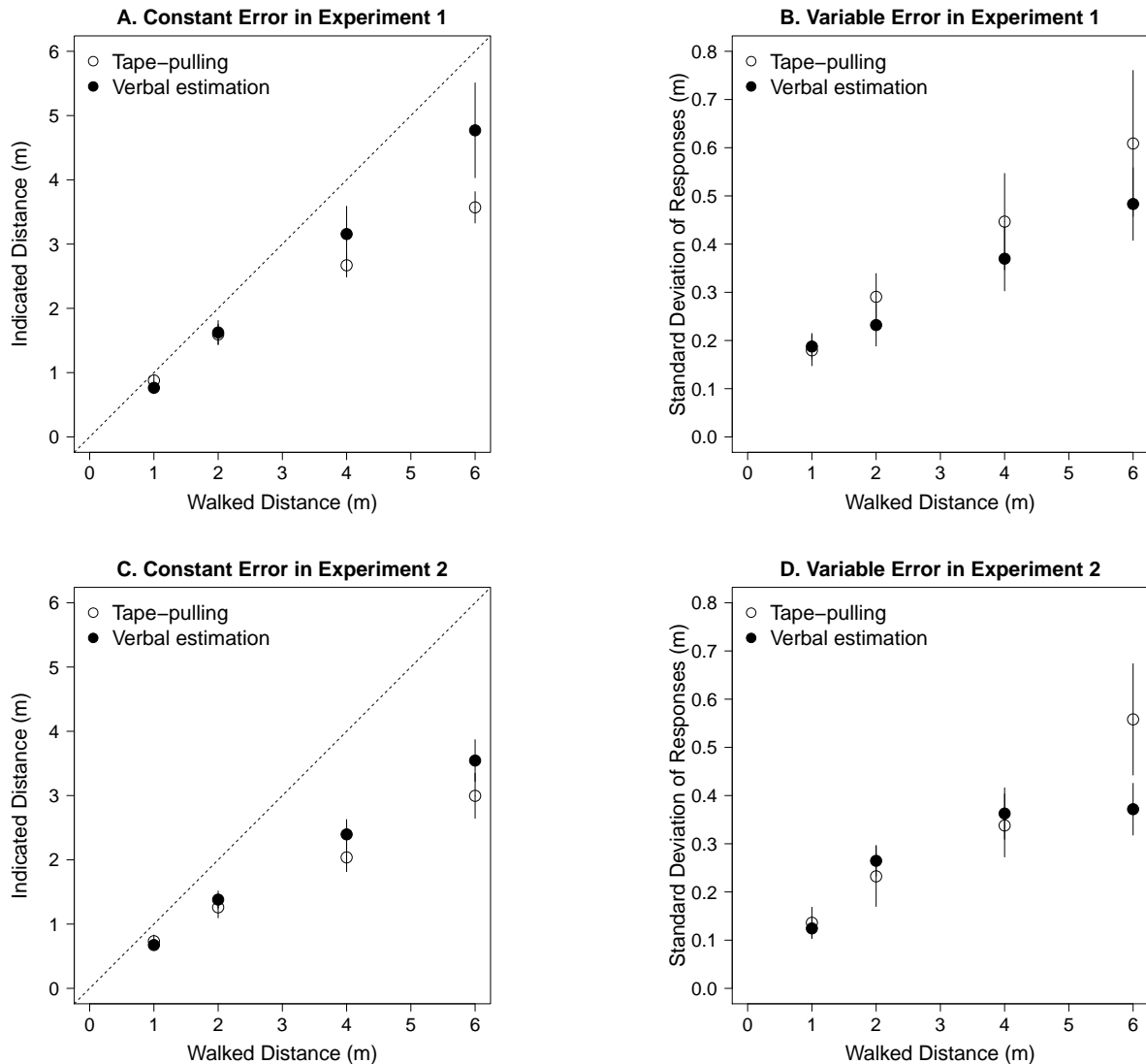


Figure 2: Mean constant and variable errors in estimation of walked distances in Experiments 1 and 2. They are shown as a function of response mode and walked distance. The dashed diagonal lines in panels A and C indicate accurate estimation of the walked distances (i.e., constant error = 0). Error bars represent  $\pm 1$  standard error of the mean.

knowledge about the environment could influence nonvisual distance perception (Philbeck & O'Leary, 2005).

In the beginning of each trial, participants stood at a fixed starting position and held one end of the measuring tape in their preferred hand. The rest of the tape was untangled and placed on the floor. In addition, participants were given a 5-digit random number and asked to remember it until they completed the trial. Because this number was typically retained in memory by rehearsal, this concurrent task was intended to interfere with subvocal counting of steps in walking and draws in tape-pulling that could otherwise be used to aid distance judgement. By discouraging participants from utilizing this counting strategy, their distance perception and its indicators were more based on somatosensory and vestibular information acquired during walking. Participants' accu-

racy in recalling this number ranged from 54.17% to 91.67% in both experiments (mean = 80.03%), suggesting that they attempted to memorize the numbers and it was sufficiently challenging to make the counting difficult. When participants were ready to start the trial, they proceeded straight ahead until they were told to stop at an appropriate distance. They then turned to face the starting position and pulled a length of the measuring tape so that it matched the walked distance. They marked the end of the pulled tape with their fingers and handed it to the experimenter. Subsequently, they gave a verbal estimate of the walked distance by using a distance unit of their choice. Most participants used ft, and a few used m or cm. They were encouraged to be as accurate as possible by using fractions (e.g., .25 m) when necessary. They were also instructed to derive the verbal estimate from the

walked distance, not from the length of tape they just pulled. Finally, participants repeated the 5-digit number, and were guided back to the starting position for the next trial. Participants walked in the same direction in all trials. No feedback was given to participants during the experiment.

**Data Analysis** Lengths of tape pulled and verbal estimates of walked distances were analyzed by calculating constant and variable errors. Constant errors represent how accurately participants indicated the walked distances by tape-pulling and verbal estimation. They were obtained by computing the mean amount of tape pulled and the mean verbal estimate for each walked distance. Variable errors characterized participants' consistency in responding to the same walked distance and were defined by standard deviations of six responses to each walked distance. Constant and variable errors were analyzed separately by split-plot analyses of variance (ANOVAs) with participants' gender as a between-subject factor and response mode (tape-pulling and verbal estimation) and walked distance (1, 2, 4, and 6 m) as within-subject factors. All *F*-tests conducted in this study were corrected for nonsphericity by using Greenhouse–Geisser epsilon when appropriate. Generalized eta squared ( $\eta_G^2$ ) values are reported as effect size statistics (Bakeman, 2005; Olejnik & Algina, 2003).

## Results

**Constant Error** Figure 2A shows mean constant errors in Experiment 1 as a function of response mode and walked distance. Participants generally made larger errors as they walked farther. In addition, tape-pulling tended to yield increasingly larger errors than verbal estimation as the walked distance increased. These observations were supported statistically by the main effect of walked distance,  $F(3,30) = 68.01, p < .001, \eta_G^2 = .60$ , and the interaction between response mode and walked distance,  $F(3,30) = 4.82, p = .049, \eta_G^2 = .056$ . A post-hoc contrast comparing the difference between the two response modes at 6 m against those at other distances had a large effect,  $F(1,11) = 5.17, p = .044, \eta_G^2 = .32$ , suggesting that tape-pulling and verbal estimation produced similar patterns of errors for the most part, except at the longest distance (6 m). Other effects and interactions did not reach statistical significance in the ANOVA.

**Variable Error** Figure 2B shows mean variable errors in Experiment 1 as a function of response mode and walked distance. Participants responded less consistently as the walked distance became longer. Although tape-pulling responses tended to be more variable than verbal estimates, the difference between them was not substantial. Consistent with these observations, only the main effect of walked distance was significant,  $F(3,30) = 16.45, p < .001, \eta_G^2 = .24$ .

## Discussion

As predicted, when participants walked under a normal condition and attempted to specify walked distance, largely comparable responses were yielded from tape-pulling and verbal estimation. These data constituted a baseline against which

findings from Experiment 2 would be evaluated: We manipulated sensory input into nonvisual walking in Experiment 2 to create an altered walking condition and investigated how constant and variable errors would change compared to those observed in Experiment 1.

## Experiment 2

To change sensory input into nonvisual walking, we asked participants to wear a heavy backpack during Experiment 2. Under this altered condition, it was expected that verbal estimation and tape-pulling would yield divergent patterns of response, if they were subserved by dissociable “what” and “how” processes. Given that these two measures mostly produced statistically indistinguishable data in Experiment 1, any differentiation between them would be indicative of the dissociation between conscious perception and action guidance in whole-body movements.

## Method

**Participants** Twelve participants (6 males and 6 females, 21–53 years of age) from the Cleveland State University community volunteered in return for monetary compensation or extra credit in psychology courses. None of them participated in Experiment 1. A new group of participants were recruited for Experiment 2 to avoid demand characteristics in the backpack manipulation. That is, if the same participants were asked to perform the tasks twice with and without the backpack, it would be relatively obvious to them that the backpack was intended to affect their performance.

**Materials, Design, Procedure, and Data Analysis** Experiment 2 was conducted in the same manner as in Experiment 1 except that each participant wore a backpack that weighed between 1/5 and 1/6 of their body weight. This weight range was adopted from previous studies in which the same backpack manipulation successfully induced measurable effects on distance perception (e.g., Proffitt, Stefanucci, Banton, & Epstein, 2003). To determine the appropriate weight of the backpack, each participant's body weight was measured before beginning the experiment. Participants put on the backpack when they were positioned at the starting position in the hallway for the first trial. They kept wearing it until all trials were completed. The backpack weight varied between 10.20 kg and 19.96 kg among participants (mean = 13.58 kg).

## Results

**Constant Error** Mean constant errors in Experiment 2 are plotted in Figure 2C as a function of response mode and walked distance. Compared to Experiment 1, constant errors observed in the present experiment, especially those yielded from verbal estimation, tended to be larger (i.e., participants showed a tendency to indicate the walked distances to be shorter). This change was more prominent in verbal estimation than in tape-pulling, resulting in comparable responses from them throughout the range of walked distance used in the present study. Consistent with this observation, only the

main effect of walked distance was significant,  $F(3,30) = 159.90, p < .001, \eta_G^2 = .68$ .

**Variable Error** Figure 2D shows mean variable errors in Experiment 2 as a function of response mode and walked distance. Variable errors in tape-pulling kept increasing as participants walked farther, just like variable errors in Experiment 1. On the other hand, variable errors in verbal estimation exhibited a qualitatively different pattern: Participants' responses at 6 m were as consistent as those at 4 m. This observation was supported statistically by the significant interaction between response mode and walked distance,  $F(3,30) = 4.98, p = .020, \eta_G^2 = .050$ , and a post-hoc contrast comparing the difference between tape-pulling and verbal estimation at 6 m with those at other distances,  $F(1,11) = 8.13, p = .016, \eta_G^2 = .42$ . The only other effect that was significant in the ANOVA was the main effect of walked distance,  $F(3,30) = 18.15, p < .001, \eta_G^2 = .29$ .

## Discussion

The backpack manipulation exerted two noticeable effects in the present experiment: (1) It caused greater underestimation of walked distance (i.e., larger constant error), especially in verbal estimation at longer distances, which removed the difference between the two measures observed in Experiment 1; and (2) variable errors in verbal estimation did not show further increase beyond the 4-m distance, while those in tape-pulling showed steady increase as a function of walked distance as in Experiment 1. Given that the overall pattern of constant errors exhibited by tape-pulling and verbal estimation was mostly the same as that observed in Experiment 1, much importance may not be attributable to the change in constant errors in Experiment 2. However, the fact that altered sensory input into nonvisual walking only affected response consistency in verbal estimation suggests that processes underlying verbal estimation and tape-pulling are dissociable. Thus, a sign of dissociation between conscious perception and action guidance within whole-body movements was found in the present experiment.

## General Discussion

The present study was designed to explore whether dissociable processes for conscious perception and action guidance can be found in whole-body movements, in which the presence of vestibular information creates a unique perceptual condition. To that end, nonvisually perceived walked distance was assessed by two modes of response: One required explicit recognition of the walked distance (verbal estimation) and the other was primarily controlled by a motor action (tape-pulling). When sensory input into nonvisual walking was altered by having participants carry additional weight, variability in verbal estimates was markedly modulated. On the other hand, variability in tape-pulling responses largely remained unchanged. This suggests that information about walked distance is processed with qualitatively different levels of precision for verbal estimation and tape-pulling,

showing a sign of dissociation between the process underlying conscious perception and that subserving action guidance in whole-body movements. This result builds upon previous findings that the same dissociation is present in visual, auditory, and somatosensory systems (e.g., Belin & Zatorre, 2000; Dijkerman & de Haan, 2007; Milner & Goodale, 1995), supporting a claim that it is a general principle with which perceptual systems are organized.

Although the present study successfully showed the initial evidence for dissociation between conscious perception and action guidance in whole-body movements, several questions are still unanswered. Most notably, it remains to be seen whether the pattern shown by variable errors in Experiment 2 is extendable to longer walked distances. A follow-up study should be carried out by expanding the range of walked distance. The follow-up study should also include a larger number of participants so that Experiments 1 and 2 can be statistically compared; such an analysis was not possible in the present study due to the lack of statistical power. Furthermore, the fact that underestimation of walked distance was exacerbated by the backpack manipulation in Experiment 2 was somewhat counterintuitive: Considering that those participants had to expend a greater amount of energy for walking a given distance, it may be more reasonable to expect that they would judge the distance to be longer, not shorter (Proffitt et al., 2003). Similarly, it is not readily clear why the additional weight increased, not decreased, precision of verbal estimates at the 6-m distance. Further research should be carried out to fully understand these important details.

## Acknowledgments

This study was supported in part by Cleveland State University Research and Creative Achievement Award to N.Y.

## References

- Aglioti, S., Beltramello, A., Bonazzi, A., & Corbetta, M. (1996). Thumb-pointing in humans after damage to somatic sensory cortex. *Experimental Brain Research*, 109, 92–100.
- Anourova, I., Nikouline, V. V., Ilmoniemi, R. J., Hotta, J., Aronen, H. J., & Carlson, S. (2001). Evidence for dissociation of spatial and nonspatial auditory information processing. *NeuroImage*, 14, 1268–1277.
- Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, 37, 379–384.
- Belin, P., & Zatorre, R. J. (2000). 'What', 'where' and 'how' in auditory cortex. *Nature Neuroscience*, 3, 965–966.
- Bottini, G., Paulesu, E., Sterzi, R., Warburton, E., Wise, R. J. S., Vallar, G., et al. (1995). Modulation of conscious experience by peripheral sensory stimuli. *Nature*, 376, 778–781.
- Bottini, G., Sterzi, R., Paulesu, E., Vallar, G., Cappa, S. F., Erminio, F., et al. (1994). Identification of the central vestibular projections in man: A positron emission tomog-

- raphy activation study. *Experimental Brain Research*, 99, 164–169.
- Dijkerman, H. C., & de Haan, E. H. F. (2007). Somatosensory processes subserving perception and action. *Behavioral and Brain Sciences*, 30, 189–239.
- Fasold, O., von Brevern, M., Kuhberg, M., Ploner, C. J., Villringer, A., Lempert, T., et al. (2002). Human vestibular cortex as identified with caloric stimulation in functional magnetic resonance imaging. *NeuroImage*, 17, 1384–1393.
- Ferrè, E. R., Bottini, G., & Haggard, P. (2011). Vestibular modulation of somatosensory perception. *European Journal of Neuroscience*, 34, 1337–1344.
- Foley, J. M. (1977). Effect of distance information and range on two indices of visually perceived distance. *Perception*, 6, 449–460.
- Goodale, M. A., Milner, A. D., Jakobson, L. S., & Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature*, 349, 154–156.
- Guldin, W. O., & Grüsser, O.-J. (1998). Is there a vestibular cortex? *Trends in Neurosciences*, 21, 254–259.
- Halligan, P. W., Hunt, M., Marshall, J. C., & Wade, D. T. (1995). Sensory detection without localization. *Neurocase*, 1, 259–266.
- Kaas, J. H., & Hackett, T. A. (1999). 'What' and 'where' processing in auditory cortex. *Nature Neuroscience*, 2, 1045–1047.
- Kammers, M. P. M., van der Ham, I. J. M., & Dijkerman, H. C. (2006). Dissociating body representations in healthy individuals: Differential effects of a kinaesthetic illusion on perception and action. *Neuropsychologia*, 44, 2430–2436.
- Maeder, P. P., Meuli, R. A., Adriani, M., Bellmann, A., Fornari, E., Thiran, J.-P., et al. (2001). Distinct pathways involved in sound recognition and localization: A human fMRI study. *NeuroImage*, 14, 802–816.
- Marcel, A. (2003). The sense of agency: Awareness and ownership of action. In J. Roessler & N. Eilan (Eds.), *Agency and self-awareness: Issues in philosophy and psychology*. New York: Oxford University Press.
- Milner, A. D., & Goodale, M. A. (1995). *The visual brain in action*. New York: Oxford University Press.
- Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, 8, 434–447.
- Paillard, J. (1999). Body schema and body image: A double dissociation in deafferented patients. In G. N. Gantchev, S. Mori, & J. Massion (Eds.), *Motor control: Today and tomorrow*. Sofia, Bulgaria: Academic Publishing House.
- Paillard, J., Michel, F., & Stelmach, G. (1983). Localization without content: A tactile analogue of 'blind sight'. *Archives of Neurology*, 40, 548–551.
- Philbeck, J. W., & Loomis, J. M. (1997). Comparison of two indicators of perceived egocentric distance under full-cue and reduced-cue conditions. *Journal of Experimental Psychology: Human Perception and Performance*, 23, 72–85.
- Philbeck, J. W., & O'Leary, S. (2005). Remembered landmarks enhance the precision of path integration. *Psicológica*, 26, 7–24.
- Philbeck, J. W., Woods, A. J., Kontra, C., & Zdenkova, P. (2010). A comparison of blindpulling and blindwalking as measures of perceived absolute distance. *Behavior Research Methods*, 42, 148–160.
- Proffitt, D. R., Stefanucci, J., Banton, T., & Epstein, W. (2003). The role of effort in perceived distance. *Psychological Science*, 14, 106–112.
- Rauschecker, J. P. (1998). Parallel processing in the auditory cortex of primates. *Audiology and Neuro-Otology*, 3, 86–103.
- Reed, C. L., Klatzky, R. L., & Halgren, E. (2005). What vs. where in touch: An fMRI study. *NeuroImage*, 25, 718–726.
- Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., & Rauschecker, J. P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nature Neuroscience*, 2, 1131–1136.
- Rossetti, Y., Rode, G., & Boisson, D. (1995). Implicit processing of somaesthetic information: A dissociation between where and how? *NeuroReport*, 6, 506–510.
- Sathian, K., Lacey, S., Stilla, R., Gibson, G. O., Deshpande, G., Hu, X., et al. (2011). Dual pathways for haptic and visual perception of spatial and texture information. *NeuroImage*, 57, 462–475.
- Schwarz, D. W. F., & Fredrickson, J. M. (1971). Rhesus monkey vestibular cortex: A bimodal primary projection field. *Science*, 172, 280–281.
- Sittig, A. C., Denier van der Gon, J. J., Gielen, C. C. A. M., & van Wijk, A. J. M. (1985). The attainment of target position during step-tracking movements despite a shift of initial position. *Experimental Brain Research*, 60, 407–410.
- Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior*. Cambridge, MA: MIT Press.
- Van Boven, R. W., Ingeholm, J. E., Beauchamp, M. S., Bikle, P. C., & Ungerleider, L. G. (2005). Tactile form and location processing in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 12601–12605.
- Westwood, D. A., & Goodale, M. A. (2003). A haptic size-contrast illusion affects size perception but not grasping. *Experimental Brain Research*, 153, 253–259.

# The Influence of Risk Aversion on Visual Decision Making

**Ruixin Yang (r4yang@cs.ucsd.edu)**

Department of Computer Science and Engineering  
9500 Gilman Drive  
La Jolla, CA 92093 USA

**Garrison W. Cottrell (gary@ucsd.edu)**

Department of Computer Science and Engineering  
9500 Gilman Drive  
La Jolla, CA 92093 USA

## Abstract

The ability to decide between multiple fixation targets in complex visual environments is essential for our survival. Evolution has refined this process to be both rapid and cheap, allowing us to perform over 100,000 saccades a day. Previous models for visual decision making have focused on maximizing reward magnitude or expected value ( $EV = \text{probability of reward} \times \text{magnitude of reward}$ ). However, such methods fail to incorporate utility, or happiness derived from reward, optimizing strictly on nominal reward values. We propose an alternative model for visual decision making, maximizing utility as opposed to value under the assumption of a decreasing marginal utility curve. To test our model, we asked 10 UCSD graduate students to participate in an eyetracking experiment where they choose between different fixation targets presented on a brief display. The reward for each target was generated from fixed, predetermined distributions with different variance that was initially unknown to the subjects. The subjects were asked to maximize their reward for each test session within the experiment. Comparing our results with expected value and reward optimizing hedge algorithms, we show that utility-based models more accurately reflect human behavior in visual decision making tasks.

**Keywords:** Visual decision making; risk aversion; utility theory; reward.

## Introduction

Target selection is a complex optimization task that the human visual system must complete thousands of times per day. Assuming a probabilistic, stationary distribution for reward, the problem is directly reducible to the multi-armed bandit problem (Lai & Robbins, 1985; Freund & Schapire, 1997; Auer et al., 2003; Chaudhuri, Freund & Hsu, 2009). Despite the complexity of the problem, an efficient, low-cost algorithm is necessary to permit rapid saccades to be made in the noisy, low-resolution perceptual environment. Previous attempts to model visual decision making involve a probabilistic, reward magnitude framework where the probability of fixation is weighted based on the expected value (EV), defined as the probability of reward multiplied by the magnitude of reward (Milstein & Dorris, 2007; Navalpakkam et al., 2010; Platt & Glimcher, 1999). Depending on the approach, the definition of the probability of reward can be taken either from the traditional economic context as the probability of obtaining a reward upon target

fixation (Milstein & Dorris, 2007; Milstein & Dorris, 2011; Platt & Glimcher, 1999) or from the perspective of noisy sensors, modifying the probability of a target's location given the noise (Navalpakkam et al., 2010). Alternative approaches that have achieved comparable accuracy to expected value strategies include a study on rhesus monkeys by Milstein and Dorris (2011), where they show reward magnitude alone may explain saccadic preparation and reaction time data.

Despite the success of expected value models, there are many real life examples where behavior does not maximize expected value. For example, for many types of large investment (e.g. automobile, home, healthcare), there is a set of insurance policies to reduce risk. Insurance companies sustain themselves by making a small profit while reducing risk for consuming individuals. If it were the case that every individual viewed reward maximization as maximizing their expected value, these institutions would no longer be profitable and would cease to exist. However, there are many instances where individuals are willing to disproportionately sacrifice some of their assets to reduce the probability of an extremely undesirable outcome. As a result, insurance is often viewed as mutually beneficial and is encouraged in many situations. Bernoulli (1738/1954) provided the first examples of deviation from expected value behavior, stating that decisions should be based on an individual's current wealth, with less wealthy individuals being more averse to risky decisions. Brocas and Carrillo (2009) presented a simple illustration where two perfectly rational individuals may come to opposite conclusions in situations of uncertainty due to differences in utility preferences. The goal of our experiment is explore the concept of utility maximization and risk aversion in the context of visual decision making, where decisions are made in quick succession, without much time for the conscious evaluation of value.

We provide an alternative model for visual decision making that attempts to maximize utility, or happiness derived from reward, rather than expected value. In the following sections, we will formally define our model, as well as provide a mathematical justification for why expected value fails to account for behavior with respect to most types of reward. We also compare the predictions of our model with those made by the expected value model and

two well-known machine learning algorithms. Our results show that a risk averse, utility maximization model performs significantly better than the other models at describing human eye movement behavior under all experimental conditions.

### Model Description

Assume that there are  $n$  targets:  $x_i$ , where  $i = 1, \dots, n$ .

- Let  $\pi_{i,l}$  be the probability of encountering target  $x_i$  at location  $l$ .
- Let  $v_i$  be the value for fixating on target  $x_i$ .

The expected value for fixating at location  $l$  can be calculated as follows:

$$EV(l) = \sum_i \pi_{i,l} v_i. \quad [1]$$

However, value alone is often a poor measurement for reward (Bernoulli, 1954; von Neumann & Morgenstern, 1953). This is due to the fact that people tend to have a decreasing marginal utility for most types of reward.

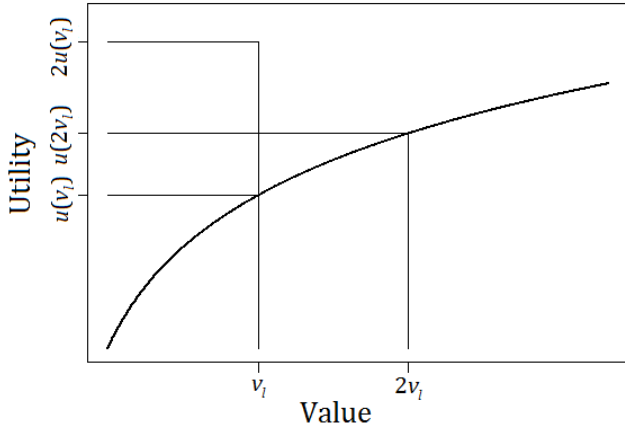


Figure 1: Individuals tend to have a decreasing marginal utility for most types of reward due to priorities in reward allocation (see text). Note that for decreasing marginal utility, the utility derived from doubling the value of reward is less than twice the utility of the original value:  $u(2v_i) < 2u(v_i)$ .

Intuitively, decreasing marginal utility arises from how consumption of reward is allocated. The first units of a reward such as money tend to be spent towards essential necessities, while later units tend to be used on luxury goods. Similar arguments could be made for other rewards such as food, shelter, and material goods. Figure 1 depicts a standard decreasing marginal utility curve.

One implication of holding a decreasing marginal utility curve is that the risk averse decision will frequently maximize utility. For example, Figure 2 shows that under the decreasing marginal utility assumption, an individual will always prefer an action that generates a guaranteed

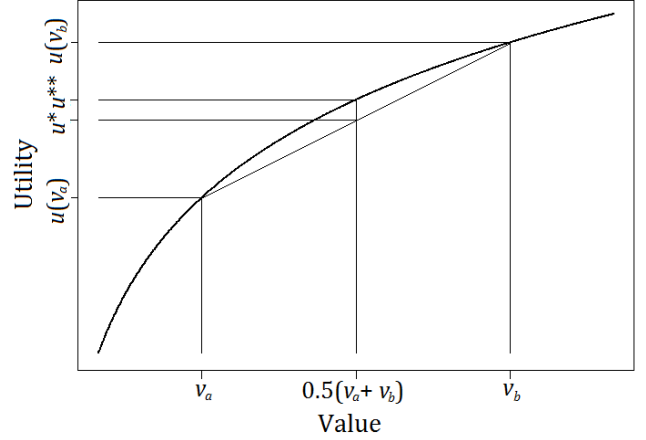


Figure 2: A property of the decreasing marginal utility curve is that individuals should demonstrate risk averse behavior in their decision making. Despite having identical expected values, a fair chance between  $v_a$  and  $v_b$  yields  $u^*$ , which is less than the utility provided by  $0.5(v_a + v_b)$ , which yields  $u^*$ . Note that the choice between  $v_a$  and  $v_b$  has greater variance than the single point,  $0.5(v_a + v_b)$ .

reward to one which provides a fair gamble between two outcomes that preserves expected value. This result generalizes: individuals with a decreasing marginal utility for reward should always prefer the choice with lower variance, given options of equal expected value.

Replacing value ( $v_i$ ) with utility  $u(v_i)$ , we obtain the following equation for expected utility:

$$EU(l) = \sum_i \pi_{i,l} u(v_i). \quad [2]$$

The goal of the subject is to find the location in the scene that contains the target which maximizes expected utility. Therefore, the objective function for maximizing utility is:

$$f = \operatorname{argmax}_l \sum_i \pi_{i,l} u(v_i). \quad [3]$$

To model  $u(v_i)$ , we note that it is monotonically increasing, but has a decreasing slope leading to diminishing returns with increased reward. For our model, we represent the curve using a natural logarithm because of its simplicity and because it shares the same properties as the utility curve. It is important to note that no two individuals hold the same utility function for reward, and the natural logarithm reflects a hypothetical utility function of the average individual.

The new objective function thus becomes:

$$f = \operatorname{argmax}_l \sum_i \pi_{i,l} \ln(cv_i + 1) \quad [4]$$

where  $c$  is a constant value reflecting the magnitude of risk aversion. In our experiment,  $v_i$  is a value that must be learned by the subject for each experimental phase.



Based on feedback, in each trial we approximate  $u(v_i)$  as a weighted average across past observed rewards,  $\ln(cr_1 + 1), \ln(cr_2 + 1), \dots, \ln(cr_t + 1)$ . Similarly,  $v_i$  is approximated by the weighted average of the observed rewards sequence  $(r_1, r_2, \dots, r_t)$ . Since the targets in our experiment are visually dissimilar,  $\pi_{i,l}$  becomes close to 0 or 1 depending on the target. This allows us to approximate  $\pi_{i,l} \in \{0, 1\}$  and simplifies our calculation.

## Experimental Methods

### Subjects

Ten naive subjects participated in the experiment after providing informed consent. All subjects were right-handed graduate students from the University of California, San Diego Computer Science and Engineering department.

### Experimental Procedure

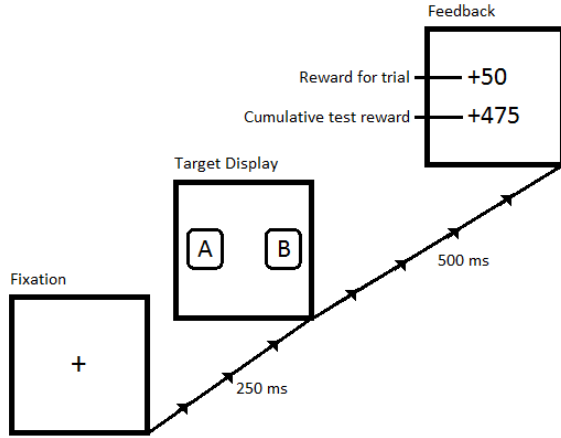


Figure 3: The basic experimental setup. At start of every trial, subjects were asked to hold a central fixation for 250 ms before being presented the target display. Once the targets are presented, the subject must make a saccade to one of the targets within 500 ms, when feedback is displayed. If no decision is made within the 500 ms, the subject receives no reward for the trial and is notified that they must make their decision faster. Subjects are permitted to spend as long as they wish on the feedback screen prior to starting the next trial to make adjustments to their strategy.

We model our experiment using similar parameters to those used in Navalpakkam et al. (2010). At the beginning of every trial, subjects were asked to indicate they were ready by fixating at a center fixation point and pressing the “enter” key on the keyboard. Each trial begins with a central fixation ‘+’ presented for 250 milliseconds. Subjects indicate their choice by saccading to one of the targets. Then, subjects were presented with an image containing two targets labeled ‘A’ and ‘B’ for 500 milliseconds. Figure 3 provides an illustrated description of each trial.

Targets in the experiment appeared at 7 degrees eccentricity from the central fixation point, and were horizontally aligned. The target stimuli were 1.8 degree in height and was each encompassed by a  $3.6 \times 3.6$  degrees square border. The location where each target appeared was randomly generated from trial to trial. Subjects viewed the display on a 19-inch cathode ray tube (CRT) monitor at a distance of 30 inches from the screen.

The experiment consists of an initial training phase and three test phases of 50 trials each. In the training phase, the rewards for the two targets were drawn from discrete uniform distributions,  $U[0, 50]$  and  $U[50, 100]$ . Subjects were required to learn which target yielded the higher reward and fixate on that target for at least 75 percent of the training trials before being allowed to proceed to the test phase. This was done to ensure the subject was accustomed to making quick and accurate saccades while wearing the eyetracking device. Subsequently, each test phase consisted of two targets of equal mean, but different variance.

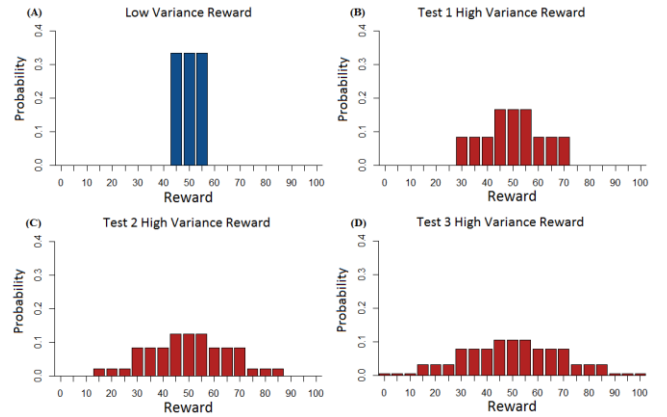


Figure 4: The distribution of target reward for each test. Of the two targets, the lower variance target (4A) was kept constant for all three tests, while the distribution of the higher variance targets (4B), (4C), and (4D) changed for Tests 1, 2, and 3 respectively.

The reward for each target was generated from the distributions described in Figure 4. For each test, the higher variance target’s distribution was adjusted, while the lower variance distribution was kept constant across all of the tests. Subjects did not know the identity or probability distribution of the target at the beginning of each trial or test phase in the experiment. They were instructed to learn the distributions and maximize their reward for each test phase.

We used an Eyelink 1000 eyetracker from SR Research to record the subjects’ eye movements. At the beginning of each test, the eyetracker was recalibrated using a nine-point calibration across the edge and center of the display.

### Modeling Procedure

Performance for each model was measured by the percentage of trials in which the model matches the human

fixation decision. For each trial, the utility-based risk averse model picked the location that maximized the objective function described in Equation [4], while the expected value model maximized  $\text{argmax}_l \sum_i \pi_{i,l} v_i$ . We chose the constant  $c$  from Equation [4] by binary search across the  $[0, 10]$  interval, finding the value of  $c$  that best predicted the human subject data. However, the results were not sensitive to the exact choice of  $c$ .

Every model prediction was compared with the decision made by the subject. Updates to both models (i.e. running averages to  $v_i$  and  $u(v_i)$ ), were done based on the experimental feedback provided to the subjects for each individual trial, as if the model had made the same choice as the subject. In addition to the greedy algorithms, where the model decision is always the one that maximized the objective function, we compared the performance of three policies that include exploration of less-valued alternatives:  $\epsilon$ -greedy, decreasing  $\epsilon$ , and softmax (Sutton & Barto, 1998). We tested epsilon and softmax temperature values of 0.05, 0.10, and 0.20. The exploration algorithms may be summarized subsequently as follows:

#### **$\epsilon$ -greedy:**

**Initialize  $\epsilon$**

*For every trial*

**Generate** random number  $r \in [0, 1]$

**If**  $r > \epsilon$

Take action maximizing the objective function

**Else**

Randomly generate action from uniform distribution

#### **Decreasing $\epsilon$ :**

**Initialize  $\epsilon$ , dec\_rate** =  $\epsilon / (\text{num\_trials} - 1)$

*For every trial*

**Generate** random number  $r \in [0, 1]$

**If**  $r > \epsilon$

Take action maximizing the objective function

**Else**

Randomly generate action from uniform distribution

**Update  $\epsilon = \epsilon - \text{dec\_rate}$**

#### **Softmax:**

**Initialize  $\tau$**

*For every trial*

**Generate** action  $i$  based on probability density:

$$\frac{e^{Q_t(i)/\tau}}{\sum_{j=1}^2 e^{Q_t(j)/\tau}}$$

where  $Q_t(i)$  is the running average of reward for choosing  $i$ .

Aside from comparing prediction performance against expected value, we also compared our results against two well-known machine learning algorithms, Hedge (Freund & Schapire, 1997) and Normal-Hedge, (Chaudhuri, Freund & Hsu, 2009) from the multi-arm bandit problem literature. Both algorithms are designed to maximize reward, given a single parameter value ( $\beta$  for Hedge and  $c_t$  for Normal-Hedge). Before we continue, it is

important to take note of one subtle difference in the objective function of the multi-armed bandit problem with our problem. The objective function of the multi-armed bandit minimizes regret (defined as the difference in reward between the ideal and the chosen action) as opposed to maximizing accumulated reward. To address this, we linearly transformed the reward obtained to range from  $[0, 1]$  and compute the loss of each action as the difference,  $1 - \text{reward}$ . In our Hedge implementation, we tested the entire range of temperature values  $[0, 1]$  in increments of 0.05. We find the best temperature setting to be  $\beta = 0.05$ , and use it for our analysis. For Normal-Hedge, the algorithm is self-adapting around a variable constraint  $c_t$  (note this variable is unrelated to the variable  $c$  from Equation [4]). We solve for  $c_t$  using line search as recommended by the authors. A detailed description of the algorithm as well as the proof on performance bounds may be found for Hedge in Freund & Schapire (1997) and Normal-Hedge in Chaudhuri, Freund & Hsu (2009).

In all our models, we excluded two subjects from our experiment as their data lay beyond two standard deviations from the mean number of saccades to the lower variance target. Note that we did not purposely remove risk seekers as this is equivalent to removing data with respect to the higher variance target due to the fact that there are two targets. Of these, one of the subjects was removed because he systematically fixated only at the target that appeared on the left side of the display, regardless of the identity of the target.

To address potential location bias concerns in making saccades, we recruited only right-handed subjects for our study. In addition, we tested our models with and without two location-based prior probabilities obtained from subject responses. The priors were the probability of fixating at each potential target location, and the probability of returning, given a previous saccade to the same location on the previous trial. In all of our models, there was no significant change in performance when we incorporated the priors under a Bayesian setting. For this and all model comparisons used in this paper, we used a paired t-test to compare the performance between models.

## **Results**

### **Behavioral Data**

The results of the human experiment are presented in Figure 5. For each test phase, we maintained a record of the number of saccade decisions to each target. The reward distribution for higher variance target in Tests 1-3 shared the same mean, but differed in variance as shown in Figure 4 (B-D) respectively, so the utility for these choices will increase. The lower variance target maintained the same distribution for all three tests. In Test 3, the subjects showed significantly risk averse behavior ( $p = 0.0321$ ), choosing the lower variance target 54.8 percent of the time. As the difference in variance between the targets decreased, subjects

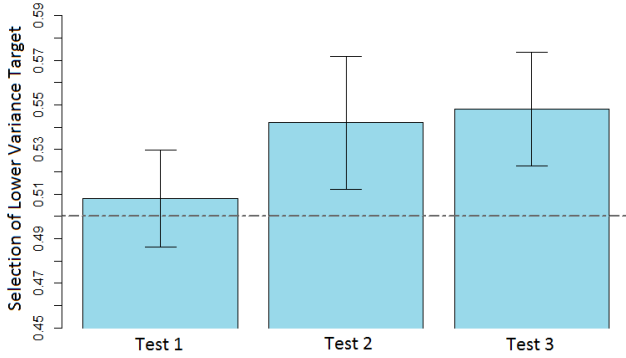


Figure 5: The results from the human experiment. The gray line represents indifference between the two targets. The reward distribution for the higher variance target in Tests 1-3 shared the same mean, but differed in variance as shown in Figure 4 (B-D) respectively. Of the three tests, subjects showed significant risk averse behavior in Test 3 (the test with greatest difference in variance between the two targets), where they chose the lower variance target 54.8% of the time.

became increasingly indifferent between the two targets. Subjects in Test 2 chose the lower variance target 54.2 percent of the time ( $p = 0.0861$ ), while subjects in Test 1 chose the lower variance target 50.8 percent of the time ( $p = 0.621$ ).

### Choosing a Value for $c$

Recall from Equation [4] in the model description where a constant,  $c$ , was included to allow for fine-tuning of the magnitude of risk-averse preferences. One interesting, and perhaps surprising result is that most reasonable settings of  $c$  outperform the expected value model in predicting human behavior. For this reason, we simply chose a local maximum using a binary search across positive values of  $c$ . Exceptions to this include  $c = 0$  (when the strategy reduces to random) and extremely large values of  $c$  (when most rewards share approximately the same value).

### Comparison with Expected Value

We simulated the expected value and utility-based risk averse strategies for 100 simulations ( $c = 2.48$ ) using greedy,  $\epsilon$ -greedy, decreasing  $\epsilon$ , and softmax exploration functions (Sutton & Barto, 1998). Our results show that all non-greedy algorithms perform significantly worse than their greedy counterparts ( $p < 0.001$ ). Simple exploration strategies yield poor performance because although they are capable of accurately capturing the probability of exploration, they fail at correctly predicting the trials on which they occur. As a result, since the probability of performing a reward maximizing action for any given trial is greater than the probability of

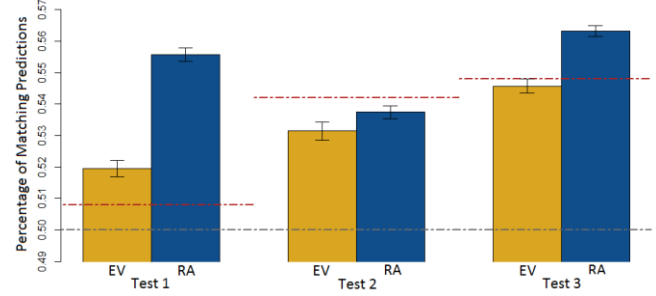


Figure 6: A comparison of the model fits between the greedy expected value and greedy utility-based risk averse (RA) strategies in predicting human data. In all three test conditions, the utility-based risk averse strategy significantly outperformed the expected value (EV) strategy for  $T = 100$  simulations ( $p < 0.001$ ). The gray dotted line represents chance performance, while the red dotted line represents fit obtained by always choosing the lower variance target (an omniscient model). Given a limited history of reward, the subject may choose the higher variance target as a result of a greedy action. Performance above the red dotted line suggests that the algorithm was fairly accurate its prediction of when the subject chose to take such greedy action as opposed to exploring the other option.

exploration, the greedy version of the algorithm will significantly outperform their exploration counterpart.

Figure 6 compares the utility-based risk averse strategy with the expected value strategy in predicting human behavior. The results show that although both models perform significantly above chance, maximizing across utility significantly outperforms value maximization ( $p < 0.001$ ) for all three test conditions. The red dotted line provides a benchmark for how much performance may be obtained from a strategy defined by choosing only the lower variance target. Note that this is an overprediction, since at the beginning of each test phase, the subject does not know which of the two targets holds lower variance (or even the value of their reward).

### Comparison with Hedge

We compare the fit of the utility-based risk averse strategy with two well-known algorithms for solving the multi-armed bandit problem from machine learning. We choose to implement hedge algorithms over an alternate strategy, Exp4 (Auer et al., 2003) due to its superior performance under conditions where the reward distribution is fixed (the reward probability distributions are fixed for our experiment as shown in Figure 4). We test twenty values for the temperature value ( $\beta$ ) in Hedge and present the result for the optimal setting,  $\beta = 0.05$ . For Normal-Hedge, we find the constraint,  $c_t$ , via line search as recommended by the authors. Figure 7 shows a summary of our results. In all three test conditions, the

utility-based risk averse strategy significantly outperforms hedge algorithms under their optimal settings ( $p < 0.001$ ).

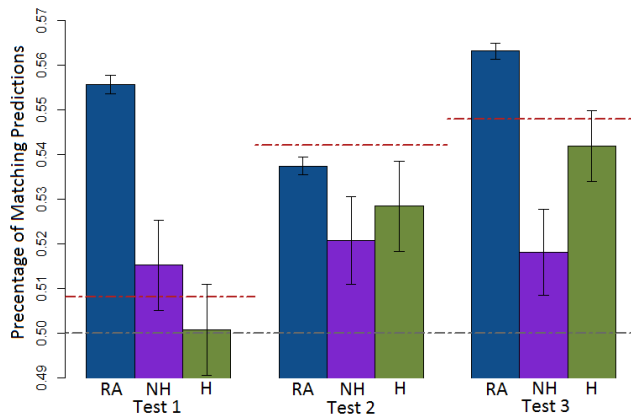


Figure 7: The results comparing the performance between the utility-based risk averse strategy with Hedge (H) and Normal-Hedge (NH) algorithms on predicting human data. In all three test conditions, the utility-based risk averse strategy significantly outperformed the Hedge and Normal-Hedge strategies for  $T = 100$  simulations ( $p < 0.001$ ). The gray dotted line represents chance performance, while the red dotted line represents performance obtained from only choosing the lower variance target.

## Discussion and Future Work

Our work shows that by constructing models from a utility maximization standpoint, we are able to make predictions regarding human behavior that would otherwise be impossible in situations involving risk. Previous models in saccadic prediction involved a direct integration of the probability and magnitude of reward, ignoring risk derived from variance in reward distributions. In this paper, we present evidence suggesting the importance of such parameters when modeling visual decision making. Our findings show that under conditions of uncertainty, the human visual system takes a risk averse approach, taking account of the variance of the reward distribution in addition to the mean.

However, the current utility-based risk averse model does not address all questions that arise with the incorporation of risk. In particular, the work does not address issues raised from prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992; Kusev et al., 2009). For example, in the context of the experiment, there is no loss associated with viewing any target, and thus the asymmetry between loss and gain perception could not be modeled. Likewise, there are many situations where risk seeking behavior is exhibited and is the utility optimizing choice. While both conditions may arise in vision, prospect theory could not be modeled under the current experimental framework, and risk seeking behavior would require a change in the shape of the utility

function. However, despite these limitations, we believe that the current work presents a starting point for analyzing visual decision making under uncertainty.

## Acknowledgements

We would like to thank the members of the GURU research lab, and the Perceptual Expertise Network for comments and feedback on this work. The work was supported in part by NSF Grant #SBE0542013.

## References

- Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. (2003). The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32, 48-77.
- Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, 22, 23-36.
- Brocas, I., & Carrillo, J. (2009). Information acquisition and choice under uncertainty. *Journal of Economics and Management Strategy*, 18, 423-455.
- Chaudhuri, K., Freund, Y., & Hsu, D. (2009). A parameter-free hedging algorithm. *Advances in Neural Information Processing Systems*, 22, 297-305.
- Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119-139.
- Kahneman, D. & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47, 263-291.
- Kusev, P., van Schaik, P., Ayton, P., Dent, J., & Chater, N. (2009). Exaggerated risk: prospect theory and probability weighting in risky choice. *JEP:LMC*, 35, 1487-1505.
- Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6, 4-22.
- Milstein, D., & Dorris, M. (2007). The influence of expected value on saccadic preparation. *Journal of Neuroscience*, 27, 4810-4818.
- Milstein, D., & Dorris, M. (2011). The relationship between saccadic choice and reaction times with manipulations of target value. *Frontiers in Neuroscience*, 5.
- Navalpakkam, V., Koch, C., Rangel, A., & Perona, P. (2010). Optimal reward harvesting in complex perceptual environments. *PNAS*, 107, 5232-5237.
- Platt, M., & Glimcher, P. (1999). Neural correlations of decision variables in parietal cortex. *Nature*, 400, 233-238.
- Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *Journal of risk and uncertainty*, 5, 297-323.
- von Neumann, J., & Morgenstern, O. (1953). *Theory of games and economic behavior*. Princeton University Press, Princeton, NJ.

# Splitting Visual Focal Attention? It Probably Depends on Who You Are

Jit Yong Yap (yapjityong@gmail.com)

Department of Psychology, National University of Singapore  
9 Arts Link, Singapore 117570

Stephen Wee Hun Lim (psylimwh@nus.edu.sg)

Department of Psychology, National University of Singapore  
9 Arts Link, Singapore 117570

## Abstract

Research evidence now suggests that the deployment of multiple attentional foci in non-contiguous locations (i.e., splitting visual focal attention) is possible under some circumstances. However, the exact circumstances under which focal attention might ‘split’ have not been well understood. The present study is the first in the literature to examine the possibility that ecological differences arising from our increasingly media-saturated environment may result in individual differences in the capacity to demonstrate splitting focal attention. Results suggest a significant relationship between the behavioural preference for consuming multiple media forms simultaneously and the capacity to employ a split mode of attention.

**Keywords:** Ecological differences; splitting focal attention; media multitasking

## Introduction

In visual attention research, whether the focus of attention can be divided is an issue of debate. Several authors espouse the view that the focus of attention is unitary and indivisible in nature (e.g. Eriksen & Yeh, 1985; McCormick & Klein, 1990; Pan & Eriksen, 1993). In order to account for the processing of multiple visual stimuli in different spatial locations, two different theories have been proposed. The *serial shifting theory of attention* suggests that the focus of attention rapidly shifts between different locations (Eriksen & Eriksen, 1974; Posner, 1980), whereas the *zoom lens theory of attention* suggests that the focus of attention is adjusted in size to accommodate multiple locations (Eriksen & St. James, 1986).

However, there is a growing body of evidence which supports the possibility of split attentional foci, or allocation of attention to noncontiguous regions (for review, see Jans, Peters, & De Weerd, 2010; cf. Cave, Bush & Taylor, 2010). For instance, Awh and Pashler (2000) presented participants with a 5 x 5 stimulus array and tasked them with identifying two targets following the onset of cues (validity = 80%). Results showed a strong accuracy advantage at cued locations that did not apply to targets appearing between cued locations, which suggest a divided focus of attention. In their appraisal of the current literature, Cave, Bush and Taylor (2010) concluded “the weight of evidence suggests that some form of split attention is possible in some circumstances”.

With the focus on establishing the possibility of split attention, there has been little research into the exact circumstances under which split attention might arise. This was the crux of a recent study by Lim and Lee (2011), who hypothesized that adoption of a unitary or split mode of attention could be a strategic choice made by the visual system, depending on whichever was the least effortful means of extracting target information from a particular visual presentation. To test this, Lim and Lee employed a modified version of the paradigm used in McCormick, Klein and Johnston (1998).

In the double cue condition of the original study, boxes 10° to the left and right of a central fixation point were used to cue the subsequent onset of a target dot. The target could appear inside either box, or between a box and the central fixation. Participants were tasked with responding to the target once it appeared in their visual field. Results showed that reaction times (RTs) for targets in cued locations (i.e. in boxes) did not differ from those that appeared in irrelevant locations (i.e. outside boxes), which was interpreted as indicative of a unitary mode of attention (See Figure 1).

For their modified version of the double cue condition, Lim and Lee introduced a vertical wall positioned in the centre of the visual display with the intent of preventing a single locus of attention from encompassing both boxes. It was believed that this would incentivize a split mode of attention and, in contrast with McCormick et al. (1998), give rise to RTs for targets in irrelevant locations that were

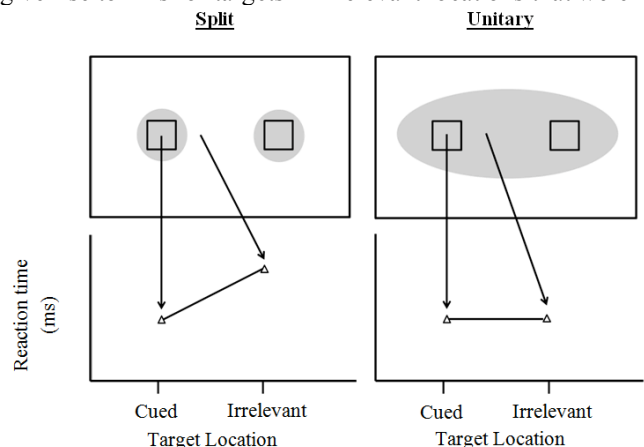


Figure 1. Predicted result trends, depending on mode of attention employed in double cue condition (McCormick et al., 1998).



slower than those in cued locations. Most intriguingly, while results revealed the expected trend, this was regardless of the presence or absence of the vertical wall. In other words, participants showed the capacity to split attention for both the experimental and control conditions – a contradiction of McCormick and colleagues' findings.

To explain this contradiction, Lim and Lee point towards the 10-odd years that have passed between the two studies. While this may not seem much in absolute terms, the 21<sup>st</sup> century has seen an exponential increase in the prevalence of digital devices in modern society. Compared to the undergraduates of 1998, those in 2011 would have grown up in a far more media-saturated environment with simultaneous exposure to multiple rich streams of information being a common occurrence (Ophir, Nass & Wagner, 2009). Studies have shown cultural differences in attentional phenomena such as change blindness (Masuda & Nisbett, 2006), while learned visual experiences have been known to result in differential allocations of visual attention (Chun, 2000). In a similar vein, by demanding the allocation of attention to multiple objects, today's media-saturated environment may encourage the development of the capacity to split attention.

In our present study, we examine the suggested relationship between media usage and splitting attention. To this end, we borrow the Media Multitasking Index (MMI) developed by Ophir et al. (2009), a measure of simultaneous media usage. High and Low media multitaskers, as indicated by the MMI, have been found to demonstrate fundamental differences in cognitive control as well as information processing approaches. We believe that similar differences may be observed in the capacity to split attention as well. Furthermore, to strengthen the case that the contradiction between Lim and Lee (2010) and McCormick et al. (1998) stems from ecological, rather than paradigm, differences, we employed a close replication of McCormick's paradigm.

It is hypothesized that:

1. MMI should not result in between subject performance differences for the single cue conditions. All participants should demonstrate faster RTs for targets that appear in cued locations, over those that appear in irrelevant locations.
2. For the double cue condition, individuals who tend not to media multitask, as indicated by their Low MMI scores, should demonstrate unitary attention: RTs for targets at cued and irrelevant locations should be comparable;
3. However, individuals who tend to media multitask, as indicated by their High MMI scores should demonstrate split attention: RTs for targets at cued locations should be significantly faster than for those at irrelevant locations

## Method

This research was conducted in two parts: a media use questionnaire and matrix, followed by a cognitive behavioral experiment.

### Participants.

66 introductory psychology students participated for course credit.

### Media use questionnaire and matrix.

The questionnaire and matrix were close replications of those employed in Ophir et al. (2009), with one slight modification. Ophir and colleagues (2009) had included items about non-visual media forms, such as music, in their original study. Given that the present study is primarily interested in the deployment of visual attention, all such irrelevant items were removed. A decision was made to retain the 'Handphone' item, as technological advances have arguably transformed the handphone into a personal digital assistant used for many functions other than for phonecalls.

The questionnaire addressed 10 different media forms: printed media, television, computer-based video (e.g. YouTube), video or computer games, non-call related mobile phone usage, online instant messaging, text messaging, e-mail, web surfing, and other computer applications (e.g. Word processor). Participants were required to report the total number of hours spent per week on each media form. Furthermore, participants filled up a 9 x 10 media-multitasking matrix, indicating the degree to which, while engaged in one media form as a primary activity, they would concurrently use other forms of media as well (1 "Never", to 4 "Most of the time"). Text messaging was excluded as a primary media form in the matrix as its usage could not be accurately described as a function of time. However, it was still available as an option under concurrent activities.

### Deriving MMI

We recoded matrix responses as follows: 1 "Never" = 0, 2 = 0.33, 3 = 0.67, and 4 "Most of the time" = 1. Summing up responses for each primary media form gave a measure of the mean number of other media used concurrently for each primary activity. Finally, to account for the different amounts of time spent on each media form, MMI was derived by calculating a sum of this measure across all primary media forms, weighted by the percentage of time spent on each primary media form. This process can be summarized with the following formula:

$$MMI = \sum_{i=1}^9 \frac{m_i \times h_i}{h_{total}}$$

where  $m_i$  is the number of media typically used while using primary media form  $i$ ,  $h_i$  is the number of hours per week reportedly spent using primary media form  $i$ , and  $h_{total}$  is the

total number of hours per week spent with all primary media forms.

### Stimuli

As far as possible, the stimuli, design, and procedure of our experiment were kept in accordance with Experiment 1 of McCormick et al. (1998). While our participants completed fewer trials, the proportion of trials for each condition remained the same (see Table 1).

The visual display comprised of a black background and centered white fixation cross ( $0.4^\circ$  by  $0.4^\circ$ ). The cue used was an empty white-bordered square ( $1.1^\circ$  by  $1.1^\circ$ ) which appeared  $10^\circ$  to the immediate left of fixation,  $10^\circ$  to the immediate right of fixation, or in both locations simultaneously. The imperative stimulus was a white dot positioned at one of four possible locations on the horizontal median ( $10^\circ$  left,  $5^\circ$  left,  $5^\circ$  right, or  $10^\circ$  right).

### Design

The experiment used a  $3 \times 4$  fully within factorial design. The two independent variables were (1) type of cue: (a) single box to the left, (b) single box to the right, and (c) double box, and (2) location of target dot: (a)  $10^\circ$  left, (b)  $5^\circ$  left, (c)  $5^\circ$  right and (d)  $10^\circ$  right.

### Procedure

The sequence of events (see Figure 2 for schematic) for each trial was as follows: the fixation cross was presented for 800ms. This was followed by a single or double-box cue. After a 515ms interval, the target dot was presented and remained on screen along with the boxes until the participant responded. Participants were instructed to press the response key (spacebar) as soon as they detected the target, but to refrain from responding when the target did not appear (catch trials).

Participants were informed that the box cues indicated the most likely location at which the target would appear, and to orient their attention to these locations. Participants were also informed that in the event of a double box cue, the target could appear at either location with equal probability. Participants were further instructed to attempt to divide their attention between the two boxes when presented with a double box cue.

Each participant took part in a single experiment session consisting of 420 trials. Four rest periods were interspersed during this session. The total number of trials for each cue-target combination is presented in Table 1.

Table 1  
Number of trials for each Cue-Target combination.

Cue Type	Target Location				Catch
	$10^\circ$ left	$5^\circ$ left	$5^\circ$ right	$10^\circ$ right	
Double Cue	60	10	10	60	20
Single Cue, Right	10	10	10	80	20
Single Cue, Left	80	10	10	10	20

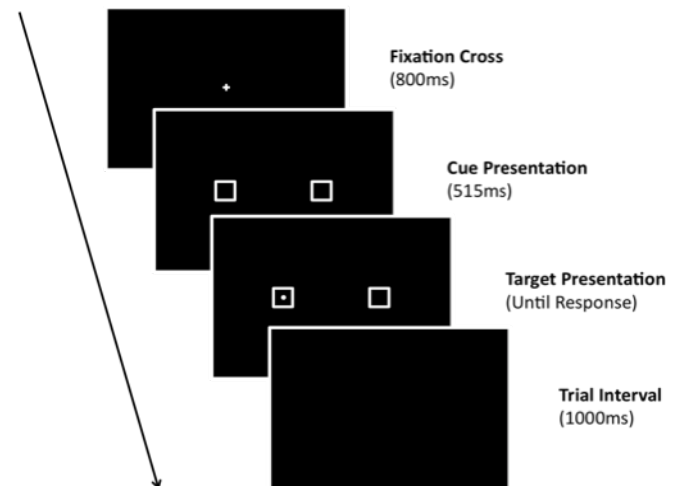


Figure 2. Schematic of the sequence of a trial.

### Results

Participant MMI scores ranged from 0.29 to 6.45, with an average of 3.15 and standard deviation of 1.37. There was no correlation between MMI score and the total number of hours spent on surveyed media forms each week,  $r(64) = .05$ ,  $p = .67$ . In preparation for further main analyses, participants with above-average MMI scores were classified as High scorers, whereas those with below-average MMI scores were considered Low scorers. Two groups of equal numbers were obtained: 33 High versus 33 Low scorers.

Responses on catch trials were only extremely rare (2.2% of catch trials), and no participants had to be excluded from analysis on this basis. For each participant, trials with RTs less than 150ms or in excess of 1000ms were regarded as errors and not analyzed (less than 1% of trials).

A repeated measures analysis of variance (ANOVA) was performed that included the variables cue type (left, right, or double) and target location ( $10^\circ$  left,  $5^\circ$  left,  $5^\circ$  right, or  $10^\circ$  right) as within-subject variables, and MMI (high or low) as a between-subject variable. Relevant means have been summarized in Table 2.

The Mauchly's test of sphericity was significant,  $p < .05$ , for the cue type  $\times$  target location interaction term, suggesting that sphericity was violated. Accordingly, a more stringent F-test (Greenhouse-Geisser) was used. Results revealed a significant main effect for cue type,  $F(2, 128) = 235.84$ ,  $p < .001$ ,  $\eta_p^2 = .90$ . There was also a significant main effect for target location,  $F(3, 192) = 34.44$ ,  $p < .001$ ,  $\eta_p^2 = .59$ . The main effect for MMI did not reach significance,  $F(1, 64) = 3.19$ ,  $p = .08$ . In terms of interactions, the cue type  $\times$  MMI interaction was non-significant,  $F(2, 128) = 0.99$ ,  $p = .37$ . Both the cue type  $\times$  target location and target location  $\times$  MMI interactions were significant,  $F(6, 384) = 189.49$ ,  $p < .001$ ,  $\eta_p^2 = .90$  and  $F(3, 192) = 4.66$ ,  $p < .01$ ,  $\eta_p^2 = .16$  respectively.

Most important, all effects were qualified by a significant cue type  $\times$  target location  $\times$  MMI interaction,  $F(6, 384) = 2.46$ ,  $p < .05$ ,  $\eta_p^2 = .21$ . Post-hoc tests revealed that the



Table 2  
Mean RTs (in ms)

MMI Score	Cue Type	Target Location			
		10° Left	5° Left	5° Right	10° Right
High	Left	307 (4.71)	358 (5.98)	358 (6.23)	380 (7.77)
	Right	397 (7.77)	380 (6.60)	361 (6.40)	304 (4.40)
	Double	319 (5.60)	328 (6.58)	332 (8.20)	313 (5.17)
Low	Left	303 (6.22)	336 (7.76)	351 (8.06)	366 (7.29)
	Right	372 (8.65)	355 (7.23)	348 (7.66)	297 (6.22)
	Double	306 (5.76)	306 (6.41)	310 (6.55)	307 (6.31)

Note: Standard errors are in parentheses.

target location x MMI interaction was non-significant for the left cue condition,  $F(3, 192) = 2.36, p = .084$ , but significant for the right and double cue conditions,  $F(3, 192) = 3.40, p < .05, \eta_p^2 = .05$  and  $F(3, 192) = 4.24, p = .01, \eta_p^2 = .22$  respectively. The significant target location x MMI interaction was examined individually for the right and double cue conditions. To test my hypothesis directly, we examined the simple main effect of target location at each level of MMI. For the right cue condition, the simple main effect of target location was significant for both High and Low MMI participants,  $F(3, 96) = 87.68, p < .001, \eta_p^2 = .90$  and  $F(3, 96) = 134.16, p < .001, \eta_p^2 = .81$  respectively. For the double cue condition, the simple main effect of target location was non-significant for Low MMI participants,  $F(3, 96) = 1.07, p = .36$ ; but significant for High MMI participants,  $F(3, 96) = 7.39, p < .001, \eta_p^2 = .19$ .

This 3-way interaction is reflected in Figure 3. A distinct RT gradient can be observed for both single cue conditions, regardless of the MMI scores of participants: RT was fastest at the cued location, and became slowed as the target was presented further from the cue. For the double cue condition, relatively constant RTs can be observed at each target location for Low MMI participants. For High MMI participants, a slight RT gradient can be observed between cued and irrelevant locations, with RTs being faster at cued locations.

As a further investigation of the 3-way interaction, and as a planned comparison, we collapsed both single cue conditions into one condition, and sorted the RT data by *trial type*. Specifically, we analysed the RT difference between *valid* trials, where the target appeared at cued locations, and *valid probe* trials, where the target appeared 5° to the left or right of cued locations. The outcomes are as follows:

For the single cue condition, the effect of trial type was significant regardless of the level of MMI. For both High and Low MMI participants: RTs on valid trials (Low: 300ms, High: 306ms) were significantly faster than those on valid probe trials (342ms, 359ms),  $t(32) = 11.05, p < .001$  and  $t(32) = 14.33, p < .001$  respectively (See Figure 4).

For the double cue condition, the effect of trial type was non-significant for Low MMI participants: RTs on valid and

valid probe trials were comparable (306ms vs 308ms),  $t(32) = 0.91, p = .37$ . In contrast, the effect of trial type was significant for High MMI participants: RTs on valid trials were faster than on valid probe trials (316ms vs 330ms),  $t(32) = 4.84, p < .001$  respectively (See Figure 5).

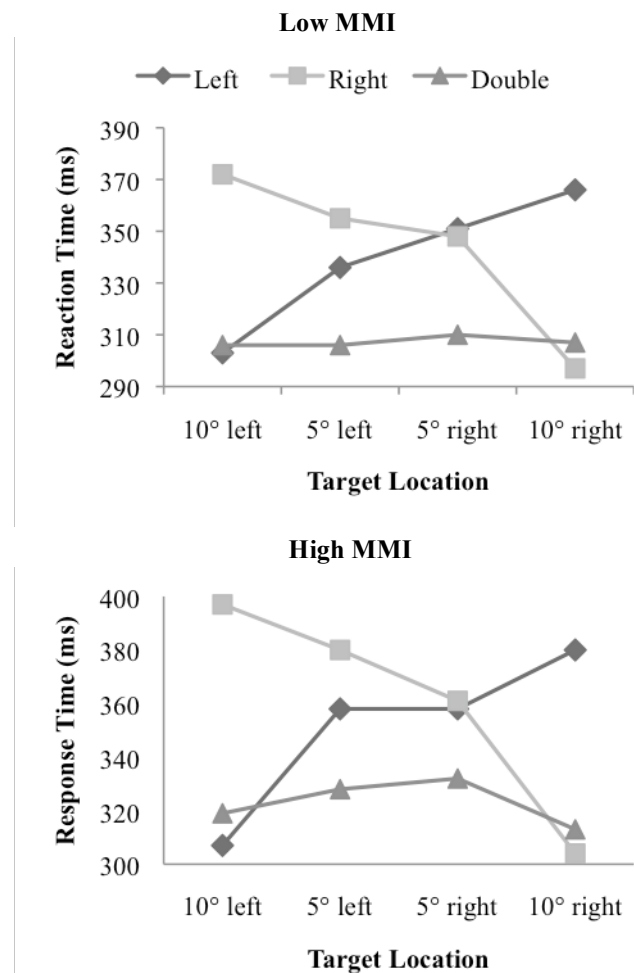


Figure 3. Response time trends (in ms) for Low and High MMI participants.

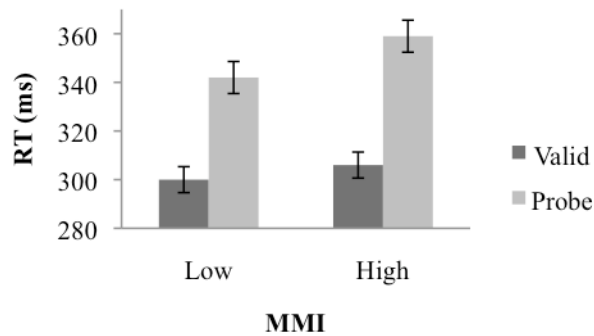


Figure 4. *Single cue condition: Effect of trial type at each level of MMI. Error bars indicate standard error.*

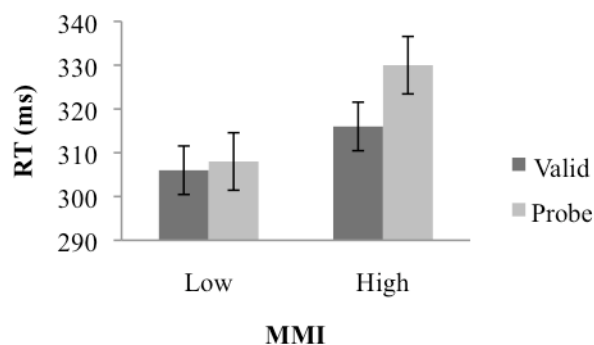


Figure 5. *Double cue condition: Effect of trial type at each level of MMI. Error bars indicate standard error.*

## Discussion

The present experiment has replicated one of the central findings in McCormick et al. (1998): Increasing RTs as targets are presented further from cued locations in the single cue conditions. This RT gradient has also been observed in prior studies which used probes to test the spatial extent of visual attention (e.g. LaBerge, 1983). Two interpretations have been offered to explain this gradient. Umiltà, Riggio, Dascola, and Rizzolatti (1991) proposed that increasing RTs could reflect the need to shift attentional focus, while McCormick and Klein (1990) suggested that the gradient could reflect diminishing concentrations of attentional resources at distances further from the cued location.

McCormick et al. (1998) did not observe such a gradient in their double cue condition. Accordingly, it was concluded that either participants had no need to shift attention to targets which appeared in irrelevant locations, or similar levels of attention had been deployed across all four potential target locations. Both interpretations are consistent with the idea of a single attentional focus expanded to encompass noncontiguous cued locations, as predicted by the zoom lens model.

Consistent with McCormick and colleagues' (1998) findings, no RT gradient was observed in the double cue condition for participants with Low MMI scores. RTs to targets appearing at cued and irrelevant locations were comparable. This suggests that a unitary mode of attention was being employed by these participants.

In contrast, the critical finding in the present experiment was the observation of an RT gradient in the double cue condition for participants with High MMI scores. RTs to targets at cued locations were faster than for those at irrelevant locations. As elaborated above, this suggests that: 1) High MMI participants had to shift attention from cued locations to attend to targets in the irrelevant region in between, or 2) High MMI participants had deployed higher concentrations of attention at cued locations, compared to irrelevant locations. Both interpretations are consistent with the idea of the deployment of two attentional foci in noncontiguous locations – a demonstration of split attention.

Taking these results together, an interpretation based on MMI variability seems to provide excellent insights into the conditions under which split attention tends to occur, which we elaborate below.

MMI is a measure of media multitasking behaviour. The 'score' obtained is an estimation of the number of media forms a participant tends to concurrently use in a typical hour of media consumption. Thus, the critical difference between Low and High MMI participants is the behavioural tendency to consume multiple visual media forms simultaneously. Our results have established the following:

1. People who tend to consume multiple visual media forms simultaneously, as indicated by High MMI scores, employed a split mode of attention when presented with cued noncontiguous locations.
2. People who tend to consume fewer visual media forms simultaneously, as indicated by Low MMI scores, employed a unitary mode of attention when presented with the same.

This suggests a relationship between simultaneous media usage and the capacity to split attention. If, as argued by Lim and Lee (2011), ecological differences in terms of the media-saturation of participants' environments account for the contradictory results between Lim and Lee (2011) and McCormick et al. (1998), this further suggests that the capacity to split attention ought to follow prolonged simultaneous media usage, and not vice versa (however, this remains an open empirical question at this juncture, as MMI scores could not be experimentally manipulated to provide a fuller claim).

How, then, might such a development take place? As suggested in Lim and Lee (2011), the employment of a split or unitary mode of attention could be a strategic decision made by the visual system, depending on whichever was the least effortful means of extracting information from a visual representation. It has also been suggested that maintaining a split mode of attention is inherently more effortful than unitary attention, and leads to performance costs (Cave et al., 2010). Pulling these trains of thought together, we

propose that prolonged simultaneous media usage acts as a form of practice which reduces the effort needed to maintain split attention. This, in turn, increases the occasions where split attention becomes strategically optimal, and thus employed in lieu of unitary attention. A new direction for future investigations would be to empirically test this hypothesis through the use of training studies.

With respect to existing split attention literature, the present study contributes to converging evidence which suggests that the deployment of multiple attentional foci is possible. Furthermore, the present study advances our understanding of the conditions under which focal attention might split. New evidence is presented, for the first time in the literature, supporting the possibility of individual differences in the capacity, or at least a tendency, to split attention. Potential individual differences, which need not necessarily stem from media multitasking behavior alone, might at least in part explain the body of conflicting results pointing to unitary attention. Prior attempts to reconcile conflicting findings have typically focused on paradigm differences (Dubois et al. 2009; Kramer & Hahn, 1995). A full picture may only be achieved by considering the interaction between individual and paradigm factors.

In addition, we wish to highlight that the present study contributes to the general finding that ecological experiences influence various visual attentional processes. An interesting direction of future research would be to explore the possibility of other ecological influences on the capacity to split attention. One possible candidate is habitual video game playing. Expert gamers are significantly better at identifying targets presented towards the periphery of their vision than non-gamers, which has been interpreted as superior ability at distributing attentional resources throughout the visual field (Green & Bavelier, 2003). Given that split attention can be conceptualized as the ability to deploy attention in a flexible manner, the attention distribution benefits stemming from habitual game playing might have an impact on the capacity to split attention as well. This possibility is both a theoretically as well as an empirically exciting one.

In conclusion, we report novel evidence that ecological differences, arising from our increasingly media-saturated environment and how we choose to interact with it, could lead to individual differences in the capacity to demonstrate split focal attention. This advances our understanding of the exact conditions under which focal attention might split, and might partly account for the conflicting evidence for and against split attention found in the literature. Further studies may consider other potential sources of individual differences in the capacity to split attention. Moving forward, we are very hopeful that ecological factors will continue to shed new light on the persistent puzzle of splitting visual focal attention.

## References

- Castiello, U., & Umiltà, C. (1992). Splitting focal attention. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 837-848.
- Cave, K. R., Bush, W. S., & Taylor, T. G. G. (2010). Split attention as part of a flexible attention system for complex scenes: Comments on Jans, Peters, and De Weerd (2010). *Psychological Review*, 117, 685-696.
- Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, 4, 170-178.
- Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, 16, 143-149.
- Eriksen, C.W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: A zoom-lens model. *Perception & Psychophysics*, 40, 225-240.
- Eriksen, C. W., & Yeh, Y. Y. (1985). Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance*, 11, 583-597.
- Jans, B., Peters, J. C., & De Weerd, P. (2010). Visual spatial attention to multiple locations at once: The jury is still out. *Psychological Review*, 117, 637-684.
- LaBerge, D. (1983). Spatial extent of attention to letters and words. *Journal of Experimental Psychology: Human Perception and Performance*, 9, 371-379.
- Lim, S. W. H., & Lee, L. N. (2011). When (and why) might visual focal attention split? In B. Kokinov, A. Karmiloff-Smith, & N. J. Nersessian (Eds.), *European perspectives on cognitive science*. New Bulgarian University Press.
- Masuda, T., & Nisbett, R. E. (2006). Culture and change blindness. *Cognitive Science*, 30, 381-399.
- McCormick, P. A., & Klein, R. (1990). The spatial distribution of attention during covert visual orienting. *Acta Psychologica*, 75, 225-242.
- McCormick, P. A., Klein, R., & Johnston, S. (1998). Splitting versus sharing focal attention: Comment on Castiello and Umiltà (1992). *Journal of Experimental Psychology: Human Perception and Performance*, 24, 350-357.
- Ophir, E., Nass, C., & Wagner, A. D. (2009). Cognitive control in media multitaskers. *Proceedings of the National Academy of Sciences*, 106, 15583-15587.
- Pan, K., & Eriksen, C. W. (1993) Attentional distribution in the visual field during same-different judgments as assessed by response competition. *Perception & Psychophysics*, 53, 134-144.
- Posner, M. I. (1980) Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32, 3-25.
- Umiltà, C., Riggio, L., Dascola, I., & Rizzolatti, G. (1991). Differential effects of peripheral cues on the reorienting of spatial attention. *European Journal of Cognitive Psychology*, 3, 247-267.

# Roles of Adults' Gestures and Eye Gaze in Whole or Object Part Presenting

**Tetsuya Yasuda (cs.yasuda@me.com)**

Saitama Prefectural University, Center for University-wide Education, 820 San-Nomiya, Koshigaya  
Saitama, 343-8540, JAPAN

**Harumi Kobayashi (h-koba@mail.dendai.ac.jp)**

Tokyo Denki University, Division of Information System Design, Ishizaka Hatoyama-machi  
Saitama, 350-0394, JAPAN

## Abstract

We investigated the use of a caregiver's actions and eye-gaze in teaching whole or part names. The experimental material consisted of two everyday objects, a toothbrush (the whole name was "haburashi", the part name was "ke") and a ball-point pen. We coded 4 action type categories and 2 eye gaze type categories based on the video data of 19 4-year-old child-mother dyads using frame-by-frame method. Results of actions showed that when the caregiver uttered a whole object name such as toothbrush ("haburashi"), the caregiver tended to present the object to the child by showing it. When she uttered a part name to teach the part name such as brush ("ke"), she pointed at the object part. Results of eye gaze analysis showed whereas the caregiver tended to look at the child's face in teaching whole names, she tended to look at the object in teaching part names. We found that caregivers use different gestures and eye gaze directions to teach whole or object part names. The study suggests that caregivers help young children's word learning using appropriate gestures and gaze directions.

**Keywords:** gesture; showing; eye gaze; teaching part names.

## Introduction

It is claimed that children acquire language with the ability of specifying adults' referential intentions. This study focuses on adults' referential actions when they were asked to teach about whole objects or object parts to examine whether they provide useful information for young children's guessing adult referential intentions.

It is necessary for young children to accurately understand adults' intentional actions and associate accurately referred objects and words (Zukow, 1990). Adamson, Bakeman, and Deckner (2004) presented a study that a caregiver looked at a young child's eyes using mutual gaze when she taught an object name for the child. It is also important for young children to know an adults gaze direction (Doherty & Anderson, 1999). Doherty, Anderson, and Howieson (2009) showed 3-year-old children understood adults' subtle gaze direction.

Children are sensitive to the referential intentions conveyed by pointing and looking and they use information they obtained for word learning (Tomasello, Carpenter, & Lizkowski, 2007; Kobayashi, 1998, 1999). Tomasello & Farrar (1986) contended that mothers' object references that follow into children's attentional focus may facilitate their lexical acquisition. However an adult may not always offer information that is unambiguous and right. Baldwin (1991,

1993) investigated how children know adults' referential intentions using a discrepant labeling situation. In the discrepant labeling situation, an adult's attention at objects does not agree with a child's attention at objects. The study showed that 18 months or older children checked an adult's eye gaze to know the object which the adult labeled. In the everyday life, maternal naming and labeling for children often includes discrepant labeling situations (Collis, 1977; Harris, Jones, & Grant, 1983). Children must be equipped with ability to know correct referential intentions of adults using eye gaze and other information even in discrepant labeling situations.

Caregivers seem to use specific ways to convey referential intentions to young children. Masur (1997) examined caregivers' natural interaction with their infants using novel, comprehended, and familiar toy animals. The results were that mothers virtually always named whole objects first. More importantly, in the first mention of novel object, they named it and designated it with physical contact such as pointing, holding, or manipulating, but such naming and actions did not occur on the first mention of comprehended or familiar toy animals.

Although these studies provided precious evidence on adults' referential actions, these studies all focus on caregivers' teaching about whole object labels. An object is composed of various parts, so in part name learning, children should find a specific object part in a whole object and associate the part with the part name. It is not known whether caregivers use any specific referential actions or eye gaze to teach children part names.

In the literature of word learning, learning part names has been more discussed in the use of linguistic cues rather than nonlinguistic cues (Markman & Wachtel, 1988; Saylor, Baldwin, & Sabbagh, 2002). Markman and Wachtel (1988) showed that children learned part names when they could use knowledge about a whole object name and by applying whole object assumption and mutual exclusivity. Saylor, Baldwin, & Sabbagh (2002) demonstrated that if an object is presented and its whole-name and one of its part names are juxtaposed, children could learn the part name of the object. However, on whether children can learn part names using nonlinguistic cues, the results are mixed. Examined nonlinguistic cues were general direction of pointing (Mervis, Golinkoff, & Bertrand, 1994), facial expression (Moll, Koring, Carpenter, & Tomasello, 2006), tracing the contour of an object part (Hansen & Markman, 2009), and

functional actions (Kobayashi, 1998). Some cues have been demonstrated to help young children, but whether children use gestures such as pointing in learning part names is not well known.

Kobayashi (1998, 2002, 2012) showed that in the task to associate novel labels with object parts, 2- and 4-year-olds accurately interpreted adults' referential actions such as moving object parts or simply touch-pointing object parts to know word meanings. Yasuda and Kobayashi (2010) reported effects of eye gaze direction in learning part names. Children learned part names accurately when the adult pointed and looked at the object part and named the object part. However, if the adult looked at child's face doing the same actions, children DID NOT associate the object part with the part name. The study suggested that children learn part names accurately when adult focused on the object part by looking at the object part in addition to pointing at the object part. We speculated that caregivers may actually use referential actions and referential gaze so that young children understand whether the whole object or the object part is named.

In this study, we examined caregivers' referential actions in teaching part names to their 4-year-old children. We examined 4-year-olds' mothers because our previous research (Yasuda & Kobayashi, 2010) showed that 4-year-olds' mothers used referential actions such as pointing in teaching object part names. In addition, our recent research (2012) shows that 4-year-old children are sensitive to adults' pointing with touching the part. We decided to examine mothers' teaching gestures in two conditions, whole object name teaching and object part name teaching.

In Analysis 1, we asked a caregiver to teach either a whole object name or object part name and analyzed the caregiver's referential actions when she uttered either a whole label or a part label. In Analysis 2, we compared the caregiver's eye gaze when she uttered labels in each situation. If the caregiver used different referential actions when she uttered a whole and a part label, and if they use different eye gaze simultaneously with labeling whole or part labels, it can be said that they actually provide different nonlinguistic information for young children to help they learn whole and part labels.

## Method

Nineteen pair of 4-year-olds (Mean=56.25 SD= 3.455 Range= 52-63 months) and their caregivers participated in the experiment in the Greater Tokyo Area, Japan. This experiment was conducted in conformity with a privacy ethical code of Tokyo Denki University.

## Material

The experimental material consisted of toothbrush and ball-point pen as familiar objects. We prepared the toothbrush ("haburashi" in Japanese) as the whole object and a brush (Ke in Japanese) as the object part. We prepared a ball-point pen ("bohru-pen" in Japanese) as the whole object and a knock bottom ("nokka" in Japanese) as the object part.

These objects in each set had distinctive shapes, and had the same color and texture as the other parts of the material object.

## Procedure

The child and the caregiver sat at a table corner face-to-face (Figure 1). The experimenter recorded the experiment by two digital video cameras (Sony, SR-60, 29.97frame/sec). One of the video cameras focused on the caregiver's face and hand, and the child's face. The other video camera took the whole view to the experimental situation. These two video cameras were appropriately synchronized.

Each pair of participants was randomly assigned to one of two conditions, whole object teaching: the caregiver teaches something about the whole object (e.g. whole object name: toothbrush or "haburashi") and object part teaching: the caregiver teaches something about the object part (e.g. object part name: brush or "ke"). In the whole object teaching condition in the brush session, the experimenter gave the toothbrush to the caregiver and said to the caregiver in Japanese: "Please teach about the toothbrush as you like." ("Jiyu ni haburashi ni tsuite oshiete kudasai."). When the caregiver finished the brush session, she return the toothbrush and received the ball-point pen and taught about the ball-point pen. In the part teaching condition, the procedure was identical with the whole object teaching condition except that the caregiver was asked to teach the object part instead of the whole object. The experimenter asked the caregiver in Japanese: "Please teach about this brush part of this object." ("Jiyu ni kono ke no bubunn ni tsuite oshiete kudasai.").

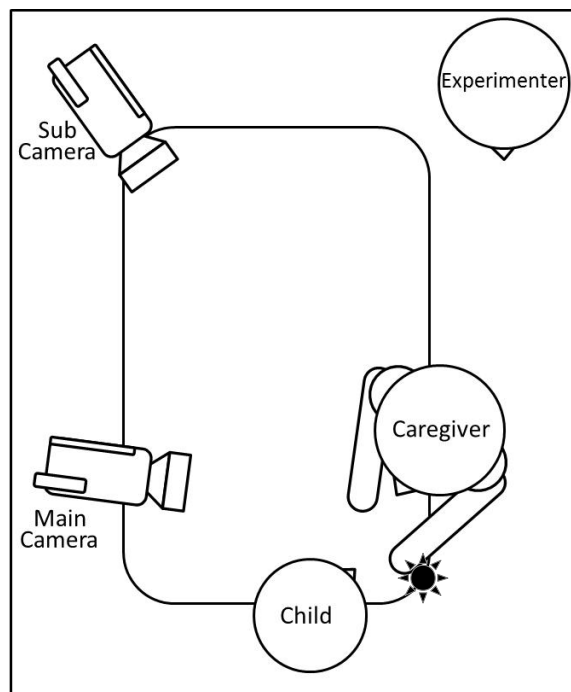


Figure 1. The experimental setting

## Analysis 1

We examined whether the caregiver looked at the object or the child's face when the caregiver taught the whole object or the object part name. We categorized the caregiver's actions into four types (pointing, stroking, demonstration, and showing) based on video data (30 frame/sec) using frame-by-frame method. We coded the caregiver's actions when she uttered the first sound of either whole object name or object part name. We coded "pointing" when the caregiver pointed at the object. We coded "stroking" when the caregiver stroked the object contour. We coded "demonstration" when the caregiver demonstrated the function of the whole object or object part to present the child. We coded "showing" when the caregiver showed the object to the child. All caregivers' actions were categorized into one of these four categories. There were no caregivers who did not do any action when she uttered the first sound of whole object name or object part name.

We first calculated the frequency of each action type in each caregiver. Figure 2 shows the ratio of action type in each caregiver.

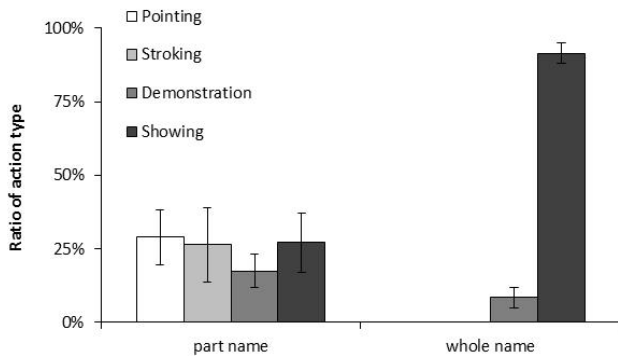


Figure 2. The ratio of action type in each caregiver.

## Results of Actions

In order to analyze each caregiver's most representative referential action, we also calculated the proportion of each caregiver's action type based on the caregiver's most frequently occurred referential action observed in each teaching condition.

To test whether caregivers' actions varied across teaching types, a 2 (Teaching: whole label naming and part label naming)  $\times$  2 (Action: pointing, stroking, demonstration, showing) chi-square test was conducted. There was a significant effect of teaching type,  $\chi^2(3)=9.6, p<.05$ .

To explore the significant effect, we conducted residual analysis by computing adjusted residual. When the caregiver pointed at the object, the caregiver uttered more part names ( $r=3.098, p<.05$ ) than whole names ( $r=-2.309, p<.05$ ). When the caregiver presented the object using showing, the caregiver uttered more whole names ( $r=2.498, p<.01$ ) than part names ( $r=-3.098, p<.01$ ).

These results indicate that 1) When the caregiver uttered the whole label, she tended to present the object by showing. 2) When the caregiver uttered the part label, she tended to point at the object.

The caregiver seemed to use appropriate gestures to convey her referential intention on the object. Caregivers pointed at the part when she uttered a part label. Here, she specified the object part and taught part names. The caregiver DID NOT point at the object when she uttered a whole name.

## Analysis 2

In Analysis 1, we examined whether a caregiver looked at the object or the child's face when caregiver uttered the whole object name or object part name. We categorized the caregiver's eye gaze into two types (object and child's face) based on video data (30 frame/sec) using frame-by-frame method. We categorized a gaze an object gaze if the caregiver looked at the object when she uttered an object name. We categorized a gaze a child's face gaze if the caregiver looked at the child's face or hand when she uttered an object name. We coded these gaze at the first sound of either whole or part names. No caregiver looked at the digital camera or others when she uttered whole or part labels.

We first calculated the frequency of each eye gaze type in each caregiver. Figure 3 shows the ratio of action type in each caregiver.

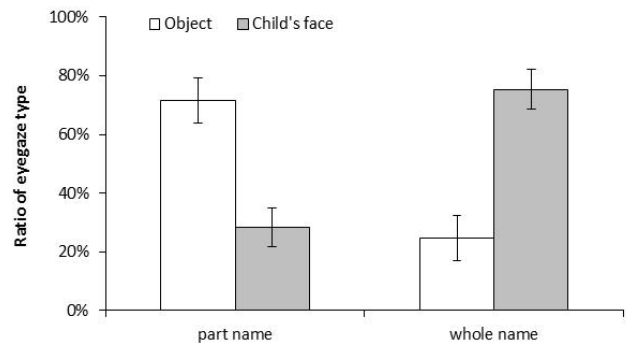


Figure 3. The ratio of gaze type in each caregiver

## Results of Eye-Gaze

In order to analyze each caregiver's most representative eye gaze, we also calculated the proportion of each caregiver's eye gaze type based on the caregiver's most frequently occurred referential gaze observed in each teaching condition.

To test whether caregivers' eye gaze varied across teaching types, a 2 (Teaching: whole label naming and part label naming)  $\times$  2 (Gaze: Object, Child's face) chi-square test was conducted. There was a significant effect of teaching type,  $\chi^2(1)=12.40, p<.01$ .

To explore the significant effect, we conducted residual analysis by computing adjusted residual. When the caregiver looked at the object, the caregiver uttered more part names ( $r = 3.528$ ,  $p < .01$ ) than whole names ( $r = -3.528$ ,  $p < .01$ ). When the caregiver looked at the child's face, the caregiver uttered more whole names ( $r = 3.528$ ,  $p < .01$ ) than part names ( $r = -3.528$ ,  $p < .01$ ).

These results indicate that 1) When the caregiver taught the whole label, she looked at the child's face or hand. 2) When the caregiver taught the part label, she looked at the object. The caregiver looked at the object when she presented the object with a showing gesture. She rarely looked at the object when she taught the whole name with a showing gesture

## Discussion

We investigated the relationship between a caregiver's actions and eye gaze in teaching whole or part object names. The experimental material consisted of two everyday objects, toothbrush (whole name is "haburashi", part name is "ke") and ball-point pen (whole name is "boru pen", part name is "knocker"). We coded 4 action type categories and 2 eye gaze type categories based on the video data using frame-by-frame method.

Results of action data showed that the caregiver presented the object by showing it when she uttered a whole object name such as toothbrush. However, the caregiver pointed at the object when she uttered a part name to teach the part name such as brush. Thus the caregiver taught part names using pointing at the object part. Caregivers may know their children can learn part names by observing adult pointing actions. Pointing at object part can appropriately attract the child's attention to the object part. That may be one reason why caregivers choose to use showing rather than pointing in whole object naming so that the child can appropriately focus on the whole object rather than any specific part of the object.

Results of eye gaze data showed that whereas caregivers tend to look at a child's face in teaching whole names, they tend to look at an object itself in teaching part names. Caregivers seem to know looking at the object part in addition to pointing the object part is important to teach children object part names.

It is suggested that the caregiver conveys referential intentions to teach whole or object part names using appropriate gestures and gaze direction. The results of this study accord with social-pragmatic approach to word learning (Clark, 2009; Tomasello, 2003). Children learn word meanings by guessing adult intentions provided by adult utterances. This study provided evidence that adult certainly provide nonlinguistic information through gestures and eye gaze to help children's understanding of adult referential intentions.

## Acknowledgments

This work was partly supported by KAKENHI (20500241) and (24530793) on Grant-in-Aid for Scientific Research (C).

## References

- Adamson, L. B., Bakeman, R., & Deckner, D. F. (2004). The development of symbol-infused joint engagement. *Child Development*, 75, 1171-1187.
- Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, 62(5), 875-890.
- Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, 20, 394-419.
- Clark, E. (2009). *First language acquisition* (2nd ed.). Cambridge, UK, Cambridge University Press.
- Collis, G. (1977). Visual co-orientation and maternal speech. In H. R. Schaffer (Ed.). *Studies in mother-infant interaction*, (pp. 355-375). London: Academic Press.
- Doherty, M.J., & Anderson, J.R. (2001). People don't keep their heads still when looking to one side, and other people can tell. *Perception*, 30, 765 - 767.
- Doherty, M.J., Anderson, J.R., & Howieson, L. (2009). The rapid development of explicit gaze judgment ability at 3 years. *Journal of Experimental Child Psychology*, 104, 296-312.
- Harris, M., Jones, D., & Grant, J. (1983). The nonverbal context of mothers' speech to infants. *First Language*, 4, 21 -30.
- Kobayashi, H. (1998). How 2-year-old children learn novel part names of unfamiliar objects. *Cognition* 68, B41-51.
- Kobayashi, H. (1999). The influence of adults' actions on children's inferences about word meanings. *Japanese Psychological Research*, 41(1), 35-49.
- Kobayashi, H. (2002). Learning the novel part names with observation of adults' gestures. *Studies in Language Sciences*, 2, 149-156.
- Kobayashi, H. (2012). Meaning of touch-pointing in part name learning. Manuscript submitted for publication.
- Masur, E. (1997). Maternal labelling of novel and familiar objects: implications for children's development of lexical constraints. *Journal of Child Language*, 24, 427-439.
- Markman, E. M. & Wachtel, G. A. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20, 120-157.
- Mervis, C. B., Golinkoff, R. M., & Bertrand, J. (1994). Two-year-olds readily learn multiple labels for the same basic level category. *Child Development*, 65, 971-991.
- Moll, H., Koring, C., Carpenter, M., & Tomasello, M. (2006). Infants determine others' focus of attention by pragmatics and exclusion. *Journal of Cognition and Development*, 7(3), 411-430.
- Tomasello, M. (1997). The pragmatics of word learning. *Japanese Journal of Cognitive Science*, 4, 59-74.
- Tomasello, M. (2001). *The Cultural Origins of Human Cognition*, Cambridge, MA: Harvard University Press
- Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.
- Tomasello, M. and Farrar, M. (1986). Joint Attention and Early Language. *Child. Development*, 57, 1454-1463.



- Tomasello, M., Carpenter, M., & Lizskowski, U. (2007). A new look at infant pointing. *Child Development*, 78, 705-22.
- Saylor, M. M., Sabbagh, M. A., & Baldwin, D. A. (2002). Children use whole-part juxtaposition as a pragmatic cue to word meaning. *Developmental Psychology*, 38, 993-1003.
- Saylor, M. M., & Sabbagh, M. A. (2004). Different kinds of information affect word learning in the preschool years: The case of part-term learning. *Child Development*, 75, 395-408.
- Yasuda, T., & Kobayashi, H. (2010). The role of adults' eye gaze direction in children's learning part names, *Handbook on the 12th Annual International Conference of the Japanese Society for Language Sciences (JSLs 2010)*, pp53-36.
- Zukow, P.G. (1990). Socio-perceptual bases for the emergence of language: an alternative to innatist approaches. *Developmental Psychobiology*, 23, 705-726.

# State-Trace Analysis of Sequence Learning by Recurrent Networks

Fayme Yeates<sup>1</sup> (fy212@exeter.ac.uk) Andy Wills<sup>1</sup> (A.J.Wills@exeter.ac.uk)  
Fergal Jones<sup>2</sup> (Fergal.Jones@canterbury.ac.uk) Ian McLaren<sup>1</sup> (I.P.L.McLaren@exeter.ac.uk)

<sup>1</sup>School of Psychology, College of Life and Environmental Sciences, University of Exeter, UK.

<sup>2</sup>School of Psychology, Canterbury Christ Church University, UK.



## Abstract

This study investigated the use of state-trace analysis (Bamber, 1979) when applied to computational models of human learning. We aimed to investigate the performance of simple recurrent networks (SRNs) on a sequence learning task. Elman's (1990) SRN and Cleeremans & McClelland's (1991) Augmented SRN are both benchmark models of human sequence learning. The differences between these models, comprising of an additional learning parameter and the use of response units activated by output units constituted our main manipulation. The results are presented as a state-trace analysis, which demonstrates that the addition of an additional type of weight component, and response units to a SRN produces multi-dimensional state-trace plots. However, varying the learning rate parameter of the SRN also produced two functions on a state-trace plot, suggesting that state-trace analysis may be sensitive to variation within a single process.

**Keywords:** Learning; state-trace analysis; SRN; sequence learning; Augmented SRN;

## Introduction

State-trace analysis (Bamber, 1979) is a method that aims to establish whether one or more underlying processes are influencing behavior on a given task. The method has been applied to a variety of paradigms, including remember-know tasks (e.g. Dunn, 2008), face recognition (e.g. Loftus, Oberg, & Dillon 2004), categorization (e.g. Newell, Dunn & Kalish, 2010) and a variety of other areas (see Prince, Brown & Heathcote, 2011).

The procedure for a state-trace analysis is to plot the relationship between two dependent variables (*dimensions*) on two or more tasks (*states*). If these points follow a single, monotonic function, it can be hypothesized that the same latent variable underlies performance on the tasks. The influence of more than one latent variable on the tasks is implied when the state plots do not follow the same function, i.e. more than one monotonic function is visualized.

Computational models are created in the full knowledge of the processes involved in their construction. Thus, the primary use of state-trace analysis, to attempt to quantify latent psychological variables, does not seem to directly lend itself to computational modeling. But the fact that we should be able to make some predictions from the nature of the models about the types of processes involved in any simulation helps us interpret any state-trace analysis of the data produced by the simulation. This paper seeks to apply state-trace analysis to the simulation results produced by computational models on a sequence learning task both to

evaluate the different types of model and as a means of evaluating the state-trace methodology itself.

The computational models chosen for this analysis are the simple recurrent network (SRN) introduced by Elman (1990) and the Augmented SRN (Cleeremans and McClelland, 1991). The basic SRN model is simple (see Figure 1), involving feed-forward input activation through a hidden layer. The activations of this hidden layer are copied back on each trial into a context layer, which is then fed back into the hidden layer as input on the next trial. This ensures that the representations of the previous trial are carried over, and gives the model the ability to learn sequential information.

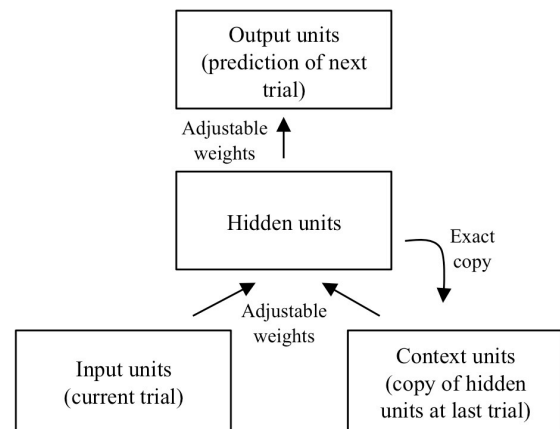


Figure 1: The architecture of the SRN (Elman, 1990).

Cleeremans and McClelland (1991) further developed the SRN in order to give a better account of the sequential effects demonstrated by human participants. This augmented simple recurrent network (AugSRN) differs from Elman's (1990) original architecture firstly in the inclusion of response units post-output. As a consequence, when making a response to the stimuli in an experiment, this response remains primed over future trials for a short time (Remington, 1969), because the response units are activated by the output units and feed activation back into the output units on the next trial.

The AugSRN also accounts for the priming of certain sequential pairings (Cleeremans & McClelland, 1991) by assuming that back propagation is implemented on not one set of connection weights, but two. One component is a set of slow weights, which produce small but permanent changes with minimal decay. These are complemented by fast weights, which have a higher learning rate but also a

greater rate of decay: simulating transient, short-term learning.

Both models have successfully simulated a range of human datasets (Cleeremans, 1993), including modeling sequence learning in serial reaction time (SRT) tasks. Jones and McLaren (2009) found that an AugSRN could produce the detailed pattern of subsequence learning demonstrated by participants in their experiments. Using a two-choice SRT task participants received continuous strings of stimuli that followed an exclusive-or rule two-thirds of the time. If the previous two responses were the same (XX or YY), then the current trial would be one response (X), and if the previous two had not been the same (XY or YX) this would lead to the other response (Y). Participants under incidental conditions found it hard to learn about the subsequence XXX compared to the other types, a result which was successfully simulated by the AugSRN.

In a later experiment Yeates, Jones, Wills, McLaren and McLaren (submitted) used the same two-choice SRT task to investigate human sequence learning. Participants in this study were divided into two groups, both of which received sequences governed by a rule that they were not informed about. In one group, the current trial could be predicted two-thirds of the time to be different to the trial before last, i.e. 'first different to third' (Group 1: XXY, XYY, YYX, YXX) and in the other the current trial would be predicted to be the same as the trial before last, 'first same as third' (Group 2: XXX, XYX, YYY, YXY). Poorer performance was predicted in Group 2 under incidental conditions based on Jones and McLaren's (2009) earlier findings that participants were unable to learn about subsequence XXX (or YYY). The manipulation did indeed produce this difference between Groups. We found that variants of the SRN and AugSRN could simulate these data to differing extents depending on the parameterisation of the model (Yeates et al., submitted), which suggests that the differences between the SRN and AugSRN may be of interest in this context.

SRNs are considered to be single-process models (e.g. Frensch & Miner, 1994; Kinder & Shanks, 2003), where parameters can be altered to produce different effects, but these involve essentially one process. The standard view is that the two connection weight components in an AugSRN represent the same kind of process; of learning through back propagation, and that their differences are of amount and not kind. Varying the learning rate affects the efficiency of learning across training (Kinder & Shanks, 2001; McClelland & Rumelhart, 1986). However, one might hold the view that the two connection weight components are in fact different processes within the AugSRN, accounting for long- and short-term learning. Similarly, as response units were introduced to take account for short term priming of the previous response (Cleeremans & McClelland, 1991), we could argue that this additional component may also represent an additional, different process.

We therefore hypothesize that when comparing performance of the SRN and AugSRN we will see a clearly

multi-dimensional state-trace plot, as the two models are different in kind. Further to this, we aim to examine the components of the AugSRN in more detail, with the aim to investigate whether state-trace analysis considers these additions to the original SRN separate processes within the model.

Given this analysis, our approach was to produce a state-trace analysis of these models' performance on a task based closely on the two-choice serial reaction time (SRT) experiments described in Jones & McLaren (2009) and Yeates, Jones, Wills, McLaren and McLaren (submitted). We aimed to compare the performance of these networks on this task, varying the free parameters of the models.

## Modeling Sequence Learning

The SRT paradigm involves participants responding to stimuli on screen that follow some sequence (Nissen & Bullemer, 1987; Lewicki, Czyzewska, and Hoffman, 1987). Therefore, faster and more accurate responses are expected for those trials that are predicted by the sequences learnt in comparison to a control group, who would receive the same task but with a pseudorandom ordering (e.g. Anastasopoulou & Harvey, 1999, Jones & McLaren, 2009).

## SRT Task Outline

The task experienced by each network follows closely that we have used with human participants, and lasted for two sessions, each with 20 blocks. Each block comprised 120 continuous trials of stimuli appearing on the right or left. The sequences making up each block were constructed differently for Group 1 and 2, and for the networks acting as control groups. For the experimental networks, all blocks in the first session and the first fifteen sequences in the second session were constructed from 40 triplets that followed the rule for each Group (Group 1: XXY, XYY, YYX, YXX; Group 2: XXX, XYX, YYY, YXY). Networks thus received ten of each subsequence type per block.

Two-thirds of experimental training trials followed the rule, as the third trial in a triplet was always consistent with the rule, as were half of first and second trials in a triplet by chance when subsequences were randomly concatenated. Test and control group training blocks were made up of pseudorandom sequences that included an equal amount of all subsequence types.

## Model Construction

The parameters varied in the model for the purpose of the state-trace analysis are the number of hidden units and the learning rates, as well as the presence or absence of response units and presence of one or two connection weight components. Two units for both input and output were chosen to represent the stimuli (right or left circle fill) and predictions for the next trial (right or left), respectively. The activation of a single input unit was set to one, with the other set to zero to correspond to a left or right stimulus presentation. The units in each layer, from input and context to hidden and to output units, fed activation forward to

every unit in the layer above (see Figure 1). The activation of the hidden and output units were determined by the logistic activation function (Rumelhart, Hinton, & Williams, 1986). Hidden unit activation was copied back to the context units on each trial with a lag of one cycle of the network. Each hidden unit also had a bias: a variable connection from a unit that had a constant activation of one. The hidden units mapped recurrently to the context units on a one-to-one basis. The feed-forward connections comprised of either one or two connection weights. These were modified by the back-propagation algorithm, which we ran without a momentum term (Rumelhart et al., 1986).

To simulate the experiment with humans reported in Yeates et al. (submitted), each model was run 128 times to match the number of participants taking part in the empirical study. Half of these simulations acted as controls (trained on pseudorandom sequences), with half receiving experimental sequences. 32 experimental networks followed Group 1 rules ('first different to third') and 32 followed Group 2 rules ('first same as third'). Initial connection weights were set for each network to random values between -0.5 and 0.5. Each simulation involved training for one session and fifteen blocks of a second session, followed by five blocks of test sequences. Therefore each simulation received 4200 training trials and 600 test trials.

The mean square error (MSE) was calculated as the difference between the location of the next trial, and the prediction of the model (see Jones & McLaren, 2009). This was taken as the measure of performance of the model on the task. As in previous simulations of these tasks, the MSE for trials consistent with the trained rule was taken from the MSE for inconsistent trials (Jones & McLaren, 2009; Yeates et al., submitted). This produces an estimate of learning about those trained sequences, and is also computed for control simulations. Half of the control simulations are assigned to the dummy variable Group 1, where 'first different to third' subsequences (XXY, XYY, YYX, YXX) are taken from the matching 'first same as third' subsequence (XXX, XYX, YYY, YXY). The remaining 32 simulations follow the Group 2 inconsistent-consistent calculation, with the MSE on 'first same as third' subsequences taken from the MSE on 'first different to third' subsequences. Comparing the differences between experimental and control groups on these scores allows learning to be assessed without any confound in terms of sequential effects (see Anastasopoulou & Harvey, 1999; Jones & McLaren, 2009). To summarize then, good learning will result in a larger difference of the form (Control network MSE for inconsistent trials - Control network MSE for consistent trials) - (Experimental network MSE for inconsistent trials - Experimental network MSE for consistent trials), as a lower MSE indicates better learning.

### State-Trace Analysis 1: SRN and AugSRN

The task was simulated on an SRN and AugSRN with 20 hidden units. The SRN had a learning rate of 0.4, the AugSRN had a slow learning rate of 0.4 and a fast learning

rate of 0.533. The AugSRN also possessed response units, unlike the SRN.

**Results.** Both models produced significant learning of both sequences, which was analyzed by means of an ANOVA with subsequences and blocks as within-subject factors, and experimental versus control as a between subject factor. The SRN exhibited a significant difference in consistent-inconsistent MSE scores between experimental and control simulations,  $F(1,124) = 613.6, p < .001$ . The AugSRN also demonstrated learning,  $F(1,124) = 1113, p < .001$ . As this difference of differences measures the learning in experimental networks compared to networks that experienced pseudorandom sequences (controlling for sequential effects), this difference is used to provide our index of learning and performance in what follows.

The SRN and AugSRN constituted the two states we wished to analyze, and we plotted performance of Group 1 against Group 2 on the axes as the dimensions. The plots follow the trace of training over collapsed blocks of five, with the seven points shown constituting the 35 training blocks. Figure 2 shows the state-trace plot of this data.

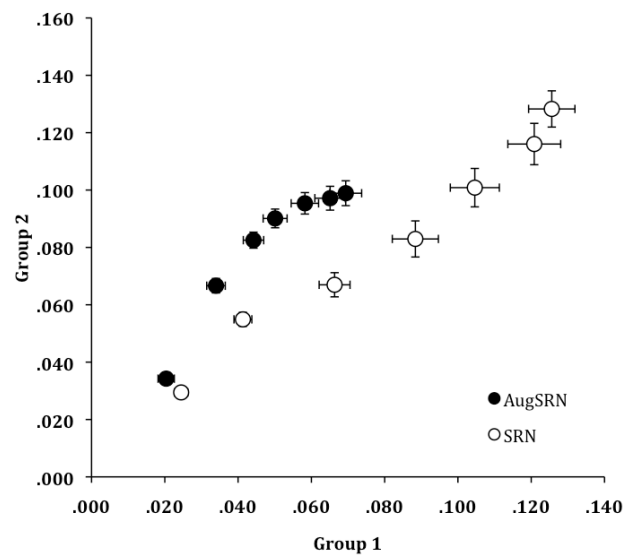


Figure 2. State-trace plot showing learning of AugSRN and SRN across training blocks of simulations.

Inspection of Figure 2 clearly suggests that there are two different, monotonic functions on the plot. We analyzed the data using hierarchical multiple regression, with the hypotheses that the model predicting Group 2 performance from Group 1 performance would be improved by the addition of Model Type as a variable, indicating a multi-dimensional model of the data. We simply coded this as a dichotomous nominal variable, with the AugSRN arbitrarily labeled as 1 and the SRN as 2. The addition of Model Type into the regression model significantly improved the  $R^2_{adj}$  value from 70.2% to 92.1%,  $\Delta R^2: F(1,11) = 34.1, p < .001$ , and overall, the model had a significant fit,  $F(2,11) = 76.7, p < .001$ ;  $\text{Group 2} = -.944 \times \text{Group 1} - .029 \times \text{Model Type} + .064$ . This corroborates the impression that the data on this plot require more than one function for a good fit, which

suggests that there is more than one underlying process governing performance in these simulations. The conclusion is that the SRN and the AugSRN differ in kind (which is perhaps not surprising – though they are very similar types of model), but the important finding here is that the state-trace methodology is sensitive to this difference.

**Discussion.** This confirms our predictions about how we expected state-trace to represent learning by these different networks on this task, producing different functions and so confirming that they are genuinely different types of model. This could easily be attributed to either or both of the two differences between the SRN and AugSRN. We now investigate whether these models are themselves best characterized as single or multi-process models of learning.

## State-Trace Analysis 2: Connection weight components

Here we ran simulations on four different models, two had response units and two had no response units. Within these dyads, we aimed to compare whether fast and slow weight components (the states in this state-trace analysis) were driving the multi-dimensional model seen in our first State-Trace Analysis. Therefore Model 1 had one connection weight component with response units, Model 2 had two connection weight components with response units (an AugSRN), Model 3 had one connection weight component with no response units (a standard SRN), and Model 4 had two connection weight components with no response units. Both had 20 hidden units and slow and fast weights of .4 and .533 respectively, as in the previous simulation.

**Results.** All four models learnt the sequences, analyzed as in Simulation 1. Model 1 showed a significant difference between experimental and control performance,  $F(1,124) = 853.6, p < .001$ . Models 2 and 3 showed learning, as seen in the results of State-Trace Analysis 1. Finally, Model 4 demonstrated the same learning,  $F(1,124) = 1634, p < .001$ .

When comparing models with one and two connection weight components we can see from Figures 3 and 4, which show the state-trace plots for models with and without response units respectively, that two monotonic functions appear.

When conducting a hierarchical linear regression as described in State-Trace Analysis 1, we this time coded Model as a predictor with the values of 1 and 2 for one and two components, respectively. Introducing Model into the regression alongside Group 1 in predicting Group 2 performance for models with response units (see Figure 3) produced a significant improvement in the  $R^2_{adj}$  value from 84.7% to 93.1%,  $\Delta R^2: F(1,11) = 15.9, p = .002$ , and overall, the model had a significant fit,  $F(2,11) = 89.4, p < .001$ ;  $\text{Group 2} = 1.231 \text{ Group 1} + .014 \text{ Model Type} - .007$ . Similarly, when there are no response units (see Figure 4) adding Model as a predictor improves the regression model, with a significant improvement in the  $R^2_{adj}$  value from 56.0% to 89.6%,  $\Delta R^2: F(1,11) = 40.0, p < .001$ , and overall, the model had a significant fit,  $F(2,11) = 56.9, p < .001$ ;  $\text{Group 2} = 1.008 \times \text{Group 1} + .042 \times \text{Model Type} - .041$ .

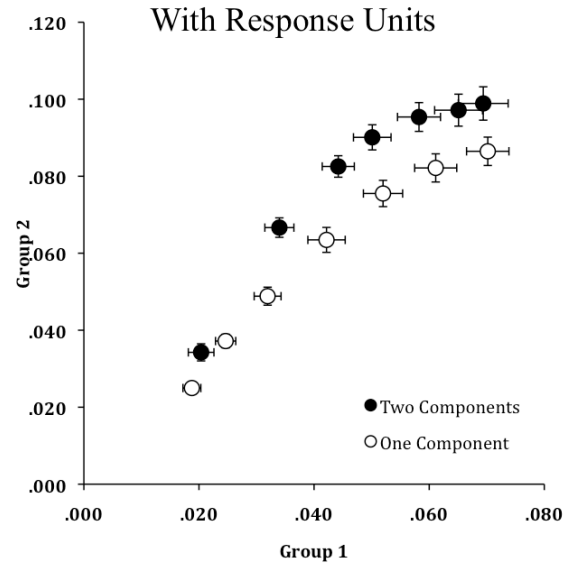


Figure 3. State-trace plot for Model 1 (one connection weight component) and Model 2 (two connection weight components) across training.

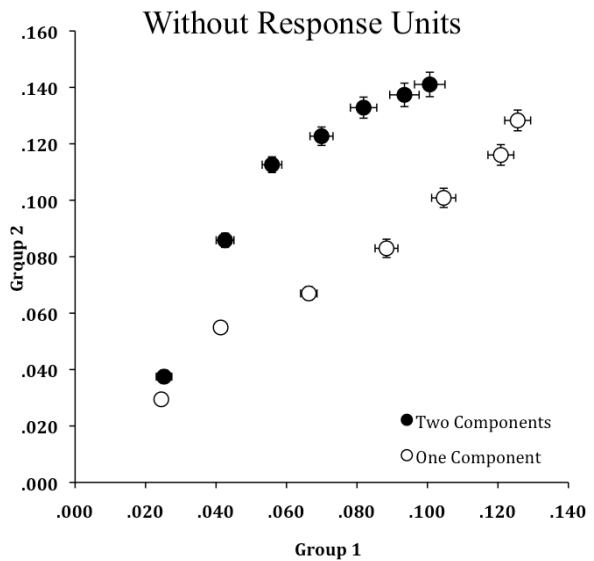


Figure 4. State-trace plot showing learning of Model 3 (one connection weight component) and Model 4 (two connection weight components), neither of which have response units, across training.

**Discussion.** Both in models with and without response units, multi-dimensional state-trace plots are produced when comparing those with one or two connection weight components. The state-trace analysis suggests that the two models are driven by different underlying processes, which in this case is due to the presence or absence of fast weights. Following the state-trace logic, this suggests that the two weight components within an AugSRN should be considered as distinct, different learning processes.

### State-Trace Analysis 3: Response units

Does the addition of response units to the basic SRN, or an SRN with two connection weight components, produce separate functions on the state-trace plot? The same four models are presented below, comparing Models 1 (no response units) and 3 (with response units), which both have one component, and Models 2 (no response units) and 4 (with response units), which both have two connection weight components.

**Results.** We compare Models depending on whether they have response units or not, coded as 1 and 0, respectively. See Figures 5 and 6 for state-trace plots of one and two connection weight component models. We find that in models with one component, adding Model as a variable improves the regression model, the  $R^2_{adj}$  value improves from 92.3% to 95.6%,  $\Delta R^2: F(1,11) = 10.3, p = .008$ , and overall, the model had a significant fit,  $F(2,11) = 143.8, p < .001$ ;  $\text{Group 2} = .942 \times \text{Group 1} + .013 \times \text{Model Type} + .006$ .

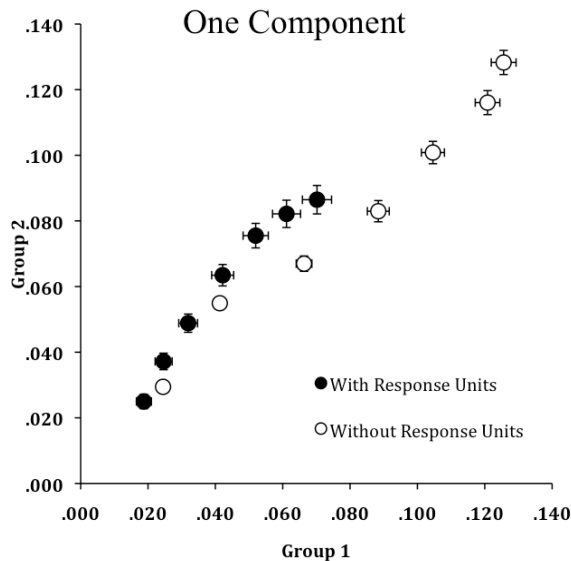


Figure 5. State-trace plot showing learning of Model 1 (with response units) and Model 3 (no response units), which both have only one connection weight component, across training blocks of simulations.

Comparing Models 2 and 4, with response units, the regression does not significantly improve when adding Model as a variable into the regression.

**Discussion.** Whilst the functions are not as distinct as in State-Trace Analysis 2, the comparison of models with and without response units still suggests a multi-dimensional structure. That the models with two connection weight components failed to reach significance is perhaps more a criticism of the linear regression method when analyzing these data. The fact that the separation of the two plots is less impressive (in size and reliability) when the presence (or not) of the response units is the manipulation than when the use of one vs. two sets of weights also suggests that for the type of model considered here, the main difference

between the AugSRN and the standard SRN is the distinction between fast and slow weights.

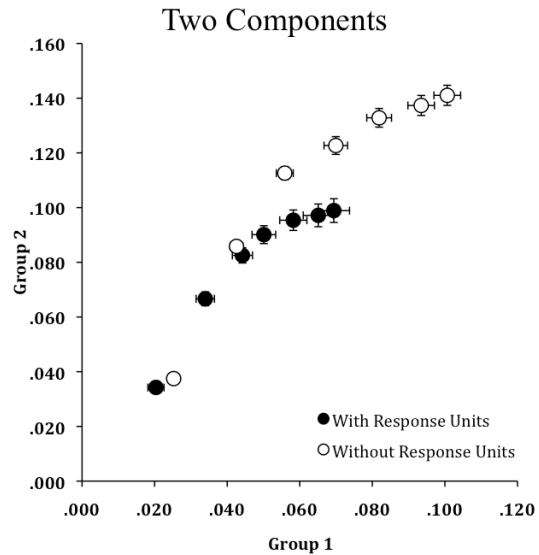


Figure 6. State-trace plot showing learning of Model 2 (with response units) and Model 4 (no response units), which both have two connection weight components, across training blocks of simulations.

### State-Trace Analysis 4: Learning Rates

Finally, to ensure that the differences seen in State-Trace Analysis 2 between one and two component models were not simply a result of the total amount or rate of learning, we varied the learning rates of the one component model, keeping the hidden units set at 20. We set the learning rate to 0.933, equal to the sum of the fast and slow learning rates employed in the AugSRN simulations to plot alongside the earlier one process model with a learning rate of 0.4. We are aware that a one component SRN with a learning-rate equal to the sum of two component's learning-rates is not a direct equivalent. Nevertheless, our manipulation should allow us to discover if varying learning rate over this range produces different state-trace plots.

**Results.** An SRN with a Learning Rate of 0.933 learns the task,  $F(1,124) = 556.8, p < .001$ . The state-trace plot of these data, alongside the original Learning Rate of 0.4 can be seen in Figure 7, which clearly shows two functions. When adding the learning rate as a regressor into a model predicting Group 2 performance from Group 1 performance, the  $R^2_{adj}$  value improves from 75.8% to 92.4%,  $\Delta R^2: F(1,11) = 27.2, p < .001$ , and this model overall had a significant fit,  $F(2,11) = 79.9, p < .001$ ;  $\text{Group 2} = .895 \times \text{Group 1} + .048 \times \text{Learn Rate} - .010$ .

**Discussion.** The state-trace plot and the regression analysis clearly demonstrate two separate functions, which according to state-trace analysis suggests the presence of multiple processes. However, the two models differ only in the values assigned to their learning rates. State-trace analysis proposes that a multi-dimensional state-trace plot will result from the presence of multiple processes in a given dataset,

which implies the influence of more than one latent variable within the SRN on performance. This suggests that either state-trace analysis is sensitive to differences within a single process, or alternatively the SRN must be considered a multi-process model of learning.

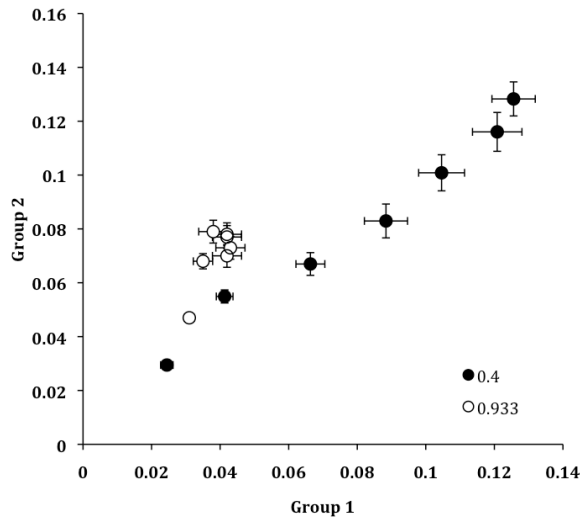


Figure 7 State-trace plot for an SRN with different learning rates (shown on graph).

## General Discussion

The analyses of SRNs with one or two learning components, and those with or without response units give separate functions on state-trace plots, suggesting the presence of more than one latent psychological variable. However, simple variation of the rate at which the SRN learnt also produced a multi-dimensional state-trace plot, which raises questions for the interpretation of multi-dimensional state-trace plots. A parameter search, varying learning rates and number of hidden units, has been conducted and, within a reasonable range for the SRN on this SRT task, produces the same functions as the data presented above. We recognise that the regression method employed in analysing the data is limited to roughly linear functions, and suggest that other methods (e.g. Newell, Dunn & Kalish, 2010; Prince, Brown & Heathcote, 2011) are also explored. But our analyses are, if anything, insensitive to the differences visualised by the plots, so this does not compromise our conclusions.

It seems, then, that not only multiple processes, but variations within a single process can produce multi-dimensional state-trace plots. The implications for state-trace analysis as a tool for the investigation of the number of latent variables underlying human behaviour needs to be considered, and further analysis of computational models with this technique is recommended.

## Acknowledgments

The research reported in this paper was supported by a Postgraduate studentship and Exeter Graduate Fellowship

awarded to Fayme Yeates, and an ESRC grant awarded to Ian McLaren and Fergal Jones.

## REFERENCES

- Anastasopoulou, T., & Harvey, N. (1999). Assessing sequential knowledge through performance measures: The influence of short-term sequential effects. *Quarterly Journal of Experimental Psychology*, 52A, 423-448.
- Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, 19, 171-181.
- Cleeremans, A. (1993). *Mechanisms of implicit learning: Connectionist models of sequence processing*. Cambridge, MA: The MIT Press.
- Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, 120, 235-253.
- Dunn, J.C. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review*, 115(2), 426-446.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.
- Jones, F. W., & McLaren, I. P. L. (2009). Human sequence learning under incidental and intentional conditions. *Journal of Experimental Psychology: Animal Behavior Processes*, 35(4), 538-553.
- Kinder, A., & Shanks, D.R. (2001). Amnesia and the declarative/non-declarative distinction: A recurrent network model of classification, recognition, and repetition priming. *Journal of Cognitive Neuroscience*, 13, 648-669.
- Lewicki, P., Czyzewska, M., & Hoffman, H. (1987). Unconscious acquisition of complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 523-530.
- Loftus, G.R., Oberg, M.A., & Dillon, A.M. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review*, 111(4), 835-863.
- McClelland, J.L., & Rumelhart, D.E. (1986). Amnesia and distributed memory. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing. Explorations in the microstructure of cognition: Psychological and biological models* (Vol. 2, pp. 503-527). Cambridge, MA: MIT Press.
- Newell, B.R., & Dunn, J.C. (2008). Dimensions in data: Testing psychological models using state-trace analysis. *Trends in Cognitive Science*, 12(8), 285-290.
- Newell, B.R., Dunn, J.C., & Kalish, M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition*, 38(5), 563-581.
- Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, 19, 1-32.
- Prince, M., Brown, S., & Heathcote, A. (2011, October 31). The Design and Analysis of State-Trace Experiments. *Psychological Methods*. Advance online publication. doi: 10.1037/a0025809
- Remington, R. J. (1969). Analysis of sequential effects in choice reaction times. *Journal of Experimental Psychology*, 82, 250-257.
- Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1, pp. 318-362). Cambridge, MA: Bradford Books.
- Yeates, F., Jones, F. W., Wills, A. J., McLaren, R., & McLaren, I. P. L. (submitted). Modelling human sequence learning under incidental conditions.



# The Role of Attention in Three-way binding in Episodic Memory

**Hyungwook Yim (yim.31@osu.edu)**

Department of Psychology & Center for Cognitive Science  
The Ohio State University  
209C Ohio Stadium East, 1961 Tuttle Park Place  
Columbus, OH 43210, USA

**Simon J. Dennis (dennis.210@osu.edu)**

Department of Psychology & Center for Cognitive Science  
The Ohio State University  
200E Lazenby Hall, 1827 Neil Avenue  
Columbus, OH 43210 USA

**Vladimir M. Sloutsky (sloutsky.1@osu.edu)**

Department of Psychology & Center for Cognitive Science  
The Ohio State University  
208C Ohio Stadium East, 1961 Tuttle Park Place  
Columbus, OH 43210, USA

## Abstract

The current study examines the role of attention in forming complex binding structures in episodic memory. Previous research (Yim, Dennis, & Sloutsky, 2011a, 2011b) indicated that three-way bindings can be formed within an explicit memory task. Here we attempt to reduce explicit attending by presenting participants with a variant of statistical learning. The paradigm was modified to accommodate the constraints in two list learning paradigms ABABr, and ABCD. Only the ABABr paradigm requires a three-way binding in order to perform above chance. Evidence of learning was derived from learning curves, accuracies and reaction time at test. Results show evidence of learning for the ABCD condition but not for the ABABr condition. This finding indicates that whereas the ABCD structure can be learned implicitly, learning of ABABr lists depends on attention and requires explicit learning.

**Keywords:** episodic memory, attention, three-way binding, context use, statistical learning

## Introduction

One of the memory tasks that we confront daily is finding where we parked our car at work. Even though the only thing that one needs to remember seems to be the parking spot, one's car, and the link between the two, we often fail. One of the reasons that we are sometimes unable to retrieve this memory is because our previous memories of parking the car interfere with today's memory. Therefore in order to correctly retrieve today's event, we also need to remember the link between the context (i.e. today) and the event (i.e. parking the car in a specific spot).

As depicted in the above example, one's success in retrieving an episode depends on what was remembered, and how it was structured when it was remembered. Depending on the episode that one has to remember, there could be differences in the number of components that have to be remembered, and the structure that has to be formed. The process by which one forms a memory representation

containing multiple components is called *binding* (Cohen & Eichenbaum, 1993; Schacter & Tulving, 1994). Along with the ability to store the bounded memory components, it has been argued that control processes such as memory strategies and metacognitive operations also play a central role in forming episodic memory (Ghetti & Lee, 2011). However, how binding occurs and the nature of the structure it forms are still not clear.

Yim, Dennis, & Sloutsky (2011a, 2011b) proposed that a key determinant of performance during episodic memory formation is simultaneously attending to the components that should be remembered (e.g. events, context). This attentional mechanism was argued to be used during encoding and/or retrieval especially when the binding structure is complex. To support the argument, they presented a list learning paradigm to different age groups (i.e. 4-year olds, 7-year olds, and adults). The participants had to remember two lists each consisting of six pairs of pictures. The two lists were separated by a retention interval and the pairs were presented one at a time. At the end of studying the two lists a cued recall test was given with a context cue and an item cue (see Figure 1-(a) left slide for an example of a trial which has a context cue and two items). The main manipulation was the structure of the lists which varied the complexity of their required binding structure.

In an ABABr condition two lists had an identical picture set with the only difference between the two lists being how they were paired (Porter & Duncan, 1953; Postman, 1964, see Figure 1-(b) left). For example, in the first list there would be two pairs, [apple]-[dog] and [chair]-[car], and in the second list there would be two different pairs, [apple]-[car] and [chair]-[dog], which is a re-arrangement of the first two pairs. It has been logically illustrated by Humphreys, Bain, & Pike (1989) that to correctly answer a given cued recall test (e.g. what was paired with apple in list1?) one must have formed, at the minimum, a *three-way binding* structure that includes the two items [apple] and

[dog], and the context [list1] together. Suppose that one formed a simpler binding structure such as a two-way binding structure (e.g. [apple] – [dog] or [apple] – [car]). In this case when cued with “apple” and “list1” in a cued recall test (e.g. what was paired with apple in list1?), the [apple] will not only elicit [dog] in list1 but also [car] in list2, which would make one’s recall ambiguous. However, if one formed a three-way binding structure, [apple] and [list1] will act as a compound cue, and will elicit the correct answer [dog].

In an ABAC condition (see Figure 1-(b) middle), the cues were identical between the two lists whereas the targets were different. In this condition the minimal binding structure that is required to remember the episode correctly are known to be *two two-way binding* structures (Barnes & Underwood, 1959; Postman, 1962). For example, one should have formed at least the binding between the two items (e.g. [apple] – [dog]) and the binding between an item and the context ([list1] – [dog]) to correctly answer at test (e.g. what was paired with apple in list1?). If one had only formed a single two-way binding structure in this case, it would be hard to recall the correct answer since [apple] has multiple two-way bindings, which are [dog] and [rat]. However, if one had formed two two-way binding structures, the item to context binding (e.g. [list1] – [dog]) would restrict the multiple bindings and lead to a correct response, [dog].

Finally, in an ABCD condition two lists contained different items (see Figure 1-(b) right). Therefore a single two-way binding structure between the two items would be sufficient at test without considering the context (provided the items are selected so that there are no preexisting bindings between them).

A multinomial process tree (MPT) model (see Batchelder & Riefer, 1999 for a review) was used to quantify the relative contributions of the cue-target, context-target and the three-way binding structures. Differences among age groups were mainly restricted to the three-way binding structure. In particular, the improved ability to form three-way bindings differentiated the 7 year olds and adults, suggesting that development of the critical mechanisms is extended, perhaps through the teenage years. In a follow up study, Yim and colleagues increased the saliency of the context cue for 4 year olds (e.g. visiting Elmo’s house instead of visiting a green house, see Figure 1-(a) right slide). By increasing the saliency of the context cue, 4 year olds increased the ability to use the context information (list). If the formation of episodic memory (or three-way binding structure) only relied on having the representational capacity to bind memory components, the manipulation of stimulus attention (or saliency) should not affect one’s performance. However, the results suggest that simultaneous attention to components is required for successful binding. It was also argued that since children have low attentional control (Zelazo, Carlson, & Kesek, 2008), the developmental mechanism underlying episodic

binding would be relying on attentional control during encoding and/or retrieval.

From the above results, it seems that attention plays an important role in binding, especially for the three-way binding structure. The goal of this research is to test this hypothesis directly. To ensure low or no attention, a statistical learning paradigm (Saffran, Aslin, & Newport, 1996) was used by modifying the task used in Yim, Dennis & Sloutsky (2011). In the modified task, participants saw a sequence of cartoon characters one at a time and their task was to distinguish whether the character was a male or a female while their accuracy and reaction time were measured. The sequence had a specific pattern resembling the ABCD condition and the ABABr condition that were used in the previous study. However, instead of presenting the two items and context together, the items and context were presented sequentially using pictures of cartoon characters (see Figure 1-(c)). Therefore, there was always a picture representing the list context followed by two pictures which represented the items. The main prediction was that if two items and context are bound together during learning, the triplet would be segmented as in the original statistical learning paradigm. Therefore, it would be possible to predict the third picture after seeing the first two pictures in the triplet (Turk-Browne, Simon, & Sederberg, in press). As a result, faster reaction time or higher accuracy at the third item in a triplet would be an indicator of learning the triplet.

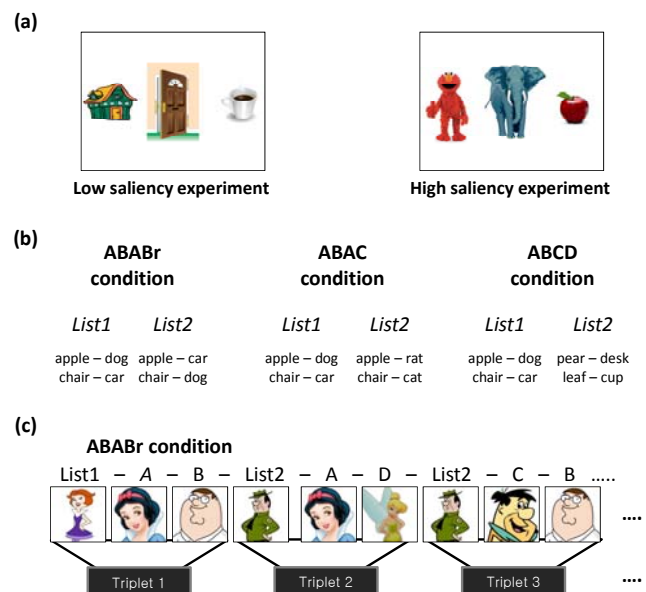


Figure 1: An illustration of experiment used in Yim, Dennis, & Sloutsky (2011a, 2011b) and its modification. (a) Stimuli used in the original experiment. The ‘green house’ and ‘Elmo’ represents the context cue whereas the ‘door’/‘apple’ represents the item cue, and the ‘cup’/‘elephant’ represents the target items (b) An example of the structure of the list in three conditions. (c) A modification of the original list learning paradigm (ABABr condition) into a statistical learning paradigm.

## Experiment

### Method

**Participants** Eighty undergraduate students at The Ohio State University participated for course credit (44 females,  $M = 19.11$  years,  $SD = 1.38$  years). Ten additional participants were excluded from the sample due to pressing the wrong key ( $N=5$ ), chance level accuracy ( $N=4$ ), and not understanding the instructions ( $N=1$ ).

**Stimuli** The stimuli were thirty six pictures of cartoon characters where half were male and the other half female. Post-experimental questions indicated that the sex of each character was easily determined by the participants. Each picture was presented on a white squared patch (10.55cm  $\times$  10.55 cm), which was centered on a black background of a 17inch computer monitor. For every participant, the pictures were randomly assigned to each experimental condition following the constraint of each condition.

The practice phase preceded the experiment and had five male and five female pictures where the order of presentation was randomized. The pictures in the practice were not used in the experiment. All experimental conditions including the baseline condition consist of triplets (i.e. context and two items). In each experimental condition, there were four unique triplets in the learning phase and an additional four unique triplets at test (see Figure 2-(a), (b)). The triplets in the learning phase followed the structure of the two list learning paradigms (i.e. ABABr, and ABCD) but were presented one at a time following the order of [context]-[item1]-[item2]. At test, participants were presented with triplets consisting of the learned elements, such that some of the triplets conformed to the learned statistics (i.e., Congruent) and some did not (Incongruent). Since the current paradigm examines the existence of a binding structure via the predictability of the third item based on the first two items, it is important to manipulate the third item at test while controlling the first two items. For example, in the ABABr condition if the first triplet (X1-A-B) was learned, which would mean that a three-way binding structure among the context (X1) and the two items (A and B) is formed, after seeing a sequence of X1 and A, it would be able to predict that the next item would be B. Therefore, response time or/and accuracy of item B would be faster or higher. On the other hand, a corresponding incongruent triplet (i.e. X1-A-D) would results in a slower response time or lower accuracy since the sequence violates the learned statistics and would interfere with the learned prediction “B”.

In each learning phase, there were ten repetitions for each unique triplet. In a repetition there were four unique triplets which result in a total of 120 (10 (repetition)  $\times$  4 (number of triplets)  $\times$  3 (pictures in a triplet)) trial per learning phase. In addition to the learning phase, the test phase consisted of 8 triplets, which had 4 congruent triplets identical to the learning phase and 4 incongruent triplets. Therefore the test phase had 24 trials (8 (triplets)  $\times$  3 (trials per triplet)) Therefore, there was a total of 144 trials for each condition.

There was also a baseline condition, which was designed to have no predictability of the third picture based on the first two pictures. Therefore, the triplets were all possible combinations using two pictures at each position (see Figure 2-(c)). The goal of the baseline condition was to measure the latency decrease that was due to task familiarity rather than to statistical learning. The total number of trials for the baseline condition was same as the experimental conditions. However, since there were eight unique triplets in the baseline condition, the whole triplet was repeated five times instead of ten. Finally, all conditions had the same number of male and female picture in each position of the triplet, and the frequency of the last picture in each triplet was equated for each condition.

(a) ABABr			(c) Baseline	
learning	test		learning	test
	congruent	incongruent		congruent
X1-A-B	X1-A-B	X1-A-D	X5-O-P	X5-O-P
X1-C-D	X1-C-D	X1-C-B	X5-Q-P	X5-Q-R
X2-A-D	X2-A-D	X2-A-B	X5-O-R	X6-Q-P
X2-C-B	X2-C-B	X2-C-D	X5-Q-R	X6-O-R
(b) ABCD			X6-O-P	
learning	test		X6-Q-P	
	congruent	incongruent	X6-O-R	
X3-G-H	X3-G-H	X3-G-K	X6-Q-R	
X3-M-H	X3-M-H	X3-M-K		
X4-J-K	X4-J-K	X4-J-H		
X4-N-K	X4-N-K	X4-N-H		

Figure 2: The structure of stimuli in each condition. Every letter represents a picture that was presented sequentially where X denotes a context while other letters denote items.

**Procedure** The experiment consisted of three blocks in addition to the practice block, where each block was assigned to a specific condition (i.e. the ABABr condition, ABCD condition and, the baseline condition). Each block consisted of a learning phase and a test phase, and the transition between learning and test was unbeknownst to the participants while there was a break between blocks. Each participant experienced all four conditions, where the order of the conditions was randomized and the participants did not know the identity of each condition. The procedure of each condition was identical except for the stimuli and the sequence of each stimuli being presented as explained in the stimuli section. Participants were told that they would see cartoon characters and their job was to distinguish whether it was a male or a female. The pictures were presented on the screen until the participants respond, and the next picture was presented after a 750msec ISI. The participants were not informed that there was a triplet structure in each condition. The test phase consist of 4 old triplets from the learning phase and 4 new triplets that where incongruent to the triplets in the learning phase. The order of the triplets was randomized.

After the experiment, participants were asked whether they saw a pattern in the sequence and whether there were any cartoon characters for which it was difficult to determine their sex.

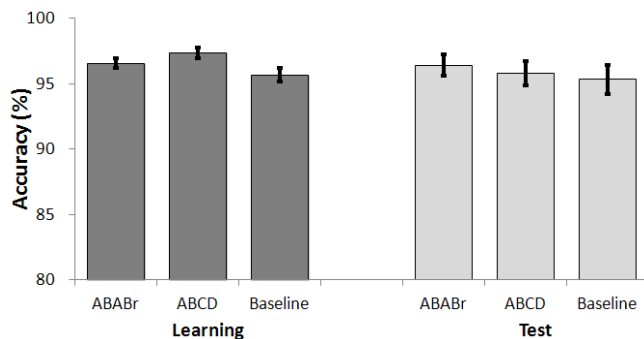


Figure 3: Accuracy for each condition at learning and test. Error bars indicate +/- one standard error.

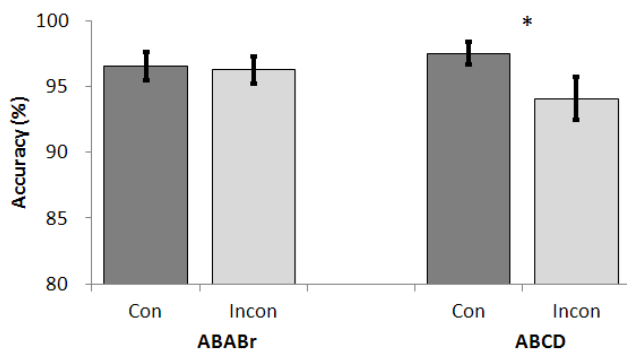


Figure 4: Comparing accuracy for congruent (Con) and incongruent (Incon) triplets at each condition at test. Error bars indicate +/- one standard error. \* indicates  $p < .05$

## Results

The results from the post-experimental questions show that all pictures were distinguishable. Moreover, there were no participants who reported finding a pattern in the sequences, thus confirming that learning was indeed implicit.

For the analysis, only the reaction time and accuracy of the third position in the triplet was used. The overall accuracy was 96.48% ( $SD = 2.71\%$ ), with 96.64% ( $SD = 2.69\%$ ) for learning and 95.66% ( $SD = 5.03\%$ ) for test. A  $2 \times 3$  (Phase: Learning vs. Test by Condition: ABABr, ABCD, and baseline) within-subjects ANOVA was conducted on accuracy showed no effect on Phase ( $p = .247$ ), nor Condition ( $p = .220$ ), and no interaction ( $p = .565$ ) (see Figure 3). A  $2 \times 2$  (Congruency  $\times$  Condition) within-subjects ANOVA conducted on the accuracy at test only showed a marginal main effect for Congruency ( $F(1, 79) = 2.95, p = .09, \eta_p^2 = .036$ ), whereas there was no significant main effect for Condition ( $p = .625$ ) nor significant interaction ( $p = .159$ ) (see Figure 4). Conducting a paired t-test between the congruent and incongruent triplets in each condition showed a significant difference in the ABCD condition ( $t(79) = 1.95, p < .05, d = .30$ ) but not in the ABABr condition ( $p = .81$ ).

Before analyzing the reaction time data (RT), values below 250msec and above 2500msec were excluded as outliers. The excluded data was .23% of the total learning data, and .08% for the test data. Also the median value was used for each subject's RT for each repetition in a condition.

The overall mean RT for learning was 577msec ( $SD = 203$ msec), 589msec ( $SD = 190$ msec) for the ABABr condition, 573msec ( $SD = 189$ msec) for the ABCD condition, and 551msec ( $SD = 80$ msec) for the Baseline condition (see Figure 5-(a)). To analyze learning during each condition, the asymptote of each learning curve was calculated. The asymptotic point was chosen by examining the last four repetitions for the experimental conditions and the last two repetitions for the baseline condition (cf. note that experimental conditions of four unique triplets and the baseline condition has eight unique triplets). The median value of each subject's RT was calculated among the asymptotic points and was analyzed using one way within-subjects ANOVA. Results show that there was a significant difference among conditions ( $F(1.95, 154.30) = 3.49, p < .05, \eta_p^2 = .042$ ). Conducting a pair-wise comparison with Bonferroni adjustments showed that the ABABr condition was significantly different from the ABCD condition ( $p < .05$ ), and not different from the Baseline condition ( $p = 1.00$ ). Difference between the Baseline condition and the ABAC condition was also marginally significant ( $p < .10$ ) (see Figure 6).

The mean RT for the ABABr condition at test was 572msec ( $SD = 121$ msec) for the congruent trials, and 563msec ( $SD = 150$ msec) for the incongruent trials. For the ABCD condition it was 543msec ( $SD = 118$ msec) for the congruent trials, and 577msec ( $SD = 162$ msec) for the incongruent trials. The congruent trials for the Baseline condition was 546msec ( $SD = 86$ msec). The analysis of test data suggests that for the ABCD conditions the congruent triplets elicited faster response than the incongruent triplets, which was not the case for the ABABr condition (see Figure 5-(b)). A  $2 \times 2$  (Condition  $\times$  Congruency) within-subjects ANOVA conducted on the test data showed no main effect for Condition ( $p < .518$ ) nor Congruency ( $p < .201$ ), but a significant interaction ( $F(1, 79) = 7.63, p < .005, \eta_p^2 = .088$ ). Further comparison between the congruent and incongruent triplets within each condition was conducted by a paired t-test. Results showed a significant difference in the ABCD condition between congruent and incongruent triplets ( $t(79) = 3.24, p < .005, d = .24$ ). However there was no significant difference between the congruent and incongruent triplets in the ABABr condition ( $p = .534$ ).

In sum, the learning data showed that the ABCD condition had a significant difference in the amount of learning compared to the baseline condition, whereas the ABABr did not. Results from the test data indicated that only the ABCD condition, but not the ABABr condition, exhibited evidence of learning.

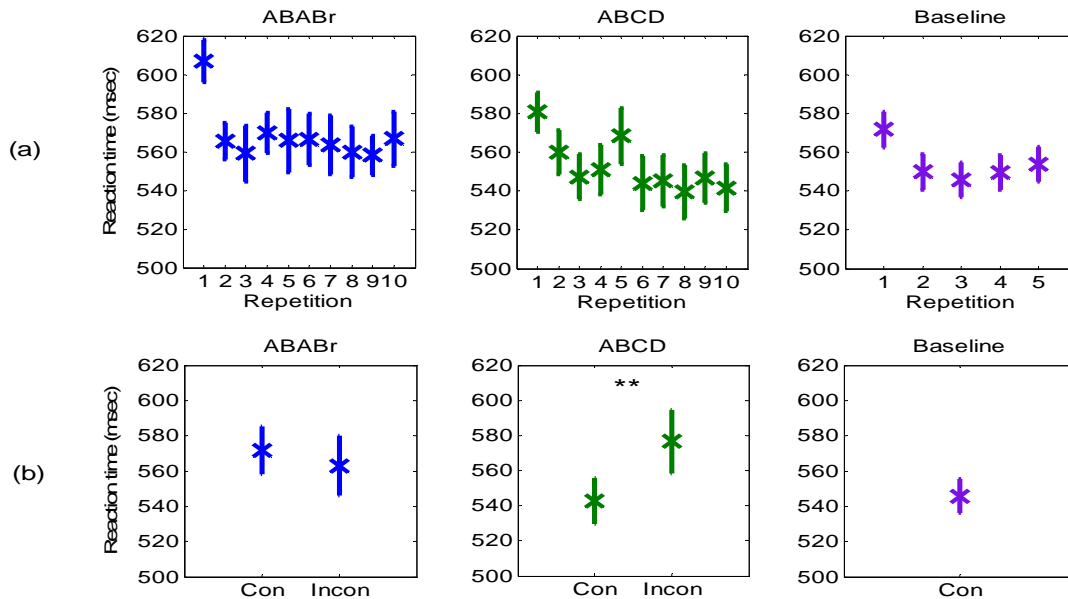


Figure 5: RT at each condition during (a) learning at each repetition, (b) test at congruent (Con) and incongruent (Incon) trials. Error bars indicate  $\pm$  one standard error. \*\* indicates  $p < .005$ .

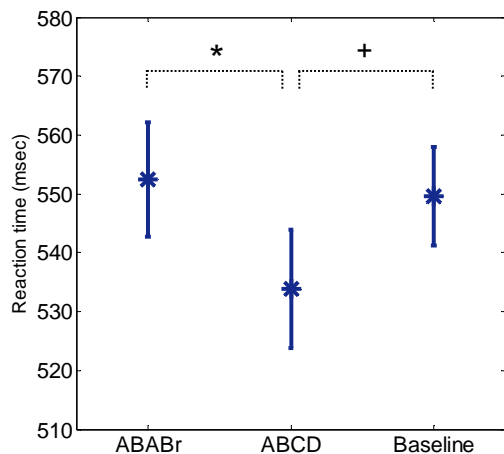


Figure 6: Comparing the asymptotic point for each condition during learning. Error bars indicate  $\pm$  one standard error. \* indicates  $p < .05$ , + indicates  $p = .10$

## General Discussion

The current study examined the possibility of forming a complex binding structure without attention by presenting participants with an implicit statistical learning task. Results showed that there was evidence of learning only in the ABCD condition but not in the ABABr condition. These findings suggest that whereas forming a two-way binding structure does not require attention, a three-way binding structure does require attention.

The results could be viewed from the point of view of the sequential learning literature (e.g. Reber, 1993) where the conditional probability of the 1<sup>st</sup> item (context item) and the 2<sup>nd</sup> item (item1) in a triplet plays a role in predicting the 3<sup>rd</sup> item (item2). Based on the predictability of each condition,

the ABABr in nature is the hardest condition to learn since the first two items conjointly should predict the third item. This argument is also consistent within the memory literature where a three-way binding structure is required to correctly recall in an ABABr condition - the two cues are conjointly bound with the target item (Humphreys, Bain, & Pike, 1989). On the other hand, the ABCD condition only requires association between the 2<sup>nd</sup> and 3<sup>rd</sup> item, which is also consistent with the memory literature - two-way binding structure.

However, predictability might not be the only factor that plays a role in the ABABr condition. The baseline condition has no predictability since it consist of all possible combinations using binary values at each position of a triplet. Thus, there should be no advantage from predicting the 3<sup>rd</sup> item. However, comparing the ABABr condition and baseline condition at test and learning shows no difference between the two conditions. Since there is evidence that 2<sup>nd</sup> order sequences are learnable (Stadler & Frensch, 1998), learning the ABABr condition should gain from predictability. Therefore, it is possible that the ABABr condition confronts additional interference which could not be alleviated without attention (cf. however, it is arguable that the ABABr condition requires more learning trials.)

The argued binding mechanism has some similarities compared to the binding mechanisms in other fields. In visual perception, binding concerns with object recognition (Treisman, 1996). To recognize an object and differentiate it from others, one should properly bind the properties (e.g. color, shape, location, etc.) to the correct object. Known that different properties are processed in different areas of the brain (Hubel & Wiesel, 2004), correct binding could not be done without attention (Treisman & Schmidt, 1982). Relational reasoning also requires a binding mechanism

(Hummel & Holyoak, 2003), where one has to bind correct fillers (e.g. *cat*, *tiger*) to correct roles (e.g. *B* is **bigger than** *A*) to form a relational representation (e.g. a *tiger* is **bigger than** a *cat*), and to further generalize or infer the representation.

The similarities among these bindings are that incoming components have to be bound together for further processing, and that incorrect binding will produce an erroneous response. However, the level of binding seems to be quite different even though they could be on the continuum. The components in the binding process of visual perception are features that are within the visual object. Moreover, due to dedicated feature detectors the binding process during visual perception would require less elaborated attention than in episodic memory. On the other hand, binding in relational reasoning requires more elaborate attention than episodic memory since the binding structure (i.e. role) is not a simple association but has a specific structure (e.g.  $\sim$  is bigger than). There is no evidence that the binding process has the same underlying mechanism across these domains. However, based on the similarities at the computational level, it is possible that there could be common mechanisms that involve the attentional system.

Finally, the fact that the participants failed to learn in the ABABr condition without attentional mechanism raises a question about the nature of the ABAC condition. Both ABAC and ABABr conditions require a complex binding structure to correctly remember the episode, which is a three-way binding and two two-way binding respectively. In forming a three-way binding structure, one should attend to all three elements that should be bound together. Therefore, attention is critical mostly at encoding but also at retrieval. However, it is possible that the two two-way binding structures are formed as two independent two-way bindings and integrated at retrieval. Thus, attention will mostly be required at retrieval. To address these issues, future studies should utilize methods that could measure attention during encoding and retrieval separately.

## References

- Barnes, J. M., & Underwood, B. J. (1959). "Fate" of first-list associations in transfer theory. *Journal of Experimental Psychology*, 58(2), 97-105.
- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychological Bulletin & Review*, 6(1), 57-86.
- Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system*. Cambridge, MA: MIT Press.
- Ghetti, S., & Lee, J. (2011). Children's episodic memory. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2, 365-373.
- Hubel, D. H., & Wiesel, T. N. (2004). *Brain and Visual Perception: The Story of a 25-Year Collaboration*. New York: Oxford University Press.
- Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, 110, 220-264.
- Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different Ways to Cue a Coherent Memory System. *Psychological Review*, 96(2), 208-233.
- Porter, L. W., & Duncan, C. P. (1953). Negative Transfer in Verbal Learning. *Journal of Experimental Psychology*, 46(1), 61-64.
- Postman, L. (1962). Transfer of training as a function of experimental paradigm and degree of first-list learning. *Journal of Verbal Learning and Verbal Behavior*, 1, 109-118.
- Postman, L. (1964). Studies of Learning to Learn II. Changes in Transfer as a Function of Practice. *Journal of Verbal Learning and Verbal Behavior*, 3, 437-447.
- Reber, A. S. (1993). *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious*. New York: Oxford University Press.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926-1928.
- Schacter, D. L., & Tulving, E. (1994). *Memory Systems 1994* (1st ed.). Cambridge, MA: The MIT Press.
- Stadler, M. A., & Frensch, P. A. (1998). *Handbook of Implicit Learning*. Thousand Oaks, CA: Sage Publications, Inc.
- Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology*, 6, 171-178.
- Treisman, A., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology*, 14, 107-141.
- Turk-Browne, N. B., Sederberg, P. B., & Simon, M. G. (in press). Scene representations in parahippocampal cortex depend on temporal context. *Journal of Neuroscience*.
- Yim, H., Dennis, S., & Sloutsky, V. (2011a). The Development of Context Use and Three-way Binding in Episodic Memory. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3000-3005). Austin, TX: Cognitive Science Society.
- Yim, H., Dennis, S., & Sloutsky, V. (2011b). The Development of Context Use and Three-way Binding in Episodic Memory. *Poster Presented at the Seventh Biennial Meeting of the Cognitive Development Society (CDS)*. Philadelphia, PA.
- Zelazo, P. D., Carlson, S. M., & Kesek, A. (2008). The Development of Executive Function in Childhood. In C. A. Nelson & M. Luciana (Eds.), *Handbook of Developmental Cognitive Neuroscience*. Cambridge, MA: MIT Press.



# Visual Context Effects on Thematic Role Assignment in Children versus Adults: Evidence from Eye Tracking in German

Lu Zhang (lzhang@cit-ec.uni-bielefeld.de)

Pia Knoeferle (knoeferl@cit-ec.uni-bielefeld.de)

Cognitive Interaction Technology Excellence Center, University of Bielefeld,  
Bielefeld, Germany

## Abstract

Prior research has shown that adults can make rapid use of visual context information (e.g., visual referential contrast and depicted agent-action-patient events) for syntactic structuring and disambiguation. By contrast, little is known about how visual context influences children's language comprehension, and some results even suggest children cannot use visual referential context for syntactic structuring (e.g., Trueswell et al., 1999). We examined whether children (unlike adults) also struggle to use other kinds of information in visual context (e.g., depicted events) for real-time language comprehension. In two eye-tracking studies we directly compared real-time effects of depicted events on children's (Exp1) vs. adults' (Exp2) processing of spoken German subject-verb-object (SVO) and object-verb-subject (OVS) sentences. Both of these word orders are grammatical, but OVS is a non-canonical structure. Five-year olds are at chance in understanding even unambiguous OVS sentences in the absence of visual context (Dittmar et al., 2008). If children can use depicted events rapidly for syntactic structuring, we should find similar visual context effects for them as have been reported for adults (Knoeferle et al., 2005), and similar gaze pattern as for the adults in the present studies. Gaze pattern in the present studies suggested that events depicting who-does-what-to-whom incrementally influenced both adults' and 5-year-olds' visual attention and thematic role assignment. Depicted-event information helped children to get rid of their initial preference for the preferred SVO structure when interpreting OVS sentences. However, visual context effects were subtly delayed in children (vs. adults), and varied as a function of their accuracy and cognitive capacity.

**Keywords:** eye tracking; child language comprehension; visual context; depicted events.

## Introduction

How visual context affects adult language comprehension has been extensively investigated. Adults can rapidly use visual context information such as referential contrast (e.g., Tanenhaus, Spivey, Eberhard, & Sedivy, 1995) and depicted events (Knoeferle, Crocker, Scheepers & Pickering, 2005; Knoeferle, Habets, Crocker, & Münte, 2008) for syntactic structuring and disambiguation. By contrast, for children's use of visual context in (real-time) language comprehension, there are conflicting results (Nation, Marshall & Altmann, 2003; Sekerina, Stromswold & Hestvik, 2004; Snedeker & Trueswell, 2004; Trueswell, Sekerina, Hill & Logrip, 1999; Weighall & Altmann, 2010).

In relating language to visual context, adults' adhere to what's been dubbed the 'referential principle' of language processing: When more than one syntactic analysis is offered

in parallel, the referentially supported (vs. unsupported) analysis is favored. Evidence for this claim comes from a study by Tanenhaus et al. (1995). Participants heard a sentence such as *put the apple on the towel into the box* and inspected related objects (e.g., an apple on a towel, another apple on a napkin, an empty towel, and a box). The prepositional phrase *on the towel* could temporarily either modify *the apple* or (ultimately incorrect) attach to the verb phrase *put*. When two apples (on a towel vs. on a napkin) were present, the phrase *on the towel* permitted adults to rapidly identify which apple the sentence was about, and to avoid a temporary misinterpretation of *on the towel* as the action destination. This was evidenced through eye gaze: adults rapidly inspected the apple on the towel and never the empty towel destination, much like they did for structurally unambiguous baseline sentences (*put the apple that's on the towel into the box*). They thus favored the referentially supported modifier over the referentially unsupported destination interpretation (Tanenhaus et al., 1995).

By contrast, 5-year old children (4;8 to 5;1 years) couldn't use the referential visual context and the underlying referential principle for online parsing decisions when tested with similar visual contexts and sentence ambiguity. In a study by Trueswell et al. (1999), participants either heard a locally structurally ambiguous sentence such as *Put the frog on the napkin in the box* or an unambiguous sentence such as *Put the frog that's on the napkin in the box*. For the ambiguous sentence, the prepositional phrase *on the napkin* can either modify the noun (location) or attach to the verb and specify the destination of the action (destination). Children inspected a table with either only one possible referent for *frog* in the 1-referent condition (e.g. a frog on the napkin, an empty napkin, a distractor and a box) or with two referents in the 2-referent condition (e.g. a frog on the napkin, another frog, an empty napkin and a box).

When 5-year old preschoolers heard the prepositional phrase *on the napkin* in the ambiguous instructions they frequently looked at the incorrect destination (the empty napkin) in both one-referent and two-referent contexts. This was interpreted as indicating that they – unlike adults in Tanenhaus et al. (1995) – incorrectly interpreted this phrase as the destination for *put*. It suggested children did not utilize the Referential Principle to guide their interpretation of the first noun phrase and the ensuing prepositional phrase. Moreover, their actions indicated that they never revised this initial misanalysis: On 60% of the trials the children performed an action that involved the incorrect destination



(e.g., moving a frog to the empty napkin before putting it in the box). The difficulty was attributable to ambiguity rather than structural complexity, since children mostly performed the action correctly (i.e., they moved the frog into the box instead of moving it to the empty napkin) with unambiguous sentences in a 2-referent context if they had moved the target animal first (the frog on the napkin).

The authors attributed the lack of visual context effects to a general inability of children to revise initial commitments made during syntactic structuring and semantic interpretation. Because five-year-old children have more limited processing capacities than adults, it may be difficult for children to make use of the referential principle (in real time) for resolving temporary syntactic ambiguities.

However, Meroni and Crain (2011) found a referential visual context effect in a post-sentence act-out task with a subtly-adjusted experimental design and procedure. Children inspected a two-referent context in which two frogs were placed on different-colored napkins. They were asked to close their eyes while listening to spoken sentences such as *Put the frog on the red napkin into the box*. In this setting, 3- to 5-year-olds exhibited adult-like performance in a post-sentence act-out task (i.e., they moved the frog that was on the red napkin directly into the box). The authors concluded that the changes in experimental setup and procedure enabled young children to inhibit their incorrect syntactic commitment and semantic interpretation. However, they only measured children's performance in a post-sentence task, and not the moment-by-moment processes of children's online language comprehension.

We argue that another reason for the lack of visual context effects in the study by Trueswell et al. may be potential difficulties in using visual referential contrast. Children heard *Put the frog on the napkin into the box*. At the napkin they had a choice in referential processing between looking at the empty napkin (a potential referent for *napkin*) or at the frog that's on the napkin. Since even young children rapidly fixate the picture of a word they have recognized (Hollich, Hirsh-Pasek & Golinkoff, 2000), it may be that children pursue primarily a "referential" strategy when hearing *on the napkin*, and prefer to look at the single empty napkin (vs. another object on a napkin). If children employ such a strategy, it would garden-path them even more since it would direct their gaze to the incorrect destination (the empty napkin). Thus, to the extent that mapping words onto objects (i.e., referential processes) governs child language comprehension, we cannot exclude that the way in which words in the utterance related to objects in visual context, rendered the use of information from that visual context unnecessarily difficult for children.

Related studies have examined the effects of other kinds of information in visual context (depicted action events) on children's comprehension accuracy for difficult-to-understand relative clause sentences (Weighall & Altmann, 2011). Children heard either a center-embedded (*The cat that bumped the bear will hug the cow*.) or a right-branching (*The cow will hug the cat that bumped the*

*bear*.) relative clause sentence. They saw clipart pictures either of the actions described by the relative clause (e.g., an orange cat bumping a bear and a striped cat bumping a sheep) or of several referents but without the actions (e.g. a bear, a sheep, an orange cat and a striped cat). When events were (vs. were not) depicted in visual context, children's accuracy on post-sentence comprehension questions was higher. These findings suggest that visual context can improve children's comprehension of difficult relative clauses and demonstrated that children can utilize extra-linguistic information such as depicted events at least for a post-sentence comprehension task.

Overall and specifically for *real-time* syntactic structuring and thematic role assignment, however, we still know little about how children use events depicting who-does-what-to-whom. We thus examined the effects of depicted events on children's versus adults' thematic role assignment in structurally unambiguous SVO and OVS sentences. Five-year-old children had difficulty in understanding unambiguous non-canonical German OVS sentences (Dittmar, Abbot-Smith, Lieven & Tomasello, 2008), and the same is true for adults (e.g., Matzke, Mai, Nager, Rüsseler, & Münte, 2002). For adults, we already know that for understanding structurally ambiguous or unambiguous OVS sentences, depicted action events can rapidly guide their visual attention and inform thematic role assignment (i.e., shortly after the verb that mediates relevant events, see Knoeferle et al., 2005, Knoeferle, 2007).

Examining effects of depicted events in children is interesting since we don't yet know whether events can rapidly guide their visual attention and thematic role assignment. Furthermore, action events involve a different relationship between the utterance (e.g., a verb such as "pushes") and visual context (e.g., a depicted action and its associated agent) than referential contrast. When children hear the verb – and assuming they incrementally establish reference to objects as has been found – they should rapidly look at the matching action, and notice its associated agent, which could help them to disambiguate / interpret the utterance.

If the lack of a visual context effects in young children results from the complexity of referential contrast rather than children's inability to use non-linguistic visual context, then we should find effects of visual context on thematic role assignment when the content of that visual context has a more direct relationship with words in the utterance (e.g., a verb that can be associated directly with a matching action). Furthermore, the studies by Meroni and Crain (2011) and Weighall and Altmann (2011) only reported children's performance in act-out and comprehension tasks, but not their continuous visual attention during spoken sentence comprehension. Continuous measures such as eye tracking can reveal more about language comprehension in real time. This measure provides a means for examining the moment-by-moment processes of children's spoken language comprehension, in the relatively natural situation of acting out spoken instructions, or answering questions.

To assess visual context effects on children's (vs. adults') processing of German SVO and OVS sentences in depicted event (vs. no-event) contexts, we conducted two eye-tracking studies - one study with children (Exp1) and the other with adults as participants (Exp2). We expected that if children benefit from the depicted events, they would show an adult-like gaze pattern and a higher accuracy in the comprehension task when events are present (vs. absent), particularly for the difficult OVS sentences. In addition, we obtained response accuracy in a post-sentence comprehension task, to ground our interpretation of eye-movement pattern in ultimate comprehension success. We also conducted cognitive tests to get insight into potential variation of context effect as a function of children's cognitive resources. We expected children with higher WM scores would show more adult-like gaze pattern.

## Experiments 1 and 2

### Participants





Thirty-two kindergarten children (15 4-year olds and 17 5-year olds, range: 4-5;10) took part in Experiment 1 and received a small toy for their participation. Thirty-two students of Bielefeld University received four euro each for taking part in Experiment 2. All participants had German as their only mother tongue and normal or corrected-to-normal vision. All were unaware of the experiment purpose. Children, one of their parents (Experiment 1), and adult participants (Experiment 2) all gave informed consent.

### Materials

We created 16 experimental items from 64 clipart pictures edited with commercially available graphics programs, and 64 sentences. An item consisted of four images and sentences. There were two version of an image; one depicted three characters as performing actions ('event', Pictures 1a and 1c, Table 1); the other depicted the same three characters without actions ('noevent', Pictures 1b and 1d). Each image was presented together with either a structurally unambiguous spoken subject-verb-object (Table 1, (1a) and (1a') SVO) or object-verb-subject sentence ((Table 1, (1b) and (1b') OVS)). In recording the sentences we took care to use neutral prosody. The design thus included two within-subject factors: *case marking* (SVO vs. OVS) and *event depiction* (event vs. no-event).

We counterbalanced the event roles of the characters; depiction of actions did not change. For example, one picture (Picture 1a) depicted the bear as pushing the bull, and the worm as painting the bear. The bear in the middle was thus role-ambiguous (agent and patient); the bull was the patient of the pushing action, and the worm the agent of the painting action. On the counterbalancing picture (Picture 1c), the bear was still role-ambiguous but now the worm was the patient of the painting action, and the bull the agent of the pushing action. This ensured visual characteristics of the characters could not confound looks to them as role fillers in an event. The two picture versions were each presented with a SVO and an OVS sentence. Event versus no-event pictures were presented with the same sentences (see e.g., Pictures 1a and 1b in Table 1).

Table 1: Example materials

Picture	Condition	Sentence
1a 	SVO-event	(1a) <i>Der Bär schubst sogleich den Stier.</i> The bear (subj) pushes immediately the bull (obj). 'The bear pushes immediately the bull.'
	OVS-event	(1b) <i>Den Bär malt sogleich der Wurm.</i> The bear (obj) paints immediately the worm (subj). 'The bear is immediately painted by the worm.'
1b 	SVO-noevent	(1a) <i>Der Bär schubst sogleich den Stier.</i> The bear (subj) pushes immediately the bull (obj). 'The bear pushes immediately the bull.'
	OVS-noevent	(1b) <i>Den Bär malt sogleich der Wurm.</i> The bear (obj) paints immediately the worm (subj). 'The bear is immediately painted by the worm.'
1c 	SVO-event	(1a') <i>Der Bär malt sogleich den Wurm.</i> The bear (subj) paints immediately the worm (obj). 'The bear paints immediately the worm.'
	OVS-event	(1b') <i>Den Bär schubst sogleich der Stier.</i> The bear (obj) pushes immediately the bull (subj). 'The bear is immediately pushed by the bull.'
1d 	SVO-noevent	(1a') <i>Der Bär malt sogleich den Wurm.</i> The bear (subj) paints immediately the worm (obj). 'The bear paints immediately the worm.'
	OVS-noevent	(1b') <i>Den Bär schubst sogleich der Stier.</i> The bear (obj) pushes immediately the bull (subj). 'The bear is immediately pushed by the bull.'

We also counterbalanced character orientation to ensure that the role-ambiguous middle character was oriented equally often to the left as to the right for the experiment items. In addition to the 24 experimental items, we constructed 8 filler items. Each filler item had a scene with two characters and a sentence accompanying it. Two started with an adverbial phrase; two with an unambiguously object case-marked noun phrase; two started with a subject case-marked noun phrase; two had two characters doing the same action. Sixteen additional fillers were created for the adults in Experiment 2. Eight of them had the same structure as the fillers in Experiment 1. Of the remaining 8, 4 had pictures that showed three characters doing different actions and 4 depicted only one character. The fillers ensured that the scene did not always depict three characters and that the verb was not always in the second position.

From the four conditions, and their two counterbalancing versions (for character role and orientation, see above), sixteen experimental lists were created, each consisting of 16 experiment and 8 filler items in Experiment 1 (children) and 16 experiment and 24 filler items in Experiment 2 (adults). Apart from the filler trials and the instructions (see Procedure), there were no differences between the child and adult experiments. Each participant saw only one of the four conditions of each item and the same number of items in each condition. Item order was pseudo-randomized individually for every participant. Adults saw at least one filler item in between two experimental trials.

## Procedure

An EyeLink1000 remote eye-tracker with a sampling rate of 500 Hz monitored participants' eye movements. Images were presented on a 22" LCD color monitor at a resolution of 1680×1050 pixels concurrently with the spoken sentences. We only tracked the right eye, but viewing was binocular. In Experiment 1, each child was instructed to play a game. In this game, children were asked to inspect the images and to listen to the sentences. After each trial, they heard a question about the previous sentence and were asked to try to answer it correctly. In Experiment 2, adult participants were instructed to listen to the sentences and inspect the pictures, and to answer the question accurately. For adults, we devised a story to cover our actual goals. The cover story was that the experiment examined how well adults can understand and concentrate on child materials.

Each trial started with the display of a central fixation dot followed by a picture. After a 2000-ms picture preview time, the sentence was played via speakers. Five hundred milliseconds after sentence offset, any depicted actions were removed (if present) and participants only saw the three characters. With only the characters present, a spoken question asked for either the subject or the object of the verb in the previous sentence (for example, *Wer malt hier?/ Wer wird hier gemalt?*, 'Who paints?/ Who is painted?'). Participants answered by naming the correct character. At the start of the experiment, each participant was shown two example images and sentences. Next, participants were set

up and calibrated manually using a five-point fixation stimulus in Experiment 1 and a nine-point fixation stimulus in Experiment 2. The black dot for adult calibration was replaced by a smiley to attract children's attention in Experiment 1. The EyeLink software validated calibration; if validation was poor, calibration was repeated until it was good. Between trials, participants fixated a centrally-located smiley (children) or black dot (adults). This allowed the software to perform a drift correction if necessary. The experiment lasted approximately 25 mins. After the eye-tracking part children completed a working memory (WM) test (a word ordering test; Kaufmann Assessment Battery for Children), and adults were debriefed.

## Analysis

We defined three analysis time windows: the verb region (from verb onset until its offset); the adverb region (from adverb onset until its offset) and the NP2 region (from NP2 onset until its offset). We coded participants' fixations to four areas of interest in the scene: the agent (e.g., the worm in Table 1 Pictures 1a and 1b); the patient (e.g., the bull in Table 1, Pictures 1a and 1b); the role-ambiguous middle character (the bear), and the background. Of those, the agent and patient were our target areas of interest. The proportions of fixation on the target areas of interest (the patient and the agent) were entered into log-ratio analyses (c.f., Arai, van Gompel & Scheepers, 2007; Carminati, van Gompel, Scheepers, & Arai, 2008; Knoeferle, Carminati, Abashidze & Essig, 2011). We computed mean log gaze probability ratios for the agent relative to the patient  $\ln(P(\text{agent})/P(\text{patient}))$  for each condition and time window. Then we entered the log probability ratios into a 2 (*case marking*) × 2 (*event depiction*) repeated measures ANOVA. Separate models were fitted for log-ratios averaged over participants and items. We report the *p*-values of these analyses.

## Results Experiment 1

**Response accuracy** Children's overall comprehension accuracy was 67%. Their accuracy was relatively high for SVO sentences in both the event and the no-event condition (87% vs. 77%). For OVS sentences, by contrast, their accuracy was higher when events were (vs. weren't) depicted (60% vs. 42%). A 2 (*case marking*) × 2 (*event depiction*) repeated measures ANOVA revealed a significant main effect of case marking ( $p < .001$ ) and of event depiction ( $p < .001$ ) by subjects in the absence of a reliable interaction. Spearman's *rho* confirmed that response accuracy was not significantly correlated with age.

**Eye-movement results** Figures 1(a) and 1(b) plot the mean log-gaze probability ratios (participants' means) of correctly answered trials per condition at the verb and adverb regions respectively. For the verb region effects involving the independent variables were not reliable (Fig. 1(a)). By contrast, during the adverb region, children looked more at the patient entity for SVO (vs. OVS) sentences and more at the agent entity for OVS (vs. SVO) sentences, but only

when events were depicted (Fig. (1b)). When there were no depicted events, the eye-gaze data showed that children didn't inspected the correct target object until it was named at NP2 region. Inferential analyses corroborated these effects, as evidenced by an interaction between case marking and event depiction for the adverb region that was significant by subjects ( $p < .005$ ). Thus, case marking at the first noun phrase was not sufficient for incremental thematic role assignment in children; the events, however, enabled correct thematic role assignment. Note that neither case marking nor prosody could account for the different gaze pattern observed in the event vs. no-event conditions since sentences were identical between these two factor levels.

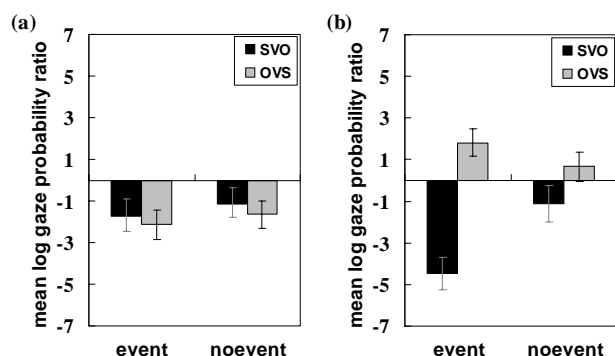


Figure 1: Children's mean log gaze probability ratios ( $\ln(P(\text{agent})/P(\text{patient}))$ ) per condition at the (a) verb region and (b) adverb region for correctly-answered trials in Experiment 1. Positive values indicate more looks to the agent; negative values indicate more looks to the patient.

We further split the subjects into high vs. low groups based on their median accuracy and working memory scores. The interaction between accuracy, case marking and events was significant ( $p < .05$ ), and there was a non-significant trend towards an interaction between WM, case marking and events ( $p < .15$ ). Separate analyses with the high vs. low groups showed that the interaction between case marking and events was significant by subjects for both the high accuracy ( $p < .005$ ) and high working memory groups ( $p < .05$ ) and marginal by items for the high accuracy group ( $p < .1$ ). Response accuracy was significantly correlated with WM scores (Spearman's  $\rho$ ,  $p < .005$ ). Furthermore, there was a hint that children's gaze pattern at the adverb region was correlated with their accuracy data as well (Spearman's  $\rho$ ,  $p < .1$ ).

## Results Experiment 2

**Response accuracy** Adults' accuracy was high (97%). For SVO sentences, their accuracy was 100% in both the event and the no-event condition. Their accuracy for OVS sentences was also high with no reliable difference for the event versus no-event condition (96% vs. 94%). Analyses

confirmed a main effect of case marking ( $p < .005$ ), but not of event depiction and no interaction of these two factors.

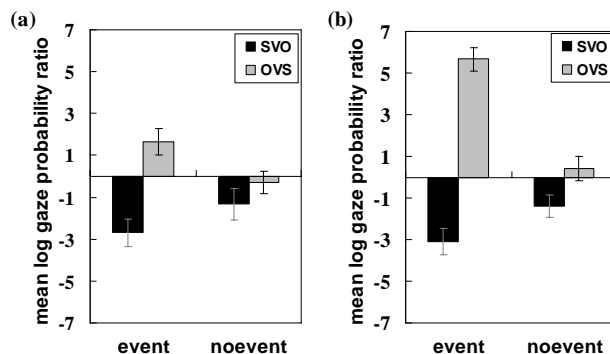


Figure 2: Adults' mean log gaze probability ratios ( $\ln(P(\text{agent})/P(\text{patient}))$ ) per condition at the (a) verb and (b) adverb region for correctly answered trials in Experiment 2. Positive values indicate more looks to the agent; negative values indicate more looks to the patient.

**Eye-movement results** Figures 2(a) and 2(b) present the mean log-gaze probability ratios (participants' means) for correct trials per condition at the verb and adverb region respectively for Experiment 2. Adults' (vs. children's) gaze data revealed an even earlier visual context effect, and a reliable interaction of case marking and event was confirmed at the verb (both  $ps < .01$ ). Adults began to look more at the target object for both SVO and OVS sentences at the verb when events were (vs. weren't) depicted (Fig. 2(a)). Figure 2(b) illustrates the continued visual context effects during the adverb, confirmed by a reliable interaction of case marking and event depiction ( $ps < .001$ ).

## Discussion

We assessed how visual context information such as depicted events influences children's online language comprehension and whether children can use visual context for language comprehension to the same extent as adults. To this end we recorded both children's (Experiment 1) and adults' (Experiment 2) eye movements to characters in clipart pictures as they listened to German subject-verb-object (SVO) and object-verb-subject (OVS) sentences. The clipart pictures depicted (vs. didn't depict) who-does-what-to-whom through action events.

The results of Experiment 2 confirm the view that visual context can influence adults' visual attention and sentence processing incrementally and rapidly. Adults looked more often toward the patient character for SVO (vs. OVS) sentences and more often towards the agent character for OVS (vs. SVO) sentences when events were depicted. This gaze pattern replicated the qualitative gaze pattern and its time-course in related adult studies (e.g., effects of case marking and event depiction in Knoeferle, 2007). Not unsurprisingly, effects for the unambiguous sentences in

Experiment 2 occurred slightly earlier (at the verb) than for initially structurally ambiguous sentences (Knoeferle et al., 2005, post-verbally). Adults' response accuracy was expectedly high in both SVO and OVS sentences.

Children's accuracy for the difficult OVS sentences was higher when action events were (vs. were not) depicted (Experiment 1). Thus, action events can improve their comprehension of unambiguous non-canonical OVS sentences for which they are otherwise at chance (Experiment 1 and Dittmar et al., 2008). Unlike suggested by prior results for locally structurally ambiguous sentences (e.g. Trueswell et al., 1999; Snedeker & Trueswell, 2004) we conclude that visual context information can help children to overcome an initial structural (SVO) preference. Future research will address whether our findings extend to locally structurally ambiguous sentences.

Furthermore, effects of the events on children's visual attention and syntactic structuring emerged *incrementally*. At the post-verbal adverb and thus one region before mention of the target entity, 5-year-old children fixated the patient more often when hearing a SVO (vs. OVS) sentence and the agent when hearing an OVS (vs. SVO) sentence but only when action events were depicted. Children's gaze pattern during the adverb suggests that visual context information (depicted agent-action-patient events) can rapidly influence their sentence processing and syntactic structuring for unambiguous but non-canonical sentences.

The interaction of case marking and event in the high (but not low) WM and accuracy groups suggests that visual context effects are sensitive to children's cognitive capacities and that they may co-vary in their emergence with increases in cognitive resources. Overall, however, visual context information can inform syntactic structuring and thematic role assignment incrementally and rapidly in both 5-year old children and adults.

## Acknowledgments

This research was funded by the Cognitive Interaction Technology Excellence Center (German research foundation, DFG) and a Fulbright fellowship to LK. We thank Linda Krull and Eva Mende for help with stimulus preparation and data collection; Maria Nella Carminati for advice on the analyses; and Helene Kreysa for advice on the Experiment Builder software. We also thank all the participating families and students for their support.

## References

- Arai, M., van Gompel, R., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, 54, 218–250.
- Carminati, M. N., Gompel, R. P. G. van, Scheepers, C., & Arai, M. (2008). Syntactic priming in comprehension: the role of argument order and animacy. *JEP: LMC*, 34, 1098–1110.
- Dittmar M, Abbot-Smith K, Lieven E, & Tomasello M. (2008). German children's comprehension of word order and case marking in causative sentences. *Child Development*, 79, 1152-1167.
- Hollich, G., Hirsh-Pasek, K., & Golinkoff, R. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs for the Society for Research in Child Development*, 65 (3, Serial No. 262).
- Knoeferle, P. 2007. "Comparing the time-course of processing initially ambiguous and unambiguous German SVO/OVS sentences in depicted events", in: R. Gompel van, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movement research: insights into mind and brain*. Oxford: Elsevier, 517 - 533
- Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment. *Cognition*, 95, 95-127.
- Knoeferle, P., Habets, B., Crocker, M. W., & Münte, T. F. (2008). Visual Scenes Trigger Immediate Syntactic Reanalysis. *Cerebral Cortex*, 18, 789-795.
- Knoeferle, P., Carminati, M., Abashidze, D., & Essig, K. (2011). Preferential inspection of recent real-world events over future events. *Front. Psychology* 2:376. doi: 10.3389/fpsyg.2011.00376
- Melchers, P. & Preuß, U. (2009). *Kaufman Assessment Battery for Children (deutsche Version)* (8., unveränd. Aufl.). Frankfurt/M.: Pearson Assessment.
- Meroni, L. & Crain, S. (2011). "How Children Avoid Kindergarten Paths". In Edward Gibson and Neal Pearlmuter (Eds.), *The processing and acquisition of reference*. Cambridge, MA: MIT Press.
- Nation, K., Marshall, C. M., & Altmann, G. (2003). Investigating individual differences in children's real-time sentence comprehension using language-mediated eye movements. *Journal of Experimental Child Psychology*, 86, 314–329.
- Sekerina, I., Stromswold, K., & Hestvik, A. (2004). How do adults and children process referentially ambiguous pronouns? *Journal of Child Language*, 31, 123-152.
- Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, 49, 238 –299.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632–1634.
- Trueswell, J., Sekerina, I., Hill, N., & Logrip, M. (1999). The kindergartenpath effect: Studying on-line sentence processing in young children. *Cognition*, 73, 89 –134.
- Weighall, A.R. & Altmann, G.T.M. (2010) The role of working memory and contextual constraints in children's processing of relative clauses. *Journal of Child Language*, 38, 579-605.

# Argument Homogeneousness and Structure Simplicity

Niina Ning Zhang (Lngnz@ccu.edu.tw)

Graduate Institute of Linguistics, National Chung Cheng University  
168 University Rd., Min-Hsiung, Chia-Yi 62102 Taiwan

## Abstract

Subjects and objects are arguments of verbs. They show either homogeneous or heterogeneous properties, with respect to time. The subject of *sleep* and the subject of *fall* are homogeneous and heterogeneous, respectively. In this research, we develop a novel analysis of the organization of arguments of various types of verbs. We argue that heterogeneous arguments are hosted in two levels of Verb Phrase, whereas homogeneous ones are hosted in one level of Verb Phrase. Therefore, homogeneous events and states are encoded in simpler syntactic structures than heterogeneous ones in natural language.

**Keywords:** argument; verb, syntax; language, simplicity; homogeneous; heterogeneous; transitive; intransitive.

## 1. Introduction

The goal of this study is to try to explore the cognitive view that different conceptualizations will lead to different syntactic structures and that the complexity of the former should correlate with the complexity of the latter. This study develops Klein's (2010) new theory of the arguments of verbs and proposes a novel analysis of the structures of the arguments of various kinds of verbs. Instead of the traditional triple division of transitive verbs (e.g., *kiss*), unaccusative intransitive verbs (e.g., *arrive*), and unergative intransitive verbs (e.g., *sleep*), we argue that verbs are first divided into homogeneous and heterogeneous ones, and then each type is further divided into transitive and intransitive subtypes. The two major types are different in the number of Verb Phrases (VPs) are involved in the syntactic structures: homogenous ones have only one-layer of VP, while heterogeneous ones have two-layers of VP. Therefore, in our understanding, homogeneous events and states are encoded in simpler syntactic structures than heterogeneous ones.

We will introduce a new classification of verbs based on Klein (2010), then present our proposal, in Section 2 and 3, respectively. In Section 4, we list our supporting facts. Finally, in Section 5, we make some general remarks about this new syntactic analysis.

## 2. Classification of Verbs

Based on the homogeneousness of the properties of an argument with respect to time, Klein (2010) discusses certain types of verb stems in their default readings. As pointed out by an anonymous reviewer, Klein's term time should be understood as state. I thus use the term state rather than Klein's term time. There are one-state arguments, which are homogeneous in the event or state, such as the subject of *laugh*. There are also two-state arguments, which

have a source state and a target state, and thus they are not homogeneous in the event or state, e.g., the subject of *fall*.

Homogeneous Intransitive (HOI) verbs have one argument with respect to one state, e.g., *sleep, dance, vibrate, be*.

Homogeneous Transitive (HOT) verbs have two arguments with respect to the same state over time, e.g., *weigh* with a measure phrase, *resemble, admire*.

Heterogeneous Intransitive (HEI) verbs have one argument with source time state and target time state, e.g., *die*, (intransitive) *drown, rise, remain*.

Heterogeneous Transitive (HET) verbs have two arguments — one at one time state, one with source time state and target time state. The state of the one-time state argument can overlap the source or target time states of the other argument, e.g., *leave, close, slay*, (transitive) *drown, observe*.

We summarize the classification in the following table (AS = Argument-State; Ss = source time state; St = target time state):

(1) Common AT-structures (cf. Klein 2010: 1231)

type	description	typical examples	AS skeleton
HOI	1 argument at one state	<i>sleep, dance, vibrate, be</i>	A   S
HOT	2 arguments at the same state	<i>weigh</i> with a measure phrase	A1 A2 \ S
HEI	1 argument with source state and target state	<i>die</i> , (intransitive) <i>drown, rise, remain</i>	A \ Ss St
HET	2 arguments—1 at one state, 1 with a source state and a target state.	<i>leave, close, slay</i> , (transitive) <i>drown, observe</i>	A1 A2   \ Ss St

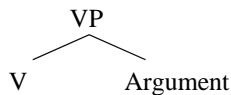
## 3. The Structure-Positions of Arguments

A well-adopted assumption is that all verbs are represented by a two-layer verbal projection: a vP to host the external argument of a transitive verb, or the unique external argument of an unergative verb, and a VP to host the internal argument of either a transitive verb or an

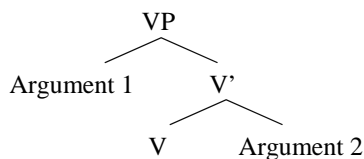
unaccusative verb. The assumption does not consider the contrast between HO and HE verbs at all. All external arguments are assumed to have the same syntactic position, i.e., Spec of vP.

We propose that arguments of the four types of verbs are base-generated as in (2).

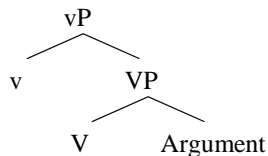
- (2) a. HOI: [<sub>VP</sub> V Argument]



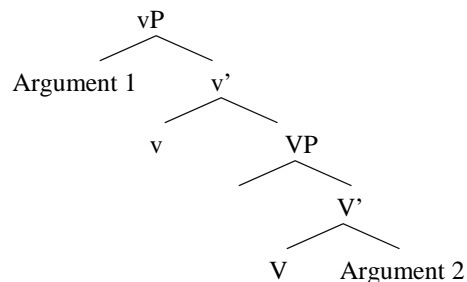
- b. HOT: [<sub>VP</sub> Argument 1 [V Argument 2]]



- c. HEI: [<sub>VP</sub> v [<sub>VP</sub> V Argument]]



- d. HET: [<sub>VP</sub> Argument 1 v [<sub>VP</sub> V Argument 2]]



One can see that HE verbs have vP, whereas HO ones do not, regardless of how many arguments occur. In (2b), the two arguments are hosted in the same VP; but in (2d), the two arguments are hosted in vP and VP, respectively.

In the structures for HO verbs, only one-layer of VP is projected, and there are maximally two positions for arguments: one is at Spec and the other is at Complement of V. In contrast, in the structure for HE verbs, there are two layers of VP, and thus one more position is available for an additional argument.

## 4. Supporting Facts

Three kinds of evidence supports our proposal in (2): the integrity of the verbs and the arguments of HO

constructions (4.1); the structural richness of HE constructions (4.2); and the special contrasts between HO and HE verb constructions (4.3).

### 4.1 Structure Integrity

All of the arguments that support severing the subject from VP or projecting of a vP shell (Marantz, 1984; Larson, 1988; Kratzer, 1996) come from HE, rather than HO, constructions. Typical tests are pseudo-cleft and the do-so replacement.

**4.1.1 Pseudo-cleft** Unlike HE verbs, HO verbs fail in pseudo-cleft (Zucchi, 1998: 349) (\* means the sentence is not acceptable).

- (3) a. What John did was eat an apple. [HE]  
b. \*What John did was resemble his father. [HO]

**4.1.2 vP-proform *do so*** Unlike HE verbs, HO verbs fail in replacement by the vP-proform *do so* (Ross, 1970; cf. Stroik 2001: 367).

- (4) a. Chris is leaving now, and Sam is doing so too. [HE]  
b. \*Mary likes Sam, and Chris does so too. [HO]  
c. \*The shoes cost 5 dollars, and the gloves do so too. [HO]

The two restrictions on HOT verbs are covered by (2b), where the whole VP is composed of two arguments and the verb. In the structure, the combination of V and Argument 2 is just part of a VP, which lacks syntactic visibility (Chomsky 1995). The combination is not a vP, either, since vP does not exist for this type of verbs. Therefore, it may not be replaced by *do so*.

### 4.2 Structure Richness

HE constructions have more argument positions than those of HO constructions. This can be seen in the following four facts.

**4.2.1 Double object constructions** Why is there no double object or applicative HO verb (5)? If only HE verbs may host their arguments in two layers of VP, the contrast is explained. (2b) does not have enough positions for three arguments.

- (5) a. I gave him the clothes. (HET)  
b. \*I like him the clothes. (HOT)  
Intended: 'I like him with respect to his clothes.'  
(6) a. John rented Bill a room. (HET)  
b. \*John resembles Bill the eyes.  
Intended: 'John looks like Bill with respect to their eyes.'

**4.2.2 Object control constructions** There is no HO object-control construction, which has two internal arguments: a nominal and a clause.



- (7) a. Mary forced John to feed the baby.  
 b. \*Mary admired John to feed the baby.  
 Intended; 'Mary admired John for his feeding of the baby.'

**4.2.3 Expansion from HO to HE constructions** Many verbs may occur in either HO or HE constructions (e.g. Dowty 1979: 60; Rosen 1999). It is easy to change an otherwise HO construction into a HE one by adding a delimitable element. But we do not add material to a HE structure to change it into a HO one (Thompson 2006: 218). HE structures are thus richer than HO ones.

(8) ADDITION OF DIRECT OBJECT

- a. Bill ran (\*in 5 minutes). [HO]  
 b. Bill ran the mile in 5 minutes. [HE]

(9) ADDITION OF INDIRECT OBJECT

- a. That book costs three dollars. [HO]  
 b. That book has cost me three dollars. [HE]

(10) ADDITION OF COGNATE OBJECT

- a. Terry sang (\*in an hour). [HO]  
 b. Terry sang the ballad in an hour. [HE]

(11) ADDITION OF X'S WAY EXPRESSION

- a. Terry sang (\*in an hour). [HO]  
 b. Terry sang her way to the Met in 10 years. [HE]

(12) ADDITION OF FAKE REFLEXIVE

- a. Terry sang (\*in an hour). [HO]  
 b. Terry sang herself to sleep in an hour. [HE]

(13) ADDITION OF RESULTATIVE

- a. Terry ran (\*in an hour). [HO]  
 b. Terry ran us ragged in an hour. [HE]

Note that heterogeneous events can be repeated (*Mary dried the dishes in an hour.* vs. *Mary dried the dishes for hours before being released from duty*) (Thompson 2006: 218; Ramchand 2008: 31). It is just like all nominals can be counted if an appropriate unit is identified (*two drops of water* as well as *two books*).

**4.2.4 Verb affix marking** Certain formatives mark the HE status of the verb, and their absence marks the HO status, e.g., *meg-* in Hungarian (Hopper & Thompson 1980: 267) (O means object):

- (48) a. *A gazda MEG-verte az inasokat.*  
 the boss PERF-beat(OBJ) the apprentices(ACC)  
 'The boss beat the apprentices.'  
 b. *A gazda verte az inasokat.*  
 the boss beat(OBJ) the apprentices(ACC)  
 'The boss would beat the apprentices.'

With the prefix *meg-* (Hetzron's 'effective aspect'), 48a means that the boss did beat all the apprentices on one occasion; the action is thus perfective and punctual, and the object is totally affected. But 48b, without *meg-* (Hetzron's 'descriptive aspect'), means that the boss was not above beating the apprentices, that he did it from time to time, but that not all the apprentices were necessarily involved; the action is claimed, then, to be imperfective and iterative, and the O is not totally affected.

From the above citation, we can see that the event reported in (48a) is bounded and thus the sentence is a HE construction, whereas the event reported in (48b) is unbounded and thus the sentence is a HO construction. Presumably, formatives such as *meg-* are licensed by vP, and thus the HE reading correlates with a richer structure.

**4.3 Special Contrasts Between HO and HE Verbs**

Three further contrasts between HO and HE verbs are reported as below.

**4.3.1 Intransitive verbs exhibit HO-HE contrasts** Only HEI verbs allow the expletive *there*, but HOI verbs may not, as seen in (14).

- (14) a. There arrived a train in the station. [HE]  
 b. \*There laughed a man in the hallway. [HO]

This contrast can be captured by the assumption that the expletive is base-generated in vP only (Deal, 2009). (2c), but not (2a), has vP, although both are for intransitive verbs.

**4.3.2 Transitive verbs exhibit HO-HE contrasts** We have seen the contrasts between HOT and HET in English in (4). Moreover, in some languages (e.g. Finnish, Hungarian) only HE structures have accusative case marker. The following Finnish examples are cited from (Hopper & Thompson 1980: 262):

- (15)  
 a. *Liikemies kirjoitti kirjeen valiokunnalle.*  
 businessman wrote letter (ACC) committee-to  
 'The businessman wrote a letter to the committee.'  
 b. *Liikemies kirjoitti kirjettä valiokunnalle.*  
 businessman wrote letter (PART) committee-to  
 'The businessman was writing a letter to the committee.'

In (15a), the presence of the ACC (accusative) marker with the direct object *kirjeen* 'letter' indicates a bound event: a letter was created at the target time, and thus the event was not homogeneous. In (15b), however, there is no ACC marker with the direct object, and the event could be homogeneous. See Rosen (1999) for more such examples.

The contrast can be captured in the contrast between (2b) and (2d): the overt accusative case marking is licensed by vP (cf. Chomsky 1995). Only in (2d), which is for HET, vP is projected and thus the ACC marker can be licensed. Since

there is no vP in (2b), which is for HOT, no ACC marker can be licensed in the structure.

**4.3.3 The prefix *re-*** Certain rules apply to HE constructions only, but not HO ones, regardless of whether the verb is transitive or intransitive. The English prefix *re-* occurs with HE verbs only, as seen in (16) (Horn, 1980). Since *again* behaves differently, the issue is not semantic.

- (16) a. The door reopened. [HE]  
 b. I reopened the door. [HE]  
 c. John {\*resmiled/smiled again}. [HO]  
 d. \*John re-admired his father. [HO]

Note that *re-* scopes over either the HE root or the affected nominal (Marantz, 2005). But the ambiguity is independent of the HE restriction.

**4.3.4 The time frame preposition phrases** Preposition phrases like *in an hour* are licensed by HE verbs, regardless of whether the verb is transitive or intransitive, as seen in (17). Thompson (2006) shows that such PPs are licensed by a projection higher than VP. Their absence in HO constructions indicates that the structure of HO constructions are lower and thus simpler than that of HE ones.

- (17) a. John walked to the store in two hours. [HE]  
 b. John destroyed the toy in two hours. [HE]  
 c. \*John slept in two hours. [HO]  
 d. \*John admired his father in two hours. [HO]

## 5. General Remarks

In this proposed new analysis of the time-argument structures of various types of verbs, the structures of homogeneous eventuality constructions are simpler than those of heterogeneous ones.

A parallel analysis of noun constructions is found in Borer (2005) and Zhang (2012). In Borer's analysis, CLP or DivP is projected for count nominals, but not for mass nominals. In Zhang (2012), DelimitP is projected for non-mass nominals, but it is absent in the structures of mass nominals. In both analyses, the structures of mass nominals are simpler than those of count nouns. The former shows homogeneousness, whereas the latter does not.

The significance of this study is that, like Borer's (2005) and Zhang's (2012) studies of nominal structures, our research of verbal structures here also show that the perceived homogeneousness in our understanding of the world, including events and individuals, correlates with the simplicity of linguistic structures.

## References

Borer, H. (2005). *In name only*. New York: Oxford University Press.

- Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.
- Deal, A. (2009). The origin and content of expletives: evidence from "selection". *Syntax* 12: 285-323.
- Dowty, D. (1979). *Word meaning and Montague Grammar*, Dordrecht: Reidel.
- Hopper, P. & S. Thompson. (1980). Transitivity in grammar and discourse. *Language* 56: 251-299.
- Horn, L. (1980). Affixation and the Unaccusative Hypothesis. *CLS* 16: 134-146.
- Klein, W. (2010). On times and arguments. *Linguistics* 48: 1221-1253.
- Kratzer, A. (1996). Servering the external argument from its verb. In Rooyck, J. and L. Zaring (eds.) *Phrase Structure and the Lexicon*. Dordrecht: Kluwer.
- Larson, R. (1988). On the double object construction, *Linguistic Inquiry* 19: 335-91.
- Marantz, A. (1984). *On the nature of grammatical relations*. Cambridge, MA: MIT Press.
- Marantz, A. (2005). Rederived Generalizations. Ms. MIT.
- Ramchand, G. (2008). *Verb meaning and the lexicon*. Cambridge: Cambridge University Press.
- Rosen, S. (1999). The syntactic representation of linguistics events. *GLOT International* 4.2: 3-11.
- Ross, J. (1970). Act. In D. Davidson & G. Harmon (eds.) *Semantics of natural language*, 70-126, Dordrecht: Reidel.
- Stroik, T. (2001). On the light verb hypothesis. *Linguistic Inquiry* 32: 362-369.
- Thompson, E. (2006). The structure of bounded events. *Linguistic Inquiry* 37: 211-228.
- Zhang, N. (2012). Numeral Classifier Structures. *lingBuzz/001197*. <http://ling.auf.net/lingBuzz/001197>.
- Zucchi, S. (1998). Aspect shift. In S. Rothstein (ed.) *Events and Grammar*, 349-370, Dordrecht: Kluwer.

# Modeling a Cognitively Limited Network in an Agent-Based Simulation

Changkun Zhao, Ryan Kaulakis, Jonathan H. Morgan, Jeremiah W. Hiam, Frank E. Ritter

(cuz111, rmk216, jhm5001, jwh189, frank.ritter@psu.edu)

The College of Information Sciences and Technology

The Pennsylvania State University

## Abstract

We investigate how cognitive capacity limits the number of group relations that a person can maintain. The simulation experiment's results using ACT-R and its memory equations replicated an effect similar to that of Dunbar's (1998) number, or the average total number of group ties capable of being supported in memory. In our study, we also examined the influences of two spatial factors (navigation strategies and map configurations) on the growth of generative networks. Our results suggest three interesting conclusions: (a) a fixed-path navigation strategy increases the speed that networks can form; (b) a higher *grid ratio* (connectivity of the agents' world) provides more chances for agents to build relations, and thus increases the network generation speed; but (c) neither factor influenced the total relations that an agent could maintain, which implies that Dunbar's number primarily depends on internal cognitive factors and less on external factors.

**Keywords:** ACT-R, Cognitive modeling, Social-cognitive network, network formation.

## Introduction

What does it mean to know someone? In colloquial English, this can imply anything from knowledge of someone's true character or secrets to a casual friendship. Nevertheless, this kind of knowing seems to imply more than a declarative association. I may know that Michelle Obama is Barrack Obama's wife, but I cannot say that I know either Michelle or Barrack Obama. Knowing in this context seems to imply knowledge not only of an individual's identity but also some knowledge of the significant relationships in their life, knowledge derived from direct interactions with that individual. I may get to *know of* Barrack Obama by reading his memoir but I get to *know* him in a social sense by speaking with him.

In this paper, we begin to explore what it means to know someone in a network, and how that knowledge influences our daily interactions. Drawing from Simon's (1991) work on bounded rationality in organizations and Dunbar's (1998) work examining the connections between cognition and language, we believe this form of knowledge reliably constrains organizations and moderates our behavior. Consequently, we seek to identify more concretely the mechanisms that underlie tie-formation, in other words the foundations of friendship. We begin by modeling the rate of tie formation in cognitively plausible generative networks.

Dunbar (1998) presents empirical evidence that suggests that human social networks are cognitively constrained. Chiefly, he argues that the neocortex size of humans limits the size of a fully connected human social network to about 150 ties. He defines a fully-connected social network as one

where all members can not only attach an identity but also a relation to all other members (Dunbar, 1998, pp. 66-68). He further argues that this constraint underlies the small-world effect observed by Milgram (1967) and others. He distinguishes this number of group ties from the number of sympathy ties, the number of intimates a person encounters in a month ( $n=11-12$ ), or the number of face-to-name matches a human can typically perform ( $n=1,500-2,000$ ). Dunbar infers these numbers and the relationship between neocortex size and social network size from empirical studies of human and non-human primates. He then compared these findings with anthropological evidence, finding his predictions basically matched the anthropological data.

McCarty, Killworth, Bernard, Johnsen, and Shelley (2001) propose a far larger number ( $n=291$ ) as an average network size. In part, this discrepancy is rooted in a difference in definitions. McCarty et al.'s (2001) definition of a social tie requires mutual identification as opposed to Dunbar's stricter definition of mutual identification and placement in the network. Also, McCarty et al. suggest other possible sources of discrepancy such as responder biases (number preferences and individual differences), size effects that influence the respondents' ability to accurately estimate the number of acquaintances associated with either very small or large subpopulations, and analysis errors arising from missing data or numerical biases introduced when combining studies. Nevertheless, neither of these potential sources of error nor the difference in definition seem to entirely account for the wide discrepancy in these estimates because, while McCarty et al. allude to ecological effects, neither they nor Dunbar systematically account for them. Also, it remains an interesting question as to what extent the difference in definitions contributes to the difference in estimates.

Ecology (defined here as an actor's physical and social environment) influences cognition not only by presenting a set of opportunities and resources but also by moderating our perceptions of those opportunities (Brantingham & Brantingham, 1993). We also know that humans are sensitive to environment when recalling sets of relations, using different approaches in different settings (Metz & Shultz, 2010). There has been far less work, however, examining to what extent ecology reliably influences tie formation in memory. For instance, one criticism of Dunbar's estimate is that it does not include the significance of environmental complexity. In other words, would the number of social ties, as Dunbar defines them, ever emerge in a distributed social structure like Suburbia, or could it? Does neocortex size impose a maximum, or is the relationship more complex?

More generally, how does memory and environment mediate and constrain social networks?

We begin to examine these questions by modeling the effects of memory on the formation of generative networks (networks arising from an initially empty set). We use McCarty's et al.'s (2001) definition of ties, that is, identification is sufficient to constitute a tie in this experiment. We believe that understanding the rate of network formation is a necessary precondition for reconciling Dunbar's and McCarty et al.'s estimates because this rate, in itself, constrains the opportunities available in the social set. In other words, the rate of network formation influences the emergence of in-and-out group dynamics which in turn mediates the formation of all subsequent groups (Festinger, Schachter, & Back, 1950).

Agent-based simulations have been used for social network studies for many years now. Carley and Newell (1994) were, to our knowledge, the first to use a cognitive architecture (Plural Soar) to study organizations. More recently, some studies have applied cognitive architectures to model human decision making in collaborative tasks (Lebiere, Gonzalez, Dutt, & Warwick, 2009; Morgan, Morgan, & Ritter, 2010; Prietula & Carley, 2001). These authors have, however, primarily focused on small group collaborations and interactions with less than 20 agents.

In this paper, we use a cognitive architecture based socio-cognitive simulation to examine the effect that memory activation thresholds, navigation strategies, and map-configurations have on the rate of network formation. Examining different memory activation thresholds for links between agents enables us to model not only the effects of memory retention on network formation but also provides us a means of representing differences in the modeled social ties' *quality*, as Dunbar defines this term (Dunbar, 1998, pp. 76-77). In other words, higher quality relationships are associated with greater cognitive investment and higher memory strength. Comparing navigation strategies and map configurations allows us to represent the social opportunities associated with activity spaces in the environment.

This study draws from previous work (Kaulakis et al., 2012) and (Zhao, Kaulakis, Morgan, Hiam, & Ritter, 2012). In Kaulakis et al.'s, we introduced an earlier version of an ecological model and modeling environment (VIPER). Kaulakis et al.'s presents a structural analysis, examining how the agents' declarative representation of their social ties reliably differed from the experiment's ground truth network, or the network formed from all the agents' room co-occurrences. Kaulakis et al. found population size had the greatest influence on network construction in memory, but that the similarity results were tentative. Zhao et al. (2012) elaborated on the model by adding navigation strategies. Zhao et al.'s primary contribution, however, showed that parameters in the simulation, world size, length of interactions, and navigation strategies, led to changes in the agents' average activation values in their social networks. While promising, these studies provided no

insight as to the rate of network formation and did not examine Dunbar's number in detail.

To reconcile Dunbar's (1998) and McCarty's et al.'s (2001) estimates, we need to understand time not only as defining all the possible social opportunities available to the network but also how previous tie formation constrains future choices. To do this, we first need some notion of a simulated network's formation baseline, when in other words does the network reach equilibrium and its members are primarily maintaining in memory as opposed to making ties? We examine this question here.

## Experiment Environment

To model multi-agent social behavior, we constructed a simulation environment, VIPER. All of our experiments were conducted on a 2GHz eight-core Linux 2.6.31 machine with Ubuntu 11.04 with 8GB of RAM, with SBCL 1.0.52 as our Lisp. We use ACT-R 6 in Anderson et al. (2004).

### ACT-R

ACT-R (Anderson et al., 2004) is a cognitive architecture and unified theory of cognition. It tries to provide a fully functional system that produces all aspects of human behavior at the cognitive level. We use ACT-R because its memory mechanisms enable us to fully implement the cognitive capacities and constraints we believe necessary to model the emergence of networks.

### VIPER

VIPER, a text-based multi-agent simulation, models physically embodied social networks (Kaulakis et al., 2012). It is designed to support multi-agent simulations used to study network science. It is lightweight in that it is text based, but is extensible and records agent behaviors over time to support studies on how networks form. VIPER represents these constraints in several ways, the chief being a strong separation between the agents and their environment. VIPER is dynamic, agent-based, and designed to be a part of a distributed model that resolves events in either real or accelerated time.

To handle large amount of agents simulation, we utilized file imaging techniques in Linux system to reduce the memory cost of ACT-R. This reduces the cost of a single ACT-R thread from 50 Mb to less than 20MB, which allows us to run 1,000 agents on one machine.

## Experiment

To explore the effects of environmental connectivity, navigation strategy, and memory activation thresholds on the pace of network formation, we ran a simulated study that examined each of these three factors.

### Map Configuration

Drawing from work in environmental psychology and crime mapping, we know environmental complexity influences network formation; we represent environmental complexity

with three room configurations. We measure the relative connectivity of our three map configurations by defining its *grid ratio*, the ratio of the number of edges over the total number of edges possible for a rectangular grid containing the same number of rooms.

We used three map configurations, shown in Figure 1. The first configuration (Figure 1a) is a two-hallway configuration with *grid ratio* 0.6. This configuration should lead to low connectivity due to the large distances between the agents. The second configuration (Figure 1b) has a central area with *grid ratio* 0.75. We predict that Figure 1b's central meeting point will lead to network connectivity that is less than that found in Figure 1c but greater than that found in Figure 1a. The last configuration (Figure 1c) is a full 5x5 grid with *grid ratio* 1.0.

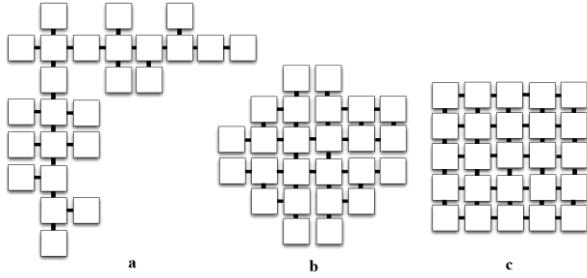


Figure 1: Maps (hallway, central, full grid) used in the simulation study.

### Navigation Strategies

In a social network the agents' movement patterns will influence the social network's topology by again influencing any one's agent's interaction opportunities. For example, a policeman walking beat will likely have a larger number of acquaintances than a person who spends most of his or her time at home because the policeman has more opportunities to meet people. To replicate human navigation behavior, we implemented two navigation strategies: random-walk and fixed-path.

- 1) *The Random-walk strategy* implements a random walk.
- 2) *The Fixed-path strategy* follows a specific path to navigation in a small area. This strategy simulates the routine navigation behavior, such as going work or going to school.

### Experiment Parameters

Zhao et al. (2012) found that the map configurations and navigation strategies influence network measures. In this experiment, we will examine the two navigation strategies for each of the map configurations in 4 runs. The total agent size is currently an arbitrary choice, 40; forty provides a populated but not crowded environment to study. The parameters of the 4 runs are shown in Table 1.

Table 1: Setting of experiment parameters

Runs	Agent size	Map configuration	Navigation strategy
Run 1	40	Hallway	Fixed
Run 2	40	Hallway	Random
Run 3	40	Central	Random
Run 4	40	Grid	Random

To examine the growth curve of the network, we captured the network growth over 18 time slices between 10 and 500 s. Those sample running times were selected by running a pilot experiment, from which we found that the curve changes significantly from 80 to 150 s. We present more sample times here, resulting in a more interesting and precise curve.

## Results and Analysis

We examine the effect of the three parameters (map connectivity, navigation strategy, and memory threshold) in order. Each run took approximately 500 seconds in real-time, with the analysis logs being analyzed by hand using ORA (Carley, Reminga, Storrick, & Columbus, 2011).

### Memory Networks

With 4 runs and 18 sampling times, we created 2,880 egocentric memory networks (one for each agent, noting who that agent thought they knew, as shown in Figure 2a), and 72 merged memory networks across a run of 40 agents (merging memories across agents in a run, as shown in Figure 2b). Both networks in Figure 1 consist of agents where no memory threshold was applied.

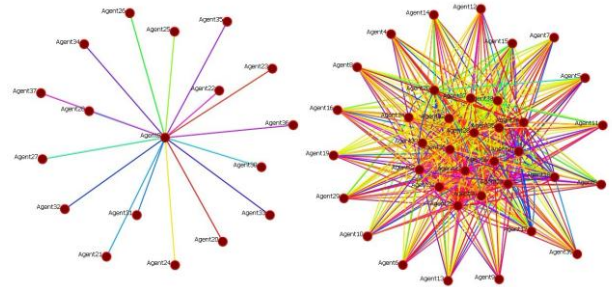


Figure 2: Example egocentric network (left) and merged memory network (right) for agents without a memory threshold applied.

### Curves of the Growth of Merged Networks

In this section, we show the effects of the model's three parameters on the rate of tie formation. Based on these figures, we will discuss how memory thresholds, map configurations, and navigation strategies influence the formation rates of simulated networks.

Figure 3 shows the growth curve of a network consisting of agents using a fixed-path navigation strategy in the Hallway map. The lower line represents the network formation rate of a network where no memory threshold was applied—if an agent met an agent, they formed a permanent tie. We find that the lower curve increase rapidly and then

flattens when it reaches 1,336 ties (the maximum is  $40 \times 39$ , or 1,560, if the agents' paths completely overlap, which they do not). This flattening occurs once the network has achieved equilibrium and is fully connected.

In Figure 3, the top solid line represents the network formation rate of a network where an activation threshold of 0.0 was applied. According to the ACT-R theory, the activation threshold represents a memory limitation, meaning that memory chunks with an activation value lower than the threshold cannot be retrieved. The top curve's more gradual progression illustrates the influence of memory on the formation rate, multiple exposures are required to remember another agent. While the difference in total number of links (800 versus 1,336) illustrates memory's effect on the network's topology. In addition, this network never achieves a fully connected state, in the sense that the agent's declarative representation at no point includes the total set of possible interactions. In other words, these agents must continue to maintain their relationships because they continue to forget. Nevertheless, this networks does eventually acheives equilibrium at 150 seconds with a network size of 800 links.

Comparing the two solid curves in the Figure 3, we noticed another difference, the time at which the rate of growth begins to increase. For the thresholded network, this time happens later than for the un-thresholded network. This is because the agents tie formation requires multiple exposures. Initially, agents are busy simply encountering other agents and building their friends list. As they, however, begin to meet more "old friends", the activation values of friendships start to increase. The dash curve in Figure 3 shows the number of relations that could not be retrieved. The curve grows fast at the beginning because most of new ties are weak and un-retrievable. It decrease after 200seconds as the network acheives equilibrium.

The x-axis of the Figure 3 represents the simulation running time in real seconds. In our experiment, we set the traval interval between rooms at 16 seconds to make the effect of memory decay more prominent. Nevertheless, this interval is still not long enough to be realistic because people might take minutes or hours to find another person. As this work only focuses on the growth pattern of the social network, we would argue that the measurement of time is a secondary factor of our study because over 80 percent of the decay happens in the first 16 seconds acording to the ACT-R decay equation, with little additional decay occouring at greater time scales. Consequently, we believe total running time of 500 seconds and a short travel interval of 16 seconds are acceptable for simulating the growth pattern.

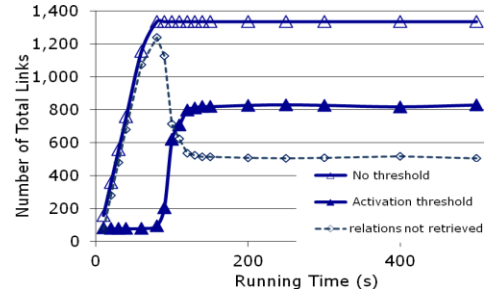


Figure 3: The effect of memory threshold on network formation over time for the fixed path navigation strategy in the hallway map.

Figure 4 shows the growth curve of a network of agents using the random navigation strategy in the Hallway map. Comparing Figure 4 with Figure 3, the non-threshold curves have the same growth pattern, but the threshold curves appear to be different. Memory appears to have different effects based on the setting in which the agents operate.

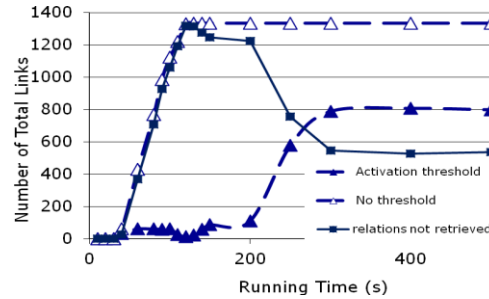


Figure 4: The effect of memory threshold on network formation over time for the random walk strategy in the hallway map.

Figure 5 compares the growth curves of two networks where a memory retrieval threshold of 0.0 was applied; these networks differ with respect to the navigation strategy used by their members. The fixed path-strategy (dash line) forms ties more quickly than the random-path strategy. We suspect that the fixed-path strategy achieves equilibrium sooner because it is more localized, and thus provides more chances for agents to meet their "old friends". On the other hand, both networks achieve equilibrium at about 800 links, suggesting that the navigation strateies in this simulation do not constrain the number of relations an agent can maintain in memory.

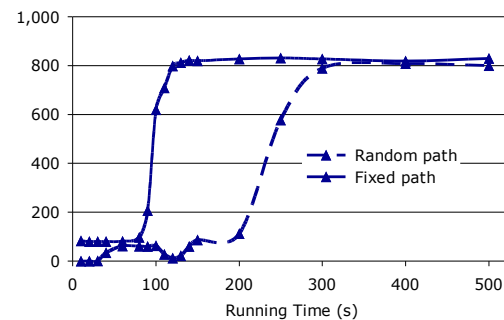


Figure 5: The effect of navigation strategy on network formation over time in the hallway map with threshold.



Figure 6 compares the network formation rates of networks occurring in each of the three map configurations (full grid, central, and hallway); all these networks consist of agents with a memory activation threshold of 0.0. We find that the map configurations have a similar influence on the networks' growth curves as the navigation strategies. Again, the map configurations influence the rate of formation but not the network's size at equilibrium.

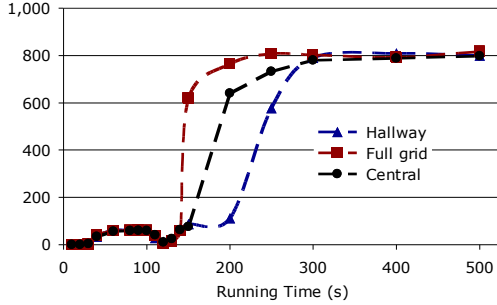


Figure 6: The effect of map configuration on network formation over time.

Comparing the three curves, we find the Hallway map (*grid ratio*=60%) is associated with the longest delay in network formation and the lowest rate of increase; the full grid map (*grid ratio*=100%) has the shortest delay and the fastest rate of link formation. These results show that delay in the network's growth rate is negatively correlated with *grid ratio*, while the network's growth rate during its growth spurt is positively correlated.

### Activation Normalization: Semantic Challenges

One of the main issues we face in the analysis and interpretation of our results is the need to assign semantic meanings to the activation values associated with our ACT-R agents' memory chunks. Because raw activation values may grow or shrink indefinitely, we see normalization as a process by which the data can be made more regular, and to help scale between time scales used in our simulation and those occurring in the real world.

In the introduction to this paper, we cited Dunbar's (1998) concerns about tie "quality". In this work, we used raw activation values in our measures, which is fine for our purposes, but is insufficient for many other questions. Activations are not portable or easily interpretable in social terms. To make sensible translations between activation levels and Dunbar's notion of tie quality, we suggest that the ties be normalized as we describe in this section. This normalization recasts activations as statements about the "probability of recall" within a particular timeframe. Using activations in this way supports the measurement of environmental parameters, and the prediction of environmental distractions that are likely to prevent tie consolidation by limiting the time available for tie maintenance. This grounding provides metrics that are empirically measurable and come closer to Dunbar's "quality" concept. Derived from the ACT-R Probability of Recall Equations (Anderson et al., 2004), where the

normalized activation value is a function of three variables internal to the agent, such that  $i$  is the current chunk,  $\tau$  is the threshold for recall, and  $s$  is noise, then the normalized value is a probability that a particular chunk will be recalled:

$$A_P(i, \tau, s) = \frac{1}{1 + e^{\frac{-(A(i) - \tau)}{s}}} \quad (\text{eq. 1})$$

This method fulfills all of the requirements above and provides a concrete interpretation of activation levels as Probabilities of Recall. Additionally, it also ties the threshold to the time of recall in seconds, like this:

$$T_i = F e^{-A_i} \quad (\text{eq. 2})$$

These properties will make the analysis of normalized activation values able to generate much stronger statements about the settings in which the agents live.

### Conclusion and Discussion

This study simulated an effect like Dunbar's number on networks of cognitive architecture-based agents. The first analysis examined to what degree **cognitive limitations** (represented by a memory activation threshold) influenced the generative process of a network. *The results suggest that cognitive limitations influence both the rate of network formation and the size of the network at equilibrium.* These findings roughly mirror what is found in empirical studies (Brantingham & Brantingham, 1993).

We can view the progression of the curves in Figures 3-6 as corresponding to three stages in network formation, though at abbreviated time scales. Between 0 and 100 seconds, the size of the network does not grow significantly, and the average number of relations stays constant at 60. This represents the tendency of people to initially remain in localized relations with a few people. During this period of the simulation, the non-thresholded network grows very fast (becoming fully connected at 100 seconds) because most the agents wander around to meet new friends and initialize new relationships. For the thresholded networks, there is greater period of latency because the agents have not yet had the time to consolidate their friendships, but rather are primarily building their friends list. Between approximately 100 to 150 seconds, we see that thresholded networks begin to rapidly increase in size as the agents become more familiar with the activity space. Finally, between 150 to 500 s, the thresholded networks stop growing because the agents have shifted from primarily establishing to maintaining their friends network. In the stable state, the number of total links remains around 800, meaning that the average number of relations per agent in these networks is about 20.

In the second analysis, we examined the influences of two **navigation strategies**. The results suggest that navigation strategies have little influence on the non-thresholded network but for growth time, and it does change the growth speed and pattern of the thresholded network. In Figure 5, we see that the network using the fixed-path strategy grows much faster. This is because the fixed-path strategy is a more focused strategy that provides more chances for



people to meet their “old friends”. In this case, people more easily form small groups associated with their starting location, such as people living on the same street or attending the same school. *We see that the fixed-path strategy facilitates the rapid creation of smaller tighter groups than the random-walk strategy.*

The third analysis focused on examining the influences of **spatial configurations** on generative networks. We defined *grid ratio* as the ratio of the number of edges over the total number of possible edges to quantify the connectivity of the map configurations. We find that delay increment is negatively correlated with *grid ratio*, while the formation rate during the growth phase is positively correlated with the *grid ratio*. This result validates our definition of *grid ratio*, because it shows the *grid ratio* does have influence on network formation; it also proves that lower *grid ratio* maps with more gaps and obstacles decrease the network’s growth rate. *We found, however, that our map configurations did not influence the final size of the network.*

Summarizing our results, we see that navigation strategies and room configurations only seem to significantly influence our networks’ delay increment and growth rate, while the final size of our thresholded networks remains at 800 links. We suspect that one possible way to adjust the final size of the network is by changing the cognitive parameters in ACT-R, for instance adjusting memory decay speed or base level learning activation. Moreover, our results also imply, at least for our world, that navigation strategies and environmental complexity do not significantly influence the number of friends that a person can maintain in memory (Dunbar’s number), as the average number of relations were same for both networks. They do, however, suggest the ecological factors significantly contribute to the degree of localization, and perhaps in a more complex world the total size and evolution of the network as defined by the number of total environmental possibilities.

## Future Work

Future work will build upon both our insights regarding the effect of cognitive resources on network topology, as well as rate of growth. First, we would like to extend our analysis of the normalized thresholds to see if there are regularities in their effects on network topology. Second, we will run more agents, because our test systems were kept deliberately small. Finally, we will extend our analysis on the effects of cognition on network measures analogous to Dunbar’s Number, such as information and knowledge diffusion.

## Acknowledgments

DTRA (HDTRA1-09-1-0054) supported this work. Jaeyhon Paik and Joseph Sanford provided technical support, while Geoffrey P. Morgan provided useful comments.

## References

- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036-1060.
- Brantingham, P. L., & Brantingham, P. J. (1993). Nodes, paths and edges: Considerations on the complexity of crime and the physical environment. *Journal of Environmental Psychology*, 13(1), 3-18.
- Carley, K. M., & Newell, A. (1994). The nature of the social agent. *Journal of Mathematical Sociology*, 19(4), 221-262.
- Carley, K., Reminga, J., Storrick, J., & Columbus, D. (2011). ORA User's Guide, Carnegie Mellon University.
- Dunbar, R. I. M. (1998). *Grooming, gossip, and the evolution of language*. Cambridge, MA: Harvard .
- Festinger, L., Schachter, S., & Back, K. (1950). The Spatial Ecology of Group Formation. In S. S. Festinger, & K. Back (Eds.), *Social Pressure in Informal Groups*.
- Kaulakis, R., Zhao, C., Morgan, J. H., Hiam, J. W., Sanford, J. P., & Ritter, F. E. Defining factors of interest for large-scale socio-cognitive simulations. In the proceedings of ICCM 2012, Germany.
- Lebiere, C., Gonzalez, C., Dutt, V., & Warwick, W. (2009). Predicting cognitive performance in open-ended dynamic tasks: A modeling comparison challenge. In *Proceedings of the 9th International Conference on Cognitive Modeling*, Manchester, UK.
- McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E. C., & Shelley, G. A. (2001). Comparing two methods for estimating network size. *Human Organization*, 60(1), 28-39.
- Metz, A., & Shultz, T. R. (2010). Spatial factors in social and asocial learning. Paper presented at the The Annual Meeting of The Cognitive Science Society, Portland. OR.
- Milgram, S. (1967). The small world problem. *Psychology Today*, 1(1), 61-67.
- Morgan, J. H., Morgan, G. P., & Ritter, F. E. (2010). A preliminary model of participation for small groups. *Computational & Mathematical Organization Theory*, 16(3), 246-270.
- Prietula, M. J., & Carley, K. M. (2001). Boundedly rational and emotional agent cooperation, trust, and rumor. In T. C. Castelfranchi, Y. H. (Eds.), *Trust and Deception in Virtual Societies*. Norwell, MA: Kluwer.
- Simon, H. A. (1991). *Models of bounded rationality: Economic analysis and public policy*. MIT Press.
- Zhao, C., Kaulakis, R., Morgan, J. H., Hiam, J. W., & Ritter, F. E. Socio-cognitive networks: modeling the effects of space and memory on generative social structures. In the Proceedings of *BRIMS 2012*, Amelia Island, FL.

# Attention Modeling for Face Recognition via Deep Learning

**Sheng-hua Zhong (csshzhong@comp.polyu.edu.hk)**

Department of Computing, Hung Hom, Kowloon  
Hong Kong, 999077 CHINA

**Yan Liu (csyliu@comp.polyu.edu.hk)**

Department of Computing, Hung Hom, Kowloon  
Hong Kong, 99907 CHINA

**Yao Zhang (csyaozhang@comp.polyu.edu.hk)**

Department of Computing, Hung Hom, Kowloon  
Hong Kong, 99907 CHINA

**Fu-lai Chung (cskchung@comp.polyu.edu.hk)**

Department of Computing, Hung Hom, Kowloon  
Hong Kong, 99907 CHINA

## Abstract

Face recognition is an important area of research in cognitive science and machine learning. This is the first paper utilizing deep learning techniques to model human's attention for face recognition. In our attention model based on bilinear deep belief network (DBDN), the discriminant information is maximized in a frame of simulating the human visual cortex and human's perception. Comparative experiments demonstrate that from recognition accuracy our deep learning model outperforms both representative benchmark models and existing bio-inspired models. Furthermore, our model is able to automatically abstract and emphasize the important facial features and patterns which are consistent with the human's attention map.

**Keywords:** face recognition; attention model; deep learning.

## Introduction

Face recognition plays an important role in the social life and attracts interest from a very broad range of researchers and scientists (Anderson, 1998). In machine learning and computer vision areas, face recognition using computational models is a classical problem. The representative models include: Eigenface (Turk, et al., 1991), Fisherfaces (Belhumer, et al., 1997), support vector machine (SVM) (Müller, et al., 2001), and so on.

In cognitive science, face recognition is a vividly researched area (Gauthier, et al., 2000) (Afraz, et al., 2006) (Civile, et al. 2011). It is argued that face perception is involved in a unique cognitive process compared with non-face object or scene perception. Researchers in cognitive science seek to understand how the visual system transforms a face image from an initial, pixel-like representation, to a new powerful form of representation, and finally induce the selectively response of the neurons in inferior temporal cortex (Afraz, et al., 2006). Hence, the researchers have utilized some signal processing techniques to simulate the response of human visual system, such as human's attention allocation. Computational attention model is utilized to measure of the conspicuity and provide the predictions about which regions are likely to attract observers' attention (Koch et al., 1985) (Parkhurst et al., 2002). Many empirical

validations have demonstrated that attention models have notable ability in various tasks, such as content aware resizing (Avidan et al., 2007), quality assessment (Zhong et al., 2010), and face recognition (Cappelli, et al., 2007) (Fang, et al., 2011).

This paper models human's attention for face recognition via deep learning technique. Deep learning models the learning tasks using deep architectures composed of multiple layers of parameterized nonlinear modules. Deep model is selected in this paper because of two considerations. First, the multiple layers deep architecture is consistent with the laminar structure of human's brain cortex and the information delivery in deep model simulates human's visual cortex. Second, deep learning has demonstrated distinguished ability of information abstraction and robust performance of data classification in various visual data analysis tasks (Hinton, et al., 2006).

This is the first paper utilizing deep learning techniques to model human's attention for face recognition. Compared with existing face recognition models, our proposed bilinear deep belief network (BDBN) has several attractive characters:

- 1) BDBN maximizes the discriminant information in a frame of simulating the human visual cortex and human's perception. As we known, nearly all existing machine learning models aims to find the discriminant solution to face recognition applications. Existing computational cognitive model emphasizes the identity between the model and the human visual system. Our model attempts to integrate the advantages of both techniques and provide a new thought of this problem.
- 2) Compared with existing computational face recognition models or representative attention models, BDBN has the ability to automatically extract and emphasize the important facial features and patterns which are consistent with the human attention map.
- 3) BDBN includes three learning stages: semiconducting bilinear discriminant initialization, greedy layer-wise reconstruction, and global fine-tuning. The rational of

three-stage learning comes from the phenomenon of two peaks activation in visual cortex areas. With regard to object recognition, the early peak is related to the activation of an “initial guess” based on the acquired discriminative knowledge, while the late peak reflects the post-recognition activation of conceptual knowledge related to the recognized object.

### Model

In this section, we design a deep learning algorithm with a deep architecture for the task of face recognition, includes bilinear discriminant initialization, greedy layer-wise reconstruction and global fine-tuning. The strategy of bilinear discriminant projection is utilized to construct a projection to map the original data into a discriminant preserving subspace. And it determines the initial parameters and sizes of the upper layer. To human, this strategy is consistent with the early peak related to the activation of “initial guess”. In the stage of greedy layer-wise reconstruction, the parameter space is refined by the greedy layer-wise information reconstruction using Restricted Boltzmann Machines (RBMs) (Smolensky, 1986) as building blocks. In the stage of global fine-tuning, we refine the parameter space for better face recognition performance. And it is consistent with the late peak related to the activation of “post-recognition”. After the deep learning model is constructed, the attention map is built based on the parameter space in the first RBM.

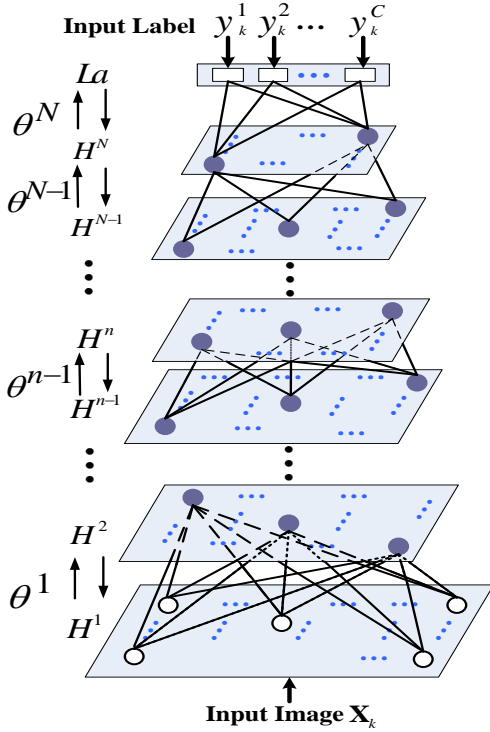


Figure 1: Architecture of the bilinear deep belief network.

Figure 1 shows the architecture of our bilinear deep belief network. A fully interconnected directed belief network

includes input layer  $H^1$ , hidden layer  $H^2, \dots, H^N$ , and one label layer  $La$  at the top. The input layer  $H^1$  has  $I \times J$  units, and this size is equal to the dimension of the input features. In our model, we use the pixel values of sample datum  $\mathbf{X}_k$  as the original input features. In the top, the label layer has  $C$  units, which is equal to the number of classes. The search of the mapping function from  $X$  to  $Y$  is transformed to the problem of finding the optimum parameter space  $\theta^*$  for the deep architecture.

In our deep learning architecture,  $X$  is a set of data samples,  $X = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k, \dots, \mathbf{X}_K]$ .  $\mathbf{X}_k$  is a sample datum in the image space  $\mathbb{R}^{I \times J}$  and  $K$  is the number of sample data.  $Y$  is a set of labels corresponding to  $X$ ,  $Y = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k, \dots, \mathbf{y}_K]$ .  $\mathbf{y}_k$  is the label vector of  $\mathbf{X}_k$  in  $\mathbb{R}^C$ ,  $y_k^c = \begin{cases} 1 & \text{if } \mathbf{X}_k \in c\text{th class} \\ 0 & \text{if } \mathbf{X}_k \notin c\text{th class} \end{cases}$ , where  $C$  is the number of classes.

Based on the given training set, the aim in face recognition is to learn a mapping function from the image set  $X$  to the label set  $Y$ , and then recognize the new coming face images according to the learned mapping function.

### Bilinear Discriminant Initialization

In order to preserve the discriminant information in the learning procedure, the objective function of bilinear discriminant initialization could be represented as follows:

$$\arg \max_{\mathbf{U}, \mathbf{V}} J(\mathbf{U}, \mathbf{V}) = \sum_{s,t=1}^K \|\mathbf{U}^T (\mathbf{X}_s - \mathbf{X}_t) \mathbf{V}\|^2 (\alpha \mathbf{B}_{st} - (1-\alpha) \mathbf{W}_{st}) \quad (1)$$

$$s.t. \mathbf{U}^T \mathbf{U} = \mathbf{I}_p, \mathbf{V}^T \mathbf{V} = \mathbf{I}_q$$

where balance weight  $\alpha \in [0,1]$  is the parameter used to balance the between-class weights  $\mathbf{B}_{st}$  and the within class weights  $\mathbf{W}_{st}$ , which are defined as follows (Yan, et al., 2007) (Sugiyama, 2007).

By simultaneously maximizing the distances between data points from different classes and minimizing the distances between data points from the same class, the discriminant information is preserved to the greatest extent in the projected feature space. Solving  $\mathbf{U}$  (or  $\mathbf{V}$ ) with fixed  $\mathbf{V}$  (or  $\mathbf{U}$ ) is a convex optimization problem. Let  $\mathbf{E}_{st} = \alpha \mathbf{B}_{st} - (1-\alpha) \mathbf{W}_{st}$ , with the fixed  $\mathbf{V}$ . The optimal  $\mathbf{U}$  is composed of the first  $P$  eigenvectors of the following eigendecomposition problem:

$$\mathbf{D}_v \mathbf{u} = \lambda \mathbf{u} \quad (2)$$

where  $\mathbf{D}_v = \sum_{st} \mathbf{E}_{st} (\mathbf{X}_s - \mathbf{X}_t) \mathbf{V} \mathbf{V}^T (\mathbf{X}_s - \mathbf{X}_t)^T$ . Similarly, with the fixed  $\mathbf{U}$ , the optimal  $\mathbf{V}$  is composed of the first  $Q$  eigenvectors of the following eigendecomposition problem:

$$\mathbf{D}_u \mathbf{v} = \lambda \mathbf{v} \quad (3)$$

where  $\mathbf{D}_u = \sum_{st} \mathbf{E}_{st} (\mathbf{X}_s - \mathbf{X}_t)^T \mathbf{U} \mathbf{U}^T (\mathbf{X}_s - \mathbf{X}_t)$ .

The above steps monotonically increase  $J(\mathbf{U}, \mathbf{V})$  and since the function is upper bounded, it will converge to a critical point with transformation matrices  $\mathbf{U}, \mathbf{V}$ .

By bilinear discriminant initialization, we obtain the discriminant initial connections in layer pair and utilize the optimal dimension to define the structure of the next layer.

$$A_{ij,pq}^n(0) = (\mathbf{U}_{ip}^n)^T \mathbf{V}_{jq}^n \quad (4)$$

$$P^{n+1} = \text{row}(\mathbf{U}^n), \quad Q^{n+1} = \text{column}(\mathbf{V}^n) \quad (5)$$

### Greedy Layer-Wise Reconstruction

In this section, we describe how to construct the first RBM between the input layer  $H^1$  and the first hidden layer  $H^2$ .

The energy of the state  $(\mathbf{h}^1, \mathbf{h}^2)$  in the first RBM is:

$$E(\mathbf{h}^1, \mathbf{h}^2; \theta^1) = -(\mathbf{h}^1 \mathbf{A}^1 \mathbf{h}^2 + \mathbf{b}^1 \mathbf{h}^1 + \mathbf{c}^1 \mathbf{h}^2) \quad (6)$$

where  $\theta^1 = (\mathbf{A}^1, \mathbf{b}^1, \mathbf{c}^1)$  are the model parameters between the input layer  $H^1$  and first hidden layer  $H^2$ . Therefore, the log-likelihood probability of the model assigned to  $\mathbf{h}^1$  in  $H^1$  is:

$$\log P(\mathbf{h}^1) = \log \sum_{\mathbf{h}^2} e^{-E(\mathbf{h}^1, \mathbf{h}^2; \theta^1)} - \log \sum_{\mathbf{h}^1} \sum_{\mathbf{h}^2} e^{-E(\mathbf{h}^1, \mathbf{h}^2; \theta^1)} \quad (7)$$

By calculating the derivative of Equation (8), we could update the parameter space with respect to the parameter  $\theta^1 = (\mathbf{A}^1, \mathbf{b}^1, \mathbf{c}^1)$ .

$$\begin{aligned} \frac{\partial \log p(\mathbf{h}^1(0))}{\partial \theta^1} &= - \sum_{\mathbf{h}^2(0)} p(\mathbf{h}^2(0) | \mathbf{h}^1(0)) \frac{\partial E(\mathbf{h}^2(0), \mathbf{h}^1(0))}{\partial \theta^1} + \\ &\sum_{\mathbf{h}^2(t)} \sum_{\mathbf{h}^1(t)} p(\mathbf{h}^2(t), \mathbf{h}^1(t)) \frac{\partial E(\mathbf{h}^2(t), \mathbf{h}^1(t))}{\partial \theta^1} \end{aligned} \quad (8)$$

The above discussion is the greedy layer-wise abstraction for the first layer  $H^1$  with its next adjacent layer  $H^2$ . Similar operations can be performed on the higher layer pairs.

### Global Fine-Tuning

In this section, we use backpropagation to adjust the entire deep network to find good local optimum parameters  $\theta = [\mathbf{A}, \mathbf{b}, \mathbf{c}]$  by minimizing the recognition error  $[-\sum_i \mathbf{y}_i \log \hat{\mathbf{y}}_i]$ , where  $\mathbf{y}_i$  and  $\hat{\mathbf{y}}_i$  are the correct recognition label and the output recognition label value of labeled sample datum  $\mathbf{X}_i$  in  $X^L$ .

Above, we utilize the greedy layer-by-layer algorithm to learn a deep model with the help of discriminant information obtained from bilinear discriminant projection. Therefore, the convergence in our algorithm obtained from backpropagation is not slow. And the result generally converges to a good local minimum on the error surface.

### Attention Modeling

The weights of first layer of BDBN are oriented, Gabor-like and resemble the receptive fields of V1 simple cell (Zhong, et al., 2011). Therefore, the first RBM is utilized to construct the attention model which is shown in Figure 2.

To every neuron in the input layer, the weight value to the one in the first hidden layer is calculated as feature map. Then, the weight value of every neuron is normalized and combined into an attention map.

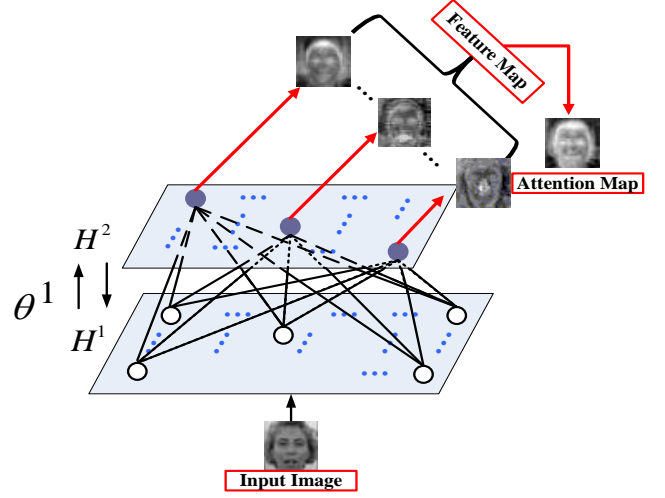


Figure 2: Construct attention model by first RBM in bilinear deep belief network.

## Experiment 1: Recognition Accuracy Analysis

### Dataset

The CMU PIE face dataset collected between October and December 2000 contains 68 subjects with a total of 41,368 face images (Sim, et al., 2002). The face images were captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination and expression. In the first experiment, we use all the images under different illuminations and expressions with five near frontal poses (C05, C07, C09, C27, C29). Thus we obtain 170 images for each individual.

### Procedure

For the CMU PIE face dataset, the preprocessing is applied following the general setting of experiment (He, et al., 2005). Original images are normalized (in scale and orientation) so that the two eyes are aligned at the same position. Then, the facial areas are cropped into the final images for matching. The size of each cropped image in all of the experiments is  $32 \times 32$  pixels. Sample images after preprocessing are shown in Figure 3.

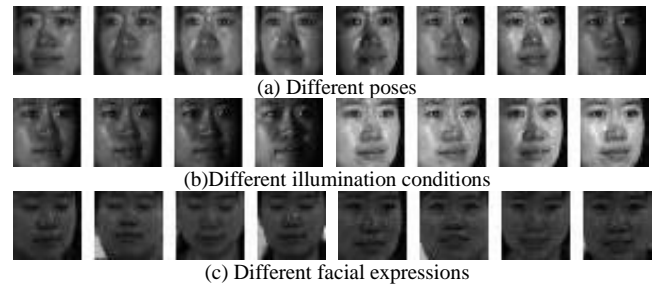


Figure 3: Sample images after preprocessing from CMU PIE.

In our experiments, the balance weight of our model is set as 0.5 for simplicity. For parameters such as the learning rate and the momentum, we simply follow the general setting of previous work on computational deep networks (Bengio, et al., 2006), although more careful choice may lead to better performance. In the fine-tuning stage, the method of conjugate gradients is utilized and three line searches are performed in each epoch until convergence.

To adapt the real-world face recognition tasks, our computational neuroscience model BDBN is applied under a semi-supervised learning framework. It makes face recognition work well when labeled images are insufficient. 120 images are randomly selected for each person to form the training set and the rest to form the test set. Of the 120 images for each person, different numbers of images are randomly selected and labeled while the others remain unlabeled. The number of labeled data per subject is equal to 5, 10, 20 and 40, respectively. We perform 10 random splits and report the average results over the 10 trials.

## Experimental Results

In the first experiment, we compare three representative face recognition models, including: Eigenface (Turk, et al., 1991), Fisherfaces (Belhumer, et al., 1997), SVM (Müller, et al., 2001), and existing bio-inspired Sparse Localized Features (SLF) model (Mutch and Lowe, 2008).

SVM, Eigenface, and Fisherfaces both are representative benchmark machine learning models for the task of face recognition. The bio-inspired Sparse Localized Features (SLF) (Mutch and Lowe, 2008) are an extensions of the C2 features from the Serre et al. HMAX model (Serre, et al., 2007). For this representation, we took advantage of the MATLAB code provided by the authors. Here, the SVM classification was based on a linear kernel with normalized training and testing data (zero-mean and unit-variance feature-wise).

The recognition accuracy rate with different numbers of labeled data is shown in Table 1. As shown in Table 1, the recognition accuracy rate of bio-inspired models SLF+SVM and BDBN is better than machine learning models Eigenfaces and SVM. And our proposed BDBN has the best performance than others.

Table 1: Recognition accuracy rate (%) on the test data with different numbers of labeled data per category on CMU PIE.

Num./Cat.	20	30	40	50
Eigenfaces	61.9 $\pm$ 0.7	72.1 $\pm$ 0.6	78.2 $\pm$ 0.5	83.8 $\pm$ 0.4
Fisherfaces	84.5 $\pm$ 0.7	92.0 $\pm$ 0.6	93.1 $\pm$ 0.5	94.8 $\pm$ 0.3
SVM	73.5 $\pm$ 0.6	80.4 $\pm$ 0.5	82.9 $\pm$ 0.5	87.1 $\pm$ 0.3
SLF+SVM	80.5 $\pm$ 0.6	86.8 $\pm$ 0.5	89.5 $\pm$ 0.5	90.2 $\pm$ 0.3
Semi_DBN	85.4 $\pm$ 0.7	92.4 $\pm$ 0.6	93.5 $\pm$ 0.5	95.0 $\pm$ 0.3
BDBN	88.4 $\pm$ 0.7	93.9 $\pm$ 0.6	94.3 $\pm$ 0.5	96.6 $\pm$ 0.3

## Experiment 2: Face Feature Points Emphasis

### Dataset

The BioID face dataset consists of 1521 gray level images collected contains 23 subjects (HumanScan, 2003). The face images in BioID are under a large variety of illumination, background.

In this dataset, the x and the y coordinate of the left eye and the right eye are provided. Furthermore, the 20 important facial feature points are manually placed, including: right eye pupil, left eye pupil, right mouth corner, left mouth corner, outer end of right eye brow, inner end of right eye brow, inner end of left eye brow, outer end of left eye brow, right temple, outer corner of right eye, inner corner of right eye, inner corner of left eye, outer corner of left eye, left temple, tip of nose, right nostril, left nostril, centre point on outer edge of upper lip, centre point on outer edge of lower lip, and tip of chin. These facial feature points are thought to be very useful for facial analysis and gesture recognition (Jesorsky, et al., 2001) (Wang, et al., 2002) (Cappelli, et al., 2007).

### Procedure

As a deep learning model for face recognition, BDBN has demonstrated the impressive recognition performance in this first experiment. In this experiment, we intend to investigate the consistency between the emphasized regions in BDBN and the attention map of human being.

The number of images in every category of BioID is varied, from 35 to 118. Therefore, firstly, we choose the categories with more than 50 face images as the subset we work on. Then, just like the procedure on face datasets, the original images are normalized (in scale and orientation) so that the two eyes are aligned at the same position. Finally, the facial areas are cropped and downsampled into the final images. The size of each final image in all of the experiments is 32 $\times$ 32 pixels, with 256 gray levels per pixel. Some sample images after preprocessing are shown in Figure 4.



Figure 4: Sample images after preprocessing from BioID.

Then, to every image in BioID face dataset, we directly input the original pixel value to the BDBN model. After

bilinear discriminant initialization and layer wise reconstruction, we evaluate the consistency between the constructed attention model and human's attention map.

## Experimental Results

Computational attention model was called saliency map first appeared in (Koch, et al., 1985). Typically, multiple low-level visual features such as intensity, color, orientation, texture and motion are extracted at multiple scales. After a feature map is computed for each of the features, they are normalized and combined into a master saliency map that represents the saliency of each pixel.

To face images, some facial areas are assessed to be attracted more attention and helpful to face recognition, for example eye, ear, nose and mouth (Hickman, et al., 2010). Fortunately, in this dataset, 20 important facial feature points are manually selected out and placed. Therefore, with marked facial feature points, the attention model based on deep learning model could be evaluated without eye tracking recordings.

Different from representative attention map which utilizes various features such as intensity, color, orientation, only the gray level pixel values are input into our model. Our attention model automatically extracts and emphasizes important features and patterns to construct facial attention model.

To demonstrate the effectiveness of our model, firstly, the visualization of the parameter space of proposed model is observed. Figure 5 (a) shows a sample image, and Figure 5 (b) shows the sample image with the facial feature points. Figure 5 (c) visualizes the parameter spaces between the input layers and the first hidden layer in BDBN. Each picture shown below represents one neuron in the hidden layer and each pixel quantizes the weight value between that neuron and the one in the input layer. Obviously, the proposed BDBN can automatically extract and emphasize the important areas of human's face, such as the eyes, eyebrows, noses, cheeks, mouths and chins.

Then, we construct the saliency regions based on the emphasized regions of BDBN. Just like the Figure 5 (c), the weight value between each neuron in the input layer to the one in the first hidden layer is calculated at first. Then, the weight value of every neuron is normalized and combined into a saliency map. According to the x and the y coordinates of the 20 important facial feature points of every face image in the dataset, we statistically analyze the percentage of all facial feature points located in the saliency regions of the saliency map.

There are 63.71% facial feature points are located inside 30% most saliency regions and only about 1% facial feature points are located outside 80% most saliency regions. It is obviously that proposed BDBN covers most of important facial feature points. From Figure 4, some of other information and regions are useful to recognize people, such as the hairstyles and the face contour, although they are not belong to the 20 important facial feature points. And as shown in Figure 5 (c), these information and regions are

also emphasized in the parameter space of proposed model. Therefore, if the importance from other important regions for face recognition is excluded, the facial feature points cover percentage in saliency map will be much better.

In Figure 6, the comparison of different computational attention maps are provided, including Graph Gabor attention map (Harel, et al., 2006), Itti classical attention map (Itti & Koch, 2000) and BDBN attention map. It is obviously that BDBN has better coverage than other models. It proves that BDBN provides a human-like judgment by referencing the human visual system.

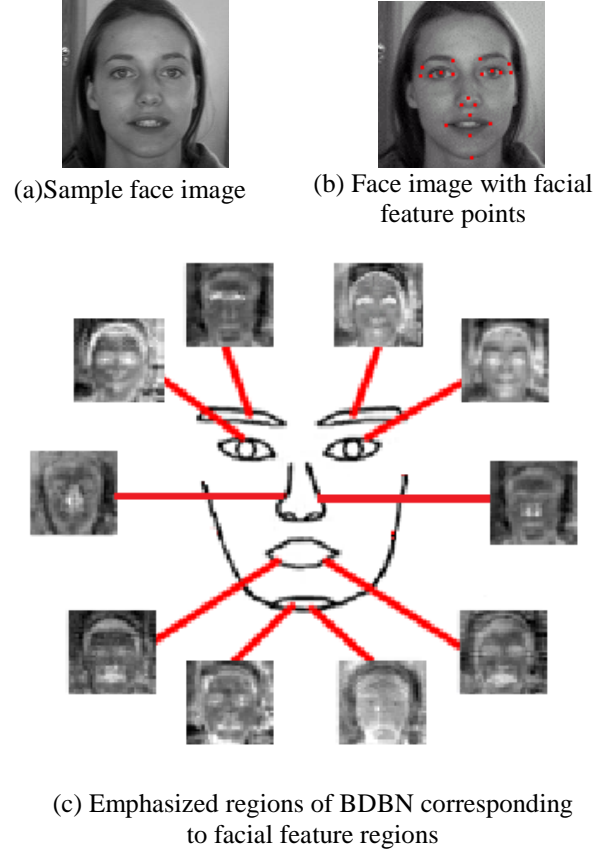


Figure 5: Samples of first layer weights learned by BDBN, and the consistency of these weights with facial feature points.

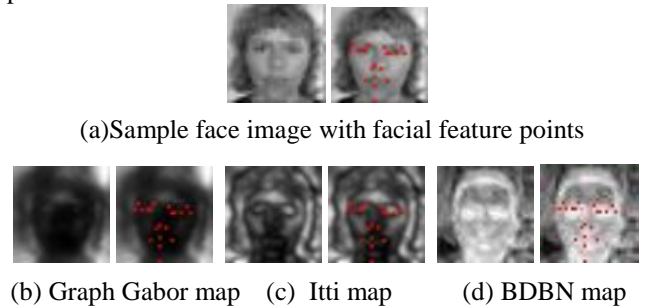


Figure 6: The comparison of different attention maps with facial feature points.



## Conclusion and Future Work

In this paper, we make an attempt to construct an attention model for face recognition in a frame of simulating the human visual cortex and human's perception. To evaluate proposed face recognition models, we do experiments on two face images' datasets, CMU PIE and BioID. Experiments results not only show the distinguishing recognition ability of our deep model but also clearly demonstrate our intention of providing a human-like face image analysis by referencing the human visual cortex and perception procedure.

It is the general opinion that advances in cognitive science especially neuroscience will provide useful insights to computer scientists into how computer models construct, and vice versa. To a certain extent our attempt is an example to prove that the computational models are not only applied into the tasks of classification and recognition just as the optimal classifier, they also can provide human-like response by referencing the human visual system. In future, we will go on this direction to propose novel computational model by referring more characters of human visual system. And vice versa, in cognitive science, we will explore whether the human visual system possess the related mechanism which is consistent with the computational model from the viewpoint of mathematics.

## References

- Anderson, JR, (1998). Social stimuli and social rewards in primate learning and cognition. *Behavioural Processes* (pp. 159–175).
- Afraz, SR, Kiani, R. and Esteky, H., (2006) Nature, 442, (pp. 692–695).
- Avidan, S. and Shamir, A. (2007). Seam carving for content-aware image resizing", In *ACM Transactions on Graphics*.
- Belhumer, P., Hespanha, P., and Kriegman, D., (1997). Eigenfaecs vs. fisherfaces: Recognition using class specific linear projection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, (pp.711-720).
- Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., (2006). Greedy layer-wise training of deep networks, *Advances in Neural Information Processing Systems*.
- Cappelli, R., Franco, A. Maio, D. (2007). Gabor Saliency Map for Face Recognition, In *Proceedings of the 14th International Conference on Image Analysis and Processing*, 443-447.
- Civile, Ciro, McLaren, R.P., McLaren, L.P.L., (2011), Perceptual learning and face recognition: Disruption of second order relational information reduces the face inversion effect. In *33th annual meeting of the Cognitive Science Society*, 2083-2088.
- Fang, F., Qing, L.Y., Wang, C.X., Miao J., Chen X.L., Gao, W.. (2011). Attention Driven Face Recognition, Learning from Human Vision System, In *International Journal of Computer Science Issues*.
- Felleman, D. J., Van Essen, D. C., (1991). Distributed hierarchical processing in the primate cerebral cortex. In *Cereb. Cortex*.
- Gauthier, I., Skudlarski, P., Gore, J.C., & Anderson, A.W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3 (2): 191–197.
- Harel, J., Koch, C. and Perona, P.. (2006). Graph-Based Visual Saliency In *NIPS*.
- He, XF, Cai, D., and Niyogi, P., (2005). Tensor subspace analysis, *Advances in Neural Information Processing Systems*.
- Hickman, L. Firestone, AR, Beck, FM, and Speer, S., (2010). Eye fixations when viewing faces. *Journal of the american dental association jada electronic resource*, (pp. 40–46).
- Hinton, G.E., and Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. In *Science*.
- Hinton, G. E., (2007). Learning Multiple Layers of Representation. In *Trends. Cogn. Sci*.
- HumanScan, (2003). BioID face database. <https://www.bioid.com/download-center/software/bioid-face-database.html>.
- Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. In *Vision Res.*.
- Jesorsky, O., Kirchberg, K., Frischholz, R.. (2001). Robust face detection using the hausdorff distance. In *Proceedings of the 3th International Conference on Audio- and Video-based Biometric Person Authentication*.
- Koch, C. & Ullman, S.. (1985). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry, In *Human Neurobiology*. pp. 219-227.
- Müller, KR, Mika, S., Räsch, G., Tsuda, K., and BSchölkopf, (2001). An introduction to Kernel-based learning algorithms, *IEEE Transactions on Neural Networks*, vol. 12, no. 2, (pp 181-201).
- Mutch, J. and Lowe, DG, (2008). Object class recognition and localization using sparse features with limited receptive fields, *International Journal of Computer Vision*.
- Parkhurst, K. Law, and Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. In *Vision Res.*.
- Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T., (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Sim, T., Baker, S., and Bsat, M., (2002). The CMU Pose, Illumination, and Expression (PIE) Database, *Proceedings of IEEE International conference on Automatic Face and Gesture Recognition*.
- Smolensky, P.. (1986). Information processing in dynamical systems: foundations of harmony theory. In *Parallel Distributed Processing: Explorations in The Microstructure of Cognition*, vol. 1: Foundations, MIT Press, (pp. 194-281).
- Sugiyama, M., (2007). Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. In *JMLR*.
- Turk, M. and Pentland, A. (1991). Face recognition using eigenfaces. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition*, Lahaina, Maui, Hawaii, (pp. 586–591).
- Yan, S., Xu, D. Zhang, B., Zhang, H.J., Yang, Q., and Lin, S., (2007). Graph embedding and extension: a general framework for dimensionality reduction. In *PAMI*.
- Yang, P., Shan, SG, Gao, W., Li, SZ, Zhang, D. (2004). Face recognition using Ada-Boosted Gabor features, *Automatic Face and Gesture Recognition*, (pp. 356-361).
- Wang, Y. Chua, C., Ho,Y., (2002). Facial feature detection and face recognition from 2D and 3D images, In *Pattern Recognition Letters*. 1191–1120.
- Zhong, S.H., Liu, Y., Liu, Y. and Chung, F.L.. (2010). A semantic no-reference image sharpness metric based on top-down and bottom-up saliency map modeling. In *IEEE International Conference on Image Processing*.
- Zhong, S.H., Liu, Y., Liu, Y..(2011). Bilinear deep learning for image classification. In *Proceedings of the 19th ACM International Conference on Multimedia*.



# Cohesion Grading Decisions in a Summary Evaluation Environment: A Machine Learning Approach

**Iraide Zipitria** (iraide.zipitria@ehu.es)

Psychology Faculty (UPV/EHU), Tolosa Etorbidea, 70  
Donostia, E-20018

**Basilio Sierra, Ana Arruarte and Jon A. Elorriaga** (b.sierra, a.arruarte, jon.elorriaga@ehu.es)

Computer Science Faculty (UPV/EHU), Manuel Lardizabal Pasealekua  
Donostia, E-20010

## Abstract

The work presented in this paper has been carried out in the context of a summary writing environment provided with automatic grading. Regarding summarisation discourse, some of the most relevant variables identified in previous work are comprehension, adequacy, use of language, coherence, and cohesion. This work is focused on cohesion. The described exploratory study starts from basic automatic measures of cohesion to further analyse which of them best reflects human expert overall cohesion grades for learner summaries written in the Basque language. For this purpose, 45 basic cohesion measures are compared to overall human cohesion grades. Machine Learning techniques are used to select the best combination for cohesion grading.

**Keywords:** Cohesion grading, machine learning, automatic scoring.

## Introduction

A summary is a short clear description that provides the main facts or ideas about a given topic. In educational contexts, a summary is an overview of the most important information on the studied theme. Summarising requires active meaning construction to a much greater degree than choosing a response in a multiple-choice test, or even than writing short answers to isolated open questions. Thus, not only is summary writing an effective means to construct and integrate new knowledge, it is also a more efficient method for assessing what students do and do not understand than traditional comprehension tests (E. Kintsch, Steinhart, Stahl, & the LSA Research Group, 2000). Thus, summaries are widely used in traditional teaching as an educational diagnostic strategy to infer comprehension, or how much information from the reading text is retained in memory (Bartlett, 1932; Garner, 1982; W. Kintsch, Patel, & Ericsson, 1999).

However, evaluating and grading summaries is a complex and time consuming task for teachers. Human judges have certain variance on summary grading. So, there is a need to systematise written summary evaluation for students. Researchers have sought to develop applications that automate summary grading and evaluation in a way that a given summary will always gain the same score.

Most of the work carried out in Computer Assisted Assessment has tried to infer the student's knowledge comprehension by analysing and comparing the answer generated by the student either explicitly represented in the system –mostly multiple choice questions– or with answers that

could be obtained using the knowledge represented in the system. The automatic evaluation of open-ended text, e.g. summaries, is a complex task strongly conditioned by text comprehension methods; statistical modelling, and Natural Language Processing (NLP) techniques. The open-ended assessment mode, although less accurate than the close-ended mode, has been present in Artificial Intelligence and Education since the very early work in Socratic dialogues systems (Clancey, 1982; Ford, 1988; Woolf, 1988; Winkels & Breuker, 1989). After these first works, there was a period when open-ended approaches had a lower profile, but new developments in NLP and cognitive modelling have seen a revival with a variety of approaches in various applications: dialogue systems (Schulze et al., 2000; A. Graesser, Person, & Harter, 2001; Zinn, Moore, & Core, 2002), feedback on narratives (Robertson & Wiemer-Hastings, 2002), and so on.

The work presented in this paper has been carried out in the context of a learner oriented summary writing environment provided with automatic grading, LEA (Zipitria, Arruarte, & Elorriaga, 2008b). Relevant variables identified when producing a summarisation environment are: text related (text type, text present/absent, theme and text length), discourse related (comprehension, adequacy, use of language, coherence, and cohesion), learner related (learner level and learner's prior knowledge) and available aid tools (dictionaries, spell and grammar check, theory in summarisation strategies, concept maps, schema, etc.). Those variables have been identified after an in-depth study of both the state of the art in summary grading and an empirical study carried out to observe human summary grading performance to model their criteria (Zipitria, Larrañaga, Armañanzas, Arruarte, & Elorriaga, 2008a).

In the context of this work, the global summary grading decisions are gained by means of a Bayesian Network based modelling approach, based on measures such as: comprehension, adequacy, use of language, coherence, and cohesion (Zipitria et al., 2008b).

- **Comprehension.** Comprehension measures the level of understanding that can be inferred from each summary.
- **Adequacy.** It refers to the use of adequate register and terminology in the written summary.
- **Use of language.** It looks at orthographic, syntactic and lexical errors (Cassany, 1993).

- Cohesion and coherence. Coherence and cohesion are closely related and often used as synonyms. A way of distinguishing both concepts is suggested by A. C. Graesser, McNamara, Lowerse, and Zhiqiang (2004), who refer to coherence as a psychological construct, whereas cohesion is referred to as a textual construct. Similarly, Todd, Khongput, and Darasawang (2007) in a connective cohesion study say that cohesion refers to explicit connective links, whereas coherence refers to implicit connections. Therefore, coherence would exist in the way that people interpret text rather than in the texts themselves, while cohesion would be provided by the text features. Cohesion has been defined by Halliday and Hasan (1976) as a set of resources for constructing relationships in discourse transcending grammatical structure (reference, ellipsis, substitution, conjunction, lexical cohesion, etc.). Hence, the aim of cohesive studies is to measure the way text discourse is tied in language. Cohesion features have been measured in this study to resemble human global cohesion grades.

In LEA, comprehension and coherence are modelled based on Latent Semantic Analysis (Zipitria, Arruarte, & Elorriaga, 2006) and adequacy and use of language are computed based on surface measures gathered from tagged text and statistical analysis. The present study describes the procedure followed searching for the best available approach to model overall cohesion grading of learner summaries written in Basque language<sup>1</sup>.

The paper is organised as follows. Section 1 is a summary of previous work that includes measures of cohesion and Section 2 describes the cohesion grading experimental setting, results and discussion.

### Previous work measuring cohesion

Cohesion has already been automatically measured under different approaches and for a variety of purposes.

Morris and Hirst (1991), in a domain independent approach, analyse lexical cohesion in text. Lexical cohesion is measured as a result of chains of related words that contribute to the continuity of lexical meaning. These lexical chains are a direct result of units on the same topic. A thesaurus is used as the knowledge base for computing lexical chains. Lexical chains are also used to determine text structure. E. Kintsch et al. (2000) took an LSA approach to cohesion, gaining sentence to sentence paraphrasing measures for learner summary grading purposes. Alonso and Fuentes (2003) describe the integration of cohesive properties with coherence for automatic summarisation purposes. An account for cohesive formation is gained by means of diagnosis of lexical cohesive chains as extra-strong, strong and medium-strong. A. C. Graesser et al. (2004) present a wide account in cohesion and coherence measures, producing over 200 measures –over 50 types

<sup>1</sup>Non Indo-European language spoken in the north of Spain and south of France. Grammatically complex, it is an agglutinative, order free and verb final language. A complete English description of the Basque grammar can be found in Hualde and Ortiz de Urbina (Hualde & Ortiz de Urbina, 2003).

of cohesion– based on surface linguistic features in a tool called Coh-Metrix. Siddharthan (2006) refers to work in automatic text production that applies a syntactic simplification process as a way to reduce comprehension complexity and maintain cohesiveness. Adequate sentence ordering, cue word selection, referring expression generation, determiner choice and pronominal use are resolved to preserve cohesiveness. Carenini, Ng, and Zhou (2008) work in the context of automatic summarisation of e-mail conversations. Cohesive measures are collected in the form of clue words or word co-occurrences between adjacent fragments, semantic similarity or subsequent sentence similarity measures based on WordNet and cosine – using TF(Term Frequency) and IDF (Inverse Document Frequency), local and global weights respectively – or segment to segment cosine similarity. Finally, Vechtomova and Karamuftuoglu (2008) measure lexical cohesion between query terms in the context of IR (Information Retrieval) term proximity. Both short distance and long distance collocation relations are measured.

### Cohesion grading experiment

As part of the modelling process to obtain global grades for each summary, the summary grading decision making model (Zipitria et al., 2008b) requires global cohesion grades. The goal of this study is to obtain a model which shows which combination of specific cohesion measures best predict cohesion. In other words, which cohesion features predict the decision of global cohesion of graders in comparison to a real-life cohesion grading task. Human cohesion grades are predicted by automatic measures of discourse cohesive features.

### Procedure

17 human experts were asked to grade the level of cohesion of summaries that had previously been gathered from university students, second language learners and primary and secondary school pupils. Experts were university lecturers or primary and secondary and L2 teachers who had been teaching summarisation strategies for more than a decade. A total of 17 summaries were written in Basque language. The goal was to obtain a wide range of different scenarios involving cohesion in summarisation. Grades were gathered on a 1 to 10. Each of the 17 raters produced grades for every summary with a between-rater agreement  $r = 0.7$  and  $p < 0.05$ . Finally, all the grades were discretized into *Fail*, *Pass* and *Distinction*.

The task for expert grading participants consisted of reading the text based on which the summaries were written. Next, they were expected to read each summary to produce global cohesion grades. In order to avoid misconception, verbal and written definitions on cohesion were provided to experts.

In parallel, cohesion measures were automatically modelled using NLP techniques. The mean scores of the graders were compared to cohesion measures in order to observe the amount of information explained by the cohesion measures.

**Cohesion measures (X)** Cohesion measures were created (see Table 1) based on theory on English discourse cohesion (Halliday & Hasan, 1976; Schiffrin, Tannen, & Hamilton, 2001) and language specific differences for Basque (Hualde & Ortiz de Urbina, 2003). In addition, previous modelling work has also been taken into account (Baayen, 2001; A. C. Graesser et al., 2004). 45 markers aiming to word variability, text structure, lexical cohesion, conjunctions and verbal cohesion have been studied:

#### *Word variability*

A total of 14 measures which refer to vocabulary variability related information: X1 Size of the sample in word tokens, X2 Number of distinct lemmas, X3 Number of distinct word tokens, X4 Distinct concept proportion in text, X5 Concept proportion among word variability, X6 Mean number of letters per word, X11 Measures on how single word measures deviate from the central word mean tendency, X12 Mean of word tokens to number of distinct word types, X13 Word proportion in text and X14 Lemma proportion in text.

#### *Text Structure*

This refers to the cohesion which is inherent to the textual structure as narrative, formal correspondence, sonnet, etc. (Halliday & Hasan, 1976). Four surface structure measures have been measured to reflect structure: X7 Mean sentences per paragraph, X8 Number of paragraphs, X9 Number of sentences and X10 Average words per sentence.

#### *Lexical cohesion*

In lexical cohesion the same word is repeated and has the same referent in both cases. It is not necessary for the second instance to be an exact repetition of the same word (Halliday & Hasan, 1976).

Two measures emulate lexical cohesion indices measured by means of overlapping concepts in subsequent sentences. Overlapping concepts are measured as word overlap and lemma overlap: X15 Cosine of overlapping words in subsequent sentence comparison and X16 Cosine of overlapping lemmas in subsequent sentence comparison.

#### *Conjunction and connectors*

Conjunctive elements are cohesive by means of their specific meaning. They express meaning which presuppose the presence of other components in discourse. It is based on the assumption that there are forms of systematic relationships between sentences (Halliday & Hasan, 1976). 16 indices have been measured with the aim of capturing the cohesion provided by conjunctive relations: X17 Average commas per sentence, X18 Measures on how single comma measures deviate from the central word mean tendency per sentence, X19 A rule based approach to the adequate use of the comma, X20 Amount of commas, X21 Number of connectives, X22 Number of additives, X23 Additive type-token ratio, X24 Number of quantifiers, X25 Connector type-token ratio, X26 Number of adversatives, X27 Adversative type-token ratio, X28 Number of distributive connectors, X29 Distributive tokens between connective tokens, X30 Connective tokens between word tokens, X31 Number of types of connectors and X32 Connective tokens times connector variety.

#### *Verbal cohesion*

Verbal forms in Basque provide important ties in discourse cohesion. Verbs can consist of single words (synthetic) or consist of a participial form and an auxiliary (analytical). Auxiliaries can also be used as the main verb. Participles carry aspectual information whereas auxiliaries convey information about argument, structure, tense and mood. Auxiliaries vary in four different tenses/aspects: present, past, hypothetical and imperative (Hualde & Ortiz de Urbina, 2003). Ties provided by verbal forms are measured by a total of 13 indices: X33 Average number of words before verb, X34 Measures on how single measures of word occurrences before verb deviate from the central tendency, X35, verbs per sentence, X36 verbs per sentence by verb variability, X37 Number of Verbs, X38 Number of distinct Verbs, X39 Verb type/token ratio, X40 Number of Transitive Verbs, X41 Number of distinct Transitive Verbs, X42 Transitive Verb type-token ratio, X43 Number of auxiliary verbs, X44 Distinct auxiliary verbs and X45 Auxiliary verb type/token.

The process from text to cohesion measure implementation starts with: (1) Text splitting and tagging. Next, (2) Texts are automatically analysed using POS (Part Of Speech) tagging with a morphosyntactic analyser (Aduriz et al., 2004) and a dependency parser (Bengoetxea & Gojenola, 2009). (3) Finally, there is a statistical processing to obtain the cohesion measures.

The human and automatic cohesion grades obtained were discretized to be analysed under several Machine Learning classification strategies.

## **Results**

This Section describes the ML analysis followed in this study.

**Experimental Design** In order to detect relevant cohesion measures (variables), we first describe how a Feature Subset Selection (FSS) can be performed in an automatic way. After applying different FSS approaches, Feature Selection allows to find the relations between the selected cohesion measures (variables) and the global cohesion grade. The relation is measured based on a set of classifiers. Finally, the goodness of the measure is considered based on the obtained grading accuracy.

In addition to the filtered and wrapper variable selections, the classifiers have also been applied to measure the cohesion for the eight most common variables in previous cohesion measures and the combination of all the 45 measures. The next Section, introduces the description of the variable sorting approach taken for this dataset in the filter approach.

It is worth mentioning that all the experiments have been carried out using the Leave One Out validation technique; this implies learning the classifier with all but one example, and then applying the obtained classifier to the example which has been left out. This process is repeated 17 times (once by example) for each classifier and feature set.

Table 1: Cohesion measures' predictive effect sizes

Type	Measure	$f^2$	$R^2$	Sig.
Word variability	X1	.161	.078	.155
	X2	.154	.072	.163
	X3	.183	.095	.131
	<b>X4</b>	<b>.602</b>	<b>.331</b>	<b>.012</b>
	<b>X5</b>	<b>.602</b>	<b>.331</b>	<b>.012</b>
	<b>X6</b>	<b>.162</b>	<b>.78</b>	<b>.154</b>
	X11	.028	.042	.538
	X12	.014	-.056	.662
	X13	.046	-.025	.438
	<b>X14</b>	<b>.291</b>	<b>.171</b>	<b>.063</b>
Text structure	X7	.029	-.4	.527
	X8	.094	.021	.27
	X9	.136	.057	.189
	X10	.016	-.055	.644
Lexical cohesion	X15	.237	.134	.090
	X16	.138	.06	.184
Conjunction and Connectors	X17	.007	-.063	.750
	X18	.117	.042	.22
	X19	.121	.044	.215
	X20	.107	.032	.242
	X21	.078	.007	.311
	X22	.057	-.013	.385
	X23	.001	-.071	.933
	X24	.001	-.07	.887
	X25	.097	.024	.261
	X26	.049	-.021	.418
	X27	.008	-.063	.737
	X28	.179	.091	.136
	X29	.077	.006	.314
	X30	.103	.029	.248
	X31	.041	-.068	.826
	X32	.091	.019	.276
Verbal cohesion	X33	.007	-.064	.756
	X34	.003	-.068	.831
	X35	.094	.021	.27
	<b>X36</b>	<b>.404</b>	<b>.237</b>	<b>.032</b>
	X37	.18	.094	.134
	X38	.169	.084	.146
	X39	.27	.157	.072
	X40	.052	-.18	.406
	X41	.18	.098	.127
	X42	.169	.084	.146
	<b>X43</b>	<b>.31</b>	<b>.183</b>	<b>.05</b>
	<b>X44</b>	<b>.291</b>	<b>.171</b>	<b>.063</b>
	<b>X45</b>	<b>.322</b>	<b>.19</b>	<b>.05</b>

**Filters** The use of classifiers requires sorting the variables prior to being classified.

In order to perform the experiment and evaluate the adequateness of the new approach, statistical measures have been used to search for the most salient variables for the cohesion problem. The formulas used with this purpose are well-known metrics in Feature Selection and behavioural research methods: *Gain Ratio*, *One Rule*, *Recursive Elimination of Features (RELIEF)*, *Support Vector Machines (SVM)*, *Chi-square ( $\chi^2$ )*, *Principal Component Analysis (PCA)* and *Effect size ( $f^2$ )*. Selected cases are marked in Table 1.

Table 2: Variable ordering obtained for each of the statistical metrics

Metric	$f^2$	GR	OneR	Relief	SVM	$\chi^2$	PCA
First	X4	X15	X35	X44	X41	X15	—
Second	X5	X16	X14	X35	X9	X16	—
Third	X6	X12	X36	X36	X19	X12	—
Forth	X14	X14	X10	X14	X44	X14	—
Fifth	X36	X13	X25	X8	X8	X13	—
Sixth	X43	X20	X5	X9	X23	X20	—
Seventh	X44	X22	X15	X43	X7	X22	—
Eighth	X45	X21	X17	X4	X26	X21	—

**Variable selection based on filtering strategies** Variable sorting under the previously described filtering strategies can

be found in Table 2; it should be noticed that the PCA approach does not give a ranking among the variables, but a set of polynomials with linear combinations of some features, this is the reason why the PCA column in Table 2 is empty.

The ML experimental phase has been organised in the following way: First, each of the five selected classifiers has been used to measure the impact of the cohesion measures for the set of student summary global cohesion grades. As shown in Table 3, the first variable in the ordering given for each metric was taken into account first. Next, a second variable is included for each metric, and the accuracy obtained with these two variables is tested with all the classifiers. The same process is run including the third, fourth, and so on variables until a decrease in the accuracy is obtained. Results in Table 3 show that the variable number for each filter is different depending on the moment when an error increase appears.

Table 3: Number of errors obtained by each approach for each subset of variables. PCA approach does not use individual variables but a linear combination of some of them.

Metric	N	Variables	BN	NB	K-NN	SVM	ANN
6*f2	1	X4	8	12	12	10	8
	2	+X5	8	12	12	10	8
	3	+X14	9	8	11	11	10
	4	+X36	9	9	11	11	10
	5	+X39	9	9	11	10	11
	6	+X43	9	8	12	11	10
4*GainR	1	X15	7	10	10	8	8
	2	+X16	7	13	12	8	8
	3	+X12	7	10	12	8	10
	4	+X14	9	10	11	11	8
8*OneR	1	X35	8	6	7	9	7
	2	+X14	9	5	7	8	7
	3	+X36	9	6	6	6	9
	4	+X10	9	6	6	6	9
	5	+X25	9	6	6	6	11
	6	+X5	9	6	9	6	10
	7	+X15	9	8	10	6	12
	8	+X17	9	8	10	6	10
	9	+X22	9	8	10	6	13
	10	+X23	9	7	10	6	11
	11	+X19	9	8	9	7	9
7*Relief	1	X44	8	5	5	13	5
	2	+X35	8	5	8	6	6
	3	+X36	9	6	6	6	8
	4	+X14	9	6	7	6	8
	5	+X8	9	6	6	6	12
	6	+X9	9	6	7	6	9
	7	+X43	9	7	7	7	9
6*SVM	1	X41	8	9	8	8	9
	2	+X9	9	10	9	7	7
	3	+X19	9	7	7	7	11
	4	+X44	9	5	4	6	10
	5	+X8	9	7	5	7	6
	6	+X23	9	8	8	8	9
4*chi <sup>2</sup>	1	X15	7	10	10	8	8
	2	+X16	7	13	12	8	8
	3	+X12	7	10	12	8	10
	4	+X14	9	10	11	11	8
8*PCA	1	—	8	10	13	9	12
	2	—	8	8	8	9	7
	3	—	8	8	10	9	8
	4	—	8	7	10	8	9
	5	—	8	7	13	8	9
	6	—	8	7	12	9	11
	7	—	8	7	11	9	9
	8	—	9	8	11	9	11

The variable subset with the best results in the filter approach is composed of the first four variables selected with the SVM filtering metric for the K-NN classifier. It shows an accuracy of 4 errors and the combination is compound by the next variables: X41, *transitive verb types*, **verbal cohesion**.

*X9, number of sentences text structure. X19 excessive use of commas connectives. Finally, X44 distinct auxiliary verbs verbal cohesion.* The same combination of variables obtains the best result with the NB paradigm.

Table 4: Number of errors obtained by each approach using the wrapper FSS.

	BN	NB	K-NN	SVM	ANN
Errors	7	5	2	6	4
Variables	X15	X44	X11, X44	X41, X44	X33, X44
ALL	9	9	10	8	12
Experts	9	9	10	5	11

**Variable selection based on wrapper strategies** The variable subset with the best results in the wrapper approach is composed of the first two variables selected with the K-NN classifier. It shows an accuracy of 2 errors and the combination is compound by the next variables: *X11 single word measures deviation from the central word mean tendency* taken from the **Word variability** related variable set, and *X44 distinct auxiliary verbs*, taken from the **verbal cohesion** feature set.

**Variable selection based on some previously used cohesion measures** Previous research (see some examples in Section ) has measured similar factors to account for cohesion. We have selected the next factor combination to observe how they account for cohesion.: *X3 (word types)*, *X6 (mean letters per word)*, *X13 (type-token ratio)*, *X15 (sentence overlap)*, *X21 (number of connectives)*, *X33 (average number of words before verb)*, *X35 (verbs per sentence)* and *X36 (verb variability)*. The combination of the eight previously studied measures (named EXPERT) has been tested under the different classifiers. It should be noticed that in this case the variables are listed using an ascending index. The reason is that the variable ordering is not known. In other words, there is no previous record of one being more relevant than another.

Results are shown in Table 4. The best approximation is provided by the *SVN* variable combination with an accuracy of 5 errors. The results are almost as accurate as the best option based on FSS filter strategies. However, they are still far from the best measures under the wrapper approach and *K-NN*.

**Using ALL the cohesion measures** Another approach is to use all the available indicators to search for cohesion grades. Here, the ALL variable option tests the 45 variables combination for classification purposes.

As shown in Table 4, the ALL variable option does not show accurate results. The accuracy shown is equal to or greater than 8 errors. Again, from the classification paradigms, *SVN* shows the best results.

## Discussion

The goal of this study was to know which measures best predict global cohesion grades. A total of 45 measures were compared to overall cohesion human grades with no previous record on which one was most relevant. The modelling analysis allows searching for the best modelling approach for Basque cohesion grading.

According to the observed results considering all the available information is not the best option for global cohesion grading decision making. The reason for this is redundancy. The EXPERT approach, which combines the most commonly used cohesion measures, has produced a good approximation under SVM classification. Nonetheless, the use of a wrapper approach and K-NN classifier seems to be the best fit for the Basque case.

The difference with the EXPERT combination is probably due to language grammar specific differences. In terms of the variable combination, the amount of auxiliary verb type (*X44*) is the most recurrent one in the best models. This is probably due to Basque grammar morphology. The Basque auxiliary verb carries a lot of grammatical information. Each auxiliary verb provides information about the subject, the two object forms – direct object and indirect object –, as well as tense and aspect. Therefore, the number of auxiliary verb types probably shows how syntactically connected the discourse is. In addition, there are other measures for text structure, word variability, and verbal cohesion which have also been salient.

The obtained model for global cohesion grading will be used as part of a summary evaluation environment (Zipitria et al., 2008a). In order to gain an overall grade for a summary each overall discourse measure is fed into the grading decision making Bayes net (Zipitria et al., 2008b). But, there still are many questions to be answered. Would results be very different if we had measured cohesion indicators that are not included in this study? Does the Basque language require further language specific analysis to better account for cohesion? Would results be very different in another language? Are there interactions among predictors?

We expect that language morphology might be responsible for language differences for cohesion. In future, we aim to analyse the impact that different languages and their morphology make in terms of results. In addition, some of the obtained results might be tied to the particular language under which the study was run. Future work will look at testing the grading scheme under more languages –e.g. Spanish and English– providing LEA with a multilingual approach.

In addition, searching for a greater scope of cohesion measures might also make differences in the results. More theoretically relevant cohesion features (Halliday & Hasan, 1976; Schiffrin et al., 2001) could be automatically modelled and empirically analysed for Basque (e.g. anaphora resolution, ellipsis, etc). A wider collection of measures and further Natural Language Processing tools could allow more in-depth analysis of discourse cohesion and probably a greater accuracy.

## Acknowledgments

This work is supported by the University of The Basque Country (EHU09/09), the Spanish Ministry of Education (TIN2009-14380), and the Basque Government (IT421-10).

## References

- Aduriz, I., Aranzabe, M., Arriola, J., Diaz de Ilaraza, A., Gojenola, K., Oronoz, M., et al. (2004). A cascaded syntactic analyser for Basque. In *Proceedings of computational linguistics and intelligent text processing* (pp. 124–135).
- Alonso, L., & Fuentes, M. (2003). Integrating cohesion and coherence for automatic summarization. In *Proceedings of the 11th meeting of the european chapter of the association for computational linguistics (eac2003)* (pp. 1–8).
- Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). Kluwer Academic Publishers.
- Bartlett, F. C. (1932). *Remembering; a study in experimental and social psychology*. Cambridge University Press.
- Bengoetxea, K., & Gojenola, K. (2009). Exploring treebank transformations in dependency parsing. In *Recent advances in natural language processing, ranlp 2009*.
- Carenini, G., Ng, R. T., & Zhou, X. (2008). Summarizing emails with conversational cohesion and subjectivity. In *Acl-08: Hlt: Proceedings of the 46th annual meeting of the association for computational linguistics: Human language technologies* (pp. 353–361).
- Cassany, D. (1993). *Didáctica de la corrección de lo escrito* (Vol. 108). Spain: Editorial Graó, de IRIF SL. (In Spanish)
- Clancey, W. J. (1982). Tutoring rules for guiding a case method dialogue. In D. Sleeman & J. S. Brown (Eds.), *Intelligent tutoring systems* (pp. 201–226). London: Academic press, inc.
- Ford, L. (1988). The appraisal of an icai system. In *Artificial intelligence and human learning* (pp. 109–123). London: Chapman and Hall, Ltd.
- Garner, R. (1982). Efficient text summarization. costs and benefits. *Journal of Educational Research*, 75(5), 275–279.
- Graesser, A., Person, B., & Harter, D. (2001). Teaching tactics and dialog in autotutor. *International Journal of Artificial Intelligence in Education*, 12, 257–279.
- Graesser, A. C., McNamara, D. S., Lowerse, M. M., & Zhiqiang, C. (2004). Coh-metrix: Analysis of text on cohesion of language. *Behavior Research Methods*, 36, 193–202.
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in english*. Longman.
- Hualde, J. A., & Ortiz de Urbina, J. (2003). *A grammar of Basque*. Mouton de Gruyter.
- Kintsch, E., Steinhart, D., Stahl, G., & the LSA Research Group. (2000). Developing summarization skills through the use of lsa-based feedback. *Interactive learning environments*, 8(2), 87–109.
- Kintsch, W., Patel, V., & Ericsson, K. (1999). The role of long-term working memory in text comprehension. *Psychologia*, 42, 186–198.
- Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17, 21–48.
- Robertson, J., & Wiemer-Hastings, P. (2002, June). Feedback on children's stories via multiple interface agents. In *Proceedings of the 6th international conference its* (pp. 923–932).
- Schiffrin, D., Tannen, D., & Hamilton, H. E. (Eds.). (2001). *The handbook of discourse analysis*. Blackwell Publishing.
- Schulze, K. G., Shelby, R. N., Treacy, D., Wintersgill, M. C., VanLehn, K., & Gertner, A. (2000). Andes: A coached learning environment for classical newtonian physics. *The Journal of Electronic Publishing*, 1(6).
- Siddharthan, A. (2006, June). Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1), 77–109.
- Todd, R. W., Khongput, S., & Darasawang, P. (2007). Coherence, cohesion and comments on students' academic essays. *Assessing Writing*(12), 10–25.
- Vechtomova, O., & Karamuftuoglu, M. (2008). Lexical cohesion and term proximity in document ranking. *Information Processing and Management*, 44, 1485–1502.
- Winkels, R., & Breuker, J. (1989). Discourse planning in intelligent help systems. In C. Frasson & G. Gauthier (Eds.), *Intelligent tutoring systems: At the crossroad of artificial intelligence and education* (pp. 124–139). Norwood, New Jersey: Ablex Publishing Corporation.
- Woolf, B. P. (1988). Representing complex knowledge in an intelligent machine tutor. In J. Self (Ed.), *Artificial intelligence and human learning* (pp. 3–28). London: Chapman and Hall, Ltd.
- Zinn, C., Moore, J. D., & Core, M. G. (2002, June). A 3-tier planning architecture for managing tutorial dialogue. In S. A. Cerri, G. Gouardères, & F. Paraguau (Eds.), *Proceedings of the 6th international conference on intelligent tutoring systems its* (pp. 574–584). Biarritz, France and San Sebastian, Spain: Springer-Verlag.
- Zipitria, I., Arruarte, A., & Elorriaga, J. A. (2006). Observing lemmatization effect in LSA coherence and comprehension grading of learner summaries. In K. Ashley & M. Ikeda (Eds.), *Proceedings of intelligent tutoring systems*. Jhonghli, Taiwan: Springer.
- Zipitria, I., Arruarte, A., & Elorriaga, J. A. (2008b). LEA: Summarization web environment based on human instructors' behaviour. In *Proceedings of 8th international conference of advanced learning technologies* (pp. 564–568).
- Zipitria, I., Larrañaga, P., Armañanzas, R., Arruarte, A., & Elorriaga, J. A. (2008a). What is behind a summary-evaluation decision? *Behavior Research Methods*, 40(2), 597–612.

# The Effects of an Incubation Period on the Metaphor Creation Process

Keiga Abe

Gifu Shoutoku Gakuen University

**Abstract:** This study examines the effects of an incubation period on the creative thinking-process, particularly on the metaphor creation process. In an experiment, participants were asked to create X like a Y type metaphors on the basis of theme phrases, which contain adjectives and nouns. Participants in the control group were asked to create metaphors immediately after they were given the theme phrase. Participants in the incubation group were asked to generate metaphors after a distracter task. Participants in the wait group were asked to create metaphors after-thinking for three minutes. The results of the experiment revealed that when the theme phrases contained connected but complex words, the incubation group created more metaphors than the other groups. When the theme phrases contained diverse but disjointed words, the wait group created more metaphors than the other groups.



# **Trial Measurement of Implicit Attitude toward Violations in Nursing by the use of Implicit Association Test**

**Yuko Adachi**  
University of Tsukuba

**Shinnosuke Usui**  
Osaka University

**Etsuko Nakagami-Yamaguchi**  
Osaka City University

**Akiko Yamada**  
Osaka City University

**Keun sik Park**  
Osaka City University

**Tatsuya Nakatani**  
Osaka City University

**Abstract:** The implicit attitude toward violations is regarded as a contributing factor in those violations. We applied the Implicit Association Test (IAT) to measure this attitude in nursing. This study examines the IAT's validity and reliability. In experiment 1, we conducted IAT and a scale of attitude toward safety for medical staff (Kamizono, 2000) for 71 students from a university of nursing. Validity was demonstrated from a difference of the reaction time between the blocks of IAT and a significant correlation between IAT score and the score of scale. Reliability was indicated by dispersion between trials. In experiment 2, we conducted the IAT for 51 medical staffs who serve as risk managers in a hospital. The IAT's validity was demonstrated from a difference of the reaction time between the blocks of IAT and reliability was indicated by dispersion between trials, too.

# Frame Augmented Language Model

**Kisuh Ahn**

Hankuk University of Foreign Studies

**Eunsuk Lim**

Hankuk University of Foreign Studies

**Abstract:** N-gram based statistical language model is widely used in NLP applications such as Automated Speech Recognition and Machine Translation due to its ease of use and effectiveness. Given the very simple assumption of this model, the effectiveness of this model is somewhat surprising, but there clearly exist deficiencies such as the inability to account for long-distance dependencies and the lack of considerations for the overall meaning.

There have been various approaches to enhance the n-gram language model by incorporating syntactic and semantic elements. In this paper, we explore the use of frames of the Berkeley FrameNet as a way of augmenting the purely statistically driven language model with semantic information; while the conventional n-gram model supplies the overall probability score of the surface form of a sentence, the candidate frames evoked by this sentence provide the means to calculate the conceptual relatedness among the words within the sentence. The two measures are combined via linear interpolation to give an overall score. In addition to boosting the overall performance, we believe that our approach brings the language model a bit closer to the reality of human language processing.

# **Multi-Voxel Pattern Analysis Applied to the Language Switch in the Bilingual Brain—An fMRI Study**

**Hiroyuki Akama**

Tokyo Institute of Technology

**Miao Mei Lei**

Tokyo Institute of Technology

**Na Li**

Tokyo Institute of Technology

**Brian Murphy**

Carnegie Mellon University

**Abstract:** Previous fMRI studies about bilingual speakers proved that the left caudate and the anterior cingulate cortex (ACC) play the important role in intralanguage task switching (Crinion et al., 2006; Abutalebi et al., 2008). Our research based on Multi-voxel pattern analysis (MVPA) suggests that these brain regions are involved in the language control not only, but also in semantic processing which encapsulates a covert translation between the first/second native languages. Using a set of stimuli which was composed of two semantic categories, Korean-Chinese bilinguals were requested to do a property generation task in Korean for a stimulus word in Chinese, vice versa. Machine learning methods applied to the fMRI datasets enabled us to create, even from the aforementioned regions of interest (relatively small with only about 1000 voxels), significant classification models with the identification accuracy of more than 60% (above chance) for predicting the semantic category of each target.

# Visual prosody: The relationships between head movements and the verbs

Haruka Amatani

University of Tokyo, Meguro-ku, Tokyo, Japan

**Abstract:** Visual prosody consists of head and facial motions which accompany speech prosody. The improvement of speech intelligibility by head movement has been reported by Munhall et al. (2003) and the increase in the perceptual rate of a sentential focus with head or facial movement by Krahmer and Swerts (2007).

Here, I will show that the syntactic structure can influence visual prosody.

With video clips of two Japanese newsreaders from thirty-nine news programs, a perceptual experiment was carried out on untrained twenty-six subjects. They were instructed to judge the timing when the newsreaders made a head nod. It turned out that the speakers had a strong tendency to nod at the matrix verbs. As for the embedded verbs, those in the sentential compliments and in the adverbial phrases frequently co-occur with head nods, whereas those in the relative clauses which modify nouns did much less frequently.

# Determining the effect of ego-involvement on causal reasoning by using contingencies in causal situations

Yoshiko Arai  
Osaka City University

**Abstract:** This study investigates the relation between contingencies and the effect of ego-involvement on causal reasoning in causal situations. In an experiment, participants were informed about their tasks that involved causality, and they were randomly assigned to four conditions: medium contingency without ego-involvement, medium contingency with ego-involvement, high contingency without ego-involvement, and high contingency with ego-involvement. Thereafter, they were asked to rate the strength of the causality of two events. The result showed that the effect of ego-involvement was clear only when the contingency was high: when ego-involvement was present, the causal relationship was judged stronger than when it was absent. This is similar to the result of an experiment in which contingency was considered as a within-subject variable. The result confirmed that ego-involvement affects causal reasoning when the contingency is high.

# Which is Stronger? : Discriminative Learning of Sound Symbolism

**Eiji Aramaki**

The University of Tokyo

**Sachi Yasuda**

The University of Tokyo

**Mai Miyabe**

The University of Tokyo

**Satoshi Miura**

Tottori University

**Masaki Murata**

Tottori University

**Abstract:** Abstract: The importance of sound symbolism of a product name has been emphasized in recent researches. However, detailed mechanisms of sound symbolism remain controversial, even in reports of recent studies. This study examines a method to detect sound symbolism using discriminative learning. First, we build a training dataset comprising name pairs: (1) name-A, (2) name-B, and (3) a label showing which has stronger sound symbolism. Next, we train a Support Vector Machine (SVM) to learn data using both character-based features and phoneme-based features. In experiments, the proposed method demonstrated almost identical performance to that of humans (72% agreement ratio to humans). This paper also presents a method to generate new names with strong sound symbolism based on greedy search. The generated names have high agreement to human judgments (84%). This the first study suggesting that machine learning can detect sound symbolism.

# The model comparison through orthography, phonology, and semantics

**Shin-ichi Asakawa**

Tokyo Woman's Christian University

**Abstract:** We tried to compare the performances of 3 neural network models. Those were perceptrons, back-propagations, and attractor networks. Perceptrons are two-layered model without any hidden layers. On the other hand, back-propagation models are 3-layered models with a hidden layer. In addition to the hidden layer, attractor networks have a cleanup layer from/to the output layer. We had all the models learnt the data sets of Hinton & Schallice(1991), Plaut & Schallice (1993), and Tyler et al(2000). The components of language processings are divided to three parts, orthography, phonology and semantics. The comparison among the models and the data sets revealed adequacy as a model of dyslectic patients. It also revealed that the cleanup layer had to play an important roll in order to process all the data set. The category specificity, which was defined as the inner correlation matrix between concepts, could be simulated, as well.



# Age-related Differences in Implicit Memory for Distractor Kanji Characters

**Akihiro Asano**

Chuo University

**Etsuko T. Harada**

University of Tsukuba

**Shoko Saito**

Chuo University

**Abstract:** Experiments investigating the effects of age and time of testing on implicit memory for distractors were executed with Japanese older and young adults from cross-cultural perspective with a modification of a Rowe, Valderrama, Hasher, and Leanartowicz (2006) procedure. Participants were required to make same or different judgments on line drawings superimposed with irrelevant Kanji characters, and then memory for distractors was tested with a Kanji perceptual identification test. Against expects according to Rowe et al. (2006), and supporting a prior research (Asano, Harada, Suto, Rowe, & Hasher, 2009) used Hiragana words as distractors, results showed that only young adults showed priming for the distractors, without any effects of time of testing. To explain these differences between studies, some hypothesis about cultural or linguistic differences between North Americans and Japanese as East Asians in cognitive processing styles will be discussed.

# **Robots as learning partners in collaborative learning research**

**Jun Ashikaga**

The University of Tokyo

**Takahiro Nakayama**

The University of Tokyo

**Sho Inaba**

The University of Tokyo

**Kenta Iyoki**

The University of Tokyo

**Naomi Miyake**

The University of Tokyo

**Abstract:** Remotely controlled robots can expand our research and practice of collaborative learning. Such robots can be used to run a controlled experiment in classroom-like settings by delivering the same information to different groups, so that we can explore the effects of a particular discourse in collaboration. Using this technique, we have tested 25 groups in jigsaw-like, knowledge-constructive collaborative learning situations, with eighty 5th and 6th graders on science topics. There was one robot for each group, which consisted of 2 to 4 children. We have found that (1) children collaborated well among themselves and (2) they could learn well with such robots. We have also found the groups whose talks focused more on the task performed better than those with less focused talks. Currently we are analyzing the protocols to see what kind of intervention by the robots would have the children focus more on the task.

# What do children hear? Japanese parents' use of numeral classifiers

**Natsuki Atagi**

University of California, Los Angeles

**Catherine Sandhofer**

University of California, Los Angeles

**Abstract:** Previous work on childrens acquisition of Japanese numeral classifiers has provided extensive information regarding classifier acquisition and its effect on cognitive development (e.g., Yamamoto, 2005). However, few studies have examined the considerable variability found in these studies. We propose that variability in childrens classifier acquisition may be due to input. Thus, we examined parent input of Japanese numeral classifiers to Japanese-speaking children. Participants were two- to five-year-old monolingual Japanese and bilingual Japanese-English children and their Japanese-speaking parents. Parents were instructed to read a wordless picture book about counting to their children. Book readings were video-recorded and coded for frequency, type, and correctness of parent classifier use. Children also participated in Give-N and counting tasks. We hypothesize that children whose input includes many generic classifiers (e.g., -ko: generic classifier for inanimate objects) will have more advanced number understanding than children whose input includes more specific classifiers (e.g., -ken: classifier for houses).

# Comparison and contrast in novel objects categorization: the role of executive functions

**Luc Augier**

University of Burgundy

**Jean-Pierre Thibaut**

University of Burgundy

**Abstract:** We investigated how 4- and 6-year-olds use within and between category comparisons when they generalize novel names for novel objects on a nonsalient dimension such as texture (rather than shape). Children first learned a novel name for one or several stimulus. In the generalization phase we pitted a texture match against a shape match and we manipulated the quantity of information (positive or negative evidence). We manipulated number of standards, presence of contrast and age. Our results confirm the role of within category comparisons in texture choices. Further, older children performed significantly better than younger children only when there were four standards. Hence, increasing the load of comparison does not hinder but may be of more help for older than for younger children. We also found a beneficial effect of between category comparisons for both age groups. This is compatible with the role of executive functions in comparisons.

# **Comparison of the brain regions activated during comprehension of action sentences referring to unimanual or bimanual actions: An fMRI study**

**Shunji Awazu**

Jissen Women's University

**Fumihiko Taya**

Keio University

**Sayako Masuda**

Keio University

**Shigeru Watanabe**

Keio University

**Abstract:** Within the framework of embodied cognition, sentences are thought to be comprehended by mental simulations in the sensorimotor system. Several fMRI studies have shown that motor regions are activated during action-sentence comprehension tasks. In this fMRI study, we examined whether all body movements are mentally simulated during these tasks. We assigned action-sentence comprehension tasks to 18 participants for identifying brain regions that are activated while reading sentences referring to unimanual or bimanual actions. When the participants read the bimanual action sentences, left SPL (Brodmann Area 7), IPL (BA40), SFL (BA6) and MCgG (BA23, 24) were activated. Direct comparison between the two conditions found significant activation of left IPL (BA40), IOG and MTG (BA18) in reading the bimanual sentences. Therefore, we conclude that while reading bimanual action sentences, mental simulations occur for bimanual movement as well as for visual experiences. Broad range of brain regions would associate with mental simulations of action sentences.

# Effectiveness of Transcranial Direct Current Stimulation on Medial Prefrontal Cortex in Aesthetic Judgement

**Leila Azari Pishkenari**

Graduate Student at Institute for Cognitive Science Studies, Tehran, Iran, Islamic Republic of

**Hamed Ekhtiari**

Neurocognitive Laboratory, Iranian National Center for Addiction Studies, Tehran, Iran

**Mohammad Javad Hatami**

Head of the MA Program in Cognitive Psychology

**Abstract:** Introduction: Among various brain regions involved in aesthetic judgment, Medial prefrontal cortex has a pivotal role in judging the beauty of visual stimuli. It seems that this regions influence in aesthetic preference is an effect of its role in affective processes. In the following study, we have used transcranial Direct Current Stimulation (tDCS) in order to evaluate the role of medial prefrontal cortex in aesthetic judgment. Method: we used three different types of tDCS stimulation, that is, anodal, cathodal, and sham. 36 participants (18 female) undertook in three experimental sessions randomly in which they received 1mA stimulation for 20 min on their medial prefrontal cortex. Active electrodes were located bilaterally on the forehead and the reference electrode was on the right arm. Ten minutes after the onset of stimulation, subjects got involved in the on-line computerized task of the aesthetic judgment. Results: In general, the effect of tDCS on medial prefrontal cortex on aesthetic judgment was significant  $F(2 \text{ \& } 37.84)=3.89$ ,  $p=0.029$ ). The results show that anodal stimulation of the medial prefrontal cortex affect the aesthetic preference significantly ( $p=0.036$ ), while no such effect was seen in cathodal stimulation ( $p=0.663$ ). There was no sex-related effect ( $F(1 \text{ \& } 33.48)=3.39$ ,  $p=0.074$ ). Discussion: Medial prefrontal cortex through its top-down control over affective side of aesthetic preference can reduce the preference.

# Learning novel words with the help of morphological information

**Sungbong Bae**

Kyungnam University, Korea

**Kwangoh Yi**

Yeungnam University, Korea

**Abstract:** Morphological information can affect the learning rate of new words. To test this hypothesis, we manipulated two variables: individual difference in Morphological Awareness (high vs. low MA groups) and sentential context (consistent vs. inconsistent sentences). 152 college students were asked to learn new words appearing within the contexts of a couple of sentences. The sentential contexts on some trials were semantically consistent with one of the possible morphological structures of words to be learned. The participants with higher MA were better at learning new words despite the definitions of words and morphemes were not given explicitly. Also, the words presented in morphologically consistent contexts were learned faster. More importantly, the effect of context differed among MA groups. The supportive context effect was much bigger for the participants of high MA. In conclusion, this study showed the significant effects of morphological representation and processing in learning new words.



# Distant Border Color Is More Preferred In a Triple Color Combination

**Ziba Bashardanesh**

Institute of Cognitive Science Studies, Iran

**Ali Yoonessi**

School of Advanced Medical Technologies, Tehran University of Medical Sciences, Iran

**Abstract:** Introduction. Previous studies for color preference are mostly limited to choices of one or two color. We used a novel stimulus, a combination of two adjacent squares with a narrow rectangle (border) in between to assess the effect of a third color in the preference for a color combination. We used this method to evaluate the distance of the border color and two peripheral colors in the DKL color space as a predictor for preference. Method. Stimuli consist of all 504 combinations of 9 selected points in the DKL space. On a rectangular representation of DKL with gray in the middle, 9 points were selected including 4 corners, the center and 4 mid-sides. Each possible combination compared with the rest in a forced two choice task in 3024. 50 subjects (25 female and 25 male) in the range of 20-40 years old participated in the experiment. Results. In general, the greater the sum of distances between colors correlated with more preference in the combinations ( $T= 3.585$ ,  $p\text{-value}=0.00$ ). In addition, if the border line settles in longer distance from two other colors, it was preferred more ( $R= -0.726$ ,  $p\text{-value}= 0.00$ ). Conclusion. Distant border color is more preferred in a triple color combination.

# **To peek and to peer: "visual" verb meanings are largely unaffected by congenital blindness**

**Marina Bedny**

MIT

**Jorie Koster-Hale**

MIT

**William Johnston**

MIT

**Lindsay Yazzolino**

MIT

**Rebecca Saxe**

MIT

**Abstract:** Congenitally blind adults learn about the world through touch, audition, and language, but not through vision. What consequences does this atypical sensory experience have for blind adults' concepts of actions and events, especially for features related to vision? One way to assess the structure of concepts is to construct similarity matrices. We elicited similarity judgements for pairs of verbs describing manner of motion (e.g. "to spin", "to strut", n=15), perceptual experience (e.g. "to peek", "to peer", n=60), or perceptible qualities (e.g. "to shimmer", "to shine", n=45). Some of the verbs described qualities or experiences linked to vision. Similarity judgements were acquired from sighted (n=22), late blind (n=9) and congenitally blind (n=24) participants. The similarity ratings for all verb categories, including "visual" verbs, were remarkably similar across groups (all  $r > .85$ ); cluster analyses on similarity ratings produces nearly identical clusters. These results suggest that: 1) verb meanings are largely unaffected by congenital blindness 2) the sensory modality of experience has little effect on conceptual structure.

# Cognitive typology

**Chuluundorj Begz**

University of the Humanities

**Abstract:** This paper presents the dynamic, psycho-cognitive approach to study of human verbal thinking on the basis of typologically different languages /as a Mongolian, English and Russian/. Topological equivalence in verbal communication serves as a basis of Universality of mental structures and therefore deep structures. Mechanism of verbal thinking consisted at the deep level of basic concepts, rules for integration and classification, neural networks of vocabulary. According to the results of psycho-cognitive analysis, semantic bootstrapping serves as a basis of grammatical categories, blending and mental mapping - as a basis for high level linguistic structures (sentence, discourse).

In terms of typologically different languages, topological structure of discourse represents the way in which the individual organizes semantic content, concepts and propositions, in his/her cognitive structure.

Comparative analysis of above named languages has shown that analogical mapping, inference between the properties (like space and time, quantity and quality, shape, color, volume structure) present a basis for semantics transformations. Comparative analysis of the influence of language on cognition proposes a pathway for new developments in psycho-cognitive research, particularly for integral generative semantics in 2D and 3D dimensions.

Key words: verbal thinking, mental structures, blending, mapping, bootstrapping, integral semantics.

# **Expert Memory in Blindfold Chess960 - An Interpretation in the light of LIDA**

**Eduardo Bermudez**

Universidad del Atlantico, Barranquilla, Atlantico, Colombia

**David Dahmen**

Universidad del Atlantico, Barranquilla, Atlantico, Colombia

**Henry Gonzalez**

Universidad del Atlantico, Barranquilla, Atlantico, Colombia

**Abstract:** Chessplayers experts perception is heavily theory laden as evidenced by the performance at recalling chess positions of an expert only allowed to glimpse the board for no more than 10 seconds. When the theoretical background is partially destroyed as in Chess 960, the recall under the same conditions is diminished from about 80% correct recalls to less than 60% (in our experiments). The view that "perception is cognition" seems to have a relation to the concept of sparse dictionaries in the software framework of learning intelligent distribution agent (LIDA). The particular experiments of this paper argue for the former concept and the probable existence of independent networks of activation of memory such as Damasio's zones of convergence-divergence, also compatible with the LIDA concept. Such an interpretation leads not only to reaffirm Ericsson & Kintsch's concept of Long Term Working Memory, but also the hypothesis of a unified theory of memory.

# Vague Linguistic Expressions and the Problem of Equidistance in Verbal Response Scales

**Franziska Bocklisch**

Chemnitz University of Technology

**Josef Krems**

Chemnitz University of Technology

**Abstract:** This contribution examines the problem of equidistance of vague linguistic terms (LT) of verbal response scales, which are used in psychological questionnaires or interviews. The study design (N = 92) employs an empirical translation procedure for the numerical translation of LTs in an example questionnaire (i.e., measuring chronic stress) that utilizes a verbal response scale with frequency expressions (e.g., sometimes). The data are modeled using fuzzy membership functions (MFs) that reflect the LTs meanings. Results show that the LTs of the original questionnaire scale are not distributed equidistantly and, therefore, that the presuppositions for data analysis are statistically violated. Further, this violation biases study results in such a way that measured stress levels are overestimated. A proposed alternative scale with different LTs shows nearly equidistant response categories. To solve the problem of equidistance, a fuzzy analysis of response data is presented and compared to conventional analyses.

# Going beyond the headlines: Narratives mitigate intergroup empathy bias

**Emile Bruneau**

Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

**Mina Cikara**

Massachusetts Institute of Technology

**Rebecca Saxe**

Massachusetts Institute of Technology

**Abstract:** Three Afghani civilians were killed in a drone strike in Khandahar.

People increasingly get their news from online sources, making it easy to see the world through the prism of headlines, instead of the longer story. What are the consequences of reading just a single sentence encapsulation of world events? We tested the effects of reading headlines versus longer narratives, about events happening to in-group versus out-group members, for people assigned to arbitrary competitive groups (Eagles or Rattlers). In four studies, reading just headlines exacerbated intergroup differences in empathy: feeling more empathy for in-group members, and more counter-empathic emotions (e.g. Schadenfreude) for outgroup members. Headlines encourage attending to, and therefore remembering, what group a person is from. However, the intergroup empathy bias can be mitigated by providing a short narrative about each individual, which draws attention away from group membership and towards the individuals experience.

# Ordered Information and Word Learning: An Associative Learning Perspective

**Joseph Burling**

University of Houston

**Hanako Yoshida**

University of Houston

**Abstract:** A computational associative learning simulation is implemented to gauge the role of temporally ordered information in regards to establishing differently weighted representations. Simulation comparisons are made between two approaches. One approach simultaneously learns multiple sets of cues, which are trained to be associated with their respective outcomes. The other approach separates training into temporally distinct early and late phases. Qualitative differences in weighted representations are observed between the two types of training. Asymmetrical learning of features occurs in the simulation with temporally distinct learning instances, and not in the simultaneous learning model. Negative associations are learned for specific cues when competing information is introduced at a later time point. These findings provide discussion about the influence of order effects in terms of word acquisition by attending to relevant features over time, and in general, the role cue competition in relation to temporal learning during early development.



# The Hierarchical Structure of Word Senses

**Hee-Rahk Chae**

Hakkuk University of Foreign Studies

**Do-Il Hong**

Hakkuk University of Foreign Studies

**Abstract:** One of the most controversial issues in lexical semantics is the distinction between polysemy and indeterminacy/vagueness. We need to figure out whether a set of meanings are realizations of multiple senses of a word or manifestations of a single sense. There have been proposed many diagnostic tests for the distinction. Although they are useful to some degree, none of them are without problems. One of the reasons for these problems is that not all senses are equal on their statuses: there are not only full senses but also sub-senses. It seems to be impossible to set up a fixed set of all-purpose criteria for the distinction. For example, the level of sense discrimination for information retrieval would not have to be as deep as that for inference. In this study, we will focus on characterizing the hierarchical nature of (Korean) word senses, in due consideration of the conceptual units which are building blocks of senses.

# Word Segmentation Difficulty in Fourth Graders with Low Reading Achievement

**Sau-chin Chen**

Tzu Chi University

**Jenn-Yeu Chen**

National Cheng Kung University

**Abstract:** A series of studies have advertized the benefit of word segmentation in Chinese sentences comprehension for children with reading difficulty. This study attempts to identify the major cognitive factor contributing to this developmental difference among the children with the same age. A sentence reading with word judgment task was designed for a group of children with reading difficulty and a control group of children with normal reading capacity. In each trial of word judgment, a stimuli sentence could follow a word appearing in the critical region, a word appearing in the critical region but without meaning, a word appearing in the uncritical region, or a word disappearing in the sentence. For the words in the critical region, there is a ambiguous segmentation between them, and the children takes more cognitive effort to deny the word without meaning. The current results show the only difference between the two groups on the words without meaning. After excluded the factors such as intelligence and vocabulary, the inefficiency of attention would be the critical to reading difficulty. An improvement method based on a cognitive processing of Chinese sentence reading is suggested in this presentation.

# Is the Proximate Unit in Chinese Word Production Motivated by the Visual Prompt of the Task?

**Train-Min Chen**

National Cheng Kung University, Tainan, Taiwan, R.O.C.

**Jenn-Yeu Chen**

National Cheng Kung University, Tainan, Taiwan, R.O.C.

**Abstract:** Previous word production research employing the implicit priming paradigm has shown that speakers can benefit from the advanced knowledge of the initial word form of the word to be produced. In Dutch and English, a single onset segment is sufficient to produce the benefit, but a complete syllable is required in Mandarin Chinese. The absence of an onset effect in Mandarin Chinese might have to do with the orthographic characteristics of the prompts, which are syllable-based and could have motivated the production system to place more emphasis on the syllable than on the segment. The present study employed the same paradigm but with spoken prompts in addition. It turned out that there was a syllable effect but not a segment effect, irrespective of the modality of the prompts. The findings suggest that the proximate unit in Mandarin Chinese word production is an intrinsic effect, and not an accidental, task-dependent artifact.

# **The Impact of Contextual Cues on Infant Categorization**

**I-Chen Chen**

National Cheng Kung University

**Marin Huang**

National Cheng Kung University

**I-Wen Yu**

National Cheng Kung University

**Pei-Ling Wang**

Taipei Municipal University of Education

**Jon-Fan Hu**

National Cheng Kung University

**Abstract:** When Infants perceive and categorize objects in natural environment, they would also found the contextual cues useful. Therefore, how this factor influences infant categorization is an important issue that worth further studying. Experiment in the present study used eye-tracking technique to accurately capture 6-month-old infants eye movements in a categorization task, investigating the role the contextual cues would play for categorizing animals and vehicles. The familiarity-novelty preference paradigm was adopted to assess infants performance of categorization. It was found that infants looked more to the pictures of vehicles than animals in both congruent and incongruent conditions during the familiarization phase. The results showed that object-context relations play a specific role in infant categorization.

Keyword: infant categorization, contextual cues, eye-tracking

# Knowing where to look: Conceptual knowledge guides fixation in an object categorization task

Lang Chen

University of Wisconsin-Madison

Timothy Rogers

University of Wisconsin-Madison

**Abstract:** Increasing evidence suggests that humans are not passive perceivers of the world but actively seek out perceptual information. A pressing question for cognitive science concerns how conceptual representations influence this active perception. To address this question, we investigated fixation patterns during visual object classification while manipulating the level of specificity at which a given item was categorized. Twenty-nine undergraduates verified whether a picture matched a preceding verbal label while their eye-movements were monitored. When pictures (e.g., beaver) were classified at a specific level (BEAVER), participants showed significantly increased dwell time to distinctive features (tail) but decreased dwell time to common features (e.g., face) compared to classification of the same picture at a general level (e.g. ANIMAL) although participants attended to both distinctive and common features in both conditions. The results suggest that conceptual knowledge can substantially influence how viewers direct their gazes and consequently focus on the relevant perceptual information.

# Mindset interacts with subjective knowledge but not fluency when affecting preferential judgment

Rongjuan Chen

Stevens Institute of Technology, Hoboken, NJ, United States

Yasuaki Sakamoto

Stevens Institute of Technology, Hoboken, NJ, United States

**Abstract:** Fluency can influence various judgments. When processing fluent information, people think more intuitively than when processing disfluent information. Recent research showed that mindset (abstract vs. concrete) would moderate fluency effects on consumers preference. Other research suggested that subjective knowledge (e.g. estimated price) could influence product judgment. In this work, we manipulated mindset (abstract vs. concrete) and the fluency of an advertisement of chocolates (easy vs. hard), tested subjective knowledge of price (high vs. low), and examined the effects of these factors on preference, measured by how much they liked the chocolates (liking), how desirable the chocolates were (desirability), and how much they wanted to eat the chocolates (wanting). We found direct effects of fluency and price, but not mindset, on preference. However, mindset interacted with price to affect liking and wanting. We conclude by discussing why mindset interacts with subjective knowledge, but not fluency, when affecting preferential judgment.

# A Study of the Syntactic Category of Rhetorical Questions

Hongbo Chen

Liaoning University of Technology

**Abstract:** A review of the literature on the study of rhetorical questions (RQs) shows that most of the previous studies are aimed to describe the pragmatic function, while the cognitive factor has been considered by relatively few researchers. The major goal of this study is to discuss the syntactic category of RQs under the Prototype Theory of Categorization. The author first set up a theoretical frame for analyzing RQs, suggesting that all grammatical structures are prototypical categories, which fall into a category of syntax, a category of semantics and a category of pragmatics and a category of syntax is a category of different semantic meanings and pragmatic functions expressed by the same syntactic form. An analysis of RQs shows that RQs and genuine questions are in the same syntactic category QUESTION because they share syntactic similarities. Keywords: rhetorical questions; the Prototype Theory of Categorization; syntactic category of RQs



# Eye Movement Patterns Reveal How Oriental People Group Objects

**Fan-Ning Cheng**

National Cheng Kung University

**Yi-Rong Wu**

National Cheng Kung University

**Jo Pan**

National Cheng Kung University

**Gert Westermann**

University of Lancaster

**Hsueh-Chih Chen**

National Taiwan Normal University

**Jon-Fan Hu**

National Cheng Kung University

**Abstract:** Past literature pointed out that oriental people tend to consider thematic relationship more than taxonomic relationship (i.e. group banana & apple together more than banana & ape) compared to western people for thematic relationship. However, how oriental people judge the relationship between objects by grouping remains unclear. An eye tracker device was used to examine whether oriental people group objects thematically as the first step to investigate the issue. It was further assumed that people would tend to look longer on the image pairs which are found surprising and unexpected since they group those objects based on specific relationship from an opposite direction. Twenty-three oriental background subjects completed the experiment. The results showed that oriental background participants do view taxonomic pairs significantly longer than the thematic pairs in terms of total fixation duration and total fixation count by which the earlier hypothesis is supported.

Keywords: oriental, eye-tracker, taxonomic, thematic, categorization

# **Reading and Writing Performance in School-aged Children with Specific Language Impairment or/and with Developmental Coordination Disorder Identified at Preschool Age**

**Rong-Ju Cherng**

Department of Physical Therapy and Institute of Allied Health Sciences, National Cheng Kung University,  
Tainan, Taiwan

**Hsiang-Chun Cheng**

Department of Physical Therapy, HungKuang University, Taichung, Taiwan and Institute of Allied Health  
Sciences, National Cheng Kung University, Tainan, Taiwan

**Jenn-Yeu Chen**

Department of Psychology and Institute of Cognitive Science, National Cheng Kung University, Tainan,  
Taiwan

**Chia-Liang Tsai**

Institute of Physical Education, Health & Leisure Studies, National Cheng Kung University, Tainan, Taiwan

**Miau-Lin Shen**

Department of Early Childhood Education, Asia University, Taichung, Taiwan

**Abstract:** Ninety-seven children with a diagnosis of specific language impairment (SLI: 43) or developmental coordination disorder (DCD: 41), or both disorders (13) at preschool and 323 typically developing (TD) children were tested with the Chinese Reading Achievement Test (CRAT) and the Basic Reading and Writing Test Battery (BRWTB) to measure their reading and writing performance when they were 7-8 years of age. The results showed that children with SLI scored lower on CRAT and BRWTB than TD children. Children with DCD scored similarly to TD children on the two reading tests, but scored lower on the writing subtest of BRWTB. The incidence of dyslexia in children with SLI, children with DCD, children with both disorders and TD children is 48.6%, 14.3%, 50% and 12.9%, respectively. Early assessment of the language and motor development in children at preschool may help to predict their reading and writing performance at the early years of schooling.

# The development of numerical comparison in 2- to 4-year-old children

**Pierina Cheung**

University of Waterloo, Waterloo, Ontario, Canada

**Mathieu Le Corre**

University of Waterloo

**Abstract:** Infants deploy two cognitive systems that support numerical thinking: one for representing small sets of objects ( $<4$ ) in parallel, and another for representing large, approximate numerical magnitudes ( $>4$ ). Despite infants demonstrated numerical competence, preschoolers fail to show which of two sets contains more elements until they have acquired some number word meanings (e.g., Brannon & Van deWalle, 2001). However, previous studies mixed up small and large numbers. The current study asked whether preschoolers failure depended on the number of elements in sets. Two- to 4-year-olds were shown either small or large numbers of rectangles, and were asked to choose the set that had more. We analyzed childrens performance based on their number word knowledge. Results found that children were better at comparing small than large sets, and numeral acquisition facilitated numerical comparison. Implications for the role of the two cognitive systems and number word learning in numerical comparison were discussed.

# **The examination of physiological factor on two context effects in multi-attribute decision making**

**Itsuki Chiba**

Rikkyo university, Niiza, Saitama, Japan

**Takashi Tsuzuki**

Rikkyo university, Niiza, Saitama, Japan

**Masashi Soma**

Rikkyo university, Niiza, Saitama, Japan

**Abstract:** This experiment used the attraction effect and compromise effect to test the influence of ingestion of sugar on reliance on intuitive, heuristic-based decision making. And also for a search of these occurrence mechanisms, we measured the eye movement and negative emotion of the subjects. In the context-dependent effect, a difficult choice between two options is swayed by the presence of a third (normatively irrelevant) alternative. Previous works showed that the attraction effect increase and the compromise effect decrease when people have depleted their mental resources performing a previous self-control task. Moreover, it is thought that the attraction effect arise for the aversion from the negative emotion which feel when the trade-off has been perceived. We replicated these findings and analyzed the eye movement and the negative emotion to test the difference of two context-dependent effects.

# **A Classroom Study of Learning to Evaluate Scientific Evidence**

**Clark Chinn**

Rutgers University, New Brunswick, NJ, United States

**Abstract:** One of the goals of education is to promote growth in reasoning; a critical subgoal is to promote competence in critically evaluating the quality of evidence (e.g., evaluating comparisons, soundness of measures, etc.). In a classroom experiment with 4 science teachers, 14 science classes, and 225 seventh graders, two methods of promoting growth in the ability to evaluate evidence were contrasted over 10 lessons in which students evaluated evidence to solve scientific problems. Half of the classes engaged in active reasoningextensive group and class argumentation about evidence quality. The other half engaged in active reasoning plus expert modeling; these students engaged in the same argumentation but also listened to and discussed short conversational exchanges in which scientists discussed how they evaluated evidence quality. Students in the active reasoning plus modeling condition outperformed students in the active reasoning condition. Implications for the development of reasoning and for instruction are discussed.

# Impact of Modified Cognitive Behavior Therapy on Children with Autism Spectrum Disorders

**Suvarna Chinta**

International Institute of Information Technology -Hyderabad

**Bipin Indurkha**

International Institute of Information Technology-Hyderabad

**Abstract:** Autism Spectrum Disorder (ASD) is Neuro Developmental in origin with triad deficit (DSM-IV, APA, 1994), co-morbid behaviors. Modified Cognitive Behavior Therapy (MCBT) techniques Multimodal program, Exposure Response prevention, have been proven successful for people with ADHD and OCD respectively. The present study aims at finding the impact of these techniques in the treatment of co-morbid behaviors (attention deficit, obsessiveness, repetitive-stereotypical behavior) in ASD. Twelve children (Age: Mean=6.25, SD=2.22) diagnosed with ASD selected, Pre-test conducted to assess co-morbid behaviors with ADHD Scale, BASIC-MR Behavioral Scale, Vineland Social Maturity Scale and Attention span (calculated time spent on a particular task), found no difference. Planned intervention with MCBT for 6 participants (experimental group), the rest 6 participants (control group) underwent traditional occupational, speech therapy. Total intervention is for a period of 16 weeks (45 minutes per day). End of 16th week post-test assessed a significant improvement in behaviors. Concludes effectiveness of MCBT ( $p < 0.001$ ).

# Can Articulating Aloud Offset Effects of Listening to L1 in a Foreign Accent?

**Kit Cho**

The University at Albany, SUNY

**Laurie Beth Feldman**

The University at Albany, SUNY & Haskins Labs

**Abstract:** The production effect refers to the finding that words that are read aloud are more distinctive and memorable relative to words that are read silently. The present study explored the production effect using auditory stimuli, spoken in two different accents. American participants repeated a word or listened passively to words spoken in either their native accent (English) or a foreign accent (Chinese). In both recall and recognition, memory was better for words that were read aloud rather than listened to, thereby extending the production effect from the visual to the auditory domain. However, the benefit to recognition associated with repeating a word, rather than listening to it, was greater in a native accent.

Evidently, the gestures of articulating a word aloud benefit memory but its impact is attenuated by phonetic mismatch between what participants hear and what they produce.



# Blending Narrative Spaces of the Flashback Scenes in the Joint Security Area

**Hye Rhang Cho**

Sogang University

**Seung Suk Nam**

Sogang University

**Sook Whan Cho**

Sogang University

**Abstract:** The film, Joint Security Area (2000), centers on a cross-border shooting incident in the flashpoint of North-South tensions. As Major Sophie Jang's investigation develops, she uncovers evidence suggesting that neither account from North and South is correct. With the use of extended flashbacks, the truth about the incident, as well as the unlikely connection between the North-South Korean soldiers, discloses. This paper is intended to propose a model of narrative blending associated with the mapping between the investigators and the suspects spaces in terms of intentionality and analogy (Fauconnier and Turner, 2002). It will be seen that the contradicting accounts of the soldiers are associated with their intentions, involving memory and fear, this input spaces of soldiers interacting with another input space of Jang. As the relations of analogy and intentionality are exhibited in the input spaces and compressed, the friendship between the soldiers reveals in the blended space.

# Blending Narrative Spaces of the Reenactment Scenes in the Thin Blue Line

**Sook Whan Cho**

Sogang University

**Seung Suk Nam**

Sogang University

**Hye Rhang Cho**

Sogang University

**Abstract:** In *The Thin Blue Line*, the story of a man (Randall Adams) convicted and sentenced to die for a murder that he did not commit, Morris (1988) presents a series of interviews demonstrating the investigations, creating reenactment scenes of the shooting built from the testimony and recollections of the detectives, suspects (David Harris), and witnesses involved in the murder. This paper is intended to propose a model of narrative blending associated with the mapping between the investigator and the suspect spaces in terms of intentionality (Fauconnier and Turner, 2002). It will be seen that Harris's failure to tell the truth explicitly is associated with his intentions and mental attitudes involving desire and fear; this input space of Harris's deceiving intention interacting with another input space depicting his criminal record and violent traits. As his intention is revealed in these space interactions, Harris is found guilty in the blended space.

# **Ten-month-old Infants Prefer Comforters, not Helpers**

**Hiu-Mei Chow**

Department of Psychology, The University of Hong Kong

**Stephanie Jui-chi Chen**

Department of Psychology, McGill University

**Geroldene Tsui**

Department of Psychology, The University of Hong Kong

**Ping-Hui Chiu**

Department of Psychology, The University of Hong Kong

**Chia-huei Tseng**

Department of Psychology, The University of Hong Kong

**Abstract:** Preverbal infants prefer characters that help others in achieving a goal (e.g. climbing up a hill), but how this preference is established remains unknown. In current study, we examined physical and emotional outcome hypotheses by presenting 6- and 10-month-old infants a social interaction similar to what was used in Hamlin et al. (2007). During the learning habituation stage, an agent emotionally comforted or upset a climber by pushing it up (helped) or down (hindered) the hill (physical outcome). On a new platform, we found that 10-month-old infants looked significantly longer when the climber approached the comforter who previously made the climber laugh regardless of what physical outcome was. This indicates infants prioritization of emotional over physical outcome and their consideration of a third partys internal state in forming a social preference, which was absent at 6-month-olds. This leads to the conclusion that this prioritization is unlikely to be innate.

# Comprehension of Representational Gestures

Kawai Chui

National Chengchi University

**Abstract:** In daily communication, the spontaneous use of hands and arms along with speech is pervasive and indispensable. The present study used real face-to-face communication materials as stimuli to investigate the comprehension of representational gestures when speech is not available. Twenty-two adults watched short, soundless video clips extracted from recordings of daily conversations, each including a spontaneous representational gesture. Participants were requested to judge whether and in what way the gestures made sense. Their responses showed that in the absence of speech, the idiosyncratic hand configurations were not incomprehensible, suggesting that speech-gesture integration is not entirely obligatory. The way representational gestures were understood in this study reveals the activated content of the gestural-action representations which consists of conceptual knowledge associated with a situation of use.

# Task-specific conflict monitoring and cognitive control in the prefrontal cortex

**Chongwook Chung**

KAIST, Daejeon, KOREA

**Chobok Kim**

Kyungpook National University, Daegu, KOREA

**Jeounghoon Kim**

KAIST, Daejeon, KOREA

**Abstract:** According to the conflict monitoring theory, the dorsal anterior cingulate cortex (dACC) and the dorsolateral prefrontal cortex (DLPFC) are involved in detecting (i.e., monitor) and regulating (i.e., controller) the conflict, respectively. However, it has been unknown whether different types of response conflicts recruit the same monitor-controller loop. To address this issue, we developed a double-response-conflict task using color-based and location-based conflict stimuli. This task evoked two types of conflict at the response level either simultaneously or separately, and we tested neural activities in dACC and DLPFC employing the conflict adaptation paradigm. Along with behavioral task-specific conflict adaptation effects, imaging results showed task-specific conflict adaptation in dACC and DLPFC. The results also demonstrated that double-conflict processing appears to be qualitatively different from single-conflicts despite the fact that the sources of double-conflict were overlapped with two single-conflict conditions. These suggest that response conflicts from two different sources are independently processed.

# **The Development of Orthographic Awareness for Radical Properties of Chinese Characters in Young Children**

**Yi-Ling Chung**

National Cheng Kung University

**Pei-Yu Luo**

National Cheng Kung University

**Hsueh-Chih Chen**

National Taiwan Normal University

**Li-Yun Chang**

University of Pittsburgh

**Jon-Fan Hu**

National Cheng Kung University

**Abstract:** Radical properties are thought as important roles in Chinese character perception. Understanding radical knowledge is useful for learning Chinese vocabulary. However, the previous results investigating radical frequency or radical position showed inconsistent conclusions. Therefore, the present study focuses on manipulating radical position-based frequency and character-like degrees to examine the effects of the regularity of the two properties on Chinese character recognition. 210 Chinese-speaking elementary students were asked to judge if one of the two words are more likely to be legal word from 123 paired Chinese pseudo-words. The results indicates that: (a) there is a radical position regularity effect on orthographic awareness, which means the accuracy of high frequency radical is better than low frequency; (b) children can identify radical position regularity; (c) the ability of radical position awareness turns into stable through the experience of character recognition growing.

Keywords: orthographic awareness, radical, position/frequency regularity, Chinese learning

# The Dynamics of Sentence (In)comprehension

**Gregory Cox**

Indiana University

**Melody Dye**

Indiana University

**Seth Frey**

Indiana University

**Abstract:** Comprehension of text develops over time, such that a single word may lead to the activation of different concepts and thus different expectations at different times after it is encountered. Since the pioneering work of Neely (1977), much evidence has been found that multiple senses of a word are activated shortly after it is encountered, but that senses that are contextually irrelevant are later inhibited. This has implications for the processing of anomalies in sentences, including malapropisms like switching antidote for anecdote: if the anomaly is primed by the semantic context (e.g., it is preceded by the word doctor) shortly before it is encountered, it should be less likely to be noticed by the reader and is less likely to impede reading (and comprehension) of the sentence. We present results from a reading-time study using sentences in which malapropisms are presented in either neutral or semantically congruent contexts, with the semantic prime at varying distances from the anomaly (e.g., doctor, above). These results are interpreted within a dynamic sentence comprehension model based on the attractor model of Tabor, Juliano, & Tanenhaus (1997).



# **Aquisition of multipliers: learning a new class of number**

**Meghan Dale**

University of Waterloo

**Mathieu Le Corre**

University of Waterloo

**Abstract:** Acquisition of number words has been examined in great detail for the count list of one through nine, but rarely do multi-digit numbers receive any study (Carey, 2009; Fuson et al., 2007). In English, once we pass into triple digits, we encounter a new class of number: multipliers. Multipliers are characterized by the unique syntactic structure of number + multiplier + noun(pl). 27 children aged 4 to 6 years old were taught a novel word (gobi) using the multiplier structure. After being trained to identify one gobi objects, children were tested on their ability to generalize gobi in four additional contexts. Participants fell into one of two distinct groups that either consistently passed or failed all tasks ( $t(25) = 13.406$ ,  $p < 0.001$ ). We conclude that prior to formal education children are able to identify the multiplier structure with minimal exposure and discuss implications for childrens acquisition of multi-digit numbers.

# Describing faces: conventionalizing ontologies through dialogic interaction

**Nicolas Davidenko**

Stanford University

**Gregory Mills**

University of Edinburgh

**Abstract:** Although we are highly efficient at remembering and recognizing faces, we find it remarkably difficult to describe faces to others. This stems in part from the holistic nature of face representation and the difficulty of expressing configural information in words. However, research on dialogic interaction has shown that when interlocutors encounter a domain for which they possess only vague descriptions, they rapidly conventionalize ad-hoc ontologies that enable efficient and systematic communication. To test the hypothesis that interlocutors will conventionalize ad-hoc ontologies for describing faces, we present data from a collaborative, computer-based task which is played by pairs of participants. Successful completion of the task requires participants to interactively describe and identify target faces from a set of distractors. The data provide evidence of dyads conventionalizing ontologies for referring to faces. Drawing on the observed interactions, we demonstrate how these conventions emerge as a consequence of participants resolving instances of miscommunication.

# **The effect of speakers identity on syntactic processing: Evidence from verb-gender agreement in Slovak**

**Doug J. Davidson**

Basque Center on Cognition, Brain and Language

**Adriana Hanulikova**

Basque Center on Cognition, Brain and Language

**Manuel Carreiras**

Basque Center on Cognition, Brain and Language

**Abstract:** An important property of speech is that it explicitly conveys speakers identity. Although previous studies have shown that speakers identity affects semantic processing, the influence of speakers identity on grammatical processing is less clear. Here we investigate subject-verb agreement in Slovak, when the agreement depends on the speakers gender (as cued by his/her voice) compared to when it depends on the formal grammatical gender of the subject.

We compared ERP responses to Slovak verbs disagreeing with the formally-marked subjects gender (e.g., (the) mother-in-lawFEM \*stoleMASC plums), and to verbs disagreeing with subjects gender as conveyed by the speakers voice (e.g., IFEM \*stoleMASC plums).

The formally-marked subject and verb disagreement resulted in a P600 preceded by an anterior negativity. However, disagreement based on the speakers voice elicited a posterior negativity but no P600. We will discuss differences in checking or repair mechanisms of these two agreement types.

# Joint Evaluation and Trend Information Mitigates the Disposition Effect

**Kyung Soo Do**

Sungkyunkwan University

**Hee-Yeon Kim**

Sungkyunkwan University

**Rae-yeop Park**

Sungkyunkwan University

**Abstract:** Disposition effect refers to the phenomena where investors hold loser stocks and sell winner stocks. Effects of information about the trend on the disposition effect were tested in two experiments, where participants decided whether they sell or not, given information about the performances of one (separate evaluation, SE) or two stocks (joint evaluation, JE). When only the prices of the previous month and the current month were given in Experiment 1, the percentage of selling the losing stock in JE was twice larger than that in SE. The differences in selling the losing stock between SE and JE disappeared when the monthly prices of previous six months were given in Experiment 2. Results of the two experiments showed that the disposition effect got weaker when the trend of loss was apparent either by knowing the losing performance over many stocks or over many months.

# Effects of Scaffolded Feedback and Confidence of Incorrect Answers on Retention

**Kyung Soo Do**

Sungkyunkwan University

**Hanna Kim**

Sungkyunkwan University

**Abstract:** The effects of scaffolded feedback and confidence of incorrect answers were tested in two experiments. As feedback imposes cognitive load, an effective feedback should help reducing the cognitive load. One possible way of reducing the load is using scaffolded feedback, in which hints are incrementally provided until participants generate the correct answer. Scaffolded feedback promoted learning and reduced the cognitive load more than the corrective feedback in Experiment 1, in which knowledge of Korean history was tested. The advantage of the scaffolded feedback disappeared when text comprehension was tested in Experiment 2. However, errors with high confidence are corrected more than errors with low confidence in both experiments. The results of the two experiments show that the scaffolded feedback promotes learning only when the cognitive load of the target task is not large.

# Computing Humorous Metaphors

**Pawel Dybala**

JSPS Research Fellow / Otaru University of Commerce

**Kohichi Sayama**

Otaru University of Commerce

**Abstract:** It was experimentally showed that humorous texts can be explained using approaches applied to metaphors, such as the salience-imbalance or the domain-interaction approach. It was also demonstrated that humorous metaphors often include a switch between positive and negative emotions.

We propose to construct a computer system able to understand and generate such metaphors. Currently we are constructing a metaphor conceptual network, in which links between concepts are calculated accordingly to their roles in metaphor understanding. This will allow the computer to process metaphors. Next, we will adjust the links distance calculation to match the humorous metaphors (to increase the salience-imbalance, as demonstrated in existing research). We will also use an emotiveness-recognition-system to detect emotive associations towards particular phrases, in order to choose pairs with the emotional switch.

The system will be evaluated in user-oriented experiments.

Acknowledgements: This work was supported by KAKENHI (Project Number: 23-01348)

# Expanding the Transformation Paths in a Mutual Transformation Mechanism of Music and Narrative

**Jun Endo**

Iwate Prefectural University, Iwate, Japan

**Taisuke Akimoto**

Iwate Prefectural University, Iwate, Japan

**Takashi Ogata**

Iwate Prefectural University, Iwate, Japan

**Abstract:** Music is an important part in the narrative generation system we have been studying, but is not merely the accompaniment of narrative. For the integration of the knowledge and techniques of narrative and music, we developed some versions of the automatic mechanisms which mutually and cyclically transform narrative structures and musical structures using narratological and musical methods. A previous prototyping system consists of various transformation modules: a music composition mechanism from a narrative or a story, a music variation mechanism from an original music, a transformation mechanism from variation music to the narrative discourse, etc. This paper's goal is to add other possible circulation paths to expand the system framework. Specifically, we implement a new original music generation mechanism by the re-formation of the original music anew and add a transformation mechanism from variation music to the new original music. Moreover, we are evaluating the corresponding relationship between narrative and music, and the quality of generated music.



# Eye tracking differences in respondent behaviour across multiple survey modes

**Tom Foulsham**

University of Essex

**Olena Kaminska**

University of Essex

**Abstract:** Research into survey methodology has revealed that respondents answer differently depending on the way in which questions are presented. To date there have been few empirical attempts to relate these differences to respondents cognition. To investigate this, we used eye tracking in three different survey modes: a computer-based mode, a pen-and-paper questionnaire and a face-to-face interview with show-cards. This novel method of obtaining cognitive measures across modes means that we are able to investigate attention and cognitive processing by coding which questions and answers are looked at the most in the different settings. The results confirm the attentional biases linked to some responses, such as a bias to the first and last response options in a list, as well as revealing differences in how response options were inspected across modes. These results are discussed and it is proposed that cognitive science can play a significant role in survey methodology.

# **Automatic facilitation of social behavior by implicit inferring of social intention**

**Haruaki Fukuda**

Aoyama Gakuin University

**Hiroaki Suzuki**

Aoyama Gakuin University

**Ayumi Yamada**

Aoyama Gakuin University

**Abstract:** It is known that we can infer others intentions and goals automatically. On the other hand, recent researches have demonstrated that people spontaneously adopt and pursue the goals perceived in others behavior. This phenomenon is called goal contagion and goal contagion can occur even when the goal is not consciously understood. In this study, we examined whether mere exposure to animated agents intentions could influence our social behavior. Participants performed social game while the short animation depicting social intention in the peripheral monitor that participants did not attend to. We found that participants exposed to an animation that implied helping behavior was more cooperative than participants exposed to an animation that implied hindering behavior. In addition, participants could not answer what was displayed in the peripheral monitor in which the animation was displayed. These results suggest that others intentions can influence our social behavior without consciousness.

# Quantifying linguistic coordination

**Riccardo Fusaroli**

Aarhus University

**Kristian Tyn**

Aarhus University

**Abstract:** Language has been defined as a social coordination device (Clark 1996) enabling innovative modalities of joint action. However, the exact coordinative dynamics over time and their effects are still insufficiently investigated and quantified. Relying on the data produced in a collective decision task (Bahrami et al 2010, Fusaroli et al. 2012) we extend to linguistic coordination dynamical measures of recurrence employed in the analysis of sensorimotor coordination (such as heart-rate (Konvalinka et al 2011), postural sway (Shockley 2005) and eye-movements (Dale, Richardson and Kirkham 2012). We employ nominal recurrence analysis (Orsucci et al 2005, Dale et al 2011) on the decision-making conversations between the participants. We report strong correlations between various indexes of recurrence and collective performance. We argue this method allows us to quantify the qualities of linguistic coordination and their effects at a fine-degree.

# How a Quantum Approach to Memory Incorporates Contextuality and Potentiality

**Liane Gabora**

University of British Columbia

**Kirsty Kitto**

Queensland University of Technology

**Abstract:** Existing models of memory incorporate context in only a limited sense, and they fail to accommodate the complementary notion of potentiality. Humans use context in a multitude of ways: (1) to resolve ambiguity between alternative pre-existing meanings, (2) to emphasize particular aspects of something, (3) to identify new and potentially meaningful relationships, and (4) to clarify an idea such that ones understanding of it goes from ill-defined to well-defined. A powerful model of memory must capture all four. Drawing on elements of vector based information retrieval models, Grdenfors geometric model of conceptual space, and matrix model of memory, we present a quantum approach to memory, and show how it can capture these four senses of contextuality. Though in its infancy, the quantum approach can provide not just exceptionally high-density memory storage but creative capacities as well.

# The Naturalization of Concepts between Computational Intractability and Cognitive Theories

Francesco Gagliardi

University of Rome "La Sapienza"

**Abstract:** In this work we present some computational considerations about the nature of concepts. After a first distinction between "per se" concepts and the epistemic or psychological aspects of concepts we focus our attention on this latter by linking their understanding to the formalization of automatic classification problems as developed in machine learning field.

We show how the problem of categorization and of formation of concepts representing categories, has to be considered a computationally intractable problem that hence can be faced only with heuristic strategies. In this perspective the different cognitive theories proposed in cognitive science to explain categorization processes (e.g. the prototype-theory by E. Rosch), can be considered as different heuristic strategies (computationally tractable solutions) to face the categorization problems and that the human mind uses in order to prevent an unsustainable cognitive-computational load.

Finally, we frame these ideas in the cognitive naturalism and in the current viewpoint that considers the most of the human reasoning as heuristic solutions to intractable problems. Our thesis is that concepts can be considered as heuristic and perspective solutions that any intelligent system, such as human mind, finds in order to represent categories using limited resources and capacities by which it organizes and gives a sense to the large variety of reality that surrounds it.

# Selection of decision rules in dynamic decision making

**Jean-Francois Gagnon**

Universit Laval

**Marie-ve St-Louis**

Universit Laval

**Sebastien Tremblay**

Universit Laval

**Abstract:** This study investigates how dynamic complexity influences the use of heuristics in a dynamic decision making task. 24 subjects performed a computerized task that required them to predict the dynamic change ( $t+1$ ) of discrete interrelated variables (range 0-20). There were four experimental sessions that involved the co-evolution of eight variables over 40 time periods. The level of complexity of the dynamic change of each variable was determined by its distance to linearity. Five heuristics, characterized by different methods of effort-reduction (Shah & Oppenheimer, 2008), were formalized. Forecasting performance of participants was compared against each heuristic. The results revealed that subjects adapt their heuristics to the level of dynamic complexity. Our findings are discussed in relation to the conditions under which individuals adapt decision rules in complex dynamic tasks and on the consequences of such heuristic changes (Hogarth & Karelaia, 2007).

# Annual Cognitive Modeling Competition

**Kevin Gluck**

Air Force Research Laboratory

**Abstract:** Although the cognitive and behavioral modeling communities now have a rather lengthy history and there is ongoing research and development in these areas across dozens of academic, industrial, and government research laboratories, very little of the work has been explicitly competitive. That said, there are precedents for modeling competitions in this area. Two examples include the PokerBot Competition and the Dynamic Stocks and Flows Model Comparison Challenge. Both of these were successful and interesting events. They were also both single shot modeling competitions that did not evolve into annual events in the spirit of something like Robocup or the NIST-sponsored automated speech recognition competitions. It is time to consider a recurring annual cognitive modeling competition. This poster is an opportunity for community discussion of the pros, cons, infrastructure requirements, and design parameters that should be considered in developing such an event within cognitive science.



# Do Young Children Habituate to their Classroom Environment?

**Karrie Godwin**

Carnegie Mellon University

**Anna Fisher**

Carnegie Mellon University

**Abstract:** Prior research has shown that visual features of the classroom environment (e.g., charts, posters) can be potential sources of distraction hindering children's ability to attend to the content of a lesson. However, it is possible that over time children may habituate to their classroom environment. To address this possibility we investigated whether more prolonged exposure in a high visual distraction learning environment would result in habituation. We presented 23 kindergartners with 10 science lessons over 2-weeks. A habituation effect was observed as evidenced by a significant decrease in the proportion of time children attended to the environment between week 1 and week 2 ( $p=0.001$ ). However, the proportion of time children attended to the environment at week 2 remained significantly above baseline levels ( $p<0.0001$ ). Thus over time children began to habituate to the classroom environment; however, the environment remained a significant source of distraction for young children.

# Phonological neighbourhood clustering effects on verbal short-term memory

Winston Goh

National University of Singapore

**Abstract:** Phonological similarity between words as measured by neighbourhood density counts the number of words differing from the target word by a single phoneme. The word cat has a denser neighbourhood (e.g. hat, cut, at, scat) than the word wag (e.g. bag, wan). However, density does not capture between-neighbour relationships (hat and at are also neighbours of each other, but bag and wan are not). A recent index called the clustering coefficient measures the proportion of neighbours of a word that are also neighbours of each other. In an immediate serial recall task of 6-word lists using non-repeated sampling, lists of words with high clustering coefficients (i.e. many neighbours were neighbours of each other) were better remembered than those with low clustering coefficients (i.e. few neighbours were neighbours of each other). The findings suggest that a network of overlapping similarities among a to-be-remembered words phonological neighbourhood enhances recall in short-term memory.

# Children's sensitivity to informant's inductive efficiency and learner's epistemic states in pedagogical contexts

**Hyowon Gweon**  
MIT

**Patrick Shafto**  
University of Louisville

**Josh Tenenbaum**  
MIT

**Laura Schulz**  
MIT

**Abstract:** Children use properties of information and epistemic states of others to socially evaluate informants in pedagogical contexts. Recent studies have shown that in pedagogical contexts, children expect teachers to provide information to support accurate inference. Do children simply prefer teachers who provide more information, or do they rationally expect teachers to provide the right amount of information? Here we present a computational model of social learning that incorporates the cost of acquiring information and a series of experiments with children (5- and 6-year-olds) to test its predictions. Children observed two teachers who either provided a minimal amount of information or exhaustive information about a multi-function toy to a nave learner. Consistent with the predictions of the model, childrens preferences rationally reflected the informativeness of the teachers demonstrations, the cost of observing the demonstrations, and even the learners epistemic state. Taken together, the results suggest that learners can use information provided by others to not only learn about the target object but also to rationally evaluate others in order to guide their future choices of informants.

# The End is Near: Anticipating the end of a sentence

Lance Hahn

Western Kentucky University

**Abstract:** Hahn (2011) demonstrated that anticipating a final word from local context benefitted from the knowledge that the sentence was ending. In this work, I hypothesized that statistical regularities of text usage would show a building anticipation of the sentences end as the ending word sequence is acquired. A text corpus analysis was conducted focusing on 570 common and 570 uncommon sentence endings. A single word ( $w_1$ ) followed by three unspecified intervening words generated a weak average anticipation ( $p(w_5 = \cdot \mid w_1 \text{ ? ? ?}) = 0.1$ ) that steadily increased with the accumulation of subsequent words in the ending sequence. The complete four-word context generated a strong average anticipation ( $p(w_5 = \cdot \mid w_1 w_2 w_3 w_4) = 0.8$ ) of the subsequent period that renders the period fairly redundant.

# **Individual differences and phonetic aptitude in the earliest stages of L2 acquisition**

**Adriana Hanulikova**

Basque Center on Cognition, Brain and Language

**Dan Dediu**

Max Planck Institute for Psycholinguistics

**Zhou Fang**

Donders Institute for Brain, Cognition and Behaviour

**Jana Basnakova**

Slovak Academy of Sciences

**Falk Huettig**

Max Planck Institute for Psycholinguistics

**Abstract:** Most learners of a foreign language (L2) struggle to acquire a native-like accent. We explored how a training study of learning consonant clusters at the very onset of the L2 acquisition can inform us about L2 learning in general and individual differences in particular. Dutch native (L1) speakers were trained on Slovak words with consonant clusters. In the session following training, participants were tested on a battery of L2 perception and production tasks. The battery of tests was repeated twice more with one week between each session. An additional battery of control tests was used to test participants L1 skills. Participants showed considerable individual differences across all L2 tasks, stable across sessions. Two participants showed L2 production performance that fell within 2 standard deviations of the mean ratings obtained for an L1 speaker. The mispronunciation detection task was the only perception task, which significantly predicted production performance in the final session.

# What do numbers tell you? The effects of data-presentation design on the prolonged use of health-related life-log tools

**Etsuko T. Harada**

University of Tsukuba

**Satomi Yoshiyama**

Hosei University

**Abstract:** Although a number of life-log tools for health care have been proposed, they share a common problem; the drop-out problem of achieving long-term continuous usage in users. Providing numerical data is believed to be useful in tackling the problem, but, that can also have adverse effects if people overly react or misinterpret the meaning of the data. To investigate how people react to numerical data, we asked 817 undergraduate students to evaluate some fictitious tools for four domains (e.g., blood pressure) with one of five different designs; with standard or precise figures (with two additional digits), combined with/without interpretive messages for the data (e.g. good condition), or no numerical data (message only). The results indicate that involvement in health-care activities is independent of tool evaluation, such that only tool evaluations were strongly influenced by the existence of numerical data. The merits and demerits of numerical presentation are discussed.

# Japanese script types in written names create the images of their referents

**Aya Hatano**

Nagoya University

**Masahiro Amagase**

Nara Women's University

**Jun Kawaguchi**

Nagoya University

**Abstract:** This study investigated how the unknown objects names written in different Japanese script types (Hiragana, Katakana, and Kanji) would create the images of their referents. Previous studies suggested that writers would creatively use the script types for communicating emotional semantic information to readers. Although there are various factors affecting images of referents, how the factors have effects on them has not been investigated sufficiently. We presented Japanese undergraduates with 10 names of ethnic foods unknown to them. The participants read each name written in one of three script types and described the images of its referent. Their descriptions were analyzed by KH Coder (a freeware for text mining). The result shows the images of referents are influenced by the Japanese writing system, general visual features of each script type, and phonological features of words.



# **Is past information useful for evaluating present covariation information? : Effect of irrelevant information on causal judgment**

**Ikuko Hattori**  
Ritsumeikan University

**Masaki Hattori**  
Ritsumeikan University

**Abstract:** This study examined how information about the past affects the estimation of present covariation in causal judgment. For example, although lottery stands advertise that they sold the tickets of a big win last year, and buyers tend to buy tickets there, winning the prize this year has no actual connection with past win. However, information about the past has a psychological influence on people in causal judgment. In this experiment, participants observed this year's records of fictitious baseball players and estimated their contract money for the next season. Participants were assigned to either powerful-team or weak-team condition, and were told that their team had finished among the best three or the worst three for ten years. Weak-team group tended to significantly discount the contribution by the player concerned. The result implies that past information should be concerned with dealing the non-cause cases in which the target cause does not occur.

# Visual cognition of "speed lines" in comics: Experimental study on speed perception

**Hiromasa Hayashi**

University of Tokyo, Meguro-ku, Tokyo, Japan

**Goh Matsuda**

University of Tokyo, Meguro-ku, Tokyo, Japan

**Yoshiyuki Tamamiya**

University of Tokyo, Meguro-ku, Tokyo, Japan

**Kazuo Hiraki**

University of Tokyo, Meguro-ku, Tokyo, Japan

**Abstract:** "Speed lines" is the abstract lines used in comic to make objects look like they're moving. The number and the length of speed lines are manipulated by comic artists to match the speed they want to express, but few studies have tested their effect on perception of motion of the viewer. To investigate the matter, we conducted an experiment employing prediction-motion task. 19 participants were shown an apparent motion of a ball moving in uniform linear motion, and were told to estimate the speed of the ball. Each ball was presented with one of six patterns of speed lines, 3(number) x 2(length). As result, we found that both the number and the length of speed lines have significant effect on subjective speed of the ball. This indicates that the number and the length of speed lines indeed affects viewer's perception of motion, thus plays significant role in comics.

# Ontology Architecture of a Neuro-psychoanalytical, Computational Model

Isabella Hinterleitner

ICT

**Abstract:** The paper introduces the ontological knowledge architecture of the psychoanalytical, hierarchical and computational model ARS (Artificial Recognition System). The ontology architecture of the model is represented by a knowledge base. Semantic and individual memories are realized by means of an ontological architecture. These memories are modeled in a triple notation by the knowledge representation languages Resource Description Framework (RDF) and Resource Description Framework Schema (RDFS). Additionally, rules are employed for reasoning in the semantic knowledge base (manipulation rules) and individual knowledge base (action rules). The approach uses RDF and RDFS for the representation of metadata describing the semantic and individual memories. The simulation shows that general concepts are needed to form semantic memory as such. In addition, individual instanced memory is needed for forming individual memory. In conclusion, the ontological approach can be further transferred to the field of building automation by representing domain-specific ontologies (individual) and upper ontologies (semantic).

# **The absence of phonetic symbolism to the novel speech sound -comparison of cross-modal correspondence between Chinese and Japanese speakers using Chinese speech sound-**

**Sachiko Hirata**

The University of Tokyo

**Shinichi Kita**

Kobe University

**Abstract:** Phonetic symbolism is a phenomenon that speech sound evokes images based on sensory experiences and thus often discussed with the similarity for cross-modal correspondence. Hirata and Kita (2012) showed cross-modal congruence between brightness/darkness and Chinese speech sound with and without aspiration in Chinese speakers by using Garner's task. In the present study, we examined whether Japanese speakers, which do not have any Chinese knowledge, show cross-modal correspondence to the Chinese speech sound or not. We conducted the same experiment as previous research to Japanese speakers with no Chinese experience. As a result, Chinese speech sounds with aspiration, which resemble voiceless consonants, were not matched with brightness, whereas those without aspiration, which resemble voiced consonants, were not matched with darkness. This result is different to its pattern in Chinese speaker and consequently suggests that phonetic symbolism is affected by the knowledge of the phonemes of its language.

# Longitudinal observation of action slips: A case study of a young child

Naoya Hirose

Kyoto Notre Dame University

**Abstract:** The purpose of this study was to investigate action slips during acquisition of a new skill. Whereas most studies on slips have collected action errors of adults using the diary method, we observed a young child's daily performance using video recordings. One child's mouse rinsing activities after toothbrushing were recorded for one year from age four to five, and analyzed in terms of action slips as well as microslips—miniature versions of slips. Slips were repeatedly observed when a child took a cup and filled it with water. Quarterly analysis reveals that the total number of slips did not decrease monotonously: whereas two types of slips, commissions and adjustments, decreased over one year, other types increased from Q1 to Q3, then decreased in Q4. This suggests that in the process of the skill acquisition some types of slip may fade out, but others may rise or last out.

# **The role of social contexts in adults' word learning**

**Masako Hirotani**

Carleton University

**Koji Shimada**

National Institute for Physiological Sciences

**Shuntaro Okazaki**

National Institute for Physiological Sciences

**Hiroki C. Tanabe**

National Institute for Physiological Sciences

**Norihiro Sadato**

National Institute for Physiological Sciences

**Abstract:** While numerous studies report the significance of social contexts in infants word learning, data concerning adults are scarce. This study aims to fill this gap. Specifically, it investigated the extent to which a number of the parameters associated with joint attention (e.g., direction and timing of eye-gaze between learners and target objects) help adults learn new words (i.e., map new objects with new labels). The study adopted a novel experimental paradigm where live interactions between two learners took place in four different learning contexts. An eye-tracker recorded learners eye-gaze during the experiment. Adult learners learned new words most successfully when they shared both the direction and timing of their eye-gaze towards the target objects. In addition, word learning initiated by a learner alone, without sharing eye-gaze with another, failed to match these results. Overall, this study supports the view that joint attention maximizes adults learning of new words.

# **A Cognitive-Educational Approach to the Verb mek-ta in Korean**

**Do-Il Hong**

Hankuk University of Foreign Studies

**Hee-Rahk Chae**

Hankuk University of Foreign Studies

**Abstract:** This study tries to analyze the basic and extended meanings of the Korean verb mek-ta 'to eat' from a cognitive point of view. The most basic meaning of the verb is 'to pass food and others through the mouth into the stomach'. This meaning has been extended to other intransitive and transitive verb meanings and auxiliary verb meanings. When presenting the verb to learners of Korean as a Foreign Language (KFL), it should be presented so that learners themselves can infer the extended meanings using a semantic network. To this end, uses of those words meaning 'to eat' in the native languages of KFL learners should be investigated, and common meanings with the Korean verb should be identified. Then, we can create a semantic network of the senses of 'to eat', a network of the basic meanings and extended meanings of mek-ta in Korean and corresponding verbs in other languages.



# Connectionist Modeling of Frequency and Regularity in Mandarin Relative Clause Processing

**Yaling Hsiao**

University of Wisconsin-Madison, Madison, Wisconsin, United States

**Maryellen MacDonald**

University of Wisconsin-Madison

**Abstract:** Studies of Mandarin relative clause processing have yielded mixed results. Subject relatives (SRCs) are found more frequent than object relatives (ORCs) in corpora; however, the word order of ORCs resembles canonical SVO simple sentences. The present study examined the experienced-based effect of structural frequency and regularity by conducting a simulation in a simple recurrent network, trained on 10000 simple and RC sentences, in the proportions found in Chinese Treebank 7.0. The model contained 18 input and output units and 36 hidden and context units. Network performance was assessed on the level of grammatical prediction error of the next word. Following 10000 training epochs, ORCs performance showed garden-pathing upon encountering disambiguating relativizer due to resemblance to the SVO order in the beginning. For SRCs, which have an irregular verb-first structure, error was high but dropped later at the relativizer, showing sensitivity to language statistics. Results reflect both frequency and regularity effects.

# A Corpus Survey of Chinese Individual Classifiers

**Shuping Huang**

National Sun Yat-sen University, Kaohsiung, Taiwan

**Jenn-Yeu Chen**

National Cheng Kung University

**Abstract:** As of today, there are scarce studies that show us how classifiers are used in real life. While information technology progresses, today's researchers are privileged to use large-size databases to explore the real usage of classifiers. We selected 386 concrete nouns in Chinese, and looked them up in Chinese GigaWord for their collocating classifiers. Seventy individual classifiers were found to pair up with these 386 nouns. The following questions are of our particular interest: (1) What are the most frequent classifiers in Chinese? (2) On average, what is the number of classifiers paired up with a noun? (3) What is the number of nouns paired up with a classifier? (4) Which semantic dimensions are particularly relevant to the Chinese classifier system?

In our data, a noun on average pairs up with 2.839 classifiers, and a classifier category generally contains 16 nouns. The most frequent classifier was *zhi1* for animate beings. The noun *di4tu2*, taking 8 classifiers, is the most flexible noun in our corpus. By corpus data we will also talk about the role of *ge0*. Concerning the semantic dimensions, Tai (1994) identified four dimensions manifested in Chinese classifier system: animacy, shape, size, and consistency. And he also found a number of miscellaneous cases that reflect part-whole or metonymic relation. Our investigations show that some classifiers are not highly generalized, such as *zun1* (for statues), *chuang2* (for buildings), *sao4* (for vehicles), etc. Those cases hint at the possibility that some classifier categories operate on low-level classification such as VEHICLE and BUILDING, instead of highly-abstract semantic dimensions.

# Property activation during the Interpretation of Noun-Noun Compounds

Jie Huang

Huazhong University of Science and Technology

**Abstract:** Noun-noun compounding is the most productive way of word formation in many languages. The past two decades have witnessed highly diversified studies in noun-noun compounds from perspectives of linguistics, psychology, computer science and so on (Sweetser 1999; Gagne 2002; Devereux & Costello 2005; Butnariu & Veale 2008; Domnguez 2009). In this paper, we carry out three experiments to examine the process of property activation during the interpretation of noun-noun compounds. The first experiment aims at testifying whether we are more inclined to interpret a novel noun-noun compound with reference to analytic or perceptual properties of the referent. The second experiment tests participants preference for metaphorical or relational association in interpreting novel noun-noun compounds. The third experiment testifies whether cognitive development has any impact on noun-noun compound interpretation. Results from the experiments indicate that cognitive factors, such as encyclopedic knowledge, conceptual similarities and contiguities, cognitive development, influence the interpretation of novel noun-noun compounds.

# Modeling How Naming Experiences Bias Sorting Performance

**Yu-Sheng Hung**

National Cheng Kung University

**Yun Li**

National Cheng Kung University

**Hsueh-Chih Chen**

National Taiwan Normal University

**Jon-Fan Hu**

National Cheng Kung University

**Abstract:** Previous studies assume that language specificity might affect naming patterns, whereas perceptual similarity based sorting might be universal. Study 1 was aimed to examine the hypothesis, but it was found that the correlations of sorting based on perceptual similarity between Chinese and Dutch (or French) are surprisingly lower than previous findings. We argue that perceptual similarity might be also affected by naming experiences. In study 2, we adopted backpropagation neural networks to model the sorting patterns observed in Study 1. During training, the inputs of models were composed of the numeric vectors of names and semantic features. Nevertheless, only semantic features were fed to networks for simulating the sorting results. After training, the simulation outcome of networks showed high correlation with the experimental results. These studies suggest that we should reconsider the involvement of naming experiences for the sorting processes.

Keywords: naming, sorting, perceptual similarity, backpropagation neural networks

# **The Effect of Priming of Individualism/Collectivism on the Miller-Lyer Illusion**

**Yiwon Hyun**

Pusan National University, South Korea

**Donghoon Lee**

Pusan National University, South Korea

**Hyunjung Shin**

Pusan National University, South Korea

**Myeong-ho Sohn**

George Washington University, U.S.A.

**Abstract:** In this study, we tried to find out whether the socio-cultural thinking styles of individualism and collectivism could affect the basic visual perceptual processes such as Miller-Lyer(ML) Illusion. If the collectivistic thinking style induces people to perceive the illusory figure more holistically and to process the relations among the components, collectivism-primed people are more susceptible to the ML illusion. On the other hand, if the individualistic thinking style induces them to perceive the illusory figure more analytically and to process the components separately, individualism-primed people are less susceptible to the illusion. To study the influence of cultural thinking styles on perceptual illusion, one group of Korean students wrote a story about myself for individualism priming and the other group about our group for collectivism priming before the illusion task. As expected, the individualism-primed group showed less ML illusion effect than the collectivism-primed group. In summary, this study suggested that cultural thinking styles could affect the basic visual perceptual processes in a top-down way.

# **How novices get skills without supervisors instructions: through analysis of skills mastery process**

**Jun Ichikawa**

Graduate School of Information Science, Nagoya University, Nagoya, Aichi, Japan

**Yugo Takeuchi**

Department of Information Science, Shizuoka University, Hamamatsu, Shizuoka, Japan

**Abstract:** This study proposes collaborative learning by which anyone can get skills without supervisors instructions. Generally, novices acquire them by being coached. Collaborative learning is defined as a method for studying and solving problems together in a group. In this study, subjects were instructed to get skills as collaborative learning for novices. The mastery process of their skills through the collaborative learning was analyzed by focusing on their speech communication and notes. And, the tacit knowledge to get skills and the process to keep it in mind were formulized from a cognitive point of view. The finding was suggested that a collaborative learning helped novices improve skills and share them, but novices could not figure out whether the outcomes were right or not. Furthermore, it was showed importance of getting skills: clues to know if skills are to decline or not, and ability to adjust opinions by collaborative learning.

# **The relationship between depressive tendency and relative metacomprehension accuracy.**

**Kenji Ikeda**

Nagoya university, Japan Society for Promotion of Science

**Yosuke Hattori**

Tokyo university, Japan Society for Promotion of Science

**Shinji Kitagami**

Nagoya university

**Abstract:** In this study, we investigated the relationship between depressive tendency and relative accuracy of metacomprehension. Many previous studies have focused on negative confidence bias or absolute accuracy of depressive persons (e.g., Fu et al., 2012). However, it has not been examined how depressive tendency affects relative accuracy. In experiment, participants read four expository texts. After reading, participants assigned a comprehension rating to each text and then completed a comprehension test. Finally, participants completed the Beck Depression Inventory (BDI) of Japanese, originally developed by Beck et al (1961). The result showed that there was negative correlation between the score of BDI and relative metacomprehension accuracy. Given that the high score of BDI means serious depressive, this result suggested that depressed persons cannot discriminate the degree of comprehension level between a numbers of texts.



# **A Narratological Mechanism for Generating Macro Structures of Story in an Integrated Narrative Generation System**

**Shohei Imabuchi**

Faculty of Software and Information Science, Graduate School of Iwate Prefectural University, Takizawa,  
Iwate, Japan

**Takashi Ogata**

Faculty of Software and Information Science, Iwate Prefectural University, Takizawa, Iwate, Japan

**Abstract:** This paper proposes a story generation mechanism based on Propp's narratology. Although there are some Propp-based story generation systems, the novel characteristics of our system are combining with a comparatively large scale conceptual dictionary, intending a comprehensive implementation of Propp's theory, and forming a part of our narrative generation system. The central part of the proposed system is a story grammar mechanism that we defined by organizing formally the description of function, which means an action seen from the result, and sub-function, which is our original name and the concrete action of each function. As other modules, the story grammar has a flexible structure which can be replaced by other story patterns and generated stories can be mutually combined. In this paper, we introduce the current version and enumerate future issues.

# A cellular automaton model of ambiguity aversion

**Kenryo Indo**

Kanto Gakuen University

**Abstract:** Daniel Ellsberg (1961) demonstrated that human decision makers tend to avoid unknown probabilities by using a pair of gamble comparisons. The ambiguity aversion has been studied by psychologists, economists, and more recently neuroscientists. For example, non-additive probability can be considered as a descriptive model to explain the ambiguity aversion. However, it lacks a cognitive processing model.

In this paper, a computational approach to modeling cognitive process of choice under ambiguous probability by using cellular automata is proposed. By designing cellular automaton in a torus consisting of "local matches" with the state transition rule by stochastic local  $q$ -majority vote, which is considered as a model of the working memory of decision maker, this paradox can be simulated. In addition, a similar but differently used version of this model can simulate the event-splitting effect which is known to induce, or eliminate, violation of first order stochastic dominance.

# The effects of frameworks and examples in learning how to solve word problems

**Miwa Inuzuka**

Taisho University

**Hirosuke Tanimoto**

Oshimanishi Junior Highschool

**Hiroko Kobayashi**

Japan Society for the Promotion of Science

**Abstract:** The study aimed at finding the way to facilitate students ability to solve proportion word problems. Previous research suggested two different ideas on abstract framework: abstract framework was less effective than examples; abstract representations would be effective in complex problems. Therefore, we focused on the abstractness of framework and the complexity of the problems. We compared problem solving performance of students in three classes with different abstractness in teaching how to solve proportion word problems. In no formula class (N=20), the teacher only showed correct expressions and answers; in abstract formula class (N=24), the teacher showed abstract formula to solve proportion problems; and in situation formula class (N=23), the teacher showed more grounded formula. The results indicated that the students in no formula did better than those in abstract formula condition. The effects of abstract frameworks and examples, and the complexity and transformation of the problems were discussed.

# **Social Projection as a Universal Strategy in Mental State Inference: Cultural Differences in Utilization of Stereotyping**

**Tatsunori Ishii**

Sophia University, Tokyo, Japan

**Masanori Takezawa**

Sophia University, Tokyo, Japan

**Abstract:** It has been argued that people selectively use two strategies, projection and stereotyping, for inferring mental states of the others. Through a series of experiments, Ames (2004) confirmed the hypothesis that, when a target person is perceived to be similar to oneself, people project ones own mental states to the other; when a target is perceived to be dissimilar, stereotype of a group/category the target person belongs to is used for mental state inferences. In this study, we replicated Ames(2004)s experiment in Japan and found that Japanese participants employed the projection unanimously regardless of the perceived similarity to the target person. We interpret this result with Yuki(2003) which argued East Asians perceive a social group as a network of independent people while Americans perceive a group as an entity. Our results thus suggest that stereotyping is a cultural-specific strategy while projection is a universal strategy employed in mental state inference.

# Functions for mutual interaction with the mind reading

**Satoru Ishikawa**

Hokusei Gakuen University, Sapporo, Hokkaido, Japan

**Abstract:** In the interaction with others, we read other's mind or estimate other's inner states, and utilize them for predicting other's behavior and deciding own behavior. Moreover, we utilize them for making someone do as we wished. In this study, we attempt to propose required functions for decision making of own and other's behavior depending on the mind reading processes and suggested mechanisms for realize these functions. Three investigational situations such as advising clothes to a customer, explanation of operating procedures, and misleading the other's belief were conducted and decision making processes of own behavior were analyzed depending on the recorded conversation during these situations or interview on video-playback. These results proposed two required functions: 1) acquisition of other's view of environment and detection of differences between own and other's view, 2) estimation of other's rules for decision making and operating them along with own will.

# **The anger superiority effect in children with and without autism**

**Tomoko Isomura**

Primate Research Institute, Kyoto University

**Hiroyasu Ito**

Primate Research Institute, Kyoto University

**Shino Ogawa**

Primate Research Institute, Kyoto University

**Miwa Fukushima**

Research Center for Advanced Science and Technology, Tokyo University

**Masahiro Shibasaki**

Primate Research Institute, Kyoto University

**Nobuo Masataka**

Primate Research Institute, Kyoto University

**Abstract:** In the face recognition research, it is well known that angry faces are detected more quickly than happy or neutral faces because of the attention-getting property provided with such threatening stimuli. This rapid processing to the threat, usually referred to as "the anger superiority effect", is thought to be automatic and inherent. Here, we examined the anger superiority effect in children with and without autism spectrum disorders (ASD) using visual search paradigm. Results revealed that the effect of emotion is correlated with age in children with ASD, but not in typically developing children. In addition, children with ASD under 9 years did not show the anger superiority effect whereas it was confirmed in typically developing children of the same age and even in younger ages. Children with ASD would be delayed for acquisition of those cognitive effects, which might be related to their delay of acquisition of variable social skills.

# **Culture, perception, and artistic visualisation: A comparative study of children's drawings in three Siberian cultural groups**

**Kirill Istomin**

Max Planck Institute for Social Anthropology, Halle (Saale), Germany

**Jaroslava Bagdasarova,**

Max Planck Institute for Social Anthropology

**Patrick Heady**

Max Planck Institute for Social Anthropology

**Abstract:** In a study of three indigenous and non-indigenous cultural groups in Northwestern and Northeastern Siberia, framed-line tests and a landscape drawing task were used to examine the hypotheses that test-based assessments of context-sensitivity and independence are correlated with the amount of contextual information contained in drawings, and with the order in which the focal and background objects are drawn. The results supported these hypotheses, and inspection of the regression relationships suggested that the inter-group variations in test-performance were likely to result from differences in the attention accorded to contextual information, as revealed by the drawings. Social and environmental explanations for the group-differences in contexts ensitivity are also discussed. The conclusions support the argument that cultural differences in artistic styles and perceptual tests reflect the same underlying perceptual tendencies, and are consistent with the argument that these tendencies reflect corresponding differences in patterns of social and environmental interaction.



# How do children with autism solve logic puzzle?

**Hiroyasu Ito**

Primate Research Institute, Inuyama, Aichi, Japan

**Nobuo Masataka**

Primate Research Institute, Inuyama, Aichi, Japan

**Abstract:** In this time, we examined that how children with autism solve logic puzzle. Seventeen school age children with autism and 17 age- and ability-matched typical children took part in the experiments. Two toys (a doll of robot and electric switch toy) were set on an architectural replica of house, and these were on a white styrofoam board. Children were instructed to search arbitrary rule that is relation between the position of robot and an on/off of electric switch toy. Children with autism solve the rule of logical AND and XOR problem (exclusive or problem) as good as typical children, but it was more difficult for children with autism to solve the rule of XOR in the big house replica on which there are many replicas of furniture than typical children. There were so many toys that Children with autism lost track of conceptualizing the pattern.

# Effects of Base Rates and Likelihoods on Intuitive Probabilistic Judgments

Tomoko Itoh

Japan Society for the Promotion of Science / Waseda University

**Abstract:** Sixty university students solved four intuitive probability problems in which two base rates (the probabilities that people in a country are genetically predisposed to a disease) and two likelihoods (the probabilities that people who are genetically predisposed to the disease will actually develop it even if they are vaccinated) were provided. Participants made judgments on whether they would get vaccinated on the basis of the information. The number of participants who responded that they would get vaccinated was significantly different across problems. More participants responded that they would get vaccinated when the base rate was 800 out of 1000 than when it was 1 out of 1000. If the base rates were the same, more participants responded that they would get vaccinated when the likelihood was 15% than when it was 75%. These results suggest that people make intuitive probabilistic judgments on the basis of both base rates and likelihoods.

# Representational Form and Metaphorical Word Use

**Anja Jamrozik**

Northwestern University

**Micah Goldwater**

Northwestern University

**Eyal Sagi**

Northwestern University

**Dedre Gentner**

Northwestern University

**Abstract:** How does semantic representation influence the likelihood that a word will be used metaphorically? We explore whether words whose meanings are defined by relations among entities (e.g., marriage, forget), are more likely to be used metaphorically than words whose meanings are defined by features of entities (e.g., bird). Verbs are generally more relational than nouns (Gentner, 1981). Relationality can also distinguish different kinds of nouns: specifically, relational nouns (e.g., marriage) vs. entity nouns (e.g., bird) (Gentner & Kurtz, 2005; Goldwater, Markman, & Stilwell, 2011; Markman & Stilwell, 2001). Prior studies have shown that the meanings of relational words are more mutable across contexts than those of entity words (Gentner & France, 1988; Asmuth & Gentner, under review). Extending this work, we find that uses of relational words (both verbs and relational nouns) tend to be more metaphorical than uses of entity nouns in natural language corpora.

# Interaction between ability and use of scaffold in EFL vocabulary learning system

**Felix Jimenez**

Chukyo University, Toyota, Aichi Pref, Japan

**Masayoshi Kanoh**

Chukyo University, Toyota, Aichi Pref, Japan

**Abstract:** This study clarified self-regulated learning processes in EFL in order to propose a more adaptive learning system. We developed an English vocabulary learning system for Japanese. It presented each word in an example sentence and the Japanese translation of the sentence except for the target word upon user request. We sought to examine how learners use this system as a scaffold. Five college students of lower and intermediate ability used the system for six weeks and took exams. We found that the intermediate-ability group learned more vocabulary than the lower-ability group, even when compared to similar-level students using another system that provided only words and translation. Process analyses revealed that the lower-ability group depended on help all through the period, while the intermediate-ability group decreased its use gradually as if to check their learnability. This suggests that the future system have both scaffolding and fading-out functions adaptable to learning states.

# Qualitative differences in sequence planning with everyday objects in traumatic brain injured individuals

**Arianne Johnson**

University of California, Santa Barbara

**Scott Grafton**

University of California, Santa Barbara

**Abstract:** The present study aimed to characterize what errors in sequential tasks that traumatic brain injury (TBI) subjects make relative to controls. Subjects (8 healthy controls, 6 TBI subjects) completed a computerized grocery bagging task on a touchscreen that required bagging items according to object properties. Relative to controls, TBI subjects produced marginally more errors overall. However, the TBI group had different error profiles than the control group. Specifically, on trials that required a nested rule (i.e. clumping items according to temperature and ordering these same items in terms of weight), the TBI group had more than twice as many errors as controls. This performance deficit is specific to nested rules as the TBI error rates did not differ relative to controls on trials that required multiple rules that were not nested. These findings suggest that planning deficits from TBI cannot simply be due to increased memory load of number of rules but instead is more specific to sequence planning.

# Differential Effects of the Cultural Orientation Dimensions on Global Precedence

**Mijung Joo**

Pusan National University

**Hyunmin Kang**

Pusan National University

**Hyunjung Shin**

Pusan National University

**Jaesik Lee**

Pusan National University

**Abstract:** This study investigated the differential effects of individualism-collectivism dimension (ICD) and horizontal-vertical dimension (HVD) in cultural orientation of individuals in the same cultural background on global precedence. The participants were classified into ICD and HVD groups, and asked to respond to compound stimuli which were varied by stimuli types (figure/letter) and stimulus-stimulus (S-S) congruence. Differences in global precedence were compared. The results showed the followings. First, although global precedence was larger in the compound figure than in the compound letter, and larger in the S-S incongruent condition than the congruent condition, none of cultural orientation dimensions made any difference. Second, ICD and HVD affected global precedence independently. Third, a significant interaction effect between HVD and S-S congruence was found, but there was no interaction between ICD and S-S congruence. These results indicated HVD rather than ICD can be a more valid dimension to compare the effect of individual differences in cultural orientation on global precedence.

# **Hierarchical Category Structures Facilitate Acquisition of Probabilistic Relational Categories.**

**Wookyoung Jung**

University of Illinois at Urbana-Champaign

**John Hummel**

University of Illinois at Urbana-Champaign

**Abstract:** Kittur et al. (2004, 2006) and Jung & Hummel (2009, 2011) showed that people have great difficulty learning relation-based categories with a probabilistic (i.e., family resemblance) structure, in which no single relation is shared by all members of a category. Yet acquisition of such categories is not strictly impossible: In all these studies, roughly half the subjects eventually reached criterion. What are these subjects doing that the other half are not? We hypothesized that successful subjects were those who divided the nominal categories into two or more sub-categories, each of which individually has a deterministic structure. We report an experiment testing and supporting this hypothesis: Explicitly presenting subjects with hierarchical (category and sub-category) structures facilitated the acquisition of otherwise probabilistic relational categories.



# Dynamic Effects of Perceptual and Categorical Similarity on Recognition Memory

**George Kachergis**

Indiana University

**Gregory Cox**

Indiana University

**Richard Shiffrin**

Indiana University

**Abstract:** Memory is sensitive to several aspects of its contents, including item similarity (Tulving, 1981) and value (Kachergis, Recchia, & Shiffrin, 2011). For the present study, we developed and scaled a new class of blob-shaped stimuli in order to manipulate item similarity. In two experiments, participants studied blob-point value pairs, with categories defined by valence (positive/negative point values) and by perceptual similarity of the stimuli. Using a dynamic 2AFC decision-making paradigm with a response deadline, we recorded response trajectories for different foil types (e.g., similar to the target, or unique). We find similarity effects in recognition memory and semi-paradoxical category size effects in accord with Tulving (1981). Interpretation of the accuracy effects is enhanced by examining the shape and timing of response trajectories, allowing us to measure phenomena such as decision reversals. Finally, we present a general Gaussian process regression framework to analyze such continuous response data (Cox, Kachergis, & Shiffrin, submitted).

# Asymmetry of McGurk Effect Depending on the Orientation of Faces

**Masayo Kajimura**

Kyoto University

**Hiroshi Ashida**

Kyoto University

**Abstract:** We examined whether the McGurk effect depends on speakers facial orientation. Speakers vocalizing scenes of, /ba/, /da/, and /ga/ were videotaped in two angles from the left and the right. A voice and a face were congruently or incongruently combined. The experimental stimuli were edited to produce normal and mirror-inverted versions of each audio-visual pair. Participants were asked to report the syllable as they heard by free description. The result replicated the McGurk effect. The error rates were higher for the mirror-inverted right-looking faces (seen as left-looking) than for the mirror-inverted left-looking faces (seen as right-looking). And the error rates were higher for the mirror-inverted right-looking faces (seen as left-looking) than for the normal right-looking faces. These results suggest that left-looking faces enhance the McGurk effect, while differences in physical movements of the left and right sides of speakers mouth could be also relevant.

# **An Optimality Theoretical Analysis of Urge Interactions in Toda's (1982) Emotional Fungus-Eater Robots**

**Yasuo Kaneko**

Kushiro Public University

**Abstract:** Toda's (1982) Urge Theory of Emotions (TUTE) is a deductive theory of emotion and cognition. The TUTE is deductive in that one of thrusts of the TUTE is a detailed analysis of individual emotions as necessary softwares of robots named Fungus-Eaters working on an imaginary planet. The Urges are built-in motivational subroutines for making decisions and executing action plans in order to solve survival problems in a wild environment. The urges are supposed to interact in a situation but not to be absolute but violable. However, the nature of the urge interactions remains to be clarified. The purpose of this study is to apply Optimality Theory (Prince & Smolensky, 2004) to an analysis of the urge interactions because OT is a theory of interactions of violable constraints. The application makes the urge interactions more visible by means of tableau and accessible for scrutiny than in the TUTE.

# **A signal detection analysis of the effects of repeated context on visual search**

**Ryan Kasper**

University of California Santa Barbara, Santa Barbara, California, United States

**Miguel Eckstein**

University of California Santa Barbara, Santa Barbara, California, United States

**Barry Giesbrecht**

University of California Santa Barbara, Santa Barbara, California, United States

**Abstract:** When searching for a target in a scene, previously learned spatial context can facilitate response times (Chun & Jiang, 1998). However, it has been debated whether this is driven by attentional guidance (Chun, 2000; Geyer et al., 2010) or response selection processes (Kunar et al., 2007; Kunar & Wolfe, 2011). To address this issue, 18 subjects performed a visual search task in which (a) the spatial configuration was either repeated or novel, and (b) the target was either present or absent. The present/absent manipulation permitted signal detection computations to decouple perceptual sensitivity ( $d$ ) and response bias ( $\beta$ ). Average RT was faster ( $p < 0.01$ ) and  $d$  was increased ( $p < 0.03$ ) for repeated displays, but there was no effect on ( $p > 0.45$ ). These results suggest that early perceptual processes, as opposed to response processes, may provide a greater contribution to the benefits of repeated context in visual search.

# Cultural variations in authority management in interaction

Yasuhiro Katagiri

Future University Hakodate

**Abstract:** Status difference, such as in seniority, wealth or social roles, plays an important role in human joint decision making situations. The degree of effectiveness of authority brought about by status difference naturally varies across different cultural groups. But, the way in which authority manifests itself in joint decision making situations also changes in different cultural groups. We have been studying cultural variations in dialogue interaction styles through the examination of Mister O corpus, a task-oriented dialogue corpus which has been collected over several languages, including Japanese, English and Arabic. We identified a contrastive manifestations of authority in Mister O dialogues, which can be characterized in terms of different emphasis on competence and deference. We discuss possible evolutionary underpinnings of the diversity of interaction style manifestations, based on the notion of multi-level selection and tribal social instinct hypothesis.

# Differences in emotional bias effect on working memory in elderly.

**Maya Katsuhara**

Kyoto university

**Mariko Osaka**

Osaka university

**Naoyuki Osaka**

Kyoto university

**Abstract:** This study investigated age-related differences in the inhibition mechanism for negative emotional information and in the facilitation mechanism for positive emotional information in working memory. Thirty-six older and thirty-six younger adults performed three RSTs: the neutral, negative, and positive emotional conditions. For each RST condition, all target words were neutral words, while the sentences themselves corresponded to the conditions. In older adults, the percentage of correct responses in the negative condition was worse than that of the neutral condition. Conversely, in younger adults, there was no significant difference between the neutral and negative conditions. In contrast, performance in the positive condition was better than the neutral condition for both age groups. These results suggest that the inhibition of negative information comes at a greater cost to older adults compared to younger adults, but the benefit of the positive information is the same in both age groups.

# **Mental rotation of pictured body stimuli: Involvement of visual representation of the stimuli**

**Tsubasa Kawasaki**

Tokyo Metropolitan University

**Takahiro Higuchi**

Tokyo Metropolitan University

**Abstract:** The present study investigated whether the time required for recognizing rotated body stimuli was matched strictly with the magnitude of correspondence regarding rotation angles between the stimulus and the body. Twelve young adults sat in front the computer monitor while their body was facing to it (no body rotation) or was rotating to the side by 90 degrees. They classified a body stimulus (hand or foot) presented with one of four orientations (0, 90, 180, -90) according to their laterality. The result showed that the time required for the classification was generally matched with the magnitude of correspondence regarding the rotation angles. However, the classification time was the longest for the 180-deg rotated stimulus even while the body was rotated by 90 degrees, demonstrating that the magnitude of rotation of the body stimuli itself affected the MR. It is likely that visual representation of the stimuli, as well as the body schema, is involved in the MR.



# Extracting the Musical Schema from Traditional Japanese, Chinese and German Folk Songs

Akihiro Kawase

Tokyo Institute of Technology

**Abstract:** In this study, we extract the pitch transition patterns from traditional Japanese, Chinese and German folk songs, and examine the characteristics of their respective schema. We sample 1,794 works from Nihon Min-yo Taikan for Japanese folk songs, 1,984 and 2,286 works from a website providing virtual musical scores for both Chinese and German folk songs, respectively. Our main method of extracting pitch transition patterns is to fit variable-length Markov chains (VLMCs) from musical data. A variable-length Markov chain model is a Markovian process having a sparse memory structure with some states that closely cohere. The structure can be characterized by a parsimonious number of transition probabilities for stationary categorical time series. The results indicate that the minimal structures of Japanese folk songs tend to create a longer schema than other two folk songs, while the minimal structures of both Chinese and German folk songs tend to create a shorter schema.

# **Implicit learning on the order of the visual stimuli under interocular suppression**

**Kaede Kido**

Osaka prefecture university

**Shogo Makioka**

Osaka prefecture university

**Abstract:** We investigated whether learning on the order of the visual stimuli occurs under interocular suppression. In learning phase, low contrast characters were presented sequentially to one of the eyes. Conscious perception of the characters was suppressed by flash streams presented to the other eye (continuous flash suppression). The character streams consisted of quadruplets of fixed order characters, and the order of the quadruplets was randomized. The characters were selected from Yi script that was new to the subjects. In testing phase, the quadruplets were presented to the both eyes, and the subjects hit the key as soon as the target character appeared. If the subjects respond faster when the position of the target was later in the quadruplets, it is likely that they had learned the order of the stimuli. Preliminary data suggests that the learning on the order of novel visual stimuli occurred under interocular suppression.

# Hierarchical Slow-Feature Models of Gesture Conversation

**Jiseob Kim**

Seoul National University, Seoul, Korea, Republic of

**Sooyong Jang**

Seoul National University, Seoul, Korea, Republic of

**Eun-Sol Kim**

Seoul National University, Seoul, Korea, Republic of

**Byoung-Tak Zhang**

Seoul National University, Seoul, Korea, Republic of

**Abstract:** How do humans interact with each other using gestures? How do they catch the semantics of gestures to predict or react to them? We explore hierarchical slow-feature models to obtain the high-level semantics in conducting gesture conversations. We adopt the hypernetwork model as a basic component to learn the elementary semantics of gestures, and combine two hypernetworks with an added upper-layer of slow features to learn the temporal transition of semantics. This hierarchical slow-feature model abstracts the low-level features of joint angles to the slowly-changing high-level features which represent higher-level semantics of the gestures. This model also learns the probability of the partners next gesture given the gesture of one person at a semantic level. We experimented with the Kinect motion capture device to record the gestures of two subjects in gesture conversation scenarios. The human gesture data was used to train the hierarchical slow-feature model to predict the gesture conversations. The trained model is then transferred to the Darwin humanoid robot. We compare the human behaviors and the robot behaviors that learned from the human gestures.

# **Is this right? or Is that wrong?: Evidence from Dynamic Eye-Hand Movement in Decision Making**

**Eun-Sol Kim**

Seoul National University, Seoul, Korea, Republic of

**Jiseob Kim**

Seoul National University, Seoul, Korea, Republic of

**Thies Pfeiffer**

Bielefeld University

**Ipke Wachsmuth**

Bielefeld University

**Byoung-Tak Zhang**

Seoul National University, Seoul, Korea, Republic of

**Abstract:** Eye tracking and hand motion (or mouse) tracking are complementary techniques to study the dynamics underlying human cognition. Eye tracking provides information about attention, reasoning, mental imagery, but figuring out the dynamics of cognition is hard. On the other hand, hand movement reveals the hidden states of high-level cognition as a continuous trajectory, but the detailed process is difficult to infer. Here, we use both eye and hand tracking while the subject watches a video drama and plays a multimodal memory game (MMG), a memory recall task designed to investigate the mechanism of recalling the contents of dramas. Our experimental results show that eye tracking and mouse tracking provide complementary information on cognitive processes. In particular, we found that, when humans make difficult decisions, they tend to ask 'Is the distractor wrong?', rather than 'Is the decision right?'.

# State Anxiety and the Processing of Covariation Information in Causal Reasoning

**Young Il Kim**

Ajou University

**Kyung Il Kim**

Ajou University

**Abstract:** The phenomena of conditionalization and discounting have been studied as reflection of attention to multiple potential causes in causal reasoning. And it has been observed that individual-difference factors significantly influence causal reasoning. Specifically, previous studies showed that sociality anxiety related factors such as self-construal influence levels of conditionalization and discounting. In this study, we further specified the relationship between anxiety and the two mechanisms in causal reasoning. We manipulated participants levels of state anxiety and found that participants with a relatively greater state anxiety showed a greater propensity to discount their judgment on an alternative cause. But this pattern was not observed in conditionalization between participants with two different levels (high vs. low) of state anxiety. First, these results suggest that conditionalization and discounting are independent to each other, further supporting previous studies. Second, we propose that different individual-difference factors specifically influence the two mechanisms in causal reasoning.

# **Cultural priming and the scene perception**

**Bia Kim**

Pusan National University

**Yoonkyoung Lee**

Pusan National University

**Donghoon Lee**

Pusan National University

**Goeun Lee**

Pusan National University

**HyunJung Shin**

Pusan National University

**Abstract:** A newly constructed 'cultural priming story writing task' was used to examine the hypothesis that priming of individualism leads to pay attention to the targets more than the grounds, whereas priming of collectivism leads to equally distribute attention to the relation between the targets and the grounds as well as the targets in the scene perception. In addition, focusing on the contradictory results of previous studies, we investigated the effect of instruction: In Experiment 1, participants were instructed to rate the preference of the scenes, whereas they were informed the recognition test beforehand in Experiment 2. The results supported the hypothesis: The collectivism-primed participants pay more attention to the figure-ground relevance information than the individualism-primed ones. However, the difference was disappeared in Experiment 2. It was suggested that intention to memorize scenes can affect the attentional allocation to the targets and the grounds regardless of the cultural dispositions.

# **Complex Network Analysis of Social Relationships and Personality from TV Drama Dialogues**

**Joon Shik Kim**

Seoul National University

**Chung-Yeon Lee**

Seoul National University

**Minsu Zhang**

Sangmoon High School

**Jun-Hee Nam**

Sangmoon High School

**Abstract:** Dialogues are linguistic interactions between people and can provide hints on social relationship and personality. We aim to analyze the social relationships of the characters in TV dramas based on dialogues. In addition to knowing just who talks to whom, the analysis of dialogue content gives more detailed information on the types of interaction, e.g. asking or informing, and the emotional status. We use complex network measures, such as path lengths, geodesics, cluster coefficient, and centrality, to analyze the structural and functional properties of the dialogue network. Among others, we found the betweenness centrality measure informative for identifying the social relationships since this assesses the importance of a node in a complex network in which small worldness is inherent. Analyzing the complete corpus of TV drama Friends aired for 10 years, we discovered the connector nodes which are essentially the social equivalent of a hub of computer network. Complexity measures of the dialogue network also help characterize the personality of the main characters in the dramas we studied. The text-based analysis of the present work can be extended to incorporate the sound modality to capture further information on the characters and their interactions.



# A Neuroethological Approach to Robotics

**DaeEun Kim**

Yonsei University, Seoul, South Korea

**Abstract:** Many animal behaviors provide a clue to animal intelligence. The underlying neural mechanism is still unsolved in many cases, but the study stimulates relevant studies in the field of artificial intelligence and robotics. Robotic researchers have interest in how sensory system is connected to motor actions and what kind of learning mechanism or what coordination in the sensorimotor mapping is helpful to the efficient operation of machines or their parts. In the field of artificial intelligence, researchers try to find optimal strategies with the acquired environmental information. A neuroethological approach to animal behaviors, which handles the interaction of neural circuitry to explain the behavioral concept, can find many clues of animal cognition. It also provides a systematic and integrative foundation of intelligence needed in robotics and artificial intelligence, which might lead to intelligent systems or robots.

# The effectiveness of English teaching integration model based on task-based approach

Eun sook Kim

Pusan National University

**Abstract:** Task-based activities have become very popular recently in both stand-alone and content-based language classes. Also, They are taken into account to improve students responsibility and sociality in English learning. This study tries to devise an integration of English teaching model that integrate reading, writing, speaking and culture, based on task-based learning approach, and investigate the improvement of interest and confidence about English productive competency like writing and speaking, which has significantly risen ( $t=0.04$ ,  $p<.05$ ) in pre- or post- application of the model. Approximately 55 percent of the students responded that the model was effective for the organization abilities of materials and the confidence about essays and presentations. There are no meaningful differences between the emotional aspects and the cognitive aspects( $t=1.61$   $p.05$ ).

# Nominal Number and Semantics of Common Nouns in Numeral Classifier Languages

**Jecheon Kim**

Hankuk University of Foreign Studies, Seoul, Korea

**Abstract:** The present study argues against the predominant view in the literature that the default semantic denotation of common nouns in numeral classifier languages, i.e., table, is similar to the denotation of the mass nouns in English, i.e., furniture. Based on this mass noun view, many linguists argue that there is no count-mass distinction in numeral classifier languages.

I show that the mass noun view based on simple morpho-syntactic analogies between English and numeral classifier languages has a number of problems both in theory and in the interpretation of the data. I provide evidence for the existence of a count-mass distinction in numeral classifier languages from various areas of grammar based on the Korean data and supporting evidence from the researches in other disciplines.

My study will be able to contribute to the renewed and correct understanding on the nature of common noun semantics of the numeral classifier languages.

# **Automaticity in Motor Learning: Evidence from Visuo-motor Tracking Performance and Pupil Dilation**

**Satoshi Kobori**

Ryukoku University

**Yosuke Abe**

Ryukoku University

**Shogo Nakazono**

Ryukoku University

**Abstract:** Motor learning has traditionally been associated with the concept of automaticity. Automaticity refers to the reduction of the cognitive effort required to perform a motor task, as learning progresses. However, there is little detailed consensus in the literature on what the process of automatization actually involves. We measured tracking performance in two groups of participants while either the target or the manual cursor was suppressed for a brief period during each trial. Subjects learned to maintain accurate tracking through periods of target or cursor suppression. We have used this approach to investigate the internal models used during tracking, and their updating during motor learning. We have simultaneously measured tracking performance and pupil dilation as a measure of cognitive load. The results showed that pupil diameter decreases with learning process of tracking performance. Decrease of pupil diameter suggests that automatization is linked specifically to the learning of internal models.

# Self-Organization of Policy by Symmetric Reasoning and its Application to Reinforcement Learning

**Yu Kohno**

Tokyo Denki University

**Tatsuji Takahashi**

Tokyo Denki University

**Abstract:** The real limitations to our cognition and our locality lead to the exploration-exploitation dilemma and hence the tradeoff between speed and accuracy. Considering that all creatures especially animals efficiently deal with the tradeoff, it is natural to suppose we can intuitively handle the dilemma. We adopt the loosely symmetric (LS) formula as a toy model of our intuitive judgment. LS, a kind of biased conditional probability, is known to precisely describe our probability judgment and to go beyond the ordinary tradeoff of speed and accuracy in two-armed bandit problems. In this study, we give analyses of LS in relation to its information theoretical and logical-probabilistic nature. Along with the analyses, we extend LS from a conditional probability to a general value function. The efficacy of LS in decision-making under risk and reinforcement learning is proven by experiments and simulations.

# Study on Facilitation of Problem Posing by Learning Examples through Reproduction

**Kazuaki Kojima**

Waseda University

**Kazuhisa Miwa**

Nagoya University

**Tatsunori Matsui**

Waseda University

**Abstract:** In general education, learning of production tasks is important but difficult due to it requires heavy cognitive activities such as generation of ideas and synthesis of structures. We proposed a method to facilitate a production task of mathematical problem posing through learning by reproducing examples. In the proposed method, learners learn essential ideas by reproducing examples based on process information indicating how to compose them. We then conducted an experimental investigation where undergraduate students posed their own problems after learning an example by reproducing or solving it. Our previous study had confirmed that undergraduate students without learning of any examples posed many problems that had simple and inappropriate solution structures. In this study, undergraduate students who learned by reproducing the example posed many complex problems, whereas those who learned by solving it didn't do. This proved that learning of ideas for a production task is more effective when it's done through a productive activity.

# Egocentric and allocentric frame of reference in virtual maze navigation

**Takatsugu Kojima**

Shiga University of Medical Science

**Abstract:** This study examines how a representation of an egocentric orientation influences selecting an allocentric frame of reference when we guide someone through a virtual maze constructed by three-dimensional computer graphics. In two experiments, participants guided a human-like agent through a virtual maze. They were instructed to use an allocentric frame of reference (an agent-centered frame of reference) to guide the agent through a maze in Experiment 1. Additionally, they were instructed to represent their own orientations during navigation and to answer their own orientations at a maze's goal in Experiment 2. We controlled experimentally sizes of mazes and analyzed participants' patterns and reaction times of selecting a direction of movement for the agent in the two experiments. The results showed that a representation of an egocentric orientation influenced selecting an allocentric frame of reference in maze navigation depending on a size of a maze.



# Consciousness and the language faculty: awareness, qualia and natural language using agents

Piotr Konderak

Maria Curie-Skłodowska University, Lublin, Poland

**Abstract:** I would like to begin with the role of language in settling the problem of consciousness. Some expressions of our everyday language (entering the consciousness, accessible to consciousness consciousness reads) suggest the wrong idea of the phenomenon, namely consciousness as something external in relation to mental states. My first claim is that mental states are conscious due to some internal property.

Accepting Chalmers' distinction between psychological and phenomenal consciousness I will analyse functioning of a natural language processing agent of the SNePS system and indicate hypothesized places and roles for psychological consciousness.

Although explaining the language faculty requires taking into account only awareness, language is also inseparably connected with qualia. Answering the question why qualia accompany some (and what) aspects of language, I will speculate that qualia (regardless their experiential character) have a function similar to Damasio's somatic marker: they allow a kind of pre-processing, triggering psychological mechanisms.

# Three co-creation stages in formation of symbol communication systems

**Takeshi Konno**

Japan Advanced Institute of Science and Technology

**Junya Morita**

Japan Advanced Institute of Science and Technology

**Akihito Kishino**

Japan Advanced Institute of Science and Technology

**Takashi Hashimoto**

Japan Advanced Institute of Science and Technology

**Jiro Okuda**

Kyoto Sangyo University

**Maki Suzuki**

Kyoto Sangyo University

**Abstract:** In Konno et al.'s (in press) experiment of co-creation of artificial language, it has been confirmed that a formation pattern had three stages: the establishment of conventional behavior, symbol system and division of roles using turn-taking. However, forty percent of participants failed the formation. We consider that a cause of the failure lies in difficulty of separation among the three stages which demand learning at different levels. Thus we designed three tasks in order to separate the three stages explicitly. Consequently, all twelve pairs completed the co-creation task, and the differences between two experiment's final scores were significant. This indicates that the order of establishment would lead to form an effective symbol communication system. We suggest that the order has an effect of scaffolding the formation. Finally, we will discuss whether the order has an effect of starting small (Elman, 1993) for the learning at the different levels.

# **Theory of Mind network encodes how you know what you know in blind and sighted adults**

**Jorie Koster-Hale**

Massachusetts Institute of Technology

**Rebecca Saxe**

Massachusetts Institute of Technology

**Marina Bedny**

Massachusetts Institute of Technology

**Abstract:** Theory of Mind (ToM) is the ability to reason about mental experiences such as beliefs and desires. ToM depends on a specific network of brain regions, including the right temporo-parietal junction (RTPJ). We investigated whether these brain regions distinguish between others' mental states experienced through seeing vs hearing ('Sarah saw that', vs 'Sarah heard that'), using multivoxel pattern analyses (MVPA). We find that the spatial pattern in the RTPJ distinguishes stories about mental states experienced through hearing vs. seeing (i.e. source) but not other salient distinctions, such as the positive vs. negative affective valence of the mental state. To investigate effects of first-person experience on these representations, we repeated the experiment in congenitally blind participants. Again, MVPA in the RTPJ distinguished source but not valence. We conclude that 1) the source information of someone else's mental state is coded in ToM brain regions and is a feature of ToM, and 2) these ToM representations of perceptual source are not based on first-person perceptual experiences.

# **A joint ideomotor effect increases the inter-brain oscillation between two people engaged in a Japanese Ouija board Kokkuri-san.**

**Kenta Kubo**

Japan Science and Technology Agency

**Kentaro Katahira**

Japan Science and Technology Agency

**Kazuo Okanoya**

University of Tokyo

**Masato Okada**

University of Tokyo

**Nobuyuki Kawai**

Nagoya University

**Abstract:** We investigated the inter-brain activity coherence recorded by an EEG between two participants when playing at Kokkuri-san, a Japanese Ouija board. The Ouija aboard, known as a spirit key board or talking board, is a flat board, marked with the letters of the alphabet, numbers, or words (yes or no). While playing at the Ouija board, two or more participants place their fingers on the planchette, and it moves about the board and spells out words. The Ouijas actions are generally considered to be due to the ideomotor effect. The EEG results showed that inter-brain activity coherence was observed more dominant in the temporo-parietal junction when the participants shared the goal (in this case, yes or no) than when they did not. These results suggest that shared ideomotor effects produce brain oscillations in the cortical area, which are used for mentalizing.

# **A Framework of Sentence Generation Mechanism for a Narrative Generation System**

**Shinya Kumagai**

Iwate Prefectural University, Iwate, Japan

**Sou Funakoshi**

Iwate Prefectural University, Iwate, Japan

**Junpei Ono**

Iwate Prefectural University, Iwate, Japan

**Taisuke Akimoto**

Iwate Prefectural University, Iwate, Japan

**Takashi Ogata**

Iwate prefectural university, Iwate, Japan

**Abstract:** For language representation in our narrative generation system, we have used a simple sentence generation mechanism. This paper discuss the more comprehensive framework. The approach contains linguistic aspect and strategic or controlling aspect. For the former, we developed programs for word order transformation, verb conjugation operation, complex sentence generation, and letter notation selection. The last function uses a language dictionary for nouns and verbs connected to a conceptual dictionary for noun and verb concepts. For the latter, we use the category of voice in Genettes narrative discourse theory, which means a concept to define the feature and relationship of narrator and narratee to decide the strategy and direction of discourse and language generation. In this paper, based on the developed sentence generation programs, we propose a comprehensive framework consisting of the definition of narrator & narratee, strategic decision rules, and concrete sentence operation modules.

# **An on-line model of quantifiers interpretation in Japanese**

**Takeo Kurafuji**

Ritsumeikan University

**Masakatsu Inoue**

Mukogawa Women's University

**Michinao Matsui**

Osaka Health Science University

**Abstract:** The present project investigates apparently paradoxical phenomena concerning quantificational expressions in Japanese sentence processing as follows. Given structurally ambiguous sentences, (i) when both the nominative and the accusative NPs are bare nouns (nouns without quantifier, interpreted existentially), the garden path (GP) effect is observed, while (ii) when either one of the NPs has a quantifier such as *\_subeteno\_* 'all/every', the GP effect reduces, but (iii) when both are quantified, the GP effect reemerges. The question is: if the existence of one quantifier reduces the GP effect, why not two? We argue that in (ii) the scope interaction of the quantifier and the bare noun is computed, which retards the syntactic processing, resulting in the low GP effect. On the other hand, in (iii), because of its complexity, the scope interaction is not computed, and those quantifiers are treated as non-quantificational, group objects.

# Beauty and Cuteness in Peripheral Vision

**Kana Kuraguchi**

Kyoto Univ., Kyoto, Kyoto, Japan

**Hiroshi Ashida**

Kyoto Univ., Kyoto, Kyoto, Japan

**Abstract:** Previous research suggests that attractiveness is not only detectable in the central visual field but also in the periphery. In Japan, visual attractiveness of people is often described in terms of beauty and cuteness (kawaii). We examined how perception of facial attractiveness in those two aspects can be different in peripheral vision where special resolution is deteriorated. Pairs of female face images were presented at several eccentricities, and participants judged which was better in terms of beauty or cuteness. The results showed that participants were able to judge both beauty and cuteness in the periphery. The discrimination performance, however, differed; beauty was more detectable in the periphery than in the parafovea, while cuteness was more detectable in the parafovea. This result suggests that detection of beauty might be ecologically more important for humans. Moreover, there were some gender differences, which may reflect different meaning of facial attractiveness for each gender.



# **Motion-related brain activity enhanced by motion representation during metaphor understanding**

**Naoko Kuriyama**

Tokyo Institute of Technology

**Asuka Terai**

Tokyo Institute of Technology

**Takashi Kusumi**

Kyoto University

**Masanori Nakagawa**

Tokyo Institute of Technology

**Kimihiko Yamagishi**

Tokyo Institute of Technology

**Koji Jimura**

Tokyo Institute of Technology

**Abstract:** Previous neuroimaging studies have suggested the involvement of motor-related brain regions during semantic processing of action-related idioms and fictive motion sentences. However, it remains unclear if such regions are also involved in understanding of metaphorical expressions. Because metaphor understanding requires abstract semantic manipulations and elaborated feature mapping, one possibility remains that it involves more semantic processing implicated in fronto-temporal regions. On the other hand, due to the conspicuous nature of active items, it may enhance representation of motion and actions, resulting in increased activations in motor-related brain regions. The current fMRI study tested these hypotheses by measuring brain activity while participants made semantic judgments about metaphor sentences involving action and motion. The motion metaphor condition, relative to the motor-related but literal condition, showed prominent brain activity in a fronto-motor region near BA 4. This result suggests that understanding of motion-related metaphor induces direct representation of action and motion.

# **The influence of redundant and idiosyncratic attributes on coherence within a category and contrast between categories**

**Ikuko Kyoya**

Ritsumeikan University

**Masaomi Oda**

Ritsumeikan University

**Abstract:** Many studies on categorization assume that not only coherence within a category but also contrast between categories is important for categorization. This study used line drawings of fictional insects to examine how redundant and idiosyncratic attributes influence categorization. In this study, these attributes were color attributes (brown or blue) that were added to only one instance for each of two categories during category learning and were mutually different between categories. After category learning, a transfer test involving categorization judgments and confidence ratings for those judgments was conducted. The results suggested that the idiosyncratic color of the own category did not help participants to categorize instances into this category, but the idiosyncratic color of the other category deterred participants from categorizing instances into the own category. These results reveal that categorization involves a complex process of drawing contrasts between categories.

# **Assessment and refinement of an intelligent tutor for complex decision making**

**Daniel Lafond**

Defence R&D Canada

**Sebastien Tremblay**

Universite Laval

**Michel DuCharme**

Defence R&D Canada

**Marie-Eve St-Louis**

Universite Laval

**Jean-Francois Gagnon**

Universite Laval

**Abstract:** We tested the effectiveness of a prototype simulation-based training procedure designed to support decision making in complex dynamic environments. An intelligent tutor was designed to monitor trainees use of over-simplifying heuristics throughout three scenarios and to discourage their use in a timely manner. Two test scenarios were used to assess training effectiveness: One involved a transparent problem structure while the other was partly opaque. Results showed a 34% decrease in the average use of the four heuristics monitored. However, goal attainment did not significantly improve in the test scenarios (heuristics were not replaced by better strategies). The training had a positive impact on participants ability to anticipate system behavior in the high transparency scenario but a negative impact in the low transparency scenario. We conclude that intelligent tutors should not only detract from using ineffective heuristics but also prescribe effective context relevant heuristics to overcome the wall of complexity.

# Evidence for a phonology-specific learning mechanism

**Regine Lai**

University of Delaware

**Jeffrey Heinz**

University of Delaware

**Abstract:** A series of comparative artificial language learning experiments were conducted to obtain evidence for or against the hypothesis that the mechanism responsible for learning phonological patterns in natural language is distinct from mechanisms responsible for pattern learning in other domains, including natural language syntax (Heinz and Idsardi, 2011, *Science*). Results from these experiments indicate that the hypothesis is correct.

Each experiment examined the learnability of one of three patterns in phonological, syntactic, non-speech auditory, and visual domains. These patterns are significant in different ways: pattern SH is attested in phonology; pattern FL is unattested in phonology; pattern NE is unattested in phonology, but attested in syntax. Results showed that SH was more readily learnable than FL in the phonological and the non-speech auditory conditions, but not the visual condition. Also, the performance of the phonological group was significantly different from the syntax group when exposed to the NE pattern.

# Deceptive strategy choice as decision making under risk

**Tei Laine**

Institute of High Performance Computing

**Kayo Sakamoto**

Institute of High Performance Computing

**Abstract:** Verbal deception often involves more subtle forms of deception than simply telling an outright lie. These different forms of deception are associated with different potential costs and benefits and thus constitute strategic choice options. To model this choice process, we treat deceptive strategy selection as decision making under risk; we assume that the speaker knows what gains can result from being deceptive (e.g., personal gain, preventing harm to others), and what she stands to lose if her deception is detected, or what losses would have resulted had she told the truth. We compare the computational model to data from human experiments in which we study whether people making deceptive strategy choices focus on inter-personal and situational factors, rather than on classical economic variables such as magnitudes and probabilities of potential losses and gains.

# Inner Speech and Task Switching in Adults

**Lucie Laurent**

MSHE, Besanon, Doubs, France

**Jean-Louis Millot**

Laboratoire de Neurosciences, Besanon, Doubs, France

**Patrice Andrieu**

Laboratoire de Neurosciences, Besanon, Doubs, France

**Valrie Camos**

Dpartement de psychologie, Universit de Fribourg, Switzerland

**Caroline Floccia**

School of Psychology, Plymouth University, United Kingdom

**fabien Mathy**

Universit de Franche-comt

**Abstract:** The aim of this study was to quantify the role of inner speech in categorical flexibility tasks. Research has shown that the suppression of inner speech in cognitive flexibility tasks leads to poorer performance by subjects. In contrast to previous studies, our objective was to offer tasks that enabled subjects to freely verbalize their reasoning, and to quantify its use. Using surface laryngeal electromyography to measure inner speech signals, we demonstrated that the more difficult the switching task, the greater the quantity of speech recruited by subjects when retaining and applying rules. Furthermore, our results show that relatively older adults ( $M = 50$  years) tend to rely on inner speech more than younger adults ( $M = 25$ ). We discuss the idea that speech acts as a support to executive functions and that it is needed to a greater extent with age due to declining executive performance.

# Enhanced Visual Processing in Perihand Space: Effects of Handedness

**Nathalie Le Bigot**

Leibniz Research Centre for Working Environment and Human Factors, Dortmund, Germany

**Marc Grosjean**

Leibniz Research Centre for Working Environment and Human Factors, Dortmund, Germany

**Abstract:** There is a growing body of evidence that visual processing is enhanced in perihand space. Some recent studies also showed that this enhancement is limited to the space around the right hand, at least in right handers. One explanation for these findings is that visual processing is facilitated at locations where action is more likely to occur. To test this notion, we had left and right handers perform a visual discrimination task under four hand-position configurations: Left only, right only, both hands, or no hands near the visual display. Results showed qualitatively different patterns for the two handedness groups. Visual sensitivity ( $d'$ ) was higher when only the left hand was near the display for left handers, whereas right handers performed better in the right- and both-hands conditions. These findings tend to confirm that visual enhancement is tied to the dominant hand, which is preferentially used for most manual actions.



# **The influence of cultural dispositions on the scene perception**

**Yoonkyoung Lee**

Pusan National University

**Bia Kim**

Pusan National University

**Yiwon Hyun**

Pusan National University

**Cheonwoo Shin**

Pusan National University

**Jaesik Lee**

Pusan National University

**HyunJung Shin**

Pusan National University

**Abstract:** We investigated whether people of individualistic disposition and those of collectivistic disposition perceive the natural scenes differently. It was hypothesized that the former pays more attention to the targets than the grounds, whereas the latter pays attention to the relation between the targets and the grounds as well as the targets themselves in the scene perception. In Experiment 1 where Korean individualists were contrasted with Korean collectivists, cultural disposition (individualism vs. collectivism), figure-ground relevance (naturalness vs. unnaturalness), and change of scene(no change vs. change of figures vs. change of grounds) were manipulated. The results of Experiment 1 showed that the collectivists respond to the unnatural scenes more sensitively than the individualists. The similar patterns were observed in Experiment 2 in which Korean as collectivists and European American as individualists were contrasted with each other. In sum, these results suggest that the collectivists who tend to see the scenes holistically respond to the unnatural scenes more sensitively than the individualists.

# **Framing Neuroethics: An Integrated-Unified Scientific Analogy (I-USA) Perspective**

**Sang Bok Lee**

Kangnam University

**Jonathan Jiseop Lee**

John F. Kennedy High School

**Abstract:** In mapping the hermeneutical scopes of neuroethics (Topics in Cognitive Science, 2(3), July 2010) the authors propose an Integrated-Unified Scientific Analogy (I-USA) model to delineate the complexities of moral dilemma we are facing critically. Based on meta-review (neuroethics in PubMed) and scientific analogy (Gentner, D., 2010) method, the authors conceptualize an I-USA model that optimally comprises multiple factors from the following disciplines: (1) neuroeconomics (2) neurogenetics (3) positive neuroscience (emotional activation, empathy and altruism, Tor Wagner, et al., 2002; Moritz de Greck, et al., 2011; Reuter, M., et al., 2010), (4) affective/cultural neuroscience (cultural values and differences, Wang, G., et al., 2011) (5) neurotheology (religious and spiritual, Ashbrook, J., 1989, 1992; Newberg, A. B., & Iverson, J., 2003; Lee, S. B., 1994, 1995), and (6) cognitive neuroscience (fMRI and neuroscientific, Tsukiura, T., & Cabeza, R., 2011; Cohen, M., 2001). The I-USA intends to optimally unify most of relevant factors for neuroethics.

# **How they pick out the answer in multiple choice questionnaire: Independent-self versus interdependent-self**

**Min-Seop Lee**

Seoul National University

**Jeong Ryu**

Seoul National University

**Dayk Jang**

Seoul National University

**Abstract:** Attentiveness to others is more likely to be a self-defining goal when the self is thought of as interdependent with others rather than independent of others. We predicted that interdependent self is more attentive in common ground than independent self in multiple choice questionnaire. In experiment 1, participants temporary self-construal was manipulated through a priming technique. As predicted, interdependence-primed participants were more likely than independence-primed participants to take the recipients knowledge into account and avoided providing redundant information in a self-administered questionnaire. Drawing on chronic differences in self-construal, experiment 2, replicated these findings with participants from independent and interdependent selves. The implications of these results for increasing understanding of behavioral priming effects in rich social contexts are discussed. Throughout, participants differential attentiveness to the common ground resulted in differential question order effects, raising important methodological issues for cross-cultural research; We expect to find culture differentiation between Korean to western people (USA or European etc.) as further study

# Information Structure, Alternatives, and Scalar Implicatures

Chungmin Lee

Seoul National University, Seoul, Republic of Korea

**Abstract:** Discourse occurs for information exchange in sentential utterances with information structure (IS). IS needs notions of Topic and Focus, Contrastive Topic (CT) and Contrastive Focus (CF). CT and CF are characterized in question answer conversation. A CT utterance comes from a QUD (Roberts 1996) with a Potential Topic (PT) consisting of a set of relevant alternative members, as a partial answer. Naturally, the question is not completely resolved and invokes a conventional scalar implicature. It is marked by a special intonation or morphemic marker and its implicature is not cancellable, hence conventional. The implicature generated is scalar with respect to the totality or qualitatively, in its DP or predicate PT.

CF is argued to invoke a closed set of disjunctive alternative possibilities. An alternative question is licensed if a pair or more of immediately relevant alternatives are available in the discourse context, as in Will you drink a tea or a coffee? It is semantically strengthened by the restriction exactly one disjunct holds (Pruitt & Roelofsens 2011) as in Inquisitive Semantics (InqSem). A correction answer is a typical CF: (1) Sue married Sam? Covertly, (2) Did SUECF marry Sam or did RITACF marry Sam? (From the immediately relevant alternatives set: Sue married Sam, Rita married Sam) (3) No, RITACF married Sam. All CF instances including pseudo-clefts, exhaustive focus, and CF-reduplication have (covert) AltQ. CF and AltQ in disjunction are likewise correlated. A conjunctively conceived CT is distinct from CF, conveying a scalar implicature, due to the unresolved partial information. CF and CT information structure is cognitively real. It is to be explored in terms of proposed possibilities in dynamic exchange.

# Observational category learning increases sensitivity to prototypical and correlational information

**Kimery Levering**  
SUNY Binghamton

**Kenneth Kurtz**  
SUNY Binghamton

**Abstract:** The traditional laboratory task of supervised classification learning tends to produce sensitivity only to information that discriminates between competing categories. Recent research broadening the study of category learning (e.g., inference learning) suggests that learners can be sensitive to structure beyond what is diagnostic. In previous work we interpreted these findings to reflect a generative (as opposed to discriminative) mode of learning and extended the phenomena to supervised observational learning of categories varying along two continuous dimensions. We now aim to demonstrate generative observational learning using binary dimensions. Categories were based on a uni-dimensional rule with within-category regularities in the form of a family resemblance structure (Experiment 1) or a perfect correlation between features (Experiment 2). Compared to classification learners, observation learners showed greater knowledge of both types of non-diagnostic structure. These results hold implications for how categories are learned and present a challenge to models grounded in discriminative category learning.

# The thinking behind decisions to spread disaster-related tweets

**Huaye Li**

Stevens Institute of Technology

**Rongjuan Chen**

Stevens Institute of Technology

**Yasuaki Sakamoto**

Stevens Institute of Technology

**Abstract:** Social media such as Twitter is playing a more and more important role in sharing information and coordinating disaster responses. However, after the 2011 Tohoku earthquake in Japan, Twitter users transmitted rumors about radiation and supplies, which Social media such as Twitter can play an important role in sharing information and coordinating disaster response. However, after the 2011 Tohoku earthquake in Japan, Twitter users transmitted rumors about radiation and supplies, which caused unnecessary panic. In the current work, we examined the thinking behind the decision to spread disaster-related tweets in Twitter. We showed subjects tweets related to the 2011 Tohoku earthquake. For each tweet, we asked them whether or not they would retweet it and the reason for their retweeting decision. Whereas thought of others (e.g., informative to others) played a major role when they decided to retweet, consideration of self (e.g., not interesting to me) dominated when they decided not to retweet. We further examined how factors such as source, familiarity, and importance of disaster-related tweets affect the retweeting decisions. Our results suggest techniques for minimizing the spread of false information during responses to disasters.

# Size Effect During Emergence of Symbolic Communication System Revealed by Agent-based Modelling

**Guanhong Li**

Japan Advanced Institute of Science and Technology

**Takashi Hashimoto**

Japan Advanced Institute of Science and Technology

**Abstract:** In the experiment proposed by Konno et al. (in press) about constructing a symbolic communication system, it had been confirmed that the emergence process includes three stages corresponding to three test conditions: no messaging, synchronous messaging and asynchronous messaging, respectively.

To investigate the mechanism of the emergence of symbolic communication system, an agent-based model is built using reinforcement learning.

The simulation result shows that: 1) for synchronous messaging condition, the number of rounds taken to get enough points to win tends to grow when increasing the size of selectable symbols; 2) for asynchronous messaging condition, the number of rounds before success remains stable even though more selectable symbols are offered.

We thus conclude that the size of selectable symbols has a negative effect on the formation of symbolic communication system under the synchronous messaging condition, while has no effect under the asynchronous messaging condition.



# Understanding human error detection as task interruptions

Simon Y. W. Li

Lingnan University, Hong Kong

**Abstract:** Experimental studies of human error detection have not made much progress since the 80s and 90s (e.g. Allwood, 1984; Rizzo et al., 1987; Sellen, 1994; Zapf et al., 1994). One of the reasons is because there is not an established methodological paradigm. A connection between interruption and error detection is proposed when an error is detected and corrected, it is similar to handling an interruption because the original task has to be suspended and resumed later. The connection is based on Trafton et al.s (2003) characterisation of interruption, which dissects interrupted activities into various measurable time-based components. A similar anatomy of the error detection process is proposed giving time-based measures such as detection lag, correction lag and resumption lag; error detection examples are discussed. The main contribution of the proposed characterisation is a methodological one postulating a set of dependent measures for future systematic investigations of the error detection process.

# **Assessment of childrens summarization ability: An alternative measure of reading comprehension**

**Chi-Shun Lien**

National Chung Cheng University, Chia-Yi, Taiwan

**Hung-Hui Chen**

National Chung Cheng University, Chia-Yi, Taiwan

**Abstract:** The ability to summarize information is important for understanding and remembering texts. Lacks of this ability, readers are not able to engage in the process of deleting, generalizing, and integrating those propositions of a text (Kintsch & van Dijk, 1978) and fail to comprehend the text. The purposes of this study were to evaluate a new diagnostic tool of text comprehension and scrutinize the development of childrens summarization ability. Sixty 3rd and 6th graders were recruited from an elementary school in Chia-Yi, Taiwan. All participants were administered 3 different versions of expository passages, which were deletion, generalization, and construction and asked to summarize them, respectively. A free-recall measure and a comprehension test were given after each summarization task. The findings indicated that the performance of summarization task was highly correlated to the performance of recall measure and comprehension test. In addition, there was developmental difference on the summarization task.

# Cross-linguistic Similarities and Differences in Causative Constructions

**Eunsuk Lim**

Hankuk University of Foreign Studies

**Kisuh Ahn**

Hankuk University of Foreign Studies

**Hee-Rahk Chae**

Hankuk University of Foreign Studies

**Abstract:** Language is learned on the basis of input and general cognitive mechanisms. We present a study on causative constructions in English, Korean and Portuguese. Causative constructions vary cross-linguistically, but we show that there is a possibility of some cross-linguistic generalizations among causative constructions in the three languages. We also point out some of the idiosyncratic constructions in each language, for example, the caused-motion construction in English. First, we divide causative constructions into three types: causation by command, causation by direct physical action and causation of emotion. Second, we group them into lexical, morphological and syntactic causation structures in terms of form-meaning pairings. Lastly, we examine the similarities and differences of the three languages focusing on their cognitive and psychological properties using a prototype model. We believe that this study will contribute to the understanding of the problems of common misuses of the causative constructions by children learning native languages as well as by adults learning foreign languages.

# Unpredictable Grunting in Tennis is More Distracting

**Ahnate Lim**

University of Hawaii at Manoa

**Alan Kingstone**

University of British Columbia

**Scott Sinnett**

University of Hawaii at Manoa

**Abstract:** Anecdotal and empirical evidence suggests that grunting in tennis can hinder an opponents performance. Here we addressed the question of whether a grunt is solely a distraction, or if instead it masks important auditory information carried by the dynamic multisensory event of the ball striking the racquet. Videos of a professional tennis player were played in silence, or included a grunt either before, during, or after impact, with the task being to judge the direction of the shot as quickly and accurately as possible. No differences were observed between the sound conditions, indicating that the adverse effects of grunting may be related to general distraction. This was replicated when manipulating the frequency of the grunts to occur 75% of the time, with overall performance improving, suggesting that the predictability of grunts is important, with more sporadic grunting leading to diminished performance.

# **Enhancing critical thinking and learning outcomes at the university: A pedagogical perspective**

**Stephen Wee Hun Lim**  
National University of Singapore

**Abstract:** Standard pedagogical strategies and assessment methods used for university undergraduate teaching normally involve knowledge transmission in classroom settings, student discussions, mid-term tests, student presentations on designated research papers and topics, and knowledge-based final exams that would not particularly help educators reach the goal of enhancing critical thinking and learning outcomes among students. This study explores dimensions beneath critical thinking and factors that would have contributed towards this goal, based on student evaluations gathered from a range of undergraduate modules taught by the author. Implications on general pedagogy and future directions will be discussed.

# Is it Possible to Train the Approximate Number System?

**Marcus Lindskog**

Uppsala University, Uppsala, Sweden

**Anders Winman**

Uppsala University, Uppsala, Sweden

**Peter Juslin**

Uppsala University, Uppsala, Sweden

**Abstract:** Humans are believed to be equipped with an Approximate Number System (ANS) that supports non-symbolic representations of numerical magnitudes. Acuity in the ANS is thought to progress developmentally from childhood to adolescence and to be related to mathematical achievement. However, from previous research, it is not clear whether it is possible to train this core ability. The present study investigates the effect of substantive corrective feedback on non-symbolic number comparisons, in an adult population, using a control group comparison design. The results indicate that even extensive training with feedback (1000 trials) has no effect on ANS-acuity. However, feedback induces a motivational effect on participants. These results suggest that some of the observed relationship between ANS-acuity and general math achievement may be mediated by a motivational factor and that a characteristic of the number sense may be its non-plasticity, at least for adults.

# **Cognitive Choreography in Mental Algebra Task**

**John Lindstedt**

Rensselaer Polytechnic Institute, Troy, New York, United States

**Wayne Gray**

Rensselaer Polytechnic Institute, Troy, New York, United States

**Abstract:** To understand the complexity of human cognition, it is important to account for how the various cognitive functions play off of one another during a complex cognitive task. This "cognitive choreography" was investigated previously in an attempt to tie cognitive functions to regional brain activity. The task used was a complex mental algebra task involving the combined use of various cognitive functions: problem representation, mental manipulation, memory retrieval, maintenance, and goal changing, as well as multiple modalities for input (auditory and visual) and output (manual and verbal). We present a faithful replication of this study's behavioral results (accuracy and latency), and discuss implications for cognitive functions in complex tasks.



# **Undergraduates' online search strategies and visual attention distribution**

**Wan-Yi Liu**

Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology

**Meng-Jung Tsai**

Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology

**Abstract:** The purpose of this study was to explore the relationship between the undergraduates online searching strategies and visual attention distribution by using eye-tracking technique. Thirty-one undergraduate subjects participated in the experiment in which they were asked to solve a task individually regarding the requirements for causing landslides. Participants visual attention distribution were measured with an eye-tracker. Students online searching strategies were assessed via a online information searching strategy inventory immediately after the online searching task. Results of this study revealed that students with higher prior knowledge can better identify relevant information from websites. And students with better problem-solving performance were found to spend more time reading relevant online information. In addition, students with better evaluation strategy were found to integrate the relevant information more efficiently. Furthermore, a significant gender difference was found in their visual attention allocated on the task problem and on the first page of search results.

# Action and affordances in nominal classifier systems

Marit Lobben

University of Oslo

**Abstract:** Although interdisciplinary research is increasingly being conducted on language and action within cognitive neuroscience, no one has looked into how the typological study of grammatical categories can be explained in terms of sensorimotor interactional learning. In this presentation cognitive semantic bases of nominal classifier systems are analysed in the light of the latest neuroscientific research within embodied cognition. Classifiers are found in unrelated languages in disparate areas of the world and are based on a restricted set of recurring semantic domains, e.g. an objects size, material, consistency, shape, dimension, inherent orientation, or function values. They therefore represent a justified area of study for general human cognition. It is demonstrated that embodied knowledge is necessary in order to use these grammatical components correctly in communication, relying on situatedness and perception in the conceptualisation process. Hypotheses of neural activation sites are proposed that accord with their modes of acquisition.

# **An ERP study on L2 grammatical aspect processing in Japanese**

**Shengyan Long**

Hiroshima University

**Yusaku Tsuyuguchi**

Hiroshima University

**Manami Sato**

Hiroshima University

**Hiromu Sakai**

Hiroshima University

**Abstract:** Second language learners process aspectual mismatch between verbs and adverbs differently from native speakers (Long et al., in press). Then, what about mismatch between grammatical aspect and adverbs? We explore into this question using event-related potentials (ERP), and compare our results with aspectual mismatch study with L1 Chinese speakers reported by Zhang & Zhang (2008).

Eleven Chinese speakers with high Japanese proficiency read Japanese sentences where aspectual property of adverbs either matched or mismatched with the progressive aspectual marker. Brain responses to mismatch were recorded by ERP.

Our results showed that violations of grammatical aspect elicited a 300-500ms posterior and left central negativity, which reflect resolving incompatibility of aspectual markers. This negativity is similar in its timing and distribution to the negativity reported in Zhang & Zhang (2008) for native speakers. This suggests that Chinese learners of Japanese process grammatical aspect as they do in their native language.

# **The Role of Embodiment on Childrens Memory Recall through LEGO Robotics Activities**

**Carol M. Lu**

Teachers College, Columbia University

**John B. Black**

Teachers College, Columbia University

**Seokmin Kang**

Teachers College, Columbia University

**Abstract:** Previous research has shown that embodied cognition involving body positions and physical enactment can enhance memory recall (Dijkstra et al., 2007; Scott et al., 2001). This study investigates the role of embodiment on childrens memory recall. Participants were fifth graders from two urban public schools attending a Robotics after-school program. They were asked to recall different LEGO Mindstorms sensors they used when constructing their robots and which of the physical science concepts were presented. Students were randomly assigned to one of the two learning conditions: science learning with robotics or science learning with robotics and embodiment. Students in embodied group outperformed the students in the robotics group. The results suggest that embodiment has remarkable potential to enhance childrens memory recall and performance on recognition tests, as compared to children without the embodied experience.

# **Social elements are not a must for preverbal infants learning in an interactive event**

**Yuen Ki Ma**

Department of Psychology, The University of Hong Kong

**Hiu Mei Chow**

Department of Psychology, The University of Hong Kong

**Jaclyn Yeung**

Department of Psychology, The University of Hong Kong

**Anna Wing Yee Ho**

Department of Psychology, The University of Hong Kong

**Chia-huei Tseng**

Department of Psychology, The University of Hong Kong

**Abstract:** Hamlin et al. (2007, 2010, 2011) showed that preverbal infants exhibit preference for animated figures in social events, and Chow, Tsui & Tseng (2011) demonstrated that infants can successfully associate visual (e.g. shape, color, and motion) cues with emotional cues (e.g. crying and laughing) which could be a prerequisite for making this social judgment. The current study examined whether infants ability of associated learning in complex sequences is limited to social-related situations only. After removing all socially relevant cues (eyes, facial expression, crying or laughing) from learning stimuli, we found 8 to 10-month-old infants could still associate agents with motion and neutral auditory outcomes. We also found the shape/color of a figure to be a more salient factor than the movement of the figure. We conclude that associated learning in animated interaction is not limited to specific social contexts in preverbal infants.

# **A possible source of phonological deficit in developmental dyslexia: Counter-evidence for universal phonological grammar**

**Norbert Maonchi-Pino**

Dept. of Developmental Cognitive Neuroscience, IDAC, Tohoku University - Japan Society for the  
Promotion of Science

**Yasuyuki Taki**

Dept. of Developmental Cognitive Neuroscience, IDAC, Tohoku University

**Satoru Yokoyama**

Dept. of Functional Brain Imaging, IDAC, Tohoku University

**Kei Takahashi**

Dept. of Functional Brain Imaging, IDAC, Tohoku University

**Annie Magnan**

Laboratoire d'étude des Mécanismes Cognitifs, Université Lumière Lyon 2 - Institut Universitaire de France

**Hiroshi Hashizume**

Dept. of Developmental Cognitive Neuroscience, IDAC, Tohoku University

**Jean Calle**

Laboratoire d'étude des Mécanismes Cognitifs, Université Lumière Lyon 2

**Ryuta Kawashima**

Dept. of Functional Brain Imaging, IDAC, Tohoku University

**Abstract:** Whether the phonological deficit in dyslexia ensues from impaired universal phonological grammar remains unanswered. Ten French dyslexic children were compared to chronological age-matched and reading level-matched controls. All were aurally-administered two syllable-counting tasks with di- and trisyllabic pseudowords (/am.al/; Exp. 1) or mono- and disyllabic pseudowords (/mal/; Exp. 2). We manipulated universal phonological sonority-related markedness within unattested onset clusters, from the phonotactically-unmarked clusters (/bal/) to the phonotactically-marked ones (/bal/). Sonority-related markedness was naturally inverted within intervocalic clusters. Across experiments, as sonority-related markedness decreased, response accuracy increased. Response patterns were similar in dyslexic children who were as accurate as both control groups, but systematically slower. Remarkably, an illusory epenthetic // vowel phonologically repaired marked clusters. Neither statistical properties nor acoustic-phonetic cues and reading skills accounted for the misperception/repair. Our results discard impaired phonological grammar as a source of dyslexics' phonological deficit and are discussed toward a deviant vs. delayed developmental course.

# **Aberrant sense of agency in patients with schizophrenia: confusion of temporal causality during intentional action**

**Takaki Maeda**

Department of Neuropsychiatry, Keio University School of Medicine

**Motoichiro Kato**

Department of Neuropsychiatry, Keio University School of Medicine

**Keisuke Takahata**

Department of Neuropsychiatry, Keio University School of Medicine

**Tsukasa Okimura**

Department of Neuropsychiatry, Keio University School of Medicine

**Hajime Asama**

Tokyo University

**Masaru Mimura**

Department of Neuropsychiatry, Keio University School of Medicine

**Abstract:** Self-disturbances in schizophrenia have been explained and studied from the standpoint of an abnormal sense of agency (SoA). We devised an original agency attribution task that evaluated explicit experiences of temporal causal relation between an intentional action and an external event, without any confounding from sense of body ownership. We demonstrated that an excessive SoA was observed in patients with prominent positive symptoms including delusion and hallucination. Moreover, those patients had a greater tendency to feel SoA even when external events were programmed to precede their action (backward causation). Therefore, patients felt both forward and backward exaggerated causal efficacy during the intentional action (Maeda et al., Psychiatry Research 2012). On the other hand, patients with predominant negative symptoms showed diminished SoA. Confusion in the experience of temporal causal relations between the self and the external world may underlie self-disturbances in schizophrenia. Aberrant SoA could be a fundamental vulnerability marker for schizophrenia.



# **Student perspectives on critical and other thinking skills: Some cultural similarities and differences**

**Emmanuel Manalo**

Waseda University

**Takashi Kusumi**

Kyoto University

**Masuo Koyasu**

Kyoto University

**Yasushi Michita**

The University of the Ryukyus

**Yuko Tanaka**

Stevens Institute of Technology

**Abstract:** The purpose of this study was to explore the views of students from different cultural backgrounds about good thinking skills, including those they perceive as required in their courses of study. Focus group interviews were conducted with undergraduate students from Kyoto (n = 8) and Okinawa (n = 7) in Japan, and from Auckland (n = 8) in New Zealand. Analysis of the students responses revealed commonalities in views about what good thinkers possess: these included many qualities associated with critical thinking. However, when specifically asked about the meaning of critical thinking, many of the students from Okinawa were uncertain, and the students from Auckland also referred to attributes not commonly associated with critical thinking such as intuition and positive thinking. The results suggest a need for more explicit instruction in critical thinking skills development as well as for greater clarity about thinking skills expectations in tertiary level courses.

# **The role of exploratory decision-making in enhancing episodic memory**

**Doug Markant**

New York University

**Sarah Dubrow**

New York University

**Lila Davachi**

New York University

**Todd Gureckis**

New York University

**Abstract:** Studies of episodic memory are typically passive in that the learner can't control the sequence and timing of stimulus presentation. This approach obscures how people alter their learning by choosing to study some items over others, a decision-making process with many implications for memory. Voss (2011) showed that when people made decisions about what to study their memory was better than when yoked to the decisions of another person, and they derived a specific benefit from revisiting recent items. The source of this benefit is unclear, however, since previous studies have not separated the effect of executing study decisions from that of resulting personalized study opportunities. By varying the decision-makers degree of control over the study experience, we show that people still benefit from making study decisions even when unable to revisit items or choose what material to study next, highlighting the role of decision-related processes in improving episodic encoding.

# Visual processing deficits due to HIV: Evidence from temporal order judgments

**Liron Marotz**

Department of Psychology, University of Hawaii at Manoa

**Scott Sinnett**

Department of Psychology, University of Hawaii at Manoa

**Cecilia M. Shikuma**

John A. Burns School of Medicine, University of Hawaii at Manoa

**Abstract:** Numerous empirical reports have demonstrated cognitive deficits after contracting HIV. While the majority of these findings have incorporated traditional neuropsychological tests (e.g., California Verbal Learning Test, Trail-Making), very little work has investigated visual temporal processing, and how spatial attention is distributed, and potentially adversely affected, by HIV infection. In this preliminary study we used a temporal order judgment task that required participants to accurately judge the successiveness of laterally presented asynchronous events. Attention was spatially directed by a peripheral cue prior to each trial, in order to assess deficits in orientation to reflexive cues. The findings demonstrated that people with HIV required more time to accurately judge temporal order when compared with a sample without HIV. However, the cue captured attention equally regardless of HIV status. The broader conclusion is that HIV may lead to impairment in visual temporal processing, without adverse effects on the distribution of spatial attention.

# Collaboration for Building Sustainable Knowledge

**Hiroyuki Masukawa**

Shizuoka University

**Ikuo Endo**

Ito City Higashi Elementary School

**Abstract:** We are investigating how collaboration leads to deeper understanding. We targeted sixth-grade math-classes. In the lesson, the teacher asked "What's the number of games of a round-robin football tournament?" In the first half, the children solved problem in groups, and in the second half, they reported their solutions and then the teacher explained the correct answer. We analyzed the protocol-data of two groups. One group discussed how to solve problems while sharing and comparing multiple levels of abstraction: conceptual, calculative, diagram, or concrete. The other group discussed only the concrete abstraction. At the end, both groups' children checked "I understand" on the self-evaluation sheet. One month later the teacher gave the retrospective test to children. The results were that one group could describe how to solve problems and the other group's ideas in detail, while the other group could provide only the answer and explain the other group members name.

# Increases in Childrens Semantic Organization Predict Category-based Reasoning

**Bryan Matlen**

Carnegie Mellon University

**Karrie Godwin**

Carnegie Mellon University

**Anna Fisher**

Carnegie Mellon University

**Abstract:** Prior research indicates a protracted developmental course in inductive reasoning based on category information (Fisher, Matlen, & Godwin, 2011). A possible explanation for the development of adult-like induction is the gradual reorganization of semantic-knowledge. To explore this possibility, we presented preschoolers (N=43), kindergarteners (N=22), first-graders (N=10) and adults (N=20) with blocks representing semantically-similar (e.g., chick-hen), physically-similar (e.g., lamb-swan) and same-habitat (e.g., whale-octopus) animals: Participants placed animal pairs on a board representing a zoo according to the degree of semantic-similarity. The distance between animals was taken as a measure of how closely participants represented the semantic concepts. Preschoolers exhibited considerable variability and were worse at discriminating between semantically-similar and non-semantically-similar dyads compared to older children and adults. Preschoolers scores on the semantic-space task were also correlated with performance on an induction task ( $r=0.46$ ;  $p=0.002$ ). These results are consistent with the possibility that the organization of semantic-knowledge underlies advances in category-based reasoning.

# Efficiency of the Cognitive Bias in Grammar Acquisition

**Ryuichi Matoba**

Toyama National College of Technology

**Makoto Nakamura**

Nagoya University

**Shingo Hagiwara**

Toyama National College of Technology

**Satoshi Tojo**

Japan Advanced Institute of Science and Technology

**Abstract:** This study investigates efficacy of cognitive biases for grammar acquisition. We present a simulation of virtual block world where a pair of a parent and an infant agent resides. When a parent takes an action and utters corresponding sentence, an infant guesses the meaning of the utterance from the situational change caused by the action. The infant tries to acquire a compositional grammar through an amount of utterances, and the grammar would be polished through generations. Through this process, we focus on how the symmetry and mutual exclusivity biases help this grammar refinement. We expect that the symmetry bias works to combine a situational change and an utterance, and that the mutual exclusivity bias contributes to solve the ambiguity between a situational change and an action; when the infant finds an utterance is already combined with a different meaning, he/she ignores the new pair of utterance/meaning by the latter bias.

# **CyclingMusic & CyclingMelody: A System for Enriching Scenery Experience in Cycling by Real-Time Synaesthetic Sonification of Passing Landscape**

**Masaki Matsubara**

University of Tsukuba, Tsukuba, Ibaraki, Japan

**Satoshi Kuribayashi**

Keio Research Institute at SFC, Fujisawa, Kanagawa, Japan

**Haruka Nukariya**

Keio University, Fujisawa, Kanagawa, Japan

**Yasuaki Kakehi**

Keio University, Fujisawa, Kanagawa, Japan

**Abstract:** When riding on a bicycle, many people enjoy looking over the landscape passing by. This study discusses how to enrich scenery experience in cycling. Since the scenery experience is based on an embodied perception, it is known as one of the tacit knowledge, thus enriching it is difficult. In order to overcome this problem, we focused on the feature of auditory sense. Human audition can accommodate dynamic information better than vision, and data sonification presumably represents time-series data more efficiently than data visualization. We constructed two systems named CyclingMusic and CyclingMelody that translate the perceived movement of the scenery and other visual impressions, such as hue, lightness and colorfulness, into music. The continuously changing view is captured with a web-cam and translated into MIDI events that are replayed instantaneously. This allows for a reflection of the visual impression, adding a sound dimension to the visual experience and deepening the state of perception. A practical experiment in cycling shows sonification makes participants deepen the spatial experience.



# Does a humanoid robot in front of you activate your mirror neuron system?

**Goh Matsuda**

The University of Tokyo

**Kazuo Hiraki**

The University of Tokyo

**Hiroshi Ishiguro**

Osaka University

**Abstract:** In the present study, we attempted to evaluate the human-likeness of a humanoid robot named Robovie by using near infrared spectroscopy (NIRS). Since the activity of the human mirror neuron system (MNS) is believed to reflect the perceived human-likeness of observed agents, we compared MNS activity during observations of an action performed by a human and Robovie. Seven males and ten females participated. There were four observation conditions such as live-human, live-robot, video-human and video-robot. In addition the participants executed the same action by their selves to confirm the location of motor-related brain areas. MNS activity was the largest in the live-human condition in line with previous studies. Our interesting finding was, however, that MNS activity was larger in the live-robot condition than the video-human condition. This suggests that we perceive a humanoid robot in front of us as a more human-like being than a videotaped human.

# The effects of exposure sequence and duration on mere exposure effect

**Ken Matsuda**

Yamaguchi University

**Eriko Sugimori**

Yale University

**Takashi Kusumi**

Kyoto University

**Abstract:** This study investigated how exposure sequence and a delay influenced mere exposure effect. We used neutral random shapes and controlled exposure sequence (heterogeneous (spaced) and homogeneous (unspaced)), exposure frequency (3, 6, and 9 times), and intervals between learning and judgment (5 min and 1 week). Fifty five participants were exposed to each stimulus, and 5 min and 1 week later, were asked to rate preference, familiarity, novelty, nostalgia, using a 5-point scale as well as recognition of old and new items. The result shows that, in judgments about preference, familiarity, and nostalgia, scores for the stimuli in homogeneous exposure condition rose after 1 week, although scores in heterogeneous exposure condition didn't change with a delay. In homogeneous and high frequent presentation, nostalgia occurred by an interval of one week, and that nostalgia raised stimulus preference and familiarity.

# The Computational Process of And-type Conditionals in Japanese

Michinao MATSUI

Osaka Health Science University, 1-9-27, Temma, Kita-ku, Osaka-Shi, Osaka, Japan

**Abstract:** Japanese has four types of conditional sentences: Nara-type, Reba-type, Tara-type and the natural consequence expression using To (And-type). Among these four types, only And-type conditionals cannot express counterfactuals. Masuoka (1993) says that the Japanese conjunction To (And) is an index which calculates the relevance between elements. This paper explains this phenomenon based on Relevance Theory (Sperber et al. 1986). Cognitive relevance between the information A and B is defined by correlation coefficient (CR1) or by Loose Symmetry Model (CR2, Shinohara and Nakano, 2007).

(1)  $CR1 = P(AB)/P(A)kP(aB)/P(a)$  (2)  $CR2 = (P(AB) + P(Ab)P(ab)/P(b)) / (P(AB) + P(Ab) + P(AB)P(aB)/P(B) + P(Ab)P(ab)/P(b))$

In the expression of counterfactuals, the relation  $P(a) = P(ab) + P(a\bar{b}) = 0$  is trivial. Therefore, CR1 is computable only when the coefficient k which means the frame range of anti-factual situation equals zero. This computation leads the result of impossibility of And-type counterfactuals. On the other hand, CR2 cannot explain this impossibility because CR2 is always infinite. Therefore, CR1 is the better definition than CR2.

# Connecting input filtering and selection in language evolution

**Luke Maurits**

University of California, Berkeley

**Tom Griffiths**

University of California, Berkeley

**Abstract:** Previous work has demonstrated the formal equivalence of simple models of cultural evolution, based on iterated Bayesian learning with symmetric Dirichlet priors, with neutral models of biological evolution. This demonstration was profitable in allowing the use of existing mathematical results from population genetics to be applied to questions of language change. We extend this work by exploring parallels between more complicated models of cultural evolution, featuring input filtering, and models of biological evolution with selection. We investigate the use of analytic expressions for steady-state distributions for predicting the outcomes of cultural learning models, and apply these methods to the explanation of historical word order change via input filtering. In addition to these practical outcomes, our work relates to the broader question of just how similar linguistic and other cultural change is to Darwinian evolution in biology, and how valid it is to speak of cultural traits being selected for.

# When Lexical Development does not Spurt; the Case of Williams Syndrome Children

Julien Mayor

University of Geneva, Geneva, Switzerland

**Abstract:** Williams Syndrome (WS) children possess relatively large vocabularies when compared to their other, impaired, cognitive skills. However, their language acquisition is delayed, their vocabulary does not spurt, categorisation skills are weak and they do not respond taxonomically.

We mimic WS categorization impairments by hindering the formation of visual categories in a model of early word learning (Mayor & Plunkett, 2010). In the absence of lesions, the model accounts for the emergence of taxonomic responding and displays a vocabulary spurt driven by the formation of categorical representations. In contrast, when categorisation is impaired, lexical acquisition is delayed, a vocabulary spurt is absent and word-object associations are not generalised. However, through repetitive labelling events, the system is still able to acquire a large "proto-lexicon" by gradually attaching several exemplars of a category to their appropriate sound pattern in an associationist mechanism, thereby leading to the formation of a surprisingly large vocabulary.

# Self-directed information selection aids learning of logical rules

**John V. McDonnell**

New York University Dept. of Psychology

**Devin Domingo**

New York University Dept. of Psychology

**Todd M. Gureckis**

New York University Dept. of Psychology

**Abstract:** In self-directed learning tasks, participants can control the sequencing and timing of information presentation. In contrast, the existing literature on category learning has focused on passive learning paradigms, wherein information is presented to the learner randomly or in a pattern determined in advance by the experimenter. To explore the impact of self-directed learning on categorization performance, we compared passive and self-directed learning in the seminal Shepard, Hovland, and Jenkins (1961) category learning tasks. Participants learned by actively querying the category membership of individual exemplars in an array or by passively viewing labeled examples. We found that active learners exhibited significantly faster learning than passive learners. In addition, the benefits of self-directed learning were not uniform, but varied as a function of the category structure. Our results suggest that differences in interaction with learning materials can alter the difficulty of learning problems, independent of the abstract structure of the underlying rule.

# Making and breaking procedural conventions in dialogue

Gregory Mills

University of Edinburgh

**Abstract:** A key problem for models of dialogue is to explain how co-ordination is established and sustained. Existing accounts emphasize the importance of interaction, demonstrating how collaborative feedback leads to more systematized, stable, arbitrary and partner-specific referring conventions.

However, in addition to conventionalizing referring expressions, recent work demonstrates how interlocutors also rapidly establish procedural conventions for resolving sequential and temporal co-ordination problems in the interaction. It is unclear, however, whether interlocutors associate these procedural conventions with specific conversational partners.

To address this question, we report a collaborative, 3-participant, computer-mediated task which presents participants with the recurrent co-ordination problem of ordering their actions and utterances into a single sequence. Artificially generated clarification requests are inserted into the dialogue, that appear, to each participant, as if they originate from either of the 2 other participants. We argue that participants' responses to these clarifications provide evidence of interlocutors associating procedural conventions with specific partners.



# **Expanding childrens perspectives on art work by robot participation**

**Masaki Miyake**

Aichi Shukutoku University, Nagoya, Aichi, Japan

**Tomoki Hirano**

Museum Plus

**Naomi Miyake**

University of Tokyo

**Abstract:** In order to promote skills of art appreciation of novices, we have been investigating the effects of dialogue. Our practice encourages each participant to take her/his own viewpoint to find something meaningful to her/him in the artwork, express it verbally to be shared by others. The effect has been recognized widely experientially at schools and museums, yet the cognitive mechanisms of this facilitation have not been fully investigated. In this study we have used a remotely operated robot to provide a group of 10 to 15 kids of 8 to 10 year olds with a perspective not shared by them, to see whether such intervention could expand the groups scope of appreciation. The robot was accepted as a friend with its own ideas, and we identified cases of such facilitation. We will report the details of these successful cases, to explore the mechanisms of this method.

# **The effect of harmonization between word meaning and typography impression on implicit memory**

**Kozue Miyashiro**

University of Tsukuba, Tsukuba, Ibaraki, Japan

**Etsuko T. Harada**

University of Tsukuba, Tsukuba, Ibaraki, Japan

**Abstract:** This study examined the effect of harmonization between word meaning and typography impression on implicit memory. In a preliminary experiment, we asked undergraduate students the degree of harmonization (how much does the typography match its word meaning?) for 153 words, showing them two lists (hiragana and kanji) with three different typographies. Based on the results, an incidental memory experiment was executed with three experimental factors: the scripts at learning (hiragana, kanji, or no learning), with/without the harmonization at learning, and the typography at testing (same as learning, different neutral typography), as within participant factors. Both a word completion task and an old/new recognition test were executed using only hiragana. The result showed that harmonization facilitated implicit memory with hiragana-learned words, but implicit memory was reduced for kanji-learned words. Morphologically-based processing occurred when word meaning and typography were harmonized. Conceptual implication of this harmonization effect will be discussed.

# **Short-term memory for tonal and verbal information: Comparison with absolute and non-absolute pitch possessors**

**Shiho Miyazawa**

Tokyo Woman's Christian University

**Akihiro Tanaka**

Tokyo Woman's Christian University

**Takehiko Nishimoto**

Waseda University

**Abstract:** This study examined the difference in the storage of pitch and phonological information in absolute pitch (AP) and non-AP (NAP) possessors. In a recognition task using musical tones (pitch information), speech sounds (phonological information), and visual patterns, participants were asked to retain two stimulus sequences. In the same type condition, the nature of the first and second stimulus set was different (e.g., one sequence was musical tones and the other was speech sounds). In the different type condition, the nature of the two sequences was the same (e.g., both sequences were musical tones). We found that, in NAP possessors, the recognition rate of musical tones and speech sounds in the different type condition was higher than in the same type condition. In AP possessors, however, the recognition rate of musical tones revealed no difference between these two conditions. These results suggest the use of different strategy in retaining musical tones between AP and NAP possessors.

# **Behavioral priming contributes to the subsequent recognition performance**

**Kiyohumi Miyoshi**

Kyoto University

**Hiroshi Ashida**

Kyoto University

**Abstract:** The present study assessed the impact of behavioral priming in several tasks (semantic, color-decision and phonological tasks) on the creation of new explicit memory. In this study, Japanese words written in Kanji were repeated once in an incidental encoding task and explicit memory for the words was unexpectedly tested afterwards. The words associated with more behavioral priming in the semantic task were better recognized in the subsequent explicit memory test than were the words with less priming. Moreover, the subjects who demonstrated greater priming in the semantic task performed better in the explicit memory test. These results suggest that behavioral priming enhances the creation of explicit memory.

# **How much do you trust me? Economic decision-making and ingroup and outgroup membership**

**Rosalba Morese**

Center of Cognitive Science, Department of Psychology, University of Turin, Italy

**Daniela Rabellino**

Center of Cognitive Science, Department of Psychology, University of Turin, Italy

**Angela Ciaramidaro**

Center of Cognitive Science, Department of Psychology, University of Turin, Italy; Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Goethe-University, Frankfurt/M

**Marco R. Elena**

Center of Cognitive Science, Department of Psychology, University of Turin, Italy

**Francesca M. Bosco**

Center of Cognitive Science, Department of Psychology, University of Turin; Neuroscience Institute of Turin, University of Turin, Italy

**Rosalba Rosato**

Department of Psychology, University of Turin; Unit of Cancer Epidemiology, San Giovanni Battista Hospital - Turin, University of Turin e CPO Piemonte

**Bruno G. Bara**

Center of Cognitive Science, Department of Psychology, University of Turin; Neuroscience Institute of Turin, University of Turin, Italy

**Abstract:** Cross-cultural studies suggest that culture can influence human decision making, especially in economic contexts, but little is known about economic decision-making comparing ingroup and outgroup membership. We studied trusting behavior in two different cultures: the Chinese one, example of collectivist culture, and the Italian one, example of individualistic culture (N=25 Italians and N=25 Chinese). Using a modified version of the Trust Game with three levels of trust (20 or 40 or 60 Monetary Units), we investigated behavioral differences in decision-making based on trust levels in ingroup (Player A and B same nationality), and outgroup condition (Player A and B different nationality). Data analysis showed that both groups tend to trust more when the amount of monetary units decreases. Furthermore, results revealed no statistically significant differences in the trusting behavior between ingroup and outgroup factor for both cultural groups.

# Using syntactic priming to facilitate the language production of novice Japanese EFL learners

Miwa Morishita

Kobe Gakuin University

**Abstract:** Syntactic priming is the tendency for language users to produce a particular syntactic structure soon after listening to or reading that structure. Previous research using Japanese EFL learners found that intermediate learners were affected significantly more by syntactic priming than novice learners in a sentence completion task mainly because the grammatical/syntactic information might not be fully represented in the mental lexicon of the latter (Morishita, 2011; Morishita, Sato, & Yokokawa, 2010). Morishita and Yokokawa (2011) found, however, that the more novice learners were exposed to certain sentence structures, the stronger the priming effects became, where repeated exposure might aid implicit learning through imitation and practice. This presentation will discuss the results of the above research, as well as those of an ongoing experiment based on a scripted interaction task, in order to assess the possibility of using syntactic priming to facilitate the language production of novice Japanese EFL learners.

# Speed reading training and visual span

Aiko Morita

Hiroshima University, Higashi-Hiroshima, Hiroshima, Japan

**Abstract:** The current study investigated the span training effect on size of visual span and on reading speed. The visual span for reading refers to the range of letters, formatted as in text, that can be recognized reliably without moving the eyes. It has been hypothesized that the size of the visual span imposes a fundamental limit on reading speed. However, the relation between visual span expansion and reading speed increase has not been clarified. The present study conducted two types of training: One was the word recognition training. Eight undergraduate students were trained to recognize three or four words simultaneously. The other was the saccade training. Ten students were trained to read texts with larger saccades. The results showed that participants reading speed increased by 30% in both training groups. The word recognition training was more effective than the saccade training on visual span expansion, especially on left visual field.



# **Analytical method of Japanese folk tale for story generation**

**Hitoshi Morita**

University of Nagasaki, Nagayo-cho, Nagasaki Pref., Japan

**Abstract:** The purpose of this research is to examine whether a new story can be created by analysing folk tales. The tools of folklore like the tale type and the motif indexes are used for the analysis. The original text of this research is "Momotaro" (Peach Boy). This story had taken a rejuvenation motif until Edo period (19C). I examine 4 factors converted from the rejuvenation to the abnormal birth motif: 1)Prosperity of popular literature, 2)Enactment of elementary school textbook, 3)Diversion of folklore to nationalism, 4)Fight uplift animation in World War II. Next, I show the structure of the story from the following points. Process of the birth, Hero's name, Attendants, Motivation of departure, Strategy of demon extermination, Return to home. I clarified how this story was influenced from the society. Then, I examine to generate the text by analyzing "Momotaro", and compared with the Japanese 5 great folk tales.

# Facilitation and inhibition in the spatiotemporal template for detecting target ring

**Masayoshi Nagai**

National Institute of Advanced Industrial Science and Technology (AIST)

**Patrick J. Bennett**

McMaster University

**Allison B. Sekuler**

McMaster University

**Abstract:** Using the classification image technique (Ahumada, 1996), this study investigated the nature of spatiotemporal facilitatory and inhibitory templates when detecting a target ring. The stimulus consisted of fifteen temporal frames of five spatial elements: one-center disk and four-surrounding concentric rings. When a target was present, the middle brighter ring than the screen background always served as the target, and the rest of the elements were darker than the background. Random luminance noises were changed on every stimulus frame. On a non-target trial the non-target pattern was presented across all 15 frames. On a target trial the target pattern was presented at the middle three temporal frames. Results showed a transient facilitatory response at the target location around the target presentation and an inhibitory response from the outer elements even after the targets actual presentation. These results suggest that spatiotemporal templates are not simple copies of the spatiotemporal target pattern.

# **Analysis method focusing on repeated and not repeated verbalizations during human interface operation**

**Yukari Nagai**

School of Knowledge Science, Japan Advanced Institute of Science and Technology, Japan

**Saori Noda**

HMI R & D Department, DENSO CORPORATION

**Georgi V. Georgiev**

Department of Mechanical Engineering, Kobe University, Japan

**Toshiharu Taura**

Department of Mechanical Engineering, Kobe University, Japan

**Deny Willy**

School of Knowledge Science, Japan Advanced Institute of Science and Technology, Japan

**Abstract:** Humans interact with and operate different devices and ultimately form impressions, such as like and dislike. In previous research, humans impressions of devices with regard to various operations, were not considered. Human impressions are difficult to capture, as they tend to be ambivalent. This study investigates an analysis method that focuses on repeated and not repeated verbalizations during human interface operation. The study included an experiment in which six subjects were asked to freely verbalize their impressions of device operation of two devices in two vehicles. The words from the verbal protocols are analyzed using concept networks that are formed of associative word pairs. This analysis outlines a complex associative layer of impressions that the verbalizations are drawn from. The result suggests that the impressions from both verbalization types repeated and not repeated identify undefined issues in the operation of interfaces. Thus, methods for resolving these issues can be formulated.

# **The effects of listeners familiarity with a talkers voice on the speech recognition in noisy condition**

**Chikako Nagaoka**

Graduate School of Human and Environmental Studies, Kyoto university, Kyoto, Japan

**Naoshi Hiraoka**

Academic Center for Computing and Media Studies, Kyoto university, Kyoto, Japan

**Shintaro Funahashi**

Kokoro research center, Kyoto university, Kyoto, Japan

**Abstract:** The present study examined whether listener's familiarity with a talker's voice improved speech recognition in noisy condition. Subjects were asked to perform word and sentence intelligibility task in noisy condition under following three conditions; (1)subjects heard familiar voice and were informed who was speaking (explicit-familiar condition), (2)subjects could hear familiar voice but were not informed who was speaking (implicit-familiar condition), (3)subjects heard unfamiliar voice and were not informed who was speaking (novel condition). We used subject's friend's voice as a familiar voice. As a result, the percentage of correct answer of explicit-familiar condition and implicit-familiar condition were better than that of novel condition, but there was no significant difference between explicit-familiar condition and implicit-familiar condition. Present results revealed that the familiarity with a talker's voice improved speech recognition in noisy condition, while knowing who was speaking did not influence speech recognition.

# **A Comparison of Experienced, Novice Counselor and Non-counselor in Recall of Client-Presented Information in Therapeutic Interview**

**Chika Nagaoka**

Kyoto University

**Sakiko Yoshikawa**

Kyoto University

**Tomoko Kuwabara**

Kyoto University

**Yasuhiro Oyama**

Kyoto University

**Chiriho Hatanaka**

Ritsumeikan University

**Motoki Watabe**

Waseda Institute for Advanced Study

**Masashi Komori**

Osaka Electro-Communication University

**Abstract:** This study examines whether counselors memory of client-presented information varies qualitatively according to the number of years of counseling experience. This study develops a methodology to measure the amount of counselors recall memory obtained from a free recall task after watching videotaped stimulus interviews. Four experienced counselors, seven novice counselors, and twelve non-counselors watched videotaped stimulus interviews and then wrote freely everything they could recall about what the client had said in the interview. Interview transcripts were employed as evaluation criteria. Independent coders judged the parts of the transcript to which the recalled items corresponded. The results indicate that the experienced counselor group scored the highest in recalling client-presented information and that recalled contents differed among the participant groups. Implications of the steps to gain counselor expertise are discussed.

This study was supported by Grant-in-Aid for Scientific Research (No. 20530569) from the Japan Society for the Promotion of Science.

# The Narrative Structure of Nostalgia Cognition and Film

**Yuya Naito**

HOSEI University Graduate School

**Akihito Kanai**

HOSEI University

**Abstract:** This study inspects re-defining nostalgia and re-modeling the narrative structure of nostalgia cognition. To re-define nostalgia, we use mainly three kinds of nostalgia by Davis (1979), which consists of simple nostalgia reflexive nostalgia and interpreted nostalgia. The narrative structure of re-defined nostalgia cognition is consisted with the three kinds of nostalgia. The re-defined nostalgia includes not only affirmative but also negative factor. This study includes analyzing of films and experiments, which are investigating whether the three kinds of nostalgia is cognized by using the programmed materials. The materials are able to change the past and current scenes of films controlled by the computer programming. These analysis of films and experimental results suggest that some films can enable a viewer experience various kinds of nostalgia. On the other hand, other films can enable a viewer experience only a factor of nostalgia or not at all.

# TCieX: An Approach toward Communicating Weight through Pseudo-Haptic Feedback Mechanisms

**Kumiyo Nakakoji**

Software Research Associates, Inc., Japan

**Yasuhiro Yamamoto**

Tokyo Institute of Technology

**Abstract:** The notion of weight is useful to represent not only the mass of an object, but also the physical feedback for a remote interaction as well as the importance of a concept. Existing human-computer interaction design has communicated weight primarily through the three means: symbolic representation ("240g"), kinesthetic interaction (by wearing a physical actuator), and haptic feedback (by applying low-frequency stimuli to the forearm). Our approach uses a fourth means that uses the visual and audio interaction to communicate weight, stiffness, or viscosity through pseudo-haptic feedback mechanisms. Pseudo-haptic feedback stimulates haptic sensations without using a haptic interface but using dynamically changing visual and audio representations in terms of the hand movement.

We have built TCieX (Touch-Centric interaction embodiment eXploratorium) as a collection of interaction test suites, with which a user produces a variety of combinations of temporal, visual, and auditory representations for different types of simple object movement through different C/D (Control/Display) ratios and mapping profiles. By using TCieX, an interaction designer explores a design space of visual and auditory representations to produce a desired effect of pseudo-haptic sensations.



# **The effects of paralinguistic cues in a teachers responses to the students utterances in a moral class**

**Keiko Nakamoto**

Bunkyo University

**Ayumi Nishiyama**

Bunkyo University

**Abstract:** This study investigated the characteristics of a teachers paralinguistic cues in a moral class in a Japanese elementary school. In the class, the children were told a story on friendship and were asked to give their views on it. In the study, the teachers responses (in particular, thats right, I see, and great) to the childrens utterances were analyzed in terms of paralinguistic characteristics. The results show that the response thats right was used fewer times in the learning activities during which the teacher asked the children to think deeply than the other activities. Moreover, during such activities, the teachers responses were made with a monotonous intonation and in a relatively low voice. Correspondingly, the children paused more often when giving their views during this activity as compared to the other activities. These results suggest that paralinguistic cues are attached consciously or unconsciously to the teachers responses, and they affect the behavior of the children during learning activities.

# **An agent-based model for the emergence of creoles**

**Makoto Nakamura**

Nagoya University

**Shingo Hagiwara**

Toyama National College of Technology

**Satoshi Tojo**

Japan Advanced Institute of Science and Technology

**Abstract:** This study investigates the emergence of creoles through computer simulations of language contact in an artificial community. Creoles are full-fledged new languages which infants growing in a multilingual community acquire as their native language. We define a creole in the agent community as a new language dominating the community. We have developed an agent-based model with Kirby (2002)'s Iterative Learning Model where agents are connected with neighbors on a social network. The infant agents acquire a compositional grammar, hearing utterances from their parents and neighbors. Assuming language exposure plays a key role for creolization, we introduce a parameter of exposure to neighbors to the model. Language groups result from clustering similar languages, each of which is spoken by an agent. Experimental results show a certain degree of the exposure is necessary for creolization. Further investigation suggests that the structure of the social network affects that of agents' grammars.

# Generating new product ideas by assemblage of different product components

**Jun Nakamura**

Kanazawa Institute of Technology

**Yukio Ohsawa**

University of Tokyo

**Abstract:** This research is focused on clarifying the effect of providing information, where synthesis of different product components provides with an implication for generating new product ideas.

We hypothesize that generating new product ideas is induced by unexpected assemblage of presented different product components. This hypothesis was evaluated by analyzing playing logs of a developed Web based tool to enjoy assemblage of elements whereof categories are consisting of three components, i.e., purpose, function and material.

We defined generating new product ideas is associated with possible architectural compatibility. As a result of 11 players, we found that generating ideas tends to be either visible products or invisible products as a service, depending on ones attention to purpose at initial stage, e.g., the more purpose is considered, the more ideas are relatively visible products. In addition, the meaning of purpose has been interpreted as means to new product ideas.

## **Postal Addresses as an Assay of Cultural Cognition**

**Hiroko Nakamura**

Nagoya University

**Hiroshi Yama**

Osaka City University

**Gary L. Brase**

Kansas State University

**Nasriah Zakaria**

University of Science, Malaysia

**Yoshiko Arai**

Osaka City University

**Norhayati Zakaria**

The University of Wollongong, Dubai

**Shafiz A. Mohd Yusof**

The University of Wollongong, Dubai

**Jun Kawaguchi**

Nagoya University

**Abstract:** Present study investigated the cultural difference in explicit and implicit semantic processing in three countries: Japan, Malaysia and United States. In the experiments, we conducted both explicit and implicit priming tasks with state and city names as stimuli. Participants were required to judge whether the target city or state name was a real one or a fake one. The results indicated direction between prime and target (state-to-city or city-to-state) had no effects on amount of priming. The cultural difference was significant only in the explicit priming task: amount of priming was larger in Malaysian participants compared to both American and Japanese participants in SOA700, while there were no cultural differences in SOA200. These results imply Malaysians are engaged in more context dependent cognition; in other words, participants consciously use a prime as a context cue to anticipate a target item. And this cultural difference was disappeared in implicit semantic processing.

# **Changes in social motivation, and learning strategies, in PBL.**

**Yoshifumi Nakanishi**

Faculty of Education, Mie University, Tsu, Japan

**Takatoyo Umemoto**

Graduate School of Education and Human Development, Nagoya University, Nagoya, Japan

**Kenshiro Tanaka**

Graduate School of Education and Human Development, Nagoya University, Nagoya, Japan

**Abstract:** Problem/Project-Based Learning (PBL) is a method of learning in which students solve problems or complete projects. To solve problems or complete projects, students work collaboratively in small groups. Through PBL, student motivation is expected to be enhanced, so this study investigated students motivational changes in PBL. Thirty-four undergraduate students who attended the PBL class were asked to complete questionnaires on 3 occasions; at the beginning, in the middle and at the end of the class. Questionnaires contained scales relating social motivation, and learning strategies. The results show that one dimension of social motivation (motive to be evaluated) increased in the process of PBL. The relationship of social motivation and learning strategies are discussed.

# **Phoneme exchange in, serial-position effect on, and lexical/semantic contributions to single-word production: An investigation using speech-error induction techniques in Japanese.**

**Masataka Nakayama**

Kyoto University, Kyoto, Kyoto, Japan and Japan Society for the Promotion of Science, Chiyoda, Tokyo, Japan

**Shogo Kajimura**

Kyoto University, Kyoto, Kyoto, Japan

**Masashi Sugimoto**

Kyoto University, Kyoto, Kyoto, Japan and Japan Society for the Promotion of Science, Chiyoda, Tokyo, Japan

**Kaori Kuraya**

Kyoto University, Kyoto, Kyoto, Japan

**Miyako Inoue**

Kyoto University, Kyoto, Kyoto, Japan

**Ryo Ishibashi**

Kyoto University, Kyoto, Kyoto, Japan and Japan Society for the Promotion of Science, Chiyoda, Tokyo, Japan

**Satoru Saito**

Kyoto University, Kyoto, Kyoto, Japan

**Abstract:** Serial-order control is a fundamental aspect of speech processing, and analyses of speech errors provide clues to the mechanisms that control this phenomenon. The present study employed a speech-error induction technique to identify speech errors that strongly resembled recall errors observed in verbal immediate serial recall. In three experiments, participants repeatedly produced a target word/nonword and, immediately before the utterance, were unexpectedly exposed to an auditory distractor word/nonword, which was either phonologically similar or dissimilar to the target. This technique successfully induced within-word phoneme exchange/transposition errors and elicited a within-word serial-position effect. On the other hand, lexical/semantic variables (e.g., lexicality of the target) led to only a very weak effect. We discussed mechanisms for the retention and production of a phoneme sequence in the context of these results.

# **An Analysis of Disfluencies in the Actor's Speech for Character Design**

**Seung Suk Nam**

Sogang University

**Hye Rhang Cho**

Sogang University

**Sook Whan Cho**

Sogang University

**Abstract:** This paper presents an analysis of quantified characteristics of Disfluency pattern of a leading South Korean film actor Kang-ho Songs Speech for character design in No.3 (Comedy, 1997), Memories of Murder (Thriller, 2003) and Thirst (Melodrama, 2009). Disfluencies, such as filled pause, repeated words, and repair or repetition utterance, are prevalent in spontaneous spoken language and show remarkably regular trends in a number of dimensions (Elizabeth Shriberg, 1994). The transcriptions of the actor's speech with disfluencies and variables in above films are analyzed, which might affect fluency rates in a corpus. It will be seen that the rates of disfluencies, induced by artificial manipulations, reflect the psychological state of a character. The actor's disfluency rates were relatively increased in Comedy and Thriller than Melodrama. Disfluencies are related with these cinematic factors: movie genre, actor role. Reflecting these features, we then recognize the acoustic features needed for character design.



# **The effect of a childbirth psychoeducation program on postnatal depression**

**Fei Wan Ngai**

The University of Hong Kong

**Sally Wai Chi Chan**

National University of Singapore

**Abstract:** The purpose of this study was to evaluate the effect of an antenatal childbirth psychoeducation program, which focused on teaching cognitive restructuring skills and problem solving strategies, on postnatal depression. Cognitive-behavioral strategies focus on challenging and changing dysfunctional cognitions and maladaptive behaviors that are believed to play a role in minimizing the risk of depression. This study used a mixed method design. The findings showed that women receiving the program had significant improvement in depressive symptoms compared to the control group during pregnancy and at 6 weeks and 6 months postpartum. Women perceived the program to be helpful in fostering the development of cognitive-behavioral skills, enhancing their confidence in taking up the maternal role, and improving their emotional control in the perinatal period. The findings provide empirical support for the feasibility of the childbirth psychoeducation program for promoting perinatal health and minimizing the risk of postnatal depression.

# **Satisfaction evaluation and time perception for waiting time in ICT usage under dual task situation**

**Sumaru Niida**

KDDI R&D Laboratories Inc. and University of Tsukuba

**Satoshi Nakamura**

University of Tsukuba

**Tomomi Moroga**

University of Tsukuba

**Etsuko T. Harada**

University of Tsukuba

**Satoshi Uemura**

KDDI R&D Laboratories Inc.

**Abstract:** Waiting is a potential source of frustration in human-computer interaction over network. The purpose of our research is to clarify a cognitive process of waiting for solving the problem, and we investigated satisfaction evaluation and time perception for waiting in human-ICT interaction under dual task situation. Participants were assigned to one of two primary tasks, evaluation of satisfaction or time length while waiting until the e-mail has been sent. Simultaneously, they executed one of two secondary tasks, calculating a consecutive numbers or solving a crossword puzzle. Result showed that the crossword task has stronger influences than calculation task in general, and surprisingly showed decrease of satisfaction with short waiting time, while showing shortening of psychological time with longer waiting time. That is, events associated with waiting process affect satisfaction evaluation and time perception in different ways. The frustration while waiting cannot be explained purely by length of psychological waiting time.

# Multimodal Interactions Development Process in Collaborative Creation

**Koshi Nishimoto**

Doshisha University

**Mamiko Sakata**

Doshisha University

**Abstract:** We have tried to quantify communication, both in terms of verbal and non-verbal processes, in collaborative activities using collaborators engaged in a creative task as our study subjects. This study involved a production task using LEGO blocks. The study subjects consisted of a total of ten groups; five groups made up of three men and another five, of three women. In an earlier study conducted by the same writer using male-only subjects, the group rated to have produced a work of greater perfection used a number of gestures to communicate, while the group rated to show a high level of originality was engaged in conversation unrelated to the task. The current study repeated a similar experiment using women subjects under the same conditions. We analyzed the modes of communication generated in collaborative activities and the differences between men and women, both in terms of verbal and non-verbal processes.

# Task complementarity and response complementarity in the social Simon effect

**Akio Nishimura**

Sophia University

**Koh Miyamoto**

The University of Tokyo

**Kazuhiko Yokosawa**

The University of Tokyo

**Abstract:** When two individuals sitting adjacently engage in separate but complementary go/no-go tasks, with each responding to only one of two non-spatial attribute (e.g., color; one to red and the other to green) of a stimulus presented on right or left sides, responses are faster when the actor and the stimulus are on the same side. The present study investigated the role of task complementarity in this social Simon effect by assigning same (e.g., red/green) or different (e.g., red/square) target feature dimensions (color/shape) to the two participants. We also investigated the role of response complementarity by manipulating the proportion of trials to which both, one, or neither of the participants responded. A comparable social Simon effect was obtained irrespective of target or response complementarity. We conclude that what is represented in the social Simon effect is the action, rather than the task, of the adjacent partner.

# **Reflexive orienting to others gaze is modulated by the accuracy of recognition of emotional facial expressions.**

**Yuka Nishiyama**

Nagoya University

**Jun Kawaguchi**

Nagoya University

**Abstract:** This study investigated the underlying mechanism of how emotional facial expressions modulate reflexive orienting to others gaze. We conducted gaze-cueing studies (e.g. Friesen & Kingstone, 1998), using dynamic emotional facial cues (i.e. happy, anger, fearful, and neutral expressions). A facial cue with gazing either left or right was presented. Participants were asked to indicate a position of the target that appear either at the looked-at (valid) location or the invalid location as quickly as possible. Subsequently, Participants were asked to classify each facial expression by a forced choice in the recognition task. It was revealed that fearful expressions facilitated the gaze-orienting effect compared to other expressions, only when the recognition of facial expressions was more accurate. The findings indicate that the accuracy of recognizing emotional facial expressions at the early perceptual stage influence the reflexive orienting to others gaze.

# **A Case Study of Meta-cognitive Exploration of Facial Expressions**

**Takeshige Nishiyama**

Keio University

**Hiromi Ochiai**

National Museum of Emerging Science and Innovation

**Yuko Toukairin**

Keio University

**Masaki Suwa**

Keio University

**Abstract:** We live among and belong to various social communities. Relationships and bonds with these communities are formed by daily communications. There are two main communication channels: verbal and non-verbal. Non-verbal communication channels are important for creating relationships within communities. Facial expressions are formed implicitly, and it is difficult for a person to know his/her expressions when having a conversation. Thus, we present a case study in which participants explore their own facial expressions by taking photos while having conversations. The process of the exploration includes taking photos, categorizing and labeling each expression, and making a pictorial book using these photos. Through this process, participants were able to know his/her expressions explicitly, which allowed them to reflect on their interactions within conversations. This indicates that, the process encourages meta-cognitive explorations by participants. Therefore, the participants gain new insight on their daily communications.

# **A cognitive emotional model for intrinsic motivation**

**Kohei Noda**

Kocoro Laboratory Inc.

**Abstract:** The author of this article previously collaborated on the analysis of the process of a coaching session and a motivation training program. Case studies were conducted from those analyses. However, those studies are separate and there is no integrated cognitive model existing to explain the cognitive emotional process of those cases. Also, case studies of entrepreneurs in which the reason they founded their companies have also been conducted. In this study, a cognitive emotional model and intrinsic motivation, a key concept, are introduced and modeled. To model the Knowledge, Skills, Abilities and Others (e.g. traits and etc.)(KSAOs), Human Resource Assessment (HRA) Ontology is introduced. To model human emotion and motivation, Cognitive Emotional Agent Architecture (CEAA) is introduced. A new integrated concept for motivation training courses and coaching sessions is developed based on the developed cognitive emotional model.



# What is the trial-and-error process of design thinking?

**Hisataka Noguchi**

freelance (retired)

**Abstract:** What is the trial-and-error process of design thinking?

Hisataka Noguchi

When we think about design creativity we should recollect the fundamental feature of the design thinking process as this:

- (1) Design thinking is essentially included in general human productive works as a whole.
- (2) Design thinking usually goes not straight to the goal but requires trial-and-error processes in which repetitive divergent and convergent thinking are taken.
- (3) For converging the trial-and-error to goal, designer needs self-evaluations in every step of getting tentative solutions.
- (4) Even though the collective thinking and discussions facilitate this process, creative result can be carried out basically depend on each designer's self-evaluation.

The author tried to make a model of trial-and-error process of design thinking in which one can analyze how the self-evaluations converge the process as a representation of designer's internal world.

# Choosing unknown goods: An fMRI study of product choice

**Ikuya Nomura**

The University of Tokyo

**Kazuyuki Samejima**

Tamagawa University

**Kazuhiro Ueda**

The University of Tokyo / CREST

**Yuichi Washida**

Hitotsubashi University

**Hiroyuki Okada**

Tamagawa University

**Takashi Omori**

Tamagawa University

**Abstract:** Choice between known goods and unknown goods is repeated in everyday life as new products go on the market one after another. Although such choice is one of the key factors of consumer behavior, very few experimental studies have been done thereon. In the present study, repetitive choices among known goods and unknown goods are performed by utilizing mineral water as stimuli and characteristics related to the choice are examined. Further, brain activity during product choice was measured with fMRI. As a result, subjects who had a tendency to seek for information tend to choose unknown goods more and activity in the right frontal pole was observed when unknown goods were chosen. These results indicate that choosing unknown goods is a behavior for the purpose of gaining information but not a consequence of balancing the profits and losses.

# Clarifying position derived from sophisticated beliefs about the nature of knowledge-to-use

Ryota Nomura  
The University of Tokyo

**Abstract:** This study examined students self-reported active clarification of their own and other members position in group discussion in order to shed light on the relation between students beliefs about the nature of knowledge-to-use and their way of participation in cooperative learning. According to their epistemic beliefs, 58 (14 male and 44 female) undergraduate students were assigned to 3 type of groups, including sophisticated (i.e., assuming a wide applicable scope of knowledge and considering prior conditions for application), nave, and mixed. The results demonstrated that students with sophisticated beliefs in mixed-group rated their own clarification higher compared to students with nave beliefs and to students in sophisticated-group. It is discussed that the gap of beliefs provided different ways of argument among students, leading a demand to clarify their stance each other and offering a comparative advantage for students with sophisticated beliefs.

# **A Method of interviewing to Constructively Generate a Narrative through Interactions between Interviewer and Interviewee - A Case Study to Examine Creative Thoughts of an Architectural Student -**

**Haruka Nukariya**

Keio University

**Masaki Suwa**

Keio University

**Abstract:** We discuss both a method of interviewing about an individuals creative thinking and a method of recording an interviewers behavior. From the viewpoint of scientific examination, interviewing is supposed to be an act of retrieving an interviewees thoughts in a raw manner without being influenced by the interviewers viewpoints or intentions. An individuals thoughts are, however, mostly tacit knowledge. Even the individual himself is not necessarily able to externalize it by himself. We argue, rather, that interviewing should be an act of generating a narrative in a constructive manner through active interactions between an interviewer and an interviewee. Based on this paradigm change, we have devised a method of interviewing using a memo tool so that the interaction becomes active. Moreover we have developed a method of recording an interviewers behavior situated in the interaction in order to examine what kind of behaviors make the interaction active.

# **Do children who experience regret make better decisions? A developmental study of the behavioral consequences of regret**

**Eimear O'Connor**

Queen's University Belfast, Belfast, Co Antrim, Northern Ireland

**Aidan Feeney**

Queen's University Belfast, Belfast, Co Antrim, Northern Ireland

**Teresa McCormack**

Queen's University Belfast, Belfast, Co Antrim, Northern Ireland

**Abstract:** To date there has been little research investigating the assumption that experienced regret influences future decision making. Three novel experiments examined the relationship between childrens ability to experience regret and the quality of their subsequent decision making. Children chose between two options on Day 1, discovered the non-chosen option was more attractive, and rated their feelings about their choice. On Day 2, to assess choice switching, children were presented with the same choice. Experiment 1 found regret and adaptive choice switching emerged around 7 years of age. Experiments 2 and 3, found that 6-and 7-year-olds who experienced regret on Day 1, engaged in profitable decision making significantly more often than those who did not experience the emotion. These findings remained even when age and verbal ability were controlled for. These findings suggest that the experience of regret influences similar future decision making in childhood.

# Analysis of Human Solving Process of Constraint Satisfaction Problem

**Hidemi Ogasawara**

Chukyo University

**Yoshiyuki Matsuzawa**

Chukyo University

**Masahide Ogawa**

Chukyo University

**Abstract:** This study discusses cognitive process in solving constraint satisfaction problem in terms of both operation and information acquisition strategies by examining human process data with eye mark tracking solving Sudoku.

Generally a Sudoku problem is defined as a CSP where a cell corresponds to a variable, and its domain of values is a set of the numbers. One of the major inference strategies to solve it is constraint propagation based on local consistency that resolves conflicts within a local variable set, and a problem that is solvable only by resolving the two-variable consistency is classified as "easy".

However, the result of human data does not support this naive problem definition and the strategy. It can be explained by another problem representation where a number corresponds to a variable, and a cell to a domain value. We check the cognitive validity of this representation and the strategies based on it.

# Towards the Development of Integrated Narrative Generation System as the Implementation of Literary Knowledge

**Takashi Ogata**

Iwate prefectural university, Iwate, Japan

**Taisuke Akimoto**

Iwate Prefectural University, Iwate, Japan

**Abstract:** Narrative contains diverse literary elements (e.g. narrator, character, story, plot, media), and narrative generation system should organically combine them. Previously, we developed a variety of the elements comparatively individually. The current goal is to develop a system in which all modules are organically integrated. This paper shows the pilot integrated system including about 800 modules and conceptual dictionaries, moreover discusses a detailed design for the revised version. The generation phase is divided into story, discourse, and expression. Story and discourse are described as each conceptual structure. The transformation by narrative techniques is corresponding to the narrative generation process. The mechanism is constructed by the integration of AI techniques such as script, story grammar, discourse relations and narratological knowledge such as Propps story structure, Genettes discourse theory and Jausss reception theory. In this sense, this system is also an approach integrating the interdisciplinary genres. The generation process is executed by a control mechanism that selects techniques to be used based on the parameters for defining structures and characteristics of the narrative to be generated.



# Effect of Age-related Decline of Task Switching on the Task Sequences that Simulate Real Job

**Keiji Ogata**

Azbil Corporation

**Satoru Suto**

Shizuoka University

**Kazutaka Ueda**

The University of Tokyo

**Takatsune Kumada**

National Institute of Advanced Industrial Science and Technology

**Tohru Ifukube**

The University of Tokyo

**Abstract:** Task switching is an important cognitive function when older adults work using ICT equipment. We focused on properties of older adults task switching function in order to design usable ICT equipment. Generally, a worker has to execute some tasks serially by switching them reciprocally. In this study, some sequences such as AB, AAB, or AAAAB were taken into consideration because there are routine tasks and rare tasks. Participants were 24 older adults and 24 younger adults. Reaction times and switch costs of various sequences were analyzed. As the result, although switch costs for older adults were heavier in AAB than AB, the switch costs did not increase for longer repetition of A (i. e., AAAB and AAAAB). The heaviest switch cost was seen in AAB when A was a difficult task and B was an easy task. These results were different from those of younger adults.

# **Rhetoric of Biopic and Viewer's Reconsideration : Isnt story an Obstacle?**

**Yukiko Ogawa**

HOSEI University

**Akihito Kanai**

HOSEI University

**Abstract:** In the cognitive process when the viewers of narrative moving images appreciate its world and to deepen their consideration, is it story or rhetoric to take an important role? This is the interdisciplinary point at issue cross-linking the domain where study such as art, literature, rhetoric, and a problem of representation with images, and a cognitive-affective mechanism are connected. Therefore this study used two biopics - "Tower of TARO" and "YUMEJI", and examined it. After having shown the 10 minutes partial images which experimenters fixed, the questionnaire was performed for 267 participants. The question was given "how you reconsidered the concept that experimenters showed to watch each image". Free variables were made by the answers about disposition to various images, ANOVA was carried out afterwards. As a result, it was suggested that the deep concept reconsideration might occur more towards the viewers which minded rhetoric than story development.

# Developmental Adjustment of Iconic Language in Care-Takers' Input

**Masato Ohba**

Tamagawa University

**Noburo Saji**

Keio University / JSPS

**Mutsumi Imai**

Keio University

**Tomoko Matsui**

Tokyo Gakugei University

**Abstract:** This study explores how care-takers use mimetics -a word class which has a clear sound symbolism- in communicating with their children to investigate the role of iconicity in the input for language development. We approached the question by combining a study of mother-child interaction corpus and a Child-Directed-Speech elicitation experiment, in which care-takers were asked to describe animated actions first to their child and then to an adult experimenter. The results showed that care-takers use mimetics more frequently and use them alone as interjections without embedding them in a sentence for younger children. As their child develop care-takers begin to use mimetics adverbially in a sentence, as they do in Adult-Directed-Speech. This change suggests that care-takers adjust the degree of iconicity in the input according to the child's developmental stage, from bare use to syntactically embedded use of mimetics, and then eventually to non-mimetic conventional words.

# Effects of Language on Asynchronized Audiovisual Speech Perception

**Hitoshi Ohnishi**

The Open University of Japan

**Kaname Mochizuki**

Teikyo University

**Abstract:** When people are exposed to asynchronized audiovisual speech, they feel that something is unnatural. We performed tests to determine whether the language used in audiovisual speech affected the observer's perception. In experiment 1, we examined whether Japanese native speakers were more sensitive to audiovisual asynchrony in Japanese audiovisual speech than in English audiovisual speech by using simultaneity judgment tasks. Results showed that the language had no effect on sensitivity to audiovisual asynchrony. In experiment 2, we examined whether audiovisual asynchrony affected the observer's perception of loudness. Results showed that the observer perceived a voice to be quieter when the speaker's face moved after the speaker had spoken. This effect was weaker when the reverse was true and the speaker's face moved first, followed by the speaker's voice. Moreover, the effect was also not as strong when Japanese native speakers observed asynchronized audiovisual speech in English as opposed to Japanese.

# **The effects of regularity in spatial inferences with and without local landmarks on spatial learning**

**Kayoko Ohtsu**

Waseda University

**Yoshihiro Ouchi**

Teikyo-Gakuen Junior College

**Abstract:** Regularity in spatial inferences during wayfinding has a different effect on spatial learning depending on clues available in an environment. Our previous study has revealed that the irregular updating, which involves multidirectional self-to-object updating, has a facilitation effect when global landmarks are available, while it impedes the learning without the landmarks. In the present study, we examined the updating mode in the context of local landmark information. In the experiment using a real maze, participants visited multiple targets in one of four conditions by a combination of two factors: an updating mode (regular or irregular) and an environmental information mode (with or without local landmarks). The results suggest that there is an interaction between factors. Although there is no difference between three of the conditions (regular & landmarks, regular & no landmarks, and irregular & landmarks), the irregular updating impedes the learning without the landmarks.

# Computational model of the meaning acquisition of sentence-final particles

**Natsuki Oka**

Kyoto Institute of Technology, Kyoto, Kyoto, Japan

**Naohiro Nonoguchi**

Kyoto Institute of Technology

**Chie Fukada**

Kyoto Institute of Technology

**Motoyuki Ozeki**

Kyoto Institute of Technology

**Abstract:** Sentence-final particles serve an important role in (spoken) Japanese, because they express the speaker's mental attitudes toward the proposition and/or the interlocutor. They are acquired at early ages and occur very frequently in everyday conversation. There has been, however, little proposal for the computational model of the acquisition of sentence-final particles. The purpose of this study is to get a robot to learn how to act upon the utterance with a sentence-final particle. The robot learns appropriate responses based on the rewards given by the interlocutor. The experimental results show that the robot learns to behave correctly in response to 'yo,' which expresses the speaker's intention to communicate new information, and to 'ne,' which denotes the speaker's desire to confirm that some information is shared. Using the learned actions as a lead, the acquisition of inner information processing such as word learning is the next research target.

# Is the breaking point in mental number line in Japanese children five or ten?

**Masahiko Okamoto**

Osaka Prefecture University

**Sari Nakamura**

Osaka Prefecture University

**Abstract:** The purpose of the present study is to examine that (1) are there breaking points in mental number line in Japanese children, and (2) does it relate childrens calculation performances. The magnitude decision task for eight number pairs, e.g. 3-5, was used in this study. Participants were 40 first graders children. An ANOVA for the decision time revealed that the main effect of number pairs is significant ( $F(1, 39) = 8.45, p < .01$ ). A Tukeys HSD test indicated that the decision time for 4-6 and 5-7 were slower than for 3-5, and the time for 9-11 and 10-12 were slower than for 8-10. These results indicated that there are the five-break and the decade-break in metal number line in Japanese children. And a negative correlation between the decision time and the performances of calculation problem were obtained. It suggested that the calculation skill is acquired based on the mental number line.



# **Influences of a potential interlocutor on the utterances of the speakers who try to describe objects**

**Junji Okamoto**  
Gakushuin University

**Saori Ushiyama**  
Gakushuin University

**Abstract:** Two sets of production tests were conducted in January and February in 2012, both in Tokyo and in Mannheim (Germany) to capture differences of utterances which subjects produced with or without the presence of another person, who sat at 45 deg. left in front of them, remained silent, but pretended to be ready to talk with them. The subjects were requested to describe or comment on objects appearing on the computer display at an interval of 10 sec. Our working hypothesis is that intersubjective expressions such as end-particles in Japanese (JEPs) and modal particles in German (GMPs) would be more frequently observed with the presence of a potential interlocutor. The result was that although the influence was confirmed, GMPs were rarely used even in her presence, while JEPs were frequently used even without her presence. This suggests that there is a crucial difference in terms of interlocutor's commitment.

# **Delay of word order development in Japanese? Evidence from a preferential looking study with 19 and 30-month-old children**

**Akira Omaki**

Johns Hopkins University

**Romy Lassotta**

University of Geneva

**Tessei Kobayashi**

NTT Communication Science Laboratories

**Luigi Rizzi**

University of Siena

**Julie Franck**

University of Geneva

**Abstract:** Word order constitutes a fundamental cue for sentence comprehension. Previous research on English and French found evidence for adult-like word order knowledge in 19-month-olds, but little is known about early word order development in languages with frequent word order alteration and argument omissions that reduce positive evidence for the canonical word order. To address this question, we conducted (i) a distributional analysis of word order variation in child-directed Japanese, and (ii) a preferential-looking study with Japanese 19- and 30-month-olds, using the same design as a previous study on French 19-month-olds. Our analysis of 17726 child-directed utterances revealed that 91% of the input was uninformative for identifying the canonical SOV order. Next, our preferential-looking experiment revealed that Japanese 19-month-olds fail to understand sentences with a canonical word order, unlike French 19-month-olds or Japanese 30-month-olds. We suggest that the sparseness of SOV in the input delays the development of word order knowledge.

# Effects of supraliminal and subliminal hint priming on insight problem solving.

**Ryo Orita**

Graduate School of Letters, Ritsumeikan University

**Masaki Hattori**

Department of Psychology, Ritsumeikan University

**Abstract:** In two experiments, a total of 135 participants engaged in a 10-coin problem for four minutes. Half of them were exposed to a hint figure one time per ten seconds. In Experiment 1, the hint figure was presented for one second, and participants were told that this figure was unrelated to the problem (supraliminal priming). Thus, they were aware of the figure, but they were unaware that it served as a hint. In Experiment 2, the hint was presented for thirty milliseconds, and it was masked by geometric configurations (subliminal priming). This ensured that participants were unaware of the hint. Both types of hint increased the solution rate. These results suggest that the conscious access to the hint is not essential for the priming effects. Exposure to a hint can activate insightful ideas without awareness, and it increases the probability of producing an appropriate strategy to escape from an impasse.

# Transfer of learning from project activities to individual leaning

Naoko Osada

Seisen Jogakuin College

**Abstract:** A number of studies indicate that project-based learning enhances a student's motivation, engagement, and deep understanding. But it is difficult to promote successive individual learning based on the experience of the project. To improve the linkage between the project and the subsequent activities, this study attempted to exploit the context-based approach to enhance the reflection by the students under the design experiment. The results showed that more than half of the second year students successfully used the findings through the project activities to define their individual research theme. Also their final reports were well-organized and the level of the consideration was relatively high than that of the first year. In the real world settings like workplace, it is important to figure out the situations carefully before applying the individual knowledge or skill. Contexts could work as scaffoldings to improve the transfer in the real world.

# **Ranges of storage item sizes in complex working memory span tasks: Latent-variable analysis of the memory load**

**Kazunori Otsuka**

University of Nagasaki

**Makoto Miyatani**

Hiroshima University

**Abstract:** To measure individual differences in working memory capacity, complex span tasks have been developed. These tasks, which predict several aspects of higher-order cognition, have associated ranges of storage item sizes that are regularly controlled to avoid ceiling and floor effects in performance distributions. To examine the memory load incurred by the ranges of storage item sizes, this study used different storage item sizes for complex span tasks across verbal, numerical, and spatial content domains. Particularly examined were the relations among processing, storage, and general fluid abilities of the different storage item sizes of complex span tasks. Results of latent-variable analysis indicate different relations among these components of participants performances in complex span tasks, for which the item sizes differed. The influences on the memory load of storage item sizes of these tasks are discussed.

# **Does the detection of mind wandering require attentional resources?**

**Sho Otsuka**

Graduate School of Education, Tokyo Gakugei University

**Takahiro Sekiguchi**

Department of Educational Psychology, Tokyo Gakugei University

**Abstract:** Individuals can observe the occurrence of mind wandering (task-unrelated thought) while doing an important task and stop it. This study examined whether mind wandering is detected through the monitoring process requiring attentional resources. In the first task, participants read a short story, where the constituent words were presented one-by-one and the next word appeared by pressing a key. The reading time for each word was measured. As a secondary task, one group reported the occurrence of mind wandering each time they noticed it during the reading. The other group performed a prospective memory task (detecting target words) that required attentional resources, and the control group performed only the reading task. The result showed that while performing the prospective memory task increased the mean reading time, the reporting of mind wandering did not influence it. This finding suggests that the detection of mind wandering did not require attentional resources.

# **Analysis of the relationship between writing skills and evaluation in an expository writing assignment**

**Hiroko Otsuka**

Future University Hakodate, Hakodate, Hokkaido, Japan

**Mio Tsubakimoto**

Future University Hakodate, Hakodate, Hokkaido, Japan

**Hiroshi Numata**

Future University Hakodate, Hakodate, Hokkaido, Japan

**Abstract:** Our research aims at acquiring the knowledge to develop an effective system of support for revising written assignments, in order to assist students to compose logical sentences by themselves. We analyzed, quantitatively and qualitatively, the problems that arose in expository writing assignments, and the relevance of grade calculations. Furthermore, we also surveyed student attitudes toward written assignments in order to examine appropriate methods of support in the revision process.

Expository writing texts are composed of a title, a body text, and a bibliography. The body text contains the definition of the theme, the reason for the choice of theme, questions that arise relating to the theme, exploration of these questions, and the conclusions.

Our results demonstrated that the quality of the revised texts was influenced by the following factors: the length of the whole text, the section pertaining to the exploration of the research questions, and the bibliography.



# Loosely symmetric heuristics as the basis for biases and the empirical Bayes methods

**Kuratomo Oyo**

Tokyo Denki University

**Tatsuji Takahashi**

Tokyo Denki University

**Abstract:** The loosely symmetric (LS) model, a conditional probability-like formula describing human symmetric cognitive biases, is shown to be effective in an ample amount of areas including causal induction, learning (reinforcement, supervised and unsupervised), game-theoretical situations and digital game AI. However, despite its interesting mathematical properties, the total rationale for the model has not been given. In this study, we analyze LS from the viewpoint of Bayesian statistics, especially of the empirical Bayes methods. As the result, we show that the bias terms in LS, which deviates LS from the ordinary conditional probability, are the hyper parameters for the prior Beta distribution LS assumes. Given this analysis, LS is shown to describe various cognitive biases including Gamblers fallacy, status quo bias, the framing effect, dependence to the reference point and the reflection effect, etc., all at the same time.

# How simple explanations change our minds and why we prefer them

**Michael Pacer**

UC Berkeley

**Tania Lombrozo**

UC Berkeley

**Abstract:** People prefer simpler explanations and judge them more satisfying, where simplicity is defined as the number of unexplained causes in the explanation (root simplicity). But what are the effects of this preference on cognition and why do people like simple explanations? We consider the answers to these questions with three experiments. Our findings suggest that the participants who select simpler explanations when the simpler explanation conflicts with available data are more likely to overestimate how often they saw data points consistent with their chosen explanation. We show that choosing a simpler explanation causes this data distortion. We find that the preference for root simplicity is stronger when root causes greatly alter the probability of their effects. This aligns with arguments suggesting a preference for root simplicity results from the desire to efficiently encode observations and to support effective interventions. These findings have implications for theories of inference and explanation.

# **The Influence of Culture: Thematic versus Taxonomic Categorization**

**Jo Pan**

National Cheng Kung University

**Yi-Rong Wu**

National Cheng Kung University

**Fan-Ning Cheng**

National Cheng Kung University

**Gert Westermann**

University of Lancaster

**Hsueh-Chih Chen**

National Taiwan Normal University

**Jon-Fan Hu**

National Cheng Kung University

**Abstract:** There is evidence that cognition processes may not be universal for people in the world. The goal of this study was to explore whether people from different culture backgrounds adopt varied concepts to categorize objects. Participants from eastern and western were required to rate association strength of word pairs in questionnaires which used five-point-scale. Two types of word relations are manipulated. These words were paired by taxonomic (e.g., banana - apple) or thematic (e.g., banana - orangutan) categories. The results showed culture differences. Relative to western people, easterners tend to put weaker emphasis on objects taxonomic categories than on thematic relations. There was an interaction between people from different cultural background and the type of relations. The tendency influenced by culture to sort objects into thematic or taxonomic categories will be discussed in further research.

Keywords: culture, thematic, taxonomic, categorization

# **Do Objects Matter for Infants Formation of a Spatial Category?**

**Youjeong Park**

Cornell University

**Marianella Casasola**

Cornell University

**Abstract:** Presenting a relation using simplified objects facilitates the generalization of that relation in older children and adults (e.g., Kaminski, Sloutsky, & Heckler, 2006, 2008). We investigated infants ability to form a category of a support relation (i.e., on) when the objects that depicted the relation were perceptually simple versus more complex. Fifty-three infants of 8 and 14 months were habituated to dynamic support events with either simple or complex objects. They were then tested with events with novel objects, a novel spatial relation, or both. Infants at each age formed a support category, looking significantly longer at test events with a novel than familiar relation. There was a marginal effect of object complexity. The current results suggest a role of object features in infant spatial categorization, and provide the first evidence that infants of 8 months can form a spatial category of support relations.

# **Cross-level Illusory Conjunction between Implied (Semantic) and Actual (Perceptual) Colors**

**C. J. Park**

Department of Psychology, The University of Hong Kong

**A. W. Y. Ho**

Department of Psychology, The University of Hong Kong

**G. H. T. Tsui**

Department of Psychology, The University of Hong Kong

**J. T. Y. Chan**

Department of Psychology, The University of Hong Kong

**X. Luo**

Department of Psychology, The University of Hong Kong

**C. H. Tseng**

Department of Psychology, The University of Hong Kong

**Abstract:** Goldfarb and Treisman (2010) found observers made more perceptual binding errors (illusory conjunctions, ICs) when the features of an object were inconsistent. They suggested that such mistakes originate from features in the same level. We investigated whether we could induce ICs across perceptual and semantic domains by manipulating the font colour of object words containing implied colours. In Experiment 1, participants saw very brief displays of four Chinese characters ((sun), (fire),(hill), (water)) printed in either yellow, red, green, or blue. Observers had to report the physical color of a randomly selected target word in each trial. We observed significantly higher error rates when the words were printed in colours incongruent with their implied color. We replicated the result with a set of English words (lemon-yellow, water-blue, blood-red, and grass-green) in Experiment 2, which led us to conclude that illusory conjunctions can arise in the gulf between semantic and perceptual domains.

# **Judgment under uncertainty is not always certainty-oriented**

**Youngjun Park**

Ajou University

**Kyungil Kim**

Ajou University

**Abstract:** Previous research of uncertainty has assumed that peoples single time- preference for alternatives is based on the degree of availability in probability. We, however, suggest that the influence of probability decreases when a global strategy includes multiple selections with a single goal. A game, which involves a tournament of 8 teams, is introduced to the participants and they were asked to select 4 teams so that their selection includes the winner. Participants choice responses were classified into 2 types. The first type is to pick all the 4 teams on one side of the bracket (a sure 50% to win) making the degree of uncertainty minimized. The second type includes all other ways of selection, except the first type, with greater uncertainty. The results indicate that only 37% chose the former strategy indicating that the majority of people used a strategy with higher uncertainty.

# The Effect of Perceptual Complexity on Affective Picture Processing

**Taejin Park**

Chonnam National University, South Korea

**Soodam Park**

Chonnam National University, South Korea

**Abstract:** Previous ERP studies using IAPS (International Affective Picture System) pictures have reported several ERP components (e.g., P1, N2, LPP) whose amplitudes are modulated by valence and arousal. But Bradley and her colleagues (2007) reported that early ERPs were modulated only by perceptual complexity, and only late ERPs (LPP) were modulated by affective factors. To elucidate the modulatory effect of perceptual complexity and affective valence of IAPS pictures on ERPs, we manipulated perceptual complexity and valence of them. Pleasant, neutral, and negative pictures from IAPS depicting either simple figure-ground compositions or complex scenes were selected. Each picture was presented for 2 s and participants were required to do valence discrimination response. 30-electrodes ERP data of 28 university students were analyzed. Mean amplitudes of early posterior P1 and P2, and LPP showed interaction effect of perceptual complexity and valence, and showed perceptual complexity effect only for neutral and positive pictures (no perceptual complexity effect for negative pictures). Time-frequency analysis showed similar results: Power of alpha and theta activities showed perceptual complexity effect only before 300ms, and showed the effect only for neutral and positive pictures. These results suggest that early ERPs can be modulated by affective valence and perceptual complexity only for neutral and positive pictures, and valence effect of negative pictures might override the effect of perceptual complexity.



# Effect of Saliency-Based Masking in Scene Classification

**Tae-Suh Park**

Seoul National University

**Byoung-Tak Zhang**

Seoul National University

**Abstract:** In this paper, the effect of attention-based local feature selection to 15-class scene classification is investigated, as an extension of the previous researches showing the different effect of each spatial scale to its performance in the early stage of human vision processing. Visual saliency is used as a criterion for selecting the local regions from where HoG features are extracted. Experimental results show that such saliency-based masking significantly affects the classification performance: contrary to the previous reports in the field of object recognition, the low-salient regions contribute more than the high-salient regions in scene classification, and that is consistent with several previous reports of insisting the importance of spatial layout in the low frequency channel, which support the scene schema hypothesis. Also, the result implies that the highest salient regions, which occupies top 20 percent in saliency, hardly contribute to classification performance.

# **fMRI evidence for sensitivity to coherence and context in the theory of mind network**

**Alexander Paunov**

Massachusetts Institute of Technology

**Jorie Koster-Hale**

Massachusetts Institute of Technology

**Rebecca Saxe**

Massachusetts Institute of Technology

**Abstract:** Much neuroimaging work converges on a set of brain regions engaged in representing the mental states of other agents. These regions, the theory of mind (ToM) network, include the bilateral temporoparietal junction, dorsomedial prefrontal cortex, and precuneus. Studies employing diverse methodologies consistently reveal greater activation in the ToM network to stimuli that prompt mental state attribution relative to various control conditions. However, attempts to elicit differential responses within the network by manipulating mental state content have typically failed. We report findings that the network is sensitive to two features of mental state stories. First, the hemodynamic response is reduced following a coherence break in the stories plot relative to a coherent condition, in both ToM and left-hemisphere language areas. Second, the amount of context provided is positively related to the strength of activation in the network. These results suggest that ToM representation depends on a sustained input of coherent information.

# **The Role of Linguistic Knowledge in Hue Perception**

**Katherine Phelps**

University of Colorado, Boulder , CO, United States

**Steve Duman**

University of Colorado, Boulder

**Kevin Gould**

University of Colorado, Boulder

**Les Sikos**

Swarthmore College, Swarthmore, PA

**Abstract:** It has been argued that real-world structure constrains the semantic representations of verbs, resulting in cross-linguistic convergence of naming patterns for motion events. This study explores the nature of this real-world structure by manipulating individual features of human locomotion in video stimuli and comparing the responses of English and German speakers in an elicitation task. We show that individual features influence naming patterns and that languages encode these features differently. Furthermore, the semantic representations of several German motion verbs sharply contrast with their English equivalents.

# **fMRI of attention and automaticity in judgments from facial appearance**

**Ramsey Raafat**

University College London, Cognitive, Perceptual and Brain Sciences

**Nikos Konstantinou**

University of Cyprus, Department of Psychology

**Chris Frith**

University College London, Wellcome Centre for Functional Neuroimaging

**Nilli Lavie**

University College London, Institute of Cognitive Neuroscience

**Nick Chater**

University of Warwick, Warwick Business School

**Abstract:** Facial social judgments such as trustworthiness, indeed the processing of facial and emotional stimuli are fast (Bar, 2006) and are often proposed to take place in an automatic fashion, that is independent of top down factors such as attention (Jonides 1983). Here we present work that challenges the automaticity account of such judgements.

In a series of experiments we apply load theory (Lavie, 1995, 2010), to investigate subjective judgements under load. Employing a combined visual search and face judgement task, where the level of attentional load in the search task was manipulated (by varying the search set size) the results indicated reduced accuracy for trustworthy judgements under the effects of attentional (perceptual) load. Finally, in a neuro-imaging task (that is a pure signal response to facial stimuli) we again illustrate the effect of load, where high load modulates (at a neural basis) trustworthy faces.

# **Cooperative Behavior in Multicultural Settings: The Contribution of Altruistic Punishment**

**Daniela Rabellino**

Center for Cognitive Science, Dept. of Psychology, University and Polytechnic of Turin, Italy

**Rosalba Morese**

Center for Cognitive Science, Dept. of Psychology, University and Polytechnic of Turin, Italy

**Angela Ciaramidaro**

Center for Cognitive Science, Dept. of Psychology, University and Polytechnic of Turin, Italy; Dept. of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Goethe-University, Frankfurt/M., Germany

**Bruno G. Bara**

Center for Cognitive Science, Dept. of Psychology, University and Polytechnic of Turin, Italy; Neuroscience Institute of Turin, Italy

**Rosalba Rosato**

Center for Cognitive Science Dept. of Psychology, University and Polytechnic of Turin; Unit of Cancer Epidemiology, San Giovanni Battista Hospital - Turin, University of Turin and CPO Piemonte

**Francesca M. Bosco**

Center for Cognitive science, Dept. of Psychology, University and Polytechnic of Turin, Italy; Neuroscience Institute of Turin, Italy

**Abstract:** In this study we want to explore the behavior of altruistic punishment, investigated through the Third Party Punishment game. This game shows the behavior of spending ones own money, without any personal benefit, to punish unfair behavior of players who violate cooperation norms. This behavior may be differently displayed depending on the in-group versus out-group setting, and typically favors ones own group (the so called parochial altruism). We compared two different cultural groups: Italian (N=26) versus Chinese people (N=26). In both groups, our results show the presence of altruistic punishment behavior, and this tendency emerges much prominent when facing another groups player who behaves unfairly with ones own group members. Furthermore the whole sample shows the propensity to spend little sums of money to punish also fair behaviors: this attitude is known as antisocial punishment. Our data also show that Italians spend significantly higher amounts than Chinese for antisocial punishment.

# **Why so Stressful? The effect of coping strategies and social support to undergraduate students in Korea**

**Young Sun Ryu**

Seoul National University

**Ha Rim Kim**

Seoul National University

**Jeong Ryu**

Seoul National University

**Abstract:** Purpose: To investigate the amount of stress that Nursing College Students for Nursing Examination acceptance competitions and campus life. Method: The data was collected from 250 students at S Nursing College in Seoul, Korea using questionnaires including the following: CES-D(Center for Epidemiologic Studies Depression), BAI(Beck Anxiety Inventory), Coping Stress Scale, Self Esteem Scale, Social Support. Results: The average score for the CES-D turned out to be 18.43(SD: 7.25) which higher than the average 14.77(SD: 9.47) which was firefighters in Seoul, Korea who had experienced many kinds of disasters. As a result, the mediating effects of positive emotion on the relationship between the social factors and CES-D /BAI were supported. CES-D was positively related only to social support( $r = 6.47$ ,  $p < .000$ ) and especially, CES-D most related family support which one of the social support and didnt any relationship with the amount of other scales. Conclusions: We must acknowledge the fact that the amount of stress being received by college students preparing for Nursing Examination is very severe. The amount of stress was even higher than that of college students who experienced a natural disaster. There should be active programs in Nursing College to alleviate this stress if we want to have brighter, healthier college students in the future.

# Automatic Reverse Engineering of Human Behavior Based on Text for Knowledge Acquisition

**Rafal Rzepka**

Hokkaido University, Sapporo, Hokkaido, Japan

**Kenji Araki**

Hokkaido University, Sapporo, Hokkaido, Japan

**Abstract:** We propose a novel approach for building cognitive architectures based on Wisdom of Crowd. As knowledge needed for an intelligent system is difficult to gather and heavily depends on a programmer's bias, we decided to automatize the process of surveying reactions, decisions, opinions, etc. For demonstrating the usability of our approach we implemented Web-mining functions into our system and tested it by confronting with over 100 ethically-significant real world problems, e.g. "killing a man", "stealing money", "bribing someone", "helping people" or "saving environment". The accuracy was 86% when used emotional consequences discovery, however we show how adding social consequences discovery, natural language processing tools and bigger data sets are able to refine the results and increase correctness of moral judgment. We discuss the importance of such algorithm based on crowd behavior, which opens a path for retrieving universal transcultural rules when the system becomes multilingual and compares results from different nations.



# **The Internal Structures of Sound-Symbolic Systems: the Universal and Language-Specific Portions of Sound Symbolism**

**Noburo Saji**  
Keio University/JSPS

**Kimi Akita**  
Osaka University

**Mutsumi Imai**  
Keio University

**Katerina Kantartzis**  
University of Birmingham

**Sotaro Kita**  
University of Birmingham

**Abstract:** This paper demonstrates a new quantitative approach to identify what is behind universally sensed sound-symbolism and sound-symbolism detected only by speakers of a particular language. For this purpose, we presented 70 locomotion videos to English and Japanese speakers and asked them to create a word that would sound-symbolically match each action. Then the participants rated each action on 6 semantic dimensions. Multivariate analysis detected what level of sound-unit (e.g., phonetic features, phoneme, mora) are linked to each semantic dimension. Results revealed that certain sound-meaning links (e.g., voicing and heaviness) are more consistent than others within and across languages. Furthermore, language-specific sound-symbolism was found in the particular sound-units (e.g., the sequence of a voiced initial consonant and a middle-low vowel and heaviness in English). This implies that the language-specific sound-symbolism is motivated by the phonological system each language possesses, whereas universal sound-symbolism appears at the more abstract level of sound component.

# **Does word order influence non-verbal event description by speakers of OS language?**

**Hiromu Sakai**

Hiroshima University

**Takuya Kubo**

Hiroshima University

**Hajime Ono**

Kinki University

**Manami Sato**

Hiroshima University

**Masatoshi Koizumi**

Tohoku University

**Abstract:** Goldin-Meadow et al. (2008) examined whether word orders of speakers own languages influence their non-verbal behaviors by asking speakers of English, Chinese, Spanish, and Turkish to describe events non-verbally using gestures. They found that speakers of all four languages preferred to perform gestures in Actor-Patient-Action order. Although they argued that this reflects a natural order of event description for humans, their research is limited in languages with Subject-Object (SOV or SVO) word order. In order to verify their claim, we examined Kaqchikel, a Mayan language spoken in Guatemala, employing Object-Subject (VOS) word order. In our non-verbal event description experiment, 32 Kaqchikel native speakers described 18 pictures of transitive events using gestures. We found predominant Actor-Patient order (86.2%) as well as Patient-Actor order (13.8%). These results revealed that event descriptions are not only governed by universal conceptual preference to Actor-Patient order but also influenced by word orders of speakers own languages.

# Embodied Skill to Activate Communication in TV Shows

**Rui Sakaida**

Keio University

**Masaki Suwa**

Keio University

**Abstract:** How can we activate communication? Although we participate in and contribute to various conversations everyday, we are not necessarily self-aware of what skills we employ to activate conversations. A good example of experts good at activating conversations is comedians. They quickly grasp what roles they are supposed to play in any changing circumstances in a TV variety show so that the whole atmosphere is enjoyable. Focusing on Ametalk, a popular Japanese TV variety show, we examined what skills comedians possess to direct conversations toward an enjoyable atmosphere. We analyzed conversations statistically from the viewpoints of roles and rhetoric, clarifying why comedians conversation makes us amused. Especially, our focus is on a popular comedian Mr. Kendo Kobayashi; we statistically showed that he not only makes an appeal of his strong personality by playing some roles he is good at playing, but also changes roles according to circumstances.

# Collective decision-making processes in online social networks

Yasuaki Sakamoto

Stevens Institute of Technology

**Abstract:** How do users process the vast amount of information available online to form opinions and make decisions? To answer this question, I analyzed individual and collective decision-making processes in a social media website, Digg. The results from behavioral experiments and computer simulations indicate that some users make decisions by following the choices of others. In particular, I found evidence of three types of following behavior: following the opinions of the whole community, following the opinions of specific individuals, and reciprocal following of pairs of users that could contribute to the emergence of collective decisions in online communities. The results further suggest that knowing others' opinion about a piece of information affects a user's decisions to follow the opinion but not her perception of the information. I conclude by discussing the implications of this work for predicting trends as well as for using and designing social media websites.

# **Innovating scuba diving education through enhanced immersion and authenticity within the eDiving environment**

**Ron Salden**

Madeira Interactive Technologies Institute, Funchal, Madeira, Portugal

**Abstract:** While recreational scuba diving is considered safe and easy to learn, people often become divers on vacation without frequent follow-up diving. Infrequent diving is ill-advised since the scuba diving skills deteriorate over time which can lead to diving accidents. We aim to address the gaps between diving activities by giving people the opportunity to practice relevant scuba diving skills at home. Although prior work on interactive e-learning environments has shown beneficial learning through adaptive personalized instruction, a missing element concerns physical education. As such, we will add two immersive technologies to enhance the authenticity within the interactive eDiving simulator (<http://www.ediving.us>). For each technology we will run a laboratory and a field study, followed by a field study combining both technologies and using Open Water students. As such, our work not only can improve scuba education but also avoid deterioration of important scuba diving skills and consequently reduce diving accidents.

# **What is promoted by imitation, what promotes imitation: Relation to understanding of others mental states**

**Wakako Sanefuji**

Osaka University

**Tomoka Yamamoto**

Osaka University

**Ikuko Mohri**

Osaka University

**Masako Taniike**

Osaka University

**Abstract:** Infants are more likely to imitate actions on objects than body movements (Rogdon & Kurdek, 1977). A recent developmental theory assumes the relationship between body imitation and understanding of others mental states such as intention (Meltzoff, 2004), but there are few evidence on such relationship. The present study was aimed to reveal the developmental linkage between body imitation and understanding of others intention in infancy. Twenty-two participants visited longitudinally at 13, 15, and 17 months of age. We conducted imitation tasks (object manipulations, gestures, and non-meaningful actions) at 13 and 17 months, and task of understanding of others intention at 15 months. Results showed a developmental path from imitation (non-meaningful actions) at 13 months to understanding others intention at 15 months, and also from understanding others intention at 15 months to imitation (gesture and non-meaningful actions) at 17 months. The relationship between imitation and understanding others intention will be discussed.

# Conditions that Modulate Perceptual Interference

**Ava Santos**

Fort Lewis College

**Lawrence Barsalou**

Emory University

**Christy Wilson**

Northeastern University

**Abstract:** Perceptual interference refers to less accurate memories due to verbalizing or imaging nonverbal experiences. We theorize that this occurs because people have prototypical representations of nonverbal stimuli, and these prototypical representations are activated when they verbalize or image exemplars of the stimuli. The activation takes place even though the prototype and the exemplar differ. When people are prompted to remember the exemplar, they mistakenly remember the prototype. In the present study, we want to identify conditions that modulate perceptual interference. We hypothesize that verbalizing nonverbal stimuli multiple times, in the absence of the actual stimuli, will increase the perceptual interference effect. We also hypothesize that viewing nonverbal stimuli multiple times will decrease the perceptual interference effect. Our sample consists of 48 students from a private Southern university and members of the surrounding community. Our hypotheses are supported, demonstrating that there are conditions that can affect the strength of perceptual inference.



# **Contribution of the positive emotional sounds to upward vection**

**Kyoshiro Sasaki**

Kyushu University, Fukuoka, Japan

**Takeharu Seno**

Kyushu University, Fukuoka, Japan

**Yuki Yamada**

Yamaguchi University, Yamaguchi, Japan

**Kayo Miura**

Kyushu University, Fukuoka, Japan

**Abstract:** We examined the effect of positive sounds on upward self-motion perception (vection). Observers listened to positive (e.g., a baby's laughter) or neutral (e.g., a sound of helicopters propeller) emotional sounds during viewing a large vertically moving sinusoidal luminance-defined grating (78 deg 62 deg). The results showed that the upward vection was enhanced by the positive emotional sounds, but not the neutral emotional sounds. This might be because positive affects are linked with spatial/directional up in our mind.

# Surmising of location with vague embodied agents instructions

**Ryo Sato**

Shizuoka University, Hamamatsu, Shizuoka, Japan

**Yugo Takeuchi**

Shizuoka University, Hamamatsu, Shizuoka, Japan

**Abstract:** In this paper we investigate whether joint attention should be achieved between a human and a vague embodied agent by conducting an experiment. Joint attention plays an important role in making conversation smoothly and fulfilling sympathy for each other. And the effect of joint attention has been proved between humans and agents. In this experiment, humans are directed their attention to the certain location by agents instructions consisting of demonstrative pronoun. We made observations of the special relation between human and agent, and the human responses to the agents instructions. What the vague embodied agent is a sphere like bowling ball and its appearance doesnt show where its attentions are turned makes people be unknown its directivity of attention. The result shows that its possible to achieve joint attention between human and the agent even if the agents appearance is too simple and doesnt show where its attentions are turned.

# Looking at nothing facilitates memory retrieval

**Agnes Scholz**

Chemnitz University of Technology

**Katja Mehlhorn**

Carnegie Mellon University

**Josef Krems**

Chemnitz University of Technology

**Abstract:** People fixate on blank locations if task irrelevant visual stimuli previously occupied that region of space. This so-called looking at nothing phenomenon has been associated to information retrieval from an integrated memory representation. However, it is unclear, whether it directly affects the retrieval of information from memory. To clarify this, 23 participants listened to four sentences, each associated to one of four areas on the screen. Subsequently, they had to verify an auditorily presented statement about one of the sentences, by retrieving the related information from memory. During retrieval, participants could either gaze freely, or had to look at a fixation cross that appeared in the area associated to the tested sentence or in one of the other three areas. This manipulation of eye movements significantly affected retrieval performance. The results conform to a grounded perspective on the looking at nothing phenomenon.

# Does Talking to a Robot in a High-Pitched Voice Strengthen an Attachment?

**Ryoko Shibata**

Kyoto Institute of Technology, Kyoto-city, Kyoto-fu, Japan

**Takatsugu Kojima**

Shiga University of Medical Science, Seta Tsukinowa-cho, Otsu City, Shiga, Japan

**Chie Fukada**

Kyoto Institute of Technology, Kyoto-city, Kyoto-fu, Japan

**Kaori Sato**

Kyoto Institute of Technology, Kyoto-city, Kyoto-fu, Japan

**Yuki Hachikura**

Kyoto Institute of Technology, Kyoto-city, Kyoto-fu, Japan

**Motoyuki Ozeki**

Kyoto Institute of Technology, Kyoto-city, Kyoto-fu, Japan

**Natsuki Oka**

Kyoto Institute of Technology, Kyoto-city, Kyoto-fu, Japan

**Abstract:** When adults talk to infants, they tend to use infant-directed speech (IDS) rather than adult-directed speech (ADS). IDS is considered to draw the infants attention more than ADS, to convey adults emotional states to infants more easily, and to make language acquisition easier for infants. It is not clear, however, whether the use of IDS has some effects on the adults as well as on the infants. In this research, we focused on one of the most distinctive features of IDS, a high-pitched voice, and conducted two human-robot interaction experiments to examine whether or not the use of a high-pitched voice by the participants him/herself can create a good impression of the robot in their minds. The results showed that the participants in our experiments seemed to be more favorably impressed with the robot when they used a high-pitched voice.

# **An explanation for status of listening to the late musical works of Morton Feldman by cognitive point of view**

**Takuro Shibayama**

Tokyo Denki University

**Tatsuji Takahashi**

Tokyo Denki University

**Abstract:** Abstract: Late musical works composed by Morton Feldman (1926-1987) have common characters, 1) a denial of the development by musical form, 2) a continuance of tranquility and monotonous texture and 3) very long duration of performance. When listening to these pieces, the listeners will just face to the each note rather than trace the musical form brought by the evaluation of result made by listener's expectation to the next musical scene in succession. Shono defines this status of listening as the process that listeners find out the meaning in each note actively, rather than understand the given meaning passively (Shono, 1991). In this study, we explain that the status of this listening are arisen by, a) forfeiture of figure and ground of time sensation, b) activation of multi and inter sensation by the forfeiture of figure and ground, and c) generation of ambiguity between listeners and musical work as an object.

# **A tool supporting conflation of video and picture for physical expression**

**Satoshi Shibuya**

Tokyo Denki University and Future University Hakodate

**Ken-ichi Kimura**

Future University Hakodate

**Abstract:** In this research, an iPad application was developed as a tool promoting physical expression. This tool can create a video picture having a unique story by conflating the pictures that a user drew to the video that one's body action was recorded. In physical expression, by repeating getting a sense by moving a body and giving meaning for that sense, it is necessary to clarify what one want to express. Placing one's physical movement in a story for oneself, it can promote to give concrete meanings for the physical movement therefore it can be useful exercise for the expression. By intuitive touch operation, this tool can perform from recording a body action to drawing the picture and making of the video picture conflated with the pictures. Therefore this tool is aimed at utilizing in physical expression workshops for apprentices.

# Narrative Blending and Tense Aspect by Learners of Korean as a Second Language

Eunji Shim

Sogang University, Seoul, Korea

**Abstract:** As reported in Shim (2011), advanced learners of Korean as a second language (L2) mostly produced background information by using various tense markers while foreground was dominantly described in past tense by learners of low proficiency. Why would proficient learners largely produce background information? This question is addressed in my paper. Unlike foreground, background information must include a description as well as views from an objective perspective, including a commentary and an explanation. It is suggested that, within the framework of Fauconnier and Turner (2002, 93-94), the perspective of the speaker as a writer in a mental space must be linked and blended with that of the other in another space by the Change and the Identity vital relations, and that the blending must be learned by the L2 learners. Viewed from this blending account, productivity of background is likely to be associated with sensitivity of these vital relations



# Reassessing the motivation effect of illustrations in text comprehension

Hideaki Shimada

Shinshu University

**Abstract:** Suppose that you find an article including an illustration that looks interesting, you might read it. My collaborator and I call this phenomenon the motivation effect of illustrations. We proposed an experimental method to evaluate the motivation effect of instructional manuals. In this method, participants were allowed to glance at one manual page for two seconds and were then asked to answer the following two questions: (a) Did the page motivate you to read? and (b) Did the page look easy to understand? The method was utilized to demonstrate that illustrations enhanced participants motivation for reading. But, in this previous study, participants glanced at the same stimuli twice or more in a single experiment, so the procedure could skew the motivation effect. The purpose of this study is reassessing the motivation effect of illustrations when the said problem is improved. As a result, findings of the study prove the validity of the motivation effect of illustrations.

# Brain activity during observation of others action in live and delayed video-mediated social interaction

Sotaro Shimada  
Meiji University, Japan

**Abstract:** Social interaction is largely facilitated by spatiotemporal contiguity among individuals. The brain area called the mirror neuron system (MNS), which is activated when an individual observes the actions of others, as well as when they perform the same action themselves, is considered to play a crucial role in social cognition. The present study investigated whether and how the MNS activity is influenced by temporal contiguity between the observer and observee: (1) live face-to-face, (2) live video-mediated, (3) 6s-delayed video-mediated, and (4) prerecorded action observation conditions. ANOVA revealed that there was a significant modulation by the conditions. MNS activity in the live face-to-face condition was significantly larger than that in the prerecorded condition. A moderate activity was observed in the live and 6s-delayed video-mediated conditions. These results indicate that MNS activity is sensitive to the temporal contiguity, which likely affects the quality of social interaction.

# Many faces of diagrams: from general properties to practical advantages and disadvantages

Atsushi Shimojima  
Doshisha University

**Abstract:** Research on diagrammatic representations has revealed that they, as opposed to sentence-based representations, have the following general properties:

(1) Expressing a certain set of information results in the expression of consequential information (e.g., Barwise & Etchemendy 1990), (2) Certain sets of information cannot be expressed alone, without the (at least selective) expression of other, non-consequential information (e.g., Stenning & Oberlander 1995), (3) Certain inconsistent sets of information cannot be expressed (e.g., Barwise & Etchemendy 1994), (4) Information of multiple levels of information can be expressed simultaneously and relatedly (e.g., Pinker 1990, Cleveland 1994).

There has been an attempt to explain the cause of these properties in a unified manner (e.g., Shimojima 2002). This presentation will focus more on their consequences, specifically, the resultant advantages and disadvantages in the actual uses of information graphics. Examples will be taken from newspaper, magazines, TV news shows, and other popular forms of information media.

# Learning Grammar via Statistical Mechanism

**WonJae Shin**

University of Notre Dame, Notre Dame, Indiana, United States

**Kathleen Marie Eberhard**

University of Notre Dame, Notre Dame, Indiana, United States

**Abstract:** Adults' learning of grammatical dependencies was investigated with an artificial language consisting of shapes (e.g., circles, squares, etc.) and a novel prediction task. The grammar yielded simple sentences and complex sentences with an embedded clause resulting in a non-adjacent agreement relation. Similar to Elman's (1993) Simple Recurrent Network (SRN), sentences were concatenated and presented one shape at a time to participants, who predicted the next shape while the preceding seven shapes remained visible. The sentences represented a staged learning condition, with simple occurring before complex, or a mixed condition with simple and complex randomly ordered. Like the SRN, the participants' token predictions were frequently incorrect; however, accuracy was assessed by whether their prediction was grammatical. Accuracy was above chance in both the staged and mixed conditions demonstrating the beneficial effect of prediction errors, and it was significantly higher in the staged condition, where the stronger local contingencies facilitated category learning.

# **Can you see things from your opponents point of view? – The relationship between critical thinking dispositions and the ability to articulate views different from ones own**

**Noriko Shingaki**

Seijo University

**Yukie Tsuzuki**

Seijo University

**Abstract:** Citizens today are increasingly called upon to critically examine current social issues from different points of view and to engage in public debate about them. Accordingly, training in critical thinking and debate is an important part of Japanese college education. In this study, 72 college students were given a newspaper article about homeless patients being "dumped" by hospitals in Los Angeles. The article was sympathetic to the patients. The students were then asked to comment on the situation from a "neoliberal" standpoint that stressed the value of competition. While 56 students had no difficulty arguing from the alternative standpoint, 16 students were unable to do so and disparaged the hospitals' immoral behavior. The latter group scored significantly lower on the "need for cognition" scale. Protocol analyses revealed that this group had trouble setting their own opinions aside when tasked with articulating views different from their own.

# Magic props induce misdirection differently from the magicians face

**Marie Shoda**

The University of Tokyo

**Kazuhiko Yokosawa**

The University of Tokyo

**Abstract:** Successfully conducting a magic performance requires that the magician guide the gaze of the audience away from the trick towards the props, which is known as misdirection. The role of the social cue, the direction of the magician's face, has been overestimated and the role of the props has been ignored until recently. We contested this assumption by using context congruent props, e.g., cards. Participants observed the same magic show twice. During the second observation, there was a mask on the magician's face, or on the prop, or there was no mask. The prop and the face independently caused misdirection during the second observation, implying that the social cue was not the only factor causing misdirection. It is suggested that top-down knowledge gradually modified the misdirection by inhibiting the fixation on the prop. We conclude that sufficient time is needed to activate top-down knowledge to cancel the misdirection.

# **Infants form expectations about others' emotions based on context and perceptual access**

**Amy Skerry**  
Harvard University

**Mina Cikara**  
Massachusetts Institute of Technology

**Susan Carey**  
Harvard University

**Elizabeth Spelke**  
Harvard University

**Rebecca Saxe**  
Massachusetts Institute of Technology

**Abstract:** Infants attend to and discriminate between other people's emotional expressions. However, it is unknown whether infants have a conceptual understanding of others' emotions and the contexts that elicit them, or merely a perceptual schema for emotional facial expressions. Using a violation of expectation paradigm, we tested whether 11 month old infants expect others' emotional reactions to be contextually congruent. Specifically, infants saw one adult (Observer) react to another adult's (Target) emotional expression. If the Observer was looking at the Target, infants expected the Observer to react congruently (positively to a happy Target, negatively to a sad Target), and looked longer at incongruent emotional reactions. Infants had no expectation of congruence if the Observer was looking away from, and could not see, the Target. This result suggests that within the first year of life, infants represent emotions in terms of the contexts that elicit them.



# **The conflict response of charitable donations on various decoy options**

**Masashi Soma**

Rikkyo University

**Itzuki Chiba**

Rikkyo University

**Yuichi Hasimoto**

Rikkyo University

**Abstract:** The purpose of this paper is to demonstrate to diminish conflict response on charitable donation by multiplying decoy option(attraction,compromise,phantom) in 2 attribute 3 option decision making. Rubaltelli & Agnoli(2011) composed 2 attribute 3 option decision making in donated victim and indicated that a conflict response had been diminished by decoy options(attraction effect). We demonstrate to decoy effect by using Rubaltelli & Agnoli(2011)s task.

# **A Neuro-Robotics Model for the Acquisition of Higher Order Concepts in Action and Language**

**Francesca Stramandinoli**

Plymouth University

**Davide Marocco**

Plymouth University

**Angelo Cangelosi**

Plymouth University

**Abstract:** Recent neurophysiological experiments have shown that motor neurons responsible for encoding specific motor acts present different activation patterns according to the final goal of the action sequence in which that particular motor act is embedded (Fogassi et al., 2005). This is consistent with the chain model hypothesis (Chersi et al., 2006) according to which the processing of action-related sentences involves the activation of the sequence of motor neurons (chain) directly involved in the sentence.

We have developed a cognitive robotic model for the learning of compositional actions from combination of motor primitives. This model uses recurrent neural networks. Simulation results have shown that motor primitives have different activation patterns according to the actions sequence in which they are contained. These results suggest that the motor chain hypothesis (Chersi et al., 2006) can be a general mechanism that explains the way in which recurrent networks represents and reuse hierarchical concepts.

# Representational Translation with Concrete and Virtual Models in Organic Chemistry

**Andrew Stull**

University of California, Santa Barbara, California, United States

**Trevor Barrett**

University of California, Santa Barbara, California, United States

**Mary Hegarty**

University of California, Santa Barbara, California, United States

**Abstract:** Organic chemists must understand the 3-D structure of molecules and be adept at relating different 2-D diagrammatic representations of molecules. Concrete (3D) models can aid chemistry students in developing these aspects of representational competence. A growing trend is to incorporate 3D virtual models into instruction. In two studies, we tested the relative effectiveness of concrete and virtual models when relating molecular representations. Participants completed tasks that involved matching either virtual or concrete models to three different types of molecular diagrams. There were no differences in accuracy for virtual and concrete models, but participants performed significantly faster with virtual models. The benefit likely resulted from the hand-held computer interface, which constrained interactivity to make the most task-relevant information salient. These results highlight the importance of interface design in promoting effective use of spatial representations, and suggest a potential benefit of teaching with virtual models rather than traditional media in chemistry.

# Language and Number Sense: ANS Representations for 'Most'

**Yasutada Sudo**

MIT

**Hadas Kotek**

MIT

**Michelle Fullwood**

MIT

**Martin Hackl**

MIT

**Abstract:** Quantificational expressions ('every', 'most', etc.) afford speakers the means to convey information about the quantity of entities pertinent to a conversation. Their formal linguistic properties have been studied in great detail (Barwise & Cooper 1981, etc.). However, little is known about how they are mapped onto cognitive representations of quantity in particular, to what extent that mapping is determined by their linguistic properties. We present evidence from a series of experiments concerning the quantifier 'most' and argue that 1) it is, as a default, mapped onto representations of the Approximate Number System (ANS) rather than the precise numbers (cf. Lidz et al. 2011); and 2) the mapping reveals cognitive correlates of the complex internal structure of 'most', which encompasses a superlative operator as well as a gradable predicate MANY expressing measure function of quantity (Hackl 2009, Kotek et al. 2011).

# **A left cerebral hemispheres superiority in processing spatial-categorical relation in a non-verbal semantic format.**

**Takashi Suegami**

University of Oslo

**Bruno Laeng**

University of Oslo

**Abstract:** It has been shown that the left and right cerebral hemispheres (LH and RH) have preferences for processing qualitative (or categorical) and metric (or coordinate) spatial relations respectively. However, categorical spatial information could be divided into semantic- and visuospatial-categorical information. We examined whether semantic- and visuospatial-categorical information were different, and whether such a distinction was differentially lateralized. We manipulated the colors and positions of the standard traffic light sign so that either the semantic- or visuospatial-categorical information changed in a non-verbal format and used these stimuli in a sample-to-match task coupled to the divided visual field method. In the semantic-categorical matching task a LHs advantage for processing semantic-categorical information was observed even in a non-verbal format. In the visuospatial-categorical matching task, however, neither left nor right visual field advantage was obtained. These results suggest the processing of semantic-categorical information is lateralized in LH and it is distinct from visuospatial-categorical information.

# **What role do you play in group activity? Objective evaluation through third parties**

**Noriko Suzuki**

Japan Society for the Promotion of Science/Doshisha University

**Tosiro Kamiya**

Osaka University

**Ichiro Umata**

NICT

**Sadanori Ito**

NICT

**Shoichiro Iwasawa**

NICT

**Mamiko Sakata**

Doshisha University

**Katsunori Shimohara**

Doshisha University

**Abstract:** Our daily life provides us a lot of opportunities to participate at work, learn and play with our peers: group learning in schools, training courses in the company, and even outdoor activities. The interactive roles of leaders and followers sometimes emerge and reorganize in our group activities. This study revealed both verbal and non-verbal cues for expressing emergent interactive roles in group activities. Objective evaluation through neutral third parties was conducted to select roles including leader, active or passive follower that the participants played in a triad assembly task. From the result of the correlation between behavioral data and third party evaluation, the interactive roles could have been characterized the amount of speech, i.e., potential leader talked more than potential followers. It was also suggested that the amount of gaze to other participants and the distance from them were useful cues for distinguishing active follower from passive one.

# Do we prefer simple realities or simple descriptions?

**Colleen Szurkowski**

University of Richmond

**David Landy**

University of Richmond

**Abstract:** Explanations play an important role in our mental lives. In most cases, however, explanations are not deducible outright; multiple explanations exist for any circumstance. Simplicity strongly biases preference. We distinguish between two cases: a preference for a simple underlying reality and a preference for a reality that can be simply described

We created a fictional disease scenario with three symptoms; participants rated possible diagnoses. One diagnosis was complex (3 viruses), but had a simple term; the other diagnosis involved 2 viruses. Participants generally favored the simple world over the more complex. Additionally, participants preferred whichever explanation was currently described using fewer terms. About half of participants opted for the simpler label when available, but the simpler explanation when unavailable. It seems that people take into account both the simplicity of the world model postulated by a theory and also the simplicity with which a theory is expressed, when deciding among possible options.



# **Developing and Testing "Visualizing Connection Note," a Constraint-Free Two-Dimensional Concept-Mapping Tool for Ideation**

**Yuisho Takafuji**

Chukyo University, Toyota, Aichi Pref, Japan

**Hajime Shirouzu**

Chukyo University

**Abstract:** Writing a research paper imposes a heavy cognitive load wherein one gathers ideas, develops relations among them, and decides their order. This complex, time-consuming process can be facilitated by collecting daily ideas and laying them out freely on a two dimensional map. We thus developed Visualizing Connection Note: ViCoNote, which has cloud application Evernote on its left side and a two-dimensional space on its right side. The user can easily draw various documents like texts, photos, and voice memos from Evernote to the space, and try to group, superimpose, juxtapose, and connect them. One participant used ViCoNote for two months as a pilot study, and found that it facilitated recording contextual information for brainstorming as well as condensing the information for concept mapping. We are now testing its effect for undergraduates writing their graduation theses.

# **Cognitive process of the children with reading difficulties: Analysis of the reading patterns with customizable digital reading software.**

**Maiko Takahashi**

Tokyo Woman's Christian University

**Mamoru Iwabuchi**

The University of Tokyo

**Kenryu Nakamura**

The University of Tokyo

**Abstract:** Children with reading difficulties, such as children with dyslexia, struggle with standard printed materials in regular classrooms. In this study, a reading support system with a text-to-speech function, Touch & Read, was developed to examine changes in textbook information comprehension among children. Using this system, readers can magnify text, or tap part of the text to listen to the words read aloud, and adjust the speed of such auditory feedback. The system can log the way that the reader uses it to be analyzed to report on points of difficult for a reader, as well as help estimate his or her reading process afterward. We provided the Touch & Read system to 175 students from the first to sixth grade in the regular class and asked them to use it in the lesson for reading comprehension. Analysis of our data shows improvements in the reading rate and accuracy of the children with reading difficulties and a consequent rise in reading scores. Through an analysis of the logs of the reading behavior, the differences of the reading process between the children with and without reading difficulties are discussed here.

# Infant biases in lexical acquisition induced by loosely symmetric reasoning

**Tatsuji Takahashi**

Tokyo Denki University

**Takumi Kamiya**

Tokyo Denki University

**Takahiro Shimizu**

Tokyo Denki University

**Shuji Shinohara**

Advanced Algorithms & Systems

**Abstract:** Four biases (or constraints) are known to guide word learning by infants: whole object bias, noun-category bias, mutual exclusivity, and shape similarity. They can be understood in relation to manifesting and invalidating the logical types between an object and the class or a token and the type. Hence we can suppose the biases have close connection to self-reference and our flexibility in thinking. Because the biases basically contradict each other, there must be an adjusting mechanism. We propose the loosely symmetric (LS) model as the plausible mechanism. LS, a heuristics describing human symmetric cognitive biases in the form of conditional probability, is shown to be effective in an ample amount of areas including causal induction, learning (reinforcement, supervised and unsupervised), game-theoretical situations and digital game AI. Infant agent inferring what to learn with LS shows not only efficient word learning but also appropriate use of the acquired knowledge for communication. The agent also shows appropriate adjusting of which bias prevails.

# Color Affects Face Perception in Schematic Faces

**Fumiyo Takahashi**

Hokkaido University, Sapporo, Hokkaido, JAPAN

**Yasuhiro Kawabata**

Hokkaido University, Sapporo, Hokkaido, JAPAN

**Abstract:** Previous research has revealed color effects in scene recognition and in object recognition by diagnostic tasks but only for particular objects with typical colors, whereas Changizi (2006) showed development of trichromatism in primates in terms of the evolution by computational models. Accordingly, it suggests the importance of face color as a social signal. In our previous study (2009), we found the tendency to typical colors for each emotion: red for anger, bluish colors for sadness, and yellow, orange and pink for joy. In this study, we investigate the color effects on cognition of emotional facial expression in colored schematic faces, which can represent prototypes of emotional facial expression and control the degree of change for each part; eyes and mouth. As a result, consistent colors for each emotional expression facilitated the task performance.

# Neural substrates of grammatical information retrieval during sentence comprehension

**Kei Takahashi**

Tohoku University

**Satoru Yokoyama**

Tohoku University

**Toshimune Kambara**

Tohoku University

**Ryuta Kawashima**

Tohoku University

**Abstract:** To comprehend a sentence, retrieval processes of both grammatical and lexical-semantic information from memory are required. However, previous studies had mainly focused on the maintenance process. The current fMRI study investigated the existence of information retrieval process on grammatical and lexical-semantic processing by manipulating retrieval load but by controlling maintenance load.

Two tasks have prepared; in the grammatical task participants read ungrammatical Japanese coordination construction sentences and judged in which conjunct incongruent item had appeared, and in the lexical task, the participants remembered six Japanese words and judged when the lastly displayed word appeared.

The result showed significant brain activity in the superior part and inferior part of left middle frontal gyrus in the grammatical task and in the lexical task, respectively ( $p < .001$ , uncorrected). This indicates the existence of retrieval system independently from maintenance process. In addition, grammatical information retrieval is processed distinctly from lexical-semantic information retrieval.

# **Soccer as Social Interaction between Observable Bodies**

**Katsuya Takanashi**

Japan Science and Technology Agency / Kyoto University

**Kazuki Sekine**

Japan Society for Promotion Science / University of Birmingham

**Abstract:** What is the most essential in soccer is that a player's orientation as to how s/he is about to play is expressed through bodily movements and becomes observable for other players in the process of game. And, this is not only true for soccer, but also for social interaction in general such as conversation. Therefore, it is indispensable for cognitive science of social interaction to build a methodology for grasping what a participant in interaction observes in other participants' behaviors in the on-going process of interaction as a starting point for modeling cognitive competence behind such behaviors. This presentation introduces an analytic framework for soccer in which soccer is conceived as a kind of social interaction and bodily movements in it are analyzed with reference to several analytic concepts developed in conversation analysis and gesture studies which can describe details of sequential organization of social interaction.

# **Tactile sensation and onomatopoeia in Japanese**

**Yufuko Takashima**

Dokkyo University, Soka, Saitama, Japan

**Abstract:** This paper explores words of tactile perception in Japanese investigated by using an elicitation task. Japanese language contains a lot of onomatopoeic words and has been focused on its expressiveness. On the other hand, their accuracy in haptic domain has not been often discussed. To examine the role of onomatopoeic words in tactile expression, I have used a task developed by MPI language and perception project (Majid and Levinson 2007) using a texture booklet contains ten different textures. The investigation has been held in a small village among mountains. The equality of their way of life is helpful to avoid over-generalization. Consequently, the number of texture onomatopoeic words is not enough to express tactile sensation accurately: one forth of the token contains onomatopoeic words in contrast to half of expression to try to guess the source in detail. Some onomatopoeic words are used to express more than one stimulus.



# **A remotely operated robot as a research tool to study the effects of different roles for successful collaborative learning**

**Nakayama Takayaro**  
The University of Tokyo

**Ashikaga Jun**  
The University of Tokyo

**Inaba Sho**  
The University of Tokyo

**Iyoki Kenta**  
The University of Tokyo

**Naomi Miyake**  
The University of Tokyo

**Abstract:** In collaborative learning situations, remotely operated robots can be used to deliver the same information to different groups, so that we can explore the roles of partners in collaboration. Using this technique, we have tested 25 collaborative learning cases with eighty 5th and 6th graders on science topics. Even though the basic script for running these experimental classes was kept constant and the children learned successfully from the first case, we have found that the operators behavior changed over the course. For the first nine cases, the robots were operated as a moderator. In the following eight cases, the operators mixed this style with operations where the robot acted more like an equally capable peer. Observing some success for soliciting more active collaboration by this mode of operation, in the recent eight cases, the operators have begun to explore different possibilities of kids roles.

# **The Influence of Cognitive Functions to Acquire Nursing Skills for Patient Transfer**

**Keiko Takeda**

Nihon Fukushi University

**Yoriko Watanabe**

Seirei Christopher University

**Taeko Harada**

Nihon Fukushi University

**Abstract:** This study aims to elucidate the influence of cognitive functions on nursing skills for patient transfer. We examined the relationship between Go/No-go task (GNG) performance and self-evaluation score in the nursing care action, magnitude of low back strain during patient transfer, and decision making time for starting patient transfer. Twenty university students (mean age 190) participated in this study. The results showed that the reaction times in the GNG was correlated with the magnitude of low back strain and the decision making time. As GNG performance is related to decision making for action selection, this cognitive process appears to be associated with nursing skill. This is observed via the magnitude of low back strain and decision time for better transfer action selection. Thus, cognitive process for selecting the nursing care action may be related to the nursing skills.

# **The effect of 3D stereoscopic display on spatial cognition: a near-infrared spectroscopy study**

**Yoshiyuki Tamamiya**

University of Tokyo

**Kazuo Hiraki**

University of Tokyo

**Abstract:** 3D stereoscopic displays have become popular in our life. Some studies show that this new technology affects our ability to recognize objects. However, it is not well known how our brain is modulated by the technology. In the current study, eighteen healthy adults participated. They were asked to make blocks same as pictures shown on a monitor. The task was similar to an intelligence test. There were two sets of pictures, 2D and 3D. The stimuli were presented through the Nvidia 3D vision system. During the task, activation in parietal area, which is related to spatial cognition, were recorded by a functional near-infrared spectroscopy (NIRS). Compared to 3D pictures, 2D pictures more activated parietal area. The result indicated that the task with 2D pictures required more spatial cognition to compute depth information of the blocks from 2D pictures, which leads to activation of parietal area.

# Understanding displacement of communication by graphical communication tasks

**Kaori Tamura**

Japan Advanced Institute of Science and Technology

**Takashi Hashimoto**

Japan Advanced Institute of Science and Technology

**Abstract:** One important feature of human linguistic communication is displacement, to communicate about absent objects. However, it has not been revealed that which aspects of displacement are unique to human beings. We claim that displacement should be considered in the context of communication and distinguished from memory to this end. We designed a graphical communication task for displaced communication based on Fay et al (2003). Two participants are paired and communicate using electric drawing pads in separate rooms. In the experiment, we investigated speakers' devices on expression to make listeners understand absent objects. We compared two different kinds of tasks: one corresponds to displacement based on memory and the other displacement not based on listeners' experience. The result shows that speakers tend to use placeholders significantly more to express events that are not based on listeners' experience. Placeholders are alternative expressions that typically have the same features with absent objects.

# Effect of embodied cognition in insight problem solving

**Masahiko Tamura**

Nagoya University, Nagoya, Aichi, Japan

**Kazuhisa Miwa**

Nagoya University

**Abstract:** We confirmed that embodied cognition improves relaxation of the fixation in insight problem solving. In this study, the embodied cognition was manipulated with eye movements in an eye-tracking task. To achieve a solution, a problem solver is required to relax the fixation that guides search in an incorrect problem space, and to shift the search into a correct one. We conducted two experiments. The participants' eye movement was controlled using an eye-tracking task prior to or during insight problem solving. Experiment 1 confirmed that the tracking task performed prior to the insight task reduced the participants' fixation guiding the incorrect search. In Experiment 2, the participants who engaged in the tracking task during solving the insight task found a correct target faster just after beginning to generate crucial hypotheses that violate the incorrect fixation than those who did not. We concluded that a stimulus accompanying embodied cognition affects fixation generation and also reduces the intensity of fixation once generated in insight problem solving.

# **For female eyes only: Comparing the effects of enlarging eyes and irises on facial attractiveness**

**Azumi Tanabe-Ishibashi**

Kyoto University, Kyoto-shi, Kyoto , Japan

**Mikina Takahashi**

Kyoto University, Kyoto-shi, Kyoto , Japan

**Maya Katsuhara**

Kyoto University, Kyoto-shi, Kyoto , Japan

**Kana Kuraguchi**

Kyoto University, Kyoto-shi, Kyoto , Japan

**Hiroshi Ashida**

Kyoto University, Kyoto-shi, Kyoto , Japan

**Abstract:** Various beauty products have been developed motivated by our general concern to facial attractiveness. The eyes and irises are quite often the main target of those cosmetic attempts to increase beauty of female faces. Many women use cosmetic items to make eyes look larger, but some Japanese women use contact lenses to make irises look larger. In this research, we compared rated attractiveness of female faces with enlarged eyes and enlarged irises. We also compared those faces in terms of likability and childishness. The result showed that faces with enlarged eyes were rated more attractive, while faces with enlarged irises were perceived more likable and childish. Moreover, female participants evaluated faces with enlarged irises more attractive, though males did not. We suggest that the eye-makeup to enlarge the appearance of irises has the effect of showing the female faces more attractive only for the viewers of the same sex.

# Gender difference of social interaction behavior in child's game

**Daisuke Tanaka**

Tottori University, Tottori, Japan

**Shinako Terakawa**

Tottori University, Tottori, Japan

**Ayumi Seki**

Tottori University, Tottori, Japan

**Hitoshi Uchiyama**

Tottori University, Tottori, Japan

**Tatsuya Koeda**

Tottori University, Tottori, Japan

**Abstract:** It is known that the Ultimatum Game is an index of human social interaction behavior with regard to inequality aversion, correction of illegitimacy and egalitarianism. Previous research found that preference for fairness tends to occur with Theory of Mind. In general, children enhance their sociability through playing with peers. However, taking into account gender difference of playing style preference, it is possible to predict that social abilities differ by gender. In this research to find a gender difference of social behavior in playing with peers, new game based on the Ultimatum Game was constructed. Participants were 9 or 10 years old and played with 3 other same gender peers (one is a friend and two are foes) according to some rules. As a result, male tends to value his friend in comparison with female. Predictable gender difference of social abilities caused by observed behavioral traits was discussed.



# **Assignment of accent patterns to nonword items in a rapid reading task by Japanese speakers of the Kansai dialect**

**Yuki Tanida**

Department of Cognitive Psychology in Education, Kyoto university, Yoshida Hon-machi, Sakyo-ku,  
Kyoto, Japan

**Yoko Higuchi**

Graduate School of Human and Environmental Studies, Kyoto University

**Yuri Yano**

Department of Cognitive Psychology in Education, Kyoto university, Yoshida Hon-machi, Sakyo-ku,  
Kyoto, Japan

**Satoru Saito**

Department of Cognitive Psychology in Education, Kyoto university, Yoshida Hon-machi, Sakyo-ku,  
Kyoto, Japan

**Abstract:** Previous studies have reported the strong influence of phonotactic knowledge on a variety of language tasks. For example, nonwords composed of phoneme combinations that occur frequently in a particular language are recalled accurately in a short-term memory task. In this study, we investigated the role of another type of phonological knowledge, prosodic knowledge, which has thus far been neglected. For this purpose, an accent-pattern assignment task with tri-mora nonwords was performed by Japanese speakers of the Kansai dialect under the assumption that the phonological system might assign the typical accent pattern, which reflects accumulated knowledge about accent patterns that occur in the language environment, to non-lexical items. However, the proportion of accent patterns produced did not reflect the frequency of each accent pattern in the Kansai dialect; instead, it reflected the proportion in the Standard Japanese language, to which all Japanese speakers have presumably been exposed by the broadcast media.

# Experience-based modulation of eye-movement behaviour in dynamic and uncertain visual environments

**Shuichiro Taya**

Taisho University

**David Windridge**

University of Surrey

**Magda Osman**

Queen Mary, University of London

**Abstract:** We examined the influence of experience in tennis on eye-movement behaviour. Participants experience with the sport was recorded, along with the number, size, and accuracy of saccadic eye-movements made around 'ball events' (i.e. hits and bounces) while watching the clips of tennis match. Overall, observers with richer experience about tennis relocated their eyes quicker and closer to the upcoming event location as compared with observers with relatively poor-experience of the sport. Moreover, even though repeated exposure to the same clips increased the efficiency of eye-movements in non-experienced observers, the main differences (e.g., saccadic amplitudes) between the two observer groups was preserved. Importantly, the influence of experience was predominantly found in ball events with higher uncertainty (bounces). These results suggest that our gaze control system utilizes knowledge accumulated through past experience to anticipate upcoming events, and this helps especially when the visual event that occurs is relatively unpredictable.

# **Constructing Social Attitudes through Persuasive Writing Training: A Cognitive Approach in Undergraduate Education of Engineering Ethics**

**Emiko Tayanagi**

Future University Hakodate

**Abstract:** This study aims to seek a new approach of educational method to enable students to cultivate ethical mind as professional engineers through strategically designed course program including persuasive essay writing. On the assumption that the social psychology theories of role-taking and attitude change derive effective basis to practice such education programs, the study develops a theoretical framework for education, a syllabus of the class and a series of short essay questions to exercise cognitive abilities concerned with ethical practices. Findings from qualitative data analysis of descriptive texts by students show that this framework succeeds in providing effective training opportunities, in which students try to learn how to construct social attitudes for ethical decision making within difficult and complex situation overcoming conflict between opposing factors such as individual and organizational value. In conclusion theoretical and practical implications of the study combining modern ethics and cognitive science would be discussed.

# **The nature of the training effects of mental rotation: the limit for transfer to novel orientation**

**Haruna Terada**

University of Tsukuba

**Hiromi Morita**

University of Tsukuba

**Abstract:** We investigated the training effects of mental rotation of two-dimensional figures. 35 participants practiced rotating the upright image into the orientation between 0 to 180 with a set of 4 figures. Then they were tested on the mental rotation into the new orientation between 0 to 180 with the same set of figures (new orientation condition), on the rotation of the upside-down image into the orientation between 180 to 360 with the old figures (novel orientation condition), and on the rotation with the different set of figures (novel figure condition). There was no difference in rotating rate between the novel orientation condition and the novel figure condition. The rotating rate for these conditions was significantly lower than that for the new orientation condition. The results suggest that the training effects do not transfer beyond the path through which participants rotated images in the practice.

# **Multi-platform Experiment to Discuss Behavioral Consistency across Laboratory and Real Situational Studies**

**Hitoshi Terai**  
Nagoya University

**Kazuhisa Miwa**  
Nagoya University

**Hiroyuki Okuda**  
Nagoya University

**Yuichi Tazaki**  
Nagoya University

**Tatsuya Suzuki**  
Nagoya University

**Kazuaki Kojima**  
Waseda University

**Junya Morita**  
Japan Advanced Institute of Science and Technology

**Akihiro Maehigashi**  
Nagoya University

**Kazuya Takeda**  
Nagoya University

**Abstract:** We constructed an innovative experimental platform to discuss behavioral consistency in driving behavior. In our experiment, the participants were required to engage a vehicle handling task using three different systems: the real (an electric vehicle), virtual (a driving simulator) and laboratory (a computer monitor and a game pad controller) systems. The results are summarized as follows. 1) In the real system, the behavioral consistencies within participants were different among three fundamental behaviors (brake, accelerator, and steering). Whereas the consistency of the brake behavior was the lowest, the steering behavior was the highest. 2) The same pattern of consistencies in the real system was confirmed in the virtual and laboratory systems. 3) The pattern of consistencies between the real and virtual systems was similar to the real system. On the other hand, the pattern of consistencies between the real and laboratory systems was different from the real system.

# **An Exploration of Crossword Skill**

**Kejkaew Thanasuan**

Michigan Technological University

**Shane Mueller**

Michigan Technological University

**Abstract:** Past research has established how different lexical skills (word-stem completion, anagram solutions, etc.) correlate with crossword puzzle expertise. However, little is understood about the cognitive processes involved in crossword play. We hypothesize that clues are solved either through an semantic or an orthographic route, but not both simultaneously. To help us test and develop this model, we conducted an initial study examining novice crossword puzzle performance, along with a secondary word-stem completion task. We found that our participants word-stem completion correlated reliably but modestly with puzzle solution, suggesting that a number of other factors may be involved. Moreover, in our study, most responses (66%) were made with little or no orthographic constraints, suggesting that novice crossword players relied heavily on the semantic route for solving clues. The results show evidence for both routes, and provide direction for testing whether a dual-route interactive memory search is possible.

# Musical thoughts behind composer's writings

**Akifumi Tokosumi**

Tokyo Institute of Technology

**Akihiro Kawase**

National Institute for Japanese Language and Linguistics

**Abstract:** Two analysis methods were applied to the essays written by Toru Takemitsu and Pierre Boulez, two prominent contemporary composers of classical music. The purpose of the network analysis was to identify keywords and their surrounding words within the text. After applying a morphological analysis to the text, a home-brewed network creation software was employed to create a network for the extracted keywords. There are clearly several node clusters and a number of measures of centrality for the network indicate that the multi-centrality of the keyword space. The text corpus for the essays was then parsed in order to carry out a content analysis at the semantic level. The aim of the content analysis was to extract the structures within the concepts employed by the composers in talking about music. The interesting results include the findings that a) their aesthetic vocabulary is strongly associated to an abstract thinking vocabulary, and b) more ordinary emotion words tend to be associated with lower level music entities. These findings seem to substantiate the layered model of affective processes proposed in our previous report. With methodological perspectives we will argue for the following points; a) Network analysis of the words within a text can provide a better basis for text analysis. b) The modeling of affective process and text analysis may be mutually beneficial.



# **Not just for consumers: Data and theory show that context effects are fundamental to decision-making**

**Jennifer Trueblood**

Indiana University

**Scott Brown**

University of Newcastle, Australia

**Andrew Heathcote**

University of Newcastle, Australia

**Jerome Busemeyer**

Indiana University

**Abstract:** Context effects—preference changes depending on the availability of other options—have attracted a great deal of attention among consumer researchers studying high-level decision tasks. Our experiments show that all three context effects from the preferential choice literature—similarity, attraction, and compromise effects—also arise in inference tasks and simple perceptual decision-making tasks. These experiments provide evidence that the effects are not confined to high-level decision tasks where the options have hedonic values such as consumer products. A new model of multi-alternative, multi-attribute choice is also developed. This model, named the Multi-attribute Linear Ballistic Accumulator (MLBA) model, extends the Linear Ballistic Accumulator model (Brown & Heathcote, 2008) and postulates that context effects arise from a combination of cognitive components including attention, a contrast mechanism, and sensitivity to indifference/dominance. The MLBA model has an analytical solution making it computationally easier to fit to experimental data than previously proposed stochastic models.

# **A collinear distractor impairs local element search regardless of its probability occurrence**

**Chiahuei Tseng**

Department of Psychology, The University of Hong Kong, Hong Kong, Hong Kong

**Li Jingling**

Graduate Institute of Neural and Cognitive Sciences, China Medical University, Taichung, Taiwan

**William Oh**

Department of Psychology, The University of Hong Kong, Hong Kong, Hong Kong

**Abstract:** Salient distractors draw our attention spontaneously even when they do not facilitate our target search. When that occurs, targets close to or overlapping with them are detected and discriminated faster. However, an opposite impairment effect is observed when the salient distractor is a column of continuous linear bars (Jingling and Tseng, 2012). One possible explanation is that observers optimize their search strategy by directing their attention away from collinear distractors but toward the area where targets are six times more likely to appear. We tested this hypothesis by arranging targets to overlap with collinear distractor columns for 60% of the trials. The same search impairment on targets overlapping with or near the collinear distractor persists, which is against the probability hypothesis. Our result suggests that the origin of this effect is at a sensory processing stage not dependent upon information to its probability occurrence.

# Can Incubation be efficient in Reviewing?

Mio Tsubakimoto

Future University Hakodate, Hakodate, Hokkaido, Japan

**Abstract:** By examining eye movements, we reveal an effect of incubation in the reviewing process of creative problem solving. Although previous research on reading comprehension considers eye movements, the reviewing process, an important stage in writing, has not been focused on. We constructed themes with different familiarities and monitored eye movements during the reviewing process on a computer screen to record saccade, fixation, and the degree of attention paid to blank spaces. While recording the reviewing, we focused on distinguished macrostructure changes to compare the quality of texts after incubation-facilitated reviewing. Additionally, we conducted a questionnaire survey on metacognition and a semi-structured interview regarding cognitive strategy in reviewing. We examined the correlations between theme familiarities, metacognition, the quality of reviewed text, and eye movements. We propose an instructional method in writing education to reveal the correlations between the reviewing aspect of complex problem solving and incubation, a semantic cognitive process.

# **Constraint discovery hint versus constraint relaxation hint in solving insight problems**

**Syoichi Tsujii**

Chyukyo University

**Syoji Hamaguchi**

Chyukyo University

**Syota Chimura**

Chyukyo University

**Hajime Shirouzu**

Chukyo University

**Abstract:** Initial failures in solving insight problems commonly cause the solver relax his problem constraints. However, such initial trials also produce usable constraints that can facilitate further solutions.

We used a tangram puzzle (constructing a T shape from four pieces) to prepare three conditions: a constraint discovery condition wherein our confederate (acting as a member of paired participants) regularly provided two combined pieces as a hint, a constraint relaxation condition wherein the confederate regularly violated the participants constraints on how to place pieces, and a control (natural collaboration) condition. Twenty four participants were divided to one of three conditions.

We found no significant difference in the number or average time of successful solutions among the conditions. However, the constraint discovery condition outperformed the other conditions in correctly recalling the solution, indicating that, when people discover usable constraints, they retain them as basic knowledge pieces.

# Heart rate synchronization in collective creative construction tasks

**Kristian Tyn**

Aarhus University

**Riccardo Fusaroli**

Aarhus University

**Abstract:** What does it mean to cooperate? In this paper we explore the effects of cooperation on heart rate. We argue that in cooperative contexts participants synchronize their heart rhythms according to two factors: the affordances of the task at hand and the gradual consolidation of collaborative practices. We instructed 6 groups of participants to construct LEGO models of six abstract notions, both individually and in groups. We employed recurrence analysis techniques to quantify the mutual adaptability of heart rates among the participants in the different tasks. During individual tasks individual heart rates synchronized both within and between groups (but not with controls), plausibly due to the affordances of the task at hand. Additionally, during collective, but not individual tasks, within group synchronization grew over time. We finally discuss how these measures of synchronization relate to the participants engagement in the tasks at hand.

# **A Longitudinal Study on the Development of Taiwanese Childrens Use of Causal and Anaphoric Cue**

**Yuhtsuen Tzeng**

National Chung Cheng University

**Chiung-hsien Tsai**

Chung Hwa University of Medical Technology

**Abstract:** Making causal connections and anaphoric inference referring expressions is important in forming a coherent discourse representation. Causal inference is central to the representation of reading comprehension. Zero pronoun sentences are grammatical in Mandarin Chinese. Therefore, this longitudinal study focused on the development of inferring causal and anaphoric coherence among elementary school children in Taiwan.

In this two-year study, we reported data from 338 children in three age groups: 104 grade second, 123 grade third and 111 grade forth children in the first year. For each subject, 36 experimental short texts and 18 fillers in randomized order were implemented in Experiment Builder. We manipulated causality (high or low) and anaphoric resolutions (overt or zero pronoun) in texts.

The patterns in this study indicated the importance of causality for the on-line inferences. There were steady developments of causal and anaphoric inference in three age groups during the study periods.

# Measuring Learners' Awareness through Persona-Conjoint Method

**Hikaru Uchida**

Tokyo Institute of Technology, Yokohama-shi, Japan

**Akiko Orita**

Keio University, Fujisawa-shi, Japan

**Masaaki Kunigami**

Tokyo Institute of Technology, Yokohama-shi, Japan

**Takao Terano**

Tokyo Institute of Technology, Yokohama-shi, Japan

**Atsushi Yoshikawa**

Tokyo Institute of Technology, Yokohama-shi, Japan

**Abstract:** This paper proposes a novel method: "Persona-Conjoint Method"(PCM), which intends to measure the effect of learning of complex situation in a business process. In practice, such as Case Learning, learners must understand how to act at a specified complex situation, such as marketing negotiations. The method is characterized by 1) a persona-set prepared by an appropriate orthogonal array used in conjoint analysis, and then 2) the measurements for changes of evaluation viewpoints of learners. Therefore, the proposed method is able to quantitatively detect and evaluate learners' awareness about the situations difficult to detect so far. From the learners' experiments, the paper describes experimental results using Manga textbook with narrative approach. The method has revealed a set of alternative personae for the particular character of the narrative material.



# **Inhibitory control in event-based prospective memory task: An examination using the retrieval-practice paradigm**

**Kenta Utsumi**

Graduate school of education, Kyoto University

**Satoru Saito**

Graduate school of education, Kyoto University

**Abstract:** The nature of forgetting in a prospective-memory (PM) task was examined through the retrieval-practice paradigm. In two experiments, participants studied a series of category-exemplar pairs that belonged to one of eight categories (e.g., FRUIT-apple) and then engaged in retrieval practice for three members in each of four categories. At the final test phase, every participant was required to detect target items that belonged to one practiced and one unpracticed category while performing an ongoing task. Results showed the worst detection performance in response to the unpracticed items in the practiced categories versus that in response to the unpracticed items in the unpracticed categories. This retrieval-induced forgetting was smaller for the detection of focal targets, for which the processing largely overlapped with the processing of the ongoing task (Experiment 2), than for non-focal targets, for which the processing was performed independently from that of the ongoing task (Experiment 1).

# Metacognition in children is specific to domain knowledge

**Vy Vo**

University of Rochester, Rochester, NY, United States

**Rosa Li**

Duke University, Durham, NC, United States

**Nate Kornell**

Williams College, Williamstown, MA, United States

**Jessica Cantlon**

University of Rochester, Rochester, NY, United States

**Abstract:** Metacognitive skills have been shown to facilitate learning. But do they develop globally or within content domains? We used an objective measure of metacognition in two different domains to investigate this question in children. 25 subjects (5 to 8 y.o.) made numerosity judgments (which picture has more dots?) and emotion judgments (which picture looks happier?) on a touchscreen. After each judgment, they placed a metacognitive bet on their accuracy using a token economy, with immediate feedback (+3/-3 high risk, +1/-1 low risk). We measured metacognition by a phi correlation of risk choice and accuracy, finding that children were significantly metacognitive on both tasks. Higher metacognitive scores on the numerosity task, and not the emotion task, predicted mathematical intelligence. However, metacognitive scores did not predict other measures of ability, such as general IQ. Our study provides evidence that metacognition develops in tandem with domain-specific knowledge, rather than globally.

# **A study of cognitive style, visual attention distributions and achievement of Web-based multimedia recipes learning**

**Ching-Yeh Wang**

Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology

**Meng-Jung Tsai**

Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology

**Abstract:** This study aims to explore, based on eye-tracking data, the effect of cognitive style on attention distributions and learning achievement in web-based multimedia recipe learning. The study uses a one way (verbal style V.S. visual style) quasi-experimental design. Subjects are 29 students of hospitality. The treatments of the study multimedia recipes are including static and dynamic representation. Students prior knowledge and learning achievement are assessed via a survey. Result of this study reveals that students cognitive styles have an intimate relationship with their visual attention distributions in reading web-based recipes. And gender could play an important role in this learning context. In addition, it shows that it is effective to improve students immediate learning achievement and retention of learning when the procedural knowledge of recipes is represented by web-based multimedia. Specifically, the static multimedia materials improve students immediate learning achievement while the dynamic multimedia materials improve their retention of learning.

# **Does self extend to video game avatars? An ERP study**

**Veronica Weser**

Vassar College

**Ken Livingston**

Vassar College

**Abstract:** The extended mind hypothesis suggests that the boundaries of self are extended to that which can be controlled directly. Avatars in video games fit this description but are not embodied in the way generally assumed by EM theory, thus providing an opportunity to test the generality of the hypothesis in a domain that actually invites mental extension of self beyond body boundaries. We compare ERP responses to images of self, own avatar, other avatar, own possessions (real & virtual), and other possessions (real and virtual) after participants have spent three weeks playing both Second Life (avatar designed and directly controlled by player) and Sims 3 (avatar generally configured but not directly controlled). We test the hypotheses that N170 waveforms are distinct for images of the self and of others, while N250 is different for personal possessions and unfamiliar objects and P300 functions as an index of attention to self-relevant stimuli.

# **Social information aids accuracy but hinders adaptation**

**Thomas Wisdom**

Center for Experimental Research in Social Sciences, Hokkaido University

**Keigo Inukai**

Department of Economics, Hokkaido University

**Wataru Toyokawa**

Department of Behavioral Science, Hokkaido University

**Kameda Tatsuya**

Center for Experimental Research in Social Sciences, Hokkaido University

**Abstract:** We introduce a simulated investment task combining a temporally varying "market" environment with a 6-armed bandit using heterogeneous payoff distributions. It includes uncertainty about the binary state of the market, as well as which of several options yields the best payoff under each hidden market state. The mean performance of grouped participants (who could view peers' choices) did not change over the course of the session, while isolated participants had lower initial performance but higher final performance than grouped participants. Further analysis showed that grouped participants benefited from relatively accurate but low-risk-biased social information, while isolated participants developed a higher tolerance for ambiguity, reducing their use of costly prediction for high-risk choices more than did grouped participants. Our results imply that social information can cushion individual performance under uncertainty, but may hinder learning and adaptation to a dynamic environment.

# Searching for something familiar or novel: ERP correlates of top-down attentional selection for specific items and categories

**Rachel Wu**

Birkbeck, University of London

**Gaia Scerif**

University of Oxford

**Richard Aslin**

University of Rochester

**Tim Smith**

Birkbeck, University of London

**Martin Eimer**

Birkbeck, University of London

**Abstract:** Visual search is often guided by top-down attentional templates that specify target-defining features. But, search can also occur at the level of object categories. We measured the N2pc component, a marker of attentional target selection, in a visual search task using familiar and novel targets. Targets were defined either categorically (e.g., any letter), or at the feature level (e.g., the letter C). An N2pc was elicited during category search, in both familiar and novel contexts, indicating that even when targets are only defined at the category level, they are selected at early sensory-perceptual stages. However, the N2pc emerged earlier and was larger during feature-based search, demonstrating the superiority of attentional guidance by feature-specific templates. Moreover, category search triggered feature-specific templates, while the inverse was not the case, suggesting that higher-order search templates automatically include lower-order templates.

# Practicing Off ice Collaborative Learning in a University Ice Hockey Team

**Masayuki Yamada**

The Center for the Promotion of Integrated Sciences, The Graduate University for Advanced Studies

**Masaki Suwa**

Faculty of Environment and Information Studies, Keio University

**Abstract:** This research is a case study of collaborative learning in a university ice hockey team. Collaborative learning were held twice a week for about two months. There were three teams of collaborative learning, each group consisting of five participants. Each group selected a theme to be discussed concerning the strategy for playing a game, analyzed videos, and made a presentation about their analysis. They had been writing memos of meta-cognitive thoughts about ice hockey on a daily basis before the meetings. We analyzed their meta-cognitive writing before and after their practice. Moreover, we carried out game analysis in order to assess the effect of performance as a team. As a result, new descriptions, had not appeared before this study, arised in their meta-cognitive writing. And it indicate a change in their behaviors. Also, the game analysis indicates that their performance improved from before the practice.



# **Design of motion using mimetic words**

**Kaori Yamada**

Kobe University, Kobe, Hyogo, Japan

**Toshiharu Taura**

Kobe University, Kobe, Hyogo, Japan

**Yukari Nagai**

Japan Advanced Institute of Science and Technology

**Abstract:** Today, with several media of expression being available, the field of design has come to address more dynamic and impressive objects such as sound and computer graphics. In this study, we attempt to design such a creative and emotional motion that resonates with deep feelings that are difficult to verbalize using mimesis in the Japanese language and we believe that mimetic word is capable of expressing deep feelings. While onomatopoeic (sound-symbolic) words imitate actual sounds, mimetic (reality-symbolic) words, for example, shiku-shiku (sobbing) and kune-kune (wriggling), express appearance, movement, feeling, and other phenomena. We extracted mimetic words which indicate movement from other mimetic words. We showed how using mimetic words can help identify suitable motions that can be combined to create a new creative and emotional motion. This motion can be applied to such elements as animated logos and be used to create appealing advertisements.

# Stochastic dynamics hidden in Japanese martial arts

**Yuji Yamamoto**

Nagoya University

**Motoki Okumura**

Shizuoka University

**Akifumi Kijima**

Yamanashi University

**Keiko Yokoyama**

JSPS Research Fellow, Hokkaido University

**Koji Kadota**

Osaka University

**Hiroo Suzuki**

Nagoya University

**Kazutoshi Gohara**

Hokkaido University

**Abstract:** Martial arts like judo or Japanese fencing (kendo) are considered typical interpersonal competitions of human motor behavior. This kind of competition requires one to attack an opponent while simultaneously avoiding the opponent's attack. To quantify the practical behavior of interpersonal distance (IPD) between two players from the viewpoint of a dynamical system, we observed players' movements in kendo matches by a motion capturing system. The participants were twelve college athletes whose team had won in All Japan championships. The time series of the IPD from the coordination phase to the attacking phase was extracted, and 419 scenes were analyzed. A return map analysis was applied to the data using peak detections. We could describe the interpersonal competition during a kendo match in terms of state transitions, as in a Markov process. This suggests that complex human movements in interpersonal competition are self-organized by a simple principle. Martial arts like judo or Japanese fencing (kendo) are considered typical interpersonal competitions of human motor behavior. This kind of competition requires one to attack an opponent while simultaneously avoiding the opponent's attack. To quantify the practical behavior of interpersonal distance (IPD) between two players from the viewpoint of a dynamical system, we observed players' movements in kendo matches by a motion capturing system. The participants were twelve college athletes whose team had won in All Japan championships. The time series of the IPD from the coordination phase to the attacking phase was extracted, and 419 scenes were analyzed. A return map analysis was applied to the data using peak detections. We could describe the interpersonal competition during a kendo match in terms of state transitions, as in a Markov process. This suggests that complex human movements in interpersonal competition are self-organized by a simple principle.

# A Classification of Manner Adverbs in Korean: A Frame-based Approach

Myung-Jin Yang

Hankuk University of Foreign Studies

**Abstract:** There have been difficulties in providing a unified account of Korean adverbs. For example, in contrast to English, most manner adverbial expressions in Korean do not occur as adverbial complements to verbs, denying one way to characterize this type of expressions. The aim of this paper is to provide some criteria for classifying and categorizing manner adverbs in Korean based on frames. In the first part of this paper, we lay out the general syntactic position and semantic interpretation of Korean manner adverbs. In the second part, we present a systematic description and classification of manner adverbs using a formal apparatus based on the Operator and Davidsonian approaches (Schaefer, 2004). Finally, we propose appropriate frame structures for adverbs of manner, making reference to the Berkeley FrameNet (BFN), which we show are useful in identifying the frame types of manner adverbs and in sub-classifying individual lexical units (esp., verbs) in Korean test data.

# **Brain activities for different cohesion type on discourse comprehension**

**Ken Yasaka**

Tohoku University

**Satoru Yokoyama**

Tohoku University

**Kei Takahashi**

Tohoku University

**Ryuta Kawashima**

Tohoku University

**Abstract:** In order to comprehend discourse, the sentences in the discourse should be appropriately combined. What combines two sentences with each other is called cohesion. Theoretically, there are four types of cohesion conjunction, ellipsis, lexical cohesion, and reference. Previous neuroimaging studies investigated brain activities for cohesive and incohesive discourse, but no studies distinguished the types of cohesion. This study compared the brain activities of four cohesion types using fMRI. Participants were asked to read cohesive/incohesive discourses, each of which were combined with one of four types of cohesion, and to judge whether the discourse was coherent or not. There was no significant difference in brain activation among the four cohesion types, suggesting that the different cohesion types are recognized by a common brain mechanism.

# **Ad hoc creature: Lost and added in translation from description to depiction**

**Sachi Yasuda**

The University of Tokyo

**Masashi Okamoto**

Ritsumeikan University

**Eiji Aramaki**

The University of Tokyo

**Abstract:** When asked to draw a rabbit, almost all people can draw a rabbit correctly. In contrast, is it possible that using only descriptions of appearance features of a rabbit could achieve the same task?

We prepared a description of rabbit from a dictionary (230 words), and masked rabbit in it. The description contains 40 appearance features of rabbit. Then, subjects (n=41) were asked to draw a creature from the description.

As the result, the average 15 features (35%; min 2%, max 53%) have appeared in each picture. This indicates that people did not use all of the description to draw the target. Moreover, some people tend to draw undescribed features: a tail (40%), whiskers (10%) and wings (4%).

The result suggests that description-to-depiction translation might involve not only bottom-up processing from appearance features of a target, but also top-down processing from an ad hoc creature image in mind.

# **The influence of biological cues on the patterns of categorization in non-mental retarded PDD**

**Hsiang-Chun Yeh**

National Cheng Kung University

**Jon-Fan Hu**

National Cheng Kung University

**Abstract:** Categorization plays an important role in the development of abstract thinking and helps subjects to cope with complex social interaction. However, in literature, there are a lot different claims about the categorization in pervasive developmental disorder (PDD). In present study, we aim to clarify whether, specifically, biological cues would influence non-mental retarded PDD in categorization. We used the short form of Wisconsin Card Sorting Test (WCST) pertaining with and without biological information to evaluate categorization ability in non-mental retarded PDD and typical developmental subjects. It was found that non-mental retarded PDD have more intact responses for physical categories than biological categories. These results allow us to highlight the patterns of deficits in prototype categorization, conceptualizing social function and accompanied specific category learning of PDD.

Keywords: categorization, biological cues, pervasive developmental disorder (PDD), WCST

# Determining people's expectations about the form of causal relationships

**Saiwing Yeung**

University of California, Berkeley

**Chris Lucas**

Carnegie Mellon University

**Tom Griffiths**

University of California, Berkeley

**Abstract:** How people make inferences about causal systems with multiple potentially interactive causes is an important but complicated question. We conducted an experiment using a blinket detector task with an iterated-learning design to study people's prior beliefs about functional form. Participants observed stimuli and generated predictions based on their observations. Some of these predictions were then observed by the next participants, ultimately converging on answers that reflect only people's prior beliefs. Our results suggest that people make judgments by balancing a preference for simple causal systems against one for adequately explaining the available data. To explain these results, we propose a novel computational model that features a grammar-based prior, expressing causal relationships as compositions of three atomic forms. This model outperformed one that is based on the noisy-OR and captured people's expectations about such causal systems, suggesting the importance of logical operations — disjunction, conjunction, and negation — in characterizing people's causal knowledge.



# **The role of linguistic inputs on bilingual language development**

**Michael C. W. Yip**

The Hong Kong Institute of Education

**Abstract:** The variable of linguistic inputs plays a critical role in language acquisition and language development. The present study examined this research question by conducting a detailed day-by-day language analysis on all the linguistic inputs perceived by a group of Chinese-English bilingual infants for three months. Through language recording and analysis, some estimates are obtained of the linguistic properties of the words as well as the language structures perceived by the infants. The preliminary findings provide a solid and realistic picture of the different categories and properties of the linguistic inputs spoken to and around the infants. Important implications from the present set of data for the study of bilingual language development will be discussed.

# Nothing-absence difference in causal induction and the pARIs rule

**Junki Yokokawa**

Tokyo Denki University

**Tatsuji Takahashi**

Tokyo Denki University

**Abstract:** In a novel environment, we establish causal relationship by inductive inference from statistical data. Although recent studies on causal induction focus on causal structure rather than intensity, how we induce the intensity from co-occurrence of an effect and a candidate cause remains important as far as the environment is of novelty. We propose a heuristic for causal induction, the proportion of assumed-to-be rare instances (pARIs), and test its rationale and the descriptive validity. The pARIs rule is based on the rarity assumption that is quite often used in rational analysis approach (e.g., Oaksford & Chater, 1994) intimately associated with the frame problem. As for the descriptive power, we show that pARIs best fits the experimental results with the highest correlation and the minimum error. In regard to the rationale, we tested how we assume the rarity of the events we focus on. We confirmed that we rigorously distinguish between rare and non-rare events, in spite of their identical status as data.

# Proficiency in foreign language reading: the relationship between proficiency test score and reading times

Satoru Yokoyama

Tohoku University

**Abstract:** The more proficient learners are at a foreign language, the faster they can process that foreign language; hence, processing speed is one of most important indices of foreign language proficiency. Yet there are few studies which examine the relationship between proficiency levels and processing speeds. In the present study, I looked at how the English proficiency test scores of 35 Japanese learners related to the speed at which the learners processed written English sentences. The results showed a statistically robust correlation between listening test scores and the processing speed of sentences which subjects read ( $p < 0.05$ ), but not between reading test scores and general processing speed ( $p = 0.5$ ). These results indicate that the reading speed of sentences in a foreign language is strongly related to listening proficiency, which proceeds to suggests that reading speed in a foreign language can be used as one index of listening proficiency.

# **Toward a history-sensitive description of syntactic development**

**Masato Yoshikawa**

Keio University

**Abstract:** Under the empiricist theory of language acquisition known as Usage-based Model it is assumed that the process of acquisition is gradual, from concrete to abstract; infants representation of grammatical knowledge is considered to be "acquired" in bottom-up fashion. This view is compatible with recent, that is, non-traditional, theories of machine learning such as connectionist models, and therefore it can be said to have a plausibility at least from a computational point of view. At the same time, however, descriptive plausibility should be achieved. In this study, an innovative method called History-Sensitive Description (HSD) is presented in order to describe the development in the usage-based fashion. This method utilizes the history of the utterances by one infant to describe the current knowledge state of the infant. HSD gives us a set of representation at any moment of development if a longitudinal data is given.

# **Pictogram Network to Support English Composition Instructors**

**Sayuri Yoshizawa-Watanabe**

Hoshi University

**Masaaki Kunigami**

Tokyo Institute of Technology

**Satoshi Takahashi**

Tokyo Institute of Technology

**Atsushi Yoshikawa**

The Japan Institute for Educational Measurement, Inc.

**Takao Terano**

Tokyo Institute of Technology

**Abstract:** Pictogram Network is a network representation with drawings, which illustrates information or knowledge that is difficult to transfer via usual conversation. Writing an essay with a complex context requires a high level of language skills, especially in foreign languages. This paper describes the basic principles of Pictogram Network and proposes a novel support method for peer review activities for English as a foreign language composition education. When applying the method to the composition learning, the learners are required to select pictograms from the provided list and link them in order to convey the main purport of their essays. The results will not be affected by inessential factors such as spelling errors and trifling grammatical or syntax errors. Therefore, ambiguities of the context about the composition topics will become clearer than in the conventional method. The proposed method also helps to indicate illogicality of the context in writings, and will be useful even in multi-lingual environment.

# Preschoolers Use Timing of Causal Actions as a Cue for Categorization

**Yue Yu**

Cornell University

**Tamar Kushnir**

Cornell University

**Abstract:** This study examined how childrens categorization is influenced by adults social cues, particularly whether adults performed causal actions before or after a sorting demonstration. In Experiment 1, preschoolers saw the experimenter sort toys with different surface features (colors) and causal properties (sound produced when shaken) into two boxes. When the experimenter shook the toy to produce a sound before sorting (shake-first condition), children were more likely to later categorize the toys based on sound. However, when the experimenter performed the same shaking action after sorting (shake-last condition), children were more likely to sort the toys by color. Experiment 2 further demonstrated that children in the shake-first condition continued to categorize a new set of toys by sound, and they did so even in free play. These results suggest that children use adults cues in particular, the timing of adults causal actions to determine whether causal properties are relevant for categorization.

# **The Rhetoric of Defamiliarization for Narrative Generation using the Constraints in a Conceptual Dictionary**

**Yike Zhang**

Graduate School of Software and Information Science, Iwate Prefectural University

**Junpei Ono**

Graduate School of Software and Information Science, Iwate Prefectural University

**Takashi Ogata**

Faculty of Software and Information Science, Iwate Prefectural University

**Abstract:** Through an analysis of advertising rhetoric, we have acquired an idea that much of the interestingness in advertising works are grounded on the rhetoric of defamiliarization for the components. The defamiliarization is originally a literary idea which was proposed by Shklovskii and Brecht and means a literary technique for changing a familiar object into unfamiliar one to reinforce the impression. For example, the impression of a familiar product is reinforced by the application of defamiliarization techniques to the objects and agents. This idea will be able to generalize to narratives other than the advertising narrative. In this paper, we summarize the techniques for operating the level of single event by the defamiliarization rhetoric combined with a conceptual dictionary for noun concepts and verb concepts. The techniques have twelve types including three types of regular rhetoric and other irregular rhetoric. The latter transforms the elements like action, actor, product (object) and location in an event into strange or preposterous things. In our integrated narrative generation system in which this module is incorporated, this function is associated with a mechanism to be able to flexibly adjust a variety of generation from realistic narratives to fantastical narratives.

## Author Index

Joshua Abbott .....	54, 60	Derrik Asher .....	90
Ahmed M. H. Abdel-Fattah .....	1242	Hiroshi Ashida .....	2714, 2740, 2786, 2886
Keiga Abe .....	2621	Jun Ashikaga .....	2630
Yosuke Abe .....	2730	Richard Aslin .....	881, 2907
Adele Abrahamsen .....	102	Natsuki Atagi .....	2631
Cengiz Acarturk .....	66	Sharona Atkins .....	1644
Margareta Ackerman .....	1870	Terry Kit-fong Au .....	2463
Edoardo Acotto .....	1248	Luc Augier .....	2632
Yuko Adachi .....	2622	Joseph Austerweil .....	54, 402
Deanne Adams .....	1254, 1260	Marios Avraamides .....	2091
Henny Admoni .....	1266	Ankit Awasthi .....	1296
Frans Adriaans .....	72	Shunji Awazu .....	2633
Kisuh Ahn .....	2623, 2757	Leila Azari Pishkenari .....	2634
Woo-Young Ahn .....	78	Sungbong Bae .....	2635
Mike Aitken .....	1185	Jaroslava Bagdasarova .....	2705
Hiroyuki Akama .....	2624	Wilma Bainbridge .....	1302
Taisuke Akimoto ...	1272, 2126, 2150, 2670, 2738, 2817	Chris Baker .....	515, 923
Kimi Akita .....	2846		
Ozge Alacam .....	66	Dare Baldwin .....	14
Ben Allison .....	1278	Jerry Ball .....	1308
Masahiro Amagase .....	2684	Bruno Bara .....	2787, 2843
Haruka Amatani .....	2625	Elizabeth Baraff Bonawitz .....	1614
Tobias Andersen .....	1284	David Barner .....	210, 1096
David Anderson .....	1013	Trevor Barrett .....	2869
John Anderson .....	767	Lawrence Barsalou .....	2852
Mark Andrews .....	16	Ziba Bashardanesh .....	2636
Patrice Andrieu .....	2746	Jana Basnakova .....	2682
Yuichiro Anzai .....	30	Peter Battaglia .....	32
Manabu Arai .....	791	Andrew Battles .....	1404
Yoshiko Arai .....	2626, 2800	Frank Baughman .....	1314
Kenji Araki .....	2845	Natalie Baughman .....	1314
Eiji Aramaki .....	2627, 2913	Philip Beaman .....	96
Joanne Arciuli .....	1013	William Bechtel .....	38, 102
Ana Arruarte .....	2615	Nicole Beckage .....	108
Burcu Arslan .....	1290	Marina Bedny .....	2637, 2736
Florian Artinger .....	84	Chuluundorj Begz .....	2638
Tomoko Asai .....	2109	Josef Behr .....	2423
Shin-ichi Asakawa .....	2628	Sieghard Beller .....	114
Hajime Asama .....	2768	Andrea Bender .....	114
Akihiro Asano .....	2629	Erin Bennett .....	354
		Patrick Bennett .....	2791



## Author Index

Benjamin Bergen .....	959	Sarah Brem .....	150
Leon Bergen .....	120, 1320	Sophie Bridgers .....	156
Kirsten Bergmann .....	1326	M. Anne Britt .....	965
Eduardo Bermudez .....	2639	Rainer Bromme .....	965
Bennett Bertenthal .....	2288	Patricia Brooks .....	1774
Vincent Berthiaume .....	402	Joshua Brown .....	78
Tarek Besold .....	1242, 1332	Meredith Brown .....	647, 1374, 1380
Catherine Best .....	2240	Scott Brown .....	2895
Krishna Bharani .....	1030	Ty Brumback .....	539
Klinton Bicknell .....	126	Emile Bruneau .....	2641
Ricardo Bion .....	1197	Peter Bruza .....	18, 1792
Elizabeth Ligon Bjork .....	683	Adam Bryant .....	1386
John Black .....	551, 1756, 1888, 2765	Daphna Buchsbaum .....	156
Mary-Jane Blais .....	1338	Cristina Burani .....	2369
Idan Blank .....	1302	Joseph Burling .....	2642
Agnes Blaye .....	384	Heather Burte .....	162
Paulo Blikstein .....	132	Jerome Busemeyer .....	18, 78, 1054, 2895
Svetoslav Bliznashki .....	1344	Carter Butts .....	108
Amber Bloomfield .....	1350	Michael Byrd .....	1392
Charles Blundell .....	1356	Benjamin Börschinger .....	2002
Franziska Bocklisch .....	905, 2640	Richard Caballero .....	1786
Rens Bogers .....	755	Zhenguang Cai .....	168, 252
Elizabeth Bonawitz .....	2180	Zhiqiang Cai .....	695
Anna Bordilovskaya .....	1362	Erik Cambria .....	174
Milena Borisova .....	408	Valérie Camos .....	2746
Lera Boroditsky .....	821, 2451	Fabian Canas .....	324
Francesca Bosco .....	2787, 2843	Angelo Cangelosi .....	2868
Peter Bossaerts .....	36	Jessica Cantlon .....	2903
Magali Boureux .....	2369	Susan Carey .....	2866
Tommy Bouwens .....	755	Peter Carnevale .....	270, 1476
Casady Bowman .....	1392	Manuel Carreiras .....	2666
Cem Bozşahin .....	1530	Emily Carrigan .....	1398
Janina Braatz .....	497	Christopher Carroll .....	180
Nick Braisby .....	2246	Alexandra Carstensen .....	827
David Braithwaite .....	138	Paulo Carvalho .....	186, 1662
Stefan Brandenburg .....	1506	Daniel Casasanto .....	306
Holly Branigan .....	228	Marianella Casasola .....	2835
Gary Brase .....	2800	Richard Catrambone .....	44
Mike Braverman .....	1368	Hee-Rahk Chae .....	2643, 2691, 2757
Micah Bregman .....	144	J. T. Y. Chan .....	2836

## Author Index

Ricky Chan .....	192	Angela Ciaramidaro .....	2787, 2843
Sally Wai Chi Chan .....	2804	Mina Cikara .....	2641, 2866
Franklin Chang .....	40	Benjamin Cipollini .....	1410
Li-Yun Chang .....	2662	Ciro Civile .....	1416, 1422
Ya-Ning Chang .....	198	Dav Clark .....	2228
Nick Chater .....	42, 48, 2842	John Clevenger .....	1368
Fudan Chen .....	1173	Moreno Coco .....	228, 1278
Hongbo Chen .....	2649	Sarah Cohen .....	2228
Hsueh-Chih Chen ...	2650, 2662, 2695, 2834	William Cohen .....	731
Hung-Hui Chen .....	2756	Ariel Cohen-Goldberg .....	52
I-Chen Chen .....	2646	Neil Cohn .....	52, 234, 240
Jenn-Yeu Chen .	204, 366, 2644, 2645, 2651, 2693	Lucia Colombo .....	2369
Lang Chen .....	2647		
Rongjuan Chen .....	2648, 2753	Eliana Colunga .....	246, 2294
Sau-chin Chen .....	2644	Louise Connell ...	168, 252, 258, 1428, 1948
Stephanie Jui-chi Chen .....	2659	Michael Connor .....	40
Train-Min Chen .....	204, 2645	Richard Cooper .....	38
Fan-Ning Cheng .....	2650, 2834	Emily Coppess .....	629
Hsiang-Chun Cheng .....	2651	Marie Coppola .....	1398
Rong-Ju Cherng .....	2651	Daniel Corral .....	1434
Pierina Cheung .....	210, 2652	Garrison Cottrell ...	1048, 1410, 1894, 2564
Itsuki Chiba .....	2653, 2867	Denis Cousineau .....	1452
Syota Chimura .....	2898	Gregory Cox .....	264, 1440, 2663, 2713
Jessie Chin .....	1404	Sarah Creel .....	144, 2174
Clark Chinn .....	2654	Matthew Crocker .....	1007
Suvarna Chinta .....	2655	Vincenzo Crupi .....	1233
Eric Chiu .....	216	Shannon Cuykendall .....	1786
Ping-Hui Chiu .....	2659	Nils Dahlbäck .....	2339
Hye Rhang Cho .....	2657, 2658, 2803	David Dahmen .....	2639
Kit Cho .....	2656	Meghan Dale .....	2664
Sook Whan Cho .....	2657, 2658, 2803	Rick Dale .....	1518
Feng-Xuan Choo .....	1018	Somayeh Danafar .....	1446
Hui-Mei Chow .....	2659, 2766	Frederic Dandurand .....	1452, 2327
Sheldon Chow .....	222	Dieter Daniëls .....	2310
Stella Christie .....	34	David Danks .....	2523, 2529
Kawai Chui .....	2660	Vivek Datla .....	491, 2434
Chongwook Chung .....	2661	Lila Davachi .....	2770
Fu-lai Chung .....	2609	Eddy Davelaar .....	426, 1458
Yi-Ling Chung .....	2662	Nicolas Davidenko .....	2665
Dorothee Chwilla .....	1596	Doug Davidson .....	2666
		Nathaniel Daw .....	36

## Author Index

Peter Dayan .....	42, 50	Nicholas Duran .....	1518
Simon de Deyne .....	1464	Frank Durgin .....	46
Celso de Melo .....	270	Kelley Durkin .....	1260
Virginia de Sa .....	26, 671, 1870, 1876	Pawel Dybala .....	1238, 2669
Dan Dediu .....	2682	Melody Dye .....	2663
Ben Deen .....	276	Kazutoshi Ebe .....	1762
Rebecca Defina .....	1470	Kathleen Marie Eberhard .....	2863
Chizuru Deguchi .....	2369	Miguel Eckstein .....	2716
Morteza Dehghani .....	563, 1476, 1482	Shimon Edelman .....	372
M. Dolores del Castillo .....	1715	Cole Edison .....	2351
Gary Dell .....	40	Brian Edwards .....	318
Sophia Deng .....	282	Yuka Egusa .....	953
Ying Deng .....	1488	Martin Eimer .....	2907
Simon Dennis .....	2587	Hamed Ekhtiari .....	2634
Jean-Louis Dessalles .....	947, 2055	Heike Elchlepp .....	1416, 1422, 1548
Ben Deverett .....	288	Marco Elena .....	2787
Andrew Dewald .....	294, 1494	Chris Eliasmith .....	22, 38, 1018
Jesse Diaz .....	16	Jon Elorriaga .....	2615
Zoltan Dienes .....	20	Ikuo Endo .....	2772
Emmanuelle-Anna Dietz .....	1500	Jun Endo .....	2670
Laura Dilley .....	1374, 1380	Jeremy Engle .....	1524
Mark Dingemanse .....	300	Kerem Eryilmaz .....	1530
Kyung Soo Do .....	2667, 2668	Arash Eshghi .....	479
Karen Dobkins .....	1096	Zhou Fang .....	2682
Sarah Dolscheid .....	306	Chris Fargen .....	977
Devin Domingo .....	2781	Caitlin Fausey .....	1668
Emanuel Donchin .....	539	Afsaneh Fazly .....	2085
Eric Doty .....	2327	Aidan Feeney .....	2351, 2815
Michael Dougherty .....	426, 1458, 1644	Laurie Beth Feldman .....	10, 2656
Leonidas Doumas .....	34, 677, 1936	Naomi Feldman .....	354, 360, 629
Steven Dow .....	635	Ying Feng .....	1524
Kenji Doya .....	36	Anne Fernald .....	1197
Uwe Drewitz .....	1506	Klaus Fiedler .....	1090
Rachel Dryer .....	2511	Alex Fine .....	599
Sarah Dubrow .....	2770	Sara Finley .....	1536
Michel DuCharme .....	2743	Hadi Firouzi .....	1578
Sam Duffy .....	1512	Anna Fisher .....	1608, 2678, 2773
Nicholas Dufour .....	312	Cynthia Fisher .....	40
Steve Duman .....	857, 2841	Branden Fitelson .....	1233
James Dungan .....	623	Nadine Fleischhut .....	84

## Author Index

Caroline Floccia .....	2746	Mengzi Gao .....	1888
Nick Flor .....	1542	Tao Gao .....	14
Stephen Flusberg .....	2451	Yue Gao .....	372
Kenneth Forbus .....	32, 40, 545, 701	Konstantina Garoufi .....	1007
Glen Forester .....	539	Maria Garraffa .....	228
Charlotte Forrest .....	1548	Albert Gatt .....	1584
James Foster .....	324	Dedre Gentner .....	34, 40, 2708
Tom Foulsham .....	330, 2671	Georgi Georgiev .....	2792
Charles Fox .....	336	Kallirroi Georgila .....	2097
Julie Franck .....	2826	Tobias Gerstenberg .....	32, 378, 1590, 1996
Michael Frank .....	342, 473, 821, 935, 989	Ray Gibbs .....	773, 2037
Mike Frank .....	2002	Edward Gibson .....	1320
Stefan Frank .....	1554	Barry Giesbrecht .....	2716
Diego Frassinelli .....	1560, 1566	Gerd Gigerenzer .....	48
Bob French .....	384	Rosa Gisladdottir .....	1596
Seth Frey .....	1572, 2663	Yannick Glady .....	384
Scott Friedman .....	40	Vladimir Glebkin .....	1602
Chris Frith .....	2842	Kevin Gluck .....	1078, 2677
Wai-Tat Fu .....	1404	Karrie Godwin .....	1608, 2678, 2773
Chie Fukada .....	2823, 2856	Ashok Goel .....	1828
Haruaki Fukuda .....	2672	Winston Goh .....	983, 2679
Itaru Fukuda .....	2126	Kazutoshi Gohara .....	2910
Miwa Fukushima .....	2704	Stephen Goldberg .....	52
Miwa Fukushima-Murata .....	2109	Aleah Goldin .....	1840
Michelle Fullwood .....	2870	Robert Goldstone .....	138, 186, 1524, 1662, 1668, 2156, 2399
Shintaro Funahashi .....	2793	Micah Goldwater .....	40, 2708
Kotaro Funakoshi .....	1816	Laura Gonnerman .....	1338
Sou Funakoshi .....	2738	Aaron Gonzalez .....	1614
Steve Furber .....	198	Cleotilde Gonzalez .....	521
Nobuhiro Furuyama .....	1810	Henry Gonzalez .....	2639
Riccardo Fusaroli .....	2673, 2899	Noah Goodman .....	24, 120, 378, 390, 665, 1590
William Fuss .....	2333	Alison Gopnik .....	156, 1108, 1114, 1614, 2180
Emily Fyfe .....	348	Kristina Gotseva .....	2061
Liane Gabora .....	1578, 2234, 2487, 2674	Martijn Goudbeek .....	1066, 1084
John Gabrieli .....	312, 911	Kevin Gould .....	2841
Annie Gagliardi .....	354, 360	Scott Grafton .....	2710
Francesco Gagliardi .....	2675	Jonathan Grainger .....	1638
Jean-François Gagnon .....	2676, 2743	Linn Gralla .....	396
Bruno Galmar .....	366	Jonathan Gratch .....	270, 563, 1476, 1482
Patricia Ganea .....	1108	Wayne Gray .....	899, 2481, 2761

## Author Index

Valerie Gray Hardcastle .....	38	Hiroshi Hashizume .....	2767
Thomas Griffiths .....	402, 893, 1918	Yuichi Hasimoto .....	2867
Tom Griffiths .....	38, 54, 60, 156, 485, 833, 1356, 2779, 2915	Mohammad Javad Hatami .....	2634
Kalanit Grill-Spector .....	420	Chiriho Hatanaka .....	2794
Michael Grimaila .....	1386	Aya Hatano .....	2684
Maurice Grinberg .....	408	Ikuko Hattori .....	2685
Marc Grosjean .....	2747	Masasi Hattori .....	2685, 2827
Suzanne Grossman .....	240	Yosuke Hattori .....	2698
Ernesto Guerra .....	1620	Robert Hausmann .....	438
Markus Guhe .....	1626	Hiromasa Hayashi .....	2686
Glenn Gunzelmann .....	414	Katsuyuki Hayashi .....	1798
Todd Gureckis ...	719, 725, 749, 1745, 2770, 2781	Yugo Hayashi .....	444, 1650
Johannes Gurlitt .....	653	Brett Hayes .....	1966
Helmar Gust .....	1242	Patrick Heady .....	2705
Hyowon Gweon .....	977, 2680	Patrick G. T. Healey .....	479, 1512, 1697
Kassandra Gynther .....	1350	Andrew Heathcote .....	2895
Jung-Woo Ha .....	1221	Benjamin Heddy .....	150
Anne Haake .....	1900	Mary Hegarty .....	162, 1240, 2162, 2869
Yoshiko Habuchi .....	1632	Jeffrey Heinz .....	2744
Yuki Hachikura .....	2856	Evan Heit .....	1656
Martin Hackl .....	2870	Christian Hempelmann .....	450, 2393
Shingo Hagiwara .....	2774, 2798	Andrew Hendrickson .....	1662, 1668
Peter Hagoort .....	467	Maria Henriksson .....	1674
Amanda Hahn .....	1852	Jonathan Herberg .....	617
Lance Hahn .....	2681	Ralph Hertwig .....	1167
Justin Halberda .....	1126	Tom Heyman .....	456
Gregory Hallman Jr. ....	1756	Jeremiah Hiam .....	2603
Syoji Hamaguchi .....	2898	Shohei Hidaka .....	1203, 1679
Rubi Hammer .....	420	Ryuichiro Higashinaka .....	2417
Simon Handley .....	2049	Andrew Higgins .....	1368
Thomas Hannagan .....	1638	Takahiro Higuchi .....	2719
Adriana Hanulikova .....	2666, 2682	Yoko Higuchi .....	2888
Etsuko Harada .....	2683, 2784	Thomas Hills .....	1167
Taeko Harada .....	2882	Isabella Hinterleitner .....	2687
J. Isaiah Harbison .....	426, 1458, 1644	Kazuo Hiraki .....	2686, 2776, 2883
Kerstin Sophie Haring .....	432	Tomoki Hirano .....	2783
Ian Harmon .....	1368	Naoshi Hiraoka .....	2793
Nigel Harvey .....	50, 1001	Sachiko Hirata .....	2688
Etsuko Haryu .....	34, 2121	Naoya Hirose .....	2689
Takashi Hashimoto ...	779, 2735, 2754, 2884	Masako Hirotani .....	2690

## Author Index

Dale Hirsch .....	2558	Shohei Imabuchi .....	2699
Anna Wing Yee Ho .....	2766, 2836	Mutsumi Imai .....	34, 2820, 2846
Seng Beng Ho .....	1685	Sho Inaba .....	2630
Marieke Hoetjes .....	461	Benjamin Inden .....	1721
Annette Hohenberger .....	1290	Kenryo Indo .....	2700
Phillip Holcomb .....	234	Bipin Indurkha .....	1727, 2655
Judith Holler .....	252, 467	Masakatsu Inoue .....	2739
Keith Holyoak .....	2144	Miyako Inoue .....	2802
Hidehito Honda .....	1691	Keigo Inukai .....	2906
Do-Il Hong .....	2643, 2691	Miwa Inuzuka .....	2701
Johan Hoorn .....	2198	Ryo Ishibashi .....	2802
Zachary Horne .....	1368	Akira Ishiguchi .....	2440
Alexandra Horowitz .....	473	Chiaki Ishiguro .....	1733
Christine Howes .....	479, 1697	Hiroshi Ishiguro ...	12, 30, 2375, 2469, 2776
Evgenia Hristova .....	408	Tatsunori Ishii .....	2702
Penka Hristova .....	1703	Satoru Ishikawa .....	2703
Janet Hui-wen Hsiao .....	2463, 2540	Shun Ishizaki .....	2417
Janet Hsiao .....	689, 1410	Phillip Isola .....	1302
Yaling Hsiao .....	2692	Tomoko Isomura .....	2704
Anne Hsu .....	485	Seiji Isotani .....	1260
Li-Chuan Hsu .....	2499	Kirill Istomin .....	2705
Jon-Fan Hu ..	2646, 2650, 2662, 2695, 2834, 2914	Hiroyasu Ito .....	2704, 2706
Jie Huang .....	2694	Hiroyuki Ito .....	2306
Lixing Huang .....	1482	Sadanori Ito .....	2872
Marin Huang .....	2646	Takane Ito .....	611
Shuping Huang .....	2693	Takeshi Ito .....	1739
Falk Huettig .....	2682	Tomoko Itoh .....	2707
John Hummel .....	2712	Mamoru Iwabuchi .....	2875
Yu-Sheng Hung .....	2695	Kunihiro Iwamoto .....	1762
Sabine Hunnius .....	306	Shoichiro Iwasawa .....	2872
Erika Hussey .....	426, 1458	Kenta Iyoki .....	2630
Sterling Hutchinson ...	491, 695, 1709, 2434	Ray Jackendoff .....	234
Yiwon Hyun .....	2696, 2748	T. Florian Jaeger .....	599, 605
Steffen Hölldobler .....	1500	Georg Jahn .....	497, 905
Jun Ichikawa .....	2697	Adel Jalabi .....	2192
Tohru Ifukube .....	2818	Azadeh Jamalian .....	503
Angel Iglesias .....	1715	Anja Jamrozik .....	2708
Keiko Ihaya .....	2306	Dayk Jang .....	2750
Tetsuya Iidaka .....	1762	Sooyong Jang .....	2722
Kenji Ikeda .....	2698	Leen Janssens .....	509, 2310

## Author Index

Julian Jara-Ettinger .....	515	Masayoshi Kanoh .....	2709
Kate Jeffery .....	46	Himanshu Kansal .....	1948
Carol Jew .....	749	Katerina Kantartzis .....	2846
Cui Jian .....	1972	Nana Kanzaki .....	557
Felix Jimenez .....	2709	Bilge Karacora .....	563
Koji Jimura .....	2405, 2741	Ryan Kasper .....	2716
Li Jingling .....	2896	Yasuhiro Katagiri .....	2717
Coley John .....	2351	Kentaro Katahira .....	2737
Arianne Johnson .....	2710	Motoichiro Kato .....	2768
Mark Johnson .....	2002	Yoichi Kato .....	1179
Philip Johnson-Laird .....	575	Maya Katsuhara .....	2718, 2886
William Johnston .....	2637	Ryan Kaulakis .....	2603
Fergal Jones .....	1185, 2581	Yasuhiro Kawabata .....	2877
Matt Jones .....	324, 1434	Jun Kawaguchi .....	2684, 2800, 2808
Mijung Joo .....	2711	Shigeto Kawahara .....	569
Ashikaga Jun .....	2881	Nobuyuki Kawai .....	2737
Wookyoung Jung .....	2712	Satoru Kawamura .....	1822
Mordechai Juni .....	1745	Naoko Kawano .....	1762
Peter Juslin .....	2760	Tsubasa Kawasaki .....	2719
Ion Juvina .....	521	Akihiro Kawase .....	2720, 2894
George Kachergis .....	527, 533, 1440, 1668, 2713	Ryuta Kawashima .....	2767, 2878, 2912
Koji Kadota .....	2910	Mark Keane .....	2252
Masayo Kajimura .....	2714	Victoria Keiser .....	731
Shogo Kajimura .....	2802	Frank Keller .....	1278, 1560
Yasuaki Kakehi .....	2775	Matthew Kelly .....	1768
Mami Kamada .....	2126	Spencer Kelly .....	467
Toshimune Kambara .....	2878	Charles Kemp .....	180, 288, 707, 713
Hiroko Kamide .....	1822	Vera Kempe .....	1774
Olena Kaminska .....	2671	Iyoki Kenta .....	2881
Jennifer Kaminski .....	1750	Sangeet Khemlani .....	575, 581, 1780
Takumi Kamiya .....	2876	Peter Khooshabeh .....	1482
Tosiro Kamiya .....	2872	Kaede Kido .....	2721
Siri-Maria Kamp .....	539	Akifumi Kijima .....	2910
Akihito Kanai .....	2795, 2819	Bia Kim .....	2725, 2748
Subu Kandaswamy .....	545	Chobok Kim .....	2661
Noriko Kando .....	953	DaeEun Kim .....	2727
Yasuo Kaneko .....	2715	Eun-Sol Kim .....	2722, 2723
Hyunmin Kang .....	2711	Eun sook Kim .....	2728
Myunggu Kang .....	1221	Ha Rim Kim .....	2844
Seokmin Kang .....	551, 1756, 2765	Hanna Kim .....	2668

## Author Index

Hee-Yeon Kim .....	2667	Alexander Koller .....	1007
Jeehoon Kim .....	2729	Takanori Komatsu .....	1816
Jeounghoon Kim .....	2661	Masashi Komori .....	1822, 2794
Jiseob Kim .....	2722, 2723	Piotr Konderak .....	2734
Jong kim .....	28	Lars Konieczny .....	432
Joon Shik Kim .....	1864, 2726	Takeshi Konno .....	779, 2735
Kyung Kim .....	2724	Nikos Konstantinou .....	2842
Kyungil Kim .....	2837	Stefan Kopp .....	1326
Young Kim .....	2724	Lily Kornbluth .....	1227
Ken-ichi Kimura .....	2858	Nate Kornell .....	2903
Alan Kingstone .....	2758	Jorie Koster-Hale ....	623, 2637, 2736, 2840
Natasha Kirkham .....	1161	Hadas Kotek .....	2870
Christo Kirov .....	587	Tsukasa Koyama .....	2019
David Kirsh .....	593, 1786, 2339	Masuo Koyasu .....	2025, 2769
Akihito Kishino .....	2735	Maria Kozhevnikov .....	2103
Shinichi Kita .....	2688	Emiel Krahmer .....	461, 1066, 1084, 1584, 2493
Sotaro Kita .....	761, 2846	Josef Krems .....	905, 2640, 2855
Shinji Kitagami .....	2698	Jeffrey Krichmar .....	90
Kirsty Kitto .....	1792, 2674	Judith Kroll .....	10
Sachiko Kiyokawa .....	1798	Yakov Kronrod .....	629
Eve Klama .....	1804	Ulf Krumnack .....	1242
Dave Kleinschmidt .....	599, 605	Nicole Krämer-Mertens .....	563
Scott Klemmer .....	635	Kenta Kubo .....	2737
Pia Knoeferle .....	1227, 1620, 2593	Takuya Kubo .....	2847
Harumi Kobayashi .....	2576	Namiko Kubo-Kawai .....	2109
Hiroko Kobayashi .....	2701	Chinmay Kulkarni .....	635
Kazuki Kobayashi .....	1816	Takatsune Kumada .....	2818
Tessei Kobayashi .....	2826	Shinya Kumagai .....	2738
Yuki Kobayashi .....	611	Neeraj Kumar .....	641
Satoshi Kobori .....	2730	Maithilee Kunda .....	1828
Kentaro Kodama .....	1810	Masaaki Kunigami .....	2901, 2920
Tatsuya Koeda .....	2887	Gina Kuperberg .....	234
Kenneth Koedinger .....	731	Takeo Kurafuji .....	2739
Bryan Koenig .....	617, 617	Kana Kuraguchi .....	2740, 2886
Jean-Pierre Koenig .....	1191	Kaori Kuraya .....	2802
Yu Kohno .....	2731	Satoshi Kuribayashi .....	2775
Masatoshi Koizumi .....	2847	Yasunari Kurisawa .....	2126
Kazuaki Kojima .....	1960, 2732, 2892	Naoko Kuriyama .....	2405, 2741
Takatsugu Kojima .....	2733, 2856	Kenneth Kurtz .....	2752
Boicho Kokinov .....	1344, 2061	Chigusa Kurumada .....	647, 2002



## Author Index

Tamar Kushnir .....	2921	Esko Lehtonen .....	1846
Ichiro Kusumi .....	2019	Miao Mei Lei .....	2624
Takashi Kusumi ....	2363, 2381, 2405, 2741, 2769, 2777	David Pierre Leibovitz .....	2535
Roman Kutlak .....	1834	Alessandro Lenci .....	1215, 1566
Tomoko Kuwabara .....	2794	Jessica Lesky .....	2300
Kenneth Kwok .....	174	Janny Leung .....	192
Søren Kyllingsbæk .....	1284	Vittoria Levati .....	84
Ikuko Kyoya .....	2742	Kimery Levering .....	2752
Kai-Uwe Kühnberger .....	1242, 1332	Stephen Levinson .....	1596
Andreas Lachner .....	653	Roger Levy .....	120, 126, 1320, 2174
Bruno Laeng .....	2871	Stephan Lewandowsky .....	1918
Daniel Lafond .....	2743	Joshua Lewis .....	26, 671, 1870, 1876
David Lagnado .....	32, 378	Daniel Leyzberg .....	1882
Regine Lai .....	2744	Guanhong Li .....	2754
Tei Laine .....	2745	Huaye Li .....	2753
Brenden Lake .....	659	Na Li .....	1888, 2624
David Landy .....	1840, 2156, 2300, 2873	Peggy Li .....	210
Otto Lappi .....	941, 1846	Rentao Li .....	1894
Frankie Lara .....	1852	Rosa Li .....	2903
Alex Lascarides .....	1626	Rui Li .....	1900
Lyuben Laskin .....	1858	Simon Y. W. Li .....	2755
Daniel Lassiter .....	665	Yun Li .....	2695
Romy Lassotta .....	2826	Jeffrey Lidz .....	354, 360, 1126
Lucie Laurent .....	2746	Chi-Shun Lien .....	2756
Nilli Lavie .....	2842	Ahnate Lim .....	677, 1906, 1912, 2758
Aureliu Lavric .....	1416, 1422	Eunsuk Lim .....	2623, 2757
Nathalie Le Bigot .....	2747	Joo-Hwee Lim .....	2505
Mathieu Le Corre .....	2192, 2652, 2664	Stephen Wee Hun Lim .....	2570, 2759
Christian Lebiere .....	521, 2168	Marcus Lindskog .....	2760
Beom-Jin Lee .....	1864	John Lindstedt .....	2761
Chung-Yeon Lee .....	1864, 2726	Daniel Little .....	1918
Chungmin Lee .....	2751	Jeri Little .....	683
Donghoon Lee .....	2696, 2725	Daniel Hsi-wen Liu .....	1930
Goeun Lee .....	2725	Tao Liu .....	1924
Jaesik Lee .....	2711, 2748	Tianyin Liu .....	689
Jonathan Jiseop Lee .....	2749	Wan-Yi Liu .....	2762
Michael Lee .....	90	Yan Liu .....	2609
Min-Seop Lee .....	2750	Ken Livingston .....	2905
Sang Bok Lee .....	2749	Katherine Livins .....	1936
Yoonkyoung Lee .....	2725, 2748	Marit Lobben .....	2763

## Author Index

John Logan .....	971	Christine Massey .....	917
Tania Lombrozo .....	1114, 1149, 2833	Vivien Mast .....	1972
Shengyan Long .....	2764	Megan Masters .....	1350
Deryle Lonsdale .....	2132	Sayako Masuda .....	2633
Max Louwerse .....	491, 695, 1709, 2434	Hiroyuki Masukawa .....	2772
Brad Love .....	38	Fabien Mathy .....	2746
Andrew Lovett .....	701	Bryan Matlen .....	1608, 2773
Carol Lu .....	2765	Ryuichi Matoba .....	2774
Shijian Lu .....	2505	Masaki Matsubara .....	2775
Christopher Lucas .....	707, 713, 2915	Shota Matsubayashi .....	2411
Pei-Yu Luo .....	2662	Goh Matsuda .....	2686, 2776
X. Luo .....	2836	Ken Matsuda .....	2777
Erkki Luuk .....	1942	Noboru Matsuda .....	731
Hendrik Luuk .....	1942	Michinao Matsui .....	2739, 2778
Dermot Lynott .....	258, 1428, 1948	Takao Matsui .....	2031
Yuen Ma .....	2766	Tatsunori Matsui .....	2732
Maryellen MacDonald .....	2692	Tomoko Matsui .....	797, 2820
Brian MacWhinney .....	773, 2037	Toshihiko Matsuka .....	1691, 1798, 2387
Naoki Maeda .....	1954	Mariko Matsumoto .....	1650
Takaki Maeda .....	2768	Takao Matsumoto .....	1978
Akihiro Maehigashi .....	1960, 2892	Takuya Matsumoto .....	2270
Annie Magnan .....	2767	Miki Matsumuro .....	1984
James Magnusson .....	1638	Yoshiyuki Matsuzawa .....	2816
Asifa Majid .....	300, 306, 1155, 1470	Gail Mauner .....	1191
Ryosaku Makino .....	1810	Luke Maurits .....	2779
Shogo Makioka .....	2721	Penelope Mavros .....	312, 911
Kai Makita .....	797	Rich Mayer .....	1254, 1260
Zofia Malisz .....	1721	Julien Mayor .....	737, 1990, 2780
Jonathan Malmaud .....	42	Ralf Mayrhofer .....	743
Laurence Maloney .....	42, 1745	Norbert Maïonchi-Pino .....	2767
Emmanuel Manalo .....	2769	Teresa McCormack .....	2815
Jaison Manjaly .....	641	John McCoy .....	1996
Doug Markant .....	719, 725, 2770	John McDonnell .....	749, 2781
Art Markman .....	1179	Keith McGregor .....	1828
Davide Marocco .....	2868	Bruce McLaren .....	1260
Liron Marotz .....	2771	Ian P. L. McLaren .	1185, 1416, 1422, 1548, 2581
Ann Martin .....	1966	Rossy McLaren .....	1416, 1422
Jay Martin .....	485	Björn Meder .....	48
Aleix Martinez .....	2240	Katja Mehlhorn .....	2855
Nobuo Masataka .....	2109, 2704, 2706	Chris Mellish .....	1834

## Author Index

Stephan Meylan .....	2002	Aiko Morita .....	2789
Yasushi Michita .....	2769	Hiromi Morita .....	2891
Jean-Louis Millot .....	2746	Hitoshi Morita .....	2790
Gregory Mills .....	2665, 2782	Junya Morita .....	779, 1960, 2735, 2892
Robert Mills .....	1386	Tomomi Moroga .....	2805
Simon Mills .....	1314	Robert Morrison .....	34, 1030
Fraser Milton .....	1804	Daniel Morrow .....	1404
Masaru Mimura .....	2768	Kinga Morsanyi .....	2049
Kayo Miura .....	2853	Shane Mueller .....	2893
Satoshi Miura .....	2627	Amitabha Mukerjee .....	1296, 2079
Kazuhisa Miwa 557, 1960, 1984, 2008, 2411, 2732, 2885, 2892		Damien Munch .....	2055
Makiko Miwa .....	953	Masaki Murata .....	2627
Mai Miyabe .....	2627	Anthony Murphy .....	539
Takayuki Miyadera .....	2014	Brian Murphy .....	2624
Maki Miyajima .....	2019	Milena Mutaftchieva .....	2061
Masaki Miyake .....	2783	Christopher Myers .....	1078, 1308
Naomi Miyake .... 30, 44, 2258, 2630, 2783, 2881		James Myers .....	52, 2067
Takashi Miyake .....	2115	Masayoshi Nagai .....	2791
Koh Miyamoto .....	2807	Yukari Nagai .....	2792, 2909
Krishna Miyapuram .....	641	Chikako Nagaoka .....	2793
Kozue Miyashiro .....	2784	Chika Nagaoka .....	1822, 2794
Makoto Miyatani .....	2829	Jonas Nagel .....	785
Shiho Miyazawa .....	2785	Yuya Naito .....	2795
Kiyohumi Miyoshi .....	2786	Etsuko Nakagami-Yamaguchi .....	2622
Ai Mizokawa .....	2025	Masanori Nakagawa .....	2405, 2741
Rika Mizuno .....	2031	Ryuichi Nakaike .....	2008
Kaname Mochizuki .....	2821	Kumiyo Nakakoji .....	2796
Shafiz Affendi Mohd Yusof .....	2800	Keiko Nakamoto .....	2797
Ikuko Mohri .....	2851	Chie Nakamura .....	791
Lisette Mol .....	755, 761	Hiroko Nakamura .....	2800
Brian Monroe .....	617, 617	Jun Nakamura .....	2799
Stephen Monsell .....	1548	Kenryu Nakamura .....	2875
Jungaa Moon .....	767	Kota Nakamura .....	2475
Adam Moore .....	1780	Kuninori Nakamura .....	803
L. Richard Moore .....	414	Makoto Nakamura .....	2774, 2798
Joseph Moran .....	312	Sari Nakamura .....	2824
Rosalba Morese .....	2787, 2843	Satoshi Nakamura .....	2805
Laura Morett .....	773, 2037, 2043	Tagiru Nakamura .....	797
Jonathan Morgan .....	2603	Yoshifumi Nakanishi .....	2801
Miwa Morishita .....	2788	Mikio Nakano .....	1816

## Author Index

Yuko Nakano .....	2073	Matthias Nückles .....	653
Momoko Nakatani .....	1179	Kerry O'Brien .....	1948
Tatsuya Nakatani .....	2622	Stephen O'Connell .....	1350
Masataka Nakayama .....	2802	Eimear O'Connor .....	2815
Takahiro Nakayama .....	2630	John O'Doherty .....	36
Shogo Nakazono .....	2730	Brian O'Donnell .....	78
Jun-Hee Nam .....	2726	Daniel O'Young .....	911
Seung Suk Nam .....	2657, 2658, 2803	Takashi Obana .....	2103
Daniele Nardi .....	46	Hiromi Ochiai .....	2809
Daniel Navarro .....	809, 1464	Masaomi Oda .....	2742
Sushobhan Nayak .....	2079	Darko Odic .....	1126
Angela Nazarian .....	1482	Hidemi Ogasawara .....	2816
Aida Nematzadeh .....	2085	Keiji Ogata .....	2818
Kleanthis Neokleous .....	2091	Takashi Ogata 1272, 2126, 2150, 2670, 2699, 2738, 2817, 2922	
Hansjörg Neth .....	48	Hitoshi Ogawa .....	1650
Nora Newcombe .....	917	Masahide Ogawa .....	2816
Elissa Newport .....	881	Shinji Ogawa .....	1727
Fei Wan Ngai .....	2804	Shino Ogawa .....	2109, 2704
Stephen Nicholson .....	1656	Yukiko Ogawa .....	2819
Ulrike Niens .....	2351	William Oh .....	2896
Sumaru Niida .....	2805	Masato Ohba .....	2820
Toyoaki Nishida .....	2115	Yoshimasa Ohmoto .....	2115
Koshi Nishimoto .....	2806	Hitoshi Ohnishi .....	2821
Takehiko Nishimoto .....	2785	Takehiko Ohno .....	1179
Akio Nishimura .....	2807	Yukio Ohsawa .....	2799
Shuichi Nishio .....	12, 2375	Yuka Ohtake .....	2121
Ayumi Nishiyama .....	2797	Kayoko Ohtsu .....	2822
Takeshige Nishiyama .....	2809	Misato Oi .....	1924
Yuka Nishiyama .....	2808	Kensuke Oishi .....	2126
Eyal Nitzany .....	372	Natsuki Oka .....	2823, 2856
Yael Niv .....	36	Hiroyuki Okada .....	34, 2812
Kohei Noda .....	2810	Masato Okada .....	2737
Saori Noda .....	2792	Takeshi Okada .....	1733, 2073, 2321
Hisataka Noguchi .....	2811	Erina Okamoto .....	2375
Ikuya Nomura .....	2812	Junji Okamoto .....	2825
Ryota Nomura .....	2813	Jun Okamoto .....	2417
Naohiro Nonoguchi .....	2823	Masahiko Okamoto .....	2824
Elnaz Nouri .....	2097	Masashi Okamoto .....	2913
Haruka Nukariya .....	2775, 2814	Yasmina Okan .....	1143
Hiroshi Numata .....	2831	Kazuo Okanoya .....	2737

## Author Index

Shuntaro Okazaki .....	2690	Jaehyon Paik .....	2168
Tsukasa Okimura .....	2768	Bozena Pajak .....	2174
Hiroyuki Okuda .....	2892	Jo Pan .....	2650, 2834
Jiro Okuda .....	2735	Peter Pantelis .....	14
Motoki Okumura .....	2910	Chan Jeong Park .....	2836
Eve Okura .....	2132	Keun sik Park .....	2622
Aude Oliva .....	1302	Rae-yeop Park .....	2667
Jesus Oliva .....	1715	Soodam Park .....	2838
Daniel Olsher .....	174, 2138	Tae-Suh Park .....	2839
Akira Omaki .....	2826	Taejin Park .....	2838
Takanobu Omata .....	2144	Youjeong Park .....	2835
Takashi Omori .....	2812	Youngjun Park .....	2837
Jia Hoong Ong .....	839	Alexander Paunov .....	2840
Luca Onnis .....	815	David Pautler .....	14, 617
Hajime Ono .....	1488, 2847	Brennan Payne .....	1404
Junpei Ono .....	2738, 2922	Roy Pea .....	44
Kou Onodera .....	2150	Lisa Pearl .....	863
Isabel Orenes .....	575	David Peebles .....	38
Akiko Orita .....	2901	Reinhart Pekrun .....	44
Ryo Orita .....	2827	Matthew Pelowski .....	1924
Andrew Ortony .....	2204, 2210	Jeff Pelz .....	1900
Naoko Osada .....	2828	Conor Pendergrast .....	2351
Mariko Osaka .....	2718	Amy Perfors .....	809, 839, 845, 1464
Naoyuki Osaka .....	2718	Gilbert Peterson .....	1386
Daniel Osherson .....	1233	Georgi Petkov .....	1703
Magda Osman .....	48, 50, 2889	Rolf Pfeifer .....	30
Hiroko Otsuka .....	2831	Thies Pfeiffer .....	851, 2723
Kazunori Otsuka .....	2829	Nadine Pfeiffer-Lessmann .....	851
Sho Otsuka .....	2830	Kathie Pham .....	2180
Erin Ottmar .....	2156	Katherine Phelps .....	857, 2841
Yoshihiro Ouchi .....	2822	Lawrence Phillips .....	863
Long Ouyang .....	821	Steven Phillips .....	869, 2014
Yasuhiro Oyama .....	2794	Katrijn Pipijn .....	456, 2186
Kuratomo Oyo .....	2832	Peter Pirolli .....	2168
Norio Ozaki .....	1762	Markus Plank .....	2469
Motoyuki Ozeki .....	2823, 2856	Patrick Plummer .....	875
Asli Ozyurek .....	467	Kim Plunkett .....	737, 1990
Michael Pacer .....	827, 833, 2833	Amanda Pogue .....	2192
Shamin Padalkar .....	2162	Howard Poizner .....	2469
Sebastian Pado .....	1215	Frank Pollick .....	14

## Author Index

Marc Pomplun .....	875, 2499, 2505	Aaron Roberts .....	2246
Matthijs Pontier .....	2198	Timothy Rogers .....	2647
Maria Popova .....	1344	Douglas Roland .....	1191
Alexandre Pouget .....	36	Rosalba Rosato .....	2787, 2843
Matthew Purver .....	479, 1697	Steven Ross .....	1350
Ting Qian .....	881	Dan Roth .....	40
Boon-Kiat Quek .....	2204, 2210	Anselm Rothe .....	743
Ramsey Raafat .....	2842	Brandon Roy .....	935
Daniela Rabellino .....	2787, 2843	Deb Roy .....	935
Daniele Radicioni .....	1248	Anna-Mari Rusanen .....	941
Anna Rafferty .....	893	Matthew Rutledge-Taylor .....	2168
Marco Ragni .....	432, 1500, 2216	Jeong Ryu .....	2750, 2844
Rohan Raizada .....	731	Young Sun Ryu .....	2844
Suparna Rajaram .....	10	Rafal Rzepka .....	2845
Jason Ralph .....	899	Okko Räsänen .....	887
Jared Ramsburg .....	2222	Norihiro Sadato .....	797, 2690
Michael Ranney .....	2228	Eyal Sagi .....	2708
Heikki Rasilo .....	887	Antoine Saillenfest .....	947
Victor Raskin .....	450, 2393	Hirofumi Saito .....	1924
Olga Rass .....	78	Hitomi Saito .....	953
Keith Rayner .....	875	Moegi Saito .....	2258
Daniel Read .....	48	Motoyuki Saito .....	2264
Stephen Read .....	270	Satoru Saito .....	2802, 2888, 2902
Felix Rebitschek .....	905	Shoko Saito .....	2629
Elizabeth Redcay .....	312, 911	Noburo Saji .....	2820, 2846
Patricia Reeder .....	881	Hiromu Sakai .....	959, 1488, 2764, 2847
Terry Regier .....	60, 827	Rui Sakaida .....	2848
Raymond Reichenberg .....	150	Kayo Sakamoto .....	2745
Daniel Reinholz .....	2228	Maki Sakamoto .....	2270, 2475, 2517
Alexander Renkl .....	2345	Yasuaki Sakamoto ..	2387, 2648, 2753, 2849
Ilyse Resnick .....	917	Mamiko Sakata .....	2806, 2872
Hilary Richardson .....	923	Ruslan Salakhutdinov .....	659
Lindsey Richland .....	34	Ron Salden .....	2850
Cory Rieth .....	929	Muniba Saleem .....	521
Sean Riley .....	2234	Mennat-Allah Saleh .....	2276
Lance Rips .....	318	Anne Pier Salverda .....	1380
Frank Ritter .....	28, 2603	Kazuyuki Samejima .....	2812
Bethany Rittle-Johnson .....	348, 1260	Adam Sanborn .....	32, 485, 1356
Samuel Rivera .....	2240	Catherine Sandhofer .....	2631
Luigi Rizzi .....	2826	Wakako Sanefuji .....	2851

## Author Index

Ava Santos .....	2852	Satoshi Shibuya .....	2858
Kyoshiro Sasaki .....	2853	Richard Shiffrin .....	264, 527, 1440, 2713
Ayumi Sato .....	2282	Kenpei Shiina .....	2315
Kaori Sato .....	2856	Cecilia Shikuma .....	2771
Manami Sato .....	959, 2764, 2847	Eunji Shim .....	2859
Ryo Sato .....	2854	Hideaki Shimada .....	2860
Rebecca Saxe .....	276, 312, 623, 923, 2637, 2641, 2736, 2840, 2866	Koji Shimada .....	2690
Kohichi Sayama .....	1238, 2669	Sotaro Shimada .....	2861
Ayse Pinar Saygin .....	2469	Tsuneo Shimazaki .....	2264
Brian Scassellati .....	1266, 1882	Daichi Shimizu .....	2321
Gaia Scerif .....	2907	Takahiro Shimizu .....	2876
Walter Schaeken .....	456, 509, 2186, 2310	Katsunori Shimohara .....	2872
Lisa Scharrer .....	965	Atsushi Shimojima .....	2862
Matthias Scheutz .....	1072, 2288	Cheonwoo Shin .....	2748
Savannah Schilling .....	2294	HyunJung Shin .....	2696, 2711, 2725, 2748
Christos Schizas .....	2091	WonJae Shin .....	2863
Ute Schmid .....	396	Yong-Wook Shin .....	78
Martin Schmidt .....	1242	Noriko Shingaki .....	2864
Mike Schoelles .....	899	Kazuko Shinohara .....	569
Jordan Schoenherr .....	971	Shuji Shinohara .....	2876
Agnes Scholz .....	905, 2855	Thomas Shipley .....	917
Herbert Schriefers .....	1596	Hajime Shirouzu .....	2874, 2898
Laura Schulz .....	977, 2680	Inaba Sho .....	2881
Rolf Schwonke .....	2345	Marie Shoda .....	2865
Kathryn Sears .....	2300	Thomas Shultz .....	1452, 2327
Elizabeth Seiver .....	156	Michael Siebers .....	396
Ayumi Seki .....	2887	Basilio Sierra .....	2615
Takahiro Sekiguchi .....	2830	Cynthia Siew .....	983
Kazuki Sekine .....	2879	Cynthia Sifonis .....	2333
Allison Sekuler .....	2791	Les Sikos .....	2841
Takeharu Seno .....	2306, 2853	Noah Silbert .....	1840
J. Ignacio Serrano .....	1715	Clare Sims .....	246, 2294
Aline Sevenants .....	2310	Gale Sinatra .....	150
Patrick Shafto .....	977, 1614, 2680	Pawan Sinha .....	911
Sadbodh Sharma .....	1296	Scott Sinnett ...	294, 677, 1494, 1906, 1912, 2758, 2771
Miau-Lin Shen .....	2651	Kenny Skagerlund .....	2339
Pengcheng Shi .....	1900	Amy Skerry .....	2866
Masahiro Shibasaki .....	2704	Irene Therese Skuballa .....	2345
Ryoko Shibata .....	2856	Lloyd Slevc .....	911
Takuro Shibayama .....	2857	Vladimir Sloutsky ....	282, 420, 1750, 2240, 2587

## Author Index

Cybelle Smith .....	989	Hide nobu Sumioka .....	2375
Kevin Smith .....	995	Yanlong Sun .....	1024, 1120
Linda Smith .....	1197, 1209	Shoji Sunaga .....	2306
Tim Smith .....	2907	Satoru Suto .....	2818
Kirsty Smyth .....	2351	Masaki Suwa .....	2809, 2814, 2848, 2908
Myeong-ho Sohn .....	2696	Hiroaki Suzuki .....	2672
Masashi Soma .....	2653, 2867	Hiroo Suzuki .....	2910
Ji Son .....	34	Maki Suzuki .....	2735
Samuel Spaulding .....	1882	Noriko Suzuki .....	2872
Maarten Speekenbrink .....	50, 1001	Tatsuya Suzuki .....	2892
Elizabeth Spelke .....	2866	Yusuke Suzuki .....	1762
Joseph Spino .....	1368	Brian Sweis .....	1030
Michael Spivey .....	216	Marc Swerts .....	461, 2493
Marie-Ève St-Louis .....	2676, 2743	Daniel Swingley .....	72
Marc Stadtler .....	965	Colleen Szurkowski .....	2873
Tom Stafford .....	336	Martin Sälzle .....	2252
Maria Staudte .....	1007	Etsuko T. Harada .....	2629, 2805
Douglas Sterling .....	713	Tomohiro Taira .....	2381
David Stevens .....	1013	Yuisho Takafuji .....	2874
Jeffrey Stevens .....	84	Fumiyo Takahashi .....	2877
Suzanne Stevenson .....	2085	Kei Takahashi .....	2767, 2878, 2912
Terrence Stewart .....	22, 1018	Maiko Takahashi .....	2875
Mark Steyvers .....	108	Mikina Takahashi .....	2886
James Stigler .....	34	Satoshi Takahashi .....	2920
Elizabeth Stine-Morrow .....	1404	Taiki Takahashi .....	18
Gert Storms .....	1464	Tatsuji Takahashi ..	2731, 2832, 2857, 2876, 2917
Francesca Stramandinoli .....	2868	Keisuke Takahata .....	2768
Andreas Stuhlmüller .....	390, 1996	Masao Takaku .....	953
Andrew Stull .....	2869	Katsuya Takanashi .....	2879
Glenda Stump .....	150	Daisuke Takano .....	1739
Christian Sturm .....	2276	Yufuko Takashima .....	2880
Gabriel Stylianides .....	731	Nakayama Takayaro .....	2881
Katja Suckow .....	2357	Kazuya Takeda .....	2892
Yasutada Sudo .....	2870	Keiko Takeda .....	2882
Takashi Sugami .....	2871	Haruhiko Takeuchi .....	1632
Eriko Sugimori .....	2777	Yugo Takeuchi .....	2697, 2854
Masashi Sugimoto .....	2363, 2802	Masanori Takezawa .....	2702
Yoko Sugioka .....	611	Yasuyuki Taki .....	2767
Simone Sulpizio .....	2369, 2369	Yoshiyuki Tamamiya .....	2686, 2883
Miho Sumihisa .....	2270	Kaori Tamura .....	2884



## Author Index

Masahiko Tamura .....	2885	Richard Tillman .....	2434
Hiroki Tanabe .....	797, 2690	Satoshi Tojo .....	2774, 2798
Azumi Tanabe-Ishibashi .....	2886	Midori Tokita .....	2440
Akihiro Tanaka .....	2785	Akifumi Tokosumi .....	2894
Daisuke Tanaka .....	2887	Jackson Tolins .....	2446
Kanji Tanaka .....	1036	Mariya Toneva .....	1882
Keiji Tanaka .....	1739	Alexia Toskos Dils .....	2451
Kenshiro Tanaka .....	2801	Yuko Toukairin .....	2809
Yuko Tanaka .....	2387, 2769	Hristina Toushek .....	1703
Michael Tanenhaus .....	647, 1374, 1380	Wataru Toyokawa .....	2906
Hiroko Taniai .....	2109	Atuhito Toyomaki .....	2019
Yuki Tanida .....	2888	Greg Trafton .....	581
Masako Taniike .....	2851	Duc Tran .....	2457
Hirosuke Tanimoto .....	2701	David Traum .....	2097
Kameda Tatsuya .....	2906	Sebastien Tremblay .....	2676, 2743
Toshiharu Taura .....	2792, 2909	Christina Triantafyllou .....	312
Fumihiko Taya .....	2633	Jennifer Trueblood .....	18, 2895
Shuichiro Taya .....	2306, 2889	Chia-Liang Tsai .....	2651
Emiko Tayanagi .....	2890	Chiung-hsien Tsai .....	2900
Julia Taylor .....	450, 2393	Meng-Jung Tsai .....	2762, 2904
Amaro Taylor-Weiner .....	240	Chia-Huei Tseng ....	2659, 2766, 2836, 2896
Yuichi Tazaki .....	2892	Ricky Van Yip Tso .....	2463
Sergiu Tcaci Popescu .....	1042	Mio Tsubakimoto .....	2831, 2897
Thora Tenbrink .....	396	Tomoki Tsuchida .....	1048
Joshua Tenenbaum 14, 24, 32, 42, 378, 515, 659, 881, 923, 1996, 2680		Geroldene H. T. Tsui .....	2659, 2836
Katya Tentori .....	1233	Syoichi Tsujii .....	2898
Haruna Terada .....	2891	Hiroya Tsukurimichi .....	2517
Asuka Terai .....	2399, 2405, 2741	Yusaku Tsuyuguchi .....	2764
Hitoshi Terai .....	1960, 2008, 2411, 2892	Takashi Tsuzuki .....	1054, 2653
Shinako Terakawa .....	2887	Yukie Tsuzuki .....	2864
Takao Terano .....	2901, 2920	Barbara Tversky .....	14, 503, 551
Takehiro Teraoka .....	2417	Matthew Twyman .....	1001
Kejkaew Thanasuan .....	2893	Kristian Tylén .....	2673, 2899
Jean-Pierre Thibaut .....	384, 2632	Yuhtsuen Tzeng .....	875, 2900
EriK Thiessen .....	815	Ilya Tëmkin .....	2487
Serge Thill .....	2423	Hikaru Uchida .....	2901
Robin Thompson .....	1554	Hitoshi Uchiyama .....	2887
John Thoresen .....	1774	Ichiro Uchiyama .....	2282
Chris Thornton .....	2429	Kazuhiro Ueda .....	2812
Yi-Min Tien .....	2499	Kazutaka Ueda .....	2818

## Author Index

Satoshi Uemura .....	2805	Kum Seong Wan .....	617
Katsuyuki Ukai .....	1762	Xiaohong Wan .....	1739
Shimon Ullman .....	14	Ching-Yeh Wang .....	2904
Tomer Ullman .....	32, 1996	Hongbin Wang .....	1024, 1120
Ichiro Umata .....	2872	Hsueh-Cheng Wang .....	875, 2499, 2505
Hiroyuki Umegaki .....	1762	Pei-Ling Wang .....	2646
Takatoyo Umemoto .....	2801	Pei Wang .....	1242
Burcu Aysen Urgan .....	2469	Nicole Ware .....	2511
Saori Ushiyama .....	2825	Yuichi Washida .....	2812
Shinnosuke Usui .....	2622	Jonathan Waskan .....	1368
Akira Utsumi .....	797, 2270, 2381, 2475	Motoki Watabe .....	2794
Kenta Utsumi .....	2902	Junji Watanabe .....	2517
Yuuka Utsunomiya .....	2517	Katsumi Watanabe .....	432, 1036
Frederic Vallee-Tourangeau .....	1060	Masataka Watanabe .....	50
Kees van Deemter .....	1584, 1834	Shigeru Watanabe .....	2633
Laurens van der Maaten .....	671	Yoriko Watanabe .....	2882
Roger P. G. van Gompel .....	1584, 2357	Steven Weisberg .....	46
Janet van Hell .....	10	Stephen Welbourne .....	198
Kurt Van Lehn .....	44	Sarah Wellen .....	2523, 2529
Koen van Lierop .....	1066	Alexis Wellwood .....	1126
Martin van Velsen .....	1260	Matthew Welsh .....	1131
Meryl Varadinov .....	1858	Markus Werning .....	1137
Richard Veale .....	1072	Veronica Weser .....	2905
Bella Veksler .....	2481	Robert West .....	1768, 2535
Vladislav Veksler .....	1078	Gert Westermann .....	2650, 2834
Tomas Veloz .....	2487	Mark Wexler .....	1042
Rineke Verbrugge .....	1290	Andrew Whalen .....	156
Jette Viethen .....	1084	Alex Wiegmann .....	1102, 1143
Mandy Visser .....	2493	Jan Wiener .....	2216
Vy Vo .....	2903	Joseph Jay Williams .....	1114, 1149
Momme von Sydow .....	1090	Mark Williams .....	1013
Stella Vosniadou .....	44	Andy Wills .....	1185, 2581
Edward Vul .....	42, 929, 995	Theodore Wills .....	917
Annalies Vuong .....	438	Deny Willy .....	2792
Ipke Wachsmuth .....	851, 1721, 2723	Christy Wilson .....	2852
Katie Wagner .....	1096	Colin Wilson .....	52, 587
Petra Wagner .....	1721	William Wilson .....	869
Michael Waldmann .....	785, 1102	Amy Winchester .....	971
Caren Walker .....	1108, 1114, 1149	David Windridge .....	2889
Dirk Walther .....	2240	Anders Winman .....	2760

## Author Index

Edwin Wirawan .....	617	Michael C. W. Yip .....	2916
Thomas Wisdom .....	2906	Junki Yokokawa .....	2917
Ewelina Wnuk .....	1155	Kazuhiko Yokosawa .....	2807, 2865
Yetta Kwailing Wong .....	2540	Keiko Yokoyama .....	2910
Elizabeth Woods .....	2546	Satoru Yokoyama ...	2767, 2878, 2912, 2918
Jennifer Wu .....	959	Ali Yoonessi .....	2636
Rachel Wu .....	1161, 1203, 2907	Hanako Yoshida .....	2457, 2546, 2642
Yi-Rong Wu .....	2650, 2834	Atsushi Yoshikawa .....	2901, 2920
Dirk Wulff .....	1167	Masato Yoshikawa .....	2919
Chengli Xiao .....	1173	Sakiko Yoshikawa .....	2794
Hiroshi Yama .....	2552, 2800	Satomi Yoshiyama .....	2683
Akiko Yamada .....	2622	Sayuri Yoshizawa-Watanabe .....	2920
Ayumi Yamada .....	2672	Robert Youmans .....	2222
Kaori Yamada .....	2909	Liane Young .....	312, 623
Masayuki Yamada .....	2908	Chen Yu .....	527, 1209
Seiji Yamada .....	1816	Erica Yu .....	1458
Yuki Yamada .....	2306, 2853	I-Wen Yu .....	2646
Kimihiko Yamagishi .....	2405, 2741	Na-Yung Yu .....	1852
Naohide Yamamoto .....	2558	Yue Yu .....	2921
Tomoka Yamamoto .....	2851	Hongoak Yun .....	1191
Yasuhiro Yamamoto .....	2796	Daniel Yurovsky .....	1197, 1203, 1209
Yuji Yamamoto .....	2910	Matei Zaharia .....	893
Takashi Yamauchi .....	1179, 1392, 1852	Nasriah Zakaria .....	2800
Mika Yamazaki .....	797	Norhayati Zakaria .....	2552, 2800
Myung-Jin Yang .....	2911	Andrew Zaldivar .....	90
Ruixin Yang .....	2564	Alessandra Zarcone .....	1215
Yuri Yano .....	2888	Byoung-Tak Zhang .	1221, 1864, 2722, 2723, 2839
Jit Yong Yap .....	2570	Lu Zhang .....	1227, 2593
Melvin Yap .....	983	Minsu Zhang .....	2726
Evelyn Yarzebinski .....	731	Niina Ning Zhang .....	2599
Ken Yasaka .....	2912	Shunan Zhang .....	90
Sachi Yasuda .....	2627, 2913	Yao Zhang .....	2609
Tetsuya Yasuda .....	2576	Yike Zhang .....	2922
Lindsay Yazzolino .....	2637	Changkun Zhao .....	2603
Fayme Yeates .....	1185, 2581	Jiaying Zhao .....	1233
Hsiang-Chun Yeh .....	2914	Desislava Zhekova .....	1972
Jaclyn Yeung .....	2766	Sheng-hua Zhong .....	2609
Saiwing Yeung .....	2915	Tom Ziemke .....	2423
Kwangoh Yi .....	2635	Iraide Zipitria .....	2615
Hyungwook Yim .....	2240, 2587	Jean Écalle .....	2767

## Reviewers

Akinori Abe	Klinton Bicknell	Nicholas Cassimatis
Caspar Addyman	Dorrit Billman	Cristiano Castelfranchi
Gabriella Airenti	Nik Nailah Binti Abdullah	Tom Caudell
John Alderete	Ricardo Bion	Nicholas Cepeda
Silvio Aldrovandi	Tamas Biro	Bhisma Chakrabarti
Afra Alishahi	Agnes Blaye	Cheri Chan
Erik Altmann	Stephen Blessing	Joel Chan
Franco Amati	Aysecan Boduroglu	Margaret Chan
Ben Ambridge	Elizabeth Baraff Bonawitz	Franklin Chang
Sarah Anderson	Jean-Francois Bonnefon	Maria Chang
Elena Andonova	Anna Borghi	Allison Chapman
Glenda Andrews	Jelmer Borst	Julia Chariker
Mark Andrews	Lewis Bott	Jenn-Yeu Chen
Florencia Anggoro	Roberto Bottini	Qi Chen
Dagmara Annaz	Jean-Michel Boucheix	Jane Childers
Jo Arciuli	Will Bridewell	Jessie Chin
Richard Ashley	Henry Brighton	Seth Chin-Parker
Richard Aslin	Stephen Briner	Eric Chiu
Marios Avraamides	Erica Briscoe	Dongkyu Choi
Harald Baayen	Andrew Brook	Ruth Church
Jonathan Back	Geoffrey Brookshire	Bill Clancey
Jonathan Bakdash	Duncan Brumby	John Clapper
Chris Baker	Angela Brunstein	Eve Clark
Joseph Baker	Tad Brunye	Timothy Clausner
Ryan Baker	Adam Bryant	Catherine Clement
Alan Bale	David Buchanan	Charles Clifton
Jerry Ball	Leandra Bucher	James Close
Raju Bapi	Simon J. Buechner	Uriel Cohen Priva
Thomas Barkowsky	Marc Buehner	Cheryl Cohen
Mike Barley	Ann Bunger	Lucia Colombo
Merja Bauters	Bruce Burns	Louise Connell
Philip Beaman	Mark Burstein	Rick Cooper
Melissa Beck	Jerome Busemeyer	Kathleen Corriveau
Anna Belardinelli	Kirsten Butcher	Rui Costa
Daniel Belenky	Ruth Byrne	Gary Cottrell
Paul Bello	Cristina Cacciari	Anna Cox
Andrea Bender	Paul Cairns	Charlie Cox
Viridiana Benitez	Paolo Canal	Gregory Cox
Francesca Benuzzi	Angelo Cangelosi	Michael Cox
Matthew Bernacki	Richard Carlson	Jed Crandall
Sven Bertel	Claudia Casadio	Christopher Crick
Luc Berthouze	Daniel Casasanto	Matthew Crossley
Brad Best	Marianella Casasola	Fred Cummins

## Reviewers

Evan Curtis	Igor Farkas	Charlotte Gerritsen
Rick Dale	Thomas Farmer	Samuel Gershman
Gregory Dam	Afsaneh Fazly	Tobias Gerstenberg
Frederic Dandurand	Aidan Feeney	Valeria Giardino
David Danks	Michele Feist	Gyslain Giguere
Drew Dara-Abrams	Naomi Feldman	Nicholas Giudice
Kasmira Dave	Gary Feng	Marco Giunti
Eddy Davelaar	Ron Ferguson	Kuba Glazek
Jim Davies	Evelyn Ferstl	Vladimir Glebkin
Colin Dawson	Katja Fiehler	Ashok Goel
Nadja De Carolis	David Field	Rob Goldstone
Paul De Palma	Mark Finlayson	Micah Goldwater
Virginia de Sa	Sara Finley	Laura Gonnerman
Marci DeCaro	Stanka Fitneva	Cleotilde Gonzalez
Morteza Dehghani	Hartmut Fitz	Geoff Goodwin
Krista DeLeeuw	Robin Flanagan	Paula Goolkasian
Gary Dell	Ken Forbus	Sandy Gould
Jean-Louis Dessalles	Tom Foulsham	Katharine Graf Estes
Marc Destefano	Ana Franco-Watkins	Collin Green
Andrew Dewald	Mike Frank	Harry Griffin
Fred Dick	Stefan Frank	Tom Griffiths
Denise Dillon	Nancy Franklin	Maurice Grinberg
Sidney DMello	Brandy Frazier	Amal Guha
Sarah Dolscheid	Bob French	Glenn Gunzelmann
Wei Dong	Daniel Freudenthal	Todd Gureckis
Tim Donovan	Marcello Frixione	Lilia Gurova
Scott Douglass	Caren Frosch	Hyowon Gweon
Leonidas Doumas	Wai-Tat Fu	Marcus Haag
John Drury	Danilo Fum	York Hagmayer
Chad Dube	Rami Gabriel	Lance Hahn
Nicholas Duran	Soniya Gadgil	Ulrike Hahn
Varun Dutt	Annie Gagliardi	Harry Haladjian
Kathleen Eberhard	Francesco Gagliardi	Graeme Halford
Alexander Eitel	Alexia Galati	Tim Halverson
Chris Eliasmith	Jean-Gabriel Ganascia	James Hampton
Michelle Ellefson	Amelia Gangemi	J. Isaiah Harbison
Stern Elsbeth	Raquel Garcia Jurado	Jack Harris
Paul Engelhardt	Simon Garrod	Kevin Harris
Eileen Entin	Ross Gayler	Anthony Harrison
Selda Eren Kanat	Hector Geffner	Joshua Hartshorne
Orlando Espino	Rosella Gennari	Guy Hawkins
Zachary Estes	Dedre Gentner	Brett Hayes
Martha Evens	LouAnn Gerken	Mary Hegarty

## Reviewers

Evan Heit	Frank Jäkel	Trent Kriete
Steve Hekkanen	Linda Kaastra	Sarah Kucker
Sebastien Helie	George Kachergis	Sven Kuehne
Michael Helms	Solene Kalenine	Maithilee Kunda
Michelle Hendricks	Deepthi Kamawar	Kenneth Kurtz
Tania Henetz	Frank Kanayet	Petko Kusev
Larry Hettinger	Vsevolod Kapatsinski	Mee-Kyoung Kwon
Thomas Hills	Themelis Karaminis	Kai-Uwe Kühnberger
Alexandra Horowitz	Michael Kaschak	David Lagnado
William Horton	Irvin Katz	Polly Lai
Andrew Howes	Artem Kaznatcheev	Tei Laine
Evgenia Hristova	Mark Keane	Brenden Lake
Penka Hristova	John Kearns	Kiran Lakkaraju
Roland Hubscher	Madeleine Keehner	David Landy
Bryce Huebner	Charles Kemp	Peter Lane
Alycia Hund	Vera Kempe	Nicholas Lange
Julie Hupp	William Kennedy	Patrick Langley
Scott Husband	Chris Kent	Lyuben Laskin
Fehmida Hussain	Liesbeth Kester	Alessandro laudanna
Jo Iacovides	Muhammad Ali Khalidi	David Leake
Wayne Iba	Sangeet Khemlani	Christian Lebiere
Mutsumi Imai	Naveen Khetarpal	N.Y. Louis Lee
Lisa Irmen	Peter Khooshabeh	Cristine Legare
Tanner Jackson	Celeste Kidd	Blair Lehman
Robert Jacobs	Dahee Kim	Benoit Lemaire
Michael Jacobson	Jihie Kim	Itamar Lerner
Georg Jahn	Walter Kintsch	Joe Levy
Azadeh Jamalian	Szabolcs Kiss	Roger Levy
Christian Janssen	Mikhail Kissine	Simon Levy
Kyle Jasmin	Sachiko Kiyokawa	Mark Lewis
Francis Jeanson	Matthew Klenk	Molly Lewis
Benjamin Jee	Alexander Klippel	Stephen Lim
Charlene Jennett	Markus Knauff	Craig Lindley
Patrick Jeuniaux	Pia Knoeferle	Robert Lindsey
William Jimenez-Leal	Ken Koedinger	Jordan Lippman
Remo Job	Judith Koehne	Damien Litchfield
Gary Jones	Bryan Koenig	Daniel Little
Joshua Jones	Boicho Kokinov	Lei Liu
Martin Jonsson	Piotr Konderak	Taosheng Liu
Jerome Scott Jordan	Stefan Kopp	Yang Liu
Kerry Jordan	Emiel Krahmer	Deryle Lonsdale
David Joyner	Adam Krawitz	Sander Los
Ion Juvina	Josef Krems	lorella lotto

## Reviewers

Ted Lougheed  
Brad Love  
Francis Lowenthal  
Christopher Lucas  
George Luger  
Christian Luhmann  
Gary Lupyan  
Dermot Lynott  
Karl MacDorman  
Naoki Maeda  
Brian Magerko  
Lorenzo Magnani  
David Majerich  
Steve Majerus  
Laurence Maloney  
Emmanuel Manalo  
Jaison A. Manjaly  
Wenji Mao  
Daniela Mapelli  
Art Markman  
Goran Martinovic  
Francesca Marzo  
Michael Matessa  
Emily Mather  
Noboru Matsuda  
Danielle Matthews  
Camillia Matuk  
Luke Maurits  
Andre Mayers  
Ralf Mayrhofer  
Devin McAuley  
Kelly McCormick  
John McCoy  
Keith McGreggor  
Marjorie McShane  
Bjoern Meder  
Robin Melnick  
Zulfiqar Ali Memon  
David Mendonca  
Stefania Mereu  
Everett Mettler  
Craig Miller  
Jelena Mirkovic

Daniel Mirman  
Robert Mitchell  
Padraic Monaghan  
Daniel Montello  
Laura Morett  
Emily Morgan  
Junya Morita  
Bradley Morris  
Anthony Morse  
Doug Morse  
Jarrod Moss  
Claudio Mulatti  
Paul Munro  
Dafne Muntanyola  
Gregory Murphy  
Christopher Myers  
Daniele Nardi  
Eduardo Navarrete  
Daniel Navarro  
Jonathan Nelson  
Hansjoerg Neth  
Hartmut Neuf  
Nora Newcombe  
Elissa Newport  
Wendy Newstetter  
Mark Nieuwenstein  
David Noelle  
Timothy Nokes-Malach  
Michael Noll-Hussong  
Ann Nordmeyer  
Timothy O'Donnell  
Padraig O'Seaghdha  
Mike Oaksford  
Tim Oates  
Natalie Obrecht  
Eric Odgaard  
Marta Olivetti  
Luca Onnis  
Santiago Ontañón  
Christine Otieno  
Michael Pacer  
Ulrike Pado  
Bozena Pajak

Massimiliano Palmiero  
Melanie Palomares  
John Pani  
Peter Pantelis  
Anna Papafragou  
Praveen Paritosh  
Woosuk Park  
Gary Parker  
David Pautler  
Francesca Pazzaglia  
Lisa Pearl  
Neal Pearlmutter  
Anna Pecchinenda  
David Peebles  
Alfredo F. Pereira  
Francesca Peressotti  
Amy Perfors  
Mark Perry  
Francesca Pesciarelli  
Eliano Pessa  
Kathrine Petersen  
Eric Peterson  
Georgi Petkov  
Steven Phillips  
Julian Pine  
Kathleen Pirog Revill  
Timothy Pleskac  
Raffaella Pocobello  
Emmanuel Pothos  
Christopher Power  
Janani Prabhakar  
Merce Prat-Sala  
Athanassios Protopapas  
Dagmar Provijn  
Cynthia Putnam  
Lisa Putzar  
Jennie Pyers  
David Pynadath  
Carolyn Quam  
Jennifer Queen  
Philip Quinlan  
Joanna Raczaszek-Leonardi  
Marco Ragni

## Reviewers

Kiruthika Ramanathan	Mike Schoelles	Michelle St Clair
Michael Ramscar	Gregor Schoener	Marc Stadler
Raj Ratwani	Christoph Schommer	Laura Staum Casasanto
Martina Rau	Lael Schooler	Mark Steedman
Florencia Reali	Holger Schultheis	Courtney Stein
Paul Reber	Anne Schutte	Amanda Stent
Stephen Reed	Angela Schwering	Neil Stewart
Jason Reiss	Rolf Schwonke	Terry Stewart
David Reitter	Claudia Scorolli	Frederik Stjernfelt
Russell Revlin	J. Ignacio Serrano	Gert Storms
Marjorie Rhodes	Carissa L. Shafto	Laurie Stowe
Rebecca Rhodes	Meredith Shafto	David Straczewski
Mike Richardson	Patrick Shafto	Bill Stubblefield
Elizabeth Richey	Priti Shah	Diana Su Yun Tham
Lindsey Richland	Michael Shelton	Jess Sullivan
Kai-Florian Richter	Hajime Shirouzu	Tom Sullivan
Chris Riesbeck	Anthony Shook	Ron Sun
Monica Riordan	Al Shorin	Julia Svoboda
Steven Ritter	Thomas Shultz	Bill Swartout
Debi Roberson	Mei Si	Germaine Symons
Etienne Roesch	Paul Siakaluk	Whitney Tabor
Hannah Rohde	Noah Silbert	Makoto Takemiya
Katharina Rohlfing	Eli Silk	Marco Tamburelli
Annelie Rothe	Massimo Silvetti	Yi-Yuan Tang
Fred Rothganger	Luca Simione	Holly Taylor
Benjamin Rottman	Clare Sims	Joshua Taylor
Caroline Rowland	Raj Singh	Julia Taylor
Brandon Roy	Scott Sinnett	Virginia Teller
Gennaro Ruggiero	Ut Na Sio	Louis ten Bosch
Maki Sakamoto	Anna Siyanova	Thora Tenbrink
William Sakas	Katrin Skoruppa	Josh Tenenbaum
Joanna Salapska-Gelleri	Steve Sloman	Jean-Pierre Thibaut
Nancy Salay	Vladimir Sloutsky	Michael Thomas
Ron Salden	Kenny Smith	Clarissa Thompson
Catherine Sandhofer	Marie Smith	Robin Thompson
Ricardo Sanz	David Sobel	Marisa Tice
Karanam Saraschandra	Melanie Soderstrom	Ida Toivonen
Brian Scassellati	Fabian Soto	Emmett Tomai
Paul Schermerhorn	Ann Speed	Brendon Towle
Matthias Scheutz	Jennifer Spenader	Greg Trafton
Franz Schmalhofer	Brian Spiering	Jan Treur
Ute Schmid	Simone Sprenger	Trin Turner
Hedda Rahel Schmidtke	Vishnu Sreekumar	Barbara Tversky



## Reviewers

Ryan Tweney  
Christina Tzeng  
Shoukat Ullah  
Tomer Ullman  
M. Afzal Upal  
Oleg Urminsky  
David Uthus  
Akira Utsumi  
Natalie van der Wal  
Emile van der Zee  
Filip Van Opstal  
Jacolien van Rij  
Hedderik van Rijn  
Marieke van Vugt  
Joachim Vandekerckhove  
Ivan Vankov  
Donald Alexander Varakin  
Argiro Vatakis  
Bella Veksler  
Tom Verguts  
Francesco Vespignani  
David Vinson

Jonathan Vitale  
Kai Vogeley  
John Voiklis  
Lisa von Stockhausen  
Stella Vosniadou  
Soroush Vosoughi  
Ed Vul  
Michael Waldmann  
Anne Warlaumont  
Walter Warwick  
Yana Weinstein  
Sarah Wellen  
Matthew Welsh  
Stefan Wierda  
Geraint Wiggins  
Andy Wills  
Nicholas Wilson  
William Wilson  
Bryan Wiltgen  
Carsten Winkelholz  
Ralph Wojtowicz  
Michael Wolfe

Sharon Wood  
James Woodson  
Ruth Wylie  
Stefan Wölfl  
Naohide Yamamoto  
Takashi Yamauchi  
Lee-Xieng Yang  
Mark Yates  
Michael Yip  
Robert Youmans  
Christopher Young  
Richard Young  
Jiwon Yun  
Diego Zapata  
Alessandra Zarcone  
Franklin Zaromb  
Matt Zeigenfuse  
Corinne Zimmerman  
Iraide Zipitria  
Filio Zourou